Poisson Model in Contingency Table

1. Two-way table. Let y_{ij} be the count collected in a two-way table with I rows and J columns. Suppose that it is fitted by a loglinear model for Poisson data as

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{1}$$

for $i \in \{1, \cdots, I\}$ and $j \in \{1, \cdots, J\}$, where $\lambda_{ij} = \mathrm{E}(y_{ij})$. Then, (1) is a saturated model. In the absence of the interaction effect, the model becomes

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j. \tag{2}$$

Let $\hat{\lambda}_{ij}$ be the fitted value of λ_{ij} in (2). Then,

$$\hat{\lambda}_{ij} = e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j} = e^{\hat{\mu}} e^{\hat{\alpha}_i} e^{\hat{\beta}_j}. \tag{3}$$

We want to show that (2) is an indepencen model, which satisfies

$$\frac{\hat{\lambda}_{ij}}{\hat{\lambda}_{++}} = \left(\frac{\hat{\lambda}_{i+}}{\hat{\lambda}_{++}}\right)\left(\frac{\hat{\lambda}_{+j}}{\hat{\lambda}_{++}}\right) = \left(\frac{y_{i+}}{y_{++}}\right)\left(\frac{y_{+j}}{y_{++}}\right), \tag{4}$$

where $\hat{\lambda}_{i+} = y_{i+} = \sum_{j=1}^{J} y_{ij}$, $\hat{\lambda}_{+j} = y_{+j} = \sum_{i=1}^{I} y_{ij}$, $\hat{\lambda}_{++} = \hat{y}_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}$ (the proof of this relationship is omitted). An interpretation of (4) is that *the joint probability equals to the product of the marginal probabilities.*

We only focus on the proof of (4). By (3), we have

$$\hat{\lambda}_{i+} = \sum_{i=1}^{I} \hat{\lambda}_{ij} = e^{\hat{\mu}+\hat{\alpha}_i} \sum_{j=1}^{J} e^{\hat{\beta}_j}$$

$$\hat{\lambda}_{i+} = \sum_{j=1}^{J} \hat{\lambda}_{ij} = e^{\hat{\mu}+\hat{\beta}_j} \sum_{i=1}^{I} e^{\hat{\alpha}_i}$$

and

$$\hat{\lambda}_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\lambda}_{ij} = e^{\hat{\mu}} \left(\sum_{i=1}^{I} e^{\hat{\alpha}_i}\right)\left(\sum_{j=1}^{J} e^{\hat{\beta}_j}\right).$$

Then, we have (4).

2. Three-way table. Let y_{ijk} with $\lambda_{ijk} = \mathrm{E}(y_{ijk})$ for $i \in \{1, \cdots, I\}$, $j \in \{1, \cdots, J\}$, and $k \in \{1, \cdots, K\}$ be the count collected from a three way table. The saturated model is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}. \tag{5}$$

We can specify it into four reduced models.

Mutual indepdence model. The main-effect model, which is the mutual independence model, is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k. \tag{6}$$

There is

$$\hat{\lambda}_{ijk} = e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}_k}.$$

Then,

$$\hat{\lambda}_{i++} = e^{\hat{\mu}+\hat{\alpha}_i}\left(\sum_{j=1}^{J} e^{\hat{\beta}_j}\right)\left(\sum_{k=1}^{K} e^{\hat{\gamma}_k}\right),$$

$$\hat{\lambda}_{+j+} = e^{\hat{\mu}+\hat{\beta}_j}\left(\sum_{i=1}^{I} e^{\hat{\alpha}_i}\right)\left(\sum_{k=1}^{K} e^{\hat{\gamma}_k}\right),$$

$$\hat{\lambda}_{++k} = e^{\hat{\mu}+\hat{\gamma}_k}\left(\sum_{i=1}^{I} e^{\hat{\alpha}_i}\right)\left(\sum_{j=1}^{J} e^{\hat{\beta}_j}\right),$$

and

$$\hat{\lambda}_{+++} = e^{\hat{\mu}}\left(\sum_{i=1}^{I} e^{\hat{\alpha}_i}\right)\left(\sum_{j=1}^{J} e^{\hat{\beta}_j}\right)\left(\sum_{k=1}^{k} e^{\hat{\gamma}_k}\right).$$

We obtain an equation for independence as

$$\left(\frac{\hat{\lambda}_{ijk}}{\hat{\lambda}_{+++}}\right) = \left(\frac{\hat{\lambda}_{i++}}{\hat{\lambda}_{+++}}\right)\left(\frac{\hat{\lambda}_{+j+}}{\hat{\lambda}_{+++}}\right)\left(\frac{\hat{\lambda}_{++k}}{\hat{\lambda}_{+++}}\right). \tag{7}$$

It means that the joint probability (i.e.,(i, j, k)) equals to produce of the three marginal probabilities (i.e., i, j, and k, respectively).

Joint indepdence model. The model with one two-factor interaction effect, which is the joint independence model, is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}. \tag{8}$$

We have an equation for joint independence as

$$\left(\frac{\hat{\lambda}_{ijk}}{\hat{\lambda}_{+++}}\right) = \left(\frac{\hat{\lambda}_{ij+}}{\hat{\lambda}_{+++}}\right)\left(\frac{\hat{\lambda}_{++k}}{\hat{\lambda}_{+++}}\right). \tag{9}$$

Please show (9) by yourself. It means that the the joint probabilities equals to the product of joint (i, j) and marginal k probabilities.

