EL MODELO LINEAL

Clase 5

Selección de modelos

Sergio Camiz

LIMA - Marzo-Mayo 2025

Asuntos de la clase 5

- Partición de la regresión
- La eliminación del promedio
- Varianza, covarianza y correlación
- Selección del modelo
- Ejemplo

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 $p_1 + p_2 = p$ regresores separados, $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$, y por esto el modelo se puede escribir como

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

donde

$$\boldsymbol{X} = (\boldsymbol{X}_1|\boldsymbol{X}_2), \quad \boldsymbol{\beta'} = (\boldsymbol{\beta}_1'|\boldsymbol{\beta}_2').$$

Bajo esta partición la matriz $\boldsymbol{X'X}$ toma la forma

$$\boldsymbol{X'X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \hline \boldsymbol{X}_2' \end{pmatrix} (\boldsymbol{X}_1|\boldsymbol{X}_2) = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 \mid \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \hline \boldsymbol{X}_2'\boldsymbol{X}_1 \mid \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}$$

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 $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{X}_2'\boldsymbol{X}_1 = \boldsymbol{0}$, o no.

Caso ortogonal

En el caso ortogonal bajo la condición $\boldsymbol{X}_1'\boldsymbol{X}_2 = \mathbf{0}$ la matriz $\boldsymbol{X}'\boldsymbol{X}$ toma la forma

$$\begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}$$

y entonces

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} \end{pmatrix}$$

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

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} \end{pmatrix}.$$

Nótese que las estimaciones de los $\hat{\boldsymbol{\beta}}$ hechas conjuntamente o separadas se quedan idénticas. Además, la varianza resulta ser

$$V(\hat{\boldsymbol{\beta}}) = V\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2\begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} \end{pmatrix},$$

así que la covarianza entre $\hat{\boldsymbol{\beta}}_1$ y $\hat{\boldsymbol{\beta}}_2$ es cero.

Por otro lado se tiene

$$\begin{aligned}\mathscr{P}_X &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \\ &= (\boldsymbol{X}_1|\boldsymbol{X}_2)\begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1} \end{pmatrix}(\boldsymbol{X}_1'|\boldsymbol{X}_2') = \\ &= \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' + \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2' = \mathscr{P}_{X_1} + \mathscr{P}_{X_2}\end{aligned}$$

Entonces, la proyección sobre X es la suma de las dos proyecciones sobre los espacios X_1 y X_2, generados respectivamente para $\boldsymbol{X}_1$ y $\boldsymbol{X}_2$. En consecuencia

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= \mathscr{P}_X\boldsymbol{y} = (\mathscr{P}_{X_1} + \mathscr{P}_{X_2})\boldsymbol{y} = \mathscr{P}_{X_1}\boldsymbol{y} + \mathscr{P}_{X_2}\boldsymbol{y} = \\ &= \mathscr{P}_{X_1}\hat{\boldsymbol{\eta}} + \mathscr{P}_{X_2}\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_1 + \hat{\boldsymbol{\eta}}_2\end{aligned}$$

y como las proyecciones son simétricas e idempotentes,

$$\begin{aligned}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} &= \boldsymbol{y}'\mathscr{P}_X\boldsymbol{y} = \boldsymbol{y}'\mathscr{P}_{X_1}\boldsymbol{y} + \boldsymbol{y}'\mathscr{P}_{X_2}\boldsymbol{y} = \\ &= \boldsymbol{y}'\mathscr{P}_{X_1}'\mathscr{P}_{X_1}\boldsymbol{y} + \boldsymbol{y}'\mathscr{P}_{X_2}'\mathscr{P}_{X_2}\boldsymbol{y} = \hat{\boldsymbol{\eta}}_1'\hat{\boldsymbol{\eta}}_1 + \hat{\boldsymbol{\eta}}_2'\hat{\boldsymbol{\eta}}_2\end{aligned}$$

Por los residuos resulta

$$\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\eta}}_1'\hat{\boldsymbol{\eta}}_1 - \hat{\boldsymbol{\eta}}_2'\hat{\boldsymbol{\eta}}_2 = \boldsymbol{y}'(I - \mathscr{P}_{X_1} - \mathscr{P}_{X_2})\boldsymbol{y}$$

Es decir que

$$SS_t = S_{yy} = \hat{\boldsymbol{\eta}}_1'\hat{\boldsymbol{\eta}}_1 + \hat{\boldsymbol{\eta}}_2'\hat{\boldsymbol{\eta}}_2 + \boldsymbol{e}'\boldsymbol{e} = SS_{\boldsymbol{X}_1} + SS_{\boldsymbol{X}_2} + SS_e$$

Se resulta la tabla de análisis de varianza siguiente:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados promedios (MS)	Esperanza de los cuadrados promedios E(MS)
$\boldsymbol{X}_1$	p_1	$SS_{\boldsymbol{X}_1} = \boldsymbol{y}'\mathscr{P}_{X_1}\boldsymbol{y}$	$MS_{\boldsymbol{X}_1} = SS_{\boldsymbol{X}_1}/p_1$	$\sigma^2 + \boldsymbol{\eta}_1'\boldsymbol{\eta}_1/p_1$
$\boldsymbol{X}_2$	p_2	$SS_{\boldsymbol{X}_2} = \boldsymbol{y}'\mathscr{P}_{X_2}\boldsymbol{y}$	$MS_{\boldsymbol{X}_2} = SS_{\boldsymbol{X}_2}/p_2$	$\sigma^2 + \boldsymbol{\eta}_2'\boldsymbol{\eta}_2/p_2$
Error	$n - p_1 - p_2$	$SS_e = \boldsymbol{y}'(I - \mathscr{P}_{X_1} - \mathscr{P}_{X_2})\boldsymbol{y}$	$MS_e = SS_e/(n - p_1 - p_2)$	σ^2
Total	n	SS_t		

