Poisson Regression

Part III

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 4 of Roback and Legler text.

Computing set up

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(patchwork)
library(viridis)

knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      dpt = 300,
                      out.width = "70%",
                      fig.align = "center")

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Topics

- Offset in Poisson regression
- Zero-inflated Poisson regression

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, Lebo, and Linden (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)

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

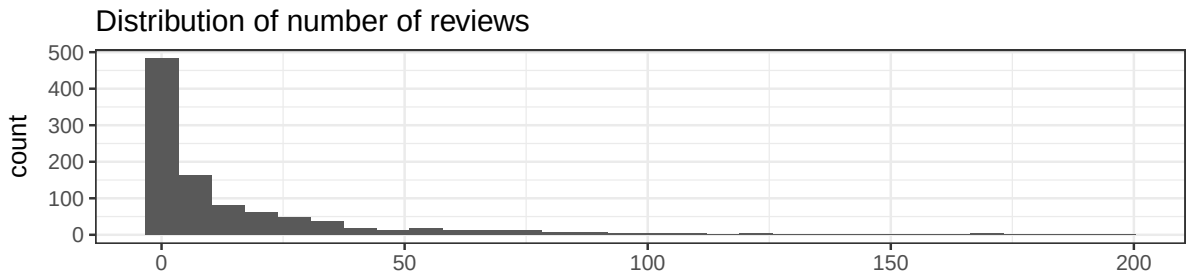
Data: Airbnbs in NYC

```
airbnb <- read_csv("data/NYCairbnb-sample.csv")
```

Data from BMLR Section 4.11.

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

EDA


Distribution of number of reviews
count
500
400
300
200
100
0
0
50
100
150
200


Log mean # of reviews vs. price
Log mean # of reviews vs. price
Airbnbs $2000 or less
Airbnbs $500 or less
Log(mean # reviews)
Log(mean # reviews)
5
4
3
2
1
0
0
500
1000
1500
0
100
200
300
400
500
Price in US dollars
Price in US dollars

EDA

mean	var
15.916	765.969

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \ \textit{price}_i + \beta_2 \ \textit{room\_type1}_i + \beta_3 \ \textit{room\_type2}_i$$

- 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?

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.

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.

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

- 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 $\log(\textit{days})$ to the model. This accounts for the fact that we would expect Airbnbs that have been listed longer to have more reviews.

$$\log(\lambda_i) = \beta_0 + \beta_1\, \textit{price}_i + \beta_2\, \textit{room\_type1}_i + \beta_3\, \textit{room\_type2}_i + \log(\textit{days}_i)$$

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 $\frac{reviews}{days}$.

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

$$\log\left(\frac{\lambda_i}{days_i}\right) = \beta_0 + \beta_1\ price_i + \beta_2\ room\_type1_i + \beta_3\ room\_type2_i$$

implies

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

implies

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

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

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`

Solution

- The mean number of **reviews per day** is predicted to decrease by 0.05% for each additional dollar in price.
- The mean number of **reviews per day** is predicted to decrease by 9.94% for each additional private room.

In summary, the mean number of **reviews per day** is not affected much by price but is by private rooms.

Key Differences in Interpretation

- Without Offset: The interpretation focuses on total reviews.
- With Offset: The interpretation focuses on reviews per unit of time (e.g., per day).

