Econ 3180 – Final Exam, April 17th 2014

Ryan Godwin

You may use a calculator. Answer all questions in the answer book provided. The exam is 3 hours long and consists of 248 marks.

A formula sheet, and a table of probabilities from the standard Normal distribution, are provided at the back of the exam booklet.

DO NOT OPEN THE EXAM UNTIL YOU ARE INSTRUCTED TO DO SO.

NAME:	
STUDENT #:	

Part A – Multiple Choice

[48 marks – 4 marks each]

1) A large *p*-value implies

a) rejection of the null hypothesis.
b) a large *t*-statistic.
c) a large estimate.
d) that the estimated value is consistent with the null hypothesis.

2) The OLS residuals

a) can be calculated using the errors from the regression function.
b) can be calculated by subtracting the predicted values from the actual values.
c) are unknown since we do not know the population regression function.
d) should not be used in practice since they indicate that your regression does not run through all your observations.

3) If you wanted to test, using a 5% significance level, whether or not a specific slope coefficient is equal to one, then you should

a) subtract 1 from the estimated coefficient, divide the difference by the standard error, and check if the resulting ratio is larger than 1.96.
b) add and subtract 1.96 from the slope and check if that interval includes 1.
c) see if the slope coefficient is between 0.95 and 1.05.
d) check if the adjusted R^2 is close to 1.

4) The formula for the OLS estimator, $\hat{\beta}_1$, when moving from one regressor to two regressors,

a) stays the same.
b) changes, unless the second regressor is a dummy variable.
c) changes, unless the second variable is uncorrelated with the first variable.
d) changes.

5) The error term is homoskedastic if

a) $var(u_i \mid X_i = x)$ is constant for $i = 1,\ldots, n$.
b) $var(u_i \mid X_i = x)$ depends on x.
c) X_i is normally distributed.
d) there are no outliers.

6) The sampling distribution is

a) a subset of the population.
b) Normal because of the Central Limit Theorem.
c) identically and independently distributed.
d) the probability distribution of an estimator.

7) Imagine you regressed earnings of individuals on a constant, a binary variable ("Male") which takes on the value 1 for males and is 0 otherwise, and another binary variable ("Female") which takes on the value 1 for females and is 0 otherwise. Because females typically earn less than males, you would expect

a) the coefficient for Male to have a positive sign, and for Female a negative sign.
b) both coefficients to be the same distance from the constant, one above and the other below.
c) none of the OLS estimators to exist because there is perfect multicollinearity.
d) this to yield a difference in means statistic.

8) In multiple regression, the adjusted R^2 ($\bar{R}^2$)

a) is always larger than R^2
b) cannot decrease when a variable is added to the regression
c) is an unbiased estimator of R^2
d) takes into account the number of regressors in the model

9) When testing joint hypothesis, you should

a) use t-statistics for each hypothesis and reject the null hypothesis is all of the restrictions fail.
b) use the F-statistic and reject all the hypothesis if the statistic exceeds the critical value.
c) use t-statistics for each hypothesis and reject the null hypothesis once the statistic exceeds the critical value for a single hypothesis.
d) use the F-statistics and reject at least one of the hypothesis if the statistic exceeds the critical value.

10) You have estimated the following equation:

$$TestScore = 607.3 + 3.85 Income - 0.0423 Income^2,$$

where *TestScore* is the average of the reading and math scores on the Stanford 9 standardized test administered to 5^{th} grade students in 420 California school districts in 1998 and 1999. *Income* is the average annual per capita income in the school district, measured in thousands of 1998 dollars. The equation

a) suggests a positive relationship between test scores and income for most of the sample.
b) is positive until a value of *Income* of 610.81.
c) does not make much sense since the square of income is entered.
d) suggests a positive relationship between test scores and income for all of the sample.

