# Trabajo Final

*Paúl Ubillús*

*2 de agosto de 2015*

**Introducción**

En el presente documento detallaremos cada uno de los pasos ejecutados para generar un modelo de regresión lineal múltiple. Además, tomaremos en cuenta las conclusiones y resultados obtenidos en el trabajo. ### Descripción información

Iniciamos cargando el archivo que contiene las varibles a utilizar, el paquete *readxl* permite leer archivos desde excel sin la necesidad de instalar complementos.

```
options(warn=-1)
library(readxl)
datarls1 <- read_excel("poblacion1.xlsx",sheet = 1,col_names = TRUE,na = "")
datarls2 <- read_excel("poblacion2.xlsx",sheet = 1,col_names = TRUE,na = "")
str(datarls1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    44 obs. of  4 variables:
##  $ identificador : num  1001 1002 1003 1004 1005 ...
##  $ poblacion     : num  18.7 13.8 13.4 11.4 10.5 10.3 10.3 10.3 10.3 9.4 ...
##  $ var.pobl.mayor: num  6.2 7.2 7.1 2.6 2.8 1.4 -0.5 3 -1.9 -1.5 ...
##  $ menores.18    : num  39.7 10.4 20.4 38.7 36.4 29.8 31.8 31.1 15.9 22.4 ...
```

```
str(datarls2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    40 obs. of  7 variables:
##  $ identificador  : num  1044 1042 1010 1009 1008 ...
##  $ part.almz.escl : num  73.8 34.4 22.2 39.9 51.7 57.2 43.7 72.3 78.2 13.9 ...
##  $ var.ingresos   : num  50.5 24.2 33.5 38.5 26.2 27.2 29.4 26 18.4 22.5 ...
##  $ tasa.crimen    : num  704.1 179.9 61.5 86.4 42.4 ...
##  $ var.tasa.crimen: num  -40.9 12.3 -32.7 -13.5 -31.9 -17.6 -10 -21.6 -12.8 -33.6 ...
##  $ region         : chr  "B" "B" "A" "A" ...
##  $ serv.bas.compl : chr  "SI" "NO" "NO" "SI" ...
```

Analizando la información disponemos en la primera data de 44 observaciones de 4 variables y en la segunda data de 40 observaciones de 7 variables.

Luego procedemos a unir los archivos leidos en un mismo objeto.

```
options(warn=-1)
poblacion <- merge(x = datarls1 ,y = datarls2)
str(poblacion)
```

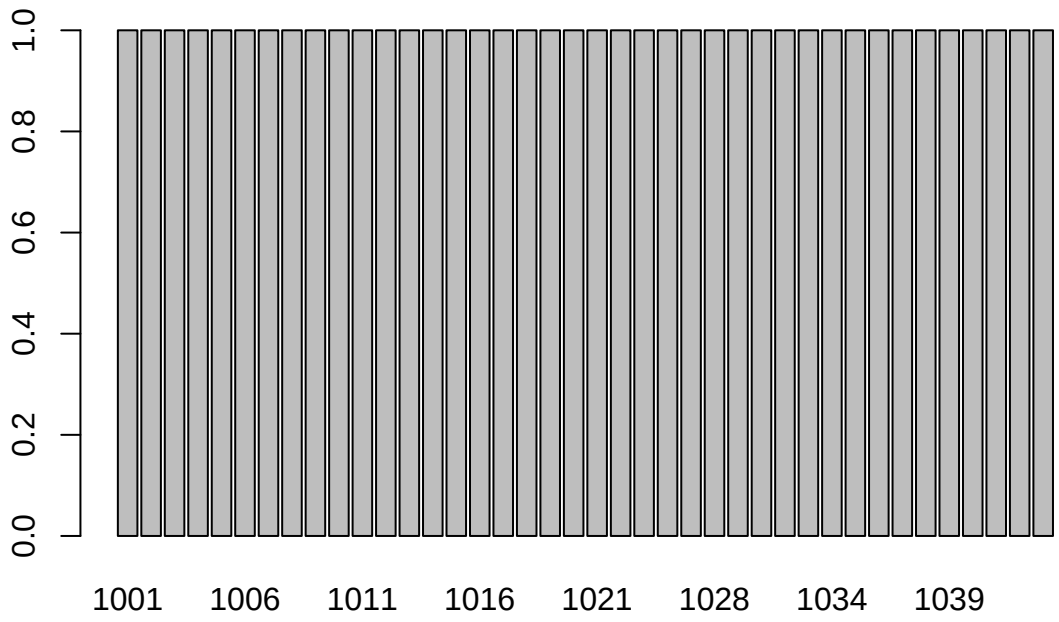
```
## 'data.frame':    40 obs. of  10 variables:
##  $ identificador  : num  1001 1002 1003 1004 1005 ...
##  $ poblacion      : num  18.7 13.8 13.4 11.4 10.5 10.3 10.3 10.3 10.3 9.4 ...
##  $ var.pobl.mayor : num  6.2 7.2 7.1 2.6 2.8 1.4 -0.5 3 -1.9 -1.5 ...
##  $ menores.18     : num  39.7 10.4 20.4 38.7 36.4 29.8 31.8 31.1 15.9 22.4 ...
```

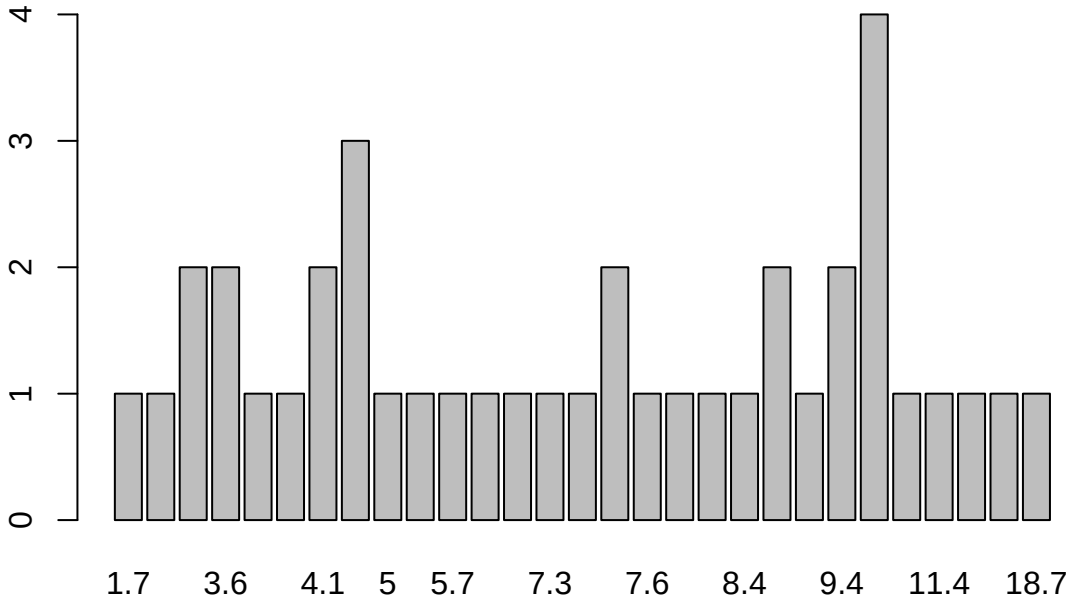
```
##  $ part.almz.escl : num  55.8 57.9 13.9 78.2 72.3 43.7 57.2 51.7 39.9 22.2 ...
##  $ var.ingresos   : num  28.7 26.2 22.5 18.4 26 29.4 27.2 26.2 38.5 33.5 ...
##  $ tasa.crimen    : num  52.6 111 38.3 86.6 77.5 54 61.2 42.4 86.4 61.5 ...
##  $ var.tasa.crimen: num  -2.9 -22.6 -33.6 -12.8 -21.6 -10 -17.6 -31.9 -13.5 -32.7 ...
##  $ region         : chr  "A" "A" "A" "A" ...
##  $ serv.bas.compl : chr  "SI" "SI" "NO" "NO" ...
```

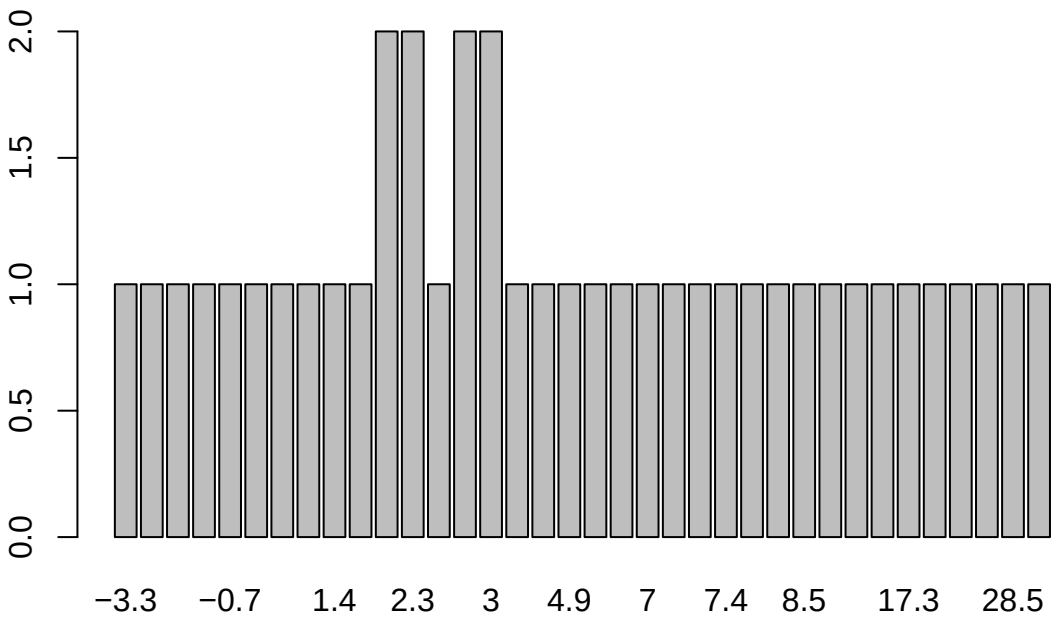
Ahora disponemos en la primera data de 40 observaciones de 10 variables.

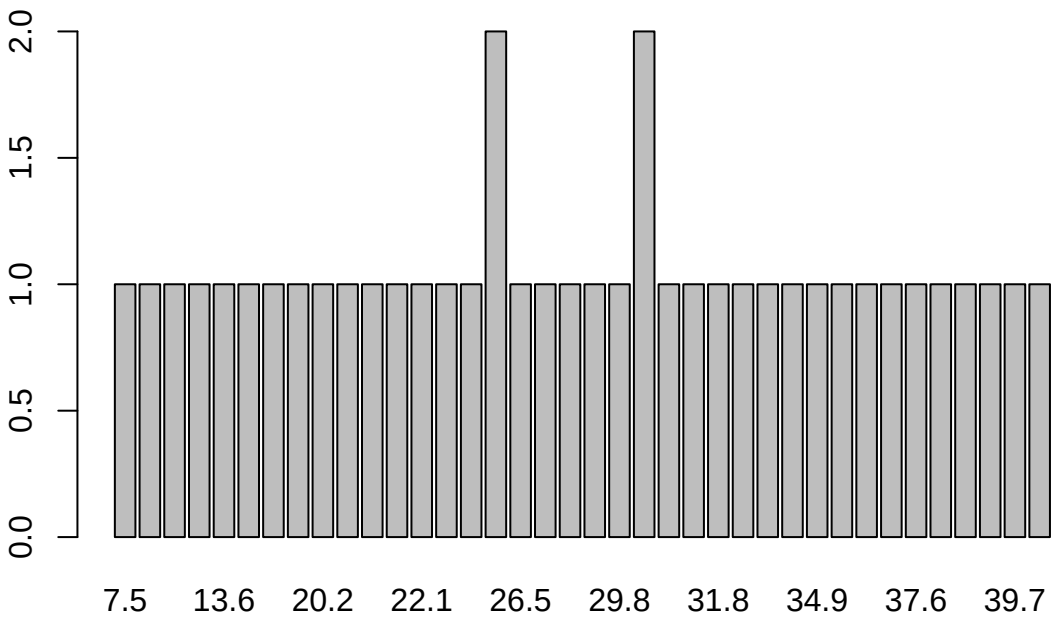
Procedemos a crear un codigo que identifique la clase de cada variable y genere diagrama de cajas para variables continuas y diagrama de barras para variables discretas.

