Sesión 3: Estadística descriptiva multivariable

EUSEBIO ANGULO SÁNCHEZ-HERRERA

LABORATORIO DE ESTADÍSTICA

ESCUELA SUPERIOR DE INFORMÁTICA CIUDAD REAL

Variables estadísticas Multivariables

- Normalmente varias características es igual a varias variables estadísticas (S.V.) del mismo individuo / experimento.
- Bivariante o de dos variables:
  - Cualitativa– Cualitativa. (Comunidad – Universidad, Grupo - Corrector)
  - Cuantitativa– Cualitativa(Nota – Universidad, Número de socios – Equipo de Fútbol)
  - Cuantitativa– Cuantitativa (Peso – Altura, Goles a favor – Goles en contra)

Variables estadísticas Multivariables

- **Ejemplo**. Se recogen las opiniones de varios clientes antes y después de probar un coche.

Clientes	1	2	3	4	5	6	7	8	9	10
1ªOpinión	b	b	g	b	b	g	b	g	g	b
2ªOpinión	g	b	g	g	b	g	b	b	g	g

b=bad; g=good

```
>primer_opinion=scan(,"") #caracteres
1: b  b  g  ...
>segun_opinion=scan()      #números (pero variable categórica)
1: 1  0  1  ...
```

Variables estadísticas Multivariables

>summary(primer_opinion)

>summary(segun_opinion)

```
Length     Class      Mode
    10 character character
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0     0.0     1.0     0.6     1.0     1.0
```

>table(primer_opinion, segun_opinion)

```
                segun_opinion
primer_opinion 0 1
             b 3 3
             g 1 3
```

> table(primer_opinion, segun_opinion)/length(segun_opinion)

Variables estadísticas Multivariables

>tab.p<-table(primer_opinion,segun_opinion)/
length(segun_opinion)*100

```
                segun_opinion
primer_opinion  0  1
             b 30 30
             g 10 30
```

>margin.table(tab.p,1)

#marginal por filas

>margin.table(tab.p,2)

#marginal por columnas

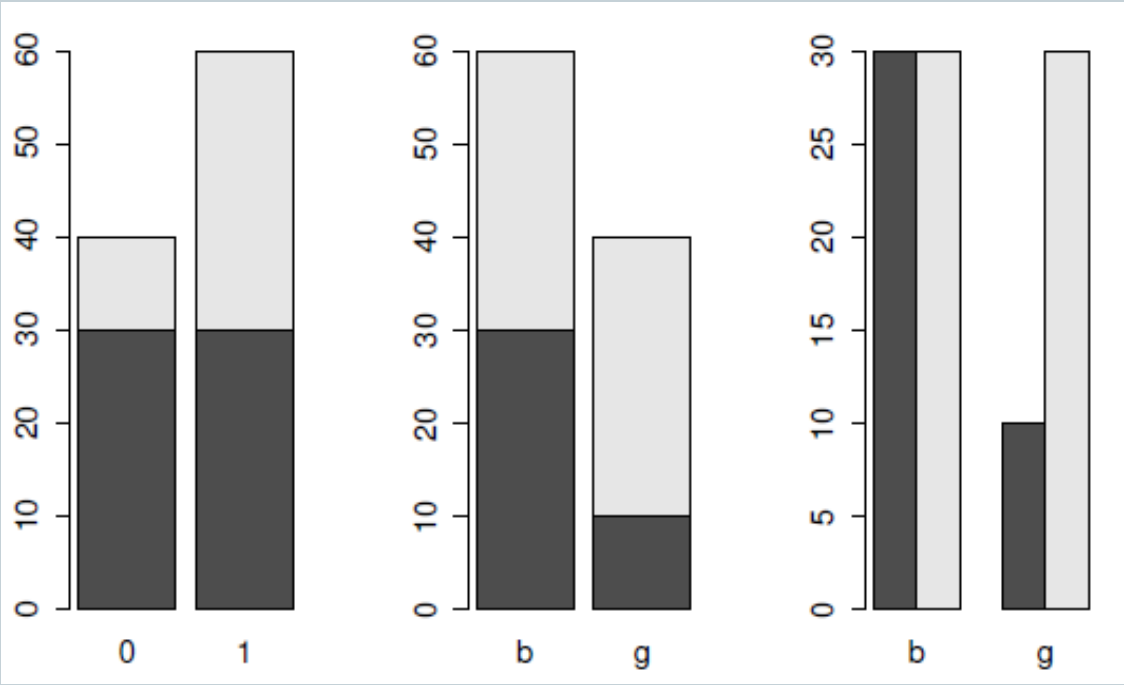
>addmargins(tab.p)

