Jai Mehrotra-Varma
12/9/22

Homework 4

1. To recap, my research question is to determine whether hospitalized COVID+ patients have a higher mortality rate than hospitalized COVID- patients, and I have filtered the data to just include 1) their age, 2) their gender, 3) their hypertension status, 4) their diabetes status, 5) their COVID status (which I eventually added manually to the pandas dataframe), and 6) their mortality outcome. To conduct my analyses, it was essential for me to correct for the inherent differences in the datasets in some way. When doing basic statistical analysis on my COVID+ dataset and COVID- dataset, I discovered that there was a statistically significant difference in several of my variables (ex: age and sex), so I knew that it wouldn't be appropriate to just do simple comparisons (ex: chi-squared test, t-test, regression, etc.) without accounting for these differences in some way. I needed to be sure that COVID status was a determining factor for differences in mortality, and it wasn't just because one group happened to have older individuals or some other major confounding difference. As a result, the two main analyses I decided to do were odds ratio analysis and propensity score matching. Odds ratios are defined as "measures of association between an exposure and an outcome" - in other words, if the odds ratio model gave us a value of 2.5 for mortality in COVID+ patients, we would be able to make the claim that COVID+ individuals were 2.5 times more likely to die than COVID- patients. Specifically, I performed what is known as an adjusted odds ratio which controls for other predictor variables in the model. From reading up on other research papers that have made use of the odds ratio, I discovered that most researchers utilize age and sex as the two major variables to control for, so I followed the same process and controlled for both age and sex while setting mortality as the dependent variable. My results were as follows:

	Odds Ratio	p-value	5%	95%	significant
covid_status	2.841336	8.996275e-32	2.386385	3.383022	True
sex	1.807317	0.000000e+00	1.765253	1.850383	True
age	1.071549	0.000000e+00	1.070798	1.072301	True

Ultimately, I observed that age and sex were indeed contributing factors to the differences in mortality between the datasets (older individuals and males were more likely to die from COVID), but even after these variables were accounted for, COVID+ patients were still more likely to die than COVID- patients (odds ratio of 2.84). The other major analysis that I did is known as propensity score matching. This is another common technique used in COVID research that also addresses the influence of confounding variables and controls for them. For example, let's imagine that we have a COVID+ patient that is 85 years old, is a male, and has both diabetes and hypertension. If we compare that man to a COVID- patient who is 24 years old, is a female, and lacks

diabetes and hypertension, and the 85-year-old man dies, the reason could have just as easily been due to age, sex, diabetes, or hypertension as opposed to his status as COVID+ (which is what I'm interested in). However, if I find another 85-year-old man who is 85 years old, is a male, and has both diabetes and hypertension **but** is COVID-, and only the COVID+ patient dies, we would be much more confident in accepting that COVID status may have played a significant role. In propensity score matching, a logistic regression is performed and propensity scores are assigned to all of the individuals in both datasets based on the present variables. Since there are substantially more COVID+ than COVID- patients, I then matched 200 COVID- patients with 600 COVID+ patients (1:3 ratio) with the goal that each of the 3 COVID+ patients matched to each COVID- patient would be as similar as possible with regards to age, sex, diabetes status, and hypertension status (which is known by comparing propensity scores), with the only real difference being COVID status. This allowed me to create a new "matched" cohort that corrected for these other variables, similar to the adjusted odds ratio model. I was able to create some simple histogram visualizations to check if this technique worked by analyzing the distribution of the propensity scores for COVID- vs Unmatched COVID+ and COVID- vs Matched COVID+. While the distributions for COVID- and Unmatched COVID+ were quite different, the distributions for COVID- and Matched COVID+ were quite similar, meaning that the matching was successful. I also performed chi-squared and t-test analyses to test for any statistically significant differences between COVID- and Unmatched COVID+ and COVID- and Matched COVID+. The differences in age, diabetes status, hypertension status, and mortality for COVID- and Unmatched COVID+ were simply a reflection of the confounding variables discussed earlier, and they served as further evidence that it was important to correct for these differences. The important results were found when comparing COVID- and Matched COVID+, shown below:

	noncovid	matched_covid	significant
Age - Mean	74.047619	73.348333	False
Age - Standard Deviation	13.437241	13.39176	-
Male	618 (52.55%)	291.0 (48.5%)	False
Diabetes	495 (42.09%)	268.0 (44.67%)	False
Hypertension	844 (71.77%)	443.0 (73.83%)	False
Mortality	159 (13.52%)	206.0 (34.33%)	True

As we expected, the propensity score matching ensured that there were not statistically significant differences in the age, sex, diabetes status, and hypertension status between the two cohorts. However, despite adjusting for all of these significant factors, there was still a statistically significant difference in mortality, with the matched cohort having a mortality rate of 34.3% and the COVID- cohort having a mortality rate of 13.5%. This significantly supports the idea that hospitalized COVID+ patients were more likely to die than hospitalized COVID- patients due to differences in their COVID status (and not just because of other confounding variables).

Jai Mehrotra-Varma
12/9/22

2. Bonus Extra Credit Visualizations:

For advanced visualizations, I was particularly interested in looking at correlations in my data in a little more depth. In my Homework 3 feedback, it was suggested to me that I could do some correlation analysis with some of my data fields, but instead of just doing a regular correlation plot, I decided to make use of the KNeighborsRegressor. This is an especially useful tool because when Linear Regression is typically performed, it strictly uses a linear model to predict the association between two variables. However, with this form of regression, we are able to estimate the association between the variables based on the proximity of the neighboring data points, which is more effective since in the real-world, data is typically not close to normal. I chose to do correlations between age and heart rate and age and blood glucose, since part of my earlier analysis was to control for these variables since I knew they were likely to affect other data fields (such as mortality). My visualizations were two interactive plots showing these two correlations and also dividing the data points into males and females to add another layer of depth. I may continue to do further correlations for other data fields before submitting my final report. My visualizations are shown here: