Tarea3

Jorge III Altamirano Astorga

July 16, 2017

Tarea 3

Jorge III Altamirano Astorga

Parte aplicada

Para esta parte pueden usar la base de datos diamonds que sugirieron, aunque hay puntos adicionales si usan alguna base original interesante.

Cargar la base que se encuentra en el paquete ggplot2. Los comandos que pueden usar para cargar la base diamonds a su ambiente de trabajo en R son:

```
library(ggplot2)
data(diamonds)
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat       cut color clarity depth table price     x     y     z
##   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
## 2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
## 3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
## 4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
## 5  0.31      Good     J     SI2  63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
```

Posteriormente deben hacer una regresión lineal. Su objetivo es explicar la variable price usando las demás variables. Noten que algunas variables no son numéricas, por lo que no pueden incluirse en un análisis crudo de regresión lineal. Para este proyecto NO es necesario saber transformar las variables no numéricas para poder usarlas en la regresión; hacerlo es optativo,

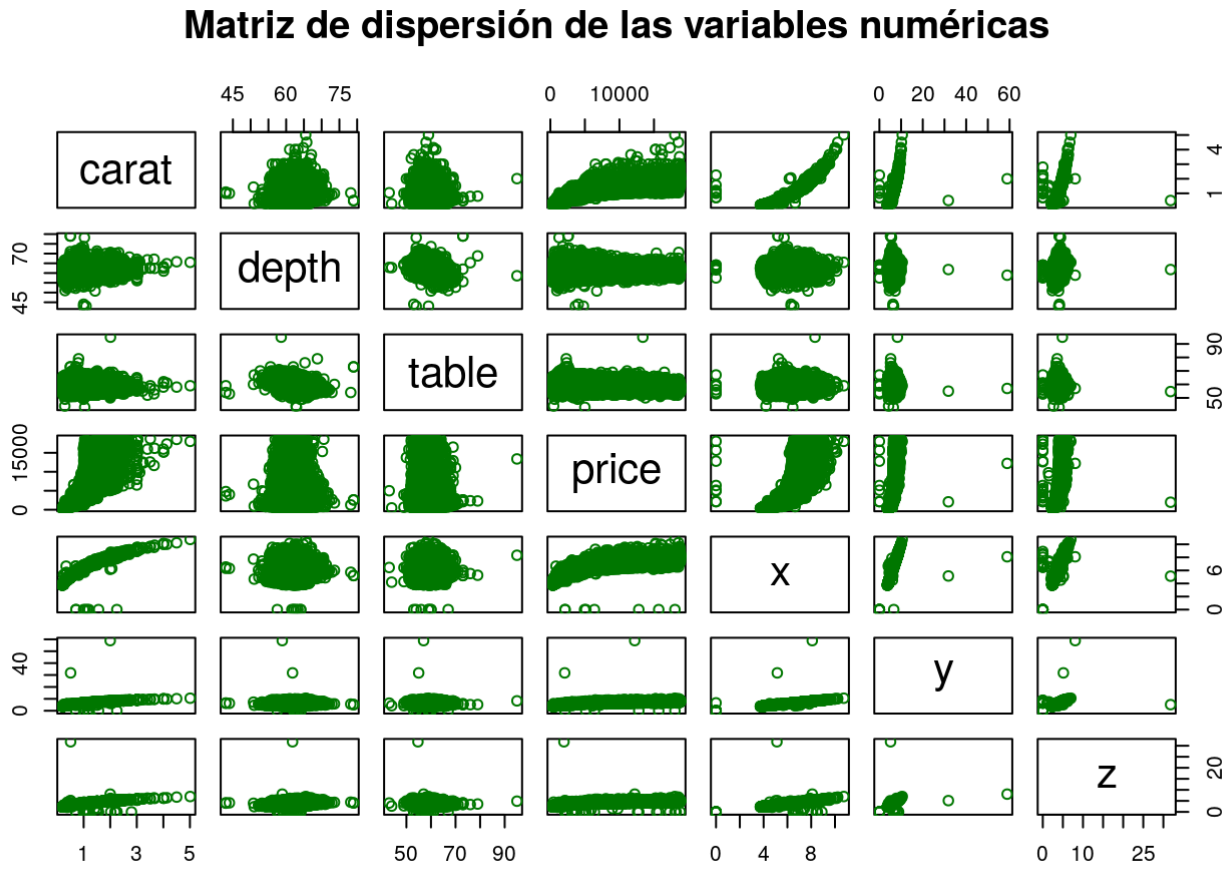
de hecho, las paqueterías lo hacen por ustedes pero deben ser cuidadosos. Pueden usar la función lm de R para su análisis de regresión.

