EL MODELO LINEAL

Clase 3

Regresión múltiple

Sergio Camiz

LIMA - Marzo-Mayo 2025

Asuntos de la clase 3

- Notación matricial
- La regresión lineal múltiple
- Estimación de los parámetros
- Análisis de la varianza
- Apalancamiento

Notación matricial

Si tenemos dos vectores, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ respuesta y $\boldsymbol{x} = (x_1, \ldots, x_n)'$ explicativo, se puede exprimir toda la teoría del modelo lineal en forma matricial, si juntamos un tercero vector $\mathbf{1} = (1, \ldots, 1)'$ y, en lugar de $(\alpha, \beta)'$, consideramos el vector de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Ahora se puede construir la matriz de los regresores $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x})$ así que el modelo lineal se escribe

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

con $\boldsymbol{\varepsilon} = (e_1, \ldots, e_n)'$ vector de residuos.

Con esta notación el modelo lineal consiste en estimar $\hat{\boldsymbol{\beta}}$ de manera que

$$SS_e(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Derivando parcialmente y igualando a cero, se consigue el sistema de ecuaciones:

$$\frac{\partial SS_e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0$$

donde $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$, así que

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

y

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

La condición de mínimo se resulta ya que

$$\frac{\partial^2 SS_e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = 2\boldsymbol{X}'\boldsymbol{X} \geq 0$$

.

Efectivamente en el caso que $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x})$ se resulta:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & S_x \\ S_x & S_{xx} \end{pmatrix} \quad y \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{S_{xx}}{nS_{xx}-S_x^2} & \frac{-S_x}{nS_{xx}-S_x^2} \\ \frac{-S_x}{nS_{xx}-S_x^2} & \frac{n}{nS_{xx}-S_x^2} \end{pmatrix}.$$

Además $\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$, así que

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} \frac{S_yS_{xx}-S_xS_{xy}}{nS_{xx}-S_x^2} \\ \frac{nS_{xy}-S_xS_y}{nS_{xx}-S_x^2} \end{pmatrix} \text{ y } \hat{\boldsymbol{\eta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Se reconoce que $\hat{\beta}_2 = \frac{nS_{xy}-S_xS_y}{nS_{xx}-S_x^2} = \frac{S_{\dot{x}\dot{y}}}{S_{\dot{x}\dot{x}}} = \hat{\beta}$, y sustituyendo esto en α se resulta $\hat{\alpha} = \bar{y} - \hat{\beta}_2\bar{x} = \frac{S_y}{n} - \frac{nS_{xy}-S_xS_y}{nS_{xx}-S_x^2}\frac{S_x}{n} = \frac{(S_yS_{xx}-\frac{S_yS_x^2}{n})-(S_xS_{xy}-\frac{S_yS_x^2}{n})}{nS_{xx}-S_x^2} = \frac{S_yS_{xx}-S_xS_{xy}}{nS_{xx}-S_x^2} = \hat{\beta}_1$.

La regresión lineal múltiple

La notación matricial permite de generalizar los resultados conseguidos al caso de un conjunto de variables explicativas.

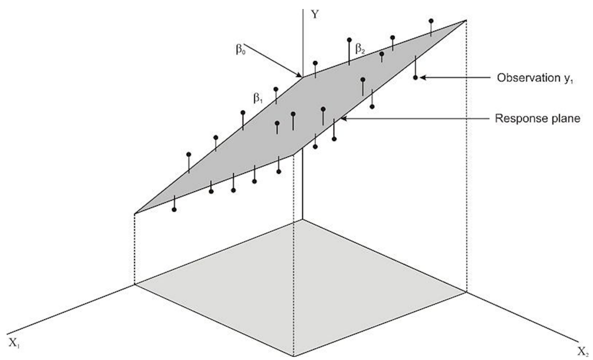
El *modelo de regresión lineal múltiple* es un modelo en el cual los parámetros aparecen linealmente en la ecuación, que se vuelve

