Poisson Regression

Zero-inflated Poisson models

Prof. Maria Tackett

02.07.22

Click for PDF of slides

Announcements

- For Mon, Feb 14: BMLR - Chapter 5: Generalized Linear Models: A Unifying Theory
- Mini-Project 01:
  - Final write up and presentations **Wed, Feb 09 at 3:30pm**
  - GitHub repo organization due **Wed, Feb 09 at 11:59pm**
- HW 02 **due TODAY at 11:59pm**

HW 02 questions?

Presentation order

(all presentations and write ups due on Wed, Feb 09 at 3:30pm)

Learning goals

- Fit and interpret the Zero-inflated Poisson model
- Write likelihood for Poisson and 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.

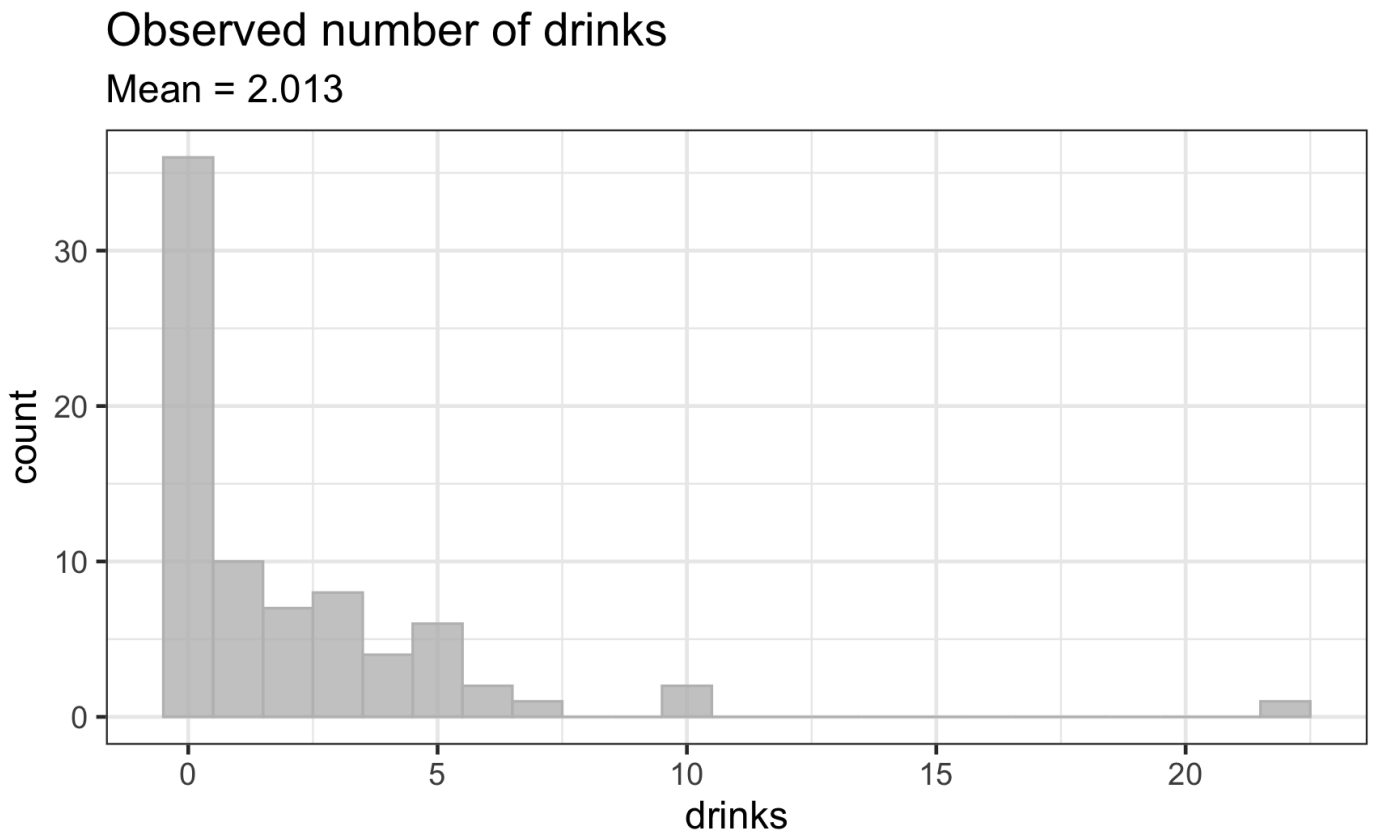
Variables

- **`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

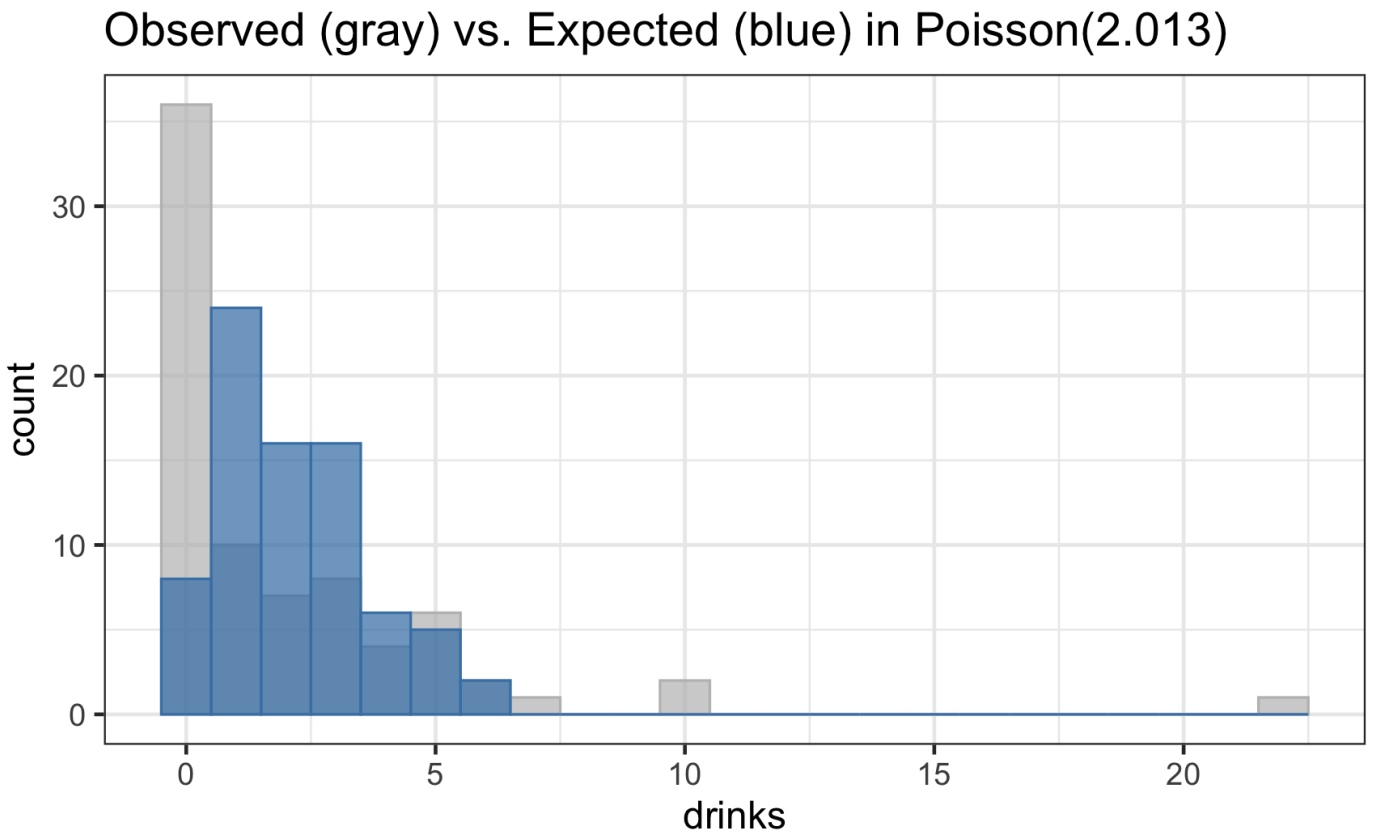
Goal: The goal is explore factors related to drinking behavior on a dry campus.

