PREGUNTA 1

Sea el modelo lineal:

$$y_i = \beta_o + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad (i = 1,2,...,n) \qquad (1)$$

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

$$y_i = \gamma_o + \sum_{j=1}^{p} \gamma_j z_{ij} + \varepsilon_i \quad (i = 1,2,...,n) \qquad (2)$$

1.1 Muestre que los modelos (1) y (2) producen las mismas predicciones y dé el estimador de γ_o.
1.2 Se supone que $\varepsilon \sim N_n(0,\sigma^2 I_n)$. 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 $\mu_o = E(y_o)$.
1.4 Deduzca para que valor de $x_o^t = (x_{o1}, x_{o2}, ..., x_{op})$ el intervalo de confianza para μ_o tiene el largo más pequeño. Dé el largo del intervalo en este caso.
1.5 En el modelo (2): $Y = Z\gamma + \varepsilon$, se supone ahora que las columnas de *Z* son ortogonales ($Z^t Z$ es diagonal). Dé la expresión individual de los coeficientes $\hat{\gamma}_j$ y muestre que son no correlacionados entre si.

PAUTA

1.1 Se tienen las predicciones para el modelo (1)

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

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

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

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

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

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

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

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

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

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

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

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

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

1.5 $\hat{\gamma} = D^{-1}Z^t y \Rightarrow$ $\boxed{\hat{\gamma}_j = \frac{Cov(x_j, y)}{var(x_j)}}$ y son no correlacionados dado que $\boxed{Var(\hat{\gamma}) = \sigma^2 D^{-1}}$.

PROBLEMA 2

Consideramos las siguientes variables obtenidas en diferentes parcelas de 4ha:

- x_1: la primera medición de volumen de madera;
- *y*: la segunda medición de volumen de madera;
- x_2: el número de pinos;
- x_3: la edad promedio de los pinos;
- x_4: el volumen promedio por pino ($x_4 = x_1 / x_2$).

1.1 Complete los resultados de la regresión de *y* sobre las 4 otras variables:

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	23.45	14.90	?	0.122
X_1	0.9321	0.08602	?	0.000
X_2	?	0.4721	1.5554	?
X_3	-0.4982	0.1520	?	0.002
X_4	3.486	?	1.533	0.132

1.2 Complete la tabla de análisis de la varianza:

Fuente de variablilidad	g.l.: Grados de libertad	SC: Suma de cuadrados	CM: SC/g.l.	F	*p-valor*
Regresión Error Total (centrado)	4 ? 54	SCR=887994 SCE=? SCtotal=902773	222000 29558	?	0.000

1.3 ¿Cuántas parcelas tenemos?

1.4 Dé el coeficiente de determinación R^2.

1.5 Haga el test de la hipótesis nula: $H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

1.6 Dé un intervalo de confianza a 95% para β_2.

1.7 Hacer el test de hipótesis $H_o : \beta_3 = 0$ contra $H_1 : \beta_3 \neq 0$ al nivel de significación $\alpha = 0.05$

1.8 Hacer el test de hipótesis $H_o : \beta_1 = 1$ contra $H_1 : \beta_1 \neq 1$ al nivel de significación $\alpha = 0.05$

1.9 Sea $r_{y,x_4|(x_1,x_2,x_3)}$ el coeficiente de correlación parcial entre x_4 e y condicionado por las otras variables x_1,, x_2 y x_3. Muestre que

$$r^2_{y,x_4|(x_1,x_2,x_3)} = \frac{SCE(y \mid x_1,x_2,x_3) - SCE(y \mid x_1,x_2,x_3,x_4)}{SCE(y \mid x_1,x_2,x_3)}$$ en donde

$SCE(y \mid x_1,x_2,x_3)$ es la suma de los residuos de la regresión de y sobre las tres variables x_1,, x_2 y x_3 y $SCE(y \mid x_1,x_2,x_3,x_4)$ es la suma de los residuos de la regresión de y sobre las cuatro variables x_1,, x_2,x_3 y x_4.

1.10 Deduzca que el estadístico F de Fisher para el test $H_o : \beta_4 = 0$ contra $H_1 : \beta_4 \neq 0$ es una función de $r^2_{y,x_4|(x_1,x_2,x_3)}$ y concluye que el estadístico F y $r^2_{y,x_4|(x_1,x_2,x_3)}$ contienen la misma información.

