EC 228: Topics Review for Mid Term Examination #2

Comparison of SLR and MLR Estimation and Inference

1. Estimation	**SLR**	**MLR**
Those (S/M)LR assumptions/ conditions	SLR.1: Linear Model $Y = \beta_0 + \beta_1 X + U$	MLR.1: Linear Model $Y = \beta_0 + \sum \beta_j X_j + U$
	SLR2: Random sampling	MLR2: Random sampling
	SLR3: Sampling variation in the independent variable	MLR3: No perfect collinearity
	SLR.4: Zero Conditional Mean $E(U \mid X = x) = 0$ for any x	MLR.4: Zero Conditional Mean $E(U \mid x_1, \ldots, x_n) = 0$ for any $x_1, \ldots$
If these four assumptions hold:	OLS = LUE: OLS estimators are Linear and Unbiased	OLS = LUE: OLS estimators are Linear and Unbiased
SRF (Sample Regression Function)	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	$\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_j x_j$
PRF (Population Regression Function)	$E(Y \mid X = x) = \beta_0 + \beta_1 x$ (conditional means)	$E(Y \mid x_1, \ldots, x_k) = \beta_0 + \sum \beta_j x_j$ (conditional means)
Add one more assumption	SLR.5: Homoskedastic errors (conditional on x)	MLR.5: Homoskedastic errors (conditional on the x_j's)
If all 5 assumptions/ conditions hold I:	$Var(B_1 \mid x's) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$ $= \frac{\sigma^2}{(n-1)S_{xx}}$	$Var(B_x \mid x, z...) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} VIF_x$ $= \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$ … where …
Variance Inflation Factors (VIF)		$VIF_x = \frac{1}{1 - R_x^2}$; $R_x^2 = 1 - \frac{1}{VIF_x}$

MT #2 Topics Review

Mean Squared Error (MSE)	$MSE = \frac{SSR}{n-2} = \frac{SSR}{dof}$	$MSE = \frac{SSR}{n-(k+1)} = \frac{SSR}{dof}$
Degrees of freedom (*dofs*)	n-2	n-k-1
If all 5 assumptions/ conditions hold II:	MSE: unbiased estimator of σ^2	MSE: unbiased estimator of σ^2
If all 5 assumptions/ conditions hold III:	Gauss-Markov Theorem: SLR.1-SLR.5 $\rightarrow$ OLS=BLUE	Gauss-Markov Theorem: MLR.1-MLR.5 $\rightarrow$ OLS=BLUE
2. Inference	**SLR**	**MLR**
Need one last assumption/condition:	SLR.6: $U_i \sim N(0,\sigma^2)$ and indept. of X	MLR.6: $U_i \sim N(0,\sigma^2)$ and indept. of the RHS variables
If all 6 assumptions/ conditions hold I:	$B_1 \sim N\left(\beta_1, \text{var}(B_1)\right)$, where $\text{var}(B_1) = \frac{\sigma^2}{(n-1)S_{xx}}$	$B_x \sim N\left(\beta_x, \text{var}(B_x)\right)$, where $\text{var}(B_x) = \frac{\sigma^2}{(n-1)S_{xx}} VIF_x$
What did the 6th condition get us?	Normal Distribution of the estimator	Normal Distribution of the estimator
Standard Deviation of B_1	$sd(B_1) = \frac{\sigma}{S_x\sqrt{n-1}}$	$sd(B_1) = \frac{\sigma}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
Standard Error of B_1 (est. of $sd\left(B_1\right)$)	$se(B_1) = \frac{RMSE}{S_x\sqrt{n-1}}$	$se(B_x) = \frac{RMSE}{S_x\sqrt{n-1}}\sqrt{VIF_x}$
If all 6 assumptions/ conditions hold II:	*t statistic:* $\frac{B_1-\beta_1}{se(B_1)} \sim t_{n-2}$	*t statistic:* $\frac{B_x-\beta_x}{se(B_x)} \sim t_{n-k-1}$

Inference follows, once you have the distribution of the t statistic…. which you have, given assumptions/conditions SLR/MLR 1-6.

