Poisson Regression in R

Andre Archer
Northwestern University
Research Computing Services

Format of the Online Workshop

- In this workshop, I will be using Google Slides and live coding in RStudio
  - I will be using the Rmd file, linear_model_code.Rmd, to teach the workshop.

- If you have an questions, please put them in the chat.
  - There are TAs monitoring the chat. They will respond to questions.
  - If necessary, I will be interrupted by a TA.

Contents

Data Description

Data we are working with

- Dataset contains 915 samples of the number of articles published by PhD student in 3 years
- The dataset set contains variables:
    - art - number of articles produced by the student in the last 3 years of their PhD
        - Discrete positive numbers (0, 1, 2, 3, …)
    - fem - sex of the student
        - Categorial: “Men” or “Women”
    - mar - martial status of the student
        - Categorical: Single, Married
    - kid5 - number of children less than 5
        - Categorical: 0, 1, 2, 3
    - phd - prestige of PhD program
        - Continuous variable
    - ment - number of articles of the mentor in the last 3 years
        - Discrete positive numbers (0, 1, 2, 3, …)

Goals of this workshop

Goals

1) Predict the average number of articles of a PhD student in the last 3 as a function of sex, marital status and number of articles of their mentors

2) Understand the effects of sex, marital status and number of articles of their mentors on the average number of articles of a student

Linear Regression vs Poisson Regression

Why Linear Models are not appropriate

- Number of articles is a discrete positive variables
    - Linear models do not restrict the response variables to positive non-negative integers
- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?

Why Linear Models are not appropriate

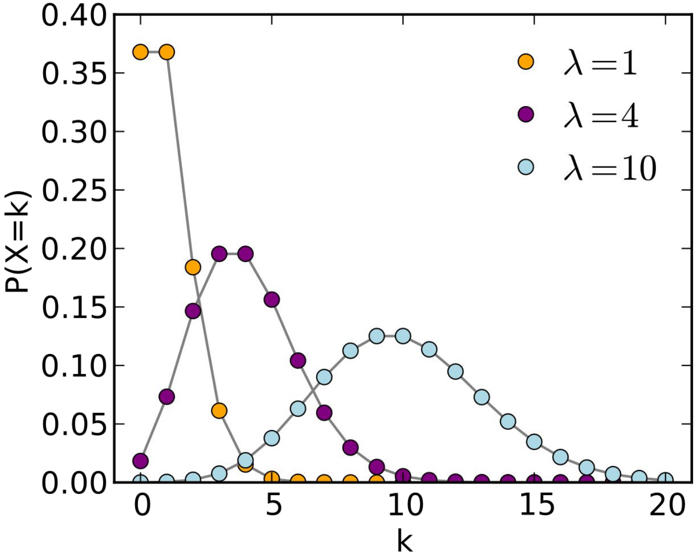
- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?
- We can assume that the response variables are Poisson distributed

$$\text{Probability that response variable is } k = \frac{\lambda^k e^{-\lambda}}{k!}$$

Why Linear Models are not appropriate

- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?
- We can assume that the response variables are Poisson distributed

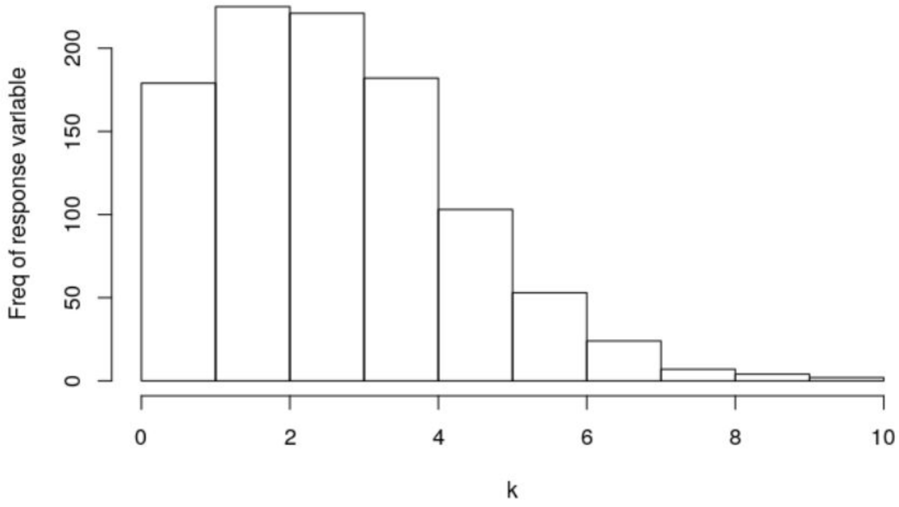
$$\text{Probability that response variable is } k = \frac{\lambda^k e^{-\lambda}}{k!}$$



Why Linear Models are not appropriate

- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?
- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset

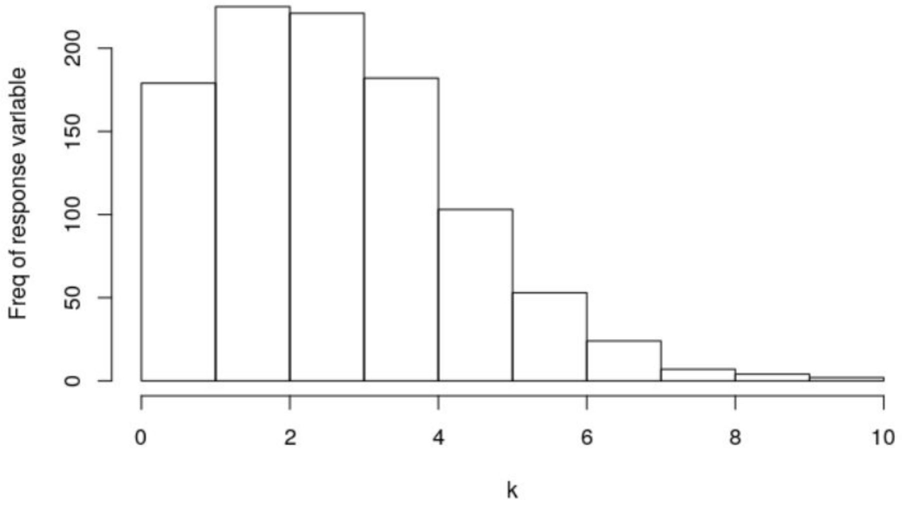
Response Variable
1
1
0
3
2
$\vdots$
0
1



