Tarea 3: Regresión Lineal

Parte teórica:

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.

$$SSR = (y - Xb)^t (y - Xb)$$

La condición del primer orden para un minimo es que el gradiente de SSR con respecto de beta debe ser cero.

$$\bigtriangledown_\beta SSR = 0$$

eso es:

$$-X^t (y - Xb) = 0$$

$$(X^t X) b = X^t y$$

Si X tiene rango completo entonces

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

¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos?

R= Porque los parámetros $$ representan la parte lineal de nuestra variable a aproximar Y en base a otra variable X. Esto es más fácil de visualizar en un ejemplo de dos dimensiones donde la fórmula para estimar Y es $\hat{Y} = B_0 + B_1 * X$ y notamos que es similar a la de una recta, donde el parámetro B_0 representa la ordenada al origen y B_1 la pendiente.

¿Podríamos usarlo para ajustar polinomios?

R= Sí, podemos ajustar polinomios del modo $y = x^2$ y otros sin ningún problema utilizando el resultado de regresión lineal. Se puede inferir de la demostración previa, que en la regresión polinomial el problema de estimación sigue siendo lineal (Beta seguiría siendo lineal, la matriz X es la que tedría dentro valos de x^k).

- 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?

R= Lo que hace el vector β es expandir el plano X, para que sobre este plano encontremos un valoy $\hat{Y}$ que es la proyección del Y en el plano X, pudiendo formar un triangulo donde el error de estimación o residual es el cateto opuesto.

También la varianza de la variable dependiente se puede descomponer en dos varianzas, la del pronóstico y la del error. 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

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

R= Al agregar la columna de 1´s lo que logramos es tener la estimación de la intercepción de Y. De esta manera, el estimador $\hat{Y}$ no necesariamente inicia desde el origen.

- 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.

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 3.4.1
```

```
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
```

```
## starting httpd help server ...
```

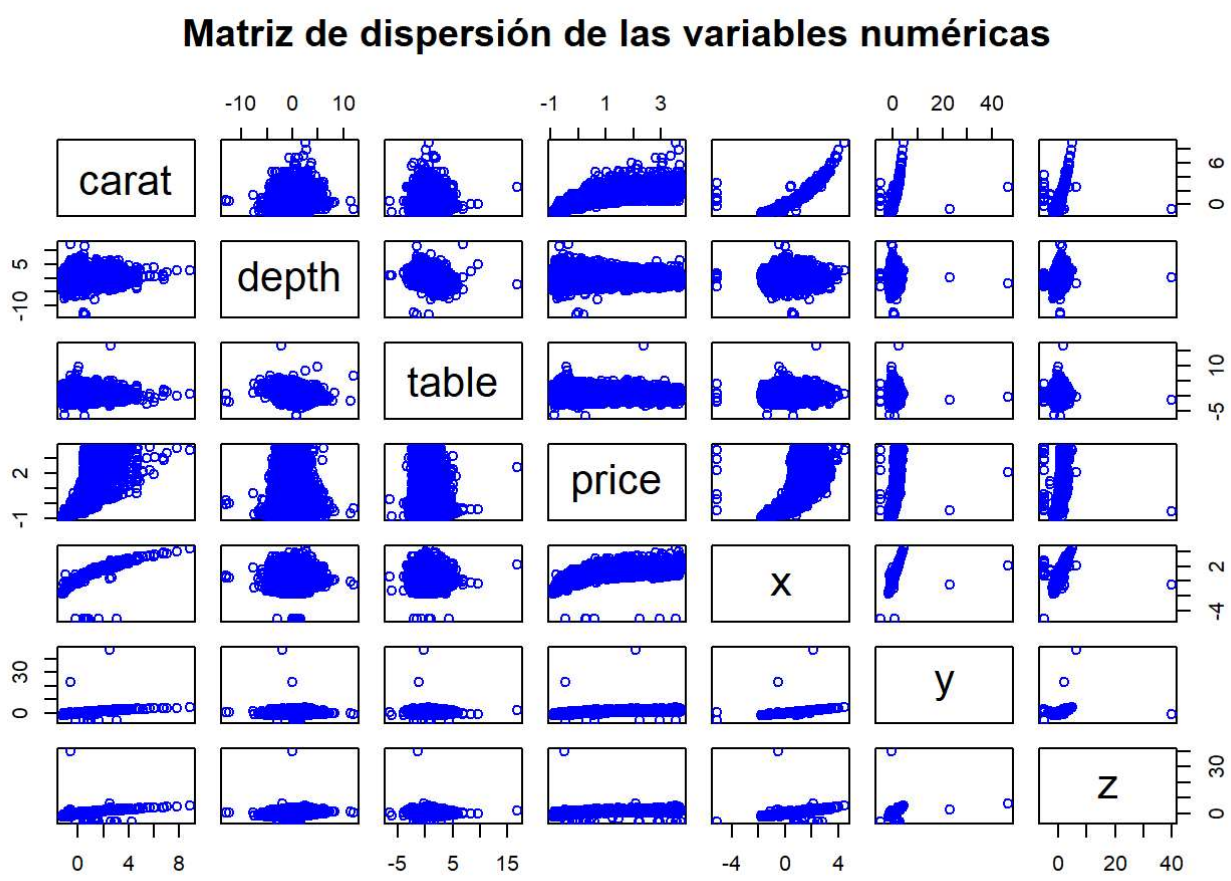
```
##  done
```

Ahora vamos a crear un nuevo data frame que solo contiene las variables numéricas centralizadas.

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

Obtenemos una matriz de dispersión para formarnos una apreciación inicial de las relaciones entre las variables.

