Notas de Microeconometría Aplicada

Autor: Arturo A. Aguilar Esteva, Colaborador: Vicente López Ramírez

Índice general

Notas de Microeconometría Aplicada

Estas notas fueron desarrolladas por el profesor Arturo A. Aguilar Esteva como material didáctico de apoyo para los cursos de *Microeconometría Aplicada* que imparte a nivel licenciatura y maestría en el Instituto Tecnológico Autónomo de México (ITAM). Todos los errores son su responsabilidad. Cualquier comentario y observación con respecto a errores se agradecerá que lo dirijan al correo arturo.aguilar@itam.mx.

Capítulo 1

Introducción

La **econometría** consiste, principalmente, en la aplicación de métodos estadísticos. Su uso se ha extendido ampliamente en los últimos años debido a que la disponibilidad de datos ha crecido exponencialmente en tiempos recientes. Su aplicabilidad abarca una amplia gama de campos que van desde la economía, políticas públicas, salud y muchos otros. El propósito de estas notas es guiar el planteamiento adecuado de un **análisis empírico**. Haremos énfasis en el alcance y limitación de dichos análisis, distinguiendo bajo que condiciones debe ser entendido como un **análisis descriptivo** y cuando como una **estimación causal**.

En las primeras secciones exploramos herramientas metodológicas comúnmente empleadas para llevar a cabo un análisis empírico. Empezamos por hacer una breve revisión de términos estadísticos importantes. Dedicamos varias páginas al método de **Mínimos Cuadrados Ordinarios**, siendo la herramienta econométrica más comúnmente utilizada al hacer análisis que involucra explorar la relación entre un resultado y una o muchas variables explicativas. Detallamos también cómo incorporar estructuras más complejas de información, como los **datos panel**. Con ello, revisamos estimaciones de primeras diferencias, efectos fijos y efectos aleatorios. Después, exploramos los **métodos de máxima verosimilitud** que, en su mayoría, se caracterizan por utilizar como resultado una variable categórica. Finalizamos esta primera parte de las notas revisando algunos métodos no paramétricos, como son las **densidades y regresiones kernel**. Al revisar estos distintos métodos establecemos las bases necesarias para la segunda parte de estas notas.

El segundo grupo de secciones se enfoca en el planteamiento adecuado de un **análisis causal**. Frecuentemente, las preguntas de mayor interés en economía aplicada suelen involucrar el análisis del efecto causal que tiene una variable sobre otra. Por ejemplo, podemos estar interesados en el efecto de la implementación de una política educativa sobre la escolaridad, el efecto de shocks

en el tipo de cambio sobre decisiones de consumo de los hogares, o el efecto de aplicar fertilizante sobre la eficiencia en producción agrícola de los hogares. Para esto, es importante poder aislar el efecto del cambio en una variable sobre otra, evitando que cambios en variables relacionadas puedan afectar o *confundir* dicho análisis. Para ello, empezamos por explorar los **experimentos aleatorizados**, entendidos como el método que *econométricamente* logra establecer dicha relación causal de la forma más limpia posible. Posteriormente, entendiendo que los experimentos aleatorizados no son adecuados o posibles de implementar en muchos contextos, dedicamos los últimos capítulos a explorar métodos **cuasi-experimentales**. Vemos como los métodos de **variables instrumentales**, **diferencias en diferencias** y **regresión discontinua** plantean distintos *supuestos* para simular la existencia de un experimento y asi estimar efectos causales. En cada capítulo respectivo, buscamos establecer claramente los supuestos y limitaciones de estos métodos para promover un uso responsable y riguroso de ellos.

Busco con estas notas transmitir que la econometría, utilizada de forma correcta, es una herramienta que tiene el potencial de dar evidencia que contribuya a avanzar nuestro conocimiento. Nos puede permitir entender la relación entre variables relevantes: ¿cuál el efecto que un año adicional de escolaridad puede tener sobre el ingreso?, ¿cómo choques de salud en la infancia pueden afectarnos para toda la vida?, ¿cómo podemos estimar una elasticdad de la demanda? Puede tambien ayudarnos a basar en evidencia bien construida la implementación de políticas públicas, estrategias de marketing y programas de ONGs para poder tomar decisiones informadas.

Capítulo 2

Repaso de Estadística

La base necesaria para la econometría consiste en tener un entendimiento importante de algunos fundamentos de **probabilidad** e **inferencia estadística**. Esta sección hace un repaso no exhaustivo de algunos conceptos fundamentales que emplearemos. Se recomienda, sin embargo, que aquellos que no se sientan cómodos con alguos fundamentos básicos hagan una revisión de dichos conceptos para fortalecer su entendimmento de los temas vistos posteriormente. Crash course statistics es una referencia útil y fácilmente accesible para aquellos que quieran hacer un repaso de conceptos básicos de estadística.

2.1. Probabilidad

Esta sección consiste es un repaso muy breve y general de conceptos de probabilidad que se utilizarán a lo largo del curso[1]. En general, cuando hablamos de probabilidad es usual distinguir entre variables discretas y continuas. En términos prácticos, las variables discretas son aquellas que toman un pequeño número de valores posibles. Por su parte, las variables continuas son aquellas que pueden tener un número infinito de valores posibles. Usualmente, se les identifica como variables que toman un valor dentro de cierto rango de valores posibles.

Las variables suelen ser descritas por una **función de densidad** (**pdf**), la cual describe la probabilidad de que una variable tome cierto valor: $f(x_j) = P(X = x_j) = p_j$. Dicha función de densidad suele estar ligada a la **función de densidad acumulada** (**cdf**), que describe la probabilidad de que una variable tenga un valor menor o igual a cierto número: $F(x_j) = P(X \leq x_j)$.

Cuando involucramos en nuestro análisis más de una variable, es común hacer referencia a la **función de densidad conjunta**: $f_{X,Y}(x_j, y_k) = P(X = x_j, Y =$

[1] Aquellos alumnos que sientan que necesitan un repaso más a detalle, se recomienda que revisen el capítulo 2 del *Stock y Watson* o el apéndice B del *Wooldridge*.

y_k) y a la **función de densidad marginal**: $f_X(x_j) = P(X = x_j)$. En el caso en que las variables X y Y sean independientes: $f_{X,Y} = (x_j, y_k) = f_X(x_j)f_Y(y_k)$. Asimismo, con más de una variable, será común referirnos a la **densidad condicional**: $f_{Y|X}(y_k|x_j) = \frac{f_{X,Y}(x_j,y_k)}{f_X(x_j)}$. [En adelante, para simplificar únicamente utilizaremos x en vez de x_j y y en vez de y_k. El único propósito de utilizar x_j y y_k era para indicar que son valores específicos, i.e. algún número. En adelante, cuando utilicemos minúsculas estaremos haciendo referencia a valores específicos.]

Valor esperado. El valor esperado es una medida de tendencia central. Si X es una variable discreta que toma k distintos posibles valores, tendremos que:

$$E[X] = \sum_{j=1}^{k} x_j f(x_j)$$

Asimismo, en el caso continuo:

$$E[X] = \int_{-\infty}^{\infty} x f(x)$$

Propiedades del valor esperado:

- $E(a) = a$, donde a es una constante
- $E(aX + bY + c) = aE(X) + bE(Y) + c$, donde a, b, c son constantes

Varianza. La varianza describe, en promedio, que tan lejos suele estar una variable del valor esperado ($\mu_X = E(X)$).

$$\sigma_X^2 = Var(X) = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

La desviación estándar es simplemente:

$$\sigma_X = \sqrt{Var(X)}$$

Propiedades de la varianza:

- $Var(a) = 0$
- $Var(aX + b) = a^2 Var(X)$
- $Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm abCov(X, Y)$

Covarianza. La covarianza mide la relación entre dos variables. Una covarianza positiva indica que ambas variables suelen moverse en la misma dirección. Una covarianza negativa indica lo contrario ($\mu_Y = E(Y)$)

$$\sigma_{X,Y} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Propiedades de la covarianza:

- $Cov(X, Y) = 0$ si X y Y son independientes
- $Cov(aX + b, cY + d) = acCov(X, Y)$, donde a, b, c, d son constantes

El coeficiente de correlación es simplemente: $\rho_{X,Y} = corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

2.2. Notación

El análisis econométrico empírico generalmente involucra llevar a cabo una inferencia a partir de los datos a los cuales se tiene acceso (la *muestra*) acerca del comportamiento de cierta población. Obtener muestras representativas de la población acerca de la cual se busca aprender algo es fundamental y suele ser en un tema a desarrollar por sí solo. A lo largo del curso daremos esta pregunta por resuelta y supondremos que tenemos acceso a datos de muestras representativas de cierta población.

La **población** es un grupo definido de *unidades de observación* que pueden estar o no restringidas de distintas formas (por ejemplo: geográfica, económica o demográficamente). Una *unidad de observación* puede ser desde una acción, como por ejemplo, un mensaje de texto enviado vía celular o una ruta específica de un punto a otro que una persona se traslada, hasta un ente con una identidad específica que puede ser desde individuos, empresas, ciudades o países. En resumen, la población es el grupo objetivo que se pretende estudiar. En gran parte de los estudios suele establecerse que nos interesa conocer algún aspecto específico de la población, típicamente un estadístico. Por ejemplo, es común indicar que nos interesa conocer la media de una variable o la correlación entre dos variables de dicha población. Siempre que nos interese un estadístico en específico, nos referiremos a éste como el parámetro poblacional de interés.

Por su parte, la **muestra** es un subconjunto de la población el cual será utilizado en el análisis empírico para poder decir algo acerca de la población. La muestra será aquella de la cual recopilaremos información que tendremos concentrada en una *base de datos* que utilizaremos con programas estadísticos para llevar a cabo el análisis. La inferencia generalmente conlleva hacer una estimación utilizando datos de la muestra con el objetivo de obtener conclusiones acerca de la población. Por ejemplo, utilizando diversas estrategias se plantean *estimadores* del parámetro poblacional, mismos que emplean datos de una muestra para dar evidencia e información específica de dicho parámetro.

Es importante conocer los detalles de la estrategia utilizada para recabar la muestra, ya que esto es indicativo de si dicha muestra es representativa de la población que se pretende[2]. Algunas causas de sesgos muestrales comunes incluyen sesgo de respuesta (i.e. el hecho de que las personas que no contestan no son

[2]Para conocer mas acerca de muestreo se recomienda ver el video https://youtu.be/RffIpB4D50

personas al azar), problemas comprendiendo las preguntas y problemas encontrando ciertos tipos de individuos en la población. Estos problemas suelen llevar a sesgos muestrales, lo cual significa que la muestra puede no ser representativa de la población objetivo. Para evaluar esto, es útil tener una base de datos de comparación (i.e. un *benchmark*) que te permita determinar si las características de tu muestra son similares a aquellas características de la población objetivo. Los censos nacionales son un buen ejemplo de este tipo de *benchmarks*.

En estadística, generalmente nos referiremos al uso de muestras aleatorias debido a que eso nos asegura representatividad. En las muestras aleatorias cada uno de sus componentes se selecciona de forma independiente y proviene de una distribución común $\{Y_1, \dots, Y_n\}$. En este caso se dice que Y_i es una **variable aleatoria independiente e idénticamente distribuida** (**i.i.d.**). Las variables aleatorias $\{Y_1, \dots, Y_n\}$ son variables desconocidas. Una vez que la muestra es recabada tendremos un conjunto de números $\{y_1, \dots, y_n\}$, los cuales utilizaremos para llevar a cabo inferencia.

Ejemplo: Supón que nuestra variable de interés es la media de la edad de los alumnos cursando educación superior en México.

- ¿Cuál es la población?
- ¿Cómo recabarías una muestra para poder obtener/hacer una estimación válida?
- ¿Qué sucedería si yo recabo una muestra basada en los alumnos del ITAM o de la clase?

A continuación buscamos utilizar nuestra muestra para conocer algo acerca del parámetro. Para ello plantearemos un estimador. El **estimador** en sí es una fórmula que utiliza como insumo los elementos de la muestra y entrega como resultado un valor que pretende aproximar al parámetro poblacional objetivo. Es incorrecto pensar en que el estimador tiene valores específicos (a pesar de que el parámetro poblacional si lo tenga) dado que el estimador en sí es una estrategia que se plantea para aproximar lo mejor posible al parámetro poblacional. Al utilizar los valores específicos de la base de datos a la cual tenemos acceso y aplicar la fórmula del estimador para obtener un valor específico, estaremos hablando de un **valor estimado**. Para calificar si un estimador es apropiado y si es mejor o peor que otro, solemos hablar de propiedades que nos ayudan a calificarlo, como **insesgadez** y **eficiencia**.

Ejemplo: Continuando con el ejemplo de la media de edad de alumnos en educación superior en México: supongamos a partir de aquí que se recaba una muestra aleatoria $\{Y_1, \dots, Y_n\}$

- Parámetro: μ. Es un número específico que típicamente no es conocido. En este caso, es la media de la edad de todos los alumnos cursando educación media superior en México.

- Estimador: $W = h(Y_1, \ldots, Y_n)$. Es una fórmula que genera un estimador del parametro.
- Valor estimado: $w = h(y_1, \ldots, y_n)$. Es un número específico que resulta de aplicar la fórmula del estimador a la muestra recabada.

Definimos a continuación las propiedades deseables de los estimadores:

1. El estimador (W) de un parámetro (μ) es **insesgado** si su valor esperado es igual al parámetro: $E(W) = \mu$. Si un estimador es insesgado, esto no quiere decir que el valor estimado será igual al parámetro (o incluso cercano a éste), ya que esto dependerá de la muestra que sea recabada.
2. Un estimador (W_1) es más **eficiente** relativamente a otro (W_2) si $Var(W_1) \leq Var(W_2)$.

Ejemplo: Consideremos dos estimadores para el parámetro μ:

a) El valor promedio: $W_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$

b) El promedio de edad de dos personas de mi muestra elegidas al azar: $W_2 = \frac{Y_a + Y_b}{2}$

¿Son estos estimadores insesgados?, ¿cuál es más eficiente?, ¿qué valor estimado estará más cerca de la media poblacional?

2.3. Propiedades asintóticas de los estimadores

En el ejemplo anterior podemos deducir que el estimador W_1 se vuelve más eficiente conforme el tamaño de muestra aumenta. Las propiedades asintóticas de los estimadores son aquellas que aplican cuando se tienen muestras *grandes*. Sin embargo, no es claro de qué tamaño necesita ser el número de observaciones (n) para que la muestra sea considerada como *grande* y sea correcto aplicar las propiedades asintóticas a los estimadores. Generalmente, esto depende de la distribución poblacional de la variable de interés, pero en la mayoría de los casos en los que utilizamos encuestas, aplicar propiedades asintóticas será razonable.

Ejemplo: Dar un ejemplo mostrando dos distribuciones normales, una más dispersa que la otra.

Consistencia. La consistencia se refiere al comportamiento de la distribución muestral del estimador conforme el tamaño de la muestra se incrementa. Conforme aumentamos el tamaño de la muestra, la distribución de W_1 se volverá más concentrada alrededor de μ. Por lo tanto, menos probable será que un valor estimado se ubique lejos de μ.

Ejemplo: ¿Es el estimador W_2 consistente?

Ley de Grandes Números (LGN). La ley de grandes números nos dice que si queremos aproximarnos a la media poblacional, podemos hacerlo en gran medida

si elegimos muestras lo suficientemente grandes y utilizamos el estimador del promedio. Sin embargo, utilizando la LGN únicamente obtenemos estimadores puntuales y no tenemos información acerca de su distribución.

Teorema Central del Límite (TCL). Sea $\{Y_1, ..., Y_n\}$ una muestra aleatoria con media μ y varianza σ^2. Entonces,

$$Z_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1) \text{ conforme } n \to \infty \tag{2.1}$$

Intuitivamente, este resultado indica que, sin importar la distribución poblacional de Y, la distribución de la variable Z_n (que es la versión estandarizada de $\overline{Y}_n$) se aproxima en gran medida a una distribución normal estándar ($N(0,1)$) conforme el tamaño de la muestra (n) aumenta[3].

2.4. Pruebas de hipótesis

En la gran mayoría de las aplicaciones empíricas de econometría tendremos que llevar a cabo pruebas de hipótesis. En dichas pruebas de hipótesis es importante notar que evaluamos si el parámetro es igual a cierto valor en el caso de la hipótesis nula. Esto es particularmente adecuado debido a que el parámetro es un número específico que es (usualmente) desconocido para el econometrista. Generalmente, durante el curso asumiremos que las muestras son grandes y por tanto podemos aplicar propiedades asintóticas, Esto quiere decir que en la mayoría de los casos podremos utilizar la distribución normal y ji-cuadrada para llevar a cabo las pruebas de hipótesis.

Para repasar cómo llevar a cabo pruebas de hipótesis supongamos que estamos interesados en evaluar si la media de edad de los alumnos en educación media superior en México es igual a 20. Cabe señalar que nuestra hipótesis es acerca del valor de un parámetro poblacional y utilizaremos una muestra para evaluar dicha hipótesis. La hipótesis nula debe establecerse como igualdad debido a que en la evaluación de la hipótesis se asume que es cierta y eso genera una distribución para el estadístico que se utilizará. En nuestro ejemplo, establecemos la siguiente hipótesis nula:

$$H_0 : \mu = 20$$

La hipótesis alternativa se establece para especificar la zona de rechazo de la hipótesis nula. Generalmente, en una prueba se busca rechazar la hipótesis nula en favor de la alternativa. En nuestro caso, la hipótesis alternativa será:

$$H_1 : \mu \neq 20$$

[3] En clase haremos simulaciones utilizando la página http://faculty.carrollu.edu/ckuster/CT/Central/%20Limit/%20Theorem/%20Simulation.html para dar una mayor intuición acerca del tema.

La hipótesis alternativa puede también ser establecida como $\mu > 20$ (o $\mu < 20$). Este sería el caso si lo que nos interesa es evaluar si la media poblacional es mayor (menor) a 20. Es importante recordar que como resultado de la prueba de hipótesis, la hipótesis nula puede ser rechazada o no rechazada. Sin embargo, **es incorrecto decir que es aceptada**. Formalmente, en el ejemplo anterior podríamos concluir ya sea que: (i) hay evidencia suficiente para rechazar que la media poblacional es igual a 20 con $x\,\%$ de significancia, o que (ii) no hay evidencia suficiente para rechazar que la media poblacional es igual a 20 con x % de significancia.

Para establecer el nivel de significancia $x\,\%$ (o alternativamente el nivel de confianza $[1-x]\,\%$), hay que tomar en cuenta los dos tipos de errores que podemos cometer al evaluar pruebas de hipótesis:

a) Error tipo I: Podemos rechazar la hipótesis nula siendo que esta es verdadera

b) Error tipo II: Podemos no rechazar la hipótesis nula siendo esta falsa

Típicamente, el nivel de significancia se establece basado en el error tipo I, que generalmente busca reducirse en las pruebas de hipótesis. Dada nuestra notación, el nivel de significancia se define como:

$$x\,\% = Pr(\text{Rechazar } H_0 | H_0) = Pr(\text{Error tipo I})$$

El valor de $x\,\%$ será una valor que tendremos que asumir para llevar a cabo la prueba de hipótesis. El valor más común es de 0.05 (o 5 %) de significancia, seguido de 0.01 y 0.1 (lo cual es equivalente a 95 %, 99 % y 90 % de nivel de confianza, respectivamente).

El error tipo II suele estar ligado al poder estadístico. Más adelante en el curso discutiremos cómo utilizar el poder estadístico para determinar el número de observaciones que se requieren para llevar a cabo un análisis estadístico experimental.

Supongamos por el momento que elegimos un nivel de significancia de 5 % y que nuestra muestra de estudiantes mexicanos es aleatoria y consta de 1000 individuos. Supongamos que el promedio muestral de edad es de 21.5 y la varianza de las edades de 500.

Tenemos 3 alternativas para llevar a cabo la prueba de hipótesis:

2.4.1. Estadístico t

Para utilizar este método utilizamos la media y la desviación estándar estimada. La idea es asumir que la hipótesis nula es verdadera. Dado que asumimos esto, querremos determinar qué tan probable es que obtengamos un valor estimado $\overline{y} = 21.5$, dado que proviene de una distribución de la variable aleatoria $\overline{Y}$ que

es normal (por propiedades asintóticas) con media 20 y desviación estándar $\sqrt{500/1000}$.

Tomando dichos supuestos y aplicando el TCL estandarizamos $\overline{Y}$, con lo cual derivamos nuestro estadístico t, mismo que tendrá una distribución $N(0,1)$:

$$t = \frac{(\overline{Y} - \mu_0)}{\sqrt{S^2/n}} \longrightarrow N(0,1) \tag{2.2}$$

Y finalmente, de dicha distribución obtenemos un valor:

$$\hat{t} = \frac{(21{,}5 - 20)}{\sqrt{500/1000}} = 2{,}1213$$

Una vez que tenemos dicho valor utilizamos la distribución de la normal estándar para comparamos este valor con un referente que esté en el límite de ser razonable (el valor crítico). Para ello empleamos el nivel de significancia. Utilizando un 5 % de significancia determinamos que valor crítico (en términos absolutos) representa el 95 % del cdf de la distribución normal estándar. Dicho valor (1.96) se compara con el valor estimado para determinar qué tan probable es observarlo dada la distribución de la cual asumimos que proviene (debido a que supusimos que la hipótesis nula es cierta). Dado que el estadístico-t es mayor que el valor crítico rechazamos la hipótesis nula con un 5 % de significancia a favor de la hipótesis alternativa.

2.4.2. *Valor-p*

El valor-p nos dice hasta qué nivel de significancia la hipótesis nula sería rechazada. Siempre que el nivel de significancia sea mayor al valor-p, la hipótesis nula sería rechazada. Para determinar significancia a los niveles usuales, este valor usualmente se compara con 0.01, 0.05 y 0.1. Sin embargo, el valor-p tiene el significado en si mismo de informar que probabilidad existe de observar un valor igual o más extremo que el obtenido de la muestra. En nuestro ejemplo:

$$\textit{valor-}p = 2 \cdot (1 - F(|t|)) = 0{,}034 \tag{2.3}$$

Es importante, tener en consideración que en el caso en que la hipótesis alternativa sea unilateral (one-sided), $\textit{valor-}p = (1 - F(|t|))$.

2.4.3. Intervalo de confianza

Si nuevamente utilizamos un nivel de significancia de 5 %, necesitaremos determinar un intervalo de confianza del 95 %. Dicho intervalo de confianza se genera de la siguiente forma:

$$\begin{aligned} Pr\left(-1{,}96 < \frac{\sqrt{n}(\overline{Y}-\mu)}{S} < 1{,}96\right) &= 0{,}95 \\ Pr\left(\overline{Y} - 1{,}96 \cdot \frac{S}{\sqrt{n}} < \mu < \overline{Y} + 1{,}96 \cdot \frac{S}{\sqrt{n}}\right) &= 0{,}95 \end{aligned} \tag{2.4}$$

En el caso del intervalo de confianza es importante recordar que la incertidumbre radica en el intervalo dado que $\overline{Y}$ es una variable aleatoria. El intervalo lo podemos interpretar como: de cada 100 muestras aleatorias que obtengamos, 95 % de ellas tendrán al valor real del parámetro poblacional μ. No podemos decir que una vez que calculemos el intervalo, con 95 % de probabilidad éste contendrá el valor real del parámetro. Recordemos que el parámetro es un valor específico (no aleatorio), por lo tanto, se encuentra o no en el intervalo.

En nuestro caso, el intervalo de 95 % será: [20.114,22.88]. Dado que 20 no se encuentra dentro del intervalo, rechazamos la hipótesis nula con un 5 % de significancia.

Para las pruebas unilaterales la alternativa al intervalo de confianza es el límite (o cota) inferior o superior. Imaginemos que en el caso de nuestro ejemplo nuestra hipótesis alternativa es:

$$H_1 : \mu > 20$$

En este caso nos interesaría comparar a el valor propuesto (20) con la cota inferior (C_I), dado que nuestro "intervalo" consistiría en: $[C_I, \infty)$. Para calcular la cota:

$$\begin{aligned} Pr\left(\mu > \overline{Y} - 1{,}64 \cdot \frac{S}{\sqrt{n}}\right) &= 0{,}95 \\ Pr\left(\frac{\sqrt{n}(\overline{Y}-\mu)}{S} < 1{,}64\right) &= 0{,}95 \end{aligned} \tag{2.5}$$

En el caso de nuestro ejemplo la cota inferior sería 20.34, y como dicho valor es mayor a 20, concluiríamos que la hipóitesis nula se rechazaría a favor de nuestra nueva hipótesis alternativa unilateral.

Imaginen que queremos estimar el número promedio de contactos de Facebook que tienen los alumnos en el ITAM. Dicho número existe y es desconocido. Le llamaremos *parámetro poblacional* y lo denotaremos como μ_x. Para poder obtener dicho número tendríaamos que entrevistar a TODOS los alumnos del ITAM y preguntarles estos datos para finalmen-

te poder calcular un promedio. Sin embargo, llevar a cabo ese ejercicio puede ser costoso (en términos de tiempo) y tedioso. Afortunadamente, la estadística nos ofrece una alternativa.

Imaginen que en lugar de entrevistar a todos los alumnos del ITAM (a los cuales nos referiremos para propósitos de este ejercicio como la *población* decidimos entrevistar a algunos. Con esto conformamos una *muestra* que denotaremos como $\{X_1, ..., X_n\}$ y utilizamos esos datos para calcular un promedio. A dicho promedio le llamaremos un *estimador* y lo denotaremos como $\bar{X}_n$.

Antes de seguir adelante detengámonos a pensar:

1. ¿Qué estrategia seguirían para elegir a esa muestra?
2. ¿Qué sucedería si, dada la situación de Covid, decido solo recopilar información de aquellos estudiantes que son amigos míos en FB?
3. ¿Cuántas personas necesito entrevistar para tener una muestra decente?

Ahora pensemos que les dejé este ejercicio de tarea y cada uno de ustedes elije una muestra y calcula el promedio. Tomemos el caso de uno de ustedes. El alumno elegido por mí (lo llamaremos alumno A) resulta que obtuvo una muestra que denotaremos como $\{x_1^A, ..., x_n^A\}$ y a partir de dicha muestra obtuvo un promedio que denotaremos como $\bar{x}_A$, al cual llamaremos *valor estimado* (en clase tendremos un número específico). Sin embargo, pudiera haber ocurrido que en vez de elegir al alumno A hubiera elegido al otro alumno (lo llamaremos alumno B). En este caso, la *muestra* y, por lo tanto, el *valor estimado* hubieran sido distintos. Es por ello, que al *estimador*, es decir, al promedio de los valores de la muestra lo llamaremos *variable aleatoria*. Con *estimador*, solo me refiero a la estrategia que seguimos, es decir, a aplicar la formula de la media utilizando los datos obtenidos por el alumno i.

Imaginemos ahora que no imponemos que tienen que elegir la media como estimador. Cada persona elige libremente su estrategia (i.e. *estimador*). Una vez que todos ustedes recabaron muestras aleatorias, es decir, independientes e idénticamente distribuidas, **i.i.d**. ¿Qué podría hacer para comparar los estimadores de dos alumnos y decidir con cuál quedarme? Para esto necesitaremos los conceptos de **insesgadez** y **eficiencia**.

Ahora, supongamos que ya elegí al mejor estimador de entre todos los que tenía disponibles en el salón. Utilizando este estimador tengo valores específicos para los estimadores de la media y de la varianza:

$$\begin{aligned} \bar{x}_n &= \sum_{i=1}^{n} x_i = 215 \\ \sigma_{\bar{x}}^2 &= \frac{\sigma_x^2}{n} = 2500 \end{aligned} \tag{2.6}$$

¿Con estos datos puedo asegurar que $\bar{x}_n$ es el valor mas cercano a μ_x de entre todos los posbiles valores estimados que me propusieron? En caso de que no, ¿por qué no simplemente elijo el valor estimado mas cercano a μ_x de entre todos los que me hayan propuesto?
Imaginemos ahora que un estudiante sabe que estamos haciendo estos cálculos y asegura ser muy popular dado que tiene 100 contactos de Facebook. Cuando les mencionamos que nuestro mejor estimador es 250 se burla de nosotros y asegura que la media realmente es 150. ¿Podríamos asegurar con certeza que se equivoca?
R: **No**. Pero el uso de la estadística se enfoca más en cuestionarse qué tan probable es que yo haya obtenido un valor estimado de 250 dado que el valor verdadero de μ_x es 150. Para esto utilizaremos *pruebas de hipótesis*. Pero antes de poder llevar a cabo pruebas de hipótesis necesitamos saber algo acerca de la distribución de $\bar{X}_n$. Para esto utilizaremos la **ley de grandes números** y el **teorema central del límite**.
Utilizando el *teorema central del límite* llego a la conclusión de que puedo utilizar una distribución $N(250, 50)$ para evaluar mis pruebas de hipótesis. Ahora puedo llevar a cabo el siguiente ejercicio. Imaginémos que le decimos al estudiante *popular*: "supongamos que estas en lo correcto y $\mu_x = 150$. Dado que eso es cierto veamos la probabilidad de observar una media mayor o igual a 250 en una muestra aleatoria."

$$\begin{aligned} Pr(\bar{X} \geq 250) &= Pr\left(\frac{\bar{X} - 150}{\sqrt{2500}} \geq \frac{250 - 150}{\sqrt{2500}}\right) \\ &= Pr(Z \geq 2) = 0{,}02275 \end{aligned} \tag{2.7}$$

Por lo tanto, no podemos asegurar que este estudiante se equivoque, si estuviera en la correcto es muy poco probable que con una muestra aleatoria de las características que tenemos, sería muy poco probable observar la media que observamos. Este cálculo que acabamos de realizar es cercano a lo que conocemos como el valor-p. El valor-p nos dice cuál es la probabildad de obtener un estadístico (en este caso la media) que sea más inusual que el observado, dado que la hipótesis planteada es cierta.
El ejercicio que llevamos a cabo anteriormente en realidad es literalmente lo que hacemos al llevar a cabo una prueba de hipótesis. En clase veremos más formalmente todos los elementos de una prueba de hipótesis y enq eu consisten las alternativas para evaluarla.

2.5. Bootstrap

En el caso de las pruebas de hipótesis anteriores estamos basando los resultados en el teorema central del límite. Para aplicar el TCL calculamos analíticamente la varianza de la media como:

$$Var(\overline{Y}) = \frac{\sigma^2}{n} \tag{2.8}$$

donde σ^2 es la varianza de Y_i.

En el caso de TCL asumimos que una versión estandarizada de $\overline{Y}$ (Z_n) converge en distribución a una normal estándar. Una alternativa a este procedimiento consiste en generar una distribución empírica de $\overline{Y}$ y utilizar dicha distribución para calcular la varianza. Un problema con esta idea radica en que para generar una distribución empírica de $\overline{Y}$ necesitamos varias observaciones de $\overline{Y}$.

El método de bootstrap genera diversas observaciones partiendo de una muestra aleatoria $\{Y_1, ..., Y_n\}$ siguiendo los siguientes pasos:

1. Utilizando las observaciones de la muestra, elige una submuestra aleatoria de tamaño n (mismo tamaño que la muestra) con reemplazo. Esto quiere decir que habrá observaciones que se repitan más de una vez
2. Usando la submuestra calcula el estimador ($\overline{Y}$ en nuestro ejemplo)
3. Repite los pasos anteriores M veces. Con esto tendrás M observaciones para $\overline{Y}$: $\{\overline{Y}_1, ..., \overline{Y}_M\}$ y habrás generado una distribución empírica
4. Genera los estimadores:

$$\begin{aligned} E(\overline{Y}) =& \frac{1}{M}\sum_{k=1}^{M}\overline{Y}_k \\ Var(\overline{Y}) =& \frac{1}{M}\sum_{k=1}^{M}(\overline{Y}_k - E(\overline{Y}))^2 \end{aligned} \tag{2.9}$$

5. Utiliza dichos estimadores para llevar a cabo pruebas de hipótesis

Este método puede ser aplicado con la mayoría de los estimadores que veremos durante el curso. Es un método de gran utilidad siempre que sea dificil calcular una varianza para llevar a cabo pruebas de hipótesis. En particular podría utilizarse para calcular errores estándar de los coeficientes en una regresión. En dicho caso los pasos a seguir son los mismos que los descritos anteriormente. Lo que sucedería en el caso de una regresión de mínimos cuadrados ordinarios es que se llevaría a cabo una regresión con cada una de las submuestras elegidas en el primer paso. Con ello se obtendrían M posibles coeficientes para cada variable. El error estándar podría calcularse como la desviación estándar para cada uno de los coeficientes.

Existe también la posibilidad de utilizar una submuestra de tamaño menor al tamaño de la muestra original (n). En dicho caso, tendría que llevarse a cabo un ajuste al cálculo de la varianza. Supongamos que se eligen submuestras de tamaño L. Todos los pasos serían los mismos que los descritos anteriormente, con la diferencia que el estimador de la varianza se calcularía como:

$$Var(\overline{Y}) = \frac{L}{n}\frac{1}{M}\sum_{k=1}^{M}(\overline{Y}_k - E(\overline{Y}))^2 \tag{2.10}$$

A continuación se presenta un ejemplo de Bootstrap realizado en R. En el ejemplo utilizaremos datos sobre la pandemia de COVID-19 provenientes del sitio Our World in Data. La base contiene información sobre la pandemia hasta el día 27 de agosto de 2020 para 188 países. En este ejemplo nos interesará estimar la media de la tasa de contagios ($TC = \frac{\text{Casos Confirmados}}{\text{Población}}$).

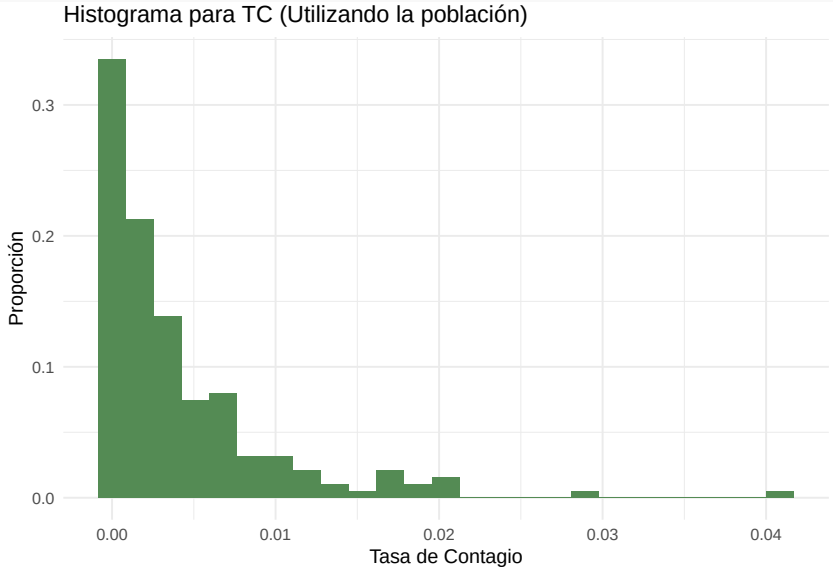
```
pacman::p_load(tidyverse)
## Iniciaremos trabajando con toda la población ##
poblacion <- as.data.frame(read.csv('COVID.csv')) # Cargamos la base poblacional

# Antes de comenzar el ejemplo
# revisemos de forma rápida las variables con las que contamos:
glimpse(poblacion)
```

```
## Rows: 188
## Columns: 25
## $ ISO_code                  <fct> ABW, AFG, AGO, ALB, AND, ARE, ARG, ARM, ATG~
## $ continent                 <fct> North America, Asia, Africa, Europe, Europe~
## $ country                   <fct> Aruba, Afghanistan, Angola, Albania, Andorr~
## $ confirmed                 <int> 1760, 38126, 2332, 8927, 1098, 68020, 37017~
## $ deaths                    <int> 8, 1401, 103, 263, 53, 378, 7839, 861, 3, 5~
## $ confirmed_per_million     <dbl> 16484.649, 979.389, 70.954, 3102.022, 14210~
## $ deaths_per_million        <dbl> 74.930, 35.989, 3.134, 91.389, 685.951, 38.~
## $ stringency_index          <dbl> 62.04, NA, NA, 53.70, 41.67, NA, NA, NA, NA~
## $ population                <int> 106766, 38928341, 32866268, 2877800, 77265,~
## $ population_density        <dbl> 584.800, 54.422, 23.890, 104.871, 163.755, ~
## $ median_age                <dbl> 41.2, 18.6, 16.8, 38.0, NA, 34.0, 31.9, 35.~
## $ aged_65_older             <dbl> 13.085, 2.581, 2.405, 13.188, NA, 1.144, 11~
## $ aged_70_older             <dbl> 7.452, 1.337, 1.362, 8.643, NA, 0.526, 7.44~
## $ gdp_pc                    <dbl> 35973.781, 1803.987, 5819.495, 11803.431, N~
## $ extreme_poverty           <dbl> NA, NA, NA, 1.1, NA, NA, 0.6, 1.8, NA, 0.5,~
## $ cardiovasc_death_rate     <dbl> NA, 597.029, 276.045, 304.195, 109.135, 317~
## $ diabetes_prevalence       <dbl> 11.62, 9.59, 3.94, 10.08, 7.97, 17.26, 5.50~
## $ female_smokers            <dbl> NA, NA, NA, 7.1, 29.0, 1.2, 16.2, 1.5, NA, ~
```

```
## $ male_smokers                <dbl> NA, NA, NA, 51.2, 37.8, 37.4, 27.7, 52.1, N~
## $ handwashing_facilities      <dbl> NA, 37.746, 26.664, NA, NA, NA, NA, 94.043,~
## $ hospital_beds_per_thousand <dbl> NA, 0.500, NA, 2.890, NA, 1.200, 5.000, 4.2~
## $ life_expectancy             <dbl> 76.29, 64.83, 61.15, 78.57, 83.73, 77.97, 7~
## $ people_tested               <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ tests_performed             <int> NA, NA, NA, NA, NA, 6755457, NA, NA, NA, 59~
## $ tests_uu                    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
# Creamos la variable de Tasa de Contagios (TC)
poblacion<-poblacion %>%
            mutate(TC = confirmed / population)
poblacion %>% ggplot(aes(TC)) +
    geom_histogram(aes(y = ..count.. / sum(..count..)),
                   fill = "palegreen4",
                   bins = 25) +
    labs(x = "Tasa de Contagio",
         y = "Proporción",
         title = "Histograma para TC (Utilizando la población)") +
    theme_minimal()
```

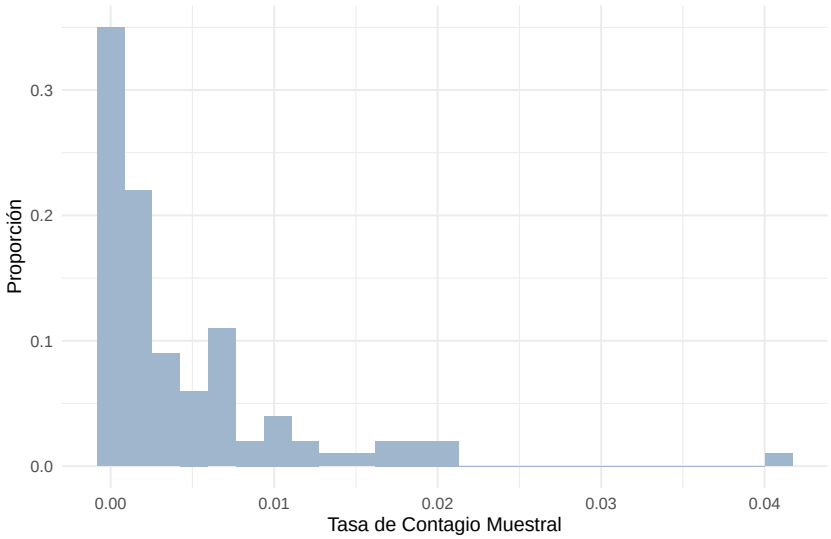



```
paste("La media poblacional de TC es ", round(mean(poblacion$TC), 3))
```

```
## [1] "La media poblacional de TC es  0.004"
```

```r
## Trabajemos ahora con una muestra de la población
# Ahora crearemos una muestra aleatoria de tamaño 100 a partir
# de la base poblacional
muestra <- sample_n(poblacion, size = 100, replace = F)

# Realizamos nuevamente un histograma
muestra %>% ggplot(aes(TC)) +
    geom_histogram(aes(y = ..count.. / sum(..count..)),
                   fill = "slategray3",
                   bins = 25) +
    labs(x = "Tasa de Contagio Muestral",
         y = "Proporción",
         title = "Histograma para TC (Utilizando la muestra)") +
    theme_minimal()
```




```r
# La gráfica previa nos muestra el histograma de la variable TC
# a partir de la muestra

paste("La media muestral de TC es ",
      round(mean(muestra$TC, na.rm = T), 3))
```

```
## [1] "La media muestral de TC es  0.004"
```

```r
paste("La varianza muestral es ",
      round(var(muestra$TC), 6))
```

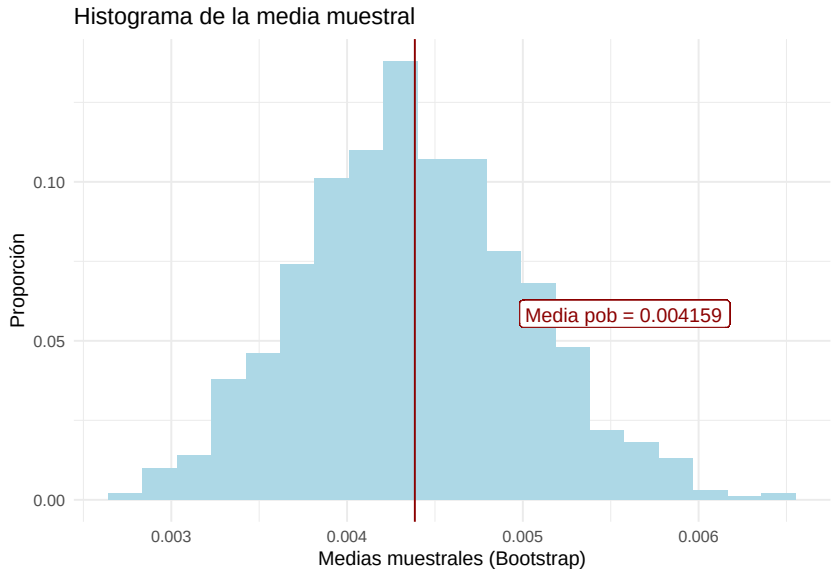
```
## [1] "La varianza muestral es  0.00004"
```

```r
paste("La varianza de la media muestral es ",
      round(var(muestra$TC) / length(muestra$TC), 10))
```

```
## [1] "La varianza de la media muestral es  0.0000003999"
```

```r
# Realizamos un Bootstrap de 1,000 submuestras del tamaño
# de la muestra original (n = 100)
meanBoot <- c() # Creamos un vector vacío
for (n in 1:1000){
    meanBoot <- c(meanBoot,
                  mean(sample(muestra$TC,
                              size = 100,
                              replace = T)))
}
meanBoot <- data.frame(mean = meanBoot)

# Ahora crearemos un histograma para la
# media muestral de TC (calculadas a partir de bootstrap)
meanBoot %>% ggplot(aes(mean)) +
    geom_histogram(aes(y = ..count../sum(..count..)),
                   fill = "lightblue",
                   bins = 20) +
    geom_vline(xintercept = mean(meanBoot$mean), colour = "red4") +
    geom_label(mapping = aes(x = mean(meanBoot$mean),
                             y = .05,
                             label = paste("Media pob =",
                                           round(mean(poblacion$TC), 6)),
                             hjust = -.5, vjust =-.5),
               colour = "red4") +
    labs(x = "Medias muestrales (Bootstrap)",
         y = "Proporción",
         title = "Histograma de la media muestral") +
    theme_minimal()
```




```
# En la gráfica anterior la línea vertical muestra
# la media poblacional de TC, dicho numero es nuestro objetivo

# Utilizando los resultados del Bootstrap, podemos estimar
# la varianza de la media muestral de TC:
paste("La varianza de la media muestral de TC es ",
      round(var(meanBoot$mean), 10))
```

```
## [1] "La varianza de la media muestral de TC es  0.0000003957"
```

```
# Comparemos los valores de la varianza de la media muestral
# de TC que hemos obtenido:
paste("De forma analítica obtuvimos: ",
      round(var(muestra$TC) / length(muestra$TC), 10))
```

```
## [1] "De forma analítica obtuvimos:  0.0000003999"
```

```
paste("A través del Bootstrap se obtuvo: ",
      round(var(meanBoot$mean), 10))
```

```
## [1] "A través del Bootstrap se obtuvo:  0.0000003957"
```

```r
# Ahora calcularemos intervalos de confianza al 95% para
# la media de TC considerando dos estrategias

# 1) Usemos el estimador de la media muestral
# y el de la varianza de la media muestral
llm_1 <- mean(muestra$TC, na.rm = T) - 1.96 * sqrt(var(muestra$TC) / length(muestra
ulm_1 <- mean(muestra$TC, na.rm = T) + 1.96 * sqrt(var(muestra$TC) / length(muestra
paste("El IC al 95%  (",
      round(llm_1, 5), ",",
      round(ulm_1, 5), ")")
```

```
## [1] "El IC al 95%  ( 0.00314 , 0.00562 )"
```

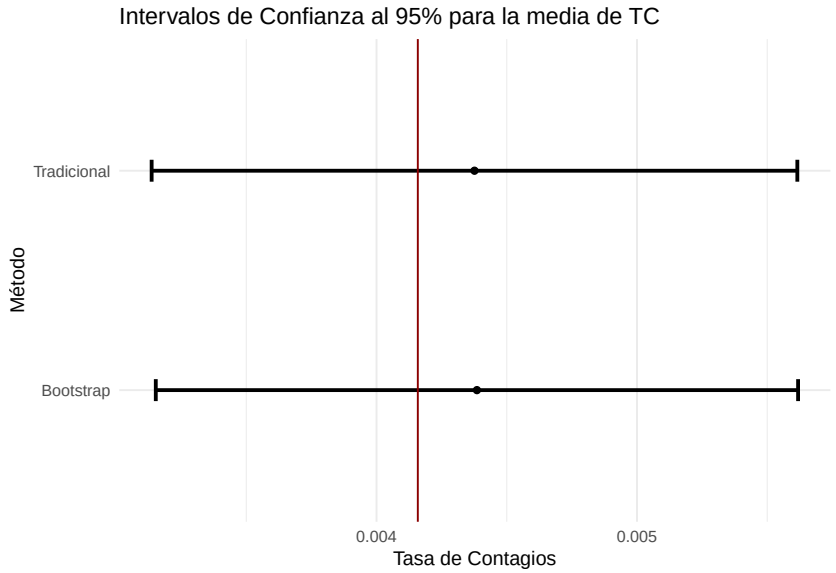
```r
# 2) Usemos los resultados que obtuvimos en Bootstrap
llm_2 <- mean(meanBoot$mean) - 1.96 * sqrt(var(meanBoot$mean))
ulm_2 <- mean(meanBoot$mean) + 1.96 * sqrt(var(meanBoot$mean))
paste("El IC al 95% (",
      round(llm_2, 5),",",
      round(ulm_2, 5),") (Bootstrap)")
```

```
## [1] "El IC al 95% ( 0.00315 , 0.00562 ) (Bootstrap)"
```

```r
#Grafiquemos los intervalos de confianza anteriores
lm <- c(llm_1, llm_2)
um <- c(ulm_1, ulm_2)
meanm <- c(mean(muestra$TC, na.rm = T),
           mean(meanBoot$mean))
tICm <- c("Tradicional", "Bootstrap")

ICm <- data.frame(tipo = tICm,
                  lower = lm,
                  media = meanm,
                  upper = um)

ICm %>% ggplot(aes(x = media, y = tipo)) +
    geom_point() +
    geom_errorbarh(aes(xmin = lower,
                       xmax = upper),
                   size = 1,
                   height = 0.1) +
    geom_vline(xintercept = mean(poblacion$TC),
               colour = "red4") +
    labs(x = "Tasa de Contagios",
         y = "Método",
         title = "Intervalos de Confianza al 95% para la media de TC") +
    theme_minimal()
```

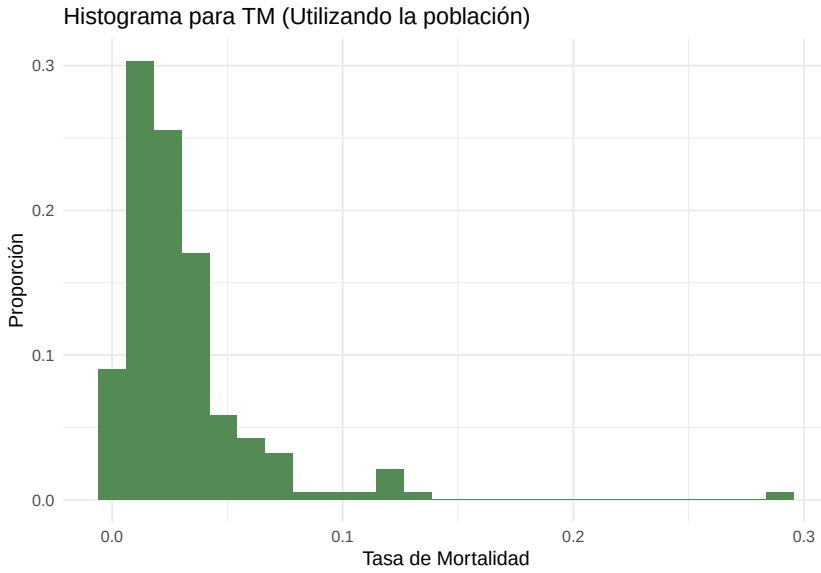



```
# La línea vertical indica el valor de la media de TC poblacional
```

Consideremos nuevamente la base de datos que utilizamos en el ejemplo anterior. Ahora estaremos interesados en calcular el *ratio del percentil 90 entre el percentil 10* $\left(\varphi = \frac{P_{90}}{P_{10}}\right)$ para la tasa de mortalidad $\left(TM = \frac{\text{Muertes}}{\text{Casos Confirmados}}\right)$. Este ratio nos indica cuántas veces mayor es la tasa de mortalidad del pais en el percentil 90 contra aquel del percentil 10.

```
## Iniciaremos trabajando con toda la población ##
# Creamos la variable de Tasa de Mortalidad (TM)
poblacion <- poblacion %>%
              mutate(TM = deaths / confirmed)

# Graficamos el histograma poblacional de TM
poblacion %>% ggplot(aes(TM)) +
    geom_histogram(aes(y = ..count.. / sum(..count..)),
                   fill = "palegreen4",
                   bins = 25) +
    labs(x = "Tasa de Mortalidad",
         y = "Proporción",
         title = "Histograma para TM (Utilizando la población)") +
    theme_minimal()
```

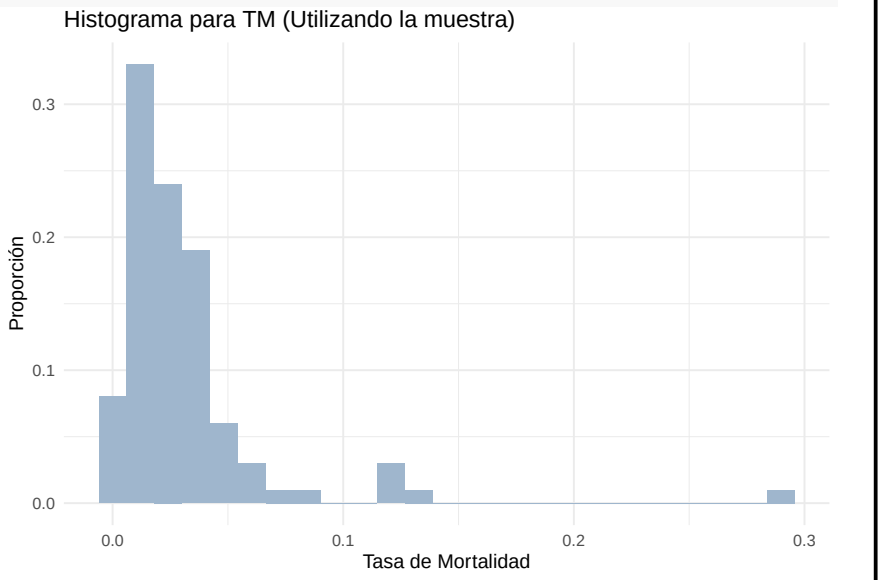



```
# Calculamos ahora el ratio del percentil 90 sobre el
# percentil 10 (P_90/P_10):
p90p <- quantile(poblacion$TM,
                 probs = 0.9,
                 type = 2)
p10p <- quantile(poblacion$TM,
                 probs = 0.1,
                 type = 2)
paste("El ratio es igual a ",
      round(p90p / p10p, 3))
```

```
## [1] "El ratio es igual a  9.514"
```

```r
## Trabajemos ahora con una muestra de la población
# Utilizaremos la muestra con la que trabajamos en
# el ejemplo anterior
muestra <- muestra %>%
            mutate(TM = deaths / confirmed)

# Graficamos el histograma de TM para la muestra
muestra %>% ggplot(aes(TM)) +
  geom_histogram(aes(y = ..count.. / sum(..count..)),
                 fill='slategray3',
                 bins = 25) +
  labs(x = "Tasa de Mortalidad", y = "Proporción",
       title = "Histograma para TM (Utilizando la muestra)") +
  theme_minimal()
```



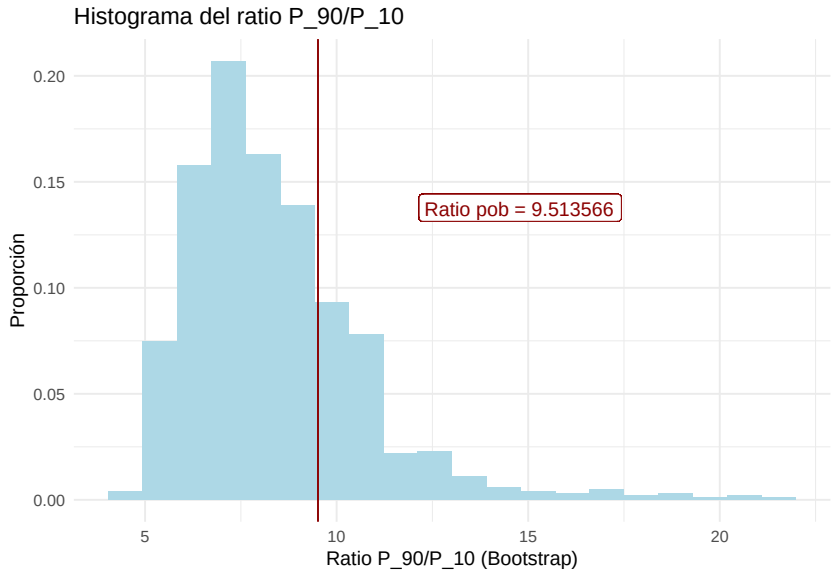

```r
# A partir de la muestra estimaremos el ratio P_90 entre P_10:
p90m <- quantile(muestra$TM,
                 probs = 0.9,
                 type = 2)
p10m <- quantile(muestra$TM,
                 probs = 0.1,
                 type = 2)
paste("El valor estimado del ratio es igual a ",
      round(p90m / p10m, 3))
```

```
## [1] "El valor estimado del ratio es igual a  8.625"
```

```
# A diferencia del ejemplo anterior, para el caso de este ratio
# es complicado obtener una expresión analítica para su varianza.
# Por tanto, para estimar dicha varianza utilizaremos Bootstrap:
rat90_10Boot <- c()
for (n in 1:1000){
    sampleboot <- sample(muestra$TM,
                         size = 100,
                         replace = T)
    rat90_10Boot<-c(rat90_10Boot,
                    quantile(sampleboot,
                             probs = 0.9,
                             type = 2) / quantile(sampleboot,
                                                  probs = 0.1,
                                                  type = 2))
}
rat90_10Boot<-data.frame(ratio = rat90_10Boot)

#Ahora crearemos un histograma para el ratio
#(calculadas a partir de bootstrap)

rat90_10Boot %>% ggplot(aes(ratio)) +
    geom_histogram(aes(y = ..count.. / sum(..count..)),
                   fill = "lightblue",
                   bins = 20) +
    geom_vline(xintercept = p90p / p10p,
               color = "red4") +
    geom_label(mapping = aes(x = p90p / p10p,
                             y = .125,
                             label = paste("Ratio pob =",
                                           round(p90p / p10p, 6)),
                             hjust = -.5, vjust = -.5),
               colour = "red4") +
    labs(x = "Ratio P_90/P_10 (Bootstrap)",
         y = "Proporción",
         title = "Histograma del ratio P_90/P_10") +
    theme_minimal()
```




```
# La línea vertical muestra el valor del ratio P_90/P_10 poblacional
# el cual es el número que tenemos como objetivo

# Al observar el histograma, podemos notar que la distribución
# del ratio P_90/P_10 muestral no corresponde a la de una Normal
# Considerando esto último, calculemos intervalos de confianza
# al 95% utilizando las submuestras del Bootstrap a través de dos estrategias:

#1) Asumamos que el ratio P_90/P_10 muestral se distribuye Normal:
llr_1 <- mean(rat90_10Boot$ratio) - 1.96 * sd(rat90_10Boot$ratio)
ulr_1 <- mean(rat90_10Boot$ratio) + 1.96 * sd(rat90_10Boot$ratio)
paste("El IC al 95% es (",
      round(llr_1, 5),",",
      round(ulr_1, 5), ") asumiendo normalidad")
```

```
## [1] "El IC al 95% es ( 3.82824 , 12.89114 ) asumiendo normalidad"
```

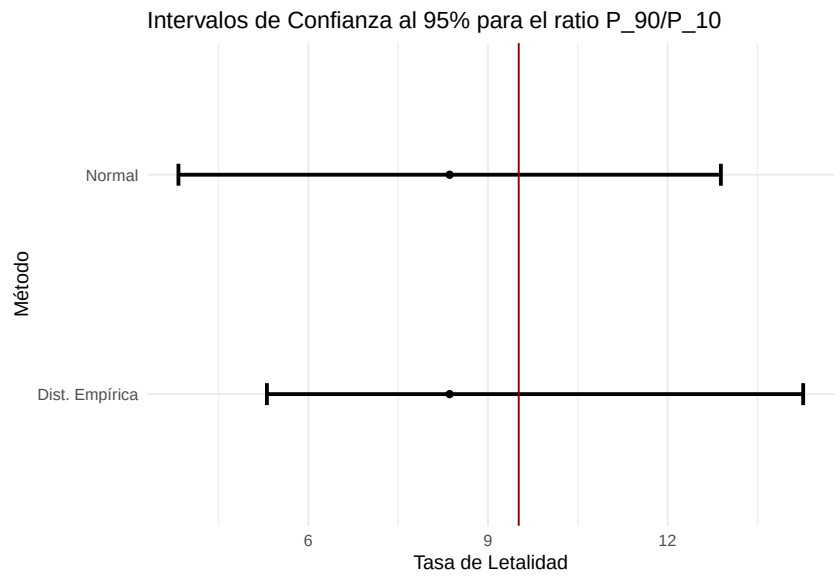
```r
# 2) Utilicemos la distribución empírica que obtuvimos a través del
# Bootstrap para el ratio P_90/P_10 muestral
# Para esto calculamos el valor del percentil 2.5 y el del percentil 97.5.
# De esta forma obtenemos el intervalo de valores en los cuales se acumula el 95%:
llr_2 <- quantile(rat90_10Boot$ratio,
                  probs = .025,
                  type = 2)
ulr_2 <- quantile(rat90_10Boot$ratio,
                  probs = .975,
                  type = 2)
paste("El IC al 95% es (",
      round(llr_2, 5), ",",
      round(ulr_2, 5), ") utilizando la distribución empírica")
```

```
## [1] "El IC al 95% es ( 5.30662 , 14.26597 ) utilizando la distribución empírica"
```

```r
#Grafiquemos los intervalos de confianza anteriores
lr <- c(llr_1, llr_2)
ur <- c(ulr_1, ulr_2)
meanr <- c(mean(rat90_10Boot$ratio),
           mean(rat90_10Boot$ratio))
tICr <- c("Normal", "Dist. Empírica")

ICr <- data.frame(tipo = tICr,
                  lower = lr,
                  media = meanr,
                  upper = ur)

ICr %>% ggplot(aes(x = media, y = tipo)) +
    geom_point() +
    geom_errorbarh(aes(xmin = lower,
                       xmax = upper),
                   size = 1,
                   height = 0.1) +
    geom_vline(xintercept = p90p / p10p,
               colour = "red4") +
    labs(x = "Tasa de Letalidad",
         y = "Método",
         title = "Intervalos de Confianza al 95% para el ratio P_90/P_10") +
    theme_minimal()
```




```
# La línea vertical indica el valor del ratio P_90/P_10 poblacional
```

Capítulo 3

Mínimos Cuadrados Ordinarios

La teoría microeconómica suele establecer, a través de modelos, relaciones funcionales entre dos variables para explicar cómo toman decisiones los agentes. Por ejemplo, supongamos que estamos interesados en analizar la relación que existe entre educación e ingreso en la sociedad mexicana. Dicha relación puede ser representada utilizando algún modelo teórico como un problema de maximización, donde cada individuo elige un nivel de escolaridad para maximizar sus ingresos. La solución a dicho problema puede establecer una relación:

$$Y = f(edu) \tag{3.1}$$

donde Y representa ingreso, edu representa educación y $f(\cdot)$ describe la relación funcional entre ambas variables (cabe señalar que no estamos restringiendo la función $f(\cdot)$ de ninguna forma). Una característica de esta relación será que a cada nivel educativo le corresponde únicamente un nivel de ingreso. Sin embargo, ¿qué sucedería si utilizamos datos empíricos para analizar esta relación?

