Tarea 3 - Juan Pablo de Botton Falcon

Parte teórica

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

$$X = [X^1 \,|\ldots|\, X^p]$$

de tama?o 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 regresi?n lineal como un problema de m?nimos cuadrados, encontrar el vector beta que resuelva

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

y encontrar la soluci?n te?rica. ?Por qu? este planteamiento nos da un ajuste lineal a nuestros datos?

La solucion del problema de minimos cuadrados nos da una aproximaci?n lineal a nuestros datos porque los par?metros β son lineales. Esto es, los par?metros capturan s?lo 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 ecuaci?n Y_hat = B_0 + B_1*X, el modelo se asemeja a una recta donde el parametro B_0 representa la ordenada al origen de dicha recta y B_1 captura la pendiente o relaci?n lineal entre las variables X y Y

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

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

$$\square || Y - X\beta ||^2 = \square \, (Y^T Y + \beta^T X^T X \beta - 2\beta^T X^T Y)$$

$$0 = 2X^T X\beta - 2X^T Y$$

Dividiendo ambos lados entre dos y resolviendo para Beta tenemos:

$$\beta_{OLS} = (X^T X)^{-1} X^T Y$$

- Argumentar la relaci?n entre la soluci?n encontrada y un problema de proyecci?n en subespacios vectoriales de ?lgebra lineal ?Cu?l es la relaci?n particular con el teorema de Pit?goras?

Cuando hay una relaci?n lineal entre dos variables, la varianza de la variable dependiente se puede descomponer en dos varianzas: la de los pron?sticos, debido a la relaci?n que la variable dependiente guarda con la variable independiente, y la de los errores o residuos. Esta relaci?n se cumple tanto para la Regresi?n Lineal Simple como para la M?ltiple. Esta descomposici?n de la varianza de la variable dependiente en dos varianzas es el "Teorema de Pit?goras" del An?lisis de Regresi?n Lineal que, para efectos del modelo anterior, la varianza de las puntuaciones observadas es igual a la varianza de las puntuaciones estimadas m?s 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 par?metro β_0 que se captura con la columna de unos dentro de la matriz X nos ayuda a incorporar la informaci?n 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 estimaci?n.

- Plantear el problema de regresi?n ahora como un problema de estad?stica

donde los errores son no correlacionados con distribuci?n

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 soluci?n de m?xima verosimilitud es la misma que la del problema de m?nimos cuadrados. La funci?n log de m?xima 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:

$$\frac{?}{\beta}\, l(\beta, \sigma^2; Y, X)$$

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

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

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

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 m?nimos cuadrados.

El Teorema de Gauss-M?rkov establece que en un modelo lineal general (MLG) en el que se cumplan los siguientes supuestos: - Correcta especificaci?n: el MLG ha de ser una combinaci?n lineal de los par?metrosβ y no necesariamente de las variables: $Y = X\beta + u$ - Muestreo aleatorio simple: la muestra de observaciones del vector $(y_i,\ x_{2i},\ x_{3i},\ \dots,\ x_{ki})$ 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 identificaci?n: 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 m?nimo cuadr?tico ordinario (MCO) de β es el estimador lineal e insesgado ?ptimo, es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.

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:

```
#install.packages("ggplot2")
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.2

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

?diamonds
```

