Inference in non-linear time series

Erik Lindström

Centre for Mathematical Sciences
Lund University

LU/LTH & DTU

Overview

General

We are interested in estimating a parameter vector θ_0 from data $\mathbf{X}$.

- Ad hoc or formal estimation?
- Properties
- Definition

$$\hat{\theta} = T_N(\mathbf{X}).$$

- Interpretation

Properties

- Bias $b = \theta_0 - \mathbf{E}[T_N(\mathbf{X})]$.
- Asympt. bias $\lim_{N\to\infty} \theta_0 - \mathbf{E}[T_N(\mathbf{X})]$.
- Consistency $\hat{\theta} \xrightarrow{p} \theta_0 \Leftrightarrow \mathbb{P}(|\hat{\theta} - \theta_0| > \varepsilon) \to 0$ as $N \to \infty$.
- Strong consistency $\hat{\theta} \xrightarrow{a.s.} \theta_0$.
- Efficiency

$$Var[T_N(\mathbf{X})] \geq I_N^{-1}(\theta_0),$$

  where $I_N^{-1}(\theta_0)$ is the Fisher information matrix.
- Asympt. normality
- Convergence

$$N^{\alpha}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\mu, \Sigma).$$

Popular estimators

How are the estimates computed?

- **Optimization:** Least squares (LS), Weighted Least Squares (WLS), Prediction Error Methods (PEM), Generalized Method of Moments (GMM) etc.
- **Solving equations:** Estimation functions (EFs), Method of moments (MM)
- **Either:** Bayesian methods (MCMC), Maximum Likelihood, Quasi Maximum Likelihood

Least squares

- Observations $y_1, \ldots, y_N$
- Predictors $\sum_{j=1}^{J} B_j(x)\theta_j$
- Vector form: $Y = Z\theta + \varepsilon, \varepsilon \sim N(0, \Omega)$

(Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{n=1}^{N} \lambda_n \left(y_n - \sum_j B_j(x_n)\theta_j \right)^2 \quad (1)$$

$$= (Y - Z\theta)^T W (Y - Z\theta) \quad (2)$$

Generalization

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \quad (3)$$

Least squares

- Observations $y_1, \ldots, y_N$
- Predictors $\sum_{j=1}^{J} B_j(x)\theta_j$
- Vector form: $Y = Z\theta + \varepsilon, \varepsilon \sim N(0, \Omega)$

(Weighted) Parameter estimation:

$$\hat{\theta}_{LS} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{n=1}^{N} \lambda_n \left(y_n - \sum_j B_j(x_n)\theta_j \right)^2 \quad (1)$$

$$= (Y - Z\theta)^T W (Y - Z\theta) \quad (2)$$

Generalization

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \quad (3)$$

- Estimate

$$\hat{\theta} = \left(Z^T W Z\right)^{-1} (ZWY) \tag{4}$$

- Bias? No, not if the correct model is used

$$\mathbf{E}\left[\hat{\theta}\right] = \left(Z^T W Z\right)^{-1} (ZW(Z\theta + \varepsilon)) \tag{5}$$

$$= \theta + \left(Z^T W Z\right)^{-1} (ZW\varepsilon) \tag{6}$$

- Variance

$$\mathrm{Var}\left[\hat{\theta}\right] = (Z^T W Z)^{-1} (Z^T W \Omega W Z)(Z^T W Z)^{-1} \tag{7}$$

Simplifies if $W = \Omega^{-1}$, and $\Omega = \sigma^2 I_N$. We then get

$$\mathrm{Var}\left[\hat{\theta}\right] = \sigma^2 (Z^T Z)^{-1}. \tag{8}$$

- Estimate

$$\hat{\theta} = \left(Z^T W Z\right)^{-1} (ZWY) \tag{4}$$

- Bias? No, not if the correct model is used

$$\mathbf{E}\left[\hat{\theta}\right] = \left(Z^T W Z\right)^{-1} (ZW(Z\theta + \varepsilon)) \tag{5}$$

$$= \theta + \left(Z^T W Z\right)^{-1} (ZW\varepsilon) \tag{6}$$

- Variance

$$\mathrm{Var}\left[\hat{\theta}\right] = (Z^T W Z)^{-1} (Z^T W \Omega W Z)(Z^T W Z)^{-1} \tag{7}$$

Simplifies if $W = \Omega^{-1}$, and $\Omega = \sigma^2 I_N$. We then get

$$\mathrm{Var}\left[\hat{\theta}\right] = \sigma^2 (Z^T Z)^{-1}. \tag{8}$$

- Estimate

$$\hat{\theta} = \left(Z^T W Z\right)^{-1} (ZWY) \tag{4}$$

- Bias? No, not if the correct model is used

$$\mathbf{E}\left[\hat{\theta}\right] = \left(Z^T W Z\right)^{-1} (ZW(Z\theta + \varepsilon)) \tag{5}$$

$$= \theta + \left(Z^T W Z\right)^{-1} (ZW\varepsilon) \tag{6}$$

- Variance

$$\mathrm{Var}\left[\hat{\theta}\right] = (Z^T W Z)^{-1}(Z^T W \Omega W Z)(Z^T W Z)^{-1} \tag{7}$$

Simplifies if $W = \Omega^{-1}$, and $\Omega = \sigma^2 I_N$. We then get

$$\mathrm{Var}\left[\hat{\theta}\right] = \sigma^2 (Z^T Z)^{-1}. \tag{8}$$

- Estimate

$$\hat{\theta} = \left(Z^T W Z\right)^{-1} (ZWY) \tag{4}$$

- Bias? No, not if the correct model is used

$$\mathbf{E}\left[\hat{\theta}\right] = \left(Z^T W Z\right)^{-1} (ZW(Z\theta + \varepsilon)) \tag{5}$$

$$= \theta + \left(Z^T W Z\right)^{-1} (ZW\varepsilon) \tag{6}$$

- Variance

$$\mathrm{Var}\left[\hat{\theta}\right] = (Z^T W Z)^{-1}(Z^T W \Omega W Z)(Z^T W Z)^{-1} \tag{7}$$

Simplifies if $W = \Omega^{-1}$, and $\Omega = \sigma^2 I_N$. We then get

$$\mathrm{Var}\left[\hat{\theta}\right] = \sigma^2 (Z^T Z)^{-1}. \tag{8}$$

F-tests

We estimate $\sigma^2 = \frac{(Y-X\theta)^T(Y-X\theta)}{N-J}$

- Define $Q(\hat{\theta}) = (Y - X\hat{\theta})^T(Y - X\hat{\theta}) = Y^T(I_N - P)^T(I_N - P)Y$ where P is the projection matrix
- We approximate Q by $\mathbf{E}[Q] = tr\left((I_N - P)^T(I_N - P)\,\mathrm{Cov}[Y,Y]\right)$.
- It holds for the standard model that $\mathrm{Cov}[Y,Y] = \sigma^2 I_N$ and $(I_N - P)^T(I_N - P) = (I_N - P)$
- It then follows that

$$tr\left((I_N - P)^T(I_N - P)\,\mathrm{Cov}[Y,Y]\right) = \sigma^2 tr\,(I_N - P) \tag{9}$$

$$= \sigma^2 tr(I_N) - \sigma^2 tr(P) \tag{10}$$

$$= \sigma^2 N - \sigma^2 tr\left(Z(Z^TZ)^{-1}Z^T\right) \tag{11}$$

$$= \sigma^2(N - tr\left((Z^TZ)^{-1}(Z^TZ)\right) = \sigma^2(N - J) \tag{12}$$

What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D (\theta - \theta_0) \qquad (13)$$

Derive on blackboard.

Another popular form is

$$\hat{\theta}_{Adaptive\ LASSO} = (Y - Z\theta)^T W (Y - Z\theta) + ||D(\theta - \theta_0)||_1 \qquad (14)$$

What about the penalized asymptotics?

What if the estimation is given by

$$\hat{\theta}_{PLS} = (Y - Z\theta)^T W (Y - Z\theta) + (\theta - \theta_0)^T D(\theta - \theta_0) \quad (13)$$

Derive on blackboard.

Another popular form is

$$\hat{\theta}_{Adaptive\ LASSO} = (Y - Z\theta)^T W (Y - Z\theta) + ||D(\theta - \theta_0)||_1 \quad (14)$$

Maximum Likelihood estimators

Defined as

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p_\theta(X_1, \ldots X_N).$$

The MLE use all information in the data, and have nice properties:

- Ia. Consistency $\hat{\theta} \xrightarrow{p} \theta_0$.
- Ib. Asympt. normality
- II. Efficiency $Var[T_N(\mathbf{X})] = I_0^{-1}(\theta_0)$,
- III. Invariant.

