Seleção de variáveis

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 20 de maio de 2024

Roteiro

Roteiro

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ é conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor de parâmetros desconhecidos a serem estimados, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2, também desconhecida, a ser estimada.

Modelo de regressão linear

Suponham que $Y_1, Y_2, \ldots, Y_n$ tais que

$$Y_\ell = \boldsymbol{x}_\ell^\top \boldsymbol{\beta} + \varepsilon_\ell, \ \ell = 1, 2, \ldots, n, \tag{1}$$

em que $\boldsymbol{x}_\ell = (x_{\ell 1}, x_{\ell 2}, \ldots, x_{\ell p})^\top$ é conhecido, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é um vetor de parâmetros desconhecidos a serem estimados, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2, também desconhecida, a ser estimada.

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \tag{2}$$

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ é a matriz de planejamento e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Forma matricial

A Equação (1) pode ser escrita de forma matricial, da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{2}$$

em que $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ é a matriz de planejamento e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, com

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \text{ e } \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Suposições

Resumindo,

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros:
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Construção do modelo

Na construção de um modelo de regressão, há dois objetivos conflitantes:

1. Selecionar o **maior número** de variáveis regressoras que contenham "toda" a informação da relação com a variável resposta;

Construção do modelo

Na construção de um modelo de regressão, há dois objetivos conflitantes:

1. Selecionar o **maior número** de variáveis regressoras que contenham "toda" a informação da relação com a variável resposta;
2. Selecionar o **menor número** de variáveis regressoras, pois a variância de $\hat{Y}$ é uma função delas. Além do custo financeiro associado.

Construção do modelo

Na construção de um modelo de regressão, há dois objetivos conflitantes:

1. Selecionar o **maior número** de variáveis regressoras que contenham "toda" a informação da relação com a variável resposta;
2. Selecionar o **menor número** de variáveis regressoras, pois a variância de $\hat{Y}$ é uma função delas. Além do custo financeiro associado.

Esperançosamente, um meio-termo entre os dois levará ao melhor modelo.

Construção do modelo

Na construção de um modelo de regressão, há dois objetivos conflitantes:

1. Selecionar o **maior número** de variáveis regressoras que contenham "toda" a informação da relação com a variável resposta;
2. Selecionar o **menor número** de variáveis regressoras, pois a variância de $\hat{Y}$ é uma função delas. Além do custo financeiro associado.

Esperançosamente, um meio-termo entre os dois levará ao melhor modelo.

Construção do modelo

As técnicas para a **seleção de variáveis** tem o segundo objetivo como meta.

Porém,

- Técnicas diferentes podem resultar em modelos distintos;

Construção do modelo

As técnicas para a **seleção de variáveis** tem o segundo objetivo como meta.

Porém,

- Técnicas diferentes podem resultar em modelos distintos;
- Nenhuma técnica garante que o modelo selecionado é o melhor;

Construção do modelo

As técnicas para a **seleção de variáveis** tem o segundo objetivo como meta.

Porém,

- Técnicas diferentes podem resultar em modelos distintos;
- Nenhuma técnica garante que o modelo selecionado é o melhor;
- A confiança total no algoritmo para resultados deve ser evitada.

Construção do modelo

As técnicas para a **seleção de variáveis** tem o segundo objetivo como meta.

Porém,

- Técnicas diferentes podem resultar em modelos distintos;
- Nenhuma técnica garante que o modelo selecionado é o melhor;
- A confiança total no algoritmo para resultados deve ser evitada. Experiência, conhecimento dos dados e do problema podem ser levados em consideração.

Construção do modelo

As técnicas para a **seleção de variáveis** tem o segundo objetivo como meta.

Porém,

- Técnicas diferentes podem resultar em modelos distintos;
- Nenhuma técnica garante que o modelo selecionado é o melhor;
- A confiança total no algoritmo para resultados deve ser evitada. Experiência, conhecimento dos dados e do problema podem ser levados em consideração.

Roteiro

Critérios para a seleção de modelos

Em um modelo sem intercepto, existem um total 2^p de possíveis modelos.

Enquanto, em um modelo com intercepto (e assumindo que ele estará em todos as possibilidades), existem 2^{p-1}.

Quando p não é tão grande, nós podemos ajustar todos os possíveis modelos e identificar o melhor ou os melhores de acordo com um ou mais critérios.

