

Tema 8. Introducción a la teoría general de modelos lineales. Regresión y análisis de la varianza

A. Hermoso Carazo

Universidad de Granada

Curso 2020/2021

La teoría de modelos lineales proporciona las bases para tratar muchos problemas reales en los que se pretende estudiar el comportamiento de un vector aleatorio en términos de otros vectores, que pueden ser también aleatorios o no.

Aquí realizaremos una breve introducción al modelo lineal básico, señalando los principales aspectos de la inferencia en estos modelos, y aplicaremos esta teoría a dos modelos concretos que resuelven los problemas de *regresión lineal* y de *análisis de la varianza de una vía*.

8.1. El modelo lineal general. Modelo de Gauss-Markov

Sea $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ un vector aleatorio n -dimensional y $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ ($k < n$) vectores n -dimensionales fijos, $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T$.

Se pretende describir el vector \mathbf{Y} mediante una combinación lineal de $\mathbf{x}_1, \dots, \mathbf{x}_k$, de manera que en esta descripción se cometa el menor error posible. Para ello, expresamos

$$\mathbf{Y} = \beta_1 \mathbf{x}_1 + \dots + \beta_k \mathbf{x}_k + \boldsymbol{\varepsilon},$$

donde β_1, \dots, β_k son constantes a determinar, que ponderan el efecto de cada vector $\mathbf{x}_1, \dots, \mathbf{x}_k$ sobre el vector \mathbf{Y} , y $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ es un vector aleatorio que representa el error que se comete si se describe \mathbf{Y} por $\beta_1 \mathbf{x}_1 + \dots + \beta_k \mathbf{x}_k$. En $\boldsymbol{\varepsilon}$ se engloban todos los factores aleatorios que afectan al comportamiento de \mathbf{Y} .

Una expresión de este tipo es lo que se conoce como *modelo lineal general*.

Por comodidad en los desarrollos, el modelo suele describirse usando notación matricial, como sigue.

Nota: El modelo descrito es un *modelo de efectos fijos*, ya que las componentes del vector de efectos son parámetros (desconocidos, pero no aleatorios). Se puede generalizar a modelos con *efectos aleatorios* y, también, con matriz de diseño aleatoria.

Dentro del modelo lineal general merece especial atención el denominado de Gauss-Markov, en cuyo estudio nos centraremos en este tema.

Modelo de Gauss-Markov

Es un modelo lineal de efectos fijos, en el que las componentes del vector de errores son variables aleatorias de segundo orden, centradas, homoce-dísticas (igual varianza) e incorreladas:

$$E[\varepsilon_i] = 0, \quad \text{Var}[\varepsilon_i] = \sigma^2, \quad i = 1, \dots, n, \quad \text{Cov}[\varepsilon_i \varepsilon_j] = 0, \quad i \neq j.$$

o, equivalentemente, expresado en forma matricial:

$$E[\boldsymbol{\varepsilon}] = 0, \quad \text{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_{n \times n}.$$

Nota: Si $Z = (Z_1, \dots, Z_n)^T$ es un vector aleatorio, $E[Z] = (E[Z_1], \dots, E[Z_n])^T$ y $\text{Cov}[Z] = E[(Z - E[Z])(Z - E[Z])^T]$.

Notemos que un modelo de Gauss-Markov está determinado por $k + 1$ parámetros, las componentes del vector de efectos, β_1, \dots, β_k , y la varianza común de los errores, σ^2 .

Ya que $E[\varepsilon_i] = 0$, $i = 1, \dots, n$, $\sigma^2 = E[\varepsilon_i^2]$, $i = 1, \dots, n$ y, por tanto, σ^2 proporciona un indicador de la magnitud de los errores cometidos al describir \mathbf{Y} por $\mathbf{X}\beta$.

Así, el estudio del modelo consiste en realizar inferencia sobre estos parámetros con objeto de:

- Determinar la relación lineal con el menor error posible (inferencia sobre los parámetros β_1, \dots, β_k).
- Analizar la adecuación del modelo, determinada por la magnitud de los errores $\varepsilon_1, \dots, \varepsilon_n$ (inferencia sobre σ^2).

La inferencia se realiza en base a una observación del vector \mathbf{Y} .

8.2. Estimación de un modelo de Gauss-Markov

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times k} \boldsymbol{\beta}_{k \times 1} + \boldsymbol{\varepsilon}_{n \times 1}, \quad E[\boldsymbol{\varepsilon}] = 0, \quad Cov[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_{n \times n}.$$

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, \quad Cov[\mathbf{Y}] = \sigma^2 \mathbf{I}_{n \times n}.$$

Hasta el momento, los problemas de inferencia se han resuelto observando un vector aleatorio (muestra de una variable) cuya distribución depende de los parámetros a estimar, y tiene una forma funcional conocida.

Ahora, la distribución del vector \mathbf{Y} , en el que se basa la inferencia, depende de los parámetros del modelo pero, en principio, no tiene una forma funcional conocida (sólo conocemos sus momentos de primer y segundo orden en términos de los parámetros).

Sin embargo, ya que el modelo lineal especifica, salvo error, la expresión de \mathbf{Y} en términos de $\boldsymbol{\beta}$, puede usarse el método de mínimos cuadrados (Tema 5) para estimar dicho vector.

Comenzamos, por tanto, estimando el vector de efectos por este método y, posteriormente, estimaremos la varianza común de los errores, σ^2 .

8.2.1. Estimación de mínimos cuadrados del vector de efectos

Se trata de minimizar la suma de los cuadrados de los errores cometidos al aproximar \mathbf{Y} por $\mathbf{X}\beta$; esto es, al aproximar cada componente de \mathbf{Y} por la correspondiente de $\mathbf{X}\beta$:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \longrightarrow Y_i = \sum_{j=1}^k x_{ij}\beta_j + \varepsilon_i, \quad i = 1, \dots, n.$$

Por tanto, se trata de minimizar la función:

$$S^2(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^k x_{ij}\beta_j \right)^2 = \|\varepsilon\|^2 = \|\mathbf{Y} - \mathbf{X}\beta\|^2.$$

Para ello, derivamos $S^2(\beta)$ respecto de las componentes de β , e igualando a cero las derivadas, obtenemos las denominadas *ecuaciones normales*:

$$\frac{\partial S^2(\beta)}{\partial \beta_h} = -2 \sum_{i=1}^n \left(Y_i - \sum_{j=1}^k x_{ij}\beta_j \right) x_{ih} = 0, \quad h = 1, \dots, k.$$

$$\sum_{i=1}^n Y_i x_{ih} = \sum_{i=1}^n \sum_{j=1}^k x_{ij} x_{ih} \beta_j, \quad h = 1, \dots, k \longrightarrow \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \beta.$$

Cualquier solución de estas ecuaciones que proporcione un mínimo absoluto de $S^2(\beta)$ es un *estimador de mínimos cuadrados de β* , y se nota

$$\hat{\beta}(\mathbf{Y}) = (\hat{\beta}_1(\mathbf{Y}), \dots, \hat{\beta}_k(\mathbf{Y}))^T.$$

- *Existencia*: mediante la teoría de proyecciones ortogonales puede probarse que siempre existe, al menos, un estimador de mínimos cuadrados de β .
- *Unicidad*: no está garantizada, salvo si el modelo es de rango máximo ya que, en tal caso, la matriz $\mathbf{X}^T \mathbf{X}$ es no singular:

$$\mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})\beta \Rightarrow \hat{\beta}(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Propiedades del estimador de mínimos cuadrados en modelos de rango máximo:

$$\hat{\beta} \equiv \hat{\beta}(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- $\hat{\beta}$ es función lineal de \mathbf{Y} .
 - * Ya que $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ es una matriz constante, cada componente de $\hat{\beta}$ es una función lineal de las componentes de \mathbf{Y} . \square
- $\hat{\beta}$ es insesgado.
 - * $E[\hat{\beta}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\mathbf{Y}] = ^1(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta$. \square
- $Cov[\hat{\beta}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$.
 - * $Cov[\hat{\beta}] = ^2(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Cov[\mathbf{Y}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$
 $= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. \square

Nota: Ya que la distribución de \mathbf{Y} depende de β y σ^2 , formalmente deberíamos notar E_{β, σ^2} , Cov_{β, σ^2} . Para simplificar, obviamos esta dependencia.

