Guía Examen MA34B-02
Profesor Cátedra: Rodrigo Abt B.
Profesor Auxiliar: Ismael Vergara C.

Problema 1

Sea el modelo lineal:

$$y_i = b_o + \sum_{j=1}^{p} b_j x_{ij} + e_i \quad (i = 1,2,...,n) \qquad (1)$$

Se denotan $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ $(j = 1,2,...,p)$ y $z_{ij} = x_{ij} - \bar{x}_j$ $(i = 1,2,...,n;\ j = 1,2,...,p)$.
Se define entonces el modelo:

$$y_i = g_o + \sum_{j=1}^{p} g_j z_{ij} + e_i \quad (i = 1,2,...,n) \qquad (2)$$

1.1 Muestre que los modelos (1) y (2) producen las mismas predicciones y dé el estimador de g_o.
1.2 Se supone que $e \sim N_n(0, s^2 I_n)$. Dé la expresión de la predicción $\hat{y}_o$ de y_o para una nueva observación $x_o^t = (x_{o1}, x_{o2}, ..., x_{op})$ a partir del modelo (2). Exprese la varianza de $\hat{y}_o$ como suma de dos varianzas.
1.3 Dé un intervalo de confianza para $m_o = E(y_o)$.
1.4 Deduzca para que valor de $x_o^t = (x_{o1}, x_{o2}, ..., x_{op})$ el intervalo de confianza para m_o tiene el largo más pequeño. Dé el largo del intervalo en este caso.
1.5 En el modelo (2): $Y = Zg + e$, se supone ahora que las columnas de Z son ortogonales ($Z^t Z$ es diagonal). Dé la expresión individual de los coeficientes $\hat{g}_j$ y muestre que son no correlacionados entre si.

Solución:

1.1 Se tienen las predicciones para el modelo (1)

$$\hat{y}_i = \hat{b}_o + \sum_{j=1}^{p} \hat{b}_j x_{ij}$$

y para el modelo (2): $\hat{y}_i = \hat{g}_o + \sum_{j=1}^{p} \hat{g}_j z_{ij} = \hat{g}_o - \sum_{j=1}^{p} \hat{g}_j \bar{x}_j + \sum_{j=1}^{p} \hat{g}_j x_{ij} = \hat{b}_o + \sum_{j=1}^{p} \hat{b}_j x_{ij}$

con $\hat{g}_o - \sum_{j=1}^{p} \hat{g}_j \bar{x}_j = \hat{b}_o$ y $\hat{g}_j = \hat{b}_j$ $(j = 1,2,...,p)$.
Luego tenemos las mismas predicciones.

Sea $Q = \sum_i (y_i - g_o - \sum_{j=1}^{p} g_j z_{ij})^2$. Luego $\frac{\partial Q}{\partial g_o} = 0 \Rightarrow \sum_i y_i - n g_o - \sum_j g_j (\sum_i z_{ij}) = 0$

Como $\sum_i z_{ij} = 0 \ (\forall j) \Rightarrow$

$$\boxed{\hat{g}_o = \bar{y}}$$

1.2 Usando el modelo (2): $\hat{y}_o = \bar{y} + \sum_{j=1}^{p} \hat{g}_j z_{oj} = \bar{y} + \sum_{j=1}^{p} \hat{g}_j (x_{oj} - \bar{x}_j)$

La varianza es $s_o^2 = Var(\hat{y}_o) = Var(\bar{y}) + Var(\sum_{j=1}^{p} \hat{g}_j (x_{oj} - \bar{x}_j))$

1.3 Se tiene $\frac{\hat{y}_o - m_o}{s_o} \sim N(0,1)$. Luego el intervalo es de la forma

$$\boxed{I = [\hat{y}_o - us_o, \hat{y}_o + us_o]}$$

en donde u es tal que: $P(|N(0,1)| \leq u) = 1 - a$.

1.4 $Var(\hat{y}_o) = Var(\bar{y}) + Var(\sum_{j=1}^{p} \hat{g}_j (x_{oj} - \bar{x}_j)) \geq Var(\bar{y}) = \frac{s^2}{n}$

Luego: $\boxed{Var(\hat{y}_o) = \frac{s}{n} \Leftrightarrow x_{oj} = \bar{x}_j \ (j = 1,2,...,p)}$

En este caso el largo del intervalo es igual a: $\boxed{2u\frac{s}{\sqrt{n}}}$.

1.5 $\hat{g} = D^{-1} Z^t y \Rightarrow$ $\boxed{\hat{g}_j = \frac{Cov(x_j, y)}{var(x_j)}}$ y son no correlacionados dado que

$\boxed{Var(\hat{g}) = s^2 D^{-1}}$.

Problema 2

2.1 Comente los resultados de la regresión lineal de una variable *Y* sobre 5 variables explicativas (Tabla 1). Complete la tabla 2 ANOVA. Deduzca el número de observaciones y el coeficiente de correlación múltiple.

Tabla 1

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	19.95	13.63	1.46	0.165
X_1	-1.793	1.233	-1.45	0.168
X_2	0.0436	0.05326	0.82	0.427
X_3	0.5558	0.09296	5.98	0.000
X_4	1.1102	0.4338	2.56	0.023
X_5	-1.811	2.027	-0.89	0.387

Tabla 2: ANOVA

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	5	582.69	116.54	?	0.000
Error	?	?	4.30		
Total	19	642.92			

2.2 Se presenta a continuación la matriz de correlaciones (Tabla 3) y las correlaciones parciales de *Y* con cada X_i, con las 4 otras variables explicativas fijas (Tabla 4). Interprete los coeficientes de correlación y correlación parcial.

Tabla 3

	X_1	X_2	X_3	X_4	X_5	*Y*
X_1	1.00000	0.18114	0.22963	0.50266	0.19677	0.19229
X_2	0.18114	1.00000	0.82718	0.05106	0.92710	0.75340
X_3	0.22963	0.82718	1.00000	0.18333	0.81906	0.92716
X_4	0.50266	0.05106	0.18333	1.00000	0.12381	0.33365
X_5	0.19677	0.92710	0.81906	0.12381	1.00000	0.73299
Y	0.19229	0.75340	0.92716	0.33365	0.73299	1.00000

Tabla 4

Corr. parcial	
X_1	-0.3622
X_2	0.2137
X_3	0.8477
X_4	0.5646
X_5	-0.2322

2.3 Los resultados de la regresión lineal de la variable *Y* sobre las variables X_3 y X_4 son dados en las tablas 5 y 6.Justifique porque se hace esta regresión. Complete la tabla ANOVA. Efectúe el test de hipótesis $H_o : b_1 = b_2 = b_5 = 0$

Tabla 5

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	14.583	9.175	1.59	0.130
X_3	0.5415	0.05004	10.82	0.000
X_4	0.7499	0.3666	2.05	0.057

Tabla 6

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	2	570.50	285.25	66.95	?
Error	?	72.43	4.26		
Total	19	?			

2.4 Dé un intervalo de confianza a 98% para el coeficiente de X_2. Comente.

Solución:

2.1 Considerando el p-valor de la tabla 2, se puede decir que el modelo es globalmente significativo. Considerando los p-valores de la tabla 1 se puede concluir que dos variables de las 5 son significativas: X_3 y X_4.

Tabla 2: ANOVA

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	5	582.69	116.54	**27.08**	0.000
Error	**14**	**60.24**	4.30		
Total	19	642.92			

2.2 Se comprueba porque no son significativos los coeficientes de las variables X_1 , X_2 y X_5 a pesar de tener coeficientes de correlación elevado con la variable *Y*: tienen un coeficiente de correlación parcial pequeño. Por otro lado la variable X_4 aumenta su correlación cuando se fijan las otras variables.

2.3 Se justifica por lo anterior esta regresión. El test de hipótesis $H_o : b_1 = b_2 = b_5 = 0$ se hace calculando el test F que compara los residuos de ambas regresión con y sin las tres variables X_1 , X_2 y X_5 : $\frac{(72.43 - 60.24)/3}{60.24/14} = 0.9443$ y. El p-valor que es igual a $P(F_{3,14} > 0.9443) = 0.4457$ confirma los resultados anterioriores: no se rechaza $H_o : b_1 = b_2 = b_5 = 0$.

Tabla 6

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	2	570.50	285.25	66.95	**0.000**
Error	**17**	72.43	4.26		
Total	19	**642.92**			

2.4 En la tabla 1 se tiene el coeficiente asociado a X_2 (*0.0436*) y su desviación estándar (*0.05326*). Por otro lado $P(|t_{19}| < 2.54) = 0.98$. Luego el intervalo de confianza buscado es:
$[0.0436 - 2.54 * 0.05326,\ 0.0436 - 2.54 * 0.05326] = [-0.0917,\ 0.1789]$. Es un intervalo que cubre el *0*, lo que confirma que X_2 no es significativa.

Problema 3

3.1 Comente los resultados de la regresión lineal de una variable *Y* sobre 4 variables explicativas (Tabla 1). Complete la tabla ANOVA (Tabla 2). Deduzca el número de observaciones.

