**Predição de fator de simultaneidade
através de modelos de regressão
para proporções contínuas**

Luiz Fernando Molinari Zerbinatti

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Área de concentração: Estatística

Orientadora: Prof^a. Dr^a. Silvia Lopes de Paula Ferrari

São Paulo, 29 de Fevereiro de 2008

Predição de fator de simultaneidade através de modelos de regressão para proporções contínuas

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Luiz Fernando Molinari Zerbinatti
e aprovada pela comissão julgadora.

São Paulo, 29 de fevereiro de 2008

BANCA EXAMINADORA

- PROFA. DRA. SILVIA LOPES DE PAULA FERRARI (ORIENTADORA) - IME/USP
- PROFA. DRA. LUCIA PEREIRA BARROSO - IME/USP
- PROFA. DRA. OLGA SATOMI YOSHIDA - IPT

AGRADECIMENTOS

- A DEUS, PELA VIDA E PELA SAÚDE.
- À MINHA ORIENTADORA, PROFESSORA SILVIA FERRARI, QUE ME CONDUZIU DE FORMA SEGURA E ACREDITOU NESTE TRABALHO. NA MINHA CURTA EXPERIÊNCIA ACADÊMICA, FORAM POUCAS AS VEZES QUE ENCONTREI TANTO ENTUSIASMO, DEDICAÇÃO E HUMILDADE EM UM PROFESSOR.
- À OLGA, QUE ACREDITOU NA MINHA CAPACIDADE, CONCEDENDO A OPORTUNIDADE DE CONVIVER DURANTE 18 MESES COM PROFISSIONAIS DE ELEVADÍSSIMO NÍVEL TÉCNICO, ALÉM DE SUGERIR E VIABILIZAR ESTA DISSERTAÇÃO.
- À PROFA. LÚCIA, QUE ENRIQUECEU ESTE TRABALHO COM SUAS VALIOSAS SUGESTÕES.
- A TODOS OS COLEGAS DO CMF / IPT - CENTRO DE METROLOGIA DE FLUIDOS DO INSTITUTO DE PESQUISAS TECNOLÓGICAS, QUE FIZERAM GRANDE PARTE DESTES TRABALHOS ACONTECER - CARLOS EDUARDO, ELAINE, MARCELO, NILSON, TERAHATA, AOS PROFISSIONAIS DA COMGÁS ENVOLVIDOS NESTE PROJETO.
- AOS MEUS PAIS. ESTE TRABALHO É FRUTO DO AMOR E DA CONFIANÇA INCONDICIONAIS QUE ELES SEMPRE DEPOSITARAM EM MIM.
- À JULCILEI QUE, À SUA MANEIRA, CONTRIBUIU PARA ESTE TRABALHO.
- AO *teacher* RENATO, PELO *help* NO *abstract*.
- AOS COLEGAS DA UNICAMP E DA USP, QUE SEMPRE ME AJUDARAM NO DIA A DIA DO MUNDO ACADÊMICO. UM AGRADECIMENTO ESPECIAL À PATRÍCIA ESPINHEIRA, QUE CONTRIBUIU DECISIVAMENTE PARA A CONCLUSÃO DESTES TRABALHOS.
- AOS FUNCIONÁRIOS DA UNICAMP E DA USP, PRINCIPALMENTE A TODOS OS MEUS PROFESSORES, SEMPRE EXCEPCIONAIS NA NOBRE MISSÃO DE ENSINAR, CONTRIBUINDO DE FORMA IMENSURÁVEL COM A MINHA FORMAÇÃO. UM AGRADECIMENTO ESPECIAL AO PROFESSOR FILIDOR, QUE ME INICIOU NA PESQUISA ACADÊMICA.

RESUMO

O fator de simultaneidade é fundamental no planejamento de redes de distribuição de gás natural. Trata-se de um multiplicador entre 0 e 1 que ajusta o consumo total teórico de um número de aparelhos de utilização em condições reais. Em 2005 o Instituto de Pesquisas Tecnológicas (IPT) e a Companhia de Gás de São Paulo (COMGÁS) realizaram um estudo no qual determinou-se o fator de simultaneidade em um conjunto de edificações residenciais. Um modelo de regressão foi proposto para expressar o fator de simultaneidade em termos da potência total instalada. O modelo ajustado pode ser utilizado para prever o fator de simultaneidade em novas edificações. O modelo em questão é um modelo de regressão linear normal no qual a variável resposta é o logaritmo do fator de simultaneidade.

Nesta dissertação, o objetivo é investigar outras possibilidades de modelos de regressão adequados aos dados obtidos pelo IPT e pela COMGÁS. Especial atenção é dada ao modelo de regressão beta proposto por Ferrari e Cribari-Neto (Journal of Applied Statistics, 2004) por possuir vantagens sobre o modelo de regressão linear normal. O modelo de regressão beta assume que, dadas as covariáveis, a variável resposta possui distribuição beta, sendo adequado para modelar dados observados no intervalo unitário. Desta forma, a transformação na variável resposta - o fator de simultaneidade - é desnecessária.

Além disso, é proposta uma nova abordagem para a predição do fator de simultaneidade, diferente de todas as abordagens pesquisadas na literatura, utilizando a técnica de *bootstrap*.

Palavras-chaves: Modelo de regressão beta; Bootstrap; Proporções contínuas; Diagnóstico; Redes de distribuição de gás; Predição; Modelos de regressão; Fator de simultaneidade.

ABSTRACT

The simultaneity factor is fundamental in planning gas distribution networks. It is a multiplier between 0 and 1 that adjusts the theoretical total consumption of a number of devices to realistic conditions. In 2005, the Instituto de Pesquisas Tecnológicas (IPT) and the Companhia de Gás de São Paulo (COMGÁS) performed a study in which the simultaneity factor of gas consumption in a set of residential buildings have been determined. A regression model was proposed to express the simultaneity factor in terms of the total power of installed equipment. The fitted model can be used to predict the simultaneity factor in new buildings. The model they proposed is a normal linear regression model in which the response variable is the logarithm of the simultaneity factor.

In the present dissertation, our aim is to investigate other possible regression models suitable to the data obtained by IPT and COMGÁS. Emphasis is given to the beta regression model proposed by Ferrari and Cribari-Neto (Journal of Applied Statistics, 2004) which has a number of advantages over normal linear regression models. The beta regression model assumes that, given the covariates, the response variable has a beta distribution, which is adequate to model data observed in the unit interval. Therefore, no transformation in the response variable, the simultaneity factor, is needed.

Additionally, we present a new approach for the prediction of the simultaneity factor, that is different from all the approaches shown in the literature, using the bootstrap technique.

Keywords: Beta regression model; Bootstrap; Continuous Proportions; Diagnostic; Gas distribution network; Prediction; Regression models; Simultaneity factor.

Sumário

1	Introdução	1
1.1	Organização da dissertação	4
2	Apresentação do problema	5
2.1	O sistema predial de gás natural	5
2.2	Fator de simultaneidade	7
2.3	Coleta de dados	11
2.4	Análise descritiva	13
3	Descrição dos modelos estatísticos	15
3.1	Modelo de regressão linear normal	16
3.2	Modelo de regressão não-linear normal	20
3.3	Modelo de regressão beta	22
3.4	Técnicas de diagnóstico	28
4	Análise dos dados	33
4.1	MRLN com transformação logito na variável resposta	34
4.2	MRLN com transformação logito na variável resposta e transformação logarítmica na variável regressora	37
4.3	MRLN com transformação logarítmica na variável resposta	40
4.4	MRLN com transformação logarítmica nas variáveis resposta e regressora	43
4.5	Modelo de regressão beta	46
4.6	Modelo de regressão beta com transformação logarítmica na variável regressora	49
4.7	Comparação entre os modelos	50

5	Predição do fator de simultaneidade	58
5.1	Introdução	58
5.2	O método <i>bootstrap</i>	59
5.3	Intervalos de predição <i>bootstrap</i>	60
6	Considerações finais	64
6.1	Conclusões	64
6.2	Sugestões para trabalhos futuros	65
A	Dados	67

Capítulo 1

Introdução

As primeiras reservas brasileiras de gás natural começaram a ser exploradas na Bahia, na década de 40. Inicialmente restrito a aplicações industriais, observou-se no decorrer da década uma tímida expansão do uso do gás em estados vizinhos bem como uma pequena diversificação de suas aplicações, passando a ser utilizado também como combustível para refinarias e pólos petroquímicos. Apenas 40 anos depois, na década de 80, assistiu-se ao primeiro marco importante na história do gás natural no Brasil, com a descoberta da bacia de Campos. Finalmente o combustível atingiu participação significativa na matriz energética brasileira. Mas a bacia de Campos não foi suficiente para impulsionar a popularização do combustível país afora.

A forte expansão do uso do gás natural que se observa hoje no Brasil é devida à conjunção de fatos bastante recentes da história da América do Sul. De um lado, a descoberta de amplas reservas de gás natural na região - incluindo-se aí a bacia de Santos - e a conclusão do gasoduto Brasil-Bolívia elevaram consideravelmente a disponibilidade do combustível no país. Do outro, a demanda pelo combustível também vem crescendo. Consumidores industriais, comerciais e residenciais estão percebendo que, além do preço competitivo, o gás natural oferece vantagens logísticas e ambientais. O recente colapso na rede elétrica do país e as incertezas sobre a disponibilidade futura deste tipo de energia também elevaram a demanda do combustível, que se apresentou como a alternativa de melhor custo-benefício no médio prazo. Por estas razões, hoje os governantes atribuem peso

elevado ao gás natural na política energética nacional. Aliado aos interesses comerciais das empresas concessionárias, o Estado vem promovendo uma forte política de expansão do gás natural no país. Não por acaso, as companhias de gás buscam formas de reduzir os custos de instalação para o gás natural ao mesmo tempo em que se assiste à expansão do uso deste combustível. Esta é uma necessidade criada pela recente política de massificação do gás natural no país, que transformou o custo de instalação em uma variável de peso nas contas das companhias, e fez com que as oportunidades de melhoria no dimensionamento da rede recebessem atenção especial dentro destas empresas.

O fator de simultaneidade insere-se no contexto de oportunidades de melhoria no dimensionamento da rede. Os textos normativos referentes às instalações prediais de gás natural definem o fator de simultaneidade como sendo "a relação percentual entre a potência verificada praticamente, com que trabalha simultaneamente um grupo de aparelhos, servidos por um determinado trecho de tubulação, e a soma da capacidade máxima de consumo desses mesmos aparelhos". Este número possui vital importância para o dimensionamento das redes de distribuição de gás, pois é a partir dele que se obtém a vazão de projeto, ou seja, a vazão máxima que provavelmente se pratica em um trecho da tubulação. Para obter a vazão de projeto em um trecho da tubulação, o fator de simultaneidade é aplicado sobre a vazão máxima possível deste mesmo trecho, de modo que

$$Q_p = F \times Q_{max},$$

onde Q_p é a vazão de projeto, F é o fator de simultaneidade e Q_{max} é a vazão máxima possível. Como o fator de simultaneidade é um número adimensional situado no intervalo (0,1), pode-se interpretar a vazão de projeto como um percentual da vazão máxima possível. Em outras palavras, o fator de simultaneidade representa o percentual da vazão máxima possível que provavelmente se pratica naquele trecho de tubulação.

Ilha (1996) explica o impacto de superestimar o fator de simultaneidade. A superestimativa do fator de simultaneidade implica não somente no superdimensionamento das tubulações internas, dos reguladores de pressão e dos medidores utilizados em um edifício residencial, - a chamada rede de distribuição interna - mas também nas estimativas de

demanda de gás natural e, conseqüentemente, na rede pública de abastecimento como um todo. Superestimar o fator de simultaneidade é gerar custos desnecessários às companhias de gás. É exatamente neste ponto em que se encontra uma fonte potencial de economia no desenvolvimento de projetos de instalação predial. Kawabe, Ogassavara e Dias (1989) mostraram indícios de que o fator de simultaneidade estava sendo superestimado com a antiga curva utilizada pela Companhia de Gás de São Paulo (COMGÁS) e publicada em Comgás (1977). Posteriormente a COMGÁS reformulou os seus procedimentos para o cálculo de vazões de projeto, que culminou em uma nova curva proposta por Carvalho (1995), e hoje publicada também na norma técnica NBR 14570.

A metodologia empregada no desenvolvimento da curva apresentada por Carvalho (1995) representou significativa redução nas estimativas do fator de simultaneidade, mas ainda sobram margens para novas melhorias. Trabalhando neste sentido, o Instituto de Pesquisas Tecnológicas (IPT) e a COMGÁS desenvolveram um novo trabalho com uma extensa campanha de medição com o objetivo de levantar fatores de simultaneidade empíricos, relacioná-los com suas respectivas vazões máximas possíveis e, assim, propor um novo modelo estatístico que relacionasse fator de simultaneidade com vazão máxima possível¹. De fato, foi verificado que o fator de simultaneidade podia ser otimizado, conforme publicado em IPT (2005). Parte dos dados deste trabalho foi disponibilizado pela COMGÁS para o desenvolvimento desta dissertação de mestrado.

O objetivo central desta dissertação consiste em analisar o problema de predição do fator de simultaneidade à luz de diversos modelos estatísticos. Particularmente, dentre os modelos considerados, especial atenção é dada ao modelo de regressão beta proposto por Ferrari e Cribari-Neto (2004) que, teoricamente, se adequa melhor à natureza dos dados coletados. Além disso, uma vez determinado o modelo mais adequado aos dados, pretende-se propor uma nova abordagem para a predição do fator de simultaneidade a partir da demanda máxima possível, diferente de todas as abordagens pesquisadas na literatura, utilizando a técnica de *bootstrap*.

¹O autor da presente dissertação foi um dos integrantes da equipe técnica responsável pela condução do trabalho de campo e pelo desenvolvimento dos modelos estatísticos deste projeto.

1.1 Organização da dissertação

Esta dissertação está organizada em seis capítulos. No Capítulo 2 encontra-se a apresentação do problema. Nele estão contidos os principais conceitos técnicos utilizados nesta dissertação, a descrição detalhada do problema e a descrição do trabalho de coleta de dados realizado pelo IPT.

No Capítulo 3 são apresentados os fundamentos teóricos dos modelos estatísticos considerados neste trabalho. Conforme será visto, só há interesse em modelos que, de alguma forma, são empregados com sucesso na modelagem de variáveis resposta cujos valores estão situados no intervalo $(0,1)$. As bases teóricas das técnicas de diagnóstico também são apresentadas neste capítulo. As técnicas de diagnóstico são um conjunto de ferramentas utilizado na avaliação da qualidade do ajuste de modelos estatísticos, e tiveram papel fundamental nesta dissertação.

O Capítulo 4 apresenta a aplicação da teoria apresentada no capítulo anterior. Cada seção mostra os resultados do ajuste do modelo, bem como uma análise de diagnóstico deste ajuste. Na última seção, apresenta-se uma comparação entre os resultados obtidos, subsidiando assim a escolha do modelo mais adequado aos dados.