Utilizando datos de la ENIGH[1] 2010, la siguiente tabla (*Tabla 1*) muestra la proporción de individuos entre 21 y 65 años que se encuentran laborando y que reportan distintos niveles de ingreso y educación al momento de la encuesta.

Ingreso Mensual	*Sin esc*	*Prim*	*Sec*	*M.S*	*N/CT*	*Prof*	*Posgdo*	Total
1000	2.07	7.56	3.95	1.46	0.56	1.06	0.03	16.67
2000	0.94	5.24	3.15	1.38	0.49	0.87	0.03	12.11
3000	0.75	5.18	4.24	1.91	0.65	1.00	0.05	13.78
4000	0.41	4.28	4.26	2.29	0.77	1.20	0.06	13.28

[1] Encuesta Nacional de Ingreso y Gasto de los Hogares, INEGI

Ingreso Mensual	*Sin esc*	*Prim*	*Sec*	*M.S*	*N/CT*	*Prof*	*Posgdo*	Total
5000	0.28	2.74	3.50	2.12	0.86	1.24	0.05	10.80
6000	0.10	1.44	2.13	1.45	0.73	1.26	0.06	7.15
7000	0.04	0.88	1.54	1.23	0.67	1.35	0.09	5.80
8000	0.02	0.46	0.88	0.76	0.41	1.22	0.09	3.84
9000	0.03	0.34	0.64	0.67	0.44	1.20	0.12	3.45
10000	0.08	0.76	1.34	1.60	1.42	6.50	1.43	13.12
Total	4.72	28.89	25.64	14.86	7.00	16.90	1.99	**100.00**

La *Tabla 1* muestra la densidad conjunta de educación e ingreso. Es decir, cada celda da el valor de $p(edu_j, y_i)$ que representa la proporción de individuos que reportan recibir un ingreso igual a y_i y tienen una educación igual a edu_j. Por ejemplo, dado que la muestra elegida incluye a un total de $44,270$ individuos, la celda que corresponde a secundaria y $\$5,000$ pesos mensuales indica que $1,549$ individuos reportan estos niveles de escolaridad e ingreso en la encuesta.

A partir de la *Tabla 1* puede notarse que, contrario a la teoría donde existe una relación funcional, la relación empírica entre educación e ingreso no es deterministica. Es decir, a cada nivel de educación no le corresponde solamente un nivel de ingreso. De lo contrario, se vería en la *Tabla 1* que, en cada columna, todos los renglones excepto uno tendrían un valor igual a cero. No obstante, lo que sí puede observarse en la *Tabla 1* es una tendencia que muestra que los individuos con mayor nivel educativo tienden a percibir mayores ingresos. Esto se ilustra más claramente con la siguiente tabla (*Tabla 2*) que muestra las frecuencias condicionales.

Cada celda de la *Tabla 2* representa $p(y|edu)$, la frecuencia de ingreso condicional a los distintos niveles educativos. Por ejemplo, de acuerdo a esta tabla el 38,39 % de los individuos que tienen educación media superior perciben un ingreso mayor a $\$5,000$ pesos mientras que solamente el 68,23 % de los individuos con educación superior reciben dicho nivel de ingreso.

Una forma de utilizar los datos empíricos para generar una relación uno a uno es utilizar la tabla de densidades condicionales para calcular la media condicional de ingresos. Dicha media es definida para cada nivel de educación j (edu_j) como:

$$m_{y|edu_j} = \sum_i y_i \cdot p(y_i|edu_j) \tag{3.2}$$

Ingreso Mensual	*Sin esc*	*Prim*	*Sec*	*M.S*	*N/CT*	*Prof*	*Posgdo*	Total
1000	43.81	26.18	15.40	9.79	7.97	6.25	1.34	16.67
2000	19.24	18.13	12.31	9.30	6.93	5.15	1.61	12.11
3000	15.97	17.94	16.53	12.84	9.27	5.93	2.42	13.78
4000	8.69	14.82	16.63	15.42	11.03	7.09	2.96	13.28
5000	5.85	9.50	13.66	14.26	12.34	7.36	2.42	10.80

Ingreso Mensual	*Sin esc*	*Prim*	*Sec*	*M.S*	*N/CT*	*Prof*	*Posgdo*	Total
6000	2.05	4.98	8.29	9.74	10.42	7.44	2.82	7.15
7000	0.91	3.06	6.00	8.28	9.58	7.98	4.44	5.80
8000	0.51	1.59	3.43	5.10	5.82	7.23	4.30	3.84
9000	0.68	1.18	2.51	4.49	6.36	7.12	6.05	3.45
10000	1.59	2.62	5.24	10.78	20.27	38.46	71.64	13.12
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	**100.00**

Utilizando estos valores se genera una relación uno a uno entre nivel educativo y de ingreso que se ilustra en las gráficas a continuación. En la Figure 3.1 se ilustra utilizando los niveles de escolaridad y en la Figure 3.2, se transforma el nivel a años de escolaridad. En la segunda gráfica puede observarse que no resultaría sencillo especificar una forma funcional para representar la relación entre ingreso y años de escolaridad.

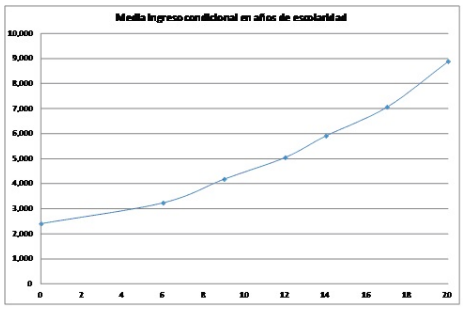


Figura 3.1: Medias condicionales por nivel educativo

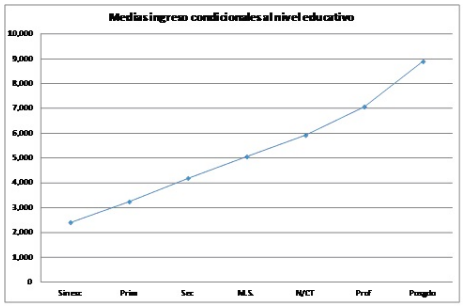


Figura 3.2: Medias condicionales por nivel educativo

En la primera parte de este curso aprenderemos cómo utilizar la metodología de *Mínimos Cuadrados Ordinarios* para poder utilizar datos empíricos para estimar una forma funcional determinada.

3.1. Derivación de Mínimos Cuadrados Ordinarios

3.1.1. Regresión simple

Partiendo con una muestra aleatoria $\{Y_i, X_i\}$ i.i.d., una manera de representar con una forma funcional la relación entre dos variables consiste en asumir que la relación entre dichas variables es lineal:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \tag{3.3}$$

Asumir que la relación es lineal implica que el cambio de la variable dependiente (Y_i) por un cambio marginal en la variable independiente (X_i) es constante. La variable U_i corresponde al error de la estimación o la parte no explicada del modelo. Nos referimos a ésta como la parte no explicada ya que el resto de la ecuación $(\beta_0 + \beta_1 X_i)$ representa una relación en la que la variable X_i pretende aproximar o explicar de la mejor manera posible a la variable dependiente, Y_i.

Ejemplo: Aumentar de 4 a 5 años el nivel de escolaridad daría el mismo rendimiento que aumentar de 16 a 17.

Una vez que asumimos esa relación lineal tenemos que determinar una manera óptima de elegir los estimadores β_0 y β_1. El método de mínimos cuadrados ordinarios consiste en minimizar el cuadrado de los errores de la estimación (U_i):

$$\min_{\beta_0, \beta_1} \quad E(U_i^2) \tag{3.4}$$

Esta minimización da como resultado:

$$\begin{aligned} \beta_0 =& E(Y_i) - \beta_1 E(X_i) \\ \beta_1 =& \frac{E(Y_i X_i) - E(Y_i)E(X_i)}{E(X_i^2) - E(X_i)^2} = \frac{Cov(X_i, Y_i)}{Var(X_i)} \end{aligned} \tag{3.5}$$

Por lo tanto, partiendo de (3.5) obtenemos los estimadores de MCO:

$$\begin{aligned} \hat{\beta}_0 =& \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ \hat{\beta}_1 =& \frac{\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \end{aligned} \tag{3.6}$$

Los valores estimados correspondientes se obtienen a partir del análogo muestral.

Las dos condiciones de primer orden que resultan de la derivación anterior son clave para obtener los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Más adelante regresaremos a estos supuestos:

$$\begin{aligned} E(Y_i - \beta_0 - \beta_1 X_i) = E(U_i) &= 0 \\ E((Y_i - \beta_0 - \beta_1 X_i)U_i) = E(X_i U_i) &= 0 \end{aligned} \tag{3.7}$$

La primera condición de primer orden ($E(U_i) = 0$) simplemente indica que por el hecho de incluir una constante en la estimación (β_0), el valor esperado de los errores debería estar centrados en cero.

La segunda condición de primer orden ($E(X_i U_i) = 0$) indica que la variable explicativa (X_i) no está correlacionada con los errores o la parte no explicada del modelo.

Es importante señalar que los errores U_i representan todo aquello que no es incluido dentro del modelo.

3.1.2. Regresión multiple

El caso anterior ilustra la derivación del modelo MCO en el caso de una regresión simple (i.e. con una sola variable explicativa). Para aumentar la cantidad de variables explicativas será útil la representación vectorial. Para esto, empecemos por notar que al incluir más variables explicativas, nuestro planteamiento para la estimación será:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + U_i \tag{3.8}$$

Sea X_i un vector de dimensión $K+1$, donde el primer componente es la constante (1) y los demás corresponden a las variables X_{li} que son incluidas como controles en la estimación (3.8). Similarmente, β será un vector de dimensión $K+1$ que incluye todos los coeficientes $[\beta_0 \quad \beta_1 \quad ... \quad \beta_K]$ de la estimación. Utilizando esta notación, la ecuación (3.8) puede ser reescrita como:

$$Y_i = X_i'\beta + U_i \quad \text{para } i = 1, ..., n \tag{3.9}$$

Siguiendo la misma metodología minimizo el valor esperado de los errores al cuadrado:

$$\min_{\beta} E(Y_i - X_i'\beta)^2 \tag{3.10}$$

Esto da como resultado la siguiente expresión para el parámetro:

$$\beta = E(\mathbf{X}_i\mathbf{X}_i')^{-1}E(\mathbf{X}_i\mathbf{Y}_i) \tag{3.11}$$

Partiendo de esto, el estimador seria :

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right) \tag{3.12}$$

Y finalmente, el valor estimado se calcula con el análogo muestral de (3.12). Utilizando las propiedades asintóticas, es posible demostrar que:

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} \beta \\ \sqrt{n}(\hat{\beta} - \beta) &\xrightarrow{d} N(0, \alpha \Sigma \alpha') \end{aligned} \tag{3.13}$$

donde,

$$\begin{aligned} \alpha =& E(X_i X_i')^{-1} \\ \Sigma =& E(U_i^2 X_i X_i') \end{aligned}$$

Demostración $\hat{\beta} \xrightarrow{p} \beta$:

$$\begin{aligned} \hat{\beta} - \beta &= \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right) - \beta \\ &= \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \beta \right) + \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i U_i \right) - \beta \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i U_i \right) \end{aligned}$$

pero,

$$\begin{aligned} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} &\xrightarrow{p} E(X_i X_i')^{-1} \\ \left(\frac{1}{n} \sum_{i=1}^{n} X_i U_i \right) &\xrightarrow{p} E(X_i U_i) = 0 \end{aligned}$$

Por lo tanto,

$$\hat{\beta} \xrightarrow{p} \beta$$

Demostracion $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \alpha\ \alpha')$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i U_i \right)$$

pero,

$$\left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \xrightarrow{p} E(X_i X_i')^{-1} = \alpha$$

Además, sea $W_i = X_i U_i$. Por lo tanto tendremos que:

$$\begin{aligned} E(W_i) &= E(X_i U_i) = 0 \\ Var(W_i) &= E(X_i U_i U_i' X_i') \\ &= E(U_i^2 X_i X_i') = \Sigma \end{aligned}$$

El segundo término se puede transformar en:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i U_i = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} W_i \xrightarrow{d} N(0, \Sigma)$$

Por lo tanto,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \alpha \Sigma \alpha')$$

3.2. Homocedasticidad y heterocedasticidad

Hasta ahora nuestros únicos supuestos en el modelo MCO han sido: (i) que nuestra muestra es aleatoria (i.i.d), y (ii) la relación lineal entre Y_i y X_i.

El supuesto de homocedasticidad indica que la varianza condicional de los errores es constante (i.e. no cambia con el nivel de X). Dicho supuesto es restrictivo, sin embargo, no es necesario para poder realizar inferencia. Si en cambio asumimos heterocedasticidad, no es necesario asumir que la varianza de los errores es constante y, por lo tanto, podemos estimar la varianza de los coeficientes con el análogo muestral de $\alpha \Sigma \alpha'$.

Empecemos con el caso de heterocedasticidad. Para obtener un estimador consistente de $\alpha \Sigma \alpha'$ basta utilizar:

$$\begin{aligned} \hat{\alpha} &= \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \xrightarrow{p} E(X_i X_i')^{-1} = \alpha \\ \hat{\Sigma} &= \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_i^2 X_i X_i' \right)^{-1} \xrightarrow{p} E(U_i^2 X_i X_i')^{-1} = \Sigma \\ \hat{U}_i &= Y_i - X_i' \hat{\beta} \end{aligned} \tag{3.14}$$

Utilizando propiedades de LGN (propiedades de Slutsky), tendremos que $\hat{\alpha} \hat{\Sigma} \hat{\alpha}' \xrightarrow{p} \alpha \Sigma \alpha'$

En el caso de homocedasticdad, estamos asumiendo que la varianza de los errores dado X_i es constante:

$$Var(U_i | X_i) = E(U_i^2 | X_i) = \sigma^2 \tag{3.15}$$

Si aplicamos este supuesto (3.15) al segundo factor de la varianza (Σ), tenemos:

$$E(U_i^2 X_i X_i') = E(E(U_i^2|X_i)X_i X_i') = \sigma^2 E(X_i X_i') \tag{3.16}$$

Por lo tanto, la varianza $\alpha\Sigma\alpha'$ se reduce a:

$$\alpha\Sigma\alpha' = E(X_i X_i')^{-1}\sigma^2 E(X_i X_i')E(X_i X_i')^{-1} = \sigma^2 E(X_i X_i')^{-1} \tag{3.17}$$

Dado que $E(U_i^2)$ es un estimador consistente de σ^2, para obtener un estimador consistente de la varianza únicamente necesitamos:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2$$

$$\hat{\sigma}^2\hat{\alpha} \xrightarrow{p} \sigma^2 E(X_i X_i')^{-1}$$

En clase ilustraremos la diferencia entre el supuesto de homocedasticidad y heterocedasticidad con un gráfico. A lo largo del curso asumiremos heterocedasticidad, que es el supuesto menos restrictivo.

3.3. Pruebas de hipótesis en el Modelo de Regresión Lineal

3.3.1. Unidimensionales. Un coeficiente o una combinación lineal de coeficientes

Si partimos del resultado que demostramos previamente:

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \alpha\Sigma\alpha') \tag{3.18}$$

Necesitamos seguir los siguientes pasos para establecer la distribución necesaria para llevar a cabo las pruebas de hipótesis. Dado que β es un vector de dimensión K, para llevar a cabo pruebas de hipótesis podemos tomar el producto $l'\beta$ donde l es un vector de dimensión K también[2]:

$$l'\beta = \sum_{j=1}^{K} l_j\beta_j \tag{3.19}$$

Con esta definición podríamos establecer la prueba de hipótesis de la siguiente forma:

[2]Previamente habíamos indicado que β tiene dimensión $K+1$. En adelante solo usamos dimensión K para simplificar la notación. En términos esctrictos podríamos decir que previamente la dimensión era $K'+1$ y definir a $K = K'+1$.

$$H_0 : \quad l'\beta = 0$$
$$H_1 : \quad l'\beta \neq 0$$

Nótese que en vez del valor "0" se puede usar cualquier constante. Este planteamiento permite hacer pruebas de hipótesis para coeficientes de forma individual. Esto es lo más común en la práctica, donde evaluamos la representatividad de un coeficiente individual. También pueden evaluarse combinaciones lineales de coefientes, como por ejemplo $\beta_1+\beta_2$, además de combinaciones lineales más complejas. La motivación para hacer pruebas de hipótesis con combinaciones lineales la analizaremos más a profundidad en la siguiente sección ya que la combinación adecuada a evaluar generalmente viene motivada por la interpretación de los coeficientes. En ese momento discutiremos algunos ejemplos. Por lo pronto basta entender que el planteamiento de $l'\beta$ permite hacer estimaciones de combinaciones tan complejas como: $\beta_1 + \frac{1}{2}\ \beta_2 - 4\ \beta_4$. En este caso, si estamos estimando una especificación con 4 variables (más la constante), $l' = [0 \ \ 1 \ \ \frac{1}{2} \ \ 0 \ \ 4]$.

Para poder evaluar pruebas de hipótesis como lo hicimos en el repaso de estadística necesitaremos asociar a $l'\beta$ con una distribución. Dado que $\hat{\alpha}\hat{\Sigma}\hat{\alpha}' \xrightarrow{p} \alpha\Sigma\alpha'$, por propiedades de LGN (Slutsky):

$$(l'\hat{\alpha}\hat{\Sigma}\hat{\alpha}'l)^{1/2} \xrightarrow{p} (l'\alpha\Sigma\alpha'l)^{1/2}$$

Entonces, por propiedades de CLT (Slutsky):

$$\frac{l'[\sqrt{n}(\hat{\beta}-\beta)]}{(l'\hat{\alpha}\hat{\Sigma}\hat{\alpha}'l)^{1/2}} \xrightarrow{d} \frac{1}{(l'\alpha\Sigma\alpha'l)^{1/2}} N(0, l'\alpha\Sigma\alpha'l) = N(0,1)$$

De manera que si definimos el error estándar como:

$$SE = \left(\frac{l'\hat{\alpha}\hat{\Sigma}\hat{\alpha}'l}{n}\right)^{1/2}$$

Tenemos que:

$$\frac{l'(\hat{\beta}-\beta)}{SE} \xrightarrow{d} N(0,1) \tag{3.20}$$

Y a partir de esto podemos generar intervalos de confianza y llevar a cabo pruebas de hipótesis.

3.3.2. Multidimensionales. Varios coeficientes o combinaciones lineales de coeficientes.

Si quisiéramos llevar a cabo pruebas de hipotesis multi-dimensionales, en vez de un intervalo de confianza necesitaríamos generar una elipse de confianza (o una

región de confianza). Llevar a cabo pruebas de hipótesis de forma independiente (como lo veremos en clase) implica dejar de considerar la covarianza que puede existir entre los coeficientes.

Supongamos que queremos evaluar la siguiente prueba hipótesis:

$$\begin{aligned} H_0: \quad & \beta_2 = 0 \\ & \beta_3 = 0 \\ H_1: \quad & e.o.c. \end{aligned} \tag{3.21}$$

Para hacer una representación matricial de esta prueba de hipótesis podemos generar una matriz L de dimensión $(h \times K)$ donde h es el número de condiciones (o ecuaciones) que estamos considerando en la prueba de hipótesis. En el caso de nuestro ejemplo $h = 2$ porque son dos condiciones las que se quieren verificar ($\beta_2 = 0$ y $\beta_3 = 0$). Por lo tanto, nuestra prueba de hipótesis sería:

$$\begin{aligned} H_0: \quad & L\ \beta = 0 \\ H_1: \quad & e.o.c. \end{aligned}$$

En este caso, si nuevamente tuviesemos una especificación con 4 variables (más la constante), L sería una matriz de tamaño 2×5:

$$L = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.22}$$

Al igual que en el caso anterior, la motivación a la prueba de hipótesis que se quiere evaluar debe venir de la interpretación de los coeficientes y veremos ejemplos relacionados. La igualdad del lado derecho nuevamente no necesariamente debe ser un vector h de ceros. Asimismo, la matriz L puede ser mas complicado, lo único necesario es que cada renglón (o condición) a evaluar debe ser una combinación lineal de coeficientes, de la misma forma que $l'\beta$. Por ejemplo, si se quisieran evaluar las siguientes condiciones:

$$\begin{aligned} H_0: \quad & 2\ \beta_1 + \beta_2 = 0 \\ & \frac{1}{3}\ \beta_1 - 3\ \beta_3 = 0 \\ & \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0 \\ H_1: \quad & e.o.c. \end{aligned}$$

La matriz L correspondiente sería:

$$L = \begin{bmatrix} 0 & 2 & 1 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & -3 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Siguiendo el mismo procedimiento que en el caso unidimensional:

$$L\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, L\alpha\Sigma\alpha' L') \tag{3.23}$$

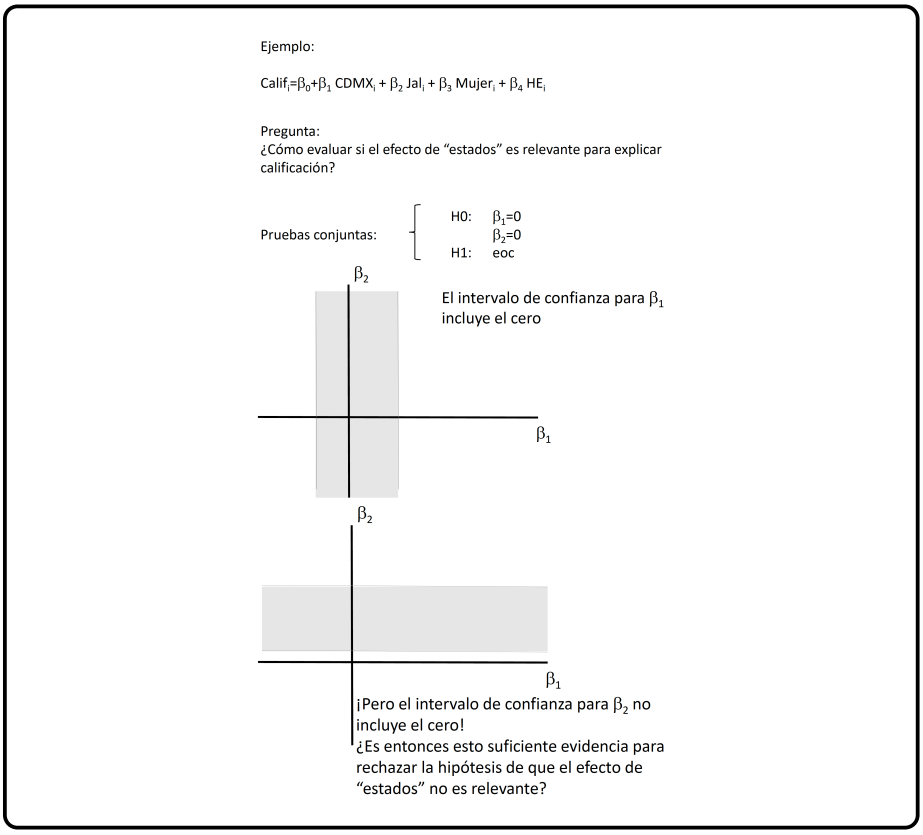
Además en este caso:

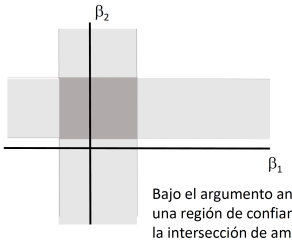
$$\widehat{Var(L\hat{\beta})} = \frac{L\hat{\alpha}\hat{\Sigma}\hat{\alpha}' L'}{n}$$

Y por teoría de probabilidad (producto de normales estandár es una ji-cuadrada con grados de libertad igual a la dimensión de la varianza):

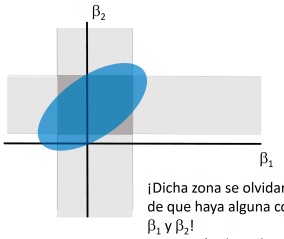
$$(L\hat{\beta} - L\beta)' \left[\frac{L\hat{\alpha}\hat{\Sigma}\hat{\alpha}' L'}{n} \right]^{-1} (L\hat{\beta} - L\beta) \xrightarrow{d} \chi^2_h \tag{3.24}$$

En clase explicaremos e ilustraremos por qué no es correcto utilizar dos intervalos de confianza generados a partir del caso de pruebas de hipótesis unidimensionales para el caso $h = 2$.

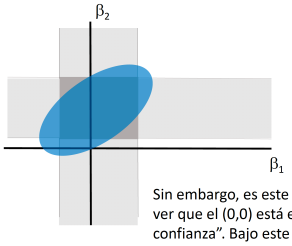




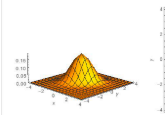
Bajo el argumento anterior, tendríamos una región de confianza producto de la intersección de ambos intervalos, como la zona gris oscuro

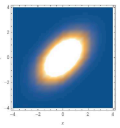


¡Dicha zona se olvidaría de la posibilidad de que haya alguna correlación entre β_1 y β_2!
Una región de rechazo como la azul si tendría en cuenta la posible correlación



Sin embargo, es este ejemplo podemos ver que el (0,0) está en la "zona de confianza". Bajo este argumento no rechazaríamos la hipótesis nula de que el efectos de "estados" no es relevante





Esta imagen representa una distribución normal bivariada (recordemos q el vector β se distribuye normal) y la imagen de la derecha es una región de confianza que la distribución de la izquierda proyecta. Noten q la forma se asemeja a la que ilustre en la lámina anterior

- ¿Cómo evaluamos entonces la prueba de hipótesis planteada?

H0: $\beta_1=0$
$\beta_2=0$
H1: eoc

- Utilizamos el estadístico F
- Esta distribución resulta del producto de dos distribuciones normales
  - Mas en específico, este es el caso de la ji-cuadrada, pero la F converge a la ji-cuadrada cuando n tiende a infinito, así como la t hacia la normal
- Intuitivamente, hacemos el producto de dos valores que se distribuyen normal y obtenemos el estadístico F con la siguiente formula:

- Siguiendo la estrategia descrita en la Nota 2 establecemos:

$$L = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{bmatrix}$$

- Recordamos ademas que $\alpha\Sigma\alpha'$ es la matriz de varianza-covarianza bajo heterocedasticidad que resulta de hacer al estimacion.
- Calculamos F y contrastamos contra valores críticos de la distribución ji-cuadrada con h grados de libertad

$$F = h \cdot (L\hat{\beta} - L\beta)' \left[\frac{L\hat{\alpha}\hat{\Sigma}\hat{\alpha}' L'}{n} \right]^{-1} (L\hat{\beta} - L\beta)$$

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
R^2	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (*STR*)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, *HiEL*)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

Tests of joint hypotheses:

***F*-Statistics and *p*-Values on Joint Hypotheses**

(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR, HiEL \times STR^2, HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
R^2	0.773	0.794	0.305	0.795	0.798	0.799	0.798

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

What can you conclude about question #1?

About question #2?

TABLE 9.2 Multiple Regression Estimates of the Student–Teacher Ratio and Test Scores: Data from Massachusetts

Dependent variable: average combined English, math, and science test score in the school district, fourth grade; 220 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (STR)	−1.72** (0.50)	−0.69* (0.27)	−0.64* (0.27)	12.4 (14.0)	−1.02** (0.37)	−0.67* (0.27)
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, $HiEL$.)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% Eligible for free lunch		−0.521** (0.077)	−0.582** (0.097)	−0.587** (0.104)	−0.709** (0.091)	−0.653** (0.72)
District income (logarithm)		16.53** (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87* (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184* (0.090)	0.165 (0.085)
District income^3			−0.0022* (0.0010)	−0.0023* (0.0010)	−0.0023* (0.0010)	−0.0022* (0.0010)
Intercept	739.6** (8.6)	682.4** (11.5)	744.0** (21.3)	665.5** (81.3)	759.9** (23.2)	747.4** (20.3)

(continued)

(Table 9.2 continued)

***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
All STR variables and interactions = 0				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2$, $Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL$, $HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
R^2	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. Individual coefficients are statistically significant at the *5% level or **1% level.

How do the Mass and California results compare?

- Logarithmic v. cubic function for STR?
- Evidence of nonlinearity in *TestScore*-STR relation?
- Is there a significant $HiEL \times STR$ interaction?

TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	−0.66* (0.29)	−0.64* (0.36)	−0.45 (0.30)	−0.69* (0.35)	−0.46 (0.31)	−0.93** (0.34)
Drinking age 18				0.028 (0.070)	−0.010 (0.083)		0.037 (0.102)
Drinking age 19				−0.018 (0.050)	−0.076 (0.068)		−0.065 (0.099)
Drinking age 20				0.032 (0.051)	−0.100+ (0.056)		−0.113 (0.125)
Drinking age						−0.002 (0.021)	
Mandatory jail or community service?				0.038 (0.103)	0.085 (0.112)	0.039 (0.103)	0.089 (0.164)
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063** (0.013)		−0.063** (0.013)	−0.091** (0.021)
Real income per capita (logarithm)				1.82** (0.64)		1.79** (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**							
Time effects = 0			4.22 (0.002)	10.12 (< 0.001)	3.48 (0.006)	10.28 (< 0.001)	37.49 (< 0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (< 0.001)		31.96 (< 0.001)	25.20 (< 0.001)
$\bar{R}^2$	0.091	0.889	0.891	0.926	0.893	0.926	0.899

These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The individual coefficient is statistically significant at the +10%, *5%, or **1% significance level.

3.4. Interpretación de coeficientes

Generalmente estamos interesados en uno de los parámetros que estamos estimando (sea β_1). La interpretación del coeficiente estimado en el caso de una regresión multivariada es que indica el cambio en la variable dependiente (Y) por un cambio marginal en la variable X_1 ($\Delta X_1 = 1$), tomando todas las demás variables de control ($X_2, \dots, X_K$) constantes (*caeteris paribus*).

3.4.1. Caso simple

En la interpretación de coeficientes es importante tomar en cuenta en qué unidades se miden las variables dependiente (Y) y la variable de control que cambia marginalmente (X_1).

Supongamos que estimamos la siguiente especificación con MCO:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + U_i \tag{3.25}$$

Veamos el resultado de aumentar en una unidad la variable X_1 ($\Delta X_1 = 1$) dejando todas las demás variables constantes:

$$Y_i' = \beta_0 + \beta_1 (X_{1i} + 1) + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + U_i \tag{3.26}$$

En este caso, el cambio de la variable dependiente será:

$$\Delta Y_i = Y_i' - Y_i = \beta_1 \tag{3.27}$$

Por lo tanto, β_1 representará el cambio en Y_i que resulta de incrementar X_1 en una unidad manteniendo todas las demás variables constantes. Durante la clase veremos ejemplos prácticos.

3.4.2. Formas funcionales

A pesar de que la forma lineal de la especificación del modelo MCO parece ser muy restrictiva en principio, es posible hacer transformaciones de las variables independientes y con ello lograr una mejor aproximación de la relación entre dos variables.

Llevar a cabo transformaciones permite hacer una estimación más exacta en términos estadísticos y más correcta en términos teóricos.

Para decidir qué tipo de transformación es más pertinente utilizar en cada caso debe considerarse:

- De acuerdo a la teoría, qué función representa de manera mas fidedigna la relación entre dos variables

- De acuerdo a los datos empíricos, qué función describe de manera más exacta la relación entre los datos observados

3.4.3. Transformación polinomial

En el caso de transformaciones para generar distintas formas funcionales (como se discute en la sección anterior) es importante considerar que las interpretaciones de los coeficientes cambian de forma importante.

Para empezar, tomemos el caso de una transformación polinomial de alguna variable de interés. Supongamos que estimamos la siguiente regresión:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_K X_{Ki} + U_i \tag{3.28}$$

Si en este caso queremos ver el cambio en la variable dependiente Y por un cambio marginal en X_1, es importante notar que la variable X_1 aparece en más de un término en la estimación (3.28). Por lo tanto, si queremos aproximar el efecto de un aumento marginal de X_1 (i.e. $\partial X_1 = 1$), calculamos la derivada respecto de X_1:

$$\frac{\partial Y_i}{\partial X_{1i}} = \beta_1 + 2\beta_2 X_{1i}$$

Lo que la ecuación anterior indica es que ahora, por la forma cuadrática, el efecto de un aumento marginal de X_1 sobre Y depende del nivel de X_1 del cual partimos. Por lo tanto, el efecto será distinto dependiendo del individuo. En particular, el coeficiente β_1 corresponde al cambio en Y asociado a un aumento de una unidad de X_1 condicional en que el valor inicial de $X_1 = 0$. Por otra parte β_2 tiene una interpretación relacionada a cómo este efecto marginal va cambiando conforme se incrementa X_1:

$$\frac{\partial^2 Y_i}{\partial X_{1i}^2} = 2\beta_2$$

En la ecuación anterior podemos ver que el signo de β_2 determina el signo de $\frac{\partial^2 Y_i}{\partial X_{1i}^2}$. Es decir, β_2 nos dice si la relación entre Y_i y X_{1i} es cóncava o convexa. En otras palabras, nos dice si el cambio de Y por un cambio marginal de X_1 es creciente o decreciente.

En (3.28) empleamos un polinomio de grado 2. Sin embargo, podemos utilizar un polinomio de mayor grado en la especificación. Cabe notar, sin embargo, que dicho cambio debe estar sustentado teóricamente o con evidencia empírica (mostrada en los datos), ya que conforme aumenta el polinomio la interpretación de algún efecto partiendo de los coeficientes se vuelve más complicado.

3.4.4. Transformación logarítmica

En el caso de una transformación logarítmica, los coeficientes cambian en su interpretación. Una característica particular de los logaritmos es que provocan que la variable sobre la cual está aplicado el logaritmo ya no utilizará sus unidades en la interpretación. Esto es de gran ayuda en casos en los cuales las unidades que se están utilizando no son fáciles de interpretar o de conocimiento universal (e.g. calificaciones en un país en específico). En vez de utilizar sus unidades, los cambios en esta variable estarán expresados en términos porcentuales. Esto es útil en casos en los cuales suelen ser comunes los cambios porcentuales (e.g. variables monetarias).

La transformación logarítmica puede aplicarse tanto a la variable dependiente como a controles de la estimación. Empecemos por ver cómo interpretar el coeficiente de algún control cuando la variable dependiente es un logaritmo:

$$\ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i \tag{3.29}$$

Para ver cómo interpretaríamos ahora un coficiente veamos el caso de β_1. Primero calculamos la derivada respecto de X_1:

$$\beta_1 = \frac{\partial \ln(Y_i)}{\partial X_{1i}} = \frac{\partial Y_i / \partial X_{1i}}{Y_i}$$

En este caso, para un cambio de X_{1i} en una unidad (i.e. $\partial X_{1i} = 1$), tenemos que β_1 representa un cambio de Y_i igual a $\frac{\partial Y_i}{Y_i}$. Para expresar esto en tasa (en términos porcentuales) tenemos que multiplicar por 100 ambos lados para obtener que: $100 \cdot \beta_1 = 100 \cdot \left(\frac{\partial Y_i}{Y_i} \right)$ %. Nótese que las unidades de Y_i en este caso ya no influyen en la interpretación del coeficiente, pero las de X_{1i} sí. En resumen, en este caso tendremos que un aumento de una unidad de X_{1i} conlleva un aumento de $100 \cdot \beta_1$ % en Y_i.

Imaginemos ahora que en vez de (3.29), hacemos una transformación logarítmica para uno de los controles:

$$Y_i = \beta_0 + \beta_1 \ln(X_{1i}) + \cdots + \beta_K X_{Ki} + U_i \tag{3.30}$$

En el caso de (3.30), la interpretación de β_1 surgiría de:

$$\beta_1 = \frac{\partial Y_i}{\partial \ln(X_{1i})} = \frac{\partial Y_i}{\partial X_{1i} / X_{1i}}$$

Sin embargo, esto llevaría a una interpretación poco intuitiva ya que si asumimos que el denominador cambia en una unidad (i.e. $\frac{\partial X_{1i}}{X_{1i}} = 1$) implicaría que X_{1i}

aumenta en 100 %. Para hacer más intuitiva la interpretación multiplicamos y dividimos entre 100 de manera tal que obtenemos:

$$\beta_1 = \frac{\partial Y_i}{\partial \ln(X_{1i})} = \frac{\partial Y_i}{\partial X_{1i}/X_{1i}} \cdot \frac{100}{100}$$

Ahora es más intuitivo asumir que el denominador cambia en una unidad (i.e. $\partial X_{1i}/X_{1i} \cdot 100 = 1$) ya que esto implicaría que X_{1i} aumenta en 1 % (i.e. $\partial X_{1i}/X_{1i} = 1/100$). Sin embargo, el cambio realizado afecta también al numerador. Si asumimos que el denominador es igual a 1, la ecuación anterior resulta en:

$$\beta_1 = \partial Y_i \cdot 100$$

Esto quiere decir que el cambio en Y_i es de $\beta_1/100$. Por lo tanto, la interpretación correcta del coeficiente β_1 en (3.30) es que, *caeteris paribus*, un cambio de X_{1i} en 1 % conlleva un cambio de $\beta_1/100$ unidades de Y_i. Cabe señalar que ahora las unidades de X_{1i} no influyen en la interpretación, pero las de Y_i sí.

Por último, vemos qué sucede si empleamos logaritmos de ambos lados, tanto en la variable dependiente como en el control:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{1i}) + \cdots + \beta_K X_{Ki} + U_i \tag{3.31}$$

En el caso de (3.31), la interpretación de β_1 surgiría de:

$$\beta_1 = \frac{\partial \ln(Y_i)}{\partial \ln(X_{1i})} = \frac{\partial Y_i/Y_i}{\partial X_{1i}/X_{1i}}$$

Nuevamente, dado que tenemos logaritmo del lado derecho, como en (3.30), siguiendo la misma estrategia multiplicamos y dividimos entre 100 para interpretar el cambio en X_{1i} como un cambio de 1 %, por lo tanto tendremos:

$$\beta_1 = \frac{\partial Y_i/Y_i}{\partial X_{1i}/X_{1i}} \cdot \frac{100}{100}$$

Para interpretar β_1 vemos que un cambio de una unidad en el denominador (i.e. $\partial X_{1i}/X_{1i} \cdot 100 = 1$) equivale a un aumento de 1 % en X_{1i} (i.e. $\partial X_{1i}/X_{1i} = 1/100$). Entonces, si el denominador es igual a 1 nos queda que $\beta_1 = 100 * \partial Y_i/Y_i$. En este caso, ya no hay necesidad de multiplicar por 100, ya que el cambio porcentual de Y_i (i.e. $\partial Y_i/Y_i$) ya está expresado en tasa debido a que esta multiplicado por 100. En resumen, para interpretar β_1 no hace falta multiplicar o dividir entre 100 como en los casos anteriores. Simplemente tendremos que, *caeteris paribus*, un aumento de 1 % en X_{1i} conlleva un cambio de β_1 % en Y_i. En este caso, las unidades de Y_i y X_{1i} no importan.

La siguiente tabla resume todas las posibles interpretaciones de la transformación logarítmica dependiendo de la posición del logaritmo en la especificación lineal:

Var. Dep (Y)	Var. Indep (X_1)	Interpretación del coef. de X_1 (β_1)
$\ln(Y)$	X_1	$\Delta X_1 = 1 \Rightarrow \Delta Y = (100\beta_1)\,\%$
Y	$\ln(X_1)$	$\Delta X_1 = 1\,\% \Rightarrow \Delta Y = (\beta_1/100)$
$\ln(Y)$	$\ln(X_1)$	$\Delta X_1 = 1\,\% \Rightarrow \Delta Y = (\beta_1)\,\%$

3.4.5. Variables Dummy

Otra interpretación interesante surge cuando la variable dependiente es una variable dummy (i.e. una variable dicotómica $X_1 = \{0, 1\}$). En estos casos no tiene sentido interpretar a β_1 como el cambio en Y tras un cambio marginal en la variable independiente ($\Delta X_1 = 1$).

Para entender la interpretación del coeficiente de una variable dummy es conveniente repasar que representa la variable en el caso de una regresión lineal simple. Supongamos que la variable dummy representa ser mujer (i.e. $X_{1i} = 1$ si el individuo i es mujer y $X_{1i} = 0$ si es hombre). Supongamos que estimamos la siguiente regresión:

$$Calif_i = \beta_0 + \beta_1 X_{1i} + U_i \tag{3.32}$$

donde $Calif_i$ representa el promedio escolar de la persona i.

En este caso la media condicional de ser mujer y hombre, respectivamente son:

$$E(Calif_i|X_{1i} = 1) = \beta_0 + \beta_1$$
$$E(Calif_i|X_{1i} = 0) = \beta_0$$

Por lo tanto, β_1 representa la diferencia en calificación de ser mujer respecto de ser hombre:

$$\beta_1 = E(Calif_i|X_{1i} = 1) - E(Calif_i|X_{1i} = 0) \tag{3.33}$$

Similarmente, supongamos que una persona se caracteriza por la región en la que vive y existen cuatro distintas posibilidades de regiones: {NE, NW, SE, SW}. Para expresar esto en una regresión podemos crear variables dummy para cada región (i.e. $NE_i = 1$ si i vive en el NE, $NE_i = 0$ *eoc*) y estimar la siguiente regresión:

$$Calif_i = \beta_0 + \beta_1 NE_i + \beta_2 NW_i + \beta_3 SE_i + U_i$$

Aquí es importante dejar una de las cuatro regiones como región omitida (o de referencia). Esto es necesario, ya que de lo contrario la variable constante será multicolinear con las cuatro variables regionales. Matemáticamente, si hacemos esto, la matriz $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)$ no sería invertible, por lo tanto, no podríamos estimar los coeficientes.

En el ejemplo anterior, β_3 será interpretada como la calificación de un individuo promedio que habita en la región SE respecto a la de un individuo promedio

que habita en la región SW. La selección de la región de referencia fue subjetiva en este caso. Cualquier región podría haber sido elegida como la de referencia.

Ejemplo: ¿Cómo cambiarían los coeficientes si la región de referencia elegida hubiera sido NW?

3.4.6. Interacciones

Las interacciones resultan de utilizar el producto de dos variables como una variable independiente adicional. Generalmente, dicho producto ocurre entre dos variables dummy o una variable dummy y una continua.

3.4.6.1. Interacción de dos variables tipo dummy

Consideremos un modelo donde queremos explicar el total de horas laboradas basado en el sexo y estado civil del individuo. Sea X_{1i} una variable indicador para mujer ($X_{1i} = 1$ si i es mujer y $X_{1i} = 0$ si es hombre) y X_{2i} una variable indicador para estado civil ($X_{2i} = 1$ si i es casado y $X_{2i} = 0$ si es soltero). Tomemos el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + U_i \tag{3.34}$$

Para entender la interpretación del coeficiente de la interacción (β_3) consideremos nuevamente medias condicionales:

$$\begin{aligned}
[A] &= E[Y_i | X_{1i} = 0, X_{2i} = 0] = \beta_0 \\
[B] &= E[Y_i | X_{1i} = 0, X_{2i} = 1] = \beta_0 + \beta_2 \\
[C] &= E[Y_i | X_{1i} = 1, X_{2i} = 0] = \beta_0 + \beta_1 \\
[D] &= E[Y_i | X_{1i} = 1, X_{2i} = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

Por lo tanto tenemos que:

$$\begin{aligned}
\beta_2 &= [B] - [A] \\
\beta_2 + \beta_3 &= [D] - [C] \\
\beta_3 &= ([D] - [C]) - ([B] - [A])
\end{aligned}$$

β_2 representa horas extra que trabaja un hombre casado respecto a un soltero. $\beta_2 + \beta_3$ representa horas extra que trabaja una mujer casada respecto a una soltera. β_3 representa si estar casado respecto a estar soltero representa una diferencia mayor para una mujer que para un hombre. Dicho de otra manera, indica si el plus de horas de trabajo por estar casado es diferente para una mujer que para un hombre.

3.4.6.2. Interacción de una variable tipo dummy y una variable continua

Supongamos ahora que una de nuestras variables es continua. En vez de la variable dummy casado/soltero supongamos que tenemos una variable que indica el nivel de educación de la persona. Podría interresarnos analizar si los retornos a un año adicional de educación son similares entre hombres y mujeres. Para ello establecemos el siguiente modelo:

$$Ing_i = \beta_0 + \beta_1 Educ_i + \beta_2 Mujer_i + \beta_3 Educ_i \cdot Mujer_i \tag{3.35}$$

En este modelo Ing_i representa el ingreso mensual de una persona medida en pesos; $Educ_i$, el nivel educativo medido en años completados; y $Mujer_i$ es una variable dummy igual a 1 si la persona es mujer.

Utilizando este modelo, los retornos a un año adicional de educación para los hombres son:

$$\begin{aligned} E(Ing_i|Mujer_i = 0) &= \beta_0 + \beta_1 Educ_i \\ \frac{\partial E(Ing_i|Mujer_i = 0)}{\partial Educ_i} &= \beta_1 \end{aligned}$$

Similarmente, los retornos a un año adicional de educación para las mujeres son:

$$\begin{aligned} E(Ing_i|Mujer_i = 1) &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Educ_i \\ \frac{\partial E(Ing_i|Mujer_i = 1)}{\partial Educ_i} &= \beta_1 + \beta_3 \end{aligned}$$

Por lo tanto, β_3 representa la diferencia en los retornos entre hombres y mujeres. Dicho de otra manera, β_3 indica qué tan mayores son los retornos a un año adicional de educación para las mujeres que para los hombres.

$$\beta_3 = \frac{\partial E(Ing_i|Mujer_i = 1)}{\partial Educ_i} - \frac{\partial E(Ing_i|Mujer_i = 0)}{\partial Educ_i}$$

3.4.6.3. Interacción de dos variables continuas

Por último, podemos utilizar un modelo con interacciones donde las dos variables involucradas en la interacción son variables continuas. En este caso, el modelo nos puede indicar si los rendimientos marginales respecto a una variable dependen de la otra variable incluida en la interacción.

Por ejemplo, supongamos que especificamos el siguiente modelo:

$$Ing_i = \beta_0 + \beta_1 Educ_i + \beta_2 Calidad_i + \beta_3 Educ_i * Calidad_i \tag{3.36}$$

donde $Calidad_i$ es un índice que mide la calidad de la educación recibida. En este caso los retornos por un año adicional de educación serán:

$$\frac{\partial E(Ing_i)}{\partial Educ_i} = \beta_1 + \beta_3 Calidad_i$$

Lo que nos indica el resultado anterior es que dichos retornos a la educación pueden depender del nivel de calidad.

