

# Taller 1

Julián Gutierrez, Rafael Cano, Mauricio Perea

Septiembre 2020

## 1 Theory Exercises: Review Stats and OLS

### 1.1

$$P(X|\epsilon_1, \epsilon_2) = \frac{P(\epsilon_1, \epsilon_2|X)P(x)}{P(\epsilon_1, \epsilon_2)}$$

Si  $\epsilon_1, \epsilon_2$  son independientes, el problema se puede reescribir como:

$$\frac{P(\epsilon_1|X)P(\epsilon_2|X)P(x)}{P(\epsilon_1, \epsilon_2)}$$

(a)

$$ii. P(\beta_1, \beta_2), P(X), P(\beta_1, \beta_2|X)$$

(b)

$$i. P(\beta_1, \beta_2), P(X), P(\beta_1|H), P(\beta_2|X)$$

### 1.2

(a)

$$E(\epsilon_i) = E(\epsilon_i|X) = E(0) = 0$$

$$E(X_i\epsilon_i) = E(X_i)E(\epsilon_i) = 0$$

(b)

$$E(X_i^T \epsilon_i) = 0$$

$$E(X_i^T (Y - X\beta)) = 0$$

$$E(X_i^T Y) - E(X_i^T X\beta) = 0$$

$$E(X_i^T Y) - E(X_i^T X)\beta = 0$$

$$- E(X_i^T X)\beta = -E(X_i^T Y)$$

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

$$\hat{\beta} = [N^{-1} \sum (X^T X)]^{-1} [N^{-1} \sum (X^T Y)] \quad (1)$$

(c) Bajo el supuesto que

$$X^T X$$

es no singular y que

$$plim[(N^{-1} \sum^N (X^T X))^{-1}] = A^{-1}$$

donde

$$A = E(X^T X)$$

Entonces, bajo la anterior expresión (1) puede escribirse como

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

### 1.3

(a) Dado que

$$R^2 = 1 - \frac{\sum^N (\hat{e}_i)^2}{\sum^N (Y_i - \bar{Y})^2}$$

Y los estimadores de MCO se obtienen mediante la minimización de la suma de los residuos cuadrados

$$\min \sum^N (\hat{e}_i)^2$$

Entonces, los estimadores de MCO maximizan el R-cuadrado, pues al observar la fórmula del R-cuadrado se puede apreciar que la suma de los residuos cuadrados está restando a el 1, por lo cual al maximizar la suma de los residuos cuadrados la expresión completa se maximiza.

(b) Teneindo en cuenta que

$$R^2 = 1 - \frac{\sum^N (\hat{e}_i)^2}{\sum^N (Y_i - \bar{Y})^2}$$

De (donde los tres términos son no-negativos)

$$\sum^N (Y_i - \bar{Y})^2 = \sum^N (\hat{e}_i)^2 + \sum^N (\hat{Y}_i - \bar{Y})^2$$

Se puede concluir que

$$\sum^N (Y_i - \bar{Y})^2 \geq \sum^N (\hat{e}_i)^2$$

Entonces

$$\frac{\sum^N (\hat{e}_i)^2}{\sum^N (Y_i - \bar{Y})^2} \leq 1$$

$$\frac{\sum^N (\hat{e}_i)^2}{\sum^N (Y_i - \bar{Y})^2} \geq 0$$

Lo cual implica que

$$0 \leq R^2 \leq 1$$

(c) Teniendo en cuenta el modelo

$$Y = \alpha + \beta X_i + U_i$$

Donde

$$\beta = \frac{Cov(X, Y)}{Var(x)}$$

Y teniendo en cuenta que

$$R^2 = \frac{SCR}{SCT}$$

$$r^2 = \frac{Cov(X, Y)^2}{Var(x)Var(Y)}$$

Si solucionamos

$$\begin{aligned} SCE &= \sum^N (Y_i - \hat{Y}_i)^2 \\ &= \sum^N (Y_i - (\alpha + \beta X_i))^2 \\ &= \sum^N (Y_i - (\alpha + \beta X_i))^2 \\ &= \sum^N ((Y_i - \bar{Y}) - \beta(X_i - \bar{X}))^2 \\ &= \sum^N (Y_i - \bar{Y})^2 + \beta^2 \sum^N (X_i - \bar{X})^2 - 2\beta^2 \sum^N (Y_i - \bar{Y})(X_i - \bar{X}) \\ &= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X, Y) \\ &= Var(Y) + \left(\frac{Cov(X, Y)}{Var(x)}\right)^2 Var(X) - 2\left(\frac{Cov(X, Y)}{Var(x)}\right) Cov(X, Y) \\ &= Var(Y) - \frac{(Cov(X, Y))^2}{Var(x)} \end{aligned}$$

Y que sabiendo que

$$\begin{aligned}
STC &= SCE + SCR \\
SCR &= STC - SCE \\
SCR &= \sum_{i=1}^N (Y_i - \bar{Y})^2 - \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \\
SCR &= Var(Y) - [Var(Y) - \frac{(Cov(X, Y))^2}{Var(X)}] \\
SCR &= \frac{(Cov(X, Y))^2}{Var(X)}
\end{aligned}$$

