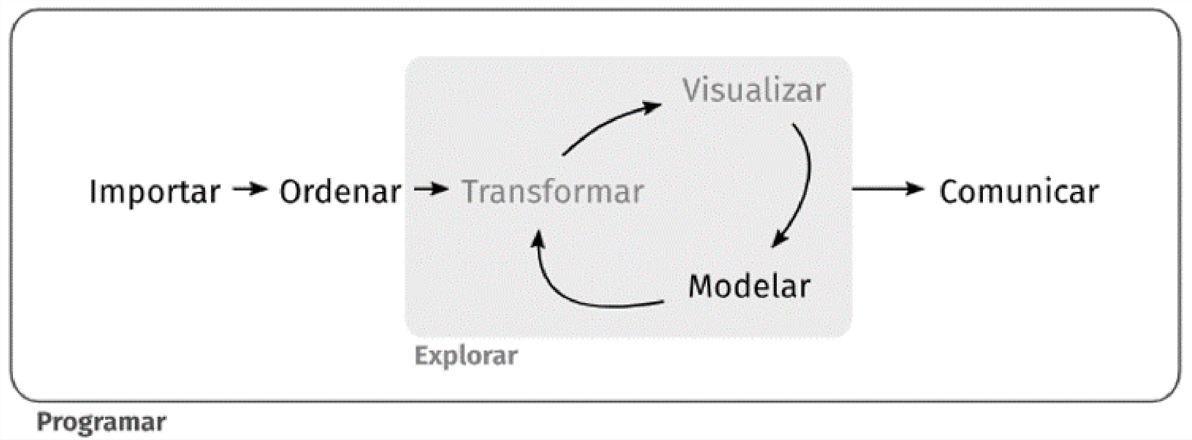
Ciencia de Datos para Políticas Públicas

Módulo 2 - Clase 5: Regresiones

Pablo Aguirre Hormann
13/07/2021

¿Qué veremos hoy?

- Visualización de datos
- Manejo de datos
- Transformación de datos
- **Inferencia Estadística/Econometría**
  - Regresiones lineales (simple y múltiple)
  - ¿Qué son? ¿Cómo se "construyen?
  - Inferencia usando regresiones



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Distribuciones conjuntas

Medir la relación entre variables

Covarianza

Medición de la variación/dependencia conjunta entre dos aleatorias.

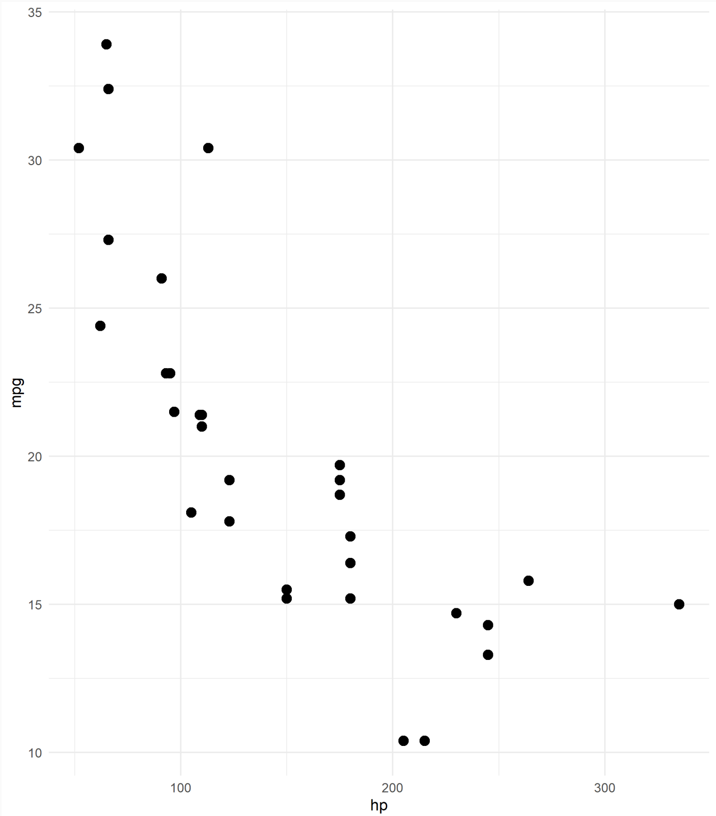
$$Cov(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}$$
$$= E(XY) - E(X)E(Y)$$

Correlación

Normalización de la covarianza. Mismo signo que $Cov(X, Y)$ pero adimensional (entre -1 y 1)

$$\rho_{x,y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$
$$= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Relaciones entre variables




```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6  ...
##  $ disp: num  160 160 108 258 360  ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123  ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44  ...
##  $ qsec: num  16.5 17 18.6 19.4 17  ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1  ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0  ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4  ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4  ...
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Relaciones entre variables

Covarianza caballos de fuerza (`hp`)/millas por galón (`mpg`)

```
mtcars %>%
  summarise(covarianza = cov(hp, mpg))
```

```
##   covarianza
## 1  -320.7321
```

Correlación caballos de fuerza (`hp`)/millas por galón (`mpg`)

```
mtcars %>%
  summarise(correlacion = cor(hp, mpg))
```

```
##   correlacion
## 1  -0.7761684
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123  ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17  ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Distribuciones conjuntas

Si X e Y son dos variables aleatorias, la **distribución conjunta** de X e Y permite calcular las probabilidades de eventos que involucren a ambas variables.

Por ejemplo, la probabilidad de que alguien mida entre 1.7 y 1.8 metros y que pese entre 60 y 80 kilogramos.

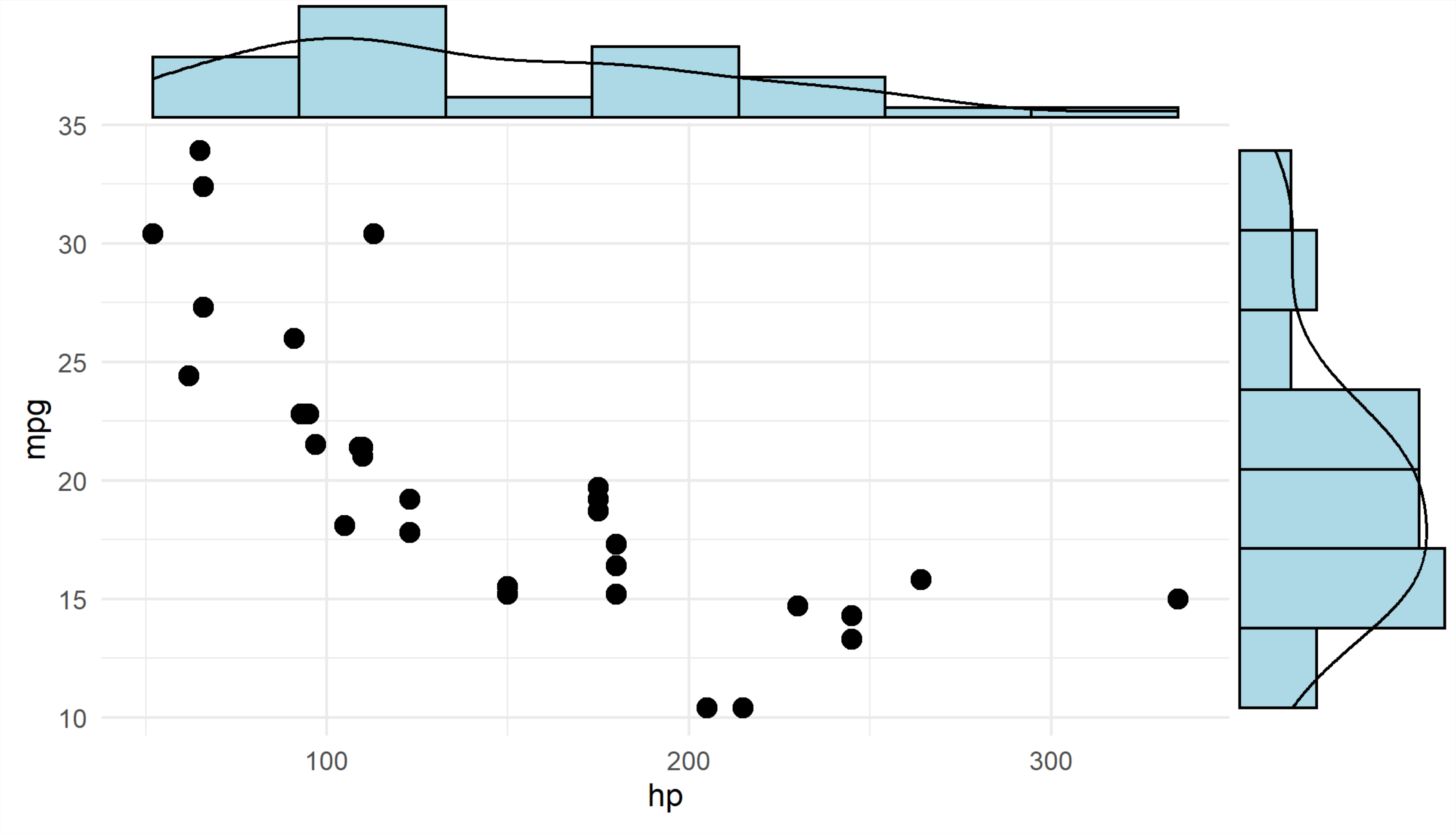
Desde una distribución conjunta podemos obtener **distribuciones marginales** y **distribuciones condicionales**.

Esperanzas condicionales

Si X e Y no son independientes, entonces saber algo de X me puede ayudar a predecir/explicar Y.

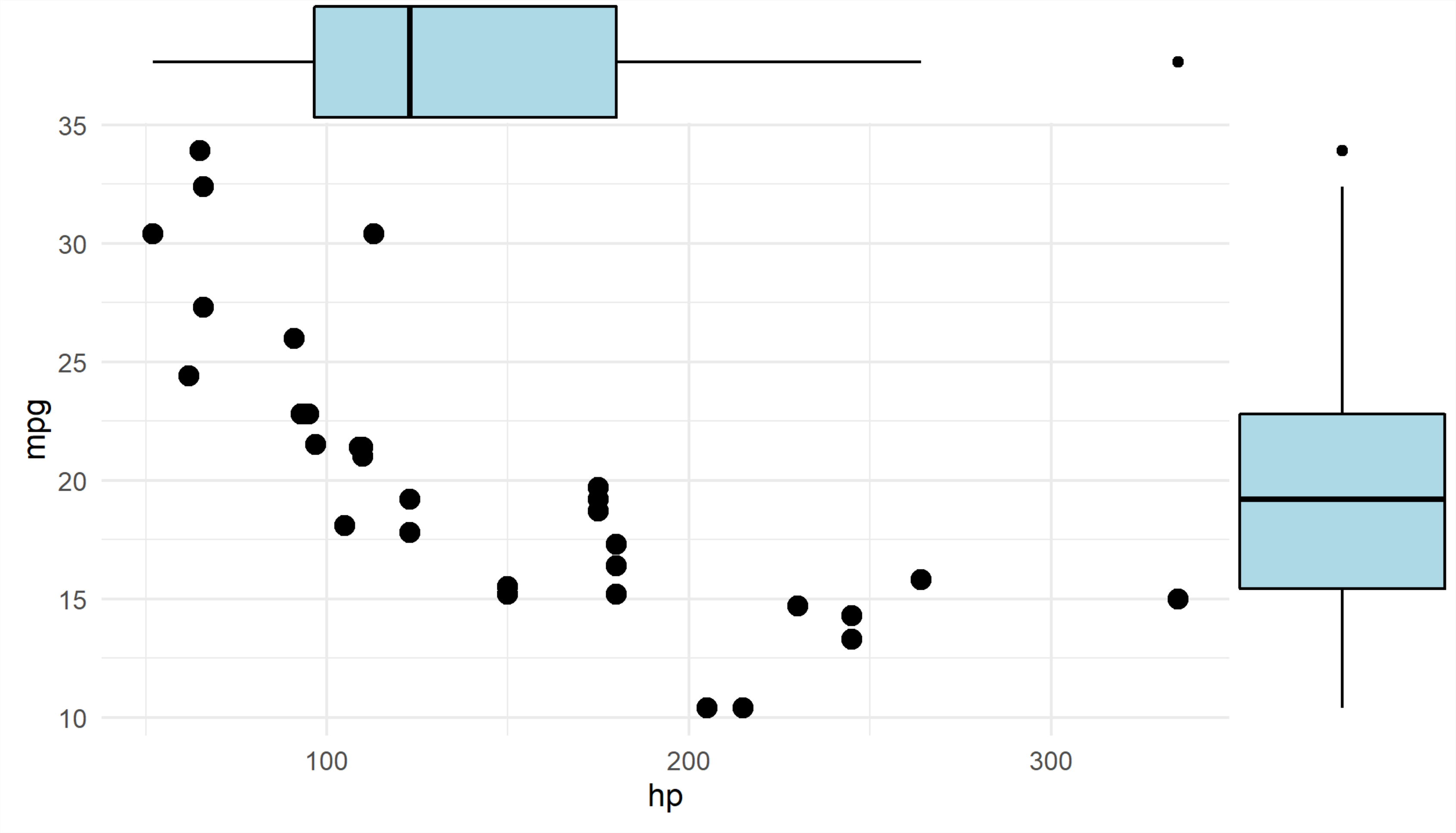
$E(Y|X)$ es una **función** que me dice para cada valor de X, la esperanza de Y de aquellos individuos con ese valor de X.

