Variations of Poisson GLM Modeling Using Airbnb Data of Melbourne, Australia
Group: Three Sigma

Xixi Chen; Lixian Chen; George Liu; Yuhao Wang
Columbia University

Abstract

We analyze a dataset drawn from Airbnb listings in Melbourne, Australia. In order to understand the drivers of conversion and engagement as measured by review count and predict the number of reviews that a listing is able to generate, we attempt to fit a loglinear model (Poisson GLM) as well as variations in order to deal with overdispersion and zero inflation in the data given a set of predictors.

In particular, because we are not privy to data about bookings given our data is scraped from the AirBnB site, we leverage factors that are likely to be predictors of historic conversion such as price, Instant Book, and availability. We also consider other variables such as months listed and property type that may have effects on reviews independent of conversion.

Introduction

The variable we are trying to analyze, the number of reviews left on a listing, naturally takes the form of count data. Thus it is most natural to treat the random component as a Poisson-distributed random variable.

Poisson distributed variables have mean and variance equal, but we find in our data that the mean number of reviews is 21.3 while the standard deviation is 39.0. With the clear presence of overdispersion we turn to the quasi-Poisson and negative binomial models in order to better fit the data at hand.

Further, the predicted number of 0s for a Poisson distribution with $\lambda = 21.3$ and $n = 22895$ is:

$$22995e^{-21.3} \approx 0.00001$$

In the data we observe approximately 5211 listings with 0 reviews, meaning that we have a significant zero-inflation problem. This can be addressed with the zero-inflated Poisson model.

