Examen Final Regresión Avanzada
Universidad Austral

Julio de 2023

Resuelva el siguiente examen utilizando el lenguaje R, entregue la resolución en dos archivos: el archivo Rmarkdown con el script completo, y el archivo html donde se pueden visualizar las salidas del código.
En el archivo *data_pancreas_resumen.csv* se han registrado los datos de pacientes sanos y pacientes que padecen cáncer de páncreas. Las variables se describen a continuación:

Nombre de las variables	Descripción
paciente	Número de Identificación Paciente
edad	Edad (años)
sexo	M = masculino, F = femenino
diagnosis	Diagnóstico: normal (no cáncer de páncreas); maligno (cáncer de páncreas)
estadío	Para los que padecen cáncer, estadío en el que están: I, II, III y IV; para los sanos: NO
creatinina	Marcador de función renal (mg/ml)
LYVE1	Proteína LYVE1 (ng/ml) en orina, receptor endotelial del ácido hialurónico, que se supone involucrado en la metástasis tumoral
REG1B	Niveles urinarios de una proteína(ng/ml) que están asociados con la regeneración pancreática
TFF1	Niveles urinarios de una proteína (ng/ml) involucrada en los procesos de reparación y regeneración del tracto urinario

Recomendaciones*: utilice el comando **read.csv2 ya que en el archivo los datos numéricos tienen coma para separar los decimales. Use el comando **as.factor** para que las variables categóricas sean consideradas como tales.

Para realizar el examen genere una muestra aleatoria estratificada por "diagnosis" de tamaño n = 300 utilizando como semilla el número del DNI/PASAPORTE. Se muestra un ejemplo a continuación.

```
library(splitstackshape)
set.seed(40123456)
strat_data <- stratified(data, "diagnosis", 300/nrow(data))
```

De ahora en más trabaje con esta base de datos.

**Indicaciones: considere un nivel de significancia del 1% para los tests de normalidad.*

Ejercicio 1

1. Construya un modelo lineal simple para explicar el valor de la creatinina en función de alguna de las restantes variables numéricas y evalúe la bondad del ajuste.
2. Realice un análisis diagnóstico y de puntos influyentes e indique si el modelo es adecuado.

3. Realice una transformación de la variable respuesta para intentar lograr normalidad en la distribución de los residuos. Indique si el modelo con esta transformación resulta adecuado.
4. Sin considerar la variable estadío, ajuste un modelo multivariado robusto para explicar el valor de la creatinina y estime el error absoluto medio cometido.
5. Sin considerar la variable estadío, utilice un método de selección de variables para proponer un nuevo modelo multivariado que explique el valor de la creatinina. Estudie el cumplimiento de los supuestos y haga una transformación en caso de ser necesario. Analice los coeficientes del modelo final.
6. Estime los errores de predicción de los 4 modelos previos y compárelos. Cuál elegiría?
7. Le parece adecuado un modelo GAMLSS en este caso? Justifique.

Ejercicio 2

Estudie analítica y gráficamente si:
1. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío.
2. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío considerando sólo la base de pacientes enfermos.
3. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto del sexo.
4. la interacción entre estadío y sexo es significativa cuando se considera la base completa.
5. se satisfacen los supuestos del modelo en 1, 2 y 3. En caso negativo intente una transformación adecuada sobre la variable respuesta en cada modelo y revise nuevamente los supuestos.
6. Obtenga conclusiones acerca de dónde se observan las diferencias si las hubiere.

Ejercicio 3

1. Ajuste un modelo logístico para predecir el diagnóstico de cáncer de páncreas en función de las variables en la base que considere razonables.
2. Evalúe la calidad de ajuste del modelo con al menos dos criterios distintos.
3. Interprete los coeficientes del modelo elegido.