%Discutir en este caso como se obtienen los coeficientes (β's). Dar un ejemplo con retornos a horas de práctica en un video juego.

3.5. El estadístico R^2

Una medida comúnmente utilizada para describir qué función describe de mejor manera los datos empíricamente es el estadístico R^2. Cabe señalar que no es recomendable elegir el modelo a utilizar basándose únicamente en el estadístico R^2. Este estadístico aumenta si incrementamos la cantidad de variables a incluir en la regresión. Sin embargo, no en todos los casos es conveniente añadir controles en una regresión, especialmente si nos interesa llevar a cabo inferencia causal. Antes de recurrir a esta alternativa se recomienda utilizar la teoría para justificar el modelo a emplear.

El estadístico R^2 se calcula como:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\ &= 1 - \frac{\sum_{i=1}^{n}\widehat{U}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\ &= 1 - \frac{SSR}{SST} \end{aligned} \tag{3.37}$$

El estadístico R^2 representa la parte de la varianza de Y_i que es explicada por el modelo MCO.

3.6. Modelo de Probabilidad Lineal

Ahora veremos qué sucede en casos en los cuales la variable dependiente es una decisión binaria. En este caso, nuestra variable dependiente será una variable

tipo dummy ($Y_i = \{0, 1\}$). En este caso, nos referiremos a la estimación como un **modelo de probabilidad lineal**, el cual consiste en simplemente estimar un modelo de MCO con errores heterocedásticos:

$$Y_i = X_i'\beta + U_i$$

En este caso, aplican los supuestos del modelo de MCO, tal como lo revisamos anterioremente. Sin embargo, una particularidad es la interpretación de los coeficientes. Dado que nuestra variable dependiente es una variable binaria, el modelo se puede especificar de la siguiente forma:

$$E(Y_i|X_i) = 1 \cdot Pr(Y_i = 1|X_i) + 0 \cdot Pr(Y_i = 0|X_i) = Pr(Y_i = 1|X_i) = X_i'\beta$$

En clase mostramos cómo se ve este modelo gráficamente.

Por lo tanto, cada coeficiente (β_j) representará (*caeteris paribus*) el cambio en la probabilidad (medida en puntos porcentuales) de que $Y_i = 1$ por un cambio marginal en X_j. Ejemplo, si nuestra variable dependiente es inscribirse o no a la universidad y una de las variables independientes es la educación de la madre, el coeficiente de esta variable representará el cambio de la probabilidad (medido en puntos porcentuales) de que el individuo se inscriba a la universidad por un aumento de un año en la educación de la madre.

En el caso del modelo de probabilidad lineal, los errores **siempre** deberán asumirse como heterocedásticos, ya que:

$$Var(Y_i|X_i) = Pr(Y_i = 1|X_i) \cdot (1 - Pr(Y_i = 1|X_i)) = X_i'\beta(1 - X_i'\beta)$$

Sin embargo, un problema de este modelo, por su diseño lineal, es que es posible que nos lleve a predecir valores imposibles (es decir, fuera del rango 0-1) para la variable dependiente. Existen modelos que corrigen este problema y estiman valores de la probabilidad estrictamente entre 0 y 1, siendo *probit* y *logit* los casos más conocidos. Estos modelos forman parte de la categoría de *modelos de estimadores de máxima verosimilitud (MLE).*

3.7. Sesgo por variables omitidas

En esta sección, el propósito es estimar cuál es la diferencia en los coeficientes entre estimar una regresión con K variables (que llamaremos regresión larga por simplicidad) utilizando MCO y estimar otra regresión que omita una de esas variables (por simplicidad omitiremos la variable X_K).

Sea la regresión larga:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K X_K + U \tag{3.38}$$

Sea la regresión omitiendo la variable X_K (regresión corta):

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1} + V \tag{3.39}$$

Queremos establecer cuál es la relación entre β_j y α_j para $j = \{0, 1, \ldots, K-1\}$.

Definamos a la regresión residual como:

$$X_K = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + W \tag{3.40}$$

Sustituyendo (3.40) en (3.38) obtenemos:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K(\gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + W) + U$$

Reordenando los términos obtenemos:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K W + U \tag{3.41}$$

donde $\tilde{\beta}_j = \beta_j + \beta_K \gamma_j$ para $j = \{0, 1, \ldots, K-1\}$.

Nótese que ésta última regresión es lo mismo que (8.1) si definimos al error como $V = \beta_K W + U$. Por lo tanto, las condiciones de primer órden que nos llevarían a resolver ésta ultima regresión y (8.1) son las mismas. Esto implica que $\tilde{\beta}_j = \alpha_j$. Por lo tanto, el sesgo se define como:

$$\alpha_j - \beta_j = \beta_K \gamma_j$$

El signo del sesgo estará entonces definido por los signos de los coeficientes β_K y γ_j.

3.8. Validez externa e interna

En esta nota hemos revisado la manera en la que se deriva el modelo MCO. En la sección anterior mencionamos el problema más común para hacer inferencia causal utilizando los resultados del modelo MCO (sesgo por variables omitidas). Sin embargo, existen otras consideraciones importantes cuando queremos llevar a cabo inferencia utilizando nuestros resultados.

3.9. Validez externa

Hay validez externa cuando los resultados de un análisis llevado a cabo con una muestra de una población específica son válidos y generalizables para otras poblaciones.

Ejemplo: Llevar a cabo análisis del impacto de una política social para hogares pobres en México (e.g. Progresa/Oportunidades) puede tratar de generalizarse para otros países en desarrollo de características similares a las de México.

Los problemas más comunes relacionados con validez externa son:

- **Diferencias en las poblaciones.** Por ejemplo, en medicina se hacen muchos estudios con ratones y siempre existe la controversia si el efecto será similar en seres humanos.
- **Diferencias en contexto.** Los resultados pueden no ser generalizables a poblaciones similares si características del contexto pueden causar diferencias importantes. Algunos ejemplos incluyen diferencias geográficas, diferencias en leyes, diferencias ambientales, etc.

3.10. Validez interna

Hay validez interna cuando los resultados de un análisis llevado a cabo con una muestra son válidos y generalizables para la población de la cual se extrajo dicha muestra.

Los problemas más comunes relacionados con validez interna son:

- **Sesgo por variables omitidas.** Discutido la sección anterior.
- **Especificación incorrecta de la forma funcional.** Discutido en la sección **??**.
- **Error de medición.** En algunos casos, es posible que una o más variables disponibles en la base de datos a utilizar en el análisis tengan errores. Dichos errores pueden surgir por errores en la medición, en la captura de datos, que los encuestados no recuerden precisamente algún dato, preguntas mal estructuradas o poco claras, respuestas falsas intencionales, etc.

Para determinar cuál puede ser el efecto de llevar a cabo una regresión con errores en las variables consideremos lo siguiente:

Sea $\tilde{X}$ la variable X medida con error, donde el error se define como: $w_i = \tilde{X}_i - X_i$.

Por simplicidad tomemos el caso de una regresión simple, donde queremos estimar el coeficiente β_1:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \tag{3.42}$$

Sin embargo, dado que en nuestra base de datos únicamente tendremos disponible a $\tilde{X}_i$, realmente estaremos estimando la siguiente regresión:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X}_i + V_i \tag{3.43}$$

Nótese que el error cambió en este caso, ya que: $V_i = \beta_1 X_i - \tilde{\beta}_1 \tilde{X}_i + U_i$.

En el caso de la regresión (3.43), el coeficiente estimado será:

$$\tilde{\beta}_1 = \frac{Cov(\tilde{X}, Y)}{Var(\tilde{X})}$$

En este caso nos interesará determinar la relación entre $\tilde{\beta}_1$ y β_1, donde:

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

Para establecer la relación:

$$\begin{aligned}
\tilde{\beta}_1 = \frac{Cov(\tilde{X}, Y)}{Var(\tilde{X})} &= \frac{Cov(X + w, Y)}{Var(X + w)} \\
&= \frac{Cov(X, Y)}{Var(X)} \frac{Var(X)}{Var(X + w)} + \frac{Cov(w, Y)}{Var(X + w)} \\
&= \beta_1 \left(\frac{\sigma_X^2}{Var(X + w)} \right) + \frac{Cov(w, \beta_0 + \beta_1 X + U)}{Var(X + w)}
\end{aligned}$$

Si suponemos que los errores no están correlacionados $Cov(w, U) = 0$:

$$\tilde{\beta}_1 = \beta_1 \left(\frac{\sigma_X^2 + Cov(w, X)}{Var(X + w)} \right)$$

Un caso particular se conoce como el error de medición clásico. Este tipo de error asume que los errores de medición (w) no están correlacionados con el valor real de X. Este tipo de errores pueden surgir, por ejemplo, si hay errores aleatorios de captura.

En este caso, tendremos que $Cov(w, X) = 0$, por lo tanto:

$$\tilde{\beta}_1 = \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \right) \tag{3.44}$$

Esto quiere decir que hay un sesgo de atenuación. Es decir, el coeficiente estimado en la regresión que utiliza las variables con errores de medición (3.43) será menor en valor absoluto que el coeficiente que resultaría si se utilizarán los valores sin error.

Debe tenerse en cuenta que en algunos casos los errores de medición pueden ser sistemáticos, por lo tanto, asumir que los errores de medición son clásicos puede ser incorrecto. Por ejemplo, en preguntas tales como ingresos mensuales, número de bebidas alcohólicas consumidas regularmente por semana, horas de ejercicio o estudio a la semana, etc. Es común pensar que hay errores sistemáticos reportando los datos por parte de los que contestan las encuestas. En este caso, la dirección del sesgo estará influido también por el factor $Cov(w, X)$.

- **Datos faltantes y sesgo muestral.** El problema de datos faltantes se refiere a que algunas observaciones en la base de datos reportaron de manera incompleta la información requerida o dicha información no estaba disponible. El sesgo muestral surge si los entes de la población seleccionados para formar la muestra no son representativos de la población en general. Estos problemas son relevantes porque amenazan el supuesto de i.i.d. Existen métodos estadísticos diseñados para tratar con estos problemas, pero dichos métodos no están incluidos en el temario del curso.
- **Causalidad simultánea.** Este problema surge cuando además de haber una relación causal de la variable X a Y, también existe una relación causal de Y a X. Dado que la estimación de los coeficientes implica calcular la correlación condicional entre las variables, la estimación estará sesgada. Ejemplo: Supongan que les interesa ver la relación entre tamaño del salón de clase y calificaciones. Supongan que el gobierno establece una iniciativa que dice que las escuelas con mejores resultados recibirán mas recursos. Esto puede llevar a la contratación de maestros y por lo tanto, a la disminución del tamaño del salón de clases. En este caso, establecer la relación causal de tamaño de clases a calificaciones requerirá que se recurran a otros métodos, como variables instrumentales y métodos experimentales que cubriremos más adelante en el curso.

3.11. Variaciones al modelo de Mínimos Cuadrados Ordinarios

3.11.1. Regresiones cuantílicas

El problema de mínimos cuadrados ordinarios consiste en minimizar:

$$E(Y_i - X_i'\beta)^2 \tag{3.45}$$

Esta estimación nos da como resultado una estimación de la media de Y_i condicional en las variables que se eligen como controles en el modelo, X_i. Sin embargo, en algunos casos nos puede interesar hacer inferencia para distintos partes de la distribución de Y_i. Por ejemplo, supongamos que nos interesa medir los retornos educativos para distintas partes de la distribución del ingreso

(para los individuos de mayores y menores ingresos). En este caso modificamos el problema de minimización planteado hacia el siguiente:

$$\min_{\beta_\tau} \quad \rho_\tau |Y_i - X_i' \beta_\tau| \tag{3.46}$$

donde ρ_τ es un ponderador de los errores absolutos y τ es un indicador del cuantil 100τ (Figure: 3.3):

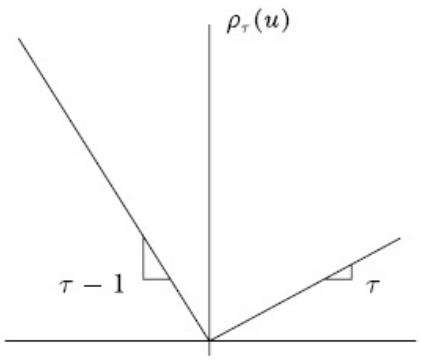


Figura 3.3: Ponderador de errores

Por ejemplo, $\rho_{0,5}$ es el ponderador para el cuantil 50. Utilizando este ponderador generaremos un estimador de la mediana de Y_i utilizando las variables explicativas X_i. Similarmente, $\rho_{0,9}$ es el ponderador para el cuantil 90 y generará un estimador del cuantil 90 de Y_i utilizando las variables explicativas X_i.

Ejemplo con retornos educativos.

3.11.2. Mínimos cuadrados ponderados

Como vimos en la sección 3.2, el supuesto de homocedasticidad asume que la varianza de los errores no depende del nivel de X_i. Como describimos, el incentivo a utilizar homocedasticidad es que conlleva mayor eficiencia en la estimación de los coeficientes de MCO. El modelo de mínimos cuadrados ponderados (Weighted Least Squares, WLS) consiste en llevar a cabo una transformación de las variables para poder estimar el modelo asumiendo homocedasticidad y obtener estimadores más eficientes.

Este modelo se basa en la idea que la varianza del error depende de una función de X_i:

$$Var(Ui|Xi) = \sigma^2 h(X_i) \tag{3.47}$$

Sin embargo, este modelo asume que la función $h(\cdot)$ es conocida y se puede estimar. Generalmente este es un supuesto restrictivo.

Suponiendo, que la función $h(\cdot)$ se puede estimar, los pasos para llevar a cabo la estimación de este modelo son los siguientes:

Cada una de las variables se divide entre la raíz de $h(X_i)$:

$$Y_i^* = \frac{Y_i}{\sqrt{h(X_i)}} \qquad X_{1i}^* = \frac{X_{1i}}{\sqrt{h(X_i)}}$$
$$X_{Ki}^* = \frac{X_{Ki}}{\sqrt{h(X_i)}} \qquad U_i^* = \frac{U_i}{\sqrt{h(X_i)}}$$

Y se puede estimar el siguiente modelo:

$$Y_i^* = \tilde{\beta}_0 + \beta_1 X_{1i}^* + \cdots + \beta_K X_{Ki}^* + U_i^* \tag{3.48}$$

En este caso puede verse que el error tendrá una varianza homocedástica (constante) y por lo tanto podemos estimar los parámetros (β's) utilizando el modelo MCO. Es importante tomar en cuenta que:

- El valor estimado de los parámetros cambiará, pero el estimador sigue siendo consistente
- Este nuevo valor estimado tendrá errores estándar menores (típicamente). Por lo tanto, la estimación será más eficiente.

Sin embargo, el supuesto importante es que la forma de la función $h(\cdot)$ es conocida. Además, otro inconveniente es que la función $h(\cdot)$ no puede ser negativa para ningún valor de X_i (de otra manera estaríamos diciendo que la varianza pudiera ser negativa para algunos valores de X_i). Una manera de evitar esto es con el modelo de *FGLS* (Feasible Generalized Least Squares - Mínimos Cuadrados Generalizados Factibles). Dicho modelo asume la siguiente especificación para la función $h(\cdot)$:

$$h(X_i) = \exp\left(\delta_0 + \delta_1 X_{1i} + \cdots + \delta_K X_{Ki}\right) \tag{3.49}$$

En este caso, la función $h(\cdot)$ cumplirá la condición de ser positiva en todo el dominio de X_i.

Los pasos a seguir para llevar a cabo la estimación FGLS son:

1. Estima la regresión $Y_i = \beta_0 + \beta_1 X1i + \cdots + \beta_K XKi$
2. Obtén los residuales: $\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X1i - \cdots - \hat{\beta}_K XKi$
3. Estima la regresión $\log(\hat{U}_i^2) = \delta_0 + \delta_1 X_{1i} + \cdots + \delta_K X_{Ki}$
4. Obtén los valores estimados de $h(X_i)$: $\widehat{h(X_i)}$
5. Sigue los pasos del modelo *WLS* utilizando los $\widehat{h(X_i)}$ estimados para modificar la variable dependiente y las independientes

Este procedimiento resulta en un estimador consistente y asintóticamente más eficiente que MCO.

Capítulo 4

Efectos Fijos y Aleatorios

En la nota anterior concluimos que el modelo MCO es extremadamente útil, ya que con muestras grandes y pocos supuestos podemos estimar relaciones funcionales entre dos variables que son sencillas de interpretar. Sin embargo, subrayamos que una limitación importante del modelo cuando queremos establecer relaciones causales es el sesgo por variables omitidas. En esta nota se explicará un modelo que para resolver el problema de sesgo por variables omitidas impone supuestos en dichas variables omitidas y hace uso de una estructura de datos de panel.

4.1. Datos de Panel

Tradicionalmente nos referimos a **datos de panel** (o datos longitudinales) cuando nuestra unidad de observación (ya sea un individuo, una familia, una empresa, un país, etc.) es captada en dos o más ocasiones a través del tiempo. Es importante que la misma unidad (i.e. el mismo individuo, la misma empresa, etc.) sea captada en cada momento a través del tiempo[1]. En este caso, la notación que utilizaremos para las variables es Y_{it}, donde i es el subíndice que indica al individuo y t el que indica tiempo. Por ejemplo, si tenemos una muestra de ingreso y educación para 100 individuos en 3 años, nuestra muestra será $(Ing_{it}, Educ_{it})_{(i=1,\ldots,100),(t=1,2,3)}$. En particular, $(Ing_{54,2}, Educ_{54,2})$ representará el ingreso y la educación para el individuo 54 en el año 2.

Sin embargo, cabe resaltar que en la especificación del modelo no es estrictamente necesario que t represente tiempo. Por ejemplo, i podría representar familias

[1] Tomar muestras independientes en distintos momentos del tiempo constituye una base de datos transversal agrupada (pooled cross-section). Esto es diferente que una base de datos de panel y no es útil para derivar los modelos de esta nota.

y t podría representar hermanos; i podría representar municipios y t ciudades; i podría representar partidos políticos y t candidatos.

La clave de la estructura de la base de datos de panel es que exista un factor en común -i- que permita agrupar a más de una observación. En el ejemplo clásico de distintas observaciones a través del tiempo, el factor común es el individuo, quien es observado en distintos momentos. En los otros ejemplos planteados, las familias, los municipios y los partidos políticos son factores que comparten distintas observaciones.

Se dice que una base de datos de panel es **balanceada** cuando cada individuo es observado en *todos* los momentos del tiempo. En el contexto de los otros ejemplos que planteamos, esto quiere decir que cada agrupación i tiene la misma cantidad de componentes (i.e. cada familia de la muestra tiene la misma cantidad de hermanos, etc.). Muchos de los modelos presentados a continuación no requieren de un panel balanceado forzosamente, pero es importante tomar en cuenta (especialmente en el ejemplo clásico de unidades observadas a lo largo del tiempo) que un panel no balanceado puede resultar de pérdida de observaciones a lo largo del tiempo. Esto se conoce como abandono de la muestra (*sample attrition*) y puede ser un factor relevante, ya que podría estar relacionado con tener una muestra sesgada.

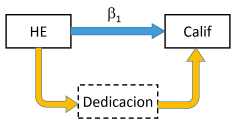
Modelo:

$$Calif_{it} = \beta_0 + \beta_1 HE_{it} + \beta_2 Mujer_{it} + \beta_3 Mat_{it} + \beta_4 Cafe_{it} + U_{it}$$

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	Tazas café al día
1	1	6.8	4	0	5	1
1	2	7.9	6.5	0	6	1.2
2	1	8.7	10	1	5	2.5
2	2	9.1	11	1	5	2.5
3	1	8.3	5	1	4	3
3	2	8.0	4.7	1	6	2.5

Problema: Sesgo por variables omitidas!

- $Calif_{it} = \beta_0 + \beta_1 HE_{it} + \beta_2 Mujer_{it} + \beta_3 Mat_{it} + \beta_4 Cafe_{it} + U_{it}$



Remover el sesgo controlando por "Dedicacion"

$$Calif_{it} = \beta_0 + \beta_1 HE_{it} + \beta_2 Mujer_{it} + \beta_3 Mat_{it} + \beta_4 Cafe_{it} + \beta_5 Dedic_{it} + U_{it}$$

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	Tazas café al día	Dedicacion
1	1	6.8	4	0	5	1	A
1	2	7.9	6.5	0	6	1.2	A
2	1	8.7	10	1	5	2.5	B
2	2	9.1	11	1	5	2.5	B
3	1	8.3	5	1	4	3	C
3	2	8.0	4.7	1	6	2.5	C

$$Calif_{i2} = \beta_0 + \beta_1 HE_{i2} + \beta_2 Mujer_i + \beta_3 Mat_{i2} + \beta_4 Cafe_{i2} + \beta_5 Dedic_i + U_{i2}$$
$$Calif_{i1} = \beta_0 + \beta_1 HE_{i1} + \beta_2 Mujer_i + \beta_3 Mat_{i1} + \beta_4 Cafe_{i1} + \beta_5 Dedic_i + U_{i1}$$

$$\Delta Calif_i = \beta_0 + \beta_1 \Delta HE_i + \beta_2 \Delta Mujer_i + \beta_3 \Delta Mat_i + \beta_4 \Delta Cafe_i + \beta_5 \Delta Dedic_i + \Delta U_i$$

Cabe notar:

- Ya no hay coeficientes de tiempo en el resultado de la resta (¿por qué?)
- Algunas variables tendrían que desaparecer (¿cuáles y por qué?)

$$\Delta Calif_i = \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \beta_4 \Delta Cafe_i + \Delta U_i$$

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	Tazas café al día	Dedicacion
1	1	6.8	4	0	5	1	A
1	2	7.9	6.5	0	6	1.2	A
2	1	8.7	10	1	5	2.5	B
2	2	9.1	11	1	5	2.5	B
3	1	8.3	5	1	4	3	C
3	2	8.0	4.5	1	6	2.5	C

Clave unica (i)	Semestre (t)	ΔCalificacion	ΔHoras estudio	Mujer	ΔNum materias	ΔTazas café al día	Dedicacion
1		1.1	2.5		1	0.2	
2		0.4	1		0	0	
3		- 0.3	-0.5		2	-0.5	

Limitaciones:

- No es posible estimar el coeficiente de ninguna variable que no cambie a lo largo del tiempo (ej. Mujer)
- ¿Por qué?
- No es posible estimar β_0
  - ¿Qué quiere decir entonces la constante en la regresión donde usamos las diferencias?

$$Calif_{i2} = \beta_0 + \beta_1 HE_{i2} + \beta_3 Mat_{i2} + \beta_4 Cafe_{i2} + \beta_5 Dedic_i + \delta_1 1\{t=2\}^{1} + U_{i2}$$
$$Calif_{i1} = \beta_0 + \beta_1 HE_{i1} + \beta_3 Mat_{i1} + \beta_4 Cafe_{i1} + \beta_5 Dedic_i + \delta_1 1\{t=2\}^{0} + U_{i1}$$

$$\Delta Calif_i = \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \beta_4 \Delta Cafe_i + \delta_1 + \Delta U_i$$

$$\Delta Calif_i = \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \delta_1 + \Delta U_i$$

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	1{t=2}	Dedicacion
1	1	6.8	4	0	5	0	A
1	2	7.9	6.5	0	6	1	A
2	1	8.7	10	1	5	0	B
2	2	9.1	11	1	5	1	B
3	1	8.3	5	1	4	0	C
3	2	8.0	4.5	1	6	1	C

Clave unica (i)	Semestre (t)	ΔCalificacion	ΔHoras estudio	Mujer	ΔNum materias	Δ1{t=2}	Dedicacion
1		1.1	2.5		1	1	
2		0.4	1		0	1	
3		- 0.3	-0.5		2	1	

- ¿Qué sucedería si tenemos más periodos de tiempo?
  - Una alternativa es producir más de un valor estimado para nuestro coeficiente de interés
  - Otra alternativa, es usar toda la información para producir un solo valor estimado
  - Veamos qué sucedería con 3 períodos de tiempo

Opción 1:

$$Calif_{i3} = \beta_0 + \beta_1 HE_{i3} + \beta_3 Mat_{i3} + \beta_5 Dedic_i + \delta_3 1\{t=3\}^{1} + \delta_2 1\{t=2\}^{0} + U_{i3}$$
$$Calif_{i2} = \beta_0 + \beta_1 HE_{i2} + \beta_3 Mat_{i2} + \beta_5 Dedic_i + \delta_3 1\{t=3\}^{0} + \delta_2 1\{t=2\}^{1} + U_{i2}$$

$$\Delta_2 Calif_i = \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \delta_3 - \delta_2 + \Delta U_i$$

Opción 2:

$$Calif_{i2} = \beta_0 + \beta_1 HE_{i2} + \beta_3 Mat_{i2} + \beta_5 Dedic_i + \delta_3 1\{t=3\}^{0} + \delta_2 1\{t=2\}^{1} + U_{i2}$$
$$Calif_{i1} = \beta_0 + \beta_1 HE_{i1} + \beta_3 Mat_{i1} + \beta_5 Dedic_i + \delta_3 1\{t=3\}^{0} + \delta_2 1\{t=2\}^{0} + U_{i1}$$

$$\Delta_1 Calif_i = \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \delta_2 + \Delta U_i$$

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	1{t=3}	1{t=2}	Dedicacion
1	1	6.8	4	0	5	0	0	A
1	2	7.9	6.5	0	6	0	1	A
1	3	8.5	8	0	5	1	0	A
2	1	8.7	10	1	5	0	0	B
2	2	9.1	11	1	5	0	1	B
2	3	9.0	10.5	1	6	1	0	B

Clave unica (i)	Semestre (t)	ΔCalificacion	ΔHoras estudio	Mujer	ΔNum materias	Δ1{t=3}	Δ1{t=2}	Dedicacion
1		0.6	1.5		-1	1	-1	
2		-0.1	-0.5		1	1	-1	

Clave unica (i)	Semestre (t)	ΔCalificacion	ΔHoras estudio	Mujer	ΔNum materias	Δ1{t=3}	Δ1{t=2}	Dedicacion
1		1.1	2.5		1	0	1	
2		0.4	1		0	0	1	

Clave unica (i)	1{Dif T3-T2}	ΔCalificacion	ΔHoras estudio	Mujer	ΔNum materias	Δ1{t=3}	Δ1{t=2}	Dedicacion
1	1	0.6	1.5		-1	1	-1	
2	1	-0.1	-0.5		1	1	-1	
1	0	1.1	2.5		1	0	1	
2	0	0.4	1		0	0	1	

$$\Delta_1 Calif_i = \beta_0 + \beta_1 \Delta HE_i + \beta_3 \Delta Mat_i + \beta_4 1\{DifT3 - T2\} + \Delta U_i$$

¿Qué representan β_0 y β_4?

- β_0 representa a δ_2
- β_4 representa a δ_3-$2\delta_2$

Otra alternativa es generar Δ1{t=3} y Δ1{t=2} y sus coeficientes correspondientes serán δ_3 y δ_2 - Esta estimación no debería incluir constante por multicolinearidad

- La estrategia anterior se puede generalizar a T periodos de tiempo.
- Hay que generar las diferencias de tiempo
  - Utilizar como variables dependientes y explicativas cambios en tiempo con respecto al periodo anterior
  - En la especificación hay que agregar dummies de tiempo

$$\Delta_t Calif_i = \beta_0 + \beta_1 \Delta_t HE_i + \beta_3 \Delta_t Mat_i + \delta_t + \Delta_t U_i$$

donde:

- $\Delta_t Calif_i = Calif_{i,t} - Calif_{i,t-1}$
- Similar para $\Delta_t HE_i$, $\Delta_t Mat_i$ y $\Delta_t U_i$ para t={2,3,...,T}
- δ_t representa agregar dummies de tiempo para las diferencias, como en el slide anterior

Otra alternativa: restar las medias a nivel individuo

$$\bar{\Delta}_t Calif_i = \beta_0 + \beta_1 \bar{\Delta}_t HE_i + \beta_3 \bar{\Delta}_t Mat_i + \delta_t + \bar{\Delta}_t U_i$$

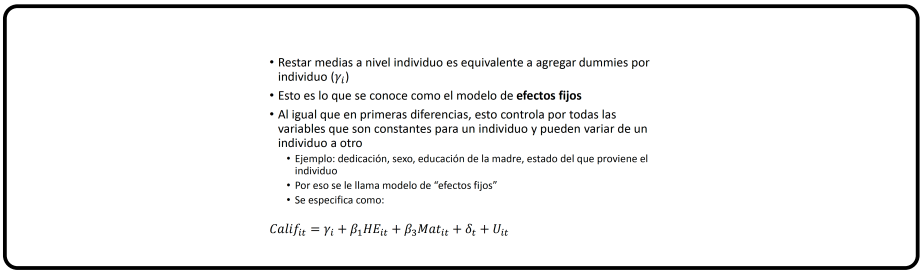
donde:

- $\bar{\Delta}_t Calif_i = Calif_{i,t} - \overline{Calif_i}$
- $\overline{Calif_i} = \frac{1}{T}\sum_{t=1}^{T} Calif_{i,t}$
- Similar para $\bar{\Delta}_t HE_i$, $\bar{\Delta}_t Mat_i$ y $\bar{\Delta}_t U_i$ para t={1,2,...,T}
- δ_t representa agregar dummies de tiempo

Clave unica (i)	Semestre (t)	Calificacion	Horas estudio	Mujer	Num materias	1{t=3}	1{t=2}	Dedicacion
1	1	6.8	4	0	5	0	0	A
1	2	7.9	6.5	0	6	0	1	A
1	3	8.5	8	0	5	1	0	A
2	1	8.7	10	1	5	0	0	B
2	2	9.1	11	1	5	0	1	B
2	3	9.0	10.5	1	6	1	0	B

7.73 8.93

Clave unica (i)	Semestre (t)	ΔCalificacion	ΔHoras estudio	ΔMujer	ΔNum materias	Δ1{t=3}	Δ1{t=2}	ΔDedicacion
1	1	-0.93	-2.17		-0.33	-0.33	-0.33	
1	2	0.17	0.33		0.67	-0.33	0.67	
1	3	0.77	1.83		-0.33	0.67	-0.33	
2	1	-0.23	-0.50		-0.33	-0.33	-0.33	
2	2	0.17	0.50		-0.33	-0.33	0.67	
2	3	0.07	0.00		0.67	0.67	-0.33	



4.2. Estimador de Primeras Diferencias (First Differences)

Empecemos por discutir un caso sencillo donde tenemos varias observaciones i en dos periodos de tiempo. Supongamos que queremos estimar los rendimientos educativos utilizando el siguiente modelo[2]:

$$\log(w_{it}) = \beta_0 + \beta_1 Educ_{it} + \beta_2 Exper_{it} + \delta_2 D2_t + \cdots + U_{it}$$

En la nota anterior señalamos que un posible problema de utilizar MCO para estimar esta especificación es que existe un problema de sesgo por variables omitidas. En particular, la variable de educación podría estar capturando el efecto de la habilidad intrínseca de cada individuo. Supongamos que dicha habilidad natural **no varia a través del tiempo** y no es observada (i.e. no esta disponible en nuestra base de datos). Esto quiere deicr que para eliminar el sesgo, nos interesaría estimar el siguiente modelo:

$$\log(w_{it}) = \beta_0 + \beta_1 Educ_{it} + \beta_2 Exper_{it} + \delta_2 D2_t + \gamma A_i + \cdots + U_{it} \qquad (4.1)$$

donde A_i representa la habilidad del individuo i. Nótese que esta variable no tiene un subíndice t, ya que la habilidad de cada individuo no varía a través del tiempo, pero si varía de un individuo a otro. En nuestro caso, solo tenemos dos momentos en el tiempo ($t = 1, 2$), por lo tanto, si tomamos la diferencia de $t = 2$ menos $t = 1$ para cada individuo:

$$\log(w_{i2}) - \log(w_{i1}) = \delta_2 + \beta_1(Educ_{i2} - Educ_{i1}) + \beta_2(Exper_{i2} - Exper_{i1}) + \cdots + (U_{i2} - U_{i1})$$

Este modelo se conoce como el **modelo de primeras diferencias**. Para estimarlo, utilizamos la metodología de MCO. Nótese que esta metodología nos permite remover cualquier sesgo por variables omitidas, siempre y cuando dichas variables omitidas no cambien a lo largo de t para cada individuo.

[2] $D2_t$ representa una dummy que indica si la observación corresponde al periodo $t = 2$.

Para poder estimar este modelo, se deben cumplir los siguientes supuestos:

1. Se deben cumplir los supuestos de MCO que establecimos en la nota anterior: (i) muestra i.i.d.; (ii) relación lineal en el modelo; (iii) no multicolinearidad

2. Los coeficientes deben ser constantes a lo largo del tiempo (eso se refleja en que los coeficientes no tienen un subíndice t en el modelo[3].

Cabe recordar que una condición de primer órden que surge al derivar el modelo MCO establece que no hay covarianza entre los errores y las variables independientes ($E(X_i U_i) = 0$). En el modelo de primeras diferencias, esto se traduce como $E((X_{it} - X_{it-1})(U_{it} - U_{it-1})) = 0$. Es decir, no debe existir covarianza entre la diferencia de los errores y la diferencia de las variables independientes.

Es importante señalar que cualquier variable omitida que no sea constante a través del tiempo para cada individuo, puede causar sesgo en este modelo. En nuestro ejemplo, una mejora alimenticia es un ejemplo de variable que podría seguir causando sesgo. Asimismo, es importante que exista variación en la variable independiente. En nuestro ejemplo, esto podría causar conflicto, ya que es de esperarse que la muestra este compuesta por individuos adultos perceptores de ingresos. En este grupo, sería de esperarse que la variable educación no cambie mucho.

Por último, cabe señalar que estos modelos no permitirán estimar el efecto de variables que no cambian a lo largo del tiempo, como género, raza, etc. Esto sucede ya que estas variables serán eliminadas también siguiendo el procedimiento que utilizamos para eliminar a A_i. Por lo tanto, la diferencia de estas variables será multicolinear con la diferencia de la variable A_i y la constante.

- Dar ejemplo de los estudios con gemelos en E.U. para eliminar el sesgo por habilidad.

- Otro ejemplo: ¿cómo el desempleo afecta las tasas de crimen?

El modelo de primeras diferencias tambien puede estimarse si se tiene más de dos observaciones para cada i a través del tiempo. Generalmente, si se tienen T periodos, el modelo sería:

$$Y_{it} - Y_{it-1} = \delta_t - \delta_{t-1} + \beta_1(X_{1it} - X_{1it-1}) + \cdots + \beta_K(X_{Kit} - X_{Kit-1}) + (U_{it} - U_{it-1})$$

En este caso, cada unidad i tendría $T - 1$ observaciones. El procedimiento consistiría en estimar la regresión utilizando MCO. Si la base de datos de panel es balanceada, se tendrían $N * (T - 1)$ observaciones.

[3]Este supuesto puede ser omitido, pero esto tendría implicaciones sobre los errores estándar de los coeficientes estimados.

4.3. Modelo de Efectos Fijos

4.3.1. Derivación

En esta sección revisaremos lo que tradicionalmente se conoce como el modelo de efectos fijos. La motivación de este modelo es la misma que la del modelo de primeras diferencias: eliminar el sesgo que puede causar una variable que sea constante dentro de un mismo grupo: A_i.

Siguiendo con nuestro ejemplo anterior (y asumiendo una base de datos de panel balanceada) tomemos la ecuación [(4.1)] y calculemos el promedio para cada individuo asumiendo en este caso que existen T periodos de tiempo:

$$\overline{\log(w_i)} = \beta_0 + \beta_1 \overline{Educ}_i + \beta_2 \overline{Exper}_i + \delta_2 \frac{1}{T} + \cdots + \delta_T \frac{1}{T} + A_i + \overline{U}_i \qquad (4.2)$$

Siguiendo el mismo procedimiento que en primeras diferencias, tomamos la diferencia entre las ecuaciones (4.1) y (4.2) y esto resulta en (para $t = 1, \ldots, T$)[4]:

$$\begin{aligned} \log(w_{it}) - \overline{\log(w_i)} = &\beta_1(Educ_{it} - \overline{Educ}_i) + \beta_2(Exper_{it} - \overline{Exper}_i) + \cdots + \\ &+ \cdots + \delta_t - \overline{\delta} + (U_{it} - \overline{U}_i) \end{aligned} \qquad (4.3)$$

Igual que en el caso del modelo de primeras diferencias, este modelo ya no incluye la variable A_i que no es observada y que potencialmente generaba sesgo. Asimismo, los supuestos establecidos en la sección de primeras diferencias deben cumplirse para poder estimar este modelo con el método de MCO. La principal diferencia entre este modelo y el de primeras diferencias radica en la condición de primer órden de la derivción de MCO, que en este caso se vuelve:

$$E((X_{it} - \overline{X}_i)(U_{it} - \overline{U}_i)) = 0$$

Esta condición es más restrictiva que la condición de primeras diferencias ya que requiere que la variable de error de cada t no este correlacionada con la variable independiente X en cada t.

[4]En este caso, $\overline{\delta}$ representa una constante y está definida como $\overline{\delta} = (\delta_2 + \cdots + \delta_T)\frac{1}{T}$. Además el modelo incluye el factor δ_t. Esto es equivalente a incluir dummies de tiempo y una constante en el modelo. Esto solo es necesario si se considera que incluir las dummies (i.e. el efecto de cada año) es relevante para el modelo. Si no se considera que el efecto de t es relevante, el modelo se simplifica mucho ya que se excluyen los factores δ_t y $\overline{\delta}$

4.3.1.1. Utilizando Variables Dummy para la estimación

Una metodología mas sencilla para estimar el modelo (4.3) partiendo del modelo (4.1) consiste en generar variables dummy para los factores i que agrupan a distintas observaciones. En la explicación del modelo (4.1) señalamos que el propósito es eliminar la variable no observada A_i que tiene la característica de variar entre distintos individuos -i- pero es constante a través del tiempo para un mismo individuo o agrupación -i-. Si modificamos el modelo agregando dummies para cada individuo o grupo -i-, el efecto de la variable no observada A_i será absorbido por estas variables dado que será colinear con dichas variables. Por lo tanto, nuestra especificación será:

$$\log(w_{it}) = \beta_0 + \beta_1 Educ_{it} + \beta_2 Exper_{it} + \eta_2 A2_i + \cdots + \eta_N AN_i + \delta_2 D2_t + \cdots + \delta_T DT_t + U_{it} \tag{4.4}$$

donde Aj_i es una dummy igual a uno si $i = j$ y cero eoc[5].

Es importante recordar la interpretación de las variables dummy: te dicen la diferencia de una media ponderada de los individuos que pertenecen al grupo identificado con la variable dummy respecto al grupo de referencia (i.e. la variable dummy omitida). En el caso de la ecuación (4.2), se calcula directamente la media para cada i usando las diferentes observaciones de i en t. Y en el modelo de efectos fijos (ecuación (4.3)), se quita dicho efecto al tomar la diferencia de la variable dependiente respecto a este promedio. Por lo tanto, la especificación del modelo utilizando variables dummy (ecuación (4.4)) será equivalente al modelo de efectos fijos (ecuación (4.3)) y estimará los mismos valores para los coeficientes y sus respectivos errores estandar.

4.3.2. Base de datos de panel no balanceados

Las derivaciones en los modelos anteriores asumen una base de datos de panel balanceada. Esto quiere decir que para todo i, existen T observaciones. Sin embargo, es posible que este supuesto no se cumpla por diversas razones dependiendo de la estructura de la base de datos. Por ejemplo, si i representa individuos y t tiempo, es posible que algunos individuos no sea posible encontrarlos para volver a entrevistarlos, ya sea por defunción, migración, que no quieran volver a contestar la encuesta, etc. En el caso de i siendo familias y t hermanos, es posible que distintas familias tengan distinto número de hermanos.

El hecho de que cada grupo tenga diferente número de observaciones t no impide aplicar el modelo de efectos fijos o la especificación utilizando variables dummies para estimar los coeficientes de interés. Lo único que es importante señalar es que aquellos i que únicamente cuenten con una observación serán

[5] Solo se incluyen $N-1$ dummies de individuos para evitar colinearidad con la constante. Si no se incluye el efecto de tiempo, podrían incluirse las N dummies dejando fuera la constante.

eliminados del modelo ya que la variable dummy predeciría perfectamente su variable dependiente.

La preocupación más importante surge desde el punto de vista del sesgo que esta selección pueda generar. Dicha selección sólamente será razón de preocupación cuando exista correlación entre algún factor no observado que cambie a lo largo del tiempo y este relacionado con la variable dependiente. Si las observaciones que causan que el panel no sea balanceado tienen valores no aleatorios de estos factores no observados, entonces podría existir una preocupación de sesgo. Por ejemplo, supongamos que queremos estimar la influencia de la altura de una persona sobre sus ingresos. La salud es una variable que cambia a los largo del tiempo en los distintos individuos y que puede estar corrleacionada con la altura y ser un factor determinante de los ingresos. En este caso es posible que las personas que causen que la base no sea balanceada tengan peor salud. Por lo tanto, esto sería una indicación de un posible sesgo en la variable de altura. Además de la salud, otra variable no observada puede ser el componente genético. El componente genético también es una variable no observada que está relacionada con la altura y puede ser un determinante de ingreso. Sin embargo, el componente genético es específico de cada individuo y no cambia a lo largo del tiempo. A pesar que el componente genético de las personas que causan que la base no sea balanceada no sea aleatorio, esto no será preocupación de sesgo, ya que el modelo de efectos fijos absorbe todos los factores no observados que no cambian para cada i.

4.4. Errores Estándar

Para tener una estimación válida de los modelos anteriores utilizando MCO es importante que se cumplan los supuestos que establecimos en el modelo MCO. Un supuesto de particular importancia es el de *i.i.d.* Dicho supuesto nos permitía asumir que las matriz de varianza-covarianza de los errores únicamente tenía valores sobre la diagonal (ya que no existía covarianza entre dos errores distintos por ser independientes las observaciones).

En el caso del uso de datos de panel este supuesto es muy restrictivo. En particular, dos observaciones utilizadas en el modelo (4.3) pueden compartir una misma i. Por ejemplo, en las observaciones ($i = 1, t = 1$) e ($i = 1, t = 2$), asumir que los errores $U_{1,1}$ y $U_{1,2}$ sean independientes es un supuesto muy restrictivo y probablemente no válido.

En los casos de homocedasticidad y heterocedasticidad asumíamos que nuestra matriz de varianza-covarianza de los errores era una diagonal dado que los errores eran independientes entre si. Sin embargo, ahora asumiremos que los errores pueden estar correlacionados dentro de un *cluster* o grupo, que estará definido por i. Esto quiere decir que nuestra matriz de varianza-covarianza de los errores tendrá algunos elementos fuera de la diagonal distintos a cero.

Existen distintas maneras de corregir este problema de los errores estándar. Uno consiste en calcular los *errores estandar clustered*. Este método es sencillo de llevar a cabo en Stata. Unicamente se tiene que especificar al final de la regresión "`, cl(var_cl)`" donde `var_cl` es la variable que indica los *clusters* (o grupos).

Esta opción calcula la varianza de la siguiente forma:

$$\widehat{Var(\beta)} = \hat{\alpha}\hat{\Lambda}\hat{\alpha}$$

donde:

$$\begin{aligned}\hat{\alpha} =& (\frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T} X_{it}X_{it}')^{-1} \\ \hat{\Lambda} =& \frac{1}{NT}\sum_{i=1}^{N} w_{it}w_{it}' \\ w_{it} =& \sum_{t=1}^{T} X_{it}\hat{U}_{it}\end{aligned}$$

En estas especificaciones, i indica los grupos o *clusters*, N es el número total de *clusters*, T es el número de observaciones por *cluster* y X_{it} es un vector $Kx1$ que incluye las K variables de control para cada par (i,t).

Ver la Nota "Ecuaciones clase 27-SEPT" (disponible en *Comunidad ITAM*) para detalles de la derivación de este tipo de errores y su comparación con los errores homocedásticos y heterocedásticos.

4.5. Modelo de Efectos Aleatorios

En la sección 3.11.2, discutimos que existe una especificación de errores estándar más eficiente que la especificación de errores heterocedasticos: mínimos cuadrados ponderados. Dicha especificación consistía en ponderar las variables dependiente y regresores y estimar una regresión utilizando las variables ponderadas bajo el supuesto de homocedasticidad. El modelo de efectos aleatorios se basa en la misma lógica.

Este modelo parte de la misma base que el modelo de efectos fijos (ecuación (4.1)):

$$\log(w_{it}) = \beta_0 + \beta_1 Educ_{it} + \beta_2 Exper_{it} + \delta_2 D2_t + \gamma A_i + \cdots + U_{it} \qquad (4.5)$$

Sin embargo, difiere del modelo de efectos fijos en el sentido de que el factor A_i no causa sesgo por variables omitidas ya que se basa en el supuesto de que:

$$Cov(X_{it}, A_i) = 0 \text{ para } t = 1, \ldots, T$$

En este caso, estimar la ecuación (4.1) utilizando MCO no generaría un estimador insesgado. El único requisito en términos de errores estándar es incluir errores tipo cluster, como discutimos en la sección anterior. La ventaja del modelo de efectos aleatorios consiste en que generará un estimador insesgado más eficiente que MCO donde además se corrige la estimación de los errores estandar por la posible covarianza dentro de un cluster.

El **modelo de efectos aleatorios** parte de la base de conjuntar el factor A_i con U_{it} en el término de error de la regresión:

$$\log(w_{it}) = \beta_0 + \beta_1 Educ_{it} + \beta_2 Exper_{it} + \delta_2 D2_t + \cdots + V_{it} \tag{4.6}$$

donde $V_{it} = \gamma A_i + U_{it}$.

Este modelo no podrá ser estimado mediante MCO ya que no cumple con la condición de *i.i.d.* En particular, la estructura de la matriz de varianza-covarianza de los errores no será una matriz con elementos distintos a cero únicamente en la diagonal (como se asume en homocedasticidad y heterocedasticidad). En lugar de esto, el modelo de efectos aleatorios consiste en asumir la siguiente estructura para la matriz de varianza-covarianza de los errores:

$$E(V_{it}V_{js}) = \begin{cases} \sigma_a^2 + \sigma_u^2 & \text{si } i = j \text{ y } t = s \\ \sigma_a^2 & \text{si } i = j \text{ y } t \neq s \\ 0 & \text{si } i \neq j \end{cases}$$

donde $\sigma_a^2 = Var(A_i)$ y $\sigma_u^2 = Var(U_{it})$. Implicitamente se está asumiendo que $Cov(A_i, U_{it}) = 0$. (Ilustrar la matriz de varianza-covarianza y compararla con el caso de errores homocedasticos y heterocedasticos).

Utilizando estos parametros (σ_a^2 y σ_u^2) se lleva a cabo la siguiente transformación al modelo base (ecuación (4.6)):

$$\begin{aligned} \log(w_{it}) - \lambda\overline{\log(w_i)} =& \beta_0(1-\lambda) + \beta_1(Educ_{it} - \lambda\overline{Educ}_i) + \\ & \beta_2(Exper_{it} - \lambda\overline{Exper}_i) + \delta_t - \lambda\hat{\delta} + (V_{it} - \lambda\overline{V}_i) \end{aligned} \tag{4.7}$$

donde: $\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}\right)^{\frac{1}{2}}$. Esto corresponde a restar una proporción de la media. Cabe notar que el modelo de efectos fijos resulta si $\lambda = 1$, mientras que un simple MCO con una base de datos transversal agrupada (*pooled OLS*) resulta si $\lambda = 0$.

Para estimar λ necesitamos estimadores para σ_a^2 y σ_u^2. Para llevar a cabo esto y estimar el modelo de efectos fijos podemos seguir el siguiente procedimiento:

1. Estimar el modelo base (ecuación (4.6)), utilizando MCO con la base de datos transversal agrupada (*pooled OLS*)

2. Utilizar los coeficientes para calcular los residuales ($\hat{V_{it}}$)
3. Estimar σ_a^2 como: $\hat{\sigma}_a^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{V}_{it}\hat{V}_{is}}{\frac{NT(T-1)}{2}}$
4. Estimar σ_u^2 como: $\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_a^2$ donde $\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{V}_{it}^2}{NT}$
5. Utilizar $\hat{\sigma}_u^2$ y $\hat{\sigma}_a^2$ para estimar $\hat{\lambda}$
6. Utilizar MCO para estimar la ecuación (4.7) utilizando el valor estimado $\hat{\lambda}$

Esto puede llevarse a cabo utilizando Stata y el comando `xtreg`. Antes de utilizar el comando `xtreg` es necesario indicarle a Stata que variables dan la estructura de base de datos de panel a la base que se utiliza (es decir, que es i y que es t). Para ello es necesario utilizar el comando `xtset variable_i variable_t`. Una vez que se llevo a cabo esto se puede indicar `xtreg y x1 x2 ... xK, re` y Stata estimará el modelo de efectos aleatorios. La mayor parte de los paquetes estadísticos (incluyendo Stata) generará como resultado además de los coeficientes de la regresión estimadores de los parametros $\hat{\lambda}$, $\hat{\sigma}_u^2$ y $\hat{\sigma}_a^2$.

Capítulo 5

Estimadores de Máxima Verosimilitud

En esta nota nos enfocamos en los casos en los que nuestra variable dependiente tiene características especiales. Iniciaremos analizando el caso en el cual la variable dependiente es una variable categórica que refleja una decisión. El primer modelo de este tipo que analizaremos será aquel en el cual la variable dependiente se puede caracterizar como una decisión binaria tales como empleado/desempleado, ir o no a la escuela, casado/soltero, o cualquier otra relación binaria. En este caso la variable dependiente podrá representarse con una variable *dummy*. Los modelos utilizados en este caso son el modelo *Probit* y *Logit*. Después de eso, veremos cómo los modelos de máxima verosimilitud son útiles también para analizar variables de decisión discretas que tengan un orden lógico de mayor a menor. Ejemplos de estas variables son: nivel máximo de estudios (básico/medio superior/superior), qué tan de acuerdo estás con cierta afirmación (nada/poco/algo/mucho), etc. Los modelos que pueden ser utilizados en este caso son: *Probit ordenado* y *Logit ordenado*. Por último, consideraremos casos los cuales la decisión es categórica pero no existe un órden de menor a mayor entre las distintas decisiones. Ejemplos pueden ser elecciones de partidos políticos (PAN/PRI/PRD/otro), selección de medio de transportación (coche/autobus/bicicleta/metro/otro), etc. En estos casos utilizaremos módelos como el *Logit multinomial*.

Además de aquellos casos en los cuales la variable dependiente es una variable categórica, los modelos de máxima verosimilitud son útiles también en casos en los cuales la variable dependiente tiene una distribución particular. Por ejemplo, podremos considerar casos en los cuales existe una alta concentración en uno de los extremos de la distribución. Por ejemplo, pudiera darse el caso que para la pregunta de horas trabajadas a la semana, todos los desempleados contestarán *cero*, lo cual podría generar una concentración importante en ese valor y poste-

riormente una distribución en valores positivos. Asimismo, podrían tener casos en las cuales la base de datos por construcción generá dicha concentración en un valor extremo. Por ejemplo, en algunos cuestionarios se reporta ingreso hasta cierto nivel (e.g. por cuestión de confidencialidad, en algunos casos cuando el ingreso es superior a 99,999 al mes se reporta como valor en la base de datos 99,999 en vez del valor verdadero del ingreso de dicha persona. En los casos anteriores utilizaremos los modelos *Tobit* y de *regresión censuarada*, respectivamente. Por último, como vimos en la sección 2, tener una muestra sesgada suele ser una preocupación. Esto provoca que la distribución de la muestra no sea necesariamente igual a la distribución poblacional. Si logramos obtener algo de información acerca de observaciones faltantes existen modelos de máxima verosimilitud que permiten hacer una corrección en la estimación para poder obtener estimadores de los parámetros poblacionales. El modelo *Heckit* es un modelo de este tipo.

Iniciaremos esta nota por mostrar cómo en un caso conocido, como es el de generar una estimación lineal, podemos emplear un modelo de máxima verosimilitud bajo ciertos supuestos para derivar estimadores de los parámetros poblacionales de nuestro interés. El único propósito de este primer ejercicio es describir el procedimiento y entender la intuición de los pasos que seguimos en los estimadores de máxima verosimilitud.

5.1. Motivación de Estimadores de Máxima Verosimilitud: Estimación Lineal

Los modelos de máxima verosimilitud consisten en encontrar los parámetros que maximizan una *función de máxima verosimilitud*.

Para entender como funcionan, veamos como se puede utilizar para generar estimadores de los parámetros (β, σ^2) en el caso de la estimación lineal:

$$Y_i = X_i'\beta + U_i$$

Si asumimos, como en el caso de teoría aisntótica que los errores se distribuyen de manera normal con media cero y varianza igual a σ^2 (como en el caso de homocedasticidad), podemos aprovechar la función de densidad normal para estimar los parámetros. Dado que asumimos que U_i se distribuye normal, los errores estarían descritos por la siguiente función de densidad:

$$f(U_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \cdot e^{-\frac{1}{2}\left(\frac{U_i}{\sigma}\right)^2}$$

Si aplicamos la transformación logarítmica y sustituimos U_i obtenemos nuestra función de máxima verosimilitud:

$$\mathcal{L}(\beta, \sigma^2) = \sum_{i=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2$$

En este caso, nuestros mejores estimadores de β y σ^2 serán los que maximizan esta función de máxima verosimilitud (Explicar intuitivamente que estamos haciendo). Si derivamos encontramos nuestras condiciones de primer orden que resultan en:

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{N} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{N} X_i Y_i\right) \\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\hat{\beta})^2\end{aligned}$$

Nótese que el estimador de β es exactamente el mismo que obtuvimos con MCO y el estimador de σ^2 es muy similar al que estimamos en el caso de homocedasticidad.

En este caso, en clase vimos cómo se derivan los estimadores de manera analítica. Sin embargo, en los casos de los estimadores que veremos a continuación, llevar a cabo dicha derivación analítica puede ser muy complicado (o imposible). En estos casos, para poder obtener valoes estimados de nuestros estimadores, se utilizan métodos numéricos, como por ejemplo el *Newton-Raphson*.

