Categorical: Poisson regression

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this lecture

- More count models
  - Too many or too few zeroes
  - Variable lengths of time
  - R^2 values
  - Comparing models

2 Zeroes

2.1 Zeroes (0)

2.1.1 Zeroes can be really important

- Conceptually, zeroes are meaningful
  - Lowest possible value of a count
  - Indicate "nothing"
- Three situations:
  - Too *few* zeroes
  - Too *many* zeroes (and **some** zeroes will always be zeroes)
  - Too *many* zeroes (and **all** zeroes will always be zeroes)

2.2 Too few zeroes

2.2.1 Too few zeroes

- Situation: *Outcome is a count*
  - But it cannot take on a value of 0
- Study of medical visits
  - Must visit the doctor to get involved in the study
- Study of substance use
  - Only recruit substance users

2.2.2 Truncated Poisson regression

- *Truncated* Poisson regression
  - Also *truncated* negative binomial
- Probability distribution removes the probability of zeroes
  - Only **positive** integer values

2.3 Too many zeroes

2.3.1 "Excess" zeroes

- Counts often display "excess" zeroes
  - More values of 0 than expected for a Poisson distribution
- Even if the rest of the distribution is approximately Poisson
  - "Excess" zeroes lead to overdispersion
  - Sometimes, what *looks like overdispersion* is really *excess zeroes*
- Several specific Poisson family models to deal with excess zeroes
  - Depending on **why** the zeroes are there

2.3.2 Why are there all these zeroes?

- This is a **substantive** question
  - Know about the outcome you're studying
- Do some people who are responding zero have **some probability** of responding otherwise?
  - Yes: **Zero-inflated Poisson regression** (or NB)
  - No (structural zeroes): **hurdle regression** (also called **with-zeroes regression**)

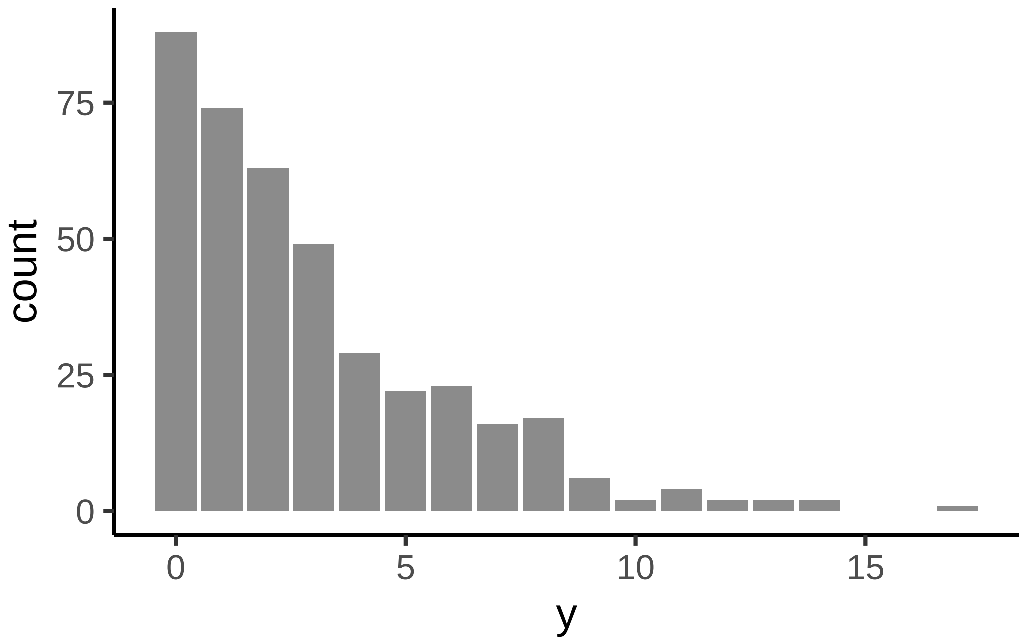
2.3.3 Why are there all these zeroes?

