INFERÊNCIA EM MODELOS HETEROSCEDÁSTICOS NA PRESENÇA DE PONTOS DE ALAVANCA

TATIENE CORREIA DE SOUZA

Orientador: Prof. Dr. Francisco Cribari Neto
Co-orientador: Prof. Dr. Klaus Leite P. Vasconcellos
Área de concentração: Estatística Matemática

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

Recife, dezembro de 2003

Aos meus pais.

Agradecimentos

A Deus, por me cercar de pessoas maravilhosas e por me abençoar a cada dia.

Aos meus pais, que são as pessoas que mais amo, mais importantes da minha vida e aos quais eu dedico todas as minhas conquistas.

A minha querida e amada, Vitória, por existir, pelos inúmeros momentos de felicidade e por me ensinar mais uma forma de amar.

Ao meu orientador, professor e amigo, Francisco Cribari Neto, pela seriedade, dedicação, estímulo, confiança, competência e por estar sempre presente nos momentos difíceis, a minha eterna gratidão.

Aos meus irmãos que tanto amo, Hélio e Humberto.

A Antônio Carlos, pelo carinho, compreensão, estímulo e por seu amor incondicional.

Aos meus familiares, em especial, ao meu avô Severino, às minhas avós Rilsa e Josefa, aos meus tios Herton, Adriana e Elizabete pelo incentivo e confiança.

Ao excelente professor Klaus Vasconcellos, por sua amizade, carinho e pela forma competente com que forneceu ótimas sugestões para realização deste trabalho.

A Tarciana Liberal, por ser uma grande amiga e por todos os momentos inesquecíveis que passamos juntas.

A Andréa, por ser tão meiga e amiga.

A minha inesquecível amiga, Valéria, a excelente secretária do Mestrado em Estatística.

A minha querida amiga, Sílvia, pelos momentos de alegria e por sua enorme amizade.

Aos meus grandes amigos que conquistei durante o mestrado: Felipe, pelo imenso carinho; Sílvia, pelos momentos de confidências e companheirismo; Keila, por me mostrar o quanto podemos ser fortes e por sua enorme confiança; Moisés, um grande irmão por quem tenho um enorme carinho; Gilson, por sua seriedade e amizade; Raydonal, por estar sempre disposto a ajudar; João Marcelo, por seu espírito fraterno e carinho; Patrícia, por sua tranqüilidade; Bartolomeu, por sua alegria e a Cristina Morais, por seu espírito materno.

Aos meus grandes amigos, Andréa, Angela, Ernando, José Ramos, Juliana, Patrícia, Tarciana, pelo amor, carinho e apoio que sempre estão dispostos a me oferecer.

A minha querida e eterna professora Cristina, por ser uma verdadeira mãe para mim.

A Claudia, por estar sempre disposta a me ajudar, por seu enorme carinho, amizade e apoio.

Aos professores Manoel e Sylvio, pelo incentivo.

Aos professores do Programa de Mestrado em Estatística da UFPE, pela socialização de seus conhecimentos.

A Adriana, Cícero e Antônio, funcionários do Departamento de Estatística, pelo carinho e amizade.

Aos meus colegas de mestrado Lenaldo, Sandra Maria, Sandra Rêgo, Gecynalda, Cherubino e Renata, em especial a André, Junior e Tatiane.

À CAPES, pelo apoio financeiro.

Resumo

Técnicas clássicas de regressão linear assumem que os erros, que representam a componente aleatória do modelo, têm variância constante, ou seja, assumem homoscedasticidade. Contudo, esta suposição é bastante forte e, em uma relevante parte dos problemas práticos, muito pouco razoável. A presente dissertação considera a estimação consistente da matriz de covariâncias do estimador de mínimos quadrados ordinários em um modelo de regressão linear sob heteroscedasticidade de forma desconhecida. O estimador mais usado é aquele proposto por Halbert White, conhecido como HC0. Consideramos também outros estimadores consistentes, a saber: o estimador HC3, que é uma aproximação do estimador jackknife, e o estimador HC4 proposto por Cribari–Neto (2004), que leva em consideração o efeito de pontos de alta alavancagem em amostras finitas. Dois estimadores consistentes obtidos a partir de esquemas de reamostragem de bootstrap são também considerados. Nós propomos, com base no estimador HC4, um novo estimador: HC5. Este estimador é o primeiro estimador na classe dos estimadores consistentes da matriz de covariâncias do estimador de mínimos quadrados a incorporar termos de descontos que se ajustam a variações no grau máximo de alavancagem dos dados. Nós apresentamos resultados de simulação de Monte Carlo sobre o desempenho de testes quasi-t cujas estatísticas são baseadas nos diferentes estimadores consistentes. A avaliação é realizada tanto sob homoscedasticidade quanto sob heteroscedasticidade e os resultados revelam que o teste construído a partir do estimador HC5 tipicamente apresenta desempenho superior aos demais testes considerados. No que se refere a inferência via bootstrap, há muito pouco ganho em amostras finitas em se usar o esquema de reamostragem de bootstrap ponderado para realizar testes bootstrap, estimando-se valores p ou valores críticos, ao invés de se utilizar o bootstrap ponderado para estimação de erros-padrão a serem utilizados em estatísticas de teste convencionais. Nossos resultados também revelam que a presença de pontos de alta alavancagem exerce um papel importante no desempenho dos diferentes estimadores consistentes em amostras de tamanho típico. Algumas aplicações empíricas são, por fim, apresentadas.

Abstract

The chief goal of this thesis is to study the finite-sample behavior of different heteroskedasticity-consistent covariance matrix estimators, under both constant and unequal error variances. We consider the estimator proposed by Halbert White (HC0), its variant known as HC3, and Wu's (1986) weighted bootstrap estimator. Recently proposed estimators, such as Cribari–Neto's (2004) HC4 and Cribari–Neto and Zarkos's (2004) inversely adjusted weighted bootstrap, are also considered. We propose a new covariance matrix estimator: HC5. It is the first consistent estimator to explicitly take into account the effect that the maximum level of leverage of the data has on the associated inference. Our numerical (Monte Carlo) results show that quasi-t inference based on HC5 is typically more reliable than inference based on other covariance matrix estimators. We also present four applications to real data.

Índice

1. Introdução 1
   1.1. Considerações iniciais 1
   1.2. Modelo e estimadores 2
   1.3. Organização da dissertação 10
   1.4. Plataforma computacional 11
2. Métodos bootstrap 12
   2.1. Introdução 12
   2.2. O método bootstrap 12
   2.3. Bootstrap em modelos de regressão 14
3. Proposta de um novo estimador e sua avaliação numérica 17
   3.1. Introdução 17
   3.2. Proposta de um novo estimador 18
   3.3. Avaliação numérica 20
   3.4. Bootstrap na estatística de teste quasi-t 44
4. Aplicações 49
   4.1. Introdução 49
   4.2. Atividade em instalações da marinha americana 50
   4.3. Gasto com educação nos EUA - I 52
   4.4. Gasto com educação nos EUA - II 58
   4.5. Graus de prestígio de atividades no Canadá 60
5. Implementação dos estimadores consistentes 63
6. Conclusões e sugestões para trabalhos futuros 70
- Apêndice 72
- Referências 81

Capítulo 1

Introdução

1.1 Considerações iniciais

A análise de regressão é um dos ramos da teoria estatística mais utilizados na pesquisa científica. O modelo clássico de regressão teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821; trata-se de uma técnica adequada para estudar o efeito de variáveis explicativas sobre uma variável resposta. Pouco ainda se conhece sobre as propriedades de inferências feitas para este modelo quando as suposições convencionais não são válidas.

Técnicas clássicas de regressão linear assumem que os erros, que representam a componente aleatória do modelo, têm variância constante, ou seja, apresentam a propriedade de *homoscedasticidade*. Porém, esta suposição é bastante forte e, em uma relevante parte dos problemas práticos, muito pouco razoável.

Sob suposições usuais para este modelo, a estimação dos parâmetros é comumente feita utilizando o método de mínimos quadrados ordinários (MQO). Este método, através de operações matriciais de fácil implementação computacional, fornece estimadores de simples interpretação e que possuem propriedades desejáveis como não-viés, consistência e eficiência. Quando a suposição de homoscedasticidade é violada, ou seja, quando a variância dos erros não é constante, dizemos que há *heteroscedasticidade* no modelo. A presença de heteroscedasticidade ocorre com freqüência quando trabalhamos com dados de corte transversal.

Sob heteroscedasticidade, o estimador de MQO mantém algumas propriedades desejáveis, tornando-se, contudo, ineficiente, i.e., não é mais o melhor estimador linear não-viesado (BLUE–*Best Linear Unbiased Estimator)*. O estimador usual da matriz de covariâncias deste estimador torna-se viesado e não-consistente, tornando pouco confiáveis estimativas intervalares e testes de hipóteses que utilizam tais valores. Uma maneira de contornar este problema é obter estimadores alternativos que sejam melhores que o estimador de MQO. Torna-se razoavelmente fácil estimar os parâmetros de um modelo de regressão quando os erros são heteroscedásticos com um padrão de heteroscedasticidade que é determinado por uma função cedástica conhecida, ou mesmo quando os parâmetros da função cedástica são desconhecidos, mas com forma de heteroscedastici-

dade conhecida. Contudo, quase sempre o processo gerador das diferentes variâncias é desconhecido.

Um procedimento de estimação muito usado na prática para o modelo de regressão linear na presença de heteroscedasticidade de forma desconhecida consiste em utilizar para o vetor de parâmetros β o estimador de mínimos quadrados ordinários, que permanece não-viesado e consistente, juntamente com um estimador para sua matriz de covariâncias que possua propriedades assintóticas desejáveis. Este estimador consistente da matriz de covariâncias do estimador de MQO possibilita a realização de inferências no modelo heteroscedástico sem que seja necessário fazer suposições sobre a forma da heteroscedasticidade.

1.2 Modelo e estimadores

O modelo de interesse é o modelo de regressão linear, definido como

$$y = X\beta + e, \tag{1.1}$$

onde y é um vetor $n \times 1$ de observações de uma variável dependente, X é uma matriz $n \times p$ $(p < n)$ de regressores fixos com posto(X)=p, β é um vetor $p \times 1$ de parâmetros desconhecidos e e é um vetor $n \times 1$ de erros aleatórios.

As seguintes suposições são usualmente feitas:

(i) $\mathrm{E}(e_i) = 0$, $i = 1, \ldots, n$;

(ii) $\mathrm{E}(e_i^2) = \sigma_i^2$, $0 < \sigma_i^2 < \infty$, $i = 1, \ldots, n$;

(iii) $\mathrm{E}(e_i e_s) = 0$, $\forall i \neq s$;

(iv) $\lim\limits_{n\to\infty} \dfrac{(X'X)}{n} = Q$, onde Q é uma matriz positiva-definida.

(v) $\lim\limits_{n\to\infty} \dfrac{(X'\Psi X)}{n} = S$, onde S é uma matriz finita positiva-definida.

Das suposições acima temos que o vetor dos erros do modelo tem média zero e matriz de covariâncias dada por $\Psi = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, onde cada elemento σ_i^2 representa a variância do respectivo termo de erro e_i, $i = 1, \ldots, n$. Sob homoscedastidade, $\sigma_i^2 = \sigma^2$, uma constante positiva, para todo i.

Um dos objetivos centrais em modelagem de regressão é fazer inferências sobre β, já que este vetor representa o efeito dos regressores considerados sobre a média da variável explicada. Mínimos quadrados ordinários é o método mais comumente utilizado para estimação dos parâmetros do modelo (1.1). O objetivo é estimar β minimizando a soma

de quadrados dos erros do modelo, expressa por

$$e'e = (y - X\beta)'(y - X\beta).$$

O estimador de mínimos quadrados ordinários (EMQO) é dado por

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + e) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'e \\ &= \beta + (X'X)^{-1}X'e. \end{aligned} \tag{1.2}$$

Utilizando o fato de que X é uma matriz de regressores fixos e a suposição (i), podemos obter o valor esperado de b:

$$\begin{aligned} \mathrm{E}(b) &= \mathrm{E}[\beta + (X'X)^{-1}X'e] \\ &= \beta + (X'X)^{-1}X'\mathrm{E}(e) \\ &= \beta. \end{aligned}$$

Convém notar que não é necessário assumir homoscedastidade para se estabelecer a propriedade de não-viés do EMQO, sendo necessário apenas assumir que os erros têm média zero. Sob heteroscedasticidade, além de ser não-viesado, o EMQO é consistente para β, i.e., quando n, o tamanho amostral, aumenta, b converge em probabilidade para β, o que será denotado por $\mathrm{plim}(b) = \beta$. É possível provar este resultado, utilizando o teorema que garante que se $g(\cdot)$ é uma função contínua e z_n é qualquer seqüência de vetores aleatórios, então

$$\mathrm{plim}\, g(z_n) = g\left(\mathrm{plim}\, z_n\right)$$

se $\mathrm{plim}\, z_n$ existe. Por (1.2), temos

$$\begin{aligned} \mathrm{plim}(b) &= \mathrm{plim}[\beta + (X'X)^{-1}X'e] \\ &= \beta + \mathrm{plim}\left(\frac{1}{n}X'X\right)^{-1} \mathrm{plim}\left(\frac{1}{n}X'e\right). \end{aligned}$$

Utilizando as suposições (i), (ii) e (v), e aplicando a Lei Fraca de Chebyshev (e.g., Rao, 1973, p.112), obtemos

$$\mathrm{plim}_{n\to\infty}\left(\frac{X'e}{n}\right) = 0,$$

pois $X'e = \sum_{i=1}^{n} X_i' e_i$, onde X_i é a i-ésima linha da matriz X, tal que $\mathrm{E}(X'e) = \sum_{i=1}^{n} X_i' \mathrm{E}(e_i) = 0$. Utilizando o fato de que o limite em probabilidade de uma seqüência não-estocástica é simplesmente igual ao seu limite e aplicando o teorema citado, segue que

$$\begin{aligned} \operatorname{plim}\left(\frac{X'X}{n}\right)^{-1} &= \left(\operatorname{plim}\frac{X'X}{n}\right)^{-1} \\ &= \left(\lim\frac{X'X}{n}\right)^{-1} \\ &= Q. \end{aligned}$$

Logo,

$$\operatorname{plim}(b) = \beta,$$

ou seja, b é consistente para β.

Para avaliar a precisão do estimador de MQO no modelo (1.1), apresentaremos a sua matriz de variâncias e covariâncias. Temos que

$$\begin{aligned} \mathrm{E}[(b-\beta)(b-\beta)'] &= \mathrm{E}[(\beta + (X'X)^{-1}X'e - \beta)(\beta + (X'X)^{-1}X'e - \beta)'] \\ &= \mathrm{E}[(X'X)^{-1}X'ee'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'\Psi X(X'X)^{-1}. \end{aligned} \tag{1.3}$$

Se o modelo for homoscedástico, a matriz de covariâncias dos erros será dada por $\Psi = \sigma^2 I$, onde I é a matriz identidade de ordem n. Assim, teremos

$$\operatorname{cov}(b) = \sigma^2 (X'X)^{-1}. \tag{1.4}$$

O Teorema de Gauss–Markov, enunciado a seguir, garante que a variância (1.4) é mínima, ou seja, garante a eficiência de b, este estimador sendo o melhor estimador não–viesado para β, sob homoscedasticidade.

Teorema (Gauss–Markov). Seja $b^* = C'y$, onde C é uma matriz de constantes tal que $C'X = I$ e sejam válidas as suposições (i), (ii) e (iii) . Então b é mais preciso do que b^* se $b \neq b^*$, ou seja,

$$\operatorname{cov}(b^*) = \operatorname{cov}(b) + A,$$

onde A é uma matriz positiva-definida.

O resultado de Gauss–Markov implica que, quando há homoscedasticidade, o estimador de MQO apresenta propriedades amostrais superiores a qualquer outro estimador de β que seja linear, não-viesado e consistente. O nosso interesse agora reside em verificar o que ocorre quando a suposição (ii) é violada, ou seja, queremos analisar agora as conseqüências de erros heteroscedásticos sobre o processo de estimação de β.

Quando os elementos diagonais da matriz Ψ não são iguais, temos que b continua sendo um estimador não-viesado e consistente para β, contudo em geral deixa de ser eficiente e, conseqüentemente, deixa de ser o melhor estimador linear não-viesado.

É possível obter um melhor estimador não-viesado para β quando os erros têm matriz de covariâncias $\mathrm{E}(ee') = \Psi = \sigma^2\Omega$. O primeiro passo nesta direção é transformar o modelo (1.1), multiplicando ambos os lados da igualdade por uma matriz P de dimensão $n \times n$ que apresenta a seguinte propriedade

$$P\Omega P' = I. \tag{1.5}$$

Como Ω é positiva-definida, a matriz P que satisfaz (1.5) sempre existe. Temos, assim, que

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

Usando a matriz P para transformar o modelo (1.1), obtemos

$$Py = PX\beta + Pe$$

ou

$$y^* = X^*\beta + e^*, \tag{1.6}$$

onde $y^* = Py$, $X^* = PX$ e $e^* = Pe$. O vetor de erro transformado e^* tem média

$$\mathrm{E}(e^*) = \mathrm{E}(Pe) = P\mathrm{E}(e) = 0$$

e matriz de covariâncias

$$\mathrm{E}(e^*e^{*'}) = \mathrm{E}(Pee'P') = P\mathrm{E}(ee')P' = \sigma^2 P\Omega P' = \sigma^2 I.$$

Assim, o modelo transformado é equivalente ao modelo (1.1), exceto pela estrutura homoscedástica dos erros. Utilizando o critério dos mínimos quadrados, o estimador para o vetor de parâmetros β no modelo (1.6) é dado por

$$\widehat{\beta}_G = (X^{*'}X^*)^{-1}X^{*'}y^*,$$

que pode ser reescrito como

$$\begin{aligned}\widehat{\beta}_G &= (X'P'PX)^{-1}X'P'Py\,. \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y\end{aligned} \tag{1.7}$$

Da expressão (1.7) e utilizando as suposições do modelo (1.1), notamos que

$$\begin{aligned}\mathrm{E}(\widehat{\beta}_G) &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\mathrm{E}(y) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta\ \ , \\ &= \beta\end{aligned}$$

e que

$$\Psi_G = \mathrm{cov}(\widehat{\beta}_G) = \sigma^2(X'\Omega^{-1}X)^{-1}.$$

Este estimador é conhecido como estimador de mínimos quadrados generalizados (EMQG) e é o melhor estimador linear não-viesado de β no modelo (1.1), onde $\mathrm{E}(e) = 0$ e $\mathrm{E}(ee') = \sigma^2\Omega$, sendo válido o teorema de Gauss–Markov, assumindo que a matriz de covariâncias Ω é conhecida.

A limitação deste estimador reside no fato de que, na maioria dos casos, a matriz Ω é desconhecida e então temos que fazer alguma suposição sobre a forma de heteroscedasticidade. Uma alternativa é substituir a matriz de covariância Ω por algum estimador consistente $\widehat{\Omega}$, obtendo-se então o estimador de mínimos quadrados generalizados viável (EMQGV)

$$\widehat{\widehat{\beta}} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y.$$

Usualmente supõe-se que os elementos de Ω são funções de um pequeno número de parâmetros desconhecidos e, desse modo, para os vários modelos de heteroscedasticidade considerados, $\widehat{\Omega}$ é obtido através da estimação destes parâmetros.

Apesar da viabilidade de $\widehat{\widehat{\beta}}$, as propriedades amostrais deste estimador não são facilmente determinadas. A dificuldade reside no fato de $\widehat{\Omega}$ e y serem correlacionados, uma vez que qualquer estimador de Ω depende das observações amostrais, não sendo possível tratar $\widehat{\Omega}^{-1}$ como uma matriz de elementos fixos, o que torna complexa até mesmo a obtenção de $\mathrm{E}(\widehat{\widehat{\beta}})$. Note que $\widehat{\widehat{\beta}}$ não é linear, e assim o Teorema de Gauss–Markov não se aplica a este estimador. Contudo, para muitos estimadores de Ω, $\widehat{\widehat{\beta}}$ é consistente, isto é, $\mathrm{plim}(\widehat{\widehat{\beta}}) = \beta$.

Um procedimento adequado de estimação no modelo de regressão linear na presença de heteroscedasticidade de forma desconhecida é utilizar o EMQO para o vetor de

parâmetros β, pois b permanece não-viesado e consistente, juntamente com um estimador de sua matriz de covariâncias que possua propriedades assintóticas desejáveis, independente de uma especificação formal para a estrutura da heteroscedasticidade (Galvão, 2000).

Sob homoscedasticidade, temos que $\sigma_i^2 = \sigma^2 > 0$, isto é, $\Omega = \sigma^2 I$. Então, a matriz de covariâncias de b é dada por $\sigma^2(X'X)^{-1}$ e pode ser facilmente estimada por $\widehat{\sigma}^2(X'X)^{-1}$, com $\widehat{\sigma}^2 = \widehat{e}'\widehat{e}/(n-p)$, onde $\widehat{e} = [I - X(X'X)^{-1}X']y$ é o vetor n-dimensional de resíduos de mínimos quadrados.

Na presença de heteroscedasticidade, o estimador $\widehat{\sigma}^2(X'X)^{-1}$ não é mais consistente nem não–viesado para a matriz de covariâncias de b. Como em um grande número de aplicações os erros são heteroscedásticos, é muito importante considerar estimadores mais confiáveis para a variância de b. Um estimador consistente da matriz de covariâncias de b muito utilizado é o proposto por Halbert White (1980), denotado por HC0. O estimador de White é obtido substituindo o i-ésimo elemento diagonal de Ψ pelo i-ésimo resíduo ao quadrado, resultando em

$$\text{HC0} = (X'X)^{-1}X'\widehat{\Psi}_0 X(X'X)^{-1},$$

onde $\widehat{\Psi}_0 = \text{diag}\{\widehat{e}_1^2, \ldots, \widehat{e}_n^2\}$.

Apesar de ser um estimador consistente da matriz de covariâncias de b tanto sob heteroscedasticidade como sob homoscedasticidade, o estimador HC0 tende a subestimar a variância do estimador de MQO, o que torna este estimador bastante viesado quando o tamanho da amostra é pequeno. Através de simulações de Monte Carlo, Long & Ervin (2000) constataram que este estimador usualmente não é apropriado quando o tamanho da amostra é menor ou igual a 250. Cribari–Neto & Zarkos (1999) e MacKinnon & White (1985), em estudos de simulação, mostraram que este estimador pode ser muito viesado quando o tamanho da amostra não é grande.

Assim como o tamanho amostral, o desenho da regressão, ou seja, a estrutura da matriz de regressores X, é outro fator que também interfere no desempenho do estimador de MQO em amostras finitas. Quando a matriz de regressores X contém observações de alta alavancagem, isto é, observações que possuem potencial para exercer grande influência sobre o valor predito, o estimador HC0 tende a ser muito viesado. Dessa maneira, o uso de estimativas obtidas a partir de HC0 pode levar a inferências errôneas.

