Adriane Machado
Cinthia Zamin Cavassola
Luiza Hoffelder da Costa

REGRESSÃO EM DADOS DE CONTAGEM: UM ESTUDO SOBRE A QUANTIDADE DE BICICLETAS ALUGADAS

Relatório técnico apresentado como atividade avaliativa na disciplina de Modelos Lineares Generalizados da Graduação em Estatística na Universidade Federal do Paraná.

Professor Dr. Cesar Augusto Taconeli

Curitiba
2017

RESUMO

Os dados se referem a informações sobre a demanda de bicicletas alugadas em Washington, D.C. no programa Capital Bikeshare, durante os anos de 2011 até 2012. O objetivo da análise foi modelar a quantidade de bicicletas alugada. Para a modelagem foram utilizados métodos adequados para tal, como por exemplo a regressão Poisson; como foi detectada superdispersão nos dados, buscou-se como melhor alternativa um modelo linear generalizado a partir da distribuição binomial negativa.

PALAVRAS-CHAVE: BICICLETAS. ALUGUEL. SUPERDISPERSÃO. POISSON. BINOMIAL NEGATIVA.

SUMÁRIO

1. INTRODUÇÃO

O trabalho a seguir tem por objetivo analisar os fatores que podem influenciar a demanda de aluguel de bicicletas no programa Capital Bikeshare em Washington, DC.

Os sistemas de compartilhamento de bicicletas são um meio de alugar bicicletas onde o processo de obtenção de membros, aluguel e retorno da bicicleta é automatizado através de uma rede de quiosques em toda a cidade. Usando esses sistemas, as pessoas podem alugar uma bicicleta a partir de um local e devolvê-lo a um lugar diferente, conforme necessário. Atualmente, existem mais de 500 programas de compartilhamento de bicicletas em todo o mundo.

Através de dados gerados por esses sistemas é possível analisar padrões históricos de uso para prever a demanda de aluguel de bicicletas. Os dados consistem em informações como a duração da viagem, local de partida, localização da chegada e tempo decorrido, padrões históricos de uso e dados meteorológicos.

A variável resposta foi definida como o número total de bicicletas alugadas por hora e o objetivo deste trabalho é, utilizando apenas informações disponíveis antes do período de locação, prever a contagem total de bicicletas alugada durante cada hora em outros períodos.

2. MATERIAL E MÉTODOS

2.1 MATERIAL

2.1.1 CONJUNTO DE DADOS

Utilizou-se uma base de dados obtida da URL:

< https://www.kaggle.com/c/bike-sharing-demand >

Esta base fornece dados de aluguel por hora que abrangem dois anos e é formada por dois conjuntos separadamente, o conjunto de treinamento é composto dos primeiros 19 dias de cada mês, enquanto o conjunto de testes é o dia 20 até o final do mês.

Conforme definido no objetivo, iremos prever a contagem total de bicicletas alugadas durante cada hora coberta pelo conjunto de teste, utilizando apenas informações disponíveis antes do período de locação.

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 bicicletas alugadas,

são elas:

Datetime: data horária + timestamp;

Season: 1 - primavera, 2 - verão, 3 - outono, 4 - inverno;

Holiday: Se o dia é considerado um feriado;

Workingday: Se o dia não é nem um fim de semana nem feriado;

Weather:

1- Céu limpo, Algumas nuvens ou Parcialmente nublado;

2- Névoa, Névoa + poucas nuvens;

3- Neve leve, Chuva leve + Trovoada + Nuvens dispersas, Chuva leve + Nuvens dispersas

4- Chuva forte + paletes de gelo + tempestade + névoa, Neve + Nevoeiro

Temp: temperatura em Celsius;

Atemp: Sensação termina em Celsius;

Humidity: Humidade relativa

Windspeed: Velocidade do vento;

Casual: Número de alugueis por usuários não registrados;

Registered: Número de alugueis por usuários registrados;

Count : Número total de alugueis.

A tabela 1 abaixo apresenta as dez primeiras observações do conjunto de dados.

Tabela 1: Conjunto de Dados

datetime	season	holiday									
01/01/2011 00:00	1	0									
01/01/2011 01:00	1	0									
01/01/2011 02:00	1	0									
01/01/2011 03:00	1	0									

2.1.2 RECURSOS COMPUTACIONAIS

A base para os ajustes da modelagem dos dados foi o software R, onde foram utilizados pacotes como car, hnp, MASS, entre outros.