¹

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \longrightarrow E[\mathbf{Y}] = \mathbf{X}\beta, \quad Cov[\mathbf{Y}] = \sigma^2 \mathbf{I}.$$

²

$$Cov[AZ] = E[(AZ - AE[Z])(AZ - AE[Z])^T] = AE[(Z - E[Z])(Z - E[Z])^T]A^T = ACov[Z]A^T.$$

8.2.2. Funciones estimables

La metodología de mínimos cuadrados para la estimación en modelos lineales está justificada por las buenas propiedades de los estimadores obtenidos, no sólo los del vector de efectos, sino los de determinadas funciones del mismo, las denominadas *funciones estimables*.

Función estimable (escalar)

Una función paramétrica escalar, $\psi(\beta) \in \mathbb{R}$, es estimable si admite un estimador insesgado, función lineal de las componentes de \mathbf{Y} ; esto es:

$$\exists \mathbf{c} \in \mathbb{R}^n / E[\mathbf{c}^T \mathbf{Y}] = \psi(\beta), \quad \forall \beta,$$

o, equivalentemente, $\psi(\beta)$ es función lineal de las componentes de $\mathbf{X}\beta$, $\psi(\beta) = \mathbf{c}^T \mathbf{X}\beta$, $\mathbf{c} \in \mathbb{R}^n$.

La equivalencia en la definición se debe a que, si $\mathbf{c}^T \mathbf{Y}$ es insesgado en $\psi(\beta)$, entonces $\psi(\beta) = E[\mathbf{c}^T \mathbf{Y}] = \mathbf{c}^T E[\mathbf{Y}] = \mathbf{c}^T \mathbf{X}\beta$.

Toda función estimable es función lineal de β (por serlo de $\mathbf{X}\beta$). Sin embargo, en general, no toda función lineal de β es estimable, salvo que el modelo sea de rango máximo. En efecto, en tal caso, si $\psi(\beta) = \mathbf{a}^T\beta$ es una función lineal arbitraria, podemos expresar:

$$\psi(\beta) = \mathbf{a}^T\beta = (\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{X}\beta$$

de forma que $\psi(\beta)$ es función lineal de $\mathbf{X}\beta$ y, por tanto, estimable. Así:

En modelos de rango máximo:

$$\psi(\beta) \text{ es estimable} \Leftrightarrow \psi(\beta) = \mathbf{a}^T\beta, \quad \mathbf{a} \in \mathbb{R}^k.$$

El siguiente teorema establece el resultado fundamental en lo que se refiere a la estimación de estas funciones, justificando el uso del método de mínimos cuadrados para la estimación de β .

Teorema de Gauss-Markov

Si la función $\mathbf{a}^T\beta$ es estimable, entonces admite un único estimador lineal insesgado uniformemente de mínima varianza en la clase de estimadores lineales insesgados.

Dicho estimador es $\mathbf{a}^T(\hat{\beta}(\mathbf{Y}))$, y se denomina estimador de mínimos cuadrados de $\mathbf{a}^T\beta$.



8.2.3. Modelo estimado y residuos mínimo cuadráticos

Una vez estimado el vector de efectos por $\hat{\beta} \equiv \hat{\beta}(\mathbf{Y}) = (\hat{\beta}_1, \dots, \hat{\beta}_k)^T$, se approxima el vector \mathbf{Y} , obteniendo lo que se denomina el *modelo estimado*:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \rightarrow \hat{Y}_i = \sum_{j=1}^k x_{ij}\hat{\beta}_j, \quad i = 1, \dots, n.$$

Los errores cometidos al aproximar cada una de las variables Y_i por \hat{Y}_i definen los *residuos mínimo cuadrático del modelo*, y el vector formado por ellos, se denomina *vector de residuos*:

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}, \quad R_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n.$$

Propiedades:

i) \hat{Y}_i es el estimador de mínimos cuadrados de $E[Y_i]$, $i = 1, \dots, n$, y, por tanto, lineal, insesgado y de mínima varianza.

- * $E[\hat{Y}_i] = \sum_{j=1}^k x_{ij}\hat{\beta}_j$ es la componente i -ésima de $E[\mathbf{Y}] = \mathbf{X}\beta$ y, por tanto, es estimable. Por el teorema de Gauss-Markov, existe su estimador de mínimos cuadrados y se obtiene por sustitución directa, $\sum_{j=1}^k x_{ij}\hat{\beta}_j$.

8.2.3. Modelo estimado y residuos mínimo cuadráticos

Una vez estimado el vector de efectos por $\hat{\beta} \equiv \hat{\beta}(\mathbf{Y}) = (\hat{\beta}_1, \dots, \hat{\beta}_k)^T$, se approxima el vector \mathbf{Y} , obteniendo lo que se denomina el *modelo estimado*:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \longrightarrow \hat{Y}_i = \sum_{j=1}^k x_{ij}\hat{\beta}_j, \quad i = 1, \dots, n.$$

Los errores cometidos al aproximar cada una de las variables Y_i por \hat{Y}_i definen los *residuos mínimo cuadrático del modelo*, y el vector formado por ellos, se denomina *vector de residuos*:

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}, \quad R_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n.$$

Propiedades:

i) \hat{Y}_i es el estimador de mínimos cuadrados de $E[Y_i]$, $i = 1, \dots, n$, y, por tanto, lineal, insesgado y de mínima varianza. \blacksquare

* $E[\hat{Y}_i] = \sum_{j=1}^k x_{ij}\hat{\beta}_j$ es la componente i -ésima de $E[\mathbf{Y}] = \mathbf{X}\beta$ y, por tanto, es estimable. Por el *teorema de Gauss-Markov*, existe su estimador de mínimos cuadrados y se obtiene por sustitución directa, $\sum_{j=1}^k x_{ij}\hat{\beta}_j = \hat{Y}_i$. \blacksquare

ii) Los residuos son variables aleatorias con media cero.

$$* E[R_i] = E[Y_i] - E[\hat{Y}_i] \stackrel{i)}{=} E[Y_i] - E[Y_i] = 0. \quad \square$$

iii) El vector de residuos es ortogonal a los vectores columna de la matriz de diseño ($\mathbf{X}^T \mathbf{R} = 0$) y, por tanto, ortogonal a $\hat{\mathbf{Y}}$ ($\hat{\mathbf{Y}}^T \mathbf{R} = 0$).

$$* \mathbf{X}^T \mathbf{R} = \mathbf{X}^T (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X}\hat{\beta} = 0 \quad (\text{ecuaciones normales}).$$

$$* \hat{\mathbf{Y}}^T \mathbf{R} = \hat{\beta}^T \mathbf{X}^T \mathbf{R} = 0. \quad \square$$

La propiedad $\mathbf{X}^T \mathbf{R} = (x_1, \dots, x_k)^T \mathbf{R} = 0$ indica que existen k relaciones lineales entre los residuos, $x_j^T \mathbf{R} = \sum_{i=1}^n x_{ij}R_i = 0$, $j = 1, \dots, k$. Si \mathbf{X} es de rango r , sólo r de estas relaciones son linealmente independientes y, por tanto, el número de residuos R_1, \dots, R_n linealmente independientes es $n - r$.

Se dice que los residuos tienen *$n - r$ grados de libertad* (dados $n - r$ de ellos, el resto está determinado por éstos).

8.2.4. Estimación de la varianza: varianza residual

Ya que $\sigma^2 = E[\varepsilon_1^2] = \dots = E[\varepsilon_n^2]$, si los errores fuesen observables, sería natural estimar σ^2 a partir de sus cuadrados.

Por esta razón, la estimación de σ^2 se realiza a partir de los cuadrados de los residuos, ya que cada R_i proporciona una aproximación del error ε_i ($\varepsilon = \mathbf{Y} - \mathbf{X}\beta$ se aproxima por $\mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{R}$).