MacKinnon & White (1985) consideraram três estimadores alternativos para melhorar as propriedades de HC0 em amostras pequenas. Um ajustamento simples para HC0, sugerido por Hinkley (1977), é usar uma correção de graus de liberdade multiplicando HC0 por $n/(n-p)$. Com esta correção obtém-se a versão conhecida como HC1, dada

por

$$\text{HC1} = \frac{n}{n-p}\left(X'X\right)^{-1} X'\widehat{\Psi}_0 X \left(X'X\right)^{-1}.$$

O ajustamento pelos graus de liberdade em HC1 não é o único caminho para compensar o fato de que os resíduos de mínimos quadrados ordinários tendem a subestimar os erros verdadeiros. Se não existe heteroscedasticidade, temos que

$$\begin{aligned}
\text{E}[\widehat{e}\widehat{e}'] &= \text{E}[(y-\widehat{y})(y-\widehat{y})'] \\
&= \text{E}\{(I-H)y[(I-H)y]'\} \\
&= \text{E}[(I-H)yy'(I-H)'] \\
&= M\text{E}[yy']M' \\
&= \sigma^2 M,
\end{aligned}$$

onde $H = X(X'X)^{-1}X'$ e $M = (I-H)$, usando o fato de que a matriz M é simétrica e idempotente. Assim,

$$\begin{aligned}
\text{E}[\widehat{e}_i^2] &= \sigma^2 m_i \\
&= \sigma^2(1-h_i),
\end{aligned}$$

onde m_i e h_i são os i-ésimos elementos diagonais de M e H, respectivamente. Então, Horn, Horn & Duncan (1975) sugeriram usar

$$\widehat{\sigma}_i^2 = \widehat{e}_i^2/(1-h_i).$$

Neste contexto, obtém-se o estimador referido como HC2 e definido da seguinte forma:

$$\text{HC2} = (X'X)^{-1}X'\widehat{\Psi}_2 X(X'X)^{-1},$$

onde

$$\widehat{\Psi}_2 = \text{diag}\left\{\frac{\widehat{e}_1^2}{1-h_1}, \ldots, \frac{\widehat{e}_n^2}{1-h_n}\right\}.$$

Uma terceira variante do estimador HC0 é baseada na técnica denominada 'jackknife'. A idéia básica desta técnica consiste em recalcular n vezes as estimativas de mínimos quadrados ordinários para o vetor β, cada vez retirando uma das observações, e então usar a variabilidade das estimativas obtidas como estimativa da variância do

estimador de MQO original. Seja $b_{(i)}$ o estimador de mínimos quadrados ordinários de β baseado em todas as observações exceto a i-ésima. É facilmente visto que

$$b_{(i)} = b - (X'X)^{-1}X_i'\widehat{\varepsilon}_i,$$

onde X_i denota a i-ésima linha de X e $\widehat{\varepsilon}_i = \widehat{e}_i/(1-h_i)$. Dessa forma, o estimador 'jackknife' da matriz de covariâncias de b é dado por

$$\frac{n-1}{n}\sum_{i=1}^{n}\left[b_{(i)} - n^{-1}\sum_{s=1}^{n}b_{(s)}\right]\left[b_{(i)} - n^{-1}\sum_{s=1}^{n}b_{(s)}\right]'.$$

Após algumas manipulações, esta quantidade pode ser escrita como

$$\frac{n-1}{n}(X'X)^{-1}[X'\Psi^*X - (n^{-1})(X'e^*e^{*'}X)](X'X)^{-1},$$

onde Ψ^* é uma matriz diagonal $n \times n$ cujo i-ésimo elemento diagonal é $\widehat{\varepsilon}_i^2$.

Davidson & MacKinnon (1993) definem um estimador, denotado por HC3, que é também uma modificação do estimador de White, cujo comportamento é similar ao do estimador 'jackknife', sendo uma aproximação simples deste estimador. O estimador HC3 é dado por

$$\text{HC3} = (X'X)^{-1}X'\widehat{\Psi}_3X(X'X)^{-1},$$

onde

$$\widehat{\Psi}_3 = \text{diag}\left\{\frac{\widehat{e}_1^2}{(1-h_1)^2}, \ldots, \frac{\widehat{e}_n^2}{(1-h_n)^2}\right\}.$$

Cribari–Neto (2004) propõe uma modificação do estimador HC3, denominada HC4, que leva em consideração o impacto de observações com alta alavancagem em amostras finitas incorporando fatores de descontos definidos pela razão entre os graus individuais de alavancagem e o grau médio de alavancagem. A importância de considerar o efeito da alta alavancagem de algumas observações pode ser verificada através dos resultados obtidos em Cribari–Neto & Zarkos (2001), que mostraram que a presença de pontos de alta alavancagem no desenho da matriz dos regressores é mais decisiva para o comportamento de estimadores consistentes para a matriz de covariâncias de b em amostras finitas do que o próprio grau de heteroscedasticidade. O estimador HC4 é definido da seguinte forma:

$$\text{HC4} = (X'X)^{-1}X'\widehat{\Psi}_4X(X'X)^{-1},$$

onde

$$\widehat{\Psi}_4 = \text{diag}\left\{\frac{\widehat{e}_1^2}{(1-h_1)^{\delta_1}}, \ldots, \frac{\widehat{e}_n^2}{(1-h_n)^{\delta_n}}\right\}$$

e

$$\delta_i = \min\left\{4, \frac{h_i}{\overline{h}}\right\} = \min\left\{4, \frac{nh_i}{\sum_{j=1}^n h_j}\right\},$$

com $\overline{h} = n^{-1}\sum_{j=1}^n h_j$, i.e., $\overline{h}$ é a média dos h_i's. Ou seja,

$$\delta_i = \min\left\{4, \frac{nh_i}{p}\right\}.$$

Aqui, usamos o fato de que a soma de todos os níveis individuais de alavancagem é igual a p, ou seja,

$$\sum_{j=1}^n h_j = \text{tr}(H) = \text{tr}(X(X'X)^{-1}X') = \text{tr}[X'X(X'X)^{-1}] = \text{tr}(I_p) = p.$$

O expoente controla o nível de desconto do i-ésimo resíduo quadrado e é determinado pela razão entre h_i e a média dos h_i's, $\overline{h}$. Como $0 < 1 - h_i < 1$ e $\delta_i > 0$, segue que $0 < (1-h_i)^{\delta_i} < 1$. O i-ésimo resíduo ao quadrado deverá ser tanto mais fortemente inflacionado quanto maior for h_i relativamente a $\overline{h}$. Este desconto linear é truncado em 4, que equivale a duas vezes o grau de desconto usado pelo estimador HC3. Ou seja, usa-se $\delta_i = 4$ quando $h_i > 4\overline{h} = 4p/n$. Os resultados numéricos obtidos por Cribari–Neto (2004) mostram que este novo estimador tem comportamento superior ao do estimador HC3 em amostras finitas no que diz respeito a testes quasi-t associados.

1.3 Organização da dissertação

A presente dissertação de mestrado está dividida em cinco capítulos. No segundo capítulo, apresentamos o método bootstrap no contexto do modelo de regressão linear. No Capítulo 3 propomos um novo estimador (HC5) e avaliamos, sob homoscedasticidade e heteroscedasticidade, o comportamento de estatísticas quasi-t construídas utilizando os diferentes estimadores da matriz de covariâncias do EMQO. Com base nas taxas de rejeição sob a hipótese nula e nos vieses relativos totais dos estimadores indicamos quais estimadores conduzem a inferências mais confiáveis. Nesta avaliação numérica também é investigado o efeito da presença de pontos de alavanca sobre o comportamento dos diversos estimadores. No quarto capítulo são apresentados os resultados de quatro aplicações empíricas. No Capítulo 5, apresentamos a implementação dos estimadores consistentes

da matriz de covariâncias do EMQO no software estatístico R. Com isto, acreditamos que o tema abordado nesta dissertação é analisado de forma ampla, incluindo aspectos teóricos, computacionais e aplicados.

1.4 Plataforma computacional

A linguagem de programação Ox, que foi criada por Jurgen Doornik em 1994, constitui a plataforma computacional utilizada no desenvolvimento desta dissertação de mestrado. Esta flexível linguagem permite a implantação de técnicas estatísticas com facilidade e atende a requisitos como precisão e eficiência. O apêndice desta dissertação contém um programa escrito em Ox que foi utilizado para obtenção dos resultados apresentados neste trabalho. Detalhes sobre esta linguagem de programação podem ser encontrados em Doornik (2001). Para a análise, manipulação e apresentação gráfica de dados, utilizamos o ambiente R, que é uma linguagem de programação de alto nível e tem a vantagem de ser distribuída gratuitamente. Para maiores detalhes sobre a plataforma R, ver Venables & Ripley (2002); ver também http://www.r-project.org.

Capítulo 2

Métodos bootstrap

2.1 Introdução

O método bootstrap, introduzido por Efron (1979), é um método de reamostragem baseado na construção de subamostras a partir de uma amostra inicial. Na verdade, trata-se tanto de uma alternativa para o processo inferencial como também de uma ferramenta de diagnóstico. É bastante útil quando se deseja avaliar, para um certo estimador, o seu erro padrão, o seu viés, ou ainda quando se quer estimar a distribuição de probabilidade do estimador. O método bootstrap pode ser construído paramétrica ou não-parametricamente. No bootstrap paramétrico, fazemos suposições distribucionais e reamostramos observações da distribuição postulada, mas usando os valores das estimativas dos parâmetros no processo de geração de pseudo-amostras. No bootstrap não-paramétrico, o processo de reamostragem se dá a partir da função de distribuição empírica dos dados (ou dos resíduos quando há uma estrutura de regressão).

Freedman (1981) e Wu (1986) discutem detalhadamente propriedades do método bootstrap em análises de regressão tratando de dois problemas, o primeiro sendo a determinação da precisão dos coeficientes estimados de regressão ou valores ajustados da resposta média e o segundo sendo o estudo da influência da seleção de variáveis ou do modelo sobre o viés de alguma medida do modelo ajustado. Efron e Tibshirani (1986) apresentam muitas aplicações do método bootstrap para procedimentos estatísticos, tais como séries temporais e dados censurados. Fisher e Hall (1989) mostram como obter regiões de confiança via bootstrap quando se utilizam dados circulares.

2.2 O método bootstrap

Considere uma amostra aleatória $y = (y_1, \ldots, y_n)$ cujos valores são realizações de variáveis aleatórias independentes e identicamente distribuídas $Y_1, \ldots, Y_n$, cada uma possuindo função de densidade de probabilidade (FDP) e função de distribuição acumulada (FDA) denotadas por f e F, respectivamente. A amostra é usada para realizar inferência sobre alguma característica da população, genericamente denotada por θ, através de uma estatística T cujo valor na amostra é t.

Há duas situações distintas para diferenciar o bootstrap paramétrico e o não-paramétrico. Quando há um modelo com constantes ajustáveis ou parâmetros ψ que determinam completamente f, tal modelo é chamado de paramétrico e métodos estatísticos baseados neste modelo são métodos paramétricos. Neste caso, o parâmetro de interesse θ é uma componente ou uma função de ψ. Quando nenhum modelo matemático deste tipo é usado, a análise estatística é não-paramétrica e usa apenas o fato de que as variáveis aleatórias Y_i's são independentes e identicamente distribuídas. Mesmo se houver um modelo paramétrico plausível, uma análise não-paramétrica pode ainda ser útil para avaliar a robustez das conclusões de uma análise paramétrica.

Um importante papel é desempenhado na análise não-paramétrica pela distribuição empírica, que coloca probabilidades iguais a n^{-1} em cada valor y_i da amostra. A estimativa usada de F é a função de distribuição empírica (FDE) $\widehat{F}$, que é definida como

$$\widehat{F}(y) = \frac{\#\{y_j \leq y\}}{n}.$$

Nota-se que o valor do salto da FDE no ponto y_i é a proporção de vezes em que y_i aparece na amostra. Se denotarmos essa proporção por f_i, e se, por exemplo, estivermos interessados em estimar a média, teremos $\widehat{\theta} = \sum_{i=1}^{n} f_i y_i$. As proporções f_i podem assumir valores $0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$, satisfazendo $\sum_{i=1}^{n} f_i = 1$.

De forma mais ampla, a estatística de interesse t é uma função simétrica de $y_1, \ldots, y_n$, significando que t não é afetada pelo reordenamento dos dados. Isto implica que t depende apenas dos valores ordenados $y_{(1)}, \ldots, y_{(n)}$ ou, equivalentemente, da FDE $\widehat{F}$. Freqüentemente isto pode ser expresso simplesmente como $t = t(\widehat{F})$, onde $t(\cdot)$ é uma função estatística — essencialmente é apenas uma expressão matemática do algoritmo para computar t a partir de $\widehat{F}$. Tal função estatística é de importância central no caso não-paramétrico porque também define a quantidade de interesse θ através de $\theta = t(F)$. Isto corresponde à idéia qualitativa de que θ é uma característica da população descrita por F. A mesma definição de θ se aplica em problemas paramétricos, onde θ é usualmente definido como um dos parâmetros em ψ.

A relação entre a estimativa t e $\widehat{F}$ pode ser geralmente expressa como $t = t(\widehat{F})$, correspondendo à relação $\theta = t(F)$ entre a característica de interesse e a distribuição. A função estatística $t(\cdot)$ é utilizada para representar a estimativa de θ baseada nos dados observados $y_1, \ldots, y_n$.

Suponha um modelo paramétrico particular para a distribuição dos dados $y_1, \ldots, y_n$. Usaremos $F_\psi(y)$ e $f_\psi(y)$ para denotar a FDA e a FDP, respectivamente. Quando ψ é estimado por $\widehat{\psi}$ — freqüentemente, mas não invariavelmente, pela sua estimativa de máxima verossimilhança — a substituição por $\widehat{\psi}$ no modelo resulta no modelo ajus-

tado, com FDA $\widehat{F}(y) = F_{\widehat{\psi}}(y)$, que pode ser usado para se obter conhecimento sobre propriedades de T, às vezes com exatidão.

A utilização do método bootstrap se justifica quando a teoria assintótica é intratável ou quando, apesar de viável, as aproximações assintóticas de primeira ordem são imprecisas para os tamanhos amostrais disponíveis. Quando, por exemplo, a teoria assintótica fornece uma aproximação imprecisa para a distribuição de uma estatística de teste, as diferenças entre o nível exato do teste (realizado com base em valores críticos assintóticos) e o nível nominal podem ser substanciais. A aplicação de bootstrap, neste caso, é de grande relevância, uma vez que o método pode reduzir consideravelmente, ou até mesmo eliminar, distorções de tamanho de testes estatísticos em amostras finitas (Espinheira, 2003). Segundo Horowitz (1997), procedimentos bootstrap simples fornecem aproximações melhoradas para a distribuição de estatísticas assintoticamente pivotais, mas não para a distribuição de estatísticas que não apresentam esta propriedade. Beran (1988) mostra que se a distribuição assintótica da estatística, sob a hipótese nula, é pivotal, então, sob algumas condições de regularidade, os tamanhos de testes bootstrap apresentam erros de ordem menor, i.e., erros cujas ordens convergem mais rapidamente para zero que as ordens dos erros dos testes baseados na teoria assintótica de primeira ordem.

2.3 Bootstrap em modelos de regressão

Bootstrap é um método que pode ser usado para avaliar a precisão de estimativas estatísticas baseado em simulações. O método bootstrap tipicamente produz uma aproximação para a distribuição da estatística de interesse que pode ser consideravelmente mais precisa do que sua aproximação assintótica de primeira ordem. A aplicação de bootstrap em modelos de regressão foi estudada em detalhes por Wu (1986). Mais recentemente, muitos autores têm investigado o uso deste método em econometria, entre eles estão Horowitz (1997), Jeong e Maddala (1993), Li e Maddala (1996) e Vinod (1993). Em sua forma mais simples, o procedimento bootstrap para o modelo de regressão linear $y = X\beta + e$ pode ser descrito como se segue:

1. Retire uma amostra aleatória $e_1^*, \ldots, e_n^*$ de $\widehat{e}$ com reposição.

2. Forme uma amostra bootstrap: $y^* = Xb + e^*$, onde $e^* = (e_1^*, \ldots, e_n^*)'$.

3. Calcule o estimador de MQO, $b^* = (X'X)^{-1}X'y^*$.

4. Repita os passos de 1 a 3 um grande número de vezes (digamos, B).

5. Calcule a variância dos B vetores das estimativas obtidas usando o esquema descrito nos passos de 1 a 4.

É comum multiplicar os resíduos no passo 1 pelo fator $\sqrt{n/(n-p)}$. O método bootstrap descrito acima é geralmente conhecido como *bootstrap não-paramétrico* e *não-ponderado* ou *bootstrap ingênuo não-paramétrico* ou *naïve*. O bootstrap paramétrico não-ponderado é semelhante ao descrito acima, mas com o passo 1 substituído por:

1a. Retire uma amostra aleatória $e_1^*, \ldots, e_n^*$ de uma distribuição normal com média zero e variância $\widehat{\sigma}^2 I_n$.

Nenhum dos esquemas bootstrap decritos acima leva em consideração a possibilidade de haver heteroscedasticidade. De fato, eles não são nem consistentes nem assintoticamente não-viesados sob heteroscedasticidade; ver Shao (1988) e Wu (1986). Um estimador robusto à presença de heteroscedasticidade foi proposto por Wu (1986) e pode ser descrito como se segue:

1b. Para cada i, $i = 1, \ldots, n$, retire t_i^* de uma distribuição com média zero e variância unitária.

2b. Forme a amostra bootstrap (y^*, X), onde $y_i^* = X_i b + t_i^* \widehat{e}_i / \sqrt{(1-h_i)}$, h_i sendo o i-ésimo elemento diagonal da 'matriz chapéu' $X(X'X)^{-1}X'$ e X_i sendo a i-ésima linha de X.

Os passos de 3 a 5 permanecem os mesmos. O mecanismo deste esquema de bootstrap consiste do seguinte: obtém-se um conjunto de n seleções independentes de uma distribuição com média zero e variância unitária; multiplica-se estes valores pelos resíduos; forma-se uma amostra bootstrap; estima-se β usando a nova amostra; repete-se o procedimento um grande número de vezes (digamos, B); então, usa-se as estimativas das B réplicas bootstrap para se obter uma estimativa para var(b). Note que a variância de $t_i^* \widehat{e}_i$ não é constante quando os erros originais não são homoscedásticos. Portanto, o esquema de bootstrap de Wu leva em consideração a não-constância potencial da variância do erro. Este esquema de reamostragem é freqüentemente chamado de *ponderado* ou *externo*.

Há diferentes maneiras de reamostrar t^*. Uma possibilidade é reamostrar de $a_1, \ldots, a_n$, onde

$$a_i = \frac{\widehat{e}_i - \overline{\widehat{e}_i}}{\sqrt{n^{-1}\sum_{t=1}^{n}(\widehat{e}_i - \overline{\widehat{e}_i})^2}}, \tag{2.1}$$

com $\overline{\widehat{e}_i} = n^{-1}\sum_{t=1}^{n}\widehat{e}_i$; quando o modelo de regressão especificado tem um intercepto sabemos que $\overline{\widehat{e}} = 0$. Esta é uma implementação não-paramétrica do esquema bootstrap de Wu, uma vez que a reamostragem é feita a partir da função de distribuição empírica dos resíduos normalizados. Em uma implementação paramétrica do bootstrap de Wu,

os t_i^*'s podem ser selecionados aleatoriamente de uma distribuição específica, e.g., a distribuição normal padrão.

Uma alternativa diferente é o bootstrap ponderado inversamente ajustado. A idéia é amostrar t_i^* no esquema de bootstrap ponderado não com probabilidade $1/n$, como no esquema de Wu, mas com probabilidade que é inversamente proporcional a h_i, como proposto por Cribari–Neto & Zarkos (2003). Seja

$$\delta = \sum_{i=1}^{n} 1/h_i.$$

Então, o esquema pode ser descrito da seguinte forma:

1c. Para cada i, $i = 1, \ldots, n$, retire t_i^*, de forma independente, de $a_1, \ldots, a_n$, ou seja, de (2.1). Aqui, a_i é selecionado com probabilidade p_i, $i = 1, \ldots, n$; ver (2.2) abaixo.

2c. Forme a amostra bootstrap (y^*, X), onde $y_i^* = X_i b + t_i^* \widehat{e_i}/\sqrt{(1-h_i)^{f_i}}$, $i = 1, \ldots, n$, onde

$$f_i = \begin{cases} 1, & \text{se} \quad h_i \leq 2p/n \\ 3, & \text{se} \quad h_i > 2p/n \end{cases}.$$

Seja $p_i = \gamma/h_i$. Então, $\sum_{i=1}^{n} p_i = 1$ implica

$$\sum_{i=1}^{n} \gamma h_i^{-1} = \gamma \sum_{i=1}^{n} h_i^{-1} = \gamma\delta = 1.$$

Logo,

$$p_i = \frac{1}{\delta h_i}. \tag{2.2}$$

O raciocínio lógico é que observações com alta alavancagem serão selecionadas no esquema de reamostragem com probabilidade mais baixa, mas mesmo assim positiva, desta forma reduzindo seu impacto sobre a inferência resultante. Isto é, quanto mais alto for o grau de alavancagem de uma observação, com menos freqüência seu correspondente resíduo será selecionado na reamostragem bootstrap. Ou seja, as probabilidades de reamostragem são controladas pelos diferentes graus de alavancagem das várias observações.

Capítulo 3

Proposta de um novo estimador e avaliação numérica

3.1 Introdução

O presente capítulo encontra-se dividido em três partes. A primeira parte propõe uma modificação do estimador HC4, que nós denominamos HC5 e que, tal como o estimador HC4, leva em consideração o impacto de observações com alta alavancagem em amostras finitas mas o estimador HC5 é o primeiro estimador na classe de estimadores consistentes da matriz de covariâncias de b a levar em consideração o nível máximo de alavancagem. A segunda parte do capítulo consiste de uma investigação numérica dos comportamentos em amostras finitas de estatísticas quasi-t baseadas nos diferentes estimadores das variâncias dos erros, através de simulações de Monte Carlo implementadas na linguagem de programação matricial `Ox` (Doornik, 2001). Com base em dois critérios de avaliação, viés relativo total dos diferentes estimadores e taxas de rejeição associadas a estatísticas quasi-t, a segunda parte deste capítulo tem por objetivo examinar o comportamento dos estimadores consistentes apresentados anteriormente, destacando aqueles que apresentam desempenho superior. Na terceira parte comparamos, através de avaliações numéricas, os desempenhos de testes quasi-t baseados nos estimadores denominados wu e invwu com os desempenhos de dois testes bootstrap, um que usa o esquema de reamostragem de bootstrap ponderado e um outro que utiliza o esquema de reamostragem ponderado com probabilidades de amostragem inversamente proporcionais às medidas de alavancagem. Os desempenhos dos testes são avaliados com base nas taxas de rejeição da hipótese nula, sendo a hipótese alternativa bicaudal e a geração dos dados realizada sob a hipótese em teste.

A avaliação numérica também será realizada em duas etapas. Na primeira etapa, o conjunto de dados utilizado consiste de preços de ações e preços ao consumidor no período pós-Segunda Guerra Mundial (até 1969) e foi extraído de Cagan (1974, p.4). Na segunda etapa, o conjunto de dados usado consiste de dados sobre despesas per capita em escolas públicas e renda per capita por estado nos Estados Unidos em 1979, estes dados tendo sido extraídos de Greene (1997).

3.2 Proposta de um novo estimador

Com o intuito de melhorar a confiabilidade de inferências realizadas a partir do estimador HC4, propomos a seguir um novo estimador, denominado HC5, modificando apenas a estrutura do expoente do termo $(1 - h_i)$ utilizado na definição de HC4. No estimador HC5, proposto aqui, usa-se

$$\alpha_i = \min\left\{\frac{h_i}{\overline{h}}, \max\left\{4, \frac{kh_{\max}}{\overline{h}}\right\}\right\}, \tag{3.1}$$

onde k é uma constante pré-especificada, $0 \leq k \leq 1$. Portanto, o novo estimador é definido da seguinte forma:

$$\text{HC5} = (X'X)^{-1}X'\widehat{\Psi}_5 X(X'X)^{-1},$$

onde

$$\widehat{\Psi}_5 = \text{diag}\left\{\frac{\widehat{e}_1^2}{(1-h_1)^{\alpha_1}}, \ldots, \frac{\widehat{e}_n^2}{(1-h_n)^{\alpha_n}}\right\}$$

e

$$\alpha_i = \min\left\{\frac{h_i}{\overline{h}}, \max\left\{4, \frac{kh_{\max}}{\overline{h}}\right\}\right\} = \min\left\{\frac{nh_i}{p}, \max\left\{4, \frac{nkh_{\max}}{p}\right\}\right\},$$

sendo $\overline{h} = n^{-1}\sum_{i=1}^{n} h_i$, i.e., $\overline{h}$ é a média dos h_i's, $h_{\max}$ sendo o h_i máximo. Note que $\overline{h} = \frac{p}{n}$, o que conduz à segunda igualdade acima.