Why Linear Models are not appropriate

- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?
- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset

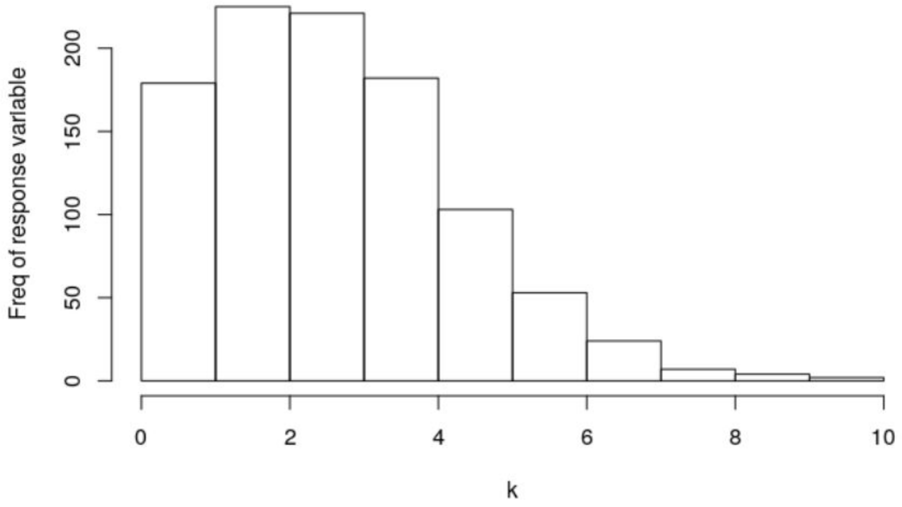
$$\text{Probability that response variable is } k = \frac{\lambda^k e^{-\lambda}}{k!}$$



Why Linear Models are not appropriate

- Linear models assume that response variables are normally distributed. How do we change the assumption on the response variable?
- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
- If Poisson distributed and with enough data, λ should be the mean and standard deviation squared of the data set

$$\text{Probability that response variable is } k = \frac{\lambda^k e^{-\lambda}}{k!}$$



How does Poisson Regression Work

- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
  - How do I include explanatory variables in this recovery process?

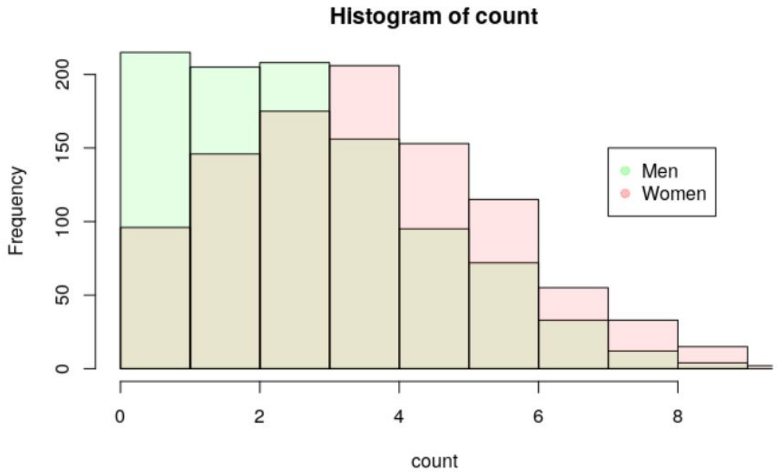
$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

How does Poisson Regression Work

- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
  - How do I include explanatory variables in this recovery process?

$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable } 1 + \beta_2 \times \text{Explanatory Variable } 2 + \ldots$$

  - For example, let's say that we have a dataset of count by sex



How does Poisson Regression Work

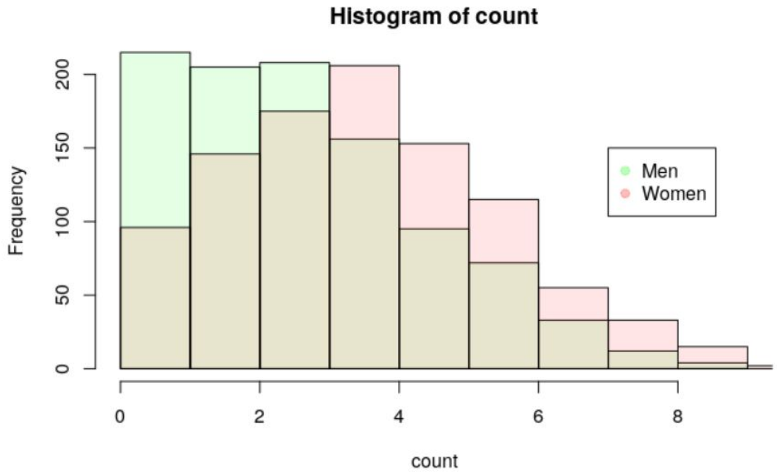
- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
  - How do I include explanatory variables in this recovery process?

$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

  - For example, let's say that we have a dataset of count by sex
  - We could fit each sex to their own Poisson distribution

$$\text{Probability that the response variable is } k = \frac{\lambda_F^k e^{-\lambda_F}}{k!}$$

$$\text{Probability that the response variable is } k = \frac{\lambda_M^k e^{-\lambda_M}}{k!}$$



How does Poisson Regression Work

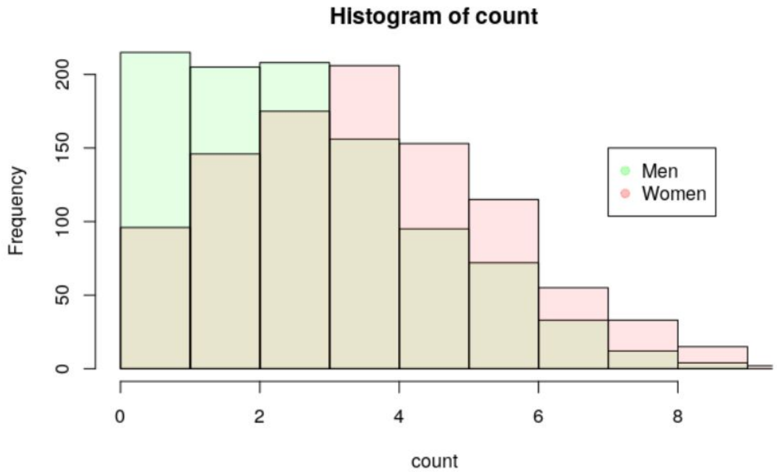
- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
  - How do I include explanatory variables in this recovery process?

