Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Problem Set 1

1. Theory Exercises: Review Stats and OLS:

1.

a. Cuál de los siguientes conjuntos de números son suficientes para el cálculo de P(X| ε 1, ε2) = P(X = 1| ε 1, ε 2). . . P(X = K| ε 1, ε 2):

Por Teorema de Bayes se tiene que:

$$\mathrm{P(A|\ B)} = \frac{\mathrm{P(B|\ A)} * \mathrm{P(A)}}{\mathrm{P(B)}}$$

Por lo tanto, el conjunto necesario es el ii:

$$\mathrm{P(X|\ \varepsilon\ 1, \varepsilon 2)} = \frac{\mathrm{P(\varepsilon\ 1, \varepsilon 2|\ X)} * \mathrm{P(X)}}{\mathrm{P(\varepsilon\ 1, \varepsilon 2)}}$$

b. Asumiendo que E1 ⊥ E2|X , el conjunto suficiente para encontrar la probabilidad es el i) dado que :

$$\mathrm{P(\ \varepsilon\ 1, \varepsilon 2\ |X)} = \mathrm{P(\varepsilon\ 1|\ X)} * \mathrm{P(\varepsilon\ 2|\ X)}$$

Por lo tanto:

$$\mathrm{P(X|\ \varepsilon\ 1, \varepsilon 2)} = \frac{\mathrm{P(\varepsilon\ 1|\ X)} * \mathrm{P(\varepsilon\ 2|\ X)} * \mathrm{P(X)}}{\mathrm{P(\varepsilon\ 1, \varepsilon 2)}}$$

2. Considerando una regresión lineal simple yi = α + βxi + ε, i = 1, ..., N

a. Muestre que E(ε i |xi) = 0 implicando que E(ε i) = 0 y E(ε i|xi) = 0

$$E(\ \varepsilon\ \mathrm{i|xi}) = \mathrm{E}(\mathrm{y_i} - \widehat{\mathrm{y_i}}|\mathrm{x_i}) = y_i - \widehat{\propto} - \hat{\beta} * x_i = 0$$

$$y_i - \widehat{\propto} - \hat{\beta} * x_i = 0$$

$$\sum_{i=1}^{N} y_i - \widehat{\propto} - \hat{\beta} * x_i = 0$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \widehat{\propto} - \sum_{\iota=1}^{n\frown} \hat{\beta} * x_\iota = 0$$

Multiplicando 1/N en los dos lados de la ecuación:

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$\frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N} \widehat{\propto} - \frac{1}{N}\sum_{\iota=1}^{\widehat{n}} \hat{\beta} * x_\iota = 0$$

Como se sabe que:

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

Por lo tanto reemplazando en la ecuación:

$$\bar{y} - \widehat{\propto} - \hat{\beta}\,\bar{x} = 0$$

Se sabe que:

$$\widehat{\propto} = \bar{y} - \hat{\beta}\,\bar{x}$$

Al reemplazar se cumple la igualdad, por lo tanto E(ε i |xi) = 0.

b. Use las implicaciones previas para encontrar el estimador por el método de momentos:

El estimador de α está dado por la siguiente expresión, solo falta encontrar el valor de β:

$$\widehat{\propto} = \bar{y} - \hat{\beta}\,\bar{x}$$

Se sabe que Cov(ε, x) =0

$$\text{Cov}(\,\varepsilon, \text{x}) = Cov(y_i - \widehat{\propto} - \hat{\beta} * x_i, x) = \frac{1}{N} * \sum_{i=1}^{N} (y_i - \widehat{\propto} - \hat{\beta} * x_i) * x_i = 0$$

Sustituyendo el valor de α:

$$\frac{1}{N} * \sum_{i=1}^{N} (y_i - \bar{y} - \hat{\beta}\,\bar{x} - \hat{\beta} * x_i) * x_i = 0$$

Reorganizando la ecuación:

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$\sum_{i=1}^{N}(y_i - \bar{y}) * x_i = \hat{\beta} \sum_{i=1}^{N}(x_i - \bar{x}) * x_i$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(y_i - \bar{y}) * x_i}{\sum_{i=1}^{N}(x_i - \bar{x}) * x_i} = \frac{\sum_{i=1}^{N}(y_i - \bar{y}) * (x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

c. Acomode los términos previos para encontrar la solución matricial

$$\hat{\beta} = Min \sum_{i=1}^{N} \varepsilon_i$$

$$\hat{\beta} = MIn \ \varepsilon' * \varepsilon$$

Siendo $\varepsilon = Y - X\hat{\beta}$

Por lo tanto:

$$\varepsilon' * \varepsilon = (Y - X\hat{\beta})' * (Y - X\widehat{\beta}) = Y'Y - Y'X\hat{\beta} - \widehat{\beta}'X'Y + \widehat{\beta}'X'X\hat{\beta}$$

$$\varepsilon' * \varepsilon = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

C.P.O

$$\frac{d\,\varepsilon' * \varepsilon}{d\hat{\beta}} = -2X'Y + 2(X'X)\hat{\beta} = 0$$

$$(X'X)\hat{\beta} = X'Y$$

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

3. Pruebe las siguientes propiedades de R2:

a. El MCO maximiza el R2:

Como se sabe la formula del R cuadrado es :

$$R^2 = 1 - \frac{SCE}{SCT}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Se sabe que el modelo de cuadrados ordinarios minimiza la suma de las distancias verticales entre los valores observados y los valores estimados, es decir:

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$Min \sum_{i=1}^{N} (y_i - \hat{y})$$

Por lo tanto al minimizar este valor se está maximizando el valor de R cuadrado.

b. 0 ≤ R2 ≤ 1

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

$$SCT = SCR + SCE$$

En primer lugar se sabe que el valor de R cuadrado es la proporción de la Suma de cuadrados de la regresión y la suma de cuadrados totales, por lo tanto es un valor positivo >=0. Por otro lado la SCR< SCT , por lo tanto el valor máximo que puede alcanzar es 1. De esta manera se demuestra que el R2 esta acotado entre 0 y 1.

c. Probar que el factor de correlación r2= R2 en una regresión simple:

$$r_{xy}{}^2 = R^2 = 1 - \frac{SCE}{SCT}$$

Dado que para una regresión simple el valor de $\widehat{\beta}$ es igual a:

$$\widehat{\beta} = \frac{Cov(XY)}{VAR(X)}$$

Por lo tanto:

$$\hat{y} = \propto + \widehat{\beta}\, x_i = (\bar{y} - \widehat{\beta}\, \bar{x}) + \widehat{\beta}\, x_i = \bar{y} + \frac{Cov(XY)}{VAR(X)}(x_i - \bar{x})$$

$$SCE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{N}\left(y_i - \bar{y} + \frac{Cov(XY)}{VAR(X)}(x_i - \bar{x})\right)^2$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$SCE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \overline{y})^2 \; + (\frac{Cov(XY)}{VAR(X)})^2(x_i - \bar{x})^2 - 2\frac{Cov(XY)}{VAR(X)}(x_i - \bar{x})(y_i - \bar{y})$$

$$SCE = VAR(Y) + (\frac{Cov(XY)}{VAR(X)})^2(VAR(X))^2 - 2\frac{Cov(XY)^2}{VAR(X)} = VAR(Y) - \frac{Cov(XY)^2}{VAR(X)}$$

$$r_{xy}{}^2 = R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{VAR(Y) - \frac{Cov(XY)^2}{VAR(X)}}{VAR(Y)} = \frac{Cov(XY)^2}{VAR(X)VAR(Y)} = r_{xy}{}^2$$

4. Considerando el modelo

$$y = \beta_1\iota + X_2\beta_2 + u$$

Muestre usando el teorema FWL, que los estimadores de MCO de β_1 y β_2 puende ser escritos como:

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}n & \iota'X_2\\ o & X_2'M_\iota X_2\end{pmatrix}^{-1}\begin{pmatrix}\iota y\\ X_2'M_\iota y\end{pmatrix}$$

Tenemos que:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$(X'X)\hat{\beta} = X'y$$

Entonces:

$$\begin{pmatrix}\iota'\iota & \iota'X_2\\ X_2'\iota & X_2'X_2\end{pmatrix}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}\iota'y\\ X_2'y\end{pmatrix}$$

Para $\hat{\beta}_1$, tenemos:

$$(\iota'\iota)\hat{\beta}_1 + \iota'X_2\hat{\beta}_2 = \iota'y$$

$$(\iota'\iota)\hat{\beta}_1 = \iota'y - \iota'X_2\hat{\beta}_2$$

$$\hat{\beta}_1 = (\iota'\iota)^{-1}\iota'y - (\iota'\iota)^{-1}\iota'X_2\hat{\beta}_2$$

$$\hat{\beta}_1 = (\iota'\iota)^{-1}\iota'\left(y - X_2\hat{\beta}_2\right) \qquad (1)$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Para $\hat{\beta}_2$, tenemos:

$$(X_2'\iota)\,\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'y$$

Reemplazando en (1), tenemos

$$(X_2'\iota)\,(\iota'\iota)^{-1}\iota'\left(y - X_2\hat{\beta}_2\right) + X_2'X_2\hat{\beta}_2 = X_2'y$$

$$(X_2'\iota)(\iota'\iota)^{-1}\iota'y - (X_2'\iota)(\iota'\iota)^{-1}\iota'X_2\hat{\beta}_2 + X_2'X_2\hat{\beta}_2 = X_2'y$$

$$X_2'X_2\hat{\beta}_2 - (X_2'\iota)(\iota'\iota)^{-1}\iota'X_2\hat{\beta}_2 = X_2'y - (X_2'\iota)(\iota'\iota)^{-1}\iota'y$$

$$\hat{\beta}_2(X_2'X_2 - (X_2'\iota)(\iota'\iota)^{-1}\iota'X_2) = X_2'y - (X_2'\iota)(\iota'\iota)^{-1}\iota'y$$

$$\hat{\beta}_2(X_2'X_2 - (X_2'\iota)(\iota'\iota)^{-1}\iota'X_2) = X_2'(I - \iota(\iota'\iota)^{-1}\iota')y$$

