STA 302f15 Assignment Eleven[1]

Problem 6 uses R. Please bring your printout for Problem 6 to the quiz. **Do not write anything on the printout in advance of the quiz, except possibly your name and student number.** The other questions are preparation for the quiz, and are not to be handed in.

1. For the general linear model with normal errors,

   (a) What is the distribution of $\mathbf{C}\widehat{\boldsymbol{\beta}}$? Note $\mathbf{C}$ is $q \times (k = 0, 1)$, and the rows of $\mathbf{C}$ are linearly independent.

   (b) If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is true, what is the distribution of $\frac{1}{\sigma^2}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})$? Please locate support for your answer on the formula sheet. For full marks, don't forget the degrees of freedom.

   (c) What other facts on the formula sheet allow you to establish the F distribution for the general linear test? The distribution is *given* on the formula sheet, so of course you can't use that. In particular, how do you know numerator and denominator are independent?

2. Suppose you wish to test the null hypothesis that a *single* linear combination of regression coefficients is equal to zero. That is, you want to test $H_0 : \mathbf{a}'\boldsymbol{\beta} = 0$. Referring to the formula sheet, verify that $F = T^2$. Show your work.

3. Starting from the formula sheet, show that the F test for comparing full and reduced models may be written

$$F = \left(\frac{p}{1-p}\right)\left(\frac{n-k-1}{q}\right),$$

   where $p = \frac{R^2(full)-R^2(reduced)}{1-R^2(reduced)}$. Show your work. It may help to compare SST from the full model to SST from the reduced model before you begin the calculation.

4. That quantity denoted by p in the last question has a useful interpretation. It's the proportion of *remaining* variation in the dependent variable that is explained when the independent variables in the second set are added to the model. That is, the variables in the reduced model explain $R^2(reduced)$, so they fail to explain $1 - R^2(reduced)$. Then the variables in the second set are added to the reduced model, yielding the full model — and R^2 goes up. The quantity p expresses this improvement as a proportion of what improvement was possible. Of course it is not the same as the p-value.

   Derive another formula for p, writing p in terms of F, n, k and q. Show your work. This formula can give an idea of how strong a set of results is, when all you are given is an F or t statistic and the degrees of freedom. The answer is on the formula sheet; prove it.

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

5. In an extended version of the SAT data, the dependent variable is first-year university Grade Point Average (GPA) again. The independent variables are

   $x_1 =$ Verbal SAT score

   $x_2 =$ Math SAT score

   $x_3 =$ High school Grade Point Average

   $x_4 =$ Mother's education, in years

   $x_5 =$ Father's education, in years

   $x_6 =$ Total family income,

   and also Location of the family home: City, Suburbs or Country.

   (a) First, write the regression equation. It is up to you which dummy variable variable scheme you use, as long as the regression planes are parallel.

   (b) Make a table with one row for each location of the family home, showing how your dummy variables are defined. Make one more column showing $E(y|\mathbf{x})$ for each location.

   (c) For each of the following questions, do three things: Give the null hypothesis in the form of a statement about the β values, Give the $\mathbf{C}$ and $\mathbf{t}$ matrices in $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$, and Give $E(y|\mathbf{x})$ for the reduced model (note that expected y for the full model is always the same).

      i. Correcting for all other variables, is location of the family home related to first-year GPA?

      ii. Controlling for all other variables, is either Verbal SAT score or Math SAT score (or both) related to GPA?

      iii. When you allow for all the other variables, is family income a useful predictor of GPA?

      iv. Controlling for all other variables, does expected GPA change faster as a function of Verbal SAT, or does it change faster as a function of Math SAT?

      v. Once you correct for the two SAT scores and High School marks, do any of the family variables matter?

      vi. Correcting for all other variables, does expected GPA change faster as a function of Mother's education, or does it change faster as a function of father's education?

      vii. Holding all the other variables constant at fixed values, is Math SAT related to first-year university GPA?

      viii. Once you allow for location of the family home, do any of the other predictors matter?

6. This is an expanded version of the `statclass` data used in Assignments 6 and 9. At the R prompt, type

```
statclass = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/LittleStatClassData2.txt")
```

Do `head(statclass)` to see what's there. Fit a regression model in which the dependent variable is mark on the Final Exam, and the independent variables are Sex, Race, Quiz Average, Computer Average, and mark on the Midterm test.

(a) What is the predicted Final Exam score for a Male student from Race A 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) Obtain a 95% prediction interval for the student described in the previous question. Do it the easy way.

(c) Controlling for the other independent variables, are there any differences in the average performance of students from the different racial groups?

i. Please test the main hypothesis 2 different ways: using the general linear test, and also the full-reduced approach. Are your F-statistics the same?

ii. What proportion of the variation in Final Exam mark is explained by the independent variables in the reduced model? The answer is a number from your printout.

iii. What proportion of the variation in Final Exam mark is explained by the independent variables in the full model? The answer is a number from your printout.

iv. After allowing for the other independent variables, what proportion of the *remaining* variation in Final Exam score is explained by race? For reference, we are talking about the quantity p from Question 4. There are two formulas for p on the formula sheet. Get both of them from the numbers on your printout.

v. Three comparisons are of interest, race A versus B, A versus C and B versus C. You already have two of them. Do the third one. Now you have three test statistics and associated p-values.

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

(d) Allowing for other predictors, is the student's sex related to mark on the Final Exam? Give the test statistic, the p-value, and answer the question in plain, non-statistical language.

(e) When other variables are held constant, an increase of one point (out of 10) in the Quiz Average results in an increase of _____ in predicted mark on the Final Exam.

(f) Controlling for other independent variables in the model, is Quiz Average a useful predictor of mark on the Final Exam? Give the test statistic, the p-value, and answer the question in plain, non-statistical language. For full marks, give a directional conclusion if possible. You will not be reminded of this on our Final Exam.

(g) After allowing for the other independent variables, what proportion of the *remaining* variation in Final Exam score is explained by Quiz Average? This is a number you could obtain with a calculator from the `summary` output of the full model. There is no need to literally fit the reduced model, though you may check your answer that way if you wish. If you are stuck, look at Question 2.