MT #2 Topics Review

2. Inference cont'd	**SLR**	**MLR**
Confidence Intervals		
Confidence level: C	*C* is confidence level	*C* is confidence level
Confidence Intervals	$\left[B_1 \pm c\ se(B_1)\right],\ P[\mid t_{n-2} \mid < c) = C$	$\left[B_x \pm c\ se(B_x)\right],\ P[\mid t_{n-k-1} \mid < c) = C$
Hypothesis Testing		
Null Hypothesis (H_0)	$H_0 : \beta_1 = 0$	$H_0 : \beta_x = 0$
t stat (under H_0)	$tstat = \frac{B_1}{se(B_1)} \sim t_{n-2}$	$tstat = \frac{B_x}{se(B_x)} \sim t_{n-k-1}$
p value (given data set and OLS estimates)	$p = P[\lvert t_{n-2} \rvert > \lvert tstat_{\hat{\beta}_1} \rvert)$ (two-tailed probability)	$p = P[\lvert t_{n-k-1} \rvert > \lvert tstat_{\hat{\beta}_x} \rvert)$ (two-tailed probability)
Significance Level	α	α
Critical value: c	$P[\lvert t_{n-2} \rvert > c) = P[\lvert t_{dof} \rvert > c) = \alpha$ (two-tailed probability)	$P[\lvert t_{n-2} \rvert > c) = P[\lvert t_{dof} \rvert > c) = \alpha$ (two-tailed probability)
Hypothesis Test	Reject H_0 if $\lvert t_{\hat{\beta}_1} \rvert > c$ or if $p < \alpha$	Reject H_0 if $\lvert t_{\hat{\beta}_j} \rvert > c$ or if $p < \alpha$
3. Convergence	**SLR**	**MLR**
t stats and …		
… R-sq	$t^2_{\hat{\beta}_1} = (n-2)\frac{R^2}{1-R^2}$	$t^2_{\hat{\beta}_x} = dofs\frac{\Delta R^2_x}{1-R^2}$
… SSE/SSRs	$t^2_{\hat{\beta}_1} = (n-2)\frac{SSE}{SSR}$	$t^2_{\hat{\beta}_x} = dofs\frac{\Delta SSE_x}{SSR}$
F stat (for the regression)	$Fstat = dof\ \frac{R^2}{1-R^2} = dof\ \frac{SSE}{SSR}$	$Fstat = \frac{dofs}{k}\frac{R^2}{1-R^2} = \frac{dofs}{k}\frac{SSE}{SSR}$

Further Discussion

Those SLR's and MLR's (also, see above):

1) Given SLR/MLR.1-4, OLS estimators are Linear Unbiased Estimators (LUEs), conditional on the x's.
   a) But OLS estimators are not alone; there are an infinite number of LUEs (e.g. any weighted average of *slopes to the sample means*).
2) Add in SLR/MLR 5 (*homoskedasticity*), and without making any distributional assumptions (e.g. no assumption of Normality)
   a) MSEs provide unbiased estimators of the conditional variance of the U's
   b) … and we can derive expressions for the variances, standard deviations and standard errors of the OLS estimators (*VIF*s factor in formulas with MLR models)
   c) Standard errors for slope coefficients
      i) SLR: increasing with RMSE and decreasing with the variance of the x, for n fixed
      ii) MLR: increasing with RMSE and VIF_x, and decreasing with the variance of x, for n fixed
   d) Gauss-Markov Theorem: *OLS* ≡ *BLUE* (minimum variance in the class of LUEs)
   e) And to repeat: We have all this with SLR/MLR 1-5, and no distributional assumptions.
3) To do *Inference* we need to make distributional assumptions:

   SLR/MLR 6: The U's are Normally distributed (with mean zero and constant variance) and independent of the x's…

   a) Given SLR/MLR 1-6, the OLS slope estimator(s) will be Normally distributed, and the associated *t statistic* will have a t distribution with n-k-1 dofs
   b) Use the t distribution with n-k-1 dofs to generate confidence intervals and do hypothesis testing (everything driven by the t statistic having a t distribution with n-k-1 dofs)

