Lab 8: Count Models

Hao Wang

March 24, 2017

Reading Materials: R for Count Models

Interaction in Logistic Regressions

1. Load essential packages

```r
library(effects)
library(ggplot2)
library(MASS)
library(Zelig)
library(ZeligChoice)
library(pscl)
```

2. Poisson

If your DV follows a poisson distribution:

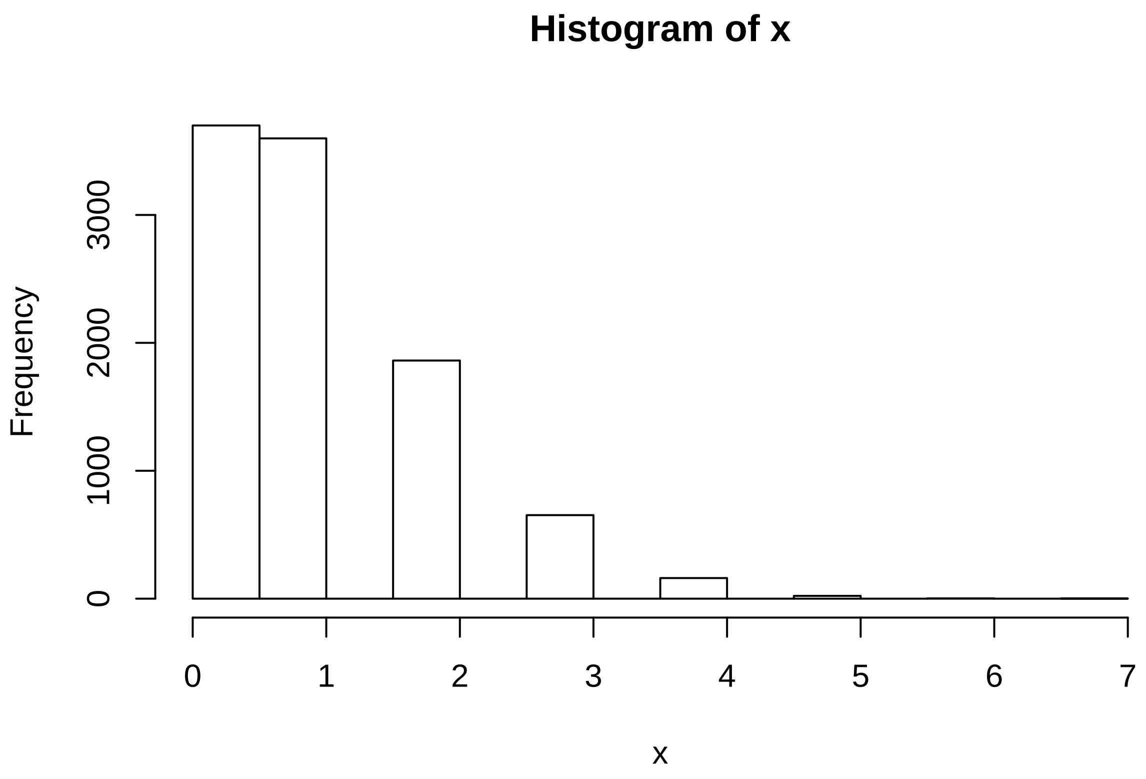
The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume. For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day.

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$x! = x(x-1)(x-2)(x-3)...2 * 1$$

λ is the mean and variance of poisson distribution. We can simulate poisson distribution in R

```r
set.seed(1)
x <- rpois(10000, 1) #1000 poisson random numbers, lambda is 1
probability <-dpois(x, 1)
hist(x)
```



