THE UNIVERSITY OF MELBOURNE

Student ID _____

Semester Two Assessment, 2018

Faculty / Dept: Faculty of Business and Economics
Department of Economics

Subject Number **ECOM20001**

Subject Name **Econometrics 1**

Writing time 2 hrs

Reading 15 minutes

Open Book status Closed Book

Number of pages (including this page) 17

Authorised Materials:

Foreign language/English dictionaries
Scientific Calculator Casio FX82 (any suffix)

Instructions to Students:

There are 100 marks available on the exam. Your exam score contributes up to 65% of your overall grade, unless the mid-semester exam is not included, in which case this exam contributes up to 75% of the overall mark in the subject.

ANSWER ALL QUESTIONS in your exam booklet. Answers to the Multiple-Choice questions are also to be written in your exam booklet.

This exam paper also contains **critical values** for a number of distributions on pages 13-14, and a **formula sheet** starting on pages 15-17.

Instructions to Invigilators:

Paper to be held by Baillieu Library: yes ____ No **X**

Extra Materials required (please supply)

Graph paper **No** Multiple Choice form **No**

Question 1: Multiple Choice (20 marks, 2 for each question)

1. Which of the following is an example of time series data?

A) Data on the unemployment rates in different parts of a country during a year.
B) Data on the consumption of wheat by 200 households during a year.
C) Data on the gross domestic product of a country over a period of 10 years.
D) Data on the number of vacancies in various departments of an organisation in a particular month.
E) None of the above.

2. The probability of an outcome

A) is the number of times that the outcome occurs in the long run.
B) equals $M \times N$, where M is the number of occurrences and N is the population size.
C) is the proportion of times that the outcome occurs in the long run.
D) equals the sample mean divided by the sample standard deviation.
E) none of the above.

3. An estimator $\hat{\mu}_Y$ of the population value μ_Y is unbiased if

A) $\hat{\mu}_Y = \mu_Y$.
B) $\overline{Y}$ has the smallest variance of all estimators.
C) $\overline{Y} \xrightarrow{p} \mu_Y$.
D) $E(\hat{\mu}_Y) = \mu_Y$.
E) none of the above.

4. What does <u>not</u> change if you multiply the dependent variable by 100 and the explanatory variable by 100,000 in a single linear regression?

A) The OLS estimate of the slope.
B) The OLS estimate of the intercept.
C) The regression R^2.
D) The variance of the OLS estimator.
E) None of the above.

5. The t-statistic is calculated by dividing

A) the OLS estimator by its variance.
B) the slope by the standard deviation of the explanatory variable.
C) the estimator minus its hypothesized value by the standard error of the estimator.
D) the slope by 1.96.
E) None of the above.

6. When there are omitted variables in the regression, which are also 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(s).
D) the OLS estimator(s) of the included variable(s) is (are) biased if the omitted variable is correlated with the included variable(s).
E) None of the above.

7. 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.
E) None of the above.

8. The interpretation of the slope coefficient in the model $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:
A) a 1% change in X is associated with a β_1 % change in Y.
B) a change in X by one unit is associated with a β_1 change in Y.
C) a change in X by one unit is associated with a 100 β_1 % change in Y.
D) a 1% change in X is associated with a change in Y of 0.01 β_1.
E) None of the above.

9. In the equation $\widehat{TestScore} = 607.3 + 3.85\ Income - 0.0423 Income^2$, which of the following income levels approximately results in the maximum test score?
A) 607.3.
B) 91.02.
C) 45.5.
D) cannot be determined without a plot of the data.
E) None of the above.

10. The AIC statistic:
A) is commonly used to test for heteroskedasticity in time series data
B) is the correct measure for testing goodness of fit in time series models
C) is an alternative to the BIC when sample size is small (typically T < 50)
D) helps in determining the number of lags to include in a time series model
E) None of the above.