11) For the polynomial regression model,

a) you need new estimation techniques since the OLS assumptions do not apply any longer.
b) the techniques for estimation and inference developed for multiple regression can be applied.
c) you can still use OLS estimation techniques, but the t-statistics do not have an asymptotic normal distribution.
d) the critical values from the normal distribution have to be changed to 1.962, 1.963, etc.

12) An example of the interaction term between two independent, continuous variables is

a) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$.
b) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.
c) $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$.
d) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$.

Part B – Short Answer [100 marks – 10 marks each]

S.1) Consider the following data:

$$Y_1 = 2 \ , \ Y_2 = 3 \ , \ Y_3 = 4 \quad ; \quad X_1 = 3 \ , \ X_2 = 4 \ , \ X_3 = 6.$$

Assume that the population model is:

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Calculate the OLS estimate for the slope coefficient (calculate $\hat{\beta}_1$).

S.2) What is the R^2 (not the $\bar{R}^2$) from the above regression?

S.3) The following regression line is estimated by OLS: $\widehat{wage} = 8.34 + 1.52 \times exper$, where $wage$ is the hourly earnings of workers and $exper$ is years of experience. What is the predicted increase in hourly earnings from an additional two years of experience?

S.4) Consider the following situations:

i) The standard error of $\hat{\beta}_1$ is estimated assuming h<u>eteroskedasticity</u>, but the random errors are actually <u>homoskedastic</u>.

ii) The standard error of $\hat{\beta}_1$ is estimated assuming <u>homoskedasticity</u>, but the random errors are actually <u>heteroskedastic</u>.

Which situation is worse? Explain.

S.5) The population model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

What might be the consequence of leaving X_2 out of the regression? Explain.

S.6) Suppose that a researcher estimates the OLS regression:

$$\underset{}{score} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28}str, \quad n = 420, \quad R^2 = 0.049$$

where $score$ is average test scores, and str is the student-teacher ratio. Conduct a formal hypothesis test to address whether or not class size affects test scores.

S.7) Suppose that I accidentally include the same variable in my regression, twice. Will OLS still be unbiased?

S.8) Suppose that you conduct an F-test and conclude that you can remove X_3 and X_4 from the regression. After you remove them, however, the R^2 decreases. Does this mean you should still include X_3 and X_4 in the regression?

S.9) Which is worse: including irrelevant regressors, or excluding relevant regressors? (Is it worse to include variables that don't matter, or to leave out variables that do matter?) Explain.

S.10) Provide an example of a restricted model, obtained from a null hypothesis.

Part C – Long Answer [100 marks total – 10 marks for each part]

This question uses data for full-time, full-year workers, age 25-34, with a high school diploma or B.A./B.S. as their highest degree.

Variables in Data Set

Name	**Description**
ahe	average hourly earnings of the worker
female	1 if female; 0 if male
age	age of the worker
bach	1 if worker has a bachelor's degree, 0 if worker has a high school degree

The sample size is 7985. Average hourly earnings (*ahe*) is the dependent variable (the right-hand-side variable) in all regressions. Below is a table of estimated models which you should use for parts (a) – (i).

	Model				
Regressor	(1)	(2)	(3)	(4)	(5)
bach	6.49** (0.18)	6.86** (0.18)	6.86** (0.18)	6.99** (0.23)	6.96** (0.24)
female		-3.16** (0.18)	-3.15** (0.18)	-3.00** (0.25)	3.50 (1.86)
age		0.44** (0.03)	2.05** (0.72)	2.06** (0.72)	0.53** (0.04)
age^2			-0.03* (0.01)	-0.03* (0.01)	
female× *bach*				-0.30 (0.36)	-0.31 (0.36)
female× *age*					-0.22** (0.06)
intercept	13.81** (0.12)	1.88* (0.92)	-21.80* (10.53)	-21.90* (10.53)	-0.89* (1.21)
R^2	0.1364	0.1898	0.1904	0.1905	0.1912
$\bar{R}^2$	0.1363	0.1895	0.1900	0.1899	0.1907

