Modelo Lineal
PRACTICA 2

1. Sean Y_i, $i = 1, \ldots, n$, variables aleatorias independientes con media común θ y varianza σ^2/w_i. Halle el estimador lineal insesgado de θ de mínima varianza y calcule esta varianza mínima.

2. Sea $Y \sim N_n(X\beta, \sigma^2 V)$, donde X es una matriz de $n \times p$ de rango p y V es una matriz conocida definida positiva de $n \times n$. Si β^* es el estimador de mínimos cuadrados generalizados de β, pruebe que

a) $Q/\sigma^2 = (Y - X\beta^*)'V^{-1}(Y - X\beta^*)/\sigma^2 \sim \chi^2_{n-p}$.

b) Q es un estimador insesgado de $(n-p)\sigma^2$.

c) Si $Y^* = X\beta^* = PY$, entonces P es idempotente, pero en general, no es simétrica. (Hint: analice el modelo transformado)

3. Sean $Y_i = \beta_o + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, 3)$, con $E(\epsilon) = 0$ y $\Sigma_\epsilon = \sigma^2 V$, donde

$$V = \begin{pmatrix} 1 & \rho a & \rho \\ \rho a & a^2 & \rho a \\ \rho & \rho a & 1 \end{pmatrix}$$

con a y $0 < \rho < 1$ desconocidos, y $x_1 = -1$, $x_2 = 0$ y $x_3 = 1$.

a) Halle los estimadores de mínimos cuadrados generalizados de β_o y β_1.

b) Pruebe que si $a = 1$ entonces los valores ajustados $Y_i^* = \beta_o^* + \beta_1^* x_i$ no pueden ser tales que todos los $Y_i - Y_i^*$ sean positivos o todos negativos.

4. Sea $Y = (Y_1, Y_2)$ un vector con densidad conjunta dada por

$$f(y_1, y_2) = k^{-1}\, exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(y_1-\theta_1)^2}{\sigma_1^2} - \frac{2\rho(y_1-\theta_1)(y_2-\theta_2)}{\sigma_1\sigma_2} + \frac{(y_2-\theta_2)^2}{\sigma_2^2}\right)\right],$$

donde $k = 2\,\pi\,\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}$, $\sigma_i > 0$, $-\infty < y_i < \infty$ y $|\rho| < 1$. Pruebe que $Y \sim N_2(\theta, \Sigma)$ y halle la correlación entre Y_1 e Y_2.

5. Si $Y \sim N_n(\theta, \Sigma)$, pruebe que $Y_i \sim N(\theta_i, \sigma^2_{ii})$.

6. Pruebe que si $Y \sim N_n(0, I_n)$, $\sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$.

7. Sean $Y \sim N_n(0, I_n)$ y A una matriz simétrica e idempotente de rango r. Si $Y'Y = Y'AY + Y'BY$, pruebe que $Y'BY \sim \chi^2_{n-r}$.

8. Pruebe que si $Y \sim N_n(\theta, \sigma I_n)$ y T es una matriz $n \times n$ simétrica de rango r. Entonces $Q = (Y-\theta)'T(Y-\theta)/\sigma^2 \sim \chi^2_r$ si y sólo si $T = T^2$ (es decir, si T es idempotente).

9. Sean $Y \sim N_n(0, \Sigma)$ y A una matriz de $n \times n$ simétrica y de rango r. Pruebe que $Y'AY \sim \chi^2_r$, si y sólo si $A\Sigma A = A$.

10. Si X e Y son dos v.a. n–dimensionales independientes con distribución normal multivariada, y si a y b son dos constantes, pruebe que $U = aX + bY$ también tiene distribución normal multivariada.

11. Sea $Y \sim N_3(0, I_3)$. Halle la esperanza de $(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2$.

12. Sean $Y_i = \beta_o + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$, donde los ϵ_i son independientes y $N(0, \sigma^2)$. Deduzca el estadístico F para testear $H_o : \beta_o = 0$

13. Dado que $\bar{x} = 0$, derive el test F para testear $H_o : \beta_o = \beta_1$ en el ejercicio anterior.

14. Sean $U_1, \ldots, U_n$ v.a. i.i.d. con distribución $N(\mu_1, \sigma^2)$ y $V_1, \ldots, V_n$ v.a. i.i.d. con distribución $N(\mu_2, \sigma^2)$ e independientes de las anteriores. Derive un test F para $H_o : \mu_1 = \mu_2$.

