

EL MODELO LINEAL

Clase 3

Sergio Camiz

LIMA - Abril-Mayo 2010

Asuntos de la clase 3

- La regresión lineal múltiple
- Estimación de los parámetros
- Propiedades estadísticas de los estimadores
- Inferencia
- Selección de modelos

La regresión lineal múltiple

- Se estudió hasta ahora el caso de la regresión lineal simple, o sea del modelo lineal simple

$$y = \alpha + \beta x = \varepsilon, \quad \forall x$$

- Ahora estudiaremos el modelo que expresa la relación de dependencia entre un carácter respuesta y *algunos* regresores, o sea

$$\eta = f(z_1, z_2, \dots, z_s; \theta_1, \theta_2, \dots, \theta_t)$$

con z_1, z_2, \dots, z_s regresores y $\theta_1, \theta_2, \dots, \theta_t$ parámetros. En forma vectorial

$$\eta = f(\mathbf{z}; \boldsymbol{\theta})$$

Así, el *modelo de regresión lineal múltiple* es un modelo en el cual los parámetros aparecen linealmente en la ecuación, que se vuelve

$$\eta = f(\mathbf{z}; \boldsymbol{\theta}) = \sum_{j=1}^k \beta_j x_j(\mathbf{z}) = \boldsymbol{\beta}' \mathbf{x}(\mathbf{z})$$

donde las componentes x_j del vector \mathbf{x} solo son función de las componentes z_h del vector \mathbf{z} *sin parámetros a estimar*.

Empezando de los regresores \mathbf{z} , hay libertad de transformarlos a condición que ningún parámetro a estimar entre en las transformaciones.

Los vectores $\boldsymbol{\beta}$ y \mathbf{x} tienen la misma dimensión k mientras la dimensión de \mathbf{z} puede ser cualquiera.

Estimación de los parámetros

- se empieza de n conjuntos de k valores $(x_{i1}, x_{i2}, \dots, x_{ik})$, $i = 1, 2, \dots, n$ de los k regresores (eventualmente transformando los $z_{i1}, z_{i2}, \dots, z_{ik}$).
- se observan valores y_i , $i = 1, 2, \dots, n$ en correspondencia de cada conjunto j .
- siempre se supone un error, así que resulta por cada i

$$y_i = \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i$$

como desvío del valor observado al modelo, o sea, $\varepsilon_i = y_i - \eta_i$

- se hace la hipótesis que los errores son errores experimentales aleatorios independientes y que por cada \mathbf{x}_i se tiene la misma distribución con media 0 y varianza σ^2 ; así que se puede escribir

$$\begin{cases} y_i = \eta_i + \varepsilon_i = \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i \\ E(y_i | \mathbf{x}_i) = \eta_i \\ V(y_i | \mathbf{x}_i) = \sigma^2 \\ y_i \text{ y } y_j \text{ independientes para cada } i \neq j \end{cases}$$

o, igualmente,

$$\begin{cases} y_i = \eta_i + \varepsilon_i \\ E(\varepsilon_i) = 0 \\ V(\varepsilon_i) = \sigma^2 \\ \varepsilon_i \text{ y } \varepsilon_j \text{ independientes para cada } i \neq j \end{cases}$$

- un error experimental siempre existe, pero esto puede depender de la influencia sobre el carácter y de otros *factores* que no están incluidos en los regresores \mathbf{x}_j , así que el modelo es un abreviado del modelo

$$y_i = \sum_{j=1}^k \beta_j x_{ij} + \sum_{l=1}^h \beta_l x_{il}$$

con otros regresores. Como los x_{il} son desconocidos, si h es grande el teorema del límite central asegura la normalidad.

Modelo geométrico

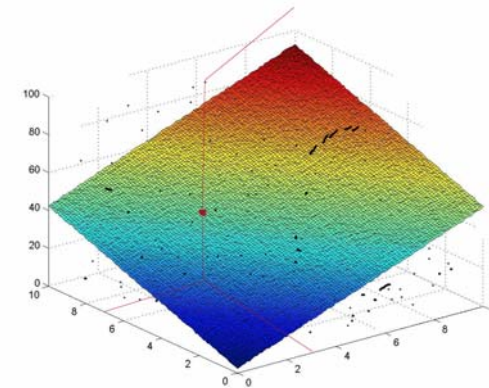
- Se quiere estimar los parámetros del modelo lineal.
- Se hicieron n observaciones, con $n \gg k$, del carácter respuesta y_i correspondiendo a valores prefijos de los regresores $x_{i1}, x_{i2}, \dots, x_{ik}$.
- Cada y_i en el espacio \mathbb{R}^n es una componente del vector respuesta \mathbf{y} .
- Los valores x_{ij} son componentes de k vectores \mathbf{x}_j de dimensión n , cada uno compuesto con los n valores observados por el j -ésimo factor;
- los valores x_{ij} son también componentes de n vectores \mathbf{x}'_i de dimensión k , conjuntos de valores de todos los regresores en la i -ésima observación.
- Se trata de una matriz X con n filas y k columnas.
- Los términos de error ε_i se pueden representar con un vector $\boldsymbol{\varepsilon}$ con n componentes.

En forma vectorial, el modelo se escribe

$$\begin{cases} \mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ E(\mathbf{y} | \mathbf{X}) = \boldsymbol{\eta} \text{ ou } E(\boldsymbol{\varepsilon}) = \mathbf{0} \\ V(\mathbf{y}) = V(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I} \end{cases}$$

- En la mayoría de los modelos es importante guardar un parámetro correspondiente a un término constante.
- A este parámetro, normalmente β_1 , se asocia en X el primero vector columna $\mathbf{x}_1 = (1, 1, \dots, 1)'$, así que en lugar de k regresores en realidad se tienen $k-1$.

Ejemplo de una respuesta y dos regresores: plan de regresión



- Las k columnas de la matriz X como vector del espacio \mathbb{R}^n generan un sub-espacio $S \subset \mathbb{R}^n$, cuya dimensión es a lo más $k < n$.
- A este espacio se lo llama *espacio solución, de regresión, de los parámetro, o de estimación*.
- Su dimensión es k solo si los regresores son *linealmente independientes* y en este caso se dice que el sistema es *de rango completo*.
- Resulta que el vector $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ es un vector del sub-espacio S , es decir que se quiere *estimar el vector \mathbf{y} en \mathbb{R}^n para un vector $\boldsymbol{\eta}$ de S* .
- Por esto se busca, en estas condiciones, *cual es la mejor estimación posible*, o sea cual es el mejor estimador de \mathbf{y} para un vector de S , como combinación lineal de vectores-columnas de X .

