Regresión

Jose Luis Vicente Villardon

9/10/2019

REGRESION

Primer ejemplo

The Federal Trade Commission cada año analiza distintas marcas de tabaco de acuerdo a su contenido en alquitrán, nicotina y monóxido de carbono. The United States Surgeon General considera cada una de estas sustancias peligrosas para la salud de los fumadores. Estudios pasados ponen de manifiesto que incrementos en el contenido de alquitrán y nicotina de los cigarros vienen acompañados por incrementos en el monóxido de carbono emitido al fumar un cigarro. Los datos que se presentan aquí fueron tomados de Mendenhall and Sincich (1992): Para más información ver el artículo "Using Cigarette Data for an Introduction to Multiple Regression" by Lauren McIntyre in Volume 2, Number 1, of the Journal of Statistics Education (1994).

Puede conseguir el artículo y los datos en: http://www.amstat.org/publications/jse/jse_archive.html

En primer lugar vamos a importar los datos desde el archivo de Excel. Puede hacerlo directamente en RStudio o correr las instrucciones siguientes teniendo en cuenta que hay que adaptar el directorio donde se encuentran los datos.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.5.2
```

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/0 Curso CSIC/Primera parte")
ciga <- read_excel("ciga.xls")
```

En primer lugar podemos realizar una descriptiva básica de las distintas variables.

```
summary(ciga)
```

```
##     Marca                Alq              Nic              Pes        
##  Length:25          Min.   : 1.00   Min.   :0.1300   Min.   :0.7851  
##  Class :character   1st Qu.: 8.60   1st Qu.:0.6900   1st Qu.:0.9225  
##  Mode  :character   Median :12.80   Median :0.9000   Median :0.9573  
##                     Mean   :12.22   Mean   :0.8764   Mean   :0.9703  
##                     3rd Qu.:15.10   3rd Qu.:1.0200   3rd Qu.:1.0070  
##                     Max.   :29.80   Max.   :2.0300   Max.   :1.1650  
##      Light              CO        
##  Min.   :0.00   Min.   : 1.50  
##  1st Qu.:0.00   1st Qu.:10.00  
##  Median :0.00   Median :13.00  
##  Mean   :0.28   Mean   :12.53  
##  3rd Qu.:1.00   3rd Qu.:15.40  
##  Max.   :1.00   Max.   :23.50
```

Observamos que la variable Light se ha imporatdo como numérica. Antes de proceder con el análisis convertiremos esta variable en nominal, en el lenguaje de R, la convertiremos en un factor. Repetimos la descriptiva.

```r
ciga$Light=factor(ciga$Light)
levels(ciga$Light)=c("no", "si")
summary(ciga)
```

```
##     Marca                Alq              Nic              Pes
##  Length:25          Min.   : 1.00   Min.   :0.1300   Min.   :0.7851
##  Class :character   1st Qu.: 8.60   1st Qu.:0.6900   1st Qu.:0.9225
##  Mode  :character   Median :12.80   Median :0.9000   Median :0.9573
##                     Mean   :12.22   Mean   :0.8764   Mean   :0.9703
##                     3rd Qu.:15.10   3rd Qu.:1.0200   3rd Qu.:1.0070
##                     Max.   :29.80   Max.   :2.0300   Max.   :1.1650
##  Light         CO
##  no:18   Min.   : 1.50
##  si: 7   1st Qu.:10.00
##          Median :13.00
##          Mean   :12.53
##          3rd Qu.:15.40
##          Max.   :23.50
```

Observamos que ahora ya está en la forma adecuada ya que el resumen es una tabla de frecuencias.

Regresión simple

Comenzamos haciendo cada una de las regresiones simples de las variables por separado, Alquitran, Nicotina y Peso. La variable Light es nominal y, de momento, no la incluiremos en la regresión. El modelo es de la forma $Y = \beta_0 + \beta_1 X_1$.

Regresión para el Alquitrán

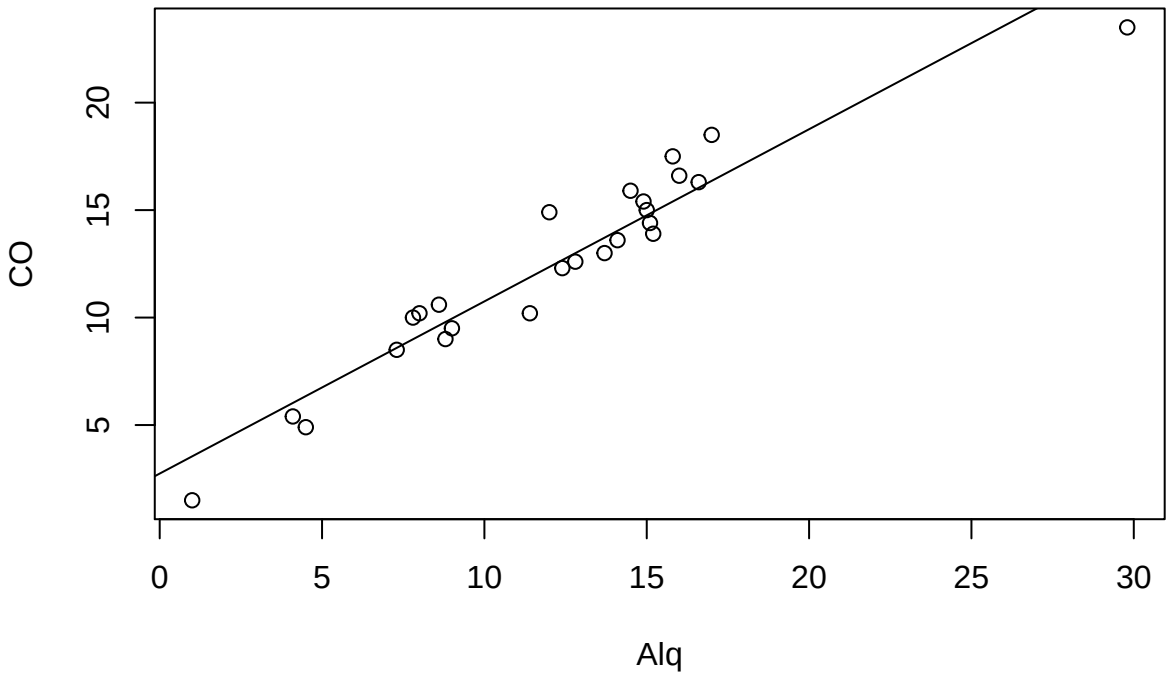
Comenzamos con el alquitrán, es decir $CO = \beta_0 + \beta_1 Alq$. Estimamos el modelo y lo interpretamos.

```r
fit1=lm(CO~Alq, data=ciga)
summary(fit1)
```

```
##
## Call:
## lm(formula = CO ~ Alq, data = ciga)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1124 -0.7167 -0.3754  1.0091  2.5450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74328    0.67521   4.063 0.000481 ***
## Alq          0.80098    0.05032  15.918 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.397 on 23 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9132
## F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

La ecuación de regresión es $CO = 2.74328 + 0.80098 * Alquitrán$ y el coeficiente de determinación es $R^2 = 0.9168$. El 91.68% de la variabilidad del monóxido de carbono está explicado por el modelo de regresión, es decir, tiene un poder explicativo alto. Podemos dibujar el diagrama de dispersión con la correspondiente recta ajustada.

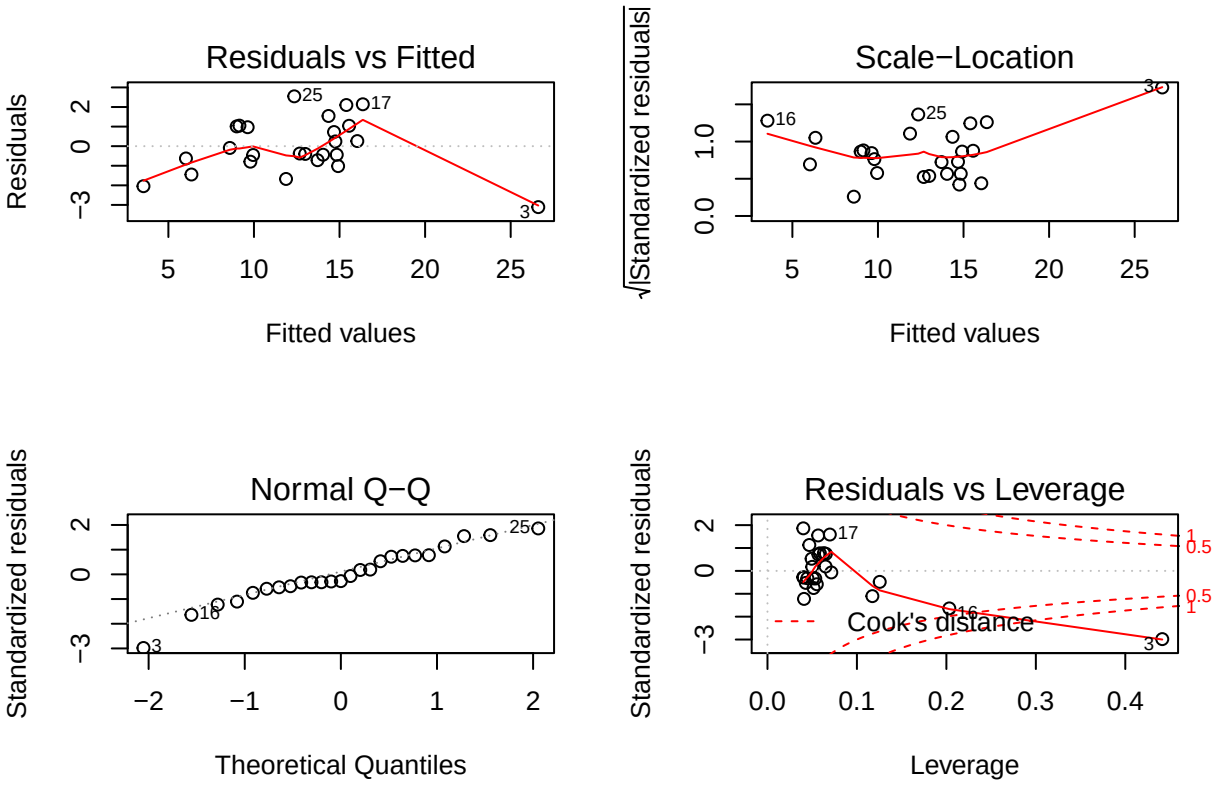
```
plot(CO~Alq, data=ciga)
abline(a=2.74328, b=0.80098)
```



La recta parece ajustarse bien a los datos. Solamente el punto de la derecha parece estar fuera del patrón de los demás, es lo que podríamos denominar un *outlier*. El modelo exlica una parte significativa de la variabilidad de la respuesta ($p = 6.552e - 14$). El coeficiente de regresión es significativamente distinto de cero ($p = 6.552e - 14$). En este caso ambos contrastes tienen el mismo $p - valor$ ya que tenemos solamente una variable.

Probamos algunos diagnósticos para comprobar si el modelo se encuentra en buenas condiciones.

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit1)
```



De nuevo parece que el punto 3 presenta algunos problemas y podría ser un outlier. Dejamos como ejercicio ajustar la regresión sin ese punto.

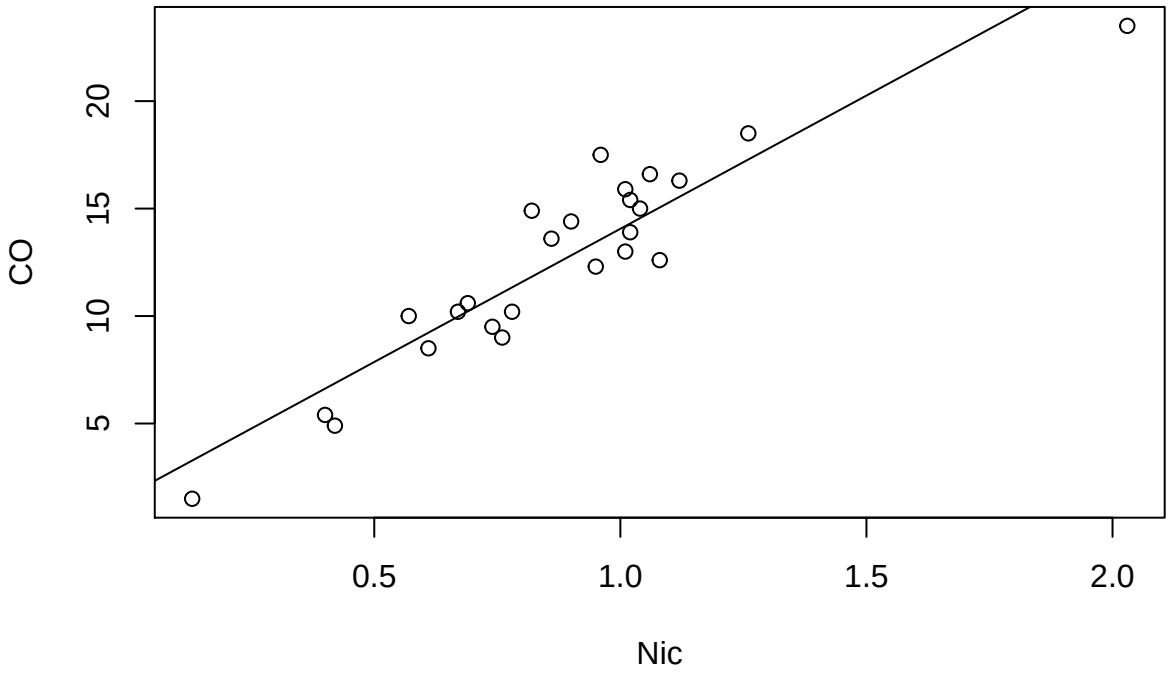
Regresión para la Nicotina

Ajustamos la regresión para la nicotina. Los comentarios se dejan al lector.

