

Especialidad en Métodos Estadísticos

MIII - Métodos Estadísticos Básicos

Periodo: Ago – Dic 2021

Profesor: M. en C. Paul Ramírez De la Cruz

Tarea 1. Análisis de Regresión

Lineal Simple

Puntaje total: 10 puntos

Fecha de entrega:

Nombre: \_\_\_\_\_

Calificación: \_\_\_\_\_

## 1. Instrucciones. Resuelva los siguientes ejercicios de regresión simple haciendo uso de R

1. [Ejercicio suplementario 11.57 Mendenhall] Un nuevo paquete de software de consulta se diseñó con el objetivo de lograr un acceso y un mantenimiento más eficientes de los conjuntos de datos a gran escala. La eficiencia se mide en términos del número de operaciones de entrada/salida (E/S) de disco (llamadas *bloques de almacenamiento*) necesarias para acceder al conjunto de datos y darle mantenimiento; cuanto menor sea el número de bloques leídos, con mayor rapidez se efectuará la operación. A fin de evaluar el desempeño del nuevo sistema de software, se registró el número de operaciones de E/S de disco necesarias para acceder a un conjunto de datos a gran escala, para cada uno de una muestra de 15 conjuntos de datos de diversos tamaños (donde el tamaño se mide como el número de registros contenidos en el conjunto de datos, en miles)

Registros	350	200	450	50	400	150	350	300	150	500	100	400	200	50	250
E/S disco	36	20	45	5	40	18	38	32	21	54	11	43	19	7	26

- a. Elabore la gráfica de dispersión. ¿Parece existir una relación lineal entre las variables?
- b. Estime puntualmente los coeficientes del modelo de regresión lineal simple e interpréte los en el contexto del problema. Grafique la recta de regresión sobre la dispersión.
- c. Proporcione una estimación por intervalo del 90% de confianza para  $\beta_0$ . Interprete el resultado
- d. Proporcione una estimación por intervalo del 90% de confianza para beta  $\beta_1$ . Interprete el resultado
- e. Calcule el valor de  $r$  y  $r^2$  e interprete ambos resultados. ¿Corresponden con lo observado en el inciso a)?
- f. Estime puntualmente el número medio de operaciones de E/S que se tendría si se procesa un archivo de 444 (miles de) registros
- g. Estime con un intervalo de confianza de 95% el número medio de operaciones de E/S (en miles) que se tendría si se procesa un archivo de 444 (miles de) registros
- h. Estime con un intervalo de confianza de 95% el número observado de operaciones de E/S (en miles) que se tendría si se procesa un archivo de 444 (miles de) registros
- i. Elabore para los residuos una gráfica de dispersión contra  $X$ , un histograma, una gráfica de cuantiles normales (probabilidad normal) y una gráfica secuencial y analícelas para

establecer si presentan violaciones claras a los supuestos del modelo de regresión lineal simple

2. Las bodegas modernas utilizan vehículos guiados computarizados y automatizados para el manejo de materiales. En consecuencia, la disposición física de la bodega debe diseñarse con cuidado a modo de evitar el congestionamiento de los vehículos y optimizar el tiempo de respuesta. En *The Journal of Engineering for Industry* (agosto de 1993) se estudio el diseño óptimo de una bodega automatizada. La disposición empleada supone que los vehículos no se bloquean entre sí cuando viajan dentro de la bodega, es decir, que no hay congestionamiento. La validez de este supuesto se verificó simulando (en una computadora) las operaciones de la bodega. En cada simulación se varió el número de vehículos y se registró el tiempo de congestionamiento (el tiempo total que un vehículo bloqueó a otro). Los datos se muestran en la siguiente tabla. Algo que interesa a los investigadores es la relación entre el tiempo de congestionamiento ( $Y$ ) y el número de vehículos ( $X$ ).

Número de vehículos	1	2	3	4	5	6	7	8
Tiempo de congestionamiento en minutos	0	0	0.02	0.01	0.01	0.01	0.03	0.03

Número de vehículos	9	10	11	12	13	14	15
Tiempo de congestionamiento en minutos	0.02	0.04	0.04	0.04	0.03	0.04	0.05

- a. Construya un diagrama de dispersión para los datos
  - b. ¿Existe evidencia de que el número de vehículos explica la variación en el tiempo de congestionamiento? Realice el contraste de hipótesis del Análisis de Varianza con  $\alpha = 0.05$ .
  - c. Encuentre la línea de mínimos cuadrados que relaciona el número de vehículos ( $X$ ) con el tiempo de congestionamiento ( $Y$ )
  - d. Trace la línea de mínimos cuadrados en la gráfica del inciso a)
  - e. Interprete los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$
3. [Milton. Sección 11.3 Prob. 16 y ss.] En los problemas de flujo de producción en las fábricas, el rendimiento suele evaluarse con base en el tiempo de producción mínimo, que es el total transcurrido desde el inicio de la primera tarea en la primera máquina hasta que se completa la última tarea en la última máquina. En un flujo de trabajo específico, este parámetro se evalúa en relación con el número de tareas realizadas. Sea  $X$  una variable independiente que denota el número de tareas, y  $Y$ , la variable dependiente que denota el tiempo de producción (en unidades estandarizadas):

Número de tareas ( $X$ )	4	5	6	7	8	9	10	11	12	13	14	15
Tiempo de producción ( $Y$ )	3.75	4.90	4.88	7.20	7.30	9.10	9.00	11.90	11.50	14.10	13.90	17.50