$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

  - For example, let's say that we have a dataset of count by sex
  - We could fit each sex to their own Poisson distribution
  - We could instead do a fit with

$$\lambda = \beta_0 + \beta_1 \times \text{Sex}$$

Sex = 1 if Women
Sex = 0 if Men



How does Poisson Regression Work

- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
    - How do I include explanatory variables in this recovery process?

$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

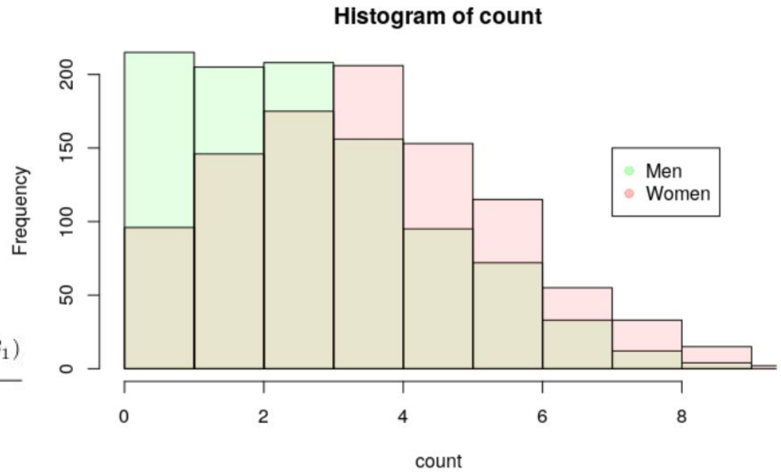
- For example, let's say that we have a dataset of count by sex
- We could fit each sex to their own Poisson distribution
- We could instead do a fit with

$$\lambda = \beta_0 + \beta_1 \times \text{Sex}$$

- Therefore,

$$\text{Probability that the response variable for men is } k = \frac{\beta_0^k e^{-\beta_0}}{k!}$$

$$\text{Probability that the response variable for women is } k = \frac{(\beta_0 + \beta_1)^k e^{-(\beta_0+\beta_1)}}{k!}$$



How does Poisson Regression Work

- We can assume that the response variables are Poisson distributed
- Given the data, Poisson regression attempts to recover the λ that generated the dataset
    - How do I include explanatory variables in this recovery process?

$$\lambda = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

    - With continuous explanatory variables, we run the risk of λ becoming negative
    - A negative λ is problematic since λ is the expected count given the explanatory variables

- To prevent negative λ , a log “link function” is used.

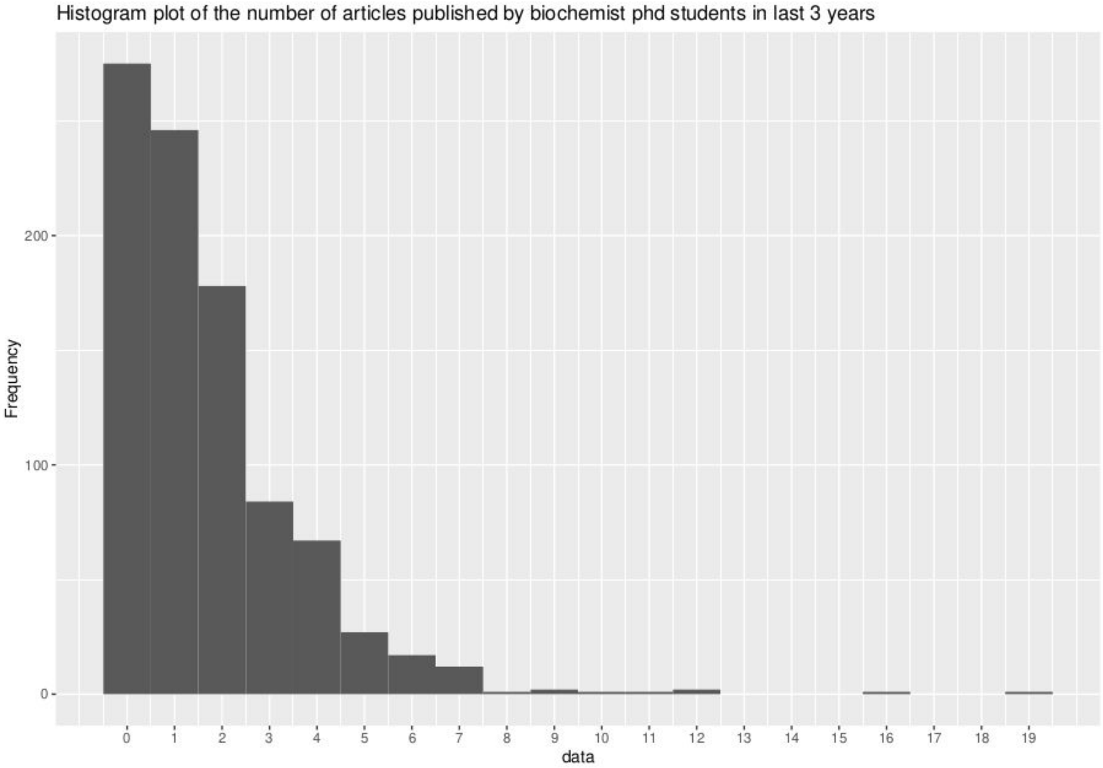
$$\log(\lambda) = \beta_0 + \beta_1 \times \text{Explanatory Variable 1} + \beta_2 \times \text{Explanatory Variable 2} + \ldots$$