O expoente controla o nível de desconto da observação e é determinado pela razão entre $h_{\max}$ e a média $\overline{h}$ dos h_i's e pela razão entre h_i e $\overline{h}$. Como $0 < (1 - h_i) < 1$ e $\alpha_i > 0$, segue que $0 < (1 - h_i)^{\alpha_i} < 1$. Se $h_i/\overline{h} \leq 4$ temos que $\alpha_i = h_i/\overline{h}$ e se $k = 0$ temos que o estimador HC5 se reduz ao estimador HC4 proposto por Cribari–Neto (2004). Se $k = 1$, temos também $\alpha_i = h_i/\overline{h} = nh_i/p$.

Note que o estimador HC5 possui uma diferença marcante em comparação com o estimador HC4 proposto por Cribari–Neto (2004): o novo estimador leva em consideração não apenas a razão entre a i-ésima medida de alavancagem (h_i) e o nível médio de alavancagem ($\overline{h}$), mas também o impacto que o *nível máximo* de alavancagem ($h_{\max}$) exerce sobre a inferência resultante.

De fato, o estimador HC5 é o primeiro estimador na classe de estimadores consistentes da matriz de covariâncias de b a incorporar em sua definição termos de desconto dos quadrados dos resíduos que se ajustam a variações no grau máximo de alavancagem.

Estes termos de desconto são adaptáveis, definindo-se apropriadamente o valor da constante k. Os resultados obtidos na avaliação numérica apresentados na Seção 3.3 sugerem que, com $k = 0.7$, este novo estimador tem comportamento superior a todos os

outros estimadores em amostras finitas no que diz respeito a testes quasi-t associados. Este valor foi escolhido após várias avaliações numéricas, mostrando-se adequado no que diz respeito à controlabilidade do erro tipo I de testes quasi-t associados em amostras de tamanho típico.

As Figuras 3.1 e 3.2 apresentam gráficos de α_i, expoente do termo $(1 - h_i)$ dos estimadores HC3, HC4 e HC5 versus a razão entre a i-ésima medida de alavancagem (h_i) e o nível médio de alavancagem $(\overline{h})$. Na Figura 3.1 consideramos um caso em que o estimador HC5 coincide com o estimador HC4 fazendo $kh_{\max}/\overline{h} = 3.89$, enquanto que na Figura 3.2 consideramos um caso mais extremo, onde $kh_{\max}/\overline{h} = 7.59$. Notamos que quando o nível máximo de alavancagem não é acentuado (Figura 3.1), o estimador HC5 coincide com o HC4. Por outro lado, quando a alavancagem máxima é extrema (Figura 3.2), o estimador HC5 difere do estimador HC4, sendo definido de forma que o valor máximo de α_i seja mais elevado, de forma a atenuar a tendência liberal dos testes associados.

O estimador HC5 proposto nesta seção poderia, alternativamente, ser definido usando

$$\alpha_i = \min\left\{\frac{h_i}{\overline{h}}, \frac{kh_{\max}}{\overline{h}}\right\}.$$

De fato, resultados de simulação não apresentados nesta dissertação sugerem que estas duas formas do estimador possuem propriedades similares em amostras finitas. Nós optamos pela (3.1) porque ela generaliza o estimador HC4 de Cribari–Neto (2004), ou seja, o estimador HC4 passa a ser um caso especial do estiamdor HC5.

3.3 Avaliação numérica

O experimento de Monte Carlo da primeira avaliação numérica é baseado no modelo de regressão linear

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n.$$

A covariável x_i corresponde à variação percentual anual dos preços ao consumidor em vinte países, no período posterior à Segunda Guerra, até 1969. Cada uma das observações de x é repetida duas, três, quatro e cinco vezes para formar amostras de tamanhos $n = 40, 60, 80, 100$, que são mantidas constantes em todo experimento. Isto garante que $\lambda = \max(\sigma_i^2)/\min(\sigma_i^2)$, uma medida do grau de heteroscedasticidade, mantém-se constante para os diferentes tamanhos amostrais, e que os resultados da simulação são apenas afetados pelo crescente número de observações. Em nossos estudos de simulação

Figura 3.1: Razão entre a i-ésima medida de alavancagem (h_i) e o nível médio de alavancagem ($\overline{h}$) versus α_i, expoente do termo $(1 - h_i)$ dos estimadores HC3, HC4 e HC5, quando $kh_{\max}/\overline{h} = 3.89$ (com $k = 0.7$).

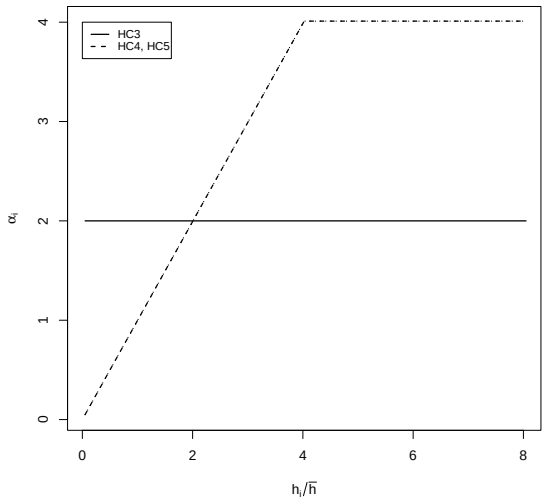
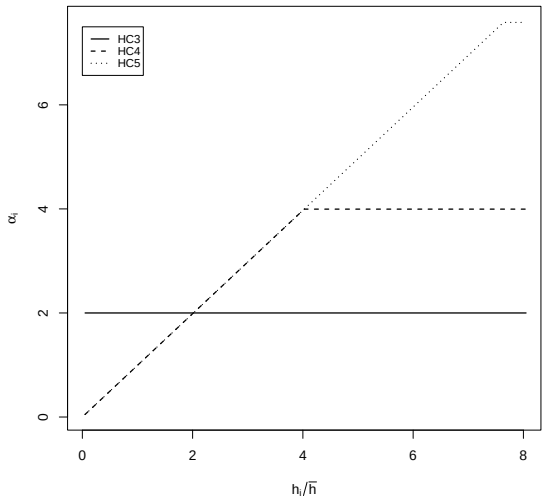


Figura 3.2: Razão entre a i-ésima medida de alavancagem (h_i) e o nível médio de alavancagem ($\overline{h}$) versus α_i, expoente do termo $(1 - h_i)$ dos estimadores HC3, HC4 e HC5, quando $kh_{\max}/\overline{h} = 7.59$ (com $k = 0.7$).



utilizamos $\beta_0 = 1$ e $\beta_1 = 0$. Sob homoscedasticidade $\lambda = 1$ e sob heteroscedasticidade o valor de λ depende da especificação da função cedástica, tendo-se $\lambda > 1$.

Nesta primeira etapa, consideramos dois conjuntos de dados; o primeiro é o conjunto de dados completos mencionado anteriormente, ao passo que, no segundo conjunto de dados, nós identificamos e retiramos os pontos de alta alavancagem. A função cedástica foi especificada como

$$\sigma_i^2 = \exp\{c_1 + c_2 x_i\}.$$

Simulações sob homoscedasticidade foram realizadas com $c_1 = c_2 = 0$. Para dados heteroscedásticos, obtivemos dois valores para λ; primeiro usamos $c_1 = c_2 = 0.110$, que resulta em um grau de heteroscedasticidade $\lambda = 14.48$, e depois usamos $c_1 = c_2 = 0.161$, que resulta em $\lambda = 50.01$ para o primeiro conjunto de dados ('dados1'). Para o segundo conjunto de dados ('dados2'), que não contém pontos de alavanca, também obtivemos dois valores para λ. Utilizamos $c_1 = c_2 = 0.430$, que produz $\lambda = 15.01$, e $c_1 = c_2 = 0.625$, que resulta em $\lambda = 51.29$. Os desempenhos dos vários estimadores da matriz de covariâncias considerados no experimento de Monte Carlo foram avaliados com base em: (a) tamanho: o tamanho do teste quasi-t associado da hipótese nula $\mathcal{H}_0 : \beta_1 = 0$ contra $\mathcal{H}_1 : \beta_1 \neq 0$; (b) viés relativo total: o viés relativo total é definido como a soma dos valores absolutos dos vieses relativos das as variâncias estimadas de b_0 e b_1. Isto é, para cada estimador foi estimado

$$\frac{\mid \mathrm{E}\{\widehat{\mathrm{var}}(\, b_0)\} - \mathrm{var}(\, b_0) \mid}{\mathrm{var}(\, b_0)} + \frac{\mid \mathrm{E}\{\widehat{\mathrm{var}}(\, b_1)\} - \mathrm{var}(b_1) \mid}{\mathrm{var}(\, b_1)},$$

onde '$\widehat{\mathrm{var}}$' denota o estimador de interesse. Os experimentos de simulação são baseados em 5000 réplicas de Monte Carlo e 500 réplicas de bootstrap (i.e., um total de 2.500.000 de réplicas por experimento).

Seis experimentos de Monte Carlo foram realizados, dois sob homoscedasticidade e quatro sob heteroscedasticidade. Como já foi mencionado anteriormente, o conjunto de dados usado é uma amostra de tamanho 20 retirada de Cagan (1974). Isto é o que chamamos de 'dados1'. O segundo conjunto de dados é obtido retirando uma observação identificada como ponto de alta alavancagem (seu h_i excede $3p/n$); esta observação corresponde ao Chile, que tem uma alavancagem muito alta $h_{\mathrm{chile}} = 0.929$ ao passo que $3p/n = 0.300$.

Os resultados das simulações relativos às taxas de rejeição dos diferentes testes quasi-t, aos níveis nominais de 10%, 5% e 1% (ou seja, as probabilidades de rejeição da hipótese nula quando esta de fato é verdadeira), sob homoscedasticidade e heteroscedasticidade, para o conjunto de dados originais de Cagan ('dados1') encontram-se no Quadro 3.1

(subdividido em 3.1a e 3.1b). O Quadro 3.2 (subdividido em 3.2a e 3.2b) apresenta resultados para o conjunto de dados sem alavancagem ('dados2'). Os principais resultados encontram-se resumidos a seguir.

Primeiro, o teste cuja estatística usa o estimador de mínimos quadrados ordinários da variância apresenta desempenho satisfatório, com ou sem presença de pontos de alta alavancagem, sob homoscedasticidade, como se esperava. Sob heteroscedasticidade, os percentuais de rejeição para o teste que utiliza este estimador não se aproximam dos níveis de significância assintóticos considerados. Por exemplo, para $n = 100$ e $\lambda = 50.01$ as taxas de rejeição do teste são iguais a 66.10%, 60.02%, 49.28% para os níveis nominais 10%, 5% e 1%, respectivamente.

Segundo, o teste construído a partir do estimador white também não apresenta comportamento confiável; tanto sob homoscedasticidade quanto sob heteroscedasticidade suas taxas de rejeição são consideravelmente altas. Para $n = 40$ e $\lambda = 14.48$, ao nível nominal de 5%, este teste rejeita a hipótese nula 36.22% das vezes.

Terceiro, o teste baseado no estimador HC3 apresenta taxas de rejeição menores que as taxas apresentadas pelo teste baseado no estimador white. É importante comentar que, apesar dos testes quasi-t associados ao estimador HC3 terem apresentado menores distorções de tamanho em relação ao teste baseado no estimador white, suas taxas de rejeição ficaram sempre acima dos níveis de significância correspondentes, sendo estes testes, portanto, liberais.

Quarto, sob homoscedasticidade, o teste baseado no estimador bootstrap naïve apresenta resultados muito bons, com tamanhos empíricos próximos aos níveis nominais. Sob heteroscedasticidade, contudo, o teste que usa o estimador bootstrap naïve é consideravelmente liberal, assim como teste baseado no estimador ols.

Quinto, os testes baseados nos estimadores ols, white, naïve e wu são consideravelmente mais liberais nos 'dados1' do que nos 'dados2'. Por exemplo, no primeiro caso, e para $n = 80$, os tamanhos correspondentes (no nível nominal de 5% e sob heteroscedasticidade, $\lambda = 14.48$) são 47.10%, 17.82%, 48.22% e 13.76%. No segundo caso ('dados2', $n = 76$ e $\lambda = 15.02$), os tamanhos empíricos correspondentes são 25.68%, 11.78%, 25.84% e 10.16%, respectivamente.

Sexto, os testes baseados nos estimadores HC5 e invwu apresentam tamanhos inferiores aos níveis nominais de significância (10% e 5%), sob homoscedasticidade e heteroscedasticidade, para $n = 20$ e $n = 40$. O teste baseado no estimador HC5 tem desempenho melhor do que o teste baseado em invwu quando os dados não contêm observações influentes, o inverso acontecendo quando os dados têm pontos influentes.

Sétimo, os testes baseados no estimadores HC3 e wu apresentam desempenho simi-

lares, sob homoscedasticidade e heteroscedasticidade na presença e ausência dos pontos de alavanca, exceto para $n = 20$ e $\lambda = 50.01$.

Oitavo, os testes quasi-t baseados nos estimadores white, HC3 e HC4 são liberais; os testes baseados no estimadores HC5 e invwu são tipicamente os mais confiáveis.

A Figura 3.3 mostra quatro gráficos correspondentes aos 'dados1' (dados com pontos de alta alavancagem); os dois gráficos superiores são relativos a erros homoscedásticos, ao passo que os gráficos inferiores são para o caso de heteroscedasticidade, $\lambda = 14.48$. Os tamanhos das amostras são $n = 40, 80$. Apresentamos as diferenças entre quantis exatos (estimados por simulação) e quantis assintóticos (da distribuição normal padrão) contra quantis assintóticos de quatro estatísticas de teste selecionadas: aquelas que utilizam em suas formulações HC3, HC4, HC5 e invwu. Quanto mais próximo as linhas estiverem da linha horizontal (sólida) desenhada em zero, mais confiável é a inferência. A Figura 3.3 claramente revela que as estatísticas de teste baseadas em HC5 e invwu têm distribuição nula em amostras finitas que são tipicamente melhor aproximadas pela distribuição nula assintótica. Esta aproximação funciona bem para os testes baseados nestes estimadores especialmente entre -2 e 2 quando $n = 80$. Podemos observar ainda que o teste baseado no estimador HC4 apresenta menores distorções de tamanho do que aquele que usa HC3.

Os Quadros 3.3 e 3.4 apresentam os vieses relativos totais dos seguintes estimadores: ols, white, HC3, HC4, HC5, naïve, wu e invwu. Estes resultados, obtidos por simulação correspondem respectivamente a 'dados1' e 'dados2'. Os resultados nos Quadros 3.3 e 3.4 levam a várias conclusões interessantes.

Primeiro, o estimador ols funciona muito bem sob homoscedasticidade, mas é bastante viesado quando as variâncias dos erros não são iguais. Sob heteroscedasticidade, para todos os tamanhos amostrais considerados, seu viés relativo total excede 100%.

Segundo, o estimador white apresenta vieses relativamente grandes em quase todos os casos. Por exemplo, para $n = 19$ e $\lambda = 51.29$ (Quadro 3.4) seu viés relativo total excede 100%.

Terceiro, o estimador HC3 mostra-se superior ao estimador white, tanto sob homoscedasticidade quanto sob heteroscedasticidade. Para exemplificar, notamos que para $n = 40$ e $\lambda = 14.48$ (Quadro 3.3), o viés relativo total do estimador white está em torno de 89%, enquanto que o estimador HC3 apresenta viés relativo total de apenas 8%.

Quarto, os estimadores HC4 e HC5 mostram-se bastante viesados quando os dados incluem pontos de alta alavancagem ('dados1'). Por exemplo, quando o tamanho da amostra é 20 e o grau de heteroscedasticidade é 14.48, o viés relativo total do estimador HC4 excede 100%, enquanto que, sob as mesmas condições, mas sem os pontos de alta alavancagem nos dados, seu viés relativo total está em torno de 43%. O viés relativo total

do estimador HC5 excede 100% em quase todos os casos, exceto sob homoscedasticidade e quando n é 80 ou 100.

Quinto, em relação aos estimadores bootstrap naïve e wu, o último mostra-se, em quase todos os casos, superior ao primeiro. Para exemplificar, notamos que para $n = 57$ e $\lambda = 15.01$, o viés relativo total do estimador naïve excede 100%, enquanto que o estimador wu apresenta viés relativo total de apenas 12%.

Sexto, o estimador invwu se mostra bastante viesado no caso de dados com pontos de alta alavancagem ('dados1'). Por exemplo, considerando $n = 40$ e $\lambda = 50.01$, o viés relativo total deste estimador fica em torno de 68%, ao passo que sem os pontos de alavancagem ('dados2') este valor é apenas 3%.

Os Quadros 3.5 e 3.6 apresentam os quatro primeiros momentos amostrais (média, variância, assimetria e curtose) das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu e invwu para os dados com os pontos de alavanca ('dados1') e para os dados sem os pontos de alavanca ('dados2').

A estatística de teste baseada no estimador HC5 apresenta uma menor dispersão na presença dos pontos de alavanca, seguida da que usa o estimador invwu, sob homoscedasticidade e heteroscedasticidade. Por exemplo, para $n = 20$ e $\lambda = 14.48$, a variância da estatística de teste baseada no estimador HC5 é 1.208, que é aproximadamente dezessete vezes menor que a variância da estatística de teste baseada no estimador HC3. Quando não há os pontos de alavanca nos dados, temos que a menor variância é da estatística de teste baseada no estimador invwu, exceto quando $n = 94$ e $\lambda = 15.01$. Para exemplificar, considere $n = 19$ e $\lambda = 15.01$; a variância de estatística de teste baseada no estimador invwu é 1.213, enquanto que a estatística do teste baseada no estimador HC5, que coincide com o estimador HC4, é 1.980. Sob homoscedasticidade e heteroscedasticidade, a estatística de teste baseada no estimador white apresenta a maior variância, tanto na presença como na ausência dos pontos de alavanca. Por exemplo, sob heteroscedasticidade ($\lambda = 14.48$) e $n = 20$, a variância da estatística de teste baseada no estimador white é 61.079, enquanto que a variância de estatística do teste baseada no estimador HC4 é 6.264.

Verificando o grau de afastamento da simetria, isto é, assimetria, notamos que em muitos casos as distribuições das estatísticas de teste baseadas nos estimadores são assimétricas negativas (à esquerda), ou seja, as distribuições apresentam cauda mais alongada à esquerda da ordenada máxima do que à direita. Em relação à curtose (grau de achatamento da distribuição) podemos notar que em todos os casos as distribuições das estatísticas de teste baseadas nos estimadores apresentados são leptocúrticas, ou seja, revelam um alto grau de afilamento, superior ao da distribuição normal padrão. As

maiores curtoses são apresentadas pela estatística de teste baseada no estimador HC5; por exemplo, sob heteroscedasticidade ($\lambda = 14.48$) e $n = 20$ a sua curtose é 108.855, que é aproximadamente quatorze vezes maior que a curtose da estatística de teste baseada no estimador HC3.

Com o propósito de ilustrar algumas conclusões anteriores, na Figura 3.4 apresentamos os gráficos das densidades estimadas de duas estatísticas de teste (aquelas baseadas em HC3 e HC5), sob heteroscedasticidade ($\lambda = 14.48$), para $n = 20, 100$. Através destes gráficos podemos notar que a distribuição da estatística de teste baseada no estimador HC5 revela um alto grau de afilamento, superior ao da distribuição normal padrão ao passo que a distribuição da estatística que usa HC3 apresenta dispersão substancialmente maior que a da distribuição assintótica de referência. Para $n = 100$ percebemos que a distribuição da estatística do teste baseado no estimador HC3 se encontra mais próxima da distribuição normal padrão.

O experimento de Monte Carlo da segunda avaliação numérica é baseado no modelo de regressão linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i.$$

As covariáveis, x_1 e x_2, são a renda per capita (reescalonada por 10^{-4}) e o quadrado da renda per capita, respectivamente, com um total de 50 observações. Cada uma das observações de x_1 e x_2 é repetida duas e três vezes para formar amostras de observações de tamanhos $n = 100, 150$, que são mantidas constantes ao longo do experimento. Em nossos estudos de simulação utilizamos $\beta_0 = \beta_1 = 1$ e $\beta_2 = 0$.

A exemplo do que foi feito anteriormente, utilizaremos dois conjuntos de dados; o primeiro é o conjunto de dados retirado de Greene (1997) e o segundo conjunto é obtido retirando dos dados os pontos de alta alavancagem. A função cedástica foi especificada como

$$\sigma_i^2 = \exp\{c_1 + c_2 x_{i1}\}.$$

Simulações sob homoscedasticidade foram realizadas com $c_1 = c_2 = 0$, ao passo que para dados heteroscedásticos consideramos dois valores para λ; primeiro usamos $c_1 = c_2 = 5.30$, que resulta em $\lambda = 15.04$, depois usamos $c_1 = c_2 = 0.93$, que resulta em $\lambda = 50.05$ para o primeiro conjunto de dados ('dados3'). Para o segundo conjunto de dados ('dados4'), também consideramos dois valores para λ: utilizamos $c_1 = 0$ e $c_2 = 7.65$, que produz $\lambda = 15.02$, e $c_1 = 0$ e $c_2 = 13.45$, que resulta em $\lambda = 50.30$.

Quadro 3.1a: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 20, 40, 60$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'.

	$n = 20$								
λ	1.00			14.48			50.01		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	11.88	6.48	1.74	61.10	54.50	42.60	76.34	72.48	64.32
white	52.98	45.58	33.24	83.02	79.68	73.40	90.24	88.48	88.38
HC3	27.96	22.02	14.38	61.68	54.08	43.42	33.98	29.02	22.30
HC4	8.32	6.52	4.26	19.40	15.70	11.58	20.48	16.40	11.70
HC5	1.18	1.00	0.56	3.68	2.86	2.34	3.56	3.18	2.20
naïve	12.00	6.58	1.80	61.20	54.64	42.62	76.50	72.50	64.32
wu	28.06	22.20	14.48	61.86	54.18	43.58	74.16	67.92	56.18
invwu	1.74	1.20	0.70	5.08	4.12	2.84	4.88	4.02	3.16
	$n = 40$								
λ	1.00			14.48			50.01		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.98	5.80	1.36	56.36	48.92	36.10	69.76	64.40	54.54
white	30.72	24.20	15.28	41.25	36.22	27.28	43.18	37.70	29.52
HC3	23.32	17.98	11.00	32.64	27.30	20.64	33.98	29.02	22.30
HC4	17.04	13.06	7.80	24.60	20.42	14.66	25.84	21.78	16.16
HC5	10.98	8.00	4.28	16.22	13.22	9.10	17.50	14.50	10.44
naïve	10.94	5.72	1.50	56.58	48.86	36.34	69.92	64.48	54.52
wu	23.50	18.00	11.06	32.56	27.44	20.60	34.02	29.06	22.36
invwu	11.60	8.54	4.50	18.16	14.32	10.16	20.48	16.40	11.70
	$n = 60$								
λ	1.00			14.48			50.01		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	11.28	5.70	1.06	55.84	48.60	36.32	67.94	62.36	52.10
white	24.10	17.98	7.06	30.26	23.42	15.96	30.56	24.44	16.46
HC3	19.38	13.94	5.38	23.50	18.32	11.60	24.20	18.80	12.24
HC4	15.38	10.76	3.66	18.36	14.14	8.44	18.64	14.36	8.90
HC5	10.64	6.84	2.44	12.84	9.26	5.46	13.12	9.70	5.74
naïve	11.50	5.96	1.10	55.84	48.54	36.16	67.92	62.30	54.04
wu	19.36	14.08	5.50	23.44	18.38	11.70	24.20	18.86	12.30
invwu	11.10	7.16	2.54	15.14	10.98	6.12	19.56	13.84	7.52

Quadro 3.1b: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 80, 100$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados1'.

