Modelos Lineales Generalizados: aplicaciones en R

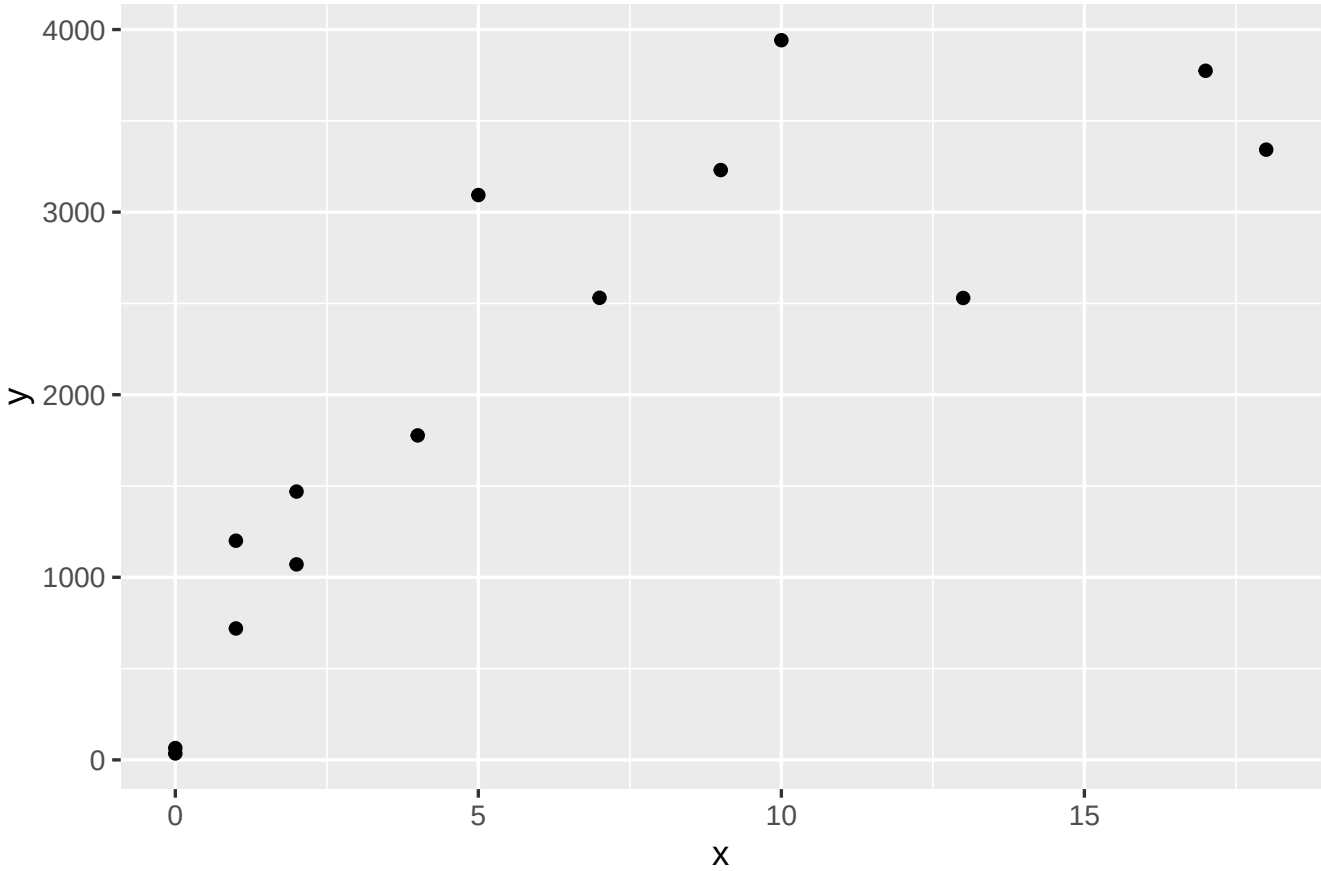
Modelos Lineales Generalizados: un enfoque aplicado

Ana M. Bianco Jemina García
(anambianco@gmail.com) (jeminagarcia@gmail.com)