Conditional indepdence model. The model with two two-factor interaction effects, which is the conditional independence model, is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}. \tag{10}$$

$$\left(\frac{\hat{\lambda}_{ijk}}{\hat{\lambda}_{i++}}\right) = \left(\frac{\hat{\lambda}_{ij+}}{\hat{\lambda}_{i++}}\right)\left(\frac{\hat{\lambda}_{i+k}}{\hat{\lambda}_{i++}}\right). \tag{11}$$

Please show it by yourself. It means that conditioning on i, the (conditional) joint probabilities (i.e., $(j,k)|i$) equals to the product of two (conditional) marginal probabilities (i.e., $j|i$ and $k|i$).

Uniform association model. The model with three two-factor interaction effects, which is the uniform association model, is

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}. \tag{12}$$

This model does not have an obvious interpretation.

3. Simpson's Paradox. Simpson's paradox is common in three-way or multi-way contingency table. We can interpret it by an example of diseases.

Assume a study compares the disease rates between two cities (e.g., A and B). The disease rates of old people (i.e, > 40) were 0.01 and 0.02, respectively. The disease rates of young people (i.e., ≤ 40) were 0.001 and 0.002, respectively. Suppose that 90% of city A and 10% of city B were old people. The overall disease rates of city A was

$$r_A = 0.01(0.9) + 0.001(0.1) = 0.0091$$

The overall disease rates of city A was

$$r_B = 0.02(0.1) + 0.002(0.9) = 0.0038.$$

Then, we have $r_B < r_A$. You should have the conclusion that city A is worse than city B when the age group ignored. The correct conclusion is drawn based on age groups. We conclude city A is better than city B.

Note: I have read an article about cancer incidence. In the comparison of lung cancer incidence rate among many countries in the world, a report showed that the lung cancer rate in North European countries (e.g., Norway, Sweden, and Finland) was higher than the lung cancer rate in South Asia countries. Note that air in North Europe was clear. One can draw a conclusion that clear air can increase the lung cancer rate. The reason is that the percentage of old people in North European countries is much higher than the percentage in South Asia countries.

4. Ordinal variables. Suppose that both row and columns are ordinal in a two-way table. We can assign scores to the two variables. Let u_i be the scores for rows and v_j be those for columns. Then, we can propose a few models between the main effects and the interaction effects models.

The main effects model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j,$$

where $\alpha_1 = \beta_1 = 0$ in the baseline constraint. It means that the rows and columns are independent. The predict count can be computed by

$$\hat{y}_{ij} = \frac{y_{i+}y_{+j}}{y_{++}}.$$

The interaction effects model

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

is the saturated model. Both models may not be interesting in practice since we are mostly interested in the relationship between rows and columns. Then, we can use the following three models.

Linear-by-linear association model. The linear-by-linear association model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma u_i v_j,$$

where the parameters are μ, α_i, β_j, and γ. It only has one degree of freedom in the term for interaction effects. The significance of the interaction effect can be determined by the p-value for γ. Let

$$\hat{y}_{ij} = \hat{\lambda}_{ij} = e^{\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j+\hat{\gamma}u_i v_j}$$

be the fitted value for y_{ij}. Based on the model, the odds ratio between rows i and i' and columns j and j' is

$$\hat{\theta}_{ii',jj'} = \frac{\hat{y}_{ij}\hat{y}_{i'j'}}{\hat{y}_{ij'}\hat{y}_{i'j}} = e^{\hat{\gamma}(u_i v_j + u_{i'} v_{j'} - u_{i'} v_j - u_i v_{j'})}.$$

Row-effect or column-effect model. The row-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_i v_j,$$

where the parameters are μ, α_i, β_j, and γ_i. To make the model well-defined, we need an additional constraint on γ_i, which can be $\gamma_1 = 0$, $\gamma_I = 0$, or $\sum_{i=1}^{I} \gamma_i = 0$. Therefore, the term for the interaction effects has $I - 1$ degrees of freedom. We can still study odds ratios by the above method.

The column-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_j u_i.$$

Study this model by yourself.

We can show that the linear-by-linear association model is a special case of the row-effect model. It is enough to study their interaction effects such that we have

$$\gamma_i v_j = \gamma u_i v_j \Rightarrow \gamma_i = \gamma u_i \Rightarrow \frac{\gamma_2 - \gamma_1}{u_2 - u_1} = \frac{\gamma_3 - \gamma_2}{u_3 - u_2} = \cdots = \frac{\gamma_I - \gamma_{i-1}}{u_I - u_{I-1}}.$$

The last equation also provides the null hypothesis in the test between the row-effect and linear-by-linear association models.

Row-column-effect model. The row-column-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_i \delta_j,$$

where the parameters are μ, α_i, β_j, γ_i, and δ_j. The constraints for the interaction effect terms are $\gamma_1 = \delta_1 = 0$, $\gamma_I = \delta_J = 0$, or $\sum_{i=1}^{I} \gamma_i = \sum_{j=1}^{J} \delta_j = 0$. Thus, it has $I + J - 2$ degrees of freedom in the interaction effect terms.

It has been pointed out that the row-column-effect model is not a generalized linear model. In fact, it is a generalized nonlinear model. The reason is that interaction effects terms are quadratic functions of unknown parameters. We cannot use the glm function to directly fit this model.