Entonces

$$\begin{aligned}
R^2 &= \frac{SCR}{SCT} \\
&= \frac{(Cov(X, Y))^2}{Var(x)} \frac{1}{Var(Y)} \\
&= \frac{Cov(X, Y)^2}{Var(x)Var(Y)} \\
&= r^2
\end{aligned}$$

En conclusión

$$R^2 = r^2$$

## 1.4

(a)

Según el teorema FWL, el problema de estimación OLS puede ser escrito como

$$\begin{aligned}
M_i Y &= M_i X_2 \beta_2 + residuals \\
\hat{\beta}_2 &= (X' M_i X)^{-1} X' M_i Y
\end{aligned}$$

De forma análoga para  $\hat{\beta}_1$ :

$$\hat{\beta}_1 = (i' M_X i)^{-1} i' M_X Y$$

A partir del teorema de FWL ( $M_X = I - P_X, P_X X = X$ ) podemos reescribir  $\hat{\beta}_1$  como:

$$\begin{aligned}
(i'(I - P_X)i)\hat{\beta}_1 &= i'(I - P_X)Y \\
\rightarrow (N - i'P_X i)\hat{\beta}_1 &= i'Y - i'P_X Y \\
\rightarrow N\hat{\beta}_1 + (i'P_X Y - i'P_X i\hat{\beta}_1) &= i'Y \\
\rightarrow N\hat{\beta}_1 + i'P_X(Y - i\hat{\beta}_1) &= i'Y \\
\rightarrow N\hat{\beta}_1 + i'P_X X\hat{\beta}_2 &= i'Y
\end{aligned}$$

De este modo, tenemos:

$$\begin{aligned} N\hat{\beta}_1 + i'X\hat{\beta}_2 &= (X'M_iX)X'M_iY \\ (X'M_iX)\hat{\beta}_2 &= X'M_iY \\ \begin{pmatrix} N & i'X \\ 0 & X'M_iX \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \begin{pmatrix} i'Y \\ X'M_iY \end{pmatrix} \end{aligned}$$

## 1.5

(a)

$$\begin{aligned} R_K^2 &= \frac{\sum^N (\hat{Y}_i - \bar{Y})^2}{\sum^N (Y_i - \bar{Y})^2} \\ &= \frac{\sum^N \sum^K (\hat{\beta}_k X_{ik} - \hat{\beta}_k \bar{X}_k)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \\ &= \frac{\sum^N \sum^K \hat{\beta}_k (X_{ik} - \bar{X}_k)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \\ &= \sum^K \hat{\beta}_k \frac{\sum^N (X_{ik} - \bar{X}_k)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \end{aligned}$$

(b)

$$\begin{aligned} R_K^2 &= \sum^{K-1} \hat{\beta}_k \frac{\sum^N (X_{ik} - \bar{X}_k)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} + \hat{\beta}_K \frac{\sum^N (X_{iK} - \bar{X}_K)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \\ &= R_{K-1}^2 + \hat{\beta}_K \frac{\sum^N (X_{iK} - \bar{X}_K)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \end{aligned}$$

Sabemos que  $R_K^2 \geq R_{K-1}^2$  ya que  $\hat{\beta}_K \frac{\sum^N (X_{iK} - \bar{X}_K)(\hat{Y}_i - \bar{Y})}{\sum^N (Y_i - \bar{Y})^2} \geq 0$ .

## 1.6

$$\begin{aligned} &E(\hat{Y}_{test} - Y_{test}) \\ &= E(\hat{Y}_{test}) - E(Y_{test}) \\ &= E(\hat{\alpha}_N + \hat{\beta}_N X_{test}) - E(\alpha + \beta X_{test} + \epsilon) \\ &= E(\hat{\alpha}_N) + E(\hat{\beta}_N X_{test}) - [E(\alpha) + E(\beta X_{test}) + E(\epsilon)] \\ &= \alpha + E(\hat{\beta}_N)X_{test} - (\alpha + \beta_N X_{test} + 0) \\ &= \alpha + \beta X_{test} - (\alpha + \beta_N X_{test}) \\ &= 0 \end{aligned}$$

$$\begin{aligned}
& E(Y_{test} - \hat{Y}_{test})^2 \\
&= E[\alpha + \beta X_{test} - (\hat{\alpha}_N + \hat{\beta}_N X_{test})]^2 + Var(\epsilon) \\
&= E[(\alpha - \hat{\alpha}_N) + (\beta - \hat{\beta}_N) X_{test}]^2 + Var(\epsilon) \\
&= E(\alpha - \hat{\alpha}_N)^2 + E[(\beta - \hat{\beta}_N) X_{test}]^2 + Var(\epsilon) \\
&= E(\alpha - \hat{\alpha}_N)^2 + E(\beta - \hat{\beta}_N)^2 (X_{test} - \bar{X})^2 + Var(\epsilon) \\
&= \frac{\sigma}{N} + \frac{\sigma}{\sum^N (X - \bar{X})} (X_{test} - \bar{X})^2 + \sigma \\
&= \sigma \left( \frac{1}{N} + \frac{(X_{test} - \bar{X})^2}{\sum^N (X - \bar{X})} + 1 \right) \\
&= \sigma \left( 1 + \frac{1}{N} + \frac{(X_{test} - \bar{X})^2}{\sum^N (X - \bar{X})} \right)
\end{aligned}$$

