MAT-266: Selección del mejor conjunto de regresores

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Bondad de ajuste y selección de modelos

Objetivo:

Se desea obtener medidas de la calidad o bondad de ajuste del modelo. En efecto, si el modelo está bien ajustado, los residuos tenderán a ser pequeños.

Consideraremos los siguientes procedimientos:

- Métodos de bondad de ajuste: R^2, s^2 y C_p.
- Criterios de información.
- Validación cruzada.
- Métodos automáticos de selección de variables.

Bondad de ajuste y selección de modelos

Una medida de bondad de ajuste ampliamente usada es el coeficiente de determinación R^2, que es definido como:

$$R^2 = \{\text{cor}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})\}^2.$$

En caso de modelos de regresión con intercepto, tenemos

$$R = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(\widehat{Y}_i - \overline{Y})}{\left\{ \sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 \right\}^{1/2}}.$$

Sabemos que

$$\sum_{i=1}^n e_i = 0, \qquad \sum_{i=1}^n e_i \widehat{Y}_i = 0.$$

Además,

$$\frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \widehat{Y}_i.$$

Tenemos

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\overline{Y})(\widehat{Y}_i-\overline{Y}) &= \sum_{i=1}^{n}(Y_i-\widehat{Y}_i+\widehat{Y}_i-\overline{Y})(\widehat{Y}_i-\overline{Y}) \\ &= \sum_{i=1}^{n}(Y_i-\widehat{Y}_i)(\widehat{Y}_i-\overline{Y})+\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2 \\ &= \sum_{i=1}^{n}e_i\widehat{Y}_i-\overline{Y}\sum_{i=1}^{n}e_i+\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2 \\ &= \sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2.\end{aligned}$$

De ahí que

$$R=\left\{\frac{\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}\right\}^{1/2}$$

Bondad de ajuste y selección de modelos

Sea

$$SYY = \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

y recordando que

$$SYY = \mathsf{RSS} + SS_{\mathsf{Regr}}$$

con $\mathsf{RSS} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ y $SS_{\mathsf{Regr}} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$. Sigue que

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \frac{SS_{\mathsf{Regr}}}{SYY} = 1 - \frac{\mathsf{RSS}}{SYY}.$$

$\sqrt{R^2}$ corresponde al coeficiente de correlación múltiple entre $\boldsymbol{Y}$ y $\widehat{\boldsymbol{Y}}$, de este modo

$$0 \leq R^2 \leq 1,$$

y un modelo será bien ajustado cuando R^2 sea cercano a 1.

Bondad de ajuste y selección de modelos

Lamentablemente, R^2 no toma en cuenta la cantidad de parámetros en el modelo. Considere un modelo con k regresores, entonces podemos usar por ejemplo

$$s_k^2 = \frac{\mathsf{RSS}_k}{n-k}.$$

Otro criterio es el R^2-ajustado, dado por

$$R^2_{\mathsf{adj}} = 1 - (1 - R^2_k)\Big(\frac{n-1}{n-p}\Big)$$

Asumiendo un modelo con intercepto, tenemos $R^2_k = 1 - \mathsf{RSS}_k\,/SYY$ y de ahí que

$$R^2_{\mathsf{adj}} = 1 - \frac{\mathsf{RSS}_k}{SYY}\Big(\frac{n-1}{n-p}\Big) = 1 - \frac{s_k^2}{SYY/(n-1)}.$$

De este modo, modelos con máximo R^2_{adj} corresponden a modelos con mínimo s_k^2.

Bondad de ajuste y selección de modelos

Considere $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$ y suponga un modelo con k regresores. Luego, $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ y sea

$$\mathrm{ME} = \|\boldsymbol{\mu} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2,$$

el error de modelo. De este modo,

$$\begin{aligned}\mathrm{ME} &= \|\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{Y}\|^2 = \|\boldsymbol{\mu} - \boldsymbol{H}(\boldsymbol{\epsilon} + \boldsymbol{\mu})\|^2 = \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{\epsilon}\|^2 \\ &= \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu}\|^2 + \|\boldsymbol{H}\boldsymbol{\epsilon}\|^2 = \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} + \boldsymbol{\epsilon}^\top\boldsymbol{H}\boldsymbol{\epsilon}.\end{aligned}$$

Esto nos permite obtener

$$\begin{aligned}\mathrm{E}(\mathrm{ME}) &= \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} + \mathrm{E}(\boldsymbol{\epsilon}^\top\boldsymbol{H}\boldsymbol{\epsilon}) = \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} + \sigma^2 \operatorname{tr}\boldsymbol{H} \\ &= \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} + \sigma^2 k\end{aligned}$$

Ahora,

$$\begin{aligned}\mathrm{E}(\mathrm{RSS}_k) &= \mathrm{E}\{\boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\} = \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} + (n-k)\sigma^2 \\ &= \mathrm{E}(\mathrm{ME}) + (n - 2k)\sigma^2.\end{aligned}$$

De este modo,

$$\frac{\mathsf{E}(\mathrm{ME})}{\sigma^2} = \frac{\mathsf{E}(\mathrm{RSS}_k)}{\sigma^2} + 2k - n.$$

Si consideramos estimar σ^2 por s_p^2, entonces podemos usar el criterio C_p de Mallows, dado por

$$C_p = \frac{\mathrm{RSS}_k}{s_p^2} + 2k - n,$$

como una estimación de $\mathsf{E}(\mathrm{ME})/\sigma^2$.

Si el modelo está bien ajustado, entonces $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu}\|^2 = \boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu}$ será pequeño. Así,

$$\begin{aligned}\mathsf{E}(C_p) &\approx \frac{\mathsf{E}(\mathrm{RSS}_k)}{\sigma^2} + 2k - n = \frac{\boldsymbol{\mu}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu}}{\sigma^2} + n - k + 2k - n \\ &\approx k.\end{aligned}$$

Es decir, escogemos aquél modelo cuyo C_p sea más cercano a k.

La discrepancia de Kullback-Leibler (KL) entre las funciones de densidad $g(x)$ y $f(x)$ es dada por:

$$D_{\mathsf{KL}}(g : f) = \int \log\Big(\frac{g(x)}{f(x)}\Big) g(x)\, \mathrm{d}x = \mathsf{E}_G\Big[\log\Big(\frac{g(x)}{f(x)}\Big)\Big].$$

Propiedades de la discrepacia KL (o información):

(a) $D_{\mathsf{KL}}(g : f) \geq 0$.

(b) $D_{\mathsf{KL}}(g : f) = 0 \Leftrightarrow g(x) = f(x)$ (casi en toda parte).

Bondad de ajuste y selección de modelos

Suponga $Y_1, \ldots, Y_n$ variables aleatorias siguiendo el modelo verdadero g y denote por $\boldsymbol{\theta}_0$ el valor verdadero de $\boldsymbol{\theta}$. Considere que se ajusta el modelo candidato $f(\boldsymbol{y}; \boldsymbol{\theta})$ maximizando

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(y_j; \boldsymbol{\theta})$$

Esto sugiere escoger aquél modelo que minimice la discrepancia $D_{\mathsf{KL}}(g : f_\theta)$.

Considere la expansión en series de Taylor,

$$\log f(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}) \approx \log f(\boldsymbol{y}; \boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \frac{\partial \log f(\boldsymbol{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$$
$$+ \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \frac{\partial^2 \log f(\boldsymbol{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Notando que $\boldsymbol{\theta}_0$ minimiza $D_{\mathsf{KL}}(g : f_\theta)$, tenemos

$$\int \frac{\partial \log f(\boldsymbol{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \, g(\boldsymbol{y}) \, \mathrm{d}\, \boldsymbol{y} = \mathbf{0}$$

De ahí que

$$nD_{\mathsf{KL}}(g : f_{\widehat{\theta}}) = n \int \log\Big(\frac{g(\boldsymbol{y})}{f(\boldsymbol{y};\widehat{\boldsymbol{\theta}})}\Big) g(\boldsymbol{y})\,\mathrm{d}\,\boldsymbol{y}$$
$$\approx nD_{\mathsf{KL}}(g : f_{\theta_0}) + \frac{1}{2}\operatorname{tr}\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{\mathcal{F}}_g(\boldsymbol{\theta}_0)\},$$

con

$$\boldsymbol{\mathcal{F}}_g(\boldsymbol{\theta}) = -n \int \frac{\partial^2 \log f(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} g(\boldsymbol{y})\,\mathrm{d}\,\boldsymbol{y} = n\,\mathsf{E}_g\Big\{-\frac{\partial^2 \log f(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big\}.$$

Sea

$$\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}) = n \int \frac{\partial \log f(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial \log f(\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top} g(\boldsymbol{y})\,\mathrm{d}\,\boldsymbol{y} = n\,\mathsf{Cov}_g(\boldsymbol{U}(\boldsymbol{\theta})).$$

Observación:

Si $g(\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\theta})$, entonces $\boldsymbol{\mathcal{F}}_g(\boldsymbol{\theta}) = \boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}) = \boldsymbol{\mathcal{F}}(\boldsymbol{\theta})$.

Cuando el modelo es mal especificado, tenemos

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta})).$$

De ahí que

$$n\,\mathsf{E}_g\{D_{\mathsf{KL}}(g : f_{\widehat{\theta}})\} \approx nD_{\mathsf{KL}}(g : f_{\theta_0}) + \frac{1}{2}\,\mathrm{tr}\{\boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\}. \qquad (1)$$

Mientras que, cuando el modelo es correcto y regular, tenemos $\boldsymbol{\mathcal{F}}_g(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)$ de modo que

$$\mathrm{tr}\{\boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\} = p$$

Para estimar (1), considere

$$\ell(\widehat{\boldsymbol{\theta}}) = \ell(\boldsymbol{\theta}_0) + \{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}.$$

De este modo

$$\begin{aligned}\mathsf{E}_g\{-\ell(\widehat{\boldsymbol{\theta}})\} &= -\,\mathsf{E}\{\ell(\boldsymbol{\theta}_0) + \tfrac{1}{2}LR(\boldsymbol{\theta})\} \\ &\approx nD_{\mathsf{KL}}(g : f_{\theta_0}) - \tfrac{1}{2}\operatorname{tr}\{\boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\} - n\int g(\boldsymbol{y})\log g(\boldsymbol{y})\,\mathsf{d}\,\boldsymbol{y},\end{aligned}$$

con $LR(\boldsymbol{\theta}_0) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0)\}$ el estadístico de razón de verosimilitudes.

Un estimador de (1) es $-\ell(\widehat{\boldsymbol{\theta}}) + c$ donde c estima $\operatorname{tr}\{\boldsymbol{\mathcal{F}}_g^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{K}}_g(\boldsymbol{\theta}_0)\}$. Dos posibles elecciones de c son p y $\operatorname{tr}(\widehat{\boldsymbol{\mathcal{J}}}^{-1}\widehat{\boldsymbol{\mathcal{K}}})$ con

$$\widehat{\boldsymbol{\mathcal{J}}} = -\sum_{j=1}^{n}\frac{\partial^2 \log f(y_j;\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top},$$

$$\widehat{\boldsymbol{\mathcal{K}}} = \sum_{j=1}^{n}\frac{\partial \log f(y_j;\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\frac{\partial \log f(y_j;\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}^\top}.$$

Bondad de ajuste y selección de modelos

Lo anterior lleva a los criterios de información de Akaike y de la red:

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p,$$

$$NIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2\,\mathrm{tr}(\widehat{\boldsymbol{\mathcal{J}}}^{-1}\widehat{\boldsymbol{\mathcal{K}}}),$$

otra posibilidad es el criterio de información de Schwarz (bayesiano), dado por

$$SIC = -2\ell(\widehat{\boldsymbol{\theta}}) + p\log n.$$

Bondad de ajuste y selección de modelos

Suponga en modelo de regresión lineal con k regresores, es decir $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top$ es vector $(k+1)$-dimensional. Tenemos

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2.$$

Además, $\mathsf{RSS} = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$, $\widehat{\sigma}^2 = \mathsf{RSS}/n$. De este modo,

$$\begin{aligned} AIC &= n\log 2\pi + n\log\widehat{\sigma}^2 + \frac{1}{\widehat{\sigma}^2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + 2(k+1) \\ &= n\log 2\pi + n\log(\mathsf{RSS}/n) + \frac{n}{\mathsf{RSS}}\mathsf{RSS} + 2(k+1) \\ &= n\log(\mathsf{RSS}/n) + 2(k+1) + n(\log 2\pi + 1) \end{aligned}$$

Bondad de ajuste y selección de modelos

Suponga $(\widetilde{\boldsymbol{x}}_1, \widetilde{Y}_1), \ldots, (\widetilde{\boldsymbol{x}}_n, \widetilde{Y}_n)$ un conjunto de datos nuevos que siguen el mismo modelo que los datos de entrenamiento $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$. Podemos considerar la siguiente medida de error (de predicción)

$$\text{PE} = \frac{1}{m} \sum_{j=1}^{m} (\widetilde{Y}_i - \widetilde{\boldsymbol{x}}_i^\top \widehat{\boldsymbol{\beta}})^2,$$

donde $\widehat{\boldsymbol{\beta}}$ es calculado usando los datos de entrenamiento.

