Seleção de Modelos

Gilberto A. Paula

Departamento de Estatística
IME-USP, Brasil
giapaula@ime.usp.br

$1^{\underline{o}}$ Semestre 2023

Sumário

Objetivos

Neste material serão apresentados os seguintes tópicos relacionados à seleção de modelos em Regressão Linear Múltipla:

- Todas Regressões Possíveis
- Métodos Sequenciais
- Estratégias para Seleção de Modelos
- Referências

Regressão Linear Múltipla

Supor o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_{i2}, \ldots, x_{ip}$ são valores observados de $(p-1)$ variáveis explicativas e $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$.

Regressão Linear Múltipla

Supor que $p = p_1 + p_2$ e que p_2 coeficientes são removidos no modelo de regressão linear múltipla. A remoção desses coeficientes diferentes de zero:

- Introduz viés nas estimativas do submodelo.
- Não aumenta as variâncias das estimativas do submodelo, nem do erro quadrático médio.

Sumário

Definição

Supor um total de $(p-1)$ variáveis explicativas a serem selecionadas num modelo de regressão e seja T o total de regressões possíveis. Tem-se que

$$T = 1 + \binom{p-1}{1} + \binom{p-1}{2} + \cdots + \binom{p-1}{p-1} = 2^{(p-1)}.$$

Por exemplo, se $p = 4$ (3 variáveis explicativas), haverá um total de $T = 1 + 3 + 3 + 1 = 8$ regressões possíveis.

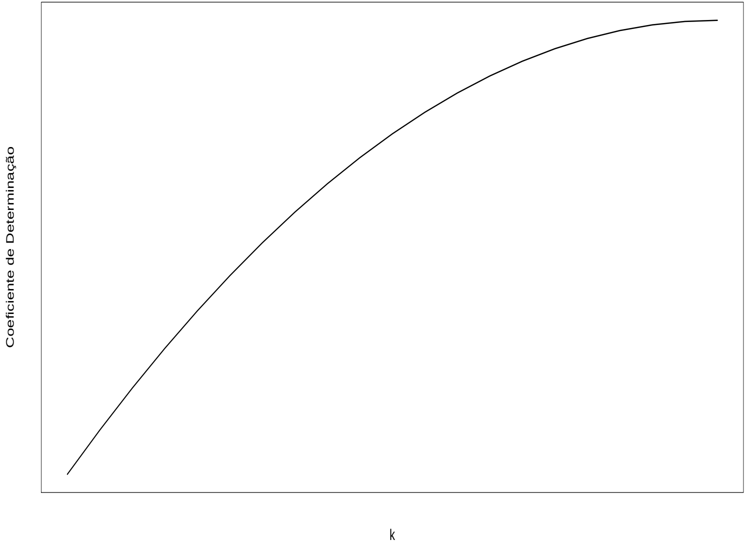
Maior R_k^2

Seja R_k^2 o coeficiente de determinação de um submodelo com k coeficientes ($(k-1)$ variáveis explicativas + o intercepto). Tem-se que

$$R_k^2 = \frac{\text{SQReg}(k)}{\text{SQT}} = 1 - \frac{\text{SQRes}(k)}{\text{SQT}}.$$

Esse critério procura um submodelo com R_k^2 alto e k pequeno.

Gráfico de R_k^2 contra k



Maior R_k^2 ajustado

Denote $\overline{R}_k^2$: R^2_{ajustado} do submodelo com k coeficientes (($k-1$) variáveis explicativas $+$ intercepto)

$$\overline{R}_k^2 = 1 - (1 - R_k^2)\frac{(n-1)}{(n-k)}.$$

Alternativamente, pode-se adotar como critério selecionar um submodelo com $\overline{R}_k^2$ alto e k pequeno. Contudo, $\overline{R}_k^2$ não necessariamente cresce com k.

