DEPARTMENT OF ECONOMICS
UNIVERSITY OF VICTORIA

ECON 345: Applied Econometrics

Summer 2010

MIDTERM EXAM - Solutions

Duration: 60 minutes **Total Marks: 50 marks**

INSTRUCTIONS:

1. This exam contains a total of 6 pages; 5 pages of exam AND 1 page for a statistical table. Please count the number of pages in this examination paper before beginning to write, and report any discrepancy IMMEDIATELY to an invigilator
2. Answer ALL questions in the booklets provided.
3. EXCEPT for a single sheet of 8.5x11 paper (with notes allowed on both sides), all notes, papers, and electronic devices other than non-programmable calculators must be put away. Please turn off any device that could make a noise.
4. Please have your Student ID out on your desk for invigilators to inspect.

SECTION 1: Multiple Choice Questions

(Total marks: 5)

Choose the correct answer for each question and record it in the booklet provided. NO explanation is required. CLEARLY record your answer (i.e. A, B, C or D); if I cannot read your writing I will not be able to mark it.

1. In cross-sectional analysis, the Classical Linear Model (CLM) assumptions state that:

    A. **The errors must have a constant variance and be normally distributed**.
    B. The errors must be normally distributed and correlated with the explanatory variables.
    C. The model must be linear in the explanatory variables and the errors must be normally distributed.
    D. None of the above.

2. Consider the following population model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$. Suppose we had a sample of data ($n = 100$) and we use this data to obtain the following estimated regression, $\hat{y} = 2.4 + 1.4x_1 + 0.02x_2 + 0.9x_3$. If we want to test the hypothesis that x_3 has no effect on y, then our null hypothesis would be:
   A. $H_0: \hat{\beta}_3 = 0$
   B. $H_0: 0.9 = 0$
   C. $H_1: \beta_3 = 0$
   D. None of the above

3. The coefficient of determination, R^2, tells us:
   A. The amount of variation in the dependent variable that is explained by the disturbance term.
   B. The amount of variation in the dependent variable that is explained by the intercept coefficient.
   C. The amount of variation in the dependent variable that is explained by the independent variables.
   D. None of the above.

4. Suppose we have the following estimated regression, $\widehat{\log(rent)} = 0.043 + 0.066\log(pop) + 0.507\log(avginc)$, where, $rent$ is the average monthly rent, pop is the total city population, and $avginc$ is the average city income. Then we can say:
   A. A 10% increase in population is associated with about a 0.066% increase in rent.
   B. A 10% increase in population is associated with about a 0.66% increase in rent.
   C. A 10% increase in population is associated with about a 6.6% increase in rent.
   D. None of the above.

5. In a multiple linear regression the assumption of homoskedastic errors means:
   A. The variance of the error term is constant, given any values of the independent variables.
   B. The variance of the error term is not constant, given any values of the independent variables.
   C. The variance of the error term is 0, given any values of the independent variables.
   D. None of the above.

SECTION 2: True/False Questions

(Total marks: 25; each question is worth 5 marks)

STATE whether each of the following statements is TRUE or FALSE. Briefly explain why.
(1 mark is for true/false statement, 4 marks are for your explanation).

1. Suppose the fitted equation is $\widehat{colGPA} = 1.29 + 0.453hsGPA + 0.0094ACT$. This regression explains college GPA ($colGPA$) in terms of high school GPA ($hsGPA$) and achievement test score (ACT). If the average high school GPA is about 3.4 and the average *ACT* score is about 24.2, then the average college GPA in the sample is 3.06.

   True. *We use the third algebraic property of OLS statistics which states that the OLS regression line always goes through the mean of the sample,* $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$. *That is, when we plug the average values of all independent variables into the OLS regression line, we obtain the average value of the dependent variable. So here we have,*

   $\overline{colgpa} = 1.29 + 0.453\,\overline{hsgpa} + 0.0094\,\overline{sat} = 1.29 + 0.453\,(3.4) + 0.0094\,(24.2) = 3.06$

2. Consider the case where you estimate parameters of the population model $y = \beta_0 + \beta_1 x + u$. For your sample, you find that the *total sum of squares,* $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 800$ and the *residual sum of squares,* $\sum_{i=1}^{n}(\hat{u})^2 = 300$. Then the R^2 is 0.375.

   False*. By definition,.* $R^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 = 1 - \sum_{i=1}^{n}\hat{u}_i^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$. *We have been given residual sum of squares as* $\sum_{i=1}^{n}\hat{u}_i^2 = 300$, *and total sum of squares as* $\sum_{i=1}^{n}(y_i -$ $y2=800$. *Therefore,.* $R2=1-300800=0.625$

3. Consider the following population regression model, $crime = \beta_0 + \beta_1 age + \beta_2 inc + \beta_3 inc^2 + u$, where $crime$ is the number of crimes committed, age is the age of the person, and inc is family income. This is an example of a linear regression model.

