Taller Práctico Regresión Lineal Múltiple (2) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al quinto taller práctico del curso de **Estadística II** para la *Universidad Nacional de Colombia*, Sede Medellín, en el periodo 2024 - 2. Se brinda una introducción al análisis de regresión. El enfoque de este taller está la comprensión del análisis de varianza y la ejecución de pruebas de hipótesis a través de la suma extra de cuadrados y el método lineal general. **Monitor:** *Santiago Carmona Hincapié.*

Keywords: regresión múltiple, pruebas hipótesis, ANOVA

Información general

Con el propósito de profundizar en los conceptos del modelo de regresión lineal múltiple vistos en clase, se propone afrontar este taller en dos partes, una de teoría básica y otra práctica.

La solución para cada uno de los problemas se efectúa a partir del software estadístico R.

Parte teórica

De respuesta a las preguntas formuladas a continuación en base a la teoría tratada en clase. **Provea una interpretación de ser necesario**.

1. Determine el valor de verdad de las siguientes afirmaciones.

   - (a) Una suma de cuadrados extra mide la reducción marginal en el SSE cuando una o varias variables predictoras son agregadas al modelo de regresión, dado que las otras predictoras ya fueron agregadas o están en el modelo.
   - (b) El estadístico T correspondiente al procedimiento de prueba empleado para probar la significancia marginal del parámetro j es:

     $$\mathrm{T}_0 = \frac{\hat{\beta}_j - 0}{\sqrt{\sigma^2 c_{jj}}} \sim t_{n-p}$$

     Con una región de rechazo asociada de $\mathrm{R}_c = \{|\mathrm{T}_0| > t_{\alpha/2,n-p}\}$ y p- valor $P(|t_{n-p}| > |T_0|)$.
   - (c) Valores grandes de R^2 implican que la superficie ajustada de respuesta es útil; sin embargo, es menos preferido que R^2_{adj} como medida de bondad de ajuste.

*El material asociado a este taller puede encontrarse en el repositorio del curso, **(https://github.com/Itsssach/Estadistica-II)**

- (d) El estadístico F correspondiente al procedimiento de prueba empleado para probar la significancia global del modelo de regresión lineal múltiple es:

$$\mathrm{F}_0 = \frac{\mathrm{SSR}/(k)}{\mathrm{SSE}/n-k} \sim f_{k,n-k}$$

Con una región de rechazo asociada de $\mathrm{R}_c = \{\mathrm{F}_{calc} > f_{\alpha,k,n-k}\}$ y p- valor $P(f_{k,n-k} > F_{calc})$.

- (e). Los grados de libertad del cuadrado medio debido a la hipótesis son iguales al rango de la matriz $\mathbf{L}$, asociada la prueba lineal general ($\mathrm{H}_0 : \mathbf{L}\beta = 0$ vs $\mathrm{H}_1 : \mathbf{L}\beta \neq 0$).

Ejercicio con datos reales

Considere el siguiente conjunto de datos que agrupa una serie de métricas enfocadas en evaluar el rendimiento en educación física de estudiantes en una institución. **Se incluyen únicamente las métricas cuantitativas,** cuya descripción puede encontrarse **en el siguiente enlace:** https://www.kaggle.com/datasets/ziya07/student-physical-education-performance

Table 1: Información en análisis

Performance	Strength	Skills	Speed
337.4335	60.22040	37.55007	54.67685
418.0430	45.23407	72.26675	53.60737
413.5214	32.09499	86.78852	44.96376
378.6311	63.44574	44.25407	53.85836
428.8907	49.40831	93.26630	43.28241

Considere a *'Overall Performance'* como la variable respuesta. *Las covariables en análisis se especifican en la tabla mostrada con anterioridad.* **Suponga que los supuestos del modelo se cumplen. De respuesta a los siguientes planteamientos:**

1. Determine cuál es el modelo empleado en esta situación, junto con sus supuestos, además, reporte la recta de regresión ajustada.

2. Determine la significancia de la regresión global. ¿Cree usted que puede realizarse esta prueba empleando otro método? De ser así, pruébelo.

3. Determine la significancia de los parámetros individuales β_j, junto con intervalo de confianza. Brinde una interpretación apropiada.

4. Determine si el efecto de la primera covariable es el mismo que el efecto de la tercera covariable; al mismo tiempo, verifique si el correspondiente efecto de la primera covariable es el mismo que el de la segunda covariable. Plantee una prueba de hipótesis para ello y realice el procedimiento adecuado. **Reporte el modelo completo y el modelo reducido.**

Tarea: Realice la prueba de significancia de los parámetros de forma marginal a través de sumas de cuadrados extra **(especificado en las notas de clase).**

Solución

Se brinda la solución para los problemas planteados. **Puede encontrar el código en solitario en el archivo 02Solucion.R.** Puede complementar estos resultados con las notas de la sesión.

Primer punto

```
# ----------------------------
#          PRIMER PUNTO
# ----------------------------
datos <- read.csv(file.choose())
# Ajustar el modelo de regresión
modelo <- lm(Performance ~ ., data = datos)
# Al ajustar Performance ~ ., el . indica que se desean
# considerar la variables restantes como regresoras
resumen_completo <- summary(modelo) # Resumen del modelo
```

Table 2: Resumen del modelo completo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.8022541	2.4566468	28.82069	0
Strength	0.8846324	0.0284840	31.05714	0
Skills	1.8736830	0.0207254	90.40506	0
Speed	3.0450232	0.0278864	109.19402	0

Puede visualizarse en la [Tabla 2] el resumen del modelo completo. **Puede emplearse la tabla para determinarse el modelo ajustado. Según los resultados, véase que se tiene:**

$$\hat{Y}_i = 70.8022 + 0.8846X_{i1} + 1.8736X_{i2} + 3.0450X_{i3}$$

Modelo de regresión ajustado

Consultar las notas de la sesión para profundizar este ajuste.