No Capítulo 5 propõe-se uma curva do fator de simultaneidade. A proposta desta nova curva é feita com base no modelo escolhido no Capítulo 4, empregando a técnica do *bootstrap*, uma abordagem diferente de todas aquelas encontradas na literatura.

Finalmente, no Capítulo 6, são apresentadas as conclusões e recomendações deste trabalho, bem como sugestões de trabalhos futuros.

Capítulo 2

Apresentação do problema

2.1 O sistema predial de gás natural

Antes de apresentar o problema, é conveniente compreender a terminologia empregada em um sistema predial de gás natural. Estas definições foram estabelecidas pela Associação Brasileira de Normas Técnicas (ABNT) e estão publicadas na norma técnica NBR 14570.

- Rede geral: rede pública de abastecimento de gás.
- Rede de alimentação: trecho de tubulação que conecta a rede pública à rede de distribuição interna.
- Rede de distribuição interna: conjunto de tubulação e acessórios situado dentro dos limites da propriedade dos consumidores, após os estágios de regulação de pressão.
- Regulador de pressão de primeiro estágio: equipamento utilizado para reduzir a pressão do gás antes de sua entrada na rede primária. Dependendo da pressão da rede de alimentação, pode ser desnecessário.
- Regulador de pressão de segundo estágio ou estágio único: equipamento utilizado para reduzir a pressão do gás antes de sua entrada na rede secundária a níveis adequados ao funcionamento dos aparelhos de utilização.

- Rede primária: trecho de tubulação compreendido entre os reguladores de primeiro e segundo estágio.
- Rede secundária: trecho de tubulação compreendido entre o regulador de segundo estágio e os pontos de utilização.
- Ponto de utilização: extremidade da tubulação à qual se conecta um aparelho de utilização a gás.
- Medidor: equipamento empregado na medição de consumo do gás natural, instalado ao longo da rede secundária.

Uma instalação predial de gás natural típica, bem como a sua integração com a rede geral, pode ser observada na Figura 2.1.

Figura 2.1: Esquema típico de um sistema predial de gás natural. Fonte: Ilha (1996)

O projeto e a execução das instalações da rede de distribuição interna de gás devem obedecer às normas conforme a NBR 14570. Nesta norma estão contidas as determinações sobre instalação, abrigo de medidores e reguladores, tubos e conexões, segurança e dimensionamento das tubulações, entre outras. Esta norma não se aplica somente às instalações prediais de gás abastecidas por canalização de rua (gás natural) mas também às aquelas abastecidas por uma central de gás (gás liquefeito).

2.2 Fator de simultaneidade

Antes de definir o conceito do fator de simultaneidade, é conveniente entender algumas relações existentes entre as unidades de grandeza de potência e vazão. Quando se fala em gás combustível, vazão e potência se inter-relacionam por meio do poder calorífico. O poder calorífico é uma propriedade física inerente a qualquer combustível e é definido como a quantidade de energia que se desprende na combustão completa de uma unidade de volume ou massa. No caso dos gases, o poder calorífico é usualmente expresso em $kcal/m^3$. A potência, por sua vez, é definida como a energia produzida ou consumida por unidade de tempo. Pode-se convenientemente expressar a potência como $kcal/h$, e assim verificar facilmente a relação

$$Q = \frac{P}{PC} \quad (2.1)$$

em que Q é a vazão nominal de um aparelho (m^3/h), P é a potência nominal de um aparelho ($kcal/h$) e PC é o poder calorífico do gás ($kcal/m^3$). Esta relação será útil nos cálculos de projeto de um sistema predial de gás combustível, conforme se verá a seguir.

Uma vez esclarecida a relação entre potência e vazão, pode-se compreender melhor o conceito e a definição do fator de simultaneidade. É consenso entre os projetistas que não é necessário projetar o sistema predial de gás combustível de modo a atender a capacidade

máxima de consumo da edificação. Basta que o sistema esteja preparado para atender à demanda máxima praticada na edificação. É usual mensurar a capacidade máxima de consumo e a demanda máxima praticada em unidades de potência, e denominá-las, respectivamente, como *potência computada* e *potência adotada*.

Os textos normativos estrangeiros e nacionais sugerem que a potência adotada seja obtida pelo método do fator de simultaneidade. Segundo a norma NBR 14570, o fator de simultaneidade é uma "relação percentual entre a potência verificada praticamente, com que trabalha simultaneamente um grupo de aparelhos, (...) e a soma da capacidade máxima de consumo desses mesmos aparelhos". Portanto, para a obtenção da potência adotada de um sistema predial de gás, segue-se o seguinte procedimento:

- apura-se a potência computada C a ser instalada em um determinado trecho, somando-se as potências nominais dos aparelhos de utilização supridos por este trecho;
- determina-se o fator de simultaneidade F em função da potência computada. Se a unidade de potência empregada nos cálculos de C for o kilowatt (kW), a relação entre F e C proposta na norma NBR 14570 é

$$F = \begin{cases} 1 & \text{se } C < 24.43; \\ \frac{1}{1 + 0.01016(C - 24.37)^{0.8712}} & \text{se } 24.43 \leq C < 670.9; \\ \frac{1}{1 + 0.7997(C - 73.67)^{0.19931}} & \text{se } 670.9 \leq C < 1396; \\ 0.23 & \text{se } C > 1396; \end{cases} \quad (2.2)$$

- calcula-se a potência adotada A fazendo $A = F \times C$.

Utilizando a relação mostrada em (2.1), pode-se redefinir o fator de simultaneidade em termos de vazão. Em outras palavras, o fator de simultaneidade é a relação percentual entre a vazão verificada na prática devida ao trabalho simultâneo de um grupo de aparelhos servido por um determinado trecho de tubulação (vazão adotada) e a vazão máxima desses mesmos aparelhos (vazão computada).

Para exemplificar o uso do fator de simultaneidade, consideremos um caso hipotético. O esquema da Figura 2.2 representa parte de um sistema predial de gás combustível referente às instalações prediais de um apartamento. Para dimensionar esta rede secundária, o projetista precisa determinar as vazões de projeto em vários trechos da tubulação.

Figura 2.2: Esquema de instalação predial de gás típica de um apartamento

Sabe-se de antemão, por meio da empreiteira responsável, que os apartamentos deste prédio disponibilizarão aos seus futuros proprietários três pontos de utilização para os seguintes aparelhos: um fogão de 6 bocas, um aquecedor de água e uma secadora de roupas. As potências nominais destes aparelhos são, respectivamente, $P_F = 11000 \text{ kcal/h}$, $P_A = 14700 \text{ kcal/h}$ e $P_S = 6000 \text{ kcal/h}$. Estes valores, que podem ser obtidos facilmente por meio das especificações técnicas contidas nos manuais dos aparelhos, representam a potência computada de cada aparelho. A norma NBR 14570 assume que o poder calorífico do gás natural é 9230 kcal/m^3 . Portanto, com base em (2.1), conclui-se que as vazões computadas do fogão, do aquecedor e da secadora são, respectivamente, $Q_F = 1.19 \text{ m}^3/\text{h}$, $Q_A = 1.59 \text{ m}^3/\text{h}$ e $Q_S = 0.65 \text{ m}^3/\text{h}$.

Os pontos em destaque na Figura 2.2 determinam ainda trechos da instalação predial com vazões computadas diferentes. Por exemplo, no trecho AB , a vazão computada é a vazão devida ao uso simultâneo dos três aparelhos, que é igual a $Q_{AB} = Q_F + Q_A + Q_S = 3.43 \text{ m}^3/\text{h}$. De forma análoga, no trecho BB' pode-se observar uma vazão computada igual à vazão devida ao fogão, ou seja, $Q_{BB'} = Q_F = 1.19 \text{ m}^3/\text{h}$. A Tabela 2.1 mostra as vazões computadas nos trechos em destaque.

Tabela 2.1 Vazões computadas nos trechos em destaque da Figura 2.2

Trecho	Vazão Computada (m^3/h)
AB	3.43
BC	2.24
CD	0.65
BB'	1.19
CC'	1.59

É importante ressaltar que a vazão computada corresponde à vazão máxima do aparelho. Portanto, se a vazão computada do fogão é $Q_F = 1.19 \text{ m}^3/h$, esta só será observada se as seis bocas e o forno do fogão forem ligados simultaneamente. O raciocínio é análogo quando avaliamos, por exemplo, a vazão computada no trecho AB . Neste trecho, a vazão $Q_{AB} = 3.43 \text{ m}^3/h$ só será observada se os três aparelhos forem utilizados, simultaneamente, em suas vazões máximas.

Uma vez definidas as potências computadas, é possível então calcular os fatores de simultaneidade para os trechos em destaque utilizando (2.2). Para tanto, deve-se converter as potências computadas de $kcal/h$ para kW ($1 \text{ kW} = 859.845 \text{ kcal/h}$). Finalmente, ao calcular o fator de simultaneidade, pode-se obter facilmente as potências adotadas. A Tabela 2.2 resume os resultados obtidos no procedimento de cálculo da potência adotada.

Tabela 2.2 Fator de simultaneidade nos trechos em destaque da Figura 2.2

Trecho	Vazão Computada (m^3/h)	Potência Computada		FS	Potência Adotada (kW)
		$kcal/h$	kW		
AB	3.43	31658.9	36.82	0.92	33.74
BC	2.24	20675.2	24.05	1.00	24.05
CD	0.65	5999.5	6.98	1.00	6.98
BB'	1.19	10983.7	12.77	1.00	12.77
CC'	1.59	14675.7	17.07	1.00	17.07

Nota-se que no trecho AB, que é o trecho no qual serão conectados todos os aparelhos de utilização, a potência adotada é inferior à potência computada. Isso ocorre porque, neste

trecho, a soma das potências computadas é suficientemente grande para se supor que ela não será atingida. De fato, para atingir esta potência, os três aparelhos de utilização deveriam ser ligados, simultaneamente, em sua potência máxima, o que configura uma situação bastante improvável. Desta forma, o fator de simultaneidade atua como um fator redutor da potência computada, gerando maior economia no dimensionamento da rede como um todo.

Esta seção teve por objetivo lançar os principais conceitos e definições, e também exemplificar a forma de utilização da curva do fator de simultaneidade para a obtenção da vazão adotada. No entanto, o foco central deste trabalho é re-discutir o atual modelo utilizado no dimensionamento de sistemas prediais de gás combustível, bem como propor outros novos modelos. O amplo trabalho de medição realizado pelo IPT e COMGÁS foi fundamental neste sentido, pois subsidiou esta dissertação com dados recentes e coletados com a tecnologia necessária para garantir a sua confiabilidade.

2.3 Coleta de dados

Conforme mencionado na Seção 2.2, o fator de simultaneidade pode ser definido como a relação percentual entre a vazão verificada na prática, devida ao trabalho simultâneo de um grupo de aparelhos servido por um determinado trecho de tubulação, e a vazão potencial desses mesmos aparelhos. Todo o trabalho de coleta de dados, portanto, foi direcionado com o objetivo principal de observar os dois números desta relação percentual e, como consequência, obter o fator de simultaneidade empírico.

Os sistemas prediais selecionados na amostra foram visitados no segundo semestre de 2004 e representam os sistemas prediais de uso exclusivamente residencial, situados na cidade de São Paulo e equipados com fogão e aquecedor de água¹. Por serem equipados com aquecedor de água, tomou-se o cuidado de coletar a maior parte dos dados nos meses de julho e agosto, meses tipicamente frios nos quais o consumo de gás aumenta, e nos

¹Em apenas quatro sistemas prediais a utilização de gás natural era exclusivamente para aquecimento de água.

quais a simultaneidade do uso de aparelhos supostamente atinge seu regime mais crítico. Foram consideradas características excludentes os sistemas prediais com índice de ocupação inferior a 50%, presença de atividade comercial e presença de caldeira. A amostra foi selecionada do cadastro de clientes COMGÁS referente ao mês de abril de 2004. Como não constam neste cadastro informações sobre as características dos sistemas prediais, muitas visitas resultaram em não-medição. Do total de 100 sistemas prediais visitados, 42 resultaram em medições válidas. Dentre os 58 sistemas prediais não monitorados, as principais causas foram ausência de aquecedor de água, presença de caldeira e presença de atividade comercial.

Para coletar dados dos sistemas prediais foram utilizadas duas ferramentas: questionários e aparelho de aquisição de dados de vazão, ou *data-logger*. Os questionários foram úteis para a obtenção de dados referentes às características dos sistemas prediais, tais como número de andares, número de apartamentos, número de apartamentos ocupados, número de moradores, entre outros, bem como dados referentes aos tipos e características dos aparelhos de utilização presentes nos apartamentos, tais como potências nominais, marcas e modelos, entre outros. Estas informações foram úteis para calcular os valores médios das potências nominais do aquecedor e do fogão em cada sistema predial. Estes valores médios foram então somados e multiplicados pelo número de apartamentos ocupados, obtendo-se assim a estimativa da potência computada do sistema predial. Utilizando a relação definida em 2.1, obtém-se a vazão máxima possível que se pode observar no sistema predial quando o grupo de aparelhos nele instalado é ligado simultaneamente. Este número é o denominador da relação percentual que define o fator de simultaneidade.

Além dos questionários, os medidores de gás dos 42 sistemas prediais monitorados foram instrumentados com *data-logger's* CDL da marca Sensus / Invensys. Estes aparelhos foram mantidos por, no mínimo, sete dias consecutivos e completos, a fim de detectar possíveis particularidades de consumo nos diferentes dias da semana. A vazão máxima observada no período de medição foi tomada como sendo a vazão verificada na prática devida ao trabalho simultâneo do grupo de aparelhos instalado no sistema predial. Este número é o numerador da relação percentual que define o fator de simultaneidade. A Figura 2.3 mostra um exemplo das informações geradas pelo *data-logger*.

(a)

DATA	HORA	VAZÃO
19/07/2004	10:15:44,96	5,556
19/07/2004	10:15:51,44	5,534
19/07/2004	10:15:57,94	5,517
19/07/2004	10:16:04,46	5,710
19/07/2004	10:16:10,76	5,975
19/07/2004	10:16:16,78	5,568
19/07/2004	10:16:23,24	5,692
19/07/2004	10:16:29,56	5,692
19/07/2004	10:16:35,88	5,674
19/07/2004	10:16:42,22	5,621

(b)

Figura 2.3: Exemplo de informações obtidas pelo *data-logger*: gráfico de vazão (a) e registros periódicos de vazão (b).

Com as estimativas da potência computada - obtidas via questionários - e as vazões máximas possíveis, - obtidas via *data-logger* - foi possível obter os fatores de simultaneidade empíricos dos 42 sistemas prediais residenciais visitados. Os dados obtidos neste trabalho, que podem ser encontrados no Apêndice A, constituem a base de dados a ser estudada nesta dissertação.

2.4 Análise descritiva

Para uma análise inicial, foi traçado o gráfico de dispersão entre o fator de simultaneidade e a potência computada, com o objetivo de verificar se existe uma relação entre estas duas variáveis e, em caso afirmativo, qual o tipo de relação sugerido. Este gráfico, que pode ser observado na Figura 2.4, evidencia uma clara relação entre estas duas variáveis, e sugere um comportamento não linear. Sob o ponto de vista da esperança do fator de simultaneidade, o gráfico sugere um decaimento exponencial à medida que a potência aumenta, com valores próximos a 0 para as potências mais elevadas. Sob o ponto de vista da variância do fator de simultaneidade, o gráfico sugere maior dispersão para potências próximas de 0, e decaimento da dispersão para valores mais elevados de potência.