Tabla 1

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	23975.69	6080.213	3.943	0.0001
X_1	0.6091	0.721	0.845	0.4000
X_2	-96.236	53.093	-1.813	0.0728
X_3	-80.621	31.328	-2.573	0.0115
X_4	-281.355	85.072	-3.307	0.0013

Tabla 2: ANOVA

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	4	2173922023	543480506	?	0.000
Error	?	?	23098393.7		
Total	106	4529958182			

Coeficiente de correlación múltiple: 0.69.

3.2 Se presenta a continuación la matriz de correlaciones (Tabla 3) y las correlaciones parciales de *Y* con cada X_i, con las 3 otras variables explicativas fijas (Tabla 4). Interprete los coeficientes de correlación y correlación parcial.

Tabla 3

	X_1	X_2	X_3	X_4	Y
X_1	1.0000	-0.0308	-0.1444	-0.1534	0.2022
X_2	-0.0308	1.0000	-0.9005	-0.8688	0.5518
X_3	-0.1444	-0.9005	1.0000	0.8714	-0.6408
X_4	-0.1534	-0.8688	0.8714	1.0000	-0.6595
Y	0.2022	0.5518	-0.6408	-0.6596	1.0000

Tabla 4

Correlación parcial	
X_1	0.0834
X_2	-0.1767
X_3	-0.2469
X_4	-0.3112

3.3 En la tabla 5 se encuentran los resultados de la regresión lineal de la variable *Y* sobre las variables X_3 y X_4. Justifique porque se hace esta regresión. Efectúe el test de hipótesis $H_o : \beta_1 = \beta_2 = 0$ utilizando los resultados de las tablas 2 y 6 con un error de tipo 1 de 5%.

Tabla 5

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	13530.32	1301.28556	10.398	0.000
X_3	-51.1909	24.688272	-2.073	0.0406
X_4	-210.224	76.494961	-2.748	0.0071

Tabla 6

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	2	2040334717	102016735	42.65516	0.0000
Error	104	2487328710	23916622.2		
Total	106	4527663427			

Coeficiente de correlación múltiple: 0.67.

Solución:

3.1 Hay *n=107* observaciones.

Tabla 2: ANOVA

Fuente de variabilidad	Grados de libertad	Suma de cuadrados	Suma de cuadrados medio	F	*p-valor*
Regresión	4	2173922023	543480506	23.5289	0.000
Error	102	2356036159	23098393.7		
Total	106	4529958182			

3.2 En presencia de las otras 3 variables la variable X_1 (y X_2) no esta correlacionada con la variable Y (correlaciones parciales), lo que confirma la falta de significación que tienen en el modelo en 1.1.

3.3 El modelo sin las dos variables X_1 y X_2 tiene casi el mismo coeficiente de correlación múltiple y las variables tienen mejor significación. El resultado del test de hipótesis $H_o : b_1 = b_2 = 0$ lo confirma. El estadístico del test se escribe tomando los residuos de ambos modelos:

- Los residuos del modelo con las 4 variables es: 2356036159
- Los residuos del modelo con las 2 variables X_3 y X_4 es: 2487328710
- El estadístico para $H_o : b_1 = b_2 = 0$ es: $\frac{(2487328710 - 2356036159)/2}{23098393.7} = 2.8420$
- La probabilidad para que un Fisher con 2 y 102 g.l. sobrepasa el valor 2.8420 es 0.0629 (el p-valor). No se puede rechazar la hipótesis nula con un riesgo de 5%.

Problema 4

Un instituto agrícola quiere comparar el efecto de dos fertilizantes F_1 y F_2 sobre el rendimiento del cultivo de trigo. Con este propósito, diseña un experimento con tres grupos de parcelas: un grupo control sin fertilizante, un grupo con el fertilizante F_1 y un grupo con el fertilizante F_2. En la tabla 1 son resumidos los resultados de la cosecha de trigo por unidad de superficie.

Tabla 1

Grupos	Media	Desviación típica	Frecuencia
Grupo control	4.8450	2.8409	120
Grupo F_1	5.3345	2.8964	80
Grupo F_2	9.0639	2.9386	75
Total	6.1380	3.4087	275

4.1 Construye la tabla ANOVA que permite decidir si se observan diferencias en el rendimiento de trigo entre los tres grupos. Interprete los resultados.

4.2 Realice los tres tests de comparación de medias sobre el rendimiento, considerando los tres pares de grupos. Precise los supuestos que hizo y las hipótesis planteadas. Concluye.

4.3 Si no cambian las medias y las desviaciones típicas de los dos primeros grupos y el tamaño del grupo control, como hay que modificar el tamaño del grupo F_1 para que cambie el resultado del test de comparación del grupo control con el grupo F_1.

Solución

4.1 La suma de los cuadrados debido a los grupos es 270*varianza intragrupos:

$$120 * 2.8409^2 + 80 * 2.8964^2 + 75 * 2.9386^2 = 2312.615$$

La suma de los cuadrados debido a los residuos es 270*varianza intergrupos:

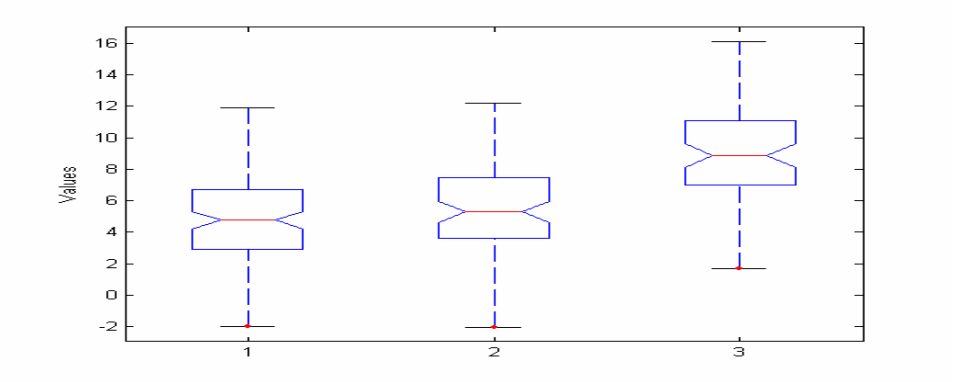
$$120 * (4.8450 - 6.1380)^2 + 80 * (5.3345 - 6.1380)^2 + 75 * (9.0639 - 6.1380)^2 = 894.346$$

Tabla ANOVA

Fuente de variación	Grados de libertad	Suma cuadrados	Cuadrado Medio	F	P_valor
Grupos	2	894.3459	447.1729	52.5946	0.000
Residuos	272	2312.615	8.5023		
Total	274	3206.9609			

Se concluye que hay diferencia entre los tres grupos.

4.2



Grupos	Hipótesis	Diferencia medias	Desv. Típica de los dos grupos	Grados de libertad	t	P_valor
Control / F_1	H_o: $m_o=m_1$ H_1: $m_o<m_1$	-0.4896	2.8922	198	-1.1728	0.1211
Control / F_2	H_o: $m_o=m_2$ H_1: $m_o<m_2$	-4.2189	2.9088	193	-9.8535	0.000
F_1 / F_2	H_o: $m_1=m_2$ H_1: $m_1<m_2$	-3.7294	2.9550	153	-7.8521	0.000

Las medias del grupo control con el grupo F_1 son significativamente diferentes. Las otras lo son. Este resultado esta validado con el boxplot.

Los supuestos:

- Se asume la normalidad
- Se supone que la varianza en cada grupo es la misma

4.3 Basta aumentar suficientemente el tamaño del grupo F_1 para que el p-valor disminuye.

$$t = -0.4896 / (2.8922 * \sqrt{\frac{1}{120} + \frac{1}{80}}) = -1.1728$$

Por ejemplo con una muestra de 500 para el grupo F_1 obtenemos un p_valor de 5%:

$$t = -0.4896 / (2.8922 * \sqrt{\frac{1}{120} + \frac{1}{500}}) = -1.6653$$

Problema 5

Se tiene la siguiente tabla con datos

x	y
1	0,08
2	1,73
3	1,31
4	2,44
5	3,07
6	2,00
7	3,17
8	5,31

a) Ajústese una recta de la forma $y = \beta_0+\beta_1 x$ para los datos usando mínimos cuadrados ordinarios. Escriba el desarrollo.
b) Suponiendo normalidad e independencia de los errores, i.e. $\varepsilon_i \rightarrow N(0,\sigma^2)$ y $Cov(\varepsilon_i, \varepsilon_j)=0$ $\forall\ i \neq j$ encuentre el estimador máximo verosímil para σ^2.
c) Encuentre el valor del coeficiente R^2 y el valor del estadístico F. Comente.
d) Efectúe y comente los siguientes tests de hipótesis (suponga α=5%):
   c.1) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
   c.1) $H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$
e) Sea $e_{n+1} = (\beta_0 - \beta_0^*)+(\beta_1-\beta_1^*)x_{n+1} + \varepsilon_{n+1}$ el error de predicción para un valor x_{n+1}, donde β_0^* y β_1^* son los estimadores MCO para β_0 y β_1 respectivamente.

   e.1) Pruebe que: $Var(e_{n+1}) = s^2\left[1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$

   e.2) Encuentre una predicción para x=8,5 y la correspondiente varianza para el error de la misma.

