Multivariate Regression: Estimating & Reporting Certainty, Hypotheses Tests

I. The C(N)LRM:
   A. Core Assumptions:
      1. \( y = X\beta + \epsilon \)
      2. \( E(\epsilon) = 0 \)
      3. \( V(\epsilon) = \sigma^2 \epsilon I_n \)
      4. \( E(\epsilon | X) = 0 \)
      5. \( X \) of full-column rank.
   B. Convenience Assumptions (Unnecessary):
      1. \( X \) non-stochastic. (Unnec. to unbiasedness; relaxed to “conditioning on \( X \)” for variance estimation.)
      2. \( \epsilon \sim MVN\left(0, \sigma^2 \epsilon I_n\right) \). (& if not: CLT & asymptopia!)

II. Properties of OLS Estimator under CLRM:

\[
\hat{\beta}_{LS} \equiv b_{LS} = (X'X)^{-1} X'y \equiv Ay
\]

A.

\[= A \left( X\beta + \epsilon \right) = \beta + A\epsilon \]
\[
E(\hat{\beta}_{LS}) = E(\beta + A\varepsilon) = \beta + E(A\varepsilon)
\]
\[
= \beta + AE(\varepsilon) = \beta, \text{ if } X \text{ non-stoch.}
\]
B.
\[
= \beta + E(A\varepsilon) = \beta, \text{ if } X \text{ stoch.} \& E(\varepsilon \mid X) = 0
\]

\[
V(\hat{\beta}_{LS}) = V(\beta + A\varepsilon) = AV(\varepsilon)A'
\]
\[
= (X'X)^{-1}X'V(\varepsilon)X(X'X)^{-1}
\]
\[
= (X'X)^{-1}X'\sigma_\varepsilon^2 I_n X(X'X)^{-1}
\]
\[
= \sigma_\varepsilon^2 (X'X)^{-1} X'X(X'X)^{-1}
\]
\[
= \begin{cases} 
\sigma_\varepsilon^2 (X'X)^{-1}, & \text{if } X \text{ non-stoch.} \\
\sigma_\varepsilon^2 E\{(X'X)^{-1}\}, & \text{if } X \text{ stoch.}
\end{cases}
\]

C.
\[
\Rightarrow V(b_k) = f\left(\sigma_\varepsilon^2, V(X), R^2_{x_k,x_{\sim k}}\right)
\]

D. NOTES:

\[
\hat{V}(\hat{b}_k) = f\left(\hat{\sigma}_\varepsilon^2, V(X), R^2_{x_k,x_{\sim k}}\right)
\]

\[
V(\hat{\beta}_{LS}) = \sigma_\varepsilon^2 (X'X)^{-1} \text{ is the Cramer-Rao lower-bound;}
\]
\[
\text{i.e., } \hat{\beta}_{LS} \text{ is efficient; i.e., } \hat{\beta}_{LS} \text{ is BLUE.}
\]

2. In fact, if \( \varepsilon \sim MVN \), then \( \hat{\beta}_{LS} \) is BUE.
E. If \( \boldsymbol{e} \sim MVN\left( \mathbf{0}, \sigma^2_\varepsilon \mathbf{I}_n \right) \), then:

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

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

if not, then

\[
\mathbf{b}_{LS} = \boldsymbol{\beta} + A\varepsilon = \text{constant} + \text{linear-sum of} \ ?
\]

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

III. Estimating \( \sigma^2_\varepsilon \) using \( \frac{1}{n-k}\mathbf{e}'\mathbf{e} \):

\[
\mathbf{e} = \mathbf{M}(X\boldsymbol{\beta} + \varepsilon) = \mathbf{MX}\boldsymbol{\beta} + \mathbf{M}\varepsilon = 0\boldsymbol{\beta} + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon
\]

\[
\Rightarrow E(\mathbf{e}'\mathbf{e}) = E\left\{ (\mathbf{M}\varepsilon)'\mathbf{M}\varepsilon \right\} = E\left\{ \varepsilon'M'M\varepsilon \right\} = E\left\{ \varepsilon'M\varepsilon \right\}
\]

\[
= E\left\{ \text{trace}(\varepsilon'M\varepsilon) \right\} = E\left\{ \text{trace}(\mathbf{M}\varepsilon\varepsilon') \right\}
\]

\[
= \text{trace}\left\{ \mathbf{M}\varepsilon\varepsilon' \right\} = \text{trace}\left\{ \mathbf{M}\sigma^2_\varepsilon \mathbf{I}_n \right\}
\]

\[
= \sigma^2_\varepsilon \times \text{trace}\left\{ \mathbf{I}_n - \mathbf{N} \right\} = \sigma^2_\varepsilon \times \text{trace}\left\{ \mathbf{I}_n \right\} - \text{trace}\left\{ \mathbf{N} \right\}
\]

\[
= \sigma^2_\varepsilon \times \{n - \text{trace}\left\{ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right\} \}
\]

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

\[
= \sigma^2_\varepsilon (n - k)
\]

\[
\Rightarrow E\left( \frac{\mathbf{e}'\mathbf{e}}{n-k} \right) = \sigma^2_\varepsilon \Rightarrow \text{use } s^2_\varepsilon \equiv \frac{\mathbf{e}'\mathbf{e}}{n-k} \text{ as LS (unbiased) est. } \sigma^2_\varepsilon.
\]
A. NOTE: \( \frac{e'e}{n-k} \) 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:

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

1. \( b_j \pm t_{n-k}^\alpha \times s.e.(b_j) = b_j \pm t_{n-k}^\alpha \times \sqrt{s_e^2 \{(X'X)^{-1}\}_{jj}} \)

2. \( \Rightarrow \{1 - \alpha\} \% \) (asympt.) conf. int.

