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()`.

```
## '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 presion 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)

```
##       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` :

```
##    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:

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

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

Evaluamos la asimetría y kurtoisis

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

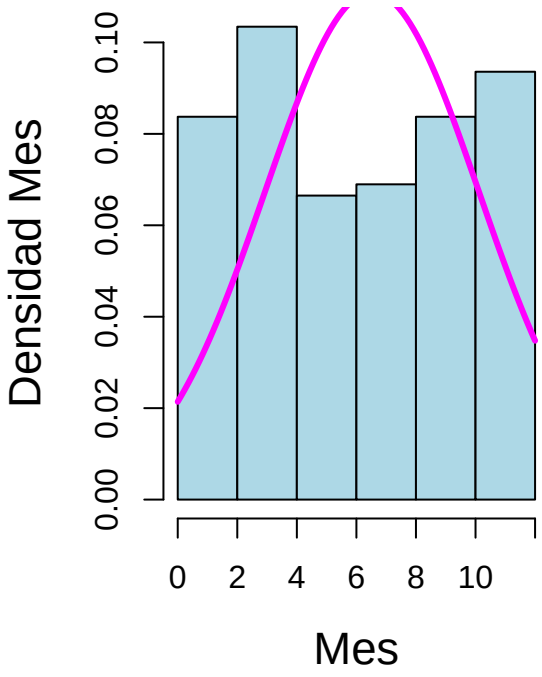
```
## [1] 1.671129
```

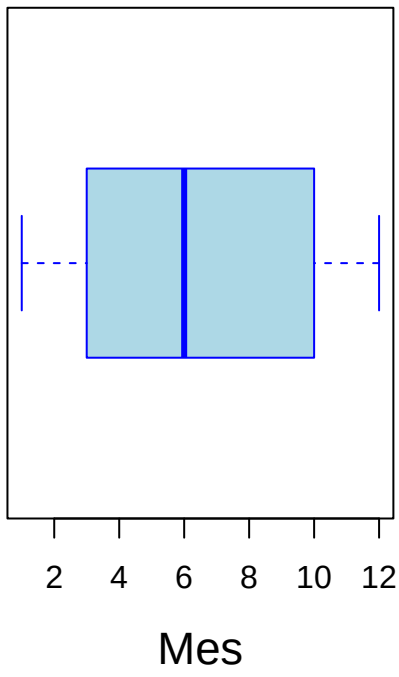
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

Vemos si hay registros atípicos

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

Como podemos ver no existe ningún registro atípico





Análisis descriptivo de la variable `DiaMes` :

```
##    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:

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

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

Evaluamos la asimetría y kurtoisis

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

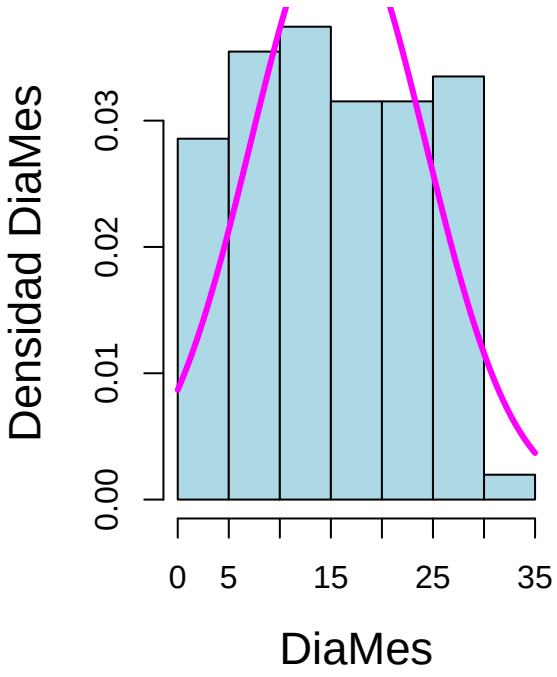
```
## [1] 1.868548
```

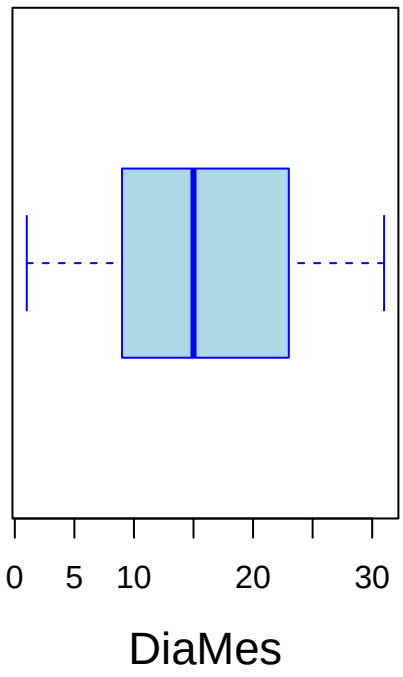
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

Vemos si hay registros atípicos

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

Como podemos ver no existe ningún registro atípico





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

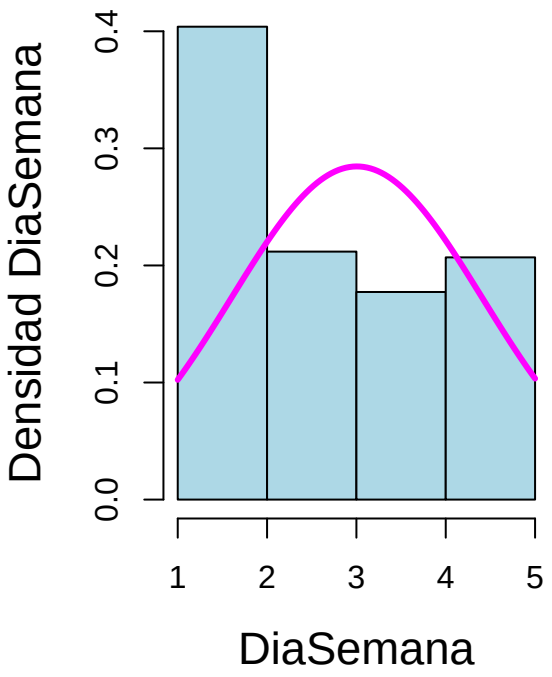
```
## [1] 1.731687
```

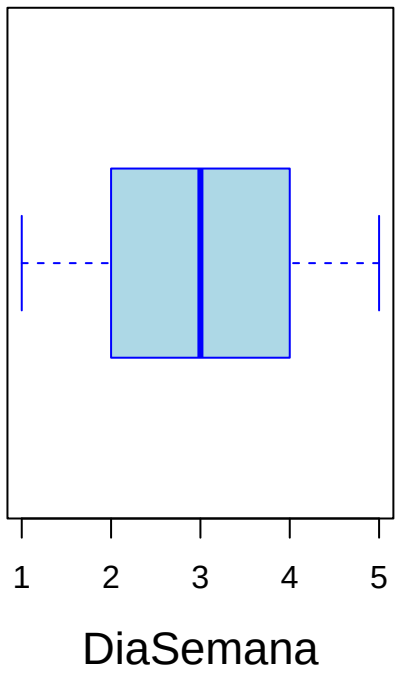
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis menor que tres, las colas de la variable comparadas con una normal son más ligeras.

Vemos si hay registros atípicos

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

Como podemos ver no existe ningún registro atípico





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

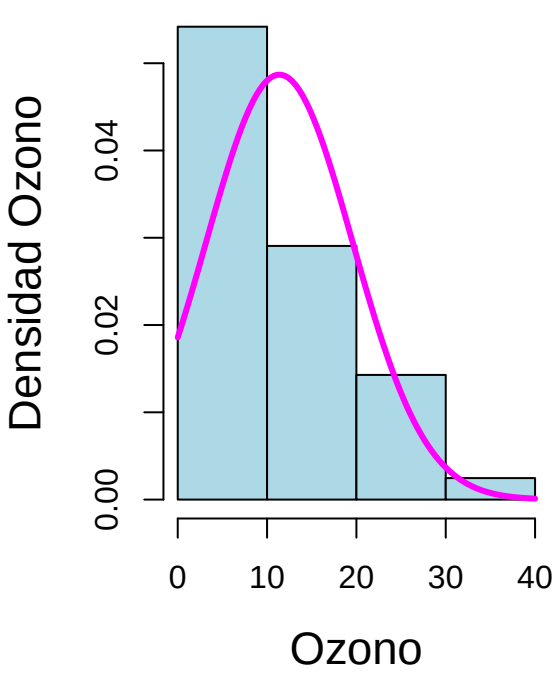
```
## [1] 3.089498
```

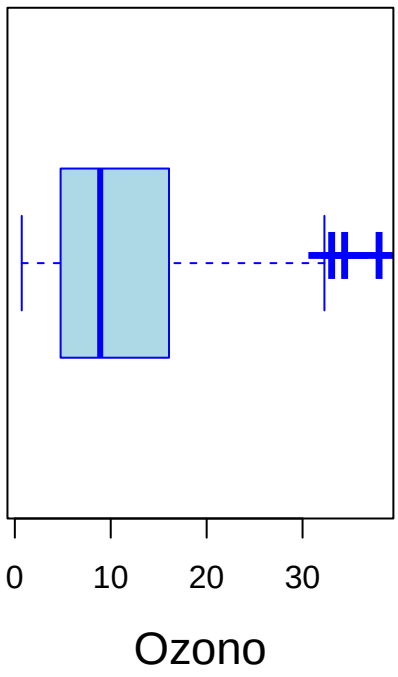
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal

Vemos si hay registros atípicos

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

Como podemos ver existen 4 registros atípicos





Análisis descriptivo de la variable `Pres_Alt` :

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

Desviación típica y rango intercuartílico:

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

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

Evaluamos la asimetría y kurtoisis

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

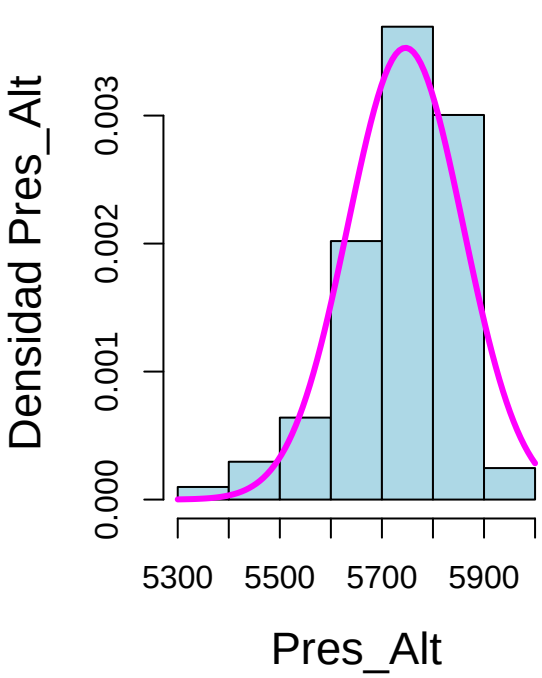
```
## [1] 4.198772
```

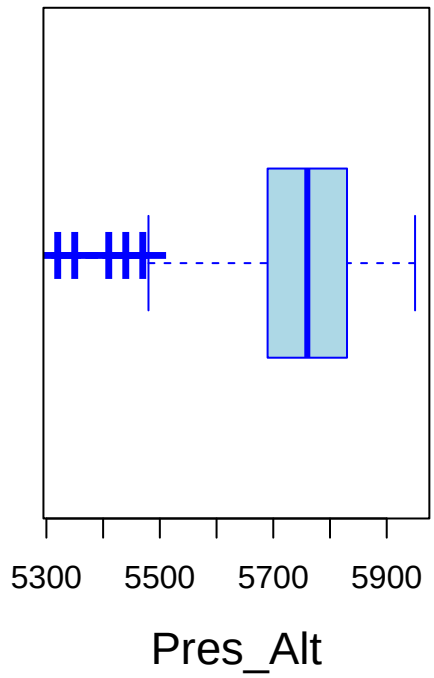
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es mayor a tres, las colas de la variable son más grandes que las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver existen 5 registros atípicos





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

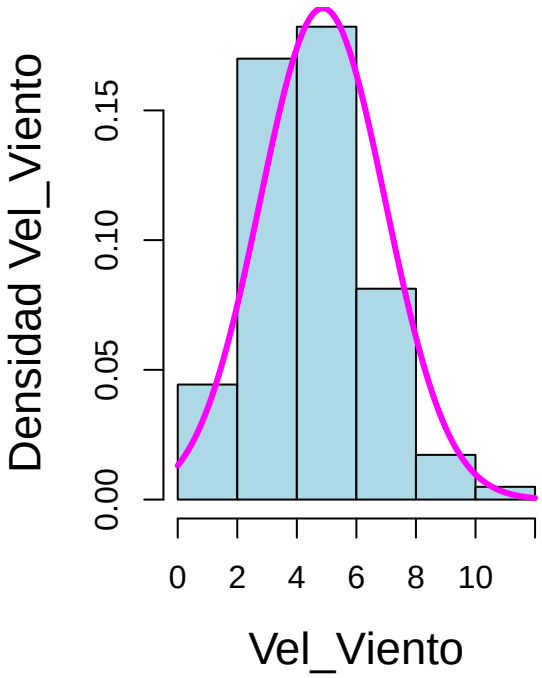
```
## [1] 3.378636
```

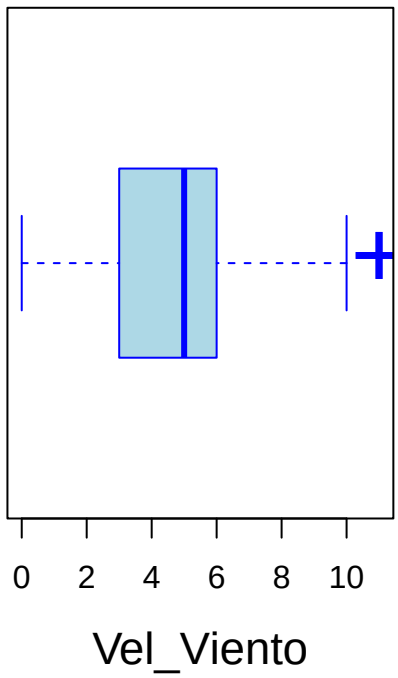
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver existen 2 registros atípicos





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

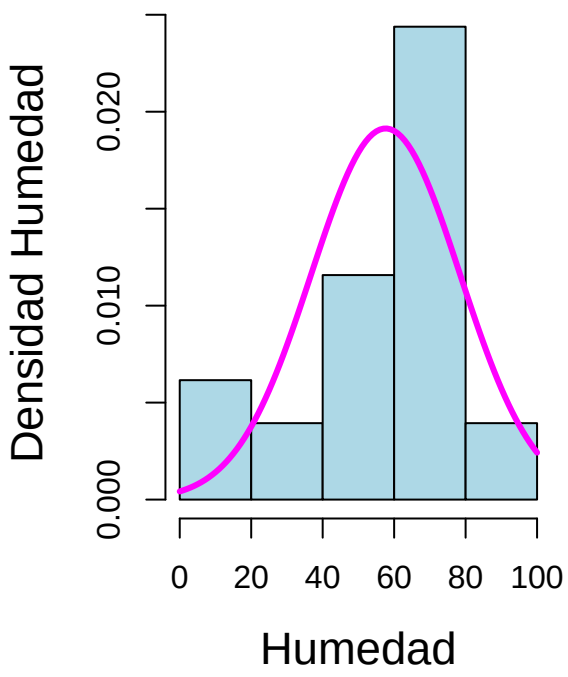
```
## [1] 2.307891
```

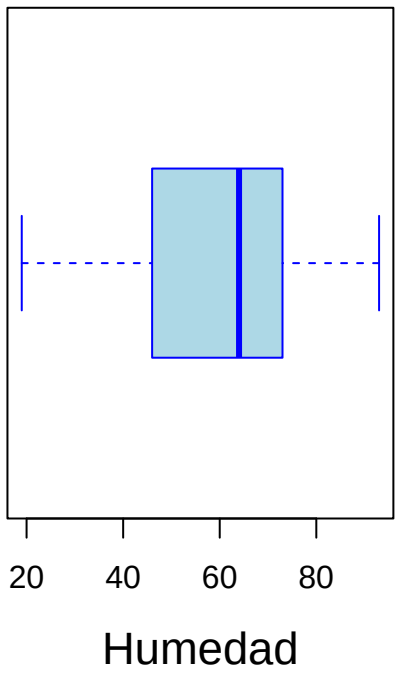
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver no existen registros atípicos





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

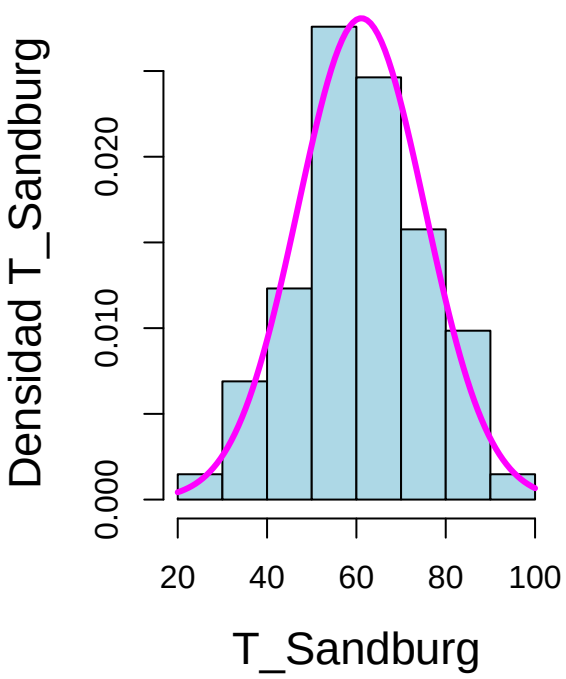
```
## [1] 2.510297
```

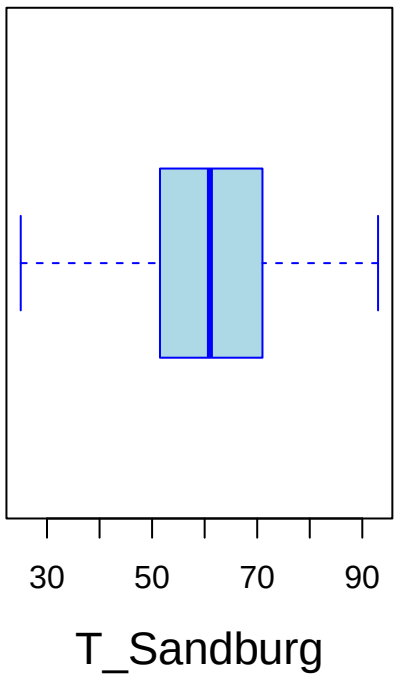
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver no existen registros atípicos





- ANÁLISIS DESCRIPTIVO VARIABLE '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:

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

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

Evaluamos la asimetría y kurtoisis

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

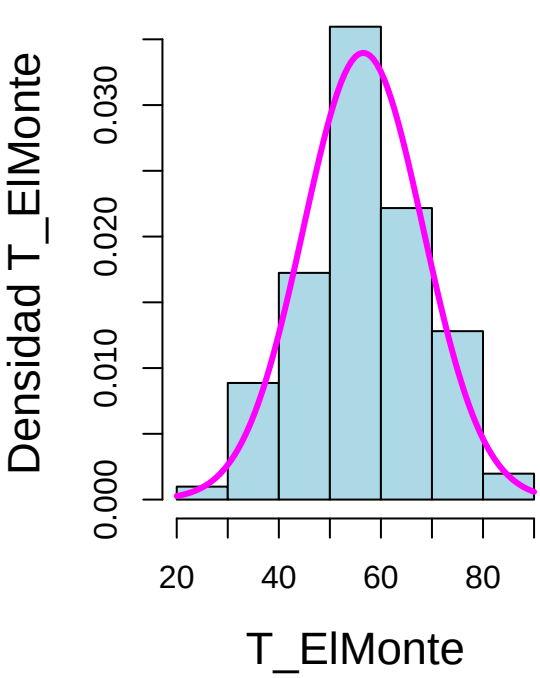
```
## [1] 2.486231
```

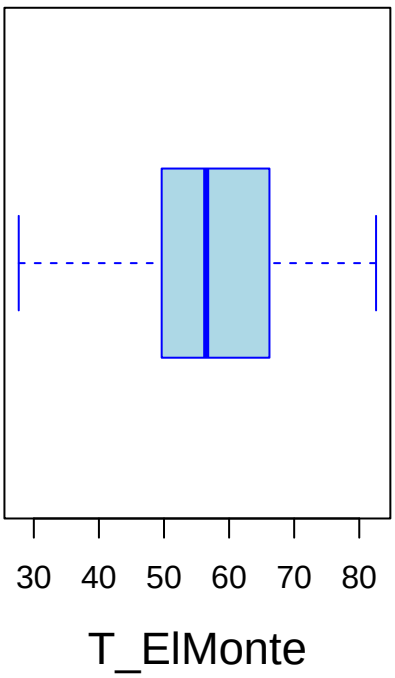
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es próximo a tres, las colas de la variable son similares a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver no existen registros atípicos





Análisis descriptivo de la variable `Inv_Alt_b` :

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

Desviación típica y rango intercuartílico:

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

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

Evaluamos la asimetría y kurtoisis

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

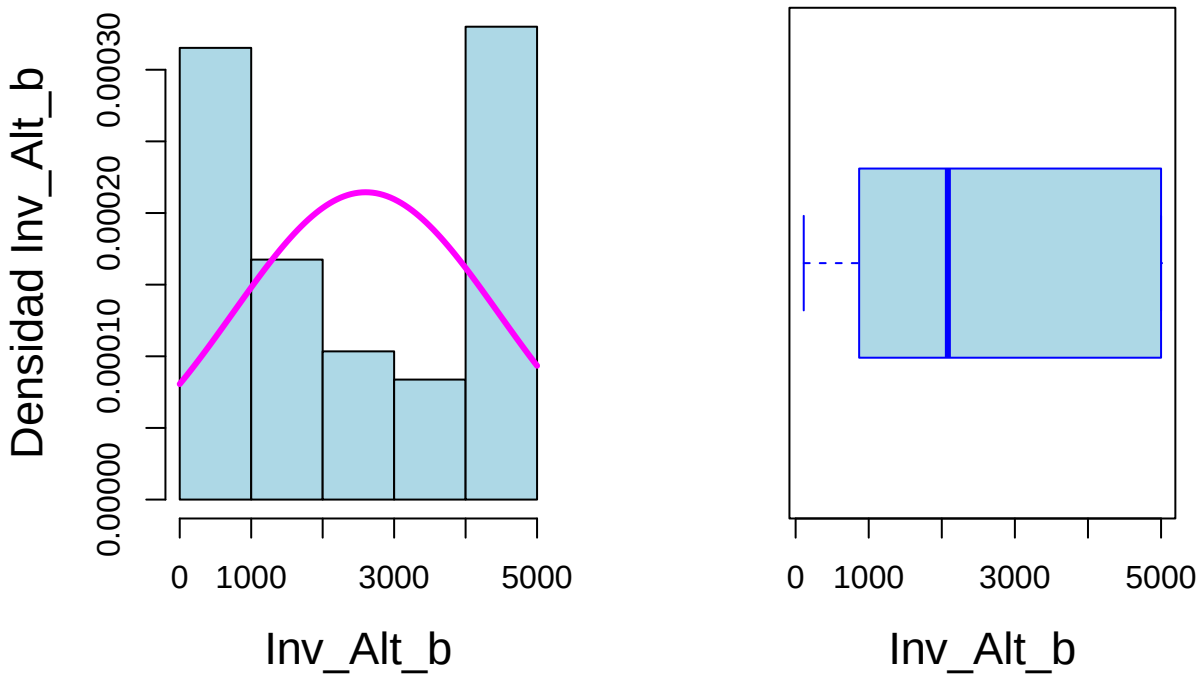
```
## [1] 1.374057
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética 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

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

Como podemos ver no existen registros atípicos



Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

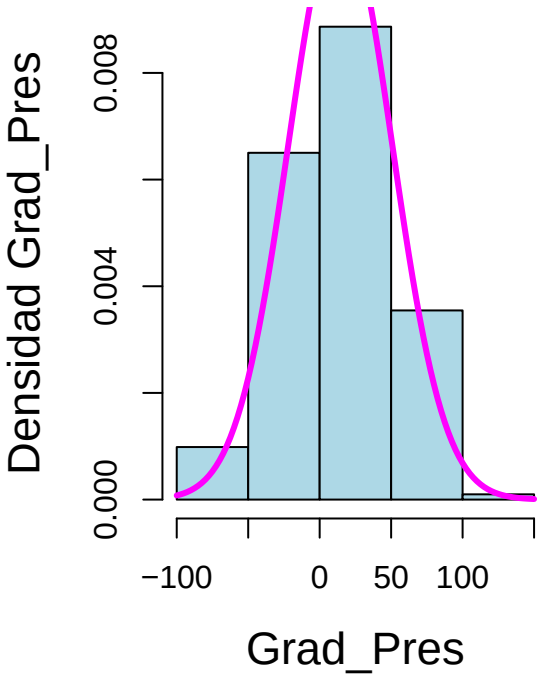
```
## [1] 2.316879
```

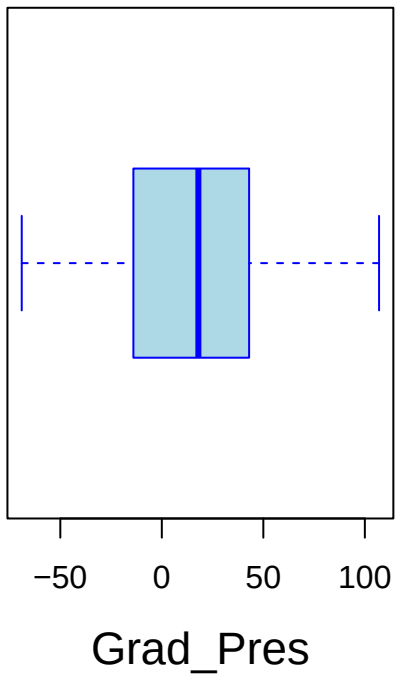
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es menor a tres, las colas de la variable son más ligeras a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver no existen registros atípicos





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

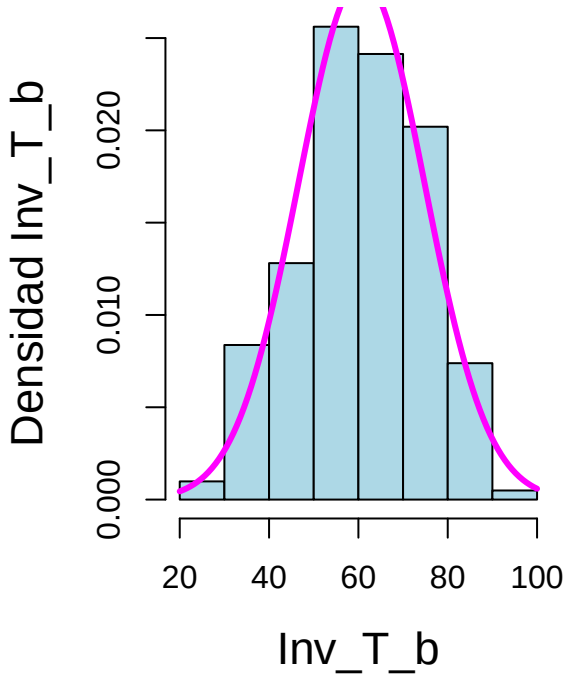
```
## [1] 2.354789
```

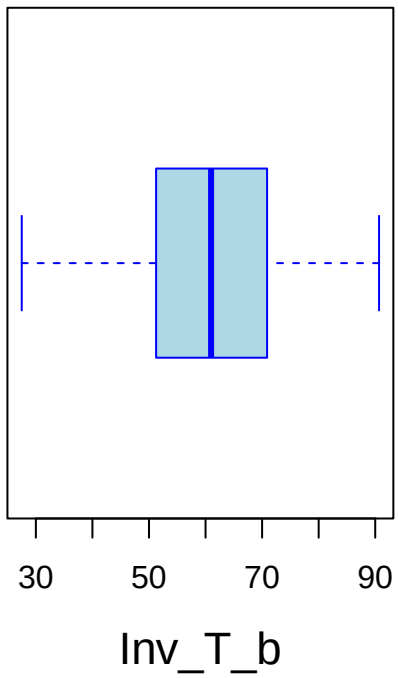
Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética y al ser el coeficiente de Kurtosis es menor a tres, las colas de la variable son más ligeras a las de una normal.

Vemos si hay registros atípicos

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

Como podemos ver no existen registros atípicos





Análisis descriptivo de la variable `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:

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

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

Evaluamos la asimetría y kurtoisis

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

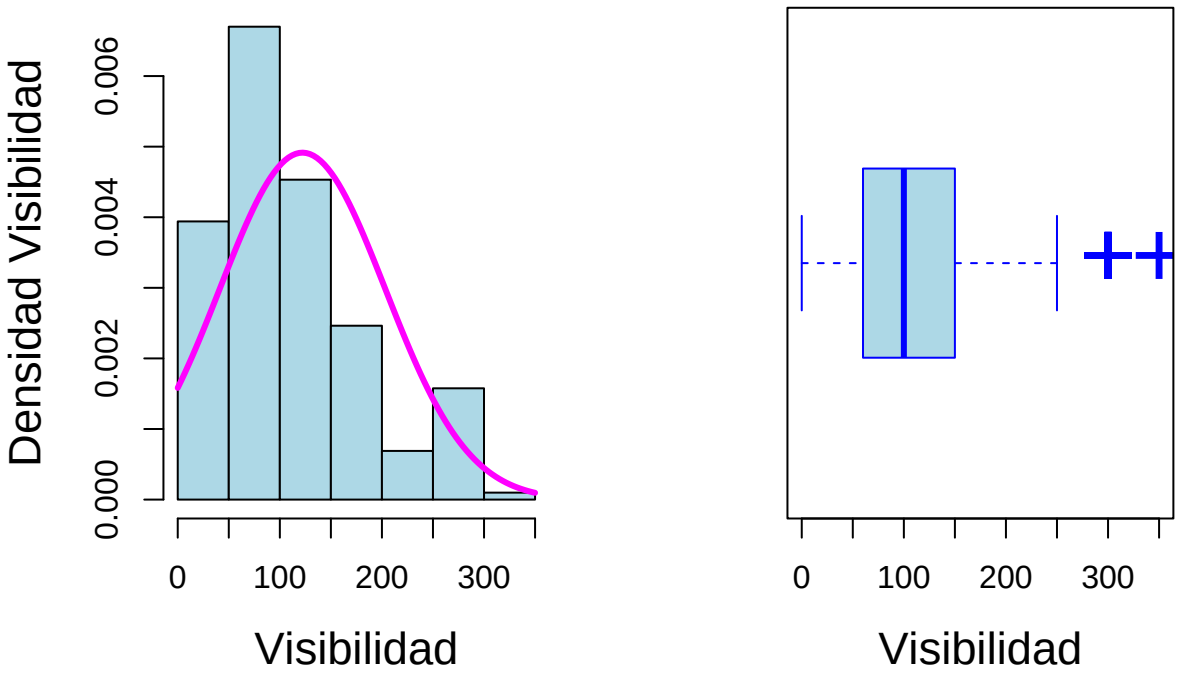
```
## [1] 2.903426
```

Podemos ver que al ser el coeficiente de asimetría cercano a 0 que puede ser una variable simética 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

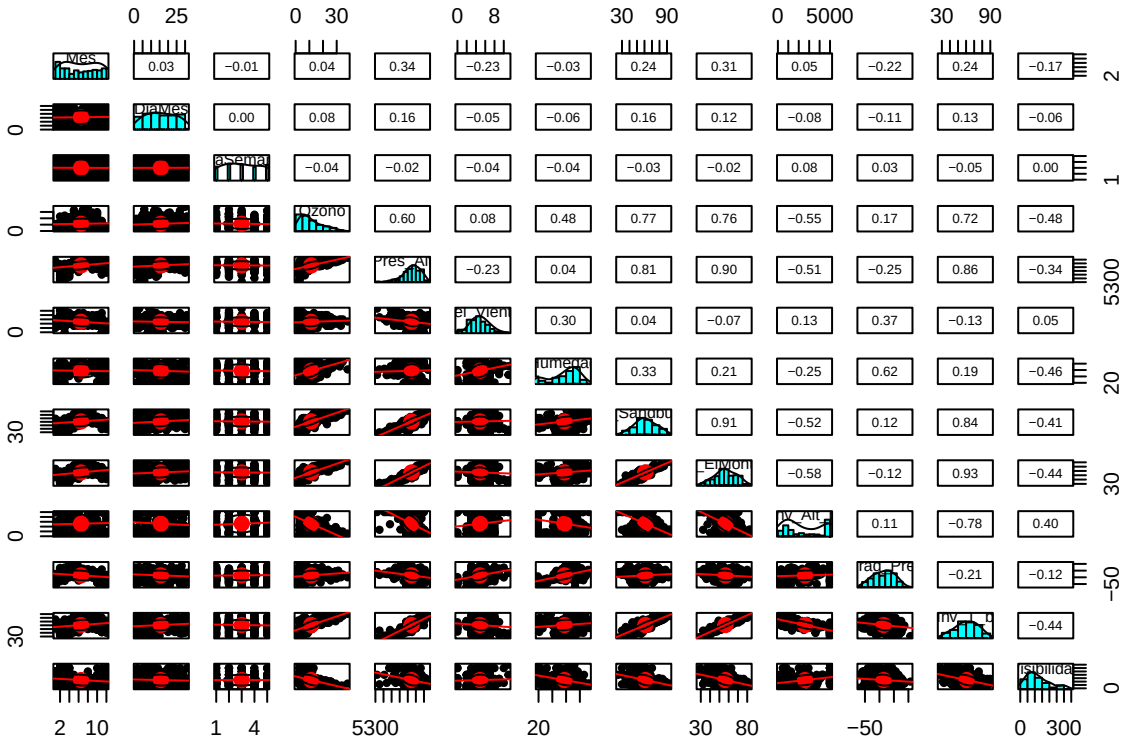
```
##  [1] 350 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300 300
```

Como podemos ver no existen registros atípicos



2. Análisis de correlación

- Correlaciones simples bivariantes(análisis gráfico y numérico):




```
## [1] 0.9308099
```

```
## [1] 8.572216e-06
```

Observamos que tenemos correlaciones muy altas y otras bajas, por lo que creemos que estamos ante un caso de posible multicolinealidad.

- Correlaciones parciales:

```
## [1] 0.5795971
```

```
## [1] 0.001780773
```

Vemos que ya no hay valores tan elevados de correlaciones parciales. Estamos ante un caso de multicolinealidad, ya que las correlaciones parciales son menores que las bivariantes.

3. Modelo matemático

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

```
##   (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_i = 55.428 - 0.343Mes_i + 0.012$Diames_i$ - 0.047$DiaSemana_i$ - 0.0133$Pres_Alt_i$ - 0.096Vel_Viento_i + 0.088$Humedad_i$ + 0.1366$T_Sandburg_i$ + 0.5598$T_ElMonte_i$ - 0.0006$Inv_Alt_b_i$ + 0.0004$Grad_Pres_i$ - 0.124$Inv_T_b_i$ - 0.005$Visibilidad_i$

Suma de residuos al cuadrado media:

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

Grados de libertad de los residuos:

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

Número de parámetros:

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

4. Análisis de multicolinealidad

```
##
## 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:

```
##
## 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.

Ahora, aplicaremos la función `step()` para obtener el modelo óptimo:

```
## 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
```

```
##
## 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_i = 51.3444845 - 0.3324536Mes_i - 0.0134013 $Pres\_Alt\_i$ + 0.0975694$Humedad_i$ + 0.1242673$T_Sandburg_i$ + 0.4743962$T_ElMonte_i$ - 0.0003211$Inv_Alt_b_i$

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:

Lo comprobaremos con un anova de modelos anidados:

```
## 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.

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

```
## 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. Posible Interacción

En este apartado analizaremos si un modelo que incluya alguna interacción entre variables originales resultaría mejor que el elegido.

Primero, definimos el modelo con todas las interacciones posibles:

```
##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes +
##     Pres_Alt + T_Sandburg:Humedad + T_Sandburg:T_ElMonte + T_Sandburg:Mes +
##     T_Sandburg:Pres_Alt + Humedad:T_ElMonte + Humedad:Mes + Humedad:Pres_Alt +
##     T_ElMonte:Mes + T_ElMonte:Pres_Alt + Mes:Pres_Alt + T_Sandburg:Humedad:T_ElMonte +
##     T_Sandburg:Humedad:Mes + T_Sandburg:Humedad:Pres_Alt + T_Sandburg:T_ElMonte:Mes +
##     T_Sandburg:T_ElMonte:Pres_Alt + T_Sandburg:Mes:Pres_Alt +
##     Humedad:T_ElMonte:Mes + Humedad:T_ElMonte:Pres_Alt + Humedad:Mes:Pres_Alt +
##     T_ElMonte:Mes:Pres_Alt + T_Sandburg:Humedad:T_ElMonte:Mes +
##     T_Sandburg:Humedad:T_ElMonte:Pres_Alt + T_Sandburg:Humedad:Mes:Pres_Alt +
##     T_Sandburg:T_ElMonte:Mes:Pres_Alt + Humedad:T_ElMonte:Mes:Pres_Alt +
##     T_Sandburg:Humedad:T_ElMonte:Mes:Pres_Alt, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8778  -2.0566   0.0579   1.9841  11.0812
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                            -2.251e+03  2.010e+03  -1.120
## T_Sandburg                              2.892e+01  4.590e+01   0.630
## Humedad                                 4.695e+01  3.521e+01   1.333
## T_ElMonte                               7.282e+01  6.300e+01   1.156
## Mes                                     7.507e+01  4.157e+02   0.181
## Pres_Alt                                3.960e-01  3.488e-01   1.135
## T_Sandburg:Humedad                     -4.482e-01  8.418e-01  -0.532
## T_Sandburg:T_ElMonte                   -9.066e-01  9.922e-01  -0.914
## T_Sandburg:Mes                          5.546e-01  8.417e+00   0.066
## T_Sandburg:Pres_Alt                    -5.163e-03  7.984e-03  -0.647
## Humedad:T_ElMonte                      -1.854e+00  1.053e+00  -1.761
## Humedad:Mes                            -2.195e+00  6.733e+00  -0.326
## Humedad:Pres_Alt                       -7.992e-03  6.119e-03  -1.306
## T_ElMonte:Mes                          -2.913e+00  1.005e+01  -0.290
## T_ElMonte:Pres_Alt                     -1.275e-02  1.073e-02  -1.188
## Mes:Pres_Alt                           -1.515e-02  7.182e-02  -0.211
## T_Sandburg:Humedad:T_ElMonte            2.275e-02  1.647e-02   1.381
## T_Sandburg:Humedad:Mes                  5.773e-03  1.404e-01   0.041
## T_Sandburg:Humedad:Pres_Alt             7.465e-05  1.464e-04   0.510
## T_Sandburg:T_ElMonte:Mes                3.796e-03  1.659e-01   0.023
```

```
## T_Sandburg:T_ElMonte:Pres_Alt              1.605e-04  1.680e-04   0.955
## T_Sandburg:Mes:Pres_Alt                   -4.994e-05  1.454e-03  -0.034
## Humedad:T_ElMonte:Mes                      1.168e-01  1.665e-01   0.702
## Humedad:T_ElMonte:Pres_Alt                 3.197e-04  1.799e-04   1.777
## Humedad:Mes:Pres_Alt                       3.706e-04  1.166e-03   0.318
## T_ElMonte:Mes:Pres_Alt                     5.445e-04  1.709e-03   0.319
## T_Sandburg:Humedad:T_ElMonte:Mes          -1.276e-03  2.664e-03  -0.479
## T_Sandburg:Humedad:T_ElMonte:Pres_Alt     -3.895e-06  2.789e-06  -1.397
## T_Sandburg:Humedad:Mes:Pres_Alt           -9.371e-07  2.431e-05  -0.039
## T_Sandburg:T_ElMonte:Mes:Pres_Alt         -1.522e-06  2.808e-05  -0.054
## Humedad:T_ElMonte:Mes:Pres_Alt            -2.006e-05  2.837e-05  -0.707
## T_Sandburg:Humedad:T_ElMonte:Mes:Pres_Alt  2.198e-07  4.516e-07   0.487
##                                           Pr(>|t|)
## (Intercept)                                 0.2643
## T_Sandburg                                  0.5294
## Humedad                                     0.1842
## T_ElMonte                                   0.2493
## Mes                                         0.8569
## Pres_Alt                                    0.2578
## T_Sandburg:Humedad                          0.5951
## T_Sandburg:T_ElMonte                        0.3621
## T_Sandburg:Mes                              0.9475
## T_Sandburg:Pres_Alt                         0.5187
## Humedad:T_ElMonte                           0.0801 .
## Humedad:Mes                                 0.7448
## Humedad:Pres_Alt                            0.1933
## T_ElMonte:Mes                               0.7723
## T_ElMonte:Pres_Alt                          0.2366
## Mes:Pres_Alt                                0.8332
## T_Sandburg:Humedad:T_ElMonte                0.1690
## T_Sandburg:Humedad:Mes                      0.9673
## T_Sandburg:Humedad:Pres_Alt                 0.6107
## T_Sandburg:T_ElMonte:Mes                    0.9818
## T_Sandburg:T_ElMonte:Pres_Alt               0.3408
## T_Sandburg:Mes:Pres_Alt                     0.9726
## Humedad:T_ElMonte:Mes                       0.4839
## Humedad:T_ElMonte:Pres_Alt                  0.0773 .
## Humedad:Mes:Pres_Alt                        0.7510
## T_ElMonte:Mes:Pres_Alt                      0.7504
## T_Sandburg:Humedad:T_ElMonte:Mes            0.6325
## T_Sandburg:Humedad:T_ElMonte:Pres_Alt       0.1643
## T_Sandburg:Humedad:Mes:Pres_Alt             0.9693
## T_Sandburg:T_ElMonte:Mes:Pres_Alt           0.9568
## Humedad:T_ElMonte:Mes:Pres_Alt              0.4804
## T_Sandburg:Humedad:T_ElMonte:Mes:Pres_Alt   0.6271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.86 on 171 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.778
## F-statistic: 23.84 on 31 and 171 DF,  p-value: < 2.2e-16
```

Ninguna variable resulta significativa. Sin embargo, haremos una selección secuencial para comprobar que ninguna interacción es significativa:

```
## 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
## + Pres_Alt    1    4818.1  8740.0 767.78
## + Humedad     1    3116.9 10441.2 803.88
## <none>                    13558.1 854.91
## + Mes         1      26.5 13531.7 856.52
##
## Step:  AIC=671.88
## Ozono ~ T_Sandburg
##
##              Df Sum of Sq     RSS    AIC
## + Humedad     1     759.0  4690.4 643.43
## + Mes         1     273.0  5176.4 663.45
## + T_ElMonte   1     233.1  5216.3 665.01
## <none>                     5449.4 671.88
## + Pres_Alt    1      29.2  5420.2 672.79
## - 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
## + T_Sandburg:Humedad   1     370.6  4319.7 628.72
## + Mes                  1     175.2  4515.2 637.70
## <none>                              4690.4 643.43
## + Pres_Alt             1      41.5  4648.9 643.63
## - 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
## + Humedad:T_ElMonte      1    591.44 3593.6 593.36
## + T_Sandburg:T_ElMonte   1    451.49 3733.6 601.12
## + Mes                    1    358.12 3827.0 606.13
## + T_Sandburg:Humedad     1    297.36 3887.7 609.33
## + Pres_Alt               1    108.61 4076.5 618.96
## <none>                               4185.1 622.29
## - 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=593.36
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Humedad:T_ElMonte
##
##                         Df Sum of Sq    RSS    AIC
## + Mes                    1    381.05 3212.6 572.61
## + T_Sandburg:T_ElMonte   1    222.72 3370.9 582.38
```

```
## + T_Sandburg:Humedad    1     86.55 3507.1 590.42
## + Pres_Alt              1     44.12 3549.5 592.86
## <none>                              3593.6 593.36
## - T_Sandburg            1     54.08 3647.7 594.40
## - Humedad:T_ElMonte     1    591.44 4185.1 622.29
##
## Step:  AIC=572.61
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte
##
##                         Df Sum of Sq    RSS    AIC
## + T_Sandburg:T_ElMonte  1    157.46 3055.1 564.41
## + T_Sandburg:Humedad    1     41.60 3171.0 571.96
## + Humedad:Mes           1     40.69 3171.9 572.02
## - T_Sandburg            1     27.57 3240.2 572.34
## <none>                              3212.6 572.61
## + Pres_Alt              1     19.78 3192.8 573.36
## + T_ElMonte:Mes         1     16.67 3195.9 573.55
## + T_Sandburg:Mes        1      0.01 3212.6 574.61
## - Mes                   1    381.05 3593.6 593.36
## - Humedad:T_ElMonte     1    614.37 3827.0 606.13
##
## Step:  AIC=564.41
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte
##
##                         Df Sum of Sq    RSS    AIC
## + T_Sandburg:Mes        1     40.29 3014.8 563.71
## + T_Sandburg:Humedad    1     36.94 3018.2 563.94
## <none>                              3055.1 564.41
## + Humedad:Mes           1     29.01 3026.1 564.47
## + T_ElMonte:Mes         1      2.81 3052.3 566.22
## + Pres_Alt              1      0.06 3055.1 566.41
## - T_Sandburg:T_ElMonte  1    157.46 3212.6 572.61
## - Mes                   1    315.78 3370.9 582.38
## - Humedad:T_ElMonte     1    402.97 3458.1 587.56
##
## Step:  AIC=563.71
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes
##
##                         Df Sum of Sq    RSS    AIC
## + T_Sandburg:Humedad    1     51.45 2963.4 562.22
## <none>                              3014.8 563.71
## + T_ElMonte:Mes         1     28.75 2986.1 563.77
## + Humedad:Mes           1     22.63 2992.2 564.18
## - T_Sandburg:Mes        1     40.29 3055.1 564.41
## + Pres_Alt              1      0.37 3014.5 565.69
## - T_Sandburg:T_ElMonte  1    197.74 3212.6 574.61
## - Humedad:T_ElMonte     1    392.69 3407.5 586.57
##
## Step:  AIC=562.22
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad
##
```

```
##                                Df Sum of Sq    RSS    AIC
## + T_ElMonte:Mes                 1    41.882 2921.5 561.33
## + Humedad:Mes                   1    34.135 2929.3 561.87
## <none>                                      2963.4 562.22
## + T_Sandburg:Humedad:T_ElMonte  1     9.571 2953.8 563.56
## - T_Sandburg:Humedad            1    51.448 3014.8 563.71
## - T_Sandburg:Mes                1    54.800 3018.2 563.94
## + Pres_Alt                      1     2.969 2960.4 564.02
## - T_Sandburg:T_ElMonte          1   206.322 3169.7 573.88
## - Humedad:T_ElMonte             1   245.208 3208.6 576.36
##
## Step:  AIC=561.33
## Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad +
##     T_ElMonte:Mes
##
##                                Df Sum of Sq    RSS    AIC
## <none>                                      2921.5 561.33
## + Humedad:Mes                   1    19.853 2901.7 561.95
## + T_Sandburg:T_ElMonte:Mes      1    19.037 2902.5 562.00
## - T_ElMonte:Mes                 1    41.882 2963.4 562.22
## + T_Sandburg:Humedad:T_ElMonte  1     9.526 2912.0 562.67
## + Pres_Alt                      1     3.917 2917.6 563.06
## - T_Sandburg:Humedad            1    64.583 2986.1 563.77
## - T_Sandburg:Mes                1    93.093 3014.6 565.70
## - T_Sandburg:T_ElMonte          1   175.325 3096.8 571.16
## - Humedad:T_ElMonte             1   266.705 3188.2 577.06
```

Que resulta en el modelo:

```
##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes +
##     Humedad:T_ElMonte + T_Sandburg:T_ElMonte + T_Sandburg:Mes +
##     T_Sandburg:Humedad + T_ElMonte:Mes, data = OzonoLA)
##
## Coefficients:
##          (Intercept)            T_Sandburg               Humedad
##            23.308869              0.284784             -0.379648
##            T_ElMonte                   Mes     Humedad:T_ElMonte
##            -0.924179              0.064954              0.014662
## T_Sandburg:T_ElMonte        T_Sandburg:Mes    T_Sandburg:Humedad
##             0.005711             -0.035228             -0.005851
##        T_ElMonte:Mes
##             0.028480
```

Ahora comprobaremos utilizando un anova de modelos anidados si el modelo con alguna interacción es mejor que el modelo seleccionado en el apartado anterior:

```
## Analysis of Variance Table
##
## Model 1: Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Pres_Alt
## Model 2: Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad +
##     T_ElMonte:Mes
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    197 3756.1
## 2    193 2921.5  4    834.61 13.784 6.605e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Mes +
##     Humedad:T_ElMonte + T_Sandburg:T_ElMonte + T_Sandburg:Mes +
##     T_Sandburg:Humedad + T_ElMonte:Mes, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0091  -2.1099  -0.3829   2.1788  13.1805
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.308869   6.670968   3.494 0.000590 ***
## T_Sandburg            0.284784   0.214590   1.327 0.186042
## Humedad              -0.379648   0.093792  -4.048 7.48e-05 ***
## T_ElMonte            -0.924179   0.240509  -3.843 0.000165 ***
## Mes                   0.064954   0.575969   0.113 0.910328
## Humedad:T_ElMonte     0.014662   0.003493   4.197 4.12e-05 ***
## T_Sandburg:T_ElMonte  0.005711   0.001678   3.403 0.000810 ***
## T_Sandburg:Mes       -0.035228   0.014205  -2.480 0.013999 *
## T_Sandburg:Humedad   -0.005851   0.002832  -2.066 0.040209 *
## T_ElMonte:Mes         0.028480   0.017122   1.663 0.097864 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.891 on 193 degrees of freedom
## Multiple R-squared:  0.7845, Adjusted R-squared:  0.7745
## F-statistic: 78.07 on 9 and 193 DF,  p-value: < 2.2e-16
```

¿Qué pasa si quitamos `Mes`?

```
##
## Call:
## lm(formula = Ozono ~ T_Sandburg + Humedad + T_ElMonte + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad +
##     T_ElMonte:Mes, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0516  -2.1517  -0.3506   2.1724  13.1922
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.415575   6.586699   3.555 0.000475 ***
## T_Sandburg            0.287394   0.212795   1.351 0.178407
## Humedad              -0.380028   0.093492  -4.065 6.98e-05 ***
## T_ElMonte            -0.923765   0.239869  -3.851 0.000160 ***
## Humedad:T_ElMonte     0.014641   0.003479   4.208 3.94e-05 ***
## T_Sandburg:T_ElMonte  0.005624   0.001484   3.790 0.000201 ***
## T_Sandburg:Mes       -0.035066   0.014097  -2.487 0.013708 *
```

```
## T_Sandburg:Humedad   -0.005826   0.002817  -2.068 0.039948 *
## T_ElMonte:Mes         0.029489   0.014558   2.026 0.044170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.881 on 194 degrees of freedom
## Multiple R-squared:  0.7845, Adjusted R-squared:  0.7756
## F-statistic: 88.28 on 8 and 194 DF,  p-value: < 2.2e-16
```

¿Qué pasa si quitamos `T_Sandburg`?

```
##
## Call:
## lm(formula = Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad +
##     T_ElMonte:Mes, data = OzonoLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3029  -2.3010  -0.3295   2.4212  12.7228
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          24.997570   6.495389   3.849 0.000161 ***
## Humedad              -0.362499   0.092783  -3.907 0.000129 ***
## T_ElMonte            -0.696112   0.171019  -4.070 6.82e-05 ***
## Humedad:T_ElMonte     0.011000   0.002204   4.992 1.32e-06 ***
## T_ElMonte:T_Sandburg  0.006328   0.001392   4.545 9.61e-06 ***
## T_Sandburg:Mes       -0.026023   0.012431  -2.093 0.037615 *
## Humedad:T_Sandburg   -0.002639   0.001542  -1.712 0.088521 .
## T_ElMonte:Mes         0.019976   0.012767   1.565 0.119289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.889 on 195 degrees of freedom
## Multiple R-squared:  0.7825, Adjusted R-squared:  0.7747
## F-statistic: 100.2 on 7 and 195 DF,  p-value: < 2.2e-16
```

¿Qué pasa si quitamos `T_ElMonte:Mes`?

```
##
## Call:
## lm(formula = Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad,
##     data = OzonoLA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.128  -2.327  -0.175   2.470  12.181
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.841847   6.477047   3.681 0.000300 ***
## Humedad              -0.381982   0.092282  -4.139 5.17e-05 ***
## T_ElMonte            -0.623647   0.165235  -3.774 0.000213 ***
## Humedad:T_ElMonte     0.012246   0.002062   5.938 1.29e-08 ***
```

```
## T_ElMonte:T_Sandburg  0.005548   0.001305   4.252 3.27e-05 ***
## T_Sandburg:Mes       -0.006707   0.001465  -4.577 8.38e-06 ***
## Humedad:T_Sandburg   -0.003587   0.001423  -2.521 0.012515 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 196 degrees of freedom
## Multiple R-squared:  0.7797, Adjusted R-squared:  0.773
## F-statistic: 115.6 on 6 and 196 DF,  p-value: < 2.2e-16
```

Obtenemos el modelo:

```
##
## Call:
## lm(formula = Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad,
##     data = OzonoLA)
##
## Coefficients:
##          (Intercept)              Humedad             T_ElMonte
##            23.841847            -0.381982            -0.623647
##    Humedad:T_ElMonte  T_ElMonte:T_Sandburg        T_Sandburg:Mes
##             0.012246             0.005548            -0.006707
##   Humedad:T_Sandburg
##            -0.003587
```

7. Inferencia modelo

Ahora ya podemos comenzar la inferencia.

```
##
## Call:
## lm(formula = Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte +
##     T_Sandburg:T_ElMonte + T_Sandburg:Mes + T_Sandburg:Humedad,
##     data = OzonoLA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.128  -2.327  -0.175   2.470  12.181
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.841847   6.477047   3.681 0.000300 ***
## Humedad              -0.381982   0.092282  -4.139 5.17e-05 ***
## T_ElMonte            -0.623647   0.165235  -3.774 0.000213 ***
## Humedad:T_ElMonte     0.012246   0.002062   5.938 1.29e-08 ***
## T_ElMonte:T_Sandburg  0.005548   0.001305   4.252 3.27e-05 ***
## T_Sandburg:Mes       -0.006707   0.001465  -4.577 8.38e-06 ***
## Humedad:T_Sandburg   -0.003587   0.001423  -2.521 0.012515 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.903 on 196 degrees of freedom
## Multiple R-squared:  0.7797, Adjusted R-squared:  0.773
## F-statistic: 115.6 on 6 and 196 DF,  p-value: < 2.2e-16
```

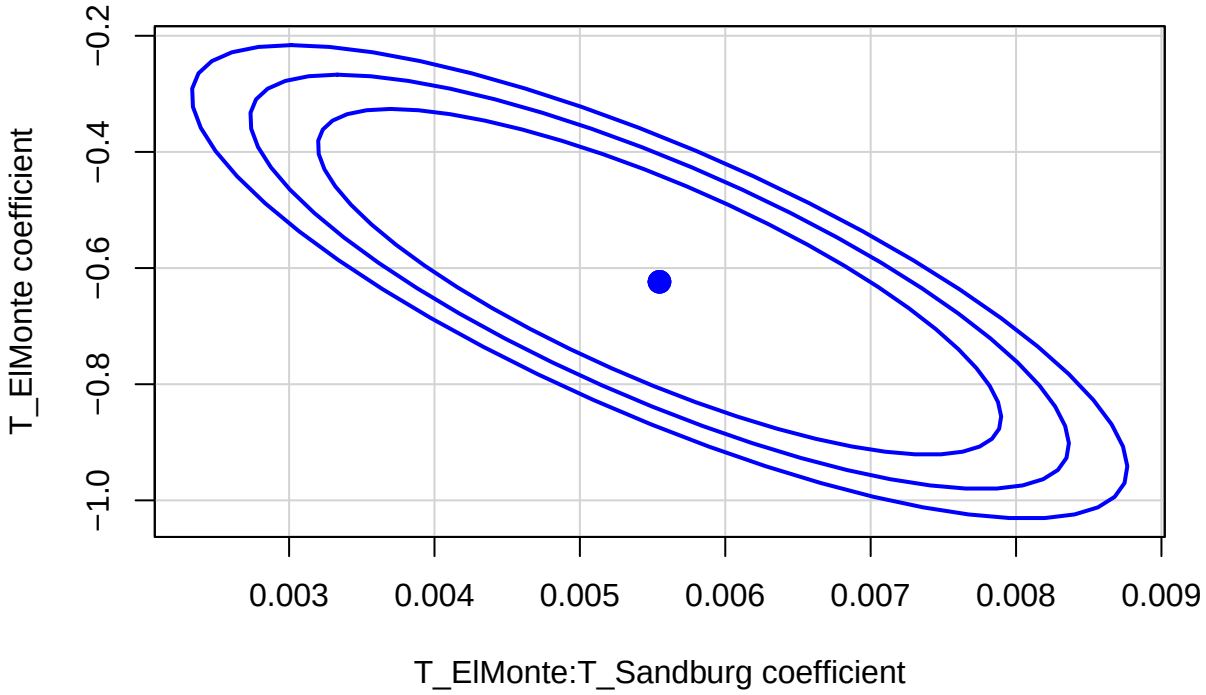
- Intervalos de confianza para los coeficientes:

```
##                              2.5 %        97.5 %
## (Intercept)          11.068194333 36.6154987621
## Humedad              -0.563975564 -0.1999880152
## T_ElMonte            -0.949513793 -0.2977804150
## Humedad:T_ElMonte     0.008178648  0.0163124929
## T_ElMonte:T_Sandburg  0.002974741  0.0081205709
## T_Sandburg:Mes       -0.009596628 -0.0038166522
## Humedad:T_Sandburg   -0.006392896 -0.0007803548
```

- Intervalos de confianza para σ^2:

```
## [1] 22.97615
```

Las elipses al 80%, 90% y 95% de confianza para el vector de coeficientes:



8. Validación modelo seleccionado

Por abreviar la notación, tenemos:

Primero, calculamos el coeficiente de robusted del ajuste:

```
## [1] 0.9318002
```

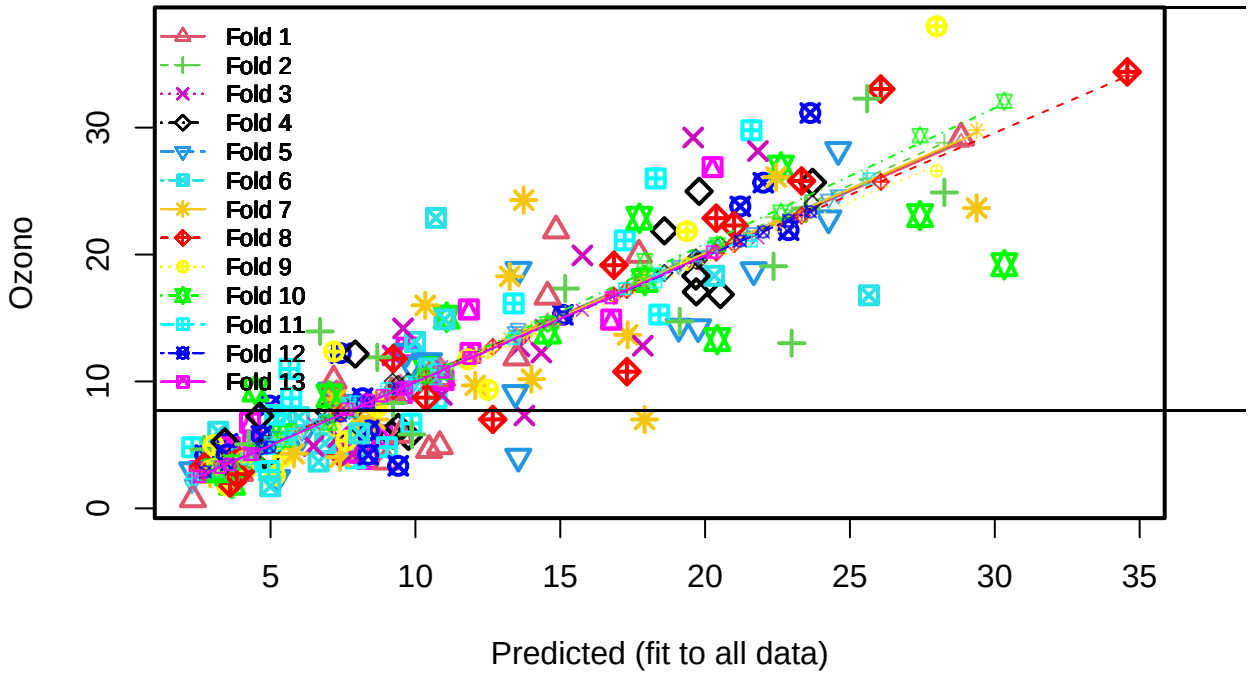
Elevado y superior al del modelo completo

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

Haremos una validación del tipo LOOCV (Leave One Out Cross Validation):

Primero, para MS:

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




```
##
## fold 1
## Observations in test set: 15
##                    13        14        26         32        37       52
## Predicted    8.982347 10.842661  9.489688 10.9198024  3.944223 7.282015
## cvpred       8.957178 11.190199  9.857542 11.1186895  3.900958 7.205498
## Ozono        3.690000  4.900000  5.800000 10.2700000  2.790000 8.900000
## CV residual -5.267178 -6.290199 -4.057542 -0.8486895 -1.110958 1.694502
##                    62        68        80        112       146       149
## Predicted    7.177872 14.856212 17.715830 28.835167 10.47097 14.562629
## cvpred       7.426509 14.857191 17.906519 28.816815 10.83238 14.658915
## Ozono       10.070000 21.900000 19.980000 29.210000  4.60000 16.680000
## CV residual  2.643491  7.042809  2.073481  0.393185 -6.23238  2.021085
##                   160       177       203
## Predicted   9.19201590 13.488849  2.322007
## cvpred      9.07027725 13.509255  2.550643
## Ozono       9.14000000 11.890000  0.720000
## CV residual 0.06972275 -1.619255 -1.830643
##
## Sum of squares = 198.55    Mean square = 13.24    n = 15
##
## fold 2
## Observations in test set: 16
##                     20       40        51        69        83        94
## Predicted    3.4671167 4.550304  6.707182 15.171474 28.258249  8.683377
## cvpred       2.9394546 4.540461  6.301865 15.312934 28.813912  8.733813
## Ozono        2.1800000 5.650000 13.940000 17.320000 24.890000 11.900000
## CV residual -0.7594546 1.109539  7.638135  2.007066 -3.923912  3.166187
##                    114       123        127       129       133       134
```

```
## Predicted    22.98818  9.883713 23.4841781 25.594831 22.363722 19.125334
## cvpred       23.50469  9.976835 23.7642416 25.965531 22.942709 19.348244
## Ozono        13.02000  5.820000 23.6200000 32.280000 19.080000 14.730000
## CV residual -10.48469 -4.156835 -0.1442416  6.314469 -3.862709 -4.618244
##                   176      186        199       202
## Predicted    9.2251370 2.699244  3.8640594 4.2430930
## cvpred       9.2925299 2.668357  3.7655344 4.1881775
## Ozono        8.3000000 4.650000  3.2100000 5.0500000
## CV residual -0.9925299 1.981643 -0.5555344 0.8618225
##
## Sum of squares = 298.91    Mean square = 18.68    n = 16
##
## fold 3
## Observations in test set: 16
##                     7        18         27         28        49        54
## Predicted    8.506684 14.346293 10.5442683 13.5887637 10.910559  9.210901
## cvpred       8.626065 14.442126 10.3045743 13.4812123 11.163694  9.321868
## Ozono        4.730000 12.280000 10.6000000 12.7700000  8.930000 12.050000
## CV residual -3.896065 -2.162126  0.2954257 -0.7112123 -2.233694  2.728132
##                   101       122      137      142       143       155       166
## Predicted   15.764898  7.839762 19.58768 17.84775 13.762564 21.823399  7.336315
## cvpred      15.613718  8.156258 19.24831 17.71678 13.629587 21.310401  7.690004
## Ozono       19.930000  4.260000 29.22000 12.81000  7.320000 28.150000  5.620000
## CV residual  4.316282 -3.896258  9.97169 -4.90678 -6.309587  6.839599 -2.070004
##                   167       168        184
## Predicted    6.483792  9.575903 2.81214283
## cvpred       6.711683  9.630773 2.99579925
## Ozono        4.910000 14.180000 3.04000000
## CV residual -1.801683  4.549227 0.04420075
##
## Sum of squares = 305.02    Mean square = 19.06    n = 16
##
## fold 4
## Observations in test set: 16
##                     1         2         6        16        24         38
## Predicted    7.737462  9.212363  9.404386  9.757764  4.844334  4.9537833
## cvpred       7.881856  9.989071  9.814686 10.023336  5.085251  5.1270986
## Ozono        5.340000  5.770000  6.390000  5.680000  4.080000  4.3200000
## CV residual -2.541856 -4.219071 -3.424686 -4.343336 -1.005251 -0.8070986
##                    43        87       109       110       124       128
## Predicted   7.0221250 20.526127 18.595954 19.794148  7.920426 23.706804
## cvpred      7.2037124 20.731316 18.548854 19.564455  7.725051 24.014729
## Ozono       7.6300000 16.850000 21.870000 24.980000 12.160000 25.690000
## CV residual 0.4262876 -3.881316  3.321146  5.415545  4.434949  1.675271
##                   136      152      189      193
## Predicted   19.695573 19.68186 4.627838 3.421309
## cvpred      19.489827 19.77399 4.789873 3.282514
## Ozono       17.060000 18.31000 7.260000 5.230000
## CV residual -2.429827 -1.46399 2.470127 1.947486
##
## Sum of squares = 152.54    Mean square = 9.53    n = 16
##
## fold 5
## Observations in test set: 16
```

```
##                    11       19        25         29       41        55       75
## Predicted    13.55223 7.096635 7.7087799  6.0812899 2.294117  9.689065 13.56228
## cvpred       14.19298 7.002620 7.7268795  5.9853765 1.888374  9.565101 13.65904
## Ozono         4.07000 9.290000 8.3200000  5.7300000 3.010000 12.330000 18.79000
## CV residual -10.12298 2.287380 0.5931205 -0.2553765 1.121626  2.764899  5.13096
##                    77        78        97        99       106       111
## Predicted    9.842958  5.229252 19.088162 24.259577 19.759550 24.593753
## cvpred       9.524743  5.653732 19.157702 24.376734 19.889952 24.710035
## Ozono       11.300000  2.390000 14.310000 22.850000 14.270000 28.240000
## CV residual  1.775257 -3.263732 -4.847702 -1.526734 -5.619952  3.529965
##                   118       121       178
## Predicted   10.47099 21.677629 13.450342
## cvpred      10.43241 21.730341 13.940592
## Ozono       11.60000 18.770000  9.090000
## CV residual  1.16759 -2.960341 -4.850592
##
## Sum of squares = 260.69    Mean square = 16.29    n = 16
##
## fold 6
## Observations in test set: 16
##                     3       31       34        36       39        56        60
## Predicted    6.657804 3.194544 10.70856  3.863841 5.980842 7.8068008  9.987907
## cvpred       6.571420 2.879091 10.22520  4.222528 6.050937 7.6137972  9.737287
## Ozono        3.690000 6.040000 22.89000  3.220000 7.190000 7.9300000 13.120000
## CV residual -2.881420 3.160909 12.66480 -1.002528 1.139063 0.3162028  3.382713
##                    67       70       108       138        144       148
## Predicted   11.004093 5.545952 25.655910 20.316026 10.4653527  6.880936
## cvpred      10.788608 5.419450 25.954103 20.594243 10.4844001  6.925148
## Ozono       14.890000 7.260000 16.790000 18.330000 11.0200000  5.140000
## CV residual  4.101392 1.840550 -9.164103 -2.264243  0.5355999 -1.785148
##                   175       183       200
## Predicted    8.058175 2.9371306  4.990653
## cvpred       8.197963 2.8401539  5.175849
## Ozono        5.910000 3.0100000  1.740000
## CV residual -2.287963 0.1698461 -3.435849
##
## Sum of squares = 322.4    Mean square = 20.15    n = 16
##
## fold 7
## Observations in test set: 16
##                    10         22        30       46        50       53
## Predicted    8.604016  2.9127922  7.396330 13.73688 14.004453 7.253434
## cvpred       8.535789  3.0496279  7.541787 13.61960 14.065571 6.949026
## Ozono        7.000000  2.7400000  4.040000 24.29000 10.180000 8.600000
## CV residual -1.535789 -0.3096279 -3.501787 10.67040 -3.885571 1.650974
##                    71       105       115       119       154       157
## Predicted   12.066415 29.372247 22.453339 17.327016  17.91151 13.253765
## cvpred      11.961353 29.800113 22.517384 17.352485  18.09001 13.354802
## Ozono        9.690000 23.660000 26.100000 13.670000   7.00000 18.280000
## CV residual -2.271353 -6.140113  3.582616 -3.682485 -11.09001  4.925198
##                   162      169       188        197
## Predicted    8.0042651 10.33919  5.807569  3.5745943
## cvpred       7.9330843 10.38315  5.910210  3.7948687
## Ozono        7.2000000 16.00000  4.310000  3.3300000
```

```
## CV residual -0.7330843  5.61685 -1.600210 -0.4648687
##
## Sum of squares = 397.76    Mean square = 24.86    n = 16
##
## fold 8
## Observations in test set: 16
##                     8        35        66        79        82        88
## Predicted    7.412348  3.746099  3.951221  9.235488 26.064440 16.854049
## cvpred       7.485404  3.942147  4.156817  9.245854 25.727345 16.762888
## Ozono        4.350000  2.260000  2.880000 11.790000 33.040000 19.160000
## CV residual -3.135404 -1.682147 -1.276817  2.544146  7.312655  2.397112
##                    90        98       104       113       126       135
## Predicted   10.37042 17.301927 34.5737363 20.384466 21.002215 23.330703
## cvpred      10.37008 17.211035 34.0596965 20.144203 20.831682 23.077218
## Ozono        8.73000 10.770000 34.3900000 22.870000 22.290000 25.800000
## CV residual -1.64008 -6.441035  0.3303035  2.725797  1.458318  2.722782
##                   179       187       192       201
## Predicted   12.670748 2.7070871  3.597482  3.833793
## cvpred      12.711497 2.9890667  3.789671  3.976447
## Ozono        7.010000 3.2900000  2.000000  2.690000
## CV residual -5.701497 0.3009333 -1.789671 -1.286447
##
## Sum of squares = 178.7    Mean square = 11.17    n = 16
##
## fold 9
## Observations in test set: 16
##                    23       58        93       117      130       153       156
## Predicted    4.115774 3.206236  3.497667 12.506908 27.99210  7.188790 19.366568
## cvpred       3.968654 3.149674  3.606809 12.316639 26.58883  6.171132 19.279886
## Ozono        2.920000 4.330000  1.800000  9.350000 37.98000 12.360000 21.840000
## CV residual -1.048654 1.180326 -1.806809 -2.966639 11.39117  6.188868  2.560114
##                   161       163        165       172       173        181
## Predicted   11.812141  5.154108  8.6865851  7.607804  5.430406  2.8104651
## cvpred      11.970608  5.078454  8.7463943  7.728425  5.204080  3.2676722
## Ozono       11.750000  2.610000  8.0100000  5.330000  4.100000  2.8200000
## CV residual -0.220608 -2.468454 -0.7363943 -2.398425 -1.104080 -0.4476722
##                   182      190        195
## Predicted    3.6387985 3.031687  4.0209263
## cvpred       3.7321039 3.457716  4.0399785
## Ozono        3.1900000 4.980000  3.6800000
## CV residual -0.5421039 1.522284 -0.3599785
##
## Sum of squares = 205.77    Mean square = 12.86    n = 16
##
## fold 10
## Observations in test set: 15
##                     17        33        42       59       73        96      103
## Predicted   10.3408262 11.076241  3.672188 4.477351 5.427949 22.615398 17.91291
## cvpred      10.4460746 10.718735  3.597516 4.394144 5.367722 23.346932 19.51779
## Ozono       11.0600000 15.060000  1.980000 9.320000 5.730000 26.890000 17.95000
## CV residual  0.6139254  4.341265 -1.617516 4.925856 0.362278  3.543068 -1.56779
##                   107       131       132      140       141        159
## Predicted   20.420556 27.419430  30.32806 7.030755 17.733296 14.5709793
## cvpred      20.765376 29.360311  32.07841 6.726947 17.854866 14.5601028
```

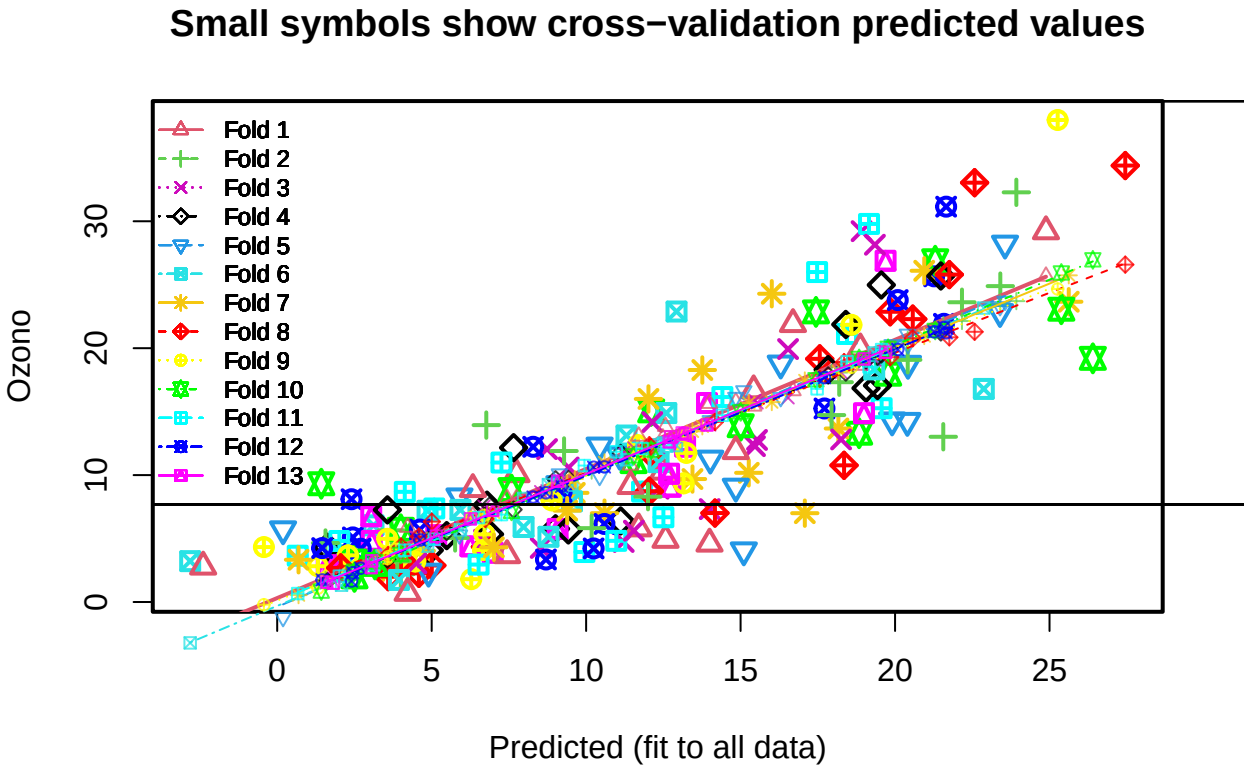
```
## Ozono       13.300000 23.070000  19.20000 8.860000 22.860000 13.8900000
## CV residual -7.465376 -6.290311 -12.87841 2.133053  5.005134 -0.6701028
##                    191        194
## Predicted    3.6672904 3.13496831
## cvpred       3.6338893 2.86135466
## Ozono        3.2300000 2.96000000
## CV residual -0.4038893 0.09864534
##
## Sum of squares = 352.62    Mean square = 23.51    n = 15
##
## fold 11
## Observations in test set: 15
##                     9        45       63         64       74        76
## Predicted    7.934666 10.735292 2.294218  4.1888174 5.712581 17.224874
## cvpred       8.373678 10.900622 2.316682  3.9170082 5.805029 17.309277
## Ozono        3.940000  8.700000 4.810000  3.6500000 8.680000 21.120000
## CV residual -4.433678 -2.200622 2.493318 -0.2670082 2.874971  3.810723
##                    86        89        91       100       150       151
## Predicted    9.015360 13.393692  9.861859 18.418469 18.296607 21.597058
## cvpred       9.413115 13.384316  9.935626 18.486438 17.859104 21.072538
## Ozono        4.820000 16.150000  6.680000 15.270000 26.000000 29.790000
## CV residual -4.593115  2.765684 -3.255626 -3.216438  8.140896  8.717462
##                   164       174       185
## Predicted   5.157344  5.652455  4.939954
## cvpred      4.993325  5.516749  4.722822
## Ozono       7.370000 10.990000  2.950000
## CV residual 2.376675  5.473251 -1.772822
##
## Sum of squares = 284.28    Mean square = 18.95    n = 15
##
## fold 12
## Observations in test set: 15
##                     15       48       61        72        84        85        95
## Predicted    8.388468 4.976956 3.515788  7.423812 23.632733 8.1757579 22.009547
## cvpred       8.303065 4.811886 3.455963  7.295744 23.366795 8.2731883 21.775998
## Ozono        6.150000 8.100000 5.090000 12.230000 31.150000 8.6800000 25.660000
## CV residual -2.153065 3.288114 1.634037  4.934256  7.783205 0.4068117  3.884002
##                    102        116       120       139       147       180
## Predicted   15.0930586 22.8809666 21.216148  9.406916  8.373777 2.746962
## cvpred      15.0861201 22.7061536 21.073878  9.421834  8.414259 2.877515
## Ozono       15.2500000 21.9200000 23.790000  3.350000  4.220000 4.200000
## CV residual  0.1638799 -0.7861536  2.716122 -6.071834 -4.194259 1.322485
##                   196       198
## Predicted   4.6828207 3.4850530
## cvpred      4.7148513 3.4713454
## Ozono       5.7100000 4.2500000
## CV residual 0.9951487 0.7786546
##
## Sum of squares = 184.12    Mean square = 12.27    n = 15
##
## fold 13
## Observations in test set: 15
##                      4         5        12       21        44        47
## Predicted    8.339044  8.848593  7.642695 2.548599 11.843792  9.694734
```

```
## cvpred       8.495788  8.880174  7.826039 2.512257 11.872735  9.535274
## Ozono        3.890000  5.760000  4.390000 2.940000 15.680000 12.670000
## CV residual -4.605788 -3.120174 -3.436039 0.427743  3.807265  3.134726
##                    57        65        81        92       125       145
## Predicted    9.5437635 4.289517 20.258380 4.4007080 16.757972 11.9035076
## cvpred       9.4199987 4.259452 20.197882 4.2845918 16.615503 11.8606816
## Ozono        9.0900000 6.760000 26.890000 5.2700000 14.880000 12.2500000
## CV residual -0.3299987 2.500548  6.692118 0.9854082 -1.735503  0.3893184
##                    158       170       171
## Predicted   10.9711614 3.591985  3.2577747
## cvpred      10.7745478 3.466788  3.2580894
## Ozono       10.1100000 4.820000  2.9000000
## CV residual -0.6645478 1.353212 -0.3580894
##
## Sum of squares = 124.94    Mean square = 8.33    n = 15
##
## Overall (Sum over all 15 folds)
##       ms
## 16.09014
```

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

```
## [1] 4.011251
```

Finalmente, para MC:




```
##
## 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
##
```

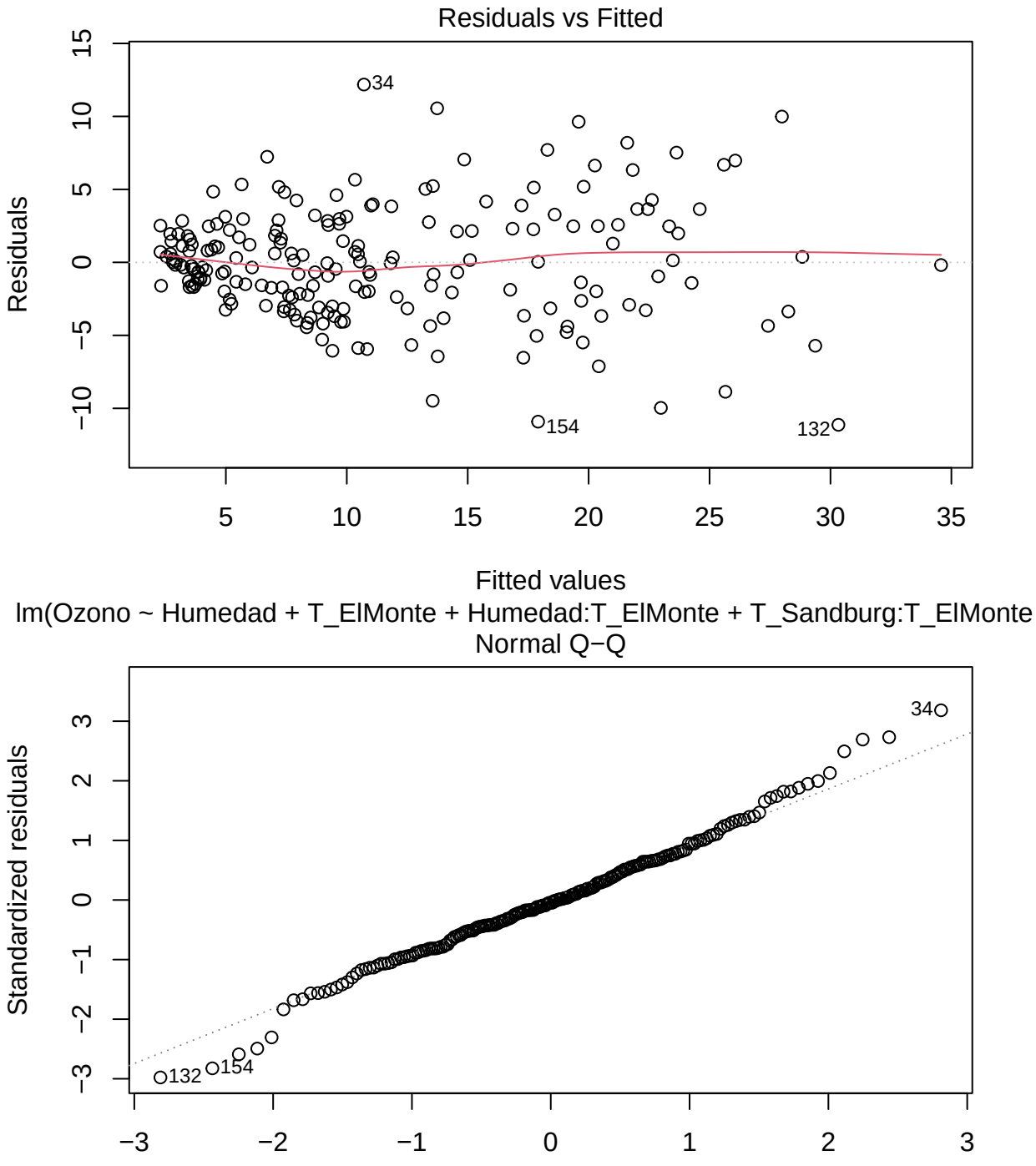
```
## fold 12
## Observations in test set: 15
##                    15       48       61        72        84         85       95
## Predicted   10.592034 2.405465 2.447951  8.284350 21.654136  9.2063365 21.26473
## cvpred      10.643806 1.652116 2.309982  8.272994 21.208759  9.1158536 21.28625
## Ozono        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

