Testes de hipóteses em modelos multivariados de covariância linear generalizada

Qualificação de Mestrado

Lineu Alberto Cavazani de Freitas
Prof. Wagner Hugo Bonat
Prof. Marco Antônio Zanata Alves

Programa de Pós-Graduação em Informática
Data Science & Big Data Research Group
Universidade Federal do Paraná

2021

Sumário

Introdução

Ciência de dados

- A ciência de dados é vista como um campo de estudo de natureza interdisciplinar que incorpora conhecimento de grandes áreas como estatística, ciência da computação e matemática (LEY; BORDAS, 2018).
- Tem diversos campos de interesse.
- Os métodos estatísticos são de fundamental importância em grande parte das etapas da ciência de dados (WEIHS; ICKSTADT, 2018).
- Neste sentido, os modelos de regressão tem papel importante.

Modelos de regressão

Para entender minimamente um modelo de regressão, é necessário compreender o conceito de **fenômeno aleatório**, **variável aleatória** e **distribuição de probabilidade**.

- Um **fenômeno aleatório** é situação na qual diferentes observações podem fornecer diferentes desfechos.
- **Variáveis aleatórias** associam um valor numérico a cada desfecho possível do fenômeno. Podem ser discretas ou contínuas.
- Existem probabilidades associadas aos valores de uma variável aleatória. Estas probabilidades podem ser descritas por funções:
  - função de probabilidade, para variáveis aleatórias discretas.
  - função densidade de probabilidade, para variáveis aleatórias contínuas.

Modelos de regressão

- Modelos probabilísticos que buscam descrever as probabilidades de variáveis aleatórias, as chamadas **distribuições de probabilidade**.
- Em problemas práticos, podemos buscar uma distribuição de probabilidades que melhor descreva o fenômeno de interesse.
- Estas distribuições são descritas por funções.
- Estas funções possuem parâmetros que controlam aspectos da distribuição.
- Os parâmetros são quantidades desconhecidas estimadas através dos dados.

Modelos de regressão

- Na análise de regressão busca-se modelar os parâmetros das distribuições de probabilidade como uma função de outras variáveis.
- Isto é feito através da decomposição do parâmetro da distribuição em outros parâmetros, chamados de parâmetros de regressão.
- Assim, o objetivo dos modelos de regressão consiste em obter uma equação que explique a relação entre as variáveis explicativas e o parâmetro de interesse da distribuição de probabilidades selecionada para modelar a variável aleatória.
- Em geral, o parâmetro de interesse da distribuição de probabilidades modelado em função das variávis explicativas é a média.

Modelos de regressão

- O processo de análise via modelo de regressão parte de um conjunto de dados.
- Pode-se usar um modelo para modelar a relação entre a média de uma variável aleatória e um conjunto de variáveis explicativas.
- Assume-se que a variável aleatória segue uma distribuição de probabilidades e que o parâmetro de média desta distribuição pode ser descrito por uma combinação linear de parâmetros de regressão associados às variáveis explicativas.
- A obtenção destes parâmetros estimados se dá na chamada etapa de ajuste do modelo.
- Fazendo uso da equação resultante do processo é possível estudar a importância das variáveis explicativas sobre a resposta e realizar predições da variável resposta com base nos valores observados das variáveis explicativas.

Modelos de regressão

- Existem modelos uni e multivariados.
- Nos modelos univariados há apenas uma variável resposta e temos interesse em avaliar o efeito das variáveis explicativas sobre essa única resposta.
- No caso dos modelos multivariados há mais de uma resposta e o interesse passa a ser avaliar o efeito dessas variáveis sobre todas as respostas.
- Existem inúmeras classes de modelos de regressão, mencionaremos neste trabalho três importantes classes:
  - Modelos lineares.
  - Modelos lineares generalizados.
  - Modelos multivariados de covariância linear generalizada.

