STA 2201/442 Assignment 6

Please bring printouts for the `R` parts to the quiz. Some may be handed in. *Do not write anything on your printouts in advance except maybe your name (and student number, if you wish).*

1. In the following regression model, the independent variables X_1 and X_2 are measured without error, but X_2, which has an impact on Y and is correlated with X_1, is not part of the data set. The true model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$. The independent variables are random, and for simplicity we'll make them normal. Let

$$\begin{bmatrix} X_{i,1} \\ X_{i,2} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{bmatrix} \right)$$

with ϵ_i independent of $X_{i,1}$ and $X_{i,2}$.

Since X_2 is not observed, it is swallowed up into the intercept and error term, as follows.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The primes just denote a new β_0 and a new ϵ.

   (a) What is $Cov(X_{i,1}, \epsilon_i')$?

   (b) All we can observe are the pairs $(X_{i,1}, Y_i)$. Their distribution is bivariate normal. Calculate the mean and covariance matrix of $(X_{i,1}, Y_i)$ under the true model.

   (c) Suppose we want to estimate β_1. Remember, X_2 is not available, so you are doing a regression with one independent variable. The usual least squares estimator for this mis-specified model is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}$$

   Is $\widehat{\beta}_1$ a consistent estimator of β_1 for all points in the parameter space under the true model? Answer Yes or no and show your work.

   (d) Are there *any* points in the parameter space for which $\widehat{\beta}_1 \stackrel{p}{\rightarrow} \beta_1$ when the true model holds?

This problem shows that the assumption of zero covariance between independent variables and the error term is critical. So when we propose a regression model, we are assuming that any important independent variables omitted from the model have zero covariance with those that are in the model. Otherwise, the error term is correlated with the independent variables and everything falls apart.

2. In the standard regression model as it is usually presented, the independent variables are constants rather than random variables, and you might think that this means they must be independent of the error term. But think of the regression model as being *conditional* on $\mathbf{X} = \mathbf{x}$. This means that for every fixed $\mathbf{X} = \mathbf{x}$, $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. That is, the conditional distribution of $\boldsymbol{\epsilon}$ does not depend on $\mathbf{X}$. Pretending for convenience that the joint distribution of $\mathbf{X}$ is continuous, this means (with a slight abuse of notation)

$$f(\boldsymbol{\epsilon}|\mathbf{x}) = f(\boldsymbol{\epsilon}).$$

How does this imply that $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are independent?

3. The following is perhaps the simplest example of what happens to regression when there is measurement error in the independent variable. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= X_i\beta + \epsilon_i \\ W_i &= X_i + e_i, \end{aligned}$$

where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma^2_x$, $Var(\epsilon_i) = \sigma^2_\epsilon$, $Var(e_i) = \sigma^2_e$, and X_i, ϵ_i and e_i are all independent. Notice that W_i is just X_i plus a piece of random noise. This is a simple additive model of measurement error.

Unfortunately, we cannot observe the X_i values. All we can see are the pairs (W_i, Y_i) for $i = 1, \ldots, n$. So we do what everybody does, and fit the *naive* (mis-specified, wrong) model

$$Y_i = W_i\beta + \epsilon_i$$

and estimate β with the usual formula for regression through the origin. Where does $\widehat{\beta}_n$ go as $n \to \infty$? Show your work.

4. In the slides on residuals and prediction intervals, the t distribution arises frequently. In particular, consider the prediction interval for Y_{n+1}.

   (a) What is the distribution of $d_{n+1} = Y_{n+1} - \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}$? Show your work. Your answer includes both the expected value and the variance.

   (b) Based on your work in the preceding part, what is a reasonable expression for $s\{d_{n+1}\}$, the estimated standard deviation of d_{n+1}? You can check your answer using the handout on the Trees data.

   (c) Now show that

$$T = \frac{d_{n+1}}{s\{d_{n+1}\}}$$

has a t distribution, and give the degrees of freedom. You may use the following without proof[1]:

- $SSE/\sigma^2 \sim \chi^2(n-p)$
- SSE and $\widehat{\boldsymbol{\beta}}$ are independent
- If $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu).$$

5. For this question, you will use the file `sat.data` from Assignment 4. There is a link on the course web page in case the one in this document does not work. 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 and a 95% prediction interval for a student who got 700 on the Math SAT.