Goodness of Fit

H_o : The model is a good fit for the data

H_a : There is a 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

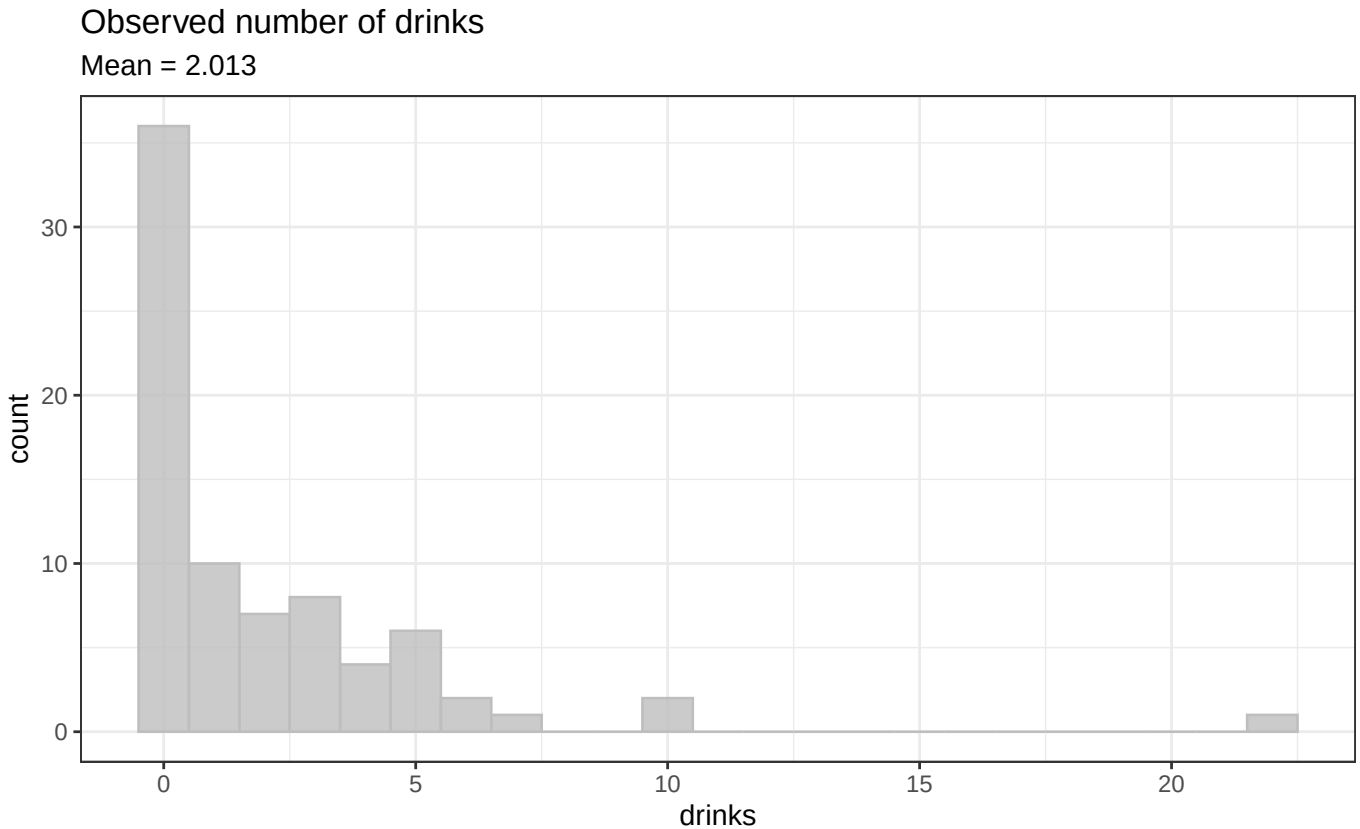
- `drinks`: Number of drinks they had in the past weekend
- `off_campus`: 1 - lives off campus, 0 otherwise
- `first_year`: 1 - student is a first-year, 0 otherwise
- `sex`: f - student identifies as female, m - student identifies as male

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

Data from case study in BMLR Section 4.10.

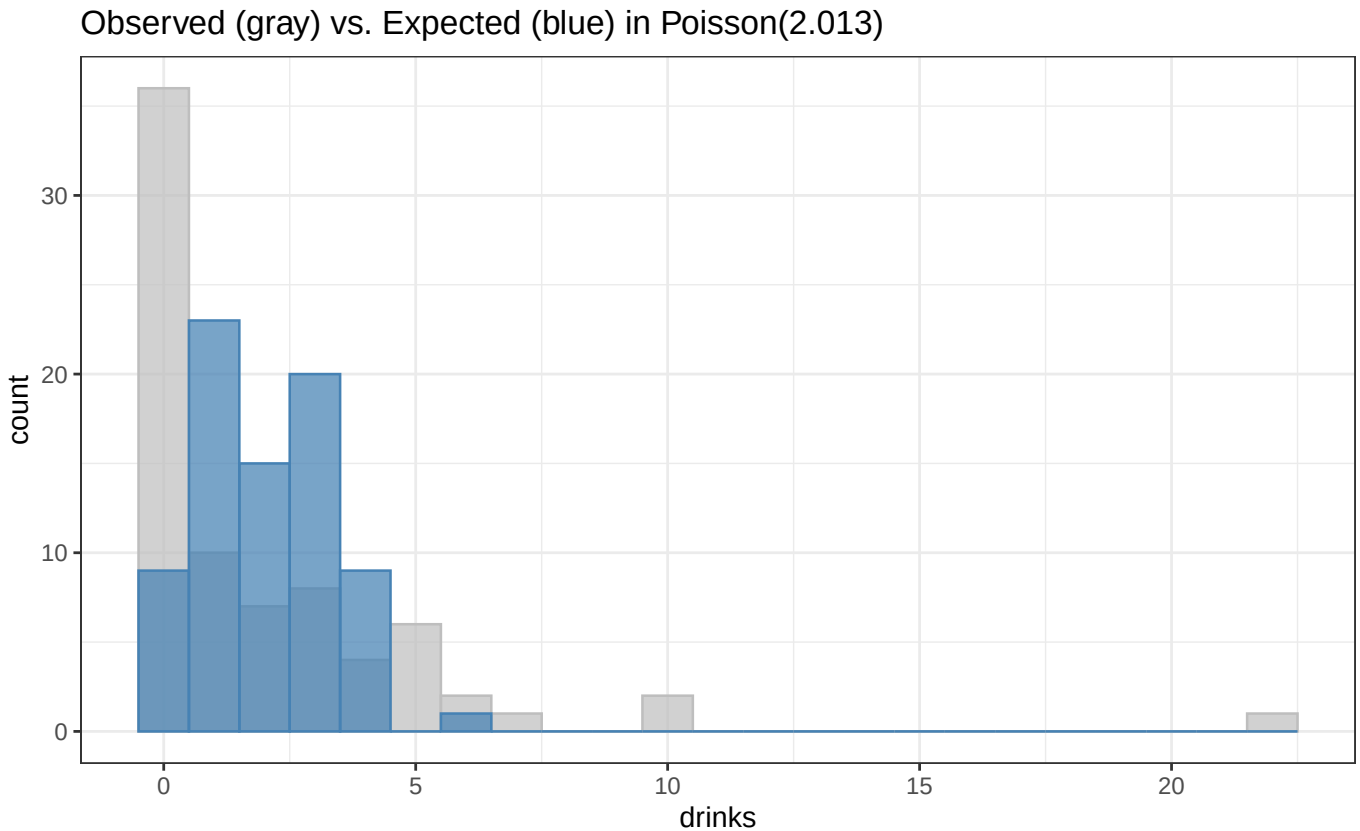
EDA: Response variable

```
drinks <- read_csv("data/weekend-drinks.csv")
```

