Modelos de Regressão PLS com Erros Heteroscedásticos

Saulo Almeida Morellato

UFSCar - São Carlos/SP

Fevereiro/2010

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de Regressão PLS com Erros Heteroscedásticos

Saulo Almeida Morellato

Orientador: Prof. Dr. Carlos Alberto Ribeiro Diniz

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos/SP
Fevereiro/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M841mr

Morellato, Saulo Almeida.
Modelos de regressão PLS com erros heteroscedásticos / Saulo Almeida Morellato. -- São Carlos : UFSCar, 2010.
49 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Análise de regressão. 2. Heteroscedasticidade. 3. Distribuição normal assimétrica. I. Título.

CDD: 519.536 (20ª)

Saulo Almeida Morellato

Modelos de Regressão PLS com Erros Heteroscedásticos

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 26 de janeiro de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar/ Orientador)

1º Examinador _____
Prof. Dr. José Antonio Cordeiro (FAMERP)

2º Examinador _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP)

Agradecimentos

Em primeiro lugar à Deus, por mais essa realização na minha vida.

À minha família, pai, mãe, irmãos e sobrinhos, pelo apoio e incentivo durante todos os anos de estudo.

Ao professor Dr. Carlos Alberto Ribeiro Diniz pela orientação e pelas idéias durante todo este trabalho.

Aos colegas, professores e funcionários do Departamento de Estatística da UFSCar, pela amizade.

À CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira.

A todos, meu muito obrigado.

Saulo Almeida Morellato

Resumo

Este trabalho aborda dois problemas relacionados aos modelos de regressão por mínimos quadrados parciais (*Partial Least Squares*, PLS): erros heteroscedásticos, ou seja, erros com variância não constante, e a assimetria na distribuição dos erros.

No método de regressão PLS uma das suposições básicas é a homocesdasticidade dos erros. Quando isso não ocorre, uma alternativa é estimar a estrutura heteroscedástica dos mesmos. Na primeira parte deste trabalho é apresentada uma técnica, baseada em PLS, que permite estimar os parâmetros do modelo de regressão linear levando em consideração a presença de heteroscedasticidade dos erros. Esta técnica é comparada com o método PLS usual na estimação em modelos com erros heteroscedásticos.

O método de regressão PLS é uma abordagem livre de distribuição, ou seja, não assume uma distribuição para os erros. Para a estimação da estrutura heteroscedástica atribuimos uma distribuição aos erros. A idéia é a mesma utilizada por Bastien *et al.* (2005).

Em geral, a análise estatística para o estudo de dados contínuos tem sido desenvolvida em grande parte com base no modelo normal. Dessa forma, esta distribuição seria a escolha comum para modelar os erros, a fim de estimar a heteroscedasticidade. Entretanto, em muitas situações práticas essa suposição de normalidade pode nos levar a inferências pouco apropriadas sobre os parâmetros de interesse. Neste trabalho, pretendemos flexibilizar essa suposição de normalidade, dispondo de uma classe de distribuições assimétricas proposta por Azzalini (1985), que inclui a distribuição normal como um caso particular, a distribuição normal assimétrica.

Para a detecção da heteroscedasticidade nos erros foram propostas adaptações de testes como o teste de White e o teste de Goldfeld-Quandt para erros normais; e testes

escore para homogeneidade dos parâmetros de escala e assimetria da distribuição normal assimétrica, proposto por Xei *et al.* (2009).

Todos os métodos decritos no trabalho são ilustrados com dados simulados e reais.

Palavras-chave: Regressão PLS; Heteroscedasticidade; Modelo Normal Assimétrico.

Abstract

Two problems related to Partial Least Squares method are considered in this work. Heteroscedastic errors and an asymmetrical error distribution. In the first part of this work a methodology is developed which allows, based in PLS methods, to estimate the model parameters in the presence of non-constant error variance. This technique is compared with the usual PLS method which considers homoscedastic erros.

The PLS method is an distribution free approach, that is, it does not assume any distribution for the error terms. In order to estimate the heterocedastic structure an probability distribution is attributed to the errors, similar to the idea of Bastien *et al.* (2005). In this work, it is proposed a class of asymmetric distributions, the asymmetric normal distribution, presented in Azzalini (1985), which includes the normal distribution as a particular case.

For the heteroscedasticity detection is proposed adaptations of the White test, the Goldfeld-Quandt test and an test proposed by Xei *et al.* (2009), which is used for testing the homogeneity of the scale parameter and/or significance of autocorrelation in skew-normal nonlinear regression model. The test methods are illustrated with two numerical examples.

All the methods present in the work are illustrated with simulated and real datasets.

Keywords: PLS method; heteroscedastic errors; asymmetric normal distribution.

Sumário

Capítulo 1

Introdução

Em vários problemas envolvendo o uso de modelos de regressão as covariáveis podem ser altamente correlacionadas ou ainda o número de covariáveis pode ser maior que o de observações. Nestas situações, uma alternativa é usar o método de regressão por mínimos quadrados parciais (PLS). A dificuldade é quando juntamente aos problemas já citados, temos ainda que os erros não tenham variância constante ou ainda que tenham um comportamento assimétrica, que são problemas não considerados na construção do modelo de regressão PLS.

O método de regressão PLS é uma técnica de estimação do modelo de regressão linear, baseada na decomposição das matrizes de variáveis resposta e de covariáveis. Este método é discutido mais amplamente na seção 2.4

Estes trabalho propõe um modelo de regressão PLS com erros heteroscedásticos, ou seja, erros com variâncias não constantes. Além de propor este modelo PLS que considera a informação da heteroscedasticidade na estimação dos coeficientes, ainda serão propostas adaptações para alguns testes de heteroscedasticidade de forma que possam ser usados na regressão PLS.

No capítulo 3 é apresentado o método de regressão PLS com erros heteroscedásticos. A idéia é escolher uma estrutura de heteroscedasticidade e depois estimar os parâmetros desta estrutura. Apesar de o PLS ser uma abordagem livre de distribuição, seguiremos a abordagem utilizada por Bastien *et al.* (2005) em regressão linear generalizada PLS, ou seja, será considerada uma distribuição para o modelo. Inserindo a estrutura

heteroscedástica na distribuição dos erros podemos estimar esta estrutura, e com isso estimar os coeficientes da regressão considerando a presença de heteroscedasticidade. A distribuição usada para os erros é a normal assimétrica, pois é uma distribuição bem geral que pode modelar dados assimétricos e tem a normal como caso particular. Estudos de simulação foram efetuados com diferentes tamanhos de amostra e diferentes graus de heteroscedasticidade para verificar se os coeficientes estimados pelo método proposto possuem menores EQM, vício, e variância que o método PLS usual. Os resultados mostram que o método proposto, que considera a heteroscedasticidade, obteve um melhor desempenho, para todos tamanhos de amostra, quando temos um alto grau de heteroscedasticidade; e um melhor desempenho para grandes amostras quando o temos um grau moderado de heteroscedasticidade.

No capítulo 4 são propostas as adaptações para alguns testes de heteroscedasticidade de forma que possam ser usados na regressão PLS. São adaptações para testes como o de White e o de Goldfeld-Quandt para erros normais, e adaptações para os testes escore para homogeneidade dos parâmetros de escala e assimetria da distribuição normal assimétrica, propostos por Xei *et al.* (2009). Foi realizado um estudo de simulação, com diferentes tamanhos amostrias e diferentes graus de heteroscedasticidade, para verificar as propriedades destes testes adaptados para a regressão PLS. Estes estudos mostraram que estes testes têm um poder elevado para grandes amostras e um alto grau de heteroscedasticidade.

Para melhor ilustrar o método proposto e os testes adaptados, foram utilizados dois conjuntos de dados. Estes exemplos de aplicação são apresentados no capítulo 5. Ao primeiro conjunto de dados é ajustado um modelo PLS heteroscedástico com erros normais, o que permite ilustrar o uso dos testes adaptados de White e de Goldfeld-Quandt. O segundo conjunto de dados é modelado via regressão PLS heteroscedástica com erros normais assimétricos, ilustrando a utilização dos testes adaptados de Xei *et al.* (2009).

Capítulo 2

Regressão por Mínimos Quadrados Parciais (PLS)

2.1 Introdução

Desenvolvido em meados dos anos 60 por Herman O. A. Wold, a regressão por mínimos quadrados parciais foi originalmente construída para o uso no campo da econometria, mas foi adotada pelo campo da quimiometria. Atualmente a regressão por mínimos quadrados parciais tornou-se uma ferramenta padrão para modelagem de relações lineares entre medições multivariadas. Algumas das referências básicas sobre o método PLS são Geladi and Kowlaski (1986), Höskuldsson (1988) e Wold (2001).

Naes *et al*. (2002) argumentam que em calibração multivariada, na quimiometria, geralmente o número de observações é inferior ao número covariáveis, e essas possuem alta correlação entre si.

O método PLS também é utilizado em estudos que usam dados de expressão gênica de DNA para classificação das amostras em categorias, tais como tipos de câncer. Segundo Nguyen e Rocke (2002), dados de expressão gênica são caracterizados por muitas variáveis observadas (gene) e poucas observações (experimentos).

O uso de PLS em econometria se deve, principalmente, ao fato de as covariáveis terem um alto grau de colinearidade.

Segundo MacGregor e Kourti (1995), em várias abordagens de controle de qualidade, é feito o monitoramento apenas das variáveis de qualidade do produto. Entretanto, frequentemente existe uma grande quantidade de covariáveis. O método PLS é utilizado na redução da dimensão dos dados referentes às covariáveis. Essa redução resulta em fatores, que são usadas na construção de cartas de controle multivariadas.

A próxima seção apresenta algumas das diferenças entre o método PLS e o de mínimos quadrados ordinários.

2.2 Diferenças entre PLS e Mínimos Quadrados Ordinários

O método de mínimos quadrados ordinários, diferentemente do PLS, apresenta resultados instáveis para tamanhos de amostra pequenos em relação ao número de variáveis independentes e o alto grau de correlação entre as covariáveis (multicolineariedade) aumenta a variância dos coeficientes estimados.

O PLS é, preferencialmente, uma técnica de predição, e não interpretação, apesar de existirem vários trabalhos que aplicam técnicas interpretativas sobre os fatores extraídos via PLS. Em contrapartida, o método de mínimos quadrados ordinários é uma técnica direcionado para ambas finalidades, predição e interpretação.

A seção a seguir aponta algumas vantagens e desvantagens do método PLS.

2.3 Vantagens e Desvantagens do PLS

O método PLS apresenta as seguintes vantagens:

- Hábil para modelar regressões com múltiplas variáveis resposta;
- Não é afetado por multicolinearidade e
- Produz fatores que tenham grandes covariâncias com as variáveis resposta, ou seja, fatores com alto poder de predição.

As desvantagens do método são:

- Dificuldade na interpretação das cargas dos fatores;
- Os estimadores dos coeficientes de regressão não possuem ditribuições conhecidas e, com isso, o teste de significância dos mesmos só pode ser realizado via métodos de reamostragem; e
- Falta de estatísticas de teste para o modelo.

2.4 Descrição do Método PLS

Nesta seção descrevemos todo o processo de estimação do método PLS, iniciando na extração dos fatores até a obtenção dos estimadores dos coeficientes da regressão. Esta seção é fortemente baseada no trabalho de Wold *et al.* (2001).

Como em regressão linear múltipla, a principal finalidade da Regressão por Mínimos Quadrados Parciais é construir um modelo linear, $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, em que $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ é uma matriz $(N \times M)$ de variáveis resposta, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)$ é uma matriz $(N \times K)$ de variáveis preditoras, $\mathbf{B}$ é uma matriz $(K \times M)$ dos coeficientes da regressão, e $\mathbf{E}$ é a matriz de ruídos para o modelo que tem a mesma dimensão de $\mathbf{Y}$. Os erros no modelo de regressão PLS têm os mesmos pressupostos que a regressão linear múltipla, exceto pela distribuição. Na regressão múltipla, para efeitos de testes de hipóteses, os erros têm distribuição normal multivariada com vetor de médias nulo e matriz de covariâncias $\sigma^2\mathbf{I}$, em que $\mathbf{I}$ é uma matriz identidade $(N \times N)$. O método PLS é uma abordagem livre de distribuição. Dessa forma, os erros da regressão possuem vetor de médias nulo e matriz de covariâncias igual a $\sigma^2\mathbf{I}$, mas sem distribuição definida. A principal consequência disso é que os estimadores dos coeficientes da regressão não possuem distribuições conhecidas. Logo, são necessárias técnicas de reamostragem para verificar a significância dos coeficientes.

O método de regressão PLS extrai um pequeno número de "novas" variáveis, que são chamadas de fatores ou componentes e denotadas por $\mathbf{t}_a$ $(a = 1, \ldots, A)$. Os fatores são preditores de $\mathbf{Y}$ e também descrevem $\mathbf{X}$ (veja equações (2.2) e (2.3) abaixo), isto é, tanto $\mathbf{X}$ como $\mathbf{Y}$ são, pelo menos em parte, modelados pelas mesmas variáveis latentes.

A idéia do método PLS é extrair componentes que consigam capturar as variâncias das covariáveis e também obter correlações com as variáveis dependentes. Isto pode ser conseguido maximizando a covariância entre os fatores de $\mathbf{X}$, $\mathbf{t}_a$, e $\mathbf{Y}$, ou seja, as variáveis latentes são modificadas para que essas covariâncias sejam maximizadas.

