

# Análisis de Datos Longitudinales

*Carlos Javier Rincón*

*versión:2019*

## Contents

<b>Introducción</b>	<b>3</b>
Métodos iniciales. . . . .	4
<b>Anova con medidas repetidas</b>	<b>13</b>
Aplicación. . . . .	16
<b>Manova con medidas repetidas</b>	<b>23</b>
Aplicación. . . . .	24
<b>Estimación de ecuaciones generalizadas - GEE</b>	<b>31</b>
Modelos lineales generalizados (MLG) . . . . .	31
GEE - Desenlace continuo. . . . .	32
Prueba Wald sobre los parámetros del modelo. . . . .	34
Aplicación. . . . .	34
<b>Modelos GEE - Desenlace dicotómico y de conteo.</b>	<b>44</b>
Desenlace Dicotómico . . . . .	44
Desenlace de conteo . . . . .	44
Aplicación. . . . .	45
<b>Modelos mixtos (MIX) - Desenlace continuo.</b>	<b>55</b>
Aplicación . . . . .	58
<b>Modelos mixtos (MIX) - Desenlace dicotómico</b>	<b>68</b>
Aplicación. . . . .	69
<b>Modelos mixtos (MIX) - Desenlace de conteo.</b>	<b>76</b>
Modelo de Poisson con afectos aleatorios y cero inflado (Mix-ZIP) . . . . .	77
Aplicación . . . . .	77



Introducción al análisis de datos longitudinales basado en el libro: **Longitudinal data analysis** de **Donald Hedeker** y **Robert D. Gibbons**. Se incluyen aplicaciones para cada uno de los temas tratados utilizando el programa **R**. Se asume un conocimiento básico por parte del lector en conceptos básicos de inferencia, modelos de regresión lineal y el uso del programa R y RStudio.

---

## Introducción

Este texto corresponde a las notas de clase de la asignatura *análisis de datos longitudinales*, donde se realiza una revisión de distintos métodos que son apropiados cuando los datos surgen de la medición de un desenlace de forma repetida sobre la unidad de análisis bajo estudio, típicamente un sujeto. Bajo la definición anterior, las mediciones obtenidas están correlacionadas entre las distintas repeticiones pero son independientes entre las unidades de análisis; es decir que los métodos basados en el supuesto de independencia no son válidos y se debe recurrir a otras herramientas, algunas que se presentarán a continuación. Adicionalmente, se trabajarán situaciones donde las características del sujetos o sus exposiciones también varíen en un periodo de tiempo.

Entre las ventajas del análisis de datos longitudinales, en comparación a realizar una sola medición del desenlace (estudios de corte transversal), tenemos:

1. Los estudios longitudinales tienen un mayor poder que los estudios de corte transversal para un número fijo de sujetos observados.
2. Si se considera a cada sujeto como su propio control (ej: estudios cruzados) y dado que la variabilidad intra sujeto es menor que la variabilidad entre sujetos, se obtienen pruebas estadísticas con mayor poder para evaluar el efecto de una exposición.
3. Permiten separar el efecto de la edad (cambios en el sujeto al pasar el tiempo) del efecto de la cohorte (diferencias entre los sujetos al inicio de su medición - línea base).
4. Pueden dar información del cambio del desenlace a nivel individual.

Entre los retos que presenta este tipo de análisis, se encuentran:

1. Requieren de métodos que consideren la correlación de las observaciones, siendo estos generalmente más complejos y computacionalmente más demandantes.
  2. Riesgo de presentar datos faltantes ya sea porque no se obtuvieron todas las mediciones programadas en un sujeto, o por abandonos o retiros del estudio antes de finalizar el seguimiento programado.
  3. Debido a la necesidad de recolección de información en distintos momentos, no solo del desenlace sino en ocasiones también de los factores de interés, se requiere de diseños de investigación que generan un mayor costo y complejidad operativa.
  4. Se debe considerar, en el caso de estudios experimentales donde se aplican varias intervenciones al mismo sujeto, el orden en que estas son aplicadas. Lo anterior ya que de esto puede depender el resultado de la relación que se está evaluando.
- Organización de los datos.

Cuando se realiza un estudio con datos longitudinales, las observaciones obtenidas se pueden organizar de dos formas o arreglos. El primer arreglo se denomina **a lo largo** o **univariado**, donde se definen dos columnas que identifican a los sujetos ( $i = 1, \dots, N$ ) y el tiempo de medición ( $j = 1, \dots, n$ ) y  $p + 1$  columnas adicionales que denotan los resultados obtenidos del desenlace  $y_{ij}$  y las  $p$  ( $k = 1, 2, \dots, p$ ) variables independientes. Un arreglo en esta estructura se representa así:

id	tiempo	$y_{ij}$	$x_{ij1}$	..	$x_{ijk}$	..	$x_{ijp}$
1	1	$y_{11}$	$x_{111}$	..	$x_{11k}$	..	$x_{11p}$
:	:	:	:	:	:	:	:
1	n	$y_{1n}$	$x_{1n1}$	..	$x_{1nk}$	..	$x_{1np}$
:	:	:	:	:	:	:	:
i	j	$y_{ij}$	$x_{ij1}$	..	$x_{ijk}$	..	$x_{ijp}$
:	:	:	:	:	:	:	:
N	1	$y_{N1}$	$x_{N11}$	..	$x_{N1k}$	..	$x_{N1p}$
:	:	:	:	:	:	:	:
N	n	$y_{Nn}$	$x_{Nn1}$	..	$x_{Nnk}$	..	$x_{Nnp}$

La segunda forma de arreglo se denomina **a lo ancho** o **multivariado** donde cada fila corresponde a un

sujeto ( $i = 1, 2, \dots, N$ ) y se coloca una columna por cada combinación de tiempo ( $j = 1, 2, \dots, n$ ) y variable (independiente y dependiente) observada ( $k = 1, 2, \dots, p$ ). Se visualiza de la siguiente forma:

id	$y_1$	..	$y_n$	$x_{11}$	..	$x_{jk}$	..	$x_{n1}$	..	$x_{np}$
1	$y_{11}$	..	$y_{1n}$	$x_{111}$	..	$x_{1jk}$	..	$x_{1n1}$	..	$x_{1np}$
2	$y_{21}$	..	$y_{2n}$	$x_{211}$	..	$x_{2jk}$	..	$x_{2n1}$	..	$x_{2np}$
:	:	:	:	:	:	:	:	:	:	:
i	$y_{i1}$	..	$y_{in}$	$x_{i11}$	..	$x_{ijk}$	..	$x_{in1}$	..	$x_{inp}$
:	:	:	:	:	:	:	:	:	:	:
N	$y_{N1}$	..	$y_{Nn}$	$x_{N11}$	..	$x_{Njk}$	..	$x_{Nn1}$	..	$x_{Nnp}$

Sobre las dos representaciones de datos anteriores, se debe considerar que también se pueden incluir variables independientes “fijas”, no cambian en el tiempo como sexo o raza, y que en ocasiones el número de medidas repetidas es diferente por cada sujeto ( $n_i \neq n$ : datos des-balanceados).

Finalmente, aunque las dos estructuras contienen la misma información, es importante reconocerlas ya que dependiendo el programa estadístico que se utilice, algunos procedimientos o funciones implementadas trabajan con una u otra forma de arreglo. En consecuencia, una habilidad importante que se requiere es poder intercambiar los datos entre los dos tipos de arreglos.

## Métodos iniciales.

Algunos métodos típicos para el análisis de datos longitudinales son:

1. Visualización de los datos a partir de gráficos y
2. Aplicación de pruebas (paramétricas y no paramétricas) apropiadas, tales como:
  - t pareada, prueba del signo, prueba de suma de rangos del signo de Wilcoxon.
  - Intervalo de confianza para estimar la diferencia de proporciones para datos pareados, prueba de McNemar.
  - Prueba de Friedman y Q de Cochran.

A continuación se presenta la aplicación de estos métodos, utilizando el siguiente ejemplo:

Escenario: Se aplicó una escala que mide la susceptibilidad a la hipnosis (a mayor puntaje mayor susceptibilidad) en 30 sujetos en tres momentos diferentes. De estos 30 sujetos, 15 recibieron un entrenamiento A y los 15 restantes un entrenamiento B. Los datos se encuentran en el archivo `intro.xlsx`.

- Vista de los primeros 5 sujetos:

```
library(rio)
d<-import("intro.xlsx");d[1:5,]
```

```
##   id antes despues final int
## 1  1    18      10     8   A
## 2  2    19      16    12   A
## 3  3    11       7     6   A
## 4  4     3       7     5   A
## 5  5     5       2     3   A
```

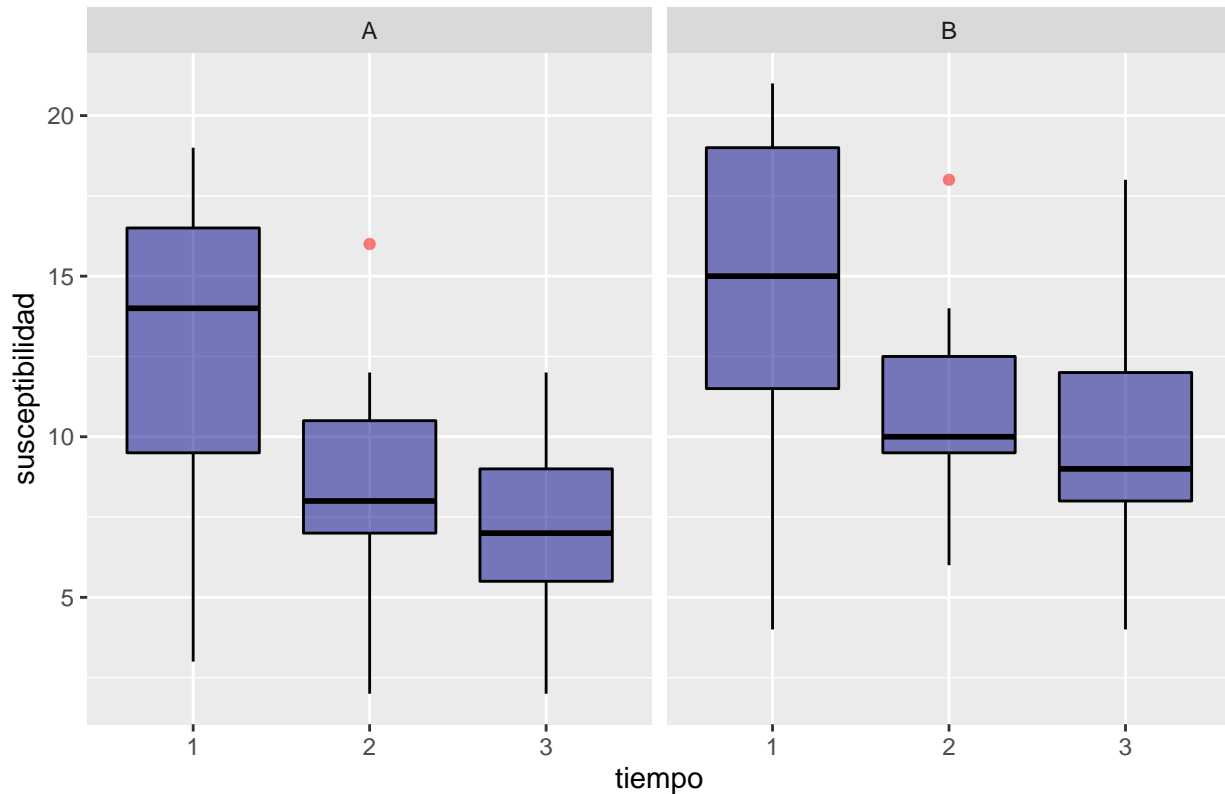
- Algunos gráficos:

*Boxplot por tiempo y entrenamiento:*

```
library(tidyverse)
dl<-reshape(d,idvar = "id",varying = list(2:4),direction = "long",v.names = "susc")
```

```
ggplot(dl,aes(x=factor(time),y=susc))+
  geom_boxplot(alpha=0.5,fill="darkblue", color="black",outlier.color="red")+
  labs(title="Boxplot susceptibilidad por tiempo y entrenamiento",x="tiempo",
        y="susceptibilidad")+
  facet_wrap(~int)
```

Boxplot susceptibilidad por tiempo y entrenamiento

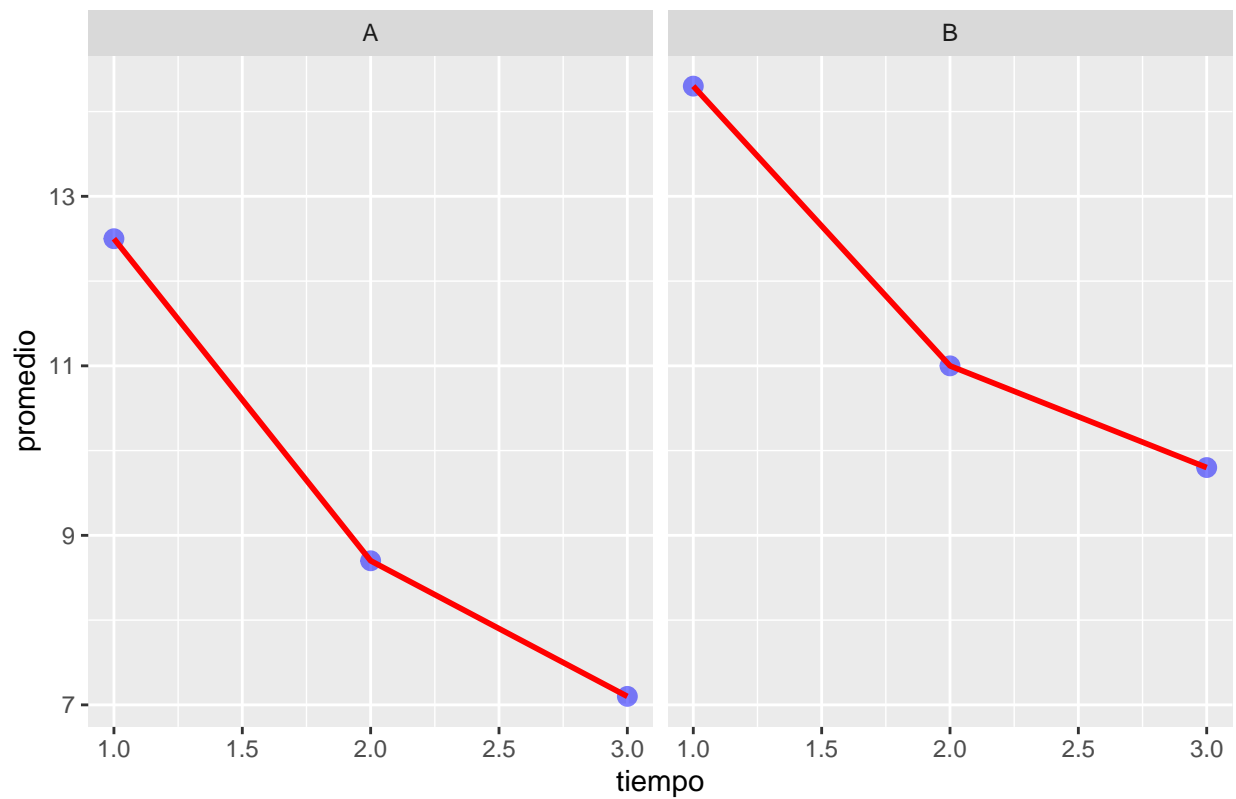


*Promedio del desenlace por cada tiempo y entrenamiento:*

```
g<-group_by(dl,int,time)
g2<-summarise(g,media=round(mean(susc),1))

ggplot(g2,aes(x=time,media))+
  geom_point(alpha=0.5,color="blue",size=3)+
  geom_line(color="red",size=1)+
  labs(title = "Promedio de susceptibilidad por tiempo y entrenamiento",
        x="tiempo",y="promedio")+
  facet_wrap(~int)
```

## Promedio de susceptibilidad por tiempo y entrenamiento



En el caso de un pequeño número de sujetos, se puede ver las tendencias de cada sujeto:

```
ggplot(dl, aes(x=time, y=susc, group=id))+  
  geom_point()+  
  geom_line(color="blue")+  
  labs(title="Tendencias del sujeto por entrenamiento", x="tiempo",  
        y="susceptibilidad")+  
  facet_wrap(~int)
```

## Tendencias del sujeto por entrenamiento

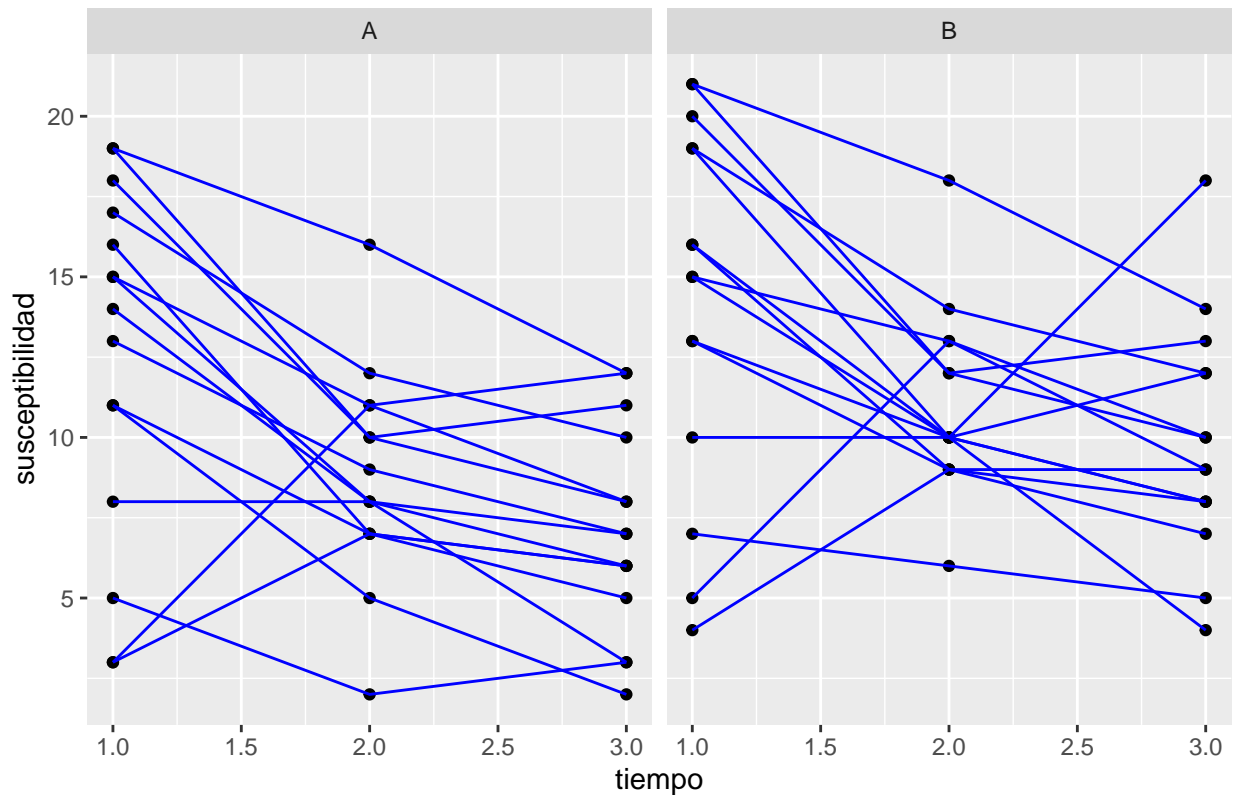


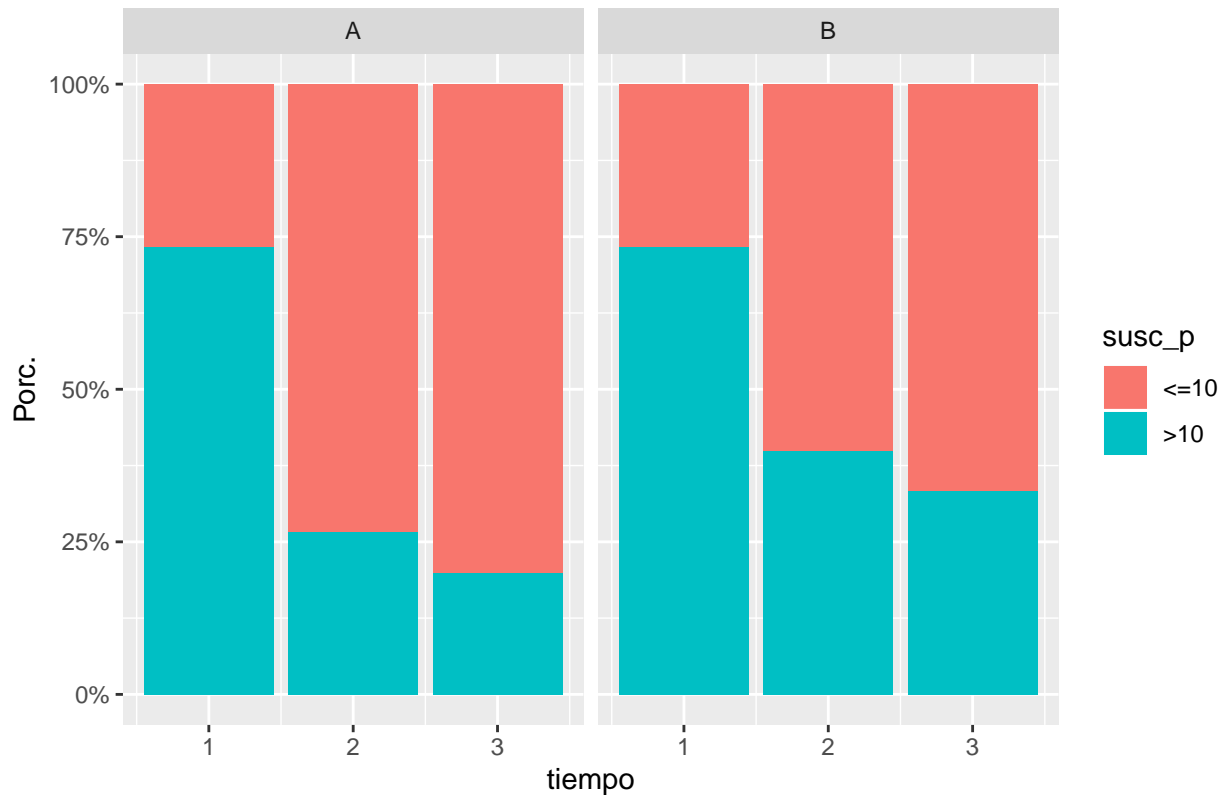
Gráfico de barras apiladas - desenlace dicotómico:

(Se define dos grupos tomando como punto de corte un puntaje de 10;  $> 10$ : susceptibles,  $\leq 10$ : no susceptibles).

```
dl<-mutate(dl,susc_p=ifelse(susc>10,">10","<=10")) # Categorización de una variable.

ggplot(dl, aes(x=time, fill = susc_p)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent)+
  labs(title = "Porcentaje de sujetos por grupos, tiempo y tratamiento",
        y="Porc.", x="tiempo")+
  facet_wrap(~int)
```

Porcentaje de sujetos por grupos, tiempo y tratamiento



- Pruebas estadísticas:

Para evaluar si existen diferencias entre la primera y última medición, se podrían aplicar las siguientes pruebas según el cumplimiento de supuestos: *t pareada*, la prueba *del signo* y la prueba de *suma de rangos del signo de wilcoxon*. Se presenta el código a continuación:

```
t.test(d$antes,d$final,paired = TRUE) # t pareada
```

```
##
## Paired t-test
##
## data: d$antes and d$final
## t = 5.7435, df = 29, p-value = 3.226e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 3.176604 6.690063
## sample estimates:
## mean of the differences
## 4.933333
```

```
library(BSDA) # paquete de la función SIGN.test
SIGN.test(d$antes,d$final) # prueba de signo
```

```
##
## Dependent-samples Sign-Test
##
## data: d$antes and d$final
## S = 25, p-value = 0.0003249
```



```
## alternative hypothesis: true median difference is not equal to 0
## 95 percent confidence interval:
## 3.129122 8.000000
## sample estimates:
## median of x-y
## 7
##
## Achieved and Interpolated Confidence Intervals:
##
## Conf.Level L.E.pt U.E.pt
## Lower Achieved CI 0.9013 4.0000 8
## Interpolated CI 0.9500 3.1291 8
## Upper Achieved CI 0.9572 3.0000 8
wilcox.test(d$antes,d$final,paired = TRUE,exact = FALSE,correct = TRUE)
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data: d$antes and d$final
## V = 425.5, p-value = 7.328e-05
## alternative hypothesis: true location shift is not equal to 0
# prueba de suma de rangos del signo de wilcoxon
```

Para evaluar si existen diferencias entre las tres mediciones de la escala, se utilizar la prueba de *Friedman*:

```
friedman.test(susc~time|id,d1) # Prueba de Friedman
```

```
##
## Friedman rank sum test
##
## data: susc and time and id
## Friedman chi-squared = 25.453, df = 2, p-value = 2.971e-06
```

Para evaluar si existen diferencias entre la proporción de puntaje de susceptibilidad (puntajes >10) en la primera y última medición podemos estimar (intervalos de confianza) la diferencia de proporciones para datos emparejados o aplicar la prueba de McNemar.

```
dw<-reshape(d1,timevar = "time",idvar = "id",direction = "wide")
dw$int.2<-NULL
dw$int.3<-NULL
```

```
table(dw$susc_p.1,dw$susc_p.3)
```

```
##
## <=10 >10
## <=10 8 0
## >10 14 8
```

*# tabla de frecuencias a dos vías con las variables antes y después.*

```
prop.table(table(dw$susc_p.1))
```

```
##
## <=10 >10
## 0.2666667 0.7333333
```

```
prop.table(table(dw$susc_p.3))
```

```
##
##      <=10      >10
## 0.7333333 0.2666667
library(PropCIs)
#Paquete para estimar diferencias de proporciones para datos emparejados.
diffpropci.mp(0,14,30,conf.level = 0.95) #Intervalo con el ajuste del Wald.

##
##
##
## data:
##
## 95 percent confidence interval:
## 0.2550339 0.6199661
## sample estimates:
## [1] 0.4375
mcnemar.test(dw$susc_p.1,dw$susc_p.3) #Prueba de McNemar.
```

```
##
## McNemar's Chi-squared test with continuity correction
##
## data: dw$susc_p.1 and dw$susc_p.3
## McNemar's chi-squared = 12.071, df = 1, p-value = 0.000512
```

Para realizar la comparación entre las tres mediciones de la escala, con respuesta dicotómica, se puede utilizar la prueba Q de Cochran.

```
library(RVAideMemoire)# paquete de la prueba q de cochran
cochran.qtest(susc_p~time|id,dl) # prueba q de cochran
```

```
##
## Cochran's Q test
##
## data: susc_p by time, block = id
## Q = 18.1053, df = 2, p-value = 0.0001171
## alternative hypothesis: true difference in probabilities is not equal to 0
## sample estimates:
## proba in group      <NA>      <NA>
##      0.7333333      0.3333333      0.2666667
```

- Modelos de regresión.

Los modelos de regresión lineal parte del supuesto de independencia por los cual no se pueden aplicar en este tipo de datos, pero existen algunas alternativas realizando una modificación de los datos observados donde no se viole el supuesto de independencia. A continuación se presentan algunas alternativas para el caso de dos medidas repetidas y una intervención con opción de respuesta dicotómica.

