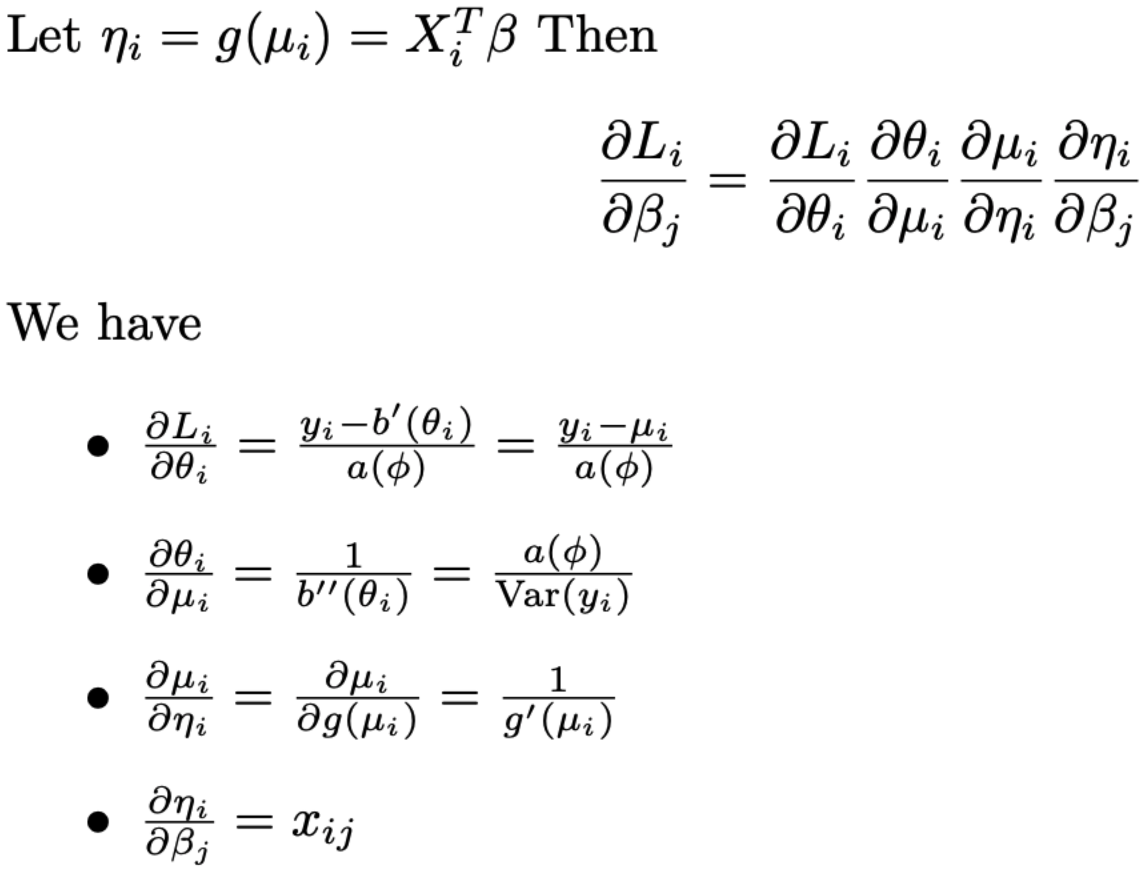
Lecture 3
Statistical estimation and hypothesis testing for exponential family GLM

Today’s topics:

- Likelihood score equation for general link
- Asymptotic distribution of the MLE estimates
- Hypothesis testing for β
- Reading: Agresti Chapter 4.3, Faraway Chapter 8.3

Likelihood score equation for a general link



Let $\eta_i = g(\mu_i) = X_i^T \beta$ Then

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

We have

- $\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{a(\phi)}{\mathrm{Var}(y_i)}$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial g(\mu_i)} = \frac{1}{g'(\mu_i)}$
- $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Likelihood score equation for a general link

