Poisson Regression

Offset & Zero-inflated Poisson models

Prof. Maria Tackett

01.31.22

Click for PDF of slides

Announcements

- Reading: BMLR - Chapter 4 Poisson regression
  - For Monday: BMLR - Chapter 5: Generalized Linear Models: A Unifying Theory
- Mini-Project 01:
  - Final write up and presentations **Wed, Feb 09 at 3:30pm**
- HW 02 **due Mon, Feb 07 at 11:59pm**

Learning goals

- Explore properties of negative binomial versus Poisson response
- Fit and interpret models with offset to adjust for differences in sampling effort
- Fit and interpret Zero-inflated Poisson models

Negative binomial regression model

Negative binomial regression model

Another approach to handle overdispersion is to use a **negative binomial regression model**

- This has more flexibility than the quasi-Poisson model, because there is a new parameter in addition to λ

Let Y be a **negative binomial random variable**, $Y \sim NegBinom(r, p)$, then

$$P(Y = y_i) = \binom{y_i + r - 1}{r - 1}(1-p)^{y_i}p^r \quad y_i = 0, 1, 2, \ldots, \infty$$

Negative binomial regression model

- **Main idea**: Generate a λ for each observation (household) and generate a count using the Poisson random variable with parameter λ
    - Makes the counts more dispersed than with a single parameter
- Think of it as a Poisson model such that λ is also random

$$\text{If } Y|\lambda \sim Poisson(\lambda)$$
$$\text{and } \lambda \sim Gamma\left(r, \frac{1-p}{p}\right)$$
$$\text{then } Y \sim NegBinom(r, p)$$

Negative binomial simulation exercise

Complete the Negative binomial regression exercise in **`lecture-07.Rmd`** (found in your lecture-07 GitHub repo).

08:00

Offset

Data: Airbnbs in NYC

The data set NYCairbnb-sample.csv contains information about a random sample of 1000 Airbnbs in New York City. It is a subset of the data on 40628 Airbnbs scraped by Awad et al. (2017).

Variables

- **`number_of_reviews`**: Number of reviews for the unit on Airbnb (proxy for number of rentals)
- **`price`**: price per night in US dollars
- **`room_type`**: Entire home/apartment, private room, or shared room
- **`days`**: Number of days the unit has been listed (date when info scraped - date when unit first listed on Airbnb)

Data set pulled from BMLR Section 4.11.3.

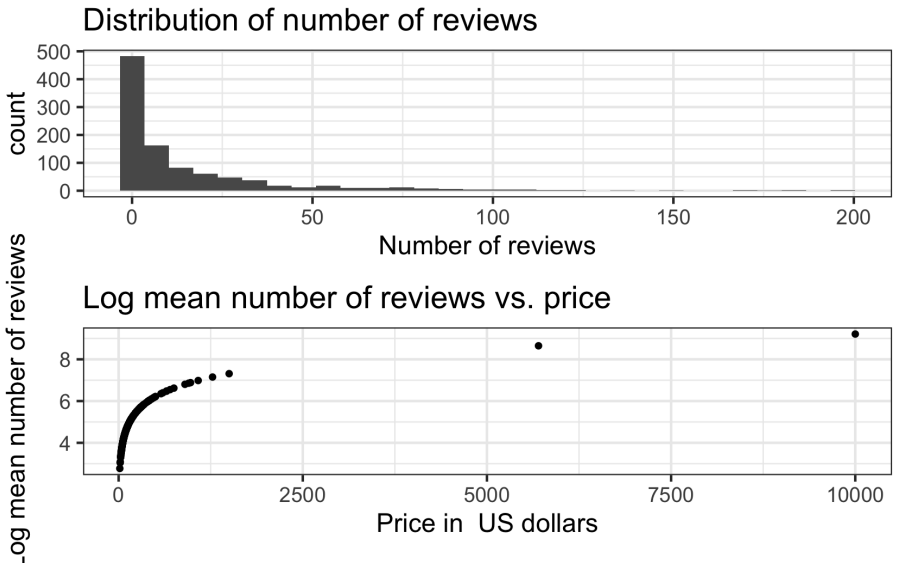
Data: Airbnbs in NYC

```
airbnb <- read_csv("data/NYCairbnb-sample.csv") %>%
  select(id, number_of_reviews, days, room_type, price)
```

id	number_of_reviews	days	room_type	price
15756544	16	1144	Private room	120
14218251	15	471	Private room	89
21644	0	2600	Private room	89
13667835	1	283	Entire home/apt	150
265912	0	1970	Entire home/apt	89

Goal: Use the price and room type of Airbnbs to describe variation in the number of reviews (a proxy for number of rentals).