- Desde el punto de vista geométrico, el espacio \mathbb{R}^n es Euclidiano, o sea tiene un producto escalar entre vectores \mathbf{v} y \mathbf{w}

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_i v_i w_i$$

- una norma de los vectores o sea su longitud

$$\|\mathbf{v}\| = \langle \mathbf{v}, \mathbf{v} \rangle = \sum_i v_i v_i$$

- una distancia Euclidiana entre puntos dada de la norma del vector que junta los dos puntos. Si O es el origen y P, Q dos puntos, entonces

$$d(P, Q) = \sqrt{\|\vec{OP} - \vec{OQ}\|}$$

- Esto permite, dados \mathbf{v} , \mathbf{w} , de construir la *proyección ortogonal* de \mathbf{v} sobre \mathbf{w} como el vector colineal a \mathbf{w}

$$\rho_{\mathbf{w}} \mathbf{v} = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2} \mathbf{w}$$

- y su *complemento ortogonal*

$$\rho_{\perp \mathbf{w}} \mathbf{v} = (\mathbf{I} - \rho_{\mathbf{w}}) \mathbf{v} = \mathbf{v} - \rho_{\mathbf{w}} \mathbf{v} = \mathbf{v} - \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2} \mathbf{w}$$

- Resultan

$$\mathbf{v} = \rho_{\mathbf{w}} \mathbf{v} + \rho_{\perp \mathbf{w}} \mathbf{v} \quad \text{y} \quad \langle \rho_{\mathbf{w}} \mathbf{v}, \rho_{\perp \mathbf{w}} \mathbf{v} \rangle = 0$$

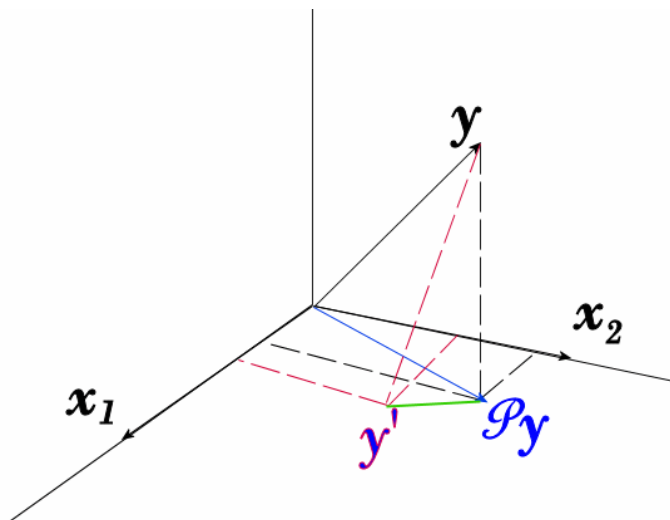
- Esta definición se basa sobre la definición de ortogonalidad:
A dos vectores \mathbf{v} , \mathbf{w} se llaman *ortogonales* si $\langle \mathbf{v}, \mathbf{w} \rangle = 0$.
- Dado un espacio S de dimensión k , siempre se puede construir una base compuesta por vectores ortogonales.
- Un vector \mathbf{v} es ortogonal a un subespacio S si el es ortogonal a todos los vectores de una base de S .
- por tanto (teorema de Pitágoras)

$$\mathbf{v} = \rho_S \mathbf{v} + \rho_{\perp S} \mathbf{v} \quad \text{y} \quad \langle \rho_S \mathbf{v}, \rho_{\perp S} \mathbf{v} \rangle = 0$$

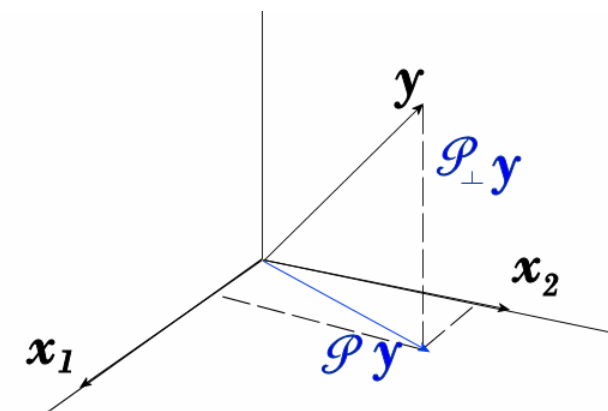
- y la distancia de un punto P a un subespacio S resulta

$$d(P, S) = \sqrt{\|\rho_{\perp S} \mathbf{OP}\|^2} = \min_{s \in S} d(P, s)$$

- En \mathbb{R}^3 por el teorema de Pitágoras la distancia $d(\mathbf{y}, \rho \mathbf{y})$ es mínima en cuanto $\mathbf{y}\mathbf{y}'$ es la hipotenusa de \mathbf{y} $\rho \mathbf{y}$ y \mathbf{y}' .



- La solución de mínimos cuadrados se encuentra para la proyección ortogonal de \mathbf{y} sobre el espacio S generado para $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$:



- **Teorema.** El punto $\hat{\boldsymbol{\eta}} = \boldsymbol{\rho}_S \mathbf{y}$ es el punto dentro de S más cercano a \mathbf{y} .
- Geométricamente la solución dada para la proyección **siempre existe y es única**.
- Esta se consigue empleando el *operador ortogonal de proyección* sobre S , $\boldsymbol{\rho}_S$.
- Como dado cualquier vector \mathbf{y} , el vector $\mathbf{y} - \boldsymbol{\rho}_S \mathbf{y}$, es ortogonal a cada vector de S ,
- cada vector de \mathbb{R}^n se encuentra compartido en dos componentes

$$\mathbf{y} = \boldsymbol{\rho}_S \mathbf{y} + (\mathbf{y} - \boldsymbol{\rho}_S \mathbf{y})$$
- el espacio también resulta compartido como *suma directa* de subespacios ortogonales:
- Si se escribe $\mathbf{y} - \boldsymbol{\rho}_S \mathbf{y} = (\mathbf{I} - \boldsymbol{\rho}_S) \mathbf{y} = \boldsymbol{\rho}_{S^\perp} \mathbf{y} = \mathcal{E}_S \mathbf{y}$ resulta por tanto

$$\mathbf{y} = \boldsymbol{\rho}_S \mathbf{y} \oplus (\mathbf{y} - \boldsymbol{\rho}_S \mathbf{y}) = \boldsymbol{\rho}_S \mathbf{y} \oplus \mathcal{E}_S \mathbf{y}.$$

- el espacio también resulta compartido como *suma directa* de subespacios ortogonales:

$$S \oplus S^\perp$$
- Cada proyector ortogonal es *idempotente*, o sea $\boldsymbol{\rho} \circ \boldsymbol{\rho} = \boldsymbol{\rho}$ y por tanto

$$\boldsymbol{\rho} \circ \mathcal{E} = \boldsymbol{\rho} \circ (\mathbf{I} - \boldsymbol{\rho}) = \boldsymbol{\rho} - \boldsymbol{\rho} \circ \boldsymbol{\rho} = \mathbf{0}.$$
- En el caso, $\mathbb{R}^n = S \oplus S^\perp$, donde S^\perp es el *complemento ortogonal de S* en \mathbb{R}^n , la proyección ortogonal de \mathbf{y} en S es $\boldsymbol{\rho}_S \mathbf{y} = \hat{\boldsymbol{\eta}} = \mathbf{X} \hat{\boldsymbol{\beta}}$, donde resulta

$$\mathbf{y} = \hat{\boldsymbol{\eta}} + \mathbf{e} \text{ con } \mathbf{y} - \hat{\boldsymbol{\eta}} = (\mathbf{I} - \boldsymbol{\rho}_S) \mathbf{y} = \mathbf{e} \in S^\perp.$$
- S^\perp se llama el *espacio de los errores* o de los *residuos*.
- **Teorema.** Un proyector ortogonal es representado para una matriz A *idempotente y simétrica*.

- utilizando el teorema de Pitágoras, resulta que el mínimo

$$\| \mathbf{e} \|^2 = \min (\mathbf{y} - \mathbf{s})'(\mathbf{y} - \mathbf{s}) = \min ((\mathbf{y} - \hat{\boldsymbol{\eta}})'(\mathbf{y} - \hat{\boldsymbol{\eta}}) + (\mathbf{s} - \hat{\boldsymbol{\eta}})'(\mathbf{s} - \hat{\boldsymbol{\eta}})) =$$

$$= \min ((\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}) + (\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\beta}}))'(\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\beta}})))$$
- se consigue cuando el segundo término es cero, o sea cuando $\boldsymbol{\theta} = \hat{\boldsymbol{\beta}}$.
- *Se puede calcular $\hat{\boldsymbol{\beta}}$ considerando que $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\eta}} \in S^\perp$, porque así \mathbf{e} es ortogonal a cada vector de S , donde desde*

$$\mathbf{X}' \mathbf{e} = \mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\eta}}) = (\mathbf{X}' \mathbf{y} - \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Resulta

$$\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{y}$$

que constituye el sistema de *ecuaciones normales*.