### Change Score Analysis

Modelo:

$$d_i = \beta_0 + \beta_1 x_1 + e_i$$

donde  $d_i = y_{i2} - y_{i1}$  y  $x_1$  es igual a 1 si recibe tratamiento y 0 si pertenece al grupo control.

Ejemplo:

```
CSA<-lm(final-antes-int,d)
summary(CSA)
```

```
##
## Call:
## lm(formula = final - antes ~ int, data = d)
##
## Residuals:
##   Min     1Q   Median     3Q      Max
## -5.600 -3.533 -1.600  2.467 10.400
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.4000     1.2299  -4.391 0.000146 ***
## intB         0.9333     1.7394   0.537 0.595786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.763 on 28 degrees of freedom
## Multiple R-squared:  0.01018, Adjusted R-squared:  -0.02517
## F-statistic: 0.2879 on 1 and 28 DF, p-value: 0.5958
```

*Ancova Post-test Score*

Modelo:

$$y_{i2} = \beta_0 + \beta_1 x_1 + \beta_2 y_{i1} + e_i$$

Ejemplo:

```
APS<-lm(final~int+antes,d)
summary(APS)
```

```
##
## Call:
## lm(formula = final ~ int + antes, data = d)
##
## Residuals:
##   Min     1Q   Median     3Q      Max
## -5.3967 -1.8273 -0.3166  1.4563  7.9665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0969     1.4603   2.121 0.04327 *
## intB         2.1602     1.0892   1.983 0.05759 .
## antes        0.3184     0.1000   3.183 0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.942 on 27 degrees of freedom
## Multiple R-squared:  0.3808, Adjusted R-squared:  0.3349
## F-statistic: 8.303 on 2 and 27 DF, p-value: 0.001547
```

*Ancova Change Score*

Modelo:

$$d_i = \beta_0 + \beta_1 x_1 + \beta_2 y_{i1} + e_i$$

Ejemplo:

```
ACS<-lm(final-antes~int+antes,d)
summary(ACS)
```

```
##
## Call:
## lm(formula = final - antes ~ int + antes, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.3967 -1.8273 -0.3166  1.4563  7.9665
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0969     1.4603   2.121  0.0433 *
## intB          2.1602     1.0892   1.983  0.0576 .
## antes        -0.6816     0.1000  -6.813 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.942 on 27 degrees of freedom
## Multiple R-squared:  0.636, Adjusted R-squared:  0.609
## F-statistic: 23.58 on 2 and 27 DF,  p-value: 1.19e-06
```

## Anova con medidas repetidas

En esta sección se presenta el análisis de varianza (anova) con medidas repetidas que corresponde al siguiente arreglo de datos (diseño de dos factores: sujeto y tiempo):

	tiempo						
sujetos	1	2	...	$j$	...	$n$	$\bar{y}_i$
1	$y_{11}$	$y_{12}$	...	$y_{1j}$	...	$y_{1n}$	$\bar{y}_1$
2	$y_{21}$	$y_{22}$	...	$y_{2j}$	...	$y_{2n}$	$\bar{y}_2$
:				:			:
$i$	$y_{i1}$	$y_{i2}$	...	$y_{ij}$	...	$y_{in}$	$\bar{y}_i$
:				:			:
$N$	$y_{N1}$	$y_{N2}$	...	$y_{Nj}$	...	$y_{Nn}$	$\bar{y}_N$
$\bar{y}_{.j}$	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.n}$	$\bar{y}$

donde  $y_{ij}$  es el desenlace de interés,  $N$  es el número de sujetos,  $n$  es el número de repeticiones,  $\bar{y}_{.j}$  es el promedio en el tiempo  $j$ ,  $\bar{y}_i$  es el promedio del individuo  $i$  y  $\bar{y}$  es el promedio global. En el diseño anterior se asume que los tiempos de medición son fijos para todos los individuos y todos los sujetos son medidos en los  $n$  tiempos. Aunque en este documento no se presenta, este método también se puede utilizar cuando no todos los sujetos son medidos en los  $n$  tiempos (datos des-balanceados).

El objetivo de este diseño es evaluar si el desenlace cambia a través del tiempo.

El análisis se basa en el siguiente modelo:

$$y_{ij} = \mu + \pi_i + \tau_j + e_{ij}$$

donde  $\mu$  es la media global,  $\pi_i$  es el efecto de cada individuo (constante en el tiempo),  $\tau_j$  es el efecto del tiempo (constante en los sujetos) y  $e_{ij}$  es el error del sujeto  $i$  en el tiempo  $j$ . El anterior es un modelo de efectos mixtos donde  $\pi_i \sim N(0, \sigma_\pi^2)$  ( $\sigma_\pi^2$ : varianza entre los sujetos - between) y  $e_{ij} \sim N(0, \sigma_e^2)$  ( $\sigma_e^2$ : varianza dentro/intra sujeto - within) son los efectos aleatorios y  $\tau_j$  es el efecto fijo.

Los supuestos del modelo son:

1.  $\sum_{j=1}^n \tau_j = 0$
2.  $E(y_{ij}) = \mu + \tau_j$
3.  $V(y_{ij}) = \sigma_\pi^2 + \sigma_e^2$
4.  $Cov(y_{ij}, y_{i'j}) = 0$  para  $i \neq i'$ ; las observaciones son independientes entre sujetos.
5.  $Cov(y_{ij}, y_{ij'}) = \sigma_\pi^2$  para  $j \neq j'$ ; la covarianza es constante entre cualquier pareja de repeticiones dentro del mismo sujeto.

Dados los anteriores supuestos, la matriz de varianzas y covarianzas de las medidas repetidas tiene la siguiente estructura que se denomina “compound symmetry”:

$$\begin{bmatrix} \sigma_\pi^2 + \sigma_e^2 & \sigma_\pi^2 & \dots & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 + \sigma_e^2 & \dots & \sigma_\pi^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\pi^2 & \sigma_\pi^2 & \dots & \sigma_\pi^2 + \sigma_e^2 \end{bmatrix}$$

Basados en el modelo y supuestos anteriores, se busca descomponer la variabilidad total del desenlace en los tres efectos: sujetos, tiempo y error. Esta descomposición, a partir del métodos de mínimos cuadrados, se presenta en la tabla de análisis de varianza:

- Tabla de anova con medidas repetidas.

<i>Fuente</i>	<i>gl</i>	<i>SC</i>	<i>MC</i>	<i>F</i>
Sujetos	$N - 1$	$SC_S = n \sum_{i=1}^N (\bar{y}_i - \bar{y})^2$	$MC_S = \frac{SC_S}{N-1}$	$F_S = \frac{MC_S}{MC_e}$
Tiempo	$n - 1$	$SC_\tau = N \sum_{j=1}^n (\bar{y}_{.j} - \bar{y})^2$	$MC_\tau = \frac{SC_\tau}{n-1}$	$F_\tau = \frac{MC_\tau}{MC_e}$
Error	$(N - 1)(n - 1)$	$SC_e = \sum_{i=1}^N \sum_{j=1}^n (y_{ij} - \bar{y}_i - \bar{y}_{.j} + \bar{y})^2$	$MC_e = \frac{SC_e}{(N-1)(n-1)}$	
Total	$Nn - 1$	$SC_t = \sum_{i=1}^N \sum_{j=1}^n (y_{ij} - \bar{y})^2$		

Con base en la tabla de análisis de varianza, se pueden evaluar las siguientes hipótesis nula con sus correspondientes estadísticos de prueba:

1.  $H_0 : \sigma_\pi^2 = 0$ , el estadístico de prueba es  $F_S \sim F_{(N-1);(N-1)(n-1)}$
2.  $H_0 : \tau_j = 0$ , el estadístico de prueba es  $F_\tau \sim F_{(n-1);(N-1)(n-1)}$

Como se mencionó anteriormente, el objetivo principal de este diseño es evaluar la hipótesis número 2.

Adicionalmente, se puede realizar la estimación del coeficiente de correlación intraclase (CCI) que es igual a:

$$CCI = \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_\pi^2 + \hat{\sigma}_e^2}$$

donde:

$$\hat{\sigma}_\pi^2 = \frac{MC_S - MC_e}{n}$$

y

$$\hat{\sigma}_e^2 = MC_e$$

Este coeficiente se interpretará como la proporción de la variabilidad total que es atribuible a los sujetos; es decir, la proporción de variabilidad que no está siendo explicada por el efecto del tiempo.

Si la  $H_0 : \tau_j = 0$  se rechaza, surge el interés de conocer como es el cambio del desenlace a través del tiempo. Es posible descomponer el efecto del tiempo a partir de la definición de  $n - 1$  contrastes definidos por la siguiente expresión:

$$L_{j'} = \sum_{j=1}^n c_{j'j} \bar{y}_{.j}$$

donde  $c_{j'j}$  son los coeficientes de cada contraste. Se han definido un conjunto de diferentes contrastes dependiendo la descomposición del efecto del tiempo que se quiera investigar.

Cuando se quiere descomponer el efecto del tiempo en polinomios de orden 1 hasta  $n - 1$ , los coeficientes  $c_{j'j}$  según el número de repeticiones ( $n$ ) son los siguientes:

link:

- Coeficientes polinomios

$$\begin{bmatrix} n = 2 & coef. \\ lineal & -1, 1 \end{bmatrix}$$

$$\begin{bmatrix} n = 3 & coef. \\ lineal & -1, 0, 1 \\ cuadrado & 1, -2, 1 \end{bmatrix}$$

$$\begin{bmatrix} n = 4 & coef. \\ lineal & -3, -1, 1, 3 \\ cuadrado & 1, -1, -1, 1 \\ cúbico & -1, 3, -3, 1 \end{bmatrix}$$

$$\begin{bmatrix} n = 5 & coef. \\ lineal & -2, -1, 0, 1, 2 \\ cuadrado & 2, -1, -2, -1, 2 \\ cúbico & -1, 2, 0, -2, 1 \\ 4orden & 1, -4, 6, -4, 1 \end{bmatrix}$$

$$\begin{bmatrix} n = 6 & coef. \\ lineal & -5, -3, -1, 1, 3, 5 \\ cuadrado & 5, -1, -4, -4, -1, 5 \\ cúbico & -5, 7, 4, -4, -7, 5 \\ 4orden & 1, -3, 2, 2, -3, 1 \\ 5orden & -1, 5, -10, 10, -5, 1 \end{bmatrix}$$

Cada coeficiente se divide por:  $\sqrt[2]{\sum_{j=1}^n c_{j',j}^2}$

La suma de cuadrados para cada contraste se define como:

$$SC_{L_{j'}} = \frac{NL_{j'}^2}{\sum_{j=1}^n c_{j',j}^2}$$

Basado en este resultado, se puede plantear la siguiente hipótesis nula para cada contraste:  $H_0$  : *El tiempo no sigue una tendencia de orden  $j'$ .*

Cada hipótesis se evalúa con el siguiente estadístico de prueba:

$$F_j' = \frac{SC_{L_{j'}}}{MC_e} \sim F_{1;(N-1)(n-1)}$$

Dado que se están realizando comparaciones múltiple, definido un nivel de significancia ( $\alpha$ ), el ajuste del nivel de significancia para las pruebas específicas por el método de bonferroni corresponde a  $\alpha^* = \frac{\alpha}{n-1}$ .

- Otros contrastes

Se pueden explorar los cambios entre los tiempos a partir de otros contrastes que tiene un sentido específico dentro de la investigación. Aquí se presentan tres casos, tomando como ejemplo un  $n = 5$ :

1. Cuando se quiere comparar cada uno de los tiempos frente al primero (línea base):

$$\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

2. Si se quiere evaluar cambios consecutivos entre dos tiempos:

$$\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

3. Cuando se quiere comparar cada medida versus el promedio de las restantes, los coeficientes son los siguiente:

$$\begin{bmatrix} 1 & -1/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 1 & -1/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & -1/4 & 1 \end{bmatrix}$$

- Supuesto de esfericidad:

Las pruebas descritas anteriormente son válidas bajo la estructura de la matriz de varianzas y covarianzas especificada al comienzo de este capítulo (supuestos 3 y 5 - compound symmetry). Esta estructura es muy restrictiva y corresponde a un caso específico de una estructura definida por el supuesto de *esfericidad* o *circularidad* que establece que la  $V(y_{ij} - y_{ij'})$  es contante para todo  $j \neq j'$ . Bajo el supuesto de esfericidad son válidas las distribuciones de los estadísticos de prueba definidos hasta este momento.

Una prueba de hipótesis desarrollada para evaluar el supuesto de **esfericidad** es la prueba de **Mauchly** que se basa en el cálculo del coeficiente  $\epsilon$ . La hipótesis nula corresponde a  $H_0 : \epsilon = 1$  que equivale al cumplimiento del supuesto de esfericidad. Si se rechaza esta hipótesis nula, una alternativa es ajustar los grados de libertad de las estadísticas  $F$  multiplicándolos por el coeficiente  $\epsilon$ . Dos propuestas para estimar el coeficiente fueron desarrolladas por Greenhouse y Geisser (1959) y Huynh y Feldt (1976) las cuales se encuentran implementadas en el programa **R**.

## Aplicación.

Escenario: 33 personas ingresaron a un programa de actividad física que tiene una duración de 5 meses. En el transcurso del programa se midió el peso de cada participante cada mes ( $n=5$ ). Se quiere saber si el programa afecta el peso de las 33 personas.

Primero, cargamos los datos que se encuentran en una estructura **ancho**.

```
library(rio)
peso<-import("peso.xlsx");peso[1:5,]
```

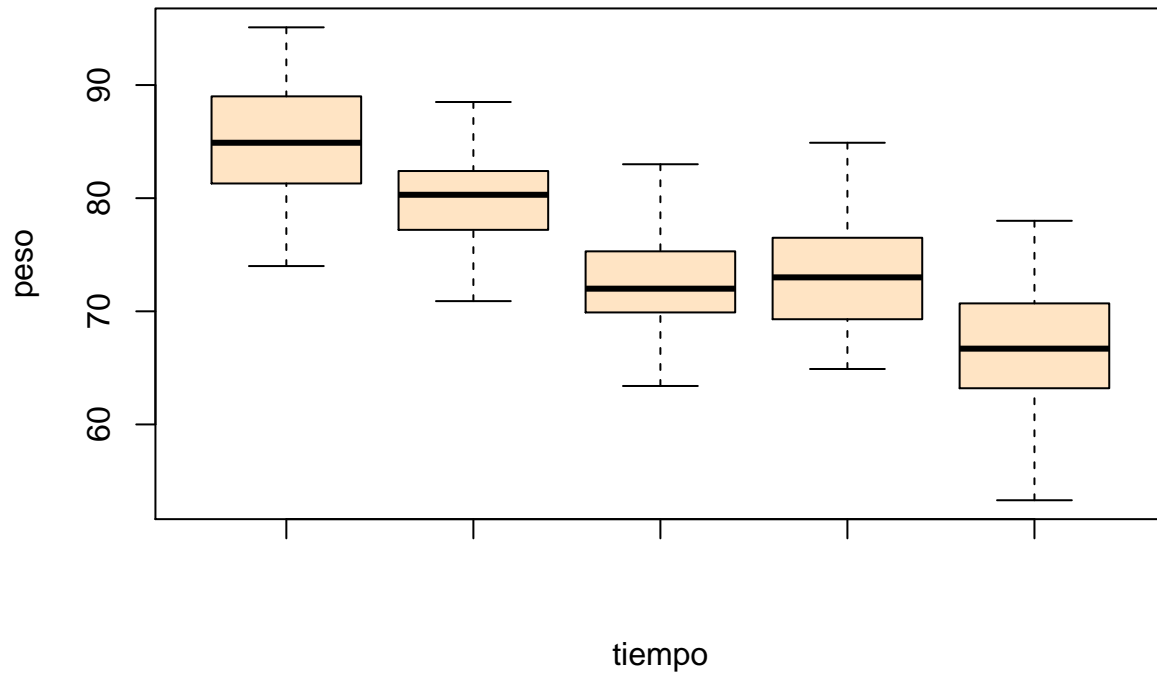
```
## ID P1 P2 P3 P4 P5
## 1 1 84.9 80.4 77.1 69.6 70.1
## 2 2 88.3 83.4 72.0 72.9 68.1
## 3 3 74.0 80.3 75.3 80.2 67.2
## 4 4 84.3 83.8 68.3 72.8 66.6
## 5 5 86.2 88.5 73.9 76.3 65.2
```

Primero, obtenemos un gráficos de cajas y bigotes por cada repetición:

```
boxplot(peso$P1,peso$P2,peso$P3,peso$P4,peso$P5,
        main = "Cajas y bigotes peso por mes",ylab="peso",xlab="tiempo",col = "bisque")
```



## Cajas y bigotes peso por mes



Evaluación del supuesto de normalidad para cada repetición con la prueba de Shapiro Wilk:

```
shapiro.test(peso$P1)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: peso$P1  
## W = 0.96945, p-value = 0.4651
```

```
shapiro.test(peso$P2)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: peso$P2  
## W = 0.9675, p-value = 0.4145
```

```
shapiro.test(peso$P3)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: peso$P3  
## W = 0.98866, p-value = 0.9757
```

```
shapiro.test(peso$P4)
```

```
##
```

```
## Shapiro-Wilk normality test
##
## data: peso$P4
## W = 0.97776, p-value = 0.7167
```

```
shapiro.test(peso$P5)
```

```
##
## Shapiro-Wilk normality test
##
## data: peso$P5
## W = 0.98436, p-value = 0.904
```

Matriz de correlación:

```
cor(peso[,2:5],method = "pearson")
```

```
##           P1           P2           P3           P4
## P1  1.00000000  0.07628701 -0.1985607 -0.1588425
## P2  0.07628701  1.00000000 -0.1346327  0.1730389
## P3 -0.19856067 -0.13463268  1.0000000  0.1201956
## P4 -0.15884247  0.17303892  0.1201956  1.0000000
```

Para utilizar la función `aov` para calcular la tabla de anova con medidas repetidas, primero se organizan los datos en una estructura **a los largo** de la siguiente forma:

```
peso2<- reshape(data = peso,varying = 2:6, v.names = "peso", timevar= "tiempo",
               idvar = "sujeto", direction = "long")
peso2$tiempo<-factor(peso2$tiempo)
peso2$sujeto<-factor(peso2$sujeto)
```

ahora el cálculo de la tabla de **anova** se obtiene así:

```
anova <- aov(peso~tiempo+Error(sujeto/tiempo), data = peso2)
summary(anova)
```

```
##
## Error: sujeto
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 32  841.4    26.29
##
## Error: sujeto:tiempo
##           Df Sum Sq Mean Sq F value Pr(>F)
## tiempo     4   6581  1645.2   65.98 <2e-16 ***
## Residuals 128   3192    24.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

dado que la  $MC_e=24.9$  y la  $MC_S=26.29$ , el coeficiente de correlación intraclase es igual a:

```
MC_S<-26.29283
MC_e<-24.93485
```

```
v.suj<-(MC_S-MC_e)/5
icc<-v.suj/(v.suj+MC_e);icc
```

```
## [1] 0.01077486
```

Para realizar la descomposición de efecto del tiempo en polinomios de orden  $n - 1 = 4$  se puede utilizar dos alternativas; la primera con la opción `contr.poly(r)` y la función `aov`:

```
#definir los contrastes con la opción contr.poly(n)
contrasts(peso2$tiempo)<-contr.poly(5)

anova<-aov(peso~tiempo+Error(sujeto/tiempo), data = peso2)
summary(anova,split=list(tiempo=list("lineal"=1, "cuadrado"= 2,
                                     "cubico"=3,"4 orden"=4)))
```

```
##
## Error: sujeto
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 32  841.4   26.29
##
## Error: sujeto:tiempo
##           Df Sum Sq Mean Sq F value  Pr(>F)
## tiempo           4   6581   1645  65.980 < 2e-16 ***
## tiempo: lineal    1   6120   6120 245.431 < 2e-16 ***
## tiempo: cuadrado  1     76     76   3.065 0.08240 .
## tiempo: cubico    1    118    118   4.731 0.03147 *
## tiempo: 4 orden   1    267    267  10.696 0.00138 **
## Residuals        128   3192     25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La segunda opción, definiendo cada uno de los 4 vectores que incluyen los coeficientes de la siguiente forma:

```
c1<-c(-2,-1,0,1,2)/sqrt(10)
c2<-c(2,-1,-2,-1,2)/sqrt(14)
c3<-c(-1,2,0,-2,1)/sqrt(10)
c4<-c(1,-4,6,-4,1)/sqrt(70)

mat<-cbind(c1,c2,c3,c4);mat
```

```
##           c1           c2           c3           c4
## [1,] -0.6324555  0.5345225 -0.3162278  0.1195229
## [2,] -0.3162278 -0.2672612  0.6324555 -0.4780914
## [3,]  0.0000000 -0.5345225  0.0000000  0.7171372
## [4,]  0.3162278 -0.2672612 -0.6324555 -0.4780914
## [5,]  0.6324555  0.5345225  0.3162278  0.1195229
```

```
contrasts(peso2$tiempo)<-mat

anova <- aov(peso~tiempo+Error(sujeto/tiempo), data = peso2)
summary(anova,split=list(tiempo=list("lineal"=1, "cuadrado"= 2,
                                     "cubico"=3,"4 orden"=4)))
```

```
##
## Error: sujeto
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 32  841.4   26.29
##
## Error: sujeto:tiempo
##           Df Sum Sq Mean Sq F value  Pr(>F)
## tiempo           4   6581   1645  65.980 < 2e-16 ***
## tiempo: lineal    1   6120   6120 245.431 < 2e-16 ***
## tiempo: cuadrado  1     76     76   3.065 0.08240 .
## tiempo: cubico    1    118    118   4.731 0.03147 *
```

```
## tiempo: 4 orden 1 267 267 10.696 0.00138 **
## Residuals 128 3192 25
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Otras forma de obtener el anova con medidas repetidas es utilizando la función `Anova` del paquete `car`. Esta función trabaja con los datos en una estructura “a lo ancho”. La ventaja de esta función es que reporta las estimaciones del coeficiente  $\epsilon$  bajo dos métodos incluyendo el valor p ajustado para evaluar el efecto del tiempo.

```
library(car)

peso3<-matrix(c(peso$P1,peso$P2,peso$P3,peso$P4,peso$P5),nrow=33, ncol=5)

modelo<-lm(peso3 ~ 1)
rep<-factor(c("p1", "p2", "p3", "p4", "p5"))

options(contrasts=c("contr.sum", "contr.poly"))
anova3 <- Anova(modelo, idata=data.frame(rep), idesign=~rep, type="III")
summary(anova3, multivariate=F)
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##          Sum Sq num Df Error SS den Df  F value    Pr(>F)
## (Intercept) 937388      1   841.4    32 35651.844 < 2.2e-16 ***
## rep          6581      4  3191.7   128   65.981 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##      Test statistic p-value
## rep      0.71925 0.34927
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
## for Departure from Sphericity
##
##      GG eps Pr(>F[GG])
## rep 0.86696 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      HF eps  Pr(>F[HF])
## rep 0.985195 8.952692e-30
```

- Otros contrastes:

*Cambio relativo a la línea de base:*

```
mean<-matrix(c(mean(peso$P1),mean(peso$P2),mean(peso$P3),mean(peso$P4),
                mean(peso$P5)),nrow=5,ncol = 1,dimnames = list(1:5,"media"))

B1<-c(-1,1,0,0,0)
B2<-c(-1,0,1,0,0)
```

```

B3<-c(-1,0,0,1,0)
B4<-c(-1,0,0,0,1)

mat1<-rbind(B1,B2,B3,B4)

SCE_L1<-33*(mat1%*%mean)^2/2
F_11<-SCE_L1/MC_e

F_11

```

```

##          media
## B1  20.90919
## B2 103.99689
## B3  92.80223
## B4 224.55027

```

```
1-pf(F_11[1,1],1,128)
```

```
## [1] 1.121364e-05
```

```
1-pf(F_11[2,1],1,128)
```

```
## [1] 0
```

```
1-pf(F_11[3,1],1,128)
```

```
## [1] 0
```

```
1-pf(F_11[4,1],1,128)
```

```
## [1] 0
```

*Cambio en tiempos consecutivos:*

```

D1<-c(-1,1,0,0,0)
D2<-c(0,-1,1,0,0)
D3<-c(0,0,-1,1,0)
D4<-c(0,0,0,-1,1)

mat2<-rbind(D1,D2,D3,D4)

