BIOS762 - Notes

Mingwei Fei

December 4, 2022

1 Poisson distribution and Regression

The binomial distribution is closely related to other distributions. If $Y_1 \sim Poisson(\mu_1)$ and $Y_2 \sim Poisson(\mu_2)$ are independent, then Y_1 given $Y_1 + Y_2 = n$ follows a $B(n; \pi)$ distribution, where $\pi = \mu_1/(\mu_1 + \mu_2)$.
The Poisson distribution is closed under addition. If $Y_1; ...; Y_n$ are independent Poisson random variables with means $\mu_1; ...; \mu_n$, respectively, then $S = Y_1 + ... + Y_n \sim Poisson(\mu_1 + ... + \mu_n)$.

1.1 Log-linear model assumption

Poisson regression is a regression technique for modeling count data, such as colony counts for bacteria or viruses and accidents, as a function of a set of covariates.
Poisson Regression assumes

(a) Distribution assumption
$y_i|x_i \sim Poisson(\mu(x_i)), i = 1, .., n$, and $y = (y_1, y_2, ..y_n)^T$ are mutually independent.

(b) Structure assumption
$\mu(x_i)$ is related to x_i.

$$g(\mu_i) = x_i^T \beta,$$

If $g(\mu) = log\mu$, then it is a log-linear model.

1.1.1 Likelihood function

The log-likelihood function of poisson regression

$$log(p(\theta)) = y_i log\lambda - \lambda, \qquad \theta = log\lambda = x_i^T \beta$$
$$ln(\beta) = \sum_{i=1}^{n} [y_i\theta - exp(\theta)] = \sum_{i=1}^{n} \left[y_i x_i^T \beta - exp(x_i^T \beta)\right]$$

Taking the first derivatives, we have

$$\begin{aligned}\frac{\partial ln(\beta)}{\partial \beta} &= \sum_{i=1}^{n} \left[y_i x_i - exp(x_i^T \beta) x_i\right] = \sum_{i=1}^{n} \left[y_i - \mu_i(\beta)\right] x_i \\ &= \sum_{i=1}^{n} \frac{y_i - E(y_i)}{Var(y_i)} \partial_\beta \mu_i(\beta)\end{aligned}$$

Here is the score function, and relate to $E[y_i], Var[y_i]$ which we don't need the parameteric distribution and can get the quasi-likelihood regarding to derivative to μ_i.

$$\frac{\partial \theta}{\partial \mu} = \frac{\partial \mu}{\partial \theta}^{-1} = \frac{\partial \phi \dot{b}_\theta}{\partial \theta}^{-1} = \frac{\phi}{\ddot{b}_\theta} = \frac{1}{Var(y_i)}$$

Fisher Information

$$\frac{\partial^2 ln(\beta)}{\partial \beta \, \partial \beta} = \sum_{i=1}^{n} \left[-\mu_i(\beta)\right] x_i^{\otimes 2}$$

The Newton-Raphson algorithm is given by

$$\begin{aligned}\frac{\partial ln(\hat{\beta})}{\partial \beta} = 0 &= \frac{\partial ln(\hat{\beta}^*)}{\partial \beta} + \frac{\partial^2 ln(\beta^*)}{\partial \beta \, \partial \beta}(\hat{\beta} - \beta^*) + o_p(1) \\ \hat{\beta} &= \beta^* - \{\frac{\partial^2 ln(\beta^*)}{\partial \beta \, \partial \beta}\}^{-1} \frac{\partial ln(\beta^*)}{\partial \beta} \\ \beta^{k+1} &= \beta^k + \{\sum_{i=1}^{n} \mu_i(\beta) x_i^{\otimes 2}\}^{-1} \sum_{i=1}^{n} \left[y_i - \mu_i(\beta)\right] x_i\end{aligned}$$

The deviance function is the twice of the difference of log-likelihood between regression model and saturated model, (treat each y_i as μ_i)

$$\begin{aligned}ln(\mu_i) &= \sum_{i=1}^{n} y_i log(\mu_i) + \mu_i \\ D(y, \hat{\mu_i}) &= 2[ln(y_i) - ln(\hat{\mu_i})] = 2 \sum_{i=1}^{n} [y_i ln(\frac{y_i}{\mu_i}) - (y_i - \mu_i))]\end{aligned}$$

Poisson regression is useful for modeling event rates as a function of a set of covariates.

2 Loglinear model in contingency table

The poisson distribution and multinomial distribution is correlated.A common use of loglinear models is in modeling cell counts in contingency tables. The models specify how the expected count in each cell depends on both levels of the categorical variables

as well as the associations and interactions among them. The loglinear model provides a way of analyzing association and interaction patterns.

