APÉNDICE 1: REGRESIÓN LINEAL SIMPLE

La técnica de regresión lineal simple se utiliza cuando se intenta explicar una variable respuesta cuantitativa en función de una variable explicativa también cuantitativa. Para esto definamos dos variables, la primera variable aleatoria Y, llamada variable dependiente, que supondremos relacionada con la segunda variable (no necesariamente aleatoria) que llamaremos variable independiente X.

A partir de una muestra de tamaño *n* para los que se cuenta con los valores de ambas variables, $(X_i, Y_i)\ ; i = 1,2,...,n$ el problema consiste en que por medio de la técnica de regresión lineal simple se trata de encontrar una recta que se ajuste a la nube de los n puntos (X_i, Y_i) dispuestos en el plano xy , y mediante esta recta se intenta predecir los valores de Y a partir de los de X.

El modelo pretende aproximar la variable respuesta mediante una función lineal de la variable explicativa de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad ; \quad i = 1,2,...,n$$

Donde:

Y_i: Es el escalar que se corresponde a la observación i-ésima de la variable dependiente del modelo.

X_i: Es la observación i-ésima de la variable independiente del modelo.

Los coeficientes β_0 y β_1 son desconocidos y deben ser estimados.

Siendo β el término independiente o constante y β el coeficiente de regresión de la variable explicativa o pendiente de la recta de regresión y el término ε es una perturbación estocástica agregada al modelo para recoger todos los posibles errores de medida tanto en las variables X e Y así como los errores en la especificación lineal del modelo, es decir recogerá todos aquellos factores que por error no se han incluido en el modelo y que pueden afectar a la variable dependiente del modelo.

Este término de perturbación ε indica en que medida las variables X e Y se apartan de la relación lineal.

De donde se supone que $\varepsilon \rightarrow \quad (0, \sigma\ )$

Por lo tanto en la técnica de regresión lineal se busca estimar los coeficientes (parámetros) de la ecuación tal que la sumatoria de los errores al cuadrado sea mínima:

$$Mín \sum_{i=1}^{n} (Y_i - \beta\ - \beta\ X_i)^2$$

Al resolver este problema aplicando cálculo diferencial, se obtienen los estimadores de mínimos cuadrados de los coeficientes de la recta de regresión:

$$\widehat{\beta}\ = \frac{\sum_{i=} \quad _i - \overline{X})(Y_i - \overline{\ })}{\sum_{i=} \quad _i - \overline{X})^2} \quad ; \qquad \widehat{\beta}\ = \overline{\ } - \beta\ \overline{X}$$

Del mismo modo que ocurre con otros estimadores, existirá cierta incertidumbre en el cálculo de las estimaciones, que se podrá reflejar mediante intervalos de confianza para ambos valores, construidos bajo la hipótesis de normalidad de los residuos, mediante las expresiones:

$$IC(1-\alpha\ \%(\beta_{\_}) = \left(\hat{\beta}_{\_} \pm t_{\alpha\ 2,n-2} S \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}} \right)$$

$$IC(1-\alpha\ \%(\beta\ ) = \left(\beta\ \pm t_{\alpha\ 2,n-2} \frac{S}{\sqrt{S_{XX}}} \right)$$

Donde:

$$S_{XX} = \sum_{i=}^{n} X_i^{\_} - \frac{\left(\sum_{i=}^{n} X_i \right)^{\_}}{n} \qquad y \quad S = \sqrt{\frac{\sum_{i=}^{n} Y_i - \bar{Y}_i)^2}{n-}}$$

y las desviaciones de los estimadores de los coeficientes del modelo están dados por:

$$\sigma_{\_} = \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}} \qquad \sigma_{\_} = \frac{S}{\sqrt{S_{XX}}}$$

El Coeficiente de determinación está dado por:

$$R^2 = \frac{SCR}{SCT}, \quad 0 \le R^2 \le$$

El Coeficiente de determinación ajustado se define así:

$$R_a^2 = -\left(\frac{n-}{n-p} * \frac{SCE}{SCT}\right), \quad 0 \le R^2 \le$$

Donde:

$$SCR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2$$

$$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$$

Análisis de la Varianza (ANOVA) en Regresión Lineal

El análisis de varianza en regresión es utilizado para verificar si el coeficiente de regresión o pendiente del modelo es igual a cero, es decir, confirmar si la variable independiente aporta explicación de la variable dependiente.

Contraste de hipótesis

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

$$EP : F = \frac{CMR}{CME}$$

$$RR : F \geq F_{(1-\alpha;1;(n-))}$$

Tabla ANOVA para Regresión Lineal

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión	1	SCR	CMR=SCR/1	F=CMR/CME
Error	n-2	SCE	CME=SCE/(n-2)	
Total	n-1	SCT		

Prueba t

En esta prueba se interesa conocer la significancia estadística del parámetro de regresión o coeficiente del modelo β respecto a un valor dado, para lo cual se establecen los siguientes contrastes de hipótesis:

$H_0 : \beta_{-} = \beta$
$H_1 : \beta_{-} \neq \beta$
$EP : t = \dfrac{\beta_{-} - \beta}{\sqrt{\dfrac{S^2}{\sum_{i=}^{n} x_i - x^{2}}}}$
$RR : |t| > t_{\frac{\alpha}{2},n-}$

$H_0 : \beta_{-} = \beta$
$H_1 : \beta_{-} > \beta$
$EP : t = \dfrac{\beta_{-} - \beta}{\sqrt{\dfrac{S^2}{\sum_{i=}^{n} x_i - x^{2}}}}$
$RR : t > t_{\alpha\ n-}$

$H_0 : \beta_{-} = \beta$
$H_1 : \beta_{-} < \beta$
$EP : t = \dfrac{\beta_{-} - \beta}{\sqrt{\dfrac{S^2}{\sum_{i=}^{n} x_i - x^{2}}}}$
$RR : t > -t_{\alpha\ n-}$

donde $S^2 = \dfrac{SCE}{n-}$.