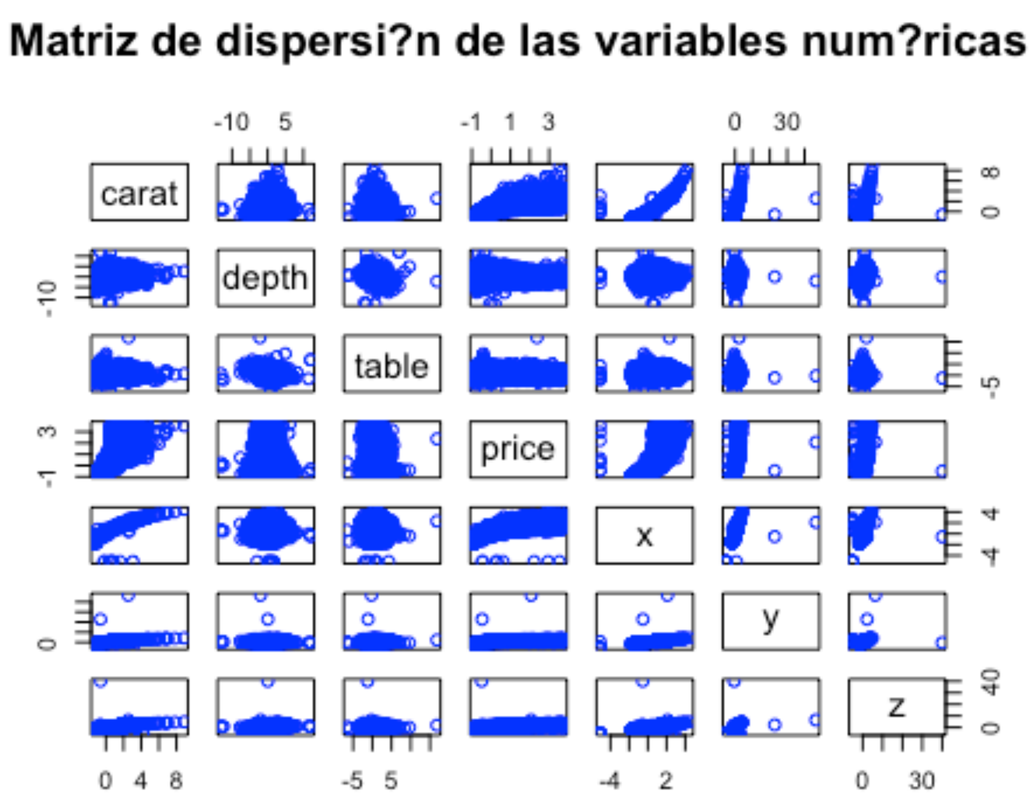
Posteriormente deben hacer una regresi?n lineal. Su objetivo es explicar la variable price usando las dem?s variables. Noten que algunas variables no son num?ricas, por lo que no pueden incluirse en un an?lisis crudo de regresi?n lineal. Para este proyecto NO es necesario saber transformar las variables no num?ricas para poder usarlas en la regresi?n; hacerlo es optativo, de hecho, las paqueter?as lo hacen por ustedes pero deben ser cuidadosos. Pueden usar la funci?n lm de R para su an?lisis de regresi?n.

Creamos un nuevo data_frame que solo contiene los datos que vamos a utilizar, es decir, las variables num?ricas centralizadas

```
diamantes = diamonds[ ,c(1,5,6,7,8,9,10)]
diamantes0 = scale(diamantes)
diamantes0 <- as.data.frame(diamantes0)
```

Obtenemos una matriz de dispersi?n para formarnos una apreciaci?n inicial de las relaciones entre las variables.

```
pairs(diamantes0, col= "blue", main="Matriz de dispersi?n de las variables nu
m?ricas")
```

