Count Response, Categorical Response and Generalized Linear Model Framework

Wei Wang

Dec 9, 2010

Outline

- Count Response
- Categorical Response
- Generalized Linear Models

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 logistic 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

$$P(y_i = k) = \exp(-\exp(X_i^T \beta)) \frac{(\exp(X_i^T \beta))^k}{k!}$$

- 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 fitting.

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.

Fitting the Model

- In R, it looks like this

```
glm(accident ~ ave_speed +
traf_light,family=poisson,
offset=log(num_vehicle))
```

Fitting the Model

- In R, it looks like this

```
glm(accident ~ ave_speed +
traf_light,family=poisson,
offset=log(num_vehicle))
```

- In Matlab, I don't really know how it looks like.

Overdispersion

- For Poisson distribution, its variance is equal to its expecation.
- But with real data, this requirement is often violated. In most of the cases, we have larger variance than expected if Poisson distribution really holds. It is called overdispersion.
- It almost always happens with count data.
- We can see the tradeoff between the parismony and flexibility of the model.

Detecting Overdispersion

- Under Poisson distribution, the standardized residuals

$$z_i = \frac{y_t - \hat{y}_t}{sd\{y_t\}} = \frac{y_t - \hat{y}_t}{\sqrt{\hat{y}_t}}$$

  should be approximately independently distributed and have mean 0 and sd 1.

- The sum of square of z_i follows a χ^2_{n-p} distribution, where p is the number of predictors (including intercept).

$$\sum_{i=1}^{n} z_i^2 \to \chi^2_{n-p}$$

- In the case of overdispersion, all the z_i's are larger than 1 in absolute values and the sum of square should be much larger than $n - p$.

Dealing with Overdispersion

- As we mentioned before, the problem of overdispersion is caused by the lack of variance parameter in Poisson distribution.
- So the solution is to loosen the Poisson distribution to Quasi Poisson distribution, which includes an overdispersion parameter ω, which is the ratio of variance and mean of the distribution.
- `glm(accident ~ ave_speed + traf_light, family=quasipoisson, offset=log(num_vehicle))`
- Overdispersed-Poisson is a class of distribution, and we commonly use Negative-Binomial distribution.

Categorical Response

- Categorical Response can be either ordered (ordinal) or unordered (nominal).
- Examples of ordered categorical response include Democrat, Independent, Republican; Yes, Maybe, No; Always, Frequently, Often, Rarely, Never.
- Examples of unordered categorical response include Football, Basketball, Baseball, Hockey; Train, Bus, Automobile, Walk; White, Black, Hispanic, Asian, Other.

Ordered Categorical: multinomial logit regression

- Outcome y can take on values $1, 2, \ldots, K$, then the multinomial logit regression is given by

$$\begin{aligned} P(y > 1) &= \text{logit}^{-1}(X\beta) \\ P(y > 2) &= \text{logit}^{-1}(X\beta - c_2) \\ P(y > 3) &= \text{logit}^{-1}(X\beta - c_3) \\ &\ldots \\ P(y > K-1) &= \text{logit}^{-1}(X\beta - c_{K-1}) \end{aligned}$$

- Equivalently, the model can be given by

$$P(y = k) = P(y > k-1) - P(y > k) = \text{logit}^{-1}(X\beta - c_{k-1}) - \text{logit}^{-1}(X\beta - c_k)$$

Latent Variable Formulation

- This model might be easier to understand if we choose the latent variable formulation point of view.

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i < 0 \\ 2 & \text{if } z_i \in (0, c_2) \\ 3 & \text{if } z_i \in (c_2, c_3) \\ \dots & \\ k-1 & \text{if } z_i \in (c_{k-1}, c_k) \\ k & \text{if } z_i > c_k \end{cases} \\
z_i &= x_i^T \beta + \xi_i \quad \xi_i \text{ i.i.d. logistic}
\end{aligned}
$$

- With different error distribution, we can have multinomial probit and multinomial robit models with latent variable formulation.

R Implementation

- `polr` (*proportional odds logistic regression*) in package `MASS` is used to fit this class of models.
- logti `polr(y ~ x_1 + x_2, method=c('logit'))`
- probit `polr(y ~ x_1 + x_2, method=c('probit'))`
- cauchit `polr(y ~ x_1 + x_2, method=c('cauchit'))`

Unordered Categorical Response

- When reponse is unordered categorical, to model is need more effort.

- Assume there are J response categories, then the ith observation y_i can be converted into J binary reponse variables

$$y_{ij} = \begin{cases} 1 & y_i \text{ is in category } j \\ 0 & \text{otherwise} \end{cases}$$

- Recall that in logistic regression, we have

$$\log\left(\frac{p_i}{1-p_i}\right) = X_i^T \beta$$

Unordered Categorical Response

- In the unordered categorical response case, there are $\frac{J(J-1)}{2}$ pairs of categories.m

$$\begin{aligned} \log\left(\frac{p_{i1}}{p_{i2}}\right) &= X_i^T \beta_{12} \\ \log\left(\frac{p_{i1}}{p_{i3}}\right) &= X_i^T \beta_{13} \\ &\cdots \end{aligned}$$

- Choose one category as baseline category (generally the last category J), the model is given by

$$\log\left(\frac{p_{ij}}{p_{iJ}}\right) = X_i^T \beta_j, j = 1, 2, \ldots, J-1$$

Unordered Categorical Response

- Equivalently, the model can be expressed as

$$\pi_{ij} = \frac{\exp(X_i^T \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(X_i^T \beta_k)}, j = 1, 2, \ldots, J-1$$

- Unordered categorical response is very unstructured, so there are a lot of parameters.

- We can use `mlogit` in package `mlogit` to fit the model `mlogit(y ~ x_1 + x_2, reflevel='1')`

Generalized Linear Models

- Generalized Linear Model is a way of unifying various kinds of regression models, including Linear Regression, Logistic Regression and Poisson Regression.
- It is DIFFERENT from General Linear Model.
- There is also an unified way to do maximum likelihood estimation for GLM, which is called Iteratively Reweighted Least Squares algorithm.

Generalized Linear Models

There are three components for GLM

- A probability distribution for response y, hopefully from the exponential family
- Linear predictor $\eta = X\beta$
- A link function g such that $Ey = g^{-1}(X\beta)$