Modelling Count Variables with R

Dublin R

Overview

1. Introduction to Modelling Count Variables
2. Poisson Regression
3. Negative Binomial Regression
4. Zero-Inflated Models and Vuong Tests
5. Zero Truncation

Introduction

- This presentation is about regression methods in which the dependent variable takes count (nonnegative integer) values.
- The dependent variable is usually the number of times an event occurs in a certain period of time.

- Linear regression is used to model and predict continuous measurement variables.
- Poisson regression is used to model and predict discrete count variables.

Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

Overview

Some examples of **event counts** are:

- number of claims per year on a particular car owners insurance policy,
- number of workdays missed due to sickness of a dependent in a one-year period,
- number of papers published per year by a researcher.

Poisson Distribution

- The number of persons killed by mule or horse kicks in the Prussian army per year.
- Ladislaus Bortkiewicz collected data from 20 volumes of Preussischen Statistik.
- These data were collected on 10 corps of the Prussian army in the late 1800s over the course of 20 years, giving a total of 200 observations of one corps for a one year period.
- The unit period of observation is thus one year.

Poisson Distribution: Prussian Cavalary

- The total deaths from horse kicks were 122, and the average number of deaths per year per corps was thus $122/200 = 0.61$.
- In any given year, we expect to observe, well, not exactly 0.61 deaths in one corps
- Here, then, is the classic Poisson situation: a rare event, whose average rate is small, with observations made over many small intervals of time.

Generating Random Numbers

```
> X <- rpois(200,lambda=0.61)
> X
[1]   1 2 0 1 0 3 0 0 1 0 0 4 0 0 0 1 0 1 0 2
[21]  0 0 0 2 2 0 0 0 1 0 0 0 0 1 0 0 0 1 2 0
[41]  0 0 1 0 1 0 1 0 0 1 1 0 1 0 0 1 0 0 3 1
.......
[141] 0 0 0 0 1 2 0 1 0 1 0 0 0 0 0 0 0 1 0 0
[161] 1 0 1 0 0 0 0 1 0 0 0 0 0 1 1 1 0 2 0 1
[181] 0 0 2 0 2 0 0 1 0 0 3 1 0 0 0 1 1 0 0 0
>
> mean(X) ;var(X)
[1] 0.53
[1] 0.5317588
```

Poisson Distribution Assumptions

- Poisson Regression is main technique used to model count variables.
- Assumption underlying Poisson Distribution : Mean and Variance are equal

$$\mathrm{E}(X) = \mathrm{Var}(X)$$

- Allow for a margin of error of about 5% . Simulation Studies can be used to determine the validity of this assumption. (see Next Slide)

Simulation Studies

```
> X=rpois(1000,lambda=1);mean(X);var(X)
[1] 1.001
[1] 1.028027
> X=rpois(5000,lambda=0.5);mean(X);var(X)
[1] 0.5074
[1] 0.5232499
> X=rpois(2500,lambda=0.7);mean(X);var(X)
[1] 0.7248
[1] 0.7317577
> X=rpois(500,lambda=3);mean(X);var(X)
[1] 3.076
[1] 2.851928
```

Simulation Studies

```
> Ratio = numeric()
> M = 10000
>
> for ( i in 1:M){
+       X=rpois(2500,lambda=5);
+       Ratio[i] = var(X)/mean(X)
+ }
>
> quantile(Ratio, c(0.025,0.975))
     2.5%     97.5%
0.9452617 1.0563994
```

Problem Areas

Over-Dispersion : Important Poisson Distributon assumption does not hold

$$\mathrm{E}(X) < \mathrm{Var}(X)$$

Zero-Inflation : More "Zeros" would occure than in conventional Poisson Process (This is actually "overdispersion" also, but we will treat them separately).

Zero-Truncation : Process does not allow for a "Zero" outcome.

Over-Dispersion

- Overdispersion is the presence of greater variability in a data set than would be expected based on a given simple statistical model.
- Poisson Distribution:

$$\mathrm{Var}(X) > \mathrm{E}(X)$$

Zero-Inflation

- One common cause of over-dispersion is excess zeros, which in turn are generated by an additional data generating process.
- In this situation, zero-inflated model should be considered.
- If the data generating process does not allow for any 0s (such as the number of days spent in the hospital), then a zero-truncated model may be more appropriate.

Over-Dispersion

- When there seems to be an issue of dispersion, we should first check if our model is appropriately specified, such as omitted variables and functional forms.
- For example, if we omitted the predictor variable prog in the example above, our model would seem to have a problem with over-dispersion.
- In other words, a misspecified model could present a symptom like an over-dispersion problem.