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{\beta}$$

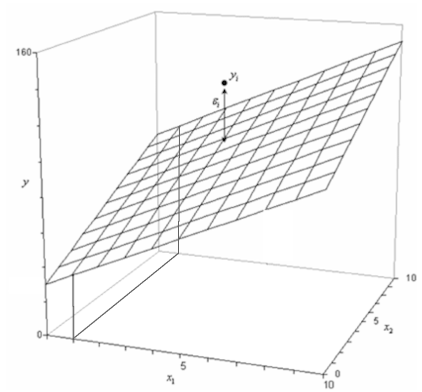
donde las componentes $\boldsymbol{x}_j$ de $\boldsymbol{X}$ solo son función de variables $\boldsymbol{z}_h$ *sin parámetros a estimar*.

Desde aquí hacia adelante, siempre se supone que entre las p variables explicativas, haya una compuesta para una columna de 1, que sirve para estimar el α, así que hablando de p columnas se estará entendiendo $p-1$ regresores más la estimación de α.

Con dos variables explicativas hay un plan de regresión en $\mathbb{R}^3$.



El caso $\boldsymbol{y} = \boldsymbol{\beta}(\mathbf{1}, \boldsymbol{x})$ se puede ver como la intersección del plano de regresión con el plano vertical $x_1 = 1$, o sea una recta.

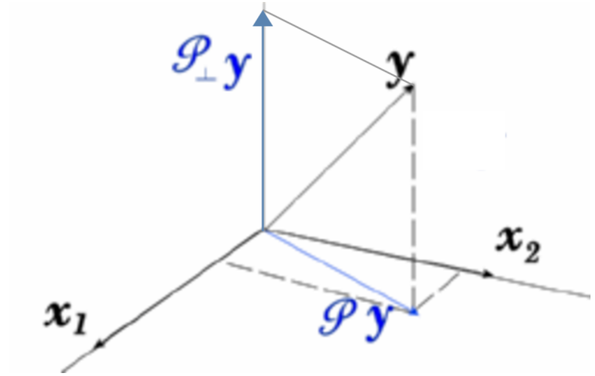


Modelo geométrico

- Se hicieron n observaciones, con $n >> p$, del carácter respuesta $y_i,\ i \in (1,n)$ correspondiendo a valores prefijos de variables descriptivas $x_{i1},\dots,x_{ip},\ i \in (1,n)$.
- Se quiere estimar los parámetros del modelo lineal con $\boldsymbol{y}$ variable respuesta y p descriptivas $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_p)$.
- Normalmente se supone que el primero sea el vector constante $\boldsymbol{x}_1 = \mathbf{1}$, que sirve para estimar la constante del modelo, resultando la matriz $\boldsymbol{X}_{n\times p}$.
- Cada variable $\boldsymbol{y}$ y $\boldsymbol{x}_j$ (columnas de $\boldsymbol{X}$) son vectores con n componentes, así que son vectores en el espacio $\mathbb{R}^n$.
- Los residuos $\varepsilon_i,\ i \in (1,n)$ forman también un vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$.
- En forma vectorial, el modelo se escribe

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

La representación de $\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2$ en $\mathbb{R}^n$ y las proyecciones ortogonales



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

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

$$< \boldsymbol{v}, \boldsymbol{w} >= \sum_i v_i w_i,$$

- A dos vectores $\boldsymbol{v}, \boldsymbol{w}$ se llaman *ortogonales* si $< \boldsymbol{v}, \boldsymbol{w} >= 0$.
- Se deduce una norma de los vectores o sea su longitud

$$\| \boldsymbol{v} \|^2 = < \boldsymbol{v}, \boldsymbol{v} >= \sum_i v_i v_i,$$

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

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

- Dados $\boldsymbol{v}, \boldsymbol{w}$, la *proyección ortogonal* de $\boldsymbol{v}$ sobre $\boldsymbol{w}$ es el vector colineal a $\boldsymbol{w}$

$$\mathscr{P}_{\boldsymbol{w}}\boldsymbol{v} = \frac{<\boldsymbol{v}, \boldsymbol{w}>}{\parallel \boldsymbol{w} \parallel}\boldsymbol{w}$$