EDA



Overall

mean	var
15.916	765.969

by Room type

room_type	mean	var
Entire home/apt	16.283	760.348
Private room	15.608	786.399
Shared room	15.028	605.971

Considerations for modeling

We would like to fit the Poisson regression model

$$\log(\lambda) = \beta_0 + \beta_1\, price + \beta_2\, room\_type1 + \beta_3\, room\_type2$$

- Based on the EDA, what are some potential issues we may want to address in the model building?
- Suppose any model fit issues are addressed. What are some potential limitations to the conclusions and interpretations from the model?

03:00

Offset

- Sometimes counts are not directly comparable because the observations differ based on some characteristic directly related to the counts, i.e. the *sampling effort*.
- An **offset** can be used to adjust for differences in sampling effort.
- Let x_{offset} be the variable that accounts for differences in sampling effort, then $\log(x_{offset})$ will be added to the model.

$$\log(\lambda_i) = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \ldots + \beta_p \, x_{ip} + \log(x_{offset_i})$$

- The offset is a term in the model with coefficient always equal to 1.

Adding an offset to the Airbnb model

We will add the offset $\log(days)$ to the model. This accounts for the fact that we would expect Airbnbs that have been listed longer to have more reviews.

$$\log(\lambda_i) = \beta_0 + \beta_1 \, price_i + \beta_2 \, room\_type1_i + \beta_3 \, room\_type2_i + \log(days_i)$$

Note: The response variable for the model is still $\log(\lambda_i)$, the log mean number of reviews

Detail on the offset

We want to adjust for the number of days, so we are interested in $\frac{reviews}{days}$.

Given λ is the mean number of reviews

$$\log\left(\frac{\lambda}{days}\right) = \beta_0 + \beta_1\, price + \beta_2\, room\_type1 + \beta_3\, room\_type2$$

$$\Rightarrow \log(\lambda) - \log(days) = \beta_0 + \beta_1\, price + \beta_2\, room\_type1 + \beta_3\, room\_type2$$

$$\Rightarrow \log(\lambda) = \beta_0 + \beta_1\, price + \beta_2\, room\_type1 + \beta_3\, room\_type2 + \log(days)$$

Airbnb model in R

```
airbnb_model <- glm(number_of_reviews ~ price + room_type,
                    data = airbnb, family = poisson,
                    offset = log(days))
```

term	estimate	std.error	statistic	p.value
(Intercept)	-4.1351	0.0170	-243.1397	0
price	-0.0005	0.0001	-7.0952	0
room_typePrivate room	-0.0994	0.0174	-5.6986	0
room_typeShared room	0.2436	0.0452	5.3841	0

The coefficient for $\log(days)$ is fixed at 1, so it is not in the model output.

Interpretations

term	estimate	std.error	statistic	p.value
(Intercept)	-4.1351	0.0170	-243.1397	0
price	-0.0005	0.0001	-7.0952	0
room_typePrivate room	-0.0994	0.0174	-5.6986	0
room_typeShared room	0.2436	0.0452	5.3841	0

- Interpret the coefficient of **`price`**.
- Interpret the coefficient of **`room_typePrivate room`**

03:00

Goodness-of-fit

$$H_0: \text{ The model is a good fit for the data}$$
$$H_a: \text{ There is significant lack of fit}$$

```
pchisq(airbnb_model$deviance, airbnb_model$df.residual, lower.tail = F)
```

```
## [1] 0
```

There is evidence of significant lack of fit in the model. Therefore, more models would need to be explored that address the issues mentioned earlier.