## 1.7

Considerare el estimador "ridge"

$$\hat{\beta}^\gamma = \frac{\hat{\beta}^{ols}}{1 + \gamma} = (X^T X + \gamma I)^{-1} X^T Y$$

(a) Sesgo del estimador "ridge"

$$\begin{aligned}
& E(\hat{\beta}^\gamma - \beta) \\
&= E(\hat{\beta}^\gamma) - E(\beta) \\
&= E[(X^T X + \gamma I)^{-1} X^T Y] - \beta \\
&= E[(X^T X + \gamma I)^{-1} X^T (X\beta - \epsilon)] - \beta \\
&= E[(X^T X + \gamma I)^{-1} X^T X\beta] + E[(X^T X + \gamma I)^{-1} X^T \epsilon] - \beta \\
&= (X^T X + \gamma I)^{-1} X^T X\beta + (X^T X + \gamma I)^{-1} X^T E[\epsilon] - \beta \\
&= (X^T X + \gamma I)^{-1} X^T X\beta - \beta \\
&= [(X^T X + \gamma I)^{-1} X^T X - I]\beta
\end{aligned}$$

Varianza del estimador "ridge"

$$\begin{aligned}
& Var(\gamma) \\
&= E[(\hat{\beta}^\gamma - \beta^\gamma)(\hat{\beta}^\gamma - \beta^\gamma)^T] \\
&= E[((X^T X + \gamma I)^{-1} X^T \hat{Y} - (X^T X + \gamma I)^{-1} X^T Y)((X^T X + \gamma I)^{-1} X^T \hat{Y} - (X^T X + \gamma I)^{-1} X^T Y)^T] \\
&= E[((X^T X + \gamma I)^{-1} X^T (\hat{Y} - Y))((X^T X + \gamma I)^{-1} X^T (\hat{Y} - Y))^T] \\
&= E[((X^T X + \gamma I)^{-1} X^T \epsilon)((X^T X + \gamma I)^{-1} X^T \epsilon)^T] \\
&= E[(X^T X + \gamma I)^{-1} X^T \epsilon \epsilon^T X (X^T X + \gamma I)^{-1}] \\
&= (X^T X + \gamma I)^{-1} X^T E[\epsilon \epsilon^T] X (X^T X + \gamma I)^{-1} \\
&= (X^T X + \gamma I)^{-1} X^T \sigma^2 X (X^T X + \gamma I)^{-1} \\
&= \sigma^2 (X^T X + \gamma I)^{-1} X^T X (X^T X + \gamma I)^{-1} \\
&= \sigma^2 (X^T X + \gamma I)^{-1} X^T X (X^T X + \gamma I)^{-1}
\end{aligned}$$

El sesgo del estimador "ridge" es:

$$E(\hat{\beta}^\gamma - \beta) = [(X^T X + \gamma I)^{-1} X^T X - I] \beta$$

Y el del estimador MCO es:

$$E(\hat{\beta}^- - \beta) = 0$$

Por lo cuál, el estimador "ridge" tiene un sesgo mayor al estimador MCO, el cual es insesgado. Lo anterior, pues gamma es un escalar positivo.

La varianza del estimador "ridge" es:

$$Var(\hat{\beta}^\gamma) = \sigma^2 (X^T X + \gamma I)^{-1} X^T X (X^T X + \gamma I)^{-1}$$

Y la varianza MCO es:

$$Var(\hat{\beta}^- - \beta) = \sigma^2 (X^T X)^{-1}$$

Por lo cuál, la varianza "ridge" es menor a la de MCO. Lo anterior, pues gamma es un escalar positivo.

(b)  
Si

$$W_\gamma = (X^T X + \gamma I)^{-1} X^T X$$

Entonces

$$\hat{\beta}^\gamma = (X^T X + \gamma I)^{-1} X^T Y = W_\gamma \hat{\beta}^{OLS}$$

