STA 2101/442 Assignment Five[1]

The questions are just practice for the quiz, and are not to be handed in. Use R for Questions 15 and 17, and bring two separate printouts to the quiz. **Your printouts should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.

1. Let $Y_1, \ldots, Y_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The sample variance is $S^2 = \frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{n-1}$.

   (a) Show $Cov(\overline{Y}, (Y_j - \overline{Y})) = 0$ for any $j = 1, \ldots, n$.

   (b) How do you know that $\overline{Y}$ and S^2 are independent?

   (c) Show that
   $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$
   Hint: $\sum_{i=1}^{n}(Y_i - \mu)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y} + \overline{Y} - \mu\right)^2 = \ldots$

2. Recall the definition of the t distribution. If $Z \sim N(0,1)$, $W \sim \chi^2(\nu)$ and Z and W are independent, then $T = \frac{Z}{\sqrt{W/\nu}}$ is said to have a t distribution with ν degrees of freedom, and we write $W \sim t(\nu)$. As in the last question, let $Y_1, \ldots, Y_n$ be random sample from a $N(\mu, \sigma^2)$ distribution. Show that $T = \frac{\sqrt{n}(\overline{Y}-\mu)}{S} \sim t(n-1)$. The key is to locate a Z for the numerator and a W for the denominator.

3. For the general fixed effects linear regression model in matrix form (see formula sheet), show that the $n \times p$ matrix of covariances $C(\mathbf{e}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$. Why does this establish that SSE and $\widehat{\boldsymbol{\beta}}$ are independent?

4. Last week you showed that $(\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{e}^\top\mathbf{e} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(X^\top X)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Dividing both sides by σ^2, show that $\mathbf{e}^\top\mathbf{e}/\sigma^2 \sim \chi^2(n-p)$. Start with the distribution of the left side.

5. Tests and confidence intervals for linear combinations of regression coefficients are very useful. Derive the appropriate t distribution and some applications by following these steps. Let $\mathbf{a}$ be a $p \times 1$ vector of constants.

   (a) What is the distribution of $\mathbf{a}^\top\widehat{\boldsymbol{\beta}}$? Show a little work. Your answer includes both the expected value and the variance.

   (b) Now standardize the difference (subtract off the mean and divide by the standard deviation) to obtain a standard normal.

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

(c) Divide by the square root of a well-chosen chi-squared random variable, divided by its degrees of freedom, and simplify. Call the result T.

(d) How do you know numerator and denominator are independent?

(e) Suppose you wanted to test $H_0 : \mathbf{a}^\top \boldsymbol{\beta} = c$. Write down a formula for the test statistic.

(f) Suppose you wanted to test $H_0 : \beta_2 = 0$. Give the vector $\mathbf{a}$.

(g) Suppose you wanted to test $H_0 : \beta_1 = \beta_2$. Give the vector $\mathbf{a}$.

(h) Letting $t_{\alpha/2}$ denote the point cutting off the top $\alpha/2$ of the t distribution with $n-p$ degrees of freedom, give a $(1-\alpha) \times 100\%$ confidence interval for $\mathbf{a}^\top \boldsymbol{\beta}$.

6. In this question you will develop a *prediction interval* (not a confidence interval) for Y_{n+1}.

(a) What is the distribution of $Y_{n+1} - \widehat{Y}_{n+1} = Y_{n+1} - \mathbf{x}_{n+1}^\top \widehat{\boldsymbol{\beta}}$? Show your work. Your answer includes both the expected value and the variance.

(b) Now standardize the difference to obtain a standard normal.

(c) Divide by the square root of a chi-squared random variable, divided by its degrees of freedom, and simplify. Call it T. Compare your answer to a slide from lecture. How do you know that numerator and denominator are independent?

(d) Using your result, derive the $(1-\alpha) \times 100\%$ prediction interval for Y_{n+1}.

7. In this question you will establish the F distribution for the general linear test. For the general linear model (see formula sheet),

(a) What is the distribution of $\mathbf{L}\widehat{\boldsymbol{\beta}}$? Note $\mathbf{L}$ is $r \times p$.

(b) If $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ is true, what is the distribution of $\frac{1}{\sigma^2}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})^\top(\mathbf{L}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{L}^\top)^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})$? 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. 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 $H_0 : \mathbf{a}^\top \boldsymbol{\beta} = 0$. Referring to earlier questions or the formula sheet, verify that $F = T^2$. Show your work.

9. For the general linear regression model with normal error terms, show that if the model has an intercept, $\mathbf{e}$ and $\overline{y}$ are independent. Here are some ingredients to start you out. For the model with intercept,

   (a) What does $X'\mathbf{e} = \mathbf{0}$ tell you about $\sum_{i=1}^{n} e_i$?

   (b) Therefore what do you know about $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} \widehat{y}_i$?

   (c) Show that the least squares plane must pass through the point $(\overline{x}_1, , \ldots \overline{x}_{p-1}, \overline{y})$.

   (d) Now show that $\mathbf{e}$ and $\overline{y}$ are independent.

