Análise de Dados Categóricos
Capítulo 5- Seleção e validação do modelo

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Novembro, 2023

- **5.1 Seleção de variáveis**
- **5.1.1 Visão geral da seleção de variáveis**
- **5.1.2 Critérios de comparação de modelos**
- **5.1.3 Regressão de todos os subconjuntos**
- **5.1.5 Métodos modernos de seleção de variáveis**
- **5.1.6 Média do modelo**

A seleção de variáveis refere-se ao processo de redução do tamanho do modelo de um número potencialmente grande de variáveis para um conjunto mais gerenciável e interpretável.

Existem muitas abordagens para selecionar um subconjunto de variáveis de um conjunto maior. Todos têm pontos fortes e fracos, e novas abordagens estão sendo desenvolvidas a cada ano. Apresentaremos primeiro os métodos que são historicamente os mais usados e, de fato, ainda são de uso comum hoje.

No entanto, esses métodos não são mais recomendados por especialistas em seleção de variáveis. Apresentaremos críticas a esses métodos e ofereceremos alternativas geralmente mais confiáveis.

5.1.1 Visão geral da seleção de variáveis

Por que fazer seleção de variáveis? Por que não simplesmente usar o modelo com todas as variáveis disponíveis nele? Há duas respostas para essas perguntas.

Primeiro, em muitas aplicações modernas, o número de variáveis disponíveis pode ser muito grande para formar um modelo interpretável. Parte do objetivo da análise pode ser tentar entender quais variáveis estão realmente relacionadas à resposta. Além disso, em áreas como genética e medicina, o número de variáveis disponíveis pode ser maior que o número de observações, problema conhecido como "$p \gg n$". Nessas aplicações, o modelo completo não pode ser ajustado, o que significa que algum tipo de seleção de variável é necessária.

A segunda resposta tem a ver com algo chamado tradeoff viés-variância.

Os erros cometidos pelas previsões de um modelo são resultado de duas fontes: viés e variância. O viés na especificação de um modelo inclui a incapacidade de adivinhar a relação correta entre a resposta e as variáveis do modelo, bem como erros na escolha das variáveis usadas no modelo.

O uso de todas as variáveis disponíveis parece ser uma solução para o segundo desses dois problemas e, de fato, minimiza o viés devido à seleção de variáveis.
A seleção de variáveis é, portanto, uma parte comum e importante de muitas análises estatísticas.

Todas as técnicas de seleção de variáveis que discutimos se aplicam a qualquer um dos modelos de probabilidade. Portanto, começamos assumindo que selecionamos um modelo de probabilidade que acreditamos ser apropriado para o nosso problema; por exemplo, Poisson, binomial ou multinomial e que existe um "conjunto" de P variáveis explicativas, $x_1, \cdots, x_P$, que são candidatos a inclusão em um modelo de regressão.

De acordo com a notação do modelo linear generalizado, assumimos que os parâmetros do modelo; por exemplo, a média μ para um modelo de Poisson ou a probabilidade π para um binomial estão ligados aos parâmetros de regressão $\beta_0, \beta_1, \cdots, \beta_P$ por meio de uma função de ligação apropriada $g(\cdot)$, por exemplo, $\log$ ou *logit*. Portanto, expressamos todos os nossos modelos em termos do preditor linear $g(\cdot) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P$.

Obviamente, queremos que o modelo final contenha todas as variáveis importantes e nenhuma das não importantes. Se o conhecimento ou alguma teoria sobre o problema nos diz que certas variáveis são definitivamente importantes, elas devem ser incluídas no modelo no início e não são mais consideradas para seleção.

Da mesma forma, se alguma variável for conhecida por não estar relacionada à resposta, ela poderá ser imediatamente excluída do modelo.

As P variáveis restantes formam o pool sobre o qual o processo de seleção deve se concentrar. Este pool de variáveis pode incluir transformações de outras variáveis no pool ou interações entre variáveis.

Precisamos de alguma forma criar modelos a partir dessas variáveis, compará-los usando algum critério e selecionar variáveis ou modelos que forneçam os melhores resultados.