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 those who drink

```
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`**.

Click here to submit our response.

03:00

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

Based on the model...

- What is the probability a first-year student is a non-drinker?
- What is the probability a upperclass student (sophomore, junior, senior) is a non-drinker?

Click here to submit your response.

02:00

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.)

Likelihoods for Poisson models

Estimating coefficients in Poisson model

- Least squares estimation would not work because the normality and equal variance assumptions don't hold for Poisson regression
- **Maximum likelihood estimation** is used to estimate the coefficients of Poisson regression.
- The likelihood is the product of the probabilities for the n *independent* observations in the data

Likelihood for regular Poisson model

Let's go back to example about household size in the Philippines. We will focus on the model using the main effect of age to understand variability in mean household size.

Suppose the first five observations have household sizes of 4, 2, 8, 6, and 1. Then the likelihood is

$$L = \frac{e^{-\lambda_1}\lambda_1^4}{4!} * \frac{e^{-\lambda_2}\lambda_2^2}{2!} * \frac{e^{-\lambda_3}\lambda_3^8}{8!} * \frac{e^{-\lambda_4}\lambda_4^6}{6!} * \frac{e^{-\lambda_5}\lambda_5^1}{1!}$$

Likelihood for regular Poisson model

We will use the log likelihood to make finding the MLE easier

$$\log(L) = -\lambda_1 + 4\log(\lambda_1) - \lambda_2 + 2\log(\lambda_2) - \lambda_3 + 8\log(\lambda_3) \\ - \lambda_4 + 6\log(\lambda_4) - \lambda_5 + \log(\lambda_5) + C$$

where

- λ is the mean number in household depending on x_i
- $C = -[\log(4!) + \log(2!) + \log(8!) + \log(6!) + \log(1!)]$

Likelihood for regular Poisson model

Given the age of the head of the household, we fit the model

$$\log(\lambda_i) = \beta_0 + \beta_1 \, age_i$$

Then we replace each λ_i in $\log(L)$ with $e^{\beta_0 + \beta_1 \, age_i}$.

Suppose the first five observations have ages $X = (32, 21, 55, 44, 28)$. Then

$$\begin{aligned}\log(L) = &[-e^{\beta_0+\beta_1 32} + 4(\beta_0 + \beta_1 32)] + [-e^{\beta_0+\beta_1 21} + 2(\beta_0 + \beta_1 21)] \\ &+ [-e^{\beta_0+\beta_1 55} + 8(\beta_0 + \beta_1 55)] + [-e^{\beta_0+\beta_1 44} + 6(\beta_0 + \beta_1 44)] \\ &+ [-e^{\beta_0+\beta_1 28}(\beta_0 + \beta_1 28)] + C\end{aligned}$$

Use search algorithm to find the values of β_0 and β_1 that maximize the above equation.

Probabilities under ZIP model

There are three different types of observations in the data:

- Observed zero and will always be 0 (true zeros)
- Observed 0 but will not always be 0
- Observed non-zero count and will not always be 0

Probabilities under ZIP model

True zeros

$$P(0|\text{true zero}) = \alpha$$

Observed 0 but will not always be 0

$$P(0|\text{not always zero}) = (1-\alpha)\frac{e^{-\lambda}\lambda^0}{0!}$$

Did not observe 0 and will not always be 0

$$P(z_i|\text{not always zero}) = (1-\alpha)\frac{e^{-\lambda}\lambda^{z_i}}{z_i!}$$

Probabilities under ZIP model

Putting this all together. Let y_i be an observed response then

$$P(Y_i = y_i|x_i) = \begin{cases} \alpha + (1-\alpha)e^{-\lambda_i} & \text{if } y_i = 0 \\ (1-\alpha)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{if } y_i > 0 \end{cases}$$

Recall from our example,

$$\lambda_i = e^{\beta_0 + \beta_1\, off\_campus_i + \beta_2\, sex_i}$$

$$\alpha = \frac{e^{\beta_{0\alpha} + \beta_{1\alpha}\, first\_year}}{1 + e^{\beta_{0\alpha} + \beta_{1\alpha}\, first\_year}}$$

- Plug in λ_i and α into the above equation obtain the likelihood function

Acknowledgements

These slides are based on content in BMLR

- Section 4.4.5: Using likelihoods to fit models
- Section 4.4.10 Case Study: Weekend Drinking