- Do the people who are responding zero have **some probability** of responding otherwise?
  - Cigarettes smoked today
    - Smoker who hasn't smoked yet today *could respond with non-zero*
    - Non-smoker **could not** respond with non-zero
  - Alcoholic beverages consumed today
    - Someone who drinks but hasn't today *could respond with non-zero*
    - Abstainer **could not** respond with non-zero

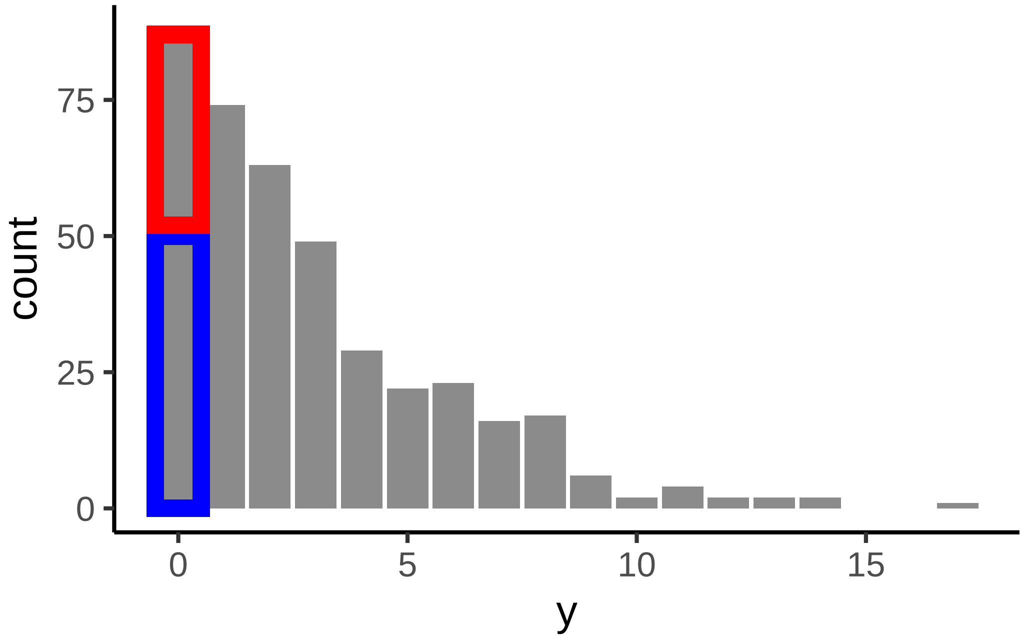
2.3.4 Zero-inflated Poisson regression

- Zeroes have some probability to be non-zero
- Two parts modeled simultaneously:
  - Logistic regression
    - **Structural zero** (must be 0) or **not**
  - Poisson regression (or OD Poisson or NB)
    - Non-structural zeroes and positive values
- Can use same set of predictors in both parts, but do not have to

