CS&SS/STAT/SOC 536: Regression for Multinomial Outcomes

Adrian Dobra
adobra@uw.edu

1 Introduction

We assume that the outcome y takes a small number of values $\{1, 2, \ldots, J\}$ ($J \geq 2$). Moreover, we assume that here is no ordering of these values. For example, y can be a set of colors: $\{1 = Blue, 2 = Red, 3 = White\}$. We are only coding each color with a number, but that does not mean that the colors are somehow ordered. We assume that we have p explanatory (independent) variables $x = (x_1, x_2, \ldots, x_p)$.

We are interested in modeling the conditional probability that y takes a value $j \in \{1, 2, \ldots, J\}$ given the values of the explanatory variables, i.e.

$$\pi_j(x) = \pi_j(x_1, \ldots, x_p) = P(y = j|x).$$

We need to have:

$$\pi_j(x) > 0, \text{ for } j = 1, 2, \ldots, J,$$

and

$$\sum_{j=1}^{J} \pi_j(x) = 1.$$

If one category never occurs, i.e. (it is impossible and $\pi_j(x) = 0$ for any x), then we simply take y to have $J - 1$ categories, not J categories. For example, if there is no "red" sample, you just say that your outcome takes two colors (Blue and White) instead of Blue, Red and White.

We assume we have observed n samples involving the outcome and the p explanatory variables:

$$\{(y^i, x^i) : i = 1, \ldots, n\},$$

where $x^i = (x_1^i, x_2^i, \ldots, x_p^i)$. This means that your data can be summarized as a $n \times (p + 1)$ matrix, where the outcome occupies the first column, while the explanatory variables occupy the remaning p columns.

We denote by n_j the number of samples that have $y = j$ (that is, we count how many

samples are red, how many samples are white, etc). We assume $n_j \geq 1$ for each $j = 1, 2, \ldots, J$ (that is, there is at least one sample of each color). We must have

$$\sum_{j=1}^{J} n_j = n.$$

2 The regression model

We consider category J to be the baseline category. Since the numbers do not imply any ordering, this means that any category can be the baseline category (i.e., white can be baseline or red can be baseline, it does not matter). Just for convenience, we take the category associated with the largest numerical code to be baseline. You must be careful here because some software take the first category to be the baseline, while other softwares take the last category to be the baseline. Again, you will get exactly the same model, except in a different parametrization.

For each $j = 1, 2, \ldots, J-1$, we consider the binary regression model of category j vs category J:

$$\log \Omega_{j|J}(x) = \log \frac{\pi_j(x)}{\pi_J(x)} = \beta_{0,j|J} + \sum_{k=1}^{p} \beta_{k,j|J} x_k. \tag{1}$$

The above binary model is based on $n_j + n_J$ samples. You can think about this as splitting your data into $J-1$ smaller datasets that will help you model the odds of the outcome being in category j vs. the baseline category J.

Once you fit these $J-1$ logit models, you would have also fit the logit model between any two other categories j_1 and j_2:

$$\log \Omega_{j_1|j_2}(x) = \log \frac{\pi_{j_1}(x)}{\pi_{j_2}(x)} = \beta_{0,j_1|j_2} + \sum_{k=1}^{p} \beta_{k,j_1|j_2} x_k \tag{2}$$

We have the simple relationship:

$$\begin{aligned} \log \Omega_{j_1|j_2}(x) &= \log \Omega_{j_1|J}(x) - \log \Omega_{j_2|J}(x), \\ &= (\beta_{0,j_1|J} - \beta_{0,j_2|J}) + \sum_{k=1}^{p} (\beta_{k,j_1|J} - \beta_{k,j_2|J}) x_k \end{aligned}$$

which implies

$$\begin{aligned} \beta_{0,j_1|j_2} &= \beta_{0,j_1|J} - \beta_{0,j_2|J}, \\ \beta_{k,j_1|j_2} &= \beta_{k,j_1|J} - \beta_{k,j_2|J}. \end{aligned}$$

Note: the above differences between regression coefficients are called contrasts. Therefore the $J-1$ binary regressions (1) fully determine the dependence between y and the p explanatory

variables. We can write (1) in an equivalent form:

$$\pi_j(x) = \pi_J(x) \cdot \exp\left(\beta_{0,j|J} + \sum_{k=1}^{p} \beta_{k,j|J} x_k\right). \tag{3}$$

Thus $\pi_j(x) > 0$ if $\pi_J(x) > 0$. At this point we need to remember we also need to satisfy the constraint $\sum_{j=1}^{J} \pi_j(x) = 1$. We use (3) to obtain:

$$\pi_J(x) = \frac{1}{1 + \sum_{l=1}^{J-1} \exp\left(\beta_{0,l|J} + \sum_{k=1}^{p} \beta_{k,l|J} x_k\right)}, \tag{4}$$

and

$$\pi_j(x) = \frac{\exp\left(\beta_{0,j|J} + \sum_{k=1}^{p} \beta_{k,j|J} x_k\right)}{1 + \sum_{l=1}^{J-1} \exp\left(\beta_{0,l|J} + \sum_{k=1}^{p} \beta_{k,l|J} x_k\right)}, \tag{5}$$

for $j = 1, \ldots, J-1$. We can unify the notations for all the categories by defining

$$\beta_{0,J|J} = \beta_{1,J|J} = \ldots = \beta_{p,J|J} = 0. \tag{6}$$

We can then write

$$\pi_j(x) = \frac{\exp\left(\beta_{0,j|J} + \sum_{k=1}^{p} \beta_{k,j|J} x_k\right)}{\sum_{l=1}^{J} \exp\left(\beta_{0,l|J} + \sum_{k=1}^{p} \beta_{k,l|J} x_k\right)}, \tag{7}$$

for $j = 1, \ldots, J$. Relations (7) and (6) OR relations (4) and (5) represent two different forms of the same regression model for the multinomial outcome y and p explanatory variables. You should remember that this model is equivalent with J regressions for binary outcomes in which the coefficients of one of these regressions are set to zero. Since we took category J to be baseline, we set to zero the regressions coefficients associated with the J-th binary regression — see (6). However, we could have chosen to set the coefficients of any other regression to zero and would have obtained the same model in an equivalent form.