Distribuciones marginales



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

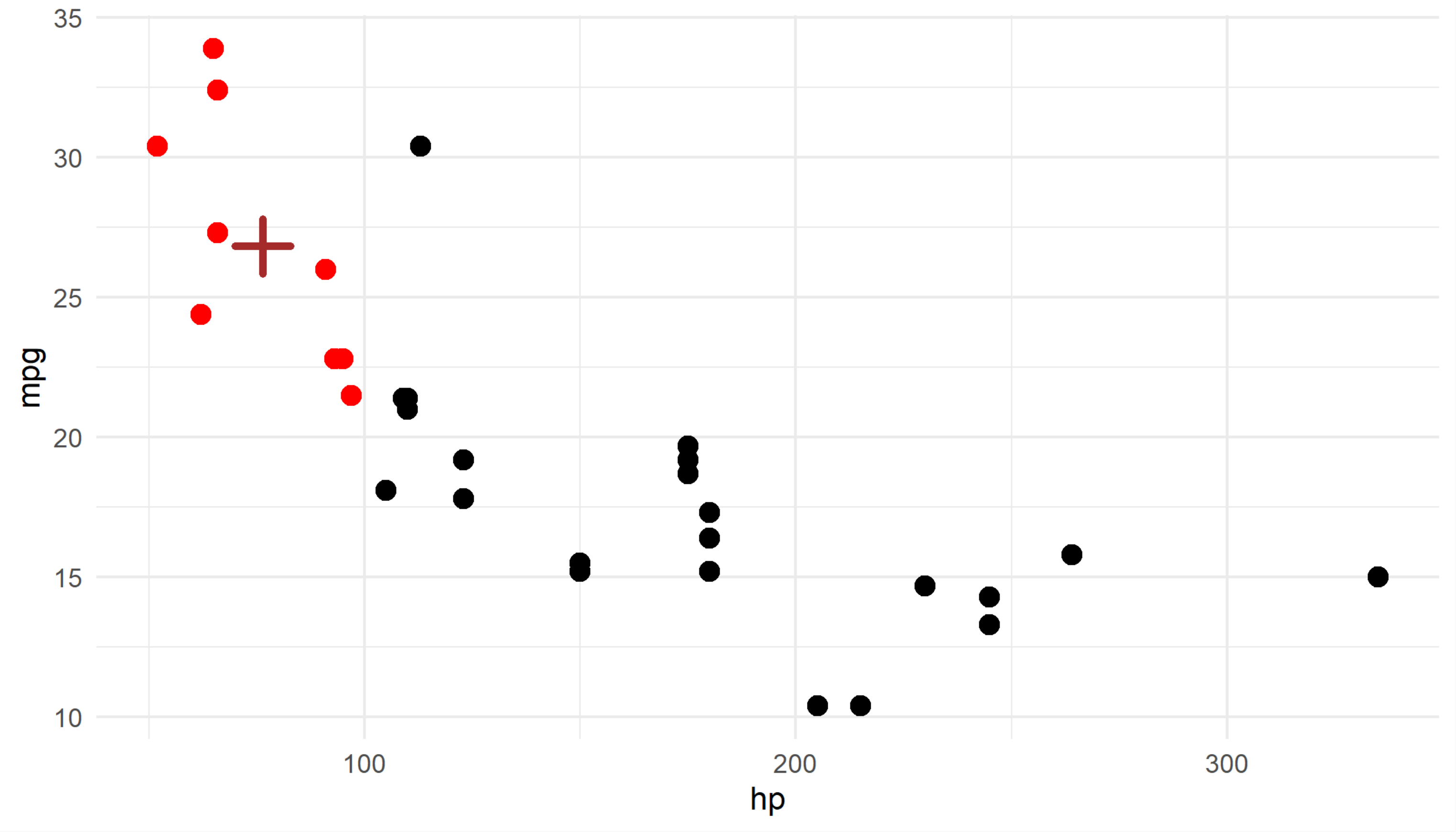
Distribuciones marginales



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

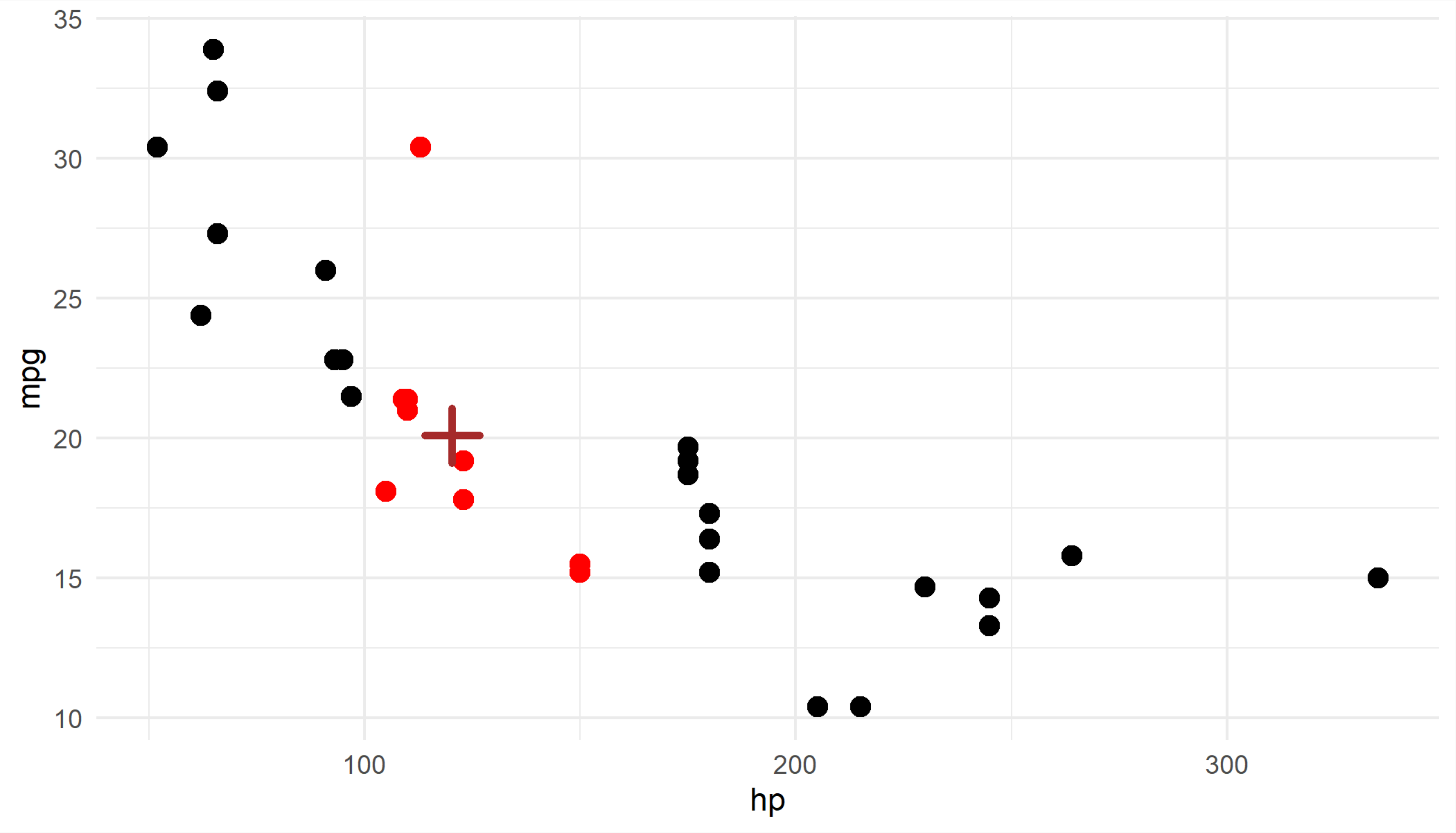
Esperanzas condicionales

$E(mpg \mid hp < 100)$



Esperanzas condicionales

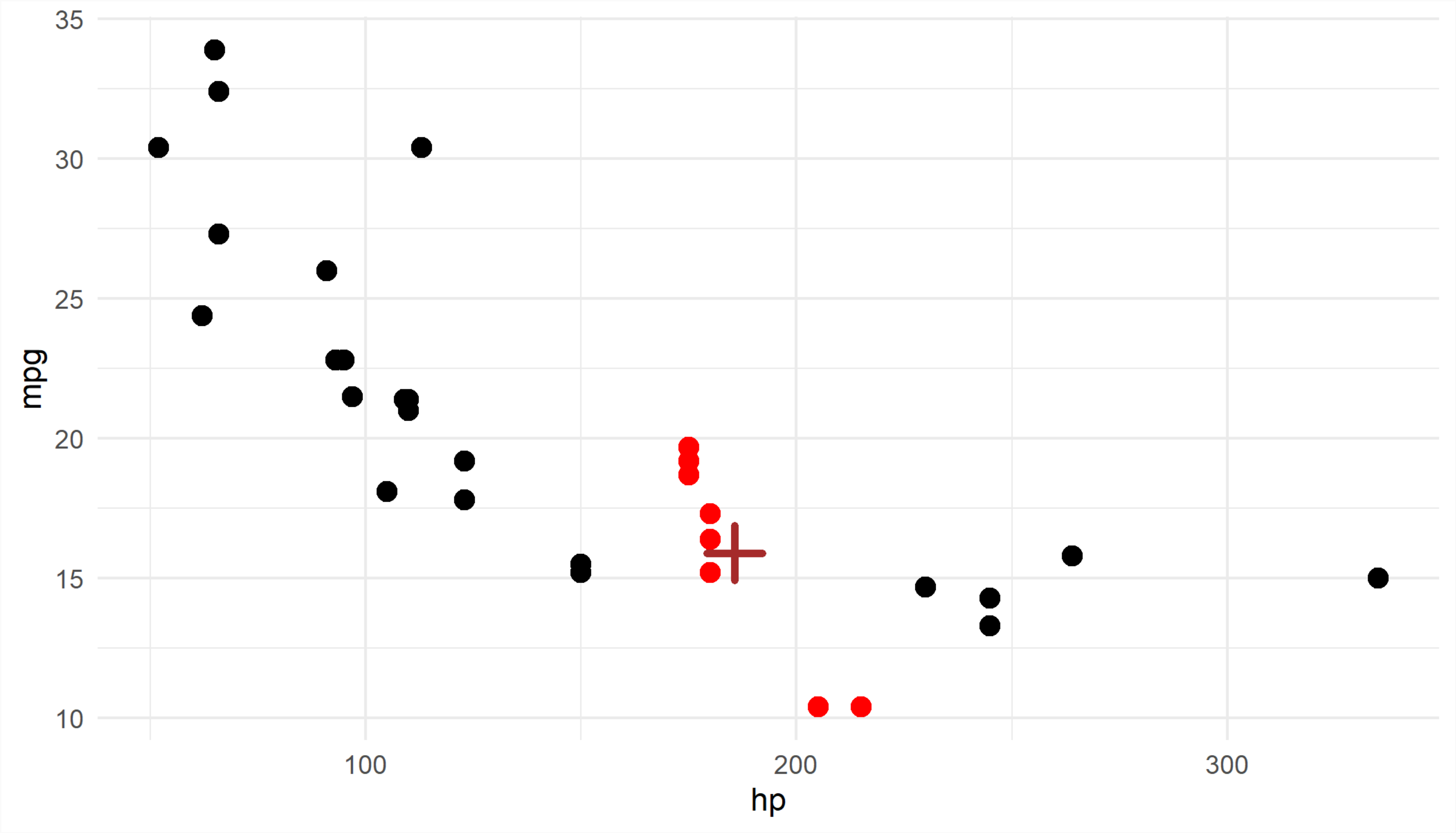
$$E(mpg \mid 100 < hp < 160)$$



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Esperanzas condicionales

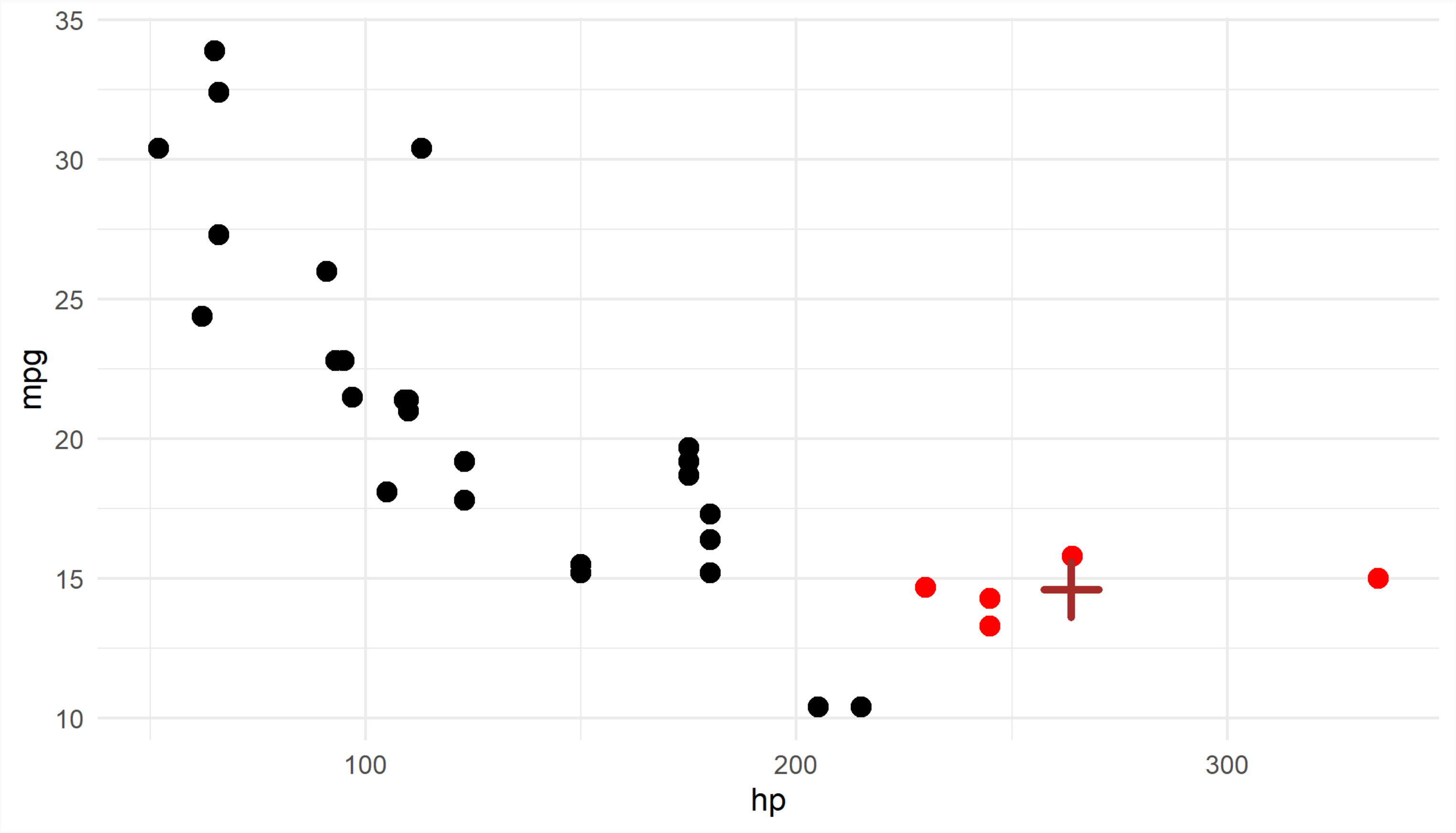
$$E(mpg \mid 160 < hp < 230)$$



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

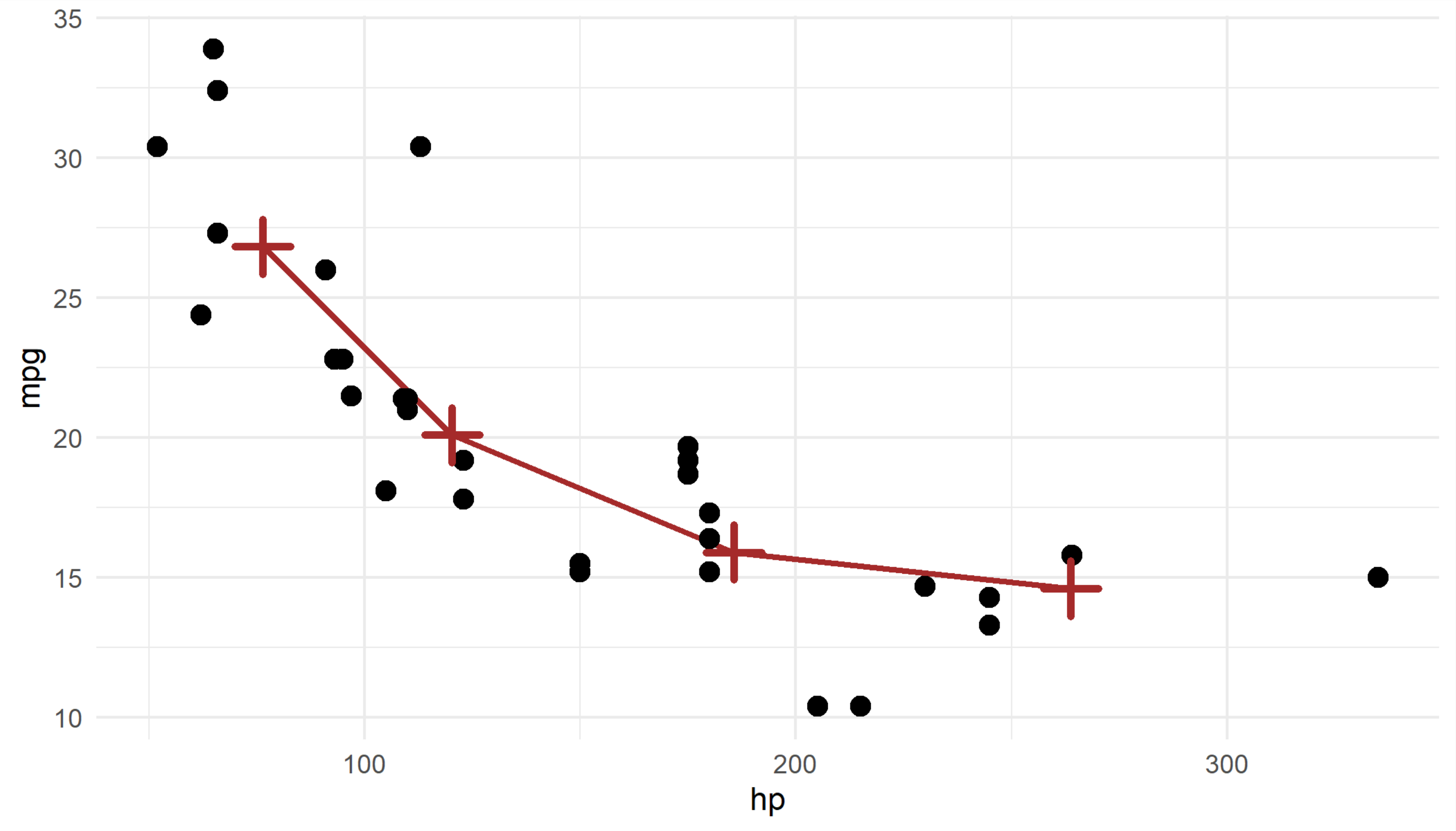
Esperanzas condicionales

$E(mpg \,|\, hp > 230)$



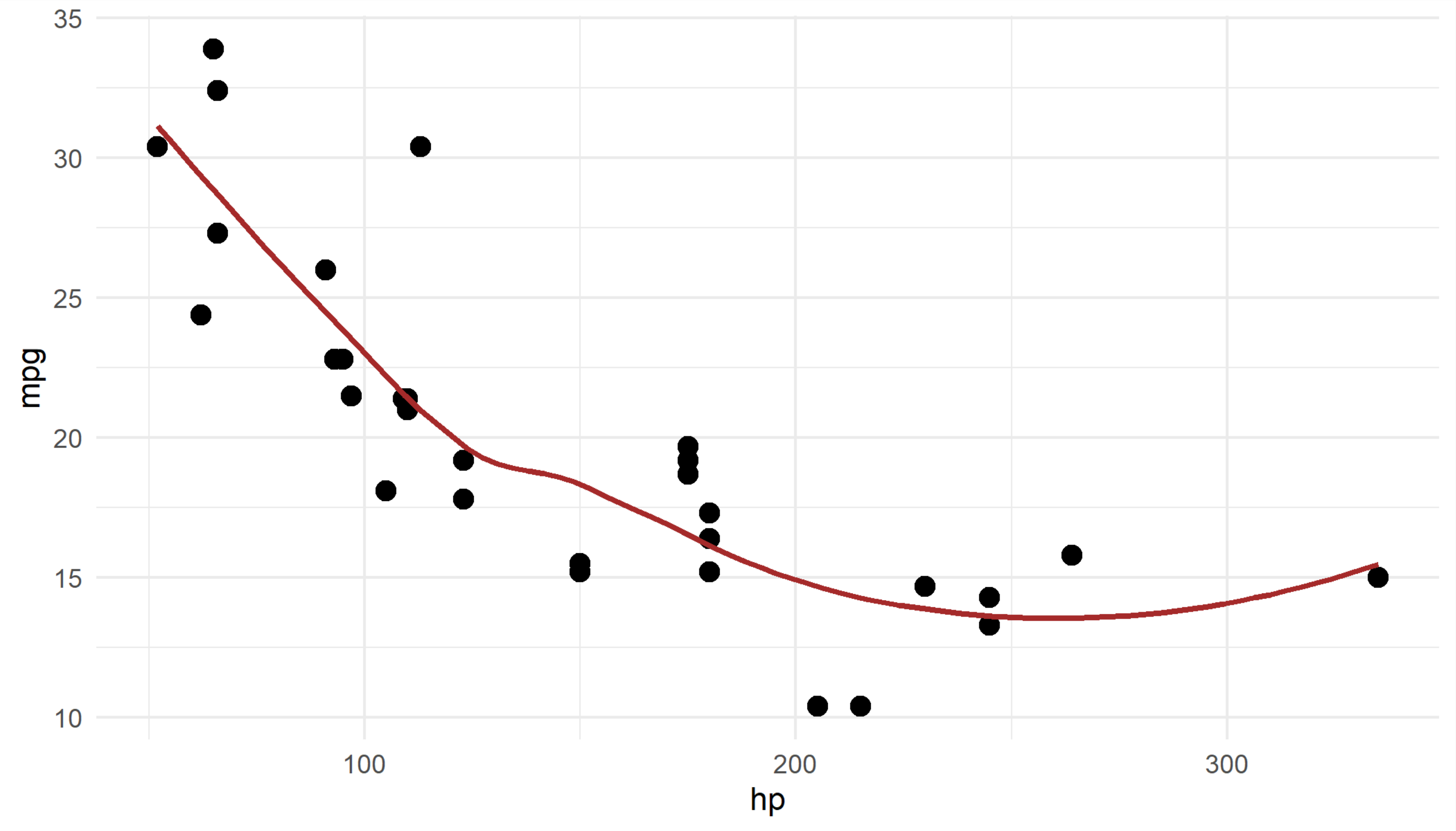
Esperanzas condicionales

$$E(mpg\,|\,hp) = \hat{f}$$



Esperanzas condicionales

ggplot2: `geom_smooth()`



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelos

Modelos

Objetivo: representar la relación entre una variable dependiente Y y una o varias variables explicativas/independientes $X_1, X_2, \ldots, X_k$.

$$\begin{aligned}\hat{Y} &= E(Y|X) \\ &= \hat{f}(X)\end{aligned}$$

- Si Y es una variable *continua*: **regresión**
- Si Y es una variable *categórica*: **clasificación** (próxima clase)

"Todos los modelos están mal... pero algunos son útiles"

Inferencia vs Predicción

Modelos para Inferencia/Explicación:

- Aprender y concluir algo sobre como se relacionan variables. Relaciones causales.
- Evitar sesgo
- Predicción *dentro de muestra*
- $\hat{f}$ / $\hat{\beta}$

Modelos para Predicción:

- Que la predicción esté lo más cerca posible del valor real
- Evitar sobreajuste al entrenar modelos
- Predicción *fuera de muestra*
- $\hat{Y}$

Algunos algorítmos pueden servir para ambos objetivos pero con diferencias en la implementación (*ej. Regresión lineal para inferencia o para predicción*).

Nosotros no hablaremos de predicción (*Machine Learning*). Eso lo verán en el módulo 3.

Regresión Lineal

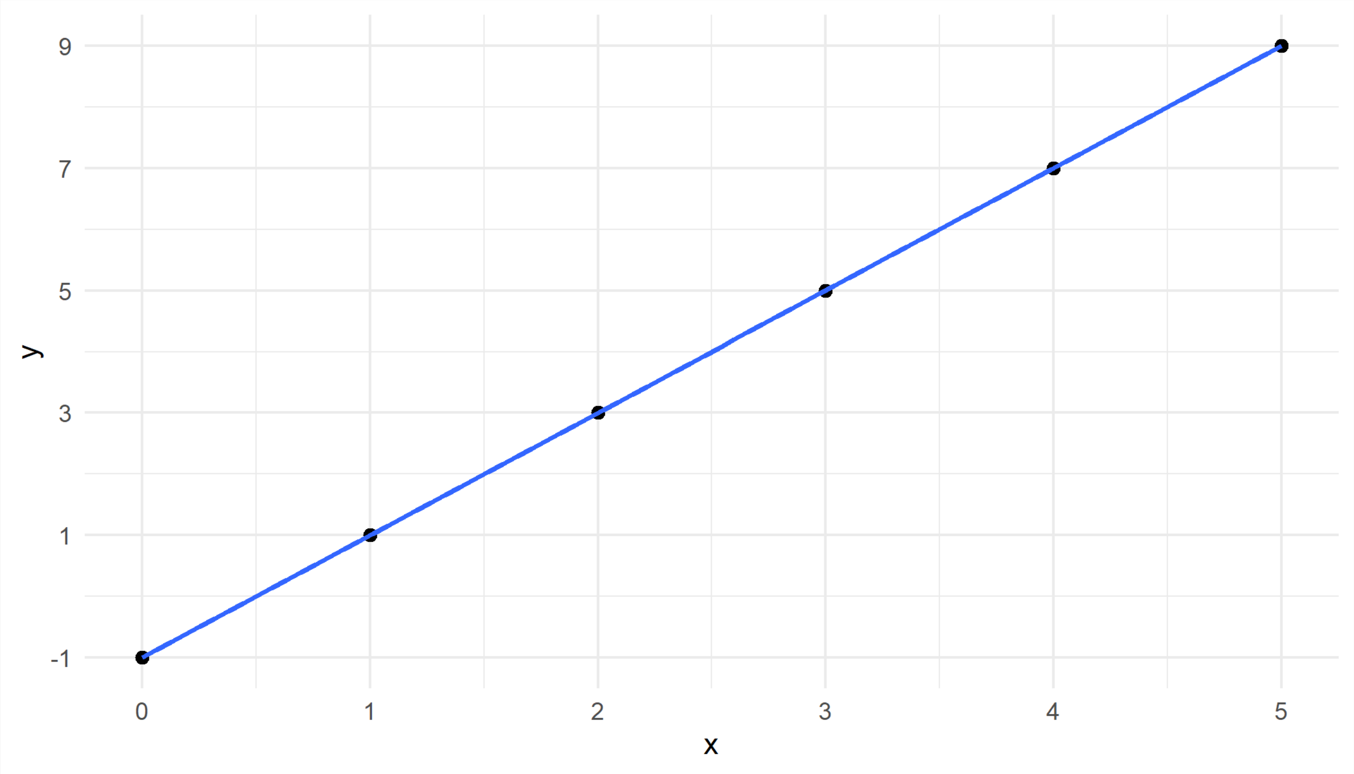
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Ecuación de la curva

Probablemente recordarán de alguna clase de matemáticas:

$$y = mx + b$$

donde m es la pendiente y b es el intercepto en el eje **y**. Por ejemplo:



$$y = 2x - 1$$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Ec. de la curva vs Regresión

Clase de **matemáticas**:

$y = b + mx$

b es el intercepto en el eje

m es la pendiente

Clase de **estadística**:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + u$

$\hat{\beta}_0$ es el intercepto en el eje

$\hat{\beta}_1$ es la pendiente

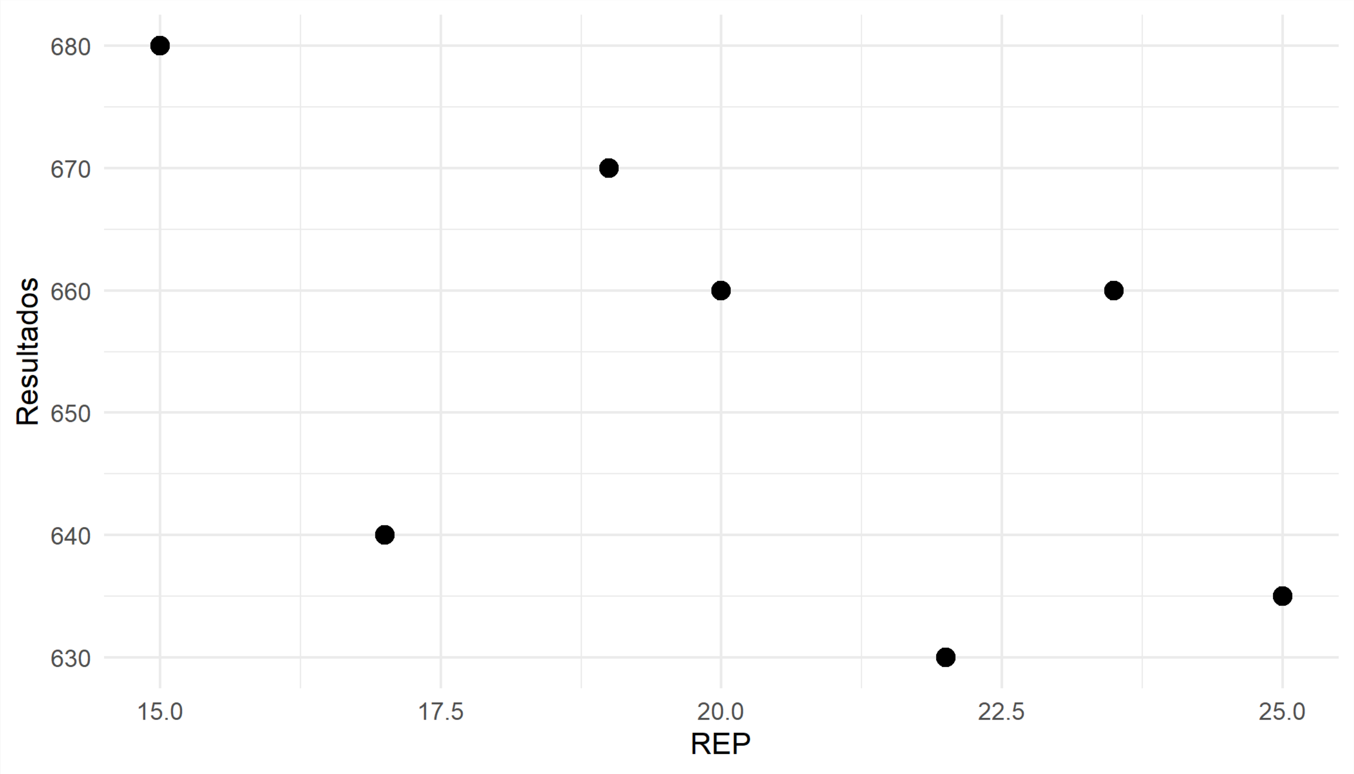
u es el error/residual

- $\hat{\beta}_0$ y $\hat{\beta}_1$ se conocen también como coeficientes de regresión y serán **parámetros a estimar**.
- $\hat{Y}$ lo denominamos como los **valores ajustados** o bien los valores estimados por nuestra función para cada valor posible de X.

Estudiantes/Profesor vs Resultados

```r
datos_colegio <- data.frame(REP = c(15, 17, 19, 20, 22, 23.5, 25),
                            Resultados = c(680, 640, 670, 660, 630, 660, 635))

ggplot(datos_colegio, aes(REP, Resultados)) +
  geom_point(size = 3) +
  theme_minimal()
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

El modelo más simple

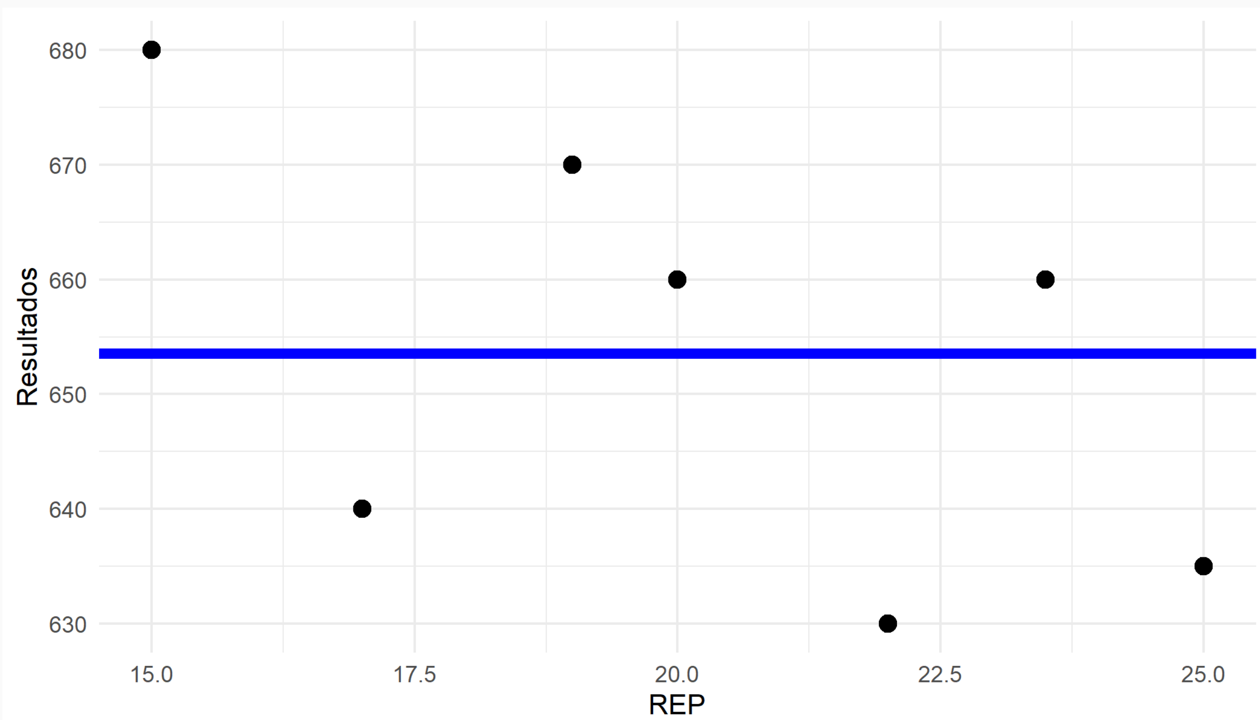
$$\hat{Y}_i = E(Y)$$

$$\hat{Y}_i = \bar{Y}$$

El modelo más simple

$$\hat{Resultados}_i = E(Resultados)$$
$$= \bar{Resultado} = 653.57$$

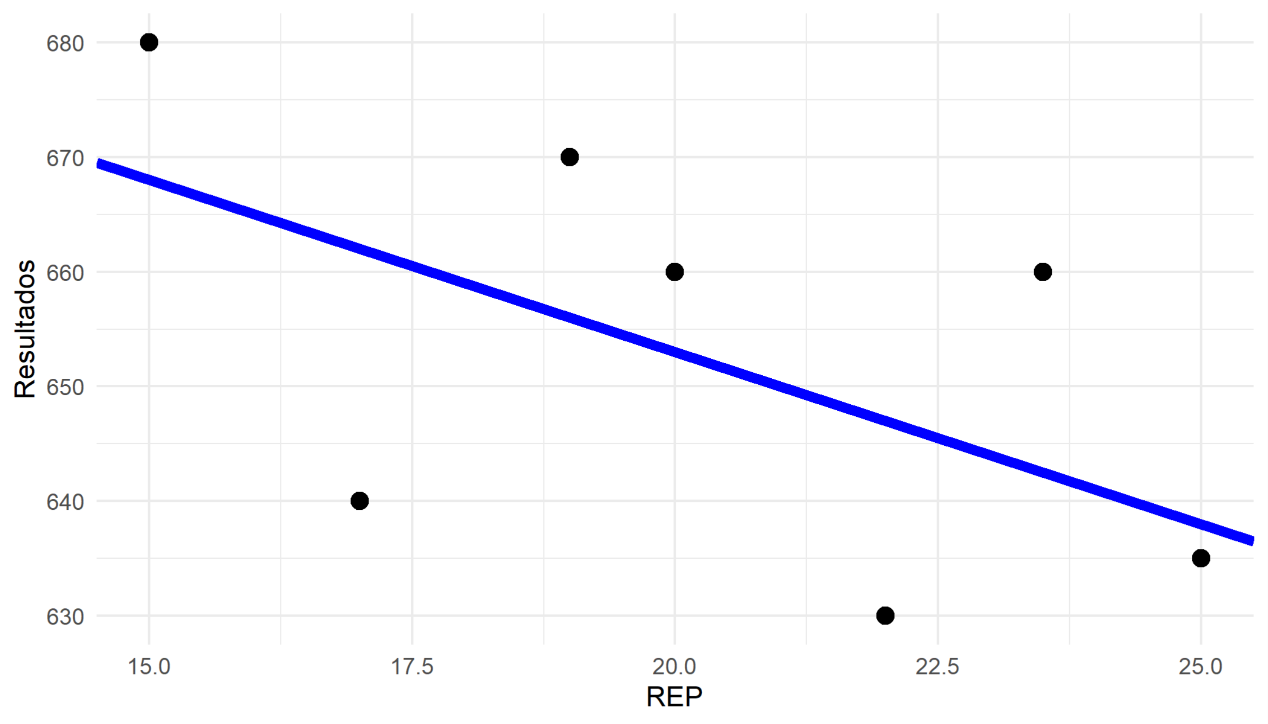
```
ggplot(datos_colegio, aes(REP, Resultados)) +
  geom_point(size = 3) +
  geom_abline(aes(intercept = mean(Resultados), slope = 0),
              size = 2, col = "blue") +
  theme_minimal()
```



Una línea que describe esta relación

$$\widehat{Resultados}_i = f(REP)$$
$$= 713 - (3 * REP_i)$$

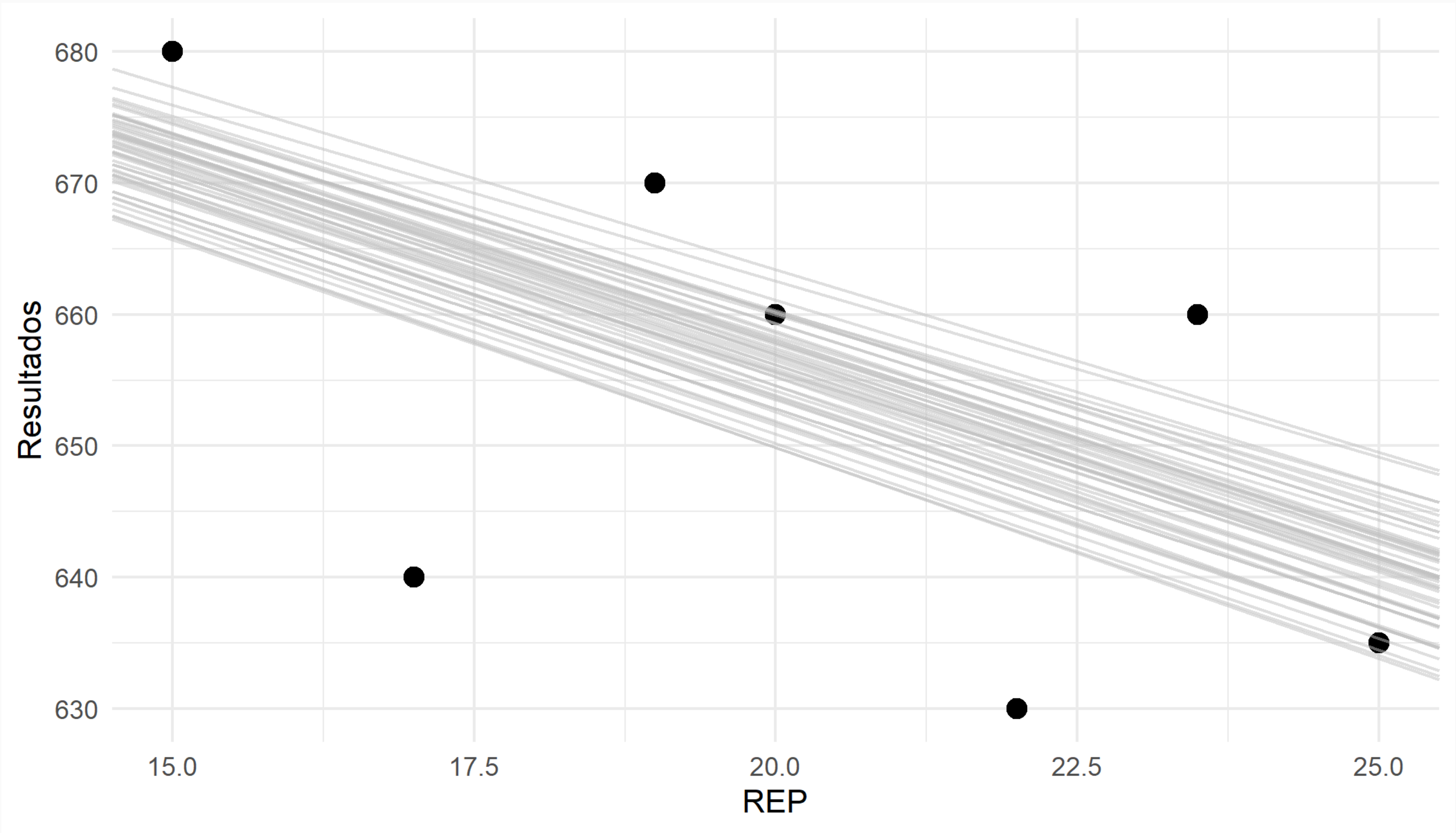
```
ggplot(datos_colegio, aes(REP, Resultados)) +
  geom_point(size = 3) +
  geom_abline(aes(intercept = 713, slope = -3),
              size = 2, col = "blue") +
  theme_minimal()
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Ahora muchas (50) líneas

