

# Clase 3

*Pachá*

*June 30, 2016*

## Modelo de Regresión Básico

- Mínimos cuadrados es una herramienta de estimación.
- Para realizar inferencia se desarrolla un modelo probabilístico de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Aquí  $\varepsilon_i$  se asume iid  $N(0, \sigma^2)$ .
- Note que  $E[Y_i | X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$
- Note que  $Var(Y_i | X_i = x_i) = \sigma^2$ .
- La estimación por ML de  $\beta_0$  y  $\beta_1$  coincide con la estimación por OLS

$$\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- $E[Y | X = x] = \beta_0 + \beta_1 x$
- $Var(Y | X = x) = \sigma^2$

## Interpretación de los coeficientes

### Intercepto

- $\beta_0$  es el valor esperado del output cuando el input es 0

$$E[Y | X = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

- Note que esto no siempre es de interés, por ejemplo cuando  $X = 0$  es imposible o está fuera del rango de los datos (e.g. Si  $X$  corresponde a presión sanguínea, estatura, etc.)
- Considere que

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + a\beta_1 + \beta_1(X_i - a) + \varepsilon_i = \tilde{\beta}_0 + \beta_1(X_i - a) + \varepsilon_i$$

Entonces, si desplazamos  $X$  en  $a$  unidades cambia el intercepto pero no la pendiente. menudo  $a$  se fija en  $\bar{X}$  tal que el intercepto se interpreta como la respuesta esperada en el valor promedio de  $X$ .

### Pendiente

- $\beta_1$  es el cambio esperado en el output cuando el input cambia en una unidad

$$E[Y | X = x + 1] - E[Y | X = x] = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x) = \beta_1$$

- Considere el impacto de cambiar las unidades (medición) de  $X$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \beta_0 + \frac{\beta_1}{a}(X_i a) + \varepsilon_i = \beta_0 + \tilde{\beta}_1(X_i a) + \varepsilon_i$$

- Entonces, la multiplicación de  $X$  por un factor  $a$  resulta en que se divide el coeficiente por el mismo factor  $a$ .
- Si queremos predecir el output dado un valor del input, digamos  $X$ , el modelo de regresión predice

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

## Ejemplo

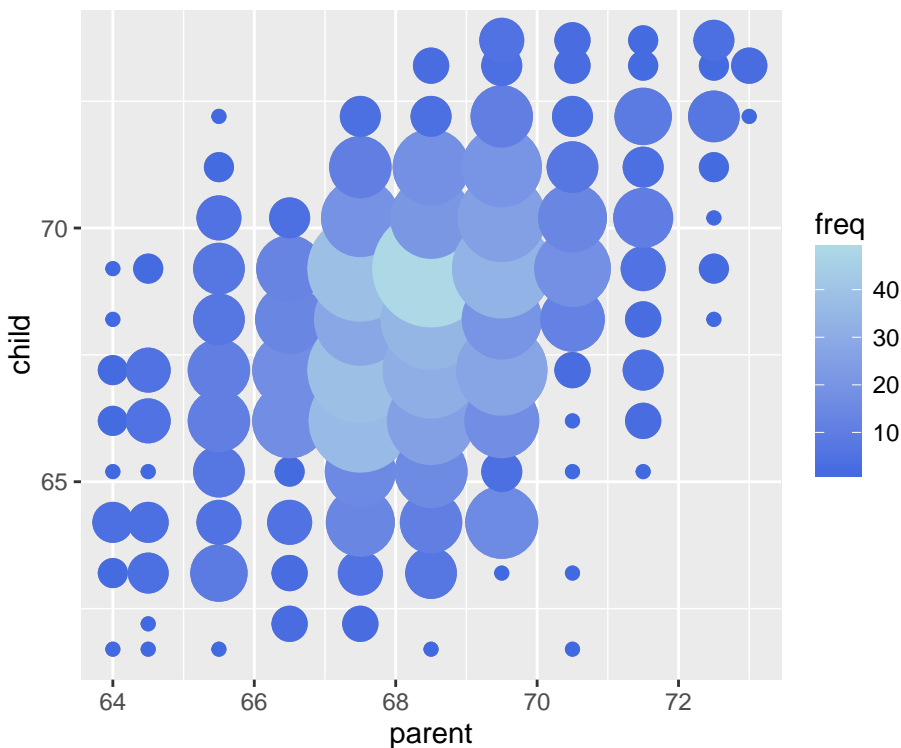
- Si  $X$  es la estatura en  $m$  e  $Y$  es el peso en  $kg$ . Entonces  $\beta_1$  es  $kg/m$ . Convirtiendo  $X$  en  $cm$  implica multiplicar  $X$  por  $100cm/m$ . Para obtener  $\beta_1$  en las unidades correctas, tenemos que dividir por  $100cm/m$  y así se tendrán las unidades correctas.