- su *complemento ortogonal*

$$\mathscr{P}_{\perp\boldsymbol{w}}\boldsymbol{v} = (I - \mathscr{P}_{\boldsymbol{w}})\boldsymbol{v} = \boldsymbol{v} - \mathscr{P}_{\boldsymbol{w}}\boldsymbol{v} = \boldsymbol{v} - \frac{<\boldsymbol{v}, \boldsymbol{w}>}{\parallel \boldsymbol{w} \parallel}\boldsymbol{w}$$

- Resultan

$$\boldsymbol{v} = \mathscr{P}_{\boldsymbol{w}}\boldsymbol{v} + \mathscr{P}_{\perp\boldsymbol{w}}\boldsymbol{v} \quad \text{y} \quad <\mathscr{P}_{\boldsymbol{w}}\boldsymbol{v}, \mathscr{P}_{\perp\boldsymbol{w}}\boldsymbol{v}> = 0$$

- Nótese que, dado un vector $\boldsymbol{v}$ y un subespacio S cualquier, siempre existen su projección ortogonal sobre S y su complemento también.

- Dado un espacio S, siempre se puede construir una base compuesta por vectores ortogonales.

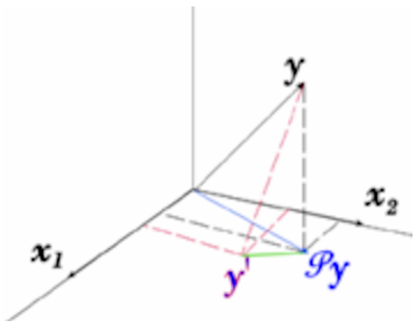
- Un vector $\boldsymbol{v}$ es ortogonal a un subespacio S si el es ortogonal a todos los vectores de una base de S.

- por tanto (teorema de Pitágoras)

$$\boldsymbol{v} = \mathscr{P}_{S}\boldsymbol{v} + \mathscr{P}_{\perp S}\boldsymbol{v} \quad \text{y} \quad <\mathscr{P}_{S}\boldsymbol{v}, \mathscr{P}_{\perp S}\boldsymbol{v}> = 0$$

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

$$d(P, S) = \|\mathscr{P}_{\perp S}(\boldsymbol{OP})\| = \min_{s \in S} d(P, s)$$



- En $\mathbb{R}^3$ por el teorema de Pitágoras la distancia $d(\boldsymbol{y}, \mathscr{P}(\boldsymbol{y}))$ es mínima en cuanto $\boldsymbol{y}\boldsymbol{y}'$ es la hipotenusa de $\boldsymbol{y}\mathscr{P}\boldsymbol{y}'$.
- La solución de mínimos cuadrados se encuentra para la proyección ortogonal de $\boldsymbol{y}$ sobre el espacio X generado para $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_p)$.

Teorema. El punto

$$\hat{\boldsymbol{\eta}} = \mathscr{P}_X(\boldsymbol{y})$$

es el punto de X, subespacio generado para $\boldsymbol{X}$ más cercano a $\boldsymbol{y}$.

- Geométricamente la solución dada para la proyección *siempre existe y es única.*

- Esta se consigue empleando el *operador ortogonal de proyección* sobre X, $\mathscr{P}_X$.