15. Una serie de $n + 1$ observaciones independientes $Y_i, i = 1, \ldots, n + 1$ son tomadas de una población con distribución normal con varianza desconocida σ^2. Después de las n primeras observaciones se sospecha que ha habido un repentino cambio en la media de la distribución. Derive un test para testear la hipótesis de que la observación $(n + 1)$–ésima tiene la misma media que las n anteriores.

16. Programe un algoritmo, usando el lm.object para un modelo lineal genérico con vector de parámetros $\beta \in \Re^p$, que permita obtener la Tabla de Análisis de la Varianza.

17. Programe un algoritmo, usando el lm.object para un modelo lineal genérico con vector de parámetros $\beta \in \Re^p$ y X de rango completo, que permita testear la hipótesis $H_o : C\beta = \delta$, donde $C \in \Re^{q \times p}$ y $\delta \in \Re^q$.

18. *Supervisores evaluados* (attitude.txt)

Recientemente en una entidad financiera se encuestó a los empleados con el fin de evaluar su satisfacción respecto a la labor de sus supervisores. Entre las preguntas había una pregunta diseñada para medir la performance total de un supervisor, asi como preguntas relacionadas a actividades específicas que involubraban la interacción entre el supervisor y el empleado. A continuación describimos las variables observadas en la encuesta:

y: puntaje asignado a la actividad global del supervisor

x_1: tiene en cuenta las quejas del empleado

x_2: no permite privilegios especiales

x_3: oportunidad de aprender cosas nuevas

x_4: aumentos basados en performance

x_5 : muy crítico a performances bajas

x_6 : velocidad para alcanzar mejores posiciones.

Las variables x_1 , x_2 y x_5 tratan de evaluar la relación interpersonal entre empleados y supervisor, mientras que x_3 y x_4 son menos personales. Por otro lado, x_6 no es una evaluación sobre el supervisor, pero sirve para medir cómo el empleado percibe su crecimiento en la empresa. Los datos de la Tabla 1 corresponden al puntaje obtenido por 30 supervisores en su respectivo departamento (de aproximadamente 35 empleados cada uno)

a) Use Y como variable dependiente y todas las variables independientes y una intercept para realizar un ajuste lineal. Calcule el estimador de mínimos cuadrados de los parámetros y para cada uno de ellos testee la hipótesis de que es 0. ¿Cuáles son significativamente distintos de 0 y con qué nivel?

b) Analizando la tabla de Análisis de Varianza, ¿es la regresión significativa?

¿Qué decisión toma? ¿Con qué nivel de significación?

¿Está testeando si todos los coeficientes son 0?

c) Como x_1 , x_2 y x_5 son de caracter más personal que x_3 , x_4 y x_6 podríamos preguntarnos si x_1 y x_3 alcanzan para explicar la regresión. Realice un test para la hipótesis nula $H_o : \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ contra la alternativa de que alguno de estos parámetros no es 0.

¿Qué decisión toma? ¿Con qué nivel de significación?

d) Suponiendo que $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$, ajuste un modelo lineal usando como únicas variables regresoras a X_1 y a X_3 . Para este modelo plantee un test para decidir si los coeficientes correspondientes a las dos variables independientes son iguales o no. ¿Cómo escribiría a H_o ?

¿Qué decisión toma? ¿Con qué nivel de significación?

e) Volviendo al modelo general planteado en a), plantee un test para decidir si β_1 y β_3 son iguales y si $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ al mismo tiempo. ¿Cómo escribiría a H_o ?

¿Qué decisión toma? ¿Con qué nivel de significación?

¿Necesariamente la conclusión va a ser igual a la tomada antes?

¿Es el test realizado en d) más sensible para detectar alejamiento de la hipótesis de igualdad entre los coeficientes de X_1 y X_3 que el planteado en este ítem?

f) Suponiendo que $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$, y con el modelo ajustado en d) usando como únicas variables regresoras a X_1 y a X_3 , decida si $\beta_1 + \beta_3 = 1$. ¿Cómo escribiría a H_o ?

¿Qué decisión toma? ¿Con qué nivel de significación?

En virtud de d) y el resultado obtenido, ¿tiene sentido pensar que $\beta_1 = \beta_3 = 0.5$?

¿Cómo testearía directamente $\beta_1 = \beta_3 = 0.5$?

