CE071 - Análise de Regressão Linear

Cesar Augusto Taconeli

05 de junho, 2018

Aula 6 - Seleção de modelos

Introdução

Princípio de Occam's Razor

Dentre as várias explicações possíveis para um fenômeno, a mais simples é a melhor.

Fuechsel, técnico da IBM

Garbage in, garbage out.

Introdução

- Neste módulo vamos tratar da seleção de covariáveis para o ajuste do modelo de regressão;
- O objetivo é identificar um modelo parcimonioso, capaz de proporcionar bom ajuste com a menor quantidade possível de parâmetros;
- Diferentes métodos podem ser aplicados na seleção de um subconjunto "ótimo" de variáveis;
- Importante ter em mente que diferentes métodos de seleção, frequentemente, remetem a modelos distintos (lembre-se: *"All models are wrong but some are useful"*).

Introdução

- Por que não incluir todas as covariáveis no modelo?

1. Um dos objetivos principais da análise de regressão é explicar a relação entre as variáveis da maneira mais simples possível;
2. Quanto maior o número de parâmetros no modelo, menos graus de liberdade para os resíduos, menor precisão para as inferências;
3. Quanto maior o número de variáveis incluídas no modelo, maior a possibilidade de colinearidade;
4. Quanto mais complexo (parametrizado) o modelo, melhor o ajuste da amostra, mas menor seu poder de generalização (baixo poder preditivo).

Como proceder a seleção do modelo?

- Antes de aplicar qualquer método analítico para seleção de modelo, é conveniente fazer uma pré-triagem de variáveis, buscando eliminar variáveis que, a título de exemplo:
  - Sejam redundantes;
  - Apresentem elevado erro de medida;
  - Não estejam no contexto do estudo.

Seleção de covariáveis - Modelos hierárquicos

- Certos modelos têm alguma hierarquia natural em suas covariáveis.
- Nesses casos, a seleção das covariáveis a compor o modelo deve respeitar a hierarquia.
- Um dos casos mais típicos de modelo hierárquicos é o **modelo polinomial**.

Seleção de covariáveis - Modelos hierárquicos

- Considere o seguinte modelo polinomial:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon. \tag{1}$$

- Em modelos polinomiais, a remoção das variáveis sempre deve iniciar pelos termos de maior ordem.
- Considere o modelo polinomial sem o termo linear:

$$y = \beta_0 + \beta_2 x^2 + \epsilon. \tag{2}$$

Seleção de covariáveis - Modelos hierárquicos

- Suponha que seja feita uma mudança de escala, trocando x por $x + c$. O modelo ficaria dado por:

$$y = \beta_0 + \beta_2 c^2 + 2\beta_2 cx + \beta_2 x^2 + \epsilon. \tag{3}$$

- Observe que o termo linear reaparece mediante simples mudança de escala.

- Em geral, é indesejado que a estrutura do modelo (e as interpretações) sejam alteradas devido à simples mudança de escala dos dados.

Seleção de covariáveis - Modelos hierárquicos

- Da mesma forma o princípio hierárquico se aplica a modelos com termos de interação entre variáveis.

- Considere o modelo de regressão com interação de segunda ordem:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon. \quad (4)$$

- Neste caso, não é recomendável remover o termo $x_1 x_2$ sem a remoção dos demais termos quadráticos (x_1^2, x_2^2);
- A remoção conjunta de $x_1 x_2$, x_1^2 e x_2^2, por outro lado, tem objetivo relevante, permitindo comparar os ajustes linear e quadrático.

Seleção de covariáveis - Métodos baseados em testes de hipóteses

- O teste F baseado na variação da soma de quadrados de resíduos pode ser usado para fins de seleção de covariáveis.

- Basicamente, se uma particular covariável não é significativa no modelo ajustado, a um nível de significância especificado α_c, então ela poderia ser removida do modelo.

- Na sequência são descritos três algoritmos para seleção de covariáveis baseados em testes de hipóteses.