- The score equations can be written as

$$\boxed{\frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(y_i)} \frac{1}{g'(\mu_i)} = 0}$$

  - μ_i and Var(y_i) are both functions of $\beta = (\beta_1, \cdots, \beta_p)$
  - The score equations only depend on the mean and variance of y_i
  - Matrix form of the score equation:

$$\dot{L}(\beta) = X^T D V^{-1}(y - \mu) = 0$$

where $V = \text{diag}(\text{Var}(y_1), \cdots, \text{Var}(y_n))$ and $D = \text{diag}(g'(\mu_1), \cdots, g'(\mu_n))^{-1}$,
$y = (y_1, \cdots, y_n)$ and $\mu = (\mu_1, \cdots, \mu_n)$.

  - L is not necessarily a concave function of β

Likelihood score equation for a general link

Special cases

- If the link function is the canonical link, then $D = \frac{1}{a(\phi)} V$, thus the score equation becomes

  $$\frac{1}{a(\phi)} X^T (y - \mu) = 0$$

  the same as we derived earlier

- If we assume that $g(\mu_i) = \mu_i = X_i^T \beta$, then the estimating (score) equation becomes

  $$\sum_i \frac{(y_i - X_i^T \beta) X_i}{\mathrm{Var}(y_i)} = 0$$

  which looks like weighted least square (difference: weights can depend on β)

Likelihood score equation for the dispersion parameter

- The MLE estimation of $\boldsymbol{\beta}$ for both the general and canonical link does not require knowing ϕ
- Statistical inference of $\boldsymbol{\beta}$ may need an estimate of ϕ (see later)
  - Example: we need to estimate σ^2 in linear regression for calculating test statistics of the coefficients: $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$

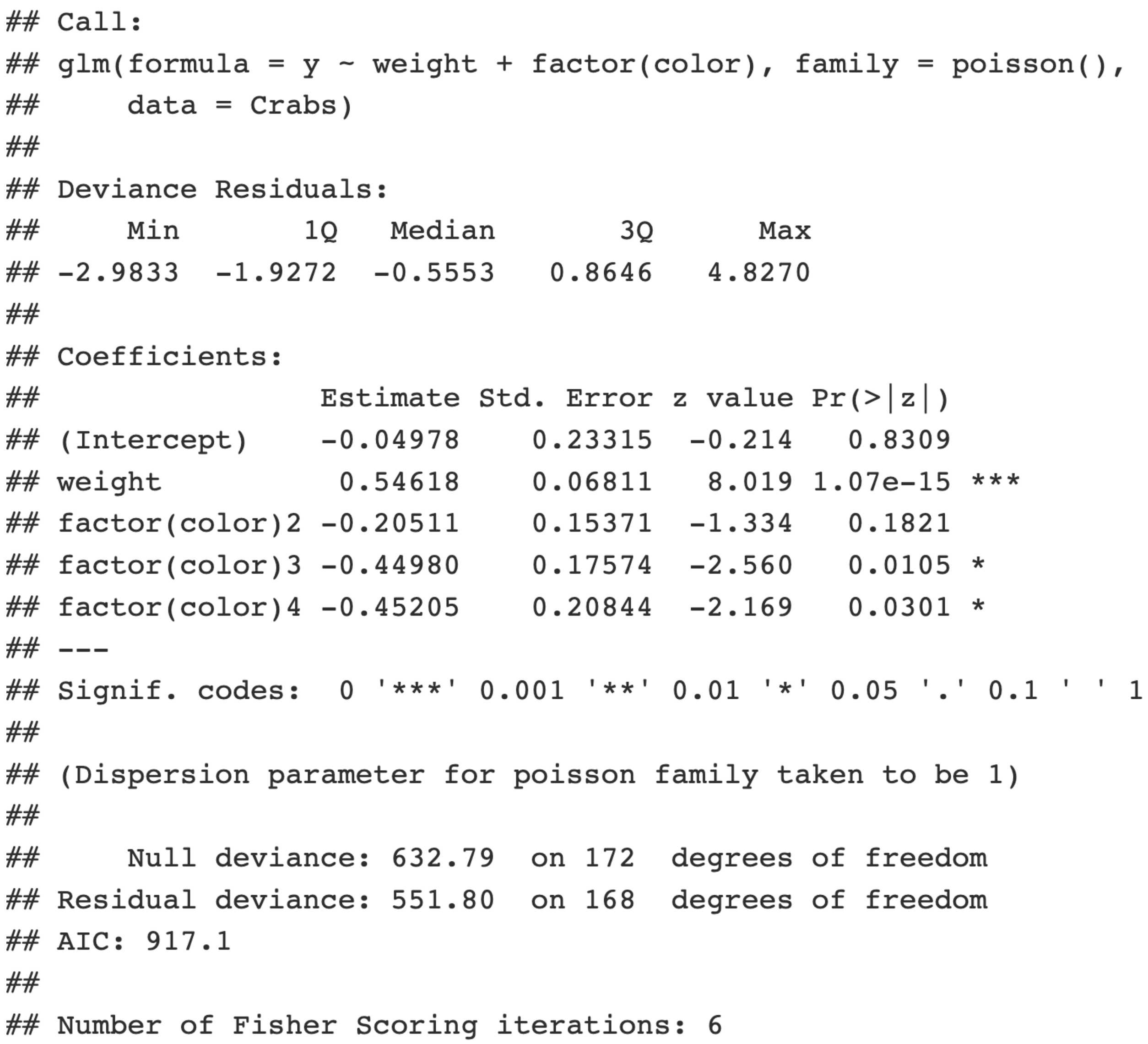
How to estimate ϕ?

