Lecture 10: GLMs: Poisson Regression, Overdispersion

Author: Nick Reich / Transcribed by Daveed Goldenberg, edited by Josh Nugent

Course: Categorical Data Analysis (BIOSTATS 743)

Poisson GLMs

- Imagine you have count data following the Poisson distribution:

$$Y_i \sim Poisson(\lambda_i)$$

- Y_i is the total count in the time interval, λ_i is $E(Y_i)$, that is, the risk/rate of occurrence in some time interval,
- We use a log link for our GLM:

$$\eta_i = X_i\beta = \log \lambda_i = g(\lambda_i) = g(E[y_i])$$

Poisson GLMs

- Key Points:
- Log link implies multiplicative effect of covariates

$$log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$\lambda = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2}$$

- Relative risk is the interpretation for e^{β}

$$\begin{aligned} log(\lambda_i | X_1 = k+1, X_2 = c) &= \beta_0 + \beta_1(k+1) + \beta_2(c) \\ -log(\lambda_i | X_1 = k, X_2 = c) &= \beta_0 + \beta_1(k) + \beta_2(c) \\ log((\lambda_i | X_1 = k+1, X_2 = c)/\lambda_i | X_1 = k+1, X_2 = c) &= \beta_1 \end{aligned}$$

Exposure / Offset Term

Often Poisson models have an 'exposure' or 'offset' term, representing a demoninator of some kind. Examples: Let u_i be offset for Y_i...

- Disease incidence: Y_i = the number of cases of flu in a population in 1 year (in location i), u_i = population size
- Accident rates: Y_i = the number of traffic accidents at site i in 1 day, u_i = average number of vehicles travelling through site i in 1 day, or u_i = the number of vehicles through site i yesterday
- The offset is used to scale the Y_i

Exposure / Offset Term

$$
\begin{aligned}
Y_i &\sim Poisson(u_i * \lambda_i) \\
E(Y_i) &= u_i * \lambda_i \\
log(E(Y_i)) &= log(u_i) + log(\lambda_i)
\end{aligned}
$$

- $log(u_i)$ is our offset (from observed data, can be thought of as an intercept)
- $\log(\lambda_i)$ is our η_i (the linear predictor)

Exposure / Offset Term

- In R, the Poisson glm can be specified with an offset
- glm(Y $\sim X_1 + X_2$, family $=$ 'poison', offset $=$ log(u), data ...)
- the log is important in order to get the correct offset
- The offset term is adding more information to the model but not estimating a coefficient

Exposure / Offset Term

$$Y_i \sim Poisson(u_i * \lambda_i)$$

- The Y_i could be cases per day
- u_i could be population (persons)
- then λ_i would be cases per day per population (persons)
- which makes this a rate for an individual

$$\begin{aligned} log(E(Y_i)) - log(u_i) &= log(\lambda_i) \\ log(E(Y_i)/u_i) &= log(\lambda_i) \end{aligned}$$

Overdispersion

- In Poisson models for $Y_i \sim Poisson(\lambda_i)$

$$Var(Y_i) = \lambda_i$$

- In GLM estimation notation

$$\mu_i = E(\lambda_i)$$
$$Var(\mu_i) = \lambda_i$$

- In an overdispersed model, the variance is higher because of some variability not captured by Poisson

$$Var(\mu_i) = \phi\lambda_i$$
$$\phi > 0$$

- Overdispersion implies $\phi > 1$

Overdispersion

- Likelihood equations for Poisson GLM

$$\sum_{i=1}^{N} \frac{(y_i - u_i)x_{ij}}{Var(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

$$j = 0, .., p$$

- Depends on the distribution of Y through μ_i and Var(μ)
- ϕ drops out of the likelihood equations - this makes sense; variability won't affect the MLE - that is, βs are identical for models with $\phi > 1$ and $\phi = 1$
- However, ϕ does impact estimated standard errors

Overdispersion

$$w_i = (\frac{\partial u_i}{\partial \eta_i})^2 / Var(Y_i)$$
$$cov(\hat{\beta}) = (X^T WX)^{-1} = \phi cov(\hat{\beta})$$

- ϕ does not affect the βs but it does affect their covariance as a scaling factor

Is Overdispersion Term Needed in a Model?

- (See example 4.7.4 in Agresti)
- Start with standardized residuals
- Assume:

$$
\begin{aligned}
z_i &= \frac{y_i - \hat{y}_i}{\sqrt{Var(\hat{y}_i)}} \\
&= \frac{y_i - \mu_i}{\sqrt{\mu_i}} \sim N(0,1) \\
\sum_{i=1}^{n} z_i^2 &\sim \chi^2_{n-k}
\end{aligned}
$$

- where k is the number of parameters
- if the sum of z_i^2 is large (compare to chi-squared), we may need an overdispersion term ϕ

Is Overdispersion Term Needed in a Model?

$$\hat{\phi} = \frac{\sum_{i=1}^{n} z_i^2}{n-k}$$

- summarizes overdispersion in data compared to the fitted model
- if $\phi^2 > 1$, we should use the "quasipoisson" family in R's glm() function
- The SEs of a quasipoisson model are equivalent to the SEs of the Poisson model miultiplied by $\sqrt{\hat{\phi}}$

$$SE_{qp}(\hat{\beta}) = SE_p(\hat{\beta}) * \sqrt{\hat{\phi}}$$