5.2. Variable Dependiente Categórica

5.2.1. Variable Dependiente: Dummy

5.2.1.1. Probit

El modelo Probit consiste en utilizar una función de densidad acumulada para aproximar la función $P(Y_i = 1|X_i)$. La ventaja de este modelo respecto al modelo de probabilidad lineal es que nunca habrá una predicción fuera del rango [0, 1]. Como mencionamos anteriormente, este modelo pertenece al grupo de estimadores de máxima verosimilitud, por lo tanto, necesitamos definir una función de máxima verosimilitud para poder derivar los parámetros que queremos estimar. Supongamos que partimos de la aproximación lineal del modelo de probabilidad lineal:

$$Y_i = X_i'\beta + U_i$$

Lo que nos interesa de este modelo son los parámetros β ya que nos dan información acerca del cambio en la probabilidad de que nuestra variable dependiente sea igual a uno por un cambio marginal en alguno de las variables incluidas en X_i.

El modelo Probit inicia por sustituir Y_i por una variable "latente" que llamaremos Y_i^*, donde $Y_i = 1\{Y_i^* > 0\}$. Esta función es una función indicador que determina que Y_i es igual a uno (cero) si Y_i^* es positivo (negativo). Nuestro modelo lineal será:

$$Y_i^* = X_i'\beta + U_i$$

El modelo Probit consiste en asumir que los errores se distribuyen de manera normal estándar. Por lo tanto, la estimación se calcula de la siguiente forma:

$$Pr(Y_i = 1|X_i) = Pr(U_i > -X_i'\beta) = Pr(U_i < X_i'\beta) = \Phi(X_i'\beta)$$

En el contexto de estimadores de máxima verosimilitud, tenemos una distribución Bernoulli ($f(y) = p^y(1-p)^{1-y}$). Sustituyendo la ecuación anterior y tomando nuevamente la transformación logarítmica obtenemos nuestra función de máxima verosimilitud (ilustrar en clase que estamos haciendo):

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} Y_i \log\left(\Phi(X_i'\beta)\right) + (1 - Y_i)\log\left(1 - \Phi(X_i'\beta)\right)$$

Como ilustramos en clase, la condición de primer orden no permite encontrar una solución cerrada, por lo tanto, para resolver este problema se emplean métodos numéricos. En cuanto al error estándar, su derivación es algebráicamente compleja y requiere además de teoría asintótica avanzada. La fórmula de los errores estándar es la siguiente y es calculada automáticamente en paquetes estadísticos como Stata:

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{i=1}^{N} \frac{\phi(X_i'\hat{\beta})X_iX_i'}{\Phi(X_i'\hat{\beta})[1 - \Phi(X_i'\hat{\beta})]}\right)^{-1}$$

Cabe mencionar que los coeficientes que resultan no tienen una interpretación intuitiva por si solos, ya que únicamente indicarán el cambio en la variable latente por un cambio marginal en una de las variables independientes X_i. Lo único que podemos saber es que si un coeficiente β_k es positivo (negativo) entonces, la probabilidad de que Y_i sea igual a uno aumentará (disminuirá) tras un cambio marginal en X_k. Sin embargo, lo que nos interesa es el estimador del cambio en la probabilidad de que Y_i sea igual a uno. Para llevar a cabo esto, necesitaremos una transformación basada en nuestra función de densidad agregada:

$$\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_{ki}} = \frac{\partial \Phi(X_i'\beta)}{\partial X_{ki}} = \phi(X_i'\beta)\beta_k$$

Posteriormente, este efecto se promedia utilizando todas las observaciones y con esto se obtiene un valor estimado del efecto de un cambio marginal en X_k sobre la probabilidad de que Y_i sea igual a uno (medido en puntos porcentuales). Esto se conoce como el *efecto parcial promedio*. Dado que en este caso tenemos una función no lineal, esto corresponde a una aproximación del efecto de un cambio marginal de X_k y es similar a lo que hicimos en la sección 3.4. Asimismo, podemos seguir el procedimiento especificado en la sección 3 para hacer una estimación exacta. La estimación exacta es particularmente interesante en los casos en que se tiene una variable dummy como variable explicativa en el modelo. Para estimar el error estándar de este efecto en ambos casos se puede utilizar el *método delta*[1] o *bootstrap*.

5.2.1.2. Logit

La derivación del modelo Logit es idéntica al modelo Probit. La diferencia entre ambos modelos radica en el supuesto que se establece para la distribución de los errores U_i. En el modelo Probit asumimos que se distribuyen normal estándar. En el caso del modelo Logit asumiremos que $\Phi(\cdot)$ es una distribución logística[2]

$$Pr(Y_i = 1|X_i) = \Phi(X_i'\beta) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}$$

La diferencia entre ambos modelos es mínima. Los coeficientes estimados suelen ser distintos, pero a diferencia de probit, en el caso de Logit los coeficientes tienen una interpretación intuitiva por si mismos. Para ver esto, calculemos el ratio de la probabilidad condicional de que Y_i sea igual a uno sobre la probabilidad condicional de que sea igual a cero:

$$\frac{Pr(Y_i = 1|X_i)}{Pr(Y_i = 0|X_i)} = \frac{Pr(Y_i = 1|X_i)}{1 - Pr(Y_i = 1|X_i)} = \exp(X_i'\beta)$$

Por lo tanto, calculando el logaritmo:

$$\log\left(\frac{Pr(Y_i = 1|X_i)}{Pr(Y_i = 0|X_i)}\right) = X_i'\beta$$

[1] El método delta resulta de una generalización del teorema central del límite. Es útil para estimar la distribución de una función continua de un parámetro siempre y cuando el parámetro converja en distribución a una normal.

[2] Cabe señalar que el modelo *Logit* también suele ser conocido como el modelo *logísitco* por el hecho de que se asume la distribución logística de los errores

En este caso, la interpretación del coeficiente β_k será: (*caeteris paribus*) un cambio marginal de X_k implica un cambio de $100\beta_k$ % en el ratio de la probabilidad que Y_i sea igual a uno sobre la probabilidad que sea igual a cero.

En algunos casos, al reportar los resultados del *Logit* es común encontrar que en vez de ver reportados los coeficientes, se reporta una transformación de dichos coeficientes que suele ser referido como el *Odds Ratio*. Para entender en qué consiste dicha transformación veamos lo siguiente. Empecemos por simplificar nuestra notación. Definamos:

$$\frac{Pr(Y_i = 1|X_i)}{Pr(Y_i = 0|X_i)} = \left(\frac{p}{1-p}\middle| X_i\right)$$

Imaginemos ahora que queremos ver el efecto de aumentar en una unidad alguna de nuestras variables explicativas. *Sin pérdida de generalidad* digamos que aumentamos la variable X_{1i} de 19 a 20. Por lo tanto, utilizando la especificación *Logit* tendríamos:

$$\begin{aligned} \log\left(\frac{p}{1-p}\middle| X_{1i} = 20\right) &= \beta_0 + \beta_1(20) + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} \\ \log\left(\frac{p}{1-p}\middle| X_{1i} = 19\right) &= \beta_0 + \beta_1(19) + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} \end{aligned} \tag{5.1}$$

Ahora supongamos que nos interesa la diferencia entre estos dos valores y que después de calcular la diferencia tomamos la exponencial:

$$\exp\left[\log\left(\frac{p}{1-p}\middle| X_{1i} = 20\right) - \log\left(\frac{p}{1-p}\middle| X_{1i} = 19\right)\right] = \frac{\exp\left[\log\left(\frac{p}{1-p}\middle| X_{1i} = 20\right)\right]}{\exp\left[\log\left(\frac{p}{1-p}\middle| X_{1i} = 19\right)\right]} \tag{5.2}$$

La expresión que hemos derivado a la derecha la podemos establecer como un cambio porcentual de manera simple:

$$\frac{\exp\left[\log\left(\frac{p}{1-p}\middle|X_{1i}=20\right)\right]}{\exp\left[\log\left(\frac{p}{1-p}\middle|X_{1i}=19\right)\right]} = \frac{\left(\frac{p}{1-p}\middle|X_{1i}=20\right)}{\left(\frac{p}{1-p}\middle|X_{1i}=19\right)}$$
$$= \frac{\left(\frac{p}{1-p}\middle|X_{1i}=19\right)+\left(\frac{p}{1-p}\middle|X_{1i}=20\right)-\left(\frac{p}{1-p}\middle|X_{1i}=19\right)}{\left(\frac{p}{1-p}\middle|X_{1i}=19\right)}$$
$$= 1 + \frac{\Delta\,\%\left(\frac{p}{1-p}\right)}{\Delta X_{1i}=1}$$

Ahora bien, si utilizamos la ecuación (5.1), sacamos la exponencial y simplificamos utilizando las propiedades de la exponencial:

$$\exp\left[\log\left(\frac{p}{1-p}\middle|X_{1i}=20\right)\right] = \exp\left[\beta_0+\beta_1(20)+\beta_2X_{2i}+\cdots+\beta_KX_{Ki}\right]]$$
$$= \exp\left[\beta_0\right]\cdot\exp\left[\beta_1\right]^{20}\cdot\exp\left[\beta_2\right]^{X_{2i}}\cdots\exp\left[\beta_K\right]^{X_{Ki}}$$
$$\exp\left[\log\left(\frac{p}{1-p}\middle|X_{1i}=19\right)\right] = \exp\left[\beta_0+\beta_1(19)+\beta_2X_{2i}+\cdots+\beta_KX_{Ki}\right]]$$
$$= \exp\left[\beta_0\right]\cdot\exp\left[\beta_1\right]^{19}\cdot\exp\left[\beta_2\right]^{X_{2i}}\cdots\exp\left[\beta_K\right]^{X_{Ki}}$$

Combinando este resultado con la ecuación (5.2):

$$1 + \frac{\Delta\,\%\left(\frac{p}{1-p}\right)}{\Delta X_{1i}=1} = \exp\left[\beta_1\right] \tag{5.3}$$

Como vemos, con este cálculo obtenemos el cambio porcentual del ratio $\frac{p}{1-p}$ que resulta de cambiar en una unidad el valor de X_1. La derivación que llevamos a cabo la hicimos asumiendo que X_1 cambia de 19 a 20, pero el mismo resultado lo hubiésemos obtenido eligiendo un cambio de una unidad con cualquier valor. Nuestro resultado es la exponencial del coeficiente original de *Logit*. Es interesante notar que en este caso nuestro resultado no depende de i, lo cual sugiere que con esta derivación el *efecto parcial promedio* y el *efecto parcial para la persona promedio* serán equivalentes nuevamente.

Sin embargo, lo mas importante que debemos señalar es que nuestro resultado reportado ($\exp\left[\beta_1\right]$) es igual a *uno mas el cambio porcentual*. Es por esto que en mucho contextos encontrarán que los resultados del *modelo Logit* (también

conicido como la *regresión logística*) suelen compararse con el 1. Es decir, ustedes querrían plantear la hipótesis nula de si su resultado es significativamente distinto a 1. En términos de interpretación, asi como en el caso del logaritmo, nuestro resultado indica un cambio porcentual del ratio $\frac{p}{1-p}$.

Es importante señalar que a pesar de lo indicado anteriormente, si queremos hacer comparable las interpretaciones del modelo Probit y el modelo Logit, de la misma manera que en la sección anterior, podemos calcular el *efecto parcial promedio* tomando el valor promedio de $\phi(X_i'\beta)\beta_k$. La diferencia es que en el caso de Logit $\phi(\cdot)$ representa la función de densidad logística. Igual que en el caso anterior, este efecto representará el cambio promedio de la probabilidad de que Y_i sea igual a uno (en puntos porcentuales) por un cambio marginal en X_k (*caeteris paribus*).

En el caso de ambos modelos, suele también reportarse el porcentaje de predicciones de Y_i correctas. Cabe recordar que en ambos casos (Probit y Logit) podemos generar un valor estimado de la variable latente (Y_i^*) para cada individuo (asumiendo errores $U_i = 0$). Dado el valor de dicha variable latente, podremos predecir basados en las variable independientes (X_i) si dicho individuo tiene una Y_i igual a uno o cero. Una manera de describir qué tan exacto es el modelo consiste en reportar el porcentaje de predicciones correctas. En algunos casos, dicho estadístico es criticado ya que un modelo puede ser muy bueno (malo) para predecir cuando Y_i es igual a cero (uno) o viceversa. Si en este caso la mayor parte de los Y_i observados es igual a cero, el estadístico puede reportar que el modelo es muy bueno para predecir a pesar de que casi nunca prediga correctamente cuando Y_i es igual a uno. Alternativas para este estadístico consisten en reportar por separado el porcentaje de predicciones correctas para cada caso (i.e. cuando Y_i es igual a cero y cuando es igual a uno).

Es importante señalar que en todos estos modelos aplican las preocupaciones de validez interna que se tienen en el modelo de MCO. En particular, el sesgo por variables omitidas sigue siendo un limitante importante para interpretar los efectos parciales promedio de forma causal.

Un ejemplo de estos modelos se pueden obtener utilizando los siguientes comandos de Stata:

- `webuse nhanes2d`
- `reg highbp height weight age female, r`
- `probit highbp height weight age female, r`
- `predict probit_hbp`
- `mfx`
- `logit highbp height weight age female, r`
- `predict logit_hbp`
- `mfx`

- `logit highbp height weight age female, r or`

5.2.2. Variable Dependiente: Multivariada Ordenada

5.2.2.1. Probit Ordenado

El modelo de Probit ordenado asume que la variable dependiente es una variable categorica y que sus valores tienen un orden lógico de menor a mayor. En clase veremos algunos ejemplos.

Este modelo es una extensión del modelo Probit. Nuevamente, esta basado en una variable latente:

$$Y_i^* = X_i'\beta + U_i \tag{5.4}$$

Sin embargo, en este caso la variable latente estará definida para $j = \{0, ..., J\}$ (donde J es el número de valores que puede tomar la variable dependiente ordenada) como:

$$Y_i = j \text{ si } \alpha_j \leq Y_i^* < \alpha_{j+1} \text{ donde } \alpha_0 = -\infty,\ \alpha_{J+1} = \infty \text{ y } \alpha_j < \alpha_{j+1} \tag{5.5}$$

Siguiendo los mismos pasos que en el caso del modelo Probit asumimos que el error se distribuye normal estándar. Por lo tanto, la probabilidad condicional de que $Y_i = 0$ será:

$$Pr(Y_i = 0|X_i) = Pr(X_i'\beta + U_i < \alpha_1) = Pr(U_i < \alpha_1 - X_i'\beta) = \Phi(\alpha_1 - X_i'\beta)$$

Y la probabilidad condicional de que $Y_i = j$ para $j = 1, ..., J-1$ será:

$$\begin{aligned} Pr(Y_i = j|X_i) =& Pr(\alpha_j < X_i'\beta + U_i < \alpha_{j+1}) = Pr(\alpha_j - X_i'\beta < U_i < \alpha_{j+1} - X_i'\beta) \\ &= \Phi(\alpha_{j+1} - X_i'\beta) - \Phi(\alpha_j - X_i'\beta) \end{aligned}$$

Finalmente, la probabilidad condicional de que $Y_i = J$ será:

$$Pr(Y_i = J|X_i) = Pr(\alpha_J < X_i'\beta + U_i) = Pr(\alpha_J - X_i'\beta < U_i) = 1 - \Phi(\alpha_J - X_i'\beta)$$

En base a esto podemos generar la función de máxima verosimilitud que utilizaremos para encontrar los valores de β y $\alpha_1, ..., \alpha_J$:

$$\mathcal{L}(\beta, \alpha_1, ..., \alpha_J) = \sum_{i=1}^{N} \Bigg[1\{Y_i = 0\} \cdot \log\left(\Phi(\alpha_1 - X_i'\beta)\right) + \sum_{j=1}^{J-1} 1\{Y_i = j\} \cdot \log\left(\Phi(\alpha_{j+1} - X_i'\beta) - \Phi(\alpha_j - X_i'\beta)\right) + 1\{Y_i = J\} \cdot \log\left(1 - \Phi(\alpha_J - X_i'\beta)\right) \Bigg]$$

Nuevamente, en este caso los coeficientes no tendrán ninguna interpretación intuitiva por si mismos. Lo interesante en este caso será predecir el cambio en la probabilidad de la ocurrencia de distintos valores de la variable dependiente por un cambio marginal en alguna de las variables independientes (X_i). Por ejemplo:

$$\frac{\partial Pr(Y_i = j|X_i)}{\partial X_{ki}} = \phi(\alpha_j - X_i'\beta)\beta_k - \phi(\alpha_{j+1} - X_i'\beta)\beta_k$$

Igual que en el caso del modelo Probit, se calcula el promedio para todas las observaciones y esto resultará en el estimador del efecto parcial promedio. En este caso lo único que sabemos a partir del signo de β_k es que si es positivo (negativo), la probabilidad de que $Y_i = J$ aumentará (disminuirá) y la probabilidad de que $Y_i = 0$ disminuirá (aumentará) por un aumento marginal en X_k.

En clase veremos un ejemplo de estos modelos utilizando los siguientes comandos de Stata:

- `webuse nhanes2f`
- `oprobit health female black age, r`
- `predict pr1 pr2 pr3 pr4 pr5`
- `mfx`
- `mfx compute, predict (outcome(#3))`

5.2.2.2. Logit Ordenado

La relación entre el logit ordenado y Probit ordenado es muy similar a la relación entre el Logit y el Probit cuando utilizamos una variable binaria como variable dependiente.

Al igual que en Probit ordenado, las ecuaciones (5.4) y (5.5) describen el planteamiento de este tipo de modelos. Al igual que en *Logit* asumiremos que los

errores tienen una distribución logísitca. Para poder ver la diferencia entre los resultados anteriores y los que tenemos en Logit ordenado vale la pena desarrollar el cáculo de algunas probabilidades:

$$\begin{aligned}
Pr(Y_i = 0|X_i) &= Pr(Y_i^* \leq \alpha_1) \\
&= Pr(X_i'\beta + U_i \leq \alpha_1) \\
&= Pr(U_i \leq \alpha_1 - X_i'\beta) \\
&= 1 - Pr(U_i > \alpha_1 - X_i'\beta) \\
&= 1 - Pr(U_i < X_i'\beta - \alpha_1) \\
&= 1 - \Phi(X_i'\beta - \alpha_1) \\
&= 1 - \frac{\exp(X_i'\beta - \alpha_1)}{1 + \exp(X_i'\beta - \alpha_1)} \\
&= \frac{1}{1 + \exp(X_i'\beta - \alpha_1)}
\end{aligned}$$

Siguiendo la misma lógica podemos derivar:

$$\begin{aligned}
Pr(Y_i = 1|X_i) &= Pr(\alpha_1 < Y_i^* \leq \alpha_2) \\
&= Pr(Y_i^* \leq \alpha_2) - Pr(Y_i^* \leq \alpha_1) \\
&= \left(1 - \frac{\exp(X_i'\beta - \alpha_2)}{1 + \exp(X_i'\beta - \alpha_2)}\right) - \left(1 - \frac{\exp(X_i'\beta - \alpha_1)}{1 + \exp(X_i'\beta - \alpha_1)}\right) \\
&= \frac{\exp(X_i'\beta - \alpha_1)}{1 + \exp(X_i'\beta - \alpha_1)} - \frac{\exp(X_i'\beta - \alpha_2)}{1 + \exp(X_i'\beta - \alpha_2)}
\end{aligned}$$

A partir de este resultado podemos derivar efectos parcial promedio y efectos parciales para la persona promedio por cambios marginales en cualquier variable explicativa X_j. Unicamente es importante mantener en cuenta cómo llevar a cabo una derivada parcial cuando tenemos una exponencial. Para poder simplificar este cálculo tomemos en cuenta lo siguiente:

$$\begin{aligned}
\Phi(X_i'\beta - \alpha_j) &= \frac{\exp(X_i'\beta - \alpha_j)}{1 + \exp(X_i'\beta - \alpha_j)} \\
\frac{\partial \Phi(X_i'\beta - \alpha_j)}{\partial X_{1i}} &= \frac{(1 + \exp(X_i'\beta - \alpha_j))\exp(X_i'\beta - \alpha_j)\beta_1 - \exp(X_i'\beta - \alpha_j)\exp(X_i'\beta - \alpha_j)\beta_1}{(1 + \exp(X_i'\beta - \alpha_j))^2} \\
&= \left[\frac{\exp(X_i'\beta - \alpha_j)}{1 + \exp(X_i'\beta - \alpha_j)} - \left(\frac{\exp(X_i'\beta - \alpha_j)}{1 + \exp(X_i'\beta - \alpha_j)}\right)^2\right]\beta_1 \\
&= \Phi(X_i'\beta - \alpha_j)(1 - \Phi(X_i'\beta - \alpha_j))\beta_1
\end{aligned}$$

Por lo tanto, podemos calcular el cambio en la probabilidad de que Y_i tenga un valor en especifico por un cambio marginal en alguna de las variables explicativas. Para ver esto, partimos de la derivación de $Pr(Y_i = 1|X_i)$ y utilizamos el resultado que acabamos de derivar:

$$\begin{aligned}
Pr(Y_i = 1|X_i) &= \frac{\exp(X_i'\beta - \alpha_1)}{1 + \exp(X_i'\beta - \alpha_1)} - \frac{\exp(X_i'\beta - \alpha_2)}{1 + \exp(X_i'\beta - \alpha_2)} \\
&= \Phi(X_i'\beta - \alpha_1) - \Phi(X_i'\beta - \alpha_2) \\
\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_{1i}} &= \frac{\partial \Phi(X_i'\beta - \alpha_1)}{\partial X_{1i}} - \frac{\partial \Phi(X_i'\beta - \alpha_2)}{\partial X_{1i}} \\
&= \beta_1 \Big[\Big(\Phi(X_i'\beta - \alpha_1)(1 - \Phi(X_i'\beta - \alpha_1)) \Big) - \Big(\Phi(X_i'\beta - \alpha_2)(1 - \Phi(X_i'\beta - \alpha_2)) \Big) \Big]
\end{aligned}$$

Por último, una característica atractiva de *Logit* es que los coeficientes tenían una interpretación específica sin la necesidad de llevar a cabo ningun cálculo o transformación. Esto se mantiene en *Logit ordenado*. Para ver esto notemos que:

$$\begin{aligned}
Pr(Y_i > j|X_i) &= Pr(Y_i^* > \alpha_{j+1}) \\
&= Pr(U_i > \alpha_{j+1} - X_i'\beta) \\
&= Pr(U_i < X_i'\beta - \alpha_{j+1}) \\
&= \Phi(X_i'\beta - \alpha_{j+1}) \\
&= \frac{\exp(X_i'\beta - \alpha_{j+1})}{1 + \exp(X_i'\beta - \alpha_{j+1})} \\
Pr(Y_i \leq j|X_i) &= 1 - Pr(Y_i > j|X_i) \\
&= 1 - \frac{\exp(X_i'\beta - \alpha_{j+1})}{1 + \exp(X_i'\beta - \alpha_{j+1})} \\
&= \frac{1}{1 + \exp(X_i'\beta - \alpha_{j+1})}
\end{aligned}$$

Por lo tanto, tenemos al igual que antes un ratio de dos probabilidades. Lo importante es que una sea complementaria de la otra. De esta forma obtenemos:

$$\begin{aligned}
\frac{Pr(Y_i > j|X_i)}{Pr(Y_i \leq j|X_i)} &= \exp(X_i'\beta - \alpha_{j+1}) \\
\log\left(\frac{Pr(Y_i > j|X_i)}{Pr(Y_i \leq j|X_i)} \right) &= X_i'\beta - \alpha_{j+1} \\
\frac{\partial \log\left(\frac{Pr(Y_i > j|X_i)}{Pr(Y_i \leq j|X_i)} \right)}{\partial X_{1i}} &= \beta_1
\end{aligned}$$

Cabe notar que este resultado que derivamos no depende de j, lo cual indica que esto se puede generalizar para el ratio de dos probabilidades para $j = \{0, 1, ..., J-1\}$.

5.2.3. Variable Dependiente: Multivariada no Ordenada

Este modelo aplica cuando la variable dependiente consiste en la selección de alternativas discretas, pero que no tienen un órden lógico de menor a mayor entre si. Ejemplos incluyen selección de forma de transporte (metro/camión/coche), selección de algún producto de diferentes características (TV bulbo, plasma, LED, no TV) tipo de gasolina (magna, premium), etc

5.2.3.1. Logit Multinomial y Logit Condicional

$Y \in \{0, 1, ..., J\}$ sin orden entre las variables. Cada valor representa una alternativa que se puede elegir.

El objetivo es generar predicciones adecuadas de dichas alternativas utilizando información de variables acerca de las alternativas o de los individuos llevando a cabo la selección. Las variables pueden depender del individuo llevando a cabo la decisión X_i o de la opción a ser elegida (posiblemente junto cada individuo) $X_{i,j}$.

McFadden desarrolló estos modelos basados en maximización de utilidad. Desarrolla un modelo para la probabilidad condicional de elegir j dados los valores de las variables explicativas:

$$Pr(Y_i = j|X) = P_j(X_i\beta)$$

Utilizando las probabilidades, deriva la siguiente función de máxima verosimilitud:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=0}^{J} \mathbf{1}\{Y_i = j\} \log P_j\left(X_{i,j}, \beta\right)$$

El *multinomial Logit* es un modelo sencillo que asume que solo tenemos variables explicativas a nivel individual. Con ello extendemos el modelo Logit y

obtenemos:

$$Pr(Y_i = j|X_i) = \frac{\exp(X_i'\beta_j)}{1 + \sum_{l=0}^{J} \exp(X_i'\beta_l)}$$
$$Pr(Y_i = 0|X_i) = \frac{1}{1 + \sum_{l=0}^{J} \exp(X_i'\beta_l)}$$
$$\implies \log\left(\frac{Pr(Y_i = j|X_i)}{Pr(Y_i = 0|X_i)}\right) = X_i'\beta_j$$

con lo cual podemos interpretar las β's.

Este modelo puede ser visto como un caso particular del modelo **Logit condicional** el cual tiene variables explicativas que varían según la alternativa a elegir.

$$Pr(Y_i = j|X_{i,0}, X_{i,1}, ..., X_{i,J}) = \frac{\exp(X_{i,j}'\beta_j)}{\sum_{l=0}^{J} \exp(X_{i,l}'\beta_l)}$$

Vinculo con maximización de utilidad:

$$U_{i,j} = X_{i,j}'\beta + \epsilon_{i,j}$$

Asumimos que un individuo elige j si le da la mayor nivel de utilidad:

$$Y_i = j \quad \text{si} \quad U_{i,j} \geq U_{i,l} \quad \forall \quad l \neq j$$

Asumimos independencia de $\epsilon_{i,j}$ para diferentes elecciones y distribución de Tipo-I Extreme Value. Esta distribución se caracteriza por:

$$F(\epsilon) = \exp(-\exp(-\epsilon))$$
$$f(\epsilon) = \exp(-\epsilon)\exp(-\exp(-\epsilon))$$

Dados estos supuestos:

$$
\begin{aligned}
&Pr(Y_i = 0|X_i) = Pr(U_{i,0} > U_{i,1}, U_{i,0} > U_{i,2}, ..., U_{i,0} > U_{i,J}) \\
&= Pr(\epsilon_{i,0} + X'_{i,0}\beta - X_{i,1}\beta > \epsilon_{i,1}, ..., \epsilon_{i,0} + X'_{i,0}\beta - X'_{i,J}\beta > \epsilon_{i,J}) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\epsilon_{i,0}+X'_{i,0}\beta-X'_{i,1}\beta} \cdots \int_{-\infty}^{\epsilon_{i,0}+X'_{i,0}\beta-X'_{i,J}\beta} f(\epsilon_{i,0}) f(\epsilon_{i,1}) \cdots f(\epsilon_{i,J}) d\epsilon_{i,0} d\epsilon_{i,1} \cdots d\epsilon_{i,J} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0})) \exp(-\exp(-\epsilon_{i,0} - X'_{i,0}\beta + X'_{i,1}\beta)) \cdots \\
&\qquad \cdots \exp(-\exp(-\epsilon_{i,0} - X'_{i,0}\beta + X'_{i,J}\beta)) d\epsilon_{i,0} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0}) - \exp(-\epsilon_{i,0} - X'_{i,0}\beta + X'_{i,1}\beta) \cdots \\
&\qquad \cdots - \exp(-\epsilon_{i,0} - X'_{i,0}\beta + X'_{i,J}\beta)) d\epsilon_{i,0} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0}) - \exp(-\epsilon_{i,0}) \exp(-X'_{i,0}\beta + X'_{i,1}\beta) \cdots \\
&\qquad \cdots - \exp(-\epsilon_{i,0}) \exp(-X'_{i,0}\beta + X'_{i,J}\beta)) d\epsilon_{i,0} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0})(1 + \exp(X'_{i,1}\beta - X'_{i,0}\beta) \cdots \\
&\qquad \cdots + \exp(X'_{i,J}\beta - X'_{i,0}\beta))) d\epsilon_{i,0}
\end{aligned}
$$

Sea $c_i = -\log(1 + \exp(X'_{i,1}\beta - X'_{i,0}\beta) + \cdots + \exp(X'_{i,J}\beta - X'_{i,0}\beta))$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0}) \exp(-c_i)) d\epsilon_{i,0} \\
&= \int_{-\infty}^{\infty} \exp(-\epsilon_{i,0}) \exp(-\exp(-\epsilon_{i,0} - c_i)) d\epsilon_{i,0}
\end{aligned}
$$

Sea $\eta_i = \epsilon_{i,0} + c_i$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \exp(c_i - \eta_i) \exp(-\exp(-\eta_i)) d\eta_i \\
&= \int_{-\infty}^{\infty} \exp(c_i) \exp(-\eta_i) \exp(-\exp(-\eta_i)) d\eta_i \\
&= \exp(c_i) \int_{-\infty}^{\infty} \exp(-\eta_i) \exp(-\exp(-\eta_i)) d\eta_i \\
&= \exp(c_i)
\end{aligned}
$$

Finalmente vemos que:

$$
\begin{aligned}
\exp(-c_i) &= \frac{1}{\exp(c_i)} \\
\exp(-c_i) &= 1 + \exp(X'_{i,1}\beta - X'_{i,0}\beta) + \cdots + \exp(X'_{i,J}\beta - X'_{i,0}\beta) \\
&= \exp(X'_{i,0}\beta - X'_{i,0}\beta) + \exp(X'_{i,1}\beta - X'_{i,0}\beta) + \cdots + \exp(X'_{i,J}\beta - X'_{i,0}\beta) \\
&= \exp(-X'_{i,0}\beta) \left[\sum_{l=0}^{J} \exp(X'_{i,l}\beta) \right] \\
&= \frac{\sum_{l=0}^{J} \exp(X'_{i,l}\beta)}{\exp(X'_{i,0}\beta)} \\
\implies \exp(c_i) = Pr(Y_i = 0|X_i) &= \frac{\exp(X'_{i,0}\beta)}{\sum_{l=0}^{J} \exp(X'_{i,l}\beta)}
\end{aligned}
$$

Para la interpretación de coeficientes, veamos un ejemplo basado en McFadden (82).

El objetivo es analizar la elección de los hogares entre comprar secadora eléctrica, secadora de gas o no comprar secadora.

Para ello plantea:

$$
\begin{aligned}
U_{i,elec} &= \beta_{0,elec} + \beta_{1,elec} own_i + \beta_{2,elec} persons_i + \beta_{3,elec} gas_i + \cdots \\
&\qquad \cdots + \beta_{oper} elec * oper_i + \beta_{cap} elec * cap_i + \epsilon_{i,elec} \\
U_{i,gas} &= \beta_{0,gas} + \beta_{1,gas} own_i + \beta_{2,gas} persons_i + \beta_{3,gas} gas_i + \cdots \\
&\qquad \cdots + \beta_{oper} gas * oper_i + \beta_{cap} gas * cap_i + \epsilon_{i,gas} \\
U_{i,no} &= \beta_{0,no} + \beta_{1,no} own_i + \beta_{2,no} persons_i + \beta_{3,no} gas_i + \epsilon_{i,no}
\end{aligned}
$$

Por lo que se asume que los costos de operación y de capital de no tener secadora son cero.

$$Pr(elec) = \frac{\exp(U^*_{i,elec})}{\exp(U^*_{i,elec}) + \exp(U^*_{i,gas}) + \exp(U^*_{i,no})}$$

MNL equivaldría a restar de cada utilidad un valor constante a nivel individuo:

$$c_i = \beta_{0,no} + \beta_{1,no} own_i + \beta_{2,no} persons_i + \beta_{3,no} gas_i$$

Supongamos aquí que nos interesa una elasticidad:

$$\epsilon_{elec,elec\cdot oper} = \frac{\partial Pr(elec)}{\partial elec \cdot oper} \cdot \frac{elec \cdot oper}{Pr(elec)}$$

Entonces:

$$\frac{\partial Pr(elec)}{\partial elec \cdot oper} = \frac{(\exp(U^*_{i,elec}) + \exp(U^*_{i,gas}) + \exp(U^*_{i,no})) \cdot \exp(U^*_{i,elec})\beta_{oper} - ... - \exp(U^*_{i,elec}) \exp(U_{i,elec})\beta_{oper}}{\left(\exp(U^*_{i,elec}) + \exp(U^*_{i,gas}) + \exp(U^*_{i,no})\right)^2}$$

$$= \left(\frac{\exp(U^*_{i,elec})}{exp(U^*_{i,elec}) + \exp(U^*_{i,gas}) + \exp(U^*_{i,no})} \right) \beta_{oper} - \left(\frac{\exp(U^*_{i,elec})}{\exp(U^*_{i,elec}) + \exp(U^*_{i,gas}) + \exp(U^*_{i,no})} \right)^2 \beta_{oper}$$

5.3. Variable Dependiente: Alta Concentración en un Extremo de la Distribución

5.3.1. Tobit

En el caso del modelo Tobit nos interesa ver cómo son afectadas las estimaciones si la variable dependiente tiene una alta concentración en un valor específico y para el resto de los valores hay una distribución relativamente continua. En la literatura suele hacerse referencia a este modelo como el de respuestas de solución de esquina. Ejemplos de este tipo de variable dependiente incluyen casos en los cuales se pregunte por la cantidad de horas trabajadas al mes o la cantidad de bebidas alcohólicas consumidas. En este caso utilizar MCO puede

llevarnos a tener predicciones ilógicas para la variable dependiente, igual que el caso del modelo de probabilidad lineal. Asimismo, la concentración de valores para la variable dependiente puede llevarnos a un sesgo por la forma funcional si queremos estimar el efecto de un cambio marginal en X_k para los valores de la variable dependiente donde existe una distribución continua (es decir, omitiendo a aquellos que no trabajan o consumen alcohol en los ejemplos).

Nuevamente, el modelo se especifica en términos de una variable latente (Y_i^*)[3]

$$Y_i^* = X_i'\beta + U_i$$

donde ahora $Y_i = \text{máx}\{0, Y_i^*\}$.

Igual que en los modelos anteriores, asumimos que los errores se distribuyen asintóticamente normal con media cero y varianza σ^2. Dado este supuesto de los errores, la probabilidad condicional de que $Y_i = 0$ será :

$$\begin{aligned} Pr(Y_i = 0|X_i) = Pr(Y_i^* < 0|X_i) = Pr(U_i < -X_i'\beta|X_i) = Pr\left(\frac{U_i}{\sigma} < \frac{-X_i'\beta}{\sigma}|X_i\right) \\ = \Phi\left(\frac{-X_i'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{X_i'\beta}{\sigma}\right) \end{aligned}$$

Y para el caso de $Y_i > 0$, la densidad de Y_i dado X_i será igual a:

$$f(U_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(\frac{-(Y_i - X_i'\beta)^2}{2\sigma^2}\right)$$

Por lo tanto, la función de máxima verosimilitud que utilizaremos para definir los valores de β y σ^2, similarmente que en los casos anteriores, resulta en:

$$\begin{aligned} \mathcal{L}(\beta, \sigma^2) = \sum_{i=1}^{N} \Bigg[& 1\{Y_i = 0\} \cdot \log\left(1 - \Phi(X_i'\beta/\sigma)\right) \\ & + 1\{Y_i > 0\} \cdot \left(-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2 \right) \Bigg] \end{aligned}$$

Como resultado obtendremos estimadores de los coeficientes β y sus errores estándar. Sin embargo, nuevamente, los coeficientes no tienen una interpretación intuitiva por si solos. En el caso del modelo Tobit existen diferentes componentes que resulta interesante analizar. En el caso de MCO, los coeficientes estimados

[3]Por simplicidad, en este caso asumimos que la distribución de la variable latente está concentrada en cero. La derivación del modelo si hay concetración en otro valor ya sea máximo o mínimo es análoga

nos daban información acerca del cambio en $E(Y_i|X_i)$ por un cambio marginal en X_k (caeteris paribus). Dicho efecto es de gran interés ya que busca establecer relaciones causales de una variable a otra. En el caso del modelo Tobit también es posible estimar dicho efecto, pero además existen otros componentes que pueden ser estimados utilizando la estructura de este modelo. Para entender esto podemos partir de especificar $E(Y_i|X_i)$ utilizando este modelo:

$$E(Y_i|X_i) = Pr(Y_i > 0|X_i)E(Y_i|Y_i > 0, X_i) = \Phi(X_i'\beta/\sigma) \cdot E(Y_i|Y_i > 0, X_i)$$

En este caso, para encontrar el término $E(Y_i|Y_i > 0, X_i)$ tomaremos en cuenta que U_i tienen una distribución normal con media cero y varianza σ^2, por lo tanto [Nota: para simplificar las siguientes ecuaciones utilizaremos el **ratio inverso de Mills** que se define como $\lambda(k) = \frac{\phi(k)}{\Phi(k)}$]:

$$\begin{aligned} E(Y_i|Y_i > 0, X_i) &= X_i'\beta + E(U_i|U_i > -X_i'\beta, X_i) \\ &= X_i'\beta + \sigma E(U_i/\sigma|U_i/\sigma > -X_i'\beta/\sigma, X_i) \\ &= X_i'\beta + \sigma\frac{\phi(-X_i'\beta/\sigma)}{1 - \Phi(-X_i'\beta/\sigma)} \\ &= X_i'\beta + \sigma\frac{\phi(X_i'\beta/\sigma)}{\Phi(X_i'\beta/\sigma)} \\ &= X_i'\beta + \sigma\lambda(X_i'\beta/\sigma) \end{aligned}$$

Una vez obtenido esto podemos calcular el efecto que nos interesa:

$$\frac{\partial E(Y_i|X_i)}{\partial X_k} = \frac{\partial Pr(Y_i > 0|X_i)}{\partial X_k} \cdot E(Y_i|Y_i > 0, X_i) + Pr(Y_i > 0|X_i) \cdot \frac{\partial E(Y_i|Y_i > 0, X_i)}{\partial X_k} \tag{5.6}$$

Para especificar esta ecuación necesitamos encontrar dos factores que por si solos pueden ser estadísticos o efectos de interés: (i) el cambio en la probabilidad condicional de que Y_i sea mayor a cero por un cambio marginal en X_k y (ii) el cambio en $E(Y_i|Y_i > 0, X_i)$ por un cambio marginal en X_k.

$$\begin{aligned} \frac{\partial Pr(Y_i > 0|X_i)}{\partial X_k} &= \frac{\partial \Phi(X_i'\beta/\sigma)}{\partial X_k} = \left(\frac{\beta_k}{\sigma}\right)\phi(X_i'\beta/\sigma) \\ \frac{\partial E(Y_i|Y_i > 0, X_i)}{\partial X_k} &= \beta_k\left[1 - \lambda(X_i'\beta/\sigma)\left(X_i'\beta/\sigma + \lambda(X_i'\beta/\sigma)\right)\right] \end{aligned} \tag{5.7}$$

Sustituyendo estas ecuaciones en la ecuación (5.6) obtenemos:

$$\frac{\partial E(Y_i|X_i)}{\partial X_k} = \Phi(X_i'\beta/\sigma)\beta_k \tag{5.8}$$

En conclusión, dependiendo de que efecto nos interese estimar, tomámos la media de nuestra muestra para alguna de las ecuaciones en (5.7) o (5.8).

5.3.2. Regresión Censuarada

Este tipo de modelos toma en cuenta que por diseño algunas bases de datos reportan un nivel máximo para algunos valores. Un ejemplo clásico es que las encuestas a hogares preguntan por el ingreso mensual de un individuo. Al reportar este nivel de ingreso en la base de datos por confidencialidad todos aquellos individuos por encima de un valor de ingreso determinado sustituyen el valor verdadero del ingreso del individuo por un ingreso tope que se determina en la encuesta (e.g. 99,999 pesos al mes). Como resultado, la persona que analiza los datos en algunos casos no observa el nivel verdadero de los ingresos para todos aquellos individuos con ingreso mayor al tope (únicamente saben que el ingreso es mayor o igual a dicho valor).

Para llevar a cabo la derivación de este tipo de modelos tomemos el caso donde se censuaran valores altos de la variable dependiente y se les da un mismo valor máximo (C_i):

$$Y_i = X_i'\beta + U_i \quad , \quad U_i|X_i, C_i \sim N(0, \sigma^2)$$

$$w_i = min\{Y_i, C_i\}$$

Ojo: U_i no solo es independiente de X_i (homocedasticidad) sino también de C_i. Usualmente C_i es una constante y no depende de i.

Para observaciones censuradas:

$$\begin{aligned} Pr(w_i = C_i|X_i) &= Pr(Y_i > C_i|X_i) \\ &= Pr(U_i > C_i - X_i'\beta) \\ &= Pr\left(\frac{U_i}{\sigma} < \frac{X_i'\beta - C_i}{\sigma}\right) \\ &= \Phi\left(\frac{X_i'\beta - C_i}{\sigma}\right) \end{aligned}$$

De forma similar a Tobit, las observaciones no censuradas: $w_i = Y_i$.

Usamos:

$$f(w_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}(w_i - X_i'\beta)^2\right)$$

La función de Máxima Verosimilitud es:

$$\mathcal{L}(\beta,\sigma) = \sum_{i=1}^{n}\left[\mathbf{1}\{w_i = C_i\}\log\left[\Phi\left(\frac{X_i'\beta - C_i}{\sigma}\right)\right] + ... \right.$$
$$\left. ... + \mathbf{1}\{w_i < C_i\}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(w_i - X_i'\beta)^2\right]\right]$$

Los coeficientes en este caso se interpretan como un OLS estándar, solo que la corrección de la censura suele ser relevante en casos en que la densidad en dicho punto es importante para hacer una corrección funcional que evite que β este sesgada.

5.4. Otros modelos

Otros modelos que están basados en funciones de máxima verosimilitud y que se derivan de manera similar a la que hemos expuesto en esta nota incluyen:

1. **Modelo de regresión truncado.** Este tipo de modelos se utiliza cuando se quiere llevar a cabo inferencia para toda una población, pero la encuesta por diseño solo incluye a un grupo restringido de la población. En este caso no se observa información para los individuos que no cumplen con la restricción establecida para la selección de la muestra y, por lo tanto, no son una muestra aleatoria de la población.
2. **Modelo de regresión poisson.** Aplica cuando la variable dependiente es una variable de *conteo* y tiene pocos valores (usualmente menor a cinco). Ejemplos incluyen número de hijos, número de trabajos, número de materias reprobadas, etc.

Capítulo 6

Kernel

En esta nota nos enfocaremos a estimar densidades de ciertas variables y una vez hecho, utilizaremos las herramientas desarrolladas para llevar a cabo estimaciones no paramétricas de medias condicionales. Daremos una intuición de cómo esto se asocia y cómo se diferencia de las estimaciones tradicionales de MCO.

6.1. Histogramas

Empezaremos por estimar densidades de variables continuas[1]. Nuestro objetivo es llevar a cabo la estimación de la densidad poblacional. Para poder llevar a cabo esto, empezaremos por estudiar en detalle la teoría detrás de los histogramas. El problema que buscamos resolver es: si consideramos las variables aleatorias $\{X_1, ..., X_n\}$ i.i.d con pdf $f_X(x)$, sea $\{x_1, ..., x_n\}$ la realización de estas variables aleatorias. Nos interesa estimar la función de densidad en un punto especifico, $f_X(x)$, donde x es una constante.

Sea $[a, b]$ el soporte de X. Para llevar a cabo un hisotgrama empezaremos por dividir este soporte en K intervalos del mismo tamaño: $\frac{b-a}{K}$. Los intervalos son, entonces:

$$\left[a + (k-1)\left(\frac{b-a}{K}\right), a + \left(\frac{b-a}{K}\right)k\right] \text{ para } k = 1, ..., K$$

Sea N_k el número de observaciones en el intervalo k, entonces:

$$N_k = \sum_{i=1}^{N} \mathbf{1}\left\{a + (k-1)\left(\frac{b-a}{K}\right) < X_i < a + k\left(\frac{b-a}{K}\right)\right\} \quad k = 1, ..., K$$

[1] El caso de variables discretas no es tan bueno para motivar el uso de regresiones kernel. En dichos casos lo mas recomendable es simplemente el uso de histogramas.

Recordemos que X_i es una variable aleatoria. Por lo tanto, podemos calcular la proporción esperada de observaciones que caeran en un intervalo dado y, con ello, la probabilidad de que una observación elegida al azar caería en dicho intervalo:

$$\begin{aligned}\frac{\mathbb{E}(N_k)}{N} &= Pr\left(a+(k-1)\left(\frac{b-a}{K}\right) < X < a+k\left(\frac{b-a}{K}\right)\right) \\ &= \int_{a+(k-1)(\frac{b-a}{K})}^{a+k(\frac{b-a}{K})} f_X(x)dx\end{aligned}$$

En el caso de los histogramas, asumimos que el estimador de la densidad de todos los puntos dentro del intervalo es igual. Por lo tanto, siendo que la probabilidad (area de cada barra en el histograma) resulta de multiplicar la densidad ($\widehat{f_X}(x)$) por el ancho del bin ($\frac{b-a}{K}$), podríamos obtener la densidad en un punto específico (x) solo despejando[2]:

$$\widehat{f_X}(x) = \frac{\frac{N_k}{N}}{\frac{b-a}{K}} = \frac{N_k}{N}\frac{K}{b-a} \quad \text{para} \quad x \in \left[a+(k-1)\left(\frac{b-a}{K}\right), a+k\left(\frac{b-a}{K}\right)\right]$$

Un aspecto que debemos decidir al llevar a cabo los histogramas es la selección del número de bins, o equivalentemente, el ancho de cada bin. Esto equivale a seleccionar K. La selección del ancho de cada bin es análogo a lo que veremos después como la selección del *bandwidth* (h) en el caso de densidades kernel. La intuición clave es que conforme mas pequeñno sea el ancho de los bins, menor será el sesgo de nuestro estimador, pero mayor será la varianza. El sesgo disminuye porque al disminuir el ancho del bin, estamos aumentando la precisión de la estimación. El aumento en la varianza surge porque corremos un mayor riesgo de que nuestro bin capture la cantidad adecuada de observaciones en el muestreo.

Empezaremos por ver el argumentos en términos de sesgo. Imaginen que nos interesa estimar la densidad en un punto específico (c). La definición de sesgo será:

$$\begin{aligned}f_X(c) - \mathbb{E}\left[\widehat{f_X}(c)\right] &= f_X(c) - \frac{\int_{a+(k-1)(\frac{b-a}{K})}^{a+k(\frac{b-a}{K})} f_X(x)dx}{\left(\frac{b-a}{K}\right)} = f_X(c) - \int f_X(x)dx\left(\frac{K}{b-a}\right) \\ &= f_X(c) - f_X(\tilde{a})\end{aligned}$$

[2]Nota que la diferencia entre el estimador de la densidad $\widehat{f_X}(x)$ y la densidad poblacional $f_X(x)$ es que la densidad poblacional es no observable y es precisamente lo que queremos estimar.

para $\tilde{a} \in \left(a + (k-1)\left(\frac{b-a}{K}\right), a + k\left(\frac{b-a}{K}\right)\right]$

Conforme $\uparrow K$ disminuye el sesgo porque hace más chicos los intervalos, ganando exactitud.

Dado lo argumentado anteriormente, uno podría pensar que lo mejor entonces es hacer lo más pequeñnos posible los intervalos. Sin embargo, esta afirmación no es del todo correcta por el hecho de que al estar llevando a cabo una estimación poblacional, lo que nos interesará es tener un intervalo de confianza acerca de nuestro estimador. A continuación lo que haremos es demostrar que existe un tradeoff entre sesgo y varianza al elegir la longitud de los intervalos: mientras más pequeñnos sean los intervalos menor será el sesgo, pero también mientras menores sean los intervalos mayor será la varianza.

El resultado anterior indica que la probabilidad de tener una observacion aleatoria en el intervalo $\left[a + (k-1)\left(\frac{b-a}{K}\right), a + k\left(\frac{b-a}{K}\right)\right]$ es:

$$\hat{p} = \hat{f}(\tilde{a})\frac{b-a}{K} \Longrightarrow \frac{K}{b-a}\hat{p} = \hat{f}(\tilde{a})$$

Por lo tanto, la probabilidad del intervalo p entre la longitud del intervalo $\left(\frac{b-a}{K}\right)$ es una forma de estimar la densidad de un punto. Esto se debe a que se distribuye uniformemente la probabilidad del intervalo.

Las observaciones aleatorias $\{X_1, .., X_n\}$ tendrán una distribución binomial donde una dummy indicará si caen en el intervalo. La estimación de p surge de promediar las dummies (cuantas caen en el intervalor) y por lo tanto, la varianza podrá surgir de estimar la varianza de p entre N debido a que estamos sacando la varianza de un promedio:

$$\begin{aligned} Var(p) = \hat{p}(1-\hat{p})\frac{1}{N} &= f(\tilde{a})\left(\frac{b-a}{K}\right)\left[1 - f(\tilde{a})\left(\frac{b-a}{K}\right)\right]\frac{1}{N} \\ \Longrightarrow Var(\hat{f}(\tilde{a})) &= Var\left(\frac{K}{b-a}p\right) = \left(\frac{K}{b-a}\right)^2 Var(p) \\ &= f(\tilde{a})\left(\frac{K}{b-a} - f(\tilde{a})\right)\frac{1}{N} \end{aligned}$$

Cabe señnalar dos cosas en la derivación de la varianza:

- $\uparrow K \Longrightarrow \uparrow Var\left(f(\tilde{a})\right)$
- $\uparrow N \Longrightarrow \downarrow Var\left(f(\tilde{a})\right)$

6.1.1. Histogramas Centrados

Un problema en el uso de histogramas para estimar la densidad de una variable es que puntos adyacentes pueden ser muy distintos si caen de uno u otro lado de

la frontera de un intervalo. Dicho de otra manera, hay cambios discontinuos en el estimador de la densidad justo en la frontera de cada bin. Esto genera mayor sesgo cerca de la frontera del intervalo que en el centro.

Una solución es estimar la densidad asumiendo que cada punto es el centro del intervalo:

$$\hat{f}(c) = \sum_{i=1}^{N} \mathbf{1}\left\{c - h \leq x_i \leq c + h\right\} \frac{1}{2hN}$$

Esta metodología es equivalente a lo que más adelante definiremos como un estimador de densidad kernel uniforme.

6.1.2. Estimador de nearest neighbor

Otra alternativa consiste en utilizar un estimador de *nearest neighbor*. Empecemos por dar una intuición. El concepto de *nearest neighbor* se define como aquella observación que se encuentra lo más cercano posible a un punto. Posteriormente, se define al k-ésimo *nearest neighbor* como el k-ésimo en estar lo mas cerca posible si las observaciones se ordenan en términos de distancia[3]. Para el caso de una distribución la definición de distancia no tiene ninguna complicación ya que estamos trabajando en una dimensión. Con ello podemos definir la distancia al k-ésimo *nearest neighbor* como:

$$\begin{aligned} d_k(x) = \text{argmin } & d \\ \text{s.t. } & d \geq 0 \\ & 1\{|X_i - x| \leq d\} \geq k \end{aligned} \tag{6.1}$$

Con este concepto de distancia, el estimador de densidad de *nearest neighbor* es similar a los histogramas. En este caso, se tiene q fijar k, donde tener una k mayor o menor replica el tradeoff de sesgo y varianza que se explicó en el caso de histogramas con el ancho del bin. Una vez especificado k el estimador de densidad se construye como:

$$\hat{f}_X(x) = \frac{k-1}{2\ N\ d_k(x)} \tag{6.2}$$

La lógica en esta ecuación es que:

- $k-1$ son el número de individuos que se observan en un intervalo específico
- Dicho intervalo tiene longitud 2 $d_k(x)$ ya que la distancia puede ir hacia dos lados

[3] Es decir, el primero es la observación más cercana a un punto determinado, el segundo es la segunda observación más cercana, y asi sucesivamente hasta llegar a la k-ésima observación más cercana a dicho punto.

- Dividir $k-1$ sobre N indica la proporción de observaciones que posteriormente se distribuyen uniformemente en el intervalo 2 $d_k(x)$

6.2. Kernel Density Estimation

Intuición: Cada observación de X_i tiene una masa 1 que distribuirá en el soporte usando una función, llamada la función Kernel. Esta función Kernel es una densidad. Una vez distribuida la masa de todas las observaciones, al sumar todas las masas, habremos distribuido una masa N. Solamente normalizamos dividiendo entre N y esto nos da el estimador de la densidad. Al distribuir la masa de cada observación será muy relevante decidir qué tan lejos se distribuirá, esto es, la selección del *bandwidth* (h).

