Estadística *descriptiva* o análisis exploratorio de datos

Primera parte
Medidas resumen.

¿Qué es la estadística?

- La ciencia de recolectar, describir y analizar datos.
  (de libro de Lock, "Unlocking the power of data")

Muestra - Datos (Observaciones)

- Muestra aleatoria: $X_1, \ldots, X_n$: variables aleatorias i.i.d.

  Datos-Observaciones: son realizaciones de las variables aleatorias

- Datos - Observaciones $x_1, \ldots, x_n$: números.

  Datos-Observaciones: son los resultados obtenidos al realizar el "experimento"

Departamentos en venta en Buenos Aires en 2016

Fuente: https://data.buenosaires.gob.ar/

```
> departamentos<-read.csv("departamentos.csv",
+                         header = TRUE)
> dim(departamentos)
[1] 7564    5
> departamentos[1:10,]
    M2    M2CUB  DOLARES  U_S_M2
1   57    50     170150   3403
2   46    46     118650   2579
3   61    56     181470   3241
4  140    76     320000   4211
5   39    33     82116    2488
6   39    34     81921    2409
7   50    45     103802   2307
8   82    58     265000   4569
9   38    38     145000   3816
10  38    35     147000   4200
```

Departamentos en venta en Buenos Aires en 2016

- $X = m^2$ de un departamento elegido al azar de entre todos los departamentos en venta en Buenos Aires en 2016

$$X \sim F$$

- Experimento: Elegir n departamentos al azar. Tengo variables aleatorias

$$X_1, \ldots X_n$$

$X_i = m^2$ del $i-$ ésimo departamento elegido.

$$X_i \sim F \text{ i.i.d}$$

Departamentos en venta en Buenos Aires en 2016

Tenemos vectores aleatorios

$$(X_1, Y_1, Z_1, W_1) \dots (X_n, Y_n, Z_n, W_n)$$

$X_i = m^2$ del i-ésimo departamento
$Y_i = m^2$ cubiertos del i-ésimo departamento
$Z_i =$ precio en dólares del i-ésimo departamento
$W_i =$ precio en dólares por m^2 del i-ésimo departamento

El conjunto de datos es una realización de estos vectores aleatorios, para $n = 7564$.

Algunos datos

- Considere los siguientes datos, generados por una distribución F.

$$165.03\ ,\ 162.37\ ,\ 156.67\ ,\ 167.84\ ,\ 172.47$$

- Recordar $F(t) = P(X \leq t)$, donde $X \sim F$
- Estime $F(160)$.
- Estime $F(168)$.
- Proponga una fórmula para estimar $F(t)$

$$\widehat{F}(t) = \ldots\ldots\ldots$$

- Grafique la función $\widehat{F} : \mathbb{R} \to [0, 1]$.

Algunos datos

- Considere los siguientes datos, generados por una distribución F.

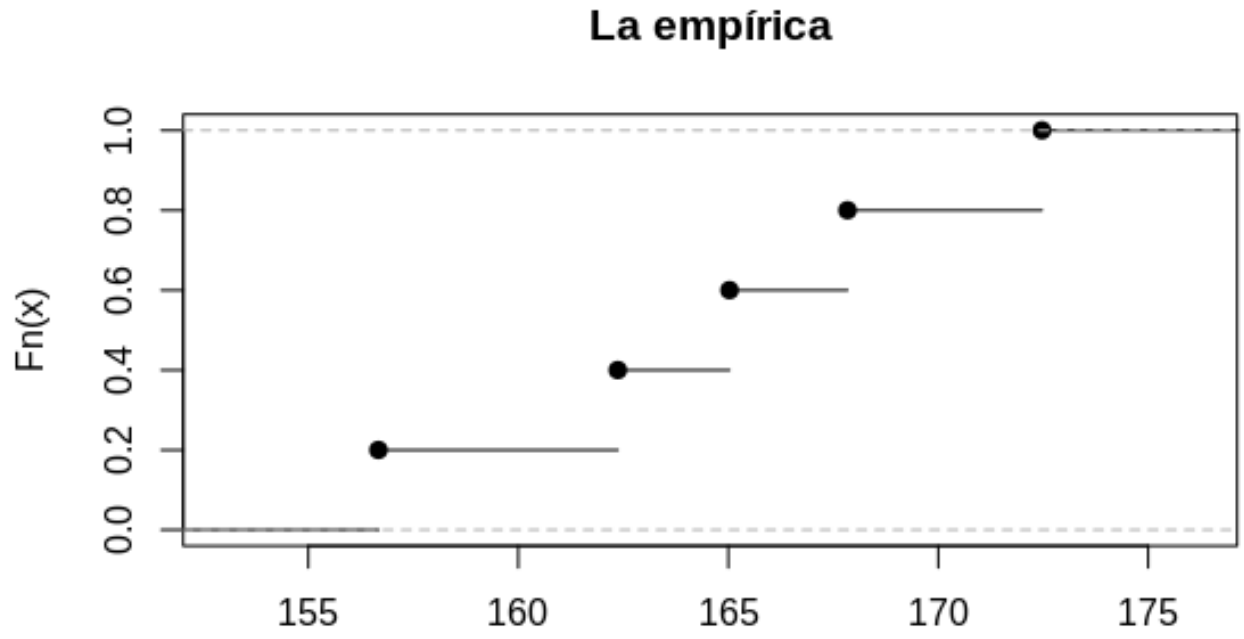
$$165.03\ ,\ 162.37\ ,\ 156.67\ ,\ 167.84\ ,\ 172.47$$

- Recordar $F(t) = P(X \leq t)$, donde $X \sim F$
- Estime $F(160)$.
- Estime $F(168)$.
- Proponga una fórmula para estimar $F(t)$

$$\widehat{F}(t) = \frac{1}{5} \sum_{i=1}^{5} I\{x_i \leq t\}.$$

- Grafique la función $\widehat{F} : \mathbb{R} \to [0, 1]$.

$$\widehat{F}(t) = \frac{1}{5}\sum_{i=1}^{5} I\{x_i \leq t\}.$$



valores	156.67	162.37	165.03	167.84	172.47
puntual	1/5	1/5	1/5	1/5	1/5

Función de distribución empírica

- Datos - Observaciones generadas con F:

$$x_1, \ldots, x_n$$

- Acumulada F. $F(t) = \mathbb{P}(X \leq t)$.
- Estimación de la acumulada : función de distribución empírica asociada a $x_1, \ldots x_n$

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\{x_i \leq t\} \qquad \texttt{mean(datos<=t)}$$

Función de distribución empírica

- Datos - Observaciones generadas con F:

$$x_1, \ldots, x_n$$

- Acumulada F. $F(t) = \mathbb{P}(X \leq t)$.
- Estimación de la acumulada : función de distribución empírica asociada a $x_1, \ldots x_n$

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\{x_i \leq t\} \qquad \texttt{mean(datos<=t)}$$

- $\widehat{F}_n$ es una función de distribución acumulada (de discreta).
- $\widehat{F}_n$ asigna peso $1/n$ a cada valor $x_1, \ldots, x_n$.

valores	x_1	$\ldots$	$\ldots$	$\ldots$	x_n
puntual	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$

La media muestral

Sea $X \sim \widehat{F}_n$ entonces
$R_X = \{x_1 \dots x_n\}$ y

$$E(X) = \sum_{x_i \in R_X} x_i p_X(x_i)$$

$$= \sum_{x_i \in R_X} x_i \frac{1}{n}$$

$$= \overline{x}_n$$

Notación: $E_F(X)$ quiere decir la esperanza de X, donde $X \sim F$.

