STATS 212: Generalized Linear Models
Lecture 3: More on Exponential Family, Model Assessment, and Diagnosis

Babak Shahbaba

UCI, Spring 2010

Exponential family and GLM

- Recall that the single parameter exponential family has the following general form:

$$P(y|\mu) = \exp\{g(\mu)T(y) + c(\mu) + h(y)\}$$

- For models with multiple parameters, we have

$$P(y|\mu) = \exp\{\sum_{k=1}^{K} g_k(\mu)T_k(y) + c(\mu) + h(y)\}$$

  where g, T, c, and h are vectors.
- We can change the parameter using the transformation $\phi_k = g_k(\mu)$, and write the distribution in terms of natural parameter ϕ:

$$P(y|\mu) = \exp\{\sum_{k=1}^{K} \phi_k T_k(y) + c^*(\phi) + h(y)\}$$

Exponential family and GLM

- In GLM, the distribution of the random component is a member of the exponential family.
- Therefore, what we previously learned about the exponential family can be extended to GLM.
- The only caveat is that we need to make inference about regression parameters β.
- To do this, we need to take the link function into account.
- We illustrate this for the Poisson regression model. All other models follow a similar process.

Poisson model

- Consider the following Poisson model:

$$\begin{aligned} P(y_i|\mu) &= e^{-\mu_i}\mu_i^{y_i}/y_i! \\ &= \exp\{\log(\mu_i)y_i - \mu_i - \log(y_i!)\} \end{aligned}$$

  where $\phi_i = g(\mu_i) = \log(\mu_i)$, $T(y_i) = y_i$, $c(\mu_i) = -\mu_i$, and $h(y_i) = -\log(y_i!)$.

- We have

$$\begin{aligned} \phi_i = \log(\mu_i) &\Rightarrow \mu_i = \exp(\phi_i) \\ c^*(\phi_i) = \exp(\phi_i) & \\ E_{\phi_i}[T(y_i)] = E(y_i) &= -\frac{\partial c^*(\phi_i)}{\partial \phi_i} = \exp(\phi_i) = \mu_i \\ var_{\phi_i}[T(y_i)] = var(y_i) &= -\frac{\partial^2 c^*(\phi_i)}{\partial \phi_i^2} = \exp(\phi_i) = \mu_i \end{aligned}$$

Poisson model

- The score function with respect to ϕ_i can be obtained as follows:

$$\begin{aligned} u(\phi_i) &= \frac{\partial L(\phi_i)}{\partial \phi_i} \\ &= T(y_i) + \frac{\partial c^*(\phi_i)}{\partial \phi_i} \\ &= y_i - \exp(\phi_i) \\ &= y_i - \mu_i \end{aligned}$$

- The total score function based on n observations is

$$u(\phi) = \sum_i y_i - \exp(\phi_i) = \sum_i y_i - \mu_i$$

- As the result, the likelihood equation is:

$$\sum_i y_i - \exp(\hat{\phi}_i) = \sum_i y_i - \hat{\mu}_i = 0$$

Poisson model

- For Poisson regression model, we are of course interested in regression parameters β.
- Therefore, we would like to write the score function in terms of β.
- To do this, we fist need to specify the link function.
- Suppose we use the log link function

$$g(\mu_i) \quad = \quad \log(\mu_i) = x_i\beta$$

- Since we have $\phi_i = g(\mu_i)$, we can write the link function as follows:

$$\phi_i \quad = \quad \log(\mu_i) = x_i\beta$$

- The link function that transforms the mean to the natural parameter is referred to as the *canonical link*.
- For Poisson model, the log link is the canonical link.

Poisson model

- Using the link function, we can now write the score function in terms of β.
- For the j^{th} element of β, we have

$$\begin{aligned} u(\beta_j) &= \sum_i \frac{\partial L(\beta)}{\partial \beta_j} \\ &= \sum_i \frac{\partial L(\phi)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \beta_j} \\ &= \sum_i [y_i - \exp(x_i \beta)] x_{ij} \end{aligned}$$

- As the result, the likelihood equation in terms of β_j is

$$\sum_i [y_i - \exp(x_i \hat{\beta})] x_{ij} = 0$$

Poisson model

- We can now easily obtain the Fisher information matrix in terms of β.

$$\begin{aligned} i(\beta_j \beta k) &= E[-\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}] \\ &= E[\sum_i x_{ij} x_{ik} \exp(x_i \beta)] \\ &= \sum_i x_{ij} x_{ik} \exp(x_i \beta) \end{aligned}$$

- In a matrix format

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

where w is a diagonal matrix whose i^{th} element is $\exp(x_i \beta)$.

