

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE ESTATÍSTICA

Modelos Lineares Generalizados para Séries Temporais com Memória Longa

Cristiano Amâncio Vieira Borges

Orientador: Prof. Dr. Mauricio Enrique Zevallos Herencia

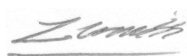
Dissertação apresentada junto ao Departamento de Estatística do Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas, para obtenção do Título de MESTRE em Estatística.

Campinas
2010

MODELOS LINEARES GENERALIZADOS PARA SÉRIES TEMPORAIS COM
MEMÓRIA LONGA

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Cristiano Amâncio Vieira Borges e aprovada pela comissão julgadora.

Campinas, 29 de janeiro de 2010.



Prof. Dr. Mauricio Enrique Zevallos Herencia
Orientador

Banca Examinadora:

1. Prof. Dr. Mauricio Enrique Zevallos Herencia (IMECC - UNICAMP)
2. Prof. Dr. Luiz Koodi Hotta (IMECC - UNICAMP)
3. Prof. Dr. Ricardo Sandes Ehlers (ICMC - USP)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de MESTRE em ESTATÍSTICA.

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO IMECC DA UNICAMP**

Bibliotecária: Crislene Queiroz Custódio – CRB8 / 7966

Borges, Cristiano Amâncio Vieira

B644m Modelos lineares generalizados para séries temporais com memória longa / Cristiano Amâncio Vieira Borges -- Campinas, [S.P. : s.n.], 2010.

Orientador : Mauricio Enrique Zevallos Herencia

Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Séries temporais não-gaussianas. 2. Longa-memória. 3. Modelos lineares generalizados. 4. Verossimilhança parcial (Estatística). 5. Previsão. I. Zevallos Herencia, Mauricio Enrique. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Generalized linear models for long memory time series.

Palavras-chave em inglês (Keywords): 1. Non-gaussian time series. 2. Long memory. 3. Generalized linear models. 4. Partial likelihood (Statistics). 5. Forecasting.

Área de concentração: Séries temporais

Titulação: Mestre em Estatística

Banca examinadora: Prof. Dr. Mauricio Enrique Zevallos Herencia (IMECC-Unicamp)
Prof. Dr. Luiz Koodi Hotta (IMECC-Unicamp)
Prof. Dr. Ricardo Sandes Ehlers (ICMC-USP)

Data da defesa: 29/01/2010

Programa de Pós-Graduação: Mestrado em Estatística

Dissertação de Mestrado defendida em 29 de janeiro de 2010 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.



Prof(a). Dr(a). MAURICIO ENRIQUE ZEVALLOS HERENCIA



Prof(a). Dr(a). LUIZ KOODI HOTTA



Prof(a). Dr(a). RICARDO SANDES EHLERS

*A meu pais Ana Maria e João Bosco,
o “Bosquinho” (in memoriam), e minhas
irmãs Mayra e Thais.*

Agradecimentos

Agradeço primeiramente a Deus, pela oportunidade do estudo, da aprendizagem, e pelo suporte constante em todas as esferas da minha vida.

À minha mãe, Ana Maria, pelo amor incondicional, apoio, incentivo e inspiração transmitidos durante todo o percurso deste trabalho.

Às minhas irmãs, Thais e Mayra, pelo carinho e apoio de sempre. À Mayra, especialmente pelas sugestões e conselhos providos frente às decisões tomadas no processo. E à Tatá, em especial pelos momentos ímpares de afeto proporcionados nos meus retornos à nossa casa em Araxá, Minas Gerais.

A Simoni Martim, pelo carinho, companheirismo, ajuda e apoio sempre presentes, além das alegrias compartilhadas no decorrer da caminhada.

Ao Prof. Mauricio Zevallos, pela amizade, orientação, dedicação, paciência e atenção dispensadas em todas as etapas do desenvolvimento deste projeto, como também pelo exemplo pessoal de conduta profissional.

Aos queridos amigos, que fizeram com que a jornada tivesse momentos tão prazerosos. Dentre os amigos da Estatística, agradeço especialmente a Rodrigo Basso (“Xester”), Alejandro Monzón, Lúcia Rolim, Márcio Diniz, Omar Muhieddine, Camila Borelli e Rodrigo Tsai. Dentre os amigos de mais longa data, especialmente a: Vinicius Conrad (“Tarzan”), Jonas Alonso, Gustavo Lima, Aender Rodrigues, Larissa Vieira, Júlio Rovigatti, Mateus Melo, Vinicius Campidelli (“Júnior”), Ariane Lopes, Marcio Caparroz e Emmanuelle Oliveira.

A toda a minha família, pela ajuda, apoio, incentivo e compreensão dos momentos em que estive ausente.

A Tânia Trinchinato, da Secretaria de Pós-graduação, pela ajuda em várias ocasiões, especialmente durante o meu inevitável afastamento em novembro de 2008.

Aos Prof.s Luiz Hotta e Ricardo Ehlers, por terem participado da banca examinadora e pelas correções e sugestões propostas.

À CAPES, pelo auxílio financeiro concedido, imprescindível ao desenvolvimento deste projeto.

Resumo

A modelagem de séries temporais não gaussianas é um tema de alta relevância na análise de séries temporais. Utilizando-se de estimação por verossimilhança parcial, Kedem e Fokianos (2002) estenderam sistematicamente a metodologia dos Modelos Lineares Generalizados (MLG) para séries temporais em que tanto a série de interesse quanto as covariáveis são estocasticamente dependentes. Entretanto, a análise estatística de séries com memória longa (ML), seja na resposta ou nas covariáveis, não é discutida em detalhes. O primeiro objetivo desta dissertação é investigar, através de simulações, as propriedades dos estimadores de máxima verossimilhança parcial dos coeficientes do MLG quando utilizado para séries temporais com ML. O segundo objetivo consiste em um estudo sobre a qualidade das previsões obtidas para vários modelos ajustados a dados de séries com ML, utilizando a metodologia proposta por Kedem e Fokianos (2002). Os modelos considerados nesta dissertação são modelos para séries de contagens, séries binárias e séries categóricas ordinais. Finalmente, as metodologias são ilustradas através de aplicações em conjuntos de dados reais de finanças e de poluição do ar.

Abstract

Non-gaussian time series modeling is a high relevance issue of time series analysis. Keddem and Fokianos (2002) have used partial likelihood estimation to extend the Generalized Linear Models (GLM) methodology systematically to time series where the response and covariate data are both stochastically dependent. However, statistical analysis of time series with long memory (LM), whether in the response or in the covariates, is not discussed in detail. The first purpose of this paper is to investigate, via simulations, the properties of the partial maximum likelihood estimators of the GLM coefficients as used for modeling LM time series. As a second purpose, we have assessed the quality of the forecasts obtained from several adjusted models (using the methodology proposed by Keddem and Fokianos (2002)) as applied to data with LM series. The models we have chosen for our work include count series, binary series, and categorical ordinal time series models. Finally, the methodologies are illustrated with applications to financial and air pollution real data.

Sumário

1	Introdução	1
1.1	Breve descrição da dissertação	4
2	Modelos Lineares Generalizados para Séries Temporais	7
2.1	Introdução	7
2.2	Verossimilhança parcial	8
2.3	Terminologia e contextualização para o modelo	9
2.4	Estimação	12
2.4.1	O algoritmo de mínimos quadrados reponderados iterativamente	16
2.5	Teoria assintótica e pressuposições do modelo	17
2.5.1	Resultados assintóticos	19
2.6	Teste de hipóteses	20
2.7	Seleção de modelos	21
2.8	Diagnóstico	22
2.8.1	Análise do desvio	22
2.8.2	Resíduos	23
3	Séries Temporais com Memória Longa	25
3.1	Processos com memória longa	25
3.2	O modelo ARFIMA	26
3.3	Estimação	28
3.4	Previsão	29

4	Regressão Logística para Séries Binárias	31
4.1	O modelo de regressão logística	32
4.2	Estimação	32
4.3	Estudos de simulação	33
4.3.1	Estudo detalhado de uma série simulada	33
4.3.2	Estudo de simulação geral	36
4.4	Modelagem das séries covariáveis	39
4.4.1	Estimação	40
4.4.2	Predição	41
4.5	Avaliação da performance preditiva para uma classificação binária	41
4.6	Aplicação em poluição do ar	44
4.6.1	Análise da série para o limiar 100	47
4.6.2	Análise da série para o limiar 120	57
4.6.3	Conclusões da aplicação	60
5	Modelo de Chances Proporcionais para Séries Categóricas Ordinais	61
5.1	Contextualização e notação	62
5.2	O modelo de chances proporcionais	64
5.2.1	Estimação	67
5.3	Estudos de simulação	70
5.3.1	Estudo detalhado de uma série simulada	70
5.3.2	Estudo de simulação geral	74
5.4	Medidas de performance preditiva para 3 categorias de resposta	76
5.5	Aplicação em poluição do ar	79
5.5.1	Análise da série com limiares 50 e 100	80
5.5.2	Análise da série com limiares 50 e 120	84
5.5.3	Conclusões da aplicação	88
6	Regressão com Séries de Contagens	91
6.1	O modelo de regressão Poisson	92
6.2	Estimação	93
6.3	Estudos de simulação	94
6.3.1	Estudo detalhado de uma série simulada	94

Sumário

6.3.2	Estudo de simulação geral	96
6.4	Aplicação a dados do setor financeiro	100
6.4.1	Ajuste	101
6.4.2	Predição	106
6.4.3	Conclusões da aplicação	109
7	Conclusões e considerações finais	111
	Referências	113

Lista de Figuras

3.1	<i>Comparação entre as autocorrelações de séries simuladas com e sem memória longa: (a), (c) X_t e sua FAC; (b), (d) W_t e sua FAC.</i>	28
4.1	<i>Séries simuladas: (a), (b) e (c) W_t, sua FAC e FACP; (d), (e) e (f) Y_t, sua FAC e FACP; (g) dispersão de $\pi_t(\boldsymbol{\beta})$ para cada nível de Y_{t-1}; (h) dispersão de $\pi_t(\boldsymbol{\beta})$ versus W_t.</i>	34
4.2	<i>Séries $\pi_t(\boldsymbol{\beta}) = P(Y_t = 1 \mathcal{F}_{t-1})$ (a), Y_{t-1} (b) e W_t (com sua linha média) (c).</i>	35
4.3	<i>Seqüências de estimativas obtidas para $\hat{\boldsymbol{\beta}}$, para o caso $N = 200$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.</i>	38
4.4	<i>Seqüências de estimativas obtidas para $\hat{\boldsymbol{\beta}}$, para o caso $N = 1000$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.</i>	39
4.5	<i>Séries de poluentes: (a) PM10, (b) SO2 e (c) NO2.</i>	45
4.6	<i>FAC (à esq.) e FACP (à dir.) dos poluentes: (a) PM10, (b) SO2 e (c) NO2.</i>	46
4.7	<i>Categorização de PM10. Indicação dos níveis de corte e do período de predição (a), $Y_{[100]}$ e sua FAC (b) e $Y_{[120]}$ e sua FAC (c).</i>	47
4.8	<i>SO2 (a) e $SO2^{(\lambda)}$ (b).</i>	49

4.9	<i>Decomposição nas componentes regular e irregular. (a) $SO_2^{(\lambda)}$ e a curva de ajuste à parte regular. (b) Destaque à componente regular. (c) Componente irregular. (d) FAC da parte irregular. (e) FACP da parte irregular. (f) Histograma da parte irregular.</i>	50
4.10	<i>Resíduos do ajuste do ARFIMA(0, d, 1) à componente irregular: (a) Série, (b) FAC, (c) Gráfico quantil-a-quantil com os da $\mathcal{N}(0, 1)$ e (d) Níveis de significância para o teste de Box-Ljung.</i>	51
4.11	<i>Séries observada e predita para SO_2.</i>	51
4.12	<i>Séries observada e predita para NO_2.</i>	52
4.13	<i>$Y_{[100]}$ observada (linha contínua) graficada junto a $\hat{\mu}_t$ (a) e $\hat{Y}_{[100]}$ (b) (linhas tracejadas).</i>	53
4.14	<i>$Y_{[120]}$ observada (linha contínua) graficada junto a $\hat{\mu}_t$ (a) e $\hat{Y}_{[120]}$ (b) (linhas tracejadas).</i>	58
5.1	<i>Séries simuladas: (a), (b) e (c) $\{W_t\}$, seu histograma e sua FAC; (d) $\{X_t\}$. 71</i>	
5.2	<i>Série $\{Y_t\}$ gerada (a), sua FAC (b), e diagramas de dispersão entre $\{Y_t\}$ e as covariáveis $\{X_t\}$ (c) e $\{W_t\}$ (d).</i>	72
5.3	<i>Séries $\{Y_t\}$ (a), $\{X_t\}$ (b) e $\{W_t\}$ (c) com suas linhas médias.</i>	73
5.4	<i>Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 200$ e $d = 0, 49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0, 1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0, 1)$: (a), (b) e (c) $\hat{\theta}_1$, (d), (e) e (f) $\hat{\theta}_2$, (g), (h) e (i) $\hat{\gamma}_1$, (j), (k) e (l) $\hat{\gamma}_2$.</i>	75
5.5	<i>Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 1000$ e $d = 0, 49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0, 1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0, 1)$: (a), (b) e (c) $\hat{\theta}_1$, (d), (e) e (f) $\hat{\theta}_2$, (g), (h) e (i) $\hat{\gamma}_1$, (j), (k) e (l) $\hat{\gamma}_2$.</i>	76
5.6	<i>Níveis de corte para PM_{10} (a), $Y_{[100]}^{Ord}$ e sua FAC (b) e $Y_{[120]}^{Ord}$ e sua FAC (c). 80</i>	
5.7	<i>$Y_{[100]}^{Ord}$ observada (linha contínua) e predita (linha tracejada).</i>	82
5.8	<i>$Y_{[120]}^{Ord}$ observada (linha contínua) e predita (linha tracejada).</i>	85
6.1	<i>Séries simuladas: (a), (b) e (c) X_t e suas funções de autocorrelação e de autocorrelação parcial; (d), (e) e (f) W_t, sua FAC e FACP; (g), (h) e (i) Y_t, sua FAC e FACP.</i>	95

6.2	<i>Resíduos do ajuste aos dados simulados: (a) Resíduos tipo componente da função desvio, (b) FAC, (c) níveis descritivos para o teste de Box-Ljung, (d) histograma de freq. rel. sobreposto pela curva da $\mathcal{N}(0,1)$, (e) gráfico quantil-a-quantil com a Normal, (f) resíduos versus valores ajustados. . . .</i>	96
6.3	<i>Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 200$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.</i>	98
6.4	<i>Estimativas das densidades conjuntas, par a par, dos valores obtidos para $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, para o caso $N = 200$ e $d = 0,49$, e correspondentes gráficos de contorno.</i>	99
6.5	<i>Séries financeiras: (a), (c) e (e) Volume de transações (Y_t) e suas funções de autocorrelação e de autocorrelação parcial, (b), (d) e (f) Volatilidade realizada (V_t) e suas FAC e FACP.</i>	101
6.6	<i>(a) Volume de transações observado na terça-feira, evidenciando os períodos do dia demarcados pela variável período. (b) Gráficos de dispersão e retas de ajuste de regressão entre Y_t e V_t, separadamente para cada período do dia.</i>	103
6.7	<i>Volatilidade e suas transformações: série, histograma de freqüências e gráfico quantil-a-quantil com os da $\mathcal{N}(0,1)$. Volatilidade (V_t) - (a), (b) e (c); log-volatilidade (Z_t) - (d), (e) e (f); volatilidade transformada pela transformação Box-Cox ($V_t^{(\lambda)}$) - (g), (h) e (i).</i>	104
6.8	<i>Volume de transações observado e ajustado.</i>	105
6.9	<i>Resíduos do ajuste: (a) histograma dos resíduos tipo componentes da função desvio (\hat{d}_t), (b) Q-Q plot, (c) dispersão entre \hat{d}_t e os valores ajustados \hat{y}_t, (d) \hat{d}_t versus t, (e) \hat{d}_t versus Y_{t-1}, (f) \hat{d}_t versus Z_t.</i>	106
6.10	<i>Resíduos do ajuste - avaliação da existência de correlação serial. Resíduos do desvio \hat{d}_t (a), suas funções de autocorrelação e de autocorrelação parcial (c) e (e), e níveis de significância para o teste de Ljung-Box (g); Resíduos de resposta \hat{e}_t (b), suas FAC e FACP (d) e (f), e níveis de significância para o teste de Ljung-Box (h).</i>	107
6.11	<i>Log-volatilidade observada (linha contínua) e predita (tracejada).</i>	108

6.12 *Predição de Y_t : (a) volume predito e observado, (b) intervalo de predição aproximado de 95% de confiança.* 109

6.13 *Predição de Y_t utilizando os valores reais de Z_t .* 110

Lista de Tabelas

4.1	<i>Estimação do modelo a partir dos dados simulados.</i>	36
4.2	<i>Resultado das $k = 1000$ simulações (valor real: $\beta = (-4, 3, 2, 7, 0, 5)'$).</i>	37
4.3	<i>Matriz de confusão.</i>	42
4.4	<i>Modelo logístico estimado para $Y_{[100]}$.</i>	48
4.5	<i>Modelo de regressão ajustado à parte regular de $SO2^{(\lambda)}$.</i>	49
4.6	<i>Matriz de erros para a predição de $Y_{[100]}$.</i>	52
4.7	<i>Performances preditivas para a modelagem de $Y_{[100]}$.</i>	54
4.8	<i>Comparação das performances preditivas de todos os modelos para $Y_{[100]}$.</i>	56
4.9	<i>Modelo logístico estimado para $Y_{[120]}$.</i>	57
4.10	<i>Matriz de erros para a predição de $Y_{[120]}$.</i>	59
4.11	<i>Performances preditivas para a modelagem de $Y_{[120]}$.</i>	59
4.12	<i>Comparação das performances preditivas de todos os modelos para $Y_{[120]}$.</i>	60
5.1	<i>Estimativas dos parâmetros do modelo obtidas para o ajuste à série simulada.</i>	73
5.2	<i>Estimativas dos parâmetros para o ajuste do modelo alternativo, com efeitos separados para cada logito.</i>	74
5.3	<i>Resultado das simulações (valor real: $\beta = (-3, 7, -2, 2, 0, 15, 0, 15)'$).</i>	74
5.4	<i>Matriz de confusão para classificação em 3 categorias.</i>	77
5.5	<i>Modelo estimado para $Y_{[100]}^{Ord}$.</i>	81
5.6	<i>Matriz de erros para a predição de $Y_{[100]}^{Ord}$.</i>	83
5.7	<i>Comparação das performances preditivas de todos os modelos para $Y_{[100]}^{Ord}$.</i>	83
5.8	<i>Modelo estimado para $Y_{[120]}^{Ord}$.</i>	85

5.9 *Matriz de erros para a predição de $Y_{[120]}^{Ord}$* 86

5.10 *Comparação das performances preditivas de todos os modelos para $Y_{[120]}^{Ord}$* . . . 86

6.1 *Resultados do ajuste aos dados simulados*. 95

6.2 *Resultados das simulações (valor real: $\beta = (1, 95, 0, 025, 0, 013)'$)*. 97

6.3 *Variável período, e a média e o desvio-padrão do volume, além das correlações entre o volume e a volatilidade, por período do dia*. 102

6.4 *Modelo estimado para o volume de transações, Y_t* 104

Introdução

A modelagem de séries temporais não gaussianas é um tema de alta relevância na análise de séries temporais. A metodologia amplamente utilizada proposta por Box e Jenkins (1976) foi desenvolvida para séries geradas de acordo com um processo ARIMA com inovações gaussianas, e apresenta limitações na modelagem de séries de contagens e séries categóricas nominais ou ordinais. Para modelar esses tipos de dados, diversas metodologias têm sido propostas. Assim, no contexto de modelos lineares generalizados para séries temporais com curta memória podemos citar, entre outros, os modelos propostos por West, Harrison e Migon (1985) e Benjamin *et al.* (2003).

Os modelos dinâmicos, propostos por West, Harrison e Migon (1985), bem como suas generalizações, permitem grande flexibilidade na modelagem de dados gerados por densidades pertencentes à família exponencial. A estimação é realizada de forma seqüencial, sendo freqüentemente necessário o uso de técnicas MCMC. Os modelos autorregressivos de médias móveis generalizados (GARMA), por sua vez, foram introduzidos por Benjamin *et al.* (2003) e são uma extensão do modelo clássico ARMA para séries temporais não gaussianas com distribuição condicional ao passado pertencente à família exponencial.

Uma outra abordagem bastante interessante para a estimação de séries temporais não gaussianas foi proposta por Kedem e Fokianos (2002), utilizando-se do conceito de verossimilhança parcial. Os autores estenderam a metodologia dos Modelos Lineares Generalizados (MLG) de Nelder e Wedderburn (1972) para séries temporais onde tanto a série de interesse (a resposta) quanto as séries covariáveis são aleatórias e estocasticamente dependentes. Baseando-se em propriedades de processos martingais, demonstraram que

o estimador de máxima verossimilhança parcial (MVP) é consistente e assintoticamente normal, permitindo que a inferência clássica de MV utilizada nos MLGs com observações independentes possa ser realizada para os estimadores de MVP. Além disso, esta classe de modelos, resultante da extensão dos MLGs para séries temporais, inclui a classe dos modelos GARMA como um caso particular, e tem ainda a vantagem, sobre o GARMA e o modelo linear generalizado dinâmico, da possibilidade de incorporação de termos de interação no preditor linear. Outra desvantagem do modelo GARMA é o fato de Benjamin *et al.* (2003) não fazerem discussão acerca de previsão para o mesmo.

O trabalho de Kedem e Fokianos (2002) vem tendo grande repercussão no cenário de pesquisa acadêmica, principalmente nos últimos anos. É possível contabilizar pelo menos 50 trabalhos citando sua metodologia, sendo a maioria bastante recente, de 2008 ou 2009. Mais da metade dos trabalhos abrange teses de doutoramento e artigos com foco teórico, relacionado ao desenvolvimento de modelos seguindo as idéias propostas por Kedem e Fokianos (2002). Assim, por exemplo, Hung *et al.* (2008) estenderam o modelo logístico para séries temporais binárias de Kedem e Fokianos (2002) para possibilitar a incorporação de efeitos aleatórios. Zhen e Basawa (2009) apresentaram alguns modelos para dados binários obtidos a partir do seccionamento de um processo autorregressivo gaussiano, e compararam através de estudos de simulação cinco métodos de estimação. Pelos estudos, concluíram que o método de MVP figurou dentre os melhores em termos de eficiência. Outros trabalhos, desta vez focados na aplicação do modelo, aplicaram diretamente o MLG para séries temporais de Kedem e Fokianos (2002) a dados de diversos contextos. Levine e Moore (2009), por exemplo, aplicaram a metodologia de Kedem e Fokianos (2002) para séries com respostas Poisson para avaliar a associação entre variáveis meteorológicas e a Síndrome da Dilatação-Vólvulo Gástrica (GDV) em uma grande população de cães do Texas.

Contudo, nenhum dos trabalhos revistos utilizando a metodologia de Kedem e Fokianos (2002) ocupa-se de séries temporais que apresentam memória longa (ML). De fato, poucos estudos existem na literatura de séries temporais sobre modelos para séries não gaussianas com ML. Esta carência é devida, em grande parte, à dificuldade inerente à realização de inferência sob a presença de longa memória. Nesse sentido, duas contribuições metodológicas que admitem o tratamento de dados contínuos ou discretos são dadas por Brockwell (2007) e Palma e Zevallos (2010).

Brockwell (2007) apresentou uma família de modelos generalizados para séries temporais com memória longa onde as observações têm uma determinada distribuição condicional, dado um processo ARFIMA latente. Esta distribuição condicional pode ser discreta, contínua ou uma mistura de ambos os tipos, e a família de modelos apresentada inclui diversos modelos existentes, como os próprios modelos ARFIMA (Granger e Joyeux, 1980; Hosking, 1981), modelos de volatilidade estocástica com ML e modelos gaussianos censurados com ML. A forma geral desta classe de modelos é bastante similar à do MLG para séries temporais de Kedem e Fokianos (2002); a diferença recai sobre o processo ARFIMA latente, o que requer estimação por MCMC para os parâmetros.

Palma e Zevallos (2010) propuseram uma classe de modelos para séries com memória longa que assume a variância condicional (à informação passada) como uma função da média condicional. Nesta classe, a resposta pode ser contínua ou discreta, e a estimação é realizada via máxima verossimilhança.

A carência de estudos na área de modelos para séries não gaussianas com memória longa motivou o desenvolvimento do presente trabalho. O objetivo desta dissertação é avaliar se a extensão do MLG para séries temporais da forma como proposta por Kedem e Fokianos (2002) funciona para séries com memória longa. Dois são os objetivos específicos. Primeiro, interessa avaliar por meio de estudos de simulação se as propriedades do estimador de MVP, de consistência e normalidade assintótica, permanecem válidas quando há memória longa na série resposta e/ou nas séries covariáveis. Segundo, avaliar, através de aplicações a dados reais, o poder preditivo do modelo quando as séries covariáveis são modeladas e preditas.

Para responder ao primeiro objetivo, realizou-se estudos de simulação para três casos especiais de MLG: o modelo logístico para séries binárias, o modelo de chances proporcionais para séries categóricas ordinais, e o modelo Poisson para séries de contagens. Estes três modelos foram escolhidos por se tratarem de modelos destinados a séries temporais que são mais raras que as gaussianas e que possuem características peculiares. As simulações foram realizadas de forma a se avaliar dois aspectos. Primeiramente, avaliar como se comportam o viés e a precisão da estimação dos modelos sob a variação do parâmetro de ML e do tamanho das séries. Em segundo lugar, verificar o comportamento assintótico da distribuição do EMVP.

Nas simulações de todos os modelos, foram utilizadas duas séries covariáveis, sendo

uma (X_t) de memória curta, determinística ou defasagem da própria resposta Y_t , e a outra (W_t) de memória longa. Consideraram-se diferentes combinações dos tipos de cada série. Nas simulações para o caso binário, $X_t = Y_{t-1}$ e $\{W_t\}$ é um ARFIMA(1,d,1). No caso categórico ordinal, X_t é um termo senoidal e $\{W_t\} \sim \text{ARFIMA}(1,d,0)$. No caso das contagens, $\{X_t\} \sim \text{AR}(1)$ e $\{W_t\} \sim \text{RF}(d)$.

O segundo objetivo específico da dissertação responde às necessidades práticas de um assunto pouco explorado na literatura e ausente também na proposta de Kedem e Fokianos (2002). Em situações onde a resposta depende de valores contemporâneos das covariáveis, para fazer previsão para a resposta precisamos das previsões das covariáveis. Portanto, faz-se necessário realizar a modelagem e previsão das séries covariáveis. Nas aplicações estudadas nesta dissertação, após as séries covariáveis terem sido modeladas e preditas, alguns modelos competidores serão comparados através de medidas de performance preditiva, para se poder avaliar a eficácia da modelagem das covariáveis sobre a predição da série de interesse.

A aplicação dos modelos a séries reais será realizada em dois contextos. Os modelos para séries binárias e para séries categóricas ordinais serão utilizados no contexto de poluição do ar, onde se deseja prever o nível diário de poluição causada por partículas em suspensão menores que $10 \mu\text{m}$ (PM10) em função de outras séries de poluentes. E o modelo Poisson será aplicado a séries financeiras, onde o objetivo é a modelagem do volume de transações em função da volatilidade.

1.1 Breve descrição da dissertação

A estrutura do trabalho está organizada de acordo com os objetivos acima descritos. No Capítulo 2 apresentaremos a metodologia proposta por Kedem e Fokianos (2002) para a análise de séries temporais através do Modelo Linear Generalizado. Serão apresentados o modelo, a forma de estimação por verossimilhança parcial e alguns métodos de diagnóstico. Apresentaremos ainda, sucintamente, as pressuposições do modelo e os importantes resultados assintóticos sobre o estimador de máxima verossimilhança parcial, resultados estes que estão vinculados ao objetivo de pesquisa desta dissertação. Assim, ficará estabelecida uma base teórica para o desenvolvimento dos capítulos subsequentes.

O Capítulo 3 apresenta brevemente os conceitos básicos de processos com memória

1.1. Breve descrição da dissertação

longa. Estes aspectos estão fortemente vinculados à modelagem das séries utilizadas nas aplicações da dissertação. Começaremos apresentando a definição de memória longa, e em seguida faremos uma introdução ao modelo ARFIMA. Um procedimento de estimação será brevemente descrito. Adicionalmente, apresentaremos a forma simples de aproximação do modelo ARFIMA(1, d , 1) - utilizado nas aplicações dos capítulos - por um AR(p), para a previsão das séries covariáveis.

No Capítulo 4, apresentaremos o modelo logístico com função de ligação logito, sua estimação por verossimilhança parcial e estudos de simulação visando à avaliação do comportamento assintótico do estimador de máxima verossimilhança parcial ($\hat{\beta}$) sob a presença de memória longa nas séries. Especificamente, as simulações serão realizadas de forma a se investigar se, sob a presença de ML, valem os resultados de que

$$\hat{\beta} \rightarrow \beta$$

em probabilidade, e

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow N_p(\mathbf{0}, \mathbf{G}^{-1}(\beta)),$$

em distribuição, à medida que $N \rightarrow \infty$, onde N é o número de observações das séries, β é o vetor de parâmetros do modelo, e $\mathbf{G}(\beta)$ é uma matriz positiva definida que será definida no Capítulo 2. Apresentaremos também a metodologia que foi utilizada em todas as aplicações da dissertação para a modelagem das séries covariáveis. Este tópico importante está relacionado ao segundo objetivo específico do presente trabalho, de avaliar a capacidade preditiva dos modelos quando as covariáveis são modeladas e preditas. Visando à avaliação da performance preditiva do modelo logístico na aplicação aos dados reais, definiremos ainda algumas medidas de performance preditiva usuais no problema de classificação binária. Por fim, encerraremos o capítulo com uma aplicação do modelo a dados de poluição do ar.

O Capítulo 5 apresenta o modelo de chances proporcionais para séries temporais categóricas ordinais, sua estimação via máxima verossimilhança parcial, e estudos de simulação para verificar se as propriedades de consistência e de normalidade assintótica do estimador de MVP valem para a modelagem de séries com ML. Adicionalmente, definiremos algumas medidas de performance preditiva para o caso de classificação em três categorias. A utilização do modelo será então ilustrada através de uma aplicação aos dados de poluição do ar também analisados no Capítulo 4 com o modelo logístico.

No Capítulo 6, apresentaremos o modelo de regressão Poisson para séries temporais de contagens e sua estimação via MVP. Um estudo de simulação será também realizado para a avaliação das propriedades e do comportamento assintótico do estimador de MVP quando estão sendo modeladas séries de memória longa. A utilização do modelo é então ilustrada através de uma aplicação a dados da área financeira.

Finalmente, o Capítulo 7 traz as conclusões da dissertação e algumas propostas para trabalhos futuros na mesma linha de pesquisa.

Recomenda-se a leitura dos Capítulos 2 e 3, inicialmente, pois neles são dadas as bases e definidos os conceitos utilizados nos capítulos posteriores. Em seguida, os Capítulos 4, 5 e 6 podem ser lidos de forma semi-independente, dado que tratam de casos específicos do MLG para séries temporais. Exceção é feita à seção 4.4 do Capítulo 4, onde é apresentada a forma de modelagem das séries covariáveis adotada em todas as aplicações da dissertação.

Modelos Lineares Generalizados para Séries Temporais

Kedem e Fokianos (2002) estenderam sistematicamente a metodologia dos modelos lineares generalizados para séries temporais, onde tanto a série resposta como as séries covariáveis são aleatórias e estocasticamente dependentes. Na adaptação, considerou-se o método de estimação por máxima verossimilhança parcial. Baseando-se em propriedades de processos martingais e em quatro pressuposições para o MLG para séries temporais, Kedem e Fokianos demonstram que o estimador de máxima verossimilhança parcial é consistente e assintoticamente normal.

O objetivo deste capítulo é apresentar a metodologia proposta por Kedem e Fokianos (2002), de forma a se estabelecer uma base teórica para o desenvolvimento dos capítulos posteriores. Serão apresentados o modelo, a forma de estimação por verossimilhança parcial e alguns métodos de diagnóstico. Apresentaremos ainda, sucintamente, as pressuposições do modelo e os importantes resultados assintóticos sobre o estimador de máxima verossimilhança parcial, que estão vinculados ao objetivo de pesquisa desta dissertação.

2.1 Introdução

Na teoria de Modelos Lineares Generalizados (MLG) introduzida por Nelder e Wedderburn (1972) e posteriormente detalhada em McCullagh e Nelder (1989), a principal suposição feita na modelagem é a de *independência* dos dados. No entanto, o método

de estimação usual pode ser estendido, assumindo-se determinadas pressuposições, para *séries temporais*, onde a variável resposta e também as covariáveis são *aleatórias* e *estocasticamente dependentes*. Tal transposição pode ser feita a partir de três quesitos:

- a idéia de uma seqüência crescente de históricos relativos a um observador;
- o conceito de **verossimilhança parcial** introduzido por Cox (1975) e posteriormente desenvolvido por Wong (1986);
- o conceito de **martingal** com relação a uma seqüência de históricos.

O primeiro tópico é pré-requisito necessário aos dois subseqüentes, pois a definição de ambos os conceitos de verossimilhança parcial e de processo martingal são dadas dentro do contexto de uma seqüência de históricos crescente. Esta idéia ficará clara na definição da função de verossimilhança parcial, a seguir. Quanto à utilização da verossimilhança parcial, que é uma forma de pseudo-verossimilhança, foi a solução encontrada por Kedem e Fokianos para a estimação consistente do MLG quando utilizado para séries temporais. Já o terceiro conceito, do processo martingal, faz-se necessário para que se possa estabelecer resultados assintóticos para os estimadores do modelo. De fato, Kedem e Fokianos (2002) se utilizaram deste conceito para aplicar a versão do Teorema Central do Limite para martingais e provar os resultados de grandes amostras para o estimador de máxima verossimilhança parcial.

2.2 Verossimilhança parcial

Como motivação para a função de verossimilhança parcial, consideremos um *par* de séries temporais conjuntamente distribuídas, (X_t, Y_t) , $t = 1, \dots, N$, onde $\{Y_t\}$ é uma série *resposta* e $\{X_t\}$ é uma *covariável aleatória dependente do tempo*. Podemos expressar a densidade conjunta de todas as observações X, Y , parametrizada por um vetor de parâmetros θ , como

$$f_{\theta}(x_1, y_1, \dots, x_N, y_N) = f_{\theta}(x_1) \left[\prod_{t=2}^N f_{\theta}(x_t | y_{t-1}, x_{t-1}, \dots, y_1, x_1) \right] \\ \times \left[\prod_{t=1}^N f_{\theta}(y_t | x_t, y_{t-1}, x_{t-1}, \dots, y_1, x_1) \right]. \quad (2.1)$$

2.3. Terminologia e contextualização para o modelo

O segundo produtório no termo à direita de (2.1) constitui uma função de verossimilhança parcial, de acordo com Cox (1975) e pode, portanto, ser utilizada para inferência.

A grande vantagem na adoção de uma forma de “condicionamento inteligente” como esta encontra-se no fato de que é possível trabalhar apenas com a verossimilhança parcial, ao invés de toda a densidade conjunta de (2.1), que normalmente é difícil de ser calculada em um grande número de situações. Sob determinadas condições, a perda de informação sobre θ devida à desconsideração da densidade marginal e do primeiro produtório em (2.1) é pequena (Kedem e Fokianos, 2002). A título de exemplo (Cordeiro, 1992), suponhamos que o vetor de parâmetros θ possa ser particionado em duas componentes, $\theta' = (\psi', \lambda')$, onde ψ é o vetor de parâmetros de interesse e λ é um vetor de parâmetros de perturbação. Então, a verossimilhança parcial será bastante útil quando envolver apenas os parâmetros de interesse, ψ , ou pelo menos um número bem menor de componentes do vetor λ , dos parâmetros de perturbação.

Com base nas idéias acima expostas, define-se a função de verossimilhança parcial com relação a uma seqüência hierárquica de históricos:

Definição 1: Sejam \mathcal{F}_t , $t = 0, 1, \dots$ uma seqüência crescente de σ -álgebras, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \dots$, e Y_1, Y_2, \dots uma seqüência de variáveis aleatórias em um mesmo espaço de probabilidade tal que Y_t seja mensurável a \mathcal{F}_t . Denotando a densidade de Y_t , dado \mathcal{F}_{t-1} , por $f_t(y_t; \theta)$, onde $\theta \in \mathbb{R}^p$ é o vetor de parâmetros de interesse, define-se a **função de verossimilhança parcial** (PL) relativa a θ, \mathcal{F}_t , e os dados Y_1, Y_2, \dots, Y_N , pelo produto

$$PL(\theta; y_1, \dots, y_N) = \prod_{t=1}^N f_t(y_t; \theta). \quad (2.2)$$

Os valores de X_t e das demais covariáveis, quando houverem, estarão subentendidos na informação do passado, i.e., estarão relacionados na σ -álgebra \mathcal{F}_t . A verossimilhança parcial leva em consideração apenas o que é conhecido ao observador até o momento da observação atual, de forma que ela permite fazer inferência condicional seqüencial.

2.3 Terminologia e contextualização para o modelo

Algumas entidades matemáticas serão endêmicas ao texto de toda a dissertação. Assim, considere-se:

- $\{Y_t\}$ - série temporal de interesse ou resposta;
- $\mathbf{Z}_{t-1} = (Z_{(t-1)1}, \dots, Z_{(t-1)p})'$ - vetor p -dimensional de variáveis explanatórias, covariáveis no passado imediato, ou defasagens da série resposta, $t = 1, \dots, N$ (\mathbf{Z}_t - processo de covariáveis);
- \mathcal{F}_{t-1} - σ -álgebra gerada por $Y_{t-1}, Y_{t-2}, \dots, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \dots$;
- $\mu_t = E(Y_t | \mathcal{F}_{t-1})$ e $\sigma_t^2 = Var(Y_t | \mathcal{F}_{t-1})$ - esperança e variância condicionais da resposta dado o “passado”.

Para séries temporais, a especificação usual do MLG dada por McCullagh e Nelder (1989) não sofre nenhuma alteração em sua forma; isto é, ele é dado pela especificação dos três elementos:

- **componente aleatória**

$$f(y_t; \theta_t, \phi | \mathcal{F}_{t-1}) = \exp \left\{ \frac{y_t \theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t; \phi) \right\}, \quad t = 1, \dots, N, \quad (2.3)$$

que determina a distribuição a ser adotada para a variável-resposta, a qual deve pertencer à família exponencial de distribuições (aqui expressa em sua forma *canônica*) onde θ_t é o chamado parâmetro *natural* da distribuição e $\alpha_t(\phi) = \phi/\omega_t$, sendo ϕ um parâmetro de *dispersão* e ω_t um parâmetro conhecido (*peso* ou *peso a priori*);

- **componente sistemática**

$$\eta_t = \sum_{j=1}^p \beta_j Z_{(t-1)j} = \mathbf{Z}'_{t-1} \boldsymbol{\beta}, \quad (2.4)$$

que discrimina as quantidades conhecidas do modelo, ou seja, as variáveis explanatórias que serão utilizadas (denomina-se η_t por *preditor linear*);

- **e função de ligação**

$$g(\mu_t) = \eta_t, \quad (2.5)$$

que é a função que estabelece a forma de relacionamento entre o preditor linear e a resposta condicionada ao passado esperada, para $t = 1, \dots, N$.

2.3. Terminologia e contextualização para o modelo

A partir das três componentes acima, portanto, define-se o **Modelo Linear Generalizado para Séries Temporais** por

$$g(\mu_t) = \mathbf{Z}'_{t-1}\boldsymbol{\beta}, \quad t = 1, \dots, N,$$

onde μ_t , \mathbf{Z}_{t-1} e g serão determinadas pelas especificações da componente aleatória, da componente sistemática e da função de ligação, respectivamente. Estas escolhas serão feitas de acordo com o tipo dos dados a serem analisados.

Em relação à componente sistemática, quando certas covariáveis X_t, W_t, \dots , forem conhecidas no tempo $t - 1$, permanecerá a notação \mathbf{Z}_{t-1} para o processo de covariáveis, i.e.,

$$\mathbf{Z}_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \dots, X_t, W_t, \dots, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \dots\}.$$

Em outras palavras, \mathbf{Z}_{t-1} é gerada por “*toda a informação do passado que é conhecida ao observador no momento $t - 1$, com a possível inclusão das informações do presente, X_t, W_t, \dots , quando forem conhecidas*”.

Como exemplos de preditor linear pode-se elicitar os abaixo, com a utilização por exemplo de covariáveis periódicas (para captar padrões cíclicos da resposta) e/ou termos de interações entre covariáveis:

- $\mathbf{Z}'_{t-1}\boldsymbol{\beta} = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_t \cos(\omega_0 t)$,
- $\mathbf{Z}'_{t-1}\boldsymbol{\beta} = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_t^2 + \beta_4 Y_{t-2} X_{t-1}$,
- $\mathbf{Z}'_{t-1}\boldsymbol{\beta} = \mathbf{X}'_t \boldsymbol{\gamma} + \sum_{i=1}^p \phi_i H_i(Y_{t-i}) + \sum_{i=1}^q \theta_i D_i(\mu_{t-i})$,

onde ω_0 é uma frequência pré-especificada, e $H_i(\cdot)$ e $D_i(\cdot)$ são funções conhecidas, para todo i . Este último é um caso especial de (2.4), fazendo

$$\mathbf{Z}_{t-1} = (\mathbf{X}_t, H_1(Y_{t-1}), \dots, H_p(Y_{t-p}), D_1(\mu_{t-1}), \dots, D_q(\mu_{t-q}))'$$

e $\boldsymbol{\beta} = (\boldsymbol{\gamma}', \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, e tem uma forma que inclui os modelos *GARMA*(p, q) (*Generalized Autoregressive Moving Average*) (Benjamin *et al.*, 2003).

