
INFERÊNCIA EM REGRESSÕES LINEARES HETEROSCEDÁSTICAS:
UMA AVALIAÇÃO NUMÉRICA

WILTON BERNARDINO DA SILVA

Orientador: Prof. Dr. Francisco Cribari Neto

Área de Concentração: Estatística Matemática

Dissertação submetida como requerimento parcial para obtenção do grau
de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, 27 de fevereiro de 2009

Universidade Federal de Pernambuco

Pós-Graduação em Estatística

27 de fevereiro de 2009
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Wilton Bernardino da Silva

intitulada

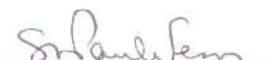
"Inferência em Modelos Heteroscedásticos: Uma Avaliação Numérica"

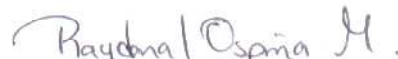
seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.


Coordenador da Pós-Graduação em Estatística

Banca Examinadora:


Francisco Cribari Neto orientador


Silvia Lopes de Paula Ferrari (USP)


Raydonal Ospina

Este documento será anexado à versão final da dissertação.

Silva, Wilton Bernardino da
Inferência em regressões lineares
heteroscedásticas: uma avaliação numérica / Wilton
Bernardino da Silva - Recife : O Autor, 2009.
vi, 74 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal
de Pernambuco. CCEN. Estatística, 2009.

Inclui bibliografia, anexo e apêndice.

1. Análise de regressão. I. Título.

519.536 CDD (22.ed.) MEI2009-032

*Aos meus queridos pais
e aos eternos amigos*

Agradecimentos

A Deus, por me conceder grandes conquistas.

Aos meus pais, José Bernardino da Silva Filho e Maria José Bernardino da Silva, que sempre incentivaram minha formação.

A todas as pessoas próximas que me aconselharam a sempre seguir em busca de objetivos.

Ao meus grandes e eternos amigos de graduação Rodrigo, Marco e Renato, que sempre estiveram próximos me dando bons conselhos.

Aos amigos Cícero, Andreia, Wallace, Isabel e Olga, por construirmos uma amizade verdadeira ao longo do mestrado.

Aos colegas de mestrado Hemílio, Marcelo e Walmir, que me ajudaram sempre que precisei.

Ao meu grande amigo Pedro Augusto Alecrin Coelho, pelos ensinamentos de vida e experiências profissionais vivenciadas.

Aos professores Ednaldo Ernerto, Antônio Carlos, César Castílho, Mascus Vinícius, Adriano Pedrosa, Paulo Figueiredo, Cristiano Ferraz e Eduardo Leandro, pelo ensino e apoio em minha vida na academia.

A Valéria Bittencourt, pela eficiência e simpatia.

Ao professor Francisco Cysneiros, pelo seu profissionalismo e pelos conselhos que sempre me deu.

Ao professor Leandro Chaves, que sempre me apoiou e me estimulou a trilhar os melhores caminhos em minha vida profissional.

Ao grandioso professor e orientador Francisco Cribari Neto, pelo apoio na orientação da tese, pelos grandes ensinamentos que me proporcionou, pela sua seriedade, pelo apoio nas cartas de recomendação, por confiar em meu potencial.

Ao CNPq, pelo apoio financeiro.

Resumo

É prática comum em estudos empíricos a realização de inferências sobre os parâmetros que indexam o modelo linear de regressão com base no estimador de mínimos quadrados ordinários (EMQO) mesmo quando se suspeita da presença de heteroscedasticidade. Nesse caso, utilizam-se estimativas assintoticamente válidas das variâncias dos estimadores no processo inferencial. Na presente dissertação, nós utilizamos integração numérica para avaliar os desempenhos de testes quase- t realizados segundo essa estratégia. Nós propomos ainda um novo estimador consistente de variâncias e covariâncias, denotado por $HC4m$. Esse estimador é uma versão modificada do estimador $HC4$. Nós propomos ainda uma versão modificada do estimador $HC5$: $HC5m$. A evidência numérica sugere que testes baseados nos estimadores propostos são tipicamente mais confiáveis que testes baseados em estimadores alternativos.

Palavras-chave: modelos lineares de regressão; heteroscedasticidade; testes quase- t .

Abstract

It is common practice to base inferences on the parameters that index the linear regression model on the ordinary least squares estimator (OLSE) even when one suspects that the data are heteroskedastic. In that case, one must use asymptotically valid variance estimates. In this thesis, we use numerical integration to evaluate the finite sample behavior of different quasi- t tests. We also propose a new heteroskedasticity-consistent covariance matrix estimator: $HC4m$. This new estimator is a modified version of $HC4$. We also propose a modified version of the $HC5$ estimator: $HC5m$. Our numerical evidence favors the two estimators we propose when the interest lies in associated hypothesis testing inference.

Key-words: linear regression models; heteroskedasticity; quasi- t tests.

1	Conceitos Iniciais	1
1.1	O Modelo de Regressão Linear e Estimadores Pontuais	1
1.2	Estimadores Consistentes para a Covariância do EMQO	8
1.3	Inferência sob Heteroscedasticidade de Forma Desconhecida	10
1.3.1	Variância de combinações lineares do vetor $\hat{\beta}$	12
1.3.2	Estatísticas quase- t	13
1.4	O Algoritmo de Imhof	13
1.4.1	A Função ProbImhof	15
1.5	A Plataforma Computacional	15
1.6	Organização da Dissertação	16
2	Avaliação Numérica Exata do Teste $HC4$	17
2.1	Introdução	17
2.2	Avaliação Numérica Inicial	18
2.2.1	Resultados numéricos	18
2.3	Avaliação Numérica com Dados Reais	22
2.4	Análise Numérica com os Quantis 0.90 e 0.95 da Distribuição Nula Limite	24
2.5	Conclusão	25

3	Um Novo Estimador $HC4$	35
3.1	Introdução	35
3.2	Uma Motivação para a Nossa Proposta	36
3.3	Uma Nova Proposta de Estimador	38
3.4	Avaliação Numérica	39
3.5	Análise Numérica com Dados Reais	43
3.6	Conclusão	44
4	Um Novo Estimador $HC5$	50
4.1	Introdução	50
4.2	Uma Versão Modificada do Estimador $HC5$	50
4.3	Avaliação Numérica	54
4.4	Análise de poder	59
4.5	Conclusão	61
5	Conclusões finais e sugestões para trabalhos futuros	63
	Apêndice A	64
	Apêndice B	69
	Referências Bibliográficas	72

1.1 O Modelo de Regressão Linear e Estimadores Pontuais

O modelo de interesse é o modelo de regressão linear definido por

$$y = X\beta + u, \quad (1.1)$$

onde y é um vetor $n \times 1$ de observações da variável dependente, X é uma matriz fixa $n \times p$ de regressores com posto p ($p < n$), β é um vetor fixo $p \times 1$ de parâmetros desconhecidos e u é um vetor aleatório $n \times 1$ de erros. Aqui, n é o número de observações, ou seja, o tamanho da amostra.

As seguintes suposições são comumente elencadas:

- S1. O modelo está corretamente especificado;
- S2. $\mathbb{E}(u_t) = 0$, $t = 1, \dots, n$;
- S3. $\text{var}(u_t) = \sigma_t^2$ e $0 < \sigma_t^2 < \infty$, $t = 1, \dots, n$;
- S3'. $\text{var}(u_t) = \sigma^2$ e $0 < \sigma^2 < \infty$, $t = 1, \dots, n$;
- S4. $\text{cov}(u_t, u_s) = 0 \quad \forall t \neq s$, i.e., $\mathbb{E}(u_t u_s) = 0 \quad \forall t \neq s$;
- S5. $\lim_{n \rightarrow \infty} n^{-1}(X'X) = Q$, onde Q é uma matriz positiva definida;

- $S5'$. $\lim_{n \rightarrow \infty} n^{-1}(X'\Omega X) = S$, onde S é uma matriz positiva definida e $\Omega = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$ é a matriz de covariâncias do vetor de erros u ;
- $S6$. $u \sim \mathcal{N}(0, \Omega)$, onde 0 é um vetor $n \times 1$ de zeros.

Observe que sob homoscedasticidade a suposição $S3$ se reduz à suposição $S3'$, ou seja, as variâncias dos erros são iguais a σ^2 , uma constante positiva e finita.

Em geral, em modelagem de regressão o interesse reside em realizar inferências sobre o vetor de parâmetros β , que representa o efeito dos regressores sobre a média da variável resposta. Um método comumente utilizado para a estimação de β é o método de mínimos quadrados ordinários, que consiste em estimar o vetor de parâmetros de regressão através da minimização da soma dos quadrados dos erros do modelo, isto é, deseja-se encontrar o vetor $\hat{\beta}$ que minimiza a função

$$SQ(\beta) = u'u = (y - X\beta)'(y - X\beta).$$

O estimador de mínimos quadrados ordinários (EMQO) de β é facilmente encontrado e possui forma fechada:

$$\hat{\beta} = (X'X)^{-1}X'y. \quad (1.2)$$

Este estimador possui algumas propriedades desejáveis sob algumas das suposições inicialmente listadas:

(i) Sob as suposições $S1$ e $S2$, $\hat{\beta}$ é não-viesado, isto é, $\mathbb{E}(\hat{\beta}) = \beta$, para todo $\beta \in \mathbb{R}^p$. Para provar esta propriedade basta substituir y por $X\beta + u$ na expressão para $\hat{\beta}$ dada na equação (1.2), o que conduz a $\hat{\beta} = \beta + (X'X)^{-1}X'u$. Tomando o valor esperado facilmente estabelecemos a propriedade de não-viés de $\hat{\beta}$.

(ii) Sob as suposições $S1$, $S2$ e $S5$, $\hat{\beta}$ é consistente para β .¹

¹O estimador $\hat{\theta}$ é dito ser consistente para o parâmetro θ se qualquer sequência de estimadores $\hat{\theta}_n$ converge em probabilidade para θ , ou seja, se, para todo $\epsilon > 0$ e para toda sequência $\hat{\theta}_n$, tivermos

$$\lim_{n \rightarrow \infty} \Pr(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

Escrevemos $\hat{\theta}_n \xrightarrow{p} \theta$ ou $\text{plim}(\hat{\theta}_n) = \theta$ para indicar que a sequência $\{\hat{\theta}_n\}$ converge em probabilidade para θ .

Para estabelecer a consistência do estimador $\hat{\beta}$, inicialmente provaremos que $\text{plim}(\frac{X'u}{n}) = 0$.

Temos que

$$X'u = \sum_{t=1}^n x'_t u_t,$$

onde x_t representa a t -ésima linha da matriz X .

Sendo a suposição $S2$ válida e usando propriedades do valor esperado concluímos que $\mathbb{E}(X'u) = 0$. Pela lei fraca dos grandes números de Chebyshev, tem-se que

$$\text{plim}\left(\frac{X'u}{n}\right) = 0.$$

Desta forma, pelo teorema de Slutsky, que estabelece que o limite de uma soma (ou produto) de seqüências estocásticas é a soma (ou produto) dos limites destas seqüências, e sendo a suposição $S5$ válida, segue que

$$\begin{aligned} \text{plim}(\hat{\beta}) &= \text{plim} [\beta + (X'X)^{-1}X'u] \\ &= \text{plim}(\beta) + \text{plim} \left[n(X'X)^{-1} \left(\frac{X'u}{n} \right) \right] \\ &= \beta + \text{plim} [n(X'X)^{-1}] \text{plim} \left(\frac{X'u}{n} \right) \\ &= \beta + Q^{-1}0 = \beta, \end{aligned}$$

o que prova a consistência de $\hat{\beta}$.

Para obter a matriz de covariâncias de $\hat{\beta}$, note que $\hat{\beta} - \beta = (X'X)^{-1}X'u$ e, assim,

$$\begin{aligned} \text{cov}(\hat{\beta}) &= \mathbb{E} \left\{ \left[\hat{\beta} - \mathbb{E}(\hat{\beta}) \right] \left[\hat{\beta} - \mathbb{E}(\hat{\beta}) \right]' \right\} \\ &= \mathbb{E} \left\{ (\hat{\beta} - \beta)(\hat{\beta} - \beta)' \right\} \\ &= \mathbb{E} \left\{ [(X'X)^{-1}X'u] [(X'X)^{-1}X'u]' \right\} \\ &= \mathbb{E} \left\{ (X'X)^{-1}X'uu'X(X'X)^{-1} \right\} \\ &= (X'X)^{-1}X' [\mathbb{E}(uu')] X(X'X)^{-1} \\ &= (X'X)^{-1}X' [\text{cov}(u)] X(X'X)^{-1} \\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1}. \end{aligned}$$

Sendo a suposição $S3'$ (homoscedasticidade) válida, temos que $\Omega = \sigma^2 I_n$, onde I_n é a matriz identidade de ordem n . Portanto,

$$\text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

O teorema a seguir estabelece que, sendo verificadas algumas das suposições inicialmente listadas, o EMQO é BLUE ('Best Linear Unbiased Estimator'), sendo portanto o melhor estimador dentre todos os estimadores lineares e não-viesados.

Teorema 1.1.1 (Gaus-Markov). *Sob as suposições $S1$, $S2$, $S3'$ e $S4$, o estimador de mínimos quadrados ordinários (EMQO) é o melhor estimador na classe de estimadores lineares² e não-viesados de β , isto é, o EMQO é o estimador de variância mínima na classe de estimadores lineares não-viesados de β .*

Demonstração. Seja $\hat{\beta}^* = Cy$, onde C é uma matriz não-estocástica de dimensão $p \times n$, com $p \leq n$ e tal que $C \neq (X'X)^{-1}X'$, isto é, $\hat{\beta}^*$ é uma função linear de y diferente de $\hat{\beta}$. A variância de $\hat{\beta}^*$ é dada por

$$\text{var}(\hat{\beta}^*) = C [\text{var}(y)] C' = C [\sigma^2 I_n] C' = \sigma^2 (CC'),$$

uma vez que a suposição $S3'$ (homoscedasticidade) é válida. Dado que $\hat{\beta}^*$ é não-viesado (as suposições $S1$ e $S2$ são válidas), devemos ter

$$\mathbb{E}(\hat{\beta}^*) = C\mathbb{E}(y) = CX\beta = \beta,$$

para todo $\beta \in \mathbb{R}^p$. Logo,

$$CX = I_n.$$

Sendo assim, temos

$$\begin{aligned} \text{var}(\hat{\beta}^*) - \text{var}(\hat{\beta}) &= \sigma^2 CC' - \sigma^2 CX(X'X)^{-1}X'C' \\ &= \sigma^2 [CC' - CX(X'X)^{-1}X'C'] \\ &= \sigma^2 C [I_n - X(X'X)^{-1}X'] C'. \end{aligned}$$

²Um estimador de β é dito ser linear se pode ser escrito como combinação linear das componentes do vetor de observações y .

Dado que $M = I_n - X(X'X)^{-1}X'$ é simétrica e idempotente temos que

$$\text{var}(\hat{\beta}^*) - \text{var}(\hat{\beta}) = \sigma^2(CM)(CM)'$$

é positiva definida.

De forma análoga, é fácil mostrar que se $\hat{\beta}^* = Cy$ e $\mathbb{E}(\hat{\beta}^*) = \beta$ então $\text{var}(\lambda'\hat{\beta}^*) > \text{var}(\lambda'\hat{\beta})$, $\forall \lambda \in \mathbb{R}^p$, onde C é uma matriz não-estocástica $p \times n$, com $p \leq n$ e tal que $C \neq (X'X)^{-1}X'$. Assim, concluímos a prova do teorema. $\square$

Para provar o Teorema de Gauss-Markov foi necessário supor homoscedasticidade. Sob heteroscedasticidade (violação de $S3'$), o resultado não é válido e o EMQO torna-se ineficiente,³ embora permaneça não-viesado, consistente e assintoticamente normal.

A fim de verificar este fato, suponhamos que no modelo (1.1) a suposição $S3'$ não seja válida, isto é, $\text{cov}(u) \neq \sigma^2 I_n$. Uma vez que Ω é positiva definida, existe uma matriz P tal que

$$P\Omega P' = I_n. \quad (1.3)$$

Assim,

$$\begin{aligned} \Omega &= P^{-1}P\Omega P'(P')^{-1} \\ &= P^{-1}I_n(P')^{-1} \\ &= (P'P)^{-1}. \end{aligned}$$

Logo,

$$P'P = \Omega^{-1}.$$

Pré-multiplicando a equação (1.1) pela matriz P obtemos

$$Py = PX\beta + Pu$$

e podemos escrever

$$y^* = X^*\beta + u^*, \quad (1.4)$$

³O termo ineficiente diz respeito ao Teorema de Gauss-Markov, que passa a não ter validade sob heteroscedasticidade.

onde $y^* = Py$, $X^* = PX$ e $u^* = Pu$. O fato interessante é que o modelo transformado (1.4) é homoscedástico, pois

$$\text{cov}(u^*) = \mathbb{E}(u^* u^{*'}) = \mathbb{E}(Puu'P') = P\mathbb{E}(uu')P' = \sigma^2 I_n.$$

Adicionalmente, a suposição $S2$ é válida, pois

$$\mathbb{E}(u^*) = P\mathbb{E}(u) = 0.$$

Nesse modelo transformado, podemos estimar β pelo método de mínimos quadrados obtendo o estimador de mínimos quadrados generalizado (EMQG):

$$\begin{aligned}\hat{\beta}_G &= (X^{*'}X^*)^{-1}X^{*'}y^* \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.\end{aligned}$$

Sobre o estimador $\hat{\beta}_G$, sabemos que:

- (i) É linear, isto é, $\hat{\beta}_G = Sy$, onde $S = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$.
- (ii) É não-viesado, pois

$$\begin{aligned}\mathbb{E}(\hat{\beta}_G) &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\mathbb{E}(y) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta \\ &= \beta.\end{aligned}$$

- (iii) Sua matriz de covariâncias é dada por

$$\begin{aligned}\text{cov}(\hat{\beta}_G) &= \mathbb{E}\left[(\hat{\beta}_G - \beta)(\hat{\beta}_G - \beta)'\right] \\ &= \sigma^2 (X^{*'}X^*)^{-1} \\ &= \sigma^2 (X'P'PX)^{-1} \\ &= \sigma^2 (X'\Omega^{-1}X)^{-1}.\end{aligned}$$

Note que

$$\text{cov}(\hat{\beta}) - \text{cov}(\hat{\beta}_G) = \sigma^2 A\Omega A',$$

onde $A = (X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}$. Sendo a diferença $\text{cov}(\hat{\beta}) - \text{cov}(\hat{\beta}_G)$ uma matriz positiva semi-definida, podemos concluir que $\hat{\beta}_G$ é um estimador melhor do que $\hat{\beta}$, em sintonia com o Teorema de Gauss-Markov. Vale observar que, sob homoscedasticidade, o EMQG coincide com o EMQO, isto é, sob a suposição $S3'$ temos $\hat{\beta} = \hat{\beta}_G$. Para verificar este fato, basta substituir a matriz $\Omega = \sigma^2 I_n$ na expressão de $\hat{\beta}_G$.

Ao considerarmos o EMQG como uma solução para a ineficiência do EMQO sob heteroscedasticidade nos deparamos com o inconveniente de que para encontrar o EMQG precisamos da matriz P e, conseqüentemente, da matriz Ω , ou seja, para usar o EMQG precisaríamos conhecer a matriz Ω , o que em geral não acontece. Uma alternativa para solucionar esse problema é substituir a matriz Ω por um estimador consistente $\hat{\Omega}$, obtendo-se assim o estimador de mínimos quadrados generalizado viável (EMQGV):

$$\hat{\beta}_V = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y. \quad (1.5)$$

Sob heteroscedasticidade, em geral supomos que os elementos da matriz Ω são funções de uma pequena quantidade de parâmetros desconhecidos e fazemos a estimação destes parâmetros.

Sobre o EMQGV sabemos que:

- (i) É não-linear. Note que a matriz $\hat{\Omega}$ depende do vetor de observações y , indicando assim que $\hat{\Omega}$ e y são correlacionados. Portanto, o Teorema de Gauss-Markov não se aplica a esse estimador.
- (ii) Não é possível mostrar, em geral, que $\hat{\beta}_V$ é não-viesado.
- (iii) Dada a consistência de $\hat{\Omega}$, em grandes amostras o EMQGV comporta-se de forma semelhante ao EMQG.

Uma desvantagem do EMQGV é que ele requer do usuário uma suposição sobre a forma funcional da matriz Ω . Uma forma de evitar tal suposição é usar o EMQO que, mesmo sendo ineficiente sob heteroscedasticidade, é consistente, não-viesado e assintoticamente normal, e estimar consistentemente sua matriz de covariâncias a fim de obter erros-padrão assintoticamente válidos. Lembremos que, sob homoscedasticidade, $\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, que pode ser facilmente estimada por $\hat{\sigma}^2(X'X)^{-1}$, onde $\hat{\sigma}^2 = \hat{u}'\hat{u}/(n - p)$, em que $\hat{u}$ é o vetor de resíduos de mínimos quadrados ordinários. Sob heteroscedasticidade, é necessário obter estimadores consistentes para

$\text{cov}(\hat{\beta})$.

Em 1980, Halbert White mostrou que para estimar $\text{cov}(\hat{\beta})$ consistentemente sob heteroscedasticidade de forma desconhecida é preciso apenas estimar $X'\Omega X$ ($p \times p$) consistentemente, ou seja, não é necessário estimar Ω ($n \times n$) de forma consistente. Adicionalmente, ele propôs um estimador consistente para a matriz $\text{cov}(\hat{\beta})$.

1.2 Estimadores Consistentes para a Covariância do EMQO

Sob suspeita de heteroscedasticidade, é comum estimar o vetor de parâmetros β usando o EMQO e utilizar algum estimador consistente para sua matriz de covariâncias. Em 1980, Halbert White propôs um estimador consistente baseando-se no fato de que para estimar $\text{cov}(\hat{\beta})$, ao invés de estimar consistentemente Ω , que possui n elementos, basta estimar $X'\Omega X$, cuja dimensão independe do tamanho amostral, de forma consistente. O estimador de White tem a seguinte forma:

$$HC0 = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1} = B\hat{\Omega}_0B' = BE_0\hat{\Omega}B',$$

onde $\hat{\Omega} = \text{diag}\{\hat{u}_1^2, \dots, \hat{u}_n^2\}$, $E_0 = I_n$ e $B = (X'X)^{-1}X'$.

