Count Data

The count data comes from Poisson:

$$Y \sim P(\mu)$$

In this part, we consider two ways to analyze the data, through Poisson and through Binomial.

Discussion of Poisson and Binomial

Poisson Model

Model Assumption:

Y: the number of events in a Poisson process of rate $\exp(x^\top \beta)$ observed for a period T , where $\mu = T\exp(x^\top \beta) = \exp(x^\top \beta + \log T)$.

Explanation:

1. The canonical link of Possion is $\mu = \exp(x^\top \beta)$;
2. The expected mean increases proportional to the T;
3. This is a log-linear model with linear predictor $\eta' = x^\top \beta + \log T$, where $\log T$, a fixed part, is a *offset* term.

Note: The offset term can also be the the amount of population.

Binomial Model derived by Poission Model

$Y_i \sim P(\mu_i), i = 1, 2$ are independent. Then,

$$\mathrm{Y}_1 | Y_1 + Y_2 = m \sim \mathrm{Bin}(n, \frac{\mu_1}{\mu_1 + \mu_2})$$

Since Y are output of Poission model, we can use the log-linear model discussed above, that is, $\mu_1 = \exp\left(\gamma + x_1^{\mathrm{T}} \beta\right)$ and $\mu_2 = \exp\left(\gamma + x_2^{\mathrm{T}} \beta\right)$. Then,

$$\pi = \exp\left\{(x_2 - x_1)^{\mathrm{T}} \beta\right\} / [1 + \exp\left\{(x_2 - x_1)^{\mathrm{T}} \beta\right\}]$$

In this case, we can use logistic model to estimate β, but we cannot estimate γ.

Note: The analysis of binomial model requires observations, otherwise it will lose some information. Therefore, $se_{Poission}(\beta) \leq se_{Binomial}(\beta)$.

Contingency Tables

Sampling Scheme

There are several sampling schemes for obtaining continegency tables ($R \times C$):

1. No constraints on the row and column totals. For the count in the (r, c) cell, $y_{rc} \sim P(\mu_{rc})$. The likelihood is:
$$\prod_{cc} \left\{ \frac{\mu_{tc}^{\text{sc}}}{y_{yc}!} e^{-\mu_{cc}} \right\}$$
2. Fix the total number $\sum_{rc} y_{rc} = m$. Then, the data are multinomially distributed. Denoting $\pi_{rc} = \mu_{rc} / \sum_{s,t} \mu_{st}$, the likelihood is:
$$\frac{m!}{\prod_{r,c} y_{rc}!} \prod_{r,c} \pi_{rc}^{y_{rc}}, \quad \sum_{r,c} \pi_{rc} = 1$$
3. Fix the row totals $m_r = \sum_c y_{rc}$. Then, the data are independently multinomial distributions for each row. Denoting $\pi_{rc} = \mu_{rc} / \sum_t \mu_{rt}$, the likelihood is:
$$\prod_r \left\{ \frac{m_r!}{\prod_c y_{rc}!} \prod_c \pi_{rc}^{y_{rc}} \right\}, \quad \sum_c \pi_{1c} = \cdots = \sum_c \pi_{Rc} = 1$$

Estimation

Noting that count data is discrete, we use GLM to analyze it. Here, we use a link $\mu_{rc} = \exp(\gamma_r + x_{rc}^\top \beta)$ and consider sampling scheme 1 (Poisson) and 2 (Multinomial).

Then, some derivations show that:

$$\widehat{\beta}_{Poiss} = \widehat{\beta}_{Mult}, \widehat{\text{sd}}(\widehat{\beta}_{Poiss}) = \widehat{\text{sd}}(\widehat{\beta}_{Mult})$$

Note: Some softwares only depends on log-linear model. With this result, data comes from sampling method 2 can be analyzed with log-linear model.

Derivation

The relation of the likelihood is shown following:

$$\begin{aligned}
\ell_{\text{Poiss}}\ (\beta, \tau) &= \sum_{r,c} (y_{rc} \log \mu_{rc} - \mu_{rc}) \\
&= \sum_r \left(m_r \gamma_r + \sum_c y_{rc} x_{rc}^\top \beta - e^{\gamma_r} \sum_c e^{x_{rc}^\mathrm{T} \beta} \right) \\
&\equiv \sum_r (m_r \log \tau_r - \tau_r) + \sum_r \left\{ \sum_c y_{rc} x_{rc}^\mathrm{T} \beta - m_r \log \left(\sum_c e^{x_r^\top \beta} \right) \right\} \Bigg\} \\
&= \ell_{\text{Poiss}}\ (\tau; m) + \ell_{\text{Mult}}\ (\beta; y \mid m)
\end{aligned}$$

where $\tau_r = \sum_c \mu_{rc} = e^{\gamma_r} \sum_c e^{x_{rc}^r}$.

So that

$$\frac{\partial \ell_{\text{Poiss}}\ (\beta, \tau)}{\partial \beta} = \frac{\partial \ell_{\text{Multi}}\ (\beta, \tau)}{\partial \beta}$$

This implies the estimation of β are equal.

The expected information for β is:

$$\hat{I}_{Poiss}(\beta) = \sum_r \hat{\tau}_r \frac{\partial^2 \log\left(\sum_c e^{x_{rc}^{\mathrm{T}}\widehat{\beta}}\right)}{\partial\beta\partial\beta^{\mathrm{T}}} = \sum_r m_r \frac{\partial^2 \log\left(\sum_c e^{x_{rc}^{\mathrm{T}}\widehat{\beta}}\right)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

$$\hat{I}_{Mult}(\beta) = \sum_r m_r \frac{\partial^2 \log\left(\sum_c e^{x_{rc}^{\mathrm{T}}\widehat{\beta}}\right)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

So that

$$\widehat{\mathrm{sd}}(\widehat{\beta}_{Poiss}) = \widehat{\mathrm{sd}}(\widehat{\beta}_{Mult})$$

Note: In fact, the expected information matrixes for these two sampling scheme are different. It's interesting to find that if τ_r is unknown and need estimation, the estimated expected information matrixes under the two circumstances are the same.

References

- Statstical Models