

Likelihood Inference

Kosuke Imai

Princeton University

POL572 Quantitative Analysis II
Spring 2011

Likelihood Function and MLE

- Joint distribution: $(Y_1, \dots, Y_n) \sim f(Y_1, \dots, Y_n | \theta)$ where $\theta \in \Theta$
- Idea: Choose the estimate of θ such that the likelihood of obtaining the sample you actually obtained is maximized
- **Likelihood function**: $L(\theta | Y_1, \dots, Y_n) = f(Y_1, \dots, Y_n | \theta)$
- Log-likelihood function: $l(\theta | Y_1, \dots, Y_n) \equiv \log L(\theta | Y_1, \dots, Y_n)$
- Function of θ *given* the data
- **Likelihood Principle**: If Y and \tilde{Y} are two samples and $L(\theta | Y) \propto L(\theta | \tilde{Y})$, then inferences about θ one would draw from Y and \tilde{Y} are the same
- Maximum likelihood estimation (MLE):

$$\hat{\theta}_n \equiv \operatorname{argmax}_{\theta \in \Theta} L(\theta | Y_1, \dots, Y_n) = \operatorname{argmax}_{\theta \in \Theta} l(\theta | Y_1, \dots, Y_n)$$

Normal Regression Model

- Model: $Y_i | X_i \stackrel{\text{indep.}}{\sim} \mathcal{N}(X_i^\top \beta, \sigma^2)$
- Or equivalently: $Y_i = X_i^\top \beta + \epsilon_i$ where $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$
- Likelihood and log-likelihood functions:

$$L_n(\beta, \sigma^2 | Y, X) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 \right\}$$

$$l_n(\beta, \sigma^2 | Y, X) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2$$

- Solving the first order condition (and checking the second order condition) yields the following MLE:

$$\hat{\beta} = \left(\sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i Y_i, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta})^2$$

Consistency

- θ_0 : True value of θ
- $\hat{\theta}_n$: MLE of θ as a function of sample size
- **Theorem:** If $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} f(Y_i | \theta_0)$, then
 $\hat{\theta}_n \equiv \operatorname{argmax}_{\theta \in \Theta} l_n(\theta | Y) \xrightarrow{P} \theta_0$ under some regularity conditions
- Identification: the maximum value of $L(\theta | Y)$ exists and is unique

Intuitive “proof”

- 1 Show $\hat{\theta}_n \xrightarrow{P} \hat{\theta} \equiv \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}\{\log f(Y_i | \theta)\}$
- 2 Recall **Jensen's inequality**: For any concave (convex) function $g(\cdot)$ and a random variable X , we have $\mathbb{E}\{g(X)\} \leq g(\mathbb{E}(X))$
($\mathbb{E}\{g(X)\} \geq g(\mathbb{E}(X))$)
- 3 Show that **Kullback-Leibler Divergence** is non-negative, i.e.,

$$\mathbb{E}(\log f(Y_i | \theta_0) - \log f(Y_i | \hat{\theta})) \geq 0$$

- 4 Argue that the equality holds if and only if $\hat{\theta} = \theta_0$

Score, Fisher Information, and Information Equality

- Score statistic:

$$\mathbf{s}_n(\tilde{\theta}) \equiv \frac{\partial}{\partial \theta} \ln(\theta | \mathbf{Y}) \Big|_{\theta=\tilde{\theta}}$$

- Mean of score is zero:

$$\begin{aligned} \mathbb{E}(\mathbf{s}_i(\theta_0)) &= \int \frac{\partial}{\partial \theta} \ln(\theta | Y_i) \Big|_{\theta=\theta_0} f(Y_i | \theta_0) dY_i \\ &= \int \frac{\partial}{\partial \theta} f(Y_i | \theta) \Big|_{\theta=\theta_0} dY_i = 0 \end{aligned}$$

- Fisher information:

$$\Omega(\theta_0) \equiv \mathbb{E}(\mathbf{s}_i(\theta_0)\mathbf{s}_i(\theta_0)^\top) = \mathbb{V}(\mathbf{s}_i(\theta_0))$$

- Hessian matrix:

$$H_n(\tilde{\theta}) \equiv \frac{\partial^2}{\partial \theta \partial \theta^\top} \ln(\theta | \mathbf{Y}) \Big|_{\theta=\tilde{\theta}}$$