- Dado cualquier vector $\boldsymbol{y}$, el vector $\boldsymbol{y} - \mathscr{P}_X(\boldsymbol{y})$, es ortogonal a cada vector de X,

- entonces cada vector de $\mathbb{R}^n$ se encuentra compartido en dos componentes

$$\boldsymbol{y} = \mathscr{P}_X \boldsymbol{y} + (\boldsymbol{y} - \mathscr{P}_X \boldsymbol{y})$$

- el espacio también resulta compartido como *suma directa* de sub-espacios ortogonales:
- Si se escribe

$$\boldsymbol{y} - \mathscr{P}_X \boldsymbol{y} = (I - \mathscr{P}_X)\boldsymbol{y} = \mathscr{P}_{X^\perp} \boldsymbol{y} = \mathscr{E}_X \boldsymbol{y}$$

resulta por tanto

$$\boldsymbol{y} = \mathscr{P}_X \boldsymbol{y} \oplus (\boldsymbol{y} - \mathscr{P}_X \boldsymbol{y}) = \mathscr{P}_X \boldsymbol{y} \oplus \mathscr{E}_X \boldsymbol{y}.$$

- el espacio también resulta compartido como *suma directa* de sub-espacios ortogonales: $X \oplus X^\perp$
- Cada proyector ortogonal es *idempotente*, o sea $\mathscr{P} \circ \mathscr{P} = \mathscr{P}$ y por tanto $\mathscr{P} \circ \mathscr{E} = \mathscr{P} \circ (I - \mathscr{P}) = \mathscr{P} - \mathscr{P} \circ \mathscr{P} = 0$.
- En el caso, $\mathbb{R}^n = X \oplus X^\perp$, donde $X^\perp$ es el *complemento ortogonal de* X en $\mathbb{R}^n$, la proyección ortogonal de $\boldsymbol{y}$ en X es

$$\mathscr{P}_X \boldsymbol{y} = \hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}},$$

donde resulta $\boldsymbol{y} = \hat{\boldsymbol{\eta}} + \boldsymbol{e}$ con $\boldsymbol{y} - \hat{\boldsymbol{\eta}} = (I - \mathscr{P}_X)\boldsymbol{y} = \boldsymbol{e} \in X^\perp$.

- A $X^\perp$ se lo llama el *espacio de los errores* o de los *residuos*.

Estimación de los parámetros

Utilizando el teorema de Pitágoras, resulta que el mínimo

$$\begin{aligned} SS_e &= \min_s (\boldsymbol{y} - \boldsymbol{s})'(\boldsymbol{y} - \boldsymbol{s}) \\ &= \min_s ((\boldsymbol{y} - \hat{\boldsymbol{\eta}})'(\boldsymbol{y} - \hat{\boldsymbol{\eta}}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{s})'(\hat{\boldsymbol{\eta}} - \boldsymbol{s})) = \\ &= \min_\theta ((\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}))'(\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}))) \end{aligned}$$

se consigue cuando el segundo término es cero, o sea cuando

$$\boldsymbol{\theta} = \hat{\boldsymbol{\beta}}$$

Se puede calcular $\hat{\boldsymbol{\beta}}$ considerando que $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\eta}} \in X^\perp$, porque así $\boldsymbol{e}$ es ortogonal a cada vector de X, donde desde

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

resulta el sistema de *ecuaciones normales*

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}. \tag{1}$$

La solución (siempre existente) de (1) simplemente sería

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

y por consecuencia

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3}$$

Para su cálculo, hay que invertir $\boldsymbol{X}'\boldsymbol{X}$, solo posible en el caso de rango completo. Si al contrario el rango es $r < p + 1$ se hace recurso a su *inversa generalizada*.

Dada una matriz $\boldsymbol{A}$ se llama su *inversa generalizada* una matriz $\boldsymbol{A}^g$ tal que $\boldsymbol{A}\boldsymbol{A}^g\boldsymbol{A} = \boldsymbol{A}$.

Dada una matriz $\boldsymbol{A}$ se llama su *inversa de Moore-Penrose* a la matriz $\boldsymbol{A}^+$ (única) tal que

1. $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$.
2. $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$.
3. $(\boldsymbol{A}\boldsymbol{A}^+)^t = \boldsymbol{A}\boldsymbol{A}^+$.
4. $(\boldsymbol{A}^+\boldsymbol{A})^t = \boldsymbol{A}^+\boldsymbol{A}$.