Teniendo en cuenta que:

$$M = (I - X(X'X)^{-1}X'$$

Entonces:

$$\hat{\beta}_2(I - X_2'\iota(\iota'\iota)^{-1}\iota'X_2) = X_2'(I - \iota(\iota'\iota)^{-1}\iota')y$$

$$\hat{\beta}_2(X_2'M_\iota X_2) = X_2'M_\iota y$$

$$\hat{\beta}_2 = (X_2'M_\iota X_2)^{-1}X_2'M_\iota y \quad (2)$$

Con (1) y (2) tenemos que :

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}n & \iota'X_2\\ o & X_2'M_\iota X_2\end{pmatrix}^{-1}\begin{pmatrix}\iota y\\ X_2'M_\iota y\end{pmatrix}$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

5. Considerando el modelo de regresión lineal con un intercepto y un regreso, asumiendo que los supuestos del modelo clásico de Gauss Markov, muestre que:

$$E(\hat{y}_{test} - Y_{test}) = 0, \text{ y que}$$

$$var(\hat{y}_{test} - Y_{test}) = \sigma_0(1 + \frac{1}{n} + \frac{(x_{test} - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2})$$

Con base en los supuestos del modelo clásico de Gauss Markov tenemos que:
$Y_{testi} = \beta_0 + \beta_1 x_i + u_i$ es la función de regresión poblacional

$Y_{testi} = \widehat{\beta_0} + \widehat{\beta_1} x_i + e_i$ es la función de regresión muestral, siendo $\widehat{\beta_0}$ y $\widehat{\beta_1}$ estimadores de $\beta_0 y\ \beta_1$respectivamente,

cumpliéndose que $\hat{y}_{test} = \widehat{\beta_0} + \widehat{\beta_1} x_i = E(y_{test}|x_i)$,

$E(\widehat{\beta_0}) = \beta_0, E(\widehat{\beta_1}) = \beta_1, E(u_i|x_i) = 0 = E(u_i) = E(E(u_i|x_i))$por la ley de expectativas iteradas, entonces

$$E(\hat{y}_{test} - Y_{testi}) = E(\hat{y}_{test}) - E(Y_{testi}) = \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 x_i - E(u_i) = 0$$

$$E(\hat{y}_{test} - Y_{testi}) = 0$$

Por otro lado,

$$var(\hat{y}_{test} - Y_{test}) = E(\hat{y}_{test} - Y_{test})^2$$

$$\hat{y}_{test} - Y_{test} = (\widehat{\beta_0} - \beta_0) + (\widehat{\beta_1} - \beta_1)x_i - u_i$$

$$E(\hat{y}_{test} - Y_{test})^2 = (\widehat{\beta_0} - \beta_0)^2 + (\widehat{\beta_1} - \beta_1)^2 x_i^2 - u_i^2 + 2(\widehat{\beta_0} - \beta_0)(\widehat{\beta_1} - \beta_1)x_i - 2(\widehat{\beta_0} - \beta_0)u_i - 2(\widehat{\beta_1} - \beta_1)x_i u_i$$

$$(\widehat{\beta_0} - \beta_0)^2 = var(\widehat{\beta_0}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\bar{x}^2 + \frac{\sigma^2}{n}$$

$$(\widehat{\beta_1} - \beta_1)^2 = var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$u_i^2 = \sigma^2$$

$$(\widehat{\beta_0} - \beta_0)(\widehat{\beta_1} - \beta_1) = cov(\widehat{\beta_0}\widehat{\beta_1}) = -var(\widehat{\beta_1})\bar{x}$$

$$E(\hat{y}_{test} - Y_{test})^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\bar{x}^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}x_i^2 - 2x_i\bar{x}\frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} + \sigma^2$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$var(\hat{y}_{test} - Y_{test}) = \sigma^2(1 + \frac{1}{n} + \frac{(x_{test} - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2})$$

7. Considerando el modelo de regresión lineal con un regreso y sin intercepto, donde xi está estandarizada para tener media cero y varianza 1. Se asumen todos los supuestos del modelo clásico de regresión lineal. Considerando $\widehat{\beta_\lambda}$ como el ridge estimator:

La solución al problema de minimización para obtener el ridge estimator da como resultado:

$$\widehat{\beta_\lambda} = (X^T X + \lambda I)^{-1} X^T y, \text{ donde I es la matriz identidad.}$$
$$\widehat{\beta_\lambda} = (X^T X + \lambda I)^{-1} X^T X\beta + (X^T X + \lambda I)^{-1} X^T U$$

A diferencia del estimador de OLS, para el ridge estimator no se requiere asumir que la matriz X es de rango completo.

Para calcular la varianza y el sesgo del ridge estimator usamos las definiciones de valor esperado y varianza condicional y los supuestos clásicos del modelo de Gauss Markov.

$$E(\widehat{\beta_\lambda}|X) = (X^T X + \lambda I)^{-1} X^T X\beta$$

Esta esperanza condicionada del ridge estimator difiere de la de OLS a menos que $\lambda = 0$. Así entonces, el sesgo del estimador está dado por:

$$E\left(\widehat{\beta_\lambda}\middle|X\right) - \beta = [(X^T X + \lambda I)^{-1} - (X^T X)^{-1}] X^T X\beta$$

En cuanto a la varianza condicional del estimador, se tiene:

$$var(\widehat{\beta_\lambda}|X) = E\left(\widehat{\beta_\lambda} - E(\widehat{\beta_\lambda})\right)^2$$

La matriz var-cov de los estimadores condicionados a las X´s está dada por:

$$var(\widehat{\beta_\lambda}|X) = E[(X^T X + \lambda I)^{-1} X^T U U^T X (X^T X + \lambda I)^{-1}]$$
$$var(\widehat{\beta_\lambda}|X) = [\sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}]$$

Considerando que:

$$var(\widehat{\beta_{MCO}}|X) = \sigma^2 (X^T X)^{-1}$$

La varianza del ridge estimator es siempre menor que la de MCO. Esto se debe a que la diferencia entre la matriz var-cov de MCO y la del ridge estimator, dada por:

$$var\left(\widehat{\beta_{mco}}\middle|X\right) - var(\widehat{\beta_\lambda}|X),$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Es definida positiva. De acuerdo a Gauss Markov, las matrices de covarianza de dos estimadores se comparan analizando si la matriz resultante de su resta es positiva definida.

El error cuadrático medio del ridge estimator está definido por la traza de la matriz de varianza-covarianza más el cuadrado de la norma del sesgo.

$$MSE(\widehat{\beta_\lambda}|X) = E[\parallel \widehat{\beta_\lambda} - \beta \parallel^2 |X]$$

$$MSE(\widehat{\beta_\lambda}|X) = trace\left(var[\widehat{\beta_\lambda}|X]\right) + \parallel bias((\widehat{\beta_\lambda}|X) \parallel^2$$

El estimador MCO tiene cero sesgos, entonces

$$MSE(\widehat{\beta_{mco}}|X) = E[\parallel \widehat{\beta_{mco}} - \beta \parallel^2 |X]$$

$$MSE(\widehat{\beta_{mco}}|X) = trace\left(var[\widehat{\beta_{mco}}|X]\right)$$

La diferencia entre los MSEs está dada por:

$$MSE(\widehat{\beta_{mco}}|X) - MSE(\widehat{\beta_\lambda}|X) = trace\left(var[\widehat{\beta_{mco}}|X] - var[\widehat{\beta_\lambda}|X]\right) - \parallel bias((\widehat{\beta_\lambda}|X) \parallel^2$$

El ridge estimator tiene menor varianza, pero mayor sesgo. La suma de las trazas de las dos matrices es igual a la traza de su suma. En este caso como solo tenemos un parámetro a estimar, la ecuación del MSE se convierte en:

$$MSE(\widehat{\beta_{mco}}|X) = \left(var[\widehat{\beta_{mco}}|X]\right)$$

$$MSE(\widehat{\beta_\lambda}|X) = \left(var[\widehat{\beta_\lambda}|X]\right) + \left(bias\left(\widehat{\beta_\lambda}|X\right)\right)^2$$

$$MSE(\widehat{\beta_{mco}}|X) - MSE(\widehat{\beta_\lambda}|X) = \left(var[\widehat{\beta_{mco}}|X] - var[\widehat{\beta_\lambda}|X]\right) - \left(bias\left(\widehat{\beta_\lambda}|X\right)\right)^2$$

Como la diferencia en las varianzas es positiva y el cuadrado del término del error es estrictamente positivo, la diferencia en los MSE puede ser negativa o positiva. Se puede demostrar que el signo de esta diferencia dependerá en el parámetro λ, siendo posible encontrar un valor de λ tal que la diferencia sea positiva. Es decir, se puede encontrar un λ tal que el ridge estimator tenga un MSE menor que el de MCO.

