

# MR - Trabajo

Alicia Losada | alicia.losada.sanchez@udc.es      María Cardoso | m.cardoso@udc.es  
Nicolás Muñiz | nicolas.muniz@udc.es

11/12/2024

## Regresión Lineal Múltiple

- Antes de empezar, cargamos los datos *OzonoLA.rda*

```
load("Datos/OzonoLA.rda")  
attach(OzonoLA)
```

### 1. Análisis descriptivo

Para el análisis descriptivo de las variables podemos comenzar con una visión general de las variables mediante las funciones `str()` y `summary()`.

```
str(OzonoLA)
```

```
## 'data.frame':   203 obs. of  13 variables:  
## $ Mes          : int  1 1 1 1 1 1 1 1 1 1 ...  
## $ DiaMes       : int  5 6 7 8 9 12 13 14 15 16 ...  
## $ DiaSemana    : int  1 2 3 4 5 1 2 3 4 5 ...  
## $ Ozono        : num  5.34 5.77 3.69 3.89 5.76 6.39 4.73 4.35 3.94 7 ...  
## $ Pres_Alt     : int  5760 5720 5790 5790 5700 5720 5760 5780 5830 5870 ...  
## $ Vel_Viento   : int  3 4 6 3 3 3 6 6 3 2 ...  
## $ Humedad      : int  51 69 19 25 73 44 33 19 19 19 ...  
## $ T_Sandburg   : int  54 35 45 55 41 51 51 54 58 61 ...  
## $ T_ElMonte    : num  45.3 49.6 46.4 52.7 48 ...  
## $ Inv_Alt_b    : int  1450 1568 2631 554 2083 111 492 5000 1249 5000 ...  
## $ Grad_Pres    : int  25 15 -33 -28 23 9 -44 -44 -53 -67 ...  
## $ Inv_T_b      : num  57 53.8 54.1 64.8 52.5 ...  
## $ Visibilidad  : int  60 60 100 250 120 150 40 200 250 200 ...
```

La salida de `str()` nos dice que los datos constan de 203 observaciones de 13 variables:

- **Mes**: Número del mes en el que se hicieron las observaciones (Entero)
- **DiaMes**: Número del día del mes en el que se hicieron las observaciones (Entero)
- **DíaSemana**: Número del día de la semana en el que se hicieron las observaciones (Entero)
- **Ozono**: Nivel de Ozono medido (Numérica)
- **Pres\_Alt**: Altura en metros a la que se alcanza una presión de 500 milibares (Entero)
- **Vel\_Viento**: Velocidad del viento en millas por hora en el Aeropuerto Internacional de Los Angeles (Entero)
- **Humedad**: Humedad en porcentaje en LAX (Entero)
- **T\_Sandburg**: Temperatura (F) en Sandburg, CA (Entero)

- T\_ElMonte: Temperatura (F) en El Monte, CA (Numérica)
- Inv\_Alt\_b: Inversion de la altura base (en pies) en LAX (Entero)
- Grand\_Pres: Gradiente de presion de LAX a Daggett, CA (Entero)
- Inv\_T\_b: Inversion de la temperatura base (F) en LAX (Numérica)
- Visibilidad: Visibilidad (millas) evaluada en LAX (Entero)

```
summary(OzonoLA)
```

```
##      Mes      DiaMes      DiaSemana      Ozono      Pres_Alt
## Min.   : 1.000   Min.    : 1.0   Min.   :1.000   Min.    : 0.72   Min.    :5320
## 1st Qu.: 3.000   1st Qu.: 9.0   1st Qu.:2.000   1st Qu.: 4.77   1st Qu.:5690
## Median : 6.000   Median :15.0   Median :3.000   Median : 8.90   Median :5760
## Mean   : 6.522   Mean    :15.7   Mean    :3.005   Mean    :11.37   Mean    :5746
## 3rd Qu.:10.000   3rd Qu.:23.0   3rd Qu.:4.000   3rd Qu.:16.07   3rd Qu.:5830
## Max.    :12.000   Max.     :31.0   Max.     :5.000   Max.     :37.98   Max.     :5950
## Vel_Viento      Humedad      T_Sandburg      T_ElMonte
## Min.   : 0.000   Min.    :19.00   Min.     :25.00   Min.     :27.68
## 1st Qu.: 3.000   1st Qu.:46.00   1st Qu.:51.50   1st Qu.:49.64
## Median : 5.000   Median :64.00   Median :61.00   Median :56.48
## Mean   : 4.867   Mean    :57.61   Mean     :61.11   Mean     :56.54
## 3rd Qu.: 6.000   3rd Qu.:73.00   3rd Qu.:71.00   3rd Qu.:66.20
## Max.    :11.000   Max.     :93.00   Max.     :93.00   Max.     :82.58
## Inv_Alt_b      Grad_Pres      Inv_T_b      Visibilidad
## Min.   : 111   Min.    : -69.00   Min.     :27.50   Min.     : 0.0
## 1st Qu.: 869   1st Qu.: -14.00   1st Qu.:51.26   1st Qu.: 60.0
## Median :2083   Median : 18.00   Median :60.98   Median :100.0
## Mean   :2602   Mean     :14.43   Mean     :60.69   Mean     :122.2
## 3rd Qu.:5000   3rd Qu.: 43.00   3rd Qu.:70.88   3rd Qu.:150.0
## Max.    :5000   Max.     :107.00   Max.     :90.68   Max.     :350.0
```

Ahora realizaremos un análisis descriptivo de cada variable:

### Análisis descriptivo de la variable Mes :

```
summary(Mes)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000  3.000   6.000   6.522 10.000   12.000
```

Desviación típica y rango intercuartílico:

```
sd(Mes)
```

```
## [1] 3.594998
```

```
IQR(Mes)
```

```
## [1] 7
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(Mes, na.rm = FALSE)
```

```
## [1] 0.03220505
```

```
kurtosis(Mes, na.rm = FALSE)
```

```
## [1] 1.671129
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

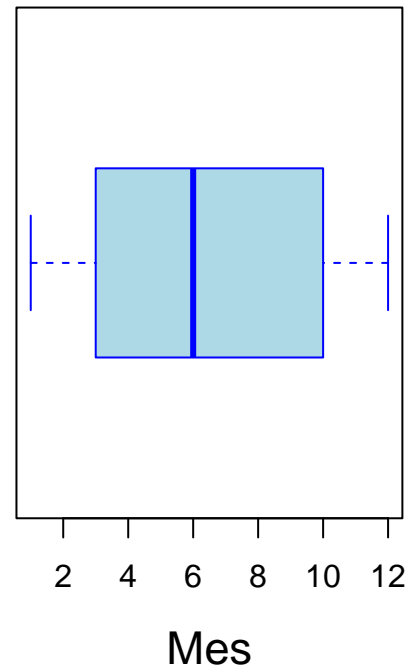
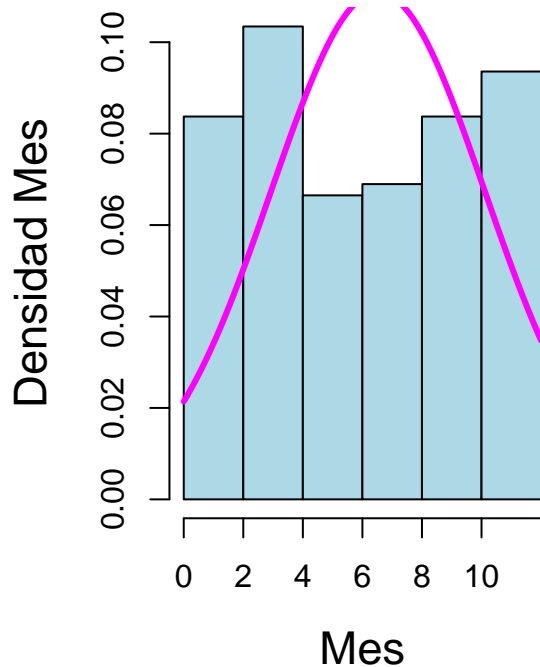
Vemos si hay registros atípicos

```
boxplot.stats(Mes)$out
```

```
## integer(0)
```

Como podemos ver no existe ningún registro atípico

```
par(mfrow=c(1,2))
hist(Mes, breaks=5, freq=FALSE, main = "", xlab="Mes",
     cex.lab=1.4, ylab = "Densidad Mes", col = "lightblue")
curve( dnorm(x, mean=mean(Mes), sd=sd(Mes)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Mes, main = "", xlab="Mes",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable DiaMes :

```
summary(Mes)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   3.000   6.000   6.522  10.000  12.000
```

Desviación típica y rango intercuartílico:

```
sd(DiaMes)
```

```
## [1] 8.569537
```

```
IQR(DiaMes)
```

```
## [1] 14
```

Evaluamos la asimetría y kurtosis

```
library(moments)
skewness(DiaMes, na.rm = FALSE)
```

```
## [1] 0.0395616
```

```
kurtosis(DiaMes, na.rm = FALSE)
```

```
## [1] 1.868548
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

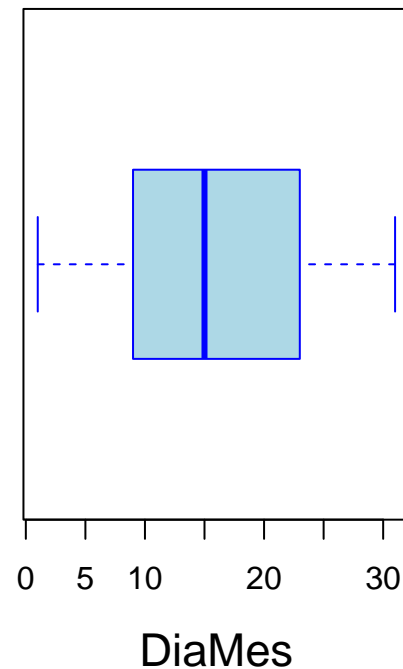
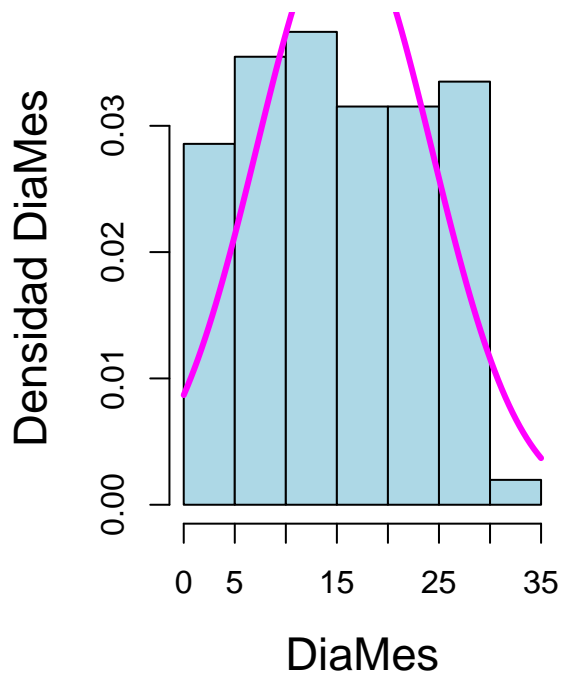
Vemos si hay registros atípicos

```
boxplot.stats(DiaMes)$out
```

```
## integer(0)
```

Como podemos ver no existe ningún registro atípico

```
par(mfrow=c(1,2))
hist(DiaMes, breaks=5, freq=FALSE, main = "", xlab="DiaMes",
     cex.lab=1.4, ylab = "Densidad DiaMes", col = "lightblue")
curve( dnorm(x, mean=mean(DiaMes), sd=sd(DiaMes)),
      col="magenta", lwd=3, add=TRUE)
boxplot(DiaMes, main = "", xlab="DiaMes",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable DiaSemana :

```
summary(DiaSemana)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000  2.000   3.000   3.005  4.000   5.000
```

Desviación típica y rango intercuartílico:

```
sd(DiaSemana)
```

```
## [1] 1.401899
```

```
IQR(DiaSemana)
```

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

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(DiaSemana, na.rm = FALSE)
```

```
## [1] 0.04527053
```

```
kurtosis(DiaSemana, na.rm = FALSE)
```

```
## [1] 1.731687
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

Vemos si hay registros atípicos

```
boxplot.stats(DiaSemana)$out
```

```
## integer(0)
```

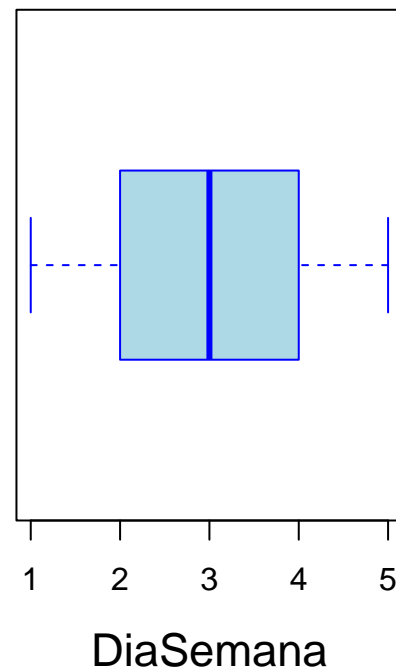
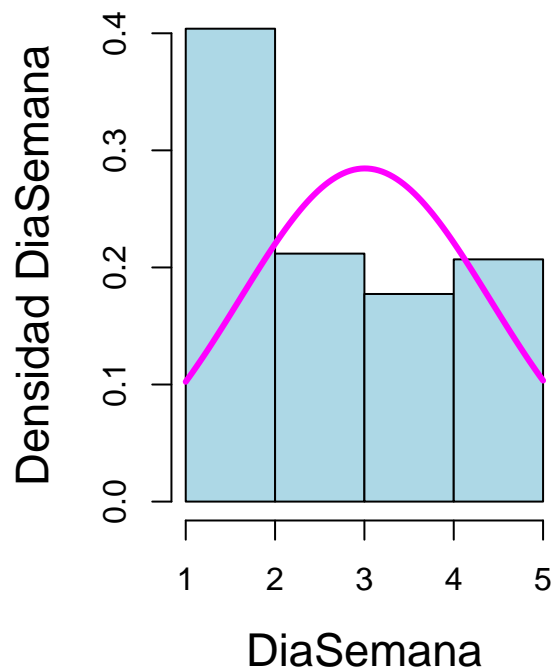
Como podemos ver no existe ningún registro atípico

```
par(mfrow=c(1,2))
```

```
hist(DiaSemana, breaks=5, freq=FALSE, main = "", xlab="DiaSemana",  
      cex.lab=1.4, ylab = "Densidad DiaSemana", col = "lightblue")
```

```
curve( dnorm(x, mean=mean(DiaSemana), sd=sd(DiaSemana)),  
       col="magenta", lwd=3, add=TRUE)
```

```
boxplot(DiaSemana, main = "", xlab="DiaSemana",  
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",  
        horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Ozono :

```
summary(Ozono)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.72   4.77   8.90  11.37  16.07  37.98
```

Desviación típica y rango intercuartílico:

