THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN
School of Management and Economics

ECONOMICS 3121
Introductory Econometrics

Spring term 2021 **FINAL EXAM** Yi Ding & Zehao Li

DATE: **Thursday May 20, 2021**

TIME: **180 minutes; 4:00 p.m. – 7:00 p.m.**

INSTRUCTIONS: The exam consists of **FOUR (4)** questions. Students are required to answer **ALL FOUR (4)** questions.

Answer all questions in the exam booklets provided. Be sure your ***name*** and ***student number*** are printed clearly on the front of all exam booklets used.

Do not write answers to questions on the front page of the first exam booklet.

Please label clearly each of your answers in the exam booklets with the appropriate number and letter.

Please write legibly.

This exam is **CONFIDENTIAL**. This question paper must be submitted in its entirety with your answer booklet(s); otherwise your exam will not be marked.

A formula sheet and tables of critical values of the t distribution and F distribution are given on the last pages of the exam.

MARKING: The marks for each question are indicated in parentheses immediately above each question. **Total marks** for the exam **equal 170**.

GOOD LUCK!

Question 1: Multiple Choice (50 points, 2 points each)
Select the **BEST** response to each of the following questions.

1. Which of the following Gauss-Markov assumptions is violated by the linear probability model?
   a) The assumption of constant variance of the error term.
   b) The assumption of zero conditional mean of the error term.
   c) The assumption of no exact linear relationship among independent variables.
   d) The assumption that none of the independent variables are constants.

2. If $\widehat{\beta_j}$, an unbiased estimator of β_j, is consistent, then the:
   a) distribution of $\hat{\beta}_j$ becomes more and more loosely distributed around β_j as the sample size grows.
   b) distribution of $\hat{\beta}_j$becomes more and more tightly distributed around β_j as the sample size grows.
   c) distribution of $\hat{\beta}_j$ tends toward a standard normal distribution as the sample size grows.
   d) distribution of $\hat{\beta}_j$ remains unaffected as the sample size grows.

3. Which of the following assumptions is required to obtain a first-differenced estimator in a two-period panel data analysis?
   a) The independent variable does not change over time for any cross-sectional unit.
   b) The independent variable changes by the same amount in each time period.
   c) The variance of the error term in the regression model is not constant.
   d) The idiosyncratic error at each time period is uncorrelated with the independent variables in both time periods.

4. What will you conclude about a regression model if the Breusch-Pagan test results in a small p-value?
   a) The model contains homoskedasticty.
   b) The model contains heteroskedasticty.
   c) The model contains dummy variables.
   d) The model omits some important explanatory factors.

5. In a multiple linear regression model with heteroskedasticity, if you use Stata to obtain the heteroskedasticity-robust standard error, you will
   a) have estimates that are unbiased, but parameter standard errors that are too small as standard errors are biased under heteroskedasticity
   b) have parameter estimates that are unbiased, but standard errors with unknown bias
   c) have the best estimates you can come up with, because OLS is the Best Linear Unbiased Estimator (BLUE)
   d) have unbiased parameter estimates and unbiased standard errors, but OLS is not the best estimator to use, as it is no longer the minimum variance estimator

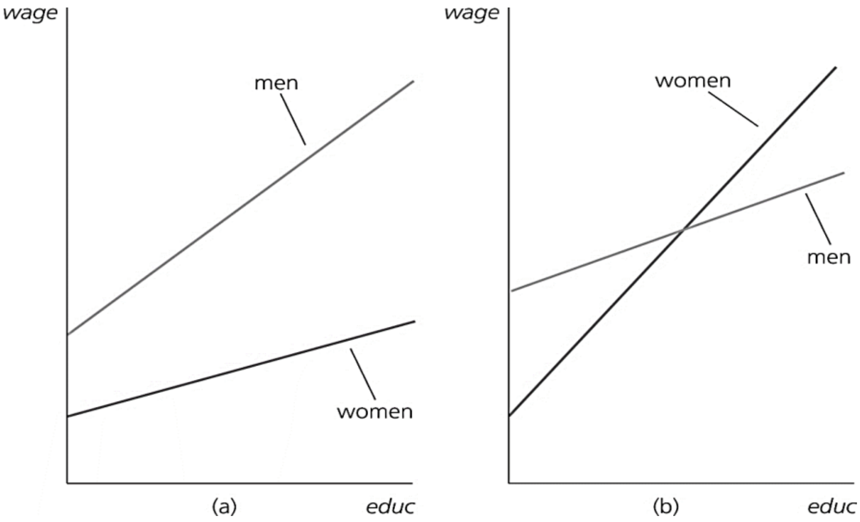
6. In a first differenced equation, which of the following assumptions is needed for the usual standard errors to be valid when differencing with more than two time periods?
   a) The regression model exhibits heteroskedasticty.
   b) The differenced idiosyncratic error or Δu_{it} is uncorrelated over time.
   c) The unobserved factors affecting the dependent variable are time-constant.
   d) The regression model includes a lagged independent variable.