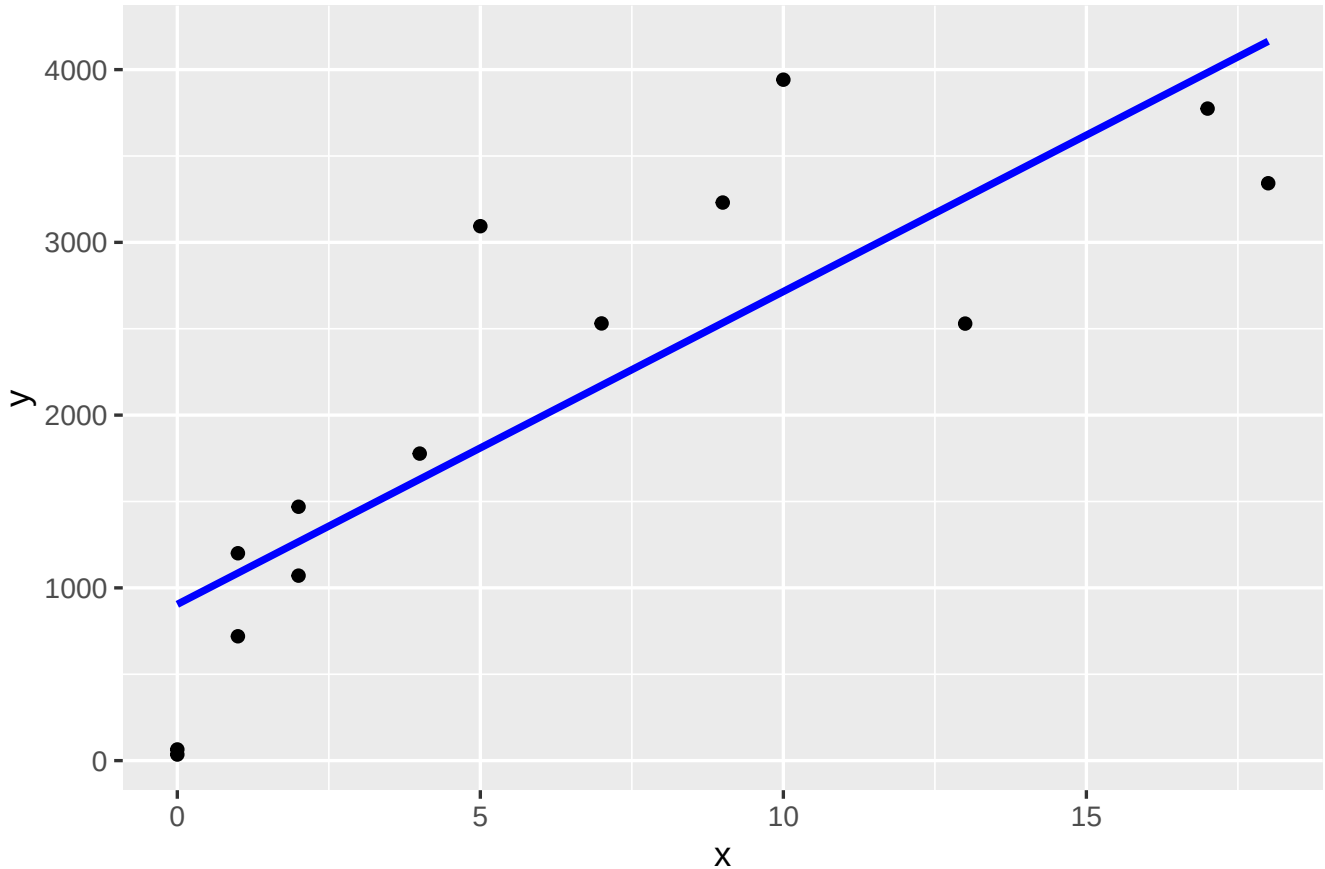
2. Modelo Lineal: Popurrí

Error de Predicción y su estimación

A partir de una muestra (muestra de entrenamiento), proponemos y estimamos un modelo.

