Teste gradiente

Victor Meneses Navarro Fernandes

5 de dezembro de 2023

Tipicamente, na teoria de testes assintóticos, discutem-se três estatísticas: razão de verossimilhanças (Wilks, 1938, *Annals of Mathematical Statistics*), Wald (Wald, 1943, *Transactions of the American Mathematical Society*) e escore (Rao, 1948, *Proceedings of the Cambridge Philosophical Society*). Recentemente, Terrell (2002, *Computing Science and Statistics*) apresenta a estatística gradiente. A proposta tem recebido considerada atenção científica devido a suas propriedades e simplicidade de obtenção. Diferentemente das estatística de Wald e escore, a estatística gradiente não depende do cálculo das matrizes de Fisher observada ou esperada, bem como de suas respectivas inversas. Diante disto, uma variedade de estudos de simulação de Monte Carlo e refinamentos para amostras de tamanho pequeno e/ou moderado têm sido realizadas para investigar e comparar o desempenho da estatística gradiente em relação as demais em diferentes classes especiais de modelos paramétricos. Dessa maneira, o presente trabalho objetiva fazer uma revisão da estatística gradiente e reproduzir os resultados de Lemonte (2016, *The Gradient Test*) para comparação do desempenho da estatística gradiente no modelo linear simétrico homoscedástico. **Palavras chave**: testes assintóticos, estatística gradiente, modelo de regressão simétrico

1 Introdução

O procedimento de testar hipóteses antecede a formalização de Fisher, Neyman e Pearson no século 20. Por exemplo, no século 11, o governo britânico demonstrava preocupação com a adulteração de moedas de ouro ou prata. Dessa maneira, visando contornar esta problemática, foi estabelecida uma cerimônia judicial nomeada *Trial of the Pyx* (The Royal Mint, 2023) para averiguar a qualidade das referidas moedas reais. Durante a cerimônia, o peso das moedas coletadas era comparado com o peso esperado destas se fossem feitas com a quantidade apropriada de ouro ou prata.

A formulação e filosofia de testes de hipóteses atualmente empregadas foram desenvolvidas no período entre 1915-1933. Historicamente, faz-se a distinção entre a abordagem de Fisher e a de Neyman-Pearson (Lehmann, 1993). De um lado, Fisher fundamentou um teste como uma "prova" por contradição, em que dados são coletadas e confrontados com as afirmações preestabelecidas. Também, a ideia de valor descritivo (ou valor-p) é atribuida a ele, como medida de evidência contra a hipótese inicial. Por outro lado, Jerzy Neyman (1894-1981) e Ergon Pearson (1985-1980) estabeleceram um teste de hipótese como uma escolha entre hipóteses nula e alternativa. A abordagem dos autores se baseiam fortemente em taxas de erros para determinar regras de decisão.

Atualmente, a teoria de testes de hipóteses engloba as contribuições destes autores, tais como conceitos relacionados a valor descritivo, erro do tipo 1 e tipo 2, testes mais poderosos (MP) e uniformemente mais poderosos (UMP) e entre outros. Geralmente, para certos tipos de hipóteses, não é possível obter um teste uniformemente mais poderoso (UMP) e a distribuição da estatística sob a hipótese nula nem sempre é conhecida. Neste contexto, comumente, utilizam-se testes assintóticos com boas propriedades assintóticas.

Tipicamente, na teoria de testes assintóticos, discutem-se três estatísticas: razão de verossimilhanças (Neyman e Pearson, 1928; Wilks, 1938), Wald (Wald, 1943) e escore (Rao, 1948). Particularmente, Wilks (1938) deduz a distribuição assintótica da razão de verossimilhanças para testes de hipóteses compostos baseado no trabalho de Neyman e Pearson (1928). Uma notável desvantagem desta estatística envolve depender tanto do estimador de máxima verossimilhança (EMV) restrito e não-restrito para sua obtenção. Posteriormente, Wald (1943) desenvolve uma estatísica que requer somente o cálculo do EMV não-restrito, embora, não seja invariante sob reparametrizações e dependa da matriz de informação de Fisher esperada. Na mesma década, Rao (1948) propõe uma estatística de teste baseada na função escore como alternativa as estatísticas supracitadas. No contexto da econometria, a estatística escore recebeu uma interpretação com base em multiplicadores de Lagrange por Aitchinson e Silvey (1958) e Silvey (1958).

Recentemente, Terrell (2002) apresenta a estatística gradiente. A proposta tem recebido considerada atenção científica devido a suas propriedades e simplicidade de obtenção. Diferentemente das estatística de Wald e escore, a estatística gradiente não depende do cálculo das matrizes de Fisher observada ou esperada, bem como de suas respectivas inversas. Diante disto, uma variedade de estudos de simulação de Monte Carlo e refinamentos para amostras de tamanho pequeno têm sido realizadas para investigar e comparar o desempenho da estatística gradiente em relação as demais em diferentes classes especiais de modelos paramétricos.

Pode-se citar Lemonte e Ferrari (2011) em modelos de regressão Birnbaum-Saunders para dados censurados, Lemonte (2011) em modelos de regressão na família exponencial não-linear, Vargas et al. (2014) em modelos lineares generalizados e dentre outros.

Dessa maneira, o presente trabalho objetiva fazer uma revisão da estatística gradiente, incluindo conceitos básicos, suas propriedades e interpretação geométrica. Ainda, descreve-se o refinamento para amostras de tamanho pequeno e moderado. Por fim, realiza-se a replicação da simulação de Monte Carlo presente em Lemonte (2016) para comparação do desempenho da estatística gradiente no modelo linear simétrico homoscedástico.

2 Conceitos básicos

Seja $\boldsymbol{X}$ uma réplica da população $P \in \mathcal{P}$, $\mathcal{P}$ uma família identificável indexada por $\boldsymbol{\theta} \subset \mathbb{R}^p$, tal que dominada por uma medida finita, λ. Denote $\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta})$ a função de verossimilhança e $l(\boldsymbol{x};\boldsymbol{\theta})$ seu logaritmo natural. O vetor escore completo é definido como

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{x};\boldsymbol{\theta})$$

e a matriz de informação de Fisher

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{U}^{\top}(\boldsymbol{\theta})\right]$$

O estimador de máxima verossimilhança (EMV) é obtido resolvendo a seguinte sistema de equações de estimação

$$\boldsymbol{U}(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$$

Sob as condições de regularidade de Cramer-Rao-Freeché (C.R.F), pelo Teorema Central do Limite (T.L.C) multivariado, temos que

$$n^{-1/2}\boldsymbol{U}(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)\right) \text{ e } n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_0)\right)$$

em que $\boldsymbol{\theta}_0$ é o valor verdadeiro do parâmetro $\boldsymbol{\theta}$. Considere $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, três testes usualmente aplicados são

1. Teste de razão de verossimilhanças

$$\mathcal{S}_{\mathrm{LR}} = 2\left[\ell(\boldsymbol{x};\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{x};\boldsymbol{\theta}_0)\right]$$

2. Teste de Wald

$$\mathcal{S}_{\mathrm{W}} = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^{\top} \boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

3. Teste escore

$$\mathcal{S}_{\mathrm{R}} = \boldsymbol{U}^{\top}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{U}(\boldsymbol{\theta}_0)$$

Estas três estatísticas, sob $\mathcal{H}_0$ têm distribuições χ^2_p assintoticamente. Observa-se que a estatística de razão de verossimilhanças (RV) requer que a log-verossimilhança seja avaliada em $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}$ e a estatística Wald depende destes valores. Ainda, está e a estatística escore dependem da matriz de informação de Fisher e sua respectiva inversa. Diante disto, Terril (2002) deduz a estatística gradiente,$\mathcal{S}_{\mathrm{G}}$, a partir da estatística de Wald modificada (Hayakawa e Puni, 1985), $\mathcal{S}^*_{\mathrm{W}}$, e $\mathcal{S}_{\mathrm{R}}$.

