

Plantear el problema de regresión como un problema de mínimos cuadrados, encontrar el vector $\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_p]^T$ que resuelva:

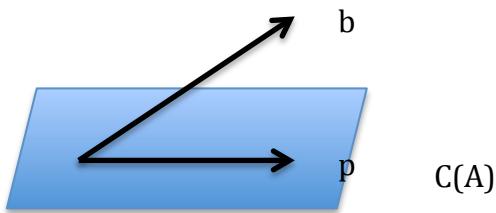
$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \| Y - X\beta \|^2$$

y encontrar la solución teórica. ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos? ¿Podríamos usarlo para ajustar polinomios (ej $y = x^2$)?

R= Dado que queremos estimar una regresión lineal para determinar la influencia de todas las variables p en la estatura de una persona. En ese sentido, plantearíamos una ecuación, a estimar, con una variable dependiente y p independientes de la siguiente manera:

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_p X_{tp}$$

Dado lo anterior podríamos simplificar a la forma $Y = X\beta$ donde Y es el vector columna que contiene la estatura X representa una matriz como la descrita en la introducción y otro vector β con p elementos. Utilizando una notación más conocida, podemos cambiar el nombre de las variables para expresarlo de la manera $Ax = b$. En otras palabras, estamos buscando el vector de las betas que al multiplicarlo por la matriz A (matriz que contiene columnas que representan variables y filas que representan cada observación, es decir cada uno de los individuos de la muestra) nos dé como resultado b . Es evidente que no tenemos una solución para la ecuación anterior, en otras palabras no existe una combinación lineal que nos permita encontrar el vector b . El vector b no se encuentra en el espacio columna de la matriz A . Gráficamente podríamos verlo de la siguiente manera.



Imaginemos que el paralelogramo azul es el espacio columna de la matriz A . La flecha negra es el vector b . Claramente, el vector b no pertenece al espacio columna de la

matriz A. En consecuencia, existe la necesidad de encontrar una aproximación de solución para este caso la gráfica denomina a este vector “solución aproximada” como p. La solución p es lo más cercano a b. No se resuelve el problema original, pero la distancia entre b y p es mínima y además el vector p sí es una combinación lineal de los vectores del espacio columna de la matriz A. Dado lo anterior intuitivamente tenemos que $\|b - Ap\|$ es la distancia es la mínima. Es claro que estamos intentando generar una proyección del vector b sobre el espacio columna de A. Para hallar la solución $Ap = v$, entonces:

$$\left\| \begin{array}{l} b_1 - v_1 \\ b_2 - v_2 \\ \vdots \\ b_n - v_n \end{array} \right\|^2 = (b_1 - v_1)^2 + (b_2 - v_2)^2 + \cdots + (b_n - v_n)^2$$

Se minimiza la suma de cuadrados presentada anteriormente descrita. Despejando nuestra ecuación original tenemos que

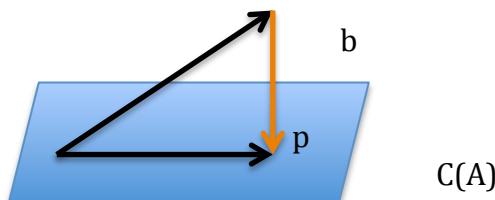
$$\begin{aligned} Ap - b &= \text{Proyección}_{C(A)}b - b \\ A^T(Ap - b) &= 0 \\ A^TAp - A^Tb &= 0 \\ A^TAp &= A^Tb \\ (A^TA)^{-1}(A^TAp) &= (A^TA)^{-1}A^Tb \\ p &= (A^TA)^{-1}A^Tb \end{aligned}$$

Dado que minimizamos la distancia entre b y v al cuadrado es que este método se denomina mínimos cuadrados.

Respecto a si podemos agregar o utilizar χ^2 , sí es posible porque los parámetros a estimar siguen siendo lineales, aunque las variables estén transformadas. La solución para encontrar los parámetros es la misma.