## [1] 4.540672
```

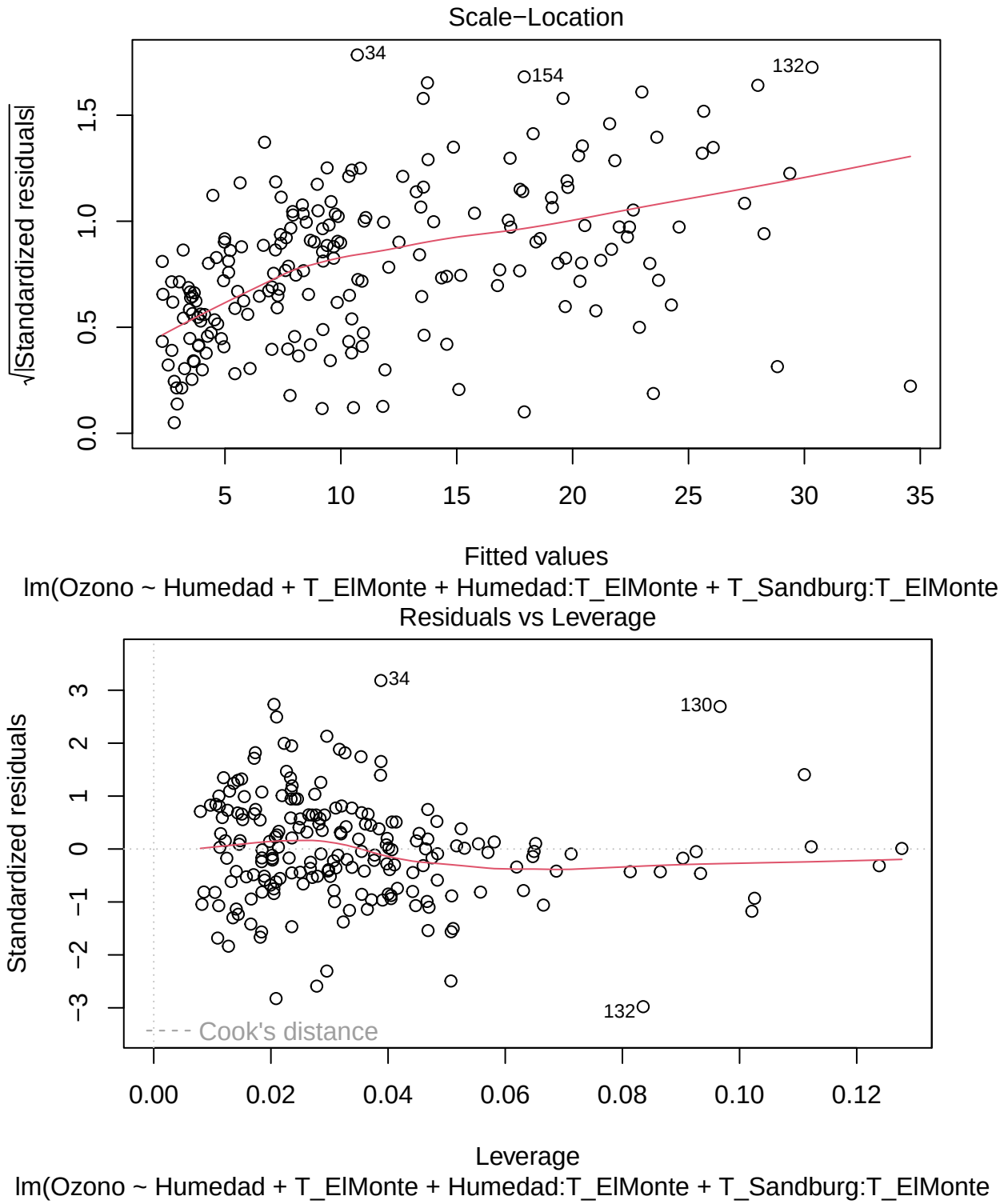
Obtenemos un comportamiento mejor con el MS que con MC, pues tenemos un menor error.

9. Análisis de residuos modelo seleccionado

Para realizar el análisis de los residuos usaremos los residuos estandarizados

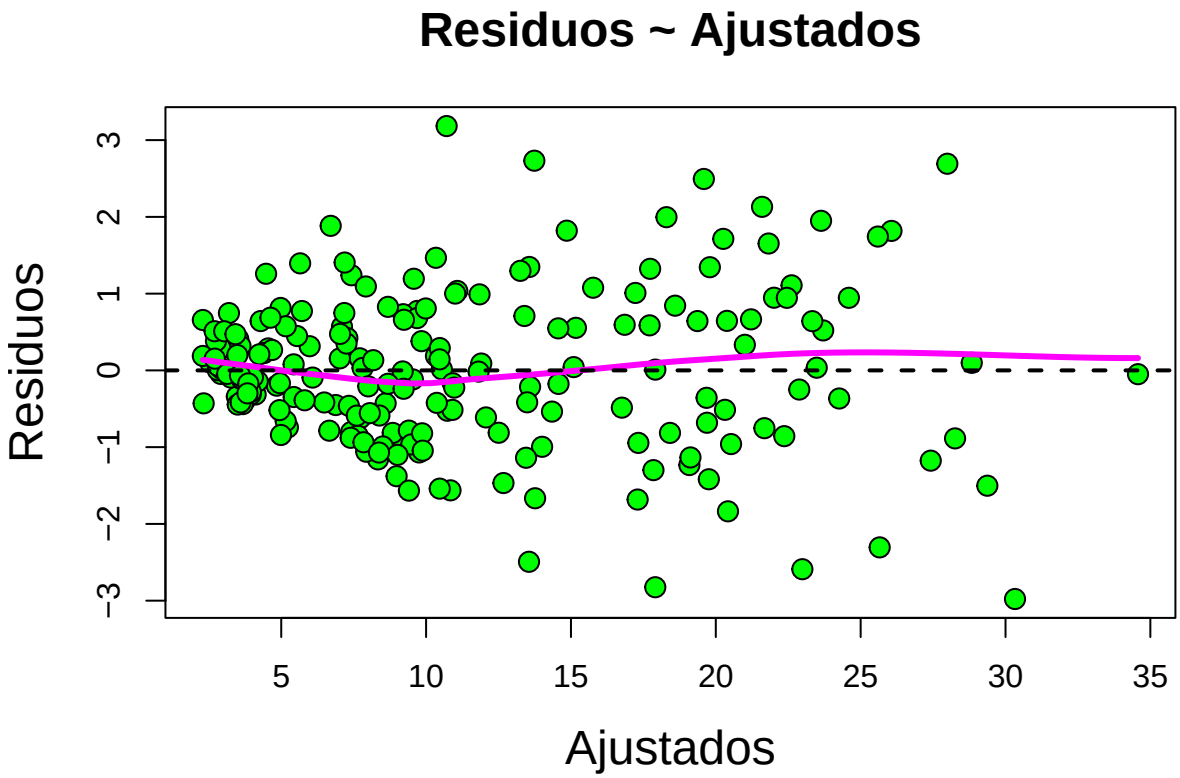
Para comenzar, haremos un análisis inicial utilizando la función plot de R:


Residuals vs Fitted
Residuals
Fitted values
lm(Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte + T_Sandburg:T_ElMonte
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Ozono ~ Humedad + T_ElMonte + Humedad:T_ElMonte + T_Sandburg:T_ElMonte
34
154
132



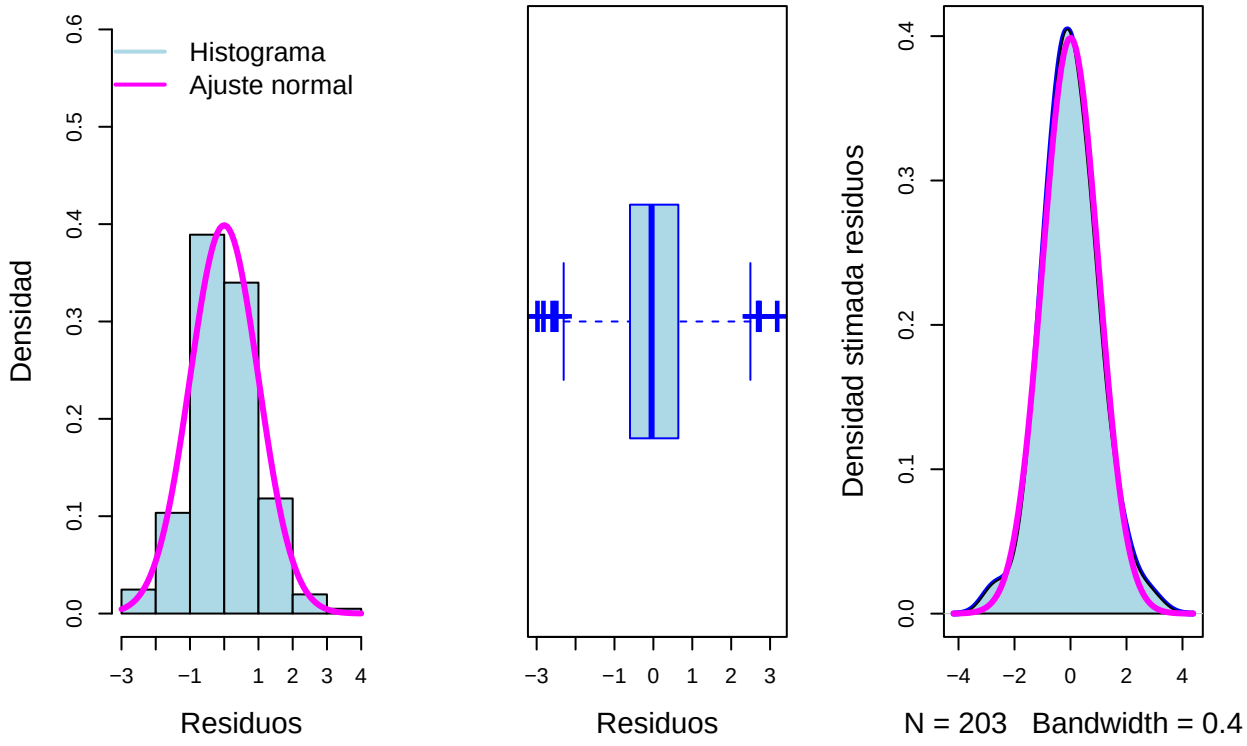
A primera vista, en el primer gráfico vemos linealidad. En el gráfico QQ-plot, que enfrenta los cuantiles teóricos con los residuos estandarizados, vemos que los residuos se apoyan en la linea, a excepción de las colas, por lo que en principio vemos normalidad. Hay tendencia en el gráfico de homoscedasticidad, por lo que decidimos que, en principio, no está presente. No vemos problema en el gráfico distancia de Cook.

Linealidad:



Como podemos observar en el gráfico, podemos llegar a creer que hay linealidad.

Normalidad:



Gráficamente, tal y como indicaba la salida del plot, nuestros datos parecen seguir claramente una distribución

normal.

Analíticamente, aplicaremos los siguientes test:

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  res.est
## D = 0.037929, p-value = 0.6773
```

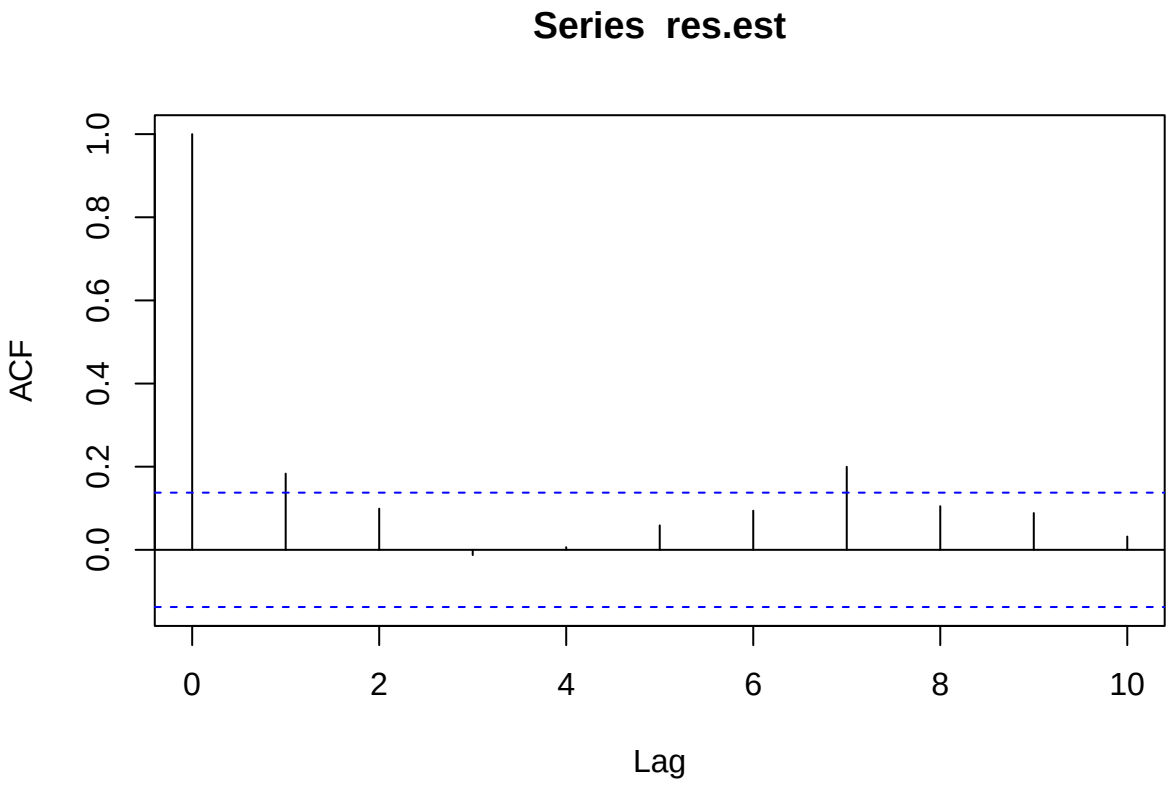
```
##
##  Cramer-von Mises normality test
##
## data:  res.est
## W = 0.077883, p-value = 0.2195
```

```
##
##  Anderson-Darling normality test
##
## data:  res.est
## A = 0.55272, p-value = 0.1524
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.est
## W = 0.98959, p-value = 0.1481
```

Con un 5% de significación, todos los test nos indican que existe normalidad.

Aleatoriedad:



```
## , , 1
##
##              [,1]
##  [1,]  1.00000000
##  [2,]  0.18333147
##  [3,]  0.09903486
##  [4,] -0.01254983
##  [5,]  0.00624498
##  [6,]  0.05882994
##  [7,]  0.09415305
##  [8,]  0.19973724
##  [9,]  0.10497941
## [10,]  0.08848945
## [11,]  0.03196118
```

Como podemos ver en el grafico, no vemos una clara tendencia, por lo que se puede asumir aleatoriedad. No obstante, vamos a aplicar los test adecuados para confirmarlo.

Comenzaremos con la prueba de Ljung-Box:

```
##
##  Box-Ljung test
##
## data:  res.est
## X-squared = 9.7232, df = 5, p-value = 0.08347
```

Los datos son normales, por lo que podemos fiarnos del resultado de este test. Se obtiene un p-valor de 0.08347, lo cual nos lleva a aceptar la hipótesis nula y a afirmar que hay aleatoriedad. Comprobaremos también el resulta de la prueba de rachas para aleatoriedad:

```
##
##  Runs Test
##
## data:  as.factor(sign(res.est))
## Standard Normal = -1.6032, p-value = 0.1089
## alternative hypothesis: two.sided
```

Obtenemos el mismo resultado.

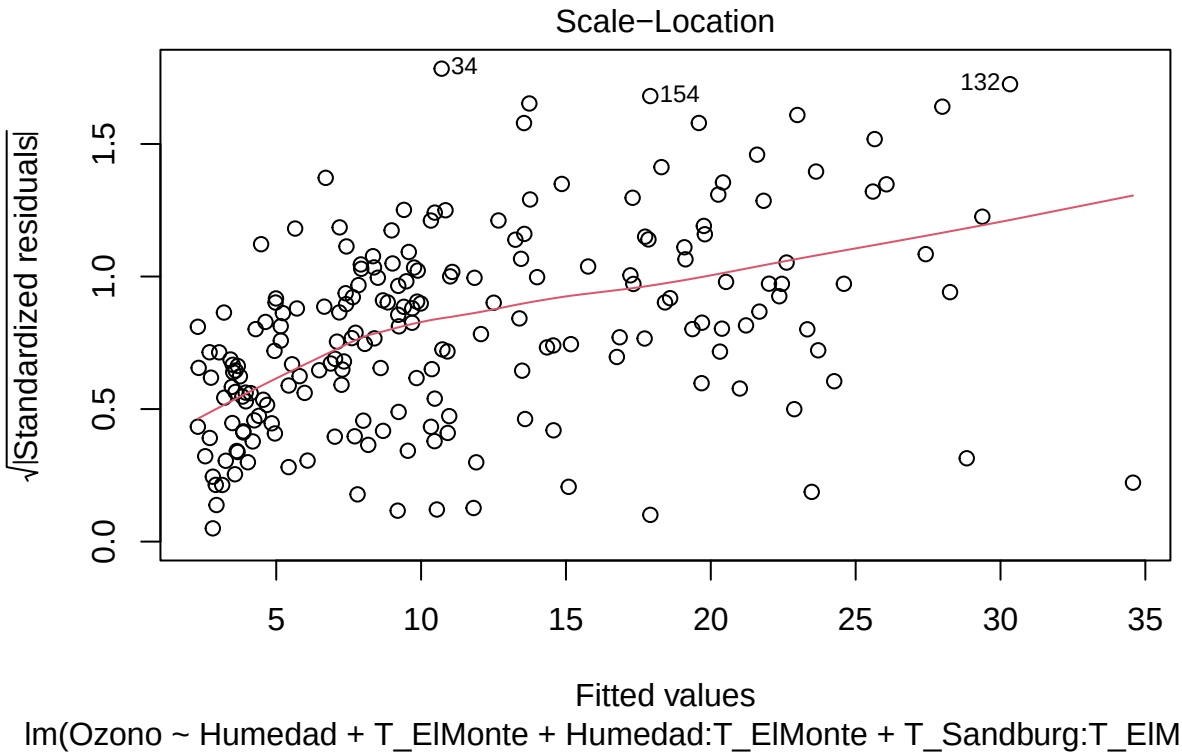
Homoscedasticidad:

El contraste a llevar a cabo es el siguiente: $H0: sigmaˆ2=cte vs H1: sigmaˆ2!=cte

$$\begin{cases} H_0 : \sigma^2 = cte \\ H_1 : \sigma^2 \neq cte \end{cases}$$

Para ello, utilizamos el test de Breusch-Pagan

```
##
##  studentized Breusch-Pagan test
##
## data:  ajuste
## BP = 37.383, df = 6, p-value = 1.483e-06
```

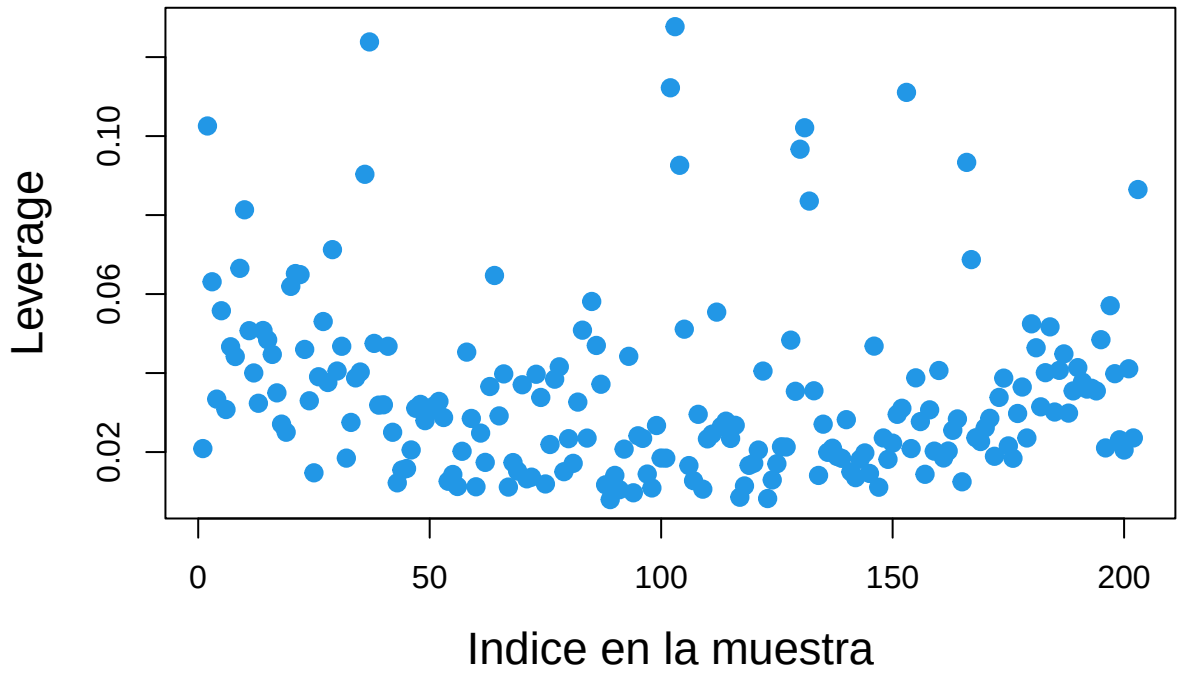


Tanto con el test de Breusch-Pagan cómo con el gráfico de los residuos podemos concluir que se rechaza la hipótesis nula de homoscedasticidad