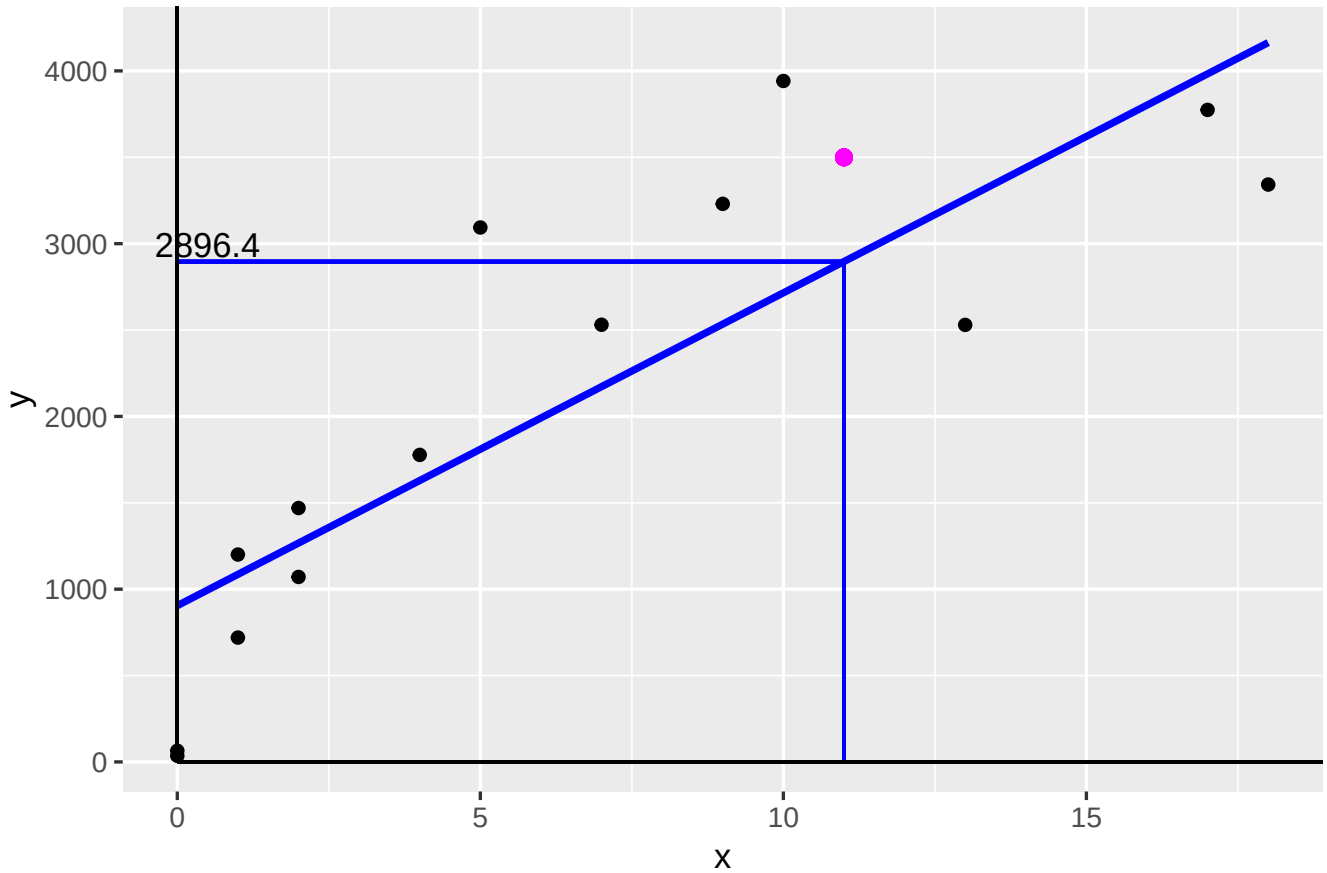


Error de Predicción y su estimación



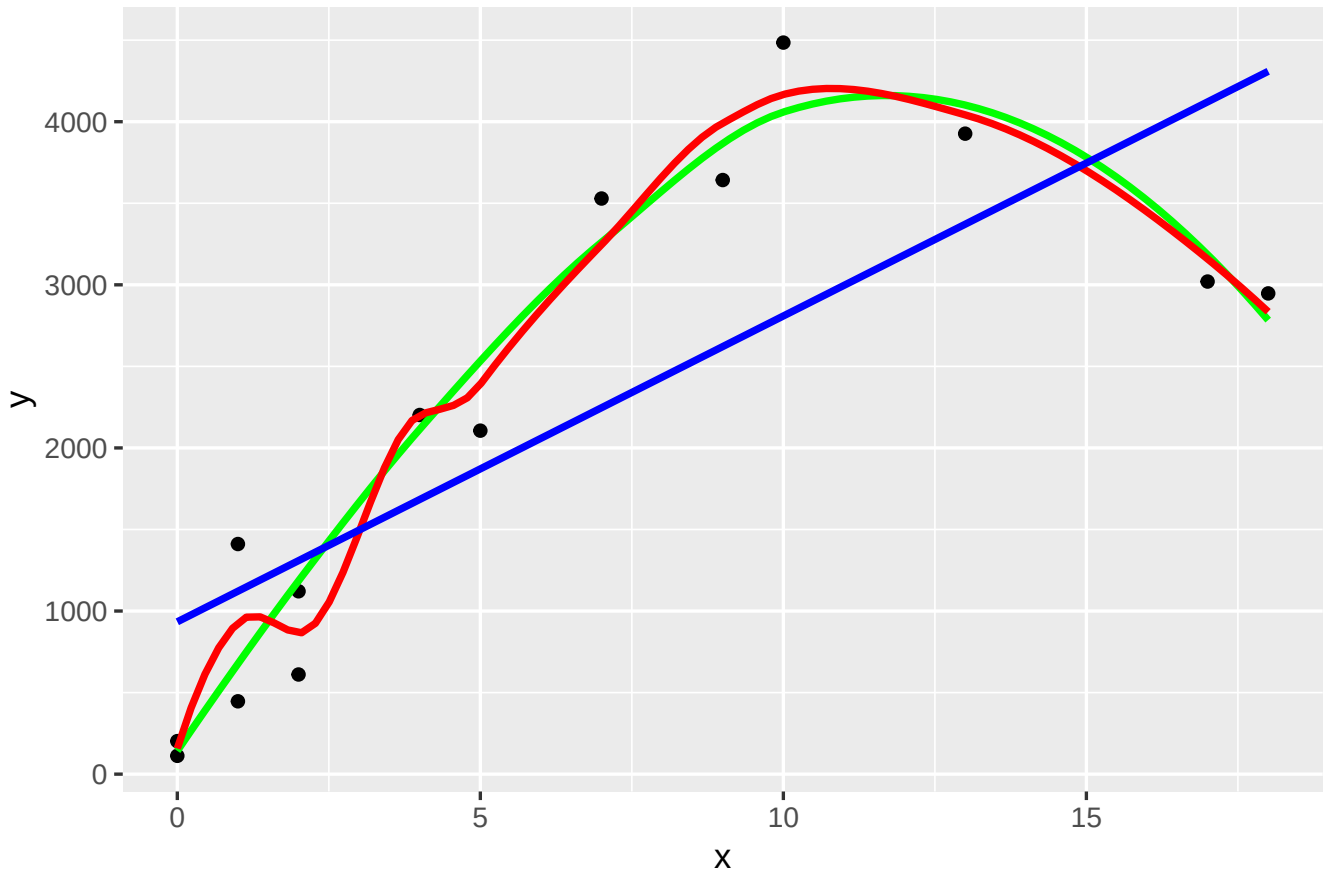
Error de Predicción y su estimación

Predecimos con el modelo estimado en $x = 11$. Realizamos una nueva observación en $x = 11$ y resulta $y = 3500$.



Error de Predicción y su estimación

Realizamos otras estimaciones a partir de otros modelos o prrocedimientos: ¿cuál elegimos?



Balance de flexibildad-rigidez

Existe una tensión entre el deseo de que el modelo sea complejo o **flexible** y que sea **rígido**.

La **flexibilidad** hará que se adapte facilmente al conjunto de datos, pero también al ruido.

La **rigidez** es deseable para tener robustez ante patrones de ruido o particulariades del presente conjunto de datos no repetibles de en otra muestra de entrenamiento.

Error de Predicción y su estimación

Supongamos que y es a quien queremos predecir a partir de x y que se cumple el modelo

$$y = f(x) + \epsilon$$

donde el error ϵ es tal que $E(\epsilon) = 0$.