Si el interés solo está concentrado sobre $\boldsymbol{\beta}_2$ se puede borrar la primera linea, sustrayendo $SS_{\boldsymbol{X}_1}$ de SS_t:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados promedios (MS)	Esperanza de los cuadrados promedios E(MS)
$\boldsymbol{X}_2$	p_2	$SS_{\boldsymbol{X}_2} = \boldsymbol{y}'\mathscr{P}_{S_2}\boldsymbol{y}$	$MS_{\boldsymbol{X}_2} = SS_{\boldsymbol{X}_2}/p_2$	$\sigma^2 + \boldsymbol{\eta}_2'\boldsymbol{\eta}_2/p_2$
Error	$n - p_1 - p_2$	$SS_e = \boldsymbol{y}'(I - \mathscr{P}_{X_1} - \mathscr{P}_{X_2})\boldsymbol{y}$	$MS_e = SS_e/(n - p_1 - p_2)$	σ^2
Total	$n - p_1$	$SS_t - SS_{\boldsymbol{X}_1} = \boldsymbol{y}'(I - \mathscr{P}_{X_1})\boldsymbol{y}$		

Esta tabla muestra que cuando el interés esta limitado a $\boldsymbol{\beta}_2$, la suma de cuadrados total es la suma de cuadrados de los residuos de la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$.

Caso no ortogonal

En el caso no ortogonal, se resulta $\boldsymbol{X}_1'\boldsymbol{X}_2 \neq \boldsymbol{0}$, así que hay que *ortogonalizar* $\boldsymbol{X}_2$ por respecto a $\boldsymbol{X}_1$. Esto porque la influencia de cada $\boldsymbol{x_{2j}}$ sobre $\boldsymbol{y}$ depende parcialmente de los $\boldsymbol{X}_1$.

Por esto, ya que los residuos de la proyección de los $\boldsymbol{X}_2$ sobre los $\boldsymbol{X}_1$ son ortogonales a $\boldsymbol{X}_1$, en lugar de los $\boldsymbol{X}_2$ hay solo que tomar los residuos de su regresiones sobre $\boldsymbol{X}_1$, $\boldsymbol{X}_2 = \boldsymbol{X}_1\hat{\boldsymbol{B}}_2 + \boldsymbol{E}$, que forman la matriz $\boldsymbol{E} = \boldsymbol{X}_{2\cdot 1} = \boldsymbol{X}_2 - \boldsymbol{X}_1\hat{\boldsymbol{B}}_2$.

Por lo tanto, el modelo que se quiere estimar resulta ser

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_{2\cdot 1}\boldsymbol{\beta}_{2\cdot 1} + \boldsymbol{\varepsilon}$$

Claro que aquí también se resultan las tablas de análisis de varianza

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados promedios (MS)	Esperanza de los cuadrados promedios E(MS)
$\boldsymbol{X}_1$	p_1	$SS_{\boldsymbol{X}_1} = \boldsymbol{y}'\mathscr{P}_{X_1}\boldsymbol{y}$	$MS_{\boldsymbol{X}_1} = SS_{\boldsymbol{X}_1}/p_1$	$\sigma^2 + \boldsymbol{\eta}_1'\boldsymbol{\eta}_1/p_1$
$\boldsymbol{X}_{2\cdot 1}$	p_2	$SS_{\boldsymbol{X}_{2\cdot 1}} = \boldsymbol{y}'\mathscr{P}_{X_{2\cdot 1}}\boldsymbol{y}$	$MS_{\boldsymbol{X}_{2\cdot 1}} = SS_{\boldsymbol{X}_{2\cdot 1}}/p_2$	$\sigma^2 + \boldsymbol{\eta}_{2\cdot 1}'\boldsymbol{\eta}_{2\cdot 1}/p_2$
Error	$n - p_1 - p_2$	$SS_e = \boldsymbol{y}'(I - \mathscr{P}_{X_1} - \mathscr{P}_{X_2})\boldsymbol{y}$	$SS_e/(n - p_1 - p_2)$	σ^2
Total	n	SS_t		

Si el interés solo está concentrado sobre $\boldsymbol{\beta}_{2\cdot 1}$ se puede borrar la primera linea, sustrayendo $SS_{\boldsymbol{X}_1}$ de SS_t:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados promedios (MS)	Esperanza de los cuadrados promedios E(MS)
$\boldsymbol{X}_{2\cdot 1}$	p_2	$SS_{\boldsymbol{X}_{2\cdot 1}} = \boldsymbol{y}'\mathscr{P}_{X_{2\cdot 1}}\boldsymbol{y}$	$MS_{\boldsymbol{X}_{2\cdot 1}} = SS_{\boldsymbol{X}_{2\cdot 1}}/p_2$	$\sigma^2 + \boldsymbol{\eta}_{2\cdot 1}'\boldsymbol{\eta}_{2\cdot 1}/p_2$
Error	$n - p_1 - p_2$	$SS_e = \boldsymbol{y}'(I - \mathscr{P}_{X_1} - \mathscr{P}_{X_{2\cdot 1}})\boldsymbol{y}$	$SS_e/(n - p_1 - p_2)$	σ^2
Total	$n - p_1$	$SS_t - SS_{\boldsymbol{X}_1} = \boldsymbol{y}'(I - \mathscr{P}_{X_1})\boldsymbol{y}$		

En este caso también, cuando el interés esta limitado a $\boldsymbol{\beta}_2$, la suma de cuadrados total es la suma de cuadrados de los residuos de la regresión de $\boldsymbol{y}$ sobre $\boldsymbol{X}_1$.

La eliminación del promedio

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

$$E(\boldsymbol{y}) = \beta_1\mathbf{1} + \beta_2\boldsymbol{x}_2 + ... + \beta_p\boldsymbol{x}_p$$

donde $\mathbf{1} = (1, 1, ..., 1)'$ tiene dimensión $n \times 1$, pero donde el interés se centra sobre los parámetros $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_p)$. Se puede observar que, si estos son todos cero, resulta $\beta_1 = \bar{y}$, el promedio de los y_i, que no es de interés, mientras que interesan otros factores. Se vuelve así al modelo

$$E(\boldsymbol{y}) = (\mathbf{1}' \mid \boldsymbol{X}_2)\binom{\beta_1}{\boldsymbol{\beta}_2}$$

con $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_p)$ de dimensión $p_2 = p - 1$, donde queremos estimar los parámetros y la suma de cuadrados de los errores, conociendo la varianza de $\boldsymbol{y}$, o sea $\sigma^2 \boldsymbol{I}$. Se debe entonces ortogonalizar previamente los $\boldsymbol{X}_2$ por respecto al vector $\mathbf{1}$, lo que corresponde a haber centrado cada regresor alrededor de su promedio.

