IECD-325: Métodos de selección automática

Felipe Osorio

felipe.osorio@uv.cl

Suponga el modelo de regresión[1]

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \epsilon_i, \qquad i = 1, \dots, n,$$

que puede ser escrito como $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ con los supuestos habituales, y $\boldsymbol{X} \in \mathbb{R}^{n\times K}$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$.

Suponga que particionamos

$$\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2),$$

donde $\boldsymbol{X}_1 \in \mathbb{R}^{n\times p}$ y $\boldsymbol{X}_2 \in \mathbb{R}^{n\times(K-p)}$ y análogamente $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. El modelo adopta la forma,

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

Objetivo:

Deseamos identificar las variables "significativas" con coeficientes no nulos.

[1] Si $x_{i1} = 1$, para $i = 1, \dots, n$, tenemos un intercepto en el modelo.

Métodos de selección automática

Suponga que sospechamos que los coeficientes asociados a las $(K-p)$-variables en $\boldsymbol{X}_2$ son cero. Es decir, deseamos probar $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$.

En otras palabras, debemos discriminar entre 2 modelos, uno con K variables y el otro con p variables. Para esto, podemos considerar el estadístico F, dado por:

$$F = \Big(\frac{n-K}{K-p}\Big)\frac{\mathsf{RSS}_p - \mathsf{RSS}_K}{\mathsf{RSS}_K}.$$

Por otro lado, suponga que $K = p + 1$, así

$$F = (n-p-1)\frac{\mathsf{RSS}_p - \mathsf{RSS}_{p+1}}{\mathsf{RSS}_{p+1}}. \tag{1}$$

Métodos de selección automática

Consideraremos los siguientes procedimientos:

- Selección forward.
- Eliminación backward.
- Método stepwise.

Adicionalmente, un procedimiento que ha ganado popularidad es regresión lasso, el cual resuelve el problema

$$\min_{\beta} Q_1(\boldsymbol{\beta}), \qquad Q_1(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

que corresponde a un método de selección de regresores.

Selección forward (SF):

Este método añade variables regresoras una en cada vez, partiendo desde el modelo más simple, y una vez que una variable ingresa al modelo, esta **no** es retirada.

La elección que cual variable ingresa al inicio es arbitraria. Sin embargo, es usual inicializar el procedimiento con el modelo que solo tiene intercepto.

Calculamos (1) con $p = 1$ para todas las $m = K - 1$ variables regresoras restantes, y escogemos aquella tal que (1) sea la mayor.

Luego de eso, repetimos el procedimiento para $p = 2, 3, \ldots$ seleccionando en cada etapa una variable que no se había seleccionado previamente.

Es decir, añadimos la variable $\boldsymbol{x}_j$ al modelo con p-regresores, si:

$$F_j = \frac{\mathsf{RSS}_p - \mathsf{RSS}_{p+1}}{\mathsf{RSS}_{p+1} / (n - p - 1)} > F_{\mathsf{IN}},$$

donde $F_{\mathsf{IN}} = F_{1-\alpha}(1, n - p - 1)$ para algún valor de α,[2] o alternativamente $F_{\mathsf{IN}} = 2$.

[2]Se recomienda usar $0.15 \leq \alpha \leq 0.25$.

Eliminación backward (EB):

Este método se inicia con aquél que tiene K variables (i.e., el modelo más complejo) y elimina una variable en cada ocasión.

Es decir, se debe calcular (1) y debemos retirar la variable que ocasiona la menor contribución.

Es decir, eliminamos la variable $\boldsymbol{x}_j$ del modelo con p-regresores, si:

$$F_j = \frac{\mathsf{RSS}_{p-1} - \mathsf{RSS}_p}{\mathsf{RSS}_p / (n-p)} < F_{\mathsf{OUT}},$$

donde $F_{\mathsf{OUT}} = F_{1-\alpha}(1, n-p)$ para algún valor de α, o alternativamente $F_{\mathsf{OUT}} = 2$.

Método stepwise:

Este método combina selección forward y eliminación backward el cual corresponde a desarrollar un paso SF seguido por un paso EB en cada etapa.

Este algoritmo inicia con el modelo más simple y añade variables si el estadístico F asociado[3] es mayor que F_{IN} o elimina variables si es menor que F_{OUT}, siempre que $F_{\mathsf{OUT}} \leq F_{\mathsf{IN}}$.

