Fundamentos de Análisis de Datos

TP3

Tarea: Realizar un informe con los ejercicios solicitados. Entregar el informe en formato electrónico PDF y los programa en R, enviándolos por mail a mgambini@itba.edu.ar.

1. Considerar los datos del archivo acath.sav, el cual se lee con la función `load`. Se trata de una muestra de 3504 pacientes que acudieron al centro con dolor en el pecho, para los que se recogieron diversas variables cuyo nombre en la base de datos y descripción es la siguiente:

   - sigdz : variable binaria que toma valores 1 y 0, indicando si el paciente presenta estrechamiento de alguna de las arterias coronarias de al menos un 75 % (sigdz = 1) o no (sigdz = 0).
   - tvdlm : lo mismo que la anterior pero corresponde a tres arterias con estrechamiento.
   - sex: variable categórica que indica el género del paciente: 0 corresponde al género masculino, 1 al género femenino.
   - age: variable continua que representa la edad en años del individuo.
   - choleste: variable continua que expresa los Mg/dl de colesterol
   - duracion: variable continua que recoge la duración, en días, de los síntomas de la enfermedad coronaria.

   a) Realizar un modelo de regresión logística simple, siendo la variable explicada sigdz, considerando como única variable explicativa la variable colesterol. Escribir el modelo de regresión logística y calcular la probabilidad de que una persona tenga estrechamiento arterial si el colesterol es 199.

   b) Realizar un modelo de regresión logística considerando todas las variables no categóricas.

   c) Realizar el modelo anterior diferenciando mujeres y varones, o sea tomando la variable sex como factor.

d) Comentar los resultados de los modelos anteriores. Explicar los resultados que devuelve el comando `summary`.

2. Resolver el siguiente problema usando un modelo de ANOVA de factor fijo. El departamento de psicología de la Universidad de Tres de Febrero ha realizado un estudio sobre la frecuencia con que los alumnos asisten a clases teóricas no obligatorias, pertenecientes a tres localidades del partido, tomadas durante un cuatrimestre. Se han utilizado los datos que tomaron los profesores a algunos estudiantes y queremos conocer si existen diferencias estadísticamente significativas entre la cantidad estudiantes universitarios que asisten a clase dependiendo de la localidad a la que pertenecen. Los datos son los siguientes:

   Caseros: 11 14 7 15 11 13 11 16 10 15 18 12 9 9 10 10 15 10 14 10 10 12 14 12 15 7 13 6 10 15 20 10 13 10 6 14 8 10 8 11

   Santos Lugares: 13 10 12 7 5 10 10 16 9 7 7 2 6 9 9 8 8 10 3 6 5 2 9 3 4 5 10 8 5 9 10 8 13 10 0 2 1 1 0 4

   Pablo Podestá: 6 7 3 5 9 6 1 6 0 2 5 6 11 6 7 0 5 7 5 4 7 4 2 8 9 6 1 4 7 7 8 9 7 5 1 6 9 4 7 6

   a) Introducir los datos en R creando las 2 variables: una que incluya las frecuencias en la asistencia a clase y otra que sea un factor, que proporcione información sobre la localidad a la que pertenecen cada uno de los estudiantes.

   b) Analizar los datos de la muestra mediante gráficos y descriptivos. ¿Observa diferencias en los valores promedios por grupos?

   c) Realizar un test ANOVA para comparar las medias de las 3 poblaciones. ¿Cuáles serían las hipótesis nula y alternativa?

   d) Describir los resultados obtenidos indicando el valor del p-valor.

   e) Extraer conclusiones en el contexto del problema.

   f) Si se han obtenido diferencias significativas entre los grupos, determinar cuales son esas diferencias utilizando el test HSD de Tukey. Representar gráficamente las diferencias encontradas e intrepretar los resultados obtenidos.

3. El archivo de datos de R 'nottem' contiene las temperaturas promedio por mes en la ciudad de Nottinham desde 1920 a 1939

   a) Leer la serie como una serie de tiempo utilizando el comando ts.

   b) Informar las fechas y temperaturas iniciales y finales.

c) Realizar el gráfico de la serie.

d) Realizar un gráfico de la tendencia y un gráfico de cajas por ciclo para decidir si la serie posee estacionalidad.

e) Dividir los datos de la serie en dos subconjuntos uno de entrenamiento y otro de prueba, dejando el último año en otra serie. Realizar un modelo ARIMA y predecir las temperaturas para el año posterior a los informados en el conjunto de datos de entrenamiento. Calcular el error.