Dpto. Matemática Aplicada y Estadística

Titulación**: Ingeniero en Organización Industrial**
Asignatura: **Estadística Industrial**
Curso: **2006/2007**
EXAMEN BLOQUE I: Análisis Multivariante

MODELO A

NOTA: Todas las respuestas deben ser razonadas para que sean puntuadas. Esto no supone explicar el desarrollo de ningún procedimiento, salvo que se indique de forma expresa.

Problema 1 (3.5 puntos)

El fichero **PRegresionEx_A.sav**, contiene los datos correspondientes a la presencia (en %) de cinco componentes químicos en un tipo de cemento, así como el calor emitido (en calorías por gramo de cemento) durante el proceso de endurecimiento. Se desea proponer un modelo que permita predecir el calor emitido en función de los componentes químicos presentes en el cemento.

1. Indica la variable respuesta y los regresores del problema.
2. Las variables del problema, ¿presentan datos atípicos? ¿Qué entiendes por dato atípico? ¿Qué método gráfico permite detectar atípicos? ¿Deben eliminarse automáticamente los datos atípicos del estudio? **(0.25 ptos)**
3. ¿Qué métodos permiten determinar la normalidad de los datos? ¿Podemos suponer que nuestra variable respuesta es Normal? Aplica las transformaciones de Box-Cox para determinar si crees adecuado NO transformar la variable respuesta. **(0.5 ptos)**
4. Calcula la matriz de correlaciones de las seis variables. ¿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). **(0.5 ptos)**
5. 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. **(0.5 ptos)**
6. Teniendo en cuenta el apartado anterior, ¿qué modelo(s) teórico(s) propondrías y por qué? **(0.25 ptos)**
7. Determina el modelo ajustado que contiene a los regresores A y B y comenta el valor del R^2 (coeficiente de determinación). ¿Para qué crees que puede servir el modelo obtenido? **(0.25 ptos)**
8. Estudia si el modelo anterior presenta observaciones influyentes (comenta los criterios utilizados). Realiza un análisis de los residuos para estudiar si se verifican las hipótesis del modelo de regresión múltiple (comenta brevemente los procesos utilizados). **(0.75 ptos)**
9. Obtén una estimación puntual del calor emitido por el cemento sabiendo que $A = 15$ y $B = 30$. Determinar también un intervalo de confianza para el calor emitido en ese caso, así como un intervalo de predicción. ¿Podemos asegurar que el calor emitido por el cemento superará las 95 cal/gr? ¿Y en promedio? **(0.5 ptos)**

Problema 2 (2.5 puntos)

En el fichero **PClusterEx_A.sav** se encuentran los datos correspondientes a la encuesta de Presupuestos Familiares del año 1990/91 para 17 provincias españolas. Las variable consideradas son: X1= alimentación, X2 = vestido y calzado, X3 = vivienda, X4= mobiliario doméstico, X5= gastos sanitarios, X6= transporte, X7= enseñanza y cultura, X8= turismo y ocio, X9 = otros gastos.

1. Realiza una clasificación de las provincias usando al menos **tres** métodos jerárquicos. Comenta la medida de distancia utilizada en cada caso así como el tipo de enlace. **(1 pto)**

2. En función de los dendogramas obtenidos en los distintos métodos jerárquicos, indica cómo agruparías a las 17 provincias españolas. Intenta explicar qué caracteriza a las provincias de cada grupo. **(0.75 ptos)**

3. Clasifica las provincias en 3 grupos usando el algoritmo de las k-medias. Determina si es más adecuado realizar sólo 2 grupos estudiando cómo varían las sumas de cuadrados dentro de grupos (SCDG). **(0.75 ptos)**

Problema 3 (4 puntos)

En el fichero **PFactorialEx_A.sav** se encuentran los datos correspondientes a la contaminación del aire en 80 ciudades de US para el año 1960. Las variables se definen a continuacion. TMR: Tasa de mortalidad, SMEAN: Media arimetica de cantidades de sulfato, SMAX: Cantidades maximas de sulfato cada dos semanas, PMEAN: Media aritmetica de particulas suspendidas, PMAX: Cantidad maxima de particulas suspendidas, PM2: Densidad poblacional por milla cuadrada, GE65: Porcentaje poblacional con 65 años o más, PERWH: Porcentaje de Blancos en la poblacion, NONPOOR: Porcentaje de familias con ingresos sobre el nivel de pobreza, LPOP: Logaritmo (base 10) de la poblacion

Se pretende aplicar una técnica de análisis mulitvariante que permita explicar la variabilidad de las ciudades en estudio respecto a las 10 variables del problema, usando un número menor de variables y de manera que se pierda la mínima información posible.

1. Explica razonadamente qué tipo de análisis parece más adecuado para este estudio. **(0.25 ptos)**

2. Las variables del problema no se miden en las mismas unidades, ¿puede esto afectar a nuestro estudio? ¿qué propones hacer con los datos del problema? **(0.25 ptos)**

3. En primer lugar, calcula las 10 componentes principales del problema a partir de la matriz de correlaciones, indicando los valores y vectores propios de la matriz de correlaciones (**indica sólo el primer vector propio y el último, así como la primera y última componentes principales**) **(0.5 ptos)**

4. ¿Con cuántas componentes principales te quedarías? Justifica tu respuesta comentando los distintos métodos de selección y los resultados obtenidos con SPSS. **(0.5 ptos)**

5. Calcula las coordenadas que tendría la ciudad de DENVER respecto a las 4 primeras componentes principales. Explica cómo has obtenido dichas coordenadas **(0.25 ptos)**

Por otra parte, se pretende estudiar si la estructura de dependencia entre las variables del problema se puede resumir a partir de unos pocos factores.

1. Explica razonadamente qué tipo de análisis parece más adecuado para este estudio. **(0.25 ptos)**

2. Determina si el análisis propuesto es adecuado en función de los datos muestrales. Indica qué procedimientos permiten responder a esta cuestión, explica en qué consisten y comenta los resultados obtenidos con SPSS. **(0.5 ptos)**

3. Lleva a cabo el análisis propuesto, respondiendo a las siguientes cuestiones y comentando los procedimientos utilizados en cada caso:

   (a) Determina el número de factores a retener. **(0.25 ptos)**

   (b) Obtén la matriz de cargas factoriales e indica el Modelo de Análisis Factorial propuesto (**indica al menos la primera y última fila del modelo**) ¿Qué método has utilizado para obtener las cargas factoriales? Comenta las ventajas e inconvenientes de este método frente a otros métodos que también permiten calcular las cargas factoriales. **(0.75 ptos)**

   (c) ¿Cómo podemos facilitar la interpretación de los factores retenidos? Aplícalo e interpreta los factores resultantes. **(0.5 ptos)**

   (d) Compara las correlaciones observadas con las reproducidas y determina si te parece adecuado el modelo factorial obtenido. ¿Qué representan las comunalidades? ¿Te parecen adecuadas en nuestro modelo? **(0.5 ptos)**