O estimador $HC0$ é consistente para $\text{cov}(\hat{\beta})$ independentemente da validade da suposição $S3$, mas tende a ser demasiadamente otimista em amostras de tamanho pequeno a moderado, isto é, tende a subestimar a matriz $\text{cov}(\hat{\beta})$ podendo apresentar alto viés em pequenas amostras e, sob alta alavancagem, vieses consideráveis. Long e Ervin (2000), por meio de simulações de Monte Carlo, concluíram que o estimador $HC0$ não é recomendável em amostras de tamanho inferior a 250. Cribari-Neto e Zarkos (1999) e MacKinnon e White (1985), também por meio de simulações, evidenciaram que em amostras pequenas esse estimador pode apresentar viés muito alto. A presença de observações de alta alavancagem pode tornar o viés do estimador ainda mais severo; ver Chesher e Jewitt (1987). Conseqüentemente, testes cujas estatísticas empregam o estimador $HC0$ tendem a ser liberais, isto é, os tamanhos efetivos desses testes tendem a ser maiores do que a probabilidade do erro tipo I especificada (nível nominal do teste).

Ao longo dos anos surgiram algumas variantes do estimador de White que procuraram atenuar

o efeito de pontos de alavanca sobre as inferências resultantes. Seja h_t o t -ésimo elemento diagonal da matriz $H = X(X'X)^{-1}X'$. Alguns estimadores propostos na literatura são os seguintes:

1- **Estimador $HC1$** (Hinkley, 1977):

$$HC1 = B\hat{\Omega}_1B' = BE_1\hat{\Omega}B', \text{ onde } E_1 = \left(\frac{n}{n-p}\right)I_n.$$

Esse estimador tenta corrigir a tendência do estimador $HC0$ de subestimar a variância do EMQO em amostras finitas trocando a matriz E_0 (matriz identidade) pelo fator corretivo $\frac{n}{n-p}$.

2- **Estimador $HC2$** (Horn, Horn e Duncan, 1975):

$$HC2 = B\hat{\Omega}_2B' = BE_2\hat{\Omega}B', \text{ onde } E_2 = \text{diag}\{1/(1 - h_t)\}.$$

A proposta do estimador $HC2$ é incorporar um ajuste baseado nas medidas de alavancagem na estimativa da matriz de covariâncias. Como observado por Chesher e Jewitt (1987), a possível existência de um viés severo no estimador de White ocorre quando há alavancagens elevadas, o que se deve ao fato de que os resíduos de mínimos quadrados associados possuem magnitude reduzida (em valor absoluto) em média e o estimador interpreta pequenos resíduos como indicação de variâncias pequenas. Sob homoscedasticidade, o estimador $HC2$ é não-viesado.

3- **Estimador $HC3$** (Davidson e MacKinnon, 1993):

$$HC3 = B\hat{\Omega}_3B' = BE_3\hat{\Omega}B', \text{ onde } E_3 = \text{diag}\{1/(1 - h_t)^2\}.$$

Um estudo numérico detalhado dos comportamentos em pequenas amostras de diferentes estimadores foi feito por Long e Ervin (2000), que concluíram que o estimador $HC3$ possui o melhor desempenho dentre os estimadores apresentados até o momento em pequenas amostras no que tange ao desempenho de testes associados. Tal estimador incorpora uma correção mais acentuada dos graus de alavancagem a fim de atenuar o fato de que os resíduos tendem a flutuar menos que os verdadeiros erros. Os testes que usam o estimador $HC3$ tendem a funcionar consideravelmente melhor do que aqueles que usam $HC0$. O estimador $HC3$ é uma aproximação ao obtido através da técnica de jackknife.⁴

4- **Estimador $HC4$** (Cribari-Neto, 2004):

$$HC4 = B\hat{\Omega}_4B' = BE_4\hat{\Omega}B', \text{ onde } E_4 = \text{diag}\{1/(1 - h_t)^{\delta_t}\} \text{ com } \delta_t = \min\{4, (nh_t)/p\}.$$

⁴Essa técnica baseia-se em recalcular n vezes as estimativas de mínimos quadrados ordinários para o vetor β retirando em cada vez uma das observações e utilizar a variância das estimativas obtidas como estimativa da variância do EMQO original.

A proposta do estimador *HC4* é acentuar a intensidade do ajuste por alavancagens feito no estimador *HC3* elevando o termo de ajuste $(1 - h_t)$ à potência δ_t , conforme explicitado acima. Desta forma, os resíduos ao quadrado serão mais inflacionados quando as observações correspondentes apresentarem maior grau de alavancagem. A motivação para considerar a influência da alta alavancagem de algumas observações pode ser encontrada nos resultados de Cribari-Neto e Zarkos (2001). Segundo esses autores, a presença de pontos de alta alavancagem na matriz de regressores é mais decisiva para o comportamento de estimadores consistentes para a matriz de covariâncias de $\hat{\beta}$ em amostras finitas do que o grau de heteroscedasticidade.

5- **Estimador *HC5*** (Cribari-Neto, Souza e Vasconcellos, 2007):

$HC5 = B\hat{\Omega}_5B' = BE_5\hat{\Omega}B'$, onde $E_5 = \text{diag}\{1/\sqrt{(1 - h_t)^{\delta_t}}\}$, sendo o expoente $\delta_t = \min\{(nh_t)/p, \max\{4, (nkh_{\max})/p\}\}$, em que $h_{\max} = \max\{h_1, \dots, h_n\}$ e k é uma constante pré-definida pertencente ao intervalo $[0, 1]$.

O estimador *HC5* foi desenvolvido no intuito de melhorar a confiabilidade de inferências realizadas a partir do estimador *HC4*. O estimador *HC5* leva em consideração não apenas a razão entre a t -ésima alavancagem e o grau médio de alavancagem, mas também o grau de influência exercido pelo *nível máximo de alavancagem* sobre a inferência resultante. Tal estimador foi o primeiro a considerar o nível máximo de alavancagem nos fatores de desconto dentre todos os estimadores consistentes da matriz de covariâncias de $\hat{\beta}$. Avaliações numéricas no desconto de todos os resíduos ao quadrado (ver Cribari-Neto, Souza e Vasconcelos, 2007) sugerem que com $k = 0.7$ o estimador *HC5* se comporta de forma superior a todos os anteriores em amostras finitas no que tange a testes quase- t associados.

1.3 Inferência sob Heteroscedasticidade de Forma Desconhecida

Sob as suposições $S1, S2, S3'$ (homoscedasticidade) e $S6$, a estatística t usada para testar a hipótese nula $H_0 : c'\beta = m$ versus $H_1 : c'\beta \neq m$, onde c é um dado vetor $p \times 1$ e m é uma dada constante real, é

$$t = \frac{c'\hat{\beta} - m}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}}, \quad (1.6)$$

onde $\widehat{\text{var}}(c'\hat{\beta})$ é obtido de $\hat{\sigma}^2(X'X)^{-1}$.

Essa estatística tem, sob H_0 , distribuição t de Student com $n-p$ graus de liberdade e converge em distribuição para $\mathcal{N}(0, 1)$. Em particular, quando $c'\beta = \beta_k$, onde β_k é o k -ésimo elemento do vetor de parâmetros β , pode-se testar a hipótese $H_0 : \beta_k = \beta_k^{(0)} = m$ versus $H_1 : \beta_k \neq \beta_k^{(0)} = m$. A estatística t nesse caso passa a ser

$$t = \frac{\hat{\beta}_k - \beta_k^{(0)}}{\sqrt{\widehat{\text{var}}(\hat{\beta}_k)}}.$$

As duas expressões da estatística t acima contêm em seus denominadores a estimativa de mínimos quadrados ordinários da variância da combinação linear do EMQO do vetor de parâmetros β determinada pelo vetor c . O fato importante a ser observado é que, quando há violação da suposição $S3'$, a estatística t dada na equação (1.6) não possui distribuição t de Student sob H_0 .

Como sabemos, $\hat{\beta}$ é consistente para β e é também assintoticamente normal mesmo sob heteroscedasticidade. Desta forma, podemos basear nossas inferências nesse estimador. Quando modificamos a estatística t dada em (1.6) substituindo a variância estimada usada no denominador por uma estimativa obtida de um estimador consistente (por exemplo, um dos estimadores HCI , $i = 0, \dots, 5$), obtemos o que é conhecido como ‘estatística quase- t ’, que será denotada por τ . Testes baseados em estatísticas quase- t são chamados de testes quase- t .

Uma vez que $\frac{\sqrt{\text{var}(c'\hat{\beta})}}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}} \xrightarrow{p} 1$, temos que

$$\begin{aligned} \tau &= \frac{c'\hat{\beta} - m}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}} \\ &= \frac{c'\hat{\beta} - m}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}} \frac{\sqrt{\text{var}(c'\hat{\beta})}}{\sqrt{\text{var}(c'\hat{\beta})}} \\ &= \frac{c'\hat{\beta} - m}{\sqrt{\text{var}(c'\hat{\beta})}} \frac{\sqrt{\text{var}(c'\hat{\beta})}}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}}. \end{aligned}$$

Dado que $\frac{\sqrt{\text{var}(c'\hat{\beta})}}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}} \xrightarrow{p} 1$, é fácil estabelecer que, sob a hipótese nula em teste,

$$\tau = \frac{c'\hat{\beta} - m}{\sqrt{\text{var}(c'\hat{\beta})}} \frac{\sqrt{\text{var}(c'\hat{\beta})}}{\sqrt{\widehat{\text{var}}(c'\hat{\beta})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Assim, sob H_0 ,

$$\tau^2 \xrightarrow{d} \chi_1^2.$$

Nosso próximo passo será escrever a estatística τ^2 como razão de formas quadráticas de variáveis aleatórias normais a fim de avaliar, de forma exata, a função de distribuição desta estatística usando o algoritmo de Imhof (1961). Para tanto, assumiremos que a suposição $S6$ (normalidade) é válida.

1.3.1 Variância de combinações lineares do vetor $\hat{\beta}$

Seja c um vetor $p \times 1$ de constantes. Consideremos a combinação linear $c'\hat{\beta}$ dos elementos de $\hat{\beta}$. Desta forma,

$$\Psi = \text{var}(c'\hat{\beta}) = c' \left[\text{cov}(\hat{\beta}) \right] c. \quad (1.7)$$

Nosso objetivo aqui reside em estimar Ψ de forma consistente. Para tanto, podemos usar a equação (1.7) e um estimador consistente para a matriz $\text{cov}(\hat{\beta})$:

$$\hat{\Psi}_i = c' B E_i \hat{\Omega} B' c, \quad i = 0, \dots, 5.$$

Seja $v_i = E_i^{1/2} B' c$, $i = 0, \dots, 5$, e defina

$$V_i = (v_i v_i')_d, \quad i = 0, \dots, 5. \quad (1.8)$$

Desta forma, podemos escrever

$$\hat{\Psi}_i = v_i' \hat{\Omega} v_i.$$

É fácil verificar que

$$\hat{\Psi}_i = \hat{u}' V_i \hat{u}, \quad i = 0, \dots, 5.$$

Sendo assim, supondo que $S6$ é válida, $\hat{\Psi}_i$ pode ser escrito como uma forma quadrática em um vetor normal (o vetor de resíduos) de média zero e cujos elementos são correlacionados. É ainda possível escrever $\hat{\Psi}_i$ como uma forma quadrática em um vetor normal de média zero e matriz de covariâncias I_n . Seguindo os desenvolvimentos em Cribari-Neto, Ferrari e Cordeiro (2000), é possível escrever

$$\hat{\Psi}_i = w' G_i w,$$

onde $w \sim \mathcal{N}(0, I)$ e $G_i = \Omega^{1/2}(I - H)V_i(I - H)\Omega^{1/2}$.

1.3.2 Estatísticas quase- t

No início da presente seção vimos que, mesmo sob heteroscedasticidade, a distribuição da estatística τ sob H_0 converge para $\mathcal{N}(0, 1)$ se considerarmos um estimador consistente para $\text{var}(c'\hat{\beta})$, o que implica que, sob H_0 ,

$$\tau^2 \xrightarrow{d} \chi_1^2. \quad (1.9)$$

Sendo assim, nos casos em que $S3'$ é violada pode-se testar H_0 versus H_1 via testes quase- t .

Considerando o numerador de τ^2 e sendo $w \sim \mathcal{N}(0, I_n)$, temos que

$$\begin{aligned} (c'\hat{\beta} - m)^2 &= \{c'\beta + c'(X'X)^{-1}X'\Omega^{1/2}w - m\}'\{c'\beta + c'(X'X)^{-1}X'\Omega^{1/2}w - m\} \\ &= \{(c'\beta - m) + c'(X'X)^{-1}X'\Omega^{1/2}w\}'\{(c'\beta - m) + c'(X'X)^{-1}X'\Omega^{1/2}w\} \\ &= (c'\beta - m)'(c'\beta - m) + 2(c'\beta - m)c'(X'X)^{-1}X'\Omega^{1/2}w \\ &\quad + w'\Omega^{1/2}X(X'X)^{-1}cc'(X'X)^{-1}X'\Omega^{1/2}w. \end{aligned}$$

Desta forma, τ^2 pode ser escrita como

$$\tau^2 = \frac{w'Rw}{w'G_iw} + \frac{(c'\beta - m)'(c'\beta - m) + 2(c'\beta - m)c'(X'X)^{-1}X'\Omega^{1/2}w}{w'G_iw},$$

onde $R = \Omega^{1/2}X(X'X)^{-1}cc'(X'X)^{-1}X'\Omega^{1/2}$, $G_i = \Omega^{1/2}(I - H)V_i(I - H)\Omega^{1/2}$ e $w \sim \mathcal{N}(0, I_n)$.

Sob H_0 , o segundo termo da estatística τ^2 se anula e, assim,

$$\Pr(\tau^2 \leq \zeta | H_0) = \Pr_0 \left(\frac{w'Rw}{w'G_iw} \leq \zeta \right), \quad (1.10)$$

onde $\Pr_0$ denota probabilidade sob a hipótese nula e $\zeta > 0$.

É possível avaliar numericamente a probabilidade $\Pr(\tau^2 \leq \zeta | H_0)$ com ζ real positivo utilizando o algoritmo de Imhof (1961) e compará-la com $\Pr(\chi_1^2 \leq \zeta)$ por meio de variações na constante ζ .

1.4 O Algoritmo de Imhof

Seja a forma quadrática $Q(x) = x'Ax$, em que A é uma matriz simétrica e x é um vetor normal $n \times 1$ com vetor de médias μ e matriz de covariâncias Ω . Inicialmente consideramos

$\mu = 0$. Se Ω é positiva definida e simétrica, então Ω possui decomposição de Choleski e pode ser escrita como $\Omega = \Gamma\Gamma'$, onde Γ é uma matriz triangular inferior não-singular. Segue que $\Gamma' A \Gamma$ é simétrica e, portanto, seus autovalores são reais, o que garante a existência de uma transformação linear ortogonal não-singular de x tal que $Q(x)$ tem a mesma distribuição de $Q(y) = \sum_{i=1}^m \lambda_i y_i^2$, em que os y_i 's são variáveis normais padrão e $\lambda_1 \geq \dots \geq \lambda_m$ representam os autovalores de $\Gamma' A \Gamma$ (ver Scheffé, 1959, Apêndice II). Para o caso não-central ($\mu \neq 0$), a forma quadrática $Q(x)$ terá a mesma distribuição de (Imhof, 1961)

$$Q(y) = \sum_{r=1}^m \lambda_r \chi_{h_r; \sigma_r^2}^2,$$

onde os h_r 's são as ordens de multiplicidade dos λ_r 's e $\chi_{h_r; \sigma_r^2}^2$ são variáveis aleatórias qui-quadrado independentes com h_r graus de liberdade e parâmetro de não-centralidade σ_r^2 .

A função característica de Q é dada por

$$\phi(t) = \prod_{r=1}^m (1 - 2it\lambda_r t)^{-1/2h_r} \exp \left\{ i \sum_{r=1}^m \frac{t\sigma_r^2 \lambda_r}{1 - 2it\lambda_r} \right\}, \quad (1.11)$$

onde i representa a unidade imaginária.

Podemos escrever a função de distribuição acumulada de $Q(x)$ (Imhof, 1961; Gil-Pelaez, 1951) como

$$F(\zeta) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty t^{-1} \text{Im} \{ \exp(-it\zeta) \phi(t) \} dt,$$

onde $\text{Im}(s)$ denota a parte imaginária do número complexo s , i é a unidade imaginária e $\phi(t)$ é a função característica de $Q(x)$ dada pela equação (1.10).

Após algumas manipulações algébricas é possível reescrever a função $F(\zeta)$ como

$$F(\zeta) = \Pr(Q(x) < \zeta) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin \theta(u)}{u \rho(u)} du, \quad (1.12)$$

onde

$$u = 2\zeta,$$

$$\theta(u) = \frac{1}{2} \sum_{r=1}^m [h_r \arctan(\lambda_r u) + \sigma_r^2 \lambda_r u (1 + \lambda_r^2 u^2)^{-1}] - \frac{1}{2} \zeta u$$

e

$$\rho(u) = \prod_{r=1}^m (1 + \lambda_r^2 u^2)^{\frac{1}{4} h_r} \exp \left\{ \frac{1/2 \sum_{r=1}^m (\sigma_r \lambda_r u)^2}{(1 + \lambda_r^2 u^2)} \right\}.$$

1.4.1 A Função ProbImhof

A função `ProbImhof`, disponível para a linguagem de programação `0x` (Doornik, 2001), avalia numericamente a probabilidade de um quociente de formas quadráticas de variáveis normais ser menor ou igual a um dado valor $\zeta > 0$. Sendo assim, considerando a razão de formas quadráticas $Rq = \frac{z'Az}{z'Bz}$ no vetor normal z de média m e matriz de covariâncias S , onde A e B são matrizes quadradas cuja ordem é o número de linhas de z , a função `ProbImhof(x,A,B,m,S)` avalia numericamente a probabilidade $\Pr(Rq \leq \zeta)$, $\zeta > 0$. Quando $B = 0$ calcula-se a probabilidade $\Pr(z'Az < \zeta)$.

A função `QuanImhof(p,A,B,m,S)` toma como primeiro argumento uma probabilidade e calcula o quantil associado.

Na presente dissertação, utilizamos a função `ProbImhof` a fim de calcular a probabilidade

$$\Pr(\tau^2 \leq \zeta | H_0) = \Pr_0 \left(\frac{w'Rw}{w'G_iw} \leq \zeta \right),$$

onde $R = \Omega^{1/2} X(X'X)^{-1} cc'(X'X)^{-1} X' \Omega^{1/2}$, $G_i = \Omega^{1/2}(I - H)V_i(I - H)\Omega^{1/2}$ e $w \sim \mathcal{N}(0, I_n)$.

O código fonte localizado no Apêndice B implementa na linguagem de programação `0x` a função `ProbImhof`.

Nele, destacamos as funções `imhofmgf(u)` e `QAGI(imhofmgf,0,1,&result,&abserr)`. A primeira calcula o integrando da expressão (1.12) e a segunda calcula a integral da expressão (1.12).

1.5 A Plataforma Computacional

Para a realização das avaliações numéricas presentes neste texto, foi usada a linguagem de programação `0x` criada por Jurgen Doornik em 1994. A linguagem `0x` permite a implementação de técnicas estatísticas com facilidade, precisão e eficiência e é distribuída gratuitamente para

uso acadêmico em <http://www.doornik.com> (para mais detalhes sobre a linguagem de programação `0x` (ver Doornik, 2001). Todos os gráficos apresentados foram produzidos no ambiente de programação, análise de dados e gráficos `R`, que se encontra disponível gratuitamente em <http://www.r-project.org>. O presente texto foi digitado usando o sistema tipográfico $\text{\LaTeX}$, também distribuído gratuitamente.

1.6 Organização da Dissertação

Como objetivo inicial, no Capítulo 2 faremos uma avaliação numérica exata utilizando o estimador $HC4$ que foi proposto por F. Cribari-Neto em 2004. A avaliação focará no estudo da constante de truncamento deste estimador, que usualmente assume valor 4.

No capítulo seguinte, proporemos um estimador alternativo (denotado por $HC4m$) que tenta corrigir o fato de que, embora as distribuições nulas exatas de estatísticas quase- t baseadas no estimador $HC4$ sejam em geral bem aproximadas na cauda pela distribuição nula limite (χ_1^2), elas são muitas vezes mal aproximadas no restante do suporte da distribuição. Por fim, no Capítulo 4, baseando-se na construção da nova versão proposta para o estimador $HC4$, proporemos também uma versão alternativa para o estimador $HC5$ (denotada por $HC5m$).

2.1 Introdução

Conforme visto no capítulo anterior, o estimador $HC4$ proposto por Cribari-Neto (2004) possui truncamento de desconto em $\gamma = 4$, o qual impede que o termo de ajuste $(1 - h_t)$ seja elevado a potências que excedem esse valor, ou seja, o estimador é dado por $HC4 = B\hat{\Omega}_4B'$, onde $B = (X'X)^{-1}X'$ e $\hat{\Omega}_4 = \text{diag}\{1/(1 - h_t)^{\delta_t}\}$ com $\delta_t = \min\{\gamma, nh_t/p\}$. A proposta do autor usou $\gamma = 4$. O presente capítulo tem por objetivo avaliar numericamente de forma exata o impacto do valor dessa constante (γ) sobre a aproximação da distribuição nula exata de estatísticas de testes baseadas nesse estimador pela distribuição nula limite (χ_1^2). Nosso interesse residirá em identificar os valores de γ que conduzem a testes quase- t associados confiáveis. Para isto, consideraremos apenas casos em que a variável resposta é normalmente distribuída e utilizaremos o algoritmo de integração numérica proposto por Imhof (1961). Os resultados que seguem sugerem o uso estimador $HC4$ com $\gamma = 4$ como uma escolha pertinente em situações de heteroscedasticidade e homoscedasticidade quando os quantis de interesse da distribuição nula limite (χ_1^2) são os quantis 0.90 e 0.95.

2.2 Avaliação Numérica Inicial

O modelo de regressão escolhido para a análise inicial é

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, \dots, n, \quad (2.1)$$

onde u_t é normalmente distribuído com média zero e variância $\sigma_t^2 = \exp(\alpha_1 + \alpha_2 x_t)$, α_1 e α_2 sendo constantes reais. Aqui, $\text{cov}(u_t, u_s) = 0 \quad \forall t \neq s$ (i.e., os erros são não-correlacionados). Foi usada a seguinte medida do grau de heteroscedasticidade:

$$\lambda = \frac{\max \sigma_t^2}{\min \sigma_t^2}.$$

Para erros homoscedásticos, teremos $\lambda = 1$; no caso de erros heteroscedásticos, quanto maior o grau de heteroscedasticidade maior o valor de λ .