Essencialmente, a proposta envolve uma matriz simétrica $\mathbf{L}(\boldsymbol{\theta}_0)$, por exemplo, $\mathbf{L}(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta}_0)$, ou seja, $\mathbf{L}^{\top}(\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\theta}_0) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)$. De modo que, $\mathcal{S}_{\mathrm{R}}$ e $\mathcal{S}^*_{\mathrm{W}}$ podem ser reescritas como, respectivamente,

$$\begin{aligned}
\mathcal{S}_{\mathrm{R}} &= \boldsymbol{U}^{\top}(\boldsymbol{\theta}_0) \left[\mathbf{L}^{\top}(\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\theta}_0)\right]^{\top} \boldsymbol{U}(\boldsymbol{\theta}_0) && (1)\\
&= \left[\mathbf{L}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0)\right]^{\top} \left[\mathbf{L}^{-1}(\boldsymbol{\theta}_0)\right]^{\top} \boldsymbol{U}(\boldsymbol{\theta}_0) && (2)\\
&= \mathbf{P}_1^{\top}\mathbf{P}_1 && (3)
\end{aligned}$$

e

$$\begin{aligned}
\mathcal{S}^*_{\mathrm{W}} &= \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^{\top} \mathbf{L}^{\top}(\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) && (4)\\
&= \left[\mathbf{L}(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right]^{\top} \mathbf{L}(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) && (5)\\
&= \mathbf{P}_2^{\top}\mathbf{P}_2 && (6)
\end{aligned}$$

Temos que, sob as condições de regularidade de C.R.F,

$$\mathbf{P}_1 \xrightarrow{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}, \mathbf{I}_p\right) \text{ e } \mathbf{P}_2 \xrightarrow{\mathcal{D}} \mathcal{N}_p\left(\mathbf{0}, \mathbf{I}_p\right)$$

pois $\mathbf{P}_1$ e $\mathbf{P}_2$ são transformações lineares dos vetores aleatórios $\boldsymbol{U}(\boldsymbol{\theta}_0)$ e $\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)$, respectivamente, cujas distribuições assintóticas são normais $p-$variadas. Além disso, note que $\mathbf{P}_1$ e $\mathbf{P}_2$ sã independentes, logo, $\mathbf{P}_1^\top\mathbf{P}_2 \xrightarrow{\mathcal{D}} \chi^2_p$, em que

$$\begin{aligned}
\mathbf{P}_1^\top\mathbf{P}_2 &= \left[\mathbf{L}^{-1}(\boldsymbol{\theta}_0\boldsymbol{U}(\boldsymbol{\theta}_0)\right]^\top \mathbf{L}(\boldsymbol{\theta}_0)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) && (7)\\
&= \boldsymbol{U}^\top(\boldsymbol{\theta}_0)\mathbf{L}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}(\boldsymbol{\theta}_0)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) && (8)\\
&= \boldsymbol{U}^\top(\boldsymbol{\theta}_0)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) && (9)\\
&= \mathcal{S}_{\mathrm{G}} && (10)
\end{aligned}$$

é a estatística gradiente. Contrário as estatísticas escore e de Wald, a estatística gradiente é obtida facilmente, tendo em vista que não requer a obtenção das matrizes de informação de Fisher observada e/ou esperada e de sua inversa, por exemplo. Cabe ressaltar situações complexas, tais como envolvendo dados censurados, não é possível obter uma forma fechada para informação de Fisher, então, a estatística gradiente torna-se uma alternativa razoável frente as estatísticas usuais.

No caso de hipóteses nulas compostas, considerando a partição $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, em que $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ são vetores de parâmetros de dimensões q e $p-q$, respectivamente. O interesse recai em testar o sistema de hipóteses $\mathcal{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)}$ contra $\mathcal{H}_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^{(0)}$. Assim os EMVs não-restritos e restritos de $\boldsymbol{\theta}$ são dados por, respectivamente, $\widehat{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$ e $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1^{(0)}, \tilde{\boldsymbol{\theta}}_2^\top)^\top$.

Portanto, a estatística gradiente assume a forma

$$\mathcal{S}_{\mathrm{G}} = \boldsymbol{U}_1^\top(\tilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^{(0)})$$

pois a partição de $\boldsymbol{\theta}$ induz $\boldsymbol{U}(\boldsymbol{\theta}) = (\boldsymbol{U}_1^\top(\boldsymbol{\theta}), \boldsymbol{U}_2^\top(\boldsymbol{\theta}))$, em que $\boldsymbol{U}_2(\tilde{\boldsymbol{\theta}}) = \mathbf{0}_{p-q}$.

3 Propriedades

Lemonte (2016) discute algumas propriedades da estatística gradiente. Dentre estas propriedades, sob $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, a estatística $\mathcal{S}_{\mathrm{G}}$ tem distribuição χ^2_p central assintótica até um erro de ordem $O(n^{-1})$. Ou seja, as estatísticas gradiente, RV, Wald e escore apresentam apoximações assintóticas de primeira ordem idênticas. Ainda, este resultado indica que pode-se rejeitar $\mathcal{H}_0$ se o valor observado da estatística gradiente excede o quantil superior de ordem $100(1-\gamma)\%$ da distribuição χ^2_p para algum nível de significância nominal.

Baseados em Hayakawa (1975) e Harris e Peers (1980), Lemonte e Ferrari (2012) obtêm a expansão assintótica da distribuição da estatística gradiente para testar hipóteses compostas sob sequências alternativas locais de Pitman, i.e, $\mathcal{H}_{1n} : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)} + n^{-1/2}\boldsymbol{\epsilon}$, em que $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_{p-q})^\top$, convergindo para a hipótese nula a uma taxa de convergêcia de $O(n^{-1/2})$. Ou seja,

$$\mathbb{P}\left(\mathcal{S}_{\mathrm{G}} \leq x\right) = G_{f,\lambda}(x) + \frac{1}{\sqrt{n}}\sum_{k=0}^{3} a_k G_{f+2k,\lambda}(x) + O(n^{-1})$$

em que $G_{m,\lambda}(x)$ é a função de distribuição acumulada de uma qui-quadrado não central com m graus de liberdade e λ parâmetro de não-centralidade. Detalhes sobre os coeficientes $a_k \; \forall \; k = 1, 2, 3$ podem ser consultados em Lemonte e Ferrari (2012).

Outro aspecto a ser considerado envolve que, ao contrário das estatísticas RV, escore e Wald, a estatística gradiente não é invariante sob reparametrizações não lineares de $\boldsymbol{\theta}$ ou, equivalentemente, dependem de como as hipóteses são formuladas. Ademais, Terrell (2002) aponta que a estatística gradiente nem sempre é não-negativa, embora seja quase-certamente. Especialmente, pode-se garantir que $\mathcal{S}_{\mathrm{G}}$ é não-negativa desde que a função de log-verossimilhança seja unimodal e diferenciável em algum $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ (Terrell, 2002, pg. 208).

Cabe ressaltar que, em geral, pode-se melhorar a aproximação da distribuição da estatística gradiente a uma distribuição χ^2 sob a hipótese nula a partir da substituição do EMV por um estimador menos viesado, $\tilde{\boldsymbol{\theta}}$, de $\boldsymbol{\theta}$.