   True. *This model is linear because it is linear in the parameters, (it satisfies assumption MLR.1). If the dependent and independent variables are expressed in various functional forms (such as squared, logarithmic etc.) the model is still considered linear as long as the beta's are linear, i.e. the beta's are not expressed as squares, logarithms or any other functional form.*

4. Including an irrelevant variable in a model has an effect on the unbiasedness property of the intercept and slope estimators.

   False. *Under assumptions MLR.1 to MLR.4 we establish that the OLS estimators are unbiased even if an irrelevant variable is included in our model. Unbiasedness means that* $E(\hat{\beta}_j) = \beta_j$ *for any value of* β_j, *including* $\beta_j = 0$. *So, if our true model is* $y = \beta_0 + \beta_1 x_1 + u$ *but we actually estimate* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ *instead (i.e. we've include* x_2, *when really we shouldn't have since* $\beta_2 = 0$ *in the true model) our estimators will still be unbiased). That is, for our intercept we have* $E(\hat{\beta}_0) = \beta_0$, *for our slope parameters we have* $E(\hat{\beta}_1) = \beta_1$ *and* $E(\hat{\beta}_2) = \beta_2 = 0$.

   Note that inclusion of an irrelevant variable in a model will generally increase the variance of the OLS estimators of the other variables because of multicollinearity, and our estimator will be less precise.

5. Consider the hypothesis set up, $H_0: \beta_1 = 0$ *vs* $H_1: \beta_1 > 0$. You obtain a $t - statistic$ of 1.08 with a sample of size 20 from a regression of a dependent variable on 4 independent variables and a constant. Conducting a hypothesis test at the 10% significance level we conclude that we fail to reject the null hypothesis.

   True. *The degrees of freedom is n – k – 1 = 20 – 4 – 1 = 15. Looking up our critical value for a one-sided t test with 15 degrees of freedom at the 10% significance level gives us c = 1.341. Now t stat is 1.08. Since 1.08 < 1.341, we fail to reject* $H_0 : \beta_1 = 0$ *at the 10% significance level. Therefore,* β_1 *is not statistically different from zero.*

SECTION 3: Short Answers Questions Based on EViews Output

(Total marks: 20)

ALL questions in this section refer to the situation below.

The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u$$

where,

$sleep$ is the number of minutes slept per week
$totwrk$ is the number of minutes worked per week
$educ$ is the number of years of schooling
age is a person's age in years

The above model was estimated in EViews using the SLEEP75.RAW data set and the following output was obtained:

Dependent Variable: SLEEP
Method: Least Squares
Date: 06/08/10 Time: 20:19
Sample: 1 706
Included observations: 706

	Coefficient	Std. Error	t-Statistic	Prob.
C	3638.245	112.2751	32.40474	0.0000
TOTWRK	-0.148373	0.016694	-8.888075	0.0000
EDUC	-11.13381	5.884575	-1.892034	0.0589
AGE	2.199885	1.445717	1.521657	0.1285

R-squared	0.113364	Mean dependent var	3266.356
Adjusted R-squared	0.109575	S.D. dependent var	444.4134
S.E. of regression	419.3589	Akaike info criterion	14.92098
Sum squared resid	1.23E+08	Schwarz criterion	14.94681
Log likelihood	-5263.106	Hannan-Quinn criter.	14.93096
F-statistic	29.91889	Durbin-Watson stat	1.942609
Prob(F-statistic)	0.000000		

a. Write out the results in equation form along with the standard errors in parenthesis, sample size and R-squared. (*You may round your results to 2 decimal places*) ***(2 marks)***

$$\widehat{sleep} = 3638.245 - 0.148\, totwrk - 11.134\, educ + 2.199 age$$
$$(112.28) \quad (0.0167) \quad (5.885) \quad (1.446)$$
$$R^2 = 0.113 \qquad n = 706$$

b. Suppose you need to explain these results to a friend. Interpret the coefficient, $\hat{\beta}_2$, making reference to correct units. ***(3 marks)***

$\hat{\beta}_2 = -11.134$. It tells us about the relationship between $sleep$ and $educ$ variables.

If the year of education increases by one year, then holding *totwrk* and *age* fixed, the number of minutes spent sleeping per week is expected to decrease by 11.135 minutes.

c. Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? ***(3 marks)***

The three explanatory variables, totwrk, educ, and *age, only explain 11.3% in the variation in sleep (this is from the R^2). This seems rather small, so perhaps there are other factors that we have not accounted for in our regression that may help explain more of the variation in the number of minutes slept during a week.*

d. Derive the unbiased estimator for the population error variance ($Var(u) = \sigma^2$). ***(3 marks)***
The unbiased estimator for σ^2 is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{1.23E+08}{706-3-1} = 175213.68$$

e. What is a confidence interval? Construct a 95% confidence interval for β_3 and interpret your result. ***(3 marks)***

A 95% confidence interval for β_3 is: $\hat{\beta}_3 \pm t_c \times se(\hat{\beta}_3)$