Modelo linear normal

- No cenário univariado, durante muitos anos o modelo linear normal (GALTON, 1886) teve papel de destaque.
- Muito usado principalmente por suas facilidades computacionais.
- Um dos pressupostos do modelo linear normal é de que a variável resposta, condicional às variáveis explicativas, segue a distribuição normal.
- Quando tal pressuposto não era atendido, uma alternativa, por muito tempo adotada, foi buscar uma transformação da variável resposta, tal como a família de transformações Box-Cox (BOX; COX, 1964).

Modelos lineares generalizados

- O avanço computacional permitiu a proposição de modelos mais complexos, que necessitavam de processos iterativos para estimação dos parâmetros (PAULA, 2004).
- A proposta de maior renome foram os modelos lineares generalizados (GLM) (NELDER; WEDDERBURN, 1972).
- Essa classe de modelos permitiu a flexibilização da distribuição da variável resposta de tal modo que esta pertença à família exponencial de distribuições.
- Em meio aos casos especiais de distribuições possíveis nesta classe de modelos estão a Bernoulli, binomial, Poisson, normal, gama, normal inversa, entre outras.

Modelos multivariados de covariância linear generalizada

- Há casos em que são coletadas mais de uma resposta por unidade experimental e há o interesse de modelá-las em função de um conjunto de variáveis explicativas.
- Neste cenário surgem os modelos multivariados de covariância linear generalizada (McGLM) (BONAT; JØRGENSEN, 2016).
- Esta classe pode ser vista com uma extensão multivariada dos GLMs que permite lidar com múltiplas respostas de diferentes naturezas e, de alguma forma, correlacionadas.
- O McGLM é uma classe flexível ao ponto de ser possível chegar a extensões multivariadas para modelos de medidas repetidas, séries temporais, dados longitudinais, espaciais e espaço-temporais.

Testes de hipóteses

- Em regressão, um interesse comum é o de verificar se a retirada de determinada variável explicativa do modelo geraria uma perda no ajuste.

- Isto é feito através dos chamados testes de hipóteses.

- Testes de hipóteses são ferramentas estatísticas que auxiliam no processo de tomada de decisão sobre valores desconhecidos (parâmetros) estimados por meio de uma amostra (estimativas).

- Podemos atribuir a teoria, formalização e filosofia dos testes de hipótese a Neyman, Pearson e Fisher.

Testes de hipóteses

No contexto de modelos de regressão, três testes de hipóteses são comuns, todos baseados na função de verossimilhança:

- O teste da razão de verossimilhanças (WILKS, 1938).
- O teste Wald (WALD, 1943).
- O teste do multiplicador de lagrange, também conhecido como teste escore (AITCHISON; SILVEY, 1958), (SILVEY, 1959), (RAO, 1948).

LRT, WT, LMT

- Os três testes podem ser usados para verificar se a retirada de determinada variável do modelo prejudica o ajuste.
- No caso do teste de razão de verossimilhanças, dois modelos precisam ser ajustados.
- Já o teste Wald e o escore necessitam de apenas um modelo.
- Os testes são assintóticamente equivalentes.
- Em amostras finitas estes testes podem apresentar resultados diferentes (EVANS; SAVIN, 1982).

Técnicas baseadas em testes de hipóteses

- Existem técnicas como a análise de variância (ANOVA) (FISHER; MACKENZIE, 1923).
- O objetivo da técnica é a avaliação do efeito de cada uma das variáveis explicativas sobre a resposta.
- Isto é feito através da comparação via testes de hipóteses entre modelos com e sem cada uma das variáveis explicativas.
- Permite que seja possível avaliar se a retirada de cada uma das variáveis gera um modelo significativamente pior quando comparado ao modelo com a variável.
- Para o caso multivariado extende-se a técnica para a análise de variância multivariada (SMITH; GNANADESIKAN; HUGHES, 1962), a MANOVA.