Primeiro discutimos critérios populares para comparar modelos. Em seguida, descrevemos algoritmos para criar modelos que podem ser comparados usando esses critérios.

5.1.2 Critérios de comparação de modelos

A comparação de dois modelos pode ser feita usando um teste de hipótese, desde que um dos modelos esteja aninhado ou encaixado no outro. Ou seja, o modelo maior deve conter todas as mesmas variáveis do modelo menor, mais pelo menos uma variável adicional.

Por exemplo, os modelos

$$g(\cdot) = \beta_0 + \beta_1 x_1$$

e

$$g(\cdot) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

podem ser comparados usando um teste de hipótese, enquanto os modelos $g(\cdot) = \beta_0 + \beta_1 x_1$ e $g(\cdot) = \beta_0 + \beta_2 x_2$ não pode.

Isso limita severamente o uso de testes de hipóteses para seleção de variáveis. Em vez disso, é necessário um critério que possa avaliar o quão bem qualquer modelo explica os dados.

Numerosas versões de critérios de informação foram propostas que usam penalidades diferentes para o tamanho do modelo.

Seja n o tamanho da amostra, r seja o número de parâmetros no modelo, incluindo parâmetros de regressão, interceptos e quaisquer outros parâmetros no modelo, como a variância na regressão linear normal e seja

$$\log\left(L\left(\widehat{\beta}\,|\,y_1,\cdots .y_n\right)\right)$$

a expressão da função de log-verossimilhança de um modelo estimado, avaliado nos MLEs para os parâmetros.

A forma geral da maioria dos critérios de informação é

$$IC(k) = -2\log\left(L\left(\widehat{\beta}\,|\,y_1, \cdots .y_n\right)\right) + kr,$$

onde k é um coeficiente de penalidade escolhido antecipadamente e usado para todos os modelos.

Para um determinado k, valores menores de $IC(k)$ indicam modelos que possuem uma grande verossimilhança em relação à sua penalidade e, portanto, são melhores modelos de acordo com o critério.

Os três critérios de informação mais comuns são:

1- Critério de informação de Akaike:

$$AIC = IC(2) = 2\log\left(L\left(\widehat{\beta}\,|\,y_1, \cdots .y_n\right)\right)) + 2r$$

2- AIC corrigido:

$$AIC_c = IC(2n/(n-r-1)) = -2\log\left(L\left(\widehat{\beta}\,|\,y_1, \cdots .y_n\right)\right))$$

$$+ \frac{2n}{n-r-1}r = AIC + \frac{2r(r+1)}{n-r-1}$$

3- Critério de Informação Bayesiana

$$BIC = IC(\log(n)) = 2\log\left(L\left(\widehat{\beta}\,|\,y_1, \cdots .y_n\right)\right) + \log(n)r$$

O coeficiente de penalidade k para o AIC_c é sempre maior do que para AIC. Aumenta à medida que o tamanho do modelo aumenta, mas torna-se mais semelhante à penalidade do AIC quando o tamanho da amostra é grande em relação ao tamanho do modelo. A penalidade BIC torna-se mais severa quando n é grande. É maior que a penalidade do AIC para todo $n \geq 8$ e maior que a penalidade do AIC_c sempre que $n - 1 - (2n/\log(n)) > r$, que é o caso em muitos problemas. Os critérios de informação são usados para a seleção de variáveis como segue.

Suponha que temos um grupo de modelos que diferem nas variáveis de regressão que usam. Selecionamos a k e calculamos $IC(k)$ em todos os modelos.

O modelo com o menor valor de $IC(k)$ é o preferido por aquele critério de informação. Observe que isso não requer o aninhamento de modelos, o que dá aos critérios de informação uma vantagem distinta sobre os testes de hipóteses para seleção de variáveis.

É possível que muitos modelos tenham valores de $IC(k)$ próximos do melhor modelo. Grosso modo, os modelos cujos valores de $IC(k)$ estão dentro de cerca de 2 unidades do melhor são considerados como tendo um ajuste semelhante ao melhor modelo, Burnham and Anderson, 2002.

Quanto mais modelos estiverem próximos do melhor, menos definitiva será a seleção do melhor modelo e maior a chance de que uma pequena alteração nos dados resulte na seleção de um modelo diferente. Por esse motivo, a média do modelo está se tornando mais popular na seleção de variáveis.