    - To interpret the effects of the each explanatory variable, we use the formula

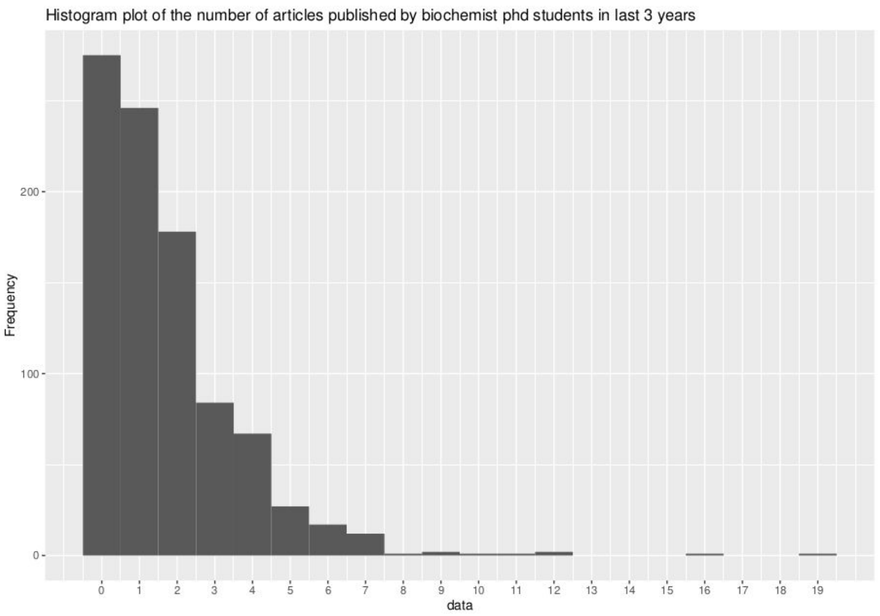
$$\lambda = e^{\beta_0} e^{\beta_1 \times \text{Explanatory Variable 1}} e^{\beta_2 \times \text{Explanatory Variable 2}} \ldots$$

Poisson Model with only a constant term

Histogram of the papers published in the last 3 years by PhD students



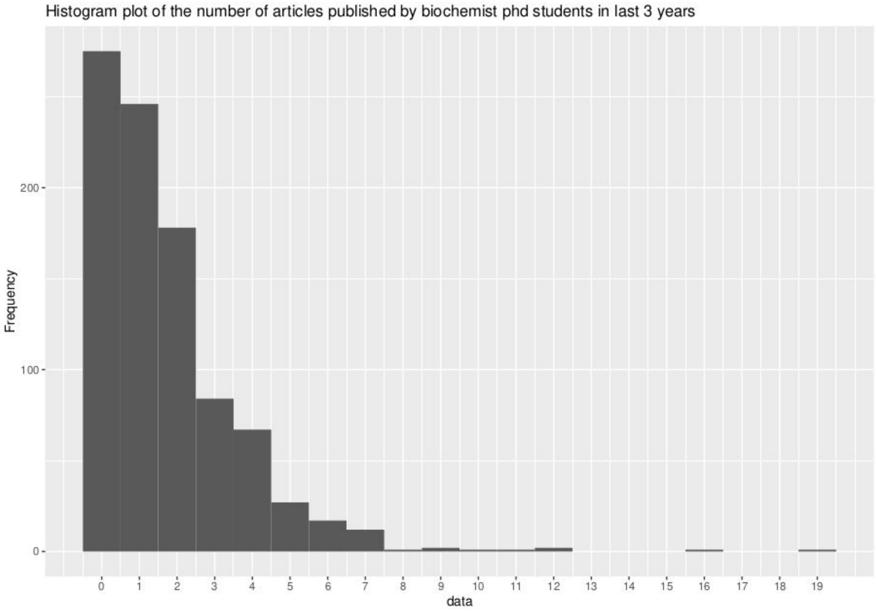
Poisson Model with only a constant term



- The data appears to be Poisson distributed
- Let's fit the number of articles to a Poisson distribution

$$\log(\lambda) = \beta_0$$

Poisson Model with only a constant term



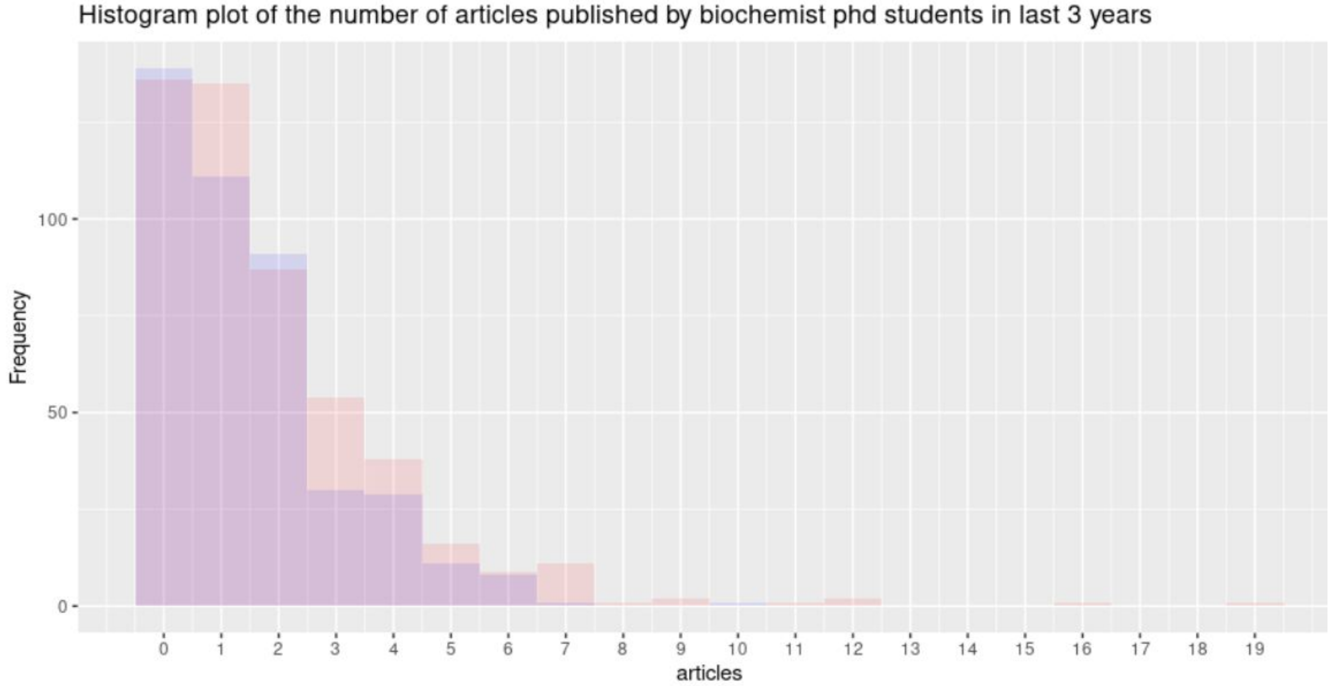
- The data appears to be Poisson distributed
- Let's fit the number of articles to a Poisson distribution

