Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, febrero de 2018

Tarea 1: Modelo de regresión lineal múltiple

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

El objetivo del estudio que vamos a desarrollar en el presente informe es obtener un modelo lineal adecuado para la variable 'Valor mediano de la casa', a partir de las variables explicativas 'Ingreso mediano', 'Edad mediana de la vivienda', 'Total de habitaciones', 'Total de dormitorios', 'Población' y 'Hogares'.
Previo a la estimación del modelo, se llevo a cabo un profundo análisis exploratorio de datos, para familiarizarnos con las variables y entender el contexto del problema.
Trabajamos con la base de datos 'cadata' que contiene un total de 20640 filas de datos, donde cada fila contiene 9 columnas (variables), estas variables corresponden a un censo realizado en Estados Unidos, más específicamente en California. Obtuvimos una muestra de 500 filas consecutivas de datos, generando un número aleatorio entre 1 y 20640, ese número aleatorio generado fue el 15529, por lo tanto, nuestra base de datos final, contenía las filas de datos desde la 15529 hasta la 16028.
Posterior al análisis exploratorio de datos, se llevo a cabo una selección de variables por distintos métodos (forward, backward y stepwise) con el fin de seleccionar el mejor modelo posible.

2. Análisis exploratorio de datos

Para un correcto análisis exploratorio de datos es necesario definir previamente las variables con las que se está trabajando para no llegar a conclusipnes erróneas. Definimos las variables de la siguiente forma:

- Valor mediano de las casas: Se define como el valor mediano de un conjunto de casas pertenecientes a un bloque.
- Ingreso mediano: Siendo el ingreso una trasformación, definimos el ingreso mediano como la cantidad mediana de dinero anual que entra a una vivienda dentro del bloque.
- Edad mediana: Corresponde a la edad mediana que tienen las viviendas dentro de un bloque.
- Total de habitaciones: Corresponde al total del numero de habitaciones que hay en las viviendas dentro de un bloque.
- Total de dormitorios: Corresponde al total del numero de dormitorios que hay en las viviendas dentro de un bloque.

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

- Población: Corresponde al numero total de habitantes dentro de un bloque.
- Hogares: Corresponde al total de familias que se auto denominan como unidad y habitan en una misma vivienda.
- Longitud y latitud: Medidas de localización donde se encuentra cada bloque.

Obtuvimos las siguientes estadísticas descriptivas de cada variable, excluyendo lógicamente las variables 'Longitud' y 'Latitud', ya que estas, al ser medidas de localización, no tiene sentido sacarles algún estadístico. Esto se hizo con el fin de conocer la dispersión, el rango de valores, la amplitud entre cuartiles y los valores medios y medianos de cada variable.

Tabla 1: Estadisticas descriptivas de las variables

Variable/Estadisticos	Mín.	1st Qu.	Mediana	Media.	3rd Qu.	Max.	Desviacion Est.
Valor mediano de las viviendas	32500	229600	294400	318400	419100	500000	120509.8
Ingreso mediano	0.7054	2.87	3.697	3.905	4.554	15	1.735087
Edad mediana	2	34.75	52	42.2	52	52	14.5956
Total de habitaciones	52	1580	2218	2537	3166	12480	1539.323
Total de dormitorios	13	346	506	583.9	735.2	2747	360.0191
Población	55	836	1168	1303	1608	5640	779.5774
Hogares	13	327.80	469.5	538.8	664.5	2538	328.173

Lo primero que notamos y que es importante a tener en cuenta es la variabilidad de los datos, si nos fijamos, las desviaciones de las variables son muy altas lo que implica que a la hora de estimar parámetros vamos a tener problemas.
También podemos ver en variables como la edad mediana, donde el valor de la mediana es igual al valor máximo, esto quiere decir que este valor se repite el 50 % de las veces, podría ser una muestra desafortunada.
Los rangos de valores que pueden tomar nuestras variables son bastante amplios por lo que puede ser un motivo de la alta variabilidad de estas.
Hay que comentar que el muestreo usado no es el más apropiado ya que aunque es aleatorio no es representativo y deja por fuera a muchos grupos de la población, lo que implica que a la hora de sacar conclusiones no vamos a ser los más certeros posibles.

