Estimating Indirect and Direct Effects of a Cancer of Unknown Primary (CUP) Diagnosis on Survival for a 6 Month-Period After Diagnosis.

A manuscript prepared in fulfillment of a B.S Honors Thesis in Statistics

Fall Quarter 2016 and Winter Quarter 2017

Department of Statistics

University of California, Davis

06/14/17

Soo Yeon Park

Supervisor: Professor Christiana Drake

Abstract

Objective: Cancer for Unknown Primary is cancer with unidentified metastases, and it is believed that CUP diagnosis affects survival in two different pathways; CUP may innately differ from other cancers with identified primaries, so the diagnosis itself can influence how long a person can survive. Another way is that CUP diagnosis affects how much people receive treatment, and the treatment, the mediator, is the factor to influence the outcome. In this project, using the dataset of 10575 cancer patients, I assessed direct and indirect effects of CUP diagnosis onto survival.

Method: I used Inverse Propensity Weight by calculating marginal probabilities, estimated from logistic models. For indirect effect, I calculated P(treatment|confounders) as marginal probabilities to estimate IPW, while using P(pcup|confounders,treatment) for direct effect. All marginal probabilities are controlled within 0.05 and 0.95 to prevent extreme weight.

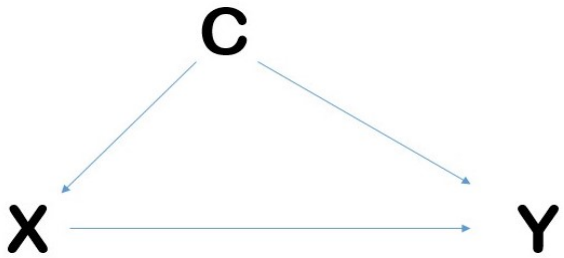
Result: IPW for direct effect is 7.376, and that for indirect effect is 1.545.

Conclusion: Direct effect of CUP diagnosis onto survival is much greater than indirect effect. CUP may innately from other cancers with identified metastases, and thus diagnosis itself can influence the patient's survival regardless of treatment.

Introduction

Most researchers know that association is different from causation; association does not require directions between variables unlike causation. For example, when ice cream sale goes up, the rate of someone drowning may go up as well. However, it would be wrong to conclude that eating more ice cream leads to a higher chance of drowning; those variables are positively associated, but one is not causing the other. In this case, the hidden variable that confound the relationship is the weather or daytime temperature. As the temperature increases and the weather gets hotter, more people go out to swim as well as eat more ice cream. Therefore, ice cream sales and the number of people drowning, each affected by the weather, increase independently.

Like the weather variable in the example above, we call variables, which distort the relationship between two variables, confounders, and they affect both the predictor and outcome variables. Confounders can either mask a relationship, or suggest there is a significant relationship between variables where, in fact, there is none. Furthermore, they can even reverse the relationship between predictor and outcome variables, a phenomenon known as Simpson's Paradox[1].



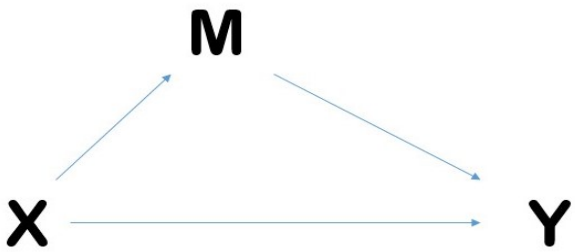
For example, let X be whether people smoke cigarettes (1 for yes, 0 otherwise), and Y be whether they have heart disease. In this case, if we ignore age, which is a confounder, C, the result might be that we need to smoke cigarettes to prevent heart disease, which would not make sense. This could be because, older people tend to smoke cigars or pipes instead of cigarettes, but also regardless of smoking, old age causes heart disease. Therefore, adjusting for confounders is necessary to investigate a causal the relationships between variables. We can infer causation between variables when all possible confounders that influence them are adjusted for[2,3].

There is also another type of variable called mediator, which bridges the relationship between the predictor and outcome. In this case, X has both direct and indirect effects on Y; indirect effect occurs when X affects mediator M, which further affects Y, while direct effect occurs when X affects Y without another variable in the causal pathway[2,3].

Lipid level is a possible mediator between smoking and heart disease (Greenland and Robbins). An indirect effect happens when smoking cigarettes increases lipid concentration in blood, leading to heart disease, while a direct effect occurs when smoking cigarettes increases heart disease through other mechanisms[3].

