Multivariate Regression: Estimating & Reporting Certainty, Hypotheses Tests

I. The C(N)LRM:

A. Core Assumptions:

1. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
3. $V(\boldsymbol{\varepsilon}) = \sigma_{\varepsilon}^{2}\mathbf{I}_{n}$
4. $E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$
5. $\mathbf{X}$ of full-column rank.

B. Convenience Assumptions (Unnecessary):

1. $\mathbf{X}$ non-stochastic. (Unnec. to unbiasedness; relaxed to "conditioning on $\mathbf{X}$" for variance estimation.)
2. $\boldsymbol{\varepsilon} \sim MVN\left(\mathbf{0}, \sigma_{\varepsilon}^{2}\mathbf{I}_{n}\right)$. (& if not: CLT & asymptopia!)

II. Properties of OLS Estimator under CLRM:

$$\hat{\boldsymbol{\beta}}_{LS} \equiv \mathbf{b}_{LS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \equiv \mathbf{A}\mathbf{y}$$

A.
$$= \mathbf{A}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$$

B.

$$E\left(\hat{\boldsymbol{\beta}}_{LS}\right) = E\left(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + E(\mathbf{A}\boldsymbol{\varepsilon})$$

$$= \boldsymbol{\beta} + \mathbf{A}E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}, \text{ if } \mathbf{X} \text{ non-stoch.}$$

$$= \boldsymbol{\beta} + E(\mathbf{A}\boldsymbol{\varepsilon}) = \boldsymbol{\beta}, \text{ if } \mathbf{X} \text{ stoch. \& } E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

C.

$$V\left(\hat{\boldsymbol{\beta}}_{LS}\right) = V\left(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}\right) = \mathbf{A}V(\boldsymbol{\varepsilon})\mathbf{A}'$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_{\varepsilon}^{2}\mathbf{I}_{n}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \begin{cases} \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}, \text{ if } \mathbf{X} \text{ non-stoch.} \\ \sigma_{\varepsilon}^{2}E\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}, \text{ if } \mathbf{X} \text{ stoch.} \end{cases}$$

$$\Rightarrow V\left(b_k\right) = f\left(\overset{+}{\sigma}_{\varepsilon}^{2}, \overset{-}{V}(\mathbf{X}), \overset{+}{R}^{2}_{x_k, x_{\sim k}}\right)$$

D. NOTES:

1. $$\hat{V}\left(\hat{b}_k\right) = f\left(\overset{+}{\hat{\sigma}}_{\varepsilon}^{2}, \overset{-}{V}(\mathbf{X}), \overset{+}{R}^{2}_{x_k, x_{\sim k}}\right)$$

$V\left(\hat{\boldsymbol{\beta}}_{LS}\right) = \sigma_{\varepsilon}^{2}(\mathbf{X}'\mathbf{X})^{-1}$ is the Cramer-Rao lower-bound;

2. i.e., $\hat{\boldsymbol{\beta}}_{LS}$ is efficient; i.e., $\hat{\boldsymbol{\beta}}_{LS}$ is BLUE.

In fact, if $\boldsymbol{\varepsilon} \sim MVN$, then $\hat{\boldsymbol{\beta}}_{LS}$ is BUE.

E. If $\boldsymbol{\varepsilon} \sim MVN\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n\right)$, then:

$\mathbf{b}_{LS} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} =$ constant + linear-sum of normals

$$\Rightarrow \mathbf{b}_{LS} \sim MVN\left(\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}\right);$$

if not, then

$\mathbf{b}_{LS} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} =$ constant + linear-sum of **?**

$$\Rightarrow \mathbf{b}_{LS} \sim^{A} MVN\left(\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}\right).$$