1.11 Se consideran los modelos lineales secuenciales:

$E(y) = \beta_o + \beta_1 x_1$ (1)

$E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2$ (2)

$E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ (3)

$E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (4)

y para cada modelo se define la cantidad:

(1): $SCR(x_1) = SCE(y) - SCE(y \mid x_1) =$ ***883880***

(2): $SCR(x_2 \mid x_1) = SCE(y \mid x_1) - SCE(y \mid x_1,x_2) =$***183***

(3): $SCR(x_3 \mid x_1,x_2) = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3) =$***3237***

(4): $SCR(x_4 \mid x_1,x_2,x_3) = SCE(y \mid x_1,x_2,x_3) - SCE(y \mid x_1,x_2,x_3,x_4) =$***694***

Sea $SCR(x_3,x_4 \mid x_1,x_2) = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3,x_4)$

Deduzca que $SCR(x_3,x_4 \mid x_1,x_2) = SCR(x_3 \mid x_1,x_2) + SCR(x_4 \mid x_1,x_2,x_3)$

1.12 Deduzca un test para el modelo (2) contra el modelo (4).

PAUTA

1.1

Variable	Estimación	Desv. típica Estimación	t-Student	*p-valor*
Constante	23.45	14.90	***1.5738***	0.122
X_1	0.9321	0.08602	***10.836***	0.000
X_2	***0.7343***	0.4721	1.5554	***0.126***
X_3	-0.4982	0.1520	***-3.2776***	0.002
X_4	3.486	***2.274***	1.533	0.132

1.2

Fuente de variablilidad	g.l.: Grados de libertad	SC: Suma de cuadrados	CM: SC/g.l.	F	*p-valor*
Regresión	4	SCR=887994	222000	***751.06***	0.000
Error	***50***	***14779***	295.58		
Total (centrado)	54	SCtotal=902773			

1.3 Hay 55 parcelas

1.4 $R^2 = \frac{SCR}{Var(y)} = \frac{SCR}{SCtotal} = \frac{883880}{902773} = 0.9836$

1.5 $F=751.06$ con 4 y 54 grados de libertad. El p-valor $P(F_{4,54} > 751.06 = 0.00$ es nulo.
Se rechaza la hipótesis nula. Hay una cierta significación del modelo.

1.6 El intervalo es: *[0.7343-2*0.4721,0.7343+2*0.4721]=[-0.21, 1.68]*

1.7 $H_o : \beta_3 = 0$ contra $H_1 : \beta_3 \neq 0$ al nivel de significación $\alpha = 0.05$:

La región crítica es de la forma: $\{ \hat{\beta}_3 < a \} \bigcup \{ \hat{\beta}_3 > a' \}$ con probabilidad 5% $\Rightarrow$

$\{ \frac{\hat{\beta}_3}{\hat{\sigma}_3} < -2 \} \bigcup \{ \frac{\hat{\beta}_3}{\hat{\sigma}_3} > 2 \} \Rightarrow \{ \hat{\beta}_3 < -0.30 \} \bigcup \{ \hat{\beta}_3 > 0.30 \}$. Como $\hat{\beta}_3$ pertenece a la región crítica, se rechaza H_o con un error de 5%.

1.8 $H_o : \beta_1 = 1$ contra $H_1 : \beta_1 \neq 1$ al nivel de significación $\alpha = 0.05$:

La región crítica es de la forma: $\{ \hat{\beta}_1 < a \} \bigcup \{ \hat{\beta}_1 > a' \}$ con probabilidad 5% $\Rightarrow$

$\{ \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_1} < -2 \} \bigcup \{ \frac{\hat{\beta}_1 - 1}{\hat{\sigma}_1} > 2 \} \Rightarrow \{ \hat{\beta}_1 < 0.828 \} \bigcup \{ \hat{\beta}_1 > 1.172 \}$. Como $\hat{\beta}_1$ no pertenece a la región crítica, no se rechaza H_o.

1.9 $r^2_{y,x_4|(x_1,x_2,x_3)} = Cor(\hat{\varepsilon}_1, \hat{\eta})$ donde $\hat{\varepsilon}_1$ es el vector de residuos de la regresión de *y* sobre x_1, x_2, x_3 y $\hat{\eta}$ es el vector de residuos de la regresión de x_4 sobre x_1, x_2, x_3.

$$r^2_{y,x_4|(x_1,x_2,x_3)} = \frac{|| P_{\hat{\eta}}(\hat{\varepsilon}_1) ||^2}{|| \hat{\varepsilon}_1 ||^2}$$

Sea $\hat{\varepsilon}$ el vector de residuos de la regresión de *y* sobre x_1, x_2, x_3, x_4. Entonces $\boxed{\hat{\varepsilon} \perp \hat{\eta}}$

Luego si $S_3 = < x_1, x_2, x_3 >$ y $S_4 = < x_1, x_2, x_3, x_4 >$, se puede escribir $S_4^{\perp} = S_3^{\perp} \oplus \Delta$ donde $\Delta = < \hat{\eta} >$. entonces $\hat{\varepsilon}_1 = \hat{\varepsilon} + P_{\hat{\eta}}(\hat{\varepsilon}_1)$ y $|| \hat{\varepsilon}_1 ||^2 = || \hat{\varepsilon} ||^2 + || P_{\hat{\eta}}(\hat{\varepsilon}_1) ||^2$. Lo que demuestre el resultado.

