Universidade do Estado do Rio de Janeiro
Instituto Politécnico
Departamento de Modelagem Computacional

Disciplina de Modelos Lineares 2012-2

Professora Ariane Ferreira

Seleção de Variáveis

Em modelos de regressão múltipla é necessário determinar um subconjunto de variáveis independentes que melhor explique a variável resposta, isto é, dentre todas as variáveis explicativas disponíveis, devemos encontrar um subconjunto de variáveis importantes para o modelo.

Construir um modelo que inclui apenas um subconjunto de variáveis explicativas envolve dois objetivos conflitantes:

1. Obter o máximo de informação por meio de um modelo com tantas variáveis independentes possíveis;
2. Diminuir a variância da estimativa e o custo da coleta por meio de um modelo com menor número possível de variáveis.

Desta forma, obter um equilíbrio entre esses dois compromissos é de interesse. Para isto, utilizamos uma técnica, denominada de *seleção de variáveis*.

Existem duas principais estratégias no processo de *seleção de variáveis*:

- **Todos os modelos possíveis:** considera todos os subconjuntos possíveis de variáveis explicativas, e considerando critérios de avaliação, seleciona o melhor deles.
- **Seleção Automática:** faz uma busca do melhor subconjunto de variáveis explicativas sem considerar todos os possíveis subconjuntos.

Na prática, assumimos que a correta especificação funcional das variáveis explicativas é conhecida (por exemplo, $1/x_1$, $ln\ x_2$) e que não há *outliers* ou pontos influentes e então, aplicamos a técnica de *seleção de variáveis*. Entretanto, o ideal seria inicialmente,

- Identificar *outliers* e pontos influentes,
- Identificar eventuais colinearidade e heteroscedasticidade,
- Realizar quaisquer transformações que sejam necessárias,

e então, aplicar *seleção de variáveis*.

A seguir apresentamos detalhadamente as estratégias **Todos os modelos possíveis** e **Seleção Automática.**

Seleção Todos os Modelos Possíveis

Considere o modelo de regressão linear múltipla

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon,$$

e suas suposições. O método de todos os modelos possíveis possibilita a análise do ajuste de todos os submodelos compostos pelos possíveis subconjuntos das p varíaveis e identifica os melhores desses subconjuntos, segundo critérios de avaliação.

Alguns critérios para avaliar os modelos são: R_p^2, R_a^2, QME (Quadrado Médio do Erro), $C_p$de Mallows, AIC, BIC e $PRESS_p$. A seguir uma abordagem de cada um deles.

Coeficiente de Determinação Múltipla

Seja R_p^2a notação do coeficiente de determinação múltipla de um modelo com p variáveis explicativas, isto é, p coeficientes e o intercepto β_0.

$$R_p^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT},$$

em que SQR, SQE e SQT são a soma dos quadrados do modelo, soma dos quadrados dos resíduos (erros) e soma dos quadrados total, respectivamente.

O critério utilizado nesse método é que se adicionarmos uma variável insignificante teremos um aumento mínimo (pequeno) de R_p^2. Assim, ele é mais usado para julgar quando parar de adicionar variáveis do que para encontrar o melhor modelo já que R_p^2nunca diminui quando p aumenta.

Coeficiente de Determinação Ajustado

Para evitar dificuldades na interpretação de R^2, alguns estatísticos preferem usar o R_a^2(R^2 ajustado) , definido para uma equação com p+1 coeficientes como

$$R_a^2 = 1 - \left(\frac{n-1}{n-(p+1)}\right)(1 - R_p^2).$$

O R_a^2não necessariamente aumenta com a adição de parâmetros no modelo. Na verdade se s variáveis explicativas são incluidas no modelo (modelo com p+s variáveis), o R_a^2desse

modelo excederá R_a^2do modelo com p variáveis apenas se a estatística parcial F para testar a significância dos adicionais s coeficientes passar de 1 (para mais detalhes ver Seber[1977]). Consequetemente, um critério para a seleção de um modelo ótimo é escolher o modelo que tem o R_a^2máximo.

Quadrado Médio dos Resíduos

O quadrado médio dos resíduos de um modelo de regressão é obtido por meio de

$$QME = SQE/(n-p-1), \qquad (2.7.1.3)$$

em que SQE é a soma dos quadrados dos resíduos. O QME sempre decresce conforme p aumenta. O quadrado médio do erro inicialmente decresce, estabiliza e eventualmente pode aumentar. Esse eventual aumento ocorre quando a redução do QME em adicionar um coeficiente para o modelo não é suficiente para compensar a perda nos graus de liberdade do denominador de (2.7.1.3).

O modelo que minimiza QME também maximizará R_a^2. Para entender isso, notamos que

$$\begin{aligned} R_a^2 &= 1 - \left(\frac{n-1}{n-p}\right)(1-R_p^2) \\ &= 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SQE}{SQT}\right) \\ &= 1 - \left(\frac{n-1}{SQT}\right)\left(\frac{SQE}{n-p}\right) \\ &= 1 - \left(\frac{n-1}{SQT}\right) QME. \end{aligned}$$

Assim, minimizar QME e maximar R_a^2são equivalentes.