Seleção de covariáveis - Método backward

1. Ajuste o modelo de regressão com todas as k covariáveis disponíveis;

2. Remova uma a uma as covariáveis do modelo e calcule os p-valores correspondentes aos respectivos testes F $(p_1, p_2, ..., p_k)$;

3. Seja $p_j = \max(p_1, p_2, ..., p_k)$. Se $p_j > \alpha_c$, então a variável x_j correspondente é eliminada do modelo (para sempre!);

4. Os passos 1 a 3 são repetidos para o novo modelo (sem x_j) e o processo continua até que $p_j = \max(p) < \alpha_c$. Quando isso ocorrer, interrompa o processo e declare o modelo atual como selecionado.

Seleção de covariáveis - Método forward

1. Ajuste o modelo nulo (sem covariáveis);

2. Adicione uma a uma as covariáveis ao modelo e calcule os p-valores correspondentes aos respectivos testes F $(p_1, p_2, ..., p_k)$;

3. Seja $p_j = \min(p_1, p_2, ..., p_k)$. Se $p_j < \alpha_c$, então a variável x_j correspondente é incorporada ao modelo (para sempre!);

4. Os passos 1 a 3 são repetidos para o novo modelo (com x_j) e o processo continua até que $p_j = \min(p) > \alpha_c$. Quando isso ocorrer, interrompa o processo e declare o modelo atual como selecionado.

Seleção de covariáveis - Método stepwise

1. Ajuste o modelo de regressão com todas as k covariáveis disponíveis. Proceda como descrito no algoritmo bacward, mas...

2. A cada passo do algoritmo, calcule os testes F correspondentes à exclusão das variáveis que estão no modelo e também os testes F referentes à inclusão das variáveis que estão no modelo.

3. Defina α_e e α_s os níveis de significância aplicados à decisão de inserir e excluir covariáveis do modelo.

4. Enquanto houver alguma variável com $p < \alpha_e$ fora do modelo e/ou $p > \alpha_s$ no modelo, continue o processo. Caso contrário interrompa o processo e declare o modelo atual como selecionado.

Seleção de covariáveis - Métodos baseados em testes de hipóteses

- Métodos de seleção de covariáveis baseados em testes de hipóteses têm suas limitações:

1. Devido à forma de seleção das covariáveis, o "modelo ótimo" pode não ser identificado ao longo do processo;

2. Os p-valores devem ser avaliados com cautela. Tenha em mente que o algoritmo envolve a avaliação de múltiplos testes e p-valores;

3. O algoritmo não está direcionado ao objetivo do estudo, seja explicação ou predição;

4. Os algoritmos apresentados tendem a selecionar modelos mais simples que o adequado para fins preditivos.

Seleção de covariáveis - Métodos baseados em testes de hipóteses

- Para contornar algumas das limitações mencionadas dos algoritmos de seleção baseados em testes de hipóteses, é comum adotar para α_c, α_e e α_s valores mais 'condescendentes', como 0,15 ou mesmo 0,20.

Critérios para avaliação e comparação de modelos

- No processo de seleção de covariáveis, diferentes critérios podem ser usados para comparar os modelos produzidos. Alguns deles são descritos na sequência.
- **Coeficiente de deteminação** - O coeficiente de determinação corresponde à proporção da variação dos dados explicada pela regressão:

$$R^2 = \frac{SQ_{Reg}}{SQ_{total}} = 1 - \frac{SQ_{Res}}{SQ_{total}}, \tag{5}$$

em que SQ_{Reg}, SQ_{Res} e SQ_{total} são as somas de quadrado de regressão e de resíduos do modelo ajustado e a soma de quadrados total.

- O coeficiente de determinação não é apropriado para comparar modelos com diferentes números de parâmetros, uma vez que R^2 sempre aumenta com a inclusão de novas covariáveis.

Critérios para avaliação e comparação de modelos

- **Coeficiente de deteminação ajustado** - O coeficiente de determinação ajustado (ou simplesmente R^2 ajustado) é definido por:

$$R^2_{Aj} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2), \tag{6}$$

em que n e p são o número de observações e o número de parâmetros do modelo.

- Diferentemente do que ocorre para R^2, o valor de R^2_{Aj} pode não aumentar mediante inclusão de novas variáveis ao modelo. Deve-se optar por modelos com maiores valores de R^2_{Aj}.

Critérios para avaliação e comparação de modelos

- **Quadrado médio de resíduos**, definido por:

$$QM_{Res} = \frac{SQ_{Res}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}. \tag{7}$$

também pode ser usado para comparação e seleção de modelos de regressão.

