Prediciendo el precio de una casa por unidad de área

Jeremías Mora Rubio

1. Resumen

Se presentará un análisis en el cual se utilizará constantemente el cálculo de modelos para poder predecir una variable en especial, la cual es Y house price of unit área. Al momento de determinar los modelos, éstos se compararán utilizando algunos valores, específicamente R^2, R^2 ajustado, AIC y BIC, que nos permitirán definir el mejor.

2. Introducción

En la actualidad, hay un gran número de construcciones de viviendas a lo largo de nuestro país y en las comunas en donde uno vive, cosa que se puede apreciar, generalmente con salir de nuestros hogares y caminar un poco. Uno suele preguntarse a cerca de la razón por la cuál se están construyendo tantas viviendas, pero la respuesta es clara y es que la población está aumentando considerablemente.

Pero la pregunta que también puede surgir, es la razón de las diferencias en los precios y las variables que influyen en ella. Ya que de cierta forma, la diferencia antes mencionada, puede permitir determinar valores justos para vender y/o comprar una vivienda y saber determinar si el precio justo. Al poder tener en cuenta las variables que se suelen considerar para determinar los precios, va a permitir estar al tanto de los posibles valores actuales y saber si es el precio que corresponde.

El informe tiene como objetivo, responder, cuáles son las principales variables que influyen en los precios de las viviendas y además, proponer un modelo múltiple para estimar los precios de por unidad de área (por metro cuadrado). Para responder estas preguntas, se analizará una base de datos históricos del mercado de bienes raíces que se recogen en New Taipei City, Taiwán para definir las variables que influyen en los precios, y dar a conocer la importancia que tiene cada una de ellas en el buscar un modelo posible para estimar los precios.

Es por ello, se presentará un marco teórico, donde se explicarán los conceptos que se deben conocer para poder comprender el análisis desarrollado. Además, se mostrará partes importantes del procedimiento realizado para comprender las variables, junto a algunos gráficos y tablas para poder volver mas ilustrativos los desarrollos obtenidos.

3. Metodología

3.1. Base de datos

Para poder realizar un análisis, primeramente debemos comprender la base de datos que se trabajará. Ésta se encuentra en la siguiente página: `https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set` que es de UC Irvine Machine Learning Repository. Podemos notar que está compuesta por las siguientes variables:

Tabla 1: Explicación base de dato

Variable	**Tipo de variable**	**Descripción**
No	Numeric	Número de observaciones
X1 transaction date	Integer	Fecha de la transacción
X2 house age	Integer	Años de la casa
X3 distance to the nearest MRT station	Integer	Distancia a la estación de metro mas cercana
X4 number of convenience stores	Numeric	Número de tiendas de conveniencias cerca
X5 latitude	Integer	Coordenada geográfica, latitud
X6 longitud	Integer	Coordenada geográfica, longitud
Y house price of unit area	Integer	Precio de la casa por unidad de área

Como vamos a evaluar el tema para algo mas general, no consideramos relevantes las variables X5 latitude y X6 longitud, ya que son lugares específicos de la New Taipei City, que también están completamente relacionados con las otras variables X3 distance to the nearest MRT station y X4 number of convenience stores, que son mas generales, razón por la cual, no consideraremos X5 ni X6 en nuestro análisis posterior.

También se debe aclarar, que la variable No sólo es el número correspondiente a la observación, cosa que no aporta nada a nuestra investigación, entonces también se elimina.

Además, aclarar que en cada momento se considerará:

X1 = X1 transaction date.

X2 = X2 house age.

X3 = X3 distance to the nearest MRT station.

X4 = X4 number of convenience stores.

Y = Y house price of unit area, siendo ésta, nuestra variable que buscaremos estimar. Notar que está 10000 New Taiwan Dollar/3.3 metros cuadrados, el cual no se cambiará, a pesar de estar en unidad de medida de otro país.

3.2. Marco Teórico

Para iniciar con el análisis de la base datos entregada, debemos tener claros una serie de conceptos, para comprender lo que se está realizando a lo largo de este proyecto.

El principal, es modelo estadístico, el cual es una forma matemática de estimar un valor mediante una serie de supuestos y valores, los cuales son determinados mediante diversos procedimientos. Existen diversos modelos, pero los principales son el modelo lineal, exponencial y logarítmico, que serán con los que trabajaremos.