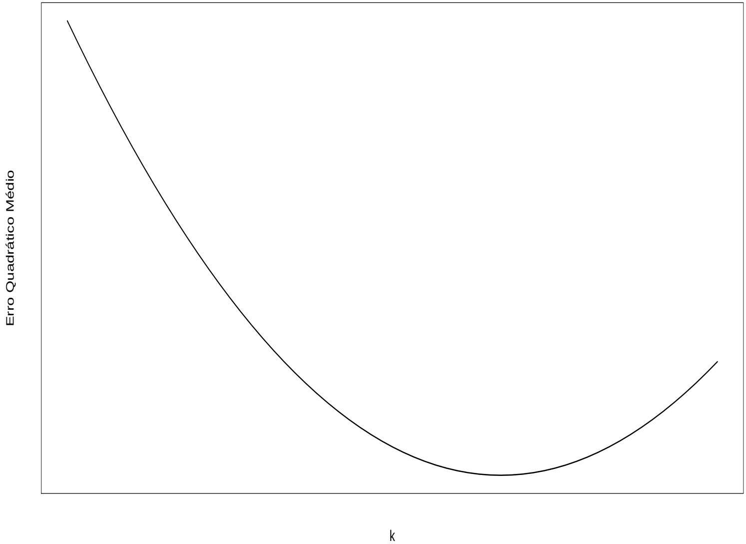
Menor s_k^2

Seja s_k^2 o erro quadrático médio de um submodelo com k coeficientes ($(k-1)$ variáveis explicativas $+$ o intercepto). Tem-se que

$$s_k^2 = \frac{\mathrm{SQRes}(k)}{n-k}.$$

Esse critério procura um submodelo com s_k^2 pequeno e k pequeno.

Gráfico de s_k^2 contra k



Relação entre $\overline{R}_k^2$ e s_k^2

Tem-se que

$$
\begin{aligned}
\overline{R}_k^2 &= 1 - \frac{(n-1)}{(n-k)}(1 - R_k^2) \\
&= 1 - \frac{(n-1)}{(n-k)} \left\{ 1 - \frac{\text{SQReg}(k)}{\text{SQT}} \right\} \\
&= 1 - \frac{(n-1)}{(n-k)} \frac{\text{SQRes}(k)}{\text{SQT}} \\
&= 1 - \frac{(n-1)}{\text{SQT}} \frac{\text{SQRes(k)}}{n-k} \\
&= 1 - \frac{(n-1)}{\text{SQT}} s_k^2.
\end{aligned}
$$

Relação entre $\overline{R}_k^2$ e s_k^2

Portanto, tem-se que

$$\overline{R}_k^2 = 1 - \frac{(n-1)}{\text{SQT}} s_k^2.$$

Assim, minimizar s_k^2 é equivalente a maximizar $\overline{R}_k^2$.

Critério C_k de Mallows

O critério de Mallows está relacionado com o erro quadrático médio do i-ésimo valor ajustado $\widehat{Y}_i$ do submodelo com k coeficientes

$$\mathrm{E}\{\widehat{Y}_i - \mathrm{E}(Y_i)\}^2 = \mathrm{Var}(\widehat{Y}_i) + \{\mathrm{E}(\widehat{Y}_i) - \mathrm{E}(Y_i)\}^2.$$

A soma dos vieses ao quadrado do submodelo com k coeficientes fica dada por

$$\{\text{Viés}(k)\}^2 = \sum_{i=1}^{n}\{\mathrm{E}(\widehat{Y}_i) - \mathrm{E}(Y_i)\}^2,$$

em que $\mathrm{E}(Y_i)$ denota o valor esperado do modelo correto.

Critério C_k de Mallows

O erro quadrático médio padronizado do submodelo com k coeficientes fica dado por

$$\text{EQM}(k) = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\{\text{E}(\widehat{Y}_i) - \text{E}(Y_i)\}^2 + \sum_{i=1}^{n}\text{Var}(\widehat{Y}_i)\right].$$

Usando o resultado que $\sum_{i=1}^{n}\text{Var}(\widehat{Y}_i) = k\sigma^2$ obtém-se

$$\text{EQM}(k) = \frac{\{\text{Viés}(k)\}^2}{\sigma^2} + k.$$

Critério C_k de Mallows

Por outro lado

$$\mathrm{E}\{\mathrm{SQRes}(k)\} = \{\mathrm{Viés}(k)\}^2 + (n-k)\sigma^2.$$

Portanto, o erro quadrático médio padronizado assume a forma

$$\mathrm{EQM}(k) = \frac{\mathrm{E}\{\mathrm{SQRes}(k)\}}{\sigma^2} - n + 2k.$$

Deve-se escolher submodelos com EQM(k)'s pequenos.