Aunque un estimador razonable podría ser la media aritmética de los cuadrados, $\sum_{i=1}^n R_i^2/n$, el que suele usarse es la suma de los cuadrados dividida por el número de residuos linealmente independientes, que se denomina **varianza residual**, y es un *estimador insesgado* de σ^2 :

$$S_R^2 = \frac{\sum_{i=1}^n R_i^2}{n-r} = \frac{\|\mathbf{R}\|^2}{n-r} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_F^2}{n-r} \quad (r : \text{rango de } \mathbf{X}).$$

Distribución normal n -dimensional

Un vector aleatorio $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ tiene distribución normal (n -dimensional), $\mathbf{Y} \rightarrow \mathcal{N}_n(\mu, \Sigma)$, si su función de densidad es:

$$f(y) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{(y-\mu)^T \Sigma^{-1} (y-\mu)}{2}}, \quad y \in \mathbb{R}^n,$$

donde $\mu = (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^n$ y $\Sigma = ((\sigma_{ij}))_{n \times n}$ es definida positiva.

Propiedades:

- $E[\mathbf{Y}] = \mu$, $Cov[\mathbf{Y}] = E[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^T] = \Sigma$.
- Las distribuciones marginales de cualquier dimensión son normales y, en particular, $Y_i \rightarrow \mathcal{N}(\mu_i, \sigma_{ii})$, $i = 1, \dots, n$.
- $\mathbf{Y} \rightarrow \mathcal{N}_n(\mu, \Sigma) \Rightarrow \forall \gamma \in \mathbb{R}^n$ (constante), $\mathbf{Y} + \gamma \rightarrow \mathcal{N}_n(\mu + \gamma, \Sigma)$.
- $\mathbf{Y} = (Y_1, \dots, Y_n)^T \rightarrow \mathcal{N}_n(\mu, \Sigma)$:
 Y_1, \dots, Y_n independientes $\Leftrightarrow \Sigma$ es diagonal $\Leftrightarrow \rho_{Y_i, Y_j} = 0$, $i \neq j$.

8.3. Inferencia bajo hipótesis de normalidad

Consideremos un modelo de Gauss-Markov, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, en el que el vector de errores (ε , equivalentemente, el vector \mathbf{Y}) tiene distribución normal:

$$\varepsilon \rightarrow \mathcal{N}_n(0, \sigma^2 \mathbf{I}_{n \times n}) \Leftrightarrow \mathbf{Y} = \mathbf{X}\beta + \varepsilon \rightarrow \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{n \times n}).$$

Al conocer la forma funcional de la distribución del vector \mathbf{Y} , en el que se basa la inferencia, podemos aplicar las técnicas estudiadas en temas anteriores, y obtener resultados adicionales en lo que se refiere a la inferencia sobre los parámetros.

En concreto, se pueden calcular los estimadores de máxima verosimilitud, aplicar el test de razón de verosimilitudes para distintos problemas de contraste y obtener intervalos de confianza para los parámetros.

Aquí nos centramos en la obtención del test de razón de verosimilitudes para contrastar lo que se denomina la *hipótesis lineal general*, que requiere la obtención de los estimadores de máxima verosimilitud.

8.3.1. Estimadores de máxima verosimilitud

La función de verosimilitud asociada a una observación de \mathbf{Y} , $y \in \mathbb{R}^n$, es:

$$L_y(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{\|y - \mathbf{X}\beta\|^2}{2\sigma^2}\right\}, \quad \forall \beta, \sigma^2.$$

- *Estimador de máxima verosimilitud de β :*

Para maximizar $L_y(\beta, \sigma^2)$ en β hay que minimizar $\|y - \mathbf{X}\beta\|^2$. Por tanto, *el estimador de máxima verosimilitud de β coincide con el de mínimos cuadrados, $\hat{\beta} \equiv \hat{\beta}(\mathbf{Y})$.*

- *Estimador de máxima verosimilitud de σ^2 :*

- $\ln L_y(\hat{\beta}(\mathbf{y}), \sigma^2) = -\ln(2\pi)^{n/2} - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \|y - \mathbf{X}\hat{\beta}(y)\|^2.$
- $\frac{\partial \ln L_y(\hat{\beta}(\mathbf{y}), \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|y - \mathbf{X}\hat{\beta}(y)\|^2 = 0$
 $\Rightarrow \hat{\sigma}^2(y) = \frac{\|y - \mathbf{X}\hat{\beta}(y)\|^2}{n}.$
- $$\hat{\sigma}^2(\mathbf{Y}) = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{n} = \frac{\sum_{i=1}^n R_i^2}{n} = \frac{(n-r)S_R^2}{n}.$$

8.3.2. Hipótesis lineal general: test de razón de verosimilitudes

Los contrastes más usuales en estos modelos se refieren a hipótesis que especifican un conjunto de relaciones lineales homogéneas entre las componentes de β , $H_0 : \mathbf{C}\beta = 0$, siendo \mathbf{C} una matriz conocida.

Bajo hipótesis de normalidad, el test de razón de verosimilitudes para contrastar una hipótesis de este tipo depende del llamado *estadístico F*, cuya distribución bajo H_0 es conocida bajo ciertas condiciones.

Hipótesis lineal general: *Es una hipótesis de la forma $H_0 : \mathbf{C}\beta = 0$, donde $\mathbf{C}_{q \times k}$ es una matriz conocida de rango $q \leq k$, verificando que todas las componentes del vector $\mathbf{C}\beta$ son funciones estimables.*

Para contrastar $H_0 : \mathbf{C}\beta = 0$ frente a $H_1 : \mathbf{C}\beta \neq 0$, usaremos el TRV:

$$\varphi(\mathbf{Y}) = \begin{cases} 1, & \lambda(\mathbf{Y}) < c \\ 0, & \lambda(\mathbf{Y}) \geq c, \end{cases} \quad \lambda(\mathbf{y}) = \frac{\sup_{H_0} L_y(\beta, \sigma^2)}{\sup_{\beta, \sigma^2} L_y(\beta, \sigma^2)} = \frac{L_y(\hat{\beta}^0, \hat{\sigma}_0^2)}{L_y(\beta, \sigma^2)}, \quad \mathbf{y} \in \mathbb{R}^n,$$

donde $(\hat{\beta}, \hat{\sigma}^2) \equiv (\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))$ y $(\hat{\beta}^0, \hat{\sigma}_0^2) \equiv (\hat{\beta}^0(\mathbf{y}), \hat{\sigma}_0^2(\mathbf{y}))$ son estimaciones máximo verosímiles de (β, σ^2) en todo el espacio paramétrico (calculadas previamente) y bajo H_0 , respectivamente.

Cálculo de $\hat{\beta}^0$ y $\hat{\sigma}_0^2$: Razonando como antes, se minimiza $\|\mathbf{y} - \mathbf{X}\beta\|^2$ bajo H_0 , y la solución, $\hat{\beta}^0$, se sustituye en la ecuación de verosimilitud de σ^2 , obteniéndose $\hat{\sigma}_0^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}^0\|^2/n$. Así:

$$\lambda(\mathbf{y}) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{n/2} = \left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}^0\|^2}{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2} \right)^{-n/2} = {}^3 \left(1 + \frac{q}{n-r} F(\mathbf{y}) \right)^{-n/2},$$

$$F(\mathbf{Y}) = \frac{n-r}{q} \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}^0\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2} \right) \xrightarrow{H_0} F(q, n-r).$$

Puesto que $\lambda(\mathbf{Y})$ decrece estrictamente con $F(\mathbf{Y})$, la región de rechazo del TRV es de la forma $F(\mathbf{Y}) > c'$, e imponiendo el tamaño exigido para determinar el punto crítico c' , se deduce el test de tamaño arbitrario:⁴

$$\text{TRV de tamaño } \alpha \in [0, 1] \longrightarrow \varphi(\mathbf{Y}) = \begin{cases} 1, & F(\mathbf{Y}) > F_{q, n-r; \alpha} \\ 0, & F(\mathbf{Y}) \leq F_{q, n-r; \alpha}. \end{cases}$$

³Sumando y restando $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2$ en el numerador.

⁴ $\alpha = \sup_{H_0} P(F(\mathbf{Y}) > c') = P(F(q, n-r) > c') \Rightarrow c' = F_{q, n-r; \alpha}$.