$MSE(\hat{\beta}^\gamma)$

$$\begin{aligned} &= E[(\hat{\beta}^\gamma - \beta)^T (\hat{\beta}^\gamma - \beta)] \\ &= E[(W_\gamma \hat{\beta} - \beta)^T (W_\gamma \hat{\beta} - \beta)] \\ &= E(\hat{\beta}^T W_\gamma^T W_\gamma \hat{\beta}) - E(\beta^T W_\gamma^T \hat{\beta}) - E(\hat{\beta}^T W_\gamma^T \beta) + E(\beta^T \beta) \\ &= E(\hat{\beta}^T W_\gamma^T W_\gamma \hat{\beta}) - E(\beta^T W_\gamma^T W_\gamma \hat{\beta}) - E(\hat{\beta}^T W_\gamma^T W_\gamma \beta) + E(\beta^T W_\gamma^T W_\gamma \beta) - E(\beta^T W_\gamma^T W_\gamma \beta) \\ &\quad + E(\beta^T W_\gamma^T W_\gamma \hat{\beta}) + E(\hat{\beta}^T W_\gamma^T W_\gamma \beta) - E(\beta^T W_\gamma^T \hat{\beta}) - E(\hat{\beta}^T W_\gamma^T \beta) + E(\beta^T \beta) \\ &= E((\hat{\beta} - \beta)^T W_\gamma^T W_\gamma (\hat{\beta} - \beta)) - \beta^T W_\gamma^T W_\gamma \beta - \beta^T W_\gamma^T W_\gamma \beta + \beta^T W_\gamma^T W_\gamma \beta + \beta^T W_\gamma^T W_\gamma \beta \\ &\quad + \beta^T W_\gamma^T W_\gamma \beta - \beta^T W_\gamma \beta - \beta^T W_\gamma^T \beta + \beta^T \beta \\ &= E((\hat{\beta} - \beta)^T W_\gamma^T W_\gamma (\hat{\beta} - \beta)) + \beta^T W_\gamma^T W_\gamma \beta - \beta^T W_\gamma \beta - \beta^T W_\gamma^T \beta + \beta^T \beta \\ &= E((\hat{\beta} - \beta)^T W_\gamma^T W_\gamma (\hat{\beta} - \beta)) + \beta^T (W_\gamma - I)(W_\gamma - I) \beta \\ &= tr[Var(\hat{\beta}^\gamma)] + Bias(\hat{\beta}^\gamma)^T Bias(\hat{\beta}^\gamma) \end{aligned}$$

Reemplazando la varianza y el sesgo que ya encontramos nos queda:

$$\begin{aligned} MSE(\hat{\beta}^\gamma) &= \sigma^2 tr((X^T X + \gamma I)^{-1} X^T X (X^T X + \gamma I)^{-1}) \\ &\quad + [(X^T X + \gamma I)^{-1} X^T X - I]^T [(X^T X + \gamma I)^{-1} X^T X - I] \end{aligned}$$

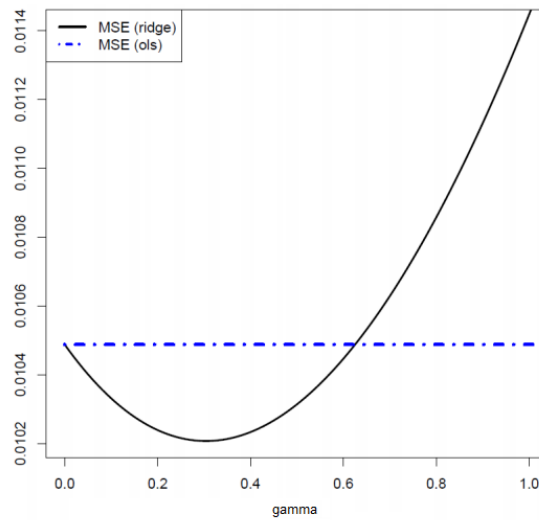
Así pues, teniendo en cuenta que el MSE de MCO es:

$$MSE(\hat{\beta}^{OLS}) = \sigma^2(X^T X)^{-1}$$

Se puede concluir que el MSE de la estimación "ridge" puede ser mayor o menor al MSE de MCO dependiendo de que valor tome gamma.

En la siguiente gráfica se puede apreciar que existe un gamma para el cual:

$$MSE(\hat{\beta}^\gamma) < MSE(\hat{\beta}^{OLS})$$



(c)

Con respecto a la práctica econométrica clásica de preferir estrictamente estimadores insesgados se puede sacar la conclusión de dicha preferencia estricta es problemática e incluso detrimental si con la estimación se pretende generar predicciones acertadas y no simplemente relaciones causales. Es así, que si se busca crear un modelo predictivo puede ser mejor sacrificar insesgadez a cambio de una varianza menor del estimador, así pues la preferencia estricta por estimadores insesgados pierde fuerza cuando obtener un estimador insesgado no es fundamental para las necesidades de la estimación y para el propósito para el cual será utilizada.



## **2 Empirical Problems**

### **2.1 Predicting House Prices in Colombia**

#### **2.1.1**

#### **2.1.2**

Observando la base de datos, y teniendo en cuenta el objetivo de construir modelos que intenten explicar el precio de apartamentos, se decidió considerar las siguientes variables como posibles variables explicativas: Número de habitaciones, Número de baños, longitud, Latitud, superficie total, superficie cubierta y tipo de propiedad. Por otro lado, la base de datos presenta un número muy grande de observaciones vacías. Para solucionar este problema, sin perder una cantidad considerable de información útil, se utilizó el método de "predictive mean matching" para reemplazar los datos faltantes de las variables de interés.