Question 2: Short Answer Questions (20 marks)

1. What are the Law of Large Numbers and the Central Limit Theorem and why are they important for regression-based empirical analysis? (6 marks)

2. Females, it is said, make 70% of the male wage. To investigate this phenomenon, you collect data on weekly earnings from 1,744 individuals, 850 females and 894 males. Next, you calculate their average weekly earnings and find that the females in your sample earned $346.98, while the males made $517.70.
   (a) How would you test whether this gender-wage gap is statistically significant? Give two approaches. (2 marks)

   (b) A peer suggests that this is consistent with the idea that there is discrimination against females in the labor market. What is your response? (2 marks)

   (c) Assuming, heroically, that education is constant across the 1,744 individuals, you consider regressing earnings on age and a binary variable for gender. You estimate two specifications initially:

$$\widehat{Earn} = 323.70 + 5.15 \times Age - 169.78 \times Female,\ R^2=0.13$$
$$\qquad\quad (21.18)\ \ (0.55)\qquad\ (13.06)$$
$$\widehat{Ln(Earn)} = 5.44 + 0.015 \times Age - 0.421 \times Female,\ R^2=0.17$$
$$\qquad\quad\ \ (0.08)\ \ (0.002)\qquad\ (0.036)$$

   where *Earn* are weekly earnings in dollars, *Age* is measured in years, and *Female* is a binary variable, which takes on the value of one if the individual is a female and is zero otherwise. Interpret the linear regression carefully. (2 marks)

   (d) Interpret the non-linear regression carefully. Based on the above regression results, for a given age, how much less do females earn on average? Should you choose the second specification on grounds of the higher regression R^2? (4 marks)

   (e) Your peer points out to you that age-earning profiles typically take on an inverted U-shape. To test this idea, you add the square of age to your log-linear regression.

$$\widehat{Ln(Earn)} = 3.04 + 0.147 \times Age - 0.421 \times Female - 0.0016\ Age^2,\ R^2 = 0.28$$
$$\qquad\quad\ \ (0.18)\ (0.009)\qquad\ (0.033)\qquad\ (0.0001)$$

   Are there strong reasons to assume that this specification is superior to the previous one? (2 marks)

   (f) What other factors may play a role in earnings determination? (2 marks)

Question 3: Lead mortality (40 marks)

Lead is toxic, particularly for young children, and for this reason government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leaked into drinking water.

You will investigate the effect of these lead water pipes on infant mortality. You are provided with a dataset called Lead_Mortality.csv on 172 U.S. cities in the year 1900 that includes the following variables:

- *infrate*: Infant mortality rate (deaths per 1,000 infants in population)
- *lead*: Indicator = 1 if city had lead pipes.
- *pH*: Water pH (a measure of water acidity)
- *pop*: City population (in 100s)
- *age*: Average age of city population

You learn that the amount of lead leaked from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water (that is, the lower its pH), the more lead is leaked. You create a new variable (*lead_pH*) which is the interaction term *lead* x *pH* and you run the following regression:

$$infrate_i = \beta_0 + \beta_1 lead_i + \beta_2 pH_i + \beta_3 lead\_pH_i + u_i$$

Figures 1 and 2 present some summary statistics for variable *infrate* and the regression results from R-Studio. For all parts of question 3, only conduct hypothesis tests based on regressions with heteroskedasticity-robust standard errors. Also, round to three decimal points for all calculations.

a. Using information from Figure 1, compute the 95% confidence intervals of the average infant mortality rate for cities with lead pipes and for cities with non-lead pipes. Note that there are 55 cities with lead pipes (lead=1) and 172 - 55 = 117 cities with non-lead pipes in the sample. (4 marks)

b. Turning to Figure 2, what is the 95% confidence interval for β_3? (2 marks)

c. What is the overall regression *F*-statistic for the regression in Figure 2, and what are its corresponding degrees of freedom. Interpret the statistical significance of this test at the 1% level, and its implication for the model. (2 marks)

d. Interpret the signs of the coefficient estimates on *lead*, *pH*, and *lead_pH* in Figure 2. Does the effect of *lead* on infant mortality depend on *pH*? Discuss statistical significance at the 5% level. (4 marks)

e. The average value of *pH* in the sample is 7.323. At this pH level, what is the estimated effect of *lead* on infant mortality implied by Figure 2? Interpret this result. (2 marks)