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

   a) For instance, \( H_0: \beta_1 + \beta_2 = 1 \). If null-hypoth. true, then estimate not far (in std. err. units) from null:

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

   But note how we could use matrix algebra to generalize:
\[ H_0: \beta_1 = 0, \beta_2 = 1 \Rightarrow \mathbf{r}' \beta = q; \] for instance, if \( \beta \) is \((4 \times 1)\), then:

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

\[ \Rightarrow H_0: \mathbf{r}' \beta = q; \text{ e.g., } H_0: \mathbf{r}' \beta = 1 \Rightarrow T = \frac{\mathbf{r}' \mathbf{b} - 1}{\sqrt{\mathbf{V}(\mathbf{r}' \mathbf{b})}} = \frac{\mathbf{r}' \mathbf{b} - 1}{\sqrt{\mathbf{r}' \hat{\mathbf{V}}(\mathbf{b}) \mathbf{r}}} \sim A t_{n-k} \]

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

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

\[ H_0: \beta_1 = q_1, \beta_2 = q_2 \Rightarrow \mathbf{R} \beta = q; \]
for instance, \( H_0: \beta_1 = 0, \beta_2 = 1 \) and \( \beta \) is \((4 \times 1)\), then:

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

\[ \Rightarrow H_0: \mathbf{R} \beta = q \Rightarrow C = (\mathbf{R} \mathbf{b} - q)' \left[ \hat{\mathbf{V}}(\mathbf{R} \mathbf{b} - q) \right]^{-1} (\mathbf{R} \mathbf{b} - q) \]

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

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

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

and \( J \times F = (\mathbf{R} \mathbf{b} - q)' \left[ \hat{\mathbf{V}}(\mathbf{R} \mathbf{b} - q) \right]^{-1} (\mathbf{R} \mathbf{b} - q) \sim A \chi^2_{n-k} \)
C. Confidence Intervals & Confidence Regions

1. Recall the simple formula for confidence intervals:

\[(1 - \alpha)\% \text{ confidence interval (c.i.) for } \beta:\]

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

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

a) c.i. overlaps 0 \(\iff\) fail-to-reject; 0 lies outside c.i. \(\iff\) reject

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

3. By latter understanding, can see how might construct a \((1 - \alpha)\%\ confidence\ region\) for \((\beta_1, \beta_2)\) as set of \((\beta_1, \beta_2)\) that would fail-to-reject at \(\alpha\) given estimates \((b_1, b_2)\):

\[
\frac{1}{J} \left[ b_1 - \beta_1 \right]' \left[ \begin{array}{cc}
\hat{V}(b_1) & \hat{C}(b_1, b_2) \\
\hat{C}(b_1, b_2) & \hat{V}(b_2)
\end{array} \right]^{-1} \left[ \begin{array}{c}
b_1 - \beta_1 \\
b_2 - \beta_2
\end{array} \right] \leq F_{J, n-k}
\]
a) Using $F_{J,n-k}$ critical value for $\alpha$ from desired $(1-\alpha)\%$ 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 $(\beta_1, \beta_2)$ that just-satisfy inequality.

4. They generally look like this:

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{confidence_region}
\caption{Confidence Region}
\end{figure}

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 $(1-\alpha_k)\%$ 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.):

\[
s_e = \sqrt{s_{e}^{2}} = \sqrt{\frac{e' e}{n-k}} \quad \text{note: If } e \sim^{A} \text{MVN, then } s_e \sim^{A} \sqrt{\frac{\chi_{n-k}^2}{n-k}}
\]

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

\[
\hat{\sigma}_e = \sqrt{\sigma_{e}^{2}} = \sqrt{\frac{e' e}{n}}
\]

b) Notes:

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

Page 8 of 11
(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 $(X\beta)$.

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

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

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

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

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

b) If $\varepsilon \sim A MVN$, 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.
\[ \bar{R}^2 = 1 - \frac{\text{unbiased}(SSE)}{\text{unbiased}(SST)} = 1 - \frac{\sum e^2 / (n-k)}{\sum (y - \bar{y})^2 / n-1} = 1 - \frac{s_e^2}{s_y^2} \]

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”: \( \text{A}^{\text{kaikeIC}} \), \( \text{B}^{\text{ayesian/SchwartzIC}} \), … http://en.wikipedia.org/wiki/Akaike%27s_information_criterion.

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

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

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 “$\Delta R^2$” Test:

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

   \[ H_0 : \beta_3 = \beta_4 = 0 \quad \Rightarrow \quad \begin{cases} H_0 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \end{cases} \]

   \[ \Rightarrow \begin{cases} H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \end{cases} \]

   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:

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

   fit-gap: $1 - R_1^2 = \frac{SSE_{n-k_i}^1}{SST_{n-1}} \quad \Rightarrow \quad$ ratio: $\frac{SSE_{n-k_0}^0 - SSE_{n-k_i}^1}{SSE_{n-k_i}^1}$

   \[ \Rightarrow \frac{\chi_{n-k_0}^2 - \chi_{n-k_i}^2}{\chi_{n-k_i}^2} \quad \Rightarrow \quad \text{free:} \quad \frac{(n-k_0)-(n-k_i)}{n-k_i} = \frac{k_1-k_0}{n-k_i} = \frac{\Delta k}{n-k_i} \]

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

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, $\lambda=0$, and $\partial \ln(L)/\partial \beta_{\text{null}}=0$ or $\partial \text{SSE}/\partial \beta_{\text{null}}=0$. 

Page 11 of 11