

Maximum likelihood

$$\text{let } Y_i \sim f(y_i | \theta)$$

$$\Pr(Y_i \leq y_i) = F(y_i | \theta)$$

Y is a random variable and has p.d.f. $f(y|\theta)$ and cdf $F(y|\theta)$, $\theta \in \Omega$

If we have a sample of size n (randomly drawn)

$$\Pr(Y_1 \leq y_1, Y_2 \leq y_2, \dots, Y_n \leq y_n) = F(\underline{y} | \theta)$$

knowledge of θ allows one to make probability statements about y . The likelihood flips this around: we know y and want to make prob statements about θ .

$$\text{let } L(\underline{\theta} | \underline{y}) \equiv f(\underline{y} | \underline{\theta}) \quad \theta \in \Omega$$

where $f(\underline{y} | \theta)$ is the joint p.d.f.

$$L(\underline{\theta} | \underline{y}) = l(\theta_1 | y_1) \cdot l(\theta_2 | y_2) \cdot l(\theta_3 | y_3) \cdots l(\theta_n | y_n)$$

where $l(\theta | y_i) \equiv f(y_i | \theta)$

For analytical & computational reasons
it's common to work with the natural
log of $L(\underline{\theta} | \underline{y})$

$$\ln L(\underline{\theta} | \underline{y}) = \sum_{i=1}^n \ln l(\underline{\theta} | y_i)$$

$$g(\underline{\theta}) = \frac{\partial \ln L(\underline{\theta} | \underline{y})}{\partial \underline{\theta}} \quad \text{likelihood E.g.'s. (gradient)}$$

$$H(\underline{\theta}) = \frac{\partial g(\underline{\theta})}{\partial \underline{\theta}^T} \quad \text{Hessian}$$

Recall that the asymptotic covariance
of $\hat{\theta}$ (MLE) is

$$\left(-E[H(\theta_T)] \right)^{-1} \text{Var}[g(\theta_T)] \left(-E[H(\theta_T)] \right)^{-1}$$

$$\text{and } \text{Var}[g(\theta_T)] = -E[H(\theta_T)]$$

Hence, the variance covariance
of the MLE is

$$\underline{\Sigma}_{MLE} \equiv -E[H(\theta_T)]^{-1} \equiv I(\theta_T)^{-1}$$

$I(\theta_T)$ is called the information
matrix. We showed that

$$-E[H(\theta_T)] = E \left[\underset{k \times 1}{g(\theta_T)} \underset{1 \times k}{g(\theta_T)^T} \right]$$

So, there are 2 equivalent expressions
for the var/cov of the MLE.

Estimating $\Sigma_{MLE} = -E[H(\theta_T)]^{-1}$

(1) $-H(\hat{\theta})^{-1}$

(2) $-E[H(\hat{\theta})]^{-1} = I(\hat{\theta})^{-1}$

(3) $[G(\hat{\theta})^T G(\hat{\theta})]^{-1}$ OPG or BHHH

$$G(\hat{\theta}) = \begin{pmatrix} \frac{\partial \ln L(\theta|y_1)}{\partial \theta_1} & \frac{\partial \ln L(\theta|y_1)}{\partial \theta_2} & \dots & \frac{\partial \ln L(\theta|y_n)}{\partial \theta_k} \\ \frac{\partial \ln L(\theta|y_2)}{\partial \theta_1} & & & \\ \vdots & & & \\ \frac{\partial \ln L(\theta|y_n)}{\partial \theta_1} & & & \frac{\partial \ln L(\theta|y_n)}{\partial \theta_k} \end{pmatrix}$$

$n \times k$

Note: Add the elements of each column to get $g(\hat{\theta})^T$.

$$g(\hat{\theta}) = \begin{pmatrix} \sum_{i=1}^n \frac{\partial \ln L(\theta|y_i)}{\partial \theta_1} \\ \vdots \\ \sum_{i=1}^n \frac{\partial \ln L(\theta|y_i)}{\partial \theta_k} \end{pmatrix}$$

$k \times 1$

1 x k

Tests

3 Classical Tests

Likelihood Ratio }
Wald } Asymptotically
Lagrange Multiplier. } Equivalent.