Para demostrar lo anterior nos basamos en Theobald (1974).

$$\begin{aligned} MSE(\widehat{\beta_{mco}}|X) - MSE(\widehat{\beta_\lambda}|X) \\ = \sigma^2(X^TX)^{-1} - \sigma^2 X^TX(X^TX + \lambda I)^{-2} \\ - \lambda^2(X^TX + \lambda I)^{-1}\beta\beta^T(X^TX + \lambda I)^{-1} \end{aligned}$$

$$= \lambda(X^TX + \lambda I)^{-1}[\sigma^2\{2I + (X^TX)^{-1}\} - \lambda\,\beta\beta^T](X^TX + \lambda I)^{-1}$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Para $\lambda > 0$ la anterior expresión es una matriz positiva definida si y solo si:

$$\sigma^2\{2I + \lambda(X^TX)^{-1}\} - \lambda\,\beta\beta^T$$

es positiva definida. Esto es cierto si:

$$2\sigma^2 I - \lambda\,\beta\beta^T$$

Es positiva definida. Teniendo en cuenta que las raíces de $2\sigma^2 I - \lambda\,\beta\beta^T$ son $2\sigma^2$ y $2\sigma^2 - \lambda\,\beta^T\beta$, la condición suficiente será:

$$\lambda < 2\sigma^2/\beta^T\beta$$

Así entonces, si $\lambda < 2\sigma^2/\beta^T\beta$ cualquier comparación entre $\widehat{\beta_{mco}}$ y $\widehat{\beta_\lambda}$ bajo el criterio de MSE va a favorecer a $\widehat{\beta_\lambda}$ (ridge estimator).

Lo que podemos concluir acerca de la práctica clásica en econometría, que favorecía los estimadores insesgados sobre los sesgados, es que permitiendo algo de sesgo podemos encontrar mejores estimadores desde una perspectiva del error cuadrático medio. Es decir, tolerando algo de sesgo encontramos estimadores con MSE menor al de estimadores insesgados.

2.1 Predicting House Prices in Colombia

1. Para predecir los precios de los inmuebles en Colombia, se cuenta con la base de datos "house_prices_ps1", la cual contiene 239.434 registros y 28 variables, las más representativas y con las que vamos a trabajar el modelo son:

- Lat: latitud de la ubicación del inmueble
- Lon: longitud de la ubicación del inmueble
- L3: Ciudad
- Rooms: Número de salones del inmueble
- Bedrooms: Número de habitaciones del inmueble
- Bathrooms: Número de baños del inmueble
- Surface_total: área del terreno del inmueble
- Surface_covered: área de la construcción del inmueble

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

- Price: Precio del inmueble
- Currency: Tipo de moneda en la cual está representado el precio
- Property_type: tipo de Propiedad

2. En la siguiente tabla se muestra las estadísticas descriptivas de las variables que potencialmente se utilizaran para la predicción de los precios de las casas en Colombia.

Tabla1. Estadísticas descriptivas

Feature	Media	Desviación estándar	Máximo	Mínimo	Mediana	Valores Nulos
rooms	3,26529	1,8614	40	1	3	157.341
bedrooms	3,20643	2,6378	336	0	3	152.718
bathrooms	2,86332	1,5330	20	1	2	31.736
surface Total	415,50307	3.567,71	198.000	-36	115	156.169
surface covered	517,82930	45.784,98	12.000.000	1	112	162.288
price	841.577.426,63	3.183.555.192	850.000.000.000	0	420.000.000	0

Como se puede observar en la tabla 1 se encuentra que existen gran cantidad de valores nulos, la base tiene en total 239.434 observaciones y para el caso de la variable Surface covered se encuentra que el 67.77% de las observaciones están sin dato, lo cual presenta grandes problemas para la estimación de un modelo a partir de la información de la base de datos "house_prices_ps1".

Con el fin de imputar los datos, se realizan dos procesos el primero consistió en un análisis de las variables en la base de datos y a partir de esta se realizó una limpieza y asignación de valores a las variables de acuerdo con la tipificación de las propiedades, y en el segundo se aplicó el proceso de Hot Deck, el cual permite a partir la formación de grupos asignar los valores de la variables según su comportamiento al interior del grupo, y asigna un valor aleatorio seleccionado dentro del grupo a los valores faltantes .

A continuación, se muestran lo que se aplico en el primer proceso.

a. Se eliminaron lo valores en 0 para las variables precio
b. A las variables rooms, bedrooms, bathrooms y Surface_covered con valores nulos se les asignó 0 en caso de que el tipo de propiedad sea lote
c. A la variable Surface_total se le asignó 0 en caso de los tipos de propiedad, oficina, apartamento o PH, esto bajo el supuesto que ninguno de ellos posee un área de terreno más allá de un coeficiente de propiedad horizontal el cual está entre 0 y 1

Para el segundo proceso como se menciono anteriormente, se aplicó el proceso de imputación de Hot Deck, sin embargo, aún quedaron algunos registros nulos, esto debido a la gran cantidad de observaciones sin información, por lo tanto, se aplicaron medias entre grupos para asignar los valores.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Finalmente se eliminaron 133 observaciones que quedaron con valor nulos en las variables Surface total y Surface covered.