   NOTA: $Cov(\beta_0^*, \beta_1^*) = \frac{-\bar{x}s^2}{\sum(x_i-\bar{x})^2}$

Solución:

a) Usando MCO:

$$Min\ Q = \sum_i \sum_j (y_i - b_0 - b_1 x_i)^2,$$

$$\frac{\partial Q}{\partial b_0} = 0 \ \text{ y } \ \frac{\partial Q}{\partial b_1} = 0 \ \Rightarrow \qquad \Rightarrow\ \hat{y} = -0{,}103 + 0{,}554x$$

$$\hat{b}_1 = \frac{\sum x_i y_i - n\bar{X}\bar{Y}}{\sum x_i^2 - n\bar{X}^2} = 0{,}554$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = -0{,}103$$

b)

Sea $f_n(y_i / b_0, b_1, s^2) = \left(\frac{1}{2ps^2}\right)^{n/2} \exp\left\{-\frac{1}{2s^2}\sum (y_i - b_0 - b_1 x_i)^2\right\}$, luego:

$$\frac{\partial \ln f_n}{\partial s^2} = 0 \ \Rightarrow \ \hat{s}^2 = \frac{1}{n}\sum (y_i - \hat{b}_0 - \hat{b}_1 x_i)^2 = 0{,}477$$

$$\tilde{s}^2 = \frac{n}{n-2}\hat{s}^2 = 0{,}636$$

c)

$$R^2 = 1 - \frac{SR}{ST}, \text{ donde } ST = \sum (y_i - \bar{y})^2 \quad \text{y} \quad F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0{,}771}{0{,}229/6} \approx 20{,}2$$

$$SR = \sum (y_i - \hat{y}_i)^2 \Rightarrow R^2 = 0{,}771$$

d) Usando los estadísticos t:

d.1)
$$\tilde{s}_{b_1} = \frac{\tilde{s}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0{,}797}{6{,}48} \approx 0{,}123$$

$$t = \frac{\hat{b}_1}{\tilde{s}_{b_1}} = \frac{0{,}554}{0{,}123} \approx 4{,}5 > t_6(2{,}5\%) = 2{,}447$$

=> Se rechaza H_0 => X es significativa.

d.2)
$$\tilde{s}_{b_0} = \frac{\tilde{s}\sqrt{\sum x_i^2}}{\sqrt{n\sum (x_i - \bar{x})^2}} = \frac{11{,}94}{18{,}33} \approx 0{,}651$$

$$t = \frac{\hat{b}_0}{\tilde{s}_{b_0}} = \left|\frac{-0{,}103}{0{,}651}\right| \approx |-0{,}16| < t_6(2{,}5\%) = 2{,}447$$

=> No se rechaza H_0 => La constante no es significativa.

e.1)

$$Var(e_{n+1}) = Var\left[(b_0 - b_0^*) + (b_1 - b_1^*)x_{n+1} + e_{n+1}\right]$$

$$= Var(b_0^*) + x_{n+1}^2 Var(b_1^*) + 2x_{n+1} Cov(b_0^*, b_1^*) + Var(e_{n+1})$$

$$= \frac{s^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2} + x_{n+1}^2 \frac{s^2}{\sum (x_i - \bar{x})^2} - 2x_{n+1} \frac{\bar{x} s^2}{\sum (x_i - \bar{x})^2} + s^2$$

$$= s^2 \left[1 + \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2} + \frac{nx_{n+1}^2}{n\sum (x_i - \bar{x})^2} - \frac{2nx_{n+1}\bar{x}}{n\sum (x_i - \bar{x})^2}\right]$$

$$= s^2 \left[1 + \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2} + \frac{nx_{n+1}^2}{n\sum (x_i - \bar{x})^2} - \frac{2nx_{n+1}\bar{x}}{n\sum (x_i - \bar{x})^2} + \frac{n\bar{x}^2}{n\sum (x_i - \bar{x})^2} - \frac{n\bar{x}^2}{n\sum (x_i - \bar{x})^2}\right]$$

$$= s^2 \left[1 + \frac{n\bar{x}^2}{n\sum (x_i - \bar{x})^2} - \frac{2nx_{n+1}\bar{x}}{n\sum (x_i - \bar{x})^2} + \frac{nx_{n+1}^2}{n\sum (x_i - \bar{x})^2} + \frac{\sum x_i^2 - n\bar{x}^2}{n\sum (x_i - \bar{x})^2}\right]$$

$$= s^2 \left[1 + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2} + \frac{1}{n}\right]$$

e.2)

$$\hat{y} = -0{,}103 + 0{,}554 * 8{,}5 = 4{,}606$$

$$Var(e_g) = 0{,}636\left[1 + \frac{1}{8} + \frac{(8{,}5 - 4{,}5)^2}{42}\right] \approx 0{,}96$$

Problema 6

Demostrar que en un modelo de regresión simple

a) Son algebraicamente equivalentes las expresiones siguientes:

$$\beta_1^* = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_1^* = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})Y_i}{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}$$

b) $R^2 = \beta_1^{*2} S_X^2 / S_Y^2$ donde:

R^2 : coeficiente de determinación

S_{xy} : covarianza muestral $= \Sigma(X_i - \bar{X})(Y_i - \bar{Y})/n$
S_x^2 : varianza muestral $= \Sigma(X_i - \bar{X})^2/n$
S_y^2 : varianza muestral $= \Sigma(Y_i - \bar{Y})^2/n$

Solución:

a) Desarrollando el numerador:

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y}) =$$

$$\sum_{i=1}^{n}(X_i Y_i - \bar{X} Y_i) - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} = \sum_{i=1}^{n}(X_{ii} - \bar{X})Y_i$$

b)

P.d.q.: $R^2 = \hat{b}_1^2 \frac{S_x^2}{S_y^2}$

Tenemos que $\hat{b}_1 = \frac{S_{xy}}{S_x^2}$

En efecto:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y})^2}$$

Se tiene que $\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{b}_0 - \hat{b}_1 x_i)^2 = \sum (y_i - (\overline{y} - \hat{b}_1 \overline{x}) - \hat{b}_1 x_i)^2$

$$= \sum (y_i - \overline{y} - \hat{b}_1 (x_i - \overline{x}))^2 = \sum (y_i - \overline{y})^2 + 2\hat{b}_1 \sum (y_i - \overline{y})(x_i - \overline{x}) + \hat{b}_1^2 \sum (x_i - \overline{x})^2$$

$$= nS_y^2 - 2n\hat{b}_1 S_{xy} + \hat{b}_1^2 nS_x^2 = nS_y^2 - 2n \frac{S_{xy}}{S_x^2} S_{xy} + \left(\frac{S_{xy}}{S_x^2} \right)^2 nS_x^2 = nS_y^2 - n \frac{S_{xy}^2}{S_x^2}, \text{luego:}$$

$$R^2 = 1 - \frac{\left[nS_y^2 - n \frac{S_{xy}^2}{S_x^2} \right]}{nS_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} = \hat{b}_1^2 \frac{S_x^2}{S_y^2}$$

Problema 7

Una compañía de telefonos celulares estudia la permanencia de sus clientes con el objeto de emprender algunas acciones. Se recogieron los datos de permanencia mensuales durante un período de 3 años. Suponiendo M el total de clientes al mes 0 se definen los datos recogidos como: $\{(x_i, y_i) / i = 0,1,2,...,35\}$ donde y_i es el número de clientes en el mes i (que quedan del total $M=y_0$ inicial); x_i es el mes ($x_i=i$).

Para explicar la permanencia de los clientes, se propone el modelo exponencial:

$$y = ae^{bx}$$

7.1) Interprete el modelo. En particular considere los cocientes y_{i+1}/y_i. Interprete el coeficiente b, precise su signo e interprete a.

7.2) Transforme el modelo de manera de encontrar un modelo lineal.

7.3) Los resultados del modelo lineal se presentan en la tabla. Complete la tabla. Opine sobre la validez local de los coeficientes del modelo. De los grados de libertad de las t-student.

7.4) Deduzca el modelo de permanencia exponencial aproximado. Comente. Estime el número de clientes del mes 0 inicial.

7.5) Después de cuantos meses se esperan que permanezcan solamente 1500 clientes?

Tabla

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	8.5092		196.3138	0.000
x	-0.0284	0.0020		0.000

Solución:

7.1 $\frac{y_{i+1}}{y_i}$ es la tasa de deserción de los clientes. Al realizar el cociente vemos que $\frac{y_{i+1}}{y_i} = \frac{a e^{b x_{i+1}}}{a e^{b x_i}} = e^b$ no depende de ***i.*** Es decir, la tasa de deserción es constante en el tiempo.

Luego $b = \log\left(\frac{y_{i+1}}{y_i}\right)$ que tiene que ser negativo ya que los y_i son decrecientes.

