STAT5144: Topics in Regression

Inyoung Kim
Outline

1. Moments, asymptotic covariance
2. Inference for GLMs
Moments and Likelihood for GLM

- Exponential dispersion family

\[ f(y_i; \theta_i, \phi) = \exp[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi) \]

This is called the exponential dispersion family and \( \phi \) is called the dispersion parameter. The parameter \( \theta_i \) is the natural parameter.

- When \( \phi \) is known, it simplifies to the form for the natural exponential family, which is

\[ f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)] \]

- We identify \( Q(\theta) \) here with \( \theta/a(\phi) \), \( a(\theta) \) with \( \exp(-b(\theta)/a(\phi)) \), \( b(y) \) with \( \exp(c(y, \phi)) \)

- \( a(\phi) \) has a form \( a(\phi) = \phi/w_i \) for a known weight \( w_i \)
Moments and Likelihood for GLM

- \( L = \sum_i L_i \), where \( L_i = \log f(y_i; \theta_i, \phi) \)
  
  \[
  L_i = \left[ y_i \theta_i - b(\theta_i) \right] / a(\phi) + c(y_i, \phi)
  \]

- \[
  \frac{\partial L_i}{\partial \theta_i} = \left[ y_i - b'(\theta_i) \right] / a(\phi)
  \]

- \[
  \frac{\partial^2 L_i}{\partial \theta_i^2} = -b''(\theta_i)] / a(\phi)
  \]

  where \( b'(\theta) \) and \( b''(\theta_i) \) denote the first two derivatives of \( b(\cdot) \) evaluated at \( \theta_i \)

- We apply the general likelihood results
  
  \[
  E\left( \frac{\partial L}{\partial \theta} \right) = 0
  \]

  \[
  -E\left( \frac{\partial^2 L}{\partial \theta^2} \right) = E^2\left( \frac{\partial L}{\partial \theta} \right)
  \]

  under the regularity conditions (Cox, and Hinkley, 1974, Sec 4.8).
Moments and Likelihood for GLM

Since \( E[Y_i - b'(\theta_i)]/a(\phi) = 0 \),

\[
\mu_i = E(Y_i) = b'(\theta_i)
\]

Since \( b''(\theta)/a(\phi) = E[(Y_i - b'(\theta))/a(\phi)]^2 = \text{Var}(Y_i)/[a(\phi)]^2 \),

\[
\text{var}(Y_i) = b''(\theta_i)a(\phi)
\]
The likelihood equations

\[
\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0
\]

Use chain-rule

\[
\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}
\]

Since \( \frac{\partial L_i}{\partial \theta_i} = \left[ y_i - b'(\theta_i) \right]/a(\phi) \) and since \( \mu_i = b'(\theta_i) \) and \( \text{var}(Y_i) = b''(\theta_i)a(\phi) \),

\[
\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a(\phi)} \quad \text{and} \quad \frac{\partial \mu_i}{\partial \theta_i} = \frac{\text{var}(Y_i)}{a(\phi)}
\]
Moments and Likelihood for GLM

Also, since $\eta_i = \sum_j \beta_j x_{ij}$, \( \partial \eta_i / \partial \beta_j = x_{ij} \)

Since $\eta_i = g(\mu_i)$, $\partial \mu_i / \partial \eta_i$ depends on the link function for the model.

\[
\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i) x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}
\]

The likelihood equations are

\[
\sum_i^N \frac{(y_i - \mu_i) x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0
\]

which is $u_j$. 
Asymptotic covariance matrix of model parameter estimators

- The likelihood function for the GLM also determines the asymptotic covariance matrix of the ML estimator \( \hat{\beta} \).
- This matrix is the inverse of the information matrix \( \mathcal{I} \), which has element \( E(-\partial^2 L(\beta)/\partial \beta_h \partial \beta_j) \).
- To find this,
  \[
  E\left( \frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j} \right) = -E\left( \frac{\partial L_i}{\partial \beta_h} \right) \left( \frac{\partial L_i}{\partial \beta_j} \right)
  \]
  which holds for exponential families.
- Thus
  \[
  E\left( \frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j} \right) = -E\left[ \frac{(Y_i - \mu_i)x_{ih}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(Y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \right] = \frac{-x_{ih}x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2
  \]
- Since \( L(\beta) = \sum_i L_i \)
  \[
  E(-\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j}) = \sum_{i=1}^{N} \frac{x_{ih}x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2
  \]
Asymptotic covariance matrix of model parameter estimators

- Generalizing from this typical element to the entire matrix, the information matrix has the form
  \[ J = X'WX \]
  where \( W \) is the diagonal matrix with main-diagonal elements
  \[ w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i) \]

- The asymptotic covariance matrix of \( \hat{\beta} \) is estimated by
  \[ \text{cov}(\hat{\beta}) = J^{-1} = (X' \hat{\mathcal{W}}X)^{-1} \]
  where \( \hat{\mathcal{W}} \) is \( W \) evaluated at \( \hat{\beta} \).

