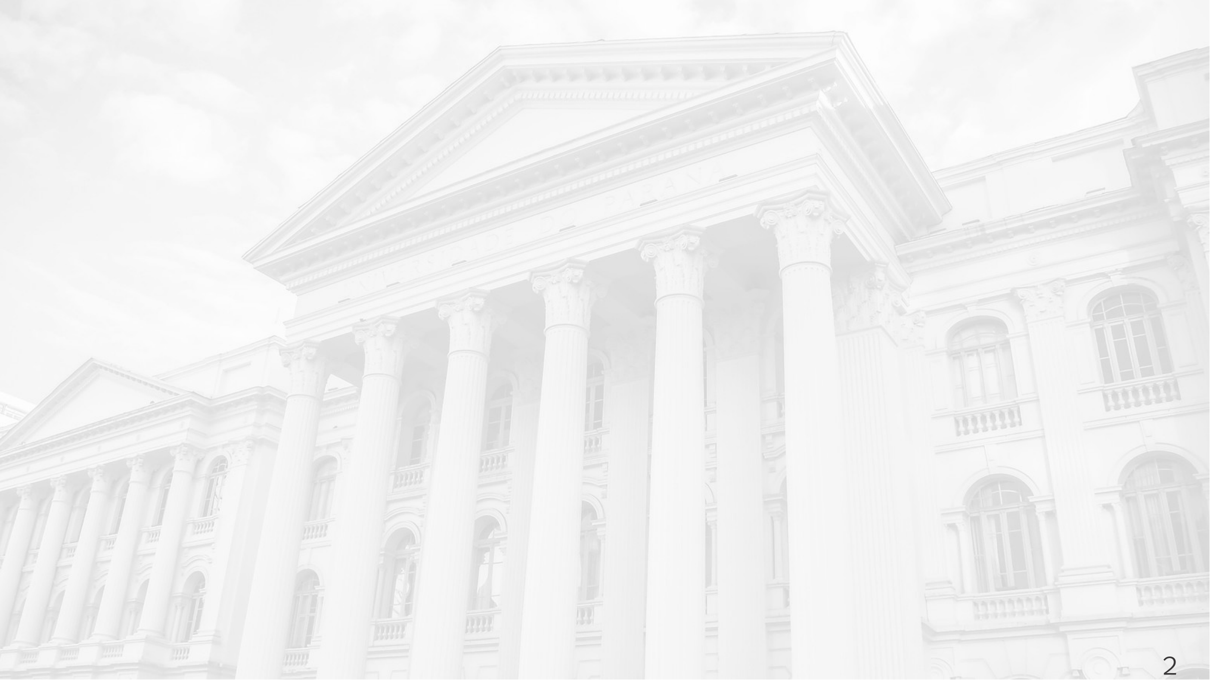
Teste Wald para avaliação de parâmetros de regressão e dispersão em modelos multivariados de covariância linear generalizada

Defesa

Lineu Alberto Cavazani de Freitas
Orientador: Prof. Dr. Wagner Hugo Bonat

PPG Informática UFPR



Sumário

Introdução

1. Motivação
   - Ciência de dados.
   - Modelos de regressão.
   - Testes de hipóteses.
   - Procedimentos baseados em testes de hipóteses.
2. Desafio e hipótese
3. Objetivo
4. Contribuição

Motivação

Motivação

Ciência de dados

- **Ciência de dados** é campo de estudo interdisciplinar que incorpora conhecimento de áreas como:
    1. **Estatística**.
    2. **Ciência da computação**.
    3. **Matemática**.

- Os **métodos estatísticos** são de fundamental importância em grande parte das etapas da ciência de dados [Weihs and Ickstadt, 2018].

- Neste sentido, os **modelos de regressão** tem papel importante.

