Poisson Regression(Loglinear Model) and Generalized Linear Model Framework

Wei Wang

Dec 9, 2010

Count Data

- Other than 0-1 binary data, count data is another discrete data type that we deal with everyday.
- number of traffic accidents, number of epidemic incidences etc.
- Again we need to link predictors X and response y.

Binomial Regression (Extension of Logistic Regression)

- If response y_i can be naturally interprested as number of success in n_i Bernoulli experiments, then we can still use logistics model.

- We only need to treat each observation as n_i data points, with y_i 1's and $(n_i - y_i)$ 0's.

Example

- We study the proportion of death penalty that were overturned in 34 states in 2000. For each state, n_i is the total number of death sentences in 2000 and y_i is the number of death penalty cases that were overturned.

- Very naturally, logistic regression is a good model for this study.

Poisson Model

- But in a lot of cases, a binomial model is not appropriate.
- No natural explanation of success/failure rate; no natural limit of maximum number of incidences.
- A more flexible way to model count data is Poisson distribution.

$$P(N = n) = \exp(-\lambda)\frac{\lambda^n}{n!}$$

- The parameter (mean) of Poisson distribution λ is positive, so we need a transformation that maps $x^T\beta$ to $[0, +\infty)$.

Poisson Regression(Loglinear Model)

- An exponential transformation is a natural choice.
- Formally, a Poisson Regression (Loglinear Model) is given by

$$\begin{aligned} Ey_i &= \lambda_i \\ \lambda_i &= \exp(x_i^T \beta) \end{aligned}$$

- Similar to logistic regression, there is no variance component in Poisson Regression.
- But unlike logistic regression, we will see that absence of variance parameter could cause us trouble in model checking.

Example: What cause traffic accidents?

- We study the number of traffic accidents at a group of street intersections. Intersections are indexed by i. y_i is the number of traffic accidents in a particular year. For predictors, we have X_1, the average speed of vehicles at that intersection, and X_2, the indicator of whether the intersection has a traffic signal.

- With standard statistical software, we have the fitted model

$$y_i \sim \text{Poisson}(\exp(2.8 + 0.012X_{i1} - 0.20X_{i2}))$$

- At first sight, the signs of the coefficient estimates make sense.

Interpretation of the Coefficients

- The slope of X_1 is the expected difference in y (on the log scale) for each additional mile-per-hour increase of the average speed of vehicles. The multiplicative increase is an $e^{0.012} - 1 = 1.012 - 1 = 1.2\%$ positive difference in the rate of traffic accidents.

- The parameter of X_2 tells us the expected difference in y (on the log scale) if the intersection has a traffic signal. The multiplicative decrease is an $1 - e^{0.20} = 18\%$ in the rate of traffic accidents.

Offset

- In most application of Poisson Regression, we want to interpret the count relative to some baseline or "exposure".
- In the traffic accidents example, it is natural to think that the rate of traffic accidents at one particular intersection should be proportional to the total number of vehicles that passed that intersection, which we denote as u_i.
- So the Poisson Regression should be expressed as

$$y_i \sim \text{Poisson}(u_i \exp(x_i^T \beta))$$

- $\log(u_i)$ is called offset in Poisson Regression terminology.

Overdispersion

- Another popular model for binary response data is Probit Model.

- In Probit Model, we choose a different mapping from $\mathbb{R}$ to $[0,1]$ unit interval: the CDF of standard normal.

- If ξ is a standard normal random variable and Φ is its CDF,

$$p = \text{probit}(x^T\beta) = P(\xi + x^T\beta > 0) = \Phi(-x^T\beta)$$

Negative Binomial Regression

- We can view probit model in a Latent Variable Formulation. Each dicrete outcome y_i is associated with a continuous, unobserved z_i

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= x_i^T \beta + \xi_i \end{aligned}$$

  ξ_i are i.i.d. standard normal random variables.

- In fact, we can incorporate logistic regression model into this formulation if we let ξ_i follows logistic distribution, which is defined by

$$P(\xi_i < x) = \text{logit}^{-1}(x)$$

Zero-Inflated Model

- A lot of statistical problems are easier to understand and tackle from a Latent Variable Formulation point of view.
- In practice, logistic regression and probit regression are not very different. In fact, logistic distribution is very close to normal distribution with mean 0 and standard deviation 1.6.
- So to use logistic or probit regression is mostly just a matter of taste.

Estimation

- Unlike simple linear regression, no explicit maximum likelihood estimators can be given for logistic regression.
- Iterative Weighted Least Squares scheme is used to estimate the parameters. In some literature, it is also called Fisher's Scoring Method.
-

We don't delve into details here.

Model Checking: Error Rate

- Another thing that we might want to look at is Error Rate.
- If we do deterministic prediction, i.e. guessing $y_i = 1$ when fitted value $\text{logit}^{-1}(x_i^T \hat{\beta}) > .5$ and $y_i = 0$ if otherwise, then we define the proportion of wrong prediciton as Error Rate.
- Error Rate need to be at least less than .5(the error rate when we just randomly guess).

Model Checking: Deviance

- Deviance is defined as -2 times the logarithm of the likelihood function. The better the model fit, the lower the deviance is.
- If a predictor that is random noise is added to the model, we expect deviance to decrease on average by 1.
- If a predictor that is informative is added to the model, we expect deviance to decrease on average by more than 1. The larger the increase, the more relevant the predictor is for our model.
- Deviance is a fundmental concept in model comparison and selection.

Model Checking: Deviance

- For the voting-income example, the null model (only has intercept) has a deviance of 1591.
- Added income as a predictor, the model has a deviance of 1557.
- This shows that income is informative for predicting vote outcome.

Software Implementations

- R: `glm`
- Matlab: `glmfit`