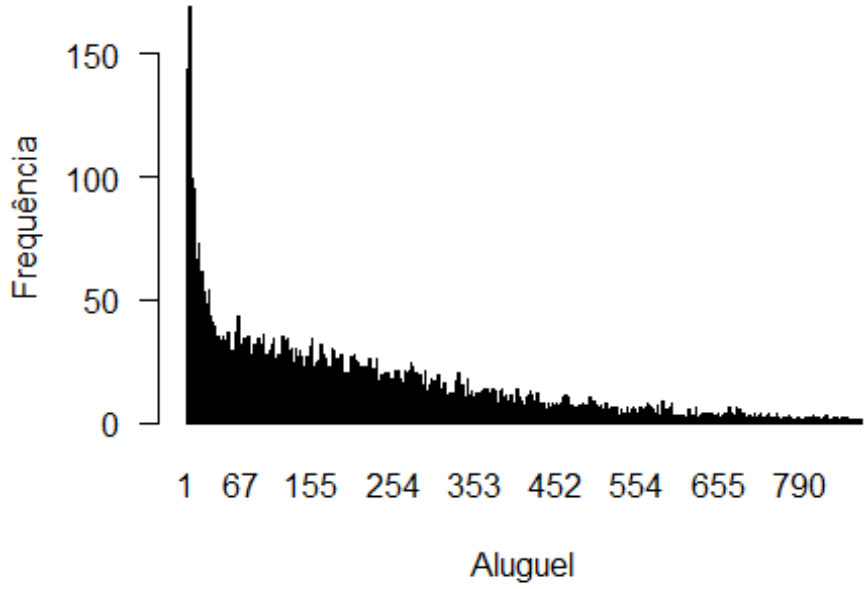
2.2 MÉTODOS

Devido a natureza de a variável resposta ser uma contagem, utilizou-se os modelos de distribuição Poisson e Binomial Negativa na tentativa de melhor ajustar os dados.

3. RESULTADOS E DISCUSSÃO

3.1 ANÁLISE DESCRITIVA

O gráfico 1 a seguir, mostra o comportamento da variável resposta, que é o numero de bicicletas alugadas. Através do gráfico podemos observar uma assimetria dos dados.



3.2 AJUSTE DOS MODELOS

Foram testados os modelos de Poisson e Binomial Negativo, na tentativa de uma melhor representação dos dados.

3.2.1 MODELO LINEAR GENERALIZADO COM DISTRIBUIÇÃO POISSON

O modelo de distribuição Poisson, é bastante usado quando se trata de observações de contagem, porém, nem sempre é o mais representativo.

A aplicação de modelo linear generalizado com distribuição Poisson, retornou uma

Deviance nula de 1.800.567 para 10.885 graus de liberdade e a *Deviance* Residual de 358.898 para 10.871 graus de liberdade. O AIC para esse ajuste foi de 428.575. A partir dos erros padrão e p-valores das covariáveis apresentados no modelo Poisson, foram detectados sinais de superdispersão; assim sendo, foi necessário fazer uso de outros modelos que acomodem um parâmetro de dispersão diferente de 1 (caso da regressão Poisson).

3.2.2 MODELO LINEAR GENERALIZADO COM DISTRIBUIÇÃO BINOMIAL NEGATIVA

Na busca de um melhor ajuste, utilizou-se a distribuição Binomial negativa para ajuste dos dados. A estimação do parâmetro de dispersão a partir do modelo apontou um $\Phi = 2{,}49$, o que inclusive justifica a ineficácia do ajuste do modelo Poisson, o qual por sua vez não possui parâmetros para acomodar uma maior dispersão nos dados.

A aplicação de modelo linear generalizado com distribuição Binomial Negativa, retornou uma *Deviance* nula de 36.084 para 10.885 graus de liberdade e a *Deviance* Residual de 12.022 para 10.871 graus de liberdade. O AIC para esse ajuste foi de 122.889.

3.2.3 MODELO FINAL

Considerando que o modelo Binomial Negativo apresentou o menor AIC, foi ajustado apenas com as variáveis que apresentaram significâncias. A Tabela 1 apresenta o sumário com as variáveis significativas para o modelo.

	Estimativa	Erro Padrão	Valor Z	P-valor
(Intercept)	3.885e+00	3.167e-02	122.662	< 2e-16 ***
season2	6.770e-02	1.850e-02	3.659	0.000253 ***
season3	1.141e-01	1.917e-02	5.952	2.65e-09 ***
season4	1.261e-01	1.854e-02	6.805	1.01e-11 ***
weather2	9.114e-02	1.510e-02	6.037	1.57e-09 ***
weather3	6.490e-02	2.541e-02	2.554	0.010655 *
weather4	6.949e-01	6.419e-01	1.082	0.279031
humidity	-5.411e-03	4.134e-04	-13.088	< 2e-16 ***
windspeed	2.672e-03	8.158e-04	3.275	0.001056 **
casual	5.443e-03	1.533e-04	35.515	< 2e-16 ***
registered	5.928e-03	4.792e-05	123.704	< 2e-16 ***

O modelo final possui uma *Deviance* nula de 35.644 para 10.885 graus de liberdade e a *Deviance* Residual de 12.020 para 10.875 graus de liberdade. O AIC para esse ajuste foi de 123.020.

Ressalta-se, ainda, que a fim de se explorar as melhores possibilidades de ajuste, foram utilizados ainda modelos de quase verossimilhança, ajuste por reamostragem *bootstrap*, e a partir do modelo Binomial Negativo, foram ainda explorados modelos com interação, termos quadráticos e o agrupamento da covariável “weather” em dois grupos (tempo 4 e o agrupamento dos outros três grupos de clima menos desafiador); tendo em

vista que não houve variação significativa nas medidas de qualidade de ajuste, permaneceu-se com o modelo Binomial Negativo com os coeficientes descritos acima.

4. CONCLUSÃO

Em face da superdispersão e assimetria encontradas nos dados trabalhados quando da tentativa de ajuste pelo modelo Poisson, lançou-se mão de diversas técnicas a fim de se encontrar um ajuste satisfatório (inserção de interações, termos quadráticos, agrupamento de covariáveis, modelos com outras distribuições, modelos de quase verossimilhança e ajustes *bootstrap*); verificou-se que no caso objeto deste estudo, realmente a estimação do parâmetro de dispersão proporcionada pela utilização do modelo Binomial Negativo mostrou-se bastante adequada para acomodar a variabilidade dos dados, já que tal modelo permite um Φ maior que 1 e consegue acomodar melhor a dispersão do conjunto de dados.