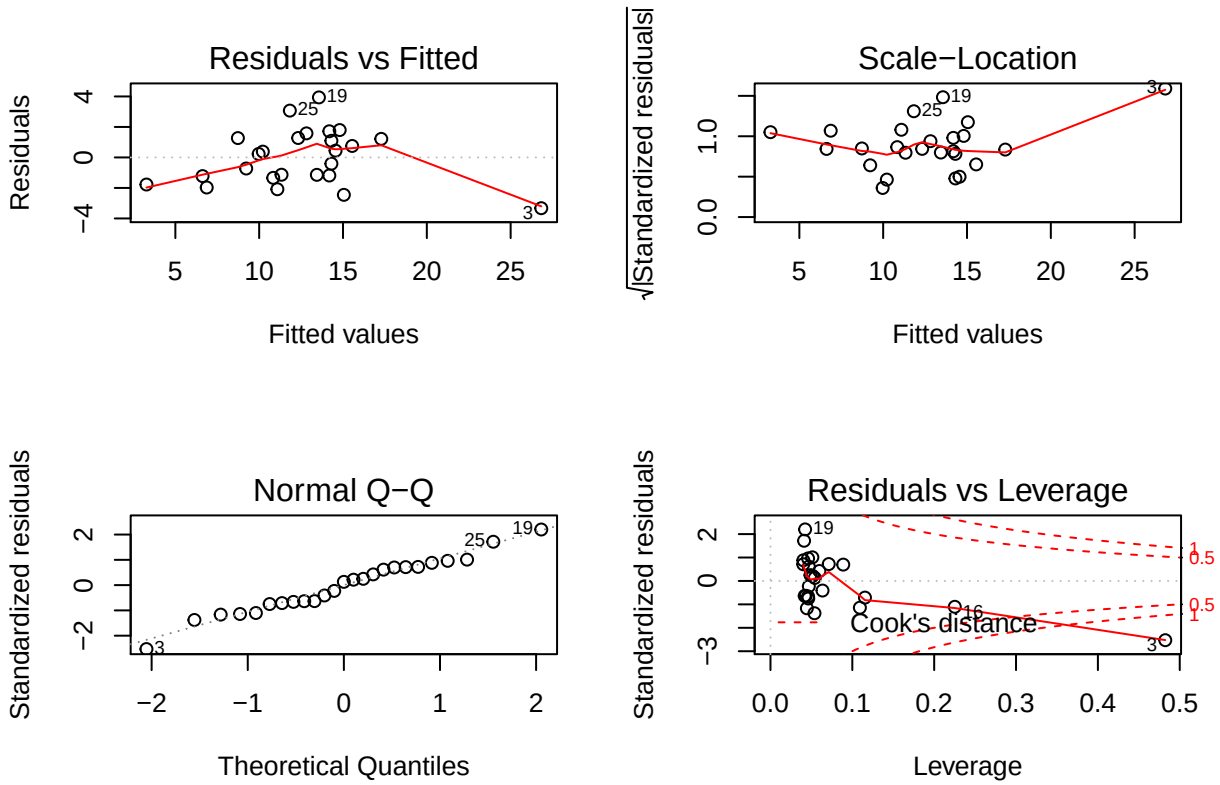
```
fit2=lm(CO~Nic, data=ciga)
summary(fit2)
```

```
##
## Call:
## lm(formula = CO ~ Nic, data = ciga)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3273 -1.2228  0.2304  1.2700  3.9357
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6647     0.9936   1.675    0.107
## Nic          12.3954     1.0542  11.759 3.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.828 on 23 degrees of freedom
## Multiple R-squared:  0.8574, Adjusted R-squared:  0.8512
## F-statistic: 138.3 on 1 and 23 DF,  p-value: 3.312e-11
```

```
plot(CO~Nic, data=ciga)
abline(a=1.6647, b=12.3945)
```




```
layout(matrix(c(1,2,3,4),2,2))
plot(fit2)
```



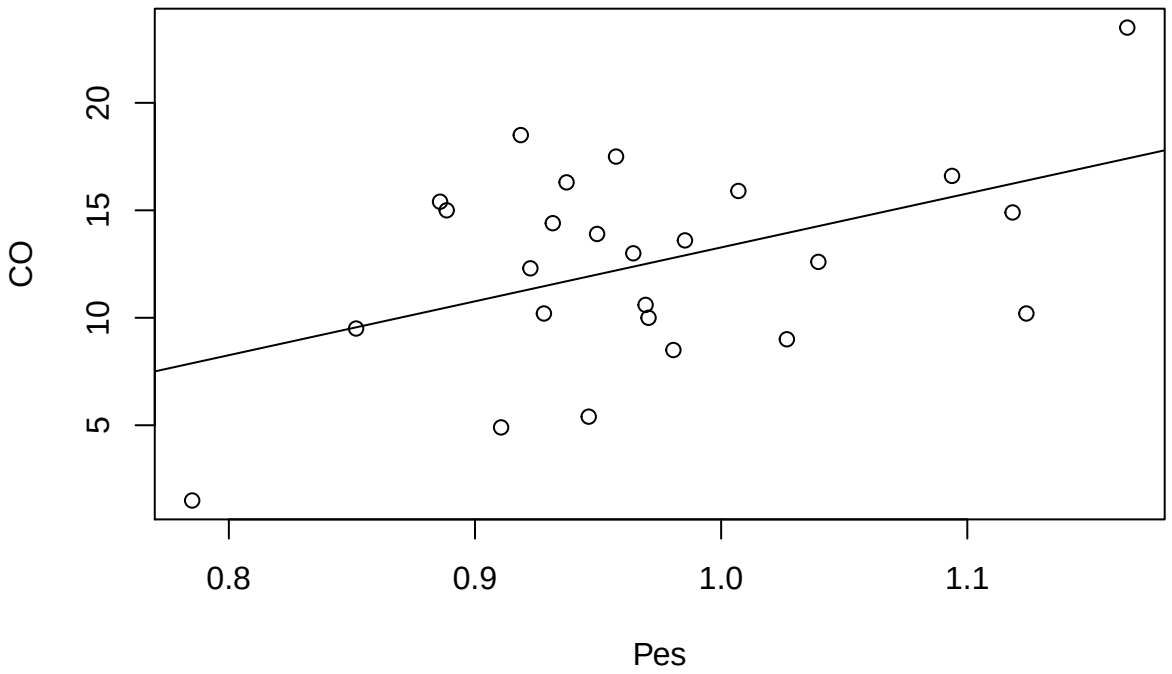
La relación con la nicotina tammbién es importante aunque su magnitud es menor.

Regresión para el Peso

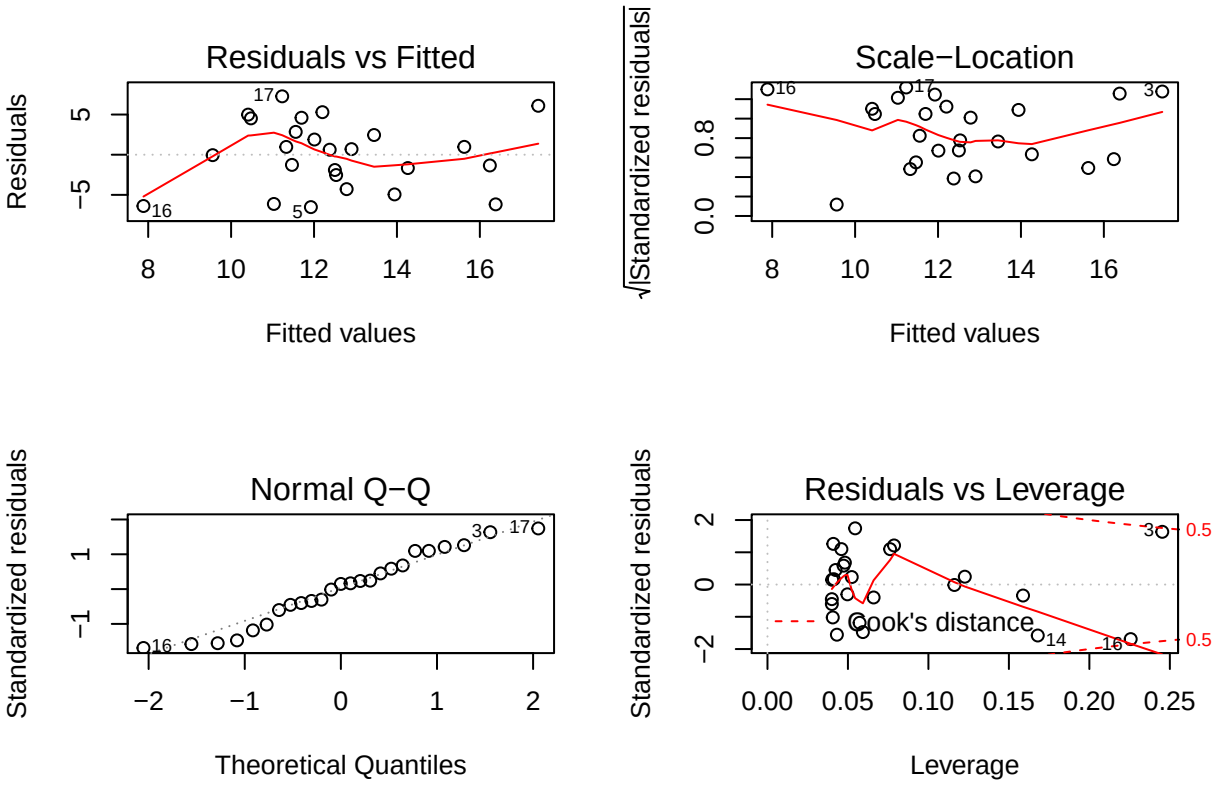
```
fit3=lm(CO~Pes, data=ciga)
summary(fit3)
```

```
##
## Call:
## lm(formula = CO ~ Pes, data = ciga)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.524 -2.533  0.622  2.842  7.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.795      9.722  -1.213   0.2373
## Pes           25.068      9.980   2.512   0.0195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.289 on 23 degrees of freedom
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.1811
## F-statistic: 6.309 on 1 and 23 DF,  p-value: 0.01948
```

```
plot(CO~Pes, data=ciga)
abline(a=-11.795, b=25.068)
```




```
layout(matrix(c(1,2,3,4),2,2))
plot(fit3)
```



Se trata de la regresión más débil de tiodas.