```
sd(Ozono)
```

```
## [1] 8.192652
```

```
IQR(Ozono)
```

```
## [1] 11.305
```

Evaluamos la asimetría y kurtosis

```
library(moments)
```

```
skewness(Ozono, na.rm = FALSE)
```

```
## [1] 0.9652702
```

```
kurtosis(Ozono, na.rm = FALSE)
```

```
## [1] 3.089498
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal

Vemos si hay registros atípicos

```
boxplot.stats(Ozono)$out
```

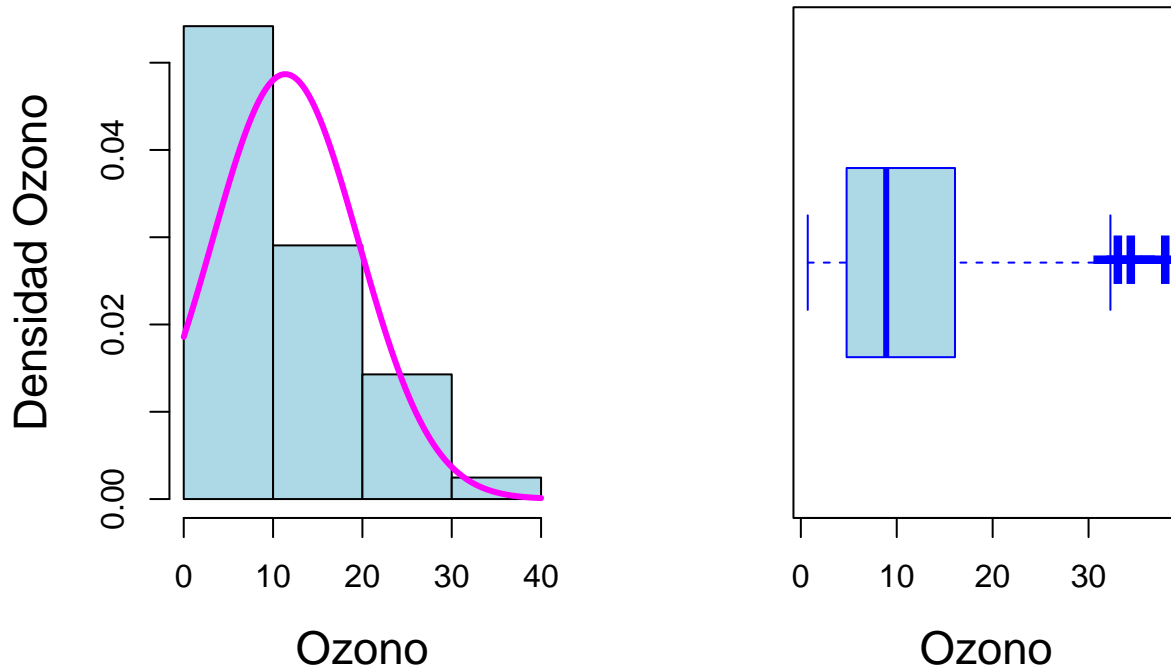
```
## [1] 33.04 34.39 37.98
```

Como podemos ver existen 4 registros atípicos

```

par(mfrow=c(1,2))
hist(Ozono, breaks=5, freq=FALSE, main = "", xlab="Ozono",
     cex.lab=1.4, ylab = "Densidad Ozono", col = "lightblue")
curve( dnorm(x, mean=mean(Ozono), sd=sd(Ozono)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Ozono, main = "", xlab="Ozono",
       cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
       horizontal = TRUE, cex=3)

```



Análisis descriptivo de la variable Pres\_Alt :

```
summary(Pres_Alt)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      5320   5690   5760   5746   5830   5950
```

Desviación típica y rango intercuartílico:

```
sd(Pres_Alt)
```

```
## [1] 113.0277
```

```
IQR(Pres_Alt)
```

```
## [1] 140
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(Pres_Alt, na.rm = FALSE)
```

```
## [1] -0.9499496
```

```
kurtosis(Pres_Alt, na.rm = FALSE)
```

```
## [1] 4.198772
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es mayor a tres, las colas de la variable son más grandes que las de una normal.

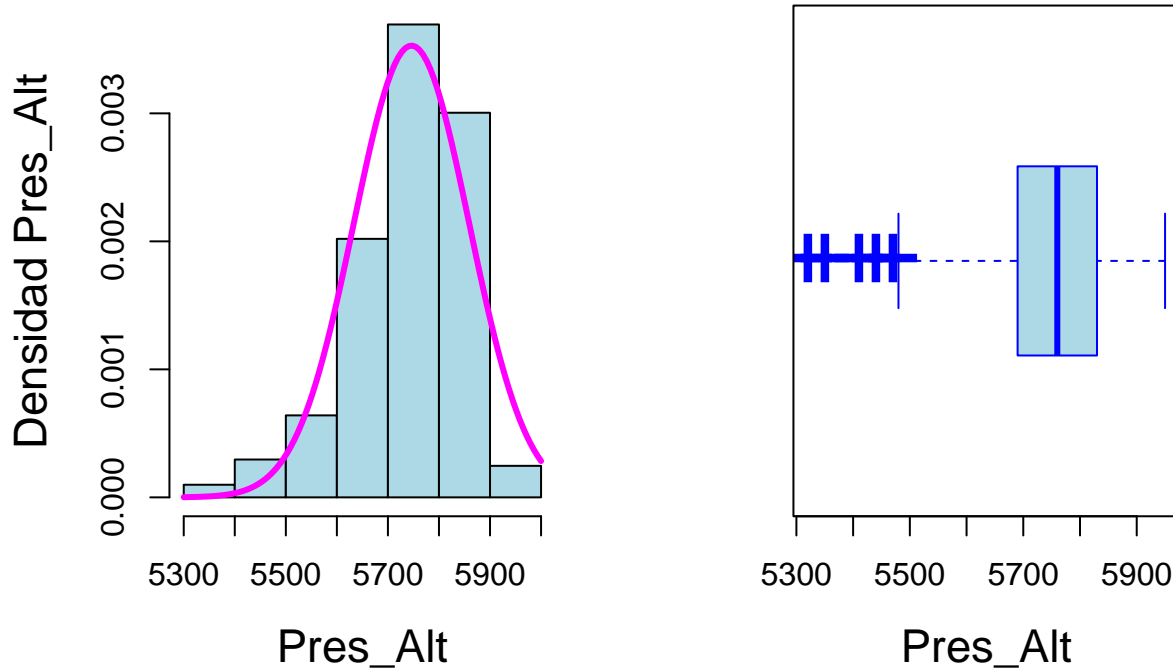
Vemos si hay registros atípicos

```
boxplot.stats(Pres_Alt)$out
```

```
## [1] 5410 5350 5470 5320 5440
```

Como podemos ver existen 5 registros atípicos

```
par(mfrow=c(1,2))
hist(Pres_Alt, breaks=5, freq=FALSE, main = "", xlab="Pres_Alt",
     cex.lab=1.4, ylab = "Densidad Pres_Alt", col = "lightblue")
curve( dnorm(x, mean=mean(Pres_Alt), sd=sd(Pres_Alt)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Pres_Alt, main = "", xlab="Pres_Alt",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Vel\_Viento :

```
summary(Vel_Viento)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.000   3.000   5.000   4.867   6.000  11.000
```

Desviación típica y rango intercuartílico:

```
sd(Vel_Viento)
```

```
## [1] 2.105402
```

```
IQR(Vel_Viento)
```

```
## [1] 3
```



Evaluamos la asimetría y kurtosis

```
library(moments)
skewness(Vel_Viento, na.rm = FALSE)
```

```
## [1] 0.09612047
```

```
kurtosis(Vel_Viento, na.rm = FALSE)
```

```
## [1] 3.378636
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

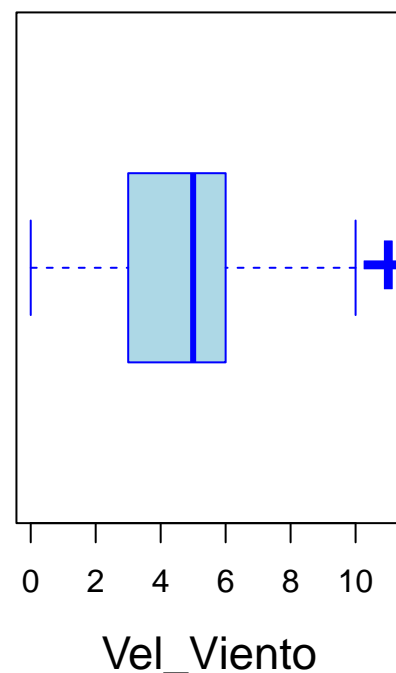
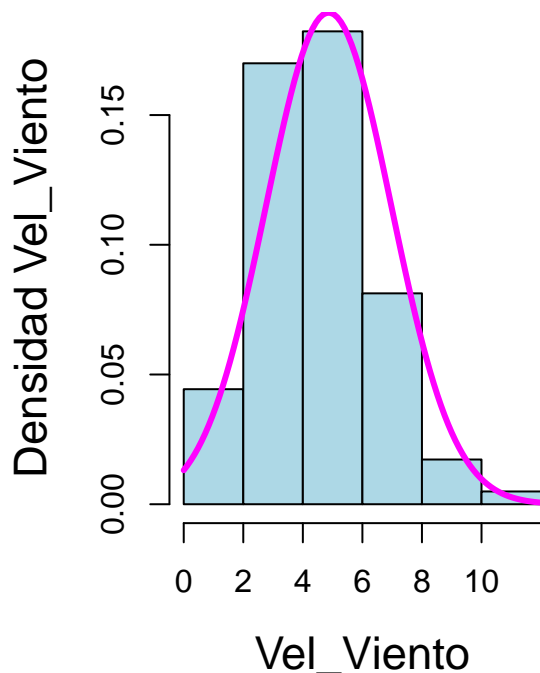
Vemos si hay registros atípicos

```
boxplot.stats(Vel_Viento)$out
```

```
## [1] 11 11
```

Como podemos ver existen 2 registros atípicos

```
par(mfrow=c(1,2))
hist(Vel_Viento, breaks=5,freq=FALSE, main = "", xlab="Vel_Viento",
     cex.lab=1.4, ylab = "Densidad Vel_Viento", col = "lightblue")
curve( dnorm(x,mean=mean(Vel_Viento),sd=sd(Vel_Viento)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Vel_Viento, main = "", xlab="Vel_Viento",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Humedad :

```
summary(Humedad)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    19.00   46.00   64.00   57.61   73.00   93.00
```

Desviación típica y rango intercuartílico:

```
sd(Humedad)
```

```
## [1] 20.84766
```

```
IQR(Humedad)
```

```
## [1] 27
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(Humedad, na.rm = FALSE)
```

```
## [1] -0.6935066
```

```
kurtosis(Humedad, na.rm = FALSE)
```

```
## [1] 2.307891
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

```
boxplot.stats(Humedad)$out
```

```
## integer(0)
```

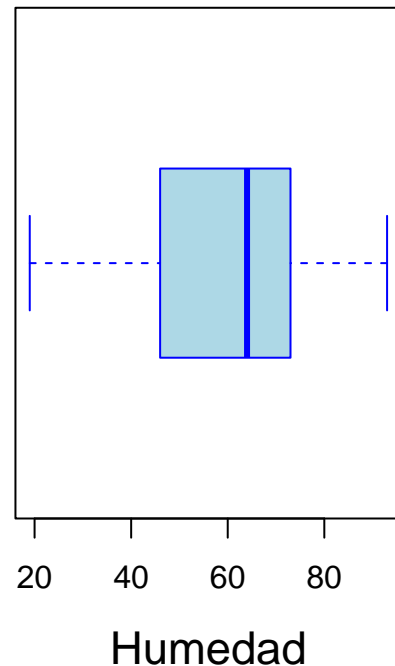
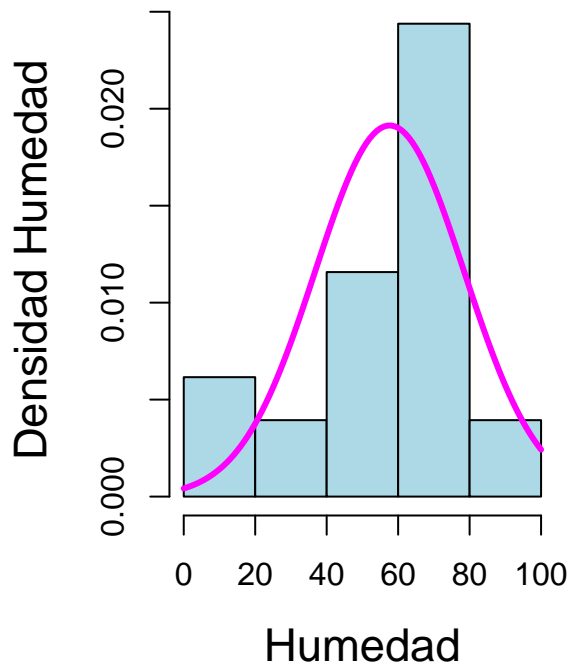
Como podemos ver no existen registros atípicos

```
par(mfrow=c(1,2))
```

```
hist(Humedad, breaks=5, freq=FALSE, main = "", xlab="Humedad",  
      cex.lab=1.4, ylab = "Densidad Humedad", col = "lightblue")
```

```
curve( dnorm(x, mean=mean(Humedad), sd=sd(Humedad)),  
       col="magenta", lwd=3, add=TRUE)
```

```
boxplot(Humedad, main = "", xlab="Humedad",  
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",  
        horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable T\_Sandburg :

```
summary(T_Sandburg)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  25.00  51.50   61.00   61.11  71.00   93.00
```

Desviación típica y rango intercuartílico:

```
sd(T_Sandburg)
```

```
## [1] 14.20647
```

```
IQR(T_Sandburg)
```

```
## [1] 19.5
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(T_Sandburg, na.rm = FALSE)
```

```
## [1] 0.006212875
```

```
kurtosis(T_Sandburg, na.rm = FALSE)
```

```
## [1] 2.510297
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

```
boxplot.stats(T_Sandburg)$out
```

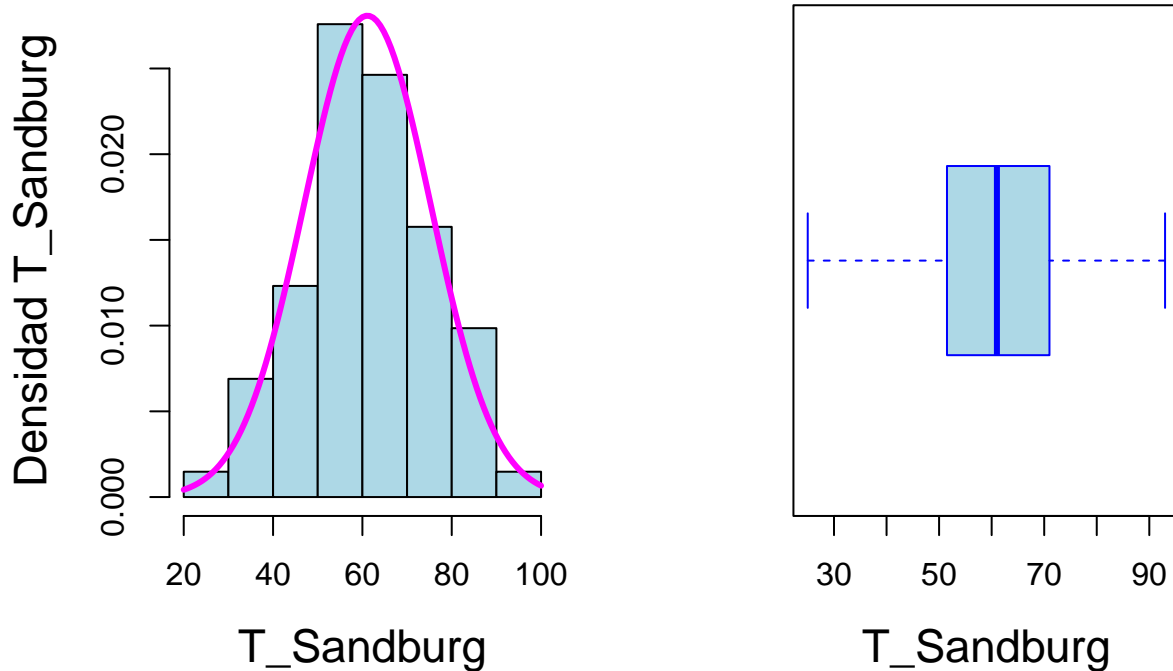
```
## integer(0)
```

Como podemos ver no existen registros atípicos

```

par(mfrow=c(1,2))
hist(T_Sandburg, breaks=5,freq=FALSE, main = "", xlab="T_Sandburg",
     cex.lab=1.4, ylab = "Densidad T_Sandburg", col = "lightblue")
curve( dnorm(x,mean=mean(T_Sandburg),sd=sd(T_Sandburg)),
      col="magenta", lwd=3, add=TRUE)
boxplot(T_Sandburg, main = "", xlab="T_Sandburg",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)

```



- ANÁLISIS DESCRIPTIVO VARIABLE 'T\_ElMonte'

```
summary(T_ElMonte)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  27.68  49.64   56.48   56.54  66.20   82.58
```

Desviación típica y rango intercuartílico:

```
sd(T_ElMonte)
```

```
## [1] 11.74267
```

```
IQR(T_ElMonte)
```

```
## [1] 16.56
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(T_ElMonte, na.rm = FALSE)
```

