Tarea 3: Regresión Lineal

Valeria Alvarez Navarro

Parte Teórica

Se quiere explicar una variable estadística Yutilizando la información de p variables $X^1 \quad X^2 \quad \dots \quad X^p$. Si se toma una muestra de N observaciones, cada variable está representada por un vector de tamaño N. La información de las variables explicativas se puede juntar en una matriz:

$$X = [X^1 \quad | \quad X^2 \quad | \quad \dots \quad | \quad X^p]$$

de tamaño $n \times p$ donde cada columna es una variable y cada fila es una observación de la muestra.

Plantear el problema de regresión como un problema de mínimos cuadrados, encontrar el vector $\hat{\beta} = \left[\widehat{\beta_1} \quad \widehat{\beta_2} \quad \dots \quad \widehat{\beta_p}\right]'$ que resuelva:

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

y encontrar la solución teórica.

Respuesta:

Se quiere explicar una variable estadística

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{p1} \\ 1 & X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{pn} \end{bmatrix}_{nxp} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{px1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{nx1}$$

La cual se puede expresar como:

$$y = X\beta + \epsilon$$

Plantear el problema como uno de Mínimos Cuadrados Ordinarios (OLS) equivale a encontrar los parámetros $\hat{\beta}$ que minimicen que la suma de los residuos al cuadrado.

El vector de los residuos es: $e = y - X\hat{\beta}$

La suma de los residuos al cuadrado (RSS) es $e'e^2$.

$$[e_1 \quad e_2 \quad \dots \quad e_n]_{1xn} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{nx1} = [e_1 \times e_1 + e_2 \times e_2 + \cdots + e_n \times e_n]_{1x1}$$

Lo anterior se puede escribir como sigue:

$$\begin{aligned} e'e = & \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) \\ = & \; y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}X'X\hat{\beta} \\ = & \; y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Para encontrar el vector $\hat{\beta}$ que minimice la suma de los residuos al cuadrado, es necesario derivar el resultado anterior con respecto a $\hat{\beta}$:

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

Para ver que esto es un mínimo, hay que volver a derivar la ecuación anterior:

$$\left(-2X'y + 2X'X\hat{\beta}\right)' = 2X'X$$

Si las columnas de X son linealmente independientes (full Rank), el resultado anterior es positivo y se tiene un mínimo.

El siguiente paso consiste en despejar $\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Donde la matriz $(X'X)$ es cuadrada $p \times p$ y simétrica.

Si la inversa $(X'X)^{-1}$ y dado que y y X son conocidos, se obtiene la estimador de $\hat{\beta}$ por OLS.

a) ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos?
   Porque los parámetros $\hat{\beta}$ son lineales o los estamos estimando.
b) ¿Se podría utilizar para ajustar polinomios (ej: $y = x^2$)
   Si, si se puede. Hay que notar que el problema se refiere a las $\hat{\beta}$'s es decir que su ajuste es lineal aunque las variables independientes no lo sean.
c) 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?
   La solución encontrada
d) ¿Qué logramos al agregar una columna de unos en la matriz X?

- Que la recta que ajusta el modelo no pase por el origen lo que hace que datos obtengan un mejor ajuste al capturar lo que las $\hat{\beta}$ asociadas a las variables independientes no logran.

Plantear el problema de regresión ahora como un problema de estadística

$$Y_i = \beta_0 + \beta_1 X_i^1 + \cdots + \beta_p X_i^p + \epsilon_i$$

donde los errores son no correlacionados, con distribución:

$$\epsilon_i \sim N(0, \sigma^2)$$

e) ¿Cuál es la función de verosimilitud del problema anterior?
Hint: empezar por escribir el problema como:

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

con

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

Escribir entonces la verosimilitud como: $L(\beta, \sigma^2) = f(Y|\beta, \sigma^2, X)$

Respesta:

$$\mathcal{L}(Y|\beta, \sigma^2, X) = \prod_{i=1}^{n} f(Y_i|\beta, \sigma^2, X)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - x_i\beta)^2\right\}$$

La función de log-verosimilitud es:

$$logL(\beta, \sigma^2; y, X)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

f) Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados.

