Introducción

Los Modelos de Regresión estudian la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas.

Cuando se estudia la relación entre una variable de interés, variable respuesta o variable dependiente, Y, y un conjunto de variables regresoras (explicativas, independientes), $X_1, X_2, \ldots, X_k$, puede darse las siguientes situaciones:

- Existe una **relación funcional** entre ellas, en el sentido de que el conocimiento de las variables regresoras determina completamente el valor que toma la variable respuesta, es decir

$$Y = f(X_1, X_2, \ldots, X_k)$$

  Ejemplo: la relación que existe entre el tiempo (Y) que tarda un móvil en recorrer una distancia y dicha distancia (X) a velocidad constante

- **No exista ninguna relación entre la variable respuesta y las variables regresoras**, en el sentido de que el conocimiento de éstas no proporciona ninguna información sobre el comportamiento de la otra.

  Ejemplo: la relación que existe entre la altura de una persona (Y) y la lluvia recogida por cm3 en un día(X).

- El caso intermedio, **existe una relación estocástica entre la variable respuesta y las variables regresoras**, en el sentido de que el conocimiento de éstas permiten predecir con mayor o menor exactitud el valor de la variable respuesta. Por tanto siguen un modelo de la forma, siendo f la función de regresión desconocida y ε una variable aleatoria de media cero (el error de observación)

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon$$

Las relaciones estocásticas son las que ocurren en la mayoría de las situaciones y su estudio corresponde a los Modelos de Regresión

El objetivo básico en el estudio de un modelo de regresión es el de estimar la función de regresión, f, y el modelo probabilístico que sigue el error aleatorio, es decir: estimar la función de distribución F de la variable de error.

Una vez estimadas estas funciones se tiene conocimiento de:

- La relación funcional de la variable respuesta con las variables regresoras, dada por la función de regresión que se define como sigue,

$$f(x_1, x_2, \ldots, x_k) = E\left[Y \middle| X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right]$$

  Esto permite tener una idea general del comportamiento de la variable respuesta en función de las regresoras.

- Se puede estimar y predecir el valor de la variable respuesta de un individuo del que se conocen los valores de las variables regresoras. Ésto es, de un individuo t se sabe que $X_1 = x_{1,t}, \ldots, X_k = x_{k,t}$, entonces se puede predecir el valor de Y_t y calcular un intervalo de predicción del mismo.

El modelo de regresión lineal simple

El modelo de regresión más sencillo es el Modelo de Regresión Lineal Simple que estudia la relación lineal entre la variable respuesta, Y, y la variable regresora, X, a partir de una muestra $\{(x_i, Y_i)\}_{i=1}^{n}$, que sigue el siguiente modelo:

$$Y = \alpha + \beta x_i + \varepsilon_i \qquad i = 1,\ 2,\ldots,\ n$$

Se supone que se verifican las siguientes hipótesis:

1. **La función de regresión es lineal**,

$$f(x_i) = E\left[Y \middle| X = x_i\right] = \alpha + \beta x_i \qquad i = 1,\ 2,\ldots,\ n$$

   o, equivalentemente, la media de los errores es cero $E(\varepsilon_i)=0$, i = 1,...,n.

2. **La varianza es constante (homocedasticidad)**,

$$V\left(Y \middle| X = x_i\right) = \sigma^2, \qquad i = 1,2,\ldots,n$$

   o, equivalentemente, $V(\varepsilon_i) = \sigma^2$, i = 1,...,n.

3. La distribución es normal,

$$Y \left| X = x_i \approx N\left(\alpha + \beta x_i, \sigma^2\right) \right. \qquad i = 1,2,\ldots,n$$

o, equivalentemente, $\varepsilon_i \cong N(0, \sigma^2)$, i = 1,...,n.

4. Las observaciones Y_i son independientes. Bajo las hipótesis de normalidad, esto equivale a que la Cov(Y_i,Y_j)=0, si i≠j. Esta hipótesis en función de los errores sería "los i son independientes", que bajo normalidad, equivale a que Cov(ε_i , ε_j) =0, si i≠j.

Estimación de los parámetros del modelo: recta de regresión

En el modelo de regresión lineal simple hay tres parámetros que se deben estimar: los coeficientes de la recta de regresión, α y β; y la varianza de la distribución normal, σ^2.

El cálculo de estimadores para estos parámetros puede hacerse por diferentes métodos, siendo los más utilizados el método de máxima verosimilitud y el método de mínimos cuadrados. Ambos métodos dan los mismos estimadores para estos parámetros.

