



UNIVERSIDADE FEDERAL DO PARANÁ  
SETOR DE CIÊNCIAS EXATAS  
DEPARTAMENTO DE ESTATÍSTICA  
CURSO DE ESTATÍSTICA

**João Matheus S. K. T. Hneda**

**Lineu Alberto Cavazani de Freitas**

## **Número de Consultas ao Médico**

Análise apresentada à disciplina Modelos Lineares Generalizados do Curso de Graduação em Estatística da Universidade Federal do Paraná.

Professor: Cesar Augusto Taconeli

**CURITIBA  
2016**

## Sumário

RESUMO .....	3
1 INTRODUÇÃO .....	4
2 MATERIAL E MÉTODOS .....	4
2.1 Material .....	4
2.1.1 Conjunto de Dados .....	4
2.1.2 Recursos Computacionais .....	5
2.2 Métodos .....	6
3 RESULTADOS E DISCUSSÃO .....	6
3.1 Análise Descritiva .....	6
3.2 Ajuste dos Modelos .....	7
3.2.1 Modelo Poisson com Função de Ligação logaritma .....	7
3.2.2 Modelo Binomial Negativo com Função de Ligação logaritma .....	8
3.2.3 Modelo Binomial Negativo com Função de Ligação logaritma apenas com variáveis significativas. ....	9
4 CONSIDERAÇÕES FINAIS .....	10

## **RESUMO**

Os dados analisados vêm do Inquérito de Saúde Australiano dos anos de 1977 e 1978 e consiste de 5190 adultos, o objetivo da análise foi modelar o número de visitas ao médico dos indivíduos em duas semanas. Devido ao curto período de coleta a base continha excesso de zeros. Para a modelagem ajustou-se o modelo log-linear de Poisson e o modelo com distribuição Binomial Negativa. Dentre os ajustes, o modelo Binomial Negativo apresentou melhor desempenho devido ao parâmetro de dispersão não ser igual a 1 (exigência do modelo Poisson); por causa do excesso de zeros nenhum dos modelos apresentou excepcional desempenho, porém o Binomial Negativo apresentou menor AIC e a maioria dos pontos do gráfico Normal de probabilidades dentro do envelope simulado, diferente do modelo Poisson.

# 1 INTRODUÇÃO

O presente trabalho tem como objetivo apresentar uma análise estatística, por meio de um modelo linear generalizado para dados de contagem, em dados referentes ao número de visitas ao médico em duas semanas de indivíduos australianos.

O trabalho contém uma breve análise descritiva (para melhor entender a base de dados), ajuste de um modelo buscando explicar o número de visitas ao médico de indivíduos em função das covariáveis disponíveis, diagnóstico para verificação se o modelo proposto nas circunstâncias se ajusta bem aos dados disponíveis, comparativo entre as distribuições propostas e quais os eventuais problemas dos dados e do método utilizado para a análise.

Entre as covariáveis disponíveis para explicar o número de visitas há, por exemplo, sexo, idade, número de doenças nas 2 últimas semanas, entre outras.

## 2 MATERIAL E MÉTODOS

### 2.1 Material

#### 2.1.1 Conjunto de Dados

Os dados utilizados para aplicação do modelo linear generalizado provêm do Inquérito de Saúde Australiano nos anos de 1977 e 1978 e consiste de 5190 adultos solteiros.

A base de dados contém uma série de covariáveis; as quais tiveram sua significância testada no que diz respeito a sua influência no número de visitas ao médico dos indivíduos, são elas:

**sex:** 1: Feminino; 0: Masculino;

**age:** Idade (em anos) dividida por 100 (medida no ponto médio de 10 grupos de idades: 15-19 anos a 65-69 anos, com 70 anos ou mais tratado como 72;

**agesq:** Idade ao quadrado;

**income:** Renda anual em dólares australianos dividido por 1000 (medida como ponto médio das classes delimitadas por: menos de 200, 200-1000, 1001-2000,..., 7001-8000, 10001-12000, 12001-14000, com 14001- tratado como 15000;

**levyplus:** 1: se tem plano de saúde particular; 0: caso contrário;

**freepoor:** 1: Se tem cobertura do governo por baixa renda; 0: caso contrário.

**freerepa:** 1: Se tem cobertura do governo devido a idade ou invalidez; 0: caso contrário;

**illness:** Número de doenças nas últimas duas semanas;

**actdays:** número de dias de atividade reduzida nas últimas duas semanas em virtude de doenças;

**hscore:** Escore de saúde usando o método de Goldberg. Quanto maior o escore, pior o estado de saúde;

**chcond1:** 1: Se o indivíduo tem condição crônica, mas sem limitação em suas atividades; 0: caso contrário;

**chcond2:** 1: Se o indivíduo tem condição crônica, com limitação em suas atividades; 0: caso contrário.

Totalizam-se 12 covariáveis com as quais tentou-se explicar o número de idas ao médico dos indivíduos estudados.

