UC | Chile

Aprendizaje estadistico y computacional

Tarea: Métodos de Regresión

¡Bienvenido(a)!

Te invitamos a realizar el primer trabajo.

- **Objetivo**: Ajustar una regresión lineal y una regresión logística en Python e interpretar sus resultados.
- **Tipo de actividad**: Grupal
- **Tipo de evaluación**: Sumativa
- **Ponderación**: 13%
- **Puntaje**: 60 puntos
- **Calificación**: Escala de 1 a 7, con una exigencia de 50%. La nota mínima para aprobar es 4.0.

Evaluación

Descarga el instrumento de evaluación y revísalo antes de realizar la actividad.

Instrucciones

1. Antes de comenzar, debes haber revisado las siguientes clases y la lectura: videos, tutoriales y lecturas de la semana 2.
2. Leer con atención los siguientes dos casos y responde según lo indicado.
3. Esta Tarea debe ser desarrollada completamente en lenguaje de programación Python, y estructurarse en formato de Notebook (seguir buenas prácticas de

escritura y programación, e incluir comentarios o celdas de markdown suficientes para explicar claramente todos los códigos computacionales).

4. Una vez finalizada la actividad, guarda un archivo con el nombre "Tarea1_ Apellidos_Integrantes", luego suba ambos archivos a la plataforma siguiendo las siguientes instrucciones:
    - Haz clic en el botón para agregar entrega. Se abrirá una nueva ventana que permite arrastrar el archivo y subirlo.
    - Comprueba que el archivo arrastrado es el correcto y presiona el botón para guardar cambios. El documento quedará guardado en la plataforma.

Enunciado

Introducción

Esta tarea está pensada en abordar y profundizar los aspectos inferencias de la regresión lineal y logística, específicamente, entorno al descubrimiento de patrones entre la variable respuesta y las otras covariables como identificar alguna forma funcional (ej. Cuadrados logaritmos, exponenciales) y también enfocada en las técnicas de selección de variables que tienen este tipo de modelos. Finalmente, se dan los primeros pasos en comprender como este tipo de modelos nos permite predecir o clasificar observaciones nuevas, cuantificando, de cierto modo el error o la calidad del ajuste.

Descripción de los problemas

I. **Caso 1 (30 puntos):** Utilice el conjunto de datos "ozone" disponibles en la librería "faraway" (buscar en https://pypi.org/project/faraway/). Considere Y=O3 como variable respuesta, todas las demás serán variables explicativas.
    1. Realizar una descomposición aleatoria de la base de datos con la proporción 80%-20% para train y test, respectivamente.
    2. Utilizando la data train, realizar un análisis descriptivo de las variables de la base de datos. Debe incluir indicadores y gráficas.

3. Utilizando la data train y alguno de los criterios de selección de variables tipo stepwise, determine el modelo lineal que mejor ajusta a la variable respuesta. Indique el criterio utilizado.
4. Analice la significancia del modelo obtenido luego del proceso de selección, y responda si:
    a. ¿Es el modelo obtenido significativo?
    b. ¿Existe alguna covariable no significativa?
    c. ¿En caso de existir alguna covariable no significativa, la quitaría del modelo? Fundamente.
5. Utilizando la data test realizar la predicción de la media (adicional: incluya los intervalos de confianza de las predicciones). Obtenga alguna medida de error de las predicciones.

II. **Caso 2 (30 puntos):** Utilice el conjunto de datos “prostate” disponibles en la libreria ``faraway" (buscar en https://pypi.org/project/faraway/). Además, considere Y= svi como la variable respuesta, donde 1 es presencia y 0 ausencia. Todas las demás serán variables explicativas.

1. Realizar una descomposición aleatoria de la base de datos con la proporción 80%-20% para train y test, respectivamente.
2. Utilizando la data train, realizar un análisis descriptivo de las variables de la base de datos. Debe incluir indicadores y gráficas.
3. Utilizando la data train y alguno de los criterios de selección de variables tipo stepwise, determine el modelo de regresión logística que mejor ajusta a la variable respuesta. Indique el criterio utilizado.
4. Analice la significancia del modelo obtenido luego del proceso de selección, y responda si
    a. ¿Es el modelo obtenido significativo?
    b. ¿Existe alguna covariable no significativa?
    c. ¿En caso de existir alguna covariable no significativa, la quitaría del modelo? Fundamente.

5. Utilizando la data test realizar la predicción de la probabilidad de presencia para cada caso (adicional: incluir los intervalos de confianza). Una vez obtenida la probabilidad realizar la respectiva clasificación

Aspectos formales

Considera los aspectos formales que se describen a continuación:

- Letra Arial 12 normal, interlineado simple.
- Extensión: Entre 1500 - 2000 palabras.
- Utilizar formato APA en citas al interior del texto y en la bibliografía.

¡Mucho éxito!

Importante: la fecha de entrega está indicada en el calendario del curso. Cuidar la redacción y la ortografía. Si tienes alguna duda sobre los contenidos o sobre cómo realizar esta actividad, puedes utilizar la herramienta "Mensajes" y enviar tu pregunta. Recibirás la respuesta de su tutor con las orientaciones correspondientes