Let $l(\hat{\theta} | y) \equiv \ln L(\hat{\theta} | y)$ be
the log-likelihood function. also,
 $r(\theta)$ is a set of J smooth
functions of the K parameters.

$$H_0: r(\theta) = 0$$

$$H_A: r(\theta) \neq 0$$

Likelihood Ratio

Let $\hat{\theta}$ be the Restricted MLE and
 $\hat{\theta}$ be the unrestricted. This
implies that you have some
way of either imposing the
 $r(\theta)$ restrictions on the model
or can do the constrained

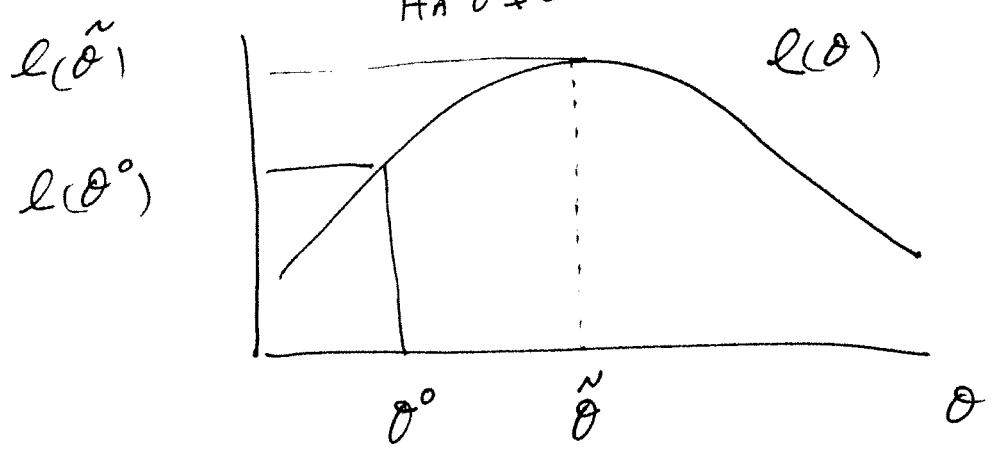
Optimization.

$$LR = 2 \left(l(\tilde{\theta}) - l(\hat{\theta}) \right) \stackrel{a}{\sim} \chi^2_5 \text{ if } H_0 \text{ True.}$$

The restricted log-likelihood will be smaller than unrestricted, \therefore

$$LR \geq 0$$

$H_0: \theta = \theta^0$
 $H_A: \theta \neq \theta^0$



As $l(\tilde{\theta})$ gets further away from $l(\theta^0)$ - The evidence against H_0 increases.

Wald Test

$$\text{Var}[\tilde{r}(\tilde{\theta})] \stackrel{a}{\sim} \underset{J \times K}{R(\tilde{\theta})} \underset{K \times K}{\text{Var}(\tilde{\theta})} \underset{K \times J}{R(\tilde{\theta})}^T$$

$$R(\tilde{\theta}) = \underset{J \times K}{\frac{\partial r(\tilde{\theta})}{\partial \tilde{\theta}^T}} \underset{1 \times K}{}$$

$$W = \tilde{r}(\tilde{\theta})^T [R(\tilde{\theta}) \text{Var}(\tilde{\theta}) R(\tilde{\theta})^T]^{-1} \tilde{r}(\tilde{\theta})$$

$\stackrel{a}{\sim} \chi^2_J$ if H_0 True.

NOTES: ① We have at least 3 different estimates of $\text{Var}(\tilde{\theta})$.

② The value of W in small samples depends on how you write the restrictions

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$H_0: \beta_2 \cdot \beta_3 = 1$$

$$H_A: \beta_2 \cdot \beta_3 \neq 1$$

$$r_1 \quad \beta_2 \cdot \beta_3 - 1 = 0$$

$$r_2 \quad \beta_2 \neq \beta_2 - 1/\beta_3 = 0$$

$$r_3 \quad \beta_3 - \frac{1}{\beta_2} = 0$$

leads to different W .