Poisson Regression with R

- Assuming that the model is correctly specified, the assumption that the conditional variance is equal to the conditional mean should be checked.
- There are several tests including the likelihood ratio test of over-dispersion parameter alpha by running the same model using negative binomial distribution.

Generalized Linear Models

The `glm()` function

- In statistics, the problem of modelling count variables is an example of generalized linear modelling.
- Generalized linear models are fit using the `glm()` function.
- The form of the `glm` function is

```
glm( modelformula,
       family=familytype(link=linkfunction),
       data=dataname)
```

Generalized Linear Models

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "$1/mu^2$")
poisson	(link = "log")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Generalized Linear Models

Texts on GLMs

- Dobson, A. J. (1990) An Introduction to Generalized Linear Models. (*London: Chapman and Hall.*)
- Hastie, T. J. and Pregibon, D. (1992) Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
- McCullagh P. and Nelder, J. A. (1989) Generalized Linear Models. (*London: Chapman and Hall.*)
- Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. *New York: Springer.*

pscl: Political Science Computational Laboratory

Author(s): Simon Jackman et al (Stanford University)

URL: http://pscl.stanford.edu/

Description

Bayesian analysis of item-response theory (IRT) models, roll call analysis; computing highest density regions; maximum likelihood estimation of zero-inflated and hurdle models for count data; goodness-of-fit measures for GLMs; data sets used in writing and teaching at the Political Science Computational Laboratory; seats-votes curves.

glm2: Fitting Generalized Linear Models

Author(s): Ian Marschner

Fits generalized linear models using the same model specification as glm in the stats package, but with a modified default fitting method that provides greater stability for models that may fail to converge using glm

VGAM: Vector Generalized Linear and Additive Models

Author(s): Thomas W. Yee (t.yee@auckland.ac.nz)

URL: http://www.stat.auckland.ac.nz/$\sim$ yee/VGAM

Vector generalized linear and additive models, and associated models (Reduced-Rank VGLMs, Quadratic RR-VGLMs, Reduced-Rank VGAMs).

This package fits many models and distribution by maximum likelihood estimation (MLE) or penalized MLE. Also fits constrained ordination models in ecology.

MA4128 - Review Questions

(i) Describe Event Counts / Count Variables.

(ii) Poisson Distribution : Asummption of Parameter Equality what is it? how to check?

(iii) State the three cases where assumption does not hold

You are not required to know anything about R implementation.

PART 2: Poisson Regression

- Poisson regression is used to model count variables.
- Poisson regression has a number of extensions useful for count models.

Poisson Regression with R

Conventional OLS regression

- Count outcome variables are sometimes log-transformed and analyzed using OLS regression.
- Many issues arise with this approach, including loss of data due to undefined values generated by taking the log of zero (which is undefined) and biased estimates.

Poisson Regression with R

If $\mathbf{x} \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form

$$\log_e(\mathrm{E}(Y \mid \mathbf{x})) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

$$\mathrm{E}(Y \mid \mathbf{x}) = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}$$

$$\mathrm{E}(Y \mid \mathbf{x}) = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}$$

$$\mathrm{E}(Y \mid \mathbf{x}) = e^{\beta_0} \times e^{\beta_1 x_1} \times \ldots \times e^{\beta_n x_n}$$

Poisson Regression : Crabs Example

The Crabs Data Set

The crabs data set is derived from Agresti *(2007, Table 3.2, pp.76-77)*. It gives 4 variables for each of 173 female horseshoe crabs.

- **Satellites** number of male partners in addition to the female's primary partner
- **Width** width of the female in centimeters
- **Dark** a binary factor indicating whether the female has dark coloring (yes or no)
- **GoodSpine** a binary factor indicating whether the female has good spine condition (yes or no)

Let the first variable be a response variable, with the other three as predictors.

Poisson Regression : Crabs Example

The data is containted in the `R` package **glm2**

```
require(glm2)

data(crabs)
head(crabs)

summary(crabs[,1:4])
```

Poisson Regression : Crabs Example

```
> head(crabs)
  Satellites   Width   Dark GoodSpine  Rep1 Rep2
1          8    28.3    no         no     2    2
2          0    22.5   yes         no     4    5
3          9    26.0    no        yes     5    6
4          0    24.8   yes         no     6    6
5          4    26.0   yes         no     6    8
....
```