f. Carefully explain the steps required to compute the standard error of the estimated effect of *lead* on infant mortality at the average *pH* level that you derived in question e. above. How would you then derive the 90% confidence interval? (4 marks)

g. The standard deviation of *pH* is 0.691. Suppose that we could decrease the pH level of water to one standard deviation lower than the average level of *pH* in the sample; what would the estimated effect of *lead* on infant mortality be? What if the *pH* level were one standard deviation higher than the average value? Interpret these results. (4 marks)

h. You are concerned about omitted variable bias. Building further on your regression model, you now estimate a second regression model:

$$infrate_i = \beta_0 + \beta_1 lead_i + \beta_2 pH_i + \beta_3 lead\_pH_i + \beta_4 age_i + \beta_5 age_i^2 + \beta_6 \log(pop) + u_i$$

The regression results are reported in Figure 3. Comparing regression results in Figures 2 and 3, should you be concerned about omitted variable bias? (2 marks)

i. Interpret the coefficient on log(pop) in Figure 3 and comment on whether it is statistically significant at the 10% level. (2 marks)

j. Looking at the coefficient estimates on *age* and age^2 in Figure 3, is the model depicting a U-shape or an inverted U-shape relationship between *infrate* and *age*? At what value of *age* is infant mortality predicted to be the lowest? What does it say about the sign of the effect of *age* on infant mortality in your sample? (4 marks)

k. What test is being conducted on Figure 4? Describe the outcome of the test using a 5% significance level, noting the relevant test statistic and degrees of freedom (if necessary). Interpret the result of this test. Based on this test alone, would you prefer the first model (Figure 2) or the second model (Figure 3)? (4 marks)

l. Would the adjusted R^2 presented in Figures 2 and 3 make you change your mind? (2 marks)

m. Using only the raw data provided, provide the **pseudo-code**[1] for an R program (e.g., the .R code) that you would write in R-Studio for estimating the elasticity of *infrate* with respect to *pop*, and its standard error, based on the model in Figure 3 (from part h) at the median of average age (27.59).
Your pseudo-code can be written in a series of bullet points. It should explicitly state all steps required in R-script to generate this test. You do not need to cite explicit R

[1] A pseudo-code consists of all the steps you would take in an R program for conducting an analysis or calculation. It is primarily written in words and not R commands or syntax.

commands, syntax, or equations, but you may do so if it helps clarify what each part of your pseudo-code does. (4 marks)

Figure 1. Mean and S.D. of *infrate* by type of water pipe

```
> mean(infrate[lead==0])
[1] 3.811679
> sd(infrate[lead==0])
[1] 1.477588
>
> mean(infrate[lead==1])
[1] 4.032576
> sd(infrate[lead==1])
[1] 1.530873
```

Figure 2. Infant mortality regression output 1

```
> lead_pH=lead*pH
> reg1=lm(infrate~lead+pH+lead_pH,data=mydata)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  9.18904    1.50494  6.1059 6.866e-09 ***
lead         4.61798    2.07614  2.2243 0.0274596 *
pH          -0.75179    0.20953 -3.5879 0.0004369 ***
lead_pH     -0.56862    0.28084 -2.0247 0.0444778 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(reg1)$adj.r.squared
[1] 0.2588579
> waldtest(reg1, vcov = vcovHC(reg1, "HC1"))
Wald test

Model 1: infrate ~ lead + pH + lead_pH
Model 2: infrate ~ 1
  Res.Df Df      F    Pr(>F)
1    168
2    171 -3 20.974 1.366e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. Infant mortality regression output 2

```
> age2=age*age
> reg2=lm(infrate~lead+pH+lead_pH+age+age2+log(pop),data=mydata)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.212831  12.437970  2.1075  0.03659 *
lead         4.362605   1.884854  2.3146  0.02187 *
pH          -0.738779   0.182627 -4.0453 8.01e-05 ***
lead_pH     -0.570530   0.253902 -2.2470  0.02596 *
age         -1.002287   0.878252 -1.1412  0.25543
age2         0.013322   0.015423  0.8638  0.38897
log(pop)     0.088242   0.085663  1.0301  0.30447
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(reg2)$adj.r.squared
[1] 0.3994807
> waldtest(reg2, vcov = vcovHC(reg2, "HC1"))
Wald test

Model 1: infrate ~ lead + pH + lead_pH + age + age2 + log(pop)
Model 2: infrate ~ 1
  Res.Df Df      F    Pr(>F)
1    165
2    171 -6 20.343 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4. Infant mortality regression test

```
> linearHypothesis(reg2,c("age=0","age2=0","log(pop)=0"),vcov = vcovHC(reg2, "HC1"))
Linear hypothesis test

Hypothesis:
age = 0
age2 = 0
log(pop) = 0

Model 1: restricted model
Model 2: infrate ~ lead + pH + lead_pH + age + age2 + log(pop)

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    168
2    165  3 15.415 6.907e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 4: Forecasting inflation (20 marks)

The Reserve Bank of Australia has hired you to develop time series models for the inflation rate. They provide you with a time series for just one variable, $PCEC_t$, which is the price index in quarter t. These data are provided from 1963:Q1 to 2013:Q4 for a total of T = 204 observations.

a. You compute the inflation rate, $Infl_t = 100 \times [\ln(PCEC_t) - \ln(PCEC_{t-1})]$. What are the units of measurement of $Infl_t$? Explain. (2 marks)

b. Figure 5 plots the autocorrelations of the inflation rate and of the first difference in inflation rate ($\Delta Infl_t = Infl_t - Infl_{t-1}$). What does this tell you about the persistence of inflation? (4 marks)

c. You now estimate autoregressive models of order 1 and 2 (AR(1) and AR(2) models) for the inflation rate. Regression results are reported in Figure 6. Interpret the regression coefficients on 1st and 2nd lags in the AR(2) model of Figure 6. Comment on their statistical significance at the 1% level. (4 marks)

d. Figure 6 also reports the BIC and AIC for both regressions. Based on these information criteria, which model should you prefer? (2 marks)

e. Using information from Figure 6, what is your best guess of the Residual Mean Square Forecast Error (RMFSE) for the AR(2) model? Explain the required condition for your guess to be valid. (4 marks)

f. You now consider the potential effect of seasonality and include quarterly dummies in your AR(2) model (*q2* = 1 if *Quarter* = 2 and 0 otherwise, *q3* = 1 if *Quarter* = 3 and 0 otherwise, *q4* = 1 if *Quarter* = 4 and 0 otherwise). Regression results are reported in Figure 7. Comment on the statistical significance of the individual quarterly dummies at the 10% level. Interpret only the regression coefficients that are statistically different from 0. (2 marks)

g. The inflation rate was 0.173% in 2013:Q4, 0.475% in 2013:Q3, -0.029% in 2013:Q2 and 0.269% in 2013:Q1. Based on the AR(2) model from Figure 7, what is your forecast for the inflation rate in 2014:Q1? (2 marks)

Figure 5. Autocorrelations in inflation rates

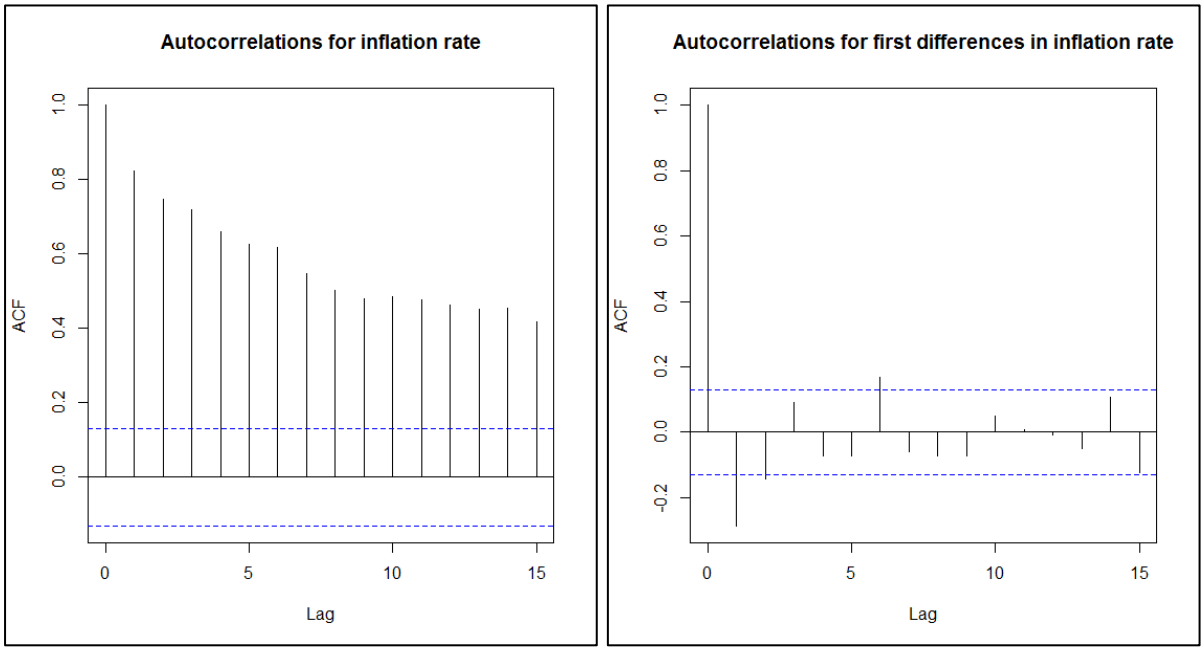

Figure 6. Inflation regression output 1

