# Econ 3180 – Final Exam, April 15th 2013

## Ryan Godwin

You may use a calculator. Answer all questions in the answer book provided. The exam is 3 hours long and consists of 300 marks.

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

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

## Easy Question

[28 marks]

**1**)

Consider a random variable, $Y$:

$$Y = 1, \text{ with probability } 0.5$$

$$Y = 2, \text{ with probability } 0.5$$

What is the expected value of $Y$? (What is $E(Y)$?)

## Part A – Multiple Choice

[104 marks – 4 marks each]

**2**) The sample average is

a. a random variable.
b. a constant.
c. don't pick this.
d. don't pick this either.

**3**) Binary variables (dummy variables)

a. are generally used to control for outliers in your sample.
b. can take on more than two values.
c. exclude certain individuals from your sample.
d. can take on only two values.

**4**) The regression $R^2$ is a measure of

a. whether or not $X$ causes $Y$.
b. the goodness of fit of your regression line.
c. whether or not $ESS > TSS$.
d. the square of the determinant of $R$.

**5**) A type I error is

a. always the same as (1-type II) error.
b. the error you make when rejecting the null hypothesis when it is true.
c. the error you make when rejecting the alternative hypothesis when it is true.
d. always 5%.

**6**) Degrees of freedom

a. in the context of the sample variance formula means that estimating the mean uses up some of the information in the data.
b. can correct for omitted variable bias.
c. are (*n*-2) when replacing the population mean by the sample mean.
d. ensure that $s_Y^2 = \sigma_Y^2$.

**7**) In the simple linear regression model, the regression intercept

a. indicates by how many percent *Y* increases, given a one percent increase in *X*.
b. when multiplied with the explanatory variable will give you the predicted *Y*.
c. indicates by how many units *Y* increases, given a one unit increase in *X*.
d. represents the expected value of *Y* when *X* is zero.

**8**) The OLS estimator is unbiased:

a. if least squares assumption #1 holds.
b. if least squares assumption #3 holds.
c. if the total sum of squares is minimized.
d. always.

**9**) Finding a large value of the *p*-value (e.g. more than 10%)

a. indicates evidence in favor of the null hypothesis.
b. implies that the *t*-statistic is less than 1.96.
c. indicates evidence against the null hypothesis.
d. will only happen roughly one in twenty samples.

**10**) In the multiple regression model, the adjusted *R*-square, $\bar{R}^2$

a. cannot be negative.
b. will never be greater than the unadjusted $R^2$.
c. equals the square of the correlation coefficient *r*.
d. cannot decrease when an additional explanatory variable is added.

**11**) Under imperfect multicollinearity

a. the OLS estimator cannot be computed.
b. two or more of the regressors are highly correlated.
c. the OLS estimator is biased even in samples of $n > 100$.
d. the error terms are highly, but not perfectly, correlated.

**12**) 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.

**13)** You have to worry about perfect multicollinearity in the multiple regression model because

a. many economic variables are perfectly correlated.
b. the OLS estimator is no longer consistent.
c. the OLS estimator cannot be computed in this situation.
d. in real life, economic variables change together all the time.

**14)** 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.

**15**) The OLS residuals in the multiple regression model

a. cannot be calculated because there is more than one explanatory variable.
b. can be calculated by subtracting the fitted values from the actual values.
c. are zero because the predicted values are another name for forecasted values.
d. are typically the same as the population regression function errors.

**16)** One of the least squares assumptions in the multiple regression model is that you have random variables which are "i.i.d." This stands for

a. initially indeterminate differences.
b. irregularly integrated dichotomies.
c. identically initiated deltas.
d. independently and identically distributed.

**17)** Omitted variable bias

a. will always be present as long as the regression $R^2 < 1$
b. is always there but is negligible in almost all economic examples
c. exists if the omitted variable is correlated with the included regressor but is not a determinant of the dependent variable
d. exists if the omitted variable is correlated with the included regressor and is a determinant of the dependent variable

**18**) In multiple regression, the $R^2$ increases whenever a regressor is

a. added unless the estimated coefficient on the added regressor is exactly zero.
b. added.
c. added unless there is heteroskedasticity.
d. greater than 1.96 in absolute value.