Critérios para a seleção de modelos

Em um modelo sem intercepto, existem um total 2^p de possíveis modelos. Enquanto, em um modelo com intercepto (e assumindo que ele estará em todos as possibilidades), existem 2^{p-1}.

Quando p não é tão grande, nós podemos ajustar todos os possíveis modelos e identificar o melhor ou os melhores de acordo com um ou mais critérios.

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **quadrado médio do resíduo**, é dada por

$$\text{QMRes}(p) = \frac{\text{SQRes}(p)}{n-p}.$$

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **quadrado médio do resíduo**, é dada por

$$\text{QMRes}(p) = \frac{\text{SQRes}(p)}{n-p}.$$

Nós selecionamos o modelo com o menor valor de QMRes(p).

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **quadrado médio do resíduo**, é dada por

$$\text{QMRes}(p) = \frac{\text{SQRes}(p)}{n-p}.$$

Nós selecionamos o modelo com o menor valor de QMRes(p). Observando que, geralmente, o QMRes(p) aumenta com mais variáveis no modelo, enquanto o SQRes(p) diminui.

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **quadrado médio do resíduo**, é dada por

$$\text{QMRes}(p) = \frac{\text{SQRes}(p)}{n-p}.$$

Nós selecionamos o modelo com o menor valor de QMRes(p). Observando que, geralmente, o QMRes(p) aumenta com mais variáveis no modelo, enquanto o SQRes(p) diminui.

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **coeficiente de determinação múltipla** é dada por

$$R_p^2 = 1 - \frac{\text{SQRes}(p)}{\text{SQT}}.$$

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **coeficiente de determinação múltipla** é dada por

$$R_p^2 = 1 - \frac{\text{SQRes}(p)}{\text{SQT}}.$$

Modelos com grandes valores de R_p^2 são preferidos, mas adicionar termos aumentará esse valor.

Critérios para a seleção de modelos

Suponha um modelo com p termos, o **coeficiente de determinação múltipla** é dada por

$$R_p^2 = 1 - \frac{\text{SQRes}(p)}{\text{SQT}}.$$

Modelos com grandes valores de R_p^2 são preferidos, mas adicionar termos aumentará esse valor.

Critérios para a seleção de modelos

O **coeficiente de determinação múltipla ajustado** é dada por

$$R^2_{\text{Aj},\,p} = 1 - \frac{n-1}{n-p}(1 - R^2_p).$$

Critérios para a seleção de modelos

O **coeficiente de determinação múltipla ajustado** é dada por

$$R^2_{\text{Aj},\,p} = 1 - \frac{n-1}{n-p}(1 - R^2_p).$$

Da mesma forma, modelos com grandes valores de $R^2_{\text{Aj},\,p}$ são preferidos.

Critérios para a seleção de modelos

O **coeficiente de determinação múltipla ajustado** é dada por

$$R^2_{\text{Aj},\,p} = 1 - \frac{n-1}{n-p}(1 - R^2_p).$$

Da mesma forma, modelos com grandes valores de $R^2_{\text{Aj},\,p}$ são preferidos.

Porém, esse valor não aumentará necessariamente à medida que termos adicionais forem introduzidos no modelo.

Critérios para a seleção de modelos

O **coeficiente de determinação múltipla ajustado** é dada por

$$R^2_{\text{Aj},\,p} = 1 - \frac{n-1}{n-p}(1 - R^2_p).$$

Da mesma forma, modelos com grandes valores de $R^2_{\text{Aj},\,p}$ são preferidos. Porém, esse valor não aumentará necessariamente à medida que termos adicionais forem introduzidos no modelo.

Critérios para a seleção de modelos

Estatística C_p de Mallows (1964)

$$C_p = \frac{\text{SQRes}(p)}{\hat{\sigma}^2} - (n - 2p).$$

Critérios para a seleção de modelos

Estatística C_p de Mallows (1964)

$$C_p = \frac{\text{SQRes}(p)}{\hat{\sigma}^2} - (n - 2p).$$

Valores pequenos de C_p são desejados.

Critérios para a seleção de modelos

Estatística C_p de Mallows (1964)

$$C_p = \frac{\text{SQRes}(p)}{\hat{\sigma}^2} - (n - 2p).$$

Valores pequenos de C_p são desejados. Valores negativos são possíveis, eles podem indicar que a variância verdadeira está sendo superestimada.