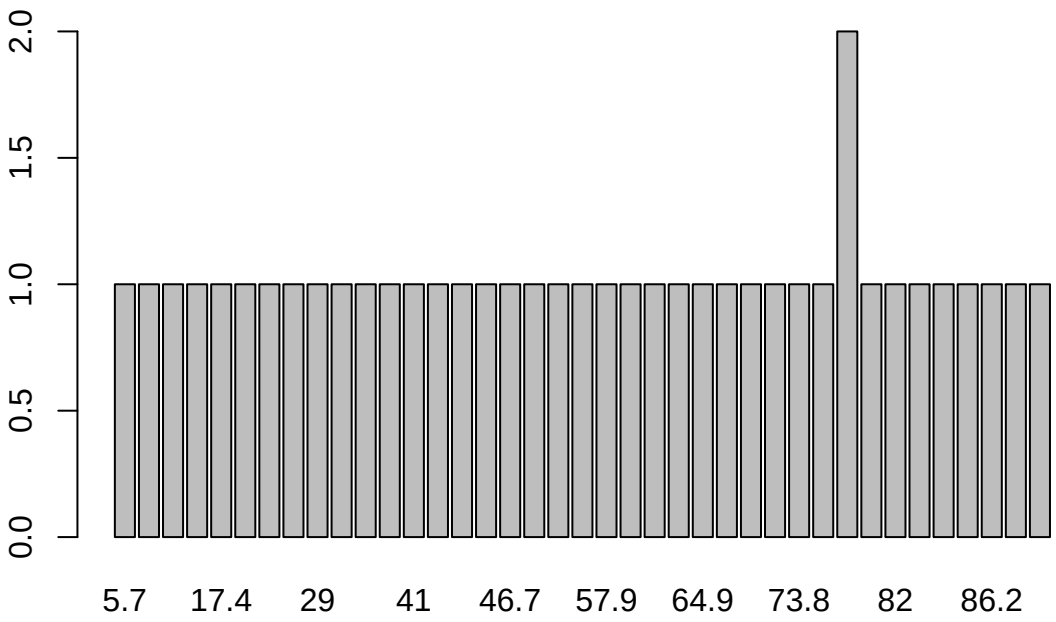
```
options(warn=-1)
for(i in 1:(ncol(poblacion))){
  if(is.numeric(poblacion[i])==T){
    hist(poblacion[,i])
  }else{
    barplot(table(poblacion[,i]))
  }
}
```

