Model Selection

① Via testing ② Via generalization criteria, e.g. C_p, AIC or cross-validation

Testing

LRT (likelihood ratio test) see ch. 9

Logic behind this → assumption of a correct null model and look for evidence against it.

Example $\theta = (\theta_1, \theta_2)$, θ_1 = parameter(s) of interest, $\dim(\theta_1) = p_1$

- Arrive at test via e.g. profile likelihood

  where we replace θ_2 by the MLE for every value of θ_1 ⇒ $\hat{\theta}_2(\theta_1)$

  ⇒ Then focus on estimation problem of θ_1

  ⇒ $L(\theta_1) = \max_{\theta_2} L(\theta_1, \theta_2) = L(\theta_1, \hat{\theta}_2(\theta_1))$

  ⇒ Now compare $\max L(\theta_1) = L(\hat{\theta}_1)$ to $L(\theta_{10})$ null hyp: $\theta_1 = \theta_{10}$

  ⇒ Wilks stat $W = 2 \log \frac{L(\hat{\theta}_1)}{L(\theta_{10})} \sim \chi^2_{p_1}$ under null

- Can also think of this as comparing $L(\theta_1, \theta_2)$ under max with restrictions on θ_1 or no restrictions

  ⇒ $W = 2 \log \max_{\theta_1, \theta_2} L(\theta_1, \theta_2) \Big/ \max_{\theta_1 = \theta_{10}, \theta_2} L(\theta_1, \theta_2) \sim \chi^2_{p_1}$

General LRT result

$$W = 2 \log \frac{\max_{\Theta} L(\theta)}{\max_{H_0} L(\theta)}$$, Null H_0 puts restrictions on θ $(\theta_{0,5}, \theta_{,-} \theta_{,0})$

$W \sim \chi^2_{\Delta_p}$ under H_0, where Δ_p = # restricted θ under H_0.

- Result follows from quad approx of log-L and Asympt. normality of $\hat{\theta}$, and (conditional) Normality of $\hat{\theta} | \hat{\theta}(H_0)$
- (see Chapter 9)

Let's focus on AIC derivation

Some previous results we need:

$$L_f(\theta | x^n) = \prod_{i=1}^{n} f_\theta(x_i)$$, $$l_f(\theta) = \log L_f(\theta | x^n) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

(log) likelihood under model f_θ (which may be incorrect)

$\Rightarrow \hat{\theta}_f = \arg\max l_f(\theta)$, maximum value $l_f(\hat{\theta}_f)$

If we now consider models f_j, $j = 1 \ldots J$

where $\{f_j = f_{\theta_j}, \theta_j \in \Theta_j\}$, $p_j = \dim(\theta_j)$

$\Rightarrow$ we cannot use $l_{f_j}(\hat{\theta}_j)$ to choose between f_j's.

Reason why $l_f(\hat{\theta}_f)$ doesn't work → always max for model f with most free parameters $\dim(\theta_f)$.

⇒ also $l_f(\hat{\theta}_f)$ only says something about the fit on this data set x^n, but we want to know if f works on future data as well - generalization.

⇒ So, a more reasonable mode of comparison is to choose the model f that maximizes $E_g[\log L_f(\theta_f|X)]$

↑ new data X^{new}

But, this doesn't work since θ_f is unknown

so use

$$E_g[\log L_f(\hat{\theta}_f(x^n)|X)] \equiv Q_f$$

↑

expectation over x^n = training data used to estimate θ_f

AND

X = test data, used to evaluate how well $f_{\hat{\theta}_f}$ works on future data

So — how to get a good estimate of

$$Q_f = E\left[\log L_f(\hat{\theta}_f(X^n) \mid X)\right] \quad ?$$

E over X and X^n

Observable: $\frac{1}{n} l_f(\hat{\theta}_f(X^n)) = \frac{1}{n} \sum_{i=1}^{n} \log f_{\hat{\theta}_f}(X_i)$ — (*)

- Not a good estimate since $X = X^n$ used for both training & testing. (*) underestimates Q_f.

How big the bias is in (*) is a function of $\dim(\Theta_f)$.

Properties we will need later

- We have $\hat{\theta}_f = \underset{\theta_f}{\operatorname{argmax}}\, l_f(\theta_f \mid X^n)$
- We have $\frac{1}{n} \sum_{i=1}^{n} \log f_{\theta_f}(X_i) \to E_g\left[\log f_{\theta_f}(X)\right]$ by WLLN and max of LH $\to$ max of RH
- Say θ_0 maximizer of $E_g\left[\log f_{\theta_f}(X)\right]$

  then $\hat{\theta}_f \to \theta_0$
- θ_0 may not be a 'true parameter' in the sense of true g.