Observe que usar valores maiores de k resulta em um critério que favorece modelos menores. Assim, o AIC tende a escolher modelos maiores que o BIC e o AIC_c, enquanto o BIC geralmente escolhe modelos menores que o AIC_c. As opiniões variam sobre qual deles é o melhor, pois existem resultados teóricos que apóiam o uso de cada um deles. O BIC tem uma propriedade chamada consistência, o que significa que, à medida que n aumenta, ele escolhe o modelo "certo", ou seja, aquele com as mesmas variáveis explicativas do modelo do qual os dados surgem com probabilidade próxima de 1, assumindo que o modelo verdadeiro está entre os que estão sendo examinados.

O *AIC* possui uma propriedade chamada eficiência, o que significa que, assintoticamente, ele seleciona modelos que minimizam o erro quadrático médio da previsão, que leva em conta tanto o erro na estimativa dos coeficientes quanto a variabilidade dos dados.

O AIC_c foi desenvolvido para modelos lineares para estender a propriedade de eficiência para amostras menores. É usado nos modelos lineares generalizados para atingir aproximadamente o mesmo objetivo, embora não exista nenhuma teoria para mostrar que a correção consegue esse propósito.

Dado que modelos parcimoniosos são geralmente preferíveis a modelos com muitas variáveis, geralmente preferimos usar AIC_c ou *BIC* para seleção de variáveis.

Tanto o *AIC* quanto o *BIC* são geralmente calculados facilmente em R usando a função genérica **AIC()**, que pode ser aplicada a qualquer objeto de ajuste de modelo que produza uma log-verossimilhança que pode ser acessada por meio da função genérica **logLik()**.

Definir o valor do argumento **k = 2** em **AIC()** fornece o **AIC**, enquanto **k =** log**(n)** fornece o *BIC*, onde *n* é um objeto contendo o tamanho da amostra do conjunto de dados.

Não há uma maneira automática de calcular o AIC_c, mas geralmente não é difícil obtê-lo de um cálculo de *AIC* usando a segunda equação do ponto 2 acima.

5.1.3 Regressão de todos os subconjuntos

Dado um critério para comparar e selecionar modelos, o próximo passo é decidir quais modelos comparar. Talvez a coisa mais óbvia a considerar seja fazer todas as combinações possíveis de variáveis e selecionar aquela que "se encaixa melhor".

De fato, essa abordagem – chamada de regressão de todos os subconjuntos – é popular onde pode ser usada e a descrevemos com mais detalhes abaixo.

Quando há P variáveis explanatórias candidatas, incluindo quaisquer transformações e interações que possam ser de interesse, então existem 2^P modelos diferentes que devem ser formados.

Referimo-nos a este conjunto de todos os modelos possíveis como o espaço do modelo. A regressão de todos os subconjuntos para modelos lineares generalizados é limitada a problemas nos quais P não é muito grande, porque cada modelo deve ser ajustado usando técnicas numéricas iterativas.

Embora os computadores modernos possam ser muito rápidos, o escopo do problema pode ser esmagador.

Por exemplo, se $P = 10$, há pouco mais de 1.000 modelos a serem ajustados e isso pode ser feito rapidamente na maioria dos casos. Se $P = 20$; o número de modelos é superior a 1 milhão, enquanto se $P = 30$; o número é superior a 1 bilhão e os requisitos de tempo ou memória podem inviabilizar a busca exaustiva em todo o espaço do modelo.

Assumindo que a tarefa é possível, então para um k escolhido um critério de informação $IC(k)$ é computado para cada um dos 2^P modelos.

O modelo com o menor $IC(k)$ é considerado "melhor", embora seja bastante comum que existam inúmeros modelos com valores de $IC(k)$ semelhantes.

Isso é especialmente provável quando P é grande, de modo que há muitas variáveis possíveis cuja importância é limítrofe.

Exemplo 5.1: Placekicking.

5.1.5 Métodos modernos de seleção de variáveis

Nas últimas duas décadas, muitos novos algoritmos foram propostos para pesquisar um espaço de modelos.