$$\log(\lambda) = \beta_0$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.52644	0.02541	20.72	<2e-16	***

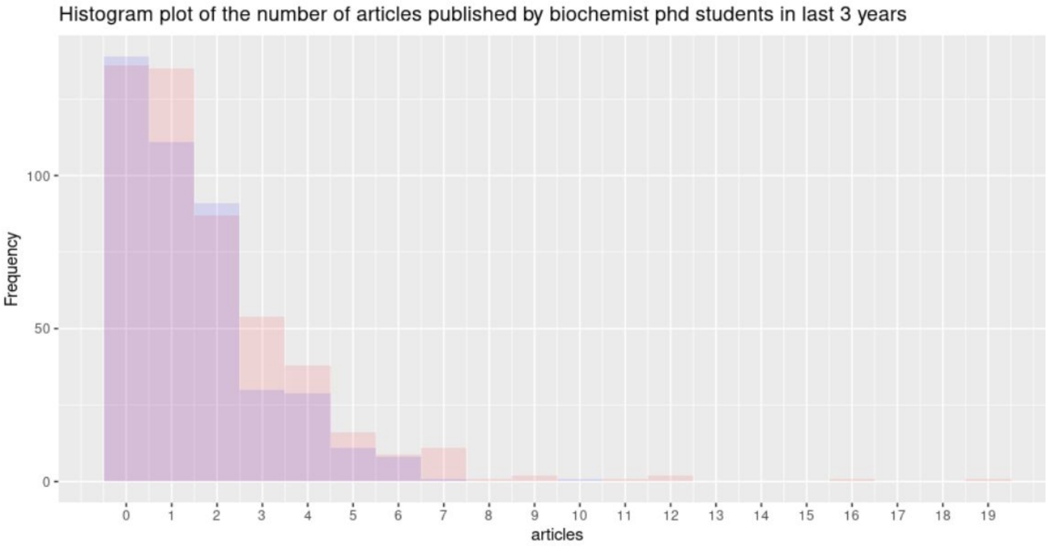
Poisson Model with sex and constant term

Poisson model with sex and constant term



- Pink - men
- Blue - women

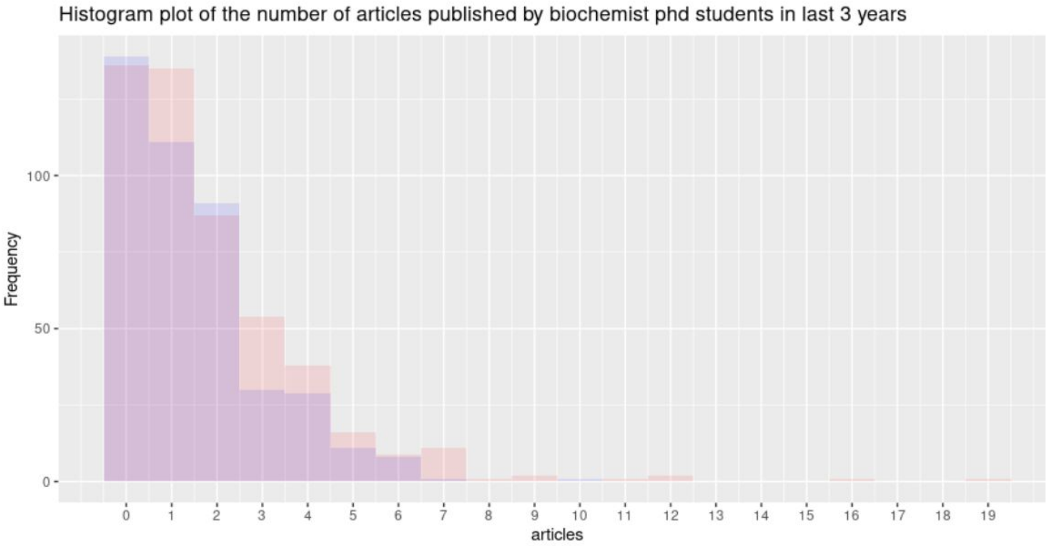
Poisson model with sex and constant term



- Each sex appears to follow their own Poisson distribution

- Pink - men
- Blue - women

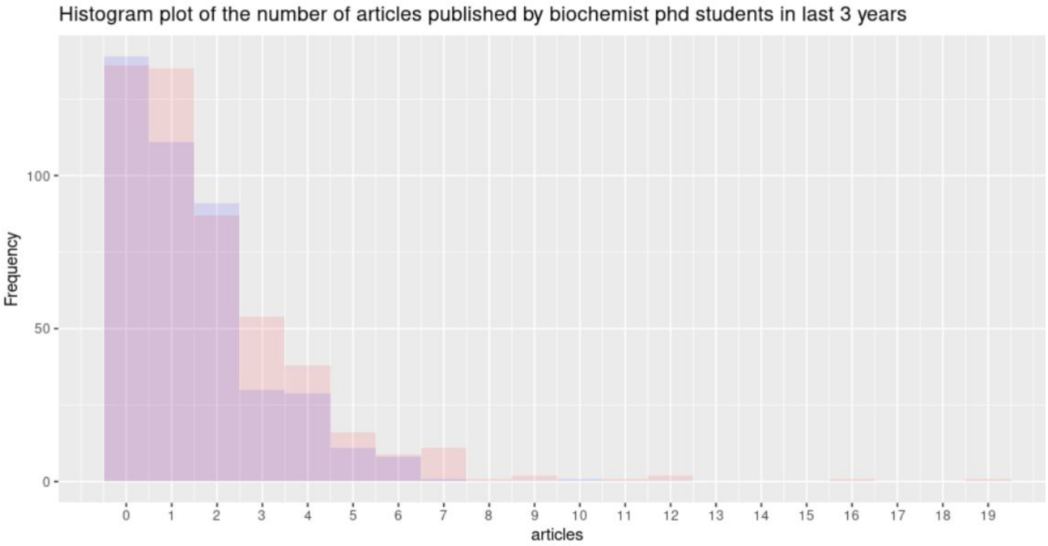
Poisson model with sex and constant term



- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution
- Let's fit the number of articles to a Poisson distribution with sex as an explanatory variable

Poisson model with sex and constant term

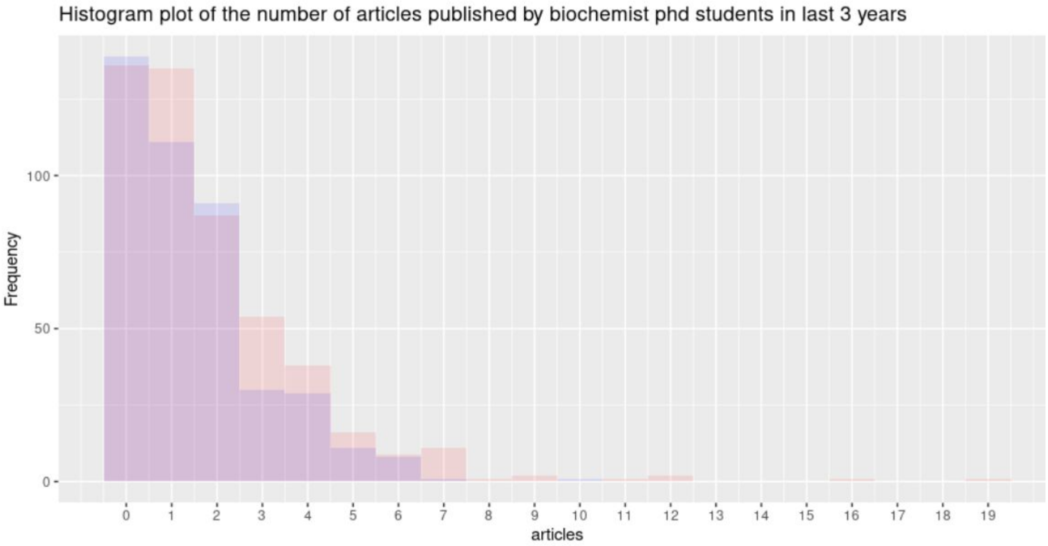


- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution
- Let’s fit the number of articles to a Poisson distribution with sex as an explanatory variable

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{fem}$$

Poisson model with sex and constant term



- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution

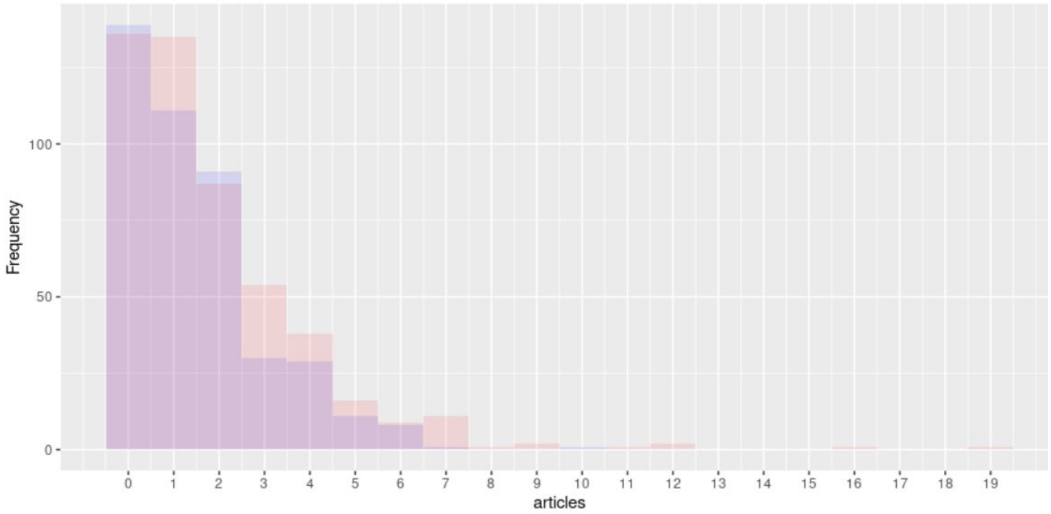
- Let’s fit the number of articles to a Poisson distribution with sex as an explanatory variable

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{fem}$$

fem = 0 if Men
fem = 1 if Women

Poisson model with sex and constant term

Histogram plot of the number of articles published by biochemist phd students in last 3 years



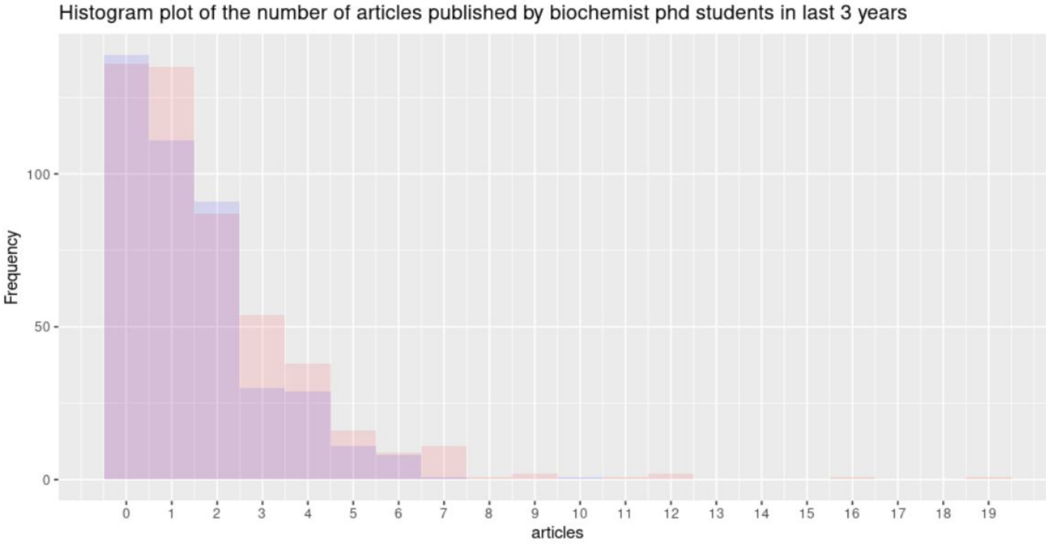
- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution
- Let's fit the number of articles to a Poisson distribution with sex as an explanatory variable

