

Problema 1 (3.5 puntos)

En el fichero **PRegFeb2010.sav** se encuentran datos relativos a partidos de la liga ACB de baloncesto. Se desea encontrar un modelo que explique los puntos por partido (**PUNTXPART**) que es capaz de anotar un jugador de baloncesto respecto a las siguientes variables regresoras: la altura del jugador (**ALTURA**), los minutos que juega por partido (**MINXPART**), los balones que pierde por partido (**BXPART**), las faltas personales cometidas por partido (**FPXPART**) y el porcentaje en tiros de campo (**POR _ TC**).

1. Las variables del problema, ¿presentan datos atípicos? ¿Podemos suponer que nuestra variable respuesta es Normal? Justifica tus respuestas **(0.3 pts)**
2. En el caso de que la variable respuesta no provenga de una distribución Normal, realiza una transformación adecuada y explica por qué has decidido dicha transformación. **(0.4 pts)**
3. Calcula la matriz de correlaciones de las variables del problema. ¿Qué regresores del modelo presentan una más estrecha relación lineal entre sí? (indica el valor que toma el coeficiente de correlación de Pearson). ¿Cuál es la primera variable que debería entrar en el modelo? (indica el valor que toma el coeficiente de correlación de Pearson). **(0.4 pts)**
4. Realiza la selección del modelo mediante regresión por pasos, hacia delante y hacia atrás. Indica, para cada uno de los tres métodos, el modelo teórico resultante. Estudia si los modelos obtenidos son reducibles (simplificables) y si presentan multicolinealidad. **(0.7 pts)**
5. ¿Qué modelo de regresión propondrías y por qué? Determina el modelo ajustado que explica los puntos por partido y comenta el valor de R^2 . **(0.4 pts)**
6. Para el modelo resultante del apartado anterior, guarda los residuos y los valores ajustados y estudia si se verifican las hipótesis del modelo de regresión múltiple y si existen observaciones influyentes, comentando los procesos utilizados. **(0.8 pts)**
7. Proporciona una estimación puntual de los puntos por partido que anotaría un jugador de altura 2.05m, con un porcentaje en tiros de campo de 0.55 y que ha jugado 18 minutos, que ha perdido 5 balones y ha cometido 3 faltas por partido. ¿Podemos asegurar que los puntos por partido de dicho jugador será superior a 4? ¿Y en promedio para los jugadores de esas características? **(0.5 pts)**

Problema 2 (2.5 puntos)

En el fichero **EPF2008_Cluster.sav** se encuentran los datos correspondientes a la Encuesta de Presupuestos Familiares del año 2008 (fuente: www.ine.es). Para la recopilación de datos se consideraron 12 grupos de gasto, que se describen a continuación. G1: Alimentos y bebidas no alcohólicas. G2: Bebidas alcohólicas, tabaco y narcóticos. G3: Artículos de vestir y calzado. G4: Vivienda, agua, electricidad, gas y otros combustibles. G5: Mobiliario, equipamiento del hogar y gastos corrientes de conservación de la vivienda. G6: Salud. G7: Transportes. G8: Comunicaciones. G9: Ocio, espectáculos y cultura. G10: Enseñanza. G11: Hoteles, cafés y restaurantes. G12: Otros bienes y servicios.

Cada una de las variables del fichero contiene el gasto medio anual por hogar para cada una de las 17 comunidades autónomas españolas, así como para Ceuta y Melilla. El objetivo del estudio es realizar una clasificación de las comunidades autónomas atendiendo a cómo distribuyen las familias su presupuesto económico anual.

1. Con los datos proporcionados en el fichero, ¿qué variable tendría mayor peso en la clasificación? ¿Y menor peso?. Justifica tus respuestas. **(0.25 pts)**
2. Realiza una clasificación de las comunidades autónomas usando tres métodos jerárquicos, justificando si conviene trabajar con los datos originales o tipificados. Comenta la medida de distancia utilizada en cada caso así como el tipo de enlace. **(0.75 pts)**

- En función de los dendogramas obtenidos en los distintos métodos jerárquicos, determina el número de grupos que harías y qué comunidades estarían en cada grupo. Intenta explicar qué caracteriza a las comunidades de cada grupo. **(0.75 ptos)**
- Usando el algoritmo de las k-medias, determina cuál sería el número de grupos adecuado estudiando cómo varían las sumas de cuadrados dentro de grupos (SCDG). Una vez determinado el número de grupos a realizar, clasifica las comunidades autónomas usando el algoritmo de las k-medias e interpreta los resultados. **(0.75 ptos)**

Problema 3 (4 puntos)

En el fichero **ferrocarril_ST.sav** se encuentran los datos correspondientes al número de viajeros (en miles) que usaron el ferrocarril como transporte interurbano, recopilados mensualmente desde enero de 1996 hasta noviembre de 2009 (fuente: www.ine.es).

- Representa los datos del número de viajeros en un gráfico temporal. ¿La serie presenta Estacionalidad? **(0.2 ptos)**
- Justifica si se trata de un modelo aditivo o multiplicativo. Para confirmarlo, realiza un gráfico de desviaciones típicas frente a medias para cada año (representalo en tu folio a mano alzada). **(0.2 ptos)**
- Extrae las componentes de la serie (Tendencia-Ciclo, Estacionalidad e Irregular). Indica los 3 primeros valores de cada componente y comenta la forma que presenta la Tendencia de la serie **(0.3 ptos)**
- Las fluctuaciones de la serie dentro de un mismo año, ¿se deben únicamente a la componente estacional? Al desestacionalizar la serie, ¿en qué medida hemos reducido las fluctuaciones anuales?. Responde en términos de la desviación típica. **(0.25 ptos)**
- Proporciona un modelo determinista que nos permita realizar predicciones, **justificando qué curva has seleccionado para representar la tendencia. (0.5 ptos)**. Expresa el modelo de la forma:

$$VIAJEROS_ESTIMADO = \dots$$

- Según el modelo propuesto en el apartado anterior, ¿qué mes presenta mayor (menor) error de predicción en promedio? **(0.3 ptos)**
- Determina las predicciones del número de viajeros hasta diciembre de 2010. (Indica sólo las predicciones para los meses de enero, agosto y diciembre). **(0.5 ptos)**

*Nos proponemos ahora analizar la serie del fichero **ferrocarril_ST.sav** mediante una técnica de alisado exponencial.*

- ¿Qué técnica de alisado exponencial te parece más adecuada para analizar esta serie? Razona tu respuesta. **(0.25 ptos)**
- Aplica la técnica de alisado que has considerado adecuada, seleccionando como parámetros de alisado aquellos que minimizan la suma de cuadrados de los errores (búsqueda en rejilla con precisión de 0.1). Determina cómo quedarían las fórmulas recurrentes de las series alisadas e interpreta el significado de los parámetros de alisado obtenidos. **(0.5 ptos)**
- Queremos determinar qué técnica predictiva es más adecuada para la serie en estudio: alisado exponencial o el método ingenuo I (véase formulario). Para ello, calcula la suma de cuadrados de los residuos para las dos técnicas. **(0.5 ptos)**
- Determina las predicciones del número de viajeros hasta diciembre de 2010. (Indica sólo las predicciones para los meses de enero, agosto y diciembre). Si comparamos estas predicciones con las obtenidas por el método clásico, ¿qué podemos destacar? **(0.3 ptos)**
- Queremos que nuestro método de alisado de aún más importancia al cambio de tendencia que se produce al final de la serie, ¿cómo deberíamos seleccionar los parámetros de alisado?. Justifica tu respuesta. **(0.2 ptos)**