Figura 2.4: Fator de Simultaneidade x Potência Computada

A Figura 2.4 acusa também a existência de duas observações isoladas da massa de dados. Na busca por uma explicação para este comportamento, analisaram-se as características dos edifícios que geraram tais observações. Constatou-se que as observações em destaque apresentam perfis bastante diferentes em relação ao perfil típico. Dos sistemas prediais visitados, o edifício típico corresponde a uma edificação de 16 andares com 4 apartamentos cada, totalizando assim 64 apartamentos. A observação 21 refere-se ao um edifício de 4 andares com 6 apartamentos cada, totalizando 24 apartamentos, configurando o menor sistema predial visitado. A observação 34 está no outro extremo, e representa o maior edifício visitado, com 208 apartamentos espalhados em 26 andares. Não foram encontrados erros de digitação ou cálculo, e as edificações não apresentaram nenhuma característica que as colocassem fora do escopo do trabalho. Embora se destaquem da massa de dados, nota-se que as observações obedecem à tendência dos dados, e não devem constituir um problema no ajuste dos modelos.

Capítulo 3

Descrição dos modelos estatísticos

Em termos estatísticos, o objetivo da pesquisa apresentada nesta dissertação resume-se em modelar o comportamento do fator de simultaneidade em função da potência computada, ressaltando que a variável resposta é um número situado no intervalo $(0,1)$.

Existem diversas alternativas para modelar uma variável resposta observada no intervalo $(0,1)$. Kieschnick e McCullough (2003), por exemplo, fizeram uma revisão bibliográfica com o propósito de apontar os modelos mais freqüentemente empregados em problemas desta natureza. Concluíram que o modelo de regressão linear normal (MRLN) foi o modelo mais utilizado, ainda que nem sempre explicitamente invocado. A sua utilização fica evidente quando se constata a aplicação dos testes t ou F , indicando que o pesquisador implicitamente assumiu a normalidade da distribuição condicional. Kieschnick e McCullough (2003) também observaram que o uso destes testes geralmente está associado à suposição de uma função linear para a esperança condicional, o que caracteriza completamente a aplicação do MRLN. Ainda segundo os autores, um outro modelo freqüentemente empregado no estudo de variáveis respostas observadas no intervalo $(0,1)$ é o modelo de regressão linear normal com variável resposta transformada. Dentre os estudos que se valeram deste modelo, todos empregaram a transformação logito, uma transformação particularmente conveniente para a interpretação dos resultados. Ao mesmo tempo em que esta alternativa ajusta a variável resposta ao domínio do MRLN, ela possibilita usufruir dos resultados já estabelecidos para este modelo.

Ainda buscando soluções para a modelagem de dados no intervalo $(0,1)$, Ferrari e Cribari-Neto (2004) propuseram um modelo de regressão no qual assumem que a variável resposta distribui-se segundo a distribuição beta. A distribuição beta caracteriza-se por ser extremamente flexível, podendo assumir diferentes formas de acordo com a combinação de seus dois parâmetros. Diferentemente das parametrizações usuais de distribuição beta, definidas por dois parâmetros de forma, Ferrari e Cribari-Neto (2004) convenientemente utilizaram uma parametrização da função de distribuição beta de modo que seus parâmetros representassem a média e a precisão, sendo este um importante passo para a interpretabilidade do modelo. Os procedimentos de modelagem e inferência considerados por Ferrari e Cribari-Neto (2004) assemelham-se àqueles propostos por McCullagh e Nelder (1989) para os modelos lineares generalizados, ou seja, uma função da média da variável resposta explicada por uma estrutura linear envolvendo variáveis regressoras. Conforme se verá a seguir, a variância da variável resposta é uma função da média e, conseqüentemente, das variáveis regressoras. Desta forma, o modelo possibilita acomodar naturalmente dados de natureza heteroscedástica.

3.1 Modelo de regressão linear normal

Sejam $x_1, \dots, x_n$ valores fixos e conhecidos, e $y_1, \dots, y_n$ observações de variáveis aleatórias independentes. O modelo de regressão linear normal simples (MRLN) é definido como

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n, \quad (3.1)$$

em que β_0 e β_1 são parâmetros desconhecidos, $x_1, \dots, x_n$ são observações fixas e conhecidas de uma variável regressora e $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios independentes e identicamente distribuídos (i.i.d.), tais que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Os estimadores dos parâmetros do modelo serão obtidos pelo método dos mínimos quadrados, ou seja, deseja-se obter uma relação linear que minimize a soma dos erros ao quadrado, dada por $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$. Para obter a solução que minimiza

esta soma, calculam-se as derivadas parciais em relação a β_0 e β_1 e iguala-se o resultado a 0, obtendo assim o sistema de equações normais cujas soluções $\hat{\beta}_0$ e $\hat{\beta}_1$ satisfazem

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

e

$$\hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$$

Desta forma, os estimadores de mínimos quadrados de β_0 e β_1 são dados, respectivamente, por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}.$$

Para fins de inferência sobre os parâmetros do modelo, é conveniente obter a distribuição dos estimadores dos parâmetros. Utilizando a suposição de normalidade dos erros e por meio de algumas manipulações algébricas, pode-se demonstrar que

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

e

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

Da teoria das probabilidades (ver, por exemplo, Casella e Berger, 1996), demonstra-se ainda que, ao substituir σ^2 por seu estimador $\hat{\sigma}^2$, as estatísticas resultantes terão distribuição t de Student, com $(n - k)$ graus de liberdade, em que k é o número de parâmetros do modelo de regressão. Logo, as estatísticas para testar as hipóteses $H_0 : \beta_0 = 0$ e $H_0 : \beta_1 = 0$ dadas, respectivamente, por

$$t(\hat{\beta}_0) = \frac{\hat{\beta}_0}{\hat{\sigma}^2} \sqrt{\frac{n \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n x_i^2}}$$

e

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\hat{\sigma}^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$$

têm uma distribuição t de Student com $(n - k)$ graus de liberdade. Aqui, um estimador não-viesado do parâmetro σ^2 é dado por

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

em que $\hat{y}_i$ é o valor predito para a variável resposta para um x_i fixo e conhecido.

Uma segunda abordagem possível para a obtenção dos estimadores dos parâmetros de regressão é utilizando o método de máxima verossimilhança. Os estimadores obtidos por este método são exatamente iguais aos obtidos pelo método dos mínimos quadrados. Além disso, pode-se demonstrar que estes estimadores são o “melhores” estimadores não-viesados, no sentido de possuírem a menor variância dentre todos os outros estimadores lineares não viciados. Para uma introdução mais detalhada sobre o modelo de regressão linear, ver Casella e Berger (1990) ou Bussab e Morettin (2002). Para um estudo mais aprofundado, consultar, por exemplo, Draper e Smith (1981), Neter, Wasserman e Kutner (1996) ou Montgomery, Peck e Vining (2001).

O modelo definido em (3.1) resulta no ajuste de uma reta, o que é evidentemente inadequado para os dados em questão, conforme se observa na Figura 2.4. Ainda que fosse adequado, Kieschnick e McCullough (2003) mencionam que o uso do MRLN para o

estudo de variáveis resposta no intervalo (0,1) carrega dois erros conceituais. Em primeiro lugar, a variável resposta não está definida no conjunto dos reais, que é o domínio sobre o qual a distribuição normal é definida. Além disso, o fato da variável resposta ser observada em um intervalo limitado implica que a função de esperança condicional precisa ser limitada e, portanto, não pode ser linear, e que a variância condicional precisa ser uma função da média pois à medida que se aproxima de seus limites, a variância se aproxima de zero. No MRLN estas duas condições são claramente violadas. Uma forma de contornar estes problemas conceituais e ao mesmo tempo obter um modelo mais adequado ao comportamento dos dados de fator de simultaneidade é transformar a variável resposta de modo que a variável transformada esteja definida no conjunto dos reais. Na revisão bibliográfica realizada por Kieschnick e McCullough (2003), este foi o segundo modelo mais freqüentemente empregado no estudo de variáveis resposta em (0,1). Os autores relatam que, na maioria dos estudos analisados, as suposições feitas sobre a variável resposta não-transformada não são claramente definidas. No entanto, pelo fato de todos os estudos empregarem a transformação logito, pode-se concluir que o modelo que está por trás destes estudos define-se como

$$\ln \left(\frac{y_i}{1 - y_i} \right) = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n, \quad (3.2)$$

em que β_0 e β_1 são os parâmetros do modelo e $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios i.i.d., tais que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. O modelo definido em (3.2) é o primeiro modelo a ser considerado neste trabalho.

Uma consideração pertinente acerca dos modelos de regressão lineares - e particularmente útil neste trabalho - diz respeito à terminologia *linear*. Embora a palavra remeta ao ajuste de uma reta (ou de seus similares em dimensões maiores) entre variável resposta e variáveis regressoras, esta característica não é necessariamente verdadeira. De modo geral, a literatura estatística emprega o termo referindo-se à suposição que se faz do valor esperado da variável resposta ser uma função linear nos parâmetros. Neste contexto, as variáveis regressoras assumem o papel de constantes que podem ser livremente transformadas, aumentando a flexibilidade do modelo. Assim, tal como definido em (3.1), o modelo

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i, \quad i = 1, \dots, n,$$

em que x_i^* representa o i -ésimo valor da variável regressora *transformada*, também é um modelo de regressão linear, independentemente da transformação empregada. Analogamente, tomando-se como base o modelo definido em (3.2), não há restrições em definir

$$\ln \left(\frac{y_i}{1 - y_i} \right) = \beta_0 + \beta_1 x_i^* + \epsilon_i, \quad i = 1, \dots, n, \quad (3.3)$$

em que ϵ_i é o erro aleatório, tal que ϵ_i i.i.d $\mathcal{N}(0, \sigma^2)$.

Em geral, o comportamento da variável resposta em função das variáveis regressoras orienta as transformações das variáveis regressoras. Neste trabalho, por exemplo, foram consideradas diversas transformações da variável regressora levando-se em conta o comportamento do fator de simultaneidade em função da potência computada, cujo decaimento é aparentemente exponencial. Dentre as diversas transformações testadas, pode-se mencionar as transformações inversa, exponencial e logarítmica, entre outras. Ao fim deste trabalho, observou-se que o logaritmo natural da variável regressora foi a transformação que apresentou os melhores resultados, e definiu-se que o modelo apresentado em (3.3) com a transformação $x_i^* = \ln(x_i)$ seria o segundo modelo a ser considerado neste trabalho.

3.2 Modelo de regressão não-linear normal

Kieschnick e McCullough (2003) citam a classe de modelos de regressão não-lineares como sendo uma alternativa aos problemas de regressão com variável resposta em $(0,1)$. A vantagem do MRNLN sobre o MRLN é o fato de permitir a utilização de uma função que limite a esperança condicional ao intervalo $(0,1)$. Por outro lado, assim como o MRLN, o MRNLN também supõe homoscedasticidade dos dados. Conforme já mencionado na seção anterior, na modelagem de dados observados no intervalo $(0,1)$, é desejável que o modelo suponha a variância como função da média. Uma outra ressalva em relação ao MRNLN é que, ao empregar o método de mínimos quadrados para a estimar os parâmetros de

modelos deste tipo, obtém-se um sistema de equações não-lineares sem solução analítica, fazendo-se necessário o uso de procedimentos de otimização não-linear, tais como Newton-Raphson e Gauss-Newton e, em algumas situações, o procedimento pode não convergir para uma solução.

Pela análise da Figura 2.4, pode-se imaginar que o fator de simultaneidade e a potência computada relacionam-se segundo uma função do tipo $f(x) = \beta_0 e^{\beta_1 x}$. Desta forma, sugere-se o seguinte modelo não-linear.

$$y_i = \beta_0 e^{\beta_1 x_i} + \epsilon_i, \quad i = 1, \dots, n,$$

em que $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios i.i.d., tais que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

A utilização de um modelo de regressão não-linear pode se tornar mais simples quando uma transformação nas variáveis de estudo permite recair em um modelo de regressão linear normal. Neste caso, uma alteração na suposição dos erros do modelo torna esta abordagem possível. Ao invés de assumir que os erros participam de forma aditiva no modelo, suponhamos que eles participem de forma multiplicativa, ou seja,

$$y_i = \beta_0 e^{\beta_1 x_i} \epsilon_i, \quad i = 1, \dots, n,$$

em que $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios i.i.d., tais que $\ln(\epsilon_i) \sim \mathcal{N}(0, \sigma^2)$. Aplicando-se o logaritmo natural em ambos os lados, tem-se que

$$\ln(y_i) = \ln(\beta_0) + \beta_1 x_i + \ln(\epsilon_i), \quad i = 1, \dots, n.$$

Por fim, pode-se definir $y_i^* = \ln(y_i)$, $\beta_0^* = \ln(\beta_0)$ e $\epsilon_i^* = \ln(\epsilon_i)$ para obter o modelo

$$y_i^* = \beta_0^* + \beta_1 x_i + \epsilon_i^*, \quad i = 1, \dots, n, \tag{3.4}$$

em que $\epsilon_1^*, \dots, \epsilon_n^*$ são i.i.d com $\epsilon_i^* \sim \mathcal{N}(0, \sigma^2)$. Desta forma, pode-se utilizar toda a teoria

do modelo de regressão linear normal para estudar a relação entre duas variáveis cuja relação funcional é claramente não-linear. O modelo definido em (3.4) é o terceiro modelo a ser considerado neste trabalho.

O quarto modelo considerado neste trabalho é o mesmo empregado em IPT (2005). Analogamente ao desenvolvimento anterior, assume-se que o fator de simultaneidade e a potência computada se relacionam segundo uma função do tipo $f(x) = \beta_0 x^{\beta_1}$ e que o erro participa de forma multiplicativa no modelo. Desta forma, sugere-se o modelo não-linear

$$y_i = \beta_0 x_i^{\beta_1} \epsilon_i, \quad i = 1, \dots, n,$$

em que ϵ_i é tal que $\ln(\epsilon_i)$ i.i.d $\mathcal{N}(0, \sigma^2)$. Aplicando-se o logaritmo natural em ambos os lados, tem-se que

$$\ln(y_i) = \ln(\beta_0) + \beta_1 \ln(x_i) + \ln(\epsilon_i), \quad i = 1, \dots, n.$$

Por fim, sejam $y_i^* = \ln(y_i)$, $x_i^* = \ln(x_i)$, $\beta_0^* = \ln(\beta_0)$ e $\epsilon_i^* = \ln(\epsilon_i)$, e obtém-se o modelo

$$y_i^* = \beta_0^* + \beta_1 x_i^* + \epsilon_i^*, \quad i = 1, \dots, n. \quad (3.5)$$

Os modelos apresentados nas Seções 3.1 e 3.2 apresentam o inconveniente de utilizar a variável resposta transformada. Desta forma, a interpretação do modelo pode ficar comprometida uma vez que os resultados obtidos referem-se à variável resposta transformada, e não à variável resposta original. O modelo de regressão proposto a seguir, por sua vez, não apresenta este inconveniente, e constitui uma terceira alternativa para a modelagem dos dados aqui apresentados.