2.1 Code in GLM

```
mydata<-read.csv("http://j.mp/PRESenergy")
energy.poisson<-glm(Energy~rmn1173+
     grf0175+grf575+jec477+jec1177+
     jec479+embargo+hostages+oilc+
     Approval+Unemploy,
     family=poisson(link=log),
     data=mydata)
summary(energy.poisson)
```

```
##
## Call:
## glm(formula = Energy ~ rmn1173 + grf0175 + grf575 + jec477 +
##     jec1177 + jec479 + embargo + hostages + oilc + Approval +
##     Unemploy, family = poisson(link = log), data = mydata)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -8.383   -2.994   -1.054    1.536   11.399
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.250093   0.329121  40.259  < 2e-16 ***
## rmn1173      0.694714   0.077009   9.021  < 2e-16 ***
## grf0175      0.468294   0.096169   4.870 1.12e-06 ***
## grf575      -0.130568   0.162191  -0.805 0.420806
## jec477       1.108520   0.122211   9.071  < 2e-16 ***
## jec1177      0.576779   0.155511   3.709 0.000208 ***
## jec479       1.076455   0.095066  11.323  < 2e-16 ***
## embargo      0.937796   0.051110  18.349  < 2e-16 ***
## hostages    -0.094507   0.046166  -2.047 0.040647 *
## oilc        -0.213498   0.008052 -26.515  < 2e-16 ***
## Approval    -0.034096   0.001386 -24.599  < 2e-16 ***
## Unemploy    -0.090204   0.009678  -9.321  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6009.0  on 179  degrees of freedom
## Residual deviance: 2598.8  on 168  degrees of freedom
## AIC: 3488.3
##
## Number of Fisher Scoring iterations: 5
```

2.2 Zelig

```
zpoisson <- zelig(Energy~rmn1173+
     grf0175+grf575+jec477+jec1177+
     jec479+embargo+hostages+oilc+
     Approval+Unemploy, model = "poisson", data = mydata)
```

```
## How to cite this model in Zelig:
##   R Core Team. 2007.
##   poisson: Poisson Regression for Event Count Dependent Variables
##   in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,
##   "Zelig: Everyone's Statistical Software," http://zeligproject.org/
```

```
summary(zpoisson)
```

```
## Model:
##
## Call:
## z5$zelig(formula = Energy ~ rmn1173 + grf0175 + grf575 + jec477 +
##     jec1177 + jec479 + embargo + hostages + oilc + Approval +
##     Unemploy, data = mydata)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -8.383  -2.994  -1.054   1.536  11.399
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 13.250093   0.329121  40.259  < 2e-16
## rmn1173      0.694714   0.077009   9.021  < 2e-16
## grf0175      0.468294   0.096169   4.870 1.12e-06
## grf575      -0.130568   0.162191  -0.805 0.420806
## jec477       1.108520   0.122211   9.071  < 2e-16
## jec1177      0.576779   0.155511   3.709 0.000208
## jec479       1.076455   0.095066  11.323  < 2e-16
## embargo      0.937796   0.051110  18.349  < 2e-16
## hostages    -0.094507   0.046166  -2.047 0.040647
## oilc        -0.213498   0.008052 -26.515  < 2e-16
## Approval    -0.034096   0.001386 -24.599  < 2e-16
## Unemploy    -0.090204   0.009678  -9.321  < 2e-16
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6009.0  on 179  degrees of freedom
## Residual deviance: 2598.8  on 168  degrees of freedom
## AIC: 3488.3
##
## Number of Fisher Scoring iterations: 5
##
## Next step: Use 'setx' method
```

3. Negative Binomial

Definition:

Assume Bernoulli trials — that is, (1) there are two possible outcomes, (2) the trials are independent, and (3) p, the probability of success, remains the same from trial to trial. Let X denote the number of trials until the rth success. Then, the probability mass function of X is:

$$f(x; r, p) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r$$

Remember that the Poisson distribution assumes that the mean and variance are the same. Sometimes, your data show extra variation that is greater than the mean. This situation is called over-dispersion and negative binomial regression is more flexible in that regard than Poisson regression (you could still use Poisson regression in that case but the standard errors could be biased). The negative binomial distribution has one parameter more than the Poisson regression that adjusts the variance independently from the mean. In fact, the Poisson distribution is a special case of the negative binomial distribution.

$$f(y; \mu, \theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta) * y!}\frac{\mu^y \theta^\theta}{(\mu+\theta)^{y+\theta}}$$