There are several ways to address mediation effects: path analysis is often used when mediator and outcome variables are continuous[2]. Another approach is using a counterfactual approach. In the counterfactual approach, actual observed responses are estimated with responses that might have been observed had the risk factor been changed to a different level. The potential outcome is counterfactual to what is observed. Therefore, if X is binary and it is 0, Y0 = actual observed

outcome, and Y1 = counterfactual outcome, what Y would have been if X were 1. Similarly, when X is 1, Y1 = observed response, Y0 = outcome if X were 0. It is possible, however, to estimate counterfactual outcomes by using observed outcomes if the samples from those are closely similar except the predictor; the estimate of Y0 when X is 1 would become Y0 when X is 0, and the estimate of Y1 when X is 0 would become Y1 when X is 1. Thus, we don't have to directly measure counterfactual outcomes with the same people, and still can estimate total causal effect by subtracting Y1 to Y0. For example, using smoking and heart disease example again, when people smoke, Y1 would be the response, while Y0 would be the response if the same people were non-smokers. For non-smokers, Y0 would be the observed outcome, and Y1 would be the response if they were smokers. To estimate counterfactual outcomes, we assume those outcomes would have been close to our observed responses if we measured from people with approximately the same backgrounds except smoking factor. Therefore, we can estimate causal effect by using the observed Y1 and Y0 even though those outcomes were not assessed from the same people.



The path diagram above is the basic model for causal relationships. If all variables are continuous, each path represents linear relationships between variables. Then,

$$\mathrm{M} = \alpha 0 + \alpha 1 X + \varepsilon 1$$

$$Y = \beta 0 + \beta 1 \mathrm{M} + \beta 2 \mathrm{X} + \varepsilon 2$$

If we plug in M into the equation of Y, the equation becomes the following:

$$\mathrm{Y} = \beta 0 + \beta 1(\alpha 0 + \alpha 1 \mathrm{X} + \varepsilon 2) + \beta 2 \mathrm{X} + \varepsilon 1$$

$$= \beta 0 + \beta 1 \alpha 0 + \beta 1 \alpha 1 \mathrm{X} + \beta 1 \varepsilon 2 + \beta 2 \mathrm{X} + \varepsilon 1$$

$$= (\beta 0 + \beta 1 \alpha 0) + \beta 1 \alpha 1 \mathrm{X} + \beta 2 \mathrm{X} + (\beta 1 \varepsilon 2 + \varepsilon 1)$$

Therefore, $\beta 1\alpha 1$ would be the indirect effect, $\beta 2$ would be the direct effect, and $\beta 1\alpha 1 + \beta 2$ would be the total causal effect from X onto Y. If we put this into the context of smoking and heart disease, $\beta 1\alpha 1$ would be how much smoking affects heart disease through changing blood lipid levels, which further affects heart disease. $\beta 2$ would be how much smoking itself affects heart disease regardless of blood lipid levels. The total of those two would be the total causal effect of smoking onto heart disease[3,7].

Barron and Kenny paper also mentioned Sobel's method to assess indirect effect[2]. Based on the equations above, using $\alpha 1$(the effect from X to M) and $\beta 1$(the effect from M to Y), the indirect effect is:

$$\sqrt{\beta 1^2 * s_{\alpha 1}^2 + \alpha 1^2 * s_{\beta 1}^2}$$

where $s_{\alpha 1}^2$ is the standard error of $\alpha 1$, and $s_{\beta 1}^2$ as the standard error of $\beta 1$.

If all variables, however, are binary, causal effect is assessed through logistic regression. The path diagram could have been more complicated with other variables such as intermediate confounder, and confounders between X and M, M and Y, and X and Y, denoted by L, U_3, U_2, U_1 respectively. However, several assumptions were made to address these variables. First, sequential conditional exchangeability assumptions state[7]:

$$Y(x, m) \perp X | C = c, \forall x, m, c$$

$$Y(x, m) \perp M | C = c, X = x, L = l, \forall x, m, c, l$$

$$M(x) \perp X | C = c, \forall x, c$$

These assumptions indicate that there is no unmeasured confounder in this model. In addition, cross-world independence assumption states that

$$M(x^*) \perp Y(x, m) | C = c, \forall x, m, c, x^*$$

and this assumption indicates that there is no L[7].

