

# Poisson Regression and Applications

Adam Dawson, Sarah Onstad-Hawes, and Asare Buahin

12/11/2019

# Introduction: Main Project Questions

- ▶ What is the Poisson Regression and what are its applications?
- ▶ Using the Poisson Regression, can we find significant relationships within our data set.

# Introduction: Our Data Set

- ▶ “Contact with Medical Doctors”
  - ▶ Cross-Sectional data collected in North Carolina between 1977-78
  - ▶ Examines 20186 observations of individuals
    - ▶ Randomly selected subset of 2000 observations for ease of computation
  - ▶ Collected from RAND Health Insurance Experiment (RHIE), which is the longest and largest socially controlled experiment regarding medical care (Price, D. 2002)
  - ▶ Our Response Variable of interest is mdu, which captures the number of times an individual visited a medical health profession during the study
  - ▶ Data Set also contains 14 other variables of interest

# Introduction: Our Research Question

- ▶ How is the number of doctor visits impacted by various factors in an individual's life?
  - ▶ Specifically, we want see how one's age, sex, income, physical limitations, and present diseases influence their ability to seek out medical care.
- ▶ Use the Poisson Regression to model the data and see what relationships, if any, exist in between our variables.

# Methodology: The Poisson Distribution

- ▶ A probability function which is especially useful for count data
  - ▶  $Y$  is a variable with discrete outcomes  $(0, 1, 2, \dots)$  where high counts for  $Y$  are rare
  - ▶  $f(Y) = \frac{\mu^Y * e^{-\mu}}{Y!}$
  - ▶  $E[Y] = \mu$
  - ▶  $P(Y = y) = \frac{\mu^y * e^{-\mu}}{y!}$
  - ▶  $\mu$  also sometimes noted as  $\lambda$

# Methodology: Necessary Conditions for the Poisson Distribution

- ▶ The Mean and Variance of  $Y$  are the equal
  - ▶  $E[Y] = V[Y] = \mu$
- ▶ Independence
  - ▶ An event  $A$  occurring does not impact event  $B$  from occurring
- ▶ Probabilities are proportional to time
  - ▶  $A$  is twice as likely to occur in two-hour window than in one hour.
- ▶ Each observation is recorder over the same fixed period of time.
- ▶ Data Set is not overloaded with zero counts.

# Methodology: The Poisson Regression

- ▶ Poisson Regression finds estimates for the linear equation relating the log of a response variable to predictors.

- ▶  $\log(\hat{\mu}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

- ▶  $\hat{\mu} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$

- ▶ Allows us to see how a marginal change in our predictor variable impacts the estimated count of our variable.
- ▶ Coefficients found using the Maximum Likelihood Estimate process
- ▶ Coefficients evaluated for statistical significance using z-statistic and corresponding p-value
- ▶ We can use R glm function where we indicate the family to be “poisson” in order to obtain our fitted equation and coefficient estimates.

## Methodology: The Poisson vs. Other Regressions

- ▶ As mentioned, Poisson deals with discrete quantitative variables, as opposed to continuous quantitative variables (Linear Regression) or Yes/No outcomes (Logistical Regression)
- ▶ Similar to the Logistical Regression in that the coefficients are interpreted in the context of the log
- ▶ Poisson is particularly helpful when looking at specific count issues, such as traffic accidents at a particular intersection, prime staffing numbers during peak hours at a business, or number of soliders killed by a mule's kick



## Results and Conclusion: An Introduction to our Data

- ▶ Observational Units: Individuals in North Carolina
- ▶ Overall, we have chosen five predictor variables to estimate our response variable
- ▶ Response Variable:
  - ▶ mdu measures the number of doctors visits one person attends in a year: Count
- ▶ Predictor Variables used:
  - ▶ Linc denotes a person's yearly log(income): Quantitative
  - ▶ Age denotes a person's age at the time of the study: Quantitative
  - ▶ Physlim denotes if a person has any sort of physical limitation: Categorical, Binary
  - ▶ ndiseases denotes the number of diagnosed diseases a person has at the time of the study: Quantitative
  - ▶ Sex denotes the sex of the patient: Categorical, Binary (Male == 1, Female == 0)