Proposta

- Considerando os McGLMs, não há discussão a respeito da construção de testes de hipóteses.
- Nosso objetivo geral é o desenvolvimento de testes de hipóteses para os McGLMs.
- Buscamos propor uma adaptação do teste de Wald clássico utilizado em modelos lineares para os McGLMs.

Proposta

Nosso trabalho tem os seguintes objetivos específicos:

- Adaptar o teste Wald para realização de testes de hipóteses gerais sobre parâmetros de McGLMs.
- Implementar funções para efetuar tais testes, bem como funções para efetuar ANOVAs e MANOVAs para os McGLMs.
- Avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação.
- Avaliar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

Revisão de literatura

Revisão de literatura

A revisão de literatura compreeende 2 temas:

- Modelos multivariados de covariância linear generalizada.
- Testes de hipóteses.

Modelos multivariados de covariância linear generalizada

- Os GLM são uma forma de modelagem para lidar com apenas uma resposta para dados de diferentes naturezas.

- É uma classe de modelos flexível e aplicável a diversos tipos de problema.

- Apresenta três importantes restrições:
  - A incapacidade de lidar com observações dependentes.
  - A incapacidade de lidar com múltiplas respostas simultaneamente.
  - Leque reduzido de distribuições disponíveis.

- Com o objetivo de contornar estas restrições, foram propostos os chamados Modelos Multivariados de Covariância Linear Generalizada (McGLM).

- Vamos discutir os McGLM como uma extensão dos GLM.

GLM

Considere:

- Y um vetor $N \times 1$ de valores observados da variável resposta.
- X uma matriz de delineamento $N \times k$
- β um vetor de parâmetros de regressão $k \times 1$.

GLM

um GLM pode ser descrito da forma

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X\beta}), \\
\mathrm{Var}(\boldsymbol{Y}) = \Sigma = \mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2}\left(\tau_0 \boldsymbol{I}\right)\mathrm{V}\left(\boldsymbol{\mu}; p\right)^{1/2},
\end{aligned} \tag{1}
$$

Em que:

- $g(.)$ é a função de ligação.
- $\mathrm{V}\left(\boldsymbol{\mu}; p\right)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor $\boldsymbol{\mu}$.
- p é o parâmetro de potência.
- τ_0 o parâmetro de dispersão.
- $\boldsymbol{I}$ é a matriz identidade de ordem $N \times N$.

GLM

1. Função de variância potência.
    - caracteriza a família Tweedie de distribuições.
    - função de variância é dada por $\vartheta(\boldsymbol{\mu}; p) = \mu^p$
    - casos particulares: normal (p = 0), Poisson (p = 1), gama (p = 2) e normal inversa (p = 3).
    - (JØRGENSEN, 1987) e (JØRGENSEN, 1997).
2. Função de dispersão Poisson–Tweedie.
    - caracteriza a família Poisson-Tweedie de distribuições
    - visa contornar a inflexibilidade da utilização da função de variância potência para respostas discretas.
    - função de dispersão dada por $\vartheta(\boldsymbol{\mu}; p) = \mu + \mu^p$
    - casos particulares os mais famosos modelos para dados de contagem: Hermite (p = 0), Neyman tipo A (p = 1), binomial negativa (p = 2) e Poisson–inversa gaussiana (p = 3)
    - (JØRGENSEN; KOKONENDJI, 2015).
3. Função de variância binomial.
    - dada por $\vartheta(\boldsymbol{\mu}) = \mu(1 - \mu)$
    - utilizada quando a variável resposta é binária, restrita a um intervalo ou quando tem-se o número de sucessos em um número de tentativas.

cGLM

- Alternativa para problemas em que a suposição de independência entre as observações não é atendida.
- A solução proposta é substituir a matriz identidade $\boldsymbol{I}$ da equação que descreve a matriz de variância e covariância por uma matriz não diagonal $\boldsymbol{\Omega}(\boldsymbol{\tau})$.
- A matriz $\boldsymbol{\Omega}(\boldsymbol{\tau})$ é descrita como uma combinação de matrizes conhecidas (ANDERSON et al., 1973) (POURAHMADI, 2000).