Nosso interesse residirá no teste da hipótese nula $H_0 : c'\beta = m$ versus $H_1 : c'\beta \neq m$, onde c é um vetor $p \times 1$ e m é um valor real. Nas análises a seguir, consideraremos $m = 0$ e $c = (0, 1)'$. Sendo assim, estaremos interessados em testar $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. A estatística de teste é

$$\tau^2 = \frac{\hat{\beta}_1^2}{\widehat{\text{var}}(\hat{\beta}_1)},$$

onde $\widehat{\text{var}}(\hat{\beta}_1)$ é baseado no estimador *HCC4* com alterações em sua constante de truncamento γ .

2.2.1 Resultados numéricos

Variações feitas no valor de γ nos permitiram realizar uma avaliação inicial do impacto dessa constante sobre as aproximações assintóticas para as probabilidades $\Pr(\tau^2 \leq \zeta | H_0)$ para os valores $\zeta = 2.706, 3.841$ e 6.635 , que correspondem aos quantis 0.90, 0.95 e 0.99 da distribuição nula assintótica (χ_1^2) da estatística τ^2 e são os valores críticos do teste mais usados em aplicações práticas. As probabilidades calculadas usando os valores $\zeta = 2.706, 3.841$ e 6.635 estão denotadas nas tabelas pelas siglas *pcalc1*, *pcalc2* e *pcalc3*, respectivamente.

Os resultados que seguem foram obtidos comparando-se a distribuição nula exata da estatística τ^2 dada acima com sua distribuição nula limite (χ_1^2), da qual são tipicamente obtidos valores

críticos aproximados. O cálculo da distribuição exata de τ^2 foi feito com base no algoritmo de Imhof (1961) a partir do resultado apresentado na Seção 1.4.1.

O tamanho da amostra considerado foi $n = 25$; nós replicamos os valores das covariáveis para obter amostras de 50 observações. O desenho de regressão considerado foi com pontos de alavancagem, tendo sido os valores dos regressores escolhidos como realizações das distribuições lognormal padrão ($\mathcal{LN}(0,1)$) e t de Student com três graus de liberdade (t_3). É importante ressaltar que, nas mesmas análises numéricas feitas considerando-se desenhos de regressão sem pontos de alavancagem, as alterações na constante de truncamento γ não resultaram em impacto notável nas aproximações obtidas, o que nos levou a considerar apenas desenhos de regressão com pontos de alavancagem.

Seja h_t o t -ésimo elemento diagonal da matriz $H = X(X'X)^{-1}X'$. Note que $\sum_{t=1}^n h_t = p$, onde p é o número de parâmetros de regressão. Desta forma, a média dos h_t 's é igual a p/n . Uma regra *ad hoc* usual é considerar como pontos de alavanca observações cujos h_t 's excedem o dobro ou o triplo da média, isto é, excedem $2p/n$ ou $3p/n$ (ver Judge et al., 1988, p. 893). A Tabela 2.1 mostra os níveis máximos de alavancagem considerados para os dois cenários analisados, incluindo também os valores limiares $2p/n$ e $3p/n$. Observa-se a nítida presença de pontos de alavanca nos dois cenários considerados.

Tabela 2.1: Análise de observações influentes, valores limiares para detectar pontos de alavancagem.

	$\mathcal{LN}(0,1)$	t_3	valor limiar	
n	$h_{\max}$	$h_{\max}$	$2p/n$	$3p/n$
25	0.418	0.472	0.16	0.24
50	0.209	0.236	0.08	0.12

As Tabelas 2.5 e 2.6 mostram as probabilidades $\Pr(\tau^2 \leq \zeta | H_0)$ calculadas para os valores de ζ anteriormente citados com os valores da constante de truncamento γ variando de 1.5 a 5.5, sendo $n = 25$ e os regressores escolhidos como realizações da distribuição t_3 (Tabela 2.5) e $\mathcal{LN}(0,1)$ (Tabela 2.6). As Tabelas 2.7 e 2.8 apresentam as mesmas probabilidades calculadas tomando $n = 50$ com regressores t_3 (Tabela 2.7) e $\mathcal{LN}(0,1)$ (Tabela 2.8). Em ambos os casos

($n = 25$ e $n = 50$), foram consideradas situações em que $\lambda = 1$ (homoscedasticidade) e $\lambda \approx 100$ (heteroscedasticidade). Foram usados $\alpha_1 = \alpha_2 = 0$ para se obter $\lambda = 1$ e, para obter $\lambda \approx 100$, foram escolhidos $\alpha_1 = \alpha_2 = 0.57$ no caso de regressores t_3 e $\alpha_1 = \alpha_2 = 0.635$ no caso de regressores lognormais.

Quando $n = 50$ (Tabelas 2.7 e 2.8), nota-se que as melhores aproximações assintóticas para as probabilidades $\Pr(\tau^2 \leq \zeta | H_0)$, considerando-se $\zeta = 2.706$ e 3.841 , ocorreram para valores de γ variando de 2.9 a 4 . Quando $n = 25$ (Tabelas 2.5 e 2.6), as melhores aproximações corresponderam a valores de γ variando de 2.8 a 4 , exceto o caso em que os regressores foram gerados da distribuição t_3 com $\lambda \approx 100$, situação em que as melhores aproximações assintóticas com $\zeta = 2.706$ corresponderam a valores de γ variando de 2.5 a 2.9 .

Observando todas as tabelas conjuntamente, percebe-se que em todos os casos os valores das probabilidades calculadas crescem à medida que γ aumenta e que valores de γ superiores a 5 e inferiores a 2 conduzem às piores aproximações, considerando-se apenas os valores $\zeta = 2.706$ e 3.841 , os quais representam os quantis 0.90 e 0.95 da distribuição nula assintótica da estatística τ^2 . Nota-se também que, para todos os casos, em linhas gerais, valores de γ no intervalo $[3, 4]$ conduziram a boas aproximações levando em consideração apenas $\zeta = 3.841$, quantil comumente usado como valor crítico do teste ao nível de significância 5% .

Seguiremos as análises numéricas considerando valores de γ iguais a 3 , 4 e 5 , uma vez que, em geral, os valores 3 e 4 conduziram a boas aproximações, tendo sido o valor 4 proposto por Cribari-Neto (2004) como valor para a constante de truncamento γ do estimador $HC4$. O valor 5 foi escolhido como limite superior para γ .

As Figuras 2.1 e 2.2 apresentam graficamente as discrepâncias relativas de probabilidade versus as correspondentes probabilidades assintóticas para os casos em que $n = 25$ (Figura 2.1) e $n = 50$ (Figura 2.2). Discrepâncias relativas de probabilidade são definidas como diferenças entre probabilidades exatas (calculadas numericamente) e as respectivas probabilidades assintóticas divididas pelas probabilidades assintóticas. Quanto mais próximas de zero estiverem as discrepâncias relativas de probabilidade, melhor será a aproximação da distribuição nula exata da estatística τ^2 pela distribuição nula limite (χ_1^2). Todos os painéis incluem uma linha de referência

horizontal indicando discrepância relativa de probabilidade nula.

Nos resultados apresentados, as estatísticas de teste usaram erros-padrão baseados no estimador *HC4* com os valores de γ indicados acima. Foram considerados erros homoscedásticos ($\lambda = 1$) e heteroscedásticos ($\lambda \approx 100$) e desenhos de regressão com pontos de alavancagem.

Notamos ao analisar as Figuras 2.1 e 2.2 que, nos casos homoscedásticos, o valor $\gamma = 3$ conduziu ao melhor desempenho ao longo da distribuição nula assintótica (χ_1^2); já nos casos heteroscedásticos ($\lambda \approx 100$), com $n = 25$ o valor $\gamma = 3$ conduziu às melhores aproximações para probabilidades assintóticas entre 0.60 e 0.90. Análise visual da Figura 2.2 revela que, para $n = 50$ e $\lambda \approx 100$, o valor $\gamma = 4$ conduz às melhores aproximações.

A Tabela 2.2 apresenta as probabilidades $\Pr(\tau^2 \leq \zeta | H_0)$ calculadas para testes baseados no estimador *HC4* considerando-se três valores de γ , a saber, 3, 4 e 5, e para valores de ζ referentes aos quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2). Foram considerados regressores cujos valores foram obtidos como realizações das variáveis aleatórias t_3 e $\mathcal{LN}(0, 1)$ com $n = 50$.

Tabela 2.2: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 50$; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas no estimador *HC4*.

λ	Pr	t_3			$\mathcal{LN}(0, 1)$		
		$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
1	0.90	0.902	0.916	0.929	0.899	0.911	0.922
	0.95	0.944	0.953	0.961	0.944	0.951	0.958
≈ 100	0.90	0.883	0.921	0.949	0.879	0.911	0.937
	0.95	0.941	0.962	0.975	0.935	0.954	0.968

Percebe-se que o estimador *HC4* com o valor $\gamma = 3$ conduziu à melhor aproximação tomando-se o quantil 0.90 nos casos homoscedásticos ($\lambda = 1$), enquanto que nos casos heteroscedásticos ($\lambda \approx 100$) os valores $\gamma = 3$ e $\gamma = 4$ conduziram a resultados muito parecidos. O estimador *HC4* com sua constante usual ($\gamma = 4$) conduziu ao melhor resultado para o quantil 0.95, exceto quando $\lambda \approx 100$ com os valores dos regressores obtidos como realizações da variável t_3 , situação em que o valor $\gamma = 3$ conduziu à melhor aproximação, mas que foi bem próxima da obtida com $\gamma = 4$.

2.3 Avaliação Numérica com Dados Reais

Seguiremos com uma avaliação numérica usando dados reais (não simulados) para os regressores do modelo. Os dados utilizados podem ser encontrados em Greene (1997, Tabela 12.1, p. 541). Aqui, a variável de interesse (y) é o gasto per capita em escolas públicas e as variáveis independentes, x e x^2 , são a renda per capita por estado em 1979 nos Estados Unidos e seu quadrado.

O modelo de regressão considerado foi

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + u_t, \quad t = 1, \dots, 50. \quad (2.2)$$

Os erros são não-correlacionados, cada u_t sendo normalmente distribuído com média zero e variância $\sigma_t^2 = \exp(\alpha_1 x_t + \alpha_2 x_t^2)$. Quando $\alpha_1 = \alpha_2 = 0$ tem-se $\lambda = 1$ (homoscedasticidade); quando $\alpha_1 = 0$ e $\alpha_2 = 4.6$, tem-se $\lambda \approx 50$ (heteroscedasticidade). Replicamos os valores das covariáveis para obter uma amostra de tamanho 100.

Nosso interesse reside em testar $H_0 : c'\beta = m$ contra $H_1 : c'\beta \neq m$. Em particular, consideraremos $m = 0$ e $c = (0, 0, 1)'$, ou seja, testamos uma especificação linear contra uma quadrática. A estatística de teste é dada por

$$\tau^2 = \frac{\hat{\beta}_2^2}{\widehat{\text{var}}(\hat{\beta}_2)},$$

onde $\widehat{\text{var}}(\hat{\beta}_2)$ é obtido do estimador *HC4* com alterações em sua constante de truncamento γ .

A Tabela 2.3 apresenta os níveis de alavancagem para dois casos considerados, a saber, com dados completos (caso 1) e sem a observação de maior alavancagem, o Alasca (caso 2). Não consideramos o caso com a remoção das três observações de maior alavancagem, pois este seria um caso praticamente sem alavancagem.

A Figura 2.3 contém os gráficos das discrepâncias relativas de probabilidade para os tamanhos de amostra $n = 50$ e $n = 100$. Foram considerados os casos em que as variâncias dos erros são iguais ($\lambda = 1$) e diferentes ($\lambda \approx 50$). Novamente, as estatísticas de teste basearam-se no estimador *HC4* com alterações na constante de truncamento γ . Os valores de γ foram os mesmos considerados anteriormente.

Tabela 2.3: Níveis de alavancagem, valores limiares para detectar pontos de alavancagem e razão entre $h_{\max}$ e $3p/n$; avaliação com dados reais.

caso	n	h_t	$2p/n$	$3p/n$	$h_{\max}/(3p/n)$
1	50	$0.651(h_{\max})$ 0.208 0.200	0.12	0.18	3.61
2	49	$0.562(h_{\max})$ 0.249	0.122	0.184	3.06

Observando-se a Figura 2.3, percebe-se que tanto quando $\lambda = 1$ (homoscedasticidade) quanto com $\lambda \approx 50$ (heteroscedasticidade), as distribuições nulas exatas de estatísticas de teste baseadas no estimador $HC4$ com constante $\gamma = 3$ foram melhores aproximadas pela distribuição nula assintótica (χ_1^2) relativamente a todas as demais distribuições nulas consideradas.

A Tabela 2.4 contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 3.841$, que representa o quantil 0.95 da distribuição nula assintótica (χ_1^2), para estatísticas de teste baseadas no estimador $HC4$ com diferentes valores de γ . Aqui, $n = 50$ e $n = 100$ (dados replicados) com $\lambda = 1$ (homoscedasticidade) e $\lambda \approx 50$ (heteroscedasticidade).

Tabela 2.4: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 50$ e 100 (dados replicados); ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC4$ com variações na contante γ .

$HC4$	$\lambda = 1$		$\lambda \approx 50$	
γ	$n = 50$	$n = 100$	$n = 50$	$n = 100$
3	0.964	0.953	0.917	0.913
4	0.978	0.965	0.952	0.940
5	0.987	0.974	0.972	0.959

Percebe-se que os valores de $\gamma = 3$ e 4 conduziram aos melhores resultados. Quando $\lambda = 1$ (homoscedasticidade), as melhores aproximações foram obtidas considerando-se $\gamma = 3$; já quando $\lambda \approx 50$ (heteroscedasticidade), as melhores aproximações ocorreram tomando-se $\gamma = 4$.

2.4 Análise Numérica com os Quantis 0.90 e 0.95 da Distribuição Nula Limite

Como próximo objetivo, nós utilizaremos integração numérica a fim de avaliar o impacto do valor da constante de truncamento γ (usualmente fixada em $\gamma = 4$) sobre a aproximação nula limite de testes baseados no estimador $HC4$. A avaliação baseia-se em modelos de regressão com 2, 4, 6 e 8 parâmetros de regressão. Os erros são não-correlacionados, cada erro tendo média zero e variância $\sigma_t^2 = \exp(\sum_{j=1}^{p-1} \alpha_j x_{t(j+1)})$, em que p é o número de parâmetros de regressão e x_{tj} é o elemento da matriz de regressores na linha t e coluna j . O tamanho da amostra é $n = 50$ tendo sido os dados replicados de uma amostra de tamanho $n = 25$ e os valores das covariáveis obtidos como realizações da distribuição lognormal padrão. Os desenhos de regressão considerados foram pontos de alavancagem: $h_{\max}/(3p/n) \approx 1.4$ quando $p = 2, 4$ e 6 , e $h_{\max}/(3p/n) \approx 1$ quando $p = 8$, em que $h_{\max}$ representa o nível máximo de alavancagem. Os resultados da avaliação numérica são mostrados graficamente na Figura 2.4, que contém os gráficos das diferenças entre $\Pr(\tau^2 \leq \zeta | H_0)$, ζ sendo o quantil 0.95 da distribuição nula limite (χ_1^2), e 0.95, a probabilidade nominal (assintótica), para os casos homoscedásticos ($\lambda = 1$) e com diferentes níveis de heteroscedasticidade, a saber, $\lambda \approx 2, 5, 20$ e 50 . As discrepâncias de probabilidade são plotadas contra os valores de γ variando de 0 a 6. Notamos que o valor $\gamma = 4$ é uma escolha bastante atraente quando o quantil 0.95 da distribuição nula limite é o de interesse.

A mesma análise numérica foi realizada considerando-se o quantil 0.90 da distribuição nula limite (χ_1^2). A avaliação pode ser vista na Figura 2.5, que contém os gráficos das diferenças entre $\Pr(\tau^2 \leq \zeta | H_0)$, ζ sendo o quantil 0.90 da distribuição nula limite (χ_1^2), e 0.90, a probabilidade nominal (assintótica), para os casos homoscedásticos ($\lambda = 1$) e com diferentes níveis de heteroscedasticidade, a saber, $\lambda \approx 2, 5, 20$ e 50 . Novamente, as discrepâncias de probabilidade são plotadas contra os valores de γ variando de 0 a 6. Embora em casos homoscedásticos o valor $\gamma = 3$ tenha conduzido às melhores aproximações, há evidência numérica de que o valor $\gamma = 4$ continua sendo uma alternativa atraente quando o quantil 0.90 da distribuição nula limite é o de interesse.

2.5 Conclusão

A partir das análises realizadas, notamos que, em geral, o valor $\gamma = 4$ da constante de truncamento do estimador *HC4* conduziu aos melhores resultados quando o grau de heteroscedasticidade foi elevado ($\lambda \approx 100$ e $\lambda \approx 50$). O valor $\gamma = 3$, sob homoscedasticidade ($\lambda = 1$), às vezes conduziu a resultados melhores do que os obtidos com $\gamma = 4$ e, sendo considerado o quantil 0.90 da distribuição nula assintótica (χ_1^2), os melhores resultados corresponderam a testes baseados no estimador *HC4* com $\gamma = 3$. Evidências numéricas nos levam a concluir que, embora algumas vezes o valor $\gamma = 3$ tenha conduzido aos melhores resultados, a escolha $\gamma = 4$ para constante de truncamento do estimador *HC4* parece ser bastante adequada.

Tabela 2.5: Probabilidades $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$; ζ igual aos quantis 0.90, 0.95 e 0.99 da distribuição χ_1^2 ; valores dos regressores obtidos como realizações da distribuição t_3 ; estatísticas $HC4$ com variações na constante de truncamento γ .

	$\lambda = 1$			$\lambda \approx 100$		
ζ	0.10	0.05	0.01	0.10	0.05	0.01
Pr	0.90	0.95	0.99	0.90	0.95	0.99
γ	pcalc1	pcalc2	pcalc3	pcalc1	pcalc2	pcalc3
1.5	0.855	0.906	0.960	0.830	0.897	0.958
1.6	0.859	0.909	0.961	0.840	0.902	0.960
1.8	0.866	0.914	0.964	0.857	0.912	0.964
1.9	0.870	0.917	0.965	0.865	0.917	0.966
2.0	0.874	0.919	0.966	0.873	0.921	0.968
2.1	0.877	0.922	0.967	0.879	0.925	0.969
2.2	0.881	0.924	0.968	0.886	0.929	0.971
2.3	0.884	0.927	0.970	0.892	0.933	0.972
2.4	0.888	0.929	0.971	0.897	0.936	0.973
2.5	0.891	0.931	0.972	0.902	0.939	0.975
2.6	0.895	0.933	0.973	0.907	0.942	0.976
2.7	0.898	0.936	0.974	0.912	0.945	0.977
2.8	0.901	0.938	0.975	0.916	0.947	0.978
2.9	0.904	0.940	0.976	0.920	0.950	0.979
3.0	0.907	0.942	0.976	0.924	0.952	0.980
3.1	0.910	0.943	0.977	0.927	0.954	0.981
3.2	0.912	0.945	0.978	0.930	0.956	0.981
3.3	0.914	0.946	0.978	0.933	0.957	0.982
3.4	0.917	0.948	0.979	0.935	0.959	0.983
3.5	0.919	0.949	0.980	0.938	0.960	0.983
3.6	0.921	0.951	0.980	0.940	0.962	0.984
3.7	0.923	0.952	0.981	0.943	0.963	0.984
3.8	0.925	0.954	0.981	0.945	0.965	0.985
3.9	0.927	0.955	0.982	0.947	0.966	0.985
4.0	0.929	0.956	0.982	0.949	0.967	0.986
4.1	0.931	0.957	0.983	0.951	0.968	0.986
4.2	0.933	0.959	0.983	0.952	0.969	0.987
4.3	0.935	0.960	0.984	0.954	0.970	0.987
4.4	0.937	0.961	0.984	0.956	0.971	0.988
4.5	0.939	0.962	0.985	0.957	0.972	0.988
4.6	0.941	0.963	0.985	0.959	0.973	0.988
4.7	0.942	0.964	0.986	0.960	0.974	0.989
4.8	0.944	0.965	0.986	0.962	0.975	0.989
4.9	0.946	0.966	0.986	0.963	0.976	0.989
5.0	0.947	0.967	0.987	0.964	0.977	0.990
5.1	0.949	0.968	0.987	0.966	0.977	0.990
5.2	0.950	0.969	0.988	0.967	0.978	0.990
5.3	0.952	0.970	0.988	0.968	0.979	0.991
5.4	0.953	0.971	0.988	0.969	0.980	0.991
5.5	0.955	0.972	0.989	0.970	0.980	0.991

Tabela 2.6: Probabilidades $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$; ζ igual aos quantis 0.90, 0.95 e 0.99 da distribuição χ^2_1 ; valores dos regressores obtidos como realizações da distribuição $\mathcal{LN}(0, 1)$; estatísticas $HC4$ com variações na constante de truncamento γ .

	$\lambda = 1$			$\lambda \approx 100$		
ζ	0.10	0.05	0.01	0.10	0.05	0.01
Pr	0.90	0.95	0.99	0.90	0.95	0.99
γ	pcalc1	pcalc2	pcalc3	pcalc1	pcalc2	pcalc3
1.5	0.869	0.920	0.971	0.819	0.885	0.949
1.6	0.871	0.922	0.972	0.827	0.890	0.951
1.8	0.875	0.924	0.973	0.841	0.898	0.954
1.9	0.878	0.926	0.974	0.847	0.902	0.956
2.0	0.880	0.927	0.974	0.854	0.906	0.957
2.1	0.882	0.929	0.975	0.860	0.910	0.959
2.2	0.884	0.930	0.975	0.865	0.913	0.960
2.3	0.886	0.932	0.976	0.871	0.917	0.962
2.4	0.888	0.933	0.976	0.876	0.920	0.963
2.5	0.891	0.934	0.977	0.881	0.923	0.964
2.6	0.893	0.936	0.977	0.885	0.925	0.966
2.7	0.895	0.937	0.978	0.890	0.928	0.967
2.8	0.897	0.939	0.978	0.894	0.931	0.968
2.9	0.899	0.940	0.979	0.898	0.933	0.969
3.0	0.901	0.941	0.979	0.901	0.935	0.970
3.1	0.903	0.942	0.980	0.905	0.938	0.971
3.2	0.905	0.944	0.980	0.908	0.940	0.972
3.3	0.907	0.945	0.981	0.912	0.942	0.973
3.4	0.909	0.946	0.981	0.915	0.944	0.973
3.5	0.911	0.947	0.982	0.918	0.946	0.974
3.6	0.913	0.949	0.982	0.921	0.947	0.975
3.7	0.915	0.950	0.982	0.923	0.949	0.976
3.8	0.917	0.951	0.983	0.926	0.951	0.977
3.9	0.919	0.952	0.983	0.928	0.952	0.977
4.0	0.921	0.953	0.984	0.931	0.954	0.978
4.1	0.923	0.954	0.984	0.933	0.955	0.979
4.2	0.924	0.956	0.984	0.935	0.957	0.979
4.3	0.926	0.957	0.985	0.937	0.958	0.980
4.4	0.928	0.958	0.985	0.939	0.959	0.980
4.5	0.930	0.959	0.986	0.941	0.960	0.981
4.6	0.931	0.960	0.986	0.943	0.962	0.981
4.7	0.933	0.961	0.986	0.945	0.963	0.982
4.8	0.935	0.962	0.987	0.946	0.964	0.983
4.9	0.936	0.963	0.987	0.948	0.965	0.983
5.0	0.938	0.963	0.987	0.950	0.966	0.983
5.1	0.939	0.964	0.988	0.951	0.967	0.984
5.2	0.941	0.965	0.988	0.953	0.968	0.984
5.4	0.941	0.966	0.988	0.953	0.968	0.985
5.5	0.941	0.966	0.988	0.953	0.968	0.985

Tabela 2.7: Probabilidades $\Pr(\tau^2 < \zeta | H_0)$ para $n = 50$; ζ igual aos quantis 0.90, 0.95 e 0.99 da distribuição χ_1^2 ; valores dos regressores obtidos como realizações da distribuição t_3 ; estatísticas *HC4* com variações na constante de truncamento γ .