$a = M$ el número de clientes inicial.

7.2 $\log(y) = \log(a) + bx = \ln(M) - b + bx$

7.3

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	8.5092	***0.0433***	196.3138	0.000
x	-0.0284	0.0020	***-13.9119***	0.000

Tengo 36 pares de datos y 2 variables que estimar por tanto los grados de libertad de las t-Student son *34*. Los coeficientes son claramente significativos (Pvalor = 0), más aún el coeficiente de correlación entre los datos observados y estimados es casi igual a 1.

Recuerden que: $t_{n-r} = \frac{\hat{b}_j}{\hat{s}_{\hat{b}_j}}$ lo que permite completar la tabla.

7.4 $y = e^{8.5092} e^{-0.0284x} = 4960.3e^{-0.0284x}$. Se estima el número de clientes del primer mes como $e^{8.5092} = 4960$.

7.5 $x = \frac{1}{b}(\log(y) - \log(a)) = -(\log(1500) - 8.5092)/0.0284 \approx 42 meses$.

Problema 8

Consideremos un modelo lineal de regresión simple: $y = b_0 + b_1 x + e$

10.1 Dé la expresión de los estimadores mínimos cuadrados $\hat{b}_0 y \hat{b}_1$.

10.2 Se hace el cambio de variable z=10x. Obtenga los estimadores del nuevo modelo:

$y = g_0 + g_1 z + e$ en función de $\hat{b}_0 y \hat{b}_1$. Compare $y = \hat{b}_0 + \hat{b}_1 x$ con $\tilde{y} = \hat{g}_0 + \hat{g}_1 z$

10.3 Dé el estadístico del test $H_0 : b_1 = 0$ y explique qué indica cuando se rechaza.

Solución:

10.1 $\hat{b}_1 = \dfrac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$ y $\hat{b}_o = \bar{y} - \hat{b}_1 \bar{x}$

10.2 $\hat{g}_o = \hat{b}_o$ y $\hat{g}_1 = \hat{b}_1 / 10$. Por lo cual $\hat{y} = \hat{b}_o + \hat{b}_1 x$ y $\tilde{y} = \hat{g}_o + \hat{g}_1 z$ son iguales.

10.3 El estadístico del test $H_o : b_1 = 0$ es $\dfrac{\hat{b}_1}{\hat{s}_1}$ que sigue una t-Student a *n-2* grados de libertad bajo la hipótesis nula. Rechazar la hipótesis nula indica que hay una cierta influencia de la variable *x* sobre la variable *y*. Se podría usar después de estimar el modelo para hacer predicciones.

Problemas de Exámenes

27.11.99 **EXAMEN**

MA 34B

PROBLEMA 1

Un banco ha decidido mejorar su sistema de atención al público. Suponga que la variable aleatoria *x* representa el número de clientes atendidos en el banco durante una mañana y que ésta tiene una distribución Poisson con parámetro desconocido q.

1.1 Suponga que el banco tiene información a priori con respecto al parámetro θ. El banco cree que este parámetro tiene una distribución Gamma(r, l). Si la función de pérdida del banco es cuadrática, demuestre que el estimador Bayesiano de q está dado por $\hat{q}_B = \frac{l\,(\sum_{i=1}^{n} x_i + r)}{ln+1}$, donde *n* es el tamaño de la muestra.

1.2 Demuestre que el estimador de q de máxima verosimilitud está dado por $\hat{q}_{MV} = \sum_{i=1}^{n} x_i / n$.

1.3 Determine si los estimadores de máxima verosimilitud y de Bayes son insesgados y consistentes para q.

1.4 Demuestre que el estimador de máxima verosimilitud es de mínima varianza entre los estimadores insesgados de q.

1.5 Calcule el error cuadrático medio de ambos estimadores ($E[(\hat{q}_B - q)^2]$ y $E[(\hat{q}_{MV} - q)^2]$).

PROBLEMA 2

La compañía de cementos CONCRETESA fabrica cementos. La dirección de la empresa está preocupada, ya que a pesar de la tendencia de precios, el volumen de ventas no sólo no se recupera sino que declina constantemente con grave impacto en la facturación de la misma. Descartando que la razón última de la situación sea una crisis en el sector de la construcción (datos a fines de 1997), el análisis se centra en la identificación de las razones internas que clarifiquen las causas.

El estudio pretende analizar las dos causas siguientes:

(a) Falta de productividad

(b) Reclamaciones de clientes: Un elevado número de reclamaciones reflejaría problemas de calidad y justificaría la caída de las ventas

Las tablas siguientes reflejan la productividad obtenida en la compañía (Tn/empleado) así como los rechazos recibidos por problemas de calidad en los últimos 12 meses. La compañía tiene valores históricos considerados como normales en la productividad (9.7 Tn/empleado) y en % de rechazos (0.75%).

2.1 ¿Puede afirmarse que la productividad de la planta es inferior a la considerada standard? (α=5%).

2.2 ¿Puede concluirse que existen problemas de calidad en el proceso de fabricación? (α=5%).

Mes	**Tn/empleado**	**% Rechazos sobre nº de pedidos**
1	10	1
2	10,5	1,25
3	9,5	1
4	8	1,25
5	9	1,5
6	9,5	1,75
7	9	1,5
8	9,25	1
9	9,5	1,5
10	9,5	1,5
11	10	1,25
12	9,5	1,25
Media	9,4375	1,3125
Des. Típica	0,5962	0,2310
	Productividad Standard	Rechazos considerados normales
	9.7	0.75

Ante la sospecha de que existan problemas de calidad se ordena hacer un estudio detallado en cada planta de la empresa obteniéndose los siguientes resultados en diez días de inspección, (% sobre nº de pedidos):

Muestra	Planta 1	Planta 2
1	1	1,2
2	0,75	0,75
3	0,5	1,2
4	0,4	1,4
5	0,75	1,4

6	1,24	0,8
7	1,2	1,2
8	0,75	0,2
9	0,8	0,6
10	0,94	0,8
Media	0,83	0,96
Varianza (Sobre n-1)	0,072	0,151

2.3 ¿Puede indicarse que existen diferencias significativas en los rechazos en función de la planta de donde provenga la producción?(α=5%)

PROBLEMA 3

Consideramos cuatro variables antropométricas para 30 niñas: y la talla; x_1 el peso, x_2 el largo del tronco y x_3 el perímetro del cráneo. Estudiamos la relación que existe entre ellas a partir del modelo: $y_i = b_o + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + e_i$. Se obtuvieron los siguientes resultados:

Variable	Coeficiente	Desviación típica	t-Student
Constante	-733.38	231.63	-3.166
Peso	0.0029	0.0147	0.195
tronco	2.052	0.241	8.509
Cráneo	1.146	0.577	1.984

Coeficiente de correlación múltiple R^2 = *0.988*	F de Fisher = *742.75*

3.1 Interpretando el coeficiente de correlación múltiple R^2 y el F de Fisher, concluye si el modelo es significativo. Vean en particular si todas las variables explicativas son realmente significativas.
3.2 Dé intervalos de confianza a 95% para los coeficientes b_1 y b_2. Comente.

Solución:

PROBLEMA 1

1.1 La f. de verosimilitud es: $f_n(x|q) \propto q^{\sum x_i} e^{-nq}$ y la f. d. a priori es: $p(q) \propto q^{r-1} e^{-q/l}$. La función de ensidad a posteriori es entonces:

$$x(q \mid x) \propto q^{\sum x_i} e^{-nq} q^{r-1} e^{-q/l} = q^{\sum x_i + r - 1} e^{-q(n+1/l)}$$

⇨ es una *Gamma* $(\sum x_i + r, \frac{l}{1+nl})$. El estimador de Bayes bajo una función de perdida cuadrática es la esperanza de esta distribución:

$$\hat{q}_B = \frac{l\,(\sum_{i=1}^{n} x_i + r)}{ln+1}$$

1.2 $f_n(x|q) \propto q^{\sum x_i} e^{-nq} \Rightarrow \frac{\partial Log(f_n)}{\partial q} = \frac{\sum x_i}{n} - n \Rightarrow \hat{q}_{MV} = \bar{x}$.

1.3 El estimador de Bayes es sesgado y consistente y el estimador de Máxima verosimilitud es insesgado y consistente. En efecto:

$\hat{q}_B = \frac{l\,(\sum_{i=1}^{n} x_i + r)}{ln+1} = \frac{n\bar{x}}{n+1/l} + \frac{r}{n+1/l}$. Como $\bar{x} \to E(X) = q$ y $\frac{n}{n+1/l} \to 1$ y $\frac{r}{n+1/l} \to 0$

cuando $n \to +\infty$. Luego $\boxed{\hat{q}_B \to q}$ es decir que $\hat{q}_B$ es consistente para q.

$E(\hat{q}_B) = E\left(\frac{\sum x_i + r}{n+1/l}\right) = \frac{nq}{n+1/l} + \frac{r}{n+1/l}$ => $\hat{q}_B$ es sesgado.

