Estadística II - Taller 05 Semestre: 2024-01

Profesores: Johnatan Cardona Jimenez, Freddy Hernández Barajas, Raul Alberto Perez

Monitor: Ronald Palencia

Parte teorica

1) Sean Ψ y Γ dos subconjuntos de parámetros o coeficientes de un modelo de RLM tales que en la $\Psi \cup \Gamma$ se encuentran todos los parámetros del modelo y la $\Psi \cap \Gamma = \emptyset$. Con respecto a la SS_{extra} de Ψ se puede afirmar que:

A) $SSR(\Psi|\Gamma) = SSR(\Psi \cup \Gamma) - SSR(\Gamma)$

B) $SSR(\Gamma|\Psi) = SSE(\Gamma) - SSE(\Psi \cup \Gamma)$

2) En el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$, para probar la hipótesis:

$$\begin{aligned} &H_0 : \beta_2 = \beta_4 = 0 \\ &H_1 : \text{Algún } \beta_j \neq 0, \quad j = 2, 4. \end{aligned}$$

¿Cuál de las siguientes fórmulas describe correctamente la estadística de prueba F_0 bajo la hipótesis nula H_0?

A. $F_0 = \frac{\text{SSR}(\beta_2,\beta_4|\beta_0,\beta_1,\beta_3,\beta_5)/2}{\text{MSE}}$ bajo $H_0 \sim \tilde{f}_{2,n-6}$

B. $F_0 = \frac{\text{SSR}(\beta_0,\beta_1,\beta_3,\beta_5)}{\text{MSE}/2}$ bajo $H_0 \sim \tilde{f}_{4,n-6}$

C. $F_0 = \frac{\text{SSR}(\beta_2,\beta_4|\beta_0,\beta_1,\beta_3,\beta_5)}{\text{MSE}}$ bajo $H_0 \sim \tilde{f}_{4,n-6}$

D. $F_0 = \frac{\text{SSE}/2}{\text{MSE}}$ bajo $H_0 \sim \tilde{f}_{2,n-6}$

3) El estadistico de prueba para la prueba de la hipótesis lineal general es:

A) $F_0 = \frac{SSH/r}{MSE}$

B) $F_0 = \frac{SSE(FM) - SSE(RM)/r}{MSE}$

C) $F_0 = \frac{SSE(FM)-SSE(RM)}{MSE}$

D) $F_0 = \frac{SSH}{MSE}$

4) ¿Cuál es el juego de hipótesis correcto para la prueba de significancia de la regresión en el análisis de varianza dentro de un modelo de regresión lineal múltiple (RLM)?

A) $H_0 : \beta_1 \neq \beta_2 \neq ... \neq \beta_K \neq 0$ vs. $H_1 : \beta_j = 0$ para algún j

B) $H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$ vs. $H_1 : \beta_j \neq 0$ para todos j

C) $H_0 : \beta_1 = \beta_2 = ... = \beta_K = 0$ vs. $H_1 : \beta_j \neq 0$ para algún $j = 1, \dots, K$

D) $H_0 : \beta_1 = \beta_2 = ... = \beta_K \neq 0$ vs. $H_1 : \beta_j = 0$ para al menos un j

5) En un modelo de Regresión Lineal Múltiple (RLM) que utiliza SSextra para la prueba de significancia de un coeficiente individual, se plantea un juego de hipótesis para el coeficiente β_j, con $j = 1, \dots, k$. Si el estadístico de prueba es $F_{j,0} = \frac{\text{SSR}(\beta_j|\beta_0,\beta_1,\dots,\beta_{j-1},\beta_{j+1},\dots,\beta_k)}{\text{MSE}}$, donde SSR representa la Suma de Cuadrados de la Regresión al añadir el coeficiente β_j y MSE es el Error Cuadrático Medio del modelo completo, ¿cuál es la región de rechazo correcta para la hipótesis nula $H_0 : \beta_j = 0$ al nivel de significancia α?

A) $F_{j,0} > f_{\alpha,1,n-k-1}$, donde $f_{\alpha,1,n-k-1}$ es el valor crítico de la distribución F con 1 y $n-k-1$ grados de libertad.

B) $F_{j,0} < f_{\alpha,1,n-k-1}$, donde $f_{\alpha,1,n-k-1}$ es el valor crítico de la distribución F con 1 y $n-k-1$ grados de libertad.

C) $F_{j,0} > f_{\alpha,2,n-k-2}$, donde $f_{\alpha,2,n-k-2}$ es el valor crítico de la distribución F con 2 y $n-k-2$ grados de libertad.

D)$F_{j,0} > t^2_{\alpha/2,n-k-1}$, donde $t_{\alpha/2,n-k-1}$ es el valor crítico de la distribución t cuadrada con $n-k-1$ grados de libertad.

Parte practica

Se tiene una base de datos con información de algunos jugadores de la NBA, se presentan algunos datos a continuación en la siguiente tabla.

Descripción de cada una de las variables:

Y : promedio de puntos anotados por juego.

X1: altura en pies.

X2: masa en libras.

X3: porcentaje de tiros de campo acertados.

