Tarea 3: Regresión Lineal

Miguel Castañeda Santiago

17 de julio de 2017

Supongamos que quieren explicar una variable estadística Y (por ejemplo altura) utilizando la información de p $X^1, X^2, \ldots X^p$ variables (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

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

de tamaño nxp donde cada columna es una variable y cada fila uno de los individuos de la muestra. Contestar

- Plantear el problema de regresión como un problema de mínimos cuadrados, encontrar el vector

$$\hat{\beta} = [\hat{\beta}_1, \ldots . \hat{\beta}_p]^p$$

- que resuelva

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

Obteniendo el gradiente de la funciòn a minimizar tenemos que

$$\nabla \|Y - X\beta\|^2 = Y^T Y + \beta^T X^T X\beta - 2\beta^T X^T Y$$

Derivando con respecto a β e igualando a 0 tenemos que la solución es:

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

Si existe $(X^T X)^{-1}$

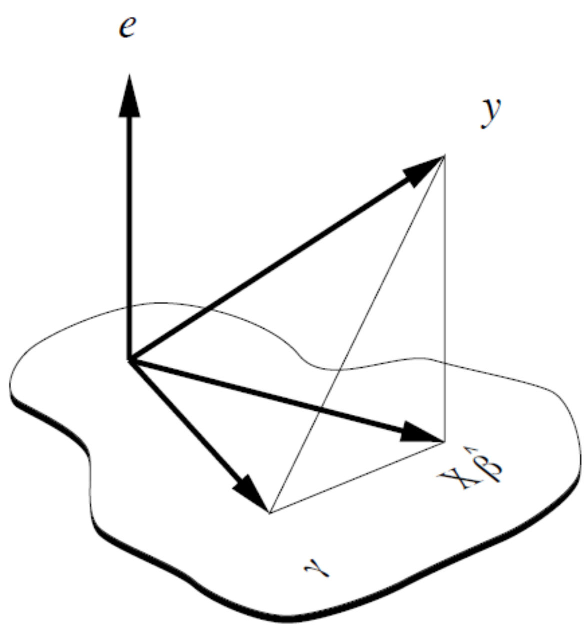
Dado que el problema se puede plantear como una combinación de los términos

$$Y = \beta_1 X_1^1 + \beta_2 X_2^2 \ldots + \beta_p X_i^P$$

El ajuste que resulta es del tipo lineal.

Como el el modelo el lineal en lo parámetros se puede definir un cambio de variable haciendo y se puede utilizar para ajustar a polinomios por ejemplo $y = x^2$

Su relación con el teorema de pitágoras es tiene una interpretación geométrica, siendo



Fuente: http://www.le.ac.uk/users/dsgp1/COURSES/THIRDMET/MYLECTURES/2MULTIREG.pdf

De la imagen se puede apreciar el triangulo formado por γ, $X\hat{\beta}$ y Y donde la distancia de $Y - \gamma$ no puede ser menor a la distancia $ Y - X$

Planteando el problema de regresión ahora como un problema de estadística se tiene que:

$$Y_i = \beta_0 + \beta_1 X_i^1 + \ldots + X_i \beta_i + \varepsilon_i$$

Escribiendo el el problema como

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

Escribiendo la verosimilitud como

$$L(\beta, \sigma^2) = f(Y|\beta, \sigma^2, X)$$

La solución de máxima verosimilitud se determina como

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

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

Aplicando el logaritmo se tiene que

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

Derivando con respecto a β

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

Igualando a cero

$$\sum_{i=1}^{N} x_i^T\,y_i - \sum_{i=1}^{N} x_i^T\,x_i\beta = 0$$

Despejando

$$\beta$$

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

Que se puede escribir tambien como

$$\beta = \frac{\sum_{i=1}^{N} x_i^T\,y_i}{(\sum_{i=1}^{N} x_i^T\,x_i)} = (X^TX)^{-1}X^Ty$$

Por lo que la solución máxima es la misma que la del problema de mínimos cuadrados