Assumindo que a ordem das operações de derivação e de integração pode ser permutada, é fácil demonstrar dois resultados clássicos e importantes sobre a classe de distribuições da família exponencial:

- $\mu_t = E(Y_t | \mathcal{F}_{t-1}) = b'(\theta_t)$ e
- $Var(Y_t | \mathcal{F}_{t-1}) = \alpha_t(\phi)b''(\theta_t)$ ($V(\mu_t) = b''(\theta_t)$ é a chamada *função de variância*).

2.4 Estimação

A seguir descrevem-se as linhas gerais do método de estimação via máxima verossimilhança parcial do MLG para dados temporais, adaptado por Kedem e Fokianos (2002). No que segue, assume-se que $\{Y_t\}$, $t = 1, \dots, N$, é condicionalmente distribuída segundo (2.3), $\{\mathbf{Z}_{t-1}\}$ é um vetor p -dimensional de covariáveis aleatórias dependentes do tempo, g é uma dada função de ligação, e ϕ é um parâmetro conhecido.

Pela definição (2.2), a função de verossimilhança parcial é dada por

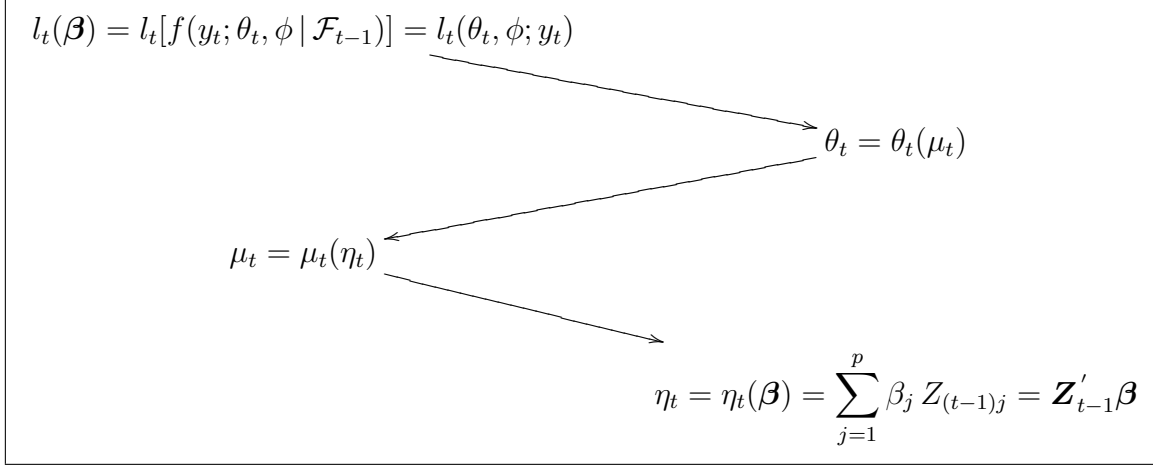
$$PL(\boldsymbol{\beta}) = \prod_{t=1}^N f(y_t; \theta_t, \phi | \mathcal{F}_{t-1}). \quad (2.6)$$

Aplicando o logaritmo em (2.6) e usando (2.3), desenvolvemos a função de log-verossimilhança parcial como

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{t=1}^N \log f(y_t; \theta_t, \phi | \mathcal{F}_{t-1}) \\ &= \sum_{t=1}^N \left\{ \frac{y_t \theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t; \phi) \right\} \\ &= \sum_{t=1}^N \left\{ \frac{y_t u(\mathbf{z}'_{t-1} \boldsymbol{\beta}) - b(u(\mathbf{z}'_{t-1} \boldsymbol{\beta}))}{\alpha_t(\phi)} + c(y_t; \phi) \right\} \equiv \sum_{t=1}^N l_t, \end{aligned} \quad (2.7)$$

onde l_t denota a t -ésima “componente” da log-verossimilhança parcial. A notação $u(\mathbf{z}'_{t-1} \boldsymbol{\beta})$ é utilizada para enfatizar a dependência da verossimilhança parcial sobre $\boldsymbol{\beta}$, através da hierarquia esquematizada no diagrama abaixo:

2.4. Estimação



lembrando que $\mu_t = E(Y_t | \mathcal{F}_{t-1}) = b'(\theta_t)$ e $g(\mu_t) = \eta_t = \mathbf{Z}'_{t-1}\boldsymbol{\beta}$. No diagrama, a notação $a = a(b)$ quer dizer que a entidade matemática a é uma função da entidade b . Ou seja, temos a relação de dependência $y_t \rightsquigarrow \theta_t \rightsquigarrow \mu_t \rightsquigarrow \eta_t \rightsquigarrow \beta_j$. Resumidamente, podemos escrever $\theta_t = (b')^{-1}(g^{-1}(\mathbf{z}'_{t-1}\boldsymbol{\beta})) = u(\mathbf{z}'_{t-1}\boldsymbol{\beta})$.

Utilizando a notação ∇ para o vetor gradiente, $\nabla \equiv \left(\frac{\partial}{\partial \beta_1}, \dots, \frac{\partial}{\partial \beta_p} \right)'$, define-se a *função escore parcial* por $\nabla l(\boldsymbol{\beta})$, para a qual será necessário o uso da regra da cadeia

$$\frac{\partial l_t}{\partial \beta_j} = \frac{\partial l_t}{\partial \theta_t} \frac{\partial \theta_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_j}, \quad j = 1, \dots, p. \quad (2.8)$$

Resolvendo as derivadas parciais em (2.8), encontra-se

$$\frac{\partial l_t}{\partial \beta_j} = \frac{(y_t - \mu_t)}{\text{Var}(Y_t | \mathcal{F}_{t-1})} \frac{\partial \mu_t}{\partial \eta_t} z_{(t-1)j}, \quad j = 1, \dots, p.$$

Da junção dos resultados encontrados, verifica-se que o *escore parcial* é um vetor de dimensão p dado por

$$\mathbf{S}_N(\boldsymbol{\beta}) \equiv \nabla l(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{Z}_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{Y_t - \mu_t(\boldsymbol{\beta})}{\sigma_t^2(\boldsymbol{\beta})}, \quad (2.9)$$

com $\sigma_t^2(\boldsymbol{\beta}) = \text{Var}(Y_t | \mathcal{F}_{t-1})$. Define-se também o *processo vetorial de escore parcial*, $\{\mathbf{S}_N(\boldsymbol{\beta})\}$, pelas somas parciais

$$\mathbf{S}_t(\boldsymbol{\beta}) = \sum_{s=1}^t \mathbf{Z}_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{Y_s - \mu_s(\boldsymbol{\beta})}{\sigma_s^2(\boldsymbol{\beta})}, \quad (2.10)$$

para $t = 1, \dots, N$.

Para os componentes da função escore, usando propriedades da esperança e variância iteradas, pode-se demonstrar que

$$E \left[\mathbf{Z}_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{Y_t - \mu_t(\boldsymbol{\beta})}{\sigma_t^2(\boldsymbol{\beta})} \mid \mathcal{F}_{t-1} \right] = \mathbf{0},$$

o que implica em $E[\mathbf{S}_N(\boldsymbol{\beta})] = \mathbf{0}$, e que os termos são ortogonais, isto é,

$$E \left[\mathbf{Z}_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{Y_s - \mu_s(\boldsymbol{\beta})}{\sigma_s^2(\boldsymbol{\beta})} \mathbf{Z}'_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{Y_t - \mu_t(\boldsymbol{\beta})}{\sigma_t^2(\boldsymbol{\beta})} \right] = \mathbf{0}, \quad s < t.$$

Ou seja, os conhecidos resultados para a função escore do MLG clássico também valem aqui, para séries temporais, sob a utilização da verossimilhança parcial.

Por fim, a solução da equação escore

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \log PL(\boldsymbol{\beta}) = \mathbf{0} \tag{2.11}$$

é denotada por $\hat{\boldsymbol{\beta}}$, e é o estimador de máxima verossimilhança parcial (EMVP) de $\boldsymbol{\beta}$.

O sistema de equações (2.11) é não-linear e é resolvido pelo tradicional método *Scoring* de Fisher, uma modificação do algoritmo iterativo do procedimento Newton-Raphson. Anteriormente à descrição do algoritmo no contexto de inferência condicional, faz-se necessário introduzir algumas matrizes importantes a ele relacionadas.

A *matriz de informação condicional cumulativa*, $\mathbf{G}_N(\boldsymbol{\beta})$, desempenha um papel importante na inferência de verossimilhança parcial, e é definida por uma soma de matrizes de covariâncias condicionais,

$$\begin{aligned} \mathbf{G}_N(\boldsymbol{\beta}) &= \sum_{t=1}^N \text{Cov} \left[\mathbf{Z}_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{Y_t - \mu_t(\boldsymbol{\beta})}{\sigma_t^2(\boldsymbol{\beta})} \mid \mathcal{F}_{t-1} \right] \\ &= \sum_{t=1}^N \mathbf{Z}_{t-1} \left(\frac{\partial \mu_t}{\partial \eta_t} \right)^2 \frac{1}{\sigma_t^2(\boldsymbol{\beta})} \mathbf{Z}'_{t-1} \\ &= \mathbf{Z}' \mathbf{W}(\boldsymbol{\beta}) \mathbf{Z}, \end{aligned} \tag{2.12}$$

com

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}'_0 \\ \mathbf{Z}'_1 \\ \vdots \\ \mathbf{Z}'_{N-1} \end{bmatrix}$$

2.4. Estimação

uma matriz $N \times p$, e $\mathbf{W}(\boldsymbol{\beta}) = \text{diag}(w_1, \dots, w_N)$ uma matriz diagonal $N \times N$ onde

$$w_t = \left(\frac{\partial \mu_t}{\partial \eta_t} \right)^2 \frac{1}{\sigma_t^2(\boldsymbol{\beta})}, \quad t = 1, \dots, N.$$

A matriz de informação incondicional é definida por

$$\text{Cov}(\mathbf{S}_N(\boldsymbol{\beta})) = E[\mathbf{G}_N(\boldsymbol{\beta})]$$

e usualmente não pode ser calculada explicitamente; entretanto, sob determinadas condições adequadas para o processo de covariáveis, ela pode ser aproximada pela matriz de informação condicional. Por último, seja $\mathbf{H}_N(\boldsymbol{\beta})$ a matriz de informação observada, definida pelo negativo da matriz hessiana de $l(\boldsymbol{\beta})$, isto é, a matriz de segundas derivadas da log-verossimilhança parcial, multiplicada por -1 ,

$$\mathbf{H}_N(\boldsymbol{\beta}) \equiv -\nabla \nabla' l(\boldsymbol{\beta}).$$

Esta matriz admite uma decomposição em termos de uma diferença entre a matriz de informação condicional e um termo residual,

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta}), \quad (2.13)$$

onde

$$\mathbf{R}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^N \mathbf{Z}_{t-1} \left[\frac{\partial^2 u(\eta_t)}{\partial \eta_t^2} \right] \mathbf{Z}'_{t-1} (Y_t - \mu_t(\boldsymbol{\beta})). \quad (2.14)$$

No caso de funções de ligação canônicas, algumas simplificações podem ser obtidas. Para a ligação canônica ($\eta_t = \theta_t$) temos que

$$\frac{\partial \mu_t}{\partial \eta_t} = \frac{\partial \mu_t}{\partial \theta_t} = b''(\theta_t),$$

e portanto o escore parcial (2.9) se transforma em

$$\mathbf{S}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^N \mathbf{Z}_{t-1} (Y_t - \mu_t(\boldsymbol{\beta})), \quad (2.15)$$

e a matriz de informação condicional (2.12) se reduz para

$$\mathbf{G}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t^2(\phi)} \sum_{t=1}^N \mathbf{Z}_{t-1} \sigma_t^2(\boldsymbol{\beta}) \mathbf{Z}'_{t-1}. \quad (2.16)$$

Também, pelo fato do uso da ligação canônica implicar em $u(\eta_t) = \eta_t$, teremos que $\frac{\partial^2 u(\eta_t)}{\partial \eta_t^2} = 0$, e portanto $\mathbf{R}_N(\boldsymbol{\beta})$ anula-se, resultando em

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}).$$

Observação: Na estimação de $\boldsymbol{\beta}$ acima descrita, considerou-se ϕ conhecido. Quando o caso não for este, sua estimação poderá ser feita, por exemplo, mediante o estimador baseado no método dos momentos

$$\hat{\phi} = \frac{1}{N-p} \sum_{t=1}^N \frac{\omega_t (Y_t - \hat{\mu}_t)^2}{V(\hat{\mu}_t)}.$$

A seguir, apresenta-se um resultado também já conhecido para o MLG clássico, de que o algoritmo de estimação de $\boldsymbol{\beta}$ por máxima verossimilhança equivale ao algoritmo de mínimos quadrados ponderados iterativamente.

2.4.1 O algoritmo de mínimos quadrados ponderados iterativamente

Como mencionado anteriormente, pelo fato de o sistema de equações-escore (2.11) ser não-linear em $\boldsymbol{\beta}$, adota-se o método *scoring* de Fisher, iterativamente, para obter sua solução. A modificação no procedimento Newton-Raphson do método Fisher é a substituição da matriz de informação observada $\mathbf{H}_N(\boldsymbol{\beta})$ por sua esperança condicional, produzindo o esquema iterativo

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{G}_N^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)}). \quad (2.17)$$

No caso de utilização da função de ligação canônica, a igualdade em (2.17) surge naturalmente, e portanto os métodos Newton-Raphson e *scoring* de Fisher se igualam.

O método *Fisher's scoring* pode ser visto como um método de estimação de mínimos quadrados ponderados. Na verdade, sob a suposição de que existe a inversa de $\mathbf{G}_N(\boldsymbol{\beta})$, (2.17) pode ser reescrita como

$$\mathbf{G}_N(\hat{\boldsymbol{\beta}}^{(k)}) \hat{\boldsymbol{\beta}}^{(k+1)} = \mathbf{G}_N(\hat{\boldsymbol{\beta}}^{(k)}) \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)}). \quad (2.18)$$

2.5. Teoria assintótica e pressuposições do modelo

Abrindo o lado direito da expressão (2.18), é possível expressá-lo em termos de uma quantidade $q_t^{(k)}$, para $t = 1, \dots, N$, definida como

$$\begin{aligned} q_t^{(k)} &= \sum_{j=1}^p Z_{(t-1)j} \hat{\beta}_j^{(k)} + (Y_t - \mu_t) \frac{\partial \mu_t}{\partial \eta_t} \\ &= \eta_t(\hat{\beta}^{(k)}) + (Y_t - \mu_t) \frac{\partial \mu_t}{\partial \eta_t} \end{aligned}$$

de forma que o lado direito de (2.18) iguala-se a $\mathbf{Z}'\mathbf{W}(\hat{\beta}^{(k)})\mathbf{q}^{(k)}$, de acordo com (2.12), e onde os elementos do vetor N -dimensional $\mathbf{q}^{(k)}$ são os $q_t^{(k)}$. A partir disto, e utilizando a expressão (2.12) para re-expressar o lado esquerdo de (2.18) temos, para esta,

$$\mathbf{Z}'\mathbf{W}(\hat{\beta}^{(k)})\mathbf{Z}\hat{\beta}^{(k+1)} = \mathbf{Z}'\mathbf{W}(\hat{\beta}^{(k)})\mathbf{q}^{(k)}.$$

Concluindo, a equação iterativa do método *scoring* de Fisher (2.17) torna-se simplificada para

$$\hat{\beta}^{(k+1)} = \left(\mathbf{Z}'\mathbf{W}(\hat{\beta}^{(k)})\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}(\hat{\beta}^{(k)})\mathbf{q}^{(k)}, \quad (2.19)$$

onde a matriz $\mathbf{W}(\hat{\beta}^{(k)})$ e o vetor $\mathbf{q}^{(k)}$ são avaliados em $\hat{\beta}^{(k)}$ em cada iteração. O limite do esquema iterativo (2.19) quando $k \rightarrow \infty$ é o estimador de máxima verossimilhança parcial $\hat{\beta}$. O procedimento iterativo (2.19) é o chamado método de *mínimos quadrados reponderados iterativamente*, onde cada iteração tem a forma de uma iteração do algoritmo de mínimos quadrados ponderados, mas com os pesos $\mathbf{W}(\hat{\beta}^{(k)})$ e as variáveis dependentes $\mathbf{q}^{(k)}$ sendo ajustados a cada iteração.

Vimos, portanto, que a maximização da verossimilhança parcial (2.6) reduz-se ao procedimento de mínimos quadrados reponderados iterativamente, válido para todos os modelos lineares generalizados, independentemente da função de ligação utilizada. A inicialização do algoritmo é feita pela simples substituição das médias condicionais por suas correspondentes respostas observadas, o que produz a primeira estimativa para a matriz de pesos \mathbf{W} e portanto um ponto de partida para $\hat{\beta}$. Cessam-se as iterações quando algum critério de convergência pré-defindo for satisfeito.

2.5 Teoria assintótica e pressuposições do modelo

Para garantir certos resultados assintóticos para a inferência via verossimilhança parcial, precisamos de algumas condições impostas sobre o modelo. Assim, Kedem e

Fokianos (2002) impõem quatro condições sobre os elementos β , \mathbf{Z}_{t-1} e g do modelo linear generalizado para séries temporais. Os pormenores concernentes a tais condições vão além do escopo desta dissertação, e portanto não foram abordados¹. Restringimo-nos, portanto, somente à menção das condições.

Pressuposições do modelo

P1. O verdadeiro parâmetro β pertence a um conjunto aberto $B \subseteq \mathfrak{R}^p$.

P2. O vetor de covariáveis \mathbf{Z}_{t-1} situa-se quase certamente em um subconjunto compacto não-aleatório Γ do espaço \mathfrak{R}^p , tal que $P(\sum_{t=1}^N \mathbf{Z}_{t-1} \mathbf{Z}'_{t-1} > 0) = 1$. Além disto, $\mathbf{Z}'_{t-1} \beta$ recai quase certamente no domínio H da função de ligação inversa $h = g^{-1}$, para todo $\mathbf{Z}_{t-1} \in \Gamma$ e $\beta \in B$.

P3. A função de ligação inversa h - definida em (**P2**) - é duas vezes continuamente diferenciável, e $|\partial h(\gamma)/\partial \gamma| \neq 0$.

P4. Existe uma medida de probabilidade ν em \mathfrak{R}^p tal que $\int_{\mathfrak{R}^p} \mathbf{z} \mathbf{z}' \nu(d\mathbf{z})$ é positiva-definida, e tal que sob (2.3) e (2.4) para conjuntos de Borel $A \subset \mathfrak{R}^p$,

$$\frac{1}{N} \sum_{t=1}^N I_{(\mathbf{Z}_{t-1} \in A)} \rightarrow \nu(A)$$

em probabilidade, à medida que $N \rightarrow \infty$, para o verdadeiro valor de β .

As condições (**P1**) a (**P4**), quando satisfeitas, garantem algumas boas propriedades ao estimador de máxima verossimilhança parcial, desde que estejam também satisfeitas as condições clássicas de regularidade da teoria de estimação de máxima verossimilhança tradicional. Uma vez satisfeitas todas estas condições, Kedem e Fokianos (2002) demonstram que o EMVP possui propriedades similares àquelas logradas pelo estimador de MV para o caso de independência dos dados. Essas propriedades incluem, para todo N suficientemente grande ($N \rightarrow \infty$), a existência, a unicidade, a consistência e a normalidade assintótica do EMVP $\hat{\beta}$.

Outro resultado importante, que é consequência das suposições do modelo, diz respeito à matriz de informação de β . Se a suposição (**P4**), de “bom comportamento” assintótico

¹Maiores informações e detalhes a respeito das suposições, e de teoremas, demonstrações e resultados sobre as mesmas, encontram-se em Kedem e Fokianos (2002), seção 1.4, e em suas referências.

2.5. Teoria assintótica e pressuposições do modelo

das covariáveis, é satisfeita, então a matriz de informação condicional $\mathbf{G}_N(\boldsymbol{\beta})$ possui um limite não-aleatório, que será aqui denotado por $\mathbf{G}(\boldsymbol{\beta})$. Isto é, existe uma *matriz de informação por observação* $p \times p$ limite, $\mathbf{G}(\boldsymbol{\beta})$, tal que

$$\frac{\mathbf{G}_N(\boldsymbol{\beta})}{N} \rightarrow \mathbf{G}(\boldsymbol{\beta}) \quad (2.20)$$

em probabilidade, quando $N \rightarrow \infty$. A condição **(P4)** garante que a matriz $\mathbf{G}(\boldsymbol{\beta})$, avaliada no verdadeiro valor de $\boldsymbol{\beta}$, seja positiva-definida, de forma que exista a sua inversa.

Por último, vale destacar que na extensão do MLG clássico para séries temporais segundo Kedem e Fokianos (2002), não se considera nenhuma suposição acerca da distribuição *conjunta* da variável resposta e das covariáveis, isto é, sobre a dinâmica conjunta de $\{Y_t, \mathbf{Z}_t\}$. As suposições e conceitos acima referidos encerram os quesitos necessários para a validade do modelo.

2.5.1 Resultados assintóticos

Kedem e Fokianos (2002) provam alguns resultados de grandes amostras para o estimador de máxima verossimilhança parcial. Suas demonstrações baseiam-se na estabilidade da matriz de informação condicional e na aplicação do teorema central do limite para martingais. Os passos da demonstração incluem (i) estabelecer a normalidade assintótica para o vetor escore parcial (2.9), usando resultados para processos martingais; (ii) mostrar que o termo residual $\mathbf{R}_N(\boldsymbol{\beta})$ (2.14) converge para zero em probabilidade, à medida que $N \rightarrow \infty$; e (iii) utilizar a expansão em série de Taylor de $\mathbf{S}_N(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ em torno de $\boldsymbol{\beta}$, truncada no primeiro termo, seguida da aplicação do teorema de Slutsky. Seguindo este percurso e provando cada um dos passos específicos, Kedem e Fokianos (2002) concluem a demonstração do teorema de substancial relevância abaixo reproduzido.

Teorema 1. *Sob as pressuposições **(P1)** a **(P4)**, o estimador de máxima verossimilhança parcial é quase certamente único para todo N suficientemente grande, e*

1. *O estimador é consistente e assintoticamente normal; isto é,*

$$\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta} \quad (2.21)$$

em probabilidade, e

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow \mathcal{N}_p(\mathbf{0}, \mathbf{G}^{-1}(\boldsymbol{\beta})), \quad (2.22)$$

em distribuição, quando $N \rightarrow \infty$.

2. A seguinte convergência em probabilidade vale, quando $N \rightarrow \infty$:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{1}{\sqrt{N}} \mathbf{G}^{-1}(\boldsymbol{\beta}) \mathbf{S}_N(\boldsymbol{\beta}) \rightarrow \mathbf{0}.$$

Este teorema é de grande importância, pois mostra que o EMVP possui as mesmas propriedades assintóticas que o estimador de MV tradicional. Isso permite que se faça a mesma inferência usual, do caso de MLGs para dados independentes, no contexto de séries temporais.

2.6 Teste de hipóteses

Uma vez estimado o modelo, é bastante comum desejar testar a hipótese de nulidade para determinado parâmetro. Isto é, testar

$$H_0 : \beta_j = 0 \quad \text{contra} \quad H_1 : \beta_j \neq 0,$$

para verificar se, dadas as demais covariáveis, a j -ésima covariável é importante para o modelo. Mais genericamente, pode-se estar interessado em testar

$$H_0 : \mathbf{p}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{versus} \quad H_1 : \mathbf{p}(\boldsymbol{\beta}) \neq \mathbf{0}, \quad (2.23)$$

onde \mathbf{p} é uma função vetorial tal que $\mathbf{p} : \Re^p \rightarrow \Re^r$, com $p > r$, sendo r o comprimento de um subvetor de $\boldsymbol{\beta}$ que se deseja testar. Por exemplo, sendo $\boldsymbol{\beta}$ um vetor de p parâmetros, considere a partição $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$, onde a dimensão de $\boldsymbol{\beta}_2$ é $r \times 1$, com $r < p$. Se estivermos interessados em testar a hipótese de nulidade do subvetor $\boldsymbol{\beta}_2$, basta fazermos $\mathbf{p}(\boldsymbol{\beta}) = \boldsymbol{\beta}_2$ em (2.23).

Apresentamos a seguir, de acordo com Kedem e Fokianos (2002), as expressões para as estatísticas de teste dos testes de razão de log-verossimilhança parcial, de Wald, e “Escore” de Rao, úteis para testar hipóteses nulas simples contra alternativas compostas da forma (2.23). Suponhamos, primeiramente, que a matriz $\mathbf{P}(\boldsymbol{\beta}) \equiv [\partial \mathbf{p}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}]$ de dimensão $p \times r$ exista, seja contínua em relação a $\boldsymbol{\beta}$, e que tenha posto r . Consideremos ainda que $\tilde{\boldsymbol{\beta}}$ seja o EMVP de $\boldsymbol{\beta}$ sob a hipótese nula H_0 de (2.23), e que $\hat{\boldsymbol{\beta}}$ seja o EMVP irrestrito de $\boldsymbol{\beta}$. Assim, temos:

2.7. Seleção de modelos

- a estatística do teste de razão de log-verossimilhança parcial,

$$\xi_{RVP} = 2 \left\{ l(\hat{\boldsymbol{\beta}}) - l(\tilde{\boldsymbol{\beta}}) \right\};$$

- a estatística de Wald,

$$\xi_W = N \mathbf{p}'(\hat{\boldsymbol{\beta}}) \left[\mathbf{P}'(\hat{\boldsymbol{\beta}}) \mathbf{G}^{-1}(\hat{\boldsymbol{\beta}}) \mathbf{P}(\hat{\boldsymbol{\beta}}) \right]^{-1} \mathbf{p}(\hat{\boldsymbol{\beta}});$$

- e a estatística do teste escore,

$$\xi_{SR} = \frac{1}{N} \mathbf{S}'_N(\tilde{\boldsymbol{\beta}}) \mathbf{G}^{-1}(\tilde{\boldsymbol{\beta}}) \mathbf{S}_N(\tilde{\boldsymbol{\beta}}).$$

Kedem e Fokianos (2002) demonstram que, sob as suposições do modelo apresentadas na seção anterior, e quando a hipótese nula de (2.23) é verdadeira, as três estatísticas de teste seguem a mesma distribuição assintótica, uma distribuição qui-quadrado com r graus de liberdade, χ_r^2 .

Um caso particular das hipóteses (2.23) que é bastante considerado em aplicações é a hipótese linear geral, testada da forma

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{contra} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

na qual \mathbf{C} é uma matriz conhecida de posto completo r , com $r < p$. Para este caso, a estatística do teste de Wald é dada por

$$\xi_W = \{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\}' \{\mathbf{C}\mathbf{G}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}'\}^{-1} \{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\},$$

ao passo que as formas para as estatísticas de razão de verossimilhança e escore permanecem as mesmas.

Outra forma de teste de hipótese, útil para modelos encaixados (*nested*), baseia-se na função desvio e será discutida subsequenteemente, no contexto de diagnóstico de modelos.

2.7 Seleção de modelos

Existem diversos métodos para comparação e seleção de modelos de regressão. Um procedimento bastante comum consiste em comparar critérios de informação calculados

para modelos competidores. Dentre estes, podem ser mencionados o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC).

Para os modelos considerados nesta dissertação, estimados por MVP, estes critérios são dados por

$$\text{AIC}(k) = -2 \log PL(\hat{\boldsymbol{\beta}}) + 2k$$

e

$$\text{BIC}(k) = -2 \log PL(\hat{\boldsymbol{\beta}}) + k \log N,$$

onde $\hat{\boldsymbol{\beta}}$ é o EMVP de $\boldsymbol{\beta}$ e k é o número de parâmetros do modelo. Por estes critérios, opta-se pelo modelo com *menor* valor calculado para as medidas AIC e BIC.

Kedem e Fokianos (2002) ressaltam que o critério BIC é preferível ao AIC. Isto é conhecido na literatura dos modelos ARMA, nos quais o AIC apresenta a tendência de indicar o modelo com maior número de parâmetros.

2.8 Diagnóstico

A análise de diagnóstico em modelos de regressão consiste no uso de procedimentos para verificar o quanto um modelo ajustado se adequa aos dados e para medir sua bondade de ajuste. No contexto de MLGs, faz-se o diagnóstico pela *análise do desvio* e pela verificação dos *resíduos* do modelo.

2.8.1 Análise do desvio

A qualidade de ajuste de um MLG é avaliada por meio da função desvio, que é uma medida usada para a comparação entre o modelo sob investigação e um modelo maior, o modelo saturado. O modelo *saturado* é um modelo com o número máximo de parâmetros que pode ser estimado, isto é, com o número de parâmetros igual ao número de observações. Neste modelo, μ_t é estimado diretamente a partir dos dados Y_1, \dots, Y_N , de maneira a termos N parâmetros, μ_1, \dots, μ_N , e de forma que o EMVP de cada um deles é dado por $\tilde{\mu}_t = Y_t$. O modelo sob investigação é denominado modelo *reduzido*, e neste caso o conjunto de N parâmetros μ_t , $t = 1, \dots, N$, é expressado em termos de um conjunto menor, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ com p parâmetros, $p < N$. Denotando por $l(\mathbf{y}; \mathbf{y})$ o valor da função de log-verossimilhança parcial avaliado nas estimativas $\tilde{\mu}_t$ do modelo saturado, e

2.8. Diagnóstico

por $l(\hat{\boldsymbol{\mu}}; \mathbf{y})$ o valor da log-VP avaliado na estimativa de MVP de $\boldsymbol{\beta}$ do modelo reduzido, define-se a **função desvio** por

$$D \equiv 2 \{ l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y}) \}. \quad (2.24)$$

Um valor pequeno para a função desvio indica que, para um número menor (p) de parâmetros, obtém-se um ajuste próximo ao obtido com o modelo saturado.

A estatística (2.24) é utilizada para proceder testes de hipóteses sobre os parâmetros de modelos encaixados, e este procedimento é denominado *análise do desvio*. Seja D_0 o valor da função desvio sob $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, e D_1 seu valor correspondente a $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1$, onde $\boldsymbol{\beta}_1$ é um vetor de dimensão p e $\boldsymbol{\beta}_0$ é um subvetor de $\boldsymbol{\beta}_1$ com dimensão $q \times 1$, $q < p$. Então, para N grande o suficiente, temos que $D_0 \sim \chi_{N-q}^2$, $D_1 \sim \chi_{N-p}^2$ e, sob H_0 ,

$$D_0 - D_1 \sim \chi_{p-q}^2, \quad (2.25)$$

aproximadamente². Rejeita-se a hipótese nula, ao nível de significância α , para valores grandes de $D_0 - D_1$, maiores que o percentil $(1 - \alpha) \times 100\%$ da distribuição χ_{p-q}^2 .

Quando ϕ for desconhecido, uma estatística de teste alternativa a (2.25) para se testar H_0 é

$$\frac{(D_0 - D_1)/(p - q)}{D_1/(N - p)},$$

que não depende de ϕ e é distribuída sob H_0 , aproximadamente, como uma $F_{p-q, N-p}$.

2.8.2 Resíduos

Os resíduos de um modelo são úteis para a avaliação de sua bondade de ajuste. Eles medem discrepâncias entre os valores observados Y_1, \dots, Y_N e seus valores ajustados μ_1, \dots, μ_N . Seja $\hat{\mu}_t = \mu_t(\hat{\boldsymbol{\beta}})$. Existem diversas formas de definir resíduos no contexto de MLGs para séries temporais. A definição mais óbvia é a dos resíduos “puros”,

$$\hat{e}_t = Y_t - \hat{\mu}_t, \quad t = 1, \dots, N, \quad (2.26)$$

chamados resíduos *de resposta*. Três outros tipos comuns de resíduos, os resíduos de Pearson, de trabalho, e os componentes da função desvio, são definidos em termos dos

²Para detalhes, vide seção 1.6.1 de Kedem e Fokianos (2002) e seções 2.3 e 4.4.3 de McCullagh e Nelder (1989).

resíduos de resposta. Os resíduos de *Pearson* são sua versão padronizada,

$$\hat{r}_t = \frac{Y_t - \hat{\mu}_t}{\sqrt{V(\hat{\mu}_t)}}, \quad t = 1, \dots, N. \quad (2.27)$$

Os resíduos de *trabalho* são uma versão diferente de resíduos padronizados,

$$\hat{w}r_t = \frac{Y_t - \hat{\mu}_t}{\partial \mu_t / \partial \eta_t}, \quad t = 1, \dots, N, \quad (2.28)$$

onde $\partial \mu_t / \partial \eta_t$ é avaliada em $\hat{\beta}$. E os resíduos *componentes do desvio* são dados por

$$\hat{d}_t = \text{sinal}(Y_t - \hat{\mu}_t) \sqrt{2[l_t(Y_t) - l_t(\hat{\mu}_t)]}, \quad t = 1, \dots, N, \quad (2.29)$$

de forma que

$$\sum_{t=1}^N \hat{d}_t^2 = D,$$

a função desvio definida em (2.24).

No contexto dos MLGs, os resíduos são usados para explorar a adequação do modelo ajustado com respeito à escolha da distribuição para os dados, da função de variância, da função de ligação e dos termos no preditor linear, e ainda para verificar a existência de observações atípicas (McCullagh e Nelder, 1989). Contudo, são desconhecidas as propriedades dos resíduos de MLGs para séries temporais com memória longa. Assim, estaremos interessados nesta dissertação em avaliar apenas alguns aspectos mais básicos dos resíduos, como os de ausência de correlação com as covariáveis, homocedasticidade, e principalmente o de ausência de correlação serial.

A ausência de correlação serial será investigada através do teste de Ljung-Box (Ljung and Box, 1978), cuja estatística é dada por

$$Q(k) = N(N+2) \sum_{j=1}^k \frac{\hat{\rho}_j^2}{N-j}, \quad (2.30)$$

sendo $\hat{\rho}_j$ a estimativa da autocorrelação entre os resíduos para a defasagem j , e k o número de defasagens escolhido para se testar. Rejeita-se a hipótese nula de que as autocorrelações até a defasagem k são nulas para valores grandes de $Q(k)$, comparados ao valor crítico da distribuição χ_k^2 .

Séries Temporais com Memória Longa

Neste capítulo serão apresentados brevemente os conceitos básicos de processos com memória longa. Estes aspectos estão fortemente vinculados à modelagem das séries utilizadas nas aplicações desta dissertação. Começaremos apresentando a definição de memória longa, e em seguida faremos uma introdução ao modelo ARFIMA. Um procedimento de estimação será brevemente descrito. Adicionalmente, apresentaremos a forma simples de aproximação do modelo ARFIMA(1, d , 1)¹ por um AR(p) utilizada nas aplicações desta dissertação para a previsão das séries covariáveis.

3.1 Processos com memória longa

Processos estacionários são ditos possuir “memória curta” se suas autocorrelações ρ_j decrescem rapidamente para zero segundo uma taxa exponencial. Isto é, se

$$|\rho_j| \leq Cr^{-j}, \quad j = 1, 2, \dots$$

onde $C > 0$ e $0 < r < 1$, de forma que

$$\lim_{n \rightarrow \infty} \sum_{j=-n}^n |\rho_j| \quad (3.1)$$

¹Neste trabalho, foi suficiente utilizar o modelo ARFIMA(p, d, q) de ordem apenas 1 para os termos AR e MA ($p = q = 1$) para modelar as séries covariáveis. Assim, apresentaremos uma forma de aproximação do ARFIMA(1, d , 1) por um modelo autoregressivo de ordem p .

seja finita. Exemplos de processos com memória curta são os ARMA (*AutoRegressivos de Médias Móveis*), amplamente conhecidos, propostos por Box e Jenkins (1976). No caso da memória longa, este decaimento é bem mais lento, e ocorre segundo uma taxa hiperbólica em j . Define-se um processo estacionário como um processo de **memória longa** (ML) aquele para o qual

$$\rho_j \sim Cj^{2d-1}, \quad j \rightarrow \infty, \quad (3.2)$$

com (3.1) não-finita, sendo $C > 0$ e d um número real.

Historicamente, o fenômeno da memória longa foi primeiramente observado na área de Hidrologia, na década de 50, por Hurst (1951, 1957). Obteve-se evidências de que séries de vazão e de capacidade de armazenamento de reservatórios, observadas por um longo período de tempo, possuíam dependência de longo alcance. Mais recentemente, passou-se a verificar a presença de ML em séries de diversas outras áreas do conhecimento, dentre as quais podem ser citadas a área de climatologia (Seater, 1993) e a área financeira (Baillie, 1996).

3.2 O modelo ARFIMA

O processo de memória longa mais utilizado em aplicações é o modelo ARFIMA (*AutoRegressivo Fracionário Integrado de Médias Móveis*), introduzido por Granger e Joyeux (1980) e Hosking (1981) como uma generalização do modelo ARIMA de Box e Jenkins (1976). Dizemos que $\{y_t\}$ é um processo ARFIMA(p, d, q) se for estacionário e satisfizer a equação

$$\phi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t, \quad (3.3)$$

onde $d \in (-0,5; 0,5)$, $\{\varepsilon_t\}$ é um processo ruído branco com média 0 e variância σ^2 , e $\phi(B)$ e $\theta(B)$ são polinômios sem raízes comuns de ordens p e q dados por

$$\begin{aligned} \phi(B) &= 1 + \phi_1 B + \phi_2 B^2 + \dots + \phi_p B^p \quad \text{e} \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q, \end{aligned}$$

e são tais que as raízes de $\phi(B) = 0$ e $\theta(B) = 0$ estão fora do círculo unitário. O operador de diferença fracionária $\nabla^d \equiv (1 - B)^d$, para qualquer número real $d > -1$, é definido

3.2. O modelo ARFIMA

pela expansão binomial

$$(1 - B)^d = \sum_{j=0}^{\infty} \binom{d}{j} (-B)^j = \sum_{j=0}^{\infty} \eta_j B^j = \eta(B), \quad (3.4)$$

onde η_j é dado por

$$\begin{cases} \eta_0 = 1, \\ \eta_j = \frac{(j-1-d)}{j} \eta_{j-1} = \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)}, \quad j \geq 1. \end{cases}$$

Dizemos ainda que $\{y_t\}$ é um processo ARFIMA(p, d, q) com média μ se $\{y_t - \mu\}$ for um ARFIMA(p, d, q).

A Figura 3.1 exemplifica uma comparação entre as funções de autocorrelação obtidas para os casos de curta e longa memória. Foram simulados um processo de memória curta $\{X_t\} \sim \text{AR}(1)$ com $\phi = 0,8$ e um processo com ML $\{W_t\} \sim \text{ARFIMA}(1, d, 0)$ com $\phi = 0,8$ e $d = 0,4$, ambos com $\mu = 10$ e inovações i.i.d. $\sim \mathcal{N}(0, 1)$. Nota-se o rápido decaimento para zero, no caso da curta memória, em oposição a um lento decaimento, hiperbólico, para o caso de memória longa.

Hosking (1981) demonstra que o processo ARFIMA(p, d, q) dado por (3.3) será:

- (i) estacionário, se $d < \frac{1}{2}$ e todas as raízes de $\phi(B) = 0$ estiverem fora do círculo unitário;
- (ii) invertível, se $d > -\frac{1}{2}$ e todas as raízes de $\theta(B) = 0$ estiverem fora do círculo unitário.

Um caso bastante especial dos modelos ARFIMA é o ARFIMA($0, d, 0$), denominado *ruído fracionário* (RF(d)) e dado por $(1 - B)^d y_t = \varepsilon_t$. Para este processo, as funções de autocovariância e de autocorrelação são dadas por

$$\gamma(k) = \text{Cov}(Y_t, Y_{t+k}) = \sigma^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}, \quad (3.5)$$

e

$$\rho(k) = \text{Corr}(Y_t, Y_{t+k}) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(k+d)}{\Gamma(1+k-d)}. \quad (3.6)$$

Para o processo ARFIMA(p, d, q) geral (3.3), $\gamma(k)$ assume uma expressão mais complexa, cujos detalhes podem ser obtidos na seção 3.2.4 de Palma (2007).