   (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 null hypothesis in symbols.

[1]The first two items are well within reach using the multivariate normal tools available for this course; it's a lot like deriving the joint distribution of $\overline{X}$ and S^2. The last item is a standard result.

B. Give the value of the test statistic. The answer is a number from your printout.

C. Give the p-value. The answer is a number from your printout.

D. Do you reject the null hypothesis?

E. Are the results statistically significant? Answer Yes or No.

F. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iii. Controlling for Verbal score, is Math score related to first-year grade point average?

A. Give the null hypothesis in symbols.

B. Give the value of the test statistic. The answer is a number from your printout.

C. Give the p-value. The answer is a number from your printout.

D. Do you reject the null hypothesis?

E. Are the results statistically significant? Answer Yes or No.

F. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iv. Math score explains _____ percent of the remaining variation in grade point average once you take Verbal score into account. Using the formula from the slides (which will be provided on the quiz if you need it), you should be able to calculate this from the output of the `summary` function. Check your answer using the `anova` function.

v. Verbal score explains _____ percent of the remaining variation in grade point average once you take Math score into account. Using the formula from the slides (which will be provided on the quiz if you need it), you should be able to calculate this from the output of the `summary` function. Check your answer using the `anova` function.

vi. Give a predicted first-year grade point average and a 95% prediction interval for a student who got 650 on the Verbal and 700 on the Math SAT. Are you confident that this student's first-year GPA will be above 2.0 (a C average)?

vii. 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 $\mathbf{C}\boldsymbol{\beta} = \mathbf{h}$. Obviously, this should be pretty routine.

B. But it's a bit more trouble than you'd think using `R`. I can think of three ways, all a little clumsy. Do the best you can. Carry out the test, producing an F statistic, degrees if freedom (a pair of numbers) and a p-value. Be able to state your conclusion in plain, non-technical language. it's something about first-year grade point average.

viii. Finally, we'll take a look at the Studentized deleted residuals. Suppose we want to carry out t-tests to locate outliers. What is the critical value? The answer is a number that you can get from `R`.

ix. How many Studentized deleted residuals are beyond the critical value? The answer is a number.

x. If the model were correct, how many Studentized deleted residuals would you expect to be beyond the critical value just by chance? The answer is a number.

xi. Make a normal QQ plot of the residuals (Studentized deleted or not). You might as well use the `qqnorm` and `qqline` functions for convenience. Do you see evidence of departure from normality? Describe any pattern you see.

xii. Plot residuals against variables in the equation. Do you see any need for a quadratic term? Any apparent outliers in 2 dimensions?

6. That SAT data set is a typical textbook example. It is very small and clean, and gives only a partial indication of what you need to do with real data. The *body fat* data are more realistic. The data are described in the file `bodyfat.txt`, and the first version of the data file is in `bodyfat1.data`. Both files are available from the course website, under the *Data sets* heading. You should probably print a copy of `bodyfat.txt`.

   The body fat data are very clean by my standards, but still there are some errors. Before analyzing any data set, we need to locate and correct all the errors we can. In practice, cleaning the data can have more effect on the results of a data analysis than choosing the best statistical method. This is all we'll do with the body fat data this week.

   The first thing you need to do with any new data set is look it over. Does it have the number of variables it's supposed to? Check the values of familiar variables like age (and height and weight, unless you live completely in the metric system). Do they make sense?

   The next step is to check the data for internal consistency. Notice that some variables are calculated from other variables. Personally, I prefer to do such calculations with software and I advise scientists not to include calculated variables in raw data sets. But it's a common practice, and provides a way of catching some errors.

   So, figure out which variables are calculated from other variables, do the calculation again with `R`, and compare the results. When I did this, my re-calculated variables had same number of decimal places of accuracy as the one in the data file. This way, I was able to just track down the cases where the difference is not zero. I made extensive use of the `cbind` function. Locate as many serious problems in the data that you can, the more serious the better. To me, differences of 0.1 between the original and re-calculated values are not important. For each problem you locate, be able to describe the inconsistency and suggest a solution.