Efectivamente se consigue $\boldsymbol{X}_{2\cdot 1} = \dot{\boldsymbol{X}}$, ya que es claro que

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

como vector de desvíos al promedio.

Entonces se puede estimar el modelo

$$\begin{cases} \boldsymbol{y} & = \mathbf{1}'\phi + \dot{\boldsymbol{X}}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) & = \mathbf{0} \\ V(\boldsymbol{\varepsilon}) & = \sigma^2 \boldsymbol{I} \end{cases}$$

Como los $\boldsymbol{x}_j$ son centrados, se resulta

$$\bar{y} = \phi + \sum_j \beta_{2,j}(\bar{\dot{x}}_j) + \bar{\varepsilon} = \phi + \bar{\varepsilon}$$

y por lo tanto

$$E(\bar{y}) = \phi \ \text{ y también } \ E(\hat{\phi}) = \bar{y} = \phi.$$

Resulta también que $\boldsymbol{\eta} = \phi + \bar{\boldsymbol{x}}'\boldsymbol{\beta}_{2\cdot 1}$ y que

$$\left(\frac{\hat{\boldsymbol{\phi}}}{\hat{\boldsymbol{\beta}}_{2\cdot 1}}\right) = \left(\frac{(\mathbf{1}'\mathbf{1})^{-1} | \quad \mathbf{0}}{\mathbf{0} \quad |(\dot{\boldsymbol{X}}'\dot{\boldsymbol{X}})^{-1}}\right)\left(\frac{\mathbf{1}'}{\dot{\boldsymbol{X}}'}\right)\boldsymbol{y} = \left(\frac{\bar{y}}{(\dot{\boldsymbol{X}}'\dot{\boldsymbol{X}})^{-1}\dot{\boldsymbol{X}}'\boldsymbol{y}}\right)$$

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

$$V\left(\frac{\hat{\boldsymbol{\phi}}}{\hat{\boldsymbol{\beta}}_{2\cdot 1}}\right) = \sigma^2 \left(\frac{1/n \,| \quad \mathbf{0}}{\mathbf{0} \quad | \,(\dot{\boldsymbol{X}}'\dot{\boldsymbol{X}})^{-1}}\right)$$

así que $\hat{\boldsymbol{\phi}}$ y $\hat{\boldsymbol{\beta}}_{2\cdot 1}$ son no correlados entre ellos.

Además, mediando, de

$$\begin{aligned} y_i &= \beta_1 + \sum_j \beta_{2,j}(x_{ij}) + \varepsilon_i \text{ se consigue} \\ \bar{y} &= \beta_1 + \sum_j \beta_{2,j}(\bar{x}_j) + \bar{\varepsilon}, \text{ así que} \\ y_i - \bar{y} &= \sum_j \beta_{2,j}(x_{ij}) - \sum_j \beta_{2,j}(\bar{x}_j) + \varepsilon_i - \bar{\varepsilon}. \end{aligned}$$

Igualmente, de

$$\begin{aligned} y_i &= \phi + \sum_j \beta_{2\cdot 1,j}(\dot{x}_{ij}) + \varepsilon_i \text{ se consigue} \\ \bar{y} &= \phi, \text{ así que} \\ y_i - \bar{y} &= \sum_j \beta_{2\cdot 1,j}(x_{ij}) - \sum_j \beta_{2\cdot 1,j}(\bar{x}_j) + \varepsilon_i - \bar{\varepsilon}, \end{aligned}$$

donde en este caso la igualdad $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2\cdot 1}$.

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

$$\boldsymbol{e}_1 = \mathscr{E}_{\mathbf{1}\boldsymbol{y}} = (I - n^{-1}\mathbf{1}\mathbf{1}')\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{y}'\mathbf{1}/n = \dot{\boldsymbol{y}}$$

correspondiendo a centrar $\boldsymbol{y}$, o sea *eliminar el promedio*. Por lo tanto, se hace la regresión de $\dot{\boldsymbol{y}}$ sobre $\dot{\boldsymbol{X}}$ para obtener $\hat{\boldsymbol{\beta}}_{2\cdot 1} = \hat{\boldsymbol{\beta}}_2$. Resulta entonces la tabla de análisis de varianza:

Fuente	Grados de libertad (DF)	Sumas de cuadrados (SS)	Cuadrados promedios (MS)	Esperanza de los cuadrados promedios E(MS)
$\dot{\boldsymbol{X}}$	$p-1$	$SS_{\dot{\boldsymbol{X}}} = \hat{\boldsymbol{\beta}}_2'\dot{\boldsymbol{X}}'\dot{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_2 = \hat{\dot{\boldsymbol{\eta}}}'\hat{\dot{\boldsymbol{\eta}}}$	$SS_e = \boldsymbol{X}_2/(p-1)$	$\sigma^2 + \dot{\boldsymbol{\eta}}'\dot{\boldsymbol{\eta}}/(p-1)$
Error	$n-p$	$SS_e = \dot{\boldsymbol{y}}'\dot{\boldsymbol{y}} - \dot{\boldsymbol{y}}'\mathscr{P}_{\dot{X}}\dot{\boldsymbol{y}}$	$SS_e/(n-p)$	σ^2
Total	$n-1$	$S_{\dot{\boldsymbol{y}}\dot{\boldsymbol{y}}} = \Sigma(y_i - \bar{y})^2$	$\mathrm{var}(\boldsymbol{y})$	

Nótese que se ha centrado tanto los $\dot{\boldsymbol{X}}$ como $\boldsymbol{y}$. En particular, como $\bar{y}$ es la esperanza de $\boldsymbol{y}$ sin regresión, esta tabla permite de establecer si hacer la regresión segura una estimación mejor.

Siguiendo esta tabla se puede definir el coeficiente de determinación:

$$R^2 = \frac{SS_{\dot{\boldsymbol{X}}}}{S_{\dot{\boldsymbol{y}}\dot{\boldsymbol{y}}}} = 1 - \frac{SS_e}{S_{\dot{\boldsymbol{y}}\dot{\boldsymbol{y}}}}$$

Su raíz cuadrada es el *coeficiente de correlación múltiple* entre y y $\dot{\boldsymbol{X}}$, o sea entre y y $\hat{\eta}$:

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

Nótese que en el caso de solo un regresor, esto se corresponde al coeficiente de correlación ordinario

$$R = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\,\Sigma(y_i - \bar{y})^2}}$$

ya que $\boldsymbol{x}$ y $\hat{\boldsymbol{\eta}}$ son proporcionales.

Varianza, covarianza y correlación

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 $r(\boldsymbol{x}, \boldsymbol{y})$, 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 $\boldsymbol{x}$ y resulta ± 1 en el caso de relación funcional perfecta positiva o negativa.

Resulta

$$\begin{aligned}\mathrm{corr}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{\mathrm{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\mathrm{var}(\boldsymbol{x})\,\mathrm{var}(\boldsymbol{y})}} = \\ &= \frac{\Sigma_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i(x_i - \bar{x})^2\,\Sigma_i(y_i - \bar{y})^2}} = \frac{S_{\dot{x}\dot{y}}}{\sqrt{S_{\dot{x}\dot{x}}S_{\dot{y}\dot{y}}}}\end{aligned}$$

En el caso de una tabla de datos $\boldsymbol{X}$ 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

$$V(\boldsymbol{X}) = \begin{pmatrix} \mathrm{var}(\boldsymbol{x}_1) & \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2) & ... & \mathrm{cov}(\boldsymbol{x}_1, \boldsymbol{x}_p) \\ \mathrm{cov}(\boldsymbol{x}_2, \boldsymbol{x}_1) & \mathrm{var}(\boldsymbol{x}_2) & ... & \mathrm{cov}(\boldsymbol{x}_2, \boldsymbol{x}_p) \\ ... & ... & ... & ... \\ \mathrm{cov}(\boldsymbol{x}_p, \boldsymbol{x}_1) & \mathrm{cov}(\boldsymbol{x}_p, \boldsymbol{x}_2) & ... & \mathrm{var}(\boldsymbol{x}_p) \end{pmatrix}$$

y la matriz de correlación

$$C(\boldsymbol{X}) = \begin{pmatrix} 1 & \mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) & ... & \mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_p) \\ \mathrm{corr}(\boldsymbol{x}_2, \boldsymbol{x}_1) & 1 & ... & \mathrm{corr}(\boldsymbol{x}_2, \boldsymbol{x}_p) \\ ... & ... & ... & ... \\ \mathrm{corr}(\boldsymbol{x}_p, \boldsymbol{x}_1) & \mathrm{corr}(\boldsymbol{x}_p, \boldsymbol{x}_2) & ... & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & r_{12} & ... & r_{1p} \\ r_{21} & 1 & ... & r_{2p} \\ ... & ... & ... & ... \\ r_{p1} & r_{p2} & ... & 1 \end{pmatrix}$$

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

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

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

$$V(\boldsymbol{X}) = \frac{1}{n} \begin{pmatrix} S_{\dot{x}_1\dot{x}_1} & S_{\dot{x}_1\dot{x}_2} & ... & S_{\dot{x}_1\dot{x}_p} \\ S_{\dot{x}_2\dot{x}_1} & S_{\dot{x}_2\dot{x}_2} & ... & S_{\dot{x}_2\dot{x}_p} \\ ... & ... & ... & ... \\ S_{\dot{x}_p\dot{x}_1} & S_{\dot{x}_p\dot{x}_2} & ... & S_{\dot{x}_p\dot{x}_p} \end{pmatrix} = \frac{1}{n}(\boldsymbol{X'X} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}')$$

Nótese que, para hacer inferencia, en lugar de dividir por n hay que dividir para $n-1$.

Entonces, si se define la matriz diagonal de los desvíos estándar $\sigma(\boldsymbol{X}) = \mathrm{diag}(\sigma_{x_1}, \dots, \sigma_{x_p}) = \sqrt{\mathrm{diag}(\frac{1}{n}S_{\dot{x}_1\dot{x}_1}, \dots, \frac{1}{n}S_{\dot{x}_p\dot{x}_p})}$, resulta

$$C(\boldsymbol{X}) = \sigma(\boldsymbol{X})^{-1} V(\boldsymbol{X}) \sigma(\boldsymbol{X})^{-1}$$

El coeficiente de correlación parcial

Supongamos ahora que dos variables $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$, con correlación $r = \text{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ tienen una relación con una tercera variable $\boldsymbol{z}$.

Queremos conocer al vínculo que hay entre $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ *eliminando el efecto de la variable* $\boldsymbol{z}$, o sea *bajo* $\boldsymbol{z}$ *constante*.

Por esto se calcula el *coeficiente de correlación parcial* entre $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$, que se escribe como:

$$r_{x_1x_2.z} = \rho(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{z})$$

Suponiendo a todas las variables centradas, su cálculo se basa sobre la hipótesis que el efecto de $\boldsymbol{z}$ sobre $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ se manifiesta para relaciones del tipo:

$$\begin{cases} \boldsymbol{x}_1 & = \beta_1 \boldsymbol{z} + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{x}_2 & = \beta_2 \boldsymbol{z} + \boldsymbol{\varepsilon}_2 \end{cases}$$

donde $\boldsymbol{\varepsilon}_1$ y $\boldsymbol{\varepsilon}_2$ son los residuos. Entonces, en función de los residuos,

$$\begin{cases} \boldsymbol{\varepsilon}_1 = & \boldsymbol{x}_1 - \beta_1 \boldsymbol{z} \\ \boldsymbol{\varepsilon}_2 = & \boldsymbol{x}_2 - \beta_2 \boldsymbol{z} \end{cases}$$

y se definen coeficientes de *covarianza parcial* y de *correlación parcial* entre $\boldsymbol{y}_1$ y $\boldsymbol{y}_2$ como los coeficientes de covarianza y de correlación entre $\boldsymbol{\varepsilon}_1$ y $\boldsymbol{\varepsilon}_2$:

$$\text{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{z}) = \text{cov}(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) \text{ y } r_{x_1x_2.z} = \text{corr}(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2).$$