La correlación nos permite poder comparar la relación que se observa entre 2 variables, la cual se evalúa entre -1 y 1, y que al tener una correlación muy cercana a 0, significa que hay muy poca relación entre ambas, mientras que cuando el valor es mas cercano a -1 o 1, se puede decir que se relacionan entre sí.

Como queremos crear una forma de poder estimar los posibles valores de los precios de las casas, deberemos crear unos modelos y que determinar que tan bien nos permiten predecir estos valores, lo que sabremos mediante diversas formas de comprar a través de ciertos valores.

Dos de ellos son, R^2 y R^2 Ajustado, son dos valores que permiten poder ver que tan bien puede nuestro modelo representar el comportamiento que tienen los datos, este valor varía entre 0 y 1, y la forma de poder determinar que modelo es mejor que otro, es viendo cual presenta un mayor R^2 y R^2 Ajustado.

Otros valores que se utilizan son AIC y BIC, que viene siendo un tipo de estimación de la información perdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos, entonces se suele elegir el modelo que presenta un menor AIC y BIC.

Para poder realizar un correcto análisis, nuestra intención es realizar una serie de modelos por cada variable que consideremos importantes en nuestro análisis y las compararemos mediante diversos métodos, para poder así determinar un modelo compuesto por varias variables que permita estimar bien lo que deseamos.

4. Resultados

Primeramente analizaremos las correlaciones que hay entre cada una de las variables, para determinar la importancia de cada una en un posible modelo.

Figura 1: Correlaciones de variables

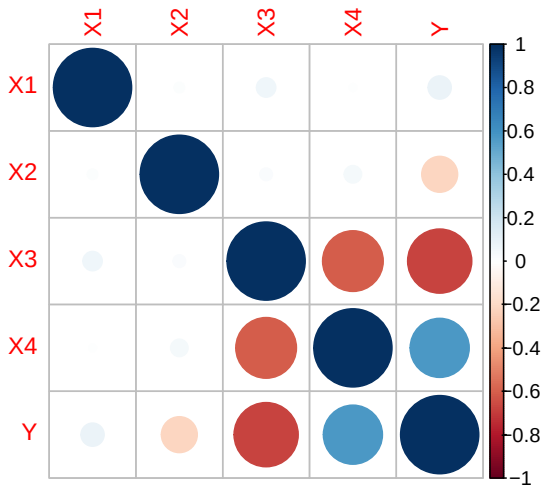


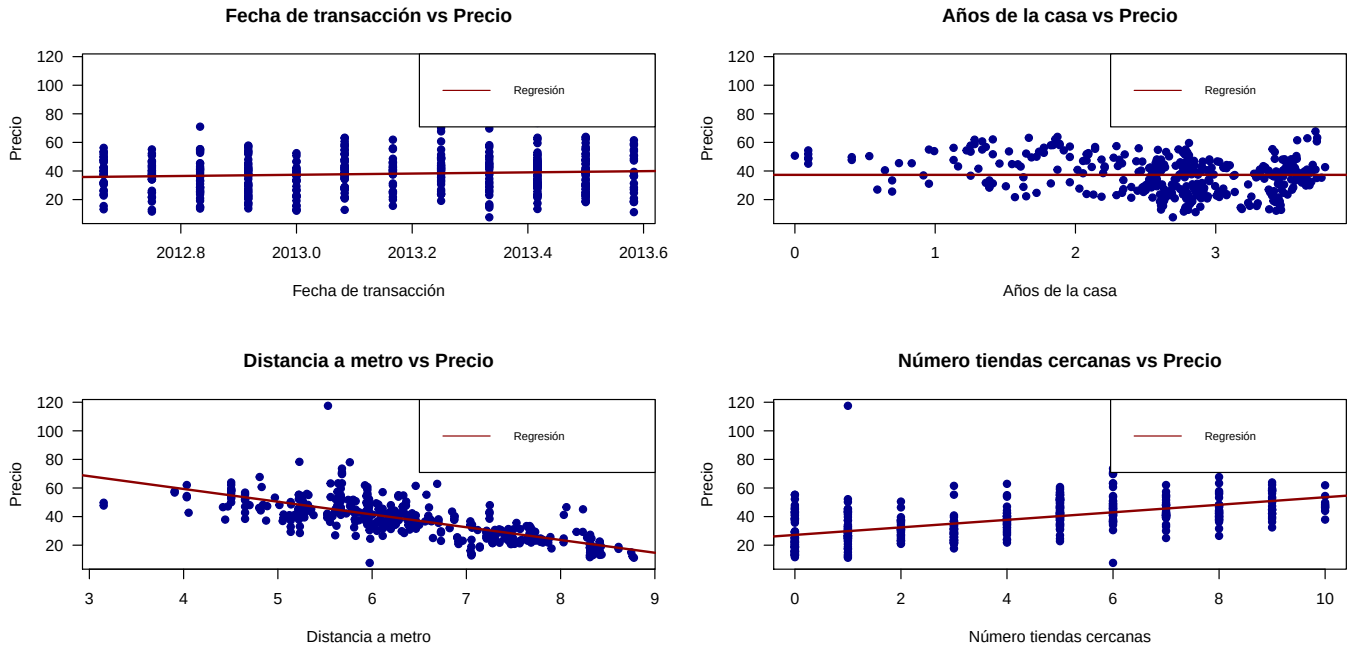
Tabla 2: Correlación de todas las variables con Y