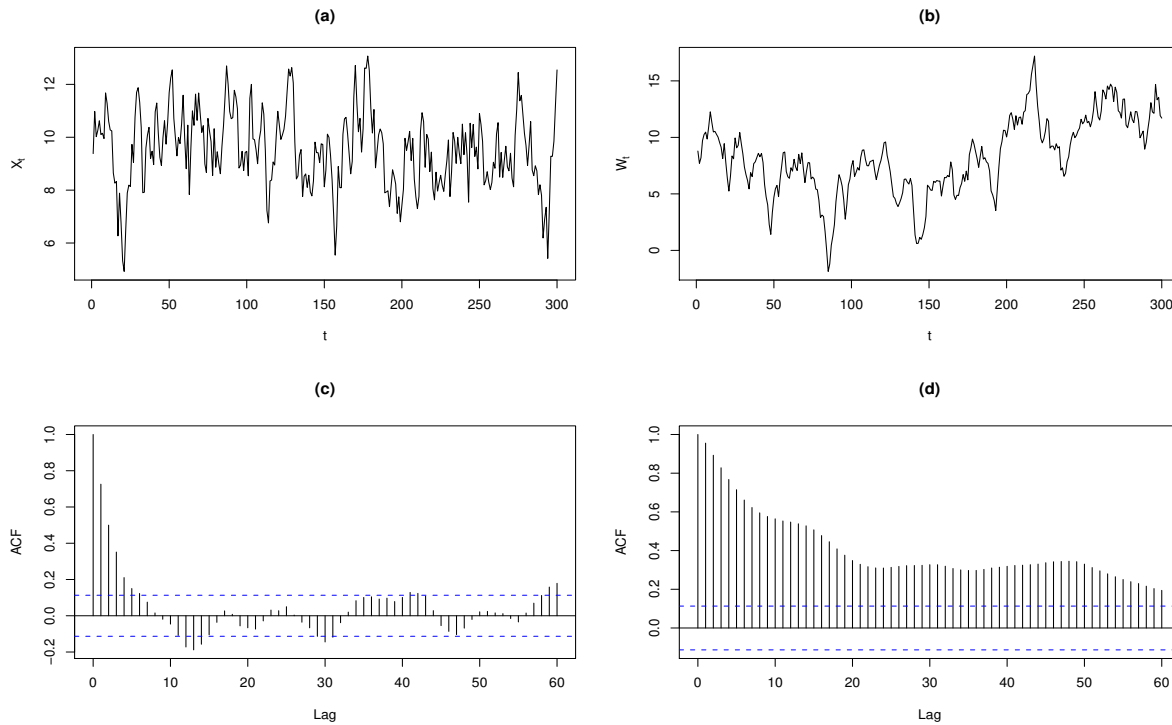


Figura 3.1: Comparação entre as autocorrelações de séries simuladas com e sem memória longa: (a), (c) X_t e sua FAC; (b), (d) W_t e sua FAC.

3.3 Estimação

Existem diversos métodos para a estimação de modelos de memória longa. Temos, por exemplo, os métodos baseados em estimação por máxima verossimilhança exata ou aproximada. Existem também métodos desenvolvidos no domínio do tempo e outros no domínio da frequência, e ainda métodos paramétricos e outros semiparamétricos.

Chan e Palma (1998) propuseram uma abordagem bastante interessante para estimação, baseada na representação de espaço de estados do processo ARFIMA(p, d, q). Embora qualquer representação do ARFIMA na forma de modelo de espaço de estados seja infinito-dimensional, Chan e Palma (1998) demonstram que a função de verossimilhança exata pode ser calculada em um número finito de passos. Em geral, para uma série de N observações y_1, \dots, y_N , as primeiras N componentes das equações de Kalman já serão suficientes. Estes fatos motivam a utilização de representação na forma de espaço de estados para se fazer estimação.

3.4. Previsão

Neste trabalho, utilizamos a estimação via filtro de Kalman da seguinte forma: o modelo ARFIMA foi primeiramente aproximado por um modelo ARMA, sendo em seguida expressado na forma de modelo de espaço de estados. A estimação é então realizada no modelo aproximado. Este procedimento tem um baixo custo computacional, e ainda permite fazer diagnóstico. O algoritmo para todo o processo de estimação e obtenção dos resíduos foi implementado na linguagem R (R Development Core Team, 2009).

Os detalhes teóricos concernentes à metodologia de estimação acima descrita não integram o foco dos objetivos deste trabalho e estão além do escopo projetado para ele. Para informações adicionais, consultar por exemplo o Capítulo 4 de Palma (2007).

3.4 Previsão

Há várias formas de se fazer previsão para o modelo ARFIMA. Uma forma bastante simples consiste em aproximar o processo ARFIMA por um modelo $AR(p)$, e então calcular as previsões para este $AR(p)$ aproximador. Nas aplicações desta dissertação, o modelo ARFIMA(1, d , 1) ou outro ARFIMA de ordem menor foi capaz de explicar a estrutura das séries. Assim, apresentamos a seguir uma forma de aproximação do ARFIMA(1, d , 1) por um $AR(p)$, e a previsão de um passo à frente associada.

Considere o modelo ARFIMA(1, d , 1), dado por

$$(1 + \phi B)(1 - B)^d y_t = (1 + \theta B)\varepsilon_t. \quad (3.7)$$

Queremos aproximar este modelo pelo $AR(p)$,

$$\pi(B) y_t = \varepsilon_t, \quad (3.8)$$

onde

$$\pi(B) = 1 + \pi_1 B + \pi_2 B^2 + \cdots + \pi_p B^p.$$

Para fazer isto, precisamos igualar os polinômios que multiplicam y_t nas equações (3.7) e (3.8),

$$\left. \begin{array}{l} \frac{(1+\phi B)}{(1+\theta B)} (1 - B)^d y_t = \varepsilon_t \\ \pi(B) y_t = \varepsilon_t \end{array} \right\} \Rightarrow (1 + \theta B) \pi(B) = (1 + \phi B) \eta(B), \quad (3.9)$$

onde $\eta(B) = (1 - B)^d$, e então determinar os coeficientes π_j do $AR(p)$ aproximador em termos dos parâmetros ϕ e θ e dos coeficientes η_j do ARFIMA(1, d , 1). Isolando e

igualando os termos das potências do operador de defasagem B de cada lado da equação (3.9), obtemos

$$\begin{aligned}\pi_1 + \theta\pi_0 &= \eta_1 + \phi\eta_0 \\ \pi_2 + \theta\pi_1 &= \eta_2 + \phi\eta_1 \\ &\vdots \\ \pi_p + \theta\pi_{p-1} &= \eta_p + \phi\eta_{p-1}.\end{aligned}$$

Portanto, o ARFIMA(1, d , 1) pode ser aproximado por um AR(p) com coeficientes

$$\pi_j = \eta_j + \phi\eta_{j-1} - \theta\pi_{j-1}, \quad j = 1, \dots, p, \quad (3.10)$$

lembrando que $\pi_0 = \eta_0 = 1$. No caso do RF(d), por exemplo, temos que $\pi_j = \eta_j$. Na literatura, adota-se comumente o valor $p = 40$ para a aproximação.

Para obter as previsões para este modelo aproximador, é útil primeiramente expressá-lo, a partir de (3.8), em sua forma autoregressiva direta,

$$y_t = -\pi_1 y_{t-1} - \pi_2 y_{t-2} - \dots - \pi_p y_{t-p} + \varepsilon_t, \quad (3.11)$$

em que o valor atual de y_t é explicado por seus valores passados. Calculando a previsão de mínimo erro quadrático médio (EQM) aproximada de um passo à frente para y_t , a partir de (3.11), obtemos

$$\begin{aligned}\hat{y}_t(1) &= E(Y_{t+1} | Y_t, Y_{t-1}, \dots, Y_1) \\ &= E(-\pi_1 Y_t - \pi_2 Y_{t-1} - \dots - \pi_p Y_{t-p+1} + \varepsilon_{t+1} | Y_t, Y_{t-1}, \dots, Y_1) \\ &= -\pi_1 y_t - \pi_2 y_{t-1} - \dots - \pi_p y_{t-p+1}.\end{aligned} \quad (3.12)$$

Portanto, denotando por T o último instante de observação da série, a previsão de um passo à frente para $t = T + 1$ é dada por

$$\hat{y}_T(1) = -\pi_1 y_T - \pi_2 y_{T-1} - \dots - \pi_p y_{T-p+1}.$$

Regressão Logística para Séries Binárias

O modelo de regressão logística é um caso particular de MLG que tem sido bastante estudado na literatura de séries temporais. Liang e Zeger (1989), por exemplo, desenvolveram uma classe de modelos de regressão logística para séries temporais binárias multivariadas, baseados em cadeia de Markov e estimados via máxima pseudo-verossimilhança. Outros autores estenderam o modelo logístico para séries temporais binárias de Kedem e Fokianos (2002) para situações mais específicas. Dentre estes, vale mencionar o trabalho de Hung *et al.* (2008), onde o modelo logístico de Kedem e Fokianos (2002) foi estendido de forma a incorporar parâmetros de efeitos aleatórios. Na verdade, a abordagem de estimação do modelo logístico para séries binárias de Kedem e Fokianos (2002) que vamos apresentar a seguir é devida a Slud e Kedem (1994).

Neste capítulo, apresentaremos o modelo logístico com função de ligação logito, sua estimação por verossimilhança parcial e estudos de simulação visando à avaliação do comportamento assintótico do estimador de máxima verossimilhança parcial sob a presença de memória longa nas séries. Especificamente, queremos avaliar se as propriedades de consistência (2.21) e normalidade assintótica (2.22) do EMVP, declaradas no Teorema 1 do Capítulo 2, permanecem válidas quando séries com ML são modeladas. Apresentaremos também a metodologia que foi utilizada em todas as aplicações da dissertação para a modelagem das séries covariáveis. Este tópico importante está relacionado ao segundo objetivo específico do presente trabalho, de avaliar a capacidade preditiva dos modelos quando as covariáveis são modeladas e preditas. Visando à avaliação da performance preditiva do modelo logístico na aplicação aos dados reais, definiremos ainda algumas

medidas de performance preditiva usuais no problema de classificação binária. Por fim, encerraremos o capítulo com uma aplicação do modelo a dados de poluição do ar.

As simulações e a análise dos dados da aplicação foram todas realizadas no *software* livre R, versão 2.9.0 (R Development Core Team, 2009).

4.1 O modelo de regressão logística

Considere que $\{Y_t\}$, $t = 1, \dots, N$, seja a série temporal binária de interesse e que o processo de covariáveis a ela relacionado seja dado pelo vetor p -dimensional $\{\mathbf{Z}_{t-1}\}$, que pode incluir valores passados de Y_t e das próprias séries covariáveis. Utilizando-se das informações contidas em \mathbf{Z}_{t-1} , interessa estudar a probabilidade de “sucesso” condicional

$$P_{\boldsymbol{\beta}}(Y_t = 1 | \mathcal{F}_{t-1}),$$

onde \mathcal{F}_{t-1} denota toda a informação do passado até o momento $t-1$. Para tal, o **modelo de regressão logística** com função de ligação logito é dado por

$$\pi_t(\boldsymbol{\beta}) \equiv P_{\boldsymbol{\beta}}(Y_t = 1 | \mathcal{F}_{t-1}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}' \mathbf{Z}_{t-1})}, \quad t = 1, \dots, N, \quad (4.1)$$

e é mais comumente expresso na forma

$$\text{logito}(\pi_t(\boldsymbol{\beta})) \equiv \log \left[\frac{\pi_t(\boldsymbol{\beta})}{1 - \pi_t(\boldsymbol{\beta})} \right] = \boldsymbol{\beta}' \mathbf{Z}_{t-1}, \quad (4.2)$$

onde se enfatiza o uso da função logito como função de ligação.

4.2 Estimação

A variável resposta binária tem função massa de probabilidade condicionada ao passado dada por uma Bernoulli($\pi_t(\boldsymbol{\beta})$),

$$P_{\boldsymbol{\beta}}(Y_t = y_t | \mathcal{F}_{t-1}) = [\pi_t(\boldsymbol{\beta})]^{y_t} [1 - \pi_t(\boldsymbol{\beta})]^{1-y_t},$$

de onde se obtém a função de verossimilhança parcial

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^N [\pi_t(\boldsymbol{\beta})]^{y_t} [1 - \pi_t(\boldsymbol{\beta})]^{1-y_t}$$

4.3. Estudos de simulação

e, através de (4.1), a função de log-verossimilhança parcial

$$l(\boldsymbol{\beta}) = \sum_{t=1}^N \{ y_t \log [1 + \exp(-\boldsymbol{\beta}' \mathbf{z}_{t-1})]^{-1} + (1 - y_t) \log [1 + \exp(\boldsymbol{\beta}' \mathbf{z}_{t-1})]^{-1} \}.$$

Derivando esta expressão com relação a $\boldsymbol{\beta}$ e realizando manipulações algébricas simples, obtém-se o vetor escore parcial de $\boldsymbol{\beta}$,

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{z}_{t-1} (Y_t - \pi_t(\boldsymbol{\beta})).$$

Devido à utilização da função de ligação logito, que é a canônica, a matriz de informação observada $\mathbf{H}_N(\boldsymbol{\beta})$ é a própria matriz de informação condicional cumulativa $\mathbf{G}_N(\boldsymbol{\beta})$, e é obtida a partir da diferenciação do negativo do vetor escore parcial, encontrando-se

$$\begin{aligned} \mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) &= \sum_{t=1}^N \mathbf{z}_{t-1} \mathbf{z}'_{t-1} \pi_t(\boldsymbol{\beta})(1 - \pi_t(\boldsymbol{\beta})) \\ &= \sum_{t=1}^N \mathbf{z}_{t-1} \mathbf{z}'_{t-1} \frac{\exp(\boldsymbol{\beta}' \mathbf{z}_{t-1})}{[1 + \exp(\boldsymbol{\beta}' \mathbf{z}_{t-1})]^2}. \end{aligned}$$

A partir das expressões para o vetor escore e para a matriz de informação, obtém-se o estimador de máxima verossimilhança parcial $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ pelo esquema iterativo (2.17).

4.3 Estudos de simulação

Apresenta-se primeiramente uma análise pormenorizada para uma única série simulada, sob o objetivo de verificar se as estimativas dos parâmetros são significativas. Em seguida, o mesmo modelo é simulado 1000 vezes para avaliarmos o comportamento assintótico do estimador de máxima verossimilhança parcial do modelo logístico para séries binárias com memória longa.

4.3.1 Estudo detalhado de uma série simulada

Gerou-se $N = 500$ observações para Y_t a partir do modelo

$$\log \left(\frac{\mu_t}{1 - \mu_t} \right) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 W_t, \quad t = 1, \dots, N, \quad (4.3)$$

onde o processo de covariáveis $\mathbf{Z}_{t-1} = (Y_{t-1}, W_t)'$ possui a primeira defasagem de Y_t e uma série covariável de memória longa, W_t , gerada de acordo com as especificações

$$\{W_t\} \sim ARFIMA(1, d, 1), \quad \text{com } \phi = 0,6, d = 0,45, \theta = 0,3, \mu_\varepsilon = 8 \text{ e } \sigma_\varepsilon = 1, \quad (4.4)$$

onde μ_ε e σ_ε são a média e o desvio-padrão utilizados na geração do processo de inovações $\{\varepsilon_t\} \stackrel{iid}{\sim} \mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon)$ do processo $\{W_t\}$. Os parâmetros para a simulação foram especificados como

$$\boldsymbol{\beta} = (-4, 3, 2, 7, 0, 5)'. \quad (4.5)$$

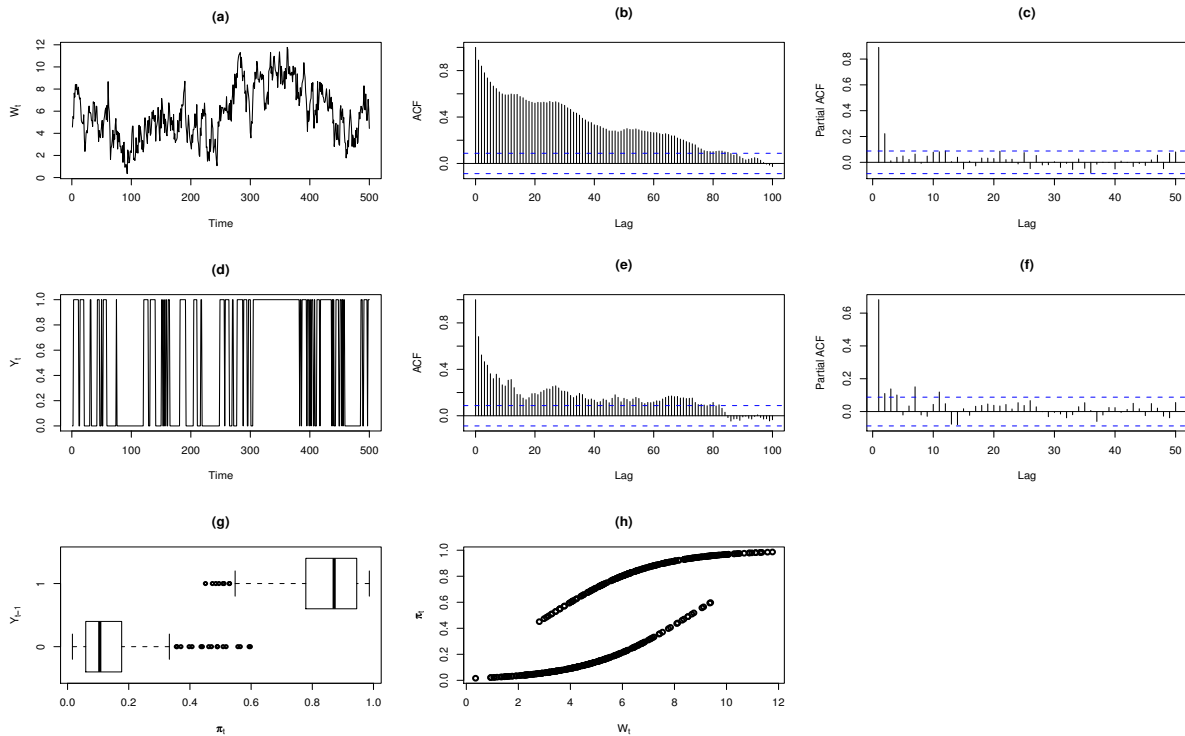


Figura 4.1: Séries simuladas: (a), (b) e (c) W_t , sua FAC e FACP; (d), (e) e (f) Y_t , sua FAC e FACP; (g) dispersão de $\pi_t(\boldsymbol{\beta})$ para cada nível de Y_{t-1} ; (h) dispersão de $\pi_t(\boldsymbol{\beta})$ versus W_t .

As Figuras 4.1 (a), (b) e (c) apresentam a série $\{W_t\}$ gerada e suas funções de autocorrelação e de autocorrelação parcial. Utilizando-se de seus valores, gerou-se os valores de $\{Y_t\}$, um a um, da maneira descrita a seguir. Primeiramente, gera-se Y_0 , um valor de “inicialização” da cadeia, de acordo com uma Bernoulli(0,5). Este valor, e o primeiro valor da

4.3. Estudos de simulação

série W_t (W_1), são então utilizados no modelo (4.3), expresso na forma (4.1), para a obtenção de $\pi_1(\boldsymbol{\beta})$. Daí, o primeiro valor de Y_t , Y_1 , é gerado segundo uma Bernoulli(π_1). Para a geração de Y_2 , $\pi_2(\boldsymbol{\beta})$ é obtida a partir dos valores Y_1 e W_2 , e então $Y_2 \sim \text{Bernoulli}(\pi_2)$. Repetindo esse processo N vezes, obtém-se a série simulada $\{Y_t\}$ (Figura 4.1 (d)). Pode-se observar pelas Figuras 4.1 (e) e (f) que Y_t “incorporou” memória longa, porém menos que W_t . Após este processo, a série covariável Y_{t-1} é obtida pelo simples defasamento em 1 tempo da série Y_t .

Uma vez que foram considerados parâmetros de efeito *positivos* na simulação, era de se esperar que a série de probabilidades $\pi_t(\boldsymbol{\beta})$, geradora da série Y_t , tivesse um relacionamento *crescente* (diretamente proporcional) com as séries covariáveis. De fato, isso se confirma pelas Figuras 4.1 (g) e (h), que indicam que valores “sucesso” (‘1’) de Y_t tendem a ocorrer para $Y_{t-1} = 1$ e para valores altos de W_t , pois μ_t assume valor alto nestas duas situações. Pode-se visualizar no relacionamento entre μ_t e W_t (Figura 4.1 (h)), por exemplo, uma curva sigmóide fragmentada, sendo sua parte superior correspondente aos tempos t em que $Y_{t-1} = 1$, e a inferior, correspondente aos instantes em que $Y_{t-1} = 0$. Verifica-se, portanto, que valores $Y_t = 1$ são gerados com probabilidade alta quando $Y_{t-1} = 1$ e W_t é alto. Estas observações são úteis para a confirmação de que a simulação foi coerente.

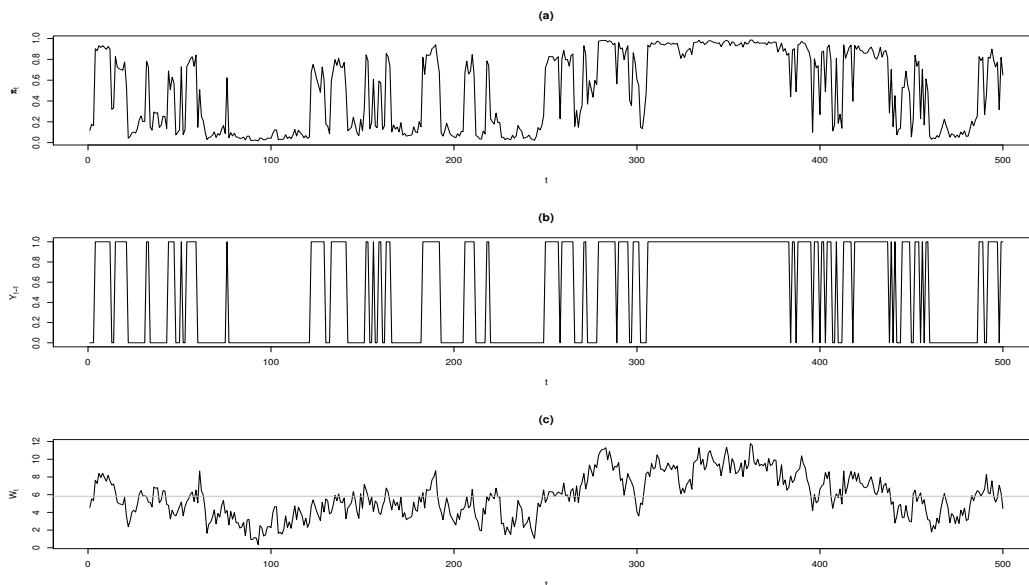


Figura 4.2: Séries $\pi_t(\boldsymbol{\beta}) = P(Y_t = 1 | \mathcal{F}_{t-1})$ (a), Y_{t-1} (b) e W_t (com sua linha média) (c).

Outra forma de visualizar e confirmar o relacionamento existente entre os valores

gerados para Y_t , e os valores das séries covariáveis, é por meio da Figura 4.2. Pode-se ver que π_t oscilou numa faixa de valores mais altos para $250 < t < 450$, aproximadamente, intervalo no qual Y_{t-1} assume predominantemente o valor 1 e W_t está bem acima de sua média. Na maior parte dos tempos restantes, W_t é baixa e $Y_{t-1} = 0$, de forma que π_t oscila mais em torno de valores baixos. A boa concordância que se observa entre Y_{t-1} e W_t na geração de π_t também se justifica, pois $\rho(Y_{t-1}, W_t) = 0,57$.

A série $\{Y_t\}$ descrita foi então utilizada para a estimação de β , segundo (4.3). As estimativas constam da Tabela 4.1. Nota-se que as estimativas são altamente significativas e apresentam valores próximos aos valores reais dos parâmetros.

Tabela 4.1: Estimação do modelo a partir dos dados simulados.

Parâmetro	Real	Estimativa	E.P.	Valor t	p -valor
β_0	-4,3	-4,71	0,485	-9,7	<0,001
β_1	2,7	2,54	0,269	9,4	<0,001
β_2	0,5	0,61	0,081	7,5	<0,001

4.3.2 Estudo de simulação geral

O modelo logístico (4.3) acima estudado foi simulado $k = 1000$ vezes para diversas combinações do tamanho da série N (200, 350, 500 e 1000) e do parâmetro de diferenciação fracionária d (0,2; 0,3; 0,4; 0,45 e 0,49), mantendo-se o mesmo valor para β . Por brevidade de apresentação, e devido à similaridade dos resultados obtidos para todos os casos, restringe-se aqui à apresentação das combinações de $N = 200, 500$ e 1000 e $d = 0,2, 0,4$ e $0,49$. Estes valores de d foram escolhidos de forma a se apresentar o comportamento dos estimadores sob a presença de série com memória longa “fraca” (0,2), “média” (0,4) e “forte” (0,49).

Foram calculadas as médias amostrais das 1000 estimativas $\hat{\beta}_k$, bem como seus desvios-padrões amostrais, os quais designaremos aqui por *erros-padrões da simulação* e representaremos por \hat{EP} . Também, para comparação com estes, obtiveram-se os erros-padrões teóricos aproximados¹, a partir da inversa da matriz de informação condicional cumulativa, $(\mathbf{G}_N(\hat{\beta}))^{-1}$, calculada a partir da estimativa obtida para $\hat{\beta}$. Utilizaremos a notação

¹De acordo com Kedem e Fokianos (2002).

4.3. Estudos de simulação

\mathbf{G}_N^{-1} para representar os erros-padrões obtidos a partir de $\mathbf{G}_N^{-1}(\hat{\boldsymbol{\beta}})$. A Tabela 4.2 apresenta os resultados das simulações. Verifica-se a partir desta que:

- independentemente do valor de d , as estimativas obtidas para os parâmetros são precisas, e parecem convergir para seus valores reais à medida em que o tamanho amostral é aumentado; isto, juntamente com o decréscimo assintótico de seus erros-padrões, sugere a *consistência* dos estimadores de MVP;
- as estimativas dos erros-padrões da simulação, \hat{EP} 's, estão muito próximas das estimativas dos erros-padrões “teóricos”, e esta proximidade parece não ser afetada pela variação do tamanho da série gerada ou do parâmetro de diferenciação fracionária; este fato sugere que a presença de memória longa nas séries do modelo não impede que se faça inferência *correta* para os erros-padrões dos estimadores de MVP, a partir da inversa da matriz de informação.

Tabela 4.2: Resultado das $k = 1000$ simulações (valor real: $\boldsymbol{\beta} = (-4, 3, 2, 7, 0, 5)'$).

		$d = 0, 2$			$d = 0, 4$			$d = 0, 49$		
N	Estat.	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
200	$\hat{\boldsymbol{\beta}}$	-4,46	2,65	0,53	-4,47	2,65	0,53	-4,52	2,65	0,54
	\mathbf{G}_N^{-1}	1,29	0,44	0,18	1,14	0,40	0,16	1,21	0,49	0,16
	\hat{EP}	1,44	0,44	0,19	1,06	0,40	0,15	1,27	0,49	0,16
500	$\hat{\boldsymbol{\beta}}$	-4,34	2,69	0,51	-4,33	2,67	0,51	-4,36	2,68	0,51
	\mathbf{G}_N^{-1}	0,86	0,26	0,11	0,63	0,25	0,10	0,71	0,29	0,09
	\hat{EP}	0,89	0,28	0,11	0,59	0,24	0,09	0,74	0,29	0,09
1000	$\hat{\boldsymbol{\beta}}$	-4,33	2,69	0,51	-4,34	2,69	0,51	-4,34	2,70	0,51
	\mathbf{G}_N^{-1}	0,64	0,20	0,08	0,34	0,18	0,06	0,35	0,17	0,05
	\hat{EP}	0,63	0,20	0,08	0,36	0,17	0,06	0,36	0,18	0,06

Interessa-nos agora avaliar o comportamento assintótico da distribuição de $\hat{\boldsymbol{\beta}}$. A Figura 4.3 apresenta as seqüências de estimativas obtidas para $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ na simulação para a combinação que representaria o pior caso, a de $N = 200$ (amostra pequena) e $d = 0, 49$ (próximo do limite de estacionariedade). Vemos que o intercepto foi subestimado diversas vezes, na mesma proporção em que o parâmetro de efeito da série de memória longa, $\hat{\beta}_2$, foi superestimado. Ou seja, vemos uma tendência de subestimações de $\hat{\beta}_0$ serem

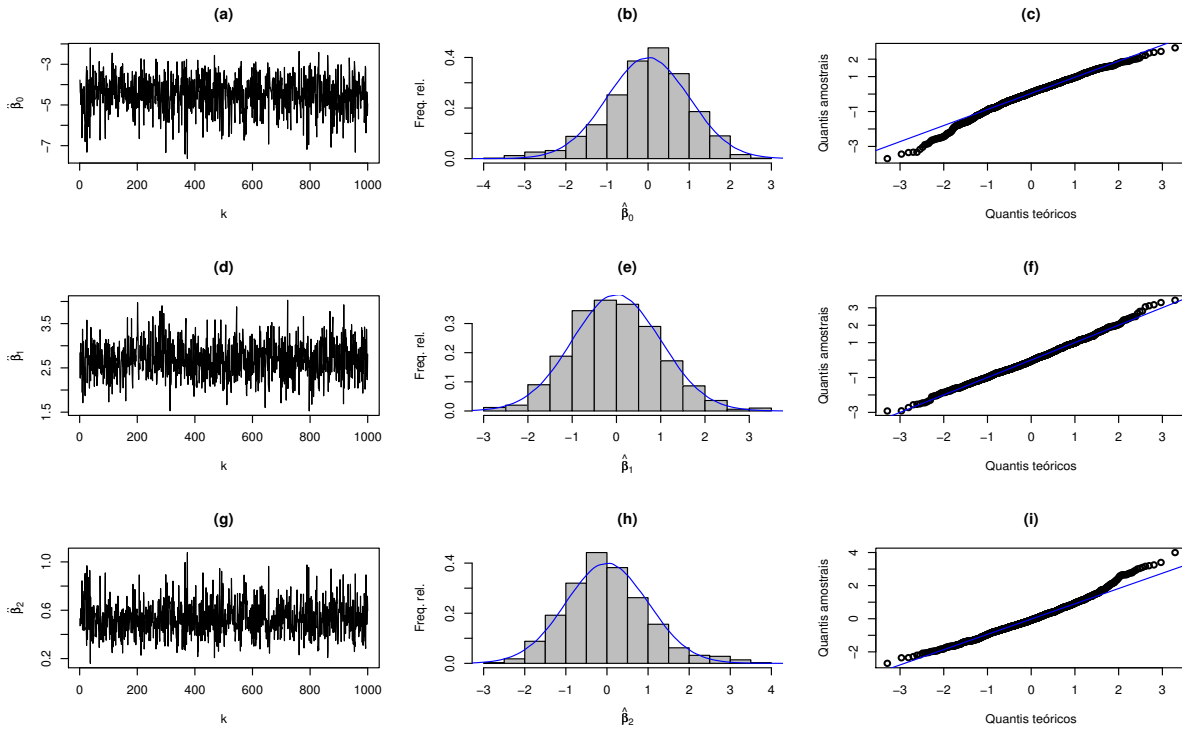


Figura 4.3: Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 200$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.

acompanhadas por superestimacões de $\hat{\beta}_2$, uma vez que tendem a ocorrer num mesmo instante da simulacão (Figuras 4.3 (a) e (g)).

Contudo, este comportamento desaparece à medida que o tamanho da série é aumentado, como podemos ver a partir da Figura 4.4, que apresenta os gráficos para a simulacão com $N = 1000$ e $d = 0,49$. Primeiramente, verifica-se pelas Figuras 4.4 (a), (d) e (g) que os valores gerados para as estimativas de cada parâmetro não apresentam *outliers*. Estimando agora as distribuições *marginais* dos parâmetros pelos histogramas de freqüência das estimativas obtidas (Figuras 4.4 (b), (e) e (h)), vemos que são levemente assimétricas. Porém, os gráficos quantil-a-quantil (Figuras 4.4 (c), (f) e (i)) sugerem que, assintoticamente, a distribuçã marginal de $\hat{\beta}_j$, $j = 0, 1, 2$, é Normal.

Concluimos portanto que, de uma forma geral, os resultados das simulacões são encorajadores, no sentido de que a teoria assintótica proposta por Kedem e Fokianos (2002) vale também quando as covariáveis possuem memória longa. Colocando de outra forma,

4.4. Modelagem das séries covariáveis

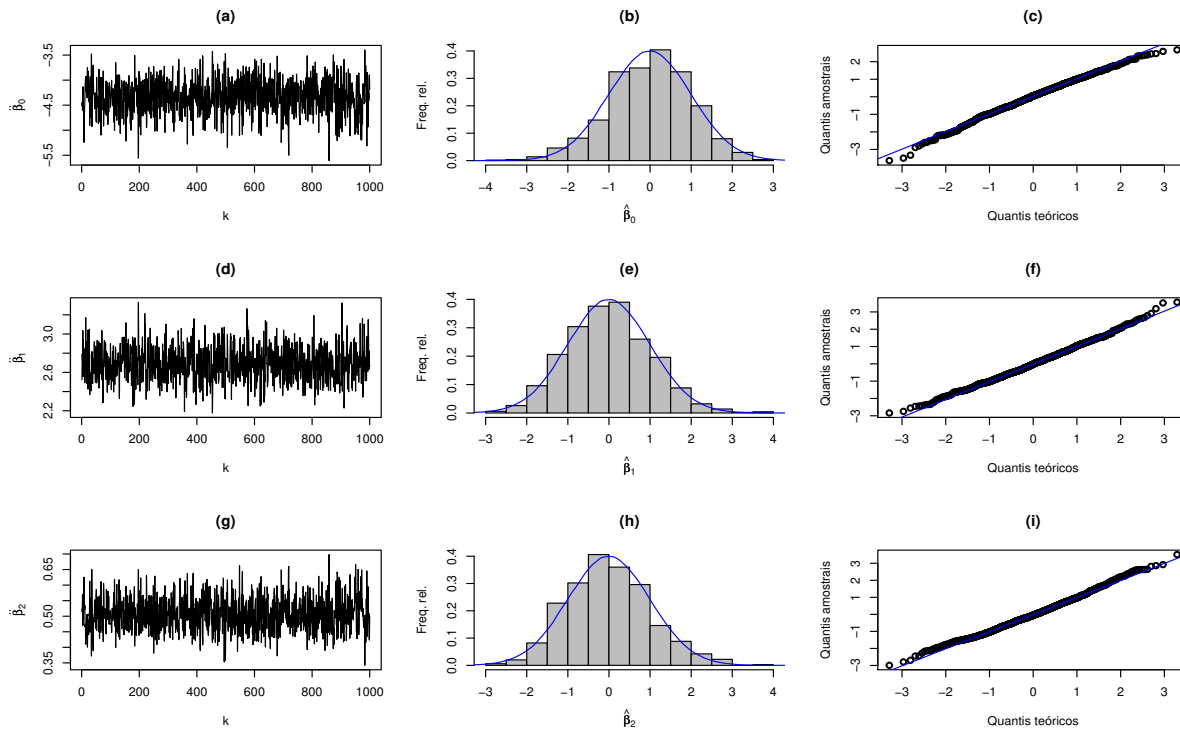
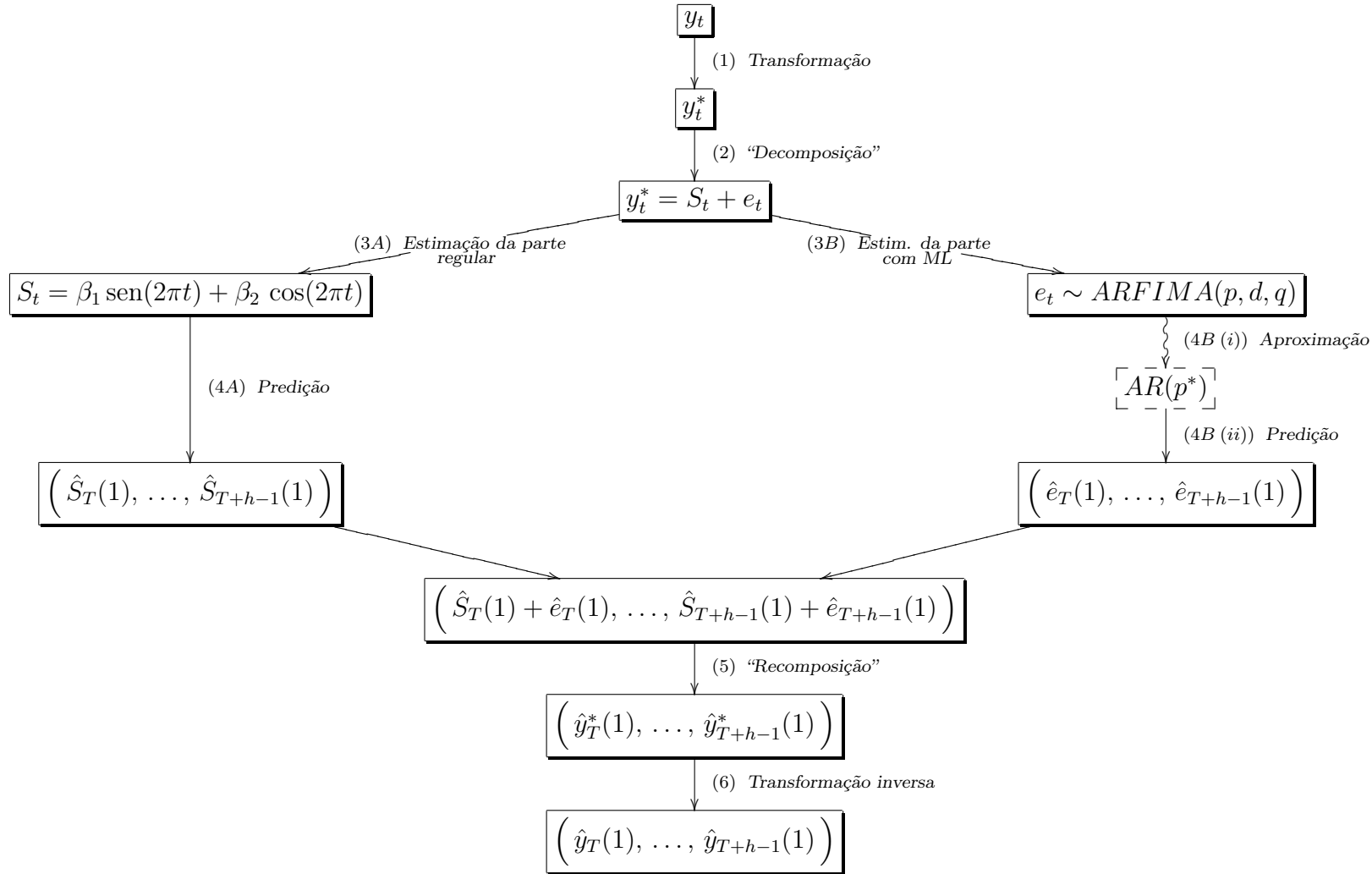


Figura 4.4: Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 1000$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.

os resultados das simulações dão algumas evidências de que (2.21) e (2.22) valem para o modelo logístico ajustado a séries temporais binárias, onde a série resposta e também as covariáveis possuem memória longa.

4.4 Modelagem das séries covariáveis

A modelagem das séries covariáveis das aplicações desta dissertação seguiu um fluxo de seis etapas, das quais as três primeiras integram o processo de estimação propriamente dito, e as outras três, o processo de predição. O diagrama abaixo ilustra o fluxograma do processo. Descrevemos, a seguir, cada uma das etapas.



4.4.1 Estimação

A fase de estimação é composta pelas seguintes etapas:

- *Etapa 1.* Verifica-se se há necessidade de usar alguma transformação, visando obter homocedasticidade para os dados. A série transformada é denotada por y_t^* .
- *Etapa 2.* A série transformada é “decomposta” em duas partes: parte regular (determinística, S_t) e parte irregular (aleatória, e_t), na forma $y_t^* = S_t + e_t$.
- *Etapa 3:* A parte regular S_t é composta pelas componentes de tendência e de sazonalidade, e é modelada por uma regressão linear com termos essencialmente si-

4.5. Avaliação da performance preditiva para uma classificação binária

nusoidais. A parte irregular e_t , que apresenta memória longa, é modelada por um ARFIMA.

4.4.2 Predição

Nesta dissertação, trabalhamos com predição **um passo à frente**, perfazendo-a h vezes, onde h é a extensão do período de predição, denotado por $[T + 1, \dots, T + h]$ (T denota aqui o último instante de observação). A fase de predição é composta pelas etapas:

- *Etapa 4.* Nesta etapa, a mais delicada do processo, faz-se a previsão de S_t e de e_t , a partir dos modelos encontrados na etapa anterior. Para a parte regular, faz-se simplesmente sua extrapolação para todo o período de predição, a partir do modelo de regressão linear ajustado, obtendo-se $(\hat{S}_T(1), \dots, \hat{S}_{T+h-1}(1))$. Já para a parte irregular, o modelo ARFIMA ajustado é aproximado por um AR(p) e através dele é feita a sua previsão de 1 passo à frente, da forma como explicada no Capítulo 3, também para todo o período de predição. Assim, obtém-se $(\hat{e}_T(1), \dots, \hat{e}_{T+h-1}(1))$.
- *Etapa 5.* Ambas as partes regular e irregular preditas são somadas, “compondo” a série predita \hat{y}_T^* .
- *Etapa 6:* Esta última etapa consiste apenas da aplicação, à série predita \hat{y}_T^* , da transformação inversa à utilizada na primeira etapa, caso alguma tenha sido utilizada.

Ao final do processo, temos os valores preditos $\{\hat{y}_t\} = (\hat{y}_T(1), \dots, \hat{y}_{T+h-1}(1))$ para a série covariável, em sua devida escala e amplitude de variação.

4.5 Avaliação da performance preditiva para uma classificação binária

Para avaliar e comparar as performances preditivas dos modelos, considerou-se algumas medidas de performance tradicionais em estudos de classificação dicotômica, derivadas da matriz de confusão (Mazucheli *et al.*, 2008). A **matriz de confusão** (ou matriz de classificação, ou matriz de erros) de um modelo, no contexto de previsão de poluição

do ar diária utilizado nesta dissertação, tem a forma da matriz apresentada na Tabela 4.3.

Tabela 4.3: Matriz de confusão.

Predita	Real		Total
	Alta (1)	Baixa (0)	
Alta (1)	a	b	$a + b$
Baixa (0)	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Os dados de entrada desta matriz têm os seguintes significados no contexto utilizado neste projeto:

- a representa a quantidade de predições corretas quando o dia é de poluição alta;
- b representa o número de dias de poluição alta, classificados incorretamente como dias de baixa poluição;
- c representa o número de dias de poluição baixa, classificados incorretamente como dias de alta poluição;
- d representa a quantidade de predições corretas quando o dia é de baixa poluição.