8.4. Modelo de regresión lineal simple

Este modelo constituye una de las principales aplicaciones de la teoría de modelos lineales. En términos generales, el problema de regresión consiste en expresar un vector aleatorio Y en términos de otros, X_1, \dots, X_k , con la mayor precisión posible. Esto es, determinar una función φ tal que al aproximar Y por $\varphi(X_1, \dots, X_k)$ se cometa el mínimo error. Así, un *modelo de regresión* se expresa como:

$$Y = \varphi(X_1, \dots, X_k) + \varepsilon,$$

donde ε es un vector que representa el error cometido si se describe Y por $\varphi(X_1, \dots, X_k)$.

Los vectores X_1, \dots, X_k se denominan *explicativos, independientes o regresores*, e Y es el *vector explicado, dependiente o respuesta*.

Aquí nos centraremos exclusivamente en el caso en que la función φ es lineal, sólo hay una variable explicativa, X , y X e Y son unidimensionales, lo que se denomina un *modelo de regresión lineal simple univariante*:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

El criterio para determinar la aproximación óptima será el de mínimos cuadrados y el problema se conoce como *regresión lineal mínimo cuadrática*.

Desde un punto de vista teórico, la solución a este problema es la *recta de regresión de mínimos cuadrados*, que requiere conocer las medias, varianzas y covarianza de las variables:

$$y = E[Y] + \frac{Cov[X, Y]}{Var[X]}(x - E[X]).$$

Ahora nos planteamos resolver el problema desde un punto de vista empírico: basándose exclusivamente en una serie de observaciones independientes del vector (X, Y) , $(x_1, y_1), \dots, (x_n, y_n)$, se trata de determinar la mejor aproximación de Y mediante una función lineal de X en el sentido de mínimos cuadrados.

Para abordar el problema desde la perspectiva de modelos lineales, se construye un modelo para aproximar Y a partir de las observaciones de X , (x_1, \dots, x_n) , y se estima dicho modelo a partir de las correspondientes observaciones de Y , (y_1, \dots, y_n) .

8.4.1. Planteamiento del modelo

Sea Y una variable aleatoria con momento de segundo orden finito, que se pretende aproximar mediante una función lineal de otra variable X .

Para abordar este problema desde la perspectiva de modelos lineales es preciso suponer que, fijado un valor arbitrario, $X = x$, la media de Y depende linealmente de dicho valor, y su dispersión respecto al valor medio es independiente de x :

$$E[Y/X = x] = \beta_0 + \beta_1 x, \quad \text{Var}[Y/X = x] = \sigma^2.$$

Bajo estos supuestos, si fijamos un conjunto de valores de X , x_1, \dots, x_n , y notamos Y_i , $i = 1, \dots, n$, a las variables que describen el comportamiento de Y cuando $X = x_i$ ($Y_i \equiv Y/X = x_i$), estas variables se pueden expresar como:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1 \dots n,$$

siendo ε_i variables aleatorias con $E[\varepsilon_i] = 0$, $\text{Var}[\varepsilon_i] = \sigma^2$, $i = 1, \dots, n$.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n \times 2} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2 \times 1} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \times 1}.$$

Por tanto, si $n > 2$, tenemos un modelo lineal. Si las variables Y_1, \dots, Y_n son independientes, los errores $\varepsilon_1, \dots, \varepsilon_n$ son independientes y, por tanto, incorrelados. De esta forma, tenemos un *modelo de Gauss-Markov* al que podemos aplicar todos los resultados previos.

Supondremos, además, que al menos dos de los valores x_1, \dots, x_n son distintos, de forma que el modelo es *rango máximo (2)*.

Notación: En lo sucesivo usaremos la siguiente notación:

$$\left| \begin{array}{l} \bar{x} = \frac{\sum\limits_{i=1}^n x_i}{n}, \quad \sigma_{x_l}^2 = \frac{\sum\limits_{i=1}^n (x_i - \bar{x})^2}{n}, \quad \bar{Y} = \frac{\sum\limits_{i=1}^n Y_i}{n}, \quad \sigma_Y^2 = \frac{\sum\limits_{i=1}^n (Y_i - \bar{Y})^2}{n}, \\ \sigma_{xY} = \frac{\sum\limits_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{n}. \end{array} \right.$$

8.4.2. Estimación del modelo

Estimador de mínimos cuadrados de $\beta = (\beta_0, \beta_1)^T$: Al ser el modelo de rango máximo, el estimador de mínimos cuadrados es único, y viene dado por $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$:

$$\hat{\beta} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} = \frac{1}{n\sigma_x^2} \begin{pmatrix} \sum_{i=1}^n x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}.$$

Notemos que $\sigma_x^2 \neq 0$ ya que $\exists x_i \neq x_j$; por tanto, $\hat{\beta}$ está bien definido:

- $\hat{\beta}_0 = \frac{\bar{Y} \sum_{i=1}^n x_i^2/n - \bar{x} \sum_{i=1}^n x_i Y_i/n}{\sigma_x^2} = \frac{\bar{Y}(\sigma_x^2 + \bar{x}^2) - \bar{x}(\sigma_x Y + \bar{x}\bar{Y})}{\sigma_x^2} = \bar{Y} - \bar{x} \frac{\sigma_x Y}{\sigma_x^2}.$
- $\hat{\beta}_1 = \frac{-\bar{x}\bar{Y} + \sum_{i=1}^n x_i Y_i/n}{\sigma_x^2} = \frac{\sigma_x Y}{\sigma_x^2}.$

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T, \quad \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sigma_x Y}{\sigma_x^2}}.$$

Propiedades de $\hat{\beta}$ (modelo de rango máximo):

- $E[\hat{\beta}] = \beta$ y, por tanto, $E[\hat{\beta}_0] = \beta_0$, $E[\hat{\beta}_1] = \beta_1$.
- $Cov[\hat{\beta}] = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$ y, por tanto:⁵

$$Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n\sigma_x^2} \right), \quad Var[\hat{\beta}_1] = \sigma^2 \frac{1}{n\sigma_x^2}, \quad Cov[\hat{\beta}_0, \hat{\beta}_1] = -\sigma^2 \frac{\bar{x}}{n\sigma_x^2}.$$

- Cualquier función lineal de β es estimable y, por el teorema de Gauss-Markov, el estimador de mínimos cuadrados de $a^T \beta$ es $a^T \hat{\beta}$.
- $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de mínimos cuadrados de β_0 y β_1 , respectivamente. Por tanto, son de mínima varianza uniformemente en la clase de estimadores lineales insesgados.

$$Cov[\hat{\beta}] = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\sigma_x^2} \begin{pmatrix} \sum_{i=1}^n x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

- **Modelo estimado, residuos mínimo cuadráticos y propiedades:**

- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \rightarrow \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} + \frac{\sigma_{xY}}{\sigma_x^2}(x_i - \bar{x}), \quad i = 1, \dots, n.$
- $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta} \rightarrow R_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} - \frac{\sigma_{xY}}{\sigma_x^2}(x_i - \bar{x}), \quad i = 1, \dots, n.$

■ *La suma de los residuos y, por tanto, su media aritmética, es nula:*

$$* \mathbf{X}^T \mathbf{R} = 0 \Rightarrow \sum_{i=1}^n R_i = 0 \Rightarrow \bar{R} = \sum_{i=1}^n R_i/n = 0. \quad \square$$

■ *La media aritmética de las variables aproximadas coincide con la media aritmética de las observadas:*

$$* \hat{Y}_i = Y_i + R_i \Rightarrow \sum_{i=1}^n \hat{Y}_i/n = \sum_{i=1}^n Y_i/n + \sum_{i=1}^n R_i/n = \bar{Y}. \quad \square$$

- $\hat{\mathbf{Y}}^T \mathbf{R} = \sum_{i=1}^n \hat{Y}_i R_i = 0.$

- **Varianza residual:** $S_R^2 = \frac{\sum_{i=1}^n R_i^2}{n-2}.$

8.4.3. Análisis de la bondad del modelo estimado

Una vez estimado el modelo, se plantea la cuestión de analizar si éste es adecuado; esto es, si las variables estimadas, \hat{Y}_i , se ajustan adecuadamente a las reales, Y_i . Para ello se usan los residuos, $R_i = Y_i - \hat{Y}_i$, $i = 1, \dots, n$ y se analiza la bondad del modelo en función de la magnitud de éstos (el modelo será tanto mejor cuanto más pequeños -más próximos a cero- sean los residuos).