$$Xm \times \frac{100cm}{m} = (100X)cm \text{ y } \beta_1 \frac{kg}{m} \times \frac{1m}{100cm} = \left( \frac{\beta_1}{100} \right) \frac{kg}{cm}$$

## Modelo Lineal Univariado

### Los Datos de Galton

Francis Galton (1882 - 1911) sentó las bases de la Econometría estudiando la estatura de padres e hijos. Los datos de su estudio están disponibles en la instalación de R. El gráfico de estatura de padres versus estatura de hijos es el siguiente:



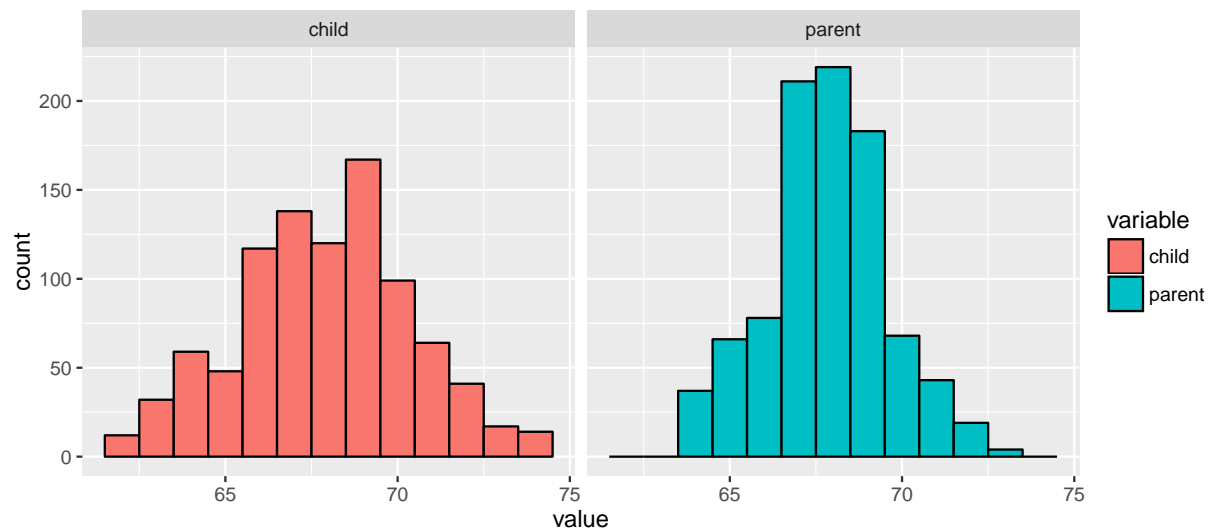
## Ejercicios

- Usar la estatura de los padres para predecir la de los hijos
- Encontrar una relación entre ambas estaturas
- Encontrar la variación de la estatura de los hijos que no depende de la estatura de los padres (variación residual)
- Los supuestos que se necesitan para generalizar más allá de los datos
- Por qué los hijos de padres muy altos tienden a ser más bajos (regresión a la media)

## Análisis de los datos

- Datos recolectados y analizados por Francis Galton en 1885.
- Galton fue un científico que creó los conceptos de correlación y regresión.
- Veremos la distribución marginal de los datos.
- La corrección por género se obtiene multiplicando la estatura de las mujeres por 1,08.

```
library(reshape); long <- melt(galton)
g <- ggplot(long, aes(x = value, fill = variable))
g <- g + geom_histogram(colour = "black", binwidth=1)
g <- g + facet_grid(. ~ variable)
g
```



## Encontrando la media usando mínimos cuadrados

- Veamos un poco los datos

```
head(galton, n=10) #primeras 10 observaciones (medidas en pulgadas)
```

```
##   child parent
## 1   61.7   70.5
## 2   61.7   68.5
## 3   61.7   65.5
## 4   61.7   64.5
## 5   61.7   64.0
## 6   62.2   67.5
## 7   62.2   67.5
## 8   62.2   67.5
## 9   62.2   66.5
## 10  62.2   66.5
```

```
dim(galton) #tamaño muestral = 928 ; variables = 2
```

```
## [1] 928  2
```

- Considere solo la estatura de los hijos. ¿Cómo se describe la media?