#añade sumatorios

```
> margin.table(tab.p,1)
primer_opinion
 b  g
60 40
> #marginal por filas
> margin.table(tab.p,2)
segun_opinion
 0  1
40 60
> #marginal por columnas
> addmargins(tab.p)
                segun_opinion
primer_opinion   0   1 Sum
             b  30  30  60
             g  10  30  40
             Sum 40  60 100
> #añade sumatorios
```

Gráficos: Variables estadísticas Multivariables


```
>barplot(tab.p)        # bar = columnas
>barplot(t(tab.p))     # t= traspuesta
>barplot(t(tab.p),beside=TRUE)    # bar = columnas
```



Variables estadísticas Multivariables

- Multivariables y cuantitativas:
  - Estudio descriptivo de las relaciones entre dos variables.
  - Análisis estadístico entre dos variables: regresión lineal.
  - Regresión multivariable

- Ejemplo de mediciones de cuatro importantes cualidades de un producto.

```
>load("Variables.Rdata")
#Variables !!
>ls()
[1] ... "Var1" "Var2" "Var3" "Var4"
>summary(Var1);summary(Var2)
```

Data Frame

- Data Frame es un objeto similar a una matriz, donde las columnas son las variables (pueden ser de diferentes tipos) y las filas el número de casos, muestra, experimentos…

```
>Data=data.frame(primer_opinion,Var4,Var1,segun_opinion)
>Data$Var4          #Acceso a variable 4
>Data[,2]           #Diferente forma de acceder a variable 4
>Data[23,4]         #… Fila 23, Columna 4
>Data[27,]
```

	primer_opinion	Var4	Var1	segun_opinion
27	b	8.011113	1.053071	0

```
> Data$Var4 == Data[,4] #label ≠ posición
```

Data Frame

- Observaciones importantes:

  1) Son diferentes:

  > length(Data);dim(Data) #nº variables y nºfilas y columnas

  [1] 4

  [1] 100 4

  2) Se pueden guardar como Rdata y luego cargar.

  >save(Data,file="DataV.Rdata")

  >load("DataV.Rdata")

  3) >attach(Data) #Genera una copia temporalmente permite
  # usar variables por el nombre sin el
  >detach(Data) # símbolo $