For 95% confidence interval, the level of significance is 5%. For a two-tailed test statistic, the critical value of t with 702 df (n-k-1) is 1.96. So the confidence interval is -0.635 to 5.033, derived as:

$$[2.199 \pm 1.96 \times 1.446]$$
$$[2.199 \pm 2.834]$$
$$[2.199 \pm 2.834]$$
$$[-0.635\ to\ 5.033]$$

f. Test the null hypothesis that *educ* has no effect on *sleep* against the alternative that *educ* has a negative effect. Carry out the test at the 5% level of significance. (You may follow the 5 steps in classical hypothesis testing. You must show all works to get the full marks, *i.e.*, the appropriate test statistic, critical value selection, the degrees of freedom, define a decision rule, etc.) ***(6 marks)***

Step 1. Specify the hypotheses

The Null Hypothesis $H_0: \beta_{educ} = 0$

$Alternative$ $H_1: \beta_{educ} < 0$

Step 2. Level of Significance, say, $\alpha = 0.05$

Step 3. Specify the test statistic. It's a *t*-statistic -

$$t_{\hat{\beta}_{educ}} = \frac{\hat{\beta}_{educ} - \beta_{educ}}{se(\hat{\beta}_{educ})}$$

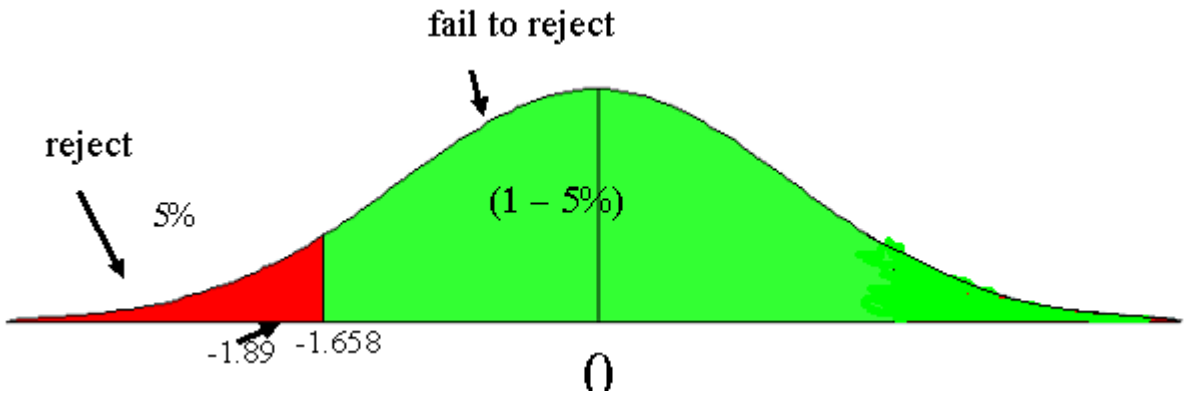
If the null is true,

$$t_{\hat{\beta}_{educ}} = \frac{\hat{\beta}_{educ} - 0}{se(\hat{\beta}_{educ})} = \frac{\hat{\beta}_{educ}}{se(\hat{\beta}_{educ})} = \frac{-11.134}{5.885} = -1.892034$$

Step 4. Formulate the decision rule

At 5% level on one-tailed test the critical value of t, t_c with 702 df is -1.658.

We reject the null hypothesis if the calculated $t - statistic < t_c$. And, here, $t - statistic\ (= -1.892) < t_c (= -1.658)$. So, we reject the null hypothesis.



Step 5. Conclusion

At 5% level we reject the null hypothesis. Meaning, the estimator for *educ* is statistically significantly different from 0. Hence, there is a significant negative relationship between *sleep* and *educ* at 5% level of significance.

One Bonus Question (3 marks)

Suppose the true model is given by, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \ (a)$. But we have estimated a misspecified model with only x_1, which is $\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1}$. Then the estimator of the slope parameter (from the misspecified model) is, $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$. Prove that $\tilde{\beta}_1$is a biased estimator for β_1.

From page 23, ch 3 lecture notes

Now we can check if this estimator is unbiased, given the actual model is (a):

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$= \frac{\beta_0(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1) + \beta_1 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i1} + \beta_2 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2} + \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

Recall: $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1) = 0$ and $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i1} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2$

$$= \frac{[\beta_1 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 + \beta_2 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2} + \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i]}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

Taking expectations (conditional on x_1) and for $E(u_i) = 0$, we get

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 E\left(\frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}\right)$$

So, the estimator $\tilde{\beta}_1$ is biased.

END OF EXAM

ONE STATISTICAL TABLE FOLLOWS

TABLE G.2

Critical Values of the *t* Distribution

		Significance Level				
1-Tailed:		.10	.05	.025	.01	.005
2-Tailed:		.20	.10	.05	.02	.01
Degrees of Freedom	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Examples: The 1% critical value for a one-tailed test with 25 *df* is 2.485. The 5% critical value for a two-tailed test with large (> 120) *df* is 1.96.

Source: This table was generated using the Stata® function invttail.

Source : Wooldridge (2009)