Poisson Regression : Crabs Example

```
> summary(crabs[,1:4])
  Satellites          Width          Dark      GoodSpine
 Min.   : 0.000   Min.   :21.0   no :107   no :121
 1st Qu.: 0.000   1st Qu.:24.9   yes: 66   yes: 52
 Median : 2.000   Median :26.1
 Mean   : 2.919   Mean   :26.3
 3rd Qu.: 5.000   3rd Qu.:27.7
 Max.   :15.000   Max.   :33.5
```

Poisson Regression : Crabs Example

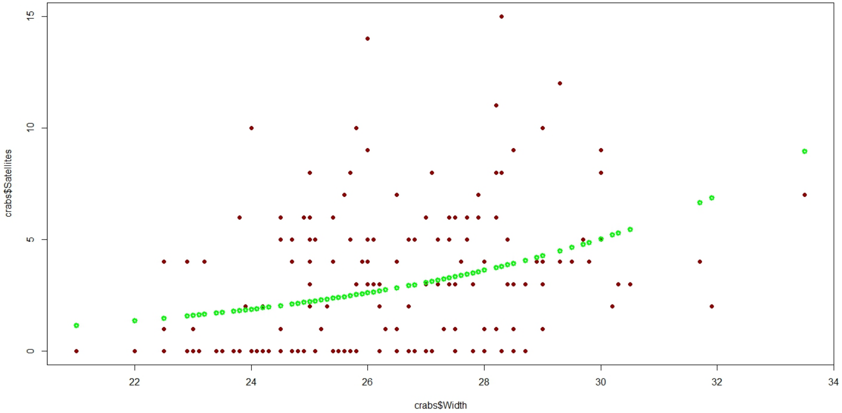
- Fit a Poisson regression model with the number of Satellites as the outcome and the width of the female as the covariate.
- What is the multiplicative change in the expected number of crabs for each additional centimeter of width?

```
crabs.pois <- glm2(Satellites ~ Width,
data=crabs, family="poisson")
summary(crabs.pois)

exp(0.164)
```

Poisson Regression : Crabs Example

```
> summary(crabs.pois)

Call:
glm2(formula = Satellites ~ Width,
family = "poisson", data = crabs)

......
......
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
Width        0.16405    0.01997   8.216  < 2e-16 ***
---
......
```

Poisson Regression : Crabs Example



Code for Crabs Data Plot

```
plot(crabs$Width, crabs$Satellites,
   pch=16, col="darkred")
   points(crabs$Width, crabs.pois$fitted.values,
   col="green", lwd=3)
```

Other Examples of Poisson regression

- The number of awards earned by students at a secondary or high school.
- Predictors of the number of awards earned include the type of program in which the student was enrolled (e.g., vocational, general or academic) and the score on their final exam in math.

Description of the data

- For the purpose of illustration, we have simulated a data set for the last example.
- The data set is called *poisreg.csv*
- In this example, **num_awards** is the outcome variable and indicates the number of awards earned by students at a high school in a year.

Predictor Variables

- **math** is a continuous predictor variable and represents students' scores on their math final exam,
- **prog** is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled.
- **prog** is coded as 1 = "*General*", 2 = "*Academic*" and 3 = "*Vocational*".

Poisson Regression with R

```
        id        num_awards            prog           math
 1      :  1   Min.   :0.00   General   : 45   Min.   :33.0
 2      :  1   1st Qu.:0.00   Academic  :105   1st Qu.:45.0
 3      :  1   Median :0.00   Vocational: 50   Median :52.0
 4      :  1   Mean   :0.63                    Mean   :52.6
 5      :  1   3rd Qu.:1.00                    3rd Qu.:59.0
 6      :  1   Max.   :6.00                    Max.   :75.0
 (Other):194
```

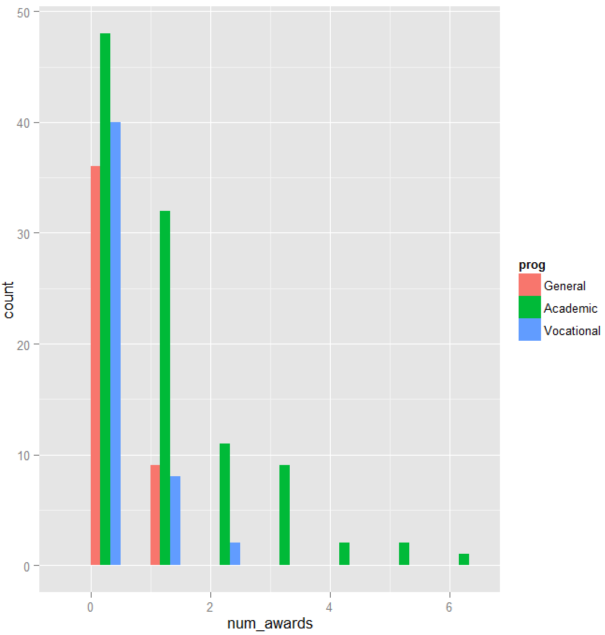

50
40
30
20
10
0
count
0
2
4
6
num_awards
prog
General
Academic
Vocational

Figure:

Poisson Regression with R

- Each variable has 200 valid observations and their distributions seem quite reasonable.
- The mean and variance of our outcome variable are more or less the same.
- Our model assumes that these values, conditioned on the predictor variables, will be equal (or at least roughly so).