- Information Equality: $\mathbb{E}(H_i(\theta_0)) = -\Omega(\theta_0)$

Asymptotic Normality

- Taylor expansion around θ_0 :

$$0 = s_n(\hat{\theta}_n) \approx s_n(\theta_0) + H_n(\theta_0)(\hat{\theta}_n - \theta_0)$$

- Asymptotic distribution:

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &\approx \underbrace{\left(-\frac{1}{n} \sum_{i=1}^n H_i(\theta_0)\right)^{-1}}_{\xrightarrow{P} \Omega(\theta_0)^{-1}} \underbrace{\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n s_i(\theta_0)\right)}_{\xrightarrow{D} \mathcal{N}(0, \Omega(\theta_0))} \\ &\xrightarrow{D} \mathcal{N}\left(0, \Omega(\theta_0)^{-1}\right) \end{aligned}$$

- Approximate variance: $\mathbb{V}(\hat{\theta}_n) \approx \frac{1}{n} \Omega(\theta_0)^{-1}$
- Variance estimator: $\widehat{\mathbb{V}}(\hat{\theta}_n) = -H_n(\hat{\theta}_n)^{-1}$

Normal Regression Model, Again

- Parameters: $\theta = (\beta, \sigma^2)$
- Score statistic:

$$s_n(\theta_0) = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \end{bmatrix}$$

- Hessian matrix:

$$H_n(\theta_0) = \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} & -\frac{1}{\sigma^4} \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta) \\ -\frac{1}{\sigma^4} (\mathbf{Y} - \mathbf{X}\beta)^\top \mathbf{X} & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \end{bmatrix}$$

- Information matrix:

$$\Omega(\theta_0) = \mathbb{E}(\mathbf{s}_i(\theta_0)\mathbf{s}_i(\theta_0)^\top \mid \mathbf{X}) = -\mathbb{E}\left(\frac{H_n(\theta_0)}{n} \mid \mathbf{X}\right) = \begin{bmatrix} \frac{1}{n\sigma^2} \mathbf{X}^\top \mathbf{X} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

- Approximate variance:

$$\mathbb{V}(\hat{\theta}_n \mid \mathbf{X}) \approx \begin{bmatrix} \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Invariance and Transformation

- If $\hat{\theta}_n$ is the MLE of θ , the MLE of $g(\theta)$ is $g(\hat{\theta}_n)$ for any function $g(\cdot)$
- But, to improve the asymptotic normal approximation of the sampling distribution, we can use transformations
 - ① $\theta \in (0, \infty) \rightarrow \log \theta$
 - ② $\theta \in (0, 1) \rightarrow \log \left(\frac{\theta}{1-\theta} \right)$ logistic
 - ③ $\theta \in (-1, 1) \rightarrow \frac{1}{2} \log \left(\frac{1+\theta}{1-\theta} \right)$ Fisher's z
- In normal regression, let's change σ^2 to $\gamma \equiv \log \sigma^2$. Then,

$$\begin{aligned} \frac{\partial}{\partial \gamma} \ln(\beta, \gamma \mid Y, X) &= \frac{\partial \sigma^2}{\partial \gamma} \frac{\partial}{\partial \sigma^2} \ln(\beta, \sigma^2 \mid Y, X) \\ &= -\exp(\gamma) \left\{ \frac{n}{2 \exp(\gamma)} - \frac{1}{2 \exp(2\gamma)} \|Y - X\beta\|^2 \right\} \end{aligned}$$

Asymptotic Efficiency

- **Cramer-Rao Inequality:** Let $\tilde{\theta}_n$ be any estimator.

$$\mathbb{V}(\tilde{\theta}_n) \geq \frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n) \left\{ n \mathbb{E}(\mathbf{s}_i(\theta_0) \mathbf{s}_i(\theta_0)^\top) \right\}^{-1} \left\{ \frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n) \right\}^\top$$

where $\mathbb{E}(\tilde{\theta}_n) = \int \tilde{\theta}_n f(Y | \theta_0) dY$

- For any asymptotically unbiased estimator $\tilde{\theta}_n$, we have

$$\mathbb{V}(\tilde{\theta}_n) \geq \frac{1}{n} \Omega(\theta_0)^{-1}$$

- MLE achieves the Cramer-Rao Lower Bound for *any* θ_0
- MLE is asymptotically uniformly minimum variance unbiased estimator (UMVUE)