3.3 Modelo de regressão beta

A família de distribuições beta é composta pelas distribuições cujas funções densidade de probabilidade têm a forma

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, \quad 0 < y < 1, \quad (3.6)$$

em que $p > 0$, $q > 0$ e $\Gamma(\cdot)$ é a função gama definida como

$$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy, \quad a > 0.$$

A média e a variância de y são dadas, respectivamente, por

$$E(y) = \frac{p}{p+q} \quad \text{e} \quad \text{Var}(y) = \frac{pq}{(p+q)^2(p+q+1)}. \quad (3.7)$$

Ferrari e Cribari-Neto (2004) ressaltam a conveniência de reparametrizar (3.6) de modo que os parâmetros representem a média e a precisão da variável resposta. Para obter esta estrutura, consideram a transformação $\mu = p/(p+q)$ e $\phi = p+q$, implicando em $p = \mu\phi$ e $q = (1-\mu)\phi$. Assim, obtém-se uma nova forma para a função densidade de probabilidade para a família de distribuições beta dada por

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \quad (3.8)$$

em que $0 < \mu < 1$ e $\phi > 0$. Além disso, de (3.7), tem-se que

$$E(y) = \mu \quad \text{e} \quad \text{Var}(y) = \frac{V(\mu)}{1+\phi}, \quad (3.9)$$

em que $V(\mu) = \mu(1-\mu)$. Desta forma, μ é o parâmetro que representa a média da variável resposta, enquanto ϕ pode ser interpretado como um parâmetro de precisão, uma vez que, para μ fixado, quanto maior o valor de ϕ menor a variância de y .

Assim, sejam y_i , $i = 1, \dots, n$ variáveis aleatórias independentes tais que cada y_i obedece a uma função densidade de probabilidade como em (3.8), com média μ_i e parâmetro de precisão ϕ . O modelo de regressão beta define-se ao assumir que

$$g(\mu_i) = \sum_{t=1}^k x_{it}\beta_t = \eta_i, \quad (3.10)$$

em que $\beta = (\beta_1, \dots, \beta_k)^\top$ é um vetor de parâmetros desconhecidos ($\beta \in \mathbb{R}^k$), $x_{i1}, \dots, x_{ik}$ são observações, fixas e conhecidas, em k variáveis regressoras, $k < n$, e $g(\cdot)$ é uma função de ligação, estritamente monótona e duas vezes diferenciável, que leva valores em $(0, 1)$ para a reta real. Desta forma, tem-se que $\mu_i = g^{-1}(\eta_i)$ e $\text{Var}(y_i) = V(\mu_i)/(1 + \phi)$. Como a variância da resposta depende de μ , é fácil concluir por estas expressões que as variâncias não são constantes, ainda que o parâmetro de dispersão ϕ seja constante. Portanto, variáveis respostas com variâncias não constantes podem ser naturalmente acomodadas pelo modelo.

Existem diversas possíveis escolhas para a função de ligação $g(\cdot)$. Atkinson (1985) e McCullagh e Nelder (1989) comparam diversas funções de ligação, tais como as funções logito, probito, log-log, entre outras. Uma função de ligação particularmente útil em termos de interpretação de parâmetros é a função logito, definida como

$$g(\mu_i) = \ln \left(\frac{\mu_i}{1 - \mu_i} \right).$$

Assumindo a função de ligação logito, suponha que a t -ésima variável regressora sofra um incremento de c unidades e todas as demais variáveis regressoras permaneçam constantes. Seja μ^+ a média de y obtida com os novos valores das variáveis regressoras, enquanto μ é média de y obtida com os valores originais das variáveis regressoras. Então, demonstra-se que

$$e^{c\beta_t} = \frac{\mu^+/(1 - \mu^+)}{\mu/(1 - \mu)},$$

ou seja, ao assumir a função de ligação logito, pode-se interpretar os parâmetros do modelo

em termos de razão de chances. No contexto deste trabalho, esta interpretação permite obter o aumento da chance de utilização simultânea média dos aparelhos de utilização de um sistema predial a cada incremento de c unidades na potência computada.

Com base em (3.8), define-se o logaritmo da função de verossimilhança como

$$\ell(\beta, \phi) = \sum_{i=1}^n \ell_i(\mu_i, \phi), \quad (3.11)$$

em que

$$\begin{aligned} \ell_i(\mu_i, \phi) = & \log \Gamma(\phi) - \log \Gamma(\mu_i \phi) - \log \Gamma[(1 - \mu_i)\phi] + \\ & + (\mu_i \phi - 1) \log y_i + [(1 - \mu_i)\phi - 1] \log(1 - y_i). \end{aligned}$$

A função escore, obtida diferenciando-se a função de máxima verossimilhança em função dos parâmetros do modelo, é dada por $(U_\beta(\beta, \phi)^\top, U_\phi(\beta, \phi)^\top)^\top$. Aqui,

$$U_\beta(\beta, \phi) = \phi X^\top T(y^* - \mu^*),$$

em que X é uma matriz $n \times k$ cuja i -ésima linha é $x_i^\top$, $T = \text{diag}[g'(\mu_1)^{-1}, \dots, g'(\mu_n)^{-1}]$ e os i -ésimos elementos de y^* e μ^* dados por $y_i^* = \log[y_i/(1 - y_i)]$ e $\mu_i^* = \psi(\mu_i \phi) - \psi[(1 - \mu_i)\phi]$, em que $\psi(\cdot)$ é a função digama, definida como $\psi(z) = d \log \Gamma(z)/dz$, $z > 0$, e

$$U_\phi(\beta, \phi) = \sum_{i=1}^n u_i,$$

em que

$$u_i = \mu_i(y_i^* - \mu_i^*) + \log(1 - y_i) - \psi[(1 - \mu_i)\phi] + \psi(\phi).$$

É também conveniente obter a matriz de informação de Fisher, denotada por

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}.$$

Ferrari e Cribari-Neto (2004) obtiveram

$$K_{\beta\beta} = \phi X^\top W X,$$

com $W = \text{diag}(w_1, \dots, w_n)$, com $w_i = \phi v_i (1/g'(\mu_i)^2)$ e $v_i = \psi'(\mu_i \phi) + \psi'[(1 - \mu_i)\phi]$,

$$K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c,$$

em que $c = (c_1, \dots, c_n)^\top$ com $c_i = \phi[\psi'(\mu_i \phi)\mu_i - \psi'((1 - \mu_i)\phi)(1 - \mu_i)]$, sendo $\psi'(\cdot)$ a função trigama, e

$$K_{\phi\phi} = \text{tr}(D),$$

em que $D = \text{diag}(d_1, \dots, d_n)$, com $d_i = \psi'(\mu_i \phi)\mu_i^2 + \psi'[(1 - \mu_i)\phi](1 - \mu_i)^2 - \psi'(\phi)$.

Ao contrário dos modelos lineares generalizados, no modelo de regressão beta os parâmetros β e ϕ não são ortogonais, ou seja, $K_{\beta\phi}$ não é nulo (ver McCullagh e Nelder, 1989, por exemplo).

Ferrari e Cribari-Neto (2004) salientam que, sob condições de regularidade e para uma amostra suficientemente grande, vale a aproximação

$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+1} \left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1} \right).$$

Aqui, $\hat{\beta}$ e $\hat{\phi}$ representam, respectivamente, os estimadores de máxima verossimilhança de β e ϕ . Como não existe uma forma fechada para estes estimadores, é necessário obtê-los numericamente, maximizando a função de log-verossimilhança (3.11) por meio de um algoritmo de otimização não-linear, tais como os algoritmos de Newton ou quase-Newton.

Para maiores detalhes, ver por exemplo Nocedal e Wright (1999).

Baseados em Rao (1973), Ferrari e Cribari-Neto (2004) utilizam expressões padrões para a inversa de matrizes particionadas para obter K^{-1} , denotada por

$$K^{-1} = K^{-1}(\beta, \phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix}.$$

Eles obtiveram

$$K^{\beta\beta} = \frac{1}{\phi} (X^\top W X)^{-1} \left(\mathcal{I}_k + \frac{X^\top T c c^\top T^\top X (X^\top W X)^{-1}}{\xi \phi} \right),$$

com $\xi = \text{tr}(D) - \phi^{-1} c^\top T^\top X (X^\top W X)^{-1} X^\top T c$ e $\mathcal{I}_k$ a matriz identidade de ordem k ,

$$K^{\beta\phi} = (K^{\phi\beta})^\top = -\frac{1}{\xi \phi} (X^\top W X)^{-1} X^\top T c,$$

e

$$K^{\phi\phi} = \xi^{-1}.$$

A inversa da matriz informação de Fisher é útil na obtenção de erros padrão assintóticos para os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão beta.

O modelo definido em (3.10) com função de ligação logito é o quinto modelo a ser considerado neste trabalho. Analogamente ao MRLN, aqui também pode ser considerada uma transformação na variável regressora. Após várias tentativas de transformação - inversa, exponencial, logarítmica, entre outras - chegou-se à conclusão que o sexto e último modelo a ser considerado neste trabalho levaria em conta uma transformação logarítmica na variável regressora, ou seja, $x_i^* = \ln(x_i)$.

3.4 Técnicas de diagnóstico

Um passo importante no processo de ajuste de modelos estatísticos é a análise de diagnóstico. De modo geral, as técnicas empregadas na análise de diagnóstico têm por objetivo validar as suposições do modelo, bem como avaliar se o modelo utilizado se ajusta bem aos dados. É por meio desta análise que são constatados desvios das suposições do modelo, sejam eles referentes às suposições da parte sistemática ou aleatória. É também por meio das técnicas de diagnóstico que são detectadas observações influentes, ou seja, observações que causam impacto desproporcional nos resultados do ajuste.

Na análise de diagnóstico, é relevante detectar pontos de alavanca. Uma observação y_i é caracterizada como ponto de alavanca quando a sua influência sobre o seu próprio valor ajustado $\hat{y}_i$ é muito grande. Uma medida desta influência foi proposta por Hoaglin e Welsch (1978) para o caso da regressão linear normal, e consiste em avaliar os valores da diagonal principal da matriz de projeção $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, ou seja,

$$h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i.$$

Segundo Paula (2003), as observações nas quais se observa a condição $h_{ii} \geq 2k/n$ são elegíveis a pontos de alavanca, e geralmente estão localizadas em regiões remotas no subespaço gerado pelas colunas da matriz $\mathbf{X}$. Ainda para o caso do modelo de regressão linear normal, mostra-se que os valores h_{ii} coincidem com a medida $\partial \hat{y}_i / \partial y_i$, em que $\hat{y}_i$ é o valor ajustado da i -ésima observação. Desta forma, pode-se interpretar h_{ii} como a variação em $\hat{y}_i$ quando se acrescenta um infinitésimo a y_i . Wei, Hu e Fung (1998) propõem uma forma geral de determinar pontos de alavanca baseando-se na medida

$$GL(\hat{\boldsymbol{\theta}}) = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}^\top},$$

onde $\mathbf{y} = (y_1, \dots, y_n)$ é o vetor de observações e $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)$ é o vetor de valores ajustados. A medida $GL(\hat{\boldsymbol{\theta}})$ denomina-se alavanca generalizada e abrange todos os casos nos quais a variável resposta é contínua. Aqui, $\boldsymbol{\theta}$ é um vetor tal que $E(\mathbf{y}) = \mu(\boldsymbol{\theta})$ e $\hat{\boldsymbol{\theta}}$ é um estimador de $\boldsymbol{\theta}$, com $\hat{\mathbf{y}} = \mu(\hat{\boldsymbol{\theta}})$. O elemento (t, u) de $GL(\hat{\boldsymbol{\theta}})$ é a taxa de variação instan-

tânea no t -ésimo valor predito com respeito ao u -ésimo valor de resposta. Seguindo esta abordagem, Ferrari e Cribari-Neto (2004) apresentam a fórmula da alavanca generalizada para o modelo de regressão beta, a qual será utilizada neste trabalho.

Também é útil avaliar resíduos na análise de diagnóstico. De modo geral, define-se o resíduo para a i -ésima observação como uma função da forma $r_i = r(y_i, \hat{\mu}_i)$, $i = 1, \dots, n$, a qual tem por objetivo medir a discrepância entre o valor observado e o correspondente valor ajustado. A maioria dos resíduos propostos na literatura se baseia na diferença entre o valor observado e o valor ajustado. De fato, a definição mais básica de um resíduo é o resíduo ordinário, que é simplesmente a diferença entre estes valores, ou seja,

$$r_i = y_i - \hat{\mu}_i, \quad i = 1, \dots, n.$$

Particularizando para o caso do modelo de regressão linear normal, verifica-se com alguma álgebra que $r_i \sim \mathcal{N}(0, \sigma^2(1 - h_{ii}))$, $i = 1, \dots, n$ e que, portanto, as quantidades r_i 's possuem variâncias diferentes. Além disso, mostra-se que $\text{Cov}(r_i, r_j) = -\sigma^2 h_{ij}$, $i \neq j$, com $i = 1, \dots, n$ e $j = 1, \dots, n$, ou seja, as quantidades r_i 's não são independentes. Para que sejam comparáveis, é conveniente expressar os resíduos ordinários em sua forma padronizada, dividindo r_i pelo seu respectivo desvio padrão estimado, e assim obter o resíduo studentizado, ou seja,

$$t_i = \frac{r_i}{s(1 - h_{ii})^{1/2}}, \quad i = 1, \dots, n,$$

em que

$$s^2 = \sum_{i=1}^n \frac{r_i^2}{n - k}, \quad (3.12)$$

sendo k o número de parâmetros do modelo. Ao contrário do que sugere o nome deste resíduo, t_i não segue uma distribuição t -Student pois r_i e s^2 não são independentes. Para contornar este problema, Paula (2003) sugere utilizar o resíduo studentizado sem a i -ésima observação, o qual define como

$$t_i^* = \frac{r_i}{s_{(i)}(1 - h_{ii})^{1/2}},$$

em que $s_{(i)} = s^2(n - k - t_i^2)/(n - k - 1)$ é o quadrado médio residual correspondente ao modelo sem a i -ésima observação. Mostra-se que t_i^* tem uma distribuição t com $(n - k - 1)$ graus de liberdade. As análises de diagnóstico referentes a modelos de regressão normal contidas nesta dissertação levam em conta o estudo de t_i^* .