- Deve-se optar por modelos com menores valores para QM_{Res};

- Pode-se mostrar que minimizar QM_{Res} é equivalente a maximizar R^2_{Aj}, de forma que os dois critérios conduzem à seleção do mesmo conjunto de covariáveis.

Critérios para avaliação e comparação de modelos

- **Cp de Mallows** - O coeficiente C_p de Mallows é definido por:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E\left[\hat{y}_i - E(y_i)\right]^2, \tag{8}$$

podendo ser estimado por:

$$C_p = \frac{SQ_{Res}}{\hat{\sigma}^2} + 2p - n, \tag{9}$$

em que $\hat{\sigma}^2$ é dado pelo quadrado médio de resídus do modelo que inclui todas as covariáveis.

Critérios para avaliação e comparação de modelos

- Para o modelo completo, com p parâmetros, $C_p = p$.
- Para submodelos, definidos por subconjunto das covariáveis, menores valores para C_p são preferíveis.
- Uma estratégia para selecionar modelos com base nos valores de C_p é plotar C_p *versus* p e adicionar ao gráfico a reta $C_p = p$.
- Modelos com viés reduzido terão pontos próximos a reta.
- Dentre os modelos com pontos próximos à reta, deve-se optar por aquele com menor C_p (e p, consequentemente).

Critérios para avaliação e comparação de modelos

- A **estatística PRESS** permite avaliar a qualidade preditiva dos modelos de regressão, sendo definida por:

$$\text{PRESS} = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2, \tag{10}$$

em que $\hat{y}_{(i)}$ é obtido com base no modelo ajustado apenas com as demais $n - 1$ observações $(i = 1, 2, ..., n)$.

- Menores valores da estatística PRESS indicam modelos com maior poder preditivo.

Critérios para avaliação e comparação de modelos

- O critério de informação de Akaike (*Akaike Information Criterion*), ou simplesmente AIC, é definido por:

$$\text{AIC} = -2l(\hat{\theta}) + 2p, \tag{11}$$

em que $l(\hat{\theta})$ é a log-verossimilhança maximizada do modelo (calculada com base nos emv's dos parâmetros) e p o número de parâmetros.

- O AIC pode ser usado para qualquer modelo ajustado por máxima verossimilhança. No caso de um modelo de regressão linear temos:

$$\text{AIC} = -n \ln(SQ_{Res}/n) + 2p. \tag{12}$$

Critérios para avaliação e comparação de modelos

- O componente $2p$, na expressão do AIC, atua como *termo de penalização* atribuído à complexidade (número de parâmetros) do modelo.
- Um critério alternativo ao AIC é o Critério de Informação Bayesiano (BIC), definido, para um modelo de regressão linear, por:

$$\text{AIC} = -n \ln(SQ_{Res}/n) + \ln(n)p. \tag{13}$$

- O BIC penaliza mais fortemente a complexidade do modelo que o AIC ao substituir p por $\ln(n)$ como fator de penalização.
- Devemos selecionar modelos com menores valores de AIC (ou BIC).

Seleção de modelos baseada em critérios de qualidade de ajuste

- Um primeiro algoritmo de seleção de modelos, baseados em critérios de qualidade de ajuste, é o método **todas as regressões possíveis**;
- Suponha uma análise de regressão com k variáveis;
- Uma primeira alternativa consiste em ajustar todos os possíveis modelos de regressão com $j \leq k$ covariáveis, para $j = 0, 1, 2, ..., k$;
- Para cada modelo de regressão ajustado calcula-se o valor do critério de seleção escolhido (AIC, C_p ou R^2_{Aj}...);
- O modelo selecionado será quele que apresentar o valor ótimo para o critério adotado (maior R^2_{Aj}, menor AIC,...).

Seleção de modelos baseada em critérios de qualidade de ajuste

- O método baseado em todas as regressões possíveis torna-se inviável quando se dispõe de um grande número de covariáveis;
- Para k variáveis o número de regressões possíveis é 2^k. Para $k = 30$, por exemplo, teríamos 1.073.741.824 regressões possíveis;
- Como alternativa ao método de todas as regressões possíveis podemos usar os algoritmos backward, forward ou stepwise substituindo o teste F por algum dos critérios apresentados.