III. Estimating σ_ε^2 using $\frac{1}{n-k}\mathbf{e}'\mathbf{e}$:

$$\mathbf{e} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{0}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon}$$

$$\Rightarrow E(\mathbf{e}'\mathbf{e}) = E\left\{(\mathbf{M}\boldsymbol{\varepsilon})'\mathbf{M}\boldsymbol{\varepsilon}\right\} = E\{\boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\varepsilon}\} = E\{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}\}$$

$$= E\left\{trace(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})\right\} = E\left\{trace(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right\}$$

$$= trace\left\{\mathbf{M}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right\} = trace\left\{\mathbf{M}\sigma_\varepsilon^2\mathbf{I}_n\right\}$$

$$= \sigma_\varepsilon^2 \times trace\left\{\mathbf{I}_n - \mathbf{N}\right\} = \sigma_\varepsilon^2 \times trace\left\{\mathbf{I}_n\right\} - trace\left\{\mathbf{N}\right\}$$

$$= \sigma_\varepsilon^2 \times \{n - trace\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right\}$$

$$= \sigma_\varepsilon^2 \times \{n - trace\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right\} = \sigma_\varepsilon^2\left[n - trace(\mathbf{I}_k)\right]$$

$$= \sigma_\varepsilon^2 (n-k)$$

$$\Rightarrow E\left(\frac{\mathbf{e}'\mathbf{e}}{n-k}\right) = \sigma_\varepsilon^2 \Rightarrow \text{use } s_e^2 \equiv \frac{\mathbf{e}'\mathbf{e}}{n-k} \text{ as LS (unbiased) est. } \sigma_\varepsilon^2.$$

A. NOTE: $\frac{\mathbf{e}'\mathbf{e}}{n-k}$ is sum of squared (asympt.) normals, divided by degrees freedom, so s_e^2 is (asympt.) $\frac{\chi^2_{n-k}}{n-k}$.

B. Therefore, std Wald *t*-tests & conf. ints. by:

1. $$T = \frac{b_j - c_0}{s.e.(b_j)} = \frac{b_j - c_0}{\sqrt{s_e^2 \left\{(X'X)^{-1}\right\}_{jj}}} \sim^{(A)} t_{n-k}$$

2. $$b_j \pm t^{\alpha}_{n-k} \times s.e.(b_j) = b_j \pm t^{\alpha}_{n-k} \times \sqrt{s_e^2 \left\{(X'X)^{-1}\right\}_{jj}}$$
$$\Rightarrow \{1-\alpha\}\% \text{ (asympt.) conf. int.}$$

3. Tests of linear restrictions (by Wald strategy):

   a) For instance, H_o: $\beta_1+\beta_2=1$. If null-hypoth. true, then estimate not far (in std. err. units) from null:

   b) $$T = \frac{(b_1+b_2)-1}{s.e.(b_1+b_2)} = \frac{(b_1+b_2)-1}{\sqrt{\hat{V}\left(\hat{b}_1+\hat{b}_2\right)}} = \frac{(b_1+b_2)-1}{\sqrt{\hat{V}\left(\hat{b}_1\right)+\hat{V}\left(\hat{b}_2\right)+2\times\hat{C}\left(\hat{b}_1,\hat{b}_2\right)}} \sim^{A} t_{n-k}$$

   But note how we could use matrix algebra to generalize:

$H_o: \beta_1 = 0, \beta_2 = 1 \Rightarrow \mathbf{r}'\boldsymbol{\beta} = q$; for instance, if $\boldsymbol{\beta}$ is (4×1), then:

$$\mathbf{r}' = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\Rightarrow H_o: \mathbf{r}'\boldsymbol{\beta} = q; \; e.g., \; H_o: \mathbf{r}'\boldsymbol{\beta} = 1 \Rightarrow T = \frac{\mathbf{r}'\mathbf{b} - 1}{\sqrt{\hat{V}(\mathbf{r}'\mathbf{b})}} = \frac{\mathbf{r}'\mathbf{b} - 1}{\sqrt{\mathbf{r}'\hat{V}(\mathbf{b})\mathbf{r}}} \sim^A t_{n-k}$$

Notice how $\mathbf{r}' \ldots \mathbf{r}$ plucks $\hat{V}(\mathbf{b}_1), \hat{V}(\mathbf{b}_2), \hat{C}(\mathbf{b}_1, \mathbf{b}_2), \hat{C}(\mathbf{b}_1, \mathbf{b}_2)$, just as in scalar!

