STA 2101/442 Assignment 2[1]

Except for Questions 9 and 10, the questions on this assignment are practice for the quiz on Friday September 21st, and are not to be handed in.

1. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants' performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. Because it is familia, convenient and not too unreasonable, we adopt the model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

   - X is an $n \times p$ matrix of known constants with $n > p$ and the columns of X linearly independent.
   - $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
   - $\boldsymbol{\epsilon}$ is an $n \times 1$ random vector with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $cov(\boldsymbol{\epsilon}) = \sigma^2 I_n$.
   - $\sigma^2 > 0$ is an unknown constant.

   The least-squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1}X^\top \mathbf{y}$, the vector of predicted y values is $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$, and the vector of residuals is $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$. Give the dimensions (number of rows and number of columns) of the following matrices, specifically for the SAT data. The answers are all numbers.

   (a) $\mathbf{y}$

   (b) $\boldsymbol{\beta}$

   (c) $X\boldsymbol{\beta}$

   (d) $(X^\top X)^{-1}$

   (e) $\widehat{\boldsymbol{\beta}}$

   (f) $\widehat{\mathbf{y}}$

   (g) $\mathbf{e}$

   (h) $\mathbf{e}^\top\mathbf{e}$

   (i) $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top$

   (j) $X^\top\mathbf{e}$

2. For the general model of Question 1 (not just the specific SAT example) show $X^\top\mathbf{e} = \mathbf{0}$.

[1] This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/appliedf17

3. Why does $X^\top \mathbf{e} = \mathbf{0}$ tell you that if a regression model has an intercept, the residuals must add up to zero?

4. For the SAT data of Question 1, what null hypothesis would you test to answer the following questions? Give the answers in symbols.

   (a) Controlling for Math score, is Verbal score related to first-year grade point average?

   (b) Allowing for Verbal score, is Math score related to first-year grade point average?

   (c) Is either Verbal score or Math score (or both) related to first-year grade point average?

   (d) Does expected GPA increase faster as a function of the Verbal SAT, or the Math SAT?

5. It is well known that people who graduate from university have higher lifetime earnings on average than those who do not. Mention at least one confounding variable that could have produced this result.

6. Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain).

   (a) What is the explanatory variable? There is just one.

   (b) What is the response variable?

   (c) Is this an experimental study, or observational?

   (d) Why is it important for the patients to be unaware of which drug they are receiving? Relate this to the idea of a confounding variable.

   (e) Is it also important for the physicians to remain unaware of what drugs their patients are getting? Why or why not?

7. In a large telecommunications firm, sales representatives working at kiosks in shopping malls get bonuses for signing up more customers to more expensive cell phone plans. In addition to financial incentives, the company decides to require additional training for the sales reps who performed worst last quarter. To assess the effectiveness of the training, they will carry out a statistical test (the mysterious matched *t*-test, in fact) to see whether the average performance of this group increases.

   (a) Your first thought is "No! Don't do it this way!" Why? Plain language is not requested here. Just let me know that you understand the issue. Keep it short and sweet.

   (b) How would you recommend that they assess the effectiveness of the training program?

8. High School History classes from across Ontario are randomly assigned to either a discovery-oriented or a memory-oriented curriculum in Canadian history. At the end of the year, the students are given a standardized test and the median score of each class is recorded. Please consider a regression model with these variables:

   X_1 Equals 1 if the class uses the discovery-oriented curriculum, and equals 0 if the class uses the memory-oriented curriculum.

   X_2 Average parents' education for the classroom.

   X_3 Average family income for the classroom.

   X_4 Number of university History courses taken by the teacher.

   X_5 Teacher's final cumulative university grade point average.

   Y Class median score on the standardized history test.

   The full regression model (as opposed to the reduced models for various null hypotheses) implies

   $$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5.$$

   For each question below, please give

   - The null hypothesis in terms of β values.
   - $E[Y|X]$ for the reduced model you would use to answer the question. Don't re-number the variables.

   (a) If you allow for parents' education and income and for teacher's university background, does curriculum type affect test scores? (And why is it okay to use the word "affect?")

   (b) Controlling for parents' education and income and for curriculum type, is teacher's university background (two variables) related to their students' test performance?

