METHODS FOR CAUSAL INFERENCE: FORMATIVE ASSIGNMENT

1.

Case 1: Early epidemiological studies have shown that women receiving combined hormone replacement therapy (HRT) have significantly lower rates of coronary artery disease than average. But we should take into account that women receiving HRT are more likely to be from a higher socioeconomic group (ABC1), meaning their diet and exercise regimen is better than average.

In this case, we have at least three types of variables, i.e., treatment is women receiving HRT, outcome is their lower rate of coronary artery disease and confounders include socioeconomic group, diet, and exercise regimen. We can see the early research work classified the relationship between receiving HRT and lower rate of coronary artery disease into a direct causal relationship is not comprehensive, we should also consider the effects from the confounders.

Case 2: The relationship between genes, smoking habits, and lung cancer. As shown in the course example, some cigarette industry stated that lung cancer is due to some specific gene expression. But it is known to the public that lung cancer is strongly associated with those people who have a smoking history.

In this case, we have three variables, i.e. smoking, the specific gene, and the ratio of catching lung cancer. Obviously, the outcome should be the rate of getting lung cancer. The industry suggests that the specific gene is the causality of easily getting lung cancer, which means that they regard specific genes as the treatment and smoking habit as a confounder. From the public perspective, most people treat having a smoking habit as treatment, and whether having that specific gene is treated as a confounder.

Case 3: The causal relationship between physical activity and chronic back pain (CBP). Recently, observational studies have reported a negative association between physical activity and CBP, but the causal relationship is different. In the study [1], Gao, Shaowei, et al. gave a conclusion that the negative relationship between physical activity and CBP is derived from reduced physical activity rather than the protective effect of physical activity.

In this case, the original research aim is figuring out the causal relationship between physical activities and CBP, where physical activities should be regarded as the variable of treatment, and CBP as the variable of the outcome. However, with deeper the research level, the confounder variables are specified into whether reduced physical activity after suffering from CBP.

2.

In case 1, the counterfactual statement would be that women from the treatment group did not receive HRT and the women in the control group receive the HRT. Because the women from the treatment group are mainly from ABC1, even they didn't get HRT, their rate of coronary artery disease may still be at a low level. Likewise, for the women from the control group, those who are not from ABC1, even they received the HRT, the rate of coronary artery disease among them would still be higher than from ABC1.

In case 2, from the industry research aspect, the counterfactual statement is people who have the specific gene turn out to receive less possibility of lung cancer. If we control people who have 'lung cancer gene' never smoke, the ratio of lung cancer decreases. From the public aspect, the counterfactual statement is people who never smoke have a high probability of receiving lung cancer. Intervention, such as making 'lung cancer gene' suddenly express, may still have little influence on people who never smoke in getting lung cancer.

In case 3, the counter fact would be people after suffering from CBP increase the frequency of physical activity have a positive effect on relief CBP. The intervention here is increasing the frequency of physical activity after getting CBP, but the result from common sense should worsen the case. Another intervention would be increasing the frequency of physical activity before getting CBP, and the result turns out to be reducing the severity of illness.

3.

Assumption:

P(cancer|abnormal) = P(cancer|history, abnormal)

Information:

P(cancer|abnormal) = 0.8

P(cancer|normal) = 0.05

P(abnormal| history) = 0.6

P(abnormal| no history) = 0.2

Answer:

P(abnormal| history, cancer)

= P(abnormal, cancer | history)/P(cancer| history)

= P(abnormal, cancer | history)/ [P(cancer, abnormal| history) + P(cancer, normal| history)]

= P(abnormal| history) * P(cancer|abnormal, history) / [P(abnormal| history) * P(cancer|abnormal, history) + P(normal| history) * P(cancer|normal, history)]

= P(abnormal| history) * P(cancer|abnormal) / [P(abnormal| history) * P(cancer|abnormal) + P(normal| history) * P(cancer|normal)]