We have under general conditions that

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_0^{-1}(\theta_0)).$$

Two useful theorems

Assume X_i iid and $\mathbf{E}[X_i] = \mu$, $Var[X_i] = \sigma^2$

- **Law of Large Numbers**

$$\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{p/a.s.} \mu.$$

- **Central Limit Theorem**

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{X_i - \mu}{\sigma} \xrightarrow{d} Z,$$

where $Z \in \mathcal{N}(0,1)$.

These theorems can be generalized further.

Fisher information

The Fisher Information matrix is defined as

$$I(\theta) = Var\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\right],$$

which is equivalent to

$$\mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^2\right]$$

and

$$-\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right].$$

Note $\mathbf{E}\left[\frac{\partial}{\partial\theta}\log p_\theta(X|\theta)\right] = 0$.

Proofs

- Ia. Kullback-Leibler and law of large number
- Ib. Second order Taylor expansion of the Likelihood.
- II. Cauchy-Schwartz inequality
- III. Direct calculations.

Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{15}$$

The estimate is given by

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta) = \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{N}\ell(\theta) \tag{16}$$

$$= \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{17}$$

Now we rewrite this as

$$\frac{1}{N}\ell(\theta) = \frac{1}{N}\ell(\theta) \pm \mathbf{E}[\ell(\theta)] \tag{18}$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)]\right) \tag{19}$$

Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{15}$$

The estimate is given by

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \frac{1}{N}\ell(\theta) \tag{16}$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\, \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{17}$$

Now we rewrite this as

$$\frac{1}{N}\ell(\theta) = \frac{1}{N}\ell(\theta) \pm \mathbf{E}[\ell(\theta)] \tag{18}$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)]\right) \tag{19}$$

Consistency

The log-likelihood function is defined as

$$\ell(\theta) = \sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{15}$$

The estimate is given by

$$\hat{\theta}_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \ell(\theta) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{N}\ell(\theta) \tag{16}$$

$$= \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x_n | x_{1:n-1}) \tag{17}$$

Now we rewrite this as

$$\frac{1}{N}\ell(\theta) = \frac{1}{N}\ell(\theta) \pm \mathbf{E}[\ell(\theta)] \tag{18}$$

$$= \mathbf{E}[\ell(\theta)] + \left(\frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)]\right) \tag{19}$$

Consistency cont.

This decomposition shows that

$$\frac{1}{N}\ell(\theta) = \underbrace{\mathbf{E}[\ell(\theta)]}_{Expected \;\; log-likelihood} + \underbrace{\frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)]}_{Random \;\; error} \qquad (20)$$

It then follows that

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta\in\Theta} \mathbf{E}[\ell(\theta)] + \left(\hat{\theta}_{MLE} - \operatorname*{argmax}_{\theta\in\Theta} \mathbf{E}[\ell(\theta)]\right) \qquad (21)$$

First we look at the random error, that is written so that we can apply the *Law of Large Numbers*. If

$$\lim_{N\to\infty} \frac{1}{N}\ell(\theta) - \mathbf{E}[\ell(\theta)] = 0 \tag{22}$$

uniformly over Θ, then it follows that

$$\lim_{N\to\infty} \hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta\in\Theta} \mathbf{E}[\ell(\theta)] \tag{23}$$

- Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.
- The difference in expected log-likelihood between the true density and any other density is given by

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int (\log q_\theta(x) - \log p_{\theta_0}(x))\, p_{\theta_0}(x)\mathrm{d}x \tag{24}$$

$$= \int \log\left(\frac{q_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\mathrm{d}x \tag{25}$$

- Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

- Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.
- The difference in expected log-likelihood between the true density and any other density is given by

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int \left(\log q_\theta(x) - \log p_{\theta_0}(x)\right) p_{\theta_0}(x)\mathrm{d}x \quad (24)$$

$$= \int \log\left(\frac{q_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\mathrm{d}x \quad (25)$$

- Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

- Finally, we denote any feasible density by $q_\theta \in Q_\theta$ and assume that the true density $p_{\theta_0} \in Q_\theta$.
- The difference in expected log-likelihood between the true density and any other density is given by

$$\mathbf{E}[\log q_\theta(X)] - \mathbf{E}[\log p_{\theta_0}(X)] = \int \left(\log q_\theta(x) - \log p_{\theta_0}(x)\right) p_{\theta_0}(x)\mathrm{d}x \tag{24}$$

$$= \int \log\left(\frac{q_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\mathrm{d}x \tag{25}$$

- Log is a concave function. We know from Jensen's inequality that $\mathbf{E}[\log(\xi)] \leq \log(\mathbf{E}[\xi])$

That means that

$$\int \log\left(\frac{q_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\mathrm{d}x \leq \log\left(\int \frac{q_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x)\mathrm{d}x\right) = 0 \quad (26)$$

We *only* get equality when $q_\theta(x) = p_{\theta_0}(x)$ for all values of x.
The conclusion is that

$$\theta_0 = \operatorname*{argmax}_{\theta\in\Theta} \mathbf{E}[\ell(\theta)] \quad (27)$$

and hence that

$$\lim_{N\to\infty} \hat{\theta}_{MLE} = \theta_0 \quad (28)$$

as the random error vanishes as $N \to \infty$.