③ Favorite method in canned software, though it may be a poor choice.

~~④ In some cases it is not~~

LM Test

This version is very useful when the model under HA. is difficult to estimate.

$L = l(\theta) - r(\theta)^T \lambda$ Lagrangian

$r(\theta)^T$ $1 \times J$
 λ $J \times 1$

vector of Lagrange multipliers.

The FOC.

$\frac{\partial L}{\partial \theta} = g(\theta) - R(\theta)^T \lambda = 0$
 $K \times 1$ $K \times J$ $J \times 1$

$\frac{\partial L}{\partial \lambda} = r(\theta) = 0$

$$R(\theta) = \frac{\partial r(\theta)}{\partial \theta^T}$$

$J \times K$ $J \times 1$
 $J \times K$ $1 \times K$

$$g(\hat{\theta}) = R(\hat{\theta})^T \lambda \quad \text{under FOC}$$

Two versions of The LM stat

$$LM_1 = g(\hat{\theta})^T I(\hat{\theta})^{-1} g(\hat{\theta})$$

~~$1 \times K$~~ ~~$K \times K$~~ $1 \times K$ $K \times K$ $K \times 1$

$\hat{\theta}$ is the restricted MLE
 i.e., the MLE of the Model
 under H_0 :

$$LM_2 = \lambda^T R(\hat{\theta}) I(\hat{\theta})^{-1} R(\hat{\theta})^T \lambda$$

Artificial Regressions.

As with W , There are as many ways to compute LM as there are ways to estimate the $I(\theta)$ information matrix.

For example, computing $I(\theta)$ using the OPG, LM, becomes

$$g(\hat{\theta})^T [G(\hat{\theta})^T G(\hat{\theta})]^{-1} g(\hat{\theta})$$

which can be computed using an artificial regression.

$$\underset{\sim}{i}_N = \underset{N \times K}{G(\theta)} \underset{\sim}{b} + \underset{\sim}{\text{res}}$$

$\underset{\sim}{b}$ are unknown parameters.

let $G \equiv G(\theta)$

$$\hat{\underset{\sim}{b}} = (G^T G)^{-1} G^T \underset{\sim}{i}_n \quad \hat{\underset{\sim}{i}}_m = G (G^T G)^{-1} G^T \underset{\sim}{i}_n$$

Model sum-of-squares $\hat{\underset{\sim}{y}}^T \hat{\underset{\sim}{y}} = \hat{\underset{\sim}{i}}^T \hat{\underset{\sim}{i}}$

$$= \underset{\sim}{i}_n^T G (G^T G)^{-1} G^T G (G^T G)^{-1} G^T \underset{\sim}{i}_n$$