- Una definición es que siendo  $Y_i$  la estatura del hijo  $i$  para  $i = 1, \dots, n$  con  $n = 928$  entonces la media es el valor de  $\mu$  que minimiza la ecuación  $\sum_{i=1}^n (Y_i - \mu)^2$
- Se tiene que  $\mu = \bar{Y}$ .

## Regresión

### Sin constante

$$Y_i = \beta_1 X_i + \varepsilon_i$$

```
lm(child ~ parent -1, data = galton)

##
## Call:
## lm(formula = child ~ parent - 1, data = galton)
##
## Coefficients:
## parent
## 0.9965
```

### Con constante

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

```
lm(child ~ parent, data = galton)

##
## Call:
## lm(formula = child ~ parent, data = galton)
##
## Coefficients:
## (Intercept)      parent
##      23.9415      0.6463
```

### Sin constante (centrando los datos)

$$(Y_i - \bar{Y}) = \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

$$\tilde{Y}_i = \beta_0 \beta_1 \tilde{X}_i + \varepsilon_i$$

```
lm(I(child - mean(child)) ~ I(parent - mean(parent)) -1, data = galton)

##
## Call:
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -
##      1, data = galton)
##
## Coefficients:
## I(parent - mean(parent))
##              0.6463
```

### Con constante (centrando los datos)

$$(Y_i - \bar{Y}) = \beta_0 + \beta_1(X_i - \bar{X}) + \varepsilon_i$$

$$\tilde{Y}_i = \beta_0\beta_1\tilde{X}_i + \varepsilon_i$$

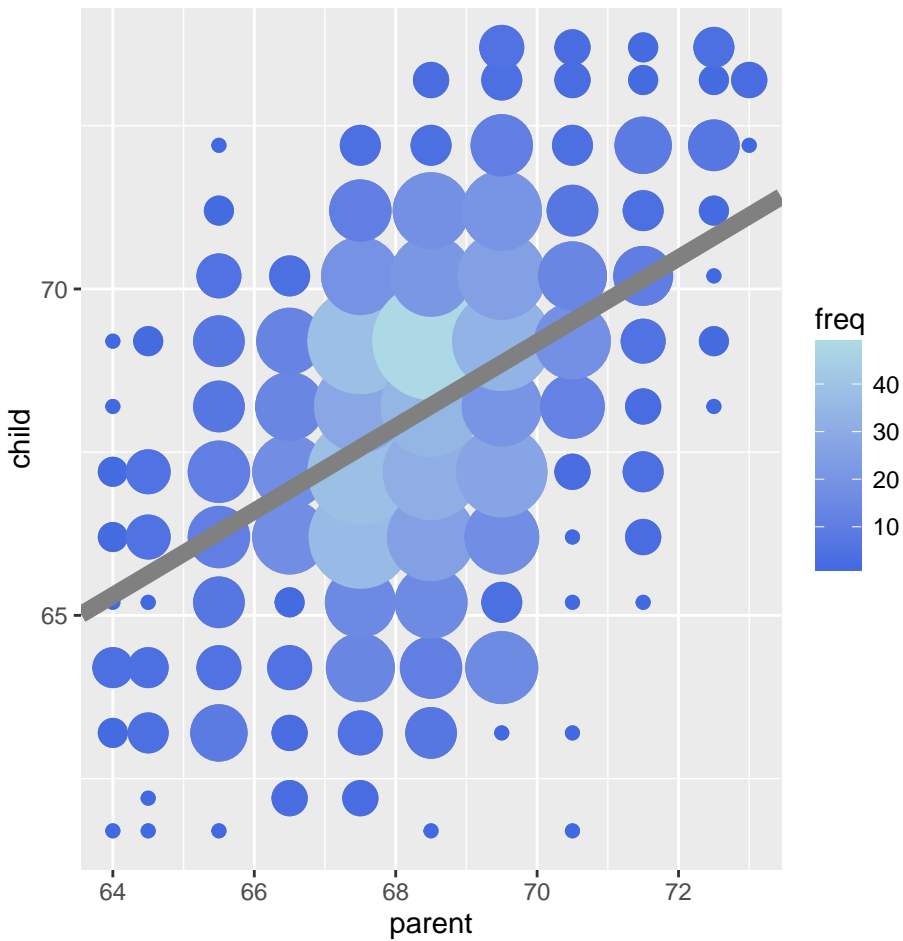
```
lm(I(child - mean(child)) ~ I(parent - mean(parent)) -1, data = galton)
```

```
##  
## Call:  
## lm(formula = I(child - mean(child)) ~ I(parent - mean(parent)) -  
##     1, data = galton)  
##  
## Coefficients:  
## I(parent - mean(parent))  
##                        0.6463
```

