Taller 1

Julián Gutierrez, Rafael Cano, Mauricio Perea

Octubre 2020

1. Theory Exercises

1.1.

Sea:

$$Y = \rho WY + \varepsilon$$
$$Y = (I - \rho W)^{-1}\varepsilon$$

- Tenemos que:

$$E(Y) = 0$$
$$V(Y) = E(YY') = (I - \rho W)^{-1}E(\varepsilon\varepsilon')(I - \rho W)'^{-1} = \Omega\sigma^2$$

- La maxima verosimilitud es:

$$L(\sigma^2, \rho|Y) = \frac{1}{\sqrt{2\pi}}|\Omega\sigma^2|^{-\frac{1}{2}} \exp -\frac{1}{2\sigma^2}Y'\Omega^{-1}Y$$

- La determinante de la matriz $|\Omega\sigma^2|^{-\frac{1}{2}} = |(I - \rho W)|^{-2}\sigma^{n2}$. - La log verosimilitud es:

$$\log(L(\sigma^2, \rho|Y)) = C - \frac{n}{2}\log(\sigma^2) + \log(|(I - \rho W)|) - \frac{1}{2\sigma^2}Y'\Omega^{-1}Y =$$
$$C - \frac{n}{2}\log(\sigma^2) + \log(|(I - \rho W)|) - \frac{1}{2\sigma^2}((I - \rho W)Y)'((I - \rho W)Y)$$

- Ahora bien, $|I - \rho W| = \prod(1 - \rho w)$ (Ord, 1975).

$$\log(L(\sigma^2, \rho|Y)) =$$
$$C - \frac{n}{2}\log(\sigma^2) + \sum \log(1 - \rho w) - \frac{1}{2\sigma^2}((I - \rho W)Y)'((I - \rho W)Y)$$

- Del proceso de optimización, tenemos:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(Y - \rho WY)'(Y - \rho WY)$$

- Ahora bien, para optener $\hat{\rho}$, se requiere usar métodos numéricos, ya que su valor no es cerrado. Remplazando $\hat{\sigma}^2_{MLE}$ en la log verosimilitud y aplicando métodos numéricos se obtiene $\hat{\rho}$.

Si suponemos que usamos MCO, el valor de $\hat{\rho}$ sería sesgado y diferente al obtenido anteriormente. La razón de que sea sesgado es debido a que existe endogeneidad de la variabla WY:

$$E(WY\varepsilon) = E(W(I-\rho W)\varepsilon\varepsilon') = W(I-\rho W)\sigma^2 \neq 0$$

Sea:

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

- Reescribimos el problema como:

$$Y = Z\alpha + \varepsilon$$

- Donde:

$$Z = \begin{pmatrix} X & WX \end{pmatrix}$$
$$\alpha = \begin{pmatrix} \beta \\ \theta \end{pmatrix}$$

- La maxima verosimilitud es:

$$L(\sigma^2, \alpha | Y, Z) = \frac{1}{\sqrt{2\pi\sigma^{n2}}} \exp -\frac{1}{2\sigma^2}(Y - Z\alpha)'(Y - Z\alpha)$$

- La log verosimilitud es:

$$\log(L(\sigma^2, \alpha | Y, Z)) = C - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - Z\alpha)'(Y - Z\alpha)$$

- Del proceso de optimización, tenemos:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}(Y - Z\alpha)'(Y - Z\alpha)$$
$$\hat{\alpha}_{MLE} = (Z'Z)^{-1}Z'Y$$

- Reescribiendo este problema usando FWL, tendríamos:

$$\hat{\beta}_{MLE} = (X'M_{WX}X)^{-1}X'M_{WX}Y$$
$$\hat{\theta}_{MLE} = ((WX)'M_X(WX))^{-1}(WX)'M_XY$$

Dodne:

$$M_{WX} = I - WX(X'W'WX)^{-1}X'W'$$
$$M_X = I - X(X'X)^{-1}X'$$

Si suponemos que usamos MCO, el valor de $\hat{\beta}$ y $\hat{\theta}$ serían los mismos estimadores que obtuvimos resolviendo MLE. Eso bajo los supuestos MELI de estimadores por MCO.

1.2.