	$\lambda = 1$			$\lambda \approx 100$		
ζ	0.10	0.05	0.01	0.10	0.05	0.01
Pr	0.90	0.95	0.99	0.90	0.95	0.99
γ	pcalc1	pcalc2	pcalc3	pcalc1	pcalc2	pcalc3
1.5	0.874	0.925	0.974	0.805	0.887	0.967
1.6	0.876	0.926	0.975	0.811	0.891	0.968
1.8	0.880	0.929	0.976	0.822	0.900	0.971
1.9	0.882	0.930	0.976	0.827	0.904	0.973
2.0	0.884	0.932	0.977	0.833	0.908	0.974
2.1	0.886	0.933	0.977	0.838	0.912	0.975
2.2	0.887	0.934	0.978	0.843	0.915	0.976
2.3	0.889	0.935	0.979	0.848	0.919	0.978
2.4	0.891	0.937	0.979	0.854	0.922	0.979
2.5	0.893	0.938	0.979	0.859	0.926	0.980
2.6	0.895	0.939	0.980	0.864	0.929	0.981
2.7	0.897	0.940	0.980	0.868	0.932	0.981
2.8	0.898	0.942	0.981	0.873	0.935	0.982
2.9	0.900	0.943	0.981	0.878	0.938	0.983
3.0	0.902	0.944	0.982	0.883	0.941	0.984
3.1	0.903	0.945	0.982	0.887	0.943	0.985
3.2	0.905	0.946	0.982	0.891	0.946	0.985
3.3	0.906	0.947	0.983	0.895	0.948	0.986
3.4	0.908	0.947	0.983	0.899	0.950	0.986
3.5	0.909	0.948	0.983	0.903	0.953	0.987
3.6	0.910	0.949	0.984	0.907	0.955	0.987
3.7	0.912	0.950	0.984	0.911	0.957	0.988
3.8	0.913	0.951	0.984	0.915	0.958	0.988
3.9	0.914	0.952	0.985	0.918	0.960	0.989
4.0	0.916	0.953	0.985	0.921	0.962	0.989
4.1	0.917	0.954	0.985	0.925	0.964	0.990
4.2	0.918	0.954	0.986	0.928	0.965	0.990
4.3	0.920	0.955	0.986	0.931	0.967	0.991
4.4	0.921	0.956	0.986	0.934	0.968	0.991
4.5	0.922	0.957	0.986	0.937	0.969	0.991
4.6	0.924	0.958	0.987	0.939	0.971	0.992
4.7	0.925	0.958	0.987	0.942	0.972	0.992
4.8	0.926	0.959	0.987	0.944	0.973	0.992
4.9	0.928	0.960	0.987	0.947	0.974	0.992
5.0	0.929	0.961	0.988	0.949	0.975	0.993
5.1	0.930	0.961	0.988	0.951	0.976	0.993
5.2	0.931	0.962	0.988	0.953	0.977	0.993
5.4	0.934	0.964	0.989	0.957	0.979	0.994
5.5	0.935	0.964	0.989	0.959	0.980	0.994

Tabela 2.8: Probabilidades $\Pr(\tau^2 < \zeta | H_0)$ para $n = 50$; ζ igual aos quantis 0.90, 0.95 e 0.99 da distribuição χ_1^2 ; valores dos regressores obtidos como realizações da distribuição $\mathcal{LN}(0, 1)$; estatísticas $HC4$ com variações na constante de truncamento γ .

	$\lambda = 1$			$\lambda \approx 100$		
ζ	0.10	0.05	0.01	0.10	0.05	0.01
Pr	0.90	0.95	0.99	0.90	0.95	0.99
γ	pcalc1	pcalc2	pcalc3	pcalc1	pcalc2	pcalc3
1.5	0.882	0.933	0.980	0.820	0.895	0.966
1.6	0.883	0.934	0.980	0.825	0.898	0.967
1.8	0.885	0.935	0.981	0.833	0.904	0.970
1.9	0.886	0.936	0.981	0.837	0.907	0.971
2.0	0.888	0.937	0.981	0.841	0.910	0.972
2.1	0.889	0.937	0.981	0.845	0.913	0.973
2.2	0.890	0.938	0.982	0.849	0.915	0.974
2.3	0.891	0.939	0.982	0.853	0.918	0.975
2.4	0.892	0.940	0.982	0.857	0.921	0.976
2.5	0.893	0.940	0.982	0.861	0.923	0.976
2.6	0.895	0.941	0.983	0.865	0.926	0.977
2.7	0.896	0.942	0.983	0.868	0.928	0.978
2.8	0.897	0.943	0.983	0.872	0.931	0.979
2.9	0.898	0.943	0.984	0.876	0.933	0.980
3.0	0.899	0.944	0.984	0.879	0.935	0.980
3.1	0.900	0.945	0.984	0.883	0.937	0.981
3.2	0.902	0.946	0.984	0.886	0.940	0.982
3.3	0.903	0.946	0.985	0.890	0.942	0.982
3.4	0.904	0.947	0.985	0.893	0.944	0.983
3.5	0.905	0.948	0.985	0.896	0.945	0.983
3.6	0.906	0.949	0.985	0.899	0.947	0.984
3.7	0.907	0.949	0.985	0.903	0.949	0.984
3.8	0.908	0.950	0.986	0.906	0.951	0.985
3.9	0.910	0.951	0.986	0.909	0.953	0.986
4.0	0.911	0.951	0.986	0.911	0.954	0.986
4.1	0.912	0.952	0.986	0.914	0.956	0.986
4.2	0.913	0.953	0.987	0.917	0.957	0.987
4.3	0.914	0.953	0.987	0.920	0.959	0.987
4.4	0.915	0.954	0.987	0.922	0.960	0.988
4.5	0.916	0.955	0.987	0.925	0.961	0.988
4.6	0.918	0.955	0.987	0.927	0.963	0.988
4.7	0.919	0.956	0.988	0.930	0.964	0.989
4.8	0.920	0.957	0.988	0.932	0.965	0.989
4.9	0.921	0.958	0.988	0.934	0.966	0.990
5.0	0.922	0.958	0.988	0.937	0.968	0.990
5.1	0.923	0.959	0.989	0.939	0.969	0.990
5.2	0.924	0.959	0.989	0.941	0.970	0.990
5.4	0.925	0.960	0.989	0.941	0.970	0.991
5.5	0.925	0.960	0.989	0.941	0.970	0.991

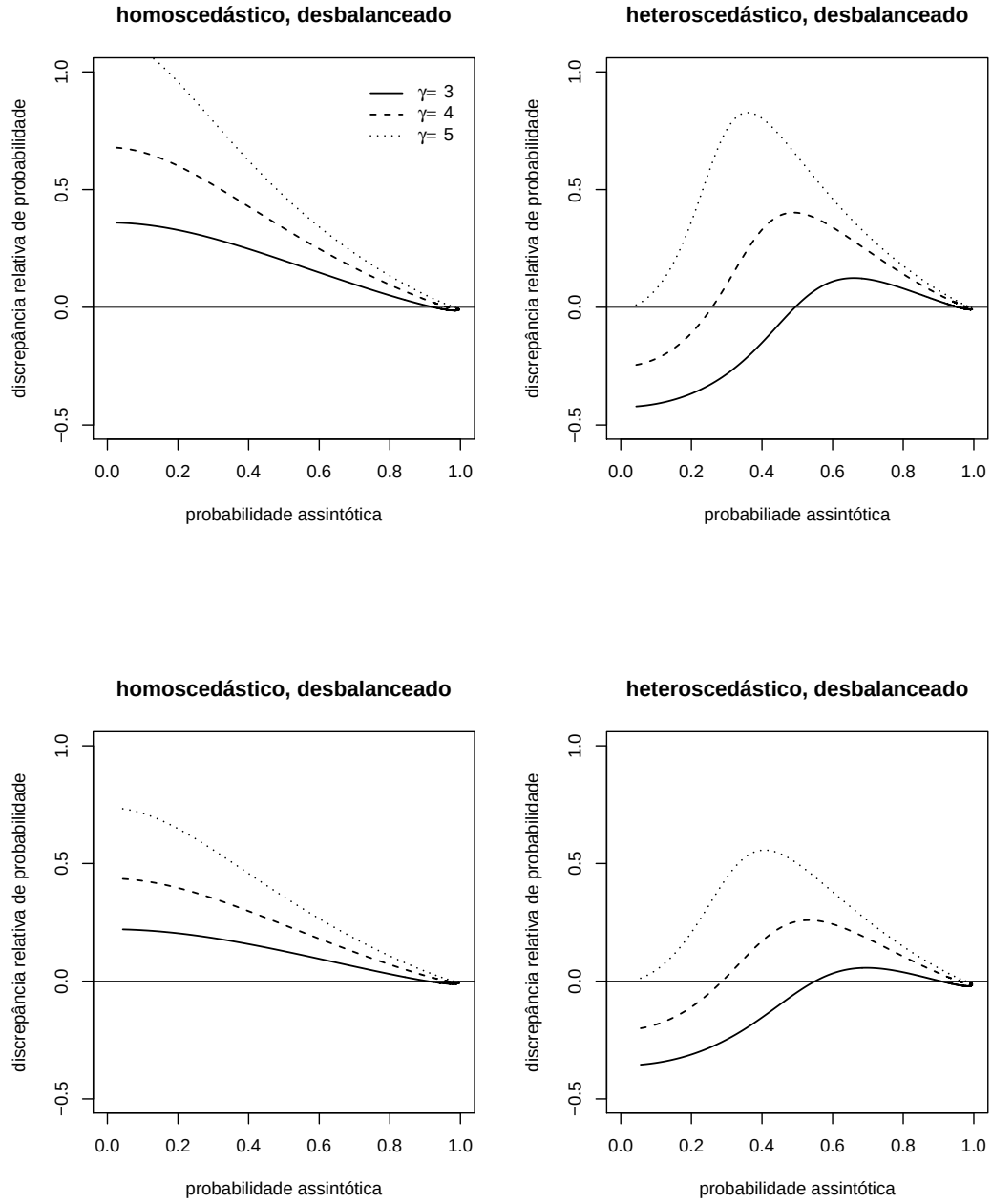


Figura 2.1: Gráficos das discrepâncias relativas de probabilidade; caso $n = 25$: valores dos regressores obtidos como realizações das distribuições t_3 (painéis superiores) e $\mathcal{LN}(0, 1)$ (painéis inferiores).

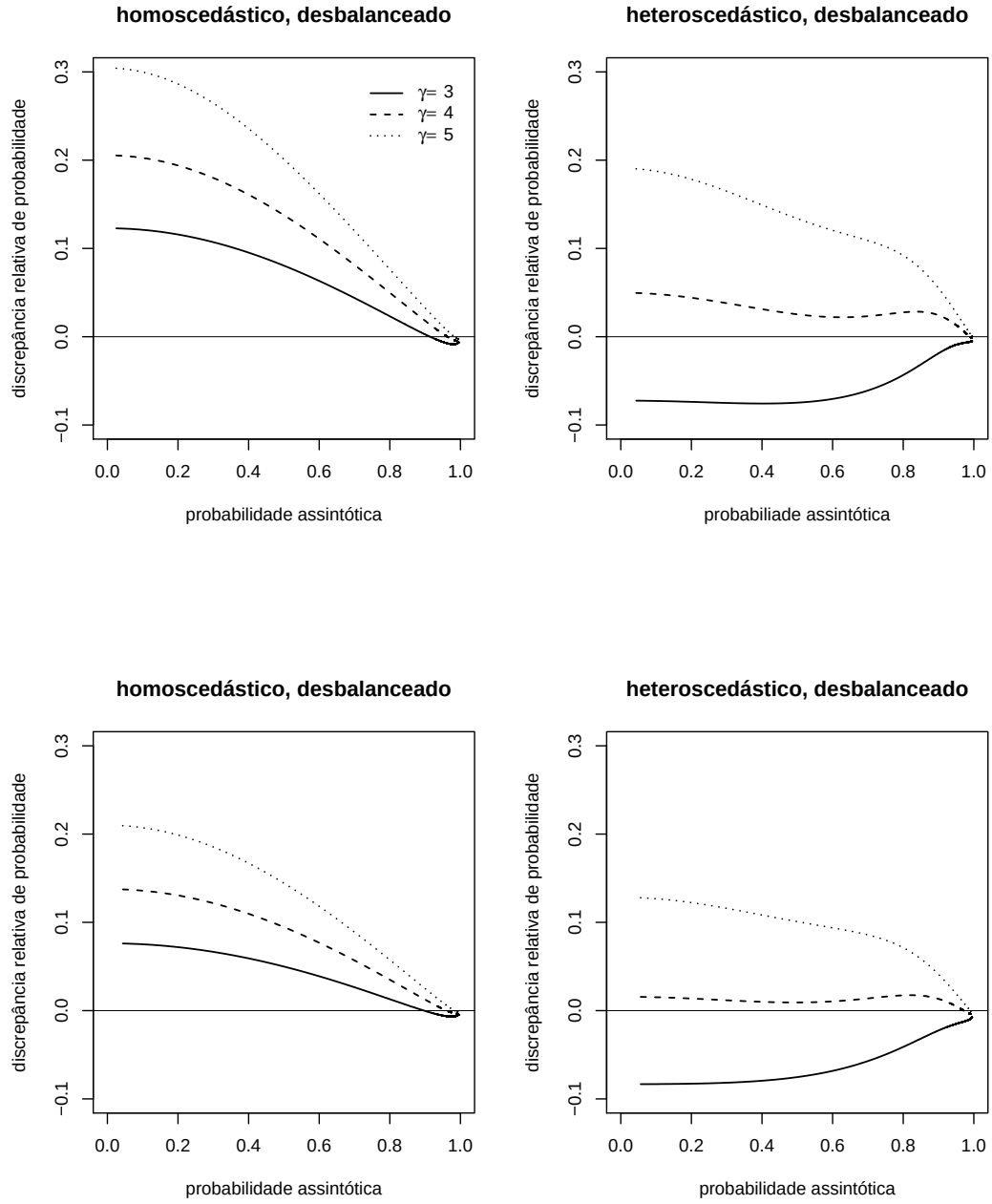


Figura 2.2: Gráficos das discrepâncias relativas de probabilidade; caso $n = 50$: valores dos regressores obtidos como realizações das distribuições t_3 (painéis superiores) e $\mathcal{LN}(0, 1)$ (painéis inferiores).

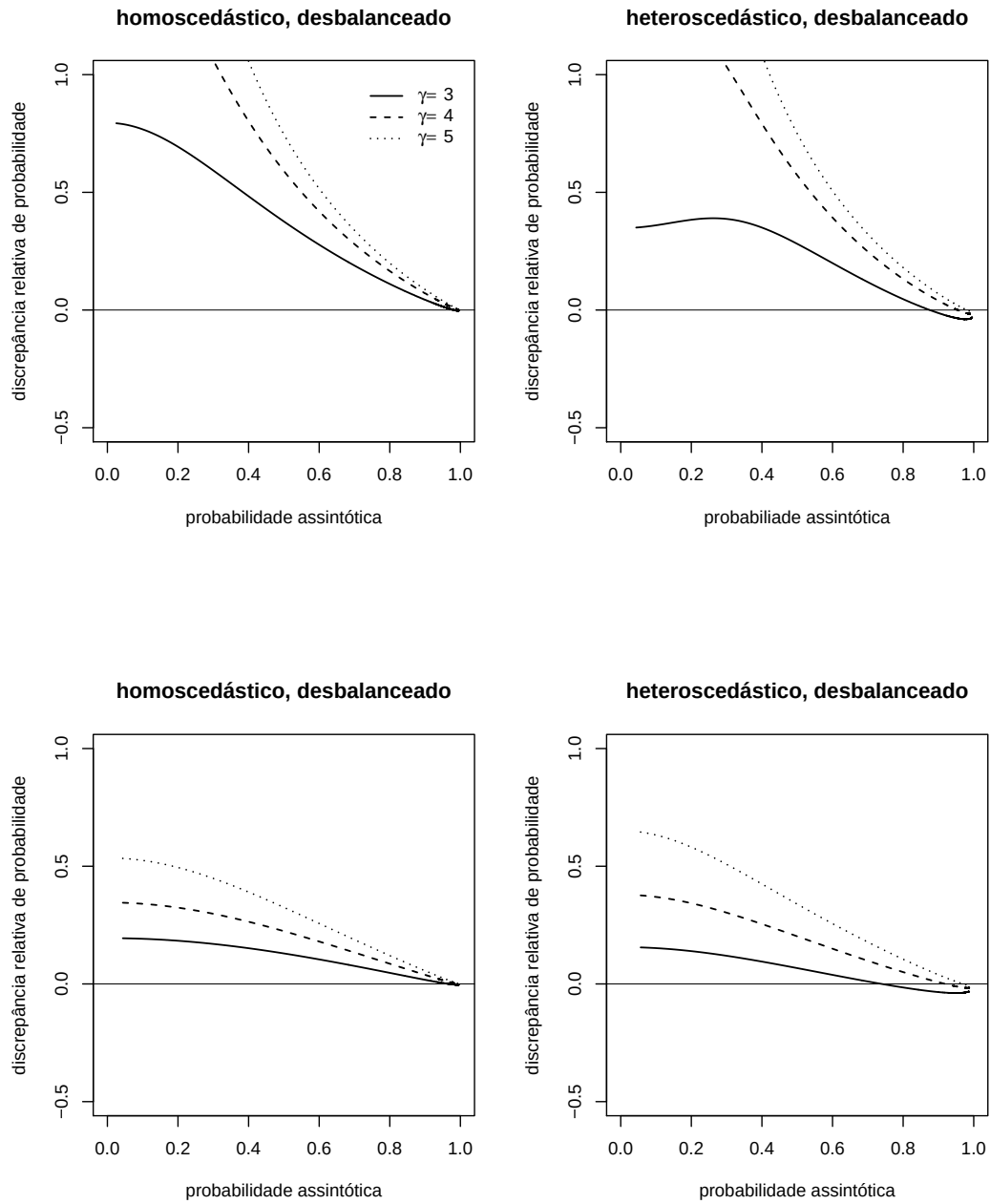


Figura 2.3: Gráficos das discrepâncias relativas de probabilidade para dados de educação; nos painéis superiores, o caso $n = 50$, nos painéis inferiores, o caso $n = 100$.

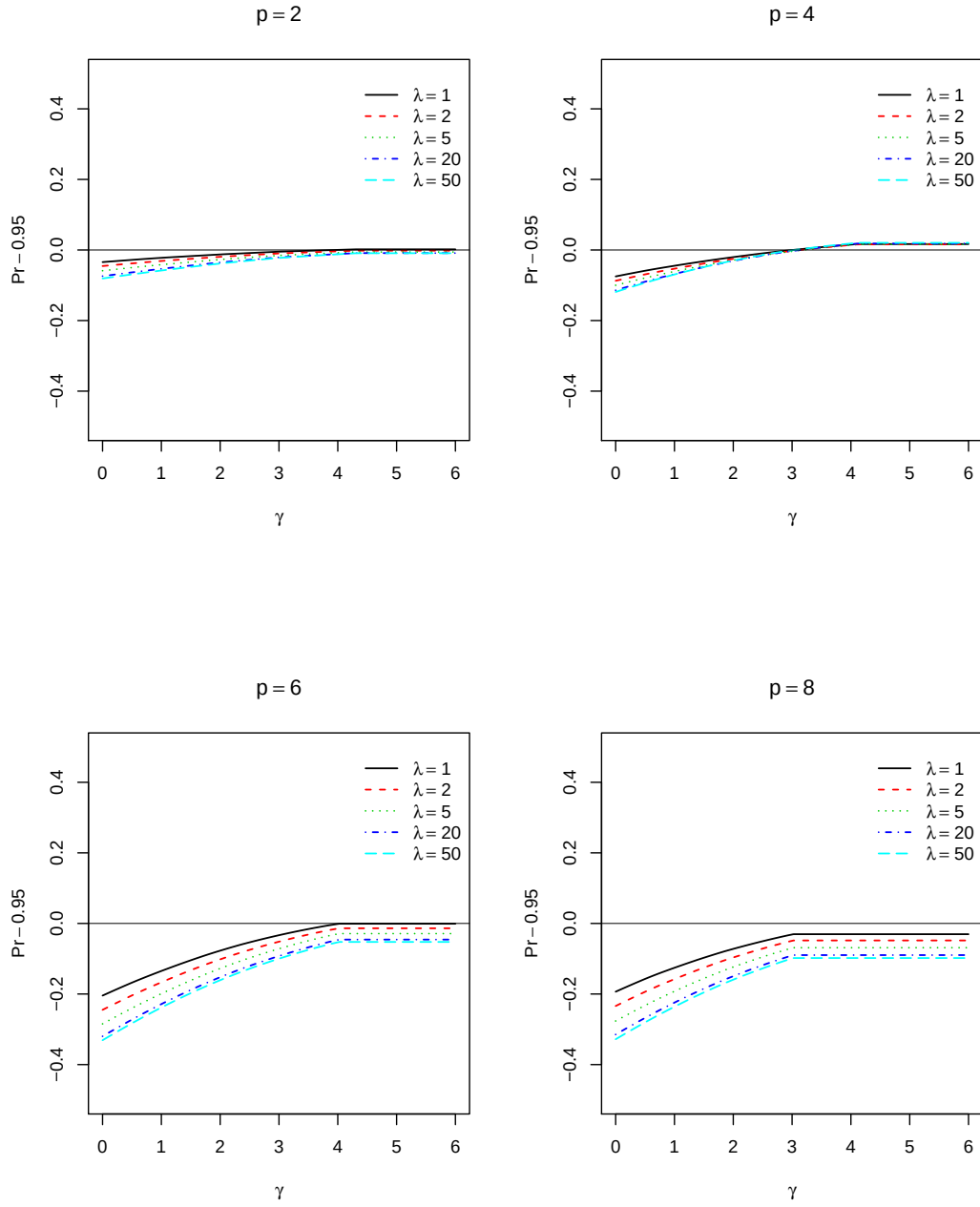


Figura 2.4: Discrepâncias de probabilidade utilizando o quantil 0.95 da distribuição nula limite (χ_1^2) de estatísticas quase- t baseadas no estimador $HC4$.