### Observaciones

- La primera regresión (sin constante) sobreestima el efecto de  $X$  sobre  $Y$  pues se fuerza que el intercepto sea cero
- No es coincidente que las últimas tres regresiones entreguen un idéntico valor de  $\beta_1$ .
- Lo anterior se debe a que al no incluir el “-1” en el código en R, no estamos forzando que el intercepto sea cero y por ende no estamos sesgando el efecto de  $X$  sobre  $Y$ .
- En las dos últimas regresiones se observa que el intercepto al centrar los datos (es decir, al cambiar la unidad de medición y hacer que el “0” la variable sea el promedio de dicha variable) sea un valor muy pequeño corresponde a un hecho atribuible a los datos y no a que se fuerza la estimación.

## Gráfico



## Ajuste de la mejor recta de regresión

- Sea  $Y_i$  la estatura del hijo  $i^{th}$  y  $X_i$  la estatura del padre  $i^{th}$ .
- Considere la recta con mejor ajuste  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ .
- La ecuación de mínimos cuadrados es

$$\sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2.$$

## Resultados

- El modelo de mínimos cuadrados ajusta la recta  $Y = \beta_0 + \beta_1 X$  a través de los pares ordenados  $(X_i, Y_i)$  e  $Y_i$  es el output que se obtiene de la recta  $Y = \hat{\beta}_0 + \hat{\beta}_1 X$  con

$$\hat{\beta}_1 = \text{Cor}(Y, X) \frac{Sd(Y)}{Sd(X)} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

- $\hat{\beta}_1$  se expresa en unidades de  $Y/X$ ,  $\hat{\beta}_0$  se expresa en unidades de  $Y$ .
- La recta de regresión pasa por  $(\bar{X}, \bar{Y})$ .

- La pendiente de la recta de regresión con  $X$  como output e  $Y$  como input es  $Cor(Y, X) \frac{sd(X)}{sd(Y)}$ .
- La pendiente es la misma que se obtiene que si se centraran los datos  $(X_i - \bar{X}, Y_i - \bar{Y})$  y se estimara una regresión que pasa por  $(0, 0)$ .
- Si se normalizan los datos  $\left(\frac{X_i - \bar{X}}{sd(X)}, \frac{Y_i - \bar{Y}}{sd(Y)}\right)$ , la pendiente es  $Cor(Y, X)$ .

## En síntesis

Es interesante notar que las estimaciones del programa coinciden con el cálculo siguiendo las fórmulas por definición.

```
y <- galton$child
x <- galton$parent
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
rbind(c(beta0, beta1), coef(lm(y ~ x)))
```

```
##      (Intercept)      x
## [1,]    23.94153 0.6462906
## [2,]    23.94153 0.6462906
```

La regresión desde el origen conserva la pendiente si primero centramos los datos

```
yc <- y - mean(y)
xc <- x - mean(x)
beta1 <- sum(yc * xc) / sum(xc ^ 2)
c(beta1, coef(lm(y ~ x))[2])
```

