STA 302f15 Assignment Nine[1]

Problems 1 and 2 are paper and pencil. They are preparation for the quiz in tutorial on Thursday November 12th, and are not to be handed in. Problem 3 uses R. Please bring your printout for Problem 3 to the quiz. Do not write anything on the printout in advance of the quiz, except possibly your name and student number.

1. The simple linear regression model is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are a random sample from a distribution with expected value zero and variance σ^2. The numbers $x_1, \ldots, x_n$ are known, observed constants, while the parameters β_0 β_1 and σ^2 are unknown constants (parameters). In a previous homework (Assignment 4), you obtained

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \qquad \text{and} \qquad \widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

   Show that for this model, R^2 is the square of the correlation coefficient

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}.$$

   You may use anything on the formula sheet.

2. This is a good test of whether you understand how t statistics are constructed. You may use the fact (a fact you have proved) that for a normal random sample,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

   Let $x_1, \ldots, x_{n_1} \stackrel{i.i.d.}{\sim} N(\mu_1, \sigma^2)$, and $y_1, \ldots, y_{n_2} \stackrel{i.i.d.}{\sim} N(\mu_2, \sigma^2)$. These two random samples are independent, meaning all the x variables are independent of all of the y variables.

   Every elementary Statistics text tells you that

$$T = \frac{\overline{x} - \overline{y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

   where

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (x_i - \overline{x})^2 + \sum_{i=1}^{n_2} (y_i - \overline{y})^2}{n_1 + n_2 - 2}$$

   This is the basis of tests and confidence intervals for $\mu_1 - \mu_2$.

   (a) Prove that T does indeed have the distribution claimed. Carefully cite material from the formula sheet when you use it. The word "independent" should appear in your answer at least *twice*.

   (b) Suppose you wanted to test $H_0 : \mu_1 = \mu_2$. Give a formula for the test statistic.

   (c) Derive a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$. "Derive" means show all the High School algebra.

[1] Copyright information is at the end of the last page.

3. The `statclass` data were used in Assignment 6. At the R prompt, type

```
statclass = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/LittleStatclassdata.txt")
```

You now have access to the `statclass` data. Fit a regression model in which the dependent variable is mark on the final exam, and the independent variables are Quiz Average, Computer Average, and mark on the Midterm test.

(a) What is the predicted Final Exam score for a student with a Quiz average of 8.5, a Computer average of 5, and a Midterm mark of 60%? The answer is a number. Be able to do this kind of thing on the quiz with a calculator from the output of `summary`.

(b) For any fixed Quiz Average and Computer Average, a score one point higher on the Midterm yields a predicted mark on the Final Exam that is _____ higher.

(c) For any fixed Quiz Average and Midterm score, an average one point higher on the Computer Average yields a predicted mark on the Final Exam that is _____ higher. Or is it lower?

(d) What is $\widehat{\beta}_3$? The answer is a number from your printout.

(e) For each of the following null hypotheses, give the value of the test statistic and the p-value. These are numbers from your printout. Also state whether you reject H_0 at $\alpha = 0.05$.

H_0	Test Statistic	p-value	Reject H_0?
$\beta_1 = \beta_2 = \beta_3 = 0$			
$\beta_0 = 0$			
$\beta_1 = 0$			
$\beta_2 = 0$			
$\beta_3 = 0$			

(f) For each of the following questions, give the null hypothesis you tested to answer the question, and also a conclusion expressed in plain, non-statistical language. Remember the rules: No statistical terminology, draw a directional conclusion if you can, be guided by $\alpha = 0.05$ but never mention it, and don't accept H_0.

i. Controlling for quiz average and computer average, is mark on the midterm test related to mark on the final exam?

ii. Allowing for mark on the midterm test and quiz average, is computer average a useful predictor of mark on the final exam?

iii. Taking into account mark on the midterm test and computer average, is quiz average related connected to mark on the final exam?

iv. Are any of the predictor variables useful?

(g) What proportion of the variation in final exam score is explained by the term work? The answer is a number from your printout.

(h) What is the largest $\widehat{\epsilon}_i$ in absolute value? The answer is on your printout.

(i) My printout has "`Residual standard error: 14.54`." What is this number?

(j) What is MSE? The answer is a number you can get with a calculator from your output.

(k) What is k for this problem? You can get it from the output of `summary`.

(l) What is n for this problem? You can calculate it from the output of `summary` without a calculator.

(m) What are the dimensions of the $\mathbf{X}$ matrix? The answer is a pair of numbers. You can calculate them from the output of `summary` without a calculator.

(n) What are the dimensions of the $\widehat{\boldsymbol{\beta}}$ matrix? The answer is a pair of numbers. You can obtain them from the output of `summary` without a calculator.

(o) What are the dimensions of the $\widehat{\boldsymbol{\epsilon}}$ matrix? The answer is a pair of numbers. You can get them from the output of `summary` without a calculator.

(p) What are the dimensions of $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$?

(q) What are the dimensions of the $\widehat{\mathbf{y}}$ matrix? The answer is a pair of numbers.

(r) What are the dimensions of the hat matrix $\mathbf{H}$? The answer is a pair of numbers.

(s) What is $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$? You can calculate this number from the output of `summary` using a calculator, if you know what `Residual standard error` is.

(t) What is SST? The answer is a single number. You can check your work with R, but calculate the number based just on the output of `summary` and the formula sheet. First show your work (there is some algebra), and then obtain the result with a calculator. Circle your final answer.

(u) The tests and confidence intervals based on the t distribution all use $t_{\alpha/2}$. By default we are using $\alpha = 0.05$, so $t_{\alpha/2}$ is the point cutting off the top 2.5% of the t distribution with $n - k - 1$ degrees of freedom. Obtain this number with R and make sure it is included in your printout.

(v) With a calculator (or using R as a calculator) calculate a 95% confidence interval for $\widehat{\beta}_3$. You can get the numbers you need from the output of `summary`. You don't need `vcov` for this one. You might want to refer to your answer to Question 6h from Assignment 8.

(w) For this question, first use the `attach` function to make the variables conveniently available for calculation. See the *Least squares with R* handout. Then calculate the means of all the independent variables. You might as well calculate $\overline{y}$ as well.

i. First, give a point estimate of $E(y|x_1 = \overline{x}_1, x_2 = \overline{x}_2, x_3 = \overline{x}_3)$. There is an easy way and a hard way. You decide: the easy way, the hard way, or both because you like to double-check everything.

ii. Give a 95% confidence interval for $E(y|x_1 = \overline{x}_1, x_2 = \overline{x}_2, x_3 = \overline{x}_3)$. For this, you will need to use `vcov`. Again, refer to your answer to Question 6h from Assignment 8. Your answer is a pair of numbers. You should do this with R and it should be on your printout.