	Overall (N=239343)
<b>bedrooms</b>	
Mean (SD)	3.08 (3.00)
Median [Min, Max]	3.00 [0, 336]
<b>bathrooms</b>	
Mean (SD)	2.92 (1.63)
Median [Min, Max]	3.00 [1.00, 20.0]
<b>lon</b>	
Mean (SD)	-75.0 (0.959)
Median [Min, Max]	-74.8 [-81.7, -73.6]
<b>lat</b>	
Mean (SD)	5.73 (2.46)
Median [Min, Max]	4.71 [3.22, 12.6]
<b>surface_total</b>	
Mean (SD)	1140 (6110)
Median [Min, Max]	220 [-36.0, 198000]
<b>surface_covered</b>	
Mean (SD)	929 (67900)
Median [Min, Max]	172 [1.00, 12000000]
<b>property_type</b>	
Apartamento	111930 (46.8%)
Casa	69225 (28.9%)
Depósito	824 (0.3%)
Finca	777 (0.3%)
Local comercial	4492 (1.9%)
Lote	2840 (1.2%)
Oficina	5362 (2.2%)
Otro	43564 (18.2%)
Parqueadero	311 (0.1%)
PH	18 (0.0%)

### 2.1.3

Los modelos a estimar son los siguientes:

$$(1) y_i = \beta_0$$

$$(2) y_i = \beta_0 + \beta_1(x_1 i^2) + \beta_2(x_2 i^2) + \beta_3(x_3 i) + \beta_4(x_4 i) + \beta_5(x_5 i^2) + \beta_6(x_6 i^2)$$

$$(3) y_i = \beta_0 + \beta_1(x_1 i) + \beta_2(x_2 i) + \beta_3(x_3 i) + \beta_4(x_4 i) + \beta_5(x_5 i) + \beta_6(x_6 i)$$

$$(4) y_i = \beta_0 + \beta_1(x_1 i^2) + \beta_2(x_2 i^2) + \beta_3(x_3 i) + \beta_4(x_4 i) + \beta_5(x_5 i^2)$$

$$(5) y_i = \beta_0 + \beta_1(x_1 i * x_2 i) + \beta_5(x_5 i) + \beta_6(x_6 i)$$

Donde:

$[y_i]$  es el precio del producto i

$x_{1i}$  es Número de habitaciones del producto i

$x_{2i}$  es Número de baños del producto i

$x_{3i}$  es la longitud producto i

$x_{4i}$  es la latitud del producto i

$x_{5i}$  es la superficie total del producto i

$x_{6i}$  es la superficie cubierta del producto i

Error cuadrático medio (out of sample)				
modelo 1	modelo 2	modelo3	modelo 4	modelo 5
5,12E+18	4,89E+18	4,89E+18	4,89E+18	4,72E+18

La tabla anterior muestra los errores cuadráticos medios afuera de muestra de los 5 modelos estimado. Es posible notar que el modelo 5 es el modelo con menor error cuadrático medio. Este modelo muestra una relación creciente entre el precio y el número de baños, el número de habitaciones, la interacción entre ambos y la superficie cubierta. En adición a esto, presenta un R cuadrado de 0.45, lo que implica que la variación en las variables explicativas explica en un 45 por ciento la variación del precio.

#### 2.1.4

#### 2.1.5

(a)

Para estimar el error por el método de LOOCV del modelo (5), se calculó la siguiente expresión:

$$CV = \frac{1}{n} \sum^N MSE_i$$

Se encontró un error de LOOCV de 9,64E+18

(b)

$$h_i = X_i(X'X)^{-1}X_i'$$

$$\sum^N TX_t'X_t = X'X$$

$$\begin{aligned} (X_i'X_i)^{-1} &= (X'X - X_i'X_i)^{-1} \\ &= (X'X)^{-1} + \frac{(X'X)^{-1}X_i'X_i(X'X)^{-1}}{1 - X_i(X'X)^{-1}X_i'}. \end{aligned}$$

$$\begin{aligned}(X_i'X_i)^{-1}X_t' &= (X'X)^{-1}X_t' + (X'X)^{-1}X_i' \left( \frac{X_i(X'X)^{-1}X_i'}{1 - X_i(X'X)^{-1}X_i'} \right) \\ &= \left( \frac{1}{1 - h_i} \right) (X'X)^{-1}X_i'.\end{aligned}$$

$$X'X\hat{\beta} = X'y,$$

$$(X_i'X_i + X_i'X_i)\hat{\beta} = X_i'y_i + X_i'y_i,$$

$$\{I_k + (X_i'X_i^{-1}X_t'X_i)\}\hat{\beta} = \hat{\beta}_i + (X_i'X_i^{-1}X_i')(X_i\hat{\beta} + \hat{u}_i).$$

$$\begin{aligned}\hat{\beta} &= \hat{\beta}_i + (X_i'X_i^{-1}X_i'\hat{u}_i) \\ &= \hat{\beta}_{(i)} + (X'X)^{-1}X_i'\frac{\hat{u}_i}{1 - h_i},\end{aligned}$$

$$\hat{u}_i = \hat{u}_i + \left( \frac{\hat{u}_i}{1 - h_i} \right) h_i$$

$$\hat{u}_i = \frac{\hat{u}_i}{1 - h_i}$$

$$CV = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

(c)

A continuación se muestra los errores cuadráticos medios de cada uno de los 5 modelos. Es posible evidenciar que estos errores son mayores que los errores calculados afuera de muestra, con una partición de datos 70-30 entre entrenamiento y prueba. Esto puede querer decir que se estaba subestimando el error de predicción al no probarse el modelo a estimar en la totalidad de la muestra, lo que es corregido por el método LOOCV.