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{fem}$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.63265	0.03279	19.293	< 2e-16	***
femWomen	-0.24718	0.05187	-4.765	1.89e-06	***

Poisson model with sex and constant term



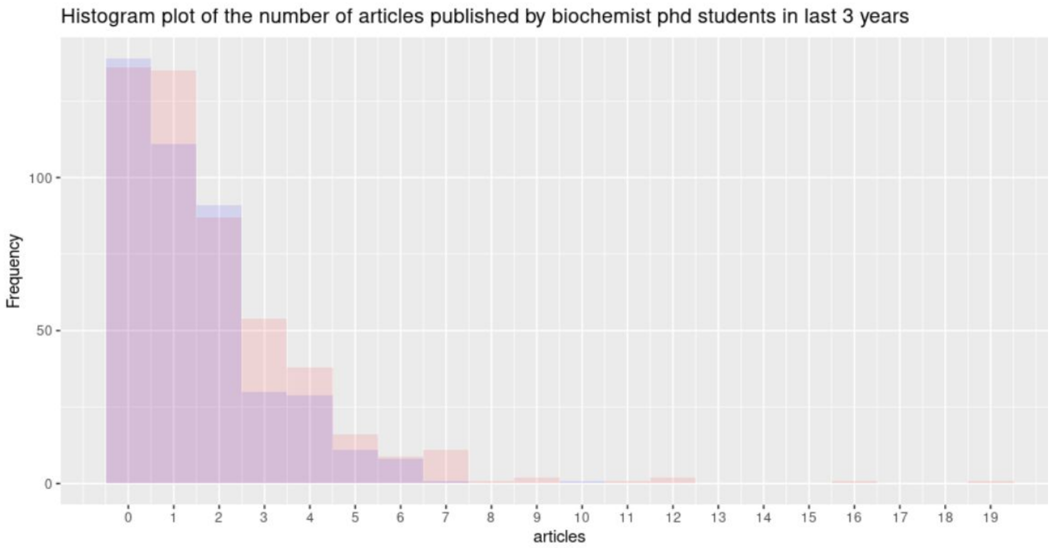
- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution
- Let's fit the number of articles to a Poisson distribution with sex as an explanatory variable

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{fem}$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.63265	0.03279	19.293	< 2e-16	***
femWomen	$\boxed{-0.24718}$	0.05187	-4.765	1.89e-06	***

Poisson model with sex and constant term



- Pink - men
- Blue - women

- Each sex appears to follow their own Poisson distribution
- Let's fit the number of articles to a Poisson distribution with sex as an explanatory variable

$$\log(\lambda) = 0.632 - 0.247 \times \text{fem}$$

- If Men,

$$\lambda = e^{0.632} = 1.88$$

- If Women,

$$\lambda = e^{0.632-0.247} = 1.47$$

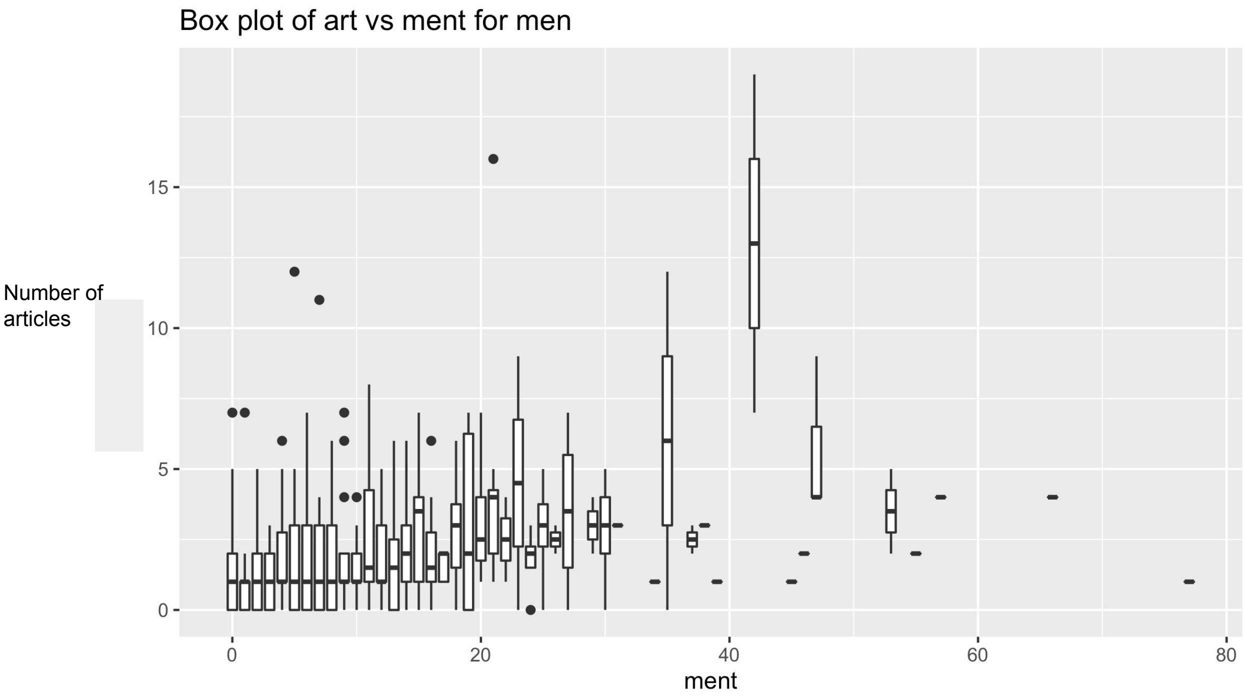
Deviance

- For a linear model, deviance is sum of squares of the residuals
- Deviance is a more generalized “sum of squares of the residuals” for GLMs, like logistic models and Poisson models
    - Significant reduction of deviance is important
    - Using ANOVA, deviance allows us to compare nested models