Modelos de regressão

Três conceitos são importantes para entender minimamente o funcionamento de um modelo de regressão:

- **Fenômeno aleatório.**
- **Variável aleatória.**
- **Distribuição de probabilidade.**

Modelos de regressão

- **Fenômeno aleatório**: situação na qual diferentes observações podem fornecer diferentes desfechos.

- **Variáveis aleatórias**: mecanismos que 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.

- **Distribuições de probabilidade**: modelos probabilísticos que buscam descrever as probabilidades de variáveis aleatórias.

Modelos de regressão

- Na prática, 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** por meio dos dados.

Modelos de regressão

- Em regressão **modelamos parâmetros** das distribuições como uma função de **variáveis explicativas**.
- O parâmetro de interesse é decomposto em uma combinação linear de novos parâmetros que associam as **variáveis explicativas** à **variável resposta**.
- Obtém-se uma **equação que explique a relação** entre as variáveis.

Modelos de regressão

Distribuição
Normal, binomial, Poisson, beta, ...

Variável resposta

$$Y | x$$

Fatores experimentais, causas de variação.

$$Q(Y|x) = \eta(x, \theta)$$

Quantidade
Média, quantil, parâmetro, ...

Parâmetros
Empíricos, interpretáveis.

Função
Linear, não linear, semiparamétrica, ...

Explicativa
Métrica, categórica.

Modelos de regressão

1. **Definição do problema.**
   - Qual o fenômeno aleatório de interesse?
   - Que fatores externos podem afetar este fenômeno?
2. **Planejamento do estudo e coleta de dados.**
   - Estudo observacional x estudo experimental.
   - Representação tabular.
3. **Análise dos dados via regressão.**
   - Escolha da distribuição de probabilidade.
   - Especificação do modelo.
   - Obtenção dos parâmetros (ajuste).
   - Diagnóstico.
4. **Interpretação dos resultados.**
   - Quais os fatores externos apresentam ou não impacto sobre o fenômeno.
   - Qual a dimensão desse impacto.

Modelos de regressão

- Existem modelos univariados e multivariados.
  - **Univariados**: apenas uma variável resposta.
  - **Multivariados**: mais de uma variável resposta.
- Em ambos os casos o interesse é avaliar o **efeito de variáveis explicativas**.
- Existem inúmeras classes de modelos de regressão, dentre elas:
  - Modelo linear normal.
  - Modelos lineares generalizados.
  - **Modelos multivariados de covariância linear generalizada.**

Modelo linear normal

- O modelo linear normal [Galton, 1886] ficou famoso por suas **facilidades computacionais**.
- Possui **pressupostos** difíceis de serem atendidos na prática.
  - Independência.
  - Normalidade.
  - Variância constante.
- Diversas técnicas foram propostas para solucionar casos em que os pressupostos fossem violados.

Modelos lineares generalizados

- O **avanço computacional** permitiu o surgimento de modelos mais gerais que necessitavam de **processos iterativos** para estimação dos parâmetros.
- Surgem os modelos lineares generalizados(GLMs) [Nelder and Wedderburn, 1972].
- Os GLMs permitem utilizar qualquer membro da **família exponencial de distribuições**.
- Casos especiais: Bernoulli, binomial, Poisson, normal, gama, normal inversa, entre outras.

Modelos multivariados de covariância linear generalizada

- Apesar do grande potencial, os GLMs apresentam três importantes **restrições**:
    1. A incapacidade de lidar com **observações dependentes**.
    2. A incapacidade de lidar com **múltiplas respostas** simultaneamente.
    3. Leque reduzido de **distribuições disponíveis**.
- Os modelos multivariados de covariância linear generalizada (McGLMs) [Bonat and Jørgensen, 2016] contornam estas restrições.

Modelos multivariados de covariância linear generalizada

- Configuram uma estrutura geral para análise via modelos de regressão.
- Comporta múltiplas respostas de diferentes naturezas.
- Pode-se ajustar modelos com diferentes preditores e distribuições para cada resposta.
- Os modelos levam em conta a correlação entre indivíduos do conjunto de dados.