3 Maximum Likelihood Estimation

The important message you need to take away from this section is that the $J-1$ binary regressions are not fit separately. Instead, they are fit together using one maximization procedure. *You could try to break your data into $J-1$ datasets, fit the binary regressions*

associated with each of these datasets, then compare the coefficient estimates with those obtained by fitting the $J-1$ regressions simultaneously. Your coefficient estimates might be close, but they will not be the same.

We denote by θ the regression coefficients of the model (4) and (5). There are $(p+1)\times(J-1)$ *free* coefficients:

$$\beta_{k,j|J},\ \text{for } k=0,1,\ldots,p \text{ and } j=1,2,\ldots,J-1. \tag{8}$$

If the outcome for the i-th sample if the j-th category, we have $y^i=j$. Equivalently, we can consider the vector form:

$$y^i=(y_1^i,\ldots,y_{j-1}^i,y_j^i,y_{j+1}^i,\ldots,y_J^i)=(0,\ldots,0,1,0,\ldots,0).$$

Thus $\sum_{j=1}^{J} y_j^i=1$ for each sample $i=1,2,\ldots,n$. The log-likelihood is:

$$\begin{aligned} l(\theta|data) &= \sum_{i=1}^{n}\left[\sum_{j=1}^{J-1} y_j^i \log\Omega_{j|J}(x^i)+\log\pi_J(x^i)\right], \\ &= \sum_{i=1}^{n}\left[\sum_{j=1}^{J-1} y_j^i\left(\beta_{0,j|J}+\sum_{k=1}^{p}\beta_{k,j|J}x_k^i\right)-\log\left(1+\sum_{l=1}^{J-1}\exp\left(\beta_{0,l|J}+\sum_{k=1}^{p}\beta_{k,l|J}x_k^i\right)\right)\right] \end{aligned}$$

This log-likelihood is concave, hence it has a unique maximum. The Newton-Raphson algorithm will converge to this maximum and return the MLEs of the regression coefficients (8).

4 Interpreting the regression parameters

We refer to equation (2) that gives the odds of categories j_1 and j_2 given the explanatory variables. It follows that

$$\frac{\partial\log\Omega_{j_1|j_2}(x)}{\partial x_k}=\widehat{\beta}_{k,j_1|J}-\widehat{\beta}_{k,j_2|J}.$$

The interpretation of the contrast $\beta_{k,j_1|J}-\beta_{k,j_2|J}$ is as follows:

For a unit change in x_k, the odds $\Omega_{j_1|j_2}(x)$ are expected to change by a factor of $\exp(\widehat{\beta}_{k,j_1|J}-\widehat{\beta}_{k,j_2|J})$.

If $j_2=J$ (recall that we took J to be the baseline category), we have $\beta_{k,J|J}=0$. Thus:

$$\frac{\partial\log\Omega_{j_1|J}(x)}{\partial x_k}=\widehat{\beta}_{k,j_1|J}.$$

The interpretation of the coefficient $\beta_{k,j_1|J}$ is as follows:

For a unit change in x_k, the odds $\Omega_{j_1|J}(x)$ are expected to change by a factor of $\exp(\widehat{\beta}_{k,j_1|J})$.

5 Variable Selection

Let $\mathcal{M}$ be a regression model that involves a variable x_{k_0}, $1 \leq k_0 \leq p$. We denote by $\mathcal{M}_{-k_0}$ the regression obtained from $\mathcal{M}$ by deleting variable x_{k_0}. Making a choice between $\mathcal{M}_{-k_0}$ and $\mathcal{M}$ involves testing the null hypothesis:

$$H_0: \quad \beta_{k_0,1|J} = \beta_{k_0,2|J} = \ldots = \beta_{k_0,J-1|J} = 0.$$

That is, we set to zero the coefficients associated with x_{k_0} in all the $J-1$ regressions. Therefore H_0 involves $J-1$ parameters. The likelihood ratio test statistics is the difference in the deviances of the two models:

$$G^2(\mathcal{M}_{-k_0}|\mathcal{M}) = \mathcal{D}(\mathcal{M}_{-k_0}) - \mathcal{D}(\mathcal{M}).$$

The asymptotic distribution of $G^2(\mathcal{M}_{-k_0}|\mathcal{M})$ is Chi-square with $J-1$ degrees of freedom because the difference in the dimensions of the two models is $J-1$. Therefore the p-value associated with the likelihood ratio test is:

$$P\left(\chi^2_{J-1} \geq G^2(\mathcal{M}_{-k_0}|\mathcal{M})\right).$$

The AIC and BIC are obtained with the same formulas given in the handout "Regression for binary outcomes". You should count properly the number of free parameters of your regression model (i.e., the dimension of the model). If $\mathcal{M}$ involves p explanatory variables, its dimension will be $p \times (J-1)$.