10. Continue assuming that the regression model has an intercept. Many statistical programs automatically provide an *overall* test that says none of the independent variables makes any difference. If you can't reject that, you're in trouble. If $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ is true,

    (a) What is the distribution of Y_i?

    (b) What is the distribution of $\frac{SST}{\sigma^2}$? Just write down the answer. Check Problem 1.

11. Still assuming $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ is true, find the distribution of SSR/σ^2? Use the formula sheet and show your work.

12. Recall the definition of the F distribution. If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ are independent, $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$. Show that $F = \frac{SSR/(p-1)}{SSE/(n-p)}$ has an F distribution under $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$? Refer to the results of questions above as you use them.

13. The null hypothesis $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ is less and less believable as R^2 becomes larger. Show that the F statistic of Question 12 is an increasing function of R^2 for fixed n and k. This mean it makes sense to reject H_0 for large values of F.

14. When you fit a full and a reduced regression model, the proportion of remaining variation explained by the additional variables in the full model is $a = \frac{R^2_F - R^2_R}{1 - R^2_R}$.

    (a) Show
    $$F = \frac{(SSR_F - SSR_R)/r}{MSE_F} = \left(\frac{n-p}{r}\right)\left(\frac{a}{1-a}\right).$$

    (b) Show $a = \frac{rF}{n-p+rF}$. This means that you can calculate the proportion of remaining variation for any F or t-test without explicitly fitting a reduced model. All you need is a calculator.

15. For a regression model with nine explanatory variables, you want the test of $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ to be statistically significant at the $\alpha = 0.05$ level provided that the variables x_1 through x_4 explain at least 3% of the remaining variation. What sample size is required? The answer is a number from your R printout.

16. In the usual univariate multiple regression model, the $\mathbf{X}$ is an $n \times p$ matrix of known constants. But of course in practice, the explanatory variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathbf{X} = \mathbf{x}$.

    (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is conditionally unbiased.

    (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally. Use double expectation.

    (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an F-test comparing full and reduced models), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory variable values. That is, $P(F > f_c|\mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. Assume that the explanatory variables are discrete, so you can write a multiple sum.

17. For this question, you will use the `sat.data` again. Get the data with

```
sat = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/openSAT.data.txt").
```

    We seek to predict GPA from the two test scores. Throughout, please use the usual $\alpha = 0.05$ significance level.

    (a) First, fit a model using just the Math score as a predictor. "Fit" means estimate the model parameters. Does there appear to be a relationship between Math score and grade point average?

        i. Answer Yes or No.

        ii. Fill in the blank. Students who did better on the Math test tended to have _____ first-year grade point average.

        iii. Do you reject $H_0 : \beta_1 = 0$?

        iv. Are the results statistically significant? Answer Yes or No.

        v. What is the p-value? The answer can be found in *two* places on your printout.

        vi. What proportion of the variation in first-year grade point average is explained by score on the SAT Math test? The answer is a number from your printout.

        vii. Give a predicted first-year grade point average and a 95% prediction interval for a student who got 700 on the Math SAT.

(b) Now fit a model with both the Math and Verbal sub-tests.

i. Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.

A. $H_0 : \beta_1 = \beta_2 = 0$

B. $H_0 : \beta_1 = 0$

C. $H_0 : \beta_2 = 0$

D. $H_0 : \beta_0 = 0$

ii. Controlling for Math score, is Verbal score related to first-year grade point average?

A. Give the null hypothesis in symbols.

B. Give the value of the test statistic. The answer is a number from your printout.

C. Give the p-value. The answer is a number from your printout.

D. Do you reject the null hypothesis?

E. Are the results statistically significant? Answer Yes or No.

F. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iii. Controlling for Verbal score, is Math score related to first-year grade point average?

A. Give the null hypothesis in symbols.

B. Give the value of the test statistic. The answer is a number from your printout.

C. Give the p-value. The answer is a number from your printout.

D. Do you reject the null hypothesis?

E. Are the results statistically significant? Answer Yes or No.

F. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iv. Math score explains _____ percent of the remaining variation in grade point average once you take Verbal score into account. Using the formula from the slides (see formula sheet), you should be able to calculate this from the output of the `summary` function. You can check your answer using the `anova` function.

v. Verbal score explains _____ percent of the remaining variation in grade point average once you take Math score into account. Using the formula from the slides (see formula sheet), you should be able to calculate this from the output of the `summary` function. You can check your answer using the `anova` function.

vi. Give a predicted first-year grade point average and a 95% prediction interval for a student who got 650 on the Verbal and 700 on the Math SAT. Are you confident that this student's first-year GPA will be above 2.0 (a C average)?

vii. Let's do one more test. We want to know whether expected GPA increases faster as a function of the Verbal SAT, or the Math SAT. That is, we want to compare the regression coefficients, testing $H_0 : \beta_1 = \beta_2$.

A. Express the null hypothesis in matrix form as $\mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

B. Carry out a two-sided t-test.

C. Carry out an F test, the easy way. Does $F = t^2$?

D. State your conclusion in plain, non-technical language. It's something about first-year grade point average.

Bring your printout to the quiz.