```
> reg1=lm(infl~infl_lag1,data=mydata)
> coeftest(reg1, vcov = vcovHC(reg1, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.147134   0.041396  3.5543 0.0004718 ***
infl_lag1   0.831424   0.043236 19.2300 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
>
> reg2=lm(infl~infl_lag1+infl_lag2,data=mydata)
> coeftest(reg2, vcov = vcovHC(reg2, "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.119215   0.043494  2.7410  0.006678 **
infl_lag1   0.672156   0.058938 11.4045 < 2.2e-16 ***
infl_lag2   0.191355   0.068591  2.7898  0.005782 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> K1=2
> SER1=sd(residuals(reg1))
> ssr1=sum(reg1$resid^2)
> BIC1=log(ssr1/T)+K1*(log(T)/T)
> AIC1=log(ssr1/T)+K1*(log(2)/T)
>
> K2=3
> SER2=sd(residuals(reg2))
> ssr2=sum(reg2$resid^2)
> BIC2=log(ssr2/T)+K2*(log(T)/T)
> AIC2=log(ssr2/T)+K2*(log(2)/T)
>
> sprintf("BIC of reg1: %f", BIC1[1])
[1] "BIC of reg1: -2.005760"
> sprintf("BIC of reg2: %f", BIC2[1])
[1] "BIC of reg2: -2.016902"
> sprintf("AIC of reg1: %f", AIC1[1])
[1] "AIC of reg1: -2.051103"
> sprintf("AIC of reg2: %f", AIC2[1])
[1] "AIC of reg2: -2.084916"
> sprintf("SER of reg1: %f", SER1[1])
[1] "SER of reg1: 0.358261"
> sprintf("SER of reg2: %f", SER2[1])
[1] "SER of reg2: 0.351657"
```

Figure 7. Inflation regression output 2

```
> reg3=lm(infl~infl_lag1+infl_lag2+q2+q3+q4,data=mydata)
> coeftest(reg3, vcov = vcovHC(reg3, "HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.165130   0.057055  2.8942  0.004227 **
infl_lag1    0.690027   0.057933 11.9108 < 2.2e-16 ***
infl_lag2    0.174432   0.067606  2.5801  0.010601 *
q2          -0.055402   0.071621 -0.7736  0.440118
q3          -0.003446   0.062268 -0.0553  0.955923
q4          -0.128206   0.076897 -1.6672  0.097048 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