$\hat{q}_{MV} = \bar{x} \to q$ => es consistente. Además $E(\hat{q}_{MV}) = E(\bar{x}) = q$ => es insesgado.

1.4 Hay que usar la desigualdad de Cramer-Rao: $Var(\hat{q}) \geq -\frac{1}{E\left(\frac{\partial^2 Log(f_n)}{\partial q^2}\right)}$ para todo estimador $\hat{q}$ insesgado de q. Calculemos la cota inferior de la varianza:

$\frac{\partial Log(f_n)}{\partial q} = \frac{\sum x_i}{q} - n$ y $\frac{\partial^2 Log(f_n)}{\partial q^2} = -\frac{\sum x_i}{q^2}$ => $-E\left(\frac{\partial^2 Log(f_n)}{\partial q^2}\right) = \frac{E(\sum x_i)}{q^2} = \frac{nq}{q^2} = \frac{n}{q}$.

Como $Var(\hat{q}_{MV}) = Var(\bar{x}) = \frac{Var(X)}{n} = \frac{q}{n}$. Luego $\hat{q}_{MV}$ es de mínima varianza entre los estimadores insesgados de q.

1.5 $E[(\hat{q}_{MV} - q)^2]$ = sesgo^2 + Varianza = $0 + \frac{q}{n}$.

$E[(\hat{q}_B - q)^2] = (q - \frac{nq}{n+1/l} + \frac{r}{n+1/l})^2 + Var(\frac{n\bar{x}}{n+1/l} + \frac{r}{n+1/l})$

$$E[(\hat{q}_B - q)^2] = \left(\frac{q + rl}{1 + nl}\right)^2 + \frac{n l^2}{(1 + nl)^2}.$$

PROBLEMA 2

2.1 Realizaremos un contraste de la media muestral con las siguientes hipótesis :

Ho. : μ=9,7 (la productividad no es inferior a la media)
H1. : μ<9,7 (la productividad es inferior a la media)

$$t_{n-1} = \frac{X - m}{\hat{S} / \sqrt{n}} = -\ 1.46$$

El valor crítico se obtiene a partir de $P(t_{11} < -1.8) = 0.05$ para un error de tipo I de 5% : -1.8
Por lo tanto no se puede decir que la productividad es inferior a la media.

2.2 Realizaremos un contraste de la media muestral con las siguientes hipótesis :

Ho. : μ=0,75 (los rechazos no son superiores a la media)
H1. : μ>0,75 (los rechazos son superiores a la media)

$$t_{n-1} = \frac{X - m}{\hat{S} / \sqrt{n}} = 8.039$$

El valor crítico se obtiene a partir de $P(t_{11} < 1.8) = 0.05$ para un error de tipo I de 5% .
Por lo tanto no se puede decir que los rechazos son superiores a la media.

PROBLEMA 3

3.1 El p-valor de la F de fisher (con 3 y 26 g.l.) es casi nula, luego las variables explican algo de la talla. Además el coeficiente de correlación múltiple es elevado. Pero parecería que el peso no es significativo, dado que su p-valor asociado es igual a 0.84, mientras que los otros son 0 y 0.05 respectivamente para el tronco y el craneo. Sin embargo habría que ver los coeficientes de correlación parcial y las correlaciones entre las variables explicativas, para poder concluir si no hay un efecto de multicolinearidad.

3.2 Intervalos de confianza
peso: [-0.0273, 0.0331]
tronco: propuesto.

PROBLEMA 1

Parte A

Sea una muestra bivariada $\{(x_i, y_i) / i = 1,...n\}$.

1.1 Define el coeficiente de correlación lineal empírico *r* y dé su recorrido. ¿Que mide? ¿Cuando *r* toma el valor *+1*? ¿Cuando r toma un valor *–1*? Dé todas las situaciones posibles.

1.2 Se supone ahora que *x* define dos poblaciones diferentes (*x* toma el valor *1* para la población P_1 y 0 para la población P_2). Se llaman $\bar{y}_1$ la media de la población P_1 e $\bar{y}_2$ para P_2. Se plantea el modelo lineal: $y_i = b_o + b_1 x_i + e_i$. Dé el estimador de mínimos cuadrados y deduzca que el modelo equivale a: $E(y) = \bar{y}_1$ si x=1 y $E(y) = \bar{y}_2$ si x = 0.

Parte B

Se considera el modelo lineal: $y = Xb + e$ con el supuesto $e \sim N_n(0, s^2 W)$ donde W es una matriz de orden *n* invertible distinta de la identidad.

1.3 Consideramos el estimador de Mínimos Cuadrados Ordinarios (MCO) $\hat{b} = (X^t X)^{-1} X^t y$. Determine si es insesgado y calcule su varianza.

1.4 Consideramos el estimador alternativo $b^* = (X^t W^{-1} X)^{-1} X^t W^{-1} y$. Determine si es insesgado y calcule su varianza.

1.5 Se supone ahora que tenemos una sola variable exógena (o explicativa) *x* y que Ud. sabe que la matriz $s^2 W$ está dada por $s^2 W_{ij} = E(e_i e_j) = \begin{cases} s^2 x_i^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$

Demuestre que $Var(\hat{b}) \geq Var(b^*)$. Discuta las consecuencias de este resultado para el teorema de Gauss-Markov. (Ayuda: utilice la desigualdad de Cramer-Rao o bien calcule las respectivas varianzas para esta caso y luego defina una nueva variable $z_i = x_i^2$ y para la comparación de las varianzas, use la definición de varianza z_i).

PROBLEMA 2

2.1 Comente la matriz de los coeficientes de correlación dados en la tabla 2 obtenidos sobre 106 países y analice los resultados globales de la regresión de la tasa de crecimiento de la población sobre las 4 otras variables (Tabla 3).

2.2 Calcule el p-valor que permite evaluar la calidad global del modelo y concluye.

2.3 ¿Qué estadísticos de la tabla permiten evaluar la calidad individual de los coeficientes del modelo? Concluye sobre la importancia de cada variable en el modelo. Comente considerando la matriz de correlaciones.

2.4 Se hace la regresión del crecimiento de población sobre la tasa de mortalidad infantil y la tasa de natalidad (Tabla 4). Justifique esta regresión y comente los resultados. Se precisará los grados de libertad de la F de Fisher.

2.5 A partir de este último modelo de regresión (tabla 4<9 estime el crecimiento de población de un país que tiene una tasa de natalidad igual a *25.9* y una tasa de mortalidad infantil de *42.3* y dé un intervalo de confianza a *95%* usando la tabla 5.

Tabla 2: matriz de correlación

	Crecimiento población	Mortalidad infantil	PNB/capita	Tasa natalidad	Fertilidad
Crecimiento población	1.000	0.60	- 0.52	0.86	0.84
Mortalidad infantil	0.60	1.000	-0.65	0.86	0.83
PNB/capita	- 0.52	-0.65	1.000	-0.66	-0.58
Tasa natalidad	0.86	0.86	-0.66	1.000	0.98
Fertilidad	0.84	0.83	-0.58	0.98	1.000

Tabla 3: resultados de la regresión

	Estimación Coeficiente	Desviación típica	T-Student	p-valor
Constante	-0.987	0.223	-4.430	0.000
Mortalidad infantil	-0.018	0.003	-6.734	0.000
PNB/capita	4.94209E-7	1.1175E-5	0.044	0.9645
Tasa natalidad	0.139	0.022	6.273	0.000
Fertilidad	-0.05	0.128	-0.392	0.696
Coeficiente de correlación múltiple R=0.90			F de Fisher = 117.48	

Tabla 4: resultados de la regresión

	Estimación Coeficiente	Desviación típica	T-Student	p-valor
Constante	-0.960	0.133	-7.236	0.000
Mortalidad infantil	-0.018	0.003	-6.983	0.000
Tasa natalidad	0.131	0.008	16.607	0.000
Coeficiente de correlación múltiple R=0.90			F de Fisher = 246.66	

Tabla 5: $\hat{s}^2(X^tX)^{-1} = 10^5\begin{pmatrix} 0.672 & -1.787 & 17.861 \\ -1.787 & 6.377 & -89.90 \\ 17.861 & -89.90 & 1828.2 \end{pmatrix}$

Solución:

PROBLEMA 1

Parte A

1.1 El coeficiente de correlación lineal empírico es igual a: $r = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

Permite medir el grado de relación lineal que existe entre dos variables. Varia entre –1 y +1.

Cuando vale +1, existe una relación lineal estricta entre las variables de pendiente positiva. Cuando es cercano a +1, existe una relación entre las dos variables tal que cuando una de las variables crece, la otra crece también de manera casi lineal, etc...

