Informe Modelo Lineal Simple para predecir el precio de los apartamentos de estrato 4 con área construida menor a 200 m^2

Camilo Andres Vega Ramírez

Contents

Estimados miembros de la inmobiliaria A&C,

Nos dirigimos a ustedes para presentarles un breve informe que resume los resultados obtenidos a partir del análisis realizado para desarrollar un modelo lineal simple que permita predecir el precio de los apartamentos de estrato 4 con un área construida menor a 200 m^2, a partir de los datos que ustedes nos proporcionaron. Este análisis fue llevado a cabo con el objetivo de proveer información relevante para la toma de decisiones en la compra, venta y valoración de propiedades.

Procedimiento

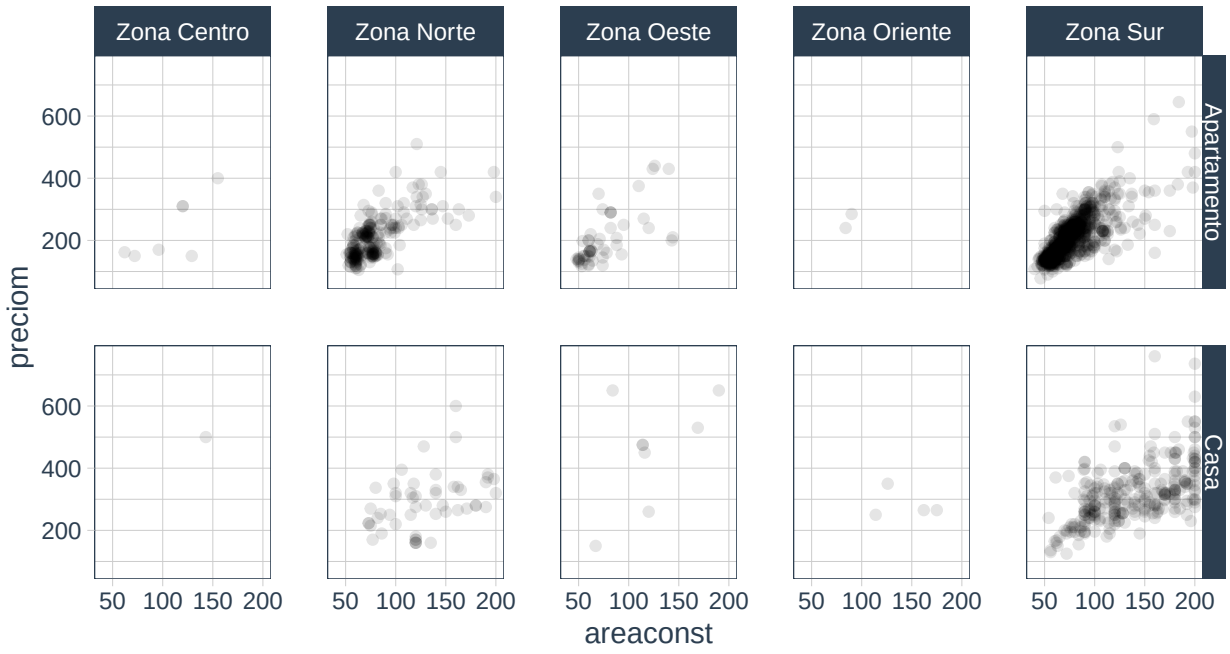
Para el desollo del modelo solicitado se llevaron acabo los siguientes pasos.

Análisis exploratorio inicial.

Para desarrollar el modelo lineal simple, se llevó a cabo un análisis exploratorio de datos inicial, cuyos resultados se encuentran en el anexo "análisis exploratorio.pdf". En dicho análisis se revisó la base de datos "vivienda4" correspondiente a 1706 observaciones de precios de vivienda, áreas, zonas y tipo de vivienda, y se llegó a las siguientes conclusiones.

Se encontró que hay una relación entre las variables precio en millones y área en metros cuadrados, pero se hace evidente la necesidad de realizar transformaciones a estas dos variables para ser utilizadas en un modelo lineal, debido a que muestran asimetría positiva. Es decir, en ambas variables hay presencia de datos con valores altos que pueden sesgar la interpretación del modelo y dificultar la identificación de patrones claros en los datos.