Se realizaron tres tipos de gráficas para cada variable con el fin de ver si se asemejaban a algún tipo de distribución en especifico, para determinar y ver gráficamente su dispersión y sus puntos atípicos, solo mostraremos los gráficos pertenecientes a una sola variable y concluiremos o contaremos lo que vimos en todas las demás.
Para la variable 'Valor mediano de las viviendas', las gráficas obtenidas fueron las siguientes:

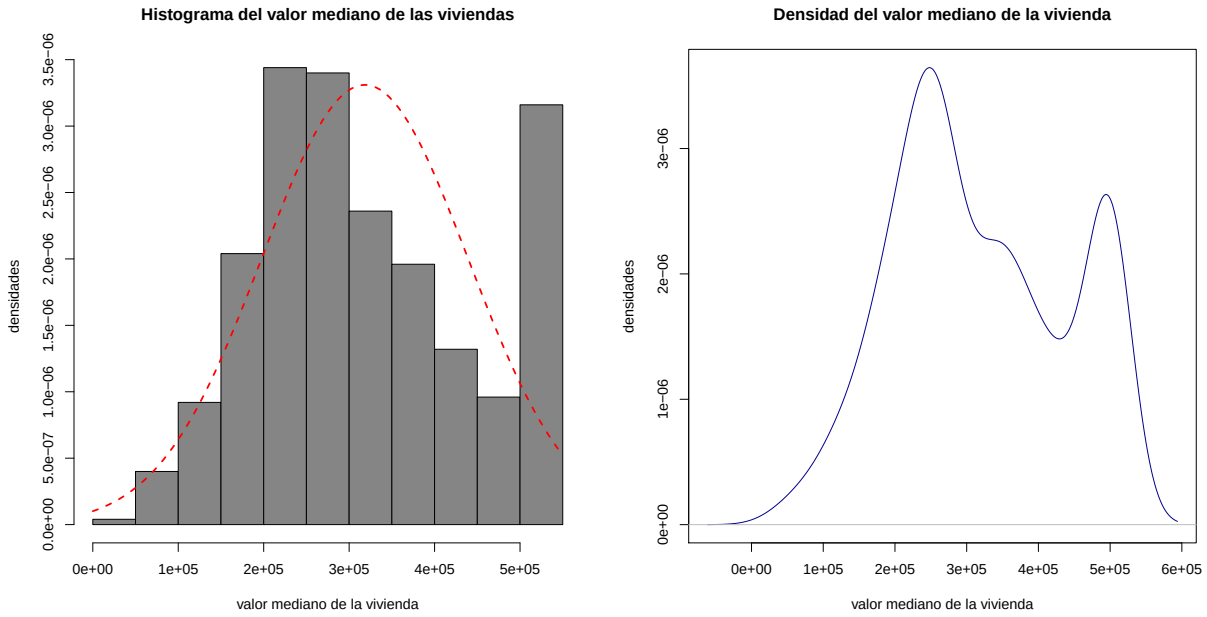

FIGURA 1: Histograma y densidad de la variable 'Valor mediano de las viviendas'

En este gráfico podemos ver claramente que en las densidades hay dos picos, al final y en el medio, esto nos indica que la mayoría de los datos se concentran en estos valores haciendo que a la larga sean complicadas algunas estimaciones y demostrando lo antes dicho que la variabilidad de los datos es considerablemente alta.