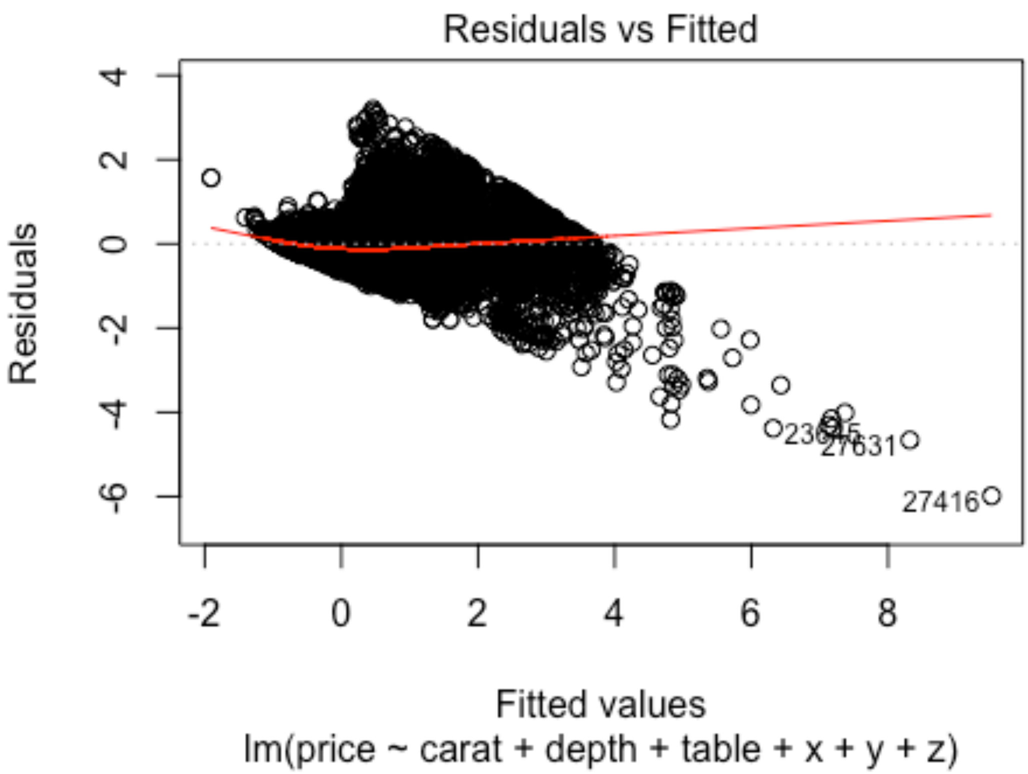


Gracias a la matriz de las variables en pares podemos discernir de forma r?pida una relaci?n positiva entre el precio y la variable "carat", aunque la varianza aumenta conforme aumenta esta variable. Tambi?n podemos observar que no hay mucha relaci?n entre el precio y las variables "y"(ancho en mm) y "z"(profundidad en mm).

```r
#Realizamos la regresi?n lineal usando todas las variables num?ricas
modelo1 = lm(price ~ carat + depth + table + x + y + z, data = diamantes0)
summary(modelo1)
```

```
##
## Call:
## lm(formula = price ~ carat + depth + table + x + y + z, data = diamantes0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9853 -0.1542 -0.0127  0.0872  3.1982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.359e-14  1.616e-03   0.000  1.00000
## carat        1.270e+00  7.509e-03 169.085  < 2e-16 ***
## depth       -7.295e-02  1.976e-03 -36.910  < 2e-16 ***
## table       -5.738e-02  1.727e-03 -33.216  < 2e-16 ***
## x           -3.699e-01  1.211e-02 -30.547  < 2e-16 ***
## y            1.899e-02  7.307e-03   2.599  0.00937 **
## z            7.364e-03  7.837e-03   0.940  0.34744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3752 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
```

```r
plot(modelo1, which =1)
```



- ?Qu? tan bueno fue el ajuste?

Del "summary" de la regresi?n podemos observar inicialmente la informaci?n de los residuales, que son todas las diferencias entre el precio estimado y el precio real; una media lo m?s cercana a cero es lo que querr?amos ver aqu?, pues indicar?a que el modelo es muy certero. En la gr?fica vemos los valores predichos contra los residuales, en un modelo perfecto, la media de los residuales ser?a 0 por lo que la l?nea roja ir?a perfectamente a lo largo de la l?nea que representa el 0. En seguida podemos apreciar los coeficientes de la intercepci?n y de cada una de las variables, usados para calcular las predicciones.As? tambi?n confirmamos que las variables "z" y "y" son las menos importantes para el modelo, esto lo sabemos por el valor p en la columna "Pr(>|t|)", que se resta a 1 para conocer la probabilidad de que NO sean relevantes al modelo, es decir, para todos las variables queremos que este sea lo m?s peque?o posible.

