Trabajo Práctico Final

Regresión Avanzada
Maestría en Explotación de Datos y Gestión del Conocimiento
Universidad Austral

Mg. Marcos Prunello - Mg. Diego Marfetán Molina

Junio 2021

Descripción del Conjunto de Datos

En las elecciones presidenciales de Estados Unidos llevadas a cabo en 1992 se enfrentaron George Bush (padre), representante del partido Republicano, y Bill Clinton, candidato por el partido Demócrata. Bush era presidente en ese momento y buscaba la reelección, pero perdió a manos de Clinton, quien finalmente gobernaría el país entre 1993 y 2001.

El conjunto de datos a analizar en este Trabajo Práctico contiene información sobre el porcentaje de votos obtenido por Clinton en cada condado de Estados Unidos, junto con información social, económica y demográfica de cada uno de ellos.

Las variables presentes en el archivo *clinton.txt* son:

1. condado: nombre del condado
2. estado: estado donde se encuentra ubicado el condado
3. pje: porcentaje de votos obtenidos por Bill Clinton en ese condado
4. edad: mediana de la edad de los habitantes del condado
5. ahorros: ahorro promedio de los habitantes del condado
6. ingpc: ingreso per cápita en ese condado
7. pobreza: porcentaje de la población bajo la línea de la pobreza en ese condado
8. veteranos: porcentaje de la población que son veteranos de guerra en ese condado
9. mujeres: porcentaje de mujeres en la población del condado
10. densidad: densidad poblacional en ese condado
11. ancianos: porcentaje de la población que vive en residencias de cuidados para personas mayores en ese condado
12. crimen: índice de criminalidad per cápita en ese condado

Consigna - Parte I

A través de un modelo de **Regresión Lineal Múltiple** ajustado por Mínimos Cuadrados Ordinarios, estudiar el porcentaje de votos obtenidos por el candidato Bill Clinton en cada uno de los condados estadounidenses. Pueden incorporar como explicativas a cualquiera de las restantes variables presentes en la base.

Detalles a tener en cuenta:

- El objetivo principal del análisis consiste en identificar cuáles son las variables más relacionadas con la respuesta. Se pretende que describan la influencia que esta(s) variable(s) poseen sobre ella, por ejemplo interpretando los coeficientes obtenidos en términos del problema, calculando intervalos de confianza u otras medidas de interés, etc.
- Para alcanzar el objetivo deberán buscar el mejor modelo posible entre una serie de candidatos, justificando su elección empleando las distintas herramientas vistas a lo largo del curso: pruebas de hipótesis para comparar modelos, técnicas de selección de variables, análisis de residuos, etc. Se espera que el reporte final incluya comentarios acerca del cumplimiento de los supuestos de los modelos postulados.

- Además de lo mencionado anteriormente, será necesario evaluar la significación de la interacción entre las variables elegidas, y también determinar la presencia de condados atípicos o influyentes. En caso de encontrar observaciones con estas características, no será necesario incursionar en el ajuste de modelos robustos.

Consigna - Parte II

Un enfoque alternativo para analizar estos datos consiste en categorizar la variable respuesta *pje* en una dicotómica, que tome los siguientes valores:

$$\begin{cases} 1 \text{ si el porcentaje de votos obtenidos por Clinton es } > 50 \\ 0 \text{ si el porcentaje de votos obtenidos por Clinton es } \leq 50 \end{cases}$$

Tomando esta nueva variable como respuesta, deberán ajustar un modelo de **Regresión Logística** con enlace canónico, cuyo componente sistemático coincida con el del modelo elegido en la sección anterior.

Se espera que evalúen la bondad del ajuste de este nuevo modelo y la significación de las variables predictoras incluidas, analizando si hubo cambios (o no) con respecto al ajuste anterior. También deberán interpretar los parámetros estimados en términos de razones de odds, para las variables que lo ameriten.

Reglas

- **Material a Entregar:**
  - Archivo en formato .pdf con 10 carillas de extensión como máximo donde figuren los análisis y resultados obtenidos. En el informe se deben incluir los resultados pertinentes, pero no es necesario mostrar los ajustes realizados que hayan sido "descartados". De todas maneras, se fomenta incluir una descripción del camino recorrido hasta llegar al modelo elegido. Si bien no es obligatorio, se promueve el uso de R Markdown para generar el documento.
  - Archivo en formato .Rmd donde se encuentren las sentencias que generan el reporte, o bien un script de R con las sentencias utilizadas para el ajuste y análisis de los modelos.
- **Fecha de Entrega:** jueves 15/07/2021.
- **Medio de Entrega:** enviar los 2 archivos por correo electrónico con copia a ambos docentes: marcosprunello@gmail.com, diego.marfetan@gmail.com.
- **Equipos:** se permite trabajar en grupos de hasta 4 personas.

¡¡¡Éxitos!!!