Poisson Regression

Dr Wan Nor Arifin

Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.
E-mail: wnarifin@usm.my

Last update: 12 March 2019

Contents

1 Objectives

1. Understand the concept of Poisson distribution.
2. Understand the concept of simple and multiple Poisson regression models.
3. Apply the models on data sets and interpret the results.

2 Poisson Distribution

- Basically, we are dealing with *count* data.
- A Poisson distribution (Fleiss et al., 2003) is defined as

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

  for non-negative integers $y = 0, 1, 2, ..., \mu > 0$.
- Y is a Poisson random variable.
- The parameter μ is the mean of Y.
- This relationship between Y and μ can be written as $Y \sim Poisson(\mu)$ i.e. read as Y follows Poisson distribution with mean μ.
- Poisson distribution comes from the binomial distribution (Rice, 1995) i.e. limit of the binomial distribution as the number of trials n approaches infinity and probability of success p approaches zero, so as $np = \mu$ (n = number of trials/samples; p = probability of success) → Read my note *Probability Distribution* to understand this better.
- In other words, the number of event is very small as compared to the denominator, thus the p/proportion/percentage is very small, e.g. 0.000000123.
- Properties

  $Mean(Y) = Var(Y) = \mu$
- Graphs of Poisson probability mass function with different p:

  Excel file → Probability Distribution.xls > Poisson (also in `poisson.R`)
- It follows the assumptions of the *Poisson process* (Daniel, 1995):

  1. The occurrences of the events are *independent*. The occurrence of an event in an interval of space or time does not affect the probability of second occurrence of the event in the same or different interval.
  2. *Infinite* number of occurrences of the event is possible in the interval.
  3. Probability of a single occurrence of the event in an interval is *proportionate* to interval length.
  4. In a *very small portion* of the interval, probability of more than one occurrence of the event is *negligible*.
- Example 2.1:

  Probability of $Y = y$

  Suppose the number of death due to motor vehicle accidents per day in Malaysia is on average 17.2 and it was found that the daily distribution follows Poisson distribution.

  What is the probability that any randomly selected day will be the one with 10 death?

- Example 2.2:

  Probability of $Y \leq y$

  What is the probability that any randomly selected day will be the one with less than 11 death?

- Example 2.3:

  Probability of $Y > y$

  What is the probability that any randomly selected day will be more than 10 death?

- Using R: `poisson.R`

3 Simple Poisson Regression

- Let say Y is the Poisson count of some events e.g. number of accidents per month, or new HIV cases per year etc.
- Suppose the count is somehow associated with some factors Xs, e.g. gender, IVDU status, age etc.
- We want to relate the Y with the X. Mean Y can be linked with X by

$$ln\, E(Y|X) = \beta_0 + \beta_1 X$$

  or its equivalent equation

$$E(Y|X) = e^{\beta_0 + \beta_1 X} = e^{\beta_0} e^{\beta_1 X}$$

  *$E(Y)$ = expected value of Y or mean of Y; $E(Y|X)$ = conditional mean of Y given X. Remember mean of Y = $E(Y)$ = μ.

- For a simple case of exposure $X = 0, 1$, for reference/non-exposed group $X = 0$,

$$ln\, E(Y|X = 0) = \beta_0$$

$$E(Y|X = 0) = e^{\beta_0}$$

  thus the exponent of β_0 is the mean of Y when $X = 0$.
  For exposed group $X = 1$,

$$ln\, E(Y|X = 1) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

$$E(Y|X = 0) = e^{\beta_0 + \beta_1} = e^{\beta_0} e^{\beta_1}$$

  thus the exponent of $\beta_0 + \beta_1$ is the mean of Y when $X = 1$.
  Then, to obtain the increase/difference in mean of Y with the change in the exposure status, which is the exponent of β_1

$$\frac{E(Y|X = 1)}{E(Y|X = 0)} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

  usually called the *rate ratio, RR*.
  The same concept is also applicable whenever the X is numerical, in which it reflects RR for $x_1 - x_0 = \Delta$ unit change in X,