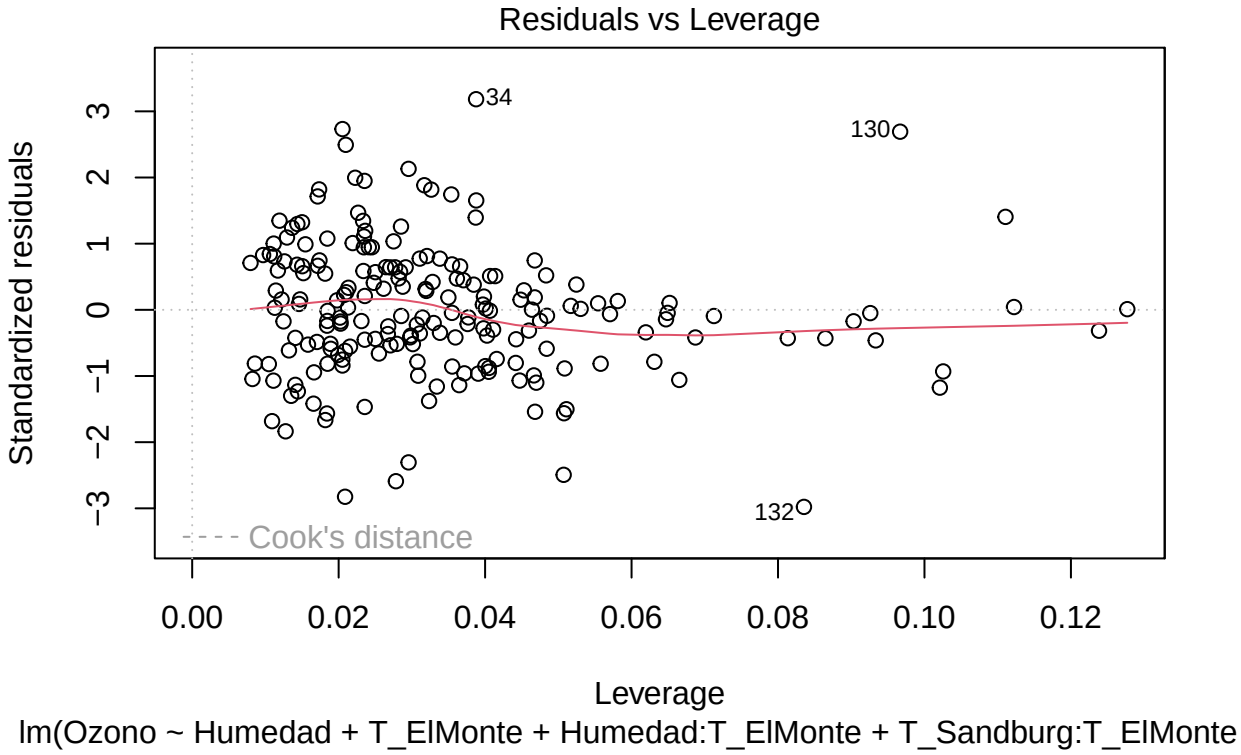
10. Análisis de influencia modelo seleccionado

En este apartado chequearemos si hay datos influyentes en nuestro ajuste. Para ello, comenzamos utilizando la siguiente función de R:

- Leverages: Los leverages nos indican en que tanto nuestros datos son extremos con respecto a las explicativas:



También es útil la salida del siguiente gráfico:



- Distancias de Cook: Los leverages nos indican cómo b¡varía el vector de betas o las predicciones con y sin el i-ésimo dato muestral.

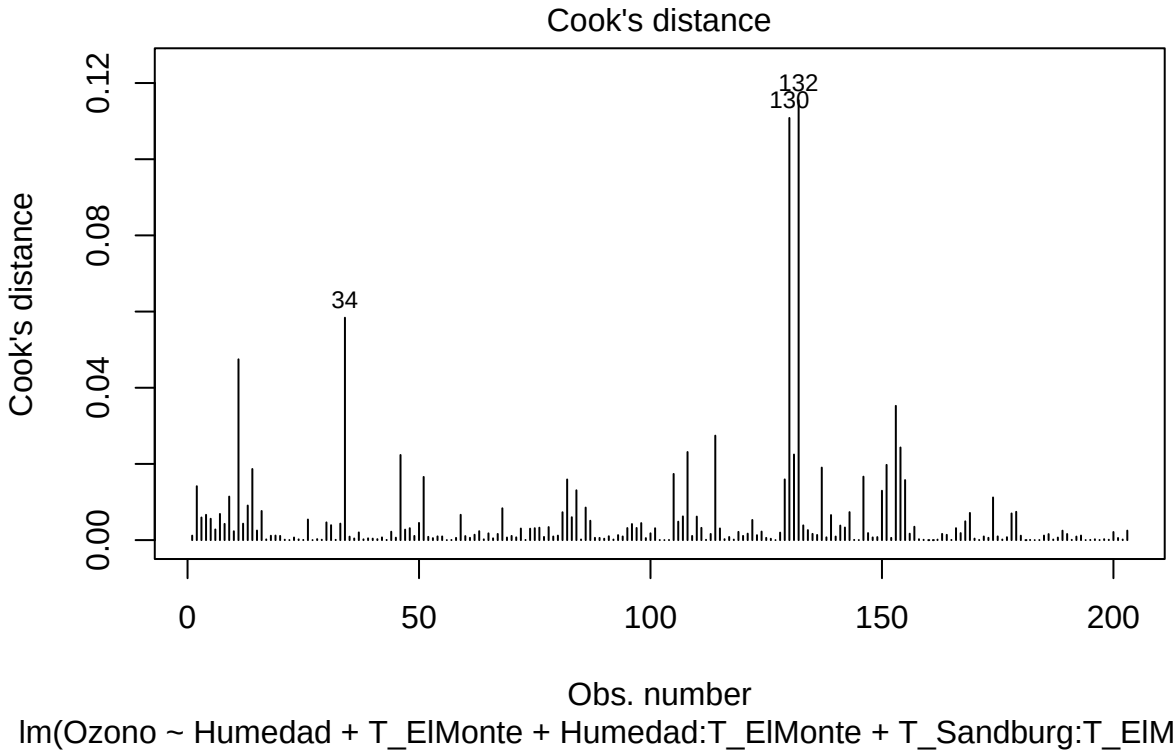
```
##            1            2            3            4            5            6
## 1.173907e-03 1.414800e-02 5.938178e-03 6.636457e-03 5.596333e-03 2.790228e-03
##            7            8            9           10           11           12
## 6.865937e-03 4.253018e-03 1.142178e-02 2.325045e-03 4.745962e-02 4.309192e-03
```

```
##           13           14           15           16           17           18
## 9.073996e-03 1.866031e-02 2.511623e-03 7.640559e-03 1.821862e-04 1.144926e-03
##           19           20           21           22           23           24
## 1.187365e-03 1.093592e-03 1.071644e-04 2.078911e-05 6.774318e-04 1.931992e-04
##           25           26           27           28           29           30
## 5.319520e-05 5.399915e-03 1.722792e-06 2.552825e-04 9.557323e-05 4.648676e-03
##           31           32           33           34           35           36
## 3.908471e-03 7.587138e-05 4.327633e-03 5.838394e-02 9.052071e-04 4.242527e-04
##           37           38           39           40           41           42
## 2.015104e-03 1.972348e-04 4.662536e-04 3.868001e-04 2.474146e-04 7.063377e-04
##           43           44           45           46           47           48
## 4.334842e-05 2.201143e-03 6.341674e-04 2.234587e-02 2.747407e-03 3.130144e-03
##           49           50           51           52           53           54
## 1.087863e-03 4.506696e-03 1.658427e-02 8.613883e-04 5.181692e-04 9.769810e-04
##           55           56           57           58           59           60
## 9.639862e-04 1.641962e-06 4.060719e-05 5.885595e-04 6.643442e-03 1.054106e-03
##           61           62           63           64           65           66
## 6.060446e-04 1.411954e-03 2.339481e-03 2.013557e-04 1.769907e-03 4.643293e-04
##           67           68           69           70           71           72
## 1.611805e-03 8.352021e-03 6.775427e-04 1.100341e-03 7.186564e-04 3.037149e-03
##           73           74           75           76           77           78
## 3.682574e-05 2.991503e-03 3.129646e-03 3.259232e-03 8.277550e-04 3.421846e-03
##           79           80           81           82           83           84
## 9.478082e-04 1.178481e-03 7.316397e-03 1.591854e-02 6.005461e-03 1.307788e-02
##           85           86           87           88           89           90
## 1.561941e-04 8.538289e-03 5.080152e-03 5.974759e-04 5.767156e-04 3.657505e-04
##           91           92           93           94           95           96
## 1.014943e-03 1.533105e-04 1.308565e-03 9.580645e-04 3.164318e-03 4.214804e-03
##           97           98           99          100          101          102
## 3.180048e-03 4.461835e-03 5.254979e-04 1.783621e-03 3.116557e-03 3.288604e-05
##          103          104          105          106          107          108
## 2.164228e-06 3.559966e-05 1.736563e-02 4.842008e-03 6.223634e-03 2.312448e-02
##          109          110          111          112          113          114
## 1.088951e-03 6.169529e-03 3.213739e-03 8.184689e-05 1.617248e-03 2.742445e-02
##          115          116          117          118          119          120
## 3.065587e-03 2.446934e-04 8.140166e-04 1.397079e-04 2.157246e-03 1.103230e-03
##          121          122          123          124          125          126
## 1.694539e-03 5.286089e-03 1.295514e-03 2.239041e-03 5.830683e-04 3.464317e-04
##          127          128          129          130          131          132
## 3.842763e-06 1.968075e-03 1.593528e-02 1.108233e-01 2.246404e-02 1.154964e-01
##          133          134          135          136          137          138
## 3.860116e-03 2.624797e-03 1.634077e-03 1.351396e-03 1.903714e-02 7.252075e-04
##          139          140          141          142          143          144
## 6.572249e-03 9.368462e-04 3.810910e-03 3.304790e-03 7.340312e-03 5.939266e-05
##          145          146          147          148          149          150
## 1.691047e-05 1.665564e-02 1.835747e-03 7.024495e-04 7.916072e-04 1.295675e-02
##          151          152          153          154          155          156
## 1.974708e-02 5.844248e-04 3.524083e-02 2.431194e-02 1.575607e-02 1.681717e-03
##          157          158          159          160          161          162
## 3.503530e-03 2.270991e-04 9.166353e-05 1.120602e-06 6.952754e-07 1.280405e-04
##          163          164          165          166          167          168
## 1.629873e-03 1.380642e-03 5.487699e-05 3.136449e-03 1.840203e-03 4.925476e-03
##          169          170          171          172          173          174
## 7.129511e-03 3.896924e-04 3.630509e-05 9.568095e-04 6.015950e-04 1.119292e-02
```

```
##          175          176          177          178          179          180
## 9.741769e-04 1.534415e-04 7.579215e-04 7.002078e-03 7.425242e-03 1.156844e-03
##          181          182          183          184          185          186
## 4.349962e-08 6.330953e-05 2.166532e-06 2.798610e-05 1.189618e-03 1.577361e-03
##          187          188          189          190          191          192
## 1.566234e-04 6.672106e-04 2.477618e-03 1.602429e-03 7.302912e-05 9.253791e-04
##          193          194          195          196          197          198
## 1.193476e-03 1.094690e-05 5.831336e-05 2.172497e-04 3.599758e-05 2.371239e-04
##          199          200          201          202          203
## 9.713963e-05 2.118885e-03 5.481748e-04 1.509543e-04 2.493401e-03
```

Se recomienda examinar los datos cuya distancia de cook supera 4/(n-k-1), con k el número de explicativas. En nuestro caso, k = 5

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



En el gráfico observamos que hay tres observaciones {11, 34, 130} con una distancia de Cook elevada, mayor que la del resto de observaciones.

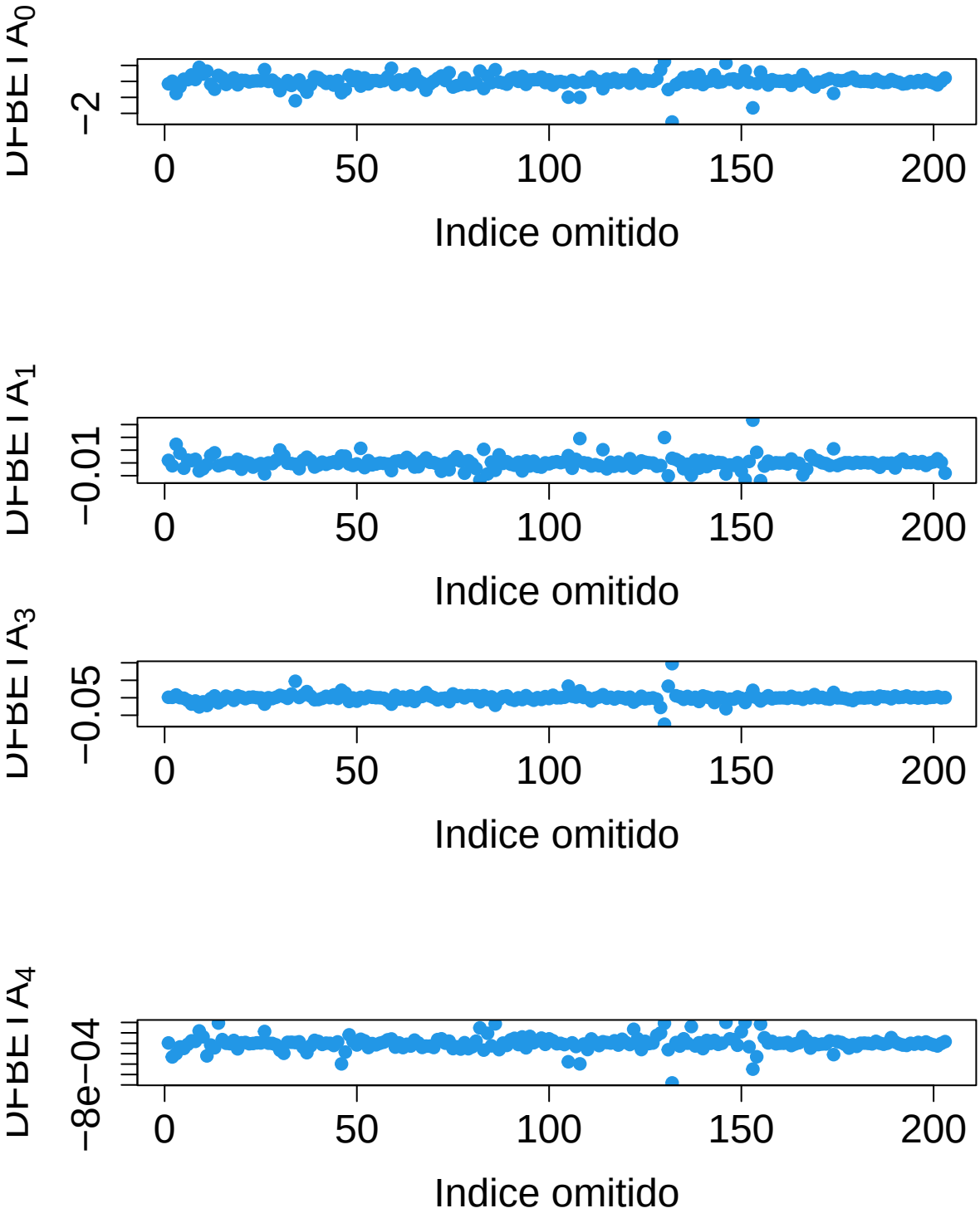
- DFFITs Los DFFITs nos indican como varía la predicción del i-ésimo dato con y sin él.

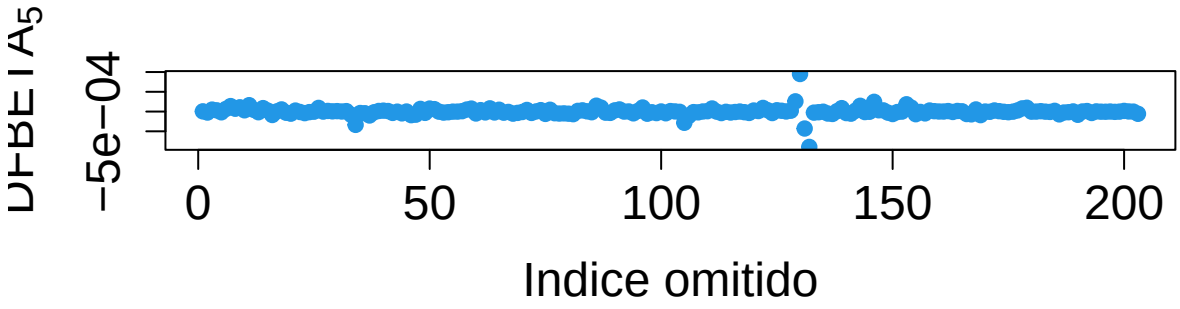
Se consideran influyentes los datos que cumplan $|\text{DFFITs}| > 2\sqrt{\frac{k+1}{n-k-1}}$:

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

De nuevo, ninguna observación sobrepasa este valor.

- DFBETAs Los DFBETA nos indican en qué medida varía la estimación del j-ésimo coeficiente con y sin el i-ésimo dato muestral.


DFBETA0
−2
0
50
100
150
200
Indice omitido
DFBETA1
−0.01
0
50
100
150
200
Indice omitido
DFBETA3
−0.05
0
50
100
150
200
Indice omitido
DFBETA4
−8e−04
0
50
100
150
200
Indice omitido



También se pueden utilizar los DFBETA estandarizados:

Se consideran influyentes los datos que cumplen |DFBETAs| > 2*sqrt(n):

```
##      row col
```

De nuevo vemos que no hay datos que cumplan esta condición.

11. 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 para cada mes con todas las demás variables fijadas en su valor medio .

```
##          fit       lwr       upr
## 1  12.594950 11.403635 13.786266
## 2  12.185085 11.140582 13.229589
## 3  11.775220 10.866854 12.683587
## 4  11.365356 10.576902 12.153809
## 5  10.955491 10.262256 11.648725
## 6  10.545626  9.911688 11.179563
## 7  10.135761  9.514821 10.756701
## 8   9.725896  9.068901 10.382891
## 9   9.316031  8.581112 10.050950
## 10  8.906166  8.062984  9.749348
## 11  8.496301  7.524605  9.467997
## 12  8.086436  6.972965  9.199907
```

```
##          fit       lwr      upr
## 1  12.594950 4.8054571 20.38444
## 2  12.185085 4.4166905 19.95348
## 3  11.775220 4.0239560 19.52648
## 4  11.365356 3.6272272 19.10348
## 5  10.955491 3.2264839 18.68450
## 6  10.545626 2.8217117 18.26954
## 7  10.135761 2.4129027 17.85862
## 8   9.725896 2.0000553 17.45174
## 9   9.316031 1.5831740 17.04889
## 10  8.906166 1.1622700 16.65006
## 11  8.496301 0.7373603 16.25524
## 12  8.086436 0.3084681 15.86440
```

Regresión Logística

- Antes de empezar, cargamos los 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()`.