**19**) In the multiple regression model, the *t*-statistic for testing that the slope is significantly different from zero is calculated

a. by dividing the estimate by its standard error.
b. from the square root of the *F*-statistic.
c. by multiplying the *p*-value by 1.96.
d. using the adjusted $R^2$ and the confidence interval.

**20**) Let $R^2_{unrestricted}$ and $R^2_{restricted}$ be 0.4366 and 0.4149 respectively. The difference between the unrestricted and the restricted model is that you have imposed two restrictions. There are 420 observations. The *F*-statistic in this case is

a. 4.61.
b. 8.01.
c. 10.34.
d. 7.71.

**21**) A 95% confidence set for two or more coefficients is a set that contains

a. the sample values of these coefficients in 95% of randomly drawn samples.
b. integer values only.
c. the same values as the 95% confidence intervals constructed for the coefficients.
d. the population values of these coefficients in 95% of randomly drawn samples.

**22**) When there are two coefficients, the resulting confidence sets are

a. rectangles.
b. ellipses.
c. squares.
d. trapezoids.

**23**) All of the following are true, with the exception of one condition:

a. a high $R^2$ or $\overline{R}^2$ does not mean that the regressors are a true cause of the dependent variable.
b. a high $R^2$ or $\overline{R}^2$ does not mean that there is no omitted variable bias.
c. a high $R^2$ or $\overline{R}^2$ always means that an added variable is statistically significant.
d. a high $R^2$ or $\overline{R}^2$ does not necessarily mean that you have the most appropriate set of regressors.

**24**) If the estimates of the coefficients of interest change substantially across specifications,

a. then this can be expected from sample variation.
b. then you should change the scale of the variables to make the changes appear to be smaller.
c. then this often provides evidence that the original specification had omitted variable bias.
d. then choose the specification for which your coefficient of interest is most significant.

**25**) 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.

**26)** 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. $Y_i^2 = \beta_0 + \beta_1 X + u_i$ .

**27)** In the model $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + u_i$, the expected effect $\frac{\Delta Y}{\Delta X_1}$ is

a. $\beta_1 + \beta_3 X_2$ .

b. $\beta_1$ .

c. $\beta_1 + \beta_3$ .

d. $\beta_1 + \beta_3 X_1$ .

## Part B – Short Answer

[80 marks – 8 marks each]

**28**) Consider a sample of three observations collected from the random variable, $Y$:

$$Y = \{2,4,6\}$$

Estimate the variance of $Y$.

**29**) Briefly explain how the OLS estimator $\hat{\beta}_1$, in a single regressor model, is derived.

**30**) Consider the following estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$, where *STR* is a variable that describes the student-teacher ratio in a classroom. What is the predicted test score in a classroom of size 30?

**31**) What problems arise when there is heteroskedasticity?

**32**) Briefly explain what it means for an estimator to be unbiased and consistent.

**33**) Suppose that a researcher, using wage data on 180 randomly selected workers *with* a university education and 200 workers *without* a university, estimates the OLS regression,

$$\widehat{Wage} = \underset{(1.45)}{12.34} + \underset{(4.22)}{6.52} \times UNI, \qquad R^2 = 0.28$$

Where *Wage* is measured in $/hour and *UNI* is a "dummy" variable that is equal to 1 if the person has a university education and 0 if the person does not have a university education.

Conduct a formal hypothesis test to determine whether or not obtaining a university education will affect hourly wage.

**34**) Consider the population regression model:

$$wage_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + u_i$$

where *male* is a dummy variable that takes on the value "1" if the individual is male, and "0" if the individual is female. Consider the variable *female*, which takes on the value "1" if the individual is female and "0" if the individual is male. What is the problem with adding *female* to the model?

**35**) Why should you use adjusted R-square ($\bar{R}^2$) instead of the unadjusted R-square ($R^2$) in the multiple regression model?