- **Media poblacional:** $E_F(X)$
- **Media muestral:** $E_{\hat{F}_n}(X)$

La mediana

Si $X \sim F$

$$\text{Med}(X) = F^{-1}(0.5)$$

¿Qué pasa si X es discreta, o, en general, si F no es inversible?
Inversa generalizada:

$$F^{-1}(p) = \inf\{x / F(x) \geq p\}$$

- **Mediana poblacional**: $F^{-1}(0.5)$
- **Mediana muestral**: $\hat{F}_n^{-1}(0.5) = x_{([(n+1)/2])}$
  donde $x_{(1)} \leq \ldots \leq x_{(n)}$
- Otra forma de definir la mediana muestral:
  - n impar $x_{(\frac{n+1}{2})}$
  - n par: $\frac{1}{2}\{x_{(n/2)} + x_{(n/2+1)}\}$

Cuantiles y percentiles

- **Cuantil α (poblacional)**: $F^{-1}(\alpha)$ es el valor v tal que $P(X \leq v) = \alpha$
- **Cuantil α muestral**: $\hat{F}_n^{-1}(\alpha) = x_{([\alpha(n+1)])}$
- Por ejemplo, si $\alpha = 0.8$, $n = 99$, el cuantil muestral 0.8 es la observación que ocupa el lugar 80 en la muestra ordenada. Estima: v tal que $P(X \leq v) = 0.8$.
- Noción equivalente: percentil
  cuantil 0.8 $=$ percentil 80
  cuantil $\alpha =$ percentil $\alpha 100$

Medidas de resumen - *Posición*

- Media muestral $\overline{x}$
- Mediana muestral $\widetilde{x}$:
- Cuantil α muestral: $x_{([\alpha(n+1)])}$
- Cuartiles: Primero: $Q_1 = \widehat{F}_n^{-1}(0.25)$ o $x_{([0.25(n+1)])}$; Segundo: Q_2 mediana; Tercero: $Q_3 = \widehat{F}_n^{-1}(0.75)$ o $x_{([0.75(n+1)])}$
- Media α- podada:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

Medidas de resumen - *Dispersión*

- Estimación de la varianza a partir de la empírica

$$\widehat{\sigma^2} = V_{\hat{F}_n}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- **Varianza muestral**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Distancia intercuartil. $Q_3 - Q_1$
- MAD: mediana$\{|x_i - \widetilde{x}|\}$

Datos de departamentos. Medidas de posición

```
> mean(departamentos$M2)
[1] 70.3834
> median(departamentos$M2)
[1] 54
> quantile(departamentos$M2)
0%  25%  50%  75% 100%
15   41   54   80  730
> mean(departamentos$M2,0.1)
[1] 60.59468
```

Datos de departamentos. Medidas de dispersión

```
> mean(departamentos$M2^2)-mean(departamentos$M2)^2
[1] 2632.525
> mean((departamentos$M2-mean(departamentos$M2))^2)
[1] 2632.525
> var(departamentos$M2)
[1] 2632.873
> sum((departamentos$M2-mean(departamentos$M2))^2)/(n-1)
[1] 2632.873
> IQR(departamentos$M2)
[1] 39
> mad(departamentos$M2, constant = 1)
[1] 16
```

Segunda parte
Gráficos.

Histograma: AREA=FRECUENCIA RELATIVA

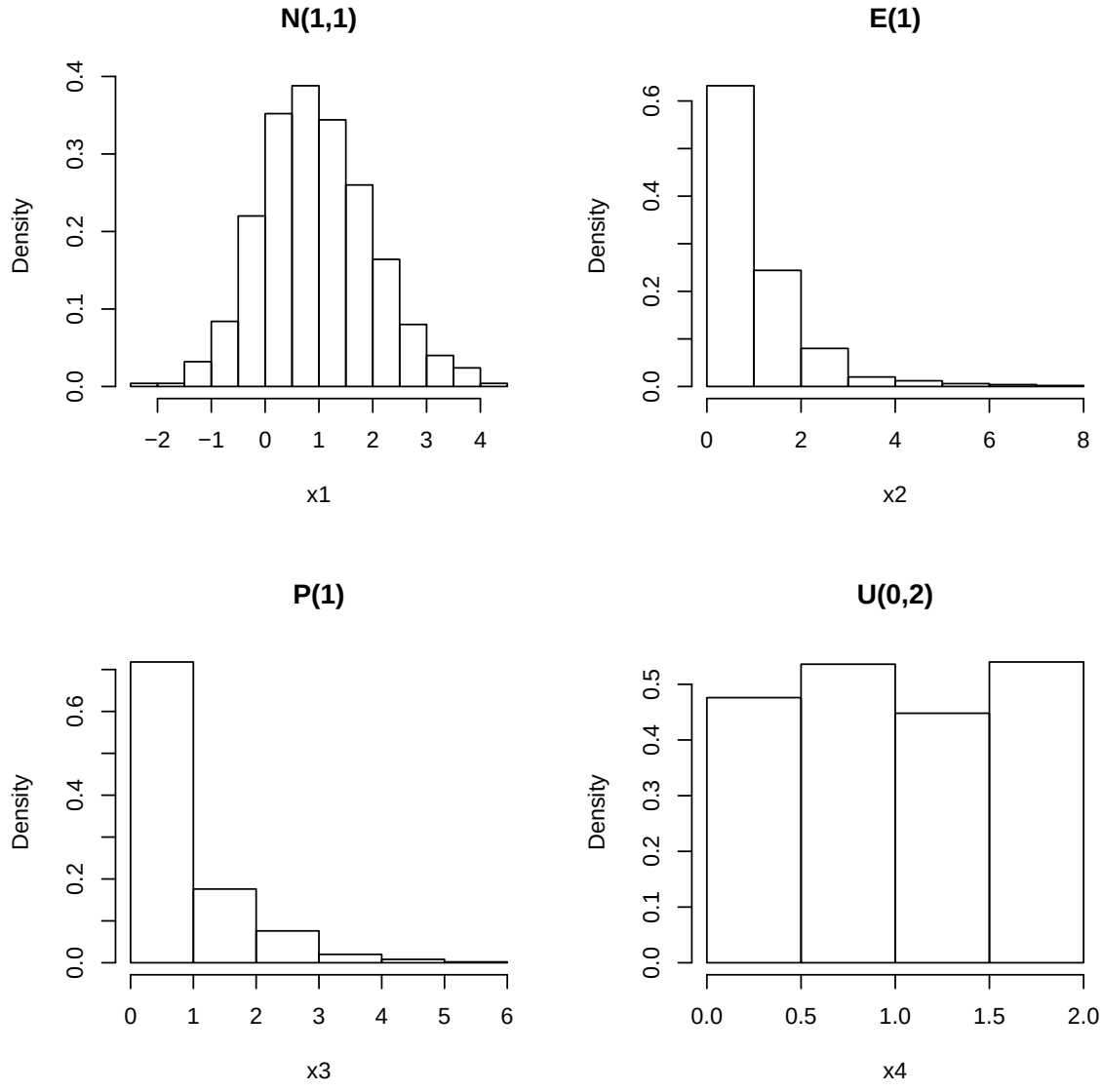
1. datos: $x_1, \ldots, x_n$
2. datos ordenados: $x_{(1)} \leq \ldots, \leq x_{(n)}$
3. $I_1, \ldots I_K$, K intervalos que particionan $[x_{(1)}, x_{(n)}]$
4. Graficamos una constante sobre cada intervalo de forma tal que el area del rectángulo coincida con la frecuencia relativa en el intervalo:

$$\text{Altura sobre } I_j \times \text{longitud } (I_j) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \in I_j}$$

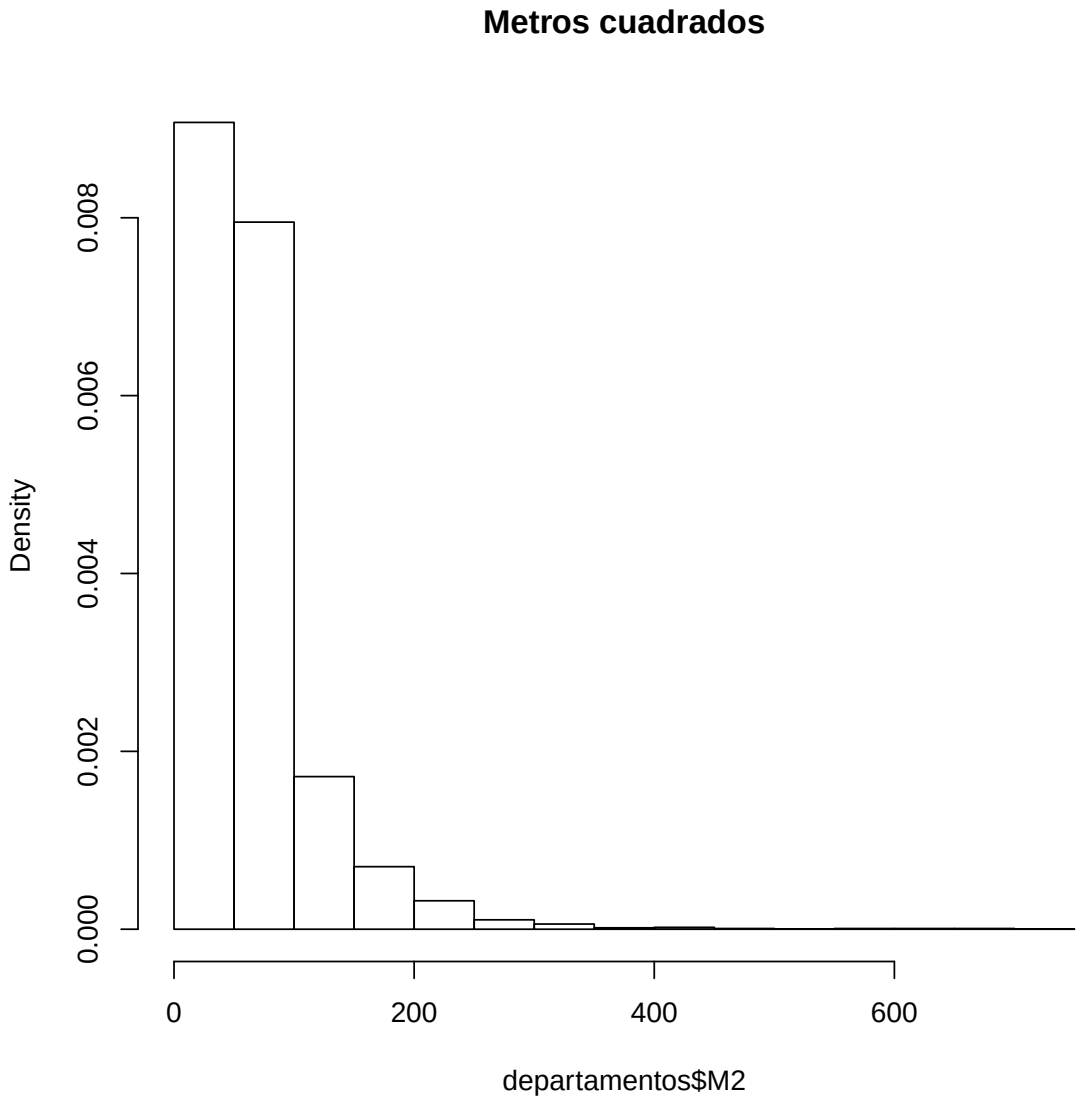
$$\text{Si } I_j = (a, b), \text{ entonces longitud } (I_j) = |I_j| = b - a$$

$$\text{Altura sobre } I_j = \frac{1}{|I_j|} \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \in I_j}$$