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= Retomando la respuesta 1, donde se planteó el problema de mínimos cuadrados y como solucionarlo tenemos lo siguiente:



Anteriormente se comentó que b no estaba en el espacio columna de la matriz A y por lo tanto el sistema $Ax = b$ no tenía una solución. En ese sentido, el método de mínimos cuadrados nos permite hacer la proyección de b sobre el espacio columna de A (en la gráfica p) que sí se encuentra en el espacio columna de A y por lo tanto es una combinación lineal de los elementos de A . Ahondando, para hallar la solución debemos restar el vector v con el vector p cuya solución es el vector que apunta de uno a otro (vector naranja en la gráfica anterior) y que además es ortogonal a todo el espacio columna. Del punto de intersección de ambos vectores se forma un ángulo de 90 grados, en combinación forman un triángulo rectángulo consistente con el teorema de Pitágoras donde: $a^2 = b^2 + c^2$, que además es consistente con que el error sigue siendo mínimo.

Entonces:

$$\begin{aligned} Ap - b &\in C(A)^\perp \\ C(A)^\perp &= N(A^T) \\ Ap - b &\in N(A^T) \end{aligned}$$

El complemento ortogonal es equivalente al espacio nulo izquierdo de la matriz A , por lo tanto $Ap - b$ se encuentra en el espacio nulo izquierdo de la matriz A . Dado lo anterior la multiplicación $A^T(Ap - b) = 0$ es igual a cero. Dada la combinación entre la respuesta 1 y lo explicado aquí, la solución por mínimos cuadrados es también una solución en un problema de subespacios vectoriales.

¿Qué logramos al agregar una columna de unos en la matriz X?

R= Retomando las conclusiones de la pregunta anterior, en este ocasión tendríamos una columna adicional de 1 en la matriz X . En ese caso la ecuación a estimar sería:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_p X_{tp}$$

En otras palabras, estaríamos agregando el intercepto u ordenada al origen. Dicha variable nueva representa, el valor promedio de la variable dependiente (en este caso la estatura) cuando todas las demás variables explicativas son cero.

Las nuevas matrices quedarían de la siguiente forma:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{12} & \dots & X_{1K} \\ \vdots & \vdots & \dots & \vdots \\ 1 & X_{N2} & \dots & X_{NK} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}$$

Para efectos prácticos la fórmula desarrollada para la estimación de una solución "cercana" permanecería sin cambios respecto a la pregunta anterior.

Plantear el problema de regresión ahora como un problema de estadística.
¿Cuál es la función de verosimilitud del problema anterior? Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados.

R= Se aprovechará para contestar dos preguntas en la misma respuesta. El modelo a estimar sería el siguiente:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_p X_{tp} + e_t$$

Dado que de acuerdo a los supuestos el término del error tiene una distribución normal, se puede usar la ecuación de la distribución normal para estimar las betas del modelo junto con el despeje del mismo en la forma $y_t = x_t\beta$.

$$N(\varepsilon_t; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y_t - x_t\beta)^2}$$

Por razones de espacio se usó una fuente mayor en la distribución anterior. Es claro que si tenemos muchas observaciones para y_t además de los vectores de x_t , estos tienen funciones de densidad $N(y_t; x_t\beta, \sigma^2)$, que tienen la misma distribución que el término del error. En ese sentido, se puede plantear una función de máxima verosimilitud para los parámetros β, σ^2 , de la siguiente forma:

$$L = \prod_{T=1}^T N(y_t; x_t\beta, \sigma^2) = (2\pi\sigma^2)^{-T/2} e^{\frac{-1}{2\sigma^2}(y_t - x_t\beta)^t(y_t - x_t\beta)}$$

Para la estimación se debe tomar el logaritmo natural de la función anterior.

$$L^*(\beta, \sigma) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y_t - x_t\beta)^t(y_t - x_t\beta)$$