Critérios para a seleção de modelos

Estatística C_p de Mallows (1964)

$$C_p = \frac{\text{SQRes}(p)}{\hat{\sigma}^2} - (n - 2p).$$

Valores pequenos de C_p são desejados. Valores negativos são possíveis, eles podem indicar que a variância verdadeira está sendo superestimada.

Critérios para a seleção de modelos

Akaike information criterion (AIC, critério de Akaike, 1973). O AIC é um critério baseado na maximização da entropia esperada de um modelo.

Critérios para a seleção de modelos

Akaike information criterion (AIC, critério de Akaike, 1973). O AIC é um critério baseado na maximização da entropia esperada de um modelo. Entropia é uma medida da informação esperada, no caso, baseada na divergência de Kullback e Leibler (1951).

Critérios para a seleção de modelos

Akaike information criterion (AIC, critério de Akaike, 1973). O AIC é um critério baseado na maximização da entropia esperada de um modelo. Entropia é uma medida da informação esperada, no caso, baseada na divergência de Kullback e Leibler (1951). O AIC é dado por

$$\text{AIC} = -2\log(L) + 2p,$$

em que L é a função de verossimilhança.

Critérios para a seleção de modelos

Akaike information criterion (AIC, critério de Akaike, 1973). O AIC é um critério baseado na maximização da entropia esperada de um modelo. Entropia é uma medida da informação esperada, no caso, baseada na divergência de Kullback e Leibler (1951). O AIC é dado por

$$\text{AIC} = -2\log(L) + 2p,$$

em que L é a função de verossimilhança.

Critérios para a seleção de modelos

Para o caso da regressão por mínimos quadrados, o AIC pode ser escrito como

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

Critérios para a seleção de modelos

Para o caso da regressão por mínimos quadrados, o AIC pode ser escrito como

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

O ideal é buscar o modelo com menor AIC.

Critérios para a seleção de modelos

Para o caso da regressão por mínimos quadrados, o AIC pode ser escrito como

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

O ideal é buscar o modelo com menor AIC.

Critérios para a seleção de modelos

A versão bayesiana do AIC, ***Bayesian information criterion*** (BIC, Schwarz, 1978) é dada por

$$\text{BIC} = -2\log(L) + p\log(n).$$

Critérios para a seleção de modelos

A versão bayesiana do AIC, ***Bayesian information criterion*** (BIC, Schwarz, 1978) é dada por

$$\text{BIC} = -2\log(L) + p\log(n).$$

Para o caso da regressão por mínimos quadrados,

$$\text{BIC}_p = n\log\left[\frac{\text{SQRes}(p)}{n}\right] + p\log(n).$$

Critérios para a seleção de modelos

A versão bayesiana do AIC, ***Bayesian information criterion*** (BIC, Schwarz, 1978) é dada por

$$\text{BIC} = -2\log(L) + p\log(n).$$

Para o caso da regressão por mínimos quadrados,

$$\text{BIC}_p = n\log\left[\frac{\text{SQRes}(p)}{n}\right] + p\log(n).$$

Da mesma forma que no AIC, o ideal é buscar o modelo com menor BIC.

Critérios para a seleção de modelos

A versão bayesiana do AIC, ***Bayesian information criterion*** (BIC, Schwarz, 1978) é dada por

$$\text{BIC} = -2\log(L) + p\log(n).$$

Para o caso da regressão por mínimos quadrados,

$$\text{BIC}_p = n\log\left[\frac{\text{SQRes}(p)}{n}\right] + p\log(n).$$

Da mesma forma que no AIC, o ideal é buscar o modelo com menor BIC.

Critérios para a seleção de modelos

Estatística PRESS (*prediction error sum of squares*), ela é definida da seguinte forma

$$\text{PRESS}(p) = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2,$$

Critérios para a seleção de modelos

Estatística PRESS (*prediction error sum of squares*), ela é definida da seguinte forma

$$\text{PRESS}(p) = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2,$$

em que $\hat{Y}_{(\ell)}$ é a variável resposta ajustada sem a observação ℓ e a ℓ-ésima linha da matriz de planejamento.

Critérios para a seleção de modelos

Estatística PRESS (*prediction error sum of squares*), ela é definida da seguinte forma

$$\text{PRESS}(p) = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2,$$

em que $\hat{Y}_{(\ell)}$ é a variável resposta ajustada sem a observação ℓ e a ℓ-ésima linha da matriz de planejamento. Valores pequenos da estatística PRESS são desejados.