**36**) Suppose you have estimated a model with multiple regressors, and two of them are individually statistically insignificant (based on the *t*-statistics). Can you test the joint hypothesis that both coefficients are equal to zero using *t*-tests? Why or why not?

**37**) Provide an example of a non-linear relationship between two variables. Why is it important to try to capture non-linear effects in our regressions?

**38**) Consider the formula:

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)}$$

Describe intuitively why a large value for "*F*" indicates that we should reject the null hypothesis.

**Part C – Long Answer**

**[88 marks total – 8 marks for each part]**

**39**) This question uses the same **CollegeDistance** data that was used in assignment #3 and #4.

These data are taken from the *HighSchool and Beyond* survey conducted by the Department of Education in 1980, with a follow-up in 1986. The survey included students from approximately 1100 high schools.

**Series in Data Set**

| **Name** | **Description** |
|---|---|
| *ed* | Years of Education Completed (See below) |
| *female* | 1 = Female/0 = Male |
| *black* | 1 = Black/0 = Not-Black |
| *hispanic* | 1 = Hispanic/0 = Not-Hispanic |
| *bytest* | Base Year Composite Test Score. (These are achievement tests given to high school seniors in the sample) |
| *dadcoll* | 1 = Father is a College Graduate/ 0 = Father is not a College Graduate |
| *momcoll* | 1 = Mother is a College Graduate/ 0 = Mother is not a College Graduate |
| *incomehi* | 1 = Family Income > $25,000 per year/ 0 = Income ≤ $25,000 per year. |
| *ownhome* | 1= Family Owns Home / 0 = Family Does not Own Home |
| *cue80* | County Unempolyment rate in 1980 |
| *stwmfg80* | State Hourly Wage in Manufacturing in 1980 |
| *dist* | Distance from 4yr College in 10's of miles |
| *tuition* | Avg. State 4yr College Tuition in $1000's |

Years of Education: Rouse (the author) computed years of education by assigning 12 years to all members of the senior class. Each additional year of secondary education counted as a one year. Student’s with vocational degrees were assigned 13 years, AA degrees were assigned 14 years, BA degrees were assigned 16 years, those with some graduate education were assigned 17 years, and those with a graduate degree were assigned 18 years.

Below is a table of estimated models which you should use for parts (a) – (k)

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| **Regressor** | ***ed*** | ***ed*** | ***ed*** | ***ed*** | |
| *dist* | -0.037**<br>(0.013) | -0.081**<br>(0.026) | -0.081**<br>(0.256) | -0.110**<br>(0.029) | -0.113**<br>(0.025) |
| $dist^2$ | | 0.005*<br>(0.002) | 0.005*<br>(0.002) | 0.007**<br>(0.003) | 0.007**<br>(0.002) |
| *tuition* | -0.191<br>(0.101) | -0.193<br>(0.101) | -0.194<br>(0.101) | -0.210*<br>(0.101) | -0.290**<br>(0.097) |
| *female* | 0.143**<br>(0.050) | 0.143**<br>(0.050) | 0.141**<br>(0.050) | 0.142**<br>(0.050) | 0.133**<br>(0.051) |
| *black* | 0.351**<br>(0.071) | 0.334**<br>(0.072) | 0.331**<br>(0.072) | 0.333**<br>(0.072) | |
| *hispanic* | 0.362**<br>(0.077) | 0.333**<br>(0.079) | 0.330**<br>(0.079) | 0.323**<br>(0.079) | |
| *bytest* | 0.093**<br>(0.003) | 0.093**<br>(0.003) | 0.093**<br>(0.003) | 0.093**<br>(0.003) | 0.087**<br>(0.003) |
| *incomehi* | 0.372**<br>(0.061) | 0.370**<br>(0.061) | 0.362**<br>(0.061) | 0.217*<br>(0.091) | 0.334**<br>(0.061) |
| *ownhome* | 0.139*<br>(0.067) | 0.143*<br>(0.067) | 0.141*<br>(0.067) | 0.144*<br>(0.067) | 0.099<br>(0.067) |
| *dadcoll* | 0.571**<br>(0.074) | 0.561**<br>(0.074) | 0.654**<br>(0.084) | 0.663**<br>(0.084) | 0.642**<br>(0.085) |
| *momcoll* | 0.378**<br>(0.082) | 0.378**<br>(0.081) | 0.569**<br>(0.117) | 0.568**<br>(0.117) | 0.591**<br>(0.118) |
| *dadcoll* × *momcoll* | | | -0.367*<br>(0.161) | -0.356*<br>(0.162) | -0.389*<br>(0.162) |
| *cue80* | 0.029**<br>(0.010) | 0.026**<br>(0.010) | 0.026**<br>(0.010) | 0.026**<br>(0.010) | 0.030**<br>(0.010) |
| *stwmfg80* | -0.043*<br>(0.020) | -0.043*<br>(0.020) | -0.042*<br>(0.020) | -0.042*<br>(0.020) | -0.052**<br>(0.020) |
| *incomehi* × *dist* | | | | 0.124<br>(0.064) | |
| *incomehi* × $dist^2$ | | | | -0.009<br>(0.007) | |
| *intercept* | 8.921**<br>(0.252) | 9.012**<br>(0.256) | 9.002**<br>(0.256) | 9.042**<br>(0.256) | 9.627**<br>(0.229) |
| | | | | | |
| F-stat. (overall) | 124.8<br>(0.000) | 115.6<br>(0.000) | 107.8<br>(0.000) | 94.73<br>(0.000) | 122.4<br>(0.000) |
| $R^2$ | 0.2836 | 0.2844 | 0.2854 | 0.2863 | 0.2796 |
| $\bar{R}^2$ | 0.2813 | 0.2819 | 0.2827 | 0.2832 | 0.2774 |

