5
Resultados

Neste capítulo serão apresentados os resultados da modelagem sob as distribuições gama e normal inversa, que seguem os procedimentos estatísticos abordados ao longo deste estudo. Para a estimação do modelo e demais análises foi utilizado o software STATA 10, com apoio do Microsoft Excel 2010.

Como visto anteriormente, os dados de perdas não técnicas não seguem a distribuição normal e segundo a análise do seu histograma, Figura 14, a variável esta definida no conjunto do zero e reais positivos, assumindo formato assimétrico à direita. Logo, surgem duas distribuições da família exponencial que seriam candidatas naturais a assumir o papel de verdadeira distribuição dos dados. São elas: a distribuição gama e a normal inversa. Seja qual for a distribuição adotada, a função de ligação utilizado no modelo será a função logarítmica (log). A função log é definida entre os reais positivos, se apresentando assim, bastante adequada para o a natureza dos dados em questão. Com isso, tem-se a necessidade de comparar modelos e escolher o melhor dos dois. Logo, na próxima etapa desse capítulo, cabe apresentar os resultados da estimação dos dois diferentes modelos candidatos a modelo final: modelo gama e modelo normal inversa, ambos com função de ligação logarítmica. O modelo que apresentar melhor desempenho e adequabilidade será escolhido como modelo final.

5.1.
Resultados da Estimação dos Coeficientes

O primeiro modelo estimado é modelo gama. A estimação dos parâmetros é dada segundo o algoritmo iterativo de Newton-Raphson definido no Capítulo 4, seguindo ainda o método de stepwise. Assim, o processo começa com o modelo vazio e a primeira variável a entrar é o percentual de domicílios com até dois moradores por cômodos (doisporcom), seguida do percentual de domicílios com cobertura de esgoto (esgoto), índice de óbitos por agressão (óbitos) e percentual de domicílios precários (precários). Na Tabela 4 são apresentados os coeficientes

estimados das variáveis que se mostraram estatisticamente significativas a um nível de confiança de 95%.

Tabela 4 – Modelo Gama Link Log

PNT	**Coeficiente**	**Z**	$\rho > \|z\|$
doisporcom	-4,613963	-4,46	0,000
esgoto	-1,139347	-3,60	0,000
óbitos	0,2041841	2,57	0,010
precários	4,493474	2,03	0,043
constante	0,6437148	0,93	0,351
Estatísticas			
AIC	-2,158325	BIC	-143,3899
R² McFadden		55.7	

O próximo passo é estimar um novo modelo só que agora com a premissa de que a variável resposta seguiria distribuição normal inversa. O mesmo procedimento foi adotado, entretanto o software STATA 10 não conseguiu obter convergência para este modelo a partir do método de stepwise. Assim, optou-se por estimar o modelo normal inversa mantendo as variáveis que foram estatisticamente significativas no o modelo gama. Os resultados são apresentados na tabela a seguir.

Tabela 5 – Modelo Normal Inversa Link Log

PNT	**Coeficiente**	**z**	$\rho > \|z\|$
doisporcom	-.2579427	-2.27	0.023
esgoto	-5.618216	-3.65	0.000
óbitos	1.068497	2.85	0.004
precários	3.150469	1.01	0.314
constante	1.218227	1.00	0.319
Estatísticas			
AIC	5.957957	BIC	357.5739
R² McFadden		38.2	

É interessante fazer uma pequena análise dos sinais dos coeficientes, pois eles permitem uma análise do sentido que a variável tem para o modelo. Tanto no modelo gama quanto no modelo inversa normal, os sinais dos coeficientes estimados foram os mesmos e estão em concordância com a relação (pelo menos empírica) de cada variável com o percentual de perdas não técnicas.

A variável percentual de domicílios com até dois moradores por cômodo, por exemplo, tem coeficiente negativo, resultado esperado uma vez que domicílios onde não há aglomeração de moradores, geralmente refletem boas condições de vida e bom nível social. Assim, quanto maior o percentual de domicílios com até dois moradores em uma área de concessão acredita-se que menor deve ser o nível de perdas não técnicas, ou seja, essas variáveis são inversamente proporcionais. O mesmo acontece na interpretação do sentido da variável percentual de domicílios com rede de esgoto, esta também apresenta coeficiente negativo, o que pode ser entendido como: quanto maior a cobertura de infraestrutura nas áreas atendidas por uma determinada distribuidora, menor é seria o nível esperado de perdas não técnicas em sua concessão.