Outline of Proof

1 Show

$$\frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n) = \mathbb{E} \left\{ \tilde{\theta}_n \frac{\partial}{\partial \theta_0} \log f(Y | \theta_0) \right\}$$

2 Show

$$\text{Cov} \left(\tilde{\theta}_n, \frac{\partial}{\partial \theta_0} \log f(Y | \theta_0) \right) = \frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n)$$

3 By the Covariance inequality,

$$\frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n) \left\{ \frac{\partial}{\partial \theta_0} \mathbb{E}(\tilde{\theta}_n) \right\}^\top \leq \mathbb{V}(\tilde{\theta}_n) \mathbb{V} \left\{ \frac{\partial}{\partial \theta_0} \log f(Y | \theta_0) \right\}$$

Model Misspecification

- What happens if the model is wrong?
- True model: $Y_i \stackrel{\text{i.i.d.}}{\sim} g(Y_i)$
- Misspecified model: $Y_i \stackrel{\text{i.i.d.}}{\sim} f(Y_i | \theta_0)$
- $\hat{\theta}_n \xrightarrow{P} \theta_0$ such that

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \underbrace{\int \log \frac{g(Y_i)}{f(Y_i | \theta)} g(Y_i) dY_i}_{\text{Kullback-Leibler distance}}$$

- Asymptotic distribution:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}\left(0, \mathbb{E}(H_i(\theta_0))^{-1} \mathbb{E}(s_i(\theta_0)s_i(\theta_0)^\top) \mathbb{E}(H_i(\theta_0))^{-1}\right)$$

- Information equality does not hold

Robust Standard Error

- Sandwich estimator:

$$\text{bread} = \left(-\frac{1}{n} H_n(\hat{\theta}_n) \right)^{-1}, \quad \text{and} \quad \text{meat} = \frac{1}{n} \sum_{i=1}^n \mathbf{s}_i(\hat{\theta}_n) \mathbf{s}_i(\hat{\theta}_n)^\top$$

- Clustering: a wrong likelihood function
- Cluster robust standard error:

$$\text{meat} = \frac{1}{n} \sum_{g=1}^G \left\{ \left(\sum_{i=1}^{n_g} \mathbf{s}_i(\hat{\theta}_n) \right) \left(\sum_{i=1}^{n_g} \mathbf{s}_i(\hat{\theta}_n) \right)^\top \right\}$$

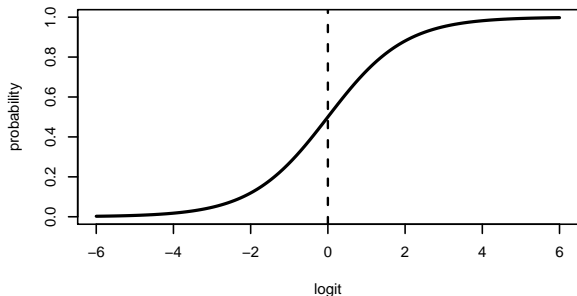
- “Correct” standard error for “wrong” estimate

Logit and Probit Models

- The logit model for binary outcome $Y_i \in \{0, 1\}$:

$$Y_i \stackrel{\text{indep.}}{\sim} \text{Bernoulli}(\pi_i)$$
$$\pi_i = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} = \frac{1}{1 + \exp(-X_i^\top \beta)}$$

- Logit: $\text{logit}(\pi_i) \equiv \log(\pi_i/(1 - \pi_i)) = X_i^\top \beta$
- Probit: $\Phi^{-1}(\pi_i) = X_i^\top \beta$



- monotone increasing
- symmetric around 0
- maximum slope at 0

Latent Variable Interpretation

- The latent variable or the “Utility”: Y_i^*
- The Model:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$
$$Y_i^* = \mathbf{X}_i^\top \beta + \epsilon_i \quad \text{with} \quad \mathbb{E}(\epsilon_i) = 0$$