La solución del sistema (siempre existente)

$$\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{y}$$

simplemente sería

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$$

Técnicamente, para conseguir directamente la solución, hay que invertir la matriz $\mathbf{X}' \mathbf{X}$, como en el caso de rango completo. Pero esto no siempre es posible, dependiendo si entre los \mathbf{x}_j hay dependencia lineal (caso incompleto, rango $< k$).

Si hay dependencia, los parámetros no son unívocamente definidos, la matriz no es directamente invertible, pero se hace recurso a su *inversa generalizada*.

Normalmente, los programas no hacen distinción entre las dos situaciones porque fueron realizados de manera de tratar sin dificultad ambas condiciones.

En nuestro estudio, nos limitaremos al caso de rango completo, pero precisando que:

- 1) *la solución siempre existe y siempre es única en cualquier situación, porque siempre se trata de una proyección ortogonal;*
- 2) además su expresión algebraica en el caso de rango no completo puede no ser única, y por esto hay que utilizar técnicas específicas.

En el caso de rango completo, como $\dim S = k$, la matriz $X'X$ es invertible y por tanto la solución es

$$\beta = (X'X)^{-1}X'y$$

donde

$$\eta = X\beta = X(X'X)^{-1}X'y = \rho y$$

Es evidente que la matriz $\rho = X(X'X)^{-1}X'$ es simétrica e idempotente, y como η se encuentra en S , η es la proyección ortogonal de y sobre S .

En consecuencia, el vector $e = y - \eta = (I - \rho)y = \mathcal{E}y$ es la proyección ortogonal de y sobre el espacio de residuos S^\perp que entonces tiene dimensión $n - k$.

Se consiguen los resultados siguientes:

- β es el estimador de mínimos cuadrados de β , que minimiza la distancia entre y y su estimador en el espacio de los regresores S ;
- η Proyección ortogonal de y sobre S es el vector de los valores ajustados para la regresión: como proyección ortogonal de y sobre S , es el punto de S más cercano de y ;
- e proyección ortogonal de y sobre S^\perp es el vector de los residuos, ortogonal a η ;

$SS_r = \eta'\eta$ es la norma del vector η , que vale

$$\eta'\eta = \beta'X'X\beta = y'X(X'X)^{-1}X'X(X'X)^{-1}X'y = y'X(X'X)^{-1}X'y =$$

y que puede verse también como el producto escalar $y'\eta$;

$SS_e = e'e$ es la norma del vector e , y resulta

$$SS_e = e'e = y'\mathcal{E}'\mathcal{E}y = y'\mathcal{E}y$$

y que puede ver también como el producto escalar $y'e$.

Como ρ y \mathcal{E} son simétricos e idempotentes, resultan

$$r(\rho) = \text{tr}(\rho) = \text{tr}(X(X'X)^{-1}X) = \text{tr}(XX(X'X)^{-1}) = \text{tr}(I_k) = k$$

$$r(\mathcal{E}) = \text{tr}(\mathcal{E}) = \text{tr}(I_n - \rho) = \text{tr}(I_n) - \text{tr}(\rho) = n - k$$

Propiedades estadísticas de los estimadores

Sabiendo que $E(\mathbf{y}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, resulta todavía que

$$E(\hat{\boldsymbol{\beta}}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\eta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$E(\hat{\boldsymbol{\eta}}) = E(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}E(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$$

$$E(\mathbf{e}) = E(\mathbf{y} - \hat{\boldsymbol{\eta}}) = E(\mathbf{y}) - E(\hat{\boldsymbol{\eta}}) = \boldsymbol{\eta} - \boldsymbol{\eta} = \mathbf{0}$$

y, considerando que $V(\mathbf{y}) = \sigma^2 \mathbf{I}$ resulta

$$V(\hat{\boldsymbol{\beta}}) = V((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$V(\hat{\boldsymbol{\eta}}) = V(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}V(\hat{\boldsymbol{\beta}})\mathbf{X}' = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2\boldsymbol{\rho}$$

$$V(\mathbf{e}) = V(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = V(\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = \sigma^2(\mathbf{I}_n - \boldsymbol{\rho}) = \sigma^2\mathcal{E}$$

Hay que observar que los estimadores de los parámetros tienen una covarianza entre ellos, dependiendo de las relaciones lineales entre regresores.

En el caso múltiple, siempre vale

Teorema (Gauss). Dados n conjuntos de observaciones, dispuestas en forma de matriz (\mathbf{X}, \mathbf{y}) , cuyos X son valores previamente elegidos y los \mathbf{y} medidas correspondientes e independientes para las cuales $E(\mathbf{y} | \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, $V(\mathbf{y} | \mathbf{X}) = \sigma^2 \mathbf{I}$; sea $\hat{\boldsymbol{\beta}}$ la estimación de mínimos cuadrados de $\boldsymbol{\beta}$. Si $\boldsymbol{\tau} = \mathbf{a}'\boldsymbol{\beta}$, donde $\mathbf{a}' = (a_1, a_2, \dots, a_n)$ es un vector de constantes, entonces entre todos los estimadores *insesgados* y *lineal* en \mathbf{y} de $\boldsymbol{\tau}$, la estimación de mínimos cuadrados $\hat{\boldsymbol{\tau}} = \mathbf{a}'\hat{\boldsymbol{\beta}}$ es la de varianza mínima.

Por tanto, cada $\hat{\boldsymbol{\beta}}$, y $\hat{\boldsymbol{\eta}}$ entre todos los estimadores insesgados y lineales en \mathbf{y} , son los de varianza mínima.

Análisis de varianza del modelo

Una vez estimados los $\boldsymbol{\beta}$ resulta que el cuadrado de la distancia promedio entre \mathbf{y} y S es

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \hat{\boldsymbol{\eta}})'(\mathbf{y} - \hat{\boldsymbol{\eta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} \end{aligned}$$

pues, según las ecuaciones normales, $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})' = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$, de donde

$$\mathbf{y}'\mathbf{y} = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} + \mathbf{e}'\mathbf{e} = \mathbf{y}'\boldsymbol{\rho}\mathbf{y} + \mathbf{y}'\boldsymbol{\mathcal{E}}\mathbf{y}$$

Esto se puede escribir también

$$SS_t = SS_r + SS_e$$

Se ha compartido la suma de cuadrados de las observaciones en dos:

- una parte SS_r debido a la regresión de \mathbf{y} sobre X
- la otra, SS_e , debido al error.

Por tanto:

- $\hat{\boldsymbol{\eta}}$ contiene la información sobre el modelo $\boldsymbol{\eta} = X\boldsymbol{\beta}$,
- \mathbf{e} solo contiene la información sobre el error,
- $\mathbf{e}'\mathbf{e}$ debe informar sobre σ^2 .