Por outro lado, a variável óbitos por agressão entrou nos modelos com sinal positivo, o que reforça a idéia de que áreas mais violentas e com maior grau de periculosidade, geralmente, são ambientes mais favoráveis a altos índices de perdas não técnicas. Por ultimo, a variável percentual domicílios precários também ficou com sinal positivo e fez valer as referências anteriores que apontavam que locais mais pobres, onde há precariedade nas condições de vida, seriam apresentariam perdas mais elevadas.

Sabendo-se que a relação entre as variáveis explicativas e a explicada tem o sentido esperado em ambos os modelos, é possível promover uma "competição" entre os modelos e ao final das análises, eleger qual deles é o melhor para estimar perdas não técnicas.

5.2.
Escolha do Modelo

Para escolher entre dois modelos estimados segundo distribuições diferentes devem-se utilizar os critérios de informação mencionados anteriormente, AIC e BIC. Neste caso em particular pode-se usar também o Log-

máxima verossimilhança já que os modelos têm o mesmo número de variáveis explicativas, logo a penalização pelo uso de muitos parâmetros é igual para ambos.

Os critérios AIC e BIC são calculados a partir do erro do modelo estimado e fazem uma penalização quanto ao nível de complexidade do modelo (número de variáveis), quanto menores forem o AIC e BIC, melhor é o modelo estimado. Geralmente se o AIC de um modelo é menor do que o AIC de outro modelo, o BIC também tende a seguir esta comparação. No caso dos dois modelos estimados, observa-se que o AIC e o BIC do modelo gama é menor.

Compara-se também, a função de log-verossimilhança dos modelos. O objetivo é saber qual dos modelos consegue maior valor na maximização da sua função de verossimilhança. A comparação pode ser feita pela função log já que a função logarítimo é uma função monotônica. Observa-se que o modelo gama apresenta maior valor na sua função de log-verossimilhança, que é definida a seguir.

Equação 17 – Função de log-verossimilhança

$$l(\theta; y) = log\,(L(\theta;y))$$

Onde:

$L(\theta; y)$ = Função de verossimilhança.

Além disso, pode-se comparar ainda o R^2 dos modelos. Como dito anteriormente, esta estatística é uma medida de qualidade de cada modelo em relação à sua habilidade de estimar corretamente os valores da variável resposta. Em outras palavras, o R^2 indica quanto da variável resposta é explicada pelas variáveis explicativas. Seu valor está compreendido no intervalo de 0 a 1: Quanto maior, mais explicativo é o modelo. Entretanto, sabe-se que para um MLG não é possível obter diretamente R^2, mas que se pode perfeitamente recorrer ao pseudo R^2 de Mcfadden. Para o modelo gama o R^2 de Mcfadden foi de 55,7%, enquanto que no modelo inversa normal o resultado foi de 38,2%.

Diante dos resultados obtidos decide-se continuar o estudo com o modelo gama. O que deve ser feito agora são os testes e gráficos para analisar os resíduos e validar o modelo, processo chamado de diagnóstico dos resíduos.

5.3.
Diagnóstico dos Resíduos

Um dos principais pressupostos dos modelos de regressão diz respeito à distribuição dos resíduos. Para estarem de acordo com a teoria estatística, os resíduos de um modelo devem ter média próxima de zero, variância pequena e serem distribuídos simetricamente, em outras palavras, os resíduos devem seguir distribuição normal. Para verificar se os resíduos são normalmente distribuídos foram elaborados alguns procedimentos gráficos, além de testes de hipóteses. A seguir são apresentados os resultados do qq-plot, histograma e do teste de Jarque-Bera para normalidade.