Histogramas de distribuciones conocidas



Datos de departamentos

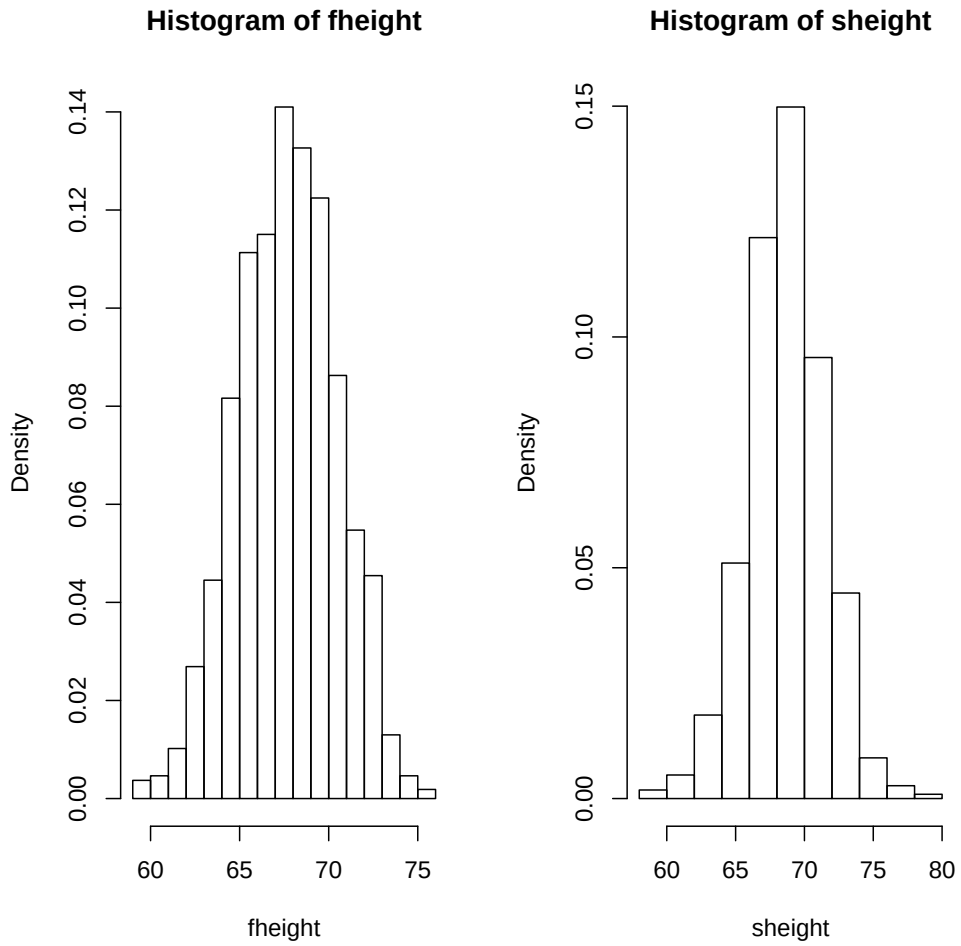


Datos de Pearson: estaturas de padres e hijos

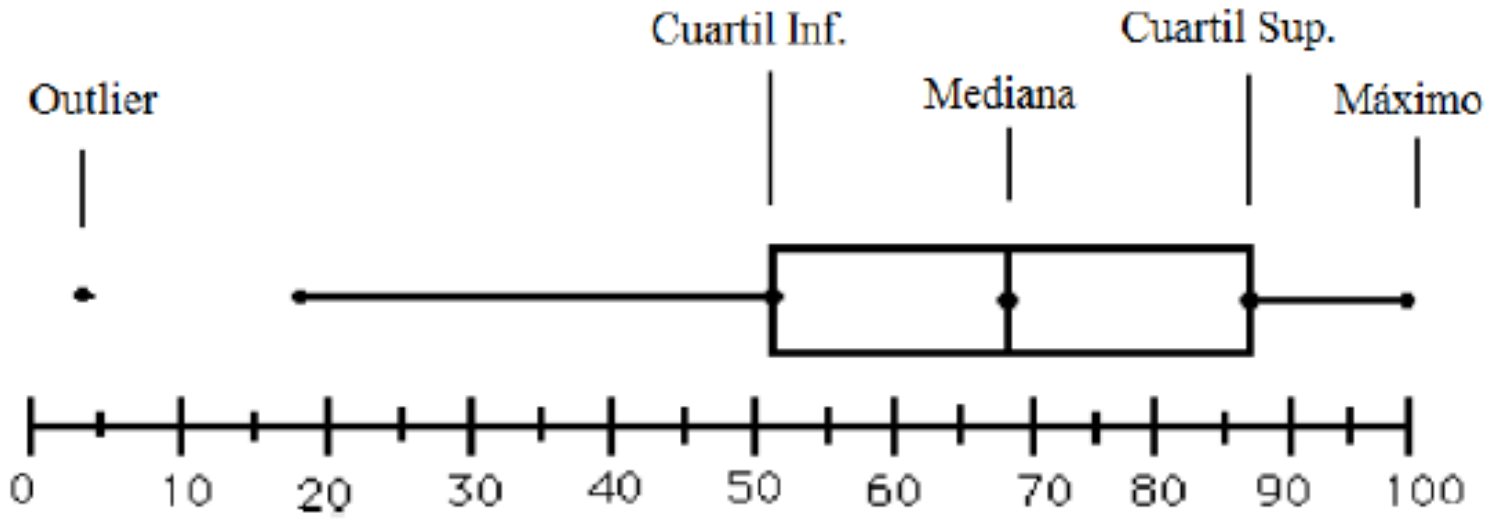
```
> library(UsingR)
> father.son[1:10,]
fheight  sheight
1  65.04851 59.77827
2  63.25094 63.21404
3  64.95532 63.34242
4  65.75250 62.79238
5  61.13723 64.28113
6  63.02254 64.24221
7  65.37053 64.08231
8  64.72398 63.99574
9  66.06509 64.61338
10 66.96738 63.97944
```

Datos de Pearson: estaturas de padres e hijos

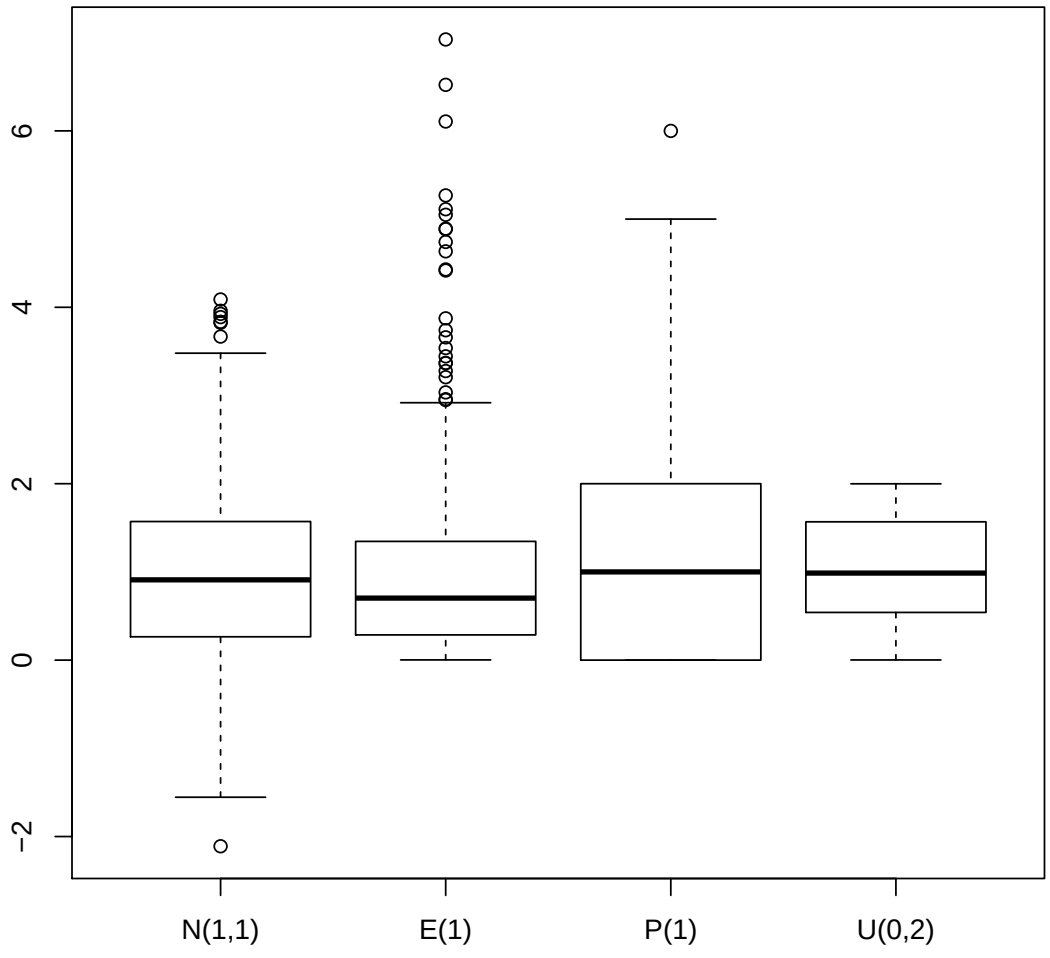
```
par(mfrow=c(1,2))
with(data=father.son, hist(fheight, probability = TRUE))
with(data=father.son, hist(sheight, probability = TRUE))
```