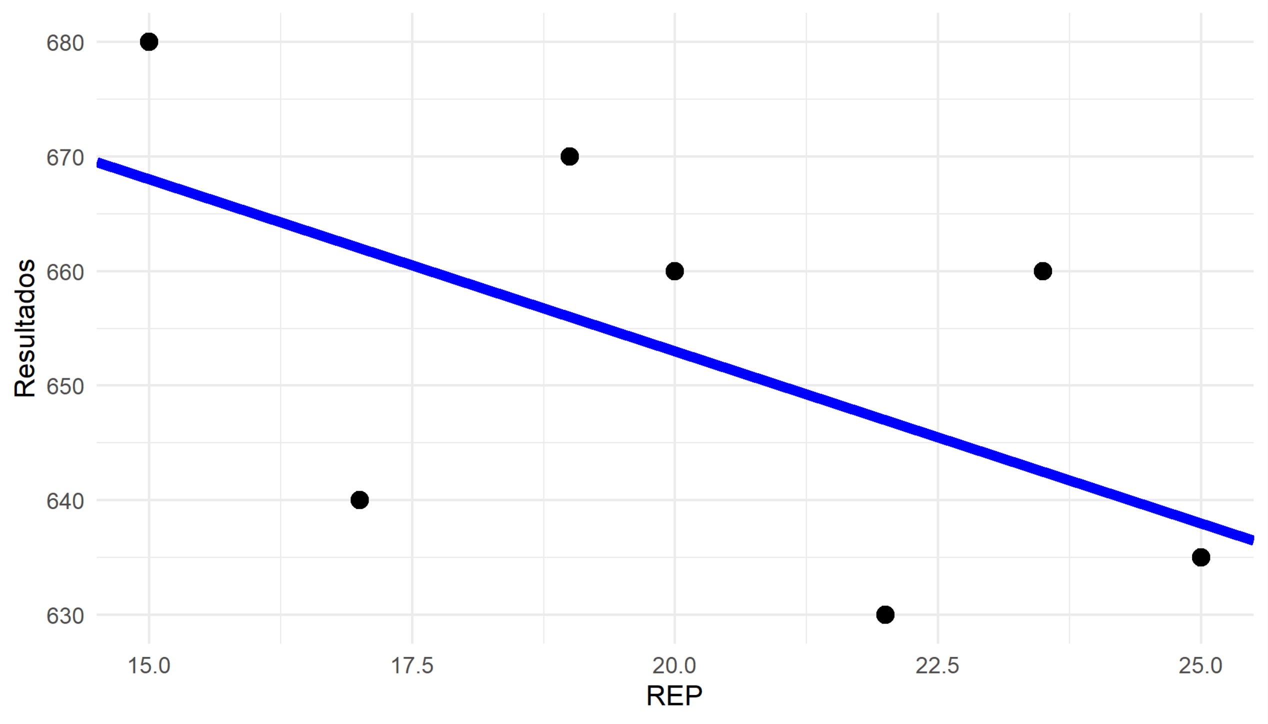
$$\widehat{Resultados}_i = ¿? - (¿? * REP_i)$$



¿Cómo podemos elegir una de estas (u otra)?

Residuales

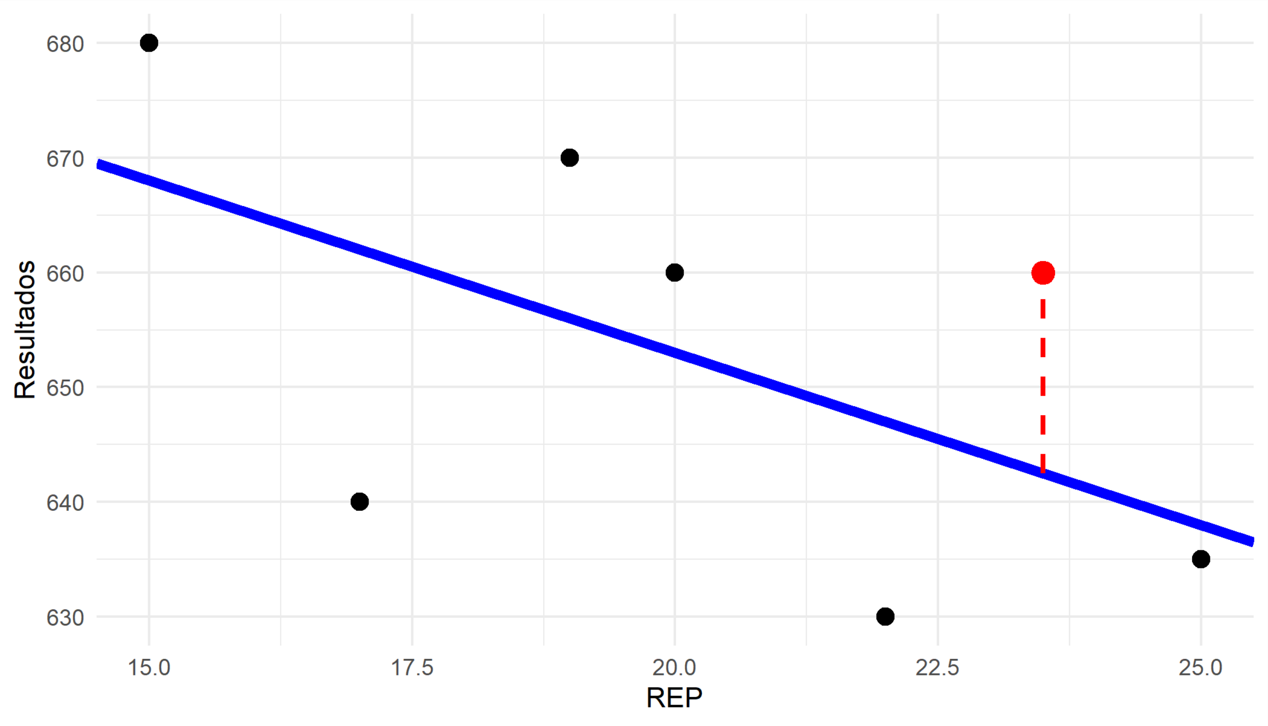
$$\widehat{Resultados}_i = 713 - (3 * REP_i)$$



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Residuales

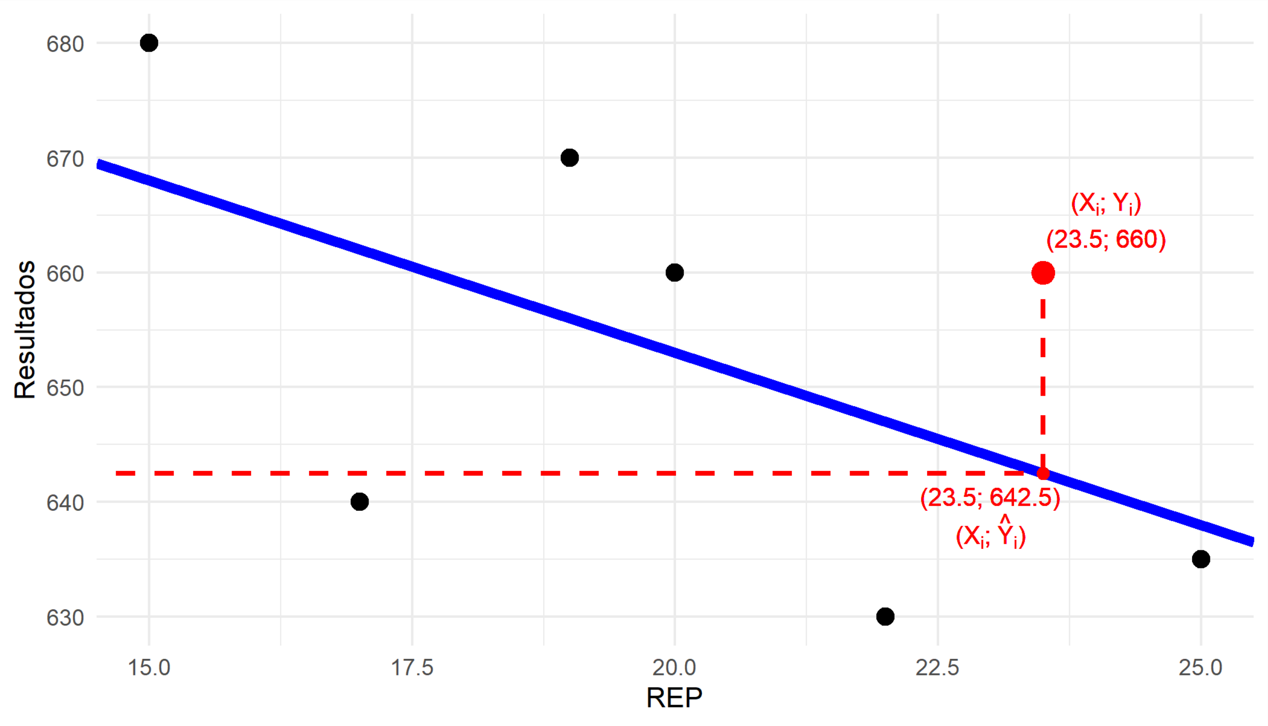
$$\widehat{Resultados}_i = 713 - (3 * REP_i)$$



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Residuales

$$\widehat{Resultados}_i = 713 - (3 * REP_i)$$



$$X_6 = 23.5; \;\; Y_6 = 660$$

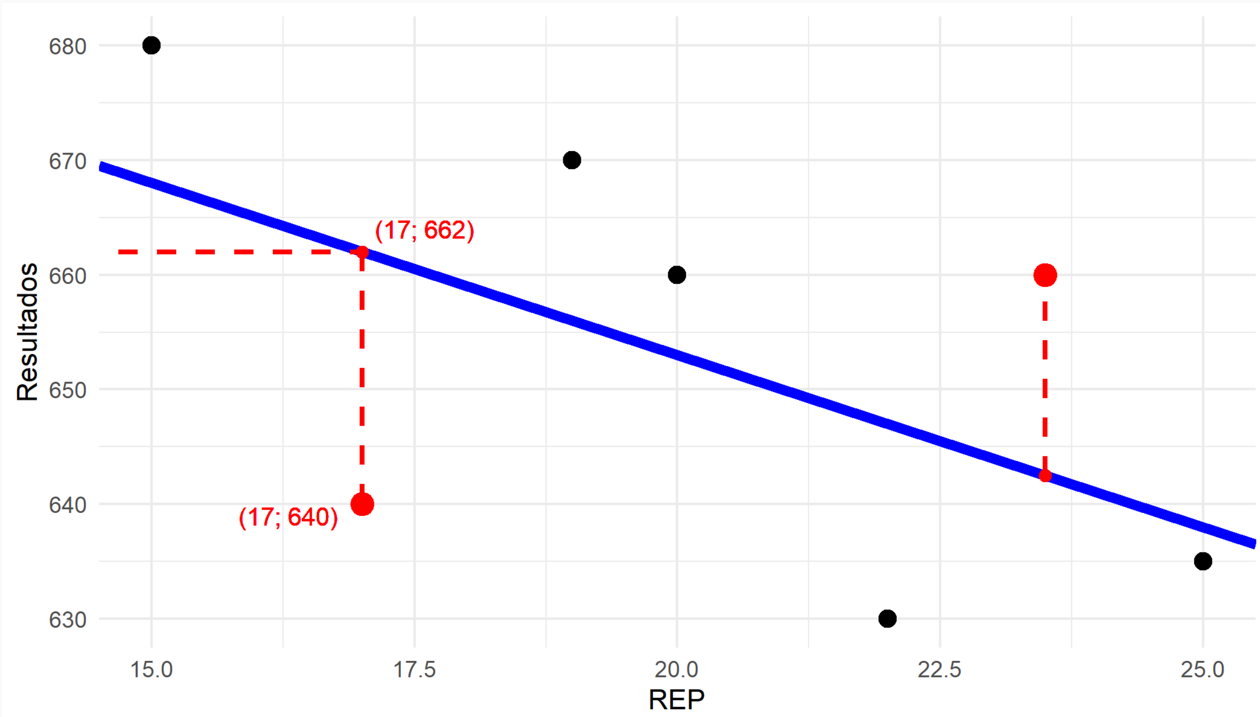
$$\hat{Y}_6 = 713 - (3 * 23.5) = 642.5$$

$$u_6 = (660 - 642.5) = 17.5$$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Residuales

$$\widehat{Resultados}_i = 713 - (3 * REP_i)$$



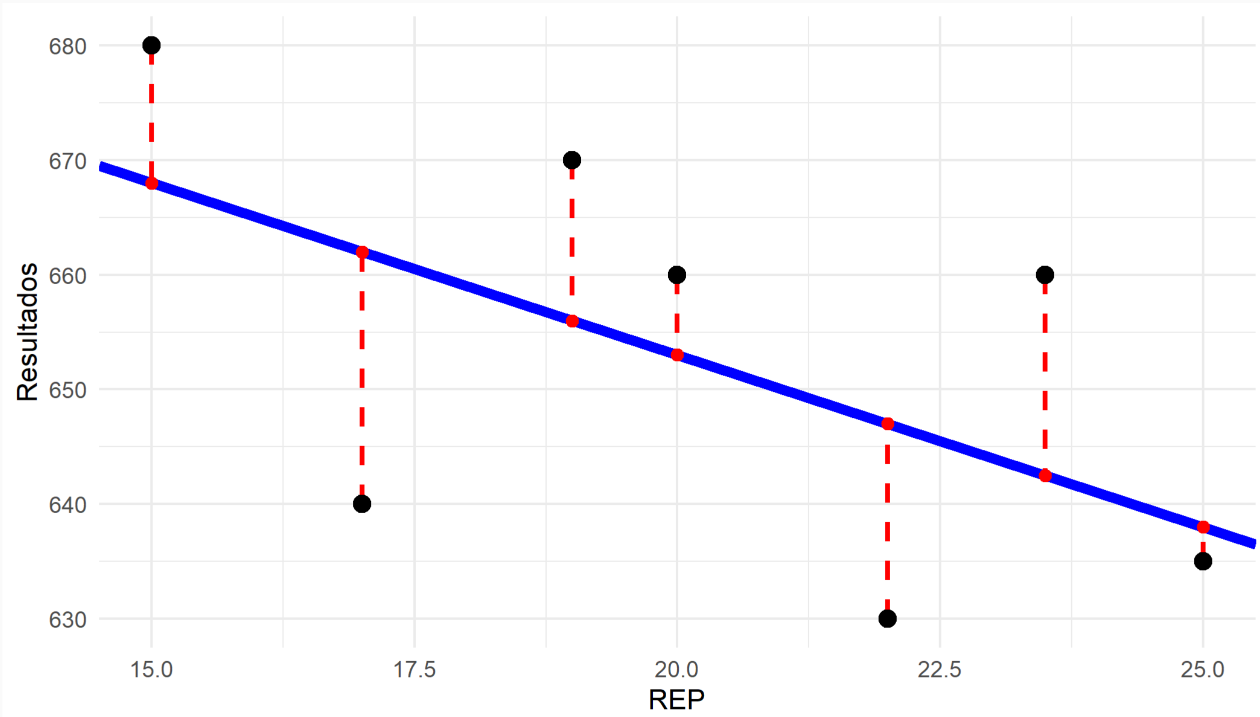
$$X_2 = 17; \ Y_6 = 640 \qquad X_6 = 23.5; \ Y_6 = 660$$

$$\hat{Y}_2 = 713 - (3 * 17) = 662 \qquad \hat{Y}_6 = 713 - (3 * 23.5) = 642.5$$

$$u_2 = (660 - 642.5) = -22 \qquad u_6 = (660 - 642.5) = 17.5$$

Suma de Cuadrados Residuales

$$\widehat{Resultados}_i = 713 - (3 * REP_i)$$

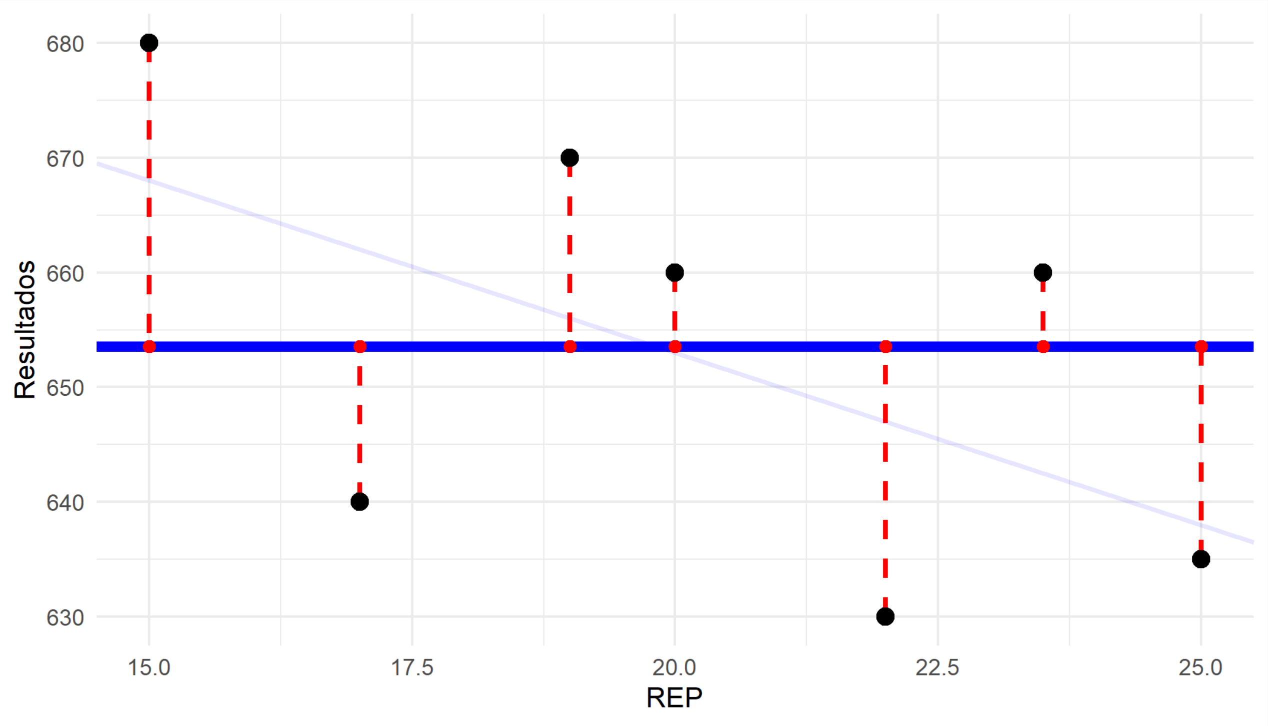


$$u_1 = 12;\ u_2 = -22;\ u_3 = 14;\ u_4 = 7;\ u_5 = -17;\ u_6 = 17.5;\ u_7 = -3$$

Suma de cuadrados residuales = $\sum_i^7(u_i^2) = \sum_i^7((Y_i - \hat{Y}_i)^2) = 1477.25$

Suma de Cuadrados Residuales

$$\widehat{Resultados}_i = E(Resultados) = 653.57$$

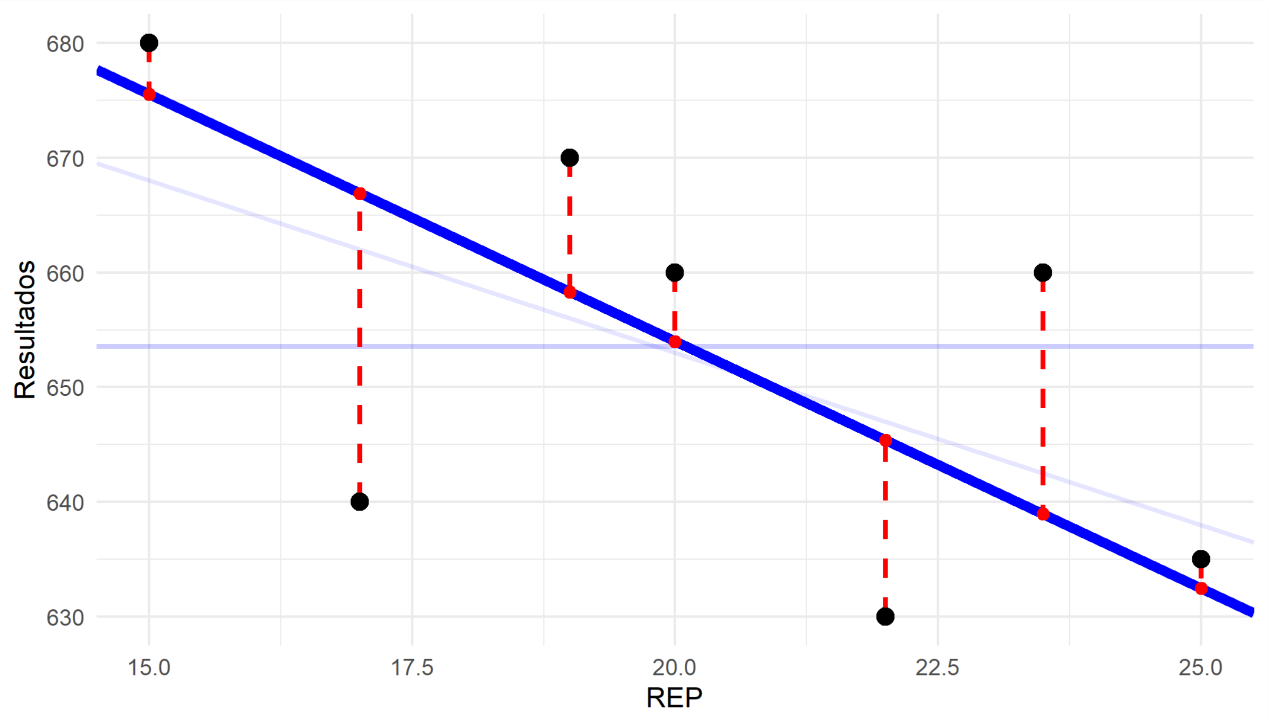


$$u_1 = 26.4;\ u_2 = -13.6;\ u_3 = 16.4;\ u_4 = 6.4;\ u_5 = -23.6;\ u_6 = 6.4;\ u_7 = -18.6$$

Suma de cuadrados residuales = $\sum_i^7 (u_i^2) = \sum_i^7 ((Y_i - \hat{Y}_i)^2) = 2135.71$

Suma de Cuadrados Residuales

$$\widehat{Resultados_i} = 740 - (4.3 * REP_i)$$



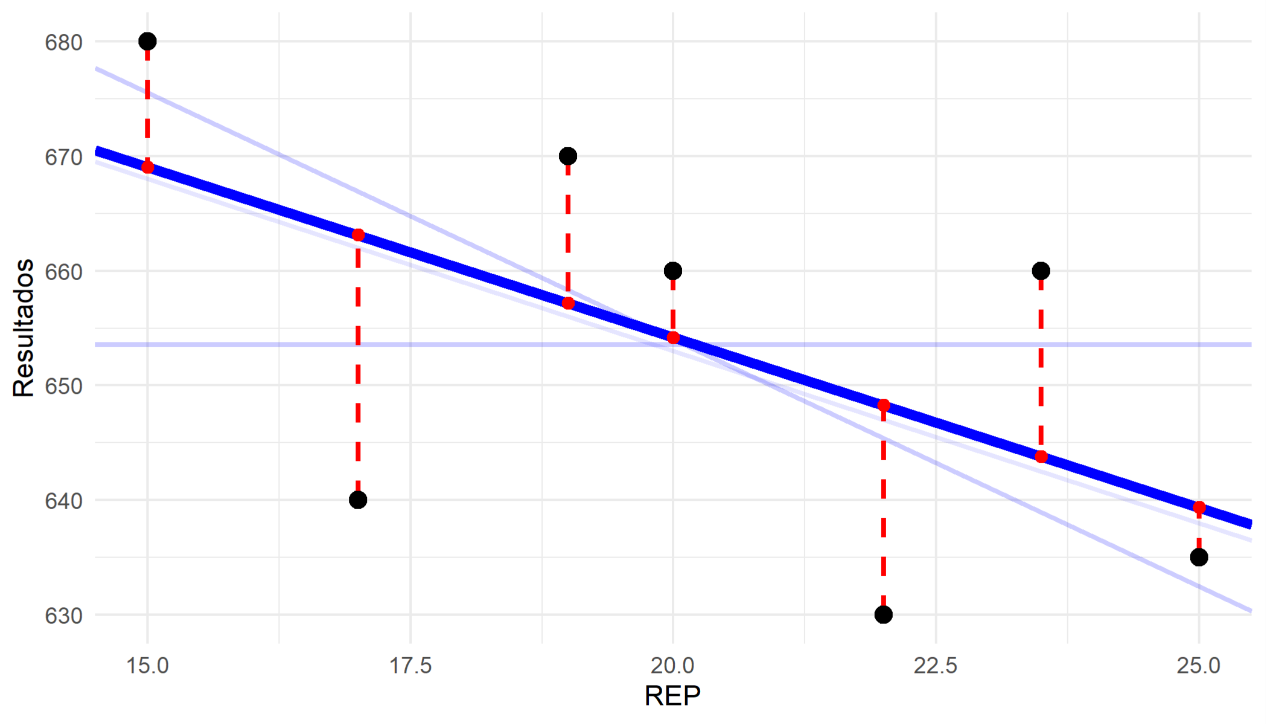
$$u_1 = 4.5;\ u_2 = -26.9;\ u_3 = 11.7;\ u_4 = 6;\ u_5 = -15.4;\ u_6 = 21;\ u_7 = 2.5$$

Suma de cuadrados residuales = $\sum_i^7(u_i^2) = \sum_i^7((Y_i - \hat{Y}_i)^2) = 1603.26$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Suma de Cuadrados Residuales

$$\widehat{Resultados}_i = 713.57 - (2.97 * REP_i)$$



$$u_1 = 10.9;\ u_2 = -23.1;\ u_3 = 12.8;\ u_4 = 5.8;\ u_5 = -18.3;\ u_6 = 16.2;\ u_7 = -4.4$$

Suma de cuadrados residuales = $\sum_i^7 (u_i^2) = \sum_i^7 ((Y_i - \hat{Y}_i)^2) = 1466.85$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

¿Cómo se estiman los coeficientes?