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

(a) Suppose we are only concerned with the effect of distance on education. Why bother including all the other variables?

(b) Using model (2), if *dist* increases from 4 to 5, how are years of education expected to change?

(c) What is the interaction term *dadcoll* × *momcoll*, measuring? Why has the inclusion of *dadcoll* × *momcoll* caused the estimated coefficients on *dadcoll* and *momcoll* to change from model (2) to model (3)?

(d) Do you think that *dist* has a non-linear or a linear effect on education?

(e) Using model (4), what is the difference between the average number of years of education obtained by women, versus the average number of years of education obtained by men?

(f) Does ethnicity affect the number of years of education obtained?

(g) What does the interaction term, *incomehi* × *dist*, measure? Why include this in the regression?

(h) Why does the $R^2$ increase when going from model (3) to model (4), but the $\bar{R}^2$ decreases?

(i) State the null and alternative hypothesis associated with the "F-stat. (overall)".

(j) Suppose you added another variable to the above regression. The p-value associated with the *t*-statistic on the estimated coefficient is 0.03. How many stars (*) would you put next to the estimated coefficient in order to make it compatible with the information in the above table?

(k) Using model (2), construct a 95% confidence interval for the variable *cue80*. Interpret this interval.

**END.**

# Econ 3180 - Final Formula Sheet

| | |
|---|---|
| expected value of $Y$ (mean of $Y$) | $\mu_Y$ |
| 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 $\mu_Y$ (for large $n$, and when $\sigma_Y^2$ is *known*) | $t = \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}} \sim N(0,1)$ |
| sample variance (estimator for $\sigma_Y^2$) | $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ |
| 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 $\mu_Y$ (for large $n$, and when $\sigma_Y^2$ is *unknown*) | $t = \frac{\bar{Y}^{act} - \mu_{Y,0}}{s_{\bar{Y}}} \sim N(0,1)$ |
| 95% confidence interval for $\mu_Y$ (for large $n$) | $conf.int. = \bar{Y} \pm 1.96 \times s_{\bar{Y}}$ |
| population linear regression model with one regressor | $Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \dots, n$ |
| OLS estimator of the slope ($\beta_1$) | $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$ |
| OLS estimator of the intercept ($\beta_0$) | $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ |
| 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 | $(X_i, Y_i), i = 1, \ldots, n$, 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 $\beta_1$ | $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)}$ |
| 95% confidence interval for $\beta_1$ (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 | $F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)}$ |