Critérios para a seleção de modelos

Estatística PRESS (*prediction error sum of squares*), ela é definida da seguinte forma

$$\text{PRESS}(p) = \sum_{\ell=1}^{n} \left\{ Y_\ell - \hat{Y}_{(\ell)} \right\}^2 = \sum_{\ell=1}^{n} \left\{ \frac{e_\ell}{1 - h_{\ell\ell}} \right\}^2,$$

em que $\hat{Y}_{(\ell)}$ é a variável resposta ajustada sem a observação ℓ e a ℓ-ésima linha da matriz de planejamento. Valores pequenos da estatística PRESS são desejados.

Critérios para a seleção de modelos

Uma vez que alguns modelos candidatos foram identificados, uma análise de regressão deve ser feita com cada dos modelos e, em seguida, compará-los.

Se o sinal das estimativas de um determinado coeficiente se alterna (entre positivo e negativo), isso pode ser uma indicação de multicolinearidade.

Critérios para a seleção de modelos

Uma vez que alguns modelos candidatos foram identificados, uma análise de regressão deve ser feita com cada dos modelos e, em seguida, compará-los.

Se o sinal das estimativas de um determinado coeficiente se alterna (entre positivo e negativo), isso pode ser uma indicação de multicolinearidade.

Roteiro

Procedimentos para seleção automática

Existem três procedimentos para seleção automática que se destacam:

1. *forward selection*;

Procedimentos para seleção automática

Existem três procedimentos para seleção automática que se destacam:

1. *forward selection*;
2. *backward elimination*;

Procedimentos para seleção automática

Existem três procedimentos para seleção automática que se destacam:

1. *forward selection*;
2. *backward elimination*;
3. Regressão *stepwise*.

Procedimentos para seleção automática

Existem três procedimentos para seleção automática que se destacam:

1. *forward selection*;
2. *backward elimination*;
3. Regressão *stepwise*.

Procedimentos para seleção automática

Forward selection

O procedimento é baseado na ideia de que nenhuma variável está no modelo originalmente, mas são adicionadas uma de cada vez.

Procedimentos para seleção automática

Forward selection

O procedimento é baseado na ideia de que nenhuma variável está no modelo originalmente, mas são adicionadas uma de cada vez.

Forward selection

O procedimento de seleção é o seguinte:

1. O primeiro regressor selecionado para ser inserido no modelo é aquele com a correlação mais alta com a resposta.

Forward selection

O procedimento de seleção é o seguinte:

1. O primeiro regressor selecionado para ser inserido no modelo é aquele com a correlação mais alta com a resposta. Se a estatística F correspondente ao modelo que contém essa variável for significativa

Forward selection

O procedimento de seleção é o seguinte:

1. O primeiro regressor selecionado para ser inserido no modelo é aquele com a correlação mais alta com a resposta. Se a estatística F correspondente ao modelo que contém essa variável for significativa (maior do que algum valor predeterminado, F_{in}), então esse regressor é deixado no modelo.

Forward selection

O procedimento de seleção é o seguinte:

1. O primeiro regressor selecionado para ser inserido no modelo é aquele com a correlação mais alta com a resposta. Se a estatística F correspondente ao modelo que contém essa variável for significativa (maior do que algum valor predeterminado, F_{in}), então esse regressor é deixado no modelo.

Forward selection

2. O segundo regressor examinado é aquele com a maior correlação parcial com a resposta.

Forward selection

2. O segundo regressor examinado é aquele com a maior correlação parcial com a resposta. Se a estatística F correspondente à adição dessa variável for significativa, o regressor é mantido.

Forward selection

2. O segundo regressor examinado é aquele com a maior correlação parcial com a resposta. Se a estatística F correspondente à adição dessa variável for significativa, o regressor é mantido.

3. Este processo continua até que todos os regressores sejam examinados.

Forward selection

2. O segundo regressor examinado é aquele com a maior correlação parcial com a resposta. Se a estatística F correspondente à adição dessa variável for significativa, o regressor é mantido.

3. Este processo continua até que todos os regressores sejam examinados.

Procedimentos para seleção automática

Backward elimination

O procedimento é baseado na ideia de que todas as variáveis estão no modelo originalmente, examinadas uma de cada vez e removidas se não forem significativas.