O qq-plot, gráfico a seguir, é utilizado para avaliar distribuições. Seus eixos são construídos contrastando os quantis teóricos de uma distribuição (neste caso, o contraste é feito com a normal) com os quantis observados a partir do conjunto de dados do estudo. O que significa dizer que quanto mais os pontos se comportam em cima da reta de 45°, mais próxima é a distribuição dos dados a aquela que se esta testando. Diante da análise do qq-plot dos resíduos percebe-se que a maioria dos pontos esta bem comportada e apenas duas observações se mostram um pouco mais distantes da reta balizadora[17]. Segundo o qq-plot há fortes indícios de que os resíduos são normais.

[17] Entretanto, essas observações só serão consideradas pontos problemáticos, ou não, depois deinvestigações futuras.

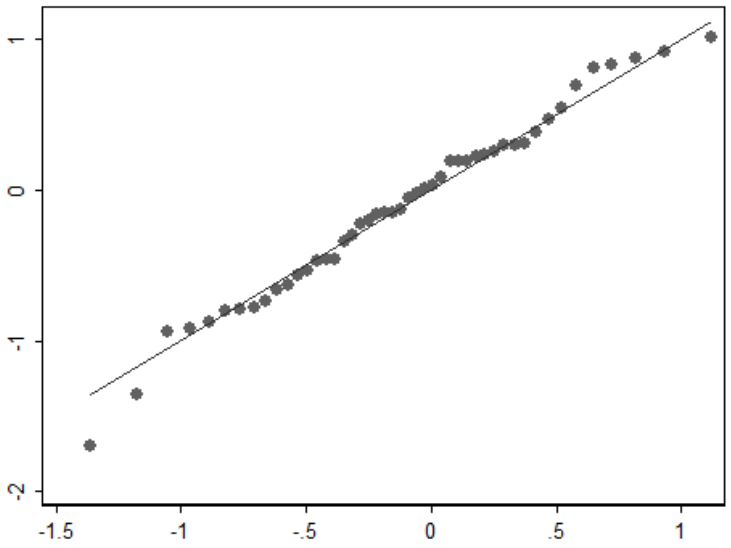

Figura 18 – QQ-Plot dos Resíduos

Outro gráfico importante para que se conheça a distribuição do conjunto de dados é o histograma. Analisando o histograma plotado na Figura 18, percebe-se que ele apresenta, quase que perfeitamente, a forma da densidade de uma normal. Há apenas uma pequena assimetria com relação à cauda direita, o que deve ser equivalente aos pontos que se distanciaram da reta no qq-plot, Figura 17.

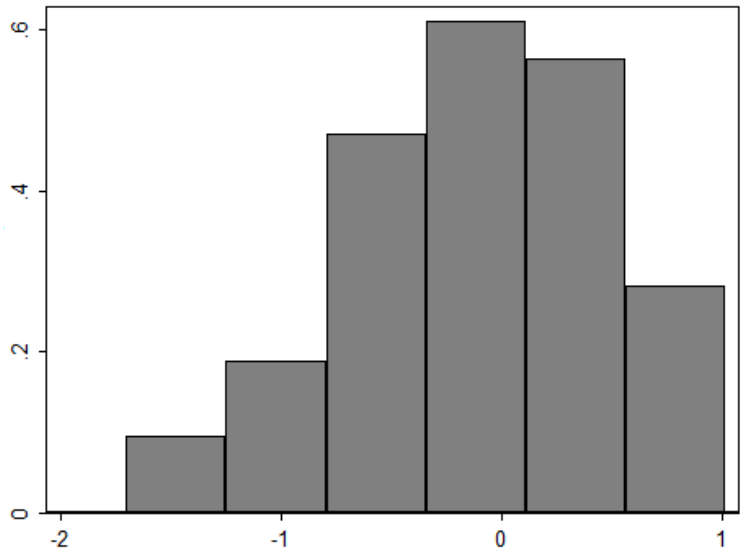

Figura 19 – Histograma dos Resíduos

Um teste formal apontará se essa assimetria compromete ou não a normalidade dos dados. Neste caso, o teste estatístico mais empregado para saber se um conjunto de dados segue distribuição normal é o teste Jarque-Bera. A estatística de decisão utilizada neste teste é conhecida com "JB" e segue da distribuição qui-quadrado assintótica com dois graus de liberdade. A estatística de teste é baseada no número de observações da amostra, na assimetria e curtose[18] de sua curva. A hipótese nula defende que a distribuição é normal enquanto a hipótese alternativa[19] que diz que a distribuição não é normal. Segundo o p-valor do teste, no caso, igual a 0,7837 não se rejeita a hipótese nula e pode-se concluir que os dados são normais.