2.3.5 ZIP: Some zeroes are always zeroes, some are not



2.3.6 ZIP: Some zeroes are always zeroes, some are not



2.3.7 Output: Zero-inflated Poisson regression

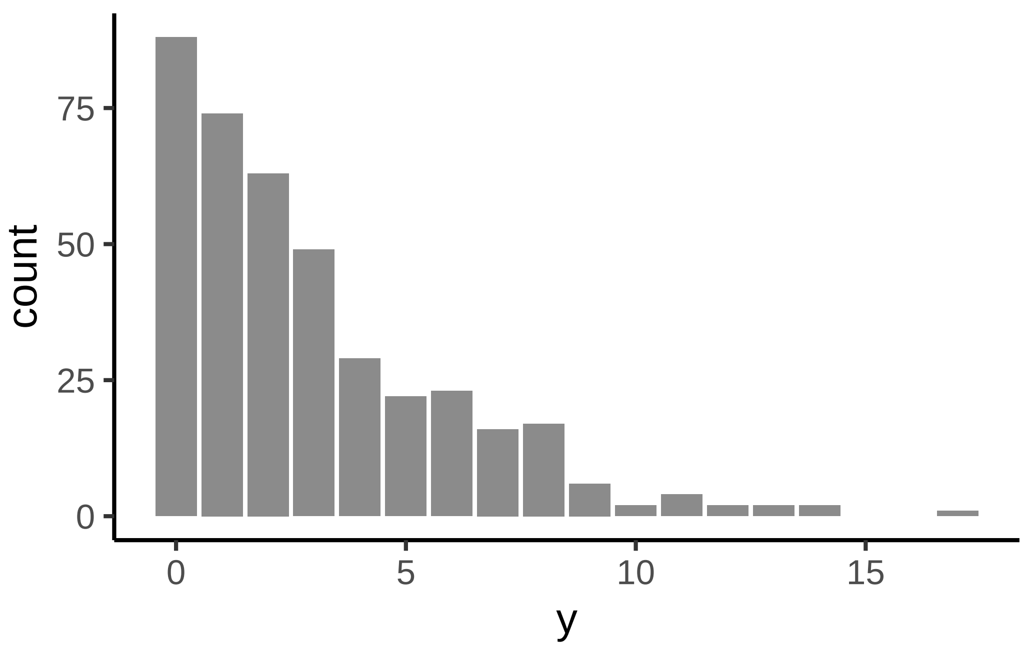
```
Call:
zeroinfl(formula = y ~ sensation4 | sensation4, data = jpa)

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.4675 -0.9658 -0.3395  0.7122  4.9638 

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.02579    0.06577  15.597  < 2e-16 ***
sensation4   0.21183    0.04304   4.922 8.57e-07 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.2836     0.2633  -4.874 1.09e-06 ***
sensation4   -0.1152     0.1850  -0.622    0.534    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Number of iterations in BFGS optimization: 8
Log-likelihood: -953.4 on 4 Df
```

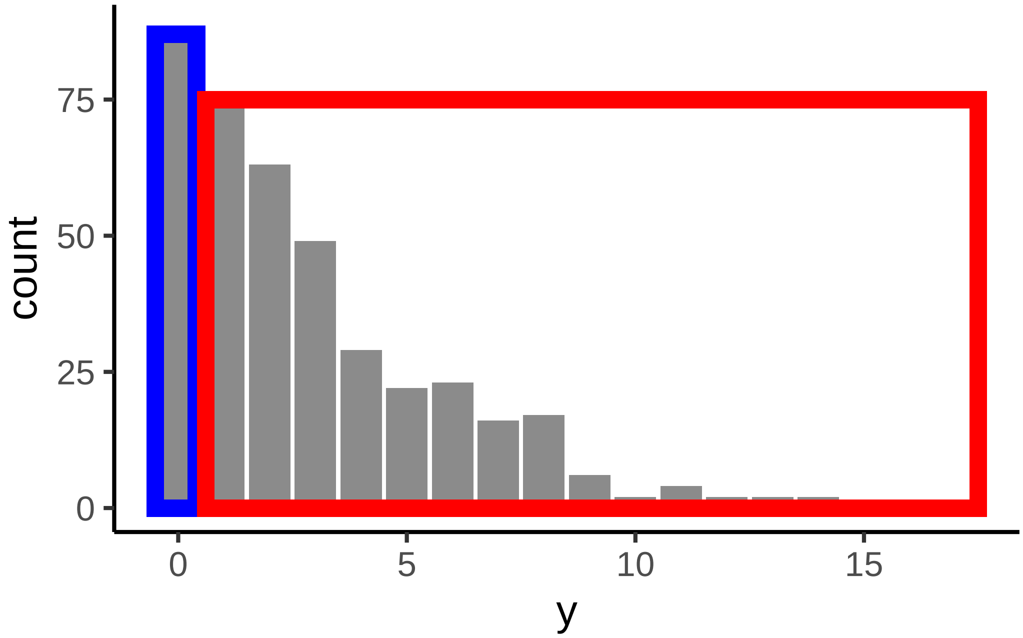
2.3.8 Hurdle regression (or with-zeroes regression)