There are different types of direct and indirect effects. Controlled direct effect is the effect of Y on X when mediator is fixed at a certain value, and the formula is the following:

$$E\{Y(1, m)\} - E\{Y(0, m)\}$$

There is also another direct called natural direct effect, which simply ignores the mediator completely. Therefore, the formula to estimate natural direct effect is:

$$E[Y\{1, M(0)\}] - E[Y\{0, M(0)\}]$$

Natural indirect effect shows how much an expected value would change when mediator plays in a relationship. This compares the values of effect on Y between when M is 0 and 1, leading to the formula,

$$E[Y\{1, M(1)\}] - E[Y\{1, M(0)\}]$$

Lastly, total causal effect is defined by the sum of the natural direct and indirect effects, with its formula[7],

$$E[Y\{1, M(1)\}] - E[Y\{0, M(0)\}]$$

In this research, we are interested in using propensity weighting to study direct and indirect effects when the outcome, the causal variable of interest, and mediator are all binary. The model consists of three main variables, X is the predictor variable, Y is the response variable, and M is the mediator. Direct effect occurs when the predictor X affects the outcome directly without passing the mediator M. An indirect effect occurs when the predictor affects the mediator, which then affects the outcome. Sequential ignorability assumption was required which stated that the data

collection method does not depend on the missing data.

Data Analysis

We will apply the methods developed to a study of Cancer of Unknown Primary (CUP), a type of cancer that is detected through metastases but for which a primary tumor cannot be identified[8]. Our variable of interest or risk factor, is a diagnosis of CUP. This diagnosis is made through a series of tests. It is, however, very hard to diagnose CUP with confidence because sometimes it can be the case where the primary can be identified later, but since doctors cannot identify the primary within a certain time frame, they diagnose CUP for patients. It can be also the case where the tests to diagnose CUP may not be sensitive enough to properly diagnose patients. Every cancer has its own specific treatment based on metastases, but since CUP patients do not know what specific cancers they have. Therefore, when receiving treatment, they go through the general treatment process, which is inefficient as well as deteriorates their bodies much quickly, and thus CUP patients usually die much earlier than regular cancer patients. In the dataset, which I investigated, 4161 out of 10575 patients were diagnosed with CUP. Surprisingly, however, unlike my anticipation, the mean survival time for CUP patients was much longer than that for regular cancer patients.

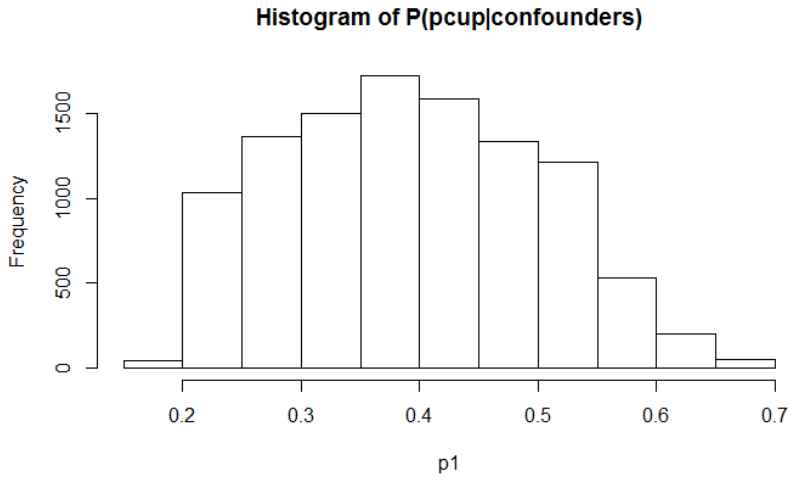
Our outcome variable is survival, and I mainly looked at survival as a binary variable. The mediator in our case is whether the patients received treatment for the cancer. We want to see if there is a direct effect of CUP diagnosis itself on the outcome. We hypothesize that regardless of treatment if there is a direct effect on survival, this indicates that a CUP is innately different from other cancers where the primary tumor is typically found first or simultaneously with metastases. The mediator in our case is treatment. Whether or not treatment is given may have an effect on survival. However, whether or not treatment is given and what type of treatment is influenced by the primary tumor diagnosis.

However, in this causal model, there are confounders and possibly moderators, which affects the variables in the causal model, and thus may affect the causal relationships we wish to study. Main confounders I considered were age, sex, the region of living, sex, ethnicity, income (variable r_cmedinc), and comorbidities. For example, one of the main confounders of our interest, age, can influence how many patients are diagnosed; other health problems tend to be present as people get old. Age also can affect whether patients would go through the rigors of cancer treatments, as well as how long they can expect to survive.