END OF EXAMINATION

Statistical Distribution Tables

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
	35	1.306	1.690	2.030	2.438	2.724
	36	1.306	1.688	2.028	2.434	2.719
	37	1.305	1.687	2.026	2.431	2.715
	38	1.304	1.686	2.024	2.429	2.712
	39	1.304	1.685	2.023	2.426	2.708
	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

95th Percentile for the F-distribution F_{ν_1,ν_2}

	Numerator ν_1											
Denominator ν_2	ν_2/ν_1	1	2	3	4	5	7	9	10	15	20	60	∞
	1	161.45	199.50	215.71	224.58	230.16	236.77	240.54	241.88	245.95	248.01	252.2	254.31
	2	18.51	19.00	19.16	19.25	19.30	19.35	19.41	19.40	19.43	19.45	19.48	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.89	8.81	8.79	8.70	8.66	8.57	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.09	6.00	5.96	5.86	5.80	5.69	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.88	4.77	4.74	4.62	4.56	4.43	4.37
	6	5.99	5.14	4.76	4.53	4.39	4.21	4.10	4.06	3.94	3.87	3.74	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.79	3.68	3.64	3.51	3.44	3.30	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.50	3.39	3.35	3.22	3.15	3.01	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.29	3.18	3.14	3.01	2.94	2.79	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.14	3.02	2.98	2.85	2.77	2.62	2.54
	15	4.54	3.68	3.29	3.06	2.90	2.71	2.59	2.54	2.40	2.33	2.16	2.07
	20	4.35	3.49	3.10	2.87	2.71	2.51	2.39	2.35	2.20	2.12	1.92	1.84
	30	4.17	3.32	2.92	2.69	2.53	2.33	2.21	2.16	2.01	1.93	1.74	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.25	2.12	2.08	1.92	1.84	1.64	1.51
	50	4.03	3.18	2.79	2.56	2.40	2.20	2.07	2.03	1.87	1.78	1.58	1.44
	60	4.00	3.15	2.76	2.53	2.37	2.17	2.04	1.99	1.84	1.75	1.53	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.09	1.95	1.91	1.75	1.66	1.43	1.25
	∞	3.84	3.00	2.60	2.37	2.21	2.01	1.88	1.83	1.67	1.57	1.32	1.00

Critical Values for the Chi-Squared Distribution

Degrees of Freedom	Critical Values		
	1%	5%	10%
1	6.64	3.84	2.71
2	9.21	5.99	4.61
3	11.35	7.81	6.25
4	13.28	9.49	7.78
5	15.09	11.07	9.24
6	16.81	12.59	10.65
7	18.48	14.07	12.02
8	20.09	15.51	13.36
9	21.67	16.92	14.68
10	23.21	18.31	15.99
11	24.73	19.68	17.28
12	26.22	21.0	18.55
13	27.69	22.4	19.81
14	29.14	23.7	21.06
15	30.58	25.0	22.31
16	32.00	26.3	23.54
17	33.41	27.6	24.77
18	34.81	28.9	25.99
19	36.19	30.1	27.20
20	37.57	31.4	28.41

Formula Sheet

Expected Values, Variances, Correlation

$E(c) = c$

$E(cx) = cE(x)$

$E(a + cx) = a + cE(x)$

$E(x + y) = E(x) + E(y)$

$E(c_1x + c_2y) = c_1E(x) + c_2E(y)$

$var(x) = \sigma^2 = E(x - E(x))^2$

$std(x) = \sigma = \sqrt{E(x - E(x))^2}$

$var(a + cx) = c^2var(x)$

$cov(x, y) = E[(x - E(x))(y - E(y))]$

$corr(x, y) = \rho = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$

$P(y = y_1|x = x_1) = \frac{P(x=x_1, y=y_1)}{p(X=x_1)}$

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$var(\bar{Y}) = \frac{\sigma_Y^2}{n}$

$std(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$

$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$

$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

$r_{xy} = \frac{s_{xy}}{s_x s_y}$

Logarithms

$x = \ln(e^x)$

$\frac{d\ln(x)}{dx} = \frac{1}{x}$

$\ln(1/x) = -\ln(x)$

$\ln(ax) = \ln(a) + \ln(x)$

$\ln(x/a) = \ln(x) - \ln(a)$

$\ln(x^a) = a\ln(x)$

$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$ (approximately equal for small Δx)