= 0.6*0.8 / (0.6*0.8 + (1-0.6)*0.05)

=0.96

P(abnormal| no history, cancer)

= P(abnormal, cancer | no history)/P(cancer| no history)

= P(abnormal, cancer | no history)/ [P(cancer, abnormal| no history) + P(cancer, normal| no history)]

= P(abnormal| no history) * P(cancer|abnormal, no history) / [P(abnormal| no history) * P(cancer|abnormal, no history) + P(normal| no history) * P(cancer|normal, no history)]

= P(abnormal| no history) * P(cancer|abnormal) / [P(abnormal| no history) * P(cancer|abnormal) + P(normal| no history) * P(cancer|normal)]

= 0.2*0.8 / (0.2*0.8 + (1-0.2)*0.05)

= 0.8

4.

(1) P(X1 = 1, X2 = 0, X3 = 1, X4 = 0)

= P(X1 = 1) * P(X2 = 0| X1 = 1) * P(X3 = 1| X2 = 0) * P(X4 = 0| X3 = 1)

= p0 * (1-p) * q * (1-p)

(2) P (X4 = 1| X1 = 1)

= P(X4 = 1, X1 = 1)/P(X1 = 1)

= [P(X4 = 1, X3 = 0, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 0, X2 = 1, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 1)]/P(X1 = 1)

= [(p0 * (1-p) * (1-q) * q) + (p0 * p * (1-p) * q) + (p0 * (1-p) * q * p) + (p0 * p * p * p)]/p0

= (1-p) * (1-q) * q + p * (1-p) * q + (1-p) * q * p + p * p * p

(3) P (X1 = 1| X4 = 1)

= P(X4 = 1, X1 = 1)/P(X4 = 1)

= [P(X4 = 1, X3 = 0, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 0, X2 = 1, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 1)] / [P(X4 = 1, X3 = 0, X2 = 0, X1 = 0) + P(X4 = 1, X3 = 0, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 0, X2 = 1, X1 = 0) + P(X4 = 1, X3 = 0, X2 = 1, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 0, X1 = 0) + P(X4 = 1, X3 = 1, X2 = 0, X1 = 1) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 0) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 1)]

= [(p0 * (1-p) * (1-q) * q) + (p0 * p * (1-p) * q) + (p0 * (1-p) * q * p) + (p0 * p * p * p)] / [(p0 * (1-p) * (1-q) * q) + (p0 * p * (1-p) * q) + (p0 * (1-p) * q * p) + (p0 * p * p * p) +((1-p0) * (1-q) * (1-q) * q) + ((1-p0) * q * (1-p) * q) + ((1-p0) * (1-q) * q * p) + ((1-p0) * q * p * p)]

(4) P (X3 = 1| X1 = 0, X4=1)

= P(X4 = 1, X3 = 1, X1 = 0) / P(X4 = 1, X1=0)

= [P(X4 = 1, X3 = 1, X2 = 0, X1 = 0) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 0)] / [P(X4 = 1, X3 = 0, X2 = 0, X1 = 0) + P(X4 = 1, X3 = 0, X2 = 1, X1 = 0) + P(X4 = 1, X3 = 1, X2 = 0, X1 = 0) + P(X4 = 1, X3 = 1, X2 = 1, X1 = 0)]

= [((1-p0) * (1-q) * q * p) + ((1-p0) * q * p * p)] / [((1-p0) * (1-q) * q * p) + ((1-p0) * q * p * p) + ((1-p0) * (1-q) * (1-q) * q) + ((1-p0) * q * (1-p) * q)]

5.

From the first glance, we can see that all the cases given completely violate the positivity, which means that there's no overlap between two groups of data, i.e. given a specific age value, there is only data from either one data group. The special cases are shown in Figure.1.

For the first case, the regression functions fitted by the two groups are linear functions, which means that if we extend the function to the whole age range, there will be an overlap between the two groups. Thus, we can say that under this case the estimate of its causal effect is more likely to be correct.