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

$$P(\beta|y,x) = \frac{f(y,x|\beta)P(\beta)}{m(y,x)} =$$
$$f(y|x,\beta)P(\beta)\frac{f(x)}{m(y,x)} = f(y|x,\beta)P(\beta)\frac{1}{P(y|x)} \propto$$
$$f(y|x,\beta)P(\beta)$$

Ahora bien, si una variable θ es proporcional a:

$$\exp -\frac{1}{2}(A\theta^2 + B\theta + C)$$

Entonces podemos decir que $\theta \sim N(m, V)$ con:

$$m = -\frac{1}{2}\frac{B}{A}$$
$$V = \frac{1}{A}$$

- la función de maxima verosimilitud y la probabilidad de β están dadas por:

$$f(y|x,\beta) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}(y - x\beta)^2 =$$
$$(\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp \sum -\frac{1}{2\sigma^2}(y - x\beta)^2$$
$$P(\beta) = \frac{1}{\sqrt{2\pi\tau^2}} \exp -\frac{1}{2\tau^2}\beta^2$$

- La función posterior es entonces:

$$P(\beta|y,x) = f(y|x,\beta)P(\beta) \propto \exp \sum -\frac{1}{2\sigma^2}(y - x\beta)^2 \exp -\frac{1}{2\tau^2}\beta^2 =$$
$$\exp \sum -\frac{1}{2\sigma^2}(y - x\beta)^2 - \frac{1}{2\tau^2}\beta^2$$

- Resolviendo el interior de la exponencial:

$$\sum -\frac{1}{2\sigma^2}(y - x\beta)^2 - \frac{1}{2\tau^2}\beta^2 =$$
$$-\frac{1}{2}(\beta^2(\frac{1}{\sigma^2}\sum x^2 + \frac{1}{\tau^2}+)\beta(-\frac{2}{\sigma^2}\sum yx) + \frac{1}{\sigma^2}\sum y^2)$$

- Por lo tanto, concluimos que $\beta|y,x \sim N(m,V)$ con:

$$m = \frac{\frac{\sum x^2}{\sigma^2}}{\frac{\sum x^2}{\sigma^2}+\frac{1}{\tau^2}}\frac{\sum yx}{\sum x^2} = \frac{\frac{\sum x^2}{\sigma^2}}{\frac{\sum x^2}{\sigma^2}+\frac{1}{\tau^2}}\hat{\beta}_{MCO}$$

La estimación de Ridge, suponiendo $\sum x^2 = 1$, está dada por:

$$\text{mín}\, R(\beta) = \sum (y - x\beta)^2 + \lambda \sum \beta^2$$

El segundo termino puede verse como la minimización del error cuadrado de la distribución prior mín $\sum(\beta - 0)^2$.

- FOC:

$$-2\sum yx + 2\beta + 2\lambda\beta = 0$$

$$\hat{\beta}_{Ridge} = \frac{1}{1+\lambda}\hat{\beta}_{MCO}$$

- Solución de Bayes

$$\hat{\beta}_{Bayes} = \frac{1}{1+\frac{\sigma^2}{\tau^2}}\hat{\beta}_{MCO}$$

La solución en ambos estimadores resulta en un $\hat{\beta}_{MCO}$ encojido a cero. Aún más, se puede observar la siguiente aproximación:

$$\lambda \approx \frac{\sigma^2}{\tau^2}$$

Entonces, λ en ridge es la analogía del ratio entre varianzas del β_{MCO} y de la prior $\frac{\sigma^2}{\tau^2}$ en la estimación Bayesiana. De esto, se puede observar que entre mayor variación del β_{MCO} haya, el parametro que encoje a cero crecerá.

1.3.

Tenemos la función donde los datos datos explicativos han sido centrados en 0 y que buscamos minimizar respecto a los coeficientes:

$$argminR(\beta,\beta_0) = (Y - X\beta - \beta_0\iota)^T(Y - X\beta - \beta_0\iota) + \lambda\beta^T\beta$$

Primero reescribimos los anterior como sumatorias y reemplazamos la X centrada por la X original menos el promedio de la X original:

$$argminR(\beta,\beta_0) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j(x_{ij} - \bar{x}_{ij}))^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Así pues derivamos respecto a los coeficientes:

$$dR(\beta,\beta_0)/d\beta = -2\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})(\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))+2\lambda\sum_{j=1}^{p}\beta_j=0 \quad (1)$$

$$dR(\beta,\beta_0)/d\beta_0 = -2\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))=0 \quad (2)$$