Mínimos Cuadrados Ordinarios (*OLS en inglés*)

- Modelo de regresión simple (una variable independiente) a estimar:

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

- ¿Cómo se estiman los parámetros?
    - **objetivo**: *minimizar la suma del cuadrado de los residuales*

$$\begin{aligned} \min\sum_{i=1}^nu_i^2 &= \min\sum_{i=1}^n(Y_i-\hat{Y}_i)^2 \\ &= \min_{{\hat{\beta}_0}, {\hat{\beta}_1}}\sum_{i=1}^n(Y_i-(\hat{\beta}_0 + \hat{\beta}_1X_i))^2 \end{aligned}$$

- Parámetros estimados:

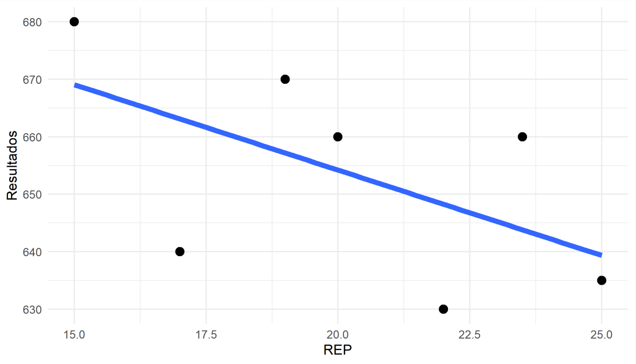
$$\begin{aligned} \hat{\beta}_1 &= \frac{Cov(X,Y)}{Var(X)}=\rho_{x,y}\frac{\sigma_Y}{\sigma_X}\\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

"Mejor" línea

La "mejor línea" (o la que **minimiza la suma de cuadrados residuales**) es, de hecho:

$$\small \begin{aligned} \widehat{Resultados}_i &= 713.57 - (2.97*REP_i) \\ E(Resultados|REP) &= 713.57 - (2.97*REP) \\ \end{aligned}$$

```
ggplot(datos_colegio, aes(REP, Resultados)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm",
              se = FALSE,
              size = 2) +
  theme_minimal()
```



$$\small \widehat{Resultados}_i = \hat{\beta_0} + \hat{\beta_1}REP_i$$

$\hat{\beta}_1=-2.97$ se interpreta como que **cada aumento en una unidad de `REP` esta asociado, en promedio, a una disminución de 2.97 unidades en `Resultados`**.

$\hat{\beta}_0=713.57$ es el valor promedio de `Resultados` cuando `REP` es igual a 0 (no necesariamente tiene interpretación práctica).

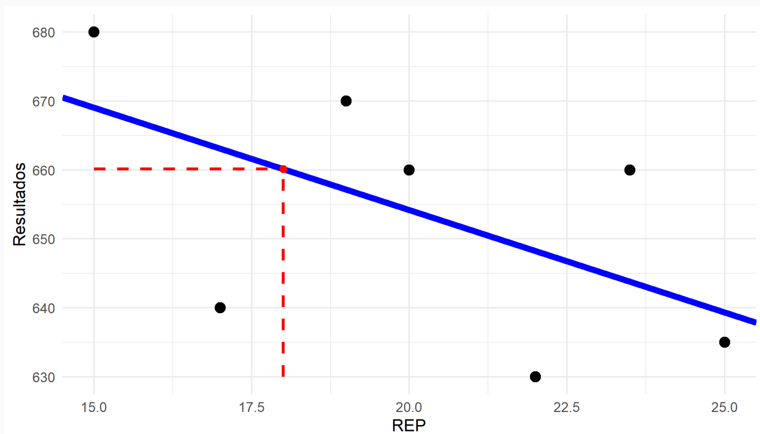
Predicción/Interporlación

\begin{aligned} \widehat{Resultados}_i &= 713.57 - (2.97*REP_i) \\ &= E(Resultados|REP) \end{aligned}

```
##    REP Resultados
## 1 15.0        680
## 2 17.0        640
## 3 19.0        670
## 4 20.0        660
## 5 22.0        630
## 6 23.5        660
## 7 25.0        635
```

Aún cuando nuestra curva **fue estimada** usando 7 valores, podemos usar la curva para "predecir" el valor de `Resultados` esperado para otros valores de `REP`.

\small \begin{aligned} E(Resultados|REP = 18) &=713.57-(2.97*18) \\ &=660.11 \end{aligned}



Ahora con "datos reales"

(Disponibles en el paquete **AER**)

Datos California (USA)

```r
library(AER)
data("CASchools")
str(CASchools)
```

```
## 'data.frame':	420 obs. of  14 variables:
##  $ district   : chr  "75119" "61499" "61549" "61457" ...
##  $ school     : chr  "Sunol Glen Unified" "Manzanita Elementary" "Thermalito Union Elementary" "Golden Feather U
##  $ county     : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6 25 ...
##  $ grades     : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
##  $ students   : num  195 240 1550 243 1335 ...
##  $ teachers   : num  10.9 11.1 82.9 14 71.5 ...
##  $ calworks   : num  0.51 15.42 55.03 36.48 33.11 ...
##  $ lunch      : num  2.04 47.92 76.32 77.05 78.43 ...
##  $ computer   : num  67 101 169 85 171 25 28 66 35 0 ...
##  $ expenditure: num  6385 5099 5502 7102 5236 ...
##  $ income     : num  22.69 9.82 8.98 8.98 9.08 ...
##  $ english    : num  0 4.58 30 0 13.86 ...
##  $ read       : num  692 660 636 652 642 ...
##  $ math       : num  690 662 651 644 640 ...
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

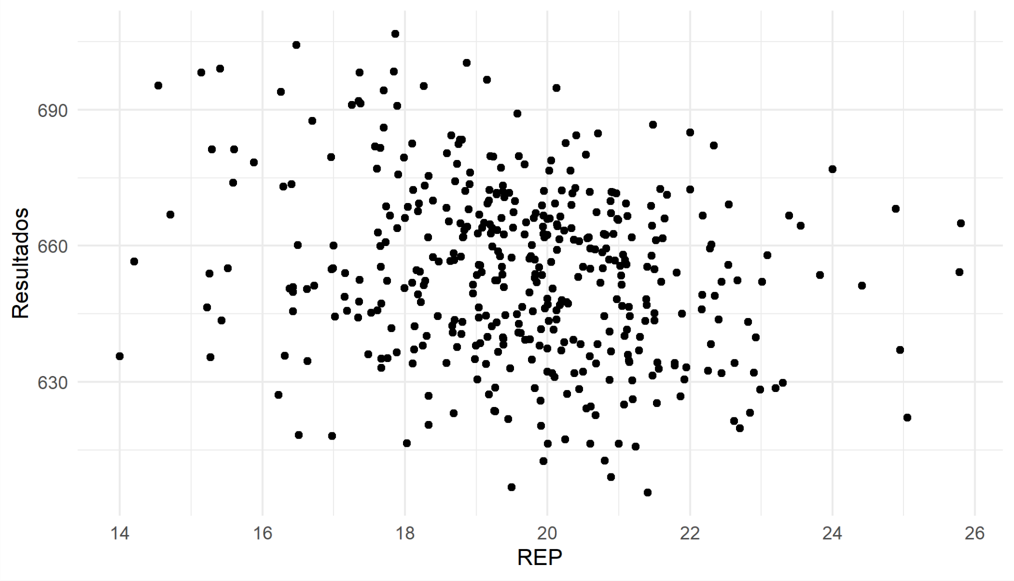
Preparar datos

```r
(datos_reg <- CASchools %>%
  rename(ingresos = income) %>%
  transmute(distrito = district, colegio = school,
            Resultados = (read + math)/2,
            REP = students/teachers,
            ingresos,
            grupo_ingresos = as.factor(ifelse(ingresos >= median(ingresos), 1, 0)),
            computadores = computer,
            almuerzo = lunch))
```

```
##     distrito                              colegio Resultados      REP
## 1      75119                   Sunol Glen Unified     690.80 17.88991
## 2      61499                 Manzanita Elementary     661.20 21.52466
## 3      61549          Thermalito Union Elementary     643.60 18.69723
## 4      61457      Golden Feather Union Elementary     647.70 17.35714
## 5      61523             Palermo Union Elementary     640.85 18.67133
## 6      62042              Burrel Union Elementary     605.55 21.40625
## 7      68536                Holt Union Elementary     606.75 19.50000
## 8      63834                  Vineland Elementary     609.00 20.89412
## 9      62331             Orange Center Elementary     612.50 19.94737
## 10     67306          Del Paso Heights Elementary     612.65 20.80556
## 11     65722            Le Grand Union Elementary     615.75 21.23810
## 12     62174               West Fresno Elementary     616.30 21.00000
## 13     71795               Allensworth Elementary     616.30 20.60000
## 14     72181           Sunnyside Union Elementary     616.30 20.00822
## 15     72298                 Woodville Elementary     616.45 18.02778
## 16     72041              Pixley Union Elementary     617.35 20.25196
## 17     63594          Lost Hills Union Elementary     618.05 16.97787
## 18     63370         Buttonwillow Union Elementary     618.30 16.50980
## 19     64709                   Lennox Elementary     619.80 22.70402
## 20     63560                   Lamont Elementary     620.30 19.91111
## 21     63230          Westmorland Union Elementary     620.50 18.33333
## 22     72058              Pleasant View Elementary     621.40 22.61905
## 23     63843               Wasco Union Elementary     621.75 19.44828
## 24     71811                   Alta Vista Elementary     622.05 25.05263
```

Relación entre variables

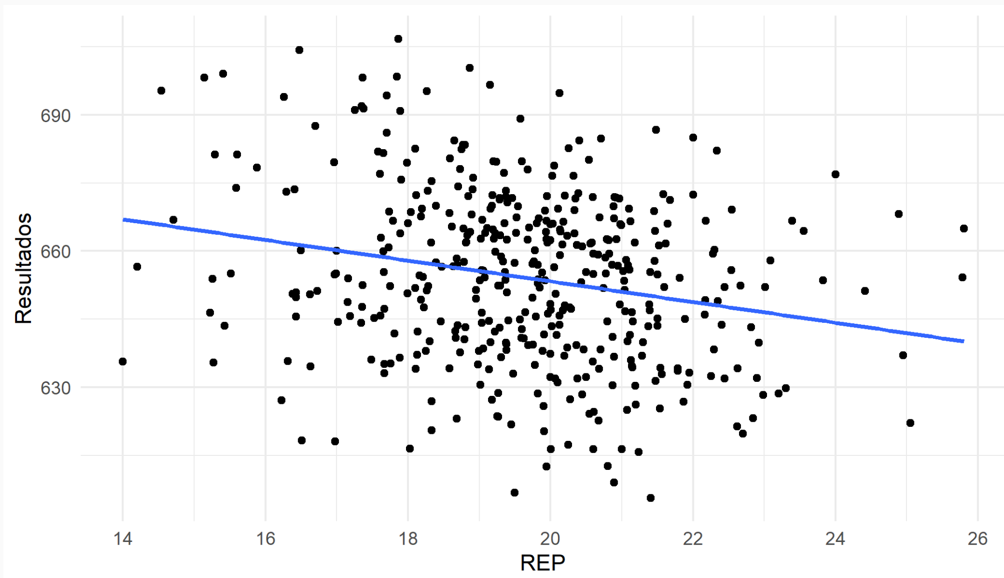
```
datos_reg %>%
  ggplot(aes(REP, Resultados)) +
  geom_point() +
  theme_minimal()
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Graficar la curva de regresión

```r
datos_reg %>%
  ggplot(aes(REP, Resultados)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal()
```



La curva graficada corresponde a $\widehat{Resultados}_i = \hat{\beta}_0+\hat{\beta}_1REP_i$

¿Cuáles son los coeficientes ($\hat{\beta}_0$, $\hat{\beta}_1$)?

Estimar coeficientes "a mano"

\begin{aligned} \hat{\beta}_1 &= \frac{Cov(X,Y)}{Var(X)} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}

```
(coeficientes <- datos_reg %>%
  transmute(Y = Resultados,
            X = REP) %>%
  summarise(beta_1 = cov(X,Y)/var(X),
            beta_0 = mean(Y)-(beta_1*mean(X))))
```

```
##      beta_1   beta_0
## 1 -2.279808 698.9329
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Por suerte **R** lo hace más simple

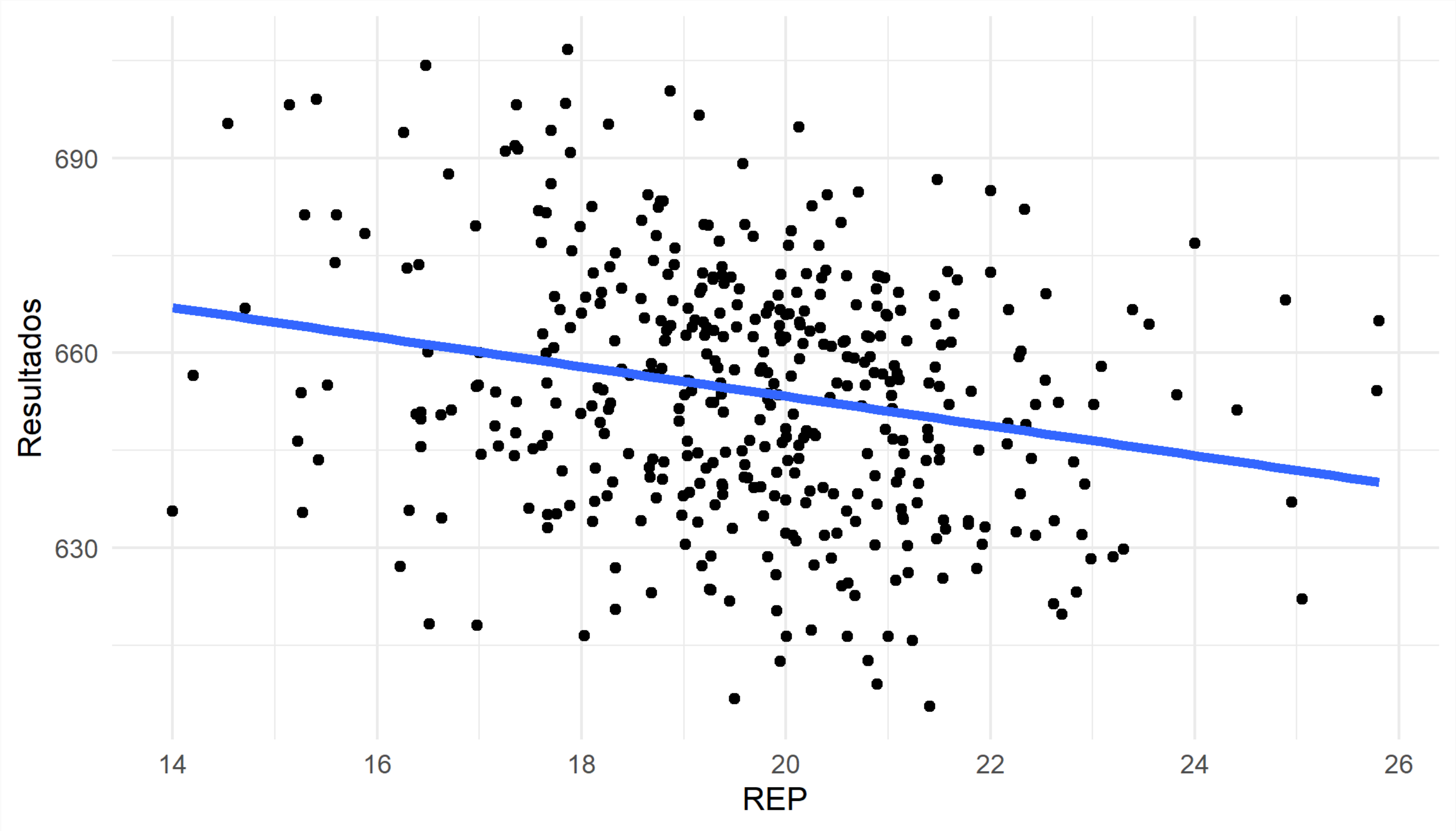
```
modelo1 <- lm(Resultados ~ REP, data = datos_reg)
summary(modelo1)
```

```
##
## Call:
## lm(formula = Resultados ~ REP, data = datos_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
## REP          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

El aumento en una unidad de `REP` esta asociada con una disminución, en promedio, de -2.28 unidades de `Resultados`.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

¿Qué tan bien **ajustada** es la curva?

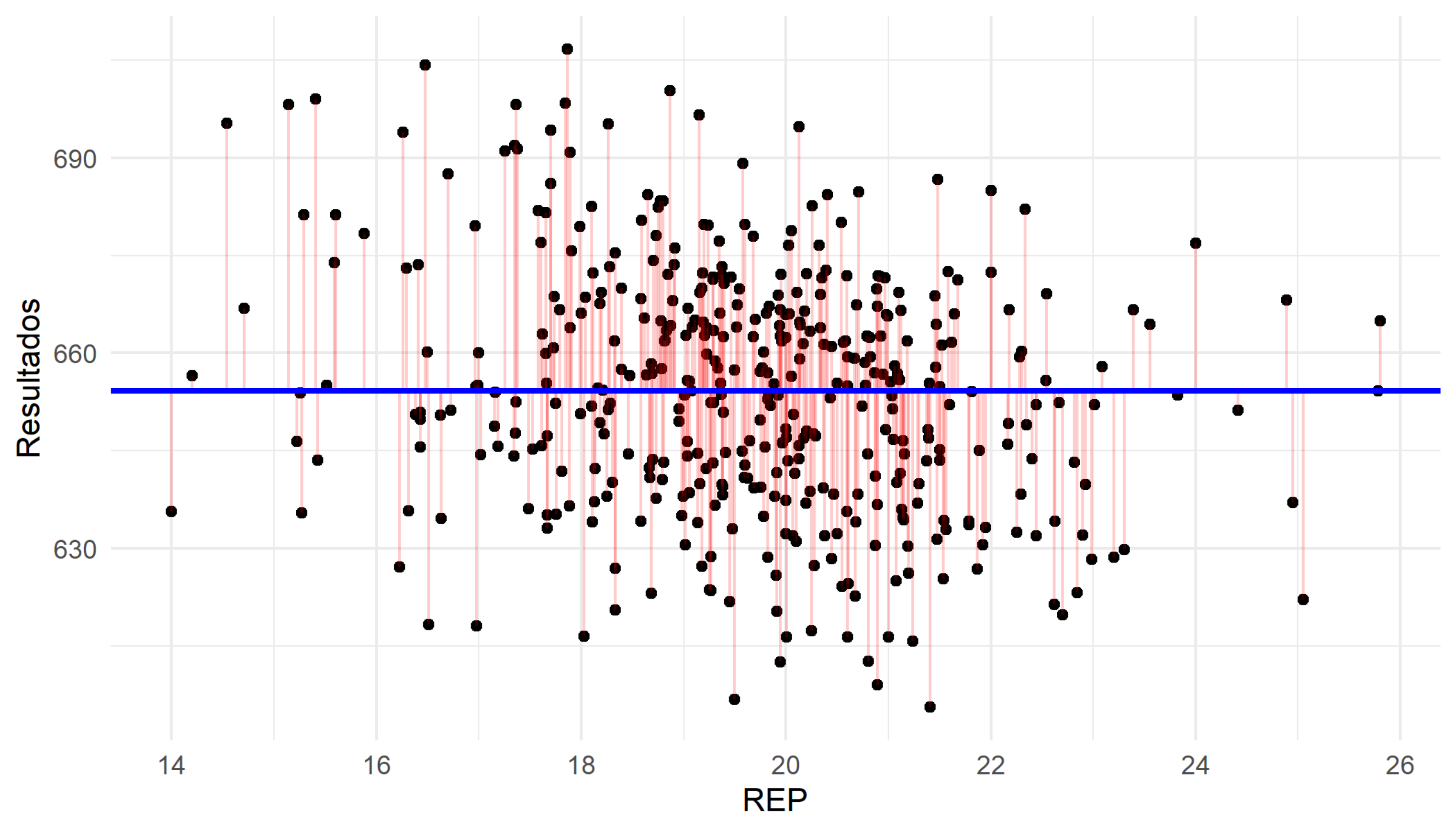


El término "ajustar" se refiere a que tan cerca de todos los puntos se encuentra la línea. En otras palabras, **¿cuánto explica la línea calculada la relación entre estas variable?**

Una de las métricas más utilizadas para representar esto es el R^2.

