BIST P8110: Applied Regression II

18. Intro to Poisson Regression

Qixuan Chen

Department of Biostatistics
Columbia University

Count Data and Poisson Distribution

- Poisson distribution is often used to model count data.
  - If Y is the number of occurrences, it can be shown that $E(Y) = \lambda$ and $var(Y) = \lambda$.
- **Poisson regression** is the simplest regression model that allows to assess the association of count data with multiple covariates simultaneously.

Linear, Logistic and Poisson Regression

- Comparing linear, logistic, and Poisson regression models:

	Linear	Logistic	Poisson
Outcome variable	Continuous	Binary	Counts
Distribution	Normal	Binomial	Poisson
Parameter of interest	$E(Y) = \mu$	$E(Y) = p$	$E(Y) = \lambda$
Range of mean	$-\infty < \mu < \infty$	$0 < p < 1$	$0 < \lambda < \infty$
Variance	σ^2	$p(1-p)$	λ

Poisson Regression for Counts

- Poisson regression model:

$$\log\{E(Y_i|X_i)\} = \log\lambda_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \qquad (1)$$

- Poisson regression is one type of GLM
  - Distribution for Y: Poisson distribution
  - linear predictor: $\eta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$
  - link function: $g(E(Y_i|X_i)) = \log\{E(Y_i|X_i)\} = \eta$

Count Data and Rate

- Events occur over time (or space), and the length of time (or amount of space) can vary from observation to observation.
- The rate is specified in terms of units of “exposure”.
  - The average number of hospital admission in a day
  - The average number of motor vehicle crashes per 100,000 kms traveled by motor vehicles
  - For occupational injuries, each worker is exposed for the period they are at work, so the rate may be defined in terms of person-years“at risk”.

Poisson Regression for Rates

- Poisson regression model:

$$\log\{E(Y_i|X_i)\} = \log(n_i) + \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \qquad (2)$$

or

$$\log\left\{\frac{E(Y_i|X_i)}{n_i}\right\} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \qquad (3)$$

where, n_i is the units of "exposure".

  - Equation (2) differs from equation (1) with the inclusion of the term $\log(n_i)$.
  - $\log(n_i)$ is called the **offset**. It is a known constant.
  - Equation (1) is a special form of equation (2) when $n_i = 1$ for all units.

Interpretation of Coefficients

- A Poisson regression model for the rate of occupational injuries

$$\log\{E(Y_i|X_i)\} = \log(n_i) + \beta_0 + \beta_1 X_i$$

$$Y_i = \text{number of injuries for worker } i$$

$$n_i = \text{number of years at work for worker } i$$

$$X_i = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$$

- The **rate ratio** (RR) for males vs females is

$$\text{RR} = \frac{E(Y_i|\text{male})/n_i}{E(Y_j|\text{female})/n_j} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

- The 95% CI for RR is then

$$\left(e^{\hat{\beta}_1 - 1.96\times\widehat{se}(\hat{\beta}_1)}, e^{\hat{\beta}_1 + 1.96\times\widehat{se}(\hat{\beta}_1)}\right)$$

Interpretation of Coefficients

- If X is a binary variable as the model in the previous slide, β_1 is interpreted as
  - The occupational injury rate among male workers is e^{β_1} times of that among female workers.
- If X is a continuous variable, β_1 is interpreted as:
  - The rate is multiplied by e^{β_1} for each unit increase in X, or
  - The rate is multiplied by $e^{10\times\beta_1}$ for each ten-unit increase in X.

Wald Tests

- To assess whether or not a covariate X is related to the rate, we have the hypothesis test:

$$H_0 : \beta_1 = 0 \text{ vs } H_\alpha : \beta_1 \neq 0$$

  when $\beta_1 = 0$, the rate ratio $\text{RR} = 1$.

- The test statistic takes the form:

$$\frac{\hat{\beta}_1^2}{\widehat{Var}(\hat{\beta}_1)}$$

  which is a Wald test and has a chi-squared distribution with 1 df under the null hypothesis.

Likelihood Ratio Tests

- When we want to test the statistical significance of a group of variables or a categorical variable with more than two levels, we can use the likelihood ratio (LR) test

  $$-2\left(\log L_{small} - \log L_{big}\right) \sim \chi^2_{df}$$

  where, “small” and “big” refers to the model without or with the variables we are testing. “df” is equal to the difference in the numbers of parameters in the two models.

Deviance

- Deviance
  - Deviance = $-2(\log L_c - \log L_s)$, where “c” denotes the current fitted model,“s” denotes the saturated model. The saturated model has a parameter for every observation so that the data are fitted exactly.
  - It is a quality of fit statistic and can be used to test the goodness of fit.
- The LR test can be conducted using Deviance
  - LR = Deviance_{small} - Deviance_{large}
  - LR follows a Chi-squared distribution
  - df = df_{large} - df_{small}

Fitted Values

- The fitted values in a Poisson regression are given by

$$\hat{y}_i = n_i e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}}.$$

Pearson Residuals

- As variance and mean are assumed to be the same for the Poisson distribution, the standard error of Y_i is estimated by $\sqrt{\hat{y}_i}$, so the **Pearson residuals** are

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}.$$

- Poorly-fit subjects are those with Pearson residuals beyond ± 2.

Goodness of Fit Test

- Hypothesis
  H_0: There is no lack of fit vs. H_α: There is lack of fit
- Test statistic

$$\sum r_i^2 = \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \sim \chi^2_{df}$$

  where, $df = n - p$, the number of observations minus the total number of parameters in the model.
- Alternative goodness of fit statistic is Deviance.

Overdispersion

- Poisson distribution assumes variance and mean to be the same, but data can have more variability than expected. Consequently, variance can be larger than mean. We call it **overdispersion** in Poisson regression.
- To identify possible overdispersion in the data for a given model
  - Use the scale parameter, defined as Deviance or Pearson Chi-Square divided by its degrees of freedom ($n - p$).
  - The scale parameter is also called the dispersion parameter.
  - If the scale parameter is close to 1, evidence of over-dispersion is lacking.

Overdispersion

- A scale parameter that is greater than 1 does not necessarily imply overdispersion. It can indicate other problems, such as
  - an incorrectly specified model (omitted variables, interactions, or non-linear terms),
  - influential or outlying observations.

Overdispersion

- If the model is correctly specified and no outliers or influential observations but the scale estimate is still greater than 1, then conclude overdispersion.
- Overdispersion needs to be “fixed”, otherwise the estimates of the standard errors are too small, which leads to smaller p-values than they should.

Overdispersion

- Possible approaches to “fix” the overdispersion
  1. Fit a zero-inflated Poisson (ZIP) model when the over-dispersion is caused by an excessive number of 0’s.
  2. Fit a negative binomial regression.
  3. Allow the variance function of the Poisson distribution to have a multiplicative overdispersion factor ϕ, such that $\text{Var}(Y) = \phi \text{E}(Y)$, where ϕ is equal to the scale parameter (e.g. Deviance/DF).