O número de componentes extraídos de $\mathbf{X}$ é menor que o número de covariáveis ($A < K$) e os mesmos são ortogonais. Estes são obtidos como combinações lineares das variáveis originais $\mathbf{x}_k$, com os coeficientes, "pesos", $\mathbf{w}_a$ ($a = 1, \ldots, A$), dados por

$$\mathbf{T} = \mathbf{XW}, \tag{2.1}$$

em que $\mathbf{T} = (\mathbf{t}_1,\ \mathbf{t}_2, \ldots,\ \mathbf{t}_A)$ é a matriz $(N \times A)$ de fatores e $\mathbf{W} = (\mathbf{w}_1,\ \mathbf{w}_2, \ldots,\ \mathbf{w}_A)$ é a matriz $(K \times A)$ de pesos.

As matrizes $\mathbf{X}$ e $\mathbf{Y}$ são decompostas, como em uma análise fatorial, da seguinte forma:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{F} \tag{2.2}$$

e

$$\mathbf{Y} = \mathbf{UC}' + \mathbf{G},$$

sendo que $\mathbf{T}$ e $\mathbf{U}$ são matrizes $(N \times A)$ de fatores de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente; $\mathbf{P}'$ e $\mathbf{C}'$ são matrizes $(A \times K)$ de cargas de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente; e $\mathbf{F}$ e $\mathbf{G}$ são matrizes de erros.

Como citado acima, na decomposição de $\mathbf{X}$ as componentes, $\mathbf{t}_a$, são obtidas de maneira que as covariâncias entre elas e as variáveis resposta da matriz $\mathbf{Y}$ sejam maximizadas.

Com a dimensão de $\mathbf{X}$ reduzida em A componentes $\mathbf{t}_a$ ($A < K$) pode-se efetuar a regressão de $\mathbf{Y}$ sobre $\mathbf{T}$ na forma

$$\mathbf{Y} = \mathbf{TC}' + \mathbf{E}. \tag{2.3}$$

Para conseguir os coeficientes da regressão PLS referentes aos dados originais, basta substituir a igualdade em (2.1), na equação (2.3), e obter

$$\mathbf{Y} = \mathbf{TC}' + \mathbf{E} = \mathbf{XWC}' + \mathbf{E} = \mathbf{XB} + \mathbf{E}.$$

Assim, os coeficientes da regressão PLS podem ser escritos como

$$\mathbf{B} = \mathbf{WC}'.$$

O estimador para $\mathbf{C}$ é obtido por mínimos quadrados, e é dado por

$$\widehat{\mathbf{C}}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}.$$

Consequentemente,

$$\widehat{\mathbf{B}} = \mathbf{W}\widehat{\mathbf{C}}' = \begin{bmatrix} \widehat{b}_{0,1} & \widehat{b}_{0,2} & \cdots & \widehat{b}_{0,M} \\ \widehat{b}_{1,1} & \widehat{b}_{1,2} & \cdots & \widehat{b}_{1,M} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{b}_{K,1} & \widehat{b}_{K,2} & \cdots & \widehat{b}_{K,M} \end{bmatrix}.$$

A j-ésima coluna da matriz $\widehat{\mathbf{B}}$ corresponde aos coeficientes estimados para o modelo referente à variável resposta $\mathbf{y}_j$, $j = 1, 2, \ldots, M$.

2.5 Algoritmo NIPALS

O algoritmo padrão usado para o cálculo dos componentes da regressão PLS é o Nonlinear Iterative Partial Least Squares (NIPALS), desenvolvido originalmente por Herman Wold (1966). Existem na literatura várias versões com pequenas alterações desse algoritmo e, assim como no algoritmo original, todas trabalham com as matrizes de dados originais $\mathbf{X}$ e $\mathbf{Y}$ padronizadas (escalonadas e centradas em zero). O algoritmo NIPALS é descrito abaixo:

1. Faça $\mathbf{u}$ igual a uma das colunas de $\mathbf{Y}$;

2. Determine uma coluna dos pesos de $\mathbf{W}$, utilizando

$$\mathbf{w} = \mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u};$$

3. Determine uma coluna dos $\mathbf{T}$, por meio de

$$\mathbf{t} = \mathbf{X}\mathbf{w};$$

4. Determine os pesos de $\mathbf{Y}$, $\mathbf{c}$, usando

$$\mathbf{c} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t};$$

5. Faça a atualização dos fatores de $\mathbf{Y}$, $\mathbf{u}$, através de

$$\mathbf{u} = \mathbf{Y}\mathbf{c}/\mathbf{c}'\mathbf{c}$$

6. Teste a convergência de $\mathbf{t}$, isto é, $\|\mathbf{t}_{velho} - \mathbf{t}_{novo}\| / \|\mathbf{t}_{novo}\| \leq \epsilon$, onde ϵ é uma constante predeterminada. Se não houver convergência retorne ao passo 2, caso contrário siga para o passo 7. Caso haja apenas uma variável resposta ($M = 1$) o procedimento converge em uma única iteração.

7. Faça

$$\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$$

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}'$$

$$\mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{c}'$$

8. Continue com o próximo componente (volte ao passo 1) até que a validação cruzada (veja seção 2.6) indique que o número de componentes é adequado.

2.6 O Número de Componentes

Com um grande número de variáveis explicativas, e possivelmente muitas destas sendo correlacionadas, há um considerável risco de ocorrer sobreajuste, isto é, obter um modelo bem ajustado e com pouco ou nenhum poder de predição. Por isso, é necessário verificar o poder preditivo para cada componente adicionado, e então parar a extração destes quando começarem a ocorrer fatores com baixo poder preditivo.

A validação cruzada é uma solução prática e confiável para verificar esse poder preditivo. Este tem sido o teste padrão na análise de regressão PLS, e está incorporado em grande parte dos softwares que tratam de PLS, como R, SAS e Matlab.

Basicamente, a validação cruzada funciona particionando o conjunto de dados em um número de grupos, G, um grupo de cada vez é omitido. Os dados dos outros grupos são usados para construir um modelo. Este modelo é usado para predizer os valores de $\mathbf{Y}$ dos dados omitidos, e as predições são comparadas com os valores omitidos, deve-se então reservar os resíduos destas predições. Este processo se repete até que cada grupo tenha sido omitido uma vez, e o total da soma dos quadrados destas diferenças é calculado, temos então a estatística $PRESS$ (preditive residual sum of squares), soma dos quadrados dos resíduos preditivos. Componentes são adicionados ao modelo até que o próximo componente aumente o valor de $PRESS$.

2.7 Número de Modelos

A regressão PLS tem a capacidade de modelar em uma análise várias respostas conjuntamente. Quando as variáveis resposta são correlacionados, eles devem ser analisados conjuntamente. Caso as variáveis resposta sejam descorrelacionados, uma única regressão PLS tende a gerar muitos componentes e a modelagem separada de cada resposta é indicada. Por isso, é usual iniciar com uma análise de componentes principais apenas na matriz $\mathbf{Y}$. Se o número de componentes resultantes dessa análise for pequeno quando comparado ao número de respostas, M, então as variáveis resposta são correlacionadas e uma única regressão PLS deve ser feita. Se entretanto, as variáveis resposta estiverem separadas em grandes grupos (na visialização dos gráficos das cargas), então faça uma regressão PLS para cada um desses grupos.

2.8 Erros Padrão e Intervalo de Confiança

Muitos esforços têm sido dedicados para construir teoricamente intervalos de confiança para os parâmetros da regressão PLS. Recentemente surgiram trabalhos que tratam desse assunto considerando a regressão PLS como um modelo de regressão de variáveis latentes.

Um caminho para estimar os erros padrão e obter intervalos de confiança diretamente dos dados é usar $jackknife$ ou *bootstrap*. $Jackknife$ foi recomendado por Wold (1966) em seu trabalho original sobre PLS, e tem recentemente sido revisto por Martens and Martens (2000) e outros.

Faremos aqui uma breve descrição do procedimento *bootstrap* para a obtenção dos intervalos de confiança para os coeficientes do PLS.

Considere a matriz de dados $\mathbf{D} = (\mathbf{t}_1, \ldots, \mathbf{t}_A, \mathbf{y})$, que é formada por uma variável resposta e pelos A fatores extraídos pelo PLS. A obtenção dos intervalos *bootstrap* seguem os seguintes passos:

- Obtenha B amostras *bootstrap* da matriz $\mathbf{D}$;

- Para $b = 1, \ldots, B$ calcule

$$\widehat{\mathbf{c}}^{(b)\prime} = (\mathbf{T}^{(b)\prime}\mathbf{T}^{(b)})^{-1}\mathbf{T}^{(b)}\mathbf{y}^{(b)} \quad \text{e} \quad \widehat{\mathbf{b}}^{(b)} = \mathbf{W}\widehat{\mathbf{c}}^{(b)\prime},$$

em que $(\mathbf{T}^{(b)}, \mathbf{y}^{(b)})$ é a b-ésima amostra *bootstrap*, $\mathbf{c}^{(b)}$ é o vetor de estimativas ligando os fatores a $\mathbf{y}$, $\mathbf{b}^{(b)}$ é o vetor de estimativas ligando as variáveis originais a $\mathbf{y}$ e $\mathbf{W}$ é a matriz de pesos, que não foi alterada.

Após obter as amostras *bootstrap*, os intervalos de confiança de $100(1-\alpha)\%$ para b_j, $j = 1, \ldots, K$ são dados por

$$\left(\mathbf{l}_j'\widehat{\mathbf{b}} - t_{(1-\alpha/2),n-1}\sqrt{\widehat{Var}_B(\mathbf{l}_j'\widehat{\mathbf{b}})}\ ;\ \mathbf{l}_j'\widehat{\mathbf{b}} + t_{(\alpha/2),n-1}\sqrt{\widehat{Var}_B(\mathbf{l}_j'\widehat{\mathbf{b}})}\right)$$

sendo $\mathbf{l}_j$ um vetor de mesma dimensão de $\mathbf{b}$, com todas as entradas iguais a zero, exceto pela j-ésima entrada que é igual a 1; $\widehat{Var}_B(\mathbf{l}_j'\widehat{\mathbf{b}})$ é a variância *bootstrap* para a j-ésima entrada de $\widehat{\mathbf{b}}$ e $t_{(\alpha),N-1}$ é o α-ésimo percentil da distribuição t de Student com $N-1$ graus de liberdade.

Capítulo 3

O Modelo de Regressão PLS com Erros Heteroscedásticos

3.1 Introdução

Os erros no modelo de regressão PLS têm os mesmos pressupostos que a regressão linear múltipla, exceto pela distribuição. Na regressão múltipla, para efeitos inferência, supõe-se que os erros têm distribuição normal multivariada com vetor de médias nulo e matriz de covariâncias $\sigma^2\mathbf{I}$, em que $\mathbf{I}$ é uma matriz identidade $(N \times N)$. O método PLS é uma abordagem livre de distribuição. Dessa forma, os erros da regressão possuem vetor de médias nulo e matriz de covariâncias igual a $\sigma^2\mathbf{I}$, mas sem distribuição definida.

Neste capítulo é desenvolvido um método de regressão PLS que construa os estimadores dos coeficientes considerando a presença de heteroscedasticidade.

Consideramos neste capítulo que o modelo é composto de uma variável resposta, ou seja,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}.$$

3.2 Método PLS

O método PLS reduz a dimensão das covariáveis, produzindo componentes que são combinações lineares das covariáveis $\mathbf{X}$ e são dadas por

$$\mathbf{T}=\mathbf{XW},$$

em que $\mathbf{T}$ são os fatores extraídos de $\mathbf{X}$ e $\mathbf{W}$ é uma matriz de pesos. O número de componentes é menor que o número de variáveis preditoras, e esses fatores extraídos são ortogonais entre si.

Após a extração dos fatores, faz-se a regressão de $\mathbf{Y}$ sobre $\mathbf{T}$, na forma

$$\mathbf{y} = \mathbf{Tc}' + \mathbf{e}, \tag{3.1}$$

encontrando os coeficientes $\mathbf{c}$. Em seguida, substitui-se $\mathbf{T}$ por $\mathbf{XW}$ em (3.1) para obter

$$\mathbf{y} = \mathbf{XWc}' + \mathbf{e} = \mathbf{Xb} + \mathbf{e}.$$

Dessa maneira, a matriz dos coeficientes da regressão para as covariáveis originais $\mathbf{X}$ seria $\mathbf{b} = \mathbf{Wc}'$. O estimador para $\mathbf{c}$ é obtido por mínimos quadrados, e é dado por

$$\widehat{\mathbf{c}}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}.$$

Consequentemente,

$$\widehat{\mathbf{b}} = \mathbf{W}\widehat{\mathbf{c}}'.$$

3.3 O Método PLS para Erros Heteroscedástico

Os pressupostos relacionados aos erros do modelo PLS incluem a correlação nula entre os termos, a média zero e a variância constante. Assim como em regressão múltipla, a presença de heteroscedasticidade nos erros causa, na regressão PLS, um aumento da variância dos estimadores dos coeficientes do modelo.