Creando el dataframe…

```
diamonds = diamonds[ ,c(1,5,6,7,8,9,10)]
diamonds0 = scale(diamonds)
diamonds0 <- as.data.frame(diamonds)
```

Matriz de dispersión de variables numéricas

```
pairs(diamonds0, col="#007700", main="Matriz de dispersión de las variables numéricas")
```



Aquí se puede discernir que hay una relación directa entre quilates (*carat*) y el precio, aunque la varianza aumenta conforme aumenta los carats.

Así mismo se observa menor relación entre precio y dimensiones (*x*, *y* y *z*)

```
model1 = lm(price ~ carat + depth + table + x + y + z, data = diamonds0)
```

¿Qué tan bueno fue el ajuste?

Se muestra un ajuste correcto en nuestro modelo si tomamos en cuenta las 6 variables, pues aparecen con una media menor de residuales (apegadas 0). Por lo que la línea roja debe ir lo más cercana a cero.

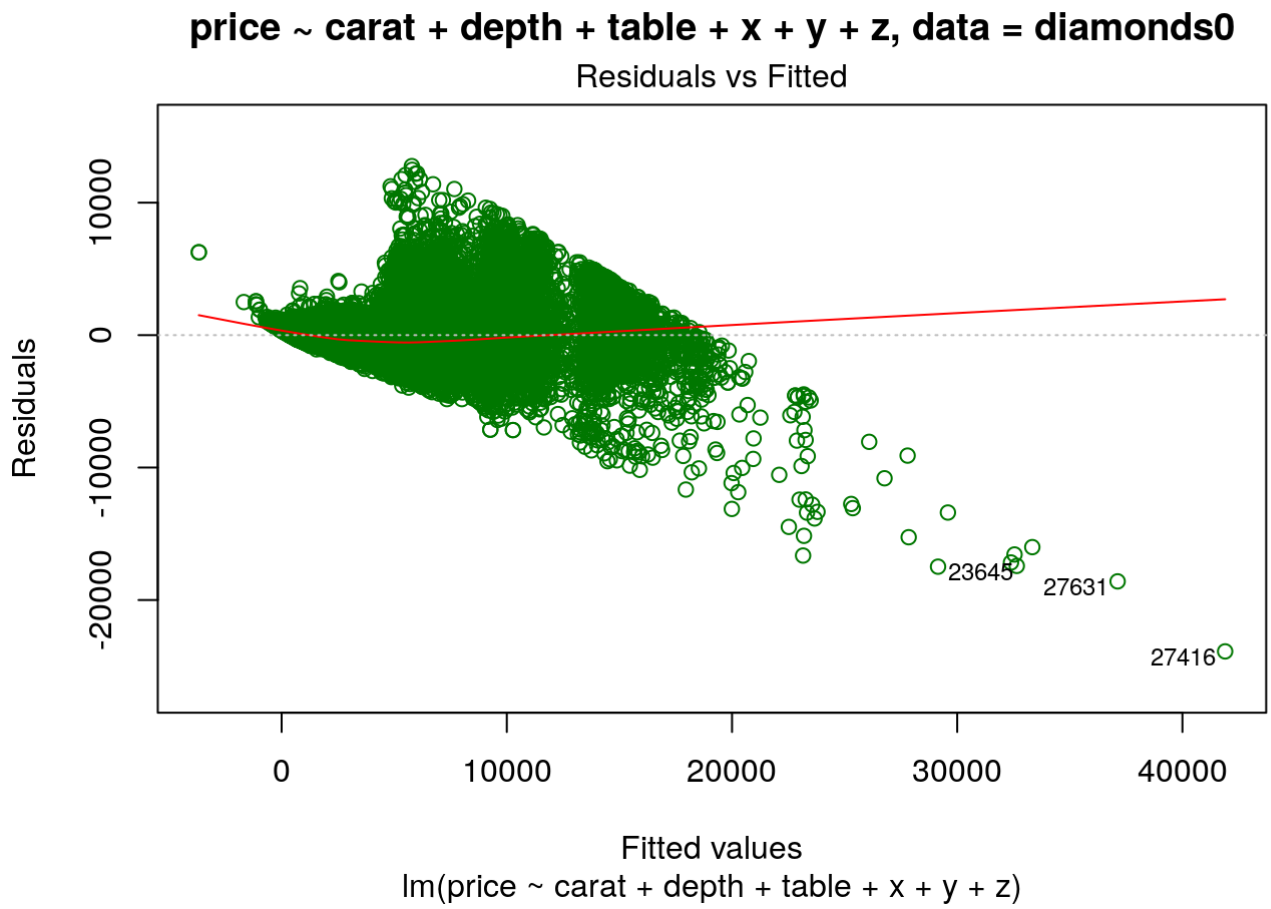
```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ carat + depth + table + x + y + z, data = diamonds0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23878.2   -615.0    -50.7    347.9  12759.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20849.316    447.562  46.584  < 2e-16 ***
## carat       10686.309     63.201 169.085  < 2e-16 ***
## depth        -203.154      5.504 -36.910  < 2e-16 ***
## table        -102.446      3.084 -33.216  < 2e-16 ***
## x           -1315.668     43.070 -30.547  < 2e-16 ***
## y              66.322     25.523   2.599  0.00937 **
## z              41.628     44.305   0.940  0.34744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 53933 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 5.486e+04 on 6 and 53933 DF,  p-value: < 2.2e-16
```