For the second and third cases, the situation depends. For example, in the second figure, if we regard the orange line as a constant function and the cyan line as a linear function, this case will be similar to the first case. But if we carefully observe the data point distribution within the orange region, we can make a bold guess that the orange line may have a right boundary like y=-log(-x). And we can also imagine the same situation on the cyan line, which has a left boundary. If the two boundaries have no overlap, we can hold that we cannot measure the causal effect due to the complete violation of positivity. The situation of case three is similar to case two, while the two fitted functions change to polynomial function. For certain polynomial functions, if the order of variable is large enough, there could also be the situation after a certain age, the change of orange line or the cyan line will tend to exponentially increase or decrease. Thus, in case two and case three, hastily estimating the causal effect would lead to a larger error.

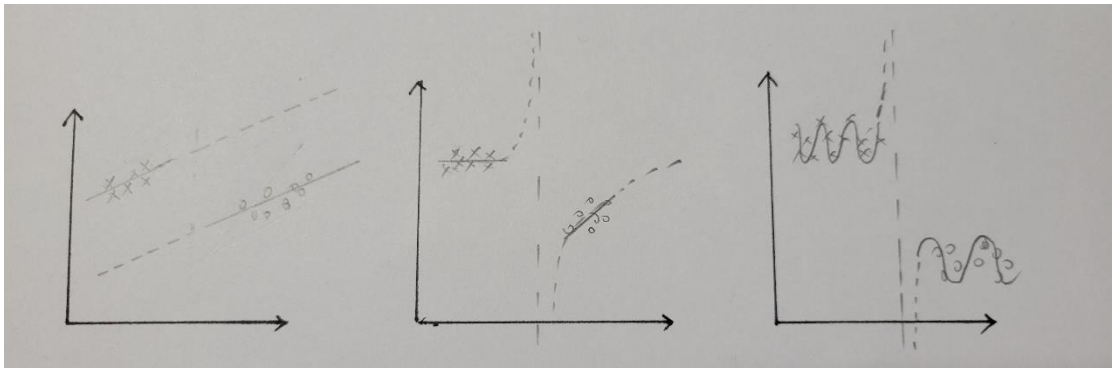

Figure.1

6.

(a)

The object function is Y = a0+b1*X1+b1*X2, our predict function is Y_hat = a0+b1*X1, then the residual error is the deviation between the observed and predicted y value, i.e., e=Y-Y_hat. The residual error is exactly the loss we want to minimize by the least square. That is to say, we need to minimize the mean squared residual e^2, where e = Y-Y_hat. We modify the right hand side by adding and subtracting the quantity (y_mean+b1*X1_mean), then we obtain e^2 = [(Y-Y_mean) - b1*(X1-X1_mean) – (a0 + b1*X1_mean -Y-mean)]^2. Because X1 and X2 are independent, we can minimize e^2 by taking partial derivatives of a0 and b1 in this function and setting them equal to zero. Finally, we could get a0 = Y_mean-b1*X1_mean and b1 = Cov(X1, Y)/Var(X1), which are regardless of X2. Figure.2 below shows the detailed derivative process.