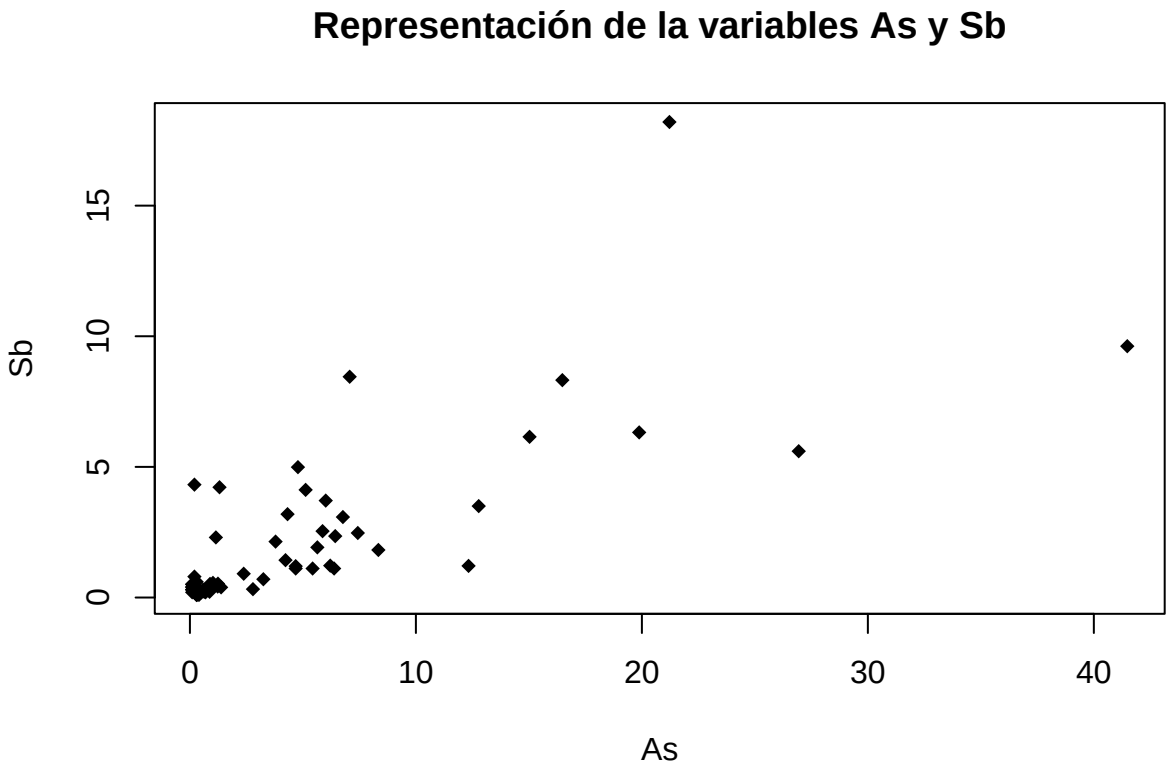
```
## 'data.frame':    64 obs. of  4 variables:
##  $ As        : num  6.77 15.03 6.43 0.1 0.1 ...
##  $ Sb        : num  3.08 6.15 2.35 0.3 0.3 9.62 0.51 3.71 4.32 0.8 ...
##  $ Corredor  : int  1 1 1 0 0 1 0 1 0 0 ...
##  $ Proximidad: int  1 1 1 0 0 1 0 1 0 0 ...
```

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

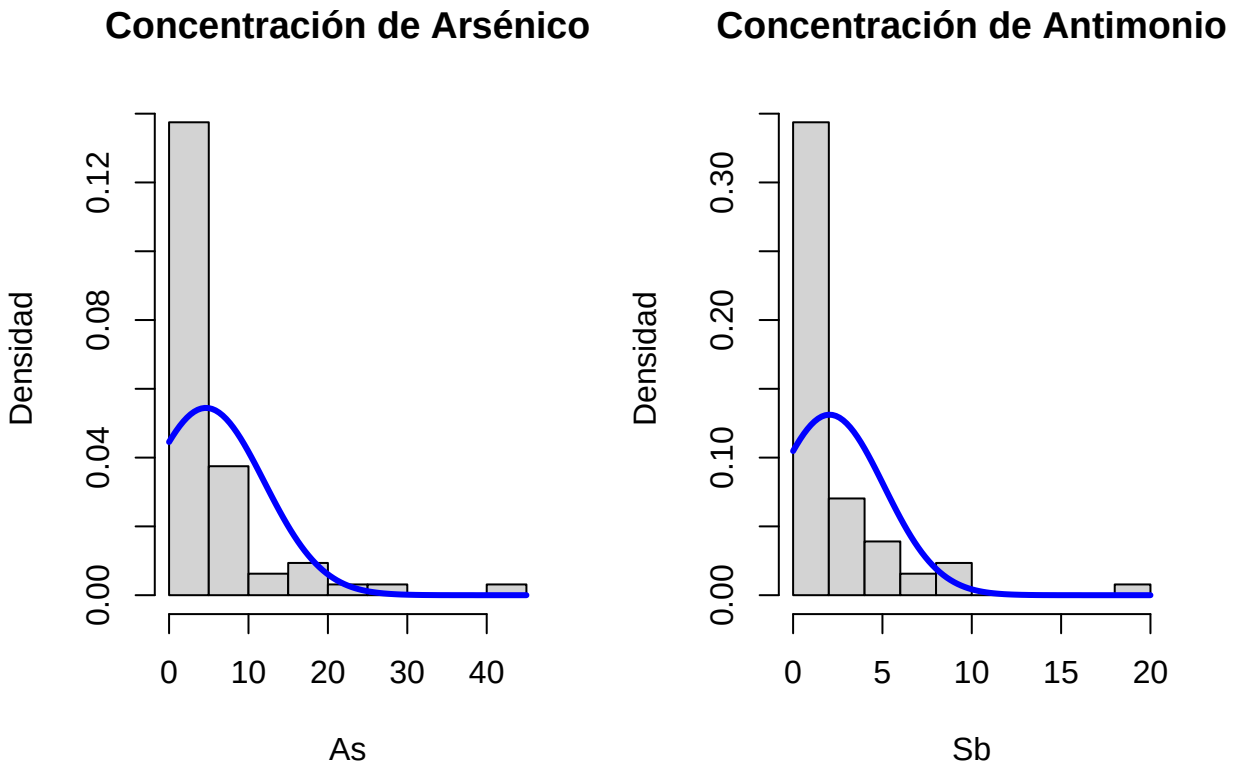
- `As`: Nivel de concentración de arsénico en la muestra de agua. (numérica)
- `Sb`: Nivel de concentración de antimonio en la muestra de agua. (numérica)
- `Corredor`: Variable binaria indicando si la zona muestreada está (1) o no está (0) en alguno de los corredores delimitados por las lineas sobre 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.

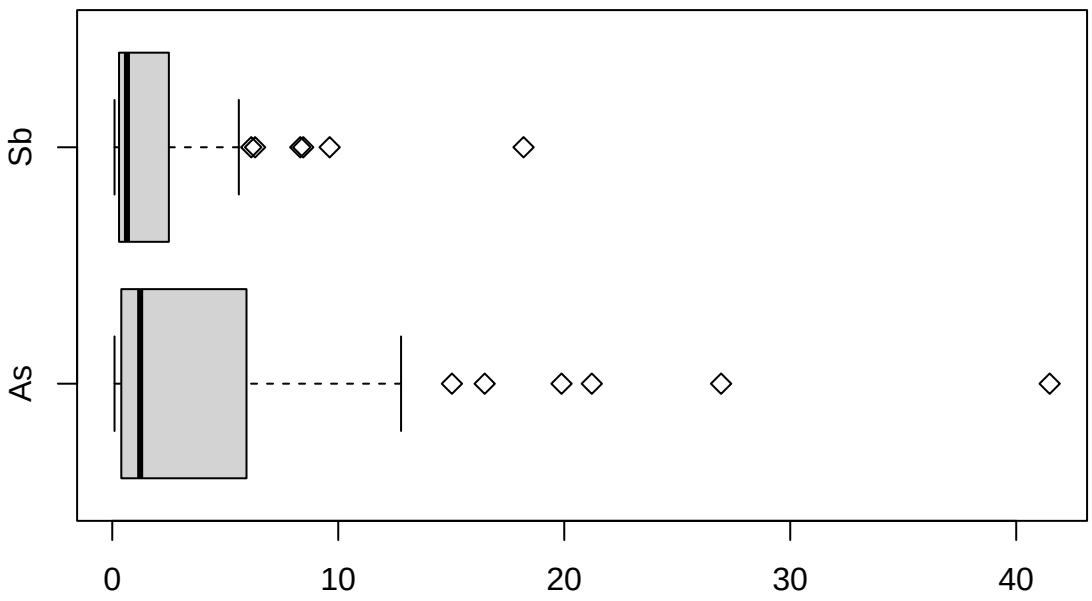
Con la salida de `summary()` y graficando `As` frente a `Sb` podemos ver que, basándonos en la diferencia entre las medias y las medianas, las variables numéricas se concentran en valores bajos, aunque deben de existir registros con valores relativamente altos:

```
##        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
```



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





Distribución de la variable `Proximidad`:

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

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

Distribución de la variable `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:

```
##            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 $\vec{X}^t = (X_1, \ldots, X_k)$ 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.

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

Por tanto, nuestro modelo logístico quedaría de la forma

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

Tal que

$$\eta = \beta_0 + \beta_1 As + \beta_2 Sb + \tau I(Corredor = 1) \tag{4}$$

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

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

3. Interpretación del modelo

Para una mejor interpretación del modelo, podemos definir el $\mathbf{odds}_i$ de manera que

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

Este es un modelo multiplicativo, en el cual e^{β_0} es la respuesta cuando $\vec{X_i} = \vec{0}$, mientras que e^{β_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 $X = 0$:

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

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

$$\ln(\frac{p_i}{1 - p_i}) = \eta = \vec{\beta}^t \vec{X_i} \tag{8}$$

Los cuales denominaremos como $\mathbf{logit}_i$. 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. Análisis de multicolinealidad

Debemos analizar si estamos ante un caso de multicolinealidad. Si así fuera, las estimaciones de los parámetros no serían correctos, y nuestro modelo solo serviría para predecir, no para explicar el comportamiento de la respuesta.

Utilizaremos los factores de inflacción de la varianza generalizada, para ver si nos encontramos con variables correlacionadas:

```
##        As        Sb Corredor1
##    1.5773    2.2937    1.8728
```

Los factores de inflacción de la varianza son todos menores que 10, por lo que no estamos ante un caso de multicolinealidad.

5. Selección del modelo

A pesar de no tener multicolinealidad en los datos, decidimos hacer una selección de variables, debido a la no significación de todas las variables.

Para ello, decidimos utilizar un método de selección exhaustiva con el BIC, ya que esta medida de selección de modelos 'castiga' a modelos con un número elevado de variables:

```
##    Intercept    As    Sb Corredor logLikelihood      BIC
## 0       TRUE FALSE FALSE    FALSE    -43.860109 87.72022
## 1       TRUE  TRUE FALSE    FALSE    -11.301429 26.76174
## 2*      TRUE  TRUE  TRUE    FALSE     -9.152897 26.62356
## 3       TRUE  TRUE  TRUE     TRUE     -7.097155 26.67096
```

Por lo tanto, definimos el ajuste sin corredor y vemos la significación del resto de las variables:

```
##
## Call:
## glm(formula = Proximidad ~ As + Sb, family = "binomial", data = Oro)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.02141  -0.19496  -0.14513   0.06255   2.60217
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.9664     1.3675  -3.632 0.000281 ***
## As            1.2490     0.3777   3.307 0.000943 ***
## Sb            0.9235     0.4486   2.059 0.039518 *
## ---
## 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: 18.306  on 61  degrees of freedom
## AIC: 24.306
##
## Number of Fisher Scoring iterations: 8
```

6. Posible Interacción

Debido a la posible necesidad de interacción, decidimos probar si un modelo que incluya interacción es mejor que nuestro modelo completo.

Comenzamos definiendo este modelo, con todas las interacciones posibles:

```
##
## Call:
## glm(formula = Proximidad ~ As + Sb + Corredor + As:Sb + As:Corredor +
##     Sb:Corredor + As:Sb:Corredor, family = binomial, data = Oro)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.9714   0.0000   0.0000   0.0000   1.9345
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -8.939  34483.934   0.000    1.000
## As                 -47.382 105299.858   0.000    1.000
## Sb                 -33.817 196896.288   0.000    1.000
## Corredor1            9.617  34483.934   0.000    1.000
## As:Sb               47.999  60183.576   0.001    0.999
## As:Corredor1        46.489 105299.858   0.000    1.000
## Sb:Corredor1        26.827 196896.289   0.000    1.000
## As:Sb:Corredor1    -44.627  60183.576  -0.001    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87.7202  on 63  degrees of freedom
## Residual deviance:  7.5068  on 56  degrees of freedom
## AIC: 23.507
##
## Number of Fisher Scoring iterations: 21
```

Ningún coeficiente es significativo. No obstante, creemos que esto puede deberse a la presencia de multicolinealidad, así que decidimos hacer una selección de variables: La haremos siguiendo el mismo método que en el apartado anterior:

```
## Start:  AIC=89.72
## Proximidad ~ 1
```

```
##            Df Deviance    AIC
## + As        1   22.603 26.603
## + Sb        1   45.332 49.332
## + Corredor  1   45.848 49.848
## <none>          87.720 89.720
```

```
##
## Step:  AIC=26.6
## Proximidad ~ As
```

```
##            Df Deviance    AIC
## + Sb        1   18.306 24.306
## + Corredor  1   19.990 25.990
## <none>          22.603 26.603
```

```
##
## Step:  AIC=24.31
## Proximidad ~ As + Sb
```

```
##            Df Deviance    AIC
## + Corredor  1   14.194 22.194
## <none>          18.306 24.306
## + As:Sb     1   17.249 25.249
```

```
##
## Step:  AIC=22.19
## Proximidad ~ As + Sb + Corredor
```

```
##               Df Deviance    AIC
## <none>             14.194 22.194
```

```
## + Sb:Corredor  1   12.253 22.253
## + As:Sb        1   12.688 22.688
## + As:Corredor  1   14.137 24.137
```

```
##
## Call:  glm(formula = Proximidad ~ As + Sb + Corredor, family = "binomial",
##     data = Oro)
##
## Coefficients:
## (Intercept)           As           Sb    Corredor1
##      -7.610        1.205        1.421        3.197
##
## Degrees of Freedom: 63 Total (i.e. Null);  60 Residual
## Null Deviance:       87.72
## Residual Deviance: 14.19     AIC: 22.19
```

Finalmente, vemos que en este caso, la interacción de las variables no aporta nada a nuestro ajuste.

Definimos el ajuste con el que nos quedamos finalmente:

7. Inferencia

Empezamos la inferencia haciendo los intervalos de confianza para los parámetros. Haremos los intervalos basados en las sd de las pruebas de Wald y en los cuantiles de una normal:

```
##                   2.5 %    97.5 %
## (Intercept) -7.64658528 -2.286183
## As           0.50875493  1.989343
## Sb           0.04431076  1.802633
```

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

$$ln(\frac{\hat{p}}{1-\hat{p}}) = \hat{\eta} = -4.9664 + 1.2490As + 0.9235Sb \tag{9}$$

Para la interpretación de los coeficientes del modelo ajustado utilizaremos los odds, que calcularemos a partir de los valores que devuelve el `summary()` del ajuste:

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -4.9663844  1.3674746 -3.631793 0.0002814590
## As           1.2490491  0.3777080  3.306917 0.0009432888
## Sb           0.9234717  0.4485597  2.058749 0.0395183372
```

Aquí podemos observar que el odds de las respuestas cuando $\beta_j = 0$, para $j \in \{1, \ldots, k\}$, tiene un valor de $e^{-4.9663844} = 0.006968297$, que indica que el cociente $\frac{p}{1-p}$ tiene una probabilidad de 0.00697:1 de estar próximo a un yacimiento de oro cuando la concentración de Arsénico y antimonio es nula.

En cuanto a los coeficientes de las variables `As` y `Sb`, indican un incremento multiplicativo del odds de $e^{1.2490491} = 3.487025$ y $e^{0.9234717} = 2.518017$ respectivamente, cuando el resto de las variables se mantiene constante.

O lo que es lo mismo, el valor de los logits cuando $\beta_j = 0$, para $j \in \{1, \ldots, k\}$ es de -4.9663844, que un incremento de una unidad en `As` representa un cambio en los logits de 1.2490491 y que un incremento de una unidad en `Sb` representa un cambio en los logits de 0.9234717 (manteniendo) el resto de las variables constantes.

8. Estimación media y probabilidad condicionada

Haremos los intervalos de confianza y de probabilidad manteniendo las dos variables en su media.

Primero, guardamos los nuevos datos en un data.frame:

Utilizamos predict para obtener la predicción estimada:

```
##         As       Sb p_est_proximidad
## 1 4.644844 2.039062        0.9380961
```

Para obtener los intervalos de confianza para estas predicciones, utilizaremos la siguiente función proporcionada en el Script de R Logísitica:

La aplicamos del siguiente modo:

```
##           p        LI        LS
## 1 0.9380961 0.6190406 0.9929738
```

9. Bondad del ajuste

Test de razón de verosimilitudes

Empezamos la inferencia con el test de razón de verosimilitudes, que chequea si la diferencia entre la deviance de nuestro modelo y la deviance nula es muy elevada.

```
## [1] 8.449853e-16
```

El test resulta significativo, por lo que tenemos que la diferencia es elevada.

R^2 de McFadden

Seguimos con el R^2 de McFadden, que se entiende como la verosimilitud explicada por el modelo respecto a la verosimilitud del modelo nulo. Se calcula del siguiente modo:

```
## [1] 0.7913161
```

Vemos que es bastante elevado (79.13%), por lo que estamos ante un buen ajuste.

Prueba de Hosmer-Lemeshown

Seguimos con la prueba de Hosmer-Lemeshown. Esta prueba, tras ordenar las observaciones en orden creciente según su probabilidad estimada, divide la muestra en G grupos homogéneos entre sí para aplicar una prueba basada en la Chi-Cuadrado de Pearson, que compara las frecuencias observadas de 1's con las frecuencias esperadas.

Esta prueba toma por defecto 10 grupos:

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  Proximidad, fitted(ajuste)
## X-squared = 2.9153, df = 8, p-value = 0.9396
```

Resulta un p-valor muy elevado, por lo que aceptamos la hipótesis nula y concluímos que nuestro modelo se ajusta a la realidad.

No obstante, probaremos con distintos números de grupos para ver si el resultado converge, teniendo en cuenta que el número de grupos tiene que ser mayor que el de explicativas. Probaremos con los siguientes números de grupos:

```
##
##  Hosmer and Lemeshow test (binary model)
```

```
##
## data:  Proximidad, fitted(ajuste)
## X-squared = 0.77131, df = 3, p-value = 0.8563
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  Proximidad, fitted(ajuste)
## X-squared = 10.756, df = 18, p-value = 0.9044
```

Obtenemos de nuevo un p-valor muy elevado, por lo que volvemos a concluír que nuestro modelo se ajusta a la realidad.

Matriz de confusión

La matriz de conclusión nos muestra de una forma visual cómo de bien predice nuestro ajuste, mostrando el número de predicciones correctas y falsas. Estableceremos como umbral de decisión entre éxito y fracaso $\hat{p} = 0.5$, el umbral natural. Así, las observaciones con $\hat{p} > 0.5$ serán éxitos, y en caso contrario se considerarán fracasos. Obtenemos la matriz del siguiente modo:

```
##     Oro$Proximidad
## pred  0  1
##    0 35  2
##    1  1 26
```

Ahora ya podemos calcular la especificidad, la sensibilidad y la tasa de clasificación correcta o precisión de nuestro ajuste:

ESPECIFICIDAD: Proporción de no admitidos bien predichos:

```
## [1] 97.22222
```

Porcentaje muy elevado, por lo que tenemos una muy buena capacidad de predecir correctamente los que no serán admitidos

SENSIBILIDAD: Proporción de admitidos bien predichos:

```
## [1] 92.85714
```

Porcentaje muy elevado, por lo que tenemos una muy buena capacidad de predecir correctamente los que serán admitidos Precisión o tasa de clasificación correcta (TCC):

```
## [1] 95.3125
```

De nuevo, porcentaje muy elevado, por lo que nuestro ajuste es muy bueno a la hora de clasificar individuos.

Error cuadrático medio

Debido a que estas medidas pueden estar sesgadas, decidimos medir el error cuadrático medio de predicción por validación cruzada. Lo hacemos del siguiente modo:

```
## [1] 0.0659519
```

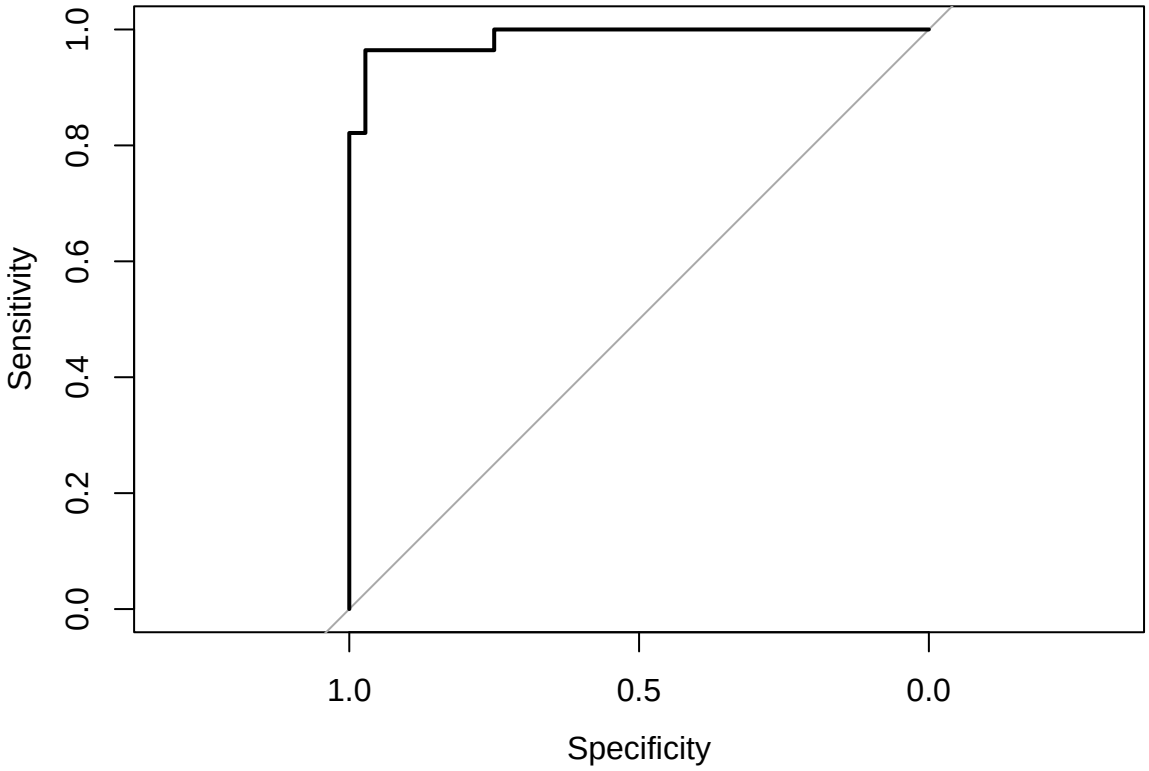
A partir de este error, podemos calcular de nuevo la tasa de clasificación correcta:

```
## [1] 0.9340481
```

Vemos que empeora un poco, pero seguimos teniendo un muy buen resultado.

Cómputo de la curva ROC y del AUC

La curva de ROC nos muestra un gráfico que enfrenta la especificidas y la sensibilidad del ajuste. Para distintos puntos de corte, calcula la especificidad y sensibilidad, y las compara. Hay que tener en cuenta que un aumento de sensibilidad conlleva a una disminución de la especificidad y viceversa. Calculamos la curva del siguiente modo:



El peor ajuste caería en la línea, y el mejor ajuste pasaría por el punto (1,1). Como se puede ver en nuestro gráfico, pasa por el punto (1,1), por lo que, como ya vimos en el apartado anterior, estamos ante un ajuste prácticamente perfecto para clasificar.

Para calcular la capacidad predictiva, calculamos el área bajo la curva (AUC). Buscamos una curva con un área elevada. La calculamos del siguiente modo:

```
## Area under the curve: 0.9871
```

Obtenemos un área muy elevada, por lo que se vuelve a confirmar que la capacidad predictiva de nuestro ajuste es muy buena, y estamos ante un muy buen modelo. ## **10.** Análisis de residuos

El modelo de regresión logísitica tiene 3 hipótesis estructurales: 1) La linealidad de los Logits. 2) La independencia de las n observaciones. 3) La respuesta Y debe ser binaria.

Tal y como sucede en regresión lineal, podemos utilizar los residuos para chequear las hipótesis estructurales. No obstante, debemos tener en cuenta que en regresión logística existen dos tipos de residuos, con fines distintos.

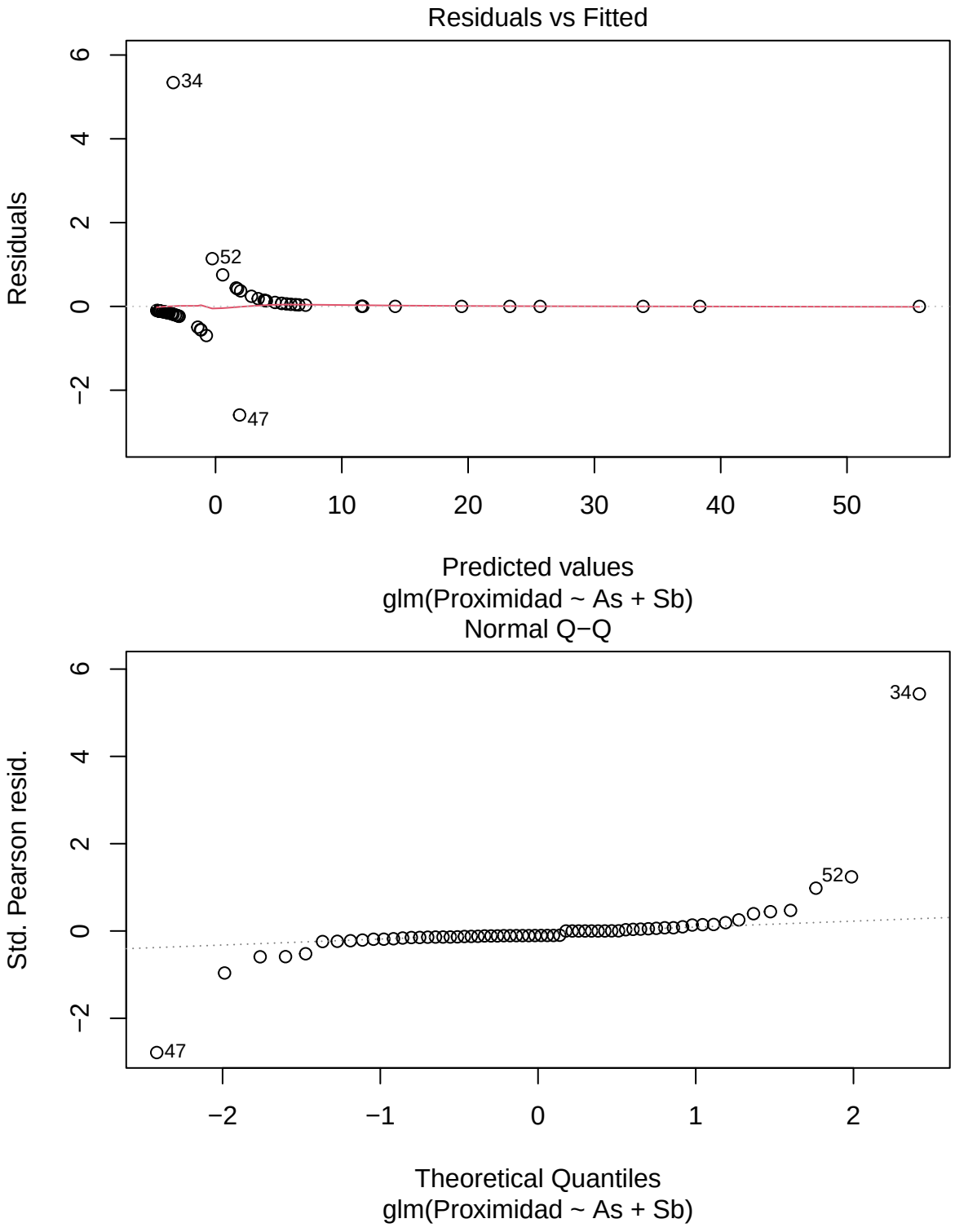
Obtención residuos de Pearson:

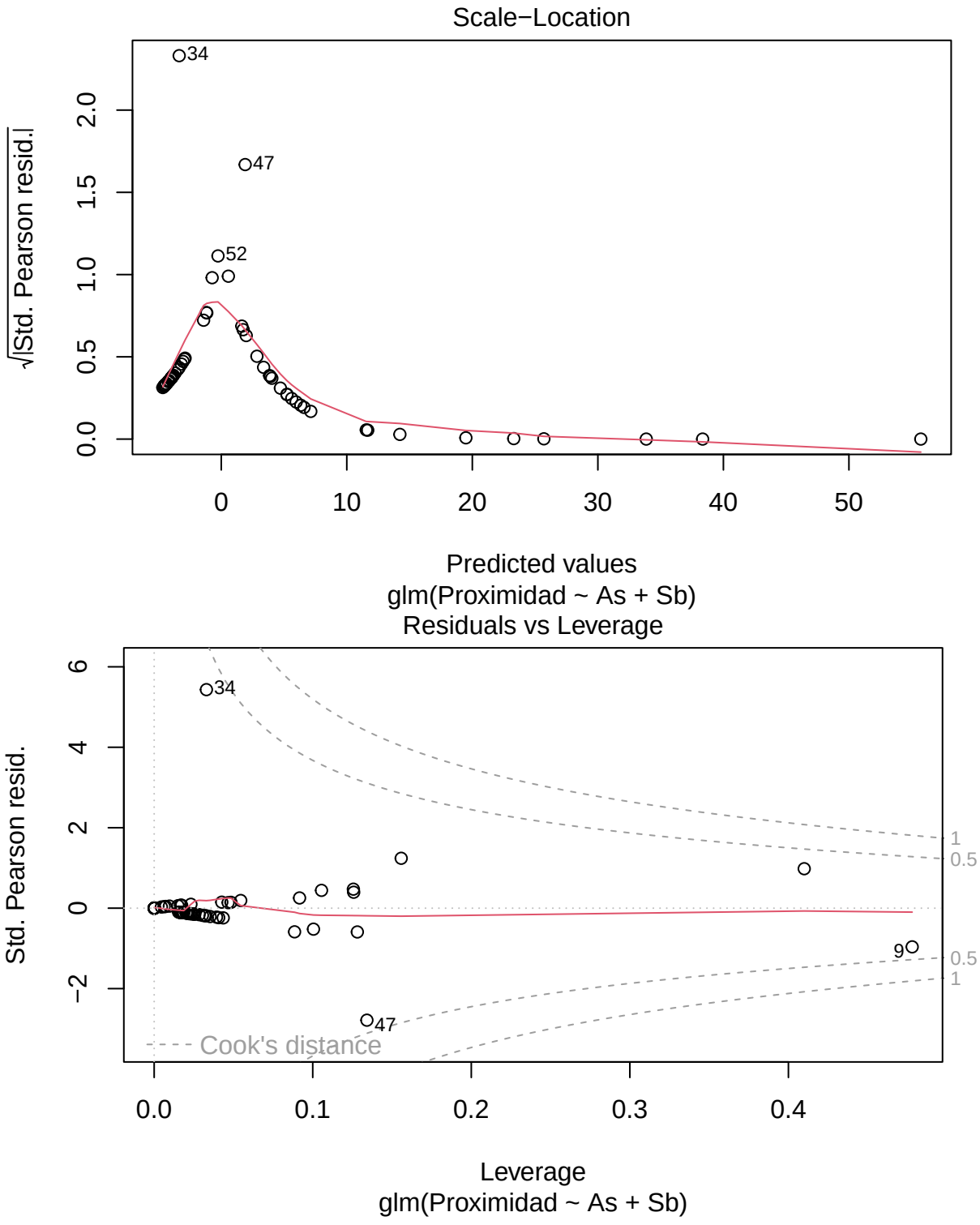
Obtención residuos de la Deviance:

Los estandarizamos: Residuos Pearson estandarizados:

Residuos deviance estandarizados:

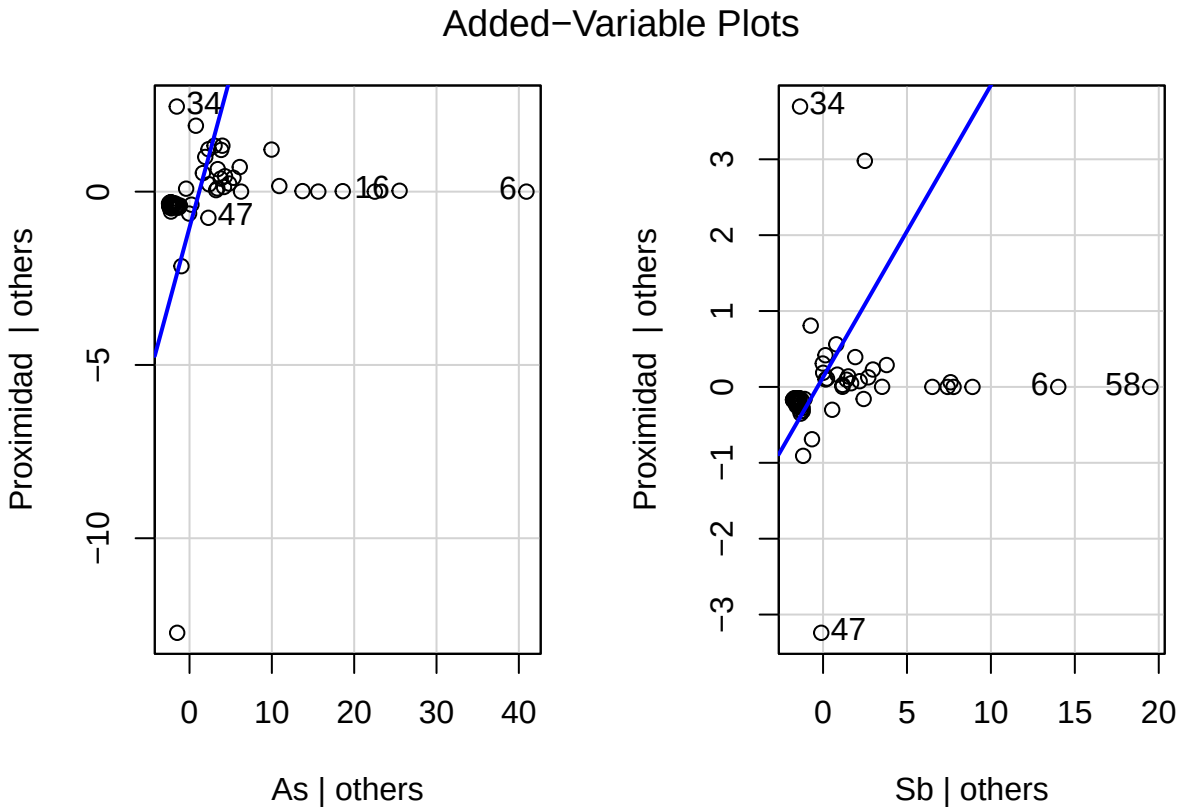
Obtenemos los gráficos de residuos:


Residuals vs Fitted
Residuals
Predicted values
glm(Proximidad ~ As + Sb)
Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(Proximidad ~ As + Sb)
34
52
47



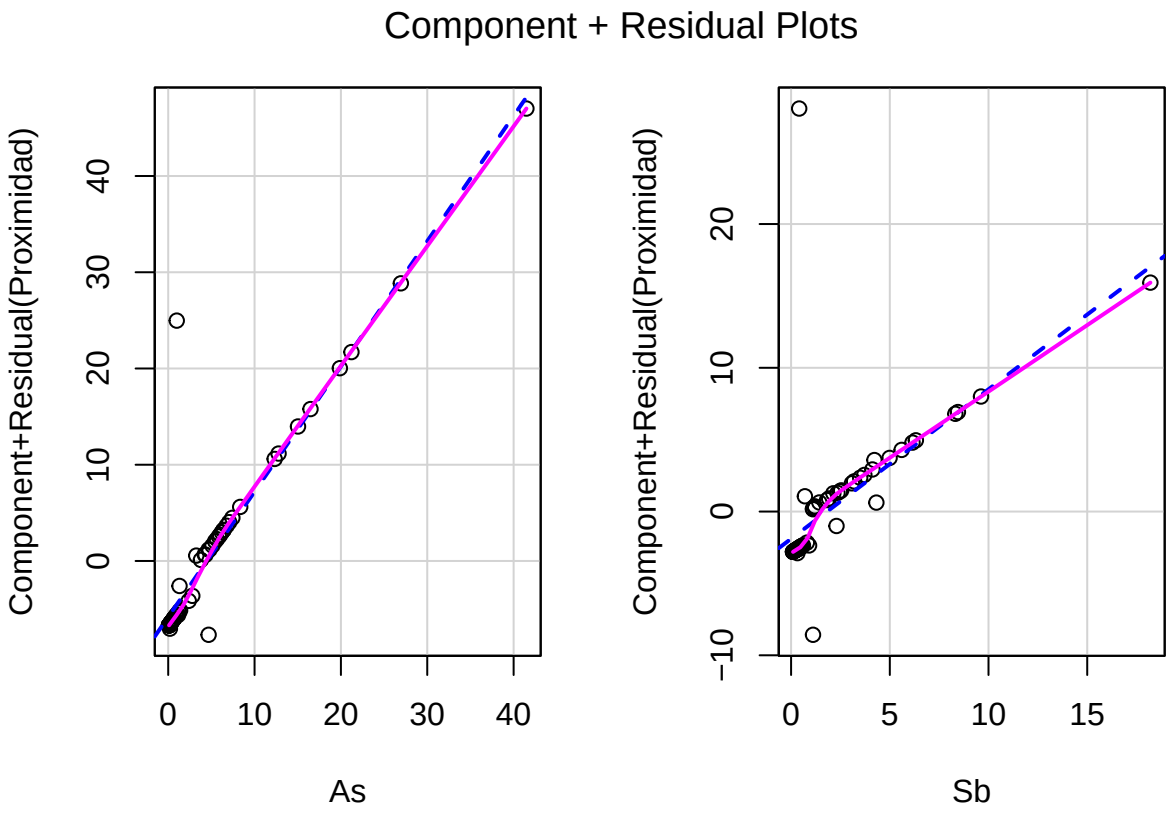
La función plot de R enfrenta los residuos estandarizados de Pearson con los logits del ajuste. Este tipo de residuo es útil simplemente para chequear la normalidad que, en este caso, evidentemente no está presente, como se aprecia en el segundo gráfico de la salida.

Para chequear la linealidad, se utilizan los residuos del segundo tipo, es decir, los de la deviance, del siguiente modo:



Podemos ver que prácticamente los datos están en torno a 0. El problema con las rectas es que, debido a la presencia de muchos datos entre 0 y 10, su pendiente varía mucho.

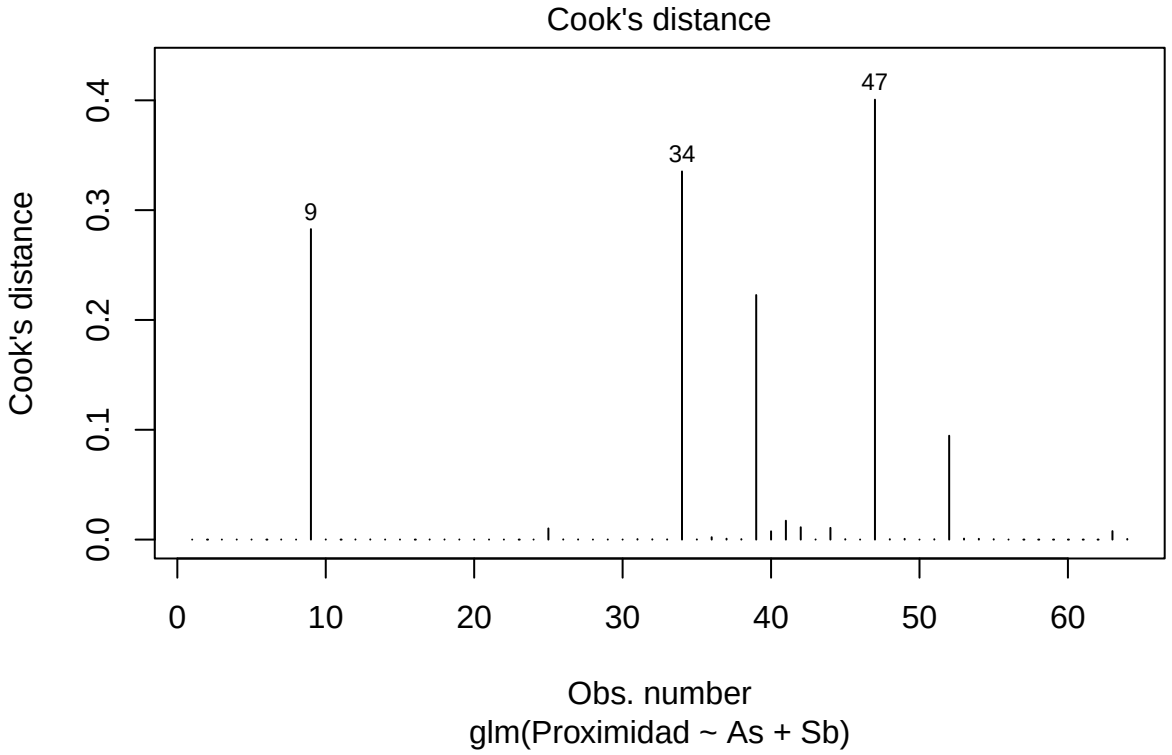
También podemos hacer gráficos de residuos parciales, para ver si la falta de linealidad es achacable a alguna variable concreta:



11. Análisis de influencia

Finalmente, los residuos de Pearson también se pueden utilizar para el análisis de influencia.

Para ver el gráfico de la distancia de Cook, se ejecuta el siguiente comando:



Vemos 3 observaciones con una distancia de Cook mayor que el resto de observaciones: {34, 39, 47}

Tal y como haciamos en regresión lineal múltiple, podemos utilizar la siguiente función de R para obtener las medidas del análisis de influencia automáticamente:

```
## Potentially influential observations of
##   glm(formula = Proximidad ~ As + Sb, family = "binomial", data = Oro) :
##
##    dfb.1_  dfb.As  dfb.Sb  dffit   cov.r   cook.d hat
## 9  -0.13    0.85   -1.64_* -2.22_*  1.55_*  0.28   0.48_*
## 34  1.10_* -0.66   -0.69    1.13_*  0.26_*  0.34   0.03
## 39 -0.28   -0.23    1.63_*  1.95_*  1.37_*  0.22   0.41_*
## 47  0.38   -1.43_*  0.08   -1.80_*  0.50_*  0.40   0.13
## 52  0.40    0.42   -0.48    1.16_*  0.88    0.09   0.16_*
```

Nos centramos en las columnas “cook.d”, “hat” y “dffit”:

Con respecto a los leverages de Pregibon, vemos que las observaciones {9, 39, 52} parecen influyentes.

Con respecto a la distancia de Cook, vemos que las observaciones que tienen mayor valor son {9, 34, 39, 47, 52}, pero ninguna parece tener una distancia lo suficientemente grande como para preocuparse.

Finalmente, con respecto a los dffit, vemos que las observaciones {9, 34, 39, 47, 52} parecen significativas, por lo que su predicción varía con y sin su observación.