- Moreover,

$$cov(\hat{\beta}) = (x' \hat{w} x)^{-1}$$

Logistic regression

- We can follow the same steps for models with binomial outcome.
- For these models, the logit link is the canonical link

$$\phi_i = g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = x_i\beta$$

- Using the canonical link, we have

$$\begin{aligned}
u(\beta_j) &= \sum_i [y_i - n_i \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}]x_{ij} \\
MLE &: \sum_i [y_i - n_i \frac{\exp(x_i\hat{\beta})}{1+\exp(x_i\hat{\beta})}]x_{ij} = 0 \\
i(\beta) &= x'wx; \qquad \text{where } w \text{ is diagnoal} \\
w_{ii} &= n_i \frac{\exp(x_i\hat{\beta})}{1+\exp(x_i\hat{\beta})} \times \frac{1}{1+\exp(x_i\hat{\beta})} \\
cov(\hat{\beta}) &= (x'\hat{w}x)^{-1}
\end{aligned}$$

Ordinary linear regression

- We can follow the same steps for models with normally distributed outcome.
- For these models, the identity link is the canonical link

$$\phi_i = g(\mu_i) = \mu_i = x_i\beta$$

- Using the canonical link, we have

$$\begin{aligned}
u(\beta_j) &= \sum_i [y_i - x_i\beta]x_{ij} \\
MLE &: \sum_i [y_i - x_i\hat{\beta}]x_{ij} = 0 \\
i(\beta) &= x'wx; \qquad \text{where } w \text{ is diagnoal} \\
w_{ii} &= 1/\sigma^2 \\
cov(\hat{\beta}) &= (x'wx)^{-1} = \sigma^2(x'x)^{-1}
\end{aligned}$$

GLM with canonical link

- In summary, when we use a GLM with canonical link, MLE can be found as follows:

$$\begin{aligned} \sum_i [y_i - \hat{\mu}_i] x_{ij} &= 0 \\ x'y - x'\hat{\mu} &= 0 \\ x'y &= x'\hat{\mu} \end{aligned}$$

- And the covariance matrix of MLE is

$$cov(\hat{\beta}) = (x'\hat{w}x)^{-1}$$

where $\hat{w}$ is a diagonal matrix.

GLM with non-canonical link

- Of course, we do not have to use canonical link.
- In general, if we use the following link

$$m(\mu_i) = x_i\beta = \eta_i = \phi_i$$

then, the score function becomes

$$\begin{aligned} u(\beta_j) &= \sum_i \frac{\partial L(\beta)}{\partial \beta_j} \\ &= \sum_i \frac{\partial L(\phi)}{\partial \phi_i} \frac{\partial \phi_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \end{aligned}$$

- We can show that (Assignment 2) $\partial\phi_i/\partial\mu_i = 1/var(y_i)$. Therefore,

$$u(\beta_j) = \sum_i \frac{\partial L(\phi)}{\partial \phi_i} \frac{1}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

GLM with non-canonical link

- We can therefore write the general form of the score function as follows regardless of whether the link function is canonical or not:

$$u(\beta_j) = \sum_i \frac{[y_i - \mu_i]x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

where $\partial \mu_i / \partial \eta_i$ depends on the link function we choose.

- As the result, the MLE can be found as the solution to the following likelihood equation

$$\sum_i \frac{[y_i - \mu_i]x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

GLM with non-canonical link

- It is easy to show that for a general link function, the Fisher information matrix becomes

$$\begin{aligned} i(\beta_j, \beta_k) &= E\Big(- \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} \Big) \\ &= \sum_i \frac{x_{ij} x_{ik}}{var(y_i)} \big(\frac{\partial \mu_i}{\partial \eta_i} \big)^2 \\ i(\beta) &= x' w x \\ w_{ii} &= \frac{\big(\frac{\partial \mu_i}{\partial \eta_i} \big)^2}{var(y_i)} \end{aligned}$$

- As before,

$$cov(\hat{\beta}) = (x' \hat{w} x)^{-1}$$

We will later discuss different methods to estimate β. For now, we assume we know how to find $\hat{\beta}$ and discuss model assessment.