Now $\hat{\theta}_f = \underset{\theta_f}{\operatorname{argmax}}\, l_f(\theta_f | X^n)$ or equivalently

the zero-crossing of the score $S_f(\theta_f) = \frac{\partial}{\partial \theta_f} \sum_{i=1}^{n} \log f_{\theta_f}(X_i)$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_f} \log f_{\theta_f}(X_i)$$

Now $S_f(\theta_f) = S_f(\hat{\theta}_f) + \frac{\partial^2}{\partial \theta_f \partial \theta_f'} \log L_f(\theta_f | X^n)\Big|_{\theta^*} (\theta_f - \hat{\theta}_f)$ (from mean-value theorem)

- $S_f(\hat{\theta}_f)$: approx near $\hat{\theta}_f$
- $S_f(\hat{\theta}_f) = 0$ by definition
- $\frac{\partial^2}{\partial \theta_f \partial \theta_f'} \log L_f$: 2nd derivative of $\log L_f$ evaluated at θ^*

where $\|\theta_f - \theta^*\| \leq \|\theta_f - \hat{\theta}_f\|$ (somewhere in between θ and $\hat{\theta}$)

(I drop notation θ_f below and just use θ so it looks less messy)

Ok, so

$$S_f(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} \log L_f(\theta | X^n)\Big|_{\theta^*} (\theta - \hat{\theta})$$

$$= \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial \theta} \log f_\theta(X_i)}_{Y_i}, \quad Y_i = \frac{\partial}{\partial \theta} (\log f_\theta(X_i)) \text{ random variable}$$

So the score $S_f(\theta)$ is a sum of iid random variables $Y_i \longrightarrow$ CLT applies. Need to figure out what mean & variance is of the asymptotic distribution of $S_f(\theta)$.

- $* E(Y_i) = E_g\left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right] = \frac{\partial}{\partial\theta} E_g[\log f_\theta(X)]$
- $= 0$ at θ_0 since θ_0 is the maximizer of $E_g[\log f_\theta(X)]$
- $* \text{At } \theta_0, \ Var(Y_i) = E(Y_i^2) = E_g\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)\Big/\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)'\Big|_{\theta_0}\right]$

$= \mathcal{J}(\theta_0)$

- $* \text{By CLT}, \ \frac{1}{\sqrt{n}} S_f(\theta_0) \to^d N(0, \mathcal{J}(\theta_0))$

We have

$$\frac{1}{\sqrt{n}} S_f(\theta_0) = \frac{1}{\sqrt{n}} \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} (\theta_0 - \hat{\theta})$$

$\to^d N(0, \mathcal{J}(\theta_0))$; want distribution of this

Focus on $\frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta^*} = (**)$

Now, since $\|\theta-\theta^*\| \le \|\theta-\hat{\theta}\|$ and $\hat{\theta} \to \theta_0$

at θ_0 as $n \to \infty$ $(**) \approx \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta_0}$

- Also $\frac{1}{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log L_f(\theta|X^n)\Big|_{\theta_0}$
- $= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(x_i)\Big|_{\theta_0}$

$\to E_g\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(X)\Big|_{\theta_0}\right]$ by WLLN

$= -FI(\theta_0)$ by definition

- $\frac{1}{\sqrt{n}} S_f(\theta_0) \approx \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(x_i)\Big|_{\theta_0}\right] \sqrt{n}(\theta_0 - \hat{\theta})$
- $\to^d N(0, J(\theta_0))$

$\to -FI(\theta_0)$

By Slutsky's $\sqrt{n}(\hat{\theta}-\theta_0) \to^d N(0, FI(\theta_0)^{-1} J(\theta_0) FI(\theta_0)^{-1})$

Example

Assume $f_\lambda(x) = \lambda e^{-\lambda x}$, $\lambda > 0$, $x \geq 0$ (exponential)

$$\log f_\lambda(x) = -\lambda x + \log(\lambda)$$

$$l_f = \sum_{i=1}^{n} -\lambda x_i + \log(\lambda) = -\lambda \Sigma x_i + n \log(\lambda)$$

$$\frac{\partial l_f}{\partial \lambda} = -\Sigma x_i + \frac{n}{\lambda} \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$$

Now $E_g \log f_\lambda(X) = E_g[-\lambda X + \log(\lambda)] = -\lambda E(X) + \log(\lambda)$

$\Rightarrow$ maximized by $\lambda_0 = \frac{1}{E(X)}$

$\hat{\lambda} \to \lambda_0$

If we believe f_λ is the model, Asympt Variance of $\hat{\lambda}$ is $\frac{FI_{\lambda_0}^{-1}}{n} = -\frac{1}{n} E\left(\frac{\partial^2}{\partial \lambda^2} \log f_\lambda(X) \Big|_{\lambda_0} \right)^{-1}$

$$= -\frac{1}{n}\left(E\left[-\frac{1}{\lambda_0^2} \right] \right)^{-1} = \frac{\lambda_0^2}{n}$$

so $\hat{\lambda} \sim N\left(\lambda_0, \frac{\lambda_0^2}{n}\right)$ in practice use $\hat{\lambda} - \lambda_0 \sim N\left(0, \frac{\hat{\lambda}^2}{n}\right)$ for testing etc

What if $f \neq g$

$\Rightarrow J(\lambda_0) = E\left(\left(\frac{\partial}{\partial \lambda} \log f_\lambda(x)\right)^2 \Big|_{\lambda_0}\right)$

$= E\left(-X + \frac{1}{\lambda_0}\right)^2 = E\left(X - \frac{1}{\lambda_0}\right)^2 = E(X - E(X))^2 = V(X)$

/

depends on g.

$\Rightarrow$ robust variance

$J I_{\lambda_0}^{-1} J J_{\lambda_0} J I_{\lambda_0}^{-1} = \lambda^2 V(X) \lambda^2 = \lambda^4 V(X) =$ (***)

Example, $g \sim \chi^2_p$ $E(X) = p$ $V(X) = 2p$, p = positive integer

$\Rightarrow \lambda = \frac{1}{p}$

$\Rightarrow$ (***) $\sim \frac{2}{p^3} = 2\lambda^3$ compared with $\lambda^2, \frac{1}{p^2}$