April 23, 2002

I. Order Determination.

(Ref: E. L. Lehman Testing Statistical Hypothesis, Wiley, N.Y).

$$Y = X\theta^* + E$$

$$Y = (y_1, \dots, y_p)^T \quad \theta^* = (b_1 \dots b_q)^T$$

$$E = (e_1 \dots e_p)^T \quad Cov(E) = \sigma^2 I$$

$$X = \begin{bmatrix} u_0 & \cdots & u_{1-q} \\ u_1 & \cdots & u_{2-q} \\ \vdots & & \vdots \\ u_{p-1} & & u_{p-q} \end{bmatrix} \quad p \times q \text{ matrix (known)}$$

$p > q$; linearly independent columns.

θ^* : vector of parameters to be estimated from data.

Let d, $0 \le d \le q$ be given

Problem: Decide on the basis of observations when the hypothesis

(H) $\theta^*_q = \theta^*_{q-1} = \cdots = \theta^*_{q-d+1} = 0$ should be rejected.

Statistical Tests

Let $k_1, k_2 > 0$. A random variable X is said to have $\chi^2(k_1)$ distribution if

$$X = V_1^2 + \cdots + V_{k_1}^2 \quad \text{where } V_i \sim N(0,1) \quad i = 1 \cdots k_1$$

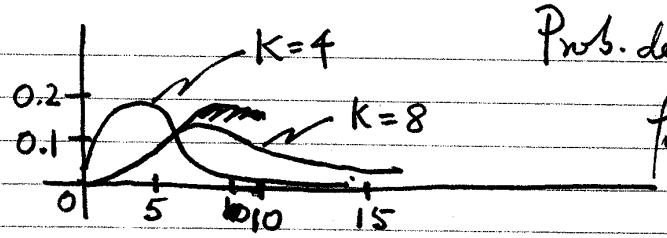
A random variable X is said to have a $F(k_1, k_2)$ distribution if it can be expressed as

$$X = \frac{\frac{S_1}{k_1}}{\frac{S_2}{k_2}}$$

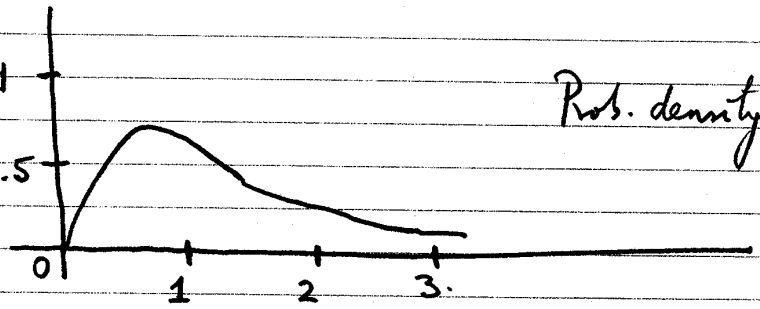
where S_1 and S_2 are independent with distributions $\chi^2(k_1)$ and $\chi^2(k_2)$ respectively.

(Define $\chi^2(0)$ to be the distribution of the degenerate random variable taking value 0, a.s.).

For large values of k_2, if X has $F(k_1, k_2)$ distribution then to good approximation $k_1 X$ has $\chi^2(k_1)$ distribution.



Prob. density fn of $\chi^2(k)$ for $k = 4, 8$



Prob. density fn. $F(4,8)$

Let $\hat{\theta}$ = least squares estimate of θ^*

$\hat{\hat{\theta}}$ = least squares estimate of θ^* under hypothesis (H).

Now

$$\hat{\hat{\theta}} = (\theta_0, 0, \dots, 0) \quad \text{where } \theta_0 = (X_0^T X_0)^{-1} X_0^T Y$$

where X_0 is $p \times q$-d matrix obtained from X by removing the last d columns. (or replacing by 0).

$$\text{let } \varepsilon(\theta) = Y - X\theta, \quad \theta \in \mathbb{R}^q \qquad (1)$$

$$\text{and } S(\theta) = \varepsilon^T(\theta)\varepsilon(\theta) \qquad (2)$$

Consider the <u>statistic</u>

$S(\hat{\hat{\theta}}) - S(\hat{\theta})$ which measures the increase in the minimum of the least squares criterion

when we decrease the number of parameters from q to $q-d$. One would reject the hypothesis if $S(\hat{\hat{\theta}}) - S(\hat{\theta})$ is large.