Si la descomposición a valores singulares de $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^t$, con $\boldsymbol{U}^t\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^t\boldsymbol{V} = I$ y $\boldsymbol{\Sigma}$ diagonal, se resulta $\boldsymbol{A}^+ = \boldsymbol{U}\boldsymbol{\Sigma}^+\boldsymbol{V}^t$ con $0 \neq \sigma_{ii}^+ = \sigma_{ii}^{-1}$.
Normalmente, los programas de modelos lineares se la utilizan directamente.

Hay que considerar que la solución geométrica siempre existe, ya que se trata de una proyección.

- Si el rango es completo, $\dim X = p$ y $\hat{\boldsymbol{\beta}}$ se estima sin problema.
- Si el rango no es completo, $\dim X = r < p$, significa que unos de los descriptores son combinación lineal de los demás y hay que tirarlos, sino las coordenadas de los puntos en el espacio resultan indeterminadas, ya que los regresores no son una base del espacio.
- Hay casos donde los programas mismos hacen esta selección automáticamente.

En seguida, consideramos el rango completo, o sea

$$\dim X = \operatorname{rank}\boldsymbol{X} = p.$$

Del punto de vista geométrico, se ha visto que $\hat{\boldsymbol{\eta}} = \mathscr{P}_X\boldsymbol{y}$ es la proyección ortogonal de $\boldsymbol{y}$ sobre el espacio X, generado para los descriptores columnas de $\boldsymbol{X}$. Ahora, como tenemos

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

desde (3) tenemos que

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \mathscr{P}\boldsymbol{y}.$$

Como $\hat{\boldsymbol{\eta}}$ se encuentra en X, es fácil de averiguar que

$$\mathscr{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

es simétrica y idempotente, y por lo tanto es el proyector ortogonal sobre X, $\mathscr{P}_X$. <

Teorema *Una matriz idempotente es un proyector ortogonal si y solo si es simétrica.*

Si $\mathbb{R}^n = X \oplus X^\perp$, ortogonales, sean $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, entonces, proyectando sobre X, se resulta
$< \mathscr{P}(\boldsymbol{x}), \boldsymbol{y} - \mathscr{P}(\boldsymbol{y}) >=< \boldsymbol{x} - \mathscr{P}(\boldsymbol{x}), \mathscr{P}(\boldsymbol{y}) >= 0$
y por lo tanto
$< \mathscr{P}(\boldsymbol{x}), \boldsymbol{y} >=< \mathscr{P}(\boldsymbol{x}), \mathscr{P}(\boldsymbol{y}) >=< \boldsymbol{x}, \mathscr{P}(\boldsymbol{y}) >$
simétrica.
Si al contrario es simétrica, entonces

$$\begin{aligned}
&< \mathscr{P}(\boldsymbol{x}), \boldsymbol{x} - \mathscr{P}(\boldsymbol{x}) >=< \mathscr{P}(\boldsymbol{x}), \boldsymbol{x} > - < \mathscr{P}(\boldsymbol{x}), \mathscr{P}(\boldsymbol{x}) >= \\
&< \mathscr{P}(\boldsymbol{x}), \boldsymbol{x} > - < \mathscr{P}^2(\boldsymbol{x}), \boldsymbol{x} >= \\
&< \mathscr{P}(\boldsymbol{x}), \boldsymbol{x} > - < \mathscr{P}(\boldsymbol{x}), \boldsymbol{x} >= 0,
\end{aligned}$$

ortogonal.

Por lo tanto, $\mathscr{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ es el proyector ortogonal sobre X, y como $\hat{\boldsymbol{\eta}}$ se encuentra en X, es la *proyección ortogonal de* $\boldsymbol{y}$ *sobre* X, que tiene dimension p.

En consecuencia, el vector $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\eta}} = (\boldsymbol{I} - \mathscr{P}_X)\boldsymbol{y} = \mathscr{E}\boldsymbol{y}$ es la proyección ortogonal de $\boldsymbol{y}$ sobre el espacio de residuos $X^{\perp}$ que entonces tiene dimensión $n - p$.

