MLG

Curso de Modelos Lineares Generalizado - DEST/UFMG
Marcos Oliveira Prates

25 de setembro de 2017

Minímos Quadrados Generalizado

- Suponha o modelo geral

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,\ \text{Var}(\varepsilon) = \sigma^2\mathbf{V}$$

  onde $\mathbf{V}$ é uma matriz $n \times n$ positiva definida (p.d.).
- Quando $\mathbf{V} = \mathbf{I}$ voltamos ao modelo linear tradicional.
- Como $\mathbf{V}$ é p.d. existe uma matriz $\mathbf{K}_{n\times n}$ tal que $\mathbf{K}\mathbf{K}' = \mathbf{V}$ e $r(\mathbf{K}) = n$.
- Defina $\mathbf{Z} = \mathbf{K}^{-1}\mathbf{Y}$, $\mathbf{B} = \mathbf{K}^{-1}\mathbf{X}$ e $\eta = \mathbf{K}^{-1}\varepsilon$.
- Como $r(\mathbf{X}) = r \leq p < n$ pela propriedade de posto temos que $r(\mathbf{B}) = r$.

- Além disso temos

$$E(\eta) = E(\mathbf{K}^{-1}\varepsilon) = \mathbf{K}^{-1}E(\varepsilon) = 0$$

$$\begin{aligned} \text{Var}(\eta) &= \text{Var}(\mathbf{K}^{-1}\varepsilon) = \mathbf{K}^{-1}\text{Var}(\varepsilon)\mathbf{K}'^{-1} \\ &= \sigma^2\mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1} = \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}'\mathbf{K}'^{-1} \\ &= \sigma^2\mathbf{I} \end{aligned}$$

- Portanto considere o modelo linear $\mathbf{Z} = \mathbf{B}\beta + \eta,\ \text{Var}(\eta) = \sigma^2\mathbf{I}$ obtemos a fórmula de regressão linear tradicional.
- Logo SSE$= (\mathbf{Z} - \mathbf{B}\beta)'(\mathbf{Z} - \mathbf{B}\beta)$ e podemos minimiza-lo com respeito a β.

- Propriedades do Resíduo
- Logo

$$\begin{aligned}\tilde{\beta}_{GLS} &= (\mathbf{B}'\mathbf{B})^{-}\mathbf{B}'\mathbf{Z} \\ &= (\mathbf{X}'\mathbf{K}'^{-1}\mathbf{K}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{K}'^{-1}\mathbf{K}^{-1}\mathbf{Y} \\ &= [\mathbf{X}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{X}]^{-}\mathbf{X}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\end{aligned}$$

- Quando $r(\mathbf{X}) = p$ temos

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

- Assim $\hat{\beta}$ é o melhor estimador não viesado para o modelo

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,\ \text{Var}(\varepsilon) = \sigma^2\mathbf{V}$$

- Nesse caso podemos calcular o estimador de σ^2_{GLS} como

$$\begin{aligned}
\hat{\sigma}^2_{GLS} &= \frac{1}{n-r}(\mathbf{Z} - \mathbf{B}\tilde{\beta}_{GLS})'(\mathbf{Z} - \mathbf{B}\tilde{\beta}_{GLS}) \\
&= \frac{1}{n-r}(\mathbf{Y} - \mathbf{X}\tilde{\beta}_{GLS})'(\mathbf{K}\mathbf{K}')^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}_{GLS})' \\
&= \frac{1}{n-r}(\mathbf{Y} - \mathbf{X}\tilde{\beta}_{GLS})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}_{GLS})' \\
&= \frac{1}{n-r}\text{SSE}_{GLS}
\end{aligned}$$

Exemplo AR(1)

- Um exemplo de modelo linear que utiliza mínimos quadrados generalizados é o modelo tradicional AR(1).
- Seja o modelo

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N(0, \mathbf{V})$$

onde

$$\mathbf{V} = \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix},$$

com $|\rho| < 1$ conhecido.

- Dessa forma podemos encontrar $\mathbf{K}$ tal que $\mathbf{K}\mathbf{K}' = \mathbf{V}$ e o modelo é agora similar

$$\mathbf{Z} = \mathbf{B}\beta + \eta, \ \text{Var}(\eta) = \sigma^2\mathbf{I}$$

- Assim podemos ajustar o modelo linear em $\mathbf{Z}$ que é equivalente ao ajuste do modelo de mínimos quadrados generalizados.
- Portanto, modelos AR(1) com ρ conhecidos pode ser ajustado como um modelo de minímos quadrados generalizados.