Calculemos ahora SS_r y SS_e

$$E(SS_r) = E(\mathbf{y}'\boldsymbol{\rho}\mathbf{y}) = \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\rho}\mathbf{X}\boldsymbol{\beta} + \text{tr}(\boldsymbol{\rho}\sigma^2\mathbf{I}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + k\sigma^2$$

$$E(SS_e) = E(\mathbf{y}'\boldsymbol{\mathcal{E}}\mathbf{y}) = \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\mathcal{E}}\mathbf{X}\boldsymbol{\beta} + \text{tr}(\boldsymbol{\mathcal{E}}\sigma^2\mathbf{I}) = (n-k)\sigma^2$$

resultando el producto $\boldsymbol{\mathcal{E}}\mathbf{X} = \mathbf{0}$ y por tanto los cuadrados promedios

$$MS_r = SS_r/k \quad E(MS_r) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/k + \sigma^2$$

$$MS_e = SS_e/(n-k) \quad E(MS_e) = \sigma^2$$

o sea MS_e es un estimador insesgado de σ^2 .

Los elementos de esta partición se representan en una *tabla de análisis de varianza* como sigue, para testar $H_0: \boldsymbol{\beta} = \mathbf{0}$:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Regresión	k	SS_r	SS_r/k	$\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/k$
Error	$n - k$	SS_e	$SS_e/(n - k)$	σ^2
Total	n	SS_t		

Los grados de libertad son efectivamente las dimensiones de los espacios:

- el espacio de los estimadores S tiene k dimensiones,
- el espacio de los residuos tiene dimensión $n - k$.

Como por la regresión simple, para testar la hipótesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, se hace una translación del origen de $\mathbf{0}$ a $\boldsymbol{\eta}_0 = X\boldsymbol{\beta}_0$ con la tabla de análisis de varianza correspondiente:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Regresión	k	$SS_{r(z)} = \hat{\boldsymbol{\phi}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\phi}}$	$SS_{r(z)}/k$	$\sigma^2 + \hat{\boldsymbol{\phi}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\phi}}/k$
Error	$n - k - 1$	$SS_{e(z)} = SS_e$	$SS_e/(n - k - 1)$	σ^2
Total	$n - 1$	$SS_{l(z)}$		

Inferencia

Para hacer inferencia, es necesario imponer una hipótesis sobre la distribución de \mathbf{y} con respecto a X .

Si se supone que la distribución sea multinormal, o sea normal por cada valor de X , se aplican los resultados que siguen, sino hay que conocer resultados parecidos por la distribución que interesa o se hace recurso a método de remuestreo (Manly, 2007).

Se analiza aquí los resultados más importantes relativos a las muestras de una distribución normal multivariada, que servirán a los test estadísticos. Para una lectura más detallada y referencias se puede ver Guttman (1982).

Definición. Se dice que el vector aleatorio $\mathbf{y} \in \mathbb{R}^n$, tiene una distribución normal si su función de densidad $p_{\mathbf{y}}(\mathbf{y})$ es

$$p_{\mathbf{y}}(\mathbf{y}) = \sqrt{\frac{|\Sigma^{-1}|}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

donde la matriz Σ^{-1} es simétrica definida positiva y $\boldsymbol{\mu}$ tiene componentes finitas. Entonces se escribe $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$, con $E(\mathbf{y}) = \boldsymbol{\mu}$ y $V(\mathbf{y}) = \Sigma$.

Teorema. Si $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y Σ es diagonal, su componentes y_i de \mathbf{y} son estadísticamente independientes.

Teorema. Sea un vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$, con $\Sigma = P'P$ y Q y considere la forma cuadrática centrada

$$Q = (\mathbf{y}-\boldsymbol{\mu})'G(\mathbf{y}-\boldsymbol{\mu})$$

donde la matriz G es simétrica y real. Entonces la ley de distribución de Q es una combinación lineal de n variables aleatorias independientes de ley chi-cuadrado con 1 grado de libertad

$$Q = \sum_i \lambda_i \chi_1^2(i)$$

donde los λ_i son los autovalores de $P'GP$ (y también de ΣG y de $G\Sigma$).

Teorema. Una condición necesaria y suficiente por que Q tenga una ley de distribución de chi-cuadrado con $k < n$ grados de libertad es que $P'GP$ sea idempotente y de rango k . Si $\Sigma = \sigma^2 I$ la condición deviene en que G sea idempotente de rango k .

Teorema de Craig. Sea un vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y las dos formas cuadráticas

$$Q_i = (\mathbf{y}-\boldsymbol{\mu})'G_i(\mathbf{y}-\boldsymbol{\mu}), \quad i=1,2$$

con G_i reales y simétricas. Entonces Q_1 y Q_2 son estadísticamente independientes si y solo si $G_1 \Sigma G_2 = 0$.

Teorema de Cochran. Sea $\mathbf{y} = N(0, I)$ una variable aleatoria y sean n observaciones de \mathbf{y} independientes, formando un vector aleatorio $\mathbf{y} = N(\mathbf{0}, I)$. Sea por otro lado

$$Q = \mathbf{y}'\mathbf{y} = Q_1 + Q_2 + \dots + Q_k$$

donde $Q_i = \mathbf{y}'A_i\mathbf{y}$ es una forma cuadrática de rango $rg(A_i) = n_i$, y A_i es una matriz simétrica $n \times n$, $i = 1, \dots, k$. Entonces las siguientes condiciones son equivalentes:

- 1) Q_1, Q_2, \dots, Q_n son estadísticamente independientes;
- 2) Q_1, Q_2, \dots, Q_n tienen individualmente distribuciones de chi-cuadrado;
- 3) $n_1 + n_2 + \dots + n_k = n$.

Teorema. Sea el vector aleatorio $\mathbf{y} = N(\boldsymbol{\mu}, \Sigma)$ y considere su distribución de \mathbf{y} condicional

$$A(\mathbf{y} - \boldsymbol{\mu}) = 0$$

donde A es una matriz $k \times n$, $k < n$, $rg(A) = k$. Entonces la distribución condicional de \mathbf{y} es tal que

$$Q_i = (\mathbf{y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \chi_{n-k}^2$$

Test del modelo y de los parámetros

Se supone que el vector \mathbf{y} tiene

- en cada punto de medida una distribución normal centrada sobre su esperanza $\boldsymbol{\eta}$
- y de varianza constante, $\mathbf{y} = N(\boldsymbol{\eta}, \sigma^2 I)$,
- bajo la hipótesis nula $H_0: \boldsymbol{\beta} = \mathbf{0}$ se tiene
- $E(\mathbf{y}) = \mathbf{0}$,
- y por tanto $\mathbf{y} = N(\mathbf{0}, \sigma^2 I)$.
- Resulta del teorema de Craig que SS_r y SS_e son estadísticamente independientes, pues son formas cuadráticas de una distribución normal centrada y reducida, de matrices ρ y \mathcal{E} tales que $\sigma^2 \rho \mathcal{E} = 0$.

En consecuencia resulta

$$SS_r = \sigma^2 \chi_k^2 \quad SS_e = \sigma^2 \chi_{n-k}^2$$

con las leyes χ^2 independientes, donde

$$\text{bajo } H_0: \boldsymbol{\beta} = \mathbf{0}, \quad F = \frac{MS_r}{MS_e} = F_{k, n-k}$$

Por tanto, fijado un nivel de probabilidad π el test resulta ser

$$\text{no aceptar } H_0: \boldsymbol{\beta} = \mathbf{0}, \text{ si } \frac{MS_r}{MS_e} > F_{k, n-k; \pi}$$

aceptar en otro caso.