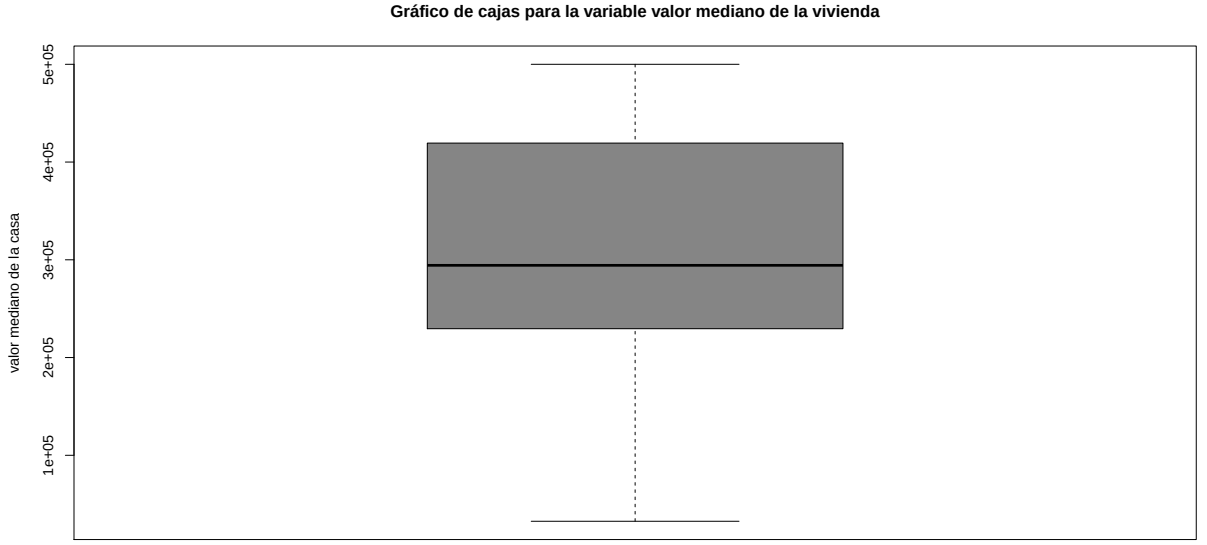

FIGURA 2: Gráfico de caja de la variable 'Valor mediano de las viviendas'

En estos tipos de gráfico (gráfico de caja) podemos ver puntos atípicos, una posible simetría y la dispersión

de los datos. En este caso en particular, podemos ver que el limite inferior o la cola inferior es mucho mas larga que el superior, esto significa que los datos son mas dispersos antes del 25 % que después del 75 %, podemos ver que la linea media de la caja, que corresponde a la mediana, no esta del todo centrada, esto significa que la distribución de los datos no es simétrica. Si notamos, la linea media de la caja, esta muy cerca del cuartil 1 y muy alejada del cuartil 3, esto nos dice que los datos son mucho mas dispersos entre el 50 % y el 75 % que entre el 25 % y el 50 % de lo datos. No se evidencias puntos atípicos, es decir, no hay datos fuera de lo 'común'.

En general con las gráficas concluimos que hay muchos datos atípicos, en algunas variables como lo son: ingreso mediano de las viviendas, el total de habitaciones, total de dormitorios, hogares y población. Además los datos que corresponden a las variables hogares, población, dormitorios y habitaciones tienen densidades muy parecidas y los datos se comportan de una manera muy similar (lo que nos llevo a evaluar posteriormente la posible correlación entre estas) aunque igualmente presentan mucha variabilidad.

Ubicamos las localizaciones en un mapa, con el software **?**, mas específicamente con la librería 'ggmap' (**?**), para tener una idea geográfica de donde provenían nuestros 500 bloques muestreados, los resultados se aprecian en las siguientes dos gráficas:

Figura 3: Mapa de San Diego con los puntos de los bloques muestreados

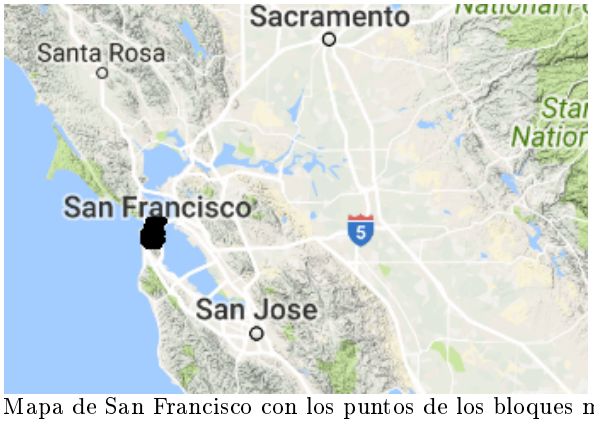

Figura 4: Mapa de San Francisco con los puntos de los bloques muestreados

En las imágenes vemos que obtuvimos datos de solo dos ciudades, San Diego (86 bloques) y San Francisco (414 bloques).

Se llevo a cabo un análisis de correlación entre las variables explicativas del modelo, con el fin de evitar que se presente el problema de la multicolinealidad en el modelo, lo que buscamos mas a fondo, es no meter una variable al modelo que quizá lo que esta explique, ya lo esta explicando la otra. Para este análisis generamos la siguiente matriz de correlaciones, para de allí extraer las combinaciones mas llamativas en cuanto a su correlación y verlas gráficamente.