```
pairs(diamantes0, col= "blue", main="Matriz de dispersión de las variables numéricas")
```



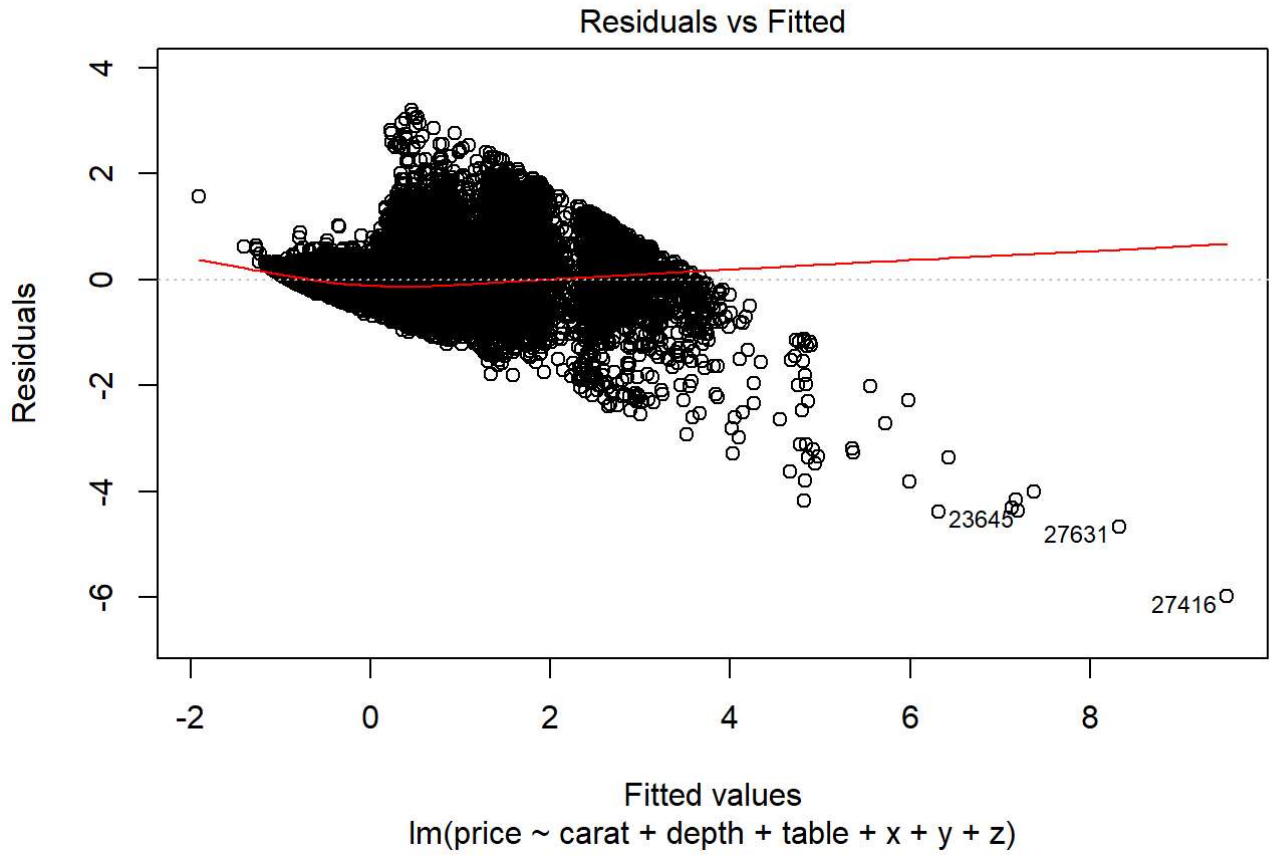
En la matriz de gráficas podemos ver de manera rápida que la variables que más se relacionan con el precio son carat.

Ahora vamos a correr la regresión lineal:

```
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
```

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



- ¿Qué tan bueno fue el ajuste?

R= En el resumen de la regresión podemos ver los residuales, que son las diferencias entre el precio estimado y el precio real; donde una media lo más cercana a cero es lo ideal, pues indicaría que el modelo se ajusta a la perfección. En la gráfica vemos los valores predichos contra los residuales, en un modelo perfecto, la media de los residuales sería 0 por lo que los puntos y la línea roja irían perfectamente a lo largo de la línea que representa el 0. En seguida podemos apreciar los coeficientes de la intercepción y de cada una de las variables, usados para calcular las predicciones.Así también 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, lo que buscamos es que este sea lo más pequeño posible.

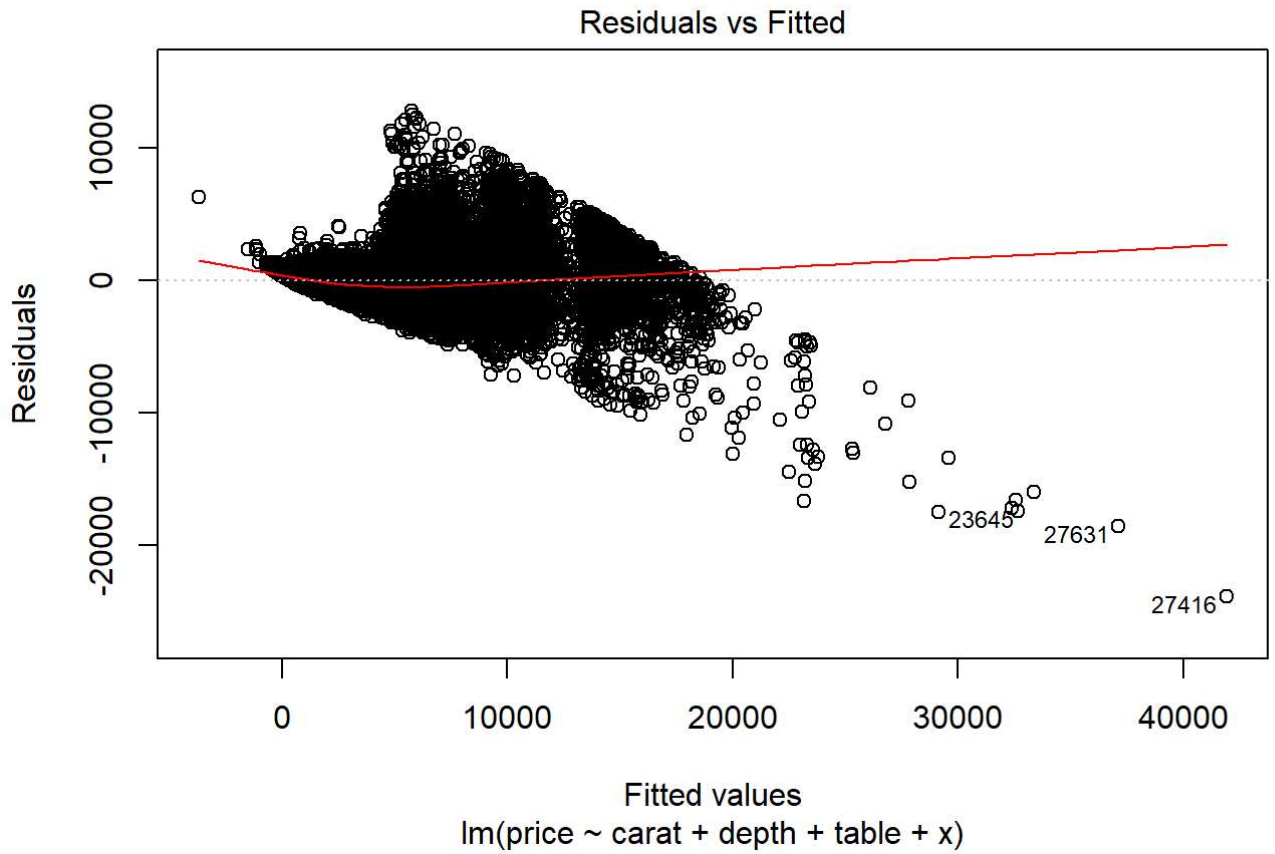
- ¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cúal fue el valor de $ ^2 $ que ajustó su modelo y que relación tienen con la calidad del ajuste?

R= En el resumen podemos ver la R cuadrada, que es la cantidad explicada por el modelo, en este caso 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.

