Regresión múltiple y otras técnicas multivariadas | Semestre 2018-2

Proyecto final

Fecha de entrega: 7 de junio de 2018

El objetivo es ajustar un modelo de regresión múltiple para explicar el precio de las viviendas a partir de algunas características de las mismas, en las entidades de la república mexicana. Todos los datos están disponibles aquí. Cada equipo debe seleccionar una entidad. Cada conjunto de datos contiene las siguientes variables:

- `Precio`: precio de la vivienda (MXN).
- `land_area`: superficie del terreno.
- `build_area`: superficie construida.
- `bedrooms`: número de dormitorios.
- `bathrooms`: número de baños.
- `half_bathrooms`: número de medios baños.
- `garages`: número de lugares de estacionamiento.
- `Pisos`: número de pisos de la vivienda.

Los puntos 1 a 6 deben realizarse con la información del archivo `ENTIDAD_entrenamiento.csv` y solamente el punto 7 se realiza con los archivos `ENTIDAD_validacionX.csv`.

1. Análisis exploratorio de datos.
   - *a*) Histogramas para las variables individuales.
   - *b*) Gráficas de dispersión del precio y cada variable explicativa continua.
   - *c*) *Boxplots* de precio condicional a cada variable explicativa discreta.
   - *d*) Proponer transformaciones en las variables para mejorar la asociación lineal entre las mismas.
2. Modelos preliminares
   - *a*) Ajustar modelos RLS para el precio con cada una de las variables explicativas. Reportar los resultados relevantes.
   - *b*) Ajustar un modelo RLM para el precio con todas las variables explicativas. Reportar los resultados relevantes.
   - *c*) Comparar los resultados de los modelos anteriores y decidir qué variables deberían incluirse en el modelo definitivo.
   - *d*) Según las observaciones del inciso anterior, ajustar el modelo indicado y presentar los resultados relevantes.
3. Selección de modelos
   - *a*) Utilizar los procedimientos de selección de modelos *backward*, *forward* y *stepwise* para construir RLM para el precio. Reportar los resultados relevantes.

b) Comparar los resultados de los modelos ajustados en el inciso anterior y el modelo final del punto anterior.

c) A partir de los resultados anteriores seleccionar un modelo para explicar el precio de las viviendas.

4. Validación de supuestos

   a) Detectar observaciones atípicas o *outliers* y si las hay, proponer alguna medida correctiva y aplicarla.

   b) Detectar problemas de multicolinealidad entre las variables explicativas incluidas en el modelo al final del punto anterior y si los hay, proponer alguna medida correctiva y aplicarla.

   c) Realizar pruebas de linealidad y falta de ajuste. De acuerdo a los resultados proponer mejoras al modelo.

   d) Realizar pruebas de homocedasticidad y comentar los resultados. De acuerdo a los resultados proponer mejoras al modelo o alternativas de estimación.

   e) Realizar pruebas de normalidad en los residuos y comentar los resultados. De acuerdo a los resultados proponer mejoras al modelo o alternativas de estimación.

   f) A partir de los resultados anteriores ajustar un nuevo modelo que incorpore las correcciones indicadas en los incisos anteriores. Presentar los resultados relevantes.

5. Pruebas de significancia

   a) Presentar la tabla ANOVA para contrastar la significancia del modelo final ajustado en el punto anterior. Interpretar los resultados.

   b) Mostrar las pruebas de significancia simultáneas de los coeficientes del modelo. Interpretar los resultados.

   c) Mostrar el coeficiente R^2_{adj} del modelo e interpretar el resultado.

6. Interpretar los resultados del modelo ajustado en el contexto de los datos.

7. Validación del modelo. Con cada uno de los cinco archivos `ENTIDAD_validacionX.csv`, realizar lo siguiente.

   a) Con la información de las variables explicativas predecir el precio de las viviendas según el modelo previamente ajustado.

   b) Calcular la suma de cuadrados del error con las predicciones del inciso anterior. Reportar los resultados obtenidos.

   c) Graficar los precios verdaderos contra los precios predichos en el inciso *a*). Interpretar los resultados.

   d) En el contexto de los datos, ¿qué tan bueno es el modelo ajustado para predecir los precios de las viviendas a partir de las características consideradas?