Tabla 2: Matriz de correlaciones entre las variables explicativas

	Ingreso	Edad	Habitaciones	Dormitorios	Población	Hogares
Ingreso	1	0.017254	0.079028	-0.181368	-0.162578	-0.160498
Edad	0.017254	1	-0.311698	-0.170245	-0.265176	-0.144566
Habitaciones	0.079028	-0.311698	1	0.862072	0.857686	0.865276
Dormitorios	-0.181368	-0.170245	0.862072	1	0.834061	0.988705
Población	-0.162578	-0.265176	0.857686	0.834061	1	0.853918
Hogares	-0.160498	-0.144566	0.865276	0.988705	0.853918	1

En la matriz anterior podemos ver que las primeras dos variables (Ingreso y edad) no tienen una correlación ni entre ellas ni con las demás, por lo cuál, se espera que ambas estén en el modelo final; todo lo contrario sucede con las restantes cuatro variables, todas ellas están correlacionadas en más del 80 % de forma lineal, por lo que esperamos que al menos una de ellas salga del modelo, ya que estaría explicando lo mismo que otra y generando una multicolinealidad en el modelo. Obtuvimos gráficas de puntos de todas las posibles combinaciones entre ellas (6 combinaciones posibles) y generamos el coeficiente de correlación de Spearman (este coeficiente no necesita supuestos acerca de la distribución de las variables, por lo tanto este es mas acertado que el de Pearson), aún así, nos arrojó en su mayoría una correlación mayor del 85 % .

3. Selección de variables

Para obtener el mejor modelo posible, aparte del modelo completo (con todas las variables), utilizamos tres métodos de selección de variables, forward, backward y stepwise por medio de **?** con el paquete 'MASS' (**?**), los modelos obtenidos por estos métodos, los comparamos con respecto al $R^2_{ajustado}$ y el cuadrado medio del error ($CME = \hat{\sigma}^2$). Los modelos obtenidos fueron los siguientes:

- Modelo completo: $Y = 57720.52 + 24261.20X_1 + 3443.94X_2 + 19.09X_3 - 67.72X_4 - 121.66X_5 + 315.92X_6$
- Forward: Este método nos arrojo el mismo modelo completo, es decir, no nos elimino ninguna variable.
- Backward: $Y = 52921.688 + 24923.244X_1 + 3484.164X_2 + 17.588X_3 - 118.652X_5 + 243.265X_6$. Nos elimino la variable 'Total de dormitorios', correspondiente a X_4
- Stepwise: Este método nos arrojo el mismo modelo que el método anterior (Backward).

Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_4 =Total de dormitorios, X_5 =Población, X_6 =Hogares.

4. Modelo final ajustado e interpretación

Como se mencionó anteriormente, se compararon los modelos con respecto al $R^2_{ajustado}$ y al $CME = \hat{\sigma}^2$, los resultados se pueden ver en la siguiente tabla:

Tabla 3: Comparación entre los modelos obtenidos

	Completo	Forward	Backward	Stepwise
$R^2_{ajustado}$	0.537	0.537	0.5371	0.5371
$CME = \hat{\sigma}^2$	6724487923	6724487923	6722361839	6722361839

Según la tabla anterior, podemos concluir que el mejor modelo que logramos obtener para nuestros 500 bloques muestreados sin hacer transformación de variables, fue el generado por el método de selección 'Backward' y 'Stepwise' (ya que su $R^2_{ajustado}$ es minimamente mayor y su $CME = \hat{\sigma}^2$ es muchísimo menor), los cuales nos eliminaron la variable explicativa X_4: Total de dormitorios.

Nuestro modelo final fue: $Y = 52921.688 + 24923.244X_1 + 3484.164X_2 + 17.588X_3 - 118.652X_5 + 243.265X_6$
Donde: Y=Valor mediano de la casa, X_1 =Ingreso mediano, X_2 =Edad mediana de la vivienda, X_3 =Total de habitaciones, X_5 =Población, X_6 =Hogares.

- β_0: Nos indica que el valor mediano de las viviendas dentro del bloque cuando todas las otras variables son 0 es aproximadamente 52921.688 dolares.
- β_1: Por cada unidad que aumente la variable ingreso mediano, el valor mediano de la casa aumenta en 24923.244 dolares.
- β_2: Por cada unidad que aumente la variable edad mediana de la vivienda, el valor mediano de la casa aumenta en 3484.164 dolares.
- β_3: Por cada unidad que aumente la variable total de habitaciones, el valor mediano de la casa aumenta en 17.588 dolares.
- β_5: Por cada unidad que aumente la variable población, el valor mediano de la casa disminuye en 1128.652 dolares.
- β_6: Por cada unidad que aumente la variable hogares, el valor mediano de la casa aumenta en 243.265 dolares.

? ? ? ?