Final Exam

August 11^{th} 2011

Student name:.....

Instructions:

- Write your name in the space provided above.
- You have 2 hours to complete the exam.
- Please read all question and possible answers carefully before choosing your answer.
- In the multiple choice questions, indicate your answer by circling the letter that corresponds to the correct answer.
- In the very short answers and short answers questions write your answer in the space provided.
- You can earn a total of 100 points:
  - 30 points total for the multiple choice questions section
  - 28 points total for the very short answers questions
  - 42 points total for the short answers questions part

Good luck!
For your convenience, A1-A6 are:

- A1: Linearity in parameters
- A2: Random sampling
- A3: Variation in the $x's$ / No perfect collinearity
- A4: Zero Conditional Mean
- A5: Homoskedasticity
- A6: Normality of errors

Multiple choice questions (3 points each)

Please circle the answer which you think is correct. There are no negative points, so do not leave any question without an answer.

1. A p-value of 0.035 for the model is obtained for a one-tailed test. What conclusions can you draw from this result?
(a) The null hypothesis will be rejected at any significance level greater than 3.5 percent for a one tailed test, or greater than 7 percent for a two tailed test
(b) The null hypothesis will be rejected at any significance level less than 3.5 percent for a one-tailed test, or less than 7 percent for a two tailed test
(c) The null hypothesis will be rejected at any significance level greater than 7 percent for a one-tailed test, or greater than 7 percent for a two tailed test
(d) c) The null hypothesis will be rejected at any significance level greater than 3.5 percent for a one-tailed test, or greater than 1.75 percent for a two tailed test
(e) None of the above

2. This question refers to a multiple linear regression model with 4 regressors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$$

The model is estimated using OLS regression; assume A1-A6 are satisfied.

Consider an F test of the null hypothesis: $H_0 : \beta_2 = 1, \beta_3 = \beta_4$ against the alternative that at least one of these restrictions does not hold. What will be the unrestricted and restricted regression used to perform this test?
(a) Unrestricted: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$
Restricted: $y_i - x_{i2} = \beta_0 + \beta_1 x_{i1} + (\beta_3 - \beta_4) x_{i3} + \beta_4 (x_{i4} + x_{i3}) + u_i$
(b) Unrestricted: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$
Restricted: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 (x_{i3} + x_{i4}) + u_i$
(c) Unrestricted: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$
Restricted: $y_i - x_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_3 (x_{i3} + x_{i4}) + u_i$
(d) Unrestricted: $y_i - x_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_3 (x_{i3} + x_{i4}) + u_i$
Restricted: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i$
(e) None of the above.

3. Consider the following multiple linear regression model using performance measures to explain the salaries of baseball players:

$$ln(s_i) = \beta_0 + \beta_1 ba_i + \beta_2 hr_i + \beta_3 rb_i + u_i$$

where s_i denotes salary, ba_i batting average, hr_i home runs per year, and rb_i runs on base. An OLS regression returns $R^2 = 0.493$; $n = 207$.

You want to test whether performance measures have any effect on salary, i.e. to test significance of the regression. Which of the following is a good approximation of the F-statistic?
(a) 67.67
(b) 65.80
(c) It cannot be computed without knowing the R^2 of the restricted model.
(d) It cannot be computed since the null hypothesis implies a change in the dependent variable between the unrestricted and restricted models.
(e) None of the above.

The next **two** questions refer to the following multiple regression model with heteroskedasticity of a known form:

$$y_i = \beta + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$
$$\text{where } Var(u_i|X_i) = \sigma^2 h(X_i)$$

The model is estimated using WLS, i.e. a regression on a transformed model:

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + u_i^*$$

4. Which of the following equations gives the transformed model for the WLS estimation?
(a) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \frac{u_i}{\sqrt{h(x_i)}}$
(b) $\frac{y_i}{h(x_i)} = \frac{\beta_0}{h(x_i)} + \beta_1 \frac{x_{i1}}{h(x_i)} + \beta_2 \frac{x_{i2}}{h(x_i)} + \beta_3 \frac{x_{i3}}{h(x_i)} + \frac{u_i}{h(x_i)}$
(c) $\frac{y_i}{\sqrt{h(x_i)}} = \frac{\beta_0}{\sqrt{h(x_i)}} + \beta_1 \frac{x_{i1}}{\sqrt{h(x_i)}} + \beta_2 \frac{x_{i2}}{\sqrt{h(x_i)}} + \beta_3 \frac{x_{i3}}{\sqrt{h(x_i)}} + \frac{u_i}{\sqrt{h(x_i)}}$
(d) $\frac{y_i}{\sqrt{\hat{h}(x_i)}} = \frac{\beta_0}{\sqrt{\hat{h}(x_i)}} + \beta_1 \frac{x_{i1}}{\sqrt{\hat{h}(x_i)}} + \beta_2 \frac{x_{i2}}{\sqrt{\hat{h}(x_i)}} + \beta_3 \frac{x_{i3}}{\sqrt{\hat{h}(x_i)}} + \frac{u_i}{\sqrt{\hat{h}(x_i)}}$
(e) None of the above.

5. What is the most accurate and appropriate interpretation of the parameter β_2
(a) The effect of x_2^* on y^*
(b) The effect of x_2^* on y
(c) The effect of x_2 on y
(d) It has no useful interpretation.
(e) None of the above.

The next **three** questions refer to the following results from OLS regression attempting to explain marital fidelity. The dependent variable is the number of extra marital affairs. Standard errors are in parentheses below the estimates.

	Specification 1	Specification 2
age	-0.0469	-0.0504
	(0.0227)	(0.0234)
years of marriage	0.1508	0.1651
	(0.038)	(0.0425)
years of education	0.011	-0.0023
	(0.0554)	(0.0598)
very happy	-0.9571	-0.9656
	(0.2775)	(0.2799)
male		0.1641
		(0.2951)
children		-0.2191
		(0.3585)
constant	1.937	2.2306
	0.9803	(1.0557)
n	601	601
R^2	0.060	0.061

veryhappy is the dummy variable equal to 1 if the individual describes his marriage as "very happy", *male* is a dummy variable equal to 1 if the individual is male, and *children* is a dummy variable equal to 1 if the individual has children.

Assume A1-A6 are satisfied.

6. Which of the following is a good approximation of the 95 percent confidence interval for the parameter on years of education in the first specification?
(a) $[-0.1195, 0.01149]$
(b) $[-0.1317, 0.1527]$
(c) $[0.0976, 0.1196]$
(d) $[-0.0976, 0.1196]$
(e) None of the above

7. Consider the parameter on the dummy variable for happiness in the second specification. What can you conclude?
(a) Holding all other regressors constant, happiness is statistically significant at the 5 percent significance level, but insignificant at 1 percent.
(b) Holding all other regressors constant, happiness is statistically significant at any reasonable significance level.
(c) Holding all other regressors constant, happiness is statistically insignificant at any reasonable significance level.
(d) It is impossible to make a statement about the significance of happiness, as it may be correlated with other variables.
(e) None of the above.

8. Perform an F-test to test the null that gender and children are jointly insignificant. Your conclusion is:
(a) Fail to reject the null at any reasonable significance level.
(b) Fail to reject the null at 1 percent significance level but reject at a 5 percent significance level.
(c) Fail to reject the null at 5 percent significance level but reject at a 10 percent significance level.
(d) The null is rejected at any reasonable significance level.
(e) None of the above.

The next **two** questions consider the following model:

$$y = \beta_0 + \beta_1 y_2 + \beta_2 x + u$$

Where y_2 is endogenous ($E(u|y_2) \neq 0$), x is exogenous ($E(u|x) = 0$) and you have the instrument z ($E(u|z) = 0$ and $Cov(z, y_2) \neq 0$).

9. You run the IV estimation and Stata returned you $R^2 = 0.15$. What can you infer from this:
(a) 15 percent of variation in y is explained by x.
(b) 0.15 percent of variation in y is explained by x and y_2.
(c) 15 percent of variation in y is explained by the instrument z.
(d) Since the R^2 is relatively low, the corresponding F statistic (for significance of the whole regression) is also low, so the regression is insignificant.
(e) Nothing, as R^2 does not have a useful interpretation in IV estimation.

10. By running the IV regression, for which parameters can we obtain the unbiased estimates?
(a) β_0, β_1 and β_2
(b) β_0 and β_2
(c) β_0, β_1, β_2 and the coefficient on the instrument z.
(d) β_0, β_2 and the coefficient on the instrument z.
(e) We cannot obtain unbiased estimates using IV estimator.