## Results and Conclusions: An Introduction to our Data continued

count	st.dev	sample_mean	med	min	max
2000	4.56	2.87	1	0	65

- ▶  $\mu = 2.87, \sigma^2 = 4.56$ 
  - ▶ Over-disperion present

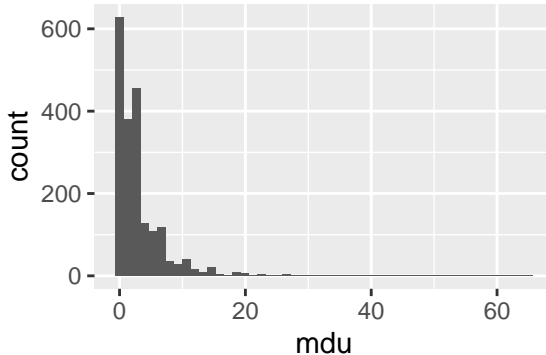


Figure 1: Count of an Individual's Doctors Visits in a Year

- ▶ Histogram shows a distribution skewed to the right
  - ▶ A high number of 0 counts in our data set

## Results and Conclusions: Another Application

- ▶ In addition to regression, we can use the poisson distribution to find probabilities of specific count occurrences or ranges of count occurrences in the data.

```
Probof4ormore = ppois(4,lambda=2.86, lower.tail=FALSE)  
Probof4ormore
```

```
## [1] 0.1617866
```

- ▶ This gives us the probability that a person will go to the doctor's office 4 or more times in a year. We can edit the lower.tail component of this code to give us the probability that a person will go to the doctor's office 4 or less times in a year

## Discussion and Critiques: Model Assumptions

- ▶ We have good reason to question a few necessary assumptions of the Poisson Distribution in regards to our data set
- ▶ Mean and Variance are not equal  $E[Y] = 2.8$  and  $V[Y] = 4.5$ 
  - ▶ Over-dispersion, which could indicate that we should use a different model
- ▶ We have a large number of observations for which the count is 0
- ▶ Consider using a different model to fit to our data
  - ▶ Zero-Inflated Poisson Regression
    - ▶ accounts for large number of 0 counts in data set
  - ▶ Zero-Inflated Negative Binomial Regression
    - ▶ helps with both larger number of 0s and over-dispersion

## Discussion and Critiques: Inference

- ▶ With our data being so significant (perhaps abnormally), we were unsure of our ability to use the fitted equation with any confidence
- ▶ With a high number of possible predictors, we could have omitted some very significant variables from the regression, which could be vital in understanding what impacts utilization of healthcare.
- ▶ Data set is from the 1970s
  - ▶ Comprehensive, but outdated
  - ▶ If using data to enact policy shifts or structural changes, our data analysis might not be “in touch” enough with the current climate of health care.
- ▶ Data is collected from N.C. only, which limits the scope in which we can make inferences.
- ▶ Perhaps some cultural/societal impacts in N.C. region impacted our data

# Conclusion

- ▶ Poisson Regression is useful for analyzing count data, but it is crucial to check the data set and see that conditions are met prior to analysis and inference
- ▶ Having a large sample size is helpful, but also must be aware of how that potentially impacts analysis.
- ▶ Working on real world data is tough and messy. Takes a lot more time and careful thought
  - ▶ No exact, clear path for analysis
- ▶ “One step forward, two steps back”

# References

<https://stats.idre.ucla.edu/r/dae/poisson-regression/>

<https://stats.idre.ucla.edu/stata/output/poisson-regression/>

<https://stats.idre.ucla.edu/r/dae/zip/> [https:](https://www.sciencedirect.com/science/article/pii/S0167629602000085)

[//www.sciencedirect.com/science/article/pii/S0167629602000085](https://www.sciencedirect.com/science/article/pii/S0167629602000085)

<https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/>

[#:~:targetText=Poisson%20Regression%20models%20are%20best,where%](#)



# R Markdown

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

## Slide with Bullets

- ▶ Bullet 1
- ▶ Bullet 2
- ▶ Bullet 3

## Slide with R Output

```
summary(cars)
```

##	speed	dist
##	Min. : 4.0	Min. : 2.00
##	1st Qu.:12.0	1st Qu.: 26.00
##	Median :15.0	Median : 36.00
##	Mean :15.4	Mean : 42.98
##	3rd Qu.:19.0	3rd Qu.: 56.00
##	Max. :25.0	Max. :120.00

# Slide with Plot