- We can also use MLE: find ϕ by solving the equation:

$$\frac{\partial L}{\partial \phi} = 0$$

  - $\frac{\partial L}{\partial \phi}$ also depends on $\boldsymbol{\beta}$: plug-in the MLE estimate $\widehat{\boldsymbol{\beta}}$
  - Example: for Gaussian linear models: $L = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2 - n\log\left(\sqrt{2\pi}\sigma\right)$
    - $\frac{\partial L}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2 - \frac{n}{2\sigma^2} \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})^2$

Statistical inference for GLM



```
## Call:
## glm(formula = y ~ weight + factor(color), family = poisson(),
##     data = Crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9833  -1.9272  -0.5553   0.8646   4.8270
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.04978    0.23315  -0.214   0.8309
## weight          0.54618    0.06811   8.019 1.07e-15 ***
## factor(color)2 -0.20511    0.15371  -1.334   0.1821
## factor(color)3 -0.44980    0.17574  -2.560   0.0105 *
## factor(color)4 -0.45205    0.20844  -2.169   0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 551.80  on 168  degrees of freedom
## AIC: 917.1
##
## Number of Fisher Scoring iterations: 6
```

- How do we get the standard error, z value and p-value of the GLM estimates?

- What does the deviance mean in this table?

Asymptotic distribution of MLE estimation

- The MLE $\widehat{\boldsymbol{\beta}}$ is consistent for the true value $\boldsymbol{\beta}_0$ when $n \to \infty$ and p is fixed

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \xrightarrow{n\to\infty} \mathbf{0}$$

- Asymptotic normality: when n is large

$$\hat{\beta} - \beta_0 \mathrel{\dot{\sim}} N(0, V_{\beta_0})$$

where β_0 is the true value of the parameter. $(nV_{\beta_0}) = O(1))$

- As an applied course, we ignore the discussions of the conditions for the above consistency and CLT results, and skip the proofs.

