Taller Práctico Regresión Lineal Múltiple (4) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al séptimo 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 problema de multicolinealidad (efectos y diagnósticos), así como el método de todas regresiones posibles para la selección de variables. **Monitor:** *Santiago Carmona Hincapié.*

Keywords: regresión múltiple, multicolinealidad, selección

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 el modelo de regresión lineal múltiple $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, la comparación de los efectos parciales de las variables debe realizarse a través del **escalamiento normal unitario de las covariables**.

   - (b) La multicolinealidad refiere a la dependencia lineal casi perfecta entre covariables, afectando la matriz $\mathbf{X'X}$. Esta puede destacartarse si la correlación entre un par de variables $X_i \neq X_j$ es pequeña.

   - (c) La multicolinealidad causa la inflación de las varianzas de los estimadores, además de estimadores $\hat{\beta}_j$ muy grandes en términos absolutos y valores de los coeficientes estimados con signo contrario a lo esperado.

   - (d) Una forma en la que se manifiesta la multicolinealidad grave es cuando el modelo de regresión ajustado es significativo (globalmente), pero los parámetros individuales no lo son.

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

- (e) Dado que el estadístico C_p es una medida del sesgo del modelo, se prefiere el estadístico más bajo, tal que $|C_p - p|$ es mínima, puesto que a mayor sesgo mayor C_p.

- (f) La suma de cuadrados de los errores de predicción $e_{(i)} = Y_i - \hat{Y}_{i(i)}$ mide qué tan bien los valores ajustados por un submodelo predicen las respuestas observadas. **Mejor se considerará el modelo entre mayor sea esta métrica**.

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. Escriba el modelo de regresión lineal múltiple, junto con sus supuestos. Deduzca a partir del modelo ajustado si podría haber problemas de multicolinealidad.

2. Realice un análisis de multicolinealidad a través del criterio del factor de inflación de varianza, número de condición, índice de condición y proporción de descomposición de varianza.

3. Use el método de todas las regresiones posibles para seleccionar los mejores submodelos en función de los criterios $R_p^2, R_{adj(p)}^2$ (o bien MSE_p) y C_p. **Posteriormente seleccione el mejor modelo.**

Observación: Es idóneo recordar que los criterios R_p^2, MSE_p permiten escoger modelos que ajusten bien a los datos, mientras que, por su parte, *C_p de Mallows'* permite escoger el mejor modelo para predecir.