- Quando

$$\mathbf{V}\sigma^2 = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix},$$

e supondo que $\sigma_i^2 = \sigma^2 / w_i$, onde w_i é conhecido, então o método de mínimos quadrados generalizados se reduz ao método de mínimos quadrados ponderados

- Nesse caso, seja

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}.$$

  logo, $\mathbf{V} = W^{-1}$.

- Com alguma álgebra temos

$$\mathbf{K}^{-1} = \mathbf{W}^{1/2} = \begin{pmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{w_n} \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} \sqrt{w_1}\, Y_1 \\ \sqrt{w_2}\, Y_2 \\ \vdots \\ \sqrt{w_n}\, Y_n \end{pmatrix}.$$

- Portanto, a soma dos resíduos dos quadrados é

$$\eta'\eta = \varepsilon'\mathbf{V}^{-1}\varepsilon = \sum_{i=1}^{n} w_i(Y_i - \mathbf{X}_i'\beta)^2,$$

onde $\mathbf{X}_i'$ é a i-ésima linha da matriz $\mathbf{X}$.

- Os resídos dos mínimos quadrados ponderados são dados por

$$e_{w,i} = \sqrt{w_i}\,(Y_i - \widehat{Y}_i)$$

para $i = 1, 2, \ldots, n$.

- No modelo de regressão simples temos

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

para $i = 1, 2, \ldots, n$. Seja

$$\overline{Y}_w = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i} \text{ e } \overline{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}.$$

Então,

$$b_1 = \frac{\sum_{i=1}^n w_i (Y_i - \overline{Y}_w)(X_i - \overline{X}_w)}{\sum_{i=1}^n w_i (X_i - \overline{X}_w)^2},$$

e

$$b_0 = \overline{Y}_w - b_1 \overline{X}_w.$$

- Escolhas de w_i
  - (i) Se $\text{Var}(Y_i) = X_i\sigma^2$ $(X_i > 0)$, então $w_i \propto 1/X_i$; e
  - (ii) Se $\text{Var}(Y_i) = X_i^2\sigma^2$ $(X_i^2 > 0)$, então $w_i \propto 1/X_i^2$.
- Suponha que os dados são coletados da seguinte forma:

Resposta		$\overline{Y}_i$
Tamanho da amostra		n_i
Variância amostral		s_i^2
Escolhas de w_i	$w_i \propto n_i$	se $\text{Var}(\overline{Y}_i) = \sigma^2/n_i$
	$w_i \propto n_i/s_i^2$	se $\text{Var}(\overline{Y}_i) = \sigma_i^2/n_i$
	$w_i \propto 1/s_i^2$	se $\text{Var}(\overline{Y}_i) = \sigma_i^2/m$

- Suponha que os dados são divididos em a grupos: $\{(Y_{ij}, \mathbf{X}_{ij}),\ i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, a\}$. Ajustamos o modelo

$$E(Y_{ij}) = \mathbf{X}'_{ij}\beta,$$

onde $\mathbf{X}'_{ij} = (1, X_{ij1}, \ldots, X_{ijk})$. Assuma que $\text{Var}(Y_{ij}) = \sigma_j^2$. Seja s_j^2 a variância amostral de $\{Y_{ij},\ i = 1, 2, \ldots, n_j\}$. Então uma possível escolha para w_i é

$$w_{ij} \propto n_j / s_j^2.$$

Seleção de Modelo

- Existem duas razões porque as vezes não nos satisfazemos com o estimador de mínimos quadrados
  1. Precisão da Predição: Os estimadores de mínimos quadrados normalmente são não viciados mas possuem grande variabilidade.
  2. Interpretação: Com um grande número de preditores, gostariamos de escolher um subconjunto que exiba os fatores prinicipais. Para entender o problema de forma geral estamos dispostos a sacrificar pequenos detalhes.