Pode-se notar que os erros, $\mathbf{e}$, na equação (3.1) são os mesmos que em $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, pois $\mathbf{y} = \mathbf{Tc}' + \mathbf{e} = \mathbf{XWc}' + \mathbf{e} = \mathbf{Xb} + \mathbf{e}$. Assim, para resolver o problema da heteroscedasticidade no modelo $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ basta resolvê-lo no modelo $\mathbf{y} = \mathbf{Tc}' + \mathbf{e}$.

Como visto anteriormente, a estimação de $\mathbf{c}$ é feita por meio de mínimos quadrados ordinários (ordinary least squares, OLS). Sabe-se que, na presença de heteroscedasticidade, deve-se substituir o estimador OLS pelo de mínimos quadrados generalizados (generalized least squares, GLS). Portanto, baseando-se nessa lógica, devemos estimar $\mathbf{c}$ como

$$\widehat{\mathbf{c}}'_{het} = (\mathbf{T}'\mathbf{\Omega}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{\Omega}^{-1}\mathbf{y},$$

em que $\mathbf{\Omega}$ é a matriz de covariâncias dos erros. Dessa forma, o estimador de $\mathbf{b}$ considerando a heteroscedasticidade é dado por

$$\widehat{\mathbf{b}}_{het} = \mathbf{W}\widehat{\mathbf{c}}'_{het}.$$

Quando a matriz de covariâncias dos erros, $\mathbf{\Omega}$, é completamente conhecida é fácil calcular $\widehat{\mathbf{c}}'_{het}$ e $\widehat{\mathbf{b}}_{het}$, mas quando $\mathbf{\Omega}$ é desconhecida é necessário estimá-la. Quando a função que determina a heteroscedasticidade é conhecida, mas com alguns parâmetros desconhecidos, uma saída seria estimar esta heteroscedasticidade.

Considere

$$Var(e_i) = \sigma_i^2 = \sigma^2 g(\mathbf{Z}_i, \boldsymbol{\rho}),$$

em que σ^2 é um parâmetro desconhecido; $\boldsymbol{\rho}$ é um vetor $q \times 1$ de parâmetros desconhecidos; $\mathbf{Z}_i$ é um vetor referente a i-ésima observação de um subconjunto das covariáveis; g é uma função diferenciável conhecida. Com $g_i = g(\mathbf{Z}_i, \boldsymbol{\rho})$, as funções mais utilizadas na literatura são $g_i = x_{ik}^{\rho}$, $g_i = \exp(\rho x_{ik})$ e $g_i = \rho_0 + \rho_1 x_{ik}$.

Com o objetivo de estimar os parâmetros de uma estrutura heteroscedástica e os coeficientes do modelo de regressão de $\mathbf{y}$ sobre $\mathbf{T}$, consideremos uma distribuição para os erros. A idéia é a mesma utilizada por Bastien *et al.* (2005) no modelo de regressão linear generalizada PLS (PLS generalised linear regression, PLS-GLR). O procedimento apresentado por Bastien *et al.* (2005) envolve os seguintes passos:

1. Extração dos A fatores através do PLS;

2. Estimação do modelo de regressão de $\mathbf{y}$ sobre $\mathbf{T}$, considerando uma distribuição para os erros, além da estrutura de heteroscedasticidade;

3. Estimando por máxima verossimilhança os coeficientes $\widehat{\mathbf{c}}$ do modelo para $\mathbf{T}$, obtêm-se as estimativas $\widehat{\mathbf{b}}$ para os dados originais $\mathbf{X}$ fazendo

$$\widehat{\mathbf{b}} = \mathbf{W}\widehat{\mathbf{c}}';$$

4. Utiliza-se um método de reamostragem para verificar a significância dos coeficientes.

Neste trabalho, o passo 2 inclui uma familia de distribuições para os erros tendo a distribuição normal como um caso particular, a distribuição normal assimétrica, e a presença de uma estrutura heteroscedástica.

3.4 Modelo Normal Assimétrico

A principal distribuição para os erros de um modelo linear é a distribuição normal. Entretanto, a suposição de normalidade nem sempre é satisfeita devido à falta de simetria dos dados. É proposta então uma família de distribuições mais geral, que consiga modelar a assimetria dos dados e, além disso, incluir a distribuição normal como um caso particular. Esta família de distribuições é denominada normal assimétrica. Azzalini (2005) apresentou uma discussão em distribuições normais assimétricas com aplicações em modelos de regressão.

3.4.1 A Distribuição Normal Assimétrica

Uma variável Y é dita ter distribuição normal assimétrica com parâmetro de locação μ, parâmetro de escala σ^2 e parâmetro de assimetria λ, denotada por $Y \sim NA(\mu, \sigma^2, \lambda)$, se esta tem uma função de densidade dada por

$$f(y) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right),$$

em que $\phi(\cdot)$ e $\Phi(\cdot)$ são a função densidade de probabilidade e a função de distribuição acumulada da distribuição normal padrão, respectivamente. Note que se $\lambda = 0$, a função densidade de Y se reduz à da distribuição normal.

3.4.2 Função de Verossimilhança

Considere o modelo

$$\begin{aligned} Y_i &= c_0 + c_1 t_{i1} + \cdots + c_A t_{iA} + e_i \\ &= \mathbf{T}_i \mathbf{c}' + e_i, \quad e_i \sim NA(0, \sigma^2, \lambda), \quad i = 1, \ldots, N, \end{aligned} \tag{3.0}$$

com $\mathbf{T}_i$ representando a i-ésima linha da matriz $\mathbf{T}$.

De (3.1), temos que $Y_i \sim NA(\mathbf{T}_i\mathbf{c}', \sigma^2, \lambda)$, demodo que

$$E(Y_i) = \mathbf{T}_i\mathbf{c}' + \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}\sigma, \quad Var(Y_i) = \left(1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}\right)\sigma^2 \quad i = 1, \ldots, N.$$

Em termos de previsão para Y, geralmente considera-se $\widehat{Y}_i = \mathbf{T}_i\widehat{\mathbf{c}}' = \mathbf{X}_i\widehat{\mathbf{b}}$ como preditor para $Y_i|\mathbf{X}_i$. Entretanto, observando a expressão de $E(Y_i)$ temos que um preditor mais adequado para $Y_i|\mathbf{X}_i$ é dado por

$$\widehat{Y}_i = \mathbf{X}_i\widehat{\mathbf{b}} + \sqrt{\frac{2}{\pi}}\frac{\widehat{\lambda}\widehat{\sigma}}{\sqrt{1+\widehat{\lambda^2}}}$$

Seja $\boldsymbol{\theta} = (\mathbf{c}, \sigma^2, \lambda)'$. Então, a função de verossimilhança para $\boldsymbol{\theta}$ pode ser escrita como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N}\left[\frac{2}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2}\right\}\Phi\left(\lambda\frac{y_i - \mathbf{T}_i\mathbf{c}'}{\sigma}\right)\right]$$

e a respectiva função log-verossimilhança é dada por

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2} + \log\Phi(\kappa_i)\right]. \tag{3.0}$$

com $\kappa_i = \lambda(y_i - \mathbf{T}_i\mathbf{c}')/\sigma$.

A matriz de informação de Fisher para $\boldsymbol{\theta}$, a partir de (3.4.2), é dada por

$$I_Y(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1+\lambda^2 a_0}{\sigma^2}\mathbf{T}'\mathbf{T} & \sqrt{\frac{1}{2\pi}}\frac{\lambda}{\sigma^3}d_1\mathbf{T}'\mathbf{1}_N & \frac{d_2}{\sigma}\sqrt{\frac{2}{\pi}}\mathbf{T}'\mathbf{1}_N \\ \sqrt{\frac{1}{2\pi}}\frac{\lambda}{\sigma^3}d_1\mathbf{1}_N'\mathbf{T} & \frac{N(2+a_2)}{4\sigma^4} & -\frac{Na_2}{2\sigma^2\lambda} \\ \frac{d_2}{\sigma}\sqrt{\frac{2}{\pi}}\mathbf{1}_N'\mathbf{T} & -\frac{Na_2}{2\sigma^2\lambda} & \frac{Na_2}{\lambda^2} \end{bmatrix}, \tag{3.0}$$

com $d_1 = \big((1+2\lambda^2)/(1+\lambda^2)^{3/2}\big) + a_1\sqrt{\pi/2}$, $d_2 = \big(1/(1+\lambda^2)^{3/2}\big) - a_1\sqrt{\pi/2}$, $\mathbf{1}_N = (\underbrace{1, \ldots, 1}_{N})'$, e $a_j = E\big\{\kappa_i^j\big[\big(\phi^2(\kappa_i)\big)/\big(\Phi^2(\kappa_i)\big)\big]\big\}$, $j = 0, 1, 2$ que devem ser obtidos numericamente. Note que está matriz esta definida para $\lambda \neq 0$.

Os estimadores de máxima verossimilhança de σ^2, λ e $\mathbf{c}$ somente são obtidos por meio de procedimentos iterativos, como Newton-Raphson ou algoritmo EM. Os estimadores dos coeficientes $b_0, b_1, \ldots, b_p$ das covariáveis originais $X_1, X_2, \ldots, X_p$ são obtidos fazendo

$$\widehat{\mathbf{b}} = \mathbf{W}\widehat{\mathbf{c}}'_{mv}.$$

em que $\widehat{\mathbf{c}}_{mv}$ é o estimador de máxima verossimilhança para $\mathbf{c}$.

Com a distribuição para os erros disponivel é possivel inserir a estrutura de hetorescedasticidade no modelo, e consequentemente estimar os seus parâmetros. A próxima seção apresenta o modelo normal assimétrico com erros heteroscedásticos.

3.5 Modelo Normal Assimétrico Heteroscedástico

Na seção anterior foi apresentado o modelo normal assimétrico, no qual os parâmetros σ^2 e λ são constantes. Desse modo, a variância da i-ésima observação é dada por $Var\Big(Y_i\Big) = \Big(1-\big(2\lambda^2/\pi(1+\lambda^2)\big)\Big)\sigma^2$. Entretanto, esse pressuposto de homoscedasticidade pode ser violado, de modo que o parâmetro σ^2 ou λ seja diferente em cada observação. Desse modo, é possivel que

$$Var(Y_i) = \left(1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}\right)\sigma_i^2$$

ou ainda

$$Var(Y_i) = \left(1 - \frac{2\lambda_i^2}{\pi(1+\lambda_i^2)}\right)\sigma^2.$$

Em ambas as formas de variância acima esta caracterizada a estrutura heteroscedástica, uma vez que cada observação tem uma variância diferente. Existe também a possibilidade de ambos os parâmetros serem não constantes.

3.5.1 Modelo com σ^2 não constante

Similar à seção 3.3, considere $\sigma_i^2 = \sigma^2 g(\mathbf{Z}_i, \boldsymbol{\rho}) = \sigma^2 g_i$, sendo σ^2 um parâmetro desconhecido; $\boldsymbol{\rho}$ um vetor $q \times 1$ de parâmetros desconhecidos; $\mathbf{Z}_i$ um vetor referente à i-ésima observação de um subconjunto das covariáveis; g uma função diferenciável conhecida.

A função log-verossimilhança de $\boldsymbol{\xi}_1 = (\boldsymbol{\rho}', \boldsymbol{\theta}')'$ pode ser escrita como

$$l(\boldsymbol{\xi}_1) = \sum_{i=1}^{N} \left[\log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2}\log g_i - \frac{(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2 g_i} + \log \Phi(\kappa_i^*) \right], \quad (3.0)$$

com $\kappa_i^* = \lambda(y_i - \mathbf{T}_i\mathbf{c}')/(\sigma g_i^{1/2})$.

Para construir a matriz de informação de $\boldsymbol{\xi_1}$ obtêm-se as derivadas de segunda ordem de $l(\boldsymbol{\xi_1})$ em relação aos vetores de parâmetros $\boldsymbol{\rho}$ e $\boldsymbol{\theta}$. Calculando a esperança do negativo das derivadas de segunda ordem obtemos

$$\begin{aligned}
I_{\boldsymbol{\rho\rho}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_1})}{\partial\boldsymbol{\rho}\partial\boldsymbol{\rho}'}\right] = \frac{2+a_2}{4}\mathbf{G}'\mathbf{G}, \\
I_{\boldsymbol{\rho c}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_1})}{\partial\boldsymbol{\rho}\partial\mathbf{c}}\right] = \sqrt{\frac{2}{\pi}}\frac{c_1}{2}\frac{\lambda}{\sigma}\mathbf{G}'\mathbf{T} \\
I_{\boldsymbol{\rho\sigma^2}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_1})}{\partial\boldsymbol{\rho}\partial\sigma^2}\right] = \frac{2+a_2}{4\sigma^2}\mathbf{G}'\mathbf{1}_N \\
I_{\boldsymbol{\rho\lambda}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_1})}{\partial\boldsymbol{\rho}\partial\lambda}\right] = -\frac{a_2}{2\lambda}\mathbf{G}'\mathbf{1}_N
\end{aligned}$$

com $\mathbf{G}' = (\mathbf{g}_1', \ldots, \mathbf{g}_N')$, $\mathbf{g}_i = \partial g_i/\partial\boldsymbol{\rho}' = (\partial g_i/\partial\rho_1, \ldots, \partial g_i/\partial\rho_q)$, $i = 1, \ldots, N$.

