STATS 212: Generalized Linear Models
Lecture 5: Multinomial Logistic Model, Poisson Model, and More

Babak Shahbaba

UCI, Spring 2010

Multinomial logistic model

- Multinomial logistic model (MNL) is a generalization of logistic regression when the outcome could have multiple values (i.e., could belong to one of K classes).
- The random component has a multinomial distribution,

$$y_i|n_i, \mu_{i1}, ..., \mu_{iK} \sim \text{multinomial}(n_i, \mu_{i1}, ..., \mu_{iK})$$

  where μ_{ik} is the probability of class k for observation i such that $\sum_{k=1}^{K} \mu_{ik} = 1$.
- y_i is also a vector of K elements with $\sum_{k=1}^{K} y_{ik} = n_i$.
- The systematic part is now a vector $\eta_{ik} = x_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a matrix of size $(p+1) \times K$.

Multinomial logistic model

- For this model, we use a generalization of the link function we used for the binary logistic regression

$$\mu_{ik} = \frac{\exp(x_i \beta_k)}{\sum_{k'=1}^{K} \exp(x_i \beta_{k'})}$$

- The likelihood in terms of β is as follows:

$$p(y|\mu) \propto \prod_{i=1}^{n} \prod_{k=2}^{K} \mu_{ik}^{y_{ik}}$$

$$P(y|x,\beta) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp(x_i \beta_k)}{\sum_{k'=1}^{K} \exp(x_i \beta_{k'})} \right)^{y_{ik}}$$

Multinomial logistic model

- $\boldsymbol{\beta}$ in general is a $(p+1) \times K$ matrix. The first row, $(\beta_{01}, ..., \beta_{0K})$ are intercepts, and $(\beta_{j1}, ..., \beta_{jK})$ in row $j+1$ are regression parameters associated with the j^{th} predictor.
- x_i is the row vector of predictors value for observation i (including the constant 1 at the beginning).
- y_{ik} is the number of cases in observation i that are in class k.
- Each column k (where $k = 1, ..., K$) corresponds to a set of $p+1$ parameters associated with class k.

Multinomial logistic model

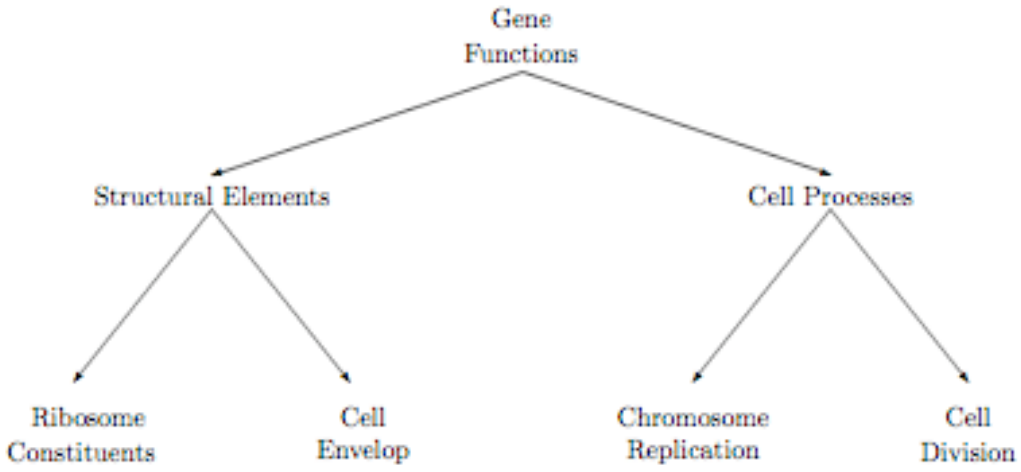
- This representation is redundant and results in nonidentifiability, since one of the β_k's (where $k = 1, ..., J$) can be set to zero without changing the set of relationships expressible with the model.
- Usually, either the parameters for $k = 1$ (the first column) or for $k = K$ (the last column) would be set to zero.
- The likelihood function for the identifiable model is

$$P(y|x, \beta) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp(x\beta_k)}{1 + \sum_{k'=1}^{K-1} \exp(x\beta_{k'})} \right)^{y_{ik}}$$

- As we do for any GLM, we obtain the score function by taking the first derivative of log-likelihood, $L(\beta)$, with respect to β.
- Fisher information is the expectation of $-\frac{\partial L^2(\beta)}{\partial \beta_j \partial \beta_k}$.