1.2 Se puede construir la matrix $X^t = \begin{pmatrix} 1 & 1 & 1 & ... & 1 \\ 1 & 1 & ... & 0 & 0 \end{pmatrix}$, $X^tX = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix}$ y

$(X^tX)^{-1} = \dfrac{1}{n_1 n_2}\begin{pmatrix} n_1 & -n_1 \\ -n_1 & n \end{pmatrix}$. Usando que $n\bar{y} = n_1\bar{y}_1 + n_2\bar{y}_2$, se deduce que $\hat{b}_0 = \bar{y}_2$ y $\hat{b}_1 = \bar{y}_1 - \bar{y}_2$. Se llega al mismo resultado derivando la suma de los cuadrados de los errores: $\sum(y_i - b_o - b_1x)^2$. Usando $E(y) = \hat{b}_o + \hat{b}_1x$ y el hecho que $x = 1$ o 0, se obtiene

$E(y) = \bar{y}_1$ si x=1 y $E(y) = \bar{y}_2$ si x=0.

Parte B

1.3 El modelo: $y = Xb + e$ con el supuesto $e \sim N_n(0, s^2W)$ donde W es una matriz de orden n invertible distinta de la identidad => $y = Xb$ y $Var(y) = s^2W$.

$E(\hat{b}) = (X^tX)^{-1}X^tE(y) = (X^tX)^{-1}X^tXb = b$ => $\hat{b}$ es insesgado

$Var(\hat{b}) = (X^tX)^{-1}X^tVar(y)X(X^tX)^{-1}$

$$\boxed{Var(\hat{b}) = s^2(X^tX)^{-1}X^tW^{-1}X(X^tX)^{-1}}$$

1.4 $E(b^*) = (X^tW^{-1}X)^{-1}X^tW^{-1}E(y) = (X^tW^{-1}X)^{-1}X^tW^{-1}Xb = b$ => b^* es insesgado.

$$Var(b^*) = (X^tW^{-1}X)^{-1}X^tW^{-1}Var(y)W^{-1}X(X^tW^{-1}X)^{-1}$$
$$= s^2(X^tW^{-1}X)^{-1}X^tW^{-1}WW^{-1}X(X^tW^{-1}X)^{-1}$$

$$= s^2 (X^t W^{-1} X)^{-1} X^t W^{-1} X (X^t W^{-1} X)^{-1}$$

$$\boxed{Var(b^*) = s^2 (X^t W^{-1} X)^{-1}}.$$

1.5 $s^2 W_{ij} = E(e_i e_j) = \begin{cases} s^2 x_i^2 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$

PROBLEMA 2

2.1 Se observan correlaciones ..., positivas o negativas, blabla..
2.2 El p-valor es $P(F_{4,101} > 117.48) \approx 0.00$.

2.3 Son t-Student con *101* grados de libertad. Los p-valores indican que solo la mortalidad infantil y la tasa de natalidad son significativa en el modelo. Esto se debe a la autocorrelación de las variables explicativas.

2.4 En este modelo de regresión ambas variables resultan significativas, el coeficiente de correlación múltiple es el mismo y el p-valor de la F es: $P(F_{2,103} > 246.66) \approx 0.0$.

2.5 La predicción es: y_o*=-0.96-0.018*42.3+0.131*25.9=1.67.* El intervalo de confianza es:

$[y_o - 1.96 s_o, y_o + 1.96 s_o]$ con $s_o = \sqrt{x_o^t V x_o}$ y $V = \hat{s}^2 (X^t X)^{-1}]$.
El intervalo: *[1.57,1.78]*.

05.07.2002 **EXAMEN**
MA34B
Duración: 3 horas

PREGUNTA 1

Sea una muestra aleatoria simple $X_1, X_2, ..., X_n$ de una distribución con densidad

$$f(x) = \begin{cases} e^{-(x-b)} & si \quad x \geq b \\ 0 & si \quad x < b \end{cases}$$

1.1 Calcule la esperanza y la varianza de X (Se recomienda usar la función generatriz de los momentos de X).

1.2 Encuentre el estimador de Máxima Verosimilitud $\hat{b}_2$ de b. ¿Es asintoticamente insesgado? ¿Es consistente? (Se recomienda usar la función generatriz de los momentos de $\hat{b}_2$).

PREGUNTA 2

Una empresa de camiones de carga sospecha que el ciclo de vida m de ciertos neumáticos es de menos de 32000 km. Para verificar este argumento, la empresa instala 30 de esos neumáticos en sus camiones y obtiene un ciclo medio de 31460 km con una desviación típica S_n de 900 km ($S_n = \sqrt{\frac{1}{n}\sum (x_i - \overline{x})^2}$).

2.1 Dé el p-valor del test $H_o : m \geq 32000$ contra $H_1 : m < 32000$. Concluye. Precisan los supuestos que tuvieron que hacer para efectuar el test.

2.2 Dé un intervalo de confianza para m de nivel de confianza igual a *0.95.* Concluye sobre la hipótesis $H_o : m \geq 32000$. Si se aumenta el nivel de confianza como cambia el largo del intervalo.

PREGUNTA 3

Se desea explicar la esperanza de vida de mujeres adultas en Chile con un modelo lineal a partir de las variables explicativas "gasto en salud", "calorías consumidas" y "tasa de alfabetización". A partir de *25* observaciones, se efectúa la regresión de la esperanza de vida (de media *67.11* años y desviación típica de *5.52*) sobre las tres otras variables; se obtiene los resultados en la Tabla 1:

Tabla 1

Variable	Media	Des. Tipica	coeficiente	Des. Tipica Coeficiente	T-student	P(|X|>T)
Constante	2.09	1.567	29.603	5.282	5.600	0.0001
Gasto salud	108.80	14.420	0.959	0.417	2.302	0.0336

Calorías Alfabetización	83.00	11.947	0.161 0.217	0.054 0.064	2.980 3.350	0.0085 0.0041

F-Fisher para las tres variables: 19.714

3.1 Dé los grados de libertad del F-Fisher y el p-valor asociado. Concluye.
3.2 Interprete los T-Student . Dé los grados de libertad.
3.3 Dé una estimación insesgada de s^2, la varianza de los errores del modelo. ¿Cuánto vale?

PREGUNTA 4

Se considera un grupo de *200* personas. Cada una lanza una moneda *5* veces.
4.1 Si la moneda fuera equilibrada, cual porcentaje de personas deberían tener *0* cara, *1* cara, ..., *5* caras.
4.2 Observando el resultado del experimento (Tabla 2), ¿puede concluir que la moneda es equilibrada con un nivel de significación de $\alpha=0.05$?
4.3 ¿El p-valor del test es mayor o menor que $\alpha=0.05$?

Tabla 2

Nº caras	0	1	2	3	4	5	Total
Nº personas	10	20	50	80	300	10	200

Solución:

PREGUNTA 1

1.1 La función generatriz de X es:

$$g(t) = E(e^{tX}) = \int_b^{+\infty} e^{tx} e^{-(x-b)} dx = \int_b^{+\infty} e^{b} e^{(t-1)x} dx = \frac{e^{tb}}{1-t} \quad \text{para} \quad t < -1$$

$$g'(t) = \frac{be^{tb}}{1-t} + \frac{e^{tb}}{(1-t)^2} \qquad g''(t) = \frac{b^2 e^{tb}}{1-t} + \frac{2be^{tb}}{(1-t)^2} + \frac{2e^{tb}}{(1-t)^3}$$

Luego: $g(0) = 1 \quad E(X) = g'(0) = b + 1 \quad E(X^2) = g''(0) = b^2 + 2b + 2$

=> $E(X) = b + 1$ y $Var(X) = 1$

1.2 La función de verosimilitud es:

$$f_n(x_1, ..., x_n) = \begin{cases} e^{-\sum (x_i - b)} & si \quad \forall x_i \geq b \\ 0 & si \quad no \end{cases}$$

Para que f_n sea máxima, b tiene que ser el mayor posible dentro los limites permitidos. Luego $\hat{b}_2 = Min\{x_i\}$. Calculemos la función de distribución *H* y la función de densidad *h* de $\hat{b}_2$:

$H(y) = 1 - (1 - F(y))^n = 1 - e^{n(b-y)} \quad \forall y \geq b$ y $h(y) = ne^{n(b-y)} \quad \forall y \geq b$.

Calculemos la función generatriz de los momentos de $\hat{b}_2$:

$$g(t) = E(e^{ty}) = \int_b^{+\infty} ne^{ty} e^{n(b-y)} dy = \int_b^{+\infty} ne^{(t-n)y} e^{nb} dy = \frac{n}{n-t} e^{tb} \quad \forall t < n$$

$$g'(t) = \frac{bne^{tb}}{n-t} + \frac{ne^{tb}}{(n-t)^2} \quad g''(t) = \frac{b^2 ne^{tb}}{n-t} + 2\frac{b\ ne^{tb}}{(n-t)^2} + \frac{2ne^{tb}}{(n-t)^3}$$

Luego $E(\hat{b}_2) = g'(0) = b + \frac{1}{n}$; $Var(\hat{b}_2) = g''(0) = b^2 + \frac{2b}{n} + \frac{2}{n^2}$.

Se concluye que $\hat{b}_2$ es un estimador asintoticamente insesgado con $Var(\hat{b}_2) \to 0$. Es cosnistente para $\hat{b}_2$.

