STA 2101/442 Assignment Nine[1]

The non-computer questions are just practice for the quiz, and are not to be handed in. Use R for Questions 1 and 2, and bring your printout to the quiz. **Your printout 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. Telephone sales representatives use computer software to help them locate potential customers, answer questions, take credit card information and place orders. Twelve sales representatives were randomly assigned to each of three new software packages the company was thinking of purchasing. The data for each sales representative include the software package (1, 2 or 3), sales last quarter with the old software, and sales this quarter with one of the new software packages. Sales are in number of units sold.

   The data are in `sales.data.txt`. Get the data with

   ```
   sales = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/sales.data.txt",header=T).
   ```

   The explanatory and response variables are what you would think.

   (a) Fit a full model in which the slopes and intercepts of the regression lines relating sales last quarter to sales this quarter might depend on the kind of software the sales representatives are using.

   (b) Carry out an ordinary F-test to determine whether the effect of software type on sales depends on the representative's performance last quarter. Be able to state your conclusion in plain, non-statistical language.

   (c) Estimate the slopes of the three regression lines. Make sure these numbers are on your printout. I don't see how you can do this without making a table.

   (d) Carry out tests to answer these questions. If they are already on the output of `summary`, use that.

   i. Are the slopes for Software 1 and 2 different?
   ii. Are the slopes for Software 1 and 3 different?
   iii. Are the slopes for Software 2 and 3 different?

   Protecting the three tests with a Bonferroni correction at the joint 0.05 significance level, what do you conclude? Plain language is not necessary, but you should say what happened.

   (e) The average (sample mean) performance last quarter was 76.56 (please use exactly this number). We are interested in whether the three software packages differ in their effectiveness for sales representatives with average performance last quarter.

   i. Estimate expected performance this quarter for sales representatives with average performance last quarter. These three numbers should appear on your printout.
   ii. State the null hypothesis in symbols.
   iii. Carry out the F-test.
   iv. In plain language, what do you conclude?

   (f) Now we will try a randomization test. Sales last quarter and sales this quarter are what they are, and furthermore the pairs stay together, to preserve the strong relationship between the covariate and response variable. We'll randomly shuffle the *pairs* against the fixed software variable, and carry out a randomization test as in Question 1e — that is, to find out whether the three software packages differ in their effectiveness for sales representatives with sales of 76.56 last quarter. Use the F-statistic as your test statistic. Your final answer is a randomization p-value. Mine was very close to what I got from the classical F-test in Question 1e.

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

2. As a student recently observed, we can easily test the null hypothesis that $\beta_1 = 0$ and $\beta_2 = 0$, but what about the null hypothesis that $\beta_1 = 0$ *or* $\beta_2 = 0$?[2] This is quite practical, because the alternative is that both parameters are non-zero. The trouble is that $H_0 : \beta_1\beta_2 = 0$ is not a linear null hypothesis, and the general linear F test only applies to collections of linear restrictions on the β values.

   Why don't we bootstrap $T = \widehat{\beta}_1\widehat{\beta}_2$, and if the 95% quantile confidence interval does not include zero, we'll reject $H_0 : \beta_1\beta_2 = 0$ at the 0.05 level. Use the SAT data again. Get the data with

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

   Your objective is to produce two numbers, the lower confidence limit and the upper confidence limit. Do you reject the null hypothesis? What do you conclude?

3. Arsenic is a powerful poison, which is why it has been used on farms for many years to kill insects. Even in very small amounts, arsenic can cause cancer in humans, and recently it has been found that rice and foods made from rice tend to be very high in arsenic. Brown rice is worse, by the way.

   In a controlled experiment, pots of rice were prepared by either washing the rice first or not, and by cooking the rice in either a low, a medium or a high amount of water. The response variable is amount of arsenic in the cooked rice.

   (a) Use a regression model with *cell means coding*. That's the model with no intercept, and one indicator dummy variable for each treatment combination. You don't have to say how the dummy variables are defined. That will become clear in the next part. Just give the regression equation.

   (b) Write the expected amounts of arsenic in the table below, in terms of the β parameters of your model.

	Amount of Water		
	Low	Medium	High
Washed			
Unwashed			

   (c) If you wanted to test whether the effect of washing the rice depended on how much water you cook it in, what is the null hypothesis? Give your answer in terms of the β values in your model.

   (d) If you wanted to test whether washing the rice before cooking has any effect if the rice is cooked in a lot of water, what is the null hypothesis? Give your answer in terms of β values.

   (e) Suppose you want to test whether the amount of water used to cook the rice makes any difference if the rice has been washed. What is the null hypothesis? Give your answer in terms of β values.

   (f) Averaging across different amounts of water used to cook the rice, does pre-washing affect the amount of arsenic in the rice. What null hypothesis would you test to answer this question? Give your answer in terms of β values.

   (g) If you wanted to test whether the effect of the amount of water used to cook the rice depends on whether you wash it first, what is the null hypothesis? Give your answer in terms of β values.

4. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression model for the problem:

$$Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,1} d_{i,2} + \epsilon_i,$$

   where $d_{i,1}$ and $d_{i,2}$ are dummy variables.

   (a) Make a two-by-two table showing the four treatment means in terms of β values. Use *effect coding*. In terms of the β values, state the null hypothesis you would use to test for

[2] Or both.

i. Main effect of the first factor
ii. Main effect of the second factor
iii. Interaction

(b) Make a two-by-two table showing the four treatment means in terms of β values. Use *indicator dummy variables* (zeros and ones). In terms of the β values, state the null hypothesis you would use to test for

i. Main effect of the first factor
ii. Main effect of the second factor
iii. Interaction

(c) Which dummy variable scheme do you like more?

5. In a study of math education in elementary school, equal numbers of boys and girls were randomly assigned to one of three training programmes designed to improve spatial reasoning. After five school days of training, the students were given a standardized test of spatial reasoning. Score on the spatial reasoning test is the response variable. You will define a regression model for this factorial analysis of variance. Don't write the model yet.

(a) In the table below, show how your dummy variables are defined. *Use effect coding.* That's the scheme with an intercept and minus ones. Write the name of each dummy variable at the head of its column.

Girls, Programme 1	
Girls, Programme 2	
Girls, Programme 3	
Boys, Programme 1	
Boys, Programme 2	
Boys, Programme 3	

(b) Give $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$ for the full model. Include the interaction terms. Notice you are *not* being asked to write expected values in the table. They are too messy.

(c) Suppose you want to test whether, averaging across training programmes, there is a difference between girls and boys in their average performance on the spatial reasoning test. State the null hypothesis in terms of the β values in your model.

(d) Suppose you want to test whether, averaging across boys and girls, there is a difference between training programmes in average performance on the spatial reasoning test. State the null hypothesis in terms of the β values in your model.

(e) Suppose you want to test whether the sex difference in average performance depends on which training programme the children are in. State the null hypothesis in terms of the β values in your model.

Please bring your printouts for Questions 1 and 2 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.