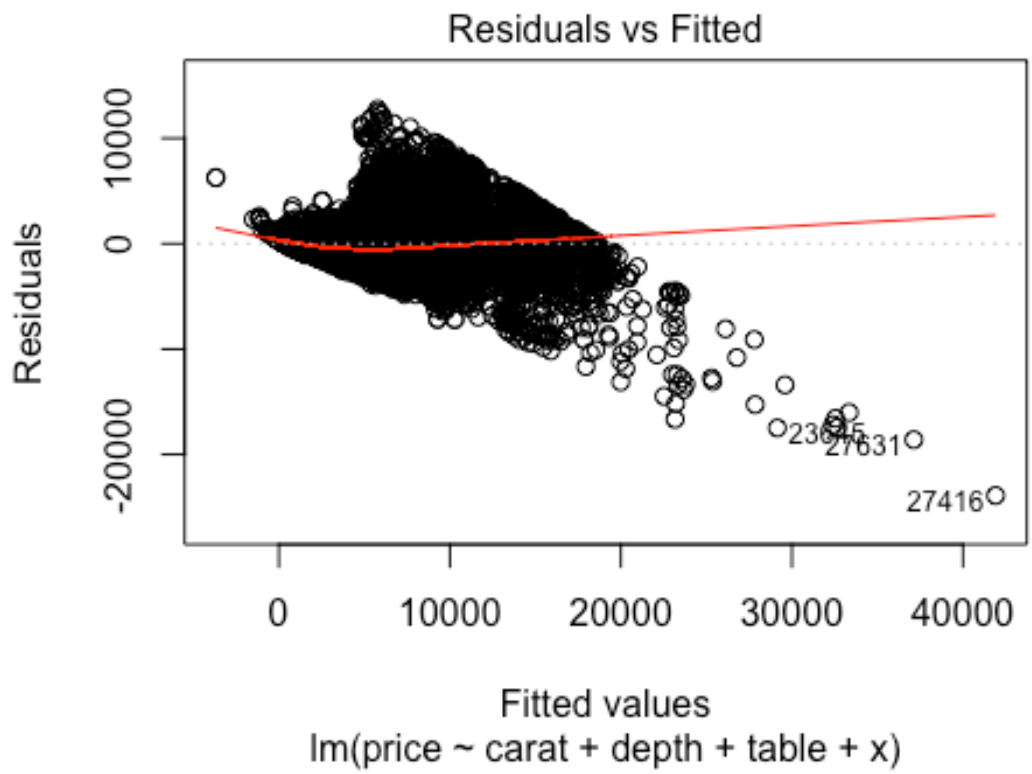
- ?Qu? medida puede ayudarnos a saber la calidad del ajuste? ?C?al fue el valor de $ ^2 $ que ajust? su modelo y que relaci?n tienen con la calidad del ajuste?

Al final del summary podemos apreciar la R cuadrada, que es la cantidad de variabilidad en lo que estas prediciendo que es explicado por el modelo, en este caso podr?amos decir que un 85.92 % del precio de los diamantes es explicado por las variables en nuestro modelo. Es una buena medida de la calidad del ajuste.

```r
# Por curiosidad, realizamos otra regresi?n lineal, dejando fuera a "y" y "z"
, que no aportaban mucho en el modelo anterior
modelo2 = lm(price ~ carat + depth + table + x, data = diamantes)
summary(modelo2)
```

```
##
## Call:
## lm(formula = price ~ carat + depth + table + x, data = diamantes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23894.2   -615.0    -50.6    346.9  12760.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20765.521    418.984   49.56   <2e-16 ***
## carat       10692.510     63.168  169.27   <2e-16 ***
## depth        -201.231      4.852  -41.48   <2e-16 ***
## table        -102.824      3.082  -33.37   <2e-16 ***
## x           -1226.773     26.678  -45.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 53935 degrees of freedom
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8592
## F-statistic: 8.228e+04 on 4 and 53935 DF,  p-value: < 2.2e-16
```

```r
plot(modelo2, which=1)
```