Proposition 1

Suppose that for some integer d, $0 \le d \le q$ hypothesis (H) is true. (For $d=0$, no restrictions on θ^*). Let $\hat{\theta}$ and $\hat{\hat{\theta}}$ be as above and let $S(\theta)$ be defined by (2). Then $\hat{\hat{\theta}}$, $S(\hat{\theta})$ and $S(\hat{\hat{\theta}}) - S(\hat{\theta})$ are independent. Moreover

$$\frac{S(\hat{\theta})}{\sigma^2} \sim \chi^2(p-q) \quad \text{and} \quad \frac{1}{\sigma^2}\left(S(\hat{\hat{\theta}}) - S(\hat{\theta})\right) \sim \chi^2(d)$$

(Note: For $d=0$, $\hat{\theta}$ and $S(\hat{\theta})$ are independent and $\frac{1}{\sigma^2} S(\hat{\theta}) \sim \chi^2(p-q)$).

Proof: First assume: $0 < d < q$. From least squares theory

$$Y - X\hat{\theta} \perp \text{Range}(X) \qquad (3)$$

$$Y - X\hat{\hat{\theta}} \perp \text{Range}(X_0) \qquad (4)$$

$$\Rightarrow X(\hat{\theta} - \hat{\hat{\theta}}) \perp \text{Range}(X_0) \qquad (5)$$

Now from (3) and (4)

$$\|Y - X\hat{\theta}\|^2 = \|(Y - X\hat{\theta}) + X(\hat{\theta} - \hat{\hat{\theta}})\|^2.$$

$$= \|Y - X\hat{\theta}\|^2 + \|X(\hat{\theta} - \hat{\hat{\theta}})\|^2.$$

i.e. $$S(\hat{\hat{\theta}}) - S(\hat{\theta}) = \|X(\hat{\theta} - \hat{\hat{\theta}})\|^2 \qquad (6)$$

Let $b_1, \ldots, b_p$ be an orthonormal basis with the following properties

(a) $b_1 \ldots, b_{q-d}$ spans Range(X_0).

(b) $b_{q-d+1} \ldots b_q$ are orthogonal to the Range(X_0) s.t. $b_1 \ldots b_q$ spans the range of X

(c) $b_{q+1}, \ldots, b_p$ are orthogonal to Range(X).

The linear independence hypothesis on the columns of X guarantees that we can choose such as basis.

Let $B := (b_1 \ldots b_p)$ and set

$$V = B^T E \qquad (7)$$

(Note: to prove something has χ^2-distribution we have to show that the r.v. can be written as the sum of indep. square of gaussian r.v's

The components of V defined by (7) will be the gaussian r.v's)

Since the columns of B are ~~independent~~ orthonormal

$$B^T = B^{-1} \qquad (8)$$

and $\mathrm{Cov}(V) = \sigma^2 I$ and hence $V \sim N(0, \sigma^2 I)$.

Now

$$E = Y - X\theta^* = (Y - X\hat{\theta}) + X(\hat{\theta} - \hat{\hat{\theta}}) + X(\hat{\hat{\theta}} - \theta^*).$$

Hence using (8)

$$V = B^{-1}(Y - X\hat{\theta}) + B^{-1}X(\hat{\theta} - \hat{\hat{\theta}}) + B^{-1}X(\hat{\hat{\theta}} - \theta^*). \qquad (9)$$

Now the mapping $X \not\Phi \longmapsto B^{-1}X$ transforms the coordinates from the standard basis of $\mathbb{R}^p$ to the coordinates w.r.t. the orthonormal basis $(b_1 \ldots b_p)$. Now

$Y - X\hat{\theta} \perp$ to $b_1 \ldots b_q$ from (3) and hence

$$B^{-1}(Y - X\hat{\theta}) \in \{\xi \in \mathbb{R}^p \mid \xi_1 = \ldots = \xi_q = 0\}. \qquad (10)$$

From (5) $X(\hat{\theta} - \hat{\hat{\theta}}) \in \mathrm{Range}(X)$ but is orthogonal to range of X_0.

and hence $B^{-1}X(\hat{\theta}-\hat{\hat{\theta}}) \in \left\{ \xi \in \mathbb{R}^p \,\middle|\, \xi_1 = \cdots = \xi_{q-d} = \xi_{q+1} = \cdots = \xi_p = 1 \right\}$.