Cp de Mallows

O critério $C_p$de Mallows é baseado no conceito do erro quadrático médio (EQM) dos valores ajustados. O erro quadrático médio da previsão é

$$\begin{aligned} EQM &= E(\hat{y}_i - E(y_i))^2 = E(\hat{y}_i - E(\hat{y}_i) + E(E(\hat{y}_i) - E(y_i)))^2 \\ &= E(\hat{y}_i - E(\hat{y}_i))^2 + (E(\hat{y}_i) - E(y_i))^2 = Var(\hat{y}_i) + vicio^2(\hat{y}_i), \end{aligned}$$

em que $E(\hat{y}_i) - E(y_i)$é o vicío. Assim, o EQM é a soma da variância de $\hat{y}_i$e o vício ao quadrado. O EQM considerando os n valores amostrais é

$$\sum E(\hat{y}_i - y_i)^2 = \sum Var(\hat{y}_i) + \sum (E(\hat{y}_i) - E(y_i))^2.$$

O critério Γ_pé o erro quadrático médio dividido pela variância dos erros σ^2.

$$\Gamma_p = \left(\frac{1}{\sigma^2}\right)[\sum Var(\hat{y}_i) + \sum (E(\hat{y}_i) - E(y_i))^2], \qquad (2.7.1.4)$$

em que $\sum Var(\hat{y}_i) = (p+1)\sigma^2$e o valor esperado da soma dos quadrados dos erros é:

$$E(SQE) = (E(\hat{y}_i) - E(y_i))^2 + (n - (p+1))\sigma^2.$$

Substituindo esses valores em (2.7.1.4), obtemos

$$\Gamma_p = \left(\frac{1}{\sigma^2}\right)[E(SQE) - (n - (p+1))\sigma^2 + (p+1)\sigma^2]$$
$$= \frac{E(SQE)}{\sigma^2} - n + 2(p+1).$$

Como σ^2é desconhecido, assumindo que o modelo que inclui todas as variáveis explicativas é tal que o QME é um estimador não viciado de σ^2e substituindo E(SQE) pelo valor observado SQE, $\Gamma_p$pode ser estimado por

$$C_p = \frac{SQE(p)}{QME} - n + 2(p+1),$$

em que SQE(p) é a soma de quadrados dos erros do submodelo e QME é o quadrado médio do modelo com todas as variáveis explicativas.

Pode também ser mostrado que quando não há vício na estimativa do modelo com as p variáveis, $E(SQE) = (n - (p+1))\sigma^2$e então,

$$E[C_p|Vicio = 0] = \frac{(n - (p+1))\sigma^2}{\sigma^2} - n + 2(p+1) = p + 1,$$

em que $p+1$é o número de parâmetros no modelo já que p é o número de variáveis explicativas mais o intercepto.

A estratégia usada para selecionar modelos com o critério C_pé identificar modelos com C_p próximo do número de parâmetros $(p+1)$.

AIC e BIC

O Critério de Informação de Akaike (AIC) é definido como

$$AIC_p = -2log(L_p) + 2[(p+1)+1],$$

em que L_pé a função de máxima verossimilhança do modelo e p é o número de variáveis explicativas consideradas no modelo.

O Critério de Informação Bayesiano (BIC) é definido como

$$BIC_p = -2log(L_p) + [(p+1)+1]log(n).$$

Tanto o AIC quanto o BIC aumentam conforme SQE aumenta. Além disso, ambos critérios penalizam modelos com muitas variáveis sendo que valores menores de AIC e BIC são preferíveis.

Como modelos com mais variáveis tendem a produzir menor SQE mas usam mais parâmetros, a melhor escolha é balancear o ajuste com a quantidade de variáveis.

Critério PRESS

O critério $PRESS_p$(Prediction Error Sum of Squares) é definido por

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2,$$

em que $\hat{Y}_{(i)}$é o valor predito da regressão sem a i-ésima observação.

Podemos escrever o PRESS como

$$PRESS_p = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2,$$

em que o h_{ii}é o i-ésimo valor da diagonal da matriz H.

No uso desse critério, modelo com menor $PRESS_p$é preferível.

Seleção Automática

Como a seleção de todas as regressões possíveis necessita de um considerável esforço computacional, outros métodos foram desenvolvidos para selecionar o melhor subconjunto de variáveis sequencialmente, adicionando ou removendo variáveis em cada passo.

O critério para a adição ou remoção de covariáveis é geralmente baseado na estatística F, comparando modelos com e sem as variáveis em questão. O AIC, assim como outros critérios, também podem ser utilizados na decisão de inserir e remover variáveis. Existem três procedimentos automáticos: (1) Método Forward, (2) Método Backward e (3) Método Stepwise.

Seleção Forward

Esse procedimento parte da suposição de que não há variável no modelo, apenas o intercepto. A ideia do método é adicionar uma variável de cada vez. A primeira variável selecionada é aquela com maior correlação com a resposta.

Procedimento:

- Ajustamos o modelo com a variável com maior correlação amostral com a variável resposta.