7. Which of the following types of variables cannot be included in a fixed effects model?
   a) Dummy variable
   b) Discrete dependent variable
   c) Time-varying independent variable
   d) Time-constant independent variable

8. A researcher has analyzed the differences in the effect of education on wages for men and women using the following regression model:

$$wage = \beta_0 + \beta_1 educ + \delta_0 female + \delta_1 \, female * educ + u$$

where $wage$ is log of hourly wages, $educ$ is the number of years of education, and $female$ is a dummy variable that is 1 for women. The following graphs show the estimated regression results from two different populations, (a) and (b).



What is the sign of δ_0 and δ_1 in the two populations?
   a) δ_0 is negative in (a) and negative in (b), δ_1 is positive in (a) and positive in (b)
   b) δ_0 is negative in (a) and negative in (b), δ_1 is negative in (a) but positive in (b)
   c) δ_0 is negative in (a) but positive in (b), δ_1 is negative in (a) and negative in (b)
   d) δ_0 is positive in (a) and positive in (b), δ_1 is negative in (a) but positive in (b)

9. Refer to the following finite distributed lag model.

$y_t = \alpha_0 + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + u_t$

$\beta_0 + \beta_1 + \beta_2 + \beta_3$ represents:
   a) the short-run change in y given a temporary increase in x.
   b) the short-run change in y given a permanent increase in x.
   c) the long-run change in y given a permanent increase in x.
   d) the long-run change in y given a temporary increase in x.

10. Which of the following is true of the OLS *t* statistics?
   a) The heteroskedasticity-robust *t* statistics are justified only if the sample size is large.
   b) The heteroskedasticty-robust *t* statistics are justified only if the sample size is small.
   c) The usual t statistics do not have exact *t* distributions if the sample size is large.
   d) In the presence of homoscedasticity, the usual *t* statistics do not have exact *t* distributions if the sample size is small.

11. In the following regression equation, y is a binary variable:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

In this case, the estimated slope coefficient, $\hat{\beta}_1$ measures _____.

a) the predicted change in the value of y when x_1 increases by one unit, everything else remaining constant
b) the predicted change in the value of y when x_1 increase by one percentage point, everything else remaining constant
c) the predicted change in the probability of success when x_1 increase by one percentage point, everything else remaining constant
d) the predicted change in the probability of success when x_1 increases by one unit, everything else remaining constant

12. Which of the following is a reason for using independently pooled cross sections?

a) To obtain data on different cross sectional units
b) To increase the sample size
c) To select a sample based on the dependent variable
d) To select a sample based on the independent variable

13. A Chow test _____.

a) is used to test the presence of heteroskedasticty in a regression model.
b) is used to determine how multiple regression differs across two groups.
c) cannot detect changes in the slope coefficients of dependent variables over time.
d) cannot be computed for more than two time periods.

14. Idiosyncratic error is the error that occurs due to _____.

a) incorrect measurement of an economic variable
b) unobserved factors that affect the dependent variable and change over time
c) unobserved factors that affect the dependent variable and do not change over time
d) correlation between the independent variables

15. Which of the following assumptions is required for obtaining unbiased fixed effect estimators?

a) The errors are heteroskedastic.
b) The errors are serially correlated.
c) The independent variables are strictly exogenous.
d) The unobserved effect is correlated with the independent variables.

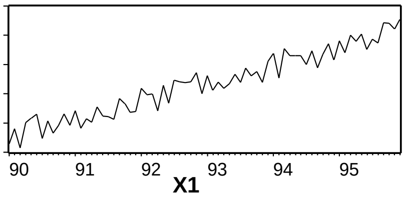
16. A pooled OLS estimator that is based on the time-demeaned variables is called the _____.

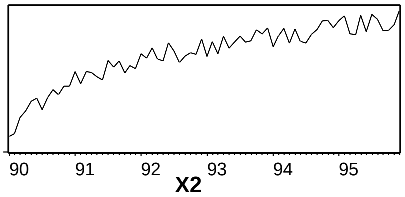
a) random effects estimator
b) fixed effects estimator
c) least absolute deviations estimator
d) instrumental variable estimator

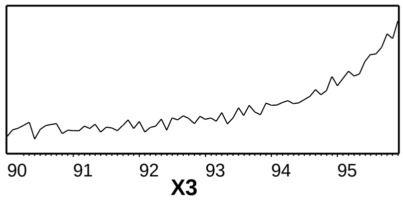
17. What should be the degrees of freedom (df) for fixed effects estimation if the data set includes 'N' cross sectional units over 'T' time periods and the regression model has 'k' independent variables?