```
# Por curiosidad, realizamos otra regresión 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
```

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

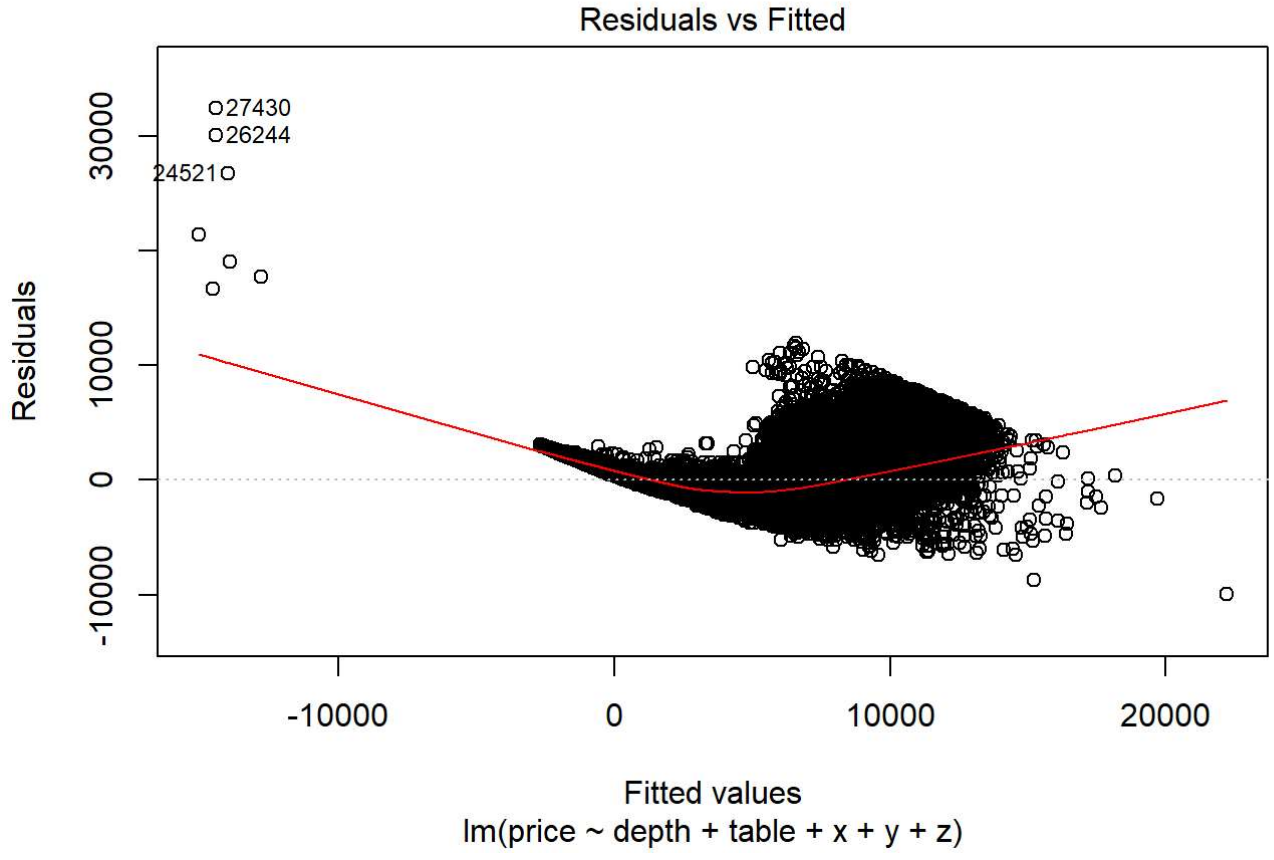


Aqui vemos, como era esperado, que deshacernos de las dos variables que no eran relevantes prácticamente no modifica valores como la media de los residuales o la R cuadrada.

```
# Por curiosidad, ahora realizamos otra regresión 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
```

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



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

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

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

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