Conocida una muestra de tamaño n, $\{(x_i , y_i)\}$, el método de mínimos cuadrados, consiste en minimizar la suma de los cuadrados de las distancias de los puntos de la nube a esa recta, medidas en la dirección del eje OY. Los estimadores que se obtienen son

$$\hat{\alpha} = \overline{y} - \frac{S_{XY}}{S_{XX}} \cdot \overline{x} \qquad \hat{\beta} = \frac{S_{XY}}{S_{XX}}$$

donde

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \qquad S_{XX} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - \left(n \cdot \overline{x}^2\right) \qquad S_{YY} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} y_i^2 - \left(n \cdot \overline{y}^2\right)$$

$$S_{XY} = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \sum_{i=1}^{n} x_i \cdot y_i - (n \cdot \overline{x} \cdot \overline{y})$$

La recta viene dada de la forma

$$\hat{y} = \overline{y} - \frac{S_{XY}}{S_{XX}} \cdot \overline{x} + \frac{S_{XY}}{S_{XX}} \cdot x \quad \text{o bien} \quad \hat{y} - \overline{y} = \frac{S_{XY}}{S_{XX}} \cdot (x - \overline{x})$$

Contrastes sobre los parámetros del modelo

Hipótesis Nula: No existe regresión lineal porque la media de Y es igual a α

$H_0 : \beta = 0$

Esto implica que la recta de regresión es $y_i = \alpha + \varepsilon_i$ y por lo tanto no existe relación lineal entre las variables X e Y

Hipótesis Alternativa: Sí existe regresión lineal porque la media de Y depende de los valores de X

$H_1 : \beta$

Este contraste se realiza mediante una prueba estadística llamada **ANOVA DE REGRESIÓN**, que consiste en comparar la variabilidad explicada por la recta de regresión lineal, que expresa la dependencia de Y respecto de la variable X, con la variabilidad residual, debida a la variación aleatoria de los individuos.

El contraste de ANOVA se basa en que es posible descomponer la variabilidad de la variable respuesta como suma de la variabilidad explicada por la recta de regresión y la variabilidad residual o no explicada por el modelo ajustado

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}_{\substack{\text{Suma de Cuadrados}\\ \text{Global (SCT)}\\ \text{g.l.=n-1}}} = \underbrace{\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2}_{\substack{\text{Suma de Cuadrados}\\ \text{Explicada (SCX)}\\ \text{g.l.=1}}} + \underbrace{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}_{\substack{\text{Suma de Cuadrados}\\ \text{Residual (SCR)}\\ \text{g.l.=n-2}}}$$

Tabla ANOVA del modelo de regresión

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

Si la hipótesis nula es cierta, es decir la variable X no influye, la recta de regresión es aproximadamente horizontal y se verifica que aproximadamente

$$\hat{y}_i = \overline{y}$$

por tanto SCX=0. Pero SCX es una medida con dimensiones y no puede utilizarse como medida de discrepancia. Para resolver este inconveniente se divide por la varianza residual y como estadístico del contraste de regresión se utiliza el siguiente

$$F_R = \frac{MCX}{MCR} \approx F_{1,n-2}$$

El contraste de la F es un contraste unilateral (de una cola)

Si el contraste resulta significativo, deben estimarse los parámetros α y β de la recta, así como las predicciones de Y para un valor x_0 y la estimación de la varianza residual mediante MCR

Se llama **coeficiente de determinación** al cociente **R^2=SCX/SCT** que indica la proporción de la variabilidad total que es explicada a través de la recta. Es un coeficiente que está entre 0 y 1. Es una medida de bondad de ajuste del modelo a los datos.

Intervalos de predicción e intervalos de confianza

Para cada valor x_o de X, el valor $\hat{y}_0 = \alpha + \beta x_0$ sirve para estimar la media de Y cuando X toma el valor x_o, pero también sirve para predecir el valor que tomará Y si $X = x_o$, pero el error que se comete al predecir el verdadero valor $Y(x_0)$ es mucho mayor que el error que se comete al estimar la media $\mu_{Y|x0}$

Para representar estos errores se construyen para cada valor x_0 de X y para cada nivel $100(1-\alpha)\%$ un intervalo de confianza de la media $\mu_{Y|x0}$ y un intervalo de confianza para la predicción $Y(x_0)$. Ambos intervalos están centrados en $\hat{y}_0$, pero el intervalo de predicción es mucho más amplio que el de la media.

Se llama **banda de confianzas de las medias** $\mu_{Y|x0}$, al nivel $100(1-\alpha)\%$ a la franja situada entre las dos líneas que forman, a ambos lados de la recta de ajuste, los extremos de los intervalos de confianza de las medias, con ese nivel, para todos los valores x_0 de X

Se llama **banda de predicción de la variable Y**, al nivel $100(1-\alpha)\%$ a la franja situada entre las dos líneas que forman a mabos lados de la recta de ajuste, los extremos de los intervalos de confianza de las predicciones con ese nivel para todos los valores x_0 de X