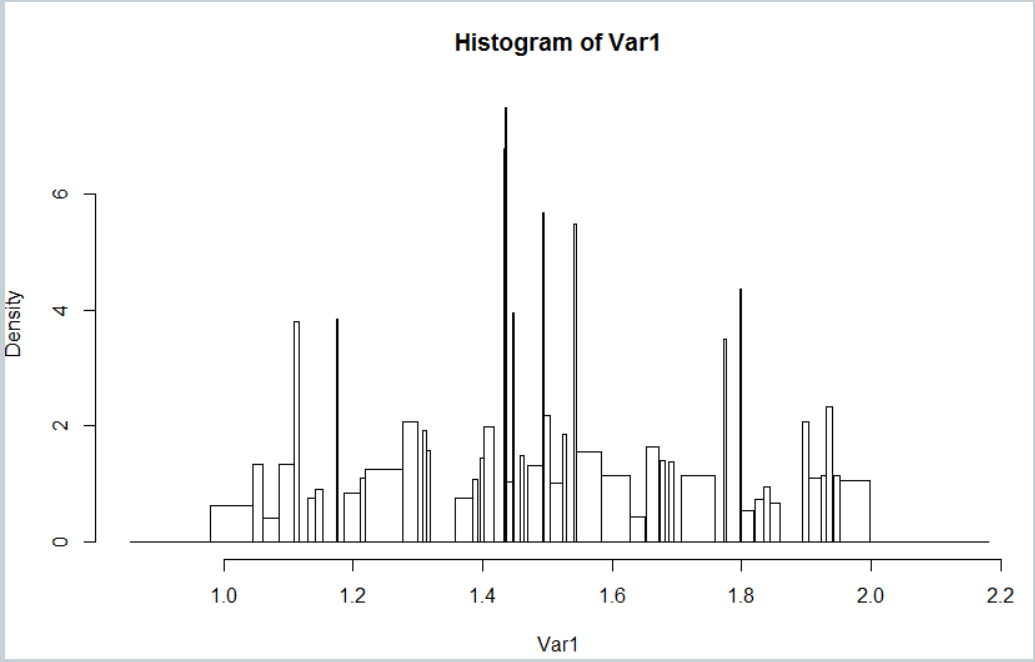
Histogramas

>hist(Var1) # Con una variable no se pueden comparar

>hist(Var1)