Además, se observó que las variables Zona y Tipo de vivienda también tienen un efecto tanto sobre los precios como sobre las áreas, como se puede apreciar en la siguiente gráfica.



Se nota que la relación entre precio y área es distinta dependiendo de la zona y el tipo de vivienda. También se hizo evidente la falta de balance en los datos para las combinaciones de zona y tipo, como se puede apreciar en las siguientes tablas de conteo y proporciones.

	Apartamento	Casa
Zona Centro	7	1
Zona Norte	237	51
Zona Oeste	52	8
Zona Oriente	2	4
Zona Sur	1065	279

	Apartamento	Casa
Zona Centro	0.0041	0.0006
Zona Norte	0.1389	0.0299
Zona Oeste	0.0305	0.0047
Zona Oriente	0.0012	0.0023
Zona Sur	0.6243	0.1635

Debido a lo anterior, se limitó el estudio y desarrollo del modelo lineal simple solo para apartamentos pertenecientes a la zona sur, ya que es donde se concentra la mayoría de los datos y donde se puede apreciar de mejor forma a nivel visual una relación entre precio y área. Somos conscientes de que esto limita el alcance del modelo a ser solamente aplicado a apartamentos de estrato 4 de la zona sur, pero debido a los datos con los que se cuenta y al requerimiento de que el modelo sea lineal simple, esta es la mejor opción para garantizar un modelo de calidad que sea útil en fase de producción.

Como alternativa, se ve en el análisis exploratorio la creación de un modelo lineal múltiple que incluya también las variables zona y tipo de vivienda, con el objetivo de ampliar el área de aplicación del modelo y tener en cuenta estas variables que, del estudio, se aprecia que tienen un efecto sobre el precio. Sin embargo, este modelo lineal múltiple solo se podría aplicar a casas y apartamentos de la zona norte y sur, ya que las demás zonas cuentan con un número limitado de observaciones que dificultan poder inferir sobre las mismas. Dicho modelo lineal múltiple también se llevó a cabo y se incluye dentro de los resultados del presente informe.

En resumen, del análisis exploratorio se concluye que se debe crear un modelo lineal simple de la relación precio-área solo para apartamentos de la zona sur, y adicionalmente, un modelo lineal múltiple de la relación precio-área tanto para apartamentos como para casas de las zonas norte y sur. Para ambos modelos, se sugiere realizar transformación de datos para los precios y áreas con el fin de reducir las asimetrías de sus distribuciones, así como posterior a estas transformaciones, retirar valores extremos.

Creación modelo lineal simple.

El procedimiento para construir el modelo lineal simple se detalla en el documento "Modelo-Lineal-Simple.pdf". En primer lugar, se llevó a cabo una transformación y filtrado de los datos según lo encontrado en el análisis exploratorio. A continuación, se probaron diversos modelos combinando datos transformados y no transformados con el objetivo de maximizar el coeficiente de determinación R^2 y ajustarce a los supuestos de regresión lineal de los residuales, para que el modelo resultante pudiera explicar la mayor cantidad posible de la variabilidad de los precios. Tras la evaluación, se concluyó que el mejor modelo es aquel que utiliza las transformaciones de Box Cox tanto para los precios como para las áreas.

Finalmente, basándonos en el modelo anteriormente mencionado, se dividieron los datos en un conjunto de prueba y entrenamiento y se sometieron a diferentes motores de regresión lineal. A continuación, se aplicó un proceso de validación cruzada para obtener un modelo capaz de predecir datos nunca antes vistos. Este procedimiento redujo el valor de R^2, algo que se esperaba debido a que la validación cruzada implica una mayor generalización del modelo, pero que se compensa al hacer al modelo más adecuado para la predicción de nuevos datos. Como resultado final, se obtuvo un modelo lineal simple basado en el motor de regresión stan, el cual usa estimación Bayesiana, con un valor de R^2 de 58.94%. Los detalles del modelo se analizarán en profundidad en la sección de resultados.

Creación modelo lineal Multiple.

