Proyecto del Segundo Parcial de Modelos Estadísticos Aplicados I

Christian Salas M., John Borbor M., Karla Silva T.

8 de septiembre de 2021

Parte I. Dataset y problema de investigación

Se seleccionó un conjunto de datos reales obtenidos de un estudio observacional. Según el repositorio, se indica que 'Real Estate Valuation' es un problema de regresión. El conjunto de datos muestra los valores históricos del mercado de valoración inmobiliaria recopilada en el distrito de Xindian, en la ciudad de Nueva Taipéi en Taiwan.

Con respecto a la información de los atributos, tenemos que para las variables predictoras, X_1 indica la fecha de transacción, X_2 la edad de la casa en años, X_3 la distancia en metros a la estación de metro MRT más cercana, X_4 el número de tiendas de conveniencia en el círculo de convivencia, X_5 la latitud en grados y X_6 longitud geográfica en grados. Finalmente, la variable de respuesta cuantitativa Y indica el precio de la vivienda por área unitaria. (10000 Nuevo Dolar Taiwanés/Ping, donde 1 Ping = 3.3 m^2)

Por lo tanto, el problema de investigación se enuncia a continuación

¿Existe asociación entre alguno de los 6 efectos de interés y el precio 'Y' de la vivienda por área unitaria?

Parte II. Construcción del modelo

Para construir el modelo de regresión lineal múltiple, utilizamos la técnica de eliminación Backward Stepwise Regression o *'Paso a paso hacia atrás'*, la cual consiste en primero definir el modelo de regresión con todas las variables predictoras de interés y en este caso, mediante la evaluación de los códigos de significancia entregadas por la función summary() del modelo y mediante el análisis de cada coeficiente de determinación R^2 y R^2 ajustado, ir eliminando paso a paso las variables que no aportan o no son significativas en el modelo. En este caso, pudimos analizar cada caso en particular aplicando la técnica de eliminación de paso a paso hacia atrás.

Procedemos a utilizar eliminación paso a paso hacia atrás, para ello primero consideramos todas las variables predictoras y utilizando la función summary y analizando los códigos de significancia y los valores de R^2, vamos eliminando una a una hasta tener el modelo con las variables de importancia. Además, evitaremos incluir términos de interacción, para no agregar complejidad al análisis.

Consideramos el siguiente modelo final resultante después de aplicar la técnica de selección de variables predictoras:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

donde ϵ_i $N(0, \sigma^2)$ para $i = 1, 2, \ldots, n$, para las variables definidas en la parte I.

Finalmente, utilizando la función `step()` que selecciona un modelo por el Criterio de Información de Akaike (AIC) en un algoritmo de Stepwise, verificamos que también nos devuelva el modelo mediante la técnica Backward Stepwise Regression. La única diferencia es que `step()` devolvió el modelo solamente sin X_6, i.e. aún incluía X_1, pero mediante el primer análisis exhaustivo decidimos que quitar X_1 del modelo también es pertinente.

Parte III. Evaluación del modelo

Para la evaluación del modelo final, utilizamos primero la técnica empírica que nos permite observar los gráficos Residuals vs Fitted, Normal Q-Q, Scale-Location y Residuals vs Leverage. En la sección de 'Apéndice' se muestran todos los gráficos que se describen a continuación.

El gráfico **Residuals vs Fitted** nos muestra que tenemos cierto grado de no linealidad. Tenemos 3 puntos que están muy alejados del modelo: 271, 313, 114, los cuales podrían ser outliers o bien valores extremos. Si los consideramos en contra de la media de la muestra podríamos decir que son outliers, pero tampoco podemos descartar la posibilidad de que podrían ser valores extremos. De todas formas, el modelo mejoró después de que los 3 puntos fueron eliminados.

El gráfico **Normal Q-Q** nos muestra cómo se comparan los residuos con los cuantiles teóricos de la distribución normal estándar, vemos que tenemos un ligero problema en ambos extremos del gráfico, sobretodo en el extremo derecho. Idealmente, todas las observaciones debe ajustarse alrededor de la recta diagonal de referencia.

