

## Práctica 2. Regresión Lineal.

1. N.H. Prater desarrolló una ecuación de regresión para estimar la producción de gasolina como una función de las propiedades de destilación de cierto tipo de petróleo crudo. Se identificaron cuatro variables de predicción: la graduación del petróleo crudo, °API ( $x_1$ ); la presión de vapor del petróleo crudo, psi ( $x_2$ ); el punto de 10% ASTM para el petróleo crudo, °F ( $x_3$ ) y el punto final ASTM para la gasolina, °F ( $x_4$ ). Los dos primeros miden la graduación y la presión de vapor del petróleo crudo. El punto de 10% ASTM es la temperatura para la cual se ha evaporado cierta cantidad de líquido, y el punto final para la gasolina es la temperatura para la cual se ha evaporado todo el líquido. La variable respuesta ( $y$ ) fue la cantidad de gasolina producida expresada como un porcentaje respecto al total de petróleo crudo. Los datos de laboratorio obtenidos por Prater se muestran en el archivo PRATER.DAT. Determinar una ecuación de regresión para la producción de gasolina como una función lineal de las propiedades de destilación de cierto tipo de petróleo crudo  $x_1$ ,  $x_2$ ,  $x_3$  y  $x_4$ . ¿Podemos eliminar alguna de las variables predictoras del modelo?
2. Se realiza un experimento para determinar el calor desarrollado en la fabricación de cemento en función de los porcentajes de 4 compuestos activos que se utilizan en la fabricación. Para ello se elige una muestra de 13 cementos, midiéndose el calor desarrollado de cal/g ( $y$ ) y los porcentajes de los compuestos referidos ( $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$ ). Los datos se encuentran en el archivo CEMENTO.DAT. Calcula la ecuación pedida, investigando la bondad del ajuste y eliminando de modo razonado alguna de las variables predictoras del modelo, si ello fuera necesario.
3. Una compañía observa que la demanda de uno de sus productos cambió debido a una variación rápida de su precio por unidad. Supóngase que la demanda  $Y$  del producto se observa en una región en particular sobre un intervalo bastante amplio de precios  $x$ . Dados los datos que se encuentran en COMPANY.DAT, determinar el grado de un polinomio que mejor ajuste estos datos.
4. Una agencia desea estimar los gastos en alimentación de una familia con base en el ingreso y su tamaño. Los datos que se encuentran en el archivo FAMILIA.DAT representan los gastos de alimentación por mes ( $y$ ) en miles de dólares, contra el ingreso mensual ( $x_1$ ) y el tamaño de la familia ( $x_2$ ), para 15 familias que se seleccionaron al azar en cierta localidad.
  - a) Ajustar todos los modelos lineales que abarcan a  $x_1$  y/o  $x_2$ , e interpretar los coeficientes de regresión estimados.
  - b) Contrastar la hipótesis  $H_0 : \beta_1 = \beta_2 = 0$  en el modelo completo.
  - c) ¿Podemos decir que alguna de las variables  $x_1$ ,  $x_2$  no es necesaria para la predicción de  $y$ ?
5. En ciertos organismos gubernamentales y compañías privadas, el problema de identificar aquellos factores que son importantes para predecir la aptitud para el trabajo de los aspirantes a obtener un empleo constituye un proceso continuo. El procedimiento usual es el de aplicar al solicitante un conjunto de pruebas apropiadas y tomar la decisión de contratarlo o no con base en los resultados de éstas. El asunto clave es conocer a priori qué pruebas pueden predecir la aptitud para el trabajo de una persona. Suponer que el personal de una compañía muy grande ha desarrollado cuatro pruebas para una determinada clasificación con respecto al trabajo. Estas pruebas se aplicaron a 20 individuos que fueron contratados por la compañía. Después de un periodo de dos años, cada uno de estos empleados se clasifica de acuerdo con su aptitud para el trabajo. La puntuación para la aptitud hacia el trabajo ( $y$ ) y la correspondiente a cada una de las cuatro pruebas ( $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$ ) se dan en el archivo TRABAJO.DAT.
  - a) Hallar la ecuación de regresión lineal de  $y$  sobre  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$ .
  - b) Hallar la tabla de análisis de la varianza mostrando las comparaciones parciales.
  - c) ¿Es posible eliminar alguna de las pruebas en el proceso de selección?
6. De manera reciente, se ha dirigido el interés hacia el desarrollo de métodos más rápidos y económicos para vigilar la concentración de sedimentos y contaminantes en los recursos acuíferos de cierta nación.