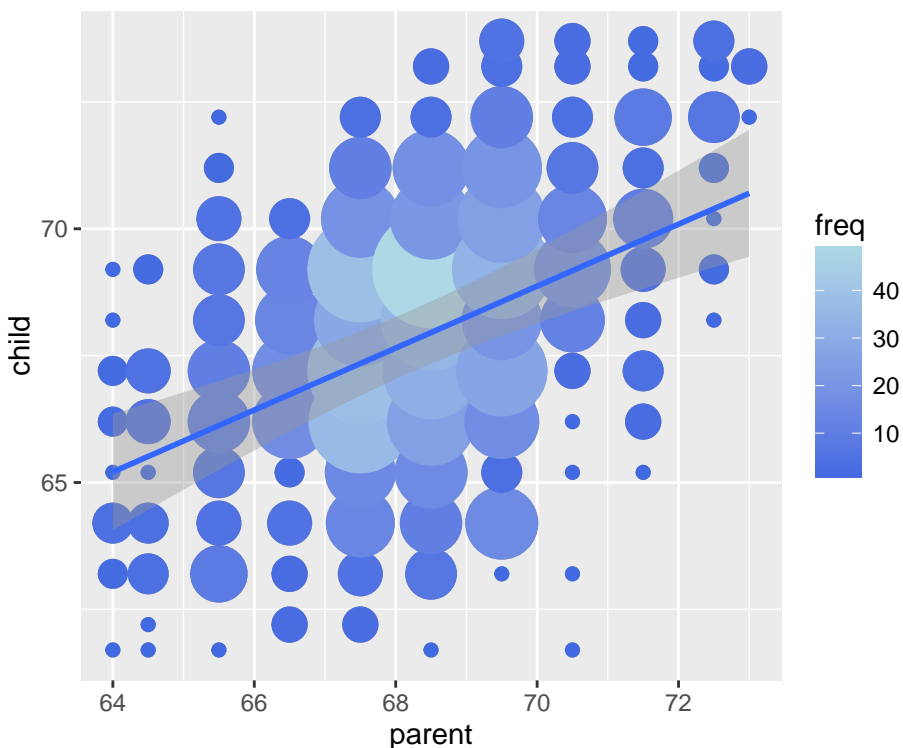
```
##      x
## 0.6462906 0.6462906
```

Si se normalizan los datos la pendiente es igual al coeficiente de correlación

```
yn <- (y - mean(y))/sd(y)
xn <- (x - mean(x))/sd(x)
c(cor(y, x), cor(yn, xn), coef(lm(yn ~ xn))[2])
```

```
##      xn
## 0.4587624 0.4587624 0.4587624
```

Mejor recta de regresión:



## Modelo Lineal Multivariado

### Extensión del caso univariado

- El modelo lineal generalizado extiende el modelo lineal simple (SLR) agregando términos linealmente al modelo. Típicamente  $X_{1i} = 1$  (se incluye un intercepto).

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^p X_{ki} \beta_k + \epsilon_i.$$

- La estimación por OLS (y también la estimación por ML bajo supuestos de iid y errores Gaussianos) minimiza

$$\sum_{i=1}^n \left( Y_i - \sum_{k=1}^p X_{ki} \beta_k \right)^2.$$

- Lo importante es la linealidad de los coeficientes, entonces

$$Y_i = \beta_1 X_{1i}^2 + \beta_2 X_{2i}^2 + \dots + \beta_p X_{pi}^2 + \epsilon_i.$$

también es un modelo lineal (aunque los regresores sean términos cuadráticos).

### Interpretación de los coeficientes

$$E[Y|X_1 = x_1, \dots, X_p = x_p] = \sum_{k=1}^p x_k \beta_k$$



$$E[Y|X_1 = x_1 + 1, \dots, X_p = x_p] = (x_1 + 1)\beta_1 + \sum_{k=2}^p x_k \beta_k$$

$$E[Y|X_1 = x_1 + 1, \dots, X_p = x_p] - E[Y|X_1 = x_1, \dots, X_p = x_p] = (x_1 + 1)\beta_1 + \sum_{k=2}^p x_k \beta_k - \sum_{k=1}^p x_k \beta_k = \beta_1$$

Un coeficiente de regresión multivariada es el cambio esperado en el output ante un cambio en una unidad en el regresor correspondiente, manteniendo todos los demás regresores fijos.

## Tasas de hambre en la población infantil

Instancia de trabajo

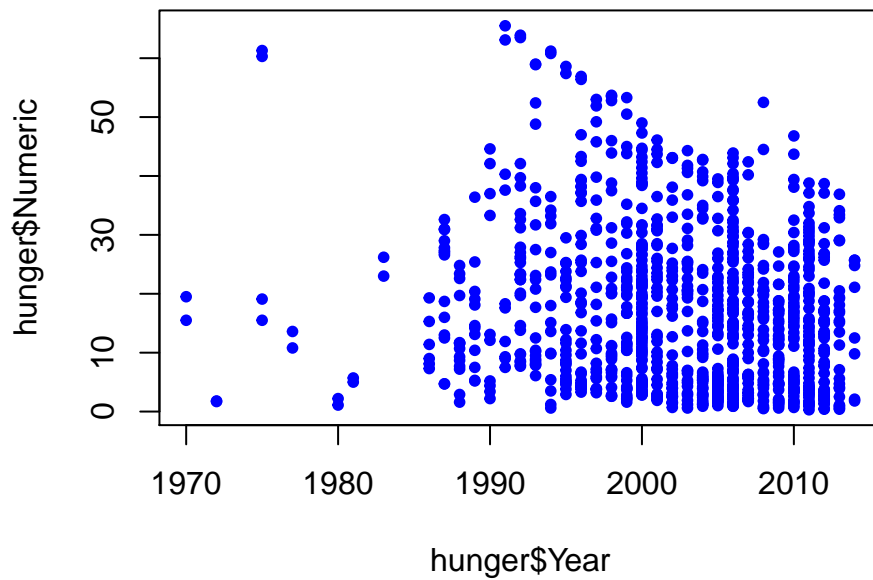
```
#link descarga
url <- "http://apps.who.int/gho/athena/data/GHO/WHOSIS_000008.csv"
file <- "hunger.csv"

if(!file.exists(file)) {
  print("descargando")
  download.file(url, file, method="curl")
}

hunger <- read.csv("hunger.csv")
hunger <- hunger[hunger$Sex!="Both sexes",]
```