Ya que los residuos son variables aleatorias con medias cero, como medida global de la magnitud de todos ellos podría usarse la suma de sus cuadrados, $\sum_{i=1}^n R_i^2$, o, por ejemplo, la varianza residual. Sin embargo, al no ser adimensionales, ni invariantes frente a cambios de escala, estas medidas no son apropiadas para hacer comparaciones.

La medida de bondad usada en estos modelos, que evita los inconvenientes anteriores, es el denominado *coeficiente de determinación lineal*.

El uso de este coeficiente se justifica a partir de la ecuación que desarrollamos a continuación, en la que se descompone la variabilidad de las observaciones aleatorias Y_1, \dots, Y_n .

Descomposición de la variabilidad de las observaciones:

Una forma común de medir de manera global la dispersión o variabilidad de un conjunto de datos numéricos es considerar la suma de las desviaciones cuadráticas respecto de su media aritmética.

Así, la variabilidad entre las variables Y_1, \dots, Y_n , que son observaciones de la variable Y para distintos valores de X , se mide por lo que se denomina la *variabilidad total de Y_1, \dots, Y_n* :

$$VT = \sum_{i=1}^n (Y_i - \bar{Y})^2 = n\sigma_Y^2.$$

Si se descompone cada observación como $Y_i = \hat{Y}_i + R_i$, $i = 1, \dots, n$, desarrollando los cuadrados, y usando que $\sum_{i=1}^n R_i = 0$ y $\sum_{i=1}^n \hat{Y}_i R_i = 0$, se tiene:

$$\begin{aligned} VT &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y} + R_i)^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n R_i^2 + 2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})R_i \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n R_i^2. \end{aligned}$$

Puesto que la media aritmética de $\hat{Y}_1, \dots, \hat{Y}_n$ es \bar{Y} , y la media aritmética de R_1, \dots, R_n es nula, los dos últimos sumandos miden la variabilidad de las aproximaciones y de los residuos, respectivamente, en el mismo sentido que se mide la de Y_1, \dots, Y_n .



Por tanto, la variabilidad total de las observaciones Y_1, \dots, Y_n se explica por dos componentes, la de las aproximaciones y la de los residuos:

$$VT = VE + VNE$$

- $VE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \frac{n\sigma_{xY}^2}{\sigma_x^2}$: *variabilidad explicada por el modelo.*
- $VNE = \sum_{i=1}^n R_i^2$: *variabilidad no explicada por el modelo.*

Así, puesto que el modelo estimado será tanto mejor cuanto menor sea $\sum_{i=1}^n R_i^2 = VNE$, su grado de bondad se mide por el siguiente coeficiente.

Coeficiente de determinación lineal: *Proporción de variabilidad total explicada por el modelo:*

$$R^2 = \frac{VE}{VT} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sigma_{xY}^2}{\sigma_x^2 \sigma_Y^2}.$$

R^2 es adimensional y $0 \leq R^2 \leq 1$. Su mayor proximidad a 1 indica más adecuación del modelo estimado.

8.4.4. Predicción a partir del modelo estimado

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ pueden usarse para *predecir futuras observaciones de la variable Y a partir de valores de X que no han intervenido en la estimación*. Esto constituye uno de los principales objetivos de la teoría de regresión.

Sea x_p un valor arbitrario de X e Y_p la variable que describe el comportamiento de Y cuando $X = x_p$. Según las hipótesis iniciales del modelo propuesto:

$$Y_p = \beta_0 + \beta_1 x_p + \varepsilon_p,$$

donde ε_p es la variable aleatoria que describe el error cometido si se approxima Y_p por $\beta_0 + \beta_1 x_p$, verificando $E[\varepsilon_p] = 0$ y $Var[\varepsilon_p] = \sigma^2$.

Por tanto, *la predicción de Y para $X = x_p$ según el modelo estimado es:*

$$\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p = \bar{Y} + \frac{\sigma_{xY}}{\sigma_x^2} (x_p - \bar{x}).$$

Propiedades:

- \hat{Y}_p es el estimador de mínimos cuadrados de $E[Y_p] = \beta_0 + \beta_1 x_p$.
 - * $E[Y_p]$ es estimable (función lineal de β). Su e.m.c. es $\hat{\beta}_0 + \hat{\beta}_1 x_p = \hat{Y}_p$.
- $E[\hat{Y}_p] = \beta_0 + \beta_1 x_p$ (Inmediato de lo anterior).
- $Var[\hat{Y}_p] = \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{n\sigma_x^2} \right]$.
 - * $Var[\hat{Y}_p] = Var[\hat{\beta}_0 + \hat{\beta}_1 x_p] = Var[\hat{\beta}_0] + x_p^2 Var[\hat{\beta}_0] + 2x_p Cov[\hat{\beta}_0, \hat{\beta}_1]$.
- Si la observación de Y_p se realiza independientemente de las observaciones Y_1, \dots, Y_n , \hat{Y}_p es independiente de Y_p y

$$ECM(\hat{Y}_p) = E[(\hat{Y}_p - Y_p)^2] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n\sigma_x^2} \right].$$

$$* E[(\hat{Y}_p - Y_p)^2] = E[(\hat{Y}_p - E[Y_p] + E[Y_p] - Y_p)^2] = Var[\hat{Y}_p] + Var[Y_p].$$

La expresión del error cuadrático medio, que se usa para comparar diferentes predicciones, indica que éstas serán mejores cuanto más próximo esté x_p a la media de los valores x_1, \dots, x_n usados en el modelo y, también, cuanto más dispersos estén dichos valores.

8.4.5. Recta de regresión estimada

En la práctica, si disponemos de un conjunto de observaciones del vector aleatorio (X, Y) tomadas de forma independiente, $(x_1, y_1), \dots, (x_n, y_n)$, consideramos el modelo lineal correspondiente a x_1, \dots, x_n :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

La estimación de este modelo a partir de observaciones independientes Y_1, \dots, Y_n , correspondientes a los valores x_1, \dots, x_n es:

$$\hat{Y}_i = \bar{Y} + \frac{\sigma_{xy}}{\sigma_x^2} (x_i - \bar{x}), \quad i = 1, \dots, n,$$

y ya que las observaciones y_1, \dots, y_n son valores concretos de Y_1, \dots, Y_n , las aproximaciones concretas obtenidas a partir de ellos son:

$$\hat{y}_i = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} (x_i - \bar{x}), \quad i = 1, \dots, n.$$

La recta determinada por los puntos $(x_1, \hat{y}_1), \dots, (x_n, \hat{y}_n)$ se denomina *recta de regresión estimada a partir de $(x_1, y_1), \dots, (x_n, y_n)$* :

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x}).$$

8.5. Análisis de la varianza de una vía

El *análisis de la varianza de una vía, o con un factor de variación (ANOVA)* es una técnica estadística propuesta por Fisher para contrastar si un supuesto factor de variación, con un número finito de niveles, afecta al comportamiento de una determinada variable aleatoria.

La técnica consiste en tomar una muestra de la variable bajo cada uno de los niveles del factor de variación y descomponer la variabilidad total de los datos en dos sumandos, uno que exprese la variabilidad entre las distintas muestras, y otro que exprese la variabilidad intrínseca de los datos muestrales por el hecho de ser observaciones de variables aleatorias.

Fisher desarrolló la técnica a partir de esta idea intuitiva y de los siguientes supuestos:

- *La variable de interés no está afectada por factores distintos al que es objeto de estudio.*
- *En cada uno de los k niveles del factor de variación, la variable tiene distribución normal, y la varianza en todos los niveles es la misma:*

$$Y_i : \text{variable de interés en el nivel } i\text{-ésimo} \rightarrow \mathcal{N}(\mu_i, \sigma^2), \quad i = 1, \dots, k.$$

Bajo estos supuestos, contrastar que el factor de variación no afecta al comportamiento de la variable equivale a contrastar que todas las variables Y_i tienen la misma distribución o, equivalentemente, la misma media.

Problema: *contrastar la igualdad de medias de k poblaciones normales con varianza común:*

$$H_0 : \mu_1 = \dots = \mu_k.$$

Para abordar este problema, se toma muestra aleatoria simple de cada variable, todas *independientes*:

$(Y_{i1}, \dots, Y_{in_i})$ muestra aleatoria simple de Y_i , $i = 1, \dots, k$,

y el objetivo es decidir si las observaciones aportan evidencia para rechazar H_0 , lo que indicará que el factor considerado afecta a la variable objeto de estudio.