Então a matriz de informação de Fisher para $\boldsymbol{\xi_1}$ é dada por

$$I_Y(\boldsymbol{\xi_1}) = \begin{bmatrix} I_{\boldsymbol{\rho\rho}} & I_{\boldsymbol{\rho\theta}} \\ I_{\boldsymbol{\rho\theta}} & I_Y(\boldsymbol{\theta}) \end{bmatrix},$$

sendo $I_{\boldsymbol{\rho\theta}} = [I_{\boldsymbol{\rho c}}, I_{\boldsymbol{\rho\sigma^2}}, I_{\boldsymbol{\rho\lambda}}]$, e $I_Y(\boldsymbol{\theta})$, como dado em (3.4.2), é a matriz de informação referente ao modelo homocedástico.

3.5.2 Modelo com λ não constante

Considere $\lambda_i = \lambda h(\mathbf{V}_i, \boldsymbol{\gamma}) = \lambda h_i$ sendo λ um parâmetro desconhecido; $\boldsymbol{\gamma}$ um vetor $q^* \times 1$ de parâmetros desconhecidos; $\mathbf{V}_i$ um vetor referente à i-ésima observação de um conjunto de covariáveis; h uma função diferenciável conhecida.

A função log-verossimilhança de $\boldsymbol{\xi}_2 = (\boldsymbol{\gamma}', \boldsymbol{\theta}')'$ pode ser escrita como

$$l(\boldsymbol{\xi}_2) = \sum_{i=1}^{N} \left[\log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2} + \log \Phi(\kappa_i^{**}) \right]. \quad (3.\text{-}5)$$

com $\kappa_i^{**} = \lambda h_i(y_i - \mathbf{T}_i\mathbf{c}')/\sigma$.

Similar à seção anterior calculamos as derivadas de segunda ordem de $l(\boldsymbol{\xi_2})$ em relação aos vetores de parâmetros $\boldsymbol{\gamma}$ e $\boldsymbol{\theta}$. Calculando a esperança do negativo destas derivadas tem-se

$$\begin{aligned} I_{\boldsymbol{\gamma\gamma}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_2})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right] = a_2\mathbf{H}'\mathbf{H}, \\ I_{\boldsymbol{\gamma c}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_2})}{\partial\boldsymbol{\gamma}\partial\mathbf{c}}\right] = \sqrt{\frac{2}{\pi}}\frac{\lambda c_1}{\sigma}\mathbf{H}'\mathbf{T} \\ I_{\boldsymbol{\gamma\sigma^2}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_2})}{\partial\boldsymbol{\gamma}\partial\sigma^2}\right] = -\frac{a_2}{2\sigma^2}\mathbf{H}'\mathbf{1}_N \\ I_{\boldsymbol{\gamma\lambda}} &= E\left[-\frac{\partial^2 l(\boldsymbol{\xi_2})}{\partial\boldsymbol{\gamma}\partial\lambda}\right] = \frac{a_2}{\lambda}\mathbf{H}'\mathbf{1}_N \end{aligned}$$

com $\mathbf{H}' = (\mathbf{h}_1', \ldots, \mathbf{h}_N')$, $\mathbf{h}_i = \partial h_i/\partial\boldsymbol{\gamma}' = (\partial h_i/\partial\gamma_1, \ldots, \partial h_i/\partial\gamma_{q^*})$, $i = 1, \ldots, N$.

A matriz de informação de Fisher para $\boldsymbol{\xi_2}$ é dada por

$$I_Y(\boldsymbol{\xi_2}) = \begin{bmatrix} I_{\boldsymbol{\gamma\gamma}} & I_{\boldsymbol{\gamma\theta}} \\ I_{\boldsymbol{\gamma\theta}} & I_Y(\boldsymbol{\theta}) \end{bmatrix},$$

com $I_{\boldsymbol{\gamma\theta}} = [I_{\boldsymbol{\gamma c}}, I_{\boldsymbol{\gamma\sigma^2}}, I_{\boldsymbol{\gamma\lambda}}]$.

Na próxima seção é feita uma comparação do modelo PLS usual com o modelo proposto.

3.6 Simulação

Na seção anterior foi apresentado um método para estimar os coeficientes da regressão PLS considerando a presença de heteroscedasticidade nos erros. Nesta seção este método é comparado com o método PLS usual via simulação. Os dados são gerados segundo o seguinte procedimento:

1. Gera-se uma amostra de tamanho N de K covariáveis ($K = N$), em que as $K/2$ primeiras seguem distribuição $U(0, 10)$, e as demais covariáveis são funções lineares de pelo menos uma das $K/2$ primeiras mais um erro;

2. Gera-se uma amostra de erros $e_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, N$, com $\sigma_i^2 = \sigma^2 x_{i1}^{\rho}$, $\sigma^2 = 0,5$;

3. Determina-se y_i através da equação $y_i = b_0 + b_1 x_{i1} + \cdots + b_K x_{iK} + e_i$, $i = 1, \ldots, N$, com $b_0 = -5$ e $b_1 = b_2 = \cdots = b_K = 1,5$; e

4. Estimam-se os coeficientes da regressão utilizando o método proposto e via PLS usual.

São usados diferentes graus de heteroscedasticidade($\rho = 0,5,\ 1,\ 2$) e diferentes tamanhos de amostra ($N = 30,\ 100,\ 200$). Para cada valor de ρ e N, são realizadas 2000 replicações com as covariáveis fixas, ou seja, repetimos o procedimento acima, do passo 2 ao passo 4, 2000 vezes.

Para comparar os modelos considera-se o ajuste e o poder de predição dos mesmos. Na seção 3.6.1 os modelos serão comparados segundo o ajuste, enquanto na seção 3.6.2 os mesmos serão comparados de acordo com o poder de predição.

3.6.1 Comparando o Ajuste

Na comparação consideramos os erros quadráticos médios (EQM), as variâncias e os vícios dos estimadores dos coeficientes de cada método. Essas quantidades são calculadas a partir das 2000 amostras geradas da seguinte forma

$$\begin{aligned} EQM(\widehat{b}_j) &= \frac{1}{1999}\sum_{l=1}^{2000}(\widehat{b}_j^{(l)} - b_{j,real})^2, \\ Var(\widehat{b}_j) &= \frac{1}{1999}\sum_{l=1}^{2000}(\widehat{b}_j^{(l)} - \bar{\widehat{b}}_j)^2 \text{ e} \\ \text{Vício}(\widehat{b}_j) &= \Big(EQM(\widehat{b}_j) - Var(\widehat{b}_j)\Big)^{1/2}, \end{aligned}$$

em que, para $j = 0, \ldots, K$, $\widehat{b}_j^{(l)}$ denota a estimativa de b_j usando a l-ésima amostra, $\bar{\widehat{b}}_j$ é a média de $\widehat{b}_j^{(1)}, \ldots, \widehat{b}_j^{(2000)}$ e $b_{j,real}$ é o verdadeiro valor do coeficiente b_j.

O motivo de usar a distribuição normal para os erros é para que a heteroscedasticidade seja a única dificuldade encontrada pelo método PLS, evitando assim a assimetria. Dessa forma, a função de verossimilhança de $\boldsymbol{\psi} = (\boldsymbol{\rho}', \mathbf{c}, \sigma^2)$, para a regressão de $\mathbf{y}$ sobre $\mathbf{T}$ é dada por

$$L(\boldsymbol{\psi}) = \prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma^2 x_{i1}^{\rho}}}\exp\left\{\frac{-(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2 x_{i1}^{\rho}}\right\}\right].$$

O logaritmo da função de verossimilhança é dado por

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2}\log x_{i1}^{\rho} - \frac{(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2 x_{i1}^{\rho}} \right].$$

Os estimadores de máxima verossimilhança de σ^2, $\boldsymbol{\rho}$ e $\mathbf{c}$ são obtidos pelo método de Newton-Raphson usando o pacote proc nlp do sistema SAS.

A Tabela 3.1 mostra, para cada medida, a proporção de coeficientes em que o método proposto obteve um melhor desempenho que o PLS. Por exemplo, se na Tabela 3.1 estiver $EQM = 0,1$, significa que em 10% dos coeficientes do modelo ocorreu $EQM(b_{pls}) > EQM(b_{modelo\ proposto})$, ou seja, em 10% dos coeficientes o modelo proposto foi melhor.

TABELA 3.1: Comparação entre o modelo proposto e o PLS usual

N	Medida	$\rho = 0.5$	$\rho = 1$	$\rho = 2$
30	VAR	0	0,1935	0,7097
	EQM	0,0645	0,1935	0,7097
	Vício	0,3548	0,3548	0,3226
100	VAR	0,4257	0,8812	1
	EQM	0,5149	0,5842	0,9901
	Vício	0,5347	0,4653	0,5347
200	VAR	0,5771	0,9900	1
	EQM	0,5373	0,5970	0.9552
	Vício	0,5274	0,4627	0,4229

Observando apenas a variância e o EQM, pode-se verificar que quando o tamanho da amostra e/ou o grau da heteroscedasticidade aumentam, melhora o desempenho do método proposto em relação ao PLS. Entretanto, analisando o vício vemos que o mesmo não se altera muito, mesmo com as variações do tamanho amostral ou do grau de heteroscedasticidade. Na maioria dos cenários o método PLS foi superior ao método proposto quando considerado o vício, ou seja, obteve menores valores de vício.

Os resultados indicam que quando a heteroscedasticidade é forte, o uso do método proposto é indicado para qualquer tamanho de amostra, e quando é moderada o método proposto deve ser usado para amostras maiores.

A comparação dos métodos foi realizada considerando os coeficientes **b** para regressão sobre as covariáveis originais **X**, e não os coeficientes **c** referentes aos fatores **T**. Medidas de comparação como EQM e vício dependem dos valores reais dos parâmetros. Na geração dos dados são atribuídos valores para **b** e não para **c**, com isso não é possível obter valores para EQM e vício de **c**.

3.6.2 Comparando o Poder de Predição

Nesta seção é verificado se o método proposto, que considera a informação de heteroscedasticidade, tem um poder de predição superior ao método PLS usual.

São geradas amostras teste em cada uma das 2000 replicações, para todos os diferentes cenários. Essas amostras são geradas segundo o procedimento descrito anteriormente. As amostra teste são usadas somente para verificar o poder de predição do modelo e não são usadas na estimação dos mesmos. A medida usada na comparação é a Diferença Relativa (DR), que é dada por

$$DR = \frac{1}{N_{teste}} \sum_{i=1}^{N_{teste}} \left| \frac{Y_i - \widehat{Y}_i}{Y_i} \right| \times 100, \quad Y_i \neq 0,$$

em que $N_{teste} = 20$ é o tamanho da amostra teste, Y_i é a i-ésima resposta da amostra teste e $\widehat{Y}_i$ é o valor predito para Y_i segundo o modelo estimado. Quanto menor o valor obtido, melhor é a predição do método.

A Tabela 3.2 mostra em cada cenário a proporção de vezes que o método proposto obteve um poder de predição superior ao método PLS usual.

TABELA 3.2: Proporção de vezes em que o método proposto fez melhores predições

N	$\rho = 0,5$	$\rho = 1$	$\rho = 2$
30	0,1150	0,1700	0,3300
100	0,5580	0,6985	0,7690
200	0,5535	0,5910	0,6625

Pode-se observar que para amostras pequenas o PLS usual obteve um desempenho superior ao método proposto. Com amostras maiores o modelo proposto foi melhor.

Pode-se verificar ainda que para qualquer tamanho de amostra, à medida que o grau de heteroscedasticidade cresce, o desempenho do método proposto melhora.

Capítulo 4

Testes para Heteroscedasticidade em Regressão PLS

É possível detectar heteroscedasticidade por meio de uma inspeção visual de diagramas de dispersão dos resíduos contra as covariáveis. Se parecer que os resíduos possuem a mesma variabilidade em torno de uma linha imaginária, então, provavelmente, não existe heteroscedasticidade. Caso contrário, uma verificação mais formal deve ser feita.

A literatura apresenta alguns testes para a detecção de heteroscedasticidade. Entretanto, todos foram desenvolvidos para regressão por mínimos quadrados ordinários. Exporemos a seguir os testes de White e de Goldfeld-Quandt, adaptados para regressão PLS.

4.1 Teste de White

O teste proposto por White (1980) examina se a variância dos erros é afetada por alguma das variáveis regressoras, seus quadrados ou seus produtos cruzados. É baseado na idéia de que a variância do estimador de mínimos quadrados ordinários é igual à variância do estimador de mínimos quadrados generalizados. O teste detecta a presença de heteroscedasticidade somente se esta afetar a consistência do estimador da matriz de

variâncias e covariâncias dado por

$$\mathbf{V} = s^2(\mathbf{X'X})^{-1}$$

com $s^2 = \widehat{\mathbf{e}}'\widehat{\mathbf{e}}/(N-K)$ e $\widehat{\mathbf{e}}$ sendo o vetor de resíduos obtidos na regressão por mínimos quadrados ordinários.

A estatística de teste para detectar a heteroscedasticidade nos dados é igual ao tamanho da amostra, N, multiplicado pelo coeficiente de determinação R^2, calculado a partir da regressão ajustada entre os quadrados dos resíduos e o produto cruzado das variáveis regressoras mais a constante. No caso de duas covariáveis, a regressão a ser ajustada seria

$$\widehat{e}_i^2 = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2}$$

A estatística de White, NR^2, segue, assintoticamente, distribuição qui-quadrado com o número de graus de liberdade igual ao número de variáveis regressoras do modelo usado para calcular R^2.

