Trimestre Enero-Marzo 2008
Departamento de Cómputo Científico y Estadística
Guía de ejercicios. Regresión Lineal Múltiple y ANOVA
Práctica N° 7

CONTENIDO
- Ajuste del modelo lineal mediante matrices.
- Propiedades de los estimadores de mínimos cuadrados.
- Inferencia respecto a los parámetros
- Predicción.
- Comparación de modelos.
- ANOVA de un factor.

1. Se sometieron 20 pacientes diabéticos a un estudio en el que se les midió el porcentaje de calorías obtenidas de los carbohidratos complejos, Y, en función de la edad, X_1, el peso, X_2, y el porcentaje de calorías obtenidas por el consumo de otros alimentos, X_3. Se ajusta el modelo: $Y_i = \beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\epsilon_i$, los resultados fueron:

$$\begin{aligned} \widehat{\beta} &= \begin{pmatrix} 36.96 \\ -0.11 \\ -0.23 \\ 1.26 \end{pmatrix}, \quad \left(X^t X\right)^{-1}_{ii} = \begin{pmatrix} 4.8158 \\ 0.0003 \\ 0.0002 \\ 0.0114 \end{pmatrix} \\ Y^t Y &= 29.368, \quad \widehat{\beta}^t X^t Y = 28800.34 \end{aligned}$$

a) Calcule los intervalos de confianza del 95% de β_1 y β_3.

b) Se supone que la edad no interviene, así que el modelo correspondiente resulta:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

y los resultados son: $\widehat{\beta}^t X^t Y = 28761.98$. Realice una prueba F para contrastar la hipótesis $\beta_1 = 0$ vs. $\beta_1 \neq 0$.

2. La siguiente tabla proporciona la latitud en grados (L), la altura en metros (A) y la temperatura anual (T) de 6 capitales marítimas españolas:

Cuidad	L	A	T
Gijón	43,4	22	13,9
Vigo	43,2	45	14,9
Barcelona	41,3	95	16,4
Valencia	39,5	24	17,2
Almería	36,8	7	18
Cádiz	36,5	30	18

a) Construir e interpretar un modelo que explique la temperatura en función de las otras variables.

b) Calcular el R^2 y la suma de cuadrados de los errores.

c) Predecir la temperatura media de una ciudad cuya latitud es 40,5 y su altura es 50 mm.

3. Se realizó un experimento para determinar si se podía predecir el peso de un animal después de un periodo dado, sobre la base de su peso inicial y la cantidad de alimento que había consumido. Se registraron los siguientes datos, en kilogramos:

Peso final (y)	Peso inicial (x_1)	Peso del alimento (x_2)
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

a) Ajuste el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

b) Prediga el peso final de un animal que tenía un peso inicial de 35 kilogramos y consumió 250 kilogramos de alimento.

4. Para estudiar la relación entre la variable y y tres variables x_1, x_2 y x_3 se toman 20 observaciones, obteniéndose:

$$X^tX = \begin{pmatrix} 20 & 11 & 8 & 9 \\ 11 & 7 & 4 & 5 \\ 8 & 4 & 4 & 4 \\ 9 & 5 & 4 & 6 \end{pmatrix}, \quad \left(X^tX\right)^{-1} = \begin{pmatrix} 1 & -1 & -1 & 0 \\ -1 & 1.4 & 0.8 & -0.2 \\ -1 & 0.8 & 1.85 & -0.4 \\ 0 & -0.2 & -0.4 & 0.6 \end{pmatrix},$$

$$X^tY = \begin{pmatrix} 327 \\ 210 \\ 138 \\ 130 \end{pmatrix}, \qquad Y^tY = 7950, \quad \overline{y} = 16.3$$

a) Estimar el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

indicar si los parámetros estimados son significativos mediante el contraste de la t de Student y calcular el R^2.

b) Estimar el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

indicar si los parámetros estimados son significativos mediante el contraste de la t de Student y calcular el R^2.

c) A la vista de los resultados anteriores, indicar qué modelo es mejor.