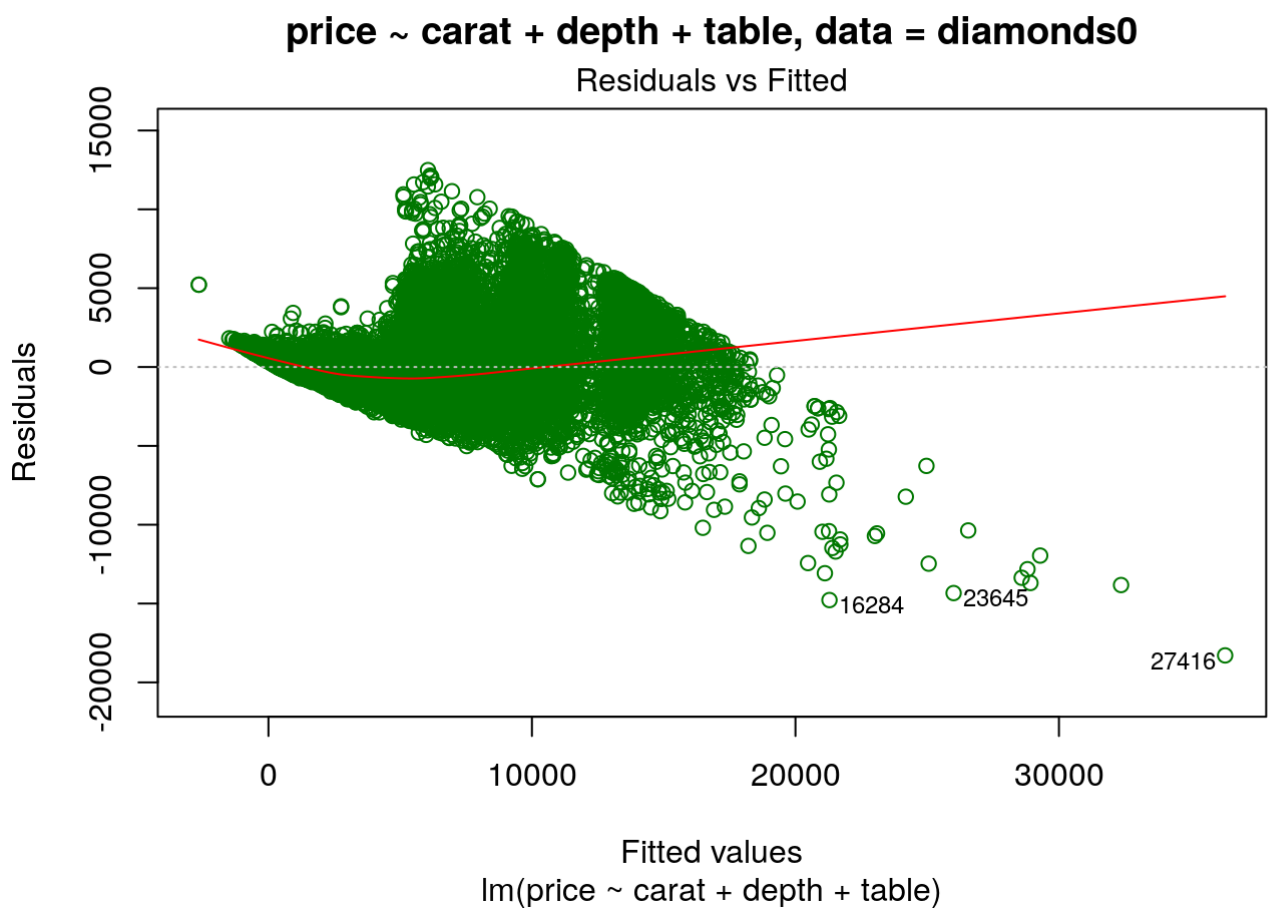
Así mismo podemos observar que el que la *R cuadrada* muestra una predicción del 85.92% del precio basado en las 6 variables del modelo.

¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cúal fue el valor de $ ^2 $ que ajustó su modelo y que relación tienen con la calidad del ajuste?

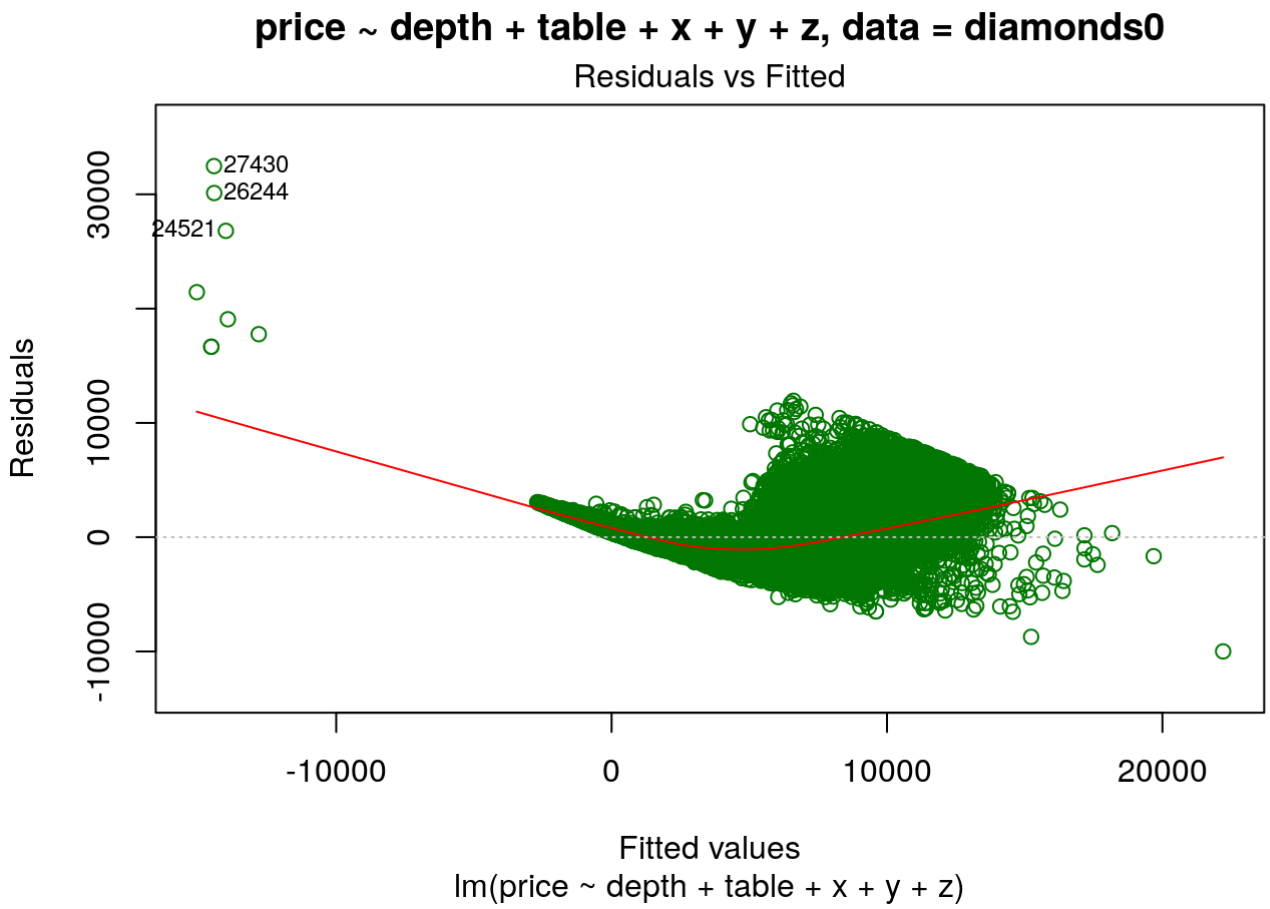
```
plot(model1,main="price ~ carat + depth + table + x + y + z, data =
diamonds0",which=1,col="#007700")
```




```
model2 = lm(price ~ carat + depth + table, data = diamonds0)
plot(model2,main="price ~ carat + depth + table, data = diamonds0",which=1,col="#00770
0")
```




```
model3 = lm(price ~ depth + table + x + y + z, data = diamonds0)
plot(model3,main="price ~ depth + table + x + y + z, data = diamonds0",which=1,col="#00
7700")
```



Dadas las gráficas anteriores que muestran cómo con nuestro modelo basado en las variables numéricas siguientes: 1. Carat

2. Depth
3. Table

Dimensiones:

4. *x*
5. *y*
6. *z*

Es el más aproximado pues al sacar las variables de dimensiones (*x*, *y* y *z*) en el modelo 2 y la variable quilate (*carat*) se muestran las líneas rojas de la media de los residuales apegada a cero. Ahora mostrando dichas variables con el sumario y sacando los R cuadradas de modelos observamos:

```
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ carat + depth + table, data = diamonds0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18288.0   -785.9    -33.2    527.2  12486.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13003.441    390.918   33.26   <2e-16 ***
## carat        7858.771     14.151  555.36   <2e-16 ***
## depth        -151.236      4.820  -31.38   <2e-16 ***
## table        -104.473      3.141  -33.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1526 on 53936 degrees of freedom
## Multiple R-squared:  0.8537, Adjusted R-squared:  0.8537
## F-statistic: 1.049e+05 on 3 and 53936 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ depth + table + x + y + z, data = diamonds0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9994  -1256   -197    945  32470
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8770.608    509.448 -17.216  < 2e-16 ***
## depth         -10.501      6.661  -1.576   0.1149
## table         -84.855      3.813 -22.255  < 2e-16 ***
## x            2918.492     43.346  67.330  < 2e-16 ***
## y             205.086     31.555   6.499 8.13e-11 ***
## z              91.814     54.802   1.675   0.0939 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1852 on 53934 degrees of freedom
## Multiple R-squared:  0.7846, Adjusted R-squared:  0.7846
## F-statistic: 3.929e+04 on 5 and 53934 DF,  p-value: < 2.2e-16
```

Tenemos los siguientes valores: - Modelo 2: 85.37% - Modelo 3: 78.46% *este es el modelo menos aproximado, pues sin los quilates como variable disminuye significativamente el valor de R cuadrada como era esperado*

¿Cual es el angulo entre Y y Ŷ estimada? Hint: usen la R^2 cuadrada y el arcocoseno?

```
acos(sqrt(0.8592))*180/pi
```

```
## [1] 22.03873
```

22.03 Grados

- Definan una funcion que calcule la logverosimilitud de unos parámetros β y σ^2.

```
library(ggplot2)
diamonds_data = data(diamonds)
diamonds_short <-  diamonds[ ,c(1,5,6,7,8,9,10)]
diamonds_x <- diamonds[ ,c(5,6,7,8,9,10)]
diamonds_m <- data.matrix(diamonds_x)
n <- length(diamonds_x )
sigma_sq <- 0.8563
mod <- lm(formula = diamonds$price ~ diamonds$carat + diamonds$x + diamonds$y + diamond
s$z + diamonds$depth)
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat       cut color clarity depth table price     x     y     z
##   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
## 2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
## 3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
## 4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
## 5  0.31      Good     J     SI2  63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
```

```
summary(mod)$coefficients[,1]
```

```
##    (Intercept) diamonds$carat     diamonds$x     diamonds$y     diamonds$z
##    12196.68697    10615.49551    -1369.67016       97.59636       64.19545
## diamonds$depth
##     -156.62430
```

```
beta1 <- c(12196.7, 10615.5, -1369.7, 97.6, 64.2, -156.6)

funcLikelyhood <- function(bet, sig){
  -(n/2)*(log(2*pi))-((n/2)*log(sig))-((1/(2*sig))*((diamonds$price-(diamonds_m*bet))*
(diamonds$price-(diamonds_m*bet))))
}
```

```
head(funcLikelyhood(beta1,sigma_sq))
```

```
##              depth         table         price             x             y
## [1,] -328247476017 -262501572209 -9.229809e+12 -1.336982e+09 -1.357507e+09
## [2,] -235061317572 -244594815596 -6.991628e+12 -9.800311e+08 -9.548015e+08
## [3,]   -3576483838   -4662362966 -1.173071e+11 -2.014903e+07 -2.033740e+07
## [4,]     -19347371     -16568263 -6.078428e+08 -3.370601e+03 -3.635205e+03
## [5,]      -8118887      -6704787 -2.617386e+08 -1.860591e+03 -1.818567e+03
## [6,]     -60398617     -50092466 -1.637326e+09 -5.303196e+05 -5.338110e+05
##                   z
## [1,] -5.016888e+08
## [2,] -3.418411e+08
## [3,] -7.116162e+06
## [4,] -3.495148e+03
## [5,] -1.466487e+04
## [6,] -3.063866e+05
```

- Utilicen la función optim de R para numericamente el máximo de la función de verosimilitud. Si lo hacen correctamente, su solución debe coincidir con la del método lm.

Parte teórica

Basado en el desarrollo matemático de Alejandro Estrada y contribuciones de trabajo en equipo de estudio del domino 16 de Julio

Esta parte del proyecto será sobre regresión lineal. Supongamos que quieren explicar una variable estadistica Y (por ejemplo altura) utiizando la información de p variables X1, …, Xp (peso, ancho de huesos, etc). Si se toma una muestra de N individuos, cada variable está representada por un vector de tamaño N. La información de las variables explicativas se pueden juntar en una matriz

$$X = [X^1 \mid \ldots \mid X^p]$$

de tamaño n x p donde cada columna es una variable y cada fila uno de los individuos de la muestra. Tienen que contestar lo siguiente:

- Plantear el problema de regresión lineal como un problema de mínimos cuadrados, encontrar el vector beta que resuelva

$$\hat{\beta} = argmin_{\beta \in R^p} \mid\mid Y - X\beta \mid\mid^2$$

y encontrar la solución teórica. ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos?

La solución del problema de mínimos cuadrados nos da una aproximación lineal a nuestros datos porque los parámetros $$ son lineales. Esto es, los parámetros capturan sólo la parte lineal que nos ayuda a explicar la variable 'Y' a partir de nuestros datos 'X'. Por ejemplo, en el caso de dos dimensiones donde estimamos la variable Y a partir de la ecuación Y_hat = B_0 + B_1*X, el modelo se asemeja a una recta donde el parámetro B_0 representa la ordenada al origen de dicha recta y B_1 captura la pendiente o relación lineal entre las variables X y Y

¿Podríamos usarlo para ajustar polinomios (ej $y = x^2$)?