19. *Caudal de río* (peak.txt)

Los datos que se muestran en la Tabla 2 son datos simulados que corresponden al caudal de agua (Y) en distintas cuencas de ríos después de episodios de tormenta. Las variables independientes son:

X_1 = área de la cuenca

X_2 = área impermeable al agua

X_3 = pendiente promedio del terreno

X_4 = máxima longitud de los afluentes de la cuenca

X_5 = índice de absorción del agua (0= absorción completa, 100= no absorción)

X_6 = capacidad de depósito del suelo

X_7 = velocidad de infiltración del agua en el suelo

X_8 = cantidad de lluvia caída

X_9 = tiempo durante el cual la cantidad de lluvia excedió 0.25 pulgada por hora

a) Calcule la matriz de correlación de todas las variables comprendidas en el problema, incluyendo a la variable dependiente Y. Inspeccionando esta matriz determine cuáles parecen ser las variables que pueden contribuir significativamente a explicar la variación de Y. Si tuviera que usar una sola variable, ¿cuál usaría?

b) Calcule la matriz de correlación de $ln(Y)$ con el ln de cada una de las variables independientes. ¿Cómo cambian las correlaciones y sus conclusiones acerca de cuáles serían las variables que contribuyen significativamente a la variación de $ln(Y)$?

c) Use $ln(Y)$ como variable dependiente y los logaritmos de las variables independientes y una intercept para realizar un ajuste lineal. Calcule el estimador de mínimos cuadrados de los parámetros y para cada uno de ellos testee la hipótesis que es 0. ¿Cuáles son significativamente distintos de 0? ¿Cuáles son las variables independientes que usted eliminaría en primera instancia para simplificar el modelo? ¿Es la regresión significativa?

d) Elimine del modelo las variables que resultan menos interesantes y reestime los parámetros. ¿Son todos los parámetros significativos al 5% ? ¿Es la regresión significativa? Si no es asi, continúe eliminando aquellas variables independientes menos importantes en cada paso y reestime los parámetros. Pare cuando todas las variables sean significativas al 5%. ¿Le parece que $\beta_o = 0$ tiene sentido en este ejemplo? Analice la tabla de análisis de la varianza para su modelo final.

20. *Cemento* (cemento.txt)

Los datos de la Tabla 3 fueron tomados en un estudio experimental para relacionar

el calor generado (Y) al fraguar 14 muestras de cemento con distinta composición. Las variables explicativas son los pesos (medidos en porcentajes del peso de cada muestra de cemento) de 5 componentes del cemento.

a) Calcule la matriz de correlación de todas las variables comprendidas en el problema, incluyendo a la variable dependiente Y. Inspeccionando esta matriz determine cuáles parecen ser las variables que pueden contribuir significativamente a explicar la variación de Y.

b) Use Y como variable dependiente y todas las variables independientes y una intercept para realizar un ajuste lineal. Calcule el estimador de mínimos cuadrados de los parámetros y para cada uno de ellos testee la hipótesis de que es 0. ¿Cuáles son significativamente distintos de 0? Analizando la tabla de Análisis de Varianza, ¿es la regresión significativa? ¿Observa alguna contradicción con el resultado obtenido en los tests individuales anteriores? ¿Vale la pena hacer un nuevo intento para seleccionar qué variables entran en la regresión?

c) Calcule la suma de las 5 variables independientes. ¿Qué observa? ¿Cómo se justifica este parecido entre los totales? A partir de este resultado, ¿le parece que la matriz de diseño puede estar bien condicionada ? ¿Puede justificar ésto que eliminemos del modelo la intercept?

d) Realice un nuevo ajuste lineal usando las 5 variables independientes y eliminando la intercept. ¿Cómo se comparan estos resultados con los obtenidos anteriormente? ¿Cuáles son significativamente distintos de 0? Plantee un nuevo modelo en el que intervengan aquellas variables independientes que contribuyen significativamente y estime los parámetros por mínimos cuadrados. ¿Qué modelo eligiría finalmente?

