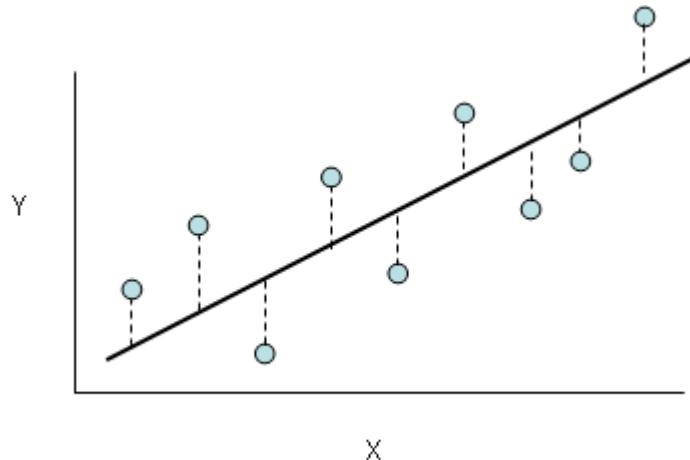


Instituto Tecnológico Autónomo de México  
Propedéutico Ciencia de Datos  
Tarea 3

Federico Riveroll

## Regresión lineal como optimización de mínimos cuadrados.

Un modelo de regresión lineal implica trazar una línea en un plano con puntos en la cual la suma de las distancias cuadradas de la línea a cada punto sea el mínimo valor posible.



Definamos este número como  $|Y-Xb|^2$ , donde Y son los valores reales y Xb son los valores que nuestra línea nos definiría. Para obtener la mejor solución posible necesitamos utilizar la siguiente ecuación:

$$\text{Argmin } \{|Y-Xb|^2\} = (X(t) * X)^{-1} * X(t) * y$$

Dicha solución para el ejemplo otorgado nos engregaría un valor equivalente a la pendiente de la línea para minimizar errores. Pero aún tenemos el problema de que tiene que pasar por el origen. Para solucionar esta situación fijamos una columna al principio de nuestra matriz X con valores constantes 1.

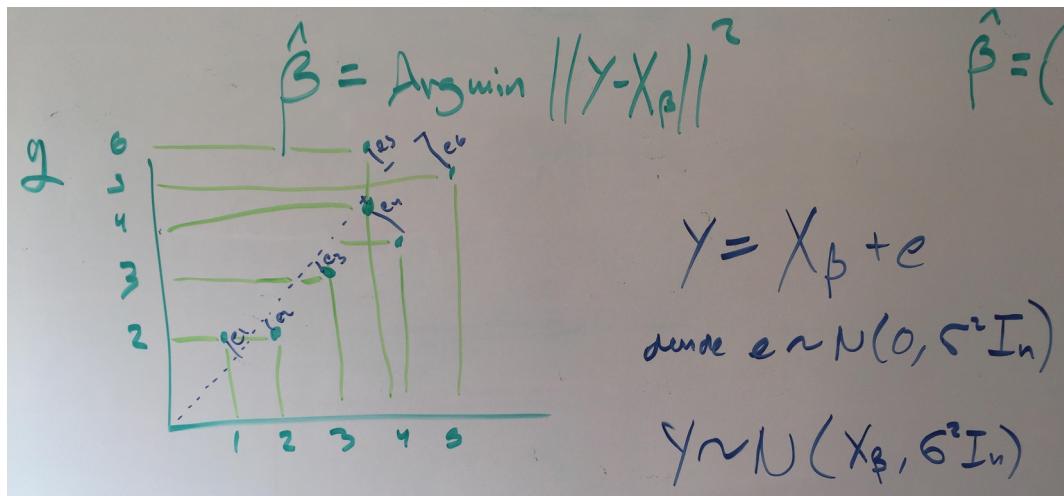
De esta forma la solución nos entregará la pendiente y el valor de intersección con el eje Y, y así nuestra solución puede desplazarse del origen y tener un menor error.