3. **The Taming of the Shrew"**

a. Modelo con solo de la constante

```
================================================
Dependent variable:
---------------------------
lnprice
------------------------------------------------
Constant                      19.954***
(0.002)

------------------------------------------------
Observations                   148,146
R2                              0.000
Adjusted R2                     0.000
Residual Std. Error     0.865 (df = 148145)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Los modelos se estimaron teniendo en cuenta el logaritmo del precio de la vivienda, en este caso se estima un promedio del error de la predicción de **0.8597777**

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

b. Modelo con múltiples variables, interacción y transformaciones

Tabla 2. Comparación de modelos n=148,146

	Dependent variable:						
	lnprice						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	18.93***	18.78***	18.79***	207.31***	207.76***	543.41***	543.91***
	(0.004)	(0.01)	(0.01)	(2.81)	(2.79)	(8.10)	(8.09)
bathrooms	0.32***	0.32***	0.32***	0.35***	0.35***	0.35***	0.35***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
rooms	0.02***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
bedrooms		-0.002***	-0.002***	0.05***	0.05***	0.05***	0.05***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
factor(property_type)Casa	-0.11***	-0.06***	-0.06***	-0.06***	-0.06***	-0.05***	-0.05***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
factor(property_type)Finca	0.31***	0.54***	0.41***	0.59***	0.46***	0.46***	0.45***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
factor(property_type)Local comercial	0.75***	0.73***	0.72***	0.81***	0.80***	0.80***	0.80***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
factor(property_type)Lote	1.77***	1.84***	1.69***	2.02***	1.87***	1.86***	1.86***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
factor(property_type)Oficina	0.65***	0.56***	0.56***	0.61***	0.60***	0.57***	0.57***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
factor(property_type)Otro	0.09***	0.14***	0.12***	0.15***	0.14***	0.14***	0.14***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

factor(property_type)Parqueadero	0.85***	0.78***	0.76***	0.89***	0.87***	0.88***	0.88***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
factor(l3)Bogotá D.C		0.32***	0.32***	1.04***	1.01***	-6.50***	-6.51***
		(0.01)	(0.01)	(0.18)	(0.18)	(0.25)	(0.25)
factor(l3)Cali		-0.24***	-0.24***	7.35***	7.32***	3.96***	3.95***
		(0.01)	(0.01)	(0.23)	(0.23)	(0.24)	(0.24)
factor(l3)Medellín		0.10***	0.10***	4.07***	4.05***	-0.03	-0.04
		(0.01)	(0.01)	(0.14)	(0.14)	(0.17)	(0.17)
surface_total			0.0000***		0.0000***	0.0000***	0.0000***
			(0.0000)		(0.0000)	(0.0000)	(0.0000)
surface_covered			0.0000			0.0000	
			(0.0000)			(0.0000)	
lat				0.42***	0.41***	-42.74***	-42.79***
				(0.03)	(0.03)	(0.98)	(0.98)
lon				2.58***	2.59***	6.93***	6.94***
				(0.04)	(0.04)	(0.11)	(0.11)
poly(surface_covered, 3)1							0.61
							(0.62)
poly(surface_covered, 3)2							-3.78***
							(0.62)
poly(surface_covered, 3)3							8.56***
							(0.63)
bathrooms:bedrooms				-0.01***	-0.01***	-0.01***	-0.01***
				(0.0002)	(0.0002)	(0.0002)	(0.0002)
lat:lon						-0.56***	-0.56***
						(0.01)	(0.01)
Note:							*p<0.1; **p<0.05; ***p<0.01

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

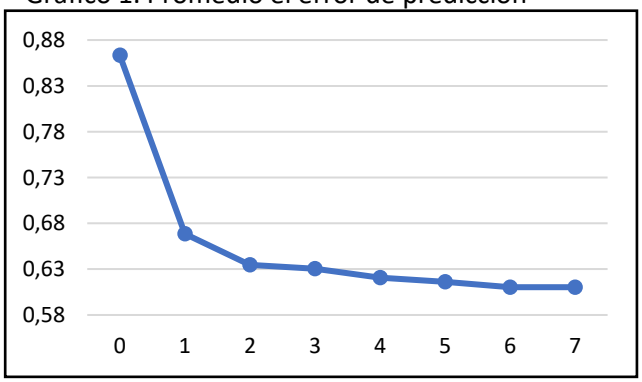
La anterior tabla nos muestra la comparación de los diferentes modelos estimados, estos modelos son:

i. Este modelo incluye 4 variables entre las cuales se encuentra el tipo de propiedad la cual se ingresa al modelo como una dummy.
ii. Este modelo incluye 5 variables entre las cuales se encuentra el tipo de propiedad y la ciudad donde se ubica la propiedad, las cuales se ingresan al modelo como una dummy
iii. Este modelo incluye 7 variables entre las cuales se encuentra el tipo de propiedad y la ci udad de la propiedad las cuales se ingresan al modelo como una dummy y las áreas de terreno y construcción.
iv. Este modelo incluye 8 variables entre las cuales se encuentra el tipo de propiedad y la ciudad de la propiedad las cuales se ingresan al modelo como una dummy y la interacc ión de bedrooms y bathrooms, adicionalmente se incluyen las variables de la ubicación geográfica.
v. Este modelo incluye 9 variables entre las cuales se encuentra el tipo de propiedad y la ciudad de la propiedad las cuales se ingresan al modelo como una dummy y la interacc ión de bedrooms y bathrooms, adicionalmente se incluyen las variables de la ubicación geográfica y área de terreno.
vi. Este modelo incluye 11 variables entre las cuales se encuentra el tipo de propiedad y la ciudad de la propiedad las cuales se ingresan al modelo como una dummy y la interacción de bedrooms y bathrooms y latitud y longitud, y las área de terreno y construcción
vii. Este modelo incluye 11 variables entre las cuales se encuentra el tipo de propiedad y la ciudad de la propiedad las cuales se ingresan al modelo como una dummy y la interacción de bedrooms y bathrooms y latitud y longitud, área de terreno y un polinomio de grado de tres del área de construcción.

Tabla 3. Promedio del error de predicción

Modelo	Promedio Error de predicción
Constante	0,8633714
1	0,668623
2	0,6345886
3	0,6304
4	0,6205398
5	0,6160614
6	0,6103244
7	0,6103244

Gráfico 1. Promedio el error de predicción



Como se observa en la tabla 3 y el gráfico1, encontramos que de manera general el promedio del error va disminuyendo al incluir más variables, a partir del modelo 5 la disminución es menor y en el modelo 6 y 7 se llega al mismo resultado, los dos modelos son similares, sin embargo el 7 incluye

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

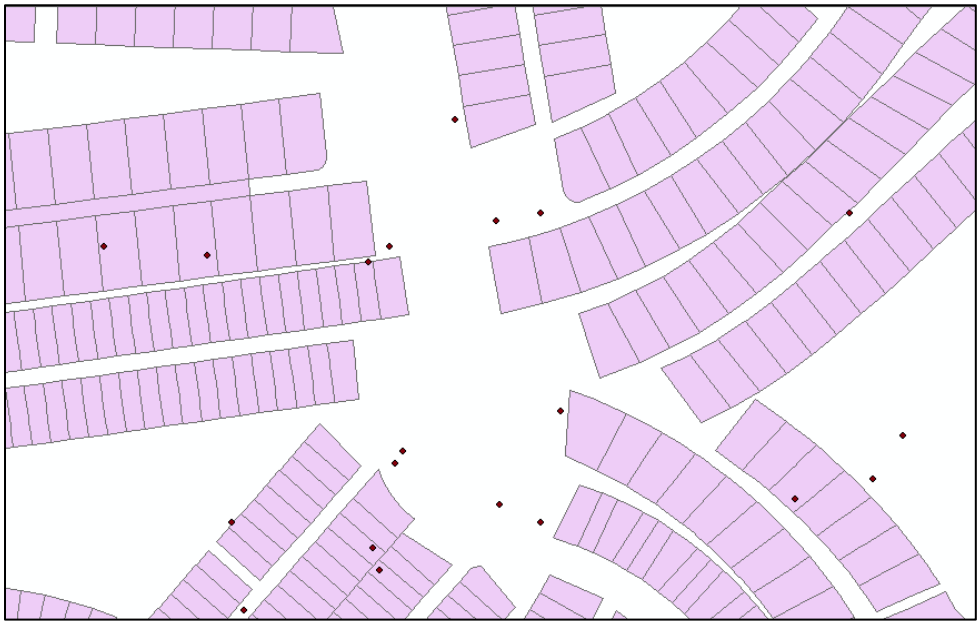
la variables área de construcción transformada con un polinomio de grado 3, por lo tanto, escogeríamos el modelo 6, el cual presenta facilidades en su interpretación.

4. Variables adicionales para mejorar la predicción.

Existen variables muy útiles para predecir los precios de la viviendas o inmuebles en la ciudades, es tas pueden estar relacionadas con la ubicación geográfica de los predios tales como , distancia a zonas de alto riesgo, distancia a estaciones de metro o buses, distancia a parques, entre otras, también podemos incluir variables relaciones con el estado de la vivienda tal como la calificación catastral la cual relaciona el estado de la estructura, pisos, paredes, conservación, entre otras y es determinada por las entidades encargadas de los catastros en el caso de Colombia el Instituto G eográfico Agustín Codazzi.

Para el desarrollo del taller, se encontraron datos de la calificación catastral de Colombia en una geodatabase del IGAC y a partir de la ubicación reportada en la base de datos, se intentó determinar esta información mediante un join espacial, sin embargo, se encuentran algunos problemas en las coordenadas reportadas en la base de datos y problemas de actualización de la información del IGAC, por lo cual quedan gran cantidad de datos vacíos haciendo poco útil su uso.

Gráfico 2. Muestra de ubicación de registros de la base de datos



Finalmente, se incluyeron datos asociados a número de hurtos a personas por municipio en el 2019, cuya fuente es la policita nacional y el total de área con licencias de construcción nuevas para el 2019, esto a partir de la información reportada por el DANE en sus Estadísticas de Licencia s de construcción (ELIC). La primera variable tendría un efecto negativo y la segunda un efecto positivo porque indica que el mercado inmobiliario presenta mayor dinamismo.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Con estas variables se estima un nuevo modelo con los siguientes resultados:

Tabla 4. Modelo con variables de hurto y área con licencia de construcción

	Dependent variable:
	lnprice
rooms	0.029***
	(0.001)
bathrooms	0.350***
	(0.001)
bedrooms	0.051***
	(0.001)
factor(property_type)Casa	-0.053***
	(0.004)
factor(property_type)Finca	0.459***
	(0.031)
factor(property_type)Local comercial	0.799***
	(0.013)
factor(property_type)Lote	1.861***
	(0.016)
factor(property_type)Oficina	0.573***
	(0.012)
factor(property_type)Otro	0.138***
	(0.005)
factor(property_type)Parqueadero	0.880***
	(0.049)
surface_total	0.00003***
	(0.00000)
surface_covered	0.00000
	(0.00000)
lat	-44.145***
	(0.795)
lon	7.037***
	(0.096)
hurto	-0.0004***
	(0.00001)
arealic	0.00001***
	(0.00000)
bathrooms:bedrooms	-0.010***
	(0.0002)

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

lat:lon	-0.581***
	(0.011)
Constant	551.270***
	(7.212)
Observations	148,146
R2	0.492
Adjusted R2	0.492
Residual Std. Error	0.615 (df = 148127)
F Statistic	7,983.028*** (df = 18; 148127)
Note:	*p<0.1; **p<0.05; ***p<0.01

Para este modelo se encuentra un error medio de predicción de **0.6102918,** el cual es menor a los modelos revisados anteriormente.

5. Leave One Out CrossValidation- LOOCV

a. LOOCV

Se realizaron las regresiones del modelo seleccionado, iterando i – th veces, a partir de esto se estimó el error de predicción y se estimo el promedio de los errores de predicción al cuadrado, obteniendo como resultado la estadística LOOCV.

LOOCV: 0.3752055

b. Leverage – LOOCV

De la misma manera se estimo el valor del estadístico de Leverage h, este permite estimar el LOOCV mediante, la siguiente formula:

$$\alpha = \frac{\hat{u}_j}{1 - h_j}$$

En donde, h_j es el estadístico Leverage para cada elemento de la matriz.

$\hat{u}_j$ es el error de predicción del modelo

Los valores estimados por los dos procesos arrojan los mismos resultados: 0.3752055

En la tabla 5, podemos observar una muestra de los valores calculas por de la estadística LOOCV de las dos maneras descritas anteriormente.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tabla 5. Promedio del cuadrado de los residuos.

observación	Valor estimado mediante iteraciones	Valor estimado mediante Leverage
1	-1.204.215.536	-1.204.215.536
2	0.259812533	0.259812533
3	-0.420576958	-0.420576958
4	1.532.345.733	1.532.345.733
5	-0.656782492	-0.656782492
6	-0.070168862	-0.070168862
7	0.016111461	0.016111461
8	0.902412702	0.902412702
9	0.005652242	0.005652242
10	0.132987709	0.132987709

En la siguiente tabla se muestran los valores del estadístico LOOCV para cada uno de los modelos estimados.

Tabla 6. Valores de LOOCV.

Modelo	Promedio Error de predicción	LOOCV
Constante	0,8633714	0,7388391
1	0,668623	0,4564172
2	0,6345886	0,4106008
3	0,6304	0,4106008
4	0,6205398	0,4064475
5	0,6160614	0,3792737
6	0,6103244	0,3759366
7	0,6103244	0.3752588

Como lo muestra la tabla la estadística LOOCV, es mayor en aquellos modelos en los cuales existen menos variables y por lo tanto las observaciones tienen una mayor influencia sobre los modelos, de la mima manera al revisar el valor obtenido para el modelo seleccionado, se encuentra entre los de menor valor lo que nos indica que este es estable y la afectación por la entrada o salida de observaciones es menor.

6. Artisanal MapReduce

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

La tabla 7, nos muestra los coeficientes estimados por los dos métodos por QR descomposición y por una regresión con el comando lm en R, como se observa los valores son igual, sin embargo, en el caso de QR el proceso es mucho más rápido.

Tabla 7. QR decomposition/estimación por lm

Variables	QR	lm
inter	20,77383441788870	20,77383441788
property_type_Casa	-0,76987878250460	-0,76987878250
property_type_Lote	0,35946788017130	0,35946788017
property_type_Otro	-0,60363915178500	-0,60363915170
property_type_Apto	-0,82314130684640	-0,82314130684
property_type_LC	-0,08496571645220	-0,08496571645
rooms	0,06888136094410	0,06888136094
property_type_Ofic	-0,41318798975650	-0,41318798975
property_type_Parq	0,16169724681330	0,16169724681
bathbed	0,00756338471160	0,00756338471
surface_total	0,00003413882180	0,00003413882
surface_covered	0,00000012137270	0,00000012137
latlon	0,00047524131690	0,00047524131
hurto	0,00003825539950	0,00003825539
arealic	-0,00000149452480	-0,00000149452

La tabla 8, nos muestra el procedimiento QR descomposición para cada una de las ciudades

Tabla 8. QR descomposición para 4 particiones por ciudad

Variables	Bogotá	Barranquilla	Medellín	Cali
	QR	QR	QR	QR
inter	-2,27772E+34	-9,14407E+34	-8,43999E+31	-4,27848E+37
property_type_Casa	-0.60141565554	-0.84545134870	-0.92807150638	-0.5410067754314
property_type_Lote	0.77540184636	0.06545767238	0.24208668688	0.4246228610558
property_type_Otro	-0.28992781326	-0.75319388801	-0.79414698716	-0.4321508855530
property_type_Apto	-0.50790103499	-106.948.830.244	-0.98832163927	-0.7131491068330
property_type_LC	0.21786189924	-0.28789323656	-0.03100252599	-0.0007916390216
rooms	0.06551291230	0.11302786482	0.07776074659	0.0579805872669
property_type_Ofic	-0.10742317705	-0.67776113758	-0.41676415568	-0.5868089360648
property_type_Parq	0.23831904333	103.011.647.871	0.72778431821	0.5146272780679
bathbed	0.00580619321	0.00762509009	0.03270546210	0.0085246932964
surface_total	0.00003166385	0.00004856222	0.00002072931	0.0000233729563
surface_covered	0.00005209819	0.00012913010	0.00001003338	0.0000001280863

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

latlon	-0.01003672097	-0.09171549221	0.11356261179	0.0179280549613
hurto	1,31127E+29	8,23106E+30	6,60738E+29	-8,8568E+32
arealic	1,89325E+27	6,5936E+27	-2,09965E+28	6,41092E+31

En esta blablá se observa que existe variables son que son más determinantes de precio de los inm uebles de acuerdo con la ciudad en donde se encuentre.

2.2 Reverend Bayes meets Web Scraping

2. Al momento de utilizar robots.txt para descargar los datos la página no permite realizarlo, por lo tanto se procedió a la descarga por scraping de las tablas de las ciudades y años solicitados.

4. Para realizar el estudio de la base de datos de las visas tipo H1B1 se eliminaron los datos de ciudades homónimas que pertenecen a otros estados y no corresponden a los solicitados de New York (NY), San Francisco (CA), Los Ángeles (CA), Chicago (IL) y Houston(TX). Luego de este filtro, se eliminaron aquellos datos que en el estatus del caso no fuera certificado ("CERTIFIED") por la embajada americana. Por último se eliminaron los datos faltantes de los nombres de las empresas, el cargo al que aplicó la persona y el salario base, dado que sin estos datos se pierde credibilidad en las estadísticas.

A partir de esto se cambió el formato de los datos de las variables "Sumbit Date" y "Start Date" a fecha y a formato numérico las variables "Base Salary". Dado esto se presenta las estadísticas descriptivas de las siguientes variables:

Tabla 9. Estadísticas descriptivas Year

Year							
n	missing	distinct	Info	Mean	Gmd	lowest	highest:
502920	0	5	0.931	2018	1.645	2016	2020

Year	2016	2017	2018	2019	2020
Frequency	77074	73634	78926	83868	189418
Proportion	0.153	0.146	0.157	0.167	0.377

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tabla 10. Estadísticas descriptivas Base Salary