4. Test of joint hypotheses (by Wald strategy):

$H_o: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$;

for instance, $H_o: \beta_1 = 0, \beta_2 = 1$ and $\boldsymbol{\beta}$ is (4×1), then:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \text{ and } \mathbf{q} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\Rightarrow H_o: \mathbf{R}\boldsymbol{\beta} = \mathbf{q} \Rightarrow C = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\hat{V}(\mathbf{R}\mathbf{b} - \mathbf{q}) \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q})$$

is a ratio of chi-squares. "Denominator" has n-k deg. free, and has n-k in its denom. "Numerator" is square 2×1 normals.

So, C/2, or, more generally, C/J where J=rows($\mathbf{R}$) is $F_{J,n-k}$!

$$F = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\hat{V}(\mathbf{R}\mathbf{b} - \mathbf{q}) \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \Big/ J \sim^A F_{J,n-k}$$

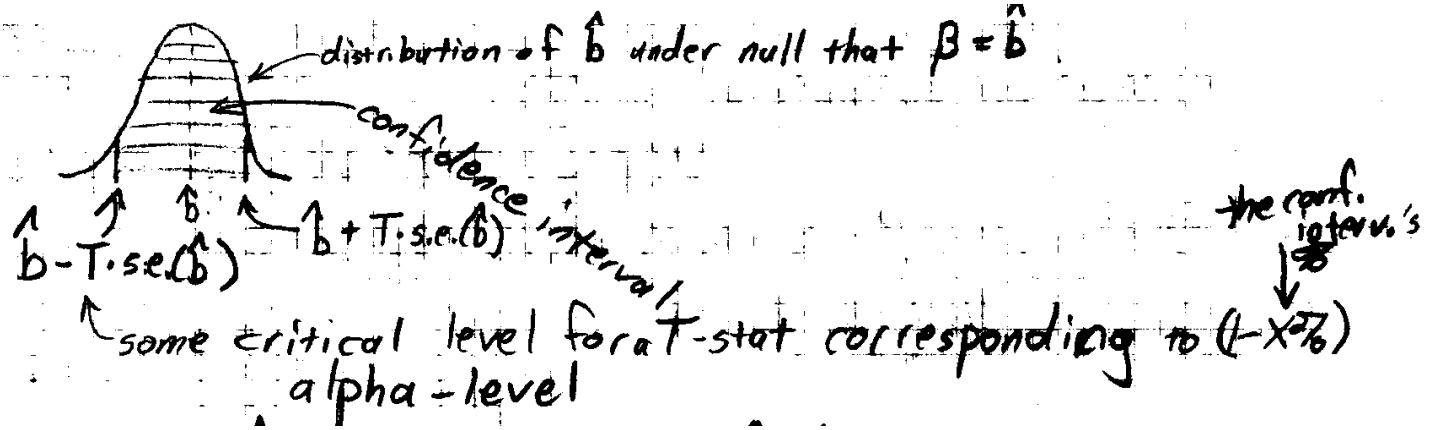
$$\text{and } J \times F = (\mathbf{R}\mathbf{b} - \mathbf{q})' \left[\hat{V}(\mathbf{R}\mathbf{b} - \mathbf{q}) \right]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{q}) \sim^A \chi^2_{n-k}$$

C. Confidence Intervals & Confidence Regions

1. Recall the simple formula for confidence intervals:

$(1-\alpha)$% confidence interval (*c.i.*) for β:

$$\hat{\beta} \pm T \times s.e.(\hat{\beta})$$



2. Recall/Note, too, 1-for-1 correspondence b/w hypothesis test @ level α and the (1-α)% *c.i.*:

   a) *c.i.* overlaps 0⇔fail-to-reject; 0 lies outside c.i.⇔reject

   b) The (1-α)% *c.i.* also corresponds to set of hypothesized β that one would fail-to-reject at level α given estimated b.

