Project Report

CS396-4 Causal Inference

[redacted], 2022

1 Group members

[redacted]

2 Code (10 points)

2.1 Running your code (4 points)

Access the code via the Canvas Assignment hand-in. The code is in a compressed file. The instructions for running the code are in readme.txt file within this compressed file. In short, relevant estimation functions are in the code file Causal Model Framingham.py. The user can also run the file Causal Model Framingham.py for the predefined program that estimates and prints expectations and the risk ratio confidence intervals.

2.2 Documentation (4 points)

See Causal Model Framingham.py and readme.txt for detailed comments and documentation. To summarize, we use external libraries, including pandas, statsmodels.formula.api, and sklearn, and had all codes in Causal Model Framingham.py. The running environment is also indicated in requirement.txt. The data file data.csv that the code depends on is also included.

2.3 Estimating function (2 points)

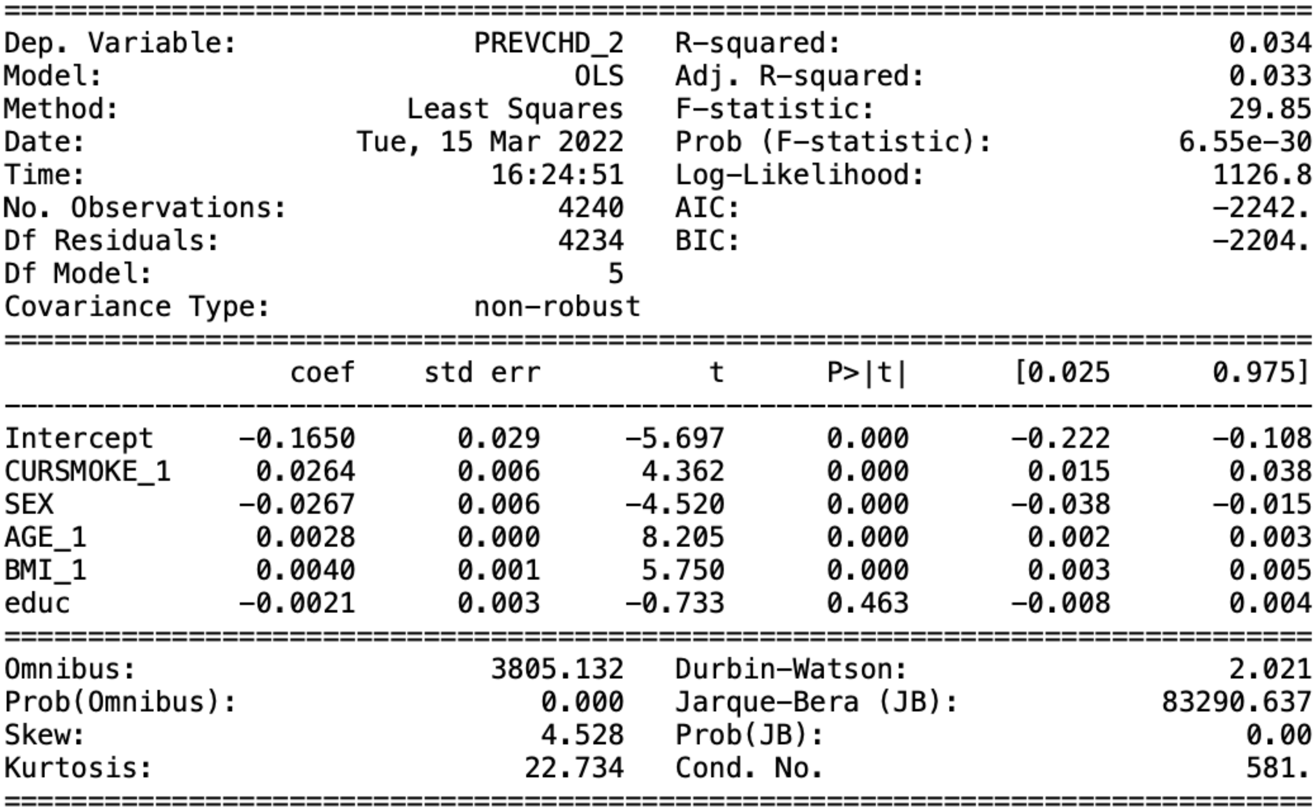
Before continuing, please see the previous project update Section 2 for the causal diagrams, variable definitions, and a description of the problem at hand. Recall from the previous project update that we are estimating three functions. The updated counterfactual functions for different periods are respectively as follows:

$$E(Y0^{A0=a0}) = \sum_{C0} E(Y0|C0, A0 = a0)P(C0) \tag{1}$$

$$E(Y1^{A0=a0,A1=a1}) = \sum_{C0,C1} E(Y1|C0, C1, A0 = a0, A1 = a1)P(C1)P(C0) \tag{2}$$

$$\begin{aligned} E(Y2^{A0=a0,A1=a1,A2=a2}) = \sum_{C0,C1,C2} & E(Y2|C0, C1, C2, A0 = a0, A1 = a1, A2 = a2) \\ & *P(C2)P(C1)P(C0) \end{aligned} \tag{3}$$