Poisson regression

- At this point, we are ready to perform our Poisson regression model analysis using the `glm()` function.
- We fit the model and save it in the object `model1` and get a summary of the model.

Poisson Regression with R

```
model1 <- glm(num_awards ~ prog + math,
family="poisson", data=poisreg)

summary(model1)
```

Poisson Regression with R

```
Call:
glm(formula = num_awards ~ prog + math,
      family = "poisson",
      data = poisreg)

Deviance Residuals:
Min       1Q   Median      3Q      Max
-2.204  -0.844  -0.511   0.256   2.680
```

Poisson Regression with R

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.2471     0.6585   -7.97  1.6e-15 ***
progAcademic     1.0839     0.3583    3.03   0.0025 **
progVocational   0.3698     0.4411    0.84   0.4018
math             0.0702     0.0106    6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Poisson Regression with R

```
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

Poisson Regression with R

Regression Coefficients

- Intercept $\beta_0 = -5.2471$
- progAcademic $\beta_1 = 1.0839$
- progVocational $\beta_2 = 0.3698$
- math $\beta_3 = 0.0702$

Exercise

Predict number of awards for Vocational Student with a maths mark of 70.

$$\hat{Y} = e^{-5.2471} \times e^{1.0839 \times 0} \times e^{0.3698 \times 1} \times e^{0.0702 \times 70} = e^{0.0367} = 1.0373$$

(i) Based on `R` output, be able to carry out calculations similar to that in previous slide.

Poisson Regression with R

glm function output

- The output begins with echoing the function call. Then the information on deviance residuals is displayed.
- Deviance residuals are approximately normally distributed if the model is specified correctly.
- Here it shows a little bit of skeweness since median is not quite zero.

glm function output

- The Poisson regression coefficients for each of the variables along with the standard errors, z-scores, p-values and 95% confidence intervals for the coefficients.
- The coefficient for math is 0.07.
- This means that the expected log count for a one-unit increase in math is 0.07.

Poisson Regression with R

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.2471     0.6585   -7.97  1.6e-15 ***
progAcademic      1.0839     0.3583    3.03   0.0025 **
progVocational    0.3698     0.4411    0.84   0.4018
math              0.0702     0.0106    6.62  3.6e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

glm function output

- The indicator variable **progAcademic** compares between **prog** = **Academic** and **prog** = **"General"** , the expected log count for **prog** = **Academic** increases by about 1.1.
- The indicator variable **prog.Vocational** is the expected difference in log count (≈ 0.37) between **prog** = **"Vocational"** and the reference group (**prog** = **"General"**).

Poisson Regression with R

- The output above indicates that the incident rate for **prog = “Academic”** is 2.96 times the incident rate for the reference group (**prog = "General"**).
- Likewise, the incident rate for **prog = “Vocational”** is 1.45 times the incident rate for the reference group holding the other variables at constant.

Poisson Regression with R

- The percent change in the incident rate of **num_awards** is by 7% for every unit increase in math.

Deviance

- In statistics, deviance is a quality of fit statistic for a model that is often used for statistical hypothesis testing.
- It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood.

glm function output

- The information on deviance is also provided.
- We can use the residual deviance to perform a goodness of fit test for the overall model.

Poisson Regression with R

glm function output

- The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed.
- Therefore, if the residual difference is small enough, the goodness of fit test will not be significant, indicating that the model fits the data.

Poisson Regression with R

glm function output

- If the test had been statistically significant, it would indicate that the data do not fit the model well.
- We could try to determine if there are omitted predictor variables, if our linearity assumption holds and/or if there is an issue of over-dispersion.

Comparing Candidate Models

- We can also test the overall effect of prog by comparing the deviance of the full model with the deviance of the model excluding prog.
- The two degree-of-freedom chi-square test indicates that prog, taken together, is a statistically significant predictor of **num_awards**.

Comparing Models

```
# update m1 model dropping prog
m2 <- update(m1, . ~ . - prog)