Em seu artigo sobre regressão beta, Ferrari e Cribari-Neto (2004) seguem o preceito dos resíduos se basearem na diferença entre os valores observados e os valores ajustados, e propõem o resíduo ordinário padronizado

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(y_i)}}, \quad (3.13)$$

em que $\widehat{\text{Var}}(y_i) = \hat{\mu}_i(1 - \hat{\mu}_i)/(1 + \hat{\phi})$, com $\hat{\mu}_i = g^{-1}(\mathbf{x}^\top \hat{\boldsymbol{\beta}})$. Aqui, $\hat{\boldsymbol{\beta}}$ e $\hat{\phi}$ são os estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e ϕ , respectivamente. No entanto, Espinheira, Ferrari e Cribari-Neto (2008) mostram, por meio de simulação de Monte Carlo, que para determinados valores de ϕ e μ os resíduos apresentados em (3.13) não são bem aproximados pela distribuição normal. Além disso, por meio de exemplos, mostram que o resíduo não é sensível o suficiente para detectar observações influentes. Com base nestas constatações, propõem dois novos resíduos para a classe de modelos de regressão beta que, diferentemente do conceito utilizado em (3.13), se baseiam na diferença entre o logito da variável resposta e seus valores ajustados. O primeiro deles, denominado resíduo ponderado padronizado 1, define-se como

$$r_i^w = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{v_i}}, \quad (3.14)$$

em que $y_i^* = \log[y_i/(1 - y_i)]$, $\hat{\mu}_i^* = \psi(\hat{\mu}_i \hat{\phi}) - \psi((1 - \hat{\mu}_i) \hat{\phi})$ e $v_i = \psi'(\hat{\mu}_i \hat{\phi}) + \psi'((1 - \hat{\mu}_i) \hat{\phi})$ com $\psi(\cdot)$ denotando a função digamma e $\psi'(\cdot)$ denotando a função trigamma, ou seja, $\psi(a) = d \log \Gamma(a)/da$ e $\psi'(a) = d\psi(a)/da$, $a > 0$. O resíduo ponderado padronizado 2 assemelha-se a r_i^w , diferenciando-se pelo fato de ser ponderado pelo elemento (i, i) da

matriz de projeção $\mathbf{H}^* = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^\top \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}}^{1/2}$, em que $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$ e $w_i = \phi v_i [1/g'(\mu_i)^2]$, e define-se como

$$r_i^{ww} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{v_i(1 - h_{ii}^*)}}. \quad (3.15)$$

Os resultados obtidos por Espinheira, Ferrari e Cribari-Neto (2008) favorecem os resíduos ponderados padronizados, especialmente o resíduo r_i^{ww} , por serem mais eficientes na identificação de observações com grande influência nas estimativas dos parâmetros do modelo na média. Por esta razão, as análises de diagnóstico referentes a modelos de regressão beta contidas nesta dissertação levam em conta o estudo de r_i^{ww} .

Como a distribuição dos resíduos não é conhecida, o gráfico de envelopes simulados proposto por Atkinson (1985) apresenta-se como uma boa alternativa para técnica de diagnóstico. A idéia consiste em construir, por meio de simulações, uma banda de confiança que possa ser utilizada para decidir se os resíduos observados são consistentes com o modelo ajustado. De modo geral, o algoritmo que permite gerar os limites do gráfico de envelope pode ser descrito assim:

1. ajustar o modelo;
2. simular uma amostra de n observações independentes, tomando o modelo ajustado como se fosse o modelo correto, e armazená-la em $\mathbf{y} = (y_1, \dots, y_n)$;
3. ajustar $\mathbf{y}$ contra as variáveis regressoras $\mathbf{X}$ e armazenar os resíduos em $\mathbf{e}_{(n \times 1)} = (e_1, \dots, e_n)$;
4. repetir m vezes os passos 1 e 2, armazenando os resultados na matriz $\mathbf{e}_{(n \times m)}$;
5. ordenar os m grupos de n resíduos;
6. ordenar os elementos da linha $i = 1$ em ordem crescente;
7. selecionar, por exemplo, o 5º percentil da linha $i = 1$ e armazená-lo em e_{inf} ;
8. selecionar, por exemplo, o 95º percentil da linha $i = 1$ e armazená-lo em e_{sup} ;

9. repetir os passos 5 a 7 para $i = 2, \dots, n$.

Ao fim do procedimento, e_{inf} e e_{sup} armazenam, respectivamente, os pontos que compõem os limites inferior e superior do envelope. Atkinson(1985) sugere $m = 19$, o que implica em uma probabilidade aproximada de $1/20$ do maior resíduo de um envelope particular exceder o limite superior. Paula (2003) adapta um programa escrito em Everitt (1994) para gerar envelopes de um modelo de regressão linear normal considerando $m = 100$.

Uma das medidas de influência mais utilizadas em modelos de regressão é a distância de Cook (Cook, 1977), que tem por princípio avaliar o impacto da exclusão de uma observação nas estimativas dos coeficientes da regressão. Originalmente desenvolvida para modelos lineares normais, diversos estudos têm sido feitos para outras classes de modelos, conforme mostram Pregibon (1981), Cook e Weisberg (1982), Atkinson (1985), Cook, Peña e Weisberg (1988), Cordeiro e Paula (1992), entre outros.

Para o modelo de regressão linear normal, quando a i -ésima observação é retirada, a distância de Cook é expressa por

$$D_i = t_i^2 \frac{h_{ii}}{1 - h_{ii}} \frac{1}{k},$$

em que t_i é o resíduo studentizado, h_{ii} é o i -ésimo elemento da matriz de projeção $\mathbf{H}$ e k é o número de parâmetros do modelo. De modo análogo, Ospina (2007) mostra o desenvolvimento da medida para o modelo de regressão beta. Neste caso, a distância de Cook é expressa na forma

$$D_i = (r_i^{ww})^2 \frac{h_{ii}^*}{1 - h_{ii}^*},$$

em que r_i^{ww} é o resíduo ponderado padronizado 2 e h_{ii}^* é o i -ésimo elemento da matriz de projeção $\mathbf{H}^*$. Em ambos os casos, nota-se que a medida D_i tem valores maiores para observações que possuem resíduo grande ou um h_{ii}^* próximo de 1.

Capítulo 4

Análise dos dados

Este capítulo contém a aplicação da base teórica apresentada nos capítulos anteriores. A parte computacional que envolve os ajustes dos modelos e a geração de gráficos foram desenvolvidas na plataforma computacional *R*, versão 2.3.1. Este programa pode ser gratuitamente adquirido por *download* no sítio <http://www.r-project.org>.

Nas seis primeiras seções são apresentados os resultados obtidos com os ajustes dos modelos descritos no Capítulo 3, por meio das estimativas dos parâmetros dos modelos e retas ajustadas sobrepostas ao gráfico de dispersão dos dados. Além disso, também são apresentados, por meio de gráficos, os resultados obtidos com as técnicas de diagnóstico descritas na Seção 3.4. O objetivo aqui é encontrar subsídios que favoreçam ou descartem a utilização dos modelos propostos. Paula (2003) lista alguns gráficos tradicionais que auxiliam na análise de diagnóstico: para a detecção de pontos aberrantes, gráfico de resíduos contra os índices das observações; para detecção de ausência de termo extra, gráfico de resíduos contra variável regressora; para detecção de pontos influentes, gráfico da distância de Cook contra os índices das observações e gráfico de pontos de alavanca contra os índices das observações; para detecção de heteroscedasticidade, gráfico dos resíduos contra o valor ajustado; e para detecção de afastamento da suposição da distribuição, gráfico de resíduos com envelope simulado.

Cada seção contém uma análise dos resultados gerados pelos gráficos supracitados.

Adicionalmente, para os modelos normais, os testes de Breusch-Pagan (Breusch e Pagan, 1979) e de Anderson-Darling (Anderson e Darling, 1954) são utilizados para reforçar as conclusões sobre, respectivamente, as suposições de homoscedasticidade e normalidade. Na última seção, conclui-se a análise dos ajustes com a escolha do modelo mais adequado aos dados desta dissertação.

4.1 MRLN com transformação logito na variável resposta

Nesta seção é considerado o modelo

$$\ln \left(\frac{y_i}{1 - y_i} \right) = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n,$$

em que $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios i.i.d. tais que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, conforme definido na Seção 3.1 em (3.2). A Tabela 4.1 mostra o resultado do ajuste do modelo e indica que o logito da potência é estatisticamente importante para explicar o fator de simultaneidade.

Tabela 4.1 Ajuste do MRLN com transformação logito na variável resposta.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.47353	0.14655	-10.055	0.0000
β_1	-0.34312	0.04496	-7.631	0.0000
σ^2	0.2276	-	-	-

A Figura 4.1.a exibe o gráfico de dispersão entre a potência e o logito do fator de simultaneidade, bem como a reta ajustada, e indica que alguns pontos não são bem acomodados pelo modelo, como é o caso das observações 12, 16, 29 e, principalmente, a 21. Também destacam-se as observações 31 e 34, mas estas por estarem localizadas em uma região remota no eixo da variável regressora. A Figura 4.1.b exibe o gráfico de dispersão entre a potência e o fator de simultaneidade, bem como a reta ajustada transformada, e

(a)

(b)

Figura 4.1: Gráficos de dispersão e reta ajustada do MRLN com transformação logito na variável resposta.

corroborar a Figura 4.1.a, ao evidenciar que o suave decaimento da reta ajustada transformada acaba por não acomodar as observações 12, 16, 21 e 29.

A Figura 4.2.a mostra que os resíduos não formam nenhuma tendência no gráfico de dispersão contra os índices das observações e põe em destaque a observação 21, caracterizando-a como observação aberrante. Também não se observa tendência claramente definida na Figura 4.2.b, sugerindo que não existe ausência de termo extra. Há indícios de que as observações 21, 31 e 34 sejam as mais influentes, conforme a Figura 4.2.c; as observações 31 e 34 também destacam-se como possíveis pontos de alavanca, conforme Figura 4.2.d, enquanto a observação 21 parece também contribuir para a violação da suposição de homoscedasticidade, conforme Figura 4.2.e. O teste de Breusch-Pagan, aplicado aos resíduos deste modelo, resulta em um nível descritivo no limiar da rejeição da hipótese nula a um nível de significância de 10% ($p\text{-valor} = 0.099$). Por fim, a Figura 4.2.f indica que os pontos encontram-se bem distribuídos dentro dos limites do gráfico de envelope, com exceção da observação 21. O teste de normalidade de Anderson-Darling aplicado sobre os resíduos padronizados deste modelo não encontra evidências estatísticas para rejeitar a hipótese de normalidade ($p\text{-valor} = 0.508$).

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.2: Gráficos de diagnóstico do MRLN com transformação logito na variável resposta.

4.2 MRLN com transformação logito na variável resposta e transformação logarítmica na variável regressora

Nesta seção é considerado o modelo

$$\ln\left(\frac{y_i}{1-y_i}\right) = \beta_0 + \beta_1 x_i^* + \epsilon_i, \quad i = 1, \dots, n,$$

em que $x_i^* = \ln(x_i)$ e $\epsilon_1, \dots, \epsilon_n$ são os erros aleatórios i.i.d. tais que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, conforme definido na Seção 3.1 em (3.3). A Tabela 4.2 mostra o resultado do ajuste do modelo e indica que o logaritmo da potência é estatisticamente importante para explicar o fator de simultaneidade.

Tabela 4.2 Ajuste do MRLN com transformação logito na variável resposta e transformação logarítmica na variável regressora.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.75537	0.09684	-18.126	0.0000
β_1	-0.82832	0.08768	-9.447	0.0000
σ^2	0.1730	-	-	-

A Figura 4.3.a exibe o gráfico de dispersão entre o logaritmo natural da potência e o logito do fator de simultaneidade, bem como a reta ajustada, e destaca as observações 21, 31 e 34 por apresentarem valores extremos na variável regressora transformada, não parecendo haver grandes desvios em relação ao modelo de regressão ajustado. A Figura 4.3.b exibe o gráfico de dispersão entre a potência e o fator de simultaneidade, bem como a reta ajustada transformada. O forte decaimento da reta ajustada transformada, principalmente para valores pequenos da potência computada, acaba por acomodar bem as observações 12, 16, 29 e, principalmente, a 21.

A Figura 4.4.a mostra que os resíduos não formam nenhuma tendência no gráfico de dispersão contra os índices das observações e põe em destaque as observações 31 e 33 por

(a)

(b)

Figura 4.3: Gráficos de dispersão e reta ajustada do MRLN com transformação logito na variável resposta e transformação logarítmica na variável regressora.

estarem pouco além do intervalo $(-2,2)$. Também não se observa tendência claramente definida na Figura 4.4.b, sugerindo que não existe ausência de termo extra. Segundo o critério da distância de Cook, existem indícios de que a observação 31 seja a mais influente, conforme indica a Figura 4.4.c, enquanto a observação 21 destaca-se como possível ponto de alavanca, conforme Figura 4.4.d. Não existem indícios de violação da suposição de homoscedasticidade, conforme Figura 4.4.e. De fato, o teste de Breusch-Pagan, aplicado aos resíduos deste modelo, não encontra evidências estatísticas para rejeitar a hipótese de homoscedasticidade dos resíduos ($p\text{-valor} = 0.70$). Por fim, a Figura 4.4.f indica que as observações não estão bem distribuídas dentro dos limites do gráfico de envelope, com algumas observações extrapolando estes limites. O teste de normalidade de Anderson-Darling aplicado sobre os resíduos padronizados deste modelo corrobora os indícios do gráfico de envelope, apresentando, para um nível de significância de 5%, um nível descritivo no limiar da rejeição da hipótese de normalidade ($p\text{-valor} = 0.0545$).

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.4: Gráficos de diagnóstico do MRLN com transformação logito na variável resposta e transformação logarítmica na variável regressora.

4.3 MRLN com transformação logarítmica na variável resposta

Nesta seção é considerado o modelo

$$y_i^* = \beta_0^* + \beta_1 x_i + \epsilon_i^*, \quad i = 1, \dots, n,$$

em que $y_i^* = \ln(y_i)$, $\beta_0^* = \ln(\beta_0)$ e $\epsilon_1^*, \dots, \epsilon_n^*$ são os erros aleatórios i.i.d. tais que $\epsilon_i^* \sim \mathcal{N}(0, \sigma^2)$, conforme definido na Seção 3.2 em (3.4). A Tabela 4.3 que mostra o resultado do ajuste deste modelo e indica que a potência, como variável regressora, contribui significativamente para explicar o fator de simultaneidade.

Tabela 4.3 Ajuste do MRLN com transformação logarítmica na variável resposta.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.6823	0.1255	-13.408	0.0000
β_1	-0.3071	0.0385	-7.976	0.0000
σ^2	0.1668	-	-	-

A Figura 4.5.a exibe o gráfico de dispersão entre a potência e o logaritmo natural do fator de simultaneidade, bem como a reta ajustada, e destaca as observações 31, 34 e 21 por situarem-se em uma região remota no eixo da variável regressora. A observação 21, por sua vez, destaca-se tanto no eixo da variável regressora quanto no eixo da variável resposta. A Figura 4.5.b exibe o gráfico de dispersão entre a potência e o fator de simultaneidade, bem como a reta ajustada transformada, e indica que as observações 12, 16, 29, e principalmente a 21 não são bem acomodadas pelo modelo.