Veamos cómo hacerlo con una distribución uniforme y empecemos por asumir que tenemos un *bandwidth* específico dado por h. Más adelante discutiremos como seleccionar dicho bandwidth.

$$K(z) = \frac{1}{2} \quad \text{si} \quad z \in (-1, 1)$$

Con esta función utilizaremos el espíritu de histogramas centrados. Tomemos un histograma centrado en x; para saber si una observación especifica le reparte densidad a este punto calculamos:

$$\begin{aligned} K\left(\frac{X_i - x}{h}\right) &= \frac{1}{2} \quad \text{si} \quad \frac{X_i - x}{h} \in (-1, 1) \\ &= \frac{1}{2} \quad \text{si} \quad X_i \in (x - h, x + h) \end{aligned}$$

Ahora repetimos con todas las observaciones, sumamos y dividimos entre N:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)$$

Además de la función Kernel Uniforme, existen también los siguientes casos populares de la función K:

1. $K(z) = (1 - |z|)$ si $z \in (-1, 1)$ Triangular
2. $K(z) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}z^2\right)$ Gaussian
3. $K(z) = \frac{3}{4\sqrt{5}}\left(1 - \frac{z^2}{5}\right)$ si $z \in (-\sqrt{5}, \sqrt{5})$ Epanechnikov

Las funciones Kernel tienen las siguientes propiedades:

- $\int K(u)du = 1 \rightarrow$ Masa de la densidad = 1

- $\int uK(u)du = 0 \rightarrow$ Valor esperado del error usando como peso la densidad Kernel es igual a 0
- $\int u^2K(u)du = k_2 > 0 \rightarrow$ permite estimar la varianza. (Similar a valor esperado de errores al cuadrado)

6.3. Selección de Bandwidth (h)

No existe una regla general para la selección óptima de un *bandwidth* (h). Muchas veces lo que se hace es probar distintos bandwidths hasta que la distribución ya no tiene un comportamiento muy escalonado, es decir, da una distribución "smooth' '.

Una alternativa ampliamente utilizada consiste en elegir el bandwidth utilizando como función objetivo minimizar la integral de los errores al cuadrado[4].

La integral de los errores al cuadrado es una función conveniente para la selección de h porque representa el tradeoff entre sesgo y varianza debido a que se puede demostrar que:

$$\text{ISE} = \int \left(\hat{f}(x) - f(x)\right)^2 dx = \int \text{sesgo}^2\left(\hat{f}(x)\right) dx + \int Var\left(\hat{f}(x)\right) dx \quad (6.3)$$

Empecemos por revisar el componente del sesgo:

$$\begin{aligned}
\text{sesgo}\left(\hat{f}(x)\right) &= \mathbb{E}\left(\hat{f}(x)\right) - f(x) = \mathbb{E}\left[\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)\right] - f(x) \\
&= \frac{1}{Nh}\sum_{i=1}^{N} \mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] - f(x) \underset{\text{iid}}{=} \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] - f(x) \\
&= \frac{1}{h}\int K\left(\frac{y-x}{h}\right) f(y)dy - f(x) = \frac{1}{h}\int K\left(u\right) f(x+hu)h{\cdot}du - f(x)
\end{aligned}$$

Donde $u = \frac{y-x}{h}$ y $dy = h \cdot du$

$$= \int K(u)f(x+hu)du - f(x) \quad (6.4)$$

[4]Esto no es lo mismo que el valor esperado de los errores al cuadrado. El valor esperado de los errores al cuadrado daría un peso distinto a las observaciones dependiendo de la densidad.

$$\therefore \quad \text{si } h \to 0$$

$$= \int K(u)f(x)du - f(x) = f(x) \overbrace{\int K(u)du}^{\nearrow 1} - f(x)$$

$$\implies \quad \text{sesgo} \to 0$$

Partiendo de la ecuación (6.5) podemos utilizar una expansión de Taylor de segundo grado en h:

$$\text{sesgo}\left(\hat{f}(x)\right) \approx \int \overbrace{K(u)f(x)}^{\nearrow f(x)} du + \int \overbrace{K(u)uhf'(x)}^{\nearrow 0} du + \frac{1}{2}\int K(u)u^2h^2f''(x)du - f(x)$$

$$= h^2k_2f''(x)$$

$$\therefore \quad \int \text{sesgo}^2\left(\hat{f}(x)\right) = \frac{1}{4}h^4k_2^2\int \left(f''(x)\right)^2 dx$$

Puede verse en esta fórmula que conforme disminuye h, el sesgo tiende a cero. Siguiendo pasos similares podemos calcular la varianza (la demostración va más alla de lo que se pretende en esta clase y se dejan las referencias para el lector interesado):

$$Var\left(\hat{f}(x)\right) = \frac{1}{Nh}f(x) \int K(u)^2du \quad - \quad \frac{1}{N}f(x)^2 \tag{6.5}$$

Cabe notar de la derrivación de esta varianza dos aspectos. En primer lugar, esta fórmula es útil en el caso de que nos interese generar una distribución o hacer una prueba de hipótesis acerca de $f(x)$. Para ello habría que asumir una función kernel $K(.)$ y emplear el h óptimo. En segundo lugar, en proceso de selección óptimo de h el segundo término de la fórmula anterior no importa debido a que h no interviene en ese componente. Por ello no lo empleamos en los siguientes términos. Usando el primer término de dicha varianza tenemos que:

$$\int \text{Var}\left(\hat{f}(x)\right) dx \approx \frac{1}{Nh}\int K(u)^2du$$

Aqui podemos notar que conforme h disminuye, la varianza aumenta. Por lo tanto, ambos componentes que hemos derivado efectivamente representan el tradeoff entre sesgo y varianza que se pretendía. Sustituyendo estos términos en la ecuación (6.3) obtenemos:

$$\text{ISE} = \frac{1}{4}h^4k_2^2\int \left(f''(x)\right)^2 dx + \frac{1}{Nh}\int K(u)^2du \tag{6.6}$$

Si minimizamos esta función con respecto a h obtenemos (despues de calcular CPO):

$$h^* = k_2^{-\frac{2}{5}} N^{-\frac{1}{5}} \left(\int K(u)^2 du \right)^{1/5} \left(\int f''(x)^2 dx \right)^{-1/5} \tag{6.7}$$

Si bien, este es el *bandwidth* óptimo tenemos el problema de que depende de $f(\cdot)$ que es lo que queremos estimar en primer lugar. Cabe señnalar algunas cosas con respecto a este valor óptimo:

- Conforme N aumenta, h^* disminuye (pero lentamente por el factor (1/5)
- La $f''(x)$ representa que tan suave es la curvatura de la función de densidad estimada. Un valor pequeñno (en términos absolutos) para la $f''(x)$ indica una curvatura suave con cambios no abruptos. Este término agrega el cuadrado de dicha curvatura a lo largo de toda la distribución de x. Curvaturas suaves estan relacionadas con menores valores de h^*.

La mayor parte de la selección de bandwidth óptimo parten de la función objetivo (6.3). Algunos métodos siguiente este proceso y utilizando el resultado de la ecuación (6.7) hacen diferentes supuestos para encontrar una solución. Empezaremos viendo dos ejemplos de este tipo de procedimientos.

6.3.1. Asumir una Distribución

Silverman (86) asume que tanto $K(\cdot)$ como $f(\cdot)$ son normales. Si tomamos este supuesto, la ecuación (6.7) puede simplificarse enormemente para obtener:

$$h^*_{Silv} = 1{,}06 \cdot \sigma_X \cdot N^{-1/5} \tag{6.8}$$

A h^*_{Silv} se le conoce como el *Silverman Rule of Thumb*.

Además, hay una alternativa más robusta a outliers que, en vez de utilizar σ_X, utiliza:

$$h^*_{robust} = 1{,}06 \cdot N^{-1/5} \cdot min \left\{ \sigma_X, \frac{R}{1{,}34} \right\}$$

donde R es el rango intercuartil $Q(X)_{75} - Q(X)_{25}$.

Cabe señnalar que la función Kernel *Epanechnikov* curiosamente es la que, dada una distribución, minimiza el valor esperado del ISE.

6.3.2. Plug-In Methods

Estos métodos son más intensivos en cálculo y consisten en los siguientes pasos:

1. tomemos un valor inicial de h, llamémoslo h_0 (pudiera ser el Silverman o un valor elegido aleatoriamente)
2. Utilizando este h_0 calculamos: $\int \hat{f}''(x)dx$
3. Con este valor calculamos $h_1 = k_2^{-2/5} N^{-1/5} \left(\int K(u)^2 du\right)^{1/5} \left(\int \hat{f}''(x)dx\right)^{-1/5}$

4.Iteramos este proceso hasta lograr convergencia en h^*.

Este método fue sugerido por *Scott, Tapia y Thompson (77).*

Existen otros *plug-in methods* que cambian la forma de aproximar $\int \hat{f}''(x)dx$ o en usar una expansión de Taylor con más términos en la varianza en ISE. Esto puede verse en Turlach (93).

6.3.3. Cross-Validation

Este es otro método clásico para determinar h^* y existe en diferentes versiones. Aquí veremos dos:

6.3.3.1. Least-Squares Cross Validation

Recordemos que nuestra función objetivo es minimizar el *ISE* (ecuación (6.3):

$$\text{ISE}(h) = \int \left(\hat{f}(x) - f(x)\right)^2 dx$$

Si desarrolamos el término cuadrático obtenemos:

$$\int \hat{f}(x)^2 dx - 2\int \hat{f}(x)f(x)dx + \int f(x)^2 dx$$

De estos términos, sólo $\hat{f}(x)$ depende de h, por lo tanto, minimizar ISE será lo mismo que minimizar $\mathbb{L}(h)$, donde:

$$\mathbb{L}(h) = ISE - \int f(x)^2 dx = \int \hat{f}(x)^2 dx - 2\int \hat{f}(x)f(x)dx$$

Notemos que $\hat{f}(x)^2 dx$ se puede calcular directamente a partir de asumir una función Kernel y agregar la estimación de las densidades a lo largo del rango

de posibles valores de x. Sin embargo, $f(x)$ no es observable. Por lo tanto, utilizamos como estimador de $\mathbb{L}(h)$ a:

$$\mathrm{CV}_{LS}(h) = \int \hat{f}(x)^2 dx - 2\frac{1}{n}\sum_i \hat{f}_{-i}(X_i)$$

El segundo término consiste en estimar la densidad de X_i tomando en cuenta todas las observaciones de la muestra, excepto i y luego promediar a través de X_i:

$$f_{-i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)$$

Finalmente *Cross Validation* resuelve numéricamente utilizando muchos valores de h y viendo cuál da un valor más chico de $CV_{LS}(h)$.

6.3.3.2. Likelihood CV

Este método consiste en preguntarse: >Qué pasaría si tuviéramos una observación más independiente? ¿Qué tan buena sería la estimación de $\hat{f}(x)$ para predecir la densidad?

Para no elegir arbitrariamente una observación, sacamos el promedio de qué pasaría si no tuviéramos cada una de las observaciones de la muestra:

$$\Longrightarrow \quad \max_h \quad CV_L(h)$$

donde $CV_L(h) = \frac{1}{n}\sum_{i=1}^{n} \log \widehat{f_{-i}}(X_i)$

Este método fue propuesto por Haberman, Herman, y Van der Brock (74) y Duvin (76), aunque tiene algunos predecesores.

6.4. Regresiones Kernel

Este es un método para estimar regresiones sin asumir ninguna forma funcional. Nota que, con variables X discretas simplemente promediamos. Este método, que se centra en suavizar la tendencia, se basa más en el caso de X continuas. La metodología es muy similar a la estimación de densidades.

Primero definamos la regresión: $g(x) = E(Y|X = x)$. Empecemos imaginando que conocemos el estimador de la densidad conjunta $\widehat{f_{YX}}(y, x)$. Con esto podríamos estimar una densidad condicional y una vez teniendo esto:

$$\widehat{g(x)} = \int y\widehat{f_{Y|X}}(y|x)dy = \int y\frac{\widehat{f_{YX}}(y,x)}{\widehat{f_X}(x)}dy = \int y\frac{\widehat{f_{YX}}(y,x)}{\int \widehat{f_{Y,X}}(z,x)dz}dy$$

Para simplificar la estimación asumiremos que tenemos un Kernel bivariado que se puede separar en dos Kernel univariados $K(u,v) = K_1(u)K_2(v)$.

El denominador:

$$\widehat{f_X}(x) = \int f_{YX}(z,x)dz = \frac{1}{nh^2}\int \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}, \frac{Y_i - y}{h}\right) dz = \frac{1}{nh^2}\int \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}, v\right) h \cdot dv$$
$$= \frac{1}{nh}\int \sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h}\right) K_2(v)\, dv = \frac{1}{nh}\sum_{i=1}^{n} K_1\left(\frac{X_i - x}{h}\right)$$

El numerador:

$$\int y\widehat{f_{YX}}(y,x)dy = \frac{1}{nh^2}\int \sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}, \frac{Y_i - y}{h}\right) dy = \frac{1}{nh}\int \sum_{i=1}^{n} y_i K_1\left(\frac{X_i - x}{h}\right) K_2(v)\, dv$$
$$= \frac{1}{nh}\sum_{i=1}^{n} y_i K_1\left(\frac{X_i - x}{h}\right)$$

Por lo tanto:

$$\hat{g}(x) = \sum_{i=1}^{n} w(X_i, x)\, Y_i$$

donde:

$$w(X_i, x) = \frac{K_1\left(\frac{X_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)}$$

Esta estimación se conoce como la *Nadaraya Watson*

6.4.1. Regresion lineal local (LL)

El estimador de *Nadaraya Watson* también resultaría de:

$$\hat{\alpha} = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(y_i - a)^2$$

$$\mathcal{L} = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(y_i - \alpha)^2$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = -2\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(y_i - \alpha) = 0 \Longrightarrow$$

$$\alpha = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

Esto sugiere que también podríamos encontrar otras formas funcionales. En vez de esto, podríamos definir una *regresión lineal local* (LL): $g(x) = \alpha + \beta x$, donde:

$$(\hat{\alpha}, \hat{\beta}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(y_i - a - bx_i)^2$$

El resultado de esto es muy cercano a OLS:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\left[w(x_i, x)\left(x_i - \sum_{j=1}^{n} w(x_j, x)\, x_j\right)\left(y_i - \sum_{j=1}^{n} w(x_j, x)\, y_j\right)\right]}{\sum_{i=1}^{n} w(x_i, x)\left(x_i - \sum_{j=1}^{n} w(x_j, x)\, x_j\right)^2}$$

$$\hat{\alpha} = \sum_{i=1}^{n} w(x_i, x)\, y_i - \hat{\beta}\sum_{i=1}^{n} w(x_i, x)\, x_i$$

Esto corresponde a una regresión local ponderada por K.

Una característica favorable es que LL suele ser un mejor estimador cerca de las fronteras.

Nadaraya Watson es mejor si la relación entre y y x es más plana. LL es mejor si es más irregular.

6.4.2. Cross Validation (Jackknife)

Para determinar h, nuevamente podemos usar CV :

1. Sea $\widehat{g_{-i}}^{h}(x) = \sum_{j \neq i} w^h(X_j, x)\, Y_j$, donde: $w^h(X_j, x) = K\left(\frac{X_j - x}{h}\right) / \sum_{k \neq i} K\left(\frac{X_k - x}{h}\right)$

2. Definimos el siguiente criterio de *Cross-Validation*

$$CV(h) = \sum_{i=1}^{n} \left(\widehat{g_{-i}}^{h}(X_i) - Y_i \right)^2$$

3. Elegimos h tal que $h^* = \text{argmin}\, CV(h)$, lo cual se puede resolver numéricamente.

6.4.3. Distribución del estimador

Para poder llevar a cabo inferencia, nos interesará la distribución del estimador. Al igual que en la discusión acerca del tradeoff de sesgo y varianza, en el caso de la varianza, esta tendrá una relación negativa con el bandwidth. No llevaremos a cabo la derivación formal de la varianza, pero hay dos cosas por señnalar. Primero, la varianza tendrá la siguiente forma:

$$Var\left(\hat{g}(x)\right) = \frac{1}{N \cdot h} \cdot \frac{\sigma^2(x)}{\hat{f}(x)} \int K(u)^2 du + o\left(\frac{1}{N \cdot h}\right) \tag{6.9}$$

donde $sigma^2(x)$ resultará de calcular los residuales $Y_i - \hat{g}(X_i)$ y posteriormente hacer una regresión kernel del cuadrado de los residuales vs X. Todos los demás términos son conocidos. El término final $o\left(\frac{1}{Nh}\right)$ es un término que conforme N aumenta va disminuyendo, lo cual hace que la distribución asintótica (muestras grandes) no sea relevante este último término.

Es común en la práctica ver estrategias que utilizan un bandwidth pequeñno para el estimador y bandwidths un poco mas grandes para la varianza. En clase discutiremos brevemente esta práctica.

Capítulo 7

Experimentos aleatorizados

En las secciones anteriores detallamos el uso de diferentes herramientas de uso común en análisis econométricos. A partir de esta sección la discusión se enfoca en el problema de identificar efectos causales de una variable (X_1) sobre la variable dependiente de nuestro modelo (Y). El análisis causal está intrinsecamente ligado a la creación de modelos que en economía se plantean para entender de foma estructurada la forma en que individuos toman decisiones y cómo el cambio de algunos factores pueden influir de forma positiva o negativa en dichas decisiones. Lo que un modelo teórico plantea como estática comparativa, que consiste en ver el cambio en alguna variable de decisión que resulta del cambio de otra variable, un análisis empírico busca estimar vía un análisis causal.

Además, el análisis causal ha ganado creciente atención en diversos contextos mediante su aplicación en la evaluación de impacto de proyectos.[1] Su uso se ha promovido intensamente como parte de una búsqueda de hacer proyectos basados en evidencia estadística. Esto ha resultado en una extensa promoción del uso de experimentos aleatorizados, implementados y analizados desde la perspectiva de ciencias sociales. En esta y las siguientes secciones nos enfocaremos en el análisis causal. Empezamos en esta sección con los experimentos aleatorizados, ya que es el método que goza de mayor aceptación por su alta validez interna. Las secciones siguientes exploran los métodos cuasi-experimentales que, de diferentes formas y bajo ciertos supuestos, replican condiciones similares a las de un experimento.

Antes de discutir el método de experimentos aleatorizados, empezamos esta sección con un conjunto de definiciones que son relevantes en cualquier análisis causal, independientemente del método utilizado. Recordamos también el problema de auto-selección como motivacion para el uso de los métodos experimentales y cuasi-experimentales.

[1] El *Poverty Action Lab* (J-PAL) ha sido un fuerte promotor en EU y a nivel mundial de su uso vinculando academicos y hacedores de política en el uso de los experimentos aleatorizados.

7.1. Fundamentos

Para facilitar el análisis causal empezaremos por comparar el impacto de la implementación de alguna intervención. Podemos entender esto como la aplicación de algun programa cuyo impacto nos interesa medir. Siendo consistentes con la terminología ampliamente usada, describiremos a la implementación de dicho programa como el **Tratamiento** (T).[2] Para poder medir su impacto, compararemos la implementación del proyecto respecto al *statu quo* o falta de programa al cual nos referiremos como el **Control** (C)[3]. Imaginemos, por ejemplo, que nos interesa medir el impacto de un programa de construcción de bibliotecas sobre las capacidades lectoras (medidas con algún examen estandarizado). En este caso, el grupo de *tratamiento* correspondería a lugares donde se contruyó una biblioteca y el *control* a lugares donde no se construyó. Para distinguir en una base de datos entre ambos grupos utilizaremos una variable dummy T_i que tendrá valor igual a uno $(T_i = 1)$ si la observación i recibe tratamiento y valor igual a cero $(T_i = 0)$ si es parte del grupo de control.

Todo diseño de una intervención efectiva es importante que venga acompañada de trabajo previo que fundamente el por qué y el cómo de la intervención. En particular, es útil documentar el problema que se busca resolver y sustentar cómo una intervención pudiera ser efectiva para resolver el problema. Siguiendo esta misma lógica, al proponer un intervención es importante desarrollar una **teoría de cambio**. La teoría de cambio es una representación gráfica que busca dar sustento y estructura a la intervención. Va desde los insumos de la intervención (en nuestro ejemplo la construcción de una biblioteca) hasta los resultados finales que la intervención busca modificar (en nuestro ejemplo, las capacidades lectoras). Sin embargo, este vínculo entre insumos y resultados finales suele venir intermediado por distintos *canales*. Por ejemplo, para que la construcción de bibliotecas logre aumentar las capacidades lectoras, podría darse el caso que la biblioteca mejora el acceso y promueve el interés en la lectura, lo cual genera mejores hábitos de lectura, mismos que logran aumentar las capacidades lectoras. En este caso podríamos ver a la *mejora en acceso* y la *promoción de interés* en la lectura como variables intermediarias de primer nivel y a los mejores *hábitos de lectura* como un intermediario de segundo nivel.

Una vez establecida la teoría de cambio, el siguiente paso es identificar cuáles son las métricas que se emplearán para medir los conceptos que sea posible en la teoría de cambio. Para ello, suele utilizarse una **matriz de indicadores** que relacione los conceptos incluidos en la teoría de cambio con las métricas a utilizar. Por ejemplo, en el caso de la *mejora en el acceso* podría utilizarse como métrica la *distancia de los hogares a la biblioteca mas cercana*; como métrica

[2] En la exposición que utilizamos en estas notas hacemos la comparación de **un tratamiento** y el control. Esta descripción la hacemos por simplicidad, aunque es posible plantear una intervención que tenga más de un tipo de tratamiento. Mas adelante en las notas describimos esta situación.

[3] El Control también puede ser entendido como una política alternativa, aunque generalmente en los análisis de impacto suele ser la falta de tratamiento.

de *hábitos de lectura* podría recopilarse información administrativa de *número de prestamos de libros* o alternativamente vía una encuesta, preguntar en los hogares el *número de libros leidos al mes*. Esta matriz de indicadores establece el vínculo entre la teoría de cambio y la base de datos que se utilizará para hacer el análisis econométrico.

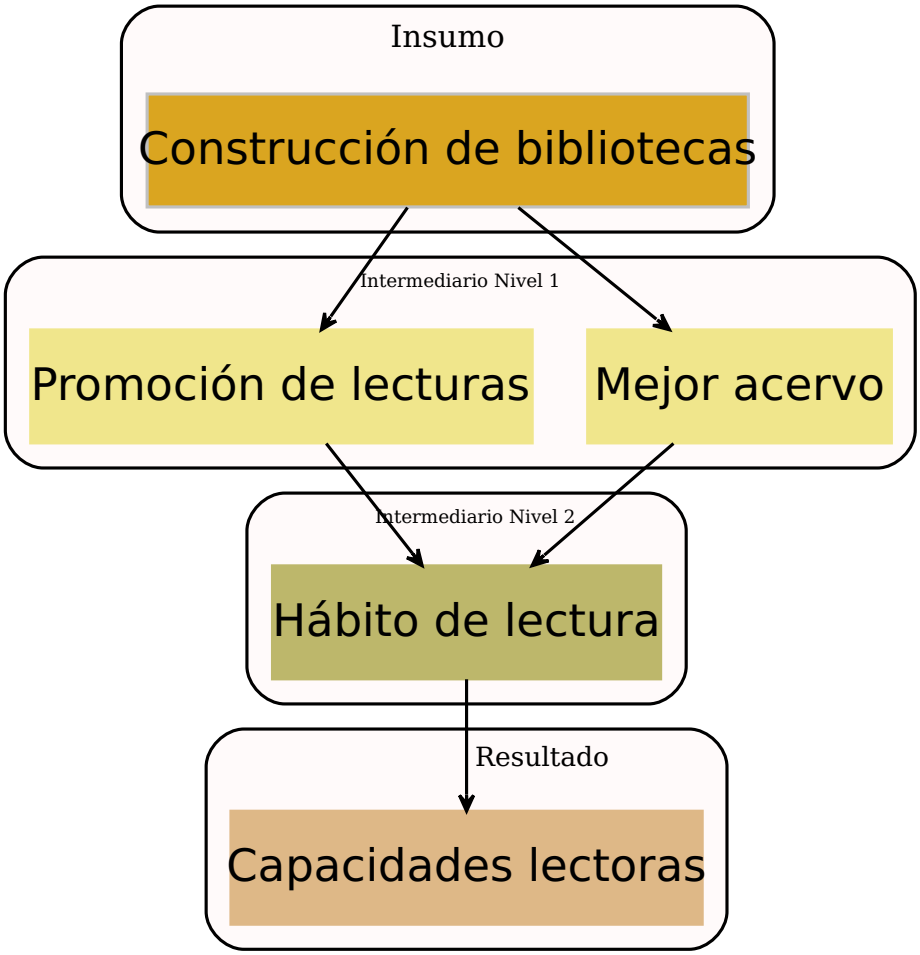


Figura 7.1: Teoría de Cambio

En las secciones subsecuentes describiremos cómo llevar a cabo el análisis econométrico del impacto de la intervención sobre el resultado final. Los resultados intermedios se pueden utilizar cómo parte de un análisis separado donde se explore si algunos de los canales propuestos parecen estar activos. Las estrategias econométricas descritas a continuación aplican para identificar el efecto agregado que la intervención tiene sobre los resultados finales. Vale la pena aclarar que estas estrategias no deben entenderse como una forma de identificar el efecto

de los distintos resultados intermedios sobre el resultado final, sino solo como el efecto de la intervención sobre el resultado final, donde varios de los canales propuestos en la teoría de cambio pudieran explicar los efectos encontrados. No será posible desagregar en que proporción cada resultado intermedio explica el resultado final, sino solo el acumulado de todos ellos. Sin embargo, las estrategias econométricas que describiremos pueden aplicarse para medir el impacto de la intervención sobre cada uno de los indicadores intermedios por separado. Para esto simplemente se tiene que utilizar cada resultado intermedio como variable dependiente (Y).

7.1.1. Resultados potenciales

Continuando con nuestro ejemplo, supongamos que estamos interesado en medir el efecto de la construcción de bibliotecas sobre las capacidades lectoras y utilizamos el resultado de una prueba como métrica del resultado final. Sea Y_i^T el resultado del niño i si recibe el tratamiento (i.e. vive en un lugar donde se construyó una biblioteca) y Y_i^C el resultado del mismo niño i si está en el control (i.e. vive en un lugar donde NO se contruyó biblioteca). A estos valores se les conoce como **resultados potenciales**. A nosotros nos interesa el valor de $Y_i^T - Y_i^C$ que corresponde al **efecto de tratamiento** sobre el individuo i.

Una limitación para poder identificar dicho efecto consiste en que para cada individuo solo es posible observar ya sea Y_i^T o Y_i^C. Para esto, necesitamos aplicar el supuesto de *no interferencia*. Este supuesto consiste en que el resultado observable Y_i solo depende de de la asignación a tratamiento del propio individuo y no es afectado por la asignación a tratamiento de algún otro individuo j. Pensemos, por ejemplo, que j es el mejor amigo del individuo i. Si j es asignado a tratamiento y tiene mejor acceso a una biblioteca, podría suceder que i, pese a estar en control, vea su resultado potencial Y_i^C afectado por j. Este caso, y otros de externalidades que discutiremos mas adelante, representan una amenaza al supuesto de *no interferencia*. Este supuesto también es descrito como el supuesto **SUTVA** (*Stable Unit Treatment Value Assumption*) por sus siglas en inglés.

Definamos a Y_i como el resultado *observable* para el econometrista. Bajo el supuesto de *SUTVA* podemos definir:

$$Y_i = Y_i^T T_i + Y_i^C (1 - T_i) \tag{7.1}$$

Dado que para cada individuo solo podemos observar uno de los dos resultados potenciales, usualmente el parámetro de interés que buscamos identificar consiste en una medida central de la distribución de los efectos de tratamiento. El parámetro mas común es el **efecto promedio de tratamiento (Average Treatment Effect, ATE)**:

$$\tau = E(Y_i^T - Y_i^C) \tag{7.2}$$

7.1.2. Sesgo de autoselección

Un aspecto fundamental para poder utilizar el resultado observable Y_i para estimar el ATE es el método de asignación del tratamiento. Si el tratamiento se elige de forma voluntaria por los individuos, al tratar de estimarlo utilizando *mínimos cuadrados ordinarios* encontraremos lo que comúnmente llamamos el **sesgo de autoselección**. El sesgo de autoselección se puede asociar directamente al sesgo por variables omitidas que previamente describimos en el capítulo de MCO.

Imaginemos que queremos estimar el ATE con una regresión simple y el resultado observable. En dicho caso estimaríamos:

$$Y_i = \beta_0 + \beta_1 \, T_i + U_i \tag{7.3}$$

Como previamente vimos en interpretación de MCO tendríamos que:

$$\begin{aligned}\beta_1 &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \\ &= E(Y_i^T T_i + Y_i^C(1 - T_i)|T_i = 1) - E(Y_i^T T_i + Y_i^C(1 - T_i)|T_i = 0) \\ &= E(Y_i^T|T_i = 1) - E(Y_i^C|T_i = 0)\end{aligned}$$

En este caso β_1 corresponde a la diferencia del promedio de niños que recibieron el tratamiento y el promedio de niños de control. Para que β_1 sea igual al ATE, en la ecuación anterior, sumemos y restemos $E(Y_i^C|T_i = 1)$. Este valor es no observado y corresponde al promedio que los niños de tratamiento hubieran tenido si no hubieran recibido el tratamiento (un *contrafactual*):

$$\beta_1 = [E(Y_i^T|T_i = 1) - E(Y_i^C|T_i = 1)] + [E(Y_i^C|T_i = 1) - E(Y_i^C|T_i = 0)] \tag{7.4}$$

El primer término de (7.4) se puede expresar como $(E(Y_i^T - Y_i^C|T_i = 1))$ y corresponde al **efecto promedio de tratamiento de los individuos tratados** (*Treatment on the Treated*, TOT). El segundo término $(E(Y_i^C|T_i = 1) - E(Y_i^C|T_i = 0))$ corresponde al **sesgo por selección**. En nuestro ejemplo representa la diferencia en el promedio de niños de tratamiento y control si el tratamiento no hubiese existido (i.e. si ambos hubieran recibido el control). Por ejemplo, supongamos que tenemos niños en localidades que decidieron construir una biblioteca y otros en localidades que no lo hicieron. Si la construcción de escuelas está explicada porque en esas localidades existía una mayor demanda de niños y padres de familia por tener mejor acceso a material de lectura, es posible que aún sin la existencia de las bilbiotecas, dichos padres y niños hubieran conseguido material de lectura de otra forma. Por lo tanto, podría ser razonable asumir que sus capacidades lectoras aún sin bibliotecas se hubieran

desarrolado mejor y tendríamos que $E(Y_i^C|T_i = 1) > E(Y_i^C|T_i = 0)$. Por lo tanto, el segundo término de la ecuación (7.4) es positivo y β_1 sobre-estimaría el TOT.[4]

7.1.3. Aleatorización del tratamiento

Los **experimentos aleatorizados (Randomized Control trials, RCTs)** resuelven el probema del sesgo de autoselección asignando el tratamiento al azar. Si T_i es una asignación aleatoria será fácil de justificar el **supuesto de independencia**:

$$\{Y_i^T, Y_i^C\} \perp T_i$$

Bajo este supuesto, el término de autoselección desaparece. La intuición de esto es que si la decisión de construir las bibliotecas es al azar, uno esperaría que tanto en lugares con y sin acceso a bilbiotecas exista una demanda similar de parte de los niños y padres por material de lectura. Por lo tanto, la distribución de Y_i^C es similar en ambos grupos y se justifica que $E(Y_i^C|T_i = 1) = E(Y_i^C|T_i = 0)$. Asimismo, respecto al primer término, esperaríamos que los efectos de tratamiento se distribuyeran de forma similar en el grupo de tratamiento y control ($E(Y_i^T - Y_i^C|T_i = 1) = E(Y_i^T - Y_i^C|T_i = 0)$). Con ello tendríamos que el TOT y el ATE deberían ser iguales en valor esperado y podríamos estimarlos con β_1 de nuestra regresión simple de la ecuación (7.3).

Cabe señalar que en este caso el ATE es una estimación del efecto agregado del tratamiento. Dicho de otra manera, no resulta de una derivada parcial (*caeteris paribus*), sino de una derivada total. Esto se debe a que es muy posible que el tratamiento, además de tener un impacto directo sobre la variable dependiente, puede tener efectos indirectos, como describimos previamente. En nuestro ejemplo de la contrucción de bibliotecas, puede ser que los niños con acceso a bibliotecas estén más motivados y fomen mejores hábitos de lectura, como discutimos con la *teoría de cambio* . Ambos canales pueden llegar a tener efectos sobre las capacidades lectoras, pero será imposible determinar que parte del efecto que identificamos se debe a cada uno de los mecanismos descritos.

7.1.4. Beneficios de datos basales

Pese a que en un ambiente de *experimentos aleatorizados* no es estrictamente necesario levantar *datos basales*, entendidos como un conjunto de variables (X_i)

[4]En una argumentación de sesgo por variables omitidas diriamos que existe una variable omitida no observada, como el *interés por la lectura*, mismo que esta correlacionado positivamente con la existencia de una biblioteca cercana y con tener mejores capacidades lectoras. Por lo tanto, una estimación de OLS tendría un sesgo positivo para estimar el efecto de las bibliotecas sobre las capacidades lectoras si no es posible controlar por la variable de *interés por la lectura*.

que existen previo a la implementación de una intervención, es considerado una buena práctica ya que trae consigo diversos beneficios:

1. **Tablas de Balance.** A pesar de que no sea posible demostrar directamente que la asignación al tratamiento cumple con el supuesto de independencia, es común utilizar controles recopilados en la línea basal para dar evidencia indirecta de esto. La idea es que en un ámbito de ausencia del tratamiento, es posible utilizar un conjunto de variables explicativas para modelar a la variable potencial Y_i^C:

$$Y_i^C = X_i'\beta + U_i \tag{7.5}$$

   Para dar evidencia de que la asignación aleatoria fue satisfactoria en el sentido de *independencia*, las tablas de balance muestran la diferencia en las medias entre tratamiento y control para las variables X_i, **esperando que la mayoría de ellas no sean estadísticamente distintas**. En algunos casos incluso se recopila la variable que se utilizará como resultado previo a la implementación del tratamiento Y_i en $(t = 0)$ y se incluye en esta comparación. Estas tablas suelen mostrar:

   - El estadístico-t del test de diferencias de medias entre tratamiento y control (empleando varianzas distintas entre ambos grupos). Se espera que en una aleatorización satisfactoria estos estadísticos no sean estadísticamente distintos de cero.

   - El estadístico-F de la significancia conjunta de todos los coeficentes de la siguiente estimación, esperando obtener un valor-p alto:

$$T_i = X_i'\gamma + U_i$$

2. **Controles.** Posibilidad de utilizar controles como parte de la estimación del ATE. Esto incluye la posibilidad de incrementar la eficiencia (aunque típicamente de forma muy limitada) y de llevar a cabo análisis de heterogeneidad en el ATE.

3. **Estratificación.** Posibilidad de llevar a cabo estratificación en la asignación de tratamiento. En el diseño de un experimento esto se hace con el propósito de incrementar la eficiencia.

4. **Atrición.** Se refiere a la pérdida de observaciones que originalmente se encontraban en la asignación del tratamiento. Esta es una preocupación importante en la práctica en muchos experimentos aleatorizados, por lo que más adelante dedicamos una sección a este problema.

7.2. Estimaciones econométricas

7.2.1. Estimación con MCO

El estimador más comunmente empleado en la práctica para el análisis de experimentos aleatorizados es una regresión simple de *mínimos cuadrados ordinarios*, como lo describimos previamente en la ecuación (7.3):

$$Y_i = \beta_0 + \beta_1 \, T_i + U_i$$

En este caso, el coeficiente β_1 corresponde a un estimador insesgado del efecto promedio de tratamiento (ATE) bajo el supuesto de SUTVA y de asignación aleatoria del tratamiento. En las siguientes secciones detallaremos ajustes posibles y estimaciones alternativas al estimador simple de mínimos cuadrados otrdinarios para el análisis econométrico adecuado de los experimentos aleatorizados.

Si en la ecuación previa sustituimos $\beta_0 = E(Y_i^C)$, $\beta_1 = (Y_i^T - Y_i^C)$ y $U_i = Y_i^C - E(Y_i^C)$ podemos ver que obtendríamos la ecuación @ref. Esto motiva que podemos agregar controles a esta especificación utilizando la ecuación @ref para obtener:

$$Y_i = \beta_0 + \beta_1 \, T_i + X_i'\beta + U_i$$

Siendo que T_i es independiente y no está relacionada a X_i, como mostramos en la tabla de balance, esto no debería afectar de forma importante el valor estimado de β_1, pero podría mejorar su eficiencia si es que los controles ayudan a explicar de forma importante la variable dependiente Y_i. Este es el típico argumento de uso de controles en RCTs para incrementar la eficiencia del estimador del ATE.

Es importante cuidar no utilizar cualquier variable como control. En particular, variables que pudieran representar *resultados intermedios* no deberían emplearse como controles ya que podrían sesgar la estimación. Imagínen que tienen una variable K_i que fue afectada por el tratamiento y representa un resultado intermedio. Para ver el efecto que el tratamiento tuvo sobre dicha variable podríamos estimar:

$$K_i = \gamma_0 + \gamma_1 T_i + V_i$$

Ahora veamos que sucedería si agregamos a K_i como control en la estimación de los efectos de T_i sobre Y_i:

$$\begin{aligned} Y_i &= \alpha_0 + \alpha_1 \, T_i + \alpha_2 K_i + U_i \\ &= \alpha_0 + \alpha_1 \, T_i + \alpha_2(\gamma_0 + \gamma_1 T_i + V_i) + U_i \\ &= (\alpha_0 + \alpha_2\gamma_0) + (\alpha_1 + \alpha_2\gamma_1) \, T_i + (U_i + \alpha_2 V_i) \\ &= \beta_0 + \beta_1 \, T_i + W_i \end{aligned}$$

Con esto podemos ver que el coeficiente de T_i que obtendríamos en el MCO simple (β_1) no es el mismo que obtendríamos como coeficiente de T_i si agregamos

a K_i como control (α_1). En particular, $\beta_1 = \alpha_1 + \alpha_2\gamma_1$. Si el resultado intermedio tiene un efecto positivo sobre Y_i ($\alpha_2 > 0$) y el tratamiento tuvo un efecto positivo sobre el resultado intermedio ($\gamma_1 > 0$), entonces tendríamos que al agregar K_i como control estaríamos subestimando el efecto verdadero β_1 ya que $\alpha_1 < \beta_1$.

7.2.2. Estimación de Neyman

El coeficiente β_1 de la estimación simple de MCO representa la diferencia promedio de la variable dependiente entre el grupo de control y tratamiento. Neyman en 1935 propuso la aplicación de experimentos en agricultura y sugirió la estimación del efecto promedio de tratamiento (ATE) usando una diferencia de medias simple, que es equivalente a lo que el estimador de MCO produce:

$$\hat{\tau} = \overline{Y^1} - \overline{Y^0} \tag{7.6}$$

donde $\overline{Y^1}$ y $\overline{Y^0}$ corresponden a los promedios simples de Y_i para los individuos asignados a los grupos de tratamiento y control, respectivamente:

$$\overline{Y^1} = \frac{1}{N_T} \sum_{i|T_i=1} Y_i$$
$$\overline{Y^0} = \frac{1}{N_C} \sum_{i|T_i=0} Y_i$$

Para demostrar que $\hat{\tau}$ es un estimador insesgado de τ, demostramos que es un estimador insesgado de τ en la muestra que elegimos. Si la muestra es representativa de la población, entonces querría decir que es un estimador insesgado de τ en la población.

Siguiendo a Athey e Imbens (2017) empezamos por definir el estadístico:

$$W_i = \left(\frac{T_i \; Y_i}{N_T/N} - \frac{(1 - T_i) \; Y_i}{N_C/N} \right) \tag{7.7}$$

La motivación para definir este estadístico es que el promedio de W_i es igual a $\hat{\tau}$:

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} W_i &= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{T_i \; Y_i}{N_T/N} \right) - \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(1 - T_i) \; Y_i}{N_C/N} \right) \\
&= \left(\frac{1}{N(N_T/N)} \sum_{i=1}^{N} T_i \; Y_i \right) - \left(\frac{1}{N(N_C/N)} \sum_{i=1}^{N} (1 - T_i) \; Y_i \right) \\
&= \frac{1}{N_T} \sum_{i|T_i=1} Y_i - \frac{1}{N_C} \sum_{i|T_i=0} Y_i \\
&= \overline{Y^1} - \overline{Y^0}
\end{aligned}$$

Sabemos que si la asignación fue aleatoria y utilizando la definición (7.2.1), $Y_i = Y_I^T$ si $T_i = 1$ y $Y_i = Y_I^C$ si $T_i = 0$. Sustituyendo esto en la definición de W_i obtenemos:

$$W_i = \left(\frac{T_i \; Y_i^T}{N_T/N} - \frac{(1 - T_i) \; Y_i^C}{N_C/N} \right)$$

Al hacer el paso anterior implicitamente estamos diciendo que toda la incertidumbre está en la aleatoriedad de la asignación de T_i. Si al azar se elige $T_i = 1$ ($T_i = 0$), entonces el individuo tendrá como resultado Y_i^T (Y_i^C). Si calculamos el valor esperado de W_i obtenemos:

$$\begin{aligned} E(W_i) &= E\left(\frac{T_i \; Y_i^T}{N_T/N} \right) - E\left(\frac{(1 - T_i) \; Y_i^C}{N_C/N} \right) \\ &= \frac{Y_i^T}{N_T/N} E(T_i) - \frac{Y_i^C}{N_C/N} E(1 - T_i) \end{aligned}$$

Utilizando el hecho de que en el experimento aleatorizado por diseño se elegirán N_T unidades para el tratamiento y N_C para el control:

$$\begin{aligned} E(W_i) &= \frac{Y_i^T}{N_T/N} E(T_i) - \frac{Y_i^C}{N_C/N} E(1 - T_i) \\ &= \frac{Y_i^T}{N_T/N} \cdot (N_T/N) - \frac{Y_i^C}{N_C/N} \cdot (N_C/N) \\ &= Y_i^T - Y_i^C \end{aligned}$$

Utilizamos este resultado en el cálculo del valor estimado del estimador propuesto por Neyman para concluir nuestra demostración de que es un estimador insesgado del efecto promedio de tratamiento para nuestra muestra:

$$\begin{aligned} E(\hat{\tau}) &= E(\overline{Y^0} - \overline{Y^1}) \\ &= E\left(\frac{1}{N} \sum_{i=1}^{N} \; W_i \right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \; E(W_i) \\ &= \frac{1}{N} \sum_{i=1}^{N} \; (Y_i^T - Y_i^C) \end{aligned}$$

Neyman tambien estaba interesado en estimar un intervalo de confianza y, por tanto, en la varianza del estimador. Esto involucra una demostración mas complicada ya que la asignación del tratamiento donde se decide que N_T individuos recibiran el tratamiento, en vez de asignar aleatoriamente a cada individuo de

forma independiente, hace que la varianza sea mas compleja. Como veremos mas adelante, en los cálculos de poder estadístico, poder controlar la proporción de individuos en tratamiento y control trae beneficios en la eficiencia del estimador. Sin embargo, si fijamos el número de individuos que recibirán tratamiento eso afecta la covarianza entre las observaciones, ya que si un individuo fue asignado a tratamiento eso afecta la probabilidad de que los individuos subsecuentes sean asignados a tratamiento.[5]

Para derivar una parte de la varianza veamos que:

$$\begin{aligned} Var(\hat{\tau}) &= Var(\overline{Y^0} - \overline{Y^1}) \\ &= Var(\overline{Y^0}) + Var(\overline{Y^1}) - 2\ Cov(\overline{Y^0}, \overline{Y^1}) \\ &= \frac{Var(Y_i^C)}{N_C} + \frac{Var(Y_i^T)}{N_T} - \frac{S_{01}^2}{N} \end{aligned}$$

Los primeros dos términos se pueden estimar simplemente utilizando los valores observados de Y_i^T y Y_i^C:

$$S_0^2 = Var(Y_i^C) = \frac{1}{N_C - 1} \sum_{i|T_i=0} (Y_i - \overline{Y^0})^2$$

$$S_1^2 = Var(Y_i^T) = \frac{1}{N_T - 1} \sum_{i|T_i=1} (Y_i - \overline{Y^1})^2$$

El último término S_{01}^2 surge del hecho de que si queremos fijar el número de observaciones que recibirán tratamiento (N_T), la asignación del tratamiento ya no es independiente entre las observaciones. Para ilustrar esto piensen que la primera observación en ser aleatorizada tiene una probabilidad $\frac{N_T}{N}$ de ser tratado. Supongamos que es asignado al tratamiento. Esto querría decir que la segunda undiad tendrá una probabilidad de $\frac{N_T - 1}{N-1}$ de ser asignado al tratamiento. El componente de la covarianza no puede estimarse ya que sería observar simultaneamente a Y_i^C y Y_i^T para algunos individuos. Athey e Imbens (2017) demuestran que este término es positivo;[6] si lo ignoramos tendremos una *sobre-estimación* de la varianza. Por lo tanto, Neyman sugiere utilizar como varianza conservadora el siguiente término:

$$Var(\hat{\tau}) = \frac{S_0^2}{N_C} + \frac{S_1^2}{N_T} \tag{7.8}$$

Este estimador de la varianza es muy cercano al resultado de asumir heterocedasticidad en la estimación de MCO. La varianza heterocedástica es ligeramente

[5] Imaginense una situación donde en una tómbola ha N_T bolitas rojas de tratamiento y N_C bolitas azules de control. El primer individuo en formarse y tomar una bolita tiene una probabilidad N_T/N de sacar una bolita roja de tratamiento. Sin embargo, si este individuo efectivamente saca una bolita roja de tratamiento, el segundo individuo ahora tendrá una probabilidad $(N_T - 1)/(N - 1)$ de sacar una bolita roja de tratamiento.

[6] Muestran que este término es: $S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i^T - Y_i^C - \tau)^2$

menor (indistinguible en muestras muy grandes) ya que para la estimación de S_0^2 y S_1^2 no utiliza el estimador insesgado de la varianza muestral y en cambio solo divide entre N_C y N_T en vez de $(N_C - 1)$ y $(N_T - 1)$. Por lo tanto, el error estándar heterocedástico de MCO se podría calcular a través de la varianza del estimador de Neyman haciendo un pequeño ajuste:

$$Var(\beta_1^{MCO}) = \frac{S_0^2}{N_C} \cdot \frac{N_C - 1}{N_C} + \frac{S_1^2}{N_T} \cdot \frac{N_T - 1}{N_T} \tag{7.9}$$

7.2.3. Estratificación

Una alternativa para buscar incrementar la eficiencia del estimador del ATE consiste en utilizar los datos basales para estratificar la muestra y asignar el tratamiento de forma aleatoria al interior de cada estrato (manteniendo constante al interior de cada estrato la proporción de individuos tratados $\frac{N_T}{N}$). Intuitivamente, esta estrategia es equivalente a utilizar controles en una estimación de MCO, ya que en la práctica eso involucra estimar el tratamiento *caeteris paribus* los controles. Sin embargo, en la estratificación el *caeteris paribus* se genera por diseño en la distribución aleatoria del tratamiento. Para elegir las variables de la estratificación se sugiere que sean controles que tengan un valor predictivo importante sobre la variable dependiente.

En el caso de estratificar, es posible calcular el efecto promedio de tratamiento al interior de cada estrato. Tomando un promedio ponderado de estos efectos promedio es posible calcular el ATE de la muestra, donde los ponderadores son el tamaño relativo del estrato. La ganancia de este tipo de diseños proviene de la varianza. La intuición es que dado que la aleatorización se realiza al interior del estrato, eso generá una independencia entre algunos individuos al momento de distribuir el tratamiento. Regresando al argumento en el cálculo de la varianza de *Neyman*, en esa ocasión si el primer individuo era asignado tratamiento el segundo individuo tendría una probabilidad $\frac{N_T - 1}{N - 1}$ de recibir tratamiento (lo cual rompía la independencia). Sin embargo, en el caso de la estratificación, si el segundo individuo se encuentra en un estrato distinto que el primer individuo, su probabilidad de recibir tratamiento dado que el primer individuo fue asignado a tratamiento, no se ve afectada (sigue siendo $\frac{N_t}{N}$). A este segundo individuo solo le afecta la asignación de tratamiento de otros individuos al interior de su mismo estrato. La ganacia en estratificación, sin embargo, es en valor esperado. Pudieran existir casos, particularmente si la vairable de estratificación tiene un bajo poder predictivo sobre la variable dependiente, donde en la práctica se obtenga una varianza mayor que la que se obtendría en una aleatorización sin estratificación. Sin embargo, este tipo de casos son lo suficientemente raros como para descartar el uso de la estratificación.

Por último, la estratificación por diseño además tiene beneficios sobre el balance de la muestra, haciendo mas improbable que la aleatorización resulte en una distribución no balanceada.

Imaginemos que al utilizar los datos basales formamos G estratos que denotaremos con el subíndice $g = \{1, 2, ..., G\}$. Siguiendo la idea del estimador de Neyman podemos imaginar que en cada estrato se asigna una proporción constante de $\frac{N_T}{N}$ individuos a tratamiento. Con esto tendremos en el estrato $N_{T,g} = N_g * \frac{N_t}{N}$ individuos de tratamiento y $N_{C,g} = N_g - N_{T,g}$ individuos de control, donde N_g representa el total de individuos en dicho estrato. Podemos calcular el efecto promedio de tratamiento al interior del estrato siguiendo la idea de la ecuación (7.6):

$$\widehat{\tau_g} = \overline{Y_g^1} - \overline{Y_g^0} \tag{7.10}$$

donde $\overline{Y_g^1}$ y $\overline{Y_g^0}$ corresponden a los promedios simples de Y_i para los individuos asignados a los grupos de tratamiento y control al interior del estrato (para identificar el estrato utilizamos una dummy $D_g = 1$ si el individuo i pertenece al estrato g y $D_g = 0$ si pertenece a otro estrato):

$$\overline{Y_g^1} = \frac{1}{N_{T,g}} \sum_{i|T_i=1, D_g=1} Y_i$$
$$\overline{Y_g^0} = \frac{1}{N_{C,g}} \sum_{i|T_i=0, D_g=1} Y_i$$

Utilizando los promedios de tratamiento para los distintos estratos podemos calcular el efecto promedio de tratamiento (ATE) con un promedio ponderado:

$$\hat{\tau} = \sum_{g=1}^{G} \widehat{\tau_g} \left(\frac{N_g}{N} \right) \tag{7.11}$$

Para la varianza primero necesitamos calcular las varianzas del estimador del efecto promedio para cada estrato. Al igual que (7.8) la varianza estará sobreestimada:

$$Var(\widehat{\tau_g}) = \frac{S_{0,g}^2}{N_{C,g}} + \frac{S_{1,g}^2}{N_{T,g}} \tag{7.12}$$

donde los componentes $S_{0,g}^2$ y $S_{1,g}^2$ se calculan como:

$$S_{0,g}^2 = \frac{1}{N_{C,g} - 1} \sum_{i|T_i=0, D_g=1} (Y_i - \overline{Y_g^0})^2$$
$$S_{1,g}^2 = \frac{1}{N_{T,g} - 1} \sum_{i|T_i=1, D_g=1} (Y_i - \overline{Y_g^1})^2$$

Dado que la elección de individuos entre distintos estratos es independiente, la varianza del estimador de ATE la podemos calcular con la suma ponderada de varianzas:

$$Var(\hat{\tau}) = \sum_{g=1}^{G} Var(\widehat{\tau_g}) \left(\frac{N_g}{N}\right)^2 \tag{7.13}$$

Una versión extrema de la estratificación se conoce como un diseño de **matched-pairs**. En este caso, llevamos la idea de la estratificación al límite buscando tener la mayor independencia posible entre las asignaciones a tratamiento y control. En este caso, con las variables de la línea basal se busca hacer *parejas* de individuos (lo cual en la práctica puede ser muy difícil), donde uno de los individuos será asignado a tratamiento y el otro a control. Podemos utilizar buena parte de la notación previa, imaginando que en este caso g es un subíndice para las G parejas. Al interior de cada pareja ya no calcularemos un promedio sino una diferencia simple:

$$\widehat{\tau_g} = Y_g^1 - Y_g^0 \tag{7.14}$$

El ATE en este caso se calcula como un promedio simple de las $\frac{n}{2}$ distintas $\hat{\tau}_g$:[7]

$$\hat{\tau} = \frac{1}{N/2} \sum_{g=1}^{G} (\hat{\tau}_g) \tag{7.15}$$

La varianza presenta una dificultad distinta dado que no será posible calcular $S_{0,g}^2$ y $S_{1,g}^2$ siguiendo las fórmulas previas porque al interior de una pareja no hay variación (recuerden que solo hay una observación de tratamiento y una de control al interior del estrato). Por lo tanto, se propone como estimador de la varianza de matched-pairs:

$$Var(\hat{\tau}) = \frac{1}{N/2} \cdot \left(\frac{1}{\frac{N}{2} - 1} \sum_{g=1}^{G} (\hat{\tau}_g - \hat{\tau})^2 \right) \tag{7.16}$$

La intuición de esta varianza es la usual de los promedios. Al interior del paréntesis estamos calculando la varianza muestral de las distintas τ_g y al exterior del paréntesis dividimos esta varianza sobre el número de observaciones que se usan en este promedio ($\hat{\tau}$).

7.2.4. FETs: Fisher Exact Test

Ronald Fisher propuso un análisis basado en el planteamiento de hipótesis nulas específcas a nivel individual (*sharp null hypothesis*). Este método se conoce como las **Pruebas Exactas de Significancia de Fisher** (FETs por sus siglas en

[7] No es necesario hacer un promedio ponderado ya que todos los estratos (parejas) tienen el mismo número de observaciones: $N_g = 2$ para toda g

inglés). La idea es que, bajo la hipótesis nula, los resultados potenciales pueden ser observados o inferidos. Pensemos, por ejemplo, que tenemos a un individuo que recibió control ($T_i = 0$) y, por lo tanto, observamos $Y_i = Y_i^C$. Tomemos la siguiente hipótesis nula:

$$H_0 : Y_i^T = Y_i^C + \tau_i \tag{7.17}$$

Con esta hipótesis nula, si asumimos un valor para τ_i[8], querría decir que con la Y_i observada podríamos inferir el valor de Y_i^T.

Para formar una intuición, imaginemos que tenemos un programa cuyo tratamiento no tiene ningún efecto ($\tau_i = 0$ para todo i). Esto querría decir que el tratamiento y el control tendrían valores de Y_i muy similares si fuesen observados simultáneamente. Entonces, si observamos a una persona bajo *control* y el tratamietno no tiene efecto, bien podríamos imaginar que su Y_i observada sería igual si hubiera recibido *tratamiento*.

