HW 4: Building Predictive and Prescriptive Models

Purpose of This Assignment

These problems will help solidify some skills that you might use for your projects. Building predictive models helps you have a better understanding of the world based on the data, and building prescriptive models helps you make decisions given your understanding of the world.

Building Predictive Models (45 points)

In the Content folder named HW 4, you'll find *btma.431.736.f2018.v2.rda*. This contains the raw final scores of students who took that class a few years ago, without accounting for any of their bonus marks. The file also contain students' final project, post-retake midterm, homework average, and textbook quiz average scores. It also contains a column specifying whether or not the student was in the BANA concentration (business analytics). Download the data to answer the following questions. Note that this question is not meant to make you anxious about grades. Rather, it gives you a chance to practice analyzing data that you're familiar with in a way that illustrates some core concepts in regression analysis.

1a) (9 points) Suppose that you wanted to predict students' raw final scores (excluding bonus marks) using all the other columns as predictors. To two decimals, what is the coefficient estimate for final.project?

1b) (9 points) In 2018, the homework scores were out of 20 and the textbook scores were out of 15. Scale them so they are both out of 100. In other words, normalize these numeric predictors so that they represent percentages. Then re-do the regression with these re-scaled predictors. What changes in the regression output compared to the previous model?

1c) (9 points) Does your regression model suggest that BANA students do statistically significantly better than non-BANA students when all the other variables are included in the model? What is the p-value of the line corresponding to that hypothesis test?

1d) (9 points) Is there evidence that the way in which the post.retake.midterm score impacts the final score is different between BANA students and non-BANA students? Add the appropriate interaction term to your previous model, and state the p-value of the line corresponding to the hypothesis test to two decimals.

1e) (9 points) Remove BANA as a predictor of your model. Create another model by predicting the log of your response variable with the log of the numeric predictors. What is the coefficient for log(final.project)? Compare the coefficients of this model to the coefficients of the same model but without the logs. Once we go through the log-log module in the slide deck and understand the interpretation of coefficients in a log-log model, why the numbers you see in your output seem similar will make more sense. This is just to plant the seed and illustrate that point with real-world data.

Building Prescriptive Models (55 points)

2a) (7 points) Farmer Jill has been selling fresh-pressed apple juice at the farmer's market for some time now. She has noticed that the higher she sets the price, fewer people bought her apple juice, all else equal. She wants to know how much to price a bottle of fresh-squeezed apple juice. From the data she has collected, she estimates (using a regression model) that the expected quantity demanded as a function of price is given by $Q(p) = 50 - 5p$. This means that if she sets price at $p = 5$, then she can expect to sell 25 bottles. If she sets price to $p = 2$, she can expect to sell 40 bottles. Her marginal cost of production (the cost of producing and packaging a single bottle of apple juice) is \$1. To the nearest ten cents, what is the optimal price? On D2L, leave out the dollars sign and write the answer to two decimals. What is the profit at the optimal price (to two decimals, leaving out the dollar sign)? In particular, search between $p = 1$ and $p = 9$.

2b) (7 points) She actually only has a rough sense of the demand function. In particular, she wants to understand how **sensitive** her decision is to **parameters of her model**. What would be the optimal price if the demand function was really $Q(p) = 45 - 5p$? What would be the optimal price if the demand function was really $Q(p) = 55 - 5p$? For your reference, this is called **sensitivity analysis**.

2c) (7 points) Let M denote the maximum demand she would see if she set her price to 0. Then $Q(p) = M - 5p$. Plot $p^*(M)$ (how the optimal price changes as M changes), where M goes from 40 to 60.

2d) (7 points) Let k denote the parameter measuring the marginal impact of price on demand. Then $Q(p) = M - kp$. On the same plot, plot $p^*(k)$ (how the optimal price changes as a function of k) for $M = 45$ and $M = 55$, where k goes from 2 to 8. Color-code your plot so that $p^*(k)$ for $M = 45$ is in some shade of red and $p^*(k)$ for $M = 55$ is in some shade of blue. **Note**: When searching for the optimal price, don't constrain your price above by $p = 10$ in this. Search for a price p between $p = 1$ and $p = 15$. Also, the following links may be useful if you want to use ggplot() to plot: https://rpubs.com/euclid/343644 and https://stackoverflow.com/questions/40833809/add-legend-to-geom-line-graph-in-r/40834306.

