Modelos no paramétricos y de regresión/Estadística II | Semestre 2018-1

Proyecto modelos de regresión

Fecha de entrega: 24 de noviembre

El objetivo es modelar el ingreso corriente en los hogares de los municipios del país a partir un conjunto de variables sociodemográficas relevantes.

- `ENT`: Clave de la entidad
- `Entidad`: porcentaje de población de 6 a 14 años que no asiste a la escuela.
- `MUN`: porcentaje de población de 15 años y más con educación básica incompleta.
- `Municipio`: porcentaje de viviendas particulares habitadas con piso de tierra.
- `Alf`: porcentaje de población de 15 o más años que sabe leer y escribir.
- `Asi`: porcentaje de población de 6 a 14 años que asiste a la escuela.
- `EBC`: porcentaje de población de 15 o más años con educación básica completa.
- `Der`: porcentaje de población con derechohabiencia a servicios de salud.
- `Pis`: porcentaje de viviendas particulares habitadas con piso diferente de tierra.
- `San`: porcentaje de viviendas particulares habitadas que disponen de excusado o sanitario.
- `Agu`: porcentaje de viviendas particulares habitadas que no disponen de agua potable.
- `Dre`: porcentaje de viviendas particulares habitadas que no disponen de drenaje.
- `Ele`: porcentaje de viviendas particulares habitadas que no disponen de electricidad.
- `Lav`: porcentaje de viviendas particulares habitadas que disponen de lavadora.
- `Ref`: porcentaje de viviendas particulares habitadas que disponen de refrigerador.
- `Ing`: Estimación de la media del ingreso corriente mensual en los hogares.

Con los datos del archivo `proyecto.csv`, responder lo siguiente.

1. Hacer un análisis exploratorio pertinente:
   - *a*) Explorar la asociación lineal entre las variables con *corrplot*.
   - *b*) Explorar la dispersión de cada variable con *boxplots*.
   - *c*) Interpretar los resultados.
2. Ajustar un modelo RLM tomando como respuesta la variable `Ing`.
   - *a*) Reportar las estimaciones, puntuales y por intervalo, de los parámetros del modelo.
   - *b*) Reportar la significancia individual y conjunta (ANOVA) de las variables incluidas.
   - *c*) Reportar los coeficientes R^2 y R^2_{adj}.
3. Multicolinealidad
   - *a*) Calcular la matriz de correlaciones y presentarla como `corrplot`.
   - *b*) Calcular indice de condición κ de la matriz $\mathbf{X}^T\mathbf{X}$.
   - *c*) Calcular los factores de inflación de la varianza de cada variable explicativa.
   - *d*) A partir de los resultados anteriores, concluir si existen problemas de multicolinealidad.

e) Si es el caso, proponer y aplicar las medidas correctivas que se consideren necesarias.

f) Si es el caso, validar que el problema de multicolinealidad se haya corregido.

g) Si es el caso, ajustar nuevamente el modelo incorporando las correcciones aplicadas.

4. Validar el supuesto de linealidad.

   a) Hacer gráficos de residuos parciales para cada variable.

   b) Hacer pruebas de falta de ajusta (*lack-of-fit*) para cada variable.

   c) Hacer una prueba de no aditividad de Tukey.

   d) A partir de los resultados anteriores, concluir si existen desviaciones graves al supuesto de linealidad.

   e) Si es el caso, proponer y aplicar las medidas correctivas que se consideren necesarias y validar que el problema de no linealidad se haya corregido.

   f) Si es el caso, ajustar nuevamente el modelo incorporando las correcciones aplicadas.

   g) Si es el caso, explorar nuevamente los supuestos anteriores y ajustar de nuevo el modelo incorporando las correcciones aplicadas.

5. Validación del supuesto de homocedasticidad.

   a) Graficar los residuos contra cada variable explicativa.

   b) Hacer una prueba de Breusch-Pagan.

   c) Hacer una prueba de White.

   d) A partir de los resultados anteriores, concluir si existen desviaciones graves al supuesto de linealidad.

   e) Si es el caso, proponer y aplicar las medias correctivas que se consideren necesarias y validar que las desviaciones al supuesto de homocedasticidad se hayan corregido.

   f) Si es el caso, ajustar nuevamente el modelo incorporando las correcciones aplicadas.

   g) Si es el caso, explorar nuevamente los supuestos anteriores y ajustar de nuevo el modelo incorporando las correcciones aplicadas.

6. Validar el supuesto de normalidad.

   a) Presentar un histograma de los residuos estudentizados.

   b) Presentar un `QQplot` de los residuos estudentiados.

   c) Aplicar alguna prueba de normalidad (K-S, A-D, etc.)

   d) A partir de los resultados anteriores, concluir si existen desviaciones graves al supuesto de linealidad.

   e) Si es el caso, proponer y aplicar las medidas correctivas que se consideren necesarias y validar que las desviaciones al supuesto de normalidad se hayan corregido.

   f) Si es el caso, ajustar nuevamente el modelo incorporando las correcciones aplicadas.

   g) Si es el caso, explorar nuevamente los supuestos anteriores y ajustar de nuevo el modelo incorporando las correcciones aplicadas.

7. Explorar la presencia de observaciones atípicas e inluyentes.

   a) Utilice las entradas de la matriz $\mathbf{H}$ para identificar observaciones influyentes.

b) Utilice los `DFFITS` y `DFBETAS` para identificar observaciones influyentes.

c) Utilice los residuales estudentizados para identificar observaciones atípicas.

d) A partir de los resultados anteriores, decida si hay presencia de observaciones atípicas o de observaciones influyentes, de ser el caso decida que medidas correctivas aplicar y valide que el problema se haya resuelto.

e) Si es el caso, ajustar nuevamente el modelo incorporando las correcciones aplicadas.

f) Si es el caso, explorar nuevamente los supuestos anteriores y ajustar de nuevo el modelo incorporando las correcciones aplicadas.

8. Del último modelo ajustado:

   a) Reportar las estimaciones, puntuales y por intervalo, de los parámetros del modelo.

   b) Reportar la significancia individual y conjunta (ANOVA) de las variables incluidas.

   c) Reportar los coeficientes R^2 y R^2_{adj}.

   d) Comparar estos resultados con los obtenidos en el punto 2. ¿Mejoró el ajuste del modelo?

9. Interpretar los resultados en el contexto de los datos.

   a) ¿Qué variables tienen algún efecto para modelar el ingreso en los hogares?

   b) ¿Qué variable tiene el mayor efecto? ¿Qué variable tiene el mejor efecto?

   c) ¿Qué signos tienen los βs y qué interpretación tienen estos signos?

   d) ¿La varianza del modelo permite obtener *buenas* predicciones?