Aquí vamos a resolver el problema a partir de teoría general de modelos lineales. Para ello, expresamos cada variable muestral, $Y_{ij} \rightarrow \mathcal{N}(\mu_i, \sigma^2)$, como:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \rightarrow \mathcal{N}(0, \sigma^2); \quad i = 1, \dots, k; \quad j = 1, \dots, n_i,$$

siendo las variables ε_{ij} *independientes y, por tanto, incorreladas*.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \hline Y_{21} \\ \vdots \\ Y_{2n_2} \\ \hline \vdots \\ \hline Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix}_{n \times 1} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline & \vdots & & \vdots \\ & \vdots & & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n \times k} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}_{k \times 1} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \hline \vdots \\ \hline \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}_{n \times 1}$$

$$E[\varepsilon_{ij}] = 0, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i; \quad E[\varepsilon_{ij}\varepsilon_{i'j'}] = 0, \quad i \neq i' \text{ ó } j \neq j'$$

$$\Downarrow \quad n = \sum_{i=1}^k n_i > k$$

Modelo de Gauss-Markov de rango máximo ($r = k$)

8.5.1. Estimación del modelo, residuos y propiedades

En lo que sigue, las medias aritméticas de todas las variables del modelo y de cada muestra serán denotadas como:

$$\bar{Y} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}}{n}; \quad \bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}, \quad i = 1, \dots, k.$$

- **Estimador de mínimos cuadrados de $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T$:**

$$S^2(\boldsymbol{\mu}) = \sum_{i=1}^k \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2.$$

$$\frac{\partial S^2(\boldsymbol{\mu})}{\partial \mu_i} = \frac{d \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2}{d \mu_i} = -2 \sum_{j=1}^{n_i} (Y_{ij} - \mu_i) = 0 \Rightarrow \hat{\mu}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} = \bar{Y}_i, \quad i = 1, \dots, k.$$

$$\Downarrow \quad \hat{\boldsymbol{\mu}} = (\bar{Y}_1, \dots, \bar{Y}_k)^T.$$

Esto es, la media de cada variable Y_i , $i = 1, \dots, k$, se estima por la media muestral correspondiente.

- **Modelo estimado:** $\hat{Y} = X\hat{\mu} = (\underbrace{\bar{Y}_1, \dots, \bar{Y}_1}_{n_1} | \cdots | \underbrace{\bar{Y}_k, \dots, \bar{Y}_k}_{n_k})^T$.

Esto es, cada variable Y_{ij} se aproxima por $\hat{Y}_{ij} = \bar{Y}_i$, la media muestral correspondiente.

- **Residuos:** $R_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i; i = 1, \dots, k, j = 1, \dots, n_i$.

- *La suma de los residuos de cada muestra es nula. Por tanto, la suma de todos los residuos y, consecuentemente, su media aritmética, es nula:*

$$* X^T R = 0 \Rightarrow \sum_{j=1}^{n_i} R_{ij} = 0, i = 1, \dots, k \Rightarrow \bar{R} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij} = 0. \quad \square$$

- *La media aritmética de las variables aproximadas coincide con la media aritmética de las observadas:*

$$* \hat{Y}_{ij} = Y_{ij} + R_{ij} \Rightarrow \sum_{i=1}^k \sum_{j=1}^{n_i} \hat{Y}_{ij} / n = \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} / n + \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij} / n = \bar{Y}. \quad \square$$

- $\hat{Y}^T R = \sum_{i=1}^k \sum_{j=1}^{n_i} \hat{Y}_{ij} R_{ij} = 0$.

8.5.2. Descomposición de la variabilidad

Como en el modelo de regresión lineal, la variabilidad de las variables del modelo, Y_{ij} , que se mide por la suma de las desviaciones cuadráticas respecto de su media aritmética, se expresa como suma de la variabilidad de las aproximaciones y la de los residuos:

$$VT = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

Para la descomposición, expresamos $Y_{ij} = \hat{Y}_{ij} + R_{ij}$. Desarrollando los cuadrados y teniendo en cuenta que $\sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij} = 0$ y $\sum_{i=1}^k \sum_{j=1}^{n_i} \hat{Y}_{ij} R_{ij} = 0$, se tiene:

$$\begin{aligned} VT &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} + R_{ij} - \bar{Y})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y}) R_{ij} \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}^2 \end{aligned}$$

Puesto que la media aritmética de las variables aproximadas es \bar{Y} y la de los residuos es nula, los sumandos de la descomposición miden la variabilidad de las aproximaciones y de los residuos, respectivamente. Por tanto, la variabilidad total de las observaciones se explica por dos componentes:

$$VT = VE + VNE$$

- $VE = \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y})^2$: *variabilidad explicada por el modelo.*
- $VNE = \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}^2$: *variabilidad no explicada por el modelo.*

Además, en este contexto, las componentes de variabilidad tienen una interpretación adicional al ser $\hat{Y}_{ij} = \bar{Y}_i$:

- $VE = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$: *variabilidad entre grupos.*

Mide la variabilidad de las medias muestrales de los distintos grupos.

- $VNE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$: *variabilidad dentro de grupos.*

Para cada $i = 1, \dots, n$, $\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ mide la variabilidad de las variables de la muestra i -ésima. Por lo tanto, VNE es una medida de la variabilidad dentro de las muestras.

8.5.3. Problema de contraste y test de razón de verosimilitudes

La hipótesis a contrastar, $H_0 : \mu_1 = \dots = \mu_k$, puede expresarse como

$$H_0 : \mu_1 - \mu_2 = 0, \mu_1 - \mu_3 = 0, \dots, \mu_1 - \mu_k = 0,$$

y especifica, por tanto, $k - 1$ relaciones lineales homogéneas entre las componentes de μ . Ya que el modelo es de rango máximo, toda función lineal es estimable y H_0 es un caso particular de la hipótesis lineal general sobre el vector de efectos:

$$H_0 : \mathbf{C}\mu = 0 \implies \mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{pmatrix}_{(k-1) \times k} \quad (\text{rango } k-1).$$

Por tanto, el contraste puede resolverse mediante el test de razón de verosimilitudes a partir del estadístico F :

$$F(\mathbf{Y}) = \frac{n-k}{q} \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\mu}^0\|^2 - \|\mathbf{Y} - \mathbf{X}\hat{\mu}\|^2}{\|\mathbf{Y} - \mathbf{X}\hat{\mu}\|^2} \right) = \frac{n-k}{k-1} \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\mu}^0\|^2 - VNE}{VNE} \right).$$

Cálculo de $\|\mathbf{Y} - \mathbf{X}\hat{\mu}^0\|^2$: Bajo $H_0 : \mu_1 = \dots = \mu_k$, si μ es el valor común:

$$Y_{ij} \rightarrow \mathcal{N}(\mu, \sigma^2); i = 1, \dots, k, j = 1, \dots, n_i, \text{ independientes (m.a.s.)} \Rightarrow \hat{\mu} = \bar{Y}.$$

Entonces, $\hat{\mu}^0 = (\bar{Y}, \dots, \bar{Y})^T$, $\mathbf{X}\hat{\mu}^0 = (\bar{Y}, \dots, \bar{Y} | \dots | \bar{Y}, \dots, \bar{Y})^T$, y

$$\|\mathbf{Y} - \mathbf{X}\hat{\mu}^0\|^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = VT.$$

Por tanto:

$$F(\mathbf{Y}) = \frac{n-k}{k-1} \frac{VT - VNE}{VNE} = \frac{VE/(k-1)}{VNE/(n-k)} = \frac{S_E^2}{S_R^2},$$

donde S_R^2 es la varianza residual, que en este contexto suele denominarse **varianza dentro de grupos** y S_E^2 es la denominada **varianza entre grupos**.