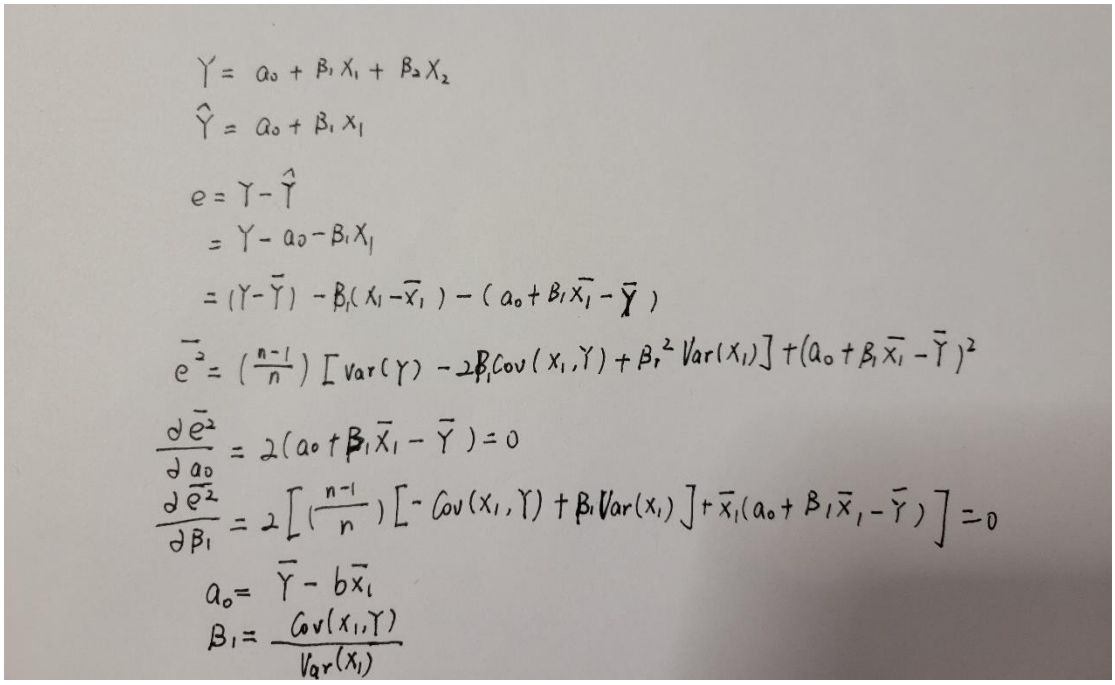

Figure.2

(b) Assuming that the relation between Y and T is linear, we can apply the Two-Stage Least Squares Estimator to get T's coefficient tau. Firstly, we get T_hat by estimating E[T|Z], where Z is independent of U. Secondly, we could estimate E[Y|T_hat] by fitting the coefficient in front of T_hat in this regression. Given the fact above, we could regard T_hat as X1, tau as b1, U as X2, and sigma_U as b2. Different form T, T_hat is independent of U, thus the above fact can be applied here to obtain tau.

(c)

Code:

```
import numpy as np
tau_list = []
for _ in range(10000):
    Z = np.random.randn(1000)
    U = 0.01 * np.random.randn(1000)

    T = 2*Z + 0.5*U
```

```
    T[T > 0] = 1
    T[T <= 0] = 0
    Y = 5*T + 2*U

    # get error bar!!!!! if the result is within the error bar, it
could be accepted
    ww1 = np.linalg.lstsq(np.array([Z]).T, T, rcond=None)[0]
    T_prime = np.array([Z]).T @ ww1
    tau = np.linalg.lstsq(np.array([T_prime]).T, Y, rcond=None)[0]
    tau_list.append(tau[0])

print('The error bar of tau (given 10000 times randomized trial):',
[np.mean(tau_list) - np.std(tau_list),np.mean(tau_list) +
np.std(tau_list)])
```

Output:

The error bar of tau (given 10000 times randomized trial): [4.99842346390062, 5.001584989414965]

Experiment original screen shot:

Result analysis:

In this simulation experiment, we performed randomized 10000 times and got the error bar of tau. Our fixed tau value is 5, the error bar is [4.99842346390062, 5.001584989414965]. Because the fixed tau value is within the error bar, we can conclude that the 2-step least-squares procedure indeed results in the correct estimate of the causal effect tau.

Reference

[1] Gao, Shaowei, et al. "Investigating the causal relationship between physical activity and chronic back pain: a bidirectional two-sample Mendelian randomization study." *Frontiers in genetics* 12 (2021).