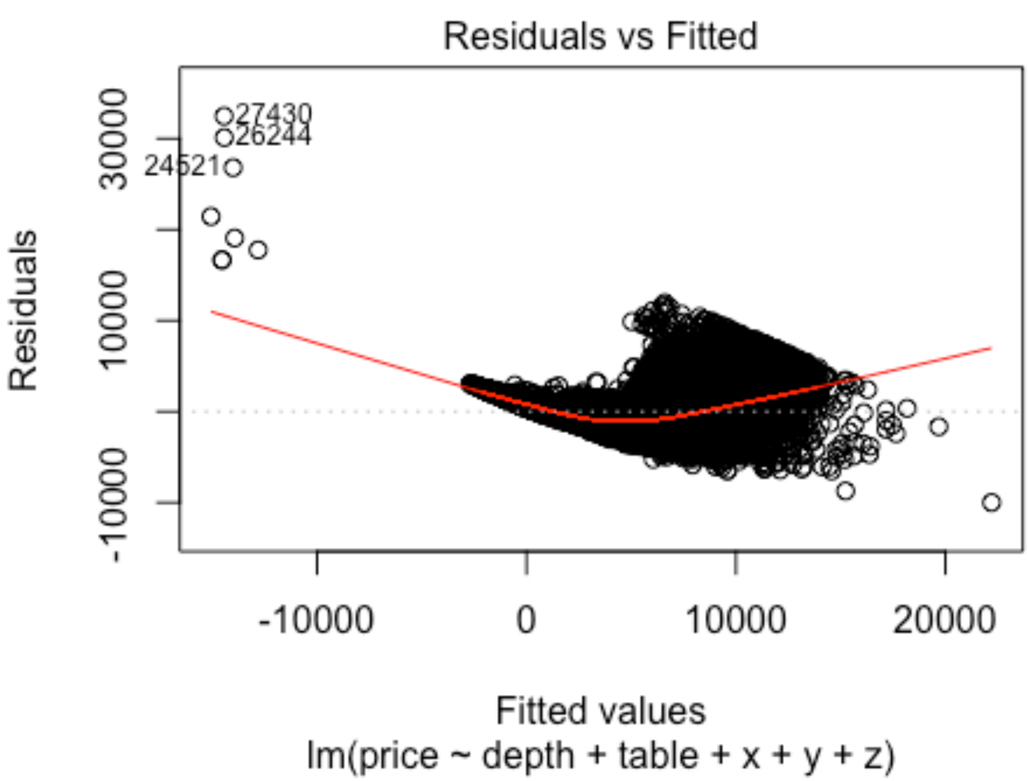
Aqui vemos, como era esperado, que deshacernos de las dos variables que no eran relevantes pr?cticamente no modifica valores como la media de los residuales o la R cuadrada.

```
# Por curiosidad, ahora realizamos otra regresi?n lineal, pero ahora dejamos
fuera la variable "carat", que es una variable relevante para el modelo
modelo3 = lm(price ~ depth + table + x + y + z, data = diamantes)
summary(modelo3)

##
## Call:
## lm(formula = price ~ depth + table + x + y + z, data = diamantes)
##
## 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
```

```r
plot(modelo3, which=1)
```



Como era esperado, remover una variable importante, disminuyo el valor de R cuadrada y aumento la media de los residuales (cuyo efecto puede verse en la recta de la grafica).

- Cual es el angulo entre Y y $\hat{Y}$ estimada? Hint: usen la R^2 cuadrada y el arcocoseno?

```r
angulo <- acos(sqrt(.8592))
angulo * 180/pi

## [1] 22.03873
```

- Definan una funcion que calcule la logverosimilitud de unos par?metros 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)]
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
```

```
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
+ diamonds$z + diamonds$depth)
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
```

```
Beta_1 <- c(12196.7, 10615.5, -1369.7, 97.6, 64.2, -156.6)

funcionVerosimilitud <- 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))))
}

#funcionVerosimilitud(Beta_1,sigma_sq)
```

- Utilicen la funci?n optim de R para numericamente el m?ximo de la funci?n de verosimilitud. Si lo hacen correctamente, su solucion debe coincidir con la del metodo lm.

Intentamos este ej pero no nos salió la máxima verosimilitud con Optim

Disclaimer: Esta tarea la hicimos en equipo con algunos compañeros Mauricio. Trabajamos en donde doy los cursos de programación. Yo trabajé sobretodo en las explicaciones teóricas del principio y en escoger las variables del modelo de los diamantes.