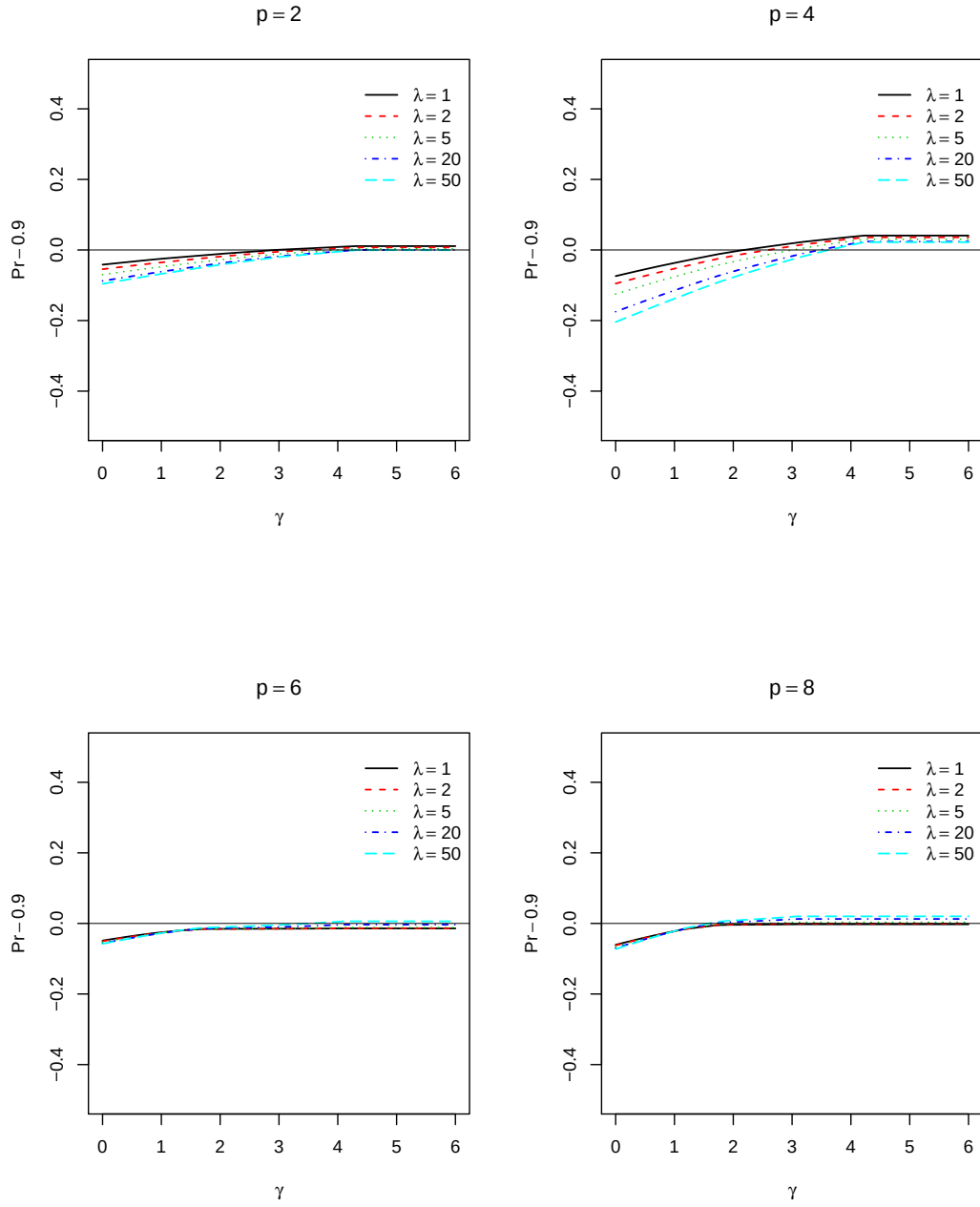


Figura 2.5: Discrepâncias de probabilidade utilizando o quantil 0.90 da distribuição nula limite (χ_1^2) de estatísticas quase- t baseadas no estimador $HC4$.

3.1 Introdução

O presente capítulo tem por objetivo apresentar e avaliar numericamente uma versão modificada do estimador *HC4*. Esta nova versão será denotada por *HC4m*. A motivação em propor esta nova versão veio do fato de que as distribuições nulas exatas de estatísticas quase-*t HC4*, embora sejam em geral bem aproximadas pela distribuição nula limite (χ_1^2) nas caudas, são muitas vezes pobremente aproximadas no restante do suporte da distribuição. O estimador proposto, segundo avaliações numéricas realizadas, torna-se bastante competitivo com o estimador *HC3* no que tange aos desempenhos de testes associados. As avaliações numéricas realizadas basearam-se no comportamento em amostras finitas de estatísticas quase-*t* ao quadrado construídas a partir de diferentes estimadores da matriz de covariâncias $\text{cov}(\hat{\beta})$, e mais uma vez foram realizadas com base no algoritmo de Imhof (1961). Os estimadores aqui considerados foram *HC3*, *HC4* e o estimador proposto (*HC4m*).

Nas avaliações, comparamos probabilidades nominais (assintóticas) com probabilidades nulas exatas calculadas por integração numérica. Consideramos diversos cenários de regressão com pontos de alavancagem. Na próxima seção deste capítulo motivamos a nossa proposta de alteração no estimador *HC4*; em seguida, mostramos a forma do novo estimador e comentamos sobre o porquê desta forma. Por fim, análises numéricas foram feitas usando dados simulados

e não-simulados (dados reais). Os dados reais escolhidos são dados sobre despesas per capita em escolas públicas e renda per capita por estado nos Estados Unidos em 1979, dados esses que foram extraídos de Greene (1997).

3.2 Uma Motivação para a Nossa Proposta

O estimador $HC4$, proposto por Cribari-Neto (2004), tem tido boa aceitação no que tange à estimação consistente da matriz $\text{cov}(\hat{\beta})$ quando em presença de pontos de alta alavancagem. Nosso interesse aqui reside em evidenciar numericamente que as distribuições nulas exatas de estatísticas quase- t $HC4$, embora sejam em geral bem aproximadas na cauda pela distribuição nula limite (χ_1^2), conforme pudemos notar nas análises numéricas realizadas no capítulo anterior, parecem ser pobremente aproximadas ao longo do restante do suporte da distribuição. Com o aumento do número de parâmetros, essa deterioração parece ocorrer também na cauda da distribuição. As avaliações numéricas aqui realizadas utilizaram o algoritmo de integração numérica de Imhof (1961).

O modelo aqui escolhido para análise foi o modelo de regressão

$$y_t = \beta_0 + \beta_1 x_{t2} + \beta_2 x_{t3} + \beta_3 x_{t4} + \beta_4 x_{t5} + u_t, \quad t = 1, \dots, n, \quad (3.1)$$

em que x_{tj} representa o elemento da matriz de regressores na linha t e coluna j . Aqui, u_t é normalmente distribuído com média zero e variância $\sigma_t^2 = \exp(\alpha_1 x_{t2} + \alpha_2 x_{t3} + \alpha_3 x_{t4} + \alpha_4 x_{t5})$, α_i , $i = 1, \dots, 4$, sendo uma constante real. Assumimos também que $\text{cov}(u_t, u_s) = 0 \forall t \neq s$ (erros não-correlacionados). Novamente, foi usada como medida de heteroscedasticidade a razão

$$\lambda = \frac{\max \sigma_t^2}{\min \sigma_t^2}. \quad (3.2)$$

Nosso interesse residirá no teste da hipótese nula $H_0 : c'\beta = m$ versus $H_1 : c'\beta \neq m$, onde c é um vetor $p \times 1$ e m é um valor real. Na análise a seguir, consideramos $m = 0$ e $c = (0, 1, 0, 0)'$. Sendo assim, estaremos interessados em testar $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Ressaltamos que nas análises apresentadas nas seções seguintes consideraremos o mesmo teste, alterando apenas

a dimensão do vetor β . A estatística de teste é

$$\tau^2 = \frac{\hat{\beta}_1^2}{\widehat{\text{var}}(\hat{\beta}_1)}, \quad (3.3)$$

onde $\widehat{\text{var}}(\hat{\beta}_1)$ é baseado no estimador *HC4*. Os valores dos regressores foram escolhidos como realizações da distribuição t_3 e o tamanho da amostra foi $n = 25$. Aqui, $h_{\max}/(3p/n) \approx 1.4$.

A Figura 3.1 apresenta um gráfico das discrepâncias relativas de probabilidade sendo a estatística de teste considerada baseada no estimador *HC4* com sua constante de truncamento usual ($\gamma = 4$), sob heteroscedasticidade ($\lambda \approx 100$) e com desenho de regressão com pontos de alavancagem. Percebe-se a pobre aproximação pela distribuição nula limite longe da cauda.

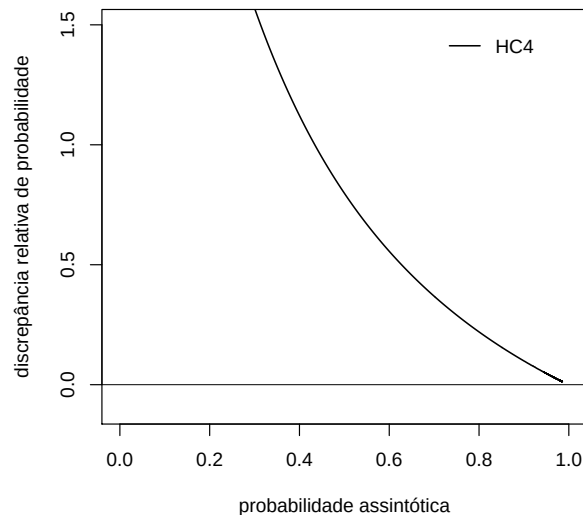


Figura 3.1: Discrepâncias relativas de probabilidade, $n = 25$: valores dos regressores obtidos como realizações da variável aleatória t_3 e estatística *HC4*.

A Tabela 3.1 mostra as probabilidades calculadas considerando-se os quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2) para o caso discutido nessa seção. Percebe-se neste caso que as aproximações obtidas a partir da distribuição nula limite foram pobres.

Conforme podemos observar na Figura 3.1 e na Tabela 3.1, tanto na cauda como ao longo da distribuição as aproximações obtidas da distribuição nula limite foram pobres, mas longe da cauda da distribuição as aproximações se deterioraram mais marcadamente.

Tabela 3.1: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas no estimador $HC4$.

Pr	$HC4$
0.90	0.989
0.95	0.994

Seguiremos com uma proposta de alteração no estimador $HC4$ a fim de obter resultados mais confiáveis ao longo de todo o suporte da distribuição nula limite de estatísticas quase- t .

3.3 Uma Nova Proposta de Estimador

Com o intuito de melhorar a confiabilidade de inferências realizadas a partir do estimador $HC4$, propomos a seguir uma nova versão deste estimador, denotada por $HC4m$. A modificação foi feita apenas na estrutura do expoente do termo $(1 - h_t)$, utilizado como fator de inflação dos resíduos ao quadrado no estimador $HC4$. No estimador aqui proposto usa-se a estrutura

$$\delta_t = \min\{\gamma_1, nh_t/p\} + \min\{\gamma_2, nh_t/p\}, \quad (3.4)$$

onde γ_1 e γ_2 são constantes reais positivas pré-especificadas. Sendo assim, o novo estimador é dado por

$$HC4m = B\hat{\Omega}_{4m}B' = BE_{4m}\hat{\Omega}B',$$

em que $E_{4m} = \text{diag}\{1/(1 - h_t)^{\delta_t}\}$, sendo δ_t conforme estabelecido em (3.4), $B = (X'X)^{-1}X'$ e $\hat{\Omega} = \text{diag}\{\hat{u}_1^2, \dots, \hat{u}_n^2\}$.

É importante lembrar que, sendo $\bar{h}$ a média dos h_t 's, temos $\bar{h} = p/n$. O expoente δ_t controla o nível de desconto aplicado ao quadrado do t -ésimo resíduo. Como no estimador $HC4$, o desconto aplicado dependerá do valor de cada h_t relativamente à média ($\bar{h}$). Considerando os casos em que $\gamma_1 \neq \gamma_2$, os valores possíveis para δ_t são: $2h_t/\bar{h}$, se $0 < h_t/\bar{h} \leq \min\{\gamma_1, \gamma_2\}$; $\min\{\gamma_1, \gamma_2\} + h_t/\bar{h}$, se $\min\{\gamma_1, \gamma_2\} < h_t/\bar{h} < \max\{\gamma_1, \gamma_2\}$ e $\gamma_1 + \gamma_2$ se $h_t/\bar{h} \geq \max\{\gamma_1, \gamma_2\}$. Se $\gamma_1 = \gamma_2$, então $\delta_t = 2\min\{\gamma_1, h_t/\bar{h}\}$. São de maior interesse os casos em que os γ_i 's são distintos, pois nesses

casos há uma alteração nos descontos para valores de $h_t/\bar{h}$ em um dado intervalo, conforme podemos observar com os possíveis valores que δ_t pode assumir.

Análises numéricas realizadas a partir de diversos desenhos de regressão com pontos de alavancagem indicaram os valores $\gamma_1 = 1.0$ e $\gamma_2 = 1.5$ como escolhas atraentes. Para o referido caso, os valores dos δ_t 's são truncados em 2.5, o que representa uma ‘tolerância’ de 25% em relação ao valor 2, usado no estimador *HC3*, e uma redução de aproximadamente 40% na constante de truncamento usada no estimador *HC4*. A Figura 3.2 apresenta o gráfico de δ_t versus a razão entre a t -ésima medida de alavancagem (h_t) e o nível médio de alavancagem para os estimadores *HC3*, *HC4* e *HC4m*, quando $\gamma_1 = 1.0$ e $\gamma_2 = 1.5$. Percebe-se uma inclinação de reta mais íngreme que a inclinação usada para o estimador *HC4* para valores de $h_t/\bar{h}$ inferiores a 1.0, o que resulta em maiores valores para os expoentes de cada termo $(1 - h_t)$, visto que nesses casos tem-se $\delta_t = 2h_t/\bar{h}$. Observando a Figura 3.2, podemos justificar a escolha adotada para a estrutura dos δ_t 's, dada na equação (3.4): essa estrutura possibilita um aumento nos níveis de desconto dos quadrados dos resíduos na região de alavancagens menos elevadas; em compensação há uma redução nos níveis de desconto na região de alavancagens mais elevadas. Adicionalmente, a fim de tentamos ‘equilibrar’ os níveis de desconto presentes nas estruturas dos estimadores *HC3* e *HC4*, é interessante que os valores dos δ_t 's sejam truncados em um valor superior ao usado no estimador *HC3* e inferior ao usado no estimador *HC4*, o que de fato ocorre.

3.4 Avaliação Numérica

Seguiremos com uma avaliação numérica baseada em diversos desenhos de regressão com pontos de alavancagem. Nosso objetivo é evidenciar numericamente que a nossa proposta de modificação no estimador *HC4* é atraente no sentido de que as distribuições nulas exatas de estatísticas de teste baseadas nesse novo estimador são melhores aproximadas pela distribuição nula limite (χ_1^2) ao longo de todo o seu suporte relativamente a estatísticas *HC4*. Além disso, quando aumentamos o número de regressores, percebemos que o teste baseado no novo estimador torna-se competitivo ao baseado no estimador *HC3*.

Na avaliação numérica realizada utilizamos o algoritmo de Imhof (1961) e comparamos as

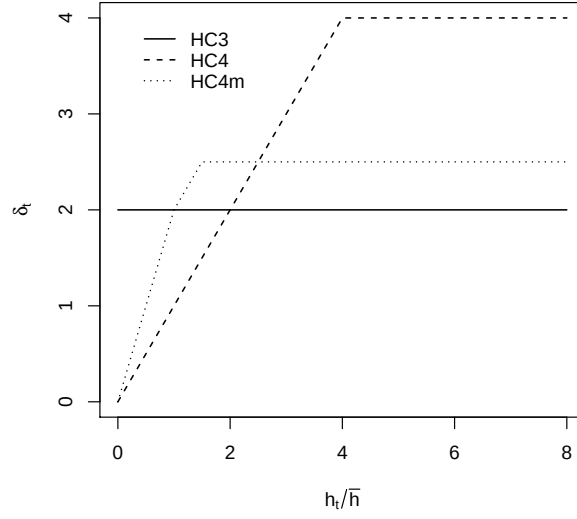


Figura 3.2: Gráfico de δ_t versus a razão entre a t -ésima medida de alavancagem (h_t) e o nível médio de alavancagem para os estimadores $HC3$, $HC4$ e $HC4m$, quando $\gamma_1 = 1.0$ e $\gamma_2 = 1.5$.

probabilidades nominais (referentes à distribuição assintótica) e as probabilidades exatas calculadas. As comparações foram feitas observando-se as discrepâncias relativas de probabilidades, como definidas na Seção 2.2.1.

Inicialmente, consideraremos o modelo de regressão

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, 2, \dots, n. \quad (3.5)$$

Os erros u_t são normalmente distribuídos com média zero e variância $\sigma_t^2 = \exp(\alpha x_t)$, α sendo uma constante real. Novamente, os erros não-correlacionados (i.e., $\text{cov}(u_t, u_s) = 0 \ \forall t \neq s$). Como medida do grau de heteroscedasticidade usaremos a medida λ dada em (3.2). Usamos $\alpha = 0$ e $\alpha = 0.68$ para obter $\lambda = 1$ e $\lambda \approx 100$, respectivamente. Os valores dos regressores foram escolhidos como realizações da distribuição lognormal padrão e o tamanho de amostra considerado foi $n = 25$; replicamos os dados a fim de obter amostras de tamanho 50. Em nossa avaliação, $h_{\max}/(3p/n) \approx 2.4$, onde $h_{\max}$ representa o nível máximo de alavancagem.

A Figura 3.3 contém gráficos de discrepâncias relativas de probabilidade. As discrepâncias relativas foram obtidas a partir de estatísticas quase- t ao quadrado baseadas nos estimadores

$HC3$, $HC4$ e $HC4m$. Percebe-se que quando $n = 25$ as aproximações obtidas a partir do estimador $HC4$ foram as mais pobres ao longo da distribuição nula limite (χ_1^2), o problema sendo mais severo no caso homoscedástico ($\lambda = 1$).

A Tabela 3.2 contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 , quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2), considerando-se estatísticas de teste baseadas nos estimadores $HC3$, $HC4$ e $HC4m$. Aqui, consideramos $n = 25$ e $n = 50$ (dados replicados) com $\lambda = 1$ (homoscedasticidade) e $\lambda \approx 100$ (heteroscedasticidade).

Tabela 3.2: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$ e 50 com $p = 2$; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC3$ e $HC4$, $HC4m$.

λ	Pr	$n = 25$			$n = 50$		
		$HC3$	$HC4$	$HC4m$	$HC3$	$HC4$	$HC4m$
1	0.90	0.875	0.941	0.897	0.880	0.920	0.892
	0.95	0.917	0.961	0.932	0.927	0.953	0.935
≈ 100	0.90	0.818	0.936	0.866	0.805	0.905	0.834
	0.95	0.869	0.952	0.902	0.872	0.944	0.895

Percebe-se que quando $\lambda = 1$ o estimador $HC4m$ conduziu a resultados atraentes para os dois quantis considerados. Já quando $\lambda \approx 100$, os melhores resultados foram obtidos utilizando o estimador $HC4$. É importante frisar que em todos os casos o novo estimador foi superior ao estimador $HC3$ no que tange à precisão da aproximação da distribuição nula exata pela distribuição limite nos quantis considerados.

Seguiremos com nossas análises agora considerando modelos de três e cinco parâmetros associados ao modelo inicialmente considerado dado na equação (3.5). Novamente, os valores dos regressores foram escolhidos como realizações da distribuição lognormal padrão. Sendo p o número de parâmetros do modelo, a variância de u_t é dada por $\sigma_t^2 = \exp(\sum_{j=1}^{p-1} \alpha_j x_{t(j+1)})$, em que x_{tj} representa o elemento da matriz dos regressores na linha t e coluna j . Quando $p = 3$, foram escolhidos os valores $\alpha_1 = \alpha_2 = 0$ para os casos homoscedásticos ($\lambda = 1$) e $\alpha_1 = \alpha_2 = 0.198$ para os casos heteroscedásticos ($\lambda \approx 100$); já quando $p = 5$, foram escolhidos os valores $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ para os casos homoscedásticos e $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.36$ para os casos heteroscedásticos. As Figuras 3.4 e 3.5 mostram gráficos similares aos apresentados na Figura 3.3 para os casos agora

sendo considerados com três (Figura 3.4) e cinco (Figura 3.5) parâmetros de regressão. Na Figura 3.4 mostramos resultados relativos a $n = 25$ e $n = 50$ enquanto que na Figura 3.5 apresentamos apenas resultados obtidos com $n = 25$. Aqui, $h_{\max}/(3p/n) \approx 2$ quando $p = 3$ e $h_{\max}/(3p/n) \approx 1.4$ quando $p = 5$.

É importante observar como as aproximações para as distribuições nulas exatas de estatísticas *HC4* deterioram ao longo do suporte da distribuição, enquanto que estatísticas baseadas nos estimadores *HC3* e *HC4m* se mostraram mais confiáveis nesse aspecto.

A Tabela 3.3 apresenta as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 , quantis assintóticos mais usados como valores críticos do teste, sendo $p = 3, 5$ e $n = 25$.

Tabela 3.3: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$ com $p = 3$ e 5 ; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores *HC3* e *HC4*, *HC4m*.

