Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística

Trabajo final

Modelo lineal general II

1. La base de datos `WHO2016.csv` recopila 7 características socio-económicas y de salud sobre 148 países. Con estos datos, se tiene como objetivo determinar que factores influyen sobre la esperanza de vida de los países, especialmente variables asociadas a inmunización de enfermedades. Esto ayudará a los países a identificar a que áreas deben reformar para mejorar la esperanza de vida de su población. Otro objetivo es proponer un modelo predictivo para esperanza de vida.

   Las variables que se encuentran en la base de datos son:

   - `Status`: estatus del país (desarrollado o en desarrollo),
   - `Life.expectancy`: esperanza de vida (en años),
   - `Infant.deaths`: número de infantes (menores de 5 años) muertos por cada mil nacidos vivos,
   - `Alcohol`: promedio de consumo de alcohol por persona (litros),
   - `percentage.expenditure`: gasto en salud como porcentaje del PIB per cápita (%),
   - `Hepatitis.B`: cobertura de vacunación contra la hepatitis B entre los niños de 1 año (%),
   - `BMI`: Índice de masa corporal promedio de toda la población,
   - `Polio`: cobertura de vacunación contra el polio entre niños de 1 año (%)
   - `Diphtheria`: Cobertura de vacunación contra Difteria, tétano, toxoide y tos ferina en niños de 1 año (%),
   - `HIV.AIDS`: número de muertes de infantes por sida por cada mil nacidos vivos,
   - `GDP`: producto interno bruto per cápita (en dólares),
   - `Thinness.10.19.years`: prevalencia de delgadez en niños y adolescentes de 10 a 19 años (%),
   - `Schooling`: promedio del número de años de escolaridad,
   - `Homicides`: tasa de homicidios por 100,000 habitantes,
   - `Fertility`: número esperado de hijos que tendría una mujer si viviera hasta el final de sus años fértiles,
   - `Unemployment.rate..women`: tasa de desempleo de las mujeres.

   Antes de empezar el análisis de datos, divida la muestra en dos partes de forma aleatoria: una para estimar el modelo - entrenamiento - (100 observaciones) y otra para hacer una evaluación del modelo ajustado - validación - (48 observaciones).

   Con la sub-muestra de entrenamiento ajuste un modelo de regresión lineal considerando la variable `Life.expectancy` como respuesta y todas las demás como covariables. Si es necesario, haga transformaciones para la variable respuesta y/o covariables (si alguna covariable tiene un comportamiento muy asimétrico es preferible realizar una transformación logarítmica). Evalúe multicolinealidad.

   Luego realice un proceso de selección de variables. Proponga al menos dos modelos diferentes. Justifique claramente los criterios de selección utilizados. A partir de los modelos propuestos (y el modelo completo) realice predicciones para los países de la sub-muestra de validación. Para evaluar las predicciones utilice el error cuadrado medio de predicción:

$$ECMP = \frac{1}{M}\sum_{i=1}^{M}(y_i^* - \widehat{y}_i^*)^2,$$

donde y_i^* es la esperanza de vida (o una transformación de ella) del i-ésimo país de la muestra de validación, $\widehat{y}_i^*$ es la predicción de la esperanza de vida (o de la transformación) del i-ésimo país usando la muestra de entrenamiento, y M es el tamaño de la muestra de la base de datos de validación.

A partir de este análisis, ¿qué factores tienen mayor influencia sobre la esperanza de vida?, particularmente, ¿la inmunización de la población tiene algún efecto sobre la esperanza de vida de los países?. ¿los modelos predictivos propuestos proporcionan buenas predicciones de la esperanza de vida?.

2. Los datos `campuscrime.csv` contienen la cantidad de robos reportados en 47 universidades públicas de los Estados Unidos en el año 2009 (`burg09`). Además, se tiene la siguiente información adicional sobre estas universidades: la región donde se encuentra (`region`), la cantidad total de estudiantes (`total`), el porcentaje de hombres (`pct.male`), puntajes promedio de las pruebas SAT (`sat.tot`) y ACT (`act.comp`), y costo de la matrícula (`tuition`).

   Ajuste un modelo para el número de robos reportados en las universidades en el año 2009 utilizando como covariables: región donde se encuentra la universidad, el porcentaje de hombres, puntajes promedio de las pruebas SAT y ACT, y costo de la matrícula (en miles de dolares). Para esto, tenga en cuenta: la distribución de probabilidad para la variable respuesta, inclusión de posible offset, sobredispersión y/o inflación de ceros.

   A partir de los resultados, que factores (y como) influyen sobre la tasa del número de robos reportados.

Aspectos a tener en cuenta

Para la entrega, tenga en cuenta lo siguiente aspectos:

- El reporte no debe exceder 10 páginas. No incluir códigos o salidas de R.
- Todas las tablas y figuras deben estar enumeradas. Solo incluya las que consideren relevantes (es decir, no incluya tablas o figuras si no van a hacer ningún comentario importante sobre ellas). Recuerde que el número de páginas es limitado, así que use el espacio de forma inteligente.
- Los modelos propuestos deben estar claramente especificados, incluyendo sus supuestos.
- Para cada caso, debe incluir una sección de conclusiones y recomendaciones.
- Fecha de entrega: 28 de junio de 2023 al medio día (en físico y a través del campus virtual).
- La sustentación del trabajo se hará el 30 de junio de 2023 (horario de clase).
- La calificación será 50 % informe y 50 % sustentación.