O mais popular deles é o operador de seleção e encolhimento absoluto mínimo (LASSO) proposto por Tibshirani (1996).

O LASSO foi refinado, aprimorado e redesenvolvido de várias maneiras diferentes. Em seguida, descrevemos o LASSO original e mencionamos várias melhorias nele.

LASSO

Um procedimento de seleção de variável tem maior probabilidade de selecionar uma variável específica quando o acaso nos dados faz com que a variável pareça mais importante do que realmente é, em comparação com os momentos em que o acaso faz com que pareça menos importante do que realmente é.

Como resultado, as estimativas de parâmetros para as variáveis selecionadas tendem a ser enviesadas: elas são frequentemente estimadas mais longe de zero do que deveriam. Esta é a principal motivação para o LASSO, que tenta simultaneamente selecionar variáveis e encolher suas estimativas de volta a zero para neutralizar esse viés.

O LASSO estima parâmetros usando um procedimento que adiciona uma penalidade à log-verossimilhança para evitar que as estimativas de parâmetros sejam muito grandes.

Especificamente, para um modelo com p variáveis explicativas, os parâmetros LASSO estimados são

$$\widehat{\beta}_{0,LASSO}, \widehat{\beta}_{1,LASSO}, \cdots, \widehat{\beta}_{p,LASSO}$$

obtidos da maximização da função

$$\log\left(\beta_0, \beta_1, \cdots, \beta_p \,|\, y_1, \cdots, y_n\right) - \lambda \sum_{j=1}^{n} |\beta_j|,$$

onde λ é um parâmetro de ajuste que precisa ser determinado.

Para um dado valor de λ, o termo de penalidade $\sum_{j=1}^{n} |\beta_j|$ no critério de verossimilhança tem dois efeitos.

Faz com que algumas das estimativas dos parâmetros da regressão permaneçam em zero, de forma que suas variáveis correspondentes sejam consideradas excluídas do modelo.

Além disso, desencoraja que outras estimativas de parâmetros se tornem muito grandes, a menos que o aumento de suas magnitudes forneça um aumento suficiente na verossimilhança de superar a penalidade adicional. Assim, o procedimento simultaneamente seleciona variáveis e reduz suas estimativas para zero, o que pode ajudar a superar o viés observado acima.

À medida que o parâmetro de penalidade λ cresce, mais encolhimento ocorre e modelos menores são escolhidos.

Esse parâmetro geralmente é escolhido por validação cruzada (CV), que divide aleatoriamente os dados em vários grupos e prevê as respostas dentro de cada grupo com base em modelos ajustados ao restante dos dados. A comparação das respostas previstas e reais de alguma forma fornece uma estimativa do erro de previsão do modelo. O erro de predição é calculado para uma sequência de valores diferentes para λ, e aquele que tiver o menor erro de predição estimado é uma possível escolha para a penalidade.

Observe que o CV é um procedimento aleatório. É provável que erros de previsão ligeiramente diferentes resultem de execuções repetidas do algoritmo CV, levando a diferentes valores escolhidos e resultando em diferentes modelos e estimativas de parâmetros.

Além disso, muitas vezes existem muitos valores λ que levam a valores semelhantes de erro de previsão. Na prática, é comum usar o menor modelo cujo erro de previsão esteja dentro de 1 erro padrão do menor erro de CV, onde o erro padrão se refere à variabilidade das estimativas de erro de previsão de CV e pode ser calculado de várias maneiras.

Embora as estimativas LASSO geralmente resultem em melhores previsões do que os MLEs comuns, os procedimentos de inferência subsequentes ainda não foram totalmente desenvolvidos. Ainda não existem intervalos de confiança para parâmetros de regressão ou valores previstos. Portanto, o LASSO é usado como uma ferramenta de seleção de variáveis ou para fazer previsões onde as estimativas de intervalo não são necessárias.

5.5: Placekicking.

5.1.6 Média do modelo

Conforme observado na seção anterior, muitas vezes há muitos modelos com valores de IC(k) muito próximos do menor valor.

