Regresión múltiple 17-2

Proyecto final

Fecha de entrega: 08/06/2017

El objetivo es modelar el PIB pp de las entidades del país a partir de las variables sociodemográficas utilizadas en el *índice de rezago social* de CONEVAL. El conjunto de datos contiene las siguientes variables.

- `P15YM_NALF`: porcentaje de población de 15 años y más que no sabe leer y escribir.
- `P6A14_NASI`: porcentaje de población de 6 a 14 años que no asiste a la escuela.
- `P15YM_EBIN`: porcentaje de población de 15 años y más con educación básica incompleta.
- `VIV_PIS`: porcentaje de viviendas particulares habitadas con piso de tierra.
- `VIV_NEXC`: porcentaje de viviendas particulares habitadas que no disponen de excusado o sanitario.
- `VIV_NAGU`: porcentaje de viviendas particulares habitadas que no disponen de agua potable.
- `VIV_NDRE`: porcentaje de viviendas particulares habitadas que no disponen de drenaje.
- `VIV_NELE`: porcentaje de viviendas particulares habitadas que no disponen de electricidad.
- `VIV_NLAV`: porcentaje de viviendas particulares habitadas que no disponen de lavadora.
- `VIV_NREF`: porcentaje de viviendas particulares habitadas que no disponen de refrigerador.
- `PIBPP`: Producto interno bruto per capita en miles de pesos.

Los datos están en el archivo `rezago.csv`. Utilizar la información de las 32 entidades y los años 2005 y 2010 para hacer lo siguiente.

1. Hacer un análisis exploratorio de los datos.
2. Identificar observaciones atípicas, *outliers*.
3. Explorar si hay multicolinealidad en las variables explicativas.
4. Ajustar un modelo RLM tomando como respuesta a `PIBPP` y considerando las variables explicativas que sugieran las exploraciones anteriores (quizá sea necesario transformar a linealidad o eliminar por multicolinealidad). De igual manera, considerar las observaciones que sugieran los análisis previos (posiblemente sea necesario eliminar observaciones influyentes o atípicas).
5. Explorar no linealidad y heterocedasticidad. Confirmar los hallazgos con pruebas de falta de ajuste (no linealidad en las v. explicativas), no aditividad (no linealidad en la respuesta) y homocedasticidad.
6. En caso que las pruebas resulten positivas para no linealidad o heterocedasticidad, aplicar las medidas correctivas que se consideren necesarias y ajustar de nuevo el modelo con las variables transformadas.
7. Explorar la normalidad de los errores. Confirmar con alguna prueba de normaliadad.
8. Si el supuesto de normalidad es razonable, hacer la prueba de significancia del modelo y presentar la tabla ANOVA completa. Si no hay normalidad, hacer la prueba de significancia del modelo utilizando *bootstrap* (se debe aproximar la distribución del estadístico F).
9. Si el supuesto de normalidad es razonable, hacer pruebas de t simultáneas para las componentes de $\boldsymbol{\beta}$. Si no hay normalidad, hacer pruebas individuales para los componentes del vector $\boldsymbol{\beta}$ utilizando *bootstrap*. Interpretar los resultados en el contexto del problema.
10. Calcular los coeficientes R^2 y R^2-ajustado. Interpretar los resultados.