# test model differences with chi square test
anova(m2, m1, test="Chisq")
```

Poisson Regression with R

```
Analysis of Deviance Table

Model 1: num_awards ~ math
Model 2: num_awards ~ prog + math
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       198        204
2       196        189  2     14.6  0.00069 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 
```

Incident Rate Ratios

- Sometimes, we might want to present the regression results as **incident rate ratios** (IRRs) and their standard errors, together with the confidence interval.
- To compute the standard error for the incident rate ratios, we will use the **Delta method** (Numerical Computation Method).
- To this end, we make use the function `deltamethod` implemented in R package **msm**.

Incident Rates

Incidence rate is the occurrence of an event over person-time, for example person-years.

$$\text{Incidence Rate} = \frac{\text{events}}{\text{Person Time}}$$

Note: the same time intervals must be used for both incidence rates.

Incident Rate Ratios

A **rate ratio** (sometimes called an incidence density ratio) in epidemiology, is a relative difference measure used to compare the incidence rates of events occurring at any given point in time.

$$\text{Incidence Rate Ratio} = \frac{\text{Incidence Rate 1}}{\text{Incidence Rate 2}}$$

Poisson Regression with R

Delta Method

```
s <- deltamethod(list(~ exp(x1), ~ exp(x2), ~ exp(x3),
      ~ exp(x4)), coef(m1), cov.m1)

#exponentiate old estimates dropping the p values

rexp.est <- exp(r.est[, -3])

# replace SEs with estimates
# for exponentiated coefficients

rexp.est[, "Robust SE"] <- s
```

Poisson Regression with R

```
rexp.est

                Estimate Robust SE       LL      UL
 (Intercept)    0.005263   0.00340 0.001484 0.01867
 progAcademic   2.956065   0.94904 1.575551 5.54620
 progVocational 1.447458   0.57959 0.660335 3.17284
 math           1.072672   0.01119 1.050955 1.09484
```

Poisson Regression with R

- Sometimes, we might want to look at the expected marginal means.
- For example, what are the expected counts for each program type holding math score at its overall mean?
- To answer this question, we can make use of the predict function.
- First off, we will make a small data set to apply the predict function to it.

Poisson Regression with R

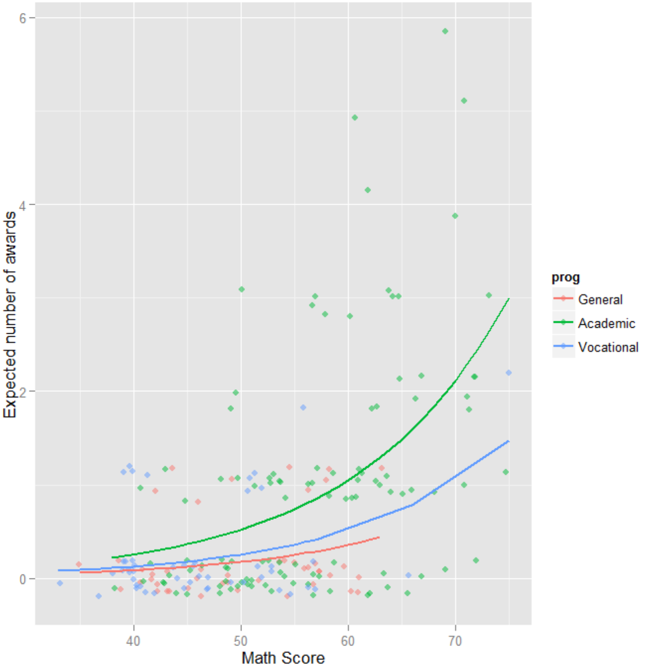
```
(s1 <- data.frame(math = mean(p$math),
  prog = factor(1:3, levels = 1:3,
  labels = levels(p$prog))))

   math       prog
1 52.65    General
2 52.65   Academic
3 52.65 Vocational
```

Poisson Regression with R

```
predict(m1, s1, type="response", se.fit=TRUE)

 $fit
      1      2      3
 0.2114 0.6249 0.3060

 $se.fit
       1       2       3
 0.07050 0.08628 0.08834

 $residual.scale
 [1] 1
```

Poisson Regression with R

- In the output above, we see that the predicted number of events for level 1 of prog is about 0.21, holding math at its mean.
- The predicted number of events for level 2 of prog is higher at 0.62, and the predicted number of events for level 3 of prog is about .31.
- The ratios of these predicted counts ($\frac{0.625}{0.211} = 2.96$, $\frac{0.306}{0.211} = 1.45$) match what we saw looking at the IRR.



Expected number of awards
Math Score
prog
General
Academic
Vocational
0
2
4
6
40
50
60
70