Model assessment as a decision problem

- In analyzing data, our choice of model is not always obvious, and we need to compare several competing models.
- Model comparison is more appropriately discussed as a special case of decision problems.
- This is specially true in the Bayesian paradigm.
- Decision theory, in general, provides a mathematical framework for making decisions under uncertainty.
- Here, our decision to accept model M_1 over the alternative model M_0 depends not only on the posterior probability of M_1 and M_0, but also on the assumed loss function for such decision.

Model assessment as a decision problem

- When a good loss function is not readily available, using a simple 0-1 loss function (i.e., 0 if we identify the correct model, and 1 if we fail to identify the correct model), simplifies the decision rule such that M_1 is accepted over its corresponding alternative M_0 if M_1 has a higher posterior probability compared to M_0.
- In this case, higher posterior probability correspond to lower posterior risk, and therefore, our decision rule is consistent with the *expected loss* principle: "in deciding between different rules, choose the one with the smallest posterior risk".
- Commonly, researchers avoid expressing prior odds in favor of either M_1 or M_0, and rely on likelihood alone.

Discrepancy measures

- Using likelihood, we can evaluate models based on the distance of the data to alternative models.
- The deviance is a common measure of discrepancy (i.e., lack of fit) between the data and the model (i.e., the lower deviance, the better the model), and it is defined as follows

$$D(y, \mu) = -2 \log[p(y|\mu)] = -2L(\mu, y)$$

- When comparing different models, it is common to use the minimum achievable deviance as the baseline. This corresponds to the deviance of the saturated model where we have one parameter for each observation, i.e., $\mu = y$. The deviance (scaled deviance in Agresti, and McCullagh and Nelder) is then defined as

$$D(y, \mu) \quad = \quad -2[L(\mu, y) - L(y, y)]$$

Discrepancy measures

- Using deviance as a measure of model performance is appealing partly due to its connection to the Kullback-Leibler (KL) divergence measure.
- In the limit of large sample sizes, a model with the lowest K-L divergence has the lowest expected deviance, and thus the highest posterior probability.
- Therefore, when assuming 0-1 loss function, choosing the model with the smallest deviance is consistent with the expected loss principle.

Deviance for exponential family

- In a Frequentist framework, we use maximum likelihood, $\hat{\mu}$ to estimate μ.
- Therefore, the deviance is in fact 2 times the difference between the maximum log-likelihood achievable and the maximum log-likelihood achieved using our model.

$$D(y, \hat{\mu}) \quad = \quad -2[L(\hat{\mu}, y) - L(y, y)]$$

- Note that in this case, the deviance is the same as likelihood ratio statistic test comparing the fitted model to the saturated model.
- In exponential family, the deviance would have the following form:

$$D(\hat{\mu}, y) = -2 \sum_i \{[g(\hat{\mu}) - g(y_i)]\, T(y_i) + c(\hat{\mu}) - c(y_i)\}$$

Deviance for Poisson

- For Poisson distribution, we have

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

- The deviance is therefore,

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

- We can write this as

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

- The second term is usually omitted (it would be zero for a Poisson model with log link that includes intercept), and the deviance reduces to another statistic called G^2

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

Deviance for Bernoulli

- Recall that for the Bernoulli distribution we had

$$P(y|\mu) = \exp\{\log(\frac{\mu}{1-\mu})y + \log(1-\mu)\}$$

- The deviance is therefor,

$$-2\sum_i \{[\log(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}) - \log(\frac{y_i}{1-y_i})]y_i + \log(1-\hat{\mu}_i) - \log(1-y_i)\}$$

- We can write this as

$$2\sum_i \{\log(\frac{y_i}{\hat{\mu}_i})y_i + \log(\frac{1-y_i}{1-\hat{\mu}_i})(1-y_i)\}$$

- Of course, since the values for y_i are either 0 or 1, it would be better to write the deviance as

$$2\sum_i \{-\log(\hat{\mu}_i)I(y_i = 1) - \log(1-\hat{\mu}_i)I(y_i = 0)\}$$

Deviance for Binomial and Poisson

- Alternatively, we can group the data (e.g., by discretizing continuous random variables) and use the deviance based on the binomial distribution to avoid observations with value of y_i equal to 0.
- The deviance for Binomial and Poisson [with log link and intercept] would have the following form:

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

Deviance for normal

- For the normal distribution, we have

$$P(y|\mu,\sigma^2) = \exp\{\frac{-(y-\mu)^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\}$$

- The deviance $D(\hat{\mu}, y)$ for a fixed σ is

$$-2\sum_i\{-(y_i-\hat{\mu})^2/(2\sigma^2)\} = \sum_i\{(y_i-\hat{\mu})^2/(\sigma^2)\}$$

- Note that this is related to the residual sum of squares.
- In in Agresti, and McCullagh and Nelder, $\sum_i\{(y_i-\hat{\mu})^2\}$ is called the deviance and $\sum_i\{(y_i-\hat{\mu})^2/(\sigma^2)\}$ is called the scaled deviance.