Dado que ya se cuenta con L^* , a través de la primera y segunda derivada se pueden estimar los parámetros que no se conocen. Las derivadas quedan de la siguiente manera:

$$\frac{\partial L^*}{\partial \beta} = \frac{1}{\sigma^2}(y_t - x_t\beta)^t X$$

$$\frac{\partial(L^*/\partial\beta)^t}{\partial\beta} = -\frac{1}{\sigma^2}X^t X$$

Con las derivadas calculadas se igualan a cero. Esto nos entrega las condiciones de primer orden para la maximización de la función de máxima verosimilitud. Con estas condiciones encontramos unos estimadores de la forma:

$$\check{\beta} = (X^t X)^{-1} X^t Y$$

Como podemos ver es claro que a través de este método de estimación se llega a una solución idéntica al método de álgebra lineal de mínimos cuadrados.

Investiga el contenido del teorema de Gauss-Markov sobre mínimos:

R=El teorema de Guass-Markov sostiene que en un modelo de regresión lineal en cual los errores se distribuyen de la forma $e \sim N(0, \sigma^2)$, el Mejor Estimador Lineal Insesgado (MELI) se obtiene aplicando el método de mínimos cuadrados ordinarios. El concepto de “mejor” se asocia a que dicho estimador da la menor varianza posible, comparado con otros parámetros estimados. Para que el teorema tenga validez se asumen los siguientes supuestos:

1. El modelo poblacional es lineal en sus parámetros:
 $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$
2. La muestra aleatoria de tamaño n ,
 $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i=1, 2, \dots, n\}$, es representativa de la población, de modo que el modelo muestral es:
 $y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_kx_{ik} + e_i$
3. $E(e/x_1, x_2, \dots, x_k) = 0$, lo cual implica que todas las variables explicativas son exógenas (no endogeneidad).
4. Ninguna variable x es constante ni tiene una correlación lineal exacta con otra (no multicolinealidad).
5. La varianza del error es constante $Var(e) = \sigma^2$.
6. El término de error es independiente y no correlacionado $Cov(e_i, e_j) = E(e_i, e_j) = 0, i \neq j$.

Parte Aplicada

¿Qué tan bueno fue el ajuste? ¿Qué medida puede ayudarnos a saber la calidad del ajuste?

R= Empezamos analizando la estructura de la base de datos diamonds de la siguiente manera:

```

983 0.70 Ideal G VVS2 b2.1 53.0 2895 5.71 5.75 3.56
984 0.74 Very Good G VS1 59.8 58.0 2896 5.85 5.89 3.51
985 0.77 Very Good G VS2 61.3 60.0 2896 5.81 5.91 3.59
986 0.77 Very Good G VS2 58.3 63.0 2896 6.00 6.05 3.51
987 0.53 Ideal F VVS1 61.6 56.0 2896 5.18 5.24 3.21
988 0.79 Ideal D SI1 61.5 56.0 2896 5.91 5.96 3.65
989 0.73 Ideal E SI2 61.5 55.0 2896 5.82 5.86 3.59
990 0.77 Ideal D SI2 62.1 56.0 2896 5.83 5.89 3.64
991 0.77 Premium E SI1 60.9 58.0 2896 5.94 5.88 3.60
992 1.01 Very Good I I1 63.1 57.0 2896 6.39 6.35 4.02
993 1.01 Ideal I I1 61.5 57.0 2896 6.46 6.45 3.97
994 0.60 Very Good D VVS2 60.6 57.0 2897 5.48 5.51 3.33
995 0.76 Premium E SI1 61.1 58.0 2897 5.91 5.85 3.59
996 0.54 Ideal D VVS2 61.4 52.0 2897 5.36 5.34 3.26
997 0.72 Ideal E SI1 62.5 55.0 2897 5.69 5.74 3.57
998 0.72 Good F VS1 59.4 61.0 2897 5.82 5.89 3.48
999 0.74 Premium D VS2 61.8 58.0 2897 5.81 5.77 3.58
1000 1.12 Premium J SI2 60.6 59.0 2898 6.68 6.61 4.03
[ reached getOption("max.print") -- omitted 52940 rows ]
> str(diamonds)
Classes 'tbl_df', 'tbl' and 'data.frame':      53940 obs. of  10 variables:
 $ carat : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut   : Ord.factor w/ 5 levels "Fair",<,"Good",<,"Very Good",< ...
 $ color : Ord.factor w/ 7 levels "D","E","F","G",<,"H",< ...
 $ clarity: Ord.factor w/ 8 levels "I1","SI2","SI1",<,:2,3,5,4,2,6,7,3,4,5 ...
 $ depth : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61.65 58.57 57.55 61.61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x     : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y     : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z     : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...