BASE SALARY												
n	missing	distinct	Info	Mean	Gmd	0.05	0.1	0.25	0.5	0.75	0.9	0.95
502920	0	29979	1	100608	45418	50086	58020	70000	90600	120702	155000	180000
lowest:	300	310	315	336	364,	highest:	1800000	1826000	1870000	1900000	4000000	

Tabla 11. Estadísticas descriptivas Location

LOCATION	CHICAGO, IL	HOUSTON, TX	LOS ANGELES, CA	NEW YORK, NY	SAN FRANCISCO, CA
Frequency	93763	93163	45453	160803	109738
Proportion	0.186	0.185	0.090	0.320	0.218

Tabla 12. Estadísticas descriptivas Sumbit Date

SUBMIT DATE						
n	missing	distinct	Info	Mean	Gmd	
502920	0	2057	1	27/11/2017	596.9	
.05	.10	.25	.50	.75	.90	.95
17/07/2015	23/02/2016	28/09/2016	31/01/2018	5/03/2019	23/10/2019	3/03/2020

Tabla 13. Estadísticas descriptivas Start Date

START DATE						
n	missing	distinct	Info	Mean	Gmd	
502920	0	2281	1	27/02/2018	596.9	
.05	.10	.25	.50	.75	.90	.95
18/09/2015	21/05/2016	7/11/2016	29/03/2018	27/06/2019	2/12/2019	16/06/2020

Andrés Bocanegra - 201310223

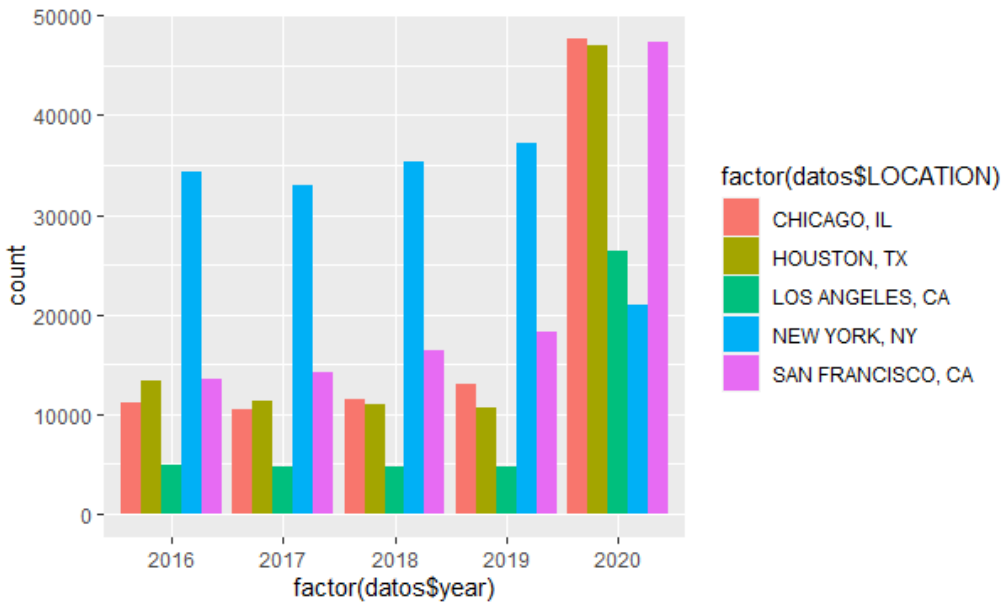
Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tabla 14. Frecuencia entre YEAR y LOCATION

	LOCATION					
YEAR	CHICAGO,	HOUSTON,	LOS ANGELES	NEW YORK	SAN FRANCISCO	TOTAL
2016	11128	13269	4819	34332	13526	77074
	14.40%	17.20%	6.30%	44.50%	17.50%	15.30%
2017	10429	11281	4749	32975	14200	73634
	14.20%	15.30%	6.40%	44.80%	19.30%	14.60%
2018	11550	10925	4679	35389	16383	78926
	14.60%	13.80%	5.90%	44.80%	20.80%	15.70%
2019	12959	10699	4765	37172	18273	83868
	15.50%	12.80%	5.70%	44.30%	21.80%	16.70%
2020	47697	46989	26441	20935	47356	189418
	25.20%	24.80%	14.00%	11.10%	25.00%	37.70%
Total	93763	93163	45453	160803	109738	502920

Gráfico 3. Freciencias



Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

6. Comportamiento empresas por estado

Para el caso de Chicago el comportamiento de las empresas se concentra en E&Y la cual esta en el ranking en todos los años evaluados, le sigue Deloitte Consulting la cual aparece en todos los años a excepción del 2019 y le sigue Tata Consultancy Services con tres apariciones en el Ranking.

Tabla 15. Comportamiento empresas Chicago

CHICAGO, IL	
EMPLOYER 2016	**Número de Aplicaciones**
CAPGEMINI AMERICA INC	718
ERNST & YOUNG US LLP	442
DELOITTE CONSULTING LLP	401
EMPLOYER 2017	**Número de Aplicaciones**
ERNST & YOUNG US LLP	581
TATA CONSULTANCY SERVICES LIMITED	386
DELOITTE CONSULTING LLP	345
EMPLOYER 2018	**Número de Aplicaciones**
DELOITTE CONSULTING LLP	909
ERNST & YOUNG US LLP	323
TATA CONSULTANCY SERVICES LIMITED	308
EMPLOYER 2019	**Número de Aplicaciones**
COGNIZANT TECHNOLOGY SOLUTIONS US CORP	744
ERNST & YOUNG US LLP	597
INFOSYS LIMITED	595
EMPLOYER 2020	**Número de Aplicaciones**
DELOITTE CONSULTING LLP	2635
ERNST & YOUNG US LLP	2310
TATA CONSULTANCY SERVICES LIMITED	1930

En el caso de Houston la empresa que lidera la presencia en el ranking es Infosys Limited la cual aparece en todos los años, le sigue Accenture LLP con 4 apariciones y le siguen E&Y y Capgemini con dos apariciones.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tabla 16. Comportamiento empresas Houston

HOUSTON, TX	
EMPLOYER 2016	**Número de Aplicaciones**
INFOSYS LIMITED	688
CAPGEMINI AMERICA INC	475
ACCENTURE LLP	360
EMPLOYER 2017	**Número de Aplicaciones**
INFOSYS LIMITED	566
ERNST & YOUNG US LLP	350
CAPGEMINI AMERICA INC	237
EMPLOYER 2018	**Número de Aplicaciones**
DELOITTE CONSULTING LLP	416
INFOSYS LIMITED	268
ACCENTURE LLP	225
EMPLOYER 2019	**Número de Aplicaciones**
INFOSYS LIMITED	666
COGNIZANT TECHNOLOGY S	261
DELOITTE CONSULTING LLP	260
EMPLOYER 2020	**Número de Aplicaciones**
ACCENTURE LLP	1462
ERNST & YOUNG US LLP	1303
INFOSYS LIMITED	1215

En el caso de Los Ángeles, E&Y tiene presencia en todos los años, y se destacan la presencia en tres años de la Universidad del Sur de California, Universidad de California e Infosys Limited.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tabla 17. Comportamiento empresas Los Angeles

LOS ANGELES, CA	
EMPLOYER 2016	**Número de Aplicaciones**
ERNST & YOUNG US LLP	552
INFOSYS LIMITED	198
UNIVERSITY OF CALIFORNIA, LA	155
EMPLOYER 2017	**Número de Aplicaciones**
ERNST & YOUNG US LLP	833
INFOSYS LIMITED	159
UNIVERSITY OF SOUTHERN CALIFORNIA, LA	125
EMPLOYER 2018	**Número de Aplicaciones**
ERNST & YOUNG US LLP	252
INFOSYS LIMITED	247
UNIVERSITY OF SOUTHERN CALIFORNIA, LA	163
EMPLOYER 2019	**Número de Aplicaciones**
ERNST & YOUNG US LLP	263
UNIVERSITY OF CALIFORNIA, LA	195
UNIVERSITY OF SOUTHERN CALIFORNIA, LA	178
EMPLOYER 2020	**Número de Aplicaciones**
ERNST & YOUNG US LLP	2267
UNIVERSITY OF CALIFORNIA, LA	883
DELOITTE CONSULTING LLP	832

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

En el caso de New York, E&Y domina en todos los periodos, Goldman Sachs está presente en tres periodos y Google, Capgemini America y JP Morgan en dos periodos.

Tabla 18. Comportamiento empresas New York

NEW YORK, NY	
EMPLOYER 2016	**Número de Aplicaciones**
CAPGEMINI AMERICA INC	1457
ERNST & YOUNG US LLP	1040
JPMORGAN CHASE & CO	520
EMPLOYER 2017	**Número de Aplicaciones**
ERNST & YOUNG US LLP	1049
JPMORGAN CHASE & CO	646
CAPGEMINI AMERICA INC	629
EMPLOYER 2018	**Número de Aplicaciones**
DELOITTE CONSULTING LLP	839
ERNST & YOUNG US LLP	780
GOLDMAN SACHS & CO	717
EMPLOYER 2019	**Número de Aplicaciones**
ERNST & YOUNG US LLP	1554
GOOGLE LLC	906
GOLDMAN SACHS & CO	746
EMPLOYER 2020	**Número de Aplicaciones**
ERNST & YOUNG US LLP	952
GOOGLE LLC	562
GOLDMAN SACHS & CO	440

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

En San Francisco, Salesforcecom Inc se encuentra en todos los años, seguido de Uber con tres presencias y por último E&Y, Infosys System y Deloitte Consulting con dos operaciones.

Tabla 18. Comportamiento empresas San Francisco

SAN FRANCISCO, CA	
EMPLOYER 2016	**Número de Aplicaciones**
INFOSYS LIMITED	557
SALESFORCECOM INC	538
UBER TECHNOLOGIC INC	386
EMPLOYER 2017	**Número de Aplicaciones**
SALESFORCECOM INC	614
UBER TECHNOLOGIC INC	524
ERNST & YOUNG US LLP	372
EMPLOYER 2018	**Número de Aplicaciones**
SALESFORCECOM INC	844
UBER TECHNOLOGIC INC	666
DELOITTE CONSULTING LLP	413
EMPLOYER 2019	**Número de Aplicaciones**
SALESFORCECOM INC	939
UBER TECHNOLOGIC INC	770
INFOSYS LIMITED	363
EMPLOYER 2020	**Número de Aplicaciones**
SALESFORCECOM INC	1786
ERNST & YOUNG US LLP	1602
DELOITTE CONSULTING LLP	1247

Descripción de las compañías:

E&Y: Firma de Servicios Profesionales (Consultoría) que incluyen auditoría, finanzas, contabilidad, manejo de impuestos, asesoría legal y asesoramiento en la gestión de la empresa.

Deloitte: Firma privada de servicios profesionales dedicada a ofrecer consultoría en cinco grandes áreas funcionales: consultoría, impuestos, asesoría jurídica, financiera y auditoría.

Infosys Limited: Empresa multinacional de servicios de tecnologías de la Información y consultoría con sede base en India.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

Tata Consultancy Services: Empresa de servicios de consultoría, de TI y organización en soluciones de negocios. TCS ofrece una cartera integrada de servicios de TI, BPS, infraestructura, ingeniería y servicios de control de calidad.

Accenture: Es una firma global de consultoría y servicios profesionales que ofrece servicios de estrategia, consultoría digital, tecnología y operaciones.

Capgemini: Es una multinacional francesa de consultoría tecnológica, la cual proporciona servicios de TI y es una de las mayores compañías del mundo de consultoría, externalización y servicios profesionales