Modelos multivariados de covariância linear generalizada

- Os parâmetros são interpretáveis:
  - **Parâmetros de regressão**: efeito das variáveis explicativas sobre as respostas.
  - **Parâmetros de dispersão**: impacto da correlação entre unidades.
  - **Parâmetros de potência**: indicativo de qual distribuição se adequa ao problema.
  - **Parâmetros correlação**: força de associação entre respostas.

Testes de hipóteses em modelos de regressão

- Usados para verificar se a **retirada de determinada variável** explicativa do modelo geraria uma **perda no ajuste**.
- Os três testes mais usados são:
  - O teste da razão de verossimilhanças [Wilks, 1938].
  - O teste Wald [Wald, 1943].
  - O teste do multiplicador de lagrange ou teste escore [Aitchison and Silvey, 1958, Silvey, 1959, Rao, 1948].
- São baseados na função de verossimilhança dos modelos.
- São **assintóticamente equivalentes** [Engle, 1984].

Testes de hipóteses em modelos de regressão

Teste da razão de verossimilhanças

- Efetuado a partir de dois modelos com o objetivo de compará-los.
- Obter um modelo com todas as variáveis explicativas e um segundo modelo sem algumas dessas variáveis.
- O teste é usado para comparar estes modelos por meio da diferença do logaritmo da função de verossimilhança.

Testes de hipóteses em modelos de regressão

Teste Wald

- Requer apenas um modelo ajustado.
- Consiste em verificar se existe evidência para afirmar que um ou mais parâmetros são iguais a valores postulados.
- Avalia quão longe o valor estimado está do valor postulado.
- É possível formular hipóteses para múltiplos parâmetros,

Teste Escore

- Requer apenas um modelo ajustado.
- O modelo ajustado não possui o parâmetro de interesse
- O que é feito é testar se adicionar esta variável omitida resultará em uma melhora significativa no modelo.

Existe uma série de **procedimentos baseados em testes de hipóteses**, tais como:

- Análise de variância (ANOVA).
- Análise de variância multivariada (MANOVA).
- Testes de comparações múltiplas.

ANOVA & MANOVA

- Formas de **avaliar a significância** de cada uma das variáveis de uma forma procedural.
- Consiste em efetuar testes sucessivos impondo **restrições ao modelo** original.
- O objetivo é testar se a ausência de determinada variável gera perda ao modelo.
- Os resultados são sumarizados numa tabela, o chamado **quadro de análise de variância**.
- Caso univariado: ANOVA [Fisher and Mackenzie, 1923]. Caso multivariado: MANOVA [Smith et al., 1962].

Testes de comparações múltiplas

- Complementar às ANOVAs e MANOVAs
- São utilizados quando a análise de variância aponta como conclusão a existência de efeito significativo dos parâmetros associados a uma variável categórica.
- A análise de variância mostrará se há efeito de uma variável no modelo.
- Os testes de comparações múltiplas determinam **quais níveis diferem entre si**.

Temas abordados até aqui

- Ciência de dados.
- Modelos de regressão.
- Classes de modelos de regressão (ênfase nos **McGLMs**).
- Testes de hipóteses em modelos de regressão (ênfase no **teste Wald**).
- Procedimentos baseados em testes de hipóteses (ANOVA, MANOVA, testes de comparações múltiplas).

Desafio e hipótese

- Não há discussão a respeito da construção de testes de hipóteses para os McGLMs.

- Contudo, os McGLMs apresentam os elementos necessários para utilizar o teste Wald:

    1. Um vetor de estimativas dos parâmetros

    2. Uma matriz de variância e covariância destas estimativas.

- Das três opções clássicas de testes de hipóteses, o teste Wald se torna o mais atrativo para os McGLMs.

Objetivo

1. Propor a utilização do teste Wald para realização de testes de hipóteses gerais sobre parâmetros de regressão e dispersão de McGLMs.