De otro lado, es fácil de ver que en

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

el primero término a la derecha vale $\mathbf{e}'\mathbf{e}$ y se puede escribir

$$\frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \boldsymbol{\rho} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' \boldsymbol{\mathcal{Z}} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Entonces resulta
$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) / k}{\mathbf{e}' \mathbf{e} / (n - k)} = \frac{\sigma^2 \chi^2_{k/k}}{\sigma^2 \chi^2_{n-k} / (n - k)} = F_{k, n-k}$$

donde se construye la región de confianza de $\boldsymbol{\beta}$ al nivel de $1 - \pi$

$$C_{1-\pi} = \left\{ \boldsymbol{\beta} \mid (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq k MS_e F_{k, n-k; \pi} \right\}$$

$\hat{\boldsymbol{\beta}}$ es distribuido normalmente con promedio $\boldsymbol{\beta}$ y varianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, por tanto cada $\hat{\beta}_i$ tiene una distribución normal $\hat{\beta}_i = N(\beta_i, \sigma^2 c_{ii})$ donde c_{ii} es elemento diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$. Entonces su *desvío estándar* es

$$SD(\hat{\beta}_i) = \sqrt{MS_e c_{ii}}$$

el test **no aceptar** $H_0: \beta_i = \beta_i$, si $\left| \frac{\hat{\beta}_i - \beta_i}{SD(\hat{\beta}_i)} \right| > t_{n-k; \pi/2}$

aceptar en otro caso.

y el intervalo de confianza (a tratar con cuidado)

$$C_{1-\pi} = \left\{ \beta_i \mid \hat{\beta}_i - SD(\hat{\beta}_i) t_{n-k; \pi/2} \leq \beta_i \leq \hat{\beta}_i + SD(\hat{\beta}_i) t_{n-k; \pi/2} \right\}$$

Para el tema de la varianza, su intervalo de confianza es dado por

$$C_{1-\pi} = \left\{ \sigma^2 \mid \frac{SS_e}{\chi^2_{n-k; \pi/2}} \leq \sigma^2 \leq \frac{SS_e}{\chi^2_{n-k; 1-\pi/2}} \right\}$$

Resulta $V(\hat{\boldsymbol{\eta}}) = V(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}V(\hat{\boldsymbol{\beta}})\mathbf{X}' = \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$,

entonces por la estimación de $\boldsymbol{\eta}_0 = \mathbf{x}_0' \boldsymbol{\beta}$, resulta su desvío

estándar $SD(\hat{\boldsymbol{\eta}}_0) = \sqrt{MS_e \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0}$

se deriva la condición del test

$$\left| \frac{\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0}{SD(\hat{\boldsymbol{\eta}})} \right| > t_{n-k; \pi/2}$$

con intervalo de confianza

$$\left\{ \boldsymbol{\eta} \mid \hat{\boldsymbol{\eta}} - SD(\hat{\boldsymbol{\eta}}) t_{n-k; \pi/2} \leq \boldsymbol{\eta} \leq \hat{\boldsymbol{\eta}} + SD(\hat{\boldsymbol{\eta}}) t_{n-k; \pi/2} \right\}$$

De manera análoga resulta que el desvío estándar del predictor \tilde{y}_0 vale

$$SD(\tilde{y}_0) = \sqrt{MS_e (1 + \mathbf{x}_0' (X'X)^{-1} \mathbf{x}_0)},$$

y por tanto su intervalo de confianza resulta

$$\{y \mid \tilde{y}_0 - SD(\tilde{y}_0) t_{n-k; \pi/2} \leq y \leq \tilde{y}_0 + SD(\tilde{y}_0) t_{n-k; \pi/2}\}$$

Partición de la regresión

En un estudio, puede resultar que el interés este concentrado solo sobre algunos parámetros. Esto significa que se consideran dos conjuntos de $k_1 + k_2 = k$ regresores separados, X_1 y X_2 , y por esto el modelo puede escribir como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

donde $\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{X}_2)$, $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1 \mid \boldsymbol{\beta}'_2)$. Bajo esta partición la matriz $X'X$ toma la forma

$$X'X = \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} (X_1 \mid X_2) = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}$$

Si solo se esta interesado en los $\boldsymbol{\beta}_2$, los $\boldsymbol{\beta}_1$ se llaman parámetros de fastidio. Tiene que distinguir en el análisis dos casos: si los dos conjuntos de regresores son ortogonales o sea $X_1'X_2 = 0$, o no.

En el caso no ortogonal, hay que *ortogonalizar* X_2 . Como se puede escribir $X_2 = X_{21} + X_{2\circ 1}$, con las columnas de X_{21} regresión de la columna correspondiente sobre X_1 y la matriz $X_{2\circ 1}$ formada por los residuos, se estudiará el modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_{2\circ 1}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

En el caso ortogonal bajo la condición $X_1'X_2 = 0$ la matriz $X'X$ toma la

forma $\begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}$ y entonces

$$[X'X]^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix}$$

En consecuencia la estimación de $\boldsymbol{\beta}$ se puede dividir en dos partes independientes, por lo que resulta

$$\beta = (X'X)^{-1}X'y = \begin{pmatrix} (X_1'X_1)^{-1}X_1'y \\ (X_2'X_2)^{-1}X_2'y \end{pmatrix}$$

con varianza

$$V(\beta) = V \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix}$$

De aquí resulta que la covarianza entre β_1 y β_2 es cero.

Por otro lado se tiene

$$\begin{aligned} \rho_S &= X(X'X)^{-1}X' = (X_1|X_2) \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} = \\ &= X_1(X_1'X_1)^{-1}X_1' + X_2(X_2'X_2)^{-1}X_2' = \rho_{S_1} + \rho_{S_2} \end{aligned}$$

Entonces, la proyección sobre S es la suma de las dos proyecciones sobre los espacios S_1 y S_2 generados respectivamente para X_1 y X_2 . En consecuencia

$$\hat{\eta} = \rho_S y = (\rho_{S_1} + \rho_{S_2})y = \rho_{S_1}y + \rho_{S_2}y = \rho_{S_1}\hat{\eta} + \rho_{S_2}\hat{\eta} = \hat{\eta}_1 + \hat{\eta}_2$$

y como las proyecciones son simétricas e idempotentes,

$$\hat{\eta}'\hat{\eta} = y'\rho_S y = y'\rho_{S_1}y + y'\rho_{S_2}y = y'\rho_{S_1}'\rho_{S_1}y + y'\rho_{S_2}'\rho_{S_2}y = \hat{\eta}_1'\hat{\eta}_1 + \hat{\eta}_2'\hat{\eta}_2$$

Por los residuos resulta

$$e'e = y'y - \hat{\eta}'\hat{\eta} = y'y - \hat{\eta}_1'\hat{\eta}_1 - \hat{\eta}_2'\hat{\eta}_2 = y'(I - \rho_{S_1} - \rho_{S_2})y$$

Así se pueden organizar los resultados en la tabla de análisis de varianza siguiente:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
X_1	k_1	$SS_{X_1} = y'\rho_{S_1}y$	SS_{X_1} / k_1	$\sigma^2 + \beta_1'X_1'X_1\beta_1/k_1$
X_2	k_2	$SS_{X_2} = y'\rho_{S_2}y$	SS_{X_2} / k_2	$\sigma^2 + \beta_2'X_2'X_2\beta_2/k_2$
Error	$n - k_1 - k_2$	$SS_e = y'(I - \rho_{S_1} - \rho_{S_2})y$	$SS_e / (n - k_1 - k_2)$	σ^2
Total	n	SS_t		

Si el interés solo está concentrado sobre β_2 se puede borrar la primera línea y cambiar la última, de manera que se obtiene:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
X_2	k_2	$SS_{X_2} = \mathbf{y}'\rho_{S_2}\mathbf{y}$	SS_{X_2} / k_2	$\sigma^2 + \beta_2' X_2' X_2 \beta_2 / k_2$
Error	$n - k_1 - k_2$	$SS_e = \mathbf{y}'(I - \rho_{S_1} - \rho_{S_2})\mathbf{y}$	$SS_e / (n - k_1 - k_2)$	σ^2
Total	$n - k_1$	$\mathbf{y}'(I - \rho_{S_1})\mathbf{y}$		

Esta tabla muestra que cuando el interés está limitado a β_2 , la suma de cuadrados total es la suma de cuadrados de los residuos de la regresión de \mathbf{y} sobre X_1 .

