CIMAT

**Centro de Investigación en Matemáticas**
Unidad Monterrey

# Análisis Multimodal

## Tarea 2

Gustavo Hernández Angeles

29 de octubre de 2025

# Índice

**Ejercicio 1:**

Generación de datos simulados y aplicación de los métodos de selección de subconjuntos.

a) Usa una función en `R` para generar una variable predictora `X` de longitud 100, así como un vector de ruido $\epsilon$ de longitud 100.

b) Genera un vector de respuesta `Y` de longitud 100, de acuerdo al modelo

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

c) Utiliza la función **`regsubsets()`** para realizar la selección de subconjuntos con el fin de elegir el mejor modelo que contenga los predictores $\mathbf{X}$, $\mathbf{X}^2$, ..., $\mathbf{X}^{10}$. ¿Qué modelo se selecciona como el mejor según el AIC, BIC y el $R^2$ ajustado? Muestra algunas gráficas que proporcionen evidencia de tu respuesta y reporta los coeficientes de mejor modelo obtenido.

d) Repite (c) usando la selección forward stepwise y backward stepwise. ¿Cómo se compara tu respuesta con los resultados obtenidos en (c)?

## 1.1 Inciso a)

Para generar la variable predictora `X` y el vector de ruido $\epsilon$, utilizamos las siguientes funciones en `R`:

```r
set.seed(123) # Para reproducibilidad
n <- 100
X <- rnorm(n, mean = 10, sd = 3) # Variable predictora
epsilon <- rnorm(n, mean = 0, sd = 1) # Vector de ruido
```

## 1.2 Inciso b)

Utilizando la variable predictora `X` y el vector de ruido $\epsilon$ generados en el inciso anterior, creamos el vector de respuesta `Y` según el modelo especificado. Asumimos los siguientes valores para los coeficientes: $\beta_0 = 5$, $\beta_1 = 2$, $\beta_2 = -0.5$, y $\beta_3 = 0.1$.

```r
beta_0 <- 5
beta_1 <- 2
beta_2 <- -0.5
beta_3 <- 0.1
Y <- beta_0 + beta_1 * X + beta_2 * X^2 + beta_3 * X^3 + epsilon
```

## 1.3 Inciso c)

Para realizar la selección de subconjuntos utilizando la función `regsubsets()` del paquete `leaps`, primero creamos un data frame que contenga la variable respuesta `Y` y los predictores `X`, `X`$^2$, ..., `X`$^{10}$.

```r
data <- data.frame(Y = Y)
for (i in 1:10) {
    data[[paste0("X", i)]] <- X^i
}
```

Luego, aplicamos la función `regsubsets()` para realizar la selección de subconjuntos y analizamos los resultados utilizando AIC, BIC y $R^2$ ajustado.

```r
regfit.full <- regsubsets(Y ~ ., data = data, nvmax = 10)
reg.summary <- summary(regfit.full)

# AIC, BIC y R^2 ajustado
aic_values <- reg.summary$bic + 2 * (1:10) # AIC approximation
cat(aic_values)
```

```
## -661.931 -850.7434 -849.675 -843.0792 -837.0971 -832.1326 -826.2726 -820.1073 -813.8585 -
```

```r
bic_values <- reg.summary$bic
cat(bic_values)
```

```
## -663.931 -854.7434 -855.675 -851.0792 -847.0971 -844.1326 -840.2726 -836.1073 -831.8585 -
```

```r
adjr2_values <- reg.summary$adjr2 #£
cat(adjr2_values)
```

```
## 0.998795 0.9998275 0.9998351 0.9998334 0.9998327 0.9998336 0.999833 0.999832 0.9998307 0.
```

**Ejercicio 2: S**

e ha visto que a medida que aumenta el número de características de un modelo, el error de entrenamiento disminuirá necesariamente, pero el error de prueba puede que no. Explorar esto con datos simulados.

a) Genera un conjunto de datos con $p = 20$ características, $n = 1000$ observaciones y un vector de respuesta cuantitativo generado de acuerdo con el modelo

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$$

donde $\beta$ tiene algunos elementos exactamente iguales a cero.

b) Divide los datos en un conjunto de entrenamiento con 100 observaciones y otro de prueba con 900 observaciones.

c) Realiza la selección del mejor *subconjunto* sobre el conjunto de entrenamiento y grafica el error de entrenamiento MSE asociado con el mejor modelo en cada tamaño.

d) Gráfica el error de prueba MSE asociado con el mejor modelo de cada tamaño.

## Referencias