En adelante, si no hay confusión, a $\mathscr{P}_X$ se lo escriberemos $\mathscr{P}$.

Se consiguen los resultados siguientes:

- $\hat{\boldsymbol{\beta}}$ *es el estimador de mínimos cuadrados* de $\boldsymbol{\beta}$, que minimiza la distancia entre $\boldsymbol{y}$ y su estimador colocado en el espacio de los regresores X;
- $\hat{\boldsymbol{\eta}}$ *proyección ortogonal de* $\boldsymbol{y}$ *sobre* X es el vector de los *valores ajustados* para la regresión: como proyección ortogonal de $\boldsymbol{y}$ sobre X, es el punto de X más cercano de $\boldsymbol{y}$;
- $\boldsymbol{e}$ *proyección ortogonal de* $\boldsymbol{y}$ *sobre* $X^{\perp}$ es el vector de los *residuos*, ortogonal a $\hat{\boldsymbol{\eta}}$;

- $SS_r = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}$ es la norma del vector $\hat{\boldsymbol{\eta}}$, que vale

$$\begin{aligned} SS_r = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} &= \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \\ &= \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \\ &= \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\mathscr{P}\boldsymbol{y} \end{aligned}$$

  y que puede verse también como el producto escalar $\boldsymbol{y}'\hat{\boldsymbol{\eta}}$;

- $SS_e = \boldsymbol{e}'\boldsymbol{e}$ es la norma del vector $\boldsymbol{e}$, y resulta

$$SS_e = \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\mathscr{E}'\mathscr{E}\boldsymbol{y} = \boldsymbol{y}'\mathscr{E}\boldsymbol{y}$$

  y que corresponde también a $\boldsymbol{y}'\boldsymbol{e}$.

- como $\mathscr{P}$ y $\mathscr{E}$ son simétricos e idempotentes, resultan

$$\begin{aligned} r(\mathscr{P}) = tr(\mathscr{P}) &= tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \\ &= tr(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = tr(\boldsymbol{I}_{p+1}) = p \\ r(\mathscr{E}) = tr(\mathscr{E}) &= tr(\boldsymbol{I}_n - \mathscr{P}) = tr(\boldsymbol{I}_n) - tr(\mathscr{P}) = n - p \end{aligned}$$

Análisis de la varianza

Cuando un modelo se estima a través de los mínimos cuadrados, se pueden descomponer las sumas de los cuadrados de los valores observados en una parte, correspondiente a la suma de los cuadrados de los valores estimados más la parte correspondiente al desvío de los valores observados a los valores estimados.

1. Las n observaciones son repartidas en q clases con numerosidad $n_1 + n_2 + \cdots + n_q = n$.

En este caso, se supone que el modelo consiste en el afectar a cada grupo el promedio de su observaciones.

$$SS_T = \sum_{i=1}^{n} y_i^2 = \sum_{k=1}^{q} n_k \bar{y}_k^2 + \sum_{k=1}^{q}\sum_{i=1}^{n_k}(y_{ik} - \bar{y}_k)^2 = SS_B + SS_W$$

ya que

$$\sum_{k=1}^{q} n_k \bar{y}_k^2 + \sum_{k=1}^{q}\sum_{i_k=1}^{n_k}(y_{i_k} - \bar{y}_k)^2 =$$
$$\sum_{k=1}^{q} n_k \bar{y}_k^2 + \sum_{k=1}^{q}\sum_{i_k=1}^{n_k} y_{i_k}^2 - 2\sum_{k=1}^{q}\sum_{i_k=1}^{n_k} y_{i_k}\bar{y}_k + \sum_{k=1}^{q}\sum_{i_k=1}^{n_k}\bar{y}_k^2 =$$
$$\sum_{k=1}^{q} n_k \bar{y}_k^2 + \sum_{i=1}^{n} y_i^2 - 2\sum_{k=1}^{q} n_k \bar{y}_k \bar{y}_k + \sum_{k=1}^{q} n_k \bar{y}_k^2 = \sum_{i=1}^{n} y_i^2$$

- SS_T La suma de cuadrados *total* se descompone en
- SS_B suma de cuadrados *between* debido al modelo, o sea a las diferencias entre clases, y
- SS_W suma de cuadrados *within* de desvío al modelo, debido a la variación dentro de las clases.