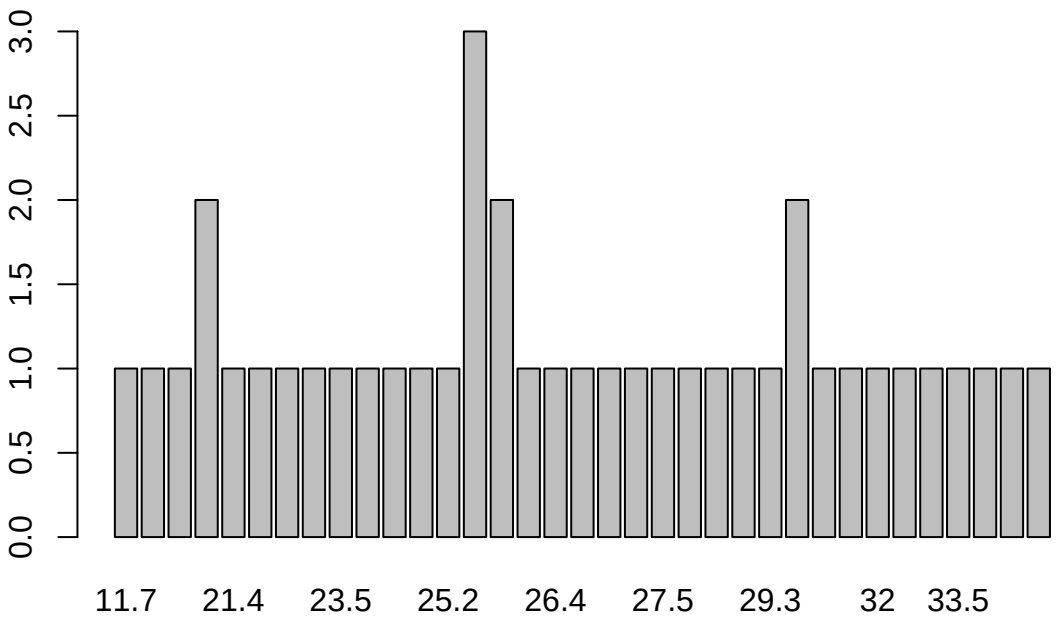


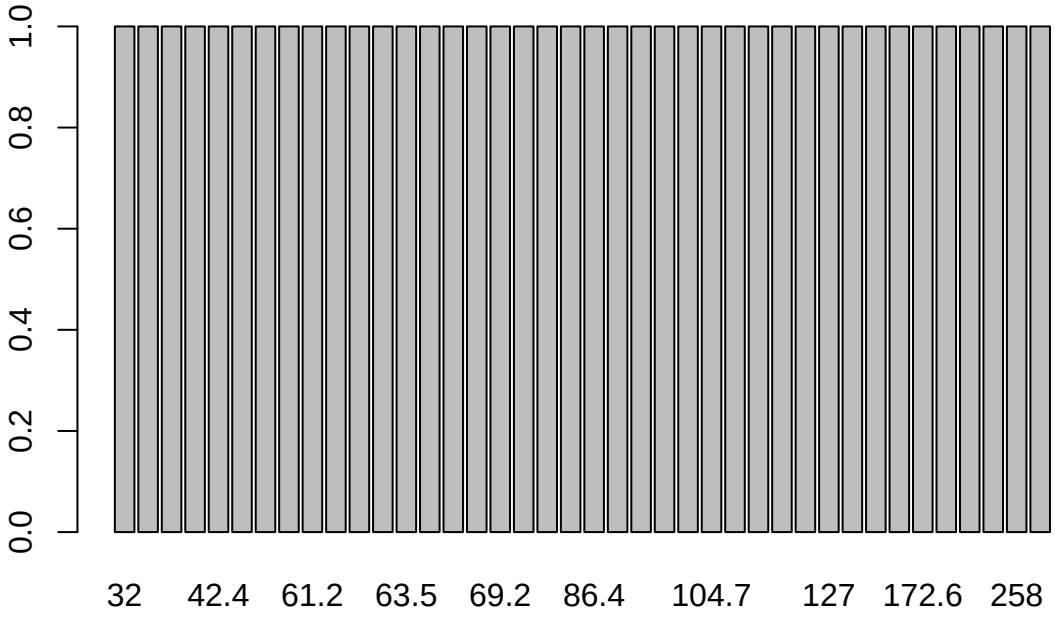

0
1
2
3
4
1.7
3.6
4.1
5
5.7
7.3
7.6
8.4
9.4
11.4
18.7

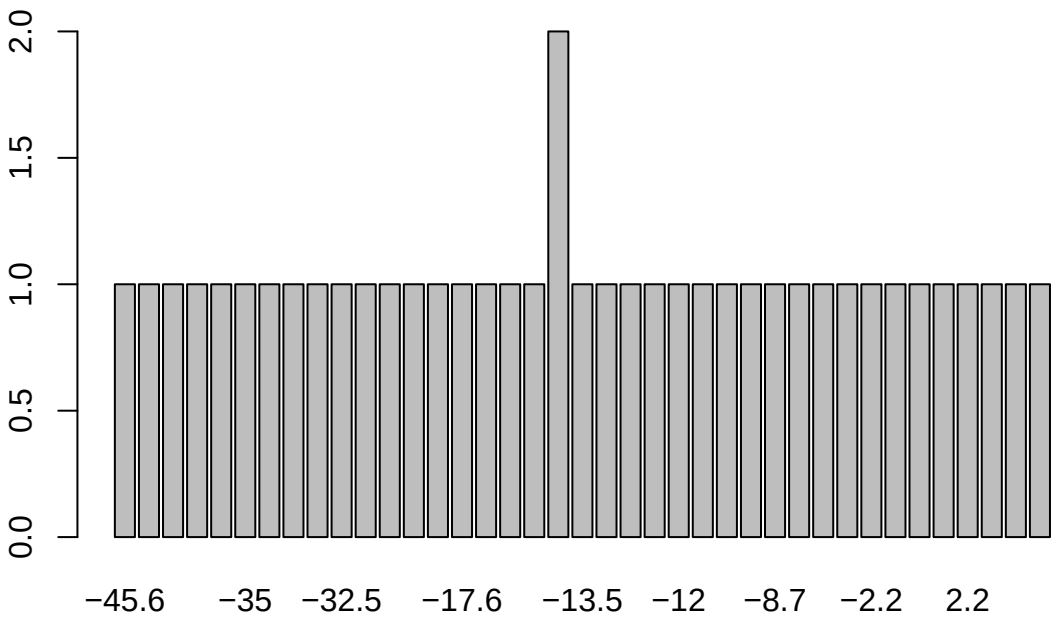

0.0
0.5
1.0
1.5
2.0
−3.3
−0.7
1.4
2.3
3
4.9
7
7.4
8.5
17.3
28.5

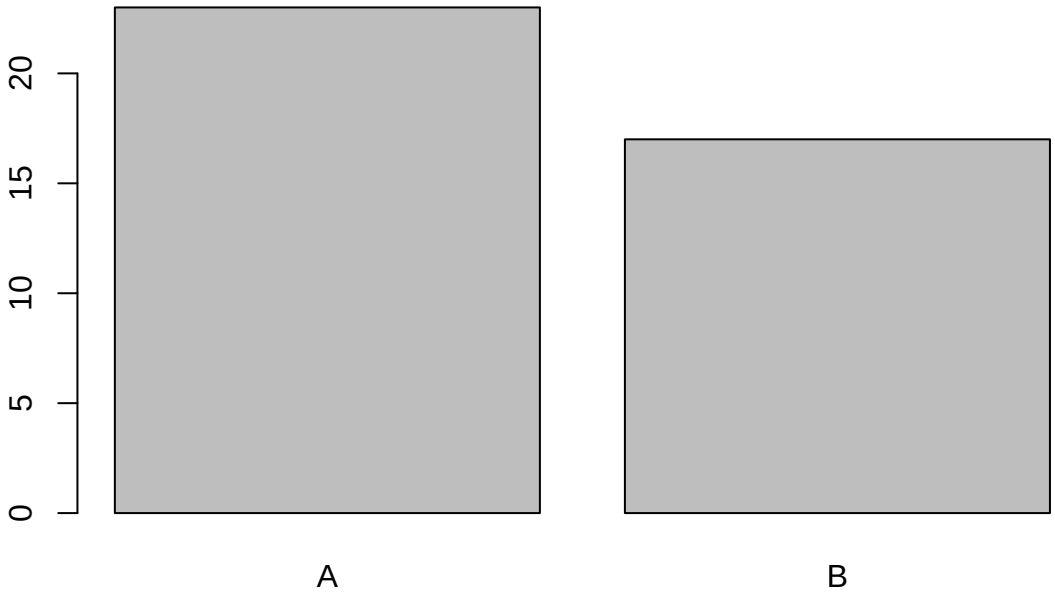

0.0
0.5
1.0
1.5
2.0
7.5
13.6
20.2
22.1
26.5
29.8
31.8
34.9
37.6
39.7

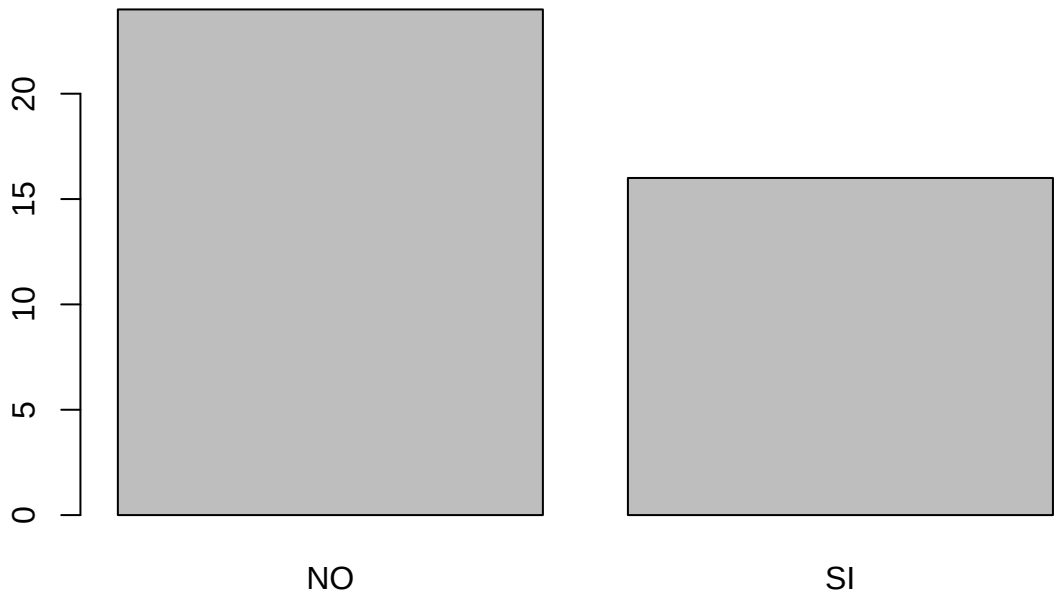

0.0
0.5
1.0
1.5
2.0
5.7
17.4
29
41
46.7
57.9
64.9
73.8
82
86.2


0.0
0.5
1.0
1.5
2.0
2.5
3.0
11.7
21.4
23.5
25.2
26.4
27.5
29.3
32
33.5


0.0
0.2
0.4
0.6
0.8
1.0
32
42.4
61.2
63.5
69.2
86.4
104.7
127
172.6
258


0.0
0.5
1.0
1.5
2.0
−45.6
−35
−32.5
−17.6
−13.5
−12
−8.7
−2.2
2.2


0
5
10
15
20
A
B




Creemos un codigo que calcule automaticamente el mínimo, media, máximo, desviación estándar, primer cuartil de cada variable numérica y frecuencia en el caso de variables categoricas.

```
options(warn=-1)
for(i in 1:(ncol(poblacion))) {
  if(is.numeric(poblacion[i])==T){
    print(names(poblacion[i]))
    print(summary(poblacion)[4,i])
    print(summary(poblacion)[4,i])
    print(summary(poblacion)[6,i])
    print(sd(poblacion[,i]))
    print(summary(poblacion)[2,i])
    print("*******")
  } else {
    print(names(poblacion[i]))
    print(summary(poblacion)[4,i])
    print(summary(poblacion)[4,i])
    print(summary(poblacion)[6,i])
    print(sd(poblacion[,i]))
    print(summary(poblacion)[1,i])
    print("********")
  }
}
```

```
## [1] "identificador"
## [1] "Mean   :1022  "
```

```
## [1] "Mean   :1022  "
## [1] "Max.   :1044  "
## [1] 12.99349
## [1] "Min.   :1001  "
## [1] "********"
## [1] "poblacion"
## [1] "Mean   : 7.232  "
## [1] "Mean   : 7.232  "
## [1] "Max.   :18.700  "
## [1] 3.621403
## [1] "Min.   : 1.700  "
## [1] "********"
## [1] "var.pobl.mayor"
## [1] "Mean   : 6.647  "
## [1] "Mean   : 6.647  "
## [1] "Max.   :32.200  "
## [1] 8.148776
## [1] "Min.   :-3.300  "
## [1] "********"
## [1] "menores.18"
## [1] "Mean   :27.07  "
## [1] "Mean   :27.07  "
## [1] "Max.   :41.50  "
## [1] 9.036337
## [1] "Min.   : 7.50  "
## [1] "********"
## [1] "part.almz.escl"
## [1] "Mean   :53.17  "
## [1] "Mean   :53.17  "
## [1] "Max.   :88.50  "
## [1] 25.42078
## [1] "Min.   : 5.70  "
## [1] "********"
## [1] "var.ingresos"
## [1] "Mean   :26.79  "
## [1] "Mean   :26.79  "
## [1] "Max.   :50.50  "
## [1] 6.854165
## [1] "Min.   :11.70  "
## [1] "********"
## [1] "tasa.crimen"
## [1] "Mean   :111.03  "
## [1] "Mean   :111.03  "
## [1] "Max.   :704.10  "
## [1] 108.8891
## [1] "Min.   : 32.00  "
## [1] "********"
## [1] "var.tasa.crimen"
## [1] "Mean   :-15.963  "
## [1] "Mean   :-15.963  "
## [1] "Max.   : 27.200  "
## [1] 15.75862
## [1] "Min.   :-45.600  "
## [1] "********"
```

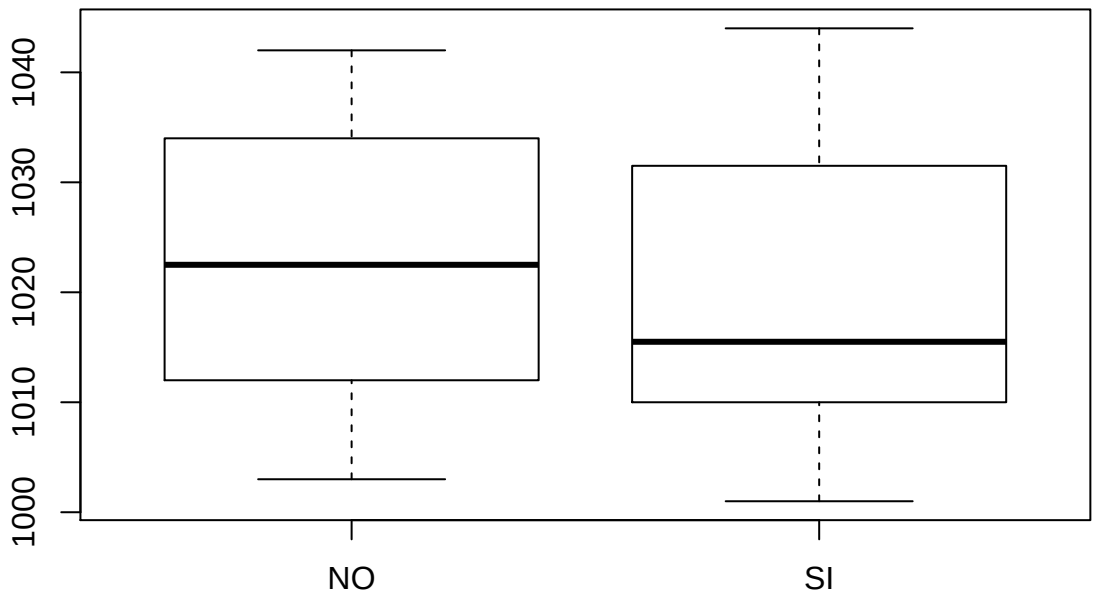
```
## [1] "region"
## [1] NA
## [1] NA
## [1] NA
## [1] NA
## [1] "Length:40         "
## [1] "********"
## [1] "serv.bas.compl"
## [1] NA
## [1] NA
## [1] NA
## [1] NA
## [1] "Length:40         "
## [1] "********"
```

Consideremos la variable categórica "serv.bas.compl" con una confiabilidad del 90% ¿Puede asumirse que la media de la variable "poblacion" en el grupo "serv.bas.compl:SI" es distinta a la media del grupo "serv.bas.compl:NO" ? Primero veamos si podemos asumir que las varianzas de los grupos que se van a tomar son iguales o no. Procedemos a realizar un diagrama de cajas:

```
options(warn=-1)
names(poblacion)
```

```
##  [1] "identificador"   "poblacion"       "var.pobl.mayor"
##  [4] "menores.18"      "part.almz.escl"  "var.ingresos"
##  [7] "tasa.crimen"     "var.tasa.crimen" "region"
## [10] "serv.bas.compl"
```

```
var1<-poblacion[,1]
var2<-poblacion[,10]
boxplot(var1~var2)
```




Gracias al diagrama de cajas podemos ver que en el grupo de "SI" existe más variación que en el grupo de "NO". Además, comparando las varianzas de los 2 grupos tenemos:

```
options(warn=-1)
var(var1[var2=="SI"])
```

```
## [1] 194.3625
```

```
var(var1[var2=="NO"])
```

```
## [1] 154.8243
```

Claramente vemos que las varianzas son distintas. Ahora procedemos a aceptar o rechazar nuestra prueba de hipotesis de las medias. Procedamos a comprobar la hipotesis si las medias son iguales. Así:

```
options(warn=-1)
t.test(var1[var2=="SI"], var1[var2=="NO"],conf.level = 0.90)
```

```
##
##  Welch Two Sample t-test
##
## data:  var1[var2 == "SI"] and var1[var2 == "NO"]
## t = -0.77776, df = 29.699, p-value = 0.4429
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 90 percent confidence interval:
##  -10.676164   3.967831
## sample estimates:
## mean of x mean of y
##  1019.688  1023.042
```

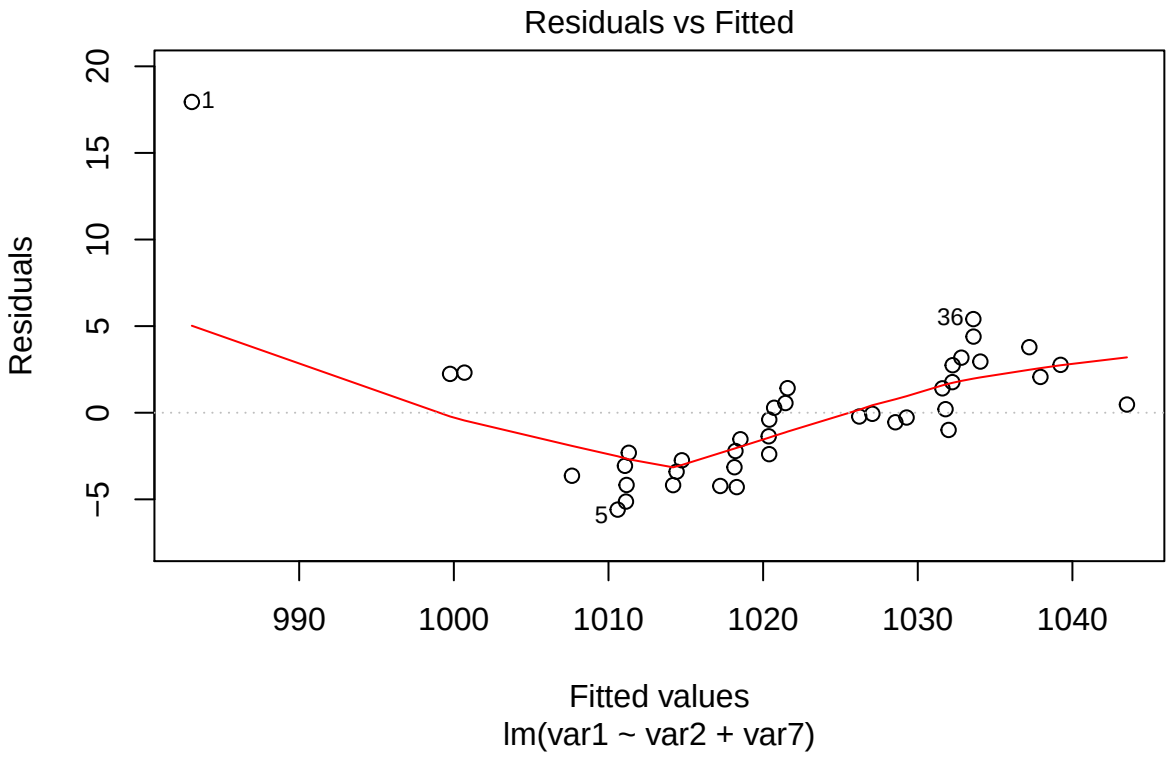
Como $t = -0.7777566$ es menor que $df = 29.6993396$ se acepta $H_0 : u1 - u2 = 0$.

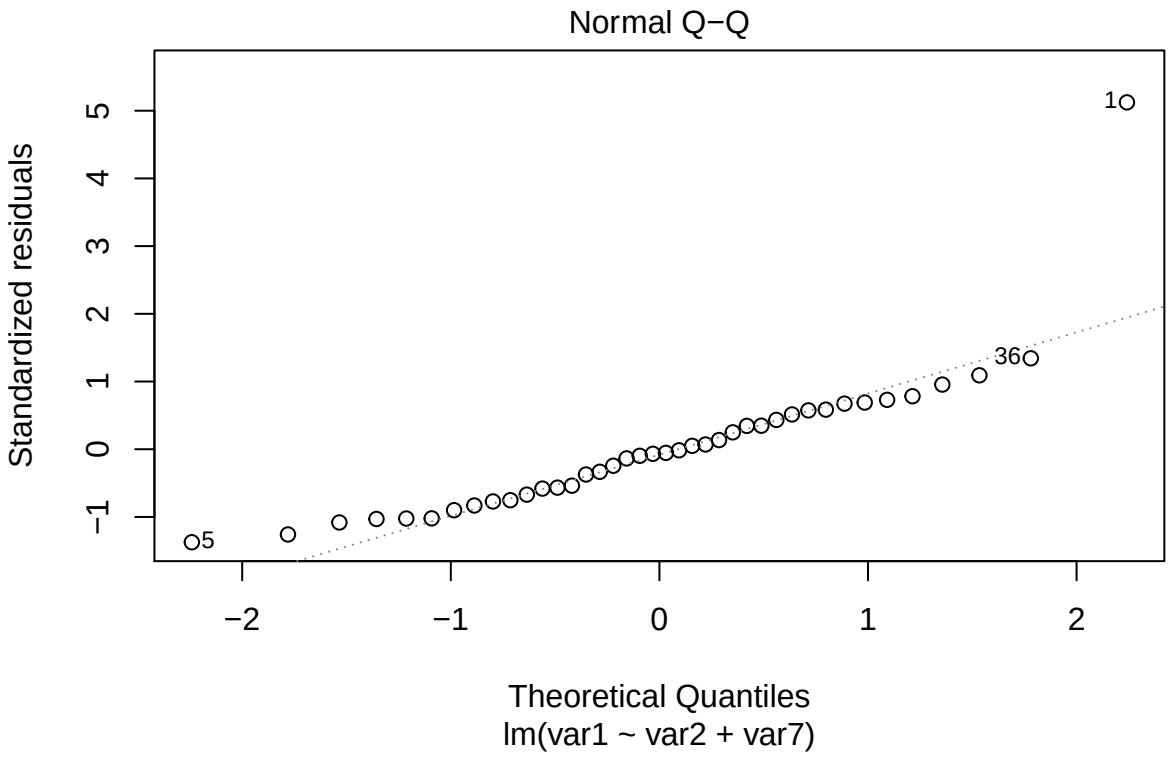
Generemos el modelo de regresión lineal múltiple que mejor se ajuste a nuestros datos.