Si llamamos $\widehat{f}(x)$ a la estimación de $f(x)$ que obtenemos a partir de la muestra, la predicción será

$$\widehat{y} = \widehat{f}(x)$$

y error de predicción será

$$y - \widehat{y}$$

Buscamos que el error de predicción sea lo más pequeño posible.

Error de Predicción y su estimación

Para ello, elegimos una función de pérdida, por lo general la cuadrática

$$L(y, \widehat{f}(x)) = (y - \widehat{f}(x))^2$$

Como $L(y, \widehat{f}(x))$ es aleatoria, entonces consideramos

$$Err = E(L(y, \widehat{f}(x))) = E(y - \widehat{f}(x))^2$$

Esta es una cantidad desconocida, que solo podremos estimar.

Error de Predicción y su estimación

$(x_1, y_1), \ldots, (x_n, y_n)$ es la **muestra de entrenamiento** a partir de la cual estimamos f: $\widehat{f}$

Consideremos $(x_o^1, y_o^1), \ldots, (x_o^m, y_o^m)$ una muestra futura (de prueba o de validación), el error de predicción cuadrático en el $j-$ésimo punto será

$$(y_o^j - \widehat{f}(x_o^j))^2$$

por lo tanto el **error de predicción estimado** lo calculamos como

$$\widehat{Err} = \frac{1}{m} \sum_{j=1}^{m} (y_o^j - \widehat{f}(x_o^j))^2$$

Error de Predicción y su estimación

Por otro lado, también podemos calcular el **error de entrenamiento promedio**

$$\overline{err} = \frac{1}{m}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

Típicamente $\widehat{Err}$ es una buena estimación de Err en tanto se acerca a él cuando crece el tamaño de la muestra, mientras que $\overline{err}$ da una estimación pobre.

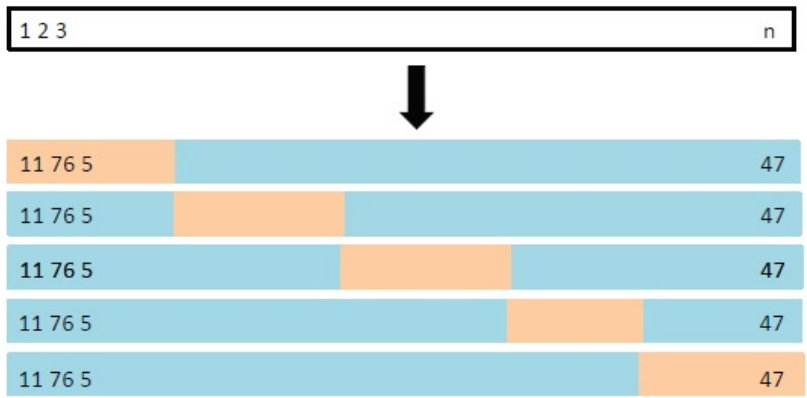
Convalidación Cruzada: ¿Muestra Futura?

Una de las formas más populares de estimar el error de predicción es el método de **convalidación cruzada**.

Para ello, dada una muestra la dividimos aleatoriamente en dos partes: muestra de entrenamiento y muestra de testeo o de validación.

Convalidación Cruzada (Cross-Validation)

Aquí hay una solución de compromiso con los tamaños muestrales. Como muchas veces los datos escasean, se suele usar una versión llamada **K-fold Cross-Validation**, en la que los datos se dividen aleatoriamente en k folds o partes como muestra la figura:



Convalidación Cruzada

Se divide al azar en K partes o folds los datos, se separa al primer fold, se *estima* a f con las $K-1$ muestras restantes y se calcula el error estimado en ese primer fold: $\widehat{Err}_1$

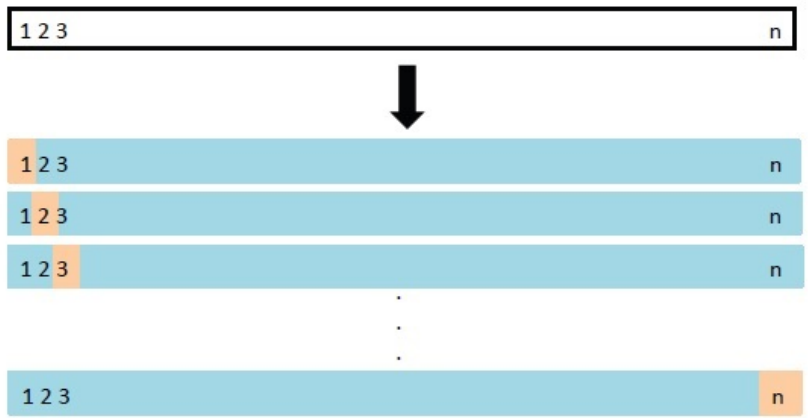
Luego se repite tomando como muestra j de validación a cada uno de los folds y se computa $\widehat{Err}_j$.

Finalmente, se estima el error de predicción como

$$\widehat{Err} = \frac{1}{K} \sum_{j=1}^{K} \widehat{Err}_j$$

Convalidación Cruzada

Muchas veces se usa $K = N$, si bien este método podría introducir mayor varianza debido a que las muestras de entrenamiento son muy parecidas.