λ	Pr	$p = 3$			$p = 5$		
		<i>HC3</i>	<i>HC4</i>	<i>HC4m</i>	<i>HC3</i>	<i>HC4</i>	<i>HC4m</i>
1	0.90	0.890	0.973	0.918	0.908	0.971	0.921
	0.95	0.930	0.984	0.949	0.948	0.985	0.957
≈ 100	0.90	0.882	0.985	0.928	0.940	0.989	0.960
	0.95	0.925	0.991	0.956	0.972	0.995	0.982

Observamos o fraco desempenho do teste *HC4* e uma competitividade atraente entre os resultados obtidos através do uso dos estimadores *HC3* e *HC4m*; neste caso, ambos os estimadores conduziram a resultados muito parecidos. Nos cenários aqui analisados, observamos como o aumento no número de regressores parece deteriorar a confiabilidade de testes baseados no estimador *HC4*.

Proseguiremos com uma última análise utilizando dados simulados (gerados computacionalmente). Desta vez obtemos os valores dos regressores como realizações da distribuição χ_2^2 e o número de parâmetros de regressão (p) é igual a 3 e 5. Quando $p = 3$, foram escolhidos os valores $\alpha_1 = \alpha_2 = 0$ para os casos homoscedásticos ($\lambda = 1$) e $\alpha_1 = \alpha_2 = 0.63$ para os casos heteroscedásticos ($\lambda \approx 100$); já quando $p = 5$, foram escolhidos os valores $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ para os casos homoscedásticos e $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.206$ para os casos heteroscedásticos. A Figura 3.6 contém gráficos das discrepâncias relativas de probabilidade para os casos agora analisados

com tamanho amostral $n = 25$. Novamente, nota-se um pobre desempenho do teste $HC4$ ao longo do suporte da distribuição nula limite (χ_1^2). Esse pobre desempenho foi mais severo quando $p = 3$. Vale destacar a boa atuação do teste $HC4m$ no caso heteroscedástico ($\lambda \approx 100$) com $p = 3$.

Tabela 3.4: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$ com $p = 3$ e 5; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC3$ e $HC4$, $HC4m$.

λ	Pr	$p = 3$			$p = 5$		
		$HC3$	$HC4$	$HC4m$	$HC3$	$HC4$	$HC4m$
1	0.90	0.889	0.955	0.911	0.897	0.939	0.916
	0.95	0.929	0.971	0.943	0.937	0.964	0.949
≈ 100	0.90	0.811	0.937	0.861	0.876	0.947	0.904
	0.95	0.863	0.953	0.899	0.918	0.967	0.938

A Tabela 3.4 mostra as probabilidades calculadas para os quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2) para as estatísticas de teste consideradas. Nesses casos, em linhas gerais, o estimador aqui proposto conduziu aos melhores resultados para os dois quantis considerados, exceto no caso heteroscedástico ($\lambda \approx 100$) com $p = 3$, em que o estimador $HC4$ conduziu às melhores aproximações.

3.5 Análise Numérica com Dados Reais

Proseguiremos nossa análise utilizando dados reais (não simulados). O conjunto de dados utilizado é o mesmo do capítulo anterior e consiste de dados sobre despesas per capita em escolas públicas e renda per capita por estado nos Estados Unidos em 1979; esses dados foram extraídos de Greene (1997). O modelo de regressão é dado na equação (2.2) da Seção 2.3. Levaremos em consideração o mesmo teste abordado na referida seção e consideraremos também estatísticas de teste baseadas nos estimadores $HC3$ e $HC4m$. Na análise aqui realizada utilizaremos o estimador $HC4$ com sua constante de truncamento usual ($\gamma = 4$).

A Figura 3.7 contém os gráficos das discrepâncias relativas de probabilidade sendo consideradas estatísticas de teste baseadas nos estimadores indicados acima. Nos painéis inferiores tem-se o caso com os dados replicados ($n = 100$). Notamos que nos casos heteroscedásticos

($\lambda \approx 50$) o estimador proposto ($HC4m$) conduziu aos melhores resultados ao longo do suporte da distribuição nula limite (χ_1^2) enquanto que o estimador $HC4$ conduziu a resultados pobres quando $n = 50$ (dados não replicados).

A Tabela 3.5 contém probabilidades calculadas considerando-se os quantis 0.90 e 0.95 da distribuição nula limite. Percebe-se, em linhas gerais, que estatísticas de teste baseadas no estimador $HC4m$ conduziram aos melhores resultados nos casos homoscedásticos ($\lambda = 1$) enquanto que estatísticas de teste baseadas no estimador $HC4$ conduziram aos melhores resultados nos casos heteroscedásticos ($\lambda \approx 50$).

Tabela 3.5: $\Pr(\tau^2 < \zeta | H_0)$ usando dados sobre gastos em educação, $n = 50, 100$ (dados replicados); ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC3$ e $HC4$, $HC4m$.

λ	Pr	$n = 50$			$n = 100$		
		$HC3$	$HC4$	$HC4m$	$HC3$	$HC4$	$HC4m$
1	0.90	0.901	0.963	0.922	0.894	0.934	0.905
	0.95	0.941	0.978	0.954	0.940	0.965	0.947
≈ 50	0.90	0.805	0.936	0.852	0.822	0.906	0.846
	0.95	0.854	0.952	0.890	0.877	0.940	0.896

3.6 Conclusão

Os resultados numéricos apresentados nesse capítulo evidenciam como o aumento no número de regressores do modelo de regressão pode deteriorar as aproximações usadas no teste $HC4$, especialmente longe da região caudal em se tratando de dados com pontos de alavancagem. Destacamos também que o novo estimador proposto, $HC4m$, procura corrigir tal deterioração tornando-se competitivo com os estimadores $HC4$ e $HC3$ na região caudal.

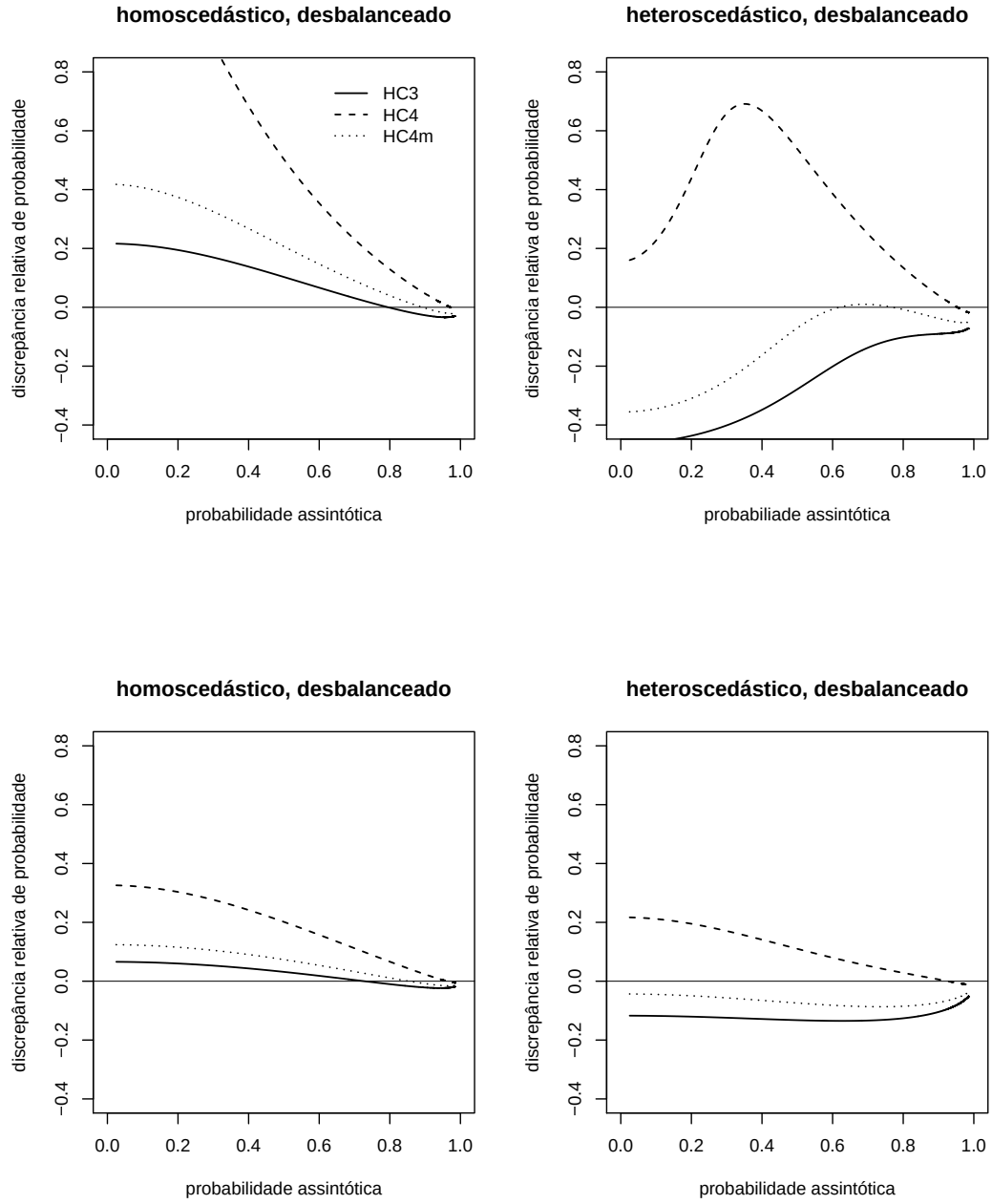


Figura 3.3: Gráficos de discrepâncias relativas de probabilidade quando os valores dos regressores vêm de realizações da distribuição $\mathcal{LN}(0, 1)$ com $p = 2$, $n = 25$ (painéis superiores) e $n = 50$ (painéis inferiores).

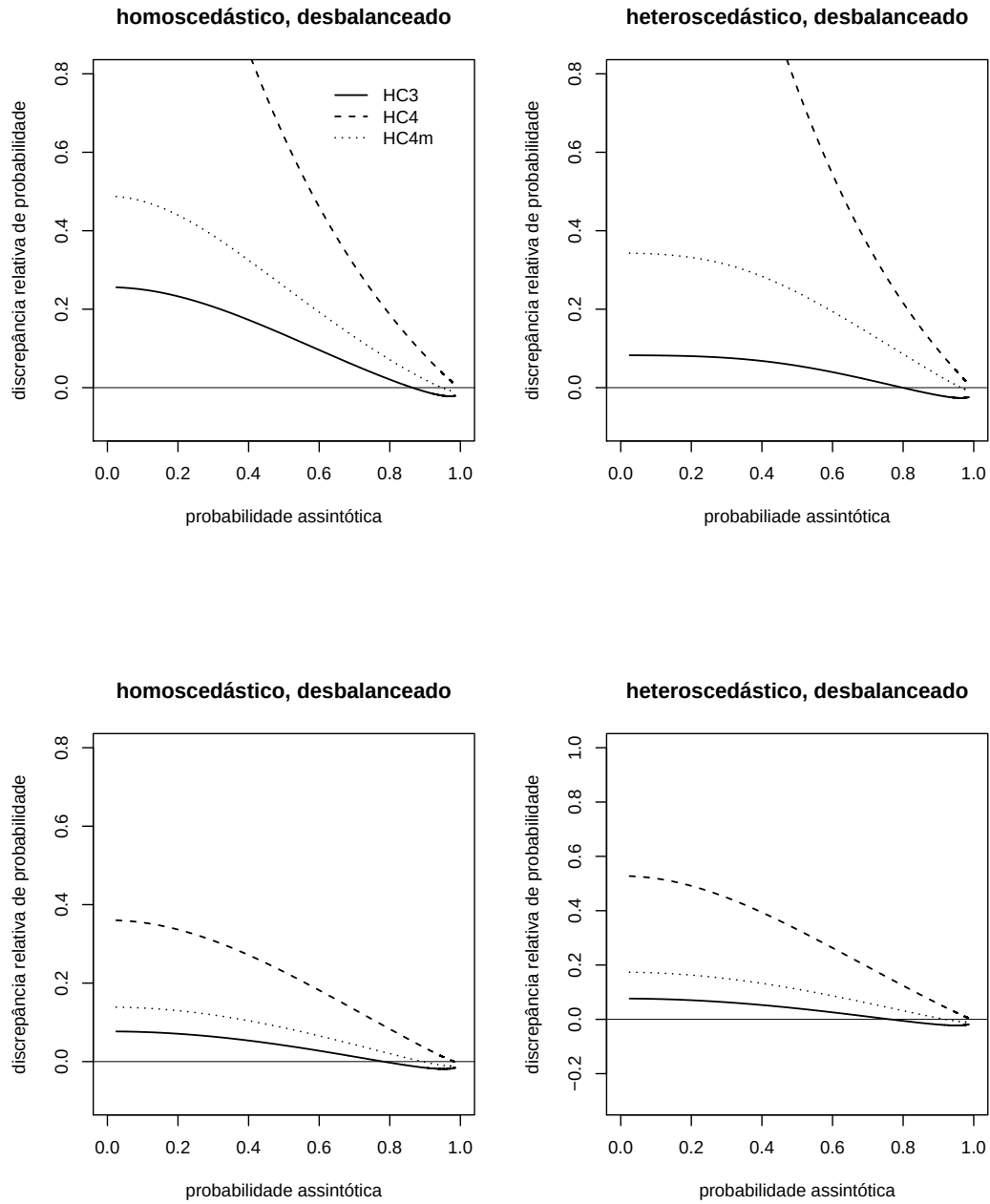


Figura 3.4: Gráficos de discrepâncias relativas de probabilidade quando os valores dos regressores vêm de realizações da distribuição $\mathcal{LN}(0, 1)$ com $p = 3$, $n = 25$ (painéis superiores) e $n = 50$ (painéis inferiores).

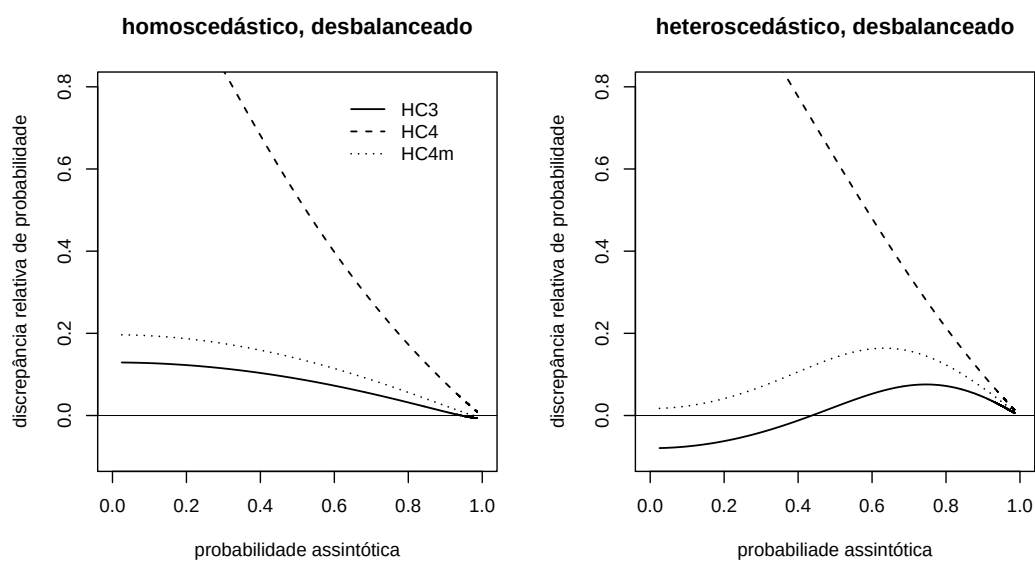


Figura 3.5: Gráficos de discrepâncias relativas de probabilidade quando os valores dos regressores vêm de realizações da distribuição $\mathcal{LN}(0, 1)$ com $p = 5$, $n = 25$.

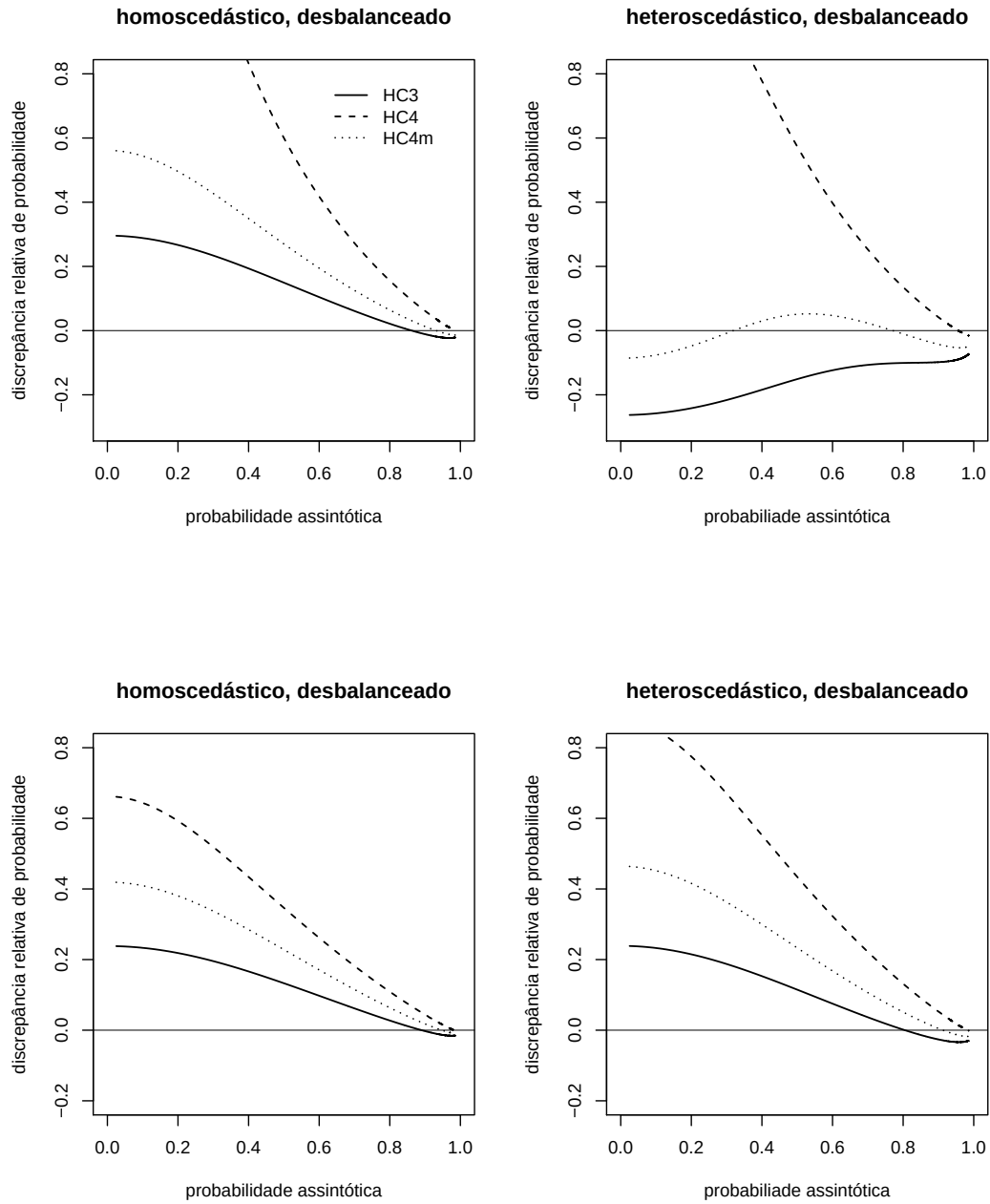


Figura 3.6: Gráficos de discrepâncias relativas de probabilidade quando os valores dos regressores vêm de realizações da distribuição χ^2_2 com $n = 25$, $p = 3$ (painéis superiores) e $p = 5$ (painéis inferiores).

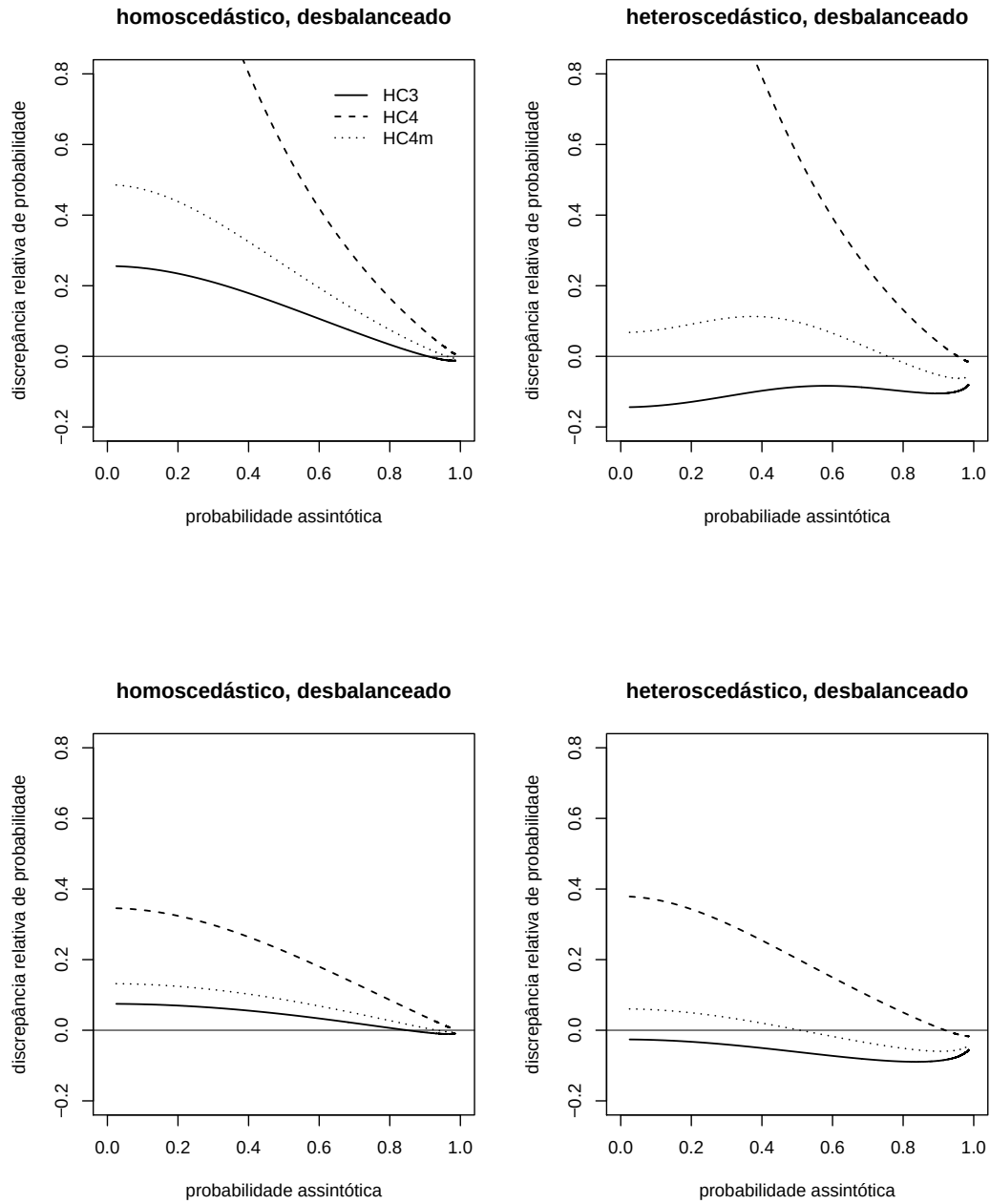


Figura 3.7: Gráficos de discrepâncias relativas de probabilidade usando dados de gastos em educação para os regressores: $n = 50$ (painéis superiores) e $n = 100$ (painéis inferiores).