PREGUNTA 2

2.1 La forma de la región crítica es: $\bar{x} \leq c$. El p-valor es $P(\bar{x} \leq 31460 \mid m = 32000)$.

Hay que suponer que $x \sim N(m, s^2)$ => $\bar{x} \sim N(m, s^2/n)$. Luego $\frac{\bar{x} - m}{\sqrt{s^2/n - 1}} \sim t_{n-1}$.

Bajo $m = 32000$, $\frac{(\bar{x} - 32000)}{\sqrt{s^2 / n - 1}} \sim t_{n-1}$.

p-valor= $P(t_{n-1} \leq \frac{(31460 - 32000)}{900 / \sqrt{29}}) = P(t_{n-1} \leq -3.123 = 0.0015$. Se puede rechazar $H_o : m \geq 32000$.

2.2 El intervalo de confianza es de la forma:

$[\bar{x} + t_{29}^{0.025} \frac{s}{\sqrt{29}}, \bar{x} - t_{29}^{0.025} \frac{s}{\sqrt{29}}]$=[31118,31802]. Se puede confirmar que se rechaza $H_o : m \geq 32000$ dado que se encuentra todo el intervalo menor que *32000*.

Si se aumenta el nivel de confianza crece el largo del intervalo.

PREGUNTA 3

3.1 Los grados de libertad son: *3 y 21.* El p-valor: $P(F_{3,21} > 19.714) = 0.000$. El modelo es significativo.

3.2 El grados de libertad de las t es *21*. Son todos significativos.

3.3 $\hat{s}^2 = \sum_i \hat{e}_i^2 / 21$ es insesgado para s^2 . Como $F = \frac{(\sum(y_i - \bar{y})^2 - \sum \hat{e}_i^2) / 3}{\sum \hat{e}_i^2 / 21}$. Se deduce que $\hat{s}^2 = \frac{\sum_i \hat{e}_i^2}{21} = \frac{nvar(y)}{3F + 21} = 9.505$

PREGUNTA 4

4.1 SI la moneda fuera equilibrada la distribución del número de caras obtenidos en 5 lanzamientos es una binomial (5,0.5). La probabilidad de sacar k caras en 5 lanzamientos es: $P(x = k) = \binom{5}{k}(0.5)^k (0.5)^{5-k} = \binom{5}{k}(0.5)^5 = 0.03125 * \binom{5}{k}$

4.2

Nº caras	0	1	2	3	4	5	Total
Proba p_i	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125	1.00
frec.teorica np_i	6.2500	31.25	62.50	62.50	31.25	6.2500	200
frec. observ. f_i	10	20	50	80	30	10	200

El estadístico es: $Q = \sum_i \frac{(f_i - np_i)^2}{np_i} \sim c_5^2$. Aquí *Q=16.* La región crítica del test es: $Q > a$ con $P(c_5^2 > a) = 0.05$. $Q > 11.07$ => se rechaza que es equilibrada.

4.3 El p-valor es más pequeño que 0.05.

Problema 3 EXAMEN 2003/02

Se hizo un estudio sobre el nivel de desarrollo poblacional en el uso de las tecnologías de información, para lo cual se recolectó la siguiente información para 10 países:

Paises	Area (Km2)	Población (millones)	Tasa Desempleo	Computadores por persona
E.E.U.U.	9629091	278,1	4,0	5,40
Namibia	825418	1,8	35,0	0,03
Francia	547030	59,6	9,7	0,89
Luxemburgo	2586	0,4	2,7	0,66
Finlandia	337030	5,2	9,8	1,20
Laos	236800	5,6	5,7	0,01
Chile	756950	15,3	8,0	0,56
Zimbawe	390580	11,4	50,0	0,03
Japan	377835	126,8	4,7	6,20
Kenia	582650	30,8	50,0	0,02

Para el estudio se utilizó un Análisis de Componentes Principales (ACP), del cual se presentan resultados en las tablas 5, 6 y 7):

Tabla 5

	Media	Desviación Estándar
Area	1368597,000	2763092,637
Población	53,504	83,497
Desempleo	17,960	18,252
Computadores	1,500	2,193

Tabla 6: Matriz de correlaciones

	Área	Población	Desempleo	Computadores
Área	1	0,895	-0,222	0,581
Población	0,895	1	-0,324	0,849
Desempleo	-0,222	-0,324	1	-0,471
Computadores	0,581	0,849	-0,471	1

Tabla 7: Valores y vectores propios normalizados

	$\lambda 1$	$\lambda 2$	$\lambda 3$	$\lambda 4$
Valor propio	2,753	0,874	0,353	0,020
	U1	U2	U3	U4
Área	0,517	0,393	-0,595	-0,473
Población	0,584	0,236	0,047	0,775
Desempleo	-0,318	0,881	0,346	-0,050
Computadores	0,539	-0,113	0,724	-0,415

3.1 Explique en qué consiste el análisis de componentes principales y cuáles son sus principales aplicaciones.
3.2 Interprete los resultados de las tablas 5 y 7.
3.3 Dibuje el círculo de correlaciones para los dos primeros ejes principales e interprete el resultado. Explique porqué es conveniente quedarse solo con estos.
3.4 Encuentre las componentes principales de cada país asociadas a los primeros ejes principales y grafique aproximadamente. Interprete los resultados.

3.5 Determine las contribuciones porcentuales de cada variable original en la construcción del eje.

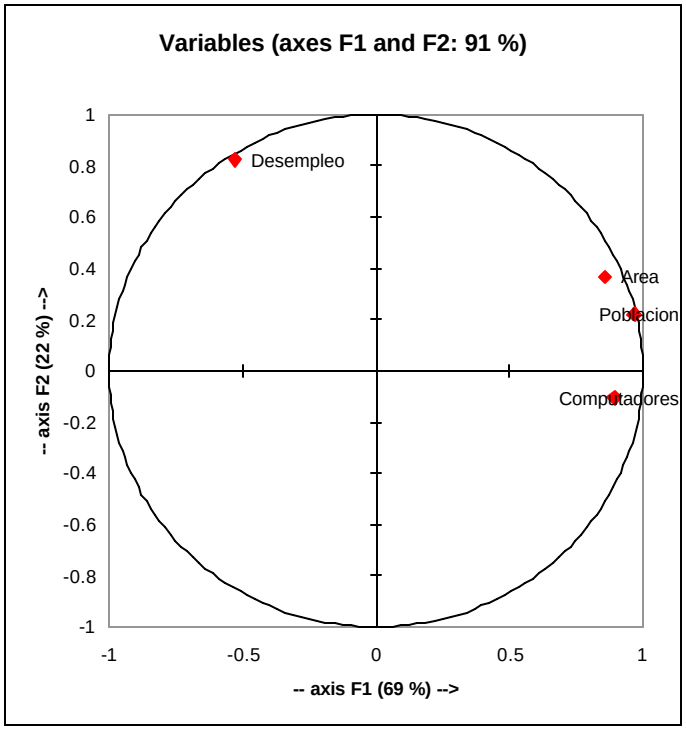
Solución:

Es técnica que permite describir un conjunto de información en función de un conjunto menor de variables no correlacionadas, tratando de preservar la mayor cantidad de variabilidad de los datos originales. Dentro de las principales aplicaciones del ACP, se encuentran: construcción de índices compuestos, reducción de multicolinealidad en modelos lineales y clasificación estadística entre otras.

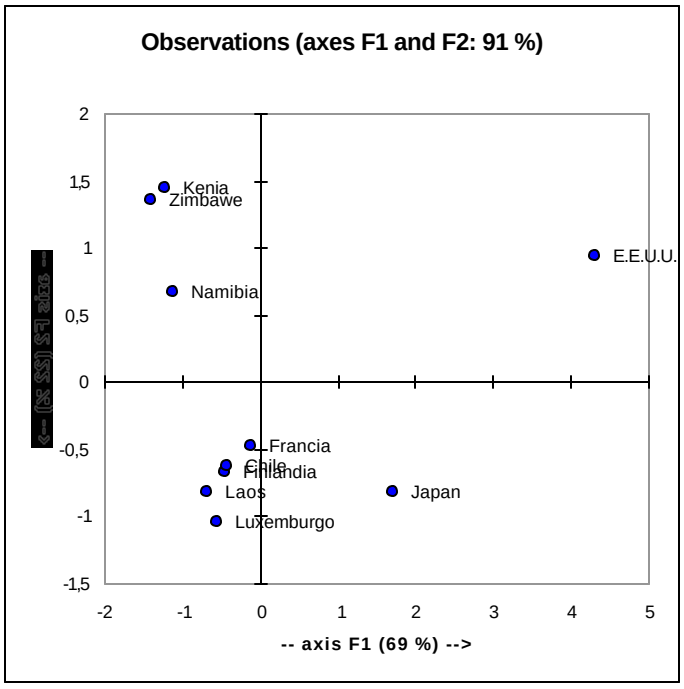
La tabla 5 muestra los valores promedio y desviaciones estándar de las variables originales, a partir de la cual se puede observar que los datos difieren bastante en escala, por lo que es importante trabajar con los datos estandarizados, y utilizar la matriz de correlaciones para llevar a cabo el ACP

Coordenadas de las variables en el círculo:

	U1	U2
Area	0,858	0,368
Poblacion	0,969	0,221
Desempleo	-0,528	0,824
Computadores	0,895	-0,105



Se puede observar que las variables Área, Población y Computadores están más vinculadas al primer eje principal, al cual podemos denominar tentativamente como “cobertura tecnológica”, mientras que el segundo eje está más relacionado al aspecto económico de los países, eje que podría denominarse “capacidad de trabajo” (leído en sentido inverso, es decir, de arriba hacia abajo). Las variables originales se encuentran cerca de la frontera del círculo, lo que les confiere un alto poder explicativo en la construcción de los ejes, es decir, se espera un elevado coeficiente de determinación. Debería esperarse que los países más pobres se encuentren cerca del II cuadrante, mientras que los países más desarrollados deberían encontrarse cerca del IV cuadrante.



