Econ 184b, Econometrics, Assignment 6

The raw .rmd file for this assignment are available at: https://raw.githubusercontent.com/f1kidd/Econ184/main/Assignment/Assignment_6.Rmd

Note: For the math problems, you can either (1) type your solution and compile it with RMarkdown; (2) type your solution use another word-processing software and convert it into a PDF file; (3) write your solution on paper, scan it and make it into a legible PDF file. TAs can decide, at their discretion, that the submitted file is illegible and thus give zero credit to a question or the entire problem set. If you type your solutions (options 1 and 2 above), you will get **a bonus equal to 5% of your performance** on the problem set. No credit will be given if you only report the final answers without showing intermediate steps whenever appropriate.

For the programming problems, you need to submit both the compiled pdf/html file and the RMarkdown (.rmd) file on LATTE. You will get **an additional bonus equal to 5% of your performance** if your RMarkdown file is completely reproducible with minimal altercation (install packages, etc.). That is, our team (or anyone else) should be able to recompile your Rmarkdown file and reach *the same* result. You need to explicitly write out your answers - just showing programming outputs will receive zero credit.

You will be receiving a total of 15% bonus if you submit one single pdf/html compiled directly from Rmarkdown, along with the .rmd file.

Question 1:

There are three kinds of candy in a huge candy bag: the M&M, the Godiva, and the Sugus. A random variable Y has the following probability distribution: Pr(Y = M&M) = p, Pr(Y = Godiva) = q, and Pr(Y = Sugus) = 1 - p - q. You reach into the bag and randomly grab n=100 pieces of candy from the bag: random variables are denoted $Y_1, Y_2, ..., Y_n$.

a. Denote N_1, N_2, N_3 as the number of M&M, Godiva, and Sugus draws from the bag ($N_1 + N_2 + N_3 = 100$). Derive the likelihood function for the parameters p and q.
b. Suppose amongst these 100 pieces, there are 20 M&Ms, 30 Godivas, and 50 Sugus. Derive the maximum likelihood estimator for p and q.

Question 2: Who tends to have extramarital affairs?

This question uses a data set (affairs.csv and affairs_description.pdf) from a 1969 survey filled out by readers of Psychology Today. The sample includes employed men and women who are married for the first time. We want to know if there are any significant relationships between the probability of having an extramarital affair and certain characteristics of married people.

a. Looking at the variables in the dataset, do you expect a relationship between extramarital affairs and any of the variables in the dataset? If yes, what sign(s) do you expect?
b. Estimate a linear probability model of affair (not affairs) on male, age, yearsmarried, kids, religiousness, education, and rating (all in one regression). Remember to use the heteroskedasticity robust standard errors.
c. Interpret the coefficient on "rating".
d. What is the expected probability of having an affair for Veronica, a 30-year-old woman who is a high school graduate, with no children, has been married for 3 years, with average religious level (= 3) and is very happy (= 5) in her marriage?

e. Now consider a very religious woman, Katelyn, 25 years old, no children, who is happily married for a year and has 9 years of education. What is Katelyn's expected probability of having an affair? What does this imply about the applicability of the linear probability model in this case?

f. Estimate the same regression as in (b) using a probit regression. Are the results similar as in (b) in terms of statistical significance of the coefficients?

g. Repeat the calculation of predicted probabilities for Veronica and Katelyn.

h. Estimate the same regression as in (b) using a logit regression.

i. Repeat the calculation of predicted probabilities for Veronica and Katelyn.

j. Which model appears to be most appropriate in this example? Explain.

Your code could start with this:

```
url = "https://github.com/f1kidd/Econ184/blob/main/Data/affairs.csv?raw=true"
affairs = read_csv(url)
```

Question 3: Predicting Breast Cancer

This question uses the Wisconsin Breast Cancer dataset (available here). For a documentation of the data set, please visit https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).

a. Our goal is to predict whether a patient's breast tumor is benign or malignant (cancer). Run a probit model, predict the class of the tumor using all features. Which of these factors point to a malignant breast tumor with 5% statistical significance?

b. What is the in-sample predictive accuracy of the model?

c. Cross-validate the above model using 10-fold cross validation. What is the out-of-sample predictive accuracy of the probit model?

d. Estimate a linear probability model with the same variable. What is the in-sample predictive accuracy of the model? (Note: it will be easier to use the *glm* function instead of the *lm* function to estimate the LPM. *glm* defaults to the OLS model identical to *lm*.)

e. Cross-validate the above model using 10-fold cross validation. What is the out-of-sample predictive accuracy of the linear probability model?

Your code could start with this:

```
url = "https://raw.githubusercontent.com/f1kidd/Econ184/main/Data/breastcancer.csv"
breastcancer = read_csv(url)
breastcancer %<>% mutate(Class=ifelse(as.character(Class)=="benign",0,1))
```