$r_{x_1x_2.z}$ se puede calcular directamente bajo los coeficientes de correlación usual. Ya que

$$r_{x_1x_2.z} = r_{e_1e_2} = \frac{\boldsymbol{e}_1'\boldsymbol{e}_2}{\sqrt{(\boldsymbol{e}_1'\boldsymbol{e}_1)(\boldsymbol{e}_2'\boldsymbol{e}_2)}}$$

y que los residuos son centrados, se obtiene

$$r_{x_1x_2.z} = \frac{(\boldsymbol{e}_1'\boldsymbol{e}_2)/n}{\sqrt{((\boldsymbol{e}_1'\boldsymbol{e}_1)/n)((\boldsymbol{e}_2'\boldsymbol{e}_2)/n)}}$$

o sea

$$r_{x_1x_2.z} = \frac{cov(\boldsymbol{e}_1, \boldsymbol{e}_2)}{\sqrt{(var(\boldsymbol{e}_1)var(\boldsymbol{e}_2))}}$$

Ahora, la covarianza parcial se escribe como:

$$\begin{aligned} \text{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{z}) &= \frac{1}{n}\boldsymbol{e}_1'\boldsymbol{e}_2 = \frac{1}{n}(\boldsymbol{x}_1 - \hat{\beta}_1\boldsymbol{z})'(\boldsymbol{x}_2 - \hat{\beta}_2\boldsymbol{z}) = \\ &= \frac{1}{n}(\boldsymbol{x}_1'\boldsymbol{x}_2 - \hat{\beta}_1\boldsymbol{z}'\boldsymbol{x}_2 - \hat{\beta}_2\boldsymbol{x}_1'\boldsymbol{z} + \hat{\beta}_1\hat{\beta}_2\boldsymbol{z}'\boldsymbol{z}) \end{aligned}$$

Si se substituyen los coeficientes de regresión para su valores estimados

$$\hat{\beta}_1 = \frac{\boldsymbol{x}_1'\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} \;\; y \;\; \hat{\beta}_2 = \frac{\boldsymbol{x}_2'\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}}$$

se obtiene luego de una simplificación:

$$\text{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{z}) = \frac{1}{n}\left(\boldsymbol{x}_1'\boldsymbol{x}_2 - \frac{(\boldsymbol{x}_1'\boldsymbol{z})(\boldsymbol{x}_2'\boldsymbol{z})}{\boldsymbol{z}'\boldsymbol{z}}\right)$$

que se puede escribir también como:

$$\operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2|\boldsymbol{z}) \;=\; \operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2) - \frac{\operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{z})\operatorname{cov}(\boldsymbol{x}_2, \boldsymbol{z})}{\operatorname{var}\boldsymbol{z}}$$

Las varianzas de los residuos se calculan de manera parecida y se encuentra, para $\boldsymbol{e}_i, i = 1, 2$:

$$\frac{1}{n}\boldsymbol{e}_i'\boldsymbol{e}_i \;=\; \operatorname{var}\boldsymbol{x}_i - \frac{\operatorname{cov}^2(\boldsymbol{x}_i, \boldsymbol{z})}{\operatorname{var}\boldsymbol{z}} \;=\; \operatorname{var}\boldsymbol{x}_i(1 - r^2(\boldsymbol{x}_i, \boldsymbol{z}))$$

y dividiendo para $\sqrt{var\boldsymbol{x}_1\, var\boldsymbol{x}_2}$ se consigue:

$$r_{x_1x_2.z} = \frac{r_{x_1x_2} - r_{x_1z}r_{x_2z}}{\sqrt{(1 - r^2{}_{x_1z})(1 - r^2{}_{x_2z})}}.$$

Caso general

Análogamente se construyen las matrizes de varianza-covarianza parcial $V(\boldsymbol{X}|\boldsymbol{Z})$ y de correlación parcial $C(\boldsymbol{X}|\boldsymbol{Z})$ entre p variables $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p$ suponiendo fijas q variables $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_q$. Por esto resulta el sistema

$$\begin{cases} \boldsymbol{x}_1 & = a_{11}\boldsymbol{z}_1 + a_{12}\boldsymbol{z}_2 + ... + a_{1q}\boldsymbol{z}_q + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{x}_2 & = a_{21}\boldsymbol{z}_2 + a_{22}\boldsymbol{z}_2 + ... + a_{2q}\boldsymbol{z}_q + \boldsymbol{\varepsilon}_2 \\ ..... & .... \\ \boldsymbol{x}_p & = a_{p1}\boldsymbol{z}_1 + a_{p2}\boldsymbol{z}_2 + ... + a_{pq}\boldsymbol{z}_q + \boldsymbol{\varepsilon}_p \end{cases}$$

y por tanto $V(\boldsymbol{X}|\boldsymbol{Z})$ y $C(\boldsymbol{X}|\boldsymbol{Z})$ van ser respectivamente la matriz de varianza-covarianza y de correlación entre las variables residuos: $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, ..., \boldsymbol{\varepsilon}_p$

$$,\begin{cases} \boldsymbol{\varepsilon}_1 & = \boldsymbol{x}_1 - (a_{11}\boldsymbol{z}_1 + a_{12}\boldsymbol{z}_2 + ... + a_{1q}\boldsymbol{z}_q) \\ \boldsymbol{\varepsilon}_2 & = \boldsymbol{x}_2 - (a_{21}\boldsymbol{z}_2 + a_{22}\boldsymbol{z}_2 + ... + a_{2q}\boldsymbol{z}_q) \\ ..... & .... \\ \boldsymbol{\varepsilon}_p & = \boldsymbol{x}_p - (a_{p1}\boldsymbol{z}_1 + \; a_{p2}\boldsymbol{z}_2 + ... + a_{pq}\boldsymbol{z}_q) \end{cases}$$

Sean $\boldsymbol{X}$ la matriz $n \times p$ cuyas lineas representan las observaciones y cuyas columnas corresponden a las p variables $\boldsymbol{x}_i$ y $\boldsymbol{Z}$ la matriz $n \times q$ cuyas columnas representan las q variables $\boldsymbol{z}_j$.