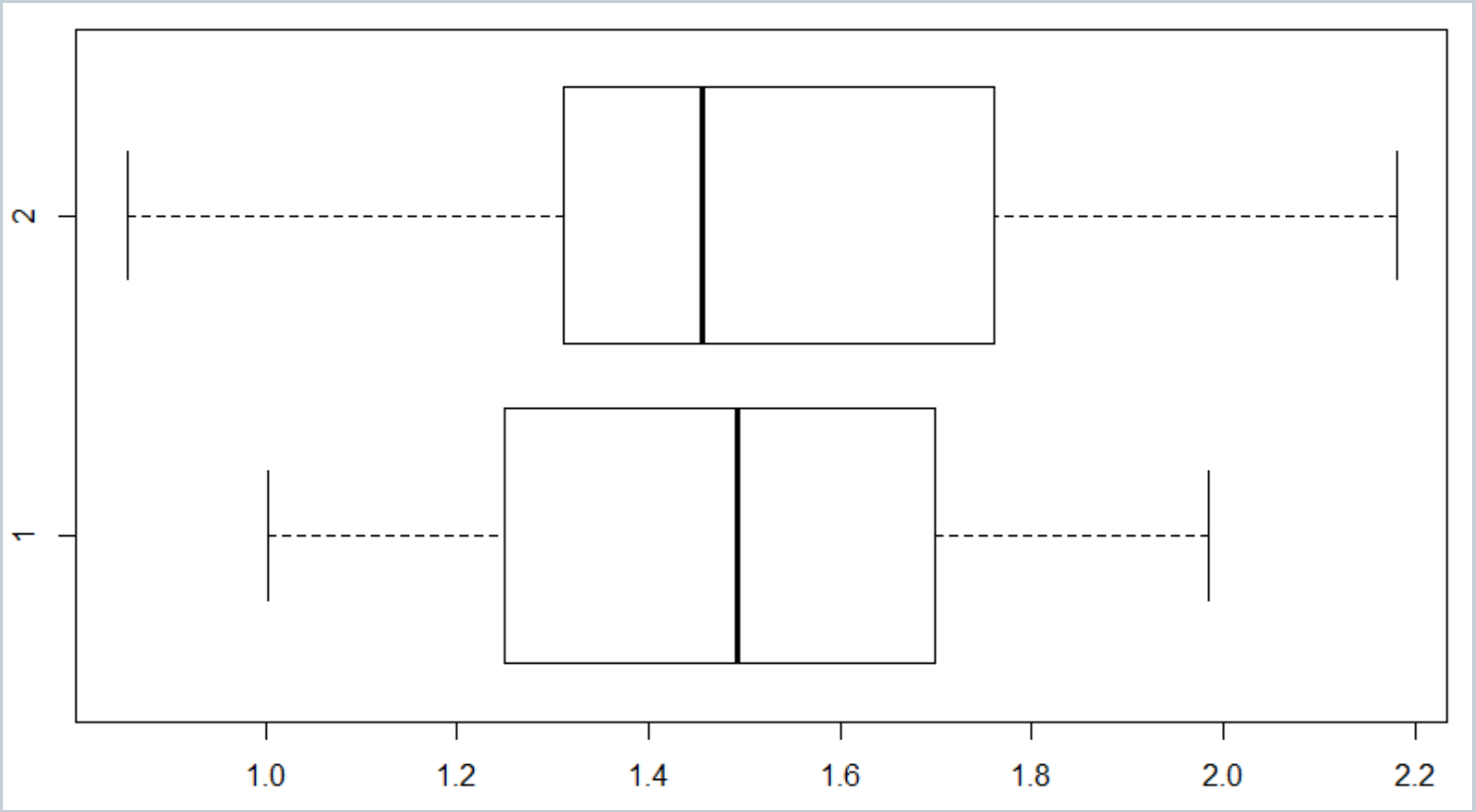
>hist(Var1,Var2) # Error

>hist(Var1,Var3) #



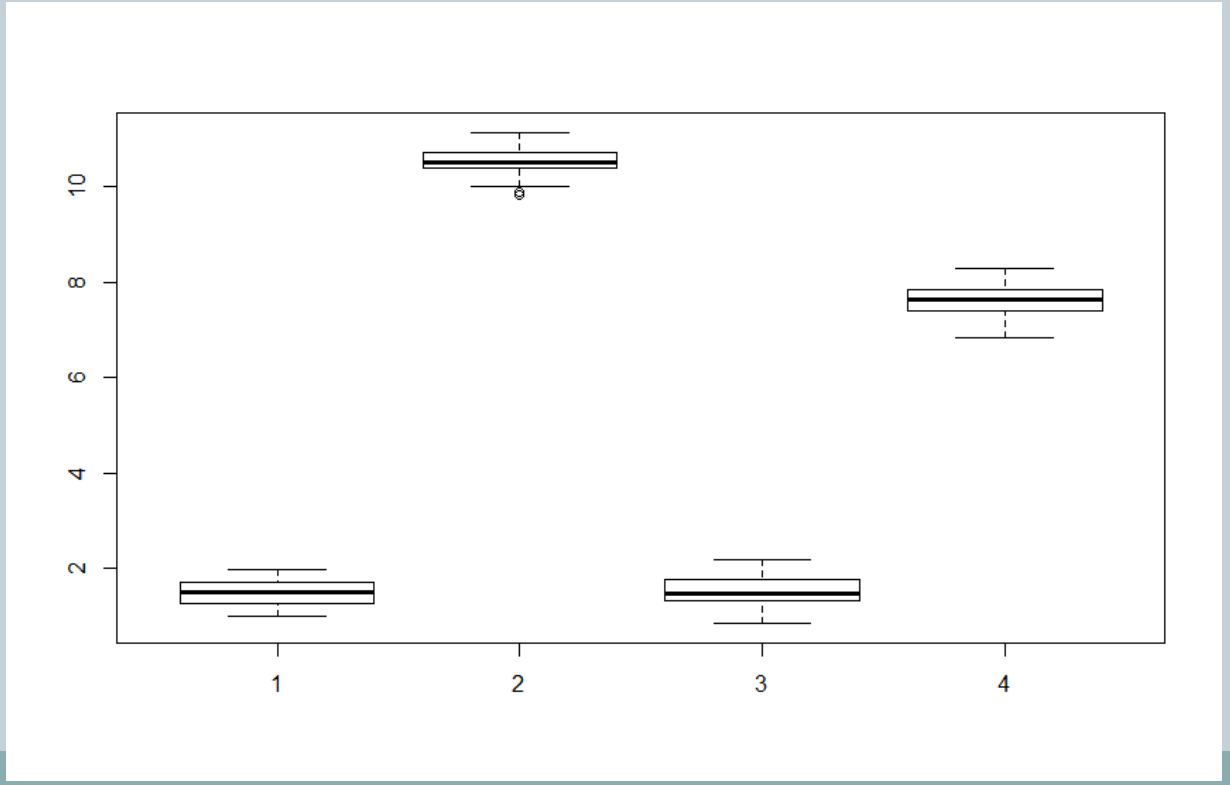
Boxplots

>boxplot(Var1,Var3,horizontal=TRUE)