Tomando la primera derivada parcial de la función de log-verosimilitud en $\beta_0, \beta_1, \ldots, \beta_p$ y σ^2.

$$\frac{\delta logL}{\delta \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i'(y_i - x_i\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i' y_i - \sum_{i=1}^{n} x_i' x_i \beta$$

$$\sum_{i=1}^{n} x_i' y_i - \sum_{i=1}^{n} x_i' x_i \beta = 0$$

$$\beta = \left(\sum_{i=1}^{n} x_i' x_i\right)^{-1} \sum_{i=1}^{n} x_i' y_i = (X'X)^{-1} X'y$$

Al igual que por OLS si se asume que X tiene full Rank entonces $X'X$ es invertible.

La derivada parcial de la log-verosimilitud con respecto a la varianza es:

$$\frac{\partial}{\partial \sigma^2} logL(\beta, \sigma^2; y, X)$$

$$= \frac{\delta}{\delta \sigma^2}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right)$$

$$= -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right]\frac{\delta}{\delta \sigma^2}\left(\frac{1}{\sigma^2}\right) = -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right]\left(-\frac{1}{(\sigma^2)^2}\right)$$

$$= -\frac{n}{2\sigma^2} + \left[\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta)^2\right]\left(\frac{1}{(\sigma^2)^2}\right) = \frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2 - n\right]$$

El resultado anterior es igual a cero, sólo si:

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

De lo anterior:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2$$

Al resolver estas ecuaciones, se llega a los mismos estimadores obtenidos por OLS.

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

Respuesta:
El Teorema de Gauss-Markov establece que de entre los estimadores insesgados de β, el obtenido por mínimos cuadrados es el que tiene varianza mínima y que es BLUE (best linear unbiased estimate).

Teorema de Gauss-Markov:
Sea $Y = X\beta + W$ donde $\beta \in \mathbb{R}^p$ es un (no conocido) vector determinístico, X una matriz conocida determinística de tamaño $n \times p$ con rango p, y ϵ W $\mathbb{R}^n$ es un vector aleatorio de varianza cero y matriz de covarianza $\sigma^2 I$. Sea $\widehat{\beta}$: $\mathbb{R}^n \to \mathbb{R}^p$ una función definida por:

$$\hat{\beta}(z) = (X'X)^{-1}X'z$$

Entonces, $\hat{\beta}(Y) \in \mathbb{R}^p$ es un vector aleatorio con las siguientes propiedades:

(i) $\mathrm{E}\left[\hat{\beta}(Y)\right] = \beta$, esto es $\hat{\beta}$ es un estimador insesgado de β.
(ii) $\mathrm{cov}\left[\hat{\beta}(Y)\right] = \sigma^2(X'X)^{-1}$
(iii) Sea $\tilde{\beta}$: $\mathbb{R}^n \to \mathbb{R}^p$ otra función lineal (que puede ser escrita como $\tilde{\beta}(z) = Pz$), que es un estimador insesgado de β tal que $\mathrm{E}\left[\hat{\beta}(Y)\right] = \beta$; entonces,