Error cuadrático medio LOOCV				
modelo 1	modelo 2	modelo3	modelo 4	modelo 5
1,01E+19	1,13E+19	9,86E+18	9,96E+18	9,64E+18

### 2.1.6

	Descomposición QR	Modelo lineal
intercept	1,65E+16	1,65E+16
bedrooms2	2,56E+11	2,56E+11
bathrooms2	2,52E+13	2,52E+13
lon	2,10E+14	2,10E+14
lat	-2,77E+13	-2,77E+13
surface_total2	1,27E+05	1,27E+05

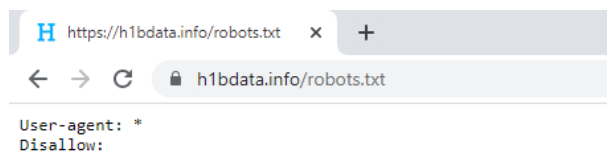
Tal como se puede observar, los resultados obtenidos entre la regresión lineal lm y los obtenidos por la descomposición QR son iguales. Esto es debido a que la estimación QR se logra a través de una transformación de matrices que no afecta los betas de las estimaciones. Esto mismo se observa al realizar la descomposición QR en subgrupos (por ciudades) y compararlos con los resultados del modelo lineal.

## 2.2 Reverend Bayes meets web scraping

### 2.2.1

### 2.2.2

Al revisar archivo robots.txt, que se puede apreciar en la siguiente image, se observa que no hay ningún tipo de restricción para hacer "scraping" a los datos.



### 2.2.3

### 2.2.4

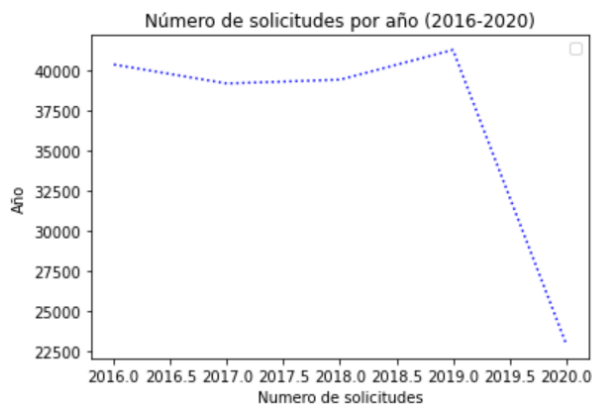
Para limpiar la base de datos primero reemplazamos todas las observaciones que que contenían valores missings NAN (NA), y segundo eliminamos todos los NAN de la base de datos. Posterior a la limpieza de la base de datos, la base de datos contiene información acerca de la promoción de empleos en 26492 compañías estadounidenses entre el año 2016 y 2020. La información que se dispone es la de las ciudades: Chicago, Houston, San Francisco, Los Angeles y New York.

count	2.283030e+05
mean	9.593226e+04
std	5.069068e+04
min	3.100000e+02
25%	6.700000e+04
50%	8.550000e+04
75%	1.150000e+05
max	1.100000e+07

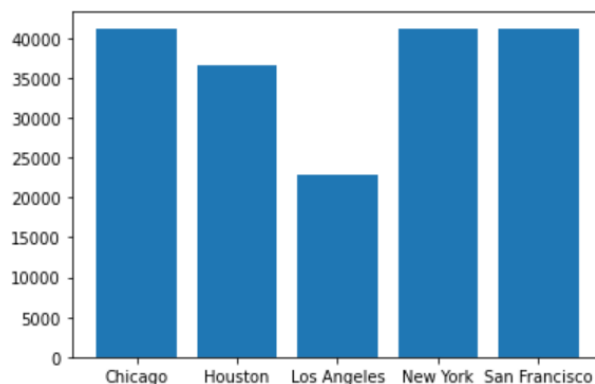
	count	mean	std	min	25%	50%	75%	max
location								
Chicago	49998.0	88445.349974	65777.969852	450.0	65000.0	80000.0	100000.0	11000000.0
Houston	49993.0	85971.124077	40610.175568	600.0	63000.0	77273.0	98821.0	1308000.0
Los Angeles	28325.0	87300.893769	43778.936260	603.0	60000.0	77200.0	102206.0	1200000.0
New York	49991.0	102282.847413	45685.182501	310.0	71700.0	92500.0	125000.0	1600000.0
San Francisco	49996.0	111920.138111	45734.450467	1777.0	80000.0	105000.0	135000.0	1826000.0

Las tablas anteriores muestran las estadísticas descriptivas del salario base en general y por ciudades.

## 2.2.5



En la anterior gráfica se puede apreciar que del año 2016 hasta el primer semestre del 2019 el número de solicitudes estuvo relativamente estable, con un pequeño aumento en el primer semestre del 2019. Sin embargo, en el segundo semestre del 2019 y el primer semestre del 2020 se ve una reducción bastante pronunciada en la cantidad de solicitudes.