Boxplots

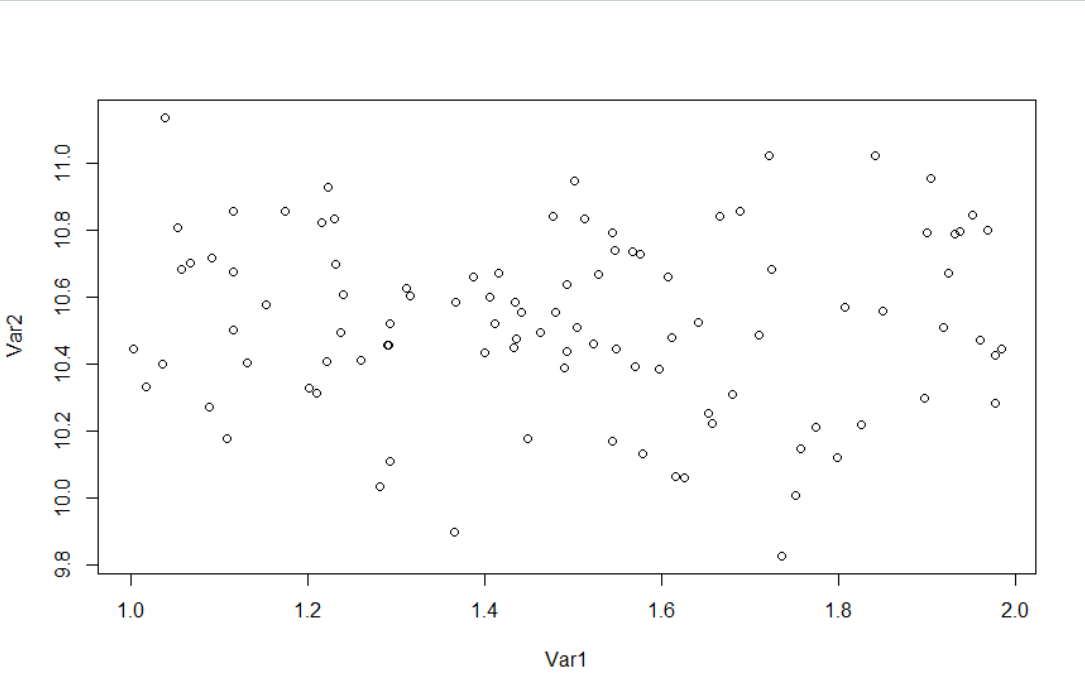

```
>boxplot(Var1,Var2,Var3,Var4) # Mejor con objetos
>Vars<-data.frame(Var1,Var2,Var3,Var4)
>boxplot(Vars)
```



Correlación entre 2 variables numéricas

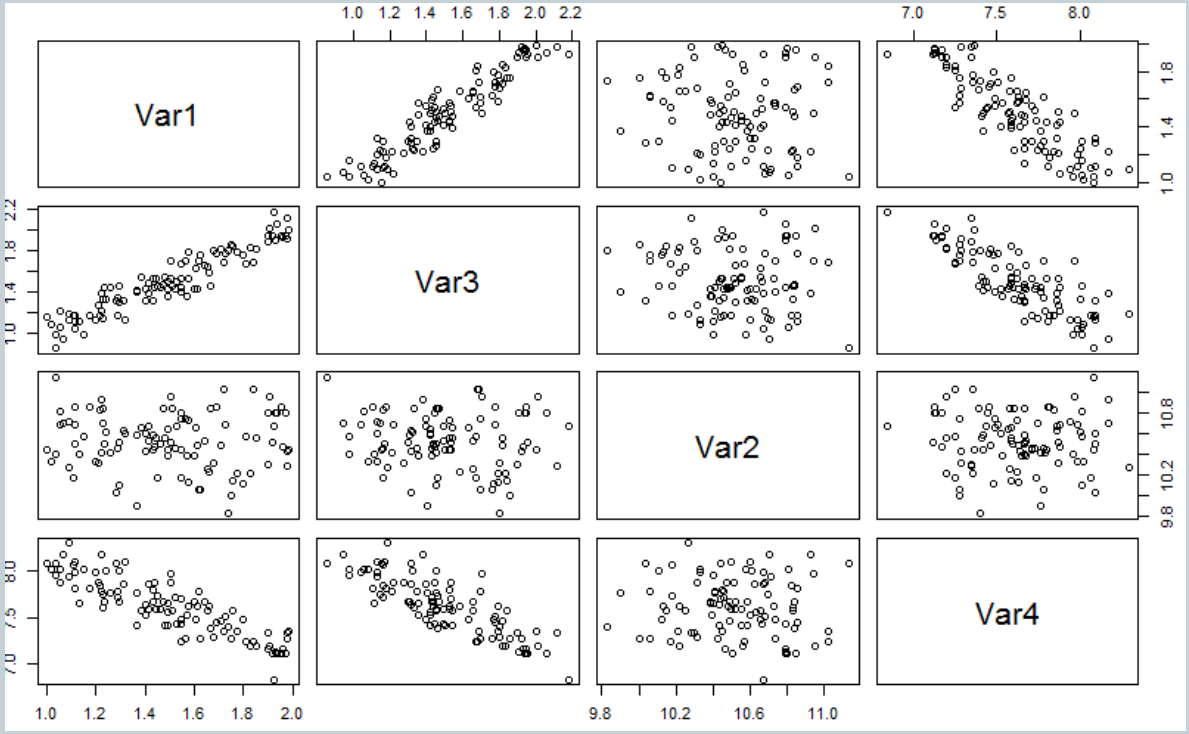
- Las variables deben ser numéricas
- **Scatterplots:**

