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 it's applications?
- Can we fit the Poisson regression to our data set and use it to make any inferences about the relationships we see?

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 $(0, 1, 2, ...)$ where high counts for Y are rare
  - $f(Y) = \frac{\mu^{Y} * e^{-\mu}}{Y!}$
  - $E[Y] = \mu$
  - $P(Y = y) = \frac{\mu^{y} * e^{-\mu}}{y!}$
  - μ also sometimes noted as λ

Methodology: Necessary Conditions for the Poisson Distribution

- The Mean and Variance of Y are equal
  - $E[Y] = V[Y] = \mu$
- Independence
  - An event A occuring does not impact event B from occuring
- 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.
  - $log(\hat{\mu}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$
  - $\hat{\mu} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}$
- 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 evalued for statisical 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

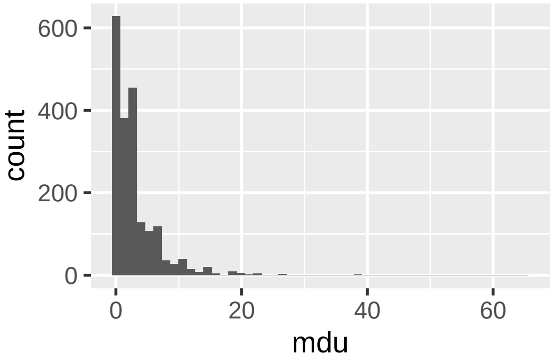

Figure 1: Count of an Individual's Doctors Visits in a Year

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

Results and Conclusions: Forward Selction:Step 1

```r
regmod1<-glm(mdu~physlim, Doc_sample, family=poisson)
regmod2<-glm(mdu~ndisease, Doc_sample, family=poisson)
regmod3<-glm(mdu~linc, Doc_sample, family=poisson)
regmod4<-glm(mdu~age, Doc_sample, family=poisson)
regmod5<-glm(mdu~sex, Doc_sample, family=poisson)
p.physlim <- summary(regmod1)$coefficients[2,4]
p.disease <- summary(regmod2)$coefficients[2,4]
p.linc <- summary(regmod3)$coefficients[2,4]
p.age <- summary(regmod4)$coefficients[2,4]
p.sex <- summary(regmod5)$coefficients[2,4]
```

Results and Conclusions: Forward Selction:Step 1

```
##                        [,1]
## p.physlim  9.880857e-95
## p.disease 2.247778e-151
## p.linc     1.925348e-25
## p.age      5.061433e-39
## p.sex      1.898671e-38
```

Results and Conclusions: Forward Selction:Step 2

```r
regmod11<-glm(mdu~ndisease+physlim, Doc_sample,
              family=poisson)
regmod12<-glm(mdu~ndisease+linc, Doc_sample,
              family=poisson)
regmod13<-glm(mdu~ndisease+age, Doc_sample,
              family=poisson)
regmod14<-glm(mdu~ndisease+sex, Doc_sample,
              family=poisson)
p.dp <- summary(regmod11)$coefficients[3,4]
p.dli <- summary(regmod12)$coefficients[3,4]
p.da <- summary(regmod13)$coefficients[3,4]
p.ds <- summary(regmod14)$coefficients[3,4]
```

```
##                [,1]
## p.dp  1.037162e-39
## p.dli 1.220258e-25
## p.da  5.939466e-10
## p.ds  9.587483e-18
```

Results and Conclusions: Forward Selction:Step 3

```
regmod21<-glm(mdu~ndisease+physlim+age,
              Doc_sample, family=poisson)
regmod22<-glm(mdu~ndisease+physlim+sex,
              Doc_sample, family=poisson)
regmod23<-glm(mdu~ndisease+physlim+linc,
              Doc_sample, family=poisson)
p.dpa <- summary(regmod21)$coefficients[4,4]
p.dps <- summary(regmod22)$coefficients[4,4]
p.dpl <- summary(regmod23)$coefficients[4,4]
```

```
##                [,1]
## p.dpa 1.367631e-09
## p.dps 2.372625e-16
## p.dpl 4.288775e-29
```

Results and Conclusions: Forward Selction:Step 4

```r
regmod31<-glm(mdu~ndisease+physlim+linc+sex,
              Doc_sample, family=poisson)
regmod32<-glm(mdu~ndisease+physlim+linc+age,
              Doc_sample, family=poisson)
p.dpls <- summary(regmod31)$coefficients[5,4]
p.dpla <- summary(regmod32)$coefficients[5,4]
```

```
##                 [,1]
## p.dpls 2.339247e-19
## p.dpla 7.851804e-09
```

Results and Conclusions: Forward Selction:Step 5

```r
regmod41<-glm(mdu~ndisease+physlim+linc+sex+age,
              Doc_sample, family=poisson)
p.dlspa <- summary(regmod41)$coefficients[6,4]
```

```
##                 [,1]
## p.dlspa 6.771765e-08
```

Results and Conclusions: Summary Output of model

```
##
## Call:
## glm(formula = mdu ~ ndisease + physlim + linc + sex + ag
##     data = Doc_sample)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1501  -1.9434  -0.8458   0.5508  16.8987
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.0282135  0.1450822  -7.087 1.37e-12 ***
## ndisease     0.0287146  0.0018776  15.293  < 2e-16 ***
## physlimTRUE  0.4349585  0.0313656  13.867  < 2e-16 ***
## linc         0.1844135  0.0159876  11.535  < 2e-16 ***
## sexmale     -0.2448111  0.0279221  -8.768  < 2e-16 ***
## age          0.0042310  0.0007839   5.397 6.77e-08 ***
## ---
```

Results and Conclusions: Final Model

- From our knowledge on Poisson Distribution, we can write out the regression model like so $log(MDU) = -1.03 + 0.029 NDisease + 0.435 PhysLim + 0.184 linc - 0.245 Male + 0.004 Age$
- Which is equivalent to $MDU = e^{-1.03+0.029 NDisease+0.435 PhysLim+0.184 linc-0.245 Male+0.004 Age}$

Results and Conclusions: Another Application

- In addition to regression, we can use the poisson distribution to find probabilities of specific count occurences or ranges of count occurences 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 probablity 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 ommitted 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:
//www.sciencedirect.com/science/article/pii/S0167629602000085
https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/
#:~:targetText=Poisson%20Regression%20models%20are%20best,where%