2. Las n observaciones tienen que ser estimadas con un modelo lineal simple y se calcula la suma de cuadrados del desvío al modelo, o sea los residuos:

$$\begin{aligned} SS_e &= \sum_i (y_i - \hat{\eta}_{x_i})^2 = \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)(y_i - \hat{\alpha} - \hat{\beta}x_i) \\ &= \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)y_i - \hat{\alpha}\sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i) \\ &\quad - \hat{\beta}\sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = \sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)y_i \\ &= S_{yy} - (\hat{\alpha}S_y + \hat{\beta}S_{xy}) = SS_t - SS_r \end{aligned}$$

Como $\hat{\alpha}$ y $\hat{\beta}$ resuelven las ecuaciones normales, dos términos de la cuarta igualidad son iguales a cero.

Se consigue

$$SS_t = S_{yy} = (\hat{\alpha}S_y + \hat{\beta}S_{xy}) + \sum_i (y_i - \hat{\eta}_{x_i})^2 = SS_r + SS_e$$

con

- $SS_t = S_{yy}$, *suma de cuadrados total*, se descompone en
- $SS_r = S_{\hat{\eta}\hat{\eta}}$ *suma de cuadrados de la regresión*,
- SS_e *suma de cuadrados del error* o de los *residuos*.

3. En el caso de la regresión múltiple, una vez estimados los $\hat{\boldsymbol{\beta}}$ resulta que el cuadrado de la distancia entre $\boldsymbol{y}$ y X es

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

pues, según las ecuaciones normales,

$$\boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y})' = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}},$$

donde $\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} + \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\mathscr{P}\boldsymbol{y} + \boldsymbol{y}'\mathscr{E}\boldsymbol{y}$.

Esto se puede escribir también

$$SS_t = SS_r + SS_e$$

Así se ha compartido la suma de cuadrados de las observaciones:

- una parte SS_r debido a la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}$
- otra, SS_e, debido al error.
  Por tanto:
- $\hat{\boldsymbol{\eta}}$ contiene la información sobre el modelo $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$,
- $\boldsymbol{e}$ solo contiene la información sobre el error,
- $\boldsymbol{e}'\boldsymbol{e}$ debe informar sobre la varianza del error.

Los elementos de esta partición se suelen representar en una *tabla de análisis de varianza* como

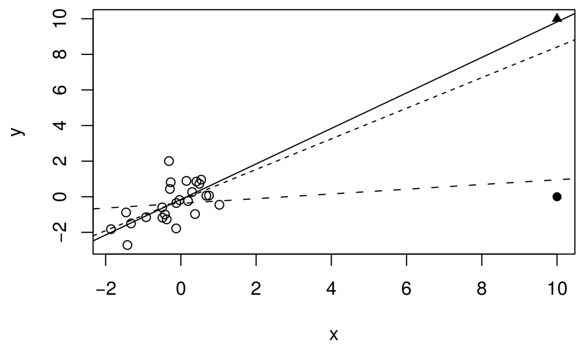
Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)
Between	q	SS_B	$MS_B = SS_B/q$
Within	$n-q$	SS_W	$MS_W = SS_W/(n-q)$
Total	n	SS_T	

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)
Regresión	p	SS_r	$MS_r = SS_r/(p)$
Error	$n-p$	SS_e	$MS_e = SS_e/(n-p)$
Total	n	SS_t	

Estaremos viendo la utilidad de los cuadrados promedios.

