DATA1001 Introduction to Data Science and Decisions

Assignment 1

Question 1 (4 marks)

Are the following statements *True*, *False* or *Uncertain*? You will also need to justify your answer. (Please keep your explanations brief. Answers should be no more than half a page each.)

(a) Outliers need to be deleted from your data because they have the potential to distort statistical analyses.

(b) Consider the following regression model designed to explain the annual electricity consumed (*elec*) by a sample of households as a function of education represented by the years of education of the household head (*educ*), household income (*income*) and other control variables:

$$elec_i = \beta_0 + \beta_1 educ_i + \beta_2 income_i + \beta_3 x_{3i} + \cdots + \beta_p x_{pi} + u_i.$$

Education has two distinct and opposite effects on electricity demand. On the one hand, households with highly educated members tend to consume less electricity because they have a stronger awareness of conservation and environmental concerns. On the other hand, higher education is normally associated with higher income, which tends to increase electricity usage. Therefore, the expected coefficient sign of β_1 is uncertain.

Question 2 (2 marks)

A hotel chain is considering changes to the services they offer and their overall pricing strategy with an aim to increasing their market share. The chain decides it needs to undertake some market research to provide input into the decision-making process. In each of the following cases describe any possible biases in the sampling approach:

(a) They hand out a survey to the customers who are currently staying in their hotels and ask them to return them when they check out.

(b) They include the survey in a follow-up email to a randomly selected sample of recent hotel guests.

(c) They randomly select telephone numbers from residents in cities served by their hotels and conduct a telephone poll between 9am and 4pm to ask the survey's questions.

(d) They used the Google Review responses for their hotels.

Question 3* (9 marks)

Many products such as artwork and houses are sold at auction. Typically, these are auctioned in the "English" or "ascending price" format. Bidding starts low, and the auctioneer subsequently calls out higher and higher prices. Sellers of individual items will set a secret reserve price, and if the bidding does not reach this level, the items will go unsold. While reserve prices not known, price estimates are typically made available to prospective buyers.

Here we are interested in contemporary art presented for sale at Christies' and Sotheby's in London. We use a subset of data collected by Professor Kathryn Graddy contained in conart_prices.txt that is available from the data section of the Moodle site. The data represent distinct works of art presented for auction and the variables of interest are:

price = Price (if sold) or highest bid (if unsold) - in thousands of English pounds

low_est = Expert low-price estimate - in thousands of English pounds

sold = 1 if the art was sold and 0 otherwise

(a) What are the key features of the distribution of *price*? **(2 marks)**

(b) Run the simple regression $y_i = \beta_0 + \beta_1 x_i + u_i$ where $y_i = price$ and $x_i = sold$ and interpret the results. **(2 marks)**

(c) Now consider the relationship between *price* and *low_est*. Using relevant summary statistics, plots, correlation and regression write a short report describing this relationship. **(5 marks)**

*** Note: You are expected to use R to complete this question.**

Notes on Assignment

The course outline states that assignments should be submitted in labs. This is not correct. Your answers to the above questions must be submitted **to Assignment Box 4, labelled DATA1001 located on the ground floor of the UNSW Business School Building in the West Wing before the end of Week 5 (that's 5pm on Friday August 24)**. Late penalties apply; see course outline for details.

These will be marked for the course assessment and will be worth 15% of your final grade. The Assignment is based on the material covered by lectures and tutorials up to the end Week 4.

You must attach your R code for Question 3 as an appendix. Your submission, including the attached cover sheet and appendix, should not exceed 7 pages. Do not use plastic sheets or binders. Do not submit loose-leaf sheets of paper. Simply staple the pages together.

All submissions will be checked for plagiarism. **The University regards plagiarism as a form of academic misconduct, and has very strict rules regarding plagiarism. For UNSW policies, penalties, and information to help you avoid plagiarism see: https://student.unsw.edu.au/plagiarism as well as the guidelines in the online ELISE tutorials for all new UNSW students: http://subjectguides.library.unsw.edu.au/elise .**