POLI 170 Data Analysis Assignment 2

Due 3:29PM Tuesday October 26, 2021

Please submit this assignment by uploading your html (`.html`) and code (`.Rmd`) files onto Canvas before the due time, with easy-to-recognize file names (e.g., `assignment2_KirkBansak.Rmd`). Your homework will be graded based on completeness, accuracy, and readability of both code and written answers.

The point allocation in this problem set is given by:

Q1.1	Q1.2	Q1.3	Q1.4	Q1.5	Q1.6	Q1.7
4	4	4	4	4	4	4

Q2.1	Q2.2	Q2.3	Q2.4	Q2.5	Q2.6	Q2.7	Q2.8
4	4	4	4	4	4	4	4

Q3.1	Q3.2	Q3.3	Q3.4	Q3.5	Q3.6 (Bonus)
4	4	4	4	4	4

Q4.1	Q4.2	Q4.3	Q4.4	Q4.5	Total (Bonus)
4	4	4	4	4	100 (4)

Intro

The dataset and data dictionary you need for this assignment, which can be found on canvas, are `data_health_synth.csv` and `data_dictionary.pdf`. The dataset is comprised of synthetic data posted by the authors of Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* Vol. 366, No. 6464 (2019)—hereafter referred to as Obermeyer et al. (2019). It is designed to reflect the true data used by the authors in their study while protecting the privacy of the subjects.

Getting a Handle on a New Dataset

Before working with a dataset, it is important to understand its structure and what the variables are, so you should begin this assignment by looking through the data dictionary. It is quite long given the number of variables in the dataset, so you do not need to commit it to memory, but you should have a general understanding of the types of variables included, as well as be able to identify the most important variables. Also be sure to distinguish variables measured at year t from those measured at year $t-1$.

In addition, while not required, you may also find it useful to read through the first 6 pages of the supplementary materials to Obermeyer et al. (2019), which provides additional details on the dataset. This document was posted in Readings on Canvas (`suppl_materials_Obermeyer_etal_2019.pdf`).

Once you have a sense of the data dictionary, load `data_health_synth.csv` into `R` using the `read.csv()` function, store it as a data frame called `dat`, and proceed with the assignment.

Part 1: Summary Statistics

1. The key variables we will focus on for much of this assignment are the following:

   (a) Total medical expenditures in year t (not in year $t-1$)

   (b) Total number of active chronic illnesses in year t (not in year $t-1$)

(c) Risk score produced by the commercial algorithm

(d) Program enrollment (i.e. whether or not a patient was enrolled in the high-risk care management program)

(e) Race

What are the names of these variables in the dataset?

2. In class, we discussed the distinction between conceptional, operational, and actualized versions of a variable. Discuss this distinction in the context of the second variable listed above.

3. Compute and report the mean values for the first four of the variables listed above.

4. Compute and report the proportion of patients who belong to each racial group.

5. Using bracket notation (i.e. not `dplyr`), compute and report the mean medical expenditures and mean number of chronic illnesses—variables (a) and (b)—separately for each of the two racial groups.

6. Repeat the previous except now, instead of using bracket notation, use functionality from the `dplyr` package. In particular, you will want to use the pipe operator `%>%` along with the appropriate functions.

7. Comment on the results. How do medical expenditures and chronic illnesses vary on average across the racial groups?

Part 2: Assessing Program Enrollment

1. Generate and display histograms of the risk score by program enrollment status. That is, you should create two risk score histograms: one for patients enrolled in the program, and one for patients not enrolled in the program. You can use whatever plotting functions you prefer, but you should make sure the x-axis ranges from 0 to 100 for both histograms for easier comparison.

2. Compare the two histograms and comment. Does there appear to be a strict threshold in the risk score that perfectly determines who gets enrolled in the program? Based on what we know about the relationship

between the risk score and program enrollment from the Obermeyer et al. (2019) reading, does this surprise you?

3. Compute, store, and report the 25th and 75th percentiles of the risk score.

4. Compute and report the mean enrollment (i.e. proportion of patients enrolled in the program) separately across three groups: (a) patients whose risk score is below the 25th percentile, (b) patients whose risk score is above/equal to the 25th percentile and below the 75th percentile, and (c) patients whose risk scores is above/equal to the 75th percentile.

5. Comment on the results. Does there appear to be a relationship between risk score and program enrollment based on this analysis?

6. Create a subsetted version of the data that includes only patients who satisfy both of the following criteria: (i) are not enrolled in the program, and (ii) have a risk score above/equal to the 75th percentile. Be sure to store this as a separate data frame rather than overwriting the full data frame. Report how many patients are in this "not-enrolled-high-risk" subset.