2. Avaliar as propriedades e comportamento dos testes propostos com base em estudos de simulação.

3. Implementar funções em R para testes de hipóteses, ANOVA, MANOVA e testes de comparações múltiplas para os McGLMs.

4. Motivar o potencial de aplicação das metodologias discutidas com base na aplicação a conjuntos de dados reais.

Contribuição

- Formas de avaliar os parâmetros estimados pelos McGLMs.
- Fornecer ferramentas para uma melhor interpretação dos parâmetros estimados.
- Fornecer uma maneira procedural e segura de responder questões comuns no contexto de modelagem.
- Extrair mais informações e conclusões a respeito dos problemas modelados por meio dos McGLMs.

Referencial teórico

Referencial teórico

1. McGLMs
    - Elementos.
    - Preditores lineares e matriciais.
    - Funções de variância.
    - Parâmetros.
    - Estimação.
2. Testes de hipóteses
    - Elementos de um teste de hipóteses.
    - Testes de hipóteses em modelos de regressão.
    - Teste Wald
    - ANOVA e MANOVA.
    - Testes de comparações múltiplas.

Modelos multivariados de covariância linear generalizada

Modelos multivariados de covariância linear generalizada

Modelos multivariados de covariância linear generalizada

Para definição de um McGLM 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.
- X_r denota uma matriz de delineamento $N \times k_r$.
- $\boldsymbol{\beta}_r$ denota um vetor $k_r \times 1$ de parâmetros de regressão.

Modelos multivariados de covariância linear generalizada

Considere ainda:

- Σ_b uma matriz de correlação entre variáveis resposta, de ordem $R \times R$.
- Σ_r, $r = 1, \ldots, R$, a matriz de variância e covariância para cada resposta r, de dimensão NxN:

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

Em que:

- $\mathrm{V}_r\left(\boldsymbol{\mu}; p\right)$ é uma matriz diagonal em que as entradas principais são dadas pela função de variância aplicada ao vetor $\boldsymbol{\mu}$.
- p_r é o parâmetro de potência.
- $\boldsymbol{\Omega}\left(\boldsymbol{\tau}_r\right)$ a matriz de dispersão que descreve a parte da covariância dentro de cada variável resposta.

Preditor linear matricial

- A matriz $\boldsymbol{\Omega}(\boldsymbol{\tau}_r)$ descreve a estrutura de correlação entre as observações da amostra.

- É modelada através de um preditor linear matricial combinado com uma função de ligação de covariância:

$$h\left\{\boldsymbol{\Omega}(\boldsymbol{\tau}_r)\right\} = \tau_{r0} Z_0 + \ldots + \tau_{rD} Z_D$$

  - $h()$ é a função de ligação de covariância.

  - Z_{rd} com $d = 0, \ldots, D$ são matrizes que representam a estrutura de covariância presente em cada variável resposta r.

  - $\boldsymbol{\tau_r} = (\tau_{r0}, \ldots, \tau_{rD})$ é um vetor $(D + 1) \times 1$ de parâmetros de dispersão.

Funções de variância

1. **Função de variância potência** [Jørgensen, 1987, 1997].
    - Família Tweedie de distribuições.
    - $\vartheta\left(\boldsymbol{\mu}; p\right) = \mu^p$.
    - Casos particulares: normal (p = 0), Poisson (p = 1), gama (p = 2) e normal inversa (p = 3).
2. **Função de dispersão Poisson–Tweedie** [Jørgensen and Kokonendji, 2015].
    - Família Poisson-Tweedie de distribuições.
    - $\vartheta\left(\boldsymbol{\mu}; p\right) = \mu + \mu^p$.
    - Casos particulares: Hermite (p = 0), Neyman tipo A (p = 1), binomial negativa (p = 2) e Poisson–inversa gaussiana (p = 3).
3. **Função de variância binomial.**
    - $\vartheta(\boldsymbol{\mu}) = \mu(1 - \mu)$.
    - Acomoda respostas binárias ou restritas a um intervalo.