3. By latter understanding, can see how might construct a (1-α)% ***confidence region*** for (β_1, β_2) as set of (β_1, β_2) that would fail-to-reject at α given estimates (b_1, b_2):

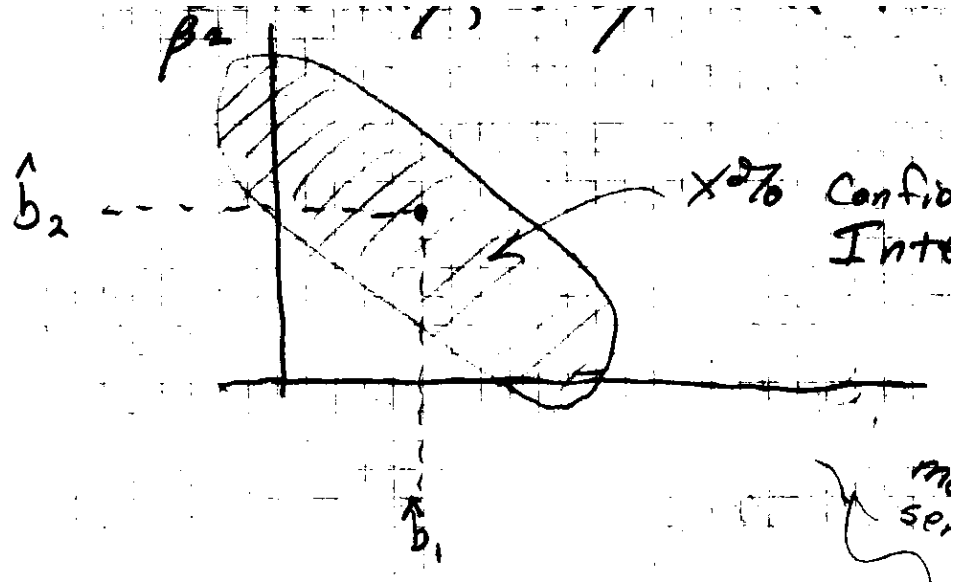
$$\frac{1}{J}\begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix}' \begin{bmatrix} \hat{V}(b_1) & \hat{C}(b_1, b_2) \\ \hat{C}(b_1, b_2) & \hat{V}(b_2) \end{bmatrix}^{-1} \begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix} \leq F_{J, n-k}$$

a) Using $F_{J,n-k}$ critical value for α from desired (1-α)% *c.i.*,

b) J is the number of estimates at issue (2 here) & so J is also the dimensionality of the resulting confidence region.

c) Multiply & solve for (β_1, β_2) that just-satisfy inequality.

4. They generally *look* like this:



5. Are *properties* of conf. reg. intuitive to you?

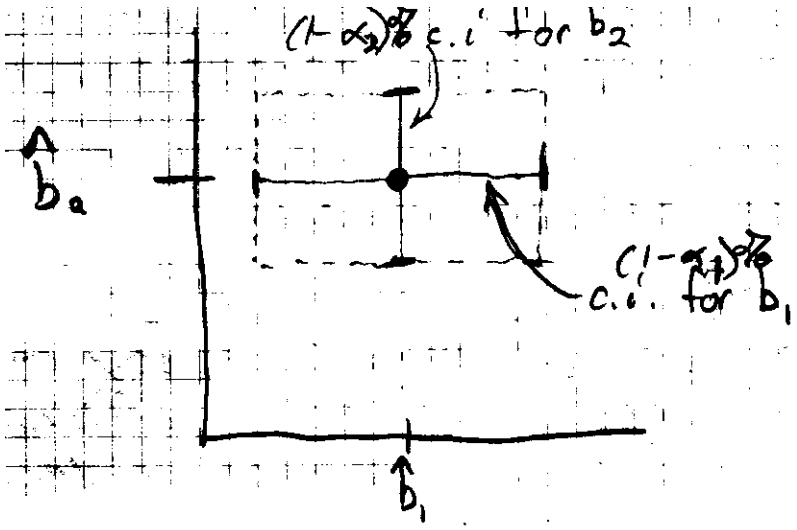
a) Ellipsoidal and centered on (b_1,b_2).