*In practice we would assess goodness-of-fit and finalize the model **before** any interpretations and conclusions.*

Zero-inflated Poisson model

Data: Weekend drinking

The data **`weekend-drinks.csv`** contains information from a survey of 77 students in a introductory statistics course on a dry campus.

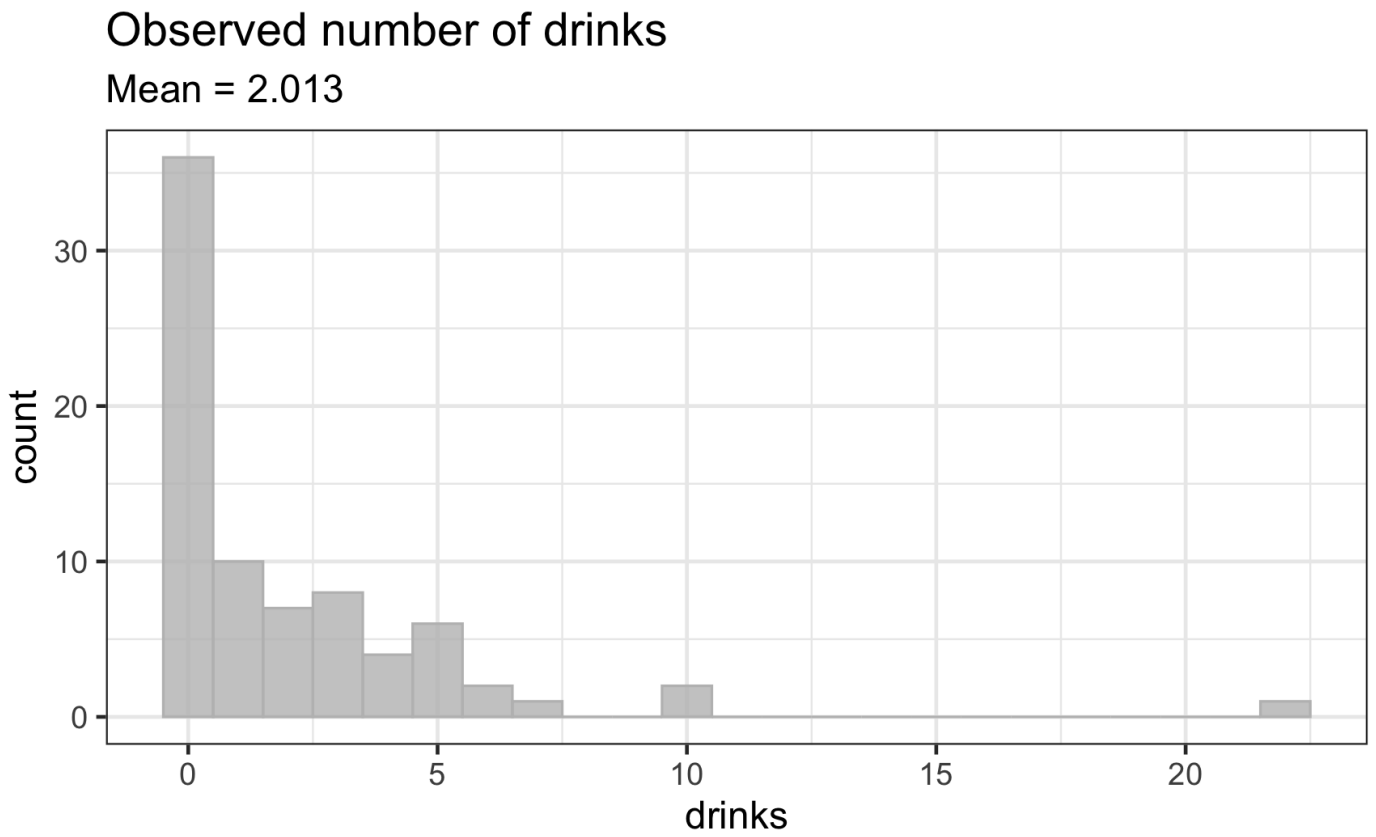
Variables

- **`drinks`**: Number of drinks they had in the past weekend
- **`off_campus`**: 1 - lives off campus, 0 otherwise
- **`first_year`**: 1 - student is a first-year, 0 otherwise
- **`sex`**: f - student identifies as female, m - student identifies as male