En la anterior gráfica se puede apreciar que las ciudades con mayor número de solicitudes son, en orden, Chicago, Nueva York y San Francisco. En cuarto lugar se encuentra Houston con una cantidad inferior, y en último y quinto lugar está Los Angeles, que recibió muchas menos solicitudes que las otras cuatro ciudades.

#### 2.2.6

El top 3 de empresas por número de aplicaciones cambia ligeramente en el tiempo, sin embargo, la mayoría de los años las siguientes empresas ocupan el podio.

##### ERNST YOUNG US LLP

EY, como también es conocida es una de las más importantes firmas de servicios profesionales del mundo. La empresa, una de las Cuatro Grandes Auditoras, se remonta a principios del siglo XX. El escocés Arthur Young y el estadounidense Alwin Ernst fundaron sus empresas independientes en 1906 y 1903, respectivamente. Las empresas se fusionaron en 1989. Actualmente, EYa emplea a más de 200.000 personas y opera en más de 150 países.

##### DELOITTE CONSULTING LLP

Deloitte es la firma privada de servicios profesionales que más factura en el mundo (36.800 millones de dólares en 2016). Es una de las llamadas Cuatro Grandes Auditoras. The firm was founded by William Welch Deloitte in London in 1845 and expanded into the United States in 1890.

##### TATA CONSULTANCY SERVICES LIMITED

Tata Consultancy Services es una empresa multinacional india de consultoría y servicios de tecnología de la información con sede en Mumbai, India. Es la segunda empresa india más grande por capitalización de mercado. La compañía fue fundada en 1968.

### 2.2.7

$$\pi(\theta|v) = \frac{f(v|\theta)P(\theta)}{m(v)}$$

We know that:

$$f(v|\theta) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n}}(v-\theta)^2\right)$$

$$P(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right)$$

So:

$$\pi(\theta|v) = \frac{1}{m(v)\sqrt{4\pi^2\tau^2\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2\frac{\sigma^2}{n}}(v-\theta)^2 - \frac{1}{2\tau^2}(\theta-\mu)^2\right)$$

$$= \frac{1}{m(v)\sqrt{4\pi^2\tau^2\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2}\left(\frac{1}{\frac{\sigma^2}{n}}(v^2 - 2v\theta + \theta^2) + \frac{1}{\tau^2}(\theta^2 - 2\theta\mu + \mu^2)\right)\right)$$

De este modo, podemos sugerir que  $\pi(\theta|v)$  se distribuye  $N(m, V)$ , donde:

$$m = \left(\frac{\frac{1}{\frac{\sigma^2}{n}}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}\right)v + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}\right)\mu$$

$$V = \left(\frac{1}{\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\tau^2}}\right)$$

### 2.2.8

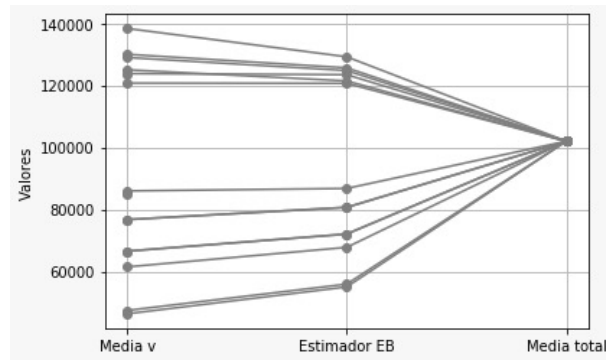
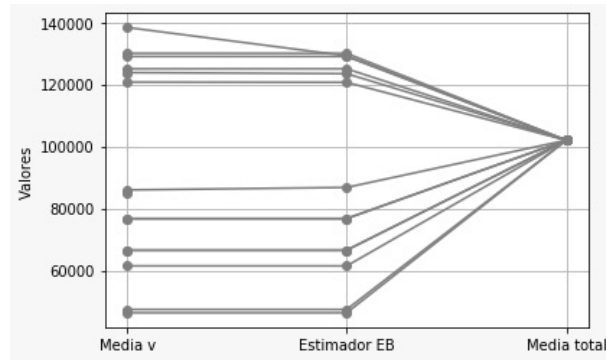
$$m(v) = \int \frac{1}{m(v)\sqrt{4\pi^2\tau^2\frac{\sigma^2}{n}}} \exp\left(-\frac{1}{2}\left(\frac{1}{\frac{\sigma^2}{n}}(v^2 - 2v\theta + \theta^2) + \frac{1}{\tau^2}(\theta^2 - 2\theta\mu + \mu^2)\right)\right) d\theta$$

$$= \frac{1}{m(v)2\pi\tau\frac{\sigma}{\sqrt{n}}} \int \exp(K_0 \frac{|Z|}{\frac{\sigma}{\sqrt{n}}\tau})$$

Donde  $K_0$  es una función Bessel modificada.