Como para la i-esima variable, el ajuste de mínimos cuadrados entre los $\boldsymbol{x}_i$ y las $\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_q$ se escribe:

$$\boldsymbol{x}_i = \hat{\beta}_{i1}\boldsymbol{z}_1 + \hat{\beta}_{i2}\boldsymbol{z}_2 + ... + \hat{\beta}_{iq}\boldsymbol{z}_q + \boldsymbol{e}_i, i = 1, ..., p$$

con $\boldsymbol{e}_i$ el vector de los residuos. Para todos los p ajustes, se puede construir la matriz $q \times p$ $\hat{\boldsymbol{B}}$ de los parámetros, y la matriz $n \times p$ $\boldsymbol{E}$ de los residuos, así que el sistema se puede escribir compactamente por

$$\boldsymbol{X}_{n\times p} = \boldsymbol{Z}_{n\times q}\hat{\boldsymbol{B}}_{q\times p} + \boldsymbol{E}_{n\times p}$$

En la matriz $\hat{\boldsymbol{B}}$ la j-ésima columna es

$$\boldsymbol{\beta}_j = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{x}_j$$

y por tanto se puede escribir $\hat{\boldsymbol{B}}$ bajo la forma

$$\hat{\boldsymbol{B}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X} \tag{1}$$

Con esta notación, la matriz $p \times p$ de varianza-covarianza parcial entre $\boldsymbol{X}$ *con* $\boldsymbol{Z}$ *constante* se puede escribir:

$$\begin{aligned} V(\boldsymbol{X}|\boldsymbol{Z}) &= \frac{1}{n}\boldsymbol{E}'\boldsymbol{E} = \frac{1}{n}(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{B}})'(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{B}}) = \\ &= \frac{1}{n}(\boldsymbol{X}'\boldsymbol{X} - \hat{\boldsymbol{B}}\boldsymbol{Z}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{Z}\hat{\boldsymbol{B}} + \hat{\boldsymbol{B}}'\boldsymbol{Z}'\boldsymbol{Z}\hat{\boldsymbol{B}}) \end{aligned}$$

y substituyendo $\hat{\boldsymbol{B}}$ con $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$ se reduce a

$$V(\boldsymbol{X}|\boldsymbol{Z}) = \frac{1}{n}(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}) \tag{2}$$

Juntando $\boldsymbol{X}$ y $\boldsymbol{Z}$ en una tabla , la matriz de varianza-covarianza se puede compartir en 4 sub-matrices de covarianza

$$V(\boldsymbol{T}) = \begin{pmatrix} V_{\boldsymbol{XX}} & V_{\boldsymbol{XZ}} \\ V_{\boldsymbol{ZX}} & V_{\boldsymbol{ZZ}} \end{pmatrix}$$

con

$$V_{\boldsymbol{XX}} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}; \quad V'_{\boldsymbol{XZ}} = V_{\boldsymbol{ZX}} = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}; \quad V_{\boldsymbol{ZZ}} = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}.$$

Resulta

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

La matriz de correlaciones parciales se calcula fácilmente empezando con $V(\boldsymbol{X}|\boldsymbol{Z})$ como se calculó la matriz de correlación ordinaria: tomando $\sigma(\boldsymbol{X}|\boldsymbol{Z}) = \sqrt{V(\boldsymbol{X}|\boldsymbol{Z})}$ se consigue

$$C(\boldsymbol{X}|\boldsymbol{Z}) = (\sigma(\boldsymbol{X}|\boldsymbol{Z}))^{-1}V(\boldsymbol{X}|\boldsymbol{Z})(\sigma(\boldsymbol{X}|\boldsymbol{Z}))^{-1}$$

Empleando la inversa de la matriz de correlación $\boldsymbol{C}^{-1}(\boldsymbol{X})$ entre las columnas de $\boldsymbol{X}$, se pueden calcular correlaciones múltiple y parciales.

Más precisamente, sean c'_{jj} su elementos diagonales.

Se demonstra que la correlación múltiple de cada columna $\boldsymbol{x}_j$ con las demás resulta ser

$$r_{j,1\ldots p} = \sqrt{1 - \frac{1}{c'_{jj}}}.$$

Además, se resulta que los coeficientes de correlación parciales entre las columnas i y j, tirando los efectos de las demás, resultan ser

$$r_{ij,1\ldots p} = \frac{-c'_{ij}}{\sqrt{c'_{ii}c'_{jj}}}.$$

Técnicas de regresión

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

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

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

El registro exhaustivo (all possible regression)

- Se estudiar todas las posibles regresiones.
- Se resultan 2^p regresiones a comparar.
- Pero, ¿Como seleccionar el mejor modelo?
- Se puede hacer recurso a unas estadísticas, el objetivo siempre siendo lo de minimizar los residuos.
- Se pueden aumentar los regresores.
- Pero, como $SS_T = SS_r + SS_e$, es fácil entender que aumentando los regresores SS_r no decrece.
- Por otro lado, el aumento de regresores seguramente puede causar un aumento de relaciones entre ellos, que va influenciar la covarianza entre los $\hat{\beta}$.
- podemos considerar las siguientes estadísticas, para n observaciones y k regresores:

- el *error cuadrático promedio (MSE, mean square error)* $MSE = MS_e = \frac{SS_e}{n-k}$, que es un estimador de σ^2 insesgado.
- el *error estándar de los residuos (RSE, residual standard error)*, su raíz cuadrada: $RSE = \sqrt{MSE}$.
- el *error promedio absoluto (MAE, minimum absolute error)* o el *MASE*, o sea *MAE escalado* (*scaled*), respectivamente: $MAE = \frac{\Sigma_i |\varepsilon_i|}{n}$, $MASE = \frac{\Sigma_i |\varepsilon_i|}{\Sigma_i |y_i - \bar{\boldsymbol{y}}|}$.
- siempre buscando el mínimo.