Boxplot - en R boxplot(datos)



Boxplots de distribuciones conocidas



Datos de departamentos

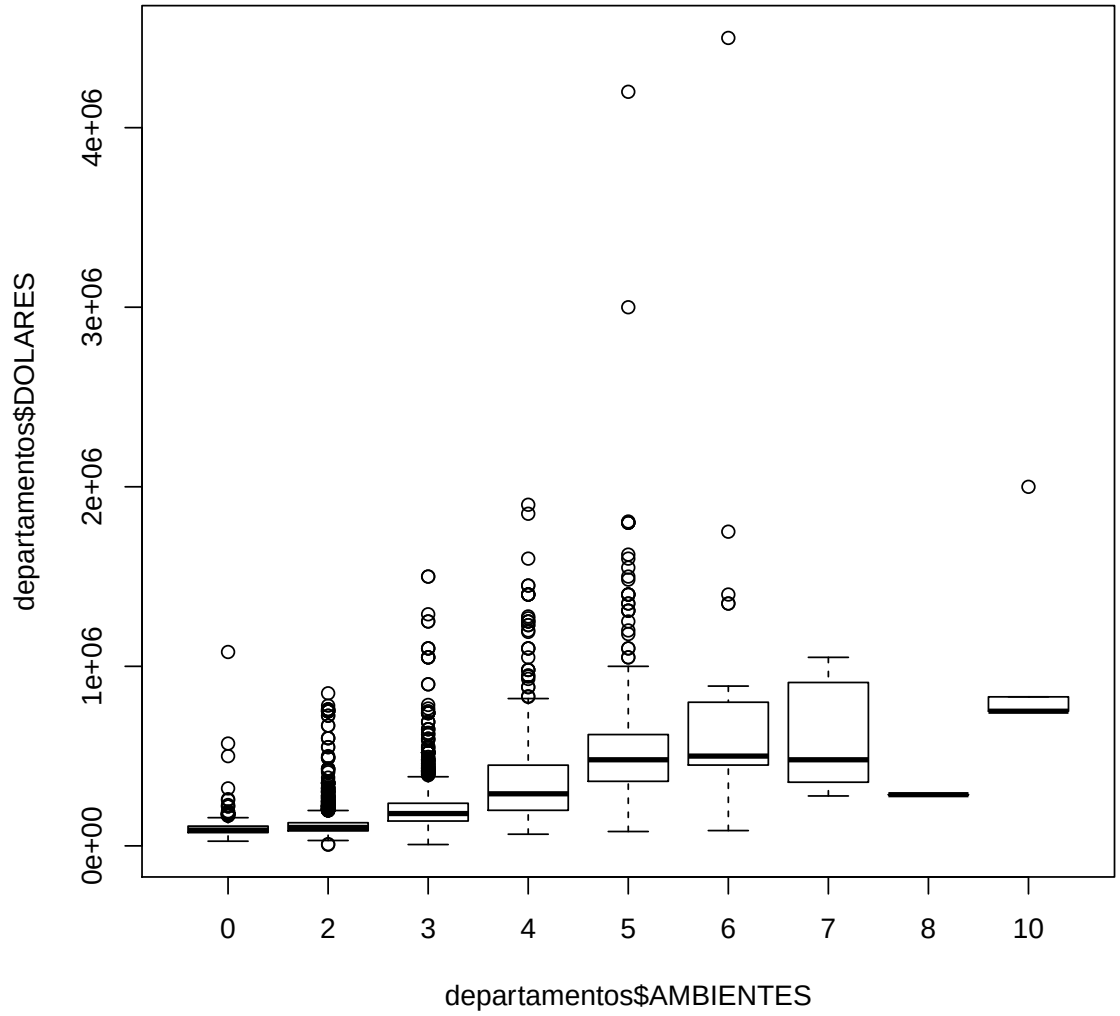
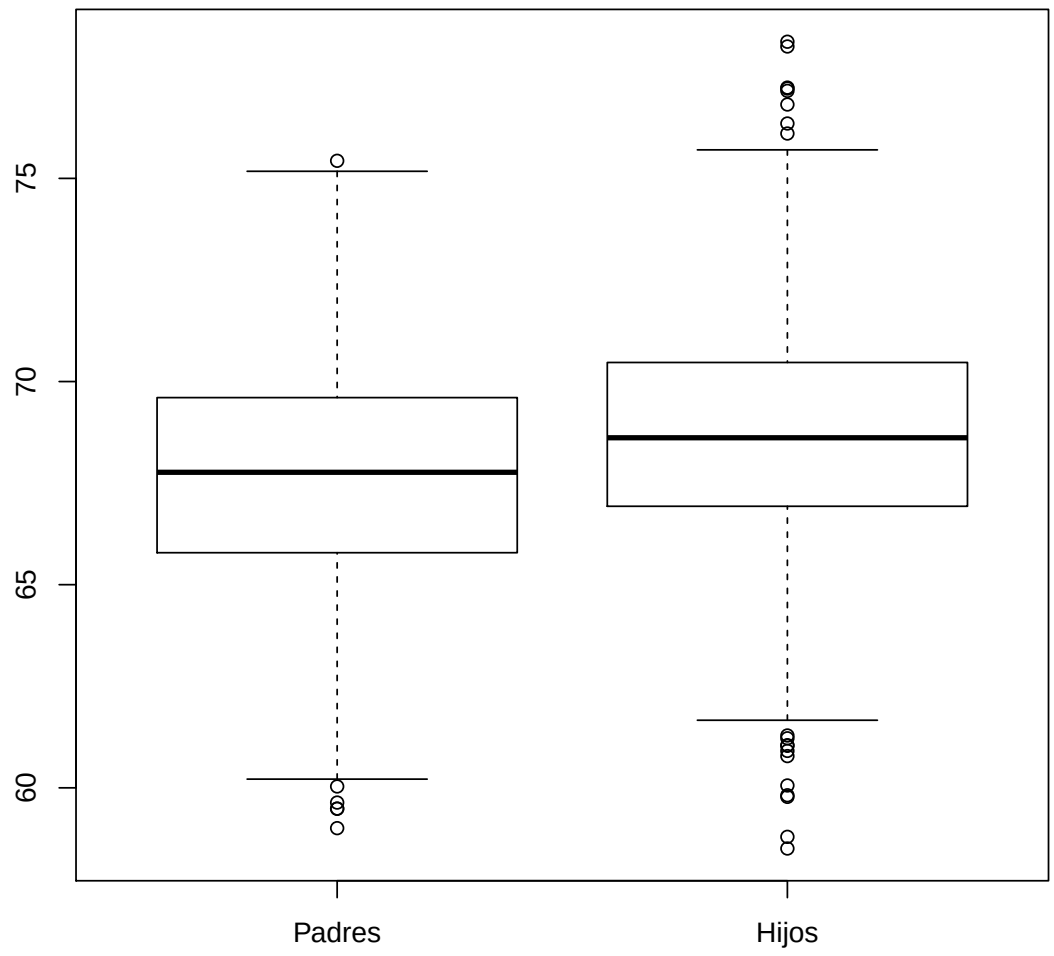