```

Después procedemos a desarrollar el modelo de la siguiente manera (como el solicitado en la pregunta de este apartado):

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_p X_{tp} + e_t$$

Si utilizamos la función names (diamonds) y str(diamonds) nos damos cuenta que la base de datos tiene 10 variables, pero no todas numéricas. Las variables numéricas son:

"carat" "depth" "table" "price" "x" "y" "z"

Las variables categóricas son:
"cut" "color" "clarity"

Se procede a utilizar el comando lm para realizar la regresión solicitada:

regresión <- lm(price~carat+depth+table+x+y+z, data=diamonds)

summary(regresion)

```

> regresion<- lm(price~carat+depth+table+x+y+z, data=diamonds)
> summary(regresion)

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

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

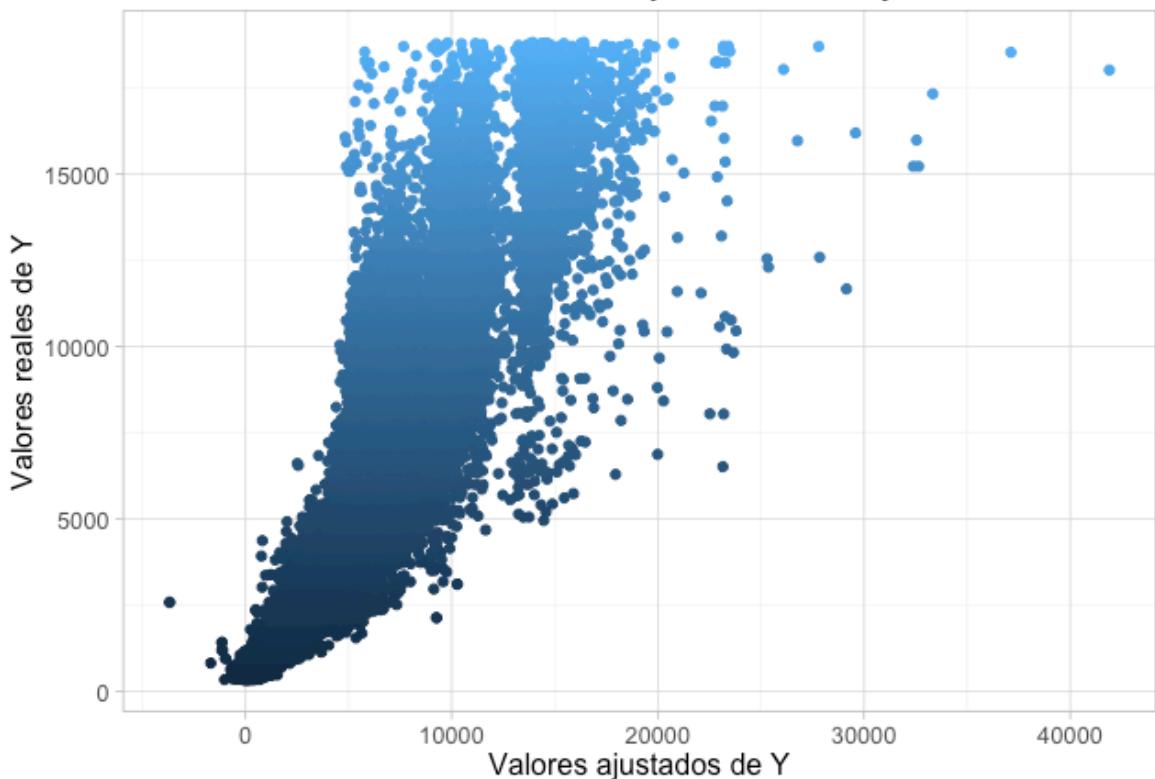
Residual standard error: 1497 on 53933 degrees of freedom
Multiple R-squared:  0.8592,    Adjusted R-squared:  0.8592 
F-statistic: 5.486e+04 on 6 and 53933 DF,  p-value: < 2.2e-16

```

> |

Para empezar un análisis de qué tan bueno fue el ajuste vamos a graficar los valores del precio de los diamantes con los valores ajustados por el modelo de regresión lineal. Si hubiéramos tenido un ajuste perfecto, tendríamos una línea

Precio de los diamantes y sus valores ajustados

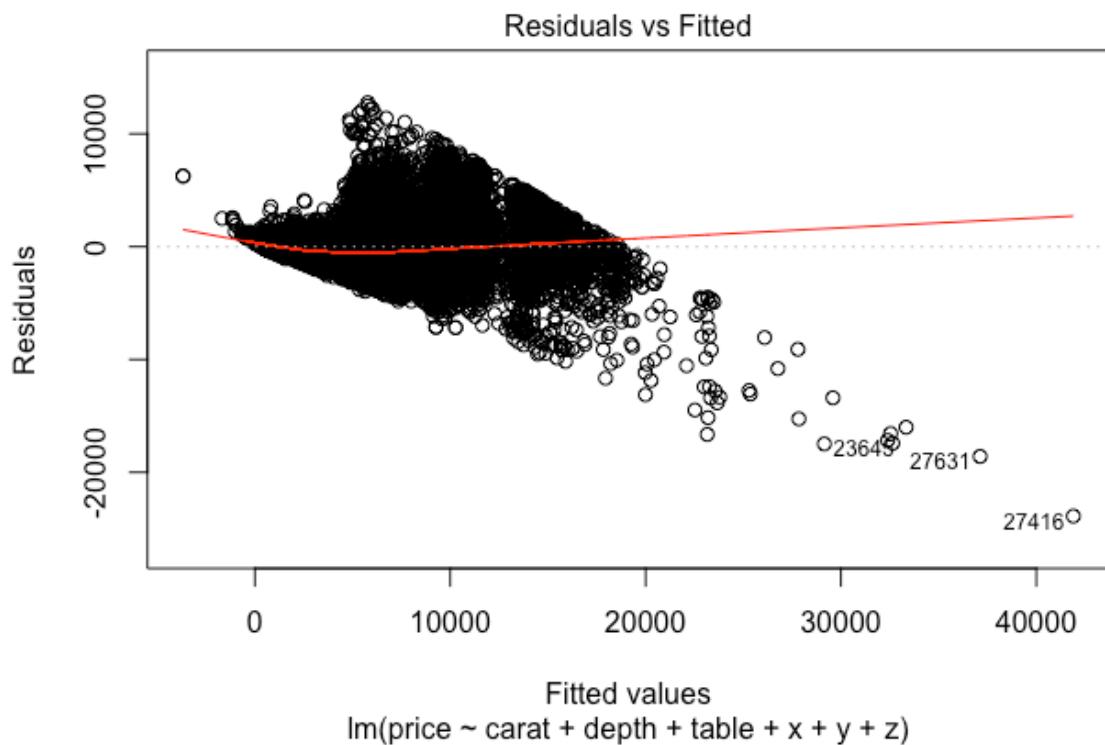


El ajuste está dado por la R^2 de la tabla de estimación mostrada anteriormente. Dicho número representa la variación de la variable dependiente explicada por la variación de las variables independientes.