- Logit: $\epsilon_i \stackrel{\text{i.i.d.}}{\sim}$ logistic (the density is $\exp(-\epsilon_i)/\{1 + \exp(-\epsilon_i)\}^2$)
- Probit: $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$
- The variance of Y_i^* is not identifiable
- The “cutpoint” is not identifiable

Inference with the Logit Model

- Likelihood and log-likelihood functions:

$$L_n(\beta \mid Y, \mathbf{X}) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i}$$

$$l_n(\beta \mid Y, \mathbf{X}) = \sum_{i=1}^n \{Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i)\}$$

- Score function: $\mathbf{s}_n(\beta) = \sum_{i=1}^n (Y_i - \pi_i) \mathbf{X}_i$
- Hessian: $H_n(\beta) = - \sum_{i=1}^n \pi_i (1 - \pi_i) \mathbf{X}_i \mathbf{X}_i^\top \leq 0$
- Approximate variance: $\mathbb{V}(\hat{\beta}_n \mid \mathbf{X}) \approx \{\sum_{i=1}^n \pi_i (1 - \pi_i) \mathbf{X}_i \mathbf{X}_i^\top\}^{-1}$
- Globally concave

Newton-Raphson Algorithm

- Find $\hat{\theta}_n$ such that $s_n(\hat{\theta}_n) = 0$
- **Mean Value Theorem**: If $f(x)$ is continuous and differentiable on $[a, b]$, then there exists $c \in [a, b]$ such that

$$\left. \frac{\partial}{\partial x} f(x) \right|_{x=c} = \frac{f(b) - f(a)}{b - a}$$

- Thus, for $\tilde{\theta} \in [\theta^{(t)}, \hat{\theta}_n]$,

$$s_n(\theta^{(t)}) = s_n(\theta^{(t)}) - s_n(\hat{\theta}_n) = H_n(\tilde{\theta})(\theta^{(t)} - \hat{\theta}_n)$$

- The algorithm: converges at $\hat{\theta}_n$

$$\theta^{(t+1)} = \theta^{(t)} - H_n(\theta^{(t)})^{-1} s_n(\theta^{(t)})$$

- **Fisher scoring algorithm**: use $\Omega(\theta^{(t)})^{-1}$ (always positive-definite)
- Global maxima vs. local maxima: different starting values

Calculating Quantities of Interest

- Logistic regression coefficients are not quantities of interest
- Predicted probability: $\pi(x) = \Pr(Y = 1 \mid X = x) = \frac{\exp(x^\top \beta)}{1 + \exp(x^\top \beta)}$
- Attributable risk (risk difference): $\pi(x_1) - \pi(x_0)$
- Relative risk: $\pi(x_1)/\pi(x_0)$
- Odds ratio: $\frac{\pi(x_1)/\{1-\pi(x_1)\}}{\pi(x_0)/\{1-\pi(x_0)\}}$
- Average Treatment Effect:

$$\mathbb{E}\{\Pr(Y_i = 1 \mid T_i = 1, X_i) - \Pr(Y_i = 1 \mid T_i = 0, X_i)\}$$

- MLE: plug in $\hat{\beta}_n$
- Asymptotic distribution: the Delta method (a bit **painful!**)

$$\sqrt{n}(\hat{\pi}(x) - \pi(x)) \xrightarrow{D} \mathcal{N}\left(0, \frac{\pi(x)^2}{\{1 + \exp(x^\top \beta_0)\}^2} x^\top \Omega(\beta_0)^{-1} x\right)$$

Quasi-Bayesian Approximation

- Bayesians: parameters are random variables

$$\underbrace{p(\theta | Y)}_{\text{posterior}} = \frac{\overbrace{p(Y | \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int p(Y | \theta) p(\theta) d\theta}_{\text{marginal likelihood}}} \propto p(Y | \theta) p(\theta)$$

- **Bernstein – Von Mises Theorem:** For a large sample, Bayes estimate is close to the MLE. The posterior distribution of the parameter around the posterior mean is also close to the distribution of the MLE around the truth,
- Sample θ from $\mathcal{N}(\hat{\theta}_n, -H_n(\hat{\theta}_n)^{-1})$ and compute “Bayes” estimate and confidence interval of $g(\theta)$ for any function $g(\cdot)$