Universidad de California en Los Ángeles: Fundada en 1919, ofrece 337 programas de grado y postgrado en un amplio rango de especialidades.

Universidad del Sur de California: Fundada a finales del siglo XIX, la cual se destaca por sus altos niveles de investigación y sus programas de humanidades, ciencias sociales y ciencias físicas y naturales.

Goldman Sachs: Es uno de los grupos de banca de inversión y de valores mas grande del mundo fundado en 1869.

Google LLC: Empresa especializada en productos y servicios relacionados con Internet Software, dispositivos electrónicos y otras tecnologías.

JP Morgan: Empresa financiera creada en el año 1799, es una de las firmas de servicios financieros especializada en inversiones bancarias, gestión de activos financieros e inversiones privadas.

Salesforce.com: Es una empresa estadounidense de software bajo demanda, más conocida por producir CRM llamado Sales Cloud.

Uber: Plataforma de tecnología especializada para la movilidad de personas y cosas.

7. New York

I. Para estimar el salario promedio, la varianza del salario y el número de aplicaciones por compañía para el año 2020 en NYC, se hizo un collapse de la base por compañía calculando los estadísticos mencionados.

II. Asuma que las medias X_i provienen del siguiente modelo:

$$X_i|\ \mu_i \sim iid\ N\left(\mu_i, \frac{\hat{\sigma}^2}{n_i}\right), i = 1, \dots\dots, n$$

$$\mu_i \sim N(\mu_0, \tau_0^2), i = 1, \dots\dots, n$$

$$donde\ i\ es\ el\ índice\ de\ las\ compañías$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

La función de densidad de densidad de probabilidad de una muestra es

$$p(X_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2}\frac{(X_i-\mu)^2}{\sigma^2}\right)$$

Siendo cada muestra independiente tenemos:

$$p(X|\mu) = \prod_{i=1}^{n} p(X_i|\mu) = (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2)$$

La prior está dada por

$$p(\mu|\sigma^2) = (2\pi\tau_0^2)^{-1/2}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\tau_0^2}\right)$$

es decir, μ tiene una distribución normal con media μ_0 y varianza τ_0^2. Dado la prior y la función de probabilidad, el posterior se aproxima de la siguiente manera:

$$p(X_i|\mu,\sigma^2)p(\mu|\sigma^2)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\exp\,(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2)\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\tau_0^2}\right)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}\exp\,(-\frac{1}{2}[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2+\frac{1}{\tau_0^2}(\mu-\mu_0)^2])$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}\exp\left(-\frac{q}{2}\right)$$

Donde definimos:

$$q = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2+\frac{1}{\tau_0^2}(\mu-\mu_0)^2$$

Notamos que:

$$\sum_{i=1}^{n}(x_i-\mu)^2$$

$$[A] = \sum_{i=1}^{n}(x_i-\tilde{x}+\tilde{x}-\mu)^2$$

$$= \sum_{i=1}^{n}(x_i-\tilde{x})^2+2(\tilde{x}-\mu)\sum_{i=1}^{n}(x_i-\tilde{x})+n(\tilde{x}-\mu)^2$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$[B] = \sum_{i=1}^{n} (x_i - \tilde{x})^2 + n\,(\tilde{x} - \mu)^2$$

En $[A]$ sumamos y restamos la media muestral

$$\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

En $[B]$ usamos que:

$$\sum_{i=1}^{n} (x_i - \tilde{x}) = \sum_{i=1}^{n} x_i - n\tilde{x} = \sum_{i=1}^{n} x_i - n\frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

Reescribimos:

$$q = \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{\tau_0^2}(\mu - \mu_0)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \tilde{x})^2 + \frac{n}{\sigma^2}(\tilde{x} - \mu)^2 + \frac{1}{\tau_0^2}(\mu - \mu_0)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \tilde{x})^2 + \frac{n}{\sigma^2}\tilde{x}^2 + \frac{n}{\sigma^2}\mu^2 - 2\frac{n}{\sigma^2}\tilde{x}\mu + \frac{1}{\tau_0^2}{\mu_0}^2 - 2\frac{1}{\tau_0^2}\mu_0\mu$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \tilde{x})^2 + \frac{n}{\sigma^2}\tilde{x}^2 + \frac{1}{\tau_0^2}{\mu_0}^2 + \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)\mu^2 - 2\left(\frac{n}{\sigma^2}\tilde{x} + \frac{1}{\tau_0^2}\mu_0\right)\mu$$

$$[A] = \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \tilde{x})^2 + \frac{n}{\sigma^2}\tilde{x}^2 + \frac{1}{\tau_0^2}{\mu_0}^2 - \frac{1}{\tau_n^2}{\mu_n}^2 + \frac{1}{\tau_n^2}(\mu^2 - 2\mu_n\mu + \mu_n^2)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} (x_i - \tilde{x})^2 + \frac{n}{\sigma^2}\tilde{x}^2 + \frac{1}{\tau_0^2}{\mu_0}^2 - \frac{1}{\tau_n^2}{\mu_n}^2 + \frac{1}{\tau_n^2}(\mu - \mu_n)^2$$

En $[A]$ definimos:

$$\tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}$$

$$\mu_n = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}\left[\frac{n}{\sigma^2}\left(\sum_{i=1}^{n} x_i\right) + \frac{1}{\tau_0^2}\mu_0\right]$$

Juntando los resultados Podemos obtener:

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$p(x|\mu)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}q\right)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}\exp\,(-\frac{1}{2}[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\tilde{x})^2+\,\frac{n}{\sigma^2}\tilde{x}^2+\frac{1}{\tau_0^2}\mu_0^2-\frac{1}{\tau_n^2}\mu_n^2])$$

$$*\exp\,(\frac{1}{2\tau_n^2}(\mu-\mu_n)^2)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}(2\pi\tau_n^2)^{\frac{1}{2}}\exp\,(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(x_i-\tilde{x})^2+\,n\tilde{x}^2+\frac{\sigma^2}{\tau_0^2}\mu_0^2-\frac{\sigma^2}{\tau_n^2}\mu_n^2])$$

$$*\,(2\pi\tau_n^2)^{\frac{1}{2}}\exp\,(\frac{1}{2\tau_n^2}(\mu-\mu_n)^2)$$

$$\mathrm{h(x)g}(\mu,x)$$

donde

$$\mathrm{h(x)} = (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau_0^2)^{-\frac{1}{2}}(2\pi\tau_n^2)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\tilde{x})^2+n\tilde{x}^2+\frac{\sigma^2}{\tau_0^2}\mu_0^2-\frac{\sigma^2}{\tau_n^2}\mu_n^2]\right)$$

depende de x pero no de μ, y

$$g(\mu,x) = (2\pi\tau_n^2)^{\frac{1}{2}}\exp\left(-\frac{1}{2\tau_n^2}(\mu-\mu_n)^2\right)$$

es la densidad de una distribución normal con media μ_n y varianza τ_n^2. Por un resultado estándar de la factorización de funciones de densidad de probabilidad, tenemos que:

$$p(\mu|x) = g(\mu,x)$$

$$p(x) = h(x)$$

Así, la distribución posterior $p(\mu_i|x)$ es una distribución normal con media μ_n y varianza τ_n^2.

donde

$$\mu_n = \left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}[\frac{n}{\sigma^2}x_i+\frac{1}{\tau_0^2}\mu_0]$$

$$\mu_n = \frac{\sigma^2}{n\tau_0^2+\sigma^2}\mu_0+\frac{\tau_0^2 n x_i}{n\tau_0^2+\sigma^2}$$

Teniendo que $\hat{\sigma}^2 = \frac{\sigma^2}{n_i}$

$$\mu_n = \frac{\hat{\sigma}^2}{\tau_0^2+\hat{\sigma}^2}\mu_0+\frac{\tau_0^2 x_i}{\tau_0^2+\hat{\sigma}^2}$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$\tau_n^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}$$

Por tanto, la posterior de μ es una distribución normal con media μ_n y varianza τ_n^2. La media de la posterior μ_n es la media ponderada de:

- El dato observado
- la prior μ_0

La distribución marginal de x está dada por:

$$p(x) = (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau^2)^{-\frac{1}{2}}(2\pi\tau_n^2)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\tilde{x})^2 + n\tilde{x}^2 + \frac{\sigma^2}{\tau_0^2}\mu_0^2 - \frac{\sigma^2}{\tau_n^2}\mu_n^2]\right)$$
$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau^2)^{-\frac{1}{2}}(2\pi\tau_n^2)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}q\right)$$

Y hemos definido:

$$q = \sum_{i=1}^{n}(x_i-\tilde{x})^2 + n\tilde{x}^2 + \frac{\sigma^2}{\tau_0^2}\mu_0^2 - \frac{\sigma^2}{\tau_n^2}\mu_n^2$$

