

Estadística Aplicada II
Tarea 3

1. Considere un modelo de regresión lineal en donde $E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 (3X_i^2 - 2)$, $i = 1, 2, 3$. Además, $X_1 = -1$, $X_2 = 0$, $X_3 = 1$.

a) Estimar β_0 , β_1 y β_2 por mínimos cuadrados.

b) Demostrar que los estimadores de β_0 y β_1 no cambian si $\beta_2 = 0$.

NOTA: Al parecer la influencia de la variable $(3X^2 - 2)$ sobre Y es independiente de la influencia de X sobre Y. Esto sucede si las variables son ortogonales, y ésto es lo que pasa en este caso.

2. Sea $Y_i = \beta_0 + \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + \epsilon_i$, $i=1,2,\dots,n$, donde $E(\epsilon) = \mathbf{0}$, $\text{Var}(\epsilon) = \sigma^2 \mathbf{I}$. Si $\hat{\beta}_1$ es el estimador de mínimos cuadrados de β_1 , demuestre que

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (X_{1i} - \bar{X}_1)^2 (1 - r_{12}^2)}$$

donde r_{12} es el coeficiente de correlación muestral entre X_1 y X_2 .

3. Considere el modelo $E(\mathbf{Y}) = \mathbf{X} \boldsymbol{\beta}$, en donde la matriz \mathbf{X} está formada por los vectores columna $\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_k$, es decir, $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_k)$. Suponiendo que todas las columnas de \mathbf{X} son ortogonales, demostrar que la suma de cuadrados del error se puede expresar como

$$\text{SCE} = \mathbf{Y}' \mathbf{Y} - \sum_{s=0}^k \hat{\beta}_s^2 \mathbf{X}_s' \mathbf{X}_s, \quad \text{con } \hat{\beta}_s = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{Y}$$

4. Sea $\mathbf{w} \in \mathfrak{R}^k$ un vector, y sea \mathbf{A} una matriz de dim $(k \times k)$, demuestre que

$$E(\mathbf{w}' \mathbf{A} \mathbf{w}) = E(\mathbf{w}') \mathbf{A} E(\mathbf{w}) + \text{tr}[\mathbf{A} \text{Var}(\mathbf{w})]$$

5. Usa el resultado del problema 4 para demostrar que $E(\text{CME}) = \sigma^2$.

6. Para el modelo de regresión lineal simple, muestra que los elementos de la matriz "gorrito" son:

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{S_{xx}}, \quad \text{y} \quad h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}$$

7. Para saber cuál es el número de gansos en cierta área, es usual utilizar métodos de muestreo aéreo, en el cual una persona experimentada sobrevuela una área específica y al detectar un grupo de animales estima, a ojo, el número X, de gansos. Para investigar la confiabilidad de este método de conteo, se realiza un experimento en el cual, además de la estimación a ojo, se toma una fotografía, de tal forma que se cuenta con el número exacto Y, de gansos. Se realizaron 45 observaciones, y éstas fueron:

X	50	25	30	35	25	20	12	34	20	10	25	10	15	20	40
Y	56	38	25	48	38	22	22	42	34	14	30	9	18	25	62

X	30	75	35	9	55	30	25	40	75	100	150	120	250	500	200
Y	26	88	56	11	66	42	30	90	119	165	152	205	409	342	200

X	50	75	150	50	60	75	150	40	25	100	200	60	40	35	20
Y	73	123	150	70	90	110	95	57	43	55	325	114	83	91	56

a) Construya un diagrama de dispersión de Y vs X. ¿Sugieren estas gráficas que un modelo de regresión lineal de Y sobre X es el adecuado?, ¿porqué?. ¿Porqué es apropiado considerar una regresión de Y sobre X y no una de X sobre Y?.

- b) Estime por máxima verosimilitud la regresión lineal de Y sobre X.
- c) ¿Se cumplen los supuestos bajo los cuales se realizó la estimación en el inciso anterior?. En particular, analice el supuesto de heterocedasticidad. Realice las correcciones pertinentes.
- d) Obtener una predicción de Y si se sabe que $X = 80$.
- e) Como resultado de todo lo anterior, ¿qué se puede decir de la práctica de usar conteo visual?

8. Se tienen los datos de la actuación de los equipos de football americano, de la liga de los E.U.A. en el año de 1976. Tomaremos como variable dependiente (Y) el número de juegos ganados en una temporada en la que se juegan 14 partidos. Las variables independientes son: (X_1) número de yardas por pase, (X_2) número de yardas por corrida y (X_3) número de yardas por corrida del oponente.

<i>Equipo</i>	<i>Y</i>	<i>X₁</i>	<i>X₂</i>	<i>X₃</i>
Washington	10	1985	59.7	2205
Minnesota	11	2855	55.0	2096
New England	11	1737	65.6	1847
Oakland	13	2905	61.4	1903
Pittsburg	10	1666	66.1	1457
Baltimore	11	2927	61.0	1848
Los Angeles	10	2341	66.1	1564
Dallas	11	2737	58.0	1821
Atlanta	4	1414	57.0	2577
Buffalo	2	1838	58.9	2476
Chicago	7	1480	67.5	1984
Cincinnati	10	2191	57.2	1917
Cleveland	9	2229	58.8	1761
Denver	9	2204	58.6	1709
Detroit	6	2140	59.2	1901
Green Bay	5	1730	54.4	2288
Houston	5	2072	49.6	2072
Kansas City	5	2929	54.3	2861
Miami	6	2268	58.7	2411
New Orleans	4	1983	51.7	2289
New York Giants	3	1792	61.9	2203
New York Jets	3	1606	52.7	2592
Philadelphia	4	1492	57.8	2053
St. Louis	10	2835	59.7	1979
San Diego	6	2416	54.9	2048
San Francisco	8	1638	65.3	1786
Seattle	2	2649	43.8	2876
Tampa Bay	0	1503	53.5	2560

- a) Ajusta el modelo de regresión lineal múltiple $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
- b) Construye una tabla de análisis de varianza para probar la significancia del modelo.
- c) Calcula R^2 .
- d) Realiza una prueba para verificar la contribución de cada regresor.
- e) Muestra que el cuadrado de la correlación muestral entre Y_i y \hat{Y}_i es igual a R^2 .
- f) Encuentra un intervalo de confianza al 95% para β_2 .
- g) Encuentra un intervalo de confianza del 95% para el número promedio de juegos ganados por un equipo cuando $X_1 = 2300$, $X_2 = 56.0$ y $X_3 = 2100$.
- h) Realiza un análisis de residuos para verificar la adecuación del modelo y realiza las correcciones pertinentes.