Homework #4

Instruction: Do all the following empirical exercises using R. Turn in your answer with tables and graphs, if any, (along with your program and output files appended at the end of document). Refer to the output file whenever appropriate when discussing your results.

Note that for all simulation exercises, set the seed number to 123456 to ensure the reproducibility of your results.

Question 1. Naive Bayes Classification

My tutorial on this topic on Github can get you started with this question. Suppose that we are interested in predicting someone's marital status using his or her geographic variables. Lets use our `wage2.csv` data to examine this issue. `married` is the outcome of interest, and `south` and `urban` are the two geographic variables for predicting the outcome. Note that both geographic variables are integers.

1. Estimate the Naive Bayes Model with `married` being the outcome and `urban` and `south` the predictors.
2. Generate a new dataset indicating that one individual is currently residing in an urban area in the South.
3. Use the model from 1.1 and the new data in 1.2 to predict the marital status for an individual who is currently residing in an urban area in the South.

Question 2. Conditional Expectation Function: Prediction and Marginal Effects

Gary Becker, who pioneered in applications of economics to many social issues and won the Nobel Prize for his contributions, proposed the so-called quality-quantity trade-off model, in which a child's quality (i.e., human capital) is negatively related with the family size (i.e., quantity). Subsequent literature testing this hypothesis finds some supportive evidence. However, such relationship may disappear once birth order is taken into account, as some research has shown. The birth order effects on wages or other labor market outcomes are an interesting (and robust) finding in the literature. In light of this, lets use our `wage2.csv` data to examine this issue as well. `brthord` is the birth order variable.

1. To simplify the calculations, let's group the higher-order births into one group. Specifically, generate a new variable called `birthorder`, equal to one when the actual birth order equals one, two when the actual order equals two, three when the actual order greater than or equal to three.

2. Using the `aggregate()` function, Obtain and write down the conditional mean function $\mathbb{E}$[wages | birthorder]. **Hint:** Note that you will see an error message regarding the length of the argument. I would like to see if you can figure out how to "fix" the problem.

3. What is the impact of being a second-born child on wages (compared to the first-borns)?

4. Obtain the conditional means using the linear regression approach for two different specifications of the conditional mean function.

1. In `R`, run a linear model with dummy variables for each possible value (that is, the prediction specification of the model).
2. In `R`, use the `predict()` command to generate predictions or conditional means for each value of birth order.
3. In `R`, run a linear model with an intercept and all dummy variables except one base category (that is, the partial-effect specification of the model).
4. In `R`, use the `predict()` command to generate predictions or conditional means for each value of birth order.
5. Using your results in 4.4., what is the impact of being a third-born (or higher) on wages compared to first-borns?
6. Can you manipulate your linear model to **directly** inform you about the impact of being a third-born (or higher) on wages compared to **second**-borns?