# Lecture 5: GLM: Inference

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

# General process and logic

The process and logic can be summarized as follows:

1. Specify a model $M_0$ corresponding to $H_0$. Specify a more general model $M_1$ (with $M_0$ as a special case of $M_1$).|

2. Fit $M_0$ and calculate the goodness of fit statistic $G_0$. Fit $M_1$ and calculate the goodness of fit statistic $G_1$.
3. Calculate the improvement in fit, usually $G_1 - G_0$ but $G_1/G_0$ is another possibility.
4. Use the sampling distribution of $G_1 - G_0$ (or some related statistic) to test the null hypothesis that $G_1 = G_0$ against the alternative hypothesis $G_1 \neq G_0$.
5. If the hypothesis that $G_1 = G_0$ is not rejected, then $H_0$ is not rejected and $M_0$ is the preferred model. If the hypothesis $G_1 = G_0$ is rejected, then $H_0$ is rejected and $M_1$ is regarded as the better model.

The basic idea is that under appropriate conditions, if $S$ is a statistic of interest, then approximately

$$\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{var}(S)}} \sim \mathrm{N}(0,1)$$

or equivalently

$$\frac{[S - \mathrm{E}(S)]^2}{\mathrm{var}(S)} \sim \chi^2(1),$$

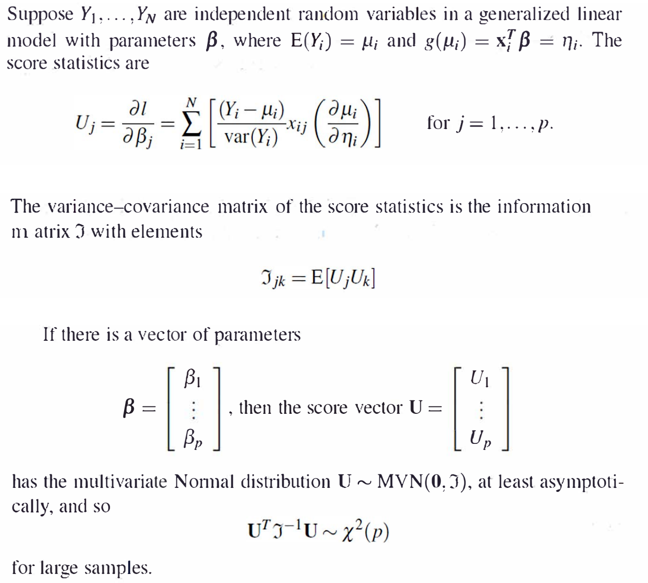
where $\mathrm{E}(S)$ and $\mathrm{var}(S)$ are the expectation and variance of $S$, respectively.

If there is a vector of statistics of interest $\mathbf{s} = \begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$ with asymptotic expectation $\mathrm{E}(\mathbf{s})$ and asymptotic variance–covariance matrix $\mathbf{V}$, then approximately

$$[\mathbf{s} - \mathrm{E}(\mathbf{s})]^T \mathbf{V}^{-1} [\mathbf{s} - \mathrm{E}(\mathbf{s})] \sim \chi^2(p),$$

provided $\mathbf{V}$ is non-singular so a unique inverse matrix $\mathbf{V}^{-1}$ exists.

# Sampling distribution of MLEs



Suppose $Y_1,\ldots,Y_N$ are independent random variables in a generalized linear model with parameters $\boldsymbol{\beta}$, where $\mathrm{E}(Y_i) = \mu_i$ and $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$. The score statistics are

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N}\left[\frac{(Y_i-\mu_i)}{\mathrm{var}(Y_i)}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \qquad \text{for } j=1,\ldots,p.$$