Variables	X1	X2	X3	X4
Correlación con Y	0.0875	-0.2106	-0.6736	0.5710

Tanto en el gráfico como en la tabla, se puede notar una gran relación entre las variables X3(X3 distance to the nearest MRT station) y X4(X4 number of convenience stores) con Y(Y house price of unit area), por lo que tenemos que ponerle principal importancia a estas 2, sin dejar de lado X1(X1 house price of unit area) y X2(X2 house age).

Iniciaremos mostrando los gráficos de cada variable en relación a Y, con la intención de ver el mejor modelo que se ajuste a cada variable.

Figura 2: Regresiones por variable

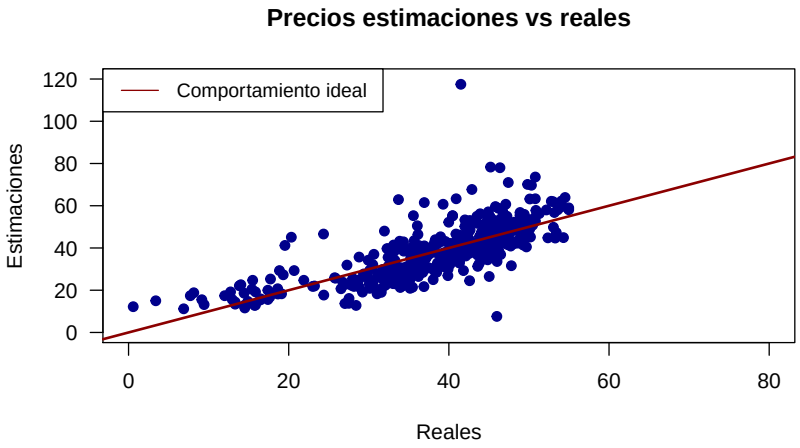


Podemos notar que los modelos ajustan bastante bien, siendo que:

- Para Fecha de transacción(X1) ajusta mejor un modelo lineal.
- Para Años de la casa(X2) ajusta mejor un modelo logarítmico.
- Para Distancia a metro(X3) ajusta mejor un modelo logarítmico.
- Para Número tiendas cercanas(X4) ajusta mejor un modelo lineal.

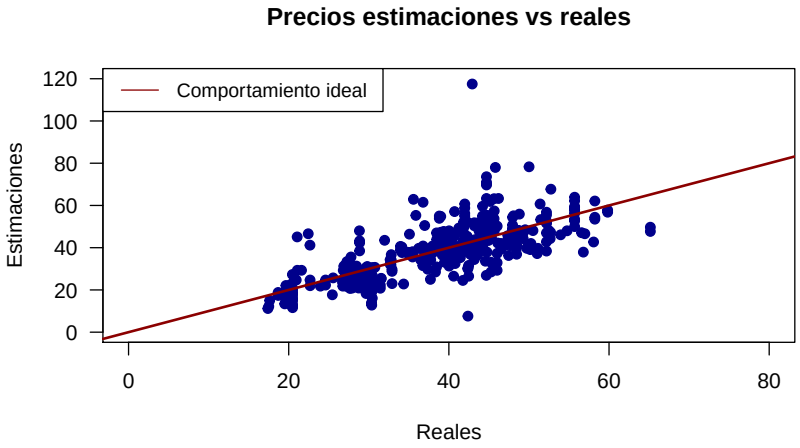
Crearemos un modelo lineal en el cual se utilice las cuatro variables con las cuales estemos trabajando, además entregaremos los valores que mas importantes que utilizaremos para comparar nuestros modelos, a este modelo lo llamaremos "modelo múltiple 1".