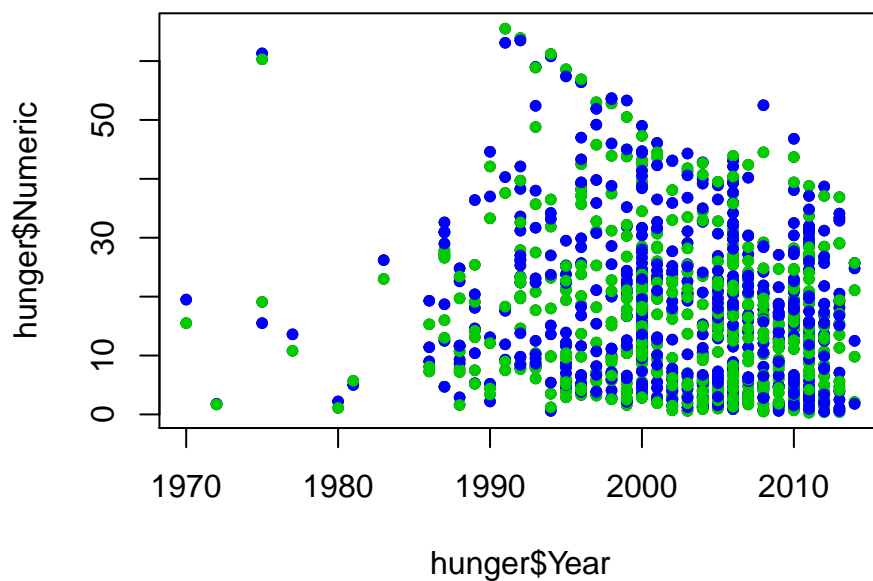
Sin controlar por género:

```
lm1 <- lm(hunger$Numeric ~ hunger$Year)
plot(hunger$Year, hunger$Numeric, pch=20, col="blue")
```



Controlando por género (azul = niñas, verde = niños):

```
plot(hunger$Year,hunger$Numeric,pch=20)
#azul=niñas verde=niños
points(hunger$Year,hunger$Numeric,pch=20,col=((hunger$Sex=="Male")*1+3))
```



## Modelo univariado

Sin controlar por género:

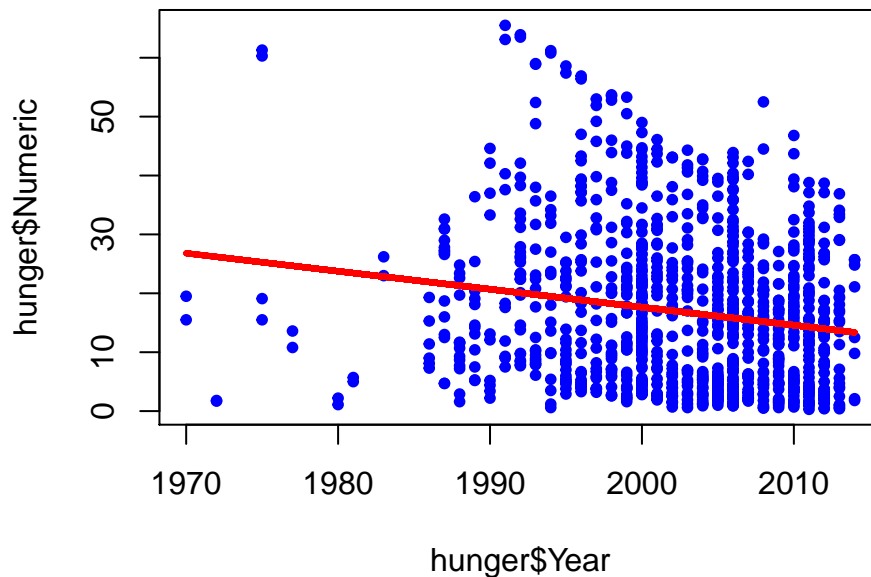
$$Hu_i = b_0 + b_1 Y_i + e_i$$

$b_0$  = % de hambre en el año 0

$b_1$  = disminución del % de hambre por año

$e_i$  = todas las variables no medidas

```
lm1 <- lm(hunger$Numeric ~ hunger$Year)
plot(hunger$Year, hunger$Numeric, pch=20, col="blue")
lines(hunger$Year, lm1$fitted, lwd=3, col="red")
```



Controlando por género:

$$HuF_i = bf_0 + bf_1 YF_i + ef_i$$

$bf_0$  = % de hambre en las niñas en el año 0

$bf_1$  = disminución del % de hambre por año en las niñas

$ef_i$  = todas las variables no medidas

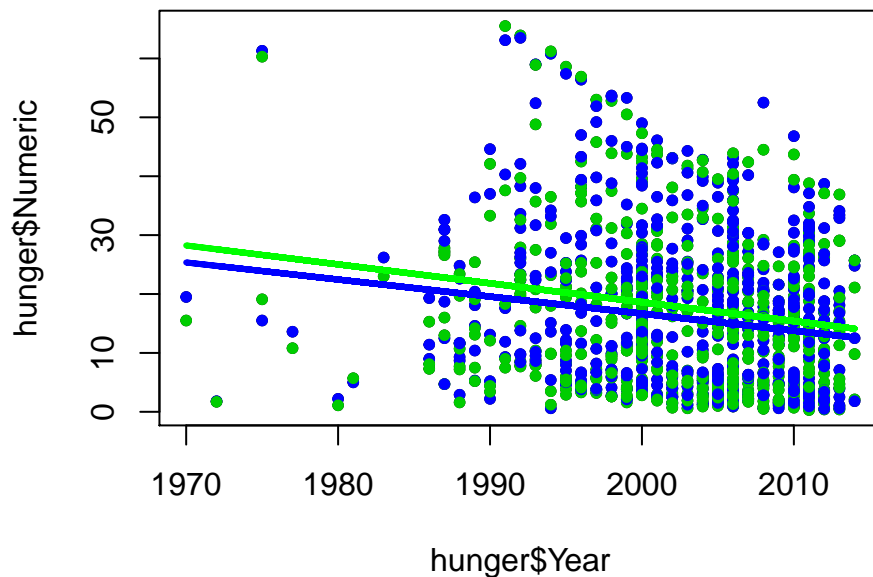
$$HuM_i = bm_0 + bm_1 YM_i + em_i$$

$bm_0$  = % de hambre en los niños en el año 0

$bm_1$  = disminución del % de hambre por año en los niños

$em_i$  = todas las variables no medidas

```
lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"])
lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex=="Female"])
plot(hunger$Year, hunger$Numeric, pch=20)
points(hunger$Year, hunger$Numeric, pch=20, col=((hunger$Sex=="Male")*1+3))
lines(hunger$Year[hunger$Sex=="Male"], lmM$fitted, col="green", lwd=3)
lines(hunger$Year[hunger$Sex=="Female"], lmF$fitted, col="blue", lwd=3)
```



## Modelo multivariado

Las dos rectas anteriores tienen la misma pendiente. Vamos a estimar el siguiente modelo:

$$Hu_i = b_0 + b_1 M_i + b_2 Y_i + e_i^*$$

$b_0$  = % de hambre en las niñas en el año 0

$$M_i = \begin{cases} 1 & \text{si es niño} \\ 0 & \text{si es niña} \end{cases}$$