A dificuldade em utilizar este teste na regressão PLS é que geralmente o número de observações é menor que o número de covariáveis. Desse modo, para utilizar o teste de White na regressão PLS, devemos primeiramente extrair os componentes $\mathbf{t}_1, \ldots, \mathbf{t}_A$ e depois disso seguir todo o procedimento do teste, substituindo as covariáveis pelo fatores extraídos.

O teste pode ser aplicado caso o número de fatores extraidos for pequeno de tal modo que, na regressão dos resíduos sobre estes fatores e seus produtos cruzados, o número de covariáveis não se aproxime demais do número de observações. Deste modo, se forem extraidos dois fatores ($A = 2$), deve-se obter o resíduos da regressão PLS e ajustar a regressão

$$\widehat{e}_i^2 = \beta_0 + \beta_1 t_{i1} + \beta_2 t_{i2} + \beta_3 t_{i1}^2 + \beta_4 t_{i2}^2 + \beta_5 t_{i1} t_{i2}.$$

Como anteriormente, deve-se calcular a estatística NR^2 para essa regressão e compará-la à distribuição qui-quadrado.

4.2 Teste de Goldfeld-Quandt

A lógica do teste proposto por Goldfeld and Quandt (1965) é dividir a amostra em duas partes e verificar se existe forte diferença entre a soma dos quadrados dos resíduos de cada subamostra. Neste teste as N observações são primeiramente ordenadas de acordo com os valores da variável regressora, X_k, da qual suspeitamos causar heteroscedasticidade. Em seguida é necessário dividir a amostra ordenada em três partes, desconsiderando os valores da subamostra central (25% aproximadamente). Duas regressões são ajustadas com as outras duas subamostras ordenadas, obtendo-se a soma do quadrado dos resíduos das mesmas. A estatística de teste é dada por

$$GQ = \frac{SQR_2/(N_2 - K)}{SQR_1/(N_1 - K)},$$

em que SQR_j e N_j, $j = 1, 2$ são, respectivamente, a soma dos quadrados dos resíduos e o tamanho da subamostra j. Se os erros forem normalmente distribuidos, então sob a hipótese de homoscedasticidade, a estatística GQ tem distribuição F com $N_2 - K$ e $N_1 - K$ graus de liberdade.

A dificuldade da utilização deste teste na regressão PLS é, novamente, que o número de observações é próximo ao número de covariáveis. Com isso, deve-se trabalhar utilizando os fatores no lugar das covariáveis na obtenção das SQR's, e além disso substituir K por A. Para aplicar este teste na regressão PLS devemos proceder da seguinte forma:

1. Faça a extração dos A fatores via PLS usual;
2. Ordene as N observações, inclusive os fatores extraídos, de acordo com a covariável X_k, responsável pela heteroscedasticidade;
3. Divida a amostra em três partes como descrito anteriormente;
4. Ajuste duas regressões de $\mathbf{y}$ sobre os fatores extraídos com as duas subamostras ordenadas, obtendo a soma do quadrado dos resíduos de cada uma;
5. A estatística de teste é dada por

$$GQ = \frac{SQR_2/(N_2 - A)}{SQR_1/(N_1 - A)},$$

e tem, sob hipótese de homocedasticidade, distribuição F com $N_2 - A$ e $N_1 - A$ graus de liberdade.

4.3 Teste de Heteroscedasticidade para o Modelo Normal Assimétrico

Os testes de heteroscedasticidade citados são sensíveis ao pressuposto de normalidade. Por isso, quando assumimos a distribuição normal assimétrica para modelar os erros, devemos utilizar um teste adequado a esta distribuição. Xei *et al.* (2009) desenvolveram um teste, baseado no teste escore, para verificar a homocedasticidade dos erros em modelos normais assimétricos de regressão não-linear.

Como mostrado na seção 3.4.2, a variância da variável resposta é dada por

$$Var(Y_i) = \left(1 - \frac{2\lambda^2}{\pi(1+\lambda^2)}\right)\sigma^2.$$

Podemos observar que essa variância depende dos parâmetros σ^2 e λ. Dessa forma, para testar a hipótese da variância ser constante, devemos testar se estes parâmetros são iguais para todo i, $i = 1, \ldots, N$.

O trabalho de Xei *et al.* (2009) apresenta dois testes, que são:

- Teste de homogeneidade para o parâmetro de escala, σ^2; e
- Teste de homogeneidade para o parâmetro de assimetria, λ.

Nas próximas seções serão apresentadas adaptações destes testes para o modelo PLS que considera erros normais assimétricos.

4.3.1 Teste de Homogeneidade para o Parâmetro de Escala

Este teste verifica se o parâmetro σ^2 é constante, ou como visto na seção 3.3, uma função das covariáveis $\sigma_i^2 = \sigma^2 g(\mathbf{Z}_i, \boldsymbol{\rho})$.

Assume-se que existe um único valor $\boldsymbol{\rho}_0$ de $\boldsymbol{\rho}$ tal que $g(\mathbf{Z}_i, \boldsymbol{\rho}_0) = 1$ para todo i. Observe que quando $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ temos $\sigma_i^2 = \sigma^2$.

Dessa forma, o teste para a homogeneidade do parâmetro σ^2 é equivalente a testar as hipóteses

$$H_0 : \boldsymbol{\rho} = \boldsymbol{\rho}_0; \quad H_1 : \boldsymbol{\rho} \neq \boldsymbol{\rho}_0. \tag{4.0}$$

De acordo com a função log-verossimilhança dada por (3.5.1), temos que

$$\frac{\partial l(\boldsymbol{\xi}_1)}{\partial \boldsymbol{\rho}} = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\frac{1}{g_i}\frac{\partial g_i}{\partial \boldsymbol{\rho}} + \frac{(y_i - \mathbf{T}_i\mathbf{c}')^2}{2\sigma^2 g_i^2}\frac{\partial g_i}{\partial \boldsymbol{\rho}} + \frac{\phi(\kappa_i^*)}{\Phi(\kappa_i^*)}\frac{\partial \kappa_i^*}{\partial \boldsymbol{\rho}} \right\},$$

com $\partial\kappa_i^*/\partial\boldsymbol{\rho} = \big(\lambda(y_i - \mathbf{T}_i\mathbf{c}')/2\sigma g_i^{3/2}\big)\big(\partial g_i/\partial\boldsymbol{\rho}\big)$. Então, a função escore sob H_0 dada em (4.3.1) é

$$\left.\frac{\partial l(\boldsymbol{\xi}_1)}{\partial \boldsymbol{\rho}}\right|_{\boldsymbol{\xi}_1=\widehat{\boldsymbol{\xi}}_1^0} = \big\{\mathbf{G}'\boldsymbol{\tau}\big\}_{\boldsymbol{\xi}_1=\widehat{\boldsymbol{\xi}}_1^0},$$

em que $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N)'$, $\tau_i = -\big(1/2\big) + \big((y_i - \mathbf{T}_i\mathbf{c}')^2/(2\sigma^2)\big) - \big(\kappa_i/2\big)\big(\phi(\kappa_i)/\Phi(\kappa_i)\big)$, e $\widehat{\boldsymbol{\xi}}_1^0 = (\boldsymbol{\rho}_0', \widehat{\boldsymbol{\theta}}')'$ denota a estimativa de máxima verossimilhança de $\boldsymbol{\xi}_1$ sob a hipótese nula H_0.

Finalmente, segundo Xei *et al.* (2009), a estatística de teste para H_0 é

$$SC_1 = \big\{\boldsymbol{\tau}'\mathbf{G}\big(I_{\boldsymbol{\rho}\boldsymbol{\rho}} - I_{\boldsymbol{\rho}\boldsymbol{\theta}}I_Y^{-1}(\boldsymbol{\theta})I_{\boldsymbol{\rho}\boldsymbol{\theta}}'\big)\mathbf{G}'\boldsymbol{\tau}\big\}_{\boldsymbol{\xi}_1=\widehat{\boldsymbol{\xi}}_1^0}.$$

Sob H_0, quando $N \to \infty$, a distribuição da estatística SC_1 tende à distribuição χ^2_q, lembrando que q é a dimensão do vetor $\boldsymbol{\rho}$.

4.3.2 Teste de Homogeneidade para o Parâmetro de Assimetria

O parâmetro de assimetria λ é muito importante, pois define a direção da assimetria da distribuição, além de levar a distribuição à normalidade quando $\lambda = 0$. Este teste verifica se o parâmetro λ é constante, ou similar à seção anterior, uma função das covariáveis $\lambda_i = \lambda h(\mathbf{V}_i, \boldsymbol{\gamma}) = \lambda h_i$.

Assume-se que exista um único valor $\boldsymbol{\gamma}_0$ de $\boldsymbol{\gamma}$ tal que $h(\mathbf{V}_i, \boldsymbol{\gamma}_0) = 1$ para todo i. Observe que quando $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$, temos $\lambda_i = \lambda$.

Dessa forma, o teste para a homogeneidade do parâmetro λ é equivalente a testar as hipóteses

$$H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0; \quad H_1 : \boldsymbol{\gamma} \neq \boldsymbol{\gamma}_0. \tag{4.0}$$

De (3.5.2), a derivada de primeira ordem de $l(\boldsymbol{\xi}_2)$ em relação a $\boldsymbol{\gamma}$ é

$$\frac{\partial l(\boldsymbol{\xi}_2)}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{N} \left\{ \frac{\phi(\kappa_i^{**})}{\Phi(\kappa_i^{**})} \frac{\lambda(y_i - \mathbf{T}_i \mathbf{c}')}{\sigma} \frac{\partial h_i}{\partial \boldsymbol{\gamma}} \right\}.$$

A função escore sob H_0 em (4.3.2) é

$$\left. \frac{\partial l(\boldsymbol{\xi}_2)}{\partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\xi}_2 = \widehat{\boldsymbol{\xi}}_2^0} = \left\{ \mathbf{H}' \boldsymbol{\omega} \right\}_{\boldsymbol{\xi}_2 = \widehat{\boldsymbol{\xi}}_2^0},$$

em que $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)'$, $\omega_i = \kappa_i \big(\phi(\kappa_i)/\Phi(\kappa_i) \big)$, e $\widehat{\boldsymbol{\xi}}_2^0 = (\boldsymbol{\gamma}_0', \widehat{\boldsymbol{\theta}}')'$ denota a estimativa de máxima verossimilhança de $\boldsymbol{\xi}_2$ sob a hipótese nula H_0.

Finalmente, segundo Xei *et al.* (2009), a estatística de teste para H_0 é

$$SC_2 = \left\{ \boldsymbol{\omega}' \mathbf{H} \big(I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - I_{\boldsymbol{\gamma}\boldsymbol{\theta}} I_Y^{-1}(\boldsymbol{\theta}) I_{\boldsymbol{\gamma}\boldsymbol{\theta}}' \big) \mathbf{H}' \boldsymbol{\omega} \right\}_{\boldsymbol{\xi}_2 = \widehat{\boldsymbol{\xi}}_2^0}.$$

Sob H_0, quando $N \to \infty$, a distribuição da estatística SC_2 tende à distribuição $\chi^2_{q^*}$, lembrando que q^* é a dimensão do vetor $\boldsymbol{\gamma}$.

4.4 Simulação

Nesta seção é examinado em um estudo de simulação o poder estimado dos testes propostos. Nas simulações para erros com distribuição normal, utiliza-se o teste apresentado na seção 4.2, teste de Goldfeld-Quandt; para erro com distribuição normal assimétrica, teste de homogeneidade do parâmetro de escala e teste de homogeneidade do parâmetro de assimetria, apresentados na seção 4.3.

4.4.1 Teste de Goldfeld-Quandt

Nesta simulação os dados são gerados seguindo o procedimento abaixo:

1. Gera-se uma amostra de tamanho N de K covariáveis ($K = N$), em que as $K/2$ primeiras tenham distribuição $U(0, 10)$, e as demais covariáveis são funções linear de pelo menos uma das $K/2$ primeiras mais um erro;

2. Gera-se uma amostra de erros $e_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, N$, em que $\sigma_i^2 = \sigma^2 x_{i1}^{\rho}$ com $\sigma^2 = 0,5$; e

3. Determina-se y_i através da equação $y_i = b_0 + b_1 x_{i1} + \cdots + b_K x_{iK} + e_i$, com $b_0 = -5$ e $b_1 = \cdots = b_K = 1,5$.

São usados diferentes graus de heteroscedasticidade ($\rho = 0,5,\ 1,\ 2$) e diferentes tamanhos de amostra ($N = 30,\ 100,\ 200$), com $K = N$. Para cada valor de ρ e N, são geradas 2000 replicações, ou seja, repetimos o procedimento acima citado 2000 vezes. Então, consideramos o poder do teste estimado como sendo a proporção de vezes em que a hipótese H_0 é rejeitada a um nível de significância de 10%. Entretanto, a realização do teste não é efetuada utilizando o ponto crítico teórico, e sim o ponto crítico ajustado proposto por Zhang and Boos (1994).

Em seu trabalho, Zhang e Boos propõem simular N^* replicações do modelo sob H_0, e com isso calcular N^* estatísticas de teste. Com esses valores simulados da estatística de teste, tem-se uma distribuição empírica da mesma. A proposta de Zhang e Boos para estimar o poder de um teste é substituir o ponto crítico teórico de α% de significância pelo $(1-\alpha)$-ésimo percentil da amostra de estatísticas de teste simuladas sob H_0. Durante o restante do trabalho é considerado o ponto crítico ajustado.

