Taller Práctico Regresión Lineal Múltiple (1) *

Estadística II *Universidad Nacional de Colombia, Sede Medellín*

Este documento corresponde al cuarto 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 modelo de regresión lineal múltiple a nivel matricial. **Monitor:** *Santiago Carmona Hincapié.*

Keywords: regresión múltiple

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. Considere el siguiente modelo de regresión lineal múltiple con k variables regresoras, $p = (k+1)$ parámetros asociados $Y_i = \beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+...+\beta_k X_{ik}+\varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

    - (a) Escriba el modelo de forma matricial junto con sus supuestos. **Especifique las dimensiones de cada componente**.
    - (b) Demuestre que el estimador $\hat{\boldsymbol{\beta}}$ que se obtiene a través del método de mínimos cuadrados es un estimador insesgado para $\boldsymbol{\beta}$. Analice $\hat{\boldsymbol{\beta}}$.

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

    - (a) Bajo los supuestos del modelo de regresión lineal múltiple, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, el estimador para los parámetros $\hat{\boldsymbol{\beta}}$ es el mismo que el de máxima verosimilitud, así como para la varianza $\hat{\sigma}^2$.
    - (b) Se requiere que la matriz $(\mathbf{X^t X})$ sea singular.
    - (c) La matriz de varianzas- covarianzas siempre es simétrica respecto a su diagonal principal, además, siempre tiene unos en su diagonal principal.
    - (d) La matriz $\mathbf{H}$ es simétrica e idempotente, al igual que $(\mathbf{I_n} - \mathbf{H})$.

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

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

function (..., list = character(), package = NULL, lib.loc = NULL, verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE) {
fileExt <- function(x) {
db <- grepl("\\.[^.]+\\.(gz|bz2|xz)$", x)

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. Determine cuál es la matriz de diseño ($\mathbf{X}$) para este problema en específico.
2. Calcule el vector de parámetros estimados $\hat{\boldsymbol{\beta}}$.
3. Calcule la estimación de la respuesta media $\hat{\boldsymbol{y}}$.
4. Calcule el vector de los errores estimados $\hat{\boldsymbol{\varepsilon}}$.

Tarea: Construya el valor de la estimación para la varianza $\hat{\sigma}^2$. (**Ayuda:** Emplee el vector anterior).

Solución

Se brinda la solución para los problemas planteados. **Puede encontrar el código en solitario en el archivo 01Solucion.R.** Puede complementar estos resultados con las notas de la sesión.

Primer punto

```
# ----------------------------
#          PRIMER PUNTO
# ----------------------------
datos <- read.csv(file.choose())
# La instrucción anterior abre una pestaña auxiliar
# O también se puede con una dirección absoluta:
```

```r
# datos <- read.csv('data/students_performance.csv')
X <- as.matrix(cbind(Intercept = 1, datos[, -1]))
# Definiendo la matriz de diseño (objeto matricial)
Y <- as.matrix(datos[, 1])
# Definiendo el vector Y (como objeto matricial)
```

Los resultados se pueden visualizar corriendo el código en R, ya que son demasiado extensos para mostrarlos aquí.

Segundo punto

```r
# -----------------------------
#          SEGUNDO PUNTO
# -----------------------------
betas <- solve((t(X) %*% X)) %*% t(X) %*% Y
# Según la definición (X'X)^{-1} X'Y
```

Table 2: Vector de betas

Intercept	70.6933175
Strength	-0.0509026
Skills	-0.0168088
Speed	0.0466670

Este es el resultado de los betas empleando el método de estimación algebraico. Verificar con la tabla de resultados final, en donde se ajusta el modelo de regresión.

Tercer punto

```r
# -----------------------------
#          TERCER PUNTO
# -----------------------------
matriz_H <- X %*% solve((t(X) %*% X)) %*% t(X)
Y_gorro <- matriz_H %*% Y
# O también se puede de la siguiente manera:
# y_gorro <- X %*% betas
```

Los resultados se pueden visualizar corriendo el código en R, ya que son demasiado extensos para mostrarlos aquí.

Cuarto punto

```r
# ----------------------------
#          CUARTO PUNTO
# ----------------------------
residuales <- Y - Y_gorro # Por definición
```

Los resultados se pueden visualizar corriendo el código en R, ya que son demasiado extensos para mostrarlos aquí.

Solución Tarea

```r
# ----------------------------
#         Solución tarea
# ----------------------------
n <- nrow(datos); p <- nrow(betas)
# Para hallar el número de datos 'n'
# Para hallar el número de parámetros 'p'
MSE <- (t(residuales) %*% residuales)/(n- p)
```

El valor obtenido corresponde a MSE = 99.7965106. Verificar con la tabla de resultados mostrada al final. **Recuerde que en la tabla final se muestra la raíz cuadrada del MSE, no el MSE.**

Verificación de resultados

```r
# VERIFICAR TODOS LOS RESULTADOS
modelo <- lm(Performance ~ ., data = datos)
# Al ajustar Performance ~ ., el . indica que se desean
# considerar la variables restantes como regresoras
summary <- summary(modelo) # Resumen del modelo
```

Table 3: Resumen del modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.6933175	3.9469262	17.9109802	0.0000000
Strength	-0.0509026	0.0457953	-1.1115241	0.2668813
Skills	-0.0168088	0.0309541	-0.5430245	0.5873568
Speed	0.0466670	0.0456462	1.0223632	0.3071074

Table 4: Raíz cuadrada del MSE

x
9.98982