Before fitting any models, I changed the outcome variable into binary by setting 1 as people who lived longer than 6 months, assuming that each month consists of 30 days, and 0 otherwise. The result showed that about one fourth of the patients lived longer than 6 months. The dataset also contained majority of patients who did not receive treatments, and over half of the patients in the sample were diagnosed with CUP. Logistic regression model was used to estimate different probabilities for each patient. I first fitted the logistic regression model for estimating $P(\text{CUP diagnosis}|\text{confounders})$

$$\text{logit}(\text{pcup}) \sim \beta_1 * \text{agegr} + \beta_2 * \text{URBAN} + \beta_3 * \text{S}_{\text{SEX}} + \beta_4 * \text{ethnicity} + \beta_5 * \text{r}_{\text{cmedinc}} + \beta_6 * \text{comorb}$$

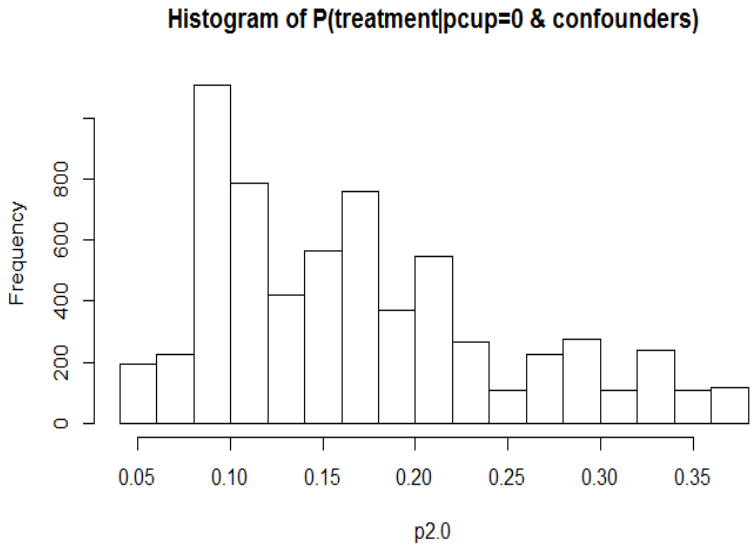
Then, I fitted P(CUP diagnosis|confounders) for each patient and displayed through histogram below.

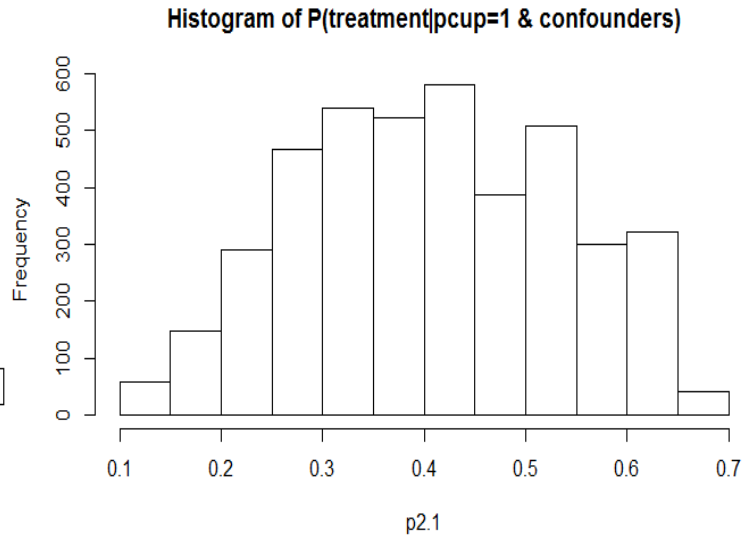


The probability distribution looks good with no extreme values. Next, I fitted two other logistic models to see if the probability of getting treatment differs whether a patient is diagnosed with CUP. Both models have the same form as below:

$$\text{logit(tx)} \sim \beta_1 * \text{pcup} + \beta_2 * \text{agegr} + \beta_3 * \text{URBAN} + \beta_4 * \text{S\_SEX} + \beta_5 * \text{ethnicity} + \beta_6 * \text{r\_cmedinc} + \beta_7 * \text{comorb}$$