Convergence: Goodness-of-Fit v. Inference

4) Convergence of the two approaches to assessment: Goodness-of-fit (how well predicteds fit actuals) and Inference (precision of estimation of unknown parameter values)
5) SLR Models: for given n, the magnitude of the t stat is directly related to the R-sq (and SSEs and SSRs) of the regression
6) MLR Models: for given n, the magnitude of the t stat for a RHS variable is directly related to the **incremental** contribution of that variable to R-sq (and SSEs and SSRs)
   a) When dropping RHS variables, adj R-sq will increase or decrease dependin on magnitude of t stat (dropping RHS var: $|t| < 1$: adj R-sq increases; $|t| > 1$: adj R-sq decreases)
   b) relative magnitudes of t stats capture relative incremental contributions to SSEs, SSRs, R-sq and adj R-sq

MT #2 Topics Review

Heteroskedasticity

7) SLR/MLR 5 is violated. OLS is still unbiased given SLR/MLR1-4; but the usual reported OLS standard errors are not usable… and so the typically reported t stats, p-values and confidence intervals are also not correct

8) To get to BLUE: run weighted least squares if possible (weights inversely related to variances)

9) But in any event run generate robust standard error using **, robust** (no harm if have homoskedasticity) to get heteroskedasticity corrected standard errors

   a) Estimated coefficients are unchanged under ,robust (OLS still a LUE is conditions 1-4 hold)

   b) standard errors, t stats and p values can increase or decrease relative to regression results absent the correction for heteroskedasticity

Dummies on the RHS

10) On the RHS

   a) allow for different intercepts (*fixed effects*) and/or slopes for differing populations (*intercept* and *slope* dummies)

   b) useful for estimating impact/bias as well as in quieting the endogeneity critics

   c) dummies pick up averages of unexplained residuals, sort of

   d) importance of the benchmark/reference group (the omitted dummy) in interpreting results

      i) estimated coefficients capture average (unexplained) differences vis-à-vis the benchmark population

      ii) your estimates are only as good/reliable as the rest of your model (what did you leave out of the model?, and how has that biased your estimated coefficient?)

   e) To understand what your SRF is telling you, just look at predicted values under different combinations of 0's and 1's

   f) Fixed Effects with a full complement of dummies: Stata will drop one dummy due to perfect-collinearity, and that population becomes the reference/baseline population for interpreting coefficients

   g) Examples, examples, examples…. and more examples: gender effects; death penalty impacts; bias in sovereign debt ratings; fixed effects and ticket prices

11) On the LHS:

   a) Bivariate dependent variable

   b) SRF = LPM (Linear Probability Model)

      i) strength: can read marginal effects right off the regression output (unlike more sophisticated models)

      ii) weaknesses: may generate predicted probabilities above 1 or below 0

MT #2 Topics Review

F statistics and F tests

12) The F statistic is an elasticity, looking at $\%\Delta SSRs$ v. $\%\Delta dofs$; also defined using R_R^2 and R_{UR}^2; when testing for a single parameter, $Fstat = (tstat)^2$

13) F stats drive convergence

   a) connect the two approaches to *Assessment, Goodness-of-Fit* (SSRs, R-sq's, SSEs) and *Inference* (t stats etc .);

   b) t-stats (squared) capture incremental impacts on R-sq's (can also express in terms of changes in SSRs or SSEs)

14) Relationship between F statistics, t statistics, adj R^2, and MSE

   a) In moving from the unrestricted model to the restricted model, adjusted R^2 decreases iff the F statistic exceeds 1, and increases if F stat < 1 (so $adjR_R^2 < adjR_{UR}^2$ *iff* $F > 1$)

   b) Previously saw this type of result with t stats.

15) F tests:

   a) identical to the t test when looking at a single RHS variable;

   b) with more than one RHS variable, use the F test to test linear restrictions on the parameters of the model

      i) watch out for *baby/bathwater* effect

   c) The reported F stat and associated p-value reported in standard OLS regression output is for the F test of the (joint) null hypothesis that the true parameter values for the RHS variables are all zero

      i) This statistics and test is usually of no interest… as you almost always will reject the Null Hypothesis

      ii) If you cannot reject that null hypothesis, you have a really really bad model!

Supplemental Material

We did not get to the unit on functional forms etc… but if we had:

16) Functional forms: exploring functional relationships with slope and intercept dummies, ln's, exp's, polynomials, percentile (e.g. quintile, decile etc) dummies and *fixed effects*… and maybe include trend effects, categorical dummies (e.g. regional dummies)

17) Three terms not mentioned for more than a nano-second in the course

   a) *Correctly specified model, identification, consistency*

   b) And one technique: *Instrumental Variables (IVs)*