cGLM

A matriz $\boldsymbol{\Omega}(\boldsymbol{\tau})$ pode ser escrita como:

$$h\left\{\boldsymbol{\Omega}(\boldsymbol{\tau})\right\} = \tau_0 Z_0 + \ldots + \tau_D Z_D, \tag{2}$$

em que

- $h(.)$ é a função de ligação de covariância.
- Z_d com d = 0,. . ., D são matrizes que representam a estrutura de covariância presente nos dados.
- $\boldsymbol{\tau}$ = $(\tau_0, \ldots, \tau_D)$ é um vetor $(D+1) \times 1$ de parâmetros de dispersão.
- Tal estrutura pode ser vista como um análogo ao preditor linear para a média e foi nomeado como preditor linear matricial.

McGLM

- Pode ser entendido como uma extensão multivariada do cGLM.

- Contorna as principais restrições presentes nos GLM.

Considere

- $Y_{N\times R} = \{Y_1, \ldots, Y_R\}$ uma matriz de variáveis resposta
- $M_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ uma matriz de valores esperados.
- Σ_b, uma martiz de ordem $R \times R$, que descreve a correlação entre as variáveis resposta

Cada uma das variáveis resposta tem sua própria matriz de variância e covariância, responsável por modelar a covariância dentro de cada resposta, sendo expressa por

$$\Sigma_r = \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2} \boldsymbol{\Omega}_r\left(\boldsymbol{\tau}\right) \mathrm{V}_r\left(\boldsymbol{\mu}_r; p\right)^{1/2}. \tag{3}$$

McGLM

Um MCGLM é descrito como

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{Y}) &= \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\} \\
\mathrm{Var}(\boldsymbol{Y}) &= \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b,
\end{aligned}
\tag{4}
$$

em que

- $\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ é o produto generalizado de Kronecker.
- a matriz $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$.
- o operador Bdiag denota a matriz bloco-diagonal.
- $\boldsymbol{I}$ uma matriz identidade $N \times N$.
- Toda metodologia do McGLM está implementada no pacote *mcglm* (BONAT, 2018) do software estatístico R.

Funções de estimação

As funções de estimação para os parâmetros de regressão (função quasi-score) e de dispersão (função de estimação de Pearson) são dadas por:

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^{\top} \boldsymbol{C}^{-1}(\mathcal{Y} - \mathcal{M})$$

$$\psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}(W_{\boldsymbol{\lambda} i}(\boldsymbol{r}^{\top} \boldsymbol{r} - \boldsymbol{C})), i = 1, .., Q$$

Em que:

- $\boldsymbol{\beta}_r$ denota um vetor $k_r \times 1$ de parâmetros de regressão.
- $\boldsymbol{\lambda}$ é um vetor $Q \times 1$ de parâmetros de dispersão.
- $\mathcal{Y}$ é um vetor $NR \times 1$ com os valores da matriz de variáveis respostas $Y_{N \times R}$ empilhados.
- $\mathcal{M}$ é um vetor $NR \times 1$ com os valores da matriz de valores esperados $M_{N \times R}$ empilhados.

- $\boldsymbol{D} = \nabla_{\boldsymbol{\beta}} \mathcal{M}$ é uma matriz $NR \times K$, e $\nabla_{\boldsymbol{\beta}}$ denota o operador gradiente.
- $W_{\boldsymbol{\lambda} i} = -\frac{\partial \boldsymbol{C}^{-1}}{\partial \boldsymbol{\lambda}_i}$
- $\boldsymbol{r} = (\mathcal{Y} - \mathcal{M})$

Distribuição assintótica e algoritmo de estimação

- Para resolver o sistema de equações $\psi_{\boldsymbol{\beta}} = 0$ e $\psi_{\boldsymbol{\lambda}} = 0$ faz-se uso do algoritmo Chaser modificado:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - S_{\boldsymbol{\beta}}^{-1} \psi \boldsymbol{\beta}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}),$$
$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} \alpha S_{\boldsymbol{\lambda}}^{-1} \psi \boldsymbol{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}).$$

