# STAT3015/7030:
# Generalised Linear Modelling
## Poisson GLM

Bronwyn Loong

Semester 2 2014

# Reference

Faraway, Ch 3

Ramsey and Schafer, Ch 22

Gelman and Hill, Ch 6.2

# Case Study - Galapagos Data

The Galapagos islands lie in the Pacific Ocean, off the west coast of South America. Despite the physical isolation of the islands, environmental threats exist from introduced species. A study was undertaken to count the number of plant species across 30 Galapagos islands, and the number which are endemic to the island.

Scientists would like to understand what geographical features of the island are associated with the number of species.

Q: What type of statistical model can we build to answer the research question??

# Case Study - Galapagos Data

Q: How would you describe the distribution of your response variable?

# Case Study - Galapagos Data

Q: How would you describe the distribution of your response variable?

- The number of species is a 'count' (positive integer)
- No definite upper bound (unlike binomial data)

Solution: Fit a Poisson GLM to model the count data

# Poisson regression

If $Y$ is Poisson with mean $\mu > 0$, then:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!} \quad y = 0, 1, 2, ...$$

$E[Y] = \mu$; $Var[Y] = \mu$
Examples of Poisson distributed random variables:

- Rare events - eg incidence of rare forms of cancer, the number of people affected is a small proportion of the population in a given geographical area.
- Occurrence of an event in a given time interval - where the probability of occurrence in a given time interval is proportional to the length of that time interval and independent of the occurrence of other events. For example - earthquakes, number of incoming telephone calls into a service center.

# Poisson or Binomial

- If the count is some number out of a fixed total, then the response might be more appropriately modelled as binomial.
- However, for small successes and large fixed totals, the Poisson is a good approximation.
- For $Y \sim Bin(n, p)$, for large $n$ and small $p$, we can show that $P(Y = y) \doteqdot e^{-\mu} \frac{(\mu)^y}{y!}$ where $\mu = np$. And logit(p) $\approx \log p$
- Also note that if the count is the number falling into some level of a given category, a multinomial response model is more appropriate.
- Also, if all counts are very big (that is, the mean of the Poisson distribution is very large), then a normal distribution assumption is a good approximation

# Poisson regression

Also note that the sum of $n$ independent Poisson random variables is also Poisson.

Suppose $Y_i \sim Pois(\mu_i)$, for $i = 1, 2.., n$, and are independent. Then

$$\sum_{i=1}^{n} Y_i \sim Pois\left(\sum_{i=1}^{n} \mu_i\right)$$

Why is this a useful result for Poisson GLM?

# Poisson regression

Q: What is the structure of our model?

Goal: build a model for the count responses $Y_i$ in terms of some predictors $x_i$. If $Y_i \sim Pois(\mu_i)$, the GLM will model some function of the mean response $g(\mu_i)$ as a linear combination of the predictors $\eta_i = x_i^T \beta$.

Choice of function $g$??

# Poisson regression - canonical link

$$f(y|\theta,\phi) = e^{\mu}\frac{\mu^y}{y!} = \exp(y\log(\mu) - \mu - \log y!)$$

Hence, $\theta = \log\mu$. That is, the canonical link is the log function. So we have

$$\log\mu_i = x_i^T\beta$$

How to obtain parameter estimates $\beta$?

# Poisson regression - likelihood

$$\log L(\beta) = \sum_{i=1}^{n} (y_i x_i^T \beta - \exp(x_i^T \beta) - \log y_i!)$$

Differentiate with respect to $\beta_j$ gives the MLE as the solution to:

$$\sum_{i=1}^{n} (y_i - \exp(x_i^T \hat{\beta})) x_{ij} = 0 \ \forall j$$

$$X^T y = X^T \hat{\mu}$$

$\rightarrow$ no closed form solution - use IRWLS.

# Deviance Goodness of Fit Test

(see R code) Is the model a good fit?

# Deviance Goodness of Fit Test

(see R code) Is the model a good fit? $\rightarrow$ Deviance goodness of fit test

**Deviance Statistic = Sum of Squared Deviance Residuals**
Approximate p-value = $\Pr(\chi^2_{n-p} >$ Deviance Statistic$)$

For the Poisson regression

$$D = \sum_{i=1}^{n} 2(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i))$$

The $\chi^2_{n-p}$ approximation is valid when the Poisson means are large. If a substantial proportion of the cases have estimated means less than 5, the approximation is questionable.