Regresión Multiple

Ajustamos ahora un modelo con todas la variables a la vez, es decir un modelo deregresión múltiple : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ ó $CO = \beta_0 + \beta_1 * Alq + \beta_2 * Nic + \beta_3 * X_3 * Pes$. Eldesarrollo de un modelo de regresión multiple es similar al de regresión simple con algunas adaptaciones en la interpretación de los parámetros.

```
fit4=lm(CO~Alq+Nic+Pes, data=ciga)
summary(fit4)
```

```
##
## Call:
## lm(formula = CO ~ Alq + Nic + Pes, data = ciga)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89261 -0.78269  0.00428  0.92891  2.45082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2022     3.4618   0.925 0.365464
## Alq           0.9626     0.2422   3.974 0.000692 ***
## Nic          -2.6317     3.9006  -0.675 0.507234
## Pes          -0.1305     3.8853  -0.034 0.973527
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.446 on 21 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.907
## F-statistic: 78.98 on 3 and 21 DF,  p-value: 1.329e-11
```

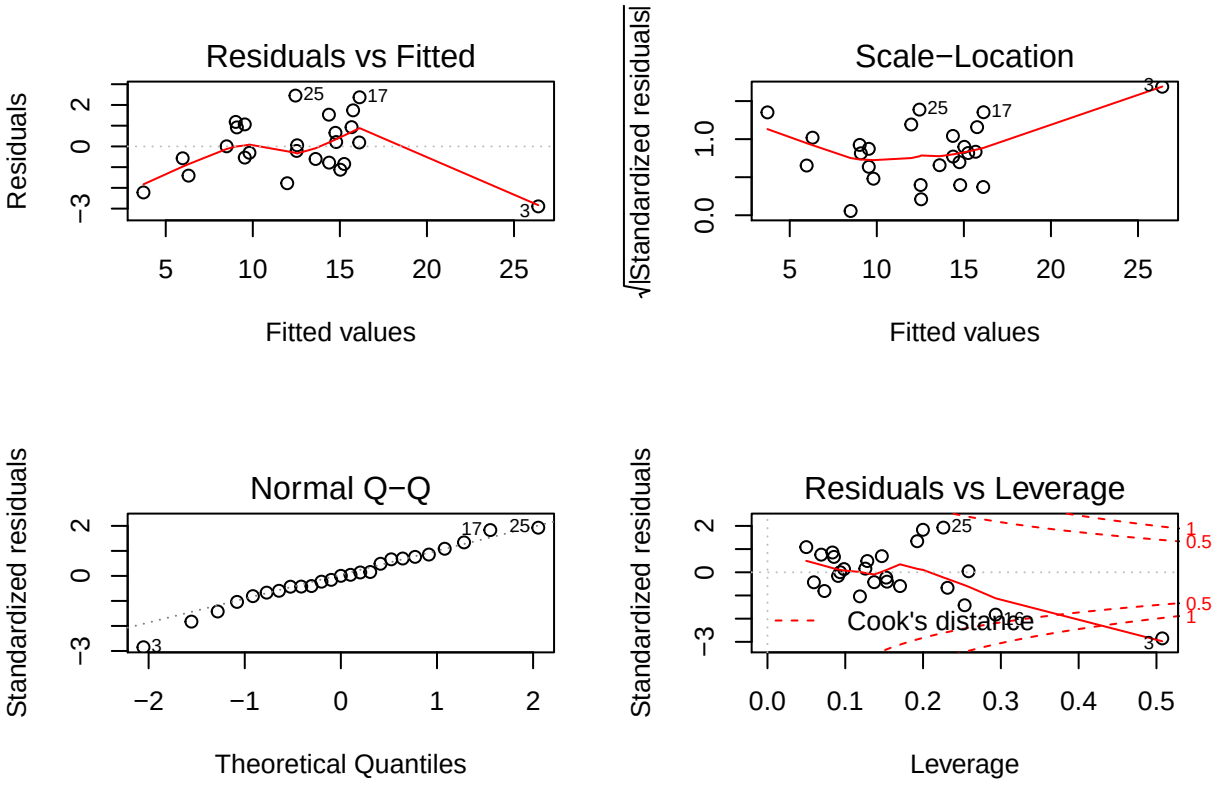
El modelo ajustado es $CO = 3.2022 + 0.9626Alq - 2.6317 * Nic - 0.1305 * X_3 * Pes$. El coeficiente de determinación es $R^2 = 0.9186$, es decir, el 91.86% de la variabilidad del CO es explicado por el modelo, el análisis de la varianza es signiificativo $p = 1.329e - 11$. De todos los coeficientes del modelo solamente el correspondiente al alquitrán es significativamente distinto de cero ($p = 0.000692$). Esto no quiere decir que el resto de las variables no estén relacionadas con el CO (como ya vimos) sino que no aportan información adicional a la que ya tiene el alquitrán. Esto se debe a que los predictores están fuertemente relacionados entre si y no es posible separar sus efectos. Para verlo mostramos la matriz de correlaciones entre ellos. Observese que el porcentaje explicado con el alquitrán sólo, es del 91.68% mientras que con todas es 91.86%, aumenta muy poco.

```
cor(ciga[,2:4])
```

```
##           Alq       Nic       Pes
## Alq 1.0000000 0.9766076 0.4907654
## Nic 0.9766076 1.0000000 0.5001827
## Pes 0.4907654 0.5001827 1.0000000
```

La correlación entre alquitrán y nicotina es bastante elevada. Dibujamos los diagnósticos.

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit4)
```



De nuevo el punto 3 presenta problemas. Dejamos como ejercicio ajustar el modelo sin el punto 3.

Introducción de una variable nominal

Supongamos que nos quedamos con el alquitrán y queremos introducir la vat¡riable light que es nominal. En lugar de ajustar el modelo conjunto $Y = \beta_0 + \beta_1 X$ podemos ajustar el modelo $Y = \beta_0 + \beta_1 X + \delta D$ donde la variable D se define de la siguiente manera - 1 Si el cigarrillo es light y 0 si el cigarrillo no lo es. La variable D se dice que es una variable ficticia ya que no ha sido medida directamente de esta forma. Veamos como la variable ficticia soluciona el problema.

La interpretación de los modelos en los que se han incluido variables ficticias es simple. Calculemos el modelo en cada uno de los grupos.

En el grupo de los normales $(D = 0)$ $Y = \beta_0 + \beta_1 X + \delta 0 = \beta_0 + \beta_1 X$ En el grupo de los light $(D = 1)$ $Y = \beta_0 + \beta_1 X + \delta 1 = (\beta_0 + \delta) + \beta_1 X$

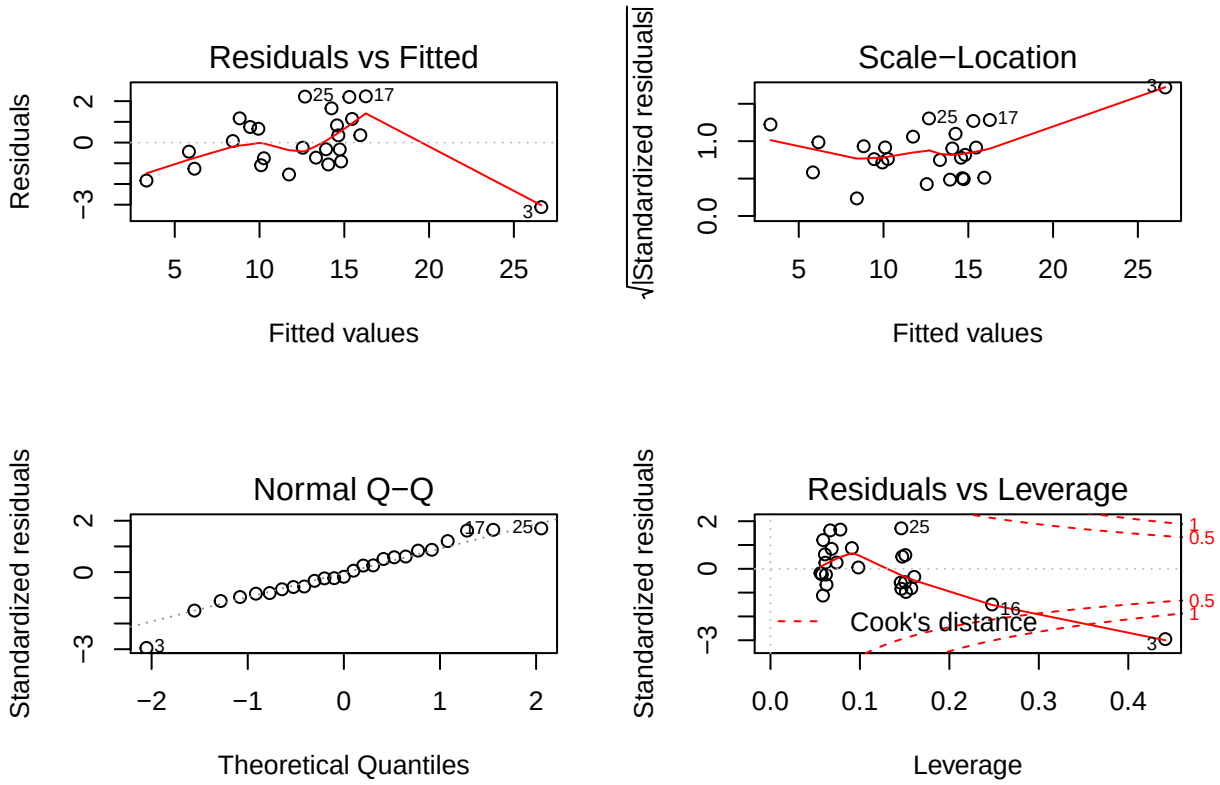
Luego β_1 es la pendiente (común) de los modelos para ambos grupos. β_0 es la constante en el modelo para el grupo de los normales, $(\beta_0 + \delta)$ es la constante en el modelo para el grupo de los light. Entonces δ es la diferencia entre el CO de los normales y los light, sea cual sea el nivel de alquitrán El contraste de igualdad a cero de δ es el contraste de que no hay diferencias en el nivel de CO entre los dos grupos.

R entiende lo que tiene que hacer cuando encuentra una variable nominal.

```
fit5=lm(CO~Alq+Light, data=ciga)
summary(fit5)
```

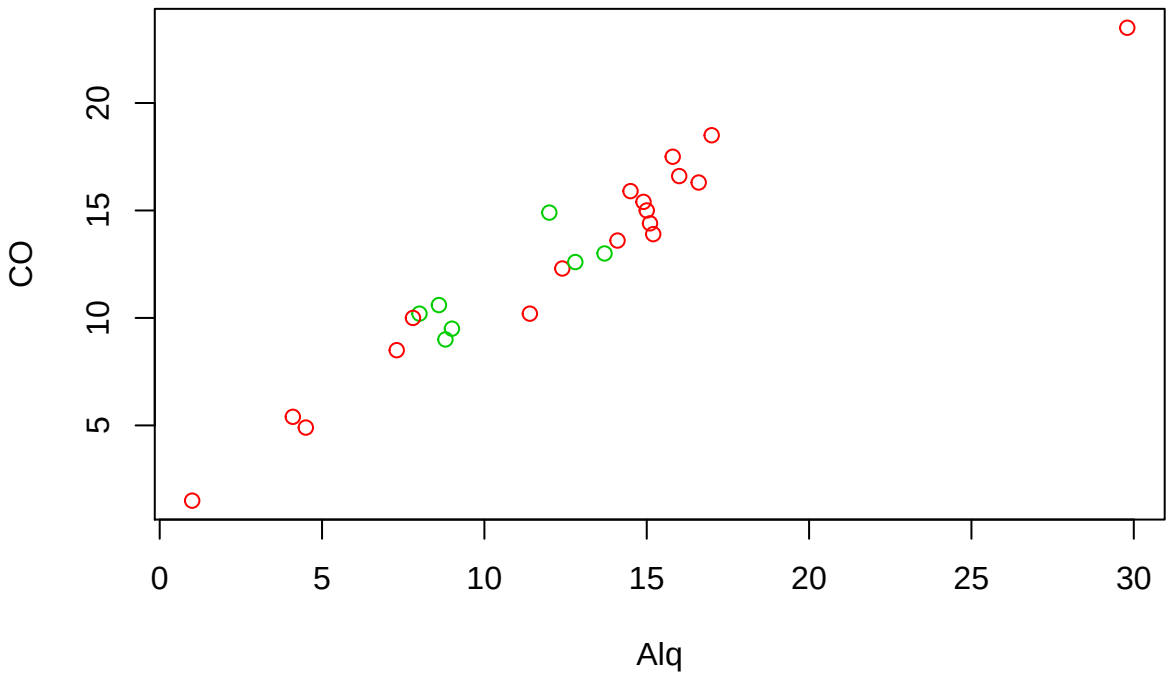
```
##
## Call:
## lm(formula = CO ~ Alq + Light, data = ciga)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1159 -0.9126 -0.2490  0.8299  2.2322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.52423    0.74897   3.370  0.00276 **
## Alq          0.80845    0.05195  15.563 2.33e-13 ***
## Lightsi      0.45638    0.64225   0.711  0.48480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.412 on 22 degrees of freedom
## Multiple R-squared:  0.9186, Adjusted R-squared:  0.9112
## F-statistic: 124.2 on 2 and 22 DF,  p-value: 1.033e-12
```

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit5)
```



El coeficiente correspondiente al tipo de cigarrillos no es significativo. No hay diferencias entre normal y light eliminado el efecto del alquitrán. L falta de relación se pone de manifiesto en el siguiente gráfico en el que se han coloreado los dos tipos d forma diferente.

```
plot(CO~Alq, data=ciga, col=as.numeric(ciga$Light)+1)
```