```
## [1] -0.1025587
```

```
kurtosis(T_ElMonte, na.rm = FALSE)
```

```
## [1] 2.486231
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

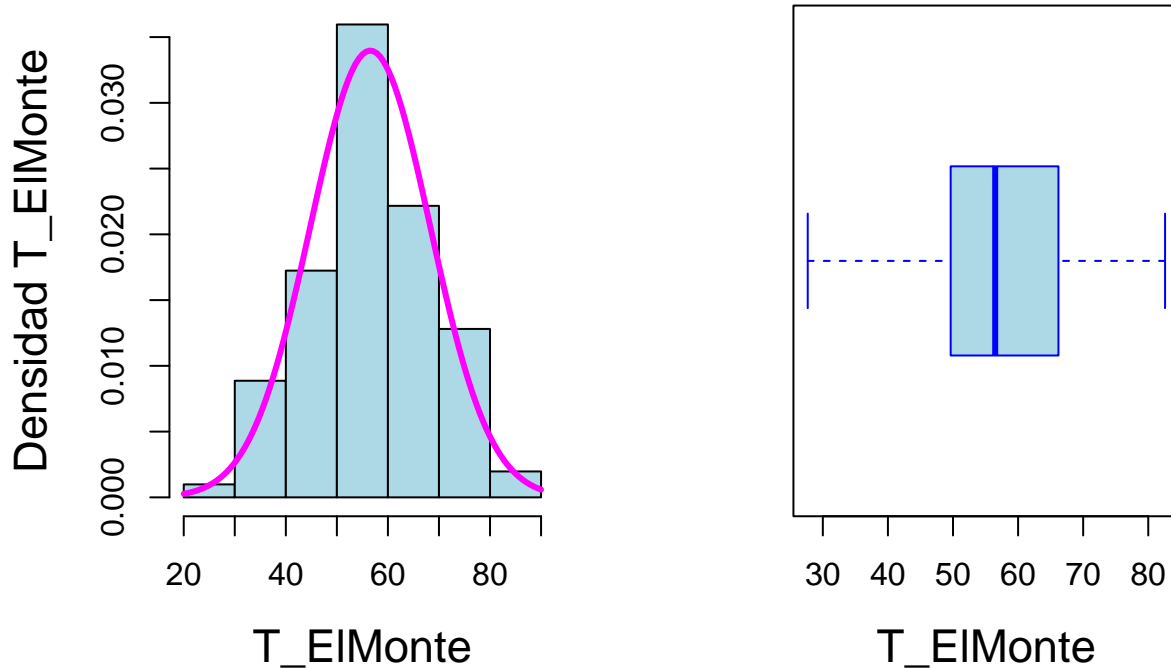
Vemos si hay registros atípicos

```
boxplot.stats(T_ElMonte)$out
```

```
## numeric(0)
```

Como podemos ver no existen registros atípicos

```
par(mfrow=c(1,2))
hist(T_ElMonte, breaks=5, freq=FALSE, main = "", xlab="T_ElMonte",
     cex.lab=1.4, ylab = "Densidad T_ElMonte", col = "lightblue")
curve( dnorm(x, mean=mean(T_ElMonte), sd=sd(T_ElMonte)),
      col="magenta", lwd=3, add=TRUE)
boxplot(T_ElMonte, main = "", xlab="T_ElMonte",
      cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
      horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Inv\_Alt\_b :

```
summary(Inv_Alt_b)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      111    869    2083    2602    5000    5000
```

Desviación típica y rango intercuartílico:

```
sd(Inv_Alt_b)
```

```
## [1] 1859.889
```

```
IQR(Inv_Alt_b)
```

```
## [1] 4131
```

Evaluamos la asimetría y kurtois

```
library(moments)
skewness(Inv_Alt_b, na.rm = FALSE)
```

```
## [1] 0.2355015
```

```
kurtosis(Inv_Alt_b, na.rm = FALSE)
```

```
## [1] 1.374057
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es menor a tres, las colas de la variable son más ligeras a las de una normal.

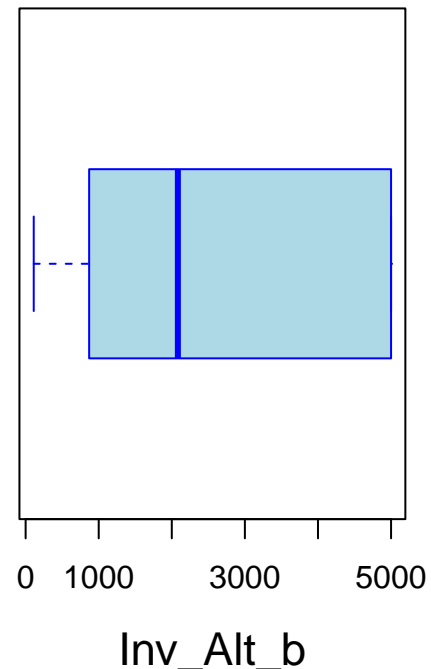
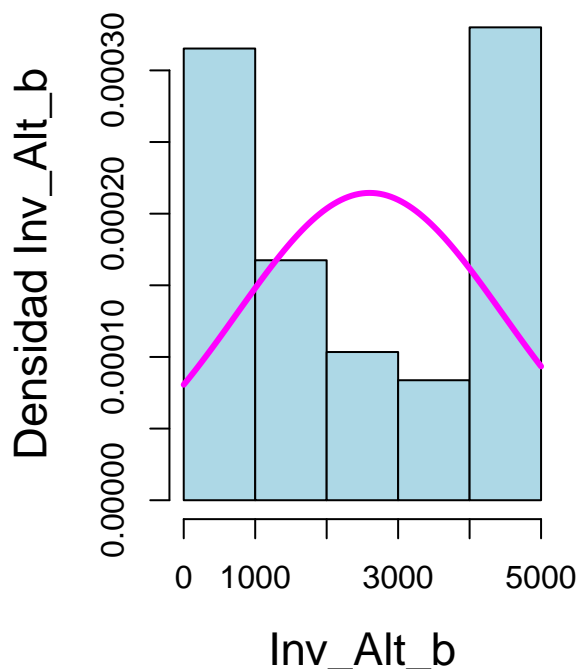
Vemos si hay registros atípicos

```
boxplot.stats(Inv_Alt_b)$out
```

```
## integer(0)
```

Como podemos ver no existen registros atípicos

```
par(mfrow=c(1,2))
hist(Inv_Alt_b, breaks=5, freq=FALSE, main = "", xlab="Inv_Alt_b",
     cex.lab=1.4, ylab = "Densidad Inv_Alt_b", col = "lightblue")
curve( dnorm(x, mean=mean(Inv_Alt_b), sd=sd(Inv_Alt_b)),
       col="magenta", lwd=3, add=TRUE)
boxplot(Inv_Alt_b, main = "", xlab="Inv_Alt_b",
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
        horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Grad\_Pres :

```
summary(Grad_Pres)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -69.00  -14.00   18.00   14.43  43.00  107.00
```

Desviación típica y rango intercuartílico:

```
sd(Grad_Pres)
```

```
## [1] 36.3172
```

```
IQR(Grad_Pres)
```

```
## [1] 57
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(Grad_Pres, na.rm = FALSE)
```

```
## [1] -0.131977
```

```
kurtosis(Grad_Pres, na.rm = FALSE)
```

```
## [1] 2.316879
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es menor a tres, las colas de la variable son más ligeras a las de una normal.

Vemos si hay registros atípicos

```
boxplot.stats(Grad_Pres)$out
```

```
## integer(0)
```

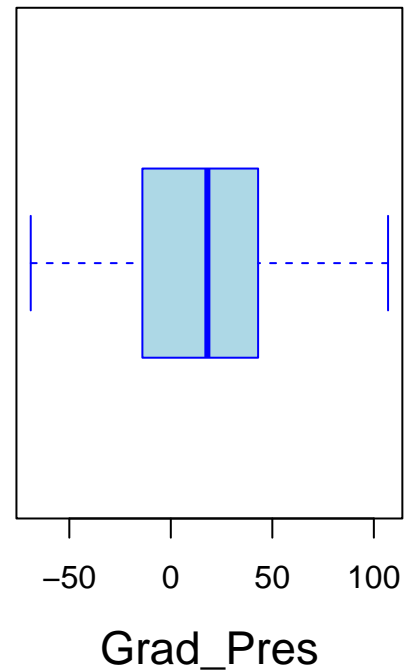
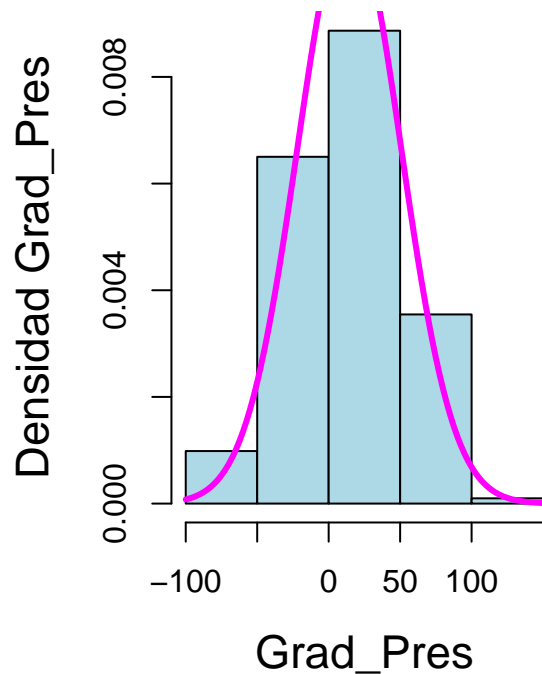
Como podemos ver no existen registros atípicos

```
par(mfrow=c(1,2))
```

```
hist(Grad_Pres, breaks=5, freq=FALSE, main = "", xlab="Grad_Pres",  
     cex.lab=1.4, ylab = "Densidad Grad_Pres", col = "lightblue")
```

```
curve( dnorm(x, mean=mean(Grad_Pres), sd=sd(Grad_Pres)),  
       col="magenta", lwd=3, add=TRUE)
```

```
boxplot(Grad_Pres, main = "", xlab="Grad_Pres",  
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",  
        horizontal = TRUE, cex=3)
```



Análisis descriptivo de la variable Inv\_T\_b :

```
summary(Inv_T_b)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  27.50  51.26   60.98   60.69  70.88   90.68
```

Desviación típica y rango intercuartílico:

```
sd(Inv_T_b)
```

```
## [1] 14.12473
```

```
IQR(Inv_T_b)
```

```
## [1] 19.62
```

Evaluamos la asimetría y kurtois

```
library(moments)
```

```
skewness(Inv_T_b, na.rm = FALSE)
```

```
## [1] -0.1886259
```

```
kurtosis(Inv_T_b, na.rm = FALSE)
```

```
## [1] 2.354789
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis es menor a tres, las colas de la variable son más ligeras a las de una normal.

Vemos si hay registros atípicos

```
boxplot.stats(Inv_T_b)$out
```

```
## numeric(0)
```

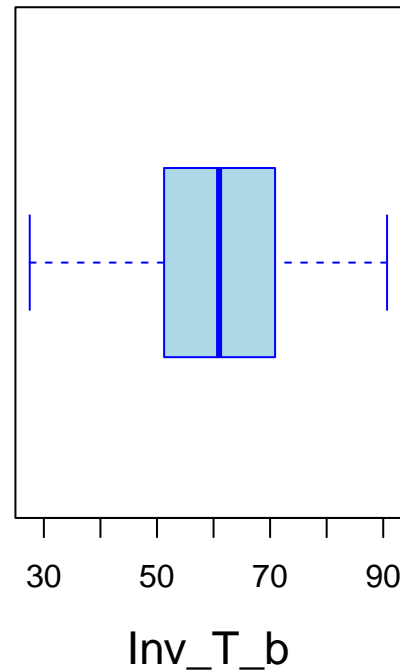
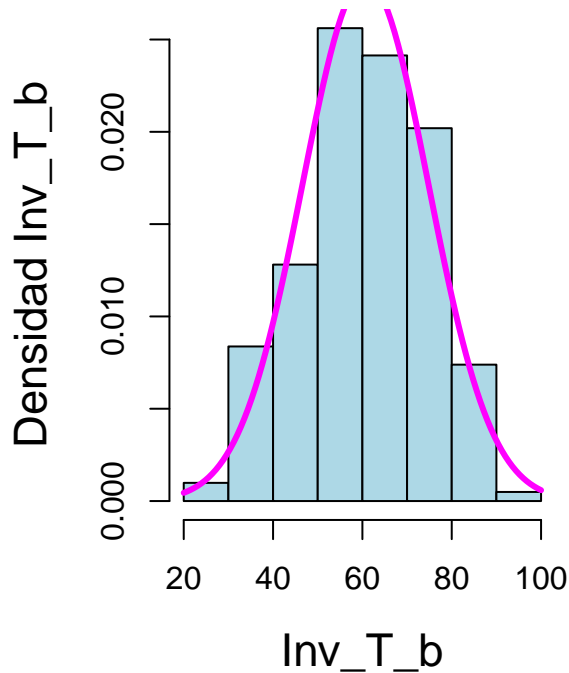
Como podemos ver no existen registros atípicos



```

par(mfrow=c(1,2))
hist(Inv_T_b, breaks=5, freq=FALSE, main = "", xlab="Inv_T_b",
     cex.lab=1.4, ylab = "Densidad Inv_T_b", col = "lightblue")
curve( dnorm(x, mean=mean(Inv_T_b), sd=sd(Inv_T_b)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Inv_T_b, main = "", xlab="Inv_T_b",
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
        horizontal = TRUE, cex=3)

```



Análisis descriptivo de la variable Visibilidad :

```
summary(Visibilidad)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.0   60.0   100.0  122.2  150.0  350.0
```

Desviación típica y rango intercuartílico:

```
sd(Visibilidad)
```

```
## [1] 81.17132
```

```
IQR(Visibilidad)
```

```
## [1] 90
```

Evaluamos la asimetría y kurtosis

```
library(moments)
```

```
skewness(Visibilidad, na.rm = FALSE)
```

```
## [1] 0.8067613
```

```
kurtosis(Visibilidad, na.rm = FALSE)
```

```
## [1] 2.903426
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simétrica y al ser el coeficiente de Kurtosis próximo a tres, las colas de la variable son próximas a las de una normal.