Despejamos Beta sub cero de (2)

$$0 = -2\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))$$

$$0 = -\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))$$

$$0 = -\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}\beta_0+\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij})$$

$$\sum_{i=1}^{N}\beta_0 = \sum_{i=1}^{N}y_i-\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij})$$

Teniendo en cuenta que :

$$\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})=0$$

Dado que $\bar{x}_{ij}$ *es la media de* x_{ij}. *Entonces obtenemos* :

$$\sum_{i=1}^{N}\beta_0 = \sum_{i=1}^{N}y_i$$

$$N*\beta_0 = \sum_{i=1}^{N}y_i$$

$$\beta_0 = \frac{\sum_{i=1}^{N}y_i}{N} \quad (3)$$

$$\beta_0 = \bar{y}$$

Resolvemos (1)

$$-2\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})(\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))+2\lambda\sum_{j=1}^{p}\beta_j=0$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})(\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij}))+\lambda\sum_{j=1}^{p}\beta_j=0$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}y_i+\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}\beta_0+\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij})$$

$$+\lambda\sum_{j=1}^{p}\beta_j=0$$

$$\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j(x_{ij}-\bar{x}_{ij})+\lambda\sum_{j=1}^{p}\beta_j=\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}y_i$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{ij})\sum_{i=1}^{N}\beta_0$$

Convertimos la lanterior expresión a forma matricial y remplazamos por la X centrada para facilitar los cálculos

$$X^TX\beta+\lambda\beta=X^TY-X^T\beta_0$$

Reemplazamos por (3): $\beta_0=\frac{\sum_{i=1}^{N}y_i}{N}$ *pero en forma matricial*: $\beta_0=\frac{Y}{N}$

$$X^TX\beta+\lambda\beta=X^TY-X^T\frac{Y}{N}$$

$$X^TX\beta+\lambda\beta=X^TY-\frac{X^TY}{N}$$

Teniendo en cuenta que:

$\bar{X}=0$

Nos queda:

$$X^TX\beta+\lambda\beta=X^TY \qquad (4)$$

Para terminar, despejamos Beta de (4):

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

Así pues, queda comprobado que los parámetros que minimizan

$$R(\beta,\beta_0)=(Y-X\beta-\beta_0\iota)^T(Y-X\beta-\beta_0\iota)+\lambda\beta^T\beta$$

Son:

$$\beta(X^TX+\lambda I)=X^TY$$
$$\beta_0=\bar{y}$$

1.4.

Suponiendo el siguiente modelo de regresión:

$$Y = X\beta + \epsilon = 0$$

A este modelo le hacemos la siguiente tranformción. Aumentamos la matriz X centrada con p filas adicionales con raiz cuadrada de lambda, y aumentamos Y con ceros. Por lo cual, nos queda el siguiente modelo:

$$Y_* = X_*\beta + \epsilon = 0$$

donde

$$X_* = \begin{pmatrix} X \\ \sqrt{\lambda}I \end{pmatrix}$$
$$Y_* = \begin{pmatrix} Y \\ 0I \end{pmatrix}$$

Si obtenemos los estimadores de este último modelo por MCO obtenemos la siguiente expresión:

$$\hat{\beta} = (X_*^T X_*)^{-1} X_*^T Y_*$$

Así pues, si remplazamos la expresiones de las nuevas matrices en esta última ecuación nos queda:

$$\hat{\beta} = \left(\begin{pmatrix} X \\ \sqrt{\lambda}I \end{pmatrix} \quad \begin{pmatrix} X \\ \sqrt{\lambda}I \end{pmatrix}^T \right)^{-1} \begin{pmatrix} X \\ \sqrt{\lambda}I \end{pmatrix}^T \begin{pmatrix} Y \\ 0I \end{pmatrix}$$

Si resolvemos esta última ecuación nos queda:

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

Esta expresión es igual al estimador Ridge", y por ende, demuestra que el procedimiento realizado al inicio genera los mismos estimadores que por el método Ridge".