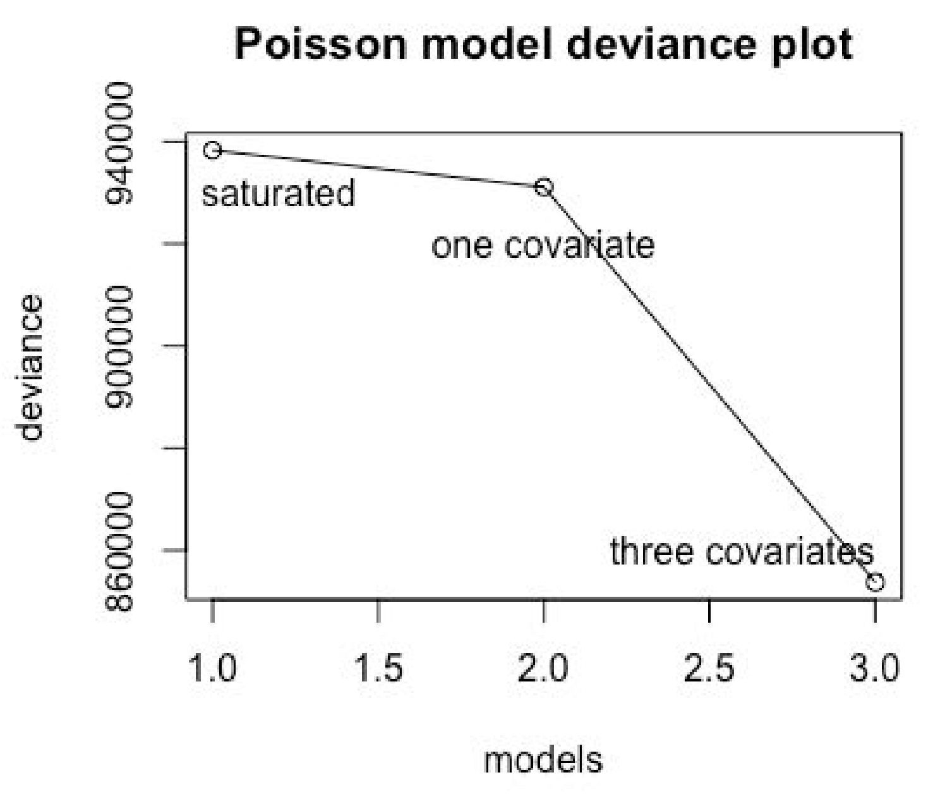

Figure 1. Deviance Comparison of Poisson Models

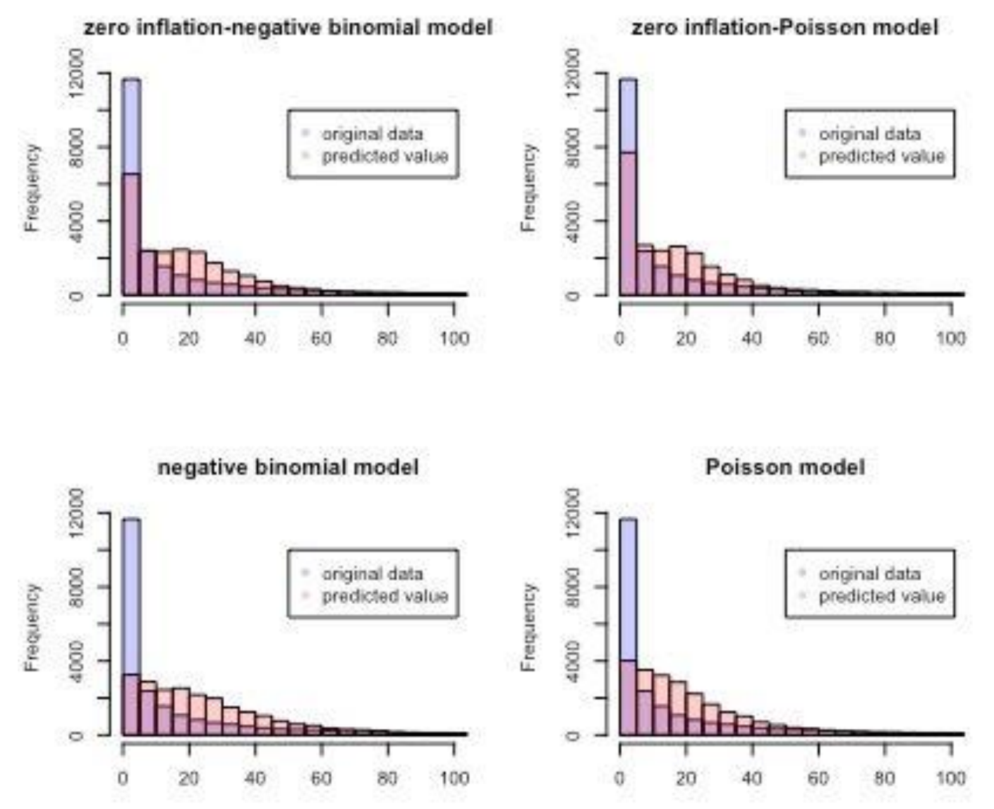

Figure 2. Histogram of Predicted values vs Original Data

Methods and Materials

We primarily leverage variations of the Poisson GLM (log-linear model). In particular, the loglinear Poisson GLM

$$\log \mu_i = \alpha + \beta x_i \ + \ \log t_i$$

The expansions of it are the quasi-Poisson model, where we introduce a parameter ϕ to the likelihood equation so we solve

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi \mu_i} \frac{\partial \mu_i}{\partial \eta_i} = 0, j \ = 1, \ldots, p$$

The negative binomial model arises when we assume that the mean is not fixed but is random with a $\mathrm{Gamma}(\mathrm{r}, \frac{1-p}{p})$ distribution. Thus $\mathrm{Y} \sim \mathrm{NegBin(r,p)}$

The zero-inflated Poisson model introduces a mixed model where with some probability π_i Y is 0 and with probability $1 - \pi_i$ is distributed as Poisson. We also analyze this with an underlying negative binomial distribution as described above.

Results

We tested the Poisson, negative binomial Poisson, quasi-Poisson, zero-inflated Poisson, zero-inflated negative binomial models. We also compared Poisson models with 0 predictors, 1 predictor, and 3 predictors.

For the quasi-Poisson, due to irregular outputs with an extremely high predicted ϕ value, we decided to not utilize this model.

We then compared the models using the Vuong test since we are comparing non-nested models

The results showed clear advantages to both the negative binomial and zero-inflated models, with the zero-inflated negative binomial being the winner obviously. All Vuong test results are summarized in Table 1 below.

Table 1. Vuong test results comparing different models

Models Compared (winner in italics)	Vuong z-statistic
Poisson and *Negative Binomial*	79.34
Poisson and *Zero-inflated Poisson*	79.74
Zero-inflated Poisson and *Negative Binomial*	62.27
Zero-inflated Negative Binomial and Negative Binomial	36.45
Zero-inflated Poisson and *Zero-inflated Negative Binomial*	63.45

Model	Predicted # of zeros	Variance
Poisson	760	19.71
Zero-inflated Poisson	5265	15.73
Negative Binomial	559	17.96
Zero-inflated Negative Binomial	4219	23.43