1.10 El *F* de Fisher para el test $H_o : \beta_4 = 0$ contra $H_1 : \beta_4 \neq 0$ es igual a:

$$\boxed{F = (n-5)\frac{r^2_{y,x_4|(x_1,x_2,x_3)}}{1-r^2_{y,x_4|(x_1,x_2,x_3)}}}$$

1.11 Como $SCR(x_3,x_4 \mid x_1,x_2) = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3,x_4)$ y $SCR(x_4 \mid x_1,x_2,x_3) = SCE(y \mid x_1,x_2,x_3) - SCE(y \mid x_1,x_2,x_3,x_4)$, se obtiene: $SCR(x_3,x_4 \mid x_1,x_2) = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3) + SCR(x_4 \mid x_1,x_2,x_3)$ y como $SCR(x_3 \mid x_1,x_2) = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3)$, se obtiene:

$$\boxed{SCR(x_3,x_4 \mid x_1,x_2) = SCR(x_3 \mid x_1,x_2) + SCR(x_4 \mid x_1,x_2,x_3)}$$

1.12 El test F para el modelo (2): H_0: $E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2$ contra el modelo (4): H_1: $E(y) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ es

$$F = \frac{\sum(\hat{y}_i - \hat{y}_{io})^2 / 2}{\sum \hat{\varepsilon}_i^2 / 50}$$

$$\sum(\hat{y}_i - \hat{y}_{io})^2 = \sum \hat{\varepsilon}_{io}^2 - \sum \hat{\varepsilon}_i^2 = SCE(y \mid x_1,x_2) - SCE(y \mid x_1,x_2,x_3,x_4) = SCR(x_3,x_4 \mid x_1,x_2)$$

Luego, $F = \dfrac{SCR(x_3,x_4 \mid x_1,x_2)/2}{SCE(y \mid x_1,x_2,x_3,x_4)/50} = \dfrac{(3237+694)/2}{295.58} = 6.65$

El p-valor es $P(F_{2,50} > 6.65) = 0.003$; se rechaza H_o.

PROBLEMA 3

1.1 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			

1.2 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

Tabla 4

	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

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

1.3 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 $H_o : \beta_1 = \beta_2 = \beta_5 = 0$

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	?			

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

PAUTA

1.1 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			

1.2 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.

1.3 Se justifica por lo anterior esta regresión. El test de hipótesis $H_o : \beta_1 = \beta_2 = \beta_5 = 0$ 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 anterior: no se rechaza $H_o : \beta_1 = \beta_2 = \beta_5 = 0$.

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**			

1.4 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 4

1.1 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.

1.2 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

1.3 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.

PAUTA

1.1 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			

1.2 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.

1.3 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 $H_o : \beta_1 = \beta_2 = 0$ 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 $H_o : \beta_1 = \beta_2 = 0$ es: $\dfrac{(2487328710 - 2356036159)/2}{23098393.7} = 2.8420$
- 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 5

El ministerio de educación quiere estudiar de qué depende el gasto anual en educación de un hogar, para ello, recolecta información en 100 hogares y plantea el modelo lineal:

$$E(y) = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \quad (1)$$

Donde x_1 es el ingreso del hogar (en miles de pesos), x_2 el número de hijos, x_3 la talla del jefe de hogar y x_4 el número de perros en la casa.

2.1 Complete los resultados de la regresión lineal (1) dados en las tablas n° 4 y 5.

2.2 Interprete los resultados.

2.3 Se plantea un modelo con el ingreso y el n° de hijos solamente:

$$E(y) = b_o + b_1 x_1 + b_2 x_2 \quad (2)$$

Se propone resolver el test: de hipótesis $H_o : E(y) = b_o + b_1 x_1 + b_2 x_2$ contra $H_1 : E(y) = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$.

Para esto, se resuelve el modelo (2) obteniéndose el conjunto de resultados presentados en las tablas 6 a 7.

Comente el cambio en la suma de los cuadrados de los residuos *SSR* del modelo (1) al modelo (2) y calcule la variación porcentual.

Se propone como estadístico para medir la significación del cambio en la suma residual a:

$$\frac{(SSR_2 - SSR_1)/(k_2 - k_1)}{SSR_1/(n - k_1)}$$

donde SSR_2 y SSR_1 corresponden a la suma de los cuadrados de los residuos de los modelos (2) y (1) respectivamente, y donde k_1 y k_2 son la cantidad de coeficientes de cada modelo. Encuentre la distribución que sigue este estadístico y concluye con un error de tipo I de 5% si las variables *nº de perros* y *talla del jefe* son significativas en el modelo (1) utilizando los resultados de las tablas 4 a 7.