$$cov\left[\tilde{\beta}(Y)\right] \geq cov\left[\hat{\beta}(Y)\right]$$

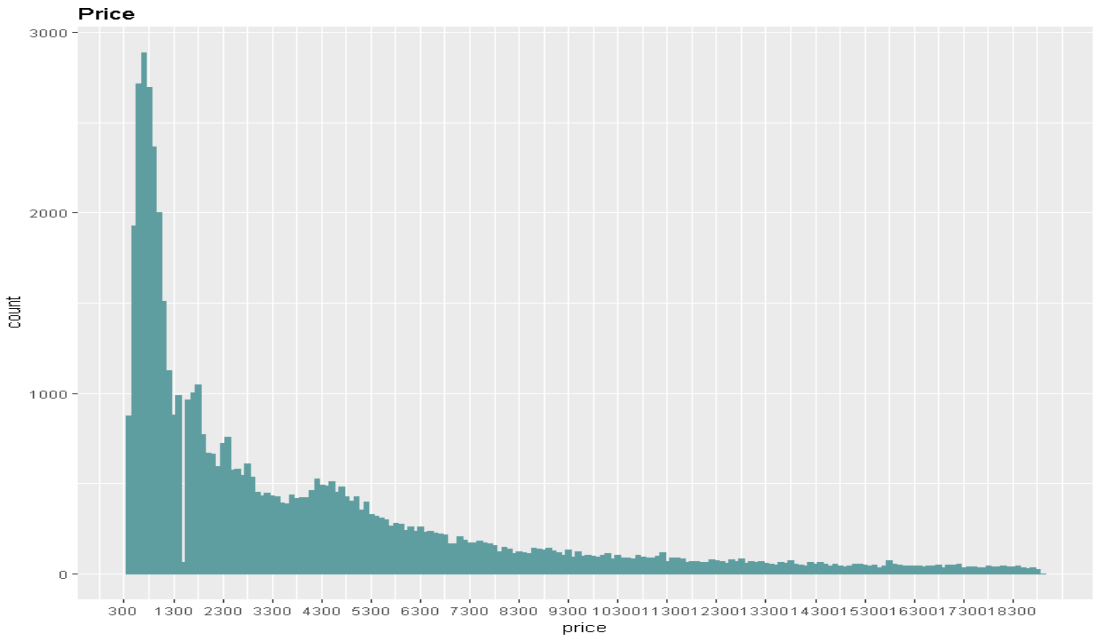
Parte Aplicada

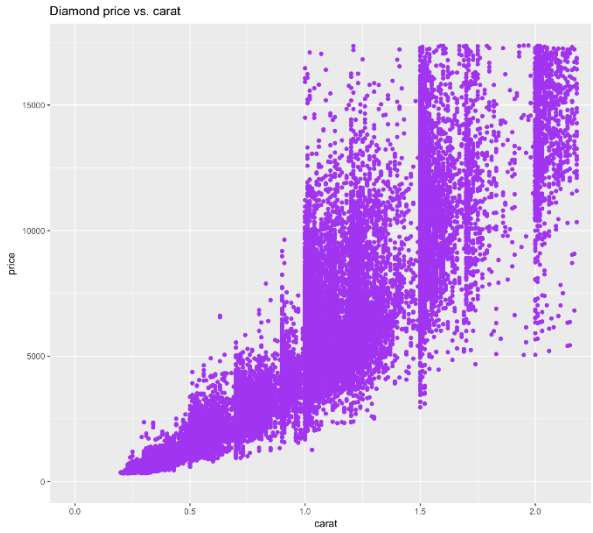
Regresión lineal en R

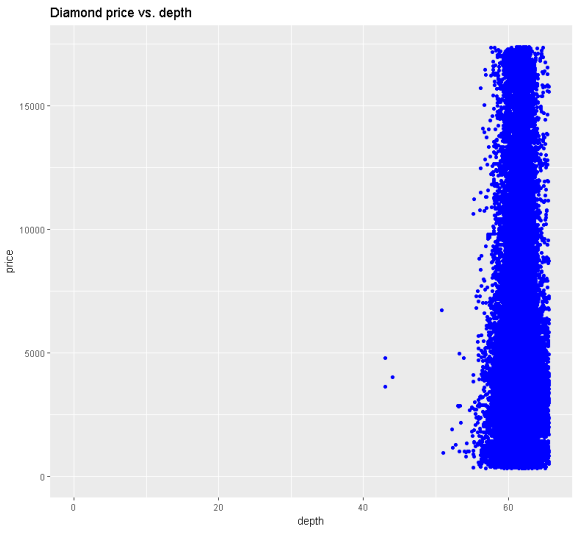
Cargar la base diamonds que se encuentra en el paquete gglot2.