Consider a 2×2 contingency table of two categorical variables X and Y , each with two levels. Assume that we observe n subjects and the cell frequency counts are $n_{i,j}$ for $i = 1, 2; j = 1, 2$. The loglinear model assumes that the cell counts are independent observations from a Poisson distribution. Since the cell probabilities are π_{ij}, the expected frequencies are $\mu_{ij} = n\pi_{ij}$. If X and Y are independent, then $\pi_{ij} = \pi_{i.}\pi_{.j}$ for all i, j = 1, 2. This is also equivalent to

$$logR_{XY} = log\pi_{11} - log\pi_{12} - log\pi_{21} + log\pi_{22} = 0$$

2.1 Model assumption

Poisson regression is a regression technique for modeling count data, such as colony counts for bacteria or viruses and accidents, as a function of a set of covariates. Poisson Regression assumes

(a) Distribution assumption
$y_{ij} \sim Poisson(\mu_{ij}), i = 1, .., n$,

(b) Structure assumption
μ_{ij} is related to λ.

$$log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$
$$\lambda_1^X + \lambda_2^X = 0, \qquad \lambda_1^Y + \lambda_2^Y = 0, \qquad \lambda_{11}^{XY} = \lambda_{22}^{XY}, \qquad \lambda_{12}^{XY} + \lambda_{21}^{XY} = 0$$

Four identifiable parameters, $\lambda_1^X, \lambda_1^Y, \lambda_{11}^{XY}$

2.1.1 Odds Ratio

2×2 contingency table

Source	
$exp(\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY})$	$exp(\lambda + \lambda_1^X - \lambda_1^Y - \lambda_{11}^{XY})$
$exp(\lambda - \lambda_1^X + \lambda_1^Y - \lambda_{11}^{XY})$	$exp(\lambda - \lambda_1^X - \lambda_1^Y + \lambda_{11}^{XY})$

$$logR_{XY} = log\pi_{11} - log\pi_{12} - log\pi_{21} + log\pi_{22} = 4\lambda_{11}^{XY}$$

2.1.2 Likelihood function

The log-likelihood function of poisson regression in contingency table

$$
\begin{aligned}
p(\pi_{ij}) &= \prod_{i=1}^{2}\prod_{j=1}^{2}\pi_{ij}^{n_{ij}}, \qquad \pi_{ij} > 0, \quad \sum_i\sum_j \pi_{ij} = 1 \\
p(\mu_{ij}) &= \prod_{i=1}^{2}\prod_{j=1}^{2}\frac{exp(-\mu_{ij})\mu_{ij}^{n_{ij}}}{n_{ij}!}, \\
ln(\pi_{ij}) &= \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} log(\pi_{ij}) \\
ln(\mu_{ij}) &= \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} log\mu_{ij} - \mu_{ij}
\end{aligned}
$$

The likelihood function is same as multinomial distribution, then $\pi_{ij} = \frac{n_{ij}}{n}$. Thus, independence between X and Y can be characterized by the hypothesis as $H_0 : \lambda_{11}^{XY} = 0$. Under H_0, the loglinear model is given by

$$
log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \qquad i = 1, 2; \quad j = 1, 2
$$

where $\theta = (\pi_{11}, \pi_{12}, \pi_{21})^T$. Under $H_0, \hat{\pi}_{ij} = n_i \times n_j/n^2$, whereas $\hat{\pi}_{ij} = n_{ij}/n$ under the alternative hypothesis. Thus, we obtain the likelihood ratio statistic given by

$$
G^2 = 2[\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} log\frac{n_{ij}}{\mu_{ij}}] \sim \chi^2(1), \qquad \tilde{\mu}_{ij} = n_{.i}n_{.j}/n
$$

The likelihood ratio test follows a chi-square distribution

3 Loglinear model in $I \times J$ contingency table

3.0.1 Likelihood function

The log-likelihood function of poisson regression in contingency table

$$
\begin{aligned}
p(\mu_{ij}) &= \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{exp(-\mu_{ij})\mu_{ij}^{n_{ij}}}{n_{ij}!}, \\
ln(\mu_{ij}) &= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} log\mu_{ij} - \mu_{ij} \\
log(\mu_{ij}) &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}
\end{aligned}
$$

So

$$
\begin{aligned}
ln(\mu_{ij}) &= \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}[\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}] - exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \\
&= n\lambda + \sum_{i=1}^{I} n_{i+}\lambda_i^X + \sum_{j=1}^{J} n_{j+}\lambda_j^Y + \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\lambda_{ij}^{XY} - \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \\
n_{i+} &= \sum_{j=1}^{J} n_{ij}, \qquad n_{+j} = \sum_{i=1}^{I} n_{ij}
\end{aligned}
$$