A variável resposta é:

**doctorco:** Número de consultas com um médico ou especialista nas últimas 2 semanas.

Os dados estão dispostos da seguinte forma:

**Tabela 1: Seis Primeiras Linhas da Base de Dados Analisada**

sex	Age	agesq	income	levyplus	Freepoor	freerepa	illness	actdays	hscore	chcond1	chcond2	doctorco
1	0,19	0,0361	0,55	1	0	0	1	4	1	0	0	1
1	0,19	0,0361	0,45	1	0	0	1	2	1	0	0	1
0	0,19	0,0361	0,90	0	0	0	3	0	0	0	0	1
0	0,19	0,0361	0,15	0	0	0	1	0	0	0	0	1
0	0,19	0,0361	0,45	0	0	0	2	5	1	1	0	1
1	0,19	0,0361	0,35	0	0	0	5	1	9	1	0	1

### 2.1.2 Recursos Computacionais

O *software* R, versão 3.3.1 (2016-06-21) foi utilizado para ajustar os modelos lineares generalizados aos dados descritos. Os pacotes utilizados com este propósito

foram: o pacote faraway (onde estão disponíveis os dados), car, statmod, hnp, entre outros.

## 2.2 Métodos

A proposta para modelar o número de consultas ao médico nas últimas 2 semanas foi o modelo linear generalizado aditivo com distribuições Poisson e Binomial negativa, tais modelos são amplamente utilizadas quando a variável de resposta é uma contagem.

## 3 RESULTADOS E DISCUSSÃO

### 3.1 Análise Descritiva

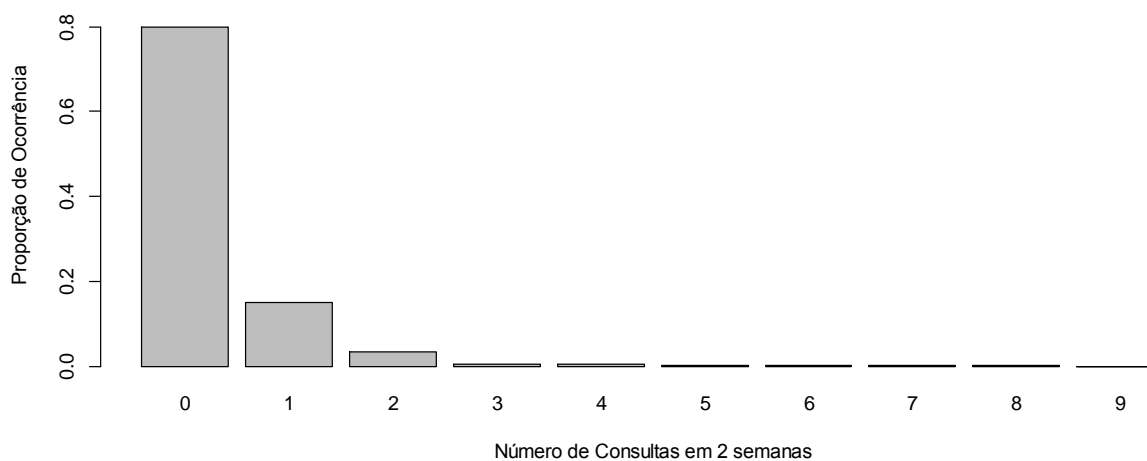
A análise descritiva dos dados, realizada antes da análise revelou uma importante informação: o excesso de zeros na variável resposta, facilmente observado na Tabela 2:

**Tabela 2: Proporção de Ocorrência Das Contagens Observadas**

Consultas em 2 semanas	0	1	2	3	4	5	6	7	8	9
Proporção	0,79	0,15	0,033	0,005	0,004	0,001	0,002	0,002	0,0009	0,0001

Logo, aproximadamente 80% dos indivíduos não foram ao médico em duas semanas, o que é ainda mais evidente no Gráfico 1:

**Gráfico 1: Proporção de Indivíduos em Função do Número de Consultas Médicas em 2 Semanas**



### 3.2 Ajuste dos Modelos

Foram testados dois modelos lineares generalizados aditivos para modelar o número de visitas ao médico em duas semanas: o modelo Poisson e o modelo Binomial negativo.