# Case Study - Galapagos Data

(see R code)

- Goodness of fit test
- Diagnostic checks - outliers, variance function, link function, choice of predictors

# Poisson GLM - OVERDISPERSION

The Poisson distribution is one of many distributions that assign probabilities to outcomes for count data. We prefer it to count the number of events in a fixed interval (time/space) if these events occur at a known average rate and independently of the time since the last event.

Recall under the Poisson model $Y \sim Pois(\lambda)$,
$E(Y) = Var(Y) = \lambda$. The variance and the mean parameter are the same.

Sometimes however, the data may exhibit more variation in the response than is predicted by the Poisson model (eg clustering of events).

We want to account for this extra variation somehow in how our model - introduce a dispersion parameter to account for overdispersion (or sometimes underdispersion).

# Poisson GLM - with dispersion parameter

$$E[Y] = \lambda = \mu$$

$$Var[Y] = \phi E[Y] = \phi\mu$$

$$\log(\mu_i) = \beta_0 + x_{1i}\beta_1 + ..... + x_{pi}\beta_p$$

The extra parameter is the dispersion parameter $\phi$.

- $\phi = 1$ - regular Poisson regression
- $\phi > 1$ - overdispersion
- $\phi < 1$ - underdispersion

# Poisson GLM - OVERDISPERSION

How to test for overdispersion - indicative checks (or things to watch out for)

- Consider the situation - are important explanatory variables not available? - Rejection of goodness of fit test
- Are the events making up the count clustered or spaced unevenly through time?
- Compare sample variance to sample averages computed for groups of responses with identical explanatory variable values (should be approximately equal).
- Too many outliers

# Poisson GLM - OVERDISPERSION

How to test for overdispersion - a numerical check

Define Pearson residual

$$z_i = \frac{y_i - \hat{\mu}_i}{sd(\hat{\mu}_i)}$$

If the Poisson model is true, the $z_i$'s should be approximately independent each with mean 0 and standard deviation 1.

If there is overdispersion, then we expect the $z_i$'s to be larger in absolute value, reflecting the extra variation beyond what is predicted under the Poisson model.

estimated overdispersion $\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} z_i^2$.

Compare $\sum_{i=1}^{n} z_i^2$ to the quantiles of a $\chi^2_{n-p}$ distribution.

# Poisson GLM - adjusting inferences for overdispersion

Multiply all standard errors by $\sqrt{\hat{\phi}}$.

OR

Fit the model using the `quasipoisson` family. Why quasi?
Because we do not fix the variance function $V(\mu) = \mu$.
Notes: because we have estimated an extra parameter $\phi$

- The drop in deviance test statistics is divided by $\hat{\phi}$ to form an F-statistic.
- Use the t-distribution as the reference distribution for tests of significance of individual coefficient estimates.

(see R code)

# Poisson GLM - Rate models

The number of events observed may depend on a size variable that determines the number of opportunities for the events to occur. Examples - the number of burglaries in a city depends on the number of dwellings; the number of customers served depends on the time interval.

We could consider a binomial model, but the size variable may not be known exactly, and for small counts, the Poisson approximation is appropriate. Furthermore, if time is the size variable, this is not a count variable (more like a measure of *exposure*)

$$y_i \sim Pois(u_i \lambda_i)$$

where $u_i$ is the exposure variable. Then $\log(u_i)$ is called the *offset*.

# Poisson GLM - Rate models

We want a model for $\lambda_i$ - the mean rate per unit of exposure.

The observed count data we use to fit the model $y_i$ depends on the exposure variable $u_i$ (or in other words, the count variable is observed after exposure to different values of variable $u_i$). That is, $E[Y_i] = \mu_i = \lambda_i u_i$. (note: $u_i$ is a fixed known quantity - doesn't require estimation).

Model:

$$\log(\lambda_i) = \log\left(\frac{\mu_i}{u_i}\right) = X_i^T \beta$$

$$\log(\mu_i) = \log u_i + X_i^T \beta$$

If we fix the coefficient $u_i$ to 1, this is known as the *offset* term.

# Case Study - Gamma radiation

Data

- *ca*: number of chromosonal abonormalities
- *cells*: number of cells exposed to gamma radiation
- *doseamt*: dose amount
- *doserate*: rate at which dose is applied

$$\log(ca/cell) = X\beta$$

$$\log(ca) = \log(cells) + X\beta$$

(see R code)

# Excess Zeroes

What if there were more zeroes in your data than a Poisson regression would predict?