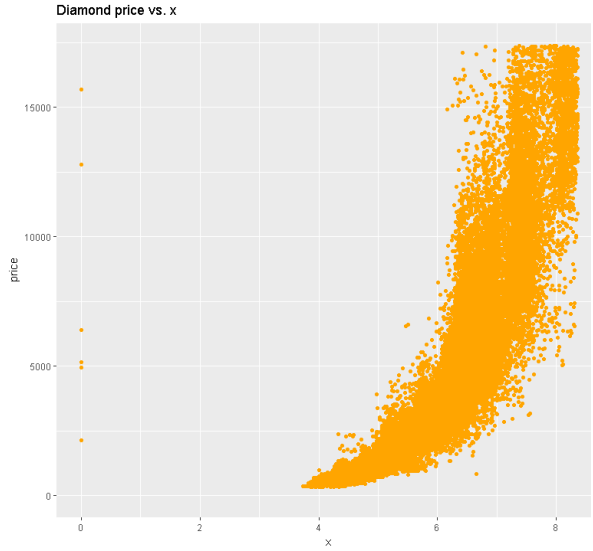
Posteriormente, hacer una regresión lineal. El objetivo es explicar la variable $price$ usando las demás variables. Notar que algunas variables no son numéricas, por lo que no pueden incluirse en el análisis crudo de la regresión lineal. Para este proyecto NO es necesario saber transformar las variables no numéricas para saberlas usar en la regresión. Se puede usar la función lm de R para hacer el análisis de regresión.

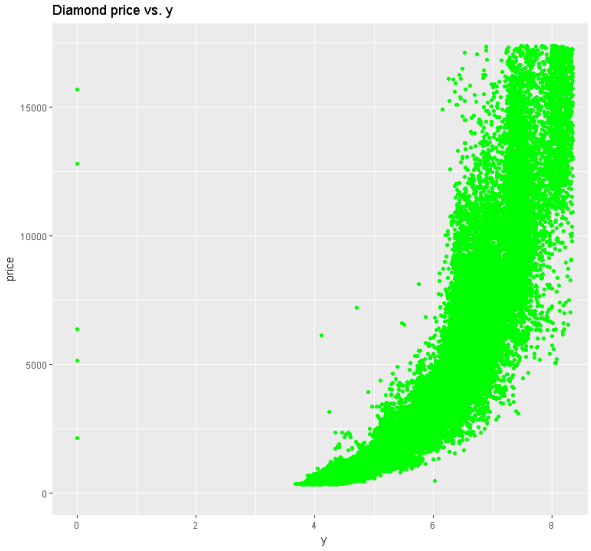
Respuesta:
Para poder hacer una regresión lineal, es recomendable realizar primero un análisis de los datos. Esto ayudará a tener un conocimiento a priori de qué variables serán útiles para ajustar la variable dependiente.

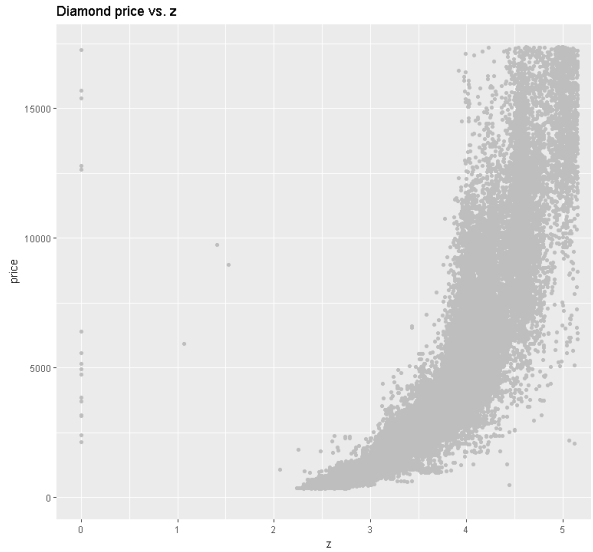


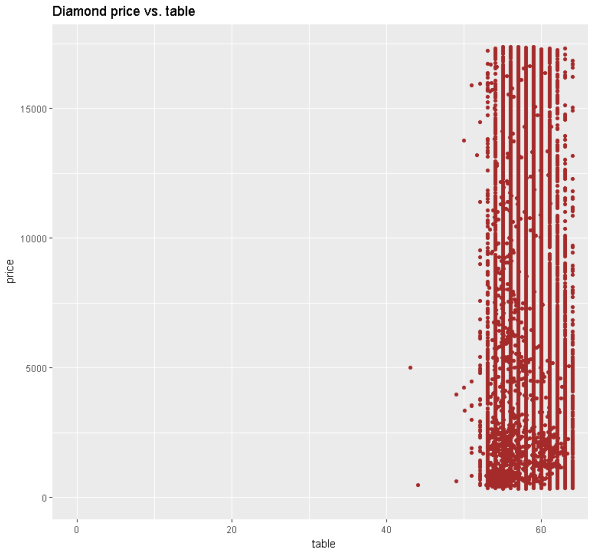

Diamond price vs. carat
price
carat
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0