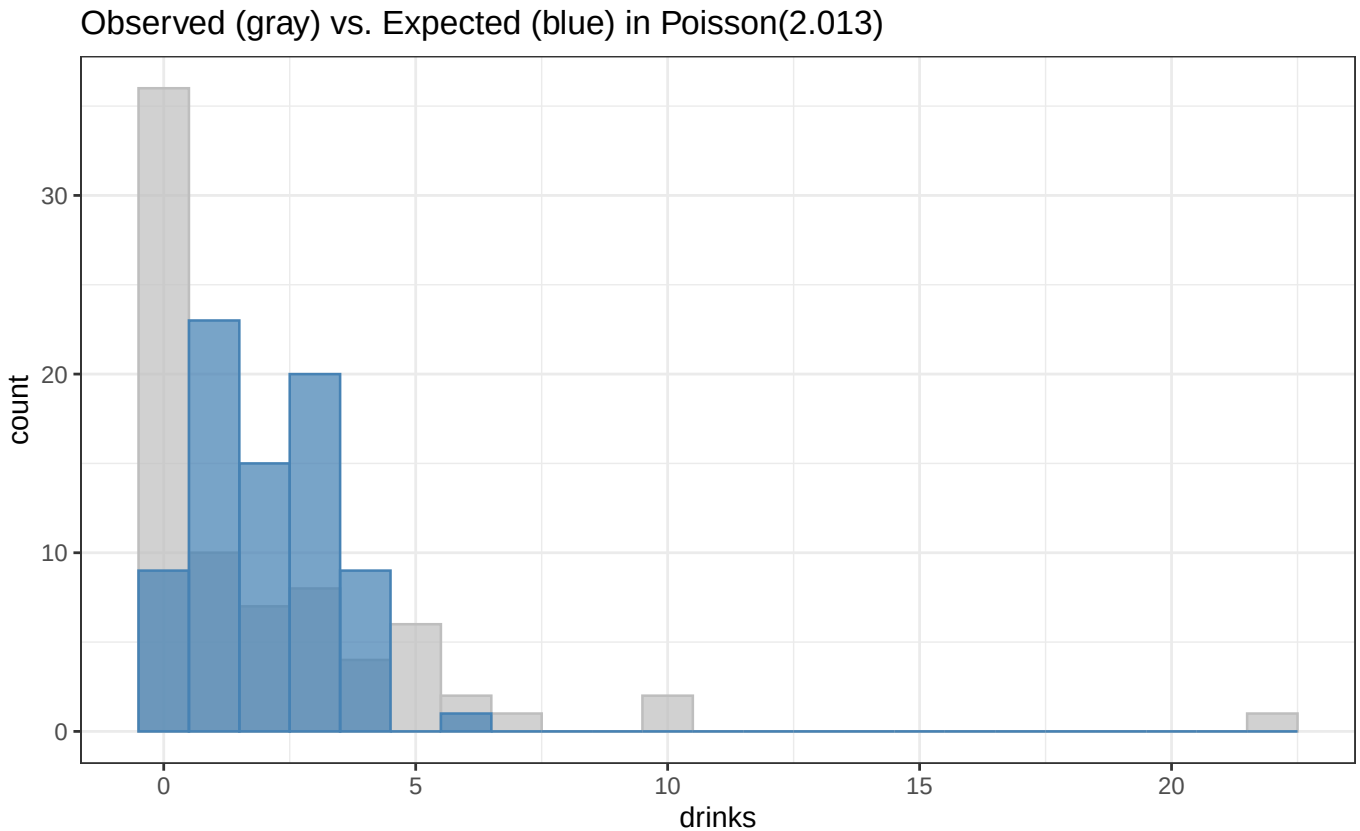


mean	var
2.013	10.75

Observed vs. expected response



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**

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:

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:

1. The association between mean number of drinks and various characteristics among those who drink
2. The estimated proportion of non-drinkers

Zero-inflated Poisson model

Zero-inflated Poisson (ZIP) model has two parts

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \ \mathit{off\_campus}_i + \beta_2 \ \mathit{sex}_i$$

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

Zero-inflated Poisson model

Zero-inflated Poisson (ZIP) model has two parts

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \ \textit{off\_campus}_i + \beta_2 \ \textit{sex}_i$$

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

2. Probability that a student does not drink

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

where α is the proportion of non-drinkers

Zero-inflated Poisson model

Zero-inflated Poisson (ZIP) model has two parts

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

$$\log(\lambda_i) = \beta_0 + \beta_1 \ \textit{off\_campus}_i + \beta_2 \ \textit{sex}_i$$

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

2. Probability that a student does not drink

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

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
```

Count model coefficients

Consider the "count model coefficients", which provide information on how the sex and off-campus status of a student who is a drinker are related to the number of drinks report by that student over a weekend.

We interpret these parameters as we have with previous Poisson models and exponentiate each coefficient.

Tidy output

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

```
library(poissonreg)
```

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	2.021	0.043
sexm	count	1.021	0.175	5.827	0.000

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	2.021	0.043
sexm	count	1.021	0.175	5.827	0.000

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

Solution: 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	2.021	0.043
sexm	count	1.021	0.175	5.827	0.000

- Interpret the intercept.

For those that drink, the expected mean number of drinks when all covariates are zero is

$$\exp(\beta_0) = \exp(0.754) = 2.12.$$

Solution: 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	2.021	0.043
sexm	count	1.021	0.175	5.827	0.000

- Interpret the coefficients `off_campus` and `sexm`.

For those that drink, the average number of drinks for males is

$$\exp(1.0209) = 2.76$$

times the number for females given we are comparing to students that live in similar settings.

For those that drink, the mean number of drinks for students living off campus is

$$\exp(0.4159) = 1.52$$

times that of students living on capus for those of the same sex.

Zero-inflated model coefficients

We now consider the "zero-inflated model coefficients", which refer to seperating drinkers from non-drinkers.

There is a first_year indicator, which is a binary indicator (0/1).

Tidy output

Proportion of non-drinkers

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?

Solution: Estimated proportion zeros

- What is the probability a first-year student is a non-drinker?

Observe that

$$\exp 1.136 = 3.11.$$

It is interpreted as the

$$\text{odds}\left(\frac{\alpha}{1-\alpha}\right)$$

that a first year student is a non-drinker is 3.11 times the odds that an upper class student is a non-drinker.

Recall that

$$\log(\frac{\alpha}{1-\alpha}) = -0.6036 + 1.1364 \textit{firstYear}$$

This implies that

$$\alpha = \frac{\exp\{-0.6036 + 1.1364 \textit{firstYear}\}}{1 + \exp\{-0.6036 + 1.1364 \textit{firstYear}\}}.$$

Solution: Estimated proportion zeros

- What is the probability a first-year student is a non-drinker?

Thus, the estimated probability that a first year student does not drink is

$$\frac{\exp\{-0.6036 + 1.1364\}}{1 + \exp\{-0.6036 + 1.1364\}} = \frac{\exp\{0.533\}}{1 + \exp\{0.533\}} = 0.63.$$

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 probability a upperclass student (sophomore, junior, senior) is a non-drinker?

The estimated probability that an upperclass student does not drink is 0.354.

Likelihoods for ZIP model

Probabilities under ZIP model

There are three different types of observations in the data:

- Observed 0 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

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

Probabilities under ZIP model

True zeros

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

Observed 0 but will not always be 0

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

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

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

Probabilities under ZIP model

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

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

Recall from our example,

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

$$\alpha_i = \frac{e^{\beta_{0\alpha} + \beta_{1\alpha} \ first\_year_i}}{1 + e^{\beta_{0\alpha} + \beta_{1\alpha} \ first\_year_i}}$$

Plug in λ_i and α_i into the above equation obtain the likelihood function

References

Awad, Annika, Evan Lebo, and Anna Linden. 2017. "Intercontinental Comparative Analysis of Airbnb Booking Factors."