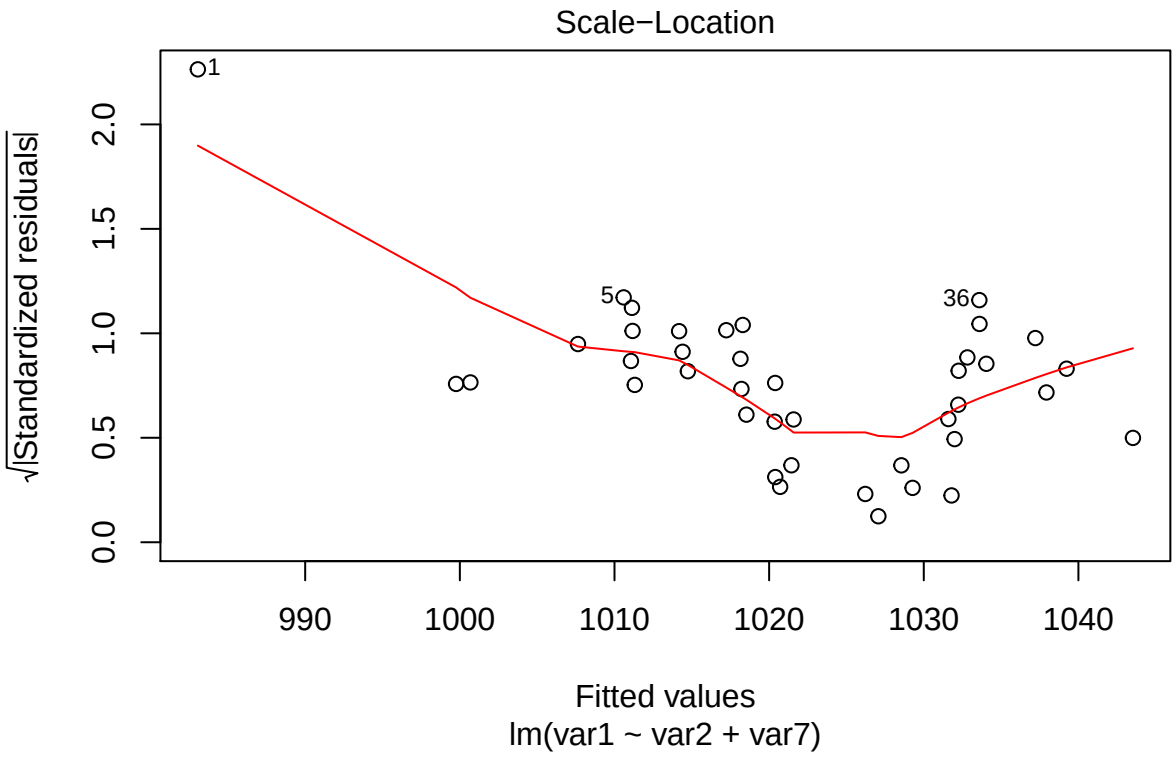
```
options(warn=-1)
var2<-poblacion[,2]
var7<-poblacion[,7]
mod1<-lm(var1~var2+var7)
summary(mod1)
```

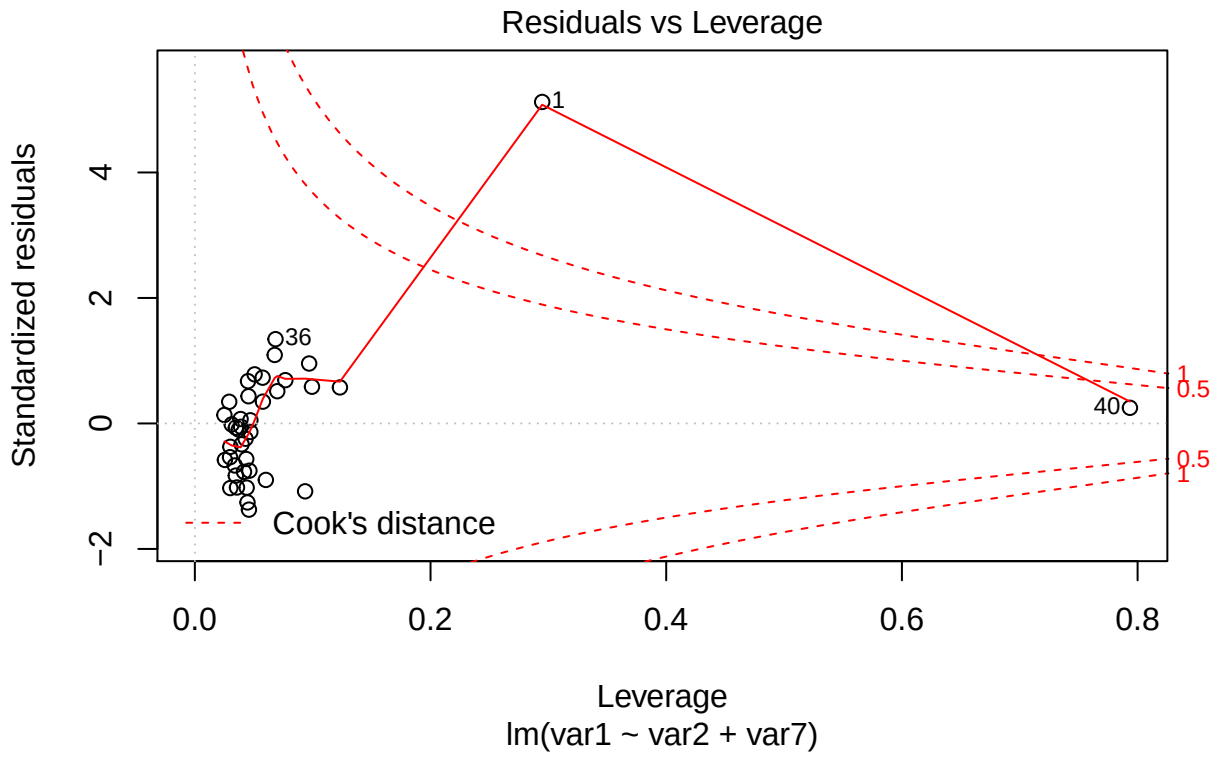
```
##
## Call:
## lm(formula = var1 ~ var2 + var7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5943 -2.8257 -0.2477  2.1106 17.9421
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.045e+03  1.959e+00 533.684   <2e-16 ***
## var2        -3.341e+00  1.989e-01 -16.797   <2e-16 ***
## var7         5.638e-03  6.615e-03   0.852      0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.17 on 37 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.897
## F-statistic: 170.8 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
plot(mod1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(var1 ~ var2 + var7)
1
5
36
20
15
10
5
0
−5
990
1000
1010
1020
1030
1040


Normal Q−Q
Standardized residuals
−1
0
1
2
3
4
5
1
36
5
−2
−1
0
1
2
Theoretical Quantiles
lm(var1 ~ var2 + var7)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1
5
36
990
1000
1010
1020
1030
1040
Fitted values
lm(var1 ~ var2 + var7)




Gracias a los gráficos podemos ver que nuestro modelo lineal múltiple siguen normalidad.

**Interpretación de Coeficientes**

Tenemos que: $B1 = 1045.2376674$, $B2 = 1.9585318$ y $B3 = 533.6842932$ son significantes. También tenemos que, $Pr1 = 1.6447373 \times 10^{-73}$, $Pr2 = 6.7286169 \times 10^{-19}$ y $Pr3 = 0.39955$ los cuales son valores muy pequeños con respecto a $t1 = 533.6842932$, $t2 = -16.7974568$ y $t3 = 0.8522653$ respectivamente y por tanto podemos decir que nuestros coeficientes son significativos.

**Interpretación de R^2**

Como $R^2 = 0.9022815$ podemos decir que aproximadamente el $90.2281517\%$ de nuestra variación en nuestro problema puede ser explicado por este modelo, además el $R^2 ajustado = 0.9022815$ por lo tanto la regresión es significativa.

***Análisis de Significancia de la Regresión***

**Gráficos de Dispersión**

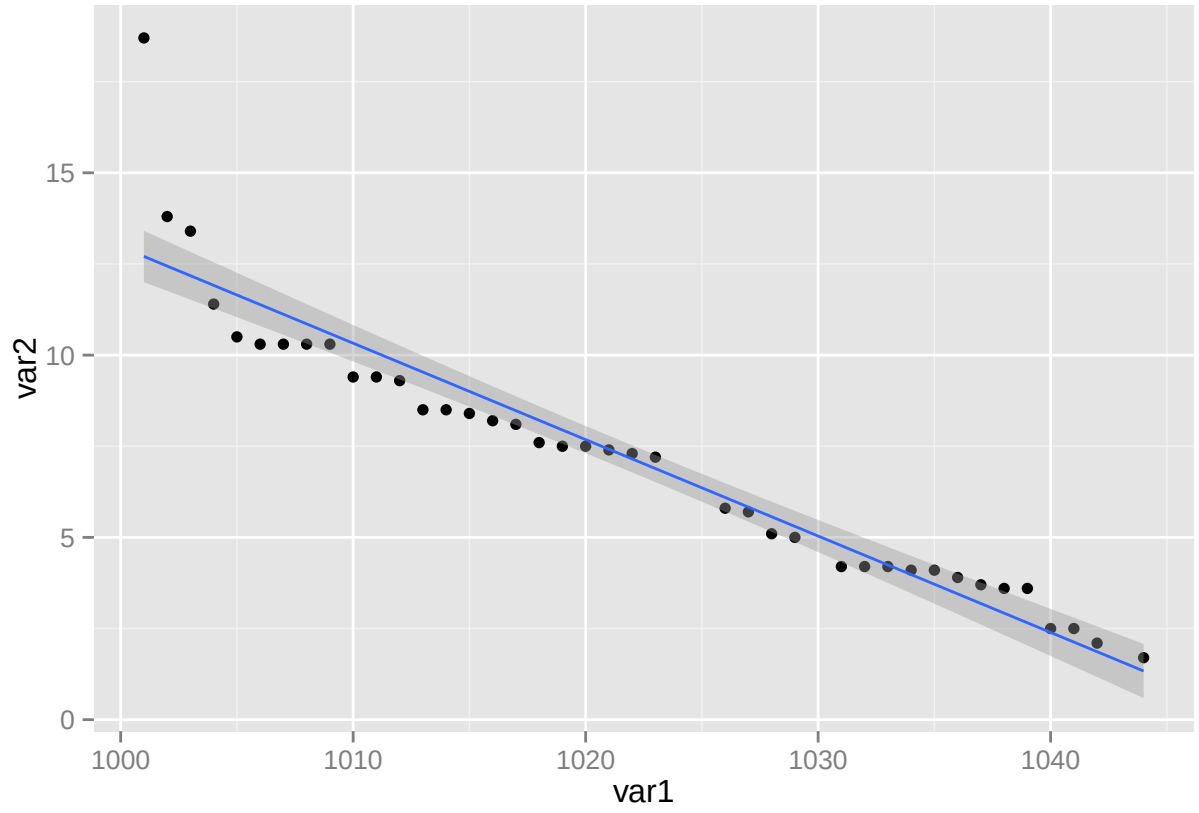
Carguemos la librería *library(ggplot2)* que nos permite realizar nuestros gráficos para concluir sobre la significancion de nuestra regresión.