Diamond price vs. depth
price
depth
15000
10000
5000
0
0
20
40
60

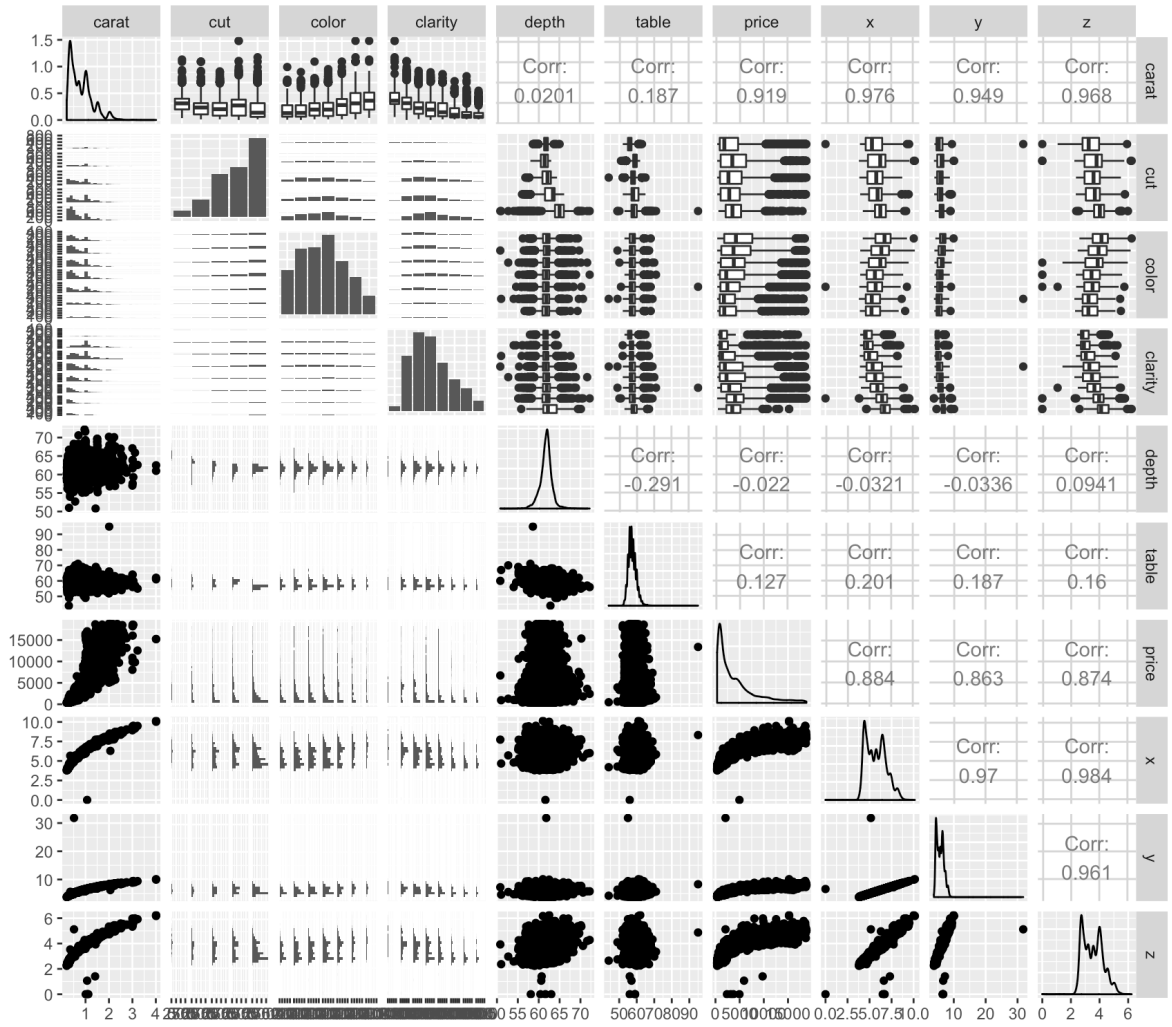

Diamond price vs. x
price
x
15000
10000
5000
0
0
2
4
6
8


Diamond price vs. y
price
y
15000
10000
5000
0
0
2
4
6
8


Diamond price vs. z
price
z
15000
10000
5000
0
0
1
2
3
4
5


Diamond price vs. table
price
table
15000
10000
5000
0
0
20
40
60

Se hace una gráfica de pares de variables para ver cómo están relacionadas.