A Tabela 4.1 lista os poderes dos testes referentes à estatística GQ para $N = 30$, 100 e 200 baseados nas 2000 replicações.

TABELA 4.1: Poder estimado para o teste de Goldfeld-Quandt

N	$\rho = 0,5$	$\rho = 1$	$\rho = 2$
30	0,1410	0,1950	0,2650
100	0,3320	0,6185	0,8900
200	0,4325	0,8180	0,9855

Os resultados obtidos mostram que o poder do teste acompanha o crescimento tanto do tamanho amostral quanto do grau da heteroscedasticidade. Estes resultados mostram ainda que o teste é indicado quando o grau da heteroscedasticidade é alto ou quando é moderado com uma amostra grande.

4.4.2 Teste de Homogeneidade do Parâmetro de Escala

Os dados são gerados da seguinte forma:

1. Gera-se uma amostra de tamanho N de K covariáveis ($K = N$), em que as $(K/2)$-ésimas primeiras tenham distribuição $U(0, 10)$, e as demais covariáveis são funções linear de pelo menos uma das $K/2$ primeiras mais um erro;

2. Gera-se uma amostra de erros $e_i \sim NA(0, \sigma_i^2, \lambda)$, $i = 1, \ldots, N$, em que $\sigma_i^2 = \sigma^2 x_{i1}^{\rho}$ com $\sigma^2 = 0,5$ e $\lambda = 1$; e

3. Determina-se y_i através da equação $y_i = b_0 + b_1 x_{i1} + \cdots + b_K x_{iK} + e_i$, com $b_0 = -5$ e $b_1 = \cdots = b_K = 1,5$.

Como na seção anterior adotamos $\rho = 0,5$, 1, 2 e calcula-se o poder do teste estimado de forma similar, ou seja, como a proporção de vezes em que a hipótese $H_0 : \rho = 0$ é rejeitada para um nível de significância de 10%.

A estrutura usada para modelar σ_i^2 é chamada de multiplicativa. Outra estrutura muito utilizada para modelar esses parâmetros é a estrutura exponencial, dada por $\sigma_i^2 = \sigma^2 \exp(\rho x_{i1})$. A Tabela 4.2 lista os poderes dos testes referentes à estatística SC_1 para $N = 30$, 100 e 200 baseados nas 2000 replicações. Esta tabela apresenta o poder do teste considerando a estrutura multiplicativa ($\sigma^2 x_{i1}^{\rho}$) e a exponencial ($\sigma^2 e^{\rho x_{i1}}$).

TABELA 4.2: Poder estimado de SC_1 para estrutura multiplicativa ($\sigma^2 x_{i1}^{\rho}$) e exponencial ($\sigma^2 e^{\rho x_{i1}}$)

N	$\rho = 0,5$		$\rho = 1$		$\rho = 2$	
	$\sigma^2 x_{i1}^{\rho}$	$\sigma^2 e^{\rho x_{i1}}$	$\sigma^2 x_{i1}^{\rho}$	$\sigma^2 e^{\rho x_{i1}}$	$\sigma^2 x_{i1}^{\rho}$	$\sigma^2 e^{\rho x_{i1}}$
30	0,1630	0,1160	0,1850	0,1340	0,1900	0,1275
100	0,4500	0,2480	0,8665	0,7680	0,9270	0,9435
200	0,5650	0,3810	0,9605	0,9355	0,9955	0,9980

A Tabela 4.2 mostra que o poder do teste cresce conjuntamente com o grau da heteroscedasticidade e o tamanho da amostra. Isto sugere que o teste seja mais indicado para grandes amostras e alto grau de heteroscedasticidade.

Assim como observado no trabalho de Xei *et al.* (2009), o poder estimado do teste não é muito sensivel à forma funcional que causa a não heterogeneidade da variância.

4.4.3 Teste de Homogeneidade do Parâmetro de Assimetria

Os dados são gerados de maneira similar à seção anterior, exceto pelos erros que seguem distribuição assimétrica, $e_i \sim NA(0, \sigma^2, \lambda_i)$, $i = 1, \ldots, N$, com $\lambda_i = \lambda x_{i1}^{\gamma}$ sendo que $\sigma^2 = 2$ e $\lambda = 0,5$.

Similar às seções anteriores, tomamos $\gamma = 0,5$, 1, 2 e o poder do teste é estimado como a proporção de vezes em que a hipótese $H_0 : \gamma = 0$ é rejeitada para um nível de significância de 10%. A Tabela 4.3 lista os poderes dos testes referentes à estatística SC_2 para $N = 30$, 100 e 200 baseados nas 2000 replicações. Esta tabela apresenta o poder do teste considerando a estrutura multiplicativa (λx_{i1}^{γ}) e a exponencial ($\lambda e^{\gamma x_{i1}}$).

TABELA 4.3: Poder estimado de SC_2 para estrutura multiplicativa (λx_{i1}^{γ}) e exponencial ($\lambda e^{\gamma x_{i1}}$)

N	$\gamma = 0,5$		$\gamma = 1$		$\gamma = 2$	
	λx_{i1}^{γ}	$\lambda e^{\gamma x_{i1}}$	λx_{i1}^{γ}	$\lambda e^{\gamma x_{i1}}$	λx_{i1}^{γ}	$\lambda e^{\gamma x_{i1}}$
30	0,1040	0,0035	0,0860	0,0050	0,0660	0,0090
100	0,1775	0,1400	0,3510	0,2500	0,4565	0,2695
200	0,2650	0,0575	0,6145	0,2510	0,7725	0,2780

Os resultados indicam que o poder do teste não é muito satisfatório, mesmo para grandes amostras e alto grau de heteroscedasticidade. Além disso, diferente do teste para a homogeneidade de σ^2, o poder deste teste se mostrou sensível à escolha da forma funcional.

Capítulo 5

Aplicação em Dados Reais

Para ilustrar a aplicação do modelo e dos testes apresentados, um modelo é ajustado a um conjunto de dados que diz respeito à composição de um composto usado na fabricação de produtos farmacêuticos. São analisadas 60 amostras deste composto. As variáveis resposta são as concentrações de 5 diferentes solventes, e como covariáveis temos 2646 leituras de propriedades físico-químicas. Ou seja, o conjunto de dados é constituído de 60 observações, com 5 variáveis respostas e 2646 covariáveis. Denotamos as variáveis resposta como $Y_1, \ldots, Y_5$ e as covariáveis como $X_1, \ldots, X_{2646}$. Este conjunto de dados é denomido CPS e pode ser obtido no site http://statmaster.sdu.dk/courses/ST02/data/index.html

Na seção 5.1 é ajustado o modelo normal com heteroscedasticidade na variável resposta Y_2 utilizando testes para justificar o uso desta distribuição e de uma estrutura heteroscedástica para os erros. Na seção 5.2 a variável resposta considerada é Y_1 e o modelo ajustado é o normal assimétrico com heterogeneidade nos parâmetros de escala e assimetria. Novamente testes são usados para justificar o uso desta distribuição e de estruturas heterogêneas para σ^2 e λ. A decisão de utilizar apenas as variáveis resposta Y_1 e Y_2 na análise se deve ao fato de que apenas esta apresentaram erros heteroscedásticos.

5.1 Modelo Normal Heteroscedástico

Para a análise desse tipo de conjunto de dados geralmente é utilizado o método PLS para estimação do modelo de regressão. Considerando apenas a variável Y_2 como

resposta e considerando o algoritmo NIPALS, a validação cruzada apontou que o número de fatores extraídos que minimiza a estatística $PRESS$ foi 4, como pode ser visto na Tabela 5.1. Estes quatro componentes explicam cerca de 89% da variação das covariáveis

TABELA 5.1: Análise do número de fatores a serem extraídos

Número de fatores	$PRESS$	Número de fatores	$PRESS$	Número de fatores	$PRESS$
0	1,1047	5	0,0845	10	0,0875
1	0,1869	6	0,0864	11	0,0874
2	0,1449	7	0,0883	12	0,0874
3	0,0954	8	0,0879	13	0,0874
4	**0,0831**	9	0,087739	14	0,087461

e 99% da variação da variável resposta. Após extrair os fatores, estima-se

$$\widehat{\mathbf{y}}_2 = \mathbf{T}\widehat{\mathbf{c}}',$$

e obtém-se os resíduos $\widehat{\mathbf{e}} = \mathbf{y}_2 - \widehat{\mathbf{y}}_2$. Como citado anteriormente, um dos pressupostos para o uso do método PLS é que o erros sejam homocesdásticos.

Deve-se agora aplicar um teste para verificar a homoscedasticidade dos erros. Para escolher qual teste aplicar, deve-se ter conhecimento da distribuição do erros. E após assumir uma distribuição, aplica-se o teste adequado.

Para decidir entre as distribuições normal e normal assimétrica aplicaremos o teste de assimetria proposto por Yamagata (2003). Este teste de assimetria é robusto à não-normalidade e à heteroscedasticidade, e testa a hipótese H_0 : "Os erros têm um comportamento simétrico". Aplicando o teste, obteve-se um valor-p de $0,8826$, o que indica uma distribuição simétrica para os erros, favorecendo a escolha da distribuição normal. Para confirmar essa escolha, os testes de Kolmogorov-Smirnov e Shapiro-Wilk não rejeitaram a hipótese de normalidade dos erros.

Como assume-se a normalidade dos erros, devemos testar a heteroscedasticidade usando testes como o de White e/ou de Goldfeld-Quandt. Aplicando o teste de White, usando os componentes extraídos como covariáveis, chegamos a um valor-p de $0,8706$, indicando erros homocedásticos.

Entretanto, este teste verifica se os componentes afetam de alguma forma os resíduos do modelo. O próximo passo agora é verificar se uma das covariáveis originais

$X_1, \ldots, X_{2646}$, afeta os resíduos. Fazendo uma análise gráfica, representando os resíduos obtidos contra cada uma das covariáveis, pode-se observar que as covariáveis X_{172} e X_{560} parecem ter algum tipo de influência sobre as variâncias dos erros. Observando a figura 5.1 podemos verificar que à medida que X_{172} e X_{560} crescem, a dispersão dos residuos aumenta.

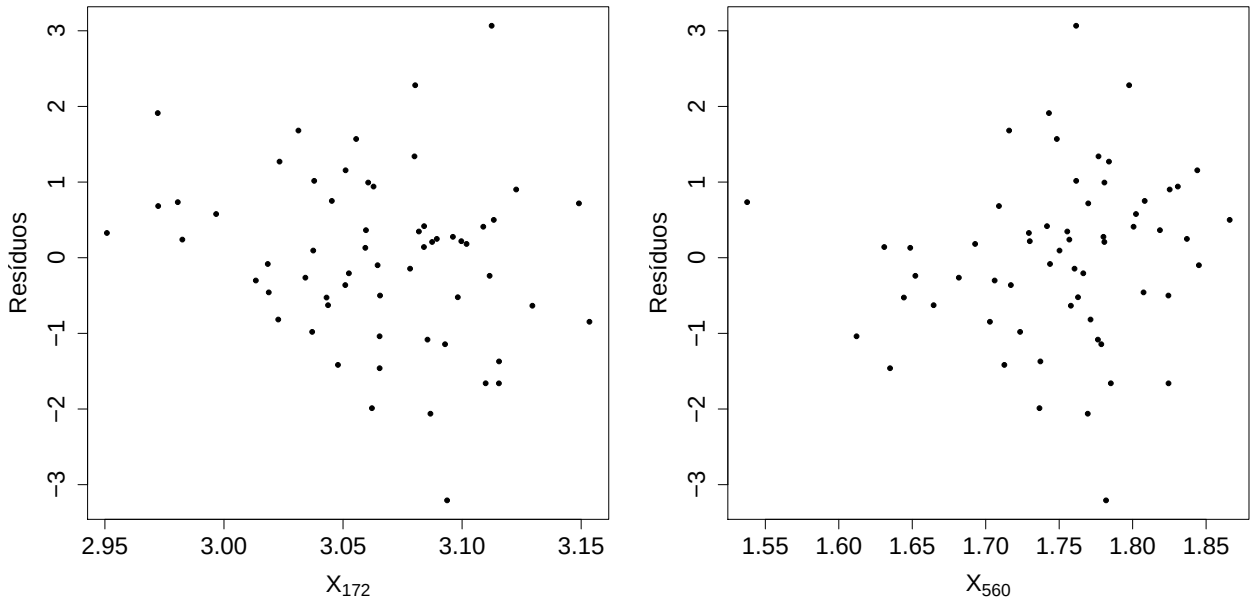

FIGURA 5.1: Gráfico de dispersão entre os resíduos e variáveis explicativas

Os gráficos mostram que as covariáveis X_{172} e X_{560} podem ser as responsáveis pela presença de heteroscedasticidade nos erros. O teste de Goldfeld-Quandt pode ser aplicado para confirmar ou não esta suspeita. Aplicando o teste considerando X_{172} e X_{560}, obteve-se os valores-p de 0.0386 e 0.0525, respectivamente. Dessa forma, há indícios de que ambas as covariáveis causam heteroscedasticidade.

