Homework 4: Multivariate Regression and Interactions

Answer the following questions in a .pdf or .docx, explaining all of your answers and putting any tables and figures in the document as necessary. When data is called for to answer applied questions, I will provide it in bblearn. Turn in your R code that created all of the tables and figures separately, and be sure that it runs from source in such a way that it loads the data and performs all the tests without me fiddling with it. Make sure to document your R source code using # comments if you want partial credit.

For multivariate regression, we will be performing and interpreting some linear regression results. We will be predicting the weekly wage of a sample of Indian wage earners. This time, our data is the format of an SPSS data file, a .sav file. Import it using the Rstudio import wizard, which also uses the haven package. This data has verbose labels on it for each variable describing the data (by default your graphs will use the labels, which is nice), but I have also put that information in an accompanying .txt codebook.

1. First, let's see the effect of education on these Indian worker wages. Predict their weekly wage with the three dummy/indicator variables of highest education level and report the regression coefficients in a table.

2. Now interpret the findings of this regression. What is our interpretation of each of the non-intercept individual coefficients in words? (Be careful, this uses your understanding of categorical variables we covered last week)

   What is the overall practical interpretation of all of the coefficients in the model as an organic whole?

3. Run and interpret a new model predicting weekly wages with all of the variables in the data frame as independent variables. Report a table with all of these coefficients. What does a six year increase in age predict about the respondent's weekly wages? What effect does being a woman have on wages in this model?

4. Now, we will have you visualize this model. Plot and report the model's predictions with weekly wages on the Y axis and age of respondent on the X axis. Please include the original data points in this graph for reference.

5. Plot the model's predictions with weekly wages on the Y axis and age of respondent on the X axis for both respondents with a permanent jobs and those without a permanent jobs. You can either do this in two separate graphs or put both of these types of

individuals on the same graph. Please include the original data points in this/these graphs for reference.

6. Now, run a bivariate regression model where age predicts weekly ages. Report the regression coefficient of age and compare it to the previous model. Why might not including the other variables change the coefficient of age in this way?

For interactions, we will be looking at some experimental data (ec.csv) on viewing a children's television show "The Electric Company" that is used frequently by GHV. Basically, some students in grades 1-4 are treated by being selected to regularly watch this educational children's programming, others are in the control group and are not selected. See the codebook for the description.
Use the code `ec$grade <- as.factor(ec$grade)` immediately after importing the data to resave the `grade` variable as a factor rather than continuous variable.

1. First, run a multivariate regression where we predict a child's score on the post-experiment test (`posttest`) with whether they were in the treatment group or the control group (`group`), also adding their pre-treatment scores on that test (`pretest`) into the regression. Report the regression coefficients in a table. How big an effect did watching the electric company treatment have on the child's score? Did scoring better on a pretest predict better post test scores? By how much?

2. Now, we want to see if there are any differences in the effect of the treatment depending on the grade of the student to which it was applied. Run a regression where you interact the treatment variable with the categorical variable measuring the grade of the respondent (R will create dummies for each grade automatically if you turned it into a factor variable, see above). Also incorporate the pretest variable as a normal additive part of the regression. Report the coefficients from the model in a table.

3. Consider the coefficients for both the treatment effect and also for the interactions of the treatment with the grade of the respondent. What do they mean? When with respect to the child's education does the treatment yield the largest increase in the child's scores?

4. What would be the logical, substantive interpretation of the differences in coefficient effects you found in the previous question?