Critério C_k de Mallows

A estatística C_k é definida por

$$C_k = \frac{\text{SQRes}(k)}{\widehat{\sigma}^2} - n + 2k.$$

em que $\widehat{\sigma}^2$ deve ser obtido de um modelo bem ajustado. Sob viés zero tem-se que

$$\text{E}(C_k|\text{Viés} = 0) = \frac{(n-k)\sigma^2}{\sigma^2} - n + 2k = k.$$

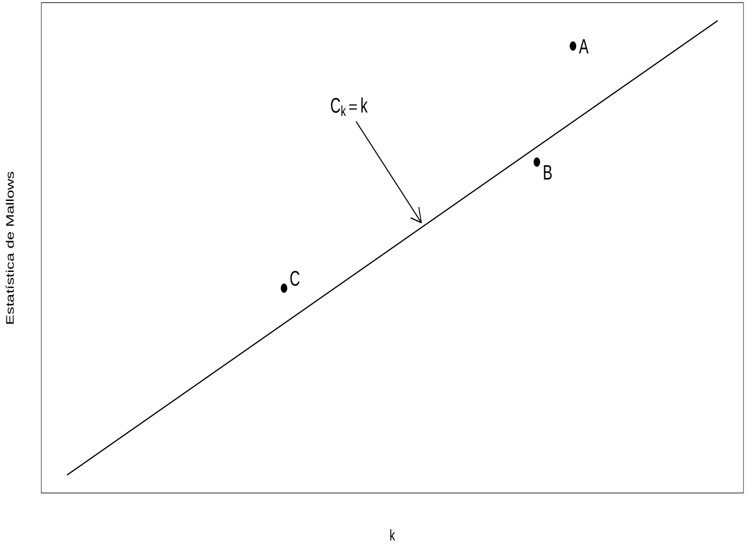
Critério C_k de Mallows

Portanto, deve-se escolher submodelos com C_k pequenos tais que

$$C_k \cong k.$$

Para um mesmo k, submodelos com $C_k < k$ têm uma SQRes menor enquanto submodelos com $C_k > k$ têm uma SQRes maior.

Descrição da Reta $C_k = k$ e Modelos A, B e C



Critério C_k de Mallows

Na figura anterior o modelo A é o pior modelo, tem C_k alto e viés alto. O modelo B tem um C_k menor e viés pequeno. Já o modelo C tem um viés um pouco maior do que o modelo B, porém um C_k bem menor, assim poderia ser o modelo escolhido.

Estatística Press

O critério Press consiste em escolher o submodelo com o menor valor para a estatística

$$\text{Press}_k = \sum_{i=1}^{n} \{y_i - \widehat{y}_{(i)}\}^2,$$

em que $\widehat{y}_{(i)}$ denota o valor predito para y_i do ajuste do submodelo com k coeficientes sem a i-ésimo observação.

Estatística Press

Desde que $\widehat{y}_{(i)} = \mathbf{x}_i^\top \widehat{\beta}_{(i)}$, em que

$$\widehat{\beta}_{(i)} = \widehat{\beta} - \frac{r_i}{(1 - h_{ii})}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{x}_i.$$

obtém-se após algumas manipulações algébricas o seguinte resultado:

$$\text{Press}_k = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2,$$

em que r_i e h_{ii} denotam, respectivamente, o i-ésimo resíduo ordinário e i-ésima medida de alavanca do submodelo com k coeficientes.

Todas as Regressões Possíveis

A fim de selecionar um submodelo usando os critérios: R_k^2 maior, s_K^2 menor, $C_k \cong k$ e pequeno e menor Press_k, deve-se ajustar todas as $T = 2^{(p-1)}$ regressões possíveis e selecionar um submodelo seguindo os 4 critérios descritos.