El gráfico **Scale-Location**, nos muestra la distribución de los residuos del modelo, en relación a los valores ajustados por el modelo $\hat{Y}$.

El gráfico **Residuals vs Leverage**, nos muestra las distancias de Cook, identifica todos los valores extremos y muestra que algunos de esos valores extremos tienen un alto impacto en el modelo. La construcción del modelo está altamente influenciada por esos valores extremos. Eso significa que la observación atípica de la derecha probablemente tiene un alto impacto en el modelo.

Por otro lado, no se identifican ni variables de confusión ni variables mediadoras ni variables colisionadoras.

Utilizando la prueba global `summary(gvlma())` pudimos notar que el único supuesto que cumple el modelo múltiple es la Heterocedasticidad.

Finalmente, para determinar si existe o no multicolinealidad, utilizamos la función `vif()` de la librería 'car' que calcula el Factor de Inflación de la Varianza. Recordemos que, de acuerdo a la bibliografía, un valor $VIF > 10$ es una referencia para identificar problemas de multicolinealidad donde $VIF = \dfrac{1}{1 - R^2}$. En este caso, obtuvimos 1.013216, 1.992371, 1.607857 y 1.575344 para $x2$, $x3$, $x4$ y $x5$ respectivamente. Como cada uno de estos valores es mucho menor que 10, concluimos que **no existe multicolinealidad** entre las variables predictoras consideradas.

Parte IV. Interpretación de resultados

Para las pruebas de hipótesis y otras inferencias, utilizamos el modelo final definido en la parte 2. Sin embargo, es importante recalcar que dado que no se cumplieron todos los supuestos en la prueba global, aún utilizando el método de Box Cox, las inferencias realizadas no serán exactas. En este estudio observacional no podemos establecer relaciones causales ya que pueden existir variables de confusión que no fueron observadas para el estudio.

Sea $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta X_{i6} + \epsilon_i$ el modelo completo y sea $Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$ el modelo reducido final. Realizamos la prueba F con la función anova() para comparar el modelo completo vs el modelo reducido con el método de paso a paso hacia atrás. Obtuvimos que (F* = 5.4819, p-value = 0.004474), es decir rechazamos H_0 con un $\alpha = 0.001$ de significancia. Sin embargo, es importante recalcar que generalmente se utiliza un alfa del 5% o del 10%. Es muy poco común usar un alfa del 0.1% ya que con ese nivel la potencia de la prueba será muy pequeña.

Según ANOVA, el modelo final `lm(y ~ x2 + x3 + x4 + x5)` es significativamente diferente al modelo completo `lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)`. El test compara la reducción de la suma cuadradática de los residuos. Podemos ver que existe un ligero aumento en la suma cuadrática de los residuos en el modelo final, es decir aumenta comparado con el modelo completo. Sin embargo, el valor p pequeño indica que quitando X_1 y X_6 del modelo, se logra un mejor ajuste en comparación con el modelo completo.

Por otro lado, con respecto a las limitaciones del modelo, desde el principio tuvimos indicios de que el modelo podría ser no paramétrico. Recordemos que en los problemas prácticos no es posible conocer el verdadero modelo que genera las respuestas. El R^2 nunca sobrepasó el 0.58 en ninguno de los modelos evaluados. Este límite de 0.58 es importante mencionarlo, pues el R^2 nos indica que tan bien el modelo explica la variación en los datos de la variable de respuesta. Sin embargo, es importante considerar que el R^2 no puede ser utilizado como un único criterio para evaluar que tan apropiado es el modelo para los datos, pues el R^2 tiene muchas limitaciones.

En el modelo completo, el Multiple R-squared es de 0.5824, es decir el modelo explica el 58% de los errores o residuos del modelo. El ajustado Adjusted R-squared: 0.5669 toma en consideración el número de variables que se usó para construir el modelo, mayor cantidad de variables implica mayor R^2 ajustado.