A Figura 4.6.a, que mostra o gráfico de dispersão entre os resíduos studentizados e os índices das observações, não revela nenhuma tendência, mas põe em destaque a observação 21, caracterizando-a como observação aberrante. Também não se observa tendência claramente definida na Figura 4.6.b, que mostra a dispersão entre os resíduos studentizados e a variável regressora, sugerindo não haver ausência de termo extra. Há

(a)

(b)

Figura 4.5: Gráficos de dispersão e reta ajustada do MRLN com transformação logarítmica na variável resposta.

indícios de que as observações 21, 31 e 34 sejam as mais influentes, conforme indica a Figura 4.6.c, na qual se observa a distância de Cook contra os índices das observações. Além disso, analisando a medida h_{ii} , há evidências de que a observação 34 seja também um ponto de alavanca, conforme indica a Figura 4.6.d. A observação 21 destaca-se na Figura 4.6.e, na qual está o gráfico de dispersão entre resíduo studentizado e valor ajustado, mas não determina um padrão a ponto de suspeitar heteroscedasticidade dos resíduos. De fato, o teste de Breusch-Pagan, aplicado aos resíduos deste modelo, não encontra evidência estatística para rejeitar a hipótese de homoscedasticidade (p-valor = 0.24). Por fim, a Figura 4.6.f mostra o gráfico de envelope simulado e indica que os pontos encontram-se todos distribuídos dentro dos limites do gráfico de envelope, sugerindo aderência à suposição de distribuição do modelo. De fato, o teste de normalidade de Anderson-Darling aplicado sobre os resíduos padronizados deste modelo não encontra evidências estatísticas para rejeitar a hipótese de normalidade (p-valor = 0.81).

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.6: Gráficos de diagnóstico do MRLN com transformação logarítmica na variável resposta.

4.4 MRLN com transformação logarítmica nas variáveis resposta e regressora

Nesta seção é considerado o modelo

$$y_i^* = \beta_0^* + \beta_1 x_i^* + \epsilon_i^*, \quad i = 1, \dots, n,$$

em que $y_i^* = \ln(y_i)$, $x_i^* = \ln(x_i)$, $\beta_0^* = \ln(\beta_0)$ e $\epsilon_1^*, \dots, \epsilon_n^*$ são os erros aleatórios i.i.d. tais que $\epsilon_i^* \sim \mathcal{N}(0, \sigma^2)$. Este modelo, já definido na Seção 3.2 em (3.5), foi também considerado em IPT (2005). A Tabela 4.4 mostra o resultado do ajuste deste modelo e indica que o logaritmo natural da potência, como variável regressora, é estatisticamente importante para explicar o fator de simultaneidade.

Tabela 4.4 Ajuste do MRLN com transformação logarítmica nas variáveis resposta e regressora.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.9575	0.0889	-22.018	0.0000
β_1	-0.7136	0.0805	-8.865	0.0000
σ^2	0.1458	-	-	-

A Figura 4.7.a exibe o gráfico de dispersão entre o logaritmo natural da potência e o logaritmo natural do fator de simultaneidade, bem como a reta ajustada, e destaca as observações 21, 31 e 34 apenas por apresentarem valores extremos na variável regressora transformada, não parecendo haver grandes desvios em relação ao modelo de regressão ajustado. O mesmo pode-se dizer em relação à Figura 4.8.b., que exibe o gráfico de dispersão entre a potência e o fator de simultaneidade, bem como a reta ajustada transformada. Nota-se que, por conta do forte decaimento da reta ajustada transformada, observações com baixos valores de potência computada são bem acomodadas pelo modelo, em especial a 21.

A Figura 4.8.a mostra que os resíduos studentizados não formam nenhuma tendência no gráfico de dispersão contra a ordem de observação e põe em destaque as observações

(a)

(b)

Figura 4.7: Gráficos de dispersão e reta ajustada do MRLN com transformação logarítmica nas variáveis resposta e regressora.

31 e 33 por estarem pouco além do intervalo $(-2,2)$. Também não se observa tendência claramente definida na Figura 4.8.b, sugerindo que não existe ausência de termo extra. Segundo o critério da distância de Cook, há indícios de que as observações 31 e 34 sejam as mais influentes, conforme indica a Figura 4.8.c, enquanto a observação 21 destaca-se como possível ponto de alavanca segundo o critério da medida h_{ii} , conforme Figura 4.8.d. Não existem indícios de violação da suposição de homoscedasticidade, conforme Figura 4.8.e, que mostra o gráfico de dispersão entre os resíduos studentizados e os valores ajustados. De fato, o teste de Breusch-Pagan, aplicado aos resíduos deste modelo, não encontra evidências estatísticas para rejeitar a hipótese de homoscedasticidade dos resíduos ($p\text{-valor} = 0.31$). Por fim, a Figura 4.8.f indica que as observações estão dentro dos limites do gráfico de envelope, com algumas no limiar destes limites, sugerindo uma leve falta de qualidade de ajuste do modelo. O teste de normalidade de Anderson-Darling aplicado sobre os resíduos padronizados deste modelo rejeita a hipótese de normalidade a um nível de 5 % de significância ($p\text{-valor} = 0.0462$).

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.8: Gráficos de diagnóstico do MRLN com transformação logarítmica nas variáveis resposta e regressora.

4.5 Modelo de regressão beta

Nesta seção é considerado o modelo de regressão beta definido na Seção 3.3. Admite-se que $y_1, \dots, y_n$ são variáveis aleatórias independentes tal que y_i tem distribuição beta de média μ_i e parâmetro de dispersão ϕ , e ainda,

$$\ln \left(\frac{\mu_i}{1 - \mu_i} \right) = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n,$$

em que β_0 e β_1 são parâmetros desconhecidos do modelo. Note que aqui é utilizada a função de ligação $g(\mu_i) = \ln(\mu_i/(1 - \mu_i))$, ou seja a função logito. Esta função de ligação foi escolhida por ser particularmente útil em termos de interpretação de resultados, uma vez que possibilita a interpretação do parâmetro β_1 em termos de razão de chances. A Tabela 4.5, que mostra o resultado do ajuste do modelo, indica que a potência contribui significativamente para explicar o fator de simultaneidade.

Tabela 4.5 Ajuste do modelo de regressão beta.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.30368	0.1408	-9.26	0.0000
β_1	-0.36479	0.0536	-6.80	0.0000
ϕ	45.391	10.027	-	-

A Figura 4.9.a exibe o gráfico de dispersão entre a potência e o logito do fator de simultaneidade, bem como a reta ajustada, e destaca as observações 21, 31 e 34. As observações 31 e 34 destacam-se por situarem-se em uma região remota no eixo da variável regressora. A observação 21, por sua vez, destaca-se tanto no eixo da variável regressora quanto no eixo da variável resposta. Tal comportamento sugere que esta observação não é bem acomodada pelo modelo. A Figura 4.9.b exibe o gráfico de dispersão entre a potência o fator de simultaneidade, bem como a reta ajustada transformada, e sugere que, de modo geral, as observações estão bem acomodadas pela reta ajustada, exceção feita à observação 21.

Figura 4.9: Gráficos de dispersão e reta ajustada do modelo de regressão beta.

A Figura 4.10.a evidencia que os resíduos r_i^{ww} definidos em (3.15) não formam nenhuma tendência no gráfico de dispersão contra os índices das observações e põe em destaque a observação 21, caracterizando-a como observação aberrante. Na Figura 4.10.b, na qual é traçado o gráfico de dispersão entre os resíduos r_i^{ww} e os valores da variável regressora, observa-se uma tendência de afunilamento nos resíduos, sugerindo a possibilidade de melhora do ajuste do modelo caso seja adicionado um termo extra, possivelmente para permitir que o parâmetro de dispersão dependa da potência. Seguindo o critério da distância de Cook, há indícios de que a observação 21 seja a mais influente, conforme indica a Figura 4.10.c. Por outro lado, utilizando o critério dos valores de alavanca generalizada, a Figura 4.10.d sugere a inexistência de pontos de alavanca. A Figura 4.10.e., que mostra o gráfico de dispersão entre r_i^{ww} e os respectivos valores ajustados, sugere um comportamento heteroscedástico, na medida em que os resíduos parecem se dispersar mais para valores ajustados maiores. Por fim, a Figura 4.10.f indica que os resíduos não estão bem distribuídos dentro do gráfico de envelope simulado, com algumas delas extrapolando os seus limites, sugerindo portanto desvio em relação às suposições do modelo.

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.10: Gráficos de diagnóstico do modelo de regressão beta.

4.6 Modelo de regressão beta com transformação logarítmica na variável regressora

Nesta seção é considerado o modelo de regressão beta definido na Seção 3.3. Admite-se que $y_1, \dots, y_n$ são variáveis aleatórias independentes tal que y_i tem distribuição beta de média μ_i e parâmetro de dispersão ϕ , e ainda,

$$\ln\left(\frac{\mu_i}{1 - \mu_i}\right) = \beta_0 + \beta_1 x_i^*, \quad i = 1, \dots, n,$$

em que β_0 e β_1 são parâmetros desconhecidos do modelo, $g(\mu_i) = \ln(\mu_i/(1 - \mu_i))$ é a função de ligação logito e $x_i^* = \ln(x_i)$, conforme definido na Seção 3.3. A Tabela 4.6, que mostra o resultado do ajuste do modelo, indica que o logaritmo natural da potência computada é estatisticamente importante para explicar o fator de simultaneidade.

Tabela 4.6 Ajuste do modelo de regressão beta com transformação logarítmica na variável regressora.

Parâmetro	Estimativa	Erro Padrão	t	p-valor
β_0	-1.71223	0.06720	-25.5	0.0000
β_1	-0.79352	0.06654	-11.9	0.0000
ϕ	79.34491	17.40425	-	-

A Figura 4.11.a exibe o gráfico de dispersão entre o logaritmo da potência e o logito da variável resposta, bem como a reta ajustada, enquanto a Figura 4.11.b exibe o gráfico de dispersão entre a potência e o fator de simultaneidade, bem como a reta ajustada transformada. Ambas as figuras destacam as observações 21, 31 e 34 por apresentarem valores extremos na variável regressora transformada, não apresentando grandes desvios em relação ao modelo ajustado.

A Figura 4.12.a mostra que os resíduos não formam nenhuma tendência claramente definida no gráfico de dispersão contra a ordem de observação. Também não se observa nenhuma tendência no gráfico da Figura 4.12.b, sugerindo que não existe ausência de

Figura 4.11: Gráficos de dispersão e reta ajustada do modelo de regressão beta com transformação logarítmica na variável regressora.

termo extra. A observação 21 apresenta o maior valor residual do ajuste deste modelo. As observações 16, 33 e 35 apresentam valores de resíduo altos, mas estão relativamente próximas da massa de dados. Segundo o critério da distância de Cook, a observação 21 destaca-se como uma possível observação influente, conforme mostra a Figura 4.12.c. A observação 21 também é destaque como ponto de alavanca, juntamente com a observação 15, conforme indica a Figura 4.12.d. A Figura 4.12.e não sugere nenhum comportamento heteroscedático dos resíduos. Também não se observam desvios relevantes em relação à suposição de distribuição do modelo, conforme Figura 4.12.f, uma vez que as observações estão todas distribuídas dentro dos envelopes ou bem próximos de seus limites.

4.7 Comparação entre os modelos

A proposta desta seção é escolher, entre os seis modelos ajustados nas seções anteriores, o modelo mais adequado aos dados. Em uma análise inicial, serão descartados os modelos que apresentaram algum desvio evidente em relação às suas suposições. Em

(a)

(b)

(c)

(d)

(e)

(f)

Figura 4.12: Gráficos de diagnóstico do modelo de regressão beta com transformação logarítmica na variável regressora.

seguida, para os modelos remanescentes, será avaliado o impacto da retirada de observações possivelmente influentes que, segundo os gráficos de diagnóstico apresentados nas seções anteriores, estão presentes em todos os modelos. Ao fim desta análise, será escolhido um único modelo que será trabalhado de modo a fornecer uma banda de predição, cujo limite superior será tomado como a proposta da curva do fator de simultaneidade.

Para o MRLN com transformação logito na variável resposta, cujos resultados são apresentados na Seção 4.1, nota-se que a observação 21 extrapola os limites do gráfico de envelope simulado, sugerindo uma leve falta de qualidade de ajuste. Além disso, fazendo uma análise conservadora deste modelo, a suposição de homoscedasticidade é violada (p -valor = 0.099 no teste de Breusch-Pagan), sendo também a observação 21 a responsável por este resultado. Desta forma, o modelo em questão caracteriza-se por não acomodar de forma satisfatória a observação 21.

Para o MRLN com transformação logito na variável resposta e transformação logarítmica nas variáveis regressora, cujos resultados são apresentados na Seção 4.2, o principal desvio fica por conta do teste de Anderson-Darling, que gerou um nível descritivo no limiar da rejeição da hipótese de normalidade a um nível de significância de 5% (p -valor = 0.0545). Desta forma, o modelo em questão caracteriza-se por violar a suposição de normalidade, comprometendo assim os resultados inferenciais.

Para o MRLN com transformação logarítmica na variável resposta, cujos resultados são apresentados na Seção 4.3, não há nenhuma evidência de desvio em relação às suposições do modelo. Desta forma, o modelo em questão constitui uma alternativa a ser considerada.

Similarmente ao modelo apresentado na Seção 4.2, para o MRLN com transformação logarítmica nas variáveis resposta e regressora, cujos resultados são apresentados na Seção 4.4, o principal desvio também fica por conta do teste de Anderson-Darling, que rejeita a hipótese de normalidade a um nível de significância de 5% (p -valor = 0.0462). Desta forma, o modelo em questão caracteriza-se por violar a suposição de normalidade, comprometendo assim os resultados inferenciais.

Para o modelo de regressão beta, cujos resultados são apresentados na Seção 4.5, foram encontrados diversos desvios. Foram encontrados indícios de ausência de termo extra, dispersão variável e desvio na suposição da distribuição do modelo, configurando-se, portanto, uma situação na qual o modelo é inadequado para os dados considerados.

Finalmente, para o modelo de regressão beta com transformação logarítmica na variável regressora, cujos resultados são apresentados na Seção 4.6, não há nenhuma evidência de desvio em relação às suposições do modelo. Desta forma, o modelo em questão constitui uma alternativa a ser considerada.

Portanto, as técnicas de diagnóstico utilizadas neste trabalho não evidenciaram razões para descartar dois dos seis modelos testados nesta dissertação: o MRLN com transformação logarítmica na variável resposta e o modelo de regressão beta com transformação logarítmica na variável regressora.

Uma grande diferença entre estes dois modelos pode ser observada na Figura 4.13, que contrapõe as duas curvas ajustadas transformadas. A queda acentuada da curva do modelo de regressão beta, observada na Figura 4.13.b, é devida à transformação logarítmica da variável regressora. Comportamento similar foi observado em todos os modelos que consideraram esta transformação. A pergunta a ser respondida agora é: qual destas duas curvas melhor se aproxima do verdadeiro comportamento do fator de simultaneidade?