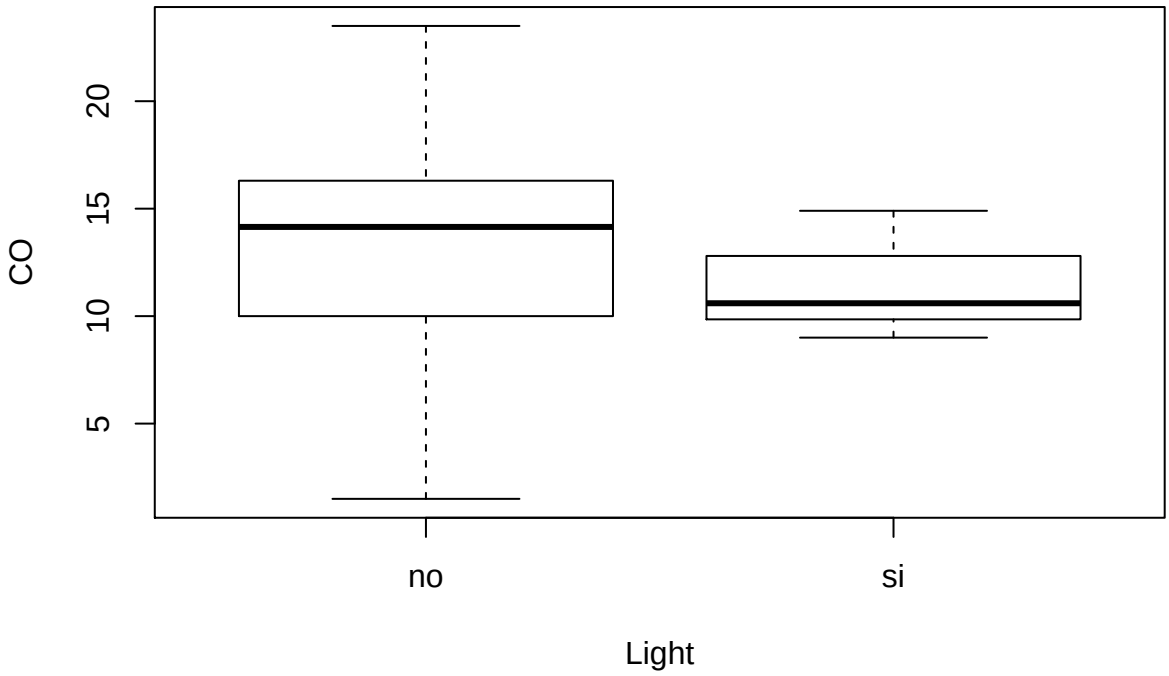


Si hacemos una regresión con una única variable nominal es exactamente equivalente a hacer un análisis de la varianza con un factor de variación.

```r
fit6=lm(CO~Light, data=ciga)
summary(fit6)
```

```
##
## Call:
## lm(formula = CO ~ Light, data = ciga)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4667  -2.4000   0.9333   2.9333  10.5333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.967      1.128  11.495 5.19e-11 ***
## Lightsi       -1.567      2.132  -0.735     0.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.786 on 23 degrees of freedom
## Multiple R-squared:  0.02294,    Adjusted R-squared:  -0.01954
## F-statistic: 0.5401 on 1 and 23 DF,  p-value: 0.4698
```

```r
plot(CO~Light, data=ciga)
```



No hay una diferencia estadísticamente signivicativa en el monóxido de carbono que producen los normales y los light.

Para el mismo propósito podemos usar también

```r
fit7=aov(CO~Light, data=ciga)
summary(fit7)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Light        1   12.4   12.37    0.54   0.47
```

```
## Residuals   23  526.8   22.90
```

que nos organiza mejor la tabla de análisis de la varianza.

El comando *anova* podemos utilizarlo para comparar modelos.

```
anova(fit4, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: CO ~ Alq + Nic + Pes
## Model 2: CO ~ Alq
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     21 43.893
## 2     23 44.869 -2   -0.9765 0.2336 0.7937
```

Como ya vimos, nicotina y peso no aportan nada al modelo que no explicara el alquitrán.

Segundo ejemplo

At the end of February 2002 the U.S. Senate considered comprehensive energy legislation. Senators John McCain and John Kerry proposed raising the Corporate Average Fuel Economy (CAFE) standard for cars and trucks. On March 13, 2002 the United States Senate voted on the Levin amendment (No. 2997), charging the National Highway Traffic Safety Administration with the development of a new standard and effectively shelving the McCain/Kerry proposal. The dataset consists of information about each of the 100 U.S. Senators regarding their vote on the Levin amendment. A senator's vote (Vote) is the response variable. Provided explanatory variables include the state represented, political party affiliation (Party), and the lifetime total amount of contributions received from auto manufacturers (Amount).

Leemos los datos

```
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.5.2
```

```
setwd("~/Library/Mobile Documents/com~apple~CloudDocs/0 Curso CSIC/Primera parte")
Votos=read.spss("Votos.sav", to.data.frame =TRUE)
```

```
## re-encoding from latin-9
```

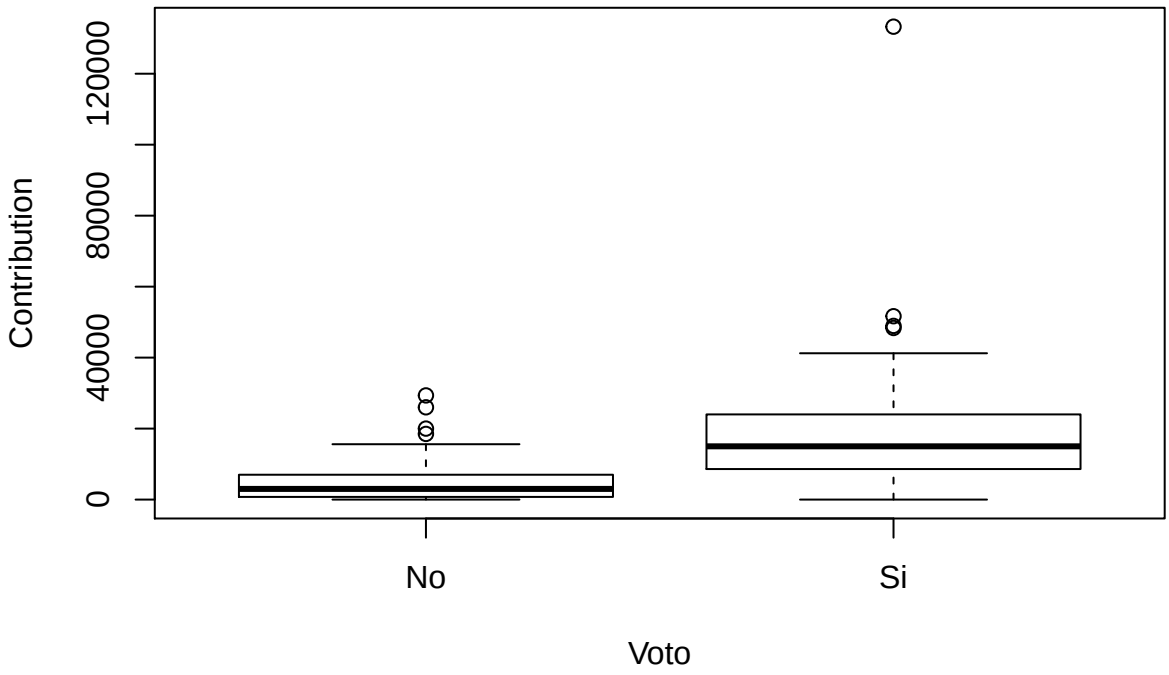
```
head(Votos)
```

```
##              Senator State     Partido Voto Contribution
## 1 Murkowski, Frank      AK Republicano   Si        19700
## 2 Stevens, Ted          AK Republicano   Si        13000
## 3 Sessions, Jeff        AL Republicano   Si         9500
## 4 Shelby, Richard       AL Republicano   Si        25000
## 5 Hutchinson, Tim       AR Republicano   Si         4900
## 6 Lincoln, Blanche      AR   Democrata   Si         5500
```

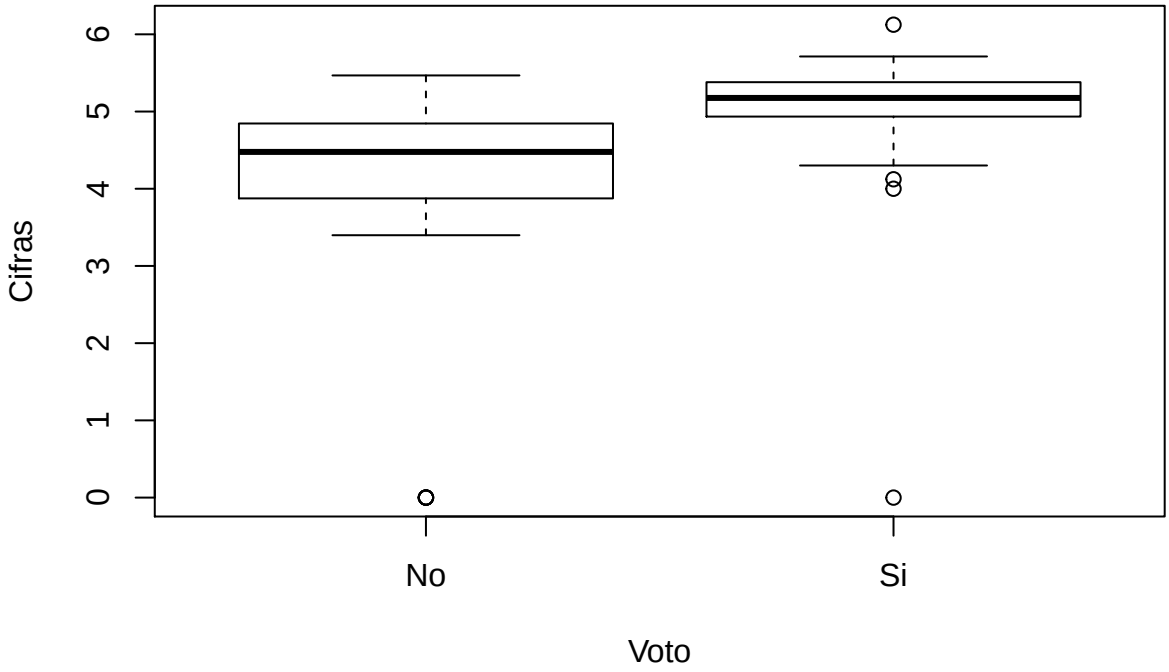
```
Votos$Cifras=log10(10*Votos$Contribution+1)
```

Añadimos una variable que es el número de cifras. Al tratarse de una cantidad monetaria, la distribución suele ser muy asimétrica, conviene transformar la variable con alguna transformación logarítmica. Se ha usado $log_{10}(10 * x + 1)$ que está relacionado con el número de cifras de la cantidad. Hacemos algunos análisis descriptivos antes de modelar los datos:

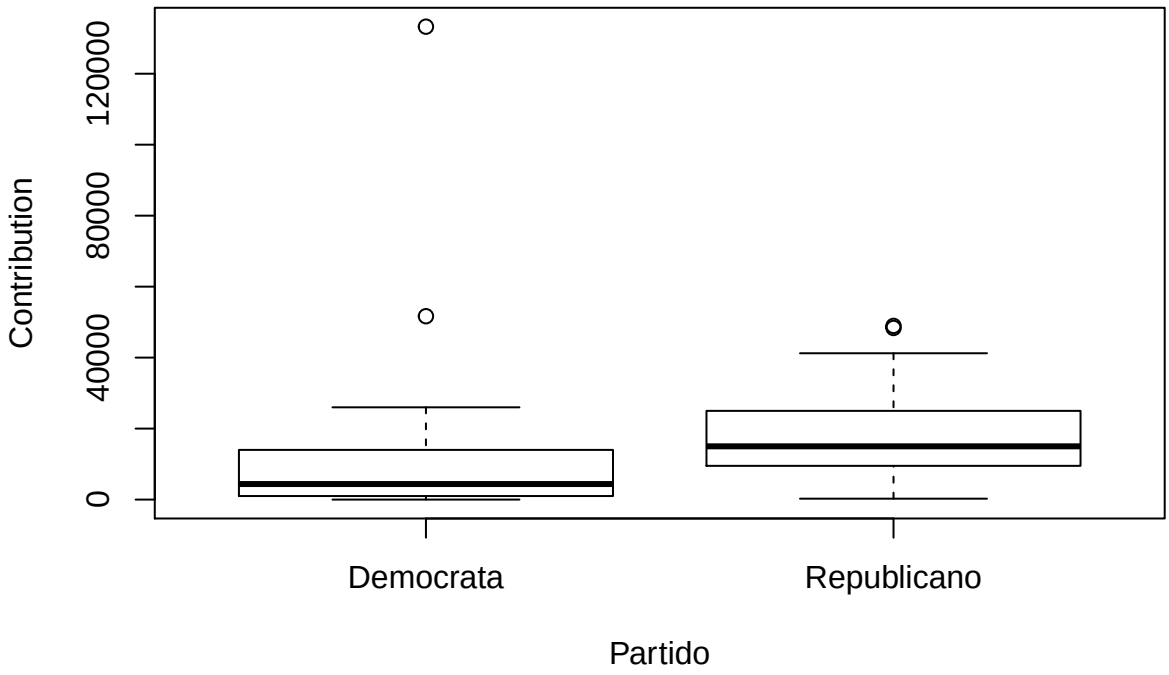
```
plot(Contribution~Voto, data=Votos)
```