Definiendo $v = \frac{\sigma^2}{\tau_0^2}$, Podemos escribir

$$q = \sum_{i=1}^{n}(x_i-\tilde{x})^2 + n\tilde{x}^2 + \frac{\sigma^2}{\tau_0^2}\mu_0^2 - \frac{\sigma^2}{\tau_n^2}\mu_n^2$$
$$= x^T\left(I + \frac{1}{v}ii^T\right)x + n\left(\frac{1}{n}i^Tx\frac{1}{n}i^Tx\right) + v\mu_0^2 - \frac{\sigma^2}{\tau_n^2}\mu_n^2$$
$$= x^T\left(I + \frac{1}{v}ii^T\right)x + \frac{1}{n}x^T(ii)^Tx + v\mu_0^2 - \sigma^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{n}{\sigma^2}\left(\sum_{i=1}^{n}x_i\right) + \frac{1}{\tau_0^2}\mu_0\right]^2$$

$$= x^TIx + v\mu_0^2 - \frac{1}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{n}{\sigma^2}\left(\sum_{i=1}^{n}x_i\right) + \frac{\sigma^2}{\tau_0^2}\mu_0\right]^2$$
$$= x^TIx + v\mu_0^2 - \frac{1}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\sum_{i=1}^{n}x_i + \frac{\sigma^2}{\tau_0^2}\mu_0\right]^2$$

$$= x^TIx + v\mu_0^2 - \frac{1}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\sum_{i=1}^{n}x_i + v\mu_0\right]^2$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$= x^T Ix + v\mu_0^2 - \frac{1}{n+v}\left[\sum_{i=1}^{n} x_i + v\mu_0\right]^2$$

$$= x^T Ix + v\mu_0^2 - \frac{1}{n+v}[x^T(ii^T)x + v^2\mu_0^2 + 2v\mu_0 i^T x]$$

$$= x^T\left(I - \frac{1}{n+v}ii^T\right)x + \left(v - \frac{v^2}{n+v}\right)\frac{1}{n}\mu_0 i^T\mu_0 + 2\frac{v}{n+v}\mu_0 i^T x$$

$$= x^T\left(I - \frac{1}{n+v}ii^T\right)x + \frac{v}{n+v}\mu_0 i^T\mu_0 + 2\frac{v}{n+v}\mu_0 i^T x$$

$$[A] = x^T\left(I - \frac{1}{n+v}ii^T\right)x + \mu_0 i^T\left(\frac{1}{n+v}ii^T\right)i\mu_0 + 2\mu_0 i^T\left(I - \frac{1}{n+v}ii^T\right)x$$

$$= (x - \mu_0 i)^T\left(I - \frac{1}{n+v}ii^T\right)(x - \mu_0 i)$$

$$[B] = (x - \mu_0 i)^T\left(I + \frac{1}{v}ii^T\right)^{-1}(x - \mu_0 i)$$

En $[A]$ usamos el hecho que:

$$i^T\left(I - \frac{1}{n+v}ii^T\right)i$$

$$= i^T i - \frac{1}{n+v}i^T ii^T i$$

$$= \frac{n+v}{n+v}i^T i - \frac{n}{n+v}i^T i$$

$$= \frac{v}{n+v}i^T i$$

y

$$\mu_0 i^T\left(I - \frac{1}{n+v}ii^T\right)x$$

$$= \mu_0 i^T x - \frac{1}{n+v}\mu_0 i^T ii^T x$$

$$= \frac{n+v}{n+v}\mu_0 i^T x - \frac{n}{n+v}\mu_0 i^T x$$

$$= \frac{v}{n+v}\mu_0 i^T x$$

En [B] usamos el hecho que:

$$\left(I - \frac{1}{n+v}ii^T\right)\left(I + \frac{1}{n+v}ii^T\right)$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$= I - \frac{1}{n+v} ii^T + \frac{1}{v} ii^T + \frac{1}{v} ii^T \frac{1}{n+v} ii^T$$

$$= I - \frac{1}{n+v} ii^T + \frac{1}{v} ii^T - \frac{n}{v} \frac{1}{n+v} ii^T$$

$$I + \frac{-v + n + v - n}{(n+v)v}$$
$$= I$$

entonces

$$\left(I - \frac{1}{v} ii^T\right)^{-1} = \left(I - \frac{1}{n+v} ii^T\right)$$

Ahora note que

$$(2\,\pi\sigma^2)^{-\frac{n}{2}} (2\,\pi\tau_0^2)^{-\frac{1}{2}} (2\pi\tau_n^2)^{\frac{1}{2}}$$
$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{\tau_n^2}{\tau^2}\right)^{\frac{1}{2}}$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{1}{\tau_0^2} \frac{\sigma^2 \tau_0^2}{n\tau_0^2 + \sigma^2}\right)^{\frac{1}{2}}$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{1}{n\,\frac{\tau_0^2}{\sigma^2} + 1}\right)^{\frac{1}{2}}$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{1}{\frac{n}{v} + 1}\right)^{\frac{1}{2}}$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{v}{n+v}\right)^{\frac{1}{2}}$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}} \left(\frac{n+v}{v}\right)^{-\frac{1}{2}}$$

$$[A] = (2\,\pi\sigma^2)^{-\frac{n}{2}} \left|\det\left(I + \frac{1}{v} ii^T\right)\right|^{\frac{1}{2}}$$

En [A] usamos el determinate de matrices

$$\det\left(I + \frac{1}{v} ii^T\right) = 1 + \frac{1}{v} i^T i$$
$$= 1 + \frac{n}{v}$$

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

$$= \frac{n+v}{v}$$

Juntando las piezas, temenos:

$$p(x) = (2\,\pi\sigma^2)^{-\frac{n}{2}}\,(2\,\pi\tau^2)^{-\frac{1}{2}}(2\pi\tau_n^2)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}q\right)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\left|\det\left(I+\frac{1}{v}ii^T\right)\right|^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_0 i\right)^T\left(I+\frac{1}{v}ii^T\right)^{-1}(x-\mu_0 i)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}\left|\det\left(\sigma^2 I+\frac{\sigma^2}{v}ii^T\right)\right|^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_0 i\right)^T\left(\sigma^2 I+\frac{\sigma^2}{v}ii^T\right)^{-1}(x-\mu_0 i)$$

$$= (2\,\pi\sigma^2)^{-\frac{n}{2}}|\det(\sigma^2 I+\tau^2 ii^T)|^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_0 i\right)^T(\sigma^2 I+\tau^2 ii^T)^{-1}(x-\mu_0 i)$$

Bajo esta distribución, una muestra de x tiene una media μ_0 y varianza $\hat{\sigma}^2+\tau_0^2$. Entonces:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{X})^2}{n-3} = \hat{\sigma}^2+\tau_0^2$$

$$E(\bar{X}) = \mu_0$$

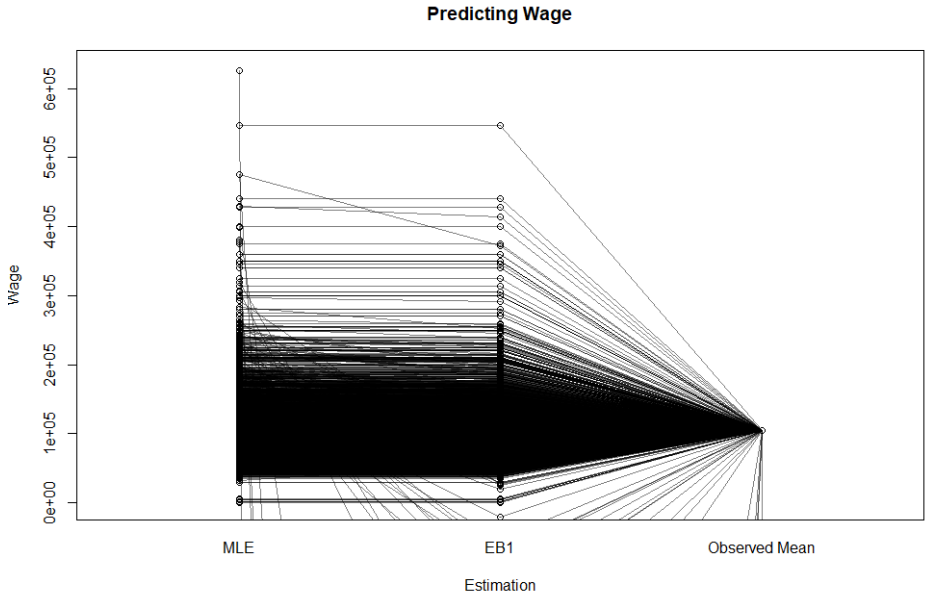
$$\mu_n = \frac{(n-3)\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i-\bar{X})^2}\bar{X} + x_i - \frac{(n-3)\hat{\sigma}^2 x_i}{\sum_{i=1}^{N}(x_i-\bar{X})^2}$$

Con base en la marginal, estimamos $\hat{\sigma}^2+\tau_0^2$ y μ_0 y con esto sacamos la media de la distribución posterior. Como la varianza $\hat{\sigma}^2$ es desconocida, usamos como estimador la varianza del salario de cada solicitud de empleo en cada grupo de compañías, divido por el número de solicitudes. Así entonces, aplicando la anterior fórmula a la base de datos obtuvimos los resultados que se muestran en el siguiente gráfico.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222

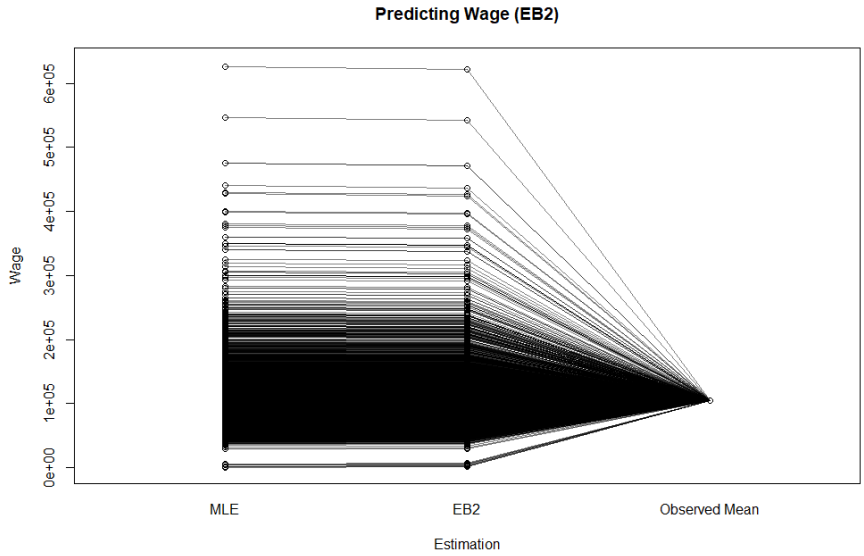


Debido a que la varianza de algunas compañías es demasiado grande, esto hace que el ponderador w sea mayor a 1. Esto hace que sea posible que el estimador de bayes tome valores negativos para algunas firmas (para 18 firmas). Otra consecuencia de haber elegido la varianza al interior de cada firma como estimador de sigma cuadrado es que al existir 3.144 firmas en las cuales solo hubo una solicitud laboral, la varianza es cero. Con varianza cero, el valor de la posterior es igual al observado. Por esta razón en la imagen se puede apreciar que hay muchas compañías para las cuales el MLE es igual al EB1 (estimador bayesiano 1). En vista de estas limitaciones, decidimos cambiar el estimador de la varianza, y escogimos la media de las varianzas de los salarios al interior de las firmas. Estos resultados se muestran en el siguiente gráfico.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222



Como se puede ver, con esta nueva estimación sobre el sigma cuadrado los valores de la media posterior no son menores que cero. Por otro lado, para las firmas que tienen varianza de los salarios de sus aplicaciones igual a cero la estimación de la media posterior es diferente a la de MLE. Sin embargo, el EB2 a diferencia del EB1 tiene los ponderadores sobre el $\bar{X}$ y $x_i$constates lo que hace que en general la tendencia del EB2 hacia el observed mean sea más parsimoniosa que en el EB1. A pesar de que los resultados sean sensibles al supuesto sobre el sigma cuadrado, el método de Empirical Bayes hace que las estimaciones se encojan hacia la media observada.

Referencias:

Taboga, Marco (2017). "Ridge regression", Lectures on probability theory and mathematical statistics, Third edition. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/ridge-regression.

Theobald, C. M. (1974) " Generalizations of mean square error applied to ridge regression", Journal of the Royal Statistical Society, Series B (Methodological), 36, 103-106.

Andrés Bocanegra - 201310223

Carlos Andrés Beltrán- 201012296

Felipe Luque Prieto– 201414222