Podemos ajustar polinomios del modo $y = x^2$ y mayores sin ningún problema utilizando el resultado de regresión lineal para las β's. Lo que es importante notar es que aunque la regresión de polinomio ajusta un modelo no-lineal sobre los datos, el problema de estimación estadística continúa siendo lineal (es lineal en las β's aunque no en las variables x) lo que es consecuencia directa de que la función E(Y|X) es lineal en los parámetros beta estimados.

$$\bigtriangledown \mid\mid Y - X\beta \mid\mid^2 = \bigtriangledown(Y^TY + \beta^TX^TX\beta - 2\beta^TX^TY)$$

$$0 = 2X^TX\beta - 2X^TY$$

Dividiendo ambos lados entre dos y resolviendo para Beta tenemos:

$$\beta_{OLS} = (X^TX)^{-1}X^TY$$

- Argumentar la relación entre la solución encontrada y un problema de proyección en subespacios vectoriales de álgebra lineal ¿Cuál es la relación particular con el teorema de Pitágoras?

Cuando hay una relación lineal entre dos variables, la varianza de la variable dependiente se puede descomponer en dos varianzas: la de los pronósticos, debido a la relación que la variable dependiente guarda con la variable independiente, y la de los errores o residuos. Esta relación se cumple tanto para la Regresión Lineal Simple como para la Múltiple. Esta descomposición de la varianza de la variable dependiente en dos varianzas es el "Teorema de Pitágoras" del Análisis de Regresión Lineal que, para efectos del modelo anterior, la varianza de las puntuaciones observadas es igual a la varianza de las puntuaciones estimadas más la varianza de los residuos.

- ¿Que logramos al agregar una columna de unos en la matriz X? es decir, definir mejor

$$X = [1_n \mid X^1 \mid \ldots \mid X^p]$$

con $1_n = [1, 1, \ldots, 1]^T$

El parámetro β_0 que se captura con la columna de unos dentro de la matriz X nos ayuda a incorporar la información no contenida en las variables de nuestro modelo. De esta manera, el estimador $\hat{Y}$ no necesariamente inicia desde el origen, lo que ayuda a reducir los errores en nuestra estimación.

- Plantear el problema de regresión ahora como un problema de estadística

donde los errores son no correlacionados con distribución

- ¿Cual es la función de verosimilitud del problema anteriror? Hint: empiecen por excribir el problema como

Sea

$$Y = X\beta + \epsilon$$

con

$$\epsilon \sim N(0, \sigma^2 I_n)$$

con I_n la matriz identidad. Y concluyan entonces que

$$Y \sim N(X\beta, \sigma^2 I_n)$$

Escriban entonces la verosimilitud como

$$L(\beta, \sigma^2; Y, X) = \prod_{i=1}^{p} f_y(y_i|X; \beta, \sigma^2)$$

$$= \prod_{i=1}^{p} (2\pi\sigma^2)^{-1/2} exp(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2})$$

$$= (2\pi\sigma^2)^{-N/2} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{p}(y_i - x_i\beta)^2)$$

- Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados. La función log de máxima verosimilitud es:

$$l(\beta, \sigma^2; Y, X) = -\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2$$

El siguiente paso es derivar respecto a cada una de las β:

$$\bigtriangledown_\beta l(\beta, \sigma^2; Y, X)$$

$$\bigtriangledown_\beta(-\frac{N}{2}ln(2\pi) - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^T(y_i - x_i\beta)$$

$$= \frac{1}{\sigma^2}(\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta)$$

Que es igual a cero solo si

$$\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i \beta = 0$$

Esto se satisface si:

$$\beta = (\sum_{i=1}^{N} x_i^T x_i)^{-1} \sum_{i=1}^{N} x_i^T y_i = (X^T X)^{-1} X^T y$$

- Investiga el contenido del Teorema de Gauss-Markov sobre mínimos cuadrados.

El Teorema de Gauss-Márkov establece que en un modelo lineal general (MLG) en el que se cumplan los siguientes supuestos: - Correcta especificación: el MLG ha de ser una combinación lineal de los parámetrosβ y no necesariamente de las variables: $Y = X\beta + u$ - Muestreo aleatorio simple: la muestra de observaciones del vector $(y_i,\ x_{2i},\ x_{3i},\ \ldots,\ x_{ki})$ es una muestra aleatoria simple y, por lo tanto, el vector (y_i, X_i') es independiente del vector (y_i, X_j')
- Esperanza condicionada de los errores nula: $E(u_i|X_i') = 0$ - Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N - Homocedasticidad: la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones:$Var(U|X) = \alpha^2 I$

El estimador mínimo cuadrático ordinario (MCO) de β es el estimador lineal e insesgado óptimo, es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.