SCE_L2<-33*(mat2%*%mean)^2/2
F_12<-SCE_L2/MC_e

F_12

```

```

##          media
## D1 20.9091863
## D2 31.6432137
## D3  0.3186547
## D4 28.6397321

```

```
1-pf(F_12[1,1],1,128)
```

```
## [1] 1.121364e-05
```

```
1-pf(F_12[2,1],1,128)
```

```
## [1] 1.112043e-07
```

```
1-pf(F_12[3,1],1,128)
```

```
## [1] 0.5734052
```

```
1-pf(F_12[4,1],1,128)
```

```
## [1] 3.893391e-07
```

*Cada repetición versus el promedio del resto:*

```
E1<-c(1,-1/4,-1/4,-1/4,-1/4)
```

```
E2<-c(-1/4,1,-1/4,-1/4,-1/4)
```

```
E3<-c(-1/4,-1/4,1,-1/4,-1/4)
```

```
E4<-c(-1/4,-1/4,-1/4,1,-1/4)
```

```
mat3<-rbind(E1,E2,E3,E4)
```

```
SCE_L3<-33*(mat3%*%mean)^2/1.25
```

```
F_13<-SCE_L3/MC_e
```

```
F_13
```

```
##      media
```

```
## E1 155.148828
```

```
## E2  27.309724
```

```
## E3  13.457152
```

```
## E4   7.705366
```

```
1-pf(F_13[1,1],1,128)
```

```
## [1] 0
```

```
1-pf(F_13[2,1],1,128)
```

```
## [1] 6.842055e-07
```

```
1-pf(F_13[3,1],1,128)
```

```
## [1] 0.0003567227
```

```
1-pf(F_13[4,1],1,128)
```

```
## [1] 0.006331956
```

## Manova con medidas repetidas

El análisis de varianza multivariado (Manova) con medidas repetidas es otra alternativa para el análisis de datos longitudinales, que a diferencia del anova con medidas repetidas, no requiere de una estructura específica (“compound-symmetric”) sobre la matriz de varianzas y covarianzas. La principal limitación es que requiere de información completa de todos los individuos seleccionados lo que puede generar posibles sesgos si se retiran los sujetos con datos faltantes; en el caso anterior se puede considerar realizar imputación de datos faltantes.

El Manova se basa en el siguiente modelo:

$$Y_i = \mu + \varepsilon_i$$

donde  $Y_i$  es un vector de dimensión  $n \times 1$  que contiene el desenlace del individuo  $i$  en los  $n$  tiempos,  $\mu$  es un vector de dimensión  $n \times 1$  que contiene las medias de cada repetición y  $\varepsilon_i$  es un vector de dimensión  $n \times 1$  que contiene los errores del individuo  $i$  que siguen una distribución  $N(0, \Sigma)$  donde  $\Sigma$  es la matriz de varianza y covarianza de los errores y no tiene ninguna estructura predefinida (a diferencia del anova con medidas repetidas).

El vector  $\mu$  se puede expresar en términos de un conjunto de contrastes con el fin de evaluar la tendencia del tiempo. Por ejemplo, para  $n = 5$  al igual que en la sección de anova con medidas repetidas, se pueden definir la matriz de polinomios ortogonales denominada por  $P$  de la siguiente forma:

$$P = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & \mapsto \div \sqrt{5} \\ -2 & -1 & 0 & 1 & 2 & \mapsto \div \sqrt{10} \\ 2 & -1 & -2 & -1 & 2 & \mapsto \div \sqrt{14} \\ -1 & 2 & 0 & -2 & 1 & \mapsto \div \sqrt{10} \\ 1 & -4 & 6 & -4 & 1 & \mapsto \div \sqrt{70} \end{bmatrix}$$

usando  $P$ , la tabla de análisis de varianza multivariado es igual a:

- Tabla de Manova

Fuente	gl	SC
tiempo	1	$SCT = NP\bar{y}.\bar{y}'P'$
Error	N-1	$SCE = P(Y'Y - N\bar{y}.\bar{y}')P'$
Total	N	$SCY = PY'YP'$

donde  $\bar{y}.$  corresponde a un vector de dimensión  $n \times 1$  que incluye los promedios de cada repetición,  $Y$  es la matriz de dimensiones  $N \times n$  que incluye los valores de los desenlaces observados en los  $N$  sujetos en las  $n$  repeticiones y  $'$  denota la traspuesta del objeto correspondiente.

Los elementos de la diagonal de las matrices  $SCT$  y  $SCE$  serán los valores de interés que se pueden denotar de la siguiente forma:

$$SCT = \begin{bmatrix} SCT_0 & & & & \\ & SCT_1 & & & \\ & & SCT_2 & & \\ & & & \dots & \\ & & & & SCT_{n-1} \end{bmatrix}$$

$$SCE = \begin{bmatrix} SCE_0 & & & & \\ & SCE_1 & & & \\ & & SCE_2 & & \\ & & & \dots & \\ & & & & SCE_{n-1} \end{bmatrix}$$

Basado en las dos matrices anteriores, se puede evaluar si el tiempo tiene una tendencia descrita por un polinomio de grado  $j$ . El estadístico de prueba es el siguiente:

$$F_{Cj} = \frac{SCT_j}{SCE_j/(N-1)} \sim F_{(1)(N-1)}$$

para el contraste  $j$ , variando de 1 a  $n-1$ .

Si se cumple el supuesto de esfericidad, la tabla de anova univariado (anova con medidas repetidas) se puede reconstruir de las matrices  $SCT$  y  $SCE$ .  $SCE_0$  (posición [1,1] de la matriz  $SCE$ ) corresponde a la suma de cuadrados del sujeto ( $SC_S$ ), la  $\sum_{j=1}^{n-1} SCE_j$  corresponde a la suma de cuadrados de los errores ( $SC_E$ ) y  $\sum_{j=1}^{n-1} SCT_j$  corresponde a la suma de cuadrados de los tiempos, donde cada elemento de la sumatoria corresponden a la descomposición de la suma de cuadrados del efecto del tiempo por polinomios ortogonales.

- Prueba global del efecto del tiempo.

Para evaluar la hipótesis nula de *no efecto del tiempo*, es decir que las medias de cada repetición son iguales ( $H_0 : \mu_1 = \mu_2 = \dots = \mu_n$ ), la prueba de hipótesis se basa en la comparación entre las dos matrices  $SCT$  y  $SCE$ , resolviendo para  $\lambda$  la siguiente expresión:

$$|SCT * -\lambda SCE*| = 0$$

donde  $SCT*$  y  $SCE*$  corresponde a las submatrices de dimensiones  $(n-1) \times (n-1)$  retirando la primera fila y la primera columna de las dos respectivas matrices.

Evaluar la hipótesis de que las media son iguales en todos los tiempos es equivalente a probar  $H_0 : \lambda = 1$ . Las pruebas de: **Wilks**, **Pillai**, **Lawley - Hotelling** y **Roy** son cuatro alternativas que evalúan la hipótesis nula y que se incluyen en varios programas estadísticos.

## Aplicación.

Utilizaremos el caso desarrollado en la sección de anova con medidas repetidas:

Caso.33 personas ingresaron a un programa de actividad física que tiene una duración de 5 meses. En el transcurso del programa se midió el peso de cada participante cada mes ( $n=5$ ). Se quiere saber si el programa afecta el peso de las 33 personas. Los datos observados fueron los siguientes:

Primero se cargan los datos obtenidos de los 33 sujetos en la 5 mediciones y los guardamos en una matriz de orden  $33 \times 5$ .

```
library(rio)
peso<-import("peso.xlsx")
Y<-matrix(c(peso$P1,peso$P2,peso$P3,peso$P4,peso$P5),nrow=33, ncol=5);Y
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] 84.9 80.4 77.1 69.6 70.1
## [2,] 88.3 83.4 72.0 72.9 68.1
## [3,] 74.0 80.3 75.3 80.2 67.2
## [4,] 84.3 83.8 68.3 72.8 66.6
```

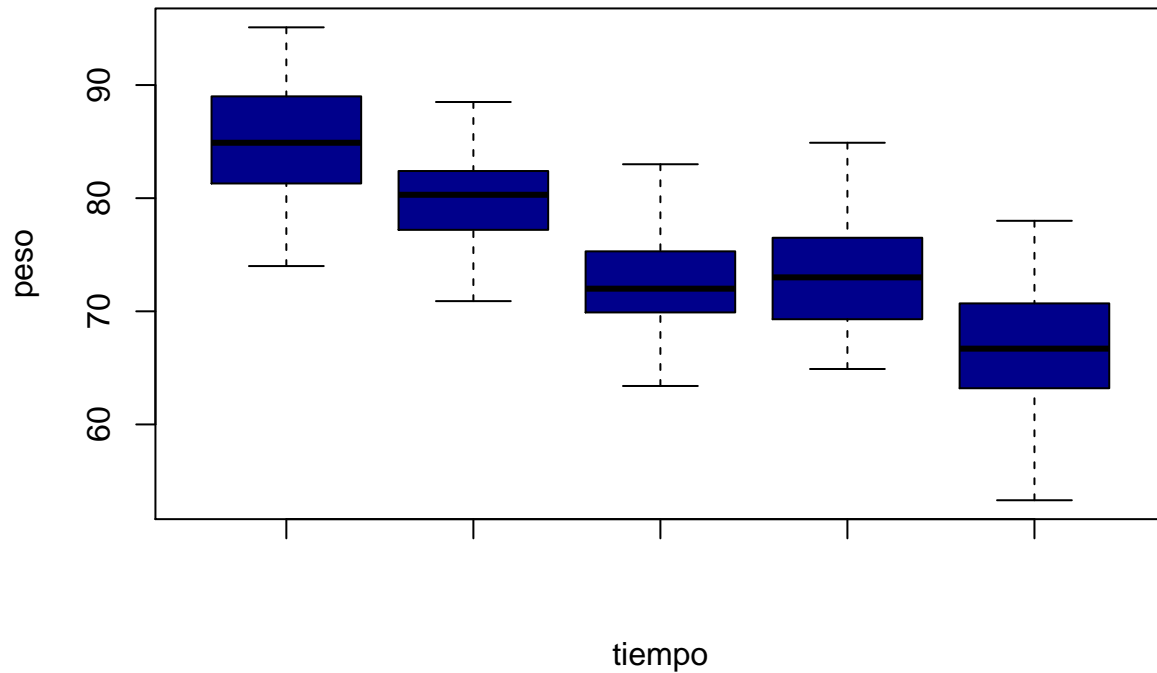


```
## [5,] 86.2 88.5 73.9 76.3 65.2
## [6,] 90.8 78.7 72.4 67.3 63.8
## [7,] 81.3 71.9 76.2 64.9 74.7
## [8,] 87.6 74.1 73.8 76.5 62.5
## [9,] 74.4 75.0 74.0 74.4 61.7
## [10,] 87.6 78.5 67.8 72.0 70.7
## [11,] 87.7 78.1 69.9 74.3 76.4
## [12,] 76.5 80.8 74.7 74.8 73.4
## [13,] 89.8 74.4 71.6 79.9 66.1
## [14,] 92.0 84.0 69.9 79.2 55.3
## [15,] 84.9 83.5 66.4 68.7 66.2
## [16,] 90.7 80.7 67.9 68.1 63.2
## [17,] 82.4 80.3 75.2 72.6 61.9
## [18,] 87.2 82.5 77.8 79.3 78.0
## [19,] 86.7 72.3 83.0 66.0 66.7
## [20,] 83.6 78.8 71.3 68.4 72.0
## [21,] 80.1 76.9 63.4 70.1 63.3
## [22,] 93.6 83.2 70.5 69.3 57.8
## [23,] 95.1 73.6 67.1 78.1 67.2
## [24,] 79.8 77.2 70.9 73.3 64.1
## [25,] 83.4 80.0 66.9 74.7 62.6
## [26,] 91.3 79.1 75.6 66.4 68.8
## [27,] 83.2 82.3 70.8 68.0 74.7
## [28,] 81.2 84.5 78.6 76.8 69.5
## [29,] 76.9 70.9 74.3 73.0 53.3
## [30,] 74.7 82.0 78.1 84.9 70.8
## [31,] 93.9 77.2 77.5 76.9 66.8
## [32,] 83.8 82.4 69.6 72.1 57.5
## [33,] 89.0 82.1 71.4 74.3 72.8
```

obtenemos un gráfico de los datos observados y evaluamos si las observaciones provienen de una población con distribución normal multivariada.

```
boxplot(peso$P1,peso$P2,peso$P3,peso$P4,peso$P5, main="Boxplot peso",
        ylab="peso",xlab="tiempo",col = "darkblue")
```

## Boxplot peso



```
library(mvShapiroTest)
mvShapiro.Test(Y) # prueba de normalidad multivariada
```

```
##
## Generalized Shapiro-Wilk test for Multivariate Normality by
## Villasenor-Alva and Gonzalez-Estrada
##
## data: Y
## MVW = 0.97479, p-value = 0.8571
```

Con esta estructura (a lo ancho) y utilizando los comando `lm` y `Anova` se puede correr el primer Manova de la siguiente forma.

```
library(car)
modelo<-lm(Y~1)
anova1<-Anova(modelo,type="III")
```

Los resultados que se obtienen son los siguientes:

```
anova1
##
## Type III MANOVA Tests: Pillai test statistic
##           Df test stat approx F num Df den Df   Pr(>F)
## (Intercept) 1  0.99922  7173.1     5    28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado anterior ( $F_{(5,28)} = 7173.1$  con valor  $p < 2.2e-16$ ) esta evaluando la hipótesis nula de que todas las medias de las repeticiones son iguales a cero, lo cual no es de mayor interés. Para obtener la prueba global del efecto del tiempo se puede ejecutar la siguiente instrucción:

```
tiempo<-factor(c("p1","p2","p3","p4","p5"))
anova2 <- Anova(modelo,idata=data.frame(tiempo), idesign=~tiempo, type="III"
,icontrasts = "contr.poly")
```

Los resultados son los siguientes:

```
anova2

##
## Type III Repeated Measures MANOVA Tests: Pillai test statistic
##           Df test stat approx F num Df den Df   Pr(>F)
## (Intercept) 1  0.99910  35652      1    32 < 2.2e-16 ***
## tiempo      1  0.85648     43      4    29 7.999e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con una  $F_{(4;29)} = 43$  y un valor  $p = 7.999e - 12$ , se rechaza la hipótesis nula de que las medias de los tiempos son iguales. Si se quiere obtener mayor detalle de la salida obtenida por el comando Anova, se puede utilizar la función summary así:

```
summary(anova2,univariate = F)

##
## Type III Repeated Measures MANOVA Tests:
##
## -----
##
## Term: (Intercept)
##
## Response transformation matrix:
## (Intercept)
## [1,] 1
## [2,] 1
## [3,] 1
## [4,] 1
## [5,] 1
##
## Sum of squares and products for the hypothesis:
## (Intercept)
## (Intercept) 4686940
##
## Multivariate Tests: (Intercept)
##           Df test stat approx F num Df den Df   Pr(>F)
## Pillai    1  0.9991 35651.84      1    32 < 2.22e-16 ***
## Wilks     1  0.0009 35651.84      1    32 < 2.22e-16 ***
## Hotelling-Lawley 1 1114.1201 35651.84      1    32 < 2.22e-16 ***
## Roy       1 1114.1201 35651.84      1    32 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----
##
## Term: tiempo
```

```
##
## Response transformation matrix:
## tiempo.L tiempo.Q tiempo.C tiempo^4
## [1,] -0.6324555 0.5345225 -3.162278e-01 0.1195229
## [2,] -0.3162278 -0.2672612 6.324555e-01 -0.4780914
## [3,] 0.0000000 -0.5345225 -4.095972e-16 0.7171372
## [4,] 0.3162278 -0.2672612 -6.324555e-01 -0.4780914
## [5,] 0.6324555 0.5345225 3.162278e-01 0.1195229
##
## Sum of squares and products for the hypothesis:
## tiempo.L tiempo.Q tiempo.C tiempo^4
## tiempo.L 6119.7734 -683.86986 849.64555 1277.5444
## tiempo.Q -683.8699 76.42080 -94.94583 -142.7625
## tiempo.C 849.6455 -94.94583 117.96148 177.3693
## tiempo^4 1277.5444 -142.76250 177.36930 266.6961
##
## Multivariate Tests: tiempo
## Df test stat approx F num Df den Df Pr(>F)
## Pillai 1 0.856478 43.26485 4 29 7.9991e-12 ***
## Wilks 1 0.143522 43.26485 4 29 7.9991e-12 ***
## Hotelling-Lawley 1 5.967566 43.26485 4 29 7.9991e-12 ***
## Roy 1 5.967566 43.26485 4 29 7.9991e-12 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La tabla de Manova con medidas repetidas con una matriz de polinomios ortogonales de orden 4, se puede reconstruir de la siguiente forma:

```
y_mu<-matrix(c(mean(peso$P1),mean(peso$P2),mean(peso$P3),mean(peso$P4),
               mean(peso$P5)),nrow=5,ncol=1,dimnames = list(1:5,c("media")))
```

```
y_mu
```

```
## media
## 1 85.05758
## 2 79.43636
## 3 72.52121
## 4 73.21515
## 5 66.63636
```

```
c0<-c(1,1,1,1,1)/sqrt(5)
c1<-c(-2,-1,0,1,2)/sqrt(10)
c2<-c(2,-1,-2,-1,2)/sqrt(14)
c3<-c(-1,2,0,-2,1)/sqrt(10)
c4<-c(1,-4,6,-4,1)/sqrt(70)
```

```
P<-t(matrix(c(c0,c1,c2,c3,c4),nrow=5,ncol=5,
             dimnames = list(1:5,c("A","B","C","D","E"))))
```

```
P
```

```
## 1 2 3 4 5
## A 0.4472136 0.4472136 0.4472136 0.4472136 0.4472136
## B -0.6324555 -0.3162278 0.0000000 0.3162278 0.6324555
## C 0.5345225 -0.2672612 -0.5345225 -0.2672612 0.5345225
## D -0.3162278 0.6324555 0.0000000 -0.6324555 0.3162278
## E 0.1195229 -0.4780914 0.7171372 -0.4780914 0.1195229
```

```
#Matrices de sumas de cuadrados con P
```

```
SCTiempo<-33*P%*%y_mu%*%t(y_mu)%*%t(P)  
SCEError<-P%*%(t(Y)%*%Y-33*y_mu%*%t(y_mu))%*%t(P)  
SCTotal<-P%*%t(Y)%*%Y%*%t(P)
```

se obtiene que las matrices *SCT* y *SCE* son iguales a:

```
SCTiempo
```

```
##           A           B           C           D           E  
## A 937387.997 -75740.3598 8463.80183 -10515.49714 -15811.3165  
## B -75740.360  6119.7734 -683.86986   849.64555  1277.5444  
## C  8463.802  -683.8699   76.42080  -94.94583  -142.7625  
## D -10515.497   849.6455  -94.94583   117.96148   177.3693  
## E -15811.317  1277.5444 -142.76250   177.36930   266.6961
```

```
SCEError
```

```
##           A           B           C           D           E  
## A 841.370667  162.2386  179.8429  63.24127  -2.478224  
## B 162.238580 1059.2516 -212.4205 207.65945 117.867534  
## C 179.842862 -212.4205 853.9342 113.16313 118.570476  
## D 63.241273  207.6595  113.1631 601.39352 63.775434  
## E -2.478224  117.8675  118.5705 63.77543 677.080892
```

Con base en estas dos matrices se puede evaluar las hipótesis correspondientes a la tendencia del tiempo de la siguiente forma:

```
test<-SCTiempo/(SCEError/32)  
test
```

```
##           A           B           C           D           E  
## A 35651.844 -14939.0577 1505.990598 -5320.827520 204163.16112  
## B -14939.058  184.8784  103.021319  130.929061  346.84209  
## C 1505.991  103.0213   2.863764  -26.848555  -38.52898  
## D -5320.828  130.9291  -26.848555   6.276701   88.99693  
## E 204163.161  346.8421  -38.528983   88.996926  12.60451
```

```
1-pf(test[2,2],1,32) # lineal
```

```
## [1] 7.549517e-15
```

```
1-pf(test[3,3],1,32) # Cuadrático
```

```
## [1] 0.100312
```

```
1-pf(test[4,4],1,32) # Cúbico
```

```
## [1] 0.01751652
```

```
1-pf(test[5,5],1,32) # Orden 4
```

```
## [1] 0.001215241
```

- Otros contrastes:

Comparación entre primera medición (línea base) y cada una de las mediciones:

```
b0<-c(1,1,1,1,1)
```

```
b1<-c(-1,1,0,0,0)
```

```

b2<-c(-1,0,1,0,0)
b3<-c(-1,0,0,1,0)
b4<-c(-1,0,0,0,1)

```

```

P1<-t(matrix(c(b0,b1,b2,b3,b4),nrow=5,ncol=5,
              dimnames = list(1:5,c("A","B","C","D","E"))))

```

```
P1
```

```

##    1 2 3 4 5
## A  1 1 1 1 1
## B -1 1 0 0 0
## C -1 0 1 0 0
## D -1 0 0 1 0
## E -1 0 0 0 1

```

```
#Matrices de sumas de cuadrados con P
```

```

SCTiempo<-33*P1%*%y_mu%*%t(y_mu)%*%t(P1)
SCErrror<-P1%*%(t(Y)%*%Y-33*y_mu%*%t(y_mu))%*%t(P1)
SCTotal<-P1%*%t(Y)%*%Y%*%t(P1)

```

Los resultados de la prueba son los siguientes:

```

test<-SCTiempo/(SCErrror/32)
test

```

```

##           A           B           C           D           E
## A  35651.84 31847.58506 31366.22457 239519.41996 -13370.21531
## B  31847.59  22.32127   68.61168    56.11474    97.62326
## C  31366.22  68.61168   84.80067   109.16480   167.29088
## D 239519.42  56.11474   109.16480    71.74318   186.20130
## E -13370.22  97.62326   167.29088    186.20130   158.02688

```

```
1-pf(test[2,2],1,32) # 1v2
```

```
## [1] 4.417758e-05
```

```
1-pf(test[3,3],1,32) # 1v3
```

```
## [1] 1.637076e-10
```

```
1-pf(test[4,4],1,32) # 1v4
```

```
## [1] 1.115775e-09
```

```
1-pf(test[5,5],1,32) # 1v5
```

```
## [1] 6.37268e-14
```

La evaluación para otros contrastes se puede realizar cambiando los coeficiente de la matriz P.

# Estimación de ecuaciones generalizadas - GEE

## Modelos lineales generalizados (MLG)

Los modelos de estimación de ecuaciones generalizadas (**GEE**) son una extensión de los modelos lineales generalizados (**MLG**) para datos correlacionados. Los **MLG** son una clase de modelos [McCullagh y Nelder - 1989] que plantean la exploración de la relación de una desenlace, ya sea este de naturaleza continua o categórica, y un conjunto de variables independientes a través de un modelo de regresión lineal. Un supuesto de estos modelos es que el desenlace corresponde a una variable que tiene una distribución que pertenece a la familia exponencial entre las que se encuentra la distribución **normal o gaussiana** para desenlaces continuos, la distribución **binomial** para variables dicotómicas y la distribución de **poisson** para una variable de conteo. La construcción de los MLG se basa en tres especificaciones:

1. Predictor lineal:

$$\eta_i = x_i\beta$$

donde  $x_i$  son los valores observados en las variables independientes en el individuo  $i$ , y  $\beta$  son los coeficientes de la regresión.

2. Función de enlace:

$$g(\mu_i) = \eta_i$$

que corresponde a una función que convierte el valor esperado de la variable desenlace en el predictor lineal. Las funciones de enlace para desenlaces continuos, dicotómicos y de conteo son las siguientes:

Desenlace	Función de enlace	$g(\mu_i)$
Continuo	identidad	$\mu_i = x_i\beta$
Dicotómico	logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i\beta$
Conteo	logaritmo	$\log(\mu_i) = x_i\beta$

3. Expresión de la varianza del desenlace:

$$V(y_i) = \phi v(\mu_i)$$

donde  $v(\mu_i)$  se conoce como la función de la varianza y  $\phi$  como el parámetro escalar. La expresión de la varianza y parámetro escalar, dependiendo el tipo de desenlace, son:

Desenlace	$\phi$	$v(\mu_i)$
Continuo	$\sigma_\epsilon^2$	1
Dicotómica	1	$\mu_i(1 - \mu_i)$
Conteo	1	$\mu_i$

Con base en las tres especificaciones anteriores, la estimación de los  $\beta$  se obtiene de resolver la siguiente ecuación:

$$\sum_{i=1}^N \left(\frac{\partial \mu_i}{\partial \beta}\right)' (V(y_i))^{-1} [y_i - \mu_i] = 0$$

Dado que la expresión anterior depende únicamente de la media y la varianza de  $y$  y no de la definición completa de la función de distribución del desenlace  $y$ , el método de estimación de los coeficientes  $\beta$  se denomina: "cuasi - verosimilitud".

## GEE - Desenlace continuo.

Los modelos **GEE** corresponde a una extensión de los **MLG** para datos correlacionados. Entre sus características se encuentra:

1. No se requiere que todos los individuos sean medidos en todos los tiempos ( $n_i \neq n$ ).
2. Se realiza el supuesto sobre la distribución del desenlace en cada tiempo (distribución marginal); no se requiere realizar el supuesto sobre la distribución conjunta de  $y_{ij}$ .
3. Se realiza una estimación independiente de los coeficientes de regresión del modelo ( $\beta$ ) y de los parámetros de asociación (estimación de la estructura de covarianza de las medidas repetidas).
4. Los **GEE** se utilizan cuando el interés principal es estudiar la relación entre las variables independientes y el desenlace (coeficientes de regresión) y no el estudio de la estructura de varianzas y covarianza entre las medidas repetidas.
5. Se obtienen estimaciones consistentes de los coeficientes de regresión y sus respectivos errores estándar incluso ante una mala especificación de la estructura de varianzas y covarianzas. Es decir, ante una mala especificación, la disminución del poder de la estimación de los coeficientes de regresión es baja para un tamaños de muestra grandes.

Dado que los **GEE** son una extensión de los **MLG** para datos correlacionados, las tres especificaciones de estos modelos se mantienen dado el tipo de desenlace.

Para el caso de un desenlace continuo, el predictor lineal y la función de enlace producen la siguiente expresión:

$$\mu_{ij} = x'_{ij}\beta$$

donde  $\mu_{ij}$  es la media del desenlace del sujeto  $i$  en el tiempo  $j$ ;  $x_{ij}$  son las  $p$  variables independientes observadas en el sujeto  $i$  en el tiempo  $j$  y  $\beta$  es un vector de dimensión  $p + 1$  [ $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ ] correspondiente a los coeficientes de regresión. La expresión de la varianza de  $y_{ij}$  es:

$$V(y_{ij}) = \phi V(\mu_{ij})$$

Adicional a las tres especificaciones anteriores, se realiza una adicional sobre la matriz de varianzas y covarianzas de las medidas repetidas: "working correlation structure"; denotada como  $R(a)$  de dimensiones  $n \times n$  para los individuos evaluados en todos los  $n$  momentos. Si el sujeto no fue observado en todas las  $n$  observaciones ( $n_i \leq n$ ), su matriz de varianzas y covarianzas se denota por  $R_i(a)$  y se toma retirando de la matriz  $R(a)$  las filas y columnas correspondientes a los momentos no observados.

Algunas de las estructuras predefinidas que se utilizan para  $R(a)$  y que se encuentran implementadas en varios programas de estadística son las siguientes:

- Independencia.

$$\begin{bmatrix} - & 0 & 0 & 0 & \dots & 0 \\ & - & 0 & 0 & \dots & 0 \\ & & - & 0 & \dots & 0 \\ & & & : & : & : \\ & & & & - & 0 \\ & & & & & - \end{bmatrix}$$

- Exchangeable:



$$\begin{bmatrix} - & \rho & \rho & \rho & \dots & \rho \\ & - & \rho & \rho & \dots & \rho \\ & & - & \rho & \dots & \rho \\ & & & \vdots & \vdots & \vdots \\ & & & & - & \rho \\ & & & & & - \end{bmatrix}$$

- m\_dependent (banded)

$$\begin{bmatrix} - & \rho_1 & \rho_2 & \dots & \rho_m & 0 & \dots & 0 \\ & - & \rho_1 & \rho_2 & \dots & \rho_m & \dots & 0 \\ & & - & \rho_1 & \dots & \rho_{m-1} & \dots & 0 \\ & & & & & \vdots & \vdots & \\ & & & & & - & \rho_1 & \rho_2 \\ & & & & & & - & \rho_1 \\ & & & & & & & - \end{bmatrix}$$

- Auto regresivo de orden 1 [AR(1)]

$$\begin{bmatrix} - & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ & - & \rho & \rho^2 & \dots & \rho^{n-2} \\ & & - & \rho & \dots & \rho^{n-3} \\ & & & & \vdots & \vdots \\ & & & & - & \rho \\ & & & & & - \end{bmatrix}$$

- No estructurada

$$\begin{bmatrix} - & \rho_{12} & \rho_{13} & \rho_{14} & \dots & \rho_{1n} \\ & - & \rho_{23} & \rho_{24} & \dots & \rho_{2n} \\ & & - & \rho_{34} & \dots & \rho_{3n} \\ & & & & \vdots & \vdots \\ & & & & - & \rho_{(n-1)n} \\ & & & & & - \end{bmatrix}$$

Una recomendación para la elección de esta matriz, es tomar la estructura que más se acerque a las correlaciones observadas y que se considere la más “sencilla”, es decir, que requiera estimar el menor número de parámetros  $\rho_{jj'}$ .

La estimación de los parámetros de regresión del modelo ( $\beta$ ) corresponden a la solución de la siguiente ecuación:

$$\sum_{i=1}^N D_i' [V(\hat{a})]^{-1} (y_i - \mu_i) = 0$$

donde  $\hat{a}$  es la estimación de  $a$  y  $D_i = \frac{\partial \mu_i}{\partial \beta}$ .