Finally, $X(\hat{\hat{\theta}}-\theta^*) \in \text{Range}(X_0)$ and hence

$$B^{-1}X(\hat{\hat{\theta}}-\theta^*) \in \left\{ \xi \in \mathbb{R}^p \,\middle|\, \xi_{q-d+1} = \cdots = \xi_p = 0 \right\}$$

Therefore:

$$B^T(Y - X\hat{\theta}) = (0, \cdots, 0, V_{q+1}, \cdots V_p)^T \quad (11)$$

$$B^T X(\hat{\theta}-\hat{\hat{\theta}}) = (0, \cdots, 0, V_{q-d+1}, \cdots, V_q, 0 \cdots 0)^T \quad (12)$$

$$B^T X(\hat{\hat{\theta}}-\theta^*) = (V_1, \cdots, V_{q-d}, 0, \cdots 0)^T. \quad (13)$$

Therefore from (6)

$$S(\hat{\hat{\theta}}) - S(\hat{\theta}) = \|X(\hat{\theta}-\hat{\hat{\theta}})\|^2 = \|B^T X(\hat{\theta}-\hat{\hat{\theta}})\|^2$$

$$= \sum_{i=q-d+1}^{q} V_i^2 \quad (\text{from } (12)). \quad (14$$

$$S(\hat{\theta}) = \|Y - X\hat{\theta}\|^2 = \|B^T(Y - X\hat{\theta})\|^2 = \sum_{i=q+1}^{p} V_i^2 \quad (15)$$

Now

$$\hat{\hat{\theta}} - \theta^* = (X^TX)^{-1}X^TX(\hat{\hat{\theta}}-\theta^*) = (X^TX)^{-1}X^TBB^TX(\hat{\hat{\theta}}-\theta^*)$$

$$= (X^TX)^{-1}X^TB(V_1 \cdots V_{q-d}, 0, \cdots 0)^T$$

that is

$$\hat{\hat{\theta}} = \theta^* + \left(X^T X\right)^{-1} X^T B \left(v_1, \ldots v_{q-d}, 0, \ldots 0\right)^T. \quad (16)$$

From (14),

$\frac{1}{\sigma^2} S\left(\hat{\hat{\theta}}\right) - S\left(\hat{\theta}\right)$ has distribution $\chi^2(d)$

From (15),

$S\left(\hat{\theta}\right)$ has distribution $\chi^2(p-q)$.

and (14), (15) and (16) shows that

$S\left(\hat{\hat{\theta}}\right) - S\left(\hat{\theta}\right)$, $S\left(\hat{\theta}\right)$ and $\hat{\hat{\theta}}$ are independent.

The case $d = 0$ and q can be dealt with using similar arguments.

Case σ^2 known

The above proposition shows that the statistic

$$\frac{1}{\sigma^2}\left[S\left(\hat{\hat{\theta}}\right) - S\left(\hat{\theta}\right)\right]$$ has distribution $\chi^2(d)$

under the hypothesis (H).

Let K_α be the upper-α percentile of the $\chi^2(d)$ distribution, that is, the event $X \geq K_\alpha$ has probability α. Then the probability that the event

$$S(\hat{\hat{\theta}}) - S(\hat{\theta}) > \sigma^2 K_\alpha$$ will occur

when (H) is true is α. Therefore there is good evidence for rejecting hypothesis (H) if the above inequality holds for some pre-set value α (0.05 say).

Case σ^2 unknown

Use the property that

$$\frac{S(\hat{\hat{\theta}}) - S(\hat{\theta})}{d} \Big/ \frac{S(\hat{\theta})}{p-q} \sim F(d, p-q). \quad (17)$$

(Want something independent of σ^2).

As before, let K_α denote the upper α-percentile of the $F(d, p-q)$ distribution

The event

$$\left(\frac{S(\hat{\theta}) - S(\hat{\theta})}{d.} \Big/ \frac{S(\hat{\theta})}{p-q} \right) > k_\alpha \qquad (18)$$

has probability α if (H) is true and there is good evidence for rejecting (H) if $\alpha = 0.05$, say.

Procedure for Model Order Selection.

The above suggests a procedure for model order selection.

Let S_n = min. value of least squares criterion over vectors of parameters of dimension n, $n = 1, 2 \ldots$.

Assume $p >> q \Rightarrow q/p$ small

$\Rightarrow F(1, p-q) \cong \chi^2(1)$.