A partir da Tabela 4.3, definimos as seguintes medidas de performance preditiva adotadas neste trabalho:

- Acurácia (AC): é uma medida global da performance de predição do modelo, definida como a proporção do total de predições corretas,

$$AC = \frac{a + d}{a + b + c + d};$$

- Sensibilidade (S): é a capacidade de acerto de positivos, e expressa a probabilidade do modelo sob investigação fornecer um resultado positivo dado que o dia tem a característica de interesse, de alta poluição. Ou seja, a sensibilidade corresponde à proporção de dias de alta poluição que são corretamente preditos, e é dada por

$$S = \frac{a}{a + c};$$

4.5. Avaliação da performance preditiva para uma classificação binária

- Especificidade (E): é a capacidade de acerto de negativos, e expressa a probabilidade do modelo sob investigação fornecer um resultado negativo dado que o dia não tem a característica de interesse, ou seja, dado que o dia é de baixa poluição. Assim, a especificidade corresponde à proporção de dias de baixa poluição que são corretamente preditos, e é dada por

$$E = \frac{d}{b + d};$$

- Precisão de positivos (Prec_POS): é a proporção de dias de alta poluição corretamente preditos, dentre todos os dias preditos como positivos,

$$\text{Prec_POS} = \frac{a}{a + b};$$

- Precisão de negativos (Prec_NEG): é a proporção de dias de baixa poluição corretamente preditos, dentre todos os dias preditos como negativos,

$$\text{Prec_NEG} = \frac{c}{c + d};$$

- Taxa de verdadeiros positivos (VP): é a proporção de dias de alta poluição corretamente identificados,

$$VP = \frac{a}{a + b + c + d};$$

- Taxa de falsos positivos (FP): é a proporção de dias de poluição alta classificados incorretamente como dias de baixa poluição,

$$FP = \frac{b}{a + b + c + d};$$

- Taxa de falsos negativos (FN): é a proporção de dias de poluição baixa classificados incorretamente como dias de alta poluição,

$$FN = \frac{c}{a + b + c + d};$$

- Taxa de verdadeiros negativos (VN): é a proporção de dias de baixa poluição corretamente identificados,

$$VN = \frac{d}{a + b + c + d}.$$

Portanto, na modelagem da poluição pela abordagem binária, desejaremos obter um modelo com valores altos de acurácia, sensibilidade, especificidade, prec_POS, prec_NEG, VP e VN, e valores tão baixos quanto possível para as taxas de erro, FP e FN.

4.6 Aplicação em poluição do ar

Nesta seção, o modelo logístico é utilizado na análise de dados de poluição do ar da cidade de Santiago do Chile. Como será discutido adiante, uma vantagem na utilização do modelo logístico no contexto de poluição é o da simplicidade na interpretação das predições, dado que o interesse recai sobre a predição de uma das classes - digamos, concentração “alta” ou “baixa” do poluente - e não sobre o próprio valor da concentração. Pelas análises, verificou-se que o modelo obteve boa performance preditiva, de forma que o mesmo pode ser considerado para efeito de utilização por órgãos de vigilância ambiental, bem como de políticas públicas.

A poluição atmosférica é um dos maiores problemas ambientais que a capital chilena enfrenta. O problema se agrava durante o inverno devido a fenômenos climáticos como a inversão térmica, as chuvas e a considerável redução das massas de ar. Tudo isso, somado ao frio próprio dessa época do ano, tem grande impacto sobre a saúde da população, causando por exemplo o aumento das infecções respiratórias ou até mesmo aumentando o número de mortes por doenças respiratórias ou cardiovasculares (Ostro *et al.*, 1995). Dentre os poluentes mais nocivos à saúde, estão as *partículas inaláveis* menores que 10 μm e as menores que 2,5 μm . Por isto, é de grande importância a busca por modelos que sejam capazes de prever os níveis diários de concentração destes poluentes.

Foram analisados os dados da estação de Pudahuel, obtidos a partir do portal SINCA² - *Sistema de Información Nacional de Calidad del Aire*, que é vinculado à CONAMA - *Comisión Nacional del Medio Ambiente*. Os dados consistem das concentrações diárias de material particulado menor que 10 μm (PM10), dióxido de enxofre (SO₂) e dióxido de nitrogênio (NO₂), todas medidas em $\mu\text{g}/\text{m}^3$, no período de 01/Janeiro/2004 a 14/Setembro/2007.

Primeiramente, foi realizado um tratamento dos dados, para verificação da existência de valores aberrantes e preenchimento de valores faltantes (*missings*). As séries com maior quantidade de *missings* foram NO₂ e SO₂, com 14% e 11%, respectivamente. Os métodos utilizados no preenchimento dos valores faltantes foram o método de suavização exponencial e o método de Winter.

As séries de poluentes e suas funções de autocorrelação e de autocorrelação parcial estão apresentadas nas Figuras 4.5 e 4.6. Verifica-se, pela Figura 4.5, que os picos de

² Website - <http://sinca.conama.cl/>, acesso em 04/08/2009.

4.6. Aplicação em poluição do ar

poluição por PM10 ocorrem nos invernos e que os mesmos são acompanhados pelos picos de SO2 e de NO2, o que evidencia o forte relacionamento entre PM10 e as séries SO2 e NO2 ($\rho = 0,76$ e $0,74$, respectivamente). Além disso, o decaimento demorado para as FAC's e a persistência de valores altos para as FACP's de todas as séries indicam a existência de memória longa (Figura 4.6).

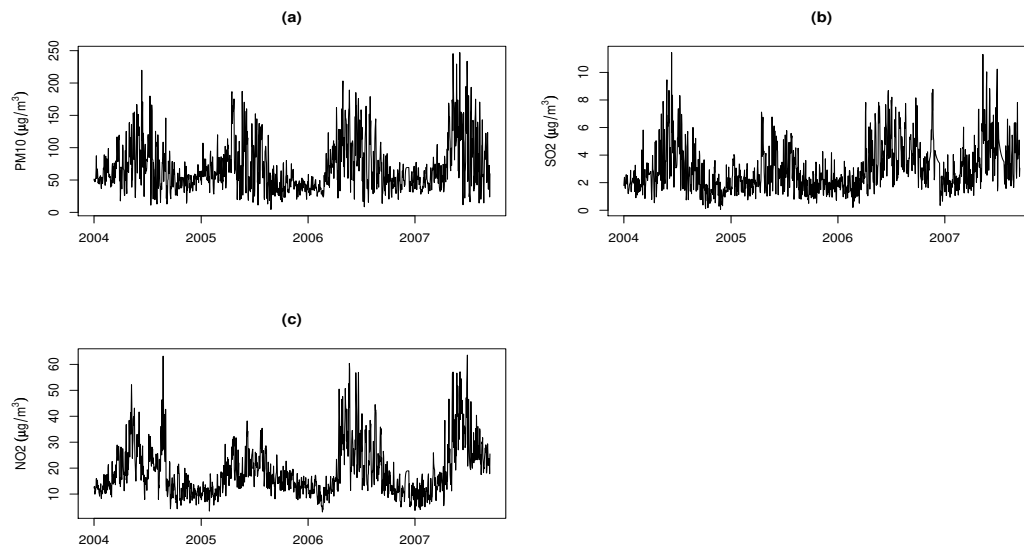


Figura 4.5: Séries de poluentes: (a) PM10, (b) SO2 e (c) NO2.

O objetivo desta análise é modelar a concentração de PM10 em termos dos demais poluentes, por meio do modelo de regressão logística. Para isso, é necessário especificar o limiar de referência para a categorização de PM10 em uma série binária. Diversos limiares têm sido adotados pelos órgãos estatais e internacionais de regulamentação da qualidade do ar, sendo alguns mais exigentes do que os outros. O padrão de qualidade da *U.S. National Ambient Air Quality Standards* (NAAQS), que é o padrão da Agência de Proteção Ambiental dos Estados Unidos (EPA), o da Organização Mundial da Saúde (meta de curto prazo), e o da maioria dos países europeus, considera o limiar de $150 \mu\text{g}/\text{m}^3$ (National Academy of Engineering and National Research Council, 2008). Esta é também a referência para países como México, Japão, China e Brasil. Um nível intermediário, de 120, é utilizado como padrão na Tailândia³. Mas, em alguns países desenvolvidos, como a

³*Pollution Control Department, Ministry of Natural Resources Environment, Thailand. Website - <http://www.pcd.go.th/indexEng.cfm>.*

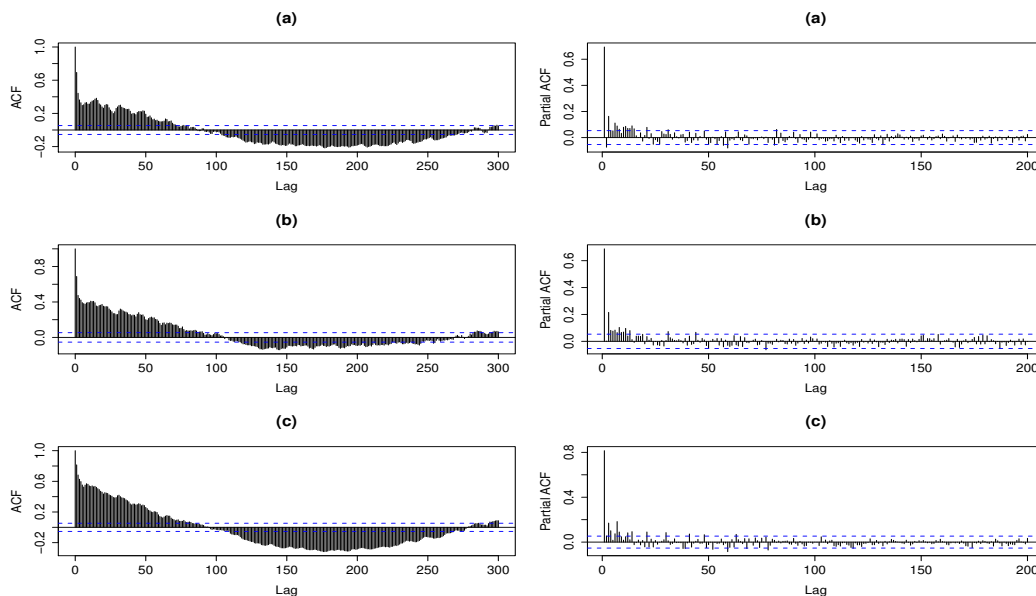


Figura 4.6: FAC (à esq.) e FACP (à dir.) dos poluentes: (a) PM10, (b) SO2 e (c) NO2.

Irlanda (The Environmental Protection Agency of Ireland⁴) e a Nova Zelândia (Ministry for the Environment, New Zealand⁵), o limite considerado como seguro para a saúde é bem mais rigoroso: $50 \mu\text{g}/\text{m}^3$, que é também a meta de longo prazo da OMS.

Para a análise da concentração de PM10 de Santiago de Chile foram adotados dois níveis de referência intermediários: 100 e $120 \mu\text{g}/\text{m}^3$. O nível de referência 100 é uma meta de médio prazo da OMS (National Academy of Engineering and National Research Council, 2008). As séries binárias resultantes da categorização - as quais designaremos por $Y_{[100]}$ e $Y_{[120]}$ - e suas FAC's apresentam-se na Figura 4.7, onde se observa a permanência da memória longa. As frequências da classe “1” (diga-se, “alta poluição”) equivalem a 19% e 12%, respectivamente, para $Y_{[100]}$ e $Y_{[120]}$.

As séries foram separadas entre dados de ajuste e dados de predição. O período de ajuste compreende os 1200 dias entre 01/Janeiro/2004 e 14/Abril/2007, e o de predição, os 153 dias a partir de 15/Abril/2007, abrangendo o dias mais críticos para a poluição do ar. O período de predição foi escolhido de forma que dificultasse ao máximo a predição por qualquer modelo, uma vez que nele há uma alta frequência de oscilação entre os níveis “0” e “1”, seja para a série do limiar 100 , seja para a do 120 (Figura 4.7 (a)). Considerando

⁴<http://www.epa.ie/whatwedo/monitoring/air/reports/pm10/>.

⁵<http://www.mfe.govt.nz/environmental-reporting/air/air-quality/pm10/>.

4.6. Aplicação em poluição do ar

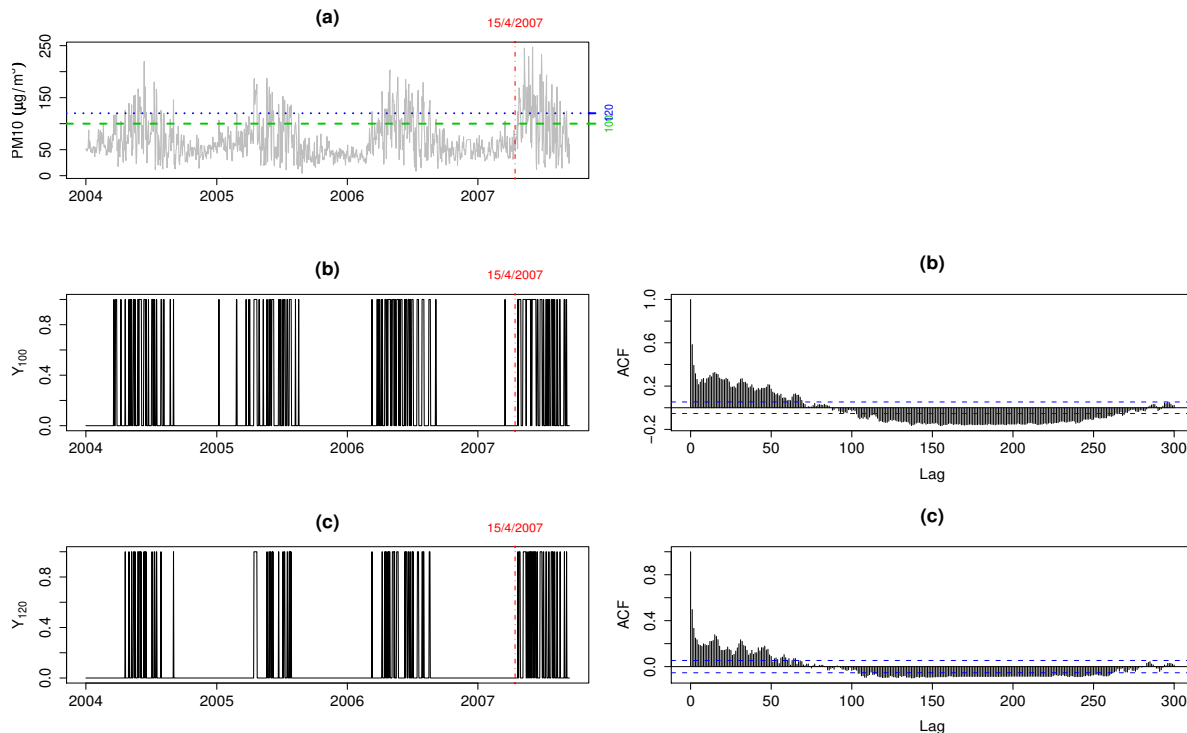


Figura 4.7: Categorização de PM10. Indicação dos níveis de corte e do período de predição (a), $Y_{[100]}$ e sua FAC (b) e $Y_{[120]}$ e sua FAC (c).

apenas os dados de predição, as frequências dos episódios de alta poluição são de 51% e 37%, para $Y_{[100]}$ e $Y_{[120]}$, respectivamente.

4.6.1 Análise da série para o limiar 100

Modelos logísticos da forma (4.2) foram testados variando-se a inclusão de covariáveis no preditor linear, sendo o melhor conseguido o modelo dado por

$$\log\left(\frac{\mu_t}{1 - \mu_t}\right) = \beta_0 + \beta_1 Y_{[100](t-1)} + \beta_2 SO2_t + \beta_3 SO2_{t-1} + \beta_4 NO2_t + \beta_5 I_{[prim-ver]}(t), \quad (4.6)$$

$t = 1, \dots, N$, onde $I_{[prim-ver]}$ é uma variável indicadora das estações primavera e verão. Vale observar que o efeito sazonal foi melhor captado por esta indicadora dicotômica do que pela variável nominal *Estação do ano*, que assume um dentre quatro valores possíveis.

A Tabela 4.4 apresenta as estimativas dos parâmetros. Verifica-se que todos os parâmetros são altamente significativos, e que a ocorrência de poluição por PM10 “alta”

no dia anterior, e o aumento na concentração dos poluentes SO2 e NO2 do mesmo dia, aumentam a chance de haver poluição alta. Além disso, a poluição por PM10 é mais alta nas estações outono e inverno, pelo fato de o coeficiente de $I_{[prim-ver]}$ ser negativo.

Tabela 4.4: Modelo logístico estimado para $Y_{[100]}$.

Termo	Estimativa	E.P.	Valor t	$Pr(> t)$
Intercepto	-7,69	0,64	-11,9	<0,001
$Y_{[100]}(t-1)$	2,58	0,41	6,2	<0,001
$SO2_t$	1,31	0,14	9,6	<0,001
$SO2_{t-1}$	-0,53	0,13	-4,1	<0,001
$NO2_t$	0,11	0,02	5,8	<0,001
$I_{[prim-ver]}(t)$	-1,42	0,50	-2,9	0,004

O modelo estimado não pode ser usado diretamente para predição, pois $Y_{[100]}$ depende dos valores contemporâneos de SO2 e NO2. Por esse motivo, é preciso primeiro modelar e prever as séries covariáveis. A seguir, ilustra-se a metodologia de modelagem apresentada na seção 4.4 através da modelagem da série SO2. Subseqüentemente, o mesmo processo é aplicado à série NO2. Esta etapa constitui um ponto de grande importância para a modelagem da poluição por PM10.

4.6.1.1 SO2

A Figura 4.5 (b) indica que a série SO2 varia de forma diferente para diferentes estações do ano, sugerindo que uma transformação dos dados pode ajudar em sua modelagem. Assim, os dados foram transformados segundo a transformação simples de Box-Cox

$$SO2_t^{(\lambda)} = \frac{SO2_t^\lambda - 1}{\lambda}, \quad (4.7)$$

onde $\lambda = 0,23$. O valor de λ foi estimado pelo método de máxima verossimilhança perfilada. A Figura 4.8 (b) evidencia que a transformação teve êxito na obtenção de homocedasticidade.

Para a parte regular de SO2 transformada, encontrou-se o modelo apresentado na Tabela 4.5. O ajuste e as componentes da decomposição em parte regular e irregular estão apresentados na Figura 4.9. Vemos que a componente irregular (Figura 4.9 (c)), resultante da subtração da parte regular (Figura 4.9 (b)) da série transformada, é aproximadamente

4.6. Aplicação em poluição do ar

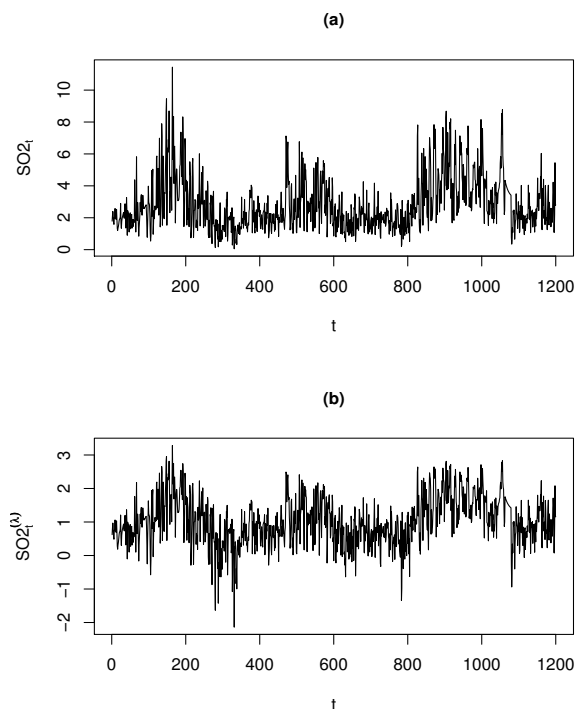


Figura 4.8: SO_2 (a) e $SO_2^{(\lambda)}$ (b).

homocedástica e dessazonalizada, e que ela apresenta memória longa (Figuras 4.9 (d) e (e)).

Tabela 4.5: Modelo de regressão ajustado à parte regular de $SO_2^{(\lambda)}$.

Termo	Estimativa	E.P.	$Pr(> t)$
Intercepto	1,06	0,02	<0,001
$\cos(2\pi t/365)$	-0,45	0,03	<0,001
$\cos(4\pi t/365)$	0,14	0,03	<0,001

O melhor modelo obtido para a componente irregular foi um ARFIMA(0, d , 1), com d estimado em 0,26 (0,01) e θ em 0,30 (0,03), onde os valores entre parênteses são os erros-padrões. Os resíduos deste ajuste não se comportam exatamente como um processo ruído branco, devido à existência de autocorrelações significativas para as defasagens a partir da 13^a, aproximadamente (Figuras 4.10 (b) e (d)). Contudo, como veremos mais adiante, isto não compromete os bons resultados obtidos em termos de previsão.

O modelo ARFIMA(0, d , 1) ajustado foi aproximado por um AR($p = 40$) para a predi-

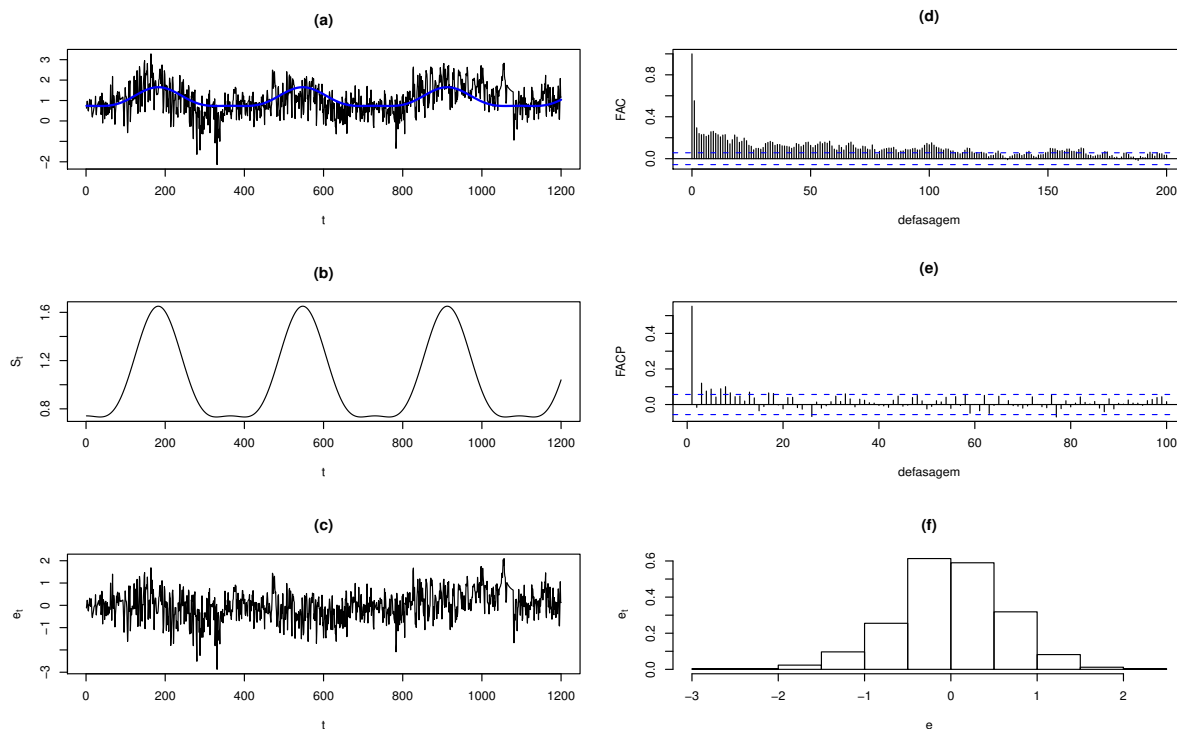


Figura 4.9: Decomposição nas componentes regular e irregular. (a) $SO_2^{(\lambda)}$ e a curva de ajuste à parte regular. (b) Destaque à componente regular. (c) Componente irregular. (d) FAC da parte irregular. (e) FACP da parte irregular. (f) Histograma da parte irregular.

ção da parte irregular, da forma como explicada na seção 3.4 do Capítulo 3. Somadas as componentes regular e irregular preditas, e aplicada a transformação inversa de Box-Cox, temos a predição desejada para os valores de SO_2 , disposta na Figura 4.11. Vemos a partir desta figura que a série predita não é capaz de acompanhar a amplitude dos valores extremos, o que é uma dificuldade natural em predição. Contudo, a modelagem de SO_2 foi eficiente no sentido de que, exceto no caso dos valores extremos, os valores preditos acompanham bem a oscilação dos valores reais, e assumem valores bem próximos destes⁶.

4.6.1.2 NO₂

Seguindo os mesmos passos da modelagem de SO_2 , a série NO_2 foi modelada e predita. A transformação utilizada foi a logarítmica, e o melhor modelo encontrado para a

⁶Observação: a “falha” que se observa logo após a primeira metade das observações do período de predição de SO_2 (Figura 4.11), é devida ao preenchimento inicial dos valores faltantes.

4.6. Aplicação em poluição do ar

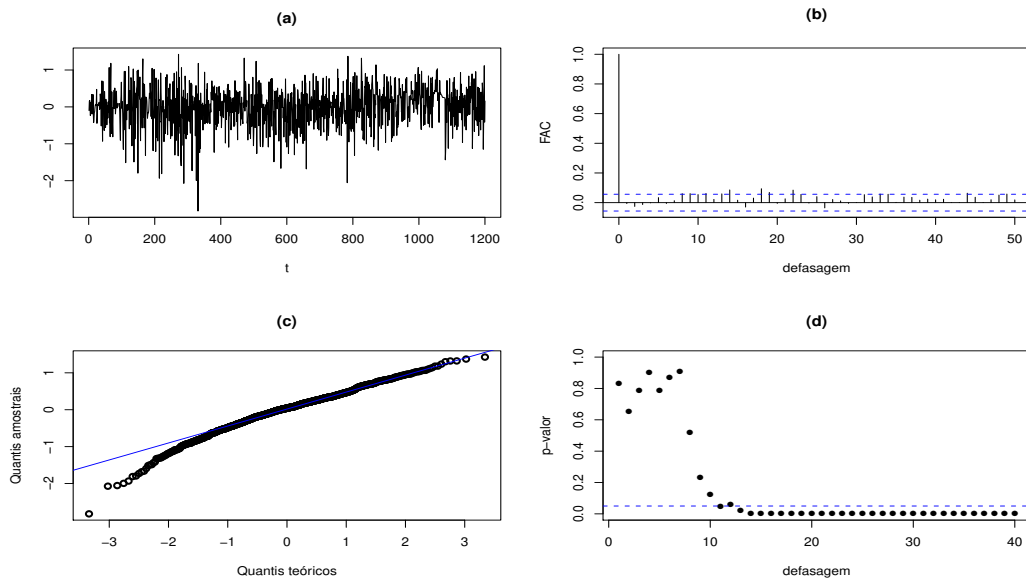


Figura 4.10: Resíduos do ajuste do $ARFIMA(0, d, 1)$ à componente irregular: (a) Série, (b) FAC, (c) Gráfico quantil-a-quantil com os da $N(0, 1)$ e (d) Níveis de significância para o teste de Box-Ljung.

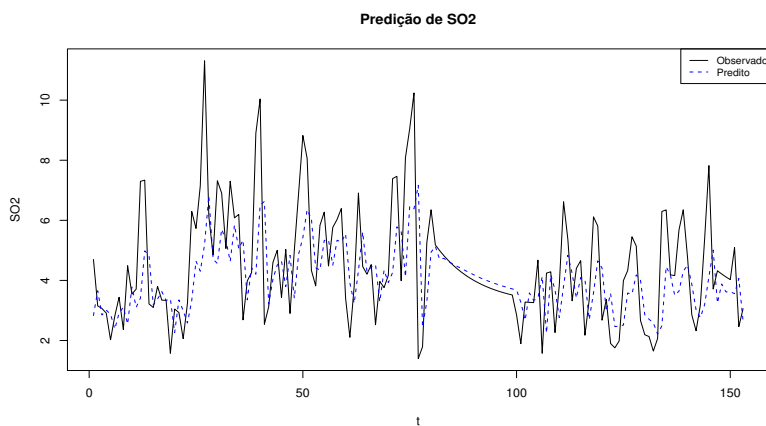


Figura 4.11: Séries observada e predita para SO_2 .

componente irregular foi novamente um $ARFIMA(0, d, 1)$, com as estimativas e respectivos erros-padrões dos parâmetros dados por $\hat{d} = 0,33$ (0,001) e $\hat{\theta} = 0,29$ (0,03). A série predita consta da Figura 4.12, a partir da qual pode-se ver que a predição de NO_2 foi também satisfatória.

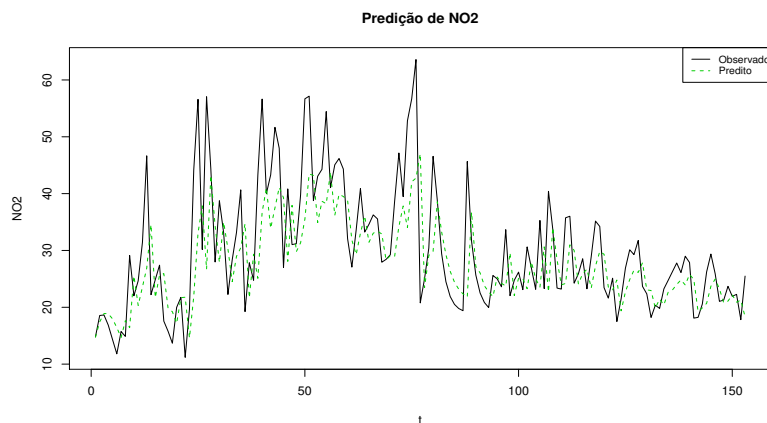


Figura 4.12: Séries observada e predita para NO2.

4.6.1.3 Predição de $Y_{[100]}$

Utilizando-se dos valores preditos de SO2 e NO2, fez-se a predição de **1 passo à frente** de $Y_{[100]}$, a partir do modelo ajustado (Tabela 4.4). Considerou-se valor predito “1” para $\hat{\mu}_t \geq 0,5$, e “0”, caso contrário. A matriz de confusão e as medidas de performance preditiva apresentam-se nas Tabelas 4.6 e 4.7 (b), respectivamente. Observa-se pelas medidas de performance (Tabela 4.7 (b)) que o modelo proporciona predições equilibradas, no sentido de que não há uma tendência a acertar mais os dias de baixa ou de alta poluição. Uma observação importante sobre a predição, disposta graficamente na Figura 4.13 (b), diz respeito ao alto grau de acerto em períodos de dias com episódios consecutivos de alta poluição, e também em dias de episódios consecutivos de baixa poluição. O mesmo não ocorre em períodos de dias onde há grande alternância entre poluição alta e poluição baixa. Vale lembrar, entretanto, que o período que foi destinado à predição é o mais crítico, i.e., um período dentro das estações outono e inverno onde há a maior oscilação anual da concentração de PM10.

Tabela 4.6: Matriz de erros para a predição de $Y_{[100]}$.

		Real		
		1	0	Total
Predito	1	58	17	75
	0	20	58	78
Total		78	75	153

4.6. Aplicação em poluição do ar

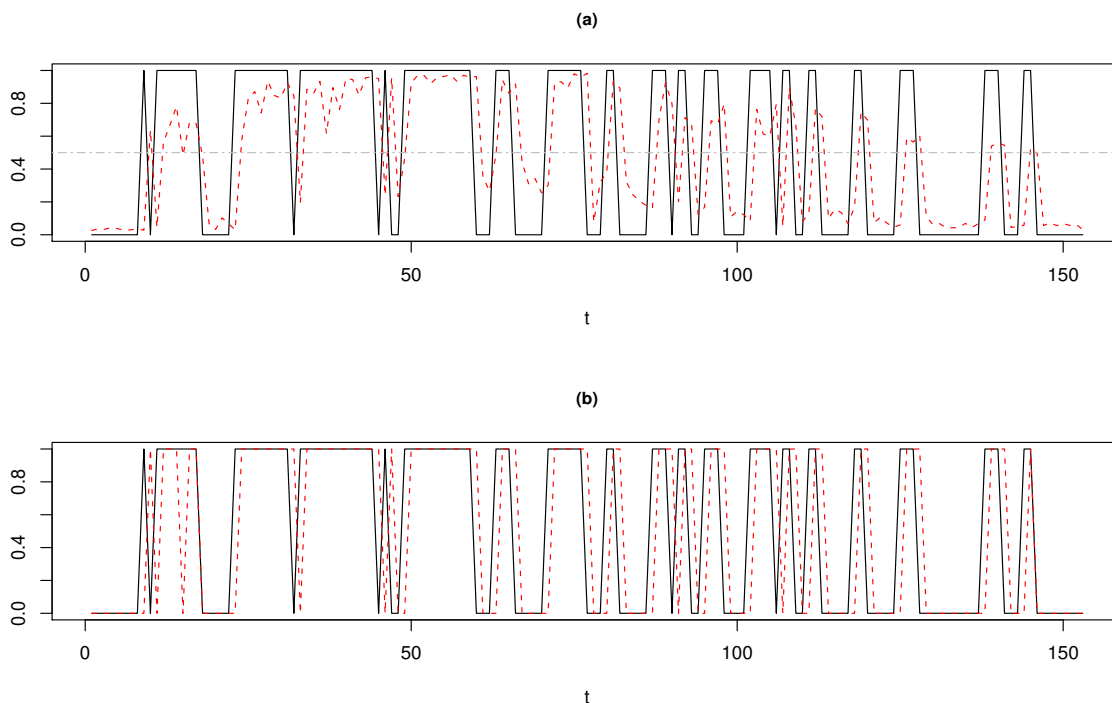


Figura 4.13: $Y_{[100]}$ observada (linha contínua) graficada junto a $\hat{\mu}_t$ (a) e $\hat{Y}_{[100]}$ (b) (linhas tracejadas).

Para permitir uma maior flexibilidade na utilização das previsões do modelo pelo usuário, pode-se alterar o nível que define se cada valor predito será “0” ou “1”, de acordo com o objetivo do usuário. Por exemplo, um modelo com foco voltado para a **saúde** da população seria aquele que tem menor taxa de falsos-negativos, uma vez que prever poluição baixa em um dia que será alta é o pior erro para a saúde. Para tal, ao invés de se considerar $\hat{y}_t = 1$ se $\hat{\mu}_t \geq 0,5$, pode-se estabelecer, por exemplo, $\hat{y}_t = 1$ se $\hat{\mu}_t \geq 0,2$. Esta medida força o modelo a prever a poluição como baixa somente quando a probabilidade predita de poluição alta ($\hat{\mu}_t$) for realmente “baixa” ($< 0,2$), predizendo poluição alta em todos os demais casos e prevenindo a população de sofrer danos à saúde.

Uma desvantagem de tal modelo seria o custo de sua implementação prática. O alerta à população sobre a poluição alta no dia seguinte implica em restrições à circulação de veículos e às atividades das indústrias. Estas medidas restritivas são ordenadas pelas autoridades competentes, para que a concentração dos poluentes não aumente e não chegue a níveis ainda mais críticos. Nos dias em que a previsão de poluição alta for errônea (falso-positivo), as empresas serão fechadas desnecessariamente, acarretando um grande

prejuízo para a economia da cidade. Assim, um modelo com foco oposto ao da saúde, seria o focado “na **economia**”, para o qual o objetivo é ter uma baixa taxa de falsos-positivos. Neste caso, pode-se considerar $\hat{y}_t = 1$ se $\hat{\mu}_t \geq 0,7$, por exemplo, de forma que o modelo irá prever $\hat{y}_t = 1$ *somente* se sua probabilidade predita for realmente “alta” ($> 0,7$). Portanto, dependendo dos objetivos do usuário, o modelo pode ser usado de uma dessas formas alternativas.

Tabela 4.7: Performances preditivas para a modelagem de $Y_{[100]}$.

	(a) Foco na saúde	(b) Modelo neutro	(c) Foco na economia
Nível para $\hat{\mu}_t$	0,2	0,5	0,7
Acurácia (%)	0,72	0,76	0,69
Sensibilidade (%)	0,83	0,74	0,54
Especificidade (%)	0,60	0,77	0,85
Precisão (Pos.) (%)	0,68	0,77	0,79
Precisão (Neg.) (%)	0,78	0,74	0,64
Verdadeiros Positivos (%)	0,42	0,38	0,27
Falsos Positivos (%)	0,20	0,11	0,07
Falsos Negativos (%)	0,08	0,13	0,24
Verdadeiros Negativos (%)	0,29	0,38	0,42

A Tabela 4.7 - colunas (a) e (c) - apresenta as medidas de performance preditiva obtidas para as duas abordagens mencionadas. Verifica-se que para o foco na saúde, a taxa de FN baixou para 8%, enquanto que para o foco na economia, a de FP baixou para 7%. Estas reduções seriam ainda maiores se tivessem sido utilizados níveis de corte mais extremos para $\hat{\mu}_t$, abaixo de 0,2 e acima de 0,7. Em contrapartida, as taxas de erro “opostas” a estas (FP em (a), e FN em (c)) aumentaram para 20 a 25%, aproximadamente. Portanto, ao optar por uma das abordagens (a) ou (c), o usuário terá a taxa de erro mais preocupante reduzida, porém estará sujeito ao aumento da outra. Comparando as três abordagens - saúde, neutra e economia - vê-se nitidamente o equilíbrio da performance preditiva obtida pelo modelo neutro: as taxas de erros específicos, e também as de acertos específicos, são próximas. Desta forma, os valores das precisões das predições de “alta” e de “baixa” poluição são também próximos e altos, o que acarreta uma alta acurácia de predição. Esta é menor se for utilizado um modelo com um dos focos propostos.

4.6. Aplicação em poluição do ar

4.6.1.4 Comparação com modelos alternativos

Sob o intuito de julgar a performance preditiva alcançada pelo modelo ajustado anteriormente, o qual designaremos por modelo “**M**”, comparou-se a mesma com as performances de outros três modelos: um modelo baseado somente em covariáveis defasadas (modelo **I**), o MLG Normal para PM10 contínua (modelo **II**), e o modelo “impraticável” (modelo **III**), que é o próprio modelo logístico apresentado anteriormente, mas com a utilização dos valores contemporâneos *reais* de SO2 e de NO2, em lugar dos preditos.

O primeiro modelo alternativo (I) tem a grande vantagem prática de não requerer a modelagem das séries covariáveis, tornando o processo de análise mais simples e curto. Desta forma, deseja-se fazer comparação para verificar o quanto se ganha, em termos de predição um passo à frente, ao modelar as covariáveis e permitir o uso de termos contemporâneos no modelo.

O modelo (II) seria o modelo “de referência”, dado que a variável de interesse, PM10, é originalmente contínua. Ajusta-se o MLG Normal com função de ligação identidade para PM10 (não-categorizada), e *após* os valores de PM10 serem preditos, eles são categorizados de acordo com o limiar 100. Para este caso, o objetivo com a comparação é verificar se há perda ou ganho de predição na utilização do modelo logístico para $Y_{[100]}$ em substituição ao modelo Normal para PM10 original. Naturalmente, esse modelo também requer a modelagem das séries covariáveis.

O terceiro modelo (III) é dito ser impraticável porque considera a utilização dos valores **reais** de $SO2_t$ e $NO2_t$ na predição de $Y_{[100]}$ para o próprio tempo t . Isto é impossível de se fazer, uma vez que no instante t dispomos apenas de informação até o tempo $t - 1$. Obviamente, pelo fato de se utilizar apenas de valores reais, este modelo terá uma performance preditiva superior a de qualquer outro. Assim, o objetivo para este caso é verificar o quão distante está a performance preditiva do modelo factível (M), da que seria obtida por este modelo “ideal” impraticável.

Vale ressaltar que os resultados apresentados para cada caso correspondem aos resultados dos **melhores** modelos obtidos em cada um, em termos de qualidade de ajuste, de BIC e de performance preditiva. O nível de corte para $\hat{\mu}_t$ utilizado na predição de todos os modelos foi 0,5. A forma de predição adotada é a de 1 passo à frente. Por brevidade de apresentação, restringimo-nos à apresentação somente das performances de predição; todavia, é importante destacar que todos os modelos se ajustaram bem aos

dados, obtendo-se sempre coeficientes altamente significativos.

Os resultados das predições de todos os modelos estão resumidos na Tabela 4.8. Verifica-se, em primeiro lugar, que as predições proporcionadas pelo modelo que seria ideal (III), mas é impraticável, são bem superiores às obtidas pelos modelos factíveis. Uma vez que a Acurácia é uma medida de performance global, um aumento de 10% ou mais é muito significativo. Além disso, comparando com os resultados da Tabela 4.7, observa-se que as taxas de erro de falsos-positivos e de falsos-negativos alcançadas pelo modelo impraticável são equilibradas e igualam-se a 7%, que é justamente o valor alcançado - individualmente - quando as abordagens de foco na saúde e foco na economia são adotadas.

Tabela 4.8: Comparação das performances preditivas de todos os modelos para $Y_{[100]}$.