Figure:

Datos de Pearson: estatura de padres e hijos

```
boxplot(father.son$fheight,father.son$sheight,
        names=c("Padres", "Hijos"))
```



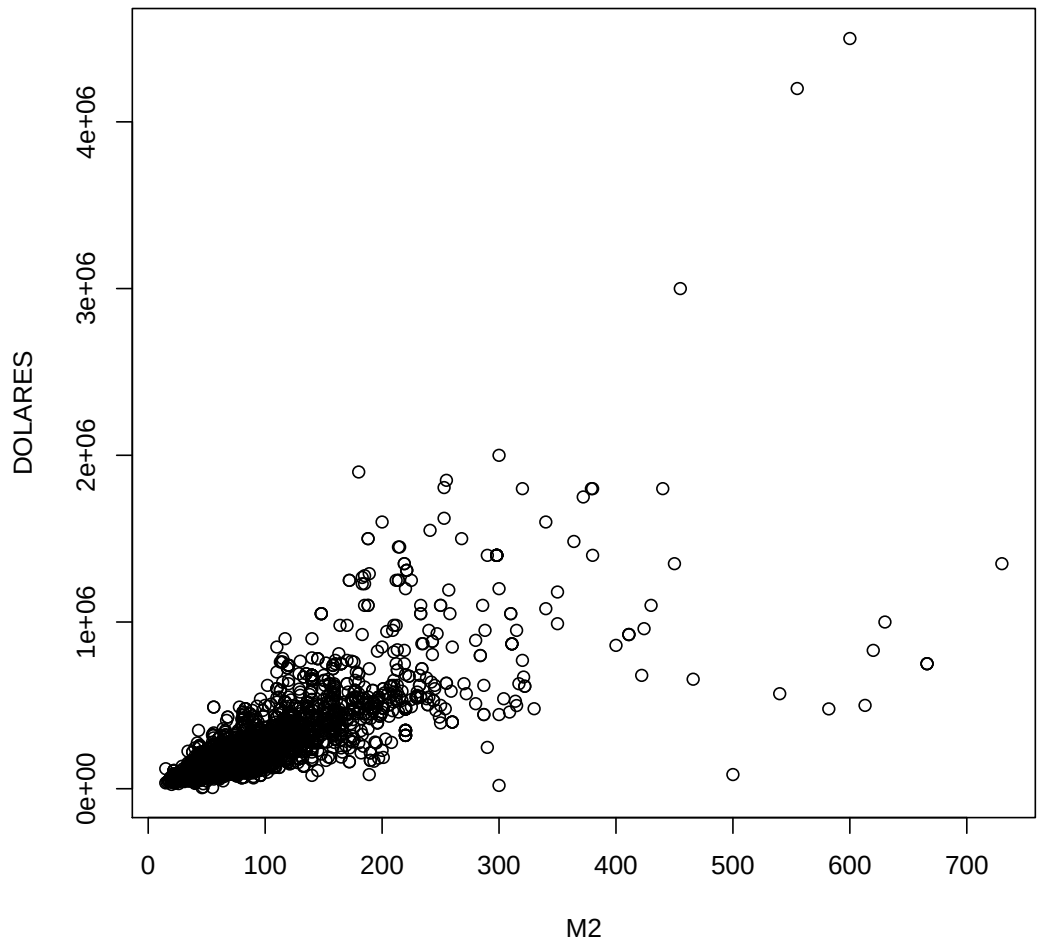
Tercera parte
Datos bivariados.

Datos bivariados: relación entre dos variables numéricas

- ¿Hay relación entre las variables?
- ¿Cuán fuerte es la relación entre las variables?
- ¿Es la relación entre las variables lineal?

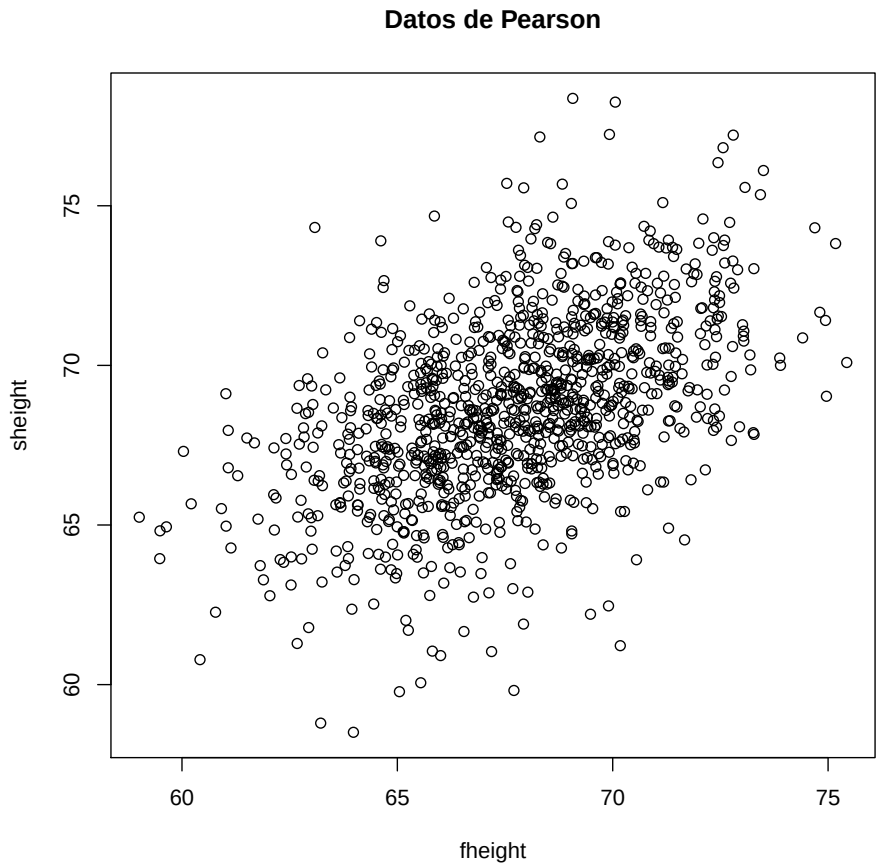
Datos de departamentos

```
departamentos<-read.csv("departamentos.txt",header = TRUE)
with(data=departamentos, plot(M2,DOLARES))
```



Datos de Pearson: estatura de padres e hijos

```
library(UsingR)
with(data=father.son, plot(fheight, sheight)
```



Estimación de la correlción

- **Recuerdo:** La correlación (poblacional) de un vector aleatorio (X, Y):

$$\rho(X,Y) = \frac{E((X - E(X)(Y - E(Y)))}{\sqrt{V(X)V(Y)}}$$

- La correlación muestral:

$$\hat{\rho}(X,Y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})/n}{\sqrt{\sum(x_i - \overline{x})^2/n \sum(y_i - \overline{y})^2/n}}$$

$$\hat{\rho}(X,Y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

Estimación de la correlación en R

```
> cor(departamentos$M2,departamentos$DOLARES)
[1] 0.8003153

> cor(father.son$fheight,father.son$sheight)
[1] 0.5013383
```

Ejercicio Verificar que da lo que indica la fórmula de la diapositiva anterior.