Segundo punto

```r
# ----------------------------
#          SEGUNDO PUNTO
# ----------------------------
# Determinar significancia regresión global
anova_completo <- anova(modelo) # Comparar ambas tablas
anova_completo2 <- model::anova_table_lm(modelo)
# Se sugiere comparar anova_completo con anova_completo2
# ¿Qué diferencias existen entre ambas líneas de código?
```

Table 3: Tabla ANOVA modelo completo

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
Regression	2132324.39	3	710774.7960	6855.288	0
Residuals	51426.62	496	103.6827	NA	NA
Total	2183751.01	499	NA	NA	NA

En la [Tabla 3] puede visualizarse la **Información ANOVA** del model de regresión lineal múltiple analizado. Es labor del estudiante comparar los códigos expuestos con anterioridad, esto es, *anova_completo* y *anova_completo2*. Aquí se muestra este último, que es el que interesa. **En las líneas de código que se muestran a continuación, se realiza el cálculo de la [Tabla 3] a mano.**

```r
# ----------------------------
# También se puede realizar con otros métodos
modelo_reducido <- lm(Performance ~ 1, data = datos)
# Revisar las características del modelo reducido
resumen_reducido <- summary(modelo_reducido) # Revisar ajuste modelo
anova_reducido <- anova(modelo_reducido) # ANOVA modelo reducido
# ¿Cómo determinar el estadístico de prueba en este caso?
# Se debe tomar en consideración la información de los
# modelos completo y reducido.
# ----------------------------
SSE_completo <- anova_completo$`Sum Sq`[4]
# SSE_completo <- 51426.62
SSE_reducido <- anova_reducido$`Sum Sq`[1]
# SSE_reducido <- 2183751
MSE <- anova_completo$`Mean Sq`[4] # MSE en general
# MSE <- 103.6827
# ----------------------------
# Determinar el número de parámetros por modelo:
parametros_completo <- length(modelo$coefficients)
# parametros_completo <- 4
parametros_reducido <- length(modelo_reducido$coefficients)
```

```r
# parametros_reducido <- 1
# ----------------------------
# Definición provista: F0 = [SSE(MR) - SSE(MF)/k]/MSE(MF)
# donde k = gl(SSE(MR) - gl(SSE(MF))) = (n- 1)- (n- 4) = 3
# Similarmente k = gl(SSR(MF) - gl(SSR(MR))) = 4 - 1 = 3
# AMBOS RESULTADOS SON LOS MISMOS. A continuación se muestra:
# ----------------------------
k <- parametros_completo- parametros_reducido
# k <- 3 # Dada la resta anterior
# ----------------------------
# # Calcular la suma parcial con el SSE
SSE_parcial <- SSE_reducido - SSE_completo
# SSE_parcial <- 2132324
F0 <- (SSE_parcial/k)/MSE # Estadístico prueba
# F0 <- 6855.288
# ----------------------------
# CALCULAR AHORA EL VALOR P:
p_value <- pf(F0, df1 = k, df2 = parametros_completo, lower.tail = FALSE)
# Se rechazará H0
# Se pueden verificar los resultados a mano con los resultados
# dados por anova_completo2.
```

Table 4: Tabla hecha a mano

MSE	F0	p_value
103.6827	6855.288	1e-07

Verificar los resultados obtenidos en la [Tabla 4] con los resultados obtenidos en la [Tabla 3]. **Todos son iguales, o bien, aproximados.** Si el estudiante quiere calcular los resultados a mano, lo puede hacer copiando y pegando los valores en una variable. Por ejemplo: *MSE <- 103.6827* en lugar de *MSE <- anova_completo$**Mean Sq**[4]*. Este último procedimiento corresponde a cómo se efectuaría el proceso usando una programación un poco más rigurosa.

Tercer punto

```r
# ----------------------------
#          TERCER PUNTO
# ----------------------------
intervalos <- confint(modelo, level = 0.95) # Intervalos confianza
```

Table 5: Intervalos confianza

	2.5 %	97.5 %
(Intercept)	65.9755369	75.6289713
Strength	0.8286681	0.9405966
Skills	1.8329626	1.9144035
Speed	2.9902333	3.0998132

La [Tabla 5] muestra los intervalos de confianza hallados para cada uno de los parámetros del modelo de regresión. **Intente brindar una interpretación apropiada.**

Cuarto punto

```
# ----------------------------
#          CUARTO PUNTO
# ----------------------------
# Ajustar para modelo reducido
X13 <- datos$Strength + datos$Speed # Nueva variable
modelo_reducido2 <- lm(Performance ~ X13, data = datos) # Ajustar
# ----------------------------
# Analizar de una forma más sencilla
comparar_modelos <- anova(modelo_reducido2, modelo)
```

Table 6: Resultados comparación modelos

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
498	1206643.41	NA	NA	NA	NA
496	51426.62	2	1155217	5570.924	0

Analizar los resultados especificados en la [Tabla 6], que corresponde a la validación de la siguiente prueba de hipótesis:

$$\begin{cases} \mathrm{H}_0 : \beta_1 - \beta_3 = 0, \beta_2 = 0 \\ \mathrm{H}_1 : \text{Algún } \beta_j \neq 0 \end{cases}$$

Complementar con las notas de la sesión para entender los resultados con mayor profundidad.