R^2 - Coeficiente de determinación

Suma de Cuadrados Totales (SCT)

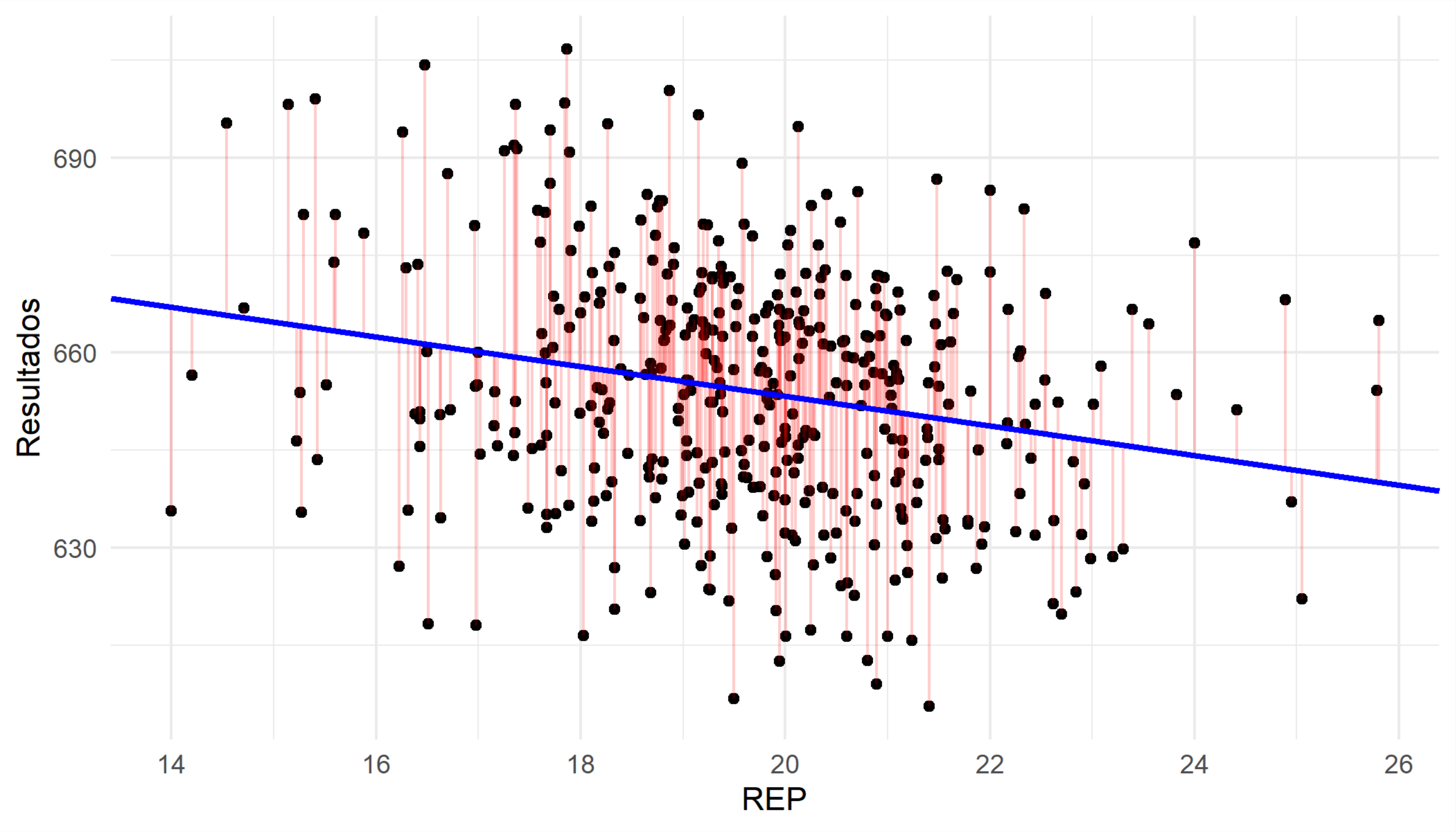


SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

R^2 - Coeficiente de determinación

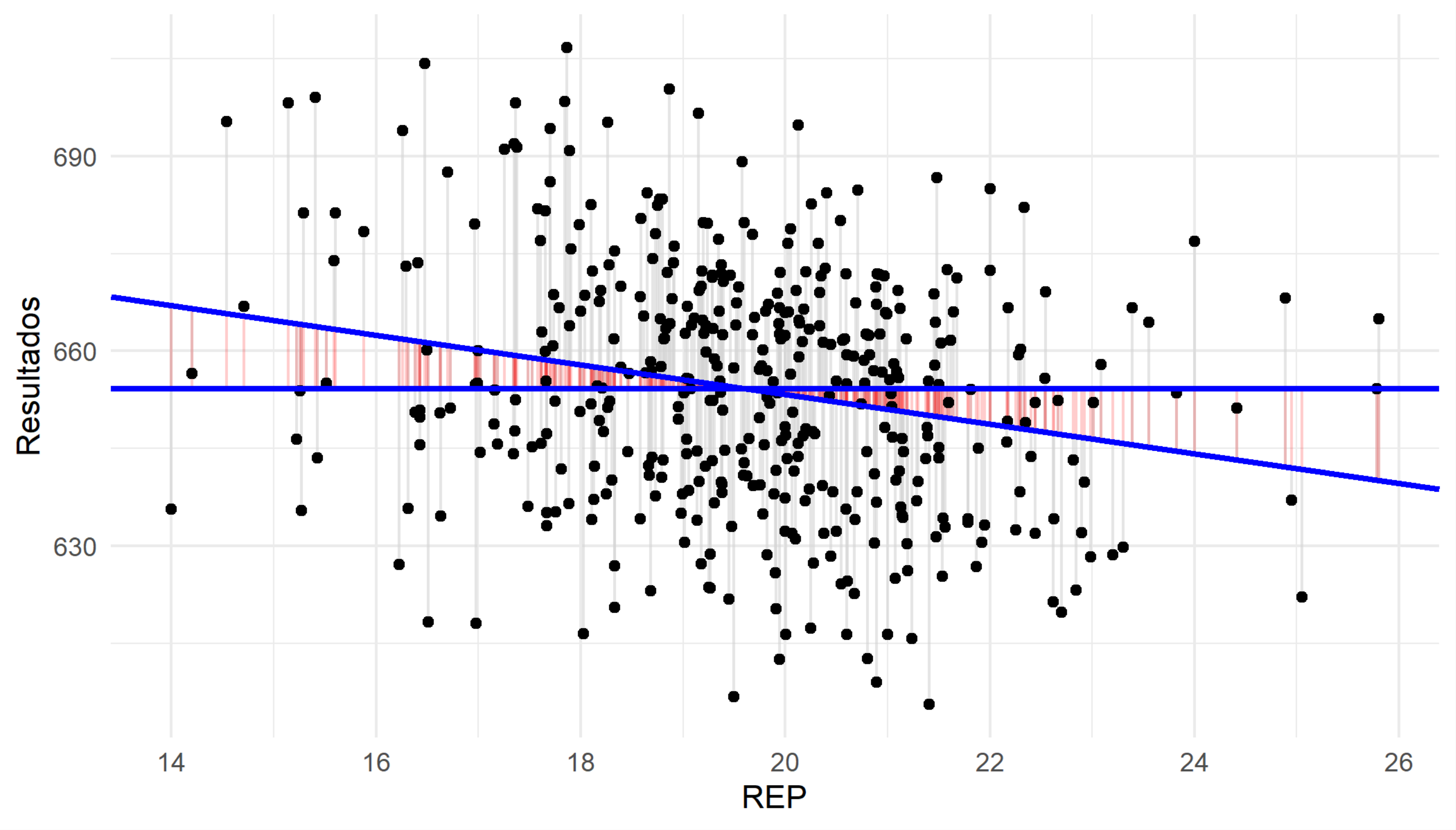
Suma de Cuadrados Residuales (SCR)



$$SCR=\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (Y_i-\hat{Y}_i)^2$$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Suma de Cuadrados Explicados (SCE)



$SCE=\sum_{i=1}^n (\hat{Y_i}-\bar{Y})^2$

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

R^2 - Coeficiente de determinación

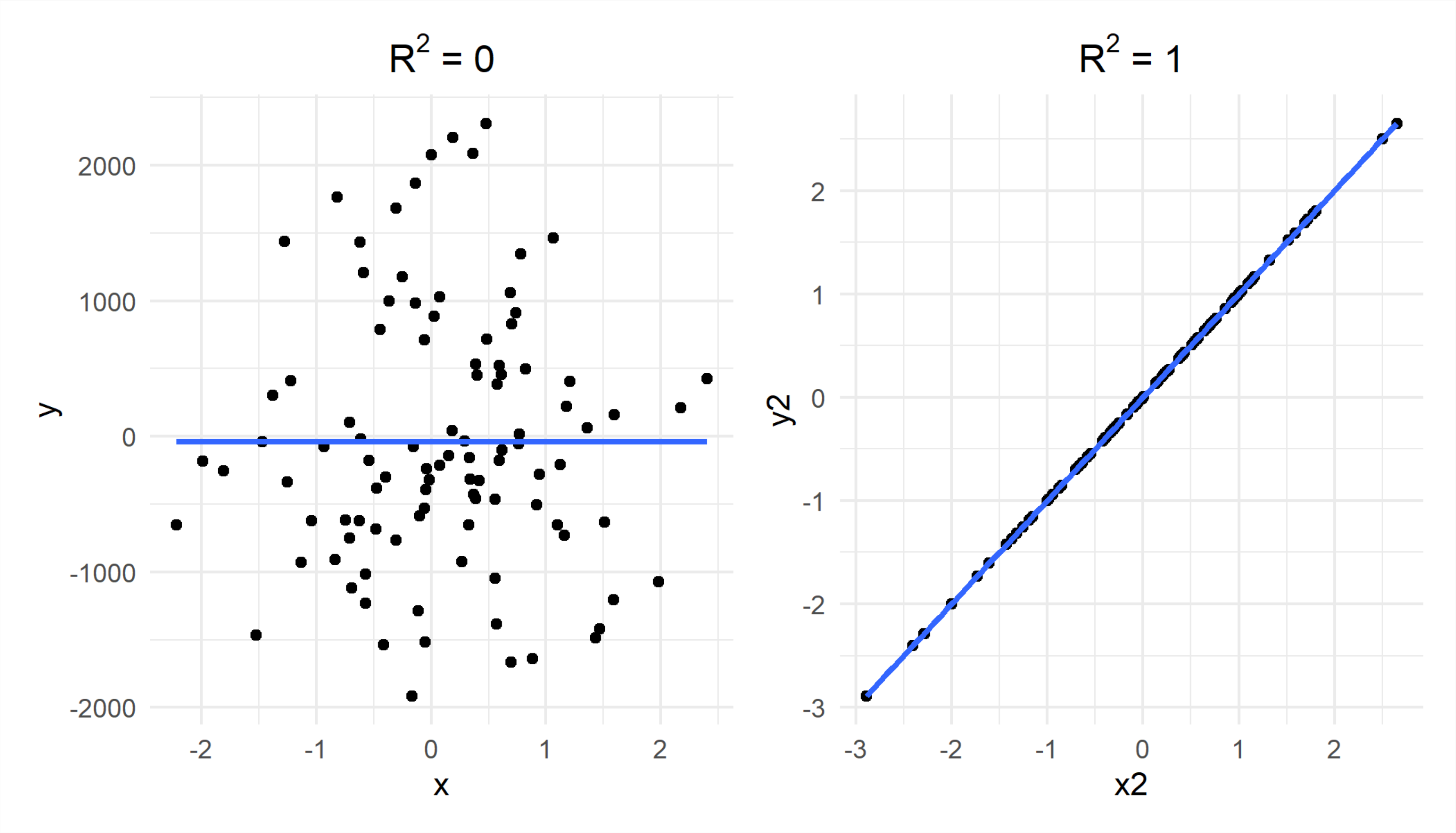
- Suma de Cuadrados Residuales (SCR): $\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (Y_i-\hat{Y}_i)^2$
- Suma de Cuadrados Explicados (SCE): $\sum_{i=1}^n (\hat{Y_i}-\bar{Y})^2$
- Suma de Cuadrados Totales (SCT): $SCT = SCE + SCR = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Teniendo estos valores, el coeficiente de determinación corresponde a:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Noten que $R^2 \in [0,1]$, con 0 correspondiente a un nulo ajuste y 1 a un ajuste perfecto (todos los puntos sobre la curva estimada)

¿Cómo se ve esto?



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Calcular R^2 "a mano"

R^2=\frac{SCE}{SCT}=\frac{\sum_{i=1}^n (\hat{Y_i}-\bar{Y})^2}{\sum(Y_i - \bar{Y})^2}

```
datos_reg %>%
  transmute(X = REP,
            Y = Resultados) %>%
  summarise(SCE = sum(((coeficientes$beta_0+(coeficientes$beta_1*X))-mean(Y))^2),
            SCT = sum((Y-mean(Y))^2),
            R2 = round(SCE/SCT, 5))
```

```
##          SCE      SCT      R2
## 1 7794.109 152109.6 0.05124
```

Interpretamos este valor como que **nuestra variable independiente, `REP`, explica un 5.1% de la variación de la variable dependiente, `Resultados`**.

Calcular R^2 en **R**

```
summary(modelo1)
```

```
##
## Call:
## lm(formula = Resultados ~ REP, data = datos_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9329     9.4675  73.825  < 2e-16 ***
## REP          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Variable independiente no numérica

Hasta ahora solo vimos un ejemplo donde tanto la variable dependiente, Y, como la variable independiente, X son numéricas.

¿Cómo sería una regresión con una variable independiente categórica?

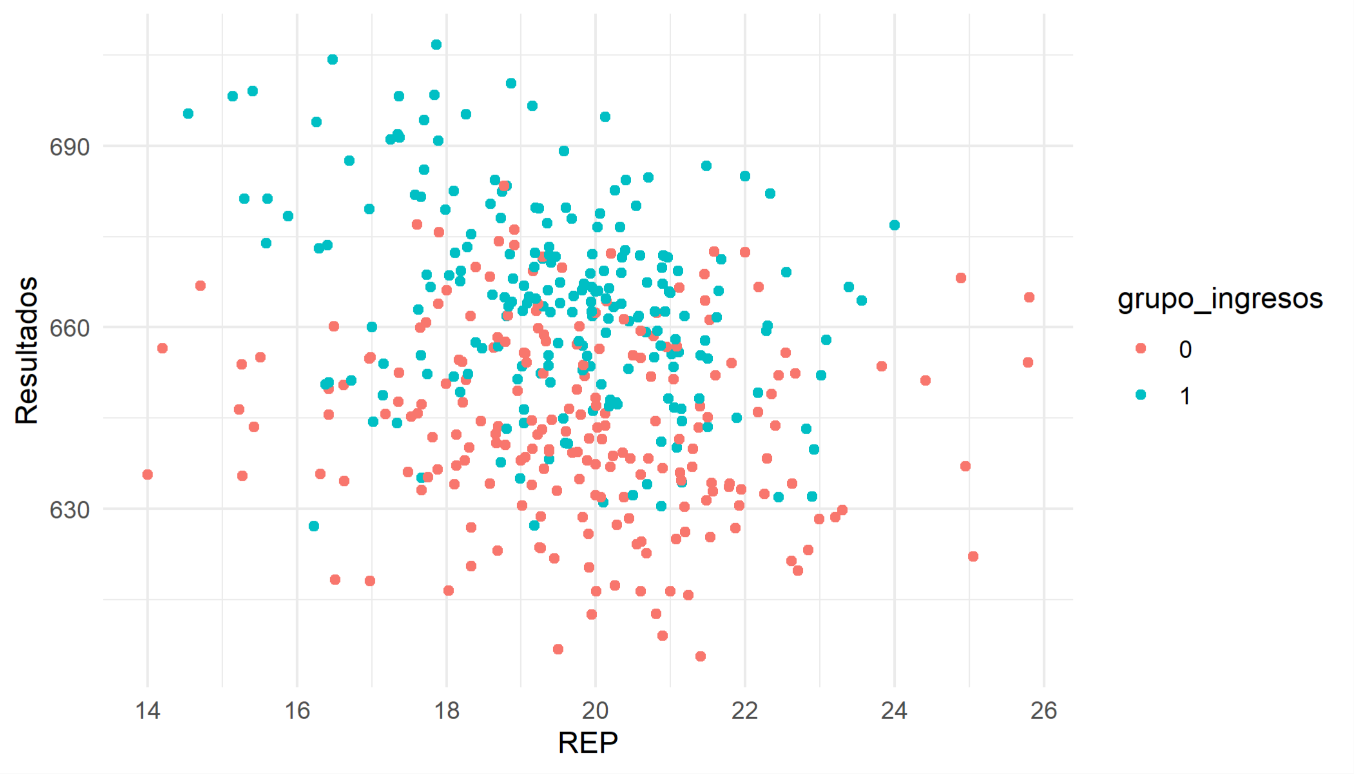
```
glimpse(datos_reg)
```

```
## Rows: 420
## Columns: 6
## $ distrito       <chr> "75119", "61499", "61549", "61457", "61523", "62042", "~
## $ colegio        <chr> "Sunol Glen Unified", "Manzanita Elementary", "Thermali~
## $ Resultados     <dbl> 690.80, 661.20, 643.60, 647.70, 640.85, 605.55, 606.75,~
## $ REP            <dbl> 17.88991, 21.52466, 18.69723, 17.35714, 18.67133, 21.40~
## $ ingresos       <dbl> 22.690001, 9.824000, 8.978000, 8.978000, 9.080333, 10.4~
## $ grupo_ingresos <fct> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

$$\begin{align} grupo\_ingresos = \begin{cases} 0 & si\ ingresos\ en\ el\ 50\%\ inferior \\ 1 & si\ ingresos\ en\ el\ 50\%\ superior \end{cases} \end{align}$$

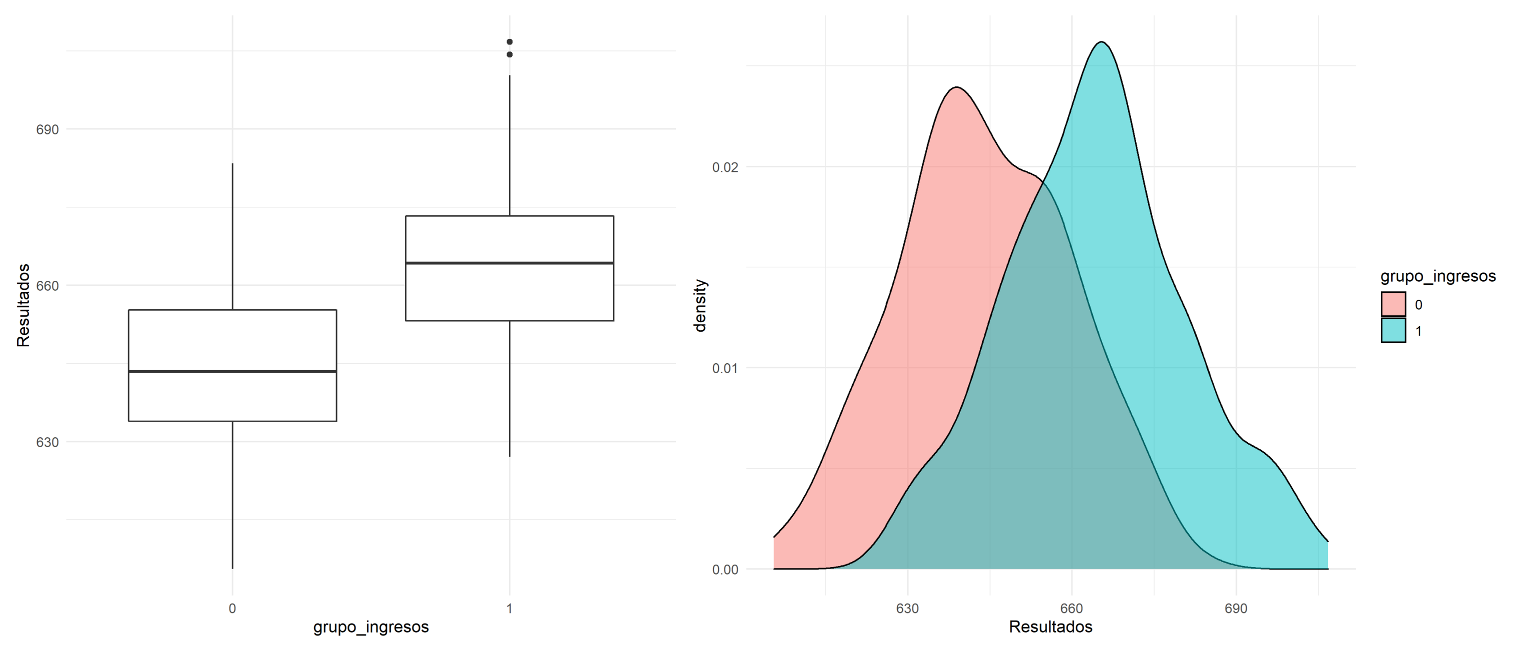
Variable independiente no numérica

```
datos_reg %>%
  ggplot(aes(x = REP, y = Resultados, col = grupo_ingresos)) +
  geom_point() +
  theme_minimal()
```



Los colegios de mayores ingresos parecieran tener mejores resultados.

Variable independiente no numérica




```
datos_reg %>%
  group_by(grupo_ingresos) %>%
  summarise(resultados_prom = mean(Resultados))
```

```
## # A tibble: 2 x 2
##   grupo_ingresos resultados_prom
##   <fct>                    <dbl>
## 1 0                         644.
## 2 1                         664.
```

Los colegios de más ingresos tienden a tener mejores resultados.

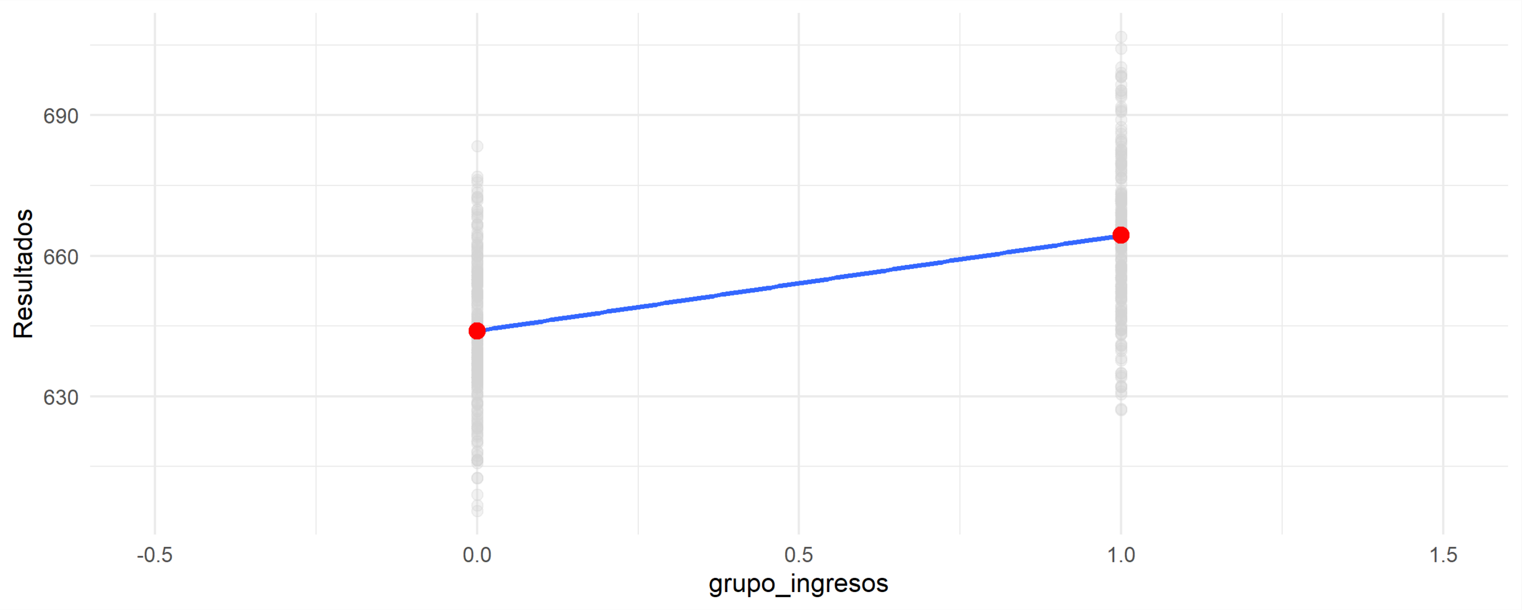
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Regresión con **grupo_ingresos**

\small \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 1_{g.ing. = 1}

```
(modelo2 <- lm(Resultados~grupo_ingresos, data = datos_reg))
```

```
##
## Call:
## lm(formula = Resultados ~ grupo_ingresos, data = datos_reg)
##
## Coefficients:
##    (Intercept)  grupo_ingresos1
##         643.96            20.39
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Regresión con **grupo_ingresos**

\small \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 1_{g.ing. = 1}

```r
(modelo2 <- lm(Resultados~grupo_ingresos, data = datos_reg))
```

```
##
## Call:
## lm(formula = Resultados ~ grupo_ingresos, data = datos_reg)
##
## Coefficients:
##    (Intercept)  grupo_ingresos1
##         643.96            20.39
```

\small \begin{align} \widehat{Resultados}_i&=643.96 + (20.39*1_{g.ing. = 1}) \\ E(Resultados|grupo_ingresos)&=643.96 + (20.39*1_{g.ing. = 1}) \end{align}

\begin{aligned} \scriptsize E(Resultados|GrupoIng=0) & \scriptsize =643.96 + (20.39*0) \\ & \scriptsize =643.96 \\ \scriptsize E(Resultados|GrupoIng=1) & \scriptsize =643.96 + (20.39*1) \\ & \scriptsize =664.35 \end{aligned}

```
## # A tibble: 2 x 2
##   grupo_ingresos resultados_prom
##   <fct>                    <dbl>
## 1 0                         644.
## 2 1                         664.
```

Pertenecer al 50% superior de ingresos está asociado con un aumento, en promedio, de 20.39 unidades en `Resultado`.

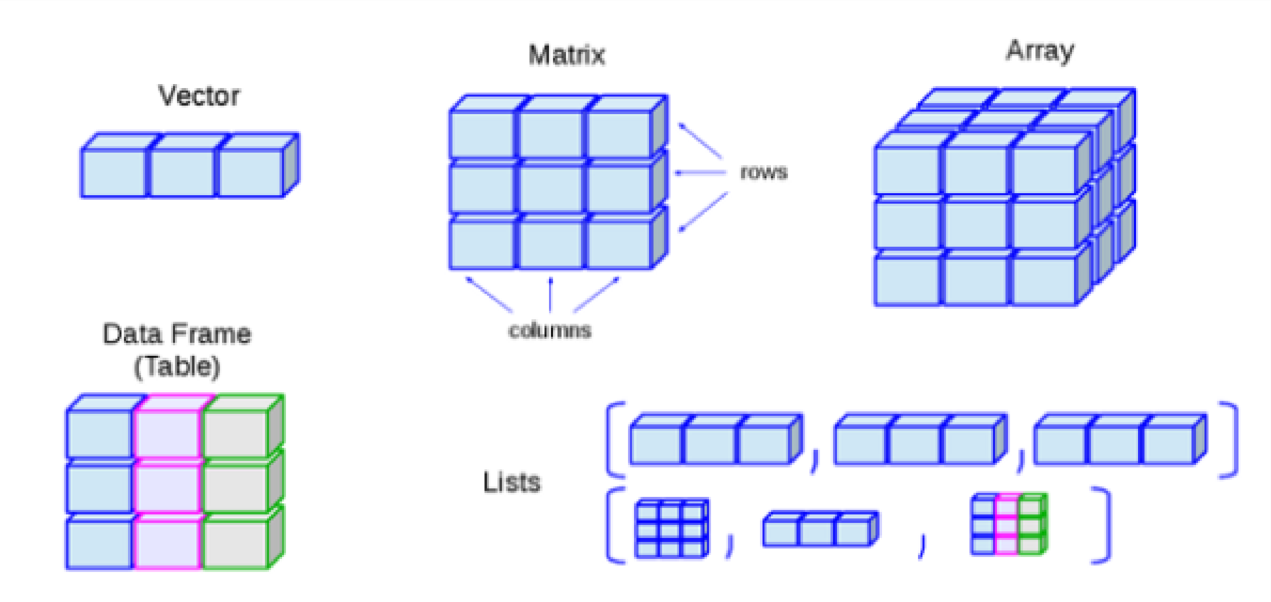
¿Qué es esto?