#### 3.2.1 Modelo Linear Generalizado com Distribuição Poisson

Quando se trata de modelos cuja resposta é uma contagem costumeiramente se utiliza o modelo Poisson; não raro tal modelo não costuma apresentar o melhor ajuste, porém dá indicativos de qual caminho seguir.

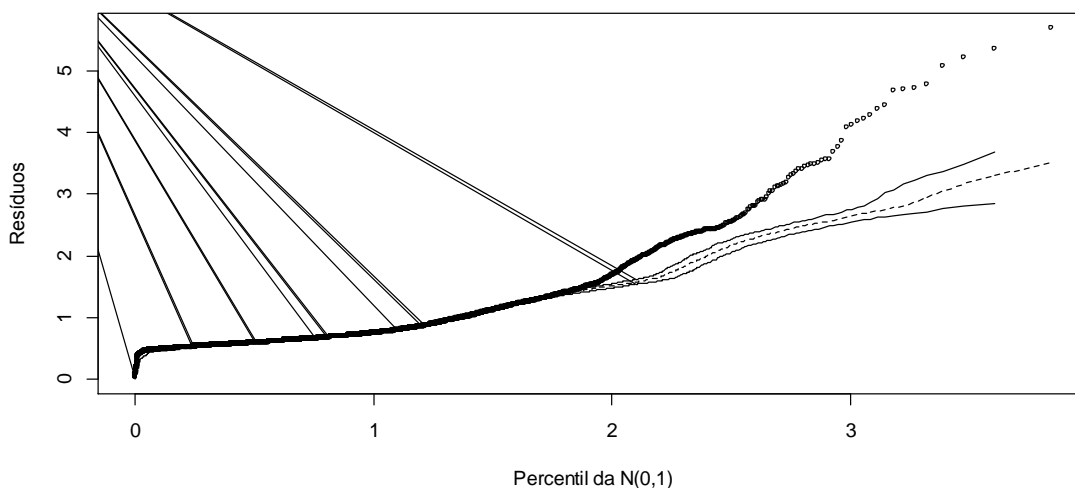
Para os dados analisados, a aplicação da função glm com distribuição Poisson retornou as seguintes informações:

**Tabela 3: Sumário do Ajuste do Modelo Linear Generalizado Poisson com link logaritmo**

	Estimativa	Erro Padrão	Valor Z	P - Valor
Intercepto	-2.223848	0.189816	-11.716	<0,001
Sex	0.156882	0.056137	2.795	0,005
Age	1.056299	1.000780	1.055	0,29
Agesq	-0.848704	1.077784	-0.787	0,43
Income	-0.205321	0.088379	-2.323	0,02
Levyplus	0.123185	0.071640	1.720	0,08
Freepoor	-0.440061	0.179811	-2.447	0,01
Freerepa	0.079798	0.092060	0.867	0,38
Illness	0.186948	0.018281	10.227	<0,001
actday	0.126846	0.005034	25.198	<0,001
Hscore	0.030081	0.010099	2.979	0,002
Chcond1	0.114085	0.066640	1.712	0,08
Chcond2	0.141158	0.083145	1.698	0,08

Este modelo apresentou Deviance nula igual a 5634,8 para 5189 graus de liberdade e Deviance residual igual a 4379,5 para 5177 graus de liberdade; uma redução na Deviance de 1255,3. Porém o modelo não se mostrou adequado, constatação baseada no envelope simulado mostrado na imagem 1:

**Imagem 1: Gráfico Normal de Probabilidades com Envelope Simulado para o Modelo Poisson**



### 3.2.2 Modelo Linear Generalizado com Distribuição Binomial Negativa

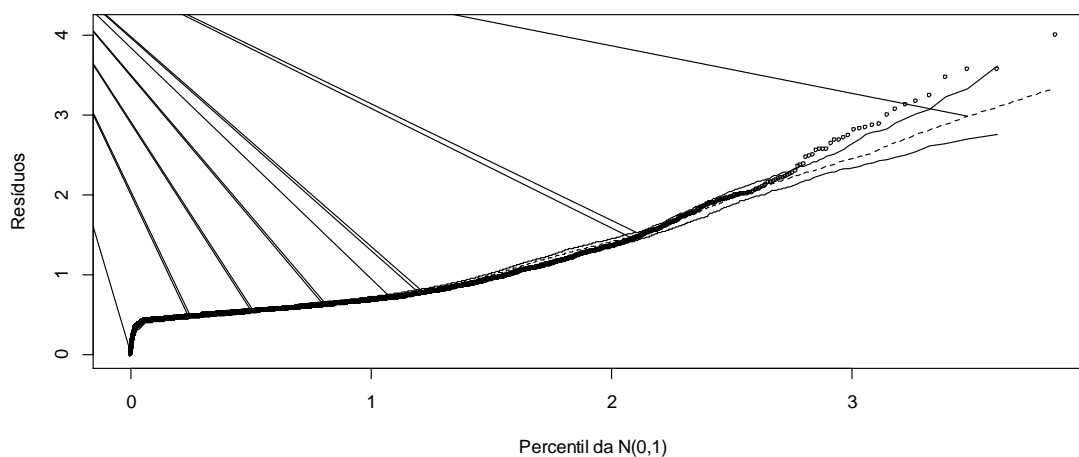
Como segunda tentativa, buscando um ajuste melhor que o proporcionado pelo ajuste Poisson, utilizou-se o modelo linear generalizado com distribuição Binomial Negativa, o qual retornou as seguintes estimativas:

**Tabela 4: Sumário do Ajuste do Modelo Linear Generalizado Binomial Negativo**

	Estimativa	Erro Padrão	Valor Z	P – Valor
Intercepto	-2.190007	0.233592	- 9.375	<0,001
Sex	0.216644	0.069697	3.108	0,001
Age	-0.216159	1.266701	- 0.171	0,86
Agesq	0.609158	1.383245	0.440	0,65
Income	-0.142202	0.108417	- 1.312	0,18
Levyplus	0.118064	0.085806	1.376	0,16
Freepoor	-0.496611	0.210803	- 2.356	0,01
Freerepa	0.144982	0.115970	1.250	0,21
Illness	0.214341	0.023579	9.090	<0,001
actday	0.143754	0.007311	19.662	<0,001
Hscore	0.038060	0.013654	2.788	0,005
Chcond1	0.099355	0.079303	1.253	0,21
Chcond2	0.190327	0.104357	1.824	0,068

Este modelo apresentou Deviance nula igual a 3928,7 para 5189 graus de liberdade e Deviance residual igual a 3028,3 para 5177 graus de liberdade; uma redução na Deviance de 900,4. Comparado ao modelo Poisson, o Binomial Negativo apresentou um melhor ajuste, com mais pontos dentro do envelope:

**Imagem 2: Gráfico Normal de Probabilidades com Envelope Simulado para o Modelo Binomial Negativo**





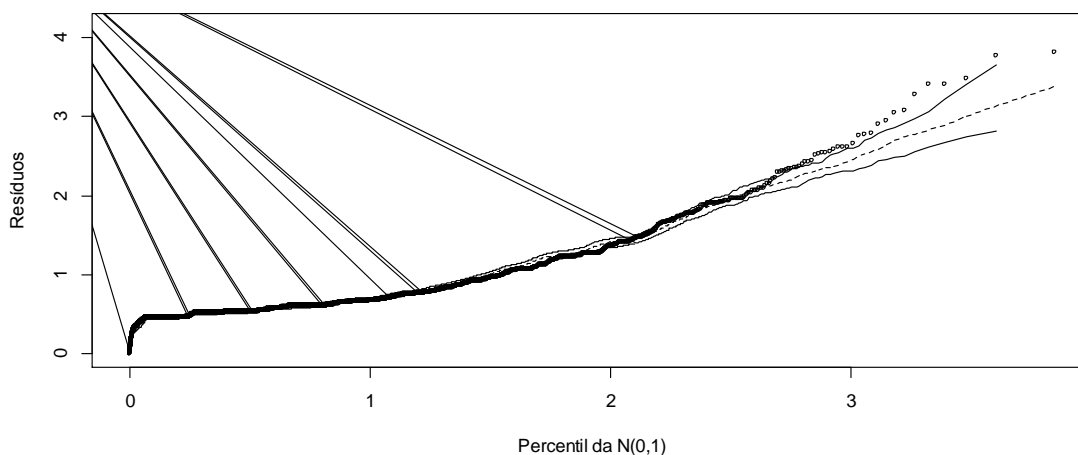
### 3.2.3 Modelo Linear Generalizado com Distribuição Binomial Negativa Apenas com Variáveis Significativas.

Considerando que o modelo Binomial Negativo apresentou o melhor diagnóstico reajustou-se o mesmo, porém sem as variáveis que não foram consideradas significativas na análise, o resultado das estimativas para este modelo, bem como o gráfico com envelope simulado estão descritos na Tabela 5 e na Imagem 3, respectivamente.