Similar to the previous model, I displayed each patient's fitted probability through histograms. Some of the values for the model with non-exposure was too low (below 0.05), so I adjusted those numbers to 0.05 to maintain the sampling weight less than 20. The histograms are displayed below:

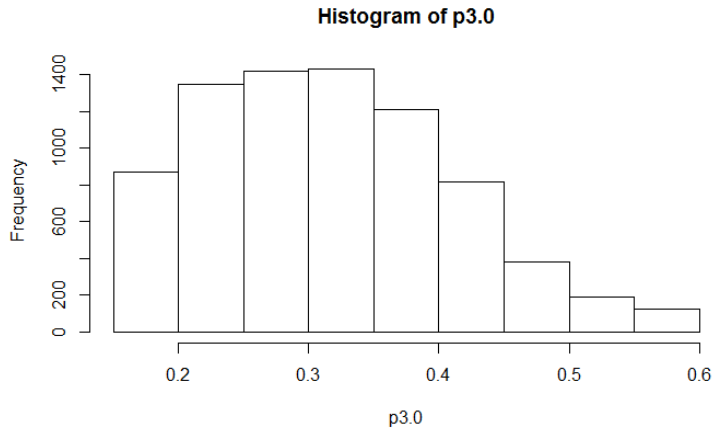




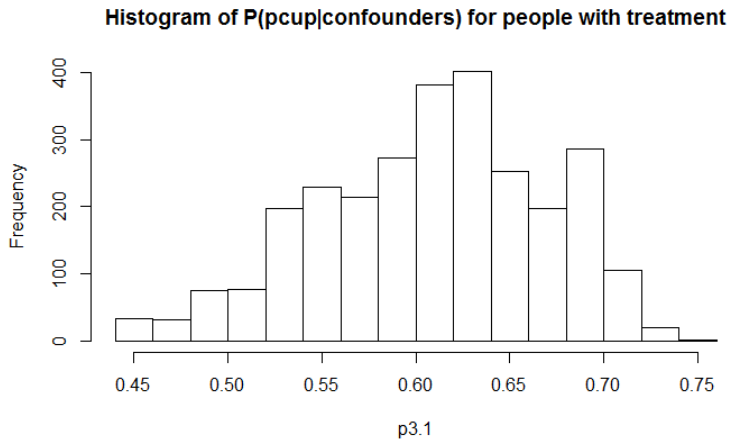
According to the histograms above, Lastly, to calculate P(predictor|confounder) depending on the mediator, I fitted the logistic model:

$$\text{logit}(\text{pcup}|\text{tx}) \sim \beta_1 * \text{agegr} + \beta_2 * \text{URBAN} + \beta_3 * \text{S\_SEX} + \beta_4 * \text{ethnicity} + \beta_5 * \text{r\_cmedinc} + \beta_6 * \text{comorb}$$

It is the same model that I fitted in the very beginning, but this time, I fitted twice based on the value of the mediator. The histogram of $\text{P}(\text{predictor}|\text{confounder})$ for people who did not receive treatment is below:



The histogram of $\text{P}(\text{pcup}|\text{confounders})$ for patients who received treatments were below:



Based on two histograms above, we can see that great proportion of people receiving treatment is diagnosed with CUP.

Using the probabilities that we calculated previously as e_i, I estimated marginal probabilities of CUP diagnosis onto survival. The probability of outcome under exposure is:

$$\widehat{\text{p0}} = \left(\sum_{1}^{n} \frac{Z_i}{e_i} \right)^{-1} \sum_{1}^{n} \frac{Z_i Y_i}{e_i}$$

And the probability of outcome under non-exposure is:

$$\widehat{p1} = \left(\sum_{1}^{n} \frac{1 - Z_i}{1 - e_i} \right)^{-1} \sum_{1}^{n} \frac{Y_i (1 - Zi)}{1 - e_{\bar{i}}}$$

Then, using these formulas, I could estimate inverse propensity weighting estimators as follow:

$$\text{IPW} = \frac{\left(1 - \widehat{p0}\right) \times \widehat{p1}}{\widehat{p0} \times (1 - \widehat{p1})}$$

The IPW for the indirect effect and direct effect turned out 1.545 and 7.376 respectively. This shows that direct effect is greater than indirect effect. Therefore, even though many people with CUP diagnosis received treatment, survival is more greatly influenced by the diagnosis itself or other mechanisms rather than the treatment.