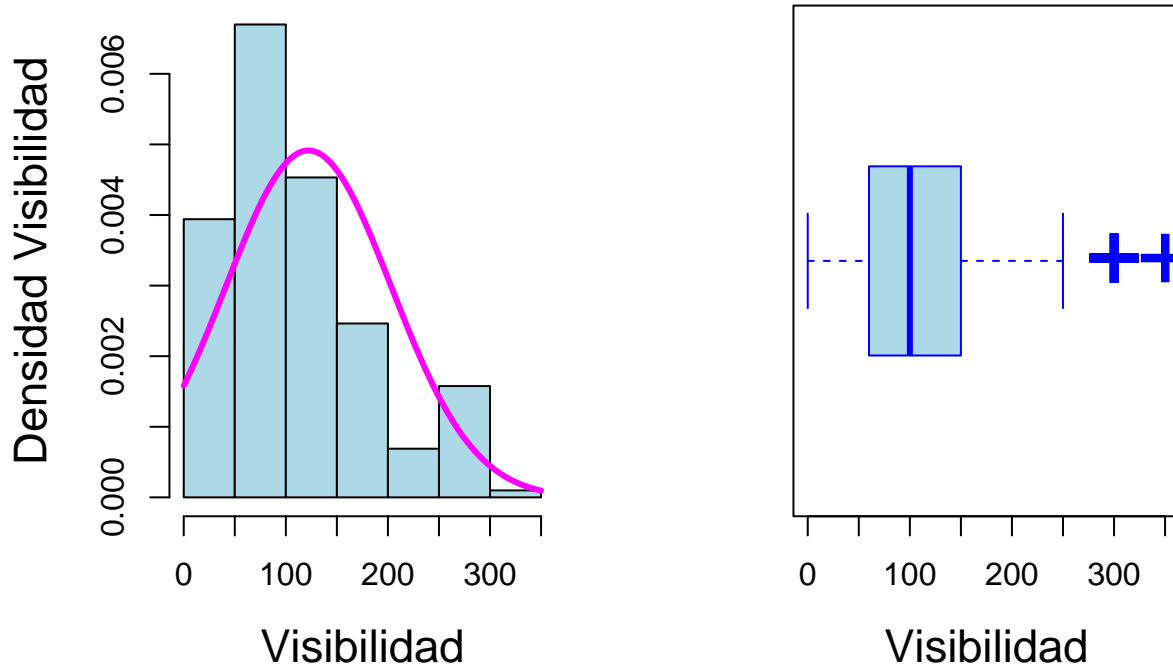
Vemos si hay registros atípicos

```
boxplot.stats(Visibilidad)$out
```

```
## [1] 350 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300
```

Como podemos ver no existen registros atípicos

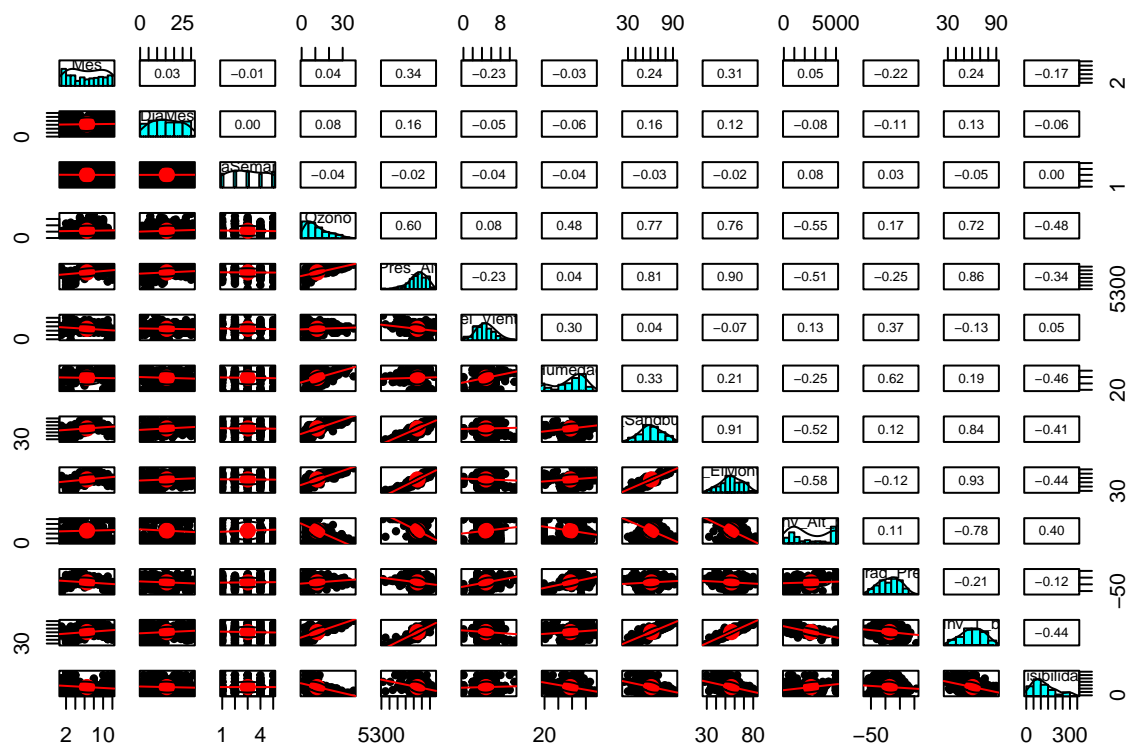
```
par(mfrow=c(1,2))
hist(Visibilidad, breaks=5, freq=FALSE, main = "", xlab="Visibilidad",
     cex.lab=1.4, ylab = "Densidad Visibilidad", col = "lightblue")
curve( dnorm(x, mean=mean(Visibilidad), sd=sd(Visibilidad)),
      col="magenta", lwd=3, add=TRUE)
boxplot(Visibilidad, main = "", xlab="Visibilidad",
        cex.lab=1.4, border = "blue", col= "lightblue", pch="+",
        horizontal = TRUE, cex=3)
```



## 2. Análisis de correlación

- Correlaciones simples bivariantes(análisis gráfico y numérico):

```
library(psych)
pairs.panels(OzonoLA, smooth = TRUE, density=TRUE, digits = 2,
             ellipses=TRUE, method="pearson", pch = 20,
             lm=TRUE, cor=TRUE)
```



```
cor(OzonoLA)
```

##	Mes	DiaMes	DiaSemana	Ozono	Pres_Alt
## Mes	1.000000000	0.029780944	-6.406562e-03	0.04417525	0.33793183
## DiaMes	0.029780944	1.000000000	3.418381e-03	0.08364060	0.15808064
## DiaSemana	-0.006406562	0.003418381	1.000000e+00	-0.03750993	-0.02206218
## Ozono	0.044175248	0.083640605	-3.750993e-02	1.000000000	0.59612683
## Pres_Alt	0.337931827	0.158080640	-2.206218e-02	0.59612683	1.000000000
## Vel_Viento	-0.226893006	-0.046090839	-3.667633e-02	0.08179858	-0.23161673
## Humedad	-0.034727288	-0.064739863	-3.855381e-02	0.47947091	0.03869121
## T_Sandburg	0.235445072	0.157156363	-3.035349e-02	0.77335204	0.80633038
## T_ElMonte	0.314323892	0.117127229	-2.481044e-02	0.76001956	0.89689385
## Inv_Alt_b	0.045305170	-0.082352709	7.998485e-02	-0.55196217	-0.50891157
## Grad_Pres	-0.218837079	-0.111239793	3.418479e-02	0.17391799	-0.24549047
## Inv_T_b	0.236540625	0.127530054	-5.365959e-02	0.71756186	0.85642134
## Visibilidad	-0.167796386	-0.057896954	-8.572216e-06	-0.47629112	-0.34272720
##	Vel_Viento	Humedad	T_Sandburg	T_ElMonte	Inv_Alt_b
## Mes	-0.22689301	-0.03472729	0.23544507	0.31432389	0.04530517
## DiaMes	-0.04609084	-0.06473986	0.15715636	0.11712723	-0.08235271
## DiaSemana	-0.03667633	-0.03855381	-0.03035349	-0.02481044	0.07998485
## Ozono	0.08179858	0.47947091	0.77335204	0.76001956	-0.55196217
## Pres_Alt	-0.23161673	0.03869121	0.80633038	0.89689385	-0.50891157
## Vel_Viento	1.00000000	0.30356343	0.04122208	-0.06983510	0.12834881
## Humedad	0.30356343	1.00000000	0.33132296	0.21158607	-0.24703914
## T_Sandburg	0.04122208	0.33132296	1.00000000	0.91396229	-0.51539621
## T_ElMonte	-0.06983510	0.21158607	0.91396229	1.00000000	-0.57965832
## Inv_Alt_b	0.12834881	-0.24703914	-0.51539621	-0.57965832	1.00000000
## Grad_Pres	0.37328762	0.62433536	0.11765666	-0.12091597	0.11350236
## Inv_T_b	-0.12959891	0.19101936	0.84310310	0.93080989	-0.78286145
## Visibilidad	0.04534341	-0.45750232	-0.41038641	-0.43897902	0.39669789
##	Grad_Pres	Inv_T_b	Visibilidad		

```
## Mes      -0.21883708  0.23654062 -1.677964e-01
## DiaMes   -0.11123979  0.12753005 -5.789695e-02
## DiaSemana 0.03418479 -0.05365959 -8.572216e-06
## Ozono     0.17391799  0.71756186 -4.762911e-01
## Pres_Alt  -0.24549047  0.85642134 -3.427272e-01
## Vel_Viento 0.37328762 -0.12959891  4.534341e-02
## Humedad   0.62433536  0.19101936 -4.575023e-01
## T_Sandburg 0.11765666  0.84310310 -4.103864e-01
## T_ElMonte -0.12091597  0.93080989 -4.389790e-01
## Inv_Alt_b  0.11350236 -0.78286145  3.966979e-01
## Grad_Pres  1.00000000 -0.20663872 -1.200549e-01
## Inv_T_b    -0.20663872  1.00000000 -4.377177e-01
## Visibilidad -0.12005488 -0.43771768  1.000000e+00
```

- Correlaciones parciales:

```
partial.r(OzonoLA)
```

```
##           Mes      DiaMes      DiaSemana      Ozono      Pres_Alt
## Mes      1.000000000 -0.01473632 -0.029646884 -0.239632308 -0.008364478
## DiaMes   -0.014736319 1.000000000  0.017131467  0.023224469  0.074079502
## DiaSemana -0.029646884 0.01713147  1.000000000 -0.015463849 -0.014083279
## Ozono     -0.239632308 0.02322447 -0.015463849  1.000000000 -0.134822542
## Pres_Alt  -0.008364478 0.07407950 -0.014083279 -0.134822542  1.000000000
## Vel_Viento -0.192898039 0.01519492 -0.052672027 -0.040039195 -0.292700944
## Humedad   0.160860221 -0.03992322 -0.050358261  0.262774072 -0.095321178
## T_Sandburg 0.008578204  0.20842819 -0.037515653  0.141155532  0.108888567
## T_ElMonte 0.131026789 -0.12847809  0.050717722  0.312487718  0.344311253
## Inv_Alt_b  0.230043843 -0.02868566  0.036820690 -0.111064127  0.120880379
## Grad_Pres -0.127208517 -0.13665426  0.068684046  0.001780773 -0.044096421
## Inv_T_b    0.048692150 -0.02999001 -0.008230412 -0.076866881  0.140848869
## Visibilidad -0.108506988 -0.06279200 -0.037003418 -0.074160846  0.014979648
##           Vel_Viento      Humedad      T_Sandburg      T_ElMonte      Inv_Alt_b
## Mes      -0.19289804  0.16086022  0.008578204  0.13102679  0.23004384
## DiaMes    0.01519492 -0.03992322  0.208428191 -0.12847809 -0.02868566
## DiaSemana -0.05267203 -0.05035826 -0.037515653  0.05071772  0.03682069
## Ozono     -0.04003920  0.26277407  0.141155532  0.31248772 -0.11106413
## Pres_Alt  -0.29270094 -0.09532118  0.108888567  0.34431125  0.12088038
## Vel_Viento 1.00000000  0.15651029  0.089387359  0.11902520  0.11170466
## Humedad   0.15651029  1.00000000 -0.044727403 -0.04353431 -0.05762633
## T_Sandburg 0.08938736 -0.04472740  1.000000000  0.35489823  0.18928541
## T_ElMonte 0.11902520 -0.04353431  0.354898232  1.00000000  0.39942102
## Inv_Alt_b  0.11170466 -0.05762633  0.189285412  0.39942102  1.00000000
## Grad_Pres 0.05542912  0.50554293  0.498084949 -0.05195235 -0.15571589
## Inv_T_b    0.01217894  0.06712657  0.229456614  0.57959707 -0.81884177
## Visibilidad 0.11148387 -0.32142715  0.085393863 -0.12200008  0.09905698
##           Grad_Pres      Inv_T_b      Visibilidad
## Mes      -0.127208517  0.048692150 -0.10850699
## DiaMes   -0.136654263 -0.029990011 -0.06279200
## DiaSemana 0.068684046 -0.008230412 -0.03700342
## Ozono     0.001780773 -0.076866881 -0.07416085
## Pres_Alt  -0.044096421  0.140848869  0.01497965
## Vel_Viento 0.055429122  0.012178940  0.11148387
## Humedad   0.505542925  0.067126570 -0.32142715
## T_Sandburg 0.498084949  0.229456614  0.08539386
```

```
## T_ElMonte -0.051952353 0.579597071 -0.12200008
## Inv_Alt_b -0.155715887 -0.818841765 0.09905698
## Grad_Pres 1.000000000 -0.326942874 0.01948577
## Inv_T_b -0.326942874 1.000000000 0.03558761
## Visibilidad 0.019485768 0.035587611 1.00000000
```

### 3. Modelo matemático

$$\mathbb{E}(\vec{Y}|\mathbf{X}) = \beta_0 + \sum_{i=1}^n \beta_i X_{ij} \quad (1)$$

```
MOD_FULL <- lm(Ozono~., data=OzonoLA)
MOD_FULL
```

```
##
## Call:
## lm(formula = Ozono ~ ., data = OzonoLA)
##
## Coefficients:
## (Intercept)      Mes      DiaMes      DiaSemana      Pres_Alt      Vel_Viento
## 55.4279486   -0.3431326   0.0120308   -0.0473689   -0.0133495   -0.0959961
## Humedad      T_Sandburg      T_ElMonte      Inv_Alt_b      Grad_Pres      Inv_T_b
## 0.0880372    0.1366231    0.5597690   -0.0006176    0.0003624   -0.1244500
## Visibilidad
## -0.0049469
```

```
coef(MOD_FULL)
```

```
## (Intercept)      Mes      DiaMes      DiaSemana      Pres_Alt
## 55.4279486216 -0.3431325880 0.0120307523 -0.0473688814 -0.0133495197
## Vel_Viento      Humedad      T_Sandburg      T_ElMonte      Inv_Alt_b
## -0.0959961221 0.0880371866 0.1366230525 0.5597690142 -0.0006175971
## Grad_Pres      Inv_T_b      Visibilidad
## 0.0003623595 -0.1244500321 -0.0049468590
```

Ozono = 55.428 - 0.343*Mes* + 0.012*DiaMes* - 0.047*DiaSemana* - 0.0133*Pres\_Alt* - 0.096*Vel\_Viento* + 0.088*Humedad* + 0.1366*T\_Sandburg* + 0.5598*T\_ElMonte* - 0.0006*Inv\_Alt\_b* + 0.0004*Grad\_Pres* - 0.124*Inv\_T\_b* - 0.005*Visibilidad*

```
(MSSR <- summary(MOD_FULL)$sigma^2)
```

```
## [1] 19.24102
```

```
(gl.R <- MOD_FULL$df)
```

```
## [1] 190
```

```
(gl.E <- MOD_FULL$rank)
```

```
## [1] 13
```

### 4. Análisis de multicolinealidad

```
summary(MOD_FULL)
```

```
##
## Call:
## lm(formula = Ozono ~ ., data = OzonoLA)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0342  -2.8582  -0.4764   2.6584  12.7160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.4279486  37.6060409   1.474  0.142161
## Mes          -0.3431326   0.1008551  -3.402  0.000815 ***
## DiaMes         0.0120308   0.0375710   0.320  0.749158
## DiaSemana     -0.0473689   0.2222014  -0.213  0.831415
## Pres_Alt      -0.0133495   0.0071178  -1.876  0.062255 .
## Vel_Viento    -0.0959961   0.1737974  -0.552  0.581361
## Humedad        0.0880372   0.0234515   3.754  0.000231 ***
## T_Sandburg     0.1366231   0.0695151   1.965  0.050828 .
## T_ElMonte      0.5597690   0.1234488   4.534  1.02e-05 ***
## Inv_Alt_b      -0.0006176   0.0004009  -1.540  0.125116
## Grad_Pres       0.0003624   0.0147623   0.025  0.980443
## Inv_T_b        -0.1244500   0.1171095  -1.063  0.289275
## Visibilidad    -0.0049469   0.0048259  -1.025  0.306638
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.386 on 190 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.7133
## F-statistic: 42.89 on 12 and 190 DF,  p-value: < 2.2e-16
```

Obtenemos que muchos de los coeficientes son no significativos, por lo que debemos hacer una selección de las variables. No obstante, como esto se puede deber a la presencia de multicolinealidad, vamos a analizarla.

Para ello, utilizaremos la librería “mctest”, que proporciona un análisis completo de multicolinealidad:

```
library(mctest)
mctest(MOD_FULL, type="o")

##
## Call:
## omcdiag(mod = mod, Inter = TRUE, detr = detr, red = red, conf = conf,
##      theil = theil, cn = cn)
##
##
## Overall Multicollinearity Diagnostics
##
##              MC Results detection
## Determinant |X'X|:           0.0001           1
## Farrar Chi-Square:        1900.8790           1
## Red Indicator:             0.3656           0
## Sum of Lambda Inverse:     85.6887           1
## Theil's Method:            -1.2174           0
## Condition Number:         586.6642           1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

Este test proporciona 6 medidas, de las cuales 4 indican que estamos ante un caso en el que la multicolinealidad está presente.

Para solucionar esto y conseguir un ajuste correcto, sobre el que hacer inferencia debemos hacer una selección de variables.

## 5. Selección del modelo

Para hacer la selección del modelo, utilizaremos la selección sistemática por STEPWISE, utilizando como criterio el AIC del modelo. Elegimos este método de selección por ser el mejor, al permitir incluir y eliminar variables a lo largo del proceso.

Primero, definimos el modelo con solo el intercept.