>plot(Var1,Var2) #plot(x,y)≠ plot(y,x)



Correlación entre 2 variables numéricas

- Vars<-data.frame(Var1,Var3,Var2,Var4) cargar fichero con scan
- > pairs(Vars) # Matriz que compara la relación de todas las # variables



Correlación entre 2 variables numéricas

- **Coeficiente de correlación de Pearson**

r	≈1	≈0	≈-1
Relación Lineal	Directa	Ausencia	Inversa

A	>cor(Var1,Var3)	B	-0.0454823
B	>cor(Var1,Var2)	C	-0.863472
C	>cor(Var1,Var4)	A	0.9386554

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Observación 1: **cor(x,y)=cor(y,x)**
- Observación 2: Se recomienda usar

```
use="pairwise.complete.obs" # necesario cuando hay NA, ya
# que calcula la correlación aun habiendo valores vacios o
# perdidos
```

Modelo de regresión lineal

$$y_i = a + b \cdot x_i + \epsilon_i$$

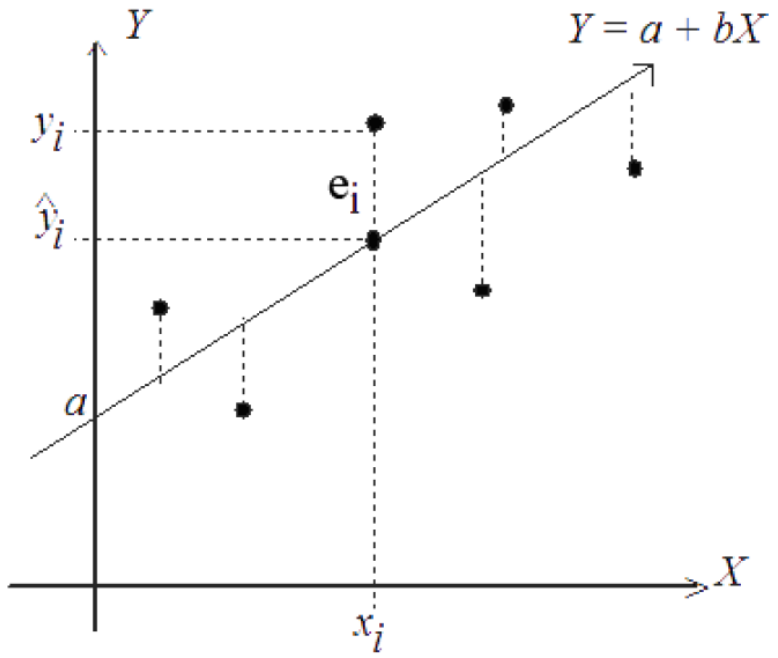
- y_i: variable dependiente o respuesta
- x_i: variable independiente
- Coeficientes de regresión:
  - a: ordenada en el origen constante
  - b: pendiente de la recta
- €i: error