Medida de perf. preditiva (%)	Impraticável (III)	Baseado somente no passado (I)	Normal, pós-categorizado (II)	Modelo final (M)
Acurácia	0,86	0,71	0,73	0,76
Sensibilidade	0,86	0,55	0,68	0,74
Especificidade	0,85	0,87	0,79	0,77
Precisão (Pos.)	0,86	0,81	0,77	0,77
Precisão (Neg.)	0,85	0,65	0,70	0,74
Verdadeiros Positivos	0,44	0,28	0,35	0,38
Falsos Positivos	0,07	0,07	0,10	0,11
Falsos Negativos	0,07	0,23	0,16	0,13
Verdadeiros Negativos	0,42	0,42	0,39	0,38

Apesar de a performance do modelo factível (M) ser inferior à do modelo irreal, em termos globais é superior à do modelo baseado somente em covariáveis defasadas, como se esperava, e é também superior à do modelo Normal para PM10. Além disso, todas as medidas de performance preditiva do modelo final são muito mais equilibradas do que as dos modelos baseado no passado e Normal, o que leva à conclusão de que o modelo conseguido foi bastante eficiente para a predição da poluição por PM10.

4.6. Aplicação em poluição do ar

4.6.2 Análise da série para o limiar 120

As análises e modelagens de $Y_{[100]}$ foram realizadas de forma similar para $Y_{[120]}$. O modelo logístico final encontrado é dado por

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_0 + \beta_1 Y_{[120](t-1)} + \beta_2 Y_{[120](t-2)} + \beta_3 SO2_t + \beta_4 SO2_{t-1} + \beta_5 NO2_t + \beta_6 \text{sen}\left(\frac{2\pi t}{365}\right) + \beta_7 \text{cos}\left(\frac{2\pi t}{365}\right), \quad (4.8)$$

$t = 1, \dots, N$. A diferença entre o modelo para $Y_{[100]}$ e o modelo (4.8) para $Y_{[120]}$ está na inclusão do termo $Y_{[120](t-2)}$, que foi significativo para este caso, e na explicação da sazonalidade da série. A versão dicotômica da estação do ano utilizada para o limiar 100 foi substituída, para o limiar 120, pelo par seno-cosseno, por este explicar melhor a sazonalidade de $Y_{[120]}$. As estimativas dos parâmetros são todas significativas (Tabela 4.9), e suas interpretações são idênticas às do modelo para $Y_{[100]}$, à exceção dos termos para a sazonalidade.

Tabela 4.9: Modelo logístico estimado para $Y_{[120]}$.

Termo	Estimativa	E.P.	Valor t	$Pr(> t)$
Intercepto	-10,29	1,00	-10,3	<0,001
$Y_{[120](t-1)}$	1,59	0,51	3,1	0,002
$Y_{[120](t-2)}$	0,95	0,42	2,3	0,02
$SO2_t$	1,35	0,17	8,2	<0,001
$SO2_{t-1}$	-0,51	0,16	-3,3	0,001
$NO2_t$	0,09	0,02	4,5	<0,001
$\text{sen}(2\pi t/365)$	1,35	0,35	3,9	<0,001
$\text{cos}(2\pi t/365)$	-1,78	0,58	-3,1	0,002

Os valores preditos de SO2 e NO2 foram utilizados no modelo para a predição de $Y_{[120]}$. A visualização da série predita pode ser feita pela Figura 4.14 (b). Verifica-se que a predição foi inferior à obtida para $Y_{[100]}$, provavelmente devido ao fato de a série $Y_{[120]}$ ser mais difícil de prever. O uso do limiar 120 fez com que a série binária oscilasse com maior frequência, dificultando a predição qualquer que seja o modelo utilizado. Mesmo assim, observa-se na figura que os dois períodos mais longos de poluição alta por dias consecutivos teve predição acurada em sua maior parte. O pior resultado do modelo foi o

fato de não ter predito nenhum dos episódios de poluição alta do último terço do período de predição.

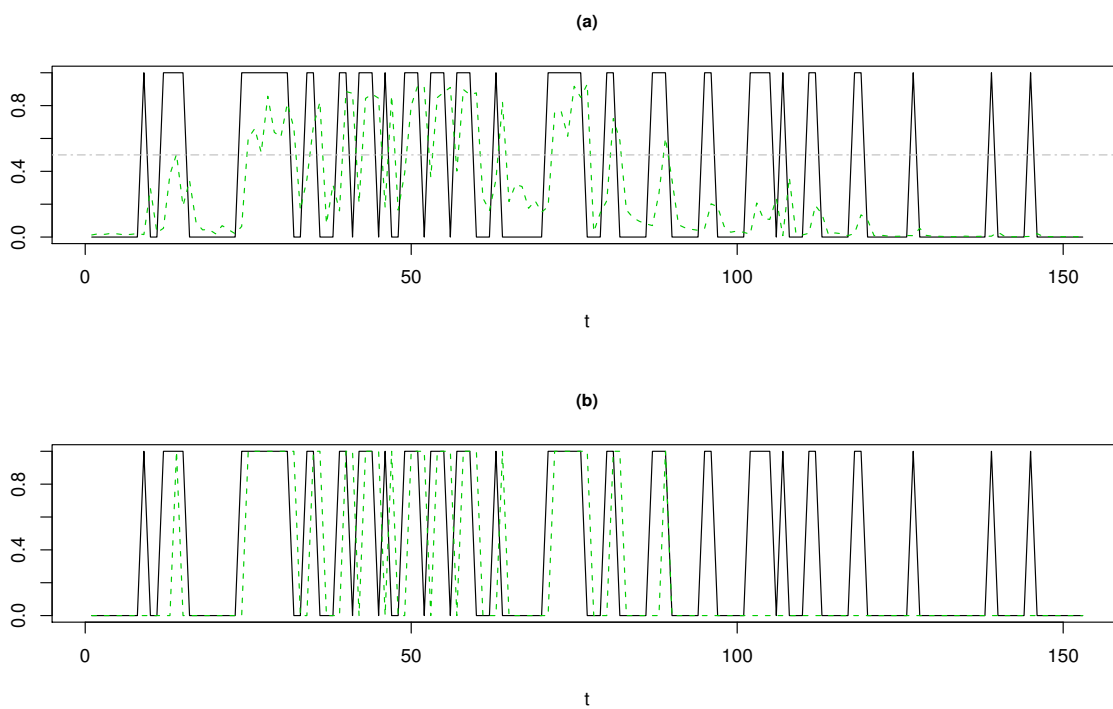


Figura 4.14: $Y_{[120]}$ observada (linha contínua) graficada junto a $\hat{\mu}_t$ (a) e $\hat{Y}_{[120]}$ (b) (linhas tracejadas).

As Tabelas 4.10 e 4.11 (b) apresentam a matriz de confusão e as medidas de performance preditiva do modelo. A acurácia obtida, de 73%, confirma o desempenho um pouco inferior ao do modelo para $Y_{[100]}$ (76%). Além disso, houve um desequilíbrio entre as medidas de sensibilidade e de especificidade, e entre as taxas de erro FP e FN, sendo a primeira bem menor. Esta assimetria na predição é provavelmente decorrente do desbalanceamento de classes - $\{0, 1\}$ ocorre com as frequências $\{63\%, 37\%\}$ - originado pela utilização de um limiar mais extremo (120) do que o de 100 para a geração da série binária para PM10. No caso de $Y_{[100]}$, estas mesmas medidas se mostraram equilibradas porque havia um perfeito balanceamento entre as classes: '1' e '0' ocorrem com as frequências de 51% e 49%, respectivamente.

Com relação à modificação do modelo para os focos na saúde e na economia, observa-se que o modelo tal como está pode ser diretamente adotado para o foco na economia, devido à baixa taxa de FP (Tabela 4.11 (b)). Para verificar a magnitude da redução que pode

4.6. Aplicação em poluição do ar

ser obtida para esta taxa, as previsões para o foco na economia foram realizadas com um nível de corte mais extremo para $\hat{\mu}_t$: 0,8. Analogamente, considerou-se um nível também extremo, de 0,1, para a previsão com o foco na saúde. Observa-se que o ganho em termos da redução dos falsos-positivos, para o foco na economia, é muito baixo (Tabela 4.11 (c)). Para o foco na saúde, porém, a redução da taxa de FN é bastante significativa: de 20 para 8% (Tabela 4.11 (a)).

Tabela 4.10: Matriz de erros para a previsão de $Y_{[120]}$.

		Real		
		1	0	Total
Preditio	1	26	11	37
	0	30	86	116
Total		56	97	153

Tabela 4.11: Performances preditivas para a modelagem de $Y_{[120]}$.

Nível para $\hat{\mu}_t$	(a) Foco na saúde	(b) Modelo neutro	(c) Foco na economia
	0,1	0,5	0,8
Acurácia (%)	0,68	0,73	0,66
Sensibilidade (%)	0,77	0,46	0,20
Especificidade (%)	0,63	0,89	0,93
Precisão (Pos.) (%)	0,54	0,70	0,61
Precisão (Neg.) (%)	0,82	0,74	0,67
Verdadeiros Positivos (%)	0,28	0,17	0,07
Falsos Positivos (%)	0,24	0,07	0,05
Falsos Negativos (%)	0,08	0,20	0,29
Verdadeiros Negativos (%)	0,40	0,56	0,59

Os modelos alternativos considerados na análise de $Y_{[100]}$ foram também utilizados para a modelagem de $Y_{[120]}$, e as performances das previsões estão dispostas na Tabela 4.12. Diferentemente dos resultados observados para $Y_{[100]}$, verifica-se que a performance preditiva do modelo final (M) para $Y_{[120]}$ foi similar, e não superior, à dos modelos alternativos baseado no passado e Normal. As configurações das matrizes de erros para estes modelos são muito similares, assim, todas as medidas de performance são estatisticamente equivalentes. Por último, comparando as performances do modelo final e do impraticável, nota-se que há uma perda de 11%, em termos de performance preditiva geral.

Tabela 4.12: Comparação das performances preditivas de todos os modelos para $Y_{[120]}$.

Medida de perf. preditiva (%)	Impraticável (III)	Baseado somente no passado (I)	Normal, pós-categorizado (II)	Modelo final (M)
Acurácia	0,84	0,73	0,73	0,73
Sensibilidade	0,66	0,45	0,48	0,46
Especificidade	0,95	0,89	0,88	0,89
Precisão (Pos.)	0,88	0,69	0,69	0,70
Precisão (Neg.)	0,83	0,74	0,75	0,74
Verdadeiros Positivos	0,24	0,16	0,18	0,17
Falsos Positivos	0,03	0,07	0,08	0,07
Falsos Negativos	0,12	0,20	0,19	0,20
Verdadeiros Negativos	0,60	0,56	0,56	0,56

4.6.3 Conclusões da aplicação

A partir dos resultados observados na aplicação, é possível concluir que:

- O modelo de regressão logística constitui uma ótima opção para a modelagem e predição 1 passo à frente de séries temporais de poluição com memória longa. Para os dados de poluição do ar da cidade de Santiago do Chile, o modelo logístico para a versão binária da concentração de PM10 obteve performance preditiva igual ou superior à do modelo Normal para PM10 original (contínua).
- A escolha do limiar de referência para a categorização da série resposta é de grande importância, pois tem impacto sobre as predições geradas pelo modelo. Para os dados chilenos, o uso do limiar $100 \mu\text{g}/\text{m}^3$ para a discretização de PM10 proporcionou predições mais acuradas e equilibradas do que o uso do limiar 120.
- O modelo logístico, adotado no contexto de poluição, tem a vantagem da simplicidade na interpretação das predições, dado que o interesse recai sobre a predição de uma das classes - concentração “alta” ou “baixa” do poluente.
- De acordo com a finalidade de utilização, como as com o foco na saúde da população e com o foco na economia da cidade, o usuário pode “calibrar” a regra de classificação adotada na predição do modelo, visando à redução da taxa de erro mais preocupante.

Modelo de Chances Proporcionais para Séries Categóricas Ordiniais

Diferentes metodologias têm sido desenvolvidas para a análise de séries temporais categóricas. Dentro da abordagem por modelos de regressão, Fahrmeir e Kaufmann (1987) foram os primeiros a estabelecer uma base sólida para a teoria de estimação por máxima verossimilhança. Outras abordagens, como as baseadas em modelos de Markov (Zeger e Qaqish, 1988), em modelos de espaço-de-estados (Fahrmeir, 1992), e a estruturada no domínio da frequência (Stoffer, Tyler e Mcdougall, 1993), também foram propostas e têm sido adotadas em diversas aplicações. Para a modelagem de séries categóricas do tipo ordinal, o modelo de chances proporcionais constitui uma ótima opção, por ser simples e de fácil interpretação.

Apresentaremos neste capítulo o modelo de chances proporcionais para séries temporais categóricas ordiniais, como dado por Kedem e Fokianos (2002), sua estimação via máxima verossimilhança parcial, e estudos de simulação para verificar se as propriedades de consistência e de normalidade assintótica do estimador de máxima verossimilhança parcial valem para a modelagem de séries com memória longa. Adicionalmente, definiremos algumas medidas de performance preditiva para o caso de classificação em três categorias. A utilização do modelo será então ilustrada através de uma aplicação a dados de poluição do ar.

As simulações e as análises da aplicação do modelo foram realizadas no *software* livre R, versão 2.9.0 (R Development Core Team, 2009). Para os ajustes do modelo de chances

proporcionais, foi utilizada a função `{vglm}` do pacote `{VGAM}` (Yee, 2008).

5.1 Contextualização e notação

Há duas representações comuns para séries temporais categóricas, a simples e a vetorial. A adoção de uma ou outra depende do objetivo da utilização e do contexto. A representação “simples” é bastante útil para a visualização gráfica da série. Nela, a cadeia $\{Y_t\}$, $t = 1, \dots, N$ denota a série temporal categórica, onde a variável Y_t assume, em cada momento t , um dentre os possíveis valores $1, 2, \dots, m - 1, m$, representando cada valor uma categoria de resposta e levando-se em consideração a ordem lógica entre elas. Contudo, por motivos de especificação e representação das variáveis aleatórias em algoritmos de estimação, é conveniente adotar a representação *vetorial*.

Uma variável categórica ordinal temporal com m categorias é geralmente representada por um *vetor* em cada momento de observação t , isto é, uma seqüência de $m - 1$ valores “0” e um valor “1”, indicando este a categoria de resposta que foi observada no tempo t . Assim, representaremos uma variável ordinal temporal com m categorias por um vetor $\mathbf{Y}_t = (Y_{t1}, \dots, Y_{tq})'$ de tamanho $q = m - 1$, com elementos

$$Y_{tj} = \begin{cases} 1, & \text{se a } j\text{-ésima categoria é observada no tempo } t \\ 0, & \text{caso contrário} \end{cases}$$

para $t = 1, \dots, N$ e $j = 1, \dots, q$. Como exemplo, suponha-se o estudo do sono de um bebê recém-nascido, que tenha sido mensurado junto à temperatura e à frequência cardíaca do mesmo. Suponha as categorias

- 1: acordado,
- 2: sono inativo (repouso),
- 3: sono indeterminado,
- 4: sono ativo, profundo.

Temos aqui $m = 4$, $q = 3$ e os vetores

$$\begin{aligned} \mathbf{Y}_t &= (0, 0, 0)' \\ \mathbf{Y}_t &= (1, 0, 0)' \\ \mathbf{Y}_t &= (0, 1, 0)' \\ \mathbf{Y}_t &= (0, 0, 1)' \end{aligned} \tag{5.1}$$

5.1. Contextualização e notação

para representar as categorias “acordado”, “repouso”, “sono indeterminado” e “sono profundo”, respectivamente. Nota-se que a “ausência” de representação ($\mathbf{Y}_t = (0, 0, 0)'$) deverá sempre denotar a primeira ou a última categoria, dado que a ordenação natural das categorias deve ser respeitada.

Na análise de dados ordinais, o interesse recai sobre a estimação das probabilidades condicionais da variável resposta assumir cada um dos valores-categoria, isto é, sobre $\boldsymbol{\pi}_t = (\pi_{t1}, \dots, \pi_{tq})'$ onde

$$\pi_{tj} = E[Y_{tj} | \mathcal{F}_{t-1}] = P(Y_{tj} = 1 | \mathcal{F}_{t-1}), \quad j = 1, \dots, q, \quad (5.2)$$

para todos os tempos $t = 1, \dots, N$. É comum nomear as probabilidades π_{tj} por “probabilidades de transição”, dado que cada uma delas denota a probabilidade da variável de interesse migrar, no momento t , de algum estado anterior para o estado (categoria) j . A σ -álgebra denotada por \mathcal{F}_{t-1} corresponde a toda informação do passado, da variável de interesse e das covariáveis explanatórias, se houverem, conhecida até o tempo t .

A representação de $q = m - 1$ categorias de resposta é feita, como mencionado, pelos q vetores não-nulos como exemplificados em (5.1); a categoria restante, tomada como “referência”, é expressada por

$$Y_{tm} = 1 - \sum_{j=1}^q Y_{tj},$$

e possui probabilidade condicional de ocorrência

$$\pi_{tm} = 1 - \sum_{j=1}^q \pi_{tj}.$$

Com relação ao processo de covariáveis $\{\mathbf{Z}_{t-1}\}$, $t = 1, \dots, N$, pelo fato de lidarmos com uma resposta *multivariada* no caso ordinal, precisaremos ter uma *matriz* $p \times q$ para sua representação, ou seja, a cada resposta Y_{tj} corresponderá um vetor de tamanho p de covariáveis aleatórias dependentes do tempo, o qual irá compôr a j -ésima coluna de \mathbf{Z}_{t-1} . O processo de covariáveis poderá incluir valores defasados da série resposta e/ou de qualquer outro processo auxiliar.

A formatação do MLG para o caso ordinal se dá da seguinte forma: assume-se que o vetor de probabilidades de transição, que (a partir de (5.2)) é a esperança condicional do vetor resposta dado o passado, é explicado pelo processo de covariáveis através da ligação

$$\boldsymbol{\pi}_t = \mathbf{h}(\mathbf{Z}'_{t-1}\boldsymbol{\beta}), \quad (5.3)$$

sendo $\boldsymbol{\beta}$ um vetor p -dimensional de parâmetros do modelo, e \mathbf{h} - que é a função de ligação *inversa* - uma função *multivariada* que faz o mapeamento da reta real para o intervalo unitário. A função de ligação inversa \mathbf{h} , aqui, desempenha o mesmo papel que nos outros casos de MLG; ela associa o valor que o preditor linear assume, definido sobre o intervalo real \Re , ao valor da resposta esperada, que neste caso é uma probabilidade e portanto deve estar contida no intervalo unitário $(0, 1)$. Ou seja, a função $\mathbf{h} : \Re^q \rightarrow \Re^q$ deve necessariamente satisfazer o mapeamento bijetor entre um subconjunto $H \subseteq \Re^q$ e o conjunto definido por $\{(w_1, \dots, w_q)' : w_j > 0, j = 1, \dots, q, \sum_{j=1}^q w_j < 1\}$, onde w_j denota a probabilidade da variável resposta assumir a categoria j , em função do processo de covariáveis.

O modelo linear generalizado *multivariado* (5.3) é melhor visualizado na forma

$$\boldsymbol{\pi}_t = \begin{pmatrix} \pi_{t1}(\boldsymbol{\beta}) \\ \pi_{t2}(\boldsymbol{\beta}) \\ \dots \\ \pi_{tq}(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} h_1(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \\ h_2(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \\ \dots \\ h_q(\mathbf{Z}'_{t-1}\boldsymbol{\beta}) \end{pmatrix} = \mathbf{h}(\mathbf{Z}'_{t-1}\boldsymbol{\beta}), \quad (5.4)$$

onde se evidencia as componentes $\{h_1, \dots, h_q\}$ de \mathbf{h} . Kedem e Fokianos (2002) citam vários trabalhos onde o modelo (5.4) foi considerado. Um caso especial de (5.4) é o binário - considerado no capítulo anterior - obtido para $m = 2$ (e portanto $q = 1$), de forma que o modelo se reduz para

$$\pi_t(\boldsymbol{\beta}) = P(Y_t = 1 | \mathcal{F}_{t-1}) = h(\boldsymbol{\beta}'\mathbf{Z}_{t-1}),$$

onde $h : \Re \rightarrow (0, 1)$ é uma função de ligação inversa escalar e monótona (geralmente, uma f.d.a.).

O modelo específico a ser ajustado aos dados é determinado pela função \mathbf{h} . Se a função adotada for a função de distribuição da densidade logística padrão (média 0 e variância $\frac{\pi^2}{3}$), que é a função de ligação inversa correspondente à função de ligação logito, teremos o *modelo logito cumulativo*, mais conhecido pelo nome *modelo de chances proporcionais*.

5.2 O modelo de chances proporcionais

Quando se lida com dados categóricos ordinais, é proveitoso modelar não as probabilidades condicionais de cada categoria (5.2), mas sim as probabilidades condicionais

5.2. O modelo de chances proporcionais

cumulativas

$$P(Y_t \leq j | \mathcal{F}_{t-1}) = \pi_{t1} + \dots + \pi_{tj},$$

para $j = 1, \dots, m$, dado que tem sentido lógico agrupar probabilidades de categorias adjacentes. Sobre elas, são definidos os *logitos cumulativos*

$$\begin{aligned} \text{logito}[P(Y_t \leq j | \mathcal{F}_{t-1})] &\equiv \log\left(\frac{P(Y_t \leq j | \mathcal{F}_{t-1})}{1 - P(Y_t \leq j | \mathcal{F}_{t-1})}\right) \\ &= \log\frac{\pi_{t1} + \dots + \pi_{tj}}{\pi_{t(j+1)} + \dots + \pi_{tm}}, \end{aligned}$$

para $j = 1, \dots, q$. Obviamente, temos apenas $q = m - 1$ logitos cumulativos, pois para a última categoria m , o denominador se anularia.

De forma bastante similar à do caso binário, e utilizando-se da função logito cumulativo acima referida, define-se o **modelo de chances proporcionais** pelas equações

$$\text{logito}[P(Y_t \leq j | \mathcal{F}_{t-1})] = \theta_j + \boldsymbol{\gamma}' \mathbf{z}_{t-1}, \quad t = 1, \dots, N, \quad (5.5)$$

para $j = 1, \dots, q$, onde $\{\theta_1, \dots, \theta_q\}$ são termos de intercepto e $\boldsymbol{\gamma}$ é um vetor d -dimensional de parâmetros de efeito das covariáveis. Desta forma, temos que o vetor $\boldsymbol{\beta}$ de parâmetros do modelo, de dimensão $p = q + d$, fica dado por $\boldsymbol{\beta} = (\theta_1, \dots, \theta_q, \boldsymbol{\gamma}')'$. O vetor de valores assumidos pelo processo de covariáveis no momento t é aqui representado por \mathbf{z}_{t-1} , em letra minúscula, para distingui-lo de \mathbf{Z}_{t-1} , que agora denota a *matriz* $(q + d) \times q$

$$\mathbf{Z}_{t-1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \mathbf{z}_{t-1} & \mathbf{z}_{t-1} & \dots & \mathbf{z}_{t-1} \end{bmatrix}.$$

No modelo (5.5), cada logito cumulativo tem seu próprio intercepto θ_j , portanto temos q termos de intercepto. Além disso, os $\{\theta_j\}$ são crescentes em j , já que as probabilidades cumulativas $P(Y_t \leq j | \mathcal{F}_{t-1})$ aumentam com o aumento de j (para \mathbf{z}_{t-1} fixado) e os logitos são funções crescentes delas. É importante destacar que este modelo tem os mesmos efeitos $\boldsymbol{\gamma}$ para cada logito.

A interpretação dos parâmetros de efeito ($\boldsymbol{\gamma}$) do modelo de chances proporcionais não é feita da forma usual. Geralmente, num modelo de regressão, se um determinado

parâmetro de efeito β_l é positivo, então o relacionamento entre sua covariável (X_l) e a resposta (Y) será positivo, isto é, X_l e Y serão diretamente proporcionais. Neste caso, a interpretação de β_l é a usual: o aumento no valor de X_l está associado a um aumento no valor de Y . No caso do modelo de chances proporcionais, entretanto, a interpretação é a contrária. Se tivermos uma única covariável - digamos, Z_t - e seu parâmetro γ for *positivo*, então Z_t e a resposta Y_t serão inversamente proporcionais, pois um aumento em Z_t estará associado a um aumento na chance de Y_t assumir seus valores *menores*. Por este motivo, alguns autores consideram uma parametrização diferente para o modelo de chances proporcionais (5.5), substituindo γ por $-\gamma$ e utilizando-o na forma

$$\text{logito}[P(Y_t \leq j | \mathcal{F}_{t-1})] = \theta_j - \gamma' z_{t-1}, \quad t = 1, \dots, N,$$

para $j = 1, \dots, q$. Outra abordagem alternativa é a utilização dos logitos de $P(Y_t \geq j | \mathcal{F}_{t-1})$, $j = 2, \dots, m$, ao invés dos logitos cumulativos (que são definidos sobre as probabilidades $P(Y_t \leq j | \mathcal{F}_{t-1})$), e assim a interpretação dos parâmetros de efeito será a usual.

A justificativa para o nome do modelo (5.5) decorre do fato abaixo explicado. Sejam $z_{t-1}^{[1]}$ e $z_{t-1}^{[2]}$ dois diferentes padrões assumidos pelo vetor de covariáveis, isto é, dois valores distintos que o processo pode assumir. Se, por exemplo, num contexto de estudo clínico temos uma única covariável no modelo, X_t , definida como a “exposição a um determinado fator de risco” na forma

$$z_{t-1} = X_t = \begin{cases} 1, & \text{se exposto} \\ 0, & \text{se não exposto} \end{cases},$$

então $z_{t-1}^{[1]} \equiv 1$ e $z_{t-1}^{[2]} \equiv 0$ são dois diferentes valores assumidos pelo processo covariável, ou, diferentes padrões de covariáveis. Se for tomada a diferença dos logitos da probabilidade da resposta assumir uma categoria inferior ou igual a j , condicionada aos respectivos padrões de covariáveis, ver-se-á que será proporcional à diferença entre os padrões. Em outras palavras, o logaritmo da razão de chances dos eventos acima será proporcional à

5.2. O modelo de chances proporcionais

diferença dos padrões, i.e. (usando (5.5)),

$$\begin{aligned}
 \text{logito}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[1]})] - \text{logito}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[2]})] &= \log \left[\frac{P(Y_t \leq j | \mathbf{z}_{t-1}^{[1]})/P(Y_t > j | \mathbf{z}_{t-1}^{[1]})}{P(Y_t \leq j | \mathbf{z}_{t-1}^{[2]})/P(Y_t > j | \mathbf{z}_{t-1}^{[2]})} \right] \\
 &= \log \left(\frac{\text{odds}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[1]})]}{\text{odds}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[2]})]} \right) \\
 &= \gamma'(\mathbf{z}_{t-1}^{[1]} - \mathbf{z}_{t-1}^{[2]}), \tag{5.6}
 \end{aligned}$$

onde *odds*, aqui, denota a função bastante comum em Estatística¹ definida por

$$\text{odds}[\text{'evento de probabilidade } \pi'] \equiv \frac{\pi}{1 - \pi}.$$

Do resultado (5.6), obtemos a chamada *razão de chances cumulativas*, dada por

$$\begin{aligned}
 \text{RCC} &\equiv \frac{\text{odds}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[1]})]}{\text{odds}[P(Y_t \leq j | \mathbf{z}_{t-1}^{[2]})]} \\
 &= e^{\gamma'(\mathbf{z}_{t-1}^{[1]} - \mathbf{z}_{t-1}^{[2]})}.
 \end{aligned}$$

Dizendo de outra forma, a chance de se obter resposta $\leq j$ para $\mathbf{z}_{t-1} = \mathbf{z}_{t-1}^{[1]}$ é

$$\exp[\gamma'(\mathbf{z}_{t-1}^{[1]} - \mathbf{z}_{t-1}^{[2]})]$$

vezes maior ou menor que (ou seja, *vezes*) a chance para $\mathbf{z}_{t-1} = \mathbf{z}_{t-1}^{[2]}$. O logaritmo da razão de chances cumulativas RCC é proporcional à distância entre $\mathbf{z}_{t-1}^{[1]}$ e $\mathbf{z}_{t-1}^{[2]}$, e esta mesma proporcionalidade (constante) se aplica também aos outros logitos cumulativos, ou seja, ela é válida para todos os q logitos do modelo. Devido a esta propriedade, o modelo (5.5) é chamado *modelo de chances proporcionais* (Agresti, 2002).

5.2.1 Estimação

A estimação será descrita somente para o caso utilizado nas simulações e na aplicação aos dados reais, no qual temos $m = 3$ categorias de resposta. O desenvolvimento segue

¹Matematicamente, as funções *odds* e logito desempenham a mesma tarefa, sendo a *odds* usada para *eventos*, e a logito usada para *probabilidades* de eventos. Contudo, historicamente, convencionou-se utilizar o nome *odds ratio* (razão de chances) para a razão de *probabilidades* condicionais de eventos; por isto a sua adoção aqui.

as mesmas linhas gerais do caso binário, considerado no Capítulo 4, exceto por algumas complicações técnicas que surgem devido à natureza multivariada do problema. Detalhes sobre o desenvolvimento de cada expressão, e sobre a estimação para $m > 3$, são apresentados em Kedem e Fokianos (2002), Capítulo 3.

Quando dispomos de $m = 3$ categorias de resposta, \mathbf{Y}_t denota o vetor $(Y_{t1}, Y_{t2})'$, lembrando que

$$Y_{tj} = \begin{cases} 1, & \text{se a } j\text{-ésima categoria é observada no tempo } t \\ 0, & \text{caso contrário} \end{cases},$$

com probabilidades de ocorrência dadas por $\boldsymbol{\pi}_t = (\pi_{t1}, \pi_{t2})'$. Para a terceira categoria de resposta, temos $Y_{t3} = 1 - (Y_{t1} + Y_{t2})$ e $\pi_{t3} = 1 - (\pi_{t1} + \pi_{t2})$.

A variância condicional de \mathbf{Y}_t é dada por uma matriz 2×2 , aqui denotada por $\boldsymbol{\Sigma}_t$, e dada por

$$\begin{aligned} \text{Var}(\mathbf{Y}_t | \mathcal{F}_{t-1}) &\equiv \boldsymbol{\Sigma}_t(\boldsymbol{\beta}) \\ &= \begin{bmatrix} \pi_{t1}(\boldsymbol{\beta})(1 - \pi_{t1}(\boldsymbol{\beta})) & -\pi_{t1}(\boldsymbol{\beta})\pi_{t2}(\boldsymbol{\beta}) \\ -\pi_{t1}(\boldsymbol{\beta})\pi_{t2}(\boldsymbol{\beta}) & \pi_{t2}(\boldsymbol{\beta})(1 - \pi_{t2}(\boldsymbol{\beta})) \end{bmatrix}. \end{aligned} \quad (5.7)$$

A função de verossimilhança parcial é dada em termos das probabilidades multinomiais como

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^N \prod_{j=1}^3 \pi_{tj}^{y_{tj}}(\boldsymbol{\beta}), \quad (5.8)$$

de forma que a log-verossimilhança parcial fica dada por

$$l(\boldsymbol{\beta}) = \sum_{t=1}^N \sum_{j=1}^3 y_{tj} \log \pi_{tj}(\boldsymbol{\beta}). \quad (5.9)$$

Podemos abrir o somatório interno de (5.9) e, após algumas manipulações algébricas, reexpressá-la como

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{t=1}^N \left\{ y_{t1} \log \left(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})} \right) + y_{t2} \log \left(\frac{\pi_{t2}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})} \right) \right. \\ &\quad \left. + \log (1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})) \right\}, \end{aligned} \quad (5.10)$$

5.2. O modelo de chances proporcionais

introduzindo uma reparametrização que facilitará os cálculos posteriores. Agora, substituindo

$$\begin{aligned}\boldsymbol{\theta}_t(\boldsymbol{\beta}) &= (\theta_{t1}(\boldsymbol{\beta}), \theta_{t2}(\boldsymbol{\beta}))' \\ &= \left(\log \left(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})} \right), \log \left(\frac{\pi_{t2}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})} \right) \right)',\end{aligned}\quad (5.11)$$

reescrevemos (5.10) como

$$\begin{aligned}l(\boldsymbol{\beta}) &= \sum_{t=1}^N \{ y_{t1} \theta_{t1}(\boldsymbol{\beta}) + y_{t2} \theta_{t2}(\boldsymbol{\beta}) - \log [1 + \exp(\theta_{t1}(\boldsymbol{\beta})) + \exp(\theta_{t2}(\boldsymbol{\beta}))] \} \\ &= \sum_{t=1}^N l_t(\boldsymbol{\beta}).\end{aligned}\quad (5.12)$$

Para o cálculo do escore parcial, fazemos uso da regra da cadeia para funções multivariadas, de forma que

$$\frac{\partial l_t}{\partial \boldsymbol{\beta}'} = \frac{\partial l_t}{\partial \boldsymbol{\theta}_t'} \frac{\partial \boldsymbol{\theta}_t}{\partial \boldsymbol{\pi}_t'} \frac{\partial \boldsymbol{\pi}_t}{\partial \boldsymbol{\eta}_t'} \frac{\partial \boldsymbol{\eta}_t}{\partial \boldsymbol{\beta}'},\quad (5.13)$$

onde $\boldsymbol{\eta}_t = \mathbf{Z}'_{t-1} \boldsymbol{\beta}$. O cálculo das duas primeiras derivadas parciais de (5.13) traz

$$\frac{\partial l_t}{\partial \boldsymbol{\theta}_t'} = (Y_{t1} - \pi_{t1}, Y_{t2} - \pi_{t2}) = (\mathbf{Y}_t - \boldsymbol{\pi}_t)'\quad (5.14)$$

e

$$\frac{\partial \boldsymbol{\theta}_t}{\partial \boldsymbol{\pi}_t'} = \begin{bmatrix} \frac{1 - \pi_{t2}}{\pi_{t1}(1 - \pi_{t1} - \pi_{t2})} & \frac{1}{1 - \pi_{t1} - \pi_{t2}} \\ \frac{1}{1 - \pi_{t1} - \pi_{t2}} & \frac{1 - \pi_{t1}}{\pi_{t2}(1 - \pi_{t1} - \pi_{t2})} \end{bmatrix} = \boldsymbol{\Sigma}_t^{-1}.\quad (5.15)$$

Denotando a terceira derivada parcial de (5.13) por $\mathbf{D}'_t(\boldsymbol{\beta})$ temos, a partir de (5.3), que

$$\mathbf{D}'_t \equiv \frac{\partial \boldsymbol{\pi}_t}{\partial \boldsymbol{\eta}_t'} = \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}} & \frac{\partial \pi_{t1}}{\partial \eta_{t2}} \\ \frac{\partial \pi_{t2}}{\partial \eta_{t1}} & \frac{\partial \pi_{t2}}{\partial \eta_{t2}} \end{bmatrix} = \frac{\partial \mathbf{h}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t'}.\quad (5.16)$$

Além disso, sabemos que

$$\frac{\partial \boldsymbol{\eta}_t}{\partial \boldsymbol{\beta}'} = \mathbf{Z}'_{t-1}.\quad (5.17)$$

Substituindo agora (5.14), (5.15), (5.16) e (5.17) em (5.13), vemos que

$$\frac{\partial l_t}{\partial \boldsymbol{\beta}'} = (\mathbf{Y}_t - \boldsymbol{\pi}_t)' \boldsymbol{\Sigma}_t^{-1} \mathbf{D}_t' \mathbf{Z}'_{t-1}, \quad (5.18)$$

e portanto o escore parcial fica dado por

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{Z}_{t-1} \mathbf{D}_t(\boldsymbol{\beta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}) (\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})). \quad (5.19)$$

Por último, pode-se mostrar que o cálculo da matriz de informação condicional cumulativa conduz a

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{Z}_{t-1} \mathbf{D}_t(\boldsymbol{\beta}) (\mathbf{D}_t(\boldsymbol{\beta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}))' \mathbf{Z}'_{t-1}. \quad (5.20)$$

As expressões para o vetor escore (5.19) e para a matriz de informação (5.20) são então utilizadas no procedimento de Newton-Raphson (2.17) para a obtenção do EMVP, $\hat{\boldsymbol{\beta}}$, de $\boldsymbol{\beta}$.

5.3 Estudos de simulação

Nesta seção, simulamos um modelo de chances proporcionais com duas covariáveis, sendo uma determinística (senoidal) e a outra com memória longa. Primeiramente, o modelo é simulado uma única vez, para a verificação de detalhes sobre a simulação e sobre a estimação a partir da série simulada. Após esta análise, o modelo é simulado 1000 vezes, de forma a avaliarmos se o estimador de máxima verossimilhança parcial do modelo de chances proporcionais, na modelagem de séries categóricas ordinais com memória longa, possui as propriedades de consistência e normalidade assintótica.

5.3.1 Estudo detalhado de uma série simulada

Foram simuladas $N = 500$ observações do modelo

$$\log \left[\frac{P(Y_t \leq j | \mathcal{F}_{t-1})}{P(Y_t > j | \mathcal{F}_{t-1})} \right] = \theta_j + \gamma_1 10 \left(\text{sen} \left(\frac{2\pi t}{100} \right) + 1 \right) + \gamma_2 W_t, \quad j = 1, 2, \quad (5.21)$$

para $t = 1, \dots, N$, onde $\mathbf{Z}_{t-1} = (X_t, W_t)'$ é composto por uma série determinística cíclica

$$X_t = 10 \left(\text{sen} \left(\frac{2\pi t}{100} \right) + 1 \right),$$

5.3. Estudos de simulação

e uma série de memória longa (W_t) gerada de acordo com o processo

$$\{W_t\} \sim ARFIMA(1, d, 0) \quad \text{com} \quad \phi = 0,75, d = 0,45, \mu_\varepsilon = 11 \text{ e } \sigma_\varepsilon = 1,$$

onde μ_ε e σ_ε são a média e o desvio-padrão utilizados na geração do processo de inovações $\{\varepsilon_t\} \stackrel{iid}{\sim} \mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon)$ da série W_t . Os parâmetros para a simulação foram especificados como

$$\beta = (\theta_1, \theta_2, \gamma_1, \gamma_2)' = (-3, 7, -2, 2, 0, 15, 0, 15)'. \quad (5.22)$$

A Figura 5.1 apresenta o gráfico da série W_t com sua FAC e histograma, e o da série determinística X_t . As séries covariáveis foram utilizadas para a simulação de Y_t a partir de (5.21), e seus valores foram gerados. Verifica-se a presença de memória longa em Y_t (Figura 5.2 (b)) e um comportamento de alta alternância entre seus valores (Figura 5.2 (a)).

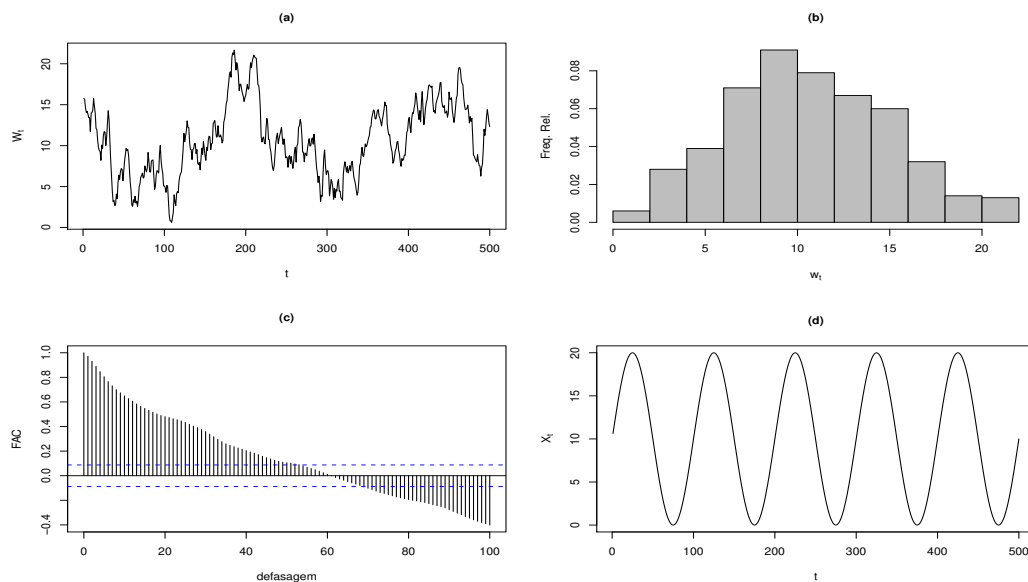


Figura 5.1: Séries simuladas: (a), (b) e (c) $\{W_t\}$, seu histograma e sua FAC; (d) $\{X_t\}$.

Para analisar o relacionamento entre Y_t e as covariáveis, não poderemos inspecionar os gráficos das duas curvas logísticas (sigmóides) do modelo, pois, pelo fato de haverem duas covariáveis, teremos uma *superfície* no espaço gerado por $\{X_t\}$ e $\{W_t\}$, e suas projeções sobre os planos $P(Y_t \leq j | \mathcal{F}_{t-1}) \times X_t$ e $P(Y_t \leq j | \mathcal{F}_{t-1}) \times W_t$ proporciona difícil visualização. Entretanto, sabemos que elas são decrescentes em ambas as covariáveis, pois

especificou-se para a simulação $\gamma_1, \gamma_2 > 0$, o que estabelece um relacionamento inverso entre Y_t e X_t e entre Y_t e W_t . Assim, devemos esperar que o evento $\{Y_t = 1\}$ ocorra para valores *altos* de X_t e de W_t , ao passo que $\{Y_t = 3\}$ deve ocorrer para valores *baixos* de X_t e de W_t . As Figuras 5.2 (c) e (d) vêm a confirmar estas relações, e portanto a simulação está coerente.