Procedimentos para seleção automática

Backward elimination

O procedimento é baseado na ideia de que todas as variáveis estão no modelo originalmente, examinadas uma de cada vez e removidas se não forem significativas.

Backward elimination

O procedimento de seleção é o seguinte:

1. A estatística F parcial é calculada para cada variável como se fosse a última adicionada ao modelo.

Backward elimination

O procedimento de seleção é o seguinte:

1. A estatística F parcial é calculada para cada variável como se fosse a última adicionada ao modelo. O regressor com a menor estatística F é examinado primeiro e será removido se este valor for menor que algum valor predeterminado F_{out}.

Backward elimination

O procedimento de seleção é o seguinte:

1. A estatística F parcial é calculada para cada variável como se fosse a última adicionada ao modelo. O regressor com a menor estatística F é examinado primeiro e será removido se este valor for menor que algum valor predeterminado F_{out}.

Backward elimination

2. Se este regressor for removido, o modelo é reajustado com as variáveis do regressor restantes e as estatísticas F parciais calculadas novamente.

Backward elimination

2. Se este regressor for removido, o modelo é reajustado com as variáveis do regressor restantes e as estatísticas F parciais calculadas novamente.

   O regressor com a menor estatística F parcial será removido se esse valor for menor que F_{out}.

Backward elimination

2. Se este regressor for removido, o modelo é reajustado com as variáveis do regressor restantes e as estatísticas F parciais calculadas novamente. O regressor com a menor estatística F parcial será removido se esse valor for menor que F_{out}.
3. O processo continua até que todos os regressores sejam examinados.

Backward elimination

2. Se este regressor for removido, o modelo é reajustado com as variáveis do regressor restantes e as estatísticas F parciais calculadas novamente. O regressor com a menor estatística F parcial será removido se esse valor for menor que F_{out}.
3. O processo continua até que todos os regressores sejam examinados.

Procedimentos para seleção automática

Stepwise regression

Este procedimento é uma modificação do *forward selection*.

Procedimentos para seleção automática

Stepwise regression

Este procedimento é uma modificação do *forward selection*.

Stepwise regression

O procedimento de seleção é o seguinte:

1. A contribuição de cada variável regressora colocada no modelo é reavaliada por meio de sua estatística F parcial.

Stepwise regression

O procedimento de seleção é o seguinte:

1. A contribuição de cada variável regressora colocada no modelo é reavaliada por meio de sua estatística F parcial.

Stepwise regression

2. Um regressor que entra no modelo também pode ser removido se for considerado insignificante com a adição de outras variáveis ao modelo.

Stepwise regression

2. Um regressor que entra no modelo também pode ser removido se for considerado insignificante com a adição de outras variáveis ao modelo.

   Se a estatística F parcial for menor que F_{out}, a variável será removida.

Stepwise regression

2. Um regressor que entra no modelo também pode ser removido se for considerado insignificante com a adição de outras variáveis ao modelo. Se a estatística F parcial for menor que F_{out}, a variável será removida.
3. Este método requer um valor F_{in} e um valor F_{out}.

Stepwise regression

2. Um regressor que entra no modelo também pode ser removido se for considerado insignificante com a adição de outras variáveis ao modelo. Se a estatística F parcial for menor que F_{out}, a variável será removida.
3. Este método requer um valor F_{in} e um valor F_{out}.

Procedimentos para seleção automática

Considerações

As três técnicas podem resultar em modelos diferentes. E nenhum deles pode ser considerado o "melhor".

Procedimentos para seleção automática

Considerações

As três técnicas podem resultar em modelos diferentes. E nenhum deles pode ser considerado o "melhor".

Roteiro

Regressão lasso

Um método para encontrar estimadores (viesados) e ao mesmo tempo selecionar as variáveis é a regressão *lasso* (*least absolute shrinkage and selection operator*, Tibshirani, 1996).

Regressão lasso

Com as variáveis resposta e explicativas padronizadas na **escala de tamanho unitário**, as estimativas dos parâmetros podem obtidas com a solução da minimização de:

$$Q = \sum_{\ell=1}^{n} [Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p})]^2 + t \sum_{m=1}^{p} |\beta_m|, \qquad (3)$$

Regressão lasso

Com as variáveis resposta e explicativas padronizadas na **escala de tamanho unitário**, as estimativas dos parâmetros podem obtidas com a solução da minimização de:

$$Q = \sum_{\ell=1}^{n} [Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p})]^2 + t \sum_{m=1}^{p} |\beta_m|, \qquad (3)$$

em que $t \geq 0$ é uma constante denominada de parâmetro de regularização.