Calculation of V_{β_0}

- Taylor expansion (local linear approximation):

$$0 = \dot{L}(\hat{\beta}) \approx \dot{L}(\beta_0) + \ddot{L}(\beta_0)(\hat{\beta} - \beta_0)$$

- Then

$$\hat{\beta} - \beta_0 \approx -\left(\ddot{L}(\beta_0)\right)^{-1} \dot{L}(\beta_0) = -\frac{1}{\sqrt{n}} \left(\frac{\ddot{L}(\beta_0)}{n}\right)^{-1} \left(\frac{\dot{L}(\beta_0)}{\sqrt{n}}\right)$$

Calculation of V_{β_0}

Under appropriate conditions, we have

$$\ddot{L}(\beta_0)/n = \sum_i \ddot{L}_i(\beta_0)/n \to \text{Const.} \quad \text{(law of large numbers)}$$

$$\frac{\dot{L}(\beta_0)}{\sqrt{n}} = \frac{\sum_i \dot{L}_i(\beta_0)}{\sqrt{n}} \xrightarrow{d} N(0, V) \quad \text{(central limit theorem)}$$

Thus we have

$$V_{\beta_0} = \left(\mathbb{E}\left(\ddot{L}(\beta_0)\right)\right)^{-1} \mathrm{Cov}\left(\dot{L}(\beta_0)\right) \left(\mathbb{E}\left(\ddot{L}(\beta_0)\right)\right)^{-1}$$

Calculation of V_{β_0}

- The above calculation also can also be used to find the variance of $\hat{\beta}$ from a general estimating equation $\varphi(\hat{\beta}) = 0$ (will discuss more in later lectures)

- Property of the likelihood score equation:

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

Thus

- We also have

$$V_{\beta_0} = -\mathbb{E}\left(\ddot{L}(\beta_0)\right)^{-1} = \mathrm{Cov}\left(\dot{L}(\beta_0)\right)^{-1}$$

$$V_{\beta_0} = (X^T W X)^{-1} \text{ where } W = D^2 V^{-1}$$

- If we use a canonical link, then $W = \frac{D}{a(\phi)} = V/a^2(\phi)$

Asymptotic distribution of any function $h(\hat{\beta})$

- $h(\hat{\beta})$ is a consistent estimator of $h(\beta_0)$

- We use Delta method to understand its uncertainty:

$$h(\hat{\beta}) \approx h(\beta_0) + \dot{h}(\beta_0)^T(\hat{\beta} - \beta_0)$$

$$\sqrt{n}\left(h(\hat{\beta}) - h(\beta_0)\right) \to N\left(0, n\dot{h}(\beta_0)^T V_{\beta_0} \dot{h}(\beta_0)\right)$$

- Example: use Delta method to obtain a CI for $\mu_i = g^{-1}(X_i^T \beta_0)$ of any individual i

Hypothesis testing

- How to test

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

- Example: $H_0: \beta_1 = 0$ V.S. $H_1: \beta_1 \neq 0$

- We will introduce three types of tests:
  - Wald test
  - Score test
  - Likelihood-ratio test

Wald test

- Test statistic

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

  - $\widehat{\text{Cov}}(A\hat{\beta}) = AV_{\hat{\beta}}A^T$

- If a_0 is a scalar, then we can rewrite the test statistic as the Wald statistic

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

- Under H_0, when n is large Wald statistic $z \dot{\sim} N(0,1)$
- We can also obtain a 95% CI for $A\hat{\beta}$: $[A\hat{\beta} - 1.96\sqrt{\widehat{\text{Var}}\left(A\hat{\beta}\right)}, A\hat{\beta} + 1.96\sqrt{\widehat{\text{Var}}\left(A\hat{\beta}\right)}]$

Wald test

- Test statistic

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

  - $\widehat{\text{Cov}}(A\hat{\beta}) = AV_{\hat{\beta}}A^T$

- If a_0 is in general d-dimensional , then under H_0, $T \dot{\sim} \mathcal{X}_d^2$

- The Wald statistic is the “z-value” in the R GLM output for each coefficient β_j

A potential issue with Wald test

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

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

- We can treat the above model as using a canonical link with X being 1, then the asymptotic variance of β_0 is