El siguiente paso consiste en proponer algún estadístico $W(Y, T)$ que nos permita formar evidencia para evaluar la hipótesis nula. Siguiendo la idea de nuestras secciones anteriores, utilizaremos la diferencia de las medias de tratamiento y control. Sin embargo, esto no es necesario para los FETs, podríamos en cambio utilizar cualquier otro estadístico.[9] Nuestro estadístico propuesto es:

$$W(Y, T) = \left(\frac{1}{N_T} \sum_{i|T_i=1} Y_i^T \right) - \left(\frac{1}{N_C} \sum_{i|T_i=0} Y_i^C \right)$$

Siguiendo un espíritu similar a lo que vimos en *bootstrap*, este test realiza J simulaciones de la asignación al tratamiento. Es decir, simulará la selección de los N_T individuos de tratamiento varias veces, pese a que sean simulaciones *falsas* en el sentido de que no coincidirán con la asignación que efectivamente se realizó. Denotaremos a cada asignación simulada T^j, que es un vector de dimensión N con N_T ($T_i^j = 1$) y N_C ($T_i^j = 0$). Mientras J sea mayor, nuestra precisión del valor-p aumentará, asi que se sugiere hacer diversas simulaciones. Para cada paso calcularemos el estadístico $W(Y, T)$:

$$W^j(Y, T) = \left(\frac{1}{N_T} \sum_{i|T_i^j=1} Y_i^T \right) - \left(\frac{1}{N_C} \sum_{i|T_i^j=0} Y_i^C \right)$$

Por último, agregamos a todos los resultados de las simulaciones el valor del estadístico observado con la asignación verdadera. Con todos estos valores calculamos un valor-p contando la cantidad de simulaciones que generan un valor del estadístico más extremo al observado con la asignación del tratamiento que verdaderamente sucedió ($W^{obs}(Y, T)$):

$$p\ value = \frac{1}{J} \sum_{j=1}^{J} 1\{\ |W^j(Y, T)| \geq |W^{obs}(Y, T)|\ \} \tag{7.18}$$

[8] típicamente utilizamos en la hipóteiss nula $\tau_i = 0$ para tener una hipótesis nula equivalente a asumir que el tratamiento no tiene efecto.

[9] Athey e Imbens (2017) sugieren, por ejemplo, el uso de *ranks* por ser menos sensibles a valores atípicos.

Para ilustrar esto veamos el siguiente ejemplo. Supongamos que nuestro tratamiento es una beca y nuestro outcome son las calificaciones finales. La tabla siguiente muestra 8 observaciones, donde las primeras 4 observaciones reciben tratamiento ($N_T = 4$) y las últimas 4 control ($N_C = 4$). Como la tabla muestra, para las primeras 4 observaciones $Y_i = Y_i^T$ y para las últimas cuatro $Y_i = Y_i^C$:

i	Y_i	T_i	Y_i^T	Y_i^C
1	9	1	9	.
2	9	1	9	.
3	10	1	10	.
4	8	1	8	.
5	7	0	.	7
6	5	0	.	5
7	6	0	.	6
8	8	0	.	8

Imaginemos que en nuestra primera simulación aleatoriamente obtenemos que las observaciones pares tienen tratamiento y las impares control. Siguiendo la estrategia de la tabla anterior podemos llenar 4 valores de Y_i^T y 4 de Y_i^C. Luego utilizamos la hipótesis nula (7.17) y un valor de $\tau_i = 1$ para ver si podemos rechazar la hipótesis de que las becas generan, en promedio, un aumento de un punto en la calificación. Con este supuesto llenamos los valores restantes de las columnas de Y_i^T y Y_i^C. Marcamos en negritas los valores contrafactuales que rellenamos con la hipótesis nula. Con esto, podremos calcular nuestro estadístico ($W^j(Y,T)$) para esta simulación.

i	Y_i	T_i^j	Y_i^T	Y_i^C
1	9	0	**9**	8
2	9	1	9	**8**
3	10	0	**10**	9
4	8	1	8	**7**
5	7	0	**8**	7
6	5	1	6	**5**
7	6	0	**7**	6
8	8	1	9	**8**

En esta simulación $W^j(Y,T) = 0{,}5$. Para simplificar nuestra ilustración supongamos que repetimos este ejercicio 10 veces, lo cual resulta en la siguiente tabla, donde cada renglón corresponde a una simulación distinta. En el primer renglón mostramos el resultado de la simulación que acabamos de realizar y en el último renglón indicamos la asignación que verdaderamente se realizó. En las columnas mostramos la asignación aleatoria de tratamiento para cada simulación y en la

última columna el valor del estadístico correspondiente.

j	T_1^j	T_2^j	T_3^j	T_4^j	T_5^j	T_6^j	T_7^j	T_8^j	W^j
1	0	1	0	1	0	1	0	1	0.5
2	0	1	1	1	0	1	0	0	1
3	1	0	1	0	1	0	0	1	2.5
4	1	0	0	0	1	1	1	0	-0.5
5	0	1	0	1	1	0	0	1	1.5
6	0	0	1	0	0	1	1	1	0.5
7	1	1	1	0	0	0	0	1	3
8	0	0	0	1	1	1	0	1	0
9	0	0	0	1	1	0	1	1	0.5
10	0	0	1	1	1	1	0	0	0.5
Real	1	1	1	1	0	0	0	0	2.5

Como podemos ver en este caso, solo las simulaciones $j = \{3, 7\}$ tienen valores iguales o más extremos que el estadístico observado $W^{obs}(Y, T) = 2{,}5$. Empleando el cálculo de (7.18) vemos que $p\ value = 0{,}2$. A los valores usuales no rechazaríamos la hipótesis nula. Cabe resaltar, por supuesto, que la cantidad de simulaciones y observaciones que tenemos en nuestro pequeño ejemplo son pocas y se hicieron solo con fines ilustrativos.

7.3. Atrición

La atrición es un problema que suele presentarse en el contexto de experimentos aleatorizados, especialmente en las ciencias sociales, donde el control que se tiene sobre la muestra para dar seguimiento es menor que en el trabajo de laboratorio. Entendemos a la atrición como la pérdida de observaciones que originalmente se encontraban en el diseño del experimento y la asignación del tratamiento. Es importante como primer paso documentar el grado de atrición y si dicha atrición fue diferencial entre tratamiento y control. Se recomienda empezar por documentarla: reportar el porcentjade de las observaciones originales que no pudo ser observada al final del experimento y dividirla netre tratamiento y control.

La atrición es un problema que suele ser grave ya que puede venir acompañada de:

- **Pérdida de eficiencia** en los estimadores (incremento en los errores estándar).
- **Problemas de validez externa** si los individuos que abandonan el experimento son distintos de los que se mantienen. Para esto suele hacerse una versión de la *tabla de balance* descrita previamente para comparar a los

individuos que se mantienen en la muestra con aquellos que salieron. Que haya diferencias significativas no quiere decir que tendremos un estimador sesgado del ATE, pero si el ATE es heterogéneo, podría querer decir que la población para la cual el estimador es representativo es distinta respecto a la población inicial que teníamos considerada en el estudio.

- **Problemas de validez interna** si los individuos que abandonan el experimento lo hacen de forma diferenciada entre tratamiento y control. Esta es la preocupación más grave y típicamente la detectamos re-haciendo la *tabla de balance* con la muestra que tenemos disponible para el análisis final, es decir, después de la pérdida de observaciones.

En el resto de esta subsección nos enfocamos en estrategias utilizadas para lograr llevar a cabo una estimación enfocada en la validez interna. Para facilitar la exposición, imaginémos que quisiéramos emplear un MCO para estimar el ATE, pero únicamente contamos con un subconjunto de las observaciones inicialmente consideradas en la aleatorización. Supondremos también que para **todas las observaciones** podemos observar un conjunto de *variables basales* X_i que no hayan sido afectadas por el tratamiento (ni por una anticipación del tratamiento). Como ya explicamos previamente, dichas variables pueden (o no) ser incluidas como controles en la estimación de MCO, sin esto afectar el sesgo de la estimación. Nuestro objetico sería estimar:

$$Y_i = \beta_0 + \beta_1 \ T_i + X_i'\beta_2 + U_i \tag{7.19}$$

Dada la asignación aleatoria, suponemos que podríamos estimar este modelo con MCO de forma insesgada si tuviéramos acceso a todos los datos $[E(T_iU_i|X_i) = E(T_iU_i) = E(T_i)E(U_i) = 0)]$. Sin embargo, el problema de atrición implica que no contamos con toda la información para algunos individuos (*attritors*): en particular tenemos la Y_i faltante. Definamos a s_i como una variable dummy que indica si para el individuo i tenemos los datos disponibles y, por lo tanto, lo podemos utilizar en la estimación.

Partiendo del modelo (7.19) podemos obtener:

$$s_iY_i = \beta_0 \ s_i + \beta_1 \ T_i + s_iX_i'\beta + s_iU_i \tag{7.20}$$

Nótese que estimar este modelo con todas las observaciones es equivalente a estimar (7.19) con la muestra restringida, es decir, con las observaciones para las cuales $s_i = 1$. Por lo tanto, estaremos interesados en determinar bajo qué condiciones podemos estimar (11.2) consistentemente. En este caso, estamos utilizando todas las observaciones, por lo tanto, aun no es un problema el sesgo muestral. Necesitamos entonces fijarnos en las condiciones de primer orden de la estimación para determinar si pudiera haber sesgo. En este caso, las condiciones de primer orden serían:

$$E[(s_i\ T_i|X_i)(s_i\ U_i|X_i)] = E[s_i\ T_i\ U_i|X_i] = 0 \tag{7.21}$$

porque $s_i^2 = s_i$.

$$E(s_i\ T_i\ U_i|X_i) = E(s_i|X_i)\ E(T_i\ U_i|X_i) = 0$$

7.3.1. Atrición aleatoria

En el caso en que la pérdida de observaciones fuese aleatoria, deberíamos poder observar que el balance en variables observables se mantiene entre el tratamiento y el control. Además, no debería representar un problema de validez externa dado que las observaciones perdidas deberían ser similares a las que se mantienen en la muestra. El único aspecto que podría afectar es la **perdida de eficiencia** en la estimación.

Para ver que esto no representa una amenaza en la identificación del ATE notemos que si la atrición es *aleatoria* o al menos independiente de variables observables y no observables se cumplirá que: $E(s_i\ T_i\ U_i|X_i) = E(s_i|X_i)E(T_i\ U_i|X_i) = 0$. Esto podría suceder si la pérdida de observaciones ocurrió por un evento exógeno, como pérdida de encuestas o imposibilidad de recopilar datos por un problemas climáticos o desastres naturales.

Esta aleatoriedad de la atrición se podría evaluar si modelamos la pérdida de observaciones con una estimación:

$$S_i^* = \delta_0 + \delta_1\ T_i + X_i'\ \delta_2 + Z_i'\ \delta_3 + V_i \tag{7.22}$$

donde S_i^* es una variable latente que sirva con una estimación de probit o logit a modelar $S_i = 1\{S_i^* \geq 0\}$. En este caso utilizamos las variables X_i y Z_i para diferenciar entre variables que podrían ser relevantes para explicar a la variable dependiente del análisis (X_i) de aquellas que no lo sean, pero si sean relevantes para modelar la atrición (Z_i). En el caso de una atrición aleatoria podríamos evaluar la hipótesis conjunta de todos los coeficientes, esperando no rechazar la siguiente hipótesis nula:

$$\begin{aligned} H_0 :& \delta_1 = 0 \\ & \delta_2 = 0 \\ & \delta_3 = 0 \\ H_1 :& \ e.o.c. \end{aligned}$$

7.3.2. Atrición no aleatoria

En los casos en los cuales la atrición no parece ser aleatoria, el diseño del experimento original tendrá problemas de *validez externa* e *interna*, como discutimos

previamente. En esta subsección enfocaremos la discusión hacia resolver el problema de validez interna para lograr obtener resultados no sesgados de nuestros estimadores. Algunos de los métodos que discutiremos pueden ser empleadas en contextos más amplios que el de experimentos aleatorizados.

Empecemos por considerar nuestra estimación de la especificación (7.19) restringiendo a los datos a los que tenemos acceso ($S_i = 1$), mismos que modelamos con nuestra especificación (7.22):

$$\begin{aligned} E(Y_i|T_i, X_i, Z_i, S_i = 1) &= \beta_0 + \beta_1\ T_i + X_i'\beta_2 + E(U_i|T_i, X_i, Z_i, S_i = 1) \\ &= \beta_0 + \beta_1\ T_i + X_i'\beta_2 + ... \\ &\quad + E(U_i|T_i, X_i, Z_i, V_i \geq -\delta_0 - \delta_1\ T_i - X_i'\delta_2 - Z_i'\delta_3) \end{aligned} \tag{7.23}$$

En este caso, el problema de sesgo se ve reflejado en que el último término no necesariamente es cero pese a que el tratamiento se haya asignado de forma aleatoria, ya que la pérdida de datos pudiera estar relacionado con T_i. Un primer paso para analizar la atrición y su relación con la asignación de tratamiento consiste en estimar (7.22) con y sin los controles (X_i) y analizar si el coeficiente de T_i es significativo. Si al controlar por X_i el coeficiente de δ_1 pierde significancia, esto sugeriría que tal vez podemos emplear variables observables para modelar la pérdida de observaciones y controlar el sesgo por atrición. De lo contrario, necesitaríamos utilizar algún supuesto de *restricción de exclusión* y modelar la atrición. Finalmente, además de las alternativas anteriores podemos llevar a cabo una estimación de cotas (**bounds**) a los efectos, bajo ciertas restricciones adicionales de validez externa.

7.3.2.1. Inverse Probability Weights

Una alternativa que frecuentemente se emplea como solución a la atrición no aleatoria consiste en llevar a cabo una reponderación de las observaciones disponibles después de la atrición. Estos métodos utilizan variables observables que no deben haber sido afectadas por el tratamiento (de preferencia recopiladas durante la línea basal). Siendo métodos que utilizan variables observables que pudieran afectar al tratamiento, nos referiremos a estas en nuestra notación a las variables X_i.

La intuición de estos modelos consiste en modificar el peso que le damos a cada una de las observaciones que tenemos disponibles para la estimación final (i.e. después de la atrición) con el propósito de obtener una muestra que *sea similar* a la muestra que teníamos originalmente. La *similitud* se define en términos de la distribución de las variables observables X_i. Para este propósito existen modelos paramétricos y no paramétricos.

El **Inverse Probability Weight** (IPW), es un método no paramétrico donde empleamos la estimación de la especificación (7.22) que llevamos a cabo con un *probit* o *logit*. Esta especificación hace que las probabilidades que ser observado sean función de X_i (en esta discusión omitimos el uso de las variables Z_i), por lo tanto, podemos corregir el cálculo de los promedios, dándole mayor peso a las observaciones que sean poco probables de observar dados sus valores de X_i. De aquí surge el nombre de *probabilidad inversa* e dichos pesos. La teoría que justifica el uso de estos pesos surge una aplicación de la ley de las esperanzas iteradas junto con la regla de Bayes.

Típicamente, cuando calculamos la media sumamos las Y_i y dividimos entre N. Todas las siguientes derivaciones las hacemos para el caso de las observaciones de tratamiento. El cálculo para el control es idéntico, pero condicionando en $T_i = 0$. Si les facilita para seguir los siguientes cálculos, pueden omitir la condicional de $(T_i = 1)$. En la exposición lo dejo explícito solo para recordarles que estos pasos corresponden al tratamiento. Desde la perspectiva de una integral, el valor esperado es:

$$\int_{T_i=1} y f(y|T_i = 1)dy$$

Esta sumatoria en un contexto de una base de datos implicaría sumar todas las Y_i de los individuos en el grupo de tratamiento y dividir entre N_T. Esto es equivalente a pensar que $f(y|T_i = 1) = \frac{1}{N_T}$ de forma uniforme. El problema de este cálculo es que no todas las observaciones están disponibles dado el problema de atrición, solo tenemos acceso a aquellas con $S_i = 1$. Con el *IPW* buscamos estimar $f(y|T_i = 1)$ utilizando información de un conjunto de variables basales X_i y las observaciones disponibles. Empezamos por considerar que podemos utilizar la *Ley de Eseranzas Iteradas*:

$$f(y|T_i = 1) = \int_{x,T_i=1} f(y, x|T_i = 1)dx$$

Para incorporar un cálculo que podamos estimar con el problema de atrición utilizamos la *regla de Bayes*:

$$\begin{aligned} g(y, x|T_i = 1, S_i = 1) &= \frac{Pr(S_i = 1|y, x, T_i = 1)}{Pr(S_i = 1|T_i = 1)} \; f(y, x|T_i = 1) \\ &= \frac{Pr(S_i = 1|x, T_i = 1)}{Pr(S_i = 1|T_i = 1)} \; f(y, x|T_i = 1) \\ &= \frac{f(y, x|T_i = 1)}{w(x, T_i = 1)} \end{aligned} \tag{7.24}$$

donde $g(.)$ representa una densidad conjunta condicional a que las variables son observables $(S_i = 1)$; la segunda igualdad resulta de que la probabilidad de atrición (condicional en X) es independiente de Y, es decir, la atrición se explica por completo con X; y la tercera igualdad sustituye $w(x, T_i = 1) =$

$\left(\frac{Pr(S_i=1|x,T_i=1)}{Pr(S_i=1|T_i=1)}\right)^{-1}$. El componente $w(x, T_i = 1)$ es el **ponderador** en el IPW. Utilizando la derivación de la regla de Bayes podemos sustituir en la densidad que nos interesaba estimar para obtener:

$$f(y|T_i = 1) = \int_{x,T_i=1,S_i=1} g(y, x|T_i = 1, S_i = 1)\ w(x, T_i = 1)dx$$

Si llevamos esto a la práctica $g(y, x|T_i = 1, S_i = 1) = \frac{1}{N_{TS}}$ en una sumatoria que esta empleando las observaciones de tratamiento después de la atrición. Agregando el ponderador podemos entonces estimar:

$$\overline{Y_s^1} = \frac{1}{N_{TS}} \sum_{i|T_i=1,S_i=1} w(X_i, T_i = 1)\ Y_i \tag{7.25}$$

donde $w(X_i, T_i = 1)$ la podemos estimar con un *probit* o *logit*. Podemos emplear una versión simplificada de (7.22) donde solo usamos nuestras variables X_i observables para modelar la atrición:

$$S_i^* = X_i'\ \delta_1 + \delta_2\ T_i + V_i$$

Sustituyendo el resultado de esta estimación para el cálculo de $w(X_i, T_i = 1)$ obtenemos:

$$w(X_i, T_i = 1) = \left(\frac{\Phi(X_i'\delta_1 + \delta_2)}{N_{TS}/N_T}\right)^{-1}$$

donde $\Phi(.)$ es la densidad acumuladad de la distribución normal o logística dependiendo de si utilizamos un probit o logit en la estimación de la atrición y N_{TS} es el número de observaciones de tratamiento después d ela atrición. Con esto, tenemos todos los elementos necesarios para estimar el valor promedio de tratamiento. La derivación del promedio para el control es similar y resulta en:

$$\overline{Y_s^0} = \frac{1}{N_{CS}} \sum_{i|T_i=0,S_i=1} w(X_i, T_i = 0)\ Y_i \tag{7.26}$$

donde los ponderadores corresponidentes son:

$$w(X_i, T_i = 0) = \left(\frac{\Phi(X_i'\delta_1)}{N_{CS}/N_C}\right)^{-1}$$

Asi pues obtenemos nuestro estimador del ATE con IPW:

$$\tau^{IPW} = \overline{Y_s^1} - \overline{Y_s^0}$$

Con este planteamiento podemos utilizar el estimador de Neyman o una regresión de mínimo cuadrados ponderados (*Weighted Least Squares*, WLS) donde el ponderador es precisamente $w(X_i, T_i)$.

7.3.2.2. Heckman

El modelo de *Heckman* se enfoca en emplear variables Z_i para modelar la atrición y estimar de forma insesgada el ATE. Una condición importante para el uso de este modelo es que es necesario identificar esta(s) variable(s) Z_i que cumpla(n) con la condición de explicar la atrición, pero no explicar a la variable dependiente Y_i en la especificación (7.19). A esto lo conocemos como la **restricción de exclusión** que necesitamos para este modelo. Una limitación con este supuesto es que no es posible evaluarlo directamente.

El método de *Heckman* es un estimador de máxima verosimilutud. Partimos de estimar con un *probit* o *logit* la especificación (7.22). Utilizando el supuesto que $(U_i, V_i) \perp \{T_i, X_i, Z_i\}$ y partiendo de (7.19) obtenemos:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1\ T_i + X_i'\beta_2 + U_i \\ E(Y_i|T_i, X_i, Z_i, V_i) &= \beta_0 + \beta_1\ T_i + X_i'\beta_2 + E(U_i|T_i, X_i, Z_i, V_i) \\ &= \beta_0 + \beta_1\ T_i + X_i'\beta_2 + E(U_i|V_i) \\ &= X_i'\beta + \rho V_i \end{aligned} \tag{7.27}$$

donde asumimos que $E(U_i|V_i) = \rho V_i$, lo cual surge del supuesto de que U_i y V_i son conjuntamente normales con media cero. Esta ecuación no puede ser estimada dado que V_i no es observada, pero podemos utilizarla como punto de partida para estimar $E(Y_i|T_i, X_i, Z_i, S_i)$:

$$E(Y_i|T_i, X_i, Z_i, S_i) = \beta_0 + \beta_1\ T_i + X_i'\beta_2 + \rho E(V_i|T_i, X_i, Z_i, S_i) \tag{7.28}$$

Dado que V_i tiene una distribución normal estándar, al igual que en el caso de Tobit, podemos mostrar que cuando $S_i = 1$[10]:

$$\begin{aligned} E(V_i|T_i, X_i, Z_i, S_i = 1) &= E(V_i|V_i \geq -\delta_0 - \delta_1 T_i - X_i'\delta_2 - Z_i'\delta_3) \\ &= \frac{\phi(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3)}{\Phi(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3)} \\ &= \lambda(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3) \end{aligned} \tag{7.29}$$

Sustituyendo este resultado en (11.7) obtenemos:

$$E(Y_i|T_i, X_i, Z_i, S_i = 1) = \beta_0 + \beta_1\ T_i + X_i'\beta_2 + \rho\lambda(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3) \tag{7.30}$$

Cabe recordar que asumimos que V_i se distribuye como una normal estándar. Este supuesto es clave para poder calcular para cada individuo $\lambda(\delta_0 + \delta_1 T_i +$

[10]Para obtener este resultado utilizamos el supuesto de que la distribución de V_i es normal y que estamos llevando a cabo una integral entre $-\delta_0 - \delta_1 T_i - X_i'\delta_2 - Z_i'\delta_3$ e ∞

$X_i'\delta_2 + Z_i'\delta_3$). Dado que V_i se distribuye como una normal estándar y la definición (7.22), tendremos que:

$$\begin{aligned} Pr(S_i = 1|T_i, X_i, Z_i) &= Pr(V_i < \delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3) \\ &= \Phi(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3) \end{aligned} \tag{7.31}$$

Por lo tanto, el procedimiento del modelo de *Heckman* consiste de los siguientes pasos:

1. Se estimará (11.10) utilizando el modelo probit. En esta estimación se utilizarán todas las observaciones (incluso aquellas que no se observan después de la atrición, i.e. aquellas para las cuales $S_i = 0$).
2. Se utilizarán los coeficientes de esta primera estimación para calcular $\lambda(\delta_0 + \delta_1 T_i + X_i'\delta_2 + Z_i'\delta_3)$ para cada individuo.
3. Se estimará la especificación (11.9). En esta estimación se utilizarán únicamente las observaciones con las observaciones disponible después de la atrición (i.e. aquellas para las cuales $S_i = 1$).

Esta última especificación generará estimadores insesgados de β_1. Puede además utilizarse esta estimación para evaluar si existía sesgo muestral. Para ello simplemente se evalúa si $\rho = 0$. En los casos en los cuales se rechaza la hipótesis y tenemos evidencia de que $\rho \neq 0$ tendríamos que la estimación de MCO con solo las observaciones que tienen $S_i = 1$ generaría estimadores sesgados de β_1 si es que el balance se hubiera perdido después de la atrición.

Para generar un ejemplo en `Stata` del uso del modelo de Heckman pueden emplear los siguientes comandos:

- `webuse womenwk`
- `sum wage education age children married`
- `gen si = (wage < .)`
- `probit si education age married children`
- `predict probit_Xb, xb`
- `gen mills = normalden(probit_Xb) / normal(probit_Xb)`
- `reg wage education age mills, r`
- `heckman wage education age, twostep select(education age married children) rhosigma first`

7.4. Asignación aleatoria

La forma mas sencilla de formar grupos comparables consiste en determinar el status de tratamiento o control de forma aleatoria. Siguiendo este procedimien-

to, es posible asegurarse que el estatus de tratamiento será independiente de cualquier variable no observable y podemos obtener un estimador insesgado de τ_1.

En la práctica algunos de los métodos de asignación aleatoria son:

1. Loterias por sobre-demanda. En situaciones en las cuales la demanda para algún programa es demasiado alta y no existen recursos suficientes para servir a toda la población, se considera que distribuir los recursos por loteria puede ser uno de los métodos éticamente más justo. Ejemplos: vouchers en Colombia, préstamos en Sudáfrica.

2. Expansión en fases. Algunos proyectos se expanden en distintas fases a lo largo del tiempo. Para determinar el órden de expansión puede considerarse justo determinarlo de forma aleatoria. Ejemplos: programa de desparasitación en Kenya. Sin embargo, este método puede tener problemas si los individuos anticipan la expansión y reaccionan y no pueden medirse efectos de largo plazo.

3. Asignación aleatoria dentro de grupos. En algunos casos se considera éticamente injusto que algunos grupos reciban beneficios de un programa y otros no. Algunas alternativas consisten en otorgar el beneficio de forma aleatoria a subgrupos dentro de cada grupo. Ejemplo: balsakis en India. Un problema de este método es el supuesto establecido en la sección anterior donde el estatus de otros individuos no te afecte.

4. Fomento al tratamiento. Por último, si negar los recursos puede determinarse no ético, una alternativa consiste en dar acceso general, pero a un grupo de individuos ofrecerles incentivos a participar en el tratamiento. Ejemplos: muestras de fertilizantes en Kenya, en E.U. enviar materiales gratis para estudiar para un examen.

Una prueba que se hace para demostrar que la asignación aleatoria funcionó y que los grupos resultantes son comparables consiste en utilizar información previa al experimento (información basal) y comparar ambos grupos. Generalmente los experimentos incluyen una tabla en la cual se comparan medias de variables contenidas en la base de datos. En teoría, no debería de poder rechazarse la hipótesis que las medias de cada variable para ambos grupos son iguales.

En la mayoría de las casos alguna de las variables resulta en diferencias significativas, solo por definición probabilística. En estos casos, un ejercicio común consiste en estimar el modelo (**??**) con y sin estas variables como control y comparar la estimación de τ_1 en ambos casos.

Un método popular para llevar a cabo la asignación aleatoria consiste en agrupar a los individuos por características similares formando grupos (matched-pairs). Esto se lleva a cabo utilizando la información basal o datos administrativos preexistentes. Una vez hecho esto se determina de manera aleatoria quien dentro del grupo o pair recibe el tratamiento. Este método además de garantizar tener grupos comparables, te permite disminuir la varianza de tu estimador.

7.5. Problemas de implementación

7.5.1. Participación parcial

En algunos casos decides de manera aleatoria los individuos que recibirán el tratamiento. En nuestro ejemplo distribuir los libros. Sin embargo, en diversos casos el tratamiento suele ser distinta de la *intención de tratamiento*. En nuestro ejemplo, si la pregunta relevante es ver como leer durante el verano mejora tus resultados, el hecho de leer es distinto de recibir libros, que es lo que impulsa la política. En este caso, supongamos que si el tratamiento (T_i) es leer durante el verano, necesitaremos crear una nueva variable que sea recibir libros (dummy Z_i si la escuela i fue aleatoriamente elegida para recibir libros). En este caso, dos cosas distintas serán el modelo (**??**) (utilizando T_i) y la siguiente estimación:

$$Y_i = \lambda_0 + \lambda_1 Z_i + V_i \tag{7.32}$$

Dado que la distribución aleatoria consisite en la distribución de libros, λ_1 será un estimador insesgado. Esto se conoce como *intención de tratamiento* (Intent to Treat, ITT). Sin embargo, la decisión de leerlos o no (T_i) no es aleatoria, por ende, estimar (**??**) utilizando MCO nos dará un estimador sesgado de τ_1. En concreto, decidir leer los libros puede ser una decisión basada en el interés en la lectura, por lo tanto, habrá sesgo por variables omitidas. En algunos casos, el estimador λ_1 puede tener un interés en si mismo. Si por ejemplo, quieres ver la eficiencia de este programa de distribución de libros. Sin embargo, si hay distintas maneras de incentivar la lectura y lo que te interesa realmente es estimar tau_1, tendremos que tomar supuestos específicos para obtener un estimador insesgado.

Para poder obtener un estimador insesgado de τ_1 necesitaremos llevar a cabo los siguientes supuestos:

1. Independencia. $\{Y_i^T, Y_i^C\}$ son independendientes de Z_i.
2. Monotonicidad. $T_i(1) \geq T_i(0)$. Es decir, no defiers. (Recuerden la tabla de always-takers, never-takers, defiers y compliers)

Bajo estos supuestos podemos utilizar el método de variables instrumentales para identificar el efecto del tratamiento (T_i) **para los compliers**. Esto se conoce como el *Efecto Promedio de Tratamiento Local* (o Local Average Treatment Effect, LATE). En clase demostraremos que IV identifica el ATE para una subpoblación (los compliers). El estimador de LATE resulta de aplicar la metodología de IV. Es decir, estimamos la forma reducida (7.32) y la primera etapa:

$$T_i = \eta_0 + \eta_1 Z_i + W_i \tag{7.33}$$

El estimador de LATE será $\tau_1 = \lambda_1/\eta_1$. Debemos recordar que Z_i debe cumplir los supuestos de variables instrumentales: (i) exogeneidad, que dado que es distribuido de manera aleatoria, no debe haber problema. Solo cabe señalar que Z_i

no debe influir directamente la variable dependiente Y_i, mas que a través de T_i; (ii) relevancia, que quiere decir que, en promedio, ser asignado aleatoriamente debe hacer mas probable que tomes el tratamiento.

El supuesto más fuerte en este caso es el de independencia. En muchos casos lo más correcto es tomar el ITT como el único estimador que es posible interpretarse de manera insesgada. Tomemos el ejemplo del programa de desparasitación en Kenya. Sea Z_i el indicador de estar en una escuela que aleatoriamente fue elegida para distribuir las pastillas de desparasitación y sea T_i una dummy que indica si el niño efectivamente tomó la pastilla de desparasitación. En este caso, estimando (7.32) podemos generar un estimador insesgado del ITT, es decir, el efecto del programa. Pero además podríamos estar tentado a calcular el LATE utilizando el IV para estimar el efecto promedio de la desparasitación sobre los niños compliers. Sin embargo, en este caso el supuesto de independencia no es correcto, ya que los niños en las escuelas de tratamiento que hayan decidido no tomar las pastillas (never-takers) se beneficiaron del tratamiento ya que a sus compañeros (algunos de ellos si las tomaron, los compliers) son en promedio menos propensos a propagar enfermeddes de parásitos. En este caso, únicamente es correcto estimar el ITT.

7.5.2. Externalidades

Hay una gran cantidad de intervenciones que provocan externalidades. Por ejemplo, la intervención que distribuyen pastillas desparasitantes beneficia no solo al grupo de tratamiento, sino también al control, ya que la probabilidad de contagio se reduce. Las externalidades pueden generarse por distribución de información, aprendizaje y reacción de los grupos de control. En estos casos, incluso el estimador ITT puede estar sesgado.

Si se espera que las externalidades surjan, el diseño del experimento puede incorporar este componente para estimar el alcance e importancia de dichas externalidades. Un ejemplo de esto lo llevan a cabo Duflo y Saez, que quiene ver el efecto de distribuir información sobre la selección de planes de retiro. Lo que hacen es hacer la asignación aleatoria en dos pasos: (i) elegir de manera aleatoria algunas instituciones donde se distribuiría la información; (ii) dentro de las instituciones seleccionadas, elegir de manera aleatoria individuos a quienes se distribuiría la información. En este caso, los autores estiman la externalidad producto de diseminación de información comparando los resultados de individuos no seleccionados en institutciones seleccionadas con individuos no seleccionados en instituciones no seleccionadas. El supuesto es que aquellos individuos no seleccionados en instituciones seleccionadas se beneficirían de la información que reciben de sus compañeros que si fueron seleccionados.

7.5.3. Pérdida de observaciones

La pérdida de observaciones puede ser problemática en el caso de los experimentos. En el caso en el que las observaciones perdidas son aleatorias, este problema implica un menor poder estadístico. Sin embargo, el principal problema se da cuando la pérdida de observaciones no es aleatoria. En particular, si aquellos que se benefician en menor medida del tratamiento deciden abandonar el experimento, llevar a cabo la estimación sin tomar en cuenta la pérdida de observaciones nos puede llevar a sobreestimar el efecto del experimento.

La pérdida de observaciones es costosa y en algunos casos muy difícil de evitar. Es una práctica común reportar en cada experimento que proporción de la muestra basal se ha perdido en las encuestas subsecuentes. En particular, la encuesta basal es útil para poder deteminar si las observaciones perdidas son similares a las que se mantienen en el experimento a través de la comparación de sus características observables antes del experimento.

En los casos en que se reconoce que la pérdida de observaciones no es aleatoria es recomendable analizar que tipo de sesgo puede generar este tipo de selección. Además existen algunas estrategias para tratar de control este tipo de selección. Una de ellas consiste en hacer un pareamiento (matching) de individuos del tratamiento y del control utilizando las características recabadas en la encuesta basal.

7.6. Críticas

- Efectos de equilibrio general. Los experimentos generalmente son de baja escala, lo cual no permite analizar los efectos de equilibrio general que dichas intervenciones implicarían. Estos efectos son importantes para poder evaluar las implicaciones de bineestar que conllevaría la aplicación de las intervenciones como política.

- Efectos en el comportamiento: *Hawthrone* y *John Henry*. La implementación de un experimento puede conllevar cambios en el comportamiento de los individuos. En particular, los individuos que reciben el tratamiento pueden simpatizar con el experimento, saber que son observados y por ende, esforzarse para que haya efectos positivos del experimento. Por otro lado, aquellos que son parte del grupo de control pueden sobreesforzarse para competir con el grupo de tratamiento. Estos comportamientos se conocen como efectos *Hawthrone* y *John Henry*, respectivamente y potencialmente no se hubieran dado en ausencia del experimento.

- Validez externa. Tres preocupaciones que generalmente surgen con los experimentos son:

- Si el experimento fue desarrollado con amplio nivel de cuidado será difícil asumir que asi será llevado a cabo si se generaliza como política pública.
- El hecho de que el experimento se haya llevado a cabo con una muestra específica genera preocupación de que el mismo resultado se daría con alguna muestra/población distinta.
- Qué tanto los restulados se deben a detalles específicos de la intervención. Es decir, qué tanto se puede aprender de intervenciones similares.

7.7. Experimentos naturales

La misma metodología cubierta en esta Nota puede seguirse en el caso de experimentos naturales. Los experimentos naturales se dan cuando algún evento exógeno forma dos grupos: tratamiento (T), es decir, aquellos individuos afectados por el evento y un grupo de control (C), aquellos no afectados. Si el evento es auténticamente exógeno, ambos grupos deben de ser comparables antes de la ocurrencia del evento. Esto puede verificarse, al igual que en un experimento social, haciendo una comparación de medias con datos recabados antes del evento.

Ejemplos de experimentos naturales incluyen:

- Vietnam Era Draft Lottery. Ser elegible para ser llamado al ejercito durnate la guerra de Vietnam se determinaba por un número que era asignado de manera aleatoria dependiendo del día de nacimiento de la persona. Esta elegibilidad fue utilizada como una fuente de variabilidad exógena para determinar el efecto de estar enlistado sobre los ingresos vitalicios de los individuos.
- La epidemia de la influenza española de 1918 se utilizó como una variación exógena para determinar la importancia del desarrollo en-utero. Se demostró que aquellos individuos que con alta probabilidad se encontraban durante su desarrollo en-utero durante el lapso de la pandemia tuvieron efectos negativos sobre años de educación, discapacidades físicas, ingreso y estatus socio-económico.
- El mes sagrado del Ramadan se utiliza para determminar la importancia de la salud fetal y la alimentación sobre el desarrollo de los individuos. Se comparan individuos de madres árabes que se desarrollaron en-utero durante estos meses contra individuos que se desarrollaron en otros meses y se encuentran efectos negativos sobre peso al nacer, mortalidad pre-natal y discapacidades en adultos.
- Ubicación de salones. Se comparan resultados de lectura de salones que por su ubicación eran afectados por el ruido del transporte público, respecto a salones no afectados. Se encuentra que la media de lectura en salones

expuestos al ruido tenían un rezago equivalente a 3-4 meses de aprendizje. En este caso, se argumenta que la localización de los salones es una fuente de variación exógena.

7.8. Tamaño de la muestra y poder estadístico

Llevar a cabo experimentos es costoso. Por esta razón, muchos experimentos vienen acompañados de cálculos de cuál tiene que ser el tamaño de una muestra para lograr identificar un efecto de manera significativa. Este ejercicio se conoce como *cálculo de poder estadístico*.

Para entender en que consiste el cálculo de poder estadístico podemos partir del planteamiento básico de nuestro modelo (ecuación (**??**)). En este caso, supongamos que el efecto verdadero del tratamiento es τ (por simplicidad asumiremos que $\tau > 0$, pero el caso de $\tau < 0$ es simétrico). El cálculo del poder estadístico querrá determinar de que tamaño tiene que ser una muestra (N) para que exista una alta probabilidad (poder estadístico) de obtener un estimador que nos permita rechazar la hipótesis nula: $H_0 : \tau = 0$ en favor de la alternativa $H_1 : \tau \neq 0$. Para empezar, para poder rechazar la hipótesis nula, tendremos que establecer un nivel de significancia α (o error tipo I). El error tipo I es la probabilidad de rechazar la hipótesis nula si dicha hipótesis es verdadera:

$$\alpha = Pr\left(\frac{\hat{\tau}}{se(\hat{\tau})} > t_\alpha | \tau = 0 \right) \tag{7.34}$$

Nuestro segundo componente es el *poder estadístico*. El *poder estadístico* (κ) es la probabilidad de obtener un estadístico-t mayor a t_α (para tener un estadístico significativo) dado que el valor verdadero de $\tau = \tau_0$ (como se mencionó previamente asumiremos que $\tau_0 > 0$, pero el caso de $\tau_0 < 0$ es simétrico):

$$\begin{aligned}
\kappa &= Pr\left(\frac{\hat{\tau}}{se(\hat{\tau})} > t_\alpha | \tau = \tau_0 \right) \\
&= Pr\left(\frac{\hat{\tau} - \tau_0 + \tau_0}{se(\hat{\tau})} > t_\alpha | \tau = \tau_0 \right) \\
&= Pr\left(\frac{\hat{\tau} - \tau_0}{se(\hat{\tau})} > t_\alpha - \frac{\tau_0}{se(\hat{\tau})} | \tau = \tau_0 \right) \\
&= 1 - \Phi\left(t_\alpha - \frac{\tau_0}{se(\hat{\tau})} \right)
\end{aligned} \tag{7.35}$$

Por lo tanto, para obtener un poder estadístico mayor o igual a κ:

$$\begin{aligned} 1-\kappa \leq & \Phi\left(t_\alpha - \frac{\tau_0}{se(\hat{\tau})}\right) \\ t_\alpha - \frac{\tau_0}{se(\hat{\tau})} \leq & -t_{1-\kappa} \end{aligned} \tag{7.36}$$

Para simplificar, asumamos errores homocedásticos en la estimación de (**??**) con un tamaño de muestra N y N_T individuos recibiendo el tratamiento, donde $N_T = N \cdot P$ (es decir $100 \cdot P\,\%$ de los individuos forman parte del tratamiento). En este caso el error estándar del estimador será:

$$se(\hat{\tau}) = \sqrt{\frac{1}{P \cdot (1-P)} \cdot \frac{\sigma^2}{N}} \tag{7.37}$$

Por lo tanto, para los valores α, κ, P y N podremos obtener un estimador significativo si:

$$\begin{aligned} \tau_0 &\geq (t_\alpha + t_{1-\kappa}) \cdot se(\hat{\tau}) \\ EMD(\tau_0) &= (t_\alpha + t_{1-\kappa}) \cdot se(\hat{\tau}) \end{aligned} \tag{7.38}$$

donde EMD es el efecto mínimo detectable, es decir, el valor mínimo de τ_0 para el cual podremos obtener un estimador significativo. - En cuanto al tamaño de muestra N, cabe resaltar que conforme mayor es N, menor es $se(\hat{\tau})$, por lo tanto, el EMD es menor. - En cuanto mayor es el poder estadístico (κ), mayor es $t_{1-\kappa}$ y por lo tanto, el EMD es mayor. - Al igual, conforme menor es el error tipo I (α), mayor será t_α y el EMD es mayor. En casos en los cuales la asignación se haga por grupos (e.g. escuelas), será importante permitir correlación en los errores de individuos en una misma escuela. Para esto se tendrán que asumir errores tipo cluster o el modelo de efectos fijos. En este caso, entre mayor sea la proporción de la varianza explicada por la correlación dentro de un grupo (cluster) (o el intra-class correlation) mayor es el EMD.

Capítulo 8

Variables Instrumentales

Como describimos en la sección 3, uno de los principales problemas de validez interna para los estimadores de MCO es el sesgo por variables omitidas. Esto provoca que los resultados de las estimaciones de MCO no puedan ser interpretadas de manera causal en la mayoría de las ocasiones. El sesgo por varibles omitidas se genera porque: (i) la variable de interés (X_1) esta correlacionada con alguna variable no observada o no incluida dentro de la estimación, y (ii) porque dicha variable no incluida está correlacionada con la variable dependiente. La primera condición implica que uno de nuestros supuestos utilizados para estimar el modelo de MCO ($E(X_{1i}U_i) = 0$) no sea un supuesto valido, ya que la variable omitida implícitamente forma parte del error de la estimación (U_i).

El método de *variables instrumentales* es una alternativa para estimar el efecto de dicha variable de interés (X_1) sobre la variable dependiente. Intuitivamente, este método consiste en encontrar un instrumento (Z) que juegue el rol de la variable de interés (X_1) sin tener el problema que dicha variable de interés tiene.

8.1. Planteamiento

Empecemos por recordar el sesgo por variables omitidas. Supongamos que queremos estimar el efecto de la educación sobre los ingresos. Para llevar a cabo esto empezamos por estimar un modelo de mínimos cuadrados ordinarios donde nuestra variable dependiente es el ingreso mensual (Ing_i)[1] y nos interesa ver el efecto de los años de escolaridad ($Educ_i$):

$$Ing_i = \alpha_0 + \alpha_1 Educ_i + U_i \tag{8.1}$$

[1]Podriamos usar el logaritmo del ingreso también, pero para simplificar la exposición utilizamos solo ingreso.

Un problema con esta estimación es que los años de educación pueden estar sesgados por omitir en esta estimación variables como educación de los padres, habilidad del individuo, mejores redes sociales, etc. Tomemos el ejemplo de habilidad. Si agregamos esta variable a nuestro modelo tendríamos:

$$Ing_i = \beta_0 + \beta_1 Educ_i + \beta_2 Habil_i + V_i \tag{8.2}$$

Y nuestro sesgo por variables omitidas estaría descrito por $\beta_2\gamma_1$ donde γ_1 tendrá el mismo signo que la correlación entre educación y habilidad. El problema en la estimación de (8.1) es que las variables omitidas implicitamente se encontraba en el error (U_i) y nuestra variable de interés ($Educ_i$) está correlacionada con ellas. Esto implica que se viola el supuesto de MCO $E(Educ_i U_i) = 0$.

Una altenativa para estimar de manera consistente α_1 utilizando el modelo (8.1) consiste en utilizar el método de variables instrumentales. Este método consiste en identificar un instrumento (Z_i). Utilizando este instrumento y nuestro modelo (8.1) podemos calcular:

$$\begin{aligned} Cov(Ing_i, Z_i) &= Cov(\alpha_0 + \alpha_1 Educ_i + U_i, Z_i) \\ &= \alpha_1 Cov(Educ_i, Z_i) - Cov(U_i, Z_i) \end{aligned} \tag{8.3}$$

Por lo tanto obtenemos[2]:

$$\alpha_1 = \frac{Cov(Ing_i, Z_i)}{Cov(Educ_i, Z_i)} + \frac{Cov(U_i, Z_i)}{Cov(Educ_i, Z_i)} \tag{8.4}$$

Para tener un estimador consistente de α_1 se deben cumplir dos condiciones, que son los supuestos fundamentales de los modelos de variables instrumentales:

1. **Relevancia** ($Cov(Educ_i, Z_i) \neq 0$). Intuitivamente, este supuesto implica que dado que queremos utilizar al instrumento (Z_i) para representar a nuestra variable de interés ($Educ_i$), dichas variables deben estar fuertemente correlacionadas. Una manera de evaluar si está condiciónes se satisface es llevar a cabo una regresión de la variable de interés ($Educ_i$) contra el instrumento (Z_i):

   $$Educ_i = \eta_0 + \eta_1 Z_i + U_i \tag{8.5}$$

   En la literatura se sugiere que para tener un buen instrumento, el estadístico F que resulte de llevar a cabo la siguiente prueba de hipótesis[3] debe ser mayor a 10:

   $$H_0 : \eta_1 = 0$$

[2] Nótese que la generalización de este modelo consiste en sustituir Ing_i con Y_i y $Educ_i$ con X_{1i}}

[3] Recuerden que cuando se evalúa la hipótesis de un solo coeficiente, el cuadrado del estadístico t es igual que el estadístico F. Establecemos esta prueba en términos del estadístico F, ya que como veremos más adelante en la Nota, puede ser que tengamos más de un instrumento

$$H_1 : \eta_1 \neq 0$$

2. **Exogeneidad o restricciones de exclusión** ($Cov(U_i, Z_i) = 0$). Exogeneidad implica que nuestro instrumento no está correlacionado con el error (U_i), que es lo que causaba el problema de sesgo por variables omitidas. Cabe recordar que U_i incluye todas aquellas variables que no incluimos en el modelo (8.1), tales como educación de los padres, habilidad, redes sociales, etc. Generalmente, el supuesto de exogeneidad es el más difícil de justificar y en términos del modelo no se puede evaluar directamente, a menos que se cuenten con más instrumentos que variables endógenas (es decir, aquellas, cuyo estimador está sesgado). Implicitamente, este supuesto además implica que el instrumento no debe de afectar directamente a la variable dependiente (Ing_i). El único efecto que identificará el modelo es el efecto indirecto de la variable de interés que estamos instrumentando ($Educ_i$)[4].

Angrist y Krueger (1991) sugieren como posible instrumento en este caso la fecha de nacimiento de la persona. Utilizando la fecha de nacimiento identificaron a aquellos que nacen en diferentes trimestres del año. La motivación para su instrumento se basa en la idea de, que de acuerdo a las leyes vigentes en E.U., una persona es requerida a estudiar hasta el momento en que cumple 16 años. Sin embargo, las generaciones escolares conjuntan a los niños nacidos entre Agosto y Julio del siguiente año. Por lo tanto, una persona que cumple años en enero ya podrá trabajar y aun no habrá completado el año escolar, mientras que una persona que cumple años en julio ya habrá terminado el grado escolar en el momento en que puede empezar a trabajar. Por lo tanto, es muy posible que la fecha de nacimiento influya sobre los años de escolaridad completados de un individuo. Esto puede ser verificado como describimos en el inciso de relevancia. El mayor reto consiste en argumentar la exogeneidad. Puede argumentarse que la fecha de nacimiento no influye los ingresos del individuo, y que no está relacionado con variables como acceso a transporte, habilidad y redes sociales. El único problema potencial es que los padres reconozcan esto y padres de familia más sofisticados decidan tener hijos de manera tal que nazcan cerca del final del ciclo escolar (pero como pueden darse cuenta es un argumento mas difícil de establecer).

Si asumimos que la fecha de nacimiento es un instrumento satisfactorio, podríamos utilizar dummies de haber nacido en distintos trimetres para estimar el efecto de la edcuación sobre el ingreso. Para ilustrar esto tomaremos solo una dummy y posteriormente veremos cómo ampliarlo a más instrumentos. Sea $Q1_i$ una dummy que indica si el individuo nació en el ultimo triemstre del año. Utilizando este instrumento debemos estimar las siguientes dos ecuaciones. Estas ecuaciones se conocen como *ecuaciones de forma reducida* cuando solo incluyen variables exógenas como regresores:

[4]De no cumplirse esta condición, no será posible distinguir que tanto de la $Cov(Ing_i, Z_i)$ se debe al efecto directo de Z_i sobre Ing_i y que tanto por el efecto a través de $Educ_i$.

$$\begin{aligned} Ing_i &= \gamma_0 + \gamma_1 Q1_i + \upsilon_i \\ Educ_i &= \eta_0 + \eta_1 Q1_i + \nu_i \end{aligned} \tag{8.6}$$

En este caso:

$$\hat{\alpha}_1 = \hat{\gamma}_1 / \hat{\eta}_1 \tag{8.7}$$

8.2. Agregar controles

El modelo de variables instrumentales puede incluir otras variables de control. Para llevar a cabo la estimación veamos como desarrollar el caso generalizado. Supongamos que tenemos k variables que queremos incluir en el modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + U_i \tag{8.8}$$

Supongamos que en este caso nos interesa estimar el efecto causal de X_1 sobre Y. Estimar el modelo (8.8) utilizando MCO genera un estimador sesgado de β_1 por haber sesgo por variables omitidas. Por lo tanto, incluimos un instrumento (Z) para X_1 que cumpla con las condiciones antes descritas. Si estimamos un modelo de forma reducida para estimar X_1 utilizando nuestro instrumento y las demás variables del modelo tendremos:

$$X_{1i} = \eta_0 + \eta_2 X_{2i} + \cdots + \eta_k X_{ki} + \phi Z_i + V_i \tag{8.9}$$

Sustituyendo (8.9) en (8.8) obtenemos:

$$Y_i = \beta_0 + \beta_1[\eta_0 + \eta_2 X_{2i} + \cdots + \eta_k X_{ki} + \phi Z_i + V_i] + \cdots + \beta_k X_{ki} + U_i \tag{8.10}$$

Reordenando los términos obtenemos:

$$Y_i = \psi_0 + \psi_2 X_{2i} + \cdots + \psi_k X_{ki} + \theta Z_i + W_i \tag{8.11}$$

donde:

$$\begin{aligned} \psi_j &= \beta_j + \beta_1 \eta_j \\ \theta &= \beta_1 \phi \\ W_i &= U_i + \beta_1 V_i \end{aligned}$$

Por lo tanto, para obtener estimadores insesgados utilizando MCO en (8.11) tendremos las siguientes condiciones de primer órden:

$$\sum_{i=1}^{N}(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \cdots - \hat{\psi}_k X_{ki} - \hat{\theta} Z_i) = 0 \tag{8.12}$$

$$\sum_{i=1}^{N} Z_i(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \cdots - \hat{\psi}_k X_{ki} - \hat{\theta} Z_i) = 0 \tag{8.13}$$

Y para $j = 2, \ldots, k$:

$$\sum_{i=1}^{N} X_{ji}(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \cdots - \hat{\psi}_k X_{ki} - \hat{\theta} Z_i) = 0 \tag{8.14}$$

Por lo tanto, para que los estimadores $\psi_0, \psi_2, \ldots \psi_k$ y γ sean insesgados, tendrá que cumplirse que la covarianza de $X_2, \ldots, X_k, Z$ con el error W del modelo (8.11) sea cero en cada caso. Esto se cumplirá si la covarianza de $X_2, \ldots, X_k, Z$ con los errores U y V de los modelos (8.8) y (8.11) son cero, respectivamente. En el caso de Z, este requisito es el supuesto de exogeneidad. Para el resto de los controles, este supuesto está imponiendo el requisito de exogeneidad. Recordemos que nuestro interés radica en obtener un estimador insesgado de β_1. Si removemos alguno de los controles porque nos preocupa que no cumple con los supuestos de exogeneidad, el requisito adicional que estamos imponiendo por no incluir dicho control es que el instrumento no deberá estar correlacionado con éste control, ya que el control pasará a formar parte del error U del modelo (8.8). Si el control no es relevante para explicar la variable dependiente, es mejor no incluirlo en la estimación.

A partir de estimar los modelo (8.8) y (8.9), se puede obtener un estimador de $\beta_0, \beta_1, \ldots, \beta_k$. Dadas las derivaciones previas tenemos que:

$$\beta_1 = \theta/\phi \tag{8.15}$$

Por lo tanto, únicamente dividimos el coeficiente que resulta de estimar (8.11) entre el que resulta de estimar (8.9).