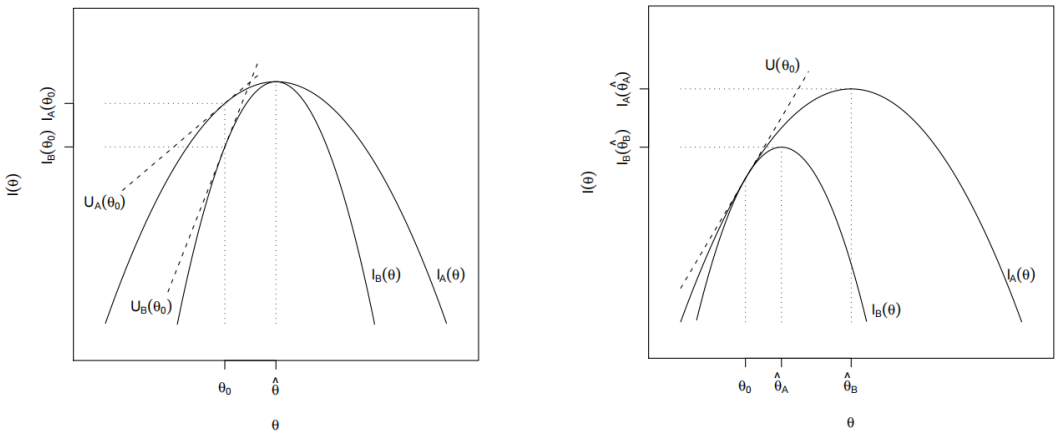
4 Interpretação geométrica

Muggeo e Lovison (2014) argumentam que a bibliografia básica sobre testes assintóticos carece de uma exposição unificada de sua interpretação geométrica. Especificamente, as estatísticas de RV, Wald e escore são medidas em escalas distintas, a mencionar, da log-verossimilhança, dos parâmetros e da função escore. Nesta abordagem, Buse (1982) discute o obtenção das estatísticas de RV, Wald e escore sob a perspectiva geométrica.

Assim, faz-se necessária uma breve revisão da interpretação geométrica das estatísticas supracitadas de modo a facilitar a elucidação da estatística gradiente neste contexto. Por simplicidade, assume-se que o parâmetro de interesse seja um escalar θ e deseja-se testar a $\mathcal{H}_0 : \theta = \theta_0$ contra $\mathcal{H}_1 : \theta \neq \theta_0$, em que $\theta \in \Theta$ e θ_0 é o valor verdadeiro do parâmetro. Ainda, suponha a existência e unicidade do EMV, $\hat{\theta}$, e denota-se $\mathcal{C}(\hat{\theta})$ a curvatura da log-verossimilhança avaliada em $\hat{\theta}$.

Pela figura, observa-se que, fixado $\mathcal{C}(\hat{\theta})$, $\mathcal{S}_{\text{RV}}/2$ aumenta conforme a distância entre o EMV restrito e o não-restrito aumenta, $\hat{\theta}-\theta_0$. Por outro lado, fixado $\hat{\theta}-\theta_0$, quanto mais acentuada a curvatura, maior a distância mencionada. Todavia, cabe ressaltar que para conjuntos de dados A e B com mesma distância entre $\hat{\theta}$ e θ e distintas curvaturas apresentam valores da estatística de RV diferentes.

Figura 1: Comparação gráfica das estatísticas de razão de verossimilhanças generalizada, Wald, escore e gradiente em dois conjuntos de dados distintos A e B



Fonte: Montoril e Souza (2013)

Dessa maneira, parece razoável ponderar $(\hat{\theta}-\theta_0)^2$ pela curvatura $\mathcal{C}(\hat{\theta})$, ou seja, obtemos $\mathcal{S}_{\text{w}} = (\hat{\theta}-\theta_0)^2\mathcal{C}(\hat{\theta})$ a estatística de Wald. Além disso, no caso da estatística escore, se a hipótese nula é verdadeira, o EMV restrito será próximo do não-restrito. Assim, dado que o EMV não-restrito satisfaz $U(\hat{\theta})=0$, enão, a função escore pode ser uma medida apropriada para mensurar a distância entre θ_0 e $\hat{\theta}$. Entretanto, similarmente, a estatística $U(\theta_0)^2$ padece da problemática observada na estatística de Wald. Pois, nota-se pela figura que dois conjuntos de dados A e B podem apresentar o mesmo valor $U(\theta_0)$, mas um destes consta θ_0 mais próximo de $\hat{\theta}$. De modo que quanto maior a curvatura da log-verossimilhança, menor $\hat{\theta}-\theta_0$. Logo, sugere-se ponderar $U(\theta_0)^2$ pelo recíproco da curvatura, isto é, $\mathcal{S}_{\text{R}} = U(\theta_0)^2\mathcal{C}^{-1}(\theta_0)$.

Quanto a estatística gradiente, Montoril e Souza (2013) reanalisam o caso da obtenção da estatística de Wald ao manter apenas $\hat{\theta}-\theta_0$ ao contrário de seu quadrado. Os autores argumentam que o problema envolvendo o sinal de $\hat{\theta}-\theta_0$ pode ser contornado ao ponderá-lo

pela função escore. Esta modificação justifica-se em virtude de ambos possuirem o mesmo sinal e de quanto maior a curvatura $\mathcal{C}(\theta_0)$, maior $U(\theta_0)$, em valor absoluto. O procedimento supracitado resulta na estatística gradiente, $\mathcal{S}_{\mathrm{G}} = U(\theta_0)(\hat{\theta} - \theta_0)$.

Diante disto, Muggeo e Lovison (2014) comparam geometricamente as quatro estatísticas sob as escalas da log-verossimilhança e da função escore. Considerando a escala da log-verossimilhança e sob as condições de regularidade apropriadas, os autores utilizam-se aproximações quadráticas de $\ell(\theta)$ baseadas na expansão de Taylor em $\theta = \hat{\theta}$ para estatística Wald, em $\theta = \theta_0$ para a estatística escore e de primeira ordem em θ_0 para a estatística gradiente, isto é

$$\mathcal{P}_{\mathrm{W}}(\theta) \approx \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})\ell''(\hat{\theta})$$

$$\mathcal{P}_{\mathrm{R}}(\theta) \approx \ell(\theta_0) + (\theta - \theta_0)\ell'(\theta_0) + \frac{1}{2}(\theta - \theta_0)\ell''(\theta_0)$$

$$\mathcal{P}_{\mathrm{G}}(\theta) \approx \ell(\theta_0) + (\theta - \theta_0)\ell'(\theta_0)$$

O fato de as estatísticas Wald e escore serem aproximadas por polinômios de segunda ordem refletem a ponderação de $(\hat{\theta} - \theta_0)^2$ e de $U(\theta_0)^2$ como discutida em Buse (1982). Então, após algumas manipulações algébricas, obtêm-se

$$\mathcal{S}_{\mathrm{W}} \approx 2\left[\mathcal{P}_{\mathrm{W}}(\hat{\theta}) - \mathcal{P}_{\mathrm{W}}(\theta_0)\right]$$

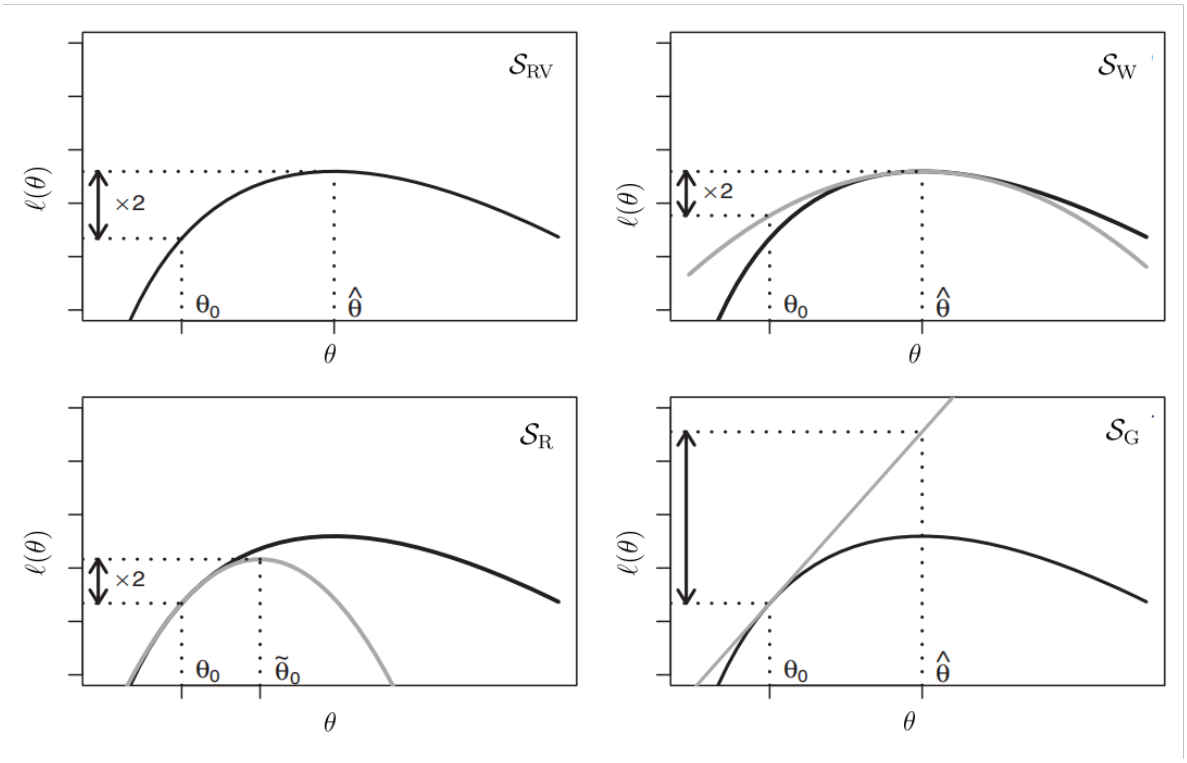
$$\mathcal{S}_{\mathrm{R}} \approx 2\left[\mathcal{P}_{\mathrm{R}}(\tilde{\theta}_0) - \mathcal{P}_{\mathrm{R}}(\theta_0)\right]$$

