Econ 3180 – Final Exam, April 25th 2015

Professor: Ryan Godwin

You may use a calculator. Answer all questions in the answer book provided. The exam is 3 hours long and consists of 100 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 – each question is worth 1 mark

1) A large p-value implies
a. rejection of the null hypothesis.
b. a large t-statistic.
c. a large $\bar{Y}^{act}$.
d. that the observed value $\bar{Y}^{act}$ is consistent with the null hypothesis.

2) To standardize a variable you
a. subtract its mean and divide by its standard deviation.
b. integrate the area below two points under the normal distribution.
c. add and subtract 1.96 times the standard deviation to the variable.
d. divide it by its standard deviation, as long as its mean is 1.

3) Suppose that a normally distributed random variable, Y, has mean 5 and variance 4. That is, $Y \sim N(5,4)$. What is the probability that $Y \geq 8.92$?
a. 0.01
b. 0.025
c. 0.05
d. None of the above.

4) The following are all least squares assumptions with the exception of:
a. The conditional distribution of u_i given X_i has a mean of zero.
b. The explanatory variable in the regression model is normally distributed.
c. $(X_i, Y_i), i = 1, ..., n$ are independently and identically distributed.
d. Large outliers are unlikely.

5) $E(u_i \mid X_i) = 0$ says that
a. dividing the error by the explanatory variable results in a zero (on average).
b. the sample regression function residuals are unrelated to the explanatory variable.
c. the sample mean of the Xs is much larger than the sample mean of the errors.
d. the conditional distribution of the error given the explanatory variable has a zero mean.

6) Heteroskedasticity means that
a. homogeneity cannot be assumed automatically for the model.
b. the variance of the error term is not constant.
c. the observed units have different preferences.
d. agents are not all rational.

7) The confidence interval for the sample regression function slope
a. can be used to conduct a test about a hypothesized population regression function slope.
b. can be used to compare the value of the slope relative to that of the intercept.
c. adds and subtracts 1.96 from the slope.
d. allows you to make statements about the economic importance of your estimate.

8)	When there are omitted variables in the regression, which are determinants of the dependent variable, then

a.	you cannot measure the effect of the omitted variable, but the estimator of your included variable(s) is (are) unaffected.
b.	this has no effect on the estimator of your included variable because the other variable is not included.
c.	this will always bias the OLS estimator of the included variable.
d.	the OLS estimator is biased if the omitted variable is correlated with the included variable.

9)	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.

10)	In the multiple regression model, the least squares estimator is derived by

a.	minimizing the sum of squared prediction mistakes.
b.	setting the sum of squared errors equal to zero.
c.	minimizing the absolute difference of the residuals.
d.	forcing the smallest distance between the actual and fitted values.

11)	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.

12) The overall regression F-statistic tests the null hypothesis that

a. all slope coefficients are zero.

b. all slope coefficients and the intercept are zero.

c. the intercept in the regression and at least one, but not all, of the slope coefficients is zero.

d. the slope coefficient of the variable of interest is zero, but that the other slope coefficients are not.

13) The interpretation of the slope coefficient in the model $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:

a. a 1% change in *X* is associated with a β_1% change in *Y*.

b. a 1% change in *X* is associated with a change in *Y* of 0.01 β_1.

c. a change in *X* by one unit is associated with a 100 β_1% change in *Y*.

d. a change in *X* by one unit is associated with a β_1 change in *Y*.

14) An example of a quadratic regression model is

a. $Y_i = \beta_0 + \beta_1 X + \beta_2 Y^2 + u_i$.

b. $Y_i = \beta_0 + \beta_1 \ln(X) + u_i$.

c. $Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + u_i$.

d. $Y_i^2 = \beta_0 + \beta_1 X + u_i$.

15) A nonlinear function

a. makes little sense, because variables in the real world are related linearly.

b. can be adequately described by a straight line between the dependent variable and one of the explanatory variables.

c. is a concept that only applies to the case of a single or two explanatory variables since you cannot draw a line in four dimensions.

d. is a function with a slope that is not constant.

16) The OLS residuals

a. can be calculated using the errors from the regression function.

b. can be calculated by subtracting the fitted 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.

Part B - Short Answer – 4 marks each

1) Describe some reasons why it is important to look at a scatterplot of the data.

2) What is the problem with assuming that the error terms (the u_i's) are homoskedastic, when they are actually heteroskedastic?

3) Explain why using unadjusted-R-square (R^2) in the multiple regression model is a bad idea. Why is it better to use adjusted-R-square ($\bar{R}^2$) in the multiple regression model?

4) Suppose you want to test $H_0: \beta_2 = 0, \beta_3 = 0$. Can you use the *t*-test? Why or why not?

5) Use the following data for this question:

$$Y = \{1,6,7\} \qquad X = \{2,4,6\}$$

The population model is:

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

What are the OLS estimates for β_0 and β_1?

6) Explain the difference between the random error term (u_i) and the OLS residual ($\hat{u}_i$).

7) Explain the difference between a restricted and an unrestricted model. Provide a simple example.

8) Consider the estimated regression model:

$$\widehat{TestScore} = 607.3 + 3.85 Income - 0.0423 Income^2$$

What is the predicted change in *TestScore* due to a change in *Income*?