El proceso de construcción del modelo lineal múltiple se detalla en el documento "Modelo-Lineal-Multiple.pdf". En primer lugar, se llevó a cabo una transformación y filtrado de los datos según lo encontrado en el análisis exploratorio. Basándonos en lo visto en el modelo lineal simple, se partió de los datos de precio y área transformados, el primero con una transformación de Box Cox y el segundo con una transformación de Yeo-Johnson. Se añadieron a estas variables las variables tipo y zona para la construcción del modelo múltiple. A continuación, se evaluó la variable precio en función de las variables área, tipo y zona, así como las interacciones entre ellas. Dependiendo de los resultados de los valores p de los coeficientes, se redujo la complejidad del modelo eliminando aquellas interacciones entre variables con coeficientes no significativos. Esto llevó a un modelo del precio en función del área, tipo, zona e interacción entre área y tipo.

Finalmente, basándonos en el modelo anteriormente mencionado, se dividieron los datos en un conjunto de prueba y entrenamiento y se sometieron a diferentes motores de regresión lineal. A continuación, se aplicó un proceso de validación cruzada para obtener un modelo capaz de predecir datos nunca antes vistos. Este procedimiento redujo el valor de R^2, algo que se esperaba debido a que la validación cruzada implica una mayor generalización del modelo, pero que se compensa al hacer al modelo más adecuado para la predicción de nuevos datos. Como resultado final, se obtuvo un modelo lineal múltiple basado en el motor de regresión Brulee, el cual se basa en herramientas de deep learning, con un valor de R^2 de 63.84%. Los detalles del modelo se analizarán en profundidad en la sección de resultados.

Cabe resaltar que el modelo lineal múltiple presenta un mejor R^2 y un mejor desempeño en las evaluaciones de los supuestos de los residuos, tal como se puede apreciar en los respectivos documentos "Modelo-Lineal-Simple.pdf" y "Modelo-Lineal-Multiple.pdf".

Resultados.

A continuación, se describen los modelos lineales simples y múltiples finales desarrollados en los estudios. Se explicará su fórmula matemática, representación gráfica, interpretación y aplicación correspondiente.

Resultado Modelo Lineal Simple.

El modelo lineal simple creado para predecir el valor de apartamentos de estrato cuatro en la zona sur se expresa mediante la siguiente fórmula:

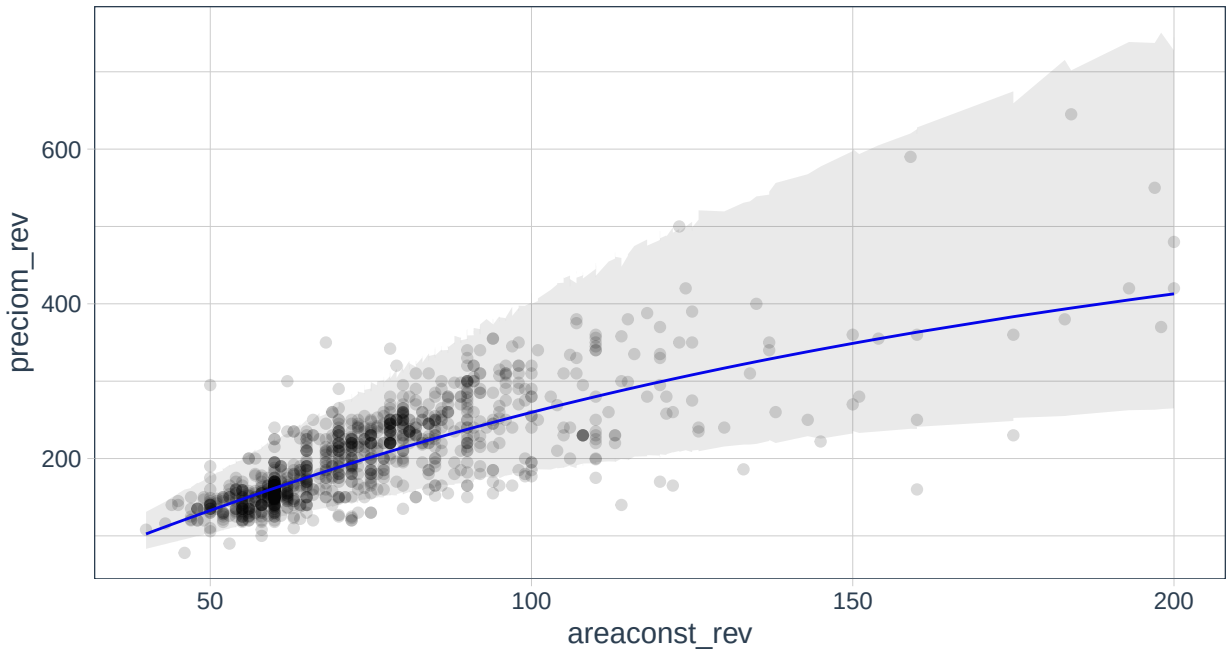
$$\hat{y} = \left(\left(0.0065604 + 0.7787920 \cdot \left(\frac{x^{-0.9999576} - 1}{-0.9999576} \right) \right) \cdot (-0.567132) + 1 \right)^{-1/0.567132}$$