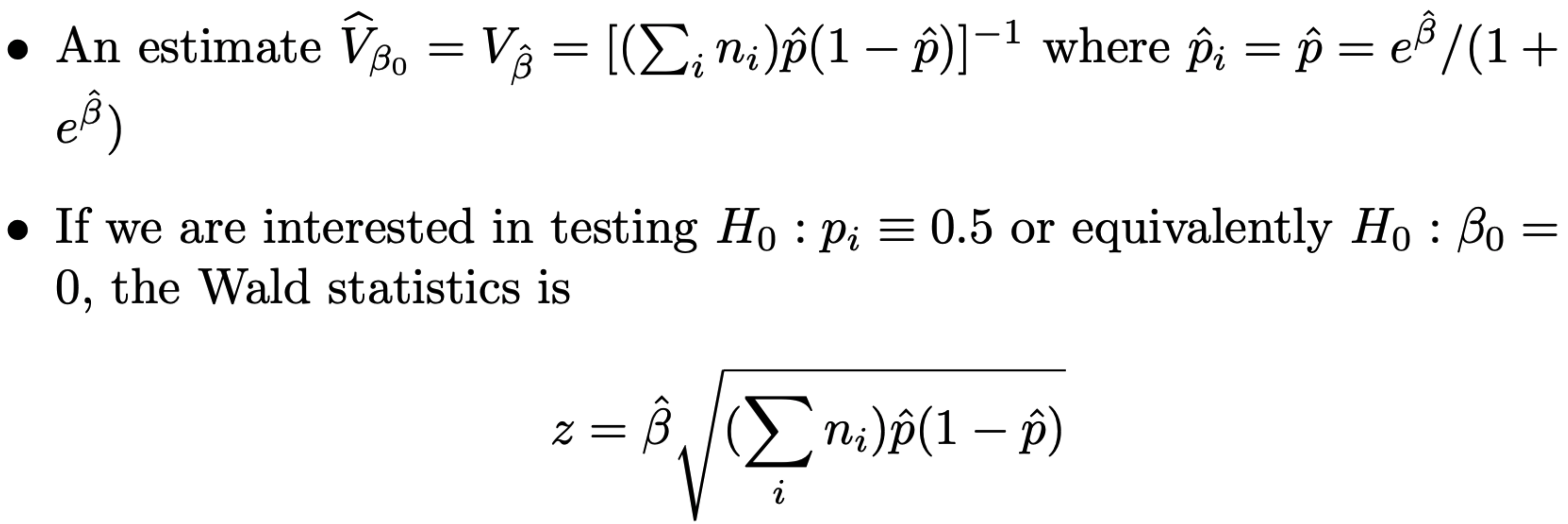
$$V_{\beta_0} = (\sum\nolimits_i V_i)^{-1} = (\sum\nolimits_i n_i p(1-p))^{-1}$$

- 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

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

A potential issue with Wald test



- 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

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

- Let's assume we only have one sample
  - Score equation: $y - np = 0$, so $\hat{p} = y/n$
  - If $y = 23$ and $n = 25$, then $z = 3.31$
  - If $y = 24$ and $n = 25$, then $z = 3.11$.
  - We have a smaller z value when we have stronger evidence against the null?

A potential issue with Wald test

- On the other hand, we use the Wald test to directly test for $H_0\colon p_i \equiv 0.5$

- In the example with only one sample, we can obtain the asymptotic distribution of $\hat{p}$ directly, which results in another Wald statistic

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

  - If $y = 23$ and $n = 25$, then $z = 7.74$
  - If $y = 24$ and $n = 25$, then $z = 11.74$.