Teorema de Gauss Markov

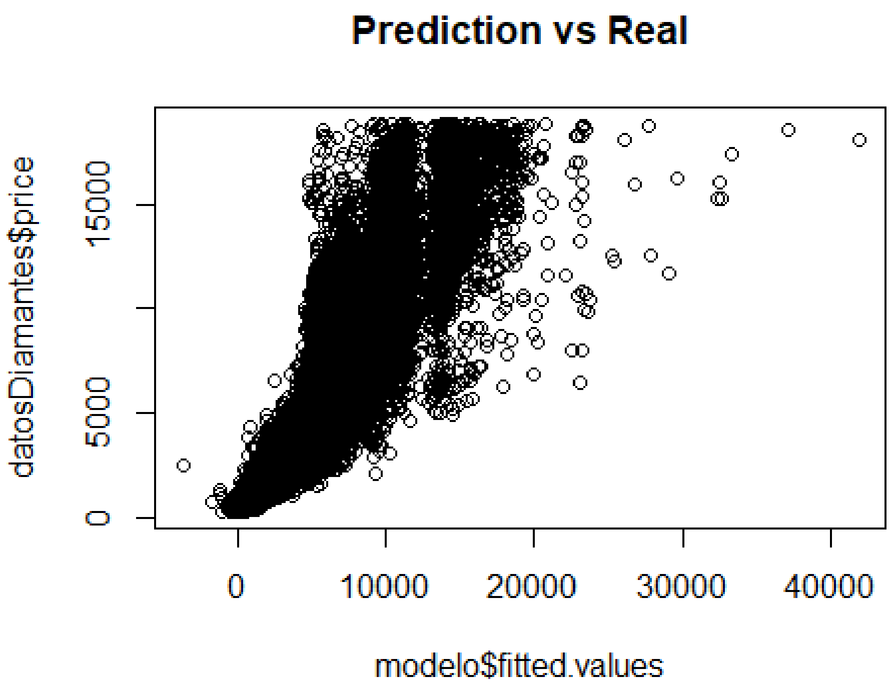
En estadística, el Teorema de Gauss-Márkov, formulado por Carl Friedrich Gauss y Andréi Márkov, establece que en un modelo lineal general (MLG) en el que se establezcan los siguientes supuestos:

Correcta especificación: el MLG ha de ser una combinación lineal de los parámetros ($ $) 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},,,,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 las perturbaciones nula: $ E(u_{i}|X'_{i})=0}$ *Correcta identificación: la matriz de regresoras (X) ha de tener rango completo: rg(X)=K<=N* Homocedasticidad: $Var(U|X)=^{2}I} $

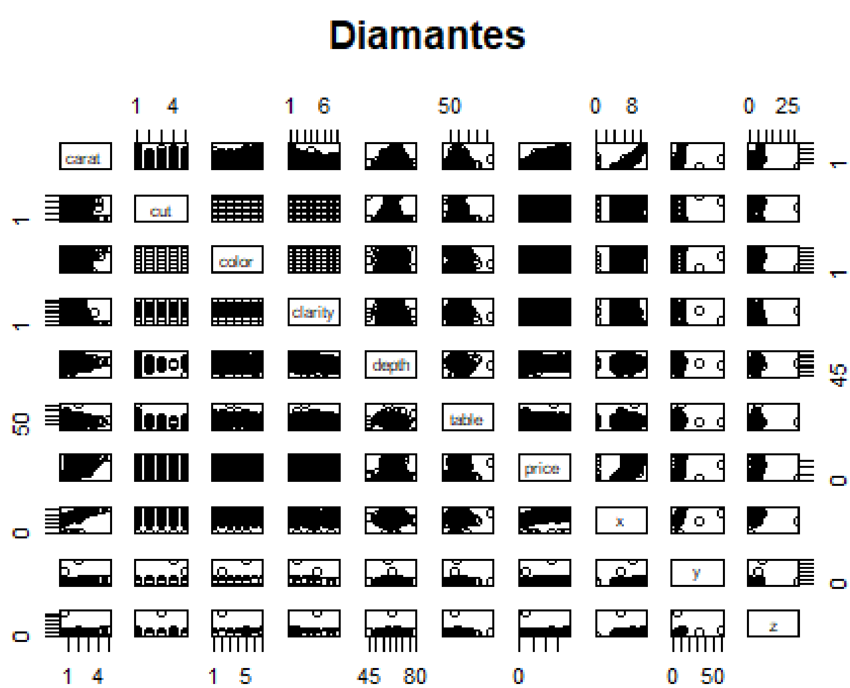
el estimador mínimo cuadrático ordinario (MCO) de B es el estimador lineal e insesgado óptimo (ELIO o BLUE: best linear unbiased estimator), es decir, el estimador MCO es el estimador eficiente dentro de la clase de estimadores lineales e insesgados.

Parte Aplicada

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.1

datosDiamantes <- diamonds
head(datosDiamantes)

## # 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

modelo <- lm (price ~ carat + depth + table + x + y + z, data=
datosDiamantes)
summary(modelo)

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

plot(modelo$fitted.values,datosDiamantes$price,main="Prediction vs Real")
```




```r
pairs(datosDiamantes,main="Diamantes")
```




```r
summary(modelo)
```

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

*¿Qué tan bueno fue el ajuste?, dada la informaciòn arrojada por la función summary vemos que el ajuste tienen un valor R-cuadrado de 0.8592 el cual podemos considerar un bueno ajute.

*¿Qué medida puede ayudarnos a saber la calidad del ajuste? el valor de r^2 ¿Cuál fue el valor de que ajustó sumodelo y que relación tiene con la calidad del ajuste? el valor obtenido por el modelo fue 0.8592

*¿Cuál es el ángulo entre y ?. Hint: usen la y el arcocoseno.

```
angulo <- acos(sqrt(0.8592))*180/pi
angulo

## [1] 22.03873
```