Apalancamiento

El *apalancamiento* se medida con la diagonal de la matriz $\hat{\boldsymbol{H}}$.



- A la matriz $\mathscr{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \hat{\boldsymbol{H}}$, proyector sobre el espacio generado para $\boldsymbol{X}$ (inclusivo del vector $\mathbf{1}$) se le llama también *hat matrix*.
- Su valor h_{ij} se lo interpreta como la influéncia del valor y_j, debido a los valores $\boldsymbol{x}_j$ correspondientes, sobre la predicción η_i.
- Al valor h_{ii} se lo llama *apalancamiento* y representa cuanto la i-ésima observación influye sobre la regresión. La verdad es que esto solo depende de la distancia de la observación del baricentro de los $\boldsymbol{x}_j$.
- Como el promedio de los h_{ii} es $(p)/n$, tien que mirar de cerca las observaciones con $h_{ii} \geq 2(p)/n$.

Ejemplo

El archivo *Linnerud* tiene mediciones físicas de 20 atletas de un gimnásio (*Peso, Cintura, Pulsaciones*) y tres prestaciones (*Tracciones, Flecciones, Saltos*). Intentamos de explicar las Tracciones a través de las mediciones físicas.

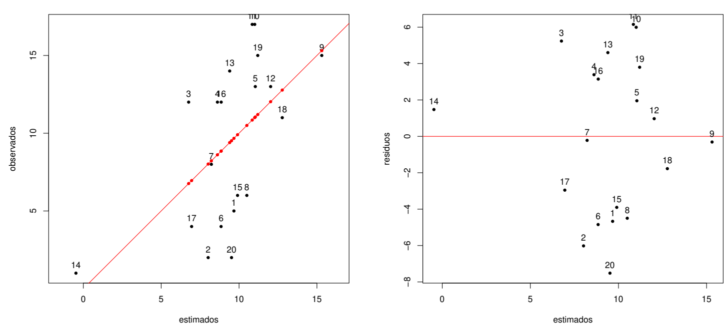
El modelo siendo: `Pulls ~ Weight+Waist+Pulse`
se resultan los $\hat{\boldsymbol{\beta}}$:

```
 (Intercept)      Weight       Waist       Pulse
47.96841291  0.07884384 -1.45584256 -0.01895002
```

y la siguiente tabla de análisis de la varianza:

	Degrees of Freedom	Sum of squares	Mean squares
Regression	4	1966.3455	491.58638
Residuals	16	350.6545	21.91591
Total	20	2317.0000	0.00000

En los gráficos se comparan los valores estimados con los observados (izquierda) y losresíduos (derecha). En rojo el modelo.



Resultan con apalancamiento relevante ($> 2p/n = 0{,}4$) los atletas 9 y 14.

Corriendo el mismo modelo sin ellos, se resultan los $\hat{\boldsymbol{\beta}}$:

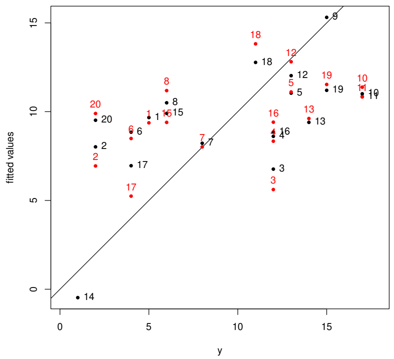
```
(Intercept)      Weight       Waist       Pulse
61.23658732  0.14053819 -2.27498972  0.06383284
```

que se pueden comparar a los del modelo completo:

```
 (Intercept)      Weight       Waist       Pulse
47.96841291  0.07884384 -1.45584256 -0.01895002
```

Las diferencias son bastante evidentes: en los gráficos se ven las diferencias entre valores estimados e residuos con respecto a los valores y observados.

En el gráfico se comparan valores observados (horizontal) con los estimados(vertical). En rojo las estimaciones sin 9 y 14.



Aquí los residuos (con signo cambiado) son representados contra los valores observados.