Using your work above, to the nearest two decimal places, for what k would the optimal price be \$5.00 when M = 45? Using the same value of k as in the above question, what would be the optimal price when M = 55 (to the nearest two decimal places)?

2e) (27 points) Over the last few years, Jill has been recording the number of bottles of juices sold each day as well as the price for that day. Assume that the marginal cost of production remained at \$1 during this time period. Based on the data, she wants to figure out what to price bottles at to maximize her expected profit. She has asked you to **build a decision support tool** so that, given a data frame with quantity and price data, she can find the optimal price to the nearest ten cents. Build this function for her. In particular, given any set of data (with data on price and quantity sold), your function should build a polynomial regression model, estimate the parameters of that model to predict profit from price (you can assume a quadratic relationship between price and profit by adding a 2nd order term when predicting profit from price), and then use those estimates to find the optimal price to the nearest ten cents. Using your function and the dataset provided by Jill (*salesData.rda* in the HW 4 folder), what would you recommend as the price of her apple juice? Round to the nearest ten cents and write to two decimals.

If you want a bit more of a challenge, instead of assuming a quadratic fit, jump ahead to the model selection lecture module and try to use model selection techniques to decide the polynomial degree based on the data.

Note: Make sure you have built a user-defined function for 2e). For any dataset the retailer provides (you can assume the dataset has the same format/columns as the one provided on D2L), your function should tell the retailer what is the best price to set to maximize expected profit. Your answer to 2e) should be the output of the functions you made. Otherwise, 25 points will be deducted from your score. Comment your code well so that anyone can understand how to use your functions.

Grading Scheme

R Code

Submit your R code as **a single R file** to the D2L dropbox folder. In the file name, include the homework number, your first name, your last name, and your section number. An example would be 'HW4_firstName_lastName_L02.R' or 'HW4_firstName_lastName_L02.Rmd' (depending on whether or not you used RMarkdown). Submit your work as a single R file with the stated naming convention. **Don't create a zip file with multiple R files.** Also, make sure you clearly denote which problem is which problem. For example, use R code comments to write: #### Problem 2a ####.

As always, you will have unlimited attempts on the D2L quiz for the assignment.

Note: Make sure you have built a user-defined function for 2e). For any set of quantity and price data, your function should tell the retailer what is the best price to set to maximize expected profit. Your answer to 2e) should be the output of this function you made. Otherwise, 25 points will be deducted from your score. Comment your code well so that anyone can understand how to use your function. You can feel free to use/modify your function from 2e) if you wanted to though.

Caution: Early on in the semester, students do not follow instructions carefully. If you follow the steps below, you can almost guarantee that you will not lose points due to your R code not running properly on my computer. Carefully making sure your code works fine and does not contain irrelevant chunks of code is good practice, and you want to have formed these habits prior to working together in teams on projects.

1) Save your current homework R file, and close RStudio.
2) Open your homework R file again.
3) Start at Line 1 (the first line in your code). Press Ctrl + Enter (Cmd + Enter for Macs).
4) Line by line, keep pressing Ctrl + Enter (Cmd + Enter) and check the outputs. Make sure that your outputs match your D2L quiz responses for the assignment.
5) Remove all extraneous bits or chunks of code that were not needed to produce your output. For example, if you had bits of code that were dead ends and the objects you created were not used in your solution, then delete those lines of code. Students have things like 'Attempt 4' or 'Attempt 10' in their code. Before doing this step, you may want to save a separate copy that contains all your dead ends (in case you accidentally delete something you needed). That's great to have a separate copy of your work (saved as a separate R file) containing the dead ends, but in your final submission, only submit code that is part of your working solution. Otherwise, if you include all prior attempts before your working solution, then the TA might accidentally mark the wrong approach and flag your work for academic misconduct (in which case, I would have to investigate).
6) **Repeat Steps 1 - 5 again (to make sure you didn't remove a needed line of code).**
7) If you use setwd() in your code, then either use the load() function to load the datasets or write as R comments what files you are using from your working directory. I don't have access to your computer, so your code will not work on my computer if you just use setwd() and load the data in the Console. I need those datasets in my working directory to run your code.
8) Save your file with your first name and last name (as specified in the instructions), and submit your work to D2L. Your work for the assignment should all be in a single R file, not multiple R files. Points will be taken off if the file name is not in the specified format or if you split your homework file into multiple R files.