```
Mod_NULL <- lm(Ozono ~ 1, data = OzonoLA)
```

Ahora, aplicaremos la siguiente función para obtener el modelo óptimo:

```
stepMod <- step(Mod_NULL, direction = "both", trace = 1,
               scope = list(lower = Mod_NULL,
                             upper = MOD_FULL) )
```

```
## Start:  AIC=854.91
## Ozono ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + T_Sandburg  1    8108.8  5449.4  671.88
## + T_ElMonte  1    7831.6  5726.6  681.95
## + Inv_T_b    1    6981.0  6577.1  710.06
## + Pres_Alt   1    4818.1  8740.0  767.78
## + Inv_Alt_b  1    4130.7  9427.5  783.15
## + Humedad    1    3116.9 10441.2  803.88
## + Visibilidad 1    3075.7 10482.4  804.68
## + Grad_Pres  1     410.1 13148.0  850.68
## <none>                13558.1  854.91
## + DiaMes      1      94.8 13463.3  855.49
## + Vel_Viento  1      90.7 13467.4  855.55
## + Mes         1      26.5 13531.7  856.52
## + DiaSemana   1      19.1 13539.1  856.63
##
## Step:  AIC=671.88
## Ozono ~ T_Sandburg
##
##           Df Sum of Sq    RSS    AIC
## + Humedad    1      759.0  4690.4  643.43
## + Inv_Alt_b  1      434.3  5015.0  657.02
## + Visibilidad 1      411.8  5037.6  657.93
## + Mes        1      273.0  5176.4  663.45
## + T_ElMonte  1      233.1  5216.3  665.01
## + Inv_T_b    1      201.4  5247.9  666.23
## + Grad_Pres  1       94.5  5354.8  670.33
## <none>                5449.4  671.88
## + Vel_Viento  1       33.8  5415.5  672.62
## + Pres_Alt    1       29.2  5420.2  672.79
## + DiaMes      1       20.0  5429.4  673.14
## + DiaSemana   1        2.7  5446.7  673.78
## - T_Sandburg  1    8108.8 13558.1  854.91
##
## Step:  AIC=643.43
```

```

## Ozono ~ T_Sandburg + Humedad
##
##           Df Sum of Sq    RSS    AIC
## + T_ElMonte  1    505.3  4185.1 622.29
## + Inv_T_b    1    371.8  4318.5 628.67
## + Inv_Alt_b  1    335.7  4354.6 630.35
## + Mes        1    175.2  4515.2 637.70
## + Visibilidad 1    116.1  4574.2 640.34
## + Grad_Pres  1     92.0  4598.4 641.41
## <none>                4690.4 643.43
## + Pres_Alt   1     41.5  4648.9 643.63
## + Vel_Viento 1      7.8  4682.6 645.09
## + DiaMes     1      1.0  4689.3 645.39
## + DiaSemana  1      0.6  4689.7 645.40
## - Humedad    1    759.0  5449.4 671.88
## - T_Sandburg 1   5750.9 10441.2 803.88
##
## Step:  AIC=622.29
## Ozono ~ T_Sandburg + Humedad + T_ElMonte
##
##           Df Sum of Sq    RSS    AIC
## + Mes        1    358.12 3827.0 606.13
## + Inv_Alt_b  1    126.22 4058.9 618.08
## + Pres_Alt   1    108.61 4076.5 618.96
## <none>                4185.1 622.29
## + Visibilidad 1     19.69 4165.4 623.34
## + Inv_T_b    1     18.92 4166.2 623.37
## + Grad_Pres  1     11.28 4173.8 623.75
## + Vel_Viento 1      3.68 4181.4 624.11
## + DiaMes     1      1.50 4183.6 624.22
## + DiaSemana  1      0.65 4184.4 624.26
## - T_Sandburg 1    100.19 4285.3 625.10
## - T_ElMonte  1     505.29 4690.4 643.43
## - Humedad    1   1031.23 5216.3 665.01
##
## Step:  AIC=606.13
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes
##
##           Df Sum of Sq    RSS    AIC
## + Pres_Alt   1     70.84 3756.1 604.34
## <none>                3827.0 606.13
## + Inv_Alt_b  1     34.70 3792.3 606.28
## + Visibilidad 1     34.59 3792.4 606.29
## - T_Sandburg 1     63.90 3890.9 607.50
## + Vel_Viento 1      2.21 3824.8 608.02
## + DiaMes     1      1.48 3825.5 608.06
## + Inv_T_b    1      1.30 3825.7 608.07
## + Grad_Pres  1      0.91 3826.1 608.09
## + DiaSemana  1      0.74 3826.2 608.09
## - Mes        1    358.12 4185.1 622.29
## - T_ElMonte  1    688.22 4515.2 637.70
## - Humedad    1    946.87 4773.8 649.01
##
## Step:  AIC=604.34

```



```

## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt
##
##           Df Sum of Sq    RSS    AIC
## + Inv_Alt_b   1      41.91 3714.2 604.06
## <none>                        3756.1 604.34
## + Visibilidad  1      36.56 3719.6 604.36
## + Vel_Viento   1      18.08 3738.0 605.36
## + Inv_T_b      1       6.40 3749.7 606.00
## + DiaMes       1       3.86 3752.3 606.13
## - Pres_Alt     1      70.84 3827.0 606.13
## - T_Sandburg   1      72.62 3828.7 606.23
## + DiaSemana    1       0.92 3755.2 606.29
## + Grad_Pres    1       0.07 3756.1 606.34
## - Mes          1     320.34 4076.5 618.96
## - T_ElMonte    1     664.43 4420.6 635.41
## - Humedad      1     678.82 4434.9 636.07
##
## Step:  AIC=604.06
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt + Inv_Alt_b
##
##           Df Sum of Sq    RSS    AIC
## <none>                        3714.2 604.06
## - Inv_Alt_b   1      41.91 3756.1 604.34
## + Inv_T_b      1      26.12 3688.1 604.63
## + Visibilidad  1      25.74 3688.5 604.65
## + Vel_Viento   1       8.67 3705.5 605.59
## + DiaMes       1       2.73 3711.5 605.91
## + Grad_Pres    1       1.61 3712.6 605.98
## + DiaSemana    1       0.19 3714.0 606.05
## - Pres_Alt     1      78.05 3792.3 606.28
## - T_Sandburg   1      87.87 3802.1 606.81
## - Mes          1     228.30 3942.5 614.17
## - T_ElMonte    1     515.95 4230.2 628.47
## - Humedad      1     596.56 4310.8 632.30
summary((stepMod))

##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes +
##     Pres_Alt + Inv_Alt_b, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0749  -3.0474  -0.1831   2.7775  12.6395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.3444845 35.1290934   1.462 0.145454
## T_Sandburg   0.1242673  0.0577088   2.153 0.032513 *
## Humedad      0.0975694  0.0173897   5.611 6.80e-08 ***
## T_ElMonte    0.4743962  0.0909164   5.218 4.59e-07 ***
## Mes         -0.3324536  0.0957810  -3.471 0.000638 ***
## Pres_Alt    -0.0134013  0.0066034  -2.029 0.043763 *
## Inv_Alt_b    -0.0003211  0.0002159  -1.487 0.138571

```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.353 on 196 degrees of freedom
## Multiple R-squared:  0.7261, Adjusted R-squared:  0.7177
## F-statistic: 86.58 on 6 and 196 DF,  p-value: < 2.2e-16
```

El modelo resultante de la selección secuencial es:  $Ozono = 51.3444845 - 0.3324536Mes - 0.0134013 Pres\_Alt + 0.0975694Humedad + 0.1242673T\_Sandburg + 0.4743962T\_ElMonte - 0.0003211Inv\_Alt\_b$

No obstante, con un 10% de significación, la variable `Inv_Alt_b` no es significativa, por lo que examinaremos si se debe excluir del modelo:

```
ajuste_sin_inv_alt_b <- update(stepMod, .~.-Inv_Alt_b)
```

Lo comprobaremos con un anova de modelos anidados:

```
anova(ajuste_sin_inv_alt_b, stepMod)
```

```
## Analysis of Variance Table
##
## Model 1: Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt
## Model 2: Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt + Inv_Alt_b
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      197 3756.1
## 2      196 3714.2   1    41.913 2.2117 0.1386
```

Prueba no significativa, por lo que nos quedamos con el modelo sin la variable.

```
ajuste <- ajuste_sin_inv_alt_b
```

Comprobaremos si es mejor que el modelo completo, utilizando un anova de modelos anidados:

```
anova(ajuste, MOD_FULL)
```

```
## Analysis of Variance Table
##
## Model 1: Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt
## Model 2: Ozono ~ Mes + DiaMes + DiaSemana + Pres_Alt + Vel_Viento + Humedad +
##           T_Sandburg + T_ElMonte + Inv_Alt_b + Grad_Pres + Inv_T_b +
##           Visibilidad
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      197 3756.1
## 2      190 3655.8   7    100.33 0.7449 0.6342
```

El resultado es no significativo, por lo que la selección ha merecido la pena.

## 6. Inferencia modelo

Ahora ya podemos comenzar la inferencia.

```
summary(ajuste)
```

```
##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes +
##     Pres_Alt, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -10.7435 -2.9604 0.0761 2.9540 12.5572
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.210973  34.993284  1.292  0.1979
## T_Sandburg  0.111767  0.057269  1.952  0.0524 .
## Humedad    0.102313  0.017147  5.967 1.11e-08 ***
## T_ElMonte   0.514331  0.087127  5.903 1.54e-08 ***
## Mes        -0.375442  0.091596 -4.099 6.06e-05 ***
## Pres_Alt   -0.012738  0.006609 -1.928 0.0554 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.367 on 197 degrees of freedom
## Multiple R-squared:  0.723, Adjusted R-squared:  0.7159
## F-statistic: 102.8 on 5 and 197 DF, p-value: < 2.2e-16
```

Las únicas variables que parecen ser significativas son Mes, Humedad y T\_ElMonte. También podemos considerar que son bastante significativas, pero no tanto, las variables T\_Sandburg y Pres\_Alt. Por otra parte, según el coeficiente de bondad, con este ajuste podemos explicar el 73,04% de la variabilidad de los datos. Por último, gracias a la última línea del summary deducimos que es mejor este ajuste en comparación al modelo que contiene únicamente el intercept, debido al p-valor < 2.2e-16.

## 7. Validación modelo seleccionado

Por abreviar la notación, tenemos:

```
MS <- ajuste # Ajuste modelo elegido.
MC <- MOD_FULL # Ajuste modelo completo
```

Primero, calculamos el coeficiente de robusted del ajuste:

```
library(DAAG)
( B2 <- sum(residuals(MS)^2)/press(MS) )
```

```
## [1] 0.9442346
```

Elevado y superior al del modelo completo

```
sum(residuals(MC)^2)/press(MC)
```

```
## [1] 0.8823052
```

Haremos una validación del tipo LOOCV (Leave One Out Cross Validation): Primero, para MS:

```
class(OzonoLA) # ya es un data frame
```

```
## [1] "data.frame"
```

```
set.seed(5198)
```

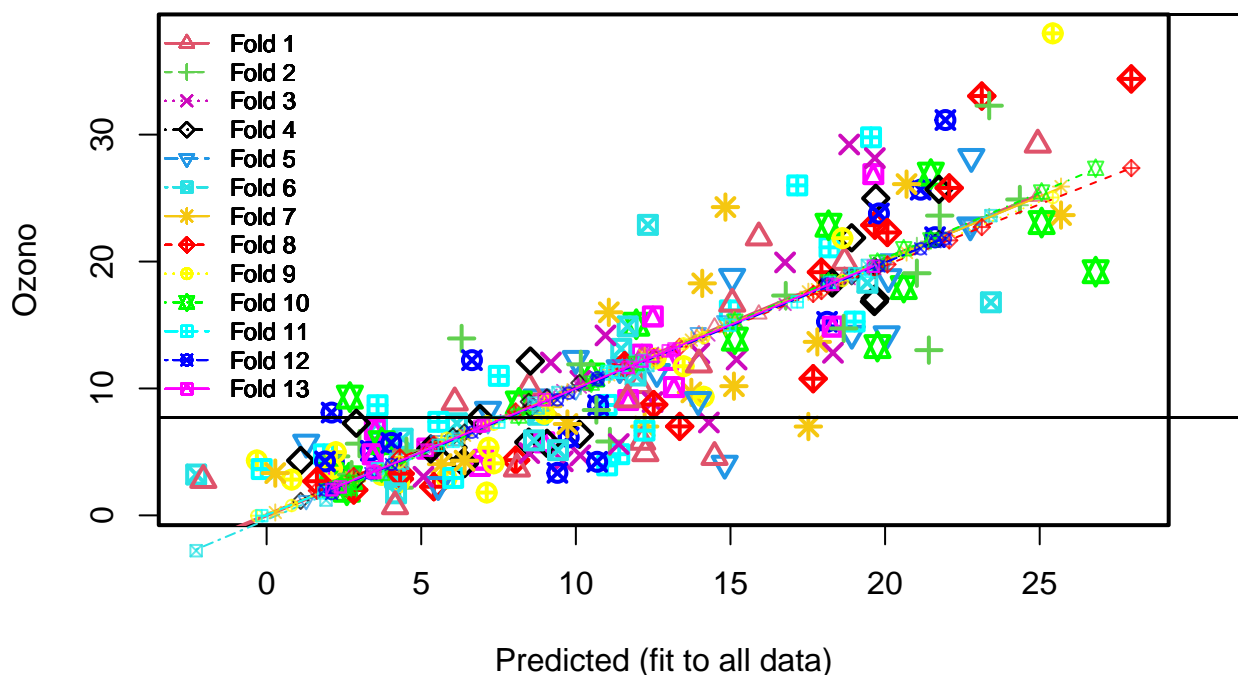
```
cv_k3_MS <- cv.lm(data=OzonoLA,form.lm= formula(MS),m=length(OzonoLA))
```

```
## Warning in cv.lm(data = OzonoLA, form.lm = formula(MS), m = length(OzonoLA)):
```

```
##
```

```
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```

## Small symbols show cross-validation predicted values



```
##
## fold 1
## Observations in test set: 15
##          13      14      26      32      37      52      62
## Predicted   8.094337 12.249041 12.14553 12.040563 -2.046856 6.083821 8.485786
## cvpred      7.928546 12.448259 12.50498 12.175213 -2.091523 5.861667 8.638667
## Ozono       3.690000 4.900000 5.80000 10.270000 2.790000 8.900000 10.070000
## CV residual -4.238546 -7.548259 -6.70498 -1.905213 4.881523 3.038333 1.431333
##          68      80      112      146      149      160
## Predicted  15.916928 18.688605 24.938009 14.48427 15.057249 11.668834
## cvpred     15.858296 18.855042 25.131695 14.83124 15.318839 11.467194
## Ozono      21.900000 19.980000 29.210000 4.60000 16.680000 9.140000
## CV residual 6.041704 1.124958 4.078305 -10.23124 1.361161 -2.327194
##          177      203
## Predicted  13.978194 4.152568
## cvpred     14.219525 4.540370
## Ozono      11.890000 0.720000
## CV residual -2.329525 -3.820370
##
## Sum of squares = 345.01    Mean square = 23    n = 15
##
## fold 2
## Observations in test set: 16
##          20      40      51      69      83      94
## Predicted  4.289757 3.001376 6.305743 16.7880673 24.3579150 10.161111
## cvpred     4.047500 3.000899 6.044460 16.8250824 24.4386217 10.180638
## Ozono      2.180000 5.650000 13.940000 17.3200000 24.8900000 11.900000
## CV residual -1.867500 2.649101 7.895540 0.4949176 0.4513783 1.719362
##          114      123      127      129      133      134
```