   (c) Correcting for teacher's university background and for curriculum type, are parents' education and family income (considered simultaneously) related to students' test performance?

   (d) Taking curriculum type, teacher's university background and parents' education into consideration, is parents' income related to students' test performance?

   (e) Here is one final question. Assuming that $X_1, \ldots, X_5$ are random variables (and I hope you agree that they are),

      i. Would you expect X_1 ro be related to the other explanatory variables?

      ii. Would you expect the other explanatory variables to be related to each other?

9. The STA data described in Question 1 are available here. We seek to predict GPA from the two test scores. Throughout, please use the usual $\alpha = 0.05$ significance level.

   (a) First, fit a model using just the Math score as a predictor. "Fit" means estimate the model parameters. Does there appear to be a relationship between Math score and grade point average?

      i. Answer Yes or No.
      ii. Fill in the blank. Students who did better on the Math test tended to have ________ first-year grade point average.
      iii. Do you reject $H_0 : \beta_1 = 0$?
      iv. Are the results statistically significant? Answer Yes or No.
      v. What is the p-value? The answer can be found in *two* places on your printout.
      vi. What proportion of the variation in first-year grade point average is explained by score on the SAT Math test? The answer is a number from your printout.
      vii. Give a predicted first-year grade point average for a student who got 700 on the Math SAT. The answer is a number you could get with a calculator from your printout.

   (b) Now fit a model with both the Math and Verbal sub-tests.

      i. Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.
         A. $H_0 : \beta_1 = \beta_2 = 0$
         B. $H_0 : \beta_1 = 0$
         C. $H_0 : \beta_2 = 0$
         D. $H_0 : \beta_0 = 0$
      ii. Controlling for Math score, is Verbal score related to first-year grade point average?
         A. Give the value of the test statistic. The answer is a number from your printout.
         B. Give the p-value. The answer is a number from your printout.
         C. Do you reject the null hypothesis?
         D. Are the results statistically significant? Answer Yes or No.
         E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.
      iii. Allowing for Verbal score, is Math score related to first-year grade point average?
         A. Give the value of the test statistic. The answer is a number from your printout.
         B. Give the p-value. The answer is a number from your printout.
         C. Do you reject the null hypothesis?
         D. Are the results statistically significant? Answer Yes or No.
         E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iv. Give a predicted first-year grade point average for a student who got 650 on the Verbal and 700 on the Math SAT.

v. Let's do one more test. We want to know whether expected GPA increases faster as a function of the Verbal SAT, or the Math SAT. That is, we want to compare the regression coefficients, testing $H_0 : \beta_1 = \beta_2$.

A. Express the null hypothesis in matrix form as $\mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

B. Carry out an F test.

C. State your conclusion in plain, non-technical language. It's something about first-year grade point average.

Please bring your printout to the quiz.

10. Before the beginning of the Fall term, students in a first-year Calculus class took a diagnostic test with two parts: Pre-calculus and Calculus. Data are in the file `math1.data.txt`. The variables are

- Identification code
- Course: 1=Catch-up 2=Mainstream 3=Elite 4=NoResponse
- Score on pre-calculus part of diagnostic test
- Score on calculus part of diagnostic test
- High School GPA
- High School Calculus mark
- High School English mark
- University Calculus mark
- First language
- Sex

These are real data, with 776 missing values coded as NA. There are other rough edges, too; these have deliberately not been fixed. I'm not even sure if I remember what they all are, which makes this more interesting.

Your job is simple. Read the data and produce frequency distributions for the categorical variables and a table of means, standard deviations and sample sizes for the quantitative variables. But first, and as you do this, make a list of the strange things you notice about the data. Decide which of them should be fixed and which should be left alone. Fix the ones that you think should be fixed. Fix them in the way that makes most sense to you, so that the simple descriptive statistics you produce will be as meaningful as possible. All other things being equal, basing statistics on more data is better.

It is okay to discuss this with classmates. In fact, we will all discuss this, identify the problems, consider the solutions, and then I will decide what to do about each problem and let you know. As you will see, some of these decisions will be semi-arbitrary, but at least that way we should get the same numerical answers when we analyze the data — assuming we do it the same way.

Please bring your printout to the quiz.