Resolviendo la expresión anterior, para el caso de un desenlace continuo se obtiene la siguiente estimación:

$$\hat{\beta} = \left[ \sum_{i=1}^N X_i' [R_i(\hat{a})]^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^N X_i' [R_i(\hat{a})]^{-1} y_i \right]$$

La expresión del modelo con los  $\hat{\beta}$  tiene la siguiente forma:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{i11} \\ x_{i21} \\ \vdots \\ x_{in1} \end{bmatrix} + \beta_2 \begin{bmatrix} x_{i12} \\ x_{i22} \\ \vdots \\ x_{in2} \end{bmatrix} + \dots + \beta_p \begin{bmatrix} x_{i1p} \\ x_{i2p} \\ \vdots \\ x_{inp} \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in} \end{bmatrix}$$

Los parámetros  $\beta$  se interpretan como la diferencia en la media del desenlace cuando cambia en una unidad la variable correspondiente a este coeficiente, controlando o manteniendo fijas el resto de variables independientes.

## Prueba Wald sobre los parámetros del modelo.

Para evaluar la hipótesis nula  $H_0 : \beta' s = 0$ , es decir que un grupo de coeficientes son iguales a cero, el estadístico de prueba es igual a:

$$\chi^2 = \hat{\beta}' C' (CV(\hat{\beta}) C')^{-1} C \hat{\beta} \sim \chi_q^2$$

donde  $C$  es una matriz de dimensiones  $q \times p + 1$ ,  $q$  es el número de parámetros a evaluar incluidos en la hipótesis nula, y  $p$  el número de variables independientes del modelo completo. Dentro de la matriz  $C$ , en cada fila se colocan ceros en todas las posiciones menos en una, que corresponde a la posición de cada coeficiente incluido dentro de la hipótesis nula; por ejemplo si se tienen  $p = 5$  variables independientes y se quiere probar  $H_0 : \beta_3 = \beta_4 = 0$ , la matriz  $C$  tendría la siguiente forma:

$$\begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Como caso específico, si  $C$  es igual a  $C = [0 \ 0 \ 0 \ 1 \ 0 \ 0]$  la hipótesis nula que se está evaluando es:  $H_0 : \beta_3 = 0$ .

## Aplicación.

Escenario: Se quiere evaluar la relación entre el número de días de estancia hospitalaria y las variables: edad, sexo (H=1, M=0), número de días de soporte ventilatorio (sv) e índice de severidad (apache). Se recolectó información de una cohorte de pacientes que fueron hospitalizados durante el año 2003. Dado que un mismo paciente pudo haber estado hospitalizado en más de una ocasión durante este año, se registro el identificador de cada sujeto.

A continuación, cargamos la base de datos en una estructura **a lo largo**:

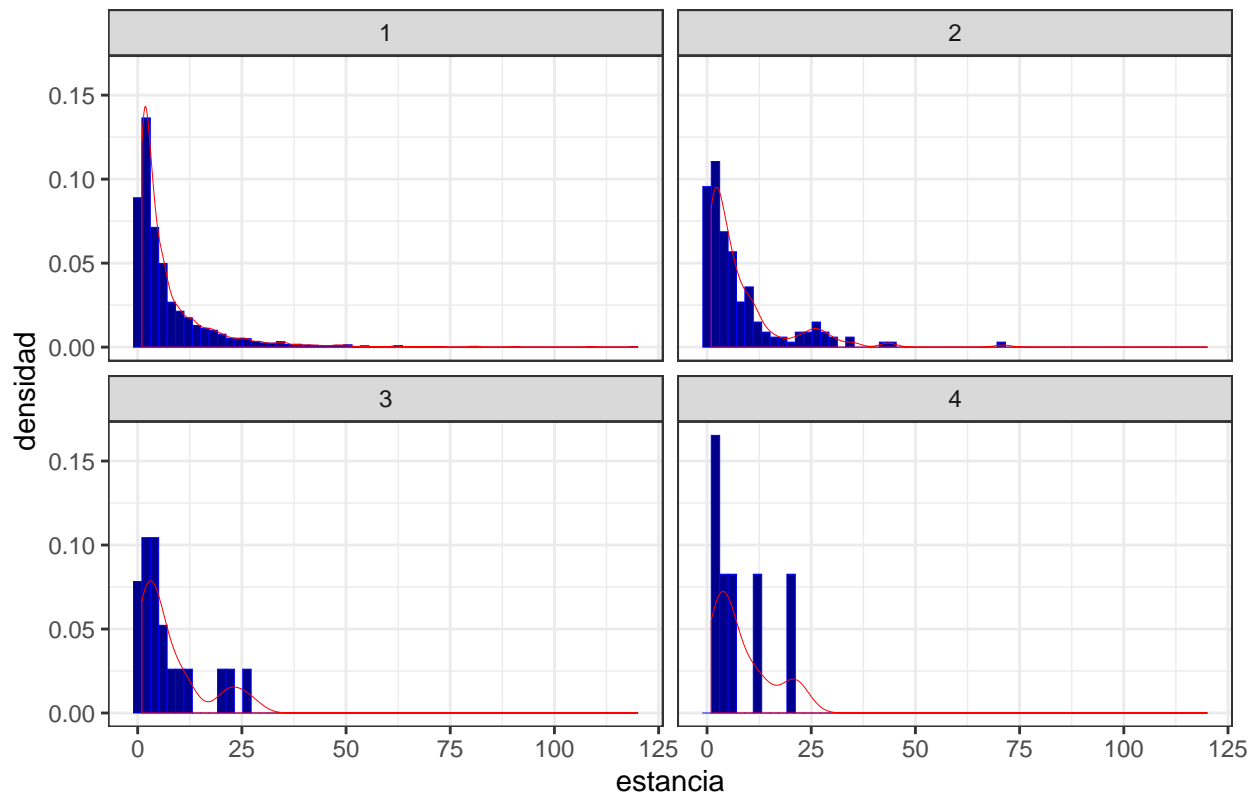
```
library(rio)
BD<-import("estancia.xlsx")
```

Observamos la forma en que se distribuye la estancia en cada tiempo y probamos la hipótesis de normalidad.

- Histograma

```
library(tidyverse)
ggplot(BD,aes(x=estancia))+
  geom_histogram(bins=60,aes(y = ..density..),fill="darkblue",
                color = "blue",size=0.1)+
  geom_density(col="red",size=0.1)+
  labs(title="Histograma - estancia por tiempos",y="densidad",
        x="estancia")+
  theme_bw()+
  facet_wrap(~tiempo)
```

## Histograma – estancia por tiempos

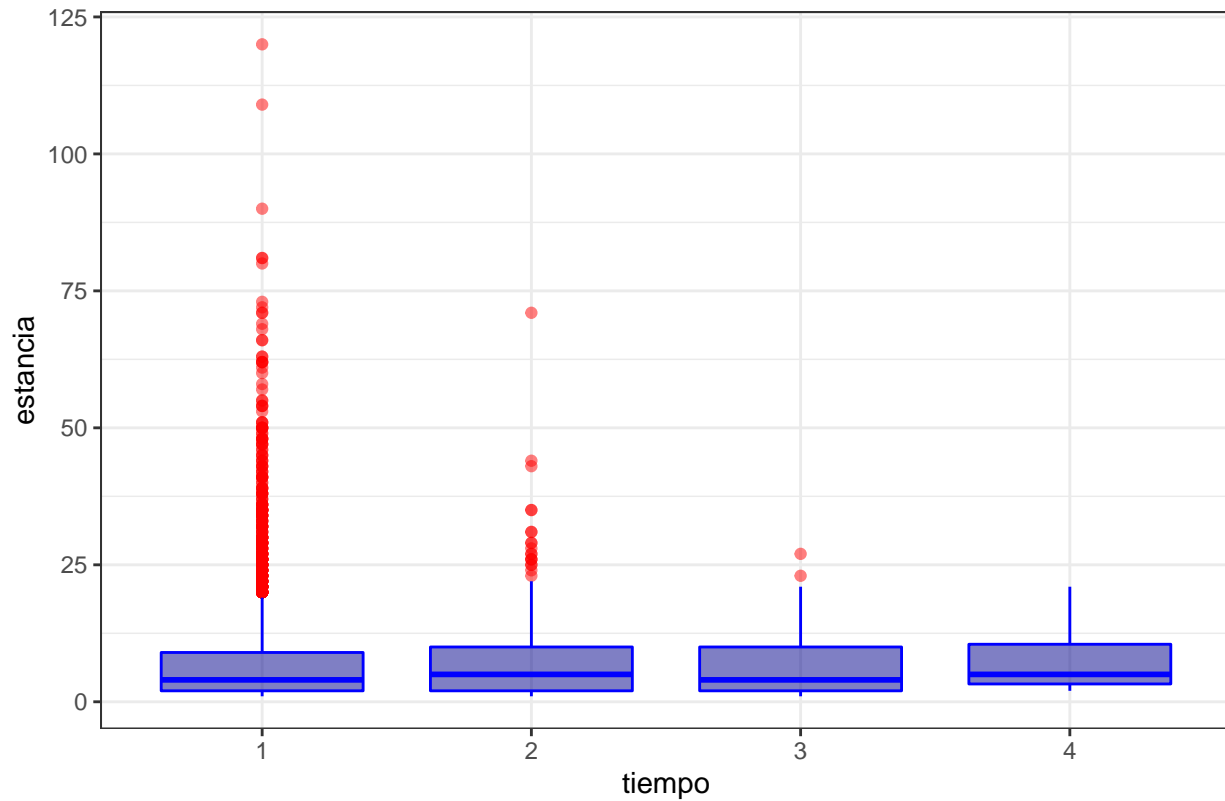


*(no se cumple el supuesto de normalidad pero continuaremos con el análisis para efectos prácticos)*

- Boxplot

```
ggplot(BD, aes(x=factor(tiempo), y=estancia)) +  
  geom_boxplot(alpha=0.5, fill="darkblue", color="blue", outlier.color="red") +  
  labs(title="Boxplot - estancia por tiempos", y="estancia", x="tiempo") +  
  theme_bw()
```

Boxplot – estancia por tiempos



- Shapiro wilk

```
by(BD,BD$tiempo, function(BD) shapiro.test(BD$estancia))
```

```
## BD$tiempo: 1
##
## Shapiro-Wilk normality test
##
## data: BD$estancia
## W = 0.6408, p-value < 2.2e-16
##
## -----
## BD$tiempo: 2
##
## Shapiro-Wilk normality test
##
## data: BD$estancia
## W = 0.70365, p-value < 2.2e-16
##
## -----
## BD$tiempo: 3
##
## Shapiro-Wilk normality test
##
## data: BD$estancia
## W = 0.77771, p-value = 0.0005519
##
```

```
## -----
## BD$tiempo: 4
##
## Shapiro-Wilk normality test
##
## data: BD$estancia
## W = 0.8368, p-value = 0.1226
```

Ahora exploramos la matriz de correlación de los desenlaces en los cuatro tiempos.

```
BD2<-reshape(BD,timevar = "tiempo",idvar=c("id"),direction = "wide")
estancia_wide<-data.frame(BD2$estancia.1,BD2$estancia.2,BD2$estancia.3,
                          BD2$estancia.4)
cor(estancia_wide,use = "pairwise.complete.obs", method = "pearson")
```

```
##                BD2.estancia.1 BD2.estancia.2 BD2.estancia.3 BD2.estancia.4
## BD2.estancia.1      1.0000000      0.2232210     -0.09602102      0.14354640
## BD2.estancia.2      0.22322099      1.0000000      0.12451061      1.00000000
## BD2.estancia.3     -0.09602102      0.1245106      1.00000000     -0.09578263
## BD2.estancia.4      0.14354640      1.0000000     -0.09578263      1.00000000
```

Trabajaremos con la opción “No estructurada”.

La función `gee` del paquete `gee` realiza la estimación de los coeficientes de regresión y de los parámetros de asociación de la siguiente forma:

```
library("gee")
gee1<-gee(estancia~edad+sexo+sv+apache,id=id, data=BD,family = gaussian,
          corstr = "unstructured")
```

```
## (Intercept)      edad      sexo      sv      apache
## 3.05955454 0.01054431 0.22272883 0.56645728 0.04115235
```

```
# working correlation forms: independence exchangeable
# unstructured stat_M_dep AR-M (Mv)
```

```
summary(gee1)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link: Identity
## Variance to Mean Relation: Gaussian
## Correlation Structure: Unstructured
##
## Call:
## gee(formula = estancia ~ edad + sexo + sv + apache, id = id,
##      data = BD, family = gaussian, corstr = "unstructured")
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -104.690686 -3.089863 -1.671585  1.296743  66.824414
##
##
```

```

## Coefficients:
##           Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept) 3.05489536 0.434854414  7.0250991 0.431425273 7.0809374
## edad        0.01048030 0.006619889  1.5831540 0.007136166 1.4686184
## sexo        0.22103298 0.268859793  0.8221124 0.287632432 0.7684564
## sv          0.56612009 0.010320208 54.8554912 0.077773602 7.2790777
## apache      0.04148005 0.016539827  2.5078888 0.020732749 2.0007018
##
## Estimated Scale Parameter:  57.6979
## Number of Iterations:  2
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.00000000  0.107181037 -0.001027667 -0.008289615
## [2,]  0.107181037  1.000000000  0.001363131 -0.001134144
## [3,] -0.001027667  0.001363131  1.000000000  0.001507596
## [4,] -0.008289615 -0.001134144  0.001507596  1.000000000

```

- Comparación de valores observados vs. ajustados

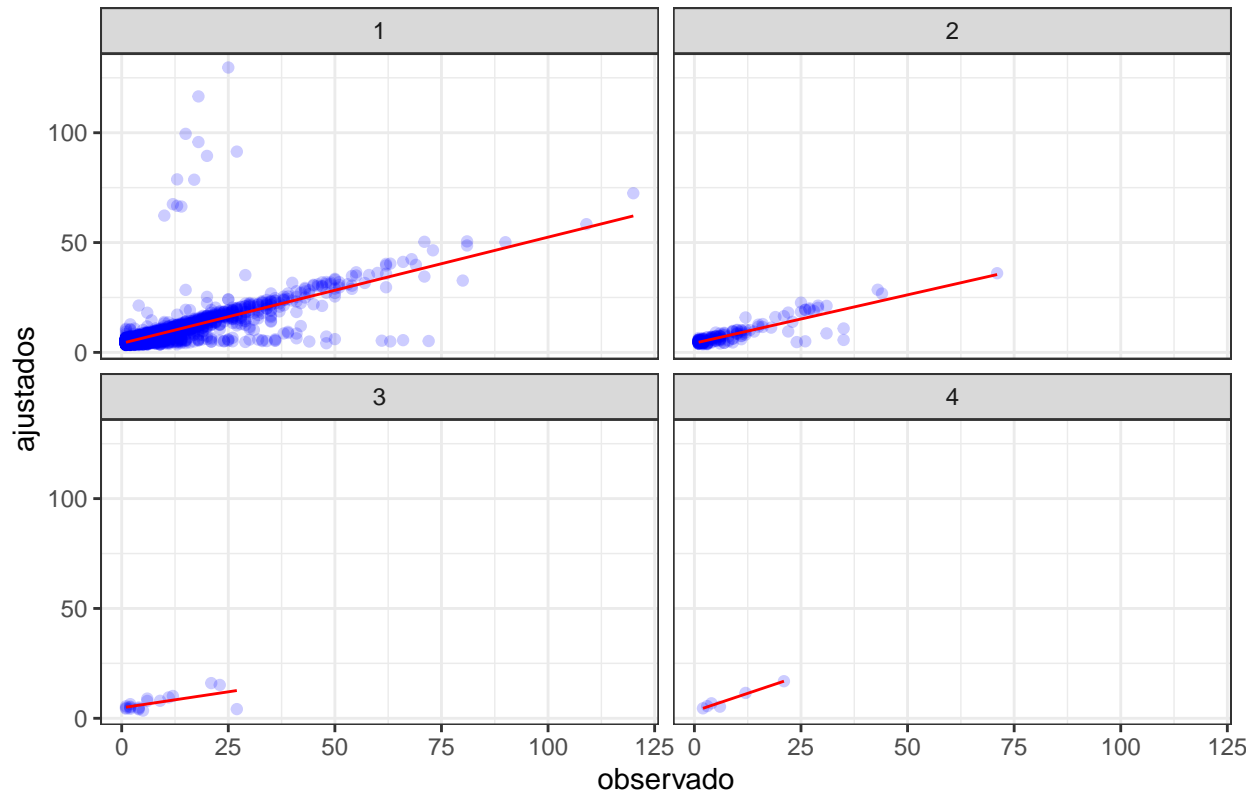
```

fit<-data.frame(BD$estancia,gee1$fitted.values,BD$tiempo,gee1$residuals)

ggplot(fit,aes(BD.estancia,gee1.fitted.values))+
  geom_point(alpha=0.2,color="blue")+
  geom_smooth(method = lm,color="red",size=0.5,se=FALSE)+
  theme_bw()+
  labs(title = "Observados vs ajustados por tiempo",
        x="observado",y="ajustados")+
  facet_wrap(~BD.tiempo)

```

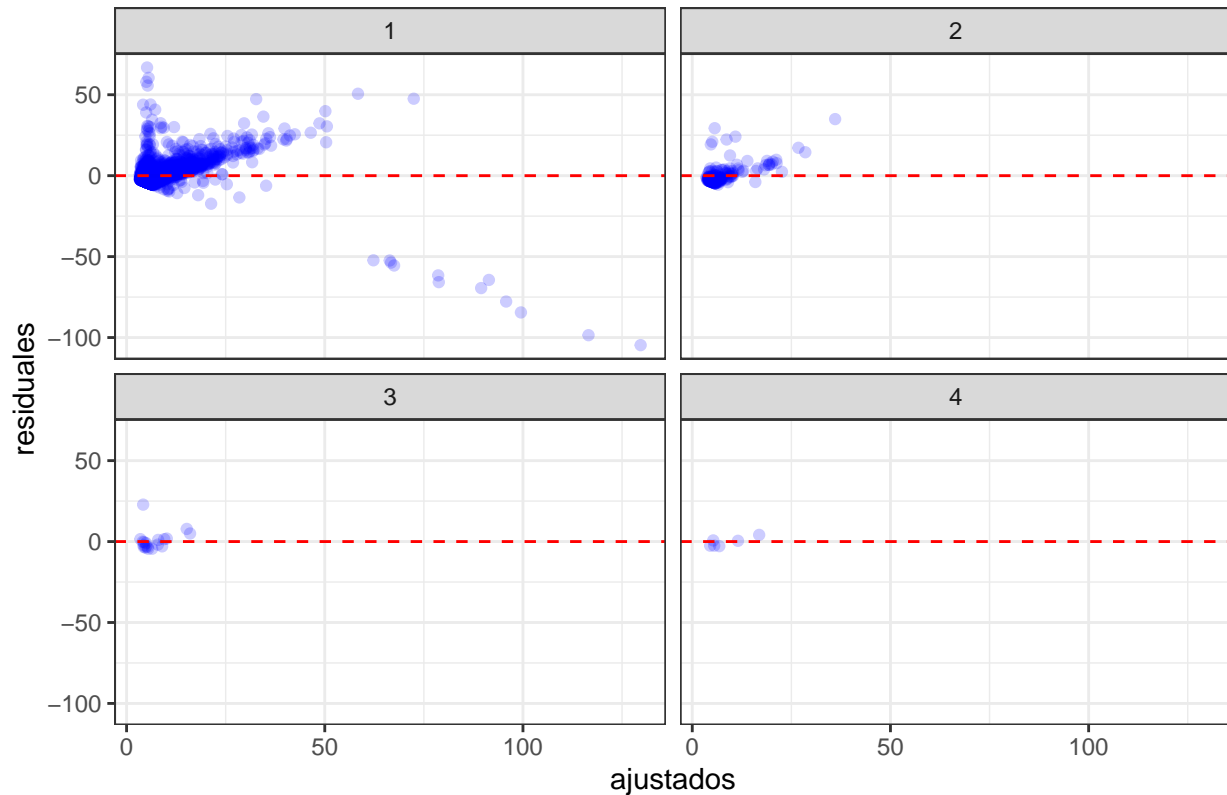
## Observados vs ajustados por tiempo



- Comparación valores ajustados versus residuales.

```
ggplot(fit,aes(gee1.fitted.values,gee1.residuals))+  
  geom_point(alpha=0.2,color="blue")+  
  geom_hline(yintercept=c(0), lty=2,col="red")+  
  theme_bw()+  
  labs(title = "Ajustados vs residuales por tiempo",x="ajustados",  
        y="residuales")+  
  facet_wrap(~BD.tiempo)
```

## Ajustados vs residuales por tiempo



- Prueba de Wald

La prueba de Wald para evaluar distintas hipótesis cambiando la definición de la matriz  $C$ , se puede obtener desarrollando la siguiente expresión:

$$\chi^2 = \hat{\beta}' C' (CV(\hat{\beta}) C')^{-1} C \hat{\beta} \sim \chi_q^2$$

Por ejemplo, si queremos reconstruir el estadístico de prueba para el coeficiente  $\beta_3$  (sv) se puede correr el siguiente código:

```
A<-c(0,0,0,1,0)
C<-rbind(A)
```

```
gee1$coefficients
```

```
## (Intercept)      edad      sexo      sv      apache
## 3.05489536  0.01048030  0.22103298  0.56612009  0.04148005
```

```
gee1$robust.variance
```

```
##          (Intercept)      edad      sexo      sv
## (Intercept)  0.1861277661 -2.025992e-03 -0.0496307889 -8.109649e-03
## edad        -0.0020259921  5.092486e-05  0.0005891417  -7.016024e-05
## sexo        -0.0496307889  5.891417e-04  0.0827324158  -7.443836e-03
## sv          -0.0081096494  -7.016024e-05 -0.0074438358  6.048733e-03
## apache     -0.0003371003  -3.914603e-05  0.0006623960  -9.107687e-04
##          apache
## (Intercept) -3.371003e-04
```



```
## edad      -3.914603e-05
## sexo      6.623960e-04
## sv        -9.107687e-04
## apache    4.298469e-04
```

```
x2_b3<-t(gee1$coefficients)%*%t(C)%*%solve(C%*%gee1$robust.variance%*%
t(C))%*%C%*%gee1$coefficients
```

```
z<-sqrt(x2_b3)
p_3<-1-pchisq(x2_b3,1)
z
```

```
##          [,1]
## [1,] 7.279078
```

```
p_3
```

```
##          [,1]
## [1,] 3.360645e-13
```

Una prueba global sobre el modelo completo dejando únicamente el intercepto, se puede obtener de la siguiente forma:

```
A1<-c(0,1,0,0,0)
A2<-c(0,0,1,0,0)
A3<-c(0,0,0,1,0)
A4<-c(0,0,0,0,1)
C<-rbind(A1,A2,A3,A4)
```

```
gee1$coefficients
```

```
## (Intercept)      edad      sexo      sv      apache
## 3.05489536  0.01048030  0.22103298  0.56612009  0.04148005
```

```
gee1$robust.variance
```

```
##          (Intercept)      edad      sexo      sv
## (Intercept) 0.1861277661 -2.025992e-03 -0.0496307889 -8.109649e-03
## edad        -0.0020259921  5.092486e-05  0.0005891417  -7.016024e-05
## sexo        -0.0496307889  5.891417e-04  0.0827324158  -7.443836e-03
## sv          -0.0081096494  -7.016024e-05 -0.0074438358  6.048733e-03
## apache      -0.0003371003  -3.914603e-05  0.0006623960  -9.107687e-04
##          apache
## (Intercept) -3.371003e-04
## edad        -3.914603e-05
## sexo         6.623960e-04
## sv          -9.107687e-04
## apache       4.298469e-04
```

```
x2_full<-t(gee1$coefficients)%*%t(C)%*%solve(C%*%gee1$robust.variance%*%t(C))%*%C%*%gee1$coefficients
p_full<-1-pchisq(x2_full,4)
```

```
x2_full
```

```
##          [,1]
## [1,] 151.1693
```

```
p_full
```

```
##      [,1]
## [1,]    0
```

- Comparación de modelos no anidados

Ajuste de un segundo modelo bajo la estructura “exchangeable”:

```
gee2<-gee(estancia~edad+sexo+sv+apache,id=id, data=BD,family = gaussian,
          corstr = "exchangeable")
```

```
## (Intercept)      edad      sexo      sv      apache
## 3.05955454 0.01054431 0.22272883 0.56645728 0.04115235
```

```
summary(gee2)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link: Identity
## Variance to Mean Relation: Gaussian
## Correlation Structure: Exchangeable
##
## Call:
## gee(formula = estancia ~ edad + sexo + sv + apache, id = id,
##      data = BD, family = gaussian, corstr = "exchangeable")
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -104.721945 -3.089815 -1.672548  1.299055  66.826965
##
##
## Coefficients:
##      Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept) 3.05578062 0.434888718  7.0265806 0.431606985 7.0800073
## edad      0.01049028 0.006619947  1.5846477 0.007137528 1.4697364
## sexo      0.21724887 0.268840056  0.8080971 0.287428007 0.7558375
## sv        0.56625705 0.010318797 54.8762667 0.077776089 7.2806059
## apache    0.04146069 0.016538901  2.5068586 0.020741209 1.9989522
##
## Estimated Scale Parameter: 57.6979
## Number of Iterations: 2
##
## Working Correlation
##      [,1]      [,2]      [,3]      [,4]
## [1,] 1.00000000 0.08550972 0.08550972 0.08550972
## [2,] 0.08550972 1.00000000 0.08550972 0.08550972
## [3,] 0.08550972 0.08550972 1.00000000 0.08550972
## [4,] 0.08550972 0.08550972 0.08550972 1.00000000
```

Quasilikelihood under the independence model criterion (QIC)

```
library(MuMIn)
```

```
## Warning: package 'MuMIn' was built under R version 3.6.1
```

```
QIC(gee1)
```

```
##      QIC  
## 3458.038
```

```
QIC(gee2)
```

```
##      QIC  
## 3458.053
```

Se elige el modelo con menor **QIC**.

# Modelos GEE - Desenlace dicotómico y de conteo.

## Desenlace Dicotómico

Los modelos **GEE** cuando el desenlace tiene dos posibles respuesta,  $y_{ij} = \begin{cases} 1 & \text{si es acierto} \\ 0 & \text{si es fracaso} \end{cases}$ , se asume que proviene de una distribución binomial y en consecuencia se realizan las tres especificaciones descritas anteriormente en los modelos lineales generalizados acorde a este desenlace (predictor lineal, función de enlace y descomposición de la varianza del desenlace).

El predictor lineal ( $X_{ij}\beta$ ) y la función de enlace (*logit*) en este caso forman la siguiente expresión:

$$\text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1 - \mu_{ij}}\right) = X_{ij}\beta$$

donde obtenemos que  $\mu_{ij}$  es igual a:

$$\mu_{ij} = \frac{1}{1 + e^{-X_{ij}\beta}}$$

La  $V(y_{ij}) = \phi V(\mu_{ij})$  es igual a  $V(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$  (parámetro escalar  $\phi = 1$ ).

La interpretación de los coeficientes  $\beta$  indican el cambio en el logaritmo del chance de presentar el desenlace cuando cambia en una unidad la covariables dejando el resto igual. Este cambio se puede presentar ya sea entre dos momentos o entre dos sujetos. Típicamente se interpreta el valor exponencial de  $\beta$  que corresponde a la razón de disparidades (Odds Ratio - OR).