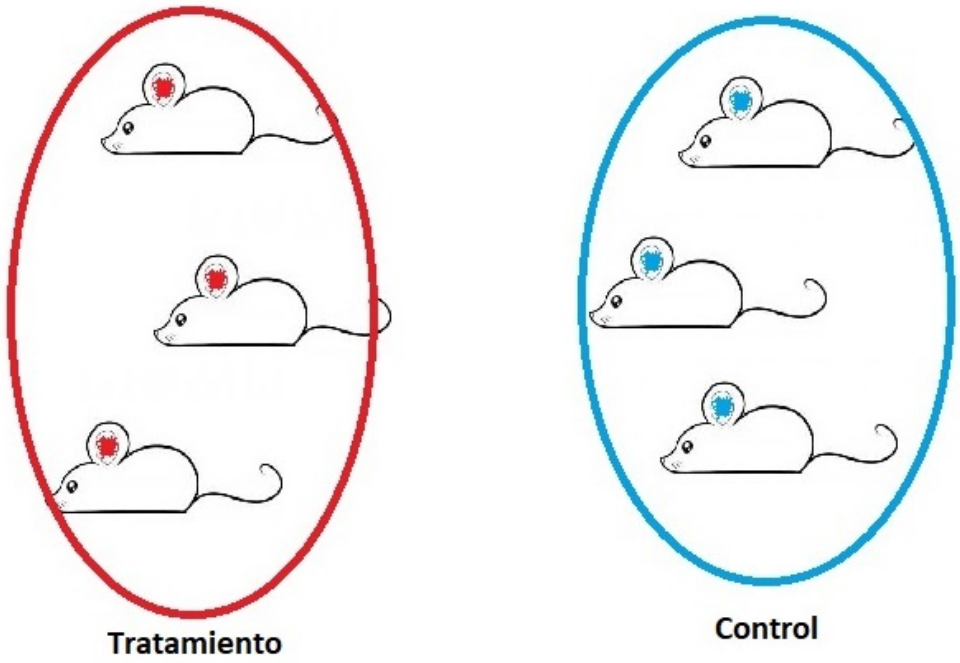
Análisis de la Varianza: ANOVA

El modelo lineal ofrece una estructura muy versátil con la que se pueden tratar problemas muy diversos. Uno de ellos, que parece muy alejado, pero no lo está, es el problema de comparar la media de 2 o más poblaciones normales.

Variables de clasificación (no numéricas), tales como tratamientos, grupos o bloques pueden ser incluidas en el modelo como variables dummies (0 y 1) y por lo tanto, el problema de comparación puede tratarse linealmente.

Comparación de dos medias

Supongamos que tenemos 2 grupos con 3 observaciones cada uno:



Análisis de la Varianza: ANOVA

$$
\begin{aligned}
y_{ij} &= \mu_i + \epsilon_{ij} \\
y_{ij} &= \mu + \beta d_i + \epsilon_{ij} \\
& \quad i = 1, 2 \quad j = 1, 2, 3
\end{aligned}
$$

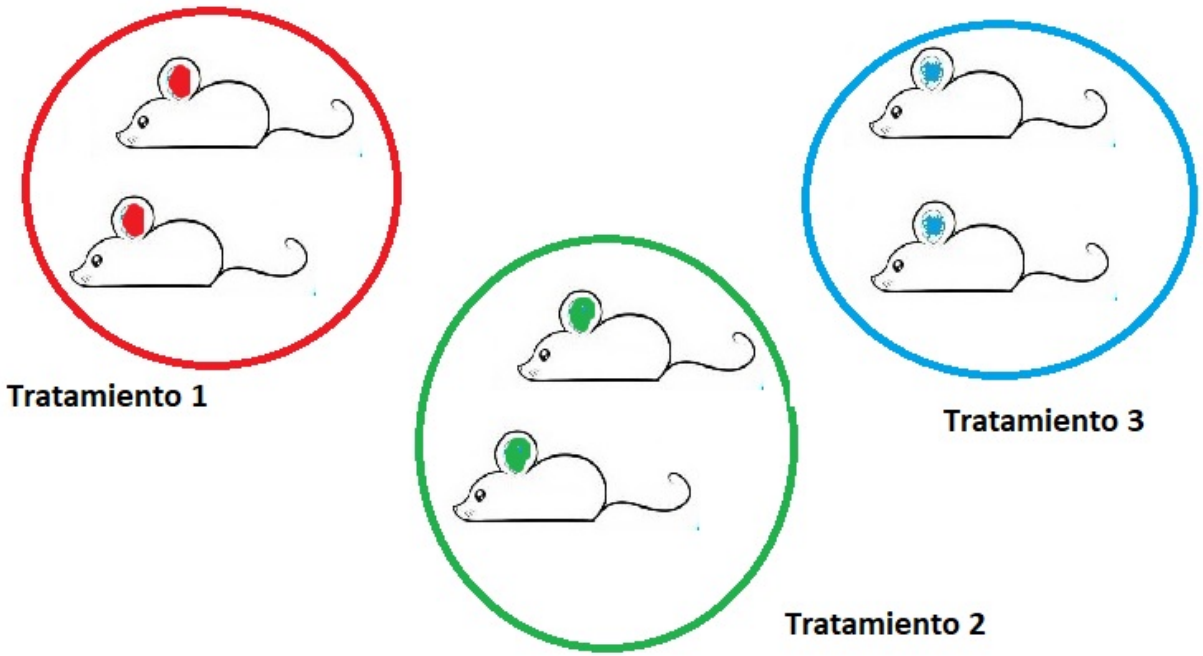
- μ: media general, β: efecto tratamiento
- $\begin{cases} d_i = 1 & \text{si la observacion pertenece al grupo tratado} \\ d_i = 0 & \text{si la observacion pertenece al control} \end{cases}$

En Notación Matricial

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

Análisis de la Varianza: ANOVA

Supongamos que tenemos 3 grupos con 2 observaciones cada uno:



Análisis de la Varianza: ANOVA

$$
\begin{aligned}
y_{ij} &= \mu_i + \epsilon_{ij} \\
y_{ij} &= \mu + \beta_i + \epsilon_{ij} \\
& \quad i = 1, 2, 3 \quad j = 1, 2
\end{aligned}
$$

μ: media general, β_i: efecto tratamiento i

Análisis de la Varianza: ANOVA

$$
\begin{aligned}
y_{ij} &= \mu_i + \epsilon_{ij} \\
y_{ij} &= \mu + \beta_i + \epsilon_{ij} \\
& \quad i = 1, 2, 3 \quad j = 1, 2
\end{aligned}
$$

μ: media general, β_i: efecto tratamiento i

$$
\begin{aligned}
y_{ij} &= \mu + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon_{ij} \\
&= \mu + \beta_i d_i \\
& \quad i = 1, 2, 3 \quad j = 1, 2
\end{aligned}
$$

- $\begin{cases} d_i = 1 & \text{si la observacion pertenece al grupo tratado i} \\ d_i = 0 & \text{caso contrario} \end{cases}$

En Notación Matricial

Notemos que si se conocen los valores de d_1 y d_2, se sabe cuanto vale d_3, asi que es redundante y podemos omitir esa componente del modelo.

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

Más sobre dummies

Estas variables tambieén pueden ser incorporadas al modelo lineal junto con una variable cuantitativa. Supongamos que tenemos dos grupos: Grupo 1 y Grupo 0.

Por ejemplo, d_i: indica la pertenencia de la observación i al Grupo 1

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i \quad i = 1, 2, \ldots, n$$

Por lo tanto:

- En el Grupo 1 ajustamos el modelo $y_i = \beta_0 + \beta_1 x_i + \beta_2 + \epsilon_i$
- En el Grupo 0 ajustamos el modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

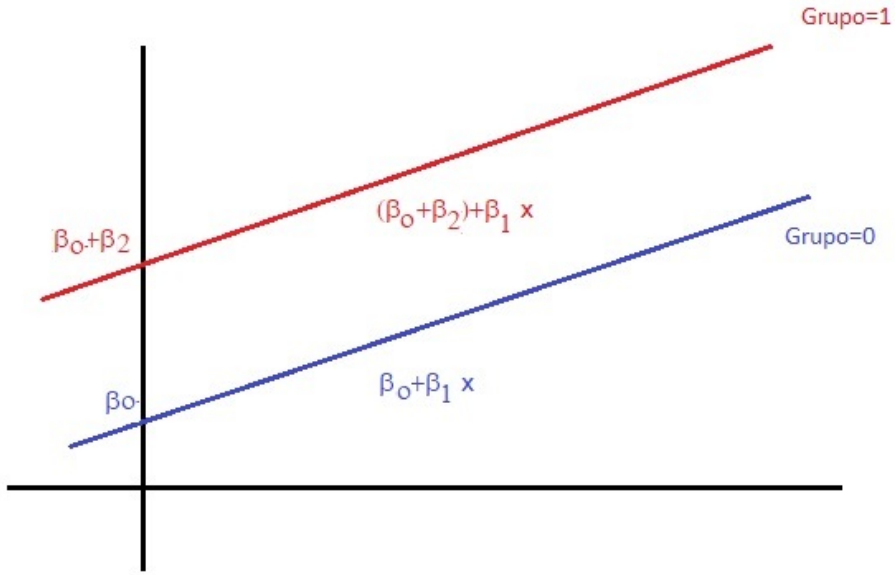
Más sobre dummies

Estas variables tambieén pueden ser incorporadas al modelo lineal junto con una variable cuantitativa. Supongamos que tenemos dos grupos: Grupo 1 y Grupo 0.

Por ejemplo, d_i: indica la pertenencia de la observación i al Grupo 1

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i \quad i = 1, 2, \ldots, n$$

- β_1 indica la variación de la respuesta y cuando x cambia en una unidad y dejamos d_i fijo
- β_2 representa el cambio en la respuesta y cuando cambiamos del Grupo 1 al Grupo 0
- Este modelo representa dos rectas paralelas con distinta ordenada al origen


Grupo=1
$(\beta_o+\beta_2)+\beta_1$ x
$\beta_o+\beta_2$
Grupo=0
$\beta_o+\beta_1$ x
β_o

Un ejemplo: Diabetes in Pima Women

Los datos scorresponde a una población de mujeres de 21 años o más de ascendencia Pima residentes en Phoenix, Arizona. Se chequeó si eran diabéticas o no de acuerdo a los criterios de la World Health Organization. Los datos fueron recogidos por la US National Institute of Diabetes and Digestive and Kidney Diseases. De los 532 registros completos el archivo Pima.tr contiene una muestra de tamaño 200.

Las variables registradas son:

- *glu*: plasma glucose concentration in an oral glucose tolerance test.
- *bp*: diastolic blood pressure (mm Hg).
- *skin*: triceps skin fold thickness (mm).
- *bmi*: body mass index (weight in kg/(height in m)^2).
- *ped*: diabetes pedigree function.
- *age*: age in years.
- *type*: Yes or No, for diabetic according to WHO criteria.