Agora que se sabe que os resíduos seguem a distribuição normal, devem-se realizar procedimentos a fim confirmar que os resíduos são homocedásticos, ou seja, apresentam com variância constante. Para isso elaboraram-se dois gráficos de resíduos. O primeiro foi o gráfico de resíduos contra as observações, Figura 19, e o segundo consiste em resíduos contra o preditor linear, Figura 20. Caso esses gráficos mostrem a presença de alguma estrutura definida, como por exemplo, uma estrutura linear, quadrática ou em formato de cone, deve-se suspeitar de heterocedasticidade (variância não constante dos resíduos), por outro lado, caso a nuvem de pontos se mostre aleatoriamente distribuídas conclui-se que a variância dos resíduos é constante, e a análise prossegue.

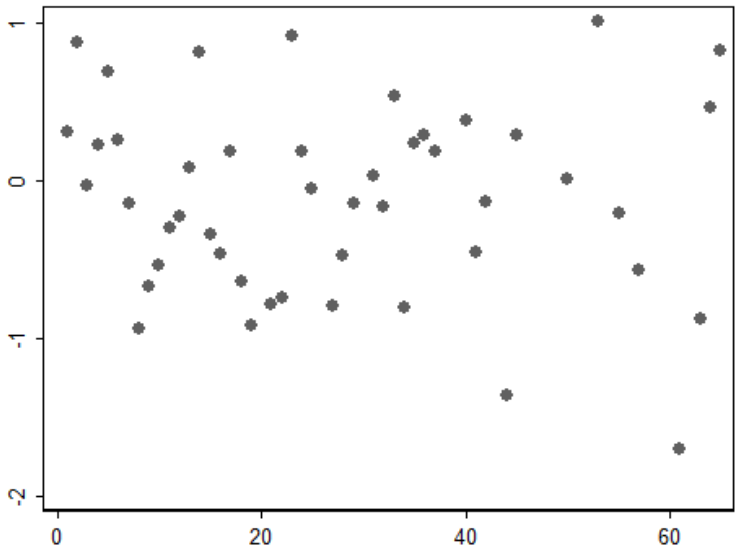

Figura 20 – Resíduos x Observações

[18] Achatamento.
[19] Gujarati (2006).

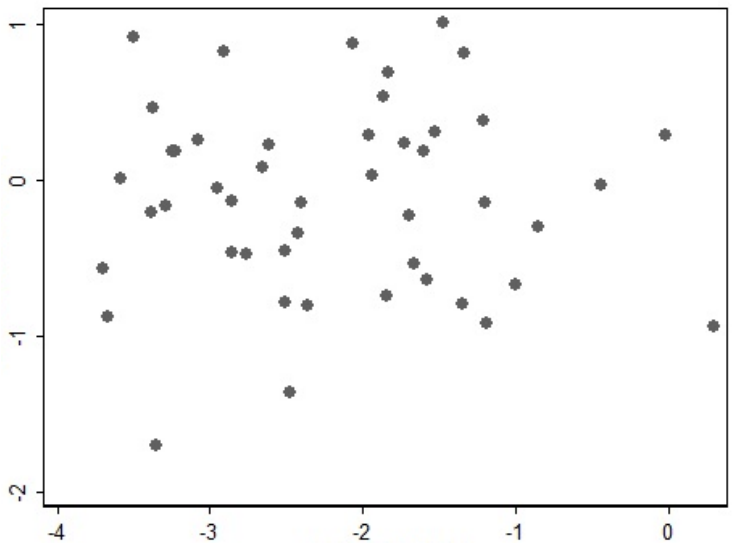

Figura 21 – Resíduos x Preditor Linear

Analisando a Figuras 19 e a Figura 20, não é possível, em princípio, identificar nenhuma estrutura definida ou tendência nos resíduos, isso dá fortes indícios para conclusão que os pontos são aleatoriamente distribuídos e que não há formação que sugerisse heterocedasticidade, sendo assim os resíduos são homocedásticos, como se desejava. Foi ainda desenvolvido um teste de White que visa corroborar a análise gráfica, e seu p-valor resultante foi de 0,8762, o que confirma que os resíduos são homocedásticos.

Deseja-se agora investigar se o modelo teve suas estimativas alteradas por observações atípicas, classificadas como “pontos influentes”. A presença de uma observação que possui grande influência nos resultados do modelo deve ser eliminada para que o ajuste volte a um padrão regular. Para identificar se determinada observação é passível de eliminação, analisa-se uma medida conhecida como distância de Cook. Esta medida tem por objetivo avaliar a influência individual de cada observação sobre a estimativa do vetor de coeficientes. Segundo Cadime e Silva (2009), devem ser investigados os pontos com valor da distância de Cook (D-Cook), maior que 0,5. Para um panorama geral da distância de Cook do modelo estimado foi elaborado um gráfico entre as distâncias e as observações.

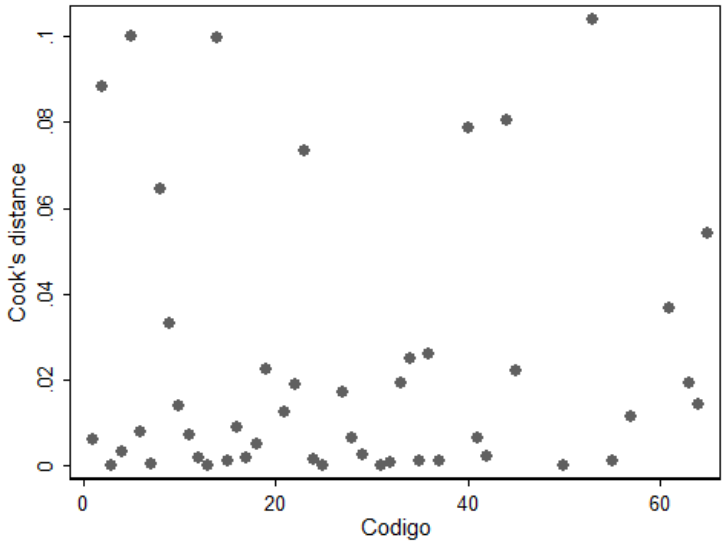

Figura 22 – Distâncias de Cook

Diante do gráfico da distância de Cook, Figura 21, percebe-se que nenhuma ultrapassa o limite de 0,5. Sendo assim, é possível concluir que o modelo esta livre de pontos influentes e que as análises podem continuar adiante, sem a necessidade de exclusão de qualquer observação, pelo menos sob o ponto de vista estatístico.

Com o modelo validado a partir das análises realizadas, podem ser calculados os resultados de perda não técnica estimados e posteriormente compará-los com os dados observados. Além dessas estatísticas, a Tabela 6 apresenta ainda valores do preditor linear do modelo e os resíduos transformados de Anscombe, para cada distribuidora presente no escopo do estudo.

