STA 2101/442 Assignment Six[1]

Please bring your R printouts to the quiz. The non-computer questions are just practice for the quiz, and are not to be handed in, though you may use R as a calculator. Bring a real calculator to the quiz.

Most of this assignment is based on the usual normal linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant. You may use facts like these without proof.

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$
- $SSE/\sigma^2 \sim \chi^2(n-p)$
- SSE and $\widehat{\boldsymbol{\beta}}$ are independent
- If $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$ are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu).$$

1. Show that the $n \times p$ matrix of covariances $C(\mathbf{e}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$. This is the key to showing that SSE and $\widehat{\boldsymbol{\beta}}$ are independent.

2. Consider the prediction interval for Y_{n+1}.
   (a) What is the distribution of $Y_{n+1} - \widehat{Y}_{n+1} = Y_{n+1} - \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}$? 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.
   (d) Using your result, derive the $(1-\alpha) \times 100\%$ prediction interval for Y_{n+1}.

3. There is a difference between a prediction interval and a confidence interval. Using the preceding question as a guide, derive a $(1-\alpha) \times 100\%$ for $E(Y_{n+1}) = \mathbf{x}'_{n+1}\boldsymbol{\beta}$. This time, rather than trapping Y_{n+1} in the interval, you are trapping $\mathbf{x}'_{n+1}\boldsymbol{\beta}$.

4. 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 $a = \frac{R^2_F - R^2_R}{1-R^2_R}$. Show

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

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

5. 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased.

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

   (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.

6. It is perfectly natural to assume that something like response to a drug might be approximately linear over some range of dosage values, but that each person in the population might have his or her own slope. Thus each time you select a random sample you'll get a different collection of slopes, and the regression coefficient corresponding to the slope would be a random variable. Here is a simple model illustrating this situation. Let

$$Y_i = b_i x_i + \epsilon_i,$$

   where $x_1, \ldots, x_n$ are known constants, and independently for $i = 1, \ldots, n$,

   b_i is a random variable with expected value β and variance σ_1^2,

   ϵ_i is a random variable with expected value zero and variance σ_2^2, and

   b_i and ϵ_i are independent.

   (a) This is a special case of the general mixed linear model (See Regression 1 slides).

      i. What is the matrix $\mathbf{X}$? What is p?

      ii. What is the matrix $\boldsymbol{\beta}$?

      iii. What is the matrix $\mathbf{Z}$? What is q?

      iv. What is the matrix $\mathbf{b}$?

      v. What is the matrix $\boldsymbol{\Sigma}_b$?

   (b) What would happen if you tried to estimate β in the usual way with

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

   Under what conditions on the x_i values is this estimator consistent?

   (c) Find another estimator of β by calculating $E(\overline{Y}_n)$. Why does the Law of Large Numbers *not* apply here? Okay, anyway, propose an estimator, and describe the conditions on the x_i values that will make it consistent for β.

(d) Now suppose that all the x_i values are equal to one. To make things as easy as possible, assume everything is normally distributed.

  i. What is the distribution of Y_i in this situation?

  ii. Propose an estimator of β that should satisfy anyone.

  iii. Give an *exact* $(1-\alpha)100\%$ confidence interval for β; you don't have to show any work.

  iv. Now suppose that you want to estimate σ_1^2 and σ_2^2. Remember the problem from last assignment in which men and women were calling a help line according to independent Poisson processes, and we tried to estimate λ_1 and λ_2? Does that problem tell you anything about your chances of success?

This last little example shows you two things. First, whether the parameters of a model can be estimated depends on how you collect the data; this is a matter of experimental design (assuming the x_i values are under the control of the investigator). Second, it is possible that some parameters can be estimated very successfully, while others cannot be estimated at all.

7. For most configurations of $x_1, \ldots, x_n$, the variance parameters in Question 6 can be estimated successfully — but it's not so easy to see how. So we'll do it numerically with maximum likelihood. Again, suppose everything is normally distributed.

Some data from the model of Question 6 are available from the class website, in the file `randslope.data`.

(a) Make a scatterplot of the data and bring it to the quiz. Does it look funny? You're guaranteed that the model is correct. *Why* does the scatterplot look the way it does? How would it look if there were also a range of negative x_i values?

(b) Estimate the parameters numerically. Your answer to this part is a set of three numbers. Show the definition of the function you're minimizing, as well as all the other input and output leading to your answer. Wondering about starting values? Well, at least you know where to start looking for $\widehat{\beta}$.

(c) Using the asymptotic variance of $\widehat{\beta}$, carry out a simple 2-sided Z-test of $H_0 : \beta = 0$. Your output should include the computed value of Z and the two-tailed p-value. Do you reject H_0 at $\alpha = 0.05$?

Bring your printout to the quiz.

8. For this question, you will use the file sat.data from Assignment 5. There is a link on the course web page in case the one in this document does not work. 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 (which will be provided on the quiz if you need it), you should be able to calculate this from the output of the `summary` function. 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 (which will be provided on the quiz if you need it), you should be able to calculate this from the output of the `summary` function. 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 $\mathbf{C}\boldsymbol{\beta} = \mathbf{h}$. Obviously, this should be pretty routine.

B. But it's a bit more trouble than you'd think using `R`. I can think of three ways, all a little clumsy. Do the best you can. Carry out the test, producing an F statistic, degrees if freedom (a pair of numbers) and a p-value. Be able to state your conclusion in plain, non-technical language. it's something about first-year grade point average.

Bring your printout to the quiz.