STA 302f20 Assignment Eight[1]

The following problems are not to be handed in. They are preparation for the quiz in tutorial and the final exam. Please try them before looking at the answers. Use the formula sheet. Please remember that the R part (Question 2) is *not a group project*. You may compare numerical answers, but do not show anyone your code or look at anyone else's.

1. Data from a STA302 class many years ago consist of quiz average, computer assignment average, midterm score and Final Exam score, all in percent. We seek to predict final exam score from the term work.

   (a) Write the regression equation in scalar form using $x_{i,j}$ and y_i variables. Assume the order of predictor variables given above.

   (b) What is the expected final exam score for a student with a 70% average on the quizzes, 85% on the computer assignments, and 65% on the midterm? Answer in terms of β_j values.

   (c) For any fixed quiz average and computer average, a score one point higher on the midterm yields an expected mark on the Final Exam that is _____ higher.

   (d) We want a hypothesis test to answer this question: Are any of the term work variables useful in predicting final exam score? This is one test.

      i. State the null hypothesis in terms of scalar β_j values.

      ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

      iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model.

   (e) Controlling for computer assignment average and midterm score, is quiz average related to Final Exam score?

      i. State the null hypothesis in scalar form.

      ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

      iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model. Do not renumber the explanatory variables or regression coefficients.

   (f) Allowing for quiz average and computer assignment average, is midterm score a predictor of Final Exam score?

      i. State the null hypothesis in terms of scalar β_j values.

      ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

      iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model. Do not renumber the explanatory variables or regression coefficients.

(g) Holding for quiz average and midterm score fixed, is computer assignment average connected to Final Exam score?

i. State the null hypothesis in terms of scalar β_j values.

ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model. Do not renumber the explanatory variables or regression coefficients.

(h) Controlling for computer assignment average, is quiz average or midterm score (or both) related to Final Exam score? This is one test.

i. State the null hypothesis in terms of scalar β_j values.

ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model. Do not renumber the explanatory variables or regression coefficients.

(i) The professor thinks that the quizzes and midterm should have equal weight, and should be worth twice as much as the computer assignments. If this idea is correct, it should be reflected in the relationship of the term marks to the final exam. Also, it makes sense that if a student got zero on all three components of the term mark, he or she should also expect a zero on the final exam — even though this extreme case is outside the range of the data. Taken together, these ideas represent an unusual but testable null hypothesis. If it is rejected, we could say that the professor's ideas are not supported by the data.

i. State the null hypothesis in terms of scalar β_j values.

ii. The null hypothesis could be written $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Give the $\mathbf{C}$ and $\mathbf{t}$ matrices.

iii. The null hypothesis could be tested using the full-reduced model approach. Give the regression equation for the reduced model. Do not renumber the explanatory variables or regression coefficients.

2. Data from the example of Question 1 are available here. The URL is `http://www.utstat.utoronto.ca/~brunner/data/legal/LittleStatclassdata.txt` . In the data file, the quiz averages and computer averages are out of ten, but you should fix them up so they are out of 100. Prepare one pdf document showing your input and output for the questions below. You might be asked to attach it to the quiz.

   (a) Start with a `summary` of the data frame, a correlation matrix, and a matrix of scatterplots using `pairs`, just so you have a general idea of what is going on.

      i. What is the median score on the computer assignments? The answer is a number.

      ii. What is the correlation between the computer average and the final exam score? The answer is a number.

      iii. You multiplied computer average and quiz average by 10 to convert to percent. Does this affect their correlations with other variables? Answer Yes or No and prove your answer.

      iv. Suppose you were to fit a simple regression model with quiz average as the single explanatory variable and final exam score as the response variable. Without actually fitting the model, what would R^2 be? The answer is a number.

   (b) Fit the full model (your answer to Question 1a) and display `summary` on it. The answers to many of the questions below are in the output of `summary`, or can be obtained from the `summary` output by a quick calculation.

   (c) What is $\widehat{\beta}_2$? The answer is a number.

   (d) What is n for this problem?

   (e) What is k for this problem?

   (f) What are the dimensions of the $\mathbf{X}$ matrix? The answer is a pair of numbers, number of rows and number of columns.

   (g) What are the dimensions of $\widehat{\boldsymbol{\beta}}$? The answer is a pair of numbers, number of rows and number of columns.