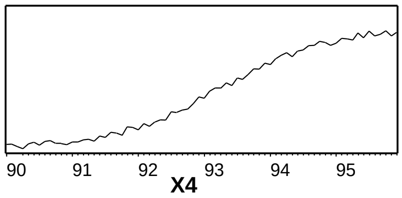
a) N-kT
b) NT-k
c) NT-N-k $N(T-1)-k$
d) N-T-k

18. The time series variables depicted below contain various types of time trend components.









which of the following time series contains an exponential time trend?

a) X1
b) X2
c) X3
d) X4

19. Which of the following is true of heteroskedasticity?

a) Heteroskedasticty causes inconsistency in the Ordinary Least Squares estimators.
b) Population R^2 is affected by the presence of heteroskedasticty.
c) The Ordinary Least Square estimators are not the best linear unbiased estimators if heteroskedasticity is present.
d) It is not possible to obtain F statistics that are robust to heteroskedasticity of an unknown form.

20. In an OLS multiple regression, if you divide the explained variable y by 10, which of the following will happen?

a) The estimated intercept parameter will be reduced by 10 times
b) the estimated slope parameter will be reduced by 10 units
c) the Total Sum of Squares for the regression will be reduced by $\sqrt{10}$ times
d) R-squared for the regression will be reduced by 10 times

21. Which of the following statements is true when the dependent variable, $y > 0$

a) Taking log of a variable often expands its range.
b) Models using $\log(y)$ as the dependent variable may satisfy CLM assumptions more closely than models using the level of y.
c) Taking log of variables make OLS estimates more sensitive to extreme values.
d) Taking logarithmic form of variables make the slope coefficients more responsive to rescaling.

22. Consider the following regression model: $y = \beta_0 + \beta_1 x + u$. Which of the following is a property of Ordinary Least Square (OLS) estimates of this model and their associated statistics?

a) The sum, and therefore the sample average of the OLS residuals, is positive.
b) The sum of the OLS residuals is negative.
c) The sample covariance between the regressors and the OLS residuals is positive.
d) The point $(\bar{x}, \bar{y})$ always lies on the OLS regression line.

23. First-differenced estimation in a panel data analysis is subject to serious biases if _____.

a) key explanatory variables vary significantly over time
b) the explanatory variables do not change by the same unit in each time period
c) one or more of the explanatory variables are measured incorrectly
d) the regression model exhibits homoscedasticity

24. Weighted least squares estimation is used only when _____.

a) the dependent variable in a regression model is binary
b) the independent variables in a regression model are correlated
c) the error term in a regression model has a constant variance
d) the form of the error variances is known

25. You have just read a paper which claims that there is strong evidence that x affects y negatively. The researcher found that the estimated coefficient for x is significantly negative at the 5% level. You suspect that there is omitted variable bias. Which of the following statements is correct?

a) If there is omitted variable bias, and the bias is positive, then the researcher's conclusions are invalid.
b) If there is omitted variable bias, and the bias is positive, then the researcher's conclusions are still valid.
c) If there is omitted variable bias, and the bias is zero, then the researcher's conclusions are invalid.
d) If there is omitted variable bias, and the bias is negative, then the researcher's conclusions are still valid.

Question 2 (40 points)
Please answer all questions below.

(6 points)
(a) Briefly explain what strict exogeneity and contemporaneous exogeneity are in time series study.

x and u are uncorrelated

(9 points)
(b) Suppose the unobserved effects panel data model is as follows:

RE: gk bljr

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \beta_3 x_{it3} + a_i + u_{it}, \quad t = 1, 2, \dots, T$$

where a_i is the individual unobserved effect, u_{it} is idiosyncratic error. Write down the first differenced (FD) equation, the fixed effect (FE) equation, and the random effect (RE) equation.

FD: delta(yit) = b1 delta(xit1) + b2 delta(xit2) + yg 3 + delta(uit)

FE: smua

(8 points)
(c) Briefly explain why the fitted value from the first stage of 2SLS can be used as an instrumental variable to consistently estimate the slope coefficient on the endogenous variable.