También se puede utilizar como función de enlace la función **log**, llegando a la siguiente expresión:

$$\log(\mu_{ij}) = X_{ij}\beta$$

donde el valor exponencial de  $\beta$  corresponde al riesgo relativo (RR), dado el caso en que los datos provengan de un estudio de cohortes.

## Desenlace de conteo

Cuando el desenlace es el número de eventos que se presentan en un lapso de tiempo determinado (variable que sigue una distribución de Poisson), el predictor lineal ( $\eta_i$ ) y la función de enlace (*log*) correspondientes, forman la siguiente expresión:

$$\log(\mu_{ij}) = X_{ij}\beta$$

obteniendo que

$$\mu_{ij} = e^{(X_{ij}\beta)}$$

La descomposición de la varianza del desenlace  $V(y_{ij}) = \mu_{ij}$  ( $V(\mu_{ij}) = \mu_{ij}$  y  $\phi = 1$ ).

La interpretación de los coeficientes  $\beta$  indican el cambio en el logaritmo de  $\mu_{ij}$  por una unidad de cambio en la covariable dejando el resto igual. Al igual que en el caso anterior, este cambio se puede dar entre tiempos o entre sujetos. La interpretación del valor exponencial del coeficiente  $\beta$  indica la razón entre el número de eventos esperados (**IRR** - incident rate ratios) cuando cambia en una unidad la covariable y dejando el resto igual.

Las características descritas en el capítulo anterior sobre los modelos **GEE** con desenlace continuo se mantienen para los desenlaces dicotómico y de conteo, por lo tanto las estructuras descritas de la matriz de varianzas y covarianza entre los tiempos también se incorporan en las estimaciones que se realizan. De igual forma, la estadística de Wald para evaluar la significancia de uno o varios coeficientes de regresión y la estadística **QIC** también se utilizan.

## Aplicación.

Escenario: En 98 niños recién nacidos se realizaron cuatro mediciones del peso, cada tres meses; este peso se categorizó en desnutrido (1) y normal (0) (`peso_dic`). También se evaluó en los mismos periodos el número de visitas al médico (`visitas`). Como variables independientes se midieron las siguientes variables: sexo (H=0, M=1), si recibió lactancia materna (No = 0, Si = 1) y horas de sueño (cuantitativa), esta última variable cambia en el tiempo.

Primero trabajaremos con el desenlace `peso_dic` y las variables independientes.

Realizamos el cargue de la base de datos:

```
library(rio)
RN<-import("RN peso.xlsx")
round(prop.table(table(RN$t1,RN$peso_dic),1),3)*100
```

```
##
##      0      1
## 0 57.1 42.9
## 1 74.5 25.5
## 2 68.5 31.5
## 3 79.6 20.4
```

- Matriz de correlación

```
RN2<-reshape(RN,timevar = "t1",idvar=c("id"),direction = "wide")
peso_wide<-data.frame(RN2$peso_dic.0,RN2$peso_dic.1,RN2$peso_dic.2,
                      RN2$peso_dic.3)
cor(peso_wide,use = "pairwise.complete.obs", method = "pearson")
```

```
##                RN2.peso_dic.0 RN2.peso_dic.1 RN2.peso_dic.2 RN2.peso_dic.3
## RN2.peso_dic.0      1.0000000      0.06081636      0.11285524     -0.05230229
## RN2.peso_dic.1      0.06081636      1.0000000      0.07644421      0.01861026
## RN2.peso_dic.2      0.11285524      0.07644421      1.0000000      0.11459627
## RN2.peso_dic.3     -0.05230229      0.01861026      0.11459627      1.0000000
```

- Modelo dicotómico con función de enlace logit

```
library("gee")
RN<-RN[order(RN$id),] # ordenar la base por id para utilizar la función gee
gee1<-gee(peso_dic~lactancia+sexo+sueno,id=id, data=RN,
          family = binomial(link="logit"),corstr = "unstructured",
          scale.fix = TRUE)
```

```
## (Intercept)  lactancia      sexo      sueno
## -0.08083343 -0.88660873  0.07011349 -0.04769036
```

```
summary(gee1)
```

```
##
```

```

## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                      Logit
## Variance to Mean Relation: Binomial
## Correlation Structure:     Unstructured
##
## Call:
## gee(formula = peso_dic ~ lactancia + sexo + sueno, id = id, data = RN,
##      family = binomial(link = "logit"), corstr = "unstructured",
##      scale.fix = TRUE)
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -0.4686475 -0.4155877 -0.3147358  0.5401590  0.7101486
##
##
## Coefficients:
##              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -0.06181840 0.37256051 -0.1659285  0.32604705 -0.1895996
## lactancia   -0.56472253 0.17194048 -3.2844070  0.15983798 -3.5330935
## sexo        -0.12956644 0.19605256 -0.6608760  0.22161893 -0.5846361
## sueno       -0.01559565 0.04707863 -0.3312682  0.04206035 -0.3707923
##
## Estimated Scale Parameter:  1
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.0000000  0.1949421 -0.2786268 -0.5575633
## [2,]  0.1949421  1.0000000  0.4870888  0.1457615
## [3,] -0.2786268  0.4870888  1.0000000  0.4277563
## [4,] -0.5575633  0.1457615  0.4277563  1.0000000

```

```

# Calculo de los OR
OR<-exp(gee1$coefficients)
EE1<-sqrt(c(gee1$robust.variance[1,1],gee1$robust.variance[2,2],
            gee1$robust.variance[3,3],gee1$robust.variance[4,4]))
LI<-exp(gee1$coefficients-qnorm(0.975)*EE1)
LS<-exp(gee1$coefficients+qnorm(0.975)*EE1)
OREst<-data.frame(OR,LI,LS); OREst

```

```

##              OR      LI      LS
## (Intercept) 0.9400536 0.4961590 1.7810838
## lactancia   0.5685179 0.4156136 0.7776756
## sexo        0.8784762 0.5689677 1.3563519
## sueno       0.9845253 0.9066195 1.0691256

```

```

# objetos que se pueden consultar:
#gee1$fitted.values

```

```
#gee1$linear.predictors
```

```
library(tidyverse)
```

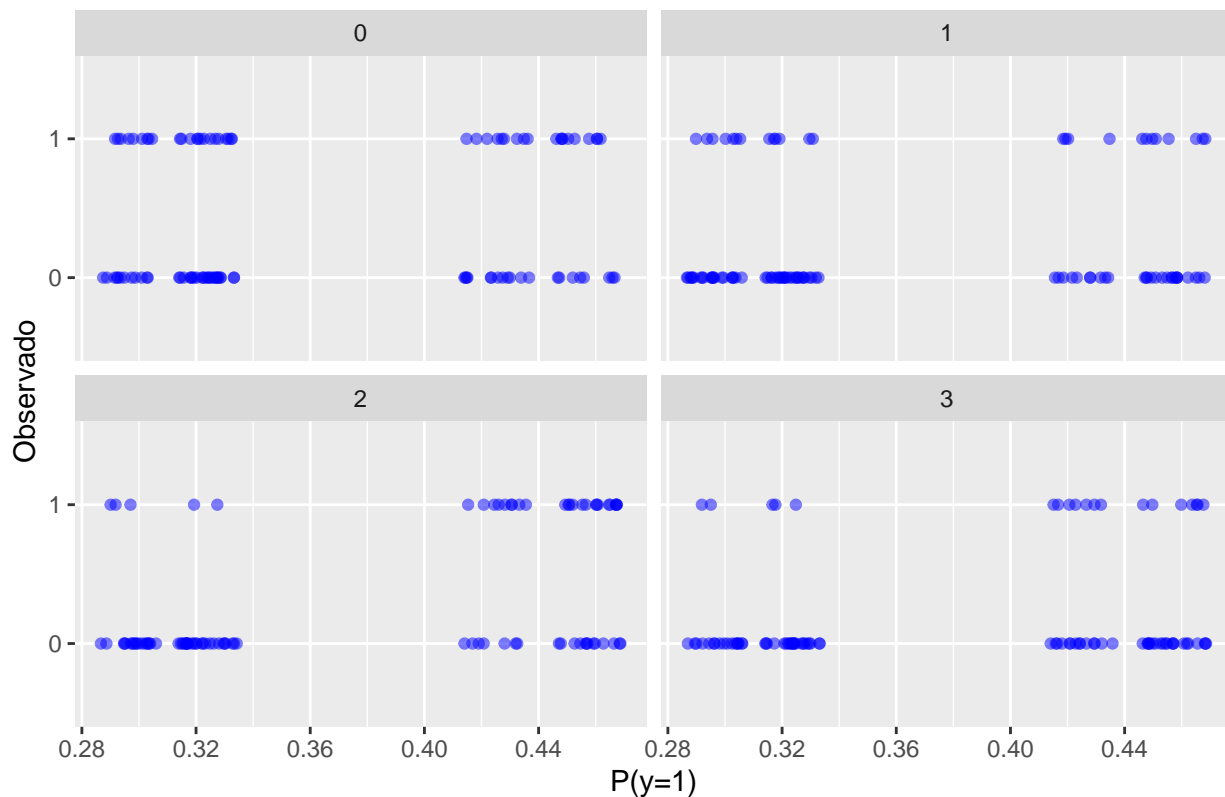
```
prob<-data.frame(RN$id,RN$t1,RN$peso_dic,RN$lactancia,gee1$fitted.values)
```

```
g<-group_by(prob,RN.t1,RN.peso_dic)  
summarise(g,mean(gee1.fitted.values))
```

```
## # A tibble: 8 x 3  
## # Groups:   RN.t1 [4]  
##   RN.t1 RN.peso_dic `mean(gee1.fitted.values)`  
##   <dbl>   <dbl>           <dbl>  
## 1     0         0           0.358  
## 2     0         1           0.371  
## 3     1         0           0.360  
## 4     1         1           0.375  
## 5     2         0           0.355  
## 6     2         1           0.423  
## 7     3         0           0.375  
## 8     3         1           0.407
```

```
ggplot(prob,aes(y=factor(RN.peso_dic),x=gee1.fitted.values))+  
  geom_point(alpha=0.5,col="blue")+  
  labs(title="Probabilidad de desnutrición vs observados",y="Observado", x="P(y=1)")+  
  facet_wrap(~RN.t1)
```

## Probabilidad de desnutrición vs observados



Se puede definir valores específicos para la matriz de trabajo.

```
WC<-matrix(c(1,.0450843,.1399795,-.0859096, .0450843,1,.105207,.0872144,.1399795,
             .105207,1, -.0000467, -.0859096,.0872144, -.0000467,1),nrow=4,ncol=4)
```

```
gee_WC<-gee(peso_dic~lactancia+sexo+sueno,id=id, data=RN,R=WC,
            family = binomial(link="logit"),corstr = "fixed",scale.fix = TRUE)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)  lactancia      sexo      sueno
## -0.08083343 -0.88660873  0.07011349 -0.04769036
```

```
summary(gee_WC)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                               Logit
## Variance to Mean Relation: Binomial
## Correlation Structure:      Fixed
##
## Call:
## gee(formula = peso_dic ~ lactancia + sexo + sueno, id = id, data = RN,
##      R = WC, family = binomial(link = "logit"), corstr = "fixed",
##      scale.fix = TRUE)
##
## Summary of Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4406287 -0.3771431 -0.2181472  0.5797063  0.8011930
##
##
## Coefficients:
##              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -0.11844329 0.49627598 -0.2386642  0.43374441 -0.2730716
## lactancia   -0.87799699 0.22682507 -3.8708111  0.23654831 -3.7117026
## sexo         0.04398739 0.24548827  0.1791833  0.24447218  0.1799280
## sueno       -0.04044444 0.06397536 -0.6321877  0.05783839 -0.6992663
##
## Estimated Scale Parameter:  1
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.0000000 0.0450843 0.1399795 -0.0859096
## [2,]  0.0450843 1.0000000 0.1052070  0.0872144
## [3,]  0.1399795 0.1052070 1.0000000 -0.0000467
## [4,] -0.0859096 0.0872144 -0.0000467  1.0000000
```

Aplicación del estadístico de Wald, para evaluar la hipótesis  $H_0 : \beta_{sexo} = \beta_{sueno} = 0$

```
A1<-c(0,0,1,0)
```

```
A2<-c(0,0,0,1)
```



```

C<-rbind(A1,A2)

x2_wald<-t(gee1$coefficients)%*%t(C)%*%solve(C%*%gee1$robust.variance%*%t(C))%*%
  C%*%gee1$coefficients

x2_wald

##           [,1]
## [1,] 0.4263943

p_value<-1-pchisq(x2_wald,2)
p_value

##           [,1]
## [1,] 0.8079968

```

Segundo modelo modificando la matriz de trabajo:

```

gee2<-gee(peso_dic~lactancia+sexo+sueno,id=id, data=RN,
          family = binomial(link="logit"),corstr = "exchangeable",scale.fix = TRUE)

```

```

## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
## (Intercept)  lactancia      sexo      sueno
## -0.08083343 -0.88660873  0.07011349 -0.04769036

```

```
summary(gee2)
```

```

##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                               Logit
## Variance to Mean Relation: Binomial
## Correlation Structure:               Exchangeable
##
## Call:
## gee(formula = peso_dic ~ lactancia + sexo + sueno, id = id, data = RN,
##     family = binomial(link = "logit"), corstr = "exchangeable",
##     scale.fix = TRUE)
##
## Summary of Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4488768 -0.3702524 -0.2164761  0.5780531  0.8073239
##
##
## Coefficients:
##              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -0.07760098 0.50049651 -0.1550480  0.44373348 -0.1748820
## lactancia   -0.88492030 0.22930261 -3.8591811  0.24103912 -3.6712725
## sexo         0.06665050 0.24592498  0.2710196  0.25232315  0.2641474
## sueno       -0.04786156 0.06449273 -0.7421233  0.05872252 -0.8150461
##
## Estimated Scale Parameter: 1

```

```
## Number of Iterations: 2
##
## Working Correlation
##      [,1]      [,2]      [,3]      [,4]
## [1,] 1.00000000 0.04849786 0.04849786 0.04849786
## [2,] 0.04849786 1.00000000 0.04849786 0.04849786
## [3,] 0.04849786 0.04849786 1.00000000 0.04849786
## [4,] 0.04849786 0.04849786 0.04849786 1.00000000
```

```
# Comparación de los dos modelos anteriores
```

```
library(MuMIn)
```

```
QIC(gee1)
```

```
##      QIC
## 456.1056
```

```
QIC(gee2)
```

```
##      QIC
## 459.3265
```

- Modelo con desenlace dicotómico y función de enlace log

```
library("gee")
```

```
gee3<-gee(peso_dic~lactancia+sexo+sueno,id=id, data=RN,
          family = binomial(link="log"),corstr = "unstructured",scale.fix = TRUE)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)  lactancia      sexo      sueno
## -0.67746415 -0.61572471  0.04285348 -0.03495958
```

```
summary(gee3)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                      Logarithm
## Variance to Mean Relation: Binomial
## Correlation Structure:     Unstructured
##
## Call:
## gee(formula = peso_dic ~ lactancia + sexo + sueno, id = id, data = RN,
##      family = binomial(link = "log"), corstr = "unstructured",
##      scale.fix = TRUE)
##
## Summary of Residuals:
##      Min      1Q      Median      3Q      Max
## -0.4738539 -0.4111695 -0.3170492  0.5353009  0.7097406
##
##
```

```
## Coefficients:
##           Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept) -0.711728132 0.22032255 -3.2303916  0.19460398 -3.6573154
## lactancia   -0.353480995 0.10737268 -3.2920945  0.10113875 -3.4950104
## sexo        -0.094635221 0.12048676 -0.7854408  0.13699827 -0.6907767
## sueno       -0.008592814 0.02853025 -0.3011825  0.02554253 -0.3364120
##
## Estimated Scale Parameter:  1
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.0000000 0.1947441 -0.2779614 -0.5574418
## [2,]  0.1947441 1.0000000  0.4884588  0.1479734
## [3,] -0.2779614 0.4884588  1.0000000  0.4293618
## [4,] -0.5574418 0.1479734  0.4293618  1.0000000
```

```
RR<-exp(gee3$coefficients)
EE3<-sqrt(c(gee3$robust.variance[1,1],gee3$robust.variance[2,2],
           gee3$robust.variance[3,3],gee3$robust.variance[4,4]))
```

```
LI_RR<-exp(gee3$coefficients-1.96*EE3)
LS_RR<-exp(gee3$coefficients+1.96*EE3)
```

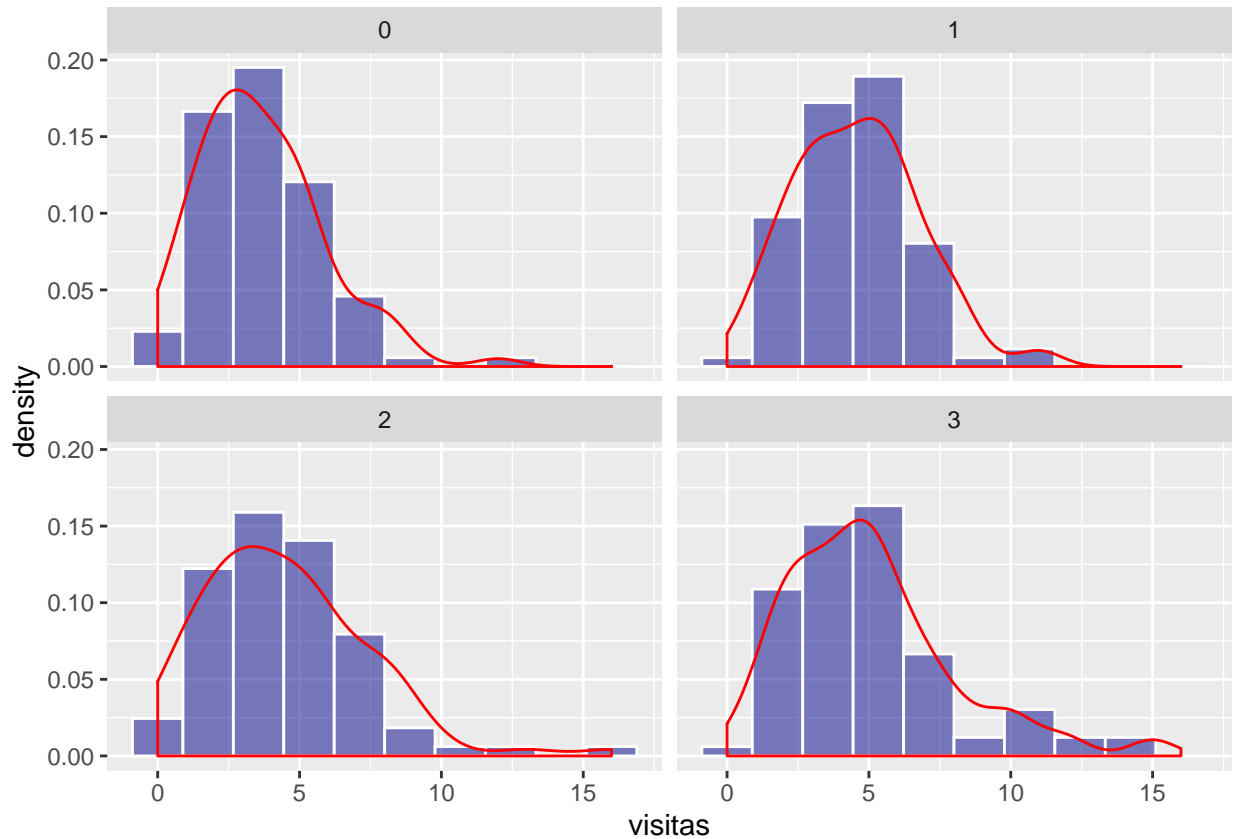
```
RREst<-data.frame(RR,LI_RR,LS_RR);RREst[2:4,]
```

```
##           RR      LI_RR      LS_RR
## lactancia 0.7022393 0.5759624 0.8562019
## sexo      0.9097047 0.6954808 1.1899144
## sueno     0.9914440 0.9430310 1.0423425
```

- Modelo con desenlace conteo

Ahora trabajaremos con el desenlace **visitas** y las variables independientes.

```
ggplot(RN,aes(x=visitas))+
  geom_histogram(bins = 10,aes(y = ..density..), alpha=0.5,
                fill="darkblue", color = "white")+
  geom_density(col="red",size=0.5)+
  facet_wrap(~t1)
```



- Ajuste del modelo

```
gee4<-gee(visitas~lactancia+sexo+sueno,id=id,data=RN,
          family=poisson(link = log),corstr="unstructured",scale.fix = TRUE)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)  lactancia      sexo      sueno
##  1.45816789 -0.15299858  0.01504248  0.01456738
```

```
summary(gee4)
```

```
##
## GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
## gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
## Link:                               Logarithm
## Variance to Mean Relation: Poisson
## Correlation Structure:               Unstructured
##
## Call:
## gee(formula = visitas ~ lactancia + sexo + sueno, id = id, data = RN,
##      family = poisson(link = log), corstr = "unstructured", scale.fix = TRUE)
##
## Summary of Residuals:
##      Min      1Q   Median     3Q      Max
```

```
## -4.8441911 -1.9601272 -0.2526101 1.3575435 11.1865763
##
##
## Coefficients:
##           Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept)  1.47726387 0.10389414 14.2189339  0.12008293 12.3020307
## lactancia   -0.13545747 0.04759116 -2.8462737  0.06161881 -2.1983136
## sexo        -0.00682737 0.04435371 -0.1539301  0.05418526 -0.1260005
## sueno        0.01305200 0.01336328  0.9767067  0.01618409  0.8064714
##
## Estimated Scale Parameter: 1
## Number of Iterations: 4
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,]  1.00000000 -0.16838653 -0.07090772 -0.21296804
## [2,] -0.16838653  1.00000000  0.04435367  0.05421648
## [3,] -0.07090772  0.04435367  1.00000000  0.01367190
## [4,] -0.21296804  0.05421648  0.01367190  1.00000000
```

```
IRR<-exp(gee4$coefficients)
```

```
EE4<-sqrt(c(gee4$robust.variance[1,1],gee4$robust.variance[2,2],
           gee4$robust.variance[3,3],gee4$robust.variance[4,4]))
```

```
LI_IRR<-exp(gee4$coefficients-1.96*EE4)
```

```
LS_IRR<-exp(gee4$coefficients+1.96*EE4)
```

```
IRREst<-data.frame(IRR,LI_IRR,LS_IRR);IRREst[2:4,]
```

```
##           IRR      LI_IRR      LS_IRR
## lactancia 0.8733163 0.7739637 0.9854227
## sexo      0.9931959 0.8931234 1.1044812
## sueno     1.0131376 0.9815044 1.0457902
```

```
#gee4$linear.predictors
```

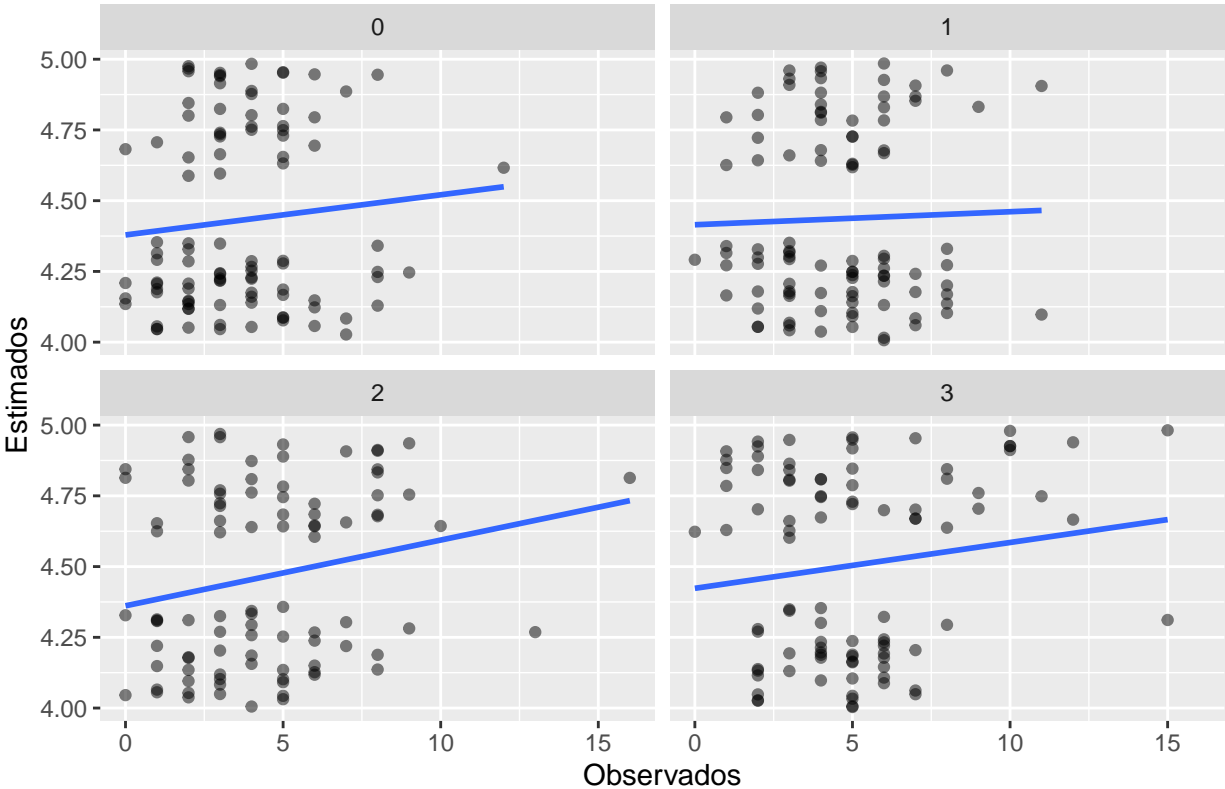
```
#gee4$fitted.values
```

- Gráfico de ajuste del modelo:

```
Fit<-data.frame(RN$id,RN$t1,RN$visitas,gee4$fitted.values)
```

```
ggplot(Fit,aes(x=RN.visitas,y=gee4.fitted.values))+
  geom_point(alpha=0.5,color="black")+
  geom_smooth(method = lm,se=FALSE)+
  labs(title="Estimados vs. observados",y="Estimados", x="Observados")+
  facet_wrap(~RN.t1)
```

Estimados vs. observados



## Modelos mixtos (MIX) - Desenlace continuo.

Los modelos de efectos mixtos (MIX) también se presentan como una extensión de los modelos lineales generalizados, pero no resuelve el incumplimiento del supuesto de independencia a través de la especificación de una “matriz de trabajo” como los modelos **GEE**, sino incluyendo efectos aleatorios para cada sujeto dentro de la expresión del modelo. Las características generales de estos modelos son las siguientes:

1. No requiere que todos los sujetos sean medidos  $n$  veces ( $n_i \neq n$ ).
2. Se pueden incluir dentro del modelo variables independientes “fijas” al igual que variables que “cambien” en el tiempo.
3. Se puede incluir el tiempo como una variable independiente continua en el modelo, por lo tanto los sujetos no tienen que ser medidos en los “mismos” tiempos.
4. Los **MIX** permiten estimar cambios en el desenlace a nivel global y a nivel individual, incluyendo efectos aleatorios para cada sujeto y para cada variable independiente que varíe en el tiempo.

A continuación se presentan las expresiones definidas para estos modelos:

- MIX - Intercepto aleatorio.