That means that

$$\int \log\left(\frac{q_\theta(x)}{p_{\theta_0}(x)}\right) p_{\theta_0}(x)\mathrm{d}x \leq \log\left(\int \frac{q_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x)\mathrm{d}x\right) = 0 \quad (26)$$

We *only* get equality when $q_\theta(x) = p_{\theta_0}(x)$ for all values of x.
The conclusion is that

$$\theta_0 = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbf{E}[\ell(\theta)] \quad (27)$$

and hence that

$$\lim_{N\to\infty} \hat{\theta}_{MLE} = \theta_0 \quad (28)$$

as the random error vanishes as $N \to \infty$.

Proof Ib

- Rewrite $L(\mathbf{X}) = \exp\left(\log p_\theta(X_1) + \sum_{n=2}^{N} \log p_\theta(X_n|X_{1:n-1})\right)$.
- Second order Taylor expansion around θ_0.
- Maximize
- $\sqrt{N}(\hat{\theta} - \theta_0) \approx \ldots$
- Consistency and asymp normality follows.

Proof I, Ext.

- Write $\ell_N(\theta) = \log L(\mathbf{X}|\theta)$.
- This is a sum consisting of N terms (think LLN and CLT!).
- Approximate $\ell_N(\theta)$ using a second order Taylor expansion around θ_0.
- Thus

$$\ell_N(\theta) = \ell_N(\theta_0) + \partial_\theta \ell_N(\theta_0)(\theta - \theta_0) + \frac{1}{2}\partial_\theta^2 \ell_N(\theta_0)(\theta - \theta_0)^2 + R.$$

- Ignore the last term, multiply ℓ_N with $1/\sqrt{N}$ and maximize wrt θ to obtain $\hat{\theta}$.
- We obtain $\frac{1}{\sqrt{N}}\partial_\theta \ell_N(\theta_0) + \frac{1}{N}\partial_\theta^2 \ell_N(\theta_0)\sqrt{N}(\hat{\theta} - \theta_0) := 0.$

Proof I, Ext.

- We have from LLN that
$$\frac{1}{N}\partial_\theta^2 \ell_N(\theta_0) \to -I(\theta_0),$$
- and from CLT that
$$\frac{1}{\sqrt{N}}\partial_\theta \ell_N(\theta_0) \to Z,$$
where $Z \in \mathcal{N}(0, I(\theta_0))$.
- Rearranging gives $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0))$.

Proof II

- Denote the score function $U = \frac{\partial}{\partial \theta} \log_\theta p(X)$ and another estimator by T.
- Assume that $E[T] = \theta_0$ (to rule out anomalies).
- Calculate $Cov(U, T)$
- $Cov(U, T) = E[UT] - E[U]E[T] = E[UT] - \theta_0 \cdot 0$
- $E[UT] = \int T(x) \frac{\partial}{\partial \theta} \log p_{\theta_0}(x) p_{\theta_0}(x) \mathrm{d}x$
- $= \frac{\partial}{\partial \theta} \int T(x) p_{\theta_0}(x) \mathrm{d}x = \frac{\partial}{\partial \theta} \theta = 1.$
- Cauchy-Schwartz states $Cov(U, T) \leq \sqrt{Var[U] Var[T]}$.
- Thus $Var[T] \geq Var[\frac{\partial}{\partial \theta} \log p_\theta(X)]^{-1}$
- I.e. $Var[T] \geq Var[\hat{\theta}_{MLE}]$!

Proof III

- Original problem

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p_\theta(X_1, \ldots X_N).$$

- Define $Y = g(X)$.
- Calculations... Ok.

Other estimators

- We can derive the asymptotical distribution using the same arguments for any *M-estimator*, i.e. for any estimator taking the estimate as the value that maximized/minimizes a function of data.
- Similar arguments can also be used for *Z-estimators*, i.e. estimator taking the estimate as the value that solves a system of equations depending on data.

