Introduction

- This talks 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.

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 Regression with R

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 period or module 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.

```
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)
[1] 0.53
> var(X)
[1] 0.5317588
```

Overview

- ▶ Poisson regression is main technique used to model count variables.
- ▶ Poisson Distribution : Mean and Variance are equal

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

- ▶ Sometimes conventional Poisson Regression is not an appropriate technique, and alternative or variant techniques are used instead.
- ▶ For example, Negative Binomial regression is for modelling "over-dispersed" count outcome variables.

Problem Areas

Over-Dispersion Important Poisson Distributon 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.

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(formula, 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")

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/∼ 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.