```
options(warn=-1)
library(ggplot2)
```

Ahora realicemos un estudio entre las diferentes variables tomadas en cuenta en nuestra regresión.

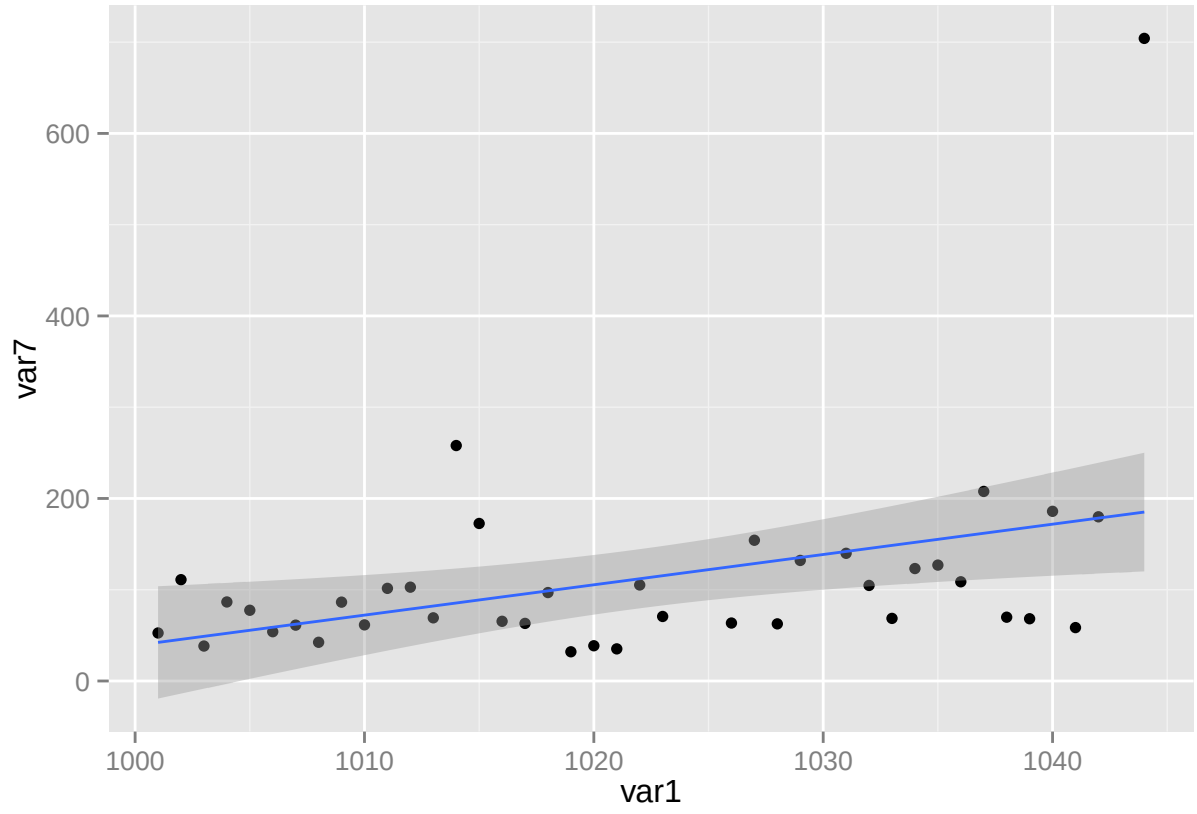
**var1 vs var2**

```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var1, y=var2))
g + geom_point() + geom_smooth(method="lm")
```



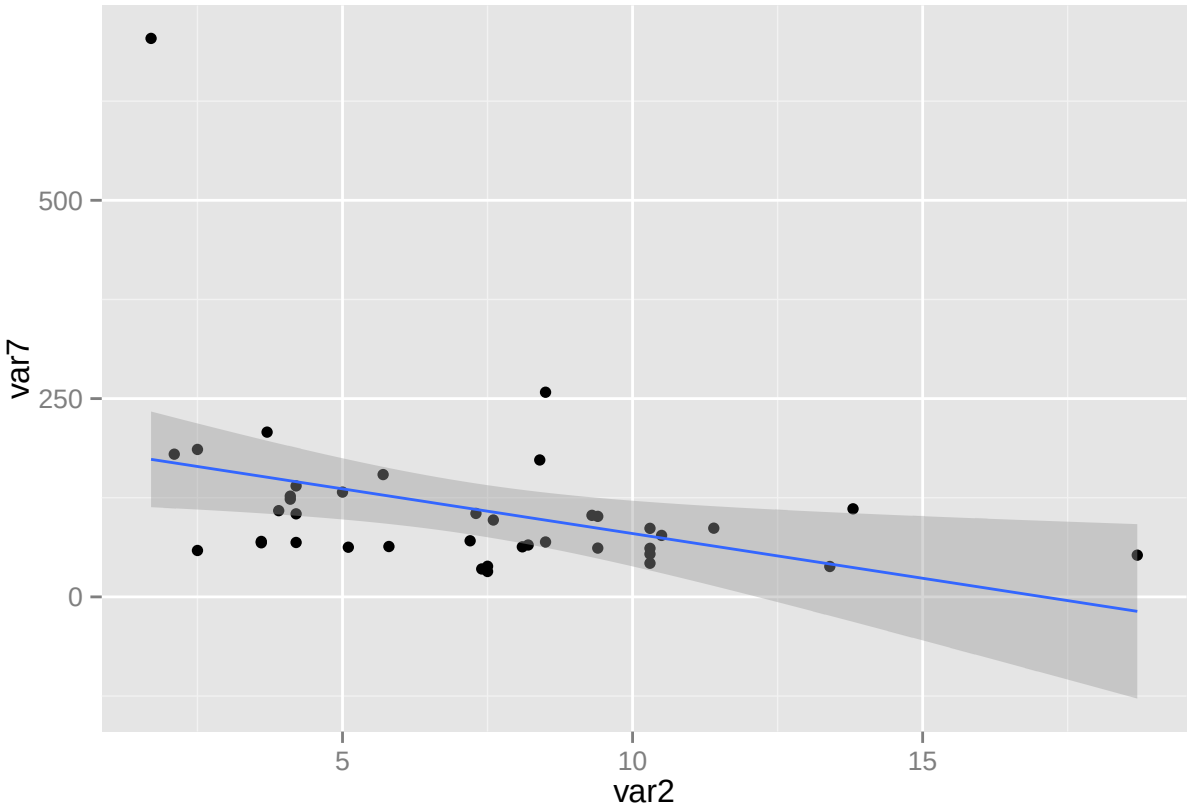

**var1 vs var7**

```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var1, y=var7))
g + geom_point() + geom_smooth(method="lm")
```




**var2 vs var7**

```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var2, y=var7))
g + geom_point() + geom_smooth(method="lm")
```

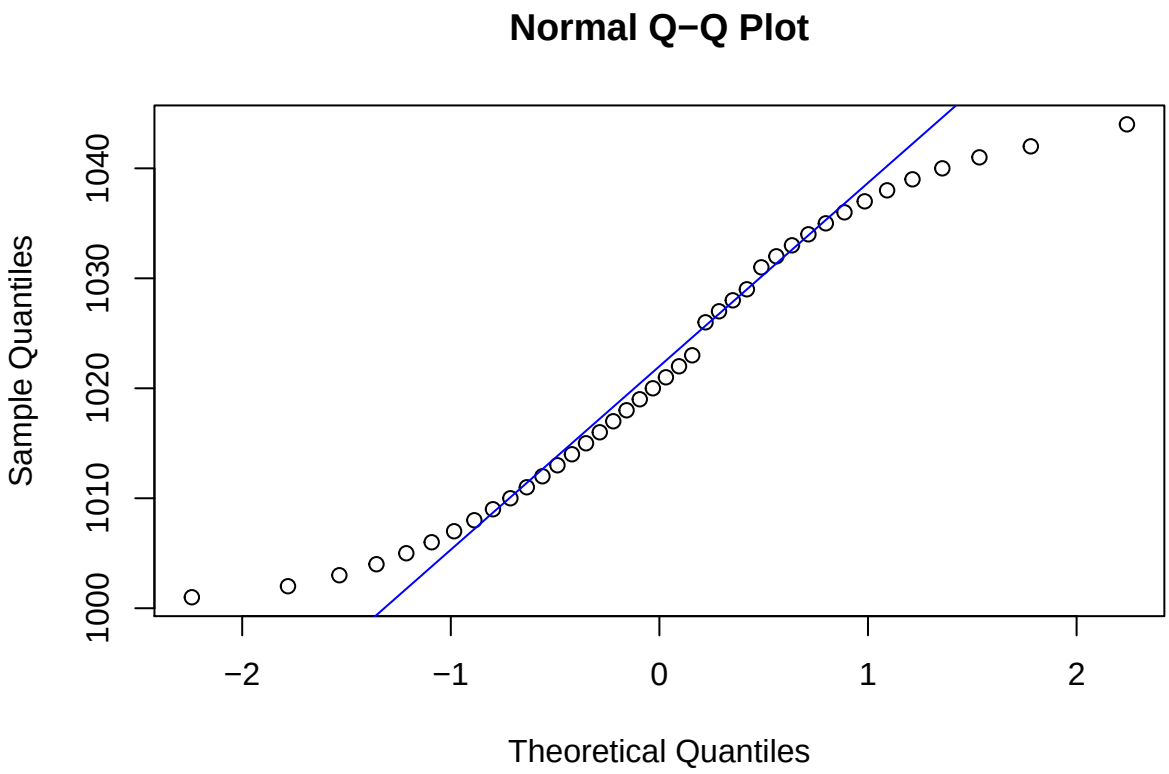



En el caso del primer gráfico podemos distinguir una relación lineal entre las variables. En el segundo y tercer caso se tiene una mayor dispersión de puntos. Tomemos en cuenta que en los 3 casos existen puntos atípicos.

**Gráfico de Normalidad**

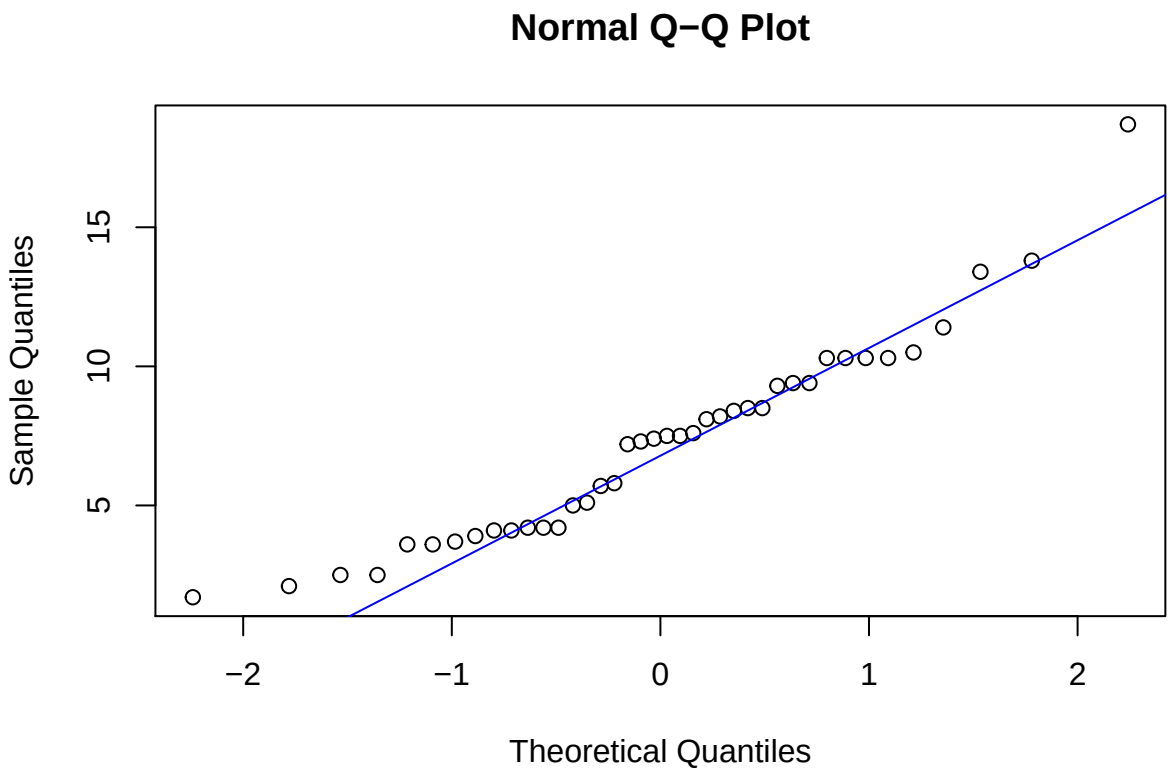
**var1**