Modelos multivariados de covariância linear generalizada}

Os McGLMs são definidos por:

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\}$$

$$\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b$$

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.
- $\tilde{\boldsymbol{\Sigma}}_r$ denota a matriz triangular inferior da decomposição de Cholesky da matriz $\boldsymbol{\Sigma}_r$.
- Bdiag() denota a matriz bloco-diagonal.
- $\boldsymbol{I}$ uma matriz identidade $N \times N$.
- $g_r()$ são as tradicionais funções de ligação.

Modelos multivariados de covariância linear generalizada

- Parâmetros estimados nos McGLMs:
  1. Regressão.
  2. Dispersão.
  3. Potência.
  4. Correlação.

- Todas estas quantidades são interpretáveis e são estimadas com base nos dados.

- A estimação é feita por meio de **funções de estimação**.
  1. **Função quasi-score** para parâmetros de regressão.
  2. **Função de estimação de Pearson** para os demais parâmetros.

Funções de estimação

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

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

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 $\psi_{\beta} = 0$ e $\psi_{\lambda} = 0$ 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 $S_{\boldsymbol{\theta}}^{-\top} = (S_{\boldsymbol{\theta}}^{-1})^{\top}$.

Testes de hipóteses

Testes de hipóteses

Testes de hipóteses

- Inferência: **inferir** conclusões válidas a respeito de uma população por meio do estudo de uma amostra.

- Problemas de inferência estatística são:
  1. **Estimação** de parâmetros com base em informação amostral.
  2. **Testes de hipóteses**.
     - Com base na evidência amostral, podemos considerar que dado parâmetro tem determinado valor?

Testes de hipóteses

- São postuladas 2 hipóteses, chamadas de **nula** e **alternativa**.
- Avalia-se uma estatística de teste.
- Com base no valor da estatística e de acordo com sua distribuição de probabilidade, toma-se a decisão de rejeitar ou não rejeitar a hipótese nula.
- Seja θ um parâmetro, um teste de hipóteses sobre θ é dado por:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Testes de hipóteses

Desfechos possíveis:

	Rejeita H_0	**Não Rejeita** H_0
H_0 **verdadeira**	Erro tipo I	Decisão correta
H_0 **falsa**	Decisão correta	Erro tipo II

Tabela 1. Desfechos possíveis em um teste de hipóteses

- A probabilidade do erro do tipo I recebe o nome de nível de significância.
- A probabilidade de se rejeitar corretamente H_0 recebe o nome de poder do teste.
- A probabilidade de a estatística de teste tomar um valor igual ou mais extremo do que aquele que foi observado recebe o nome de valor-p.

Testes de hipóteses em modelos de regressão

- Modelos de regressão: modelar uma ou mais variáveis em função de um conjunto de variáveis explicativas.
- Modelos contêm parâmetros que são quantidades desconhecidas que estabelecem a relação entre as variáveis sob o modelo.
- Pode ser de interesse verificar se a retirada de uma ou mais variáveis do modelo gera um modelo significativamente pior que o original.
- Verificar se há evidência suficiente nos dados para afirmar que determinada variável explicativa não possui efeito sobre a resposta.

Teste Wald em modelos de regressão

- A ideia do teste consiste em verificar se existe evidência suficiente nos dados para afirmar que um ou mais parâmetros são iguais a valores especificados.
- Avalia a distância entre as estimativas dos parâmetros e um conjunto de valores postulados.
- Esta diferença é ainda padronizada por medidas de precisão das estimativas dos parâmetros.
- Quanto maior for esta distância padronizada, menores são as evidências a favor da hipótese de que os valores estimados são iguais aos valores postulados.