Nota-se que, entre as duas curvas, quase não há diferença em estimar o fator de simultaneidade médio para potências computadas superiores a 1 *MW*. A maior diferença entre as duas curvas, portanto, envolve a estimativa de fatores de simultaneidade inferiores a 1 *MW*. Como o escopo deste trabalho é apenas sistemas prediais cuja utilização de gás natural é para o aquecimento de água e cocção, as chances de observar potências computadas com valores baixos ficou bastante reduzida. De fato, nota-se que os valores observados de potência computada são superiores a 0.5 *MW*, com exceção da observação 21, cujo valor de potência computada foi 0.17 *MW*. Portanto, este trabalho não levantou massa de dados que possa evidenciar, com algum grau de certeza, o comportamento do

(a)

(b)

Figura 4.13: Gráficos de dispersão e reta ajustada do MRLN com transformação logarítmica na variável resposta (a) e do modelo de regressão beta com transformação logarítmica na variável regressora (b).

fator de simultaneidade no intervalo que vai de 0 a 0.5 MW .

A primeira evidência de que a transformação logarítmica é a mais adequada aos dados vem da análise das observações que estão no intervalo (0.5 MW , 1 MW) de potência computada. Estas observações são melhores acomodadas pelo modelo de regressão beta, cuja tendência de decaimento é mais acentuada nesta região. A observação 21 insere-se nesta interpretação como uma continuação natural desta tendência. A segunda evidência vem dos diversos outros estudos ligados ao fator de simultaneidade que evidenciam que, à medida que os valores de potência computada se aproximam de 0, o valor do fator de simultaneidade tende a aumentar, e que para valores muito baixos de potência computada, é bastante provável que o fator de simultaneidade tenda ao valor 1 (ver por exemplo, Ilha, 1996 e Kawabe, Ogassavara e Dias, 1989). Esta situação é claramente contemplada no modelo de regressão beta, mas não no MRLN. Portanto, embora não existam razões estatísticas para descartar os modelos apresentados nas Seções 4.3 e 4.6, a literatura evidencia que o MRLN com transformação logarítmica na variável resposta não esteja retratando o verdadeiro comportamento do fator de simultaneidade devido a um

problema de falta de informação em um intervalo de valores da potência. Desta forma, o modelo de regressão beta com transformação logarítmica na variável regressora apresenta-se como o modelo mais adequado neste trabalho.

Há ainda de se retomar a discussão sobre as premissas que tornam o modelo de regressão beta teoricamente mais adequado a dados medidos de forma contínua e limitados ao intervalo $(0, 1)$, conforme já mencionado no Capítulo 3. O MRLN apresenta duas inconsistências conceituais quando aplicados em dados compreendidos no intervalo $(0, 1)$. Primeiro, a variável resposta não está definida no domínio sobre o qual a distribuição normal é definida. Segundo, dados deste tipo não admitem que a função de esperança condicional seja linear e a função de variância condicional seja independente da média. A transformação da variável resposta busca contornar estes erros conceituais, mas ao mesmo tempo adicionam dificuldades na interpretabilidade do modelo, uma vez que os resultados obtidos referem-se à variável resposta transformada, e não à variável resposta original. O modelo de regressão beta, por sua vez, apresenta uma série de vantagens em relação aos dados compreendidos entre $(0, 1)$, conforme também mencionado no Capítulo 3. Em primeiro lugar, todas as suposições do modelo são referentes à variável resposta original, e não à variável resposta transformada, facilitando a interpretação dos parâmetros do modelo. A utilização da ligação logito restringe a esperança condicional ao intervalo $(0, 1)$. Por fim, variáveis respostas com variâncias não-constantes, situação muito comum em dados compreendidos no intervalo $(0, 1)$, podem ser naturalmente acomodadas pelo modelo de regressão beta. Desta forma, as adequações quase que naturais da regressão beta a dados compreendidos no intervalo $(0, 1)$ e a facilidade de interpretação dos resultados também são fatores que favorecem a escolha do modelo apresentado na Seção 4.6.

Uma vez definido o modelo a ser considerado nesta dissertação, é importante avaliar o impacto de observações possivelmente influentes sobre as estimativas do modelo, conforme apontado pelos gráficos de diagnóstico da Seção 4.6. A Tabela 4.7 mostra a variação das estimativas mediante a retirada das observações possivelmente influentes. Sob o efeito da retirada da observação 21, não há grandes impactos sobre as estimativas de β_0 e β_1 . A variação observada no parâmetro β_1 é inferior a 8% e as conclusões sobre a importância da potência para explicar o fator de simultaneidade permanece inalterada. Em relação ao

parâmetro de dispersão ϕ , a variação observada com a retirada da observação é inferior a 1%. A Figura 4.14 mostra as curvas ajustadas com e sem a observação 21 evidenciando que, de fato, quase não há impacto no ajuste das curvas mediante retirada da observação 21.

Tabela 4.7 Estimativas do modelo de regressão beta com transformação logarítmica na variável resposta.

Parâmetro	Todas	Sem 21	$\Delta\%$
β_0	-1.7122	-1.7641	3.0
p-valor	0.0000	0.0000	
β_1	-0.7935	-0.7305	-7.9
p-valor	0.0000	0.0000	-
ϕ	79.3449	78.7392	-0.8

(a)

(b)

Figura 4.14: Efeito da retirada da observação 21 no modelo de regressão beta com transformação logarítmica na variável resposta.

Outro passo importante consiste na interpretação do modelo ajustado, ou seja,

$$\ln \left(\frac{\hat{\mu}}{1 - \hat{\mu}} \right) = -1.71 - 0.79 x^*$$

em que $x^* = \ln(x)$, com x representando a potência computada (em MW). Isolando $\hat{\mu}$, temos que

$$\hat{\mu} = \frac{e^{(-1.71 - 0.79 x^*)}}{1 + e^{(-1.71 - 0.79 x^*)}}.$$

Como é usual em modelos que levam em conta a função logito como função de ligação, poder-se-ia interpretar o modelo ajustado em termos de razão de chances por meio da medida $e^{(c \times \hat{\beta}_1)}$, em que c é um incremento na variável regressora. No entanto, como a variável resposta aqui estudada não representa uma probabilidade, não é conveniente interpretá-la como tal.

Nota-se, porém, que $e^{(-0.1 \times \hat{\beta}_1)} = e^{(0.1 \times 0.79)} \approx 1.08$, o que significa que, a cada redução de 0.1 no logaritmo natural da potência (aproximadamente 1MW na potência computada), a utilização simultânea média dos aparelhos de utilização de um sistema predial aumenta em 8%.

Capítulo 5

Predição do fator de simultaneidade

5.1 Introdução

O modelo de regressão beta com transformação logarítmica na variável resposta foi o modelo que melhor se ajustou aos dados de fator de simultaneidade, conforme Seção 4.7. As técnicas de diagnóstico apresentadas na Seção 3.4 contribuíram para concluir que este modelo definiu a curva média que melhor acomodou as observações no gráfico de dispersão entre fator de simultaneidade e potência computada.

No entanto, na prática, não se pode utilizar a estimativa da esperança condicional do fator de simultaneidade para dimensionar as instalações de um sistema predial de gás natural. Se fosse adotado o modelo da Seção 4.6 para definir os fatores de simultaneidade dos novos projetos de sistema predial de gás natural, seria observado um grande número de sistemas prediais com fator de simultaneidade subestimado, como mostra a Figura 4.11.b. Conforme já mencionado no Capítulo 1, subestimar o fator de simultaneidade compromete a qualidade do serviço de fornecimento de gás natural e a segurança dos usuários.

É necessário, portanto, que na escolha do fator de simultaneidade, assegure-se, com algum grau de confiança, que o novo sistema predial a ser projetado não tenha seu fator de simultaneidade subestimado. Estatisticamente, este desafio traduz-se em obter um intervalo de predição para o fator de simultaneidade de novas edificações. Nesta dissertação

será utilizado o método de *bootstrap*, o qual será escrito nas seções seguintes.

5.2 O método *bootstrap*

O *bootstrap* é um método computacional de inferência estatística proposto por Efron (1979) que vem se mostrando útil em responder a questões práticas sem a necessidade de cálculos analíticos complexos ou inviáveis. A fim de descrever as suas premissas básicas, consideremos uma amostra $y = (y_1, \dots, y_n)$ de uma variável aleatória populacional Y , cujo comportamento probabilístico é completamente descrito por sua função de distribuição acumulada $\mathbb{F}$. Consideremos ainda que $\theta = t(\mathbb{F})$ é um parâmetro de interesse e que $\hat{\theta} = S(x)$ é seu estimador. Supondo ser inviável a obtenção das propriedades da distribuição de $\hat{\theta}$ por meio da teoria assintótica, pode-se considerar a utilização do *bootstrap*.

A aplicação do *bootstrap* consiste basicamente em, a partir de y , obter um grande número de amostras e um igual número de estimativas *bootstrap* de $\hat{\theta}^* = S(x^*)$. Com base na distribuição empírica de $\hat{\theta}^*$ pode-se então estimar a função de distribuição de $\hat{\theta}$.

A forma de obtenção das amostras *bootstrap* definem duas versões para o método: não-paramétrica e paramétrica. No *bootstrap* não-paramétrico, a amostra original y assume *status* de população, sobre a qual são extraídas, aleatoriamente e com reposição, as amostras *bootstrap* y^* . Formalizando este procedimento em termos estatísticos, as amostras *bootstrap* são obtidas de uma estimativa não-paramétrica de $\mathbb{F}$, que é empiricamente estimada por

$$\hat{\mathbb{F}}(t) = \frac{\#\{y_i \leq t\}}{n}, t \in \mathbb{R}, \quad (5.1)$$

que atribui probabilidade $1/n$ para cada elemento y_i , $i = 1, \dots, n$. A versão paramétrica do *bootstrap* pode ser utilizada quando a distribuição $\mathbb{F}$ pertence a uma família paramétrica finita e conhecida de distribuições $\mathbb{F}(\tau)$. Neste caso, na possibilidade de obter $\hat{\tau}$, uma estimativa consistente para τ , pode-se definir a estimativa paramétrica de $\mathbb{F}$, $\mathbb{F}(\hat{\tau})$. Neste trabalho será utilizada a versão não-paramétrica do método de *bootstrap*.

5.3 Intervalos de predição *bootstrap*

Na prática, a curva do fator de simultaneidade é utilizada para a determinação do fator em novas edificações. Em termos estatísticos, deseja-se prever o fator de simultaneidade de uma nova edificação utilizando um modelo de regressão previamente ajustado com base em um conjunto de n pares de dados $((x_1, y_1), \dots, (x_n, y_n))$. Sejam x_+ a potência computada de uma nova edificação, pertencente ou não ao conjunto de dados original, e y_+ o respectivo fator de simultaneidade não observado. A predição pontual deste novo valor do fator de simultaneidade é dada por $\hat{\mu}_+ = g^{-1}(\hat{\beta}_0 + x_+ \hat{\beta}_1)$, em que $g^{-1}(\cdot)$ é a função inversa da função de ligação e $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de máxima verossimilhança dos parâmetros do modelo, obtidos com base no conjunto de dados original. Além da predição pontual, é também conveniente estabelecer os limites de predição. Estes limites formam um intervalo de predição baseado em níveis de confiança, e são construídos com base na distribuição aproximada do erro de predição.

Seja $\mathcal{R}(y, \mu)$ uma função monótona crescente em y com variância constante. Suponha que a média μ_+ e a distribuição de $\mathcal{R}(y, \mu)$ sejam conhecidas e que q_α seja o α -ésimo quantil desta distribuição. Logo, os limites de predição para um intervalo com nível nominal $1 - \alpha$ são os valores $y_{+, \alpha/2}$ e $y_{+, 1-\alpha/2}$ que satisfazem, respectivamente, $\mathcal{R}(y, \mu_+) = q_{(\alpha/2)}$ e $\mathcal{R}(y, \mu_+) = q_{(1-\alpha/2)}$. Se μ é estimado por $\hat{\mu}$ independentemente de y_+ e se $\mathcal{R}(y_+, \hat{\mu})$ tem quantis conhecidos, o mesmo método se aplica. Se a distribuição de $\mathcal{R}(y_+, \hat{\mu})$ é desconhecida, a mesma pode ser aproximada por métodos assintóticos ou por métodos de reamostragem, tal como o método *bootstrap*, por exemplo. Desta forma, por meio do método *bootstrap* é possível obter uma aproximação da distribuição de $\mathcal{R}(y_+, \hat{\mu})$, e assim obter os quantis empíricos utilizados na obtenção dos limites de predição. A função $\mathcal{R}(y, \mu)$ corresponde a alguma definição de resíduo do modelo. O algoritmo *bootstrap* utilizará para o processo de reamostragem uma versão padronizada da função $\mathcal{R}(y, \mu)$, cuja distribuição tenha variância aproximadamente constante.

Para o modelo de regressão beta, Ospina (2007) propõe que

$$\mathcal{R}(y, \hat{\mu}) = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{v_i}},$$

ou seja, o resíduo padronizado 1 tal como definido em (3.14). No processo de reamostragem, propõe-se utilizar a padronização do resíduo padronizado 1, ou seja,

$$r_i^{ww} = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{v_i(1 - h_{ii}^*)}} \quad ,$$

que é o resíduo padronizado 2 tal como definido em 3.15. Para desenvolver o método *bootstrap* para a obtenção de limites de predição empíricos para o modelo de regressão beta, Ospina (2007) baseou-se no método proposto por Davison e Hinkley (1997) para modelos lineares generalizados. Em sua tese, a autora detalha o algoritmo para a obtenção dos limites de predição empíricos, mostra um exemplo de aplicação e conclui que o intervalo de predição *bootstrap* tem, em geral, bom desempenho, mas tende a falhar quando usado para prever casos que seriam considerados influentes se estivessem presentes no conjunto de dados observados.

O método proposto por Ospina (2007) foi a ferramenta utilizada para a obtenção do intervalo de predição do modelo ajustado na Seção 4.6. Implementada na plataforma computacional Ox, a saída do método computacional são os valores referentes ao intervalo de predição *bootstrap*. Dentro do intervalo (0, 10] (em kW), a cada incremento de 1W no valor da potência computada, foi empregado o método *bootstrap*, gerando assim 10000 valores preditos do fator de simultaneidade com seus respectivos limites inferiores e superiores de intervalo de predição *bootstrap*. Aqui, foram considerados limites de predição com coeficiente de confiança nominal igual a 95%.

Os resultados gerados pelos limites inferiores do intervalo de predição não têm utilidade prática neste trabalho. O maior interesse está no limite superior do intervalo de predição, cujos valores formam o gráfico exibido na Figura 5.1.a.

Aparentemente, a “curva” exibida na Figura 5.1.a é contínua e monotonicamente decrescente. Na verdade, esta figura é formada pela sucessão de pontos bem próximos que carregam a variação natural existente em uma técnica de *bootstrap*, e que pode ser observada por alguns pontos que se distanciam da curva principal. A fim de eliminar

esta variação, foi empregada a técnica de suavização de gráficos de dispersão descrita em Cleveland (1979) e Cleveland (1981). Esta técnica, conhecida como *LOWESS* (das iniciais de *Locally Weighted Robust Scatterplot Smoothing*), utiliza modelos de regressão ponderados robustos para gerar, com base nos pontos de um gráfico de dispersão, uma curva suavizada. Aqui, foram tomados os pares formados pelos valores das potências utilizadas no método *bootstrap* e os respectivos valores obtidos para o limite superior do intervalo de predição, totalizando assim 10000 pontos. O resultado deste procedimento pode ser observado na Figura 5.1.b.