	$n = 80$								
λ	1.00			14.48			50.01		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.84	5.54	1.06	55.56	47.90	35.74	67.16	61.38	50.60
white	20.22	13.86	7.06	23.74	17.82	10.12	23.90	17.88	10.32
HC3	16.32	11.02	5.38	19.20	13.70	7.84	19.48	13.98	7.96
HC4	13.00	8.42	3.66	15.00	10.48	5.72	15.32	10.66	5.88
HC5	9.44	6.22	2.44	10.84	7.54	3.64	11.00	7.62	3.82
naïve	10.80	5.52	1.10	55.58	48.22	35.88	67.18	61.36	50.60
wu	16.56	11.06	5.50	19.30	13.76	7.84	19.66	14.12	8.04
invwu	9.82	6.34	2.54	14.02	9.62	4.50	20.24	14.20	6.74
	$n = 100$								
λ	1.00			14.48			50.01		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.30	5.72	1.12	54.92	47.42	34.54	66.10	60.02	49.28
white	18.12	11.92	5.02	21.00	14.56	7.58	21.36	14.90	7.74
HC3	15.02	9.68	3.80	16.96	11.84	5.34	17.08	11.92	5.62
HC4	12.30	7.58	3.04	13.92	9.12	4.12	14.08	9.14	4.32
HC5	9.18	5.26	2.10	10.26	6.48	2.80	10.28	6.46	2.86
naïve	10.50	5.58	1.14	54.98	47.16	34.56	66.20	60.32	49.16
wu	15.06	9.58	3.88	17.38	11.68	5.40	17.56	11.86	5.62
invwu	9.64	5.42	2.26	14.50	8.84	3.92	22.02	14.84	6.26

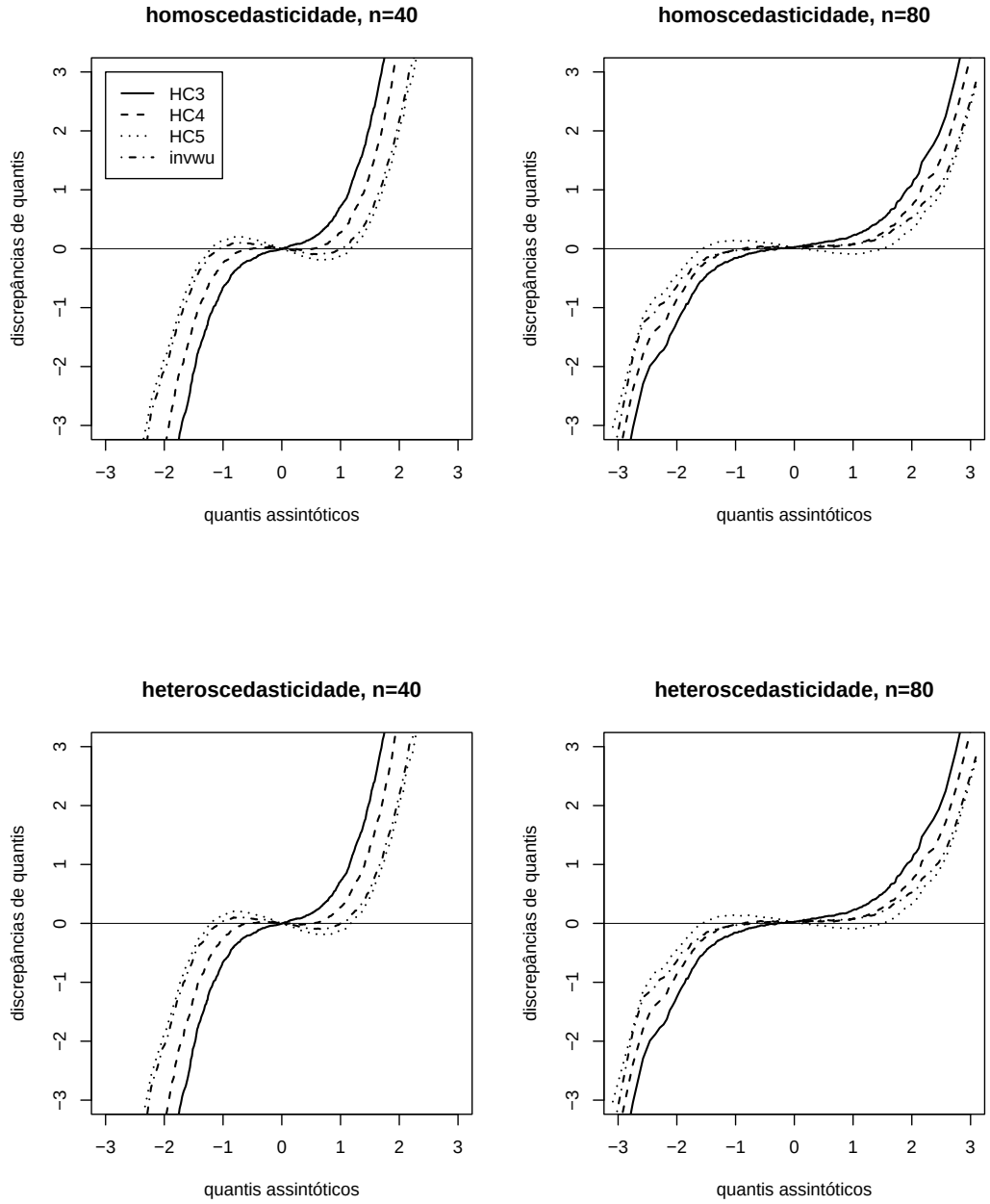
Quadro 3.2a: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 19, 38, 57$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (sem pontos de alavanca) - 'dados2'.

	$n = 19$								
λ	1.00			15.02			51.29		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	12.20	6.80	1.98	38.00	29.52	14.76	53.98	45.46	29.40
white	23.46	16.94	8.78	42.22	32.56	18.70	50.88	38.84	21.42
HC3	17.70	12.24	5.56	28.22	20.26	10.38	30.14	20.98	10.36
HC4	14.00	9.60	4.30	18.18	12.56	6.64	16.22	11.38	5.72
HC5	14.00	9.60	4.30	18.18	12.56	6.64	16.22	11.38	5.72
naïve	12.34	6.90	1.98	38.04	29.50	14.92	53.82	45.34	29.36
wu	17.88	12.28	5.56	28.22	20.52	10.42	30.34	20.88	10.34
invwu	9.02	5.94	2.20	10.08	7.46	3.68	8.46	5.86	2.98
	$n = 38$								
λ	1.00			15.02			51.29		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	11.02	5.56	1.30	35.10	26.40	14.00	48.34	39.86	26.12
white	17.08	11.04	4.04	27.38	19.40	9.80	31.04	22.84	11.36
HC3	14.40	8.94	3.10	21.64	14.92	6.58	24.88	17.10	7.28
HC4	12.32	7.54	2.64	16.84	11.48	4.64	19.36	11.98	4.28
HC5	12.32	7.54	2.64	16.84	11.48	4.64	19.36	11.98	4.28
naïve	10.92	5.66	1.28	34.98	26.24	14.08	48.16	39.94	26.04
wu	14.38	8.94	3.10	21.72	14.86	6.60	25.18	17.12	7.26
invwu	10.02	5.36	1.66	13.54	8.40	2.92	16.74	9.46	2.88
	$n = 57$								
λ	1.00			15.02			51.29		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	11.08	5.92	1.20	34.74	25.80	13.74	46.92	39.08	24.70
white	15.64	9.20	3.24	21.58	14.84	6.62	23.22	16.46	7.74
HC3	13.44	7.68	2.42	18.32	12.10	4.98	19.50	13.80	5.60
HC4	12.10	6.92	2.14	15.66	10.06	3.82	16.46	11.14	3.84
HC5	12.10	6.92	2.14	15.66	10.06	3.82	16.46	11.14	3.84
naïve	11.06	5.88	1.20	34.66	25.96	13.74	46.96	38.92	27.74
wu	13.74	7.84	2.46	18.36	12.08	5.00	19.46	13.66	5.78
invwu	10.00	5.70	1.54	13.72	8.54	2.90	17.48	11.10	3.80

Quadro 3.2b: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n =$ 19, 38, 57. Os níveis nominais considerados são α=0.10,0.05,0.01 (sem pontos de alavanca) - 'dados2'.

	$n = 76$								
λ	1.00			15.02			51.29		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	9.96	5.18	1.22	34.18	25.68	13.22	46.10	37.70	24.54
white	12.98	7.74	2.34	18.68	11.78	4.60	20.38	13.10	5.20
HC3	11.92	6.72	1.94	16.38	10.04	3.66	17.16	11.02	4.08
HC4	11.02	5.98	1.66	14.06	8.24	2.82	14.78	8.68	3.22
HC5	11.02	5.98	1.66	14.06	8.24	2.82	14.78	8.68	3.22
naïve	9.92	5.18	1.10	34.34	25.84	13.32	46.40	37.88	24.42
wu	11.98	6.72	2.04	16.42	10.16	3.78	17.48	10.98	4.14
invwu	9.64	4.86	1.34	13.82	7.82	2.66	17.68	10.64	3.66
	$n = 95$								
λ	1.00			15.02			51.29		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	9.90	5.06	1.04	33.98	25.50	13.56	45.32	37.72	23.86
white	12.54	7.02	1.90	16.52	10.22	3.66	17.94	11.04	4.08
HC3	11.38	6.14	1.58	14.64	8.86	3.04	15.86	9.32	3.20
HC4	10.48	5.50	1.46	12.86	7.56	2.50	13.46	7.98	2.64
HC5	10.48	5.50	1.46	12.86	7.56	2.50	13.46	7.98	2.64
naïve	10.20	4.94	1.10	34.10	25.46	13.30	45.56	37.36	24.14
wu	11.26	6.28	1.72	14.50	9.16	3.08	15.82	9.52	3.26
invwu	9.60	4.86	1.22	13.62	7.58	2.36	17.46	10.74	3.50

Figura 3.3: Discrepâncias de quantis para o conjunto de 'dados1' (com pontos de alavancagem). Diferenças entre quantis exatos (estimados por simulação) e quantis assintóticos (computados da distribuição $\mathcal{N}(0,1)$) dispostos graficamente contra quantis assintóticos de quatro estatísticas de teste: HC3, HC4, HC5 e invwu.



Quadro 3.3: Vieses relativos totais dos estimadores ols, white, HC3, HC4, HC5, naïve, wu, invwu para dados com alavancagem - 'dados1'.

n	λ	ols	white	HC3	HC4	HC5	naïve	wu	invwu
20	1	0.009	1.120	0.040	14.817	459.316	0.038	0.070	235.657
	14.48	1.624	1.790	1.502	2.488	121.812	1.631	1.509	60.541
	50.01	1.853	1.914	1.725	0.944	80.676	1.856	1.727	38.625
40	1	0.001	0.554	0.007	0.993	3.576	0.013	0.012	2.794
	14.48	1.528	0.888	0.082	1.410	5.263	1.532	0.084	1.967
	50.01	1.741	0.950	0.096	1.496	5.600	1.744	0.096	0.676
60	1	0.003	0.374	0.002	0.501	1.489	0.010	0.004	1.231
	14.48	1.500	0.597	0.050	0.736	2.250	1.501	0.050	0.552
	50.01	1.708	0.639	0.057	0.784	2.400	1.709	0.057	0.326
80	1	0.000	0.277	0.002	0.341	0.933	0.004	0.003	0.786
	14.48	1.485	0.445	0.029	0.507	1.421	1.486	0.031	0.139
	50.01	1.691	0.476	0.034	0.540	1.516	1.691	0.035	0.616
100	1	0.002	0.219	0.005	0.262	0.681	0.002	0.007	0.583
	14.48	1.476	0.350	0.015	0.393	1.043	1.476	0.014	0.063
	50.01	1.680	0.374	0.018	0.419	1.114	1.680	0.017	0.763

Quadro 3.4: Vieses relativos totais dos estimadores ols, white, HC3, HC4, HC5, naïve, wu, invwu para dados sem alavancagem - 'dados2'.

n	λ	ols	white	HC3	HC4	HC5	naïve	wu	invwu
19	1	0.009	0.617	0.015	1.073	1.073	0.016	0.009	3.460
	15.01	1.276	1.312	0.749	0.431	0.431	1.279	0.750	2.454
	51.29	1.582	1.440	0.904	0.255	0.255	1.586	0.905	1.759
38	1	0.002	0.303	0.011	0.321	0.321	0.002	0.010	0.758
	15.01	1.223	0.644	0.203	0.380	0.380	1.224	0.206	0.505
	51.29	1.523	0.707	0.248	0.375	0.375	1.523	0.250	0.030
57	1	0.004	0.203	0.005	0.186	0.186	0.003	0.004	0.416
	15.01	1.210	0.430	0.119	0.247	0.247	1.210	0.119	0.146
	51.29	1.506	0.471	0.144	0.250	0.250	1.506	0.144	0.276
76	1	0.001	0.148	0.009	0.137	0.137	0.001	0.006	0.288
	15.01	1.201	0.317	0.078	0.189	0.189	1.201	0.081	0.002
	51.29	1.496	0.348	0.095	0.193	0.193	1.496	0.100	0.399
94	1	0.001	0.119	0.006	0.105	0.105	0.002	0.003	0.213
	15.01	1.197	0.253	0.059	0.151	0.151	1.198	0.059	0.080
	51.29	1.491	0.277	0.072	0.154	0.154	1.491	0.070	0.468

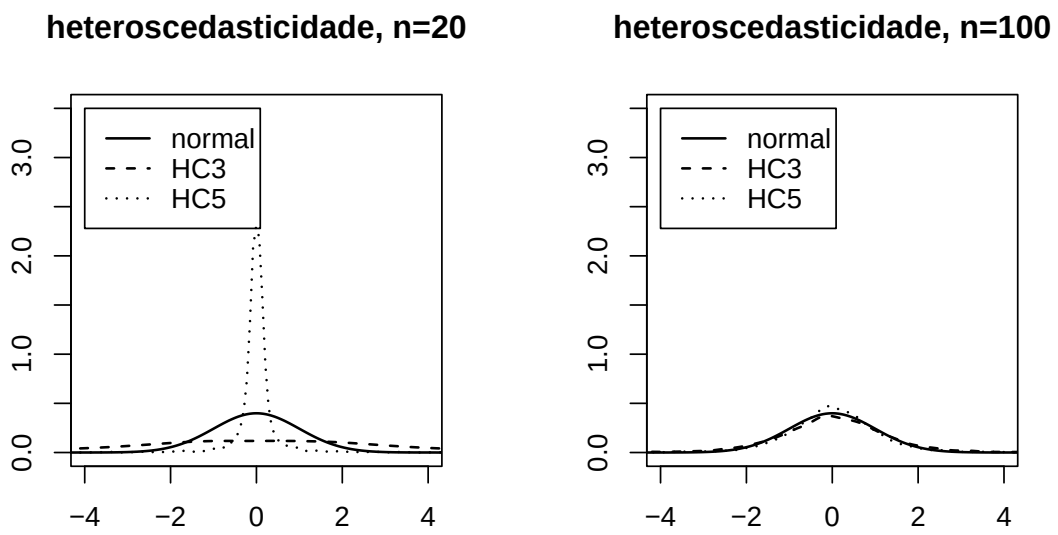
Quadro 3.5: Momentos amostrais: média, variância, assimetria e curtose das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu, invwu para dados com alavancagem - 'dados1'.

	λ	1.00				14.48			
n	teste	média	variância	assimetria	curtose	média	variância	assimetria	curtose
20	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	−0.005	8.560	−0.195	4.599	0.006	61.079	−0.044	3.626
	HC3	−0.007	3.682	−0.309	7.508	−0.013	21.270	0.027	7.565
	HC4	−0.003	1.083	−0.265	17.136	−0.003	6.264	0.347	22.797
	HC5	−0.003	0.183	0.998	98.908	−0.000	1.208	0.956	108.855
	wu	−0.007	3.740	−0.285	7.549	−0.016	21.537	0.012	7.546
	invwu	−0.003	0.243	0.212	63.279	−0.000	1.534	0.782	75.220
100	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.024	1.684	−0.054	4.194	0.022	2.159	−0.094	5.658
	HC3	0.022	1.440	−0.053	4.316	0.020	1.774	−0.092	5.738
	HC4	0.020	1.230	−0.051	4.454	0.018	1.457	−0.090	5.821
	HC5	0.018	0.999	−0.048	4.622	0.016	1.135	−0.086	5.912
	wu	0.023	1.441	−0.046	4.265	0.021	1.778	−0.086	5.671
	invwu	0.019	1.022	−0.040	4.501	0.018	1.433	−0.072	4.843

Quadro 3.6: Momentos amostrais: média, variância, assimetria e curtose das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu, invwu para dados sem alavancagem - 'dados2'.

	λ	1.00				15.01			
n	teste	média	variância	assimetria	curtose	média	variância	assimetria	curtose
19	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.014	2.287	−0.083	4.405	0.020	4.220	−0.037	3.574
	HC3	0.011	1.717	−0.072	4.816	0.018	2.801	−0.016	4.239
	HC4	0.009	1.398	−0.054	5.633	0.018	1.980	0.015	5.643
	HC5	0.009	1.398	−0.054	5.633	0.018	1.980	0.015	5.643
	wu	0.011	1.723	−0.084	4.830	0.019	2.808	−0.004	4.256
	invwu	0.007	0.949	−0.019	6.236	0.015	1.213	0.018	7.000
94	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.008	1.173	0.034	3.303	0.013	1.461	0.044	3.496
	HC3	0.008	1.101	0.034	3.327	0.013	1.325	0.043	3.533
	HC4	0.008	1.051	0.034	3.361	0.012	1.208	0.041	3.582
	HC5	0.008	1.051	0.034	3.361	0.012	1.208	0.041	3.582
	wu	0.008	1.106	0.045	3.343	0.013	1.332	0.048	3.547
	invwu	0.007	0.982	0.034	3.300	0.012	1.223	0.043	3.328

Figura 3.4: Densidades estimadas das estatísticas de teste baseadas nos estimadores HC3 e HC5 para o conjunto 'dados1' (com pontos de alavancagem) e densidade da distribuição $\mathcal{N}(0,1)$. A estimação das densidades foi feita usando um kernel gaussiano.



Os desempenhos dos vários estimadores da matriz de covariâncias considerados no experimento de Monte Carlo foram avaliados com base em: (a) tamanho do teste quasi-t associado para o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra $\mathcal{H}_1 : \beta_2 \neq 0$; (b) viés relativo total.

Seis experimentos de Monte Carlo foram realizados, dois sob homoscedasticidade e quatro sob heteroscedasticidade. Como mencionado anteriormente, o conjunto de dados usado é uma amostra de tamanho 50 retirada de Greene (1997); isto é o que chamamos de 'dados3'. O segundo conjunto de dados é obtido retirando três observações identificadas como pontos de alta alavancagem (seus h_i's excedem $3p/n$); estas observações correspondem aos estados do Alasca, Mississipi e Washington, D.C.. Em particular, o Alasca é um ponto de alta alavancagem, pois h_{Alasca}=0.651 ao passo que $3p/n = 0.180$.

Os resultados relativos às taxas de rejeição dos diferentes testes quasi-t, aos níveis nominais de 10%, 5% e 1%, sob homoscedasticidade e heteroscedasticidade, para o conjunto de dados originais de Greene ('dados3') encontram-se apresentados no Quadro 3.7. O Quadro 3.8 apresenta resultados para o conjunto de dados sem alavancagem ('dados4'). Os principais resultados podem ser resumidos como se segue. Primeiro, o teste que usa o estimador ols apresenta desempenho satisfatório apenas sob homoscedasticidade. Sob

heteroscedasticidade, os percentuais de rejeição para o teste não se aproximam do nível assintótico considerado; por exemplo, para $n = 150$ e ao nível assintótico de 5%, este teste apresenta rejeição de 4.76% quando $\lambda = 1.00$; na presença de forte heteroscedasticidade ($\lambda = 50.05$) e de pontos influentes, o mesmo teste apresenta rejeição de 34.72%.

Segundo, o teste construído com base no estimador white não apresenta comportamento confiável tanto sob homoscedasticidade quanto sob heteroscedasticidade; suas taxas de rejeição são consideravelmente superiores àquelas verificadas para os testes associados aos demais estimadores. O teste que usa o estimador wu apresenta taxas de rejeição superiores às esperadas, porém inferiores às taxas de rejeição apresentadas pelo teste que usa o estimador white.

Terceiro, o teste baseado no estimador HC3 apresenta taxas de rejeição inferiores às do teste baseado no estimador white, sendo estas taxas, contudo, superiores aos níveis nominais selecionados.

Quarto, sob homoscedasticidade, o teste baseado no estimador bootstrap naïve apresenta bons resultados; sob heteroscedasticidade, todavia, este teste é liberal. Por exemplo, ao nível de 5% para $\lambda = 1$ e $n = 100$, a taxa de rejeição é de 5.40%, enquanto que para $\lambda = 50.05$ a taxa de rejeição é de 35.58%.

Quinto, os testes baseados no estimadores HC3 e wu apresentam desempenho similares, sob homoscedasticidade e heteroscedasticidade na presença e ausência dos pontos de alavanca, por exemplo, para $n = 50$, $\lambda = 50.05$ e ao nível de 5%, o teste baseado no estimador HC3 apresenta taxa de rejeição de 24.18%, enquanto que o teste baseado no estimador wu rejeita a hipótese nula 24.14% das vezes.

Sexto, quando pontos de alta alavancagem são introduzidos nos dados, as distorções de tamanho dos testes baseados nos diversos estimadores crescem substancialmente. A presença destes pontos exerce um efeito negativo sobre os desempenhos dos estimadores avaliados quando estes são utilizados conjuntamente com testes quasi-t, elevando as taxas de rejeição; a exceção é o estimador HC5. Por exemplo, o teste baseado no estimador HC5, aos níveis nominais considerados (10%, 5% e 1%), para $n = 50$ e $\lambda = 50.05$, apresenta taxas de rejeição 7.58%, 5.74% e 3.26%, enquanto que sem considerar as observações influentes ($n = 47$) e com $\lambda = 50.30$, temos taxas de rejeição maiores que as anteriores: 14.04%, 8.58%, 3.08%, respectivamente.

Sétimo, o teste baseado no estimador invwu apresenta taxas de rejeição próximas aos níveis nominais tanto na presença quanto na ausência das observações influentes. O teste que usa o estimador HC4 tem comportamento melhor na ausência das observações influentes. Por exemplo, aos níveis nominais considerados (10%, 5% e 1%), para $n = 100$ e $\lambda = 15.04$, as taxas de rejeição são 15.42%, 10.36% e 4.62%, enquanto que sem

considerar as observações influentes ($n = 94$) e com $\lambda = 15.02$, temos taxas de rejeição menores que as anteriores: 11.56%, 6.22%, 1.64%, respectivamente.

Com o propósito de ilustrar algumas conclusões anteriores, foram construídos gráficos das diferenças entre quantis exatos de algumas estatísticas de teste, estimados por simulação, e quantis assintóticos para $n = 50, 100$, sob homoscedasticidade e sob heteroscedasticidade. As discrepâncias de quantis são plotadas contra quantis assintóticos. Estes gráficos estão apresentados na Figura 3.5. Quanto mais próximas as linhas estiverem da linha horizontal (zero), melhor será a aproximação pela distribuição assintótica $\mathcal{N}(0, 1)$ da distribuição da estatística de teste sob a hipótese nula. A análise dessas figuras evidencia que os testes baseados nos estimadores HC5 e invwu se comportam muito bem. De fato, nota-se que as distribuições exatas, em geral, são melhor aproximadas pela distribuição limite do que as distribuições exatas das demais estatísticas.

Os Quadros 3.9 e 3.10 ('dados3' e 'dados4', respectivamente) apresentam os vieses relativos totais dos seguintes estimadores: ols, white, HC3, HC4, HC5, naïve, wu e invwu. Os resultados nos Quadros 3.9 e 3.10 conduzem a várias conclusões interessantes. Primeiro, o estimador ols sob homoscedasticidade funciona muito bem, mostrando-se, todavia, bastante viesado com o aumento do grau de heteroscedasticidade. Segundo, o estimador wu apresenta vieses reduzidos quando há homoscedasticidade; à medida que o grau de heteroscedasticidade aumenta, seu viés relativo total aumenta. Para exemplificar, considere $n = 100$; sob homoscedasticidade ($\lambda = 1$) o viés relativo total deste estimador é praticamente nulo, sob heteroscedasticidade seu viés relativo total passa a ser 19.6% ($\lambda = 15.04$) e 26.4% ($\lambda = 50.05$).

