# STA 302f13 Assignment Eight[1]

This assignment assumes you are using the new Formula sheet. There is a link on the course home page in case the one in this document does not work. The formula sheet (or part of it) will be provided with the quiz.

1. 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 $\beta$ 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?

2. For each part of Question 1, Give $E(Y)$ for the reduced model, and give $E(Y)$ for the full model.

3. 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 $\frac{1}{\sigma^2}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})$?

   (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?

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

4. Suppose you need 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.

5. 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. You may use $SST = SSR + SSE$ and $R^2 = \frac{SSR}{SST}$, which are not on the formula sheet (yet).

6. 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.

7. This question uses the data file `CensusTract.data` from Assignment 7. There is a link on the course home page in case the one in this document does not work. 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 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?

---