Table 2. Residual plots for final models

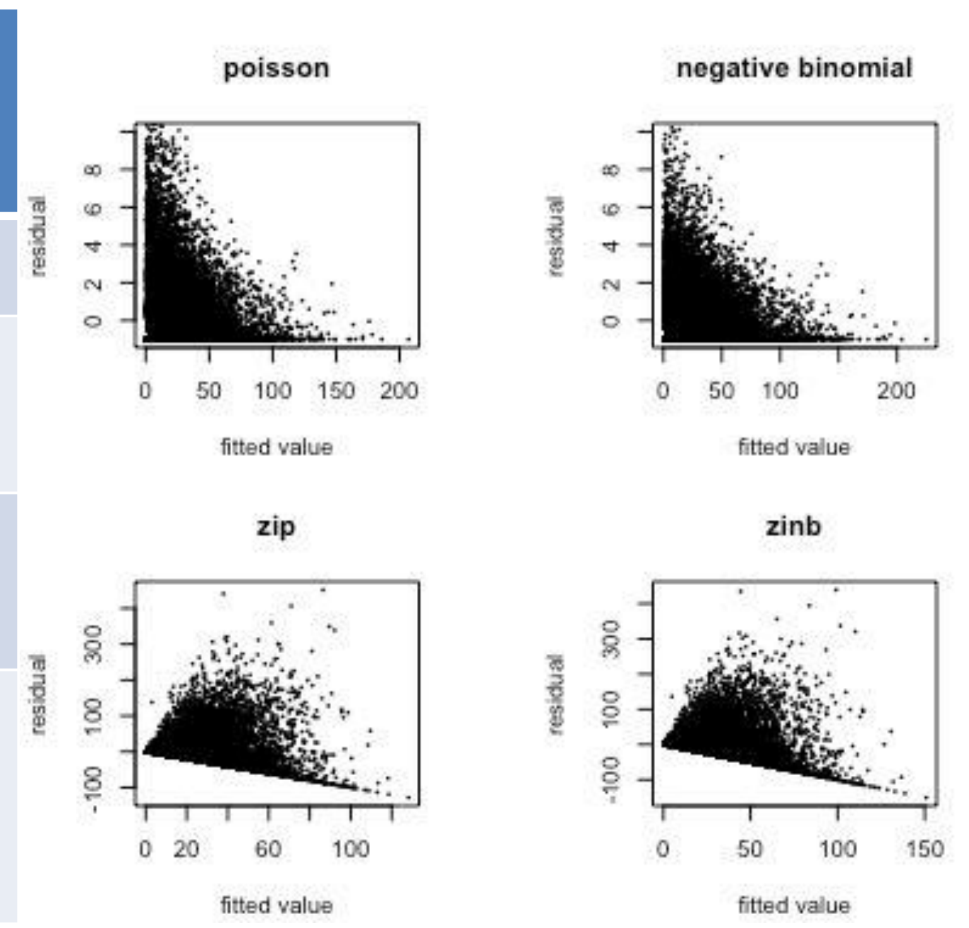

Figure 3. Residual plots for final models

Model Output

```
Call:
zeroinfl(formula = number_of_reviews ~ price + availability_365 + instant_bookable + offset(log(months)),
    data = listings_joined, dist = "negbin")
Pearson residuals:
     Min       1Q   Median       3Q      Max 
 -0.8221  -0.6355  -0.4021   0.3047 440.1621 

Count model coefficients (negbin with log link):
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       3.993e-02  1.846e-02   2.163   0.0306 *  
price            -5.161e-04  7.039e-05  -7.333 2.25e-13 ***
availability_365  8.291e-04  7.478e-05  11.087  < 2e-16 ***
instant_bookablet  5.285e-01  1.765e-02  29.943  < 2e-16 ***
Log(theta)       -2.863e-01  1.226e-02 -23.358  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                   Estimate Std. Error  z value Pr(>|z|)    
(Intercept)      -4.5297124  0.0372372 -121.645  < 2e-16 ***
price             0.0008192  0.0001209    6.776 1.23e-11 ***
availability_365 -0.0026005  0.0001724  -15.086  < 2e-16 ***
instant_bookablet  0.1995159  0.0426965    4.673 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 0.751 
Number of iterations in BFGS optimization: 15 
Log-likelihood: -8.057e+04 on 9 Df
```

Conclusions

We were able to produce a model which we felt had reasonable predictive power, in which we see a negative relationship between price and reviews, but a positive one in availability and Instant Book status. None of the coefficients provides a shocking conclusion, so this model is one that we would feel comfortable with.

However, as we can see from the histogram there are still areas which can be refined. For instance, we can see that our model provides higher weights in the mid-range of review count. We also observe that there are certain residuals that are quite large.

The natural next steps are to understand the points with large residuals to see whether they have excessive leverage, and to refine our understanding of the number of low-review count and whether we could introduce additional variables to improve the fit.

Contact

Lixian Chen
Department of Statistics, Columbia University
Email: lc3359@Columbia.edu

References

1. Tyler Xie. (2019). *Melbourne AirBnB Open Data*. (Version 10) [Data files]. Retrieved from https://www.kaggle.com/tylerx/melbourne-airbnb-open-data/
2. Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and non-nested Hypotheses". Econometrica. 57 (2): 307–333. JSTOR 1912557.
3. Agresti, Alan. (2012) *Categorical Data Analysis (3rd ed.)*. Hoboken, NJ: John Wiley and Sons