Teste Wald em modelos de regressão

Considere um modelo de regressão em que:

- β um vetor $k \times 1$ parâmetros de regressão.
- $\hat{\beta}$ as estimativas dos parâmetros.
- c um vetor de valores postulados de dimensão s.
- L uma matriz de especificação das hipóteses, de dimensão $s \times k$.

Teste Wald em modelos de regressão

- As hipóteses podem ser descritas como:

$$\begin{cases} H_0 : \boldsymbol{L\beta} = \boldsymbol{c} \\ H_1 : \boldsymbol{L\beta} \neq \boldsymbol{c} \end{cases}$$

- A estatística de teste é dada por:

$$WT = (\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T \, (\boldsymbol{L} \, \boldsymbol{Var}^{-1}(\hat{\boldsymbol{\beta}}) \, \boldsymbol{L}^T)^{-1} \, (\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}).$$

- WT $\sim \chi^2_s$.

ANOVA e MANOVA

- Testes sucessivos impondo restrições ao modelo original.
- O objetivo é testar se a ausência de determinada variável gera um modelo significativamente inferior.
- O quadro de ANOVA ou MANOVA contém em cada linha:
  1. A variável.
  2. O valor de uma estatística de teste referente à hipótese de nulidade de todos os parâmetros associados a esta variável.
  3. Os graus de liberdade desta hipótese.
  4. Um valor-p associado à hipótese testada naquela linha do quadro.
- É possível gerar quadros de análise de variância por meio do teste Wald.

Testes de comparações múltiplas

- Usado quando a ANOVA aponta para efeito significativo de uma variável categórica.
- Comparações aos pares a fim de detectar para quais níveis da variável categórica os valores da resposta se alteram.
- Pode ser avaliada utilizando o teste Wald.
- Por meio da correta especificação da matriz L, é possível avaliar hipóteses sobre qualquer possível contraste entre os níveis de uma determinada variável categórica.

Trabalhos relacionados

Trabalhos relacionados

Trabalhos relacionados

- Propostas que visam contornar restrições dos GLMs e como são efetuados testes de hipóteses para estas propostas.
- Propostas univariadas e multivariadas.
- Efeitos aleatórios.
- Correção de erros padrões.
- Grande variedade de modelos de regressão multivariados para fins específicos.

Trabalhos relacionados

- Efeitos aleatórios para acomodar correlação entre observações.
- Modelos lineares generalizados mistos (GLMM). Modelos lineares generalizados multivariados mistos (MGLMM).
- A interpretação dos parâmetros de regressão dependem de manter o efeito aleatório fixado.
- A estimação destes modelos não é simples. Envolve integrais complexas e é uma tarefa computacionalmente desafiadora.
- É possível usar máxima verossimilhança.
- Testes de hipóteses tradicionais costumam ser empregados.

Trabalhos relacionados

- Equações de estimação generalizadas (GEE).
- Alternativa para acomodar a correlação entre observações.
- Incluir no processo de estimação uma matriz de correlação de trabalho.
- O foco do método não é a modelagem da estrutura de correlação entre os indivíduos, mas sim a correção dos erros padrões.
- Testes de hipóteses tradicionais costumam ser empregados.

Trabalhos relacionados

- Modelos aditivos generalizados para locação, escala e forma (GAMLSS).
- Classe de modelos de regressão univariados com um número considerável de distribuições disponíveis.
- É possível modelar todos os parâmetros distribucionais.
- É possível incluir efeitos aleatórios e termos suavizadores.
- A estimação é feita com base verossimilhança penalizada.
- Testes de hipóteses tradicionais costumam ser empregados.

Trabalhos relacionados

Grande variedade de modelos de regressão multivariados para fins específicos.