Pudimos comprobar que efectivamente, un modelo no paramétrico es más apropiado para analizar las relaciones de interés. Recordemos que cuando utilizamos la función `summary(gvmla())`, pudimos descubrir que algunos de los supuestos no se cumplián. Más aún, ni si quiera se cumplían todos los supuestos de la prueba global utilizando la transformación de Box Cox. Por lo tanto, esto implica que las inferencias solo permitan concluir asociaciones, y nunca establecer causalidad. Sólo se puede establecer causalidad con supuestos adicionales. De todos modos, se respondió a la pregunta de investigación satisfactoriamente.

Bibliografía

Kutner M., Nachtsheim C., Neter J., Li W. (2004) *Applied Linear Statistical Models Fifth Edition.* McGraw Hill Irwin. New York.

I-Cheng Yeh. (2018) *Real Estate Valuation Data Set.* Department of Civil Engineering Tamkang University, Taiwan. Recuperado de: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Apéndice

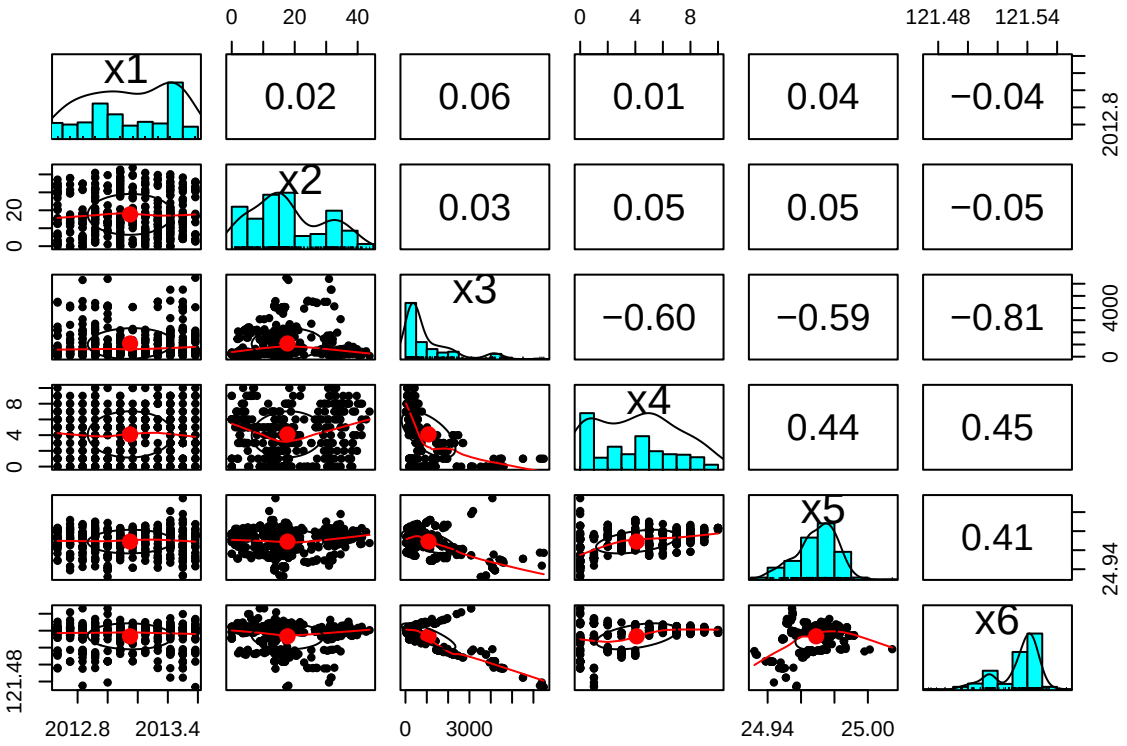

Figure 1: Análisis General

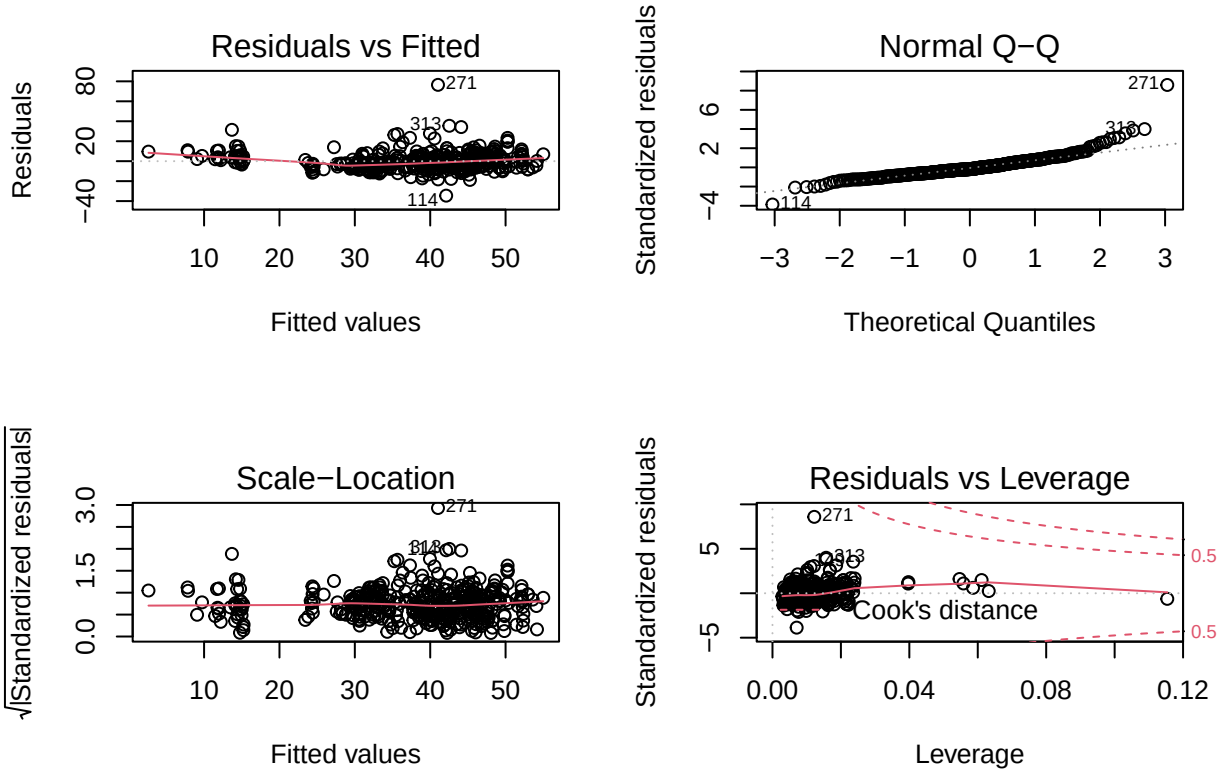

Figure 2: Evaluación del Modelo

```
summary( lm(y ~  x2 + x3 + x4 + x5 )  )
```

```
##
## Call:
## lm(formula = y ~ x2 + x3 + x4 + x5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.522  -5.292  -1.579   4.264  76.466
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.916e+03  1.113e+03  -5.317 1.74e-07 ***
## x2          -2.687e-01  3.893e-02  -6.903 1.95e-11 ***
## x3          -4.175e-03  4.928e-04  -8.473 4.37e-16 ***
## x4           1.165e+00  1.897e-01   6.141 1.94e-09 ***
## x5           2.386e+02  4.456e+01   5.355 1.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.954 on 409 degrees of freedom
## Multiple R-squared:  0.5711, Adjusted R-squared:  0.5669
## F-statistic: 136.2 on 4 and 409 DF,  p-value: < 2.2e-16
```

```
anova( lm(y ~ x1 + x2 + x3 + x4 + x5 + x6)  , lm(y ~ x2 + x3 + x4 + x5 ))
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6
## Model 2: y ~ x2 + x3 + x4 + x5
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    407 31931
## 2    409 32792 -2   -860.18 5.4819 0.004474 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```