Esto sugiere proceder en dos etapas:

- 1) Calcular la regresión de \mathbf{y} sobre X_1
- 2) Calcular la regresión de \mathbf{e}_1 sobre X_2 .

Se puede observar que, cuando se incluyó X_2 en el modelo la suma de cuadrados de la regresión aumentó y la suma de cuadrados de los residuos disminuyó en la misma cantidad, así que siempre se puede escribir

$$SS_r(\beta_2) = SS_e(\beta_1) - SS_e(\beta_1, \beta_2)$$

$$SS_r(\beta_2) = SS_r(\beta_1, \beta_2) - SS_r(\beta_1)$$

La eliminación del promedio

Ocurre a menudo que el modelo más apropiado es de la forma

$$\mathbf{E}(\mathbf{y}) = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \dots + \beta_{k-1} \mathbf{x}_{k-1}$$

donde $\mathbf{1} = (1, 1, \dots, 1)'$ tiene dimensión $n \times 1$, y donde el interés se centra sobre los parámetros $\beta_1, \beta_2, \beta_{k-1}$. Se puede observar que, si estos no son todos cero, resulta $\beta_0 = \bar{y}$ el promedio de los y_i , que no es de interés, mientras que interesan los otros factores. Se vuelve así al modelo

$$\mathbf{E}(\mathbf{y}) = (\mathbf{1}'\mathbf{X}_1) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

donde $\beta_1 = (\beta_1, \beta_2, \dots, \beta_{k-1})$ es de dimensión $k_1 = k - 1$, donde queremos estimar los parámetros y la suma de cuadrados de los errores, conociendo la varianza de \mathbf{y} , o sea $\sigma^2 I$. Se debe entonces ortogonalizar previamente los X_1 por respecto al vector $\mathbf{1}$, lo que corresponde a haber centrado cada regresor alrededor de su promedio.

Entonces es claro que

$$\mathbf{1}'\mathbf{X}_{1\cdot 0} = (\mathbf{1}'(\mathbf{x}_i - \bar{x}_i \mathbf{1}))'_{i=1,2,\dots,n} = \mathbf{0}$$

como vector de desvíos al promedio.

Entonces se puede estimar el modelo

$$\begin{cases} \mathbf{y} &= \mathbf{x}_0' \boldsymbol{\phi} + \mathbf{X}_{1,0} \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ V(\boldsymbol{\varepsilon}) &= \sigma^2 \mathbf{I} \end{cases}$$

Resulta que $\bar{y} = \boldsymbol{\phi} + \bar{\boldsymbol{\varepsilon}}$

con $E(\bar{y}) = \boldsymbol{\phi}$

pero también $E(\hat{\boldsymbol{\phi}}) = \boldsymbol{\phi} = \bar{y}$

Resulta también que $\boldsymbol{\phi} = \boldsymbol{\beta}_0 + \bar{\mathbf{x}}' \boldsymbol{\beta}_1$ y que

$$\begin{pmatrix} \hat{\boldsymbol{\phi}} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = \begin{pmatrix} (\mathbf{1}'\mathbf{1})^{-1} & \mathbf{0} \\ \mathbf{0} & |(\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1}| \end{pmatrix} \begin{pmatrix} \mathbf{1}' \\ \mathbf{X}'_{1,0} \end{pmatrix} \mathbf{y} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1} \mathbf{X}'_{1,0} \mathbf{y} \end{pmatrix}$$

Sobre la base del teorema de Gauss, $\hat{\boldsymbol{\phi}}$ es un estimador insesgado de varianza mínima lineal en \mathbf{y} . Se verifica fácilmente que las varianzas de los parámetros valen

$$V \begin{pmatrix} \hat{\boldsymbol{\phi}} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & |(\mathbf{X}'_{1,0}\mathbf{X}_{1,0})^{-1}| \end{pmatrix}$$

así que $\hat{\boldsymbol{\phi}}$ y $\hat{\boldsymbol{\beta}}_1$ son no correlacionados entre ellos.

Se sigue el proceso haciendo la regresión de \mathbf{y} sobre $\mathbf{1}$ para calcular los residuos. Esto nos deja

$$\mathbf{e}_0 = \boldsymbol{\mathcal{E}}_1 \mathbf{y} = (\mathbf{I} - \mathbf{n}^{-1} \mathbf{1} \mathbf{1}') \mathbf{y} = \mathbf{y} - \bar{y} \mathbf{1}$$

correspondiendo a centrar \mathbf{y} , o sea *eliminar el promedio*. Por tanto se hace la regresión de \mathbf{e}_0 sobre $\mathbf{X}_{1,0}$ para obtener $\hat{\boldsymbol{\beta}}_1$. Resulta entonces la tabla de análisis de varianza

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios E(MS)
$\mathbf{X}_{1,0}$	$k - 1$	$SS_{\mathbf{X}_{1,0}} = \boldsymbol{\beta}'_1 \mathbf{X}'_{1,0} \mathbf{X}_{1,0} \boldsymbol{\beta}_1$	$SS_{\mathbf{X}_{1,0}} / (k-1)$	$\sigma^2 + \boldsymbol{\beta}'_1 \mathbf{X}'_{1,0} \mathbf{X}_{1,0} \boldsymbol{\beta}_1 / (k-1)$
Error	$n - k$	$SS_e = \mathbf{e}'_0 \mathbf{e}_0 = \mathbf{y}' \mathbf{Q}_{\mathbf{1},0} \mathbf{y}$	$SS_e / (n - k)$	σ^2
Total	$n - 1$	$\mathbf{e}'_0 \mathbf{e}_0 = \sum (y_i - \bar{y})^2$		

Siguiendo esta tabla se puede definir el coeficiente de determinación de la misma manera que en la regresión simple:

$$R^2 = \frac{SS_{S_{1.0}}}{e_0' e_0} = 1 - \frac{SS_e}{e_0' e_0}$$

Se puede mostrar que su raíz cuadrada es el *coeficiente de correlación múltiple* entre \mathbf{y} y X_1 , o sea entre \mathbf{y} y $\boldsymbol{\eta}$:

$$R = \frac{\sum (y_i - \bar{y})(\hat{\eta}_i - \bar{\eta})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{\eta}_i - \bar{\eta})^2}}$$

La falta de ajuste del modelo lineal

Análogamente que en el caso simple, en el caso múltiple se hizo la hipótesis de saber que la relación entre X y \mathbf{y} era lineal o era una buena aproximación lineal. Sin embargo hay situaciones donde esto tiene que comprobarse.

Como siempre se empieza con el hecho que $E(\mathbf{y} | \mathbf{X}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ y la varianza de los \mathbf{y} es $V(\mathbf{y}) = \sigma^2 I$. Se sabe que si el modelo ajusta bien a los datos, el cuadrado medio de los errores MS_e es un estimador insesgado de esta varianza.