Sumário

Método de Akaike

Seja L(θ) o logaritmo da função de verossimilhança de um modelo de regressão com p coeficientes a serem estimados. O método de Akaike consiste em maximimizar L(θ) com o menor número p de coeficientes da regressão. Isso é equivalente a minimizar a função penalizada abaixo

$$\text{AIC} = -2\text{L}(\widehat{\theta}) + 2p.$$

No caso de regressão linear múltipla tem-se que

$$\text{AIC} = n \log\left(\frac{SQRes}{n}\right) + 2p.$$

Método de Schwartz

Similarmente ao método de Akaike o método de Schwartz consiste em maximimizar $\mathrm{L}(\boldsymbol{\theta})$ com o menor número p de coeficientes da regressão, porém com uma penalização diferente. O método é equivalente a minimizar a função abaixo

$$\mathrm{BIC} = -2\mathrm{L}(\widehat{\boldsymbol{\theta}}) + p\log(n).$$

No caso de regressão linear múltipla tem-se que

$$\mathrm{BIC} = n\log\left(\frac{SQRes}{n}\right) + p\log(n).$$

Método LASSO

O método LASSO é utilizado para a seleção de variáveis explicativas (na forma padronizada) eliminando coeficientes da regressão cujas estimativas estejam próximas de zero. No contexto de mínimos quadrados o método é equivalente a minimizar a função abaixo

$$\mathrm{S}(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=2}^{p} |\beta_j|,$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ e $\lambda \geq 0$ é o parâmetro de penalização. Quando $\lambda = 0$ tem-se o método de mínimos quadrados e quando $\lambda \to \infty$ todos os coeficientes tendem a zero.

Passo 1

Ajustar todas as regressões possíveis com apenas 1 variável explicativa. Isto é ajustar as regressões

$$y_i = \beta_1 + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 2, \ldots, p$.

Passo 1

Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$F_j = \frac{\text{SQReg}(x_j)}{s^2(x_j)} \overset{\text{H}_0}{\sim} F_{1,(n-2)}.$$

Denote P_j: valor-P do teste.

Passo 1

Seja

$$P_{\min} = \min\{P_2, \ldots, P_p\}.$$

Se $P_{\min} \leq P_E$ então a variável explicativa correspondente entra no modelo. Supor que X_2 entra no modelo.

Passo 2

Ajustar todas as regressões possíveis com apenas X_2 mais uma variável explicativa. Isto é ajustar as regressões

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 3, \ldots, p$.

Passo 2

Testar $\mathrm{H}_0 : \beta_j = 0$ contra $\mathrm{H}_1 : \beta_j \neq 0$ e obter a estatística

$$F_j = \frac{\mathrm{SQReg}(x_j|x_2)}{s^2(x_2, x_j)} \overset{\mathrm{H}_0}{\sim} F_{1,(n-3)}.$$

Denote P_j: valor-P do teste.

Passo 2

Seja

$$P_{\text{min}} = \min\{P_3, \ldots, P_p\}.$$

Se $P_{\text{min}} \leq P_E$ então a variável explicativa correspondente entra no modelo. Supor que X_3 entra no modelo.

Passo 3

Ajustar todas as regressões possíveis com apenas X_2 e X_3 mais uma variável explicativa. Isto é ajustar as regressões

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_j x_{ij} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$, para $i = 1, \ldots, n$ e $j = 4, \ldots, p$.

Passo 2

Testar $\mathrm{H}_0 : \beta_j = 0$ contra $\mathrm{H}_1 : \beta_j \neq 0$ e obter a estatística

$$F_j = \frac{\mathrm{SQReg}(x_j | x_2, x_3)}{s^2(x_2, x_3, x_j)} \overset{\mathrm{H}_0}{\sim} F_{1,(n-4)}.$$

Denote P_j: valor-P do teste.

Passo 3

Seja

$$P_{\min} = \min\{P_4, \ldots, P_p\}.$$

Se $P_{\min} \leq P_E$ então a variável explicativa correspondente entra no modelo. Se $P_{\min} > P_E$ parar o processo, nenhuma variável entra no modelo.