GLM.nb

```
energy.nb<-glm.nb(Energy~rmn1173+
     grf0175+grf575+jec477+jec1177+
     jec479+embargo+hostages+oilc+
     Approval+Unemploy,
     data=mydata)
summary(energy.nb)
```

```
##
## Call:
## glm.nb(formula = Energy ~ rmn1173 + grf0175 + grf575 + jec477 +
##     jec1177 + jec479 + embargo + hostages + oilc + Approval +
##     Unemploy, data = mydata, init.theta = 2.149960724, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7702  -0.9635  -0.2624   0.3569   2.2034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 15.299318   1.291013  11.851  < 2e-16 ***
## rmn1173      0.722292   0.752005   0.960  0.33681
## grf0175      0.288242   0.700429   0.412  0.68069
## grf575      -0.227584   0.707969  -0.321  0.74786
## jec477       0.965964   0.703611   1.373  0.16979
## jec1177      0.573210   0.702534   0.816  0.41455
## jec479       1.141528   0.694927   1.643  0.10045
## embargo      1.140854   0.350077   3.259  0.00112 **
## hostages     0.089438   0.197520   0.453  0.65069
## oilc        -0.276592   0.030104  -9.188  < 2e-16 ***
## Approval    -0.032082   0.005796  -5.536  3.1e-08 ***
## Unemploy    -0.077013   0.037630  -2.047  0.04070 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.15) family taken to be 1)
##
##     Null deviance: 393.02  on 179  degrees of freedom
## Residual deviance: 194.74  on 168  degrees of freedom
## AIC: 1526.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.150
##           Std. Err.:  0.242
##
##  2 x log-likelihood:  -1500.427
```

Zelig

```
znb <- zelig(Energy~rmn1173+
     grf0175+grf575+jec477+jec1177+
     jec479+embargo+hostages+oilc+
     Approval+Unemploy, model = "negbin", data = mydata)
```

```
## How to cite this model in Zelig:
##   William N. Venables, and Brian D. Ripley. 2008.
##   negbin: Negative Binomial Regression for Event Count Dependent Variables
##   in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,
##   "Zelig: Everyone's Statistical Software," http://zeligproject.org/
```

```
summary(znb)
```

```
## Model:
```

```
##
## Call:
## z5$zelig(formula = Energy ~ rmn1173 + grf0175 + grf575 + jec477 +
##     jec1177 + jec479 + embargo + hostages + oilc + Approval +
##     Unemploy, data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7702  -0.9635  -0.2624   0.3569   2.2034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 15.299318   1.291013  11.851  < 2e-16
## rmn1173      0.722292   0.752005   0.960  0.33681
## grf0175      0.288242   0.700429   0.412  0.68069
## grf575      -0.227584   0.707969  -0.321  0.74786
## jec477       0.965964   0.703611   1.373  0.16979
## jec1177      0.573210   0.702534   0.816  0.41455
## jec479       1.141528   0.694927   1.643  0.10045
## embargo      1.140854   0.350077   3.259  0.00112
## hostages     0.089438   0.197520   0.453  0.65069
## oilc        -0.276592   0.030104  -9.188  < 2e-16
## Approval    -0.032082   0.005796  -5.536  3.1e-08
## Unemploy    -0.077013   0.037630  -2.047  0.04070
##
## (Dispersion parameter for Negative Binomial(2.15) family taken to be 1)
##
##     Null deviance: 393.02  on 179  degrees of freedom
## Residual deviance: 194.74  on 168  degrees of freedom
## AIC: 1526.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.150
##           Std. Err.:  0.242
##
##  2 x log-likelihood:  -1500.427
## Next step: Use 'setx' method
```

4. Zero-inflated Models

check this

Zero inflated models deal with excess zero counts. They are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial. Thus, there are two sources of zeros: zeros may come from both the point mass and from the count component. For modeling the unobserved state (zero vs count), a binary model is used: in the simplest case only with an intercept but potentially containing regressors.

