# MATH3091: Statistical Modelling II

## Problem Sheet 6 (Solution)

1. Suppose that $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ follows the Multinomial distribution with parameters $N$ and $\mathbf{p} = (p_1, \ldots, p_n)^\top$ with probability function given by

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \mathbf{p}) &= P(Y_1 = y_1, \cdots, Y_n = y_n) \\ &= \begin{cases} N! \frac{p_1^{y_1} \cdots p_n^{y_n}}{y_1! \cdots y_n!} & \text{if } \sum_{i=1}^n y_i = N \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

For given values of $x_1, \cdots, x_n$, consider the model

$$\log p_i = \beta_1 + \beta_2 x_i, \ 1 \leq i \leq n,$$

where $\beta_1$ is chosen so that $\sum_{i=1}^n p_i = 1$.

a. Find an expression for $\beta_1$ in terms of $\beta_2$ and $x_1, \ldots, x_n$.

b. Given observed cell counts $y = (y_1, \cdots, y_n)$, find the log-likelihood function for $\beta_2$.

c. Derive an equation for finding $\hat{\beta}_2$, the MLE of $\beta_2$.

d. Write down an expression for the fitted probabilities $\hat{p}_i$ under the model, in terms of $\hat{\beta}_2$.

Solution:

a. We have

$$1 = \sum_{i=1}^n p_i = \sum_{i=1}^n e^{\beta_1 + \beta_2 x_i} = e^{\beta_1} \sum_{i=1}^n e^{\beta_2 x_i},$$

so

$$\beta_1 = -\log \left( \sum_{i=1}^n e^{\beta_2 x_i} \right).$$

b. The log-likelihood is

$$\begin{aligned}\ell(\beta_2) &= \sum_{i=1}^{n} y_i \log(p_i) + \log(N!) - \sum_{i=1}^{n} \log(y_i!) \\ &= \sum_{i=1}^{n} y_i(\beta_1 + \beta_2 x_i) + \log(N!) - \sum_{i=1}^{n} \log(y_i!) \\ &= \beta_1 \sum_{i=1}^{n} y_i + \beta_2 \sum_{i=1}^{n} y_i x_i + \log(N!) - \sum_{i=1}^{n} \log(y_i!) \\ &= -N \log(\sum_{i=1}^{n} e^{\beta_2 x_i}) + \beta_2 \sum_{i=1}^{n} y_i x_i + \log(N!) - \sum_{i=1}^{n} \log(y_i!)\,.\end{aligned}$$

c. The score is

$$u(\beta_2) = \frac{\partial l}{\partial \beta_2} = \sum_{i=1}^{n} y_i x_i - N \sum_{j=1}^{n} x_j e^{\beta_2 x_j} \left( \sum_{i=1}^{n} e^{\beta_2 x_i} \right)^{-1}.$$

So $\hat{\beta}_2$ satisfies

$$u(\hat{\beta}_2) = \sum_{i=1}^{n} y_i x_i - N \sum_{j=1}^{n} x_j e^{\hat{\beta}_2 x_j} \left( \sum_{i=1}^{n} e^{\hat{\beta}_2 x_i} \right)^{-1} = 0\,.$$

d. The fitted probabilities are

$$\hat{p}_i = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \exp\left( -\log(\sum_{j=1}^{n} e^{\hat{\beta}_2 x_j}) + \hat{\beta}_2 x_i \right) = \frac{e^{\hat{\beta}_2 x_i}}{\sum_{j=1}^{n} e^{\hat{\beta}_2 x_j}}\,.$$

2. Suppose that $Z \sim \text{Binomial}(10, p)$, and that we have a single observation $z$ from this distribution, for some unknown value of the parameter $p$.

   a. Write down the MLE $\hat{p}$ of $p$.

   b. Show that this model may be written as a special case of the model in Question 1 with $n = 2$, where you should specify $N$, define $Y_1$ and $Y_2$ in terms of $Z$, and choose appropriate values of $x_1$ and $x_2$ (there may be more than one valid choice).

   Find $\hat{p}_i$ $(i = 1, 2)$ in this case, and express $\hat{p}_i$ in terms of $\hat{p}$.

Solution:

a. We have $\hat{p} = z/10$.

b. Write $Y_1 = Z$ and $Y_2 = 10 - Z$. Then for $N = 10$, $n = 2$, this is a special case of (a), where we want

$$\log p_1 = \beta_1 + \beta_2 x_1 = \log p\,,$$

and

$$\log p_2 = \beta_1 + \beta_2 x_2 = \log(1-p)\,.$$

To achieve this, we could set $x_1 = 0$ and $x_2 = 1$ (other choices are also fine). Then we have

$$p_1 - p_2 = \beta_2 = \log(1-p) - \log(p) = -\log\left(\frac{p}{1-p}\right) = -\text{logit}(p)\,,$$

and $\beta_1$ is already fixed as in Question 1 (a).

