Curso Propedéutico en Ciencias de Datos ITAM

Roberto Acevedo Orozco

2017-07-17

Tarea 3: Regresión Lineal

Parte teórica. Esta parte del proyecto será sobre regresión lineal. Supongamos que quieren explicar una variable estadística ***Y*** (por ejemplo, altura) utilizando la información de p variables $\boldsymbol{X^1}, \dots, \boldsymbol{X^p}$ (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:

$$\boldsymbol{X} = [\boldsymbol{X^1}| \dots |\boldsymbol{X^p}]$$

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:

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

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

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

$$\|Y - X\beta\|^2 = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta$$

$$\frac{\partial(\|Y - X\beta\|^2)}{\partial\beta} = -2X^TY + 2X^TX\beta$$

Igualando a cero y despejando:

$$X^TX\beta = X^TY$$

Y si X^TX es matriz no singular y por lo tanto tiene inversa, tenemos:

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

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

El planteamiento nos da un ajuste lineal porque en este caso se busca reducir las distancias de los puntos a la línea, plano o hiperplano que más se acerca a las observaciones y la solución se deriva de una combinación lineal de ***X*** y ***Y***. Es posible utilizar el planteamiento para ajustar polinomios si se realizan las operaciones a las variables ***X*** previo a realizar el método, sin embargo, el resultado se debe de interpretar de manera cautelosa.

2) 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?

En clase se observó que:

$$\widehat{H} = X(X^T X)^{-1} X^T Y$$

Es la proyección de ***Y*** en el espacio generado por las variables de ***X*** ($span[\boldsymbol{X^1}| \dots |\boldsymbol{X^p}]$), por lo que la relación que existe entre la solución y la matriz de proyección es solamente multiplicar por la matriz ***X***. Su relación con el teorema de Pitágoras es que justamente al ser la menor distancia del entre el vector ***Y*** y el espacio generado por ***X***, se trata de la norma generada por el ángulo entre ellos.

3) ¿Qué logramos al agregar una columna de unos en la matriz ***X***? Es decir, definir mejor:

$$\boldsymbol{X} = [\boldsymbol{1_n}|\boldsymbol{X^1}| \dots |\boldsymbol{X^p}]$$

Con:

$$\boldsymbol{1_n} = [\boldsymbol{1}, \boldsymbol{1}, \dots., \boldsymbol{1}]^{\boldsymbol{T}}$$

Al incorporar una columna de unos en la matriz *X,* se agrega un factor constante a la solución del vector $\hat{\beta}$, es decir, la solución ya no estará centrada en cero para cualquier X.

4) Plantear el problema de regresión ahora como un problema de estadística

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

Donde los errores son no correlacionados con distribución:

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

Para este problema es necesario encontrar el error esperado de todas las Y_i, sin embargo, utilizando la LGN se puede deducir que:

$$E\left[\sum_{i=1}^{n} \varepsilon_i\right] = 0$$

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

$$Y = X\beta + \varepsilon$$

Con

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

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

$$L(\beta, \sigma^2; Y, X) = \prod_{i=1}^{p} (2\pi\sigma^2)^{-1/2} e^{\frac{-(y_i - x_i\beta)^2}{2\sigma^2}}$$

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

En términos matriciales:

$$L(\beta, \sigma^2; Y, X) = (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}(Y - X\beta)^2}$$

En dónde los términos $X\beta$ y $\sigma^2 I_n$ corresponden a la media y varianza de una distribución normal, por lo que:

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

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

$$L(\beta, \sigma^2; Y, X) = (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}(Y - X\beta)^2}$$

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

$$\frac{\partial \log\big(L(\beta, \sigma^2; Y, X)\big)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^T (y_i - x_i\beta)$$

$$\frac{\partial \log\big(L(\beta, \sigma^2; Y, X)\big)}{\partial \beta} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta\right)$$

Igualando a cero y despejando queda:

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

Por lo tanto:

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

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

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

El Teorema de Gauss-Markov establece que en un modelo lineal general (MLG) en el que se cumplan los siguientes supuestos:

a. Correcta especificación: el MLG ha de ser una combinación lineal de los parámetros β y no necesariamente de las variables: Y=Xβ+u
b. 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')
c. Esperanza condicionada de los errores nula: $E(u_i|X'_i) = 0$
d. Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N
e. 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, aunque hay puntos adicionales si usan alguna base original interesante.