Expresión del modelo:

$$y_{ij} = \beta_{0i} + \beta_1 t_{ij} + e_{ij}$$

donde

$$\beta_{0i} = \beta_0 + v_{0i}$$

$\beta_0$  es el intercepto promedio en la población y  $v_{0i} \sim N(0, \sigma_{v0}^2)$  son las desviaciones de cada sujeto con relación a la población.  $\beta_1$  es la pendiente de la población que es igual para todos los sujetos,  $t_{ij}$  es una variable que representa el tiempo  $j$  en el cual el desenlace fue medido para el sujeto  $i$ , y  $e_{ij}$  son los errores que ahora se consideran condicionalmente independientes con distribución  $N(0, \sigma_e^2)$ . Independencia condicional quiere decir, condicional al efecto aleatorio individual  $v_{0i}$ .

El modelo con intercepto aleatorio, asume “compound symmetry” por lo cual se puede obtener el índice de correlación intraclase a partir de la siguiente expresión:

$$ICC = \frac{\sigma_{v0}^2}{\sigma_{v0}^2 + \sigma_e^2}$$

Este coeficiente representa el nivel de asociación de los datos longitudinales en cada sujeto y también, la proporción de variabilidad no explicada por el modelo atribuible a los sujetos.

- MIX - Intercepto y pendiente aleatoria

Expresión del modelo:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$$

donde:

$$\beta_{0i} = \beta_0 + v_{0i}$$

y

$$\beta_{1i} = \beta_1 + v_{1i}$$

Es decir, se está incluyendo como término adicional  $v_{1i}$  que representa la desviación de la pendiente  $\beta_1$  para el sujeto  $i$  y ahora  $e_{ij}$  se consideran condicionalmente independientes a los efectos  $v_{0i}$  y  $v_{1i}$  con distribución  $N(0, \sigma_e^2)$ .

Ahora, con dos efectos aleatorio,  $v_{0i}$  y  $v_{1i}$ , se asume que:

$$v_{0i}, v_{1i} \sim NM(0, \Sigma_v)$$

donde:

$$\Sigma_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v0,v1} \\ \sigma_{v0,v1} & \sigma_{v1}^2 \end{bmatrix}$$

El término  $\sigma_{v0,v1}$  corresponde a la covarianza entre  $v_{0i}$  y  $v_{1i}$  que al expresarla como una correlación su interpretación es:

- + valores positivo indica que sujetos con intercepto alto toma pendientes altas, y
- + valor negativo indica que sujetos con interceptos altos toma valores bajo la pendiente.

Adicionalmente, se pueden estimar intervalos de confianza para los interceptos y pendientes de los sujetos:

$$\beta_0 \pm z_{(1-\frac{\alpha}{2})} \sigma_{v0}$$

y

$$\beta_1 \pm z_{(1-\frac{\alpha}{2})} \sigma_{v1}$$

En este modelo, se ha incluido la variable tiempo ( $t$ ) como una variables independiente continua.  $t$  puede esta codificada de diferentes formas; por ejemplo si tenemos  $n$  mediciones,  $t$  puede tomar el valor de:  $0, 1, 2, \dots, n - 1$ , donde la primera medición  $t = 0$  representa la línea base. Si se mantiene la misma magnitud de cambio entre los tiempos, en el modelo con intercepto aleatorio, pero el punto cero representa otro momento distinto al inicial, la estimación de los parámetros relacionados al intercepto cambian pero la pendiente es la misma. La selección de la codificación depende de la pregunta que se quiera contestar.

- MIX - Intercepto y pendiente aleatoria, y una variable fija con respuesta dicotómica.

Expresión del modelo:

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}$$

donde

$$\beta_{0i} = \beta_0 + \beta_2x_i + v_{0i}$$

y

$$\beta_{1i} = \beta_1 + \beta_3x_i + v_{1i}$$

y  $x_i$  es una variable indicador que toma valor de 0 y 1 para identificar cada grupo.

$\beta_0$  corresponde al intercepto de la población del grupo  $x_i = 0$  y  $\beta_2$  corresponde al cambio en el intercepto para el grupo  $x_i = 1$  con relación al grupo  $x_i = 0$ .  $\beta_1$  es la pendiente del grupo  $x_i = 0$  y  $\beta_3$  corresponde al cambio en la pendiente del grupo  $x_i = 1$  con relación al grupo  $x_i = 0$ .

- Expresión matricial del modelo



Expresión del modelo:

$$y_i = X_i\beta + Z_iv_i + \epsilon_i$$

donde:

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}_{(n_i \times 1)}$$

$$X_i = \begin{bmatrix} 1 & x_{i11} & x_{i12} & \dots & x_{i1p} \\ 1 & x_{i21} & x_{i22} & \dots & x_{i2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{in_i1} & x_{in_i2} & \dots & x_{in_ip} \end{bmatrix}_{(n_i \times (p+1))}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1}$$

$$Z_i = \begin{bmatrix} 1 & z_{i11} & z_{i12} & \dots & z_{i1r} \\ 1 & z_{i21} & z_{i22} & \dots & z_{i2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_{in_i1} & z_{in_i2} & \dots & z_{in_ir} \end{bmatrix}_{(n_i \times (r+1))}$$

$$v_i = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_r \end{bmatrix}_{((r+1) \times 1)}$$

y

$$\epsilon_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}_{(n_i \times 1)}$$

donde  $n_i$  es el número de repeticiones del sujeto  $i$ ,  $p$  es el número de covariables y  $r$  es el número de coeficientes aleatorios.

Los supuestos del modelo bajo esta expresión son:

$$\epsilon_i \sim NM(0, \sigma^2 I_{n_i})$$

y

$$v_i \sim NM(0, \Sigma_v)$$

Por lo tanto,

$$\begin{bmatrix} y_i \\ v_i \end{bmatrix} \sim NM\left(\begin{bmatrix} X_i\beta \\ 0 \end{bmatrix}, \begin{bmatrix} Z_i\Sigma_v Z_i' + \sigma^2 I_{n_i} & Z_i\Sigma_v \\ \Sigma_v Z_i' & \Sigma_v \end{bmatrix}\right)$$

- Inferencia.

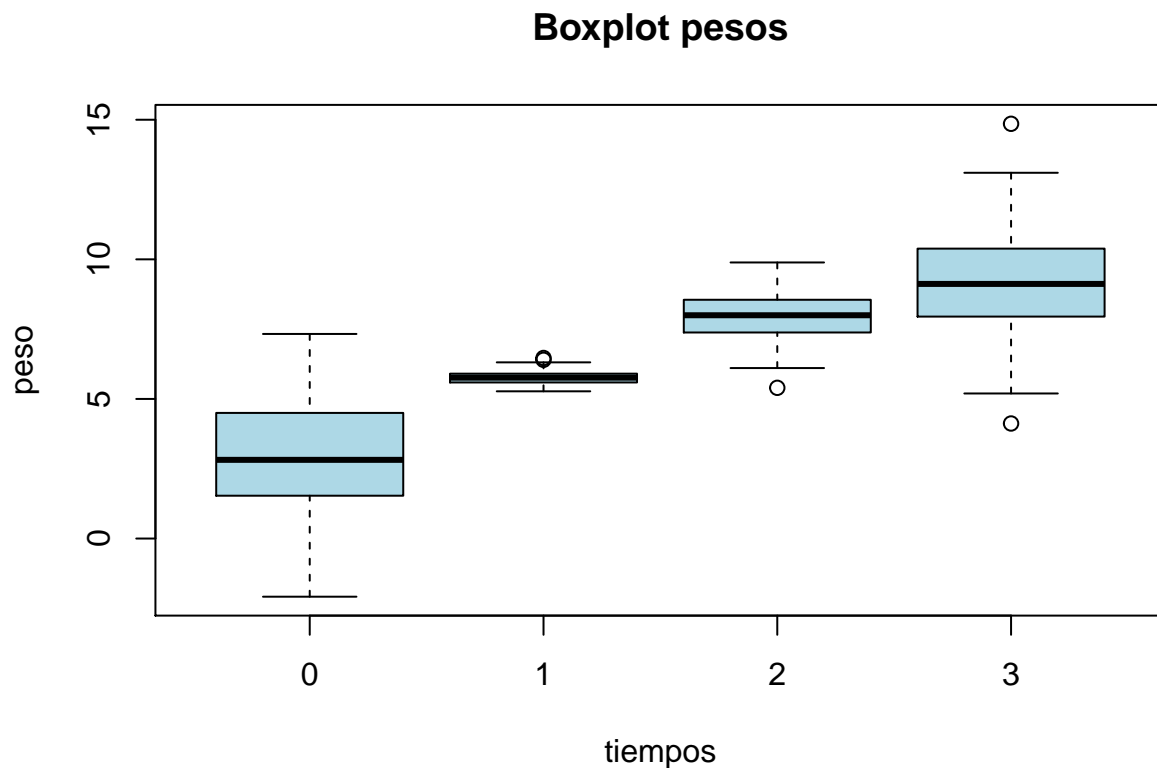
La evaluación de los parámetros fijos ( $H_0 : \beta = 0$ ) se realiza con el estadístico de Wald, que corresponde a  $\frac{\hat{\beta}_j}{SE_{\hat{\beta}_j}} \sim N(0, 1)$ .

La evaluación de los parámetros de varianzas y covarianzas para modelos anidados, se puede realizar con la diferencia entre las estadísticas de *deviance* ( $-2\log L$ ) de cada modelo; este resultado tiene una distribución  $\chi^2$  con grados de libertad igual a la diferencia entre el número de parámetros estimados de cada modelo.

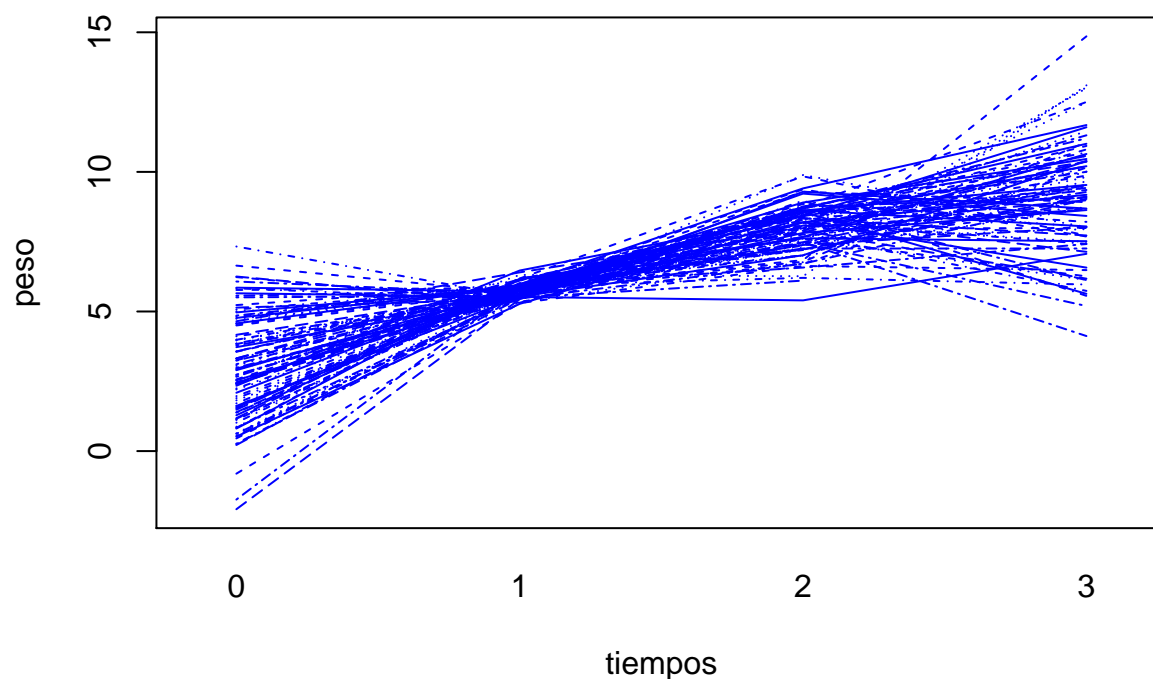
## Aplicación

Escenario: Se midió una cohorte de niños recién nacidos donde se registro el peso (desenlace), el sexo (H=0, M=1), si recibió lactancia materna (No = 0, Si = 1) y el tiempo en que se realizó la medición.

1. Descripción:



## Tendencia del peso por sujeto



2. Ajuste modelo: Intercepto aleatorio y tiempo como variable independiente.

```
library(nlme)
```

```
m1<-lme(peso~t1,data=datos,random=~1|id)
```

```
summary(m1)
```

```
## Linear mixed-effects model fit by REML
## Data: datos
##      AIC      BIC    logLik
## 1396.167 1411.917 -694.0837
##
## Random effects:
## Formula: ~1 | id
##      (Intercept) Residual
## StdDev: 0.004338571 1.486517
##
## Fixed effects: peso ~ t1
##              Value Std.Error DF t-value p-value
## (Intercept) 3.316143 0.12587545 282 26.34463    0
## t1          2.082661 0.06806577 282 30.59777    0
## Correlation:
## (Intr)
## t1 -0.796
##
## Standardized Within-Group Residuals:
##      Min      Q1      Med      Q3      Max
```

```
## -3.6615448 -0.4352411 0.1639113 0.4934864 3.5590871
##
## Number of Observations: 381
## Number of Groups: 98
```

```
# Calculo índice de correlación intraclase:
```

```
ICC<-0.003268628^2/(0.003268628^2+1.48652^2);ICC
```

```
## [1] 4.834899e-06
```

- Explorar efectos aleatorios del intercepto.

```
library(tidyverse)
```

```
coef.ale<-data.frame(m1$coefficients$random$id)
```

```
shapiro.test(coef.ale$X.Intercept.)
```

```
##
```

```
## Shapiro-Wilk normality test
```

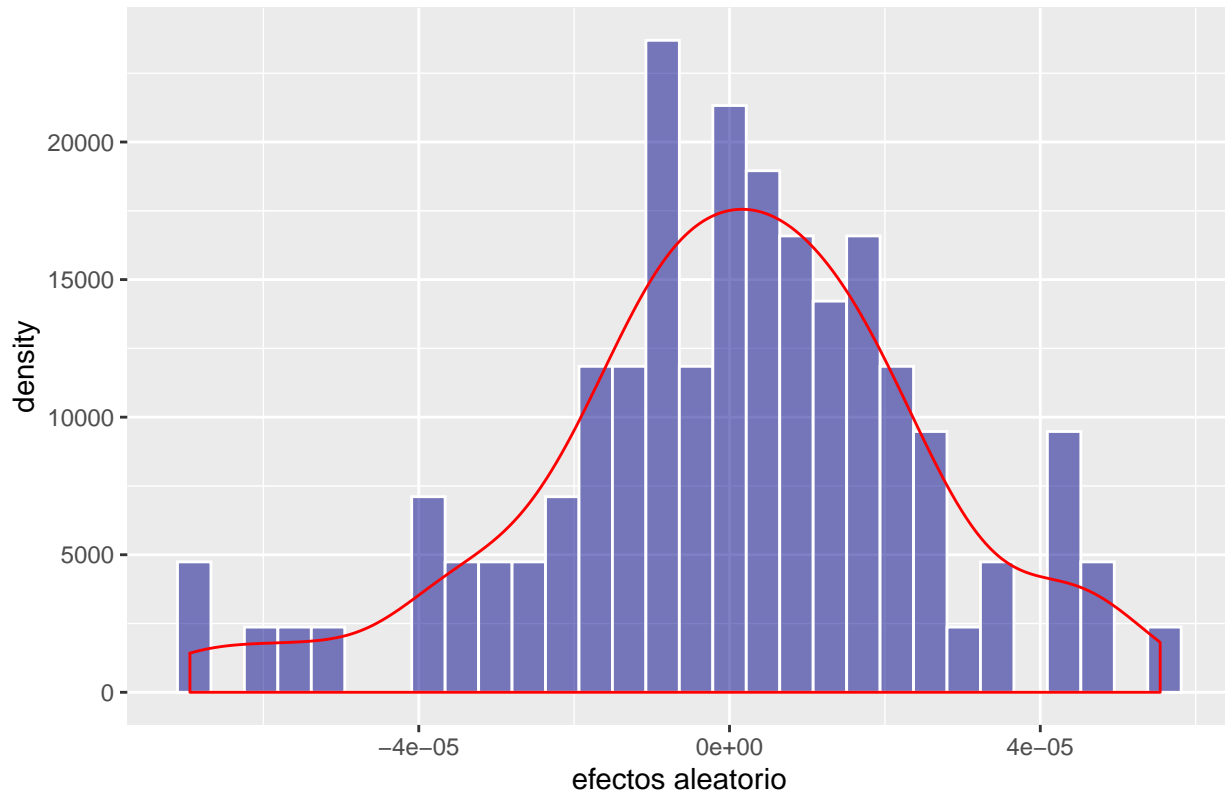
```
##
```

```
## data: coef.ale$X.Intercept.
```

```
## W = 0.98063, p-value = 0.1579
```

```
ggplot(coef.ale,aes(X.Intercept.))+
  geom_histogram(bins=30,aes(y = ..density..), alpha=0.5,fill="darkblue",
                color = "white")+
  geom_density(col="red",size=0.5)+
  labs(title="Histograma de los efectos aleatorios del intercepto",
        x="efectos aleatorio")
```

## Histograma de los efectos aleatorios del intercepto



- valores ajustados versus observados.

```
m1fit<-data.frame(m1$data$id,m1$fitted[,2],m1$residuals[,2],m1$data$t1,m1$data$peso)
names(m1fit)
```

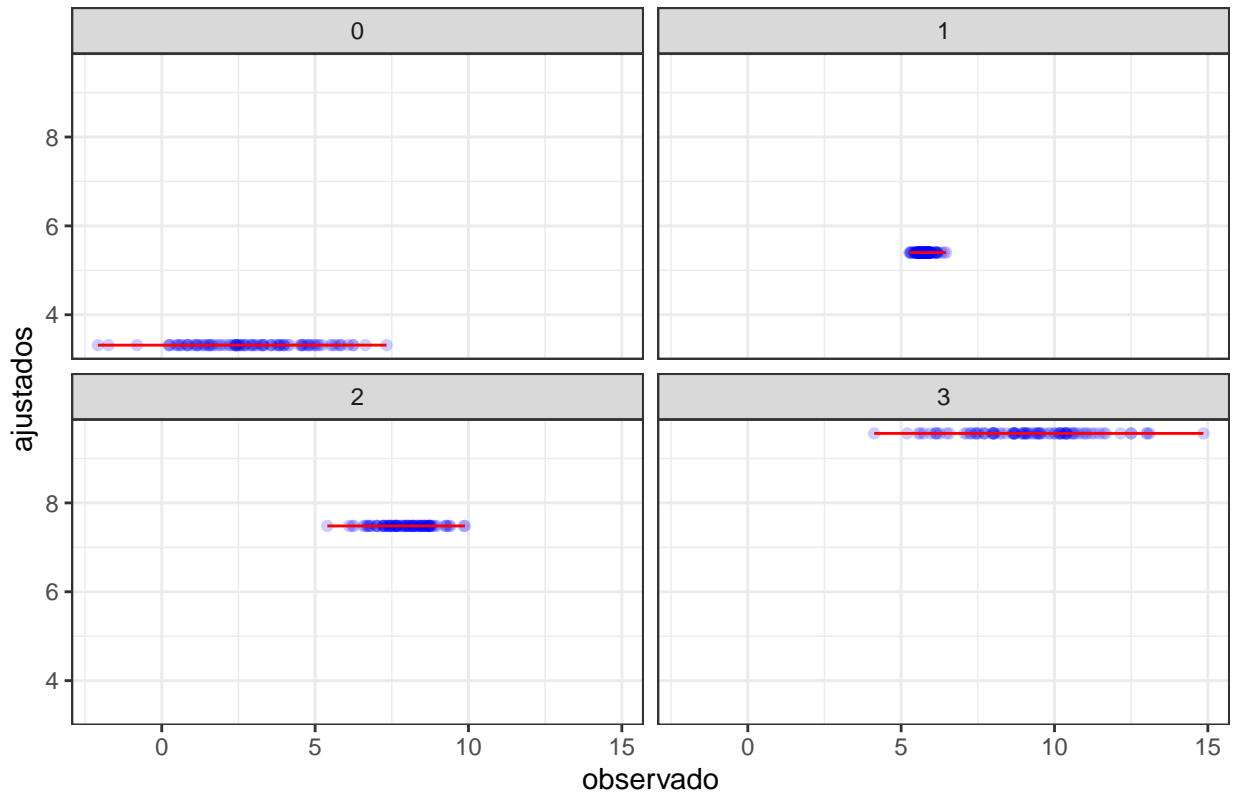
```
## [1] "m1.data.id"      "m1.fitted...2."  "m1.residuals...2."
## [4] "m1.data.t1"     "m1.data.peso"
```

```
m1fit<-rename(m1fit,id=m1.data.id,fit=m1.fitted...2.,
              res=m1.residuals...2.,t=m1.data.t1,obs=m1.data.peso)
names(m1fit)
```

```
## [1] "id" "fit" "res" "t" "obs"
```

```
ggplot(m1fit,aes(obs,fit))+
  geom_point(alpha=0.2,color="blue")+
  geom_smooth(method = lm,color="red",size=0.5,se=FALSE)+
  theme_bw()+
  labs(title = "Observados vs ajustados por tiempo",
       x="observado",y="ajustados")+
  facet_wrap(~t)
```

## Observados vs ajustados por tiempo



- Exploración de los residuales

```
m1res.w<-reshape(m1fit,timevar = "t",idvar=c("id"),direction = "wide")
m1res.w<-select(m1res.w,res.0,res.1,res.2,res.3)
```

```
m1res.w.complete<-m1res.w%>%filter(complete.cases(.))
res.matrix<-matrix(c(m1res.w.complete$res.0,m1res.w.complete$res.1,
                    m1res.w.complete$res.2,m1res.w.complete$res.3)
                  ,nrow = 87,ncol = 4)
```

```
library(mvShapiroTest)
mvShapiro.Test(res.matrix)
```

```
##
## Generalized Shapiro-Wilk test for Multivariate Normality by
## Villasenor-Alva and Gonzalez-Estrada
##
## data: res.matrix
## MVW = 0.98672, p-value = 0.6623
```

3. Ajuste modelo: Intercepto y pendiente aleatoria para el tiempo t.

```
m2<-lme(peso~t1,data = datos,random=~t1|id)
summary(m2)
```

```
## Linear mixed-effects model fit by REML
## Data: datos
##      AIC      BIC    logLik
```

```

## 1388.392 1412.017 -688.1958
##
## Random effects:
## Formula: ~t1 | id
## Structure: General positive-definite, Log-Cholesky parametrization
##          StdDev   Corr
## (Intercept) 0.8154545 (Intr)
## t1          0.5358887 -0.914
## Residual    1.3198515
##
## Fixed effects: peso ~ t1
##          Value Std.Error DF t-value p-value
## (Intercept) 3.314326 0.13889257 282 23.86252 0
## t1          2.085589 0.08133503 282 25.64195 0
## Correlation:
## (Intr)
## t1 -0.838
##
## Standardized Within-Group Residuals:
##          Min      Q1      Med      Q3      Max
## -3.3483145 -0.4982481 0.1877081 0.5139058 3.0382057
##
## Number of Observations: 381
## Number of Groups: 98

```

*# (Ahora utilizando una codificación diferente del tiempo!!)*

```

m3<-lme(peso~t2,data = datos,random=~t2|id)
summary(m3)

```

```

## Linear mixed-effects model fit by REML
## Data: datos
##          AIC      BIC    logLik
## 1388.392 1412.017 -688.1958
##
## Random effects:
## Formula: ~t2 | id
## Structure: General positive-definite, Log-Cholesky parametrization
##          StdDev   Corr
## (Intercept) 0.3920316 (Intr)
## t2          0.5358881 -0.533
## Residual    1.3198516
##
## Fixed effects: peso ~ t2
##          Value Std.Error DF t-value p-value
## (Intercept) 5.399915 0.08344408 282 64.71298 0
## t2          2.085589 0.08133499 282 25.64196 0
## Correlation:
## (Intr)
## t2 -0.421
##
## Standardized Within-Group Residuals:
##          Min      Q1      Med      Q3      Max
## -3.3483128 -0.4982486 0.1877079 0.5139060 3.0382077
##

```

```
## Number of Observations: 381
## Number of Groups: 98
```

- explorar los efectos aleatorios

```
coef.ale<-data.frame(m2$coefficients$random$id)

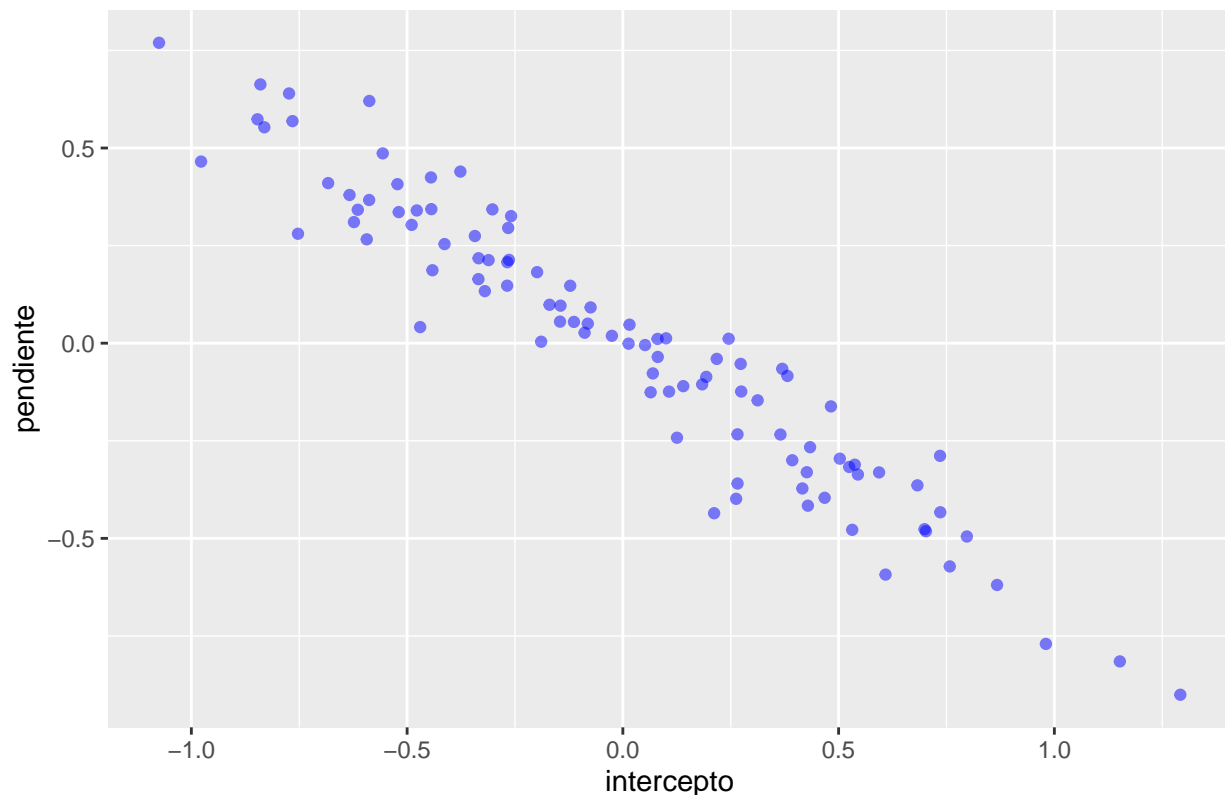
ale.matrix<-matrix(c(coef.ale$X.Intercept.,coef.ale$t1),
                   nrow = 98,ncol = 2)

mvShapiro.Test(ale.matrix)
```

```
##
## Generalized Shapiro-Wilk test for Multivariate Normality by
## Villasenor-Alva and Gonzalez-Estrada
##
## data: ale.matrix
## MW = 0.99076, p-value = 0.8476
```

```
ggplot(coef.ale,aes(x=X.Intercept.,y=t1))+
  geom_point(alpha=0.5,color="blue")+
  labs(title="Gráfico de dispersión efectos aleatorio intercepto vs. pendiente",
       x="intercepto",y="pendiente")
```