$$\mathcal{S}_{\mathrm{G}} \approx \mathcal{P}_{\mathrm{G}}(\hat{\theta}) - \mathcal{P}_{\mathrm{G}}(\theta_0)$$

em que $\tilde{\theta}_0 = \theta_0 - \ell'(\theta_0)/\ell''(\theta_0)$.

Comparando graficamente as quatro estatísticas de teste na escala da log-verossimilhança, percebe-se na figura a curva da log-verossimilhança e da aproximação das respectivas estatísticas em linhas sólidas preta e cinza, respectivamente. Nota-se, pelas setas os valores observados para cada estatística de teste. Quanto maior este valor, maior a evidência contra a hipótese nula. Assim, evidencia-se a diferença em magnitude destes valores, sobretudo para estatística gradiente. Dado que $\ell'(\theta) = U(\theta)$, as quatro estatísticas também podem ser expressas na

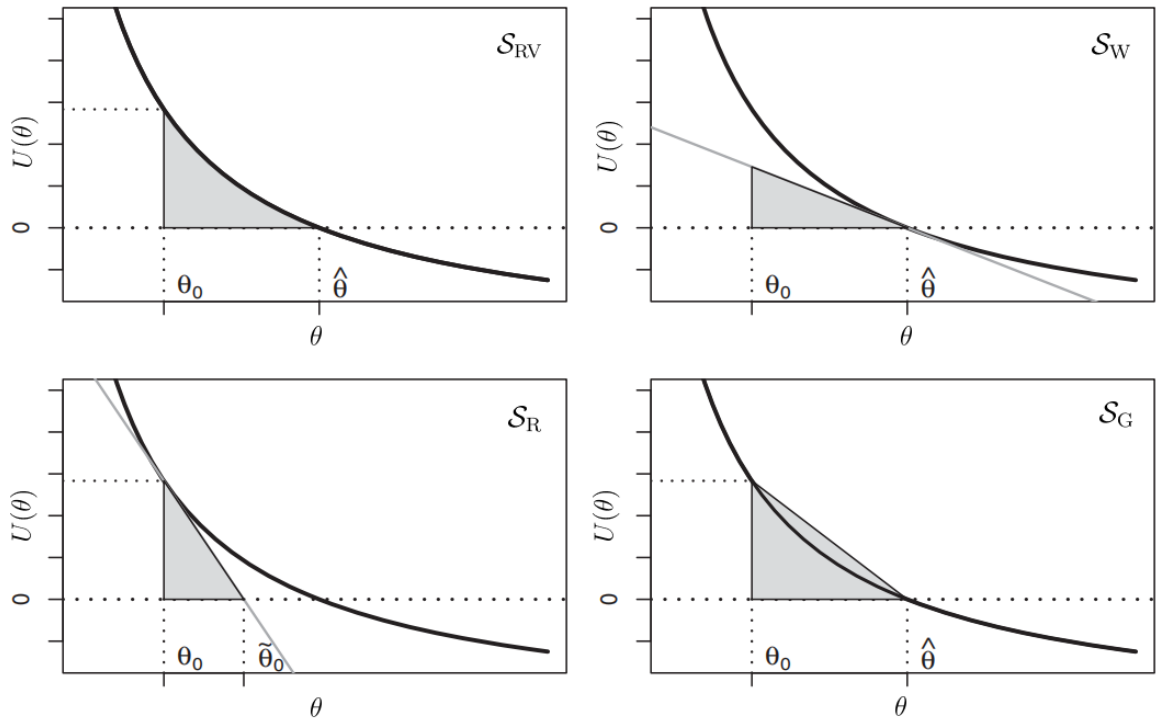
Figura 2: Comparação gráfica das estatísticas de razão de verossimilhanças generalizada, Wald, Escore e Gradiente na escala da log-verossimilhança



Fonte: Adaptado de Muggeo e Lovison (2014)

escala da função escore. Ou seja, os valores observados das estatísticas mencionadas são dadas por áreas sobre a função escore, conforme a figura.

Figura 3: Comparação gráfica das estatísticas de razão de verossimilhanças generalizada, Wald, Escore e Gradiente na escala da função escore



Fonte: Adaptado de Muggeo e Lovison (2014)

5 Refinamentos

Embora os testes assintóticos usuais e gradiente sejam equivalentes assintoticamente, para amostras de tamanho pequeno ou moderado podem não ser confiáveis, dado que a probabilidade do erro do tipo 1 pode não estar próxima do tamanho nominal estabelecido. Neste contexto, usualmente, correções Bartlett e tipo Bartlett atenuam as distorções de tamanho dos testes. Aplicações destas correções para a estatística gradiente podem ser consultadas em Vargas et al. (2014) para modelos lineares generalizados, Medeiros e Ferrari (2017) para modelos de dispersão, Magalhães e Gallardo (2020) e Medeiros e Lemonte (2021) para modelos de regressão exponencial e weibull em dados censurados e heteroscedásticos generalizados (Barros

et al. 2023).

Inicialmente, a ideia de aprimorar a aproximação da distribuição da estatística de testes assintóticos pela distribuição χ^2 originou-se em Barlett (1937). A proposta envolve multiplicar a estatística de razão de verossimilhanças por uma constante apropriada c^{-1}. Posteriormente, Lawley (1956) exibiu um método geral para obtenção deste fator de correção baseado no primeiro momento de $\mathcal{S}_{\mathrm{RV}}$.

Para o aprimento da estatísica escore, Cordeiro e Ferrari (1991) introduziram um fator de correção tipo Bartlett para melhorar sua aproximação para uma χ^2. Baseado no referido trabalho, Vargas et al. (2013) obtêm um fator de correção para a estatística gradiente. Assim, sob $\mathcal{H}_0$, a estatística gradiente distribui-se segundo uma χ^2 até um erro de ordem $O(n^{-3/2})$, enquanto a não corrigida tem distribuição χ^2 até um erro de ordem $O(n^{-1})$. Ou seja, a correção do tipo Bartlett reduz o erro de aproximação de $O(n^{-1})$ para $O(n^{-3/2})$. Em outras palavras, para amostras de tamanho pequeno e/ou moderado, espera-se que a estatística gradiente corrigida apresente resultados superiores em relação a usual.

A estatística gradiente corrigida é definida como:

$$\mathcal{S}_{\mathrm{G}}^{*} = \mathcal{S}_{\mathrm{G}}\left[1 - (c + b\mathcal{S}_{\mathrm{G}} + a\mathcal{S}_{\mathrm{G}}^{2})\right]$$

As constantes a, b e c são obtidas através da expansão assintótica da distribuição acumulada sob $\mathcal{H}_0$ para testar hipóteses compostas, mais detalhes em Vargas et al. (2013).