- Análise de contagens multivariadas em que os testes usuais se aplicam [Zhang et al., 2017].
- Modelo de regressão multivariado com distribuição Poisson inversa gaussiana em que testes de hipóteses ao estilo da razão de verossimilhanças se aplicam [Mardalena et al., 2020].
- Modelo de regressão Poisson zero inflacionado multivariado em que o teste da razão de verossimilhanças e o teste Wald se aplicam [Sari et al., 2021].
- Modelo de regressão multivariado gamma em que um análogo do teste da razão de verossimilhanças e o teste Wald se aplicam Rahayu et al. [2020].

Teste Wald para McGLMs

Teste Wald para McGLMs

1. Definição das hipóteses.

2. Estatística de teste.

3. Distribuição.

4. Hipóteses comuns.

5. Construção da matriz L.

6. Exemplos.

Hipóteses

$$H_0 : \boldsymbol{L\theta}^* = \boldsymbol{c} \text{ vs } H_1 : \boldsymbol{L\theta}^* \neq \boldsymbol{c}.$$

Em que:

- $\boldsymbol{\theta}^*$ é o vetor de dimensão $h \times 1$ de parâmetros de regressão, dispersão e potência do modelo.
- Em que $\boldsymbol{L}$ é a matriz de especificação das hipóteses a serem testadas, tem dimensão $s \times h$.
- $\boldsymbol{c}$ é um vetor de dimensão $s \times 1$ com os valores sob hipótese nula.

Estatística de teste

$$W = (\boldsymbol{L}\hat{\boldsymbol{\theta}}^* - \boldsymbol{c})^T \; (\boldsymbol{L} \; \boldsymbol{J}^{*-1} \; \boldsymbol{L}^T)^{-1} \; (\boldsymbol{L}\hat{\boldsymbol{\theta}}^* - \boldsymbol{c}).$$

Em que:

- $\boldsymbol{L}$ é a matriz da especificação das hipóteses, tem dimensão $s \times h$.
- $\hat{\boldsymbol{\theta}}^*$ é o vetor de dimensão $h \times 1$ com todas as estimativas dos parâmetros de regressão, dispersão e potência.
- $\boldsymbol{c}$ é um vetor de dimensão $s \times 1$ com os valores sob hipótese nula.
- E $\boldsymbol{J}^{*-1}$ é a inversa da matriz de informação de Godambe desconsiderando os parâmetros de correlação, de dimensão $h \times h$.
- $W \sim \chi^2_s$

Hipóteses comuns

- Costuma ser de interesse formular:
  1. Hipóteses para parâmetros individuais.
  2. Hipóteses para múltiplos parâmetros.
  3. Hipóteses para avaliar igualdade de parâmetros
  4. Hipóteses sobre parâmetros de regressão ou dispersão para respostas sob mesmo preditor.
  5. Hipóteses sobre contrastes.
- O elemento chave é a correta especificação da matriz L.

Construção da matriz L

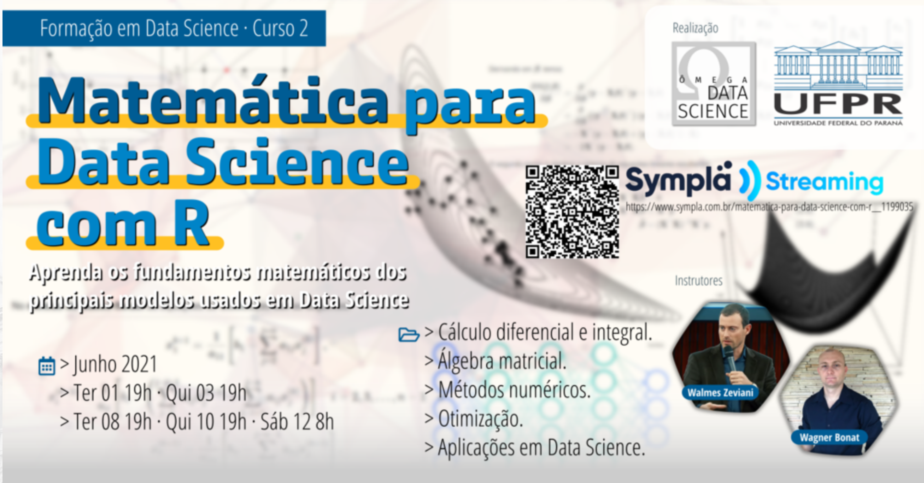
- Cada coluna da matriz L corresponde a um dos h parâmetros de θ^*.
- Cada linha corresponde a uma restrição.
- A matriz é composta por valores iguais a 0, 1 e eventualmente -1.
- O produto $L\theta^*$ deve resultar nas hipóteses de interesse.