- el *R cuadrado* $R^2 = \frac{SS_r}{SS_t} = 1 - \frac{SS_e}{SS_t} = 1 - \frac{SS_e/n}{SS_t/n}$.
- R^2 mide la cuota de varianza debido a la regresión con respecto al total.
- $\rho = \sqrt{R^2} = \rho(\boldsymbol{y}, \hat{\boldsymbol{\eta}})$ es el coeficiente de correlación múltiple de $\boldsymbol{y}$ con los regresores $\boldsymbol{X}$.
- R^2 siempre es creciente aumentando los regresores, así que no ayuda en la selección.
- R^2 *ajustado*: $R^2_{adj} = 1 - \frac{SS_e/(n-k)}{SS_t/n}$, buscando el máximo.
- R^2_{adj} es el complemento a la cuota de varianza *estimada* sobre los residuos debido a la regresión, dependiendo también de k.

- C_p *de Mallows* (1964): se basa sobre la esperanza de la suma de cuadrados de los errores de predicción (*SSPE, sum of squared prediction errors*) $SSPE = \frac{E[\sum_i(\hat{\eta}_i - E[y_i|\boldsymbol{x}_i])^2]}{\sigma^2}$.
- σ^2 se estima como $MS_e = SS_e/(n-k)$, así que se resulta para un subconjunto de p regresores:
- $C_p = \frac{SS_{e,p}}{SS_{e,k}/(n-k)} - n + 2p$, las sumas calculadas sobre los residuos de la regresión con un subconjunto y la con el total de regresores.
- se busca el mínimo, C_p indicando la precisión de la estimación.

- *AIC criterio de información de Akaike* (1974, *an information criterion*), un estimador del error de previsión, de uso general.
- $AIC = 2k - 2\log(\hat{L})$, con k el número de parámetros y $\hat{L}$ el valor maximizado de la función de verosimilitud del modelo.
- se resulta comparando la perta de información debido al empleo de dos modelos g_1, g_2 de una función desconocida f, a través de la divergencia de Kullbach-Lebler

$$D_{KL}(f \parallel g) = \sum_x f(x) \log \frac{f(x)}{g(x)}$$

que mide el exceso de sorpresa esperado usando $g(x)$ en vez de la distribución verdadera $f(x)$.

- En el caso de la regresión, la $\log(L)$ vale

$$\log(L) - \frac{m}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i e_i^2$$

así que, estimando σ^2 con MS_e se resulta

$$AIC = 2k + n\log MSe + n(1 + \log(2\pi)) = 2k + n\log MS_e + C,$$

buscando el mínimo.

- en el caso de muestras pequeñas hay que corregir así:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}.$$

Método paso a paso: selección para adelante (forward)

- Para acelerar el proceso, hay métodos que automáticamente construyen un modelo *paso a paso*, claramente sub-optimo, a través de una secuencia de regresiones encapsuladas, obtenidas ajuntando o quitando un predictor a cada paso.
- El método de selección *para adelante* empieza con solo el α y junta al modelo a cada paso la variable que optimiza el criterio elegido.
- Es de suponer que la capacidad predictiva de la variable elegida sobrepase la penalización debido al número de variables.

La selección para atrás (backwards)

- La selección para adelante tiene el defecto que alguna variable, aunque importante para describir el fenómeno, podría no resultar incluida en el modelo final, debido a la presencia de otros predictores que esconden su influencia.
- Por esta razón unos consideran mejor la *selección para atrás*, porque con esto criterio se puede averiguar que ningún predictor importante sea olvidado.
- En este caso se empieza con todas las variable y se va tirando paso a paso la que optimiza el criterio.
- Es de suponer que su capacidad predictiva es menor de la mejoría del criterio debido a la reducción de variables.

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.
- El criterio adoptado, calculado a mismo tiempo ajuntando o quitando variables, indica a cada paso la mejor elección posible.
- Esto depende si tirando una variable otra tiene una capacidad predictiva mayor o viceversa ajuntando una, otra pierde de relevancia.

Siempre hay que considerar 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 los procesos iterativos no implica un mejor modelo posible.

Una buena manera para averiguar la calidad del modelo propuesto, en particular con respecto a un modelo con más o menos variables, que pero aparece más convincente, es de utilizar un análisis de varianza entre modelos.

Fuente	*Grados de libertad (DF)*	*Sumas de cuadrados (SS)*	*Cuadrados promedios (MS)*	*Esperanza de los cuadrados promedios E(MS)*	*F*	*Prob*
Variación	$h = k - p$	$SS_v = SS_m - SS_t$	$MS_v = SS_v/h$	$\sigma^2 + \boldsymbol{\eta}_h'\boldsymbol{\eta}_h/h$	$\frac{MS_v}{MS_m}$	π
Mod.menor	$n - p - 1$	SS_m	$MS_m = SS_m/(n-p-1)$	$\sigma^2 + \boldsymbol{\eta}_p'\boldsymbol{\eta}_p/(p-1)$		
Mod. mayor	$n - k - 1$	SS_t	$MS_t = SS_m/(n-k-1)$	$\sigma^2 + \boldsymbol{\eta}_k'\boldsymbol{\eta}_k/(k-1)$		

Un test F de Fisher con $k - p$ y $n - p - 1$ grados de libertad permite de testar la hipótesis de igualdad entre los dos modelos.

Ejemplo

El archivo ***storks.csv*** es sacado de Matthews, R. (2000), Storks Deliver Babies (p= 0.008). *Teaching Statistics*, 22: 36-38. doi: 10.1111/1467-9639.00013.