Terceiro, o estimador white apresenta-se menos viesado sob heteroscedasticidade do que o estimador naïve, exceto quando o tamanho da amostra é 50 e o grau de heteroscedasticidade é 15.04. Notamos ainda que os estimadores white e naïve são mais viesados sob heteroscedasticidade do que sob homoscedasticidade.

Quarto, o estimador HC3 mostra-se superior ao estimador white, tanto sob homoscedasticidade quanto sob heteroscedasticidade. Por exemplo, para $n = 47$ e $\lambda = 15.02$, o viés relativo total do estimador white encontra-se em torno de 43%, enquanto que o estimador HC3 apresenta viés relativo total de apenas 4.6%.

Quarto, os estimadores HC4 e HC5 apresentam-se bastante viesados quando os dados incluem pontos de alta alavancagem ('dados3'). Por exemplo, quando o tamanho da amostra é 50 e o grau de heteroscedasticidade é 15.04, o viés relativo total do estimador HC4 excede 100%, enquanto que sob as mesmas condições mas sem considerar os pontos de alta alavancagem seu viés relativo total está em torno de 8%. O viés relativo total do estimador HC5, considerando as observações influentes, excede 100% sob

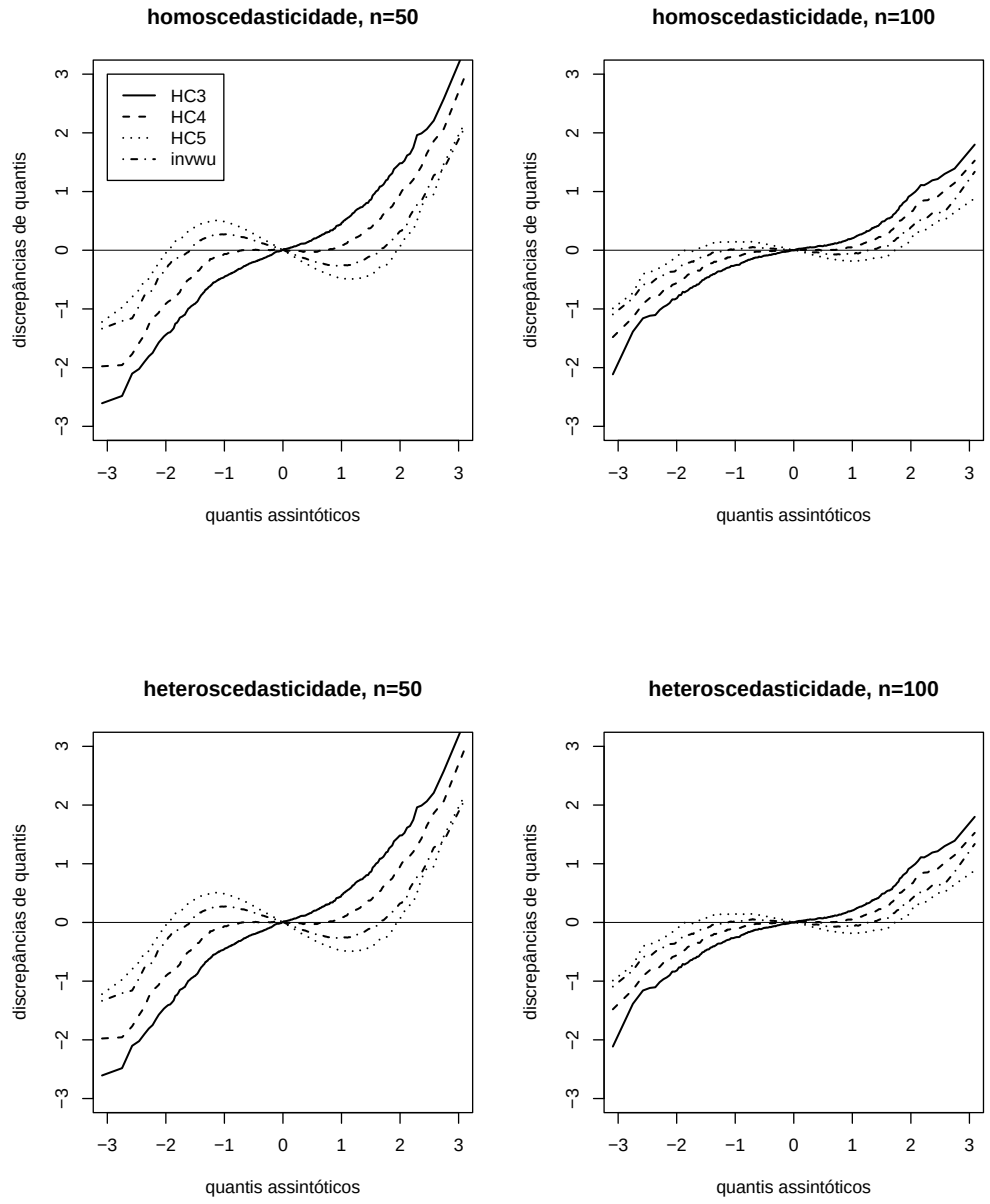
Quadro 3.7: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 50, 100, 150$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados3'.

	$n = 50$								
λ	1.00			15.04			50.05		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.64	6.16	1.42	33.34	24.80	13.22	45.20	37.32	24.20
white	20.70	14.04	6.26	37.60	29.64	18.10	44.66	36.56	24.22
HC3	15.34	9.74	4.20	26.60	19.94	11.24	31.64	24.18	14.66
HC4	10.68	6.48	2.72	17.18	12.52	6.80	19.52	14.50	8.50
HC5	4.66	2.68	1.08	6.88	5.06	2.64	7.58	5.74	3.26
naïve	10.70	6.12	1.48	33.38	29.94	13.46	45.04	37.54	24.30
wu	15.28	9.98	4.20	26.76	20.04	11.48	31.84	24.14	14.68
invwu	5.86	3.40	1.32	10.32	6.94	3.34	11.42	8.16	4.66
	$n = 100$								
λ	1.00			15.04			50.05		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.62	5.28	1.04	32.22	24.10	12.44	43.70	35.86	22.52
white	16.12	9.92	3.36	24.70	17.52	8.98	27.92	20.54	11.22
HC3	13.50	7.84	2.40	19.62	13.58	6.58	22.18	15.62	8.30
HC4	11.08	6.10	1.92	15.42	10.34	4.62	16.80	11.64	5.68
HC5	7.22	4.04	1.12	9.46	5.70	2.42	10.22	6.56	3.02
naïve	10.70	5.40	1.12	32.36	24.38	12.60	44.02	35.58	22.80
wu	13.66	7.86	2.54	19.82	13.74	6.48	22.36	15.88	8.32
invwu	8.22	4.48	1.24	11.86	7.96	3.26	14.16	9.48	4.40
	$n = 150$								
λ	1.00			15.04			50.05		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.28	4.76	0.98	32.04	23.80	11.64	43.16	34.72	22.34
white	14.28	8.28	2.18	19.70	13.06	6.04	21.08	14.66	7.24
HC3	12.26	6.72	1.86	16.50	10.72	4.62	17.42	12.12	5.74
HC4	10.80	5.58	1.66	13.52	8.62	3.56	14.40	9.66	4.40
HC5	7.98	3.98	0.80	9.44	5.92	2.12	9.98	6.34	2.58
naïve	10.44	4.82	1.00	32.00	23.88	12.10	43.46	34.82	22.50
wu	12.50	6.84	1.78	16.50	10.72	4.62	17.72	12.08	5.84
invwu	8.56	4.22	1.10	11.92	7.30	2.86	13.74	8.76	3.78

Quadro 3.8: Porcentagens de rejeição de testes quasi-t baseados nos estimadores ols, white, HC3, HC4, HC5, naïve, wu e invwu no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 47, 94, 141$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (sem pontos de alavanca) - 'dados4'.

	$n = 47$								
λ	1.00			15.02			50.30		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	10.92	5.76	1.34	17.60	10.12	3.38	21.44	13.48	5.40
white	13.30	8.00	2.16	16.04	9.48	3.38	17.30	10.98	4.12
HC3	11.64	6.54	1.70	13.42	7.78	2.64	14.64	8.96	3.18
HC4	11.52	6.48	1.76	12.86	7.52	2.64	14.04	8.58	3.08
HC5	11.52	6.48	1.46	12.86	7.52	2.64	14.04	8.58	3.08
naïve	10.94	5.82	1.36	17.54	10.56	3.34	21.58	13.88	5.36
wu	11.80	6.88	1.70	13.74	8.16	2.92	14.92	9.28	3.40
invwu	9.42	4.84	1.08	11.22	6.08	1.92	12.86	7.52	2.54
	$n = 94$								
λ	1.00			15.02			50.30		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	9.90	5.08	0.96	16.88	10.10	2.84	20.52	13.34	4.52
white	11.40	6.32	1.44	12.90	6.22	1.64	13.64	8.18	2.44
HC3	10.32	5.50	1.18	11.80	6.44	1.68	12.36	7.12	1.88
HC4	10.36	5.48	1.18	11.56	6.22	1.64	12.00	6.86	1.82
HC5	10.36	5.48	1.18	11.56	6.22	1.64	12.00	6.86	1.82
naïve	10.14	5.16	0.98	17.00	10.12	2.90	20.78	13.62	4.82
wu	10.64	5.74	1.16	11.94	6.64	1.78	12.50	7.38	1.96
invwu	9.34	4.68	0.94	11.68	6.24	1.54	13.44	7.88	2.22
	$n = 141$								
λ	1.00			15.02			50.30		
Teste	10%	5%	1%	10%	5%	1%	10%	5%	1%
ols	9.42	4.78	1.10	16.20	9.48	2.84	19.86	12.32	4.44
white	10.38	5.66	1.50	11.60	6.02	1.78	11.98	6.72	2.16
HC3	9.80	5.24	1.28	10.78	5.50	1.60	11.16	6.30	1.74
HC4	9.78	5.18	1.30	10.64	5.48	1.56	10.94	6.20	1.64
HC5	9.78	5.18	1.30	10.64	5.48	1.56	10.94	6.20	1.64
naïve	9.52	4.98	1.06	11.30	6.00	1.68	20.30	12.28	4.52
wu	10.10	5.10	1.34	10.98	5.72	1.66	11.18	6.48	1.94
invwu	9.04	4.62	1.06	11.30	6.00	1.68	13.00	7.36	2.32

Figura 3.5: Discrepâncias de quantis para o conjunto de 'dados1' (com pontos de alavancagem). Diferenças entre quantis exatos (estimados por simulação) e os quantis assintóticos (computados da distribuição $\mathcal{N}(0,1)$) dispostos graficamente contra quantis assintóticos de quatro estatística de teste: HC3, HC4, HC5 e invwu.



homoscedasticidade e sob heteroscedasticidade.

Quinto, o estimador invwu mostra-se bastante viesado no caso de dados com pontos de alta alavancagem ('dados3'). Por exemplo, quando $n = 150$ e $\lambda = 15.04$, o viés relativo total deste estimador fica em torno de 80%, ao passo que sem pontos de alavancagem ('dados4') este valor é apenas 15.3%.

A seguir estão apresentados os quatro primeiros momentos amostrais (média, variância, assimetria e curtose) das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu e invwu para dados com pontos de alavanca ('dados3') e para dados sem pontos de alavanca ('dados4'). Estas medidas estão apresentadas nos Quadros 3.11 e 3.12, respectivamente.

Quadro 3.9: Vieses relativos totais dos estimadores ols, white, HC3, HC4, HC5, naïve, wu, invwu para dados com alavancagem - 'dados3'.

n	λ	ols	white	HC3	HC4	HC5	naïve	wu	invwu
50	1	0.002	0.931	0.005	2.195	20.625	0.010	0.020	8.472
	15.04	1.782	1.806	0.758	2.041	25.780	1.786	0.762	8.508
	50.05	2.262	2.036	1.013	1.764	25.34	2.265	1.015	7.094
100	1	0.004	0.460	0.005	0.590	2.346	0.010	0.002	1.506
	15.04	1.752	0.902	0.186	0.834	3.993	1.756	0.196	1.666
	50.05	2.223	1.018	0.252	0.855	4.294	2.225	0.264	1.187
150	1	0.002	0.294	0.018	0.358	1.174	0.080	0.011	0.833
	15.04	1.741	0.583	0.083	0.534	2.071	1.743	0.087	0.801
	50.05	2.208	0.659	0.118	0.558	2.247	2.210	0.121	0.397

Quadro 3.10: Vieses relativos totais dos estimadores ols, white, HC3, HC4, HC5, naïve, wu, invwu para dados sem alavancagem - 'dados4'.

n	λ	ols	white	HC3	HC4	HC5	naïve	wu	invwu
47	1	0.004	0.306	0.003	0.025	0.025	0.028	0.026	0.431
	15.02	0.721	0.434	0.046	0.081	0.081	0.739	0.090	0.299
	50.30	1.055	0.497	0.079	0.089	0.089	1.071	0.121	0.080
94	1	0.000	0.149	0.003	0.013	0.013	0.006	0.003	0.181
	15.02	0.698	0.210	0.014	0.047	0.047	0.700	0.026	0.059
	50.30	1.023	0.239	0.025	0.055	0.055	1.025	0.038	0.286
141	1	0.001	0.100	0.001	0.008	0.008	0.010	0.004	0.117
	15.02	0.694	0.141	0.010	0.029	0.029	0.698	0.018	0.153
	50.30	1.016	0.161	0.017	0.034	0.034	1.020	0.025	0.380

A distribuição da estatística de teste baseada no estimador HC5 apresenta a menor dispersão na presença dos pontos de alavanca, seguindo-se a distribuição da estatística pautada em invwu, sob homoscedasticidade e heteroscedasticidade. Por exemplo, para $n = 50$ e $\lambda = 15.04$, a variância da estatística de teste baseada no estimador HC5 é de apenas 0.787, que é quatro vezes menor que a variância da estatística de teste baseado no estimador white. Quando os pontos de alavanca não estão presentes nos dados, temos que a menor variância é a da estatística que usa o estimador invwu, exceto quando $n = 141$ e $\lambda = 15.02$. Para exemplificar, considere $n = 47$ e $\lambda = 15.02$; a variância da estatística de teste baseada no estimador invwu é 1.099, enquanto que a variância da estatística de teste baseada no estimador HC5, que coincide com o estimador HC4, é 1.224. Sob homoscedasticidade e heteroscedasticidade, a estatística de teste baseada no estimador white apresenta a maior variância, tanto na presença como na ausência dos pontos de alavanca. Por exemplo, sob heteroscedasticidade ($\lambda = 15.04$) e com $n = 150$, a variância da estatística de teste baseada no estimador white é 1.727, enquanto que a variância da estatística de teste baseada no estimador HC4 é apenas 1.300.

Em relação à curtose (grau de achatamento da distribuição) podemos notar que em todos os casos as distribuições das estatísticas de teste baseadas nos estimadores apresentados são leptocúrticas, ou seja, revelam um alto grau de afilamento, superior ao da distribuição normal padrão. As maiores curtoses são apresentadas pela estatística de teste baseada no estimador HC5; por exemplo, sob heteroscedasticidade ($\lambda = 15.04$), sua curtose é 10.429, que é quase o dobro da curtose da estatística de teste baseada no estimador HC3.

Com o propósito de ilustrar algumas conclusões anteriores, apresentamos na Figura 3.6 os gráficos das densidades estimadas de duas estatísticas de teste, sob heteroscedasticidade ($\lambda = 15.04$), para $n = 50, 150$. Através destes gráficos podemos notar que a estatística de teste baseada no estimador HC5 apresenta alto grau de afilamento, superior ao da distribuição normal padrão. Para $n = 150$ percebemos que as distribuições das estatísticas de teste baseadas nos estimadores HC3 e HC5 se encontram proxima da distribuição normal padrão.

3.4 Bootstrap na estatística de teste quasi-t

O método bootstrap fornece refinamentos para quantidades de interesse envolvendo estatísticas assintoticamente pivotais, ou seja, estatísticas cujas distribuições limites são independentes de parâmetros populacionais desconhecidos. Seja $T_n = T_n(Y_1, \ldots, Y_n)$

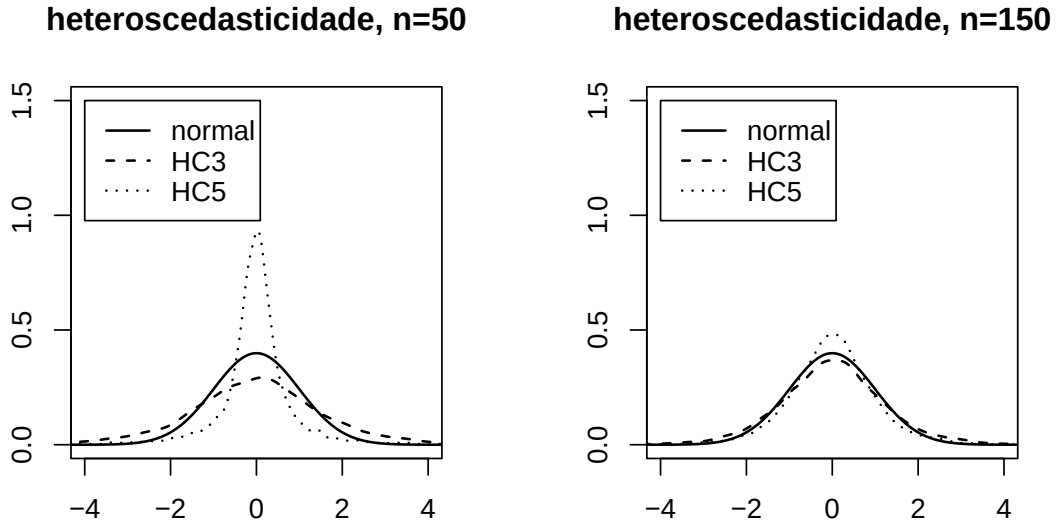
Quadro 3.11: Momentos amostrais: média, variância, assimetria e curtose das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu, invwu para dados com alavancagem - 'dados3'.

	λ	1.00				15.04			
n	testes	média	variância	assimetria	curtose	média	variância	assimetria	curtose
50	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.001	1.838	0.034	3.590	0.002	3.875	0.069	3.595
	HC3	0.003	1.414	0.046	3.991	0.002	2.656	0.074	4.233
	HC4	0.005	1.060	0.065	4.814	0.002	1.776	0.086	5.494
	HC5	0.004	0.538	0.142	8.019	0.003	0.787	0.295	10.429
	wu	0.002	1.427	0.034	3.987	0.000	2.682	0.063	4.299
	invwu	0.004	0.681	0.079	6.130	0.001	1.073	0.123	7.354
150	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.006	1.268	−0.001	3.360	0.002	1.727	0.012	3.839
	HC3	0.005	1.148	−0.002	3.417	0.002	1.514	0.014	3.950
	HC4	0.004	1.047	−0.003	3.489	0.001	1.300	0.017	4.089
	HC5	0.004	0.877	−0.003	3.639	0.001	0.981	0.029	4.410
	wu	0.005	1.156	−0.000	3.434	0.000	1.527	0.007	3.978
	invwu	0.004	0.922	−0.004	3.504	0.002	1.165	0.029	4.064

Quadro 3.12: Momentos amostrais: média, variância, assimetria e curtose das estatísticas de teste baseadas nos estimadores white, HC3, HC4, HC5, wu, invwu para dados sem alavancagem - 'dados4'.

	λ	1.00				15.02			
n	testes	média	variância	assimetria	curtose	média	variância	assimetria	curtose
47	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.003	1.222	0.035	3.266	−0.003	1.415	0.030	3.485
	HC3	0.004	1.104	0.035	3.281	−0.003	1.253	0.031	3.530
	HC4	0.004	1.099	0.036	3.296	−0.003	1.224	0.031	3.575
	HC5	0.004	1.099	0.036	3.296	−0.003	1.224	0.031	3.575
	wu	0.003	1.120	0.032	3.305	−0.003	1.285	0.028	3.548
	invwu	0.004	0.965	0.030	3.300	−0.003	1.099	0.032	3.472
141	assintótico	0.000	1.000	0.000	3.000	0.000	1.000	0.000	3.000
	white	0.005	1.036	0.012	3.134	0.005	1.114	0.021	3.218
	HC3	0.005	1.000	0.012	3.135	0.005	1.067	0.022	3.224
	HC4	0.005	0.999	0.013	3.135	0.005	1.055	0.023	3.231
	HC5	0.005	0.999	0.013	3.135	0.005	1.055	0.023	3.231
	wu	0.006	1.008	0.019	3.156	0.005	1.075	0.029	3.226
	invwu	0.005	0.960	0.008	3.119	0.004	1.099	0.020	3.184

Figura 3.6: Densidades estimadas das estatísticas de testes baseadas nos estimadores HC3 e HC5 para o conjunto de 'dados1' (com pontos de alavancagem) e densidade da distribuição $\mathcal{N}(0,1)$. A estimação foi feita usando um kernel gaussiano.



uma estatística usada para testar uma hipótese particular sobre a distribuição da qual a amostra aleatória $y = (y_1, \ldots, y_n)$ é extraída. Seja

$$G_n(z, F) = P(T_n \leq z)$$

a função de distribuição acumulada de T_n sob $\mathcal{H}_0$. Então, um procedimento de teste bicaudal rejeita $\mathcal{H}_0$, ao nível α, se $|T_n| > z_{n\alpha}$, onde o valor crítico $z_{n\alpha}$ satisfaz

$$G_n(z_{n\alpha}, F) - G_n(-z_{n\alpha}, F) = 1 - \alpha.$$

Se algum aspecto de F é desconhecido, sob $\mathcal{H}_0$, o valor exato de $z_{n\alpha}$ não pode ser obtido, a não ser que T_n seja pivotal. Na prática, a existência de estatísticas que apresentam esta propriedade é rara. No modelo de regressão, por exemplo, a estatística usada para testar hipóteses sobre alguns dos coeficientes do modelo baseada no estimador de mínimos quadrados ordinários será pivotal apenas sob a suposição de normalidade dos erros.

Outra possibilidade é substituir F em $G_n(; F)$ por um estimador consistente, por exemplo, $\widehat{F}$. Suponha que optamos por usar a função de distribuição empírica da amostra

original, i.e., $\widehat{F}$. Então, $G_n(;F)$ é aproximada por $G_n(;\widehat{F})$ e o valor crítico aproximado, $z^*_{n\alpha}$, tal que $\mathcal{H}_0$ é rejeitada ao nível α se $|T_n| > z^*_{n\alpha}$, satisfaz

$$G_n(z^*_{n\alpha}, \widehat{F}) - G_n(-z^*_{n\alpha}, \widehat{F}) = 1 - \alpha.$$

Segundo Horowitz (1997), sob algumas condições de regularidade padrão,

$$\sup_y |\widehat{F}(y) - F(y)|$$

e

$$\sup_z |G_n(z, F_n) - G_n(z, F)|$$

convergem para zero em probabilidade ou quase certamente. Isto garante que o método de bootstrap fornece boas aproximações para $G_n(z, F)$ e para $z_{n\alpha}$ se n é suficientemente grande, sendo estas aproximações tipicamente mais precisas que as fornecidas pela teoria assintótica de primeira ordem se a distribuição de T_n é assintoticamente pivotal.