The variance–covariance matrix of the score statistics is the information matrix $\mathfrak{I}$ with elements

$$\mathfrak{I}_{jk} = \mathrm{E}[U_jU_k]$$

If there is a vector of parameters

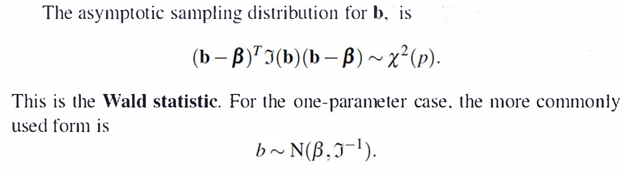
$$\boldsymbol{\beta} = \begin{bmatrix}\beta_1\\ \vdots \\ \beta_p\end{bmatrix}, \text{ then the score vector } \mathbf{U} = \begin{bmatrix}U_1\\ \vdots \\ U_p\end{bmatrix}$$

has the multivariate Normal distribution $\mathbf{U} \sim \mathrm{MVN}(\mathbf{0},\mathfrak{I})$, at least asymptotically, and so

$$\mathbf{U}^T\mathfrak{I}^{-1}\mathbf{U} \sim \chi^2(p)$$

for large samples.

# Sampling distribution of scoring statistics



The asymptotic sampling distribution for **b**, is

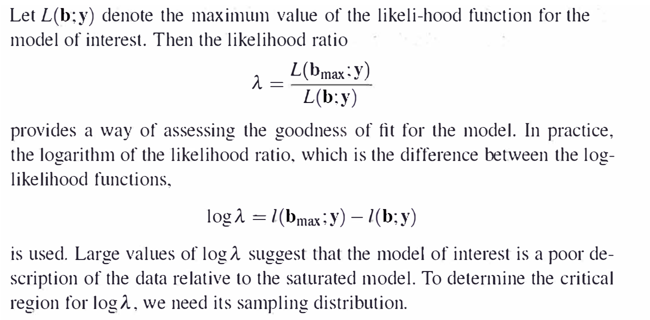
$$(\mathbf{b}-\boldsymbol{\beta})^T \Im(\mathbf{b})(\mathbf{b}-\boldsymbol{\beta}) \sim \chi^2(p).$$

This is the **Wald statistic**. For the one-parameter case, the more commonly used form is

$$b \sim \mathrm{N}(\beta, \Im^{-1}).$$

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated. This is called a **saturated model**. It is a generalized linear model with the same distribution and same link function as the model of interest.

# Sampling distribution for the deviance



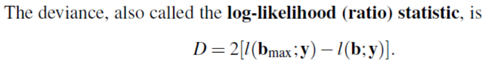
Let $L(\mathbf{b};\mathbf{y})$ denote the maximum value of the likeli-hood function for the model of interest. Then the likelihood ratio

$$\lambda = \frac{L(\mathbf{b}_{\max};\mathbf{y})}{L(\mathbf{b};\mathbf{y})}$$

provides a way of assessing the goodness of fit for the model. In practice, the logarithm of the likelihood ratio, which is the difference between the log-likelihood functions,

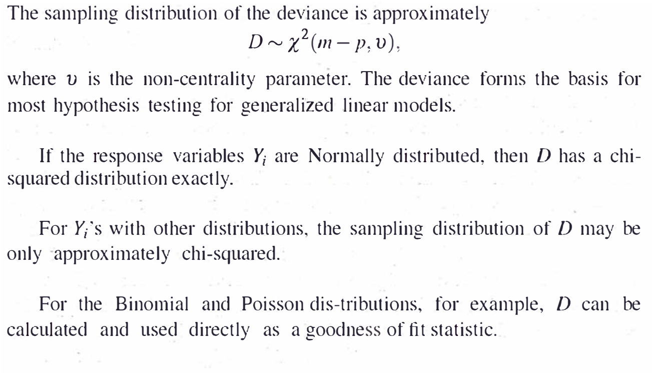
$$\log\lambda = l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})$$

is used. Large values of $\log\lambda$ suggest that the model of interest is a poor description of the data relative to the saturated model. To determine the critical region for $\log\lambda$, we need its sampling distribution.