Regressão lasso

Com as variáveis resposta e explicativas padronizadas na **escala de tamanho unitário**, as estimativas dos parâmetros podem obtidas com a solução da minimização de:

$$Q = \sum_{\ell=1}^{n} [Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p})]^2 + t \sum_{m=1}^{p} |\beta_m|, \qquad (3)$$

em que $t \geq 0$ é uma constante denominada de parâmetro de regularização.

A segunda parte do lado de direito de (3), é um termo de "encolhimento".

Regressão lasso

Com as variáveis resposta e explicativas padronizadas na **escala de tamanho unitário**, as estimativas dos parâmetros podem obtidas com a solução da minimização de:

$$Q = \sum_{\ell=1}^{n} [Y_\ell - (\beta_1 x_{\ell 1} + \beta_2 x_{\ell 2} + \cdots + \beta_p x_{\ell p})]^2 + t \sum_{m=1}^{p} |\beta_m|, \qquad (3)$$

em que $t \geq 0$ é uma constante denominada de parâmetro de regularização. A segunda parte do lado de direito de (3), é um termo de "encolhimento".

Regressão lasso

Os estimador lasso não tem forma fechada. Uma aproximação da forma fechada pode ser obtida se o termo de encolhimento for o seguinte

Regressão lasso

Os estimador lasso não tem forma fechada. Uma aproximação da forma fechada pode ser obtida se o termo de encolhimento for o seguinte

$$\sum_{m=1}^{p} \frac{\beta_m^2}{|\beta_m|}.$$

Regressão lasso

Os estimador lasso não tem forma fechada. Uma aproximação da forma fechada pode ser obtida se o termo de encolhimento for o seguinte

$$\sum_{m=1}^{p} \frac{\beta_m^2}{|\beta_m|}.$$

Dessa forma, o estimador *lasso*, $\widetilde{\beta}$, é aproximado por:

Regressão lasso

Os estimador lasso não tem forma fechada. Uma aproximação da forma fechada pode ser obtida se o termo de encolhimento for o seguinte

$$\sum_{m=1}^{p} \frac{\beta_m^2}{|\beta_m|}.$$

Dessa forma, o estimador *lasso*, $\widetilde{\boldsymbol{\beta}}$, é aproximado por:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}}^{*} = (\boldsymbol{X}^{\top}\boldsymbol{X} + k\boldsymbol{W}^{-})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y},$$

Regressão lasso

Os estimador lasso não tem forma fechada. Uma aproximação da forma fechada pode ser obtida se o termo de encolhimento for o seguinte

$$\sum_{m=1}^{p} \frac{\beta_m^2}{|\beta_m|}.$$

Dessa forma, o estimador *lasso*, $\widetilde{\beta}$, é aproximado por:

$$\widehat{\beta}_{\mathrm{R}}^{*} = (\boldsymbol{X}^{\top}\boldsymbol{X} + k\boldsymbol{W}^{-})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y},$$

Regressão lasso

em que $\boldsymbol{W}^{-}$ é a inversa generalizada de $\boldsymbol{W}$, uma matriz diagonal, com o m-ésimo elemento dado por $|\tilde{\beta}_m|$ e k é escolhido de forma $\sum_{m=1}^{p} |\beta_m|^* = t$.

Essa aproximação obtém as estimativa da regressão lasso a partir de um método iterativo utilizando a regressão *ridge*.

Regressão lasso

em que $\boldsymbol{W}^{-}$ é a inversa generalizada de $\boldsymbol{W}$, uma matriz diagonal, com o m-ésimo elemento dado por $|\tilde{\beta}_m|$ e k é escolhido de forma $\sum_{m=1}^{p} |\beta_m|^* = t$.

Essa aproximação obtém as estimativa da regressão lasso a partir de um método iterativo utilizando a regressão *ridge*.

Regressão lasso

Mesmo tendo uma expressão fechada, Tibshirani (1996) sugere o uso dos estimadores obtidos a partir de (3), ao invés dos oriundos da forma fechada.