Supongamos entonces que el modelo no ajuste bien, o sea que

$$E(\mathbf{y} | \mathbf{X}) = \boldsymbol{\gamma} \neq \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

En este caso se puede proyectar $\boldsymbol{\gamma}$ sobre X y se consigue su proyección

$$\boldsymbol{\gamma}^\circ = \boldsymbol{\rho}\boldsymbol{\gamma}$$

y el vector

$$\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ = \boldsymbol{\mathcal{E}}\boldsymbol{\gamma}$$

que se puede llamar *vector residual del modelo*. Este informa sobre el desvío entre la esperanza verdadera y la supuesta. Su cuadrado es

$$\Lambda^2 = (\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ)'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ)$$

así que, si $\boldsymbol{\gamma} = \boldsymbol{\eta}$, entonces $\Lambda^2 = 0$.

Análogamente al caso lineal se llega a la tabla de análisis de varianza

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados medios (MS)	Esperanza de los cuadrados medios $E(MS)$
Inter	k	$\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \boldsymbol{\gamma}'\boldsymbol{\rho}\boldsymbol{\gamma}$		$\sigma^2 + \boldsymbol{\gamma}'\boldsymbol{\rho}'\boldsymbol{\gamma}^\circ$
Inter falta de ajuste	$m-k$	$SS_M = \sum_j n_j (\bar{y}_j - \hat{\eta}_j)^2$	$MS_M = SS_M / (m-k)$	$\sigma^2 + \Lambda^2 / (m-k)$
Intra	$n - m$	$SS_W = \sum_{j=1}^m \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	$MS_W = SS_W / (n - m)$	σ^2
Total	n	$SS_T = \boldsymbol{y}'\boldsymbol{y}$		

Para testar el ajuste del modelo, se puede rechazar la hipótesis de linealidad a nivel de probabilidad π si

$$F_M = \frac{MS_M}{MS_W} > F_{m-k, n-m; \pi}$$

y aceptarla en caso contrario.

Selección de modelos

Para seleccionar entre modelos diferentes, resulta conveniente utilizar matrices que sintetizan las relaciones entre el conjunto de variables.

Ya sabemos que el coeficiente de correlación $r(\mathbf{x}, \mathbf{y})$, resultando del cálculo de los residuos de una regresión lineal, informa sobre la intensidad de la relación lineal entre las variables mismas.

En particular, resulta cero en el caso que ninguna parte de SS_y sea explicada para una recta de regresión sobre \mathbf{x} y ± 1 en el caso de relación funcional perfecta positiva o negativa.

Resulta

$$r = \text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}(\mathbf{x})\text{var}(\mathbf{y})}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

En el caso de una tabla de datos X con p variables en columna,, se usa sintetizar este conjunto a través de matrizes (simétricas y semi-definidas positivas). Específicamente, se introducen la matriz de varianza-covarianza

$$V(X) = \begin{pmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \dots & \text{cov}(x_1, x_p) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \dots & \text{cov}(x_2, x_p) \\ \dots & \dots & \dots & \dots \\ \text{cov}(x_p, x_1) & \text{cov}(x_p, x_2) & \dots & \text{var}(x_p) \end{pmatrix}$$

y la matriz de correlación

$$C(\mathbf{X}) = \begin{pmatrix} 1 & \text{corr}(\mathbf{x}_1, \mathbf{x}_2) & \dots & \text{corr}(\mathbf{x}_1, \mathbf{x}_p) \\ \text{corr}(\mathbf{x}_2, \mathbf{x}_1) & 1 & \dots & \text{corr}(\mathbf{x}_2, \mathbf{x}_p) \\ \dots & \dots & \dots & \dots \\ \text{corr}(\mathbf{x}_p, \mathbf{x}_1) & \text{corr}(\mathbf{x}_p, \mathbf{x}_2) & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

A veces puede ser útil saber como se pueden calcular de manera sintética dichas matrices.

Sabiendo que el producto de la tabla de datos con su traspuesta vale

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} S_{x_1x_1} & S_{x_1x_2} & \dots & S_{x_1x_p} \\ S_{x_2x_1} & S_{x_2x_2} & \dots & S_{x_2x_p} \\ \dots & \dots & \dots & \dots \\ S_{x_px_1} & S_{x_px_2} & \dots & S_{x_px_p} \end{pmatrix}$$

la matriz de varianza-covarianza resulta del centrado de los datos alrededor de su promedio dividido por $n - 1$, o sea, definiendo el vector $\mathbf{S}' = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$ de los promedios

$$V(\mathbf{X}) = \frac{1}{n-1} \begin{pmatrix} S_{\hat{x}_1\hat{x}_1} & S_{\hat{x}_1\hat{x}_2} & \dots & S_{\hat{x}_1\hat{x}_p} \\ S_{\hat{x}_2\hat{x}_1} & S_{\hat{x}_2\hat{x}_2} & \dots & S_{\hat{x}_2\hat{x}_p} \\ \dots & \dots & \dots & \dots \\ S_{\hat{x}_p\hat{x}_1} & S_{\hat{x}_p\hat{x}_2} & \dots & S_{\hat{x}_p\hat{x}_p} \end{pmatrix} = \frac{1}{n-1} (\mathbf{X}'\mathbf{X} - n\mathbf{S}\mathbf{S}')$$

Entonces, si se define la matriz diagonal de los desvíos estándar $\sigma(\mathbf{X}) = \text{diag}(\sigma_{x_1}, \sigma_{x_2}, \dots, \sigma_{x_p})$, resulta

$$C(\mathbf{X}) = \sigma(\mathbf{X})^{-1} V(\mathbf{X}) \sigma(\mathbf{X})^{-1} \quad (8)$$

El coeficiente de correlación parcial

Supongamos ahora que dos variables aleatorias \mathbf{x}_1 y \mathbf{x}_2 dependen linealmente de una misma variable aleatoria \mathbf{z} . Es posible de medir directamente el coeficiente de correlación $r = \text{corr}(\mathbf{x}_1, \mathbf{x}_2)$ sobre una muestra de tamaño n representada en \mathbb{R}^n para los vectores con n componentes \mathbf{x}_1 y \mathbf{x}_2 , pero queremos conocer el vínculo entre \mathbf{x}_1 y \mathbf{x}_2 eliminando el efecto de la variable \mathbf{z} cuyas n observaciones son las componentes del vector \mathbf{z} .

Existe una estadística que permite calcular este vínculo entre x_1 y x_2 bajo z constante de manera sencilla, también con muestras pequeñas, bajo una hipótesis de linealidad de los vínculos. Se trata del *coeficiente de correlación parcial* entre x_1 y x_2 , que se escribe como:

$$r_{x_1 x_2 \cdot z} = \rho(x_1, x_2 | z)$$

Suponiendo a todas las variables centradas, su cálculo se basa sobre la hipótesis que el efecto de z sobre x_1 y x_2 se manifiesta para relaciones del

tipo:

$$\begin{cases} x_1 = \beta_1 z + \varepsilon_1 \\ x_2 = \beta_2 z + \varepsilon_2 \end{cases}$$

donde ε_1 y ε_2 son los residuos. Entonces, en función de los residuos,

$$\begin{cases} \varepsilon_1 = x_1 - \beta_1 z \\ \varepsilon_2 = x_2 - \beta_2 z \end{cases}$$

y se define el coeficiente de *correlación parcial* entre y_1 y y_2 como el coeficiente de correlación entre ε_1 y ε_2 :

$$r_{x_1 x_2 \cdot z} = \text{corr}(\varepsilon_1, \varepsilon_2)$$

$r_{x_1 x_2 \cdot z}$ se puede escribir también haciendo aparecer los coeficientes de

correlación usual, $r_{x_1, x_2} = \frac{e_1' e_2}{\sqrt{(e_1' e_1)(e_2' e_2)}}$

desde que se obtiene $r_{x_1 x_2 \cdot z} = \frac{(e_1' e_2)/n}{\sqrt{((e_1' e_1)/n)((e_2' e_2)/n)}}$

y por tanto

$$r_{x_1 x_2 \cdot z} = \frac{r_{x_1 x_2} - r_{x_1 z} r_{x_2 z}}{\sqrt{(1 - r_{x_1 z}^2)(1 - r_{x_2 z}^2)}}$$