## Regresión como problema de estadística.

Una función o muestra puede ser descrita como una combinación lineal de predicciones escalares ( $B_i X_i$ ) más el error que estas conllevan;

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n + e, \text{ o bien,}$$

$$Y_i = B_i X_i + e_i$$



Lo cual es exactamente igual a una regresión lineal salvo que ahora se suman los errores, y se asume que dichos errores se distribuyen Normalmente (por teorema Gauss-Markov) con media 0, lo que implica que el promedio de los errores debe tender a la función original.

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

$$\text{EMV}(\mu) \rightarrow \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

para  $X_1, X_2, X_3, \dots, X_n$

① Multiplicar todos:  $\prod_{i=0}^n \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right)$

Necesitamos minimizar para maximizar ecuación

$\sum (x_i - \mu)^2$  Es una diferencia cuadrática

Por lo tanto,  $Y$  misma se distribuye normal, con media  $\mathbf{XB}$  (obviamente) y la misma varianza de  $e$ . Si sacamos la verosimilitud como se muestra en la imagen anterior, llegamos de nuevo al problema de minimizar los errores cuadráticos para maximizar la verosimilitud.

## Trabajo Práctico

Para la tabla diamonds se utilizarán las variables numéricas para explicar la variable de precios “price”:

```
lmMultiple <- lm(price ~ carat + depth + table + x + y + z, data=diamantes)
```

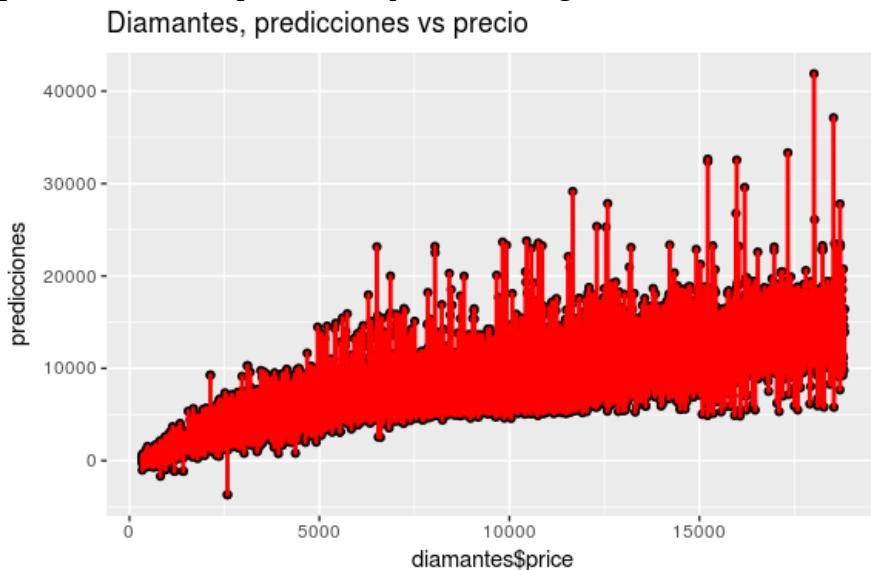
Los coeficientes de la regresión arrojados por R fueron los siguientes:

Coefficients:

(Intercept)	carat	depth	table	x	y	z
20849.32	10686.31	-203.15	-102.45	-1315.67	66.32	41.63

Para probar la calidad del ajuste podemos multiplicar cada coeficiente por sus respectivos valores y sumarlos y así obtener el “precio predecido”, y luego calcular el error cuadrático con respecto al verdadero valor. Finalmente, podemos sumar los errores cuadráticos y así ver el error cuadrático general y tener algún parámetro para medir la calidad del modelo.

La gráfica de las predicciones vs el precio real quedó de la siguiente manera:



Para la realización se utilizó la librería GGPlot. Podemos ver que hay una concordancia moderada entre el precio y la predicción, pero hay muchos valores incoherentes, negativos o muy altos. El **error cuadrático medio** fué de 1.20857e+11.