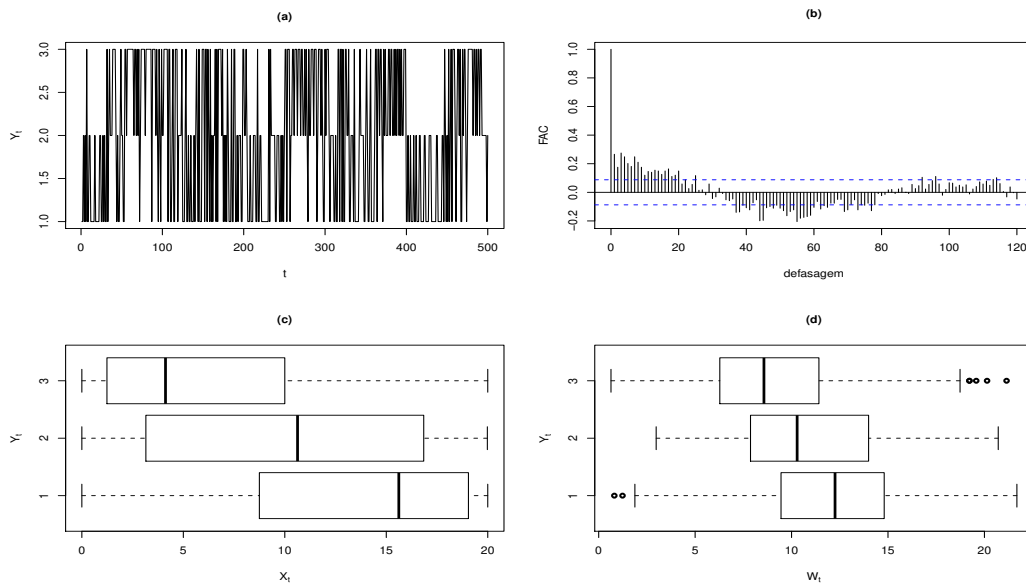


Figura 5.2: Série $\{Y_t\}$ gerada (a), sua FAC (b), e diagramas de dispersão entre $\{Y_t\}$ e as covariáveis $\{X_t\}$ (c) e $\{W_t\}$ (d).

O relacionamento entre as séries covariáveis e a série resposta simulada pode ser também visualizado pela Figura 5.3, de onde se verifica que valores altos de Y_t tendem a ocorrer para valores baixos de X_t e de W_t . Por exemplo, para t aproximadamente entre 50 e 100, Y_t assume predominantemente seu valor médio ou alto, 2 ou 3, sendo que X_t e W_t assumem valores sempre abaixo de suas médias. O inverso ocorre para $400 < t < 450$.

Estimando o modelo a partir da série Y_t simulada, obtemos as estimativas apresentadas na Tabela 5.1. Nota-se a boa qualidade da estimação: estimativas altamente significativas e muito próximas dos valores reais dos parâmetros.

É interessante fazer uma comparação visando à confirmação da validade da suposição de chances proporcionais do modelo simulado. Para isso, ajusta-se aos dados simulados um modelo alternativo, mais complexo, e os ajustes dos dois modelos são comparados. Este modelo alternativo considera efeitos separados para cada logito, isto é, ele considera

5.3. Estudos de simulação

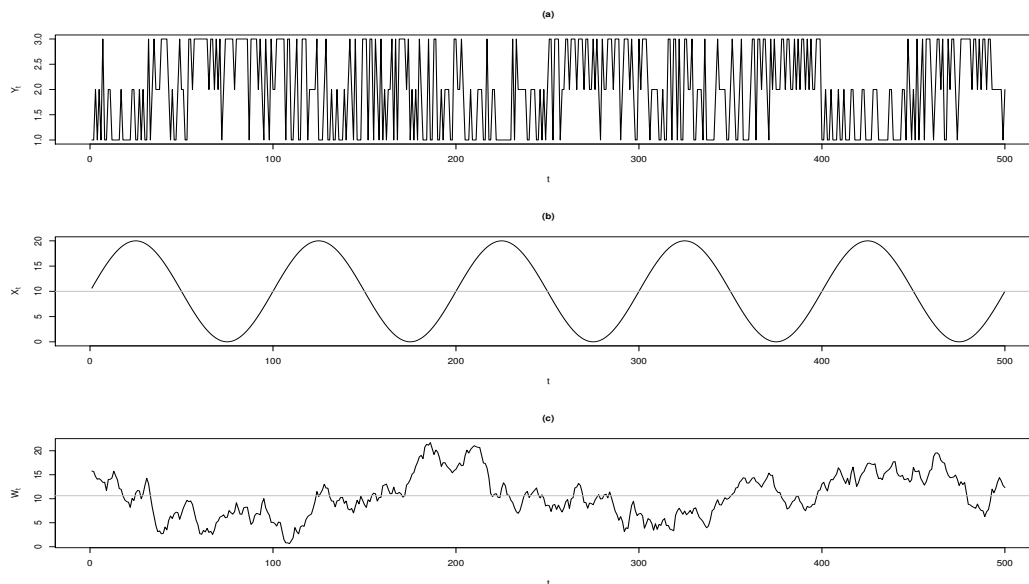


Figura 5.3: Séries $\{Y_t\}$ (a), $\{X_t\}$ (b) e $\{W_t\}$ (c) com sua linhas médias.

Tabela 5.1: Estimativas dos parâmetros do modelo obtidas para o ajuste à série simulada.

Parâmetro	Real	Estimativa	E.P.	Valor t	p -valor
θ_1	-3,7	-3,622	0,317	-11,44	<0,001
θ_2	-2,2	-1,977	0,282	-7,00	<0,001
γ_1	0,15	0,124	0,013	9,33	<0,001
γ_2	0,15	0,147	0,021	7,05	<0,001

um parâmetro de efeito γ_{pj} da p -ésima covariável para cada logito j . A verificação é feita por meio de um teste de razão das verossimilhanças (TRV) estimadas para cada modelo, o conhecido teste assintótico *score*. O ajuste do modelo alternativo trouxe as estimativas apresentadas na Tabela 5.2. Verifica-se que as estimativas dos parâmetros de X_t de cada logito, γ_{x1} e γ_{x2} , são muito próximas. O mesmo ocorre para as estimativas dos parâmetros de W_t . Isto sugere que não há a necessidade de se especificar efeitos diferentes para cada logito, ou seja, que vale a suposição de chances proporcionais, de um único efeito para cada covariável. Realizando o TRV, obtemos a confirmação da validade da suposição ($\chi^2 = 1,43$ em 2 g.l.; $p = 0,49$).

Capítulo 5. Modelo de Chances Proporcionais para Séries Categóricas Ordinais

Tabela 5.2: Estimativas dos parâmetros para o ajuste do modelo alternativo, com efeitos separados para cada logito.

Parâmetro	Estimativa	E.P.	Valor t	p -valor
θ_1	-3,401	0,365	-9,31	<0,001
θ_2	-2,177	0,336	-6,48	<0,001
γ_{x1}	0,120	0,016	7,68	<0,001
γ_{x2}	0,129	0,016	7,95	<0,001
γ_{w1}	0,133	0,024	5,55	<0,001
γ_{w2}	0,165	0,026	6,30	<0,001

5.3.2 Estudo de simulação geral

O modelo (5.21) foi simulado 1000 vezes para as combinações de $N = 200, 500$ e 1000 , e $d = 0,2, 0,4$ e $0,49$. Foram calculadas as médias e os desvios-padrões amostrais (\hat{EP}) das estimativas obtidas para cada parâmetro, e também seus desvios-padrões teóricos aproximados (segundo Kedem e Fokianos, 2002), obtidos a partir da inversa da matriz de informação condicional cumulativa, $(\mathbf{G}_N(\boldsymbol{\beta}))^{-1}$. Os resultados constam da Tabela 5.3.

Tabela 5.3: Resultado das simulações (valor real: $\boldsymbol{\beta} = (-3, 7, -2, 2, 0, 15, 0, 15)'$).

N	Estat.	$d = 0, 2$				$d = 0, 4$				$d = 0, 49$			
		θ_1	θ_2	γ_1	γ_2	θ_1	θ_2	γ_1	γ_2	θ_1	θ_2	γ_1	γ_2
200	$\hat{\boldsymbol{\beta}}$	-3,77	-2,25	0,15	0,15	-3,78	-2,25	0,15	0,15	-3,75	-2,22	0,15	0,15
	\mathbf{G}_N^{-1}	0,81	0,78	0,02	0,07	0,43	0,38	0,02	0,03	0,55	0,50	0,02	0,04
	\hat{EP}	0,80	0,77	0,02	0,07	0,47	0,41	0,03	0,04	0,54	0,49	0,02	0,04
500	$\hat{\boldsymbol{\beta}}$	-3,70	-2,19	0,15	0,15	-3,73	-2,21	0,15	0,15	-3,74	-2,23	0,15	0,15
	\mathbf{G}_N^{-1}	0,52	0,50	0,01	0,05	0,38	0,36	0,01	0,03	0,32	0,29	0,01	0,03
	\hat{EP}	0,50	0,48	0,01	0,05	0,40	0,38	0,01	0,03	0,33	0,30	0,01	0,03
1000	$\hat{\boldsymbol{\beta}}$	-3,73	-2,22	0,15	0,15	-3,70	-2,20	0,15	0,15	-3,71	-2,20	0,15	0,15
	\mathbf{G}_N^{-1}	0,37	0,35	0,01	0,03	0,24	0,22	0,01	0,02	0,27	0,25	0,01	0,02
	\hat{EP}	0,36	0,33	0,01	0,03	0,24	0,23	0,01	0,02	0,27	0,25	0,01	0,02

Da Tabela 5.3, constata-se que:

- os resultados da simulação sugerem a consistência dos estimadores de MVP, dado o alto grau de proximidade entre seus valores reais e suas estimativas assintóticas, e o decréscimo dos erros-padrões à medida em que aumenta o tamanho amostral;

5.3. Estudos de simulação

- é notável a enorme similaridade entre os erros-padrões obtidos segundo a inversão da matriz de informação, $\mathbf{G}_N^{-1}(\boldsymbol{\beta})$, e os obtidos por simulação, inclusive para o estimador do parâmetro da série de memória longa, γ_2 ; tal fato permite concluir que a presença de ML *não* inviabiliza a obtenção de estimativa para os erros-padrões dos parâmetros do modelo pela forma tradicional, através de \mathbf{G}_N^{-1} .

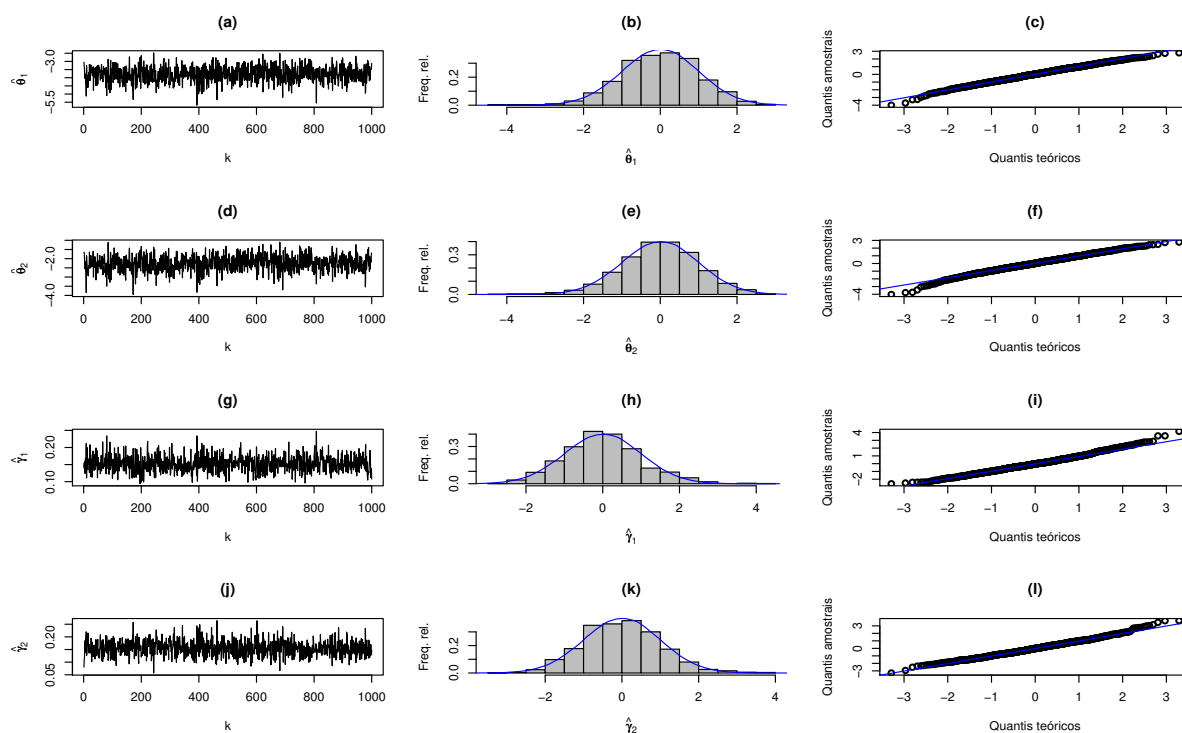


Figura 5.4: Seqüências de estimativas obtidas para $\hat{\boldsymbol{\beta}}$, para o caso $N = 200$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\theta}_1$, (d), (e) e (f) $\hat{\theta}_2$, (g), (h) e (i) $\hat{\gamma}_1$, (j), (k) e (l) $\hat{\gamma}_2$.

Em termos da distribuição assintótica dos estimadores, os $Q-Q$ plots da Figura 5.4, obtidos para a simulação com $N = 200$ e $d = 0,49$, apontam o mesmo comportamento observado para o caso binário, visto no Capítulo 4: existe uma leve assimetria nas distribuições marginais das seqüências de estimativas obtidas, causadas por um relacionamento entre os termos de intercepto e os de efeito (superestimação de um associado à subestimação do outro). Contudo, aumentando o tamanho das séries, é possível observar que tal assimetria diminui e tende à simetria, como indicam os gráficos da Figura 5.5 ($N = 1000$ e $d = 0,49$).

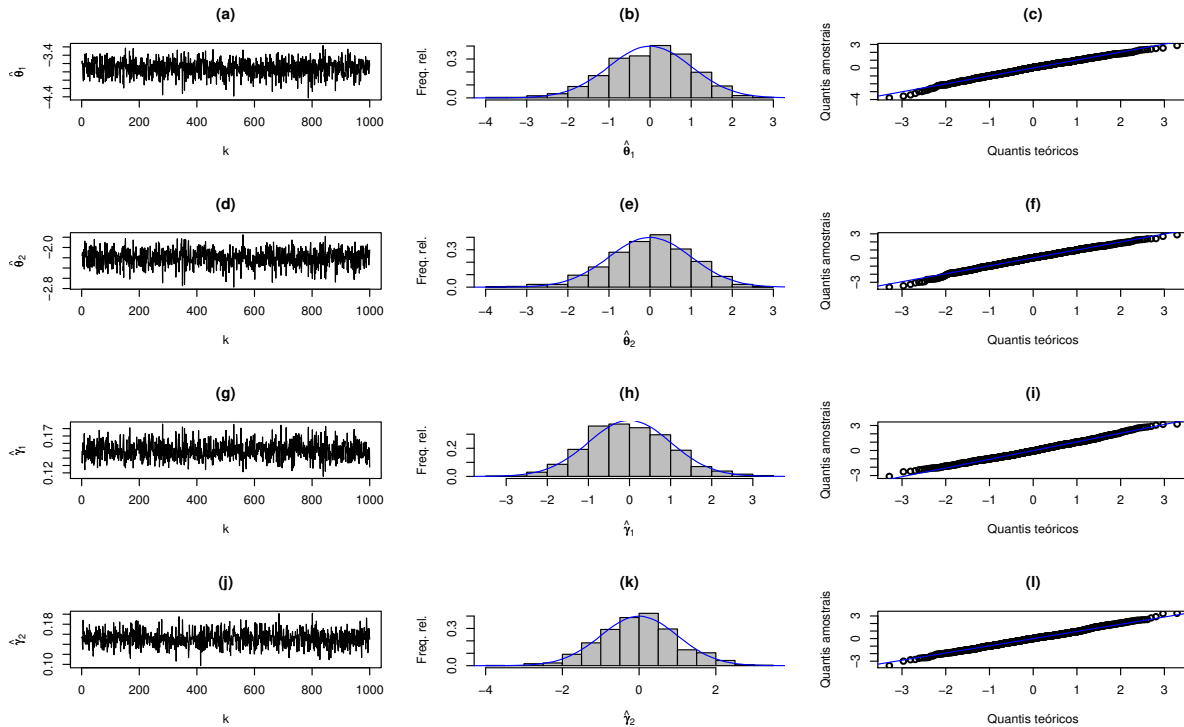


Figura 5.5: Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 1000$ e $d = 0,49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0,1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0,1)$: (a), (b) e (c) $\hat{\theta}_1$, (d), (e) e (f) $\hat{\theta}_2$, (g), (h) e (i) $\hat{\gamma}_1$, (j), (k) e (l) $\hat{\gamma}_2$.

Assim sendo, temos evidências de que a distribuição assintótica marginal de cada estimador é muito próxima da normal, e portanto a inferência estatística usual, do caso do MLG clássico para dados independentes, pode ser feita também para o modelo de chances proporcionais para séries temporais que apresentem memória longa na série resposta e/ou nas covariáveis.

5.4 Medidas de performance preditiva para 3 categorias de resposta

No estudo da poluição do ar pela abordagem categórica ordinal, com as 3 categorias de resposta dadas por poluição “baixa”, “média” e “alta”, teremos uma matriz de erros de dimensão 3×3 , como a apresentada na Tabela 5.4. Neste caso, estaremos interessados nas seguintes medidas de performance:

5.4. Medidas de performance preditiva para 3 categorias de resposta

Tabela 5.4: Matriz de confusão para classificação em 3 categorias.

Predita	Real			Total
	Baixa (1)	Média (2)	Alta (3)	
Baixa (1)	a	b	c	$a + b + c$
Média (2)	d	e	f	$d + e + f$
Alta (3)	g	h	i	$g + h + i$
Total	$a + d + g$	$b + e + h$	$c + f + i$	$a + \dots + i$

- Acurácia (AC): definida de forma similar à do caso binário, e agora dada por

$$AC = \frac{a + e + i}{a + \dots + i};$$

- Precisão de baixa poluição (Prec_BAIXA): estabelecida como a proporção de dias de baixa poluição corretamente preditos, dentre todos os dias observados com poluição baixa, e dada por

$$Prec_BAIXA = \frac{a}{a + d + g};$$

- Precisão de média poluição (Prec_MÉDIA): estabelecida como a proporção de dias de média poluição corretamente preditos, dentre todos os dias de poluição média observada, e dada por

$$Prec_MÉDIA = \frac{e}{b + e + h};$$

- Precisão de alta poluição (Prec_ALTA): estabelecida como a proporção de dias de alta poluição corretamente preditos, dentre todos os dias observados com poluição média, e dada por

$$Prec_ALTA = \frac{i}{c + f + i};$$

- Taxa de falsos negativos do tipo “13” (FN13): é o pior erro do tipo falso negativo que se pode cometer, prever baixa poluição para um dia de poluição alta, e é dada por

$$FN13 = \frac{c}{a + \dots + i};$$

- Taxa de falsos negativos do tipo “23” (FN23): é o segundo pior tipo de falso negativo, prever poluição média para um dia de poluição alta, e é dada por

$$FN23 = \frac{f}{a + \dots + i};$$

- Taxa de falsos negativos do tipo “12” (FN12): é outro tipo de falso negativo, prever baixa poluição para um dia de poluição média, e é dada por

$$\text{FN12} = \frac{b}{a + \dots + i};$$

- Taxa de falsos positivos do tipo “31” (FP31): é o pior erro do tipo falso positivo, prever poluição alta para um dia de baixa poluição, e é dada por

$$\text{FP31} = \frac{g}{a + \dots + i};$$

- Taxa de falsos positivos do tipo “32” (FP32): é o segundo pior tipo de falso positivo, prever poluição alta para um dia de poluição média, e é dada por

$$\text{FP32} = \frac{h}{a + \dots + i};$$

- Taxa de falsos positivos do tipo “21” (FP21): é outro tipo de falso positivo, prever poluição média para um dia de baixa poluição, e é dada por

$$\text{FP21} = \frac{d}{a + \dots + i}.$$

As medidas de sensibilidade e especificidade anteriormente definidas para a classificação binária (Capítulo 4) não podem ser diretamente consideradas para o caso tricotômico. Além disso, as precisões específicas de cada classe para o caso tricotômico, foram definidas em relação ao total observado para cada classe, ao invés do total predito, por se considerar mais relevantes desta forma.

Vale comentar que as taxas de erro FNs e FPs para a Tabela 5.4 têm importâncias distintas para diferentes objetivos de utilização. Para a utilização do modelo tendo como foco a saúde da população, por exemplo, deseja-se que o modelo não preveja baixa (média) poluição em dias em que ela será média (alta), pois nenhum alerta será feito à população, e assim a saúde das pessoas estará em risco. Portanto, as taxas FN são as mais preocupantes quando o objetivo da utilização do modelo é prevenir doenças e complicações respiratórias.

Em outro contexto, o foco da utilização poderia estar voltado para as medidas de política pública tomadas pelo governo local. Sempre que a poluição predita atinge níveis alarmantes, são adotadas medidas de segurança e de prevenção, como a restrição à circulação de carros e às atividades industriais, de forma a evitar a elevação do nível de

poluição para patamares ainda mais altos. Sendo assim, quando ocorrerem os erros do tipo FP, da previsão de alta (média) poluição em dias de poluição média (baixa), a economia local será fortemente afetada, pois os erros de previsão causarão prejuízo. Portanto, mantendo-se o foco na economia, os erros do tipo FP é que são os mais preocupantes.

5.5 Aplicação em poluição do ar

O modelo de chances proporcionais foi aplicado aos dados de poluição do ar da cidade de Santiago do Chile, os quais foram analisados no Capítulo 4 pelo modelo de regressão logística. O objetivo com a re-utilização dos dados de poluição é a comparação das performances preditivas das duas abordagens: a predição de PM10 sob categorização binária *versus* categorização em três classes: poluição “baixa”, “média” e “alta”. Verificou-se que as predições obtidas para as séries binárias foram superiores às obtidas para as categóricas ordinais, mas que o modelo de chances proporcionais constitui também uma boa opção para a modelagem de séries de poluição com memória longa.

Duas séries ordinais oriundas da discretização de PM10 são analisadas: $Y_{[100]}^{\text{Ord}}$, categorizada a partir dos limiares 50 e 100, e $Y_{[120]}^{\text{Ord}}$, utilizando os limiares 50 e 120 $\mu\text{g}/\text{m}^3$. O limiar 50 corresponde à meta de longo prazo da OMS, e já é utilizado atualmente em países mais preocupados com a questão ambiental, como a Irlanda e a Nova Zelândia, daí sua adoção.

A Figura 5.6 apresenta as séries $Y_{[100]}^{\text{Ord}}$, $Y_{[120]}^{\text{Ord}}$, e suas FAC's, que indicam a presença de memória longa em ambas as séries. Seguindo as mesmas idéias das análises do Capítulo 5, o modelo de chances proporcionais foi aplicado aos subconjuntos de dados de ajuste de $Y_{[100]}^{\text{Ord}}$ e $Y_{[120]}^{\text{Ord}}$, que abrangem as observações de 01/Janeiro/2004 a 14/Abril/2007, e posteriormente utilizado para prever seus valores no período destinado à predição, que vai de 15 de abril a 14 de setembro de 2007. As classes “1”, “2” e “3”, que representam os episódios de poluição “baixa”, “média” e “alta”, respectivamente, ocorrem com as frequências de 36, 45 e 19% para $Y_{[100]}^{\text{Ord}}$, e de 36, 52 e 12% para $Y_{[120]}^{\text{Ord}}$. Dentro do período de predição, as frequências equivalem a 18, 31 e 51% para $Y_{[100]}^{\text{Ord}}$, e de 18, 45 e 37% para $Y_{[120]}^{\text{Ord}}$.

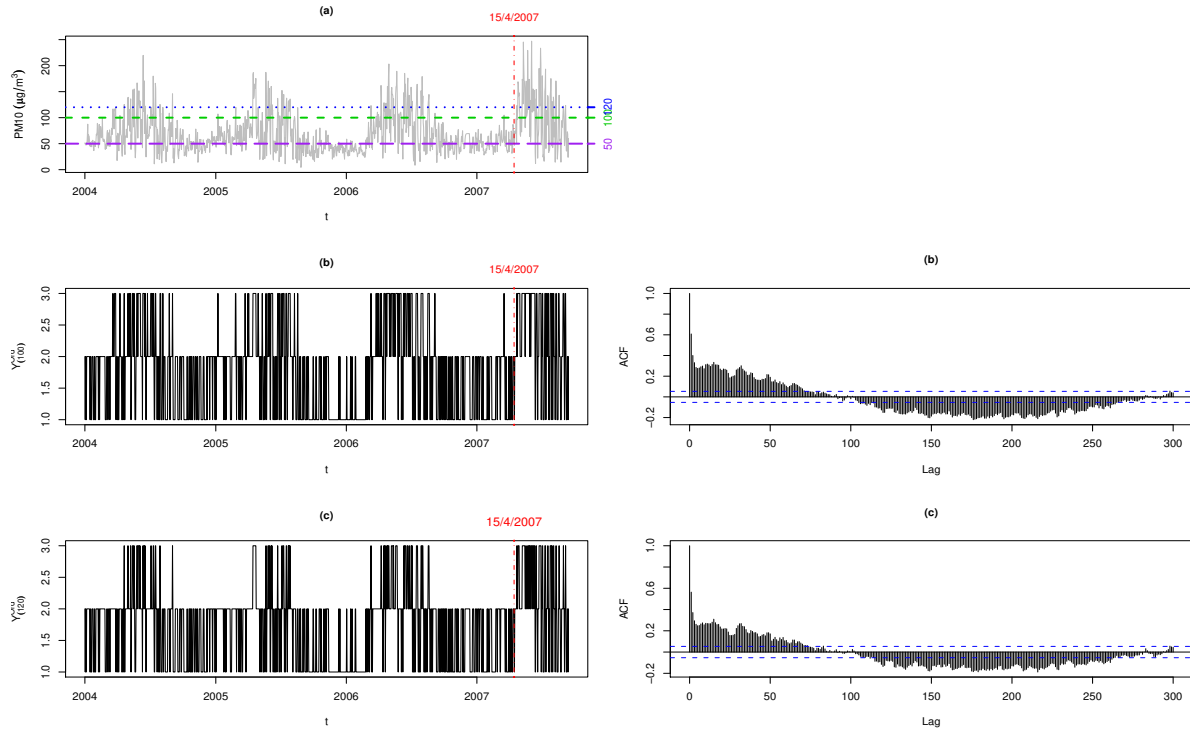


Figura 5.6: Níveis de corte para PM10 (a), $Y_{[100]}^{Ord}$ e sua FAC (b) e $Y_{[120]}^{Ord}$ e sua FAC (c).

5.5.1 Análise da série com limiares 50 e 100

O melhor modelo encontrado para $Y_{[100]}^{Ord}$, em termos de melhor qualidade de ajuste e de menor valor para o critério BIC, será denotado por “M^{Ord}” e é dado por

$$\begin{aligned} \log \left[\frac{P(Y_{[100]}^{Ord}(t) \leq j | \mathcal{F}_{t-1})}{P(Y_{[100]}^{Ord}(t) > j | \mathcal{F}_{t-1})} \right] &= \theta_j + \gamma_1 Y_{[100]}^{Ord}(t-1)_2 + \gamma_2 Y_{[100]}^{Ord}(t-1)_3 + \gamma_3 Y_{[100]}^{Ord}(t-2)_2 \\ &+ \gamma_4 Y_{[100]}^{Ord}(t-2)_3 + \gamma_5 SO2_t + \gamma_6 SO2_{t-1} + \gamma_7 SO2_{t-2} \\ &+ \gamma_8 NO2_{t-1} + \gamma_9 NO2_{t-2} + \gamma_{10} I_{[out]}(t) + \gamma_{11} I_{[prim]}(t) \\ &+ \gamma_{12} I_{[ver]}(t), \quad j = 1, 2, \end{aligned} \quad (5.23)$$

$t = 1, \dots, N$, onde $Y_{[100]}^{Ord}(t-1)_k$ é uma variável indicadora da ocorrência da k -ésima categoria no dia anterior, e $I_{[out]}$ a $I_{[ver]}$ são as indicadoras das estações do ano “outono”, “primavera” e “verão”, sendo a estação de inverno a estação de referência.

Os coeficientes estimados são todos altamente significativos (Tabela 5.5), à exceção da indicadora da estação primavera. Em comparação com os modelos obtidos para o caso binário (Tabelas 4.4 e 4.9), observa-se que um maior número de termos é utilizado para

5.5. Aplicação em poluição do ar

a explicação da versão categórica tricotômica de PM10. Com relação à interpretação dos parâmetros, verifica-se pelos sinais negativos que a ocorrência de poluição por PM10 média ou alta nos dois dias anteriores, e as poluições por SO2 e NO2 do mesmo dia, diminuem a chance de poluição baixa. Ou, colocando de outra forma, aumentam a chance de poluição média ou alta no dia corrente.

Tabela 5.5: Modelo estimado para $Y_{[100]}^{\text{Ord}}$.

Termo	Estimativa	E.P.	Valor t	$Pr(> t)$
θ_1	5,18	0,42	12,3	<0,001
θ_2	10,67	0,57	18,6	<0,001
$Y_{[100]}^{\text{Ord}}(t-1)_2$	-1,98	0,19	-10,2	<0,001
$Y_{[100]}^{\text{Ord}}(t-1)_3$	-3,82	0,40	-9,5	<0,001
$Y_{[100]}^{\text{Ord}}(t-2)_2$	-0,82	0,19	-4,3	<0,001
$Y_{[100]}^{\text{Ord}}(t-2)_3$	-1,90	0,40	-4,8	<0,001
$SO2_t$	-1,34	0,09	-14,6	<0,001
$SO2_{t-1}$	0,41	0,09	4,5	<0,001
$SO2_{t-2}$	0,29	0,08	3,4	<0,001
$NO2_t$	-0,18	0,02	-10,5	<0,001
$NO2_{t-2}$	0,07	0,02	4,4	<0,001
$I_{[\text{out}]}(t)$	-0,97	0,23	-4,3	<0,001
$I_{[\text{prim}]}(t)$	-0,30	0,25	-1,2	0,12
$I_{[\text{ver}]}(t)$	-1,26	0,25	-5,0	<0,001

Para a predição de $Y_{[100]}^{\text{Ord}}$, utilizou-se dos valores preditos de SO2 e NO2, determinados no Capítulo 4. Quando a variável predita assume uma dentre três ou mais categorias, não faz sentido trabalhar com um único nível de corte para $\hat{\pi}_t$. No caso binário, por exemplo, é fácil estabelecer uma regra para a predição, pois a variável resposta possui apenas dois níveis, “0” ou “1”. Assim, é razoável estabelecer a regra

$$\hat{y}_t = 1 \quad \text{se} \quad \hat{\mu}_t = P(Y_t = 1 | \mathcal{F}_{t-1}) \geq P(Y_t = 0 | \mathcal{F}_{t-1}) = (1 - \hat{\mu}_t).$$

Em outras palavras, $\hat{y}_t = 1$ se $\hat{\mu}_t \geq 0,5$, caso contrário, $\hat{y}_t = 0$. Desta maneira, determina-se que a categoria predita é aquela cuja probabilidade predita seja a maior. Agora, aplicando esta idéia para o caso de três categorias temos, por exemplo, que

$$\hat{y}_t^{\text{Ord}} = 2 \quad \text{se} \quad \hat{\pi}_{t2} = P(Y_t^{\text{Ord}} = 2 | \mathcal{F}_{t-1}) \geq \hat{\pi}_{t1} \quad \text{e} \quad \hat{\pi}_{t2} \geq \hat{\pi}_{t3},$$

e assim por diante. Ou seja, $\hat{y}_t^{\text{Ord}} = k$ se $\hat{\pi}_{tk} = \max\{\hat{\pi}_{t1}, \hat{\pi}_{t2}, \hat{\pi}_{t3}\}$, $k \in \{1, 2, 3\}$. Esta foi a regra utilizada nas previsões deste capítulo.

A série predita é apresentada na Figura 5.7. Observa-se que, em geral, a série predita está “atrasada” em um dia em relação à observada, provavelmente pelo fato de o termo preditor mais importante para a previsão $Y_{[100]}^{\text{Ord}}(t)$ ser o seu valor do dia anterior, $Y_{[100]}^{\text{Ord}}(t-1)$. Entretanto, em períodos de dias com episódios consecutivos de alta poluição, a série predita “acompanha” a série real, sendo portanto útil ao alerta à população. Outro aspecto que merece atenção é a baixa quantidade dos erros mais graves, que são as previsões de poluição alta em dia de baixa, e as de baixa em dia de alta (Figura 5.7 (c)). Verifica-se que estes erros ocorrem geralmente nos dias de mudança entre os níveis alto e baixo de $Y_{[100]}^{\text{Ord}}$, pelo fato de as previsões estarem defasadas em um dia.

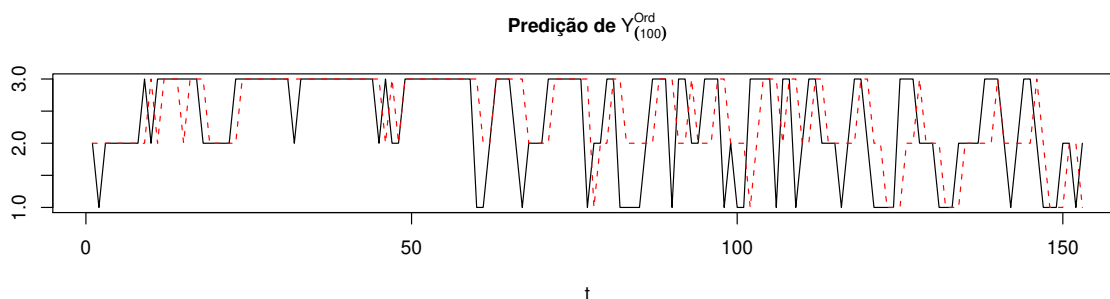


Figura 5.7: $Y_{[100]}^{\text{Ord}}$ observada (linha contínua) e predita (linha tracejada).

A qualidade das previsões pode ser melhor julgada a partir da matriz de erros (Tabela 5.6). A taxa de erro do tipo FN_{13} é de $2/153 = 1\%$, e a de FP_{31} , de $8/153 = 5\%$. Verifica-se também a baixa quantidade de acertos obtida para a previsão de baixa poluição. As principais medidas de performance preditiva de interesse apresentam-se na Tabela 5.7 (modelo “ M^{Ord} ”). A taxa de acerto geral é de 60%, e as precisões das previsões das classes “média” e “alta” poluição são de aproximadamente 70%. A precisão para baixa poluição é bem inferior, 21%. As taxas dos segundos piores erros, que são a previsão de média para um dia de alta, e a de alta para um dia de média poluição, são razoavelmente baixas, equivalendo a 14% e 7%, respectivamente.

Modelos alternativos foram ajustados para $Y_{[100]}^{\text{Ord}}$, seguindo as mesmas idéias gerais discutidas na análise dos dados feita no Capítulo 4. Foram ajustados um modelo baseado somente no passado, no qual apenas covariáveis *defasadas* são consideradas; o MLG

5.5. Aplicação em poluição do ar

Tabela 5.6: Matriz de erros para a predição de $Y_{[100]}^{Ord}$.

		Real			
		1	2	3	Total
Predito	1	6	4	2	12
	2	14	32	22	68
	3	8	11	54	73
	Total	28	47	78	153

Normal para PM10 original (contínua), categorizando posteriormente os valores preditos de acordo com os limiares 50 e 100; e o modelo “impraticável”, que é o próprio modelo (M^{Ord}) anteriormente apresentado, mas se utiliza dos valores contemporâneos *reais* das séries covariáveis, o que não é possível de se fazer na prática. A Tabela 5.7 apresenta os resultados das predições obtidas por cada um. Em termos de performance global, verifica-se que o modelo final (M^{Ord}) produziu predições 13% menos acuradas que o modelo “impraticável”, mas superiores ou equivalentes às produzidas pelo modelo baseado em covariáveis defasadas e pelo modelo Normal. Aliás, a equivalência aproximada entre as performances preditivas do modelo final e do modelo Normal (Tabela 5.7, modelos M^{Ord} e II) é de notável interesse, e conta a favor da consideração do modelo de chances proporcionais para a modelagem destes dados.

Tabela 5.7: Comparação das performances preditivas de todos os modelos para $Y_{[100]}^{Ord}$.

Medida de perf. preditiva (%)	Impraticável (III)	Baseado somente no passado (I)	Normal, pós-categorizado (II)	Modelo final (M^{Ord})
Acurácia	0,73	0,54	0,61	0,60
Precisão (Baixa)	0,39	0,25	0,21	0,21
Precisão (Média)	0,81	0,02	0,72	0,68
Precisão (Alta)	0,79	0,96	0,68	0,69
FN13 (PIOR erro)	0,00	0,02	0,00	0,01
FN23 (2º pior)	0,10	0,00	0,16	0,14
FN12	0,01	0,05	0,02	0,03
FP31 (PIOR erro)	0,02	0,13	0,04	0,05
FP32 (2º pior)	0,05	0,25	0,07	0,07
FP21	0,09	0,01	0,10	0,09

As predições do modelo baseado somente em informações do passado (I) foram ruins.

Este modelo teve uma forte tendência a prever a poluição como “alta”, o que justifica a alta precisão observada para esta categoria e as baixas precisões para as demais (Tabela 5.7, coluna (I)). Inspeccionando os detalhes da previsão deste modelo, vimos que $\hat{\pi}_{t3} \geq 0,6$ para a maior parte do período de predição, fazendo com que a categoria “poluição alta” dominasse as outras. Pudemos observar que, dos 153 dias da predição, 133 foram preditos como ‘3’ (alta poluição), 18 como ‘1’ e apenas 2 como ‘2’ (poluição média); daí a performance desequilibrada. É difícil levantar os motivos que levaram a este resultado, mas acreditamos que estejam relacionados preponderantemente ao fato de o modelo utilizar apenas valores passados das covariáveis SO2 e NO2. De alguma forma, a combinação destes valores no preditor linear do modelo teve um efeito ruim sobre as probabilidades preditas de cada classe, de maneira que elas não se alternaram como deveriam, e deixaram as predições enviesadas.

Por último, vale destacar que todos os modelos tiveram dificuldade na predição dos dias de baixa poluição.

5.5.2 Análise da série com limiares 50 e 120

Para a modelagem de $Y_{[120]}^{\text{Ord}}$, o melhor modelo encontrado é o dado por

$$\begin{aligned} \log \left[\frac{P(Y_{[120]}^{\text{Ord}}(t) \leq j | \mathcal{F}_{t-1})}{P(Y_{[120]}^{\text{Ord}}(t) > j | \mathcal{F}_{t-1})} \right] &= \theta_j + \gamma_1 Y_{[120]}^{\text{Ord}}(t-1)_2 + \gamma_2 Y_{[120]}^{\text{Ord}}(t-1)_3 + \gamma_3 Y_{[120]}^{\text{Ord}}(t-2)_2 \\ &+ \gamma_4 Y_{[120]}^{\text{Ord}}(t-2)_3 + \gamma_5 SO2_t + \gamma_6 SO2_{t-1} + \gamma_7 NO2_t \\ &+ \gamma_8 NO2_{t-2} + \gamma_9 \text{sen} \left(\frac{2\pi t}{365} \right), \quad j = 1, 2, \end{aligned} \quad (5.24)$$

$t = 1, \dots, N$, que é bastante similar ao obtido para a série $Y_{[100]}^{\text{Ord}}$ (5.23). A principal diferença está no termo utilizado para a explicação da sazonalidade, que neste caso é o termo senoidal. Os coeficientes estimados são todos altamente significativos (Tabela 5.8), e suas interpretações são idênticas às das estimativas para o modelo de $Y_{[100]}^{\text{Ord}}$.

Os valores preditos de SO2 e NO2 no Capítulo 4 foram utilizados para a predição de $Y_{[120]}^{\text{Ord}}$ a partir do modelo (5.24). Verifica-se pela Figura 5.8 que a predição de $Y_{[120]}^{\text{Ord}}$ obteve menos êxito do que a de $Y_{[100]}^{\text{Ord}}$ (Figura 5.7), devido à alta alternância de $Y_{[120]}^{\text{Ord}}$ entre as categorias ‘1’, ‘2’ e ‘3’. Similarmente ao caso da variável binária correspondente, $Y_{[120]}$, analisada no Capítulo 4, o corte de PM10 a partir do limiar 120 faz com que $Y_{[120]}^{\text{Ord}}$ mude

5.5. Aplicação em poluição do ar

Tabela 5.8: Modelo estimado para $Y_{[120]}^{Ord}$.

Termo	Estimativa	E.P.	Valor t	$Pr(> t)$
θ_1	3,98	0,28	14,2	<0,001
θ_2	10,76	0,55	19,5	<0,001
$Y_{[120]}^{Ord}(t-1)2$	-2,14	0,20	-10,9	<0,001
$Y_{[120]}^{Ord}(t-1)3$	-3,84	0,46	-8,4	<0,001
$Y_{[120]}^{Ord}(t-2)2$	-0,51	0,19	-2,8	0,003
$Y_{[120]}^{Ord}(t-2)3$	-1,80	0,41	-4,4	<0,001
$SO2_t$	-1,23	0,09	-13,7	<0,001
$SO2_{t-1}$	0,54	0,08	6,6	<0,001
$NO2_t$	-0,17	0,02	-10,4	<0,001
$NO2_{t-2}$	0,08	0,01	6,0	<0,001
$\text{sen}(2\pi t/365)$	-0,68	0,12	-5,9	<0,001

de nível com uma frequência muito alta, dificultando a predição por qualquer modelo. Por este motivo é que se observa na Figura 5.8 (c) muitos erros de previsão. Nota-se também o “atraso” de um dia da série predita em relação à observada.

