Taller Práctico Regresión Lineal Múltiple (3) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al tercer 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 de los supuestos del modelo, puntos de balanceo, atípicos e influenciales, así como la respuesta media y predicción. **Monitor:** *Santiago Carmona Hincapié.*

Keywords: regresión múltiple, supuestos, influenciales, balanceo

---

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) Bajo los supuestos del modelo de regresión lineal múltiple, se cumple que **la respuesta media estimada** $\hat{Y}_0 \sim N(\mathrm{E}[Y|\underline{\mathbf{x}}_0], \sigma^2\underline{\mathbf{x}}_0(\mathbf{X}'\mathbf{X})^{-1}\underline{\mathbf{x}}_0)$; donde además, $\hat{Y}_0$ es un estimador insesgado para Y_0.

   - (b) El error de predicción $\hat{Y}_0 - Y_0$ tiene una varianza asociada dada por $\sigma^2\underline{\mathbf{x}}_0(\mathbf{X}'\mathbf{X})^{-1}\underline{\mathbf{x}}_0$, al igual que $\hat{Y}_0$, de aquí que si se sabe que $\underline{\mathbf{x}}_0 = [1, x_{01}, x_{02}, ..., x_{0k}]'$ es un punto en que no se comete extrapolación, entonces es correcto afirmar que $\underline{\mathbf{x}}_0(\mathbf{X}'\mathbf{X})^{-1}\underline{\mathbf{x}}_0' < 1$.

   - (c) El procedimiento analítico *Shapiro- Wilk*, cuyo juego de hipótesis está dado por $\mathrm{H}_0 : \varepsilon_i \sim N(0, \sigma^2)$ vs $\mathrm{H}_1 : \varepsilon_i \nsim N(0, \sigma^2), i = 1, ..., n$ permite determinar la normalidad de los residuales del modelo de regresión.

   - (d) Una observación atípica está separada del resto de las observaciones en su valor de respuesta Y aunque no afecta los resultados del ajuste. **Su evaluación se realiza a través del residual estandarizado** $|d_i| > 3$.

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

- (e). Se cumple que una observación i es de balanceo si está definida en el espacio de la respuesta Y y se cumple que $h_{ii} > 2p/n$, afectando estadísticas como el R^2 y los errores estándar de los coeficientes del modelo.

- (f). Se cumple que $D_i > 1, |\text{DFBETAS}_{j(i)}| > (2/\sqrt{n})$ y $|\text{DFFITS}_i| > (2\sqrt{p/n})$ simultáneamente para toda observación categorizada como influencial en un conjunto de datos.

2. Seleccione las expresiones adecuadas que se muestra a continuación, interprételas y corrija las expresiones incorrectas.

a. $\mathbf{d_i} = \frac{e_i}{\sqrt{\text{MSE}}}$ **b.** $\mathbf{r_i} = \frac{d_i}{\sqrt{1-h_{ii}}}$

c. $\mathbf{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\text{MSE}_{(i)} c_{jj}}}$ **d.** $\mathbf{DFFITS}_i = \frac{\hat{y}_i - y_{i(0)}}{\sqrt{c_{jj}\text{MSE}_{(i)}}}$

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.* **De respuesta a los siguientes planteamientos:**

1. Verifique los supuestos del modelo de regresión, esto es, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ a partir de los procedimientos apropiados para ello.

2. Identifique puntos atípicos y puntos de balanceo a través de un criterio gráfico y analítico. **¿Podrían ser estos puntos a su vez influenciales?**

3. Identifique puntos influenciales. **Compare los criterios empleados para ello.**

4. Realice inferencia para $\underline{\mathbf{x}}_{01} = [1, 45.03, 80.88, 60.33]'$ y $\underline{\mathbf{x}}_{02} = [1, 77.08, 100, 13.76]'$ con su respectivo intervalo de predicción. **Verifique primero si no se trata de un punto de extrapolación.**