5. Se llevó a cabo un estudio sobre el uso de cierto rodamiento y y su relación con

$$\begin{aligned} x_1 &: \textit{viscosidad del aceite} \\ x_2 &: \textit{carga} \end{aligned}$$

Se obtuvieron los siguientes datos.

y	x_1	x_2
193	1.6	851
172	22.0	1058
113	33.0	1357
230	15.5	816
91	43.0	1201
125	40.0	1115

a) Ajuste el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

b) Prediga el uso para una viscosidad del aceite de 20 y una carga de 1200.

c) Estime σ^2 usando regresión múltiple de y sobre x_1 y x_2.

d) Calcule un intervalo de confianza de 95% para la media del uso y un intervalo de predicción de 95% para el uso observado, si $x_1 = 20$ y $x_2 = 1000$.

6. Se llevó a cabo un experimento para analizar el efecto de cuatro factores, *temperatura* T_1, *presión* P, *catálisis* C y *temperatura* T_2, en la producción Y de un químico.

a) Los valores (o niveles) de los cuatro factores empleados en el experimento aparecen en la siguiente tabla. Si cada uno de los cuatro factores se codifica para generar las cuatro variables x_1, x_2, x_3 y x_4, respectivamente, indique la transformación que relaciona cada variable codificada con su original correspondiente.

T_1	x_1	P	x_2	C	x_3	T_2	x_4
50	-1	10	-1	1	-1	100	-1
70	1	20	1	2	1	200	1

b) Ajuste el modelo lineal

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

a la siguiente tabla de datos.

				x_4 = +1		x_4 = −1	
				x_3		x_3	
				−1	1	−1	1
x_1	−1	x_2	−1	22.2	24.5	24.5	25.9
			1	19.4	24.1	25.2	28.4
	+1	x_2	−1	22.1	19.6	23.5	16.5
			1	14.2	12.7	19.3	16.0

c) ¿Proporcionan estos datos suficiente evidencia que indique que T_1 aporta información para la estimación de Y? ¿Lo hace P? ¿Lo hace C? ¿Lo hace T_2? (Pruebe las hipótesis, respectivamente, de que $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$ y $\beta_4 = 0$) Encuentre límites para el valor p asociado con cada prueba. ¿Qué concluiría usted si utiliza $\alpha = 0.01$ en cada caso?

7. Los datos que contiene la tabla siguiente provienen de la comparación de la tasa de crecimiento de bacterias tipo A y tipo B. En la tabla se muestra el crecimiento Y registrado en cinco puntos de tiempo equidistantes (y codificados).

	Tiempo				
Tipo de bacteria	-2	-1	0	1	2
A	8.0	9.0	9.1	10.2	10.4
B	10.0	10.3	12.2	12.6	13.9

a) Ajuste el modelo lineal

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

a los $n = 10$ datos. Sea $x_1 = 1$ si el dato corresponde a la bacteria tipo B y $x_1 = 0$ si corresponde al tipo A. Sea $x_2 =$ tiempo codificado.

b) Localice los datos en el plano y trace una gráfica de las dos rectas de crecimiento. Observe que β_3 es la diferencia entre las pendientes de las dos rectas y representa la interacción tiempo-bacteria.

c) Prediga el creciemiento de bacterias tipo A en el tiempo $x_2 = 0$ y compare su resultado con la gráfica. Repita el proceso para las bacterias tipo B.

d) ¿Presentan los datos evidencia suficiente que indique una diferencia en las tasas de crecimiento para los dos tipos de bacterias?

e) Obtenga un intervalo de confianza de 90% para el crecimiento esperado de las bacterias tipo B en el tiempo $x_2 = 1$.

f) Encuentre un intervalo de prediccioón de 90% para el crecimiento Y de las bacterias tipo B en el tiempo $x_2 = 1$.