X4: porcentaje de tiros libres acertados.

Table 1: Algunos datos de jugadores de la NBA

Y	X1	X2	X3	X4
9.2	6.8	225	0.442	0.672
11.7	6.3	180	0.435	0.797
15.8	6.4	190	0.456	0.761
8.6	6.2	180	0.416	0.651
23.2	6.9	205	0.449	0.900
27.4	6.4	225	0.431	0.780

a) El resultados despues de ajustar un modelo usando todas las covariables es el siguiente:

```
Call:
lm(formula = Y ~ ., data = datos)

Residuals:
   Min     1Q Median     3Q    Max
-8.966 -3.545 -1.187  2.613 15.211

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.148707  14.855006   0.279  0.78121
X1          -3.690499   2.970780  -1.242  0.22005
X2           0.009458   0.046297   0.204  0.83897
X3          47.940199  15.709131   3.052  0.00367 **
X4          11.371019   7.868536   1.445  0.15479
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.411 on 49 degrees of freedom
Multiple R-squared:  0.2223,    Adjusted R-squared:  0.1588
F-statistic: 3.501 on 4 and 49 DF,  p-value: 0.01364
```

Usando el resutado anterior:

- Escribir la ecuacuión del modelo general en formato largo y en formato matricial, ademas la ecuación del modelo ajustado junto con sus supuestos.

- Haga un análisis de los coeficientes de manera marginal, es decir, analice estos de manera individual, ¿cuáles de estos son significativos?
- Construya un intervalo de confianza para el coefiente significativo.
- Basado en los resultados de un modelo de regresión que predice el promedio de puntos anotados por juego a partir de varias características de jugadores de baloncesto, ¿cuál de las siguientes afirmaciones se puede interpretar correctamente de los coeficientes estimados?

  A. Por cada incremento de un pie en la altura, se espera, en promedio, que el número de puntos anotados por juego disminuya en aproximadamente 3.69 puntos, manteniendo las demás variables constantes.

  B. Por cada incremento de una libra en la masa, se espera, en promedio, un aumento de 9.458 puntos anotados por juego, manteniendo las demás variables constantes.

  C. Un aumento de un 1% en el porcentaje de tiros de campo acertados está asociado con un incremento de aproximadamente 47.94 puntos anotados por juego, manteniendo las demás variables constantes.

  D. El coeficiente de porcentaje de tiros libres acertados es estadísticamente significativo al nivel del 5%.

- ¿Cuál es la interpretación correcta del coeficiente de determinación (R^2) en este modelo de regresión lineal múltiple que predice el promedio de puntos anotados por juego a partir de la altura, la masa y porcentajes de tiros acertados de jugadores de baloncesto?

  A. El 22.23% de la variabilidad en el promedio de puntos anotados (Y) por juego puede ser explicado por la altura, la masa, los porcentajes de tiros de campo acertados y porcentaje de tiros libres acertados..

  B. El 15.88% de la variabilidad en la altura, la masa y los porcentajes de tiros acertados es explicada por el promedio de puntos anotados por juego.

  C. El modelo de regresión no es significativo ya que el coeficiente (R^2) es menor que 0.5.

  D. Hay una correlación del 22.23% entre el promedio de puntos anotados por juego y las variables predictoras incluidas en el modelo.

b) Se tiene como resultado del analisis de varianza la siguiente tabla

```
      Sum_of_Squares DF Mean_Square F_Value   P_value
Model        409.934  4    102.4834 3.50058 0.0136397
Error       1434.532 49     29.2762
```

- Haga la prueba de la significancia de la regresión, usando el valor p y la región de rechazo.

d) La siguiente tabla recibe el nombre de tabla de todas las regresiones posibles. La cual sera de ayuda para validar algunos juegos de hipótesis.

	k	R_sq	adj_R_sq	SSE	Cp	Variables_in_model
1	1	0.115	0.098	1632.80	5.772	X3
2	1	0.060	0.042	1733.89	9.225	X4
3	1	0.005	-0.014	1835.71	12.703	X1
4	1	0.000	-0.019	1844.29	12.996	X2
5	2	0.189	0.157	1495.73	3.090	X1 X3
6	2	0.178	0.146	1516.42	3.797	X3 X4
7	2	0.161	0.128	1547.07	4.844	X2 X3
8	2	0.064	0.027	1726.34	10.967	X2 X4
9	2	0.060	0.023	1733.83	11.223	X1 X4
10	2	0.012	-0.027	1821.93	14.233	X1 X2
11	3	0.222	0.175	1435.75	3.042	X1 X3 X4
12	3	0.198	0.150	1479.71	4.543	X2 X3 X4
13	3	0.189	0.140	1495.67	5.088	X1 X2 X3
14	3	0.074	0.019	1707.18	12.313	X1 X2 X4
15	4	0.222	0.159	1434.53	5.000	X1 X2 X3 X4

- Usando Sumas de cuadrados Extra, realizar la siguiente prueba de hipóstesis usando la región de rechazo:

$$H_0 : \beta_1 = \beta_4 = 0 \text{ vs } H_1 : \text{al menos un } \beta_j \neq 0, j = 1, 4$$