Exemplo seção

Exemplo subseção

- bullet.

Exemplo figura



Outra seção

Exemplo slide 2 colunas

- R para data science.
  - Data: 04 a 15/05/2021.
  - **Consulte condições de acesso.**
- Matemática para data science.
  - Data: 01 a 12/05/2021.
- Probabilidade e Estatística para data science.
  - Data: 22/06 a 01/07/2021.
- Visualização, dashboards e relatórios dinâmicos.
  - Data: 06 a 17/07/2021.
- Modelagem estatística para data science.
  - Data: 03 a 14/08/2021.
- Planejamento de experimentos.
  - Data: 07 a 18/09/2021.
- Cursos previstos
  - Machine learning com R.
  - Mineração de texto com R.
  - Web Scraping com R.
  - R avançado.
  - R para big data (spark).

Exemplo texto e figura 2 colunas

- A Ômega Data Science é um projeto que objetiva construir, capacitar e conectar pessoas em uma comunidade focada em Data Science.
- Instagram: @omegadatascience
- Twitter: @omegadatascienc
- Telegram: <cutt.ly/omega_grupo_telegram>.
- YouTube: /OmegaDataScience

John Aitchison and SD Silvey. Maximum-likelihood estimation of parameters subject to restraints. *The annals of mathematical Statistics*, pages 813–828, 1958.

Wagner Hugo Bonat and Bent Jørgensen. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5): 649–675, 2016.

Robert F Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826, 1984.

R. A. Fisher and W. A. Mackenzie. Studies in crop variation. ii. the manurial response of different potato varieties. *The Journal of Agricultural Science*, 13(3):311–320, 1923. doi: 10.1017/S0021859600003592.

Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Bent Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145, 1987.

Bent Jørgensen. *The theory of dispersion models.* CRC Press, 1997.

Bent Jørgensen and Célestin C Kokonendji. Discrete dispersion models and their tweedie asymptotics. *AStA Advances in Statistical Analysis*, 100(1):43–78, 2015.

Selvi Mardalena, Purhadi Purhadi, Jerry Dwi Trijoyo Purnomo, and Dedy Dwi Prastyo. Parameter estimation and hypothesis testing of multivariate poisson inverse gaussian regression. *Symmetry*, 12(10):1738, 2020.

John Ashworth Nelder and Robert William Maclagan Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135:370–384, 1972.

Anita Rahayu, Dedy Dwi Prastyo, et al. Multivariate gamma regression: Parameter estimation, hypothesis testing, and its application. *Symmetry*, 12(5):813, 2020.

C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press, 1948.

Dewi Novita Sari, Purhadi Purhadi, Santi Puteri Rahayu, and Irhamah Irhamah. Estimation and hypothesis testing for the parameters of multivariate zero inflated generalized poisson regression model. *Symmetry*, 13(10):1876, 2021.

Samuel D Silvey. The lagrangian multiplier test. *The Annals of Mathematical Statistics*, 30 (2):389–407, 1959.

H Smith, R Gnanadesikan, and JB Hughes. Multivariate analysis of variance (manova). *Biometrics*, 18(1):22–41, 1962.

Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

Claus Weihs and Katja Ickstadt. Data science: the impact of statistics. *International Journal of Data Science and Analytics*, 6(3):189–194, 2018.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Yiwen Zhang, Hua Zhou, Jin Zhou, and Wei Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13, 2017.