STA 312f22 Assignment Nine[1]

Please bring your R printout for Question 3 to the quiz. The non-computer questions are practice for the quiz on Friday Nov. 25th, and are not to be handed in.

1. 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 (especially rice grown in the United States) 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. It's a continuous variable, and this is normal theory regression.

   (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_j 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_j 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_j 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_j 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_j 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_j values.

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

2. 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

      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?

3. This question uses the built-in R table `Titanic`. You may want to get the data in shape for logistic regression using the brutal approach illustrated in lecture (Logistic regression on the Berkeley data). There are other ways, but at least this one works and you have an example. Start by constructing a data frame with only the adults. My data frame has 2,092 rows.

   (a) The first analysis uses just adult males, because the main goal of this analysis is to compare the survival of passengers to crew, and most of the crew were men. There's a variable Class, which is 1st, 2nd, 3d and Crew. You'll test the relation of this variable to survival, and then compare the survival of Crew to the individual passenger classes.

      i. First just as a check, test the relationship of Class to survival with a contingency table. Calculate the likelihood ratio test statistic and the p-value. I used the `table` and `chisq.test` functions. Also, please use `prop.table` on your contingency table, so you can see what actually happened!

      ii. Now do the likelihood ratio test with logistic regression and dummy variables. *Make crew the reference category.*

      iii. What do you conclude from the tests for β_1, β_2, and β_3? Use plain, non-statistical language.

      iv. The summary output also includes a test of $H_0 : \beta_0 = 0$. What does this null hypothesis mean in terms of survival? In plain, non-statistical language, what do you conclude from the test?

(b) Now we are going to look for a possible Sex by Class interaction, just for the adult passengers.

i. The first job is to take a look at the data and see what happened. This should be the first step in any data analysis. Making use of the fact that sample proportions are just sample means computed on the 0-1 outcome y, use `tapply` the way I did on the rotten potato data, and get a Sex by Class table of sample proportions. Each number in the table is the proportion of passengers who survived.

ii. Using effect coding (that's the dummy variable setup with 0, 1 and -1), test for the Sex by Class interaction with a likelihood ratio test and a Wald test. What is the critical value? (Check the formula sheet.) I get $G^2 = 64.074$.

iii. The test for interaction definitely indicates that the odds ratios are unequal for the three passenger classes. So, it's important to look at the estimated odds ratios. I think the best way to express them is odds of survival for women divided by odds of survival for men. There are several good ways to do this. It does not matter how you get the job done, but those three numbers should appear on your printout. My answer for first class is 72.45614.

iv. Now use Wald tests to carry out all three pairwise comparisons of the odds ratios. The easiest way is to use cell means coding. That's the model with no intercept and an indicator dummy variable for each treatment combination. I didn't actually make the dummy variables myself. I constructed a combination variable using `paste`, like the variable `TB` in the rotten potato lecture. In plain, non-statistical language, what do you conclude from the pairwise comparisons? As usual, be guided by the 0.05 significance level.

v. When you fit a model with just your combination variable, you get some z-tests. They are meaningful. Be able to say what each one means, in plain, non-statistical language.

vi. Just to verify that you know what's going on, transform the $\widehat{\beta}_j$ values into estimated survival probabilities with one line of code. Compare your answer to Question 3(b)i.

Please bring hard copy of your full R input and output to the quiz. Some of it may be handed in.