Modeling structured categorical response

- The MNL model discussed above treats classes as unrelated entities without any hierarchical structure.
- This is not always a realistic assumption. In many classification problems, one can arrange classes in a hierarchical form.
- For example, gene functions are usually presented in a hierarchical form starting with very general classes (eg, cell processes) and becoming more specific in lower levels of the hierarchy (eg, cell division).



Modeling ordered classes

- If the classes have in fact the assumed structure, one would expect to obtain a higher performance by using this additional information.
- A special case is when the classes are ordered (e.g., education level).
- For these problems, we could use a more parsimonious model to improve the power.
- One such model is the *cumulative logit* defined as follows

$$
\begin{aligned}
P(y_i \leq k|x_i, \beta) &= \mu_1 + \mu_2 + ... + \mu_k \\
logit[P(y_i \leq k|x_i, \beta)] &= \log[\frac{\mu_1 + \mu_2 + ... + \mu_k}{\mu_{k+1} + \mu_{k+2} + ... + \mu_K}] \\
&= \log[\frac{P(y_i \leq k|x_i, \beta)}{1 - P(y_i \leq k|x_i, \beta)}] \\
&= \alpha_k + x_i\beta, \qquad k = 1, ..., K-1
\end{aligned}
$$

Modeling ordered classes

- Note that in this model, we denote the intercept as α, therefore β denotes regression coefficients only, and x_i does not include a constant 1 as its first element.
- In this model, while the regression coefficients are shared between all catogories, each category has its own unique intercept α_j.
- Note that $\mu_k = P(y_i \leq k|x_i, \beta) - P(y_i \leq k-1|x_i, \beta)$.
- Therefore, the likelihood is as follows

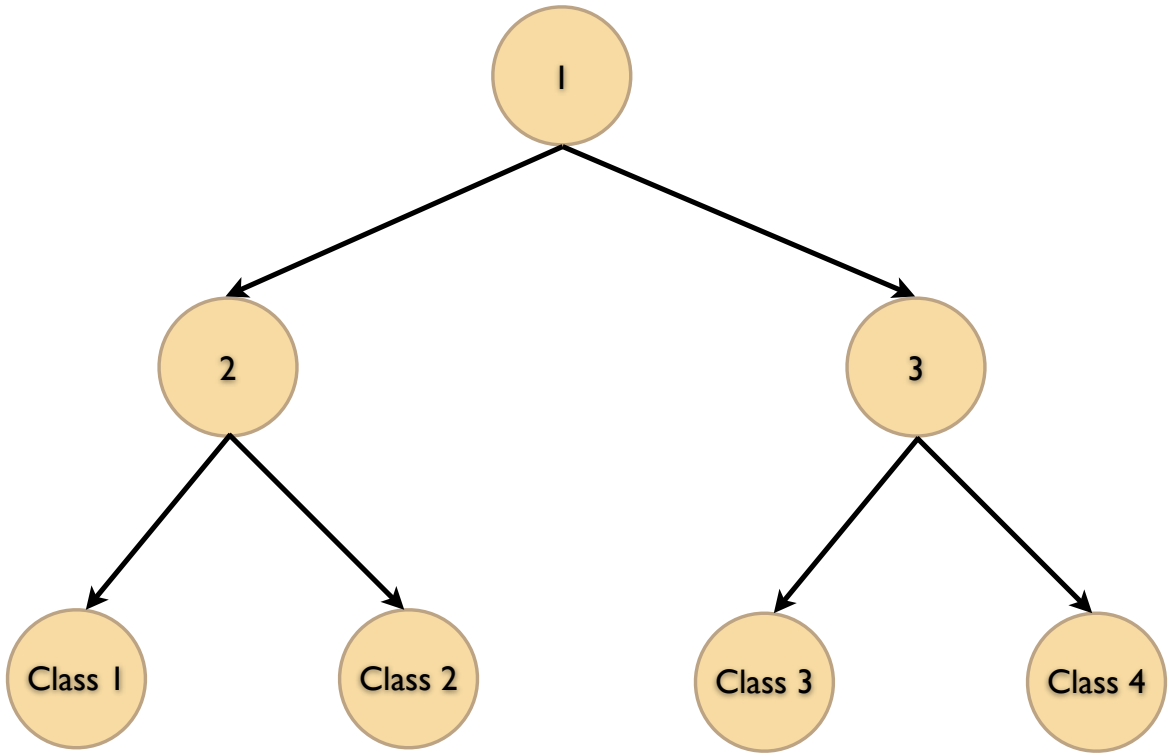
$$\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{\exp(\alpha_k + x_i\beta)}{1+\exp(\alpha_k + x_i\beta)} - \frac{\exp(\alpha_{k-1} + x_i\beta)}{1+\exp(\alpha_{k-1} + x_i\beta)}\right)^{y_{ik}}$$