Em suma, uma profunda revisão bibliográfica sobre correções Bartlett e tipo Bartlett foi realizada por Cordeiro e Cribari-Neto (1996) para estatísticas de razão de verossimilhanças, Wald e escore.

6 Aplicação

Objetiva-se reproduzir os resultados de Lemonte (2016, pg. 52-55) para comparação da performance dos testes de RV, Wald, escore e gradiente em amostras de tamanho pequeno e moderado sob o modelo de regressão linear simétrico homoscedástico.

Considere o modelo simétrico linear homoscedástico dado por:

$$y_i = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon_i \ \forall \ i = 1, \ldots, n$$

em que $\varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d}{\sim} \mathcal{S}(0, 1)$, então, a densidade de Y_i é da forma

$$f(y_i) = \frac{1}{\sqrt{\phi}} g(u_i) \mathbb{I}_{\mathbb{R}}(y_i) \ \forall i = 1, \ldots, n$$

Ou seja, $Y_1, \ldots, Y_n \overset{ind}{\sim} \mathcal{S}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \phi)$ com $u_i = (y_i - \boldsymbol{X}\boldsymbol{\beta})^2/\phi$. Denomina-se $g : \mathbb{R} \to [0, \infty]$ função geradora de densidade.

Assim, a log-verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$ é expressa como

$$\ell(\boldsymbol{y}; \boldsymbol{\theta}) = -\frac{n}{2} \ln \phi + \sum_{i=1}^{n} \ln g(u_i)$$

A função escore de $\boldsymbol{\beta}$ e ϕ são definidas, respectivamente, por

$$\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \frac{1}{\phi} \boldsymbol{X}^\top \mathbf{D}(\boldsymbol{v})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \text{ e } U_\phi(\boldsymbol{\theta}) = \frac{1}{2\phi} \left(\frac{1}{\phi} Q - n \right)$$

em que $\mathbf{D}(\boldsymbol{v}) = \text{diag}\{v_1, \ldots, v_n\}$, $v_i = -2W_{g_i}$, $W_{g_i} = g'(u_i)/g(u_i)$ e $Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \mathbf{D}(\boldsymbol{v})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Obtêm-se o EMV de $\boldsymbol{\theta}$ a partir da resolução iterativa do sistema de equações formado por $\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{0}$ e $U_\phi(\boldsymbol{\theta}) = 0$. A solução simultânea é dada por:

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}^\top \mathbf{D}(\boldsymbol{v}) \boldsymbol{X} \right]^{-1} \boldsymbol{X}^\top \mathbf{D}(\boldsymbol{v}) \boldsymbol{y} \text{ e } \hat{\phi} = \frac{1}{n} \hat{Q}$$

Pode-se mostrar que $\boldsymbol{\beta}$ e ϕ são parâmetros globalmente ortogonais, logo, a matriz de informação de Fisher é bloco diagonal, ou seja, $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}) = \text{bloco diag}\left\{ \boldsymbol{\mathcal{J}}(\boldsymbol{\beta}), \mathcal{J}(\phi) \right\}$, em que

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \frac{4d_g}{\phi} \boldsymbol{X}^\top \boldsymbol{X} \text{ e } \mathcal{J}(\phi) = \frac{n}{4\phi^2}(4f_g - 1)$$

em que $d_g = \mathbb{E}_U \left[W_g^2(U^2) U^2 \right]$ e $f_g = \mathbb{E}_U \left[W_g^2(U^2) U^4 \right]$, $U \sim \mathcal{S}(0, 1)$. Cysneiros e Paula (2005) fornecem expressões para $g(u), W_g(u)$ e $W_g'(u)$ e valores para d_g, f_g e ξ para algumas distribuições simétricas. Ainda, os autores avaliam o comportamento dos pesos v contra u para diferentes valores dos parâmetros ν e k das distribuições t de Student e exponencial potência, respectivamente.

Considerando hipóteses compostas da forma $\mathcal{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0$ contra $\mathcal{H}_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^0$, as estatísticas de razão de verossimilhanças, Wald, escore e gradiente para o modelo de regressão simétrico homoscedástico baseadas em Cysneiros e Paula (2005) e Lemonte (2012) são, respectivamente

$$\mathcal{S}_{\text{RV}} = 2 \left[\ell(\boldsymbol{y}; \hat{\boldsymbol{\beta}}_1, \hat{\phi}) - \ell(\boldsymbol{y}; \boldsymbol{\beta}_1^{(0)}, \tilde{\phi}) \right]$$

$$\mathcal{S}_{\mathrm{W}} = \frac{4d_g}{\hat{\phi}} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right)^{\top} \left[\mathbf{C}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\mathbf{C}^{\top}\right]^{-1} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right)$$

$$\mathcal{S}_{\mathrm{R}} = \frac{\tilde{\phi}}{4d_g} \boldsymbol{U}_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}_1^{(0)}, \tilde{\phi}\right)^{\top} (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1} \boldsymbol{U}_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}_1^{(0)}, \tilde{\phi}\right)$$

e

$$\mathcal{S}_{\mathrm{G}} = \frac{1}{\tilde{\phi}} (\boldsymbol{y} - \boldsymbol{X}_2\tilde{\boldsymbol{\beta}}_2)^{\top} \mathbf{D}(\boldsymbol{v}) \boldsymbol{X}_1 \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}\right)$$

em que $\mathbf{C} = \left(\mathbf{I}_q \ \mathbf{0}_{(q\times p-q)}\right)$.

O esquema de simulação consiste em testar a hipótese $\mathcal{H}_0 : \beta_1 = \cdots = \beta_q = 0$ contra $\mathcal{H}_1 : \exists_j \beta_j \neq 0 \ \forall \ j = 1, \ldots, p$, em que $q \leq p$. Considerou-se o modelo de regressão linear simétrico homoscedástico tal que $x_{ij} \sim \mathcal{U}(0,1) \ \forall \ j$, $\beta_{q+1} = \cdots = \beta_p = 1$ e $\varepsilon \sim \mathcal{S}(0,\phi)$ sob distribuições $\mathcal{N}(0,4)$ e t_4. Ainda, determinou-se os tamanhos amostrais $n = 15, 25, 40, 80, 120$, níveis de significância nominais de $\alpha = 0,05$ e $0,01$, $p = 3, \ldots, 6$ fixado $q = 2$ e $q = 1, \ldots, 6$ fixado $p = 7$. Utilizou-se 5000 réplicas de Monte Carlo.

Para a comparação do poder empírico dos quatro testes assintótico, avaliou-se a hipótese alternativa $\beta_1 = \beta_2 = \delta, \delta \in (-5, 5)$, com $n = 30$, $p = 3$ e $\alpha = 0,01$.

Figura 4: Taxa de rejeição empírica do teste $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$ com $p = 3, \ldots, 6$ e nível de significância $\alpha = 0,10$ no modelo de regressão linear simétrico sob distribuição $\mathcal{N}(0, 4)$ e t_4, respectivamente

