

Critérios para seleção de modelos baseados na razão de verossimilhança

Prelecionista: Claudomiro Moura Gomes André

Orientador: Adair José Regazzi

A escolha do modelo apropriado, do ponto de vista estatístico, é um tópico extremamente importante na análise de dados (Bozdangan, 1987). Busca-se o modelo mais parcimonioso, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta. Nesta linha, diversos critérios para seleção de modelos são apresentados na literatura (ver Bozdogan, 1987; Wolfinger, 1993, Littel et. al, 2002). Dentre os critérios para seleção de modelos, os critérios baseados no máximo da função de verossimilhança (MFV) são os mais utilizados, com maior ênfase o Teste da Razão de Verossimilhança (TRV), o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC).

O teste da razão de verossimilhança é apropriado para testar dois modelos, desde que um dos modelos seja um caso especial do outro (modelos aninhados). O TRV usa a estatística L_R dada por: $L_R = 2 \times [L_2 - L_1]$ sendo L_2 o máximo do logaritmo natural da função de verossimilhança (MLFV) para o modelo mais parametrizado (M_2) e L_1 MLFV para modelo mais simples (M_1). Se M_1 for o modelo de melhor ajuste ao conjunto de dados L_R tem distribuição assintótica $\chi^2_{\delta, \nu}$, com δ o parâmetro de não centralidade e ν graus de liberdade, e ν a diferença entre o número de parâmetros dos modelos. A hipótese que o modelo M_1 apresenta melhor ajuste é rejeitada caso $L_R > \chi^2_{\alpha, \nu}$.

O Critério de Informação de Akaike (AIC) admite a existência de um modelo “real” que descreve os dados que é desconhecido, e tenta escolher dentre um grupo de modelos avaliados, o que minimiza a divergência de Kullback-Leibler (K-L). O valor de K-L para um modelo f com parâmetros θ , em relação ao modelo “real” representado por g é

$$I(g, f; \theta) = \int g(y) \ln \left(\frac{g(y)}{f(y | \theta)} \right) dy$$
. Esta divergência está relacionada à informação perdida por se usar um modelo aproximado e não o “real”. A estimativa do AIC para um determinado modelo é dada por: $AIC = -2L + 2k$ em que, L o MLFV do modelo com os parâmetros θ e k o número de parâmetros. O modelo com menor valor de AIC é considerado o modelo de melhor ajuste.

O Critério Bayesiano de Schwarz (BIC) tem como pressuposto a existência de um “modelo verdadeiro” que descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção. Assim o critério é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O valor do critério BIC para um determinado modelo é dado por: $BIC = -2L + 2k \ln(n)$, com n o número de observações. O modelo com menor BIC é considerado o de melhor ajuste.

Os três critérios apresentados apesar de conceitualmente diferentes acerca dos modelos em avaliação, utilizam o mesmo critério estatístico, o máximo da função de verossimilhança como medida do ajustamento, entretanto, definem valores críticos diferentes. Esta é a diferença fundamental entre os três métodos. Com o teste da razão de verossimilhança, considera-se por hipótese que o modelo mais simples é o de melhor ajuste, até que se observem, dado um nível α de significância, diferenças estatísticas para um modelo mais completo. Utilizando-se o AIC admite-se que dentre os modelos avaliados nenhum é considerado o que realmente descreve a relação entre a variável dependente e as variáveis explanatórias, ou o “modelo verdadeiro” e então, tenta-se escolher o modelo que minimize a divergência (K-L). Com o Critério Bayesiano de Schwarz (BIC), está implícito que existe o modelo que descreve a relação entre as variáveis envolvidas e o critério tenta maximizar a

probabilidade de escolha do verdadeiro modelo.

Referências

- Bozdongan. H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. **Psychometrika**. v.52, n.3, 345-370, Sep. 1987.
- Wolfinger, R. D. Covariance estrutura selection in general mixed models. **Communications in Statistics**. V.22. p1079-1106. 1993.
- Littell, R. C.; Milliken, G. A. Stroup, W. W & Wolfinger, R. D. **SAS System for Mixed Models**. Cary: Statistical Analysis System Institute, 2002. 633p.

Claudomiro Moura Gomes André

Adair José Regazzi