Neste sentido, um enfoque alternativo ao que foi apresentado anteriormente é usar o estimador de White para construir estatísticas quasi-t e então reamostrar esta quantidade, que é assintoticamente pivotal. O esquema de reamostragem é realizado impondo a restrição em teste na geração das amostras de bootstrap, a estatística de teste sendo calculada em cada réplica de bootstrap. No final do esquema de reamostragem de bootstrap, obtemos ou um valor crítico para o teste (a ser usado em substituição ao valor crítico assintótico obtido da distribuição normal padrão) ou um p-valor bootstrap. O teste bootstrap pode ser realizado como se segue. De início, calcule a estatística quasi–t, digamos τ. Então:

1. Para cada i, $i = 1, \ldots, n$, retire um número aleatório t^*_i de uma população que tem média zero e variância um.

2. Construa uma amostra bootstrap (y^*, X), onde $y^*_i = X_i\widetilde{b} + t^*_i\widetilde{e_i}/(1 - h_i)$. Aqui, $\widetilde{b}$ e $\widetilde{e_i}$ são as estimativas restritas dos parâmetros e os resíduos de MQO sob a hipótese nula. X_i representa a i-ésima linha de X.

3. Calcule o EMQO de β, $b^* = (X'X)^{-1}X'y^*$, e calcule a estatística de teste quasi-t associada, τ^*.

4. Repita os passos 1 a 3 um grande número de vezes (digamos, B).

5. Calcule o quantil de interesse da distribuição empírica de $B + 1$ realizações da estatística de teste.

6. Realize o teste usando a estatística quasi-t calculada inicialmente (τ) conjuntamente com o valor crítico bootstrap obtido no passo 5 acima.

Observe que no teste bootstrap não utilizamos valores críticos obtidos da distribuição assintótica da estatística de teste sob a hipótese nula, i.e., não são utilizados valores críticos da distribuição normal padrão. Aqui são usados valores críticos obtidos através da reamostragem bootstrap.

A regra de decisão pode ser mais convenientemente expressa usando o p-valor bootstrap. A aproximação do p-valor obtida a partir do esquema de bootstrap, para um teste bilateral, é dada por

$$p = \frac{1 + \#\{|\tau_b^*| \geq |\tau|\}}{B + 1},$$

onde τ_b^*, $b = 1, \ldots, B$, são realizações de bootstrap da estatística de teste. Rejeitamos a hipótese nula quando o p-valor bootstrap é menor do que o tamanho nominal selecionado para o teste.

Com o objetivo de comparar numericamente os desempenhos dos testes quasi-t baseados nos estimadores wu e invwu com os desempenhos de dois testes bootstrap, um que usa o esquema de reamostragem de bootstrap ponderado e outro que utiliza o esquema de reamostragem ponderado com probabilidades de amostragem inversamente proporcionais às medidas de alavancagem, utilizamos o conjunto de dados retirado de Greene (1997), 'dados3', e também 'dados4', que não contém pontos de alavanca. Os desempenhos dos testes são avaliados com base nas taxas de rejeição da hipótese nula, sendo a hipótese alternativa bicaudal e a geração dos dados realizada sob a hipótese em teste. Os resultados apresentados a seguir são baseados em 5000 réplicas de Monte Carlo e 500 réplicas de bootstrap.

Os Quadros 3.13 e 3.14 apresentam porcentagens de rejeição de testes quasi-t baseados nos estimadores wu e invwu e também porcentagens de rejeição de dois testes bootstrap, a saber: o teste que usa o esquema de reamostragem de bootstrap ponderado e o teste que usa o esquema de reamostragem ponderado com probabilidades de amostragem inversamente proporcionais às medidas de alavancagem. Nestes testes bootstrap, a reamostragem foi realizada a partir de estatísticas de teste construídas com base no estimador white, que são assintoticamente pivotais, e objetiva obter valores críticos de bootstrap, que são usados em substituição aos valores críticos normais (assintóticos). Nós denotaremos o primeiro teste por wu* e o segundo por invwu* na análise que segue.

O teste bootstrap realizado com base no esquema de bootstrap ponderado (wu*) apresenta muito pouco ganho em amostras finitas relativamente ao uso do bootstrap para estimação de variâncias. Para exemplificar, considere $\lambda = 1$ e $n = 100$, ao nível nominal de 10%. O teste quasi-t que usa o esquema de bootstrap ponderado para estimação de variância (wu) rejeita a hipótese nula 13.66% das vezes ao passo que o

teste bootstrap wu* o faz 13.58% das vezes. Sob heteroscedasticidade ($\lambda = 15.04$), as conclusões são similares e até há casos em que a distorção de tamanho do teste cuja estatística usa o estimador wu é menor que a do teste bootstrap wu*; por exemplo, ao nível de 5% e $n = 150$, o primeiro teste apresenta taxa de rejeição de 10.80% enquanto o segundo apresenta taxa de rejeição de 10.86%.

O teste bootstrap realizado com base no esquema de reamostragem ponderado com probabilidades de amostragem inversamente proporcionais às medidas de alavancagem (invwu*) apresenta taxas de rejeição mais baixas do que se esperava. O resultado central favorece o uso do teste quasi-t baseado no estimador invwu. Por exemplo, sob homoscedasticidade, ao nível nominal de 5% e $n = 100$ a distorção de tamanho do teste bootstrap invwu* é aproximadamente cinco vezes maior que a distorção de tamanho do teste baseado no estimador invwu. Sob heteroscedasticidade ($\lambda = 15.04$), ao nível nominal de 10% e $n = 50$, o teste baseado no estimador invwu apresenta distorção de 0.32 enquanto que a distorção do teste bootstrap invwu* é 5.10. Podemos notar ainda que quando os dados não contêm observações influentes (Quadro 3.10), o teste baseado no estimador invwu apresenta menor distorção de tamanho do que o teste bootstrap invwu*, exceto quando $n = 94, 141$ e $\lambda = 15.02$.

Quadro 3.13: Porcentagens de rejeição de testes quasi-t baseados nos estimadores wu e invwu e porcentagens de rejeição dos testes bootstrap wu* e invwu* no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 50, 100, 150$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (com pontos de alavanca) - 'dados3'.

	$n = 50$					
λ	1.00			15.04		
α	10%	5%	1%	10%	5%	1%
wu	15.28	9.98	4.20	26.76	20.04	11.48
wu*	15.22	9.44	4.32	26.92	19.60	11.54
invwu	5.86	3.40	1.32	10.32	6.94	3.34
invwu*	2.70	1.78	0.60	4.90	3.34	1.72
	$n = 100$					
λ	1.00			15.04		
α	10%	5%	1%	10%	5%	1%
wu	13.66	7.86	2.54	19.82	13.74	6.48
wu*	13.58	8.04	2.80	20.04	13.58	6.72
invwu	8.22	4.48	1.24	11.86	7.96	3.28
invwu*	4.32	2.18	0.52	5.74	3.52	1.20
	$n = 150$					
λ	1.00			15.04		
α	10%	5%	1%	10%	5%	1%
wu	12.50	6.84	1.78	16.50	10.80	4.72
wu*	12.60	7.00	2.06	16.58	10.86	4.62
invwu	8.56	4.22	1.10	11.92	7.30	2.86
invwu*	5.12	2.56	0.62	7.04	4.04	1.40

Quadro 3.14: Porcentagens de rejeição de testes quasi-t baseados nos estimadores wu e invwu e porcentagens de rejeição dos testes bootstrap wu* e invwu* no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $e_i \sim (0, \exp\{c_1 + c_2 x_i\})$, $i = 1, \ldots, n$, para $n = 50, 100, 150$. Os níveis nominais considerados são $\alpha = 0.10, 0.05, 0.01$ (sem pontos de alavanca) - 'dados4'.

	$n = 47$					
λ	1.00			15.02		
α	10%	5%	1%	10%	5%	1%
wu	11.80	6.88	1.70	13.74	8.16	2.92
wu*	11.74	6.80	1.80	13.66	7.90	2.66
invwu	9.42	4.84	1.08	11.22	6.08	1.92
invwu*	6.58	2.86	0.72	7.72	3.90	1.20
	$n = 94$					
λ	1.00			15.02		
α	10%	5%	1%	10%	5%	1%
wu	10.64	5.74	1.16	11.94	6.64	1.78
wu*	10.68	5.72	1.34	12.10	6.48	1.76
invwu	9.34	4.68	0.94	11.68	6.24	1.54
invwu*	7.80	3.62	0.84	10.66	5.38	1.26
	$n = 141$					
λ	1.00			15.02		
α	10%	5%	1%	10%	5%	1%
wu	10.10	5.10	1.34	10.98	5.72	1.66
wu*	10.06	5.20	1.52	10.96	5.66	1.74
invwu	9.04	4.62	1.06	11.30	6.00	1.68
invwu*	8.06	4.06	0.94	11.60	5.96	1.68

Capítulo 4

Aplicações

4.1 Introdução

Neste capítulo apresentaremos quatro exemplos empíricos onde as variâncias dos estimadores de MQO dos coeficientes de um modelo de regressão linear são estimadas a partir dos distintos procedimentos considerados nesta dissertação. Os dados do primeiro exemplo foram extraídos de Myers (1990, Tabela 5.2, p.218). A variável dependente (y) representa o número de homens-hora mensais em instalações da marinha americana e as variáveis independentes (x_1, x_2) correspondem à ocupação média diária de instalação e ao número de alas construídas na instalação. Os dados encontram-se apresentados no Quadro 4.1. Os dados do segundo exemplo, que são apresentados no Quadro 4.5, foram extraídos de Greene (1997, Tabela 12.1, p.541) e sua fonte original é o Departamento de Comércio dos Estados Unidos. A variável dependente (y) representa o gasto per capita em escolas públicas e as variáveis independentes (x, x^2) correspondem à renda per capita por estado, em 1979, nos Estados Unidos, e seus valores quadrados. Os valores para a covariável 'renda' estão reescalonados por 10^{-4}. O estado de Wisconsin foi retirado da amostra por apresentar informação incompleta, de modo que a amostra final é de tamanho 50. O terceiro exemplo utiliza os dados que foram apresentados por Chatterjee & Price (1991, Tabela 5.2a., p.131) e que estão apresentados no Quadro 4.9. A variável dependente (y) é o gasto per capita em educação projetado para 1975 e as variáveis independentes (x_1, x_2, x_3) correspondem à renda per capita em 1973, ao número de milhares de residentes abaixo de 18 anos de idade em 1974 e ao número de milhares de residentes vivendo em áreas urbanas em 1970, respectivamente. No quarto exemplo, a variável dependente (y) representa os graus de prestígio de diferentes atividades calculados a partir de uma pesquisa realizada no Canadá na década de 60 e as variáveis independentes (x_1, x_2, x_3) correspondem à percentagem de empregados do sexo feminino, ao quadrado do nível médio de escolaridade das pessoas ativas na profissão em 1971 e ao quadrado da renda média das pessoas que se ocupam daquela profissão em 1971.

Inicialmente, é útil investigar a presença de heteroscedasticidade, o que pode ser feito aplicando-se um teste de heteroscedasticidade. O teste aqui empregado foi o teste

de Koenker (1981), considerando-se a hipótese nula de homoscedasticidade. Este teste rejeitou a hipótese de que as variâncias dos erros são constantes nos três modelos considerados para os níveis usuais de significância. Dada a indicação da presença de heteroscedasticidade, a realização de testes de hipótese sobre os parâmetros lineares dos modelos deve ser conduzida empregando-se estatísticas de teste quasi–t associadas a estimadores para as matrizes de covariâncias dos estimadores de MQO que sejam consistentes sob heteroscedasticidade de forma desconhecida.

4.2 Atividade em instalações da marinha americana

O modelo proposto inicialmente para descrever a relação entre a variável dependente e os regressores é da forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \tag{4.1}$$

para $i = 1, \ldots, 25$. Os parâmetros lineares em (4.1) são estimados pelo método de MQO. As estimativas pontuais de β_0, β_1 e β_2 são $b_0 = 610.83$, $b_1 = 4.23$, $b_2 = 89.71$, respectivamente. O coeficiente de determinação (R^2) mede a variabilidade da variável aleatória y que pode ser explicada ao se levar em consideração o efeito que as variáveis independentes têm sobre ela. Neste caso o $R^2 = 0.645$.

Quadro 4.1: Dados sobre o número de homens-hora mensais, ocupação média diária e o número de alas construídas.

local	y	x_1	x_2	local	y	x_1	x_2
1	180.23	2.00	1	14	1387.82	54.58	6
2	182.61	3.00	1	15	3559.92	113.88	6
3	164.38	16.60	1	16	3115.29	149.58	14
4	284.55	7.00	1	17	2227.76	134.32	12
5	199.92	5.30	3	18	4804.24	188.74	26
6	267.38	16.50	2	19	2628.32	110.24	12
7	999.09	25.89	3	20	1880.84	96.83	10
8	1103.24	44.42	18	21	3036.63	102.33	14
9	944.21	39.63	10	22	5539.98	274.92	58
10	931.84	31.92	6	23	3534.49	811.08	17
11	2268.06	97.33	6	24	8266.77	354.50	24
12	1489.50	56.63	4	25	1845.89	95.00	9
13	1891.70	96.67	14				

No Quadro 4.2 encontram-se os erros-padrão dos elementos de $b = (b_0, b_1, b_2)$, obtidos

a partir dos diferentes estimadores consistentes apresentados anteriormente. Os erros-padrão dos elementos de b diferem bastante quando calculados com base nos diferentes estimadores considerados. Observa-se que os maiores erros-padrão são provenientes de HC5 e HC4, seguidos por invwu, HC3, wu, white, naïve e ols, nesta ordem. O estimador ols é aquele que fornece as menores estimativas, distanciando-se notadamente das demais. Por exemplo, o erro padrão de b_2 obtido a partir do estimador HC4 é quase vinte seis vezes maior do que aquele obtido a partir do estimador white.

Quadro 4.2: Erros-padrão para b_0, b_1 e b_2 em um modelo de regressão linear da forma $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $i = 1, \ldots, 25$.

estimador	ols	white	HC3	HC4
$\sqrt{\widehat{\text{var}}(b_0)}$	340.21	233.55	691.59	3553.80
$\sqrt{\widehat{\text{var}}(b_1)}$	1.67	2.90	17.79	123.76
$\sqrt{\widehat{\text{var}}(b_2)}$	23.53	28.28	127.46	743.53
estimador	HC5	naïve	wu	invwu
$\sqrt{\widehat{\text{var}}(b_0)}$	8783.8	330.32	358.86	1128.7
$\sqrt{\widehat{\text{var}}(b_1)}$	327.57	1.62	6.79	35.41
$\sqrt{\widehat{\text{var}}(b_2)}$	1892.80	23.14	56.03	226.68

Considere o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra uma hipótese alternativa bicaudal. Uma vez que β_2 é o coeficiente do termo x_{i2} no modelo apresentado acima, o teste de tal hipótese avalia a forma funcional que melhor explica a relação entre a variável dependente 'número de homens-hora mensais' e a covariável 'número de alas construídas'. O Quadro 4.3 apresenta os diferentes p-valores para este teste. Mesmo a tamanhos nominais muito pequenos, verifica-se que os testes cujas estatísticas usam os estimadores ols, white e naïve rejeitam a hipótese nula, isto é, sugerem que o número de alas construídas é importante no sentido de que variações neste regressor levam a variações significantes, em média, no número de homens-hora mensais. O teste quasi-t baseado no estimador HC5 tem o maior p-valor dentre todos os testes, seguido pelos testes baseados em HC4, invwu e HC3, nesta ordem.

Uma análise de influência das observações correspondentes aos locais 22 e 23 revela que estas observações são pontos de alta alavancagem, uma vez que, para o modelo (4.1), o valor de referência $3p/n$ é igual a 0.360 e os elementos diagonais da matriz H referentes a estas observações são 0.724 e 0.857, respectivamente. Quando estas observações são removidas da amostra e o teste de hipótese $\mathcal{H}_0 : \beta_2 = 0$ é refeito, esta hipótese não é rejeitada aos níveis de significância usuais, qualquer que seja o estimador empregado para a matriz de covariâncias. Neste contexto, os estimadores ols, white e naïve mostram-se pouco confiáveis, pois os testes a eles associados conduzem a conclusões provavelmente enganosas sobre a estrutura do modelo, o que não acontece com os estimadores HC3, HC4, HC5, wu e invwu, cujos testes associados apresentam as mesmas conclusões independente dos dados apresentarem ou não observações influentes. Ou seja, as inferências realizadas a partir de testes construídos com base nos estimadores invwu, HC3, HC4 e HC5 não são sensíveis à presença nos dados de observações de extrema alavancagem (locais 22 e 23). Em particular, os p-valores dos testes realizados a partir de invwu, HC4 e HC5 são bastante elevados.

Quadro 4.3: Inferência quasi-t, p-valores.

com pontos de alavanca, n=25		sem pontos de alavanca, n=23	
teste	p-valor	teste	p-valor
ols	0.000	ols	0.598
white	0.002	white	0.594
HC3	0.482	HC3	0.691
HC4	0.904	HC4	0.756
HC5	0.962	HC5	0.756
naïve	0.000	naïve	0.602
wu	0.109	wu	0.637
invwu	0.692	invwu	0.726

Com o objetivo de examinar o impacto que observações correspondentes aos locais 22 e 23 têm na inferência resultante, estimamos o modelo (4.1) 25 vezes, cada vez retirando uma observação. As estimativas dos parâmetros resultantes estão apresentadas no Quadro 4.4. O grande impacto que estas observações (observações 22 e 23) têm nas estimativas é evidente. Quando a observação 22 não está na amostra, a estimativa de β_2 torna-se próxima do dobro da média (85.236); quando a observação 23 não está na amostra a estimativa de β_2 se torna negativa (-7.698). As inferências realizadas a partir dos testes que usam os estimadores ols, white e naïve são dominadas por duas

observações. As inferências realizadas através de outros testes consistentes, por outro lado, não sofrem do mesmo mal.

4.3 Gastos com educação nos EUA - I

Na segunda aplicação o modelo proposto para descrever a relação entre a variável dependente e os regressores é da forma

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \tag{4.2}$$

para $i = 1, \ldots, 50$. Os parâmetros lineares em (4.2) são estimados pelo método de MQO. As estimativas pontuais são $b_0 = 832.91$, $b_1 = -1834.20$, $b_2 = 1587.04$ e o $R^2 = 0.655$.

No Quadro 4.6 encontram-se os erros-padrão dos elementos de $b = (b_0, b_1, b_2)$, obtidos a partir dos diferentes estimadores consistentes apresentados anteriormente. O estimador ols é aquele que fornece os menores erros-padrão. Comparando-se os erros-padrão dos estimadores bootstrap e HC's, observa-se que as as maiores variâncias estimadas correspondem a HC5, seguido por HC4, invwu, HC3, wu, white e naïve, nesta ordem.

Quadro 4.5: Dados sobre gasto per capita em escolas públicas e a renda per capita por estado em 1979 nos Estados Unidos.

estado	gasto	renda	estado	gasto	renda	estado	gasto	renda
Alab.	275	6247	Alasca	821	10851	Ariz.	339	7374
Ark.	275	6183	Cal.	387	8850	Colo.	452	8001
Ct.	531	8914	Del.	424	8640	D.C.	428	10022
Fla.	316	7505	Ga.	265	6700	Hawaii	403	8380
Idaho	304	6813	Ill.	437	8745	Ind.	345	7696
Iowa	431	7873	Kans.	355	8001	Ky.	260	6615
La.	316	6640	Maine	327	6333	Md.	427	8306
Mass.	427	8063	Mich.	466	8442	Minn.	477	7847
Miss.	259	5736	Mo.	274	7342	Mont.	433	7051
Nebr.	294	7391	Nev.	359	9032	N.H.	279	7277
N.J.	423	8818	N.Mex.	388	6505	N.Y.	447	8264
N.C.	335	6607	N.Dak.	311	7478	Ohio	322	7812
Okla.	320	6951	Oreg.	397	7839	Pa.	412	7733
R.I.	342	7526	S.C.	315	6242	S.Dak.	321	6841
Tenn.	268	6489	Texas	315	7697	Utah	417	6622
Vt.	353	6541	Va.	356	7624	Wash.	415	8450
W.Va.	320	6456	Wisc.	$*$	7597	Wyo.	500	9096

$*$ Dado faltante.

Quadro 4.4: Estimativas dos parâmetros sem a i-ésima observação.

obs.	b_0	b_1	b_2
1	654.228	4.182	88.400
2	654.326	4.183	88.388
3	659.727	4.193	88.087
4	647.179	4.195	88.576
5	663.787	4.151	88.520
6	656.709	4.186	88.372
7	610.132	4.233	89.730
8	650.265	3.904	94.79
9	648.719	4.129	90.178
10	663.300	4.197	89.413
11	569.820	4.213	90.942
12	591.936	4.240	90.269
13	624.558	4.201	90.263
14	610.357	4.233	89.720
15	501.441	4.114	93.512
16	591.615	4.219	89.275
17	611.899	4.233	89.715
18	616.763	4.233	89.715
19	519.889	4.246	89.495
20	612.553	4.232	89.706
21	584.884	4.285	88.712
22	107.907	2.590	171.154
23	176.036	21.876	-7.698
24	680.015	2.702	83.432
25	609.566	4.233	89.725

Considere agora o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra uma hipótese alternativa bicaudal. Uma vez que β_2 é o coeficiente do termo quadrático no modelo (4.2), o teste de tal hipótese avalia se a forma funcional que melhor explica a relação entre a variável dependente 'gasto' e a covariável 'renda' é linear ou quadrática. Verifica-se que os testes cujas estatísticas empregam os estimadores ols, white e naïve rejeitam a hipótese nula, ao nível nominal de 10%, sugerindo, então, a existência de uma relação quadrática entre o gasto médio per capita com educação pública e a renda per capita. A partir dos testes baseados nos demais estimadores conclui-se o contrário, i.e., não se obtém evidência suficiente para a rejeição de $\mathcal{H}_0$ aos níveis usuais de significância, o que sugere a existência de relação linear entre as variáveis 'gasto' e 'renda'.

Quadro 4.6: Erros-padrão para b_0, b_1 e b_2 em um modelo de regressão linear da forma $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$, $i = 1, \ldots, 50$.

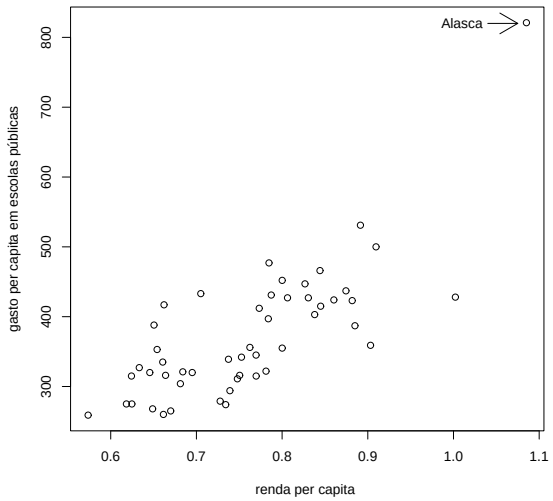
estimador	ols	white	HC3	HC4
$\sqrt{\widehat{\text{var}}(b_0)}$	327.29	460.89	1095.00	3008.00
$\sqrt{\widehat{\text{var}}(b_1)}$	829.99	1243.00	2975.40	8183.20
$\sqrt{\widehat{\text{var}}(b_2)}$	519.08	829.99	1995.2	5488.90
estimador	HC5	naïve	wu	invwu
$\sqrt{\widehat{\text{var}}(b_0)}$	19755.00	330.36	689.85	1652.9
$\sqrt{\widehat{\text{var}}(b_1)}$	53745.00	836.91	1870.10	4495.70
$\sqrt{\widehat{\text{var}}(b_2)}$	36045.00	524.05	1252.60	3015.90

A Figura 4.1 apresenta graficamente os dados, de onde se nota que a observação referente ao estado do Alasca encontra-se destacada das demais. Uma análise da medida da influência desta observação revela ser este um ponto de muito alta alavancagem, uma vez que, para o modelo (4.2), o valor de referência $3p/n$ é igual a 0.180 e o elemento diagonal da matriz H referente ao Alasca é igual a 0.651. Quando a observação correspondente ao Alasca é removida da amostra e o teste da hipótese $\mathcal{H}_0 : \beta_2 = 0$ é refeito, esta hipótese não é rejeitada ao nível nominal de 10%, qualquer que seja o estimador empregado para a matriz de covariâncias. Portanto, para o exemplo considerado, não se pode rejeitar a especificação linear para descrever a relação entre as variáveis do mode-

lo. Neste contexto, os estimadores ols, white e naïve mostram-se pouco confiáveis, pois seus testes associados levam a conclusões provavelmente enganosas sobre a estrutura do modelo. O teste baseado em HC5 é o que apresenta maior p-valor nas duas situações, ou seja, com e sem a observação correspondente ao estado do Alasca. Quando retiramos as três observações identificadas como pontos de alta alavancagem (seus h_i's excedem $3p/n$); estas observações correspondem aos estados do Alasca, Mississipi e Washington, D.C., a hipótese $\mathcal{H}_0 : \beta_2 = 0$ não é rejeitada qualquer que seja o estimador empregado para a matriz de covariâncias.