8. Se propuso el siguiente modelo para probar si existía evidencia de discriminación salarial en contra de las mujeres en una universidad estatal:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2 + \epsilon$$

donde Y = salario anual (en miles de dólares)

$$\begin{aligned} x_1 &= \begin{cases} 1, & \text{si es mujer} \\ 0, & \text{si es hombre} \end{cases} \\ x_2 &= \text{experiencia (en años)} \end{aligned}$$

Al ajustar este modelo a los datos que se obtuvieron de los expedientes de 200 miembros de la facultad, se obtuvo $SSE = 783.90$. También se ajustó el modelo reducido $Y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \epsilon$, que dio como resultado un valor de $SSE = 795.23$. ¿Presentan los datos evidencia suficiente que apoye la afirmación de que el salario promedio depende del género de los miembros de la facultad? Utilice $\alpha = 0.05$.

9. Trece alumnos de características análogas se asignan al azar a 3 métodos de aprendizaje, obteniendo los siguientes resultados en el examen posterior.

Método 1	Método 2	Método 3
7.72	8.01	7.91
7.98	7.93	8.32
7.85	8.15	8.12
7.87	8.09	8.28
		8.23

¿Podemos concluir que los métodos son distintos? Use un nivel de significación del 0.05.

10. En un análisis de la varianza con cinco grupos se conoce que las medias estimadas en cada grupo son 20, 22, 24, 26 y 28. Hay diez observaciones en cada grupo y la varianza de estimación de las medias de grupo es 6. Calcular la tabla ANOVA.

11. Un psicólogo clínico quería comparar tres métodos para reducir los niveles de hostilidad en estudiantes universitarios. Cada prueba sicológica (PNH) fue usada para medir el grado de hostilidad. Las puntuaciones altas en esta prueba su usaron como indicación de gran hostilidad. En el experimento se usaron 11 estudiantes que obtuvieron puntuaciones altas y muy cercanas entre sí. De los 11 estudiantes se seleccionaron 5 al azar y se trataron con el método A, de los 6 restantes se tomaron tres al azar y se trataron con el método B y el resto se trató con el método C. Todos los tratamientos se realizaron durante un trimestre. Cada estudiante tomó la prueba PNH nuevamente al final del trimestre, con los siguientes resultados:

Método A	73, 83, 76, 68, 80
Método B	54, 74, 71
Método C	79, 98, 87

a) Realice un análisis de varianza para este experimento.

b) Presentan los datos suficiente evidencia para concluir que hay diferencias entre las respuestas medias de los estudiantes de los tres métodos, después del tratamiento?

Use un nivel de significación del 0.05.

12. Un estudio mide la tasa de absorción de tres tipos diferentes de solventes químicos orgánicos. Estos solventes se utilizan para limpiar partes metálicas industriales labradas y son desechos peligrosos potenciales. Se prueban muestras independientes de solventes de cada tipo y se registran sus tasas de absorción como porcentaje molar.

Arómaticos		Cloroalcanos		Ésteres		
1.06	0.95	1.58	1.12	0.29	0.43	0.06
0.79	0.65	1.45	0.91	0.06	0.51	0.09
0.82	1.15	0.57	0.83	0.44	0.10	0.17
0.89	1.12	1.16	0.43	0.55	0.53	0.17
				0.61	0.34	0.60

a) ¿Existe una diferencia significativa en la tasa de absorción media para los tres solventes? Utilice un p-valor para sus conclusiones.

13. Se llevó a cabo un experimento para analizar el efecto de la edad sobre la frecuencia cardiaca cuando un individuo hace ejercicio. Se eligieron 10 hombres

aleatoriamente de 4 grupos de edades. Cada individuo caminó por una banda sinfín a una velocidad fija durante un periodo de 12 minutos, y se registró (en latidos por minuto) el incremento en la frecuencia cardiaca, la diferencia antes y después del ejercicio.Estos datos aparecen en la siguiente tabla. ¿Proporcionan estos datos suficiente evidencia que indique una diferencia en el incremento medio de la frecuencia cardiaca entre los 4 grupos de edades? Lleva a cabo la prueba con un nivel de significancia de $\alpha = 0.05$.

	Edad			
	10-19	20-39	40-59	60-69
	29	24	37	28
	33	27	25	29
	26	33	22	34
	27	31	33	36
	39	21	28	21
	35	28	26	20
	33	24	30	25
	29	34	34	24
	36	21	27	33
	22	32	33	32
Total	309	275	295	282