DEPARTAMENTO DE INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE SANTIAGO DE CHILE

EJERCICIO PRÁCTICO 13: REGRESIÓN LINEAL SIMPLE Y MÚLTIPLE

CONTEXTO

Si bien hoy en día existen muchas herramientas que facilitan la búsqueda y construcción de un modelo de regresión lineal múltiple (RLM), conseguir un modelo adecuado suele ser un desafío.

El objetivo de este ejercicio es practicar el proceso de creación y evaluación de un modelo de regresión lineal simple (RLS) para predecir una variable numérica y su extensión a multivariado.

OBJETIVOS DE APRENDIZAJE

1. Preparar un conjunto de datos para la construcción de modelos RLS y RLM.
2. Iterar en el proceso de selección de variables, creación y evaluación de modelos RLM, hasta conseguir uno que sea confiable y satisfactorio.

ÉXITO DE LA ACTIVIDAD

El equipo es capaz de encontrar modelos RLS y RLM confiables y de buen desempeño al predecir una variable dependiente.

ACTIVIDADES

Un estudio recolectó muestras de distintas botellas de vino de una importante viña francesa. Estas mediciones están disponibles en el archivo EP13 Datos.csv que acompaña a este enunciado. El estudio incluyó 12 mediciones de características del vino y la indicación de si cada vino es tinto o blanco:

Columna	Descripción	Unidad
clase	Tipo de vino	Categórica (Tinto, Blanco)
calidad	Calidad del vino	Entera [0; 10]
acidez.fija	Ácidos fijos [mg/L]	Real [0; 20]
acidez.volatil	Ácidos volátiles [g/L]	Real [0; 2]
acido.citrico	Ácidos cítrico [mg/L]	Real [0; 2]
azucar.residual	Azúcar residual [g/L]	Real [0; 80]
cloruros	Cloruros [g/L]	Real [0; 1]
dioxido.azufre.libre	Dióxido de azufre libre [g/L]	Real [0; 300]
dioxido.azufre.total	Dióxido de azufre total [g/L]	Real [0; 500]
densidad	Densidad total [g/L]	Real [0; 1,5]
ph	Acidez (pH)	Real [1; 5]
sulfatos	Sulfatos [g/L]	Real [0; 3]
alcohol	Porcentaje de alcohol	Real [0; 20]

1. El equipo copia el enunciado del problema asignado como comentarios de un script R.
2. El equipo lee el enunciado, descarga el archivo de datos (EP13 Datos.csv) desde UVirtual y selecciona las columnas para trabajar de acuerdo a las instrucciones.

3. El equipo construye los modelos solicitados usando la muestra correspondiente.
4. El equipo sube el script con las actividades anteriores comentando en detalle los pasos seguidos.

Fuera del horario de clases, cada equipo debe subir el script realizado UVirtual con el nombre "EP13-respuesta-grupo-i", donde i es el número de grupo asignado. Las respuestas deben subirse antes de las 23:30 del lunes 3 de julio.

PREGUNTA (todos los grupos)

Se pide construir un modelo de regresión lineal simple y otro de regresión lineal múltiple para predecir la variable calidad, de acuerdo con las siguientes instrucciones:

1. Definir la semilla a utilizar, que corresponde a los últimos cuatro dígitos del RUN (sin considerar el dígito verificador) del integrante de menor edad del equipo.
2. Seleccionar una muestra de 100 vinos.
3. Seleccionar de forma aleatoria 6 posibles variables predictoras.
4. Seleccionar, entre las variables que no fueron escogidas en el punto anterior, una que el equipo considere que podría ser útil para predecir la variable calidad, justificando bien esta selección.
5. Usando el entorno R, construir un modelo de regresión lineal simple con el predictor seleccionado en el paso anterior.
6. Agregue la variable seleccionada en el paso 4 al conjunto obtenido en el punto 3.
7. Usando herramientas para la exploración de modelos del entorno R, escoger entre dos y cinco predictores de entre las variables presentes en el conjunto obtenido en el paso anterior para construir un modelo de regresión lineal múltiple.
8. Evaluar los modelos y “arreglarlos” en caso de que tengan algún problema con las condiciones que deben cumplir.
9. Evaluar el poder predictivo del modelo en datos no utilizados para construirlo (o utilizando validación cruzada).

CRITERIOS DE EVALUACIÓN

- Obtienen una muestra de datos para poder crear y evaluar modelos de regresión lineal, cumpliendo las restricciones indicadas en el enunciado (semilla, tamaño, género, posibles variables predictoras, etc.).
- Seleccionan, justificando su utilidad de forma convincente, una variable no elegida anteriormente que utilizan para construir correctamente un modelo de RLS para predecir la variable solicitada.
- Seleccionan un conjunto de variables relevantes para predecir la variable solicitada, utilizando correctamente gráficos y/o utilidades en paquetes de R para explorarlas, desde el conjunto de ocho variables elegidas aleatoriamente como posibles predictores y respetando las otras restricciones indicadas en el enunciado.
- Construyen correctamente un modelo de RLM para predecir la variable solicitada agregando las variables seleccionadas anteriormente al modelo RLS que se tiene.
- Escriben comentarios y código en R correcto que verifica las condiciones que garantizan que tanto el modelo de RLS como el modelo de RLM obtenidos tienen un buen nivel de ajuste y son generalizables, interpretando explícita y correctamente los resultados obtenidos en cada paso y tomando acciones correctivas apropiadas de ser necesarias o comentando los riesgos asociados.

- Escriben código R correcto que evalúa la calidad predictiva tanto del modelo de RLS como del modelo de RLM obtenidos, en datos no utilizados para su construcción, y comparando correctamente los desempeños observados.
- Entregan conclusiones correctas y completas, basadas en las evaluaciones realizadas y el proceso de búsqueda de predictores seguido, respecto del modelo de RLM. obtenido
- El script está completo, ordenado y bien indentado, se comenta paso a paso el procedimiento implementado, con buena redacción (basta una lectura para entender) y con buena ortografía (no más de 3 errores), logrando un programa que es fácil de seguir y que no requiere cambios para que funcione
- Escriben con buena ortografía y redacción (<3 errores), usando vocabulario propio de la disciplina y el contexto del problema.