Donde:

- $\hat{y}$ es el precio en millones predicho para apartamentos de estrato cuatro de la zona sur.
- x es el área en metros cuadrados del apartamentos de estrato cuatro de la zona sur.

Es importante mencionar que la fórmula ya tiene en cuenta la transformación de Box-Cox aplicada al área, así como la reversión de dicha transformación para el precio. Esto permite que los resultados se expresen en la escala original de los datos.

La siguiente es la representación gráfica del modelo aplicado a las 1344 observaciones de apartamentos de estrato 4 en la zona sur:



Con el objetivo de tener una idea acerca de las predicciones del modelo, se presenta una tabla que contiene los valores predichos junto con sus intervalos de confianza (95%) para áreas de $50m^2$, $75m^2$, $100m^2$, $150m^2$ y $200m^2$ de apartamentos de estrato cuatro en la zona sur.

Área	Precio Predicho	Rango Inferiror	Rango Superior
50	132.8274	103.6903	178.3814
75	201.6537	148.1501	290.7226
100	259.6624	183.9330	399.3537
150	348.8482	234.0549	586.5988
200	412.8764	269.3719	743.4498

Como se puede observar, el modelo predice un comportamiento logarítmico de los datos, lo que significa que el efecto del área construida sobre el precio es positivo, pero su influencia disminuye gradualmente a medida que el área aumenta. Esto también afecta los rangos de predicción del modelo, que son cada vez más amplios a medida que el área de la vivienda aumenta.

Como se mencionó anteriormente, el valor de R^2 del modelo es del 0.5894, lo que indica que el modelo puede explicar el 58.94% de la variabilidad en los precios de las viviendas en función de su área. Por lo tanto, se puede concluir que el modelo tiene una habilidad moderada/baja para la predicción y que se necesitarán tener en cuenta variables adicionales así como la intervención de un experto en vivienda para interpretar sus resultados.

Se adjuntan a este informe los archivos binarios en formato .RDS “area_box_cox.rds”, “precio_box_cox.rds” y “modelo_lineal_simple.rds”, que podrán ser utilizados en futuras implementaciones de automatización del modelo.

Resultado Modelo Lineal Multiple.

El modelo lineal simple creado para predecir el valor de apartamentos y casas de estrato cuatro en las zonas sur y norte se expresa mediante la siguiente fórmula:

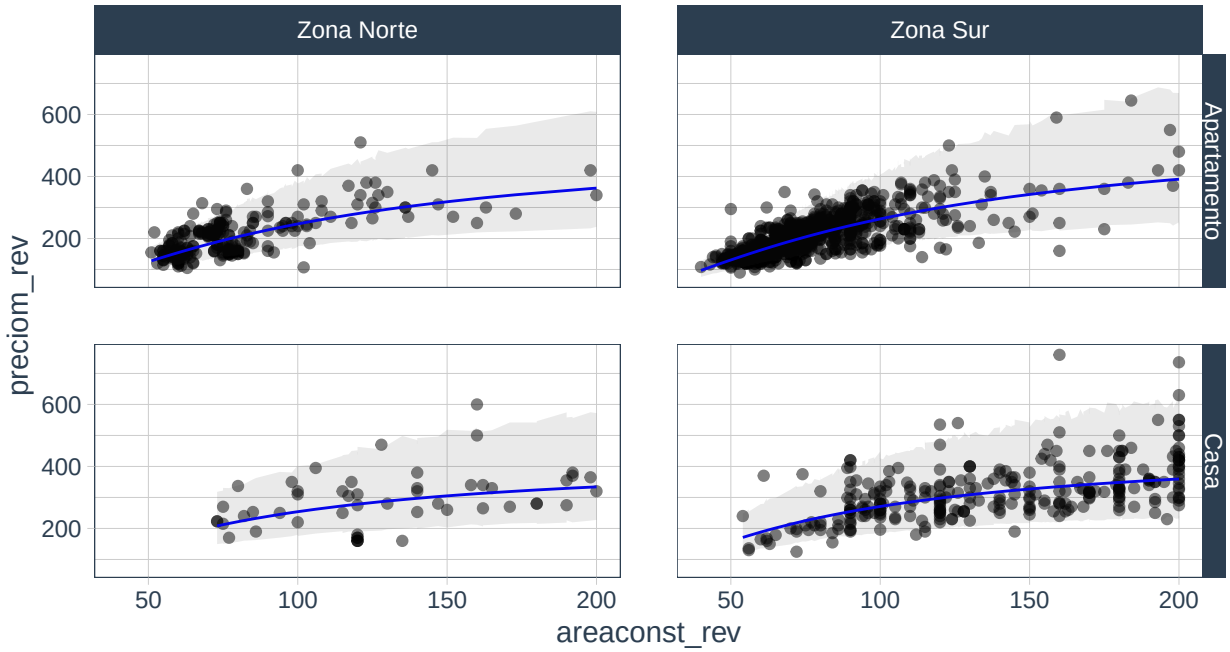
$$\hat{y} = \left(\left(-0.1226 + 0.848 \left(\frac{((x_1 + 1)^{-1.2272}) - 1}{-1.2272} \right) + 0.1987 x_2 + 0.0121 x_3 - 0.2354 x_1 x_2 \right) \cdot (-0.4822) + 1 \right)^{-\frac{1}{0.4822}}$$

Donde:

- $\hat{y}$ es el precio en millones predicho para la vivienda.
- x_1 es el área en metros cuadrados de la vivienda.
- x_2 es `0` si la vivienda es apartamento o `1` si la vivienda es una casa.
- x_3 es `0` si la vivienda es de la zona norte o `1` si la vivienda de la zona sur

Es importante mencionar que la fórmula ya tiene en cuenta la transformación de Yeo-Johnson aplicada al área, así como la reversión de la trasnformación Box Cox para el precio. Esto permite que los resultados se expresen en la escala original de los datos.

La siguiente es la representación gráfica del modelo aplicado a las 1632 observaciones de viviendas de estrato 4 en las zonas sur y norte:



Con el objetivo de tener una idea acerca de las predicciones del modelo, se presenta una tabla que contiene los valores predichos junto con sus intervalos de confianza (95%) para áreas de $50m^2$, $75m^2$, $100m^2$, $150m^2$ y $200m^2$ en todas las combinaciones de apartamento, casa, para la zona norte y sur, de estrato 4.

Área	tipo	zona	Precio Predicho	Rango Inferior	Rango Superior
50	Apartamento	Zona Norte	124.9432	96.06598	170.6114
75	Apartamento	Zona Norte	193.8948	140.21813	285.1075
100	Apartamento	Zona Norte	246.9602	174.93821	384.2688
150	Apartamento	Zona Norte	318.4655	206.65170	519.9193
200	Apartamento	Zona Norte	362.4841	237.92747	612.7785
50	Casa	Zona Norte	151.0702	111.02822	219.8004

Área	tipo	zona	Precio Predicho	Rango Inferior	Rango Superior
75	Casa	Zona Norte	212.2277	147.36826	315.2464
100	Casa	Zona Norte	254.0515	176.78322	395.7507
150	Casa	Zona Norte	305.1029	208.25581	488.4456
200	Casa	Zona Norte	334.0935	223.39854	564.9571
50	Apartamento	Zona Sur	130.7331	99.40005	176.7818
75	Apartamento	Zona Sur	205.0906	148.77376	305.9302
100	Apartamento	Zona Sur	263.0687	181.12007	415.1216
150	Apartamento	Zona Sur	342.0949	225.52959	599.2593
200	Apartamento	Zona Sur	391.2126	247.59156	689.6565
50	Casa	Zona Sur	158.7672	114.25817	232.6113
75	Casa	Zona Sur	225.0518	156.09460	349.0665
100	Casa	Zona Sur	270.8613	189.63989	426.6907
150	Casa	Zona Sur	327.2531	215.04095	552.5314
200	Casa	Zona Sur	359.4941	236.68636	617.2778

Como se puede observar, el modelo predice un comportamiento logarítmico de los datos. Esto significa que el efecto del área construida sobre el precio es positivo, pero su influencia disminuye gradualmente a medida que el área aumenta. Además, esto afecta los rangos de predicción del modelo, que se vuelven cada vez más amplios a medida que el área de la vivienda aumenta.

