# STAT347: Generalized Linear Models
## Lecture 3

> Today's topics: Chapters 4.3-4.4
>
> - Hypothesis testing for $\beta$
>
> - Deviance analysis of a GLM

# 1 Wald, likelihood-ratio and score tests

In last lecture, we have mentioned that when $n$ is large

$$\hat{\beta} - \beta_0 \overset{\cdot}{\sim} N(0, V_{\beta_0})$$

How to test

$$H_0 : A\beta_0 = a_0 \quad V.S. \quad H_1 : A\beta_0 \neq a_0$$

## 1.1 Wald test

Test statistics:

$$T = (A\hat{\beta} - a_0)^T \left[ \widehat{\mathrm{Var}}(A\hat{\beta}) \right]^{-1} (A\hat{\beta} - a_0)$$

- $\widehat{\mathrm{Var}}(A\hat{\beta}) = AV_{\hat{\beta}} A^T$

- If $a_0 \in \mathbb{R}^1$, Wald statistic can also be written as

$$z = \frac{A\hat{\beta} - a_0}{\sqrt{\widehat{\mathrm{Var}}(A\hat{\beta})}}$$

- Under $H_0$, Wald statistic $z \overset{\cdot}{\sim} N(0,1)$

- We can also obtain a 95% CI for $A\beta_0$ as $[A\hat{\beta} - 1.96\sqrt{\widehat{\mathrm{Var}}(A\hat{\beta})}, A\hat{\beta} + 1.96\sqrt{\widehat{\mathrm{Var}}(A\hat{\beta})}]$

- When $a_0 \in \mathbb{R}^d$, then under $H_0$, $T \overset{\cdot}{\sim} \mathcal{X}_d^2$

- This is the GLM R function output for the analysis of each component $\beta_j$

## 1.2 A potential issue with Wald test

Let's look at an example of using Wald test for Binomial data $y_i \sim \mathrm{Binomial}(n_i, p_i)$ where we work on the null model:

$$\log \frac{p_i}{1 - p_i} = \log \frac{\mu_i}{n_i - \mu_i} = \beta_0$$

- As we use a canonical link, the asymptotic variance is $V_{\beta_0} = (X^T W X)^{-1}$ where $W = D^2 V^{-1} = D/a(\phi) = D$ (Lecture 2, section 2.2 and $a(\phi) = 1$ for Binomial data).

- $D_{ii} = \frac{1}{g'(\mu_i)} = \mu_i(n_i - \mu_i)/n_i$

- An estimate $\widehat{V}_{\beta_0} = V_{\hat{\beta}} = [(\sum_i n_i)\hat{p}(1-\hat{p})]^{-1}$ where $\hat{p}_i = \hat{p} = e^{\hat{\beta}}/(1 + e^{\hat{\beta}})$

- If we are interested in testing $H_0 : p_i \equiv 0.5$ or equivalently $H_0 : \beta_0 = 0$, the Wald statistics is

$$ z = \hat{\beta} \sqrt{(\sum_i n_i)\hat{p}(1-\hat{p})} $$

- If we only have one sample with $y = 23$ and $n = 25$, then $z = 11$. If $y = 24$ and $n = 25$ then $z = 9.7$. Why do we have a smaller $z$ when we have stronger evidence against the null?

- In the above specific example with only one sample, we can also obtain the CLT of $\hat{p} = y/n$, which result in another Wald statistics

$$ z = \frac{\hat{p} - 0.5}{\sqrt{\hat{p}(1-\hat{p})/n}}. $$

So the Wald statistics is not unique and depends on parameterization.