Predicción

Vimos que el mejor predictor lineal de Y basado en X según el criterio del ECM es:

$$\hat{Y} = \mu_Y - \frac{\sigma_Y}{\sigma_X}\rho_{XY}\mu_X + \frac{\sigma_Y}{\sigma_X}\rho_{XY}X$$

Ahora podemos estimar μ_X,μ_Y, σ_X, σ_Y y ρ_{XY}.

Predicción

Vimos que el mejor predictor lineal de Y basado en X según el criterio del ECM es:

$$\hat{Y} = \underbrace{\mu_Y - \frac{\sigma_Y}{\sigma_X}\rho_{XY}\mu_X}_{\alpha} + \underbrace{\frac{\sigma_Y}{\sigma_X}\rho_{XY}}_{\beta} X$$

$$\hat{\beta} = \sqrt{\frac{\sum(y_i - \overline{y})^2}{\sum(x_i - \overline{x})^2}} \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

$$= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

La recta de regresión

- La recta

$$y = \alpha + \beta x,$$

  con α y β definidos como en la diapositiva anterior, se llama recta de regresión
- **Prop:** $(\alpha, \beta) = \mathit{argmin}_{a,b}\, E\left((Y - a - bX)^2\right)$

La recta de regresión estimada

- La recta

$$y = \hat{\alpha} + \hat{\beta}x,$$

  se llama recta de regresión estimada

- **Prop:**

$$(\hat{\alpha}, \hat{\beta}) = \underset{a,b}{\textit{argmin}} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

  Dem: no la haremos en este curso.

- $(\hat{\alpha}, \hat{\beta})$ se llama estimación por mínimos cuadrados de (α, β).

Estimación por mínimos cuadrados en R

```
> ajuste1 <- lm(DOLARES~M2, data=departamentos)
> ajuste1

Call:
lm(formula = DOLARES ~ M2, data = departamentos)

Coefficients:
(Intercept)           M2
-24860          2973
```

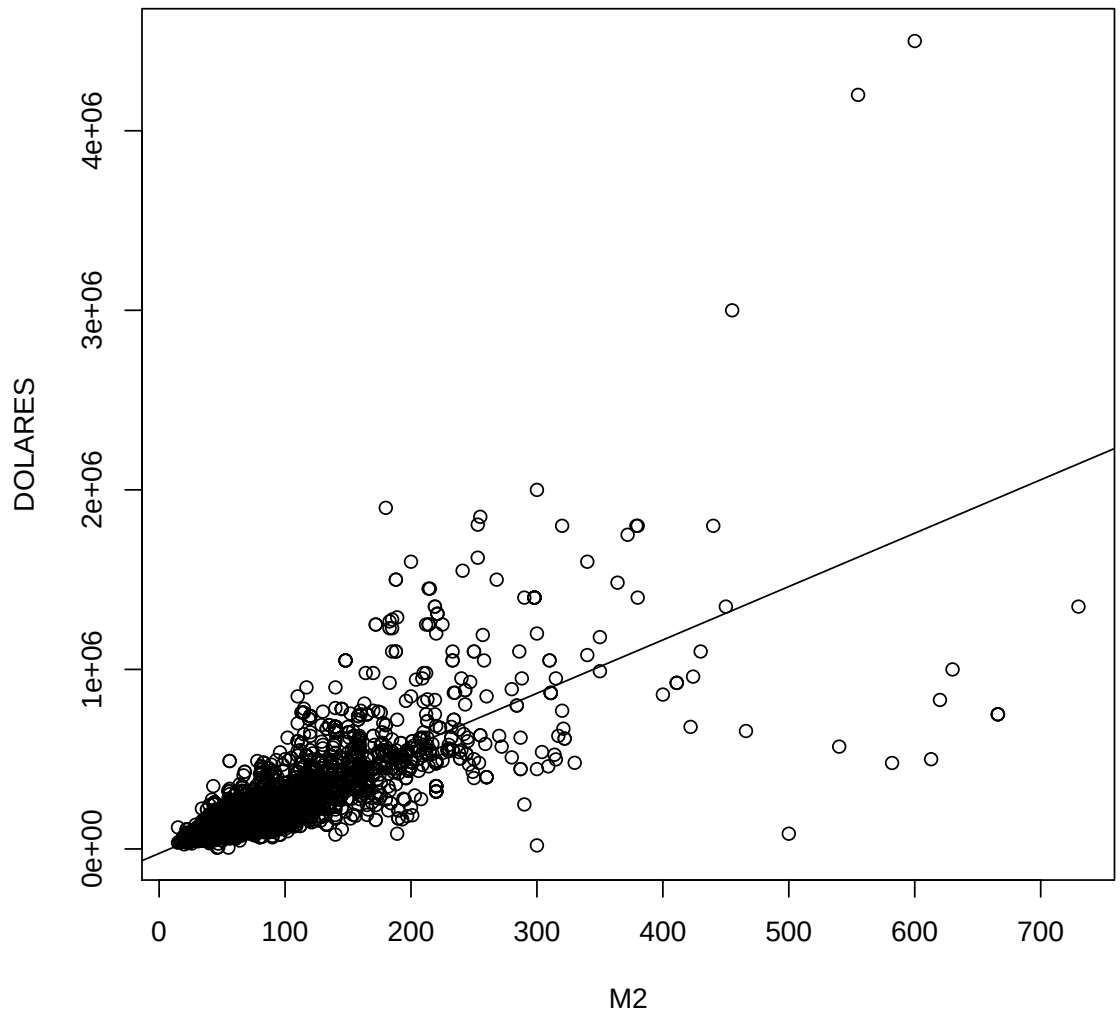
La recta de regresión estimada de DOLARES vs M2 basada es