**Tabela 5: Sumário do Ajuste do Modelo Linear Generalizado Binomial Negativo Apenas com Variáveis Significativas**

	Estimativa	Erro Padrão	Valor Z	P - Valor
Intercepto	- 2.182410	0.063114	- 34.579	<0,005
Sex	0.321139	0.065049	4.937	<0,005
Freepoor	- 0.601198	0.200700	- 2.996	<0,005
Illness	0.247901	0.022035	11.250	<0,005
actday	0.146973	0.007118	20.648	<0,005
Hscore	0.039384	0.013378	2.944	<0,005

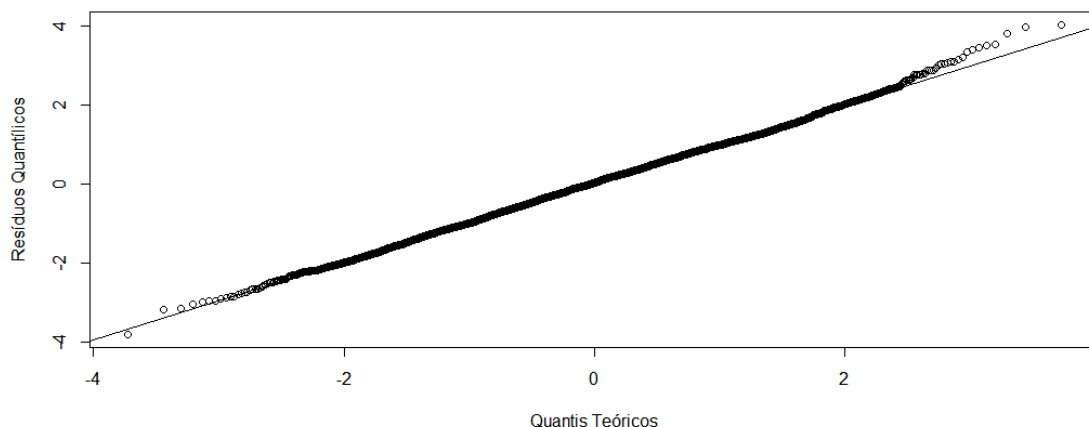
**Imagem 3: Gráfico Normal de Probabilidades com Envelope Simulado para o Modelo Binomial Negativo**



Este modelo apresentou Deviance Nula igual a 3927,6 para 5189 graus de liberdade e Deviance residual de 3051,5 para 5184 graus de liberdade; apresentando diferença de 876.

Após os ajustes serem feitos, como outra medida para verificação de qualidade de ajuste avaliou-se a normalidade dos resíduos quantílicos aleatorizados. A normalidade deste tipo de resíduo é um indicativo de bom ajuste do modelo proposto.

Imagem 4: Gráfico de Normalidade para Resíduos Quantílicos Aleatorizados



A avaliação desse gráfico deve ser feita pela comparação entre os resíduos quantílicos dos dados e os quantis teóricos da distribuição normal. Pela imagem, é possível dizer que os resíduos quantílicos aleatorizados seguem distribuição normal com leve fuga nas caldas, e portanto, não há evidências de que o modelo binomial negativo com as variáveis significativas está com falta de ajuste aos dados por este critério.

## 4 CONSIDERAÇÕES FINAIS

O modelo que apresentou melhor ajuste dentre os selecionados foi o modelo com distribuição Binomial Negativa com função de ligação logaritma. Os AICs observados para cada modelo foram 6737,083 para o modelo log-linear de Poisson; 6425,488 para o modelo com distribuição Binomial Negativa e 6435,542 para o modelo com distribuição Binomial Negativa com apenas as variáveis significativas, considerando que o melhor modelo é aquele com menor AIC, o modelo melhor ajustado é o Binomial Negativo sem as variáveis não significativas.

O principal problema observado na base de dados foi o excesso de respostas 0, fato este decorrente do curto período de coleta dos dados (2 semanas); outra situação constatada foi a super-dispersão dos dados, o parâmetro de dispersão observado foi igual a 1,33, fato esse que explica o motivo pelo qual o modelo log-linear de Poisson não apresentou bom ajuste, já que para funcionar o modelo Poisson exige parâmetro de dispersão igual a 1.

Apesar do melhor ajuste, o modelo Binomial Negativo não apresentou um ajuste “excelente”, uma alternativa para melhorar este ajuste seria buscar modelos lineares generalizados que considerem o excesso de zeros, pois tal característica influencia toda a análise.

Quanto ao problema, pode-se dizer que, a partir da aplicação do modelo linear generalizado com distribuição Binomial Negativa, há evidência de que sexo, cobertura do plano de saúde pelo governo, número de doenças nas duas últimas semanas, número de dias

de atividade reduzida nas últimas duas semanas em virtude de doenças e o escore de saúde influenciam o número de consultas com médicos ou especialistas na área de saúde nas últimas duas semanas .