- We will discuss this more when we learn binary GLM (Chapter 5.3.3)

## 1.3   Score test

We only discuss the simple case

$$ H_0 : \beta = \beta_0 \in \mathbb{R}^p \quad V.S. \quad H_1 : \beta \neq \beta_0 $$

Last time we used the property of the likelihood that:

$$ \text{Var}\left(\dot{L}(\beta_0)\right) = \mathbb{E}\left(\left(\frac{\partial L}{\partial \beta}|_{\beta=\beta_0}\right)^2\right) = -\mathbb{E}\left(\ddot{L}(\beta_0)\right) $$

where $\beta_0$ is the true value of the parameter. We construct the test statistics:

$$ T = -\dot{L}(\beta_0)^T \left(\ddot{L}(\beta_0)\right)^{-1} \dot{L}(\beta_0) $$

We make use of the asymptotic normal distribution of $\dot{L}(\beta_0)$. Under $H_0$, we have $T \to \mathcal{X}_p^2$ when $n \to \infty$.

## 1.4   Likelihood ratio test

We test for the null

$$ H_0 : A\beta_0 = a_0 \quad V.S. \quad H_1 : A\beta_0 \neq a_0 $$

where $a_0 \in \mathbb{R}^d$. The likelihood ratio test statistics is

$$ -2\log\Lambda = -2\left(L(\tilde{\beta}) - L(\hat{\beta})\right) $$

where $\tilde{\beta}$ is the MLE of $\beta$ under the constraint $A\beta = a_0$, and $\hat{\beta}$ is our original MLE of $\beta$ without any constraint. As $n \to \infty$, under $H_0$

$$-2 \log \Lambda \to \mathcal{X}_d^2$$

- Relationship among the three tests: Agresti Chapter 4.3.4

- Construct CI: invert tests (illustrate more in later lectures)

## 2    Deviance analysis

Remember that in linear regression, we use $R^2$, defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{\mu}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{\mu}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

to evaluate how well the model fits the data. We have an analogy in GLM, which is the deviance analysis.

### 2.1    Definition (more general than the textbook)

Consider density function $f(y; \theta) = e^{\frac{y\theta - b(\theta)}{a(\phi)}} f_0(y; \phi)$ at two values $\theta_1$ and $\theta_2$. Measure the "distance" between two distributions:

$$D(\theta_1, \theta_2) = 2\mathbb{E}_{\theta_1} \left\{ \log \frac{f(y; \theta_1)}{f(y; \theta_2)} \right\} = 2\mathbb{E}_{\theta_1} \left\{ y(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2) \right\} / a(\phi)$$

$$= 2 \left[ \mu_1(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2) \right] / a(\phi)$$

Remember the 1-to-1 mapping between $\mu$ and $\theta$, we also write $D(\mu_1, \mu_2) = D(\theta_{\mu_1}, \theta_{\mu_2})$

- $D(\mu_1, \mu_2) \geq 0$ and the equality holds only when $\mu_1 = \mu_2$

- Generally, $D(\mu_1, \mu_2) \neq D(\mu_2, \mu_1)$

- KL divergence: $D(\mu_1, \mu_2)/2$

- If $f$ is the normal density, then $D(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2 / \sigma^2$

Saturated model: imagine the case that we collect an infinite number of covariates, then we can perfectly fit the data and obtain $\hat{\mu}_i = y_i$ for all samples. Then this is called a saturated model.

Deviance between the saturated model (saturated when there is only one observation $y$): $\hat{\mu} = y$ and another model with $\mu$:

$$D(y, \mu) = 2 \left[ y(\theta_y - \theta) - b(\theta_y) + b(\theta) \right] / a(\phi)$$
$$= -2 \log \left[ f(y, \theta) / f(y, \theta_y) \right]$$

With samples $(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)$, the total deviance in GLM (the deviance definition in the text book)

$$D_+(y, \hat{\mu}) = \sum_i D(y_i, \hat{\mu}_i)$$

$$= -2 \sum_i \log \left[ f(y_i, \hat{\theta}_i) / f(y_i, \theta_{y_i}) \right]$$

This is also called the residual deviance, and compares the estimated GLM model with the saturated model Null deviance:

$$\sum_i D(y_i, \bar{y})$$

where $\bar{y} = \sum_i y_i/n$. The null deviance compares the null model ($\mu_i \equiv \mu$) with the saturated model.