$$y = -24860 + 2973x$$

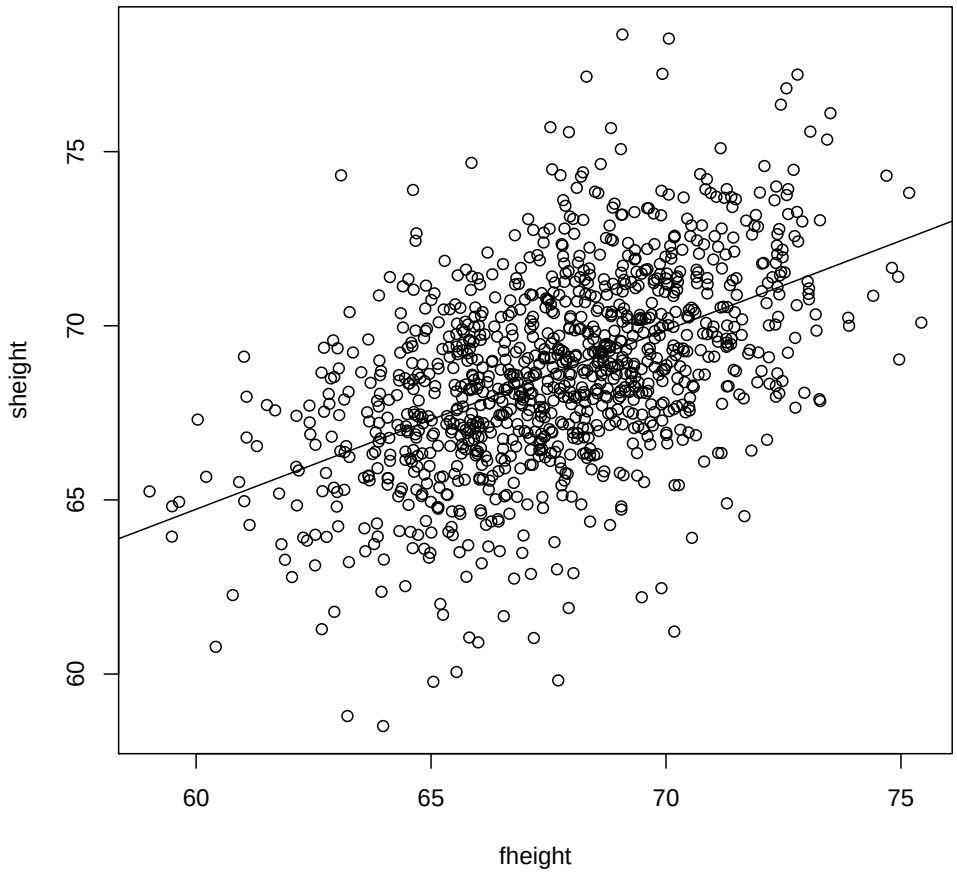
Predicción del precio de un departamento de $120M2$:

$$\hat{y}_{120} = -24860 + 2973 * 120 = 331900$$

```
with(data=departamentos, plot(M2,DOLARES))
abline(ajuste1$coefficients)
```



```
with(data=father.son, plot(fheight,sheight))
ajuste2 <- lm(sheight~fheight, data=father.son)
abline(ajuste2$coefficients)
```



Para explorar: el método L1

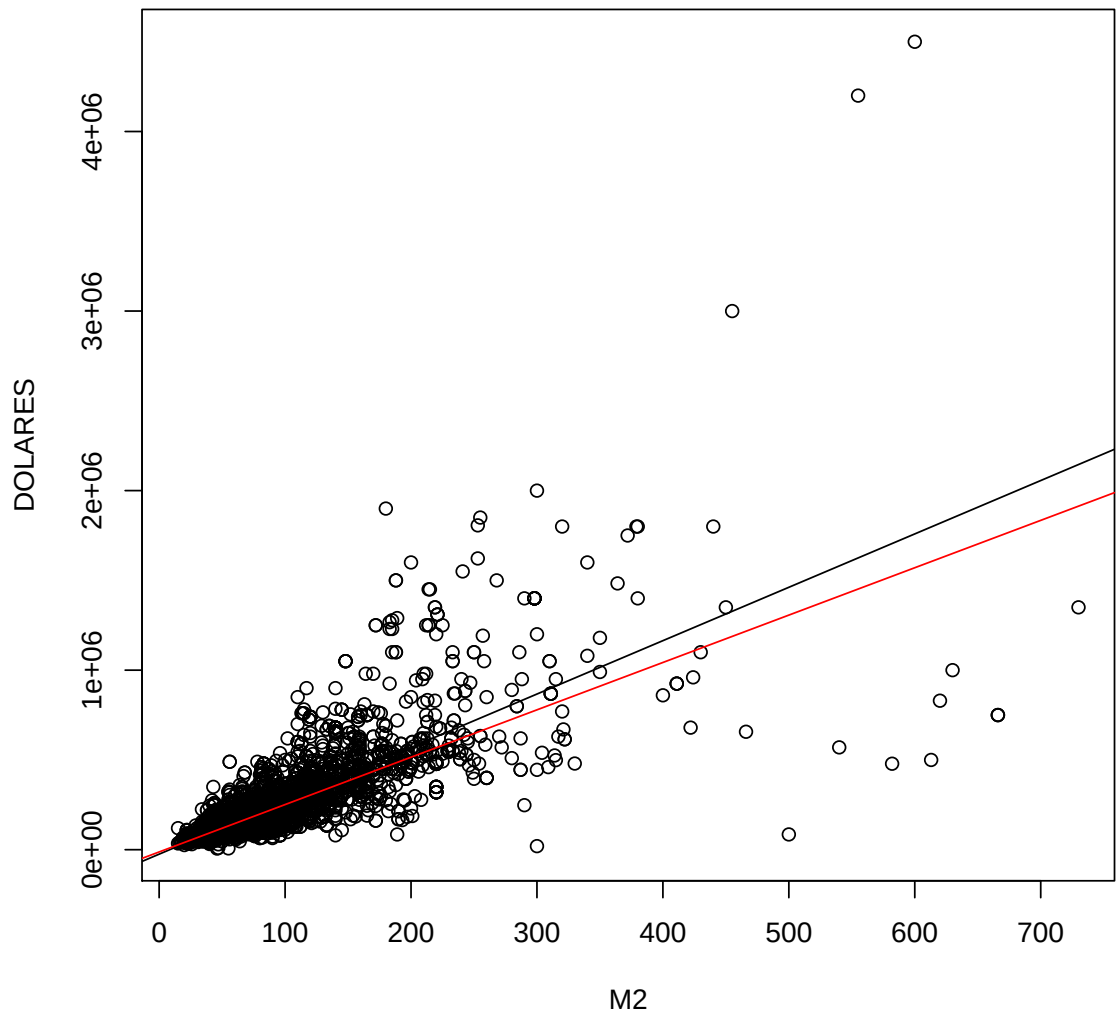
Con este método se minimiza el EAM

$$(\hat{\alpha}, \hat{\beta}) = \underset{a,b}{\textit{argmin}} \sum_{i=1}^{n} (|y_i - a - bx_i|)$$

$(\hat{\alpha}, \hat{\beta})$ no se pueden calcular en forma explícita.
Pero sí se pueden calcular por método numéricos:

```
> ajuste3 <- l1fit(departamentos$M2,departamentos$DOLARES)
> ajuste3$coefficients
Intercept           X
-12542.435   2638.224
```

```
with(data=departamentos, plot(M2,DOLARES))
abline(ajuste1$coefficients)
abline(ajuste3$coefficients)
```



```
with(data=father.son, plot(fheight,sheight))
ajuste2 <- lm(sheight~fheight, data=father.son)
ajuste4 <- l1fit(father.son$fheight,father.son$sheight)
abline(ajuste2$coefficients)
abline(ajuste4$coefficients, col=2)
```