```
options(warn=-1)
qqnorm(var1)
qqline(var1,col="blue",size=2)
```



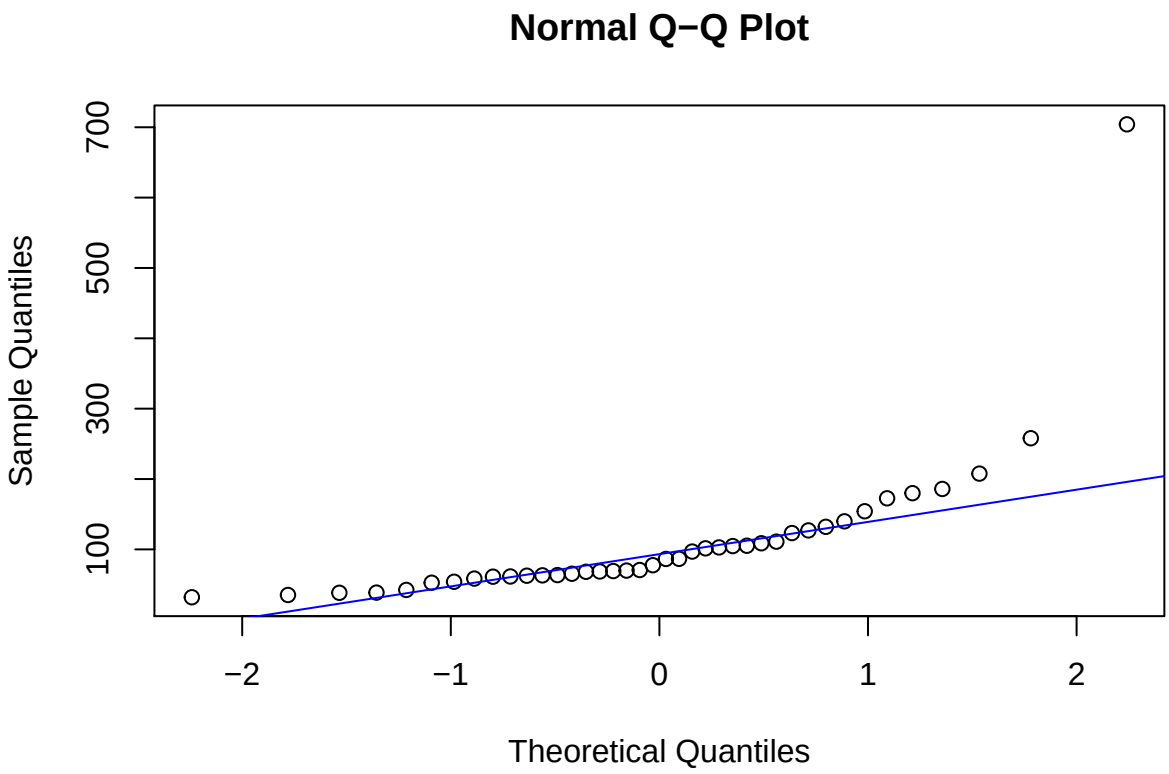

**var2**

```
options(warn=-1)
qqnorm(var2)
qqline(var2,col="blue",size=2)
```




**var7**

```
options(warn=-1)
qqnorm(var7)
qqline(var7,col="blue",size=2)
```

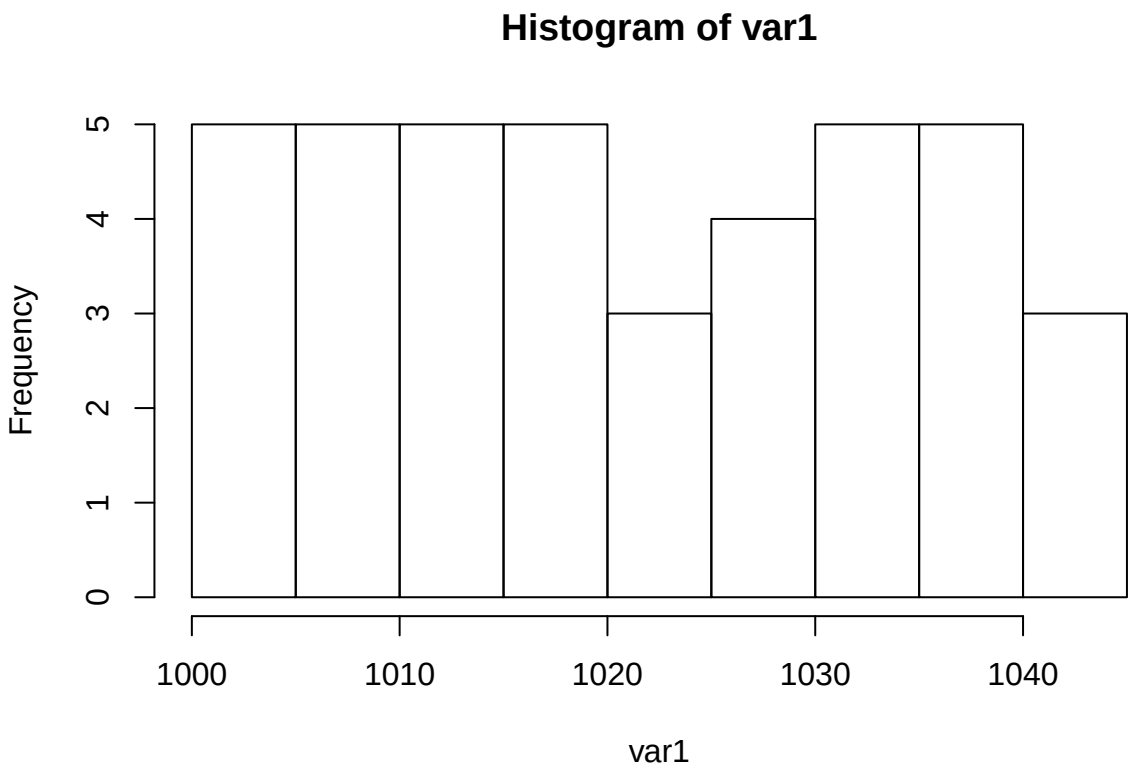



En el tercer gráfico se puede notar claramente normalidad mietras que en los graficos 1 y 2 se viola el supuesto de normalidad.

**Histogramas**

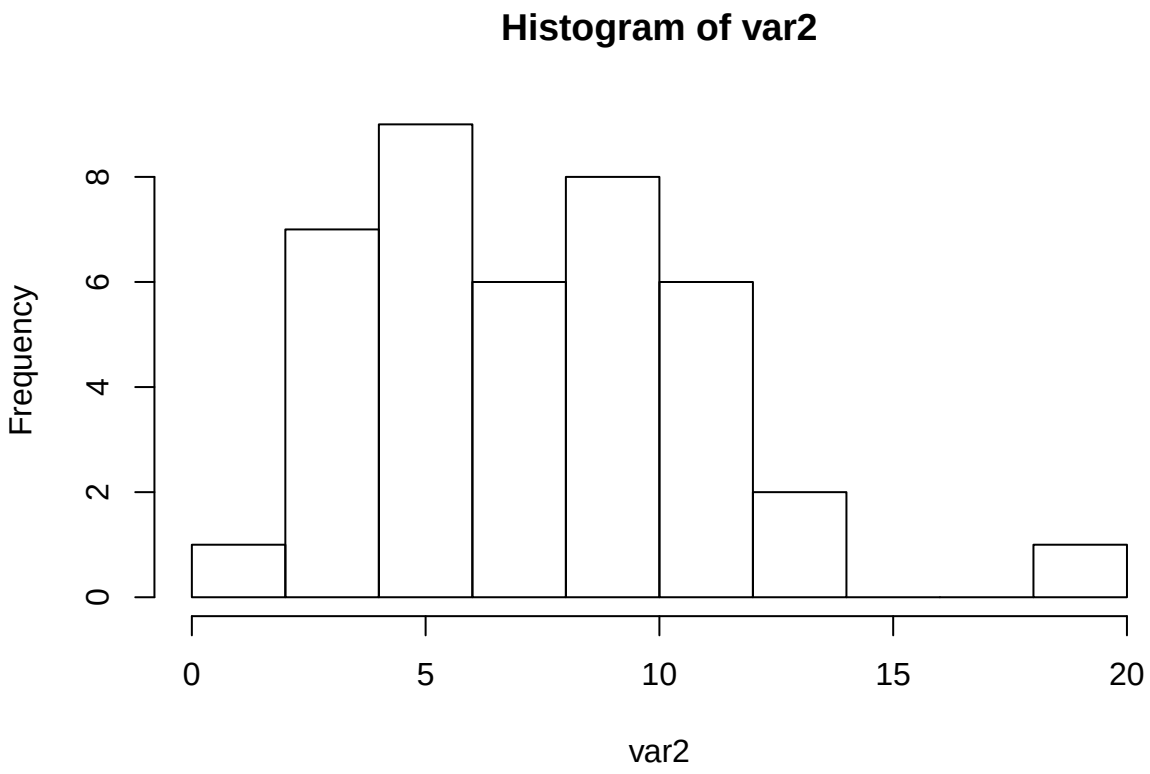
**var1**

```
options(warn=-1)
hist(var1)
```



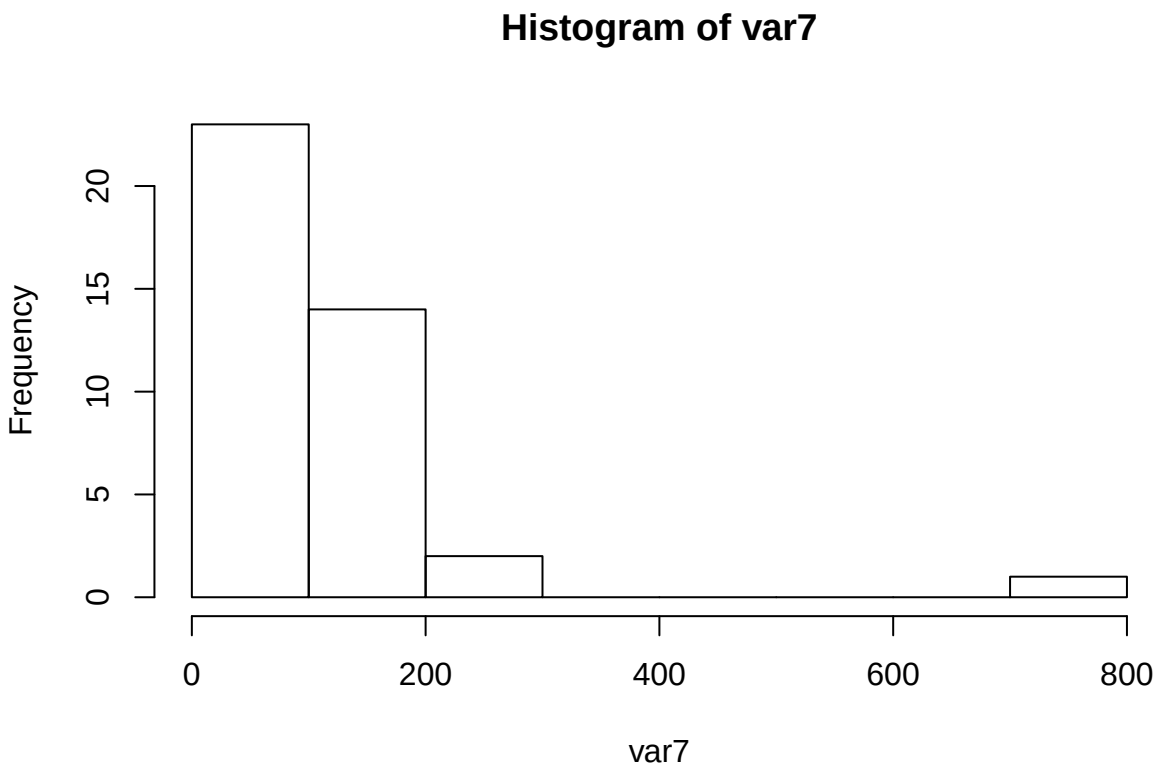

**var2**