From Question 1 (c), we know

$$\sum_{i=1}^{2} y_i x_i - 10 \sum_{j=1}^{2} x_j e^{\hat{\beta}_2 x_j} \left( \sum_{i=1}^{2} e^{\hat{\beta}_2 x_i} \right)^{-1} = 0\,,$$

which simplifies to

$$y_2 - 10 \frac{e^{\hat{\beta}_2}}{1 + e^{\hat{\beta}_2}} = 0\,,$$

so

$$\hat{\beta}_2 = \text{logit}\left(\frac{y_1}{10}\right)\,.$$

This gives

$$\hat{p}_1 = \frac{1}{1 + e^{\hat{\beta}_2}} = 1 - \frac{e^{\hat{\beta}_2}}{1 + e^{\hat{\beta}_2}} = 1 - \frac{y_2}{10} = \frac{z}{10} = \hat{p}\,,$$

and

$$\hat{p}_2 = \frac{e^{\hat{\beta}_2}}{1 + e^{\hat{\beta}_2}} = \frac{y_2}{10} = 1 - \frac{z}{10} = 1 - \hat{p}\,.$$

3. Suppose we are interested in which factors might affect whether people go on to develop a disease. A study on this recruits 220 healthy volunteers, and monitors how many people go on to develop the disease in a one-year period, cross-classified by smoking status and gender. Suppose that there are a total of 100 men (20 smokers and 80 non-smokers) and 120 women (30 smokers and 90 non-smokers). We now consider four possible experiments which might be use to collect this data.

   a. The total number of people recruited into the study is fixed at 220, with no constraints on gender or smoking status.

   b. The number of men recruited into the study is fixed at 100, and the number of women is fixed at 120, with no constraints on smoking status.

c. The number of smokers recruited into the study is fixed at 50, and the number of non-smokers is fixed at 170, with no constraints on gender.

d. The numbers of male smokers recruited into the study is fixed at 20, male non-smokers fixed at 80, female smokers at 30 and female non-smokers at 90.

In each case, we can model the cell counts $(y_1, \cdots, y_8)$ (the cross-classification by disease, gender and smoking status) by using a Multinomial distribution, with cell probabilities $(p_1, \cdots, p_8)$. The various experimental setups (a)—(d) provide different restrictions on marginal totals.

A saturated log-linear model for $\mu_i = 220 \cdot p_i$ may be written as

$$\log(\mu_i) = \alpha + \beta_D(d_i) + \beta_G(g_i) + \beta_S(s_i) + \beta_{DG}(d_i, g_i) + \beta_{DS}(d_i, s_i) + \beta_{GS}(g_i, s_i) + \beta_{DGS}(d_i, g_i, s_i)\,,$$

where

$$d_i = \begin{cases} 1 & \text{if group } i \text{ have disease} \\ 0 & \text{otherwise,} \end{cases} \qquad g_i = \begin{cases} 1 & \text{if group } i \text{ male} \\ 0 & \text{if group } i \text{ female,} \end{cases}$$

and

$$s_i = \begin{cases} 1 & \text{if group } i \text{ smokers} \\ 0 & \text{if group } i \text{ non-smokers.} \end{cases}$$

Each $\beta$ term is non-zero only if all its arguments are 1, e.g. $\beta_{DG}(1,0) = \beta_{DG}(0,1) = \beta_{DG}(0,0) = 0$.

We could fit this model as a Poisson log-linear model, and conduct hypothesis tests to determine whether or not to drop each term. Which terms must be kept in the model in each case (a)—(d)?

Suppose that our final preferred model is

$$\log(\mu_i) = \alpha + \beta_D(d_i) + \beta_G(g_i) + \beta_S(s_i) + \beta_{DG}(d_i, g_i) + \beta_{GS}(g_i, s_i)\,.$$

What interpretation would you make about conditional independence of variables?

Solution:

In order to fit the Multinomial log-linear model using a Poisson log-linear model, we need some restrictions

a. The total is fixed in advance, so must include the intercept $\alpha$.
b. The gender margin is fixed in advance, so must include the gender main effect $\beta_G$ and the intercept $\alpha$.
c. The smoking margin is fixed in advance, so must include the smoking main effect $\beta_S$ and the intercept $\alpha$.
d. The gender/smoking margin is fixed in advance, so must include the gender/smoking interaction $\beta_{GS}$, the gender main effect $\beta_G$, the smoking main effect $\beta_S$ and the intercept $\alpha$.

If the final preferred model is

$$\log(\mu_i) = \alpha + \beta_D(d_i) + \beta_G(g_i) + \beta_S(s_i) + \beta_{DG}(d_i, g_i) + \beta_{GS}(g_i, s_i)\,,$$

we would conclude that disease is conditionally independent of smoking status, given gender.