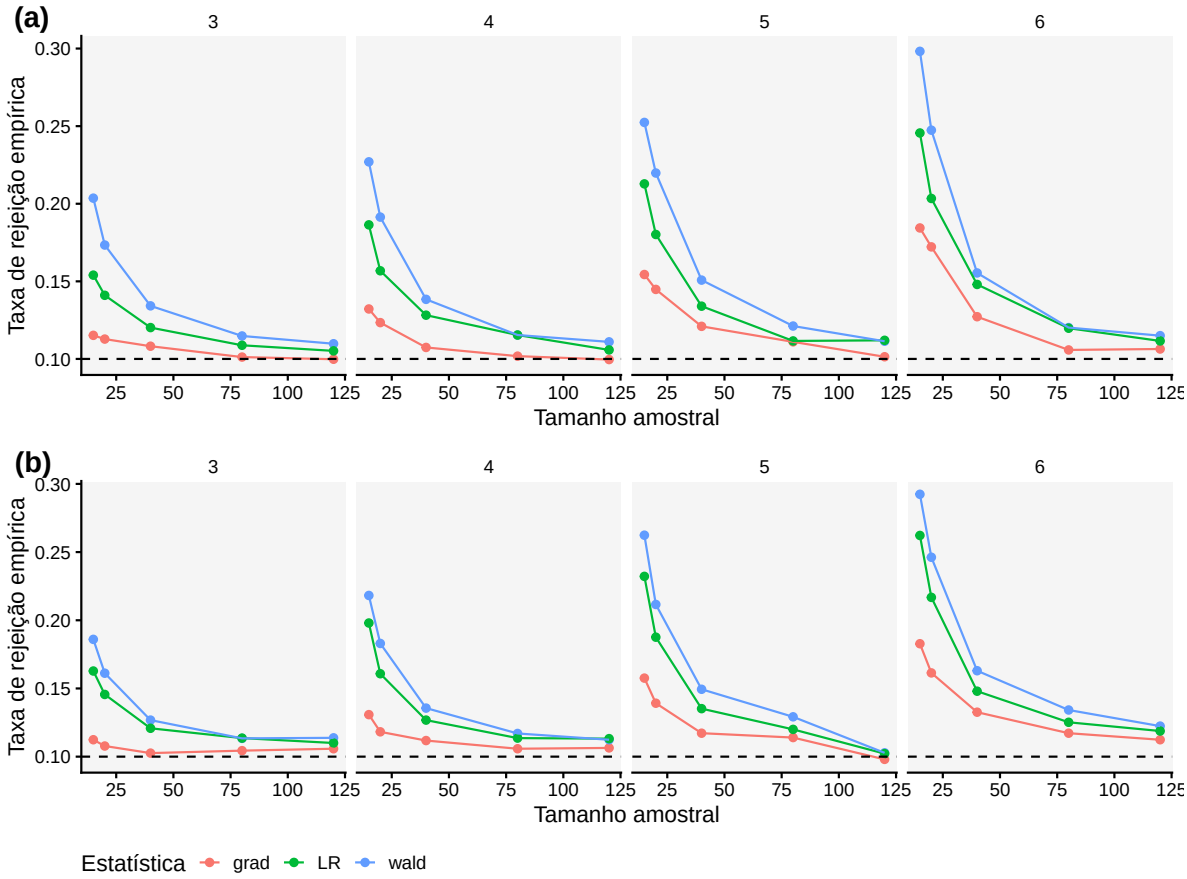


Figura 5: Taxa de rejeição empírica do teste $\mathcal{H}_0 : \beta_1 = \beta_2 = 0$ com $p = 3, \ldots, 6$ e nível de significância $\alpha = 0,05$ no modelo de regressão linear simétrico sob distribuição $\mathcal{N}(0, 4)$e t_4, respectivamente

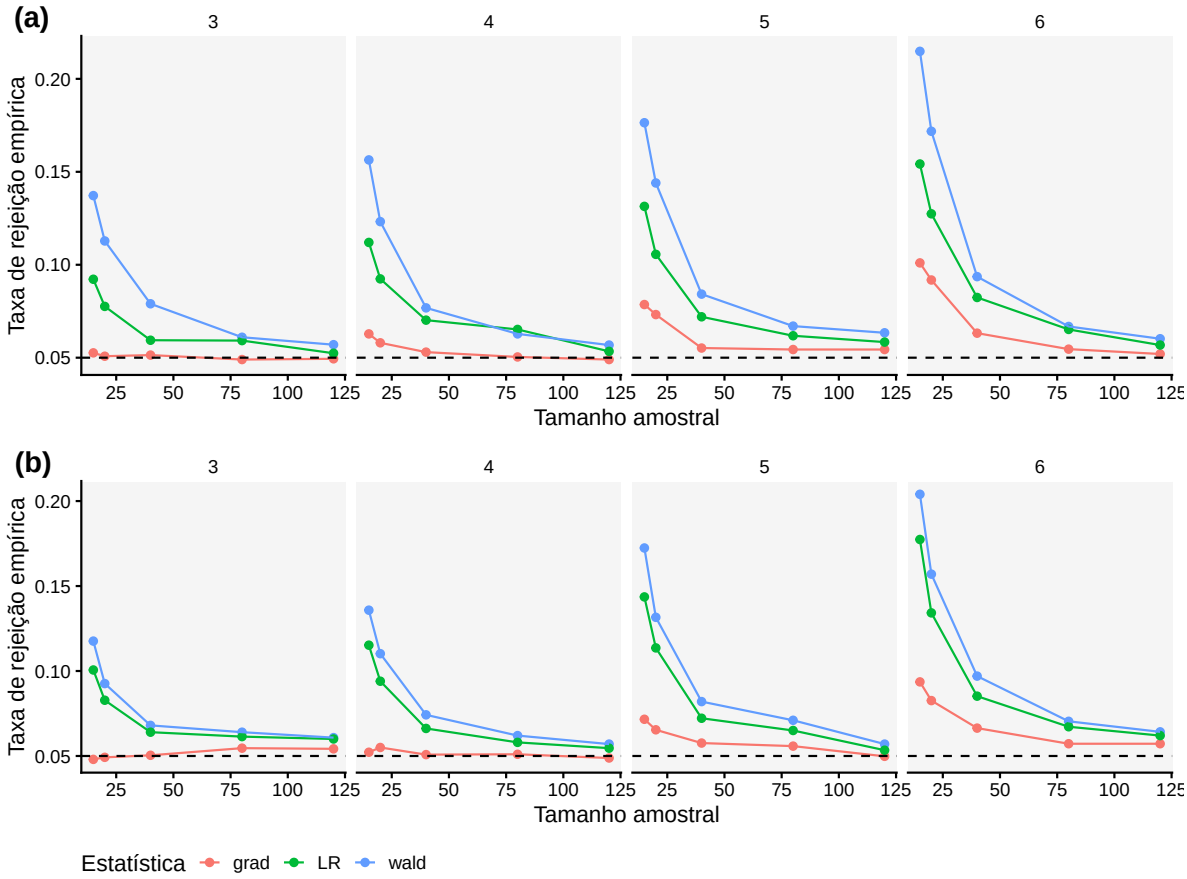


Figura 6: Taxa de rejeição empírica do teste $\mathcal{H}_0 : \beta_1 = \cdots = \beta_q = 0$ com $q = 1, \ldots, 6$ fixado $p = 2$ e nível de significância $\alpha = 0,05$ no modelo de regressão linear simétrico sob distribuição $\mathcal{N}(0, 4)$ e t_4, respectivamente