4.1 Introdução

O presente capítulo tem por objetivo propor uma modificação do estimador $HC5$ (Cribari-Neto, Souza e Vasconcellos, 2007), à luz do que foi proposto com base no estimador $HC4$ no capítulo anterior. Objetivamos, ainda, realizar uma avaliação numérica exata dos testes associados. As avaliações numéricas mostram que o problema que atinge o estimador $HC4$ também pode atingir o estimador $HC5$, porém com menor intensidade, ou seja, distribuições nulas exatas de estatísticas quase- t $HC5$ podem ser pobremente aproximadas pela distribuição nula assintótica (χ_1^2) fora da cauda da distribuição. A versão modificada do estimador $HC5$ procura atenuar tal deficiência.

4.2 Uma Versão Modificada do Estimador $HC5$

Para a versão modificada do estimador $HC5$, denotada por $HC5m$, usaremos uma estrutura semelhante à adotada para o estimador $HC4m$ (versão modificada do estimador $HC4$). Novamente, a modificação foi feita apenas na estrutura do expoente do termo $(1 - h_t)$, utilizada como fator de inflação dos resíduos ao quadrado no estimador $HC5$. A estrutura do estimador $HC5m$

é determinada por

$$\delta_t = \min\{(nh_t)/p, \max\{\gamma_1, nk_1h_{\max}/p\}\} + \min\{(nh_t)/p, \max\{\gamma_2, nk_2h_{\max}/p\}\}, \quad (4.1)$$

onde γ_1 e γ_2 são constantes reais positivas pré-especificadas, k_1 e k_2 sendo constantes pré-definidas pertencentes ao intervalo $[0, 1]$. O novo estimador é dado por

$$HC5m = B\widehat{\Omega}_{5m}B' = BE_{5m}\widehat{\Omega}B',$$

em que $E_{5m} = \text{diag}\{1/\sqrt{(1-h_t)^{\delta_t}}\}$, sendo δ_t conforme estabelecido em (4.1), $B = (X'X)^{-1}X'$ e $\widehat{\Omega} = \text{diag}\{\hat{u}_1^2, \dots, \hat{u}_n^2\}$.

Consideramos os valores $\gamma_1 = 1$ e $\gamma_2 = 1.5$, os mesmos valores considerados na obtenção do estimador $HC4m$. Resultados numéricos obtidos e não apresentados no texto sugerem $k_1 = 0.1$ e $k_2 = 0.5$ como escolhas atraentes para os valores de γ_1 e γ_2 considerados. Vale observar que com $k_1 = k_2 = 0$ o estimador $HC5m$ reduz-se ao estimador $HC4m$, proposto na Seção 3.3.

É importante notar a semelhança entre os estimadores $HC4m$ e $HC5m$ e os estimadores $HC4$ e $HC5$. A diferença marcante entre os estimadores $HC4m$ e $HC5m$ é a mesma encontrada quando comparamos os estimadores $HC4$ e $HC5$. Os estimadores $HC5$ e $HC5m$ levam em consideração não apenas a razão entre a t -ésima alavancagem (h_t) e o nível médio de alavancagem ($\bar{h}$), mas também o impacto que o nível *máximo* de alavancagem ($h_{\max}$) exerce sobre a inferência resultante.

As Figuras 4.1 e 4.2 apresentam gráficos de δ_t versus a razão entre a t -ésima medida de alavancagem (h_t) e nível médio de alavancagem ($\bar{h}$) para os estimadores $HC4m$ e $HC5m$. Na Figura 4.1 mostramos um caso em que os expoentes que compõem a estrutura dos estimadores $HC4m$ e $HC5m$ coincidem fazendo $k_1h_{\max}/\bar{h} = 0.3$ e $k_2h_{\max}/\bar{h} = 1.5$, enquanto que na Figura 4.2 consideramos um caso em que $h_{\max}/\bar{h} > 3$ fazendo $k_1h_{\max}/\bar{h} = 0.97$ e $k_2h_{\max}/\bar{h} = 4.85$. Sendo $k_1, k_2 \neq 0$, se considerarmos um nível de alavancagem máxima ($h_{\max}$) não-acentuado com $h_{\max}/\bar{h} \leq 3$ notamos que os expoentes δ_t 's que compõem a estrutura dos estimadores $HC4m$ e $HC5m$ coincidem; caso contrário, isso não acontece. Note que, diferentemente do estimador $HC5$, onde para níveis de alavancagem máxima tais que $h_{\max}/\bar{h} < 5$ os valores dos expoentes

δ_t 's coincidem com os do estimador $HC4$, para o estimador $HC5m$ tais expoentes coincidem com os do estimador $HC4m$ se e somente se $h_{\max}/\bar{h} \leq 3$.

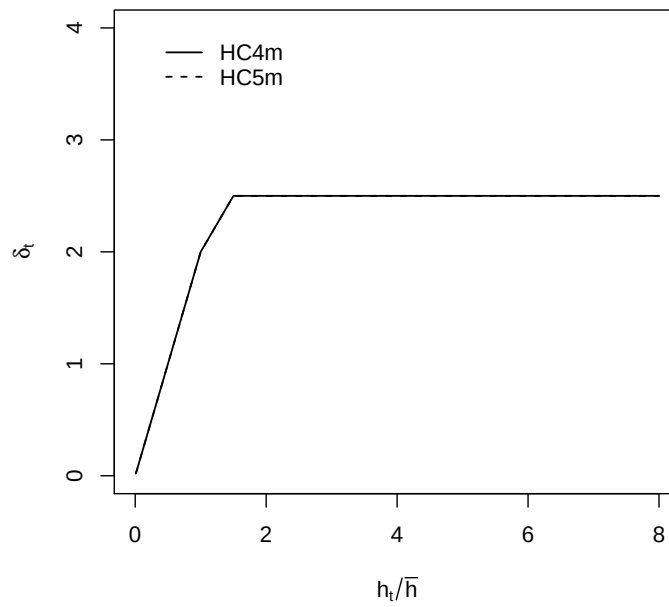


Figura 4.1: Gráfico de δ_t versus a razão entre a t -ésima medida de alavancagem (h_t) e o nível médio de alavancagem para os estimadores $HC4m$ e $HC5m$, quando $\gamma_1 = 1.0$ e $\gamma_2 = 1.5$, $k_1 = 0.1$ e $k_2 = 0.5$ com $h_{\max}/\bar{h} \leq 3$.

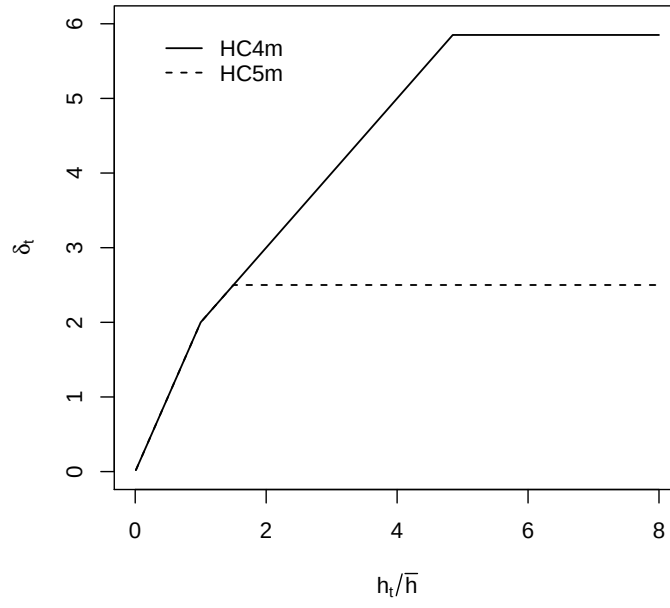


Figura 4.2: Gráfico de δ_t versus a razão entre a t -ésima medida de alavancagem (h_t) e o nível médio de alavancagem para os estimadores $HC4m$ e $HC5m$, quando $\gamma_1 = 1.0$ e $\gamma_2 = 1.5$, $k_1 = 0.1$ e $k_2 = 0.5$ com $k_1 h_{\max}/\bar{h} = 0.97$ e $k_2 h_{\max}/\bar{h} = 4.85$.

4.3 Avaliação Numérica

Na presente seção, nosso interesse residirá em avaliar numericamente o impacto das alterações feitas nos estimadores *HC4* e *HC5*. Adicionalmente, estaremos interessados em avaliar numericamente o estimador proposto no presente capítulo: *HC5m*.

Inicialmente consideraremos o modelo de regressão linear simples

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, \dots, n, \quad (4.2)$$

em que u_t é normalmente distribuído com média zero e variância $\sigma_t^2 = \exp(\alpha x_t)$, α sendo uma constante real. Novamente, assumiremos erros não-correlacionados e normais. A medida de heteroscedasticidade utilizada é a mesma dada em (3.2). A estatística de teste considerada é a mesma dada em (3.3), onde $\widehat{\text{var}}(\hat{\beta}_1)$ é baseado nos estimadores *HC4*, *HC4m*, *HC5* e *HC5m*. Ao longo da avaliação numérica em foco utilizamos o algoritmo de Imhof (1961), discutido na Seção 1.4.

Os valores dos regressores foram escolhidos como realizações da distribuição lognormal padrão ao quadrado (cenário 1), da distribuição lognormal padrão (cenário 2) e da distribuição χ_1^2 (cenário 3). A Tabela 4.1 contém os valores da razão $h_{\max}/(3p/n)$ nos cenários considerados.

Tabela 4.1: Valores de $h_{\max}/(3p/n)$ para os cenários considerados.

desenho	$h_{\max}/(3p/n)$
cenário 1	3.49
cenário 2	3.23
cenário 3	3.28

A Figura 4.3 apresenta os gráficos das discrepâncias relativas de probabilidade sendo considerados os estimadores *HC4*, *HC5* e as suas versões modificadas para os três cenários identificados acima. O tamanho da amostra considerado foi $n = 25$ com $\lambda \approx 100$ e os desenhos de regressão foram com pontos de alavancagem. Os valores tomados para a constante α nos cenários 1, 2 e 3 foram 0.096, 0.39 e 0.476, respectivamente. Percebe-se a pobre aproximação dada pela distribuição nula limite longe da cauda quando as estatísticas de teste são baseadas nos estimadores *HC4* e *HC5*, o problema sendo mais severo para estatísticas *HC4*. Podemos

observar, com maior destaque para o cenário 1, como a utilização dos estimadores $HC4$ e $HC5$ conduziu a pobres aproximações longe da região caudal.

Fato interessante é que as versões modificadas dos estimadores $HC4$ e $HC5$ procuram atenuar o problema das aproximações pobres ao longo do restante do suporte da distribuição nula exata de estatísticas quase- t $HC4$ e $HC5$. A Tabela 4.2 contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 , quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2), considerando-se estatísticas de teste baseadas nos estimadores $HC4$, $HC4m$, $HC5$ e $HC5m$.

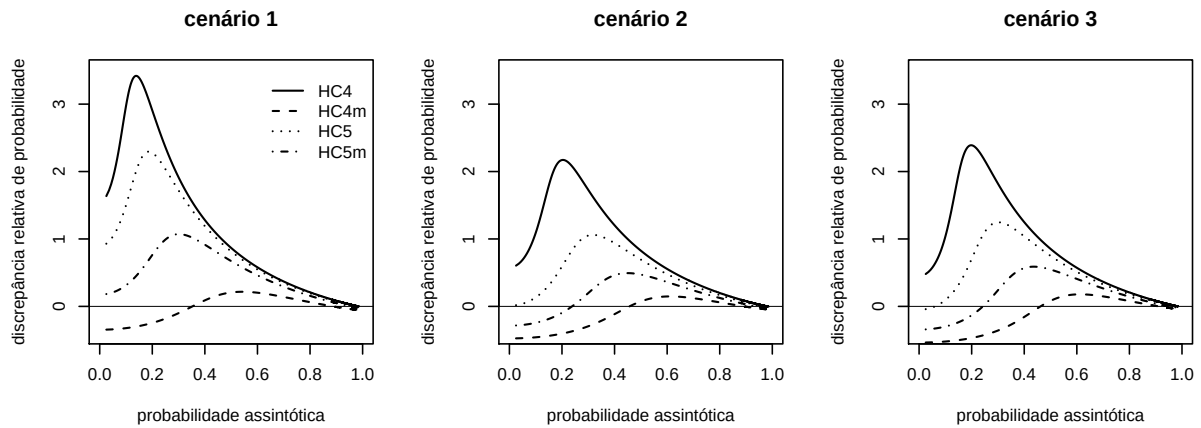


Figura 4.3: Gráficos de discrepâncias relativas de probabilidade para os cenários 1, 2 e 3.

Tabela 4.2: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC4$, $HC4m$, $HC5$ e $HC5m$ considerando-se os cenários 1, 2 e 3.

desenho	Pr	$HC4$	$HC4m$	$HC5$	$HC5m$
cenário 1	0.90	0.974	0.892	0.965	0.944
	0.95	0.979	0.914	0.972	0.954
cenário 2	0.90	0.969	0.895	0.951	0.928
	0.95	0.975	0.919	0.961	0.943
cenário 3	0.90	0.974	0.907	0.960	0.939
	0.95	0.979	0.928	0.968	0.952

Percebe-se na Tabela 4.2 que as novas versões modificadas dos estimadores conduziram a

resultados bastante atraentes, sendo os melhores resultados para o quantil 0.90 obtidos através de estatísticas $HC4m$ e para o quantil 0.95 obtidos através de estatísticas $HC5m$. Notamos que nesses casos o nível máximo de alavancagem foi bastante elevado.

Seguiremos a análise numérica usando dados reais. O primeiro conjunto de dados consiste de dados sobre a variação percentual anual dos preços das ações (y) e preços ao consumidor (x) para um corte de 20 países no períodos pós-Segunda Guerra (até 1969) (ver Gujarati 2000, p. 398). O modelo de regressão considerado é o mesmo dado em (4.2) e deseja-se testar $H_0 : \beta_1 = 0$. Os dados foram extraídos de Gujarati (2000). O segundo conjunto de dados utilizado consiste de dados sobre despesas per capita em escolas públicas por estado nos Estados Unidos em 1979; esses dados foram utilizados no capítulo anterior e foram extraídos de Greene (1997). O modelo e teste considerados são os mesmos dados na Seção 3.5. Em ambos os casos, $\lambda \approx 50$, onde para o primeiro conjunto de dados escolhemos $\alpha = 0.161$.

A Figura 4.4 apresenta gráficos de discrepâncias relativas de probabilidade para os dois casos analisados, a saber: o caso 1 (referente ao primeiro conjunto de dados) e o caso 2 (referente ao segundo conjunto de dados). Percebe-se que as distribuições nulas exatas de estatísticas baseadas nos estimadores $HC4m$ e $HC5m$ foram melhores aproximadas pela distribuição nula limite (χ_1^2) ao longo de todo seu suporte relativamente a estatísticas $HC4$ e $HC5$, respectivamente. Note que as distribuições nulas exatas de estatísticas quase- t $HC4$ e $HC5$ foram pobremente aproximadas pela distribuição nula limite (χ_1^2) ao longo de todo o seu suporte, o problema sendo mais severo no caso 1.

A Tabela 4.3 contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 , quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2), considerando-se estatísticas de teste baseadas nos estimadores $HC5$ e $HC5m$ para os casos agora analisados. No caso 1, observa-se que o estimador $HC5m$ conduziu a resultados melhores do que os obtidos com o estimador $HC5$ enquanto que no caso 2 este fato ocorreu apenas para o quantil 0.90 da distribuição nula limite (χ_1^2).

Com base nas análises numéricas aqui realizadas podemos concluir que não apenas o aumento no número de regressores do modelo de regressão, mas também o nível extremo de alavancagem

máxima pode deteriorar as aproximações usadas no teste $HC4$. Esse fato parece ocorrer com menor intensidade com a utilização do estimador $HC5$.

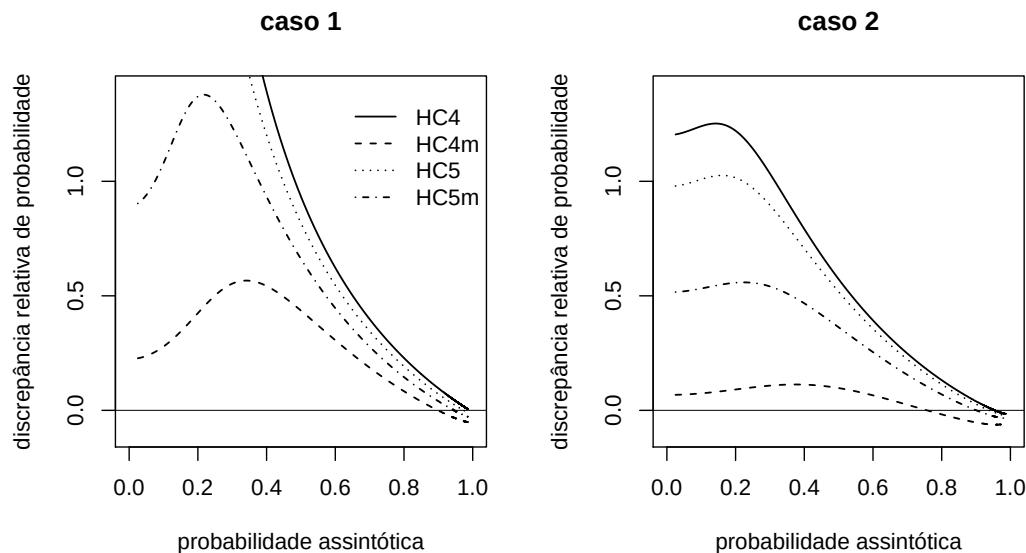


Figura 4.4: Gráficos das discrepâncias relativas de probabilidade para os casos 1 (painel à esquerda) e 2 (painel à direita).

Tabela 4.3: $\Pr(\tau^2 < \zeta | H_0)$ usando dados reais (casos 1 e 2); ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC5$ e $HC5m$.

caso	Pr	$HC5$	$HC5m$
1	0.90	0.964	0.935
	0.95	0.970	0.946
2	0.90	0.924	0.900
	0.95	0.943	0.925

Seguiremos com nossa análise numérica agora considerando o modelo de cinco parâmetros na equação (4.2). Desta vez, os valores dos regressores foram escolhidos como realizações das distribuições lognormal padrão (caso 1) e t_3 (caso 2). Aqui, consideramos apenas estatísticas $HC5$ e $HC5m$ e modelos heteroscedásticos com $\lambda \approx 100$ e com pontos de alavancagem. Para obtermos $\lambda \approx 100$ tomamos $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.36$ no caso 1 e $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.235$ no caso 2, com $h_{\max}/(3p/n) \approx 1.5$ para ambos os casos.

A Figura 4.5 contém os gráficos das discrepâncias relativas de probabilidade para os casos 1 e 2 com $n = 25$. Observa-se a superioridade do estimador $HC5m$ relativamente ao estimador $HC5$ no que tange à precisão da aproximação da distribuição nula exata pela distribuição nula limite ao longo de todo o seu suporte. Podemos observar na Tabela 4.4, a qual contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 , quantis 0.90 e 0.95 da distribuição nula limite (χ_1^2) sendo $n = 25$ e $n = 50$ (dados replicados), que tal superioridade também ocorreu nos quantis considerados.

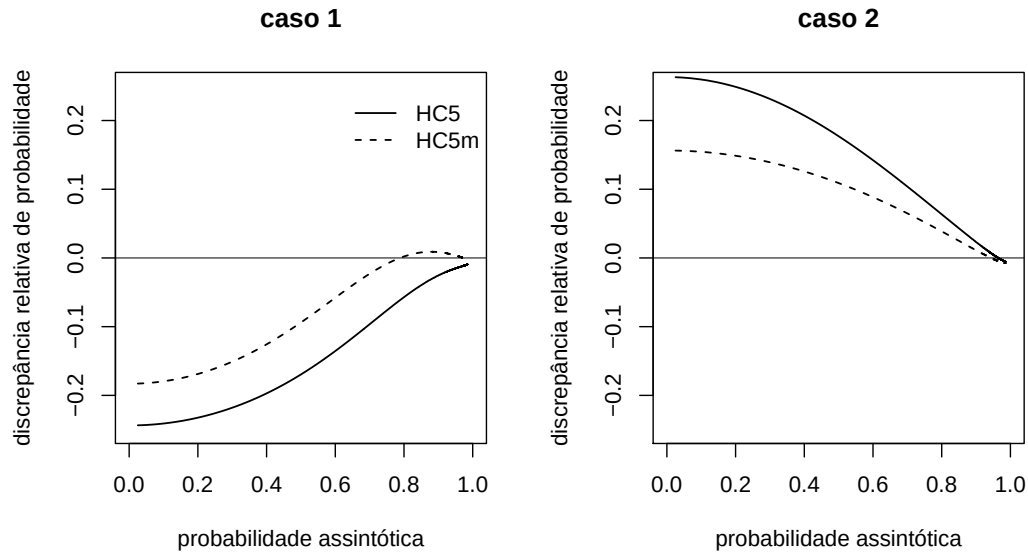


Figura 4.5: Gráficos das discrepâncias relativas de probabilidade para os casos 1 (painel à esquerda) e 2 (painel à direita).