El leverage statistic busca mirar qué tanto una observación empuja la estiamción del parámetro beta. Esto está dado por:

$$\hat{\beta}_{-j} - \hat{\beta} = -\frac{1}{1 - h_j}(X'X)^{-1} X_j \hat{U}_j$$

Ahora bien, el link que se puede observar entre ridge y leverage es en la estimación de ese tal $\hat{\beta}_j$ y $\hat{\beta}_{-j}$.

$$\hat{\beta}_{-j} = (X'_{-j} X_{-j}) X_{-j}{}' Y_{-j}$$
$$\hat{\beta} = (X'X) X' Y$$

Donde:

$$X = \begin{pmatrix} X_{-j} \\ X_j \end{pmatrix}$$

$$Y = \begin{pmatrix} Y_{-j} \\ Y_j \end{pmatrix}$$

Ambos estimadores se pueden entender como la influencia sobre el vector de estimadores que puede surgir al incluir una observación más en los datos. En el caso de leverage, esa observación es un dato outlayer y, en el caso de ridge, esa observación se da por aumentar la data con la matriz $\sqrt{\lambda I}$ y con ceros en p nuevas filas. Visto de otro modo, se puede estimar leverage de ridge como:

$$\hat{\beta}_{MCO} - \hat{\beta}_{Ridge}$$

De tal modo que se observe la influencia o qué tanto empuja las estimaciones de los beta a cero dada la penalización λ.

1.5.

$$EL(\beta) = (Y - X\beta)^2 + \lambda_1|\beta| + \lambda_2\beta^2$$

- Suponemos que $\beta = (1+\lambda_2)^{-\frac{1}{2}}\breve{\beta}$:

$$EL(\beta) = (Y - X(1+\lambda_2)^{-\frac{1}{2}}\breve{\beta})^2 + \lambda_1|(1+\lambda_2)^{-\frac{1}{2}}\breve{\beta}| + \lambda_2((1+\lambda_2)^{-\frac{1}{2}}\breve{\beta})^2 =$$
$$(Y - (1+\lambda_2)^{-\frac{1}{2}}X\breve{\beta})^2 + (0 - \lambda_2^{\frac{1}{2}}(1+\lambda_2)^{-\frac{1}{2}}\breve{\beta})^2 + \lambda_1(1+\lambda_2)^{-\frac{1}{2}}|\breve{\beta}|$$

- La anterior ecuación se puede escribir como:

$$EL(\beta) = (\breve{Y} - \breve{X}\breve{\beta})^2 + c\lambda_1|\breve{\beta}|$$

- Donde:

$$\breve{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

$$\breve{X} = (1+\lambda_2)^{-\frac{1}{2}} \begin{pmatrix} X \\ I\lambda_2^{\frac{1}{2}} \end{pmatrix}$$

$$\breve{\beta} = (1+\lambda_2)^{\frac{1}{2}}\beta$$

$$c = (1+\lambda_2)^{-\frac{1}{2}}$$

2. Empirical Problems

2.1.

