Taller Práctico Regresión Lineal Múltiple (2) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al segundo taller práctico del curso de **Estadística II** para la *Universidad Nacional de Colombia*, Sede Medellín, en el periodo 2025 - 1. 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 j- ésimo parámetro es:

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

     Con una región de rechazo asociada de $\mathrm{R}_c = \{|\mathrm{T}_0| > t_{\alpha/2,n-p}\}$ y p- valor $P(|t_{n-p}| > |T_{j,0}|)$.

   - (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) Rechazar la hipótesis nula $\mathrm{H}_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ permite determinar la significancia global de un modelo de regresión $\mathrm{Y}_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ a través de un procedimiento de sumas de cuadrados extra o procedimiento lineal general.

- (e) 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 $P(f_{k,n-k} > F_{calc})$.

- (f) Se puede demostrar que $\mathrm{F}_{j,0} = \mathrm{T}_{j,0}^2$ únicamente si se pretende determinar la significancia marginal de un coeficiente de regresión, lo que implica que $P(f_{1,n-p} > \mathrm{F}_{j,0}) \equiv P(|t_{n-p}| > |\mathrm{T}_{j,0}|)$.

- (g) Los grados de libertad del cuadrado medio debido a la hipótesis son iguales al rango de la matriz **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 a través del método de sumas de cuadrados extra. ¿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).**