[3] ver Ecuación (1).

Observaciones:

- El conjunto de regresores obtenidos mediante cada uno de los métodos puede ser diferente.
- Estos métodos pueden fallar en detectar el mejor subconjunto de regresores.
- El método EB asume que $\boldsymbol{X}$ tiene rango completo, en caso contrario (o si el número de potenciales regresores es muy alto), las únicas opciones factibles son SF o stepwise.
- Criterios de selección de modelos, como C_p, s_p^2 y AIC_p pueden ser usados como mecanismos de búsqueda.[4]

[4] En **R** están disponibles las funciones `step` o `stepAIC`.

usando stepAIC

```
# ajustando el modelo de regresión
fm <- lm(y ~ ., data = portland)

# cargando la biblioteca MASS
library(MASS)

# ejecutando eliminación backward
> EB <- stepAIC(fm, trace = TRUE, direction = "backward")
Start:  AIC=26.94
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS    AIC
- x3    1    0.1091 47.973 24.974
- x4    1    0.2470 48.111 25.011
- x2    1    2.9725 50.836 25.728
<none>              47.864 26.944
- x1    1   25.9509 73.815 30.576

Step:  AIC=24.97
y ~ x1 + x2 + x4

       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- x4    1      9.93  57.90 25.420
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629
```

usando stepAIC

```
# resumen del proceso
> EB$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ x1 + x2 + x3 + x4

Final Model:
y ~ x1 + x2 + x4


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                          8   47.86364 26.94429
2 - x3  1  0.10909         9   47.97273 24.97388
```

usando stepAIC

```
# modelo inicial y 'full'
> empty <- lm(y ~ 1, data = portland)
> full <- formula(y ~ x1 + x2 + x3 + x4)

# ejecutando selección forward
> SF <- stepAIC(empty, direction = "forward", scope = full)
Start:  AIC=71.44
y ~ 1

        Df Sum of Sq     RSS    AIC
+ x4     1   1831.90  883.87 58.852
+ x2     1   1809.43  906.34 59.178
+ x1     1   1450.08 1265.69 63.519
+ x3     1    776.36 1939.40 69.067
<none>               2715.76 71.444

...
```

usando stepAIC

```
# ... continuación

Step:  AIC=58.85
y ~ x4

       Df Sum of Sq    RSS    AIC
+ x1    1    809.10  74.76 28.742
+ x3    1    708.13 175.74 39.853
<none>              883.87 58.852
+ x2    1     14.99 868.88 60.629

Step:  AIC=28.74
y ~ x4 + x1

       Df Sum of Sq    RSS    AIC
+ x2    1    26.789 47.973 24.974
+ x3    1    23.926 50.836 25.728
<none>              74.762 28.742

Step:  AIC=24.97
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
<none>              47.973 24.974
+ x3    1   0.10909 47.864 26.944
```

usando stepAIC

```
# resumen del proceso
> SF$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ 1

Final Model:
y ~ x4 + x1 + x2


  Step Df   Deviance Resid. Df Resid. Dev      AIC
1                           12 2715.76308 71.44443
2 + x4  1 1831.89616        11  883.86692 58.85164
3 + x1  1  809.10480        10   74.76211 28.74170
4 + x2  1   26.78938         9   47.97273 24.97388
```

usando stepAIC

```
# ejecutando el método stepwise
> SW <- stepAIC(empty, direction = "both", scope = full)
Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444

Step:  AIC=58.85
y ~ x4

       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
<none>               883.87 58.852
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444

...
```

usando stepAIC

```
# ... continuación

Step:  AIC=28.74
y ~ x4 + x1

       Df Sum of Sq     RSS    AIC
+ x2    1     26.79   47.97 24.974
+ x3    1     23.93   50.84 25.728
<none>                74.76 28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190.92 1265.69 63.519

Step:  AIC=24.97
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- x4    1      9.93  57.90 25.420
+ x3    1      0.11  47.86 26.944
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629
```

```
# resumen del proceso
> SW$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
y ~ 1

Final Model:
y ~ x4 + x1 + x2


  Step Df    Deviance Resid. Df Resid. Dev       AIC
1                            12 2715.76308 71.44443
2 + x4  1 1831.89616         11  883.86692 58.85164
3 + x1  1  809.10480         10   74.76211 28.74170
4 + x2  1   26.78938          9   47.97273 24.97388
```