Cargar la base diamonds 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 3.3.3

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
```

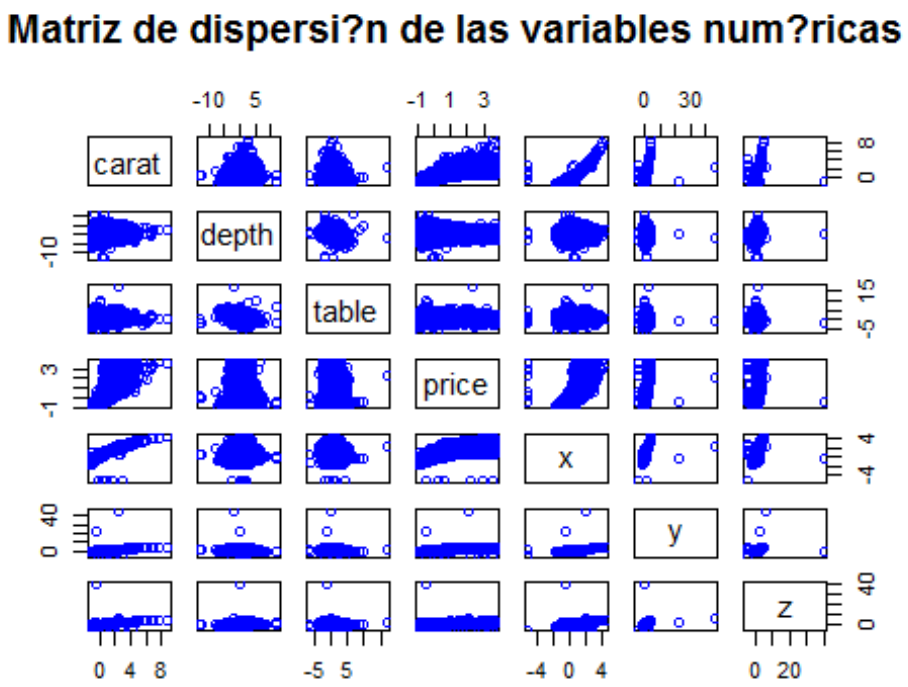
Posteriormente deben hacer una regresión lineal. Su objetivo es explicar la variable price usando las demás variables. Noten que algunas variables no son numéricas, por lo que no pueden incluirse en un análisis crudo de regresión lineal. Para este proyecto NO es necesario saber transformar las variables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

Creamos un nuevo data_frame que solo contiene los datos que vamos a utilizar, es decir, 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 variable
s numéricas")
```



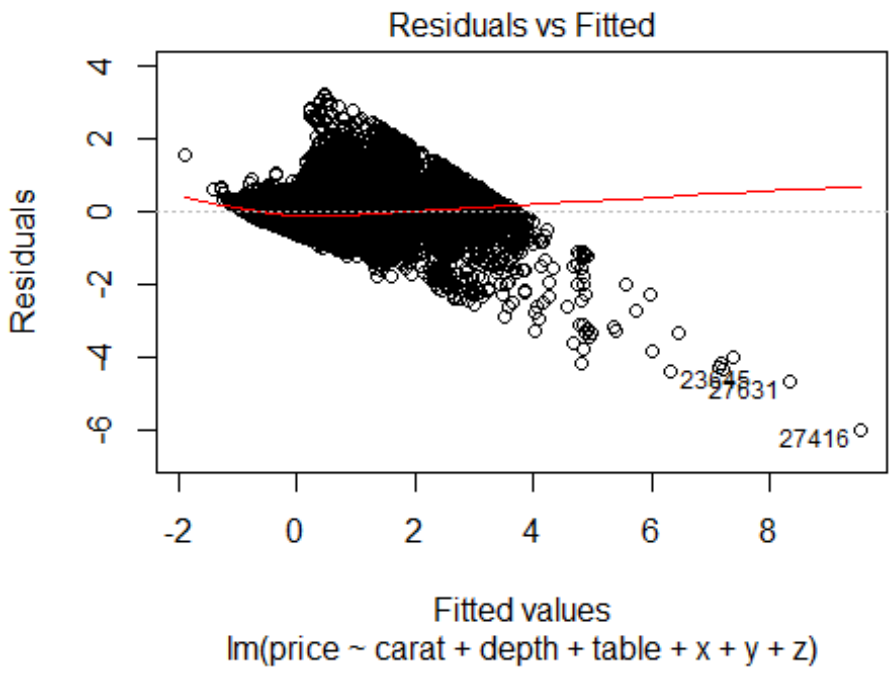
Gracias a la matriz de las variables en pares podemos discernir de forma rápida una relación positiva entre el precio y la variable "carat", aunque la varianza aumenta conforme aumenta esta variable. También podemos observar que no hay mucha relación entre el precio y las variables "y"(ancho en mm) y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

```
#Realizamos la regresión lineal usando todas las variables numéricas
modelo1 = lm(price ~ carat + depth + table + x + y + z, data = diamantes0
)
summary(modelo1)

##
## Call:
## lm(formula = price ~ carat + depth + table + x + y + z, data = diamant
es0)
##
```

```
## 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)
```



8) ¿Qué tan bueno fue el ajuste?