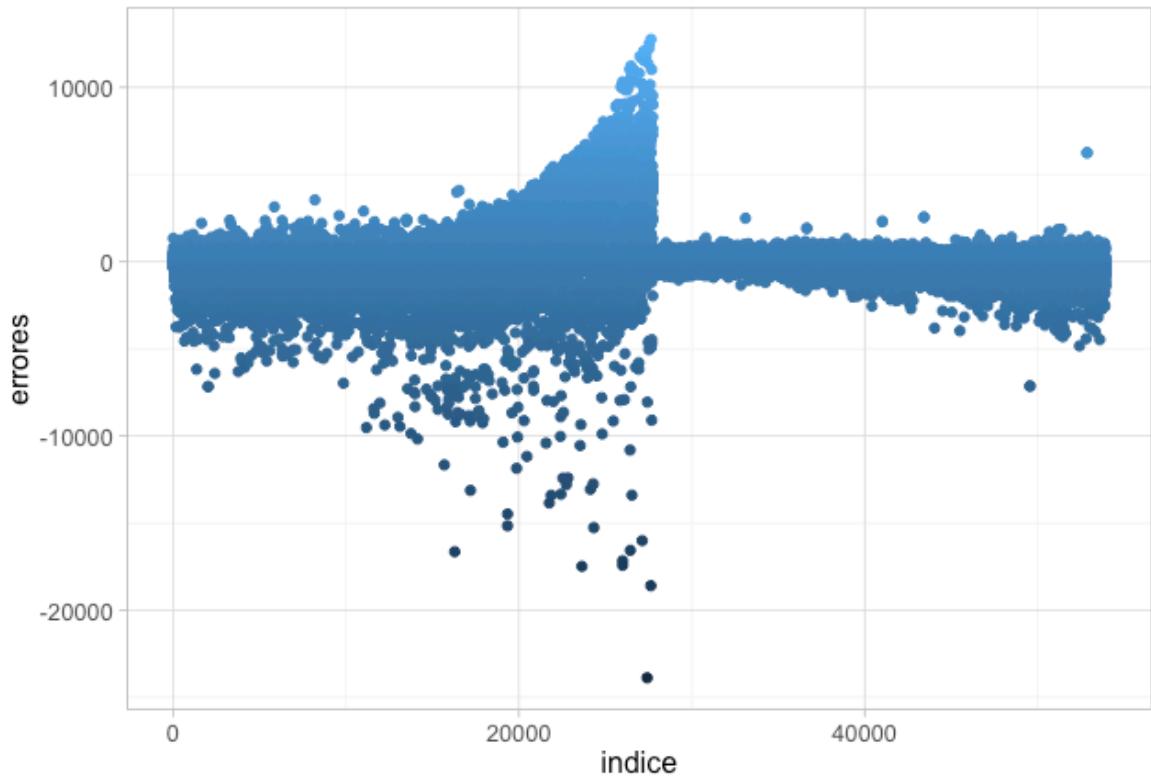
$$R^2 = 1 - \frac{\text{variación del error}}{\text{variación total en } y} = 1 - \frac{\sum \hat{e}_t^2}{\sum (y_t - \bar{y})^2}$$

Dadas las variables numéricas en la estimación (carat, depth, table, x, y, z) la variación de la variable dependiente explicada por la variación de estas variables es de 0.8592.

Otra forma de ver el ajuste es a través del error. En ese sentido, un buen ajuste a través de una regresión lineal significa una distancia mínima entre la estimación y los valores originales del precio del diamante.



Distribucion de los errores



¿Cuál fue el valor de sigma^2 que ajustó su modelo y que relación tiene con la calidad del ajuste?