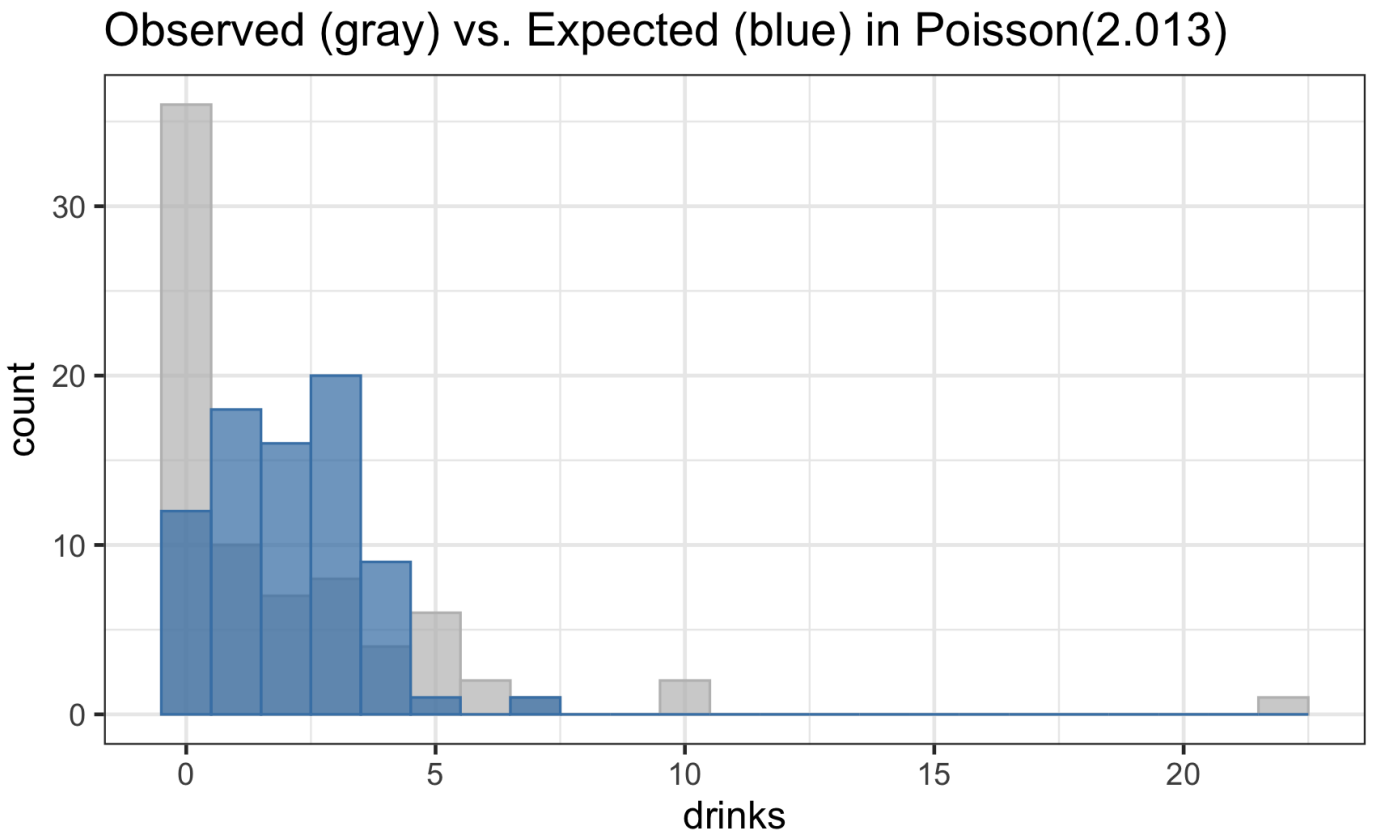
Goal: The goal is explore factors related to drinking behavior on a dry campus.