```
str(modelo2)
```

```
## List of 13
##  $ coefficients : Named num [1:2] 644 20.4
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "grupo_ingresos1"
##  $ residuals    : Named num [1:420] 26.451 17.236 -0.364 3.736 -3.114 ...
##   ..- attr(*, "names")= chr [1:420] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:420] -13406.22 208.89 -2.37 1.73 -5.12 ...
##   ..- attr(*, "names")= chr [1:420] "(Intercept)" "grupo_ingresos1" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:420] 664 644 644 644 644 ...
##   ..- attr(*, "names")= chr [1:420] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:420, 1:2] -20.4939 0.0488 0.0488 0.0488 0.0488 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:420] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "grupo_ingresos1"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   .. ..- attr(*, "contrasts")=List of 1
##   .. .. ..$ grupo_ingresos: chr "contr.treatment"
##   ..$ qraux: num [1:2] 1.05 1.05
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 418
##  $ contrasts    :List of 1
##   ..$ grupo_ingresos: chr "contr.treatment"
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Tipos de objetos



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Paquete **broom**

```r
library(broom)
tidy(modelo1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   699.       9.47      73.8  6.57e-242
## 2 REP            -2.28     0.480     -4.75 2.78e-  6
```

```r
glance(modelo1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    0.0512        0.0490  18.6      22.6 2.78e-6     1 -1822. 3650. 3663.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
augment(modelo1)
```

```
## # A tibble: 420 x 8
##    Resultados   REP .fitted .resid .std.resid    .hat .sigma  .cooksd
##         <dbl> <dbl>   <dbl>  <dbl>      <dbl>   <dbl>  <dbl>    <dbl>
##  1       691.  17.9    658.   32.7      1.76  0.00442   18.5 0.00689
##  2       661.  21.5    650.   11.3      0.612 0.00475   18.6 0.000893
##  3       644.  18.7    656.  -12.7     -0.685 0.00297   18.6 0.000700
##  4       648.  17.4    659.  -11.7     -0.629 0.00586   18.6 0.00117
##  5       641.  18.7    656.  -15.5     -0.836 0.00301   18.6 0.00105
##  6       606.  21.4    650.  -44.6     -2.40  0.00446   18.5 0.0130
##  7       607.  19.5    654.  -47.7     -2.57  0.00239   18.5 0.00794
##  8       609   20.9    651.  -42.3     -2.28  0.00343   18.5 0.00895
##  9       612.  19.9    653.  -41.0     -2.21  0.00244   18.5 0.00597
## 10       613.  20.8    652.  -38.9     -2.09  0.00329   18.5 0.00723
## # … with 410 more rows
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Comparar modelos

R^2 de \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 REP

```
glance(modelo1) %>% select(r.squared) %>% pull(1)
```

```
## [1] 0.05124009
```

`REP` **explica un 5.1%** de la variación en `Resultados`.

R^2 de \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 1_{g.ing. = 1}

```
glance(modelo2) %>% select(r.squared) %>% pull(1)
```

```
## [1] 0.286863
```

`grupo_ingresos` **explica un 28.7%** de la variación en `Resultados`.

Analizar modelos

\small \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 REP

```
tidy(modelo1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   699.        9.47      73.8  6.57e-242
## 2 REP            -2.28      0.480     -4.75 2.78e-  6
```

\small \widehat{Resultados}_i=\hat{\beta}_0 + \hat{\beta}_1 1_{g.ing. = 1}

```
tidy(modelo2)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        644.       1.11      579.  0
## 2 grupo_ingresos1     20.4      1.57       13.0 1.49e-32
```

- Para cada parámetro de están realizando las siguientes pruebas de hipótesis: \small \begin{align} H_0: \beta_j=0 \\ H_A: \beta_j \neq 0 \end{align}
- Para probar estas hipótesis tenemos las estimaciones y sus errores estándar. Con estos podemos generar una métrica estandarizada (**estadístico t**), \small \frac{\hat{\beta_j}-\beta_j}{EE_{\beta_j}}.
- Estos estadísticos nos permiten calcular **p-values** para ver que tan extremas son nuestras estimaciones bajo la hipótesis nula, \small \beta_j=0.
- Noten que todos los **p-value** son MUY bajos. Decimos entonces, que **los parámetros estimados son estadísticamente significativos**.

Regresión Lineal Múltiple

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

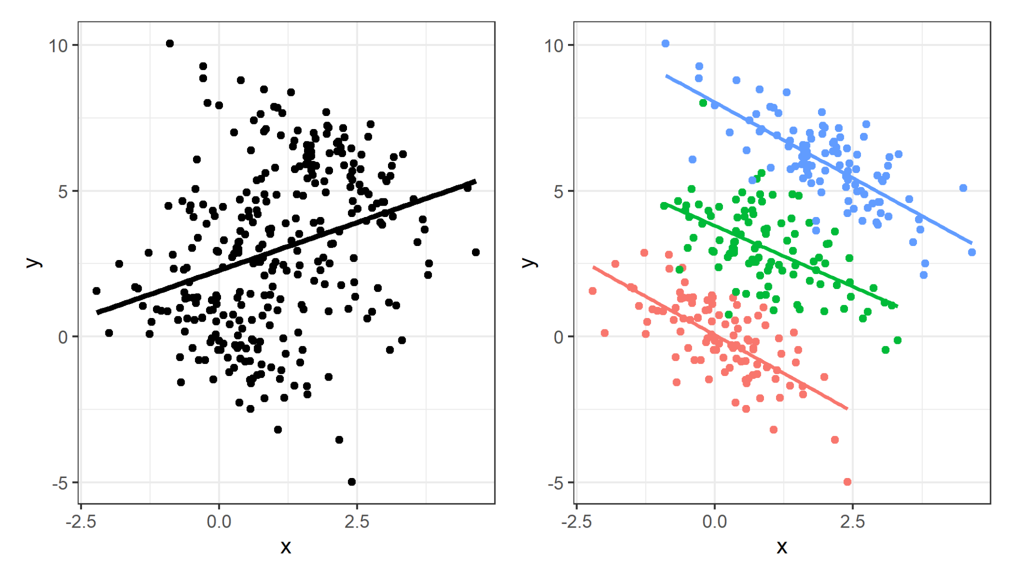
Regresión Lineal Múltiple

- ¿Es factible que solo `REP` o solo `grupo_ingresos` influyan en `Resultados`?

\small Resultados=\beta_0+\beta_1REP+\beta_2A+\beta_3B+\epsilon

- ¿Qué pasa si no se incluyen otras variables relacionadas?

Paradoja de Simpson



- "Sesgo de variable omitida"

Regresión Lineal Múltiple

¿Cómo se estiman los parámetros?

En forma matricial:

$$Y=X\beta+\epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Lo que debemos hacer es estimar el vector de parámetros $\hat{\beta}$

$$\hat{\beta}=(X'X)^{-1}X'Y$$

Terminando finalmente con $\hat{Y}=X\hat{\beta}$

Estimar coeficientes "a mano"

\widehat{Resultados}_i=\hat{\beta_0} + \hat{\beta}_1REP_i + \hat{\beta}_2ingresos_i

\hat{\beta}=(X'X)^{-1}X'Y

```
Y ← datos_reg %>%
  select(Resultados) %>%
  as.matrix()

X ← datos_reg %>%
  mutate(intercepto = rep(1, nrow(.))) %>%
  select(intercepto, REP, ingresos) %>%
  as.matrix()

solve(t(X) %*% X) %*% (t(X) %*% Y)
```

```
##             Resultados
## intercepto 638.7291572
## REP         -0.6487401
## ingresos     1.8391120
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Por suerte **R** lo hace más simple

\widehat{Resultados}_i=\hat{\beta_0} + \hat{\beta}_1REP_i + \hat{\beta}_2ingresos_i

```
modelo3 <- lm(Resultados ~ REP + ingresos, data = datos_reg)
summary(modelo3)
```

```
##
## Call:
## lm(formula = Resultados ~ REP + ingresos, data = datos_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.608  -9.052   0.707   9.259  31.898
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 638.72916    7.44908  85.746   <2e-16 ***
## REP          -0.64874    0.35440  -1.831   0.0679 .
## ingresos      1.83911    0.09279  19.821   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.35 on 417 degrees of freedom
## Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5091
## F-statistic: 218.3 on 2 and 417 DF,  p-value: < 2.2e-16
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Interpretación

```
tidy(modelo3)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  639.       7.45       85.7  5.70e-267
## 2 REP           -0.649    0.354      -1.83 6.79e-  2
## 3 ingresos       1.84     0.0928     19.8  4.38e- 62
```

- El aumento en una unidad de `REP` esta, en promedio, asociado con una disminución de -0.65 en `Resultados` **manteniendo `ingresos` constante** (o controlando por `ingresos`).

- El aumento en una unidad de `ingresos` esta, en promedio, asociado con un aumento de 1.84 en `Resultados` **manteniendo `REP` constante** (o controlando por `REP`).

- Pero **OJO**, el p-value asociado a `REP` es 0.065 por lo que **no es estadísticamente significativo al 0.05** nivel de significancia (si lo sería considerando un nivel de significancia de 0.1).

- Por otro lado, `ingresos` **si es estadísticamente significativo** (p-value <<< 0.05).

Comparemos

Comparemos los valores de R^2 para `modelo1` (simple) y `modelo2` (simple) y `modelo3` (múltiple)

```
glance(modelo1) %>% select(r.squared) %>% pull(1)
```

```
## [1] 0.05124009
```

```
glance(modelo2) %>% select(r.squared) %>% pull(1)
```

```
## [1] 0.286863
```

```
glance(modelo3) %>% select(r.squared) %>% pull(1)
```

```
## [1] 0.511483
```

- `modelo3` explica más del 50% de la variación en `Resultados`.
- Pero **OJO** con el R^2: aumentará siempre que sumemos variables (o al menos no bajará).

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

R^2 ajustado

$$R^2_{adj}=1-\left(\frac{SCE}{SCT}\frac{n-1}{n-k-1}\right)=1-\left(\frac{\sum_{i=1}^n (Y_i-\hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}\frac{n-1}{n-k-1}\right)$$

n es el número de observaciones y k es el número de variables independientes.

- Si la nueva variable no "aporta nueva información", R^2_{adj} no aumenta
- Debido a lo anterior, R^2_{adj} suele ser mejor para la comparación entre modelos

```r
glance(modelo1) %>% select(adj.r.squared) %>% pull(1)
```

```
## [1] 0.04897033
```

```r
glance(modelo2) %>% select(adj.r.squared) %>% pull(1)
```

```
## [1] 0.2851569
```

```r
glance(modelo3) %>% select(adj.r.squared) %>% pull(1)
```

```
## [1] 0.50914
```

R^2 no es la única métrica

R^2, R^2_{adj}, **AIC**, **BIC**, **LogLikehood**, **deviance**.

```
glance(modelo1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    0.0512        0.0490  18.6      22.6 2.78e-6     1 -1822. 3650. 3663.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(modelo2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.287         0.285  16.1      168. 1.49e-32     1 -1762. 3531. 3543.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
glance(modelo3)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.511         0.509  13.3      218. 1.35e-65     2 -1683. 3374. 3390.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Ejercicio

Ejercicio

- `EjercicioRegresion.R`

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

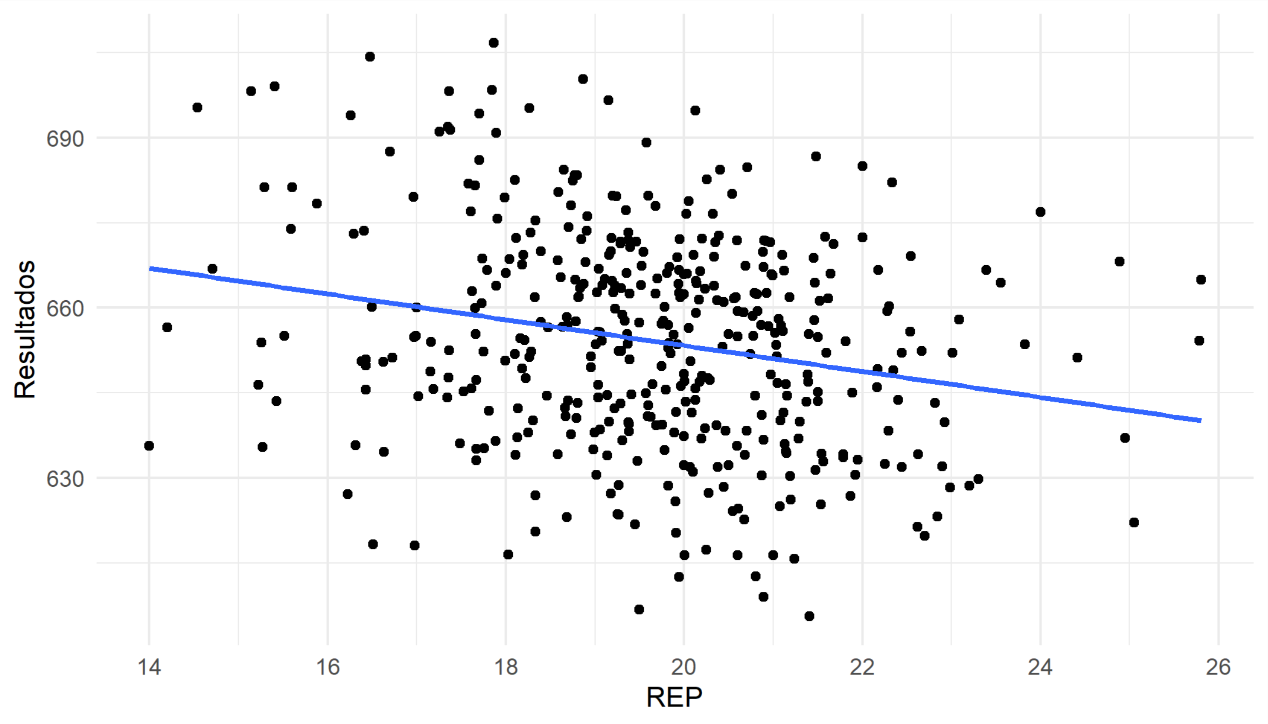
Interacciones y transformaciones

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo simple

Resultados_i=\hat{\beta_0} + \hat{\beta_1}REP_i

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   699.       9.47      73.8  6.57e-242
## 2 REP            -2.28     0.480     -4.75 2.78e-  6
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo múltiple

Resultados_i=\hat{\beta_0} + \hat{\beta_1}REP_i + \hat{\beta_2}1_{g.ing=1}

```
## # A tibble: 3 x 5
##   term            estimate std.error statistic   p.value
##   <chr>              <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)       682.       8.11      84.1  1.34e-263
## 2 REP                -1.92     0.407     -4.73 3.05e-  6
## 3 grupo_ingresos1    19.9      1.54      12.9  1.88e- 32
```

\small \begin{align} E(Resultados|grupo_ingresos=1)&=682-(1.92*REP)+(19.9*1) \\
E(Resultados|grupo_ingresos=1)&=701.9-(1.92*REP) \end{align}

\small \begin{align} E(Resultados|grupo_ingresos=0)&=682-(1.92*REP)+(19.9*0) \\
E(Resultados|grupo_ingresos=0)&=682-(1.92*REP) \end{align}

Modelo múltiple

Resultados_i=\hat{\beta_0} + \hat{\beta_1}REP_i + \hat{\beta_2}1_{g.ing=1}

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      682.        8.11      84.1  1.34e-263
## 2 REP               -1.92      0.407     -4.73 3.05e-  6
## 3 grupo_ingresos1   19.9       1.54      12.9  1.88e- 32
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo con interacción

Resultados_i=\hat{\beta_0} + \hat{\beta_1}REP_i + \hat{\beta_2}1_{g.ing=1}+\hat{\beta_3}(REP_i*1_{g.ing=1})

```
## # A tibble: 4 x 5
##   term                   estimate std.error statistic   p.value
##   <chr>                     <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)              664.      10.5        62.9  1.08e-214
## 2 REP                        -1.01     0.531     -1.90 5.83e-  2
## 3 grupo_ingresos1           62.5     16.1         3.88 1.21e-  4
## 4 REP:grupo_ingresos1       -2.17     0.818     -2.66 8.16e-  3
```

\small \begin{align} E(Resultados|grupo_ingresos=1)&=664-(1.01*REP)+(62.5*1)-(2.17*REP*1) \\
E(Resultados|grupo_ingresos=1)&=726.5-(3.18*REP) \end{align}

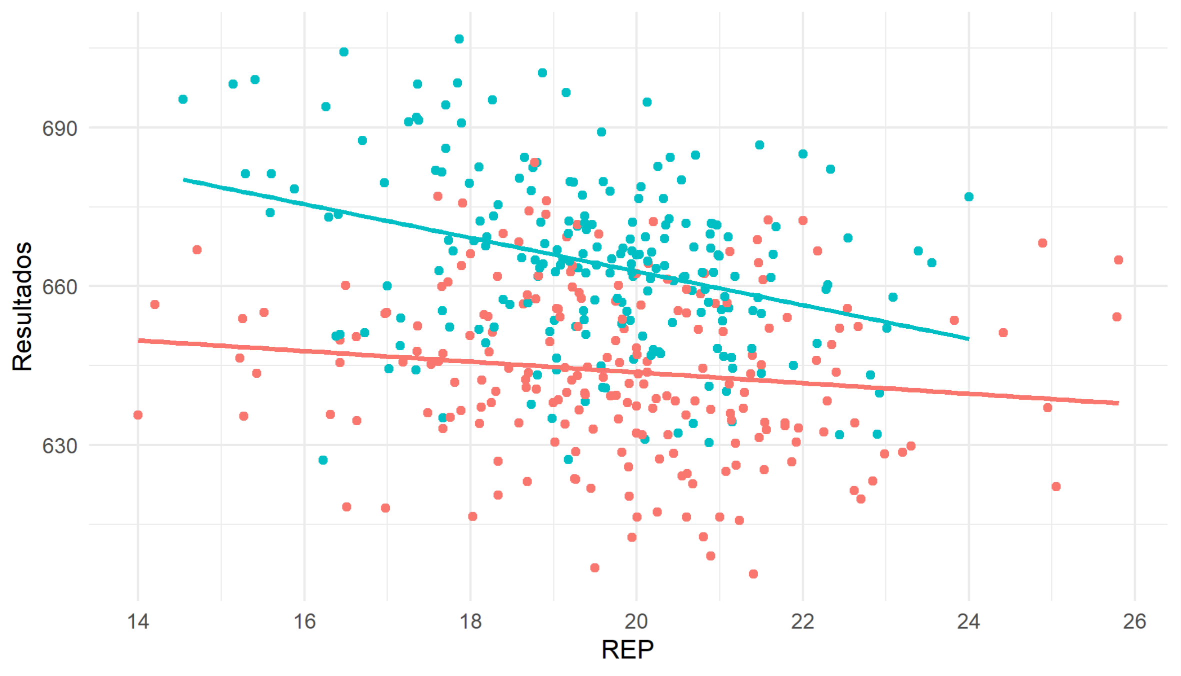
\small \begin{align} E(Resultados|grupo_ingresos=0)&=664-(1.01*REP)+(62.5*0)-(2.17*REP*0) \\
E(Resultados|grupo_ingresos=0)&=664-(1.01*REP) \end{align}

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo con interacción

Resultados_i=\hat{\beta_0} + \hat{\beta_1}REP_i + \hat{\beta_2}1_{g.ing=1}+\hat{\beta_3}(REP_i*1_{g.ing=1})

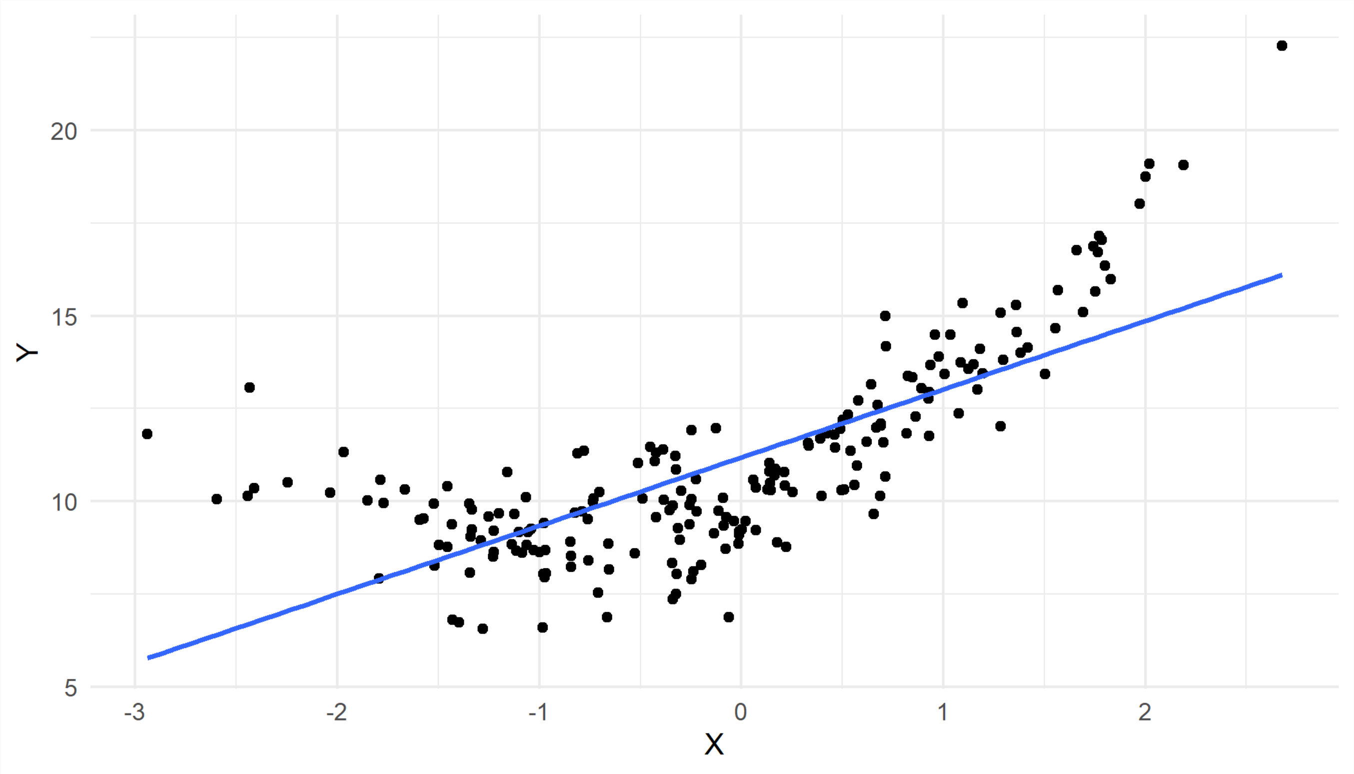
```
## # A tibble: 4 x 5
##   term                  estimate std.error statistic   p.value
##   <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)             664.      10.5        62.9  1.08e-214
## 2 REP                       -1.01     0.531     -1.90 5.83e-  2
## 3 grupo_ingresos1          62.5     16.1         3.88 1.21e-  4
## 4 REP:grupo_ingresos1      -2.17     0.818     -2.66 8.16e-  3
```



Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Otro tipo de interacción

\widehat{Y}_i=\hat{\beta_0}+\hat{\beta_1}X_i

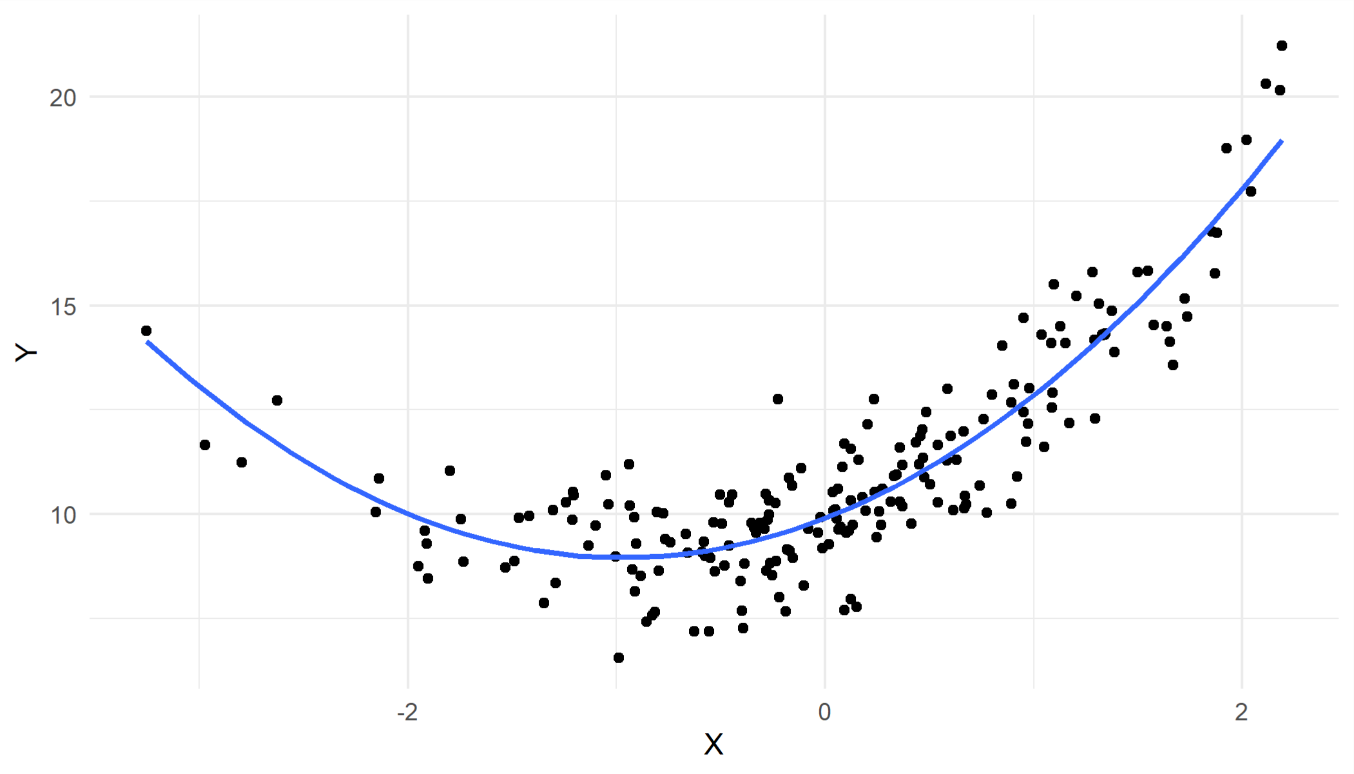



```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.546
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Otro tipo de interacción

\widehat{Y}_i=\hat{\beta_0}+\hat{\beta_1}X_i+\hat{\beta_2}X_i^2




```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.866
```

Transformaciones de variables

Transformación "cosméticas"

```
modelo3 <- lm(Resultados ~ REP + ingresos, data = datos_reg)
tidy(modelo3)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  639.       7.45       85.7  5.70e-267
## 2 REP           -0.649    0.354      -1.83 6.79e-  2
## 3 ingresos       1.84     0.0928     19.8  4.38e- 62
```

Si dividimos la variable ingresos (miles de USD) por 10, lo resultante es una variable que representa decenas de miles de USD.