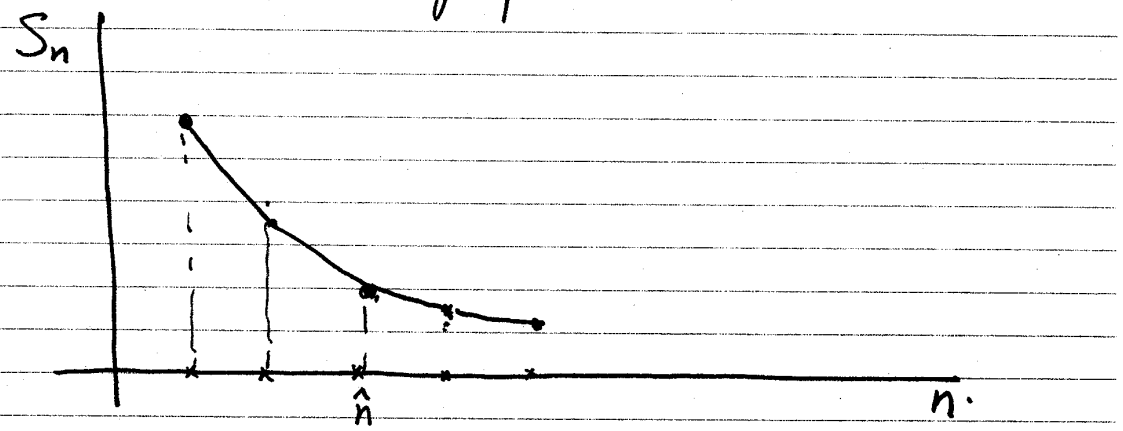
From the proposition, and the above, if n is a possible ~~code~~ model order

$$\frac{\frac{S_n - S_{n+1}}{S_{n+1}}}{p} \quad \text{has } \chi^2(1) \text{ distribution}$$

The 0.01 percentile of $\chi^2(1) \simeq 4$ (See Statistical tables). There are therefore grounds then for rejecting n as a possible model order at approx. 5% risk level if the inequality

$$S_n - S_{n+1} > K \frac{S_{n+1}}{p} \quad \text{where } K = 4.$$

Consider now the graph



Above suggests, estimate $\hat{n}$ of model order be chosen to satisfy

$$S_{n-1} - S_n > K \frac{S_{n.}}{p}$$

$$S_n - S_{n+1} \leq K \frac{S_{n+1}}{p} \qquad (19)$$

for some pre-set $K = 4$ (say).

Let us try to interpret these inequalities

View S_n, $n = 1, \cdots p$ as a uniform discretization of a C^1-function $g : [0,1] \longrightarrow \mathbb{R}$

i.e.

$$g\left(\frac{n}{p}\right) = S_n \quad , \quad n = 1, 2, \cdots p.$$

Now $\dfrac{S_n - S_{n-1}}{1/p}$ is a finite difference approx.

to $-\dfrac{d}{dx} g(x)$ at $\hat{x} = \dfrac{\hat{n}}{p}$.

Interpret (19) as

$$\left. -\frac{\frac{d}{dx} g(x)}{g(x)} - K \right|_{x = \frac{\hat{n}}{p}} = 0$$

or

$$\left. \frac{d}{dx} \left[\log g(x) + Kx \right] \right|_{x = \frac{\hat{n}}{p}} = 0$$

This says that

$\log g(x) + Kx$ assumes a minimum at $X = \dfrac{\hat{n}}{p}$

(Necessary condn).

Interpret as :

$\hat{n}$ minimizes $f(n) = \log S_n + \frac{K \cdot}{p} n$

Now p is fixed and hence $\hat{n}$ minimizes

$$A(n) = \log \frac{S_n \cdot}{p} + \tilde{K} \cdot n \quad \text{where } \tilde{K} = \frac{K \cdot}{p \cdot}$$

for $\tilde{K}$ at some pre-set level.

Accuracy of Estimates

Trustworthiness of estimated components $\hat{\theta}_i$ can be gauged from the α-confidence region for $\hat{\theta}_i$, which is an interval $I(\hat{\theta})$ and has the property that the event

$$\{\theta_i^* \in I(\hat{\theta})\} \text{ occurs with probability } \alpha$$

Case: σ^2 known

The estimate $\hat{\theta}$ is linear, unbiased and has covariance matrix $\sigma^2 (X^T X)^{-1}$ and $E \sim N(0, \sigma^2 I)$ it follows that

$$\hat{\theta} \sim N\left(\theta, \sigma^2 (X^T X)^{-1}\right)$$

Hence

$$\hat{\theta}_i \sim N\left(\theta_i^*, \sigma^2 C_{ii}\right) \quad i = 1, 2 \cdots q$$

where $(C_{ij}) = (X^T X)^{-1}$. Hence

$$\frac{\left(\hat{\theta}_i - \theta_i^*\right)}{\sqrt{\sigma^2 C_{ii}}} \sim N(0,1)$$

Let K_β be the upper $\beta/2$ percentile for the distribution $N(0,1)$. Then the confidence region can be constructed as follows:

$$\left(\hat{\theta}_i - K_\beta \sqrt{\sigma^2 C_{ii}},\ \hat{\theta}_i + K_\beta \sqrt{\sigma^2 C_{ii}}\right)$$

is a $(1-\beta)$ confidence region for θ_i^*, $i = 1, \cdots q$.