Figura 3: Modelo Múltiple 1



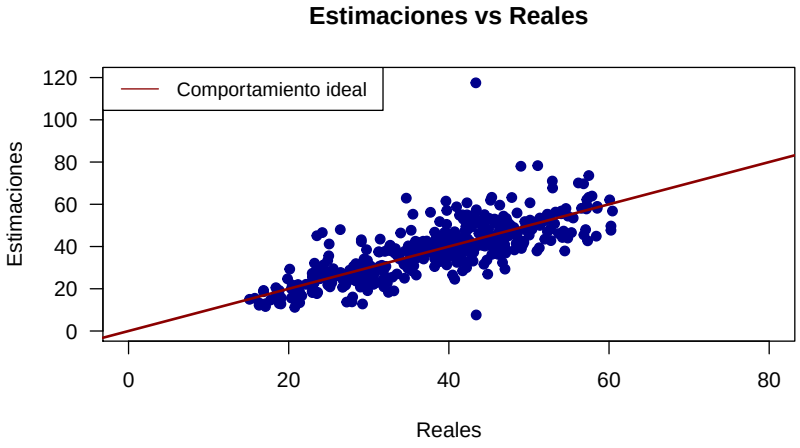
Como podemos ver en la Figura 1 y en la Tabla 2, tenemos que hay 2 variables que están muy correlacionadas con Y, entonces se creará también un modelo con las variables X3 y X4. Para poder crear este modelo haremos uno utilizando la regresión logarítmica de X3 y la regresión lineal de X4 para que nos ajuste mejor. A este modelo le llamaremos "modelo múltiple 2".

Figura 4: Modelo Múltiple 2



Como estamos trabajando solo con 4 variables para estimar Y, podemos decir que son pocas, entonces crearemos un modelo con todas estas variables. Pero la diferencia con el modelo múltiple 1, es que se utilizará la regresión que se ajuste mejor a cada variable, es decir, regresión logarítmica para X2 y X3, mientras que para X1 y X4 se usará una regresión lineal. A este modelo lo llamaremos "modelo múltiple 3".

Figura 5: Modelo Múltiple 3



Pero para cada uno de ellos determinaremos los valores R^2, R^2 ajustado, AIC y BIC. Estos permiten diferenciar los modelos y determinar el mejor modelo.

Tabla 3: Valores comparativos de modelos

	R^2	R^2 **ajustado**	**AIC**	**BIC**
Modelo múltiple 1	0.555	0.551	3011.914	3036.069
Modelo múltiple 2	0.548	0.546	3014.780	3030.884
Modelo múltiple 3	0.596	0.592	2972.584	2996.739

5. Conclusiones

Pudimos ver a lo largo de los resultados la importancia de todas las variables para determinar el precio de una casa, pero principalmente, las variables X3 distance to the nearest MRT station y X4 number of convenience stores, ya que son las que presentaron una correlación mas alta (se puede ver en la Figura 1 y Tabla 2), y que al presentar un modelo compuesto por éstas dos (modelo múltiple 2), tienen un comportamiento muy similar al modelo múltiple 3. como se puede ver en las Figura 4 y Figura 5.

Además, nos propusimos varios modelos, pero claramente, son mucho mas eficientes los 3 modelos múltiples presentados al final, pero para poder validar cúal de ellos permite predecir con menor error, es el modelo múltiple 3.

Por lo tanto, el modelo que mejor permite predecir, es el modelo múltiple 3, porque como vemos en Tabla 3, es el modelo con mayor R^2 y R^2 ajustado, y también, porque tiene el menor valor AIC y BIC, lo que nos permite poder decir directamente, que es el mejor modelo que hemos presentado. Esto se debe a que está compuesto por la regresión que mejor ajusta a cada una de las variables, es decir, que está compuesto por dos regresiones logarítmicas (por las variables X2 y X3) y otras dos regresiones lineales(variables X1 y X4), lo que permite ajustar de la mejor manera posible para predecir.

6. Referencias

Yeh, I. C., Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

Criterio de información de Akaike. (s.f). En Wikipedia. Recuperado el 13 de diciembre de 2020 de `https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_de_Akaike`

Criterio de información bayesiano. (s.f). En Wikipedia. Recuperado el 13 de diciembre de 2020 de `https://es.wikipedia.org/wiki/Criterio_de_informaci%C3%B3n_bayesiano`