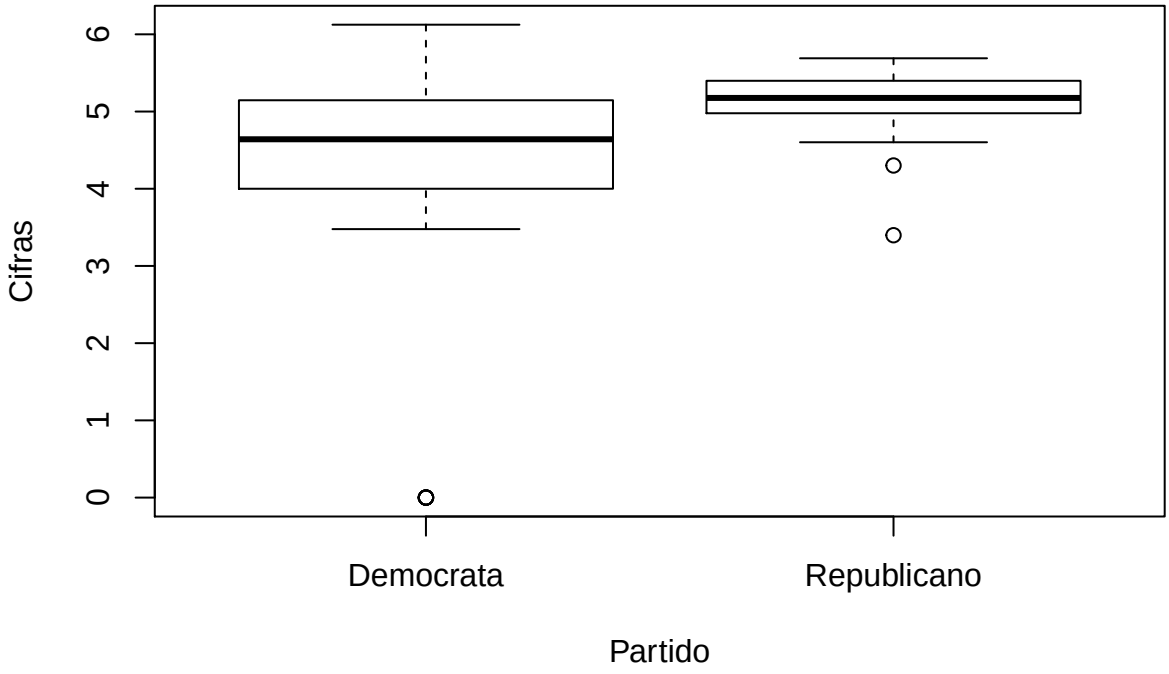

```
plot(Cifras~Voto, data=Votos)
```




```
plot(Contribution~Partido, data=Votos)
```




```
plot(Cifras~Partido, data=Votos)
```



Descripción básica

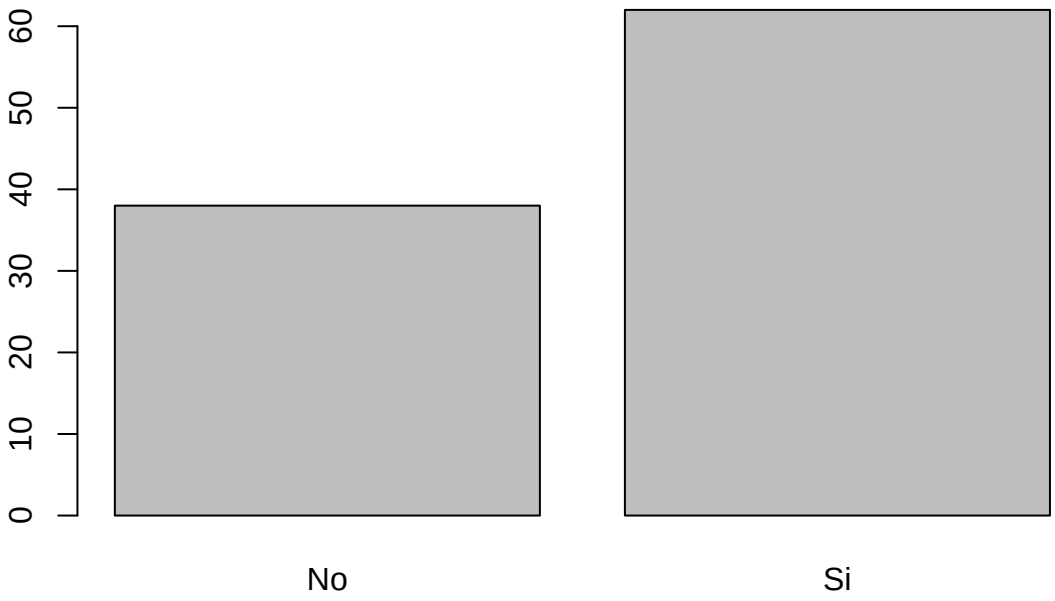
Pasamos a realizar una descripciómn básica de los datos. Por ejemplo, el número de votos favorables y desfavorables a la ley.

```
tv=table(Votos$Voto)
tv
```

```
##
```

```
## No Si
## 38 62
```

```
barplot(tv)
```

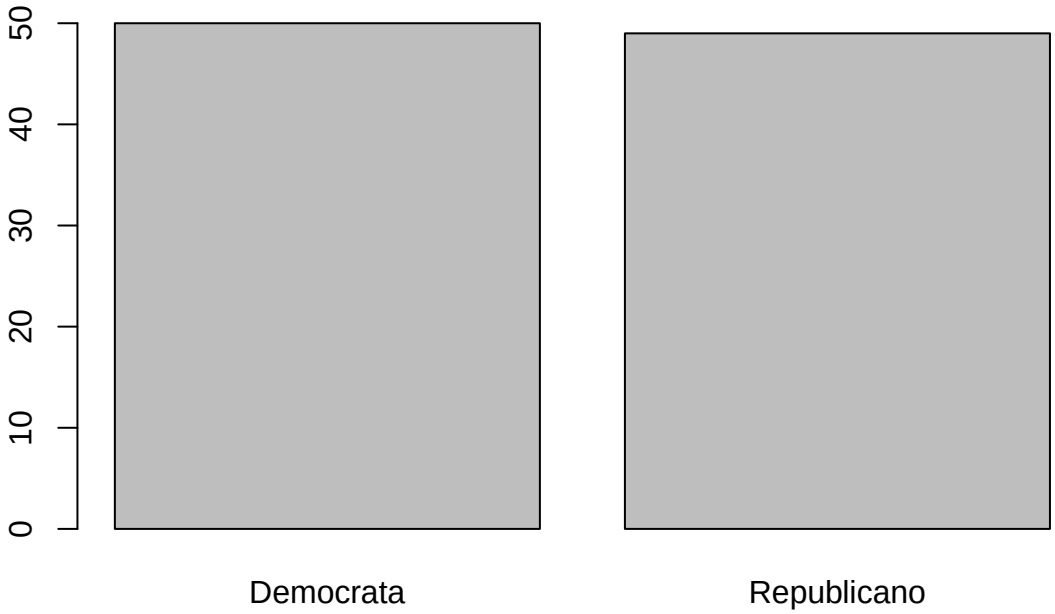


Tenemos 38 votos positivos y 62 negativos y los mismos porcentajes yaque hay 100 senadores.

```
tp=table(Votos$Partido)
tp
```

```
##
##   Democrata Republicano
##          50          49
```

```
barplot(tp)
```

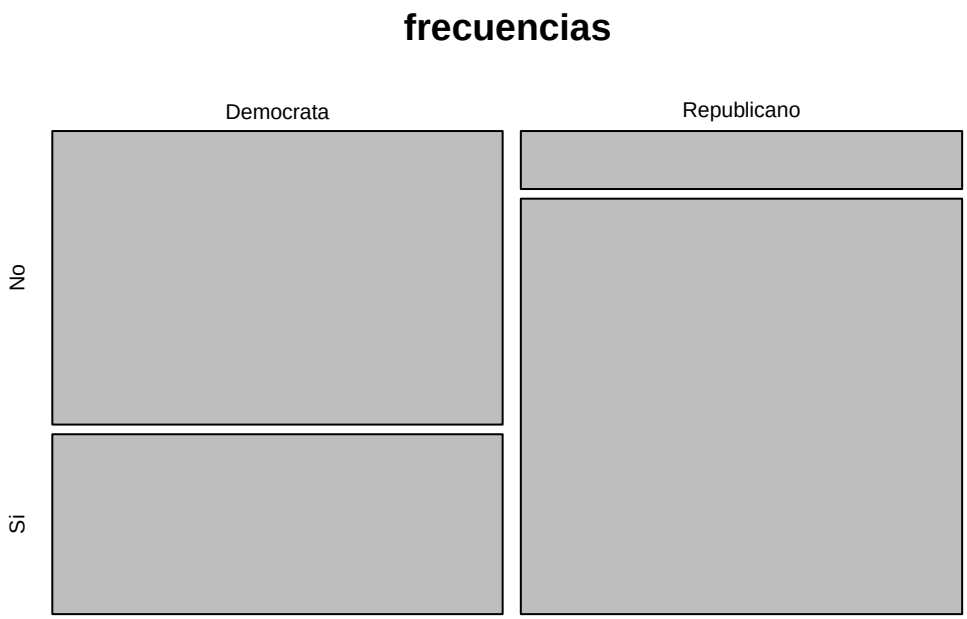


tenemos 50 senadores del partido Demócrata, 49 del Republicano y un independiente que lo hemos contabilizado como dato perdido. Podemos cruzar ambas variables:

```
frecuencias=table(Votos$Partido,Votos$Voto )
frecuencias
```

```
##
##                No Si
##   Democrata    31 19
##   Republicano   6 43
```

```
plot(frecuencias)
```



Podemos calcular los porcentajes por fila y los porcentajes por columna respectivamente.

```
prop.table(frecuencias, 1) #Porcentajes por filas
```

```
##
##                      No       Si
##   Democrata    0.620000 0.380000
##   Republicano 0.122449 0.877551
```

```
prop.table(frecuencias, 2) #Porcentajes por columnas
```

```
##
##                       No        Si
##   Democrata    0.8378378 0.3064516
##   Republicano 0.1621622 0.6935484
```

De los 50 demócratas, 19 votan afirmativamente mientras que de los 49 republicanos lo hacen, es decir votan positivamente el 38% de los demócratas y el 87% de los republicanos.

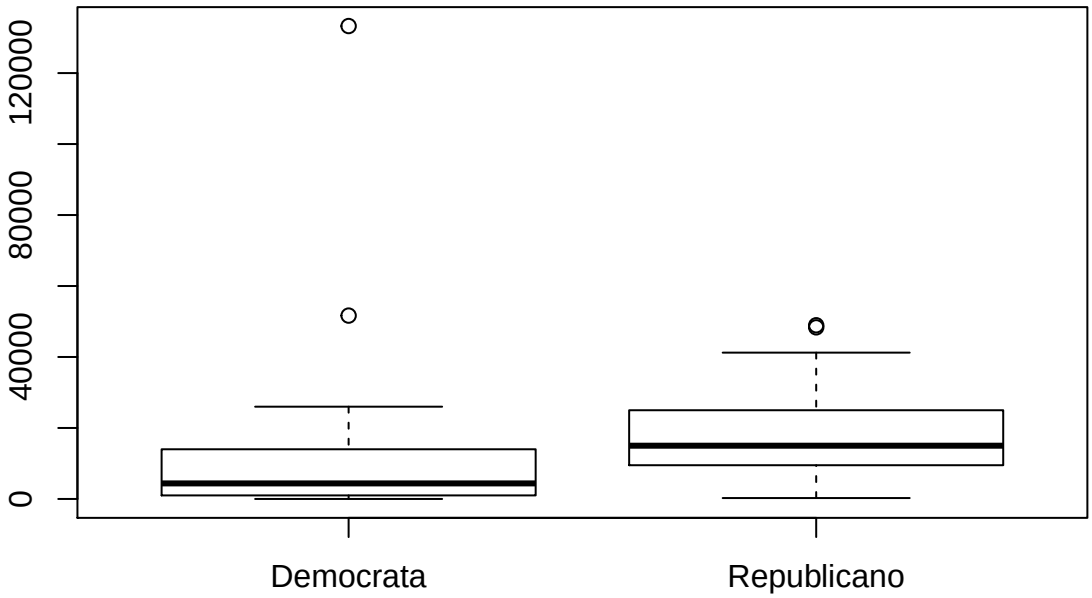
```
tapply(Votos$Contribution, Votos$Partido, mean)
```

```
##   Democrata Republicano
##    10025.28    17782.92
```

Los senadores demócratas reciben una media de 10025.28$ mientras que los republicanos reciben 17782.92$.