   (h) What are the dimensions of $\widehat{\boldsymbol{\epsilon}}$? The answer is a pair of numbers, number of rows and number of columns.

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

   (j) What are the dimensions of the $\widehat{\mathbf{y}}$ matrix? The answer is a pair of numbers, number of rows and number of columns.

   (k) What are the dimensions of the hat matrix $\mathbf{H}$? The answer is a pair of numbers, number of rows and number of columns.

   (l) What is $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$? You can calculate this number from the output of `summary` employing R as a calculator, using the fact that `Residual standard error` from your printout is the square root of *MSE*. The answer is subject to a bit of rounding error, but it's okay. Get a more accurate answer, too.

   (m) What is SST? The answer is a single number. First, obtain the number from the output of `summary`, using R as a calculator. There is some algebra; show your work. Then, check the result by a more direct calculation. I used the `var` function. The second way is more accurate, but the first one is more interesting.

(n) What is the predicted final exam score for a student with a 70% average on the quizzes, 85% on the computer assignments, and 65% on the midterm? The answer is a number.

(o) 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. The answer is a number.

(p) What is the largest $\widehat{\epsilon}_i$ in absolute value?

(q) For each of the following null hypotheses, give the value of the test statistic and the p-value. The answers are numbers that appear in the output from `summary`. Also state whether you reject H_0 at $\alpha = 0.05$.

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

(r) What proportion of the variation (sum of squares) in final exam mark is explained by the term work? The answer is a number.

(s) There is a hypothesis test to answer the question: Controlling for computer assignment average and midterm score, is quiz average related to Final Exam score?

i. State the null hypothesis in symbols.

ii. A nice 2-sided t-test is part of the default output. Give the value of the t statistic, the degrees of freedom, and the p-value.

iii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

iv. Are the results statistically significant at the $\alpha = 0.05$ level? Answer Yes or No.

v. In plain, non-statistical language, what do you conclude?

vi. You can test this same null hypothesis in the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$, using the general linear F-test. Do it using my `ftest` function. Does $F = t^2$? Compare the p-values.

vii. Carry out the same test using the full-reduced model approach.

viii. Once you have allowed for computer assignment average and midterm score, what proportion of the remaining variation does quiz average explain? The answer is a number between zero and one.

(t) Give a 95% confidence interval for β_1. Why does this confidence interval provide one more way of testing $H_0 : \beta_1 = 0$?

(u) Consider a student who is "average" on all three explanatory variables. The expected final exam score for such a student would be

$$E(y|x_1 = \overline{x}_1, x_2 = \overline{x}_2, x_3 = \overline{x}_3) = \beta_0 + \beta_1\overline{x}_1 + \beta_2\overline{x}_2 + \beta_3\overline{x}_3,$$

where the $\overline{x}_j$ are the sample means of quiz average, computer average and midterm test.

i. Give a point estimate of the expected value. The answer is a number.

ii. You knew it was equal to $\overline{y}$ all along. Why?

iii. Give a 95% confidence interval for the expected value. The answer is a pair of numbers, a lower confidence limit and an upper confidence limit.

iv. Use the `t.test` function (see `help(t.test)`) to get the usual 95% confidence interval around $\overline{y}$. The two confidence intervals are a little bit different. Why?

(v) 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. All the information you need is in the output of `summary` from the full model.

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?

(w) For Question 2v, suppose we treated the last test (Are any of the predictor variables useful?) as an overall test, and treated the other tests as follow-ups with a Bonferroni correction. What is the conclusion now? Only mention findings that are statistically significant with the Bonferroni correction.

(x) Controlling for mark on the midterm test, are the other two variables (either or both) related to mark on the Final exam?

i. State the null hypothesis in terms of scalar β_j values.

ii. State the null hypothesis in matrix terms. That is, give the matrices $\mathbf{C}$, $\boldsymbol{\beta}$ and $\mathbf{t}$ in $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

iii. Write the reduced model. Please do not re-number the variables or the β_j parameters.

iv. Obtain the F^* test statistic my `ftest` function.

v. Carry out the same test using the full-reduced model approach.

vi. Give the p-value. The answer is a number.

vii. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

viii. Are the results statistically significant at the $\alpha = 0.05$ level? Answer Yes or No.

ix. Allowing for mark on the midterm test, what proportion of the remaining variation in final exam score is explained by computer average and quiz average?

x. State your conclusions (if any) in plain, non-statistical language.