- Seja $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}^\top)^\top$ o estimador baseado em funções de estimação de $\boldsymbol{\theta}$.
- A distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, J_{\boldsymbol{\theta}}^{-1}),$$

$J_{\boldsymbol{\theta}}^{-1}$ é a inversa da matriz de informação de Godambe, dada por

$$J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-\top},$$

em que $S_{\boldsymbol{\theta}}^{-\top} = (S_{\boldsymbol{\theta}}^{-1})^\top$.

Exemplo tópicos

- item 1.
- item 2.
- item 3.
- item 4.
- item 5.
- item 6.
- item 7.
- item 8.

Slides modelo

Exemplo tópicos

- item 1.
- item 2.
- item 3.
- item 4.
- item 5.
- item 6.
- item 7.
- item 8.

Slide sem rodapé

- item 1.
- item 2.
- item 3.
- item 4.

Slide com imagem na pasta pics

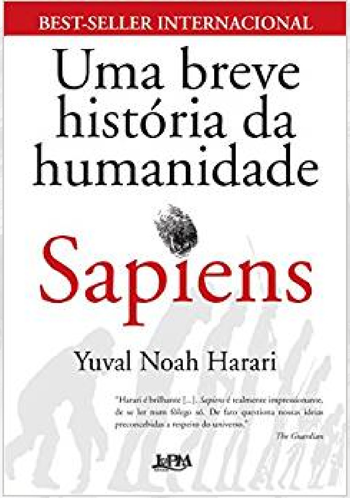

Figura 1. Harari, 2018

Slide com referencia

(DIGGLE; CHETWYND, 2011)

Slide com código R

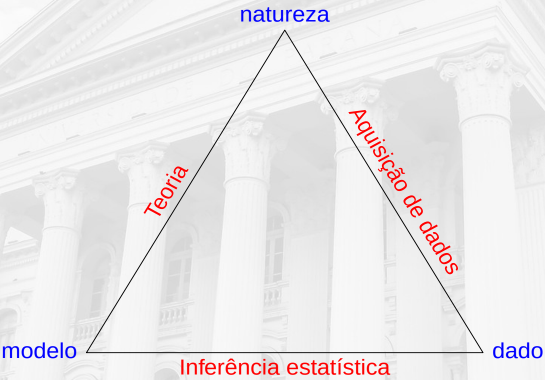

Figura 2. Estatística e o método científico (adaptado de (DIGGLE; CHETWYND, 2011)).

Referências bibliográficas

AITCHISON, J.; SILVEY, S. Maximum-likelihood estimation of parameters subject to restraints. **The annals of mathematical Statistics**, JSTOR, p. 813–828, 1958.

ANDERSON, T. et al. Asymptotically efficient estimation of covariance matrices with linear structure. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 1, n. 1, p. 135–141, 1973.

BONAT, W. H. Multiple response variables regression models in R: The mcglm package. **Journal of Statistical Software**, v. 84, n. 4, p. 1–30, 2018.

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 65, n. 5, p. 649–675, 2016.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 211–252, 1964.

DIGGLE, P. J.; CHETWYND, A. G. **Statistics and Scientific Method: An Introduction for Students and Researchers**. 1. ed. Oxford: Oxford University Press, 2011.

EVANS, G.; SAVIN, N. E. Conflict among the criteria revisited; the w, lr and lm tests. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 737–748, 1982.

FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation. ii. the manurial response of different potato varieties. **The Journal of Agricultural Science**, Cambridge University Press, v. 13, n. 3, p. 311–320, 1923.

GALTON, F. Regression towards mediocrity in hereditary stature. **The Journal of the Anthropological Institute of Great Britain and Ireland**, JSTOR, v. 15, p. 246–263, 1886.