## 2.2    Deviance analysis for nested models

Let $\beta = \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \end{pmatrix}$ where $\beta^{(1)} \in \mathbb{R}^{p_1}$ and $X = \begin{pmatrix} X^{(1)} & X^{(2)} \end{pmatrix}$.

We call $\mathcal{M}^{(1)}$ with

$$g(\mu_i) = X^{(1)}\beta^{(1)}$$

a nested model of the full model $\mathcal{M}$ where

$$g(\mu_i) = X\beta.$$

Let $\hat{\beta}^{(1)}$ be the MLE solution of the model $\mathcal{M}^{(1)}$ and $\hat{\mu}^{(1)}$ be the corresponding estimated expectations of $y$ in the fitted model.

Then,

$$D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}) = -2\left[L(\hat{\beta}^{(1)}) - L(\hat{\beta})\right]$$

is the likelihood ratio between two models.

- Test for $H_0 : \beta^{(2)} = 0$. Under $H_0$,

$$D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}) \to \mathcal{X}^2_{p-p_1}$$

- Compare with the null model, we can also define "$R^2$" in GLM:

$$1 - \frac{D_+(y, \hat{\mu})}{\sum_i D(y_i, \bar{y})}$$

## 2.3    Model comparison with deviance analysis table

Say we partition our covariates as

$$X = (1, X_{(1)}, X_{(2)}, \cdots, X_{(J)})$$

and $X_{(j)} \in \mathbb{R}^{d_j}$. We can sequentially add each partition of covariates into the model (in some pre-determined order) and understand each partition's "relative contribution" with a deviance analysis table.

Define the following quantities:

- $\hat{\beta}^{(j)}$ is the MLE solution of the GLM model with covariates $X^{(j)} = (1, X_{(1)}, X_{(2)}, \cdots, X_{(j)})$

- $\hat{\mu}^{(j)}$ is the corresponding vector of expectations of $y = (y_1, \cdots, y_n)$ in the fitted model.

| Model | twice log-likelihood | residual deviance | difference | df | Compare with |
|-------|---------------------|-------------------|------------|-----|--------------|
| $\hat{\beta}^{(0)}$ (null) | $2L(\hat{\beta}^{(0)})$ | $D_+(y, \hat{\mu}^{(0)}) = \sum_i D(y_i, \bar{y})$ | | | |
| $\hat{\beta}^{(1)}$ | $2L(\hat{\beta}^{(1)})$ | $D_+(y, \hat{\mu}^{(1)})$ | $D_+(y, \hat{\mu}^{(0)}) - D_+(y, \hat{\mu}^{(1)})$ | $d_1$ | $\chi^2_{d_1}$ |
| $\hat{\beta}^{(2)}$ | $2L(\hat{\beta}^{(2)})$ | $D_+(y, \hat{\mu}^{(2)})$ | $D_+(y, \hat{\mu}^{(1)}) - D_+(y, \hat{\mu}^{(2)})$ | $d_2$ | $\chi^2_{d_2}$ |
| $\vdots$ | | | | | |
| $\hat{\beta}^{(J)}$ | $2L(\hat{\beta}^{(J)})$ | $D_+(y, \hat{\mu}^{(J)})$ | $D_+(y, \hat{\mu}^{(J-1)}) - D_+(y, \hat{\mu}^{(J)})$ | $d_J$ | $\chi^2_{d_J}$ |

Table 1: Deviance analysis table.

Then the deviance analysis table is shown in Table 1.

The difference of two residual deviances

$$D_+(y, \hat{\mu}^{(j-1)}) - D_+(y, \hat{\mu}^{(j)}) = 2L(\hat{\beta}^{(j)}) - 2L(\hat{\beta}^{(j-1)})$$

so that we can use the likelihood ratio test.

Next time: Chapters 4.4.6, 4.5 and 4.7, residuals, computation and data examples