R= El valor de sigma cuadrada estimada está dada por la siguiente ecuación:

$$\hat{\sigma}^2 = \frac{\hat{e}' \hat{e}}{T - K} = \frac{\sum \hat{e}_t^2}{T - K}$$

Donde T es el tamaño de la muestra y K el número de parámetros estimados. También $\hat{e} = y - Xb$.

Entre menor se la varianza del error mejor será nuestra estimación (mayor calidad) ya que estaremos prediciendo con más precisión los valores poblacionales de lo que estemos analizando a partir de la muestra dada. Una varianza grande de los errores quiere decir que nuestra línea de regresión no es confiable. En este caso R nos estima directamente la raíz cuadrada de sigma cuadrada número que podemos ver en el Residual Standard Error que para este caso es de 1497.

Incluso podríamos llamar el comando `summary.lm(regresion)$sigma` obteniendo como resultado:

```

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

> summary.lm(regresion)$sigma
[1] 1496.954
> summary.lm(regresion)$sigma
[1] 1496.954
> summary.lm(regresion)$df
[1] 7 53933    7
>

```

R: Summarizing Linear Model I
Details

`print.summary.lm` tries to additionally give 'significance' information. Aliased coefficients are omitted. Correlations are printed to the screen by `summary(object)$correlation`.

Value

The function `summary.lm` creates a list containing information about the object, using the components:

- `residuals`: the weighted residuals specified in the call.
- `coefficients`: a $p \times 4$ matrix corresponding to the estimated coefficients.
- `aliased`: named logical vector indicating which coefficients are aliased.
- `sigma`: the square root of the residual sum of squares, where $R^2 = 1 - \frac{\text{RSS}}{\text{SS}_{\text{total}}}$.
- `df`: degrees of freedom.

Los números son similares a lo presentado anteriormente.

¿Cuál es el ángulo entre la Y y la Y estimada?

R=Recordando de clases, la R^2 es una medida del ángulo entre Y , \hat{Y} . En ese sentido podemos obtener el ángulo de la siguiente manera:

$$\begin{aligned}\theta &= \arccos R^2 \\ \theta &= \arccos(0.8592) \\ \theta &= 59.22^\circ\end{aligned}$$

Definan una función que calcule logverosimilitud de unos parámetros

$$L^*(\beta, \sigma) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y_t - x_t\beta)^t (y_t - x_t\beta)$$

donde: $Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \dots + \beta_p X_{tp} + e_t$

Utilicen la función optim de R para numéricamente el máximo de la función de verosimilitud .

Para la solución de este ejercicio es importante tomar en cuenta que al igual que el anterior se busca minimizar la función de error. Es decir $\sum(Y - \hat{Y})^2$. Para correr la función "optim" se inicializa las variables en sus respectivas medias, mismas que podemos ver en la siguiente tabla:

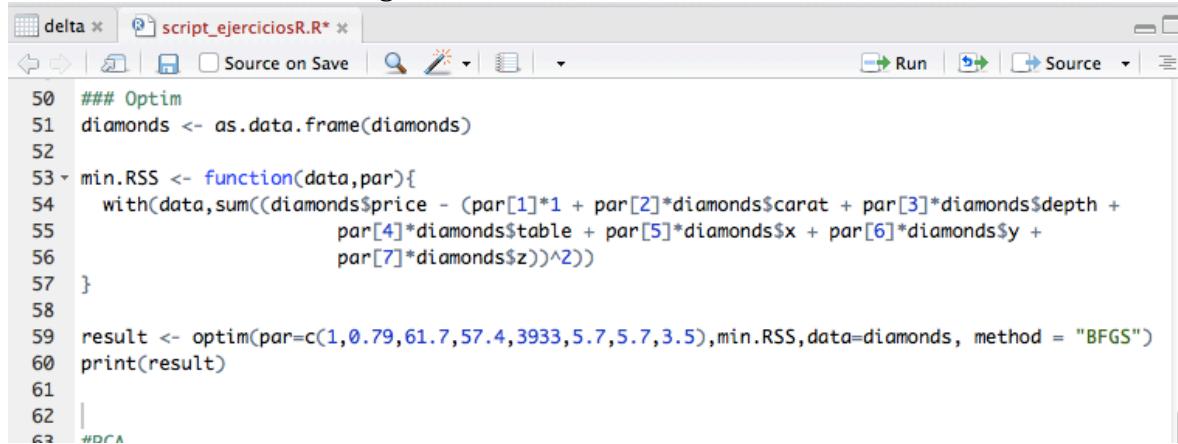
```

> summary(diamonds)
      carat          cut       color      clarity      depth       table
Min. :0.2000  Fair     : 1610  D: 6775  SI1    :13065  Min.  :43.00  Min.  :43.00
1st Qu.:0.4000 Good    : 4906  E: 9797  VS2    :12258  1st Qu.:61.00  1st Qu.:56.00
Median :0.7000 Very Good:12082 F: 9542  SI2    : 9194  Median :61.80  Median :57.00
Mean   :0.7979 Premium :13791  G:11292  VS1    : 8171  Mean   :61.75  Mean   :57.46
3rd Qu.:1.0400 Ideal   :21551  H: 8304  VVS2   : 5066  3rd Qu.:62.50  3rd Qu.:59.00
Max.  :5.0100                    I: 5422  VVS1   : 3655  Max.   :79.00  Max.   :95.00
                                J: 2808  (Other): 2531

      price          x         y         z
Min.  : 326  Min.   : 0.000  Min.   : 0.000  Min.   : 0.000
1st Qu.: 950  1st Qu.: 4.710  1st Qu.: 4.720  1st Qu.: 2.910
Median :2401  Median  : 5.700  Median  : 5.710  Median  : 3.530
Mean   :3933  Mean   : 5.731  Mean   : 5.735  Mean   : 3.539
3rd Qu.:5324  3rd Qu.: 6.540  3rd Qu.: 6.540  3rd Qu.: 4.040
Max.  :18823  Max.   :10.740  Max.   :58.900  Max.   :31.800

```

Después programamos un código en R para una función (min.RSS) que minimice los errores al cuadrado de la siguiente forma:



```

50  ### Optim
51 diamonds <- as.data.frame(diamonds)
52
53 min.RSS <- function(data,par){
54   with(data,sum((diamonds$price - (par[1]*1 + par[2]*diamonds$carat + par[3]*diamonds$depth +
55                 par[4]*diamonds$table + par[5]*diamonds$x + par[6]*diamonds$y +
56                 par[7]*diamonds$z))^2))
57 }
58
59 result <- optim(par=c(1,0.79,61.7,57.4,3933,5.7,5.7,3.5),min.RSS,data=diamonds, method = "BFGS")
60 print(result)
61
62
63 #DRA

```

Después en la variable “result” se almacena el resultado de la utilización del comando “optim” que nos permite estimar los parámetros de la regresión deseada. Como argumentos del comando “optim” incluimos la función “min.RSS” programada en el apartado anterior además de la inicialización de los parámetros buscados en las medias de cada variable respectivamente. También se utilizó el método “BFGS” para minimizar la función del error que funciona a través del método de gradiente. Los resultados de este procedimiento se observan a continuación:

```
> result <- optim(par=c(1,0.79,61.7,57.4,3933,5.7,5.7,3.5),min.RSS,data=diamonds, method = "BFGS")
> result <- optim(par=c(1,0.79,61.7,57.4,3933,5.7,5.7,3.5),min.RSS,data=diamonds, method = "BFGS")
> print(result)
$par
[1] 20849.31671 10686.30909 -203.15406 -102.44565 -1315.66785    66.32160   41.62771    3.50000

$value
[1] 1.20857e+11

$counts
function gradient
    78        12

$convergence
[1] 0

$message
NULL
```

Como se puede observar se llegan a los mismos parámetros estimados que por el método “lm”.