8.3. Mínimos Cuadrados en 2 Etapas (Two-Stage Least Squares, 2SLS)

Supongamos ahora que queremos estimar el modelo (8.8) y que tenemos dos instrumentos ($Z1$, $Z2$) que cumplen con los supuestos de relevancia y las restricciones de exclusión. En este caso, podríamos llevar a cabo dos estimaciones

de las ecuaciones (8.9) y (8.11) para obtener dos valores estimados insesgados de β_1. En el caso del primer estimador utilizaríamos las condiciones de primer órden dadas por (8.12), (8.14) y:

$$\sum_{i=1}^{N} Z1_i(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \cdots - \hat{\psi}_k X_{ki} - \hat{\theta} Z1_i) = 0 \tag{8.16}$$

En el segundo caso utilizaríamos las condiciones de primer órden dadas por (8.12), (8.14) y:

$$\sum_{i=1}^{N} Z2_i(Y_i - \hat{\psi}_0 - \hat{\psi}_2 X_{2i} - \cdots - \hat{\psi}_k X_{ki} - \hat{\theta} Z2_i) = 0 \tag{8.17}$$

Sin embargo, existe una manera de agrupar la información de manera eficiente para producir un solo estimador. Para esto es útil el método de 2SLS. Este método lleva a cabo la estimación en dos etapas, donde la primera etapa combina los instrumentos de manera eficiente y la segunda utiliza el supuesto de exogeneidad para derivar coeficientes insesgados del modelo (8.8). (Cabe señalar que el métdo de 2SLS puede ser aplicado también en el caso que tengamos un solo instrumento y un estimador y resultará en el mismo coeficiente estimado que el derivado utilizando el método antes descrito)

8.3.1. Primera Etapa (First Stage)

La primera etapa esta relacionada con el supuesto de relevancia. Esta etapa consiste en utilizar los instrumentos para predecir el valor de la variable de interés (X_1). Este paso es el descrito en la ecuación de forma reducida (8.9), pero incluyendo todos los instrumentos válidos disponibles. Para llevar a cabo esto utilizamos un modelo de MCO:

$$X_{1i} = \eta_0 + \phi_1 Z1_i + \phi_2 Z2_i + \eta_2 X_{2i} + \cdots + \eta_k X_{ki} + V_i \tag{8.18}$$

En este caso, para evaluar si los instrumentos son relevantes, calculamos el estadístico F que se relaciona con la siguiente prueba de hipótesis:

$$\begin{aligned} H_0 :& \phi_1 = 0 \\ & \phi_2 = 0 \\ H_1 :& \phi_1 \neq 0 | \phi_2 \neq 0 \end{aligned}$$

Utilizando los resultados de esta estimación podemos predecir el valor de X_{1i} basado únicamente en la información que proporcionan los instrumentos y las variables exógenas:

$$\widehat{X}_{1i} = \hat{\eta}_0 + \hat{\phi}_1 Z1_i + \hat{\phi}_2 Z2_i + \hat{\eta}_2 X_{2i} + \cdots + \hat{\eta}_k X_{ki} \tag{8.19}$$

8.3.2. Segunda Etapa (Second Stage)

La segunda etapa consiste en utilizar el supuesto de exogeneidad de los instrumentos para derivar estimadores insesgados de los coeficientes del modelo (8.8). Para esto utilizaremos las condiciones de primer orden dadas por:

$$\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}) = 0 \tag{8.20}$$

$$\sum_{i=1}^{N} \widehat{X}_{1i}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}) = 0 \tag{8.21}$$

Y para $j = 2, \ldots, k$:

$$\sum_{i=1}^{N} X_{ji}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 \widehat{X}_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki}) = 0 \tag{8.22}$$

Esto resultará en los mismos estimadores que los obtenidos por estimar el siguiente modelo utilizando MCO:

$$Y_i = \beta_0 + \beta_1 \widehat{X}_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + W_i \tag{8.23}$$

Nótese que dado que $\widehat{X}_{1i}$ es una función de $Z1_i$ y $Z2_i$, no tendremos el problema de sesgo por variables omitidas y los coeficientes que resulten de esta estimación serán insesgados si $Z1_i$ y $Z2_i$ cumplen con los supuestos para ser instrumentos válidos.

8.4. Inferencia - Errores estándar

Tomando notación matricial, sea el modelo (8.8):

$$Y_i = X_i'\beta + U_i \tag{8.24}$$

donde $X_i = [\ 1 \quad X_{1i} \quad X_{2i} \quad \ldots \quad X_{ki}\ ]'$

Y sea:

$$X_i = Z_i'\Pi + W_i \tag{8.25}$$

donde $Z_i = [\ 1 \quad Z1_i \quad Z2_i \quad X_{2i} \quad ... \quad X_{ki}\ ]'$; la segunda columna de Π es la primera etapa: $\Pi(.,2) = [\ \eta_0 \quad \phi_1 \quad \phi_2 \quad \eta_2 \quad ... \quad \eta_k\ ]'$ y el resto de las columnas tienen un coeficiente de 1 en la columna correspondiente a las variables exógenas del modelo original (i.e. $X_2, ..., X_k$) dado que éstas variables están en X_i y en Z_i; y: $W_i = [\ 0 \quad V_i \quad 0 \quad ... \quad 0\ ]'$ donde V_i es el error de la primera etapa.

Las ecuaciones (8.24) y (8.25) en términos matriciales se convierten en:

$$Y = X\beta + U \tag{8.26}$$

donde Y es el vector de variables dependientes que tiene una dimensión de $(n \times 1)$; X es la matriz de variables independientes o regresores que tiene una dimensión de $(n \times k)$; β es un vector de coeficientes de dimensión $(k \times 1)$; y U es un vector de errores del modelo estructural con dimensión $(n \times 1)$.

$$X = Z\Pi + W \tag{8.27}$$

donde Z es una matriz que incluye los instrumentos y variables exógenas de X y tiene una dimensión $(n \times L)$ (L es el número de variables exógenas más el número de instrumentos); Π es la matriz descrita antes y tiene dimensión $(L \times k)$; y W es una matriz de $(n \times k)$ que conjunta a las W_i antes descritas ($W = [\ 0 \quad V \quad 0 \quad ... \quad 0\ ]'$).

Partiendo de (8.27) tenemos:

$$\begin{aligned} X &= Z\Pi + W \\ Z'X &= Z'Z\Pi + Z'W \end{aligned} \tag{8.28}$$

y bajo el supuesto de que $E(Z'W) = E(Z'V) = 0$ por exogeneidad de los instrumentos y las variables exógenas en la primera etapa:

$$\Pi = E(Z'Z)^{-1}E(Z'X) \tag{8.29}$$

Por lo tanto, el estimador será:

$$\widehat{\Pi} = (Z'Z)^{-1}Z'X \tag{8.30}$$

Para simplificar la notación subsecuente utilizaremos la matriz de proyección $P_Z = Z(Z'Z)^{-1}Z'$ que nos permite seguir la metodología descrita en el método de *2SLS*. Utilizando el resultado de la primera etapa generamos una matriz X^* que corresponde a la parte de X explicada por los instrumentos y las variables exógenas de X (i.e. $X^* = Z\Pi$). Esta matriz deja intactas las variables exógenas de X y sustituye la variable endógena con el valor predicho por la primera

etapa utilizando los instrumentos y las variables exógenas de X. La contraparte muestral de X^* será:

$$\widehat{X} = Z\widehat{\Pi} = P_Z X \tag{8.31}$$

Utilizando X^* en (8.26) obtenemos:

$$\begin{aligned} Y &= X\beta + U \\ X^{*\prime}Y &= X^{*\prime}X\beta + X^{*\prime}U \end{aligned} \tag{8.32}$$

Por lo tanto, si se cumple el supuesto de exogeneidad ($E(X^{*\prime}U) = \Pi' E(Z'U) = 0$):

$$\beta = E(X^{*\prime}X)^{-1}E(X^{*\prime}Y) \tag{8.33}$$

Y el estimador será[5]:

$$\begin{aligned} \hat{\beta} &= (\widehat{X}'X)^{-1}\widehat{X}'Y \\ &= (X'P_Z'X)^{-1}\widehat{X}'Y \\ &= (X'P_Z'P_ZX)^{-1}\widehat{X}'Y \\ &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y \end{aligned} \tag{8.34}$$

Con esto hemos demostrado que el estimador de β resulta de hacer una regresión de Y como variable dependiente y $\widehat{X}$ como variables independientes.

La derivación de los errores estándar bajo los supuestos de homocedasticidad y heterocedasticidad sigue los mismos pasos que los descritos en la sección 3 tan solo sustituyendo X por $\widehat{X}$. En este caso tendremos los siguientes estimadores y convergencias en probabilidad:

$$\hat{\alpha}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{X}_i\widehat{X}_i'\right)^{-1} \longrightarrow E(X_i^*X_i^{*\prime})^{-1} = \alpha_{IV} \tag{8.35}$$

$$\hat{\Sigma}_{IV} = \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{U}_i^2\widehat{X}_i\widehat{X}_i'\right)^{-1} \longrightarrow E(U_i^2X_i^*X_i^{*\prime})^{-1} = \Sigma_{IV} \tag{8.36}$$

donde $\widehat{U}_i = Y_i - X_i'\hat{\beta}$.

[5]Estos pasos asumen que la matriz P_Z es idempotente ($P_Z = P_Z P_Z$) y simétrica ($P_Z = P_Z'$).

Y de la misma forma que la sección 3 en el caso de muestras grandes (teoría sintótica) tendremos convergencia en distribución para el estimador de β:

$$\sqrt{N}(\hat{\beta} - \beta) \longrightarrow N(0, \alpha_{IV}\Sigma_{IV}\alpha'_{IV}) \tag{8.37}$$

8.5. Problemas de instrumentos débiles

Los principales problemas del método de variables instrumentales (además de lograr encontrar un instrumento que cumpla con los supuestos establecidos) son:

- **Sesgo.** A partir del resultado mostrado en (8.4) podemos ver que si el supuesto de exogeneidad falla (i.e. $Cov(U_i, Z_i) \neq 0$) y nuestro instrumento es débil (i.e. $Cov(Educ_i, Z_i)$ es pequeño) el sesgo que resultaría en el estimador podría ser peor que en el caso de MCO.
- **Errores estándar.** Los instrumentos débiles provocan que los errores estándar estimados del coeficiente sean grandes. Por lo tanto, el intervalo de confianza será amplio y la capacidad de determinar que un coeficiente es significativo será menor. Para una explicación de por qué los errores estándar son aumentan con instrumentos débiles se recomienda consultar (Wooldridge 2002, pp. 101-105).

Por último, una cualidad adicional que comúnmente se otorga al método de variables instrumentales es que evita el sesgo de atenuación causado por errores de medición como los discutidos en la sección 3.10.

Capítulo 9

Diferencias en Diferencias

El método de diferencias en diferencias (*Diff-in-Diff*) se ha vuelto un método muy popular para hacer inferencia causal en microeconometría aplicada. El planteamiento de este método requiere observar dos grupos de individuos (o entidades) en al menos dos momentos distintos del tiempo, siendo uno de esos dos grupos afectado por un cambio, cuyo efecto causal se pretende estimar. Ejemplo: si el propósito fuera evaluar el efecto de una reforma o cambio legislativo utilizando el método de diferencias en diferencias, sería necesario observar a un grupo de individuos afectados por la reforma y a otro no afectado, antes y después del cambio. Existe también otra alternativa para aplicar el método de diferencias en diferencias sin tener observaciones en dos momentos del tiempo. Ésta consiste en tomar dos subgrupos de individuos tanto en el grupo afectado y el no afectado por el cambio: individuos elegibles y no elegibles para ser afectados por el cambio. Ejemplo: supongamos que se implementa un programa de construcción de escuelas en un Estado que beneficia a niños indigenas. Podríamos aplicar el método de diferencias en diferencias si tenemos una base de datos con observaciones de niños indígenas y no indígenas en el Estado donde se aplico la reforma y en al menos otro Estado no afectado.

Es común pensar en aplicar el método de diferencias en diferencias cuando se tiene un *experimento natural*. Un *experimento natural* consiste en un evento exógeno que afecta ciertas variables económicas (en términos econométricos, alguna variable independiente). La exogeneidad del evento consiste en que éste debe afectar a la variable independiente en cuestión y no a otros variables independientes que pudieran afectar la variable dependiente que se analiza. Ejemplos de experimentos naturales: eventos naturales como sequías, proliferación de algún virus, desastres naturales, eventos históricos etc.

9.1. Planteamiento básico

Empezando por el caso más simple, supongamos que observamos a individuos que se dividen en dos grupos: $G_i = \{0, 1\}$, donde el grupo 1 es afectado por el cambio y el 0 no; y que pueden ser observados en dos momentos del tiempo $T_i = \{0, 1\}$, donde $T_i = 0$ es antes del cambio y $T_i = 1$ después. (Nota: No es necesario que la base de datos sea de panel, es decir, que sea **el mismo** individuo el que sea observado en ambos momentos del tiempo. únicamente se requiere el supuesto de que la muestra sea aleatoria) Para dichos individuos observamos una variable de interés: Y_i. Por lo tanto, cada observación estará caracterizada por tres variables (Y_i, G_i, T_i). Supongamos que queremos ver el impacto de una reforma laboral sobre horas trabajadas y suponemos que dicha reforma se puede aplicar en unos Estados si $(G_i = 1)$ y en otros no $(G_i = 0)$. Un individuo que trabaja 20 horas a la semana, observado antes de la reforma en un Estado donde si se aplica la reforma, tendrá las variables: $(20, 1, 0)$.

Para calcular el efecto de la reforma laboral utilizando el modelo de diferencias en diferencias primero tendremos que calcular las medias de cada grupo en ambos momentos del tiempo. Es decir, generaremos el siguiente estadístico:

$$\overline{Y}(j,t) = \frac{\sum_{i=1}^{N} 1(G_i = j) \cdot 1(T_i = t) \cdot Y_i}{\sum_{i=1}^{N} 1(G_i = j) \cdot 1(T_i = t)}, \text{ para } j = \{0,1\} \text{ y } t = \{0,1\}$$

En particular, $\overline{Y}(0, 1)$ es la media de horas trabajadas de los individuos observados después de la reforma en el Estado B.

El estimador de diferencias en diferencias es:

$$\tau = [\overline{Y}(1,1) - \overline{Y}(1,0)] - [\overline{Y}(0,1) - \overline{Y}(0,0)]$$

Intuitivamente, τ representa la diferencia del cambio promedio de horas trabajadas en el estado afectado respecto al no afectado. Como vimos en la sección 3, este mismo efecto lo podemos estimar con una regresión de MCO si utilizamos la siguiente especificación:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 T_i + \tau G_i T_i + U_i \tag{9.1}$$

En este caso, si τ es significativo, podremos decir que la reforma laboral tuvo un efecto estadísticamente significativo en las horas trabajadas.

9.2. Supuesto de tendencia paralela

En el planteamiento del modelo, el grupo no afectado por el cambio (el grupo $G_i = 0$ en nuestro ejemplo) funciona como un "grupo de control". Si no tuviéramos la información del grupo de control, nuestra única alternativa consistiría en comparar la variable dependiente Y_i antes y después del cambio en el grupo afectado. Sin embargo, en este caso, no sería posible distinguir si el cambio en el nivel medio de Y_i se debió a un cambio económico a través del tiempo o a el cambio específico que estamos analizando.

En el ejemplo de la reforma laboral, si no contaramos con el grupo de control y quisiéramos estimar el efecto de la reforma tomado el cambio en el promedio de horas trabajadas antes y después de la reforma laboral en el grupo $G_i = 1$, calcularíamos $[\overline{Y}(1,1) - \overline{Y}(1,0)]$. Sin embargo, utilizando este estimador sería imposible distinguir si el cambio en horas trabajadas se debe a la reforma laboral o a cualquier otro cambio sucedido entre el año 0 y 1 (e.g. una desaceleración económica, elecciones políticas, etc.).

Nuestro grupo de control nos permite controlar precisamente por esos efectos. Para ello, nuestro supuesto clave es que "en ausencia del cambio (reforma laboral), el grupo afectado por el cambio ($G_i = 1$), hubiera tenido una tendencia igual a la que tuvo el grupo de control ($G_i = 0$)". Este supuesto se conoce como el **supuesto de tendencia paralela**. En términos de nuestras ecuaciones, este supuesto genera un *contrafactual*. Es decir, asume que si no hubiese habido reforma, el cambio de horas trabajadas en el grupo $G_i = 1$ hubiera sido $[\overline{Y}(0,1) - \overline{Y}(0,0)]$. Como si hubo reforma, el cambio observado fue $[\overline{Y}(1,1) - \overline{Y}(1,0)]$. Por lo tanto, el efecto de la reforma es la diferencia entre dichos valores, es decir, τ.

Para que el supuesto de tendencias paralelas sea válido el grupo de control y el afectado deben ser lo más parecidos posibles antes del cambio. En función de nuestra especificación (9.1) esto debería querer decir que preferentemente β_1 debe ser no significativo. Otra manera de comprobar esto es hacer una comparación de medias con otras variables observables entre ambos grupos. Esta es una prueba sencilla y consiste en calcular los valores medios de distintas variables disponibles en nuestra base de datos antes del cambio (i.e. cuando $T_i = 0$). Si ambos grupos son similares antes del cambio o la reforma, la diferencia de medias no debe ser estadísticamente significativo o distinto a cero en la mayoría de las variables.

Otra posibilidad para hacer el calculo del efecto de diferencias en diferencias consiste en utilizar el modelo de *efectos fijos*. Supongamos que se tiene información para todos los estados j de algún país y que la reforma se aplicó en un subconjunto de esos estados. Sea R_{jt} una dummy que indica si en el año t y el estado j, la ley ya se encontraba vigente. Para estimar el efecto de la ley utilizando el modelo de efectos fijos únicamente estimamos la siguiente especificación utilizando MCO:

$$Y_{ijt} = \alpha_j + \delta_t + \tau R_{jt} + X'_{ijt}\beta + U_{ijt} \tag{9.2}$$

En este caso, los subíndices i corresponden al individuo, j al estado y t al tiempo; α_j es un efecto fijo por estado e indica que se debe incluir una dummy por estado; δ_t indica similarmente que se debe incluir una dummy por año para controlar por el efecto de tiempo; $X'_{ijt}\beta$ son controles a nivel individuo que cambian de estado a estado y a lo largo del tiempo; U_{ijt} es el error que en este caso debe incluir *cluster* a nivel estado; y τ es el efecto de diferencias en diferencias que nos interesa estimar.

El modelo de efectos fijos además permite estimar el efecto de algún cambio cuando este se da en distintas intensidades en diferentes grupos. Por ejemplo, supongamos que queremos estimar el efecto de un impuesto al consumo de alcohol que se empezó a establecer en los estados, pero cada estado lo implementó con diferente intensidad. En la especificación anterior, esto querría decir que ahora R_{jt} es simplemente el nivel del impuesto en el estado j y año t; Y_{ijt} puede ser consumo de alcohol; y el resto de las variables se describe igual.

9.3. Pruebas de robustez

Para verificar qué tan robustos son los resultados del método de diferencias en diferencias es común (y recomendable) llevar a cabo las siguientes pruebas:

1. Utilizar datos de más de un periodo previo al cambio y llevar a cabo un ejercicio de diferencias en diferencias. Con esta prueba se puede evaluar si la tendencia de ambos grupos antes del cambio sigue una tendencia paralela. Si se cuentan con varios periodos de información, puede utilizarse la siguiente especificación:

$$Y_{ijt} = \alpha_j + \delta_t + \sum_{t=1}^{T} \tau_t T_t R_j + X'_{ijt}\beta + U_{ijt}$$

   donde T_t es una dummy igual a uno si la observación se lleva a cabo en el periodo t y R_j es una dummy igual a uno si el estado j es del grupo de estados que implementaron la reforma. En este caso, si el cambio tuvo un impacto significativo y se dio en el año $t' \in (1, T)$, deberían observarse τ_s no significativas si $s < t'$ y τ_r significativas si $r \geq t'$.

2. Utilizar un grupo de control alternativo. Si los resultados son distintos utilizando un grupo de control u otro, esto puede representar un problema de validez. En clase ejemplificaremos esto utilizando triples diferencias. La idea de triples diferencias (DDD) es comparar dos estimadores de diferencias en diferencias (DD).

3. Hacer el mismo ejercicio de diferencias en diferencias utilizando una variable dependiente que no debería haber sido afectada por el cambio. Esto se conoce como *prueba de falsificación*.

9.4. Problemas comunes de diferencias en diferencias

Los críticas más usuales a tener en cuenta con el modelo de diferencias en diferencias son:

1. Que el cambio (o reforma) se produzca como resultado de condiciones pre-existentes.

    a. Targeting basado en las condiciones preexistentes. Supongamos que queremos ver el efecto de una reforma educativa sobre la cobertura escolar. Esta reforma consiste en construir escuelas y para llevarla a cabo se eligió a los 5 estados de la república. El modelo de diferencias en diferencias consistirá en comparar la cobertura escolar en los estados elegidos con los no elegidos antes y después de la implementación del programa. Sin embargo, si la selección de los estados para la reforma se llevó a cabo por que en dichos estados había mas presión por parte de los padres de familia, habría un problema de identificación. No será posible distinguir si fue la exigencia y preocupación de los padres, características de dichos estados que impulsaron la exigencia de los padres o la propia construcción de escuelas lo que generó los resultados del modelo de diferencias en diferencias.

    b. *Ashenfelter dip*. Este efecto toma su nombre de trabajos de investigación de Orley Ashenfelter y David Card que analizaron el efecto de programas de entrenamiento sobre el salario. Los investigadores se dieron cuenta que los participantes del programa habían decidido inscribirse a los programas de entrenamiento ya que su salario había disminuido recientemente. Por lo tanto, al estimar un modelo de diferencias en diferencias el efecto se incrementaba debido a que el grupo que participaba en el entrenamiento tenía salarios que habían disminuido recientemente antes del programa.

2. Uso de la forma funcional. Tomaremos un ejemplo en el cual después de una intervención el desempleo bajo de 30 % a 20 % en el grupo que fue afectado por una reforma, mientras que el desempleo en el grupo de control paso de 10 % a 5 %. Compararemos que pasa si utilizamos los valores medidos en tasa o si utilizamos logaritmos.

3. Comparación de efectos en corto versus largo plazo. En muchos casos, la pregunta relevante de alguna política es interesante respecto a su efecto

de largo o mediano plazo. Sin embargo, el supuesto de tendencia paralela es mas realista y fácil de justificar en el corto plazo.

4. Efectos heterogeneos entre ambos grupos. El modelo de diferencias en diferencias puede también aplicarse si los dos grupos fueron afectados por un cambio pero de distintas magnitudes (como se explico en el ejemplo del impuesto al consumo de alcohol en la Sección 2). En este caso, podría existir un problema si los efectos del cambio son heterogeneos entre los grupos. Este es un caso particular de una violación al supuesto de tendencia paralela.

Capítulo 10

Regresión Discontinua

El método de regresión discontinua es un método cuasiexperimental utilizado para identificar efectos causales de algun tratamiento. Este método se basa en cortes que surgen por ley o por diseño y que implican una discontinuidad en la implementación del tratamiento, mismo que viene definido a lo largo de alguna variable, G (normalmente llamada la *variable definitoria*. Este método hará posible determinar la relación causal de cambios en el estatus de tratamiento sobre alguna otra variable (la variable dependiente).

Un ejemplo tradicional que se ha utilizado son las votaciones por mayoría. En casos donde hay dos contrincantes en una elección, si se usa la regla de mayoría sabremos que un candidato gana si obtiene mas del 50 % de las votaciones. Supongamos que nos interesa ver si los gobiernos de izquierda hacen una diferencia en la política fiscal o económica (Pettersson-Lidbom, 2007). Si el partido de izquierda recibe más del 50 % votos gana la elección y si recibe menos la pierde. Llevar a cabo una estimación de MCO donde la variable dependiente es la tasa impositiva y tu variable de interés es que el gobierno de izquierda haya ganado muy probablemente te llevará a un sesgo por variable omitida (en particular, qué tan liberal es el electorado puede llevar a un sesgo). Intuitivamente, el método de regresión discontinua compara situaciones en las cuales el gobierno de izquierda apenas gana (i.e. reciba poco mas del 50 % de los votos) con situaciones donde apenas pierde (i.e. reciba poco menos del 50 % de los votos). En este caso se puede argumentar que justo en la discontinuidad pasamos de una situación en la que el partido de izquierda pierde a una en la que gana. Sin embargo, en ambos casos el electorado debe ser similarmente liberal (de igual manera no debe haber un brinco abrupto en cualquier otra variable que provoque sesgo). Lo que haremos entonces es ver si en esa discontinuidad hay un brinco discontinuo en la variable dependiente. De haberlo, el método de regresión discontinua atribuirá dicho brinco al hecho de que el gobierno de izquierda haya ganado la elección.

10.1. Planteamiento

Este método nos permitirá investigar el efecto de la variable T_i sobre Y_i, donde T_i será una variable dummy que especifica si el individuo i forma parte del grupo de "tratamiento" ($T_i = 1$) o de "control" ($T_i = 0$). El planteamiento utiliza nuevamente el concepto de resultados potenciales. Cada individuo tiene dos resultados potenciales, de los cuales únicamente observamos uno (el realizado). Y_i^T es el nivel de la variable dependiente que i tendría si forma parte del grupo de tratamiento, Y_i^C el que tendría si forma parte del grupo de control y Y_i el nivel observado.

A diferencia que en el caso de experimentos donde ambos grupos (tratamiento y control) eran determinados de forma aleatoria, en este caso son determinados por una regla objetiva de decisión. Dicha regla deberá especificar cortes especificos en los cuales la probabilidad de formar parte del grupo de tratamiento o control cambie de forma abrupta (i.e. discontinua). La regla deberá estar basada en una variable, que llamaremos la *variable definitoria* (G_i), misma que no debe poder ser manipulable por los individuos. Esta variable puede estar relacionada directamente con los resultados potenciales (y por tanto, con la variable dependiente), sin embargo, dicha relación (así como con el resto de las variables de control) **se asume como continua**.

El cambio discontinuo en la probabilidad de formar parte del grupo de tratamiento puede ser de dos tipos:

- **Sharp.** Donde la probabilidad de formar parte del tratamiento pasa de 0 a 1 en la discontinuidad.
- **Fuzzy.** Donde la probabilidad de formar parte del tratamiento cambia abruptamente en la discnotinuidad, pero no pasa de 0 a 1 debido a que existe la posibilidad de que debajo de la discontinuidad haya unidades de observación recibiendo el tratamiento y después de la discontinuidad puede haber unidades de observación no recibiendo el tratamiento (i.e. existe la posibilidad de que existan *never-takers* y *always-takers*.

En esta nota desarrollaremos gran parte de la identificación teórica basado en el caso *sharp* y al final se incluye una sección del caso *fuzzy*.

10.2. Regresión Discontinua Sharp

Este tipo de discontinuidad se refiere al caso en el cual la probabilidad de formar parte del grupo de tratamiento pasa de 0 a 1 después del corte que determina la discontinuidad. Ejemplos de este tipo de cortes incluyen: (i) márgenes por los cuales se pierde una elección; (ii) cortes administrativos que definen diferencias en precio (e.g. adulto mayor de 65 años); (iii) política en México de apoyo a los 125 municipios más pobres.\

En este caso T_i es definida por la *variable definitoria* (G_i) y el punto de discontinuidad es k:

$$T_i = 1\{G_i \geq k\} \tag{10.1}$$

Todos los individuos con $G_i \geq k$ estarán en el grupo de tratamiento (i.e. su participación es obligatoria) y todos con $G_i < k$ estarán en el grupo de control (i.e. su participación en el tratamiento está prohibida).

10.2.1. Análisis preliminar de datos

Una primera aproximación suele consistir en combinar dos componentes en una misma gráfica:

- **Medias condicionales locales.** Esto consiste en formar distintos bins y para cada bin, calcular la media condicional de la variable dependiente (Y_i). Esto equivale a llevar a cabo una regresión kernel con un kernel uniforme y solo gráficar los puntos medios de cada bin. Existen dos alternativas para la selección de los bins: (i) *equidistantes*, que quiere decir que la distancia sobre G de un punto a otro será el mismo y (ii) {cuantil-espaciado}, que quiere decir que cada bin tendrá el mismo número de observaciones[1].
- **Polinomio global.** Pese a que nuestra estimación será local en naturaleza, para tener una idea de posibles discontinuidades y una intuición acerca de qué tan extrapolable es el resultado local, muchas veces se hace un análisis exploratorio (que no siempre se presenta) donde se estima un polinomio global (de alto grado) antes y después de la discontinuidad.

Usualmente, estas estrategias muestran la existencia de la discontinuidad. Cuando estas alternativas fallan a mostrar evidencia de la existencia de la discontinuidad, rara vez se identifican en las estimaciones más precisas. Las medias condicionales y el polinomio global se pueden calcular con el comando `rdplot`.

10.2.2. Estimación con regresión lineal local

Para identificar el efecto del tratamiento sobre la variable dependiente tendremos que asumir:

- **Independencia** (condicional en la variable definitoria). $Y_i^T, Y_i^C \perp T_i \mid G_i$
- **Continuidad.** $E(Y_i^T|G_i = g)$ y $E(Y_i^C|G_i = g)$ son continuas en $g = k$.

[1] La ventaja de tener bins equidistantes es que visualmente, la gráfica es más clara. La ventaja de tener la gráfica cuantil-espaciada es que la misma gráfica te da una idea de la distribución de la variable definitoria (G_i).

Además de esos supuestos utilizaremos la *ley de esperanzas iteradas* que indica que el valor esperado de Y_i puede ser calculado como:

$$\begin{aligned} E(Y_i|G_i = g) = E(Y_i^T|T_i = 1, G_i = g) * Pr(T_i = 1|G_i = g) + \\ + E(Y_i^C|T_i = 0, G_i = g) * Pr(T_i = 0|G_i = g) \end{aligned} \tag{10.2}$$

Utilizando los supuestos establecidos y (10.2) tenemos que:

$$\begin{aligned} E[Y_i^C|G_i = k] &= \lim_{g \to k^-} E[Y_i^C|G_i = g] \\ &= \lim_{g \to k^-} E[Y_i^C|T_i = 0, G_i = g] \\ &= \lim_{g \to k^-} E[Y_i|G_i = g] \end{aligned} \tag{10.3}$$

$$E[Y_i^T|G_i = k] = \lim_{g \to k^+} E[Y_i|G_i = g] \tag{10.4}$$

Por lo tanto, utilizando estos supuestos podemos obtener el efecto de tratamiento en el punto $G_i = k$:

$$\tau_s = E[Y_i^T - Y_i^C|G_i = k] \tag{10.5}$$

En este caso será necesario utilizar extrapolación, ya que si G_i es continua, la probabilidad de observar alguna unidad con $G_i = k$ será cero. Por lo tanto, no tendríamos observaciones para llevar a cabo la estimación. En este caso utilizaremos observaciones con G_i arbitrariamente cerca del valor de corte k.

Para estimar el valor de τ_s, el *state-of-the-art* sugiere el uso de regresiones lineales locales utilizando únicamente las observaciones con $|G_i - k| < h$ (es decir, observaciones con G_i a menos de h de distancia del valor de corte k). El uso de regresiones paramétricas globales son un buen complemento para evaluar la continuidad y forma de los resultados potenciales a lo largo de G. El uso de regresiones lineales locales se debe a que nos interesa la estimación en un punto que además es una frontera. NW suele no ser tan adecuado para esos propósitos.

Para estimar el efecto de Regresión Discontinua con regresión lineal local:

$$\min_{\alpha_L, \beta_L} \sum_{i|k-h \leq G_i < k} K\left(\frac{G_i - k}{h}\right) \left(y_i - \alpha_L - \beta_L \left(G_i - k\right)\right)^2$$

y...

$$\min_{\alpha_R, \beta_R} \sum_{i|k \leq G_i < k+h} K\left(\frac{G_i - k}{h}\right) \left(y_i - \alpha_R - \beta_R \left(G_i - k\right)\right)^2$$

Y, por lo tanto, $\hat{\tau}_s = \widehat{\alpha_R} - \widehat{\alpha_L}$. Además suele verse también en algunas aplicaciones evaluar la sensibilidad al grado del polinomio local, de forma tal, que en vez de usar una regresión lineal local, se use un polinomio local.

En diversas aplicaciones, la estimación de la regresión lineal local, suele acompañarse de estimaciones que utilizan una forma paramétrica global. Sin embargo, siendo un estimador local identificado en $G_i = k$, el uso de estimadores locales como (LL) son preferidos. Las estimaciones paramétricas suelen utilizarse como parte de ejercicios de robustez. La estimación paramétrica consistiría en:

$$\min \sum_{i=1}^{N} \left(Y_i - \beta_0 - \tau_s T_i - f_r\left(G_i - k, T_i; \beta\right)\right)^2$$

donde $f\left(G_i - k, T_i; \beta\right)$ es un polinomio grado r de $(G_i - k)$ que puede tener (o no) sus componentes interactuados con T_i para indicar que las pendientes antes y después del corte pueden (o no) ser distintas. Para flexibilidad del estimador, se sugiere incluir las interacciones. Asimismo, las pruebas de robustez suelen incluir diversos grados del polinomio para evaluar la sensibilidad del estimador al uso de diferentes grados.

Para llevar a cabo la estimación de LL hay dos componentes que debemos seleccionar:

1. **Seleccionar la función kernel, $K(\cdot)$.** De acuerdo a Cattaneo et al. (2018) se sugiere utilizar la función triangular debido a que, si se elige un bandwidth óptimo que minimice el error medio cuadrático (MSE por sus siglas en inglés), la selección de un kernel triangular derivaría en un estimador puntual con *propiedades asintóticas óptimas*. Una alternativa que suele utilizarse también es el kernel uniforme debido a que, *bajo ciertas condiciones*, minimiza la varianza asintótica de un polinomio local (incluyendo LL). Usar un kernel uniforme equivale a llevar a cabo dos regresiones OLS simples de Y contra G, una con datos en el bandwidth por debajo del corte (i.e. $k - h \leq G_i < k$), y la otra con datos en el bandwidth por encima del corte (i.e. $k \leq G_i < k + h$). Por último, en diversos casos se ha visto también el uso del kernel Epanechnikov. Dicho esto, Cattaneo et al. (2018) indican que las estimaciones suelen ser poco sensibles a la selección de la función kernel. Se motiva a los alumnos a probar los distintos kernel y comparar sus resultados en la práctica debido al bajo costo computacional de llevar a cabo esto.

2. **Selección del bandwidth, h.** Como vimos en la *Nota 2* 3, la selección de h involucra un tradeoff entre sesgo y varianza. Es importante resaltar que en esta parte nos enfocamos en la selección óptima de h para obtener un estimador *eficiente* de τ_s. Dado que hay un tradeoff entre sesgo y varianza, este estimador no será insesgado debido a que una selección muy pequeña de h llevaría a un estimador muy *volatil*. Imbens y Kalyanaraman (2012) fueron los primeros en sugerir que se minimizara la media de los errores de la estimación al cuadrado (MSE por sus siglas en inglés)[2]. Posteriormente Calonico, Cattaneo y Titiunik (2014) refinaron el calculo de la h óptima

[2]Los errores en este caso son $Y_i - g(G_i)$ para observaciones dentro del bandwidth: $k - h < G_i < k + h$, donde $g(G_i)$ se estima utilizando el LL.

utilizando también como función objetivo la minimización de MSE. La función objetivo es:

$$\begin{aligned} MSE(\hat{\tau}_s) &= Sesgo^2(\hat{\tau}_s) + Var(\hat{\tau}_s) = \mathcal{B}^2 + \mathcal{V} \\ &= h^{2(p+1)}\ \mathcal{B}^2 + \frac{1}{nh}\mathcal{V} \end{aligned} \tag{10.6}$$

donde los términos $\mathcal{B}$ y $\mathcal{V}$ son los componentes del sesgo y la varianza que no dependen directamente de h y n. Estos componentes se relacionan con la estimación de $E\left(Y_i^T - Y_i^C | G_i = k\right)$. Por lo tanto, tenemos que: $\mathcal{B} = \mathcal{B}_+ - \mathcal{B}_-$ y $\mathcal{V} = \mathcal{V}_+ + \mathcal{V}_-$. Donde los componentes del sesgo se pueden *aproximar* como: $\mathcal{B}_+ \approx \mu_+^{(p+1)}\ B_+$, $\mathcal{B}_- \approx \mu_-^{(p+1)}\ B_-$. En este caso los componentes B_+ y B_- dependen de la selección de la función kernel $(K(\cdot))$ y del grado del polinomio local (p). Dado que en nuestra estimación asumimos estos componentes, el paquete estadístico de `rdrobust` utilizará métodos plug-in para estimar estos componentes. Y además:

$$\begin{aligned} \mu_+^{(p+1)} &= \lim_{g \to k^+} \frac{d^{p+1} E\left(Y_i^T | G_i = g\right)}{dG_i^{p+1}} \\ \mu_-^{(p+1)} &= \lim_{g \to k^-} \frac{d^{p+1} E\left(Y_i^C | G_i = g\right)}{dG_i^{p+1}} \end{aligned}$$

Estos componentes $\mu_+^{(p+1)}$ y $\mu_-^{(p+1)}$ dependen de la *curvatura* de la media de la variable dependiente. En el caso de la LL, estos componentes serán estimados con un polinomio local de segundo grado (o mayor) y utilizando un bandwidth b (que no necesariamente es igual a h). Esto lo veremos gráficamente en clase. En este caso, nos estamos enfocando en un bandwidth h en común, pero existen extensiones sencillas que permiten calcular un diferente bandwidth para observaciones a la izquierda y derecha del corte. Esto vale la pena cuando la curvatura de ambas funciones $E\left(Y_i^T | G_i\right)$ y $E\left(Y_i^C | G_i\right)$ es muy distinta (lo cual se puede corroborar gráficamente), que la densidad sea distinta de forma importante (lo cual es raro dado que no puede haber un cambio abrupto en la densidad alrededor de la vecindad de k) o que la variación de la variable dependiente sea muy distinta antes y después del corte (como veremos abajo).

En cuanto al término de la varianza, los componentes se pueden *aproximar* como: $\mathcal{V}_+ \approx \frac{\sigma_+^2}{f(G_i=k)}\ V_+$, $\mathcal{V}_- \approx \frac{\sigma_-^2}{f(G_i=k)}\ V_-$. Nuevamente, los componentes V_+ y V_- dependen de la selección de la función kernel $(K(\cdot))$ y del grado del polinomio local (p). El término $f(G_i = k)$ representa la densidad de la variable definitoria en el corte k. Y además:

$$\begin{aligned} \sigma_+^2 &= \lim_{g \to k^+} V\left(Y_i^T | G_i = g\right) \\ \sigma_-^2 &= \lim_{g \to k^-} V\left(Y_i^C | G_i = g\right) \end{aligned}$$

Como resultado, el h óptimo tendrá la siguiente forma:

$$h^*_{MSE} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2} \right)^{\frac{1}{2p+3}} \; n^{\frac{-1}{(2p+3)}} \tag{10.7}$$

Esta ecuación exhibe el tradeoff sesgo-varianza y ademaás muestra que la h óptima disminuye conforme aumenta el tamaño de muestra (n). Una precisión final a hacer es que si el sesgo estimado de la h^*_{MSE} es demasiado pequeño (tiende a cero), esto generará un problema en la definición de h^*_{MSE}. Por ello, en algunos casos la estimación del h^*_{MSE} incluye un *término de regularización* ($\mathcal{R}$).

$$h^*_{MSE} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2 + \mathcal{R}} \right)^{\frac{1}{2p+3}} \; n^{\frac{-1}{(2p+3)}} \tag{10.8}$$

Por default, rdrobust asume que este término es igual a 1, pero se puede modificar como parte de las opciones para hacerlo cero.

3. **¿Agregar controles?** Aunque no es necesario agregar controles en el contexto de regresión discontinua para evitar sesgo, pudiera existir una motivación desde el punto de vista de eficiencia. Para llevar a cabo esto hay dos alternativas: (i) hacer partial-out de la variable dependiente y estimar la regresion LL utilizando como variable dependiente los errores de la estimación; (ii) utilizar el comando `rdrobust` que permite hacer un ajuste local de los controles para llevar a cabo la estimación LL.

10.2.3. Inferencia

Como hemos discutido anteriormente, típicamente estaremos interesados en llevar a cabo una prueba de hipótesis acerca del parámetro poblacional de interés. En este caso, τ_s. Para esto, será necesario derivar la distribución de nuestro estimador $\hat{\tau}_s$.

Una particularidad que es importante tomar en cuenta al momento de calcular el estimador puntual es que el sesgo no es igual a cero. Existen dos alternativas que típicamente se utilizan para corregir el hecho de que en el estimador puntual hay sesgo:

10.2.3.1. Estimar y corregir el sesgo

En este caso, veremos una solución que emplea solo los datos que están dentro del bandwidth elegido por h^*_{MSE}. La distribución del estimador (τ_s) es:

$$\frac{\hat{\tau}_s - \tau_s - \mathcal{B}}{\sqrt{\mathcal{V}}} \sim N(0,1) \tag{10.9}$$

Hay dos aspectos a tomar en cuenta con respecto a la distribución:

- Ignorar el término del sesgo es incorrecto al derivar la distribución. Esto puede llevar a sobre-estimar significancia. Dicho de otra manera, tendremos efectos significativos en mayor proporción que la realidad porque nuestro estimador está centrado en $\tau_s + \mathcal{B}$. Es decir, tiene sesgo positivo.
- La varianza a su vez debe estar afectada por la corrección de agregar el sesgo

Para corregir por el primer punto, una posibilidad es generar un estimador de $\mathcal{B}$ y centrar los intervalos de confianza en $\hat{\tau}_s - \widehat{\mathcal{B}}$. Previamente, para el calculo de h^*_{MSE} se había detallado que un estimador del sesgo se podía realizar utilizando la curvatura de la media de la variable dependiente con respecto a G_i, por lo que se utiliza un polinomio de mayor grado al que se usa en la estimación puntual para tener un estimador del sesgo. Esto lo ilustraremos en clase. En cuanto al segundo punto de corrección de la varianza, la introducción del estimador del sesgo $\mathcal{B}$ hace que la varianza sea mayor a la que sería estimada por métodos tradicionales (e.g. OLS). El comando `rdrobust` presenta el intervalo de confianza con esta corrección de sesgo y varianza en la líniea indicada como *robust*. Es importante notar que el intervalo proporcionado por el comando está centrado en $\hat{\tau}_s - \widehat{\mathcal{B}}$ y no en el estimador puntual que discutimos previamente. Al igual que en el caso de OLS, la varianza puede ser ajustada para el caso de cluster SEs. Es importante que el estimador del sesgo en este caso se lleva a cabo con un bandwidth igual que el estimador puntual (h^*_{MSE}).

10.2.3.2. Utilizar un distinto bandwidth para generar intervalos de confianza

En la sección anterior se utilizó el bandwidth h^*_{MSE} para el estimador puntual y para llevar a cabo inferencia. Sin embargo, dicho bandwidth se selección por ser óptimo como estimador puntual. Una alternativa es utilizar otro bandwidth para llevar a cabo inferencia: uno que sea óptimo para dicho propósito.

Esta alternativa consiste en utilizar el h^*_{MSE} para el estimador putual y posteriormente seleccionar uno que minimice un estimador del *error de cobertura*, que representa el porcentaje de las veces en als cuales el intervalo de confianza que resulta de una estimación no incluye el valor verdadero. Esto da lugar al h^*_{CER} que se estima como parte del paquete `rdrobust`. Típicamente $h^*_{CER} < h^*_{MSE}$. También existe una alternativa como parte del paquete estadístico `rdrobust` de estimar un distinto bandwidth a la derecha y a la izquierda de la discontinuidad.

10.2.4. Tests de robustez

A continuación enumeramos los tests de robustez más comunmente aplicados en las estimaciones de regresión discontinua.

1. **McCrary test.** Una condición fundamental en la estimación con el método de RD es que la variable definitoria no sea manipulable. Si lo fuera, las unidades de observación podrían elegir su posición con respecto al corte (posiblemente basado en sus caracterísitcas observables y no observables) para influenciar si recibirían el tratamiento o no. esto haría imposible distinguir si las diferencias en la variable dependiente se deberían al tratamiento a alguna de estas características. Un test típico se basa en observar la distribución de la variable definitoria vía histogramas o densidades kernel. El *McCrary test* (2008) consiste en llevar a cabo una densidad kernel usando solo las observaciones a la derecha de la discontinuidad y otra con las observaciones a la izquierda y evaluar si el estimador de la densidad en k es distinta usando ambos estimadores.

2. **Tests de falsificación.** Estos consisten en utilizar controles como variables dependientes. Otra alternativa de distinguir que caracterísitcas de os individuos no provocan que se seleccionen a un lado u otro de la discontinuidad es utilizar variables pre-existentes de los individuos y llevar a cabo las estimaciones de RD empleándolas como variables dependientes. En general, deberían encontrar continuidad, es decir, no debería haber un cambio abrupto en el nivel de ninguno de los controles en la discontinuidad.

3. **Diferentes bandwidths.** Pese a que ya hemos discutido a detalle la selección del bandwidth, para el caso en el que usamos un mismo bandwidth para inferencia y estimación puntual, podemos llevar a cabo un test de sensibilidad para mostrar qué tan sensibles son los resultados a la selección del bandwidth. La literatura típicamente emplea para estas sensibilidades un bandwidth igual al doble y la mitad del bandwidth óptimo que se usa en los resultados principales. Idealmente, el resultado no debería ser muy sensible al bandwidth, pero es importante indicar que de ser sensible al bandwidth (especialmente al doble del caso base), esto no quiere decir que los resultados deban invalidarse, sin embargo, sugeriría que posiblemente hay una *curvatura* importante en las medias de los resultados potenciales que se están estimando. Una alternativa recomendable para tener un bandwidth adicional es seleccionar uno utilizando *cross-validation*. Para esto, podemos establecer como función objetivo la media del cuadrado de los errores y seguir los pasos establecidos en la sección 5.2 del Capítulo 3 . Imbens sugiere también que dada la naturaleza local del RD, se puede hacer cross validation quitando observaciones de las colas de la distribución de G_i. Algunos autores, por ejemplo, usan reglas de dedo como quitar la mitad de las observaciones de cada lado de la discontinuidad.

4. **Diferentes cortes (placebo cutoffs).** Este test consiste en llevar a cabo simulaciones en las cuales se modifica artificialmente el nivel del cutoff k y se llevan a cabo las estimaciones suponiendo que dicho corte está en otro valor de la variable definitoria. Al ser falsa esta aseveracón, los efectos obtenidos deberían ser muy pequeños y no significativos. Este test suele

ser parecido en espiritu a lo que se conoce como *tests de permutación* (e.g. Fischer Exact Tests).

5. **Quitar observaciones cerca de la discontinuidad.** Este test no es muy popular debido a que la teoría de RD indica que la identificación del efecto es local justo en el punto de la discontinuidad, k. Sin embargo, en casos en los cuales se sospecha de manipulación este test se basa en que los valores más cerca de la discontinuidad potencialmente son los más sujetos a manipulación. En este se caso se quitan algunas observaciones cerca de la discontinuidad y se repite la estimación. Obviamente, estamos refiriendonos a intervalos a remover muchos más pequeños que el bandwidth óptimo.

6. **Estimaciones paramétricas globales.** Como se mencionó en la sección **??**, los resultados de la estimación que utiliza LL, suele compararse con estimaciones paramétricas globales (i.e. no restringidas por el bandwidth) con distintos grados de polinomio.

10.3. Extensiones del modelo RD

En esta sección comentamos brevemente y con pocos detalles técnicos algunas extensiones del modelo de RD.

10.3.1. Regresión Discontinua "Fuzzy"

Como comentamos al inicio de este capítulo, existen en la práctica casos en los cuales algunos individuos que son elegibles para recibir el tratamiento deciden no tomarlo e individuos que no son asignados al tratamiento, consiguen una manera de tener acceso a él. Esto se conoce como *imperfect compliance*. Algunos ejemplos de fuzzy RD incluyen: (i) PMT, como Progresa/Oportunidades; (ii) becas basadas en la calificación de algún examen de admisión estandarizado.

Para esta sección será necesario entonces distinguir entre *asignación al tratamiento* y *tomar el tratamiento*. Para ello, cabe hacer una aclaración en la notación a utilizar: T_i seguirá siendo una dummy para identificar si el individuo i recibe o toma el tratamiento, mientras que una varible adicional ahora será Z_i, que es una dummy que indica si el individuo i es *asignado* al tratamiento. En el caso de RD, dicha asignación es vía una regla que es función de la variable definitoria G_i. Al igual que en las secciones anteriores, asumiremos que existe un corte y que los indiviuos no pueden modificar su posición de G_i respecto al punto de corte (k) que establece la elegibilidad. A diferencia que el caso de RD Sharp, ahora, $Z_i = 1\{G_i \geq k\}$.

En cuanto al tratamiento, este es ahora una función de la asignación al tratamiento, y de forma similar a los resultados potenciales, cada individuo tendrá

dos posibles escenarios: ser o no asignado al tratamiento. En cada posible realización de Z_i, el individuo podrá tomar una decisión con respecto a si toma o no el tratamiento. En este caso $T_i(Z_i)$ será una dummy que indica si el individuo i toma el tratamiento y esta decisión es función de la dummy Z_i que indica asignación. Dada esta notación, existen cuatro tipos de individuos, que previamente hemos descrito en el contexto de experimentos aleatorios:

(a) *Always takers*: aquellos individuos que sean o no asignados, deciden sí tomar el tratamiento. Es decir, $T_i(1) = T_i(0) = 1$

(b) *Never takers*: aquellos individuos que sean o no asignados, deciden no tomar el tratamiento. Es decir, $T_i(1) = T_i(0) = 0$

(c) *Compliers*: aquellos individuos que si son asignados al tratamiento deciden sí tomarlo, pero si no son asignados, no lo toman. Es decir, $T_i(1) = 1$ y $T_i(0) = 0$

(d) *Defiers*: aquellos individuos que si son asignados al tratamiento deciden no tomarlo, pero si no son asignados, sí lo toman. Es decir, $T_i(1) = 0$ y $T_i(0) = 1$

En la *fuzzy RD*, la probabilidad de formar parte del grupo de tratamiento tiene un cambio discontinuo en el punto de corte $G_i = k$. Sin embargo, dado el *imperfect compliance*, la probabilidad no cambia de 0 a 1. El hecho de que haya un cambio abrupto en la discontinuidad en el punto k indica que en esa parte de la distribución existe una proporción razonable de *compliers*. \

En los contextos antes descritos es posible estimar dos parámetros simplemente siguiendo la estrategia de sharp RD:

- **Intent to Treat (ITT)**: $\tau_{ITT} = E\left[(T_i(1) - T_i(0))(Y_i^T - T_i^C)|G_i = k\right]$
- **First Stage (FS)**: $\tau_{FS} = E\left[T_i(1) - T_i(0)|G_i = k\right]$

Si además de estos parámetros, nos interesara obtener el efecto promedio de tratamiento, será necesario agregar dos supuestos adicionales a los establecidos en el *sharp RD*:

1. **Independencia de asignacion.** Y_i^T y Y_i^C no dependen de Z_i.
2. **Monotonicidad.** $T_i(g)$ es no-creciente en x si $x = k$. Este supuesto es similar al supuesto de variables instrumentales que indica que no deben existir los *defiers*.

Cabe destactar respecto a los supuestos, que contrario al caso de variables instrumentales, en este caso no es necesario asumir exogeneidad. Este supuesto se sustituye por el hecho de que las observaciones son muy similares en una vecindad alrededor de la discontinuidad.

Siguiendo con la estrategia de variables instrumentales, en este caso podremos estimar el efecto del tratamiento para los **compliers** que además tienen $G_i = k$. Este estimador corresponderá a:

$$\tau_f = \frac{\lim_{g \to k^+} E[Y_i|G_i = g] - \lim_{g \to k^-} E[Y_i|G_i = g]}{\lim_{g \to k^+} E[T_i|G_i = g] - \lim_{g \to k^-} E[T_i|G_i = g]} \tag{10.10}$$

Para estimar el valor de τ_f nuevamente podremos utilizar el resultado de dos regresiones locales lineales, ya que como explicamos anteriormente, utilizando los supuestos de *sharp RD* podemos estimar el τ_{ITT} y el τ_{FS}. Para obtener el estimados del efecto de tratamiento τ_f basta solo utilizar el cociente de los parámetros previos:

$$\hat{\tau}_f = \frac{\hat{\tau}_{ITT}}{\hat{\tau}_{FS}} (\#eq : tfrd_est) \tag{10.11}$$

Con respecto a la selección del *bandwidth* ahora la selección es más compleja ya que no solo existe la posibilidad de elegir distintos bandwidths a la derecha e izquierda de la discontinuidad, sino que además tenemos dos outcomes involucrados: Y_i y T_i. En general, la literatura sugiere elegir un solo bandwidth: aquel que sea el mínimo bandwidth óptimo del MSE que se eligiría entre usar Y_i o T_i. Típicamente, T_i suele tener un comportamiento mas plano (menos variable), lo cual suele implicar que de ambos outcomes, suele ser el que está relacionado con un bandwidth menor.