Gráfico de dispersión efectos aleatorio intercepto vs. pendiente



- valores ajustados vs. observados.

```
m2fit<-data.frame(m2$data$id,m2$fitted[,2],m2$residuals[,2],m2$data$t1,m2$data$peso)
names(m2fit)
```

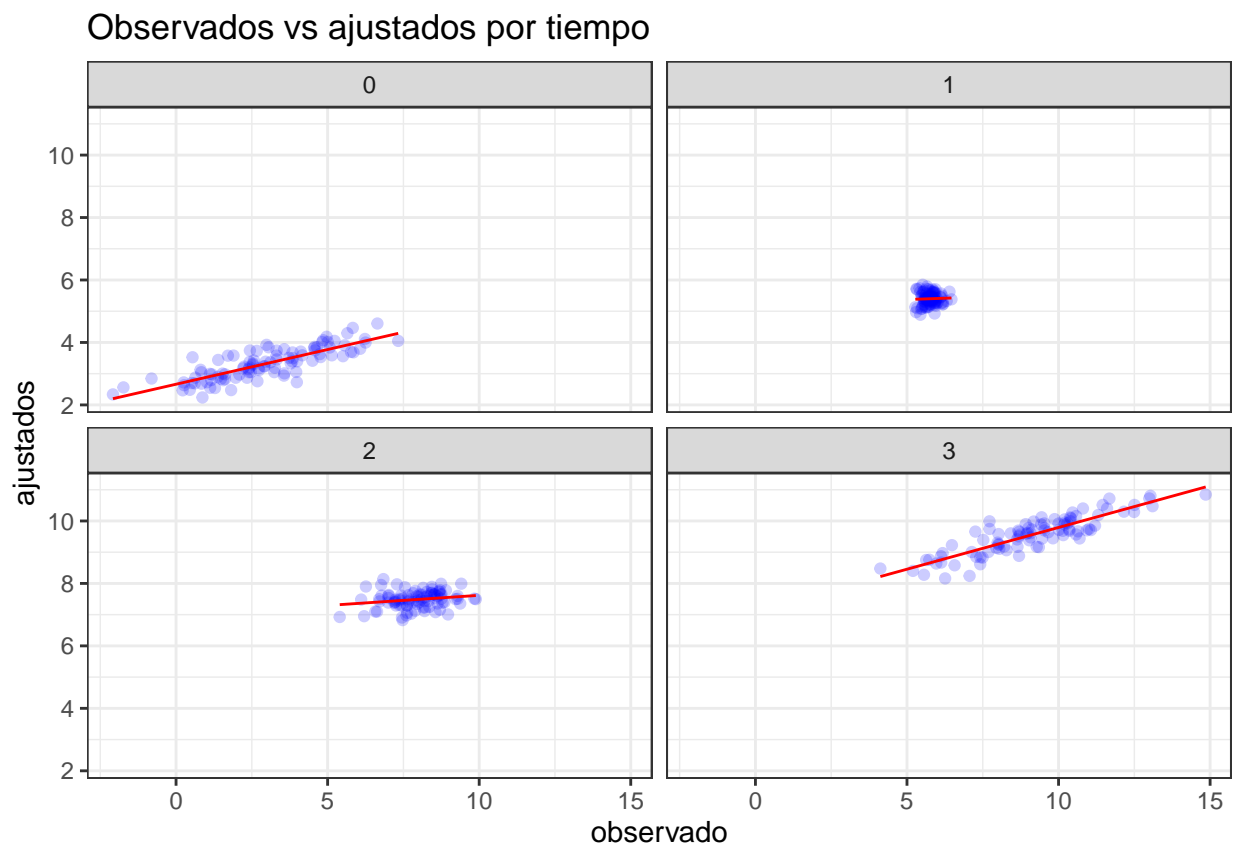


```
## [1] "m2.data.id"          "m2.fitted...2."    "m2.residuals...2."
## [4] "m2.data.t1"          "m2.data.peso"

m2fit<-rename(m2fit,id=m2.data.id,fit=m2.fitted...2.,res=m2.residuals...2.,
              t=m2.data.t1,obs=m2.data.peso)
names(m2fit)
```

```
## [1] "id" "fit" "res" "t" "obs"
```

```
ggplot(m2fit,aes(obs,fit))+
  geom_point(alpha=0.2,color="blue")+
  geom_smooth(method = lm,color="red",size=0.5,se=FALSE)+
  theme_bw()+
  labs(title = "Observados vs ajustados por tiempo",
       x="observado",y="ajustados")+
  facet_wrap(~t)
```



- exploración de los residuales

```
m2res.w<-reshape(m2fit,timevar = "t",idvar=c("id"),direction = "wide")
m2res.w<-select(m2res.w,res.0,res.1,res.2,res.3)
```

```
m2res.w.complete<-m2res.w%>%filter(complete.cases(.))
res.matrix<-matrix(c(m2res.w.complete$res.0,m2res.w.complete$res.1,
                    m2res.w.complete$res.2,m2res.w.complete$res.3),
                  nrow = 87,ncol = 4)
```

```
library(mvShapiroTest)
```

```
mvShapiro.Test(res.matrix)
```

```
##  
## Generalized Shapiro-Wilk test for Multivariate Normality by  
## Villasenor-Alva and Gonzalez-Estrada  
##  
## data: res.matrix  
## MVW = 0.98623, p-value = 0.6048
```

4. Ajuste modelo: Intercepto y pendiente aleatoria para el tiempo t con una variable fija (sexo = dos grupos)

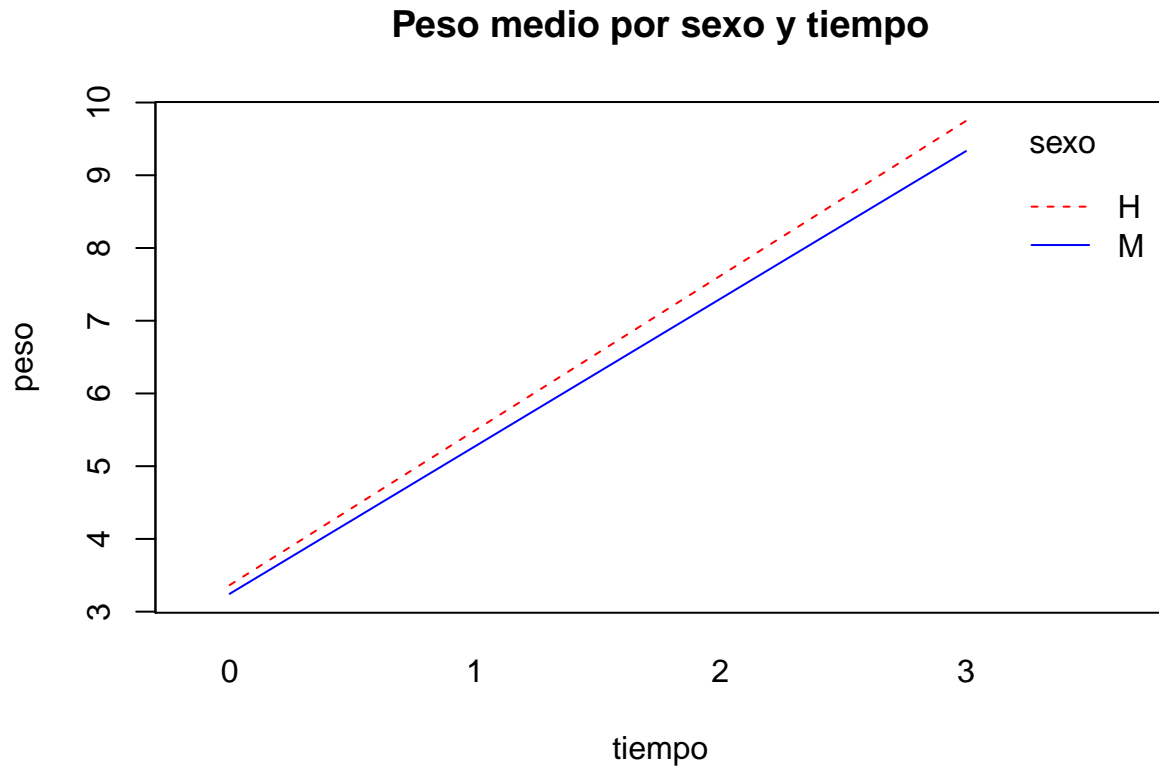
```
m4<-lme(peso~t1+sexo+sexo*t1,data=datos,random=~t1|id)  
summary(m4)
```

```
## Linear mixed-effects model fit by REML  
## Data: datos  
##      AIC      BIC    logLik  
## 1392.798 1424.256 -688.399  
##  
## Random effects:  
## Formula: ~t1 | id  
## Structure: General positive-definite, Log-Cholesky parametrization  
##           StdDev   Corr  
## (Intercept) 0.8261773 (Intr)  
## t1           0.5404779 -0.926  
## Residual    1.3202855  
##  
## Fixed effects: peso ~ t1 + sexo + sexo * t1  
##           Value Std.Error DF   t-value p-value  
## (Intercept) 3.364851 0.1846024 281 18.227562 0.0000  
## t1          2.127335 0.1079860 281 19.700106 0.0000  
## sexo        -0.119795 0.2820368 96 -0.424750 0.6720  
## t1:sexo      -0.098353 0.1650114 281 -0.596039 0.5516  
## Correlation:  
##           (Intr) t1      sexo  
## t1        -0.846  
## sexo      -0.655 0.553  
## t1:sexo   0.553 -0.654 -0.844  
##  
## Standardized Within-Group Residuals:  
##           Min      Q1      Med      Q3      Max  
## -3.2832373 -0.5074115 0.1880119 0.5201811 2.9781861  
##  
## Number of Observations: 381  
## Number of Groups: 98
```

```
# gráfico peso medio de los dos grupos - sexo
```

```
t<-c(0,1,2,3)  
peso_x0<-m4$coefficients$fixed[1]+m4$coefficients$fixed[2]*t  
peso_x1<-m4$coefficients$fixed[1]+m4$coefficients$fixed[3]+  
  (m4$coefficients$fixed[2]+m4$coefficients$fixed[4])*t  
  
sexo<-c("H", "H", "H", "H", "M", "M", "M", "M")
```

```
g1<-data.frame(c(t,t),c(peso_x0,peso_x1),sexo)
with(g1,interaction.plot(c.t..t.,sexo,c.peso_x0..peso_x1.,col = c("red","blue"),
  main="Peso medio por sexo y tiempo",xlab="tiempo",
  ylab="peso"))
```



#### 5. Comparación de modelos anidados (likelihood ratio test)

```
anova(m1,m2)
```

```
##      Model df      AIC      BIC   logLik  Test L.Ratio p-value
## m1      1  4 1396.167 1411.918 -694.0837
## m2      2  6 1388.392 1412.017 -688.1958 1 vs 2 11.7757  0.0028
```

## Modelos mixtos (MIX) - Desenlace dicotómico

La regresión logística es uno de los métodos más utilizados para evaluar la relación de un desenlace con respuesta dicotómica

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

y un conjunto de variables independientes. Este método tiene como supuesto que las observaciones del desenlace son independientes entre sí, lo cual no se cumple cuando el desenlace (ahora  $Y_{ij}$ ) se está evaluado en  $n_i$  momentos diferentes por cada  $N$  sujetos. La regresión logística con efectos mixtos es una alternativa de análisis en esta situación.

Inicialmente, podemos evaluar la relación entre el tiempo, medido como una variable cuantitativa, y el desenlace  $Y_{ij}$ . Incorporando un intercepto aleatorio; la expresión del modelo es el siguiente:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_1 t + v_{i0}$$

donde  $p_{ij}$  es la probabilidad de  $Y_{ij} = 1$ ,  $\beta_0$  y  $\beta_1$  son el intercepto y la pendiente fijas y  $v_{i0}$  es el efecto aleatorio del sujeto  $i$  donde  $v_{i0} \sim N(0; \sigma_{v_{i0}}^2)$ . La interpretación de  $\beta_0$  y  $\beta_1$  se realizan con relación al logaritmo del odds de presentar el desenlace ( $Y_{ij} = 1$ ), entonces el valor exponencial de  $\beta_1$  corresponde a la razón de disparidades (odds ratio - OR) de presentar el desenlace cuando aumenta en una unidad la variable tiempo. La interpretación de los parámetros fijos cuando la función de enlace es no lineal, como es este caso, representan el efecto de la covariable en el desenlace cuando se mantienen constantes los efectos aleatorios del sujeto, es decir que la estimación de los parámetros fijo es condicional al efecto aleatorio del sujeto.

Bajo el modelo anterior, estimar la proporción de la variabilidad no explicada por el modelo que es explicada por los sujetos se obtiene a partir del coeficiente de correlación intraclase ( $ICC$ ). Bajo el supuesto de que la varianza del error ( $\sigma_e^2$ ) sigue una distribución logística estándar con  $\sigma_e^2 = \pi^2/3$ , el  $ICC$  es igual a:

$$ICC = \frac{\sigma_{v_{i0}}^2}{\sigma_{v_{i0}}^2 + (\pi^2/3)}$$

Un segundo modelo se puede obtener incorporando la pendiente aleatoria para el tiempo. La expresión es la siguiente:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + v_{i0} + \beta_1 t + v_{i1} t$$

donde ahora se agrega el término  $v_{i1}$  que corresponde al cambio en la pendiente de cada sujeto ( $v_{i1} \sim N(0, \sigma_{v_{i1}}^2)$ ). En este modelo también se puede estimar la correlación entre los efectos aleatorios ( $v_{i0}$  y  $v_{i1}$ ).

Ahora, incluiremos una variable dicotómica fija,  $X_i = \begin{cases} 1 \\ 0 \end{cases}$ , dentro del modelo al igual que su interacción con el tiempo. La expresión es la siguiente:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_1 t + \beta_2 X + \beta_3 X t + v_{i0} + v_{i1} t$$

donde  $\beta_0$  y  $\beta_1$  corresponde al intercepto y pendiente fijas del grupo  $X = 0$ , y  $\beta_2$  y  $\beta_3$  corresponden a cambio tanto del intercepto como de la pendiente para el grupo de  $X = 1$  con relación al grupo  $X = 0$ . Recordar que la interpretación se realiza en términos del logaritmo del chance de presentar el desenlace.

Finalmente, un modelo general puede ser el siguiente: intercepto y pendiente aleatoria, con una variable dicotómica fija ( $X_{1i}$ ) y una variable cuantitativa dependiente del tiempo  $X_{2ij}$  (incluida como un efecto aleatorio). La expresión del modelo es la siguiente:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + v_{i0} + \beta_1 t + v_{i1} t + \beta_2 X_{1i} + \beta_3 X_{2ij} + v_{2i} X_{2ij}$$

- Todos los modelos anteriores, se expresan para la función  $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)$ . Si el interés es conocer la  $p_{ij}$  para un sujeto específico, se utiliza la expresión:

$$p_{ij} = \frac{1}{1 + e^{-(\text{logit}(p_{ij}))}}$$

## Aplicación.

Escenario: 98 niños recién nacidos donde se midió su peso y se categorizó en desnutridos (1) y normales (0). Se realizaron mediciones cada tres meses para un total de 4. Se midieron otras variables: sexo (H=0, M=1), si recibió lactancia materna (No = 0, Si = 1) y horas de sueño (cuantitativa - variable que cambia en el tiempo).

Cargue de la base de datos:

```
library(rio)
data<-import("RN peso.xlsx")

tabla<-round(prop.table(table(data$t1,data$peso_dic),1),2);tabla

##
##      0      1
## 0 0.57 0.43
## 1 0.74 0.26
## 2 0.68 0.32
## 3 0.80 0.20
```

1. Modelo con intercepto aleatorio y tiempo como variable independiente.

```
library(lme4)

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following object is masked from 'package:tidyr':
##
##      expand
##
## Registered S3 methods overwritten by 'lme4':
##   method                               from
##   cooks.distance.influence.merMod      car
##   influence.merMod                     car
##   predict.merMod                       MuMIn
##   dfbeta.influence.merMod              car
##   dfbetas.influence.merMod             car
##
## Attaching package: 'lme4'
```

```

## The following object is masked from 'package:nlme':
##
##   lmList

## The following object is masked from 'package:RVAideMemoire':
##
##   dummy

## The following object is masked from 'package:rio':
##
##   factorize

mix1<-glmer(peso_dic~t1+(1|id),data=data,family = binomial)
summary(mix1, corr=FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: peso_dic ~ t1 + (1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##  462.8    474.7   -228.4   456.8     378
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -0.9506 -0.6737 -0.5241  1.1560  2.0225
##
## Random effects:
##  Groups Name      Variance Std.Dev.
##  id      (Intercept) 0.2261   0.4755
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.4487     0.1893  -2.370  0.01780 *
## t1           -0.3094     0.1056  -2.931  0.00338 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Calculo de OR con su intervalo de confianza.

se1 <- sqrt(diag(vcov(mix1)))
tab1 <- cbind(Est=fixef(mix1),LL=fixef(mix1)-1.96*se1,UL=fixef(mix1)+1.96*se1)
exp(tab1)[2,]

##      Est      LL      UL
## 0.7339026 0.5967302 0.9026072

# Calculo del coeficiente de correlación intraclase.

ICC<-0.2261/(0.2261+pi^2/3)

#coef(mix1) coeficientes por sujeto organizados por id.hist
#ranef(mix1)

#fitted(mix1)(con el orden de la base de datos original)

```

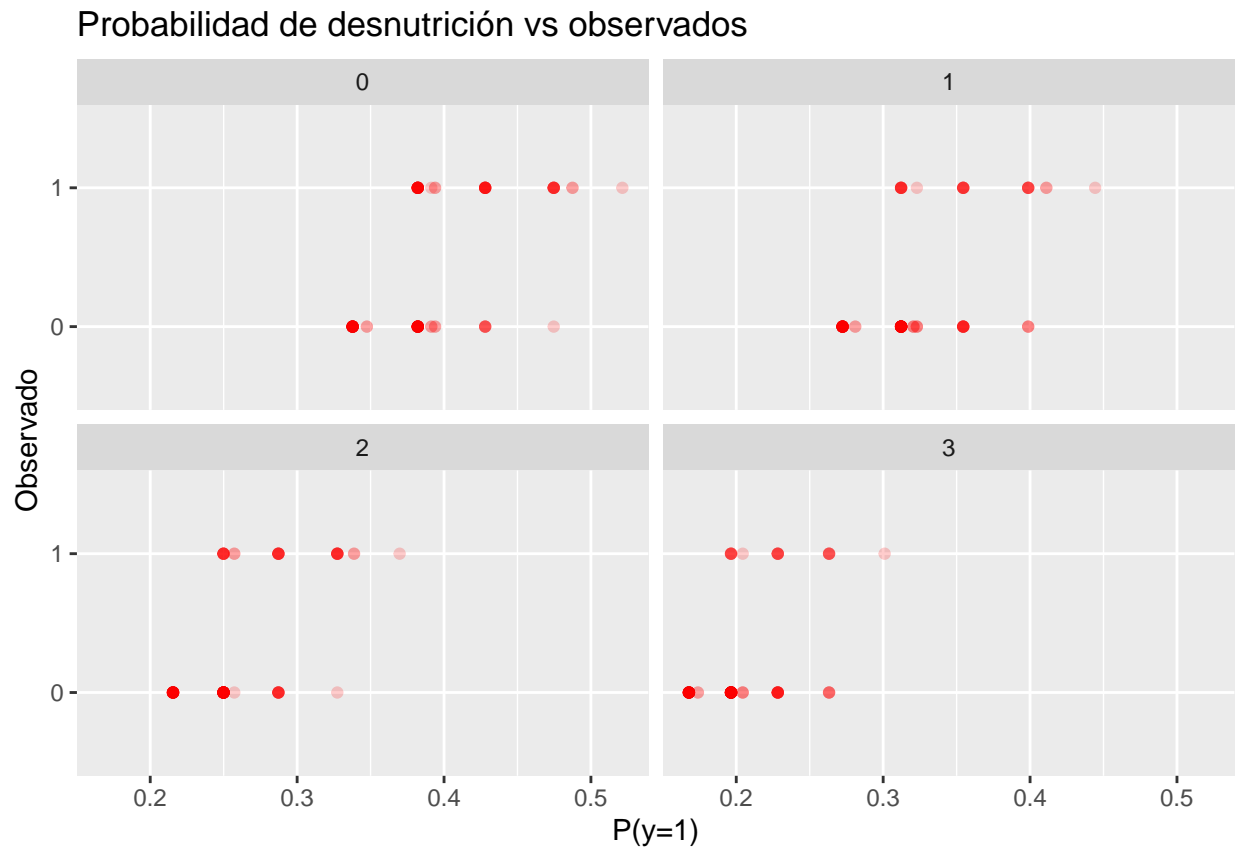
- Comparación entre observados y probabilidades estimadas:

```
library(tidyverse)

fit1<-data.frame(fitted(mix1),data$t1,data$peso_dic)
names(fit1)

## [1] "fitted.mix1." "data.t1"      "data.peso_dic"

ggplot(fit1,aes(fitted.mix1.,factor(data.peso_dic)))+
  geom_point(alpha=0.2,color="red")+
  labs(title="Probabilidad de desnutrición vs observados",y="Observado", x="P(y=1)")+
  facet_wrap(~data.t1)
```



## 2. Modelo con intercepto y pendiente aleatoria

```
mix2<-glmer(peso_dic~t1+(t1|id),data = data,family = binomial)
summary(mix2, corr=FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: peso_dic ~ t1 + (t1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##  466.4    486.1  -228.2   456.4     376
##
## Scaled residuals:
```

```

##      Min      1Q  Median      3Q      Max
## -0.8476 -0.6577 -0.4941  1.1432  1.7736
##
## Random effects:
##   Groups Name      Variance Std.Dev. Corr
##   id      (Intercept) 0.2685  0.5182
##   t1      t1          0.1106  0.3326  -0.44
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4290    0.2003  -2.142  0.0322 *
## t1           -0.3564    0.1386  -2.571  0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

se2<-sqrt(diag(vcov(mix2)))
tab2<-cbind(Est = fixef(mix2), LL = fixef(mix2) - 1.96 * se2, UL = fixef(mix2) + 1.96*se2)
exp(tab2)

##              Est      LL      UL
## (Intercept) 0.6511365 0.4397582 0.9641178
## t1          0.7002008 0.5336137 0.9187942

#coef(mix2)
#ranef(mix2)
#fitted(mix2)

```

- Observados vs. probabilidades estimadas.

```

fit2<-data.frame(fitted(mix2),data$t1,data$peso_dic)
names(fit2)

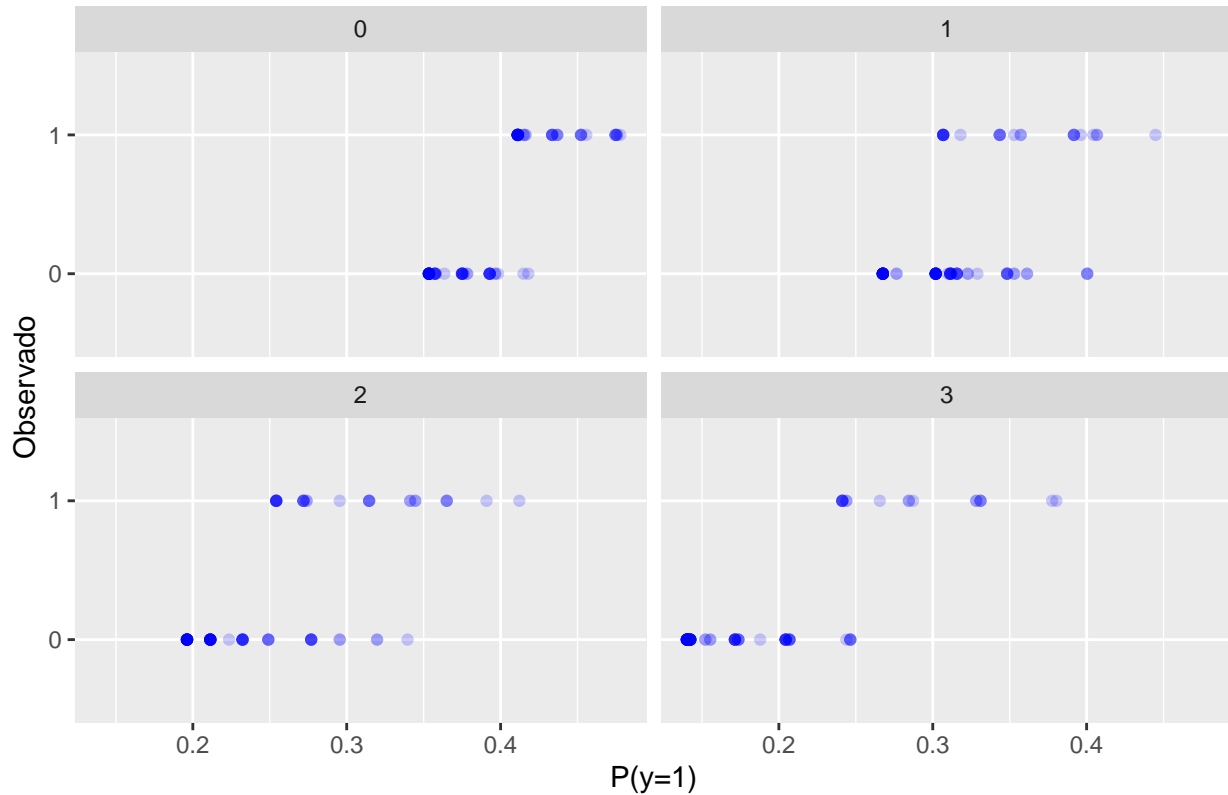
## [1] "fitted.mix2." "data.t1"      "data.peso_dic"

ggplot(fit2,aes(fitted.mix2.,factor(data.peso_dic)))+
  geom_point(alpha=0.2,color="blue")+
  labs(title="Probabilidad de desnutrición vs observados",y="Observado", x="P(y=1)")+
  facet_wrap(~data.t1)