```

## Predicted    21.415164 11.100279 21.766706 23.36661 21.028438 18.660654
## cvpred      21.572202 11.159753 21.842375 23.51191 21.238023 18.726176
## Ozono       13.020000  5.820000 23.620000 32.28000 19.080000 14.730000
## CV residual -8.552202 -5.339753  1.777625  8.76809 -2.158023 -3.996176
##              176      186      199      202
## Predicted    10.663269 2.234342 2.7478914 4.5226442
## cvpred      10.753336 2.109278 2.7259569 4.5451227
## Ozono       8.300000 4.650000 3.2100000 5.0500000
## CV residual -2.453336 2.540722 0.4840431 0.5048773
##
## Sum of squares = 291.53      Mean square = 18.22      n = 16
##
## fold 3
## Observations in test set: 16
##              7      18      27      28      49      54
## Predicted    10.145667 15.198364 10.1461421 13.99231 12.481257  9.194827
## cvpred      10.327833 15.224939 10.1021928 13.90347 12.774365  9.468645
## Ozono       4.730000 12.280000 10.6000000 12.77000  8.930000 12.050000
## CV residual -5.597833 -2.944939  0.4978072 -1.13347 -3.844365  2.581355
##              101      122      137      142      143      155
## Predicted    16.765867  9.577206 18.83929 18.314264 14.302186 19.669771
## cvpred      16.619596  9.838587 18.65523 18.205594 14.209082 19.436405
## Ozono       19.930000  4.260000 29.22000 12.810000  7.320000 28.150000
## CV residual  3.310404 -5.578587 10.56477 -5.395594 -6.889082  8.713595
##              166      167      168      184
## Predicted    11.388138  8.49343 10.957545  5.072194
## cvpred      11.787592  8.99322 11.169943  5.381118
## Ozono       5.620000  4.91000 14.180000  3.040000
## CV residual -6.167592 -4.08322  3.010057 -2.341118
##
## Sum of squares = 438.43      Mean square = 27.4      n = 16
##
## fold 4
## Observations in test set: 16
##              1      2      6      16      24      38
## Predicted    6.0272008  8.476699 10.114204  9.066450  6.395332  1.101268
## cvpred      6.2316117  8.955479 10.430674  9.301757  6.529657  1.142138
## Ozono       5.3400000  5.770000  6.390000  5.680000  4.080000  4.320000
## CV residual -0.8916117 -3.185479 -4.040674 -3.621757 -2.449657  3.177862
##              43      87      109      110      124      128
## Predicted    6.9002835 19.65340 18.918020 19.698579  8.527456 21.740586
## cvpred      7.0397051 19.66444 18.833878 19.557512  8.386945 21.779251
## Ozono       7.6300000 16.85000 21.870000 24.980000 12.160000 25.690000
## CV residual  0.5902949 -2.81444  3.036122  5.422488  3.773055  3.910749
##              136      152      189      193
## Predicted    19.678253 18.28949833 2.901051 5.30822110
## cvpred      19.502586 18.23452174 2.769524 5.16059425
## Ozono       17.060000 18.31000000 7.260000 5.23000000
## CV residual -2.442586  0.07547826 4.490476 0.06940575
##
## Sum of squares = 159.05      Mean square = 9.94      n = 16
##
## fold 5
## Observations in test set: 16

```

```

##          11          19          25          29          41          55          75
## Predicted    14.81354  8.7324036  7.158984  1.280414  2.5217442  9.998969  15.053715
## cvpred       15.34371  8.7472667  7.149033  1.173599  2.2954502  9.819741  15.154112
## Ozono        4.07000  9.2900000  8.320000  5.730000  3.0100000  12.330000  18.790000
## CV residual  -11.27371  0.5427333  1.170967  4.556401  0.7145498  2.510259  3.635888
##          77          78          97          99          106          111
## Predicted    12.619515  5.548119  18.924004  22.75617745  20.051111  22.789989
## cvpred       12.521553  5.900876  18.966341  22.87034274  20.152481  22.936382
## Ozono        11.300000  2.390000  14.310000  22.85000000  14.270000  28.240000
## CV residual  -1.221553 -3.510876 -4.656341 -0.02034274 -5.882481  5.303618
##          118          121          178
## Predicted    11.4187177  20.104689  13.955777
## cvpred       11.4109341  20.239842  14.409956
## Ozono        11.6000000  18.770000  9.090000
## CV residual   0.1890659 -1.469842 -5.319956
##
## Sum of squares = 298.29      Mean square = 18.64      n = 16
##
## fold 6
## Observations in test set: 16
##          3          31          34          36          39          56          60
## Predicted    1.920606  4.399085  12.32896 -2.278041  3.570676  8.7810354  11.467473
## cvpred       1.208680  4.100656  11.80690 -2.783057  3.180221  8.4552205  11.270727
## Ozono        3.690000  6.040000  22.89000  3.220000  7.190000  7.9300000  13.120000
## CV residual   2.481320  1.939344  11.08310  6.003057  4.009779 -0.5252205  1.849273
##          67          70          108          138          144          148
## Predicted    11.675348  6.153595  23.422133  19.427262  11.931665  9.422161
## cvpred       11.467761  6.015144  23.616471  19.696974  12.159854  9.678004
## Ozono        14.890000  7.260000  16.790000  18.330000  11.020000  5.140000
## CV residual   3.422239  1.244856 -6.826471 -1.366974 -1.139854 -4.538004
##          175          183          200
## Predicted    8.646439  4.150858  4.185553
## cvpred       8.846143  4.179757  4.416318
## Ozono        5.910000  3.010000  1.740000
## CV residual  -2.936143 -1.169757 -2.676318
##
## Sum of squares = 289.34      Mean square = 18.08      n = 16
##
## fold 7
## Observations in test set: 16
##          10          22          30          46          50          53
## Predicted    12.225524  2.7442540  5.657275  14.836293  15.111495  9.0952917
## cvpred       12.592491  2.9789799  5.788564  14.859092  15.379864  9.0000181
## Ozono        7.000000  2.7400000  4.040000  24.290000  10.180000  8.6000000
## CV residual  -5.592491 -0.2389799 -1.748564  9.430908 -5.199864 -0.4000181
##          71          105          115          119          154          157
## Predicted    13.735185  25.690079  20.694212  17.808745  17.51887  14.087396
## cvpred       13.811725  25.908061  20.707044  17.787173  17.61701  14.161069
## Ozono        9.690000  23.660000  26.100000  13.670000  7.00000  18.280000
## CV residual  -4.121725 -2.248061  5.392956 -4.117173 -10.61701  4.118931
##          162          169          188          197
## Predicted    9.727227  11.065375  6.402289  0.2796873
## cvpred       9.710282  11.074217  6.305709  0.2271546
## Ozono        7.200000  16.000000  4.310000  3.3300000

```

```

## CV residual -2.510282  4.925783 -1.995709 3.1028454
##
## Sum of squares = 392.47      Mean square = 24.53      n = 16
##
## fold 8
## Observations in test set: 16
##           8           35           66           79           82           88
## Predicted   8.053183  5.406914  4.296046 11.5765500 23.12714 17.94095
## cvpred      8.184036  5.577890  4.533929 11.5300011 22.71633 17.67660
## Ozono       4.350000  2.260000  2.880000 11.7900000 33.04000 19.16000
## CV residual -3.834036 -3.317890 -1.653929  0.2599989 10.32367  1.48340
##           90           98           104           113           126           135
## Predicted  12.517893 17.674731 27.961562 19.668473 20.074647 22.073606
## cvpred     12.446652 17.450143 27.364771 19.380569 19.774315 21.668364
## Ozono      8.730000 10.770000 34.390000 22.870000 22.290000 25.800000
## CV residual -3.716652 -6.680143  7.025229  3.489431  2.515685  4.131636
##           179          187          192          201
## Predicted  13.358434  4.316830  2.816824  1.6256739
## cvpred     13.279167  4.577333  3.102041  1.9485752
## Ozono      7.010000  3.290000  2.000000  2.6900000
## CV residual -6.269167 -1.287333 -1.102041  0.7414248
##
## Sum of squares = 323.38      Mean square = 20.21      n = 16
##
## fold 9
## Observations in test set: 16
##           23           58           93           117           130           153
## Predicted   5.821553 -0.31621478  7.117428 14.133350 25.42285 12.5794175
## cvpred      6.075590 -0.02684803  7.383264 14.015243 25.04355 12.2398938
## Ozono       2.920000  4.33000000  1.800000  9.350000 37.98000 12.3600000
## CV residual -3.155590  4.35684803 -5.583264 -4.665243 12.93645  0.1201062
##           156          161          163          165          172          173
## Predicted  18.620086 13.485664  4.284826  8.981806  7.175938  7.341584
## cvpred     18.505847 13.511314  4.205370  9.016919  7.248433  7.177546
## Ozono     21.840000 11.750000  2.610000  8.010000  5.330000  4.100000
## CV residual  3.334153 -1.761314 -1.595370 -1.006919 -1.918433 -3.077546
##           181          182          190          195
## Predicted   0.808824  3.7122724  2.237552  2.119022
## cvpred      0.778277  3.6065147  2.210584  2.068595
## Ozono       2.820000  3.1900000  4.980000  3.680000
## CV residual  2.041723 -0.4165147  2.769416  1.611405
##
## Sum of squares = 294.78      Mean square = 18.42      n = 16
##
## fold 10
## Observations in test set: 15
##           17           33           42           59           73           96
## Predicted  10.5005827 11.953390  2.5946826  2.692560  3.721097 21.479187
## cvpred     10.4251573 11.969931  2.1307498  2.307281  3.309010 21.708719
## Ozono     11.0600000 15.060000  1.9800000  9.320000  5.730000 26.890000
## CV residual  0.6348427  3.090069 -0.1507498  7.012719  2.420990  5.181281
##           103          107          131          132          140          141          159
## Predicted  20.602838 19.753000 25.06198 26.809476 8.154963 18.170346 15.141476
## cvpred     21.005845 19.951867 25.47833 27.347923 7.824355 18.409577 15.299591

```

```

## Ozono      17.950000 13.300000 23.07000 19.200000 8.860000 22.860000 13.890000
## CV residual -3.055845 -6.651867 -2.40833 -8.147923 1.035645 4.450423 -1.409591
##           191      194
## Predicted   2.064364 2.6622221
## cvpred      1.956526 2.6030435
## Ozono       3.230000 2.9600000
## CV residual 1.273474 0.3569565
##
## Sum of squares = 242.25      Mean square = 16.15      n = 15
##
## fold 11
## Observations in test set: 15
##           9      45      63      64      74      76
## Predicted   11.011052 10.996926 1.853942 -0.15812249 3.587772 18.203266
## cvpred      11.344612 11.149834 1.766860 -0.02451469 3.668608 18.356396
## Ozono       3.940000 8.700000 4.810000 3.65000000 8.680000 21.120000
## CV residual -7.404612 -2.449834 3.043140 3.67451469 5.011392 2.763604
##           86      89      91      100      150      151      164
## Predicted   11.402628 14.980089 12.21948 19.015507 17.159701 19.54534 5.552941
## cvpred      11.278811 14.979919 12.21746 19.037446 16.823109 19.22993 5.523302
## Ozono       4.820000 16.150000 6.68000 15.270000 26.000000 29.79000 7.370000
## CV residual -6.458811 1.170081 -5.53746 -3.767446 9.176891 10.56007 1.846698
##           174      185
## Predicted   7.508760 6.033263
## cvpred      7.343206 5.833228
## Ozono       10.990000 2.950000
## CV residual 3.646794 -2.883228
##
## Sum of squares = 425.04      Mean square = 28.34      n = 15
##
## fold 12
## Observations in test set: 15
##           15      48      61      72      84      85      95
## Predicted   9.768927 2.100862 3.369735 6.640398 21.950700 10.721300 21.149123
## cvpred      9.627881 1.721019 3.292343 6.556732 21.787628 10.667742 21.002762
## Ozono       6.150000 8.100000 5.090000 12.230000 31.150000 8.680000 25.660000
## CV residual -3.477881 6.378981 1.797657 5.673268 9.362372 -1.987742 4.657238
##           102      116      120      139      147      180
## Predicted   18.124340 21.6153762 19.800978 9.403464 10.689221 2.174727
## cvpred      18.222198 21.5626454 19.647804 9.104369 10.846232 2.130822
## Ozono       15.250000 21.9200000 23.790000 3.350000 4.220000 4.200000
## CV residual -2.972198 0.3573546 4.142196 -5.754369 -6.626232 2.069178
##           196      198
## Predicted   4.033710 1.887455
## cvpred      4.015462 2.011241
## Ozono       5.710000 4.250000
## CV residual 1.694538 2.238759
##
## Sum of squares = 316.8      Mean square = 21.12      n = 15
##
## fold 13
## Observations in test set: 15
##           4      5      12      21      44      47
## Predicted   6.892439 8.978100 7.011831 2.1112054 12.489117 12.1398415

```



```
## cvpred      7.028784  9.115879  7.121245  2.0296193 12.559004 12.1678026
## Ozono       3.890000  5.760000  4.390000  2.9400000 15.680000 12.6700000
## CV residual -3.138784 -3.355879 -2.731245  0.9103807  3.120996  0.5021974
##           57      65      81      92     125     145
## Predicted   11.691632  3.499894 19.61933  5.1938660 18.305004 12.9865216
## cvpred     11.664949  3.377166 19.63630  5.1447008 18.200942 12.9734517
## Ozono       9.090000  6.760000 26.89000  5.2700000 14.880000 12.2500000
## CV residual -2.574949  3.382834  7.25370  0.1252992 -3.320942 -0.7234517
##           158     170     171
## Predicted   13.171724  3.432383  2.3829847
## cvpred     13.120489  3.454883  2.2531721
## Ozono      10.110000  4.820000  2.9000000
## CV residual -3.010489  1.365117  0.6468279
##
## Sum of squares = 133    Mean square = 8.87    n = 15
##
## Overall (Sum over all 15 folds)
##      ms
## 19.45499
```

Se calcula la raíz cuadrada de la media de los cuadrados de las diferencias entre predicciones y observaciones:

```
errores <- cv_k3_MS$cvpred - cv_k3_MS$Ozono # predicho por cv - predicción real
( error_cv_k3_MS <- sqrt(mean(errores^2)) ) # estimador RMSE (raíz media suma residuos al cuadrado)
```

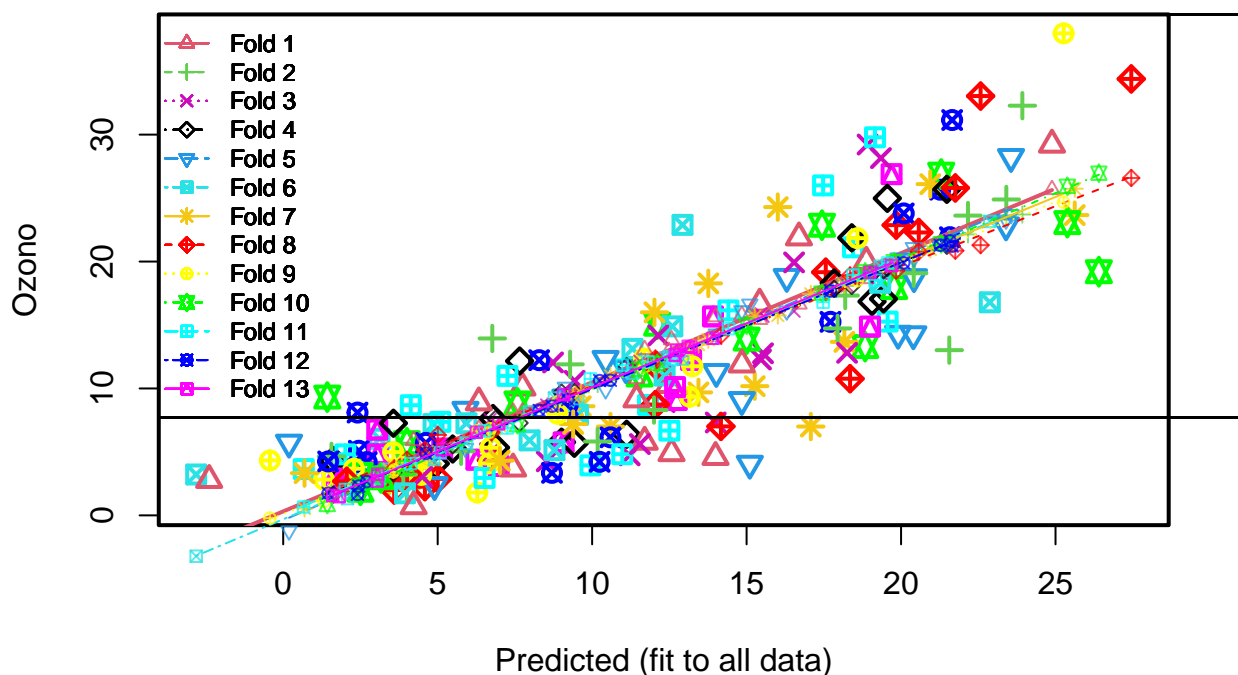
```
## [1] 4.410781
```

Finalmente, para MC:

```
set.seed(5198)
cv_k3_MC <- cv.lm(data=OzonoLA,form.lm=formula(MC),m=length(OzonoLA))
```

```
## Warning in cv.lm(data = OzonoLA, form.lm = formula(MC), m = length(OzonoLA)):
##
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```

## Small symbols show cross-validation predicted values



```
##
## fold 1
## Observations in test set: 15
##          13      14      26      32      37      52
## Predicted   7.440524 12.570232 11.700192 12.481891 -2.396811 6.332799
## cvpred      8.153843 13.829313 13.062716 12.921002 -3.355326 6.788498
## Ozono       3.690000 4.900000 5.800000 10.270000 2.790000 8.900000
## CV residual -4.463843 -8.929313 -7.262716 -2.651002 6.145326 2.111502
##          62      68      80      112      146      149
## Predicted   7.744519 16.695646 18.876144 24.885585 13.99472 15.42288
## cvpred      8.150007 16.643705 18.650796 25.670862 15.64258 15.43417
## Ozono      10.070000 21.900000 19.980000 29.210000 4.60000 16.68000
## CV residual  1.919993  5.256295  1.329204  3.539138 -11.04258  1.24583
##          160      177      203
## Predicted  11.436518 14.844856  4.222157
## cvpred     10.779545 15.770028  5.884059
## Ozono       9.140000 11.890000  0.720000
## CV residual -1.639545 -3.880028 -5.164059
##
## Sum of squares = 415.17    Mean square = 27.68    n = 15
##
## fold 2
## Observations in test set: 16
##          20      40      51      69      83      94
## Predicted   4.001573 3.698056  6.767830 18.1923441 23.406972  9.285851
## cvpred      3.938462 3.759339  6.400434 18.0794809 23.527989  9.271308
## Ozono       2.180000 5.650000 13.940000 17.3200000 24.890000 11.900000
## CV residual -1.758462 1.890661  7.539566 -0.7594809  1.362011  2.628692
##          114      123      127      129      133      134      176
```

```

## Predicted    21.56123 10.178986 22.16576 23.926548 20.410373 17.945077 12.002891
## cvpred      21.82530 10.353186 22.14034 23.707092 20.662457 18.076038 11.995393
## Ozono       13.02000  5.820000 23.62000 32.280000 19.080000 14.730000  8.300000
## CV residual -8.80530 -4.533186  1.47966  8.572908 -1.582457 -3.346038 -3.695393
##              186          199          202
## Predicted    1.562951  3.8866690  5.7592135
## cvpred      1.740678  3.8349488  5.6766484
## Ozono       4.650000  3.2100000  5.0500000
## CV residual  2.909322 -0.6249488 -0.6266484
##
## Sum of squares = 283.22      Mean square = 17.7      n = 16
##
## fold 3
## Observations in test set: 16
##              7          18          27          28          49          54
## Predicted    11.215916 15.465775  9.436245 15.531555 12.347251  8.734629
## cvpred      11.777745 15.887994  9.404568 15.689903 12.176068  8.761862
## Ozono       4.730000 12.280000 10.600000 12.770000  8.930000 12.050000
## CV residual -7.047745 -3.607994  1.195432 -2.919903 -3.246068  3.288138
##              101         122         137         142         143         155
## Predicted    16.536223  8.529697 18.91607 18.252788 14.000615 19.347816
## cvpred      16.088214  8.695824 18.72904 18.071583 13.947722 19.063656
## Ozono       19.930000  4.260000 29.22000 12.810000  7.320000 28.150000
## CV residual  3.841786 -4.435824 10.49096 -5.261583 -6.627722  9.086344
##              166         167         168         184
## Predicted    11.546461  6.694141 12.142312  4.504243
## cvpred      12.145192  7.550493 12.330433  5.106427
## Ozono       5.620000  4.910000 14.180000  3.040000
## CV residual -6.525192 -2.640493  1.849567 -2.066427
##
## Sum of squares = 449.9      Mean square = 28.12      n = 16
##
## fold 4
## Observations in test set: 16
##              1          2          6          16          24          38
## Predicted    6.873243  9.009641 11.117544  9.423672  4.9232014 1.622697
## cvpred      7.604531  9.681745 11.756068  9.702572  4.9079724 1.636612
## Ozono       5.340000  5.770000  6.390000  5.680000  4.0800000 4.320000
## CV residual -2.264531 -3.911745 -5.366068 -4.022572 -0.8279724 2.683388
##              43          87          109          110          124          128
## Predicted    6.7912501 19.060610 18.405295 19.555138  7.652500 21.48177
## cvpred      6.9089036 19.039561 18.075293 19.495939  7.281611 21.23630
## Ozono       7.6300000 16.850000 21.870000 24.980000 12.160000 25.69000
## CV residual  0.7210964 -2.189561  3.794707  5.484061  4.878389  4.45370
##              136         152         189         193
## Predicted    19.431825 17.8343745 3.567057  5.4861801
## cvpred      19.309081 17.8576265 3.347544  5.5191732
## Ozono       17.060000 18.3100000 7.260000  5.2300000
## CV residual -2.249081  0.4523735 3.912456 -0.2891732
##
## Sum of squares = 187.37      Mean square = 11.71      n = 16
##
## fold 5
## Observations in test set: 16

```

```

##          11          19          25          29          41          55
## Predicted  15.10421  9.1336905  5.899857  0.1895359  2.2981852  10.436443
## cvpred    16.64953  10.1034438  5.288593 -1.2187132  2.3202081  9.903537
## Ozono      4.07000  9.2900000  8.320000  5.7300000  3.0100000  12.330000
## CV residual -12.57953 -0.8134438  3.031407  6.9487132  0.6897919  2.426463
##          75          77          78          97          99          106
## Predicted  16.299885  14.018014  4.898545  19.904221  23.3981119  20.397884
## cvpred    16.114479  14.296348  4.779801  20.286428  23.2796716  21.108762
## Ozono      18.790000  11.300000  2.390000  14.310000  22.8500000  14.270000
## CV residual  2.675521 -2.996348 -2.389801 -5.976428 -0.4296716 -6.838762
##          111          118          121          178
## Predicted  23.55944  11.0902103  20.417815  14.845379
## cvpred    23.70526  11.1936349  20.276326  16.129878
## Ozono      28.24000  11.6000000  18.770000  9.090000
## CV residual  4.53474  0.4063651 -1.506326 -7.039878
##
## Sum of squares = 399.82      Mean square = 24.99      n = 16
##
## fold 6
## Observations in test set: 16
##          3          31          34          36          39          56          60
## Predicted  2.081988  3.117250  12.92545 -2.817305  4.866487  9.541240  11.305685
## cvpred    1.321883  2.932481  12.32884 -3.220900  4.400694  9.199472  11.214526
## Ozono      3.690000  6.040000  22.89000  3.220000  7.190000  7.930000  13.120000
## CV residual  2.368117  3.107519  10.56116  6.440900  2.789306 -1.269472  1.905474
##          67          70          108          138          144          148
## Predicted  12.596424  5.930993  22.872328  19.325532  12.344723  8.779764
## cvpred    12.280983  5.686535  23.163803  19.632142  12.384675  9.249603
## Ozono      14.890000  7.260000  16.790000  18.330000  11.020000  5.140000
## CV residual  2.609017  1.573465 -6.373803 -1.302142 -1.364675 -4.109603
##          175          183          200
## Predicted  7.975095  3.3944500  3.940621
## cvpred    8.159317  3.6304783  4.016822
## Ozono      5.910000  3.0100000  1.740000
## CV residual -2.249317 -0.6204783 -2.276822
##
## Sum of squares = 262.29      Mean square = 16.39      n = 16
##
## fold 7
## Observations in test set: 16
##          10          22          30          46          50          53
## Predicted  10.594280  3.4312358  4.6632159  16.009568  15.264726  9.633851
## cvpred    11.002061  3.6307681  5.0326898  15.751721  15.678303  9.694815
## Ozono      7.000000  2.7400000  4.0400000  24.290000  10.180000  8.600000
## CV residual -4.002061 -0.8907681 -0.9926898  8.538279 -5.498303 -1.094815
##          71          105          115          119          154          157
## Predicted  13.442198  25.622650  20.944487  18.16524  17.07908  13.761385
## cvpred    13.448517  25.739853  20.956935  18.16042  17.45495  13.841444
## Ozono      9.690000  23.660000  26.100000  13.67000  7.00000  18.280000
## CV residual -3.758517 -2.079853  5.143065 -4.49042 -10.45495  4.438556
##          162          169          188          197
## Predicted  9.386545  12.017325  7.013945  0.6901916
## cvpred    9.377751  11.795743  7.041877  0.5420600
## Ozono      7.200000  16.000000  4.310000  3.3300000

```

```

## CV residual -2.177751  4.204257 -2.731877  2.7879400
##
## Sum of squares = 353.86      Mean square = 22.12      n = 16
##
## fold 8
## Observations in test set: 16
##           8           35           66           79           82           88
## Predicted   6.739690  4.586791  5.004902 12.0791409 22.57508 17.562176
## cvpred      6.545341  5.159333  6.333695 12.1911483 21.29180 17.567112
## Ozono       4.350000  2.260000  2.880000 11.7900000 33.04000 19.160000
## CV residual -2.195341 -2.899333 -3.453695 -0.4011483 11.74820  1.592888
##           90           98           104           113           126           135
## Predicted  12.047731 18.349383 27.452640 19.842784 20.573563 21.757166
## cvpred     12.360251 18.910913 26.584818 19.288651 20.780328 20.855775
## Ozono      8.730000 10.770000 34.390000 22.870000 22.290000 25.800000
## CV residual -3.630251 -8.140913  7.805182  3.581349  1.509672  4.944225
##           179           187           192           201
## Predicted  14.169995  4.008337  3.591144  2.0571540
## cvpred     14.187255  4.382413  3.705436  2.4157289
## Ozono      7.010000  3.290000  2.000000  2.6900000
## CV residual -7.177255 -1.092413 -1.705436  0.2742711
##
## Sum of squares = 401.49      Mean square = 25.09      n = 16
##
## fold 9
## Observations in test set: 16
##           23           58           93           117           130           153
## Predicted   3.839587 -0.4308846  6.281735 13.155307 25.26400 11.6791979
## cvpred      4.154127 -0.2118862  6.664557 13.108784 24.66399 11.6941871
## Ozono       2.920000  4.3300000  1.800000  9.350000 37.98000 12.3600000
## CV residual -1.234127  4.5418862 -4.864557 -3.758784 13.31601  0.6658129
##           156           161           163           165           172           173
## Predicted  18.588454 13.244652  3.6072028  8.8837739  6.710209  6.661469
## cvpred     18.494933 13.285754  3.5938733  8.9481276  6.831756  6.527415
## Ozono     21.840000 11.750000  2.6100000  8.0100000  5.330000  4.100000
## CV residual  3.345067 -1.535754 -0.9838733 -0.9381276 -1.501756 -2.427415
##           181           182           190           195
## Predicted   1.318443  4.4697715  3.564775  2.291775
## cvpred      1.089418  4.1187978  3.420703  2.321299
## Ozono       2.820000  3.1900000  4.980000  3.680000
## CV residual  1.730582 -0.9287978  1.559297  1.358701
##
## Sum of squares = 269.38      Mean square = 16.84      n = 16
##
## fold 10
## Observations in test set: 15
##           17           33           42           59           73           96
## Predicted  11.5364232 12.10590  2.5007446  1.4263141  4.000462 21.299458
## cvpred     11.5187686 11.97613  2.4263812  0.8591298  3.754198 21.398705
## Ozono     11.0600000 15.06000  1.9800000  9.3200000  5.730000 26.890000
## CV residual -0.4587686  3.08387 -0.4463812  8.4608702  1.975802  5.491295
##           103           107           131           132           140           141
## Predicted  19.776046 18.837340 25.382205 26.402518  7.583810 17.439592
## cvpred     19.989241 19.190002 25.927378 26.961438  7.143562 17.509398

```

```

## Ozono      17.950000 13.300000 23.070000 19.200000 8.860000 22.860000
## CV residual -2.039241 -5.890002 -2.857378 -7.761438 1.716438 5.350602
##           159      191      194
## Predicted   15.002898 2.4738283 3.1660519
## cvpred      14.983142 2.2911007 3.1850573
## Ozono       13.890000 3.2300000 2.9600000
## CV residual -1.093142 0.9388993 -0.2250573
##
## Sum of squares = 256.52      Mean square = 17.1      n = 15
##
## fold 11
## Observations in test set: 15
##           9      45      63      64      74      76      86
## Predicted   9.943837 11.809295 2.011643 0.6636727 4.129266 18.445270 10.975906
## cvpred      10.751684 12.098156 2.038901 0.6654128 4.087515 18.933147 11.840547
## Ozono       3.940000 8.700000 4.810000 3.6500000 8.680000 21.120000 4.820000
## CV residual -6.811684 -3.398156 2.771099 2.9845872 4.592485 2.186853 -7.020547
##           89      91      100      150      151      164      174
## Predicted   14.411125 12.501165 19.565430 17.477059 19.15705 5.084568 7.258785
## cvpred      14.713225 12.524334 19.890372 16.797102 18.53135 4.954099 6.889160
## Ozono       16.150000 6.680000 15.270000 26.000000 29.79000 7.370000 10.990000
## CV residual 1.436775 -5.844334 -4.620372 9.202898 11.25865 2.415901 4.100840
##           185
## Predicted   6.518272
## cvpred      6.655734
## Ozono       2.950000
## CV residual -3.705734
##
## Sum of squares = 455.1      Mean square = 30.34      n = 15
##
## fold 12
## Observations in test set: 15
##           15      48      61      72      84      85      95
## Predicted   10.592034 2.405465 2.447951 8.284350 21.654136 9.2063365 21.26473
## cvpred      10.643806 1.652116 2.309982 8.272994 21.208759 9.1158536 21.28625
## Ozono       6.150000 8.100000 5.090000 12.230000 31.150000 8.6800000 25.66000
## CV residual -4.493806 6.447884 2.780018 3.957006 9.941241 -0.4358536 4.37375
##           102      116      120      139      147      180
## Predicted   17.71477 21.5816040 20.086598 8.700437 10.243076 2.703505
## cvpred      17.67977 21.6362454 19.901511 8.603268 10.618833 2.661679
## Ozono       15.25000 21.9200000 23.790000 3.350000 4.220000 4.200000
## CV residual -2.42977 0.2837546 3.888489 -5.253268 -6.398833 1.538321
##           196      198
## Predicted   4.6184035 1.454674
## cvpred      4.8832021 1.743392
## Ozono       5.7100000 4.250000
## CV residual 0.8267979 2.506608
##
## Sum of squares = 302.28      Mean square = 20.15      n = 15
##
## fold 13
## Observations in test set: 15
##           4      5      12      21      44      47
## Predicted   6.976510 9.076506 6.253474 1.604130 13.912924 13.1096748

```

```
## cvpred      7.228818  9.521362  6.557988  1.572653  13.956235  13.3257148
## Ozono       3.890000  5.760000  4.390000  2.940000  15.680000  12.6700000
## CV residual -3.338818 -3.761362 -2.167988  1.367347  1.723765 -0.6557148
##           57      65      81      92     125     145
## Predicted  12.736124  3.044066  19.693546  5.14694398  18.994232  13.2040056
## cvpred     13.006213  2.718416  19.683888  5.18736734  19.161706  13.2412375
## Ozono       9.090000  6.760000  26.890000  5.27000000  14.880000  12.2500000
## CV residual -3.916213  4.041584  7.206112  0.08263266 -4.281706 -0.9912375
##           158     170     171
## Predicted  12.685143  3.060696  1.816891
## cvpred     12.627759  3.251042  1.485275
## Ozono      10.110000  4.820000  2.900000
## CV residual -2.517759  1.568958  1.414725
##
## Sum of squares = 148.99      Mean square = 9.93      n = 15
##
## Overall (Sum over all 15 folds)
##      ms
## 20.61771
```

```
errores <- cv_k3_MC$cvpred - cv_k3_MC$Ozono
( error_cv_k3_MC <- sqrt(mean(errores^2)) )
```

```
## [1] 4.540672
```

Obtenemos un comportamiento mejor con el MS que con MC, pues tenemos un menor error.