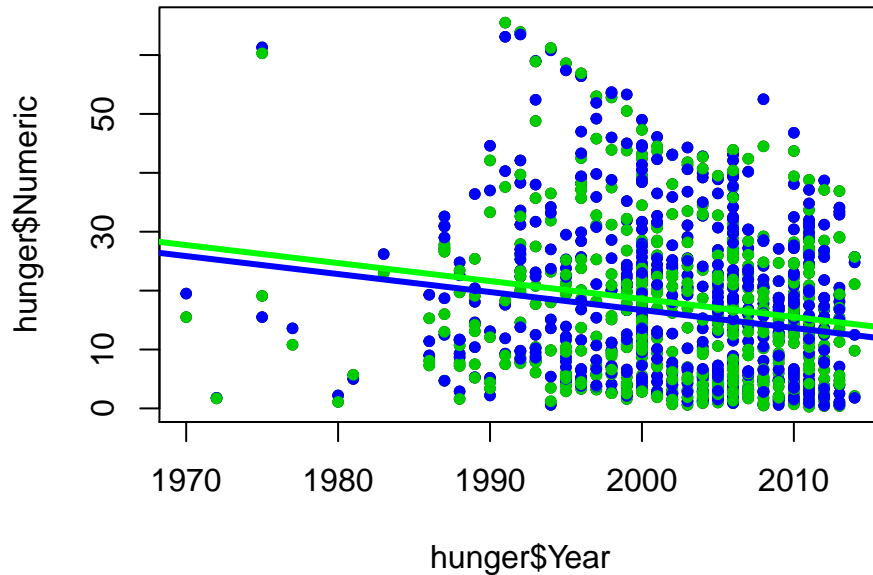
$b_0 + b_1$  = % de hambre en las niños en el año 0

$b_2$  = disminución del % de hambre por año en niños o niñas

$e_i^*$  = todas las variables no medidas

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex)
plot(hunger$Year, hunger$Numeric, pch=20)
points(hunger$Year, hunger$Numeric, pch=20, col=((hunger$Sex=="Male")*1+3))
```

```
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="blue",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] ),col="green",lwd=3)
```



$$Hu_i = b_0 + b_1 M_i + b_2 Y_i + b_3 (M_i \cdot Y_i) + e_i^+$$

$b_0$  = % de hambre en las niñas en el año 0

$$M_i = \begin{cases} 1 & \text{si es niño} \\ 0 & \text{si es niña} \end{cases}$$

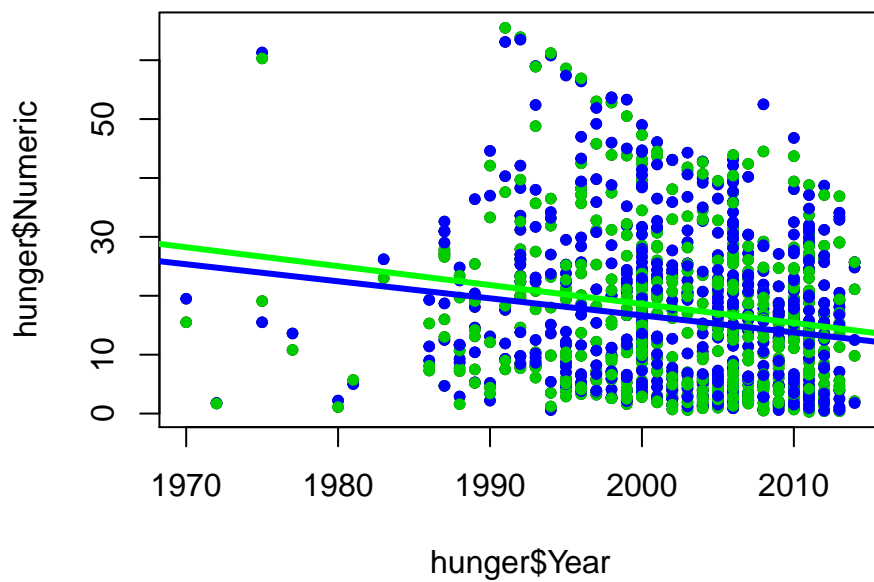
$b_0 + b_1$  = % de hambre en las niños en el año 0

$b_2$  = disminución del % de hambre por año en niños o niñas

$b_2 + b_3$  = disminución del % de hambre por año en los niños

$e_i^+$  = todas las variables no medidas

```
lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Sex*hunger$Year)
plot(hunger$Year,hunger$Numeric,pch=20)
points(hunger$Year,hunger$Numeric,pch=20,col=((hunger$Sex=="Male")*1+3))
abline(c(lmBoth$coeff[1],lmBoth$coeff[2]),col="blue",lwd=3)
abline(c(lmBoth$coeff[1] + lmBoth$coeff[3],lmBoth$coeff[2] +lmBoth$coeff[4]),col="green",lwd=3)
```



## Resultados

```
coefficients(lmBoth)
```

```
##          (Intercept)          hunger$Year
##          595.83543620          -0.28958348
## hunger$SexMale hunger$Year:hunger$SexMale
##          64.74249171          -0.03139868
```