STA 302f16 Assignment Seven[1]

These problems are preparation for the quiz in tutorial on Thursday November 3d, and are not to be handed in.

1. This question exercises your understanding of 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 $x_1, \ldots, x_{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$, and $y_1, \ldots, y_{n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$. 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.

2. For the general linear model with normal errors,

   (a) Let C be an $m \times (k+1)$ matrix of constants with linearly independent rows.What is the distribution of $C\mathbf{b}$?

   (b) If $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$ is true, what is the distribution of $\frac{1}{\sigma^2}(C\mathbf{b}-\boldsymbol{\gamma})'(C(\mathbf{X}'\mathbf{X})^{-1}C')^{-1}(C\mathbf{b}-\boldsymbol{\gamma})$? 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?

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

3. 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 $H_0 : \boldsymbol{\ell}'\boldsymbol{\beta} = 0$. Referring to the formula sheet, verify that $F = t^2$. Show your work.

4. The exact way that you express a linear null hypothesis does not matter. Let A be an $m \times m$ nonsingular matrix (meaning A^{-1} exists), so that $C\boldsymbol{\beta} = \boldsymbol{\gamma}$ if and only if $AC\boldsymbol{\beta} = A\boldsymbol{\gamma}$. This is a useful way to express a logically equivalent null hypothesis, because any matrix that is row equivalent to C can be written as AC. Show that the general linear test statistic F for testing $H_0 : (AC)\boldsymbol{\beta} = A\boldsymbol{\gamma}$ is the same as the one for testing $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

5. 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, you obtained the least squares estimates

$$b_0 = \overline{y} - b_1\overline{x} \qquad \text{and} \qquad b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

while the correlation coefficient r is given by

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

   (a) Start by writing b_1 as a function of r.

   (b) Now show that for this model, R^2 is the square of the correlation coefficient. Start with $R^2 = \frac{SSR}{SST} = \cdots$.

6. Show that for a multiple regression model with an intercept,

$$F = \frac{SSR_F - SSR_R}{ms^2} = \left(\frac{n-k-1}{m}\right)\left(\frac{a}{1-a}\right),$$

where $a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$ and $s^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$ from the full model. Show your work. It may help to compare SST from the full model to SST from the reduced model before you begin the calculation.

7. 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^2_R, so they fail to explain $1 - R^2_R$. 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 another formula for a, writing a in terms of F, n, k and m. 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; you are being asked to prove it.

8. In Assignment 6, Questions 14 through 18 developed an F-test for $H_0 : \beta_1 = \cdots = \beta_k = 0$. Write the regression equation for the reduced model in scalar form.

9. In Assignment 5, you fit a regression model to data from a statistics class (actually, STA302) many years ago. Here you will just answer questions about some possible tests without actually doing them yet. Recall that the data consist of Quiz average, Computer assignment average, Midterm score and Final Exam score.

   (a) Write the regression equation in scalar form. The dependent variable is Final Exam score; keep the independent variables in the order given above.

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

      i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

      ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

      iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.

   (c) Holding quiz average and midterm score constant, is computer assignment average related to final exam score?

      i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

      ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

      iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.

(d) Correcting for performance on the quizzes and computer assignments, is performance on the midterm related to performance on the final exam?

i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.

(e) We want to test computer average and quiz average simultaneously, allowing for score on the midterm test.

i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.

(f) It is claimed that once you correct for quiz average, neither computer average nor score on the midterm test is a useful predictor of score on the final exam. Test this proposition.

i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.

(g) Are *any* of the independent variables related to the dependent variable?

i. Give the null hypothesis in scalar form; this is a statement or collection of statements about the β_j values.

ii. Write the null hypothesis in matrix form as $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Write the regression equation for the *reduced model* in scalar form. Don't re-number the independent variables or β_js.