Notemos que $F(\mathbf{Y})$ proporciona una medida de la variabilidad entre grupos respecto a la variabilidad dentro de grupos. Así, de acuerdo a la idea original de Fisher, el TRV rechaza H_0 si la variabilidad entre grupos es grande respecto a la variabilidad dentro de grupos:

$$\text{TRV de tamaño } \alpha \rightarrow \varphi(\mathbf{Y}) = \begin{cases} 1, & F(\mathbf{Y}) > F_{k-1, n-k; \alpha} \\ 0, & F(\mathbf{Y}) \leq F_{k-1, n-k; \alpha}. \end{cases}$$

TABLA ANOVA DE UNA VÍA

Fuentes de Variación	Variabilidad	Grados de libertad	Varianzas
Entre grupos	$VE = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$k-1$	$S_E^2 = VE/(k-1)$
Dentro de grupos	$VNE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$n-k$	$S_R^2 = VNE/(n-k)$
Total	$VT = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = n\sigma_Y^2$	$n-1$	

Nota: Si el problema planteado no especifica un nivel de significación se trabaja con el denominado **p-nivel o p-valor asociado a los datos**:

$$p-\text{nivel} = P(F(k-1, n-k) > F_{exp}), \quad F_{exp} = F(\mathbf{y}).$$

En general, si el **p-nivel** es pequeño (0.05 o menor) se rechaza H_0 ; si es grande (0.15 o mayor), se acepta H_0 . Para valores intermedios hay que tratar cada situación particular, aunque, normalmente, es aconsejable tomar más datos y rehacer los cálculos.

El **p-nivel** marca los niveles de significación para los que se acepta o rechaza H_0 .

Problema 5

Una compañía farmacéutica investiga los efectos de 5 compuestos. El experimento consiste en inyectar los compuestos a 13 ratas de características similares y anotar los tiempos de reacción. Los animales se clasifican en 5 grupos de 4, 2, 2, 3 y 2 ratas, respectivamente, y a cada grupo se le administra un compuesto diferente, obteniéndose los resultados de la siguiente tabla:

Grupo	Tiempo de reacción (en minutos)			
1	8.3	7.6	8.4	8.3
2	7.4	7.1		
3	8.1	6.4		
4	7.9	8.5	10.0	
5	7.1	8		

Suponiendo que se verifican las hipótesis de normalidad, aleatoriedad, independencia e igualdad de varianzas, contrastar la hipótesis de que los tiempos medios de reacción coinciden en los cinco grupos y, por tanto, la eficacia de los cinco compuestos es la misma.

Al ser las ratas de características similares, se considera que no hay otros factores, salvo el tipo de compuesto, que influyan en el tiempo de reacción.

Y_i : Tiempo de reacción de las ratas del grupo i -ésimo, $i = 1, 2, 3, 4, 5$.

* $Y_i \rightarrow \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4, 5$.

* Muestras aleatorias simples de cada variable, independientes, de tamaños 4, 2, 2, 3, 2.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{alguna difiere.}$$

$$\bar{y}_1 = \frac{\sum_{j=1}^4 y_{1j}}{4} = 8.15, \quad \bar{y}_2 = \frac{\sum_{j=1}^2 y_{2j}}{2} = 7.25, \quad \bar{y}_3 = \frac{\sum_{j=1}^2 y_{3j}}{2} = 7.25, \quad \bar{y}_4 = \frac{\sum_{j=1}^3 y_{4j}}{3} = 8.8,$$

$$\bar{y}_5 = \frac{\sum_{j=1}^2 y_{5j}}{2} = 7.55; \quad \bar{y} = \frac{1}{13} \sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} = 7.93077.$$

- $VT = \sum_{i=1}^5 \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij}^2 - 13\bar{y}^2 = 826.91 - 817.66231 = 9.24769$.
- $VE = \sum_{i=1}^5 n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^5 n_i \bar{y}_i^2 - 13\bar{y}^2 = 822.265 - 817.66231 = 4.60269$.
- $VNE = VT - VE = 4.645$. | I

Fuente de variación	Variabilidad	G.L.	Varianzas
Entre grupos	$VE = 4.60269$	4	$s_E^2 = VE/4 = 1.15067$
Dentro de grupos	$VNE = 4.645$	8	$s_R^2 = VNE/8 = 0.58062$

$$F_{exp} = \frac{s_E^2}{s_R^2} = 1.98178.$$

$$p - nivel = P(F(4,8) > 1.98178) = 0.19037.^1$$

Los datos no aportan evidencia para rechazar H_0 (a cualquier nivel de significación menor que 0.19037, se acepta H_0). Por tanto, a partir de los datos se concluye que el tipo de compuesto no afecta al tiempo de reacción.

¹

$P(F(4,8) > 1.98178) = 0.2048$ (TABLAS, interpolando.)

Problema 6

Se quiere estudiar la eficacia de tres fertilizantes, A, B y C, en la producción de cierto fruto. Para ello se aplica el A en 8 parcelas, el B en 6, y el C en 12 parcelas. Las parcelas son de características similares en cuanto a fertilidad, por lo que se considera que las diferencias en la producción, si las hay, serán debidas al tipo de fertilizante. Las toneladas producidas en cada parcela en una determinada temporada son:

A	6	7	5	6	5	8	4	7
B	10	9	9	10	10	6		
C	3	4	8	3	7	6	3	6

Suponiendo que las tres muestras proceden de poblaciones normales con varianzas iguales, contrastar la hipótesis de que los abonos son igualmente eficaces.

Y_i : Número de toneladas producidas con cada fertilizante, $i = 1$ (A), 2 (B), 3 (C).

* $Y_i \rightarrow \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2, 3$.

* Muestras aleatorias simples de cada variable, independientes, de tamaños 8, 6 y 12.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H₁ : alguna difiere.

$$\bar{y}_1 = \frac{\sum_{j=1}^8 y_{1j}}{8} = 6, \quad \bar{y}_2 = \frac{\sum_{j=1}^6 y_{2j}}{6} = 9, \quad \bar{y}_3 = \frac{\sum_{j=1}^{12} y_{3j}}{12} = 5, \quad \bar{y} = \frac{1}{26} \sum_{i=1}^3 \sum_{j=1}^{n_i} y_{ij} = 6.23077.$$

- $VT = \sum_{i=1}^3 \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^3 \sum_{j=1}^{n_i} y_{ij}^2 - 26\bar{y}^2 = 1136 - 1009.38462 = 126.61538.$
- $VE = \sum_{i=1}^3 n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^3 n_i \bar{y}_i^2 - 26\bar{y}^2 = 1074 - 1009.38462 = 64.61538.$
- $VNE = VT - VE = 62.$

Fuente de variación	Variabilidad	G.L.	Varianzas
Entre grupos	$VE = 64.61538$	2	$s_E^2 = VE/2 = 32.30769$
Dentro de grupos	$VNE = 62$	23	$s_R^2 = VNE/23 = 2.69565$

$$F_{exp} = \frac{s_E^2}{s_R^2} = 11.98511 \rightarrow p-nivel = P(F(2, 23) > 11.98511) = 0.00027.^2$$

Los datos aportan evidencia para rechazar H_0 . A partir de ellos se concluye que los abonos no son iguales en cuanto a la cantidad de toneladas producidas.

2

$$P(F(2, 23) > 11.98511) < 0.0005 \text{ (TABLAS, para } F(2, 20) \text{ y } F(2, 24)).$$



Los siguientes datos corresponden a observaciones del consumo medio (en Kw/h) realizado por 5 tipos de calefactores para mantener una habitación a una temperatura determinada durante todo un día:

Tipo	Consumo (en kw/h)				
1	14.5	14.1	14.6	14.2	
2	13.2	13.4	13.0		
3	13.7	13.6	14.1	13.8	14.0
4	12.7	13.1	12.8	12.9	13.3
5	14.6	15.2	14.4	14.8	14.3

Contrastar la hipótesis de igualdad de los consumos medios de los diferentes tipos de calefactores. ¿Bajo qué hipótesis se puede realizar este contraste?

Debe suponerse que las habitaciones usadas en el experimento tienen las mismas características y no hay otros factores, salvo el tipo de calefactor, que influya en la temperatura de la habitación durante el día.

Y_i : Consumo medio por hora de cada tipo de calefactor, $i = 1, 2, 3, 4, 5$.

$$* Y_i \rightarrow \mathcal{N}(\mu_i, \sigma^2), i = 1, 2, 3, 4, 5.$$

* Muestras aleatorias simples de cada variable, independientes, de tamaños 4, 3, 5, 6 y 5.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{alguna difiere.}$$

$$\bar{y}_1 = \frac{\sum_{j=1}^4 y_{1j}}{4} = 14.35, \quad \bar{y}_2 = \frac{\sum_{j=1}^3 y_{2j}}{3} = 13.2, \quad \bar{y}_3 = \frac{\sum_{j=1}^5 y_{3j}}{5} = 13.84,$$

$$\bar{y}_4 = \frac{\sum_{j=1}^6 y_{4j}}{6} = 13, \quad \bar{y}_5 = \frac{\sum_{j=1}^5 y_{5j}}{5} = 14.66; \quad \bar{y} = \frac{1}{26} \sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} = 13.80435.$$

- $VT = \sum_{i=1}^5 \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij}^2 - 23\bar{y}^2 = 4393.93 - 4382.88043 = 11.04957$.
- $VE = \sum_{i=1}^5 n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^5 n_i \bar{y}_i^2 - 23\bar{y}^2 = 4392.716 - 4382.88043 = 9.83557$.
- $VNE = VT - VE = 1.214$.

Fuente de variación	Variabilidad	G.L.	Varianzas
Entre grupos	$VE = 9.83557$	4	$s_E^2 = VE/4 = 2.45889$
Dentro de grupos	$VNE = 1.214$	18	$s_R^2 = VNE/18 = 0.06744$

$$F_{exp} = \frac{s_E^2}{s_R^2} = 36.45803.$$

$$p-nivel = P(F(4, 18) > 36.45803) = 21 \times 10^{-9}.^3$$

Por tanto, a la vista de los datos, debe rechazarse H_0 . O sea, los 5 tipos de calefactores no tienen el mismo consumo.

Problema 8

En un tratamiento contra la hipertensión se seleccionaron 35 enfermos de características similares. Los enfermos se distribuyeron en cuatro grupos de 10 (P, A, B y AB). El grupo P tomó "placebo" (fármaco inocuo), el grupo A tomó un fármaco "A", el grupo B un fármaco "B" y el grupo AB una asociación entre "A" y "B". Para valorar la eficacia de los tratamientos, se registró el descenso de la presión diastólica desde el inicio del tratamiento hasta después de una semana de tratamiento. Los resultados, después de registrarse algunos abandonos, fueron:

P	10	0	15	-20	0	15	-5			
A	20	25	33	25	30	18	27	0	35	20
B	15	10	25	30	15	35	25	22	11	25
AB	10	5	-5	15	20	20	0	10	I	

A la vista de estos datos, ¿Puede afirmarse que el descenso de la presión diastólica coincide en los cuatro grupos? ¿Bajo qué hipótesis?

Al ser los enfermos de características similares, se considera que no hay otros factores, salvo el tipo de medicación, que influyan en el descenso de la presión.

Y_i : Descenso de la presión diastólica del grupo i -ésimo, $i = 1, 2, 3, 4$.

* $Y_i \rightarrow \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2, 3, 4$.

* Muestras aleatorias simples de cada variable, independientes, de tamaños 7, 10, 10 y 8.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \text{alguna difiere.}$$

Fuente de variación	Variabilidad	G.L.	Varianzas
Entre grupos	$VE = 2492.61071$	3	$s_E^2 = VE/3 = 830.87024$
Dentro de grupos	$VNE = 3020.93214$	31	$s_R^2 = VNE/31 = 97.44942$

$$F_{exp} = \frac{s_E^2}{s_R^2} = 8.52617.$$

$$p - \text{nivel} = P(F(3, 31) > 8.52617) = 0.00028.^4$$

Por tanto, a la vista de los datos, debe rechazarse H_0 .¹ El descenso de la presión varía según los grupos.

4

$P(F(3, 31) > 8.52617) < 0.0005$ (TABLAS, para $F(3, 30)$ y $F(3, 40)$)

Problema 4

En cierto estudio sobre la relación entre el diámetro de los guisantes (X) y el diámetro medio de sus descendientes (Y), Galton obtuvo los siguientes resultados:

D. Padres	21	20	19	18	17	16	15
D. Descendientes	17.26	17.07	16.37	16.4	16.13	16.17	15.98

- a) Determinar el modelo de regresión lineal estimado de Y sobre X e interpretar el valor estimado de la pendiente. Dar la predicción del diámetro de los guisantes cuyos progenitores tienen un diámetro de 18.5. Dar una medida de la bondad del ajuste de los datos a la recta estimada.
- b) Suponiendo las hipótesis adecuadas de normalidad, ¿puede deducirse, al nivel 0.05, que no hay relación lineal entre las variables consideradas? Relacionar este resultado con las conclusiones anteriores.
-
-

Valores	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
	21	17.26	441	362.46	297.9076
	20	17.07	400	341.4	291.3849
	19	16.37	361	311.03	267.9769
	18	16.4	324	295.2	268.96
	17	16.13	289	274.21	260.1769
	16	16.17	256	258.72	261.4689
	15	15.98	225	239.7	255.3604
Sumas	126	115.38	2296	2082.72	1903.2356

$$\bar{x} = \frac{\sum_{i=1}^7 x_i}{7} = 18, \quad \bar{y} = \frac{\sum_{i=1}^7 y_i}{7} = 16.48286.$$

$$\sigma_x^2 = \frac{\sum_{i=1}^7 (x_i - \bar{x})^2}{7} = \frac{\sum_{i=1}^7 x_i^2}{7} - \bar{x}^2 = 4.$$

$$\sigma_y^2 = \frac{\sum_{i=1}^7 (y_i - \bar{y})^2}{7} = \frac{\sum_{i=1}^7 y_i^2}{7} - \bar{y}^2 = 0.20622.$$

$$\sigma_{xy} = \frac{\sum_{i=1}^7 (x_i - \bar{x})(y_i - \bar{y})}{7} = \frac{\sum_{i=1}^7 x_i y_i}{7} - \bar{x}\bar{y} = 0.84.$$

a) Determinar el modelo de regresión lineal estimado de Y sobre X e interpretar el valor estimado de la pendiente. Dar una predicción para el diámetro previsto de un guisante cuyo progenitor tenga un diámetro de 18.5. Dar una medida de la bondad del ajuste de los datos a la recta estimada.

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x}) \rightarrow y = 12.70286 + 0.21x.$$

Interpretación de la pendiente: A partir de los datos se estima que cada unidad de crecimiento en el diámetro de los progenitores dará un aumento de 0.21 unidades en el diámetro de los descendientes.

Predicción para $x = 18.5 \rightarrow \hat{y} = 12.70286 + (0.21 \times 18.5) = 16.58786.$

Medida de bondad del ajuste → coeficiente de determinación:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = 0.8554.$$

El alto valor de r^2 significa que el modelo lineal es bastante adecuado para predecir el diámetro de los descendientes a partir del de los progenitores.

b) Suponiendo las hipótesis adecuadas de normalidad, ¿puede deducirse, al nivel $\alpha = 0.05$, que no hay relación lineal entre las variables consideradas? Relacionar este resultado con las conclusiones anteriores.

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \rightarrow \text{Estadístico de contraste : } F(\mathbf{Y}) = \frac{VE}{VNE/(n-2)} \stackrel{H_0}{\rightarrow} F_{1,n-2}$$

$$\left. \begin{aligned} \bullet VT &= n\sigma_Y^2 &\rightarrow 1.44354. \\ \bullet VE &= \frac{n\sigma_{xY}^2}{\sigma_x^2} &\rightarrow 1.2348. \end{aligned} \right\} \rightarrow VNE = VT - VE = 0.20874.$$

$$F_{exp} = 29.57706, \quad F_{1,5; 0.05} = 6.61.$$

Dado que $F_{exp} > F_{1,5; 0.05}$, se concluye que *los datos observados aportan evidencia para rechazar H_0* , lo que indica que puede suponerse relación lineal. De hecho, el p -nivel asociado a los datos es

$$p-nivel = P_{H_0}(F(\mathbf{Y}) > 29.57706) = 0.00285$$

lo que claramente conduce al rechazo de H_0 . Esta conclusión concuerda con el alto valor de r^2 , que indica un buen ajuste de los datos a la recta estimada.