Tabela 4.4: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$ e 50 ; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC5$ e $HC5m$ considerando-se os casos 1 e 2.

		caso1		caso 2	
n	Pr	$HC5$	$HC5m$	$HC5$	$HC5m$
25	0.90	0.877	0.907	0.920	0.910
	0.95	0.935	0.953	0.954	0.948
50	0.90	0.874	0.889	0.877	0.885
	0.95	0.934	0.944	0.929	0.935

Para uma última análise numérica consideramos o modelo de regressão linear dado em (4.2). Os valores do regressor foram escolhidos como pontos igualmente espaçados em $[0, 1]$. Em seguida, substituímos o último ponto da sequência (que é igual a um) pelos valores $a = 1.5, 2, 3, 4$ (casos a, b, c e d , respectivamente), de forma a induzir alavancagem. Em todos os casos, $\lambda \approx 100$ e $n = 25$. A Tabela 4.5 contém os valores da razão $h_{\max}/(3p/n)$ nos casos considerados.

Tabela 4.5: Valores de $h_{\max}/(3p/n)$ para os casos considerados.

caso	$h_{\max}/(3p/n)$
a	1.50
b	2.27
c	3.18
d	3.59

A Figura 4.6 contém os gráficos das discrepâncias relativas de probabilidade considerando-se estatísticas baseadas nos estimadores $HC3$, $HC4$, $HC4m$, $HC5$ e $HC5m$. Percebe-se como fortes alavancagens deterioram a aproximação da distribuição nula exata pela distribuição nula assintótica fora da região caudal, especialmente para as estatísticas $HC4$ e $HC5$. Nota-se também como estatísticas baseadas nos estimadores $HC4m$ e $HC5m$ atenuam esse comportamento.

A Tabela 4.6 contém as probabilidades calculadas $\Pr(\tau^2 \leq \zeta | H_0)$ com $\zeta = 2.706$ e 3.841 considerando-se estatísticas baseadas nos estimadores aqui analisados. Nota-se que a deterioração ocorre também na região caudal, este problema sendo atenuado com o uso dos estimadores propostos no presente texto. Vale ressaltar o fraco desempenho do teste $HC3$ na região caudal relativamente aos testes $HC4m$ e $HC5m$.

4.4 Análise de poder

A seguir realizaremos uma análise dos poderes dos testes aqui considerados usando simulação de Monte Carlo. Foram realizadas 10,000 réplicas de Monte Carlo. O modelo de regressão utilizado foi o que está apresentado na equação (4.2). Os valores dos regressores foram obtidos como realizações da distribuição lognormal padrão ao quadrado. Aqui, $\lambda \approx 100$ e $h_{\max}/(3p/n) \approx 1.6$. Nosso interesse reside em testar $H_0 : \beta_1 = 0$ contra $H_1 : \beta_1 \neq 0$. A geração de dados foi

Tabela 4.6: $\Pr(\tau^2 < \zeta | H_0)$ para $n = 25$; ζ igual aos quantis 0.90 e 0.95 da distribuição χ_1^2 ; estatísticas de teste baseadas nos estimadores $HC3$, $HC4$, $HC4m$, $HC5$ e $HC5m$ considerando-se os diferentes níveis de alavancagem dados nos casos a , b , c e d .

caso	Pr	HC3	HC4	HC4m	HC5	HC5m
a	0.90	0.866	0.925	0.887	0.856	0.840
	0.95	0.919	0.953	0.931	0.910	0.902
b	0.90	0.859	0.951	0.897	0.886	0.874
	0.95	0.907	0.965	0.930	0.922	0.915
c	0.90	0.843	0.972	0.908	0.955	0.935
	0.95	0.887	0.978	0.930	0.964	0.949
d	0.90	0.829	0.982	0.913	0.977	0.961
	0.95	0.872	0.985	0.931	0.981	0.968

realizada usando $\beta_1 = 0.4$ e $\beta_i = 1$, $i = 0, 2, 3, 4$. As estatísticas de teste foram baseadas nos estimadores $HC3$, $HC4$, $HC4m$, $HC5$ e $HC5m$.

A Tabela 4.7 contém as taxas de rejeição da hipótese nula (poderes dos testes) considerando-se os quantis 0.90, 0.95 e 0.99 da distribuição nula exata. Os níveis de significância estão denotados por α . O tamanho da amostra considerado foi $n = 25$; replicamos os dados a fim de obtermos amostras de tamanho 50 e 75. Todas as entradas da tabela são percentagens. Nota-se que os poderes do teste $HC3$ foram os maiores dentre todos os testes considerados. Destacamos também que os poderes do teste $HC4m$ foram superiores aos do teste $HC4$.

Tabela 4.7: Poder dos testes $HC3$, $HC4$, $HC4m$, $HC5$ e $HC5m$ para valores de $n = 25$, 50 e 75; $\lambda \approx 100$ e $h_{max}/(3p/n) \approx 1.6$.

	$n = 25$			$n = 50$			$n = 75$		
α	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
HC3	76.43	56.16	29.59	85.24	70.53	44.60	99.44	95.65	69.23
HC4	67.35	41.33	11.34	80.56	63.11	35.22	99.31	94.45	62.23
HC4m	70.67	47.12	20.43	83.57	67.97	41.46	99.37	95.27	67.07
HC5	75.56	54.71	28.07	84.77	69.87	43.85	99.42	95.55	68.70
HC5m	70.54	46.89	20.25	83.48	67.83	41.23	99.37	95.26	66.90

4.5 Conclusão

Com base nos resultados apresentados no presente capítulo, notamos como não apenas o aumento no número de parâmetros mas também o nível extremo de alavancagem máxima pode deteriorar as aproximações usadas, o que ocorre com menor intensidade quando usamos o estimador $HC5$. As versões modificadas dos estimadores $HC4$ e $HC5$ procuram atenuar esse problema.

Em linhas gerais, os resultados numéricos apresentados aqui e no capítulo anterior evidenciaram como os estimadores $HC4m$ e $HC5m$ se mostraram superiores aos correspondentes estimadores $HC4$ e $HC5$ em modelos com muitos parâmetros ou em situações de alavancagem máxima elevada sob forte heteroscedasticidade uma vez que as distribuições nulas exatas de estatísticas quase- t baseadas nesses estimadores foram melhores aproximadas pela distribuição nula assintótica ao longo de todo o seu suporte. Em muitos casos esse fato se deu também nas caudas da distribuição.

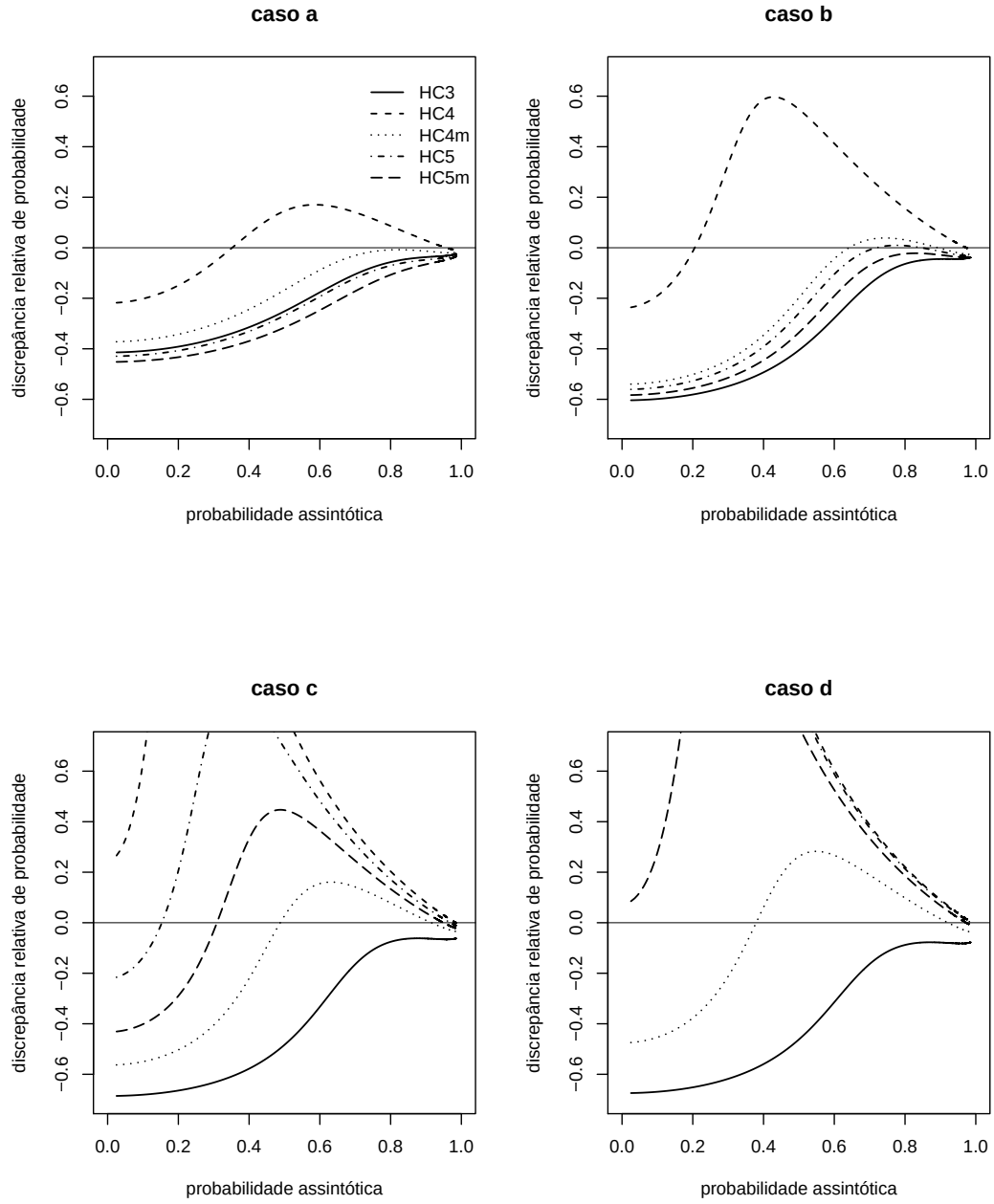


Figura 4.6: Discrepâncias relativas de probabilidade para os casos a , b , c e d ; $n = 25$, $\lambda \approx 100$; estatísticas $HC3$, $HC4$, $HC4m$, $HC5$ e $HC5m$.

Conclusões finais e sugestões para trabalhos futuros

É comum haver padrões heteroscedásticos em dados de regressão de corte transversal. Uma estratégia comumente utilizada em trabalhos aplicados que empregam o modelo linear de regressão é a estimação pontual dos parâmetros através do método de mínimos quadrados ordinários, já que o estimador resultante é consistente, não-viesado e assintoticamente normal mesmo quando a suposição de igualdade de variâncias é violada. Há vários estimadores consistentes para a estrutura de covariâncias desse estimador. Eles têm a propriedade de serem consistentes tanto sob homoscedasticidade quanto sob heteroscedasticidade de forma desconhecida e são comumente usados em estatísticas de teste, conduzindo a testes que se baseiam em aproximações assintóticas. Na presente dissertação, nós usamos um algoritmo de integração numérica para avaliar, de forma exata, a qualidade de tais aproximações. Consideramos diferentes estimadores consistentes ($HC0$, $HC3$, $HC4$ e $HC5$) e mostramos que muitas vezes a aproximação se deteriora sobremaneira à medida em que nos distanciamos da região caudal das distribuições nulas das estatísticas de teste correspondentes. Para minorar esse problema, propusemos dois estimadores consistentes que são variações dos estimadores $HC4$ e $HC5$. Avaliações numéricas realizadas mostraram que o uso desses novos estimadores pode conduzir a aproximações assintóticas mais precisas, inclusive fora da região caudal. Pesquisas futuras devem avaliar o comportamento desses estimadores quando usados para estimação intervalar, já que na pre-

sente dissertação o nosso foco residiu exclusivamente em testes de hipóteses. Adicionalmente, é importante desenvolver inferências robustas à presença de heteroscedasticidade em modelos não-lineares de regressão.

Abaixo se encontra um dos programas utilizados na obtenção dos resultados apresentados.

```
#include <oxstd.h>
#include "probimhof.ox"
#include <oxprob.h>
main()
{
    decl F4,F5,F4m,F5m,i,n,p,c,alpha1,alpha2,alpha3,alpha4;
    decl X1,X,X2,X3,X4,invx,P,H,I,sigma2,Omega,Pc,R;

    /*Etapa 1- Calculo das matrizes H, X e R */

    n=25;//tamanho da amostra;
    p=5;//numero de parametros
    c=(0|1|0|0|0);//vetor c
    alpha1=alpha2=alpha3=alpha4=0.36;
    ranseed("GM");// gerador GM
    ranseed({1510,2030});//semente escolhida
    X1=ranlogn(n,1);//coluna 1 de regressores
```

```

X2=ranlogn(n,1);//coluna 2 de regressores
X3=ranlogn(n,1);//coluna 3 de regressores
X4=ranlogn(n,1);//coluna 4 de regressores
X1=X1|X1; X2=X2|X2; X3=X3|X3; X4=X4|X4;//replicando os dados
n=2*n;//dobrando o tamanho amostral
X=1~X1~X2~X3~X4;//matriz X do modelo
invx=invertsym(X'*X);//inversa da matriz x'x
P=invx*X';
H=X*P;//matriz H das alavancagens
I=unit(n);
sigma2=exp(alpha1*(X[][1])+alpha2*(X[][2])+alpha3*(X[][3])+alpha4*(X[][4]));//
função cedástica
Omega=diag(sigma2);//matrix Omega
Pc=P'*c*c'*P;//matriz usada para computar as matrizes V_i
R=sqrt(Omega)*X*invx*c*c'*invx*X'*sqrt(Omega);//matriz R

/* Etapa 2- computando as matrizes $G_i$ */

decl w,m,mx1,mx2,mx3,g,g1,g2,g3,gnew,gnew2,weight4,weight5,weight4m,weight5m,
gnew5,gnew6,gnew7,gnew8,D4,D4m,D5,D5m,G4m,G4,G5,G5m,V4m,V4,V5,V5m;

w=diagonal(H);//vetor das alavancagens
m= max(diagonal(H));//h_max
mx1=max(1,(n*0.1*m)/p);
mx2=max(1.5,(n*0.5*m)/p);
mx3=max(4,(n*0.7*m)/p);
g=(n.*w)./(p);//computando (h_t)/(p/n), t= 1,...,n e armazenando em g
g1=g.>mx1 .?mx1 .:g;

```

```

g2=g.>mx2 .?mx2 .:g;
g3=g.>mx3 .?mx3 .:g;
gnew=g.>4 .? 4 .:g;//em gnew armazenando os min{4, (h_t)/(p/n)}
gnew2=g.>1 .? 1 .: g;//em gnew2 armazenando os min{1.0, (h_t)/(p/n)}
gnew5=g1./2+g2./2;
gnew6=g.>1.5 .? 1.5 .: g;//em gnew6 armazenando os min{1.5, (h_t)/(p/n)}
gnew7=gnew2+gnew6;
gnew8=g3./2;
weight4=sqrt(1.0./((1.0-w').^gnew'));//usado para o HC4
weight4m=sqrt(1.0./((1.0-w').^gnew7'));//usado para o HC4m
weight5=sqrt(1.0./((1.0-w').^gnew8'));//usado para o HC5
weight5m=sqrt(1.0./((1.0-w').^gnew5'));//usado para o HC5m

/* computando as matrizes V_i e G_i */

D4=diag(weight4);//usada para o HC4
D4m=diag(weight4m);//usada para o HC4m
D5=diag(weight5);//usada para o HC5
D5m=diag(weight5m);//usada para o HC5m
V4=diagonalize(D4*Pc*D4);
V4m=diagonalize(D4m*Pc*D4m);
V5=diagonalize(D5*Pc*D5);
V5m=diagonalize(D5m*Pc*D5m);
G4=sqrt(Omega)*(I-H)*V4*(I-H)*sqrt(Omega);
G4m=sqrt(Omega)*(I-H)*V4m*(I-H)*sqrt(Omega);
G5=sqrt(Omega)*(I-H)*V5*(I-H)*sqrt(Omega);
G5m=sqrt(Omega)*(I-H)*V5m*(I-H)*sqrt(Omega);

```

```

/*Laço para cálculo das probabiliades exatas */

F4=(ProbImhof(0, R, G4, 0, I));
F4m=(ProbImhof(0, R, G4m, 0, I));
F5=(ProbImhof(0, R, G5, 0, I));
F5m=(ProbImhof(0, R, G5m, 0, I));
for(i=0;i<6;i=i+0.001)
{
    F4=(ProbImhof(i, R, G4, 0, I));
    F4m=(ProbImhof(i, R, G4m, 0, I));
    F5=(ProbImhof(i, R, G5, 0, I));
    F5m=(ProbImhof(i, R, G5m, 0, I));
    println(probchi(i,1)~F4~F4m~F5~F5m);// probabilidades
    computadas para HC4, HC4m, HC5 e HC5
}

/*imprimindo dados importantes*/

println("\t OX PROGRAM: ", oxfilename());
println("\t OX VERSION: ", oxversion());
println("\t DATE: ", date());
println("\t TIME: ", time());
println("\t SAMPLE SIZE: ", n);
println("\t LAMBDA: ", max(sigma2)/min(sigma2));
println("\t MAXIMAL LEVERAGE: ", max(diagonal(H)));
println("\t LEVERAGE LIMIT 1: ", 2*(p/n));
println("\t LEVERAGE LIMIT 2: ", 3*(p/n));
}

```

Abaixo segue o código fonte que implementa a função `ProbImhof` na linguagem de programação `0x`.

```
// The function ProbImhof(x, A, B, m, S) calculates the cumulative
// distribution function, evaluated at x, of the ratio of quadratic
// forms,  $(z'Az)/(z'Bz)$ , in a normal random vector z with mean vector
// m and variance matrix S.
// If B=0, then the distribution of  $z'Az$  is computed.
//
// ProbImhof(x, A, B, m, S)
//   x in: scalar, x-value at which distribution is evaluated;
//   A in: nxn matrix (is transformed into symmetric  $(A+A')/2$ );
//   B in: nxn matrix (is transformed into symmetric  $(B+B')/2$ ) or 0;
//   m in: nx1 mean vector;
//   S in: nxn positive definite variance matrix (pd is not checked).
//
// Return value
//   Returns the probability
```

```

// P [ (z'Az) <= x ] if B == 0,
// P [ (z'Az)/(z'Bz) <= x ] if B <> 0,
//      with z ~ N[m,S].
//
// QuanImhof(p, A, B, m, S)
//      p in: scalar, probability at which quantile is evaluated;
// returns x;

#include <oxstd.h>
#include <oxfloat.h>
#include <quadpack.h>

static decl s_l, s_d, s_c;
const decl QUANT_MAXIT = 200;
const decl QUANT_EPS   = 1e-8;

static imhof_mgf(const u)
{
decl eps = 0.5*(sumc(atan(s_l*u)+(s_d.^2).*s_l*u./(1+((s_l*u).^2)))-s_c*u);
decl gam = prodc(((1+((s_l*u).^2)).^0.25).*exp(0.5*((s_d.*s_l*u).^2)/(1+(s_l*u).^2)));
return (sin(eps)/(u*gam));
}

ProbImhof(const x, const A, const B, const m, const S)
{
decl Q, V, result, abserr;
decl P = choleski(S);
  if (B == 0)

```

```

{
    Q = P'((A+A')/2)*P;
    eigensym(Q,&s_l,&V);
    s_c = x;
}
else
{
    Q = A - B*x;
    Q = (P'((Q+Q')/2)*P);
    eigensym(Q,&s_l,&V);
    s_c = 0;
}
s_d = V'solveLU(P, 0, 0, unit(rows(P))) *m;
s_l = s_l';
s_d = selectifr(s_d,s_l);
s_l = selectifr(s_l,s_l);
QAGI(imhof_mgf,0,1,&result,&abserr);
return (0.5 - result/M_PI);
}

QuanImhof(const p, const A, const B, const m, const S)
{
    // not very efficient, but it works
    decl i, pa, pb, xa, xb, w, diff, pn, xn;
    if (p <= 0)
        return 0;
    // find an initial bracket
    pb = xb = 0.0;

```



```

do
{
    pa = pb; xa = xb;
    xb = (xb + 1) * 2;
    pb = ProbImhof(xb, A, B, m, S);
} while (pb < p);

// now narrow bracket
for (i = 0; ; ++i)
{
diff = pb - pa;
w = diff > 0.01 ? 0.5 : (pb - p) / diff;
xn = xa * w + xb * (1 - w);
pn = ProbImhof(xn, A, B, m, S);
if (pn < p)
    xa = xn, pa = pn;
else
    xb = xn, pb = pn;
if (pb - p < QUANT_EPS)
    return xb;
else if (p - pa < QUANT_EPS)
    return xa;
if (i >= QUANT_MAXIT)
    return .NaN;
} }

```

Referências Bibliográficas

- [1] Chesher, A.; Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, **55**, 1217-1222.
- [2] Cribari-Neto, F.; Zarkos, S.G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, **18**, 211-228.
- [3] Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computacional Statistics and Data Analysis*, **45**, 215-233.
- [4] Cribari-Neto, F.; Ferrari, S.L.P.; Cordeiro, G.M. (2000). Improved heteroskedasticity-consistent covariance matrix estimators. *Biometrika*, **87**, 907-918.
- [5] Cribari-Neto, F.; Souza, T.C.; Vasconcellos, K.L.P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics, Theory and Methods*, **36**, 1877-1888. Errata: 37, 3329-3330, 2008.
- [6] Cribari-Neto, F.; Zarkos, S.G. (2001). Heteroskedasticity-consistent covariance matrix estimation: White's estimator and bootstrap. *Journal of Statistical Computation and Simulation*, **68**, 391-411.
- [7] Davidson, R.; MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

- [8] Doornik, J.A. (2001). *Ox: an Object-oriented Matrix Programming Language*, 4^a ed. Londres: Timberlake Consultants e Oxford: <http://www.doornik.com>.
- [9] Gil-Pelaez, J. (1951). Note on the inversion theorem. *Biometrika*, **38**, 481-482.
- [10] Greene, W.H. (1997). *Econometric Analysis*, 3^a ed. Upper Saddle River: Prentice Hall.
- [11] Gujarati, D.N. (2000). *Econometria Básica*, 3^a ed. São Paulo: Pearson Makron Books.
- [12] Hinkley, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, **19**, 285-292.
- [13] Horn, S.D.; Horn, R.A.; Duncan, D.B. (1975). Estimating heteroskedastic variances in linear models. *Journal of the American Statistical Association*, **70**, 380-385.
- [14] Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419-426.
- [15] Judge, G.C.; Hill, R.C.; Griffiths, W.E.; Lutkepohl, H.; Lee, T.C. (1988). *Introduction to the Theory and Practice of Econometrics*, 2^a ed. New York: Wiley.
- [16] Long, J.S.; Ervin, L.H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *The American Statistician*, **54**, 217-224.
- [17] MacKinnon, J.G.; White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics*, **29**, 305-325.
- [18] Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- [19] White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48**, 817-838.