2.4 Dé intervalos de confianza de nivel 95% para los tres parámetros del modelo (2).

2.5 Se tiene un nuevo hogar con un ingreso de 400 y 3 hijos. Dé una estimación de su gasto en educación.

Table n°4

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	20.387	20.384	1.000	0.319
Ingreso	0.189		9.242	0.000
N° hijos	17.379	2.978	5.836	0.000
Talla jefe	8.869	6.176		0.154
N° perros		0.107	1.749	0.083

Coeficiente de correlación múltiple R=0.785

Estimación insesgada de la varianza del error $\hat{\sigma} = 29.12$

Tabla n°5

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión		129489.083	38.185	0.0000
Residuos	95	80539.635		
Total	99			

Table n°6

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	54.03477	8.575475	6.301	0.000
Ingreso	0.197514	0.019715	10.019	0.000
N° hijos	17.804395	2.969696	5.995	0.000

Coeficiente de correlación múltiple R=0.772

Estimación insesgada de la varianza del error $\hat{\sigma} = 29.56$

Tabla n°7

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión	2	125292.851	71.713473	0.0000
Residuos	97	84735.8665		
Total	99	210028.718		

PAUTA

2.1

Table n°3

Variable	Estimación	esviación típica	t-Student	P-Valor
Constante	20.387	20.384	1.000	0.319
X_1	0.189	0.020	9.242	0.000
X_2	17.379	2.978	5.836	0.000
X_3	8.869	6.176	1.436	0.154
X_4	0.188	0.107	1.749	0.083

Tabla n°4

Fuente	Grados libertad	Suma cuadrados	Cuadrados medio	F	p-valor
Regresión	4	129489.083	32372.271	38.185	0.0000
Residuos	95	80539.635	847.786		
Total	99	210028.718			

2.2 El modelo (1) es globalmente significativo. Sin embargo se observa que las tercera y cuarta variables no parecen significativas.
2.3 El modelo (2) es tan bueno que el modelo (1). El coeficiente de correlación múltiple disminuye muy poco (0.66 -> 0.64) y un test F lo permite confirmar:

$$\frac{(\sum(\hat{\varepsilon}_i^{H_o})^2 - \sum(\hat{\varepsilon}_i^{H_1})^2)/2}{\sum(\hat{\varepsilon}_i^{H_1})^2/(100-5)}$$

Como $\sum(\hat{\varepsilon}_i^{H_o})^2 = 97 * \hat{\sigma}^2 = 97 * 29.56^2 = 84758$; en la tabla es 84735.8665 exactamente.

$\sum(\hat{\varepsilon}_i^{H_1})^2 = 95 * \hat{\sigma}^2 = 95 * 29.12^2 = 80558$ en la tabla es exactamente 80539.635

F=(84735-80539)*95/2*80539=2.4748

Luego el estadístico es el p-valor del test es: $Pr(F_{2,95} > 2.4748) \approx 0.09$

No se rechaza la hipótesis nula. El modelo (2) es tan significativo que el modelo (1).

2.4 De la tabla nº6 se deduce el intervalo: estimación ±1.96Desv. típica:

Variable	Estimación	Desviación típica	t-Student	Intervalo
Constante	54.03477	8.575475	6.301	[37.23, 70.84]
Ingreso	0.197514	0.019715	10.019	[0.159, 0.236]
N° hijos	17.804395	2.969696	5.995	[11.98, 23.62]

2.5 La estimación del gasto es: y^*=(1 400 3).(54.0348 0.19751 17.8044)'=186.45.

PROBLEMA 6

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

1.1 Construye la tabla ANOVA que permite decidir si se observan diferencias en el rendimiento de trigo entre los tres grupos. Interprete los resultados.
1.2 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.
1.3 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.

PAUTA

1.1 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:

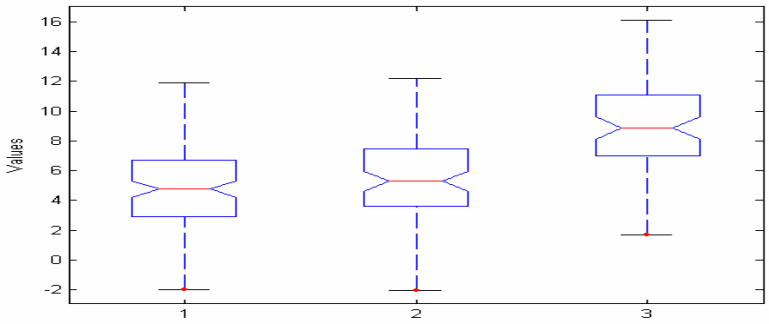
$$120 * (4.8450 - 6.1380)^2 + 80 * (5.3345 - 6.1380)^2 + 75 * (9.0639 - 6.1380)^2 = 894.346$$

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.

1.2



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

3.3 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%:

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