9) Consider the following population model:

$$ahe = \beta_0 + \beta_1 bach + \beta_2 female + \beta_3 bach \times female + u\ ,$$

where *ahe* is average hourly earnings, *female* is a dummy variable which equals 1 if the individual is female, and *bach* is a dummy variable which equals 1 if the individual has a bachelor's degree (BA). The *ahe* among various groups are:

ahe for men without a BA	= 14.88
ahe for women without a BA	= 11.92
ahe for men with a BA	= 21.98
ahe for women with a BA	= 18.47

What are the OLS estimates for β_0, β_1, β_2, and β_3?

10) In class, the effect of an additional fireplace on house price was estimated. The following two regression outputs have been reproduced from class:

```
summary(lm(Price ~ Fireplaces))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  171.824      3.234   53.13   <2e-16 ***
Fireplaces    66.699      3.947   16.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.21 on 1726 degrees of freedom
Multiple R-squared: 0.142,      Adjusted R-squared: 0.1415
F-statistic: 285.6 on 1 and 1726 DF,  p-value: < 2.2e-16

--------------------------------------------------------------

summary(lm(Price ~ Fireplaces + Living.Area))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.730146   5.007563   2.942  0.00331 **
Fireplaces   8.962440   3.389656   2.644  0.00827 **
Living.Area  0.109313   0.003041  35.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.98 on 1725 degrees of freedom
Multiple R-squared: 0.5095,     Adjusted R-squared: 0.5089
F-statistic: 895.9 on 2 and 1725 DF,  p-value: < 2.2e-16
```

Why is the estimated effect of an additional fireplace so different between the two regressions? Explain in detail.

11) Explain why it might be important to capture non-linear effects in a regression model. Explain what a non-linear effect means. Describe different ways of capturing non-linear effects.

Part C – Long Answer – each part is worth 4 marks

This question uses the Current Population Survey (CPS) dataset from assignment #3 and from the class lectures. Several estimated models have been reported below.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable							
	AHE	***ln(AHE)***	***ln(AHE)***	***ln(AHE)***	***ln(AHE)***	***ln(AHE)***	***ln(AHE)***	***ln(AHE)***
Age	0.439** (0.031)	0.024** (0.002)		0.147** (0.042)	0.146** (0.042)	0.191** (0.054)	0.117* (0.057)	0.160* (0.064)
Age^2				-0.002** (0.001)	-0.002** (0.001)	-0.003** (0.001)	-0.002 (0.001)	-0.002* (0.001)
ln(*Age*)			0.725** (0.052)					
Female×*Age*						-0.097 (0.084)		-0.123 (0.085)
Female×Age^2						0.002 (0.001)		0.002 (0.001)
Bachelor×*Age*							0.064 (0.083)	0.091 (0.084)
Bachelor×Age^2							-0.001 (0.001)	-0.001 (0.001)
Female	-3.158** (0.180)	-0.181** (0.011)	-0.180** (0.011)	-0.180** (0.011)	-0.210** (0.014)	1.358 (1.238)	-0.209** (0.014)	1.764 (1.251)
Bachelor	6.865** (0.178)	0.405** (0.010)	0.405** (0.010)	0.405** (0.010)	0.378** (0.014)	0.377** (0.014)	-0.770 (1.223)	-1.186 (1.236)
Female×*Bachelor*					0.064** (0.021)	0.063** (0.021)	0.067** (0.021)	0.066** (0.021)
Intercept	1.884* (0.920)	1.857** (0.054)	0.128 (0.177)	0.059 (0.611)	0.078 (0.610)	-0.633 (0.799)	0.604 (0.831)	-0.095 (0.939)
R^2	0.1900	0.1924	0.1927	0.1933	0.1942	0.1950	0.1957	0.1968
$\bar{R}^2$	0.1897	0.1921	0.1924	0.1929	0.1937	0.1943	0.1949	0.1959

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

You may also need the following critical values for F-statistics in large samples:

q	5% critical value
1	3.84
2	3.00
3	2.60
4	2.37
5	2.21

a) In all of the models, $R^2 > \bar{R}^2$. Why is this?

b) Can R^2 be used to compare models (1) and (2)? Why or why not?

c) What is the interpretation of the intercept in model (3)? Since the intercept appears to be statistically insignificant, should it be dropped from the model?

d) Interpret the estimated coefficient of 0.725 (on *ln(Age)*) in model (3).

e) Using model (4), what is the predicted *ln(AHE)* for a 25 year old male with a bachelor's degree?

f) Does *Age* appear to have a non-linear effect on *ln(AHE)*? Use a formal hypothesis test in your answer.

g) Test the hypothesis that the effect of having a bachelor's degree on *ln(AHE)* is the same for women as it is for men.

h) Test the hypothesis that the effect of *Age* on *ln(AHE)* is the same for women as it is for men.

i) Using model (8), what is the estimated effect on *ln(AHE)* of obtaining a bachelor's degree, for 25 year old women? For 25 year old men?

j) Approximately how much more valuable is a bachelor's degree for women than for men?

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	$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots, n$
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	$E(u\|X = x) = 0$
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	$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)}$

Standard Normal Probabilities

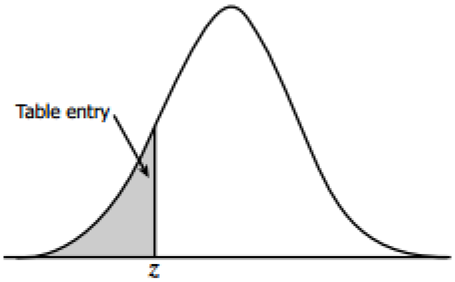

Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641