El gráfico muestra precisamente lo que se intuye del círculo de correlaciones. Países como Japón y E.E.U.U. son los países con cobertura tecnológica de información (están más a la derecha en el primer eje principal), mientras que los países africanos tienen la menor puntuación, por ser los más poblados, con mayor desempleo y menor cantidad de computadores por persona, por ende tiene menor cobertura tecnológica y menor capacidad de trabajo. Ahora, E.E.U.U. se encuentra en el I cuadrante básicamente porque es un país altamente poblado y con mucha área geográfica. Los países intermedios como Chile, Francia y Luxemburgo tienen menor cobertura tecnológica que E.E.U.U. y Japón básicamente porque tienen una menor población y área geográfica; al igual que E.E.U.U. se diferencia de Japón por el mismo motivo.

Cosenos cuadrados de las variables:

	F1	F2
Area	0,735	0,135
Población	0,938	0,049
Desempleo	0,279	0,679
Computadores	0,801	0,011

Contribución (%):

	F1	F2
Area	26,711	15,460
Población	34,082	5,592
Desempleo	10,125	77,677
Computadores	29,082	1,272

Los cuadrados de las coordenadas de las variables originales en el plano principal determinan el aporte para cada eje, del cual se pueden deducir los aportes porcentuales de las mismas. Se puede observar que la población y el número de computadores son los principales responsables del valor obtenido en cobertura tecnológica, lo cual es bastante esperable desde el punto de vista intuitivo.

EXAMEN 2004/01

PROBLEMA 1

Se considera un examen de sangre en ***500*** pacientes antes y después de un tratamiento. Llamemos ***x*** el examen antes del tratamiento e ***y*** el examen después del tratamiento. Se busca estudiar la efectividad del tratamiento. Diremos que el tratamiento es efectivo para un paciente cuando ***y>x***.

a) ¿Qué gráfico propondría hacer para ayudar al estudio? Comente.
b) Suponiendo que ***x*** e ***y*** siguen distribuciones normales, dé la distribución de la diferencia: ***d=y-x***.
c) Estime los parámetros de la distribución de ***d*** utilizando los datos de la tabla 1.
d) Construya un intervalo de confianza al 95% para la diferencia media *d=E(d)*. Interprete.
e) Efectué un test de hipótesis para $H_o : d = 0$ contra $H_1 : d > 0$ con un error de tipo I de 5%. ¿El tratamiento fue efectivo?

Tabla1

	Media	Desviación estándar	Varianza	Covarianza entre x e y	$\sqrt{0.0087 / 500}$
ANTES	8.01	0.1962	0.0385	0.0381	0.00417
DESPUÉS	8.91	0.2155	0.0464		

PROBLEMA 2

Se estudia la relación entre el Coeficiente Intelectual (CI) y el rendimiento escalar de los 30 alumnos de un curso de sexto básico: la variable y representa el CI, x_1 la nota de castellano, x_2 la nota de matemática, x_3 la nota de biología y x_4 la nota de Inglés. Se presentan la matriz de las correlaciones de todas las variables en la tabla 2 y los resultados de 2 modelos de regresión lineal distintos en las tablas 3 y 4.

a) Analicé los resultados de la regresión dados en la tabla 3. Considerando la tabla 2 ¿qué opina del efecto de la nota de matemática en el modelo?
b) Dé un intervalo de confianza de nivel 95% para el coeficiente b_1 de la nota de castellano.
c) Realice el test $H_0 : b_1 = 4$ contra $H_1 : b_1 < 4$ con un error de tipo I de 5%.
d) En el modelo de la tabla 4, se eliminó la variable "Matemática". Examine y compare las dos regresiones.
e) ¿Cuáles son los supuestos usuales sobre los errores de un modelo lineal? ¿Cómo se relacionan estos supuestos con las propiedades de los $\hat{b}$?

Tabla 2: Matriz de correlaciones

	CI	Castellano	Matemática	Biología	Ingles
CI	1.00	0.67	0.82	0.75	0.83
Castellano	0.67	1.00	0.45	0.14	0.47
Matemática	0.82	0.45	1.00	0.76	0.65
Biología	0.75	0.14	0.76	1.00	0.61
Ingles	0..83	0.47	0.65	0.61	1.00

Tabla 3: Modelo $\boldsymbol{y} = b_o + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + b_3\boldsymbol{x_3} + b_4\boldsymbol{x_4}$

Variable	Coeficiente	Des. Típica Coeficiente	T-student	P(|X|>T)
Constante	-6130.72	584.06	-10.50	0.000
Castellano	3.54	0.54	6.51	0.000
Matemática	1.33	1.05	1.28	0.214
Biología	7.58	1.62	4.67	0.000
Ingles	7.43	1.70	4-36	0.000
Coeficiente de correlación múltiple: 0.94 F-Fisher para las tres variables: 95.24 con P(X>F)=0.0001				

Tabla 4: Modelo $\boldsymbol{y} = b_o + b_1\boldsymbol{x_1} + b_3\boldsymbol{x_3} + b_4\boldsymbol{x_4}$

Variable	Coeficiente	Des. Tipica Coeficiente	T-student	P(|X|>T)
Constante	-6034.00	586.03	-10.30	0.000
Castellano	3.84	0.50	7.74	0.000
Biología	8.98	1.20	7.47	0.000
Ingles	7.74	1.70	4.54	0.000
Coeficiente de correlación múltiple: 0.93 F-Fisher para las tres variables: 123.47 con P(X>F)=0.0000				

Problema 1

a) Gráfico de dispersión: con x en abscisa e y en ordenada, el gráfico mostrará si los pacientes están en su mayoría arriba de la primera bisectriz.

b) $\boldsymbol{d} \in \boldsymbol{N}(\mu_y - \mu_x, s_{\boldsymbol{d}}^{\boldsymbol{2}})$ con $s_d^2 = s_x^2 + s_y^2 - 2cov(x,y)$

c) $\hat{m}_x = \boldsymbol{8.01}$; $\hat{m}_y = \boldsymbol{8.91}$; $\hat{m}_d = 0.90$;
$s_d^2 = 0.0385 + 0.0464 - 2 * 0.0381 = 0.0087$

d) $\bar{\boldsymbol{d}} \in \boldsymbol{N}(\mu_{\boldsymbol{d}}, \frac{s_{\boldsymbol{d}}^{\boldsymbol{2}}}{\boldsymbol{n}})$; $d \in [\bar{d} - 1.96 * \frac{s_d}{\sqrt{n}}, \bar{d} + 1.96 * \frac{s_d}{\sqrt{n}}] = [0.8918, 0.9082]$. El 1.96 por el tamaño de la muestra, se aproxima la Student a la Normal(0,1); además $\frac{s_d}{\sqrt{500}} = 0.00417$. El intervalo no contiene el 0 y de lejos, lo que permite decir que el tratamiento es **efectivo**.

e) La hipótesis nula es el tratamiento no es efectivo y la alternativa que es efectivo. La región crítica del test es de la forma $\bar{\boldsymbol{d}} > \boldsymbol{c}$:
$Prob(N(0,1) > .164) = 0.05 \Rightarrow \frac{\bar{d}}{\hat{s}_d / \sqrt{500}} > 1.64 \Rightarrow \bar{d} > 1.64 * \hat{s}_d / \sqrt{500} = 0.0068$
El valor encontrado en la muestra es de 0.90 => Se confirma que el tratamiento es **efectivo.**

Problema 2

a) La nota en matemática no resulta significativa en el modelo. Sin embargo esta variable es la más correlacionada con el CI. Esto se debe a que esta muy relacionadas con las otras variables también.
b) [3.54-1.95*0.54, 3.54+1.95*0.54]=[2.49 , 4.59] .
c) $\boldsymbol{Prob(t_{25} < -1.71) = 0.05}$ Þ **(3.54 - 4)/0.54 = - 0.8519 > -1.71** Þ **No se rechaza H_o**

d) La calidad global del modelo sin la nota de matemática no cambia casi nada y todas las variables son significativas.
e) Se supone que los errores son normales de media nula, de misma varianza e independientes entre si. Eso implica que los $\hat{b}_j$ son insesgados y de distribución normal, lo que permite calcular los intervalos de confianza y efectuar test de hipótesis.