Passo 1

Ajustar a regressão com todas as variáveis explicativas. Isto é ajustar o seguinte modelo:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

Passo 1

Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$F_j = \frac{\text{SQReg}(x_j|\text{demais})}{s^2(x_2, \ldots, x_p)} \overset{\text{H}_0}{\sim} F_{1,(n-p)}.$$

Denote P_j: valor-P do teste, para $j = 2, \ldots, p$.

Passo 1

Seja

$$P_{\text{max}} = \max\{P_2, \ldots, P_p\}.$$

Se $P_{\text{max}} \geq P_S$ então a variável explicativa correspondente sai do modelo. Supor que X_2 sai do modelo.

Passo 2

Ajustar a regressão sem a variável explicativa X_2. Isto é ajustar o seguinte modelo:

$$y_i = \beta_1 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

Passo 2

Testar $\text{H}_0 : \beta_j = 0$ contra $\text{H}_1 : \beta_j \neq 0$ e obter a estatística

$$F_j = \frac{\text{SQReg}(x_j|\text{demais})}{s^2(x_3, \ldots, x_p)} \overset{\text{H}_0}{\sim} F_{1,(n-p-1)}.$$

Denote P_j: valor-P do teste, para $j = 3, \ldots, p$.

Passo 2

Seja

$$P_{\text{max}} = \max\{P_3, \ldots, P_p\}.$$

Se $P_{\text{max}} < P_S$ o processo é terminado, nenhuma variável sai do modelo.

Descrição

O método stepwise é uma combinação dos métodos forward e backward.

Passo 1

Ajustar todas as regressões com apenas uma variável explicativa, além do intercepto. Verificar se alguma variável explicativa entra no modelo. Supor que X_2 entrou no modelo.

Passo 2

Ajustar todas as regressões com X_2 mais uma variável explicativa, além do intercepto. Verificar se alguma variável explicativa entra no modelo. Supor que X_3 entrou no modelo. Verificar se X_2 sai do modelo dado que X_3 está no modelo.

Passo 3

O processo stepwise deve continuar até que não seja possível incluir nenhuma variável no modelo, nem retirar nenhuma variável do modelo.

Critérios de Parada

Não há um consenso na área de regressão a respeito de critérios de parada para os processos sequenciais forward, backward e stepwise. Alguns critérios mais utilizados:

- Usar $F_E = F_S = 4$ que equivale aproximadamente a usar $P_E = P_S = 0,05$.
- Considerar $P_E = 0,25$ e $P_S = 0,10$, ou mesmo ainda ser mais flexível com $P_E = P_S = 0,15$.

Sumário

Considerações

A seleção de modelos é uma combinação de técnicas e bom senso e não há uma receita pronta para seleção de modelos a partir de um conjunto de variáveis explicativas. Em Montgomery, Peck e Vining (2021, Seção 10.3) há uma longa discussão a respeito de possíveis estratégias para seleção de modelos através dos critérios propostos. A seguir sintetizamos algumas dessas estratégias.

Todas as Regressões Possíveis

Quando o número de variáveis explicativas é relativamente pequeno pode ser factível ajustar todas as regressões possíves e selecionar algumas candidatas segundo os critérios R_k^2 maior, s_K^2 menor, $C_k \cong k$ e pequeno e menor Press_k. Para as regressões selecionadas sugere-se fazer uma análise de diagnóstico e levar em conta aspectos como a importância, custo e facilidade de interpretação das variáveis explicativas bem como da capacidade de predição do modelo.

Métodos Sequencias

Os métodos sequenciais forward, backward e stepwise são recomendados quando há um número médio ou alto de variáveis explicativas, contudo exigem os níveis de significância de entrada e saída das variáveis explicativas. Já os métodos AIC, BIC e LASSO são mais recomendados quando há um grande número de variáveis explicativas no sentido de se fazer uma pré-seleção de variáveis explicativas sem a necessidade de estabelecer níveis de significância. Todos os métodos sequencias podem ser combinados com o ajuste de todas as regressões possíveis. Recomenda-se análise de diagnóstico antes de escolher o modelo final.

Sumário

Referências

- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021). *Introduction to Linear Regression Analysis, 6th Edition*. Hoboken: Wiley.