$$= \underset{\sim}{i}_n^T G (G^T G)^{-1} G^T \underset{\sim}{i}_n$$

$$\hat{C}_m^T b = g \quad \therefore$$

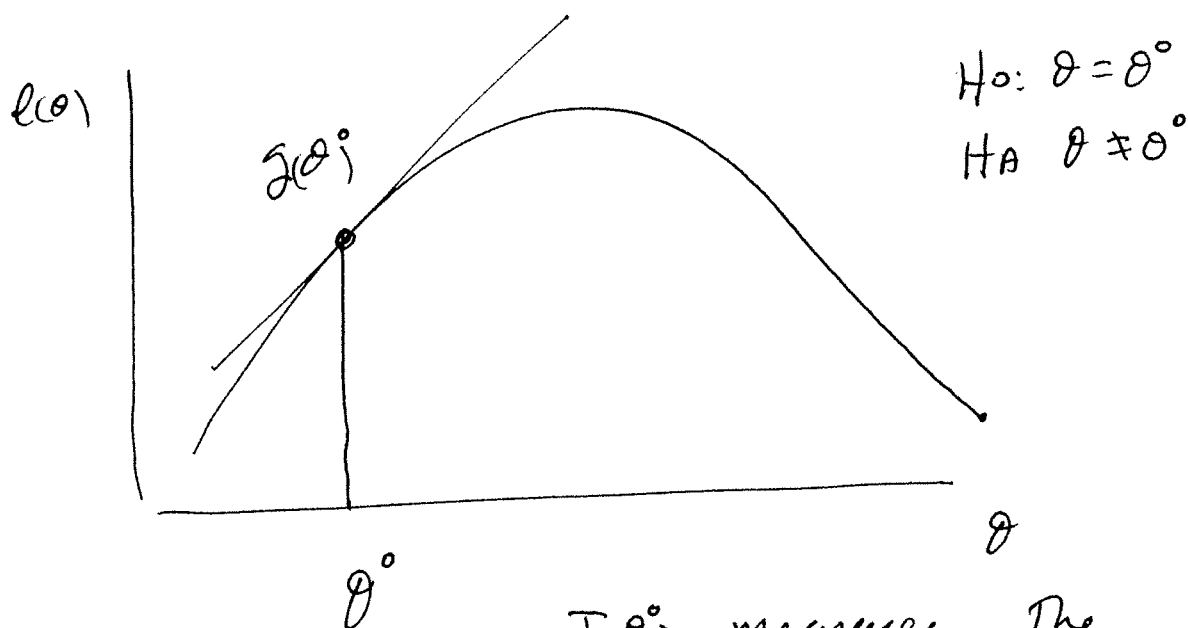
Model sum of squares in Three
Regression is LM_1 !

OR Recall that in linear regression

$$TSS = MSS + SSE$$

$$TSS = y^T y = \hat{C}_n^T \hat{C}_n = n$$

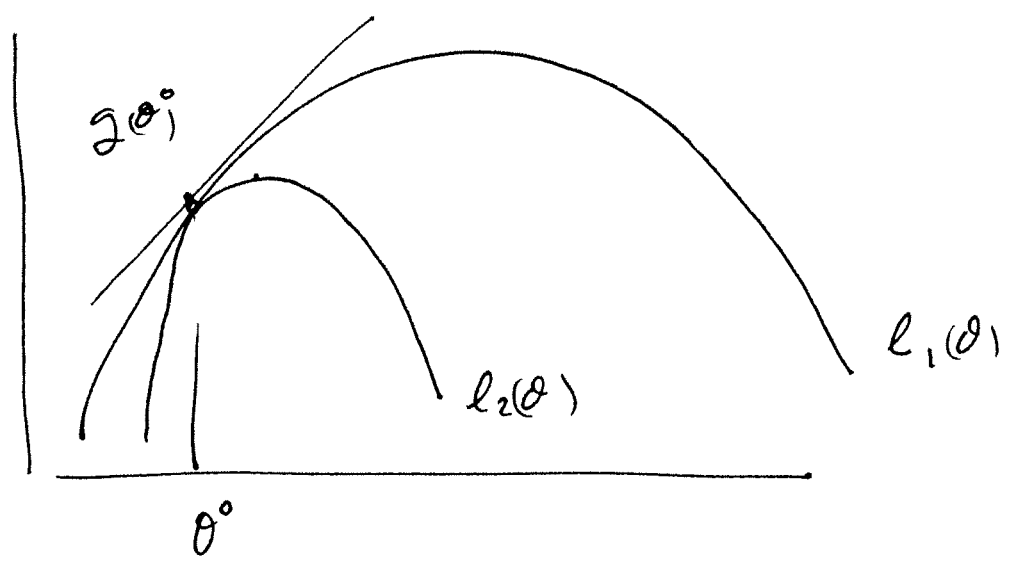
$LM_1 = n - SSE$ from the artificial
regression.



$I(\theta^0)$ measures the
curvature of $l(\theta)$ at
the restriction. Hence
 LM is based on slope of
 l at restriction

When $I(\theta)$ is "small"
 $I(\theta)^{-1}$ is "large".

small $I(\theta) \Rightarrow$ low information
low information \Rightarrow l is relatively flat.



$$I_1(\theta) < I_2(\theta)$$

$$\Rightarrow I_1(\theta)^{-1} > I_2(\theta)^{-1}$$

$$LM_1 > LM_2$$

\Rightarrow Being Farther from the MAX θ .