Con esta gráfica se comprueba que el precio de los diamantes muestra fuerte correlación con x, y, z y con carat. Asímismo, carat también se encuentra correlacionada con x, y, z. Estas relaciones son casi lineales. Por otra parte las variables depth y table no muestran una relación clara con ninguna otra variable por lo que posiblemente no sean significativas para el modelo.

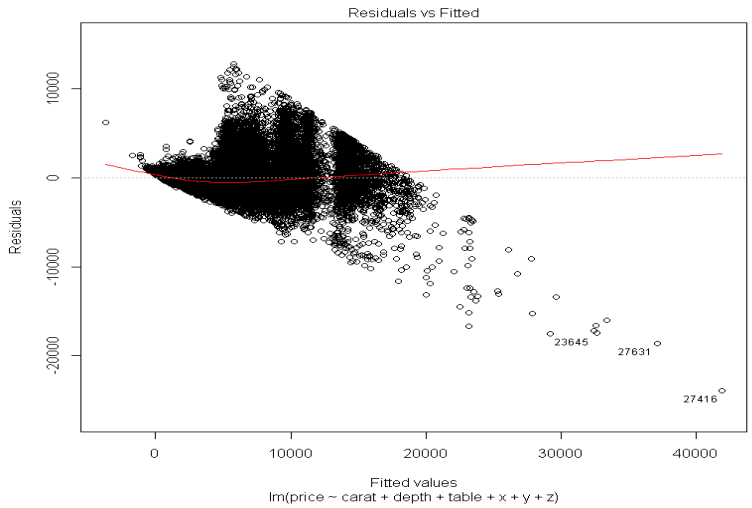
A continuación se muestran los modelos que se corrieron para estimar el precio:

Modelo con todas las variables numéricas:

```
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
```



Del resumen de la regresión, se puede apreciar que no se tiene un buen ajuste, pues el error residual es grande. Visualmente, la línea roja debería estar lo más pegada al cero posible. Por otra parte, los coeficientes de la regresión indican que la variable "z" no es significativa, la variable "y" permanecer significativa pero puede presentar algunos problemas. A pesar de que las variables "depth" y "table" son significativas según los resultados, en el análisis de los datos se visualizaba diferente.

Lo que se puede hacer para mejorar el ajuste es i) hacer una transformación a la variable dependiente para "estandarizar" sus incrementos o cambios; ii) quitar las variables que no parezcan significativas. En cuanto a esta segunda parte, al quitar variables, la estimación no mejoró y el error estándar creció. Por lo que aun cuando la R^2 sea alta, el modelo no parece

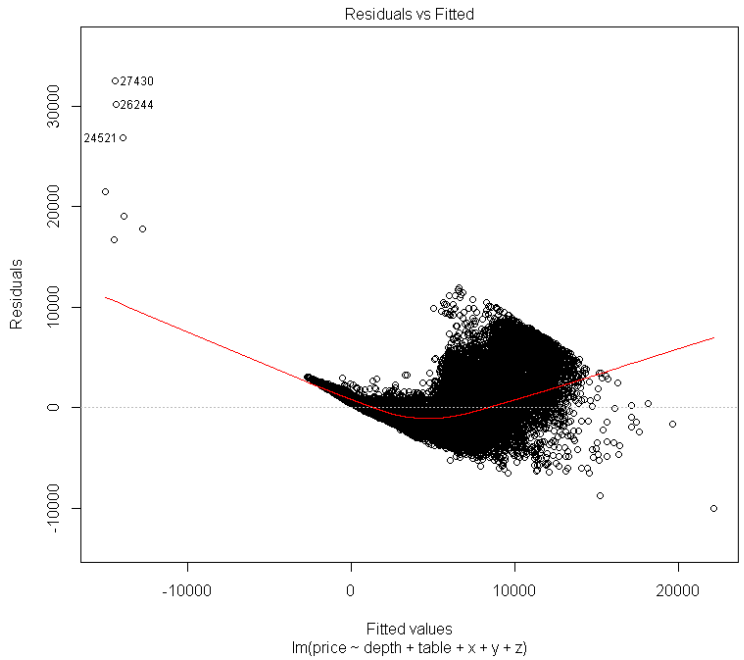
ajustar muy bien. Esto se puede deber a que la variable que parece explicar más el precio de los diamantes es “carat”, la cual no parece tener una relación lineal con el precio, sin embargo la β esta capturando la parte lineal que existe. Es posible que haya que hacer también una transformación a la variable “carat” para encontrar esta relación con el precio (también transformado), aun así “carat” es la variable más importante del modelo por lo que eliminarla incrementaría significativamente los residuales; mientras que eliminar cualquier otra variable no impactaría mucho la estimación.

```
lm(formula = price ~ depth + table + x + y + z, data = diamonds)

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
```