Objetivo:

Podemos subdividir el conjunto de entrenamiento en dos conjuntos disjuntos uno para obtener $\widehat{\boldsymbol{\beta}}$ y otro para medir el error.

Bondad de ajuste y selección de modelos

Considere seleccionar un subconjunto D con d observaciones, donde usamos las $n - d$ observaciones restantes para calcular PE(D).

Una alternativa es repetir el proceso, seleccionando $D_1, D_2, \ldots$, y promediar PE(D_1), PE(D_2), $\ldots$. Este método es llamado validación cruzada.

Existen diversos procedimientos para elegir tales subconjuntos:

- Métodos exhautivos: leave-p-out, leave-one-out.
- Métodos no exhautivos: k-fold, Monte Carlo CV.

La versión más simple de validación cruzada es eliminar una observación a la vez[1] y obtener $\widehat{\boldsymbol{\beta}}_{(i)}$, $i = 1, \ldots, n$. Esto lleva al error de predicción leave-one-out o CV

$$\text{CV} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)})^2$$

Observación:

La estadística

$$\text{PRESS} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2 = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)})^2$$

es conocida como suma de cuadrados (residual) de predicción.

[1] Seleccionar conjuntos D con $d > 1$ es computacionalmente intenso.

Es fácil mostrar que

$$Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{(i)} = Y_i - \boldsymbol{x}_i^\top \left[\widehat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i\right]$$
$$= Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} + \frac{e_i h_{ii}}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}.$$

De ahí que

$$\mathsf{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}{1-h_{ii}}\right)^2.$$

Observación:

Anteriormente usamos el criterio de validación cruzada generalizada para seleccionar el parámetro de sesgo k en regresión ridge

$$V(k) = \frac{1}{n}\frac{\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k\|^2}{\{\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}(k))/n\}^2}.$$

Ejemplo:

Considere datos de ventas en 15 regiones con 3 regresores X_1, X_2 y X_3, donde se obtuvo

p	Variables	RSS_p
1	—	428 144.64
2	X_1	88 473.00
	X_2	44 683.00
	X_3	32 483.00
3	X_1, X_2	43 968.00
	X_1, X_3	32 086.00
	X_2, X_3	535.00
4	X_1, X_2, X_3	273.00

Tenemos que

$$SYY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 428144.64, \qquad s^2 = \frac{\text{RSS}}{n-p} = \frac{273.00}{15-4} = 24.818$$

De este modo, los criterios de selección de modelos resultan:

Variables	R_p^2	s_p^2	R_{adj}^2	C_p	AIC	SIC
—	0.000	30581.760	0.000	17238.249	157.887	159.304
X_1	0.793	6805.615	0.777	3553.846	136.236	138.360
X_2	0.896	3437.154	0.888	1789.414	125.989	128.114
X_3	0.924	2498.692	0.918	1297.839	121.206	123.330
X_1, X_2	0.897	3664.000	0.880	1762.604	127.748	130.580
X_1, X_3	0.925	2673.833	0.913	1283.842	123.022	125.854
X_2, X_3	0.999	44.583	0.999	12.557	61.613	64.445
X_1, X_2, X_3	0.999	24.818	0.999	4.000	53.521	57.062

Selección automática de variables regresoras

```
# ajuste LS para los datos de fecundidad en Suiza
> fm <- lm(Fertility ~ ., data = swiss)
> fm

Call:
lm(formula = Fertility ~ ., data = swiss)

Coefficients:
     (Intercept)       Agriculture       Examination
         66.9152           -0.1721           -0.2580
       Education          Catholic  Infant.Mortality
         -0.8709            0.1041            1.0770
```

Selección automática de variables regresoras

```
# stepwise regression
> z <- step(fm, direction = "both")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
+ Examination       1     53.03 2105.0 190.69
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```

Selección automática de variables regresoras

```
# backward selection
> z <- step(fm, direction = "backward")
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                         2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                         2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```