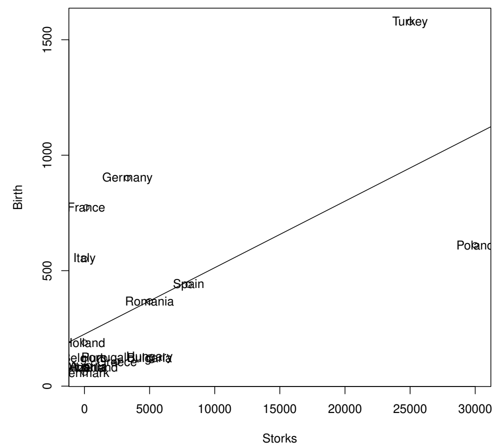
Este tiene datos de 17 países europeos de área (*Area*), pares de cigueñas (*Storks*), miliones de personas (*Humans*), y miles nacimientos (*Birth*).

De una regresión, se resulta que ¡las cigueñas traen los niños! pues


```
> attach(cic)
> rcis=lm(Birth~Storks, data=cic); summary(rcis)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.250e+02  9.356e+01   2.405   0.0295 *
Storks      2.879e-02  9.402e-03   3.063   0.0079 **
Residual standard error: 332.2 on 15 degrees of freedom
Multiple R-squared:  0.3847, Adjusted R-squared:  0.3437
F-statistic:  9.38 on 1 and 15 DF,  p-value: 0.007898

> plot(Storks,Birth)
> text(Storks,Birth,lab=rownames(cic))
> abline(rcis)
```

Vemos que tal, utilizando todas las posibles regresiones:

```
>library(leaps)  >cic1 = leaps(x=cic[,1:3], y=Birth, int=TRUE,
        method=c("r2", "adjr2",  "Cp"), nbest=1) ; cic1
$which          1          2          3
Area         TRUE       TRUE       TRUE
Storks      FALSE      FALSE       TRUE
Humans      FALSE       TRUE       TRUE
$r2     0.8510883  0.8815985  0.9048943
$adjr2  0.8411609  0.8646840  0.8829469
$Cp      7.354748   5.184304   4.000000
```

Nótese que la tablita con *TRUE* y *FALSE* ha sido traspuesta por respecto a la salida original y los nombres incluidos, porque sino no se entendia nada.

Vemos que tal, utilizando los métodos de selección para adelante, para tras y paso a paso:

```
rcic <- lm(Birth~.,data=cic)
rcic0 <- lm(Birth~1,data=cic)
cicf <- step(rcic0, scope=list(lower=rcic0, upper=rcic),
direction="forward"); summary(cicf)
cicb <- step(rcic, direction="backward"); summary(cicb)
cics <- step(rcic0,scope=list(lower=rcic0, upper=rcic),
direction="both"); summary(cics)
```

En seguida los resultados de la selección para adelante:


```
Start:  AIC=205.52
Birth ~ 1
Df Sum of Sq      RSS     AIC
+ Area     1    2289604  400603 175.15
+ Humans   1    1949210  740997 185.60
+ Storks   1    1035001 1655206 199.27
<none>                  2690208 205.52

Step:  AIC=175.15
Birth ~ Area
Df Sum of Sq    RSS    AIC
+ Humans  1      82079 318525 173.25
<none>                 400603 175.15
+ Storks  1      29807 370796 175.83

Step:  AIC=173.25
Birth ~ Area + Humans
Df Sum of Sq    RSS    AIC
+ Storks  1      62671 255854 171.53
<none>                 318525 173.25

Step:  AIC=171.53
Birth ~ Area + Humans + Storks
```

Los otros métodos dejan los mismos resultados, o sea que se necesitan todos los tres regresores. ¿Como es posible?

Vemos el modelo completo de cerca:

```
>rcic <- lm(Birth~.,data=cic)
>summary(rcic)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.824e+01  5.172e+01  -0.933   0.3680
Area         9.596e-04  3.240e-04   2.962   0.0110 *
Storks       8.965e-03  5.024e-03   1.784   0.0977 .
Humans       6.369e+00  2.635e+00   2.417   0.0311 *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 140.3 on 13 degrees of freedom
Multiple R-squared:  0.9049, Adjusted R-squared:  0.8829
F-statistic: 41.23 on 3 and 13 DF,  p-value: 6.644e-07
```

Efectivamente, se resulta que el $\hat{\beta}$ de las cigueñas no es significativo, por tanto es el caso de comparar directamente los tres mejores modelos.

```
> rcic1 <- lm(Birth~Area + Humans,data=cic); summary(rcic1);
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.137e+01  5.545e+01  -0.746 0.467933
Area         1.269e-03  2.944e-04   4.309 0.000721 ***
Humans       5.218e+00  2.747e+00   1.899 0.078318 .
Residual standard error: 150.8 on 14 degrees of freedom
Multiple R-squared:  0.8816, Adjusted R-squared:  0.8647
F-statistic: 52\cdot 12 on 2 and 14 DF,  p-value: 3.262e-07

> rcic2 <- lm(Birth~Area,data=cic) summary(rcic2)
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.7754992 56.9376784  -0.137    0.893
Area         0.0017229  0.0001861   9.259 1.36e-07 ***
Residual standard error: 163.4 on 15 degrees of freedom
Multiple R-squared:  0.8511, Adjusted R-squared:  0.8412
F-statistic: 85.73 on 1 and 15 DF,  p-value: 1.362e-07
```


También las personas parecen no ser significativas. Efectivamente, comparando directamente los modelos, se resulta

```
> anova(rcic1,rcic)
Model 1: Birth ~ Area + Humans
Model 2: Birth ~ Area + Storks + Humans
Res.  Df    RSS Df Sum of Sq      F  Pr(>F)
1     14 318525
2     13 255854  1     62671 3.1843 0.09769 .

> anova(rcic2,rcic1)
Model 1: Birth ~ Area
Model 2: Birth ~ Area + Humans
Res.  Df    RSS Df Sum of Sq      F  Pr(>F)
1     15 400603
2     14 318525  1     82079 3.6076 0.07832 .
```



```
> anova(rcic2,rcic)
Analysis of Variance Table
Model 1: Birth ~ Area
Model 2: Birth ~ Area + Storks + Humans
Res.  Df    RSS Df Sum of Sq      F  Pr(>F)
1     15 400603
2     13 255854  2    144749 3.6774 0.05424 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Se resulta que ambos *Storks* y *Humans* no ajuntan nada de relevante en el modelo básico, que solo considera la influencia del *Area* sobre *Birth*.