O valor de t deve ser encontrado através de **validação cruzada**.

Regressão lasso

Mesmo tendo uma expressão fechada, Tibshirani (1996) sugere o uso dos estimadores obtidos a partir de (3), ao invés dos oriundos da forma fechada.

O valor de t deve ser encontrado através de **validação cruzada**.

Roteiro

Aplicação

(Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y),

Aplicação

(Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5).

Aplicação

(Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5). Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 62{,}405 + 1{,}551 x_{\ell 2} + 0{,}510 x_{\ell 3} + 0{,}102 x_{\ell 4} - 0{,}144 x_{\ell 5},$$

$\ell = 1, 2, \ldots, 13.$

Aplicação

(Hald, 1952) Um conjunto de dados, com 13 observações, relacionando o calor transformado em calorias por grama de cimento (Y), com a quantidade de quatro tipos de mistura (x_2 a x_5). Após o ajuste dos dados, nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 62{,}405 + 1{,}551 x_{\ell 2} + 0{,}510 x_{\ell 3} + 0{,}102 x_{\ell 4} - 0{,}144 x_{\ell 5},$$

$\ell = 1, 2, \ldots, 13.$

Exemplo

Nós temos também que:

Tabela 1: Estimativas do parâmetros.

Parâmetro	Estimativa	EP	t_c
β_1	62,405	70,071	0,891
β_2	1,551	0,745	2,083
β_3	0,510	0,724	0,705
β_4	0,102	0,755	0,135
β_5	-0,144	0,709	-0,203

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}306$, com QMRes $= 5{,}983$.

Exemplo

Tabela 2: Medidas de multicolinearidade.

	x_2	x_3	x_4	x_5
VIF	38,50	254,42	46,87	282,51

Tabela 3: Medidas para seleção.

Var	p	SQRes	R2	R2aj	QMRes	AIC	BIC	Cp
-	1	2715,76	0,00	0,00	226,31	110,34	111,47	442,92
x2	2	1265,69	0,53	0,49	115,06	102,41	104,11	202,55
x3	2	906,34	0,67	0,64	82,39	98,07	99,77	142,49
x4	2	1939,40	0,29	0,22	176,31	107,96	109,65	315,15
x5	2	883,87	0,67	0,64	80,35	97,74	99,44	138,73
x2, x3	3	57,90	0,98	0,97	5,79	64,31	66,57	2,68
x2, x3	3	1227,07	0,55	0,46	122,71	104,01	106,27	198,09
x2, x4	3	74,76	0,97	0,97	7,48	67,63	69,89	5,50
x2, x5	3	415,44	0,85	0,82	41,54	89,93	92,19	62,44
x3, x4	3	868,88	0,68	0,62	86,89	99,52	101,78	138,23
x4, x5	3	175,74	0,94	0,92	17,57	78,74	81,00	22,37
x2, x3, x4	4	48,11	0,98	0,98	5,35	63,90	66,73	3,04
x2, x3, x5	4	47,97	0,98	0,98	5,33	63,87	66,69	3,02
x2, x4, x5	4	50,84	0,98	0,98	5,65	64,62	67,44	3,50
x3, x4, x5	4	73,81	0,97	0,96	8,20	69,47	72,29	7,34
Todas	5	47,86	0,98	0,97	5,98	65,84	69,23	5,00

Exemplo

Tabela 4: Estimativas do parâmetros após o *stepwise* (coincidiu com o *lasso*).

Parâmetro	Estimativa	EP	t_c
β_1	71,648	14,142	5,066
β_2	1,452	0,117	12,410
β_3	0,416	0,186	2,242
β_5	-0,237	0,173	-1,365

Região crítica, para $\alpha = 5\%$: $|t_c| > 2{,}262$, com QMRes $= 5{,}33$.

Roteiro

Referências bibliográficas I

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov e F. Csáki, eds, '2nd International Symposium on Information Theory', Akadémiai Kiadó, Budapest, pp. 267–281.

Hald, A. (1952), *Statistical theory with Engineering applications*, Wiley, New York.

Kullback, S. e Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Referências bibliográficas II

Mallows, C. L. (1964), 'Choosing variables in a linear regression: A graphical aid', Central Regional Meeting of the Institute of Mathematical Statistics. Manhattan, KS.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Obrigado!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Departamento de Estatística, Sala 319