Case study in BMLR - Section 4.10

EDA: Response variable



Observed vs. expected response



What does it mean to be a "zero" in this data?

Two types of zeros

There are two types of zeros

- Those who happen to have a zero in the data set (people who drink but happened to not drink last weekend)
- Those who will always report a value of zero (non-drinkers)
    - These are called **true zeros**

We introduce a new parameter α for the proportion of true zeros, then fit a model that has two components:

1️⃣ The association between mean number of drinks and various characteristics among those who drink

2️⃣ The estimated proportion of non-drinkers

Zero-inflated Poisson model

Zero-inflated Poisson (ZIP) model has two parts

1️⃣ Association, among those who drink, between the mean number of drinks and predictors sex and off campus residence

$$\log(\lambda) = \beta_0 + \beta_1 \, off\_campus + \beta_2 \, sex$$

where λ is the mean number of drinks among those who drink

2️⃣ Probability that a student does not drink

$$\text{logit}(\alpha) = \log\left(\frac{\alpha}{1-\alpha}\right) = \beta_0 + \beta_1 \, first\_year$$

where α is the proportion of non-drinkers

Note: The same variables can be used in each component

Details of the ZIP model

- The ZIP model is a special case of a **latent variable model**
    - A type of **mixture model** where observations for one or more groups occur together but the group membership unknown
- Zero-inflated models are a common type of mixture model; they apply beyond Poisson regression

ZIP model in R

Fit ZIP models using the **`zeroinfl`** function from the **pscl** R package.

```
library(pscl)

drinks_zip <- zeroinfl(drinks ~ off_campus + sex | first_year,
                       data = drinks)
drinks_zip
```

```
##
## Call:
## zeroinfl(formula = drinks ~ off_campus + sex | first_year, data = drinks)
##
## Count model coefficients (poisson with log link):
## (Intercept)   off_campus         sexm
##      0.7543       0.4159       1.0209
##
## Zero-inflation model coefficients (binomial with logit link):
## (Intercept)   first_year
##     -0.6036       1.1364
```

Tidy output

Use the **`tidy`** function from the **poissonreg** package for tidy model output.

```
library(poissonreg)
```

Mean number of drinks among drinkers

```
tidy(drinks_zip, type = "count") %>% kable(digits = 3)
```

term	type	estimate	std.error	statistic	p.value
(Intercept)	count	0.754	0.144	5.238	0.000
off_campus	count	0.416	0.206	2.020	0.043
sexm	count	1.021	0.175	5.827	0.000

Tidy output

Proportion of non-drinkers

```
tidy(drinks_zip, type = "zero") %>% kable(digits = 3)
```

term	type	estimate	std.error	statistic	p.value
(Intercept)	zero	-0.604	0.311	-1.938	0.053
first_year	zero	1.136	0.610	1.864	0.062

Interpreting the model coefficients

term	type	estimate	std.error	statistic	p.value
(Intercept)	count	0.754	0.144	5.238	0.000
off_campus	count	0.416	0.206	2.020	0.043
sexm	count	1.021	0.175	5.827	0.000

- Interpret the intercept.
- Interpret the coefficients of **`off_campus`** and **`sexm`**.

Estimated proportion zeros

term	type	estimate	std.error	statistic	p.value
(Intercept)	zero	-0.604	0.311	-1.938	0.053
first_year	zero	1.136	0.610	1.864	0.062

- What is the estimated proportion of non-drinkers among first years?
- What is the estimated proportion of non-drinkers among sophomores, juniors, seniors?

These are just a few of the many models...

- Use the Vuong Test to compare the fit of the ZIP model to a regular Poisson model
    - Why can't we use a drop-in-deviance test?
- We've just discussed the ZIP model here, but there are other model applications beyond the standard Poisson regression model (e.g., hurdle models, Zero-inflated Negative Binomial models, etc.)

Acknowledgements

These slides are based on content in BMLR - Chapter 4 Poisson regression