b) Go top-left to bottom-right if C(b_1,b_2)<0, and this will be when (partial) C(x_1,x_2)>0; go from bottom-right to top-left if C(b_1,b_2)>0, and this will be when (partial) C(x_1,x_2)<0.

c) Appear thinner as |C(b_1,b_2)| &/or |V(b_1)-V(b_2)| 'greater'.

d) Circular if C(b_1,b_2)=0 and V(b_1)=V(b_2).

6. *Short-cut Approximation*: Rectangular region given by k univariate (1-α_k)% *c.i.*'s contains at least $(1-\Sigma\alpha_k)\%$:



a) Ex: two 95% *c.i.*'s ⇒ region w/ min. (1-.05-.05)%=90%

b) Worse approx. (area too big) the more 'slanted' and thinner the actual confidence region. Never "right" area (ultimately, b/c rectangular not ellipsoidal).

D. Measures of Fit ("Goodness of Fit" Statistics)

1. Std. Err. Est./Reg. (S.E.E., S.E.R., s.e.e., s.e.r.):

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n-k}} \qquad \text{note: If } \boldsymbol{\varepsilon} \sim^{A} \text{MVN, then } s_e \sim^{A} \sqrt{\frac{\chi^2_{n-k}}{n-k}}$$

a) Sometimes also denoted σ or $\hat{\sigma}$, w/ or w/o sub e or ε, but best to reserve $\hat{\sigma}_e$ for ML est. and to use the hat & e not ε:

$$\hat{\sigma}_e = \sqrt{\sigma_e^2} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n}}$$

b) Notes:

(1) *Kinda* measure of typical or avg error or mistake. (Act'ly, measures square root of average squared mistake…)

(2) In same units as dep var. E.g., if dep var in \$, s.e.e. in \$.

(3) Not construct models to min s_e any more than to max R^2.

2. R^2: share of the variation in **y** 'explained' (linearly accounted) by the model ($\mathbf{X}\boldsymbol{\beta}$).

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\sum[(y-e)-\bar{y}]^2}{\sum(y-\bar{y})^2} = \frac{\sum[(y-\bar{y})-e]^2}{\sum(y-\bar{y})^2}$$

$$= \frac{\sum(y-\bar{y})^2 - \sum(2(y-\bar{y})e) + \sum e^2}{\sum(y-\bar{y})^2}$$

$$= 1 - \frac{2\sum(\hat{y}e + e^2 - \bar{y}e) - \sum e^2}{\sum(y-\bar{y})^2}$$

$$= 1 - \frac{2\sum e^2 - \sum e^2}{\sum(y-\bar{y})^2} = 1 - \frac{\sum e^2}{\sum(y-\bar{y})^2} = 1 - \frac{SSE}{SST}$$

a) R^2 is also the square of the correlation of **y** & $\hat{\mathbf{y}}$, i.e., $r^2_{\mathbf{y},\hat{\mathbf{y}}}$.

b) If $\boldsymbol{\varepsilon} \sim^A MVN$, then $R^2 \sim^A \frac{\chi^2}{\chi^2} = F$.