Bootstrap

- Unknown data generating process: $Y_i \stackrel{\text{i.i.d.}}{\sim} F$
- Want to know $\mathbb{V}_F(\hat{\theta}_n)$ where $\theta_n \equiv g(Y_1, \dots, Y_n)$
- Approximate it with $\mathbb{V}_{\hat{F}_n}(\hat{\theta}_n)$ where \hat{F}_n is the empirical CDF
- Real world: $F \implies Y_1, \dots, Y_n \implies \hat{\theta}_n$
- Bootstrap world: $\hat{F}_n \implies Y_1^*, \dots, Y_n^* \implies \hat{\theta}_n^* \equiv g(Y_1^*, \dots, Y_n^*)$

$$\hat{\sigma}_{boot}^2 \equiv \frac{1}{B} \sum_{b=1}^B \left(\hat{\theta}_{n,b}^* - \frac{1}{B} \sum_{b'=1}^B \hat{\theta}_{n,b'}^* \right)^2$$

- Asymptotic approximation:

$$\mathbb{V}_F(\hat{\theta}_n) \quad \underbrace{\approx}_{\text{may not be so small}} \quad \mathbb{V}_{\hat{F}_n}(\hat{\theta}_n) \quad \underbrace{\approx}_{\text{small}} \quad \hat{\sigma}_{boot}^2$$

- **Bootstrap percentile confidence intervals**
- **Parametric bootstrap**: Replace \hat{F}_n with $F_{\hat{\theta}_n}$

Bootstrap and Estimation of Bias

- Real world: $F \implies Y_1, \dots, Y_n \implies \text{bias}_F = \mathbb{E}_F(\hat{\theta}_n) - \theta(F)$
- Bootstrap world: $\hat{F}_n \implies Y_1^*, \dots, Y_n^* \implies \text{bias}_{\hat{F}_n} = \mathbb{E}_{\hat{F}_n}(\hat{\theta}_n^*) - \theta(\hat{F}_n)$
- An example: Ratio estimator $\hat{\theta}_n = \sum_{i=1}^n Y_i / \sum_{i=1}^n X_i$
- $\text{bias}_F = \mathbb{E}(\hat{\theta}_n) - \mathbb{E}(Y_i) / \mathbb{E}(X_i) \neq 0$
- $\theta(\hat{F}_n) = \hat{\theta}_n$
- $\mathbb{E}_{\hat{F}_n}(\hat{\theta}_n^*) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_{n,b}^*$

Bootstrapping Residuals

- Calculate residuals: $\hat{\epsilon}_i = Y_i - X_i^\top \hat{\beta}$
- Bootstrap residuals: $\hat{\epsilon}_i^* \stackrel{\text{i.i.d.}}{\sim} \hat{F}$
- $\hat{\beta}_b^* = (X^\top X)^{-1} X^\top Y^*$ where $Y_i^* = X_i^\top \hat{\beta} + \hat{\epsilon}_i^*$
- Bootstrap is not a uniquely defined concept
- Bootstrap residuals: assumes the model
- Bootstrap data: does not assume the model
- See Examples 3 and 4 of Freedman

Likelihood Ratio Test (LRT)

- Null hypothesis: $H_0 : \theta \in \Theta_0 \subset \Theta$ with $H_1 : \theta \in \Theta \setminus \Theta_0$
- “Nested” models: $H_0 : g_1(\theta) = \dots = g_K(\theta) = 0$
- Unlike F test, nonlinear constraints are allowed
- **Likelihood ratio statistic:**

$$\lambda_n(Y) \equiv \frac{\overbrace{\sup_{\theta \in \Theta_0} L_n(\theta | Y)}^{\text{restricted MLE}}}{\underbrace{\sup_{\theta \in \Theta} L_n(\theta | Y)}_{\text{MLE}}} = \frac{L_n(\bar{\theta}_n | Y)}{L_n(\hat{\theta}_n | Y)} \in (0, 1)$$

- Asymptotic distribution:

$$-2 \log \lambda_n(Y) \xrightarrow{D} \chi_K^2$$

where $K = \dim(\Theta) - \dim(\Theta_0)$: the difference between the number of free parameters in Θ and Θ_0