- Zeroes have **no probability** to be non-zero
  - Two different populations: Smokers vs not, drinkers vs not
- Two parts modeled simultaneously:
  - Logistic regression
    - **Zero** or **not zero**
  - Truncated Poisson regression (or OD Poisson or NB)
    - **Positive values only**
- Can use same set of predictors in both parts, but do not have to

2.3.9 Hurdle: All zeroes are structural zeroes



2.3.10 Hurdle: All zeroes are structural zeroes



2.3.11 Output: Hurdle regression

```
Call:
hurdle(formula = y ~ sensation4, data = jpa)

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.4825 -0.9675 -0.3363  0.7135  4.9816 

Count model coefficients (truncated poisson with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.02798    0.06547  15.701  < 2e-16 ***
sensation4   0.21022    0.04283   4.908 9.22e-07 ***
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.0158     0.2112   4.810 1.51e-06 ***
sensation4    0.2177     0.1556   1.399    0.162    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 7 
```

```
Log-likelihood: -953.5 on 4 Df
```

3 Miscellaneous

3.1 Pseudo R^2

3.1.1 Pseudo R^2 for count models

- Many of the same issues as logistic regression
  - No sums of squares
  - Not always between 0 and 1
  - Don't always increase with added predictors
  - Several options

3.1.2 Poisson regression model

```
Call:
glm(formula = y ~ sensation4, family = poisson(link = "log"),
    data = jpa)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7912  -1.5001  -0.4624   0.8418   4.9122

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.78560    0.05977  13.144  < 2e-16 ***
sensation4   0.23148    0.03966   5.837 5.33e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1186.8  on 399  degrees of freedom
Residual deviance: 1151.7  on 398  degrees of freedom
AIC: 2079

Number of Fisher Scoring iterations: 5
```

3.1.3 (Squared) correlation between predicted & observed

- Literally, the (squared) correlation between the *observed* (Y) values and the *predicted* ($\hat{Y}$) values
  - For the Poisson regression example, this value is 0.031
- Mathematically the same as Efron's R^2

3.1.4 McFadden R^2 (a.k.a. Likelihood ratio R^2, pseudo R^2)

- $R^2_{McFadden} = 1 - \frac{LL_{model}}{LL_{null}}$
- For the Poisson regression example, this value is 0.017
- Proportion of variance accounted for
  - Proportion of the way from null model to perfect model

3.1.5 Caution about deviance and log-likelihood

Warning

- For many models (e.g., logistic regression)
  - Deviance = -2 * log-likelihood
  - Use either deviance or LL for calculations
- **Count models don't work like that**
  - Deviance $\neq$ -2 * log-likelihood
  - Much more complicated, due to scaling, LL value in null model
  - Here are some links with more info
- Don't calculate things like R^2 or LR test by hand
  - Let the program do it for you: It will use the correct values

3.2 Variable length of time

3.2.1 Poisson distribution assumption

- Poisson distribution (and extensions) model the number of events in a **fixed length of time**

- Everyone is measured for the same time frame
  - Number of aggressive acts committed by a child *in 1 hour*
  - Number of cigarettes smoked *per day*
  - Number of alcoholic drinks consumed *on Saturday*

3.2.2 Variable length of time

- Often, we measure a count over some *variable period of time*
  - Number of aggressive acts committed by a child *while playing*
  - Number of cigarettes smoked *today before you came in*
  - Number of alcoholic drinks consumed *the last time you drank*

3.2.3 Variable length of time

- Extend Poisson-type model to incorporate *variable time period*
  - Include $ln(time)$ (measurement interval) as a predictor with regression coefficient $= 1$
    * $ln(\hat{\mu}) = ln(time) + b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$
  - The "offset" option
    * `offset(logtime)` in `glm()`, where `logtime` is the log(time)