Esta é uma indicação de que há alguma incerteza sobre qual modelo é realmente melhor. Nesses casos, alterar ligeiramente os dados pode resultar na seleção de um modelo diferente como o melhor.

Essa incerteza é difícil de medir se a seleção de variáveis for aplicada usando a abordagem tradicional de selecionar um único modelo e usá-lo para todas as inferências posteriores. Em particular, se um modelo for selecionado e então ajustado aos dados, qualquer variável que não esteja nesse modelo implicitamente terá seus parâmetros de regressão estimados como zero, com um erro padrão de zero.

Ou seja, agimos como se soubéssemos que essas variáveis não pertencem ao modelo, quando na verdade os dados não podem determinar isso com certeza.

Seria útil poder estimar melhor a incerteza em cada uma de nossas estimativas de parâmetros. Quando a seleção variável de todos os subconjuntos é viável, a média do modelo pode levar em conta a incerteza da seleção do modelo em inferências subsequentes.

Em particular, a média do modelo bayesiano *BMA*, Hoeting et al. (1999) usa a teoria bayesiana para calcular a probabilidade de que cada modelo possível seja o modelo correto, supondo que um deles seja, de fato, correto.

Acontece que essas probabilidades podem ser aproximadas por funções relativamente simples do valor *BIC* para cada modelo.

Suponha que um total de M modelos sejam ajustados e denote o *BIC* para o modelo m por $BIC_m, m = 1, \cdots, M$.

Denote o menor valor de *BIC* entre todos os modelos por BIC_0 e defina

$$\Delta_m = BIC_m - BIC_0 \geq 0.$$

Assumindo que todos os modelos sejam igualmente prováveis antes do início da análise de dados, então a probabilidade estimada de que o modelo m esteja correto, τ_m, é aproximadamente

$$\widehat{\tau}_m = \frac{\exp\left(-\frac{1}{2}\Delta_m\right)}{\sum_{a=1}^{M}\exp\left(-\frac{1}{2}\Delta_a\right)}.$$

Para o modelo com o menor *BIC*, $\exp(-\Delta_m/2) = 1$; para todos os outros modelos, $\exp(-\Delta_m/2) < 1$ e esse número diminui muito rapidamente à medida que m cresce. Observe que quando $\Delta_m = 2$, $\exp(-\Delta_m/2) = 0.37$, de modo que a probabilidade do modelo m é cerca de um terço maior que a do modelo de melhor ajuste.

Esta não é uma grande diferença, considerando que os procedimentos de seleção de variáveis que selecionam um único modelo declaram essencialmente que um modelo tem probabilidade 1 e o resto tem probabilidade zero.

Este também é o significado por trás do relatório do número de modelos dentro de 2 unidades IC dos melhores em **glmulti()**.

Agora, seja θ qualquer quantidade que queremos estimar a partir dos modelos; como um parâmetro de regressão, uma razão de chances ou um valor previsto.

Denote sua estimativa no modelo m por $\widehat{\theta}_m$ com a variância correspondente estimada por $\widehat{\text{Var}}(\widehat{\theta}_m)$.

Em seguida, a estimativa média deste parâmetro na média do modelo é

$$\widehat{\theta}_{MA} = \sum_{m=1}^{M} \widehat{\tau}_m \widehat{\theta}_m,$$

e a variância deste estimador é estimada por

$$\widehat{\text{Var}}(\widehat{\theta}_{MA}) = \sum_{m=1}^{M} \widehat{\tau}_m \left((\widehat{\theta}_m - \widehat{\theta}_{MA})^2 + \widehat{\text{Var}}(\widehat{\theta}_m) \right).$$

Se um modelo é claramente melhor do que todos os outros, então é $\widehat{\tau}_m \approx 1$.

Nesse caso, a estimativa da média do modelo, $\widehat{\theta}_{MA}$, não é muito diferente da estimativa desse modelo e a variância de $\widehat{\theta}_{MA}$ é apenas a variância estimada de $\widehat{\theta}_m$ de seu modelo.

Por outro lado, se muitos modelos têm probabilidades comparáveis, então a variância da estimativa $\widehat{\theta}_{MA}$ depende tanto de quão variáveis são as estimativas dos parâmetros de modelo para modelo até $(\widehat{\theta}_m - \widehat{\theta}_{MA})^2$ e da variância das estimativas dos parâmetros de cada modelo.