Gráficamente

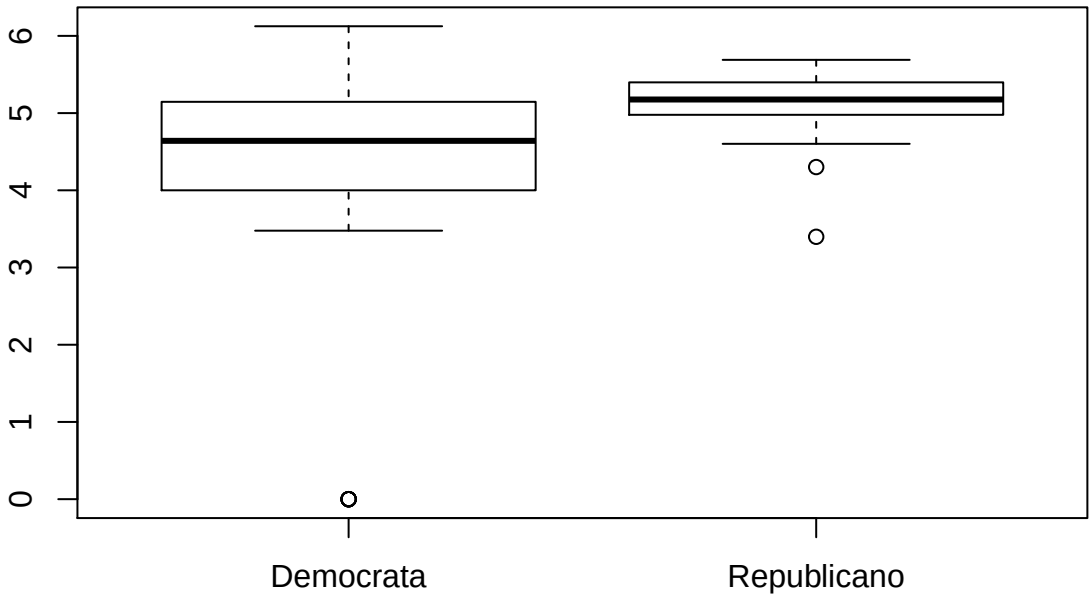
```r
boxplot(Contribution ~ Partido, data=Votos)
```



Observamos como la cantidad de dnero recibida es mayor en los republicanos. Observamos también que hay dos valores muy extremos en las cantidades recibidas por dos senadores del partido demócrata.

Con el número de cifras

```r
boxplot(Cifras ~ Partido, data=Votos)
```

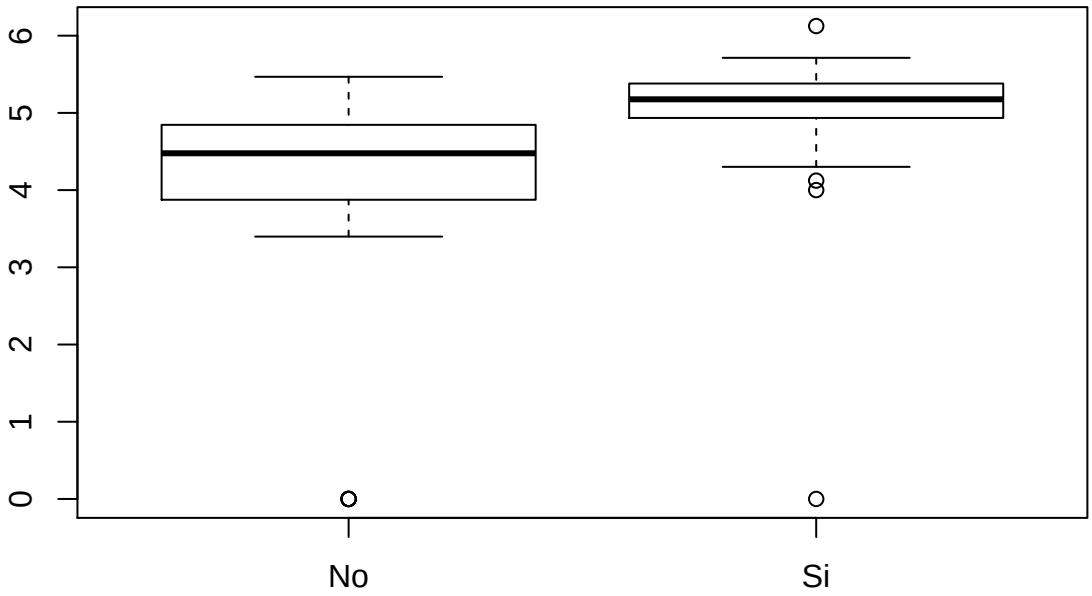


Vemos como se reduce la asimetría de las distribuciones.

Relación del voto con el resto de las variables: Regesión Logística

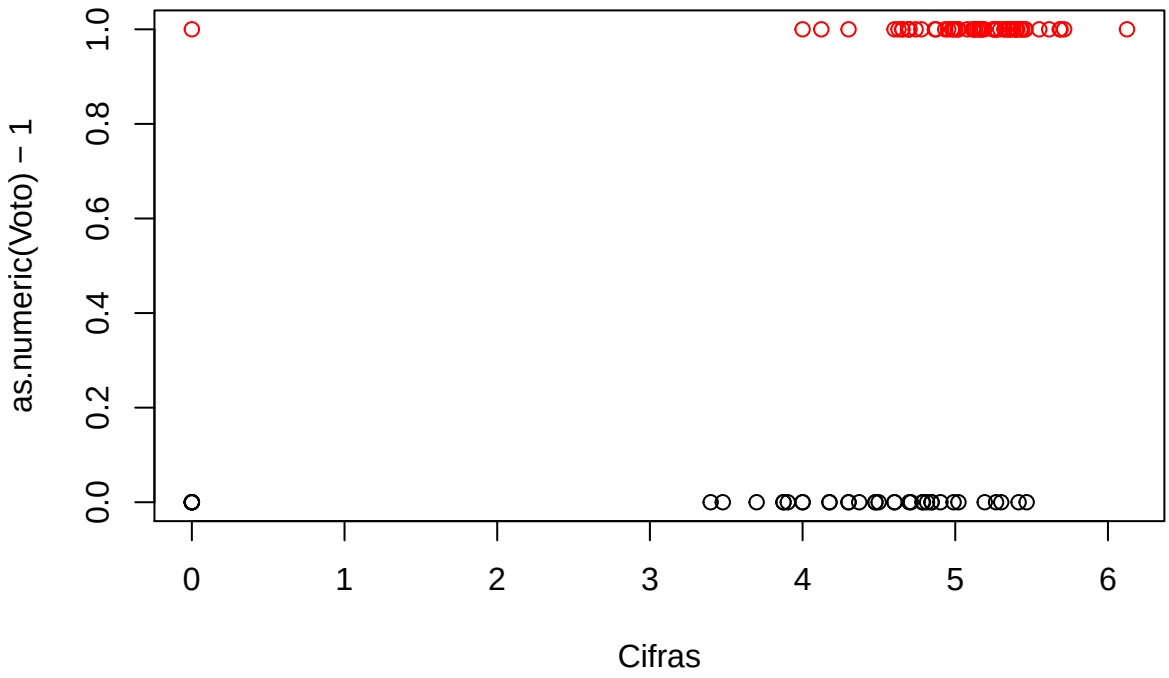
Dibujamos un box-‘plot del número de cifras y el voto.

```
boxplot(Cifras ~ Voto, data=Votos)
```



Observamos que hay una cierta relación entre ambos ya que los que votan positivo parece que recibieron mayores cantidades de dinero. Si fueran variables continuas, dibujaríamos un diagrama de dispersión con el voto en el eje Y (el voto lo pondremos como numérico con valores 0 y 1).

```
plot(as.numeric(Voto)-1 ~ Cifras, data=Votos, col=as.numeric(Voto))
```



Observamos cierta acumulación de votos positivos en la parte alta, para senadores que reciben mayores cantidades de dinero. La relación no puede describirse mediante una línea recta ya que esta nos daría valores negativos y mayores que 1 que no pueden ser probabilidades.

Vamos a tratar de ver como puede ser esa relación usando probabilidades o porcentajes. El número exacto de cifras puede calcularse como

```
Votos$Ncifras=floor(Votos$Cifras)
```

Podemos cruzarlo con el voto

```
tablecifras=table(Votos$Ncifras, Votos$Voto)
tablecifras
```

```
##
##     No Si
##   0  5  1
##   3  6  0
##   4 21 18
##   5  6 42
##   6  0  1
```

```
prop.table(tablecifras, 1)
```

```
##
##           No        Si
##   0 0.8333333 0.1666667
##   3 1.0000000 0.0000000
##   4 0.5384615 0.4615385
##   5 0.1250000 0.8750000
##   6 0.0000000 1.0000000
```

Observamos que el porcentaje de votos positivos aumenta con el número de cifras.

Modelo de regresión logística

Trataremos de modelar la probabilidad de voto positivo en función de la cantidad recibida y el partido al que pertenece cada uno de los senadores. Comenzaremos con la cantidad monetaria.

El modelo logístico es

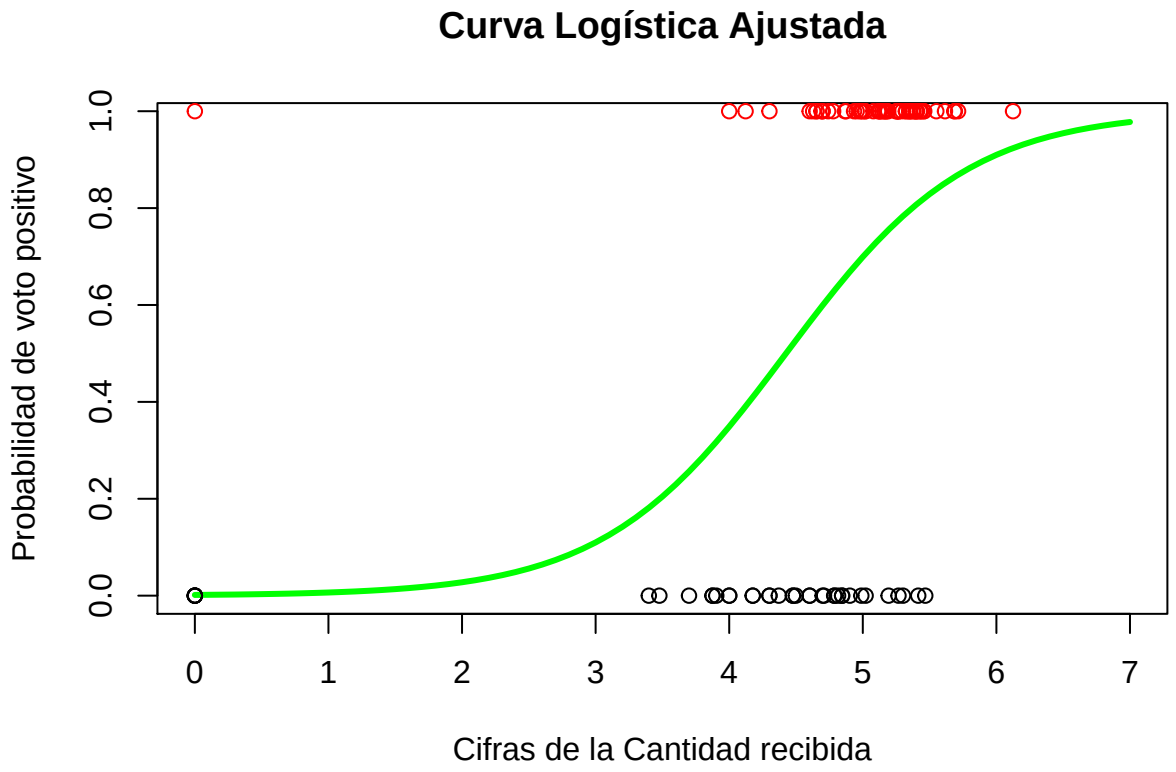
$$p_i = P\left(Y = 1/X = x_i\right) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

se trata de un modelo lineal en el *logit* de la probabilidad.

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i$$

Vamos adibujarla

```
voto=as.numeric(Votos$Voto)-1
x=seq(0,7,0.1)
linterm=-6.496 + 1.468*x
pvoto=exp(linterm)/(1+exp(linterm))
plot(x, pvoto, type="l", lwd=3, main = "Curva Logística Ajustada", xlab = "Cifras de la Cantidad recibi
points(Votos$Cifras, voto, col=voto+1)
```



Ajustamos el modelo

```
gfit1=glm(Voto~Cifras, data=Votos, family=binomial)
summary(gfit1)
```

```
##
## Call:
## glm(formula = Voto ~ Cifras, family = binomial, data = Votos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8587  -0.9518   0.6534   0.7808   3.6048
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.496      2.208  -2.942  0.00326 **
## Cifras         1.468      0.453   3.241  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 132.81  on 99  degrees of freedom
## Residual deviance: 109.09  on 98  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 5
```

El modeloestimado es

$$p_i = \frac{1}{1 + e^{-(-6.496+1.468x_i)}}$$

Por cada incremento de una unodad en el número de cifras se incrementa 1.468 unidades el *lpgit* de p_i. Además del coeficiente, a veces interpretamos también $e^{\beta_1} = e^{1.468} = 4.34115$ como un *odds-ratio*. Tratemos de explicarlo, supongamos que tenemos un senador que recibe un númeo de 4 cifras y llamamos p_4 a la probabilidad de votar positivo; el cociente $odds_4 = \frac{p_4}{(1-p_4)}$ se denomina *odds* o *ventaja* de voto positivo frente a negativo y representa el *número de veces* que es más probable el voto positivo que el negativo para un senador que recibe una contribución de cuatro cifras. De la misma forma, podríamos calcular la ventaja para 5 cifras $odds_5 = \frac{p_5}{(1-p_5)}$. El cociente entre los dos $OR_{5/4} = \frac{odds_5}{odds_4}$ se denomina *odds ratio* y es el número de veces que la ventaja de 5 cifras es mayor que la ventaja de 4 cifras. Si el denominador es mayor, el OR es menor que 1 y la ventaja es a favoir del denominador. En este caso, la ventaja de voto positivo frente a negativo es 4.34115 veces mayor para un senador que recibe 5 cifras frente a uno que recibe 4. Hay una asociación positiva, es decir, a medida que aumenta el número de cifras aumenta la probabilidad de voto positivo. Además el incremento es estadísticamente diferente de 0 ($p = 0.00119$).

