STA 302f14 Assignment Seven[1]

See the formula sheet. The formula sheet will be provided with the quiz. You may use anything from the formula sheet unless you are explicitly asked to prove it, or are instructed otherwise.

1. Show that for a simple regression with an intercept and one independent variable, $R^2 = r^2$, where r is the ordinary correlation coefficient given on the formula sheet. You may use the formulas $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$ and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x}$, which you derived back in Assignment Four. It helps to start with the formula for R^2 from the formula sheet, and then substitute for $\widehat{Y}_i$ right away.

2. Suppose you fit (estimate the parameters of) a regression model, obtaining $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{Y}}$ and $\widehat{\boldsymbol{\epsilon}}$. Call this Model One.

   (a) Then just for fun, you fit a second regression model, using $\widehat{\mathbf{Y}}$ from Model One as the dependent variable, and exactly the same $\mathbf{X}$ matrix as Model One. Call this Model Two.

      i. What is $\widehat{\boldsymbol{\beta}}$ for Model Two? Show your work and simplify.
      ii. What is $\widehat{\mathbf{Y}}$ for Model Two? Show your work and simplify.
      iii. What is $\widehat{\boldsymbol{\epsilon}}$ for Model Two? Show your work and simplify.
      iv. What is MSE for Model Two?

   (b) Now you fit a *third* regression model, this time using $\widehat{\boldsymbol{\epsilon}}$ from Model One as the dependent variable, and again, exactly the same $\mathbf{X}$ matrix as Model One. Call this Model Three.

      i. What is $\widehat{\boldsymbol{\beta}}$ for Model Three? Show your work and simplify.
      ii. What is $\widehat{\mathbf{Y}}$ for Model Three? Show your work and simplify.
      iii. What is $\widehat{\boldsymbol{\epsilon}}$ for Model Three? Show your work and simplify.

3. Consider a linear regression model with $n > k+1$, which is always the case in practice. Since $\widehat{\boldsymbol{\epsilon}} \sim N\left(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H})\right)$, it is tempting to write $\frac{1}{\sigma^2}\widehat{\boldsymbol{\epsilon}}'(\mathbf{I}-\mathbf{H})^{-1}\widehat{\boldsymbol{\epsilon}} \sim \chi^2(n)$. Please locate support for this idea on the formula sheet. But it only works if the $n \times n$ matrix $\mathbf{I}-\mathbf{H}$ has an inverse.

   (a) Look again at the brief discussion of rank in the "More linear algebra" slide show. How do you know that the hat matrix $\mathbf{H}$ has no inverse?

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

(b) But it's not so obvious for $\mathbf{I} - \mathbf{H}$.

i. Calculate $(\mathbf{I} - \mathbf{H})\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

ii. At this point, maybe you suspect that the columns of $\mathbf{I} - \mathbf{H}$ must be linearly dependent so the inverse can't exist, but for a conclusive demonstration assume that $(\mathbf{I} - \mathbf{H})^{-1}$ *does* exist, and arrive at an impossible conclusion.

4. Based on the general linear model with normal error terms,

(a) Prove the t distribution given on the formula sheet for a new observation Y_0. Use earlier material on the formula sheet. For example, how do you know numerator and denominator are independent?

(b) Derive the $(1 - \alpha) \times 100\%$ prediction interval for a new observation from this population, in which the independent variable values are given in $\mathbf{x}_0$. "Derive" means show the High School algebra.

5. In an extended version of the SAT data, 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

The dependent variable is first-year university Grade Point Average (GPA) again. For each of the following questions, give the null hypothesis in the form of a statement about the β values, and then give the $\mathbf{C}$ and $\mathbf{t}$ matrices in $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

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

(b) When you allow for all the other variables, is family income a useful predictor of GPA?

(c) 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?

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

(e) 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?

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

6. For each part of Question 5, Give $E(Y)$ for the reduced model, and give $E(Y)$ for the full model.

7. For the general linear model (see formula sheet),

   (a) What is the distribution of $\mathbf{C}\widehat{\boldsymbol{\beta}}$? Note $\mathbf{C}$ is $q \times (k+1)$.

   (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?

8. 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.

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

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

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

10. That quantity denoted by a 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_r^2, so they fail to explain $1 - R_r^2$. Then the variables in the second set are added to the reduced model, yielding the full model — and R^2 goes up. The quantity a expresses this improvement as a proportion of what improvement was possible.

    Derive a formula for a, writing a 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. After this assignment, it will be on the formula sheet.

11. This question uses the data file `CensusTract.data` from the last assignment. Start with the model in which the dependent variable is crime rate, and the independent variables are `area`, `urban`, `old`, `docs`, `beds`, `hs`, `labor` and `income`.

    (a) According the the t-tests, the independent variables `old`, `labor` and `income` don't appear to be doing much. Test them simultaneously, the easiest way you can. Your R printout will include an F statistic, degrees of freedom and p-value. What do you conclude? Is there a case for dropping these variables from the model?

(b) Do an F-test for percent of high school graduates, controlling for all other variables. Again, do it the easiest way you can. Compare the p-value to that of the the t-test. Does $F = T^2$? Are the test statistics (the specific numbers) equally informative? If not, which one tells you more?

(c) Holding all other independent variables constant at fixed values, estimate the amount by which the crime rate changes when the percent of adults in a census tract who are High School graduates is increased by one. The answer is a number in the default output from the `summary` function.

(d) A confidence interval is the estimate plus or minus a margin of error. Give the 95% margin of error for the estimate in the last question. Your answer is a number. Calculate it with R, but realize that *except for the critical value, everything you need is part of your default output*, and you could do this with a calculator on a quiz or final exam if you had the critical value.

(e) Estimate the expected crime rate for a census tract with an area of 2,500 square miles, 50 percent urban, 10 percent senior citizens, 2,000 doctors, 6,000 hospital beds, 50 percent finished high school, a labour force of 450 thousand, and a total income of 6,500 million dollars. Give both a predicted value (a single number that you could get from the default output with a calculator) and a 95% confidence interval. Do it the easiest way you can.

(f) Predict the crime rate for a *new* census tract with an area of 2,500 square miles, 50 percent urban, 10 percent senior citizens, 2,000 doctors, 6,000 hospital beds, 50 percent finished high school, a labour force of 450 thousand, and a total income of 6,500 million dollars. Give both a predicted value (a single number) and a 95% prediction interval. Do it the easiest way you can.

Bring your printout to the quiz.