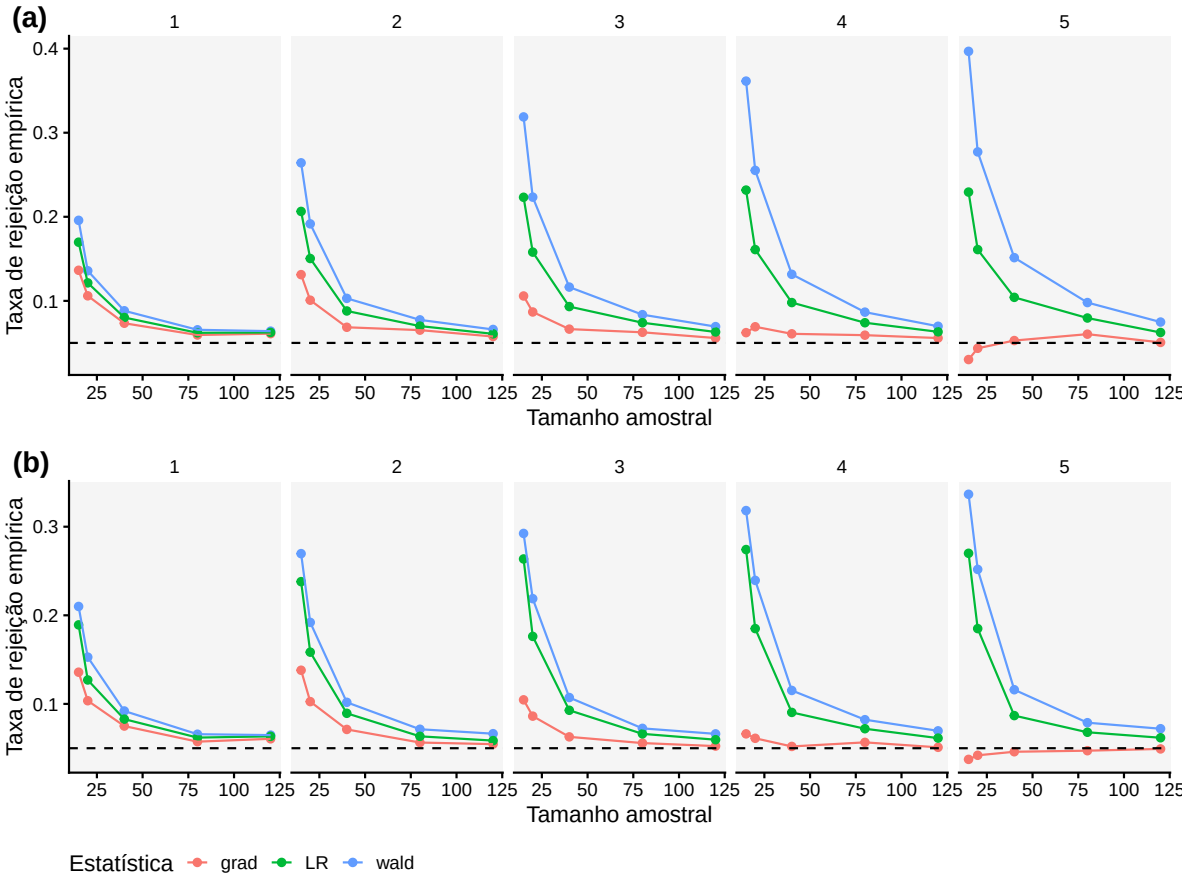
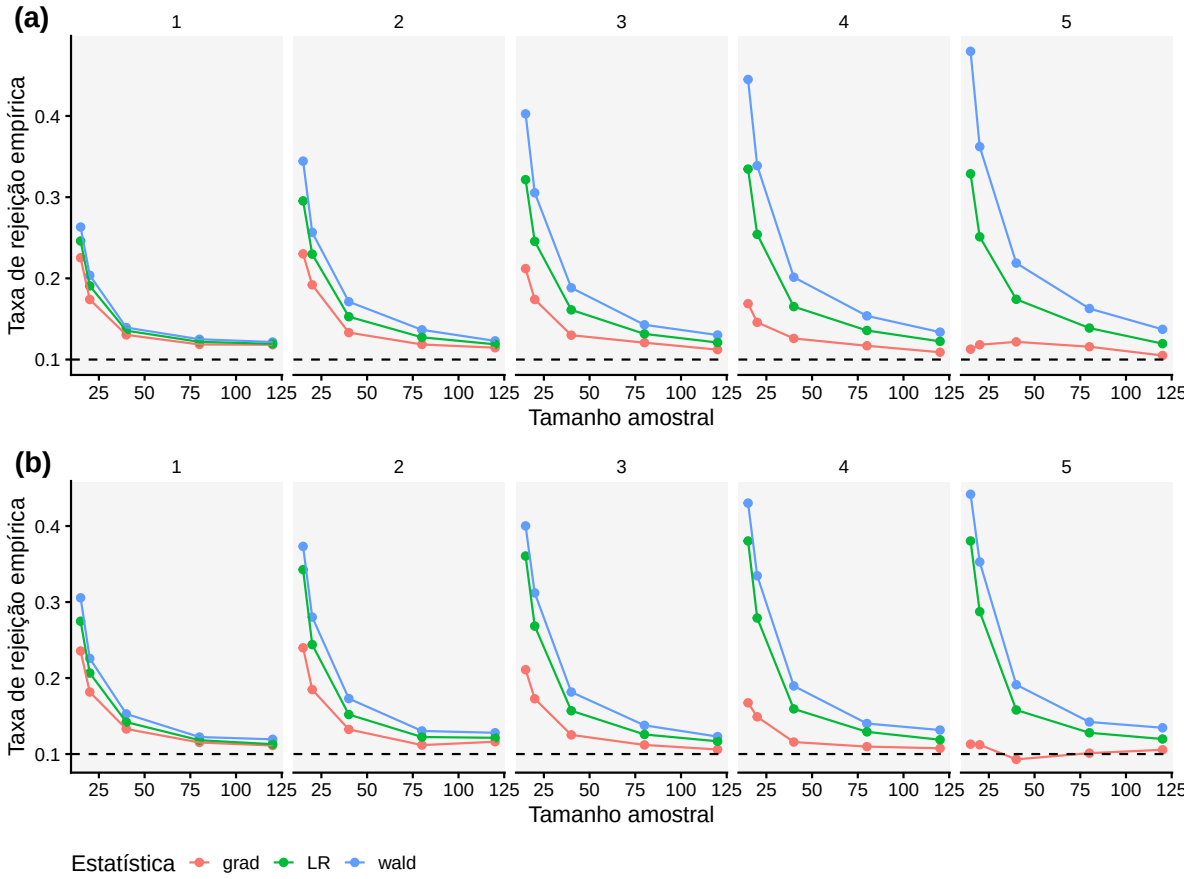


Figura 7: Taxa de rejeição empírica do teste $\mathcal{H}_0 : \beta_1 = \cdots = \beta_q = 0$ com $q = 1, \ldots, 6$ fixado $p = 2$ e nível de significância $\alpha = 0,10$ no modelo de regressão linear simétrico sob distribuição $\mathcal{N}(0, 4)$ e t_4, respectivamente



7 Considerações finais

No presente trabalho foi realizada uma revisão geral sobre a estatística gradiente em testes de hipóteses, com um foco particular no modelo de regressão linear simétrico homoscedástico. Neste âmbito, inicialmente, foi apresentado conceitos básicos em testes assintóticos. Quanto a estatística gradiente em si deu-se maior relevância a suas propriedades, interpretação geométrica e refinamentos com base nos trabalhos de Lemonte (2016), Buse (1982), Muggeo e Lovison (2014) e entre outros.

Subsequentemente, visando comparar a taxa de rejeição e o poder empírico do teste gradiente em relação aos testes de razão de verossimilhança, Wald e escore, empregou-se simulações

de Monte Carlo envolvendo hipóteses compostas sobre os parâmetros de posição do modelo de regressão linear simétrico homoscedástico sob diferentes tamanhos amostrais, números de parâmetros restritos e irrestritos. Assim, os resultados mostram a superioridade da estatística gradiente tanto sob distribuição normal como t de Student para a fonte de variação.

Considerando a performance da estatística gradiente em diversos estudos de simulação, é inegável a iminência da sua consideração na análise estatística dos dados, especialmente em cenários nos quais o tamanho da amostra seja pequeno ou moderado e os recursos computacionais sejam escassos. Ademais, estudos posteriores podem revisar as extensões da estatística gradiente, tais como no caso de má especificação do modelo como fez Lemonte (2013). Neste contexto, também, pode ser abordada a busca progressiva de Atkinson (1994) no caso simétrico para avaliar a robustez da estatística diante da presença de valores discrepantes, alavancas e/ou influentes.

8 Referências

AITCHISON, J.; SILVEY, S. D. Maximum-Likelihood Estimation of Parameters Subject to Restraints. **Annals of Mathematical Statistics**, [*S. l.*], v. 29, p. 813–828, set. 1958.

ATKINSON, A. C. Fast Very Robust Methods for the Detection of Multiple Outliers. **Journal of the American Statistical Association**, [*S. l.*], v. 89, p. 1329–1339, dez. 1994.

BARROS, F. U. *et al.* Improved gradient statistic in heteroskedastic generalized linear models. **Journal of Statistical Computation and Simulation**, [*S. l.*], v. 93, p. 2052–2066, fev. 2023.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences**, [*S. l.*], v. 160, p. 268–282, maio 1937.