IV can eliminate bias

(6 points)
(d) In a simple static model $y_t = \beta_0 + \beta_1 x_t + u_t$, $t = 1, 2, \dots$, both x_t and y_t have exhibited an increasing trend, which leads to the spurious regression problem. How to eliminate this problem? and how to compute an appropriate R-squared for the model?

kt ipan g msk

(11 points)
(e) Consider the OLS regression function $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. Explain how to use partialling out interpretation of multiple regression to find the formula of OLS estimator $\hat{\beta}_1$. Derive the formula of $Var(\hat{\beta}_1)$. Be specific.

regress x1 on other indepedent var and get residual r, regress y on x12. check di mid sol

Question 3 (40 marks)
You are investigating the relationship between the birth weights of newborn babies and four of their determinants: mother's average daily cigarette consumption during pregnancy; the number of prenatal visits made by the mother to a physician or medical facility during pregnancy; the mother's age; and the mother's race. You have sample data for 1656 babies born during a given year on the following variables

$bwght_i$ = birth weight of the baby born to the i-th mother, measured in *hundreds of grams*;

$cigs_i$ = average number of cigarettes per day smoked by the i-th mother during pregnancy, measured in *cigarettes per day*;

$npvis_i$ = number of prenatal visits to a doctor or medical facility made by the i-th mother during pregnancy;

age_i = age of the i-th mother, measured in *years*;

$white_i$ = an indicator variable defined such that $white_i = 1$ if the i-th mother is white, and $white_i = 0$ the i-th mother is non-white;

Using the given sample data on 1656 newborn babies, your trusty research assistant has estimated regression equation (1) and obtained the following estimation results (with estimated standard errors given in parentheses below the coefficient estimates):

$$bwght_i = \beta_0 + \beta_1 cigs_i + \beta_2 npvis_i + \beta_3 npvis_i^2 + \beta_4 age_i + \beta_5 age_i^2$$

$$+\beta_6 white_i + \beta_7 white_i age_i + u_i \qquad (1)$$

$$\begin{array}{llll} \hat{\beta}_0 = 16.076 & \hat{\beta}_1 = -0.1010 & \hat{\beta}_2 = 0.2971 & \hat{\beta}_3 = -0.006045 \\ \quad (4.618) & \quad (0.03338) & \quad (0.1061) & \quad (0.003429) \\ \hat{\beta}_4 = 0.8083 & \hat{\beta}_5 = -0.0101 & \hat{\beta}_6 = 6.8850 & \hat{\beta}_7 = -0.2039 \\ \quad (0.2825) & \quad (0.004534) & \quad (2.587) & \quad (0.08734) \end{array}$$

$$SSR = 53099.886 \qquad SST = 54516.777 \qquad N = 1656$$

F-test

(7 marks)
(a) Use the estimation results for regression equation (1) to test the joint significance of the slope coefficient estimates $\hat{\beta}_j$ $(j = 1, \ldots, 7)$in regression equation (1). Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null and alternative hypotheses, and show how you calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level. h0 smua beta = 0 | h1: ho is not true

df : 1656-7-1

(3 marks)
(b) Interpret the slope coefficient estimate $\hat{\beta}_1$ in sample regression equation (1). That is, explain in words what the numerical value of the slope coefficient estimate $\hat{\beta}_1$ means.

Assume all other estimator are fixed. 1 unit increase in cigar, decrease 10 gram on baby weight.

(9 marks)
(c) Use the estimation results for regression equation (1) to compute the two-sided 95 percent confidence interval for the slope coefficient β_1. Use the computed confidence interval for β_1 to perform a test of the proposition that mothers' cigarette consumption during pregnancy ($cigs_i$) has no effect on the birth weight of their babies. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null hypothesis H_0 and the alternative hypothesis H_1. State the inference you would draw from the test at the 5 percent significance level. H0: b1 = 0

t-student test

(7 marks)
(d) Compute a test of the proposition that the number of prenatal visits by the mother to a doctor or medical facility has no effect on the birth weight of newborn babies. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0. OLS estimation of this restricted regression equation yields a Residual Sum-of-Squares value of **SSR = 53542.009**. Use this information, together with the results from OLS estimation of equation (1), to calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

F test H0: b2 b3 = 0

(7 marks)
(e) Use the estimation results for regression equation (1) to test the proposition that the marginal effect of mother's age (age_i) on the birth weight of newborn babies is negatively related to age_i. Perform the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$). State the null hypothesis H_0 and the alternative hypothesis H_1 implied by this proposition. Show how you calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level.

(7 marks)
(f) Compute a test of the proposition that the marginal effect of mother's age (age_i) on birth weight is zero for babies born to white mothers. State the coefficient restrictions on regression equation (1) implied by this proposition; that is, state the null hypothesis H_0 and the alternative hypothesis H_1. Write the restricted regression equation implied by the null hypothesis H_0. OLS estimation of this restricted regression equation yields a R-squared value of $\boldsymbol{R^2}$**=0.0193**. Use this information, together with the results from OLS estimation of equation (1), to calculate the required test statistic. State the inference you would draw from the test at the 5 percent significance level (i.e., for significance level $\alpha = 0.05$)

Question 4 (40 points)
Please answer all questions below.