$$RR = \frac{E(Y|X = x_1)}{E(Y|X = x_0)} = \frac{e^{\beta_0} e^{\beta_1 (x1)}}{e^{\beta_0]} e^{\beta_1 (x0)}} = e^{\beta_1 - \beta_0} = e^{\beta_1 \Delta}$$

  or for 1 unit change in X,

$$RR = e^{\beta_1 (1)} = e^{\beta_1}$$

- Example 3.1, count data: `poisson.R`

- In medicine, it is more common to describe the count in term of prevalence, incidence, person-years i.e the rate. The equation has to be modified to include the denominator/person-years $a(X)$ by,

$$E(Y|X) = a(X)\, e^{\beta_0+\beta_1 X}$$

$$ln\, E(Y|X) = ln\, a(X) + \beta_0 + \beta_1 X$$

the $ln\, a(X)$ is specifically called the offset. This will be specified when we fit rate data.

- Example 3.2, rate data: `poisson.R`

4 Multiple Poisson Regression

- Recall our equation for simple Poisson regression,

$$ln\, E(Y|X) = ln\, \mu = \beta_0 + \beta_1 X$$

which can be extended as

$$ln\, E(Y|\mathbf{X}) = ln\, \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} = \beta_0 + \sum \beta_{p-1} X_{p-1}$$

where the $\mathbf{X}$ (in bold) denotes a collection of Xs. p is the number of estimated parameters. We minus 1 in the subscript since p also includes the intercept β_0, thus $p-1$ is the number of Xs.

- The rate ratio, RR is,

$$RR = e^{\beta_{p-1}}$$

- Similarly, to include the offset

$$ln\, E(Y|\mathbf{X}) = ln\, a(\mathbf{X}) + \beta_0 + \sum \beta_{p-1} X_{p-1}$$

- Now β_j (i.e. the specific β coefficient) is interpreted similarly to the simple regression case, while holding all other variables constant, or adjusted/controlling for the other variables.

- Similar to other multiple regressions,
  - create dummy variables for a categorical variable with > 2 categories. However, it is automatically created in R, if the variable is specified as a factor. (i.e. using the `factor()` function)
  - also consider the effect of two-way interaction terms in the model.

5 Model-building Steps for Multiple Poisson Regression

1. Variable selection.

   (a) **Univariable analysis.**
   - Determine the significance of the variables by
     - Wald's test.
     - LR test.

   (b) **Multivariable analysis.**

   i. Fit using selected variables.
   - All variables P-value $< .25$.
   - Clinically important variables

ii. Fit a smaller model by removing non-significant variables.

(c) **Interactions among variables.**

- Among clinically plausible pairs.

2. Model fit assessment.

(a) **Goodness-of-fit.**

i. **Chi-square goodness-of-fit.**

- In R, based on residual deviance (`poisgof()` function).
- $df = n - p$
- P-value > 0.05 indicates good fit.

ii. **Model-to-model AIC comparison.**

iii. **Scaled Pearson chi-square statistic.**

- Pearson chi-square statistic is given as

$$\chi^2_P = \sum \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

with $df = n - p$.

- Scaled Pearson chi-square statitic $= \chi^2_P/df$. The closer the value is to 1, the better is the fit Fleiss et al. (2003). Large value indicates *overdispersion problem (i.e. $VAR(Y) > Mean(Y)$).
- We have to calculate manually, but easy with R. Y_i is the observed counts, $\hat{\mu}_i$ is the fitted/predicted counts, obtained by `model$fitted` or `predicted(model)` functions.

(b) **Regression diagnostics.**

- We may use the standardized residuals, obtained by `rstandard()` function. Since it is in form of standardized z score, we may use specific cutoff e.g. > 1.96 ($\alpha = .05$) to > 3.89 ($\alpha = .0001$).

References

Daniel, W. W. (1995). *Biostatistics: A foundation for analysis in the health sciences.* USA: John Wiley & Sons, 6th ed. edition.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions.* USA: John Wiley & Sons, 3rd ed. edition.

Rice, J. A. (1995). *Mathematical statistics and data analysis.* USA: Duxbury Press, 2nd ed. edition.