Significance at the *5% and **1% significance level.

a) Does it make sense to interpret the intercept in this model? Explain.

b) Using model (2), what is the estimated effect of *age* on earnings? Construct a 95% confidence interval for the coefficient on *age* using model (2).

c) Does the regression in (1) seem to be suffering from important omitted variable bias?

d) Does the effect of age on earnings appear to be linear or non-linear? Support your answer with evidence.

e) Using model (3), predict the earnings for a 26-year-old male worker with a high school diploma.

f) Is the effect of a having bachelor's degree different for women than it is for men? Support your answer with evidence.

g) Using model (5), what is the predicted effect of *age* on earnings for women, and for men? Test the null hypothesis that the difference is zero.

h) Using model (2), calculate the F-statistic for the null hypothesis that both *female* and *age* have no effect on earnings.

i) Going from model (3) to model (4): how is it possible that R^2 increases while $\bar{R}^2$ decreases?

j) What model would you estimate next? **END.**

Econ 3180 - Final Formula Sheet

expected value of Y (mean of Y)	μ_Y
variance of Y	$\sigma_Y^2 = E(Y-\mu_Y)^2 = E(Y^2) - (\mu_Y)^2$
standard deviation of Y	$\sigma_Y = \sqrt{\sigma_Y^2}$
covariance between X and Y	$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$
correlation coefficient (between X and Y)	$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
expected value of the sample average, $\bar{Y}$	$E(\bar{Y}) = \mu_Y$
variance of the sample average, $\bar{Y}$	$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n}$
t-statistic for testing μ_Y (for large *n*, and when σ_Y^2 is *known*)	$t = \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \sim N(0,1)$
sample variance (estimator for σ_Y^2)	$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
sample covariance (estimator for covariance)	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
sample correlation (estimator for correlation)	$r_{xy} = \frac{s_{xy}}{s_x s_y}$
standard error of $\bar{Y}$ (estimator for the standard deviation of $\bar{Y}$)	$s_{\bar{Y}} = \sqrt{\frac{s_Y^2}{n}}$
t-statistic for testing μ_Y (for large *n*, and when σ_Y^2 is *unknown*)	$t = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_{\bar{Y}}} \sim N(0,1)$
95% confidence interval for μ_Y (for large *n*)	$conf.int. = \bar{Y} \pm 1.96 \times s_{\bar{Y}}$
population linear regression model with one regressor	$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \ldots, n$
OLS estimator of the slope (β_1)	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$
OLS estimator of the intercept (β_0)	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
OLS predicted values	$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
OLS residuals	$\hat{u}_i = Y_i - \hat{Y}_i$

explained sum of squares	$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$
total sum of squares	$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
sum of squared residuals	$SSR = \sum_{i=1}^{n} \hat{u}_i^2$
regression R^2	$R^2 = \frac{ESS}{TSS}$
standard error of regression	$\sqrt{\frac{1}{n-2} \times SSR}$
L.S.A. #1	$E(u \mid X = x) = 0$
L.S.A. #2	$(X_i, Y_i), i = 1, \dots, n$, are i.i.d.
L.S.A. #3	Large outliers are rare.
The sampling distribution of $\hat{\beta}_1$ (for large *n*)	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{var[(X_i - \mu_X)u_i]}{n\sigma_X^4}\right)$
t-statistic for testing β_1	$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)}$
95% confidence interval for β_1 (for large *n*)	$conf.int. = \hat{\beta}_1 \pm 1.96 \times SE\left(\hat{\beta}_1\right)$
alternative regression R^2	$R^2 = 1 - \frac{SSR}{TSS}$
adjusted R-square ($\bar{R}^2$)	$\bar{R}^2 = 1 - \frac{SSR}{TSS}\left(\frac{n-1}{n-k-1}\right)$
F-statistic	$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n - k_U - 1)}$
F-statistic	$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)}$