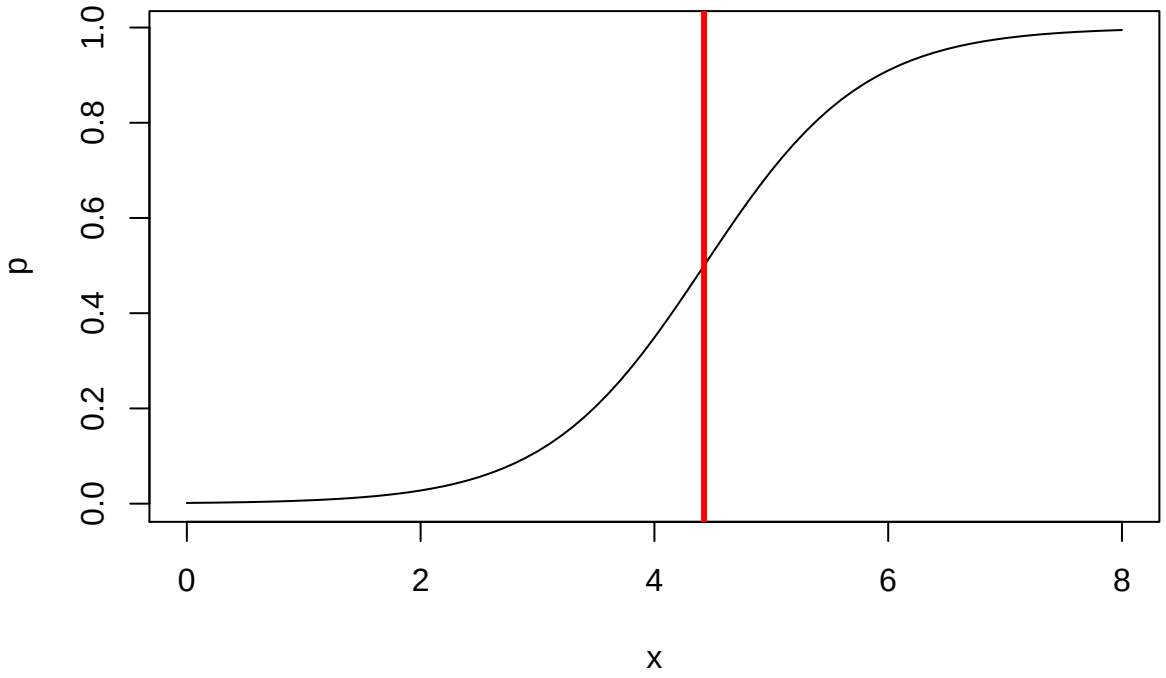
Porejemplo, para un senador que recibe 10000 dólares (5 cifras) la probabilidad esperada de voto positivo es

$$p_i = \frac{e^{-6,496+1,468x5}}{1 + e^{-6,496+1,468x5}} = 0,6993$$

Es decir, el 69.93% de los candidatos que recibe 10000$ votaría positivamente a la ley.

Podemos dibujar el modelo

```
x=seq(0, 8, 0.1)
z=- 6.496 + 1.468*x
p=exp(z)/(1+exp(z))
cut=-1*- 6.496/1.468
plot(x,p, type = "l")
abline(v=cut, col="red", lwd=3)
```



El número $\frac{-\beta_0}{\beta_1}$ es la cantidad a partir de la cual es más probable votar positivo que negativo. Se necesita un número de más de 4.425068 para que sea más probable votar positivo que negativo.

Podemos obtener un reumen más detallado

```r
summary(gfit1)
```

```
##
## Call:
## glm(formula = Voto ~ Cifras, family = binomial, data = Votos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8587  -0.9518   0.6534   0.7808   3.6048
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.496      2.208  -2.942  0.00326 **
## Cifras         1.468      0.453   3.241  0.00119 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 132.81  on 99  degrees of freedom
## Residual deviance: 109.09  on 98  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 5
```

A partir de estos resultados podemos obtener contrastes individuales para los parámetros. Por ejemplo para contrastar la hipótesis de que $H_0 : \beta_1 = 0$. El pvalor paar esta hipótesis es $p = 0.00119$. Es decir β_1 es significativamente distinto de cero o tenemos una relación estadísticamente significativa entre la proporción de voto positivo y la contribucón dineraria a la campaña del senador.

Por cada incremento de una unidad en el número de cifras el logit (logaritmo del odds) se incrementa en 1,468 unidades. A la cantidad $e^{1.468} = 4.341$ se le da una interpretación en términos de *odds-ratio*, a saber, un incremento de una unidad en el número de cifras supone un odds-ratio de 4,341, es decir, la ventaja de votar "si" frente a "no" es 4,341 veces mayor si la contribución tiene una cifra más. Por ejemplo, supongamos que un senador recibe una contribución de 4 cifras y llamemos p_5 a la probabilidad de voto positivo (para este modelo es 0.699307). El odds, o ventaja de voto positivo frente a negativo se define como $odds_5 = \frac{p_5}{1-p_5}$. Para este modelo $odds_5 = \frac{0.699307}{0.300693} = 2.325651$, es decir, para un senador que recibe 5 cifras es 2.325 veces más probable votar positivamente que hacerlo negativamente. Si recibe 6 cifras $odds_6 = \frac{0.909866}{0.09013399} = 10.09459$, la ventaja es 10.09459. El cociente de ambas ventajas se llama *odds-ratio*, en este caso $OR_{6/5} = \frac{odds_6}{odds_5} = \frac{10.09459}{2.325651} = 4.340544$, es decir, la ventaja de voto positivo frente a negativo es 4.340544 veces mayor con un número de 6 cifras que con uno de 5. Si calculamos el OR para cantidades que difieren en una unidad, obtenemos siempre el mismo valor, por ejemplo $OR_{5/4} = 4.340544$. Esta cantidad es una medida de la asociación entre ambas variables.

Podemos comparar el modelo con el modelo *nulo* que no contiene ninguna variable ($logit(p_i) = log(\frac{p_i}{1-p_i}) = \beta_0$). De esta forma comprobamos si la inclusión de la nueva variable ha proporcionado un incremento significativo en la explicación de la respuesta. Esto es el equivalente al análisis de la varianza en el modelo lineal, la distribución es una chi-cuadrado en lugar de ua F de Snedecor. El ANOVA compara el modelo ajustado con el modelo constante

```r
anova(gfit1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Voto
```

```
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     99     132.81
## Cifras  1    23.72        98     109.09 1.114e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La *Deviance* (diferencia con el modelo perfecto) del modelo nulo es 132.81 mientras que la del modelo ajustado es 109.09 (más pequeña ya que este estaría más cerca del modelo perfecto), con 99 y 98 grados de libertad respectivamente. La diferencia entre ambas es 23.72 (sigue una chi cuadrado con 1 grado de libertad) y es significativa (p=1.114e-06), en otras palabras, el modelo ajustado es significativamente mejor que el nulo (p=0.000001114).

Podríamos hacer la diferencia ajustando manualmente el odelo nulo y comparándolo con el modelo actal. El resultado es exactamnte el mismo.

```
gfit0=glm(Voto~1, data=Votos, family=binomial)
gfit0
```

```
##
## Call:  glm(formula = Voto ~ 1, family = binomial, data = Votos)
##
## Coefficients:
## (Intercept)
##      0.4895
##
## Degrees of Freedom: 99 Total (i.e. Null);  99 Residual
## Null Deviance:       132.8
## Residual Deviance: 132.8     AIC: 134.8
```

```
anova(gfit0, gfit1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Voto ~ 1
## Model 2: Voto ~ Cifras
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99     132.81
## 2        98     109.09  1    23.72 1.114e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La función base de *R* no calcula los equivalentes al R^2, podemos encontrarlos en el paquete *rms*

```
## Loading required package: rms
```

```
## Warning: package 'rms' was built under R version 3.5.2
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

## Error: package or namespace load failed for 'rms':
##  object 'plotp' not found whilst loading namespace 'rms'
```

Y ajustando de nuevo el modelo

```
#library(rms)
#gfit2=lrm(Voto~Cifras, data=Votos)
#gfit2
```

Ahora en el resumen aparece toda la información incluyendo el pseudo coeficiente de Nagelkerke, $R^2 = 0.287$.

Calculamos las predicciones de as probabilidades y clasificamos como voto positivo a aqellos cuya probabilidad sea mayor de 0.5 con lo que tenemos una predicción del voto. Cruzándolo con los votos reales, tenemos la que se denomina *tabla de confusión* que nos sirve para valorar en porcentaje de predicciones correctas.

```
pi=predict(gfit1, type="response")
pred=as.numeric(pi>0.5)
conf=table(Votos$Voto, pred)
conf
```

```
##     pred
##       0  1
##   No 18 20
##   Si  4 58
```

```
prop.table(conf, 1)
```

```
##     pred
##                0          1
##   No 0.47368421 0.52631579
##   Si 0.06451613 0.93548387
```

```
predtot=sum(diag(conf))/sum(conf)
predtot
```

```
## [1] 0.76
```

Predecimos correctamente el 76% de los votos, el 47.37% de los negativos y el 93.55% de los positivos, es decir el modelo predice mejor los votos positivos que los negativos.

Modelo con variables nominales

Como en el caso de los modelos lineales, podemos incluir variables cualitativas exactamente de la misma forma, es decir, mediante variables indicadoras.

```r
gfit2=glm(Voto~Partido, data=Votos, family=binomial)
summary(gfit2)
```

```
##
## Call:
## glm(formula = Voto ~ Partido, family = binomial, data = Votos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0494  -0.9778   0.5111   0.5111   1.3911
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.4895     0.2914  -1.680   0.0929 .
## PartidoRepublicano   2.4590     0.5242   4.691 2.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.86  on 98  degrees of freedom
## Residual deviance: 102.84  on 97  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 106.84
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(gfit2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Voto
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                       98     130.86
## Partido  1   28.021        97     102.84  1.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coefficients(gfit2)
```

```
##        (Intercept) PartidoRepublicano
##         -0.4895482          2.4589889
```

```r
exp(coefficients(gfit2)[2])
```

```
## PartidoRepublicano
##           11.69298
```

El modelo ajustado es ahora

$$p_i = P\left(Y = 1/X = x_i\right) = \frac{e^{-0.4895+2.4590 partido_i}}{1+e^{-0.4895+2.4590 vpartido_i}} = \frac{1}{1+e^{0.4895-2.4590 partido_i}}$$

Para un senador republicano $partido_i = 1$ la probabilidad estimada es $p_i = \frac{1}{1+e^{0.4895-2.4590}} = 0.8775574$. Para un senador democrata $p_i = \frac{1}{1+e^{0.4895}} = 0.3800114$. Es mucho más probable que un senador republicano vote positivo.

Observamos que el modelo copmpara el partido republicano con el demócrata en la variable ficticia. Usando el criterio anterior, el coeficiente correspondiente al partido (Republicano) es 2.459. Si tomamos $e^{\beta_1} = e^{2.459} = 11.69298$ podemos interpretarlo en términos del *odds ratio*, la ventaja de voto positivo frente a negativo es 11.69298 veces mayor en el partido republicano, loq ue quiere decir que el voto positivo es más elevado en este partido. Esto sería el equivalente al análisis de la varianza cuando la respuesta es binaria.

La tabla de confusión es:

```
pi=predict(gfit2, type="response")
pred=as.numeric(pi>0.5)
isna=is.na(Votos$Partido)
conf=table(Votos$Voto[-isna], pred)
conf
```

```
##     pred
##       0  1
##   No 24 14
##   Si 26 35
```

```
prop.table(conf, 1)
```

```
##     pred
##               0         1
##   No 0.6315789 0.3684211
##   Si 0.4262295 0.5737705
```

```
predtot=sum(diag(conf))/sum(conf)
predtot
```

```
## [1] 0.5959596
```

Ha mejorado la predicción del "No" pero ha empeorado la del "Si". La global es peor que la del número de cifras.

Varias variables

Podemos incluir varias variables en un único modelo para obtener el modelo de regresión logística múltiple

$$p_i = P(Y = 1/\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}$$

El modelo de regresión logística es un modelo lineal en escala *logit*.