Para examinar o impacto que a observação correspondente ao estado do Alasca (observação 2) tem na inferência resultante, Cribari–Neto (2004) estimou o modelo (4.2) 50 vezes, cada vez retirando uma observação distinta. As estimativas dos parâmetros resultantes estão apresentadas no Quadro 4.8. O grande impacto que esta observação (observação 2) tem nas estimativas é evidente. Quando esta observação não está na amostra a estimativa de β_2 se torna negativa (-314.139). Nos outros casos, as estimativas variam entre 1526.776 e 2113.170, com média 1603.681. Isto revela que a relação entre a média de y e x é provavelmente linear, e que a rejeição da hipótese nula de que β_2 é igual a zero pelo teste que usa o estimador white é determinada por uma única observação. As inferências realizadas através de testes construídos a partir de outros estimadores consistentes, por outro lado, não são determinadas por uma única observação e apontam para uma relação linear entre as duas variáveis.

Figura 4.1: Gasto per capita em escolas públicas e e renda per capita.



Quadro 4.7: Inferência quasi-t, p-valor.

com pontos de alavanca, n=50		sem pontos de alavanca, n=47		sem Alasca, n=49	
teste	p-valor	teste	p-valor	teste	p-valor
ols	0.002	ols	0.380	ols	0.649
white	0.056	white	0.404	white	0.616
HC3	0.426	HC3	0.463	HC3	0.776
HC4	0.773	HC4	0.476	HC4	0.892
HC5	0.965	HC5	0.476	HC5	0.960
naïve	0.004	naïve	0.402	naïve	0.650
wu	0.194	wu	0.400	wu	0.700
invwu	0.603	invwu	0.450	invwu	0.844

4.4 Gastos com educação nos EUA - II

Nesta teceira aplicação, o modelo proposto para descrever a relação entre a variável dependente e os regressores é da forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \tag{4.3}$$

para $i = 1, \ldots, 50$. Os parâmetros lineares em (4.3) foram estimados pelo método de MQO. As estimativas pontuais são $b_0 = -561.740$, $b_1 = 0.075$, $b_2 = 1.530$, $b_3 = -0.003$ e o $R^2 = 0.586$.

O Quadro 4.10 apresenta os erros-padrão dos elementos de $b = (b_0, b_1, b_2, b_3)$ obtidos a partir dos estimadores consistentes apresentados anteriormente. Estes erros-padrão diferem quando calculados com base nos diferentes estimadores considerados. Observa-se que os erros-padrão HC5 coincidem com os HC4, e que os maiores erros-padrão são provenientes dos estimadores HC4 e invwu, seguidos por HC3, wu, white, ols e naïve, nesta ordem.

Considere o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ contra uma hipótese alternativa bilateral, onde β_2 é o coeficiente da covariável x_2 no modelo apresentado acima. O Quadro 4.11 apresenta os p-valores dos diferentes testes com e sem os pontos de alavanca detectados. Mesmo a tamanhos nominais muito baixos, nota-se que os testes baseados nos estimadores ols, white, naïve e wu rejeitam a hipótese nula, ou seja, sugerem que número de milhares de residentes abaixo de 18 anos de idade em 1974 é importante no sentido de que variações nesta variável levam a variações significantes, em média, no gasto per capita em educação. Por outro lado, as conclusões baseadas nos testes cujas estatísticas usam os estimadores HC4, HC5 e invwu sugerem que a hipótese nula não deve ser rejeitada.

Quadro 4.8: Estimativas dos parâmetros sem a i-ésima observação.

obs.	b_0	b_1	b_2	obs.	b_0	b_1	b_2
1	870.359	−1920.294	1635.783	26	806.920	-1760.717	1538.670
2	−209.034	1000.534	−314.139	27	844.085	-1878.860	1622.080
3	831.140	1829.240	1583.799	28	811.747	−1755.305	1548.696
4	873.876	−1929.117	1641.177	29	809.698	−1708.300	1584.117
5	802.219	−1782.417	1564.448	30	813.273	−1777.132	1548.792
6	881.051	−1956.335	1660.508	31	822.243	−1811.343	1576.591
7	856.050	−1879.980	1604.295	32	781.409	−1721.918	1526.776
8	829.843	−1827.056	1583.229	33	863.572	−1909.745	1630.905
9	1094.365	−2580.285	2113.170	34	822.396	−1812.254	1575.810
10	815.792	−1787.849	1557.467	35	814.681	−1784.563	1555.235
11	850.940	−1869.162	1603.390	36	802.973	−1756.683	1539.429
12	828.203	−1822.768	1580.537	37	832.957	−1833.495	1586.222
13	835.901	−1838.716	1588.358	38	850.278	−1879.024	1614.490
14	830.361	−1828.527	1584.248	39	862.737	−1912.144	1635.334
15	823.128	−1808.506	1571.053	40	828.091	−1821.197	1578.770
16	871.455	−1933.316	1647.514	41	822.521	−1810.289	1573.495
17	814.980	−1788.700	1559.671	42	832.894	−1834.176	1587.038
18	859.074	−1888.543	1614.704	43	863.008	−1900.141	1622.620
19	832.348	−1833.046	1586.464	44	805.052	−1761.049	1541.529
20	813.662	−1790.499	1562.605	45	784.346	−1733.740	1536.150
21	847.917	−1870.989	1608.263	46	808.079	−1780.752	1558.741
22	861.173	−1905.438	1629.284	47	832.352	−1832.710	1586.105
23	864.225	−1909.478	1629.284	48	832.173	−1832.423	1586.046
24	902.230	−2012.966	1696.420	49	825.478	−1817.755	1578.082
25	954.456	−2124.937	1757.911	50	836.025	−1837.164	1584.778

Reproduzido de Cribari–Neto (2004).

Quadro 4.9: Dados sobre despesas per capita em educação projetadas para 1975, renda per capita em 1973, número de residentes abaixo de 18 anos em 1974 e número de residentes vivendo em áreas urbanas em 1970 nos Estados Unidos.

estado	y	x_1	x_2	x_3	estado	y	x_1	x_2	x_3
Maine	235	3944	325	508	N.C.	245	4120	321	450
N.H.	231	4578	323	564	S.C.	233	3817	342	476
Vt.	270	4011	328	322	Ga.	250	4243	339	603
Mass.	261	5233	305	846	Fla.	243	4647	287	805
R.I.	300	4780	303	871	Ky.	216	3967	325	523
Ct.	317	5889	307	774	Tenn.	212	3946	315	588
N.Y.	387	5663	301	856	Alaska	208	3724	332	584
N.J.	285	5759	310	889	Miss.	215	3448	358	445
Pa.	300	4894	300	715	Alab.	221	3680	320	500
Ohio	221	5012	324	753	La.	244	3825	355	661
Ind.	264	4908	329	649	Okla.	234	4189	306	680
Ill.	308	5753	320	830	Texas	269	4336	335	797
Mich.	379	5439	337	738	Mont.	302	4418	335	534
Wisc.	342	4634	328	659	Idaho	268	4323	344	541
Minn.	378	4921	330	664	Wyo.	323	4813	331	605
Iowa	232	4869	318	572	Colo.	304	5046	324	785
Mo.	231	4672	309	701	N.Mex.	317	3764	366	698
N.Dak	246	4782	333	443	Ariz.	332	4504	340	796
S.Dak.	230	4296	330	446	Utah	315	4005	378	804
Nebr.	268	4827	318	615	Nev.	291	5560	330	809
Kans.	337	5057	304	661	Wash.	312	4989	313	726
Del.	344	5540	328	722	Oreg.	316	4697	305	671
Md.	330	5331	323	766	Cal.	332	5438	307	909
Va.	261	4715	317	631	Ark.	546	5613	386	484
W.Va.	214	3828	310	390	Havaii	311	5309	333	831

Uma análise da influência das observações correspondentes aos estados Novo México (NM), Utah (UT) e Alasca (AK) revela que estas observações são pontos de alta alavancagem, uma vez que para o modelo (4.3) o valor de referência $2p/n$ é igual a 0.160 e os elementos diagonais da matriz H referentes a estas observações são 0.174, 0.284 e 0.445, respectivamente. Comparado a outros estados, o Alasca (AK) representa uma situação especial; é um estado com população muito pequena e em 1975 ocorreu um incremento substancial em suas receitas provenientes da produção de petróleo. A Figura 4.2 apresenta o gráfico dos resíduos (eixo vertical) versus valores ajustados (eixo horizontal). Uma análise visual deste gráfico sugere a presença de heteroscedasticidade. Quando os

Quadro 4.10: Erros-padrão para b_0, b_1, b_2 e b_3 em um modelo de regressão linear da forma $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$, $i = 1, \ldots, 50$.

estimador	ols	white	HC3	HC4
$\sqrt{\widehat{\text{var}}(b_0)}$	126.25	178.33	301.22	527.18
$\sqrt{\widehat{\text{var}}(b_1)}$	0.01	0.02	0.03	0.05
$\sqrt{\widehat{\text{var}}(b_2)}$	0.32	0.43	0.71	1.24
$\sqrt{\widehat{\text{var}}(b_3)}$	0.05	0.06	0.08	0.14
estimador	HC5	naïve	wu	invwu
$\sqrt{\widehat{\text{var}}(b_0)}$	527.18	125.87	227.65	402.36
$\sqrt{\widehat{\text{var}}(b_1)}$	0.05	0.01	0.02	0.03
$\sqrt{\widehat{\text{var}}(b_2)}$	1.24	0.31	0.54	0.95
$\sqrt{\widehat{\text{var}}(b_3)}$	0.14	0.05	0.07	0.11

pontos de alavanca são removidos da amostra e o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ é reavaliado, esta hipótese não é rejeitada ao nível nominal de 5%, qualquer que seja o estimador empregado para a matriz de covariâncias. Cumpre notar que dentre todos os testes realizados com base em estimadores consistentes, aqueles baseados em HC4 e HC5 são os que apresentam maior p-valor independentemente da presença de pontos de alavanca nos dados.

4.5 Graus de prestígio de atividades no Canadá

O modelo proposto inicialmente para descrever a relação entre a variável dependente e o regressor é da forma

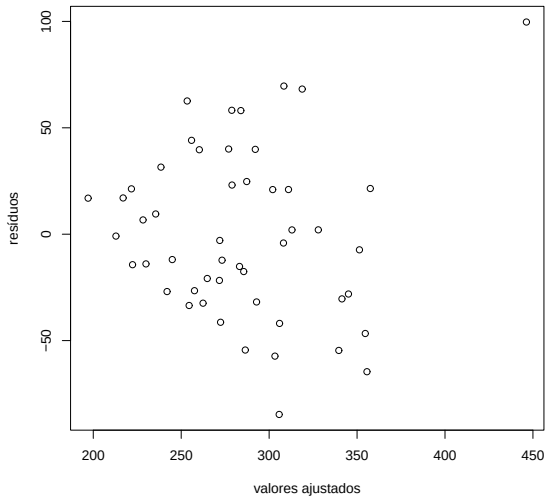
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \tag{4.4}$$

para $i = 1, \ldots, 102$. Os parâmetros lineares em (4.4) são estimados pelo método de MQO. O teste de Koenker (1981) sugere que não há heteroscedasticidade. Nesta

Quadro 4.11: Inferência quasi-t, p-valores.

com os pontos de alavanca, n=50		sem os pontos de alavanca, n=47	
teste	p-valor	teste	p-valor
ols	0.000	ols	0.128
white	0.000	white	0.056
HC3	0.032	HC3	0.083
HC4	0.216	HC4	0.073
HC5	0.216	HC5	0.073
naïve	0.000	naïve	0.128
wu	0.005	wu	0.066
invwu	0.108	invwu	0.117

Figura 4.2: Resíduos versus valores ajustados.



aplicação queremos testar a hipótese nula $\mathcal{H}_0 : \beta_3 = 0$ contra uma hipótese alternativa bilateral, onde β_3 é o coeficiente da covariável x_3 no modelo apresentado acima. O Quadro 4.12 apresenta os p-valores dos diferentes testes. Mesmo a tamanhos nominais muito baixos, nota-se que os testes baseados nos estimadores ols, white, HC3, HC4, naïve, wu e invwu rejeitam a hipótese nula, ou seja, sugerem que a renda média ao quadrado é importante no sentido de que leva a variações significantes, em média, no grau de prestígio de diferentes ocupações. Por outro lado, a inferência baseada no teste cuja estatística usa o estimador HC5 sugere que a hipótese nula não deve ser rejeitada.

Uma análise da influência das observações correspondentes às profissões de administrador geral e médico revela que estas observações são pontos de alta alavancagem, uma

vez que para o modelo (4.4), o valor de referência $3p/n$ é igual a 0.118 e os elementos diagonais da matriz H referentes a estas observações são 0.463 e 0.345, respectivamente. Quando os pontos de alavanca são removidos da amostra e o teste da hipótese nula $\mathcal{H}_0 : \beta_2 = 0$ é reavaliado, esta hipótese não é rejeitada ao nível nominal de 1% qualquer que seja o estimador empregado para a matriz de covariâncias, exceto quando o teste é baseado no estimador naïve. Ao nível de 5% quase todos não rejeitam $\mathcal{H}_0 = 0$. Cumpre notar que dentre todos os testes realizados com base em estimadores consistentes, aquele baseado em HC5 é o que apresenta o maior p-valor independentemente da presença de pontos de alavanca nos dados.

Quadro 4.12: Inferência quasi-t, p-valores.

com pontos de alavanca, n=102		sem pontos de alavanca, n=100	
teste	p-valor	teste	p-valor
ols	0.009	ols	0.028
white	0.000	white	0.093
HC3	0.000	HC3	0.217
HC4	0.003	HC4	0.393
HC5	0.192	HC5	0.631
naïve	0.000	naïve	0.000
wu	0.000	wu	0.088
invwu	0.000	invwu	0.080

Capítulo 5

Implementação dos estimadores consistentes

O objetivo do presente capítulo é mostrar como podem ser realizadas na prática inferências em modelos lineares de regressão utilizando estimadores consistentes para a matriz de covariâncias do estimador de mínimos quadrados ordinários. Para tanto, usaremos o pacote `CAR` ('Companion to Applied Regression') desenvolvido por John Fox para os ambientes `R` e `S-PLUS`. Para maiores detalhes sobre a plataforma `R`, ver Cribari–Neto e Zarkos (1999) e Venables e Ripley (2002); para detalhes sobre as metodologias implementadas no pacote `CAR`, ver Fox (2002).

Considere o segundo exemplo apresentado no capítulo anterior, ou seja, o exemplo onde a variável dependente são gastos per capita em educação pública nos estados americanos e os regressores são as rendas per capita destes estados (devidamente escalonadas) e seus quadrados. O modelo de regressão, assim, é

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, 50.$$

O primeiro passo é ler os dados e armazená-los em um objeto dentro do ambiente `R`, reescalonando os valores da variável independente:

```
alasca <- read.table("c:/alunos/Tati/tese/Programa/Programas_simulacao/alasca.mat",
header=F)
x1 <- alasca[,1]/10000
x2 <- x1^2
y  <- alasca[,2]
```

O próximo passo é ajustar a regressão por mínimos quadrados ordinários:

```
# Fit linear model
ajuste<-lm(y~1+x1+x2)

# Result summaries of the results of model fitting
summary(ajuste)
Call:
lm(formula = y ~ 1 + x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-160.709  -36.896   -4.551   37.290  109.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    832.9      327.3   2.545  0.01428 *
```

```
x1           -1834.2      829.0  -2.213  0.03182 *
x2            1587.0      519.1   3.057  0.00368 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.68 on 47 degrees of freedom
Multiple R-Squared: 0.6553,     Adjusted R-squared: 0.6407
F-statistic: 44.68 on 2 and 47 DF,  p-value: 1.345e-11
```

A fim de verificar se existe heteroscedasticidade, usaremos a versão do teste de Breusch–Pagan proposta por Koenker (1981). Este teste encontra-se implementado no pacote `lmtest`:

```
# Load add-on packages
library(lmtest)

# Koenker's test for heteroskedasticity
bptest(y~1+x1+x2)

        studentized Breusch-Pagan test

data:  y ~ 1 + x1 + x2
BP = 15.8338, df = 2, p-value = 0.0003645
```

Notamos que o teste aponta para a presença de heteroscedasticidade, o p-valor do teste sendo inferior a 1%. Isto indica que devemos basear nossas inferências em testes quasi-t cujas estatísticas empregam estimadores consistentes das variâncias de b_0, b_1 e b_2.

Os estimadores HC0, HC1, HC2, HC3 e HC4 encontram-se implementados na função `hccm` do pacote `CAR`:

```
# Load add-on packages
library(car)

# Compute heteroskedasticity-corrected covariance matrices - HC0
 hccm(ajuste, type="hc0")

             (Intercept)         x1         x2
  (Intercept)    212421.1  -571699.2   379407.4
  x1            -571699.2  1545155.9 -1029609.9
  x2             379407.4 -1029609.9   688887.8

# Compute heteroskedasticity-corrected covariance matrices - HC1
hccm(ajuste, type="hc1")

             (Intercept)         x1         x2
  (Intercept)    225979.9  -608190.6   403624.9
  x1            -608190.6  1643782.9 -1095329.6
  x2             403624.9 -1095329.6   732859.4

# Compute heteroskedasticity-corrected covariance matrices - HC2
 hccm(ajuste, type="hc2")
```

```
             (Intercept)       x1         x2
  (Intercept)    474006.6 -1283633   857209.2
  x1           -1283633.0  3483472 -2330937.3
  x2             857209.2 -2330937  1562867.7

# Compute heteroskedasticity-corrected covariance matrices - HC3
 hccm(ajuste, type="hc3")

             (Intercept)       x1       x2
  (Intercept)     1199026 -3256564  2180884
  x1             -3256564  8853073 -5934046
  x2              2180884 -5934046  3980990

# Compute heteroskedasticity-corrected covariance matrices - HC4
hccm(ajuste, type="hc4")

             (Intercept)        x1        x2
  (Intercept)     9048125 -24613470  16506471
  x1            -24613470  66964620 -44914080
  x2             16506471 -44914080  30128344
```

Os erros-padrão de b_0, b_1 e b_2 são as raízes quadradas dos elementos diagonais das matrizes acima. Estas quantidades devem ser utilizadas nos denominadores das estatísticas quasi-t correspondentes.

Note que uma vez que o código fonte da função `hccm.lm` (usada por `hccm`) encontra-se disponível, podemos alterá-lo para incluir nele a implementação do estimador HC5. Para tanto, copiamos a função `hccm.lm` em um objeto chamado `hccm.new` e editamos este objeto:

```
# hccm.new receives hccm.lm
hccm.new = hccm.lm

# edit hccm.new and implement the estimator HC5
> fix(hccm.new)

> hccm.new
function (model, type = c("hc3", "hc0", "hc1", "hc2", "hc4","hc5"),
    ...)
{
    require(car)
    if (!is.null(weights(model)))
        stop("requires unweighted lm")
    type <- match.arg(type)
    sumry <- summary(model, corr = FALSE)
    s2 <- sumry$sigma^2
    V <- sumry$cov.unscaled
    if (type == FALSE)
        return(s2 * V)
    e <- na.omit(residuals(model))
    X <- model.matrix(model)
```

```
    df.res <- df.residual(model)
    n <- length(e)
    h <- hat(X)
    p <- ncol(X)
    k <-0.7
    hmax<-max(h)
    factor <- switch(type, hc0 = 1, hc1 = df.res/n, hc2 = 1 -
        h, hc3 = (1 - h)^2, hc4 = (1 - h)^pmin(4, n * h/p),
        hc5=(1 - h)^pmin(n*h/p, pmax(4,(n*k*hmax)/p)))
    V %*% t(X) %*% apply(X, 2, "*", (e^2)/factor) %*% V
}
```

Note que `factor` agora inclui a opção do estimador HC5, proposto nesta dissertação. Para usar este estimador, basta especificá-lo como opção de `type` em `hccm.new`, como mostrado a seguir:

```
# Compute heteroskedasticity-corrected covariance matrices - HC5
> hccm.new(ajuste,type="hc5")

             (Intercept)          x1          x2
  (Intercept)   390247068 -1061708291   712057581
  x1          -1061708291  2888498336 -1937239516
  x2            712057581 -1937239516  1299259219
```

Podemos automatizar a inferência definindo uma nova função de sumário; para tanto, copiamos a função `summary.lm` em um objeto chamado `summary.hccm` o editamos para que ele utilize erros-padrão assintoticamente corretos. Note que as alterações feitas estão destacadas com comentários que iniciam com `###`.

```
# summary.hccm receives summary.lm
summary.hccm = summary.lm

# now, edit summary.hccm
fix(summary.hccm)

> summary.hccm
function (object, correlation = FALSE, symbolic.cor = FALSE, ...)
{
    ### we will need the package CAR ###
    require(car)
    ####################################
    ### print a warning to the user ###
    print("\nUsing heteroskedasticity-consistent standard errors.\n")
    ####################################
    z <- object
    Qr <- object$qr
    if (is.null(z$terms) | is.null(Qr))
        stop("invalid 'lm' object:  no terms or qr component")
    n <- NROW(Qr$qr)
    p <- z$rank
    rdf <- n - p
    if (rdf != z$df.residual)
```

```
    warning("inconsistent residual degrees of freedom. -- please report!")
p1 <- 1:p
r <- z$resid
f <- z$fitted
w <- z$weights
if (is.null(w)) {
    mss <- if (attr(z$terms, "intercept"))
        sum((f - mean(f))^2)
    else sum(f^2)
    rss <- sum(r^2)
}
else {
    mss <- if (attr(z$terms, "intercept")) {
        m <- sum(w * f/sum(w))
        sum(w * (f - m)^2)
    }
    else sum(w * f^2)
    rss <- sum(w * r^2)
    r <- sqrt(w) * r
}
resvar <- rss/rdf
R <- chol2inv(Qr$qr[p1, p1, drop = FALSE])
### comment out this line ###
#se <- sqrt(diag(R) * resvar)
##############################
### use HC5 standard errors ###
se<-sqrt(diag(hccm.new(z,type=c("hc5"))))
##############################
est <- z$coefficients[Qr$pivot[p1]]
tval <- est/se
ans <- z[c("call", "terms")]
ans$residuals <- r
ans$coefficients <- cbind(est, se, tval, 2 * pt(abs(tval),
    rdf, lower.tail = FALSE))
dimnames(ans$coefficients) <- list(names(z$coefficients)[Qr$pivot[p1]],
    c("Estimate", "Std. Error", "t value", "Pr(>|t|)"))
ans$sigma <- sqrt(resvar)
ans$df <- c(p, rdf, NCOL(Qr$qr))
if (p != attr(z$terms, "intercept")) {
    df.int <- if (attr(z$terms, "intercept"))
        1
    else 0
    ans$r.squared <- mss/(mss + rss)
    ans$adj.r.squared <- 1 - (1 - ans$r.squared) * ((n -
        df.int)/rdf)
    ans$fstatistic <- c(value = (mss/(p - df.int))/resvar,
        numdf = p - df.int, dendf = rdf)
}
else ans$r.squared <- ans$adj.r.squared <- 0
ans$cov.unscaled <- R
dimnames(ans$cov.unscaled) <- dimnames(ans$coefficients)[c(1,
    1)]
if (correlation) {
```

```
        ans$correlation <- (R * resvar)/outer(se, se)
        dimnames(ans$correlation) <- dimnames(ans$cov.unscaled)
        ans$symbolic.cor <- symbolic.cor
    }
    class(ans) <- "summary.lm"
    ans
}
```

Note que para usar, e.g., HC3 ao invés de HC5, basta especificar este estimador na linha `se<-sqrt(diag(hccm.new(z,type=c("hc5"))))` da função acima. Para usar esta função:

```
# Fit linear model - heteroskedasticity (estimator HC4)
> summary.hccm(ajuste)

"Using heteroskedasticity-corrected standard errors."

Call:
lm(formula = y ~ 1 + x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-160.709  -36.896   -4.551   37.290  109.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    832.9    19754.7   0.042    0.967
x1           -1834.2    53744.8  -0.034    0.973
x2            1587.0    36045.2   0.044    0.965

Residual standard error: 56.68 on 47 degrees of freedom
Multiple R-Squared: 0.6553,     Adjusted R-squared: 0.6407
F-statistic: 44.68 on 2 and 47 DF,  p-value: 1.345e-11
```

Capítulo 6

Conclusões e sugestões para trabalhos futuros

Sob heteroscedasticidade, o estimador de mínimos quadrados ordinários mantém algumas propriedades desejáveis, tornando-se, contudo, ineficiente, i.e., não é mais o melhor estimador linear não-viesado. O estimador usual da matriz de covariâncias deste estimador torna-se viesado e não-consistente, tornando pouco confiáveis estimativas intervalares e testes de hipóteses que utilizam este estimador. Uma maneira de contornar este problema é obter estimadores alternativos que sejam melhores que o estimador de MQO.