The deviance, also called the **log-likelihood (ratio) statistic**, is

$$D = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})].$$



The sampling distribution of the deviance is approximately

$$D \sim \chi^2(m-p, \upsilon),$$

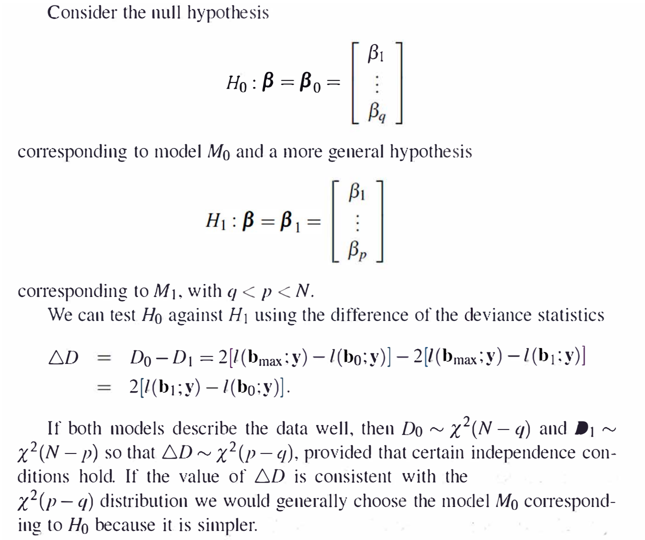
where $\upsilon$ is the non-centrality parameter. The deviance forms the basis for most hypothesis testing for generalized linear models.

If the response variables $Y_i$ are Normally distributed, then $D$ has a chi-squared distribution exactly.

For $Y_i$'s with other distributions, the sampling distribution of $D$ may be only approximately chi-squared.

For the Binomial and Poisson dis-tributions, for example, $D$ can be calculated and used directly as a goodness of fit statistic.

# Hypothesis testing



Consider the null hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

corresponding to model $M_0$ and a more general hypothesis

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

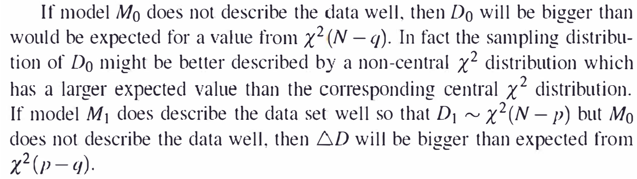
corresponding to $M_1$, with $q < p < N$.

We can test $H_0$ against $H_1$ using the difference of the deviance statistics

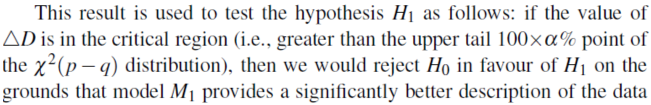
$$\begin{aligned} \triangle D &= D_0 - D_1 = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})] - 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_1;\mathbf{y})] \\ &= 2[l(\mathbf{b}_1;\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})]. \end{aligned}$$

If both models describe the data well, then $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$ so that $\triangle D \sim \chi^2(p-q)$, provided that certain independence conditions hold. If the value of $\triangle D$ is consistent with the $\chi^2(p-q)$ distribution we would generally choose the model $M_0$ corresponding to $H_0$ because it is simpler.

# Hypothesis testing



If model $M_0$ does not describe the data well, then $D_0$ will be bigger than would be expected for a value from $\chi^2(N-q)$. In fact the sampling distribution of $D_0$ might be better described by a non-central $\chi^2$ distribution which has a larger expected value than the corresponding central $\chi^2$ distribution. If model $M_1$ does describe the data set well so that $D_1 \sim \chi^2(N-p)$ but $M_0$ does not describe the data well, then $\triangle D$ will be bigger than expected from $\chi^2(p-q)$.



This result is used to test the hypothesis $H_1$ as follows: if the value of $\triangle D$ is in the critical region (i.e., greater than the upper tail $100\times\alpha\%$ point of the $\chi^2(p-q)$ distribution), then we would reject $H_0$ in favour of $H_1$ on the grounds that model $M_1$ provides a significantly better description of the data