Very short answers questions (4 points each)

1. (4 points) What is are the consequences of heteroskedasticity for the properties of the OLS estimator? (Is OLS unbiased? Is it efficient?) Explain.

2. (4 points) WLS or FGLS is used instead of OLS if an important variable has been omitted from the model. True or false? Explain.

3. (4 points) Look at the stata output (1) at the end of the exam. Which test has been used to test for heteroskedasticity in this example (Breush and Pagan or White test)? What do you conclude about presence of heteroskedasticity in this model? Explain.

4. (4 points) Using the same Stata output (1) test the hypothesis that, holding everything else constant, boys and girls do not differ in terms of weight at birth.

5. (4 points) Consider the Stata output (2) provided at the end of the test. What can you conclude about the whole regression model based on these results? Explain.

6. (4 points) Consider the following model:

$$y_1 = \beta_0 + \beta_1 y_{2i} + \beta_2 x_i + u_i$$

Where we suspect that y_2 is endogenous. When testing for endogeneity, we first obtain residuals from regressing the endogenous variable (y_2) on all exogenous variables in the model ($x's$ and $z's$) and obtain the residuals from this regression ($\hat{v}$). Than we estimate the main model with the residuals from the first regression ($\hat{v}$). So we estimate the following model:

$$y_1 = \beta_0 + \beta_1 y_{2i} + \beta_2 x_i + \delta \hat{v}_i + error$$

What do we conclude if we find that δ is not significantly different from zero?

7. (4 points) Consider the following model with endogeneity:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Where we suspect that x is endogenous. Suppose we have one good instrument available for x, z. Instrumental variable estimator, when estimating the effect of x on y ($\hat{\beta}_1$) ignores the information on x and only uses information on z. True or false. Explain.

Short answers questions

1. (10 points) Consider the following model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$ and assume that $Var(u_i|x_{i1}, x_{i2}) = \sigma^2(x_{i1} + x_{i2})^2$. Write down the transformed model and show that the new error term is homoskedastic.

2. (10 points) Consider the following model relating savings to household income:

$$sav_i = \beta_0 + \beta_1 inc_i + \beta_2 M_i + \beta_3 C_i + u_i$$

where sav_i denotes household savings, inc_i household income, M_i is a dummy variable that takes value 1 if the household has a mortgage, and C_i is a dummy variable that take on value 1 if there are children in the household.

Assume that A1-A4 hold but the model exhibits heteroskedasticity of an unknown form: $Var(u_i|X_i) = \sigma^2 h(X_i)$, where X_i denotes all independent variables in the model.

Explain how you would obtain unbiased and efficient estimator for this model, providing description of any additional regressions you will run or transformation you will make.

3. (10 points) Consider the following model:

$$log(rent) = \beta_0 + \beta_1 log(pop) + \beta_2 log(avginc) + \beta_3 pctstu + u$$

where: *rent* is the average monthly rent in a college town in the US, *pop* is the population of this town, *avginc* is the average city income and *pctstu* is the student population as a percentage of the total population.
Stata output with estimates of this model is provided at the end of the test (Stata output 3).

(a) (5 points) Test the hypothesis that size of the student population relative to the total population has no effect on monthly rent. State the null and the alternative hypothesis clearly.

(b) (5 points) **Show how you would test** the hypothesis that 1 percent increase in average city income increases the average monthly rent by 1 percent **and** once average income is controlled for, the population of the town and the size of the student population does not matter for the monthly rent. State the null and alternative hypothesis clearly. (You don't have to perform the actual test - but show cleraly how you would do it.)

4. (12 points) Consider a simple regression model with endogeneity:

$$\ln(wage) = \beta_0 + \beta_1 education + u$$

You also have a measure of IQ available in your data set and years of father schooling. (In your answers to these questions, you don't have to provide equations - but explain the logic of each method and what is identified in each estimation.)

(a) (4 points) Explain what causes endogeneity in this model.

(b) (8 points) Explain how you would estimate it using an instrumental variable method. Which parameters are identifed (estimated)? Give interpretation of the estimated coefficients (you don't have to interprete the constant).