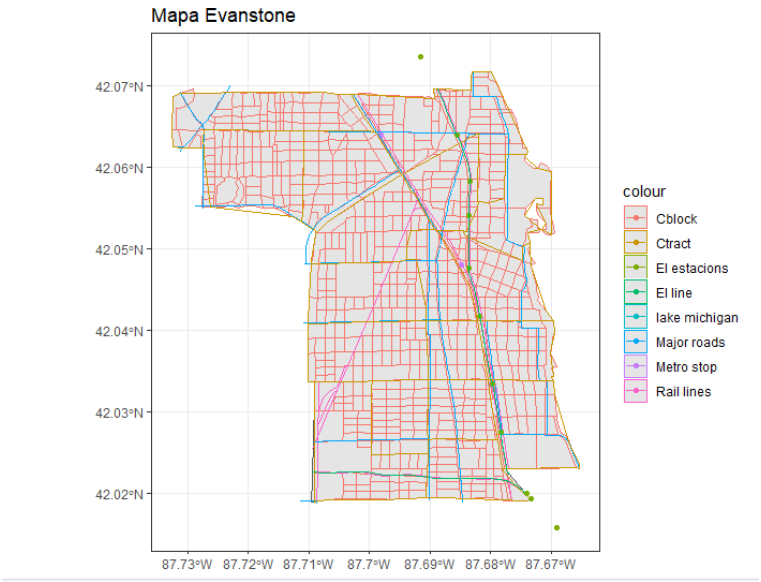

Figura 1

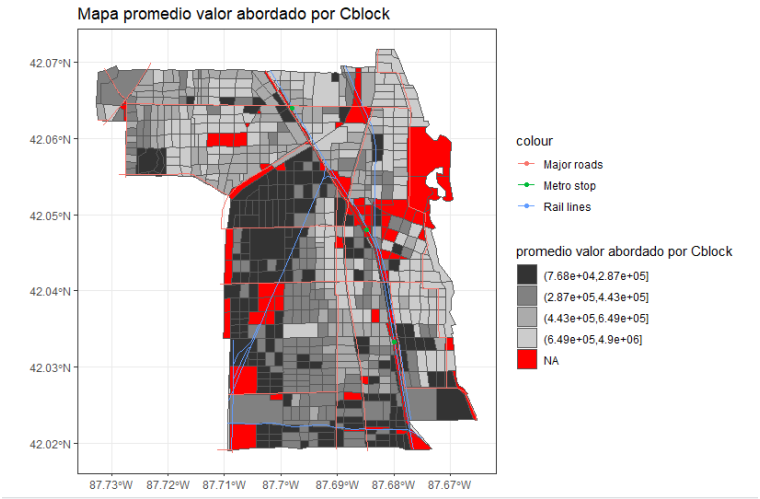

Figura 2

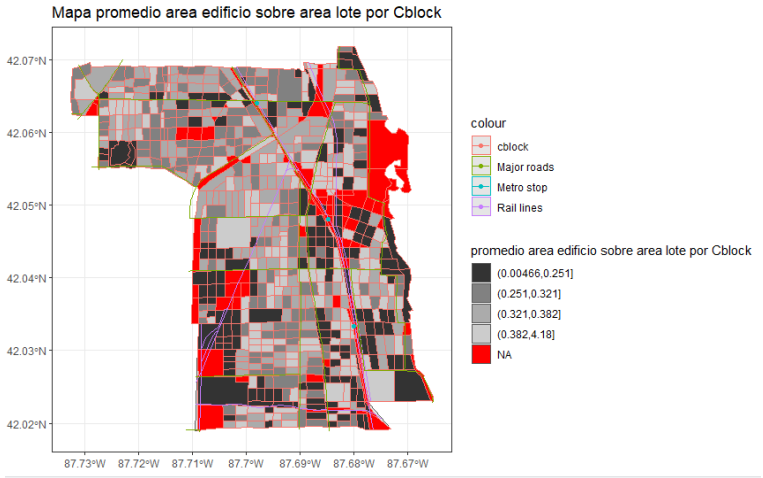

Figura 3

El mapa de Evanston (Figura 2) refleja la media de la valoración de las edificaciones por cblock (parcelas en Evanston hechas por la oficina de asesores del condado). Es dicha ilustración es posible observar que las zonas donde el valor abordado es más alto (color gris claro) coinciden con las zonas donde por donde pasan proyectos de infraestructura de transporte, en especial calles principales y la línea de metro. Es también importante resaltar que las líneas de tren no parecen tener el mismo efecto sobre la valorización, más bien parecen dividir zonas con mayores valuaciones de zonas menos favorecidas. Por otro lado, un factor que está relacionado con las valuaciones más altas es la costa lago Michigan (frontera este del mapa), entre más cerca se esté del lago mayor el valor abordado promedio del cblok.

El mapa de Evanston (Figura 3) refleja la media de la relación del área total de la edificación respecto al área del predio en el que se encuentra. Así pues, se puede observar que los cblock con un valor más bajo de la relación anteriormente mencionada (color negro) están próximos a zonas por donde pasan proyectos de infraestructura de transporte, como calles principales, la línea de metro y líneas de tren. Es decir, que en dichas zonas donde el área total de la edificación respecto al área del predio en el que se encuentra tiende a ser baja son, paradójicamente, zonas con fácil acceso a medios de transporte. Adicionalmente, un factor que está relacionado con las valuaciones más altas es la costa lago Michigan (frontera este del mapa), entre más cerca se esté del lago mayor el promedio de la relación del área total de la edificación respecto al área del predio.