- Mínimos cuadrados: $\min \sum e_i^2 = \min \sum (y_i - (a + b \cdot x_i))^2$

$$e_i = y_i - \hat{y}_i$$

Modelo de regresión lineal

- Estimación de coeficientes

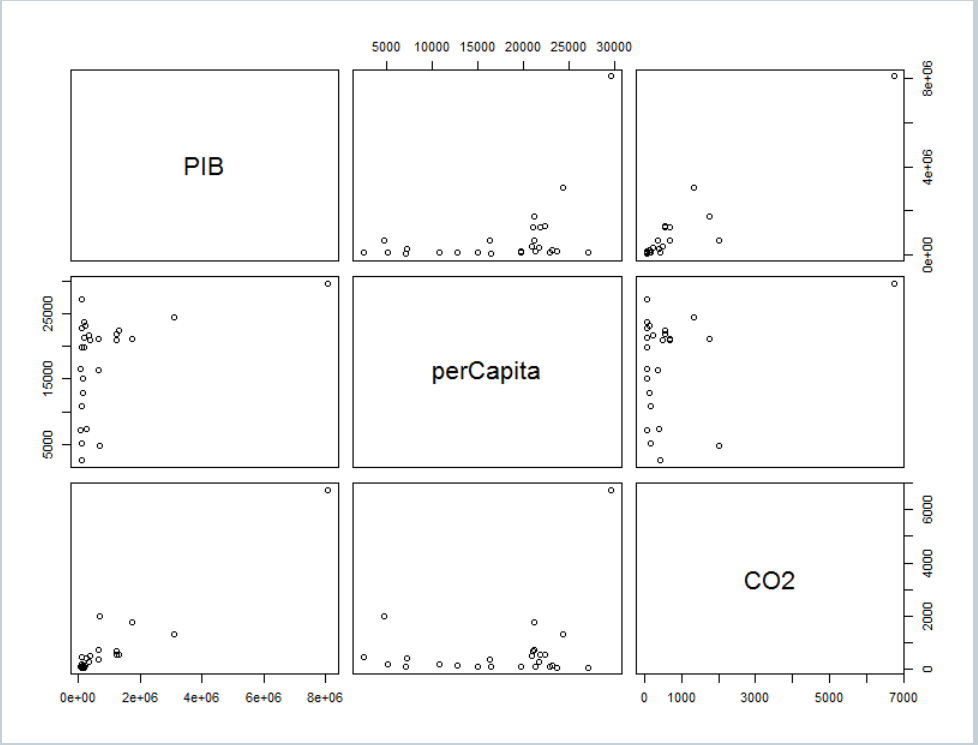


$$\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2};$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

Modelo de regresión lineal

- Datos de emisiones de CO_2, PIB(GDP) y renta per cápita de varios países
  - >load("emisiones.Rdata")
  - head(emisiones)
  - >summary(emisiones)
  - >pairs(emisiones)
  - >cor(emisiones)



Modelo de regresión lineal

- Modelo Lineal a estudiar: $CO2 = a + b \cdot PIB + €$
  - >CO2<-emisiones$CO2;PIB<-emisiones$PIB
  - >model1<-CO2~PIB #defición del modelo co2 variable y pib variable x
  - >reg1<-lm(model1) #ajustar el modelo
  - >summary(reg1)