```
datos_reg %>%
  mutate(ingresos_nuevo = ingresos/10) %>%
  lm(Resultados ~ REP + ingresos_nuevo, data = .) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     639.        7.45      85.7  5.70e-267
## 2 REP              -0.649     0.354     -1.83 6.79e-  2
## 3 ingresos_nuevo   18.4       0.928     19.8  4.38e- 62
```

- El coeficiente, \hat{\beta}, cambia en orden de magnitud: la interpretación es que un aumento en 10.000 USD de ingreso se asocia en promedio con un aumento de 18.4 en `Resultados`.
- Pero noten que el estadístico t y su respectivo p-value no cambian.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformaciones de variables

Transformación logarítmica

¿Qué es un logarítmo?

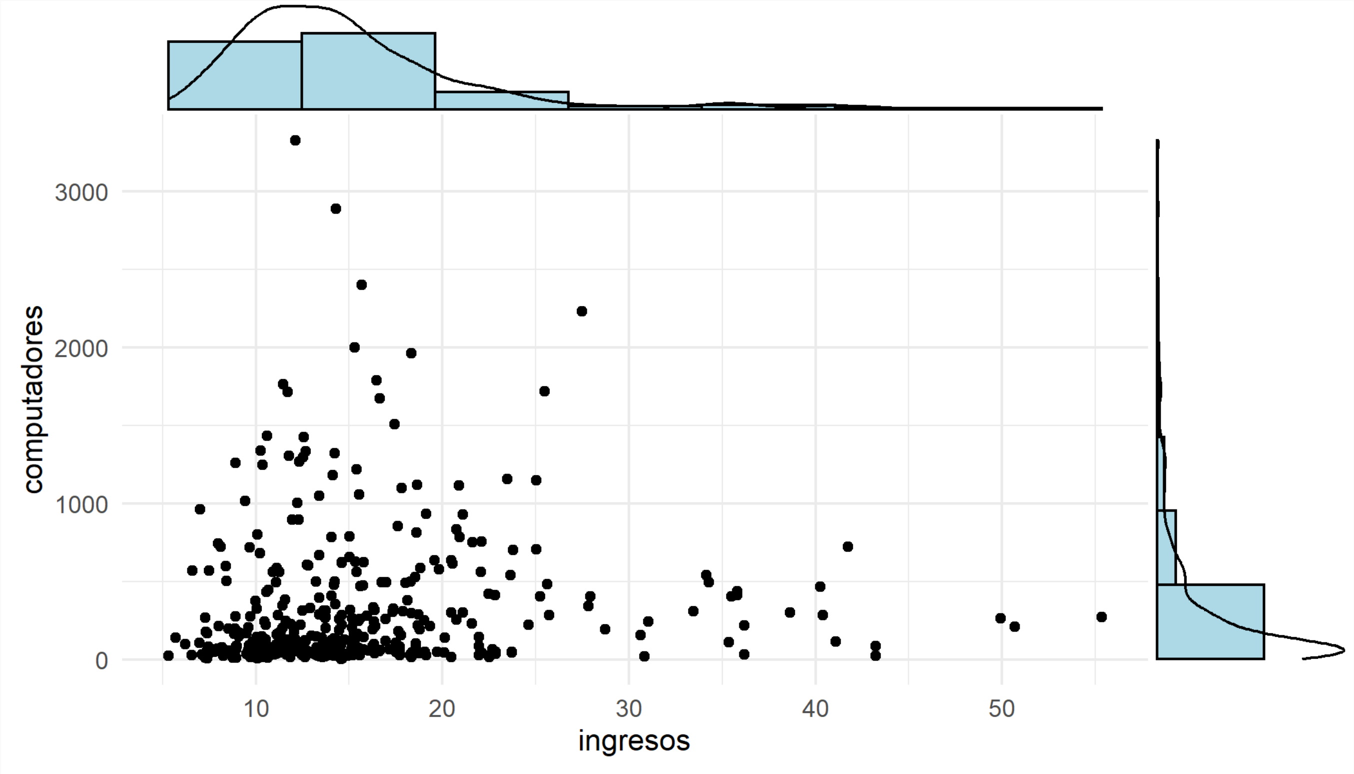
Forma exponencial	Forma logarítmica
10^2=100	log_{10}(100)=2
10^3=1000	log_{10}(1000)=3
10^4=10000	log_{10}(10000)=4
e^1=2.718...	log_{e}(2.718)=1
e^2=7.389...	log_{e}(7.389)=2
e^3=20.085...	log_{e}(20.085)=3

Transformaciones de variables

Transformación logarítmica

Modelo	Ecuación	Interpretación
Lineal-Lineal	$Y=\beta_0+\beta X$	Un cambio de una unidad en X esta asociado con un cambio de β unidades en Y
Log-Lineal	$log(Y)=\beta_0+\beta X$	Un cambio de una unidad en X esta asociado con un cambio de $(100*\beta)\%$ en Y
Lineal-Log	$Y=\beta_0+\beta log(X)$	Un cambio de 1% en X esta asociado con un cambio de $\frac{\beta}{100}$ unidades en Y
Log-Log	$log(Y)=\beta_0+\beta log(X)$	Un cambio de 1% en X esta asociado con un cambio de $\beta\%$ en Y

Transformaciones de variables



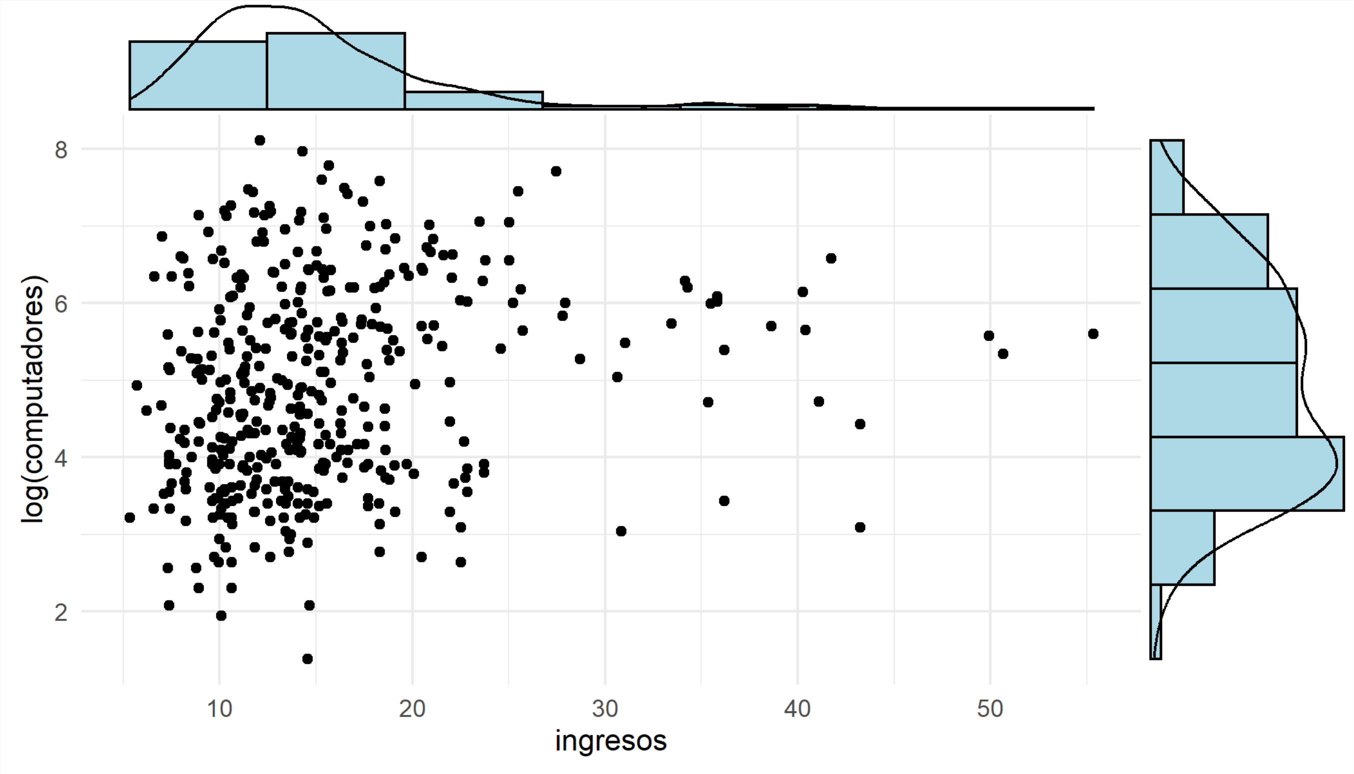

```
coef(lm(computadores ~ ingresos, data = datos_reg))
```

```
## (Intercept)    ingresos
##  221.787048    5.704378
```

Un cambio de una unidad en `ingresos` esta asociado con un cambio de 5.7 unidades en `computadores`.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformaciones de variables



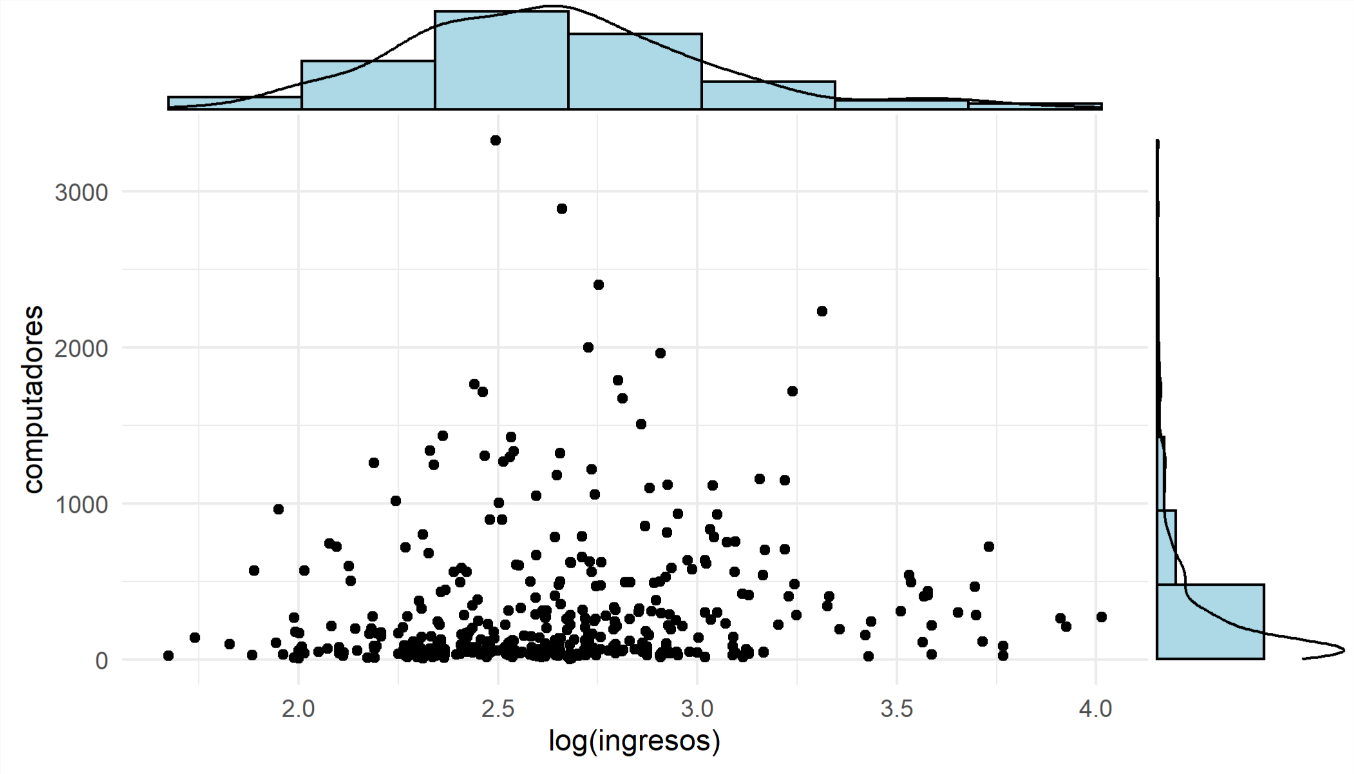

```
coef(lm(log(computadores) ~ ingresos, data = datos_reg))
```

```
## (Intercept)    ingresos
##   4.38259289  0.03453272
```

Un cambio de una unidad en `ingresos` esta asociado con un cambio de 3.45% en `computadores`.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformaciones de variables



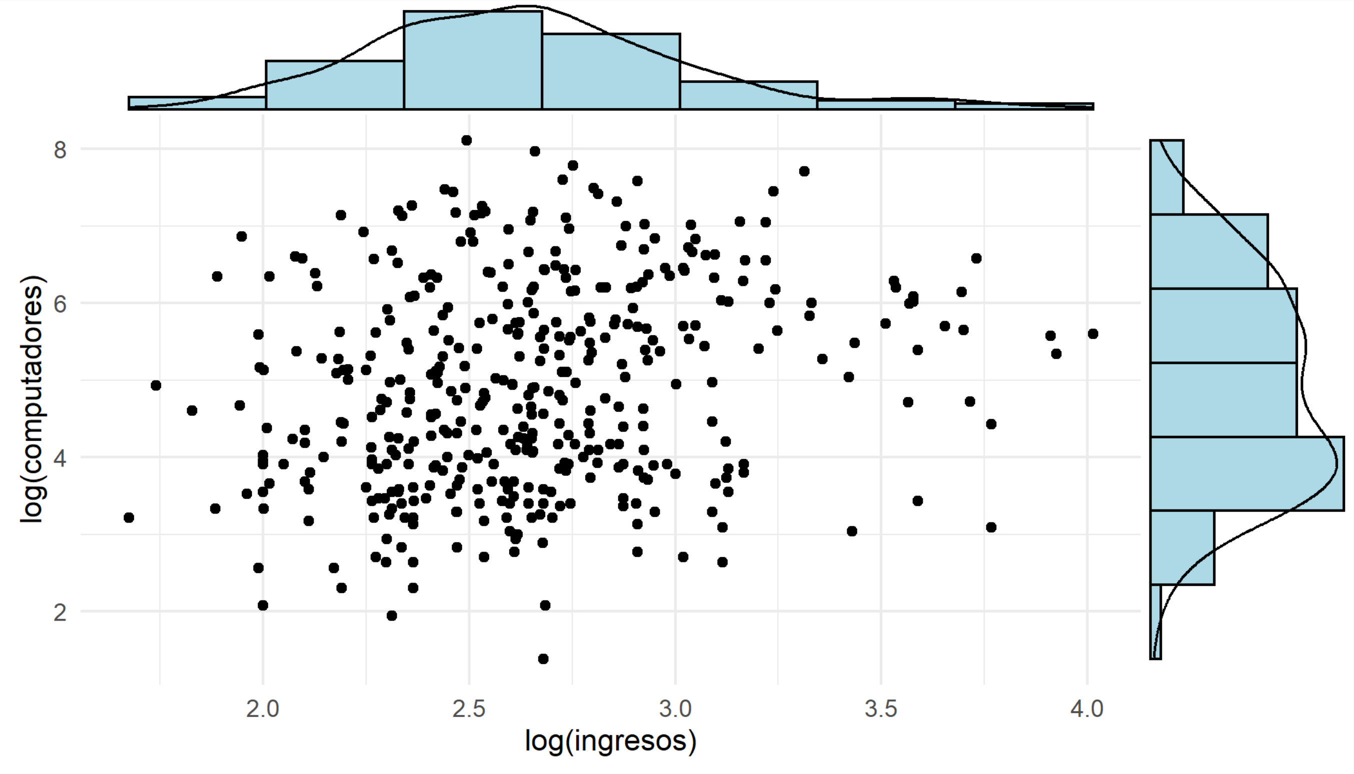

```r
coef(lm(computadores ~ log(ingresos), data = datos_reg))
```

```
##   (Intercept) log(ingresos)
##     -68.21289     142.67547
```

Un cambio de 1% en `ingresos` esta asociado con un cambio de 1.4 unidades en `computadores`.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformaciones de variables




```
coef(lm(log(computadores) ~ log(ingresos), data = datos_reg))
```

```
##   (Intercept) log(ingresos)
##     2.9946112     0.7247802
```

Un cambio de 1% en `ingresos` esta asociado con un cambio de 0.7% en `computadores`.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformaciones de variables

¿Importa la base del logaritmo?

```
coef(lm(log(computadores, base = exp(1)) ~ log(ingresos, base = exp(1)), data = datos_reg))
```

```
##                    (Intercept) log(ingresos, base = exp(1))
##                      2.9946112                    0.7247802
```

```
coef(lm(log(computadores, base = 10) ~ log(ingresos, base = 10), data = datos_reg))
```

```
##                (Intercept) log(ingresos, base = 10)
##                  1.3005431                0.7247802
```

```
coef(lm(log(computadores, base = 1234) ~ log(ingresos, base = 1234), data = datos_reg))
```

```
##                  (Intercept) log(ingresos, base = 1234)
##                    0.4207087                  0.7247802
```

Noten que el coeficiente, \hat{\beta}, no cambia.

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Juntando todo

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Censo USA 2000

```
(census <- read_csv("../datos/census2000.csv") %>%
  filter(hours > 500, income > 5000, age <60 ) %>%
  mutate(ingresos_hora = income/hours) %>%
  group_by(edad = age, sexo = sex) %>%
  summarise(ingresos_hora = mean(ingresos_hora)) %>%
   mutate(log_ingresos_hora = log(ingresos_hora)))
```

```
## # A tibble: 84 x 4
## # Groups:   edad [42]
##     edad sexo  ingresos_hora log_ingresos_hora
##    <dbl> <chr>         <dbl>             <dbl>
##  1    18 F             10.4               2.34
##  2    18 M              8.92              2.19
##  3    19 F              7.90              2.07
##  4    19 M              8.51              2.14
##  5    20 F              8.77              2.17
##  6    20 M              8.65              2.16
##  7    21 F              8.90              2.19
##  8    21 M              9.99              2.30
##  9    22 F              9.44              2.24
## 10    22 M              9.82              2.28
## # ... with 74 more rows
```

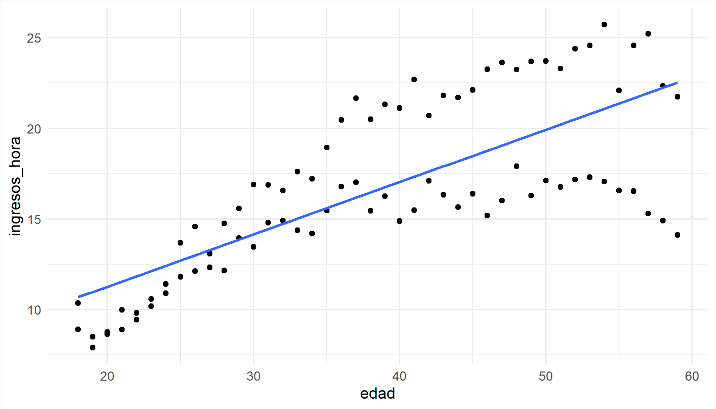
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo simple

\widehat{ingresos_ hora}=\hat{\beta_0}+\hat{\beta_1}edad

```
lm(ingresos_hora ~ edad, data = census)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)    5.49       4.83 6.17e- 6
## 2 edad           0.289     10.3  2.36e-16
```



\widehat{ingresos_ hora}=5.49+(0.289*edad)

El aumento en una unidad de `edad` esta asociado a un aumento, en promedio, de 0.289 unidades en `ingresos_hora`.

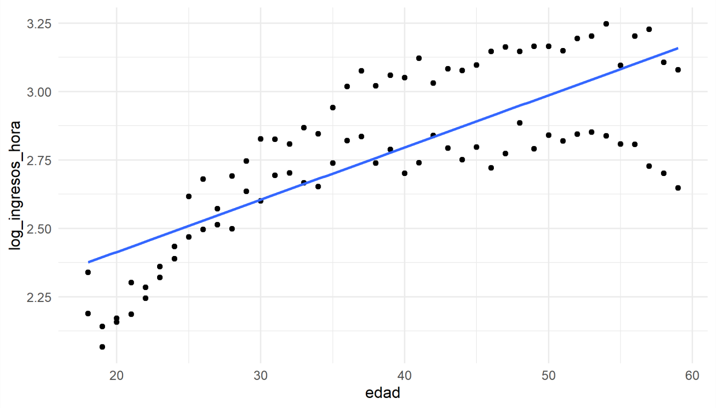
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Transformación a Y

\widehat{log(ingresos_ hora)}=\hat{\beta_0}+\hat{\beta_1}edad

```
lm(log_ingresos_hora ~ edad, data = census)
```

```
## # A tibble: 2 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)   2.03        29.1 5.11e-45
## 2 edad          0.0191      11.0 6.66e-18
```



\widehat{log(ingresos_ hora)}=2.03+(0.019*edad)

El aumento en una unidad de `edad` esta asociado a un aumento, en promedio, de 1.9% en `ingresos_hora`.

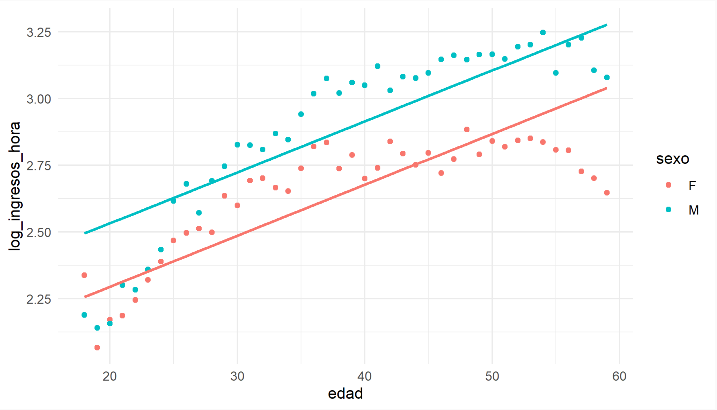
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Curvas por grupo

$$\widehat{log(ingresos\_hora)}=\hat{\beta_0}+\hat{\beta_1}edad+\hat{\beta_2}1_{M}$$

```
lm(log_ingresos_hora ~ edad + sexo, data = census)
```

```
## # A tibble: 3 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)   1.91        33.5  3.29e-49
## 2 edad          0.0191      14.1  1.80e-23
## 3 sexoM         0.237        7.22 2.48e-10
```



$$\widehat{log(ingresos\_hora)}=1.91+(0.019*edad)+(0.237*1_{M})$$

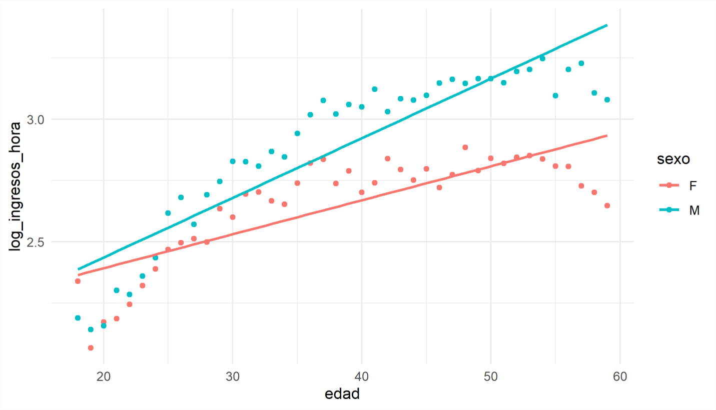
$$\begin{align} E(log(ingresos\_hora)|sexo=F)&=1.91+(0.019*edad) \\ E(log(ingresos\_hora)|sexo=M)&=2.15+(0.019*edad) \end{align}$$

Permitir pendientes distintas

$$\widehat{log(ingresos\_hora)}=\hat{\beta_0}+\hat{\beta_1}edad+\hat{\beta_2}1_{M}+\hat{\beta_3}(edad*1_{M})$$