- a. Elabore el contraste de hipótesis del Análisis de Varianza para establecer si el modelo es útil para explicar la variación en el tiempo de producción.
- b. Elabore las gráficas de residuos y analícelas por si encuentra alguna violación grave a los supuestos del modelo

4. La Comisión Federal de Comunicaciones (FCC, por sus siglas en inglés) de Estados Unidos especifica que las emisiones electromagnéticas radiadas por dispositivos digitales se deben medir en un sitio de prueba de campo abierto. A fin de verificar la aceptabilidad de un sitio de prueba es preciso evaluar la atenuación del sitio (es decir, la pérdida de transmisión desde un dipolo de media onda hasta la salida de otro cuando ambos dipolos se colocan sobre el plano de tierra). Un estudio realizado en un sitio de prueba en Fort Collins, Colorado, produjo los siguientes datos de atenuación del sitio (en decibeles) y frecuencia de transmisión (en megahertz) para dipolos a una distancia de 3 metros.

Frecuencia de transmisión, $x$ , MHz	50	100	200	300	400	500
Atenuación del sitio, $y$ , dBL	11.5	15.8	18.2	22.6	26.2	27.1

Frecuencia de transmisión, $x$ , MHz	600	700	800	900	1000
Atenuación del sitio, $y$ , dBL	29.5	30.7	31.3	32.6	34.9

- Elabore un diagrama de dispersión para los datos. ¿Es evidente una relación lineal entre  $x$  y  $y$ ?
  - Encuentre la línea de mínimos cuadrados que relaciona la atenuación del sitio  $y$  con la frecuencia de transmisión  $x$ .
  - Trace la línea de mínimos cuadrados sobre el diagrama de dispersión para verificar sus cálculos.
  - Interprete los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ .
5. [Milton. Sección 11.1. Ejercicio 7] Se estudia la relación del consumo de energía eléctrica con el ingreso familiar y se obtienen los datos siguientes sobre el ingreso familiar  $X$  (en unidades de \$1,000/año) y el consumo de energía  $Y$  (en  $10^8$  Btu/año):

Consumo de energía ( $Y$ )	1.8	3.0	4.8	5.0	6.5	7.0	9.0	9.1
Ingreso familiar ( $X$ )	20.0	30.5	40.0	55.1	60.3	74.9	88.4	95.2

- Realice un diagrama de dispersión de los datos.
  - Estime la ecuación de regresión lineal  $y = \hat{\beta}_0 + \hat{\beta}_1 x$ .
  - Si  $x = 50$  (ingreso familiar de \$50,000), estime el consumo promedio de energía de las familias con ese ingreso. ¿Cuál sería su estimación para una sola familia?
  - ¿Cuánto esperaría que cambie el consumo si el ingreso familiar aumenta en \$2,000 al año?
  - Obtenga un intervalo de confianza del 98% para el valor de  $\theta_1$ .
6. Se ha utilizado con éxito un sistema automatizado para calificar grandes cantidades de programas para computadora escritos por estudiantes de la McMaster University de Ontario, Canadá, llamado AUTOMARK, el que toma en cuenta tanto la corrección como el estilo de programación al calificar las tareas de los estudiantes. A fin de evaluar la efectividad del sistema automatizado, se usó AUTOMARK para calificar las tareas, escritas en FORTRAN77, de un grupo de 33 estudiantes (*Communication of the ACM*, febrero de 1986). Estas calificaciones se compararon después con las calificaciones asignadas por el profesor. Los resultados se muestran en la tabla.

Calificación AUTOMARK, $x$	12.2	10.6	15.1	16.2	16.6	16.6	17.2	17.6	18.2	16.5	17.2
Calificación del profesor, $y$	10	11	12	12	12	13	14	14	14	15	15

Calificación AUTOMARK, $x$	18.2	15.1	17.2	17.5	18.6	18.8	17.8	18.0	18.2	18.4	18.6
Calificación del profesor, $y$	15	16	16	16	16	16	17	17	17	17	17

Calificación AUTOMARK, $x$	19.0	19.3	19.5	19.7	18.6	19.0	19.2	19.4	19.6	20.1	19.2
Calificación del profesor, $y$	17	17	17	17	18	18	18	18	18	18	19

- Obtenga una ecuación de mínimos cuadrados para el modelo de línea recta que relaciona la calificación del profesor con la calificación de AUTOMARK  $x$ .
- ¿Hay pruebas suficientes que indiquen que el modelo resulta útil para predecir  $y$ ? Pruebe con  $\alpha = 0.10$  usando el contraste de hipótesis T.
- Calcule un intervalo de predicción de 95% para la calificación asignada por el profesor a una tarea de FORTRAN77 que recibió una calificación de AUTOMARK de 17.5. Interprete el intervalo

## Bibliografía

- Lind, Marchal & Mason.** *Estadística Para Administración Y Economía*. Alfaomega. México, 2004.
- Mendenhall, William & Sincich, Terry.** *Probabilidad y estadística para ingeniería y ciencias*. Prentice-Hall Hispanoamericana. México, 1995
- Milton, J. Susan & Arnold, Jesse C.** *Probabilidad y estadística con aplicaciones para ingeniería y ciencias computacionales*. McGraw-Hill. México, 2004