Datos Diabetes

```
> library(MASS)
> head(Pima.tr)
```

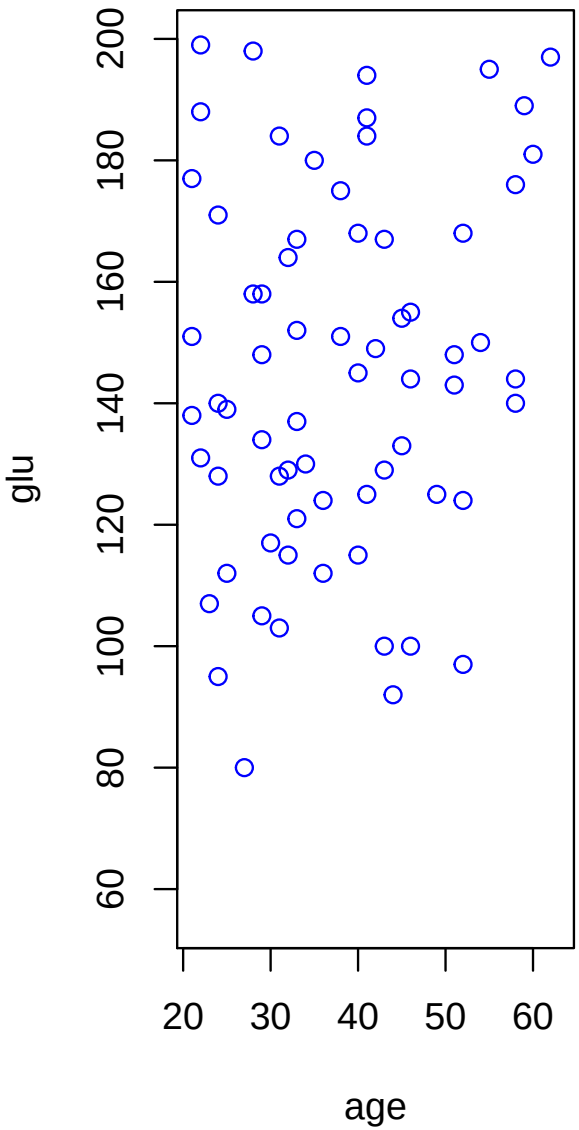
	npreg	glu	bp	skin	bmi	ped	age	type
1	5	86	68	28	30.2	0.364	24	No
2	7	195	70	33	25.1	0.163	55	Yes
3	5	77	82	41	35.8	0.156	35	No
4	0	165	76	43	47.9	0.259	26	No
5	0	107	60	25	26.4	0.133	23	No
6	5	97	76	27	35.6	0.378	52	Yes

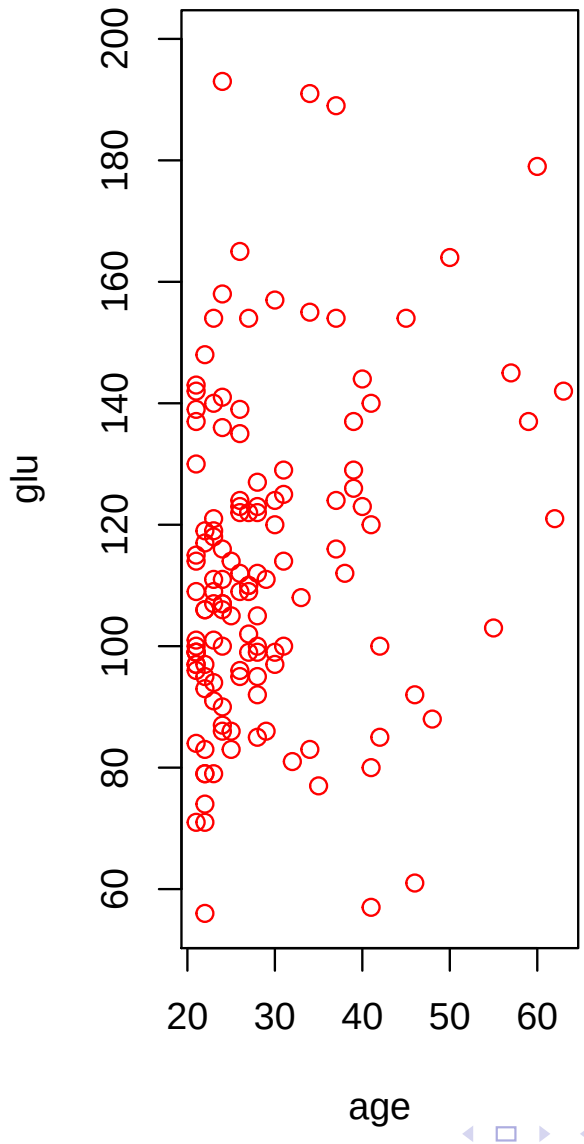
```
> attach(Pima.tr)
```

Datos Diabetes

```
> modelo.simple<- lm(glu~age)
> summary(modelo.simple)

Call:
lm(formula = glu ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-76.733 -18.093  -4.806  20.320  85.047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.1546     6.5332  14.106  < 2e-16 ***
age           0.9908     0.1926   5.145 6.41e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.82 on 198 degrees of freedom
Multiple R-squared:  0.1179, Adjusted R-squared:  0.1135
F-statistic: 26.47 on 1 and 198 DF,  p-value: 6.415e-07
```