Figura 5.1: Limite superior do intervalo de predição *bootstrap*: dados brutos (a) e sob efeito da técnica *LOWESS* de suavização de curvas (b).

A utilização da técnica do *LOWESS* é o passo final para a obtenção de uma nova curva de predição para o fator de simultaneidade com base na potência computada. No entanto, diferentemente das abordagens da norma técnica NBR 14570 e de IPT(2005), a abordagem desta dissertação não apresenta como resultado final uma fórmula fechada para o fator de simultaneidade. O resultado final deste trabalho é uma tabela que relaciona o fator de simultaneidade com a potência computada, cujos valores constituíram o gráfico da Figura 5.1.b. Estes valores podem ser observados na Tabela 5.1.

Tabela 5.1 Potências computadas (kW) e seus respectivos fatores de simultaneidade

MW	F.S.	MW	F.S.	MW	F.S.	MW	F.S.	MW	F.S.	MW	F.S.
0.00	0.842	0.34	0.417	0.68	0.304	1.02	0.251	1.36	0.219	3.50	0.143
0.01	0.823	0.35	0.412	0.69	0.302	1.03	0.250	1.37	0.218	3.60	0.142
0.02	0.803	0.36	0.407	0.70	0.299	1.04	0.249	1.38	0.218	3.70	0.140
0.03	0.784	0.37	0.402	0.71	0.298	1.05	0.248	1.39	0.217	3.80	0.138
0.04	0.764	0.38	0.397	0.72	0.296	1.06	0.247	1.40	0.216	3.90	0.137
0.05	0.745	0.39	0.391	0.73	0.294	1.07	0.245	1.41	0.215	4.00	0.136
0.06	0.726	0.40	0.387	0.74	0.292	1.08	0.244	1.42	0.215	4.10	0.134
0.07	0.706	0.41	0.383	0.75	0.290	1.09	0.243	1.43	0.214	4.20	0.133
0.08	0.687	0.42	0.379	0.76	0.289	1.10	0.242	1.44	0.213	4.30	0.132
0.09	0.667	0.43	0.375	0.77	0.287	1.11	0.241	1.45	0.213	4.40	0.131
0.10	0.648	0.44	0.372	0.78	0.285	1.12	0.240	1.46	0.212	4.50	0.130
0.11	0.635	0.45	0.368	0.79	0.283	1.13	0.239	1.47	0.211	4.60	0.128
0.12	0.621	0.46	0.364	0.80	0.281	1.14	0.238	1.48	0.210	4.70	0.128
0.13	0.607	0.47	0.361	0.81	0.280	1.15	0.237	1.49	0.210	4.80	0.127
0.14	0.594	0.48	0.357	0.82	0.278	1.16	0.236	1.50	0.209	4.90	0.125
0.15	0.580	0.49	0.353	0.83	0.277	1.17	0.235	1.60	0.203	5.00	0.125
0.16	0.567	0.50	0.350	0.84	0.275	1.18	0.234	1.70	0.197	5.25	0.122
0.17	0.553	0.51	0.347	0.85	0.273	1.19	0.233	1.80	0.192	5.50	0.120
0.18	0.540	0.52	0.344	0.86	0.272	1.20	0.233	1.90	0.187	5.75	0.119
0.19	0.526	0.53	0.341	0.87	0.270	1.21	0.232	2.00	0.183	6.00	0.117
0.20	0.513	0.54	0.339	0.88	0.269	1.22	0.231	2.10	0.179	6.25	0.115
0.21	0.505	0.55	0.336	0.89	0.267	1.23	0.230	2.20	0.175	6.50	0.114
0.22	0.498	0.56	0.333	0.90	0.266	1.24	0.229	2.30	0.172	6.75	0.112
0.23	0.490	0.57	0.330	0.91	0.264	1.25	0.228	2.40	0.169	7.00	0.111
0.24	0.482	0.58	0.327	0.92	0.263	1.26	0.227	2.50	0.166	7.25	0.109
0.25	0.475	0.59	0.325	0.93	0.262	1.27	0.226	2.60	0.163	7.50	0.109
0.26	0.467	0.60	0.322	0.94	0.261	1.28	0.225	2.70	0.160	7.75	0.107
0.27	0.460	0.61	0.320	0.95	0.259	1.29	0.225	2.80	0.157	8.00	0.107
0.28	0.452	0.62	0.317	0.96	0.258	1.30	0.224	2.90	0.155	8.25	0.106
0.29	0.445	0.63	0.315	0.97	0.257	1.31	0.223	3.00	0.153	8.50	0.105
0.30	0.437	0.64	0.313	0.98	0.256	1.32	0.222	3.10	0.151	8.75	0.104
0.31	0.432	0.65	0.311	0.99	0.254	1.33	0.221	3.20	0.149	9.00	0.103
0.32	0.427	0.66	0.308	1.00	0.253	1.34	0.221	3.30	0.147	9.50	0.102
0.33	0.422	0.67	0.306	1.01	0.252	1.35	0.220	3.40	0.145	10.00	0.101

Capítulo 6

Considerações finais

6.1 Conclusões

Neste trabalho, foram apresentadas diversas abordagens possíveis para a predição do fator de simultaneidade por meio de modelos de regressão para proporções contínuas. O objetivo foi identificar quais destas abordagens melhor se adapta ao conjunto de dados utilizado e, assim, propor uma “curva” do fator de simultaneidade.

De modo geral, pode-se dividir os modelos utilizados neste trabalho em três classes: modelos de regressão lineares normais, modelos de regressão lineares normais motivados por modelos de regressão não-lineares normais, e modelos de regressão beta. As duas primeiras classes de modelos são alternativas largamente empregadas na literatura científica, conforme mostram Kieschnick e McCullough (2003). A comparação entre os modelos ocorreu por meio das diversas técnicas de diagnósticos apresentadas na Seção 3.4, que ajudaram a identificar desvios nas suposições dos modelos e observações influentes. Por razões já apresentadas na Seção 4.7, o modelo mais adequado aos dados deste trabalho foi o modelo de regressão beta com transformação logarítmica na variável regressora, conclusão esta que vai ao encontro das expectativas iniciais do trabalho.

A simples escolha de um modelo que se ajustou melhor aos dados não fornece uma solução ao problema de predição do fator de simultaneidade, uma vez que um modelo bem ajustado forneceria apenas uma boa estimativa do fator de simultaneidade médio

para um dado valor de potência. Na prática, é necessário que, ao se determinar o fator de simultaneidade para uma nova edificação, assegure-se com algum grau de confiança, que o fator de simultaneidade não seja subestimado. Em termos estatísticos, foi necessário obter um intervalo de predição para o fator de simultaneidade baseado no modelo de regressão beta. Para tanto, foi aplicada a teoria apresentada na Seção 5.1 - e desenvolvida em detalhes por Ospina (2007) - que trata do desenvolvimento da técnica de *bootstrap* para predição em modelos de regressão beta. A utilização do *bootstrap* em conjunto com a técnica de suavização *LOWESS* forneceu o resultado final deste trabalho, apresentado na Figura 5.1 e na Tabela 5.1.

Por fim, pode-se afirmar que a curva *bootstrap* obtida neste trabalho está alinhada aos estudos similares realizados recentemente, nos quais se alimenta a expectativa de que o fator de simultaneidade possa ser cada vez menor. Este trabalho está entre aqueles que vem comprovando esta expectativa, e que só foi possível ser realizado principalmente por dois fatores. Em primeiro lugar, atualmente existe maior facilidade de acesso aos dados de comportamento de consumo de gás natural graças à expansão da rede de distribuição. Em segundo lugar, o desenvolvimento computacional tem possibilitado a aplicação de diversas técnicas estatísticas que dependem de procedimentos computacionalmente intensivos, como é o caso do *bootstrap*.

6.2 Sugestões para trabalhos futuros

A partir dos resultados obtidos neste trabalho, surgem duas novas perspectivas de trabalho futuro, uma focada no aspecto prático dos problemas envolvendo o fator de simultaneidade e outra privilegiando o aspecto teórico dos modelos estatísticos a serem utilizados.

Sob o ponto de vista prático, a oportunidade surge quando se constata o escopo restrito do modelo obtido. Os dados levantados referem-se a sistemas prediais exclusivamente residenciais, situados na cidade de São Paulo e equipados com fogão e aquecedor de água. Portanto, os resultados obtidos não devem ser extrapolados para sistemas prediais com outras características. A inclusão de sistemas prediais com outras características pode

melhorar a predição do fator de simultaneidade e aumentar a sua abrangência. Em termos estatísticos, a sugestão é considerar outras variáveis regressoras que levem em conta a diversidade dos sistemas prediais, tais como região (fria ou quente, rica ou pobre), fins de utilização do gás (coocção e aquecimento, somente coocção, somente aquecimento), atividade (somente residencial, somente comercial, residencial e comercial), entre outras.

Sob o ponto de vista teórico, existem vários outros modelos que podem ser utilizados para modelar o fator de simultaneidade. Além dos já citados aqui, Kieschnick e McCullough (2003) citam o modelo de regressão normal censurado (Tobit), o modelo simplex e os modelos de quase-verossimilhança. Smithson e Verkuilen (2006) apresentam uma variação do modelo de regressão beta que considera a modelagem do parâmetro de dispersão. Deve-se ressaltar, no entanto, que o problema não se restringe ao ajuste do modelo aos dados. Deve-se também buscar uma solução para a obtenção de intervalos de predição, que é o passo no qual efetivamente se obtém a predição do fator de simultaneidade para novas edificações.

Apêndice A

Dados

ID	Potência(MW)	FS Observado	ID	Potência(kW)	FS Observado
1	4.210	0.064	22	1.202	0.112
2	1.499	0.092	23	1.666	0.125
3	3.268	0.031	24	2.379	0.076
4	3.268	0.079	25	3.453	0.086
5	1.955	0.099	26	2.962	0.116
6	3.167	0.062	27	4.896	0.029
7	2.135	0.114	28	3.546	0.063
8	2.690	0.068	29	0.812	0.222
9	1.387	0.070	30	2.837	0.067
10	1.304	0.069	31	5.949	0.016
11	2.130	0.041	32	1.148	0.124
12	0.635	0.249	33	3.124	0.147
13	3.066	0.089	34	8.786	0.018
14	2.593	0.047	35	0.918	0.088
15	0.591	0.148	36	2.682	0.066
16	0.823	0.276	37	3.206	0.066
17	4.989	0.060	38	2.947	0.059
18	1.957	0.123	39	4.495	0.047
19	3.081	0.062	40	3.975	0.058
20	3.524	0.070	41	4.863	0.061
21	0.169	0.464	42	4.078	0.069

Referências Bibliográficas

- [1] ABNT - Associação Brasileira de Normas Técnicas. (2002) NBR 14570 - Instalações Internas para Uso Alternativo dos Gases GN e GLP - Produto e Execução. Rio de Janeiro, ABNT.
- [2] Anderson, T. W., Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, 765 - 769.
- [3] Atkinson, A. C. (1985). *Plots, Transformations and Regression: An introduction to Graphical Methods of Diagnostic Regression Analysis*. New York, Oxford University Press.
- [4] Breusch, T. S., Pagan, A.R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47, 1287 - 1294.
- [5] Bussab, W. O., Morettin, P. A. (2002). *Estatística Básica*. 5a ed., São Paulo, Saraiva.
- [6] Carvalho, H. C. (1995). *Instalações Prediais de Gás: Manual de Dimensionamento*. Rio de Janeiro, Fundação Biblioteca Nacional, 1995.
- [7] Casella, G., Berger, R. L. (1996). *Statistical Inference*. California, Duxbury Press.

- [8] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829 - 836.
- [9] Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35, 54.
- [10] Comgás - Companhia de Gás de São Paulo. (1977). *Normas Técnicas para Utilização de Gás Combustível nos Edifícios e Construções em Geral*. São Paulo.
- [11] Cook, R. D. (1977). Detection of influential observations in linear regressions. *Technometrics*, 19, 15 - 18.
- [12] Cook, R.D., Peña, D., Weisberg, S. (1988). The likelihood displacement: A unifying principle for influence measures. *Communications in Statistics, Theory and Methods*, 17, 623 - 640.
- [13] Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regressions*. London, Chapman and Hall.
- [14] Cordeiro, G.M., Paula, G.A. (1992). Estimation, large-sample parametric tests and diagnostics for non-exponential family nonlinear models. *Communications in Statistics, Simulation and Computation*, 21, 149 - 172.
- [15] Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. New York, Cambridge University Press.
- [16] Draper, N. R., Smith, H. (1981). *Applied Regression Analysis*. New York, Wiley.

- [17] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- [18] Espinheira, P. L., Ferrari, S. L. P., Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistical*. A aparecer.
- [19] Everitt, B., S. (1994). *A Handbook os Statistical Analysis using S-Plus*. Chaaman and Hall, London.
- [20] Ferrari, S. L. P., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31, 799 - 815.
- [21] Hoaglin, D. C., Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32, 17-22.
- [22] Ilha, M. S. O. (1996). *Formulação de modelos para determinação da demanda e consumo de gás combustível em edifícios residenciais*. São Paulo, 1996. Tese (Doutorado) - Escola Politécnica, USP.
- [23] IPT - Instituto de Pesquisas Tecnológicas. (2005). Determinação do fator de simultaneidade em edificações residenciais multi-familiares. *Relatório Técnico*, São Paulo, 2005.
- [24] Kawabe, L., Ogassavara, G., Dias, R. P. (1989) Fator de simultaneidade. In: *Simpósio Brasileiro de Medição de Vazão*, 1., São Paulo, 1989. Anais. São Paulo, IPT, 117 - 123.

- [25] Kieschnick, R., McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. *Statistical Modelling*, 3, 193 - 213.
- [26] Mittlböck, M., Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15, 1987 - 1997.
- [27] Montgomery, D. C., Peck, E. A., Vining, G. G (2001). *Introduction to Linear Regression Analysis*. New York, Wiley.
- [28] Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*. Chicago, Irwin.
- [29] Nocedal, J., Wright, S. J. (1999). *Numerical Optimization*. New York, Springer-Verlag.
- [30] Oliveira, M., S. (2004). *Um Modelo de Regressão Beta: Teoria e Aplicações*. São Paulo, 2004. Dissertação (Mestrado) - IME, Universidade de São Paulo.
- [31] Ospina, P. L. E. (2007). *Regressão Beta*. São Paulo, 2007. Tese (Doutorado) - IME, Universidade de São Paulo.
- [32] Paula, G. A. (2003). *Modelos de Regressão com Apoio Computacional*. São Paulo, IME/USP. Disponível em <http://www.ime.usp.br/~giapaula>.
- [33] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705 - 724.
- [34] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York, Wiley.

- [35] Smithson, M., Verkuilen, J. (2006). A better lemon-squeezer? Maximum likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11, 54-71.

- [36] Wei, B. C., Hu, Y. Q., Fung, W. K. (1998). Generalized leverage and its applications. *Scandinavian Journal os Statistics*, 25, 25 - 37.