```



## Probabilidad de desnutrición vs observados



3. Modelo con intercepto y pendiente aleatoria y una variable fija dicotómica.

```
mix3<-glmer(peso_dic~t1+sexo+t1*sexo+(t1|id),data = data,family = binomial)
summary(mix3, corr=FALSE)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: peso_dic ~ t1 + sexo + t1 * sexo + (t1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##  470.2   497.8  -228.1   456.2     374
##
## Scaled residuals:
##   Min       1Q   Median       3Q      Max
## -0.8501 -0.6657 -0.4945  1.1433  1.8311
##
## Random effects:
##   Groups Name      Variance Std.Dev. Corr
##   id      (Intercept) 0.2691  0.5188
##          t1          0.1099  0.3316 -0.45
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.42180    0.25904  -1.628  0.1035
```

```
## t1          -0.38352    0.17077   -2.246    0.0247 *
## sexo        -0.01676    0.38332   -0.044    0.9651
## t1:sexo     0.06534    0.22683    0.288    0.7733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#coef(mix3)
#ranef(mix3)
```

4. Modelo con intercepto y pendiente aleatoria, y cuatro variables independiente como efectos fijos.

```
mix4<-glmer(peso_dic~t1+sexo+lactancia+sueno+(t1|id),data = data,family = binomial)
summary(mix4,corr = FALSE)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: peso_dic ~ t1 + sexo + lactancia + sueno + (t1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##  453.0   484.5  -218.5   437.0     373
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.2326 -0.6205 -0.4545  0.9359  3.1208
##
## Random effects:
## Groups Name      Variance Std.Dev. Corr
## id      (Intercept) 0.56612  0.7524
## t1      t1          0.02896  0.1702  -1.00
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.60978    0.58148   1.049  0.29432
## t1           -0.36030    0.11609  -3.104  0.00191 **
## sexo         0.07454    0.26514   0.281  0.77860
## lactancia   -1.10108    0.26682  -4.127 3.68e-05 ***
## sueno       -0.06521    0.07018  -0.929  0.35276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
se4<-sqrt(diag(vcov(mix4)))
tab4<-cbind(Est = fixef(mix4), LL = fixef(mix4) - 1.96 * se4, UL = fixef(mix4) + 1.96*se4)
exp(tab4)
```

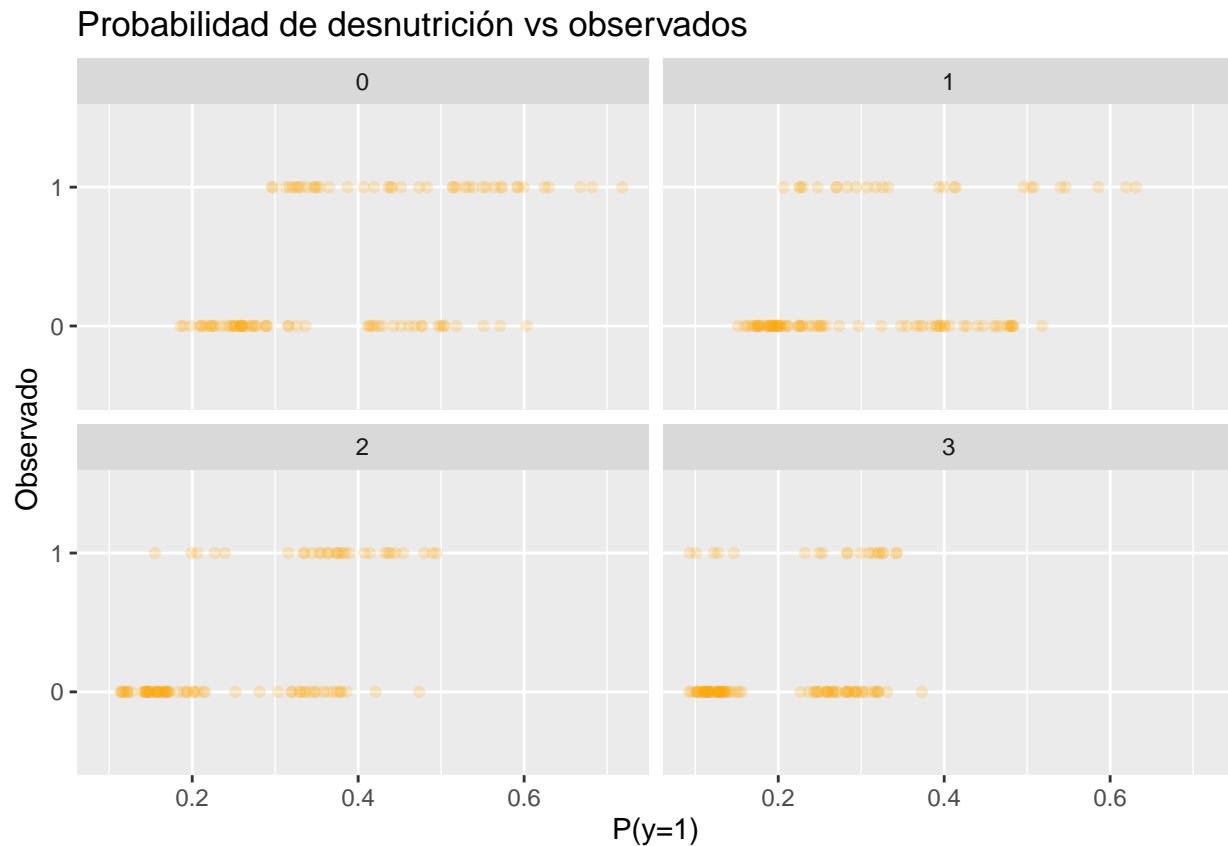
```
##              Est      LL      UL
## (Intercept) 1.8400314 0.5886587 5.7515772
## t1          0.6974670 0.5555281 0.8756716
## sexo        1.0773922 0.6407410 1.8116117
## lactancia   0.3325115 0.1971001 0.5609529
## sueno       0.9368667 0.8164694 1.0750180
```

- Observados vs. probabilidades estimadas.

```
fit4<-data.frame(fitted(mix4),data$t1,data$peso_dic)
names(fit4)
```

```
## [1] "fitted.mix4." "data.t1" "data.peso_dic"
```

```
ggplot(fit4,aes(fitted.mix4.,factor(data.peso_dic)))+
  geom_point(alpha=0.2,color="orange")+
  labs(title="Probabilidad de desnutrición vs observados",y="Observado", x="P(y=1)")+
  facet_wrap(~data.t1)
```



#### 5. Comparación de modelos:

```
anova(mix1,mix4)
```

```
## Data: data
## Models:
## mix1: peso_dic ~ t1 + (1 | id)
## mix4: peso_dic ~ t1 + sexo + lactancia + sueno + (t1 | id)
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mix1  3 462.84 474.67 -228.42   456.84
## mix4  8 452.98 484.52 -218.49   436.98 19.868     5 0.001323 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Modelos mixtos (MIX) - Desenlace de conteo.

Cuando un desenlace  $y_i$  es el número de veces que se presenta un evento en un periodo de tiempo, se puede asumir que esta variable tiene una distribución de Poisson, cuyo parámetro ( $\lambda$ ) corresponde a la media y varianza de la variable desenlace.

La exploración de la relación de este desenlace y un conjunto de variables independientes puede realizarse utilizando un modelo de regresión de Poisson. En el escenario de datos longitudinales, donde el desenlace es medido en distintos momentos sobre el mismo sujeto, el modelo de Poisson no se puede utilizar ya que no se cumple el supuesto de independencia; la extensión del modelo de Poisson al incluir efectos aleatorios se presenta como una alternativa de análisis.

Entonces, si  $y_{ij}$  es el número de veces (conteo) que se presenta un desenlace en el sujeto  $i = 1, 2, \dots, N$  en el tiempo  $j = 1, 2, \dots, n_i$ , la expresión del modelo considerando el tiempo como una variable cuantitativa independiente e introduciendo inicialmente un intercepto aleatorio, es el siguiente:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 t_{ij} + v_{i0}$$

donde  $\beta_0$  y  $\beta_1$  corresponde a los parámetros fijos, y  $v_{i0}$  es el efecto aleatorio donde  $v_{i0} \sim N(0, \sigma_{v_{i0}}^2)$ .  $\beta_1$  corresponde al cambio en el logaritmo de la tasa de incidencia ( $\lambda_i$ ) cuando cambia en tiempos consecutivos. En consecuencia, el valor exponencial de  $\beta_1$  corresponde a la razón de tasas de incidencia cuando cambia en una unidad el tiempo.

Otros modelos que se pueden ajustar son los siguientes:

- Intercepto y pendiente aleatoria para el tiempo.

$$\log(\lambda_{ij}) = \beta_0 + v_{i0} + \beta_1 t_{ij} + v_{i1} t_{ij}$$

donde  $\beta_0$  y  $\beta_1$  corresponde a los parámetros fijos, y  $v_{i0}$  es el efecto aleatorio donde  $v_{i0} \sim N(0, \sigma_{v_{i0}}^2)$ .  $\beta_1$  corresponde al cambio en el logaritmo de la tasa de incidencia ( $\lambda_i$ ) cuando cambia en tiempos consecutivos. El valor exponencial de  $\beta_1$  corresponde a la razón de tasas de incidencia cuando cambia en una unidad el tiempo.

- Intercepto y pendiente aleatoria para el tiempo, una variable dicotómica como efecto fijo ( $x_i$ ) y su interacción con el tiempo.

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 x_i t_{ij} + v_{i0} + v_{i1} t_{ij}$$

- Intercepto y pendiente aleatoria, variable dicotómica fija ( $x_{1i}$ ) y una variable cuantitativa dependiente del tiempo  $x_{2ij}$  (incluida como un efecto aleatorio)

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{1i} + \beta_3 x_{2ij} + v_{i0} + v_{i1} t_{ij} + v_{3i} x_{2ij}$$

Con base en las expresiones anteriores, se puede obtener la probabilidad de que se presente un número de eventos ( $y_{ij}$ ) para el sujeto  $i$  en el tiempo  $j$ . Por ejemplo si el modelo ajustado incorpora el intercepto y la pendiente aleatoria para el tiempo, la expresión es la siguiente:

$$f(y_{ij}|v_i) = \frac{e^{(-\lambda_{ij})} \lambda_{ij}^{y_{ij}}}{y_{ij}!}$$

donde  $\lambda_{ij} = e^{(\beta_0 + \beta_1 t_{ij} + v_{i0} + v_{i1} t_{ij})}$ .

## Modelo de Poisson con afectos aleatorios y cero inflado (Mix-ZIP)

Se puede presentar que algunos de los resultados de cero eventos en un periodo,  $y_{ij} = 0$ , observados correspondan a un grupo de personas que no puedan presentar el desenlace de interés pero que hacen parte de la población de estudio. En este caso ajustar un modelo de Poisson con efectos aleatorios subestima la probabilidad de presentar cero. Una posibilidad es ajustar un modelo que “contiene” dos modelos donde el primero sería un modelo logístico de efectos mixtos para la probabilidad de presentar un cero, y el segundo un modelo Poisson con efectos mixtos para estimar la probabilidad de  $y_{ij}$ . La expresión de este modelo es la siguiente:

$$Pr(y_{ij}) = (1 - p_{ij})f(y_{ij}) + I(y_{ij})p_{ij}$$

donde:

$$f(y_{ij}) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!}$$

$$\text{logit}(p_{ij}) = w_{ij}\gamma + \nu_i$$

con  $\nu_i \sim N(0, \sigma_{\nu_i}^2)$ ,

$$\log(\lambda_{ij}) = x_{ij}\beta + v_i$$

con  $v_i \sim N(0, \sigma_{v_i}^2)$  y finalmente  $I(y_{ij}) = 1$  si  $y_{ij} = 0$  y  $I(y_{ij}) = 0$  si  $y_{ij} = 1$  o más.

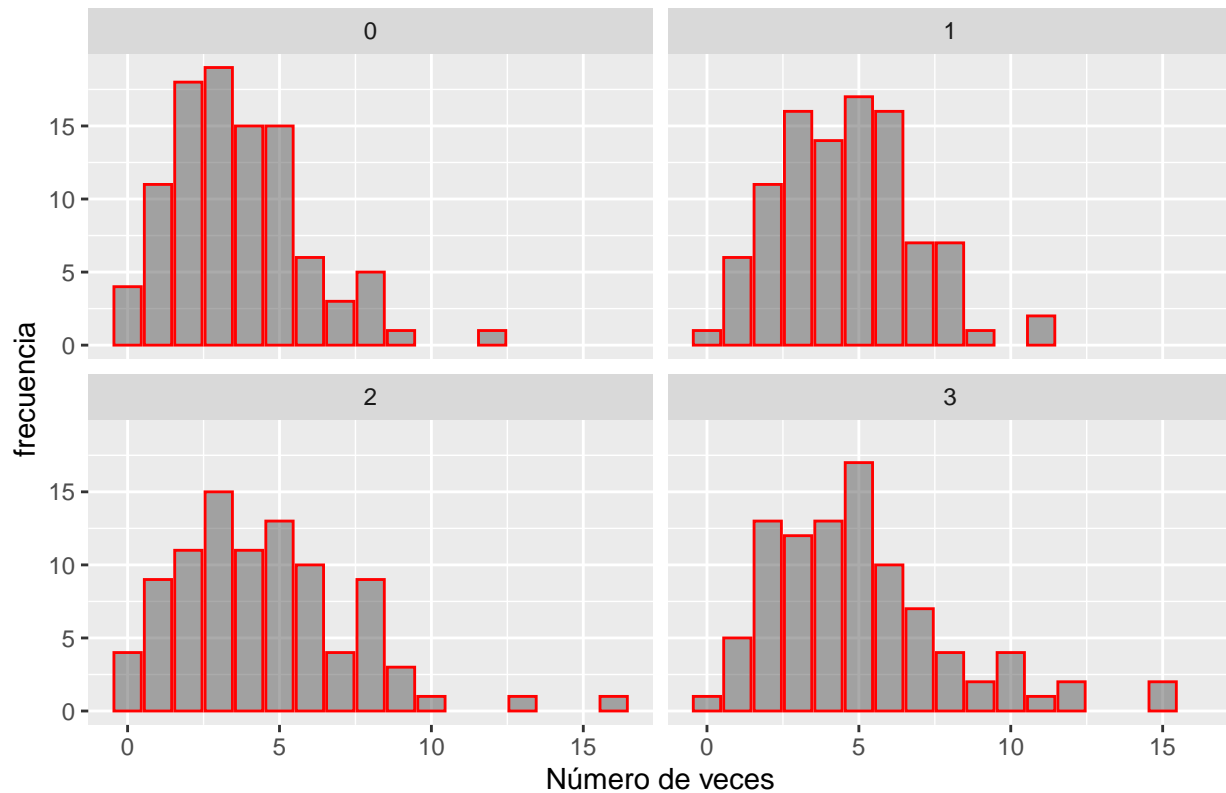
## Aplicación

```
library("rio")
data<-import("RN peso.xlsx")

library(tidyverse)

ggplot(data, aes(visitas))+
  geom_bar(alpha=0.5, color="red")+
  labs(title = "Distribución de conteos por tiempo", x="Número de veces",
        y="frecuencia")+
  facet_wrap(~t1)
```

## Distribución de conteos por tiempo



- Modelo con intercepto aleatorio

```
library(lme4)
```

```
mix1<-glmer(visitas~t1+(1|id),data=data,family = poisson)
summary(mix1, corr=FALSE)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: poisson ( log )
## Formula: visitas ~ t1 + (1 | id)
## Data: data
##
##      AIC      BIC    logLik deviance df.resid
## 1784.7  1796.5   -889.3  1778.7    378
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.2801 -0.8544 -0.1240  0.6301  4.9629
##
## Random effects:
## Groups Name      Variance Std.Dev.
## id      (Intercept) 0.01734  0.1317
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
```

```

## (Intercept) 1.33668 0.04470 29.901 < 2e-16 ***
## t1 0.09338 0.02183 4.278 1.88e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#coef(mix1) Coeficientes organizados por el id
#ranef(mix1)
#fitted(mix1) Valores ajustados en el mismo orden de la base de datos original

# Calculo razón de tasas de incidencia

se1 <- sqrt(diag(vcov(mix1)))
tab <- cbind(Est=fixef(mix1),LL=fixef(mix1)-1.96*se1,UL=fixef(mix1)+1.96*se1)
exp(tab)

## Est LL UL
## (Intercept) 3.806403 3.487078 4.154969
## t1 1.097882 1.051903 1.145869

# Ajustados versus observados

conteos<-data.frame(fitted(mix1),mix1@frame$visitas,mix1@frame$t1)
names(conteos)

## [1] "fitted.mix1." "mix1.frame.visitas" "mix1.frame.t1"

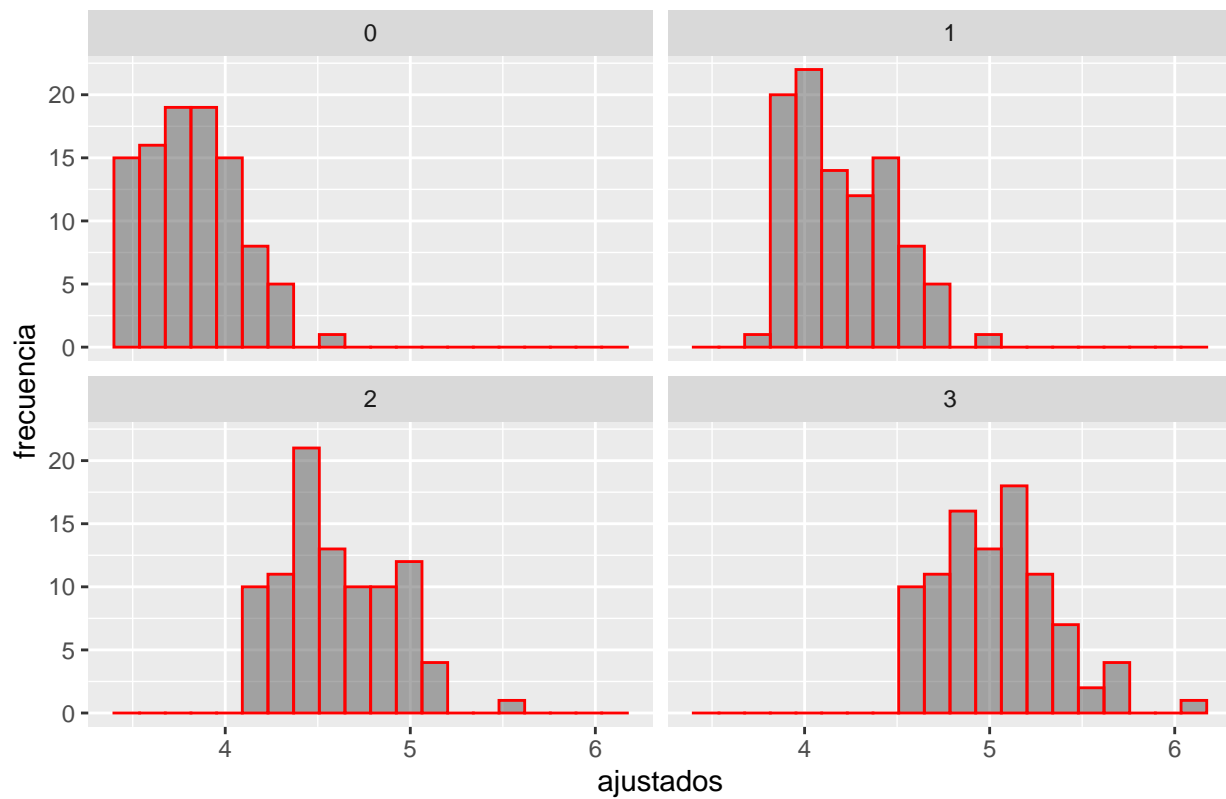
conteos<-rename(conteos,ajus=fitted.mix1.,obs=mix1.frame.visitas,
t1=mix1.frame.t1)

library(tidyverse)

ggplot(conteos,aes(x=ajus))+
  geom_histogram(bins = 20,alpha=0.5,color="red")+
  labs(title = "distribución de los eventos ajustados",x="ajustados",
y="frecuencia")+
  facet_wrap(~t1)

```

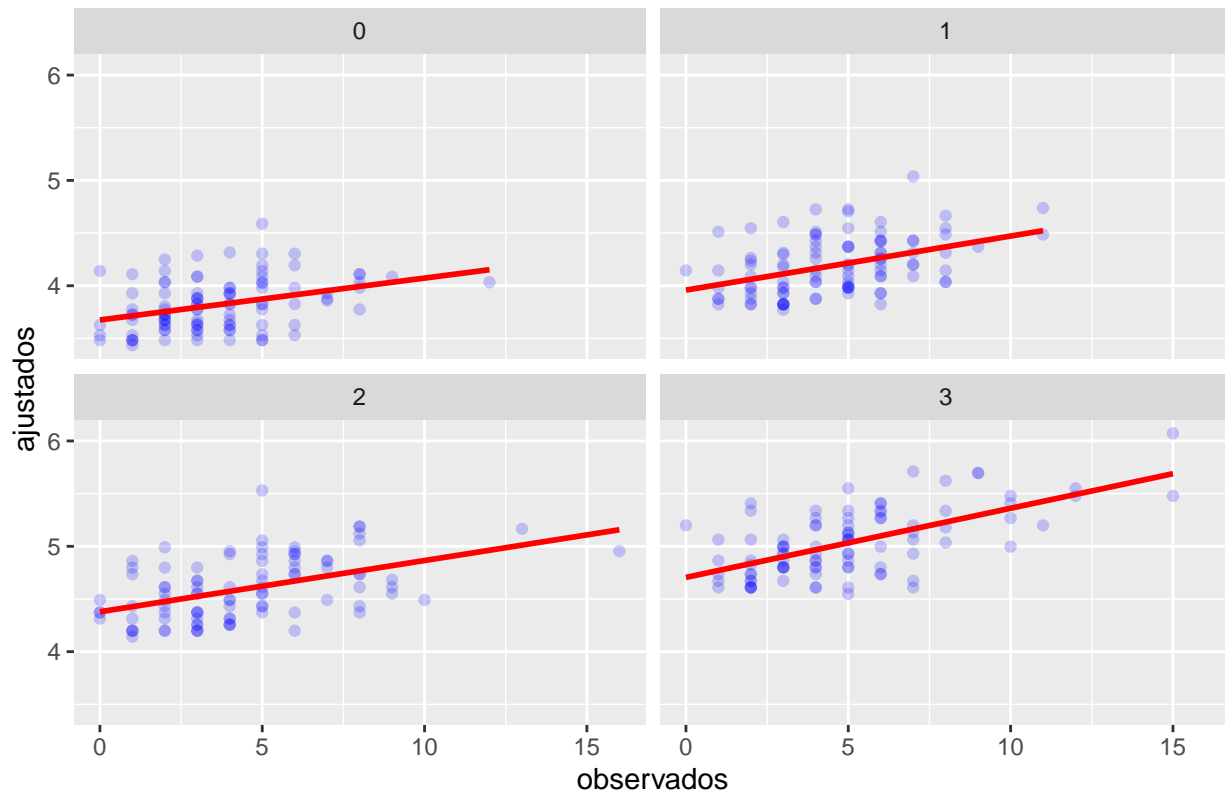
## distribución de los eventos ajustados



```
ggplot(conteos,aes(y=ajus,x=obs))+  
  geom_point(alpha=0.2,col="blue")+  
  geom_smooth(method = lm,se=FALSE,color="red")+  
  labs(title = "Observados vs ajustados", x="observados",y="ajustados")+  
  facet_wrap(~t1)
```



## Observados vs ajustados



- Modelo con intercepto y pendiente aleatoria

```
library(lme4)
```

```
mix2<-glmer(visitas~t1+(1+t1|id),data=data,family = poisson)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(mix2, corr=FALSE)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: poisson ( log )
## Formula: visitas ~ t1 + (1 + t1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
## 1777.8  1797.5  -883.9  1767.8    376
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.1370 -0.8815 -0.0752  0.5845  4.1508
##
## Random effects:
## Groups Name          Variance Std.Dev. Corr
## id      (Intercept)  0.00000  0.0000
##        t1           0.01014  0.1007  NaN
## Number of obs: 381, groups: id, 98
```

```

##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.34877    0.04232  31.869 < 2e-16 ***
## t1           0.08110    0.02439   3.325 0.000883 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## convergence code: 0
## boundary (singular) fit: see ?isSingular

#coef(mix2)

se2 <- sqrt(diag(vcov(mix2)))
tab2 <- cbind(Est=fixef(mix2),LL=fixef(mix2)-1.96*se2,UL=fixef(mix2)+1.96*se2)
exp(tab2)

##           Est      LL      UL
## (Intercept) 3.852667 3.545976 4.185885
## t1          1.084477 1.033857 1.137576

# Comparar modelos

anova(mix1,mix2)

## Data: data
## Models:
## mix1: visitas ~ t1 + (1 | id)
## mix2: visitas ~ t1 + (1 + t1 | id)
##      Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## mix1  3 1784.7 1796.5 -889.33  1778.7
## mix2  5 1777.8 1797.5 -883.89  1767.8 10.876     2  0.004349 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- Modelo con intercepto aleatorio y variables fijas

```

mix3<-glmer(visitas~t1+lactancia+sexo+sueno+(1|id),data=data,family = poisson)

## Warning in checkConv(attr("opt", "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00522653
## (tol = 0.001, component 1)

summary(mix3, corr=FALSE)

## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: poisson ( log )
## Formula: visitas ~ t1 + lactancia + sexo + sueno + (1 | id)
## Data: data
##
##      AIC      BIC    logLik deviance df.resid
## 1782.2  1805.9  -885.1  1770.2     375
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -2.2877 -0.8676 -0.1242  0.6687  4.6750
##

```

```

## Random effects:
## Groups Name      Variance Std.Dev.
## id      (Intercept) 0.01582  0.1258
## Number of obs: 381, groups: id, 98
##
## Fixed effects:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.29137    0.12064  10.704 < 2e-16 ***
## t1           0.08803    0.02196   4.008 6.13e-05 ***
## lactancia   -0.13336    0.05087  -2.622 0.00875 **
## sexo        0.01614    0.05567   0.290 0.77181
## sueno       0.01662    0.01431   1.162 0.24540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## convergence code: 0
## Model failed to converge with max|grad| = 0.00522653 (tol = 0.001, component 1)

```

- Modelo ZIP

```
#install.packages("glmmTMB")
```

```
library(glmmTMB)
```

```
## Warning: package 'glmmTMB' was built under R version 3.6.1
```

```
zip<-glmmTMB(visitas~t1+(1|id),data=data,ziformula=~sexo+(1|id),family=poisson)
summary(zip)
```

```

## Family: poisson ( log )
## Formula:      visitas ~ t1 + (1 | id)
## Zero inflation:      ~sexo + (1 | id)
## Data: data
##
##      AIC      BIC   logLik deviance df.resid
##  1785.1  1808.7  -886.5  1773.1    375
##
## Random effects:
##
## Conditional model:
## Groups Name      Variance Std.Dev.
## id      (Intercept) 0.01786  0.1336
## Number of obs: 381, groups: id, 98
##
## Zero-inflation model:
## Groups Name      Variance Std.Dev.
## id      (Intercept) 7.284e-07 0.0008535
## Number of obs: 381, groups: id, 98
##
## Conditional model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.34901    0.04593  29.368 < 2e-16 ***
## t1           0.09410    0.02217   4.244 2.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:

```

```

##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.9039     0.6078  -6.423 1.33e-10 ***
## sexo        -1.5918     2.6776  -0.594  0.552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#coef(zip)
#ranef(zip)
#predict(zip,type = "link") reporta el log(lambda), opción = "response" reporta
#lambda*(1-pij)

#exp(confint(zip,component = c("cond", "zi")))

```