```
lm(log_ingresos_hora ~ edad*sexo, data = census)
```

```
## # A tibble: 4 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)   2.11       30.1  2.43e-45
## 2 edad          0.0139      7.95 1.02e-11
## 3 sexoM        -0.165      -1.66 1.01e- 1
## 4 edad:sexoM    0.0105      4.24 5.90e- 5
```

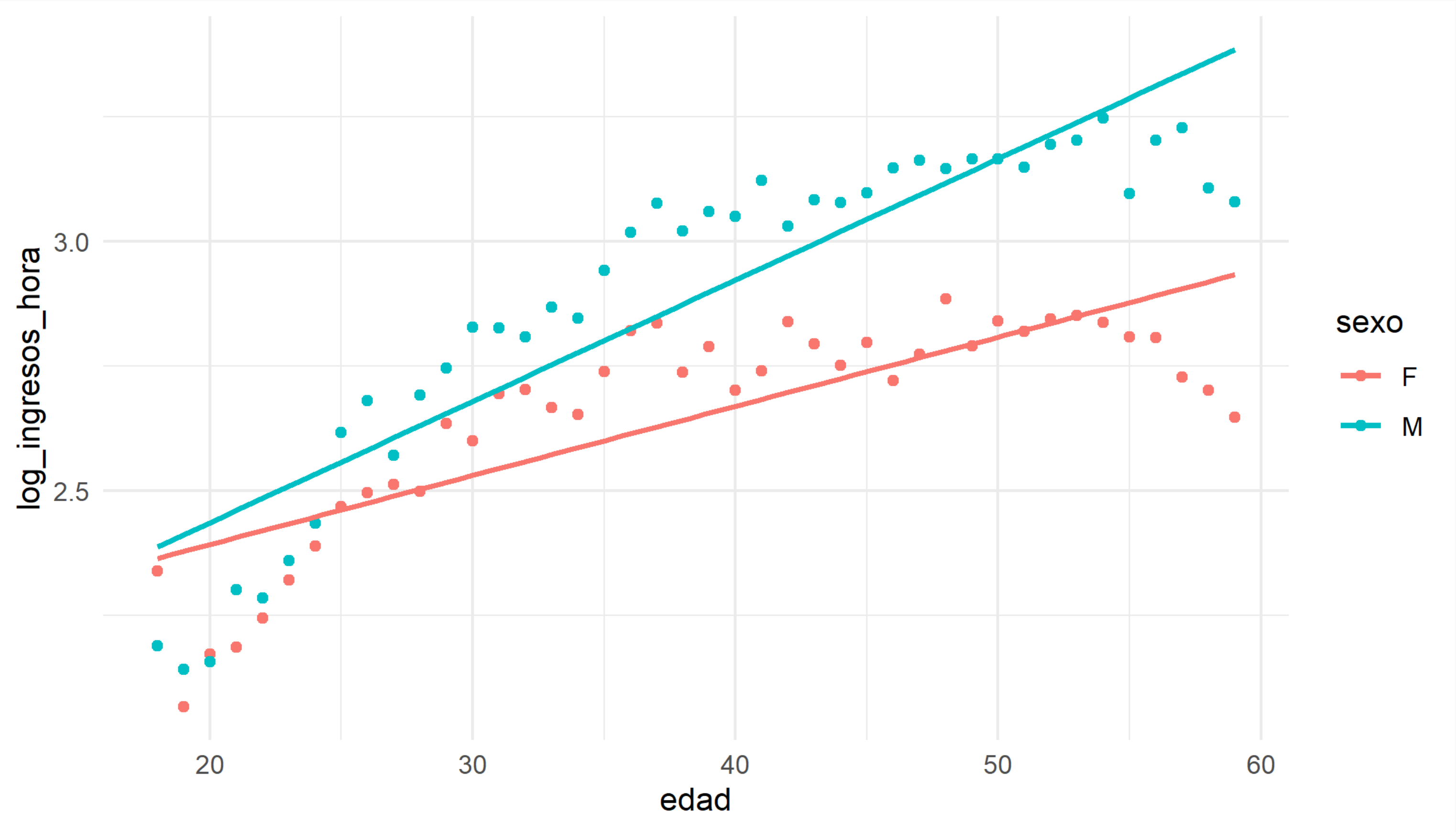


$$\widehat{log(ingresos\_hora)}=2.11+(0.014*edad)-(0.165*1_{M})+(0.01*edad*1_{M})$$

$$\begin{align} E(log(ingresos\_hora)|sexo=F)&=2.11+(0.014*edad) \\ E(log(ingresos\_hora)|sexo=M)&=1.95+(0.024*edad) \end{align}$$

Ojo, se puede visualizar directamente

```
census %>%
  ggplot(aes(x = edad, y = log_ingresos_hora, col = sexo)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal()
```



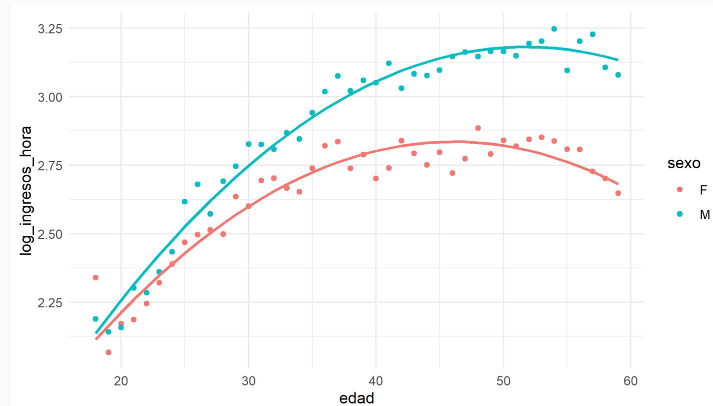
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Capturar la curvatura

\small \widehat{log(ingresos_hora)}=\hat{\beta_0}+\hat{\beta_1}edad+\hat{\beta_2}1_{M}+\hat{\beta_3}edad^2+\hat{\beta_4}(edad*1_{M})

```
lm(log_ingresos_hora ~ edad*sexo + I(edad^2), data = census)
```

```
## # A tibble: 5 x 4
##   term        estimate statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.888        12.2  6.46e-20
## 2 edad         0.0846       21.8  3.23e-35
## 3 sexoM       -0.165        -3.83 2.55e- 4
## 4 I(edad^2)   -0.000919    -18.6  1.26e-30
## 5 edad:sexoM   0.0105        9.79 2.85e-15
```



\small \widehat{log(ingresos_hora)}=0.89+(0.084*edad)-(0.165*1_{M})-(0.001*edad^2)+(0.01*edad*1_{M})

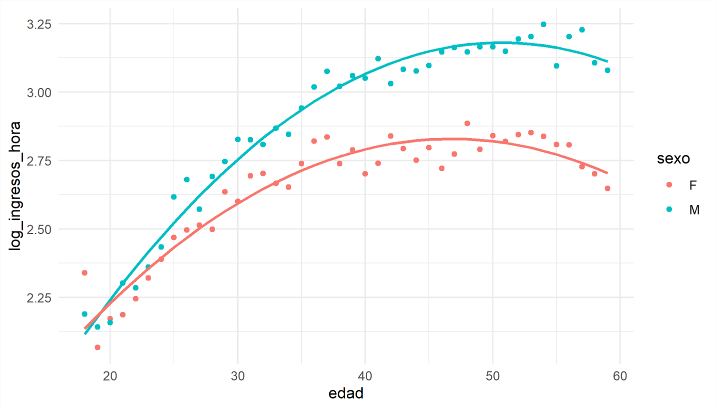
\begin{align} E(log(ingresos_hora)|sexo=F)&=0.89+(0.084*edad)-(0.001*edad^2) \\ E(log(ingresos_hora)|sexo=M)&=0.73+(0.094*edad)-(0.001*edad^2) \end{align}

Permitir curvaturas distintas

\small \widehat{log(ingresos_hora)}=\hat{\beta_0}+\hat{\beta_1}edad+\hat{\beta_2}1_{M}+\hat{\beta_3}edad^2+\hat{\beta_4}(edad*1_{M})+\hat{\beta_5}(edad^2*1_{M})

```
lm(log_ingresos_hora ~ edad*sexo + I(edad^2)*sexo, data = census)
```

```
## # A tibble: 6 x 4
##   term             estimate statistic  p.value
##   <chr>               <dbl>     <dbl>    <dbl>
## 1 (Intercept)      0.994        10.2  4.28e-16
## 2 edad             0.0785       14.6  4.81e-24
## 3 sexoM           -0.377        -2.75 7.44e- 3
## 4 I(edad^2)       -0.000839    -12.2  1.11e-19
## 5 edad:sexoM       0.0227        2.99 3.77e- 3
## 6 sexoM:I(edad^2) -0.000159     -1.62 1.08e- 1
```



\small \begin{align} \widehat{log(ingresos_hora)}=&0.99+(0.078*edad)-(0.377*1_{M})-(0.001*edad^2)+(0.02*edad*1_{M})\\&-(0.0001*edad^2*1_{M}) \end{align}

\small \begin{align} E(log(ingresos_hora)|sexo=F)&=0.99+(0.078*edad)-(0.001*edad^2) \\ E(log(ingresos_hora)|sexo=M)&=0.61+(0.098*edad)-(0.0099*edad^2) \end{align}

¿Qué modelo elegir?

R^2_{adj}

```r
glance(reg1) %>% select(2)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.593
```

```r
glance(reg2) %>% select(2)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.750
```

```r
glance(reg3) %>% select(2)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.793
```

```r
glance(reg4) %>% select(2)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.961
```

```r
glance(reg5) %>% select(2)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.962
```

Estadístico F

$$F=\frac{(SCR_R-SCR_{SR})/q}{SCR_{SR}/(n-k-1)}=\frac{(R^2_{SR}-R^2_{R})/q}{(1-R^2_{SR})/(n-k-1)}$$

El estadístico F es una forma de comparar **pares** de modelos. Se compara un modelo **Restringido (R)** y con un modelo **Sin Restricción (SR)**.

- Modelo SR: $\hat{Y}=\hat{\beta_0}+\hat{\beta_1}X_1+\hat{\beta_2}X_2$
- Modelo R: $\hat{Y}=\hat{\beta_0}+\hat{\beta_1}X_1$

Esta prueba entrega su respectivo **p.value** considerando una hipótesis nula, H_0, del tipo *"El modelo con más variables (SR) no agrega información útil en comparación al modelo R"*.

Estadístico F

```r
anova(reg1, reg2)
```

```
## Analysis of Variance Table
##
## Model 1: log_ingresos_hora ~ edad
## Model 2: log_ingresos_hora ~ edad + sexo
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     82 3.0226
## 2     81 1.8382  1    1.1845 52.194 2.477e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(reg2, reg3)
```

```
## Analysis of Variance Table
##
## Model 1: log_ingresos_hora ~ edad + sexo
## Model 2: log_ingresos_hora ~ edad * sexo + sexo
##   Res.Df    RSS Df Sum of Sq  F    Pr(>F)
## 1     81 1.8382
## 2     80 1.5005  1   0.33762 18 5.896e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(reg3, reg4)
```

```
## Analysis of Variance Table
##
## Model 1: log_ingresos_hora ~ edad * sexo + sexo
## Model 2: log_ingresos_hora ~ edad * sexo + sexo + I(edad^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     80 1.50054
## 2     79 0.27843  1    1.2221 346.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(reg4, reg5)
```

```
## Analysis of Variance Table
##
## Model 1: log_ingresos_hora ~ edad * sexo + sexo + I(edad^2)
## Model 2: log_ingresos_hora ~ edad * sexo + sexo + I(edad^2) * sexo
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     79 0.27843
## 2     78 0.26932  1 0.0091163 2.6403 0.1082
```

Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Criterios de información

Los criterios de información son medidas de calidad relativa para nuestros modelos estadísticos. En general, hablamos de dos: *Akaike Information Criterion* (AIC) y *Bayesian Information Criterion* (BIC). **Valor más bajo es mejor**.

```
(BIC ← data.frame(reg1 = glance(reg1) %>% pull(BIC), reg2 = glance(reg2) %>% pull(BIC),
                  reg3 = glance(reg3) %>% pull(BIC), reg4 = glance(reg4) %>% pull(BIC),
                  reg5 = glance(reg5) %>% pull(BIC)))
```

```
##        reg1      reg2      reg3      reg4      reg5
## 1 -27.59987 -64.94733 -77.56358 -214.6226 -212.9881
```

Estos indicadores se pueden modelar probabilísticamente.

$$\small P(M_i) \approx \frac{e^{-\frac{1}{2}[BIC(M_i-BIC_{min})]}}{\sum_{r=1}^{R}e^{-\frac{1}{2}[BIC(M_r-BIC_{min})]}}$$

```
probs ← exp(-0.5*(BIC-min(BIC)))/sum(exp(-0.5*(BIC-min(BIC))))
round(probs, 3)
```

```
##   reg1 reg2 reg3  reg4  reg5
## 1    0    0    0 0.694 0.306
```

BIC nos da el mismo resultado que el **estadístico F**.

Análisis de residuales

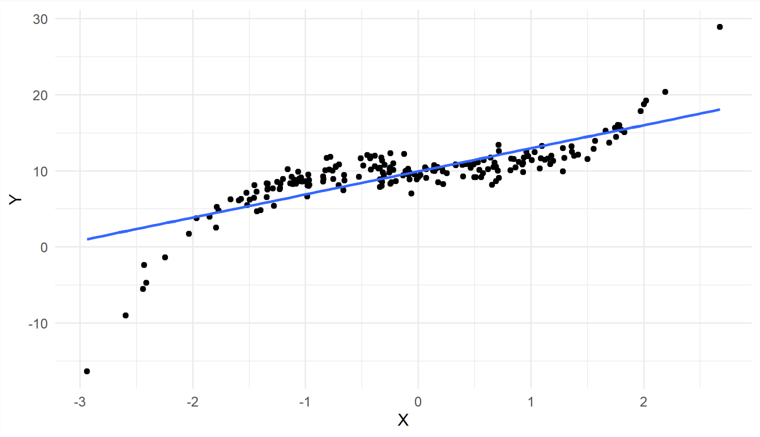
```
augment(reg0)
```

```
## # A tibble: 84 x 8
##    ingresos_hora  edad .fitted .resid .std.resid    .hat .sigma  .cooksd
##            <dbl> <dbl>   <dbl>  <dbl>      <dbl>   <dbl>  <dbl>    <dbl>
##  1         10.4     18    10.7 -0.320     -0.105 0.0460   3.15 0.000264
##  2          8.92    18    10.7 -1.77      -0.579 0.0460   3.14 0.00806
##  3          7.90    19    11.0 -3.08      -1.01  0.0427   3.13 0.0226
##  4          8.51    19    11.0 -2.47      -0.807 0.0427   3.13 0.0145
##  5          8.77    20    11.3 -2.49      -0.814 0.0396   3.13 0.0137
##  6          8.65    20    11.3 -2.62      -0.854 0.0396   3.13 0.0150
##  7          8.90    21    11.6 -2.66      -0.865 0.0367   3.13 0.0143
##  8          9.99    21    11.6 -1.56      -0.509 0.0367   3.14 0.00493
##  9          9.44    22    11.8 -2.40      -0.782 0.0340   3.14 0.0108
## 10          9.82    22    11.8 -2.03      -0.659 0.0340   3.14 0.00765
## # … with 74 more rows
```

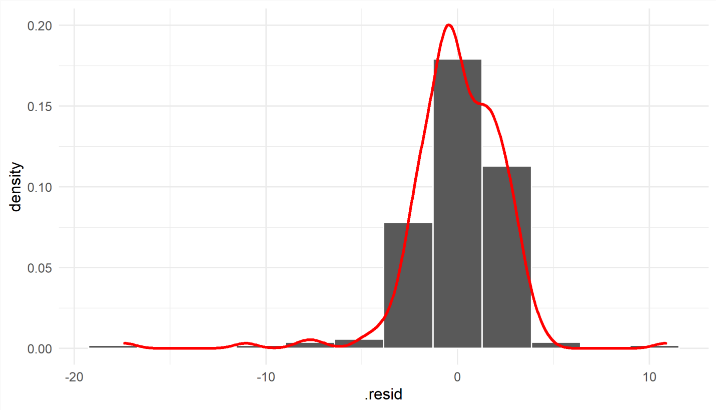
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

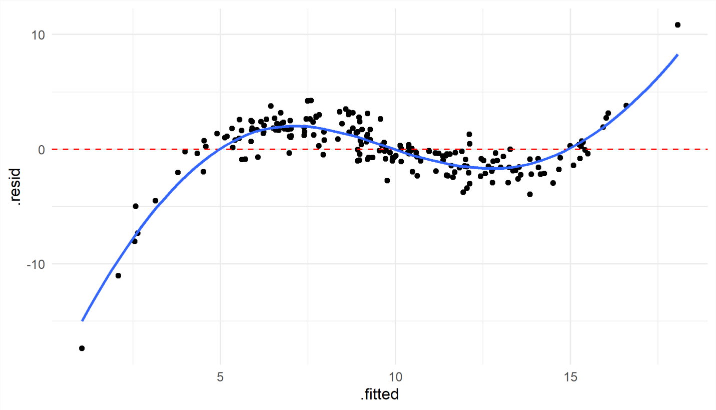
No tan buen ajuste

Modelo



Análisis residuales

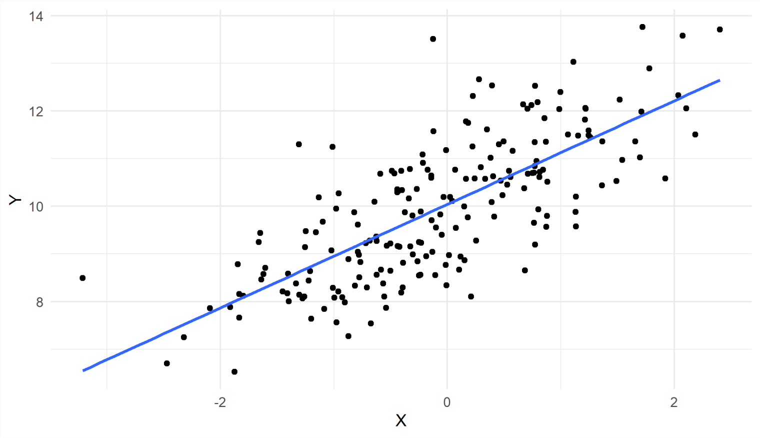




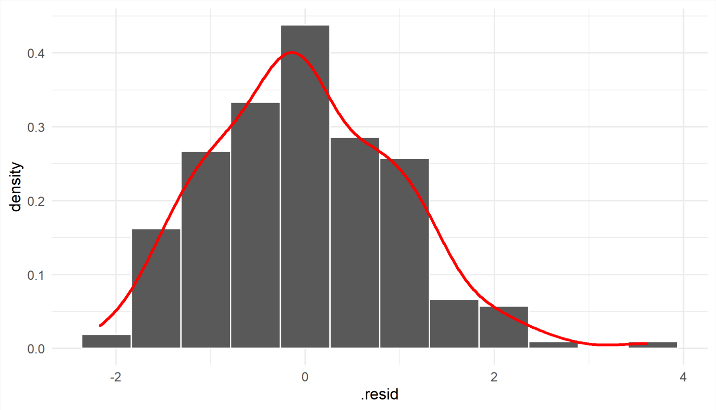
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

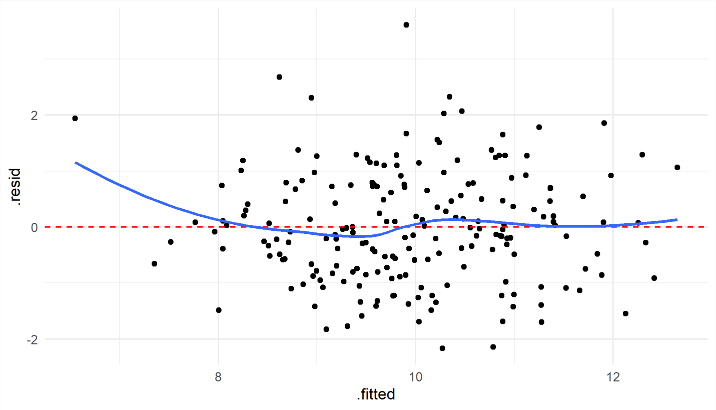
Mejor ajuste

Modelo



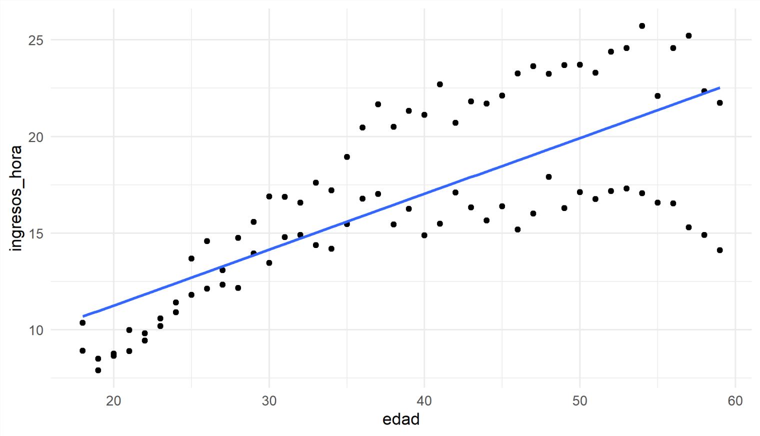
Análisis residuales



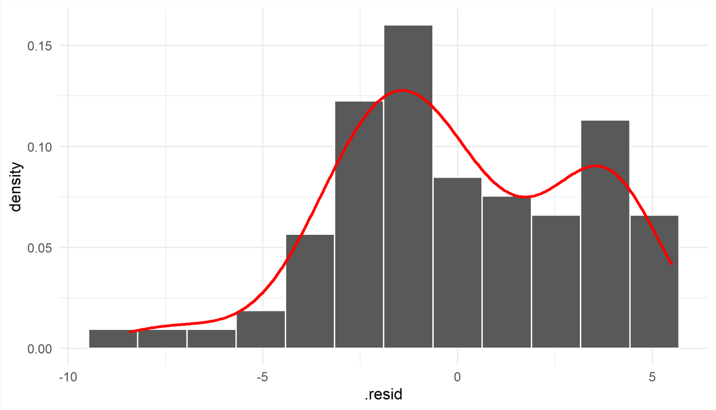


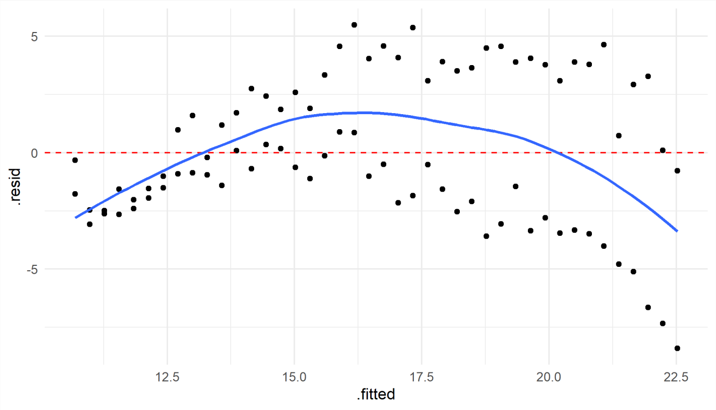
Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Modelo



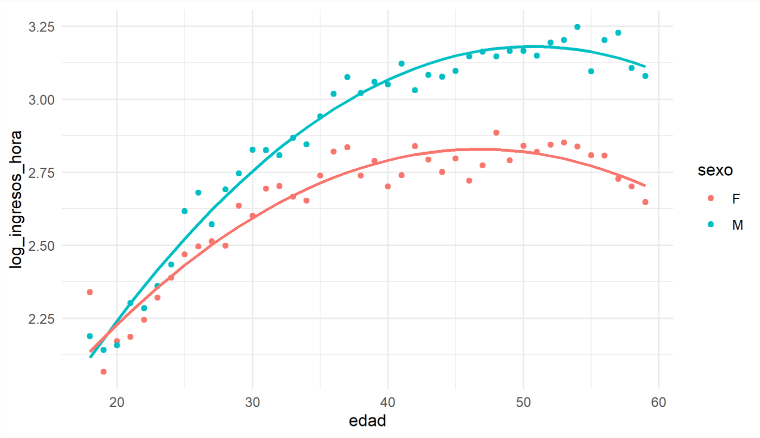
Análisis residuales



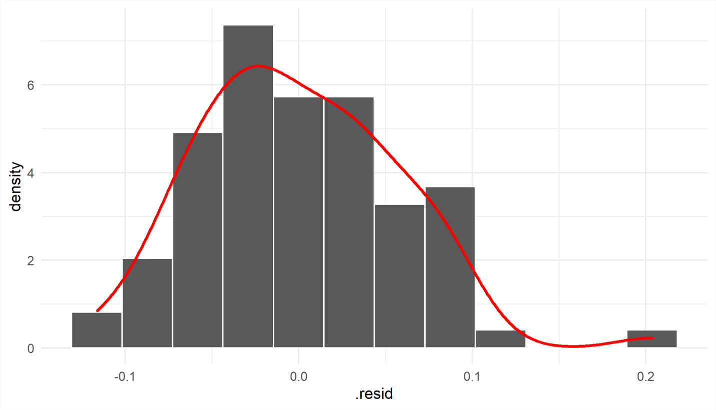


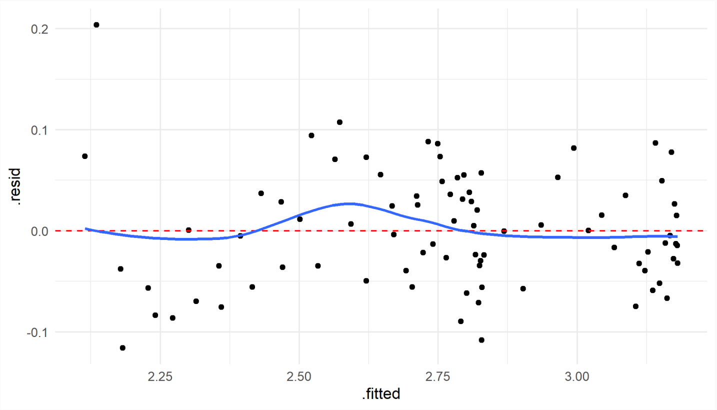
Último modelo ingresos_hora/edad

Modelo



Análisis residuales





Loading [MathJax]/jax/element/mml/optable/BasicLatin.js

Para concluir

Una idea general

- Dada una verdad: $Y = \beta_0 + \beta_1X + \epsilon$
- Realizamos una estimación: $\hat{Y} = \hat{\beta_0} + \hat{\beta_1}X$

$$Datos \rightarrow Cálculos \rightarrow Estimación \xrightarrow[]{si\ todo\ sale\ bien} Verdad$$

$$X,Y \rightarrow (X'X)^{-1}X'Y \rightarrow \hat{\beta} \xrightarrow[]{si\ todo\ sale\ bien} \beta$$

Próxima clase

- Modelo Logit
- Casos prácticos