Para los encargados de vigilar el medio ambiente, el interés principal recae en la necesidad de cuantificar los valores de concentración en el agua con base en datos de percepción remota. El uso de las técnicas de percepción remota para vigilar distintos parámetros que miden la calidad del agua parece ser prometedor. Un tipo de sistema de percepción remota, la variedad pasiva, mide el flujo total de radiación emitido por el sistema agua-atmósfera. Una componente muy grande de este flujo de radiación es el flujo de luz emitido por el agua, el cual, bajo condiciones normales, es una función de los constituyentes que se encuentran presentes en el agua. Para medir el espectro de esta radiación se encuentran disponibles un gran número de sistemas de rastreo multiespectral. Sin embargo, cada sistema tiene diferentes localizaciones de las bandas y anchos diferentes.

Se piensa que un cambio en la concentración de un contaminante causará un cambio en el valor del flujo de radiaciones, es decir, si se conocen los valores de la radiación para diferentes bandas espectrales, entonces es posible predecir la concentración de un contaminante en una fuente de agua dada. El problema reside en el hecho de identificar, de entre todas las bandas, cuál es la que puede predecir la concentración del contaminante. En un laboratorio, se obtuvieron datos reales de percepción remota, bajo condiciones controladas, que empleó cinco bandas y varios constituyentes, entre ellos el sedimento del feldespato. Los datos de la muestra se proporcionan en el archivo ACUIF.DAT.

- a) Empléense las cinco bandas ( $x_1, x_2, x_3, x_4, x_5$ ) como variables de predicción y las concentraciones de feldespato ( $y$ ) como la respuesta, para ajustar un modelo de regresión lineal.
  - b) Calcúlese la matriz de correlación para las cinco bandas de radiación. Interpretese el resultado.
  - c) Determinar el mejor conjunto de variables de predicción.
7. Los datos que se encuentran en el archivo TEMPERA.DAT representan la temperatura atmosférica promedio  $Y$  en enero para 50 estaciones climatológicas situadas en el estado de Virginia, donde cada estación se identifica por medio de su latitud  $x_1$ , longitud  $x_2$  y altitud  $x_3$ .
- a) Ajustar dichos datos por un modelo de regresión de segundo orden completo y analizar los resultados obtenidos.
  - b) ¿Deben retenerse todos los términos que aparecen en la ecuación anterior? Justificar la respuesta.
8. Se toman datos de 209 ordenadores con el objeto de predecir el rendimiento de la cpu ( $y$ ) en función de una serie de variables. El rendimiento se mide tomando como punto de referencia la cpu de un determinado modelo de la firma IBM. Las variables predictoras son el tiempo de cada ciclo en nanosegundos ( $x_1$ ), la memoria principal máxima ( $x_2$ ), la memoria principal mínima ( $x_3$ ), ambas en kilobytes, el tamaño de la memoria caché ( $x_4$ ), también en kilobytes, el número mínimo ( $x_5$ ) y máximo ( $x_6$ ) de canales. Los datos se encuentran en el archivo CPUS.DAT.
- a) Obtener una ecuación que permita predecir el rendimiento en función de las variables anteriormente citadas.
  - b) Determinar la bondad del ajuste.
  - c) Determinar el grupo de variables que resultan significativas mediante un procedimiento de entrada o salida.