Modeling hierarchical classes

- In general, categorical response variables can have hierarchical structures like the one we showed for gene functions.
- One approach for modelling hierarchical classes is to decompose the classification model into nested models (e.g., logistic or MNL).
- Nested MNL models are extensively discussed in econometrics in the context of estimating the probability of a person choosing a specific alternative (i.e., class) from a discrete set of options (e.g., different modes of transportation).

Modeling hierarchical classes

- For hierarchical classification problems with simple binary partitions, we can use successive logistic models for each binary class.
- In the figure below, for example, these partitions are $\{12, 34\}$, $\{1, 2\}$, and $\{3, 4\}$.



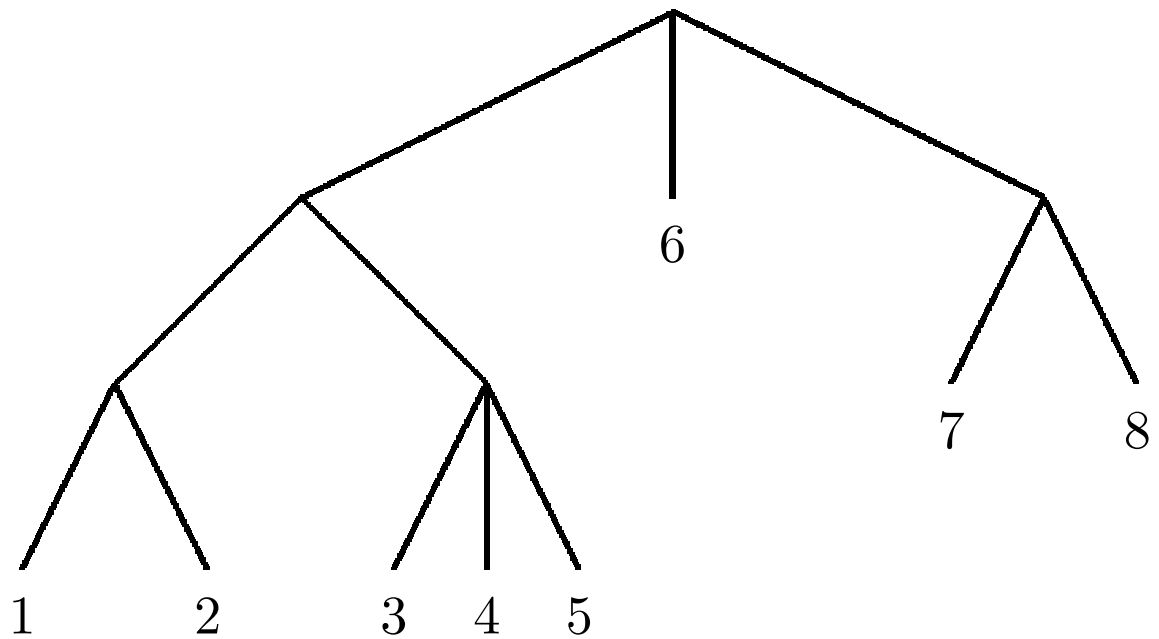
Modeling hierarchical classes

- The resulting nested binary models are statistically independent, conditioned on the upper levels. The likelihood can therefore be written as the product of the likelihoods for each of the binary models.
- For example, for the above hierarchical structure we have

$$P(y = 1|x) \quad = \quad P(y \in \{1, 2\}|x) \times P(y \in \{1\}|y \in \{1, 2\}, x)$$

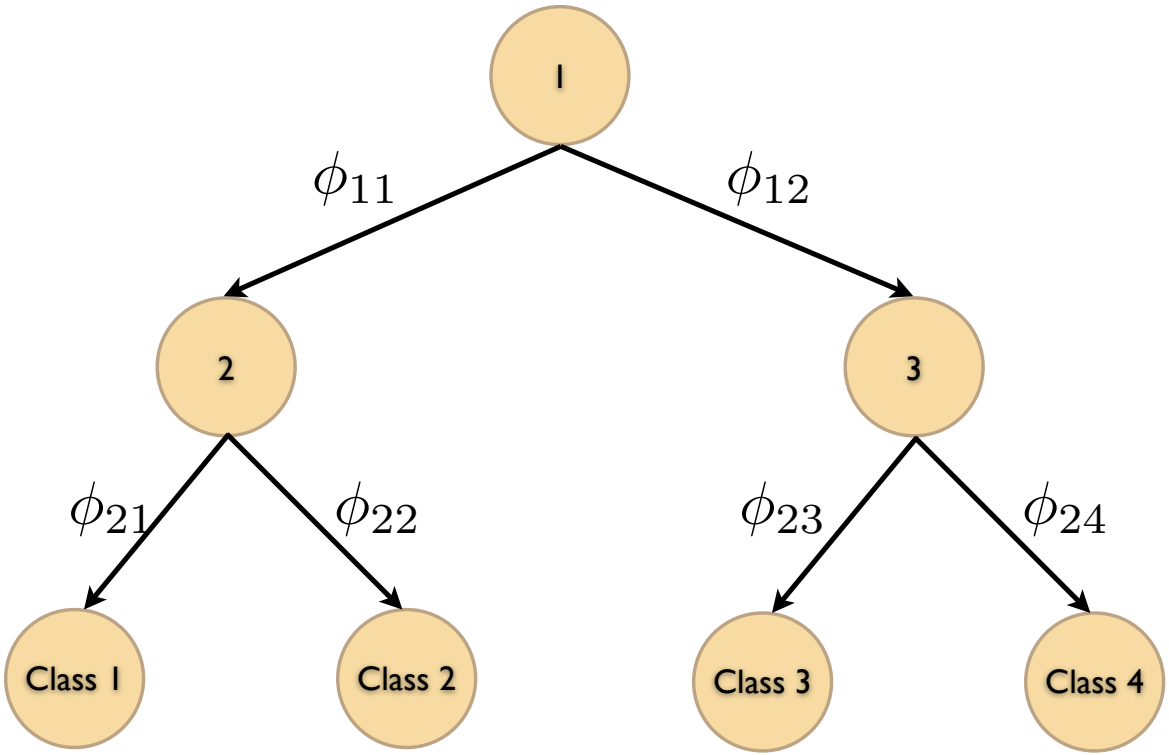
Modeling hierarchical classes

- Restriction to binary models is unnecessary.
- At each level, classes can be divided into more than two subsets and MNL can be used instead of logistic regression.



Modeling hierarchical classes- corMNL

- Shahbaba and Neal (2007) proposed an alternative method for modeling hierarchical classes.
- Consider the the following figure:



- For each branch in the hierarchy, they define a different set of parameters: ϕ_{11} and ϕ_{12} for branches in the first level and ϕ_{21}, ϕ_{22}, ϕ_{23} and ϕ_{24} for branches in the second level.

Modeling hierarchical classes- corMNL

- The model, which is referred to as corMNL, assigns objects to one of the end nodes using an MNL model whose regression coefficients for class j are represented by the sum of parameters on all the branches leading to that class.
- In the above figure, these coefficients are $\beta_1 = \phi_{11} + \phi_{21}$, $\beta_2 = \phi_{11} + \phi_{22}$, $\beta_3 = \phi_{12} + \phi_{23}$ and $\beta_4 = \phi_{12} + \phi_{24}$ for classes $1, 2, 3$ and 4 respectively.
- Sharing the common terms, ϕ_{11} and ϕ_{12}, introduces prior correlation between the parameters of nearby classes in the hierarchy.