- The form of \( W \) also depends on the link function
  - For Poisson, since \( w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i) = \mu_i \), the estimated covariance matrix of \( \hat{\beta} \) is \((X' \hat{\mathcal{W}}X)^{-1}\), where \( \hat{\mathcal{W}} \) is the diagonal matrix with elements of \( \hat{\mu} \) on the main diagonal.
The saturated GLM has a separate parameter for each observation.

- It gives a perfect fit.
- This sounds good, but it is not a helpful model.
- It does not smooth the data or have the advantage that a simpler model has.
- It serves as a baseline for other models, such as for checking model fit.
A saturated model explains all variation by the systematic component of the model.

Let \( \tilde{\theta} \) denote the estimate of \( \theta \) for the saturated model, corresponding to estimated means \( \tilde{\mu}_i = y_i \) for all \( i \).

For a particular unsaturated model, denote the corresponding ML estimates by \( \hat{\theta} \) and \( \hat{\mu}_i \).

For maximized log likelihood \( L(\hat{\mu}; y) \) for that model and maximized log likelihood \( L(y; y) \) in the saturated case,

\[
-2[L(\hat{\mu}; y) - L(y; y)] = -2\log \frac{\text{maximum likelihood for model}}{\text{maximum likelihood for saturated model}}
\]

describes lack of fit.

It is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the alternative that a more general model holds

\[
-2[L(\hat{\mu}; y) - L(y; y)] = 2 \sum_i [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2 \sum_i [y_i \tilde{\theta}_i - b(\hat{\theta}_i)]/a(\phi)
\]

Usually, \( a(\phi) \) has the form \( a(\phi) = \phi/\omega_i \), and this statistic equals

\[
2 \sum_i \omega_i [y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(y; \hat{\mu})/\phi
\]

This is called the scaled deviance and \( D(y; \hat{\mu}) \) is called the deviance.

The greater the scaled deviance, the poor the fit.
Deviance residual

- When a GLM fits poorly according to an overall goodness-of-fit test, examination of residuals highlights where the fit is poor. One type of residual uses components of the deviances \( D(y; \hat{\mu}) = \sum d_i \), where

\[
d_i = 2\omega_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]
\]

- The deviance residual for observation \( i \) is

\[
\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i)
\]
Deviance for Poisson Models

- $\hat{\theta}_i = \log \hat{\mu}_i$ and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$
- Similarly, $\tilde{\theta}_i = \log y_i$ and $b(\tilde{\theta}_i) = y_i$ for the saturated model
- $a(\phi) = 1$
- The deviance and scaled deviance

$$D(y; \hat{\mu}) = 2 \sum_i [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i]$$

- When a model with log link contains an intercept term, the likelihood equation implied by that parameter is $\sum y_i = \sum \hat{\mu}_i$
- The deviance simplifies to

$$D(y; \hat{\mu}) = 2 \sum y_i \log(y_i/\hat{\mu}_i)$$

- For some GLMS the scaled deviance has an approximate chi-squared distribution.
For two-way contingency tables, this reduces to the $G^2$ statistic substituting cell count $n_{ij}$ for $y_i$ and the independence fitted value $\hat{\mu}_{ij}$ for $\hat{\mu}_i$.

For a Poisson or multinomial model applied to a contingency table with a fixed number of cells N, the deviance has an approximate chi-squared distribution for large $\{\mu_i\}$. 

---

**Deviance for Poisson Models**

- For two-way contingency tables, this reduces to the $G^2$ statistic substituting cell count $n_{ij}$ for $y_i$ and the independence fitted value $\hat{\mu}_{ij}$ for $\hat{\mu}_i$.
- For a Poisson or multinomial model applied to a contingency table with a fixed number of cells N, the deviance has an approximate chi-squared distribution for large $\{\mu_i\}$. 