Tabela 67 – Resumo das Estimativas

DISTRIBUIDORA	PNT OBSERVADA	PNT ESTIMADA	PREDITOR LINEAR	RESÍDUOS ANSCOMBE
CELPA	0,438	1,354	0,303	-0,940
ADESA	1,300	0,982	-0,018	0,294
CEAL	0,628	0,646	-0,437	-0,028
CEMAR	0,312	0,428	-0,849	-0,298
CELPE	0,173	0,367	-1,002	-0,666
COELCE	0,102	0,305	-1,189	-0,920
ELETROACRE	0,260	0,302	-1,198	-0,147
LIGHT	0,426	0,297	-1,212	0,381
CEPISA	0,539	0,263	-1,335	0,810
ENERGISA BORB	0,040	0,259	-1,350	-0,788
CERON	0,551	0,230	-1,471	1,014
AMPLA	0,292	0,218	-1,525	0,308
COELBA	0,102	0,207	-1,574	-0,634
ENERGISA SERG	0,037	0,201	-1,605	0,188
CELTINS	0,105	0,189	-1,665	-0,536
CEMAT	0,145	0,184	-1,695	-0,226
ENERG PARAÍBA	0,040	0,178	-1,728	0,233
CEEE	0,298	0,160	-1,832	0,690
COSERN	0,068	0,159	-1,839	-0,740
ENERSUL	0,255	0,155	-1,862	0,537
ELETROPAULO	0,149	0,144	-1,935	0,028
ESCELSA	0,186	0,141	-1,960	0,294
BANDEIRANTE	0,274	0,127	-2,060	0,872
ENERGISA N FRIB	0,022	0,094	-2,360	-0,801
CELG	0,078	0,091	-2,400	-0,148
CFLO	0,062	0,089	-2,423	-0,339
AES-SUL	0,014	0,084	-2,473	-1,357
COPEL	0,033	0,081	-2,509	-0,777
PIRATININGA	0,050	0,081	-2,509	-0,456
CEB	0,091	0,073	-2,615	0,227
CEMIG	0,077	0,070	-2,654	0,087
ELEKTRO	0,038	0,063	-2,767	-0,472
RGE	0,050	0,058	-2,852	-0,132
CHESP	0,035	0,057	-2,859	-0,465
SANTA MARIA	0,021	0,055	-2,905	0,828
CSPE	0,050	0,052	-2,947	-0,055
CELESC	0,059	0,046	-3,080	0,254
JAGUARI	0,048	0,040	-3,227	0,187
CPFL PAULISTA	0,022	0,039	-3,246	0,189
ENERGISA MINAS	0,022	0,037	-3,284	-0,162
MOCOCA	0,003	0,035	-3,351	-1,700
SANTA CRUZ	0,022	0,034	-3,371	0,469
DME-PC	0,028	0,034	-3,384	-0,201
CPEE	0,067	0,030	-3,502	0,915

CAIU	0,028	0,028	-3,588	0,007
NACIONAL	0,009	0,025	-3,672	-0,874
VALE	0,022	0,025	-3,703	-0,563

Os valores foram ordenados de forma decrescente. É interessante notar que a perda não técnica estimada para a CELPA ficou acima de 100%, esse valor que *à priori* seria estranho é perfeitamente aceitável como resultado por haver no escopo dos dados observados uma distribuidora, ADESA, que teve perdas não técnicas consideradas pela ANEEL em 130%. Esse montante, que em princípio não parece lógico, vem de uma medida corretiva da ANEEL que incluiu perdas de média tensão na definição de perdas não técnicas em algumas distribuidoras.

Uma vez obtidos e apurados os principais resultados derivados da estimação do modelo, a primeira etapa para construção da meta regulatória, segundo a metodologia proposta pela ANEEL, esta completa, entretanto deve-se ter em mente que para os resultados finais de meta regulatória, ou seja, aqueles que as distribuidoras devem atingir ao final de cada ano, não são as estimativas resultantes desse modelo. Essas estimativas são a "base da pirâmide", ou seja, a partir desses resultados é que a ANEEL dá continuidade ao cálculo da meta regulatória.

O desenvolvimento dos passos subseqüentes, que completam a definição da meta regulatória, foge ao escopo desse trabalho, porém é interessante comentar que a metodologia vigente propõe que após calcular percentual de perda não técnica estimado para cada distribuidora, seja construída uma matriz de probabilidades com os resultados do modelo servindo de inputs. Essa matriz tem por objetivo a criação de um intervalo de confiança aonde empresas cujos intervalos possuem pontos em comum seriam diretamente comparáveis. O último passo da metodologia da ANEEL é clusterizar as empresas segundo porte e em seguida definir diferentes velocidades de redução que cada empresa deve atingir, segundo o seu cluster[20].

Voltando ao modelo econométrico desenvolvido no estudo, pode-se dizer que uma vez já analisadas métricas de aderência e feito o diagnóstico de seus resíduos, é possível partir para suas conclusões e considerações finais.

[20] Para maiores detalhes, ver Nota Técnica 271/2008 da ANEEL.