Seleção de Modelo

- Iremos discutir 4 métodos
  1. Melhor subconjunto
  2. Forward Selection
  3. Bacward Selection
  4. Stepwise Selection

Melhor Subconjunto

- O método de melhor subcojunto baseia-se em minimizar o SSE, ou equivalentemente maximizar o $R^2 = 1 - \frac{SSE}{SST}$.
- Para isso faz-se uma busca exaustiva entre as $\binom{p}{k}$, onde p é o número total de preditores e k é o número de preditores no modelo
- O método de subconjunto não pode ser usado para determinar k, pois o SSE (R^2) sempre diminui (aumenta) com o acréscimos de preditores.

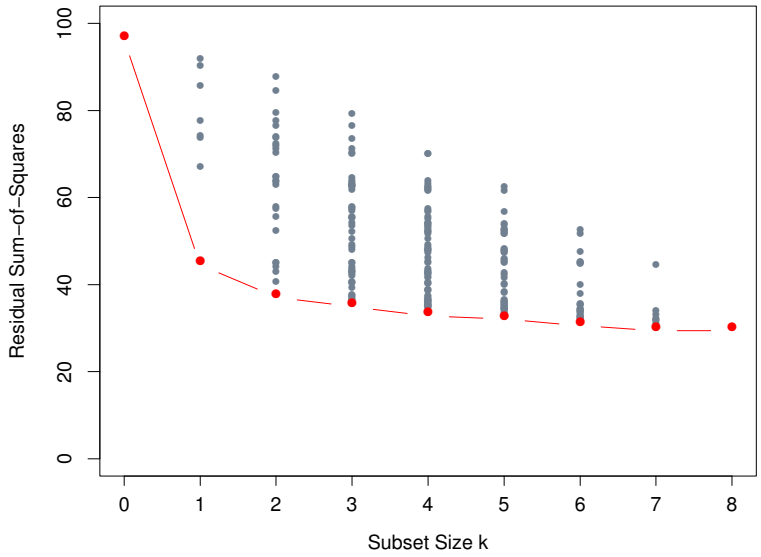

FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

- Dessa forma fixando-se k pode-se determinar o “melhor” modelo sob o critério SSE (R^2)
- Uma opção é utilizar o princípio da parsimonia para determinar k

Akaike Information Criterion

1. A “Akaike information criterion” (AIC) é uma medida de qualidade ajuste para modelos estatísticos
2. O AIC é baseado no conceito de entropia, que tenta medir a perda de informação dado um modelo para descrever a realidade.
3. Pode se dizer que o AIC é uma troca entre o viés e a variabilidade do modelo construido, ou de forma mais geral, uma troca entre a complexidade e a precisão do modelo.

$$\mathrm{AIC} = 2k - 2\ln(L)$$

onde L é a função de verossimilhança e k é o número de parâmetros no modelo.

4. Portanto, o AIC penaliza pela complexidade do modelo (k) e mede a precisão através da funçao de verossimilhança (L).

Forward Selection

1. Determine todos os fatores que serão usados, o modelo completo;
2. A primeira variável que entra no modelo é aquela no qual o AIC foi minimizado. Adiciona a variável caso o AIC seja inferior ao AIC do modelo com somente o intercepto;
3. Encontre os AIC referentes a adição de uma variável no presente modelo;
4. Se tiver algum fator que diminua o AIC, o inclua e retorne para 3. Caso contrário pare.

Backward Selection

1. Determine o modelo completo;
2. Encontre o AIC referentes a remoção de uma variável por vez no presente modelo;
3. Retire do modelo o fator no qual a remoção diminua mais o AIC e retorne para 2. Caso contrário pare.

Stepwise Selection

- O método stepwise é uma mistura entre os métodos forward e o backward.
- Basicamente o método permite que se adicione e/ou retire fatores passo a passo.

1. Defina o modelo completo;
2. Determine se a remoção de algum fator reduz o AIC
3. Encontre o AIC para a remoção das variáveis presentes e o AIC para a adição das variáveis que já foram removidos;
4. Inclua ou retire o fator que reduz o AIC e volte para 3. Caso contrário pare.