## 8. Análisis de residuos modelo seleccionado

## 9. Análisis de influencia modelo seleccionado

## 10. Estimación media condicionada y predicción

Finalmente, obtengamos el intervalo de confianza y de predicción para el nivel de ozono medio al 95% de confianza con el modelo seleccionado con todas las variables fijadas en su valor medio.

```
new.dat <- data.frame(T_Sandburg = mean(T_Sandburg), Humedad = mean(Humedad),
                     T_ElMonte = mean(T_ElMonte), Mes = mean(Mes),
                     Pres_Alt = mean(Pres_Alt), Inv_Alt_b = mean(Inv_Alt_b)) # tiene que aparecer valor
predict(ajuste, newdata = new.dat, interval="confidence", level = 0.95)
```

```
##      fit      lwr      upr
## 1 11.37399 10.76961 11.97837
```

```
predict(ajuste, newdata = new.dat, interval="prediction", level = 0.95)
```

```
##      fit      lwr      upr
## 1 11.37399  2.741654 20.00633
```

# Regresión Logística

- Antes de empezar, cargamos los datos *Oro.rda*

```
load("Datos/Oro.rda")
```

## 1. Análisis descriptivo

Para el análisis descriptivo de las variables podemos comenzar con una visión general de las variables mediante las funciones `str()` y `summary()`.

```
str(Oro)
```

```
## 'data.frame':    64 obs. of  4 variables:
## $ As          : num  6.77 15.03 6.43 0.1 0.1 ...
## $ Sb          : num  3.08 6.15 2.35 0.3 0.3 9.62 0.51 3.71 4.32 0.8 ...
## $ Corredor    : int   1 1 1 0 0 1 0 1 0 0 ...
## $ Proximidad  : int   1 1 1 0 0 1 0 1 0 0 ...
```

La salida de `str()` nos dice que los datos constan de 64 observaciones de 4 variables:

- **As**: Nivel de concentración de arsénico en la muestra de agua. (numérica)
- **Sb**: Nivel de concentración de antimonio en la muestra de agua. (numérica)
- **Corredor**: Variable binaria indicando si la zona muestreada está (1) o no está (0) en alguno de los corredores delimitados por las líneas sobre el mapa. (categórica)
- **Proximidad**: Variable de respuesta que toma los valores 1 o 0 según que el depósito esté próximo o esté muy lejano al lugar.

```
attach(Oro)
```

```
Oro$Corredor <- as.factor(Oro$Corredor) # Convertimos la variable Corredor a factor
numericas.oro <- Oro[1:2]               # Almacenamos las variables numéricas
respuesta.oro <- Proximidad              # Almacenamos la variable de respuesta
```

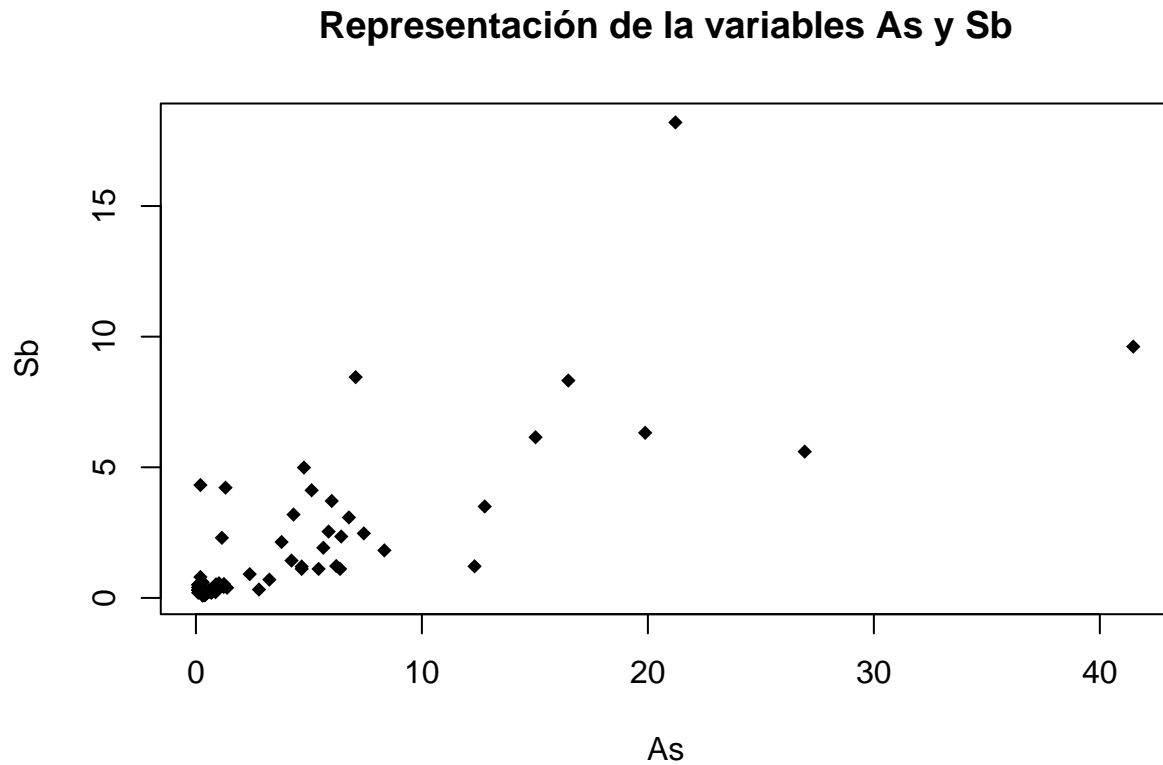
Con la salida de `summary()` y graficando **As** frente a **Sb** podemos ver que, basándonos en la diferencia entre las medias y las medianas, las variables numéricas se concentran en valores bajos, aunque deben de existir registros con valores relativamente altos:

```
summary(Oro)
```

##	As	Sb	Corredor	Proximidad
## Min.	: 0.100	Min. : 0.100	0:32	Min. :0.0000
## 1st Qu.:	0.400	1st Qu.: 0.300	1:32	1st Qu.:0.0000
## Median :	1.235	Median : 0.650		Median :0.0000
## Mean :	4.645	Mean : 2.039		Mean :0.4375
## 3rd Qu.:	5.905	3rd Qu.: 2.487		3rd Qu.:1.0000
## Max.	:41.480	Max. :18.200		Max. :1.0000



```
plot(numericas.oro, pch=18,
     main="Representación de la variables As y Sb")
```

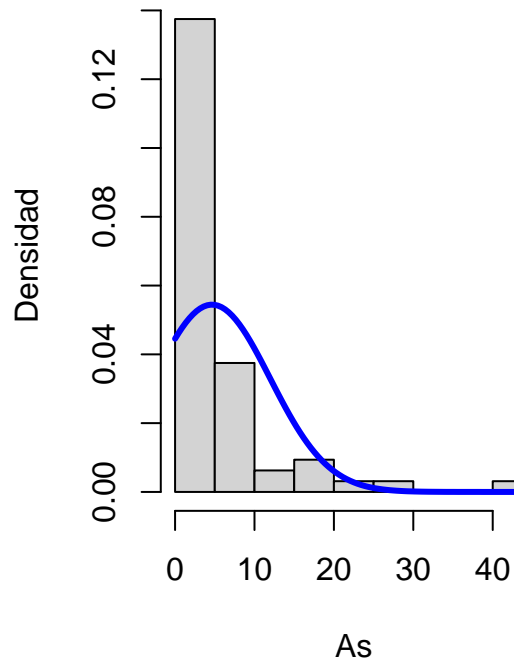


Este hecho se confirma también al mirar los histogramas y diagramas de cajas:

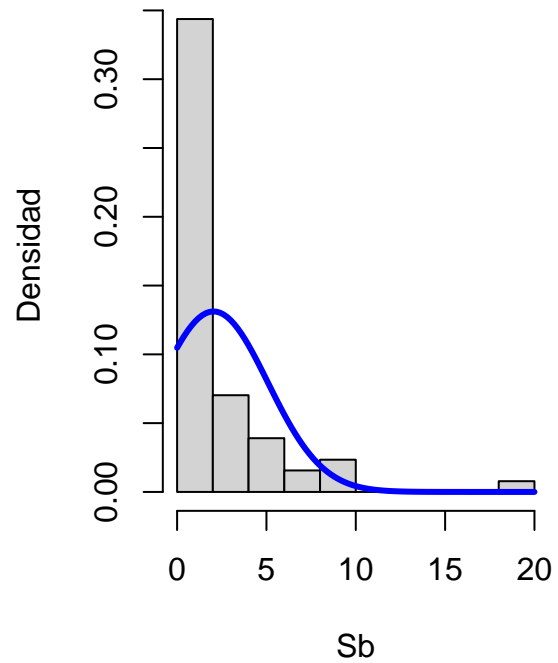
```
old.par <- par(mfrow=c(1,2))
hist(As, freq=F, xlab="As", ylab = "Densidad",
     main="Concentración de Arsénico")
curve(dnorm(x,mean=mean(As), sd=sd(As)),
      col="blue", lwd=3, add=TRUE)

hist(Sb, freq=F, xlab="Sb", ylab = "Densidad",
     main="Concentración de Antimonio")
curve(dnorm(x,mean=mean(Sb), sd=sd(Sb)),
      col="blue", lwd=3, add=TRUE)
```

### Concentración de Arsénico



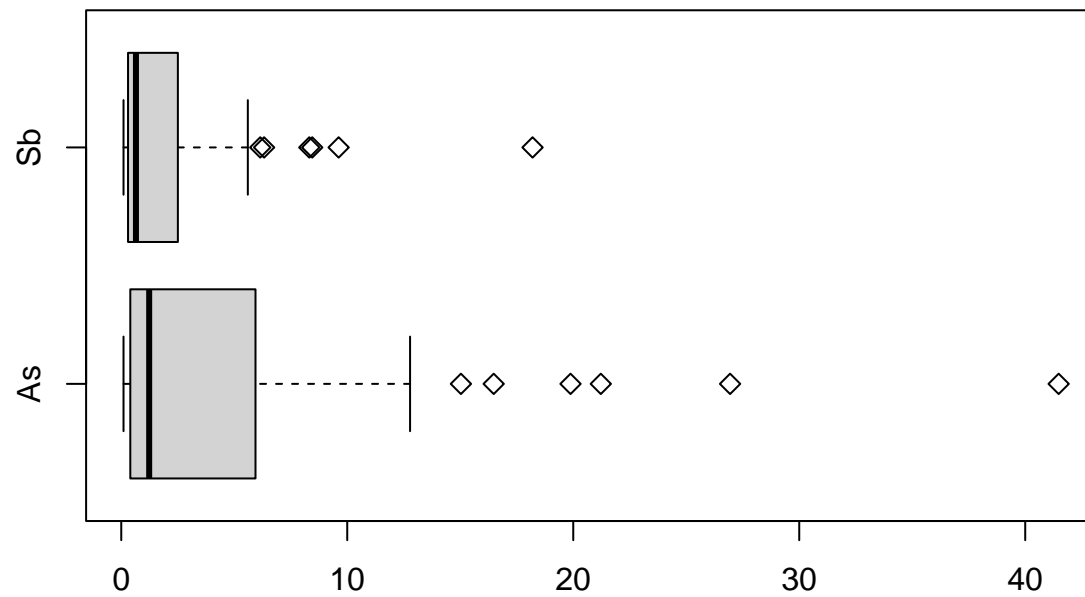
### Concentración de Antimonio



```
par(old.par)

boxplot(numericas.oro, horizontal=T, pch=5,
        main="Diagrama de cajas de las variables numéricas")
```

### Diagrama de cajas de las variables numéricas



Distribución de la variable Proximidad:

```
table(Proximidad); table(Proximidad)/nrow(Oro)
```

```
## Proximidad
##  0  1
## 36 28

## Proximidad
##      0      1
## 0.5625 0.4375
```

Distribución de la variable Corredor:

```
table(Corredor)
```

```
## Corredor
##  0  1
## 32 32
```

Observamos que si los datos se encuentran en alguno de los corredores, suelen estar próximos a un depósito de oro y lejanos si no es así:

```
xtabs(~Proximidad + Corredor, data=Oro)
```

```
##           Corredor
## Proximidad  0  1
##           0 30  6
##           1  2 26
```

## 2. Modelo matemático

Dado que contamos con una muestra de  $n$  realizaciones  $(\vec{X}^t, Y)$  con  $\vec{X}^t = (X_1, \dots, X_k)$  que asumimos independientes, y que la variable respuesta, **Proximidad**, es binaria (0 o 1), debemos de elegir un modelo que tenga esto en cuenta. En nuestro caso hemos elegido una transformación del modelo lineal, definida por la distribución logística de la ecuación 2.

$$F(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \quad (2)$$

Por tanto, nuestro modelo logístico quedaría de la forma

$$Y | (\vec{X} = \vec{X}_i) \sim Be(p_i), \quad p_i = \mathbb{P}(Y = 1 | \vec{X}_i) = \frac{1}{1 + e^{-\eta}} \quad (3)$$

Tal que

$$\eta = \beta_0 + \beta_1 A s + \beta_2 S b + \tau I(\text{Corredor} = 1) \quad (4)$$

siendo  $I(\text{Corredor} = 1)$  la variable indicadora para cuando Corredor toma el valor 1. Además,

$$1 - p_i = \mathbb{P}(Y = 0 | \vec{X}_i) = 1 - \frac{1}{1 + e^{-\eta}} = \frac{e^{-\eta}}{1 + e^{-\eta}} \quad (5)$$

### 3. Interpretación del modelo

Para una mejor interpretación del modelo, podemos definir el **odds**<sub>*i*</sub> de manera que

$$odds_i = odds(Y|\vec{X}_i) = \frac{p_i}{1-p_i} = e^\eta = e^{\vec{\beta}^t \vec{X}_i} = e^{\beta_0} e^{\beta_1 X_{i1}} \dots e^{\beta_k X_{ik}}, \quad 1 \leq i \leq n \quad (6)$$

Este es un modelo multiplicativo, en el cual  $e^{\beta_0}$  es la respuesta cuando  $\vec{X}_i = \vec{0}$ , mientras que  $e^{\beta_j}$ , para  $1 \leq j \leq k$ , es el incremento multiplicativo  $(e^{\beta_j})^l$  en el odds para algún incremento  $l$  en  $X_j$

Si resulta que existe una variable binaria podemos utilizar el **odds-ratio**, que indica en qué medida el suceso  $Y = 1$  es más posible que  $Y = 0$  si  $X = 1$  que si  $X = 0$ :

$$OR = \frac{\mathbb{P}(Y = 1|X = 1)/\mathbb{P}(Y = 0|X = 1)}{\mathbb{P}(Y = 1|X = 0)/\mathbb{P}(Y = 0|X = 0)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} \quad (7)$$

Si  $X$  es cualitativa podemos seguir aplicando el  $OR$  con  $g - 1$  variables *dummy*, siendo  $g$  el número de categorías.

También podemos expresar el modelo aplicando logaritmos a la ecuación 6, de manera que

$$\ln\left(\frac{p_i}{1-p_i}\right) = \eta = \vec{\beta}^t \vec{X}_i \quad (8)$$

Los cuales denominaremos como **logit**<sub>*i*</sub>. Estos logits son interpretables mucho más fácilmente ya que son interpretables linealmente.

Finalmente, por lo comentado en el apartado del modelo matemático y en este, este modelo sigue las tres siguientes hipótesis estructurales:

1. Linealidad de los logits.
2. Respuesta binaria de la  $Y$ .
3. Independencia de las observaciones.

### 4. Inferencia

```
ajuste <- glm(Proximidad~., data=Oro, family="binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
summary(ajuste)

##
## Call:
## glm(formula = Proximidad ~ ., family = "binomial", data = Oro)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28138  -0.06006  -0.04071   0.02446   2.32651
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.6096     3.1661  -2.403   0.0162 *
## As            1.2046     0.4899   2.459   0.0139 *
## Sb            1.4210     0.7301   1.946   0.0516 .
## Corredor1     3.1973     1.8911   1.691   0.0909 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 87.720  on 63  degrees of freedom
## Residual deviance: 14.194  on 60  degrees of freedom
## AIC: 22.194
##
## Number of Fisher Scoring iterations: 9
```

Teniendo en cuenta la ecuación 8, los coeficientes ajustados y las variables significativas, el modelo quedaría como en la ecuación 9

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\eta} = A + BAs + CSb + DI(Corredor = 1) \quad (9)$$

## 5. Bondad del ajuste