Proof for a Special Case

- $H_0 : \theta = \theta_0$ with $H_1 : \theta \neq \theta_0$
- Taylor expansion (again!):

$$l_n(\theta_0 | Y) \approx l_n(\hat{\theta}_n | Y) + s_n(\hat{\theta}_n | Y)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\top H_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)$$

- Under the null hypothesis,

$$\begin{aligned} -2 \log \lambda_n(Y) &= -2[\log l_n(\theta_0 | Y) - \log l_n(\hat{\theta}_n | Y)] \\ &\approx \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)^\top}_{\xrightarrow{D} \mathcal{N}(0, \Omega(\theta_0)^{-1})} \underbrace{-\frac{1}{n} \sum_{i=1}^n H_i(\hat{\theta}_n)}_{\xrightarrow{P} \Omega(\theta_0)} \underbrace{\sqrt{n}(\hat{\theta}_n - \theta_0)}_{\xrightarrow{D} \mathcal{N}(0, \Omega(\theta_0)^{-1})} \\ &\xrightarrow{D} \chi_K^2 \end{aligned}$$

Wald and Score Tests

- $H_0 : g_1(\theta) = \dots = g_K(\theta) = 0$
- Wald test: No need to calculate $\bar{\theta}_n$

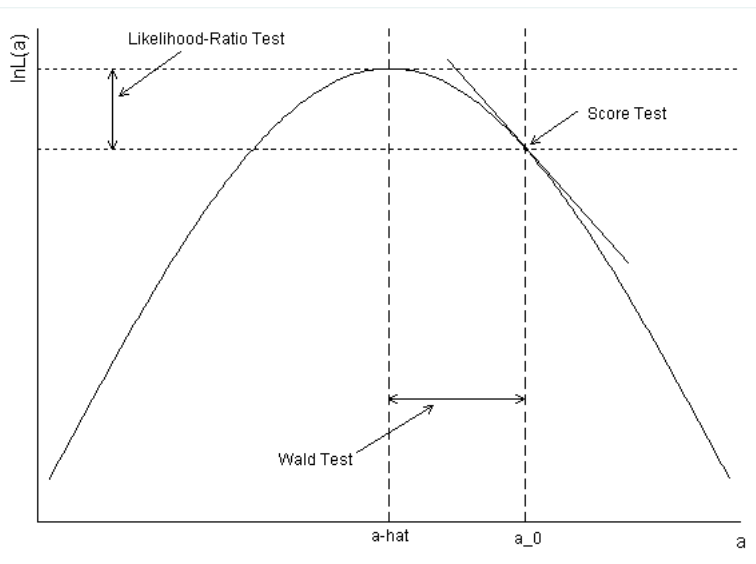
$$\begin{aligned}W_n &\equiv ng(\hat{\theta}_n)^\top \{g^{(1)}(\hat{\theta}_n)^\top \Omega(\hat{\theta}_n)^{-1} g^{(1)}(\hat{\theta}_n)\}^{-1} g(\hat{\theta}_n) \xrightarrow{D} \chi_K^2 \\ &\equiv -g(\hat{\theta}_n)^\top \{g^{(1)}(\hat{\theta}_n)^\top H_n(\hat{\theta}_n)^{-1} g^{(1)}(\hat{\theta}_n)\}^{-1} g(\hat{\theta}_n) \xrightarrow{D} \chi_K^2\end{aligned}$$

- (Rao's) Score test: No need to calculate $\hat{\theta}_n$

$$R_n \equiv -s_n(\bar{\theta}_n)^\top H_n(\bar{\theta}_n)^{-1} s_n(\bar{\theta}_n) \xrightarrow{D} \chi_K^2$$

- LRT requires the calculation of both $\bar{\theta}_n$ and $\hat{\theta}_n$
- But, they are all asymptotically equivalent!

Geometry of Likelihood Ratio, Wald, and Score Tests



Concluding Remarks

- Likelihood inference: very general and powerful tool
- Asymptotic consistency and efficiency
- Parametric assumptions need to be made with care
- Important topics that are not covered here: model (variable) selection, model validation
- Report quantities of interest rather than model parameters
- Be aware of what “robust” standard error can and cannot do