Finalmente, para llevar a cabo inferencia podemos utilizar nuestro resultado del *método Delta* y emplear la siguiente distribución:

$$\sqrt{NH}\left(\hat{\tau}_f - \tau_f\right) \Longrightarrow N\left(0, \frac{1}{\tau_{FS}^2} \cdot V_{ITT} + \frac{\tau_{ITT}^2}{\tau_{FS}^4} \cdot V_{FS} - 2 \cdot \frac{\tau_{ITT}}{\tau_{FS}^3} \cdot Cov_{ITT,FS}\right) \tag{10.12}$$

donde

$$\begin{aligned} V_{ITT} &= \frac{4}{f_G(k)} \cdot \left(\sigma^2_{Y(r)} + \sigma^2_{Y(l)}\right) \\ V_{FS} &= \frac{4}{f_G(k)} \cdot \left(\sigma^2_{T(r)} + \sigma^2_{T(l)}\right) \\ Cov_{ITT,FS} &= \frac{4}{f_G(k)} \cdot \left(\sigma_{Y(r),T(r)} + \sigma_{Y(l),T(l)}\right) \end{aligned}$$

Necesitamos obtener estimadores de estos componentes: $\sigma^2_{Y(r)}$, $\sigma^2_{Y(l)}$, $\sigma^2_{T(r)}$, $\sigma^2_{T(l)}$, $\sigma_{Y(r),T(r)}$, $\sigma_{Y(l),T(l)}$ y $f_G(k)$. El último término se puede estimar fácilmente utilizando una densidad kernel. En cuanto a los otros términos, se pueden estimar con los términos de error que resultan de las regresiones (LL) que se utilizaron para estimar τ_{ITT} y τ_{FS}. Para clarificar la notación supongamos que $\tau_{ITT} = \alpha_{Y(r)} - \alpha_{Y(l)}$ y $\tau_{FS} = \alpha_{T(r)} - \alpha_{T(l)}$[3]. Podemos definir a los términos de

[3] $Y(r)$ corresponde a utilizar Y como variable dependiente en la regresión (LL) con los datos a la derecha de la discontinuidad (i.e. aquellos entre k y $k+h$). $Y(l)$ es similar, pero utiliza los datos a la izquierda de la discontinuidad. Asimismo $T(r)$ y $T(l)$ corresponden a utilizar T_i como variable dependiente en la regresión (LL) del *first stage*.

error como $\epsilon_{Y(r),i} = Y_i - \mu_{Y(r)}(X_i)$ para los valores a la derecha de la discontinuidad al usar Y_i como variable dependiente. Similarmente, podemos obtener los términos de error $\epsilon_{Y(l),i}$, $\epsilon_{T(r),i}$ y $\epsilon_{T(l),i}$. Y con ello podríamos estimar:

$$\hat{\sigma}^2_{Y(l)} = \frac{1}{N_l}\sum_{i=1}^{N} 1\{k-h < G_i < k\}\ \hat{\epsilon}^2_{Y(l),i}$$

$$\hat{\sigma}^2_{Y(r)} = \frac{1}{N_r}\sum_{i=1}^{N} 1\{k \leq G_i < k+h\}\ \hat{\epsilon}^2_{Y(r),i}$$

$$\hat{\sigma}^2_{T(l)} = \frac{1}{N_l}\sum_{i=1}^{N} 1\{k-h < G_i < k\}\ \hat{\epsilon}^2_{T(l),i}$$

$$\hat{\sigma}^2_{T(r)} = \frac{1}{N_r}\sum_{i=1}^{N} 1\{k \leq G_i < k+h\}\ \hat{\epsilon}^2_{T(r),i}$$

$$\hat{\sigma}_{Y(l),T(l)} = \frac{1}{N_l}\sum_{i=1}^{N} 1\{k-h < G_i < k\}\ \hat{\epsilon}^2_{Y(l),i} \cdot \hat{\epsilon}^2_{T(l),i}$$

$$\hat{\sigma}_{Y(r),T(r)} = \frac{1}{N_r}\sum_{i=1}^{N} 1\{k \leq G_i < k+h\}\ \hat{\epsilon}^2_{Y(r),i} \cdot \hat{\epsilon}^2_{T(r),i}$$

Aqui N_l es el número de observaciones que caen dentro del bandwidth $(k-h, k)$ y N_r el número de observaciones que caen dentro del bandwidth $[k, k+h)$.

10.3.2. Multiples cortes

Una desventaja de las estimaciones de RD es su naturaleza *local*. Esta característica puede ser parcialmente subsanada en contextos en los cuales existen diversos puntos de corte. Ejemplos de estas situaciones incluyen: (i) votaciones con mas de dos partidos, (ii) reglas escalonadas para dar recursos públicos o privados y (iii) mecanismos de asignación a escuelas con sobre-demanda.

Una estrategia que comúnmente se sigue en estos contextos es recentrar el punto de corte en un único punto (0) que resulta de restar el valor de la variable definitoria para la observación i de su punto de corte correspondiente k_i. De esta forma se crea una variable definitoria nueva: $\tilde{G}_i = G_i - k_i$. Y utilizando $\tilde{G}_i$ se puede llevar a cabo el procedimiento de RD como hemos explicado en esta nota. El resultado de esto será (en el caso de *sharp RD*:

$$\tau_s = E\left(Y_i^T - Y_i^C | \tilde{G}_i = 0\right) = \sum_{k \in \mathcal{K}} \tau_s(k)\omega(k) \tag{10.13}$$

donde $\mathcal{K}$ es el conjunto de todos los puntos de corte, $\tau_s(k)$ es el efecto de tratamiento en el punto de corte k y $\omega(k)$ es un ponderador que resulta del peso

definido por la densidad en el punto $G_i = k$ como porcentaje de la suma de todas las densidades $\sum_{\mathcal{K}} f_{G|\mathcal{K}}(k|\mathcal{K})$.

Explotar los diferentes cortes equivale a estimar individualmente cada uno de los $\tau_s(k)$ y posteriormente analizar cómo se comporta la heterogeneidad de estos efectos de tratamiento a lo largo de la distribución de G_i. Aqui en vez de tener una sola observación del efecto de tratamiento, tendremos tantas como puntos de corte existan.

10.3.3. Kink-RD

El *kink-RD* es una pequeña extensión de *RD* y su intuición puede entenderse mejor desde el punto de vista de IV. Este tipo de RD consiste en situaciones en las cuales en el punto de corte hay un cambio discontinuo en la pendiente de la variable variable dependiente y potencialmente del estatus de tratamiento. Algunos ejemplos suelen darse cuando hay topes en cierta cantidad de beneficios, por ejemplo, beneficios de desempleo que dependen del tiempo de desempleo o transferencias monetarias (como Progresa) que dependen del numero de hijos, pero están topadas para desincentivar la fertilidad.

Para poder entender la estimación y la intuición de como se estima el *kink-RD* empecemos por definir que nuevamente, el parámetro de interés es el efecto del tratamiento sobre alguna variable dependiente Y_i. Para poder llevar a cabo esta estimación nos enfocamos en cambios en pendientes de la variable Y justo en el punto de la discontinuidad y dividimos dicho cambio de pendiente sobre cambio en la probabilidad de tratamiento. Denotaremos esto como:

$$\tau_k = \frac{\lim_{g\to k^+} \frac{dE[Y_i|G_i=g]}{dG_i} - \lim_{g\to k^-} \frac{dE[Y_i|G_i=g]}{dG_i}}{\lim_{g\to k^+} \frac{dE[T_i|G_i=g]}{dG_i} - \lim_{g\to k^-} \frac{dE[T_i|G_i=g]}{dG_i}} \tag{10.14}$$

Esta estimación se puede llevar a cabo nuevamente via regresiones LL, solo que en esta ocasión el enfoque esta sobre la pendiente de la variable definitoria en el punto de la discontinuidad. Se sugiere al lector interesado en profundizar acerca de este tema que revise Card et al. (2016).

10.4. Local randomization

Una última alternativa a la estrategia de regresión discontinua consiste en tratar a las observaciones cercanas al punto de corte como un *experimento aleatorio local*. Es decir, asumir que aquellas observaciones cerca de la discontinuidad definen de forma exógena su estatus de tratamiento. Esta interpretación tienen que ver con la motivación de RD de asumir que las observaciones un poco antes de la discontinuidad son casi idénticas a las observaciones un poco después

de la discontinuidad. Dado que asumimos que estas observaciones cerca de la discontinuidad se comportan aleatoriamente con respecto al experimento, la estrategia de análisis suele ser similar a la utilizada en los experimentos. Basta con llevar a cabo una regresión de la siguiente forma:

$$Y_i = \tau_1 \, T_i + X_i' \, \beta + U_i \tag{10.15}$$

donde Y_i es la variable dependiente, T_i es la dummy de tratamiento y X_i es un grupo de controles que pueden o no ser incluidos. En este caso, la estimación se lleva a cabo sólo con las observaciones que están en una ventana alrededor de la discontinuidad. Es decir, aquellas cuya variable definitoria $G_i \in (k - w, k + w)$.

Sin embargo, para poder llevar a cabo un análisis de *local randomization* es importante que se cumplan dos supuestos:

1. **No manipulación de la variable definitoria.** De forma similar al caso continuo, debe ser el caso que una unidad no pueda modificar o actuar estratégicamente para afectar su posición de la variable definitoria con el propósito de afectar su elegibilidad al tratamiento.

2. **Independencia de T dentro del bandwidth.** Dentro del bandwidth, los resultados potenciales dependen de la variable definitoria únicamente a través del tratamiento (T_i) y no directamente: $\{Y_i^T, Y_i^C\} \perp G_i | T_i$. En la práctica este supuesto quiere decir que los resultados potenciales no deben estar relacionados con la variable definitoria G_i dentro del bandwidth. Es decir, a diferencia que en el caso continuo, no de debe ser el caso que los resultados potenciales tengan una tendencia (positiva o negativa) conforme cambia G_i. Por el contrario, debe haber una tendencia constante (plana). Esto se puede verificar gráficamente y a través de un test de exogeneidad similar al de una tabla de balance que se usa en experimentos aleatorizados (por discutirse mas adelante en la selección del bandwidth). En caso de que esto no se cumpla, se puede hacer una transformación de la variable dependiente (partial-out) para remover la tendencia. Esto requerirá que se asuma una forma funcional paramétrica (tendencia lineal o cuadrática, por ejemplo).

10.4.1. Fischer Exact Test

Una clara desventaja del método de *local randomization* es la pérdida de poder estadístico que resulta de que pocas observaciones quedan en la muestra una vez que se elige una ventana pequeña de forma que se cumpla con el supuesto (2). Un método que suele utilizarse con muestras pequeñas es el *Fischer Exact Test*. Para llevar a cabo este método hay tres componentes que deben elegirse:

- Un bandwidth, w. Cabe señalar que este *bandwidth* es distinto en espíritu de aquel del caso continuo ya que aquí no enfrentamos el tradeoff sesgo

vs varianza. En cambio, aquí el bandwidth debe asegurar que las observaciones que acaban en la muestra se comporten como un experimento local. Por lo pronto asumiremos que w es conocido y más adelante discutiremos su selección.

- Un mecanismo de asignación aleatoria para llevar a cabo simulaciones, ψ. Dos alternativas comunes son: aleatorización completa o asignación *Bernoulli*. El ejemplo disponible en Cattaneo et al. (2018) utiliza aleatorización completa. La asignación *Bernoulli* toma una probabilidad de asignación a tratamiento específica y cada unidad tiene dicha probabilidad de ser asignada independientemente al tratamiento[4].

- El estadístico de interés ($S(Y_i, \psi, w)$). Típicamente será la diferencia de medias entre tratamiento y control. Noten que el estadístico es función del bandwidth (w), de la variable dependiente (Y_i) y del mecanismo de asignación (ψ). Otros estadísticos disponibles en el software `locrand` son: (i) *Kolmogorov-Smirnov* (KS), que calcula la máxima distancia de la densidad acumulada de los resultados bajo tratamiento y control; y (ii) *Wilcox rank sum* (WR), que corresponde a la suma de las posiciones de los individuos en tratamiento (donde la posición se calcula después de ordenar de menor a mayor las observaciones usando Y_i)[5].

El objetivo del Fischer Exact Test (FETs) es determinar qué tan atípico es el estadístico observado (S^*, que resulta de la asignación verdadera de tratamiento) si se le compara con una asignación artificial de tratamiento, donde se simulan distintos escenarios de asignación de tratamiento. La intuición detrás de este método es que si el tratamiento no tuviese ningun efecto, el estadístico observado no debería ser muy distinto a simular el tratamiento, ya que en ambos casos el resultado potencial observado debería ser muy similar al no observado. Esto es similar a un ejercicio de falsificación donde se simulan escenarios falsos de asignación del tratamiento. Dichos escenarios falsos deberían resultar en efectos bajos o nulos del tratamiento. Por lo tanto, si el tratamiento no tiene efecto, dichas simulaciones no serían muy distintas al estadístico observado en la realidad.

El estadístico que resumirá qué tan atípico es el estadístico observado es el valor-p de una distribución empírica de estadísiticos que se generan a través de las distintas simulaciones. Imaginemos que las simulaciones generan la siguiente lista de estadísticos: $\{S_1, ..., S_M\}$, donde M es la cantidad de simulaciones. En este caso el valor-p sería (en el caso de una prueba bilateral)[6]

$$p = 2 \cdot \text{mín} \left\{ \frac{\sum_{m=1}^{M} \mathbf{1}\left(S_m \geq S^*\right)}{M}, \frac{\sum_{m=1}^{M} \mathbf{1}\left(S_m \leq S^*\right)}{M} \right\} \tag{10.16}$$

[4]Una desventaja de este método es que podría darse el caso que todas las unidades sean asignadas a tratamiento o control, en cuyo caso no se podrá producir el estadístico de interés.

[5]La ventaja de este estadístico es que es menos sensible a outliers en Y.

[6]En el caso de una prueba unilateral no se multiplica por 2.

Para generar las simulaciones se siguen los siguientes pasos:

1. Se utiliza el método seleccionado de asignación aleatoria (ψ_m) que define para cada simulación m a que observaciones les corresponde el tratamiento y el control. En este caso ψ_m es un vector que para designa tratamiento o control a las observaciones que caen dentro del bandwidth: $G_i \in (k - w, k + w)$.
2. En cada simulación se calcula el estadístico de interés $S_m = S(Y_i, \psi_m, w)$

Un aspecto relevante de las simulaciones es que una unidad i podría estar designada al tratamiento siendo que en realidad recibió el control (o viceversa). Para poder calcular su resultado potencial contrafactual (el no observado), será necesario establecer una hipótesis nula. En FETs, dicha hipótesis nula corresponde a un efecto homogeneo de tratamiento, τ (al ser homogeneo es igual al efecto promedio de tratamiento):

$$H_0 : Y_i^T = Y_i^C + \tau \tag{10.17}$$

Cabe destacar que τ bajo la hipótesis nula será una constante. Típicamente en evaluación de proyectos se utiliza $\tau = 0$ para evaluar si se puede rechazar un efecto nulo del tratamiento. Utilizando esta constante se podrá generar en las simulaciones el resultado potencial para cada observación aun si esta no es observada. Por ejemplo, imaginemos que el individuo 1 recibió tratamiento y tiene $Y_1 = 10$. Si dicho individuo en la simulación ψ_j recibe control, necesitaríamos conocer Y_1^C, pero solo observamos Y_1^T. Bajo la hipótesis nula $Y_1^T = Y_1^C$, por lo tanto, en la simulación ψ_j, dicho individuo tendrá $Y_1^C = 10$. Si en cambio la hipótesis nula estableciera que $\tau = 2$, en este mismo ejemplo tendríamos (bajo H_0): $Y_1^C = Y_1^T - \tau = 10 - 2 = 8$. En Cattaneo et al. (2018) hay un ejemplo que vale la pena repasar y tener claro en las pp. 16-18.

Por útlimo, es posible generar un intervalo de confianza de $1 - \alpha$ haciendo de forma recurrente el procedimiento anterior. Para esto, se harán varias pruebas de hipótesis para un conjunto definido de $\{\tau_1, ..., \tau_L\}$. Todos aquellos valores de τ que tengan un valor-p mayor a α serían incluidos en el intervalo de confianza.

Todos estos pasos descritos se pueden llevar a cabo utilizando el paquete `locrand` en Stata o R. Dicha función fue creada por Cattaneo et al. y sigue la notación descrita en Cattaneo et al. (2018) que es la principal referencia de esta nota.

10.4.2. Definición de w

Por último, la definición del bandwidth w en el caso de *local randomization* sigue la lógica de identificar la existencia de un escenario similar al de un experimento aleatorio. Por lo tanto, lo que se hace es restringir las observaciones a una ventana $G_i \in \{k - w, k + w\}$. Utilizando estas observaciones se genera una tabla

de balance y se calcula el valor-p de la prueba de que los coeficientes $\{\gamma_1, ..., \gamma_K\}$ son conjuntamente iguales a cero utilizando la siguiente especificación:

$$T_i = \gamma_0 + \gamma_1 \, X_{1i} + \cdots + \gamma_K \, X_{Ki} + U_i \tag{10.18}$$

Si efectivamente tuvieramos algo cercano a una asignación aleatoria dicho valor-p debería ser alto. Para definir entonces la ventana se prueban distintos valores para w y el valor-p resultante nos dará información acerca de valores de w que es razonable considerar. Recomiendo ver la Figura 2.5 en Cattaneo et al. (2018). Asimismo, una tabla balanceada da evidencia a favor de que para las observaciones dentro del bandwidth, Y_i no debería de estar relacionada con G_i mas que a través de T_i, como lo establece el supuesto (2) de *local randomization*.

10.4.3. Consideraciones finales

Muchas de las pruebas de robustez que describimos para el caso continuo de RD aplican también en el contexto de *local randomization*, tales como pruebas de falsificación utilizando controles, revisar la densidad de la variable definitoria, sensibilidad al bandwidth y puntos de corte placebo.

El uso de *local randomization* en vez del RD tradicional suele surgir bajo dos situaciones: pocas observaciones cerca de la discontinuidad y tener una variable definitoria discreta. Una ventaja es que en esta nota hemos revisado diversas pruebas para determinar si una u otra identificación son adecuadas.

Capítulo 11

Missing Data

Uno de nuestros supuestos principales en MCO es que nuestra muestra es aleatoria y, por lo tanto, es representativa de la población en general. De no cumplirse este supuesto, nos encontrariamos con un problema de **sesgo muestral**. Este problema puede no ser generado únicamente por deficiencias de muestreo al momento de elegir las unidades para recabar los datos. También puede surgir de que los individuos no reporten o no puedan reportar sus datos. Por ejemplo, si estamos haciendo un estudio de los determinantes de las calificaciones de alguna prueba estandarizada en niños de secundaria. Es posible que no tengamos el dato de la prueba para algunos niños elegidos en la muestra porque dichos niños faltaron a clases el dia de la prueba, se negaron a tomarla o los padres no dieron consentimiento para el estudio. Otra causa de sesgo muestral surge en el contexto de *datos panel*. Pese a que la muestra original cumpla con las condiciones de ser aleatoria y representativa para la población que se pretende, al hacer seguimiento de los datos es común que algunos individuos migren, hayan fallecido o ya no quieran participar en el seguimiento de datos. En estos casos, encontrarse con una muestra mas restringida para la cual toda la información está disponible, puede crear limitaciones importantes de validez de los resultados. La población para la cual, la muestra con información completa es representativa ya no es la misma si es que los individuos que faltan siguen un patrón o tienen características específicas (lo cual suele ser el caso).

Por último, pudiera darse el caso que se quiera hacer un análisis representativo para una población distinta respecto a la que se contempló inicialmente. Esto puede ser visto como un problema de falta de datos también siendo que faltan datos de individuos en particular que podría hacer que la muestra fuese representativa para esta otra población. Este es un problema de *validez externa*.

En esta nota primero describimos bajo qué condiciones la falta de datos no genera sesgo. Posteriormente, describe el modelo de *Heckit*, el cual utiliza variables explicativas para agregar una variable explicativa en una regresión que busque

remover el sesgo por selección de individuos. Por último, mostramos el método de ponderación invesa de probabilidades (*inverse probability weight*). Este método consiste en reponderar cada observación para que la nueva distribución que resulta de dicha reponderación sea significativa de la población objetivo.

11.1. Planteamiento general

Imaginemos que contamos con una base de datos, la cual tiene información no disponible para algunas variables. En estos casos nos encontramos con que solo podremos llevar a cabo la estimación para un subconjunto de nuestra muestra elegida. En teoría estariamos interesados en los resultados de la estimación con la muestra completa, pero por razones logísticas, no es posible hacer esta estimación.

Primero necesitamos determinar en qué casos nuestra estimación tendría sesgo. Para ello necesitamos partir de nuestro modelo que queremos estimar:

$$Y_i = X_i'\beta + U_i \tag{11.1}$$

Suponemos que podríamos estimar este modelo con MCO de forma insesgada si tuviéramos acceso a todos los datos. Esto querría decir que $E(U_i|X_i) = 0$. Sin embargo, tenemos algunos individuos para los cuales no contamos con todos sus datos. Sea s_i una variable dummy que indica si para el individuo i tenemos los datos disponibles y, por lo tanto, lo podemos utilizar en la estimación.

Partiendo del modelo (11.1) podemos obtener:

$$s_iY_i = s_iX_i'\beta + s_iU_i \tag{11.2}$$

Nótese que estimar este modelo con todas las observaciones es equivalente a estimar (11.1) con la muestra restringida, es decir, con las observaciones para las cuales $s_i = 1$. Por lo tanto, estaremos interesados en determinar bajo qué condiciones podemos estimar (11.2) consistentemente.

En este caso, estamos utilizando todas las observaciones, por lo tanto, aun no es un problema el sesgo muestral. Necesitamos enotnces fijarnos en las condiciones de primer orden de la estimación para determinar que no haya sesgo. En este caso, las condiciones de primer orden serían:

$$E[(s_iX_i)(s_iU_i)] = E[s_iX_iU_i] = 0 \tag{11.3}$$

porque $s_i^2 = s_i$. Por lo tanto, no habrá sesgo si $E(s_iU_i|s_iX_i) = 0$.

Situaciones bajo las cuales no tendremos sesgo entonces son:

1. Si $s_i = f(X_i)$. En este caso $E(s_i U_i | s_i X_i) = s_i E(U_i | s_i X_i) = 0$ porque por condiciones de MCO $E(U_i|X_i) = 0$. Imaginémos que en el ejemplo propuesto anteriormente las autoridades escolares deciden "esconder'á lo niños con promedio menor a 7. Por lo tanto $s_i = 1\{Prom_i \geq 7\}$. Siempre que incluyamos como control en nuestra estimación el promedio del niño ($Prom_i$) no habrá una preocupación de sesgo muestral[1].

2. Si $s_i \perp (X_i, U_i)$. Dicho de otra manera si la selección es "aleatoria'ó al menos independiente de variables observables y no observables del individuo que influyan sobre la variable dependiente. En este caso se cumplirá que: $E(s_i X_i U_i) = E(s_i)E(X_i U_i) = 0$. Esto podría suceder si tenemos una muestra muy grande y decidimos omitir observaciones al azar o si se pierden exámenes de algunos niños de forma accidental.

Un caso en el cual habría sesgo en la estimación es si tenemos observaciones truncadas basado en los valores de Y_i. Imaginense, por ejemplo, que las escuelas deciden esconder los resultados de los examenes más bajos. Supongamos que $s_i = 1$ solo si $Y_i > c$. En este caso tendremos que $s_i = 1$ si $U_i > c - X_i'\beta$. Por lo tanto, s_i no será independiente de variables no observadas (o del error) y tendremos sesgo en la estimación de (11.1) con la muestra restringida.

11.2. Heckit

Una solución a casos en los cuales la variable dependiente no es observada para algunos individuos y si parece haber sesgo muestral selectivo, consiste en el modelo de Heckman (**Heckit**). Este método es un estimador de máxima verosimilutud. Se basa en la idea que la selección de observaciones disponibles se puede determinar como una función de X_i y algunas otras variables que no afectan a Y_i. Es decir, especifican un modelo de selección:

$$s_i = 1\{Z_i'\gamma + V_i\} \tag{11.4}$$

donde asumiremos que Z_i incluye todas las variables de X_i y otras adicionales, que el error V_i es independiente de Z_i y que:

$$Z_i'\gamma = \gamma_0 + \gamma_1 Z_{1i} + \cdots + \gamma_M Z_{Mi} \tag{11.5}$$

[1]En este caso debe cumplirse además que la calificación de corte (7 en este caso) no este relacionada con términos no observados que puedan afectar la calificación. Por ejemplo, si se excluye a los de promedio reprobatorio y ser un alumno con promedio reprobatorio te influye en el ánimo, lo cual a su vez afecta tu rendimiento en el examen, entonces controlar por el promedio no será suficiente.

Basados en el supuesto que $(U_i, V_i) \perp Z_i$ y partiendo de (11.1) obtenemos:

$$\begin{aligned} Y_i &= X_i'\beta + U_i \\ E(Y_i|Z_i, V_i) &= X_i'\beta + E(U_i|Z_i, V_i) \\ &= X_i'\beta + E(U_i|V_i) \\ &= X_i'\beta + \rho V_i \end{aligned} \tag{11.6}$$

donde asumimos que $E(U_i|V_i) = \rho V_i$, lo cual surge del supuesto de que U_i y V_i son conjuntamente normales con media cero.

Esta ecuación no puede ser estimada dado que V_i no es observada, pero podemos utilizarla como punto de partida para estimar $E(Y_i|Z_i, s_i)$:

$$E(Y_i|Z_i, s_i) = x_i'\beta + \rho E(V_i|Z_i, s_i) \tag{11.7}$$

Dado que V_i tiene una distribución normal estándar, al igual que en el caso de Tobit, podemos mostrar que cuando $s_i = 1$:

$$\begin{aligned} E(V_i|Z_i, s_i = 1) &= E(V_i|V_i \geq Z_i'\gamma) \\ &= \frac{\phi(Z_i'\gamma)}{\Phi(Z_i'\gamma)} = \lambda(Z_i'\gamma) \end{aligned} \tag{11.8}$$

Sustituyendo este resultado en (11.7) obtenemos:

$$E(Y_i|Z_i, s_i = 1) = X_i'\beta + \rho\lambda(Z_i'\gamma) \tag{11.9}$$

Cabe notar que en este caso asumimos que V_i se distribuye como una normal estándar. Este supuesto es clave para poder estimar γ y asi calcular para cada individuo $\lambda(Z_i'\gamma)$. Dado que V_i se distribuye como una normal estándar y la definición (11.4), tendremos que:

$$Pr(s_i = 1|Z_i) = Pr(V_i < Z_i'\gamma) = \Phi(Z_i'\gamma) \tag{11.10}$$

Por lo tanto, el procedimiento del modelo Heckit consistirá de los siguientes pasos:

1. Se estimará (11.10) utilizando el modelo probit para estimar γ. En esta estimación se utilizarán las variables de Z_i y todas las observaciones (incluso aquellas que no cuentan con la variable dependiente, i.e. aquellas para las cuales $s_i = 0$).
2. Se utilizará el estimador de γ para calcular $\lambda(Z_i'\gamma)$ para cada individuo.
3. Utilizando X_i y $\lambda(Z_i'\gamma)$ se estimará la especificación (11.9). En esta estimación se utilizarán únicamente las observaciones con variable dependiente disponible (i.e. aquellas para als cuales $s_i = 1$).

Esta última especificación generará estimadores insesgados de β. Puede además utilizarse esta estimación para evaluar si existía sesgo muestral. Para ello simplemente se evalúa si $\rho = 0$, donde ρ es el coeficiente de la variable $\lambda(Z_i'\gamma)$. En los casos en los cuales se rechaza la hipótesis y tenemos evidencia de que $\rho \neq 0$ tendríamos que la estimación de MCO con solo las observaciones que tienen $s_i = 1$ generaría estimadores sesgados de β.

En clase veremos un ejemplo de estos modelos utilizando los siguientes comandos de Stata:

- `webuse womenwk`
- `sum wage education age children married`
- `gen si=(wage$<$.)`
- `probit si education age married children`
- `predict probit_Xb, xb`
- `gen mills=normalden(probit_Xb)/normal(probit_Xb)`
- `reg wage education age mills, r`
- `heckman wage education age, twostep select(education age married children) rhosigma first`

11.3. Métodos de Descomposición

Los métodos de descomposición se desarrollaron en los 70s para cuantificar diferencias promedio en salarios por sexo y determinar qué proporción de dicha diferencia se debe efectivamente a cuestiones de discriminación y qué tanto se puede explicar porque ambos grupos son distintos en diversas características, siendo dichas características (y no la discriminación) las que podría explicar las diferencias entre ambos grupos.

La primera metodología fue propuesta por Ronald Oaxaca (73) y Alan Blinder (73). Conocido como el método Oaxaca-Blinder, consiste en separar la diferencia promedio de una variable dependiente entre dos grupos en dos componentes: (i) la parte *explicada*, que corresponde a la parte de esta diferencia que corresponde a diferencias en las características promedio entre ambos grupos y (ii) la parte *estructural* (o no explicada) que corresponde al remanente de dicha diferencia.

A continuación desarrollamos algo de notación para poder explicar la descomposición Oaxaca-Blinder. Esta notación será nuestra base en el tema de *inverse* probability weight}, que es nuestro objetivo principal. Esta notación se basa en el artículo de Fortin et al. (2011).

11.3.1. Notación general

En este método se hace la comparación entre dos grupos **mutuamente exclusivos**. Dependiendo del contexto, los grupos pueden ser: control-tratamiento, datos disponibles-datos faltantes, no migrantes-migrantes, mujeres-hombres, etc. Denotaremos a los grupos como $g = A, B$. Por lo tanto, podemos defnir una variable dummy para identificar a qué grupo pertenece un individuo i como $D_{Ai} + D_{Bi} = 1$, donde $D_{gi} = \mathbf{1}\{i \in g\}$. Asimismo, volveremos a la definición de resultados potenciales que habíamos discutido en el contexto de experimentos aleatorios. Aquí, Y_{Ai}, Y_{Bi} serán los resultados potenciales, es decir, el nivel de la variable dependiente Y que el individuo i tendrá si pertenece al grupo A y B, respectivamente. Por ende, el resultado observado será: $Y_i = D_{Ai}Y_{Ai} + D_{Bi}Y_{Bi}$. Definiremos al *contrafactual* como el nivel de la variable dependiente que un individuo del grupo B recibiría si *mantuviera sus características*, pero perteneciera al grupo A. Con sus características nos referimos al valor de las variables *observables* (X_i) que utilizamos para predecir o explicar el nivel de la variable dependiente (piensen en las variables explicativas de un MCO). Denotaremos al contrafactual como $Y_{Ai|D_{Bi}}$.

El objetivo de los métodos de descomposición es cuantificar la diferencia en algún estadístico. Priomordialmente, este estadístico es la media (μ), pero con el método de IPW podrá ser cualquier estadístico distribucional, como algún cuantil (τ) o alguna función basada en la distribución, como un índice de Gini. Definimos entonces al estadístico de interés como $\upsilon(F_{Y_{gi}|D_{si}})$, para $g, s = \{A, B\}$. Aquí $F_{Y_{gi}|D_{si}}$ es la distribución del resultado potencial Y_{gi} para individuos del grupo s. Por lo tanto, $F_{Y_{gi}|D_{si}}$ es observado (contrafactual) si $g = s \quad (g \neq s)$. Entonces, el objetivo será cuantificar la diferencia observada del estadístico υ entre ambos grupos:

$$\Delta^{\upsilon} = \upsilon(F_{Y_{Bi}|D_{Bi}}) - \upsilon(F_{Y_{Ai}|D_{Ai}}) \tag{11.11}$$

Los métodos de descomposición dividen esta diferencia observada en dos componentes: (i) Δ^{υ}_{S}, que se define como la *diferencia estructural*, es decir, aquella que se debe a características no observadas o a que las características de un grupo tienen distintos rendimientos que las características del grupo de comparación y (ii) Δ^{υ}_{X}, que la *diferencia observada*, es decir, aquella que resulta de que ambos grupos tienen diferencias en las características explicativas de Y. Para descomponer la diferencia observada Δ^{υ} en ambos componentes utilizamos el contrafactual:

$$\begin{aligned} \Delta^{\upsilon} &= \Big(\upsilon(F_{Y_{Bi}|D_{Bi}}) - \upsilon(F_{Y_{Ai}|D_{Bi}})\Big) + \Big(\upsilon(F_{Y_{Ai}|D_{Bi}} - \upsilon(F_{Y_{Ai}|D_{Ai}})\Big) \\ &= \Delta^{\upsilon}_{S} + \Delta^{\upsilon}_{X} \end{aligned} \tag{11.12}$$

Por lo tanto, el tema de fondo consiste en la forma en que se debe estimar

el contrafactual dado que los otros componentes de la ecuacion ($v(F_{Y_{Bi}|D_{Bi}})$ y $v(F_{Y_{Ai}|D_{Ai}})$) son observados.

11.3.2. Oaxaca-Blinder

La descomposición Oaxaca-Blinder utiliza algunos supuestos para estimar Δ^μ (la diferencia en la media entre ambos grupos). En particular, emplea un modelo lineal que separa los componentes observados y no observados:

$$Y_{gi} = X_i'\beta_g + \epsilon_{gi} \quad , \quad g = A, B \tag{11.13}$$

y supone que los errores son independientes de las variables observadas, $E(\epsilon_{gi}|X_i) = 0$.

Dados estos supuestos, utilizando la ley de esperanzas iteradas (LIE), desarrolla Δ^μ de la siguiente forma:

$$\begin{aligned}
\Delta^\mu &= E(Y_{Bi}|D_{Bi}) - E(Y_{Ai}|D_{Ai}) \\
&= E[E(Y_{Bi}|X_i, D_{Bi})|D_{Bi}] - E[E(Y_{Ai}|X_i, D_{Ai})|D_{Ai}]] \\
&= [E(X_i|D_{Bi})'\beta_B + E(\epsilon_{Bi}|X_i, D_{Bi})] - [E(X_i|D_{Bi})'\beta_A + E(\epsilon_{Ai}|X_i, D_{Ai})] \pm E(X_i|D_{Bi})'\beta_A \\
&= \underbrace{E(X_i|D_{Bi})'(\beta_B - \beta_A)}_{\Delta_S^\mu} + \underbrace{[E(X_i|D_{Bi}) - E(X_i|D_{Ai})]\beta_A}_{\Delta_X^\mu}
\end{aligned}$$

Podemos entonces utilizar el siguiente estimador y derivar el valor estimado con la contraparte muestral:

$$\widehat{\Delta}^\mu = \overline{X}_B(\hat{\beta}_B - \hat{\beta}_A) + (\overline{X}_B - \overline{X}_A)\hat{\beta}_A \tag{11.14}$$

Para llevar a cabo esta estimación pueden estimarse dos MCO (uno para cada grupo) y posteriormente hacer los calculos o utilizar el comando oaxaca desarrollado por Jann (2008).

11.3.3. Inverse Probability Weight (IPW)

El método IPW es un método de descomposición en el cual se genera toda la distribución contrafactual, con lo cual se puede calcular cualquier estadístico que utilice como insumo la distribución acumulada. La notación presentada aquí se desarrolla para cuantiles (τ). Por lo tanto, empezamos por definir a un cuantil en este contexto como $Q_{g,\tau}$ y se propone estimarlo utilizando la ley de las probabilidades iteradas:

$$\begin{aligned}\tau &= F_{Y_g}(Q_{g,\tau}) \\ &= E\left(F_{Y_g|X_{gi}}(Q_{g,\tau}|X_{gi})\right) \\ &= \int F_{Y_g|X_{gi}}(Q_{g,\tau}|X)dF_{X_{gi}}(X), \quad g = A, B\end{aligned} \tag{11.15}$$

En este caso, si quisieramos estimar la diferencia entre el cuantil τ para ambos grupos y descomponer dicha diferencia nos interesaría:

$$\Delta^\tau = \left(F^{-1}_{Y_{Bi}|D_{Bi}}(\tau) - F^{-1}_{Y_{Ai}|D_{Bi}}(\tau)\right) + \left(F^{-1}_{Y_{Ai}|D_{Bi}}(\tau) - F^{-1}_{Y_{Ai}|D_{Ai}}(\tau)\right) \tag{11.16}$$

Por lo tanto, el componente que necesitamos estimar es el contrafactual $F^{-1}_{Y_{Ai}|D_{Bi}}(\tau)$. Para llevar a cabo esto, el método IPW utiliza la siguiente estrategia:

$$\begin{aligned}F_{Y_{Ai}|D_{Bi}}(y) &= \int F_{Y_A|X_{Ai}}(y|X)dF_{X_{Bi}}(X) \\ &= \int F_{Y_A|X_{Ai}}(y|X)\Psi(X)dF_{X_{Ai}}(X)\end{aligned} \tag{11.17}$$

Esto hace que la distribución contrafactual sea simplemente una versión reponderada de la distribución original de Y para el grupo A, donde el reponderador es:

$$\Psi(X) = \frac{dF_{X_{Bi}}(X)}{dF_{X_{Ai}}(X)} \tag{11.18}$$

DiNardo et al. (1996) propusieron este estimador y sugirieron utilizar la regla de Bayes:

$$Pr(X|D_{Bi}) = \frac{Pr(D_{Bi}|X_i)}{Pr(D_{Bi})} \tag{11.19}$$

Con lo cual:

$$\Psi(X) = \frac{Pr(D_{Bi}|X_i)/Pr(D_{Bi})}{Pr(D_{Ai}|X_i)/Pr(D_{Ai})} \tag{11.20}$$

De esta forma $\Psi(X)$ se puede estimar utilizando probit o logit y las proporciones de cada grupo, con lo cual tendríamos el estimador que buscamos. Por último, cabe señalar que en el caso de datos faltantes, no contamos con Y_B observada (asumiendo que B es el grupo no observado), pero si contamos con caracterísitcas de este grupo, es decir, X_B. En este caso, la intuición es que las características del grupo A se reponderan para que su distribución sea la misma que las del grupo By utilizando este ponderador se usan las Y_A para generar la distribución del grupo B.

Apéndice A

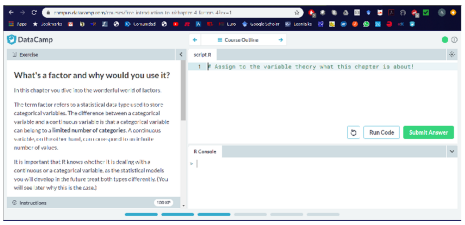
Introdución a R

A.1. Ventajas de R

- Open Source (gratuito)
- Usado por muchos desarrolladores y en muchos trabajos, especialmente aquellos intensivos en "data analysis"
- Sustituto perfecto de `Stata` aunque es menos amigable
- Visualización más flexible que `Stata`
- Es más eficiente en algunas funcionalidades como uso de datos geográficos

A.2. DataCamp

- DataCamp es una plataforma que contiene cursos y contenido para aprender `R` y otros softwares de Data Analysis (ej. `Python`)
- Recibirán una invitación a su correo del ITAM. Tendrán acceso a los cursos por 6 meses. El uso de DataCamp es opcional para el curso.
- Al usar DataCamp podrán practicar como programar en `R` y el sistema les da feedback inmediato
- Para usarlo no es necesario instalar `R` en su computadora, aunque para las Tareas sí lo necesitarán



A.3. Instalando R

1. Entrar a CRAN e instalar R siguiendo el link relevante del cuadro que indica "Download and install R"
   - Windows: seguir los links "install R for first time" y "Download R 4.0.2 for Windows. Una vez instalado, regresar a la página donde aparecía el link de "install R for first time" y elige el link que dice "Rtools". En la nueva página elige "Rtools40.exe" e instala.
   - Mac: seguir el link "R-4.0.2.pkg" (notarized and signed)
2. Entrar a RStudio. Bajen en la página hasta encontrar la sección "All Installers" y ahí dependiendo si tienen Windows o Mac eligen descargar el archivo "RStudio-1.3.1056.exe" (Windows) o "RStudio-1.3.1056.dmg" y siguen las instrucciones de instalación.

En DataCamp hay un tutorial para instalar R y RStudio también

A.3.1. R y RStudio

- R es un lenguaje de programación con enfoque estadístico y matricial. R ejecuta todas las operaciones que le indicaremos.
- RStudio es una interfaz que nos da flexibilidad al realizar distintas tareas. Podemos integrar todo nuestro proceso de trabajo en el mismo ambiente.
- Para utilizar RStudio es necesario descargar primero R.

A.3.2. R Packages

- A diferencia de Stata, R no tiene los comandos en menús y su uso si requiere del conocimiento del nombre de los comandos relevantes
- Los packages representan un conjunto de funciones, datos y comandos de R.

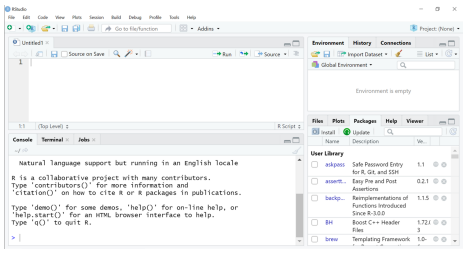


Figura A.1: Pantalla principal de RStudio

- Para instalar un nuevo package se utiliza el comando `install.packages('nombre del package')`
- Para utilizar las funciones que incluye dicho package hay que cargarlos en la sesión utilizando el comando `library(package)`
- Se recomienda que se carguen todos los packages que se utilizaran en la sesión al inicio del script de `R`

A.3.3. Abriendo bases de datos

- Utilizar el comando `setwd` para indicar el directorio de trabajo donde pondrán la base de datos por abrir (ojo con la dirección de las diagonales "`/`")
  - Ej. `setwd(Ç:/Users/aaeag/Microeconometria_aplicada/Bases Stata")`
- Para abrir bases en formato:
  - Stata -> utilizamos `library(haven)` -> comando `read_dta("nombre_archivo")`
  - CSV -> comando `read.csv("nombre_archivo")`
  - Excel -> utilizamos `library(gdata)` -> comando `read.xlsx("nombre_archivo")`

En el script de ejemplo se muestra como abrimos una base de `Stata`

Otra alternativa consiste en utilizar la ventana superior derecha, pestaña "`Environment`", botón "`Import Dataset`"

Apéndice B

Métodos Experimentales y Cuasi-experimentales

B.1. Motivación

- Evidence-based programs
- Statistical methods:
  - RCTs – promoted by agencies (e.g. J-PAL)
  - Quasi-experiments – look for situations "simulating" an experiment
- Increased access to data and measurement of concepts

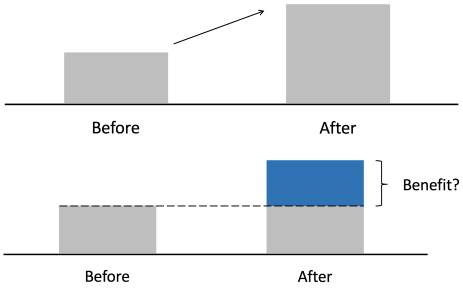
B.2. RCTs in practice

Evidence-based science

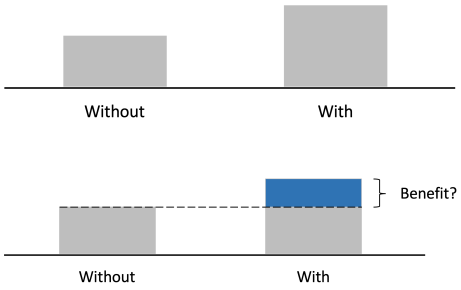
B.3. Some ideas

Objective: Measure the benefits (or lack of) that the program gives

- **Approach #1:** Before versus after



- **Approach #2:** Compare people with and without the program

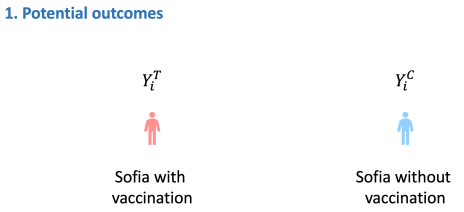


B.4. Statistical concepts

Objective: Measure the benefits (or lack of) that the program gives

Ideal approach: Compare the same person with and without the program

1. **Potential Outcomes**



2. **Treatment Effect**

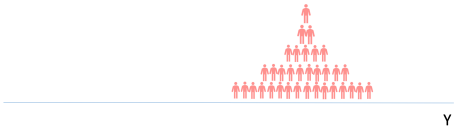
$$TE = Y_i^T - Y_i^C$$

However… we can only observe "one" version of Sofia, not both!

Let's imagine now that we compare health (Y) in two alternative worlds:

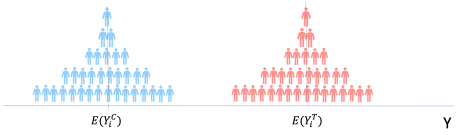
- No one gets vaccination: everyone gets their Y_i^C level of health

- Everyone gets vaccination: everyone gets their Y_i^T level of health



If we could look at both "versions" of the world we could calculate:

Average Treatment Effects: $ATE = E[Y_i^T] - E[Y_i^C]$

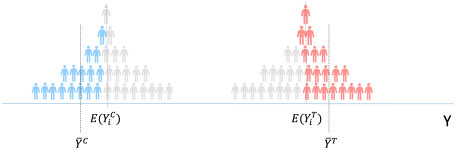


Now it looks impossible + extremely costly

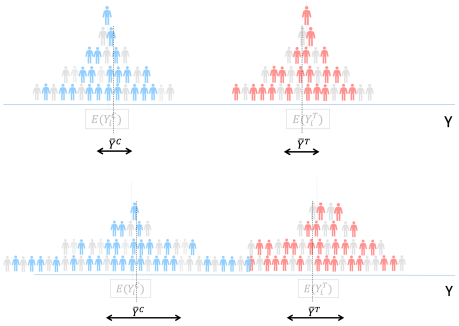
But, what if we take a sample instead of looking at "everyone"

We can still get very close to $E[Y_i^T]$ and $E[Y_i^C]$ if we choose "wisely"

Problem with self-selection $\Longrightarrow$ **Bias**



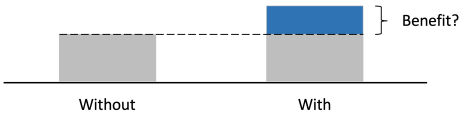
Standard errors. In this case, $\bar{Y}^T - \bar{Y}^C$ gives us an "estimate" of ATE. Thus we, need to create intervals in which ATE could be.



B.5. RCTs

Objective: Measure the benefits (or lack of) that the program gives

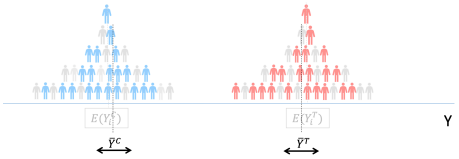
- Solution: Randomization



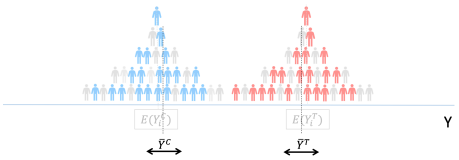
Challenges found on RCTs:

- Externalities
- John Henry and Hawthrone effects
- Attrition
- Partial participation

Externalities. Mean that it is not possible to observe "control" in a pure form



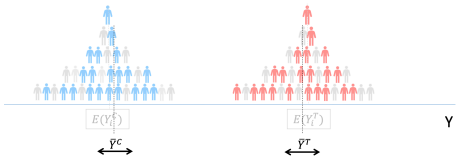
Externalities. Depending on the type of externality, we would be over- or under-estimating ATE



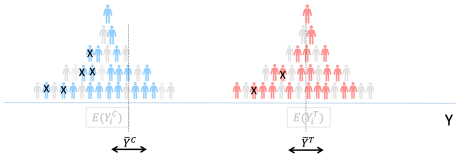
John Henry effects: a legendary American steel driver in the 1870s who, when he heard his output was being compared with that of a steam drill, worked so hard to outperform the machine that he died in the process.

Hawthrone effects: Hawthorne Works (Western Electric factory outside Chicago) commissioned a study to see if their workers would become more productive in higher or lower levels of light. The workers'productivity seemed to improve when changes were made, and slumped when the study ended.

Attrition. Need to be careful with "selective attrition"



Attrition. Could result from deaths, migration, unwillingness to continue participating



Partial participation: treatment is not always enforceable

		Not Assigned to Treatment	
		Doesn't take Treatment	Takes Treatment
Assigned to Treatment	Doesn't take Treatment	Never-taker	Defier
	Takes Treatment	Complier	Always-taker

Partial participation

$$Y_1 = \gamma_0 + \gamma_1 Z_i + U_i$$
$$T_i = \eta_0 + \eta_1 Z_i + V_i$$

LATE (ATE on compliers)

$$\frac{\partial Y}{\partial T} = \frac{\frac{\partial Y}{\partial Z}}{\frac{\partial T}{\partial Z}} = \frac{\gamma_1}{\eta_1}$$

RCT in practice

1. Identify the causes of a problem that you want to solve

   -Complement this with field work

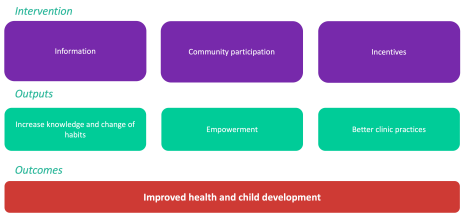


Figura B.1: Theory of change

2. Build a theory of change
3. Associate your theory of change with a list of indicators

Ouput / outcome	Measure	Source	Frequency
Knowledge	Questions asked to women about child care	SMS questions	Measured once
Empowerment	Instrument to ask women	Survey at clinics	Measured once
Better clinic practices	Satisfaction questionnaire	Administrative data	Measured each time they visit their clinic
Improved child health	Length at birth	Vital statistics	On date of birth

Figura B.2: List of indicators

Trial registry

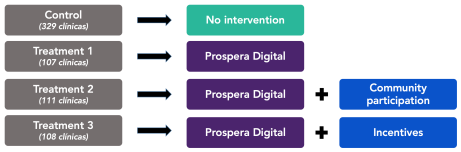
1. `Identify the causes of a problem that you want to solve`
2. `Build a theory of change`
3. `Associate your theory of change with a list of indicators`
4. `IRB / Ethics board (example)`
5. `Pre-pilot`
6. `Monitor / FOI`
   - Selecting your level of treatment

     -Individual level

     -Group level

     -Locality level
   - Selecting your sample size
   - What to do if you have several “treatment arms”?
   - Setup your monitoring tools

Example:



B.6. Quiz

- In an educational project some treatment schools decide after you randomize that they cannot participate. Your implementing partner wants to substitute those schools with new schools from the control which really want to receive treatment.
- Should you allow for this?
- In your financial intervention you gave debit cards to people in treatment localities. You realize that some people in the treatment are not using the debit card at all.
- Should you define them as part of the treatment or the control when making your analysis?
- In a health intervention you are giving workshops to patients as part of an intervention. Your intervention takes place in two States (one poor and one rich), half of the clinics in each State is treatment and the other half is control.
- Your implementing partner tells you that it is very inefficient and expensive doing half and half. That they can only afford to do one full State treatment and the other control. If this is a take-it or leave-it situation, what would you do?
- In a gender violence project you are giving one-to-one advisories to victims. Your treatment is distributed at the locality level. Some of your beneficiaries seem to be improving and ask you that they want to invite their friends in the community who have been also victims, but are not receiving "treatment".
- Should you allow for this?
- Quasi-experiments
    - Difference-in-difference
    - Regression discontinuity
    - Natural experiments
    - Matching

- Objective: Try to find a credible scenario for the "control"

Referencias

- Abadie, A. (2020). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Article prepared for the Journal of Economic Literature*.
- Almond, D. (2020). Is the 1918 Influenza Pandemic Over? Long-Term Effects of In Utero Influenza Exposure in the Post-1940 U.S. Population. *Journal of Political Economy 114* (4), 672-712.
- Angrist, J. D. and W. N. Evans (1998). Children and Their Parents'Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review 88* (3), 450-477.
- Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*. Princeton University Press.
- Athey, S. and G. W. Imbens (2017). Chapter 3 The Econometrics of Randomized Experiments. *Handbook of Economic Field Experiments 1*, 73-140.
- Athey, S. and G. W. Imbens (2020). Machine Learning Methods Economists Should Know About. *Journal of Political Economy 114* (4), 672-712.
- Belles, C. and M. Lombardi (2020). Will you marry me, later? Age-of-marriage laws and child marriage in Mexico. *The Journal of Human Resources*, 1219-10621R2.
- Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review 94* (4), 991-1013.
- Bharadwaj, P., K. V. Loken, and C. Neilson (2013). Early Life Health Interventions and Academic Achievement. *American Economic Review 103* (5), 1862-1891.
- Caliendo, M. (2008). Some Practical Guidance For The Implementation Of Propensity Score Matching. *Journal of Economic Surveys 22* (1), 31-72.
- Card, D. and A. B. Krueger (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review 84* (4), 672-712.

- Cattaneo, M. D., N. Idrobo, and R. Titiunik (2018a). A Practical Introduction to Regression Discontinuity Designs: Volume I. *Monograph prepared for Cambridge Elements: Quantitative and Computational Methods for Social Science.*
- Cattaneo, M. D., N. Idrobo, and R. Titiunik (2018b). A Practical Introduction to Regression Discontinuity Designs: Volume II. *Monograph prepared for Cambridge Elements: Quantitative and Computational Methods for Social Science.*
- Davis, L. W. (2008). The Effect of Driving Restrictions on Air Quality in Mexico City. *Journal of Political Economy 116* (1), 38-81.
- de Laat, J. (2015). Matching techniques. Impact Evaluation Research Workshop - World Bank.
- Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review 91* (4), 795-813.
- Duflo, E., R. Glennerster, and M. Kremer (2007). Using Randomization in Development Economics Research: A Toolkit. *Handbook of Development Economics 4*, 3895-3962.
- Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. *The Journal of Human Resources 33* (2), 251-299.
- Heckman, J. J. and E. J. Vytlacil (2007a). Chapter 70 Econometric Evaluation Of Social Programs,Part I: Causal Models, Structural Models And Econometric Policy Evaluation. *Handbook of Econometrics Volume 6* (B), 4779-4874.
- Heckman, J. J. and E. J. Vytlacil (2007b). Chapter 71 Econometric Evaluation Of Social Programs,Part II: Using The Marginal Treatment Effect To Organize Alternative Econometric Estimators To Evaluate Social Programs, And To Forecast Their Effects In New Environments. *Handbook of Econometrics Volume 6* (B), 4875-5143.
- Heckman, J. J. and E. J. Vytlacil (2007c). Chapter 72 Econometric Evaluation Of Social Programs,Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, And General Equilibrium Policy Evaluation. *Handbook of Econometrics Volume 6* (B), 5145-5303.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient Estimation Of Average Treatment Efects Using The Estimated Propensity Score. *Econometrica 71* (4), 1161-1189.
- Imbens, G. W. (2014). Matching Methods in Practice: Three Examples. *The Journal Of Human Resources 50* (2), 373-419.

- Imbens, G. W. (Fall 2007). Nonparametric Density Estimation. Department of Economics - Harvard University.
- Lee, D. S. (2002). Trimming for Bounds on Treatment Effects with Missing Outcomes. *NBER Technical Working Paper* (277).
- Millan, T. M. and K. Macours (2017). Attrition in randomized control trials: Using tracking information to correct bias. NOVAFRICA Working Paper Series wp1702, Universidade Nova de Lisboa, Faculdade de Economia, NOVAFRICA.
- Mullainathan, S. and J. Spiess (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives 31* (2), 87-106.
- Stock, J. and M.Watson (2011). *Introduction to Econometrics* (3 ed.). Addison Wesley.
- Wooldridge, J. M. (2012). *Introductory Econometrics. A Modern Approach* (5 ed.). Southwestern Cengage Learning.
- Zambom, A. Z. and R. Dias (2012). A review of kernel density estimation with applications to econometrics. Universidade Estadual de Campinas.