---
Deviance for Binomial Models

- Consider binomial GLMs with sample proportions \( \{y_i\} \) based on \( \{n_i\} \) trials.
- \( \hat{\theta} = \log[\hat{\pi}_i/(1 - \hat{\pi}_i)] \)
- \( b(\hat{\theta}_i) = \log[1 + \exp(\hat{\theta}_i)] = -\log(1 - \hat{\pi}_i) \)
- \( \tilde{\theta}_i = \log[y_i/(1 - y_i)] \) and \( b(\tilde{\theta}_i) = -\log(1 - y_i) \) for the saturated model
- \( a(\phi) = 1 \) and \( \omega_i = n_i \)
Deviance for Binomial Models

The deviance equals

\[ 2 \sum_i n_i \left\{ y_i \left( \log \frac{y_i}{1 - y_i} - \log \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \log(1 - y_i) - \log(1 - \hat{\pi}_i) \right\} \]

\[ = 2 \sum_i n_i y_i \log \frac{n_i y_i}{n_i - n_i y_i} - 2 \sum_i n_i y_i \log \frac{n_i \hat{\pi}_i}{n_i - n_i \hat{\pi}_i} + 2 \sum_i n_i \log \frac{1 - y_i}{1 - \hat{\pi}_i} \]

\[ = 2 \sum_i n_i y_i \log \frac{n_i y_i}{n_i \hat{\pi}_i} + 2 \sum_i (n_i - n_i y_i) \log \frac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i} \]
Deviance for Binomial Models

- At setting $i$, $n_i$, $y_i$ is the number of successes and $(n_i - n_i y_i)$ is the number of failure, $i = 1, \ldots, N$.
- The deviance is a sum over the $2N$ cells of successes and failures and has the same form,

$$D(y; \hat{\mu}) = 2 \sum \text{observed} \times \log(\frac{\text{observed}}{\text{fitted}})$$

as the deviance for Poisson loglinear models with intercept term.
With the binomial responses, it is possible to construct the data file as expressed here with the counts of successes and failures at each setting for the predictors, or with the individual Bernoulli 0-1 observations at the subject level. The deviance differs in the two cases. In the first case the saturated model has a parameter at each setting for the predictors, whereas in the second case it has a parameter for each subject. We refer to these as grouped data and ungrouped data cases.
Likelihood for Saturated model

\[ l(\text{saturated\ model}) = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{1-y_i} = 1 \]
For a Poisson or binomial model $M$, $\phi = 1$, so the deviance equals

$$D(y; \hat{\mu}) = -2[L(\hat{\mu}; y) - L(y; y)]$$  \hspace{1cm} (1)$$

Consider two models, $M_0$ with fitted values $\hat{\mu}_0$ and $M_1$ with fitted values $\hat{\mu}_1$, with $M_0$ a special case of $M_1$.

Model $M_0$ is said to be nested within $M_1$.

Since $M_0$ is simpler than $M_1$, a smaller set of parameter values satisfies $M_0$ than satisfies $M_1$.

Maximizing the log likelihood over a smaller space can not yield a larger maximum.

Thus, $L(\hat{\mu}_0; y) \leq L(\hat{\mu}_1; y)$ and it follows from (1) with the same $L(y; y)$ for each model that

$$D(y; \hat{\mu}_1) \leq D(y; \hat{\mu}_0)$$

Simpler models have larger deviance.
Assuming that model $M_1$ holds, the likelihood-ratio test of the hypothesis than $M_0$ holds uses the test statistic

$$-2[L(\hat{\mu}_0; y) - L(\hat{\mu}_1; y)]$$

$$= -2[L(\hat{\mu}_0; y) - L(y; y)] - \{ -2[L(\hat{\mu}_1; y) - L(y; y)] \}$$

$$= D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1)$$

The likelihood-ratio statistic comparing the two models is simply the difference between the deviances.

This statistic is large when $M_0$ fits poorly compared to $M_1$

Since the part involving the saturated model cancels, the difference between deviances

$$D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) = 2 \sum \omega_i [y_i(\hat{\theta}_1i - \hat{\theta}_0i) - b(\hat{\theta}_1i) + b(\hat{\theta}_0i)]$$

also has the form of the deviance.

Under regularity conditions, this difference has approximately a chi-squared null distribution with df equal to the difference between the numbers of parameters in the two models.
For binomial GLMs and Poisson loglinear GLMs with intercept, from expression for the deviance, the difference in deviances uses the observed counts and the two sets of fitted values in the form

\[ D(y; \hat{\mu}_0 - D(y; \hat{\mu}_1)) = 2 \sum \text{observed} \times \log\left(\frac{\text{fitted}_1}{\text{fitted}_0}\right) \]

With binomial response, the test compared models does not depend on whether the data file has grouped or ungrouped form.

The saturated model differs in the two cases, but its log likelihood cancels when one forms the difference between the deviances.
For binomial GLMs and Poisson loglinear GLMs with intercept, from expression for the deviance, the difference in deviances uses the observed counts and the two sets of fitted values in the form

\[ D(y; \hat{\mu}_0 - D(y; \hat{\mu}_1)) = 2 \sum \text{observed} \times \log(\text{fitted}_1 / \text{fitted}_0) \]

With binomial response, the test compared models does not depend on whether the data file has grouped or ungrouped form.

The saturated model differs in the two cases, but its log likelihood cancels when one forms the difference between the deviances.