Poisson Model with Sex and the number of mentor's papers in the last 3 years

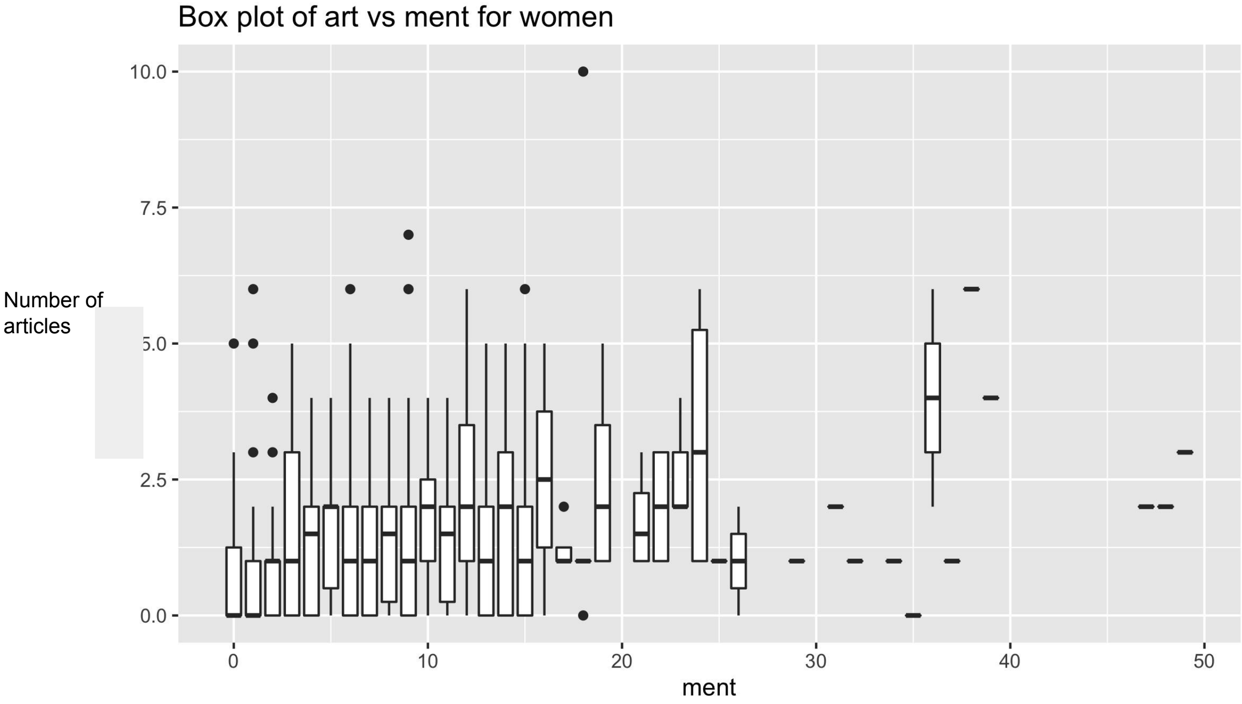
Poisson Model with sex and ment

- In this case, we are assume that each ment and sex pair has it’s own Poisson-like histogram



Poisson Model with sex and ment

- In this case, we are assume that each ment and sex pair has it’s own Poisson-like histogram



Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it's own Poisson-like histogram
  - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = \beta_0 + \beta_1 \text{fem} + \beta_2 \text{ment}$$

Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it’s own Poisson-like histogram
  - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = \beta_0 + \beta_1 \text{fem} + \beta_2 \text{ment}$$

- If fem = Men,

$$\lambda = \beta_0 + \beta_2 \text{ment}$$

- If fem = Women,

$$\lambda = \beta_0 + \beta_1 + \beta_2 \text{ment}$$

Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it's own Poisson-like histogram
  - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = \beta_0 + \beta_1 \text{fem} + \beta_2 \text{ment}$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.34909	0.04191	8.329	< 2e-16	***
femWomen	-0.18445	0.05235	-3.523	0.000426	***
ment	0.02510	0.00193	13.005	< 2e-16	***

Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it's own Poisson-like histogram
  - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = \beta_0 + \beta_1 \text{fem} + \beta_2 \text{ment}$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.34909	0.04191	8.329	< 2e-16	***
femWomen	$\boxed{-0.18445}$	0.05235	-3.523	0.000426	***
ment	0.02510	0.00193	13.005	< 2e-16	***

Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it’s own Poisson-like histogram
    - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = \beta_0 + \beta_1 \text{fem} + \beta_2 \text{ment}$$

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.34909	0.04191	8.329	< 2e-16	***
femWomen	-0.18445	0.05235	-3.523	0.000426	***
ment	0.02510	0.00193	13.005	< 2e-16	***

Poisson Model with sex and ment

- It actually appears that each ment and sex pair has it’s own Poisson-like histogram
  - The mean (λ) of the distribution seems to be increase function of ment for each sex

- We fit the data to a Poisson distribution using ment and sex as explanatory variables

$$\lambda = 0.349 - 0.184 \times \mathrm{fem} + 0.025 \times \mathrm{ment}$$

- Using ANOVA, we can determine the significance of the improvement in deviance from adding the ment variable

Conclusion and Next Steps

- We found reasonable Poisson models of the mean number of articles a student publishes in 3 years
- Exercises will allow you to experiment further with the phd, ment and sex variables
- Going further, you might want to consider
  - Model selection for Poisson regression
  - Goodness of fit measures: AIC, BIC
  - Statistical tests for goodness of fit
  - Issues with Poisson Models: Zero inflation, Dispersion
  - Modeling the rate (number of articles published per year) rather the average over the time

Exercises

1. Open the file poisson_model_exercises.Rmd
2. Get cracking!