PRÁCTICA PUNTUABLE EN R

BIOESTADÍSTICA

El trabajo se desarrollará en grupos de 3 personas. Os debéis inscribir en uno de los grupos a través de la actividad "Grupos práctica puntuable en R" del campus virtual antes del 21 de mayo. Los aspectos que se van a valorar del trabajo son: el uso de R, la utilización de la metodología adecuada al problema y las conclusiones.

Conjunto de datos Wisconsin Diagnostic Breast Cancer (WDBC)

El conjunto de datos original está disponible en el repositorio UCI, en el enlace (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)

Contiene información de 569 enfermos de cáncer de mama invasivo y sin evidencia de metástasis en el momento del diagnóstico. Se recogen 30 variables calculadas a partir de una imagen digitalizada de una masa mamaria, que describen características de los núcleos celulares presentes en la imagen. Además, se recoge el diagnóstico (maligno o benigno) y la edad del paciente. Las variables se resumen en la siguiente tabla,

Variable	Descripción
id	Identificador del paciente.
age	Edad en años.
diagnosis	M = maligno, B = benigno.
radius	Distancias del centro a los puntos del contorno.
texture	Desviación estándar del nivel de gris de los pixeles en la imagen.
perimeter	Perímetro.
area	Área.
smoothness	Variación de las longitudes del radio.
compactness	perimeter^2 / area - 1.0
concavity	Profundidad de las secciones cóncavas del contorno.
concave points	Número de secciones cóncavas del contorno.
symmetry	Simetría.
fractal_dimension	Dimensión fractal.

Para cada característica de cada imagen se calculan la media (_mean), el error estándar (_se) y la media de los tres valores más grandes (_worst, el peor valor).

Instrucciones

1. Seleccionar aleatoriamente una muestra sin reemplazamiento de 50 individuos de cada diagnóstico (50 individuos con diagnosis = M y 50 individuos con diagnosis = B). Para poder reproducir los resultados debéis usar, dependiendo de vuestro grupo, la siguiente semilla para el generador de números aleatorios,

Grupo A	Grupo B	Grupo C	Grupo D	Grupo E	Grupo F	Grupo G	Grupo H	Grupo I
1872	4755	4178	1832	4744	2015	3919	34	4399
Grupo J	**Grupo K**	**Grupo L**	**Grupo M**	**Grupo N**	**Grupo E**	**Grupo O**	**Grupo P**	
2420	1555	4838	4607	571	1418	1634	4636	

Los siguientes apartados se llevarán a cabo con la muestra seleccionada.

2. Análisis descriptivo:
   a. Análisis descriptivo univariante de aquellas variables que vais a utilizar en los análisis posteriores. Debéis considerar, al menos, una variable de cada tipo. Se requiere resúmenes numéricos y representaciones gráficas adecuadas.

      Para las variables cuantitativas se deben incluir también conclusiones sobre su ajuste a la distribución normal, existencia o no de asimetría y conclusiones sobre apuntamiento.
   b. Análisis descriptivo bivariante que debe contener al menos un análisis cualitativa vs cualitativa, uno cualitativa vs cuantitativa y uno cuantitativa vs cuantitativa con los gráficos y tablas apropiados para cada caso y conclusiones sobre la posible existencia o no de asociación entre cada pareja.

      Nota: podéis categorizar alguna de las variables cuantitativas con la función `cut()`.

3. Análisis inferencial.

   En todos los apartados que siguen especificar con detalle las hipótesis de los test, los cálculos hechos y las conclusiones. Considerar una de las variables continuas del conjunto de datos en la que sea posible asumir normalidad. Asumiendo normalidad:
   a. ¿Entre qué valores se mueve la media de la distribución con una confianza del 95%? Suponer varianza desconocida. Interpretar los resultados.
   b. Hacer un contraste sobre la media de nivel 0.1 tomando como hipótesis nula el extremo superior del intervalo bilateral del apartado a. Dar el p-valor e interpretar el resultado.

   Considerar esa misma variable cuantitativa y una variable cualitativa con dos niveles. Suponiendo normalidad:
   c. Hacer un contraste de igualdad de varianzas a nivel 0.05. Dar el p-valor e interpretar el resultado.
   d. Hacer un contraste de comparación de medias con varianzas desconocidas, dar el p-valor e interpretar el resultado.
   e. Comprobar, a través de simulación, que la distribución en el muestreo del estadístico test utilizado en d) es la que se espera.

   Categorizar, en dos grupos, una de las variables recogidas en la imagen. Considerando la variable diagnosis:

f. Hacer un contraste chi-cuadrado. Dar el p-valor e interpretar el resultado.
g. Calcular una medida de asociación junto con un intervalo de confianza del 95%. Interpretar los resultados.

4. ANOVA y Regresión:
   a. Categorizar, en tres grupos, una de las variables cuantitativas medidas en las imágenes y llevar a cabo un ANOVA.
   b. Efectuar un análisis de regresión lineal simple de dos de las variables continuas del conjunto de datos.
      I. Dar la ecuación del modelo. Interpretar el modelo.
      II. Estudiar la significación del modelo y la bondad de ajuste.
      III. Hacer un análisis residual, incluir los gráficos apropiados y estudiar la adecuación del modelo.
      IV. Si el modelo no es apropiado tratar de encontrar transformaciones que corrijan el modelo obteniendo un modelo aceptable.
   c. A partir de la variable `diagnosis`, ajustar un modelo que nos permita predecir la probabilidad de tener un tumor maligno. El modelo debe incluir al menos 5 variables independientes. Identifica e interpreta factores de riesgo/protección.

Entrega

Se deben entregar dos ficheros:

- Fichero con el código R empleado para la ejecución del trabajo. Debe estar documentado con comentarios que permitan fácilmente su lectura. Este archivo debe tener extensión .txt.
- Informe, en formato pdf, con los resultados y conclusiones de las cuestiones detalladas en la sección anterior. Debéis incluir en el archivo un informe completo que incluya las salidas numéricas y gráficas obtenidas con R que comentéis y utilicéis para responder a las preguntas planteadas, junto con su interpretación.

Habrá dos actividades en Moodle para la entrega, una para cada uno de los archivos. Se verificará la originalidad de los archivos entregados con la herramienta turnitin. Basta con que un miembro del grupo suba los archivos al campus.

Fecha máxima de entrega: miércoles 2 de junio (incluido). No se aceptarán entregas posteriores a esa fecha.