- Esses modelos são gulosos e podem parecer sub-ótimos se comparado ao melhor subcojunto.
- Porém existem razões para escolher esses métodos
  1. Não é preciso determinar k;
  2. É computacionalmente eficiente se comparado ao método do subcojunto;
  3. Estatísticamente os métodos gulosos fazem uma busca mais restrita nos modelos; incluindo um viés maior mas reduzindo a variabilidade;

Regressão Não-Linear

- Nem sempre a suposição da relação de linearidade entre a resposta e os coeficientes é adequada
- Nesse caso, podemos propor uam função $f(\beta, \mathbf{X})$ não linear em β.
- Exemplos de regressões não lineares são modelo TRI de 3 parâmetros e modelos físicos
- A classe de regressão não linear tem regressão linear como caso especial quando $f(\beta, \mathbf{X}) = \mathbf{X}\beta$ mas oferece uma maior flexibilidade nos modelos

Regressão Não-Linear

- Modelos não lineares podem ser escolhidos de diversas formas:
  - $f(\beta, X) = \beta_1 \exp(-\beta_2 X)$
  - $f(\beta_1, X) = X^{\beta_1}$
  - $f(\beta, X) = \frac{\beta_1}{1+e^{\beta_2+\beta_3 X}}$
- Dessa forma, um modelo não linear pode ser respresentado por

$$Y = f(\beta, \mathbf{X}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- Esses modelos tendem a produzir um bom ajuste com um menor número de parâmetros do que modelos lineares

- Temos portanto que $Y \sim N(f(\beta, \mathbf{X}), \sigma^2)$ e portanto

$$L(\beta) = (2\pi\sigma^2)^{-n/2} \exp \frac{-\sum_{i=1}^{n} [y_i - f(\beta, \mathbf{X}_i)]^2}{2\sigma^2}$$

- Dessa forma para maximizar a verossimilhança basta minimizar

$$SSE(\beta) = \sum_{i=1}^{n} [y_i - f(\beta, \mathbf{X}_i)]^2$$

- Diferenciando $SSE(\beta)$ com respeito a β temos

$$\frac{\partial SSE(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} [Y_i - f(\beta, \mathbf{X}_i)] \frac{\partial f(\beta, \mathbf{X}_i)}{\partial \beta}$$

- Em geral as esquações obtidas diferenciado $SSE(\beta)$ são não lineares
- Isso nos deixa com as seguintes opções
  - utilizar métodos numéricos para resolver as equações
  - "linearizar" $f(\beta, \mathbf{X})$ através de uma expansão de Taylor de primeira ordem e iterar
  - utilizar algoritmos EM
- Vamos focar nossas estimativas baseando na linearização

- Usando expansão de Taylor ao redor de β^0 temos que

$$f(\beta, \mathbf{X}) = f(\beta^0, \mathbf{X}) + \sum_{j=1}^{p} \left(\frac{\partial f(\mathbf{X}, \beta)}{\partial \beta_j} \right) \Big|_{\beta=\beta^0} (\beta_j - \beta_j^0)$$

- Podemos reescrever $f(\beta, \mathbf{X})$ aproximadamente como

$$f(\beta, \mathbf{X}) = f(\beta^0, \mathbf{X}) + \mathbf{Z}|_{\beta^0} . \triangle \beta^0$$

- Portanto

$$\mathbf{Y} \approx f(\beta^0, \mathbf{X}) + \mathbf{Z}|_{\beta^0} . \triangle \beta^0 + \varepsilon$$

- Assim temos que

$$\mathbf{Y} - f(\beta^0, \mathbf{X}) \approx \mathbf{Z}|_{\beta^0} . \triangle \beta^0 + \varepsilon$$

- Portanto podemos estimar

$$\hat{\triangle}\beta^0 = \left(\mathbf{Z}'|_{\beta^0}\mathbf{Z}|_{\beta^0}\right)^{-1}\mathbf{Z}'|_{\beta^0}(\mathbf{Y} - f(\beta^0, \mathbf{X}))$$

- Logo

$$\beta^{(k+1)} = \beta^{(k)} + \left(\mathbf{Z}'|_{\beta^k}\mathbf{Z}|_{\beta^k}\right)^{-1}\mathbf{Z}'|_{\beta^k}(\mathbf{Y} - f(\beta^k, \mathbf{X}))$$

- Assim itermaos até β^{k+1} convergir
- Possíveis critérios de parada
  - $||\beta^{k+1} - \beta^k|| < \varepsilon$
  - $\beta^{k+1}/\beta^k - 1 < \varepsilon$
  - $L(\beta^{k+1}) - L(\beta^k) < \varepsilon$
- os pacotes *car*, *nlsmsn*, *nlme*, *gln*, *nrwlr*, *nlmrt*, ... são alguns exemplos de pacotes no R para ajuste não lineares.