APÉNDICE 2: ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales, ACP, es una técnica del análisis estadístico multivariado que se clasifica entre los métodos de simplificación o reducción de la dimensión y se aplica cuando se dispone de un conjunto elevado de variables, con datos cuantitativos persiguiendo obtener un menor número de variables, combinación lineal de las variables originales, que se denominan componentes principales o factores, cuya posterior interpretación permitirá un análisis más simple del problema estudiado.

El ACP tiene como finalidad transformar un conjunto de variables, a las que se las denomina variables originales interrelacionadas, en un nuevo conjunto de variables que son combinación lineal de las originales, denominadas componentes principales. Estas nuevas variables tienen la característica de estar incorrelacionadas entre sí.

En el ACP, se persigue explicar la mayor parte de la variabilidad total con el menor número de componentes, en donde cada componente como se dijo anteriormente está expresada en función de las variables observadas y es muy adecuado para resumir y reducir datos.

Algebraicamente, las componentes principales son una combinación lineal de las p variables aleatorias originales $X_1, X_2, \ldots, X_p$ y geométricamente esta

combinación lineal representa la elección de un nuevo sistema de coordenadas obtenidas al rotar el sistema original. Estos nuevos ejes representan la dirección de máxima variabilidad. Por lo tanto el ACP permite describir la estructura e interrelación de variables originales consideradas simultáneamente, determinando *q* combinaciones lineales de las *p* -variables originales que expliquen la mayor parte de la variación total, y de esta forma resumir y reducir los datos.

Sea $X^T = [X_1 \ X_2 \ ... \ X_p]$ un vector aleatorio p-variado, donde las variables que lo componen son las variables aleatorias originales y no necesariamente normales. El vector *p*-variado **X** tiene como matriz de varianzas y covarianzas a Σ, donde se tiene que $\lambda \geq \lambda \geq \ldots \geq \lambda \geq$) y $a_1, a_2, \ldots a_p$ son los valores y vectores propios de Σ, respectivamente.

Ahora, consideremos las siguientes combinaciones lineales:

$$Y_1 = \iota_1^T X = \iota_{11} X_1 + \iota_{12} X_2 + \cdots + \iota_{1p} X_p$$
$$Y_2 = \iota_2^T X = \iota_{21} X_1 + \iota_{22} X_2 + \cdots + \iota_{2p} X_p$$
$$\vdots$$
$$Y_p = \iota_p^T X = \iota_{p1} X_1 + \iota_{p2} X_2 + \cdots + \iota_{pp} X_p$$

Entonces las variables Y_1, Y_2,…,Y_p son las componentes principales, las mismas que no están correlacionadas entre sí, son ortonormales entre ellas y además se cumple que:

$$Var(Y_i) = a_i^T \sum \quad = \quad i = ,2,\ldots,p$$

$$Cov(Y_i, Y_j) = a_i^T \sum \quad = \quad i \neq \quad i, j = ,2,\ldots,p$$

Donde se cumple que:

$\|a_i\| =$ para i=1,2,….,p y $\langle a_i, a_j \rangle =$ para $i \neq$.

$\|a_i\|$ es la norma del vector a_i y $\langle a_i, a_j \rangle$ es el producto interno entre los vectores a_i y a_j.

La primera componente principal es la combinación lineal de $Y_1 = a_1^T X$ que maximiza la varianza de Y_1, donde $\|a_1\|$=1.

La segunda componente principal es la combinación lineal $Y_2 = a_2^T X$ que maximiza la varianza de Y_2, donde $\|a_2\|$=1 y la Cov (Y_1, Y_2)=0.

En general, la i-ésima componente principal es la combinación lineal que maximiza la varianza de $Y_i = a_i^T X_i$, sujeta a que la norma del vector a_i sea unitaria y que la $Cov(Y_i, Y_k) = 0$ para k < i.

Resumiendo tenemos que $\sum$ es la matriz de varianzas y covarianzas asociada con el vector aleatorio, $X^T = [X_1 \; X_2 \ldots X_p] \in \Re^p$, y que $\sum$ tiene los pares de valores y vectores propios $(\lambda_1, a_1), (\lambda_2, a_2), \ldots, (\lambda_p, a_p)$ donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

El porcentaje total de la varianza contenida por la i-ésima componente principal o su explicación está dado por:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

y el porcentaje total de la varianza contenida por las *q* primeras componentes principales se define así:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Existen algunos criterios para determinar el número de componentes principales a retener, los cuales son:

- **En general,** el criterio más sencillo para obtener el número *m* de componentes principales a retener debe ser tal que $\lambda_1, \lambda_2, \cdots, \lambda_m$ en conjunto expliquen más del 75% de la información total de la muestra.

- **Gráfico de sedimentación**. En este gráfico en el eje Y se representan los valores propios o raíces características y en el eje X el número de componentes principales correspondientes a cada valor propio en orden decreciente, de acuerdo a este gráfico se retienen aquellas componentes que se encuentran antes de que el gráfico presente un "quiebre" o “codo”.

- **Media aritmética.** según este criterio se retienen aquellas componentes tales que :

$$\lambda_i > \lambda = \frac{\sum_{i=1}^{p} \lambda}{p}$$

y se seleccionan aquellas componentes cuya raíz característica excede de la media de las raíces características.