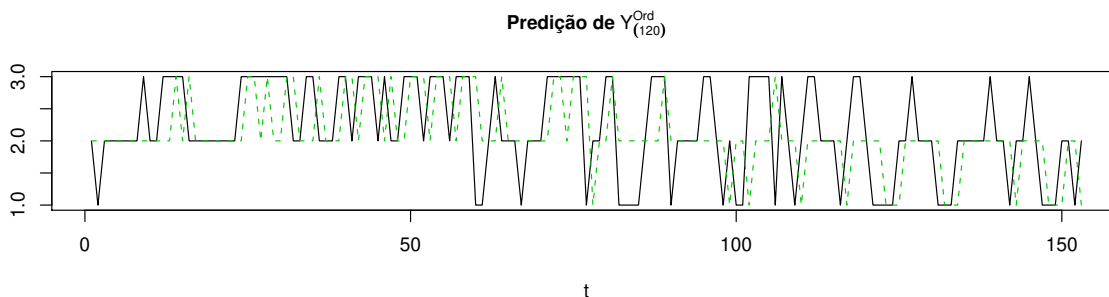


Figura 5.8: $Y_{[120]}^{Ord}$ observada (linha contínua) e predita (linha tracejada).

A performance preditiva do modelo pode ser melhor visualizada pela matriz de erros (Tabela 5.9) e pelas medidas de performance preditiva de maior interesse (Tabela 5.10, modelo “ M^{Ord} ”). As taxas de erros mais preocupantes, FN_{13} e FP_{31} , são bastante baixas, satisfazendo um dos quesitos desejados para a predição. Contudo, a acurácia do modelo é ruim - apenas 50%. E este resultado é consequência das baixas precisões das predições das classes “baixa” e “alta” poluição. Comparando com os resultados de predição do modelo para $Y_{[100]}^{Ord}$ (Tabela 5.7 (M^{Ord})), vemos que de fato o uso do corte em 120 dificulta

Capítulo 5. Modelo de Chances Proporcionais para Séries Categóricas Ordinais

mais a predição do que a utilização do corte em $100 \mu\text{g}/\text{m}^3$. Esta divergência entre as performances para cada caso fica bem evidenciada pela diferença de 10% observada entre suas acurácias.

Tabela 5.9: Matriz de erros para a predição de $Y_{[120]}^{\text{Ord}}$.

		Real			
		1	2	3	Total
Predito	1	6	9	1	16
	2	19	51	35	105
	3	3	9	20	32
	Total	28	69	56	153

Comparando com os modelos alternativos (Tabela 5.10), em termos de predição, vemos que o modelo final foi bem inferior ao impraticável: as acurácias diferem em 22%. Esta diferença é devida principalmente à melhora na precisão da poluição “alta” obtida pelo modelo impraticável, que equivale ao dobro da obtida pelo modelo final. A interpretação para este fato é a de que o uso dos valores contemporâneos reais de SO₂ e NO₂ propiciam uma predição muito mais acurada dos valores mais altos de PM₁₀ do que a utilização de seus valores preditos. Ainda sobre a performance do modelo impraticável, vale destacar as taxas nulas para os piores erros, FN₁₃ e FP₃₁.

Tabela 5.10: Comparação das performances preditivas de todos os modelos para $Y_{[120]}^{\text{Ord}}$.

Medida de perf. preditiva (%)	Impraticável (III)	Baseado somente no passado (I)	Normal, pós-categorizado (II)	Modelo final (M^{Ord})
Acurácia	0,72	0,41	0,59	0,50
Precisão (Baixa)	0,32	0,25	0,21	0,21
Precisão (Média)	0,88	0,01	0,83	0,74
Precisão (Alta)	0,71	0,96	0,48	0,36
FN13 (PIOR erro)	0,00	0,01	0,00	0,01
FN23 (2º pior)	0,10	0,00	0,19	0,23
FN12	0,02	0,06	0,02	0,06
FP31 (PIOR erro)	0,00	0,13	0,02	0,02
FP32 (2º pior)	0,03	0,39	0,06	0,06
FP21	0,12	0,01	0,12	0,12

Similarmente ao caso de $Y_{[100]}^{\text{Ord}}$ (Tabela 5.7 (II)), o modelo baseado em termos defasados

5.5. Aplicação em poluição do ar

teve um viés de predição, predizendo poluição alta para a maioria dos dias (133), e poluição média para apenas dois dias, o que levou às precisões de 96 e 1%, respectivamente. Por este motivo, sua performance foi inferior à do modelo final.

De todos os modelos utilizáveis na prática (Tabela 5.10, modelos I, II e M^{Ord}), o Normal (II) obteve a melhor performance preditiva. Sua precisão da predição dos dias de baixa poluição foi a mesma obtida pelo modelo final (M^{Ord}), mas as precisões para os dias de média e alta poluição foram superiores, de forma que sua acurácia foi 9% superior. Portanto, no caso do uso do limiar $120 \mu\text{g}/\text{m}^3$, o modelo de chances proporcionais não traz vantagens sobre o tradicional MLG Normal para PM10.

Com relação ao uso do modelo para objetivos específicos, vemos que o mesmo pode ser bem utilizado com o foco na economia, pelo fato das taxas de falsos-positivos serem baixas, menores que 7%.

Por último, numa comparação direta entre as performances preditivas do modelo logístico para a versão binária de PM10 (estudado no Capítulo 4) e do modelo de chances proporcionais para sua versão ordinal com 3 categorias, ficamos inclinados a concluir, a partir das colunas “M” e “ M^{Ord} ” das Tabelas 4.8, 4.12, 5.7 e 5.10, que o modelo logístico foi o mais bem sucedido. Isto porque as acurácias de predição dos modelos logísticos para $Y_{[100]}$ e $Y_{[120]}$ oscilaram em torno de 75%, e as dos modelos de chances proporcionais para $Y_{[100]}^{\text{Ord}}$ e $Y_{[120]}^{\text{Ord}}$ oscilaram, em média, ao redor de 55% (lembrando que a “acurácia” é uma medida de performance preditiva global). Entretanto, é importante observar que estamos comparando dois modelos destinados a dados de naturezas distintas, os binários e os categóricos ordinais com três categorias. Agora, adotando uma outra abordagem e utilizando um raciocínio simplista, podemos julgar que o modelo logístico teve uma performance muito boa, pois a utilização de um modelo “cego”² (ou *dummy*) para os dados binários, supondo classes balanceadas, resultaria em uma acurácia de 50%. Ou seja, a acurácia do modelo conseguido para os dados binários (M) foi 50% superior à que seria obtida pelo modelo cego. No caso ordinal, sob a suposição de classes balanceadas (onde cada categoria ocorreria com uma frequência de 1/3), o modelo cego teria uma acurácia de 33%.

²Modelo “cego” é aquele que prediz os valores de forma determinística e fixa, sem levar outras informações em consideração. No caso binário, o modelo cego seria aquele que prediz todos os valores como ‘1’, ou todos como ‘0’, inflexivelmente. Assim, sob a hipótese de classes balanceadas - 50% de valores ‘0’ e 50% de ‘1’s - no conjunto de predição, o modelo cego terá uma proporção geral de acertos (acurácia) de 50%.

E o modelo ordinal que conseguimos (M^{Ord}) obteve uma acurácia de 55%, um valor *67% acima do “esperado”*, de 33%, tomando o modelo cego como referência. Então, sob este ponto de vista, o modelo de chances proporcionais foi superior ao logístico, contrariando a primeira impressão que tivemos. Portanto, concluímos que a comparação direta entre as performances dos modelos logístico e de chances proporcionais não pode ser realizada de forma trivial; assim, por um critério de neutralidade, optamos por não apontar nenhum dos modelos como o mais bem sucedido na predição .

5.5.3 Conclusões da aplicação

Diante dos resultados apresentados, podemos concluir que:

- Na escolha de um modelo para a análise de dados temporais de poluição, onde as séries possuem memória longa, o modelo de chances proporcionais pode ser levado em consideração. Contudo, sua performance de predição irá depender dos limiares de referência utilizados para a categorização da série de interesse. Na análise dos dados de poluição de Santiago do Chile, a categorização a partir dos níveis 50 e 100 resultou em melhor performance preditiva do que o uso dos limiares 50 e 120 $\mu\text{g}/\text{m}^3$, porque o uso do limiar 120 gera uma série categórica com maior alternância entre seus valores assumidos em cada instante de tempo, dificultando a predição por qualquer modelo que fosse utilizado. Já o uso dos limiares 50 e 100 $\mu\text{g}/\text{m}^3$ propiciou uma boa performance preditiva, equivalente à do tradicional modelo de regressão Normal para PM10 em sua forma original, não-categorizada.
- O tratamento de dados de poluição como uma série categórica ordinal possui as vantagens de simplicidade na interpretação dos resultados e facilidade de implementação em um sistema autômato de alerta à população.
- A comparação das performances preditivas obtidas pelo modelo logístico no caso binário (Capítulo 4) e pelo modelo de chances proporcionais no caso ordinal não pode ser feita diretamente, de maneira simples. Os modelos são destinados a dados de naturezas distintas, sendo difícil estabelecer uma forma de comparação válida. Portanto, pelos resultados observados em ambos os casos e já destacados nas conclusões de cada um, optamos por não eleger um modelo sobressalente, e concluímos ape-

5.5. Aplicação em poluição do ar

nas que ambos obtiveram performances preditivas muito boas em seus respectivos contextos.

Regressão com Séries de Contagens

Modelos para séries temporais de contagens têm sido extensivamente estudados desde pelo menos os anos 1990. Zeger (1988), por exemplo, propôs um modelo orientado a parâmetro onde a correlação serial surge de um processo estacionário, autoregressivo de primeira ordem, não-observável, e a estimação dos parâmetros é feita via equação de estimação. Outras abordagens, que incluem por exemplo modelos de cadeia de Markov, modelos de espaço de estados, modelos autoregressivos condicionais e modelos baseados no operador *binomial thinning* para séries de valor inteiro, compõem a extensa bibliografia sobre modelos para séries temporais de contagens.

Dentro do contexto de séries de contagens com memória longa, entretanto, o número de trabalhos encontrados na literatura estatística é reduzido. Entre os estudos de modelos mais gerais, para séries não gaussianas, encontramos o de So (1999) e o de Brockwell (2007). So (1999) propõe uma extensão do modelo de espaço de estados clássico, de forma a comportar modelos para séries temporais não gaussianas com memória longa na equação de estado. Brockwell (2007) apresenta uma família de modelos generalizados para séries temporais com memória longa onde as observações têm uma determinada distribuição condicional, dado um processo ARFIMA latente. Agora, com relação a trabalhos voltados especificamente para o caso de contagens, temos os estudos de Quoreshi (2006a, 2006b). No primeiro, o autor propõe o modelo BINMA (*bivariate integer-valued moving average*), e sugere extensões deste modelo para a inclusão de variáveis explanatórias. No segundo trabalho, o modelo INARFIMA (*integer-valued ARFIMA*) é introduzido, combinando-se características do modelos INARMA e ARFIMA. Este modelo não permite a inclusão de

covariáveis. Por esses motivos, temos grande interesse em verificar se o modelo de regressão Poisson, estimado por máxima verossimilhança parcial, é adequado à modelagem de séries de contagens com memória longa, onde esta característica pode estar presente também nas séries covariáveis.

Neste capítulo, apresentaremos o modelo de regressão Poisson para séries temporais de contagens, com estimação via verossimilhança parcial, da forma como proposta por Kedem e Fokianos (2002). Um estudo de simulação será também realizado para a avaliação do comportamento assintótico do estimador de MVP, para se avaliar se suas propriedades de consistência e normalidade assintótica valem quando estão sendo modeladas séries com memória longa. Finalmente, o modelo é aplicado a dados do ramo financeiro.

As simulações e a modelagem dos dados financeiros foram realizadas no *software* livre R, versão 2.9.0 (R Development Core Team, 2009). Para a obtenção e o pré-tratamento dos dados de finanças, foram utilizados o *software* S-PLUS (S-PLUS 3.4 - Release 1, 1996) e seu pacote econométrico *HF* (*High Frequency*), de acordo com o procedimento exposto em Zivot (2005).

6.1 O modelo de regressão Poisson

Seja $\{Y_t\}$, $t = 1, \dots, N$, uma série temporal de contagens, aqui tomada como o processo de interesse ou resposta. Assumindo que as contagens seguem uma distribuição, condicionada ao passado, Poisson com média μ_t

$$f(y_t; \mu_t | \mathcal{F}_{t-1}) = \frac{\exp(-\mu_t) \mu_t^{y_t}}{y_t!}, \quad t = 1, \dots, N, \quad (6.1)$$

o **modelo de regressão Poisson** com função de ligação canônica $g(\mu_t) = \log(\mu_t)$ tem a forma

$$\log(\mu_t) = \eta_t = \mathbf{Z}'_{t-1} \boldsymbol{\beta}, \quad t = 1, \dots, N. \quad (6.2)$$

O mesmo é mais comumente expresso na forma

$$\mu_t(\boldsymbol{\beta}) = \exp(\mathbf{Z}'_{t-1} \boldsymbol{\beta}), \quad t = 1, \dots, N, \quad (6.3)$$

em termos da função de ligação inversa $h(\eta_t) = \exp(\eta_t)$.

6.2. Estimaco

Aqui, \mathcal{F}_{t-1} denota o “passado”, i.e., toda a informao do passado concernente à srie resposta e às sries covariveis conhecida pelo observador no momento t . Desta forma, o processo de covariveis, denotado pelo vetor p -dimensional $\{\mathbf{Z}_{t-1}\}$, $t = 1, \dots, N$, pode englobar valores passados da srie resposta e de sries covariveis. Por exemplo, \mathbf{Z}_{t-1} pode ser definido como

$$\mathbf{Z}_{t-1} = (1, X_t, Y_{t-1}, X_t Y_{t-1})',$$

onde $\{X_t\}$ é um processo adicional observado junto ao processo de interesse $\{Y_t\}$, e $X_t Y_{t-1}$ é um termo de interao. Para esta escolha, o modelo (6.3) é dado por

$$\mu_t(\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \beta_3 X_t Y_{t-1}), \quad t = 1, \dots, N, \quad (6.4)$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$.

6.2 Estimaco

Considerando o modelo (6.1)-(6.3), a funo de verossimilhana parcial é dada por

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^N f(y_t; \mu_t(\boldsymbol{\beta}) | \mathcal{F}_{t-1}) = \prod_{t=1}^N \frac{\exp[-\mu_t(\boldsymbol{\beta})] \mu_t^{y_t}(\boldsymbol{\beta})}{y_t!},$$

de onde se obtm a log-verossimilhana parcial

$$l(\boldsymbol{\beta}) = \sum_{t=1}^N \{y_t \mathbf{Z}'_{t-1} \boldsymbol{\beta} - \exp(\mathbf{Z}'_{t-1} \boldsymbol{\beta}) - \log y_t!\}. \quad (6.5)$$

Diferenciando (6.5) em $\boldsymbol{\beta}$, temos a funo score parcial,

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{Z}_{t-1} (Y_t - \exp(\mathbf{Z}'_{t-1} \boldsymbol{\beta})). \quad (6.6)$$

Obtm-se o estimador de mxima verossimilhana parcial $\hat{\boldsymbol{\beta}}$ pela soluo do sistema no-linear de equaes score, que é obtido igualando-se (6.6) ao vetor-nulo. Para sua soluo atravs do mtodo *scoring* de Fisher, resta encontrar a matriz de informao de $\boldsymbol{\beta}$. Por diferenciao direta de (6.6) ou, dado que se est fazendo uso da funo de ligao cannica, por meio de (2.16), obtm-se a matriz de informao condicional cumulativa

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^N \mathbf{Z}_{t-1} \mathbf{Z}'_{t-1} \exp(\mathbf{Z}'_{t-1} \boldsymbol{\beta}). \quad (6.7)$$

As expresses (6.6) e (6.7) so ento utilizadas no esquema iterativo (2.19), atravs de (2.18), para que se obtenha a estimativa de mxima verossimilhana parcial de $\boldsymbol{\beta}$.

6.3 Estudos de simulação

Nesta seção, simulamos um modelo de regressão Poisson com duas covariáveis, sendo uma de memória curta e a outra de memória longa. Analisamos primeiramente a estimação de uma única série simulada, para verificação da significância das estimativas e da qualidade dos resíduos. Em seguida, o modelo é simulado 1000 vezes, para se poder avaliar o comportamento assintótico dos estimadores de máxima verossimilhança parcial na presença de séries com memória longa.

6.3.1 Estudo detalhado de uma série simulada

Foram simuladas $N = 500$ observações do modelo

$$\mu_t(\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 X_t + \beta_2 W_t), \quad t = 1, \dots, N, \quad (6.8)$$

onde o processo de covariáveis $\mathbf{Z}_{t-1} = (X_t, W_t)'$ é composto por uma série de memória curta (X_t) e outra de memória longa (W_t), geradas de acordo com as escolhas

$$\begin{aligned} \{X_t\} &\sim AR(1), \text{ com } \phi = 0,8, \mu_\varepsilon = 10, \sigma_\varepsilon = 2, \quad \text{e} \\ \{W_t\} &\sim RF(d), \text{ com } d = 0,45, \mu_\varepsilon = 20, \sigma_\varepsilon = 3, \end{aligned} \quad (6.9)$$

onde μ_ε e σ_ε são a média e o desvio-padrão utilizados na geração do processo de inovações $\{\varepsilon_t\} \stackrel{iid}{\sim} \mathcal{N}(\mu_\varepsilon, \sigma_\varepsilon)$ de cada série covariável. Os parâmetros para a simulação foram especificados como

$$\boldsymbol{\beta} = (1,95, 0,025, 0,013)'. \quad (6.10)$$

A Figura 6.1 apresenta as séries geradas, suas FAC's e FACP's. A persistência de autocorrelações significativamente diferentes de zero para a FAC de Y_t , e a existência de valores maiores para os primeiros *lags* de sua FACP em comparação com os demais, indicam que Y_t “incorporou” memória longa, mesmo que pouca (Figuras 6.1 (h) e (i)). Estimando o parâmetro de diferenciação fracionária para Y_t por um método de máxima verossimilhança aproximada, obtemos $\hat{d} = 0,14$, com alta significância estatística.

Estimando o modelo a partir dos dados simulados, obtemos as estimativas dispostas na Tabela 6.1, onde os erros-padrões foram obtidos a partir da inversão de $\mathbf{G}_N(\boldsymbol{\beta})$ em

6.3. Estudos de simulação

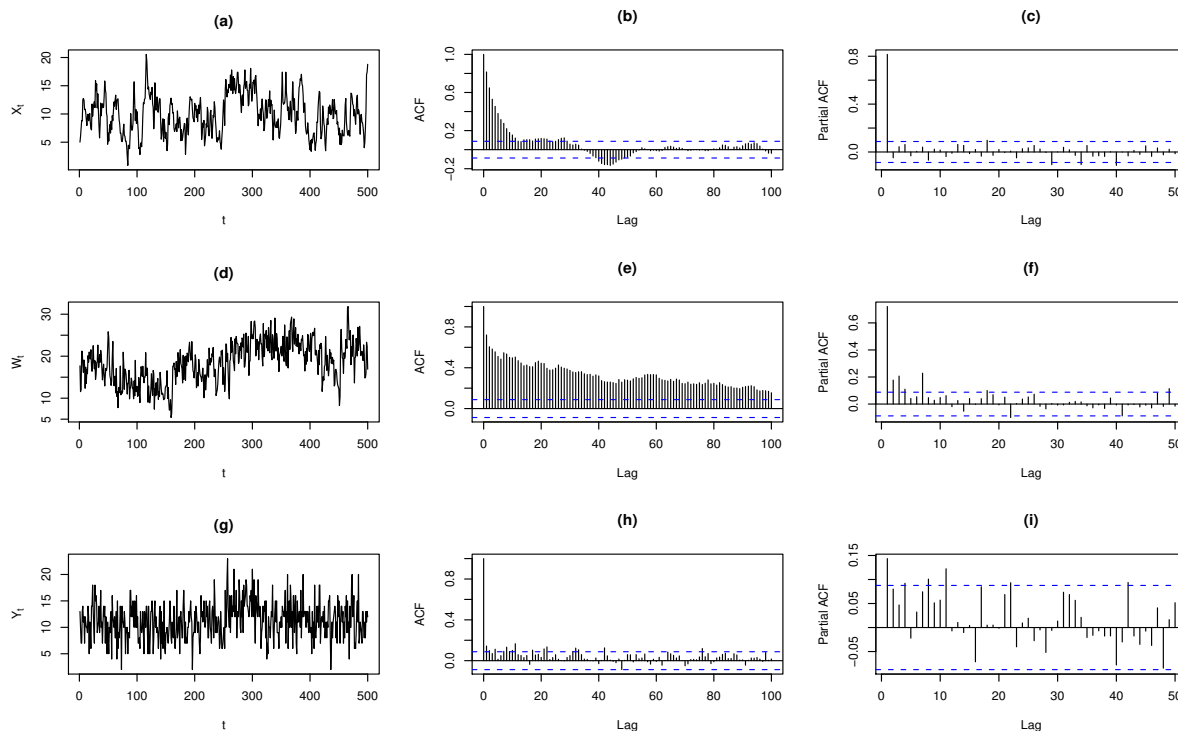


Figura 6.1: Séries simuladas: (a), (b) e (c) X_t e suas funções de autocorrelação e de autocorrelação parcial; (d), (e) e (f) W_t , sua FAC e FACP; (g), (h) e (i) Y_t , sua FAC e FACP.

(6.7). Vemos que a estimação foi bastante precisa e que os coeficientes estimados são todos altamente significativos.

Tabela 6.1: Resultados do ajuste aos dados simulados.

Parâmetro	Real	Estimativa	E.P.	Valor t	p -valor
β_0	1,95	1,98	0,063	31,5	<0,001
β_1	0,025	0,024	0,004	6,2	<0,001
β_2	0,013	0,011	0,003	3,9	<0,001

Interessa agora avaliar a qualidade dos resíduos, principalmente com relação à existência de autocorrelação para grandes defasagens. A partir da Figura 6.2 (a), vemos que os resíduos se comportam de forma satisfatória, sem nenhum padrão aparente e, mais do que isso, não apresentam autocorrelações significativas (Figuras 6.2 (b) e (c)). São também não-correlacionados com os valores ajustados (Figura 6.2 (f)) e, ainda, são distribuídos normalmente (Figura 6.2 (e)). Calculou-se também as correlações entre os resíduos e as

covariáveis, obtendo-se valores menores que 0,0005, em módulo.

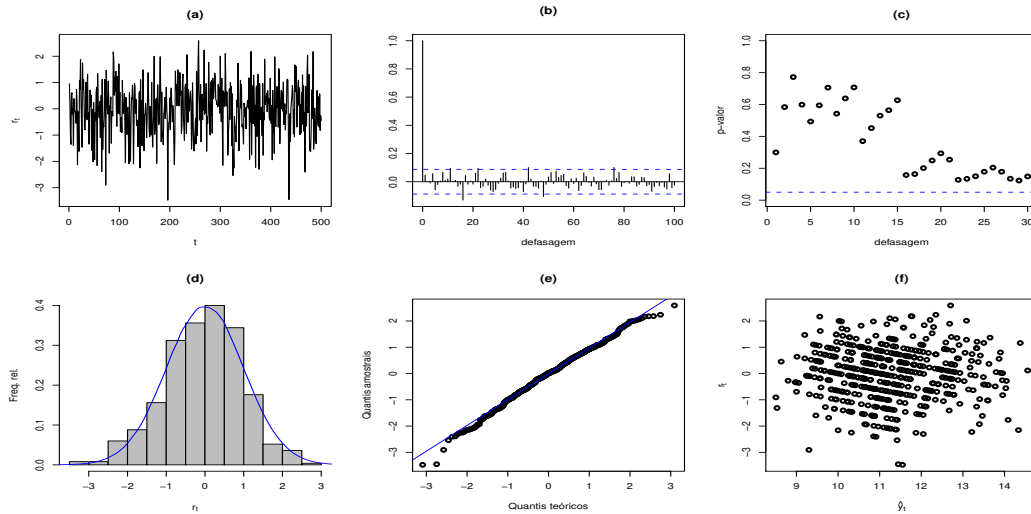


Figura 6.2: Resíduos do ajuste aos dados simulados: (a) Resíduos tipo componente da função desvio, (b) FAC, (c) níveis descritivos para o teste de Box-Ljung, (d) histograma de freq. rel. sobreposto pela curva da $\mathcal{N}(0, 1)$, (e) gráfico quantil-a-quantil com a Normal, (f) resíduos versus valores ajustados.

Portanto, pudemos confirmar a qualidade da estimação do modelo para a série simulada, mesmo com a presença de memória longa tanto em uma das séries covariáveis quanto na série resposta. Vale ressaltar apenas que ainda não existe na literatura comprovação teórica de que as técnicas usuais de diagnóstico para regressão se adequam a dados temporais como estes; entretanto, as análises e conclusões acima sugerem que a adequação, ou pelo menos uma adaptação, seja viável.

6.3.2 Estudo de simulação geral

O modelo (6.8) foi simulado 1000 vezes, variando o tamanho da série, N , entre 200, 500 e 1000, e o parâmetro de diferenciação fracionária, d , entre 0,2; 0,4; e 0,49. Foram calculadas as médias e os desvios-padrões amostrais (\hat{EP}) para cada seqüência de estimativas dos parâmetros, além de seus desvios-padrões teóricos aproximados, obtidos a partir de $\mathbf{G}_N^{-1}(\boldsymbol{\beta})$. A Tabela 6.2 apresenta os resultados das simulações, os quais evidenciam que:

- os estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, para todas as combinações de d e N , apresentaram um viés muito pequeno ou nulo, e que diminui com o aumento do tamanho amostral;

6.3. Estudos de simulação

isso sugere a *consistência* dos estimadores de máxima verossimilhança parcial;

- há uma diferença quase imperceptível entre os erros-padrões da simulação, \hat{EP} 's, e os teóricos, aproximados a partir de $\mathbf{G}_N^{-1}(\boldsymbol{\beta})$, o que indica que a obtenção dos erros-padrões dos parâmetros a partir da matriz de informação pode ser feita naturalmente, pois *não* fica invalidada pela presença de memória longa nas séries.

Tabela 6.2: Resultados das simulações (valor real: $\boldsymbol{\beta} = (1, 95, 0, 025, 0, 013)'$).

N	Estat.	d = 0, 2			d = 0, 4			d = 0, 49		
		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
200	$\hat{\boldsymbol{\beta}}$	1,945	0,0249	0,0132	1,951	0,0252	0,0128	1,948	0,0248	0,0132
	\mathbf{G}_N^{-1}	0,144	0,0056	0,0065	0,130	0,0060	0,0062	0,135	0,0061	0,0068
	\hat{EP}	0,145	0,0056	0,0065	0,124	0,0058	0,0060	0,138	0,0059	0,0070
500	$\hat{\boldsymbol{\beta}}$	1,948	0,0250	0,0131	1,951	0,0249	0,0130	1,947	0,0249	0,0132
	\mathbf{G}_N^{-1}	0,096	0,0043	0,0042	0,087	0,0043	0,0033	0,089	0,0041	0,0036
	\hat{EP}	0,095	0,0041	0,0042	0,084	0,0044	0,0033	0,087	0,0040	0,0035
1000	$\hat{\boldsymbol{\beta}}$	1,950	0,0250	0,0130	1,951	0,0250	0,0130	1,948	0,0250	0,0131
	\mathbf{G}_N^{-1}	0,063	0,0028	0,0028	0,059	0,0029	0,0025	0,054	0,0030	0,0023
	\hat{EP}	0,067	0,0028	0,0029	0,057	0,0029	0,0024	0,054	0,0029	0,0023

Com relação à distribuição assintótica dos estimadores, apresentamos na Figura 6.3 as seqüências de estimativas obtidas para $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ na simulação para a combinação que representaria, talvez, o caso mais difícil de se obter estimativas corretas, a de $N = 200$, amostra “pequena”, e $d = 0, 49$, “forte” memória longa. Podemos ver que nenhuma das 1000 estimativas, para nenhum dos parâmetros, apresentou grande viés de estimação, pois não vemos *outliers* nas seqüências de valores obtidos (Figuras 6.3 (a), (d) e (g)). Além disso, os gráficos de probabilidade normal evidenciam que a distribuição marginal dos estimadores é bem aproximada por uma distribuição gaussiana.

Desejando inspecionar a forma das distribuições *conjuntas bivariadas* das estimativas de $\hat{\boldsymbol{\beta}}$, utilizou-se o método de estimação de densidades bidimensionais por *kernel*¹. As funções densidade de probabilidade bivariadas estimadas, para o mesmo caso de $N = 200$ e $d = 0, 49$, com seus respectivos gráficos de contorno, estão dispostos na Figura 6.4. É difícil julgar qual seria a real distribuição de cada par de parâmetros simplesmente a partir

¹Implementado na função `kde2d` da biblioteca `{MASS}` do R.

da visualização das densidades estimadas. A quantidade de observações utilizada (neste caso, igual a 200) e o valor do parâmetro de suavização adotado para o *kernel*, têm grande influência sobre o resultado visual obtido para cada superfície estimada. Assim, é difícil julgar, por exemplo, os aspectos de existência de uma ou de mais modas e de decaimento exponencial da superfície, de forma que não se pode caracterizar a distribuição por detrás dos dados pela simples análise visual de suas densidades estimadas.

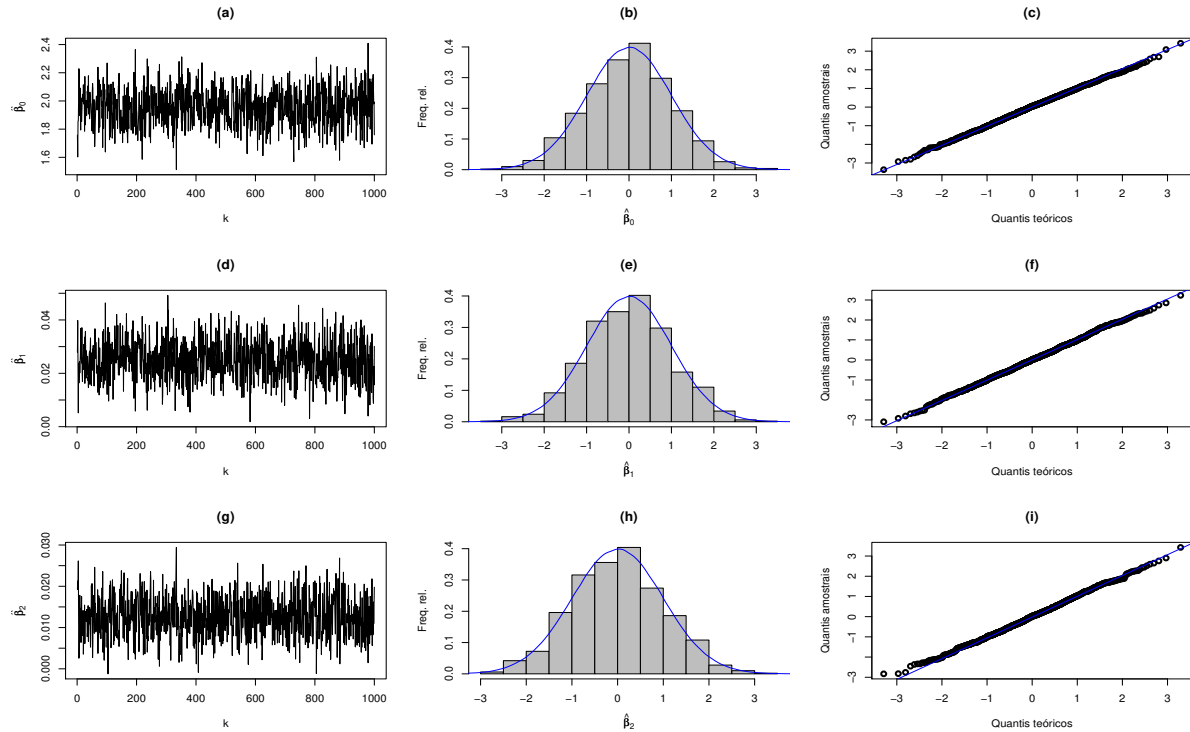


Figura 6.3: Seqüências de estimativas obtidas para $\hat{\beta}$, para o caso $N = 200$ e $d = 0, 49$; histogramas dos valores normalizados das mil estimativas, sobrepostos pela curva da densidade $\mathcal{N}(0, 1)$; gráficos quantil-a-quantil com os da $\mathcal{N}(0, 1)$: (a), (b) e (c) $\hat{\beta}_0$, (d), (e) e (f) $\hat{\beta}_1$, (g), (h) e (i) $\hat{\beta}_2$.

Contudo, sabemos que duas variáveis distribuídas segundo a Normal, se forem independentes, terão distribuição conjunta também Normal. É esta a idéia, aproximadamente, que temos a partir da superfície de probabilidade estimada para $\hat{\beta}_1$ e $\hat{\beta}_2$ (Figura 6.4). Para este par, a correlação não é nula, mas é baixa (-0,11), e sua densidade conjunta estimada lembra a de uma Normal bivariada.

Portanto, a partir das evidências obtidas nas simulações acima apresentadas, temos indícios de que o estimador de máxima verossimilhança parcial $\hat{\beta}$ do modelo Poisson, na

6.3. Estudos de simulação

presença de séries com memória longa, goza das mesmas propriedades de consistência e de Normalidade assintótica conhecidas para o estimador de máxima verossimilhança clássico, desde que as condições estabelecidas na Seção 2.5 do Capítulo 2 (pressuposições do modelo) estejam satisfeitas.

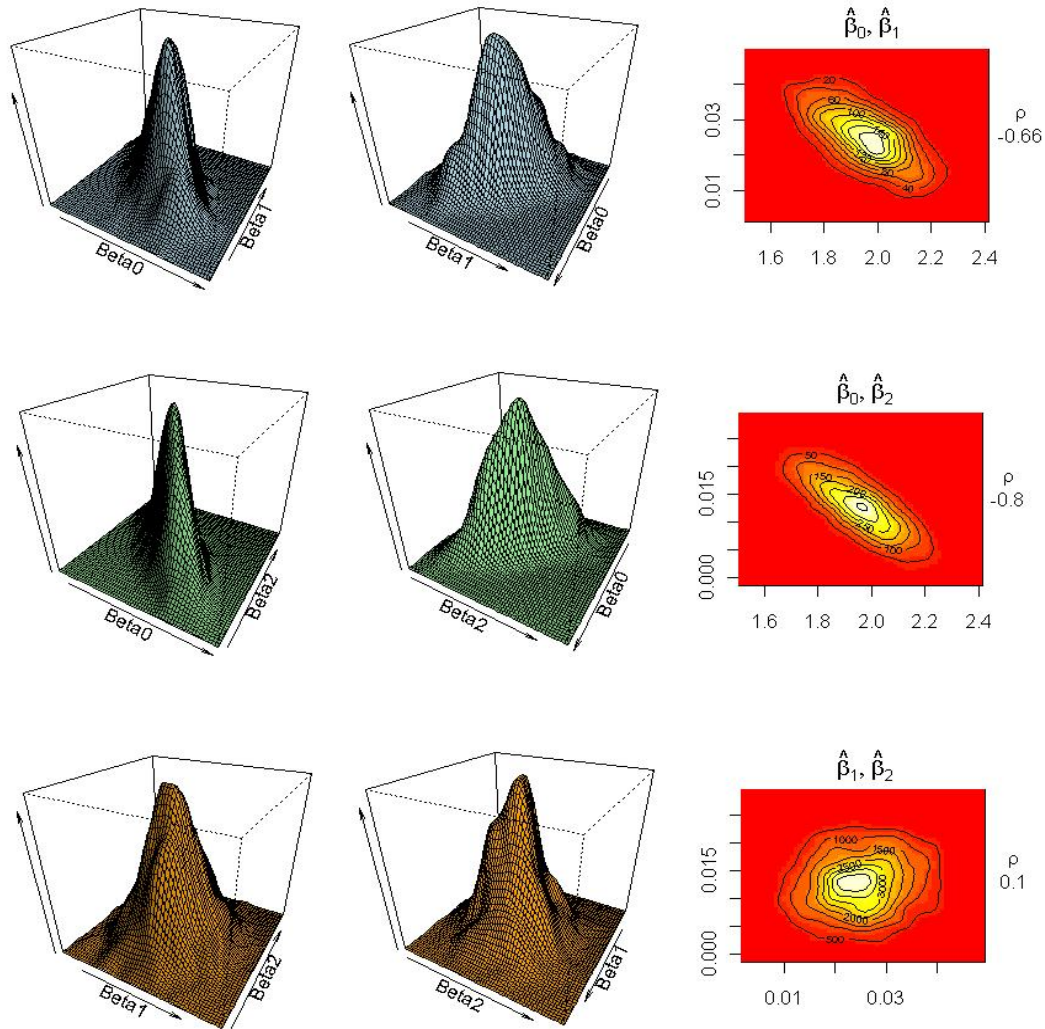


Figura 6.4: Estimativas das densidades conjuntas, par a par, dos valores obtidos para $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, para o caso $N = 200$ e $d = 0,49$, e correspondentes gráficos de contorno.

6.4 Aplicação a dados do setor financeiro

O modelo de regressão Poisson é utilizado nesta seção para a análise de dados de transações financeiras. Obteve-se uma qualidade de ajuste satisfatória e uma boa performance preditiva, levando à conclusão de que este modelo constitui uma alternativa boa e simples aos modelos mais sofisticados que são normalmente adotados no contexto financeiro.

Analistas financeiros estão geralmente interessados na modelagem do *preço* de determinado ativo financeiro em função da volatilidade de seus retornos. Contudo, o interesse pela modelagem do *volume* de negociações efetuadas, no lugar da própria série de preços, tem crescido consideravelmente, desde por exemplo os estudos de Karpoff (1987) e Schwert (1989). Desde então, o par de séries volume-volatilidade de transações tem sido alvo de estudo de diversos trabalhos, geralmente envolvendo modelos sofisticados, seja na modelagem conjunta das séries, seja na modelagem da série de volume individualmente.

Mais recentemente, Bollerslev and Jubinski (1999) e Lobato and Velasco (2000) ressaltaram a existência de memória longa nas séries do par volume-volatilidade, medidos tanto diária quanto intradiariamente. Em virtude destes fatos, visando a oferecer uma análise mais simples e direta do volume de transações em função da volatilidade, escolhemos dados desta natureza para a aplicação do modelo de regressão Poisson.

Os dados utilizados correspondem à série de valores intradiários de câmbio entre as moedas Euro e Dólar Americano, efetuados ao longo do período de 7 dias, na semana de 11 a 17 de Março de 2001 (domingo a sábado)². Da mesma forma como considerado por Zivot (2005), as observações foram limitadas aos cinco dias úteis da semana, de maneira a se retirar o efeito do fim de semana sobre o comportamento das séries. Assim, o período de observação considerado vai das 22h00 do domingo às 22h00 da sexta-feira.

A partir da série de valores, efetuou-se a contagem do *número de transações ocorridas* em cada intervalo de *5 minutos*, Y_t , $t = 1, \dots, 1440$. Com base nesta série de contagens, obteve-se a série de estimativas da *volatilidade realizada* para cada intervalo de tempo, V_t , de acordo com o procedimento exposto em Zivot (2005) e implementado na biblioteca *HF (High Frequency)* do *software* S-PLUS.

Nesta análise, desejamos modelar a série do volume de transações Y_t em termos da série de volatilidades V_t . Assim, deseja-se obter o modelo da forma (6.3) que melhor se ajuste aos dados, onde o vetor de covariáveis \mathbf{Z}_{t-1} incluirá V_t e, possivelmente, defasagens

²Dados obtidos a partir de Zivot (2005).

6.4. Aplicação a dados do setor financeiro

de Y_t e de V_t , além de termos determinísticos. Para o ajuste, foram utilizados os dados dos dias de domingo a quinta-feira, somando 1176 observações. Para a previsão, foram considerados os dados de sexta-feira, contendo 264 observações.

6.4.1 Ajuste

As séries do volume de transações e da volatilidade estão dispostas na Figura 6.5, juntamente com suas funções de autocorrelação e de autocorrelação parcial. Verifica-se o padrão sazonal diário em ambas as séries, cujo período vale 288^3 . Observa-se também o decaimento hiperbólico das autocorrelações, e a magnitude das primeiras autocorrelações parciais das séries, evidenciando a existência de memória longa. Nas Figuras 6.5 (a) e (b) está indicado o instante de partição das séries para a separação entre dados de ajuste (domingo a quinta-feira) e dados para a previsão (sexta-feira).