Com a detecção da heteroscedasticidade, deve-se agora propor uma estrutura para a mesma e em seguida estimar os parâmetros dessa estrutura. A variância dos erros é dada por

$$Var(e_i) = \sigma_i^2 = \sigma^2 g_i.$$

Consideraremos duas estruturas, multiplicativa e exponencial, ou seja,

$$\begin{aligned} g_i &= x_{i172}^{\rho_1} x_{i560}^{\rho_2}, \\ g_i &= \exp(\rho_1 x_{i,172} + \rho_2 x_{i,560}). \end{aligned}$$

Considerando a função log-verossimilhança dada por

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2}\log g_i - \frac{(y_{i2} - \mathbf{T}_i\mathbf{c})^2}{2\sigma^2 g_i} \right],$$

em que $\boldsymbol{\psi} = (\boldsymbol{\rho}', \mathbf{c}, \sigma^2)$, as estimativas de máxima verossimilhança obtidas aplicando o algoritmo de Newton-Raphson são dadas na Tabela 5.2.

TABELA 5.2: EMVs dos parâmetros do modelo normal para as estruturas multiplicativa e exponencial

Parâmetro	Estrutura para g_i	
	Multiplicativa	Exponencial
$\widehat{c}_0$	59,9862	59,9893
$\widehat{c}_1$	0,6870	0,6869
$\widehat{c}_2$	0,0375	0,0378
$\widehat{c}_3$	0,2054	0,2056
$\widehat{c}_4$	0,2503	0,2511
$\widehat{\sigma}^2$	1,00E-08	1,00E-08
$\widehat{\rho}_1$	13,4907	4,3012
$\widehat{\rho}_2$	6,2113	3,0854
$l(\widehat{\boldsymbol{\psi}})$	-89,8122	-89,8727

O parâmetro σ^2 apresenta um valor extremamente pequeno. Entretanto, este parâmetro é apenas um dos que compõem a variância. Vale lembrar que a variância é dada por $Var(e_i) = \sigma^2 x_{i172}^{\rho_1} x_{i560}^{\rho_2}$ ou $Var(e_i) = \sigma^2 \exp(\rho_1 x_{i,172} + \rho_2 x_{i,560})$. Pode-se observar que os parâmetros ρ_1 e ρ_2 obtiveram estimativas com valores grandes, o que acaba sendo uma compensação para a estimativa de σ^2. Para melhor ilustrar isso, temos que para estrutura multiplicativa $0,3165 < \widehat{Var}(e_i) < 2,5811$ e para estrutura exponencial $0,3736 < \widehat{Var}(e_i) < 2,4610$.

Para obter as estimativas para as covariaveis originais basta fazer

$$\widehat{\mathbf{b}} = \mathbf{W}\widehat{\mathbf{c}}'_{mvs},$$

em que $\widehat{\mathbf{c}}_{mvs}$ é a estimativa de máxima verossimilhança obtida para $\mathbf{c}$. Na análise desse tipo de dados o mais importante é a predição e não a significância das covariáveis, pois todas

as propriedades físico-químicas usadas aqui são consideradas necessárias para modelagem de compostos analisados.

5.2 Modelo Normal Assimétrico Heteroscedástico

Nesta seção é apresentada a análise do mesmo conjunto de dados, porém desta vez é considerada a variável resposta Y_1. Ajustando a regressão pelo método PLS, a validação cruzada apontou que o número de fatores extraídos que minimiza a estatística $PRESS$ novamente foi 4, como pode ser visto na Tabela 5.3. Estes quatro componentes explicam

TABELA 5.3: Análise do número de fatores a serem extraídos

Número de fatores	$PRESS$	Número de fatores	$PRESS$	Número de fatores	$PRESS$
0	1,1049	5	0,096415	10	0,102158
1	0,1842	6	0,100183	11	0,102095
2	0,1440	7	0,102882	12	0,102047
3	0,1104	8	0,102078	13	0,102103
4	**0,0944**	9	0,10242	14	0,10211

89,02% da variação das covariáveis e 99,68% da variação da variável resposta.

Aplicando o teste de assimetria proposto por Yamagata (2003), obtemos um valor-p de 0,0439 que, a um nível de significância de 5%, nos leva a rejeitar a hipótese de um comportamento simétrico para os erros, favorecendo a escolha da normal assimétrica. Para confirmar essa escolha, a um nível de significância de 10%, os testes de Kolmogorov-Smirnov e Shapiro-Wilk rejeitam a hipótese de normalidade dos erros.

Assumindo a distribuição normal assimétrica para os erros, deve-se testar a homogeneidade dos parâmetros σ^2 e λ usando os testes de escore propostos por Xei *et al.* (2009). Calculando o resíduo como sendo

$$\widehat{e_i} = y_i - \mathbf{T}_i\widehat{\mathbf{c}} - \sqrt{\frac{2}{\pi}}\frac{\widehat{\sigma}\widehat{\lambda}}{1+\widehat{\lambda}^2},$$

faz-se uma análise gráfica, exibindo os resíduos obtidos contra cada uma das covariáveis. Dessa maneira, pode-se observar que as covariáveis X_{172} e X_{561} parecem ter algum tipo de influência sobre a variância dos erros. Observando a Figura 5.2 podemos verificar que à medida que X_{172} e X_{561} crescem, a dispersão dos resíduos também cresce.

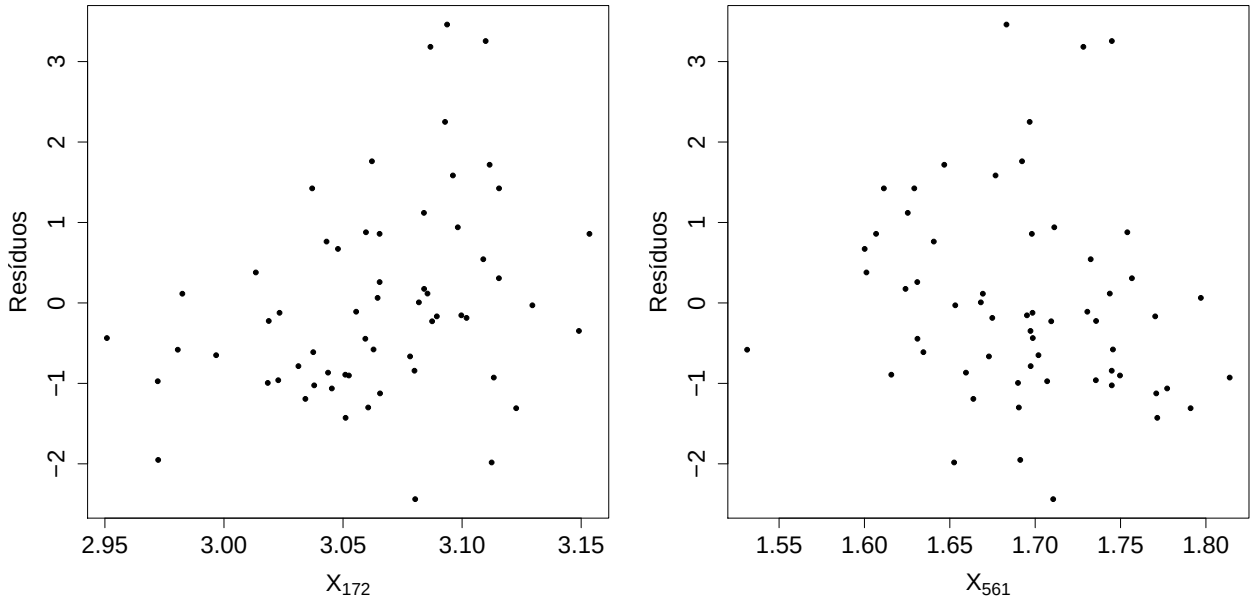

FIGURA 5.2: Gráfico de dispersão entre os resíduos e variáveis explicativas

A Tabela 5.4 mostra os valores-p dos testes de homogeneidade do parâmetro de escala com diferentes estruturas, considerando as variáveis X_{172} e X_{561} como responsáveis. Observando a Tabela 5.4 vê-se que as covariáveis individualmente não afetam a homogeneidade do parâmetro de escala, mas conjuntamente numa estrutura exponencial, estas covariáveis causam a heterogeneidade das variâncias.

TABELA 5.4: valores-p dos testes de homogeneidade do parâmetro de escala com diferentes estruturas

Covariável	Estrutura para g_i	
	Multiplicativa	Exponencial
X_{172}	0,6257	0,6161
X_{561}	0,8143	0,7674
X_{172} e X_{561}	0,3088	0,0874

A Tabela 5.5 apresenta o resultado dos testes referentes ao parâmetro de assimetria. Analisando a Tabela 5.5 pode-se verificar que as covariáveis afetam a homogeneidade do parâmetro de assimetria tanto individual quanto conjuntamente.

Assim, a um nível de significância de 10%, temos indícios que as covariáveis

TABELA 5.5: valores-p dos testes de homogeneidade do parâmetro de assimetria com diferentes estruturas

Covariável	Estrutura para h_i	
	Multiplicativa	Exponencial
X_{172}	$< 0,001$	$< 0,001$
X_{561}	$0,0054$	$< 0,001$
X_{172} e X_{561}	$< 0,001$	$< 0,001$

X_{172} e X_{561} atuam conjuntamente numa estrutura exponencial de heterogeneidade para o parâmetro de escala. Quanto à estrutura de heterogeneidade do parâmetro de assimetria, ambas (multiplicativa e exponencial) podem ser adotadas. Dessa forma, a função log-verossimilhança do modelo é dada por

$$l(\boldsymbol{\xi}) = \sum_{i=1}^{N} \left[\log 2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2}\log g_i - \frac{(y_{i2} - \mathbf{T}_i\mathbf{c})^2}{2\sigma^2 g_i} + \log \Phi(\kappa_i^{***}) \right],$$

em que $\boldsymbol{\xi} = (\boldsymbol{\rho}', \boldsymbol{\gamma}', \mathbf{c}, \sigma^2, \lambda)$, $\kappa_i^{***} = \lambda h_i (y_i - \mathbf{T}_i\mathbf{c})/(\sigma g_i^{1/2})$, $g_i = \exp(\rho_1 x_{i,172} + \rho_2 x_{i,561})$ e

$$\begin{aligned} h_i &= x_{i,172}^{\gamma_1} x_{i,561}^{\gamma_2}, \text{ ou} \\ h_i &= \exp(\gamma_1 x_{i,172} + \gamma_2 x_{i,561}). \end{aligned}$$

A Tabela 5.6 apresenta as estimativas dos parâmetros do modelo para as duas estruturas de h_i (referentes à assimetria).

Novamente, para obter as estimativas para as covariaveis originais basta fazer

$$\widehat{\mathbf{b}} = \mathbf{W}\widehat{\mathbf{c}}'_{mvs}.$$

Mais uma vez salientamos que na análise desse tipo de dados o mais importante é a predição e não a significância das covariáveis.

TABELA 5.6: EMVs dos parâmetros do modelo normal assimétrico heteroscedástico

Parâmetro	Estrutura para h_i	
	Multiplicativa	Exponencial
$\widehat{c}_0$	36,1814	36,3369
$\widehat{c}_1$	0,6842	0,6862
$\widehat{c}_2$	0,0413	0,0396
$\widehat{c}_3$	0,1938	0,1926
$\widehat{c}_4$	0,2855	0,2824
$\widehat{\sigma}^2$	0,0023	0,0021
$\widehat{\lambda}$	0,9037	0,4322
$\widehat{\rho}_1$	4,5789	4,4347
$\widehat{\rho}_2$	4,6618	4,8968
$\widehat{\gamma}_1$	1,6928	0,5807
$\widehat{\gamma}_2$	-0,0152	0,2600
$\mathrm{l}(\widehat{\boldsymbol{\xi}})$	-95,0476	-93,9998

Capítulo 6

Considerações Finais

6.1 Conclusões

Neste trabalho foram desenvolvidos um método de regressão PLS heteroscedástico e adaptações de testes de heteroscedasticidade para o método PLS. Para estimar a estrutura de heteroscedasticidade atribuímos a distribuição normal assimétrica aos erros do modelo. As propriedades dos modelos e testes propostos foram verificadas através de simulação e, para ilustrar a aplicação dos mesmos, foram usados dados reais.

Na seção 3.6 são feitas comparações entre o método proposto e o método PLS usual. As medidas usadas para fazer essas comparações são EQM, vício e variância dos coeficientes da regressão. O modelo em que os coeficientes obtiveram menores valores para essas medidas é considerado melhor. Para a comparação foram simulados vários cenários com diferentes tamanhos de amostra e diferentes graus de heteroscedasticidade. Em cada um destes cenários foram geradas 2000 replicações de amostras de tamanho $N = K$ de um modelo heteroscedástico. Observando apenas a variância e o EQM, pode-se verificar que quando o tamanho da amostra e/ou o grau da heteroscedasticidade aumentam melhor o desempenho do modelo proposto em relação ao PLS. Entretanto, analisando o vício observa-se que o mesmo não se altera muito, mesmo com as variações do tamanho amostral ou do grau da heteroscedasticidade. Na maioria dos cenários o PLS usual é superior ao método proposto quando observado o vício. Os resultados indicam que quando a heteroscedasticidade é forte o uso do método proposto é indicado para qualquer tamanho

de amostra, e quando é moderada, o método proposto deve ser usado para amostras maiores. No caso das predições, o método PLS usual se mostra superior para amostras pequenas e heteroscedasticidade baixa.