Adicionalmente, se puede apreciar que el modelo tiene un efecto en las predicciones donde las casas tienen mayores predicciones de precios que los apartamentos. Asimismo, las predicciones de precios de las viviendas en la zona sur son más altas que las de la zona norte. Por último, a medida que crecen los metros cuadrados de las casas, estas presentan una disminución más acelerada en su efecto sobre los precios en comparación con la disminución en el aumento del metro cuadrado de los apartamentos.

Como se mencionó anteriormente, el valor de R^2 del modelo es de 0.6384, lo que indica que el modelo puede explicar el 63.84% de la variabilidad en los precios de las viviendas en función de su área. Por lo tanto, se puede concluir que el modelo tiene una habilidad moderada para la predicción, y que se necesitarán tener en cuenta variables adicionales, así como la intervención de un experto en vivienda, para interpretar sus resultados.

Se adjuntan a este informe los archivos binarios en formato .RDS "area_yeo_johnson_multiple.rds", "precio_box_cox_multiple.rds" y "modelo_lineal_multiple.rds". Estos archivos podrán ser utilizados en futuras implementaciones de automatización del modelo.

Recomendaciones Finales

Es importante tener en cuenta que los datos suministrados no contienen información importante para calcular el precio de una vivienda, como el número de habitaciones, número de baños, antigüedad del inmueble, acabados del inmueble, zonas comunes, ciclos económicos, inflación, entre otros. Estos factores pueden explicar la variabilidad en los precios que los modelos suministrados no pueden explicar. Por lo tanto, es de suma importancia que al aplicar los modelos suministrados, siempre se haga con el acompañamiento de un experto en vivienda de la inmobiliaria, quien, con ayuda de los resultados del modelo, pueda dar luces respecto a cuál es el precio más probable, dado el rango suministrado por los modelos.

Recordamos que el modelo simple está acotado solo para apartamentos de la zona sur y el modelo múltiple para apartamentos y casas de las zonas norte y sur. Otra recomendación importante es la temporalidad de los datos. Es necesario tener en cuenta que los resultados de las predicciones serán de precios en el mismo año de los datos originales. Por lo tanto, será necesario ajustar los precios dependiendo de la inflación, la demanda de vivienda y otras variables para llevar los precios a valores de años posteriores.

Por último, se recomienda en lo posible repetir el estudio incorporando el mayor número de variables mencionadas en este apartado, así como las recomendadas por expertos. También es recomendable considerar la utilización de algoritmos y motores de machine learning avanzados para obtener el mejor modelo predictivo posible.

Lista de anexos

- ***"analisis_exploratorio.pdf":*** Documento que contiene el estudio exploratorio inicial de los datos.
- ***"Modelo-Lineal-Simple.pdf":*** Documento que explica el proceso de creación del modelo lineal simple.
- ***"Modelo-Lineal-Multiple.pdf":*** Documento que explica el proceso de creación del modelo lineal múltiple.
- ***"area_box_cox.rds":*** Archivo binario para aplicar la transformación y reversión de transformación a los datos de área para el modelo simple.
- ***"area_yeo_johnson_multiple.rds":*** Archivo binario para aplicar la transformación y reversión de transformación a los datos de área para el modelo múltiple.
- ***"precio_box_cox.rds":*** Archivo binario para aplicar la transformación y reversión de transformación a los datos de precio para el modelo simple.
- ***"precio_box_cox.rds_multiple":*** Archivo binario para aplicar la transformación y reversión de transformación a los datos de precio para el modelo múltiple.
- ***"modelo_lineal_simple.rds":*** Archivo binario en el motor Stan que contiene el modelo lineal simple para poder ser implementado en R.
- ***"modelo_lineal_multiple.rds":*** Archivo binario en el motor Brulee que contiene el modelo lineal múltiple para poder ser implementado en R.