Tabla 1. Supervisores evaluados.

obs.	X_1	X_2	X_3	X_4	X_5	X_6	Y
1	51	30	39	61	92	45	43
2	64	51	54	63	73	47	63
3	70	68	69	76	86	48	71
4	63	45	47	54	84	35	61
5	78	56	66	71	83	47	81
6	55	49	44	54	49	34	43
7	67	42	56	66	68	35	58
8	75	50	55	70	66	41	71
9	82	72	67	71	83	31	72
10	61	45	47	62	80	41	67
11	53	53	58	58	67	34	64
12	60	47	39	59	74	41	67
13	62	57	42	55	63	25	69
14	83	83	45	59	77	35	68
15	77	54	72	79	77	46	77
16	90	50	72	60	54	36	81
17	85	64	69	79	79	63	74
18	60	65	75	55	80	60	65
19	70	46	57	75	85	46	65
20	58	68	54	64	78	52	50
21	40	33	34	43	64	33	50
22	61	52	62	66	80	41	64
23	66	52	50	63	80	37	53
24	37	42	58	50	57	49	40
25	54	42	48	66	75	33	63
26	77	66	63	88	76	72	66
27	75	58	74	80	78	49	78
28	57	44	45	51	83	38	48
29	85	71	71	77	74	55	85
30	82	39	59	64	78	39	82

Tabla 2. Caudal de ríos.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
.03	.006	3.0	1	70	1.5	.25	1.75	2.0	46
.03	.006	3.0	1	70	1.5	.25	2.25	3.7	28
.03	.006	3.0	1	70	1.5	.25	4.00	4.2	54
.03	.021	3.0	1	80	1.0	.25	1.60	1.5	70
.03	.021	3.0	1	80	1.0	.25	3.10	4.0	47
.03	.021	3.0	1	80	1.0	.25	3.60	2.4	112
.13	.005	6.5	2	65	2.0	.35	1.25	.7	398
.13	.005	6.5	2	65	2.0	.35	2.30	3.5	98
.13	.005	6.5	2	65	2.0	.35	4.25	4.0	191
.13	.008	6.5	2	68	.5	.15	1.45	2.0	171
.13	.008	6.5	2	68	.5	.15	2.60	4.0	150
.13	.008	6.5	2	68	.5	.15	3.90	3.0	331
1.00	.023	15.0	10	60	1.0	.20	.75	1.0	772
1.00	.023	15.0	10	60	1.0	.20	1.75	1.5	1268
1.00	.023	15.0	10	60	1.0	.20	3.25	4.0	849
1.00	.023	15.0	10	65	2.0	.20	1.80	1.0	2294
1.00	.023	15.0	10	65	2.0	.20	3.10	2.0	1984
1.00	.023	15.0	10	65	2.0	.20	4.75	6.0	900
3.00	.039	7.0	15	67	.5	.50	1.75	2.0	2181
3.00	.039	7.0	15	67	.5	.50	3.25	4.0	2484
3.00	.039	7.0	15	67	.5	.50	5.00	6.5	2450
5.00	.109	6.0	15	62	1.5	.60	1.50	1.5	1794
5.00	.109	6.0	15	62	1.5	.60	2.75	3.0	2067
5.00	.109	6.0	15	62	1.5	.60	4.20	5.0	2586
7.00	.055	6.5	19	56	2.0	.50	1.80	2.0	2410
7.00	.055	6.5	19	56	2.0	.50	3.25	4.0	1808
7.00	.055	6.5	19	56	2.0	.50	5.25	6.0	3024
7.00	.063	6.5	19	56	1.0	.50	1.25	2.0	710
7.00	.063	6.5	19	56	1.0	.50	2.90	3.4	3181
7.00	.063	6.5	19	56	1.0	.50	4.76	5.0	4279

Tabla 3. Cemento.

obs.	X_1	X_2	X_3	X_4	X_5	Y
1.0000	6.0000	7.0000	26.000	60.000	2.5000	85.500
2.0000	15.000	1.0000	29.000	52.000	2.3000	76.000
3.0000	8.0000	11.000	56.000	20.000	5.0000	110.40
4.0000	8.0000	11.000	31.000	47.000	2.4000	90.600
5.0000	6.0000	7.0000	52.000	33.000	2.4000	103.50
6.0000	9.0000	11.000	55.000	22.000	2.4000	109.80
7.0000	17.000	3.0000	71.000	6.0000	2.1000	108.00
8.0000	22.000	1.0000	31.000	44.000	2.2000	71.600
9.0000	18.000	2.0000	54.000	22.000	2.3000	97.000
10.000	4.0000	21.000	47.000	26.000	2.5000	122.70
11.000	23.000	1.0000	40.000	34.000	2.2000	83.100
12.000	9.0000	11.000	66.000	12.000	2.6000	115.40
13.000	8.0000	10.000	68.000	12.000	2.4000	116.30
14.000	18.000	1.0000	17.000	61.000	2.1000	62.600