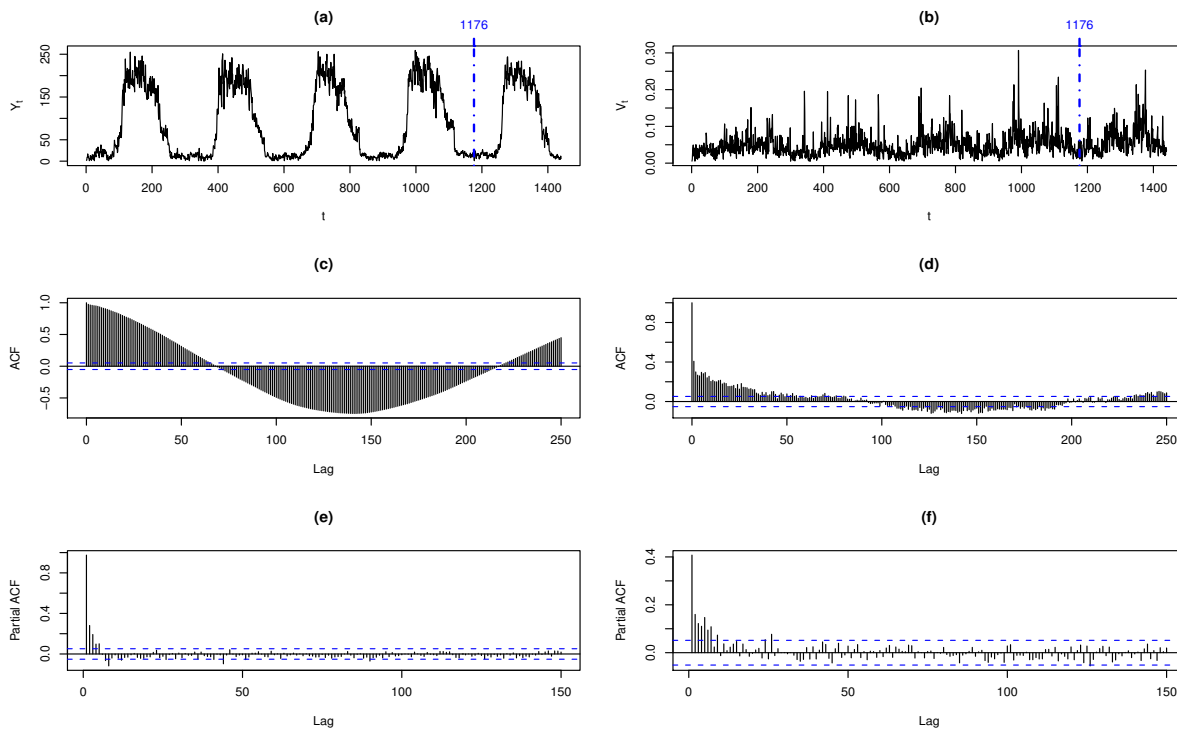


Figura 6.5: Séries financeiras: (a), (c) e (e) Volume de transações (Y_t) e suas funções de autocorrelação e de autocorrelação parcial, (b), (d) e (f) Volatilidade realizada (V_t) e suas FAC e FACP.

³1 dia = 24 horas, 1 hora = 12×5 minutos; portanto, 1 dia = $24 \times 12 = 288$ intervalos de 5 minutos.

Um aspecto da série de volume que merece destaque é seu comportamento dentro de cada dia da semana. Os altos valores de transação que se vê, oscilando em torno de uma média de 190 transações, correspondem ao período do “dia”, enquanto que os valores mais baixos, variando ao redor de 15 transações, correspondem ao período da “noite”. Portanto, será útil na análise a criação de uma variável que indique o período do dia em que a série se encontra, de forma a diferenciar o volume principalmente entre os períodos diurno e noturno.

Para tal, a variável *periodo* foi criada a partir dos intervalos de horário indicados na Tabela 6.3, onde constam também a média e o desvio-padrão de Y_t em cada período, e as correlações entre Y_t e V_t por período. A visualização dos períodos é mais fácil a partir da Figura 6.6 (a), que ilustra o comportamento de Y_t para o dia de terça-feira. Verifica-se que o volume médio de negociações efetuadas durante o dia é em torno de 14 vezes maior do que o da noite (Tabela 6.3). Além disso, é notável a menor correlação entre Y_t e V_t que se observa no período diurno, que chega a equivaler a aproximadamente à metade do valor das correlações para os outros períodos do dia. Este fato é devido à alta instabilidade do volume durante o dia, o que gera uma maior dispersão da série nesta faixa de horários. Tal comportamento está ilustrado na Figura 6.6 (b).

Tabela 6.3: Variável *periodo*, e a média e o desvio-padrão do volume, além das correlações entre o volume e a volatilidade, por período do dia.

Período do dia	Faixa de horário	Volume médio (desvio-padrão)	$\rho(Y_t, V_t)$
“outro”	06h00 às 07h00	50 (16)	0,42
DIA	07h00 às 16h00	186 (31)	0,21
“outro”	16h00 às 19h00	81 (28)	0,39
NOITE	19h00 às 06h00	13 (7)	0,38

Uma prática comum na modelagem do Volume de transações em função da Volatilidade é a utilização, no lugar desta, do logaritmo da volatilidade (log-volatilidade) pelo fato de normalmente apresentar uma variância mais homogênea do que a série original. De fato, como vemos na Figura 6.7 (d), sua variação é mais homogênea do que a da própria série de volatilidade (Figura 6.7 (a)). Além disso, sua distribuição se aproxima mais da distribuição Normal do que a de V_t (Figura (f)), e sua correlação com Y_t (0,47) é levemente superior à de V_t com Y_t (0,42). Estas mesmas conclusões se aplicam à série resultante da

6.4. Aplicação a dados do setor financeiro

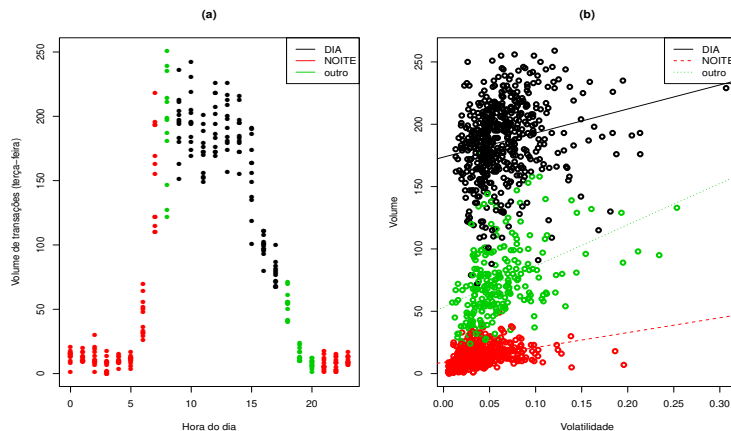


Figura 6.6: (a) Volume de transações observado na terça-feira, evidenciando os períodos do dia demarcados pela variável *periodo*. (b) Gráficos de dispersão e retas de ajuste de regressão entre Y_t e V_t , separadamente para cada período do dia.

transformação simples⁴ de Box-Cox de V_t , $V_t^{(\lambda)}$, utilizando a potência $\lambda = 0,19$, que por sua vez é bastante similar à série da (log)volatilidade (Figura 6.7 (g)). A propósito, a distribuição de $V_t^{(\lambda)}$ é mais simétrica e é a que mais se aproxima da Normal (Figuras 6.7 (h) e (i)). Por esses motivos, e por ser habitual nas modelagens em finanças, utilizamos em nossa análise $Z_t = \log(V_t)$ em substituição à série original V_t .

Utilizando-se de Z_t e de suas defasagens, das defasagens de Y_t , e da variável *periodo*, diversos modelos da forma (6.3) foram testados para a modelagem de Y_t , obtendo-se como melhor em termos de qualidade de ajuste e de resíduos, o modelo com os termos dispostos na Tabela 6.4. Os termos mais importantes para a explicação do volume de transações são o volume dos 5 minutos anteriores, a log-volatilidade contemporânea e o período do dia. Completam o modelo o par senoidal, para o ajuste da sazonalidade, as interações do volume anterior e de um termo senoidal ambas com o período, e um termo que descreve o efeito da semana sobre o fim de semana (efeito dos dias úteis). Este termo binário foi utilizado em lugar do dia da semana porque os efeitos médios de cada dia são quase todos idênticos.

A interpretação dos efeitos principais do modelo é bem condizente com o que se esperava a partir dos gráficos das séries (Figura 6.5). O volume de transações aumenta com o aumento da volatilidade, do volume anterior, e em dias úteis, em relação ao domingo.

⁴Versão simples da transformação, sem a utilização da média geométrica das observações.

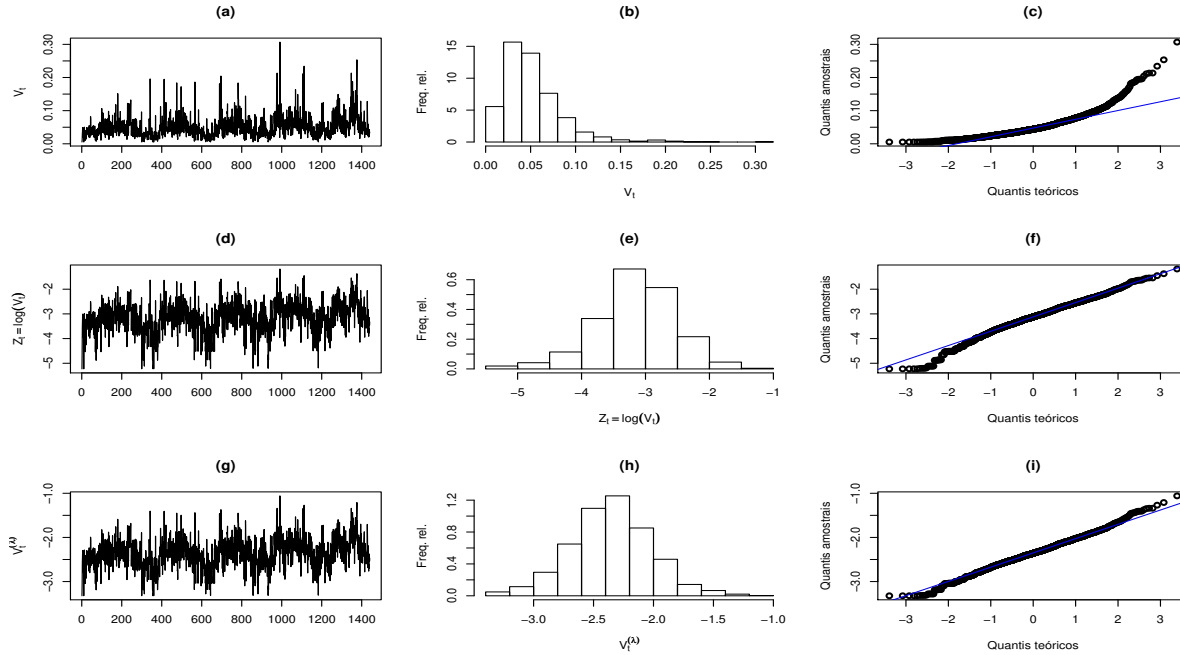


Figura 6.7: Volatilidade e suas transformações: série, histograma de freqüências e gráfico quantil-a-quantil com os da $\mathcal{N}(0, 1)$. Volatilidade (V_t) - (a), (b) e (c); log-volatilidade (Z_t) - (d), (e) e (f); volatilidade transformada pela transformação Box-Cox ($V_t^{(\lambda)}$) - (g), (h) e (i).

Tabela 6.4: Modelo estimado para o volume de transações, Y_t .

Termo	Estimativa	E.P.	Valor t	$Pr(> t)$
Intercepto	2,56	0,089	28,8	<0,001
Y_{t-1}	0,02	0,001	21,1	<0,001
Z_t	0,15	0,006	23,3	<0,001
$I_{[\text{dia útil}]}(t)$	0,43	0,078	5,5	<0,001
$I_{[\text{periodo=outra}]}(t)$	1,05	0,036	29,3	<0,001
$I_{[\text{periodo=DIA}]}(t)$	1,94	0,037	52,1	<0,001
$\text{sen}(2\pi t/288)$	-0,09	0,009	-9,8	<0,001
$\text{cos}(2\pi t/288)$	-0,31	0,026	-12,1	<0,001
$Y_{t-1} \cdot I_{[\text{periodo=outra}]}(t)$	-0,01	0,001	-12,6	<0,001
$Y_{t-1} \cdot I_{[\text{periodo=DIA}]}(t)$	-0,02	0,001	-18,6	<0,001
$\text{cos}(2\pi t/288) \cdot I_{[\text{periodo=outra}]}(t)$	0,22	0,032	6,9	<0,001
$\text{cos}(2\pi t/288) \cdot I_{[\text{periodo=DIA}]}(t)$	0,003	0,028	0,1	0,92

Todas as estimativas dos efeitos são altamente significativas, com exceção ao coeficiente da interação entre o termo sinusoidal e o período diurno (Tabela 6.4). Além disso, vale

6.4. Aplicação a dados do setor financeiro

ressaltar que a explicação dos dados pelo modelo é satisfatória, evidenciada pela redução obtida no valor da função desvio. Para o modelo “nulo”, que conta apenas com o intercepto e pode ser tomado como uma referência, o valor da função desvio é de 92924. Para o modelo apresentado, para o qual se obteve a maior redução, o valor é de 2827.

A qualidade do ajuste pode ser percebida em termos visuais, a partir da Figura 6.8. O grau de proximidade entre as séries observada e predita sobrepostas é notável, e sugere visualmente que o modelo foi capaz de explicar o volume médio de transações de maneira satisfatória. Verifica-se que os resíduos componentes do desvio são aproximadamente Normais, apesar de terem uma leve assimetria à esquerda (Figura 6.9 (b)). São também não-correlacionados, em qualquer período do dia, com os valores ajustados, com o volume do instante anterior e com a volatilidade contemporânea (Figuras 6.9 (c), (e) e (f)).

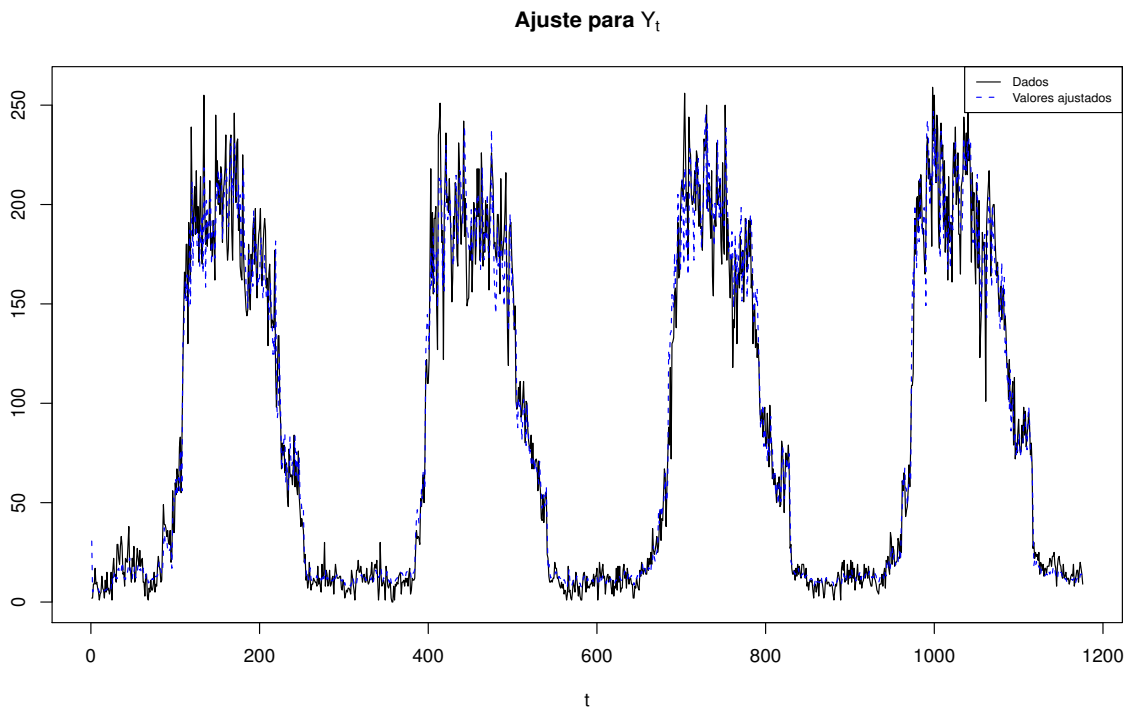


Figura 6.8: Volume de transações observado e ajustado.

Desejando avaliar a existência de autocorrelação serial nos resíduos, quisemos visualizar tanto os resíduos componentes do desvio quanto os de resposta. Os primeiros apresentam valores altos e significativos para as primeiras autocorrelações totais e parciais, de forma que não constituem uma seqüência ruído branco (Figuras 6.10 (a), (c), (e)

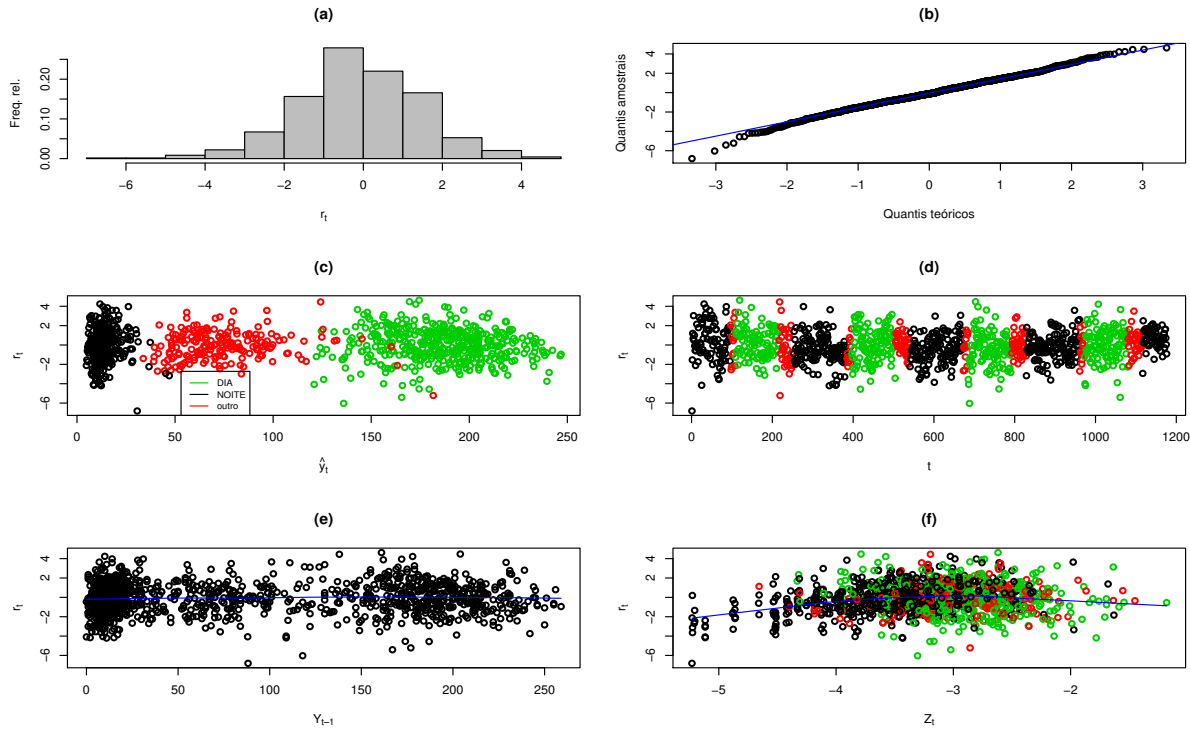


Figura 6.9: Resíduos do ajuste: (a) histograma dos resíduos tipo componentes da função desvio (\hat{d}_t), (b) Q-Q plot, (c) dispersão entre \hat{d}_t e os valores ajustados \hat{y}_t , (d) \hat{d}_t versus t , (e) \hat{d}_t versus Y_{t-1} , (f) \hat{d}_t versus Z_t .

e (g)). Já os segundos apresentam menor quantidade de autocorrelações significativas, e estas em menor magnitude, de forma que se comportam de maneira mais próxima à de um processo ruído branco, mas ainda significativamente diferente (Figuras 6.10 (b), (d), (f) e (h)). O teste de Box-Ljung não rejeita a hipótese de ausência de correlação apenas para as duas primeiras defasagens dos resíduos de resposta. Portanto, o modelo se ajustou bem aos dados, mas não foi capaz de explicar toda a dependência existente entre as observações do volume de transações.

6.4.2 Predição

Uma vez que o modelo ajustado depende do valor da log-volatilidade contemporânea, Z_t , faz-se necessária a modelagem e a previsão dos valores desta, para utilização na previsão dos valores de Y_t . Assim, a série de log-volatilidade foi modelada em seis etapas, da forma como descrita na seção 4.4 do Capítulo 4.

6.4. Aplicação a dados do setor financeiro

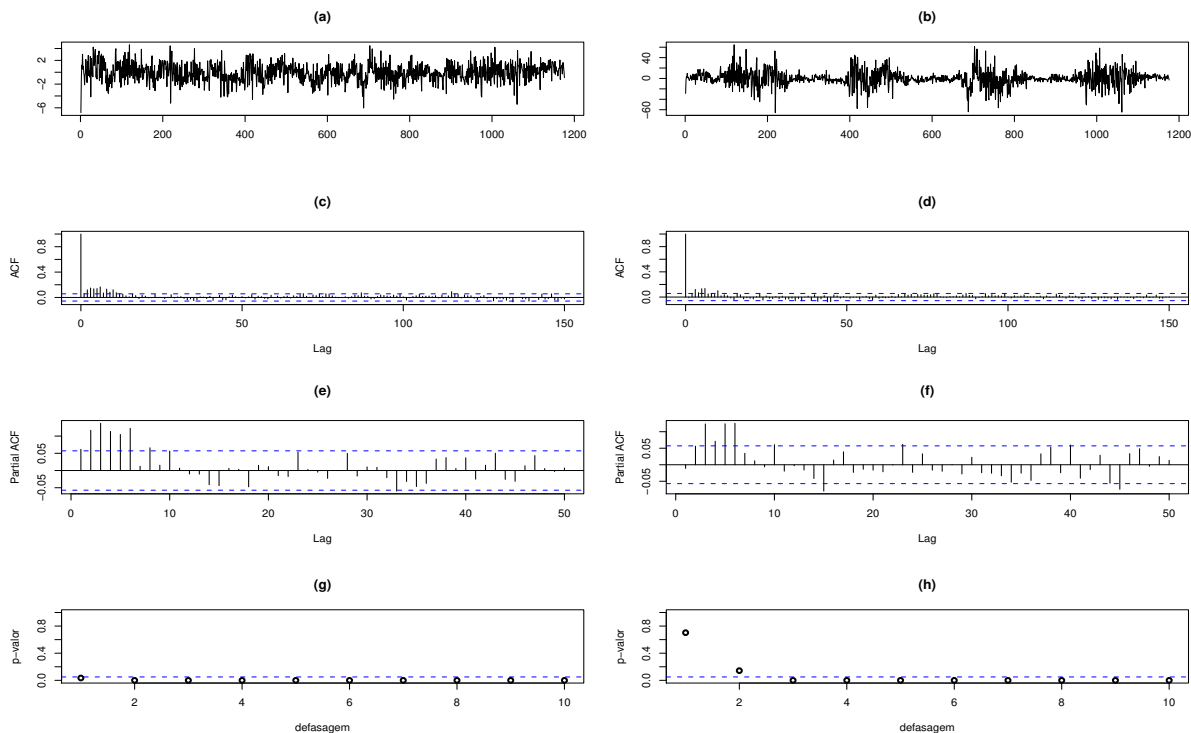


Figura 6.10: Resíduos do ajuste - avaliação da existência de correlação serial. Resíduos do desvio \hat{d}_t (a), suas funções de autocorrelação e de autocorrelação parcial (c) e (e), e níveis de significância para o teste de Ljung-Box (g); Resíduos de resposta \hat{e}_t (b), suas FAC e FACP (d) e (f), e níveis de significância para o teste de Ljung-Box (h).

O modelo de regressão linear para o ajuste da parte regular de Z_t é basicamente composto por termos sinusoidais, harmônicos (termos sinusoidais de maior frequência) e pela indicadora do período do dia, sendo todos os termos altamente significativos. Para a parte irregular foi utilizado um ruído fracionário, com $\hat{d} = 0,18$ (EP < 0,001), que foi aproximado por um modelo AR(40), da forma como apresentada na seção 3.4 do Capítulo 3, para se fazer a predição. A série predita obtida ao final do processo está apresentada na Figura 6.11. Verifica-se que ela acompanha bem a média do processo Z_t real, porém, oscila numa amplitude muitas vezes menor.

A série Z_t predita foi então utilizada para a predição de Y_t , a partir do modelo apresentado (Tabela 6.4). O volume predito e seu intervalo de predição estão dispostos na Figura 6.12. Utilizou-se aqui o intervalo de predição (IP) *aproximado* para a predição a partir do modelo Poisson. Para um nível de confiança de 95%, este IP é dado por $\hat{\mu}_t \pm 1,96\sqrt{\hat{\mu}_t}$,

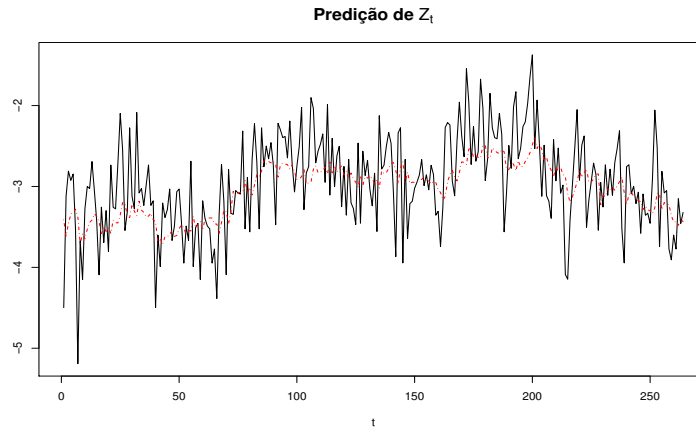


Figura 6.11: Log-volatilidade observada (linha contínua) e predita (tracejada).

uma vez que a média e a variância da distribuição condicional de Poisson são iguais. Contudo, vale destacar que para o período da noite, $\hat{\mu}_t$ assume valores pequenos, e portanto a aproximação para o intervalo é um pouco inferior.

A Figura 6.12 (a) evidencia que foi mais difícil prever no período diurno, onde a oscilação de Y_t é maior. Mas, de uma forma geral, o comportamento da série predita se aproximou bastante do comportamento do volume observado, de forma que a predição é visualmente satisfatória. Confirmando a qualidade da predição através do IP (Figura 6.12 (b)), temos uma cobertura de 84%, que é baixa em relação à esperada, de 95%. Para o IP de 99% de confiança, a cobertura é de 92%. Apesar das coberturas observadas terem sido um pouco inferiores às esperadas, consideramos que o resultado da predição é bom, por tratar-se de um modelo simples utilizado para séries relativamente difíceis de modelar.

Com o objetivo de fazer uma última comparação para avaliar a performance preditiva do modelo, fez-se a predição do volume utilizando os valores *reais* da log-volatilidade, ao invés de seus valores preditos. Desejou-se verificar se a modelagem preditiva de Z_t foi ruim no sentido de prever valores que seriam ineficientes para a previsão de Y_t . Procedendo desta forma, obtivemos a predição apresentada na Figura 6.13. Na figura, as linhas suavemente tracejadas indicam os instantes de tempo onde o IP de 95% de confiança não abrangeu o volume real. Podemos ver que a predição foi mais difícil para o modelo nos períodos “outro” e diurno, provavelmente devido às mudanças de regime da série, e por ela se comportar de maneira mais instável nestes períodos. De fato, as coberturas do IP em cada período equivalem a 93% no período noturno, 87% no diurno

6.4. Aplicação a dados do setor financeiro

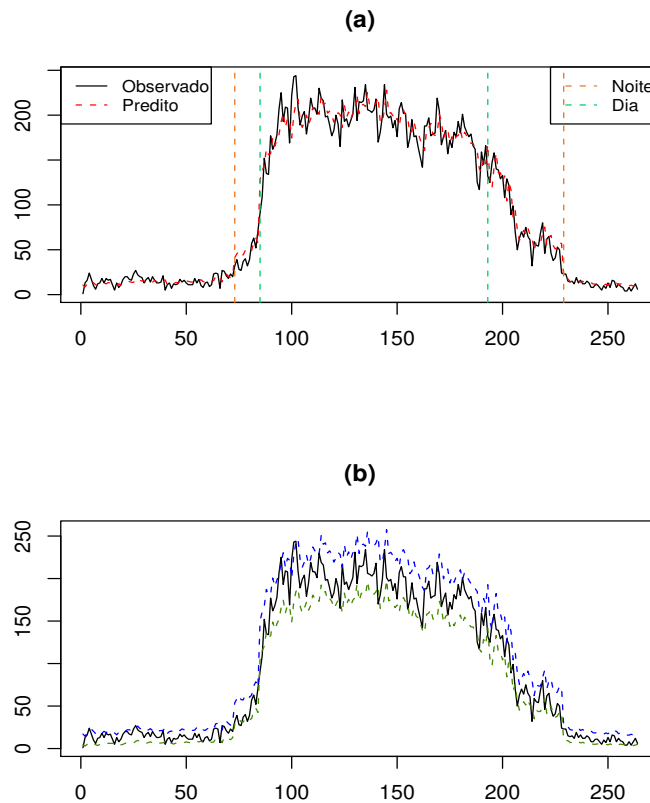


Figura 6.12: Predição de Y_t : (a) volume predito e observado, (b) intervalo de predição aproximado de 95% de confiança.

e 71% no denominador como “outro”. Por fim, em relação a todo o período de predição, a cobertura do IP foi de 86%, apenas 2% acima da obtida com o uso dos valores preditos de Z_t , e portanto concluímos que a modelagem e predição de Z_t se mostrou adequada para a utilização na predição de Y_t .

6.4.3 Conclusões da aplicação

Na modelagem da série de volume de transações em função da log-volatilidade através do modelo de regressão Poisson, obtivemos uma qualidade de ajuste pouco satisfatória, pois os resíduos apresentaram autocorrelações significativas para algumas defasagens. Entretanto, a performance preditiva do modelo foi boa. Os intervalos de predição aproxi-

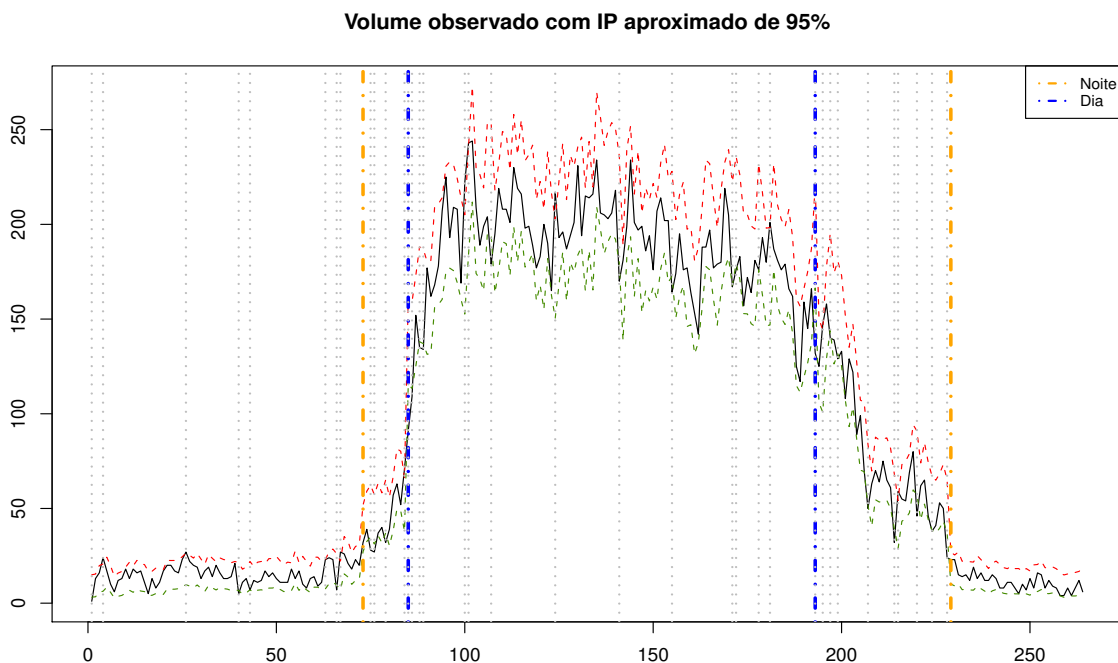


Figura 6.13: Predição de Y_t utilizando os valores reais de Z_t .

mados de 95% e 99% de confiança obtiveram coberturas de, respectivamente, 84% e 92% dos valores reais. Apesar destas diferenças, julgamos como boa a performance preditiva do modelo, considerando sua simplicidade em face da dificuldade de modelagem que estas séries financeiras oferecem. As vantagens na utilização de um modelo simples como este residem na flexibilidade de modelagem e na facilidade de utilização e de interpretação. Vale destacar ainda que a modelagem da série covariável se mostrou bastante eficiente, uma vez que a predição do volume foi praticamente a mesma quando da utilização dos valores preditos da log-volatilidade em substituição aos observados. Por tudo isso, concluímos que o modelo de regressão Poisson pode ser bem considerado para a análise de séries financeiras com memória longa, e apresenta a vantagem da simplicidade em relação a outros modelos mais sofisticados geralmente utilizados neste contexto.

Conclusões e considerações finais

Esta dissertação tem duas contribuições relacionadas à modelagem de séries temporais via Modelos Lineares Generalizados proposta por Kedem e Fokianos (2002). A primeira contribuição é avaliar mediante simulações se a estimação por máxima verossimilhança parcial (MVP) funciona quando as séries apresentam memória longa (ML). A segunda contribuição é um estudo do poder preditivo dessa classe de modelos quando as séries covariáveis precisam ser modeladas. Isto acontece quando, no modelo ajustado, a variável resposta depende de valores contemporâneos das covariáveis.

Com relação à primeira contribuição, obtiveram-se evidências de que a estimação por verossimilhança parcial do MLG pode ser confiavelmente efetuada para séries com memória longa. Os estudos de simulação realizados forneceram evidências de que: (i) os estimadores de máxima verossimilhança parcial (EMVP) são consistentes, mesmo para grandes valores do parâmetro de memória longa das séries; (ii) a presença de ML nas séries não invalida a obtenção dos erros-padrões corretos dos EMVP a partir da inversa da matriz de informação condicional cumulativa, que é a matriz de informação usada no caso da verossimilhança parcial; (iii) a distribuição assintótica dos EMVP é normal. Por estes fatos, vemos que a inferência clássica de máxima verossimilhança pode ser prontamente realizada para os estimadores de MVP do modelo.

Estas constatações foram obtidas para três casos específicos de MLGs para séries não gaussianas: o modelo logístico para séries binárias, o modelo de chances proporcionais para séries categóricas ordinais e o modelo Poisson para séries de contagens. Contudo, os resultados podem ser estendidos a outros casos, como o modelo de regressão Normal e o

modelo Gama para séries contínuas.

A segunda contribuição, da avaliação do poder preditivo dos modelos, foi realizada através de aplicações a dados reais. Os modelos logístico e de chances proporcionais foram aplicados a dados de poluição do ar, para a análise das versões binária e categórica ordinal da série de material particulado (PM10), e o modelo Poisson foi aplicado a dados financeiros, para a modelagem de uma série de volume de transações em função do logaritmo da volatilidade. Pelo fato de todos os modelos ajustados dependerem de valores contemporâneos das séries covariáveis, foi necessário fazer a modelagem e predição à parte de cada uma delas, para então usar seus valores preditos na predição da série de interesse. Tanto na aplicação dos modelos logístico e de chances proporcionais aos dados de poluição do ar, quanto na aplicação do modelo Poisson aos dados financeiros, obteve-se boa performance preditiva. Nos casos binário e categórico ordinal, a performance preditiva foi avaliada através de medidas de performance usuais calculadas a partir da matriz de confusão da predição. Para o modelo Poisson, avaliou-se a qualidade das predições por meio da porcentagem de cobertura obtida pelo intervalo de predição.

Ainda com relação às aplicações, vale ressaltar que o modelo logístico para a versão binária de PM10 obteve performance preditiva igual ou superior à do modelo de regressão Normal para PM10 contínua, não-categorizada. Além disso, o usuário final do modelo poderá adaptar a regra utilizada nas predições de acordo com seu objetivo de utilização, de maneira a reduzir a taxa de erro de classificação que for mais preocupante.

Portanto, a partir das constatações acima, concluímos que os MLGs sob estimação por MVP constituem uma boa alternativa para a modelagem de séries temporais não gaussianas com memória longa. Diante da escassez de modelos encontrados na literatura para dados desta sorte, vale destacar seus pontos fortes. Suas maiores vantagens são sua simplicidade, flexibilidade de modelagem e o fato de as rotinas de estimação já estarem implementadas em diversos *softwares* estatísticos padrão, como o R, o SAS e o SPSS.

Como sugestões de trabalhos futuros, propomos:

1. A demonstração teórica dos resultados assintóticos observados para o estimador de máxima verossimilhança parcial na presença de memória longa;
2. A realização de estudos de simulação de modelos para casos ainda mais raros de séries temporais com memória longa, como o caso Gama.

Referências

- Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. John Wiley & Sons, Inc., New Jersey.
- Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics*, 73, 5-59.
- Benjamin, M. A., Rigby, R. A. and Stasinopoulos, M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, 98, 214-223.
- Bollerslev, T. and Jubinski, D. (1999). Equity trading volume and volatility: latent information arrivals and common long-run dependencies. *Journal of Business & Economic Statistics*, 17(1), 9-21.
- Box, G. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd edition. Holden-Day, San Francisco.
- Brockwell, A. (2007). Likelihood-based analysis of a class of generalized long-memory time series models. *Journal of Time Series Analysis*, 28(3), 386-407.
- Chan, N. H. and Palma, W. (1998). State space modeling of long-memory processes. *Annals of Statistics*, 26, 719-740.
- Cordeiro, G. M. (1992). *Introdução à Teoria de Verossimilhança*. Livro texto do 10^o Simpósio Nacional de Probabilidade e Estatística, Rio de Janeiro, 1992.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 69-76.
- Fahrmeir, L. and Kaufmann, H. (1987). Regression models for non-stationary categorical time series. *Journal of Time Series Analysis*, 8(2), 147-160.

- Fahrmeir, L. (1992). State space modelling and conditional mode estimation for categorical time series. In D. R. Brillinger *et al.*, editor, *New Directions in Time Series*. Springer, New York, pp. 87-110.
- Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1, 15-29.
- Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, 68, 165-176.
- Hung, Y., Zarnitsyna, V., Zhang, Y., Zhu, C. and Wu, C. F. J. (2008). Binary time series modeling with application to adhesion frequency experiments. *Journal of the American Statistical Association*, 103(483), 1248-1259.
- Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 16, 770-799.
- Hurst, H. E. (1957). A suggested statistical model of time series that occur in nature. *Nature*, 180, 494.
- Karpoff, J. M. (1987). The relation between price changes and trading volume: a survey. *The Journal of Financial and Quantitative Analysis*, 22(1), 109-126.
- Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*. John Wiley & Sons, Inc., New Jersey.
- Levine, M. and Moore, G. E. (2009). A time series model of the occurrence of gastric dilatation-volvulus in a population of dogs. *BMC Veterinary Research*, 5, 12.
- Liang, K. Y. and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*, 84(406), 447-451.
- Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303.
- Lobato, I. N. and Velasco, C. (2000). Long memory in stock-market trading volume. *Journal of Business & Economic Statistics*, 18(4), 410-427.
- Mazucheli, J., Louzada-Neto, F., Guirado, L. e Martinez, E. Z. (2008). Algumas medidas do valor preditivo de um modelo de classificação. *Revista Brasileira de Biometria*, 26(2), 83-91.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman and Hall, London.

-
- National Academy of Engineering and National Research Council. (2008). *Energy Futures and Urban Air Pollution: Challenges for China and the United States*. The National Academies Press, Washington, D.C. ISBN-13: 978-0-309-11140-9.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370-384.
- Ostro, B., Sanchez, J. M., Aranda, C. and Eskeland, G. S. (1995). *Air Pollution and Mortality: Results from Santiago, Chile*. The World Bank, Policy Research Working Paper (No. 1453), Washington D.C.
- Palma, W. (2007). *Long-Memory Time Series*. John Wiley & Sons, Inc., New Jersey.
- Palma, W. and Zevallos, M. (2010). Fitting non-gaussian persistent data. *Applied Stochastic Models in Business and Industry*. (No prelo.)
- Quoreshi, S. (2006a). Bivariate time series modelling of financial count data. *Communications in Statistics. Theory and Methods*, 35, 7.
- Quoreshi, S. (2006b). A long memory, count data, time series model for financial application. *Umeå Economic Studies*, 673.
- R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Schwert, G. W. (1989). Why does stock market volatility change over time? *The Journal of Finance*, 44(5), 1115-1153.
- Seater, J. J. (1993). World temperature - Trend uncertainties and their implications for economic policy. *Journal of Business and Economic Statistics*, 11, 265-277.
- Slud, E. and Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4, 89-106.
- So, M. K. P. (1999). Time series with additive noise. *Biometrika*, 86(2), 474-482.
- S-PLUS 3.4, Release 1 (1996). *Insightful Corporation*, Seattle, WA. URL <http://www.insightful.com>.
- Stoffer, D. S., Tyler, D. E. and Mcdougall, A. J. (1993). Spectral analysis for categorical time series: scaling and the spectral envelope. *Biometrika*, 80(3), 611-622.

- West, M., Harrison, P. J. and Migon, H. (1985). Dynamic generalized linear model and Bayesian forecasting (with discussion). *Journal of the American Statistical Association*, 80(389), 73-97.
- Wong, W. H. (1986). Theory of partial likelihood. *Annals of Statistics*, 14, 88-123.
- Yee, T. H. (2008). *The {VGAM} Package*. R News, 8(2-2), 28-39. URL <http://CRAN.R-project.org/doc/Rnews/>.
- Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75(4), 621-629.
- Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, 44, 1019-1031.
- Zhen, X. and Basawa, I. V. (2009). Estimation for binary models generated by Gaussian autoregressive processes. *Journal of Statistical Computation and Simulation*, 1563-5163.
- Zivot, E. (2005). *Analysis of High Frequency Financial Data: Methods, Models and Software*. Minicurso para a 11ª Escola de Séries Temporais e Econometria. Vila Velha, ES, Agosto de 2005.