Supondo que essa variável seja x_1, calculamos a estatística F para testar se ela realmente é significativa para o modelo. A variável entra no modelo se a estatística F for maior do que o ponto crítico, chamado de F_{in}ou F para entrada. Notemos que F_{in}é calculado para um dado αcrítico.

- Considerando que $x_1$foi selecionado para o modelo, o próximo passo é encontrar uma variável com maior correlação com a resposta considerando a presença da primeira variável no modelo. Esta é chamada de correlação parcial e é a correlação dos resíduos do modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$com os resíduos do modelo $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1$, j=2,3,...,p.

Vamos supor que a maior correlação parcial com y seja x_2. Isso implica que a maior estatística F parcial é:

$$F = \frac{SQR(x_2|x_1)}{QME(x_1, x_2)}.$$

Se o valor da estatística é maior do que F_{in}, x_2é selecionado para o modelo.

- O processo é repetido, ou seja, variável com maior correlação parcial com y é adicionada no modelo se sua estatística F parcial for maior que F_{in}, até que não seja incluída mais nenhuma variável explicativa no modelo.

Seleção Backward

Enquanto o método Forward começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método Backward faz o caminho oposto; incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada.

A decisão de retirada da variável é tomada baseando-se em testes F parciais, que são calculados para cada variável como se ela fosse a última a entrar no modelo.

Procedimento:

- Para cada variável explicativa calcula-se a estatística F. Para a variável x_k, por exemplo,

$$F = \frac{SQR(x_k|x_1, ..., x_{k-1})}{QME}.$$

O menor valor das estatísticas F parciais calculadas é então comparado com o F crítico, F_{out}, calculado para um dado valor αcrítico. Se o menor valor encontrado for menor do que F_{out}, elimina-se do modelo a covariável responsável pelo menor valor da estatística F parcial.

- Ajusta-se novamente o modelo, agora com as $p-1$variáveis. As estatísticas F parciais são calculadas para esse modelo e o processo é repetido.
- O algoritmo de eliminação termina quando a menor estatística F parcial não for menor do que F_{out}.

Seleção Stepwise

Stepwise é uma modificação da seleção Forward em que cada passo todas as variáveis do modelo são previamente verificadas pelas suas estatísticas F parciais. Uma variável adicionada no modelo no passo anterior pode ser redundante para o modelo por causa do seu relacionamento com as outras variáveis e se sua estatística F parcial for menor que F_{out}, ela é removida do modelo.

Procedimento:

- Iniciamos com uma variável: aquela que tiver maior correlação com a variável resposta.

- A cada passo do forward, depois de incluir uma variável, aplica-se o backward para ver se será descartada alguma variável.
- Continuamos o processo até não incluir ou excluir nenhuma variável.

Assim, a regressão Stepwise requer dois valores de corte: F_{in}e F_{out}. Alguns autores preferem escolher $F_{in} = F_{out}$mas isso não é necessário. Se $F_{in}<F_{out}$: mais difícil remover que adicionar; se $F_{in}>F_{out}$: mais difícil adicionar que remover.

Algumas Considerações

A seleção de variáveis é um meio para se chegar a um modelo, mas não é a etapa final. O objetivo é construir um modelo que seja bom para obter predições ou que explique bem o relacionamento entre os dados.

Os métodos de seleção automática têm a vantagem de não necessitar de grande esforço computacional. Mas eles não indicam o melhor modelo respeitando algum critério (não retorna um conjunto de modelos em que o pesquisador tem o poder de escolha).

Já o método de todos os modelos possíveis identifica modelos que são melhores respeitando o critério que o pesquisador quiser.

É indicado, então, usar métodos passo a passo combinados com outros critérios.

Se por acaso existe um grande número de variáveis, é recomendado usar métodos de seleção automática para eliminar aquelas com efeitos insignificantes E o conjunto reduzido de variáveis pode então ser investigado pelo método de todos os modelos possíveis.

A escolha do modelo final não é uma tarefa fácil. Além dos critérios formais, devemos responder às seguintes questões:

- O modelo faz sentido?
- O modelo é útil para o objetivo pretendido? Se, por exemplo, o custo da coleta dos dados de uma variável é exorbitante e impossível de ser obtido, isso resultará em um modelo sem utilidade.
- Todos os coeficientes são razoáveis, ou seja, os sinais e magnitude dos valores fazem sentido e os erros padrões são relativamente pequeno?
- A adequabilidade do modelo é satisfatória? Sem outliers, tem variância constante, normalidade e os dados são independentes?

Um princípio a ser levado em consideração é o "princípio da parcimônia": modelos mais simples devem ser escolhidos aos mais complexos, desde que a qualidade do ajuste seja similar.

Bibliografia:

Neter ,J.; Wasserman, William; Kutner, M.H., Applied linear statistical models;
Draper,N.R.; Smith,H., Applied Regression Analysis.
Montgomery and Peck, Introduction to Linear Regression Analysis;
Seber, G.A.F., Linear Regression Analysis.
Myers and Montgomery, Generalized Linear Models.