estimacion A

Estimacion B

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.043e+01  9.441e+01   0.216     0.83
PIB         7.815e-04  5.233e-05  14.933  1.2e-13 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 427.4 on 24 degrees of freedom
Multiple R-squared:  0.9028,    Adjusted R-squared:  0.8988
F-statistic:   223 on 1 and 24 DF,  p-value: 1.197e-13
```

Bondad de ajuste

20,43

Modelo de regresión lineal

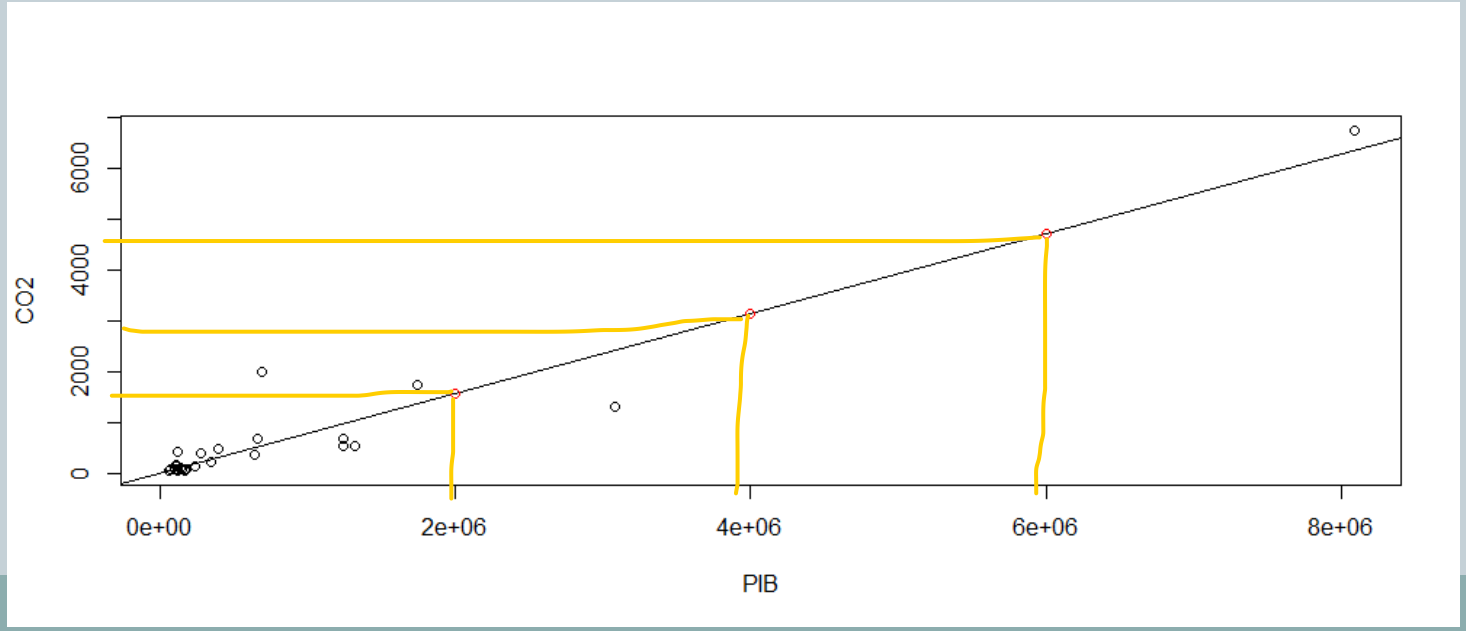
- Los resultados devuelven los valores a y b, con la significación que se han estimado.
- La **bondad de ajuste** o coeficiente de determinación es el "Multiple R-squared" (R^2).
- El **valor p-value**, es para indiciar si el coeficiente del modelo es significativo o no (esto se verá en inferencia).

Modelo de regresión lineal

- Modelo ajustado: $CO2 = 20.43 + 7.815 \cdot 10^{-4} \cdot PIB + €$
  - Predicción (PIB=2e6, 4e6 and 6e6)

2 millones

```
>plot(model1);abline(reg1)
>pred<-predict(reg1,data.frame(PIB<-c(2e6,4e6,6e6)))
>points(cbind(PIB,pred),col="red")
```



Sumario

- Repasar conceptos vistos en Tema 2 de Estadística Descriptiva. Se utilizan diferentes funciones con Variables Cualitativas y Cuantitativas
- Bivariante (2 categorías) y categóricas (más de 2 categorías): Diagramas de barras y de frecuencias
- Multivariante y cuantitativas: Con representaciones gráficas *scatterplot* (diagramas de dispersión) y medidas utilizadas coeficientes correlación (Pearson).
- Los *dataframe* muy utilizados en R.
- Modelos de Regresión Lineal