Pearson X^2 statistic

- For the normal distribution, the residual sum of squares is equivalent to another commonly used measure of discrepancy called generalized Pearson X^2 statistic defined in general as follows:

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

  where $V(\mu)$ is the estimated variance of y_i given $\hat{\mu}_i$.
- For the normal distribution, $V(\mu) = \sigma^2$; i.e., it is constant.
- for Binomial and Poisson, the original form of Pearson X^2 is used

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

Deviance in GLM

- For generalized linear models, we use the same definition for deviance with $\hat{\mu}_i = g^{-1}(x_i\hat{\beta})$.

Deviance for nested models

- In generalized linear models, we are mainly interested in comparing nested models as opposed to comparing a model of interest to the saturated model.
- We denote the full model, which is a model that includes all covariates, as M_1.
- We denote the reduced model, which is the model similar to the full model except that one or more covariates are removed from the model, as M_0.
- The deviance for M_1 and M_0 are

$$\begin{aligned} D_1 &= -2[L(\hat{\mu}, y, M_1) - L(y, y)] \\ D_0 &= -2[L(\hat{\mu}, y, M_0) - L(y, y)] \end{aligned}$$

- The difference between the two deviance measures is

$$D_0 - D_1 = -2[L(\hat{\mu}, y, M_0) - L(\hat{\mu}, y, M_1)]$$

Deviance for nested models

- This difference is the amount of improvement in the model due to the additional parameters included in M_1.
- Also, not that the difference between deviance measures is the same as likelihood ratio test statistic.
- The asymptotic null distribution of $D_0 - D_1$ is χ^2 with df equal to the difference between the number of parameters in the two nested models.

Deviance for nested Poisson and Binomial

- When comparing nested binomial and Poisson [with log link and intercept] models, the deviance has the following form:

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

Residuals

- While deviance and Pearson X^2 measure the overall discrepancy of the model, we could look at their individual components for each observation separately to identify suspicious values responsible for lack of fit.
- Using the above two statistics, we define two types of residuals: 1) deviance residuals, and 2) Pearson residual.

Deviance residuals

- Deviance residual for each observation is defined as

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

  where d_i is defined such that

$$D(\hat{\mu}, y) = \sum_i d_i$$

- For example, for Poisson distribution,

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

Pearson residual

- Pearson residual is simply defined as

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\mu_i)}}$$

where $V(\mu_i)$ is the estimated variance of y_i given μ_i

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(y_i)}}$$

- Note that

$$X^2 = \sum_i e_i^2$$

Leverage and standardized Pearson residual

- Instead of dividing the residuals by their standard deviation, the above Pearson residual divides them by the estimated $\sqrt{var(y_i)}$.
- To obtain the standardized Pearson residuals, we need to divide the residuals by their asymptotic standard errors, $\sqrt{var(y_i - \hat{\mu}_i)}$, which can be obtained by taking the square root of the diagonal elements of $cov(y - \hat{\mu})$:

$$cov(y - \hat{\mu}) = [cov(y)]^{1/2}[I - H][cov(y)]^{1/2}$$

where

$$H = w^{1/2}x(x'wx)^{-1}xw^{1/2}$$

and w as we discussed before is a diagonal matrix with $w_{ii} = (\partial\mu_i/\partial\eta_i)^2/var(y_i)$.

Leverage and standardized Pearson residual

- The diagonal elements of our estimate of H for each observation is called the *leverage*, denoted as $\hat{h}_i$, for that observation. The greater an observations's leverage, the greater its potential influence on the fit.
- The leverages fall between 0 and 1, and sum to the number of parameters in the model.
- As we can see, the value of $\hat{h}_i$ depends on the value of covariates as well as the model fit. Therefore, just having an extreme value for an observation does not necessarily mean that the observation has high leverage.
- The standardized Pearson residual is then defined as

$$\begin{aligned} r_i &= \frac{y_i - \hat{\mu}_i}{\sqrt{var(y_i)(1-\hat{h}_i)}} \\ &= \frac{e_i}{\sqrt{1-\hat{h}_i}} \end{aligned}$$