(4 points)
(a) Consider a time series model $y_t = \rho y_{t-1} + \epsilon_t$ with $-1 < \rho < 1$. $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is a sequence of i.i.d. random variables with mean 0 and variance $\sigma^2 > 0$. Suppose that $\boldsymbol{E}[y_{t-1}\epsilon_t] = 0$ for all $t = 1,2,\ldots,\mathrm{T}$. Suppose we estimate ρ using OLS. Is it likely that the OLS estimator $\hat{\rho}$ is unbiased? Explain.

(4 points)
(b) Compute $Cov(y_t, \epsilon_t)$ and $Cov(y_t, y_{t-1})$. Are y_t and y_{t-1} likely to be uncorrelated?

For the rest of the question, consider a panel data model

$$y_{i,t} = \alpha + \rho y_{i,t-1} + \beta x_{i,t} + u_i + \epsilon_{i,t}, -1 < \rho < 1$$

where u_i is an unobserved error term that is correlated with $x_{i,t}$ and constant within each cross-sectional unit i, and $\epsilon_{i,t}$ is an error term that changes across i and t. Assume that $\epsilon_{i,t}$ is uncorrelated with $y_{i,t-1}$ and $x_{i,t}$.

(8 points)
(c) Write down the first-difference estimator for ρ and β. That is, write down the equation that the first-difference method estimates, and state how to estimate ρ and β. You do not need to write down the explicit formulas for the estimators.

(4 points)
(d) What is the key assumption that guarantees the consistency of the first-difference estimators?

(20 points)
(e) Some econometricians propose using $y_{i,t-2}$ as an instrumental variable. Explain why we need such an instrumental variable and why it helps to correct the endogeneity problem. Notice that you need to explain why it satisfies (1) the exogeneity condition and (2) the relevance condition.

Scratch Paper

Scratch Paper

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 for a two-tailed test with large (> 120) *df* is 1.96.

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

5% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Example: The 5% critical value for numerator *df* = 4 and large denominator *df* (∞) is 2.37.
Source: This table was generated using the Stata® function invfprob.

ECONOMICS 3121

FORMULA SHEET

Basic statistics

$$\sigma_x^2 \; or \; \mathrm{Var}(x) = E[(x - E(x))^2]$$
$$= E(x^2) - (E(x))^2$$

$$\mathrm{Cov}(x,y) \; = \; E\,[(x - E(x))(y - E(y))]$$
$$= \; E(xy) - E(x)E(y)$$

$$corr(x,y) \; \equiv \rho_{xy} \; = \; \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$$

$$E(ax + by) = aE(x) + bE(y)$$

$$\mathrm{Var}(ax + by) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(y) + 2ab\mathrm{Cov}(x,y)$$

$$\text{sample Mean } (x) \; or \; \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_x^2 \; or \, \text{sample Var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\text{sample } \; \mathrm{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)$$

Simple OLS:

$$\hat{\beta}_0 \; = \; \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 \; = \; \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \; = \; \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$$

$$\mathrm{se}(\hat{\beta}_1) \; = \; \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\hat{\sigma}^2 \; = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

$$SST \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$SSE \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SSR \equiv \sum_{i=1}^{n}\hat{u}_i^{\,2}$$

$$R^2 \; = \frac{SSE}{SST} = \; 1 - \left(\frac{SSR}{SST}\right)$$

$$\text{adjusted } R^2 \; \equiv \; \bar{R}^2 \; = 1 - \left(\frac{SSR/(n-k-1)}{SST/(n-1)}\right)$$
$$= 1 - (1 - R^2)\,\frac{(n-1)}{(n-k-1)}$$

Multivariate OLS:

Variance of estimate of β_j is:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

$SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and R_j^2 is the R^2 from regressing x_j on all other x's

Omitted Variables:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$

$\tilde{\delta}$ is the expected value of OLS estimator from regressing x_2 on x_1

$\beta_2 \tilde{\delta}_1$ is omitted variable bias

$$\text{t-statistic } or \;\; t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

The $100(1-\alpha)$ percent Confidence Interval is:

$$[\hat{\beta}_j - c_\alpha se(\hat{\beta}_j), \;\; \hat{\beta}_j + c_\alpha se(\hat{\beta}_j)]$$

$$F \; \equiv \; \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \; \equiv \; \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)}$$