3.2.4 How does this work?

$$ln(\hat{\mu}) = ln(time) + b_0 + b_1X_1$$

- Subtract $ln(time)$ from both sides

$$ln(\hat{\mu}) - ln(time) = b_0 + b_1X_1$$

- Subtraction converts to **division** (i.e., $ln(x) - ln(y) = ln(x/y)$)

$$ln(\frac{\hat{\mu}}{time}) = b_0 + b_1X_1$$

- Predict $ln(count\ per\ unit\ of\ time)$ instead of $ln(count)$

3.2.5 Code for offset

```r
m1 <- glm(y ~ sensation4 + offset(logtime),
          data = jpa,
          family = poisson(link = "log"))
```

- Where `logtime` is ln(time variable)

3.3 Comparing models

3.3.1 Nested models

- We've already talked about comparing **nested models** using **LR test**
  - Linear regression, logistic regression, ordinal logistic, SEM
- Whether models are nested is more complicated for count models
  - Mostly because of overdispersion parameter

3.3.2 What's in each model?

Model	Coefficients	ψ	α
Poisson	count	fixed at 1	fixed at 0
Overdispersed Poisson	count	ψ	fixed at 0
Negative Binomial	count	fixed at 1	α
ZIP	count, logistic	fixed at 1	fixed at 0
ZIOD Poisson	count, logistic	ψ	fixed at 0
ZINB	count, logistic	fixed at 1	α

3.3.3 What is nested?

- Poisson **is nested** within overdispersed Poisson
  - But Poisson and OD Poisson have the *same degrees of freedom*
  - So LR tests don't work
- Poisson **is nested** within negative binomial
  - Do a LR test!

3.3.4 What is not nested?

- OD Poisson and NB are **not nested** (either direction)
  - Different overdispersion parameters
- Non-inflated models are **not nested** in zero-inflated models
  - Different sets of coefficients in the models
  - Logistic regression coefficients in inflated models

3.3.5 What is not nested?

- Overdispersed models (OD Poisson or NB) with *different predictors*
  - Model 1: Overdispersed Poisson with predictor X_1
  - Model 2: Overdispersed Poisson with predictors X_1 and X_2
- These models have **different overdispersion parameters**
  - Model 1: ψ parameter based on X_1
  - Model 2: ψ parameter based on both X_1 and X_2

3.3.6 LR tests

- Overdispersed Poisson vs Poisson
  - Technically nested, but *same degrees of freedom*
    - 0 df for the LR test

```r
lrtest(poi, odpoi)
```

```
Likelihood ratio test

Model 1: y ~ sensation4
Model 2: y ~ sensation4
  #Df  LogLik Df Chisq Pr(>Chisq)
1   2 -1037.5
2   2          0
```

3.3.7 LR tests

- Negative binomial vs Poisson
    - *If same predictors*: Test of **overdispersion** parameter

```
lrtest(poi, negbin)
```

```
Likelihood ratio test

Model 1: y ~ sensation4
Model 2: y ~ sensation4
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   2 -1037.51
2   3  -883.05  1 308.93  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3.8 Non-nested models?

- Vuong test: `vuong()` function in **pscl** package
    - Don't use to compare zero-inflated to non-inflated models: Why
- AIC and BIC
    - Function of log-likelihood and number of parameters
        - Penalizes models with more parameters
    - Smaller values are better
    - No associated test

3.3.9 Compare using AIC: Smaller is better

Model	AIC
Poisson	2079.025
OD Poisson	NA
Negative binomial	1772.091

4 Summary

4.1 Summary

4.1.1 Summary of this week

- Zeroes are interesting and important
  - Too few or too many zeroes
- Pseudo R^2: Similar to logistic regression
- Variable length of time for observations
- Comparing models
  - Nested or (more likely) not

4.1.2 Next week

- **Interactions** in GLiMs
  - Nonlinear models: What is an *interaction* if lines can't be parallel?
- **Mediation** with GLiMs
  - Just use the a and b paths like "normal"? Nope
  - What are those coefficients, conceptually?
    - What in a GLiM corresponds to that?
- Wrap up any other details of GLiMs that I have time for
  - Residuals? Diagnostics?