Modeling hierarchical classes- corMNL

- By introducing prior correlations between parameters for nearby classes, the model can better handle situations in which these classes are hard to distinguish.
- If the hierarchy actually does provide information about how distinguishable classes are, the model is expected to perform better.
- This would be especially true when the training set is small and the prior has relatively more influence on the results.
- Using an inappropriate hierarchy will likely lead to worse performance than a standard MNL model, but since the hyperparameters can adapt to reduce the prior correlations to near zero, the penalty may not be large.

Model assessment

- Since MNL is a generalization of logistic regression, to evaluate its goodness-of-fit, we can use generalization of model assessment measures we discussed for logistic regression.
- More specifically, to evaluate significance of $\beta_j = (\beta_{j1}, ..., \beta_{jK})$, we can use the multivariate versions of the likelihood based tests we discussed for logistic regression.

Model selection for prediction

- To compare the performance of MNL models, we can use average log-probability and accuracy rate as explained for simple logistic regression.
- It is also common to use other measurements such as F_1 and precision.
- F_1 is a common measurement in machine learning and is defined as:

$$F_1 = \frac{1}{K} \sum_{k=1}^{K} \frac{2A_k}{2A_k + B_k + C_k}$$

where A_k is the number of cases which are correctly assigned to class k, B_k is the number cases incorrectly assigned to class k, and C_k is the number of cases which belong to the class k but are assigned to other classes.

Model selection for prediction

- While accuracy measurements are based on the top-ranked (i.e., highest probability) category only, *precision* measures the quality of ranking and is defined as follows:

$$precision = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\sum_{k}^{K} I[P(y=k|x_i) \geq P(y=c|x_i)]}\right)$$

  where c is the correct class of test case i.
- The denominator is, therefore, the number of classes with equal or higher rank compared to the correct class.
- For categorical models with hierarchical structures, there are model assessment measures that take the structure into account. Some of these are discussed in Shahbaba and Neal (2007).

Baseline model

- To provide a baseline for interpreting the performance of your model, you can present the performance of a baseline model that ignores the covariates and whose likelihood is solely based on the observed frequency of classes.
- Such model, assigns all test cases to the class with the highest frequency in the training set.

Poisson model

- When the outcome variable, y, represents counts within a specific time period, space limit, or any other index, we usually use the Poisson model.

$$y_i|\mu_i \sim \text{Poisson}(\mu_i)$$

- The systematic components are defined as before: $\eta_i = x_i\beta$.
- The canonical link for this model is the log link:

$$g(\mu_i) = \log(\mu_i) = \eta_i$$

- We therefore have

$$\mu_i = \exp(\eta_i) = \exp(x_i\beta)$$

Poisson model

- The likelihood in terms of β can obtained as follows:

$$\begin{aligned} p(y_i|\mu_i) &\propto \prod_i^n \exp(-\mu_i)\mu_i^{y_i} \\ p(y_i|x_i,\beta) &\propto \prod_i^n \exp[-\exp(x_i\beta)][\exp(x_i\beta)]^{y_i} \end{aligned}$$

- Recall that for the Poisson model, we obtained the following score function and Fisher information

$$\begin{aligned} u_j(\beta) &= \sum_i [y_i - \exp(x_i\beta)]x_{ij} \\ i_{jk}(\beta) &= \sum_i x_{ij}x_{ik}\exp(x_i\beta) \end{aligned}$$

Poisson model

- Similar to logistic regression, we can use either Newton-Raphson algorithm (which is the same as Fisher-scoring algorithm for log link) or iterative weighted least squares.
- For inference about the significance of β, we can use one of the three likelihood base tests.
- The interpretation β_j is that $\exp(\beta_j)$ is the amount increase in the expected value of response variable for one unit increase in x_j when other covariates are fixed.

Poisson model

- When the response variable, y_i, represents the counts over time, space, or any other index, t_i (where t_i could vary from one observation to another) it would be more reasonable to model the adjusted rate of occurrence, μ_i/t_i such that

$$\begin{aligned} \log(\mu_i/t_i) &= x_i\beta \\ \log(\mu_i) - \log(t_i) &= x_i\beta \end{aligned}$$

  where the adjustment term $-\log(t_i)$ is called an *offset*.
- Note that based on this model,

$$\mu_i = t_i \exp(x_i\beta)$$

- Therefor,

$$\hat{\mu}_i = t_i \exp(x_i\hat{\beta})$$

  which would be compared to y_i to calculate the deviance.