3. Adjusted R^2, Adj. R^2, R-bar squared:

a) Can always increase R^2 just by adding variables. Want some penalty for lack parsimony. Common adj. to R^2 is to replace numerator & denominator w/ unbiased estimates.

$$\bar{R}^2 = 1 - \frac{unbiased(SSE)}{unbiased(SST)} = 1 - \frac{\sum e^2 / (n-k)}{\sum (y-\bar{y})^2 / n-1} = 1 - \frac{s_e^2}{s_y^2}$$

b) Weak penalty. Can show that adding variable w/ coeff. having t>1 increases $\bar{R}^2$. Alternative adj.'s with stronger penalties, based on "Information Criterion": A$^{\text{kaike}}$IC, B$^{\text{ayesian}}$/S$^{\text{chwartz}}$IC, … http://en.wikipedia.org/wiki/Akaike%27s_information_criterion .

c) Although not directly used (that I'm aware), notice that:

$$\text{If } \boldsymbol{\varepsilon} \sim^{A} MVN, \text{ then } \bar{R}^2 \sim^{A} \frac{\chi^2_{n-k} / (n-k)}{\chi^2_{n-1} / (n-1)} = F_{n-k,n-1}$$

4. (Log) Likelihood from ML est. is also measure of fit.

5. Use & Abuse of Fit Statistics/Measures:

a) Can use to compare model performance *in same sample*; use to compare across samples only w/ great attention and care how much V(y) to explain varies across samples.

b) At end, (relative) fit of model more something to estimate given a model, than something to model to maximize.

E. "Degradation of Fit" Strategy for Testing:

1. *Logic*: If null hypothesis were true, then imposing it as true rather than estimating its parameters should result in "little" loss of fit.

2. *Strategy:* Measure fit-loss, then determine how that measure or some function of it would be distributed

under the null hypothesis, so we can determine how likely this much fit-loss is to have occurred by chance.

3. The "change-in-R^2" or "delta-R^2" or "ΔR^2" Test:

a) Determine how to "impose the null hypothesis". Example:

$$\left.\begin{array}{c} H_o: \beta_3 = \beta_4 = 0 \\ H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \end{array}\right\} \Rightarrow \begin{cases} H_o: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ H_1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \end{cases}$$

b) Measure loss of explanatory power relative to gap from big-model explanatory power to one, and divide each numerator and denominator by its degrees of freedom:

$$\text{loss of fit: } \Delta R^2 = R_1^2 - R_0^2 = \left(1 - \frac{SSE^1_{n-k_1}}{SST_{n-1}}\right) - \left(1 - \frac{SSE^0_{n-k_0}}{SST_{n-1}}\right) = \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SST_{n-1}}$$

$$\text{fit-gap: } 1 - R_1^2 = \frac{SSE^1_{n-k_1}}{SST_{n-1}} \quad \Rightarrow \quad \text{ratio: } \frac{SSE^0_{n-k_0} - SSE^1_{n-k_1}}{SSE^1_{n-k_1}}$$

$$\Rightarrow \frac{\chi^2_{n-k_0} - \chi^2_{n-k_1}}{\chi^2_{n-k_1}} \Rightarrow \ ^\circ\text{free: } \frac{(n-k_0)-(n-k_1)}{n-k_1} = \frac{k_1 - k_0}{n-k_1} = \frac{\Delta k}{n-k_1}$$

$$\text{So: } F = \frac{\Delta R^2 / \Delta k}{(1-R_1^2)/(n-k_1)} \overset{(A)}{\sim} F_{\Delta k, n-k_1}$$

4. Tests using other measures of fit, s_e^2 or *ln(L)*, also…

5. Third logic, *Lagrange-Multiplier Tests*: if null hypoth true, then impose it as constraint on max *ln(L)* or min *SSE* should not bind, implying: Lagrange multipliers, $\lambda = \mathbf{0}$, and $\partial \ln(L)/\partial \boldsymbol{\beta}_{\text{at } \boldsymbol{\beta} \text{ null}} = \mathbf{0}$ or $\partial SSE/\partial \boldsymbol{\beta}_{\text{at } \boldsymbol{\beta} \text{ null}} = \mathbf{0}$.