M-estimators

- Take any loss function $J(\theta)$, such that

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} J(\theta). \tag{29}$$

- E.g. PEM: $J(\theta) = \sum_{i=1}^{N} \varepsilon_i(\theta)^2$.
- The corresponding problem is

$$\frac{1}{\sqrt{N}} \partial_\theta J(\theta_0) + \frac{1}{N} \partial_\theta^2 J(\theta_0) \sqrt{N}(\hat{\theta} - \theta_0) := 0.$$

- Assume that $\mathbf{E}[\frac{1}{\sqrt{N}} \partial_\theta J(\theta_0)] = 0$ (Why?), $Var[\frac{1}{\sqrt{N}} \partial_\theta J(\theta_0)] = W$ and $\mathbf{E}[\frac{1}{N} \partial_\theta^2 J(\theta_0)] = V$. Applying the limit theorems as in the Maximum likelihood setup gives the asymptotics.
- Result: $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^{-1} W (V^{-1})^T)$.

Z-estimators

- Take any loss function $G(\theta)$, such that

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, G(\theta) = 0. \tag{30}$$

- E.g. Estimation functions $G(\theta) = \sum g_i(\theta)(X_i - E[X_i])$
- Compute a first order Taylor expansion around θ_0, and solve wrt θ.
- Thus $G(\theta_0) + \partial_\theta G(\theta_0)(\hat{\theta} - \theta_0) := 0$.
- Multiply by $1/\sqrt{N}$ and apply limit theorems.
- Let $\mathbf{E}[\frac{1}{\sqrt{N}} G(\theta_0)] = 0$ (why?), $Var[\frac{1}{\sqrt{N}} G(\theta_0)] = W$ and $\mathbf{E}[\frac{1}{N}\partial_\theta G(\theta_0)] = V$.
- Then $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V^{-1} W (V^{-1})^T)$.

The likelihood framework can be used to test the fit of models.

- Confidence intervals
- Likelihood ratio tests

Confidence intervals.

- Assume $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.
- Then $I_{\theta_0} \approx \hat{\theta} \pm 2\sqrt{\frac{\mathtt{diag}(\Sigma)}{N}}$.
- Better accuracy: Use profile likelihood!

Likelihood based tests.

- Likelihood ratio tests

$$\Lambda = \frac{\{\sup L(\mathbf{X}|\theta), \theta \in \Theta^F\}}{\{\sup L(\mathbf{X}|\theta), \theta \in \Theta^R\}}$$

  where $\Theta^R \subset \Theta^F$. Then

$$2\log(\Lambda) \in \chi^2(k),$$

  k being the degrees of freedom.
- LR tests are optimal, cf. Neyman-Pearson.
- and are closely related to AIC, BIC etc.

Likelihood ratios

- Compare $L(\mathbf{X}|\hat{\theta})$ to $L(\mathbf{X}|\theta_0)$.
- Taylor expand

$$2\log(\Lambda) = 2\left(\ell_N(\hat{\theta}) - \ell_N(\theta_0)\right).$$

- We have that

$$\ell_N(\theta) = \ell_N(\theta_0) + \partial_\theta \ell_N(\theta_0)(\theta - \theta_0) + \frac{1}{2}\partial_\theta^2 \ell_N(\theta_0)(\theta - \theta_0)^2 + R.$$

- The estimate must solve

$$\partial_\theta \ell_N(\theta_0) + \partial_\theta^2 \ell_N(\theta_0)(\hat{\theta} - \theta_0) = 0.$$

- Thus

$$2\log(\Lambda) = 2\left(\frac{1}{2}\partial_\theta \ell_N(\theta_0)(\hat{\theta} - \theta_0)\right).$$

Likelihood ratios, cont.

- Plugging in

$$(\hat{\theta} - \theta_0) = -(\partial_\theta^2 \ell_N(\theta_0))^{-1} \partial_\theta \ell_N(\theta_0)$$

  gives

- $$2\log(\Lambda) = -\partial_\theta \ell_N(\theta_0)(\partial_\theta^2 \ell_N(\theta_0))^{-1} \partial_\theta \ell_N(\theta_0)$$

- Or nicer written

$$2\log(\Lambda) = \frac{1}{\sqrt{N}} \partial_\theta \ell_N(\theta_0)(-\frac{1}{N} \partial_\theta^2 \ell_N(\theta_0))^{-1} \frac{1}{\sqrt{N}} \partial_\theta \ell_N(\theta_0).$$

- This is a quadratic form, distributed as $\chi^2(k)$, k being the degrees of freedom.

Feedback

Send feedback, questions and information about typos to `erikl@maths.lth.se`.