Del "summary" de la regresión podemos observar inicialmente la información de los residuales, que son todas las diferencias entre el precio estimado y el precio real; una media lo más cercana a cero es lo que queramos ver aquí, pues indicaría que el modelo es

muy certero. En la gráfica vemos los valores predichos contra los residuales, en un modelo perfecto, la media de los residuales sería 0 por lo que la línea roja iría 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. 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, para todos las variables queremos que este sea lo más pequeño posible.

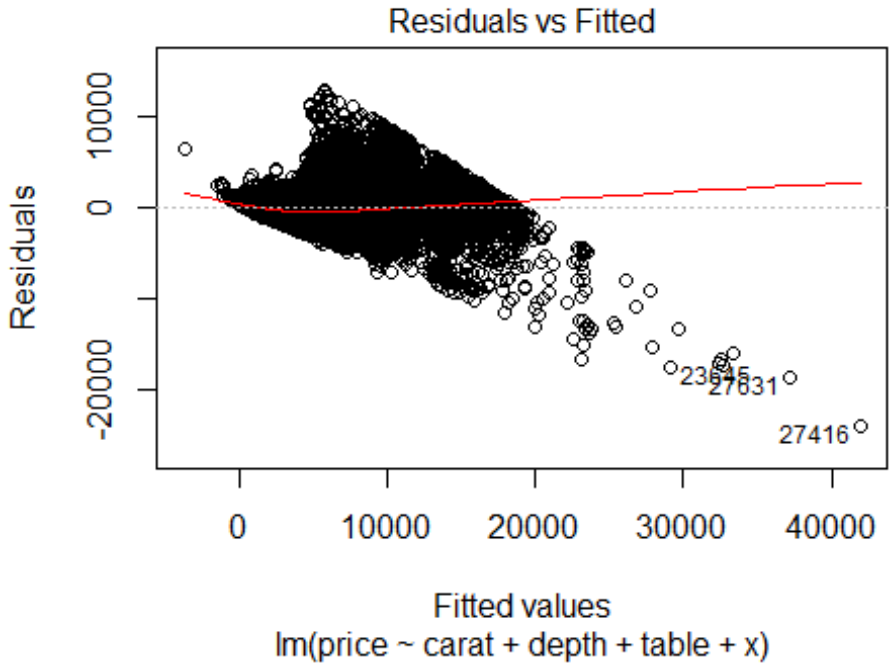
9) ¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cuál fue el valor de que ajustó su modelo y qué relación tiene con la calidad del ajuste?

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 podríamos 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.

```
# 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)
```



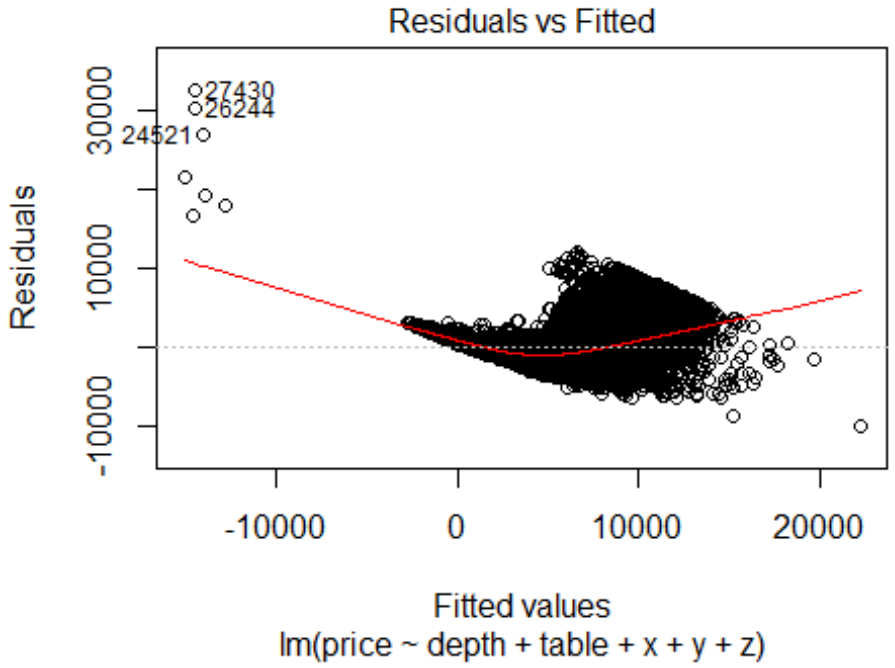
Aquí vemos, como era esperado, que deshacernos de las dos variables que no eran relevantes prácticamente no modifica valores como la media de los residuales o la R cuadrada.

```
# Por curiosidad, ahora realizamos otra regresión lineal, pero ahora deja
mos fuera la variable "carat", que es una variable relevante para el mode
lo
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, disminuye el valor de R cuadrada y aumenta la media de los residuales (cuyo efecto puede verse en la recta de la gráfica).

¿Cuál es el ángulo entre $\boldsymbol{Y}$ y $\hat{Y}$?. Hint: usen la y el arcocoseno

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

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