Um procedimento de inferência muito usado na prática para o modelo de regressão linear na presença de heteroscedasticidade de forma desconhecida consiste em utilizar para o vetor de parâmetros β o estimador de mínimos quadrados ordinários, que permanece não-viesado e consistente, juntamente com um estimador para sua matriz de covariâncias que possua propriedades assintóticas desejáveis. Este estimador consistente da matriz de covariâncias do estimador de mínimos quadrados ordinários possibilita a realização de inferências sem que seja necessário fazer suposições sobre a forma da heteroscedasticidade.

Nós propomos, com base no estimador HC4, um novo estimador: HC5. Este estimador é o primeiro na classe dos estimadores consistentes da matriz de covariâncias do estimador de mínimos quadrados a incorporar termos de descontos que se ajustam a variações no grau máximo de alavancagem dos dados. Nós apresentamos resultados de simulação de Monte Carlo sobre o desempenho de testes quasi-t cujas estatísticas são baseadas nos diferentes estimadores consistentes. A avaliação foi realizada tanto sob homoscedasticidade quanto sob heteroscedasticidade e os resultados revelaram que o teste construído a partir do estimador HC5 tipicamente apresenta desempenho superior aos demais testes considerados. No que se refere a inferência via bootstrap, há muito pouco ganho em amostras finitas em se usar o esquema de reamostragem de bootstrap ponderado para realizar testes bootstrap, estimando-se valores p ou valores críticos, ao invés de se utilizar o bootstrap ponderado para estimação de erros-padrão a serem utilizados em estatísticas de teste convencionais. Nossos resultados também revelaram que a presença de pontos de alta alavancagem exerce um papel importante no desempenho dos diferentes estimadores consistentes em amostras de tamanho típico.

Dentre as proposta de trabalhos futuros podemos destacar: (i) Estudo do comporta-

mento dos estimadores consistentes da matriz de covariâncias de b quando resíduos oriundos de regressões robustas são usados em substituição a resíduos de mínimos quadrados ordinários. (ii) Utilização dos estimadores consistentes em situações que envolvem modelos de séries temporais, e.g., o modelo AR(p).

Apêndice

Neste apêndice apresentamos o programa Ox usado na realização de simulações com dados de educação.

```
#include <oxstd.h>
#include <oxprob.h>


fbias(const test, const truevalue);
fmse(const estimatorvar, const estimator, const trueval);
fhii(const prob, const emp_prob, const newemp_prob, const obs);
rejection(const series, const criticalvalue, const nrepls);
dataread(const number);

main()
{
/* Declaration of variables used in the program */
decl et, Y, x11, x22, x1, x2, nrepls, nboots, nvar, obs, alpha_1,
alpha_2, X, XtX, invXtX, M, P, Pt, H, systematic, beta, design,
sample, num_samples, num_datasets, sigma, omega, w, bols_mc,
resid, X_bt, Y_bt, invXX_bt, P_bt, M_bt, d_first, d_final,
ratio, rndint, true_var, i, j, true, dataset, a_j, systemat,
weigh_res, res_adj, Y_Wu, Y_Wu2, weight3, naive, wu, inv_wu,
var_naive, bias_naive, mse_naive, var_wu, bias_wu, mse_wu,
var_invwu, var_white, bias_invwu, mse_invwu, var_HC4, HC4_var,
HC4, biasHC4, mseHC4, var_HC5, HC5_var, HC5, biasHC5, mseHC5,
ols_var, biasOLS, olsvar_var, mseOLS, var_HC3, HC3_var, HC3,
biasHC3, mseHC3, white_var, biasWhite, mseWhite, mse_white,
white, relativebias, bias_all, mse_all, sum, t_beta_3,
quasi_t_stat_cvs, kk, kkk, results, hii, Y_hii, X_hii,
Xt_bt, XtX_bt, invXtX_bt, weigh2_res, d_t, var_denom, b_4bt,
X_boot, Y_boot, weight4, weight_mix, Y_mix, weigh_mix, resid2,
weight_22, weight_wu, resid_factor, limit1, limit2, Y_Wu222,
estimators, data, g, weight_expon, gnew, coef, mix, gnew1,
prob, emp_prob, newemp_prob, weight5, hmax, hbarra, alpha,
delta, quoc, k;

/* Initialize the timer */
et = timer();

/* Random number generator */
ranseed("GM");

/* # of Monte Carlo and bootstrap replications */
nrepls = 5000; nboots = 500;

/* initial and final designs to be used (1 through 2) */
d_first = 1;   d_final = 2;

/* # of different sample sizes to be used in the simulation */
```

```
num_samples = 3; num_datasets = 2;

decl file = fopen("lambda15_greene.out", "w");

/* Specifying the model */

estimators = 8;
bias_all = zeros(num_samples, estimators);
mse_all = zeros(num_samples, estimators);
results = zeros(estimators, 3);
t_beta_3 = zeros(nrepls,estimators);


for (dataset = 1; dataset <= num_datasets; ++dataset)
    {
     [data, alpha_1, alpha_2]=dataread(dataset);
     x11 = data[][0] ./ 10000;   x22 = x11 .^ 2;
     print("\nRandom seed George Marsaglia's: ", ranseed(0), "\n");

     /* Loop over designs: */
     for (design = d_first; design <= d_final; ++design)
         {
         /* Loop over different sample sizes */
         for (sample = 1; sample <= num_samples; ++sample)
             {
              x1 = vec(x11 * ones(1, sample));
              x2 = vec(x22 * ones(1, sample));
              X = 1~x1~x2;

              obs = rows(X);
              nvar = columns(X);

              XtX = X'*X;
              invXtX = invertsym(XtX);
              P = invXtX*X';
              Pt = P';

              M = X*P;
              H = unit(obs) - M;

              beta = ones(nvar-1,1)|0;
              coef = 2;

              systematic = X*beta;
              var_denom = (1.0 / (obs-nvar)) .* invXtX;


             /* Initializing the matrices to be used by the program */

              ols_var = zeros(nvar,nrepls); HC4 = ols_var;
              HC5 = ols_var; white = ols_var; HC3 = white;
              mse_naive = HC4; mse_wu = HC4; mse_invwu = HC4;
              var_naive = zeros(nvar,nrepls); var_wu = var_naive;
```

```
var_invwu = var_naive;naive = zeros(nvar,nboots);
wu = naive; inv_wu = naive;


/* Selecting the desings of regression */
if (design == 1)
sigma = ones(obs,1);
else if (design == 2)
sigma = sqrt(exp(alpha_1+alpha_2*x1));

/* Computing the true var-cov matrix */
omega = diag(sigma .^ 2);
true_var = P*omega*Pt;
true = diagonal(true_var)';

/* Returns the diagonal of the product as a row vector */
w = diagonal(M);

prob = 1 ./ w';
emp_prob = prob .* (1.0 / double(sumc(prob)) );
newemp_prob = 0.0|cumulate(emp_prob);

g =  ( obs .*w ) ./ nvar;
gnew1 = g .> 4 .? 4 .: g; /* Minimum between 4 and g */

hmax=max(w);
hbarra=meanr(w);
k = 0.7;
quoc = k*hmax/hbarra;
delta = max(quoc,4);
alpha = w./hbarra .>  delta .? delta .: w./hbarra ;

weight4 = sqrt( 1.0 ./ ((1.0-w') .^ gnew1') ); /* usado para o HC4 */
weight3 = sqrt( 1.0 ./ ((1.0-w') .^ 2) );      /* usado para o HC3*/
weight5 = sqrt( 1.0 ./ ((1.0-w') .^ alpha') ); /* usado para o HC5 */

resid_factor = sqrt( obs / (obs-nvar) );
limit1 = 2*nvar/obs;
limit2 = 3*nvar/obs;

/* Weights */
 d_t = w .> limit1 .? 3 .: 1;
 weight_wu = sqrt( 1.0 ./ (1.0-w') );
 weight_22 = sqrt( 1.0 ./ ((1.0-w') .^ (d_t')) );

/* Begin the Monte Carlo loop */
for (i = 0; i < nrepls; ++i)
    {
    /* Generate the data */
     ranseed({2.0*i+50, 5.0*i+100});
     Y=systematic+sigma .* rann(obs,1);

     /* Run ols regression */
```

```
bols_mc = P*Y;
resid = H*Y;
resid2 = (resid .^ 2);

/* Calculating ols's covariance matrix */
ols_var[][i]=diagonal( double(sumc(resid2)) * var_denom )';

 /* Calculating White's covariance matrix */
 var_white = P * (resid2 .* Pt);
 white[][i] = diagonal( var_white )';

/* Calculating HC4's covariance matrix */
var_HC4 = P * ( (resid2 .* weight4) .* Pt );
HC4[][i] = diagonal(var_HC4)';

/* Calculating HC5's covariance matrix */
var_HC5 = P * ( (resid2 .* weight5) .* Pt );
HC5[][i] = diagonal(var_HC5)';

/* Calculating HC3's covariance matrix */
var_HC3 = P * ( (resid2 .* weight3) .* Pt );
HC3[][i] = diagonal( var_HC3 )';

/* Preparations for bootstrap loop */

systemat = X*bols_mc;
/* Wu's bootstrap */
a_j = resid ./ sqrt( meanr((resid2)') );
weigh_res = ( resid .* weight_wu );
/* new Wu's non-parametric boot */
weigh2_res = resid .* ( weight_22 );
/* naive bootstrap */
res_adj = resid * resid_factor;

/* Begin the bootstrap loop */
for (j = 0; j < nboots; ++j)
    {
    ranseed({2.0*(i+j*555+1), i+j});
    rndint = ranu(obs,1)*obs;

    /* Naive bootstrap */
    naive[][j] = P*( systemat+res_adj[rndint][0] );

    /* Wu's bootstrap */
    Y_Wu = systemat+(weigh_res .* a_j[rndint][0]);
    wu[][j] = P*Y_Wu;

    /* Invwu's bootstrap */
    hii =fhii(prob,emp_prob, newemp_prob, obs);
    Y_Wu222 = systemat+(weigh2_res .* a_j[hii][0]);
    inv_wu[][j] = P*Y_Wu222;
    }  /* End of bootstrap loop */
```

```
/* Obtaining variance and mean-squared error of bootstrapped estimators */

var_naive[][i] = varr(naive);
mse_naive[][i] = fmse(var_naive[][i], naive, beta);

var_wu[][i] = varr(wu);
mse_wu[][i] = fmse(var_wu[][i], wu, beta);

var_invwu[][i] = varr(inv_wu);
mse_invwu[][i] = fmse(var_invwu[][i], inv_wu, beta);

/* Obtaining the t-statistics */

t_beta_3[i][0]= (bols_mc[coef][0]-0.0) / sqrt(ols_var[coef][i]);
t_beta_3[i][1]= (bols_mc[coef][0]-0.0) / sqrt(var_naive[coef][i]);
t_beta_3[i][2]= (bols_mc[coef][0]-0.0) / sqrt(var_wu[coef][i]);
t_beta_3[i][3]= (bols_mc[coef][0]-0.0) / sqrt(var_invwu[coef][i]);
t_beta_3[i][4]= (bols_mc[coef][0]-0.0) / sqrt(HC4[coef][i]);
t_beta_3[i][5]= (bols_mc[coef][0]-0.0) / sqrt(HC3[coef][i]);
t_beta_3[i][6]= (bols_mc[coef][0]-0.0) / sqrt(white[coef][i]);
t_beta_3[i][7]= (bols_mc[coef][0]-0.0) / sqrt(HC5[coef][i]);

} /* End of Monte Carlo simulations */

/* obtaining variance and mean-squared error of all estimators */

olsvar_var = varr(ols_var);
mseOLS = fmse(olsvar_var, ols_var, true);

HC4_var = varr(HC4);
mseHC4 = fmse(HC4_var, HC4, true);

HC5_var = varr(HC5);
mseHC5 = fmse(HC5_var, HC5, true);

HC3_var = varr(HC3);
mseHC3 = fmse(HC3_var, HC3, true);

white_var = varr(white);
mseWhite = fmse(white_var, white, true);

mse_all[sample-1][] = sqrt(sumc((mseOLS)~meanr(mse_naive)~
meanr(mse_wu)~meanr(mse_invwu)~(mseHC4)~(mseHC3)~(mseHC5)));

/* obtaining bieses of estimators */

bias_naive = fbias(var_naive, true);
bias_wu = fbias(var_wu, true);
bias_invwu = fbias(var_invwu, true);
biasOLS = fbias(ols_var, true);
biasHC4 = fbias(HC4, true);
biasHC5 = fbias(HC5, true);
biasHC3 = fbias(HC3, true);
```

```
biasWhite = fbias(white, true);

relativebias=biasOLS~bias_naive~bias_wu~bias_invwu~
biasHC4~biasHC3~biasWhite~biasHC5;
bias_all[sample-1][] = sumc(fabs(relativebias));

/* Print simulation results */

print( "\n\t\t SIMULATION RESULTS: HETEROSKEDASTIC BOOTSTRAP");
print( "\n\t\t ------------------\n");
print( "\n\t\t OX PROGRAM: ", oxfilename(0) );
print( "\n\t\t OX VERSION: ", oxversion() );
print( "\n\t\t NUM. REPLICATIONS: ", nrepls );
print( "\n\t\t NUM. BOOT. REPLICATIONS: ", nboots );
print( "\n\t\t NUM. OBSERVATIONS: ", obs );
print( "\n\t\t DESIGN: ", design );
print( "\n\t\t DATASET: ", dataset );
if(dataset == 1) print(" * GREENE * ");
else if(dataset == 2) print(" * GREENE NO LEVERAGE * ");

print( "\n\t\t DATE: ", date() );
print( "\n\t\t TIME: ", time() );
print( "\n" );

format(200);

if (sample == 1)
   {
   print( "\n\nDesign: ", design );
   print( "\nDesigns 1 and 2 are for normal errors.",
   "\n(homoskedasticity and heteroskedasticity, respectively).");
   print( "\n");
   print( "\nAlpha's: ", alpha_1~alpha_2 );

   }

print( "\n\nLeverage in X matrix:", min(w)~max(w) );
print( "\nLimits: 2p/n and 3p/n: ", limit1~limit2 );
print( "\nNumber of w's beyond limits: ", (sumr(w .> limit1))~
(sumr(w .> limit2)) );


if (dataset==1)
   {
   print("\nRMSE (N=", obs, ")","%c",{"OLS", "naive", "wu","invwu",
   "HC4","HC3","white", "HC5"},"%12.3f", sqrt(meanr(mseOLS)~
   meanr(mse_naive)~meanr(mse_wu)~meanr(mse_invwu)~meanr(mseHC4)~
   meanr(mseHC3)~meanr(mseWhite)~meanr(mseHC5)));

   }
else
   {
   print("\nRMSE (N=", obs, ")", "%c",{"OLS", "naive", "wu","invwu",
```

```
        "HC4","HC3","white", "HC5"},"%12.1f", sqrt(meanr(mseOLS)~
        meanr(mse_naive)~meanr(mse_wu)~meanr(mse_invwu)~meanr(mseHC4)~
        meanr(mseHC3)~meanr(mseWhite)~meanr(mseHC5)));

        }

    print("\nRelative bias (N=", obs, ")", "%c", {"OLS", "naive", "wu","invwu",
    "HC4","HC3","white","HC5"},"%12.3f", relativebias);

    /* Tallying rejections */

    quasi_t_stat_cvs = < 1.644854, 1.959964, 2.575829 >;

    for (kk = 0; kk < rows(results); ++kk)
       {
       for (kkk = 0; kkk <= 2; ++kkk)
           {
           results[kk][kkk] = rejection(t_beta_3[][kk],
           quasi_t_stat_cvs[0][kkk], nrepls);
           }
       }

    fprint(file, "%9.4f",t_beta_3[][2:7]);
    print("\n* Estimated sizes of quasi-t tests: ", "%r", {"OLS","naive",
    "wu","invwu", "HC4","HC3","white", "HC5"},"%c", {"10%", "5%", "1%"},
    "%10.3f", results );

    }  /* End loop over sample sizes */

    print( "\n This is design ", design );

    /* Printing the variance ratio for each design*/

    ratio = ( max(sigma) ./ min(sigma) ) ^2;
    print( "\n Variance ratio (lambda) = ", ratio, "\n" );

if (dataset==1)
{
print("\n* Aggregate RMSE of:", "%c", {"OLS", "naive", "wu","invwu",
"HC4","HC3","white", "HC5"}, "%r",{"N =  50", "N = 100", "N = 150"},
"%12.3f", mse_all);
}
else
{
print("\n* Aggregate RMSE of:", "%c", {"OLS", "naive", "wu", "invwu",
"HC4","HC3","white", "HC5"}, "%r",{"N =  47", "N = 94", "N = 141"},
"%12.1f", mse_all);
}

print("\n* Aggregate relative bias of:", "%c", {"OLS","naive","wu","invwu",
"HC4","HC3","white", "HC5"}, "%r",{"N = 50", "N = 100", "N = 150",
"N = 200"}, "%12.3f", bias_all );
```

```
        }       /* End loop over designs */

        }       /* End of loop over datasets */

        fclose(file);

        print( "\n\nDate: ", date() );
        print( "\nTime: ", time() );
        print( "\nOx version: ", oxversion() );
        print( "\nExecution time: ", timespan(et) );

        }

fbias(const test, const truevalue)
{
        return ((meanr(test)-truevalue) ./ truevalue);

}

fmse(const estimatorvar, const estimator, const trueval)
{

       decl mse;
       mse =  estimatorvar + ( (meanr(estimator)-trueval) .^ 2.0 );
       return mse;
}

rejection(const series, const criticalvalue, const nrepls)
{
        return (sumc( fabs(series) .> criticalvalue ) / nrepls)*100;
}

dataread(const number)
{
        decl data, alpha_1, alpha_2;
        if (number==1)
        {
        data = loadmat("greene.mat");
        alpha_1 = 5.3; alpha_2 = 5.3;
        }
        else if (number==2)
        {

        data = loadmat("greene_no.mat");
        alpha_1 = 9.3; alpha_2 = 9.3;
        }
        return {data, alpha_1, alpha_2};
        }

fhii(const prob, const emp_prob, const newemp_prob, const obs)
   {
    decl rdraw, test, i, j;
```

```
    rdraw = ranu(obs,1);
    test = zeros(obs, 1);

    for (j=0; j<obs; ++j)
    {
    for(i=0; i < obs; ++i)
    {
        if(newemp_prob[i][0] < rdraw[j][0] && rdraw[j][0] < newemp_prob[i+1][0])
        {
              test[j][0] = i;
              break;
        }
    }
    }

    return test;
}
```

Referências

[1] Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83, 687–697.

[2] Cagan, P. (1974). *Common Stock Values and Inflation: the Historical Record of Many Countries*. Boston: National Bureau of Economic Research.

[3] Cribari–Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, a aparecer.

[4] Cribari–Neto, F. & Zarkos, S.G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, 18, 211–228.

[5] Cribari–Neto, F. & Zarkos, S.G. (2001). Heteroskedasticity-consistent covariance matrix estimation: White's estimator and the bootstrap. *Journal of Statistical Computation and Simulation*, 68, 391–411.

[6] Cribari–Neto, F. & Zarkos, S.G. (2004). Leverage–adjusted heteroskedastic bootstrap methods. *Journal of Statistical Computation and Simulation*, 74, 215-232.

[7] Chatterjee, S. & Price, B. (1991). *Regression Analysis by Example*. New York: John Willey & Sons.

[8] Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

[9] Doornik, J.A. (2001). `Ox`*: an Object-oriented Matrix Programming Language,* 4 ª ed. Londres: Timberlake Consultants.

[10] Efron, B. (1979). Bootstrapping methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

[11] Efron, B. & Tibshirani, R.J. (1986). Bootstrap methods for standard error, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54–77.

[12] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[13] Espinheira, P.L. (2003). *Estimação e Inferência sob Heteroscedasticidade de Forma Desconhecida.* Dissertação de Mestrado, Programa de Pós-graduação em Estatística, Universidade Federal de Pernambuco.

[14] Fisher, N.I. & Hall, P. (1989). Bootstrap confidence regions for directional data. *Journal of the American Statistical Association*, 84, 996–1002.

[15] Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. London: Sage.

[16] Freedman, D.A. (1981). Bootstrapping regression models. *The Annals of Statistics,* 9, 1218–1228.

[17] Galvão, N.M.S. (2000). *Estimação Consistente de Matrices de Covariâncias sob Heteroscedasticidade de Forma Desconhecida.* Dissertação de Mestrado, Programa de Pós-graduação em Estatística, Universidade Federal de Pernambuco.

[18] Greene, W.H. (1997). *Econometric Analysis*, 3ª ed. Upper Saddle River: Prentice Hall.

[19] Hinkley, D.V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19, 285–292.

[20] 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.

[21] Horowitz, J.L. (1997). Bootstrap methods in econometrics: theory and numerical performance. Em Kreps, D.M. & Wallis, K.F., eds., *Advances in Economics and Econometrics: Theory and Applications* (Seventh World Congress), 3, 188–222. New York: Cambridge University Press.

[22] Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis ans graphics. *Journal of Computational and Graphical Statistics*, 5, 299–315.

[23] Jeong, J. & Maddala, G.S. (1993). A perspective on application of bootstrap methods in econometrics. Em Maddala, G.S., Rao, C.R. & Vinod, H.D., eds., *Handbook of Statistics: Econometrics*, 11, 573–610. Amsterdam: North-Holland.

[24] Koenker, R. (1981). A note on Studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17, 107–112.

[25] Li, H. & Maddala, G.S. (1996). Bootstrapping time series models. *Econometric Reviews*, 15, 115–158.

[26] Long, J.S. & Ervin, L.H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *The American Statistician*, 54, 217–224.

[27] MacKinnon, J.G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics*, 29, 305–25.

[28] Myers, R.H. (1990). *Classical and Modern Regression with Applications*. Belmont: Duxbury Press.

[29] Shao, J. (1988). On resampling methods for variance and bias estimation in linear models. *Annals of Statistics*, 16, 986–1008.

[30] Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statitics with S*, 4 ª ed. New York: Springer-Verlag.

[31] Vinod, H.D. (1993). Bootstrap methods: Applications in econometrics. In Maddala, G.S., Rao, C.R. and Vinod, H.D., eds., *Handbook of Statistics: Econometrics*, 11, 629–661. Amsterdam: North-Holland.

[32] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817–838.

[33] Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14, 1261–1295.