# Modelo de regresión lineal múltiple

- Los modelos de regresión estudian la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas. La formulación matemática de estos modelo es la siguiente

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \varepsilon$$

donde **ε** es el error de observación debido a variables no controladas

- La expresión matricial de este modelo viene dada de la forma

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

## El modelo de regresión lineal múltiple es útil

- **Cuando la respuesta depende de varias variables explicativas cuantitativas**
- **La regresión múltiple es mejor que la simple porque se mejora la predicción de la variable respuesta**
- **Cuando la respuesta depende de más de una variable, la regresión simple las considera una a una y se pueden producir fácilmente sesgos en la estimación de los efectos que tienen cada una de ellas en la respuesta**
- **Las ideas de la regresión simple se extienden casi automáticamente a la regresión múltiple**

# Hipótesis del modelo

1. Normalidad: los errores tienen una distribución Normal ($e_i \sim N(0; \sigma^2)$)
2. Linealidad: Le esperanza de los errores es cero ($E(e_i)=0$)
3. Homocedasticidad: La varianza de los errores son iguales
4. Independencia: los errores son independientes

## Requisitos adicionales de la regresión múltiple

a) Hay más datos que parámetros desconocidos
b) n es igual o mayor que k+2
c) Ninguna de las variables explicativas es combinación lineal exacta de las restantes (colinealidad)

$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_k x_{ki} + e_i$

## Interpretación de los parámetros:

$\alpha_0$ **:**Representa el valor medio de la respuesta (y) cuando todas las variable explicativas (x) valen cero

$\alpha_i$ **:**Representa el incremento de la respuesta media (y) cuando la variable explicativa ($x_i$) aumenta en una unidad y el resto de las variables explicativas permanecen constantes

## Estimación de los parámetros

$$\hat{\alpha} = \left(X'X\right)^{-1} X'Y$$

$$e_i = y_i - \hat{y}_i = y_i - \left(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots \alpha_k x_{ik}\right)$$

$$\hat{\sigma}^2 = S_R^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2$$

## Contrastes de los coeficientes

Dado que se conoce la distribución del vector de estimación α, es posible realizar los contrastes siguientes, en los que se comprueba si la respuesta no depende linealmente de cada una de las variables regresoras $x_i$

$$H_0 : \alpha_i = 0 \ \text{(la respuesta no depende linealmente de } x_i\text{)}$$

$$H_1 : \alpha_i \neq 0 \ \text{(la respuesta sí depende linealmente de } x_i\text{)}$$

## Contraste de regresión

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

$$H_1 : \text{existe algún } \alpha_i \neq 0 \text{ para algún i}$$

## Tabla ANOVA

Razonando como en el modelo de regresión lineal simple y en base a las propiedades geométricas del modelo, podemos efectuar la siguiente descomposición

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{\text{Suma de cuadrados}\\ \text{Total (SCT)}\\ \text{g.l=n-1}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{\text{Suma de cuadrados}\\ \text{explicada por el}\\ \text{modelo (SCX)}\\ \text{g.l=k}}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\substack{\text{Suma de cuadrados}\\ \text{Residual (SCR)}\\ \text{g.l=n-(k+1)}}}$$

# Tabla ANOVA

| | Tabla ANOVA del modelo de regresión múltiple | | | |
|---|---|---|---|---|
| **Fuentes de variación** | **SUMA DE CUADRADOS** | **g.l.** | **MEDIAS DE CUADRADOS** | **$F_R$** |
| **Por el modelo** | $SCX = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ | **k** | $\hat{S}_X^2 = MCX = \frac{SCX}{k}$ | $F_R = \frac{MCX}{MCR} \approx F_{k,n-(k+1)}$ |
| **Residual** | $SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ | **n-(k+1)** | $\hat{S}_R^2 = MCR = \frac{SCR}{n-(k+1)}$ | |
| **Total** | $SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ | **n-1** | $\hat{S}_Y^2 = \frac{SCT}{n-1}$ | |

| Caso | Contraste conjunto | Contraste individual |
|---|---|---|
| 1 | Significativo | Todos significativos |
| 2 | Significativo | Alguno significativo |
| 3 | Significativo | Ninguna significativo |
| 4 | No Significativo | Todos significativos |
| 5 | No Significativo | Alguno significativo |
| 6 | No Significativo | Ninguna significativo |

**Caso 1:** Todas las variables explicativas influyen en la variable respuesta

**Caso 2:** Influyen algunas variables explicativas, otras no

**Caso 3:** Las variables explicativas son dependientes entre sí. Entonces influyen de forma conjunta. Multicolinealidad. Se soluciona eliminando algunas variables regresoras (Análisis de Componentes Principales)

**Caso 4:** Multicolinealidad. Las variables presentan una correlación fuerte y negativa

**Caso 5:** Igual que el caso anterior

**Caso 6:** Ninguna de las variables regresoras influye en la variable de respuesta o la influencia no la detecta la muestra tomada

# Coeficiente de determinación – $R^2$

¿Cómo evaluamos la fuerza del ajuste de un modelo de regresión?

El **COEFICIENTE DE DETERMINACIÓN** es la proporción de variabilidad explicada por la regresión

$$R^2 = \frac{SCX}{SCT} = 1 - \frac{SCR}{SCT}$$

Inconveniente de $R^2$:

Siempre aumenta cuando introducimos nuevas variables, aunque no sirvan para explicar la respuesta

**Coeficiente de determinación CORREGIDO $R^2$**

Se corrige $R^2$ por los grados de libertad

$$\overline{R}^2 = 1 - \frac{MCR}{MCT} = 1 - \frac{\frac{SCR}{n-(k+1)}}{\frac{SCT}{n-1}}$$

Lo usaremos únicamente para comparar modelos con distinto número de variables

(siempre es más pequeño que $R^2$ y puede ser negativo)

# Multicolinealidad

•Cuando las variables explicativas están muy correlacionadas

•El caso extremo es cuando una variable es combinación lineal exacta de otras

•Intuitivamente, el problema que se presenta es que cada variable que incluimos en el modelo supone un parámetro nuevo a estimar y necesitamos más información. Si los datos no aportan casi nada nuevo es difícil estimar los parámetros

Presenta algunos inconvenientes que pueden ser importantes:

– Gran varianza de los estimadores α

– Cambio importante en las estimaciones al eliminar o incluir regresores en el modelo

– Cambio de los contrastes al eliminar o incluir regresores en el modelo

– Contradicciones entre el contraste F y los contrastes individuales