Deviance for Poisson

- For Poisson distribution, we have

$$P(y|\mu) = \exp\{\log(\mu)y - \mu - \log(y!)\}$$

- The deviance is therefore,

$$-2\sum_i \{[\log(\hat{\mu}_i) - \log(y_i)]y_i - \hat{\mu}_i + y_i\}$$

- We can write this as

$$2\sum_i \{\log(\frac{y_i}{\hat{\mu}_i})y_i + (\hat{\mu}_i - y_i)\}$$

- A related statistic is called G^2 and is defined as

$$G^2 = 2\sum_i \{y_i \log(\frac{y_i}{\hat{\mu}_i})\}$$

Deviance for Poisson

- The deviance for Poisson model has the following form

$$2 \sum \text{Observed} \times \log(\text{Observed} \ / \ \text{Fitted})$$

- For comparing nested models, the deviance has the following form:

$$2 \sum \text{Observed} \times \log(\text{Fitted using } M_1 \ / \ \text{Fitted using } M_0)$$

- That is,

$$G^2(M_0|M_1) = 2 \sum_i \{y_i \log(\frac{\hat{\mu}_i^{(1)}}{\hat{\mu}_i^{(0)}})\}$$

where $\hat{\mu}_i^{(0)}$ and $\hat{\mu}_i^{(1)}$ are fitted values based on M_0 and M_1 respectively.

Pearson X^2 statistic

- Pearson X^2 has the following form:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- For nested models

$$X^2(M_0|M_1) = \sum_i \frac{(\hat{\mu}_i^{(1)} - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}}$$

- When M_0 holds, both $G^2(M_0|M_1)$ and $X^2(M_0|M_1)$ have asymptotic χ^2 distribution with df equal to the difference between parameters.

Deviance residuals

- For Poisson model, deviance residual for each observation is defined as

$$dr_i = sign(y_i - \mu_i)\sqrt{d_i}$$

where d_i is defined as

$$d_i = 2\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (\hat{\mu}_i - y_i)\}$$

Pearson residual

- Pearson residual is simply defined as

$$pr_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

- Note that

$$X^2 = \sum_i pr_i^2$$

- The standardized Pearson residual is then defined as

$$spr_i = \frac{pr_i}{\sqrt{1-\hat{h}_i}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{h}_i)}}$$

where $\hat{h}_i$ is the leverage.

Other generalized linear models

- So for, we have looked at the four most commonly used GLMs: Normal, logistic, MNL, and Poisson.
- There are other GLMs which are more specialized such as Gamma (and its special case Exponential) and inverse-Gaussian.
- The approach to fit these models is very similar to what we discussed so far.
- We look at Exponential model for example.

Exponential model with log link

- For continuous positive response variables, we can use an Exponential model (or more generally a Gamma model).
- The Exponential distribution has the following form

$$\begin{aligned} P(y|\theta) &\sim \theta \exp(-\theta y), \qquad \theta > 0, \qquad y \geq 0 \\ &\sim \exp[-\theta y + \log(\theta)] \end{aligned}$$

  with mean $\mu = 1/\theta$ and $var(y) = 1/\theta^2$.
- As we can see, the natural parameter $-\theta$ is bounded, whereas $\eta_i = x_i\beta$ is in $\mathcal{R}$.
- Therefore, instead of setting $-\theta_i = \eta_i$, we could set $-\log(\theta_i) = \eta_i$, which is the same as $\log(\mu_i) = \eta_i$.

Exponential model with log link

- Then, we can obtain the score function as follows:

$$u_j(\beta) = \sum_i \frac{[y_i - \exp(x_i\beta)]x_{ij}}{\exp(x_i\beta)}$$

- Note that this has the general form of score function for non-canonical link

$$u_j(\beta) = \sum_i \frac{[y_i - \mu_i]x_{ij}}{var(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i}$$

- The observed Fisher information is

$$i(\beta) = x'wx$$

where w is a diagonal matrix with $w_i = \frac{y_i}{\exp(x_i\beta)}$

- The expected Fisher information is

$$i(\beta) = x'x$$