Portanto, o estudo mostra que quando o método PLS é necessário e os erros do modelo são heteroscedásticos, o mais indicado é usar o método proposto, pois este usa a informação da estrutura heteroscedástica para construir os estimadores do modelo.

Na seção 4.4 são verificadas as propriedades dos testes propostos. O objetivo é investigar o comportamento do poder destes testes em diferentes cenários, com diferentes tamanhos amostrais de diferentes graus de heteroscedasticidades. Novamente, para cada cenários são geradas 2000 replicações de amostras de tamanho $N = K$ de um modelo heteroscedástico. Para todos os testes, os resultados obtidos mostram que o poder acompanha o crescimento tanto do tamanho amostral quanto do grau de heteroscedasticidade. Estes resultados mostram ainda que o teste é indicado quando o grau de heteroscedasticidade é alto ou quando é moderado com uma amostra grande. Uma ressalva deve ser feita para o teste de homogeneidade para λ no modelo normal assimétrico, que não é muito satisfatório, mesmo para grandes amostras e alto grau de heteroscedasticidade. Além disso, diferente do teste para a homogeneidade de σ^2, o poder deste teste se mostrou sensivel à escolha da forma funcional da estrutura heteroscedástica.

No capítulo 5 todos os modelos e testes propostos foram ilustrados com dados reais. Na seção 5.1 testes indicaram o uso da distribuição normal como a mais adequada e na seção 5.2, usando outros dados, os testes indicaram a distribuição normal assimétrica como a mais indicada para os erros do modelo.

6.2 Propostas de Trabalhos Futuros

Algumas propostas de futuros trabalhos são listadas abaixo:

- Expandir os modelos aqui apresentados para o caso multivariado;
- Utilização de outras distribuições para os erros do modelo; e
- Cálculo de intervalos de confiança para os coeficientes baseados na matriz de informação, e não em técnicas de reamostragem.

Referências Bibliográficas

[1] Azzalini, A. (1985), *A class of distributions which includes the normal ones* Scandinavian Journal Statistics 12: 171-178.

[2] Azzalini, A. (2005), *The skew-normal distribution and related multivariate families (with discussion).* Scandinavian Journal Statistics 32: 159-188.

[3] Barker, M.; Rayens, W. (2003), *Partial least squares for discrimination*, Journal of Chemometrics, 17: 166-173.

[4] Bastien, P.; Vinzi, V. E.; Tenenhaus, M. (2005), *PLS generalised linear regression*, Computational Statistics and Data Analysis, 48: 17-46.

[5] Berwin A.; Turlach (1993), *Bandwidth selection in kernel density estimation: A review*, Discussion Paper 9317, Institut de Statistique, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve.

[6] Ding, B.; Gentleman, R. (2005), *Classification using generalized partial least squares*, Journal of Computational and Graphical Statistics, 14: 280-298.

[7] Geladi, P.; Kowlaski B. (1986), *Partial least square regression: A tutorial*, Analytica Chimica Acta, 35: 117.

[8] Goldfeld, S. M.; Quandt, R. E. (1965), *Some tests for homoscedasticity*, Journal of the American Statistical Association, 60: 539-547.

[9] Höskuldsson, A. (1988), *PLS regression methods.* Journal of Chemometrics, 2: 211-228

[10] MacGregor, J. F.; Kourti, T. (1995), *Statistical process control of multivariate processes*, Control Engineering Practice, 3: 403-414.

[11] Martens, H.; Martens, M. (2000), *Modified jack-knife estimation of parameter uncertainty in bilinear modeling (PLSR)*, Food Qual. Preference, 11: 5-16.

[12] Mendonça, M. (2005), *Classificação de gasolinas comerciais através de métodos estatísticos multivariáveis*, Dissertação (Mestrado)- Escola Politécnica da Universidade de São Paulo.

[13] Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. (2002), *A user-friendly guide to multivariate calibration and classification*, NIR Publications, Chichester UK.

[14] Nguyen, D. V.; Rocke, D. M.; (2002), *Tumor classificationsby partial least squares using microarray gene expression data*, Bioinformatics, 18: 39-50.

[15] Pérez, N. F.; Ferré, J.; Boqué, R.; (2008), *Calculation of the reliability of classification in discriminant partial least-squares binary classification*, Chemometrics and Intelligent Laboratory Systems, 95: 122-128.

[16] Sharma, S. (1996), *Applied multivariate techniques*, Jonh Wiley Sons, New York.

[17] Wang, C. Y.; Chen, C. T.; Chiang, C. P.; Young, S. T.; Chow, S. N.; Chiang, H. K. (1999), *A probability-based multivariate statistical algorithm for autofluoresence spectroscopic identification of oral carcinogenesis*, Photochemistry and Photobiology, 69: 471-477.

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

[19] Wold H. (1966), *Nonlinear estimation by iterative least squares procedures*, Research Paper in Statistics: Festschrift for J. Neyman (ed. F. N. David), New York: Wiley, pp. 411-444.

[20] Wold, S.; Sjostron, M.; Eriksson, L. (2001), *PLS-regression: a basic tool of chemometrics*, Chemometrics and Intelligent Laboratory Systems, 58: 109-130.

[21] Xei, F. C.; Wei, B. C.; Lin J. G. (2009), *Homogeneity diagnostics for skew-nomal nonlinear regression models*, Statistics and Probability Letters, 79: 821-827.

[22] Yamagata, T. (2003), *A nonnormality and heteroskedasticity robust test for skewness in regression models*, The School of Economics Discussion Paper Series, 0328. URL: http://migre.me/iQ4H

[23] Zhang, J.; Boos, D. D. (1994), *Adjusted power estimates in Monte Carlo experiments*, Communication in Statistics Simulation and Computation, 23: 165-173.

Apêndice A

Programa em SAS

A.1 Programa para a simulação na seção 3.6

Segue abaixo o código em SAS para a geração e estimação do modelo normal com erros heteroscedásticos. Para obter os diferentes graus de heteroscedasticidade e tamanhos de amostra, basta substituir os valores de ρ e N.

```
/* ##### Obtendo as covariáveis ##### */

proc iml;
n=30; /* tamanho amostral */
k=30;
semente= 7777;
x1= J(n,k/2,0);
x2= J(n,k/2,0);
do j=1 to k/2;
  do i=1 to n;
  x1[i,j]= ranuni(semente)*10;
  end;
end;
do j=1 to k/2;
  do i=1 to n;
  x2[i,j]= x1[i,j]+(rannor(semente)*0.5);
  end;
end;
```

```
x= x1||x2;
cname= {"x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9" "x10" "x11"
        "x12" "x13" "x14" "x15" "x16" "x17" "x18" "x19" "x20"
        "x21" "x22" "x23" "x24" "x25" "x26" "x27" "x28" "x29" "x30"};
create dados1 from x [ colname=cname ];
append from x;
quit;

/* ##### Gerando os erros e as respostas ##### */

proc iml;
use dados1;
read all var _num_ into A;
n= nrow(A);
k= ncol(A);
semente= 2385361;
beta= J(k+1,1,1.5);
beta[1,1]= -5;
rho= 0.5;
sigma2= 0.5;
sigma= sqrt(sigma2);
e= J(n,1,0);
do i=1 to n;
e[i]= rannor(semente)*(((sigma2)*(A[i,1]**rho))**0.5); /* estrutura heteroscedástica */
end;
um= J(n,1,1);
x= um||A;
y= (x*beta)+e;
C= y||A;
cname= {"y" "x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9" "x10" "x11"
        "x12" "x13" "x14" "x15" "x16" "x17" "x18" "x19" "x20"
        "x21" "x22" "x23" "x24" "x25" "x26" "x27" "x28" "x29" "x30"};
create dados from C [ colname=cname ];
append from C;
quit;

/* ##### Regressão PLS (extração dos componentes) ##### */

proc pls data=dados cv=split;
```

```
model y= x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20
        x21 x22 x23 x24 x25 x26 x27 x28 x29 x30 /solution;
run;
proc pls data=dados nfac=15 outmodel=sss noprint;
model y= x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20
        x21 x22 x23 x24 x25 x26 x27 x28 x29 x30;
output out=pls xscore=t yscore=u stdy=sy yresidual=res ;
run;

/* ##### Maximização da Verossimilhança ##### */

proc nlp data=pls cov=2 VARDEF=N MAXIT=10000 MAXFUNC=10000 OUTEST=par_mv_ori;
max l;
parms c0= 200,
      c1= 20,
  c2= 8,
  c3= 5,
  c4= 3,
  c5= 2,
  c6= 1,
  c7= 0.5,
  c8= 0.5,
  c9= 0.5,
  c10= 0.5,
  c11= 0.5,
  c12= 0.5,
  c13= 0.5,
  c14= 0.5,
  c15= 0.5,
  rho= 1,
  sigma2= 1;
bounds 0 < sigma2;
l= (-1/2)*log(2*3.1415926535897932384626)-(1/2)*log(sigma2)-(rho/2)*log(x1)-
   ((y-c0-c1*t1-c2*t2-c3*t3-c4*t4-c5*t5-c6*t6-c7*t7-c8*t8-c9*t9-c10*t10-
   c11*t11-c12*t12-c13*t13-c14*t14-c15*t15)**2)/
   (2*sigma2*(x1**rho));
run;
quit;
```

A.2 Programa para a simulação na seção 4.4

Segue abaixo o código em SAS para a geração e estimação do modelo normal assimétrico com erros heteroscedásticos. Para obter os diferentes graus de heteroscedasticidade e tamanhos de amostra, basta substituir os valores de ρ, γ e N.

```
/* ##### Obtendo as covariáveis ##### */

proc iml;
n=30; /* tamanho amostral */
k=30;
semente= 7777;
x1= J(n,k/2,0);
x2= J(n,k/2,0);
do j=1 to k/2;
  do i=1 to n;
  x1[i,j]= ranuni(semente)*10;
  end;
end;
do j=1 to k/2;
  do i=1 to n;
  x2[i,j]= x1[i,j]+(rannor(semente)*0.5);
  end;
end;
x= x1||x2;
cname= {"x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9" "x10" "x11"
        "x12" "x13" "x14" "x15" "x16" "x17" "x18" "x19" "x20"
        "x21" "x22" "x23" "x24" "x25" "x26" "x27" "x28" "x29" "x30"};
create dados1 from x [ colname=cname ];
append from x;
quit;

/* ##### Gerando os erros e as respostas ##### */

proc iml;
use dados1;
read all var _num_ into A;
n= nrow(A);
k= ncol(A);
```

```
semente= 2385361;
beta= J(k+1,1,1.5);
beta[1,1]= -5;
rho= 0;
sigma2= 0.5;
sigma= sqrt(sigma2);
e= J(n,1,0);
y= J(n,1,0);
lambda= 1;
alfa= lambda/sqrt(1+(lambda**2));
do i=1 to n;
   u0= rannor(semente);
   v = rannor(semente);
   u1= (alfa*u0)+v*(sqrt(1-(alfa**2)));
   if u0>0 then e[i]=u1;
   else e[i]=-u1;
   e[i]= e[i]*(((sigma**2)*(A[i,1]**rho))**0.5);
end;
um= J(n,1,1);
x= um||A;
y= (x*beta)+e;
cname= {"y" "x1" "x2" "x3" "x4" "x5" "x6" "x7" "x8" "x9" "x10" "x11"
        "x12" "x13" "x14" "x15" "x16" "x17" "x18" "x19" "x20"
        "x21" "x22" "x23" "x24" "x25" "x26" "x27" "x28" "x29" "x30"};
create dados from C [ colname=cname ];
append from C;
quit;

/* ##### Regressão PLS (extração dos componentes) ##### */

proc pls data=dados cv=split;
model y= x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20
        x21 x22 x23 x24 x25 x26 x27 x28 x29 x30;
run;
proc pls data=dados nfac=15 outmodel=sss noprint;
model y= x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16 x17 x18 x19 x20
        x21 x22 x23 x24 x25 x26 x27 x28 x29 x30;
output out=pls xscore=t yresidual=res stdy=sy;
run;
```

```
/* ##### Maximização da Verossimilhança ##### */

proc nlp data=pls cov=2 VARDEF=N MAXIT=100000 MAXFUNC=100000 OUTEST=par_mv_pad noprint;
max l;
parms c0= 1,
      c1= 1,
    c2= 1,
     c3= 1,
    c4= 1,
    c5= 1,
    c6= 1,
    c7= 1,
    c8= 1,
    c9= 1,
    c10= 1,
    c11= 1,
    c12= 1,
    c13= 1,
    c14= 1,
    c15= 1,
    lambda= 3,
    sigma2= 3;
bounds 0 < sigma2;
l= log(2)-(1/2)*log(2*3.141592653589793238 46)-(1/2)*log(sigma2)-
(((y-c0-c1*t1-c2*t2-c3*t3-c4*t4-c5*t5-c6*t6-c7*t7-c8*t8-c9*t9-
c10*t10-c11*t11-c12*t12-c13*t13-c14*t14-c15*t15)**2)/(2*sigma2))+
log(CDF('NORMAL',lambda*(y-c0-c1*t1-c2*t2-c3*t3-c4*t4-c5*t5-c6*t6-
c7*t7-c8*t8-c9*t9-c10*t10-c11*t11-c12*t12-c13*t13-c14*t14-c15*t15)/(sqrt(sigma2)),0,1));
run;
quit;
```