En conclusión, los cblock cercanos a por donde pasan proyectos de infraestructura de transporte y próximos al lago Michigan son más propensos a tener una mayor valoración de las edificaciones y menor relación del área total de la edificación respecto al área del predio en el que se encuentra.

2.2.

Para explicar el valor abordado de las parcelas se se escogieron las siguientes variables: precio de venta, tamaño del garage, cuartos, tipo de recidencia, mínima distancia al lago, mínima distancia al El line, Mínima distancia a alguna carretera principal, mínima distancia a alguna parada del metro y el builduing size to floor size.

	Overall (N=22339)
building_size_to_floor_size	
Mean (SD)	0.279 (0.525)
Median [Min, Max]	0.215 [0.00318, 32.2]
sale_price	
Mean (SD)	421000 (356000)
Median [Min, Max]	333000 [1.00, 4550000]
garage_size	
Mean (SD)	3.84 (2.17)
Median [Min, Max]	3.00 [1.00, 8.00]
rooms	
Mean (SD)	7.04 (2.64)
Median [Min, Max]	6.00 [2.00, 42.0]
type_residence	
Mean (SD)	2.16 (1.09)
Median [Min, Max]	2.00 [1.00, 5.00]
distancia_mínima_lago	
Mean (SD)	1680 (1000)
Median [Min, Max]	1490 [25.0, 4450]
distancia_mínima_major_roads	
Mean (SD)	210 (159)
Median [Min, Max]	175 [13.5, 794]
distancia_mínima_El_line	
Mean (SD)	844 (844)
Median [Min, Max]	515 [11.9, 3440]
distancia_mínima_metrastops	
Mean (SD)	1020 (612)
Median [Min, Max]	912 [52.7, 2850]

Figura 4

Tabla 1	
Tipo de Estimación	**RMSE**
Validation aproach	236521.4
LM k-fold cross validation (k=10)	218065.6
Ridge k-fold cross validation (k=10)	218118.1
Lasso k-fold cross validation (k=10)	218376.5
LM SAR	224116.5

Estimamos la siguiente regresión:

$$AssesedValue = Batfa + Sale_Price + GarageSize + Rooms + TypeResidence + MinDistanceLake + MinimymdsElline1 + MinimymdsMajo + MinimymdsMetroStop \quad (1)$$

Tal como se puede observar en la tabla 1, el modelo que presentó el error cuadrado medio fue la regresión k-fold cross validation usando un k=10 (RMSE=218065.6). Este mostró ser el modelo con mejor ajuste para predecir el assessment value. El modelo de Ridge fue el segundo modelo con mejor ajuste, con un RMSE=218118.1 no muy lejano al LM k-fold CV. Por último, el modelo LM SAR arrojó el RMSE más alto (224116.5).

```
Subset selection object
Call: regsubsets.formula(assesed_value ~ batfa + sale_price + garage_size +
    rooms + type_residence + mindistancelake + minimymds_elline1 +
    minimymds_majo + minimymds_metrastops1, data = pnts, nvmax = 19)
9 Variables  (and intercept)
Selection Algorithm: exhaustive
     batfa sale_price garage_size room       esidence mindistancelake minimymds_elline1
1  ( 1 ) " "   "*"        " "          " "                 " "              " "
2  ( 1 ) " "   "*"        " "          "*"   " "           " "              " "
3  ( 1 ) " "   "*"        " "          "*"   " "           " "              "*"
4  ( 1 ) " "   "*"        " "          "*"   " "           "*"              "*"
5  ( 1 ) " "   "*"        " "          "*"   " "           "*"              "*"
6  ( 1 ) " "   "*"        "*"          "*"   " "           "*"              "*"
7  ( 1 ) " "   "*"        "*"          "*"   "*"           "*"              "*"
8  ( 1 ) "*"   "*"        "*"          "*"   "*"           "*"              "*"
9  ( 1 ) "*"   "*"        "*"          "*"   "*"           "*"              "*"
```

La tabla anterior muestra los resultados de la selección por bestsubsets. De acuerdo con esto, se observan las combinaciones de variables que modelan el mejor ajuste de acuerdo al número de predictores p entre 1 y 19. El mejor modelo es de nueve predictores (p=9), el cual se obseva en la ecuación (1).