```
options(warn=-1)
hist(var2)
```




**var7**

```
options(warn=-1)
hist(var7)
```




Podemos ver que solo el segundo gráfico tiene tendencia a seguir una ley normal y tener simetría.

***Análisis de Residuos***

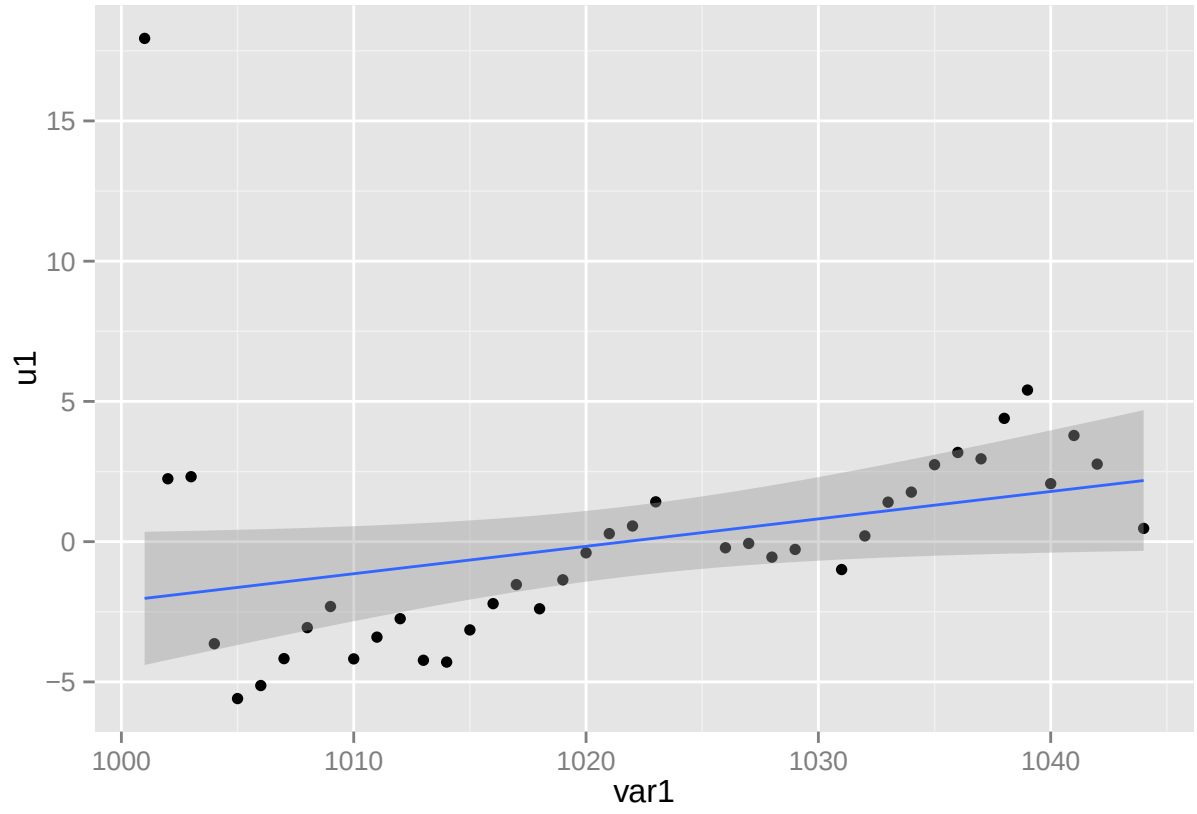
En primer lugar calculemos los residuos de neustra regresion lineal los cuales son:

```
options(warn=-1)
u1<- mod1$residuals
```

Ahora estudiemos los gráficos residuales con relación a los residuos que acabamos de calcular.

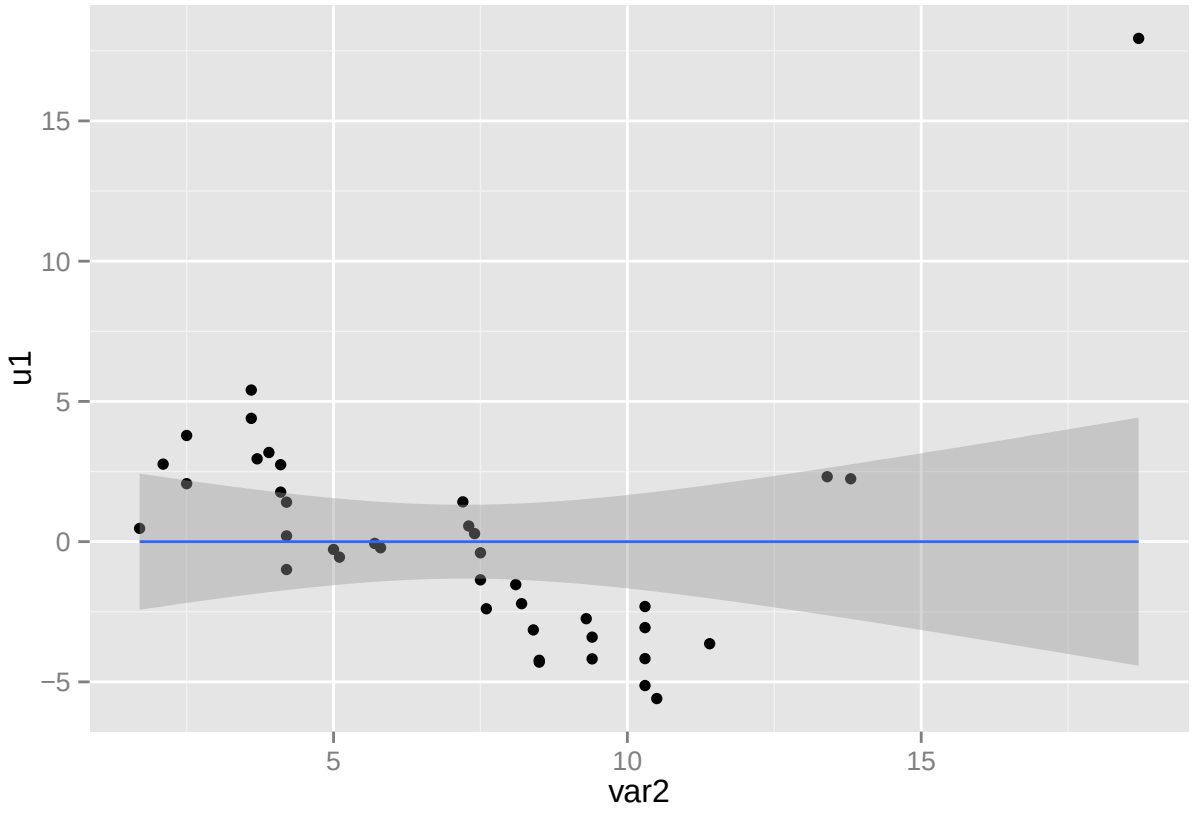
**var1 vs u1**

```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var1, y=u1))
g + geom_point() + geom_smooth(method="lm")
```



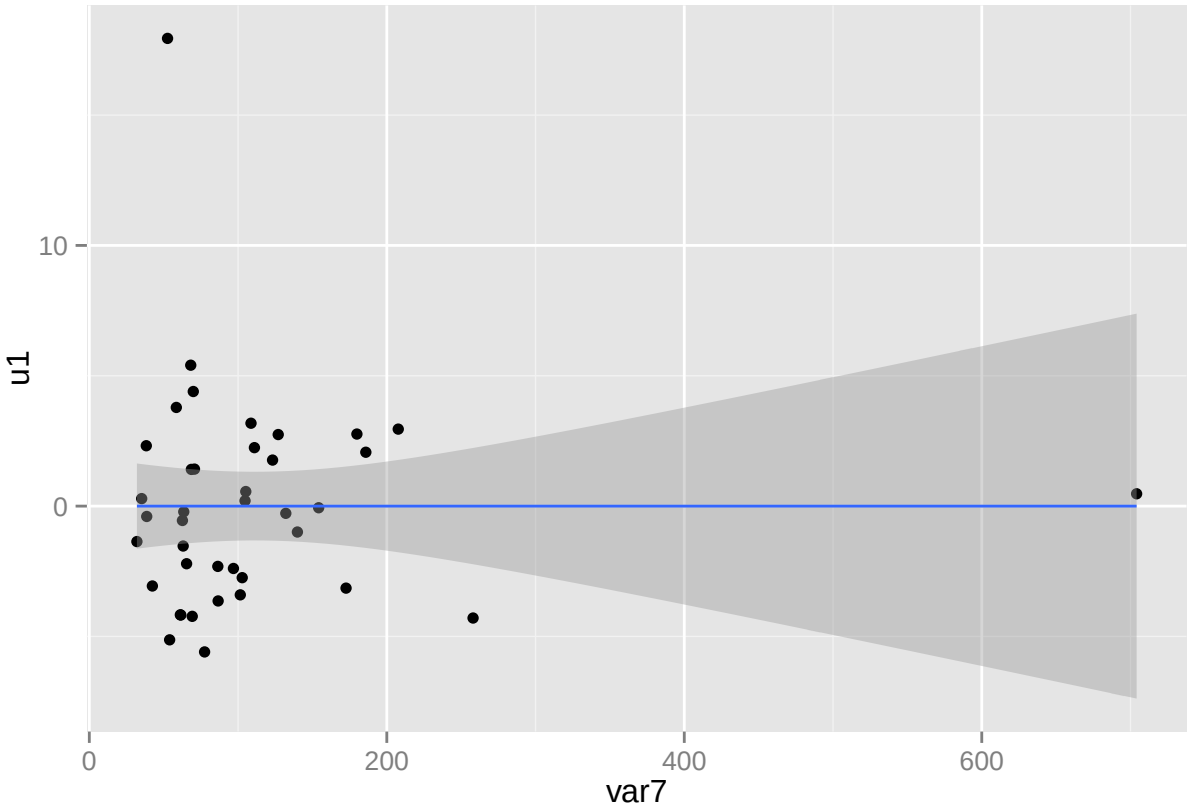

**var2 vs u1**

```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var2, y=u1))
g + geom_point() + geom_smooth(method="lm")
```




**var7 vs u1**

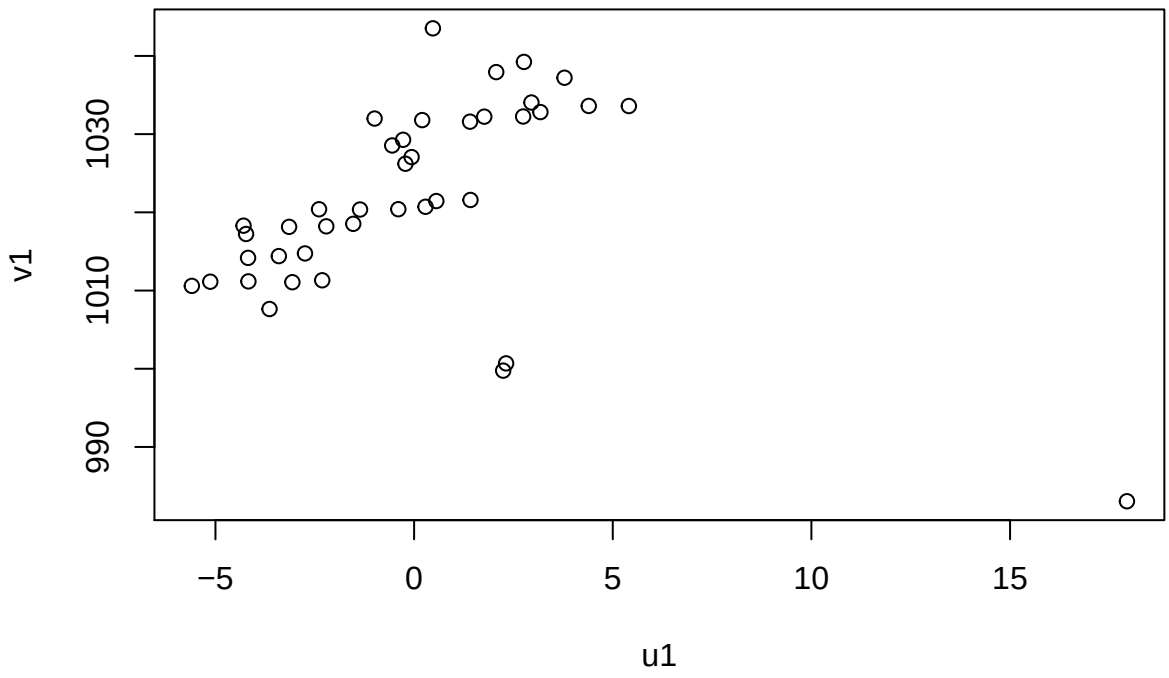
```
options(warn=-1)
g <- ggplot(data = poblacion, aes(x=var7, y=u1))
g + geom_point() + geom_smooth(method="lm")
```




Podemos observar que no estan aleatoriamente distribuidos en una banda centrada en 0. Realicemos un estudio de los pronósticos de nuestro modelo lineal vs los residuos del mismo.

_pronóstios vs Residuos