4.1 Example

The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish

```
zinb <- read.csv("http://www.ats.ucla.edu/stat/data/fish.csv")
zinb <- within(zinb, {
  nofish <- factor(nofish)
  livebait <- factor(livebait)
  camper <- factor(camper)
})

summary(zinb)
```

```
##  nofish  livebait camper     persons          child      
##  0:176   0: 34    0:103   Min.   :1.000   Min.   :0.000  
##  1: 74   1:216    1:147   1st Qu.:2.000   1st Qu.:0.000  
##                           Median :2.000   Median :0.000  
##                           Mean   :2.528   Mean   :0.684  
##                           3rd Qu.:4.000   3rd Qu.:1.000  
##                           Max.   :4.000   Max.   :3.000  
##        xb                  zg                count        
##  Min.   :-3.275050   Min.   :-5.6259   Min.   :  0.000  
##  1st Qu.: 0.008267   1st Qu.:-1.2527   1st Qu.:  0.000  
##  Median : 0.954550   Median : 0.6051   Median :  0.000  
##  Mean   : 0.973796   Mean   : 0.2523   Mean   :  3.296  
##  3rd Qu.: 1.963855   3rd Qu.: 1.9932   3rd Qu.:  2.000  
##  Max.   : 5.352674   Max.   : 4.2632   Max.   :149.000  
```

A zero inflated model assumes that zero outcome is due to two different processes. For instance, in the example of fishing presented here, the two processes are that a subject has gone fishing vs. not gone fishing. If not gone fishing, the only outcome possible is zero. If gone fishing, it is then a count process. The two parts of the a zero inflated model are a binary model, usually a logit model to model which of the two processes the zero outcome is associated with and a count model, in this case, a negative binomial model, to model the count process. The expected count is expressed as a combination of the two processes. We are going to use the variables child and camper to model the count in the part of negative binomial model and the variable persons in the logit part of the model.

```
m1 <- zeroinfl(count ~ child + camper | persons,
  data = zinb, dist = "negbin", EM = TRUE)
summary(m1)
```

```
## 
## Call:
## zeroinfl(formula = count ~ child + camper | persons, data = zinb, 
##     dist = "negbin", EM = TRUE)
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -0.5861 -0.4617 -0.3886 -0.1974 18.0129 
## 
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   1.3711     0.2561   5.353 8.63e-08 ***
## child        -1.5152     0.1956  -7.747 9.42e-15 ***
## camper1       0.8790     0.2693   3.264   0.0011 **
## Log(theta)   -0.9854     0.1759  -5.600 2.14e-08 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6028     0.8363   1.916   0.0553 .
## persons      -1.6663     0.6790  -2.454   0.0141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.3733
## Number of iterations in BFGS optimization: 2
## Log-likelihood: -432.9 on 6 Df
```

5. Interaction in Logistic Regressions

In linear regression we have the general interaction form:

$$\hat{Y} = \beta_0\beta_1X_1 + \beta_2X_2 + \beta_{12}X_1X_2$$

In logistic regression, we model our desired outcome as a logistic function of latent variable Y^*

$$L(Y^*) = \frac{1}{1+e^{-y^*}}$$

$$Pr(Y) = L(Y^*) = L(\beta_0\beta_1X_1 + \beta_2X_2 + \beta_{12}X_1X_2)$$
$$Pr(Y) = \frac{1}{1+e^{-(\beta_0\beta_1X_1+\beta_2X_2+\beta_{12}X_1X_2)}}$$

Assume our product term is meanful, the magnitude of interaction term X_1X_2 is the second derivative

$$\frac{\partial^2 Pr(Y)}{\partial X_1X_2} = Pr(Y)(1-Pr(Y))\beta_{12}+$$
$$Pr(Y)(1-Pr(Y))(1-2Pr(Y))(\beta_1+\beta_{12}X_2)(\beta_2+\beta_{12}X_1)$$

The magnitude depends on both X_1 and X_2!

If you REALLY want to include an interaction:

- You have theoretical expectation that there is an interaction effect between X_1 and X_2 to the latent value Y^* Note: this is not equivalent to the expected outcome!
- Compare your interaction model with the model without interaction, does that AIC decrease?
- Plot your effect on marginal probability.