- 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)

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:

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

- The score test uses the test statistic

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

  and makes use of the asymptotic normal distribution of $\dot{L}(\beta_0)$

- Under the null, we have

$$T \to \mathcal{X}_p^2 \text{ when } n \to \infty$$

- Benefit of using score test: does not involve any estimation

Likelihood ratio test

- We test for the null

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

- The likelihood ratio test statistic is

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

  - $\tilde{\beta}$ is the MLE of under the constraint $A\beta = a_0$, and $\hat{\beta}$ is our original MLE without any constraints (under the alternative). As $n \to \infty$, under the null

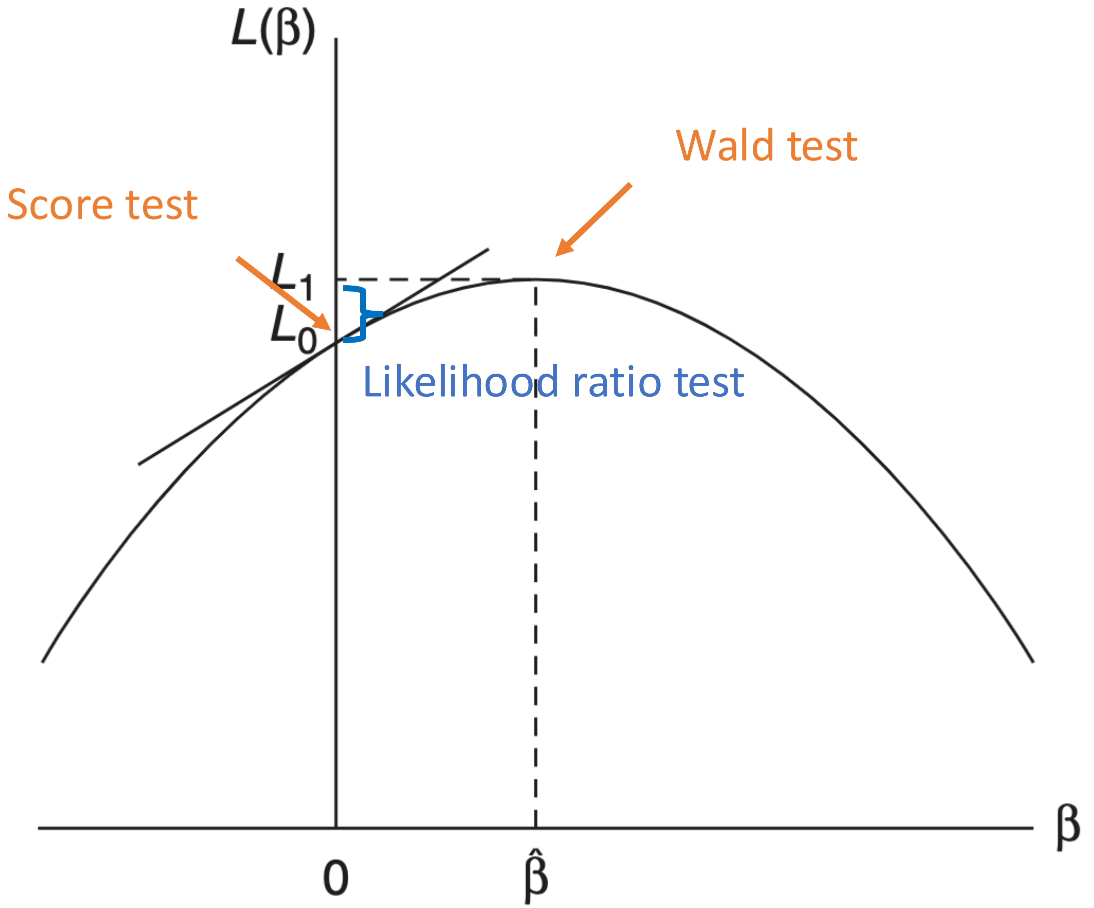
$$-2\log\Lambda \to \mathcal{X}_d^2$$

- Reference of more mathematical details: Chapter 12.4 of Lehmann and Romano, *Testing statistical hypotheses* (3rd edition)

Comparison of the three tests

- We test for the null

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



- Three tests are asymptotically ($n \to \infty$) the same under the null
- We can also construct CI from score and likelihood ratio tests by inverting the tests