Datos Diabetes





Datos Diabetes

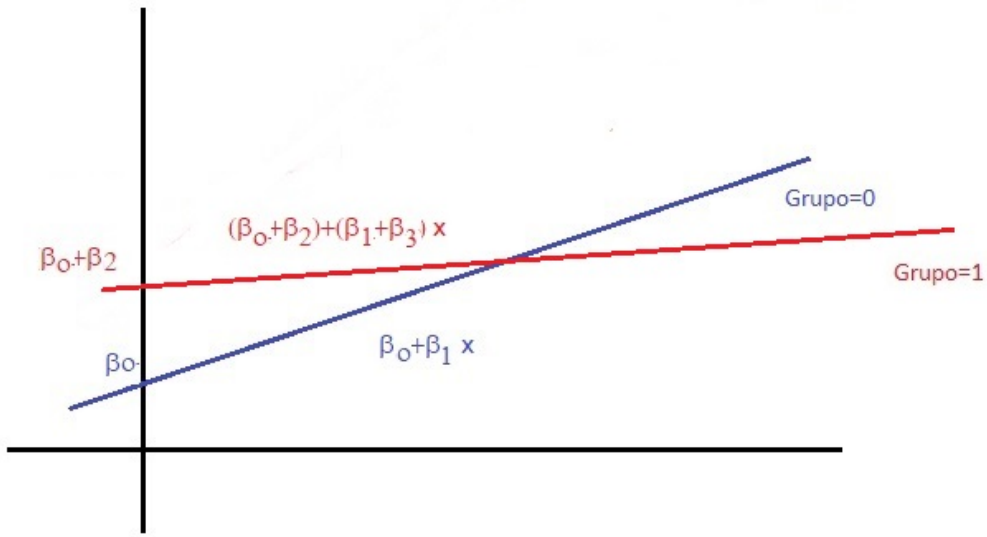
```
> modelo.dumy<- lm(glu~age+factor(type))
> summary(modelo.dumy)

Call:
lm(formula = glu ~ age + factor(type))

Residuals:
    Min      1Q  Median      3Q     Max
-62.693 -16.755  -2.135  15.552  82.825

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      96.7372     6.0355  16.028  < 2e-16 ***
age               0.5599     0.1897   2.951  0.00355 **
factor(type)Yes  27.2180     4.3848   6.207 3.13e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.34 on 197 degrees of freedom
Multiple R-squared:  0.2622, Adjusted R-squared:  0.2547
F-statistic: 35.01 on 2 and 197 DF,  p-value: 9.778e-14
```

Datos Diabetes

¿ Cómo resulta el modelo ajustado en cada grupo?

Más sobre dummies: Interacción

Sin embargo, las dos rectas podrían no ser paralelas.....



Más sobre dummies: Interacción

Sin embargo, las dos rectas podrían no ser paralelas, en este caso para reflejar el cambio en la pendiente la variable dummy podría también ser usada en el término que acompaña a x.

Por ejemplo, d_i: indica la pertenencia de la observación i al Grupo 1

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \epsilon_i \quad i = 1, 2, \ldots, n$$

Por lo tanto, en cada grupo ajustamos el modelo

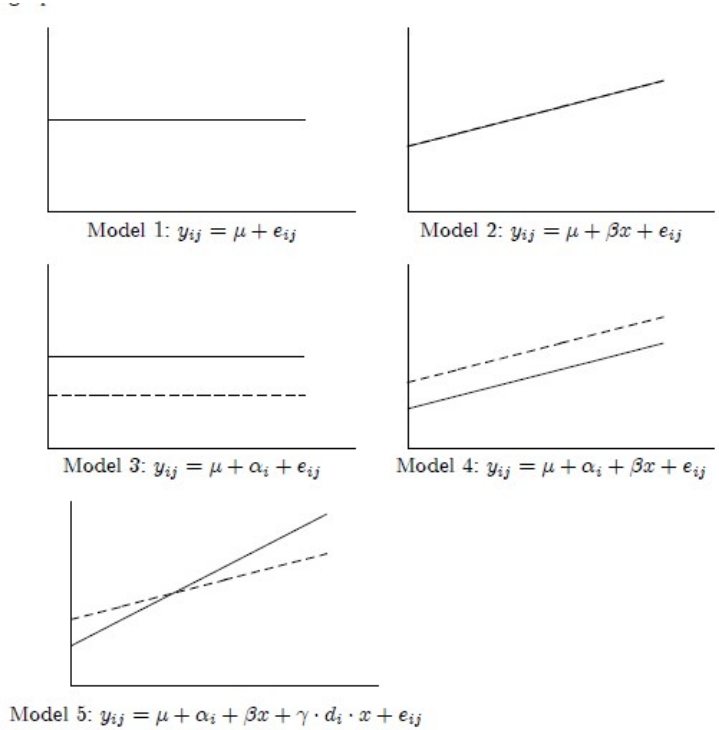
- Grupo 1: $y_i = \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i + \epsilon_i$
- Grupo 0: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$x_i d_i$ es el término de la interacción

Volvamos a los Datos Diabetes

```
Call:
lm(formula = bp ~ age + factor(type) + age * factor(type))

Residuals:
    Min      1Q  Median      3Q     Max
-33.652  -7.268   0.171   6.004  36.466

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         54.52808    2.95331  18.463  < 2e-16 ***
age                  0.51368    0.09607   5.347 2.47e-07 ***
factor(type)Yes     12.53792    5.29687   2.367   0.0189 *
age:factor(type)Yes -0.31411    0.14731  -2.132   0.0342 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.49 on 196 degrees of freedom
Multiple R-squared:  0.177, Adjusted R-squared:  0.1644
F-statistic: 14.05 on 3 and 196 DF,  p-value: 2.474e-08
```

En resumen...



Model 1: $y_{ij} = \mu + e_{ij}$

Model 2: $y_{ij} = \mu + \beta x + e_{ij}$

Model 3: $y_{ij} = \mu + \alpha_i + e_{ij}$

Model 4: $y_{ij} = \mu + \alpha_i + \beta x + e_{ij}$

Model 5: $y_{ij} = \mu + \alpha_i + \beta x + \gamma \cdot d_i \cdot x + e_{ij}$