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = M_i$$

Por ejemplo, podemos incluir las cifras y el partido ajustando el modelo correspondiente.

```
gfit3=glm(Voto~Cifras + Partido, data=Votos, family=binomial)
summary(gfit3)
```

```
##
## Call:
## glm(formula = Voto ~ Cifras + Partido, family = binomial, data = Votos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1971  -0.8493   0.4461   0.5496   3.0070
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.5100     2.0094  -2.244 0.024804 *
## Cifras               0.8990     0.4230   2.125 0.033574 *
## PartidoRepublicano   1.9146     0.5584   3.429 0.000606 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.861  on 98  degrees of freedom
## Residual deviance:  93.816  on 96  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 99.816
##
## Number of Fisher Scoring iterations: 5
```

```
anova(gfit3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Voto
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       98    130.861
## Cifras   1   24.004        97    106.858 9.616e-07 ***
## Partido  1   13.042        96     93.816 0.0003046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coefficients(gfit3)
```

```
##        (Intercept)             Cifras PartidoRepublicano
##         -4.5100423          0.8989789          1.9146340
```

```
exp(coefficients(gfit3)[2:3])
```

```
##             Cifras PartidoRepublicano
##           2.457093           6.784455
```

De nuevo parece que explicamos el comportamiento del voto en función del número de cifras y del partido, ambas variables tienen coeficientes significativamente distintos de cero. Ahora el número de cifras parece que ha perdido importancia al introducir también el partido. Hay que tener en cuenta que ambas variables (cifras y partido) están relacionadas (recordemos que reciben más dinero los republicanos) por lo que no es posible

separar los efectos de cada una de ellas.Si las ponemos en orden secuencial, es decir, ponemos primero las Cifras y luego el partido

El modelo ajustado es ahora

$$p_i == \frac{1}{1+e^{-4.510-0.899cifras_i-1.915partido_i}}$$

Y la descomposición dela *deviance*

```
anova(gfit3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Voto
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       98    130.861
## Cifras   1   24.004        97    106.858 9.616e-07 ***
## Partido  1   13.042        96     93.816 0.0003046 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

vemos que el partido añade una parte significativa a las cifras.

La interpretación de los parámetros es similar. La tabla de confusión es ahora

```
pi=predict(gfit3, type="response")
pred=as.numeric(pi>0.5)
isna=is.na(Votos$Partido)
conf=table(Votos$Voto[-isna], pred)
conf
```

```
##     pred
##       0  1
##   No 20 18
##   Si 16 45
```

```
prop.table(conf, 1)
```

```
##     pred
##               0         1
##   No 0.5263158 0.4736842
##   Si 0.2622951 0.7377049
```

```
predtot=sum(diag(conf))/sum(conf)
predtot
```

```
## [1] 0.6565657
```

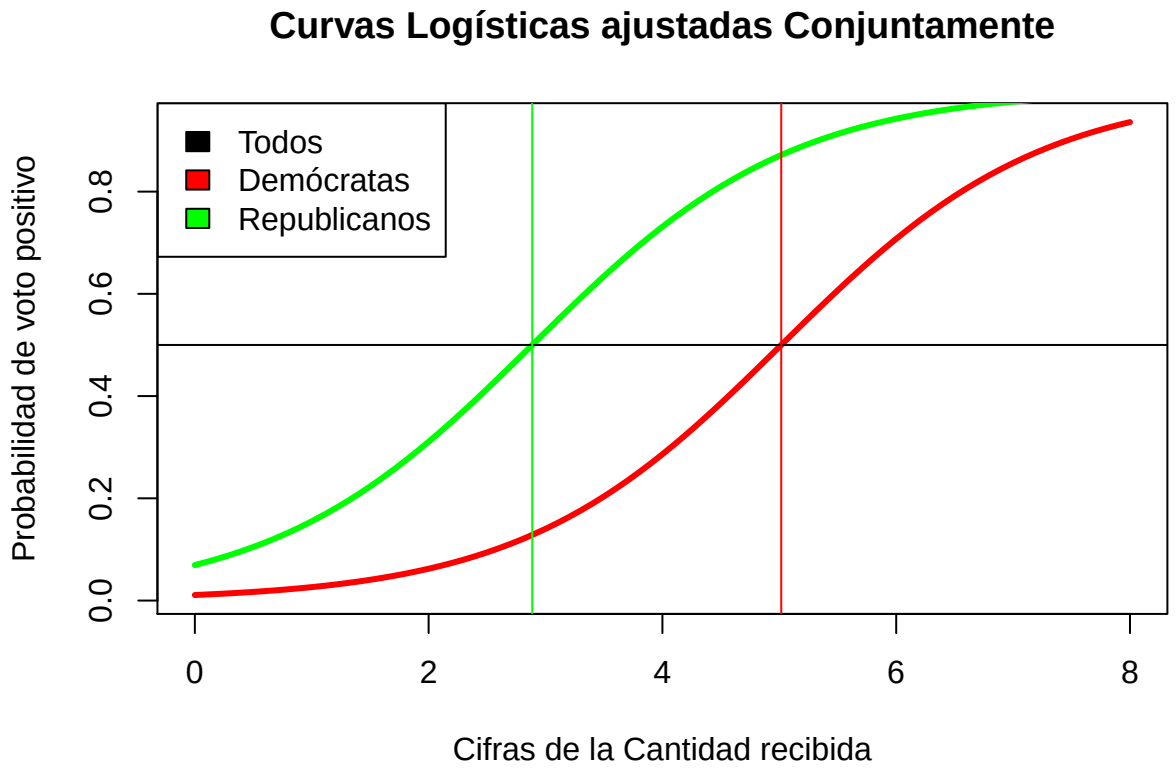
La predicción total es peor que la obtenida con solo las cifras.

Podemos dibujar las curvas

```
voto=as.numeric(Votos$Voto)-1
linDem= -4.510 + 0.899*x
```

```r
pvotoDem=exp(linDem)/(1+exp(linDem))
cutdem=4.510 / 0.899

linRep=-2.595 + 0.899*x
pvotoRep=exp(linRep)/(1+exp(linRep))
cutrep= 2.595 / 0.899
isna=is.na(Votos$Partido)
plot(x, pvotoDem, type="l", col="red", lwd=3, main = "Curvas Logísticas ajustadas Conjuntamente", xlab =
points(x, pvotoRep, type="l", col="green", lwd=3)
legend("topleft", legend=c("Todos", "Demócratas", "Republicanos"), fill=c("black", "red", "green"))
abline(h=0.5)
abline(v=cutdem, col="red")
abline(v=cutrep, col="green")
```



Observamos que las dos curvas tienen la misma pendiente, es decir, discriminan igual los votos positivos y los negativos. Podemos ajustar curvas diferentes para ambos grupos introduciendo un término de interacción.

Modelo con interacción

También podemos incluir la interacción entre las variables

```r
gfit4=glm(Voto~Cifras + Partido + Cifras:Partido, data=Votos, family=binomial)
summary(gfit4)
```

```
##
## Call:
## glm(formula = Voto ~ Cifras + Partido + Cifras:Partido, family = binomial,
##     data = Votos)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.5355  -0.8873   0.3133   0.5826   2.5466
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -3.2027     1.6453  -1.947   0.0516 .
## Cifras                     0.6184     0.3504   1.765   0.0776 .
## PartidoRepublicano        -8.0460     6.4621  -1.245   0.2131
## Cifras:PartidoRepublicano  2.0193     1.3058   1.546   0.1220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.861  on 98  degrees of freedom
## Residual deviance:  90.688  on 95  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 98.688
##
## Number of Fisher Scoring iterations: 5
```

```
anova(gfit4, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Voto
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            98    130.861
## Cifras          1  24.0036        97    106.858 9.616e-07 ***
## Partido         1  13.0417        96     93.816 0.0003046 ***
## Cifras:Partido  1   3.1284        95     90.688 0.0769406 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coefficients(gfit4)
```

```
##               (Intercept)                    Cifras
##                -3.2027442                 0.6184139
##        PartidoRepublicano Cifras:PartidoRepublicano
##                -8.0459939                 2.0193364
```

```
exp(coefficients(gfit4)[2:4])
```

```
##                    Cifras        PartidoRepublicano
##              1.8559819783              0.0003203828
## Cifras:PartidoRepublicano
##              7.5333238416
```

El modelo ajustado es

Vemos que ahora las variables han dejado de ser significativas. Al añadirlas de forma secuencial, la interacción

no ñade nada a lo que aportan las otras dos. Podemos hacerlo tambien con

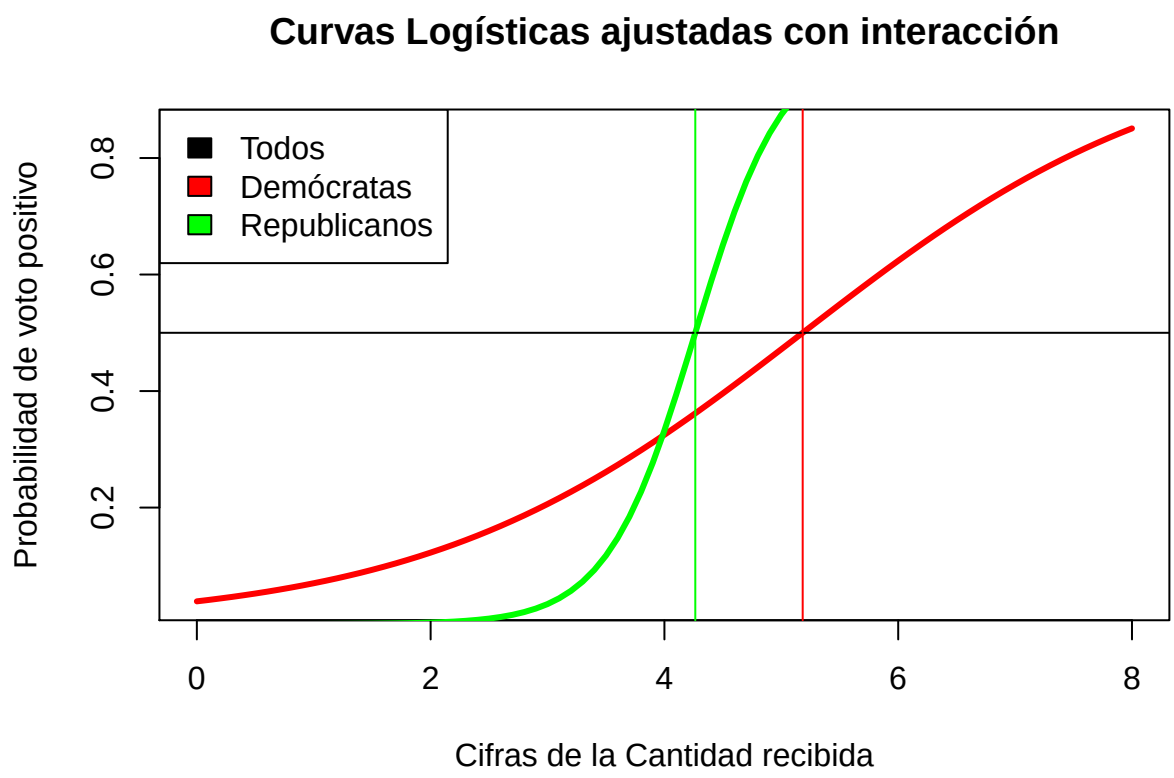
```
anova(gfit4, gfit3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Voto ~ Cifras + Partido + Cifras:Partido
## Model 2: Voto ~ Cifras + Partido
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        95     90.688
## 2        96     93.816 -1  -3.1284  0.07694 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linDem= -3.203 + 0.618*x
pvotoDem=exp(linDem)/(1+exp(linDem))
cutdem=3.203 / 0.618

linRep=-11.249 + 2.638*x
pvotoRep=exp(linRep)/(1+exp(linRep))
cutrep= 11.249 / 2.638
plot(x, pvotoDem, type="l", col="red", lwd=3, main = "Curvas Logísticas ajustadas con interacción", xlab
points(x, pvotoRep, type="l", col="green", lwd=3)

legend("topleft", legend=c("Todos", "Demócratas", "Republicanos"), fill=c("black", "red", "green"))
abline(h=0.5)
abline(v=cutdem, col="red")
abline(v=cutrep, col="green")
```



La tabla de confusión es:

```r
pi=predict(gfit4, type="response")
pred=as.numeric(pi>0.5)
isna=is.na(Votos$Partido)
conf=table(Votos$Voto[-isna], pred)
conf
```

```
##     pred
##       0  1
##   No 21 17
##   Si 19 42
```

```r
prop.table(conf, 1)
```

```
##     pred
##               0         1
##   No 0.5526316 0.4473684
##   Si 0.3114754 0.6885246
```

```r
predtot=sum(diag(conf))/sum(conf)
predtot
```

```
## [1] 0.6363636
```

Por lo que parece que el modelo a elegir es el tercero.