There are three major steps to calculating the team's estimator. First, the team uses imputes missing values in the data using MICE-like algorithm from sklearn.impute.IterativeImputer. Second, use the statsmodel linear regression API to output a linear model, providing linear terms that correspond to the expected value terms within each of the sums above. For example, for Group B (Participants whose first record of heart disease occurs at the second check-up), the corresponding linear model is generated from: fit c1 = smf.ols('PREVCHD 2 ∼ CURSMOKE 1 + SEX + AGE 1



```
==============================================================================
Dep. Variable:              PREVCHD_2   R-squared:                       0.034
Model:                            OLS   Adj. R-squared:                  0.033
Method:                 Least Squares   F-statistic:                     29.85
Date:                Tue, 15 Mar 2022   Prob (F-statistic):           6.55e-30
Time:                        16:24:51   Log-Likelihood:                 1126.8
No. Observations:                4240   AIC:                            -2242.
Df Residuals:                    4234   BIC:                            -2204.
Df Model:                           5
Covariance Type:            non-robust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1650      0.029     -5.697      0.000      -0.222      -0.108
CURSMOKE_1     0.0264      0.006      4.362      0.000       0.015       0.038
SEX           -0.0267      0.006     -4.520      0.000      -0.038      -0.015
AGE_1          0.0028      0.000      8.205      0.000       0.002       0.003
BMI_1          0.0040      0.001      5.750      0.000       0.003       0.005
educ          -0.0021      0.003     -0.733      0.463      -0.008       0.004
==============================================================================
Omnibus:                     3805.132   Durbin-Watson:                   2.021
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            83290.637
Skew:                           4.528   Prob(JB):                         0.00
Kurtosis:                      22.734   Cond. No.                         581.
==============================================================================
```

Figure 1: Example Output From smf.ols

+ BMI_1 + educ', data=df_c1).fit(). In this linear model for period one, PREVCHD_2 corresponds to Y0, CURSMOKE_1 corresponds to A0, and C0 includes SEX, AGE_1, BMI_1, educ. The first parameter tells the model to output PREVCHD_2 as a linear function of all the variables on the right hand side of the tilde. These are the same as in the project update, except we added BMI as a confounder (detailed later). The second parameter is the imputed data. This code outputs the coefficients for a linear regression based model, which are used in the corresponding Expected value for the appropriate confounder in equation 1 (Figure 1). A similar process is followed for the estimators for equations 2 and 3. Finally, the team calculated the probability of each confounder occurring, using simple tally counts within the imputed data. At this stage, every term in the sums above is decided and thus the calculation is executed.

3 Updates since your update (5 points)

3.1 BMI as a Confounder

Although BMI is not a perfect estimate of dietary and exercise habits, it is the best variable available in the dataset to measure these important variables which greatly affect cardiovascular health. So, the team included it as a confounder in C0, C1, and C2.

3.2 Missing Value Imputation

A major downside of the Framingham Dataset is that there are a lot of missing values. Indeed, there are two types of missing values. First, there are instances when patients attend a check-up but not all the information is recorded and second there are instances when patients miss entire check-ups. This data is likely not missing completely at random and so the team decided to impute the missing data. The team used the following process:

1. Import the entire dataset into a Pandas DataFrame.

2. Use a MICE-like algorithm provided through the default sklearn.impute.IterativeImputer API methods in order to impute the missing values.

3. Use the returned dataset instead of the original dataset to complete the processes outlined in the previous project update.

4 Interpreting your results (5 points)

4.1 Before

The team is interested in finding the causal risk ratio between participants who reported smoking at all check-ins (up to the present) to those who reported not smoking at all check-ins (up to the present). Specifically, with A# = 1 signifying the participant reporting that they smoke at check-in (#+1), and A# = 0 signifying the participant reporting that they don't smoke at check-in (#+1), the team found the following ratios:

1. $E(Y0^{A0=1})/E(Y0^{A0=0}) = 0.038/0.029 = 1.310$ (n=3053)

2. $E(Y1^{A0=1,A1=1})/E(Y1^{A0=0,A1=0}) = 0.060/0.056 = 1.071$ (n=2952)

3. $E(Y2^{A0=1,A1=1,A2=1})/E(Y2^{A0=0,A1=0,A2=0}) = 0.181/0.160 = 1.131$ (n=2782)

These ratios give us the percent increase in Coronary Heart Disease caused by smoking. To calculate these ratios, the team used a backdoor estimator as described in the equations in Part 3. The team used linear regression to determine the internal expectations within the summands ($E(Y2|C0, C1, C2, A0 = a0, A1 = a1, A2 = a2)$, $E(Y1|C0, C1, A0 = a0, A1 = a1)$, and $E(Y0|C0, A0 = a0)$). The team then calculated the remaining probabilities from the data and computed the sums. The corresponding 95% confidence intervals for different periods are as follows:

1. $E(Y0^{A0=1})/E(Y0^{A0=0}) : [0.870, 1.927]$ (n=3053)

2. $E(Y1^{A0=1,A1=1})/E(Y1^{A0=0,A1=0}) : [0.789, 1.468]$ (n=2952)

3. $E(Y2^{A0=1,A1=1,A2=1})/E(Y2^{A0=0,A1=0,A2=0}) : [0.960, 1.360]$ (n=2782)

The team sees that smoking increases the risk of smoking across the board (regardless of how years smoked). It appears that the causal effect is highest in the short-term with a near 30% increase and over time decreasing to around 10%. Thus, individuals should choose to not smoke if they desire to decrease their risk of heart disease and government policies should discourage smoking. However, none of the results are significant at a 95% confidence level.

4.2 After

With the new improvements to the model, the team obtained the following ratios:

1. $E(Y0^{A0=1})/E(Y0^{A0=0}) = 0.054/0.035 = 1.54$ (n=4240)
2. $E(Y1^{A0=1,A1=1})/E(Y1^{A0=0,A1=0}) = 0.071/0.060 = 1.18$ (n=3642)
3. $E(Y2^{A0=1,A1=1,A2=1})/E(Y2^{A0=0,A1=0,A2=0}) = 0.189/0.155 = 1.22$ (n=2903)

Additionally, the team obtained the following confidence intervals:

1. $E(Y0^{A0=1})/E(Y0^{A0=0}) : [1.168, 2.002]$ (n=4240)
2. $E(Y1^{A0=1,A1=1})/E(Y1^{A0=0,A1=0}) : [0.925, 1.505]$ (n=3642)
3. $E(Y2^{A0=1,A1=1,A2=1})/E(Y2^{A0=0,A1=0,A2=0}) : [1.020, 1.469]$ (n=2903)

The team sees that smoking increases the risk of smoking across the board (regardless of how years smoked). It appears that the causal effect is highest in the short-term with a near 50% increase and over time decreasing to around 18%. Thus, individuals should choose to not smoke if they desire to decrease their risk of heart disease and government policies should discourage smoking. Both Group B and Group D have statistically significant results at a 95% confidence interval.

4.3 Comparing the two

The changes implemented to the model (adding BMI as a confounder and missing value imputation), increased the causal ratios and lowered the variance of the bootstraps, which overall pushed most ratios into statistical significance. This makes sense because BMI indirectly measures dietary and exercise habits, which are big confounders of cardiovascular health. Additionally, the data that is missing is likely not missing completely at random, so filling it in gives us more accurate data and more data. This drives down the variance.

5 Reflections (5 points)

5.1 What you learned

The most interesting part of the project was creating the causal graph. It required a lot of thinking about which variables were confounders, which were mediators, which were treatments, and which were outcomes. It was also particularly interesting due to the longitudinal nature of this study. The most challenging part of this project was determining the counterfactual functions from the graph. The graphs were quite complex and it was difficult to determine which estimator to choose and why it was valid.

5.2 Unaddressed challenges

The biggest area of weakness in the team's project is that no one has medical domain knowledge. Thus, although the team put extensive analysis into constructing the causal graph, it is very plausible that a medical professional would decide that some of the team's mediators were more likely confounders. This would hugely impact the causal estimate, and potentially change the structure of the causal graph and thus the estimator used to produce the expectation of the counterfactual. This has undoubtedly the largest impact on the team's results. Another smaller challenge is that the team assumed a linear relationship between the outcomes and the confounders/treatments. It is possible that the relationship is non-linear which would undoubtedly impact the team's results (the expected values within the counterfactual functions would have different values).

5.3 What's left to do?

If the team had to spend another month on this project, the team would examine some of the assumptions the team made in an attempt to verify or disqualify them. This would largely focus on the relationship between the outcome and the confounders and treatments as well as the causal graph. The team would run distribution tests on the outcomes given the treatments and confounders, so that the team could discover if the relationship is linear or not. Additionally, the team would perform causal discovery to verify or disqualify the assumptions made about causal links between variables when constructing the causal graph. The team would look into using the PC algorithm and explore other possibilities. Overall, this additional work would grant more credibility to the team's work by analytically showing the assumptions the team selected were well-founded.

5.4 A follow-up study

If the team had lots of time and money, the team would like some additional information on dietary habits and exercise habits. Diet and exercise have a big impact on a person's cardiovascular health and these variables are not included in the current dataset. The most closely related variable is BMI. However, the relationship between dietary habits and exercise habits and cardiovascular problems via BMI is not as clear. For example, having a very low BMI could be a sign that someone is lean and work out often. However, someone with a large BMI could also be in shape, but very muscular. So, at best, BMI is a confounder with large measurement error. At worst, it isn't a confounder.

Besides collecting additional data, the team would not randomly assign treatments because it would be hugely unethical to instruct people to smoke or not smoke. Still, this new dataset would allow us to stratify out a big confounder which is not currently dealt with.