STAT 528 - Advanced Regression Analysis II

Count response regression (part I)

Daniel J. Eck
Department of Statistics
University of Illinois

Last time

- basic diagnostics for binary response models
- probit regression and threshold modeling
- basic causal inference

Learning Objectives Today

- Poisson regression

Background

We suppose that we have a sample of data (y_i, x_i), $i = 1, \ldots, n$ where

- y_i is a scalar response variable
- x_i is a vector of predictors.

Recall from the exponential family notes that the log likelihood of the exponential family is of the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{1}$$

where

- $y \in \mathbb{R}^n$ is a vector statistic having components
- $\theta \in \mathbb{R}^n$ is the canonical parameter vector.

In those notes θ is unconstrained and the likelihood (1) corresponds to a saturated regression model, one parameter for every observation.

A canonical linear submodel of an exponential family is a submodel having parameterization

$$\theta = M\beta,$$

and log likelihood

$$l(\beta) = \langle M'y, \beta \rangle - c(M\beta). \tag{2}$$

In an exponential family GLM, the saturated model canonical parameter vector θ is "linked" to the saturated model mean value parameter vector through the change-of-parameter mappings $g(\theta)$.

We can write

$$\mu = \mathrm{E}_\theta(Y) = g(M\beta)$$

which implies that we can write

$$g^{-1}\left(\mathrm{E}_\theta(Y)\right) = M\beta.$$

Poisson regression model

The Poisson regression model [and its variants] is one of the more widely used and studied exponential family GLMs in practice.

The Poisson regression model is used for analyzing a count response variable, $y_i \in \{0, 1, 2, 3, \ldots\}$.

The Poisson regression model allows for users to model the rate as a function of covariates.

For a count response variable Y and a vector of predictors X, let $\mu(x) = \mathsf{E}(Y|X = x)$. The Poisson regression model is then

$$\mu(x) = \mathsf{E}(Y|X = x) = \exp\left(x'\beta\right). \tag{3}$$

Equivalently,

$$\log\left(\mu(x)\right) = x'\beta.$$

In vector notation, we can express the above as

$$\boldsymbol{\mu} = \exp(M\beta) \quad \text{and} \quad \log(\boldsymbol{\mu}) = M\beta$$

where the above $\exp(\cdot)$ and $\log(\cdot)$ operations are understood as componentwise operations.

Let's consider the log likelihood of a sample of independent Poisson random variables

$$\sum_{i=1}^{n} y_i \log(\mu_i) - \mu_i = \sum_{i=1}^{n} y_i \theta_i - \exp(\theta_i)$$

where

$$\theta_i = \log(\mu_i) \qquad \text{and} \qquad \mu_i = \exp(\theta_i) = g(\theta_i).$$

We see that the Poisson regression model with log link is the same as the canonical linear submodel of an exponential family.

The link function g^{-1} is the logarithmic function. Hence the name log-linear models.