Caso general

Análogamente se construyen las matrices de varianza-covarianza parcial $V(X|Z)$ y de correlación parcial $C(X|Z)$ entre p variables x_1, x_2, \dots, x_p suponiendo fijas q variables z_1, z_2, \dots, z_q . Por esto resulta el sistema

$$\begin{cases} x_1 = a_{11} z_1 + a_{12} z_2 + \dots + a_{1q} z_q + \varepsilon_1 \\ x_2 = a_{21} z_1 + a_{22} z_2 + \dots + a_{2q} z_q + \varepsilon_2 \\ \dots \\ x_p = a_{p1} z_1 + a_{p2} z_2 + \dots + a_{pq} z_q + \varepsilon_p \end{cases}$$

Resulta

$$V(X|Z) = V_{XX} - V_{ZX}V_{ZZ}^{-1}V_{XZ}$$

La matriz de correlaciones parciales se calcula fácilmente empezando con la de $V(X|Z)$ como se calculó la matriz de correlación ordinaria:

tomando $\sigma(\mathbf{x}|Z) = \sqrt{\text{var}(\mathbf{x}|Z)}$ se obtiene

$$C(X|Z) = \sigma(X|Z)V(X|Z)\sigma(X|Z)$$

Técnicas de regresión

Supongamos que para modelar \mathbf{y} , *variable que explicar o criterio*, se dispone de p predictores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$.

En vez de explicar \mathbf{y} con todas las p variables explicativas, se puede intentar de explicar \mathbf{y} solo con un sub-conjunto de q variables extraídas de las p disponibles, de manera de conseguir una explicación casi igualmente satisfactoria de \mathbf{y} .

Existen muchas razones para esta operación:

- reducir el numero de predictores,
- seleccionarlos entre un numero demasiado grande,
- obtener fórmulas más estables pero con buena capacidad de predicción.

Considerando también que

- aumentando los predictores el coeficiente de correlación múltiple siempre aumenta,
- también aumenta la varianza de los estimadores;
- aumentando los predictores aumenta el riesgo de colinealidad y por consecuencia la inestabilidad de los parámetros.

El registro exhaustivo (all possible regression)

Es un método que consiste en estudiar todas las posibles regresiones. Considerando que el término de intersección β_0 se encuentra en todas las ecuaciones, resultan 2^p regresiones a comparar.

Pero, ¿Cómo seleccionar el mejor modelo?

Aumentando los regresores, siempre aumenta R^2 así que probablemente R^2 no sirve.

$$R^2 \text{ ajustado: } R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1} = 1 - \frac{SS_e df_t}{SS_t df_e}$$

buscando el máximo

Un criterio de información de Akaike (1974: *an information criterion*)

$$AIC = 2p + n(\log(2\pi SS_r/n) + 1)$$

buscando el mínimo

$$C_p \text{ de Mallows (1964) } C_p = \frac{SS_E(p)}{\sigma^2} - n + 2p$$

buscando el mínimo.

Método paso a paso: selección para adelante

En lugar de seleccionar entre todas las regresiones posibles, hay métodos que automáticamente construyen un modelo *paso a paso*, claramente sub-optimal, a través de una secuencia de regresiones ligadas, obtenidas juntando o quitando un predictor a cada paso.

El método de selección para adelante consiste en juntar a cada paso una variable en el modelo, en base a su capacidad predictiva.

Claro que no conviene juntar variables que no aumentan la calidad del modelo de manera significativa.

Para esto se utiliza la estadística *F de entrada*

$$F_E = \frac{SS_R(x_s | x_1, \dots, x_{s-1})}{MS_E(x_1, \dots, x_s)} > F_{\pi, 1, n-s+1}$$

como el ratio entre la suma de cuadrados de la regresión de los residuos con una nueva variable x_s y el promedio de cuadrados de los errores y se testa con la *F* de Fisher a nivel π , con 1 y $n - s - 1$ grados de libertad.

El proceso es el siguiente:

- 1) se define previamente un valor de probabilidad π (.10, .05, ...)
- 2) se estima el término constante β_0
- 3) se selecciona la variable cuyo coeficiente de correlación con y sea más grande en valor absoluto, digamos x_1 , a condición que su *F* de entrada sea significativa
- 4) se selecciona como variable siguiente en el modelo la variable cuya correlación parcial con y , bajo los efectos conocidos de los predictores ya en el modelo, es más grande, a condición que su *F* de entrada sea significativa
- 5) Se repite 4) hasta que no hay más variables con *F* de entrada significativa.

La selección para atrás

La selección para adelante tiene el defecto que alguna variable, aunque importante para describir el fenómeno del punto de vista causal, podría no resultar incluida en el modelo final, debido a la presencia de otros predictores que *cobren* influencia, así que resulta una F_E no significativa por esta.

Por esta razón algunos consideran mejor la *selección para atrás*, porque con este criterio se puede averiguar que ningún predictor importante sea olvidado.

La selección *para atrás* busca un buen modelo empezando con todas las p variables y quitando un predictor paso a paso.

Para elegir el predictor a quitar, se utiliza F de salida, F_S , o sea

$$F_S = \frac{SS_R(x_s | x_1, \dots, x_{s-1})}{MS_E(x_1, \dots, x_s)} < F_{\pi, 1, n-s+1}$$

quedando la variable x_s cuya F_S sea mas baja.

El proceso es el siguiente:

- 1) se define previamente un valor de probabilidad π (.10, .05, ...)
- 2) se hace la regresión con todos los p estimadores
- 3) se selecciona la variable cuyo coeficiente de correlación parcial con y sea más pequeño en valor absoluto, a condición que su F de salida no sea significativa
- 3) Se repite 4) hasta que no hay más variables con F de salida no significativa.

La regresión paso a paso (stepwise)

Los dos métodos, para adelante y para atrás sugieren muchas combinaciones posibles. La más conocida consiste en el método *stepwise*, en el cual a cada paso se puede ingresar o igualmente quedar una variable.

Claro que para que el método funcione, se necesita que las estadísticas de referencia sean diferentes. Por esto se elige el nivel de la F de entrada un poco más grande que el nivel de la F de salida, o sea

$$\pi_E > \pi_S$$

Entonces, se procede de esta manera:

- 1) se definen previamente dos valores de probabilidad, $\pi_E > \pi_S$
- 2) iterativamente ...
- 3) se testa si entre las variables no incluidas en el modelo hay que tienen una F_E significativa a nivel π_E y si hay se incluye la que tiene el valor mas grande
- 3) se testa si entre las variables en el modelo hay que tienen una F_S no significativa a nivel π_S y si hay se queda la que tiene el valor mas pequeño
- 4) Se repite desde 2) hasta que no hay más variables con F_E o F_S significativas.

Debe considerarse siempre que:

- el orden de entrada o de salida de una variable en el modelo no implica su importancia relativa o absoluta con respecto de las otras
- alguna variable importante puede ser quedada porque entraron otras cuya calidad explicativa global resulte mejor.
- los tres métodos no producen el mismo modelo de regresión y por esto se sugiere de emplearlos todos juntos
- no hay razón porque exista un acuerdo entre los métodos paso a paso y lo de todas las regresiones, porque el vínculo dado para la iteratividad de los procesos no implica un mejor modelo posible.