```
options(warn=-1)
v1 <- mod1$fitted.values
plot(u1,v1)
```




podemos observar que los puntos estan dispersos por ser puntos atípicos e influyentes por tanto no se tiene una buena linealidad. Se debe corregir el modelo lineal. Así,

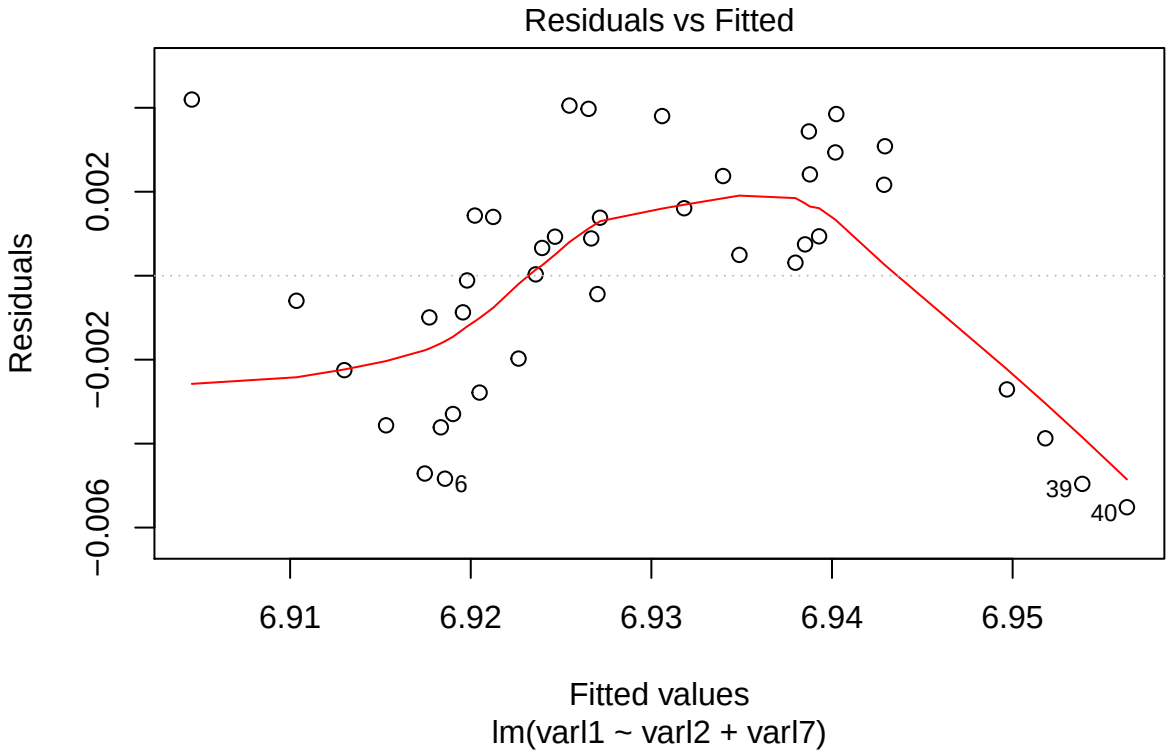
**Reajustando el modelo inicial**

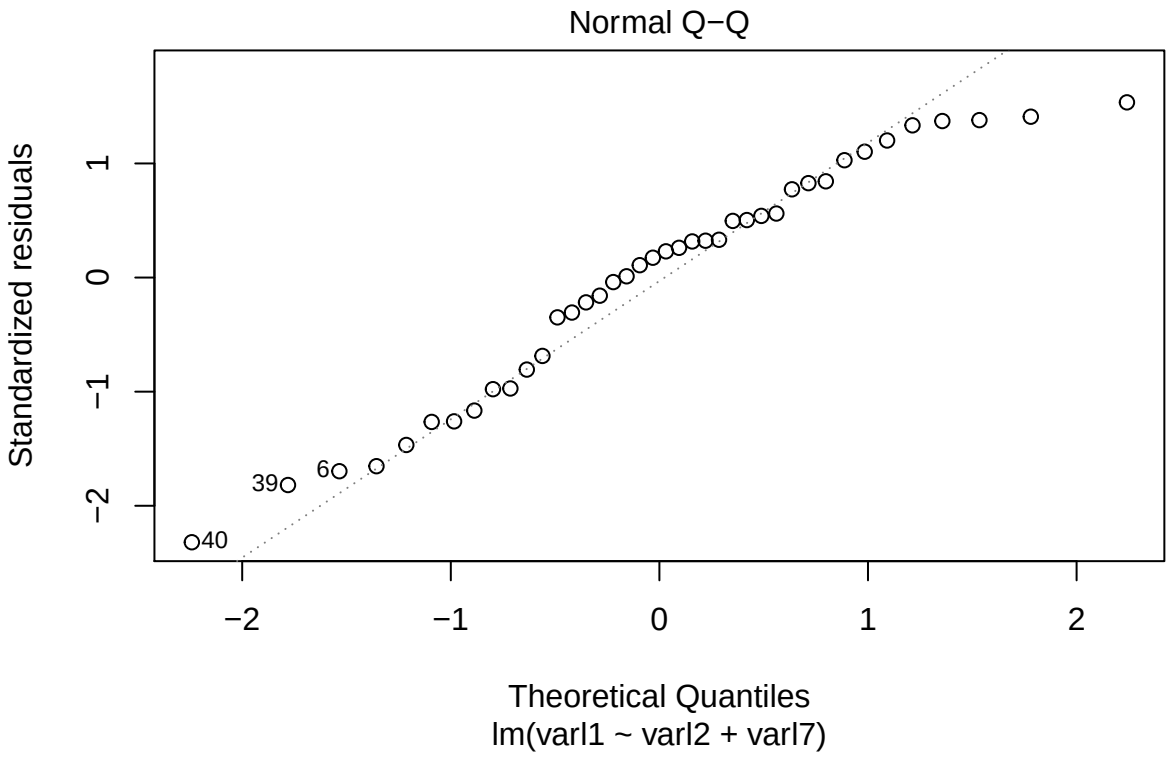
```
options(warn=-1)
varl1<-log(var1)
varl2<-log(var2)
varl7<-log(var7)
mod2<-lm(varl1~varl2+varl7)
summary(mod2)
```

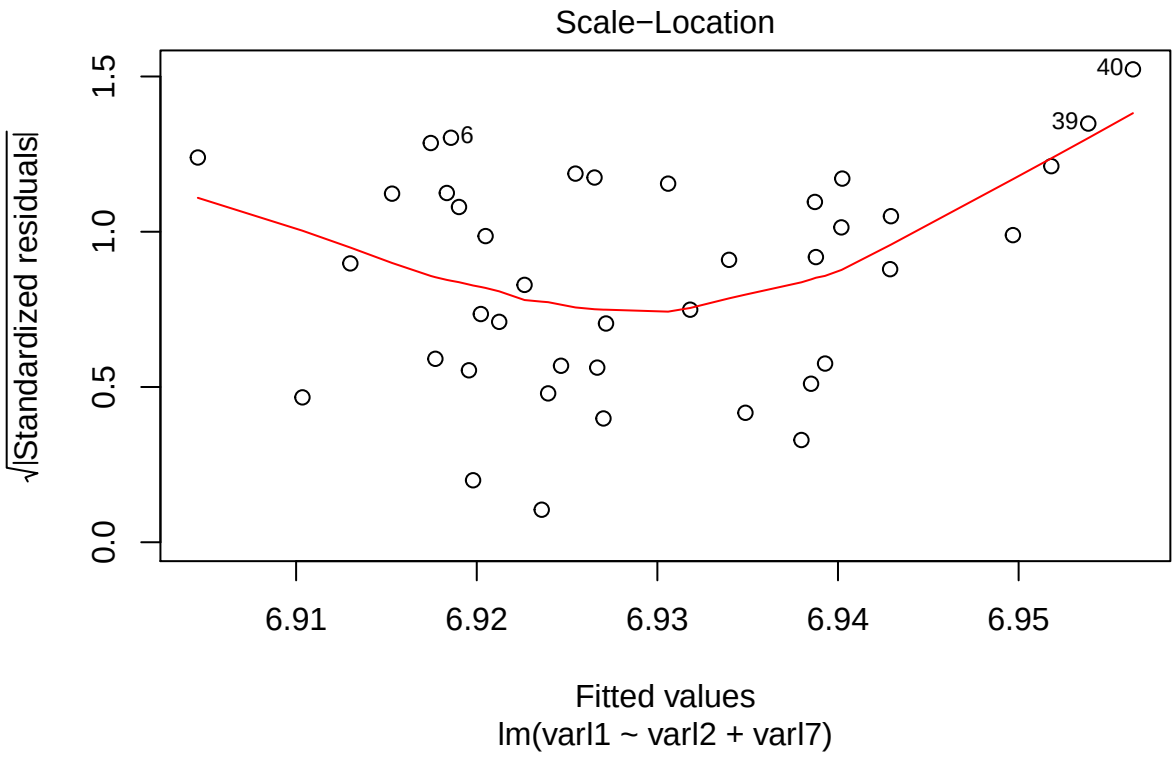
```
##
## Call:
## lm(formula = varl1 ~ varl2 + varl7)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0055150 -0.0023646  0.0005776  0.0022177  0.0041958
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  6.9808987  0.0052579 1327.707   <2e-16 ***
## varl2       -0.0235795  0.0009966  -23.659   <2e-16 ***
## varl7       -0.0018388  0.0009018   -2.039   0.0486 *
## ---
```

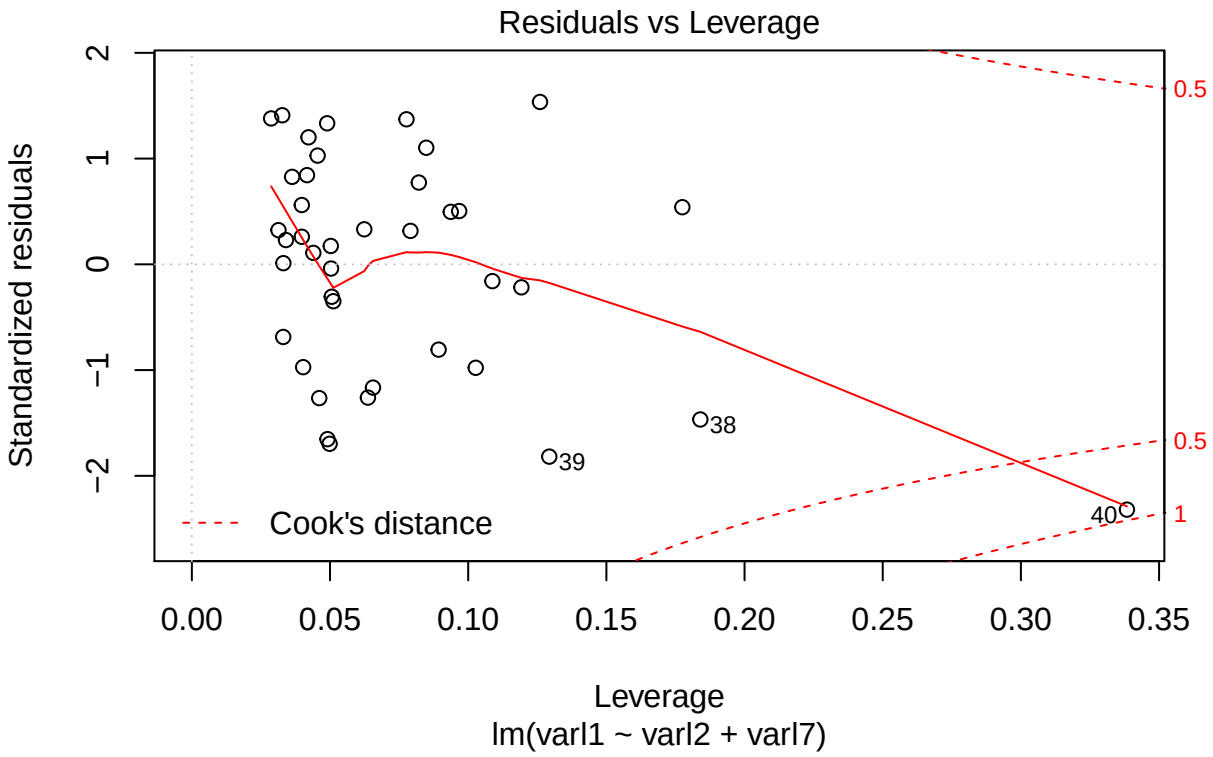
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002923 on 37 degrees of freedom
## Multiple R-squared:  0.9498, Adjusted R-squared:  0.9471
## F-statistic: 350.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

```
plot(mod2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(varl1 ~ varl2 + varl7)
40
39
6
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
6.91
6.92
6.93
6.94
6.95
Fitted values
lm(varl1 ~ varl2 + varl7)
40
39
6




Luego de realizar un ajuste en el modelo original podemos observar que los valores se ajustan mas a la linealidad y aceptacion del modelo de regresión múltiple.