<u>Case 2</u> σ^2 Unknown

Construct regions from the statistic

$$\hat{\theta}_i \Big/ \sqrt{\hat{\sigma}^2 C_{ii}}, \text{ where}$$

$\hat{\sigma}$ is the unbiased estimate of σ, given by

$$\hat{\sigma}^2 = (p-q)^{-1} \left(Y - X\hat{\theta}\right)^T \left(Y - X\hat{\theta}\right)$$

(Deduce this)

We now have to introduce another distribution. Given an integer $k>0$, a random variable V is said to have the $t(k)$-distribution if it can be expressed as

$$V = \frac{U}{\sqrt{W/k}}$$

where U and W are independent random variables with distributions $N(0,1)$ and $\chi^2(k)$ respectively.

Now for $i=1, 2, \ldots q$

$$\frac{\hat{\theta}_i - \theta_i^*}{\sqrt{c_{ii}}} \Big/ \sqrt{\hat{\sigma}^2} = \frac{\hat{\theta}_i - \theta_i^*}{\sqrt{\sigma^2 c_{ii}}} \Big/ \frac{\|Y - X\hat{\theta}\|^2}{\sqrt{(p-q)\sigma^2}}$$

Now $\dfrac{\hat{\theta}_i - \theta_i^*}{\sqrt{\sigma^2 c_{ii}}} \sim N(0,1).$

From the Proposition with $d=0$, we know $\hat{\theta}_i$ is independent of $\|Y - X\hat{\theta}\|^2$ and

$$\frac{1}{\sigma^2}\|Y - X\hat{\theta}\|^2 \sim \chi^2(p-q).$$

$$\Rightarrow \quad \frac{\hat{\theta}_i - \theta_i^*}{\sqrt{c_{ii}}} \Big/ \sqrt{\hat{\sigma}^2} \sim t(p-q).$$

Let k_β be the upper $\beta/2$ percentile for the distribution $t(p-q)$. Hence

$\left(\hat{\theta}_i - k_\beta \sqrt{\hat{\sigma}^2 c_{ii}}, \hat{\theta}_i + k_\beta \sqrt{\hat{\sigma}^2 c_{ii}}\right)$ is a $(1-\beta)$ confidence region for $\hat{\theta}_i$, $i = 1, 2 \ldots q$.

II. Maximum Likelihood Estimation

Predictor Model

$$Y_K = f_K\left(\theta; Y^{K-1}, U^{K-1}\right) + E_K \quad , K=1,2,\ldots N \qquad (1)$$

$(E_1, \ldots, E_N)$ has joint density function which depends on the parameter θ; Assume that $U_0, \ldots U_{N-1}$ are independent of $E_1, E_2, \ldots E_N$

Let $p\left(y_N, \ldots, y_1 \middle| \theta, U^{N-1}\right)$ be the joint density function of Y^N given θ and U^{N-1}.

ML-estimate is given by.

$$\max_{\theta} V_N\left(\theta, Y^N, U^{N-1}\right)\left[p\left(y_N, \ldots, y_1 \middle| \theta, U^{N-1}\right)\right]$$

Let $E_K \sim N(0, \Lambda_K)$, $\Lambda_K > 0$

and E_K's independent. Then a calculation using Bayes rule gives us

$$\log p\left(y_N, \ldots, y_1 \middle| U^{N-1}, \theta\right)$$

$$\log p\left(y_N, \dots, y_1 \middle| u^{N-1}, \theta\right).$$

$$= -\frac{Nr}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{N} \log\det \Lambda_k.$$

$$-\frac{1}{2}\sum_{k=1}^{N} \left\| y_k - \hat{Y}_k(\theta) \right\|^2_{\Lambda_k^{-1}}.$$

Max. Likelihood estimate criterion

$$\max_{\theta} \; L(\theta)$$

where $L(\theta) = \sum_{k=1}^{N} \left\| \varepsilon_k(\theta) \right\|^2_{\bar{\Lambda}_k^{-1}(\theta)} + \sum_{k=1}^{N} \log\det \Lambda_k(\theta).$