BUSE, A. The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. **The American Statistician**, [*S. l.*], v. 36, p. 153, ago. 1982.

CORDEIRO, G. M.; FERRARI, S. L. A modified score test statistic having chi-squared distribution to order n−1. **Biometrika**, [*S. l.*], v. 78, p. 573–582, 1991.

CRIBARI-NETO, F.; CORDEIRO, G. M. On bartlett and bartlett-type corrections. **Econometric Reviews**, [*S. l.*], v. 15, p. 339–367, jan. 1996.

CYSNEIROS, F. J. A.; PAULA, G. A. Restricted methods in symmetrical linear regression models. **Computational Statistics & Data Analysis**, [*S. l.*], v. 49, p. 689–708, jun. 2005.

HARRIS, P.; PEERS, H. W. The local power of the efficient scores test statistic. **Biometrika**, [*S. l.*], v. 67, p. 525–529, 1980.

HAYAKAWA, T. The likelihood ratio criterion for a composite hypothesis under a local alternative. **Biometrika**, [*S. l.*], v. 62, p. 451–460, ago. 1975.

LAWLEY, D. N. A GENERAL METHOD FOR APPROXIMATING TO THE DISTRIBUTION OF LIKELIHOOD RATIO CRITERIA. **Biometrika**, [*S. l.*], v. 43, p. 295–303, 1956.

LEHMANN, E. L. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? **Journal of the American Statistical Association**, [*S. l.*], v. 88, p. 1242–1249, 1993. Disponível em: https://www.jstor.org/stable/2291263?seq=2#metadata_info_tab_contents.

LEMONTE, A. **The Gradient Test**. [*S. l.*]: Academic Press, 2016.

LEMONTE, Artur J. Local power of some tests in exponential family nonlinear models. **Journal of Statistical Planning and Inference**, [*S. l.*], v. 141, p. 1981–1989, maio 2011.

LEMONTE, Artur J. Local power properties of some asymptotic tests in symmetric linear

regression models. **Journal of Statistical Planning and Inference**, [*S. l.*], v. 142, p. 1178–1188, maio 2012.

LEMONTE, Artur J. On the gradient statistic under model misspecification. **Statistics & Probability Letters**, [*S. l.*], v. 83, p. 390–398, jan. 2013.

LEMONTE, Artur J.; FERRARI, S. L. P. Testing hypotheses in the Birnbaum–Saunders distribution under type-II censored samples. **Computational Statistics & Data Analysis**, [*S. l.*], v. 55, p. 2388–2399, jul. 2011.

LEMONTE, Artur J.; Silvia. The local power of the gradient test. **Annals of the Institute of Statistical Mathematics**, [*S. l.*], v. 64, p. 373–381, out. 2010.

MAGALHÃES, F. M. C.; FERRARI, S.; LEMONTE, A. J. Improved inference in dispersion models. **Applied Mathematical Modelling**, [*S. l.*], v. 51, p. 317–328, nov. 2017.

MAGALHÃES, T. M.; GALLARDO, D. I. Bartlett and Bartlett-type corrections for censored data from a Weibull distribution. **Sort-statistics and Operations Research Transactions**, [*S. l.*], v. 44, p. 0127–140, jan. 2020.

MEDEIROS, F. M. C.; LEMONTE, A. J. Likelihood-based inference in censored exponential regression models. **Communications in Statistics - Theory and Methods**, [*S. l.*], v. 50, p. 3214–3233, dez. 2019.

MONTORIL, M. H.; SOUZA, E. A. Estatística gradiente: propriedades e aplicações. **Revista Brasileira de Biometria**, [*S. l.*], v. 31, p. 43–60, 2013.

MUGGEO, V. M. R.; LOVISON, G. The "ThreePlusOne" Likelihood-Based Test Statistics: Unified Geometrical and Graphical Interpretations. **The American Statistician**, [*S. l.*], v. 68, p. 302–306, out. 2014.

NEYMAN, J.; PEARSON, E. S. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. **Biometrika**, [*S. l.*], v. 20A, p. 175, jul. 1928.

RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. **Mathematical Proceedings of the Cambridge Philosophical Society**, [*S. l.*], v. 44, p. 50–57, jan. 1948.

SILVEY, S. D. The Lagrangian Multiplier Test. **The Annals of Mathematical Statistics**, [*S. l.*], v. 30, p. 389–407, jun. 1959.

TERRELL, G. The Gradient Statistic. **Computational Science and Statistics**, [*S. l.*], p. 206–215, 2002.

THE ROYAL MINT. **Trial of the Pyx - Ensuring Coin Accuracy & Quality | The Royal Mint**. [*S. l.*]: www.royalmint.com, [*s. d.*]. Disponível em: https://www.royalmint.com/discover/uk-coins/history-of-the-trial-of-the-pyx.

VARGAS, T. M.; Silvia; LEMONTE, A. J. Gradient statistic: Higher-order asymptotics and Bartlett-type correction. **Electronic Journal of Statistics**, [*S. l.*], v. 7, jan. 2013.

VARGAS, T. M.; Silvia; LEMONTE, A. J. Improved likelihood inference in generalized linear models. **Computational Statistics & Data Analysis**, [*S. l.*], v. 74, p. 110–124, jun. 2014.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. **Transactions of the American Mathematical Society**, [*S. l.*], v. 54, p. 426–482, 1943.

WILKS, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. **The Annals of Mathematical Statistics**, [*S. l.*], v. 9, p. 60–62, mar. 1938.

9 Apêndice A — Rotinas computacionais em R

```
library(gwer)

simul_null_rej_rate <- function(stat_type, alpha,
                                p, q, n, n_rep,
                                delta, seed, family) {
  set.seed(seed)

  trace <- c()
  b = matrix(c(rep(delta, q), rep(1, p-q)), nrow=p, ncol=1)
  b1 = b[1:q, ]
  b2 = b[(q+1):p, ]
  b1_null = matrix(0, nrow=q)

  for (j in 1:n_rep) {
    e = switch(family["family"][[1]],
               "Normal" = rnorm(n, sd=2),
               "Student" = rt(n, df=4))
    x = cbind(1, matrix(runif((p-1)*n), nrow=n, ncol=(p-1)))
    y = x %*% b + e
    x1 = as.matrix(x[, 1:q])
    x2 = as.matrix(x[, (q+1):p])

    mod_unrestr <- elliptical(y ~ x - 1, family=family)
    mod_restr <- elliptical(y ~ x2 - 1, family=family)
    b_hat <- mod_unrestr$coefficients
    b1_hat <- b_hat[1:q]
    b2_tilde <- mod_restr$coefficients
    Dv <- diag(mod_restr$v)
    dg <- family$g2(df=family$df)
    C <- cbind(diag(1, q), matrix(0, q, p-q))
    R <- C %*% solve(t(x) %*% x) %*% t(C)
```

```
    l_restr <- mod_restr$loglik
    l_unrestr <- mod_unrestr$loglik

    phi_hat <- (1/n) * t(y - x %*% b_hat) %*% Dv %*% (y - x %*% b_hat)
    phi_tilde <- (1/n)*t(y-x2 %*% b2_tilde) %*% Dv %*% (y-x2 %*% b2_tilde)

    if (stat_type == "score") {
      U_restr <- (1/phi_tilde) * t(x2) %*% Dv %*% (y - x2 %*% b2_tilde)
    }

    stat <- switch(stat_type,
                   "grad" = (1/phi_tilde) * t(y - x2 %*% b2_tilde) %*% Dv %*% x1 %*% (b1_h
                   "wald" = ((4*dg)/phi_hat)*t(b1_hat - b1_null) %*% solve(R) %*% (b1_hat
                   "score" = (phi_tilde/(4*dg)) * t(U_restr) %*% solve(t(x2) %*% x2) %*% U
                   "LR" = 2 * (l_unrestr - l_restr))
    chi_q <- qchisq(1-alpha, df=q)
    rej_h0 <- stat >= chi_q
    trace[j] <- rej_h0
  }

  null_rej_rate <- sum(trace)/n_rep

  return(null_rej_rate)
}
```