### 2.2.9



La primera estimación resulta en la aproximación a la media de solo algunas variables. Específicamente, aquellas variables que tienen varianza cero permanecen iguales. Esto es debido a que la estimación bayeciana resulta en una suma ponderada por las varianzas de la media prior y la media observada. Dado que hay empresas con una una observación, la varianza de la media observada resulta en cero y la estimación bayesiana de la media de estas compañías resulta siendo la misma que la media observada. Sin embargo, en una segunda estimación, se remplazó la varianza de estas compañías con una sola observación con el promedio de varianzas de toda la muestra. De este modo, se le da más peso a la media prior en la estimación bayesiana y, por lo tanto, se observa un encogimiento más pronunciado en este segundo escenario.

### 3 Pedes in terra ad sidera visus

The screenshot displays the RStudio IDE interface. The top pane shows a script with R code for data manipulation and matrix operations. The bottom-left pane shows the console output, which includes several error messages related to singular matrices and computational issues. The bottom-right pane shows the Environment pane, listing variables and their attributes.

**Script Content:**

```
1 # Welcome! This RStudio Amazon AMI contains RStudio Server version 1.3.1073,
2 # running R 4.0.2 on Ubuntu 18.04 LTS.
3 # Includes support for Shiny (add shiny/rstudio to URL).
4 # NEW: experimental support for CUDA 10.1 (incl. cuDNN 7.6.5) and Magma 2.5.3
5 # enabling use of GPU packages in R and higher performance for deep
6 # learning frameworks such as TensorFlow.
7 # AMI created by Louis Aslett (OASIS://www.louisaslett.com/). If you've
8 # any comments or suggestions please mail louis.aslett@durham.ac.uk
9
10 # NOTE: It is recommended that you change the password for logging into RStudio,
11 # which you can do by changing from the Console to the Terminal tab below,
12 # launching a new terminal, and running the passwd command.
13 # Alternatively, since this AMI was created to make RStudio Server accessible to
14 # those who are less comfortable with Linux commands you can follow the
15 # instructions below to change it without touching Linux.
16
17 # There is a mini package where functions to manipulate the server will be
18 # placed. This includes a function to change the password. First load the
19 # package:
20 library("RStudioAMI")
21
22 # Now you can change the password by just running the following function. It
23 # will prompt you to provide the existing password (just the instance ID) and
24 # then type in a new password
25 passwd()
```

**Console Output:**

```
~/ -
system is computationally singular: reciprocal condition number = 9.1573e-24
> h0<-subset(x2, select = c(h0a, surface, total ))
> xml<- data.matrix(x6)
> hml<-data.frame()
> for (i in (1:nrow(xoc))) {
+   hml[i,1]=c(xml[i,1])*ksolve(c(xml[i]*xkm+i)*xkm[i,1]
+ }
> View(xml)
> dse<-solve(c(cml[i]*xkm[i]))
Error in solve.default(c(cml[i]*xkm[i])) :
system is computationally singular: reciprocal condition number = 9.1573e-24
> dse<-solve(c(cml[i]*xkm[i]), tol = 1e-20)
Error in solve.default(c(cml[i]*xkm[i]), tol = 1e-20) :
system is computationally singular: reciprocal condition number = 9.1573e-24
> dse<-solve(c(cml[i]*xkm[i]), tol = 1e-60)
Error in solve(c(cml[i]*xkm[i]), tol = 1e-60) :
system is computationally singular: reciprocal condition number = 9.1573e-24
> for (i in (1:nrow(xoc))) {
+   hml[i,1]=c(xml[i,1])*xkm[i,1]
+ }
```

**Environment Pane:**

Variable	Attributes
coc	239343 obs. of 27 variables
do	num [1:6, 1:6] 2.78e-11 -7.02e-11 -4.37e-12 -1.07e-11 7.98e-20 ...
hml	85685 obs. of 1 variable
hml4	85 obs. of 1 variable
x2	239343 obs. of 32 variables
x3	239343 obs. of 6 variables
x6	239343 obs. of 2 variables
xml	Large matrix (343688 elements, 11.5 MB)
xml4	Large matrix (476866 elements, 3.8 MB)

**Files Pane:**

Name	Description	Version
RStudioAMI	Control of RStudio AMI Instance	0.3.0
arrow	Integration to 'Apache' Arrow	1.0.0
askpass	Safe Password Entry for R, Git, and SSH	1.1
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.8
base	The R Base Package	4.0.2
base64enc	Tools for base64 encoding	0.1-3
bayesplot	Plotting for Bayesian Models	1.7.2
bit	Boost C++ Header Files	1.72.0-3
bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
bit64	A 53 Class for Vectors of 64bit Integers	4.0.2
bitops	Bitwise Operations	1.0-6
blns	A Simple S3 Class for Representing Vectors of Binary Data (BLOBs)	1.2.1
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-25
broom	Convert Statistical Objects into Tidy Tibbles	0.7.0
cairo	R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PDF, JPEG, TIFF), Vector (PDF, SVG, Postscript) and Display (X11 and Win32) Output	1.5-12.2
callr	Call R from R	3.4.3
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
checkmate	Fast and Versatile Argument Checks	2.0.0
class	Functions for Classification	7.3-17
cli	Helpers for Developing Command Line Interfaces	2.0.2
clipr	Read and Write from the System Clipboard	0.7.0