UNIVERSITY OF CAMBRIDGE
Judge Business School

Updated: November 21, 2011

Lecturer: Thilo Klein
Contact: tk375@cam.ac.uk

MPO1 Workbook 1

Michaelmas 2011

- Answer all questions in no more than 14 pages (excluding appendix). Please be brief. There is no credit at all for extended answers.
- The weighting of Workbook 1 is 40% of the total for this module.
- The figure in square brackets after each question is the weight carried in Workbook 1.
- Documentation of your analysis: Your empirical analysis must be capable of being replicated. For this reason, you should annotate and include your $R-script$ file (or batch file in any other statistical package that you prefer to use) in the appendix. If you use packages that do not have batch command facility, you should report the sequence of your computational steps, such as data transformations etc in the appendix. This document / your batch file / your $R-script$ file) should be printed and attached as the **only** appendix.
- Note: http://www.intranet.jbs.cam.ac.uk/students/plagiarism.html
- Two copies of the completed work, together with a standard signed cover-sheet must be submitted to Ms. Nena Nelson at the MPhil Office by noon on 12 December 2011. Electronic submission is not permitted.
- If work is submitted after the deadline without valid tutorial reasons the mark for the submission will be reduced by $10n$ percentage points, where n is the number of whole, or part, weeks late.

Exercise 1

A) The net weight of a bag of specialty coffee at the CJBS café is "guaranteed" to be 10 pounds with a standard deviation of 1.5 pounds. Stephen, who manages the coffee shop is concerned that the actual weight is lower. To help him test for this, you sample 40 bags randomly.

(i) State the null and alternative hypotheses for your test.

(ii) Determine a critical value such that the size of the test does not exceed 5%.

(iii) You find that the average weight of your sample of 40 bags is 9.5 pounds. What do you conclude?

[7.5%]

B) Now suppose that under the alternative hypothesis, the expected value is 10.5 pounds.

(i) Sketch freehand the probability density function for the null and the alternative hypotheses in the same figure.

(ii) Pick the critical value such that the p-value is approximately 5%. Mark the areas that show the size and the power of the test.

(iii) Very briefly, what happens to the power of the test if the alternative hypothesis moves closer to the null hypothesis? (i.e., $\mu = 10.4, 10.3, 10.2$, etc.)

[12.5%]

Exercise 2

A) *scores.csv* contains information on educational performance of students in a university whose students often apply to JBS. The variables are:

fgpa	overall grade point average at end of freshman year (on scale from 0 to 4)
satm	score on the SAT Mathematics test divided by 100 (on scale from 0 to 10)
fem	gender (1 for females, 0 for males)

(i) Find the sample mean, median, standard deviation, skewness and kurtosis for each of the three variables.

(ii) Find sample covariances and sample correlation coefficients between the variables.

(iii) Plot histograms and scatter plots for these variables (assume the dataset is the entire population of interest).

(iv) Find the expectation of *FGPA*, and the expectation of *FGPA* conditional on *FEM*.

[5%]

B) Now assume that the data constitute an *i.i.d.* sample drawn from a population with distribution given by $N(\mu, \sigma^2)$.

(i) Test the null that the population mean is equal to 2.72 against the alternative that is less than 2.72. Calculate the p-value (using *R*, or any other software).

(ii) Test the null that the population mean is equal to 2.72 against the two-sided alternative that $\mu \neq 2.72$.

(iii) What is the relation between the p-values of one-sided and the two-sided tests?

[10%]

Exercise 3

During the last few days before any election, many voting intention surveys tend to be carried out. On any given day, quite often there are conflicting results from these polls. Think of a poll as reporting the fraction of successes (1s) of a random variable Y that can take values of success (1) or failure (0) for a candidate of interest. The probability of success is $Pr(Y = 1) = p$. Let $\hat{p}$ be the fraction of successes in the sample. We assume that this estimator is normally distributed with a mean of p and a variance of $\frac{p(1-p)}{n}$.

A) (i) Given that the estimator of the variance of $\hat{p}$ is given by $\frac{\hat{p}(1-\hat{p})}{n}$, construct a 95% confidence interval for p.

(ii) For which value of $\hat{p}$ is the standard deviation the largest?

(iii) What value does the standard deviation take in the case of a maximum $\hat{p}$?

[5%]

B) (i) When the results from the polls are reported, you are told, typically in the small print, that the "margin of error" is plus or minus two percentage points. Using the approximation of $1.96 \approx 2$, and assuming, "conservatively," the maximum standard deviation derived above, what sample size is required to add or subtract two percentage points ("margin of error") from the point estimate?

(ii) What sample size would you need to halve the margin of error?

[10%]

Exercise 4

The news-magazine *The Economist* regularly publishes data on the so called *Big Mac index* and exchange rates between countries. The data for 45 countries from the July 16, 2009 issue is listed in *bigmac.csv*.

The idea that similar foreign and domestic goods should have the same price in terms of the same currency is called *purchasing power parity*. This suggests that the ratio of the *Big Mac* priced in the local currency to the U.S. dollar price should equal the exchange rate between the two countries. Calculate the predicted exchange rate per U.S. dollar by dividing the price of a Big Mac in local currency by the U.S. price of a *Big Mac* ($3.57).

(i) Run a regression of the actual exchange rate on the predicted exchange rate. If purchasing power parity held, what would you expect the slope and the intercept of the regression to be? Test hypotheses about the slope and the intercept.

(ii) Explain the consequence of disregarding any extreme observations in the independent variable on the standard deviation of the OLS estimator of the slope?

[12.5%]

Exercise 5

rlms.csv contains earnings, experience and education data from the Russia Longitudinal Monitoring Survey for 1,204 Russian individuals in the year 2004. The variables are:

gender	male, female
height	height in cm
totwaglm	total pay received in past month
age	age
wtimeyears	years in current job
exp	potential labor force experience (age - years education - 6)
edyrs	years of education completed

(i) Estimate a model that explains Total Pay received in the past month as a linear function of *gender*, *age*, *wtimeyears*, *exp*, and *edyrs*. Explain why the number of observations used in the estimation is less than the number of observations in the sample.

(ii) Explain why *edyrs* is dropped in estimation?

(iii) Test whether this model suffers from imperfect multicollinearity.

(iv) Re-specify the model in the light of your test, and comment on the results.

(v) Include *height* as an additional explanatory variable. Height is sometimes considered a proxy for childhood nutrition, *ceteris paribus*. Compare the estimates you obtain for gender with the results of the model you estimated for section (iii) above. Explain why these coefficients may be different?

(vi) Are there any variable transformations that might give more useful results?

[37.5%]