La R^2 disminuye signficativamente

Se recuerda que en este ejercicio no se están considerando las variables cualitativas, que pueden dar más información acerca del precio de los diamantes.

A pesar de que la R^2 es la medida que daría el ajuste al modelo, a veces puede no ser precisa, si el error es muy grande. En este caso nos diría que nuestro modelo está explicado al 85% por las variables que introdujimos, sin embargo por los errores mostrados se puede ver que la estimación no es muy buena a pesar de que los parámetros parecen significativos.

¿Cuál es el ángulo entre Y y $\hat{Y}$?

El ángulo es 22.03873, que equivale a obtener el arcosen de la raíz de R^2

Definan una función que calcula la log-verosimilitud de unos parámetros
El siguiente ejercicio es propiedad del compañero Leonardo Marín quien hizo favor de realizarlo en equipo con algunos compañeros.

```
parametros <- as.vector(parametros)
 matriz.x <- cbind(1,X)  ## Agregar la columna de unos (1s)
 matriz.x <- as.matrix(matriz.x)
 Y <- as.vector(Y)

# En el objeto parámetros, incluir los coeficientes de beta a estimar más la varianza al en la última entrada del vector
  k <- length(parametros)
  beta <- parametros[c(1:(k-1))]    # Extrae los elementos excepto el último
  varianza <- parametros[k]        # Extrae el último elemento

# P es en número de parámetros a estimar
p <- length(beta)

## Notar que en R ya existe por default el valor para pi = 3.14

error <- sum((Y - matriz.x%*%beta)^2)

# Nota: La función "optim" minimiza por default, definir nuestra función con un menos
# Nota: se definió de forma positiva "mle" pues la versión original tiene signo negativo, esto para poder minimizar
mle <-  (p/2)*(log(2*pi)+log(varianza)) + (1/(2*varianza))*error

return(mle)

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

```
## Utilizar la función

## iniciar los parámetros
# Deben ser 7 betas + 1 varianza
valores.iniciales <- c(100,100,100,100,100,100,100,100)
length(valores.iniciales)

## Checar que la función definida previamente se ejecute e forma correcta
log.mle(valores.iniciales)

## Utilizar la función optim
optim(par=valores.iniciales, log.mle, method = "L-BFGS-B")

## Comparar con los resultados de la función lm
diamonds.lm$coefficients

# Parece que la solución encontró óptimos locales

#############################################################################
### Utilizar algunos valores iniciales cercanos a los óptimos

beta <- as.vector(diamonds.lm$coefficients)
matriz.x <- as.matrix(cbind(1,datos.X))  ## Agregar la columna de 1s

## Sabemos que el estimador por MLE (sesgado) de la varianz es:
varianza <-  sum((Y - matriz.x%*%beta)^2)/(n - p)

################
casi.optimos <- c(diamonds.lm$coefficients,varianza) + 1000
casi.optimos

optim(par=casi.optimos, log.mle, method = "L-BFGS-B")

###############
## Comparar con los resultados de la función lm
diamonds.lm$coefficients
varianza


## Notar que para la variables  depth, table, x, y , z se tomó un valor inicial alejado del optimo y la función los aproxima a los ya conocidos y correctos
```