Calculus

x^* that maximizes (minimizes) a strictly concave (convex) function, $f(x)$, solves $\frac{df(x)}{dx} = 0$

OLS Estimator

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s^2_X}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var((X_i - \mu_X)u_i))}{(var(X_i))^2}$

$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{var(H_i u_i)}{(E(H_i^2))^2}$; where $H_i = 1 - (\frac{\mu_X}{E(X_i^2)})X_i$

$\hat{\beta}_1 \to \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$

Hypothesis Testing

Different populations

$H_0 : \mu_w - \mu_m = d_0; \quad vs. \quad H_1 : \mu_w - \mu_m \neq d_0$

$SE(\bar{Y}_w - \bar{Y}_m) = \sqrt{s_w^2/n_w + s_m^2/n_m}$

$t^{act} = \frac{(\bar{Y}_w - \bar{Y}_m) - d_0}{SE(Y_w - Y_m)}$

Linear Regression

$t^{act} = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 \neq \beta_{1,0}$, p-value $= 2\Phi(-|t^{act}|)$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 < \beta_{1,0}$, p-value $= \Phi(t^{act})$

$H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 > \beta_{1,0}$, p-value $= 1 - \Phi(t^{act})$

t^α is the critical value for a two-sided test with α significance level

$\alpha = 2\Phi(|t^\alpha|)$

$(1-\alpha)$ CI: $[\hat{\beta}_1 - t^\alpha SE(\hat{\beta}_1), \hat{\beta}_1 + t^\alpha SE(\hat{\beta}_1)]$

For testing means, replace β with μ_X and $\hat{\beta}$ with $\bar{X}$

Joint-testing

$H_0 : \beta_j = \beta_{j,0}, \ \beta_m = \beta_{m,0}, \ldots$ for a total of q restrictions

H_1 : one or more of the q restrictions under H_0 does not hold

the F-statistic is distributed $F_{q,n-k-1}$

p-value $= \Pr[F_{q,n-k-1} > F^{act}] = 1 - G(F^{act}; q, n-k-1)$

$F = \frac{1}{2}\left(\frac{(t_1^{act})^2 + (t_2^{act})^2 - 2\hat{\rho}_{t_1^{act}, t_2^{act}} t_1^{act} t_2^{act}}{1 - \hat{\rho}_{t_1^{act}, t_2^{act}}}\right)$ if $q = 2$

$F^{act} = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n-k-1)} = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n-k-1)}$

Goodness of Fit

$SSR = \sum_{i=1}^{n} u_i^2$

$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

$$SER = s_{\hat{u}} = \sqrt{s^2_{\hat{u}}},\ s^2_{\hat{u}} = \frac{SSR}{n-k-1}$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$$

Nonlinear and Time Series Regression

$$E[Y|X_1, X_2, \ldots, X_k] = f(X_1, X_2, \ldots, X_k)$$

$$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k)$$

$$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}$$

$$(1-\alpha)\text{ CI: } [\Delta\hat{Y} - t^{\alpha}SE(\Delta\hat{Y}), \Delta\hat{Y} + t^{\alpha}SE(\Delta\hat{Y})]$$

$$\text{RMSFE} = \sqrt{E[(Y_{T+1} - \hat{Y}_{T+1|T})^2]}$$

$$SE(Y_{T+1} - \hat{Y}_{T+1|T}) = \widehat{RMSFE} = \sqrt{var(\hat{u}_t)} = SER$$

$$(1-\alpha)\text{ CI: } [\hat{Y}_{T+1|T} - t^{\alpha} \times SE(Y_{T+1} - \hat{Y}_{T+1|T}), \hat{Y}_{T+1|T} + t^{\alpha} \times SE(Y_{T+1} - \hat{Y}_{T+1|T})]$$

$$\text{BIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{\ln(T)}{T}$$

$$\text{AIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{2}{T}$$