Este processo pode ser aplicado a qualquer número de parâmetros associados ao modelo. Como exemplo, considere o caso de um único parâmetro de regressão, $\theta = \beta_j$.

Existe uma estimativa, $\widehat{\beta}_{j,m}$, em cada modelo. A estimativa média do modelo de β_j é

$$\widehat{\beta}_{j,MA} = \sum_{m=1}^{M} \widehat{\tau}_m \widehat{\beta}_{j,m}.$$

Observe aqui que metade dos modelos na seleção de todos os subconjuntos exclui X_j , então $\widehat{\beta}_{j,m} = 0$ para esses modelos.

Em modelos onde X_j aparece e que também possuem alta probabilidade a eles ligada, é provável que β_j tenha sua magnitude superestimada.

Esses fatos têm várias implicações:

1- A média do modelo geralmente resulta em uma estimativa diferente de zero para todos os parâmetros de regressão.

2- Variáveis realmente sem importância provavelmente não aparecerão nos modelos com alta probabilidade. Assim, suas estimativas médias de modelo são próximas de zero, porque as probabilidades para os modelos em que aparecem são todas pequenas e há pouca variabilidade nessas estimativas. Na prática, eles podem ser considerados zero, porque seu impacto nas previsões médias do modelo é insignificante.

3- Por outro lado, as variáveis que aparecem em alguns, mas não em todos, dos modelos com maior probabilidade normalmente têm suas estimativas reduzidas ao serem calculadas com zeros nos modelos em que não aparecem. Isso reduz o viés pós-seleção.

Exemplo 5.6: Placekicking.

Quando ajustamos um modelo linear generalizado (GLM), assumimos que as três coisas a seguir são verdadeiras:

(1) a distribuição ou componente aleatório, é aquela que especificamos (por exemplo, binomial ou Poisson);

(2) a média dessa distribuição está ligada às variáveis explicativas pela função g que especificamos, por exemplo, $g(\pi) = \beta_0 + \beta_1 x$, onde π é a média de uma variável de resposta binária Y e

(3) a função de ligação relaciona-se com as variáveis explicativas especificadas de forma linear.

Qualquer uma dessas suposições pode ser violada; na verdade, esperamos que um modelo seja apenas uma aproximação da verdade. Nesta seção apresentamos ferramentas que podem nos ajudar a identificar quando um modelo não é uma boa aproximação.

Os métodos que usamos são amplamente análogos aos usados na regressão linear normal, particularmente em sua dependência de resíduos. Para uma atualização sobre o uso de resíduos no diagnóstico de problemas com modelos lineares com erros normalmente distribuídos. Nesta seção, mostraremos como calcular e interpretar resíduos apropriados para diagnosticar problemas em distribuições em que a média e a variância estão relacionadas.

Também ofereceremos alguns métodos para testar o ajuste de um modelo. Essas técnicas sempre devem ser usadas como parte de um processo iterativo de construção de modelo. Um modelo é proposto, ajustado e verificado. Se estiver faltando, um novo modelo é proposto, ajustado e verificado e esse processo é repetido até que um modelo satisfatório seja encontrado.

Esta seção discute ferramentas que são usadas principalmente com modelos para uma única resposta de contagem por observação, como modelos binomiais e de Poisson. Historicamente, mais esforço tem sido gasto no desenvolvimento de ferramentas de diagnóstico para modelos de regressão binomial - particularmente, regressão logística - do que para outros modelos. Além disso, os modelos binomiais requerem consideração especial porque as contagens que são modeladas podem ser baseadas em diferentes números de tentativas para diferentes assuntos.

Portanto, concentramos nossa discussão em modelos binomiais e generalizamos para outros modelos para contagens conforme apropriado. Diagnósticos especializados para os modelos de regressão multinomial das Seções 3.3 e 3.4, que modelam respostas múltiplas simultaneamente, são menos desenvolvidos e não estão prontamente disponíveis em R.

Exemplo 5.7: Placekicking.

Exemplo 5.8: Placekicking.