**EC380 Final Exam Study Guide**

**Exam date: Tuesday, May 6th at 8 AM**

Extra office hours:Monday, May 5th 9:30-11:30

***Ordinary Least Squares***

Derive the OLS estimator for the slope and intercept for the model y=Bo+B1x+u

* Using the method of least squares
* Using the method of moments

Define/graphically depict: yi, yhati, B’s, Bhat’s, errors (ui), residuals (uhati), E(y|x)

Explain how OLS achieves *ceteris paribus* estimates of the effect of x on y (If y= Bo+B1x1+B2x2+u, OLS partials out influence of x2 on x1, of x2 on y, and then estimates purified y on purified x1 to get B1).

***Gauss-Markov Theorem***: If all of the Gauss-Markov assumptions hold, OLS provides estimates that are BLUE.

* Define unbiasedness, consistency, efficiency, BLUE
* List each of the Gauss-Markov assumptions.
	+ Data is randomly collected; X values are fixed in repeated sampling and errors/y are random.
	+ Model is linear in the parameters.
	+ n>k
	+ Sufficient variation in X; var(x)>0
	+ E(u|x)=0
	+ cov(x,u)=0
	+ No perfect multicollinearity between x’s
	+ Homoskedasticity: var(ui)=var(u)=constant across values of x
	+ No autocorrelation: cov(ui,uj)=0 for i≠j
* Explain why each one is important, i.e. what happens if the assumption does not hold?
	+ Is estimation possible?
	+ Is OLS still BLUE?
	+ How is estimate of β affected? (unbiased and consistent?)
	+ How is var(βhat) affected? (correct, incorrect, too big, too small?)
* Explain how to figure out (when possible) if one of the assumptions does not hold.
	+ Tests to detect heteroskedasticity
	+ Tests to detect autocorrelation
	+ When to suspect multicollinearity
* Explain how to deal with the problem if an assumption does not hold.
	+ What is the direction of bias (up or down), if any?
	+ Effects on hypothesis testing of omitting a relevant variable or including irrelevant ones.
		- When the X’s are correlated? Uncorrelated?
	+ Correcting for heteroskedasticity (GLS, WLS, White’s correction)
	+ Correcting for autocorrelation (GLS, FGLS (including Cochrane-Orcutt procedure), Newey-West correction)

***Interpreting regression output***

* Interpret the estimated coefficients for
	+ level-level, log-level, log-log models
	+ models with interaction terms or with nonlinear (squared, cubed, etc.) terms
	+ models with intercept and slope dummy variables
* Goodness of fit
	+ R2, adjusted R2
	+ RSS, TSS, ESS
* Hypothesis testing
	+ Significance of individual variable: t-test, p-value, confidence interval
	+ Joint significance: F-test
	+ Linear restrictions (B1=B3, B1=1-B2, etc.): F-test (choose correct F-statistic formula)

**New material since Exam 2**

***Heteroskedasticity***

* Heteroskedasticity
	+ Definition
	+ What are the potential sources of heteroskedasticity?
	+ Pure vs. impure heteroskedasticity
	+ Consequences:
		- OLS still LUE but no longer BLUE
	+ Detection:
		- Graphical method
		- Goldfeld-Quandt Test
		- White's Test
	+ Remedial measures & the outcome:
		- GLS when variance of residuals is known
		- GLS when variance of residuals is not known but can be speculated (WLS)
		- White's correction when variance of residuals unknown and cannot be speculated.
		- Which of these is BLUE?
* **\*** Note: While *pure* heteroskedasticity and *pure* autocorrelation do not bias the OLS estimates of β, *impure* heteroskedasticity or autocorrelation (caused by model misspecification) *can bias* the OLS estimates. Therefore, always consider an omitted variable first.

***Time series data & Autocorrelation***

* Autocorrelation
	+ What is an autoregressive process? AR(1)? AR(p)?
	+ What is stationarity?
	+ Consequences:
		- OLS still LUE but no longer BLUE
	+ Detection:
		- Regress ut on ut-1. Then t-test for significance of ρ.
		- Durbin-Watson test for AR(1)
		- Breusch-Godfrey test for AR(p)
	+ Remedial measures:
		- GLS
		- FGLS/EGLS
			* Estimate ρ from
				+ Regression of ut on ut-1
				+ DW statistic (only for AR(1))
			* Cochrane-Orcutt procedure (converges toward true ρ, better method)
		- Newey-West correction (like White’s correction for heteroskedasticity but also BLUE)
		- If sample is small, might be better to ignore autocorrelation and stick to OLS.
* Time trend
	+ De-trending time series data can avoid spurious correlation problems.

***Endogeneity and Instrumental Variables***

Missing a relevant variable that is correlated with included explanatory variable 🡪 cov(x,u)≠0 🡪 bias!

**Option 1** (OLS): Use a proxy for the missing variable.

* A proxy is highly correlated with the missing/omitted variable. The proxy replaces the missing variable.
* Using a lagged dependent variable (y) as a proxy for missing variables.

**Option 2** (2SLS): Replace the included variable with an instrumental variable (IV), where

cov(x,IV) ≠0 (IV is relevant)

cov(IV,u)=0 (IV is exogenous)

Step 1: Reduced form model - Regress x on IV and all exogenous variables from the structural model.

Step 2: Regress structural model, replacing x with its predicted value (x-hat) from the reduced form regression in step 1.

**Testing for endogeneity** (Do we really need to use 2SLS, which tends to yield larger var(Bhat)🡪small t-stats?)

Conduct the Hausman test to see if BOLS and BIVdiffer a lot. If they do, 2SLS is necessary.

1. Run reduced form regression.
2. Save the residual (the part of x that is correlated with u in original/structural equation).
3. Run structural form regression with this residual as extra variable.
4. If this is significant (t-test), then there is substantial cov(x,u) and you should use 2SLS.

**Testing for exogeneity of the IVs**

Can only do this if the model is overidentified (#IVs>#endogenous right-hand-side variables), not if “just identified” (#IVs=#endog).

1. Run 2SLS and save residual from the final regression.
2. Regress this residual on all exogenous variables (exogenous Xs and IVs).
3. If nR2> chi-squared critical value with q=(#IV’s-#endog) degrees of freedom, at least some IVs are NOT exogenous. Consequence: BIV biased and inconsistent.

**Causes of endogeneity bias:**

* There’s a missing variable correlated with included explanatory variable.
* Reverse causality (x causes y, but also y causes x)
* Simultaneous equations (supply-demand example)
	+ Indirect least squares estimation
* Measurement error in x