7. Focusing only on the "not-enrolled-high-risk" subset of data you just created, compute and report the mean number of chronic illnesses separately for each racial group.

8. Comment on the results. Does there seem to be a potential problem with the program enrollment decision-making based on this analysis?

Part 3: Building Predictive Models

Now we will train simplified versions of the commercial algorithm that produced the risk scores in the dataset using linear regression modeling.

1. Using the full dataset, fit a linear regression model in which the outcome variable (Y) is total medical expenditures in year t and the predictors (X) are all of the following:

    - Female indicator variable

- All of the age indicator variables
- The following comorbidity variables at time $t-1$
  - deficiency anemia
  - blood loss anemia
  - arrhythmia
  - complicated diabetes
  - uncomplicated diabetes
  - depression
  - hypertension
  - hypothyroid
  - pulmonary disease
  - renal failure
  - tumor
- All of the cost variables at time $t-1$

There should be a total of 32 predictors in your model, and you should only include them on their own (i.e. do not include any interaction or polynomial terms). Store the model and name it `mod1`. Report the R^2 value of the model, which you can easily extract with the following: `summary(mod1)$r.squared`

2. The model object `mod1` contains many components. This includes the "fitted values" for all of the patients in the data, stored as a vector in `mod1$fitted.values`. These are the model's predictions of the outcome for all of the patients in the sample. Compute the in-sample mean-squared-error (MSE) of the model by calculating the following in `R`:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where N denotes the number of observations, y denotes the true outcome in the data (total medical expenditures in year t), $\hat{y}$ denotes the outcome predictions stored in the model object (`mod1$fitted.values`), and i is an index that represents unique observations.

3. Now repeat (1) and (2) above except this time, instead of using medical expenditures as the outcome, use the total number of active chronic

illnesses in year t as the outcome. Store this second model (which should use all of the same predictors) as `mod2`, and again report the R^2 and in-sample MSE for `mod2`.

4. Calculate and report the following two correlations (using the `cor()` function):
   (a) the correlation between the fitted values from `mod1` and the risk scores from the commercial algorithm
   (b) the correlation between the fitted values from `mod2` and the risk scores from the commercial algorithm
5. Comment on the results. Based on what we know from the Obermeyer et al. (2019) reading, would you have expected correlation (a) or (b) to be higher?
6. (Bonus) Why do you think the correlations are not closer to 1?

Part 4: Recreating Figures

In this part, we will re-create simplified versions of key figures in Obermeyer et al. (2019). Begin by copying the following lines of code, which include four "chunks" or sections, into your `R Markdown` file:

```
#Chunk 1:
risk_deciles <- quantile(dat$risk_score_t,
                         probs = seq(from = 0, to = 1, by = 0.1))

#Chunk 2:
dat$risk_decile_bin <- as.numeric(
  cut(dat$risk_score_t, breaks = risk_deciles, include.lowest = TRUE)
)

#Chunk 3:
some_results <-
  as.data.frame(
    dat %>%
      group_by(risk_decile_bin,race) %>%
      summarise(mean_illness = mean(gagne_sum_t))
  )
```

```
#Chunk 4:
library(ggplot2)
ggplot(some_results,
       aes(x = risk_decile_bin, y = mean_illness, color = race)) +
  geom_point() + geom_line()
```

Part of becoming a good programmer involves being able to figure out what someone else's code is doing by looking through it line by line, and being able to re-use or adapt code for similar tasks. In the code above, the fourth chunk uses the `ggplot2` package (be sure to install the package if you have not already!) to plot the results of the first 3 chunks, and the resulting figure is a simplified re-creation of Figure 1(A) from Obermeyer et al. (2019). In the following questions, you will be asked to explain what each of the first 3 chunks of code are doing.

Useful strategies for understanding someone else's code include running individual lines and excerpts from individual lines in the console to see what they do, inspecting the outputs or objects that different bits of code create, and looking at the documentation of any functions you are not completely familiar with (e.g. you can look at the documentation of the `cut()` function by entering `?cut` into the console).

1. Explain what code chunk 1 is doing.

2. Explain what code chunk 2 is doing.

3. Explain what code chunk 3 is doing.

4. Now re-create a simplified version of Figure 3(A) Obermeyer et al. (2019). You will be able to do so by recycling and making very minor modifications to the code you already have above.

5. Recall that in class, we compared and discussed Figures 1(A) and 3(A), and this comparison is a central finding in Obermeyer et al. (2019). In your own words, describe the key difference between the patterns exhibited in Figure 1(A) vs. 3(A), what accounts for that difference, and its implications for racial bias.