Discussion

For this project, I fitted marginal models not conditional models, and since directly estimating $P(Y_1 = 1)$ and $P(Y_0 = 1)$ is not possible, additional assumptions were necessary[9]. Direct and indirect effects are assessed separately as I calculated inverse propensity weight by using marginal probabilities estimated from the logistic models. The benefits of the approach I took is that it does not require high levels of computing, as well as so it can be easily used by others. However, my result may not be accurate as I had to adjust the fitted probabilities between 0.05 and 0.95 to prevent the weight of each probability from getting too large. Therefore, the result may have been different if I didn't adjust the probabilities.

Conclusion

Direct effect of CUP diagnosis onto survival is much greater than indirect effect. CUP may innately from other cancers with identified metastases, and thus diagnosis itself can influence the patient's survival regardless of treatment.

Reference

1. Argesti, A. (2013). *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons.
2. Barron, R. M., Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182.
3. Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960.
4. Imai, K., Keele, L. (2010). A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4), 309–334.
5. Loux, T. M., Drake, C., Smith-Gagen, J. (2014). A comparison of marginal odds ratio estimators. *Statistical Methods in Medical Research*, 0(0), 1–21.
6. Lunceford, J. K., Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statist. Med*, 23, 2937–2960.
7. Rhian, D. (2017). Counterfactual-based Mediation Analysis Workshop. *London School of Hygiene and Tropical Medicine*.
8. Smith-Gagen, J., Mnatsakanyan, E., Tung, W. C., Caine, B. (2014). Cancer of unknown primary: time trends in incidence, United States. *Cancer Causes Control*, 25, 747–757.
9. Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statist. Med*, 4567-4580.

Code Appendix

```
#making the outcome binary(1 when survjs>180, 0 otherwise)
    pcup$survival = ifelse(pcup$survjs>180,1,0)
#subsetting patients based on the predictor(pcup)
    x1 = subset(pcup, pcup == 1)
    mean(x1$survjs)
    x0 = subset(pcup, pcup == 0)
    mean(x0$survjs)
#subsetting patients based on the mediator(treatment)
    m0 = subset(pcup, tx == 0)
    mean(m0$survjs)
    m1 = subset(pcup, tx == 1)
    mean(m1$survjs)

#indirect effect
#fitting the model to get P(predictor|confounder)
    logit.model1 = glm(pcup ~ agegr+URBAN+S_SEX+ethnicity+r_cmedinc+comorb, data =
    pcup, family = binomial)
#fitting the model to get P(treatment|confounder) when predictor is 0 or 1
```

```
    logit.model2.x0 = glm(tx ~ pcup+agegr+URBAN+S_SEX+ethnicity+r_cmedinc+comorb,
    data = x0, family = binomial)
    logit.model2.x1 = glm(tx ~ pcup+agegr+URBAN+S_SEX+ethnicity+r_cmedinc+comorb,
    data = x1, family = binomial)
#obtaining fitted values for models
    p1 = logit.model1$fitted.values
    p2.0 = logit.model2.x0$fitted.values
    p2.1 = logit.model2.x1$fitted.values
#histograms of probabilities for each model
    hist(p1,main="Histogram of P(pcup|confounders)")
    hist(p2.0,main="Histogram of P(treatment|pcup=0 & confounders)")
    hist(p2.1,main="Histogram of P(treatment|pcup=1 & confounders)")
#fixing the weighting (<0.05 to 0.05 and >0.95 to 0.95)
    p2.0[p2.0<0.05] = 0.05
    summary(p1)
    summary(p2.0)
    summary(p2.1)

#direct effect
#fitting the models based on the treatment(mediator)
    logit.model3.m0 = glm(pcup ~ agegr+URBAN+S_SEX+ethnicity+r_cmedinc+comorb, data
    = m0, family = binomial) #when treatment is 0
    logit.model3.m1 = glm(pcup ~ agegr+URBAN+S_SEX+ethnicity+r_cmedinc+comorb, data
    = m1, family = binomial) #when treatment is 1
#obtaining fitted values for both models
    p3.0 = logit.model3.m0$fitted.values
    p3.1 = logit.model3.m1$fitted.values
    summary(p3.0)
    summary(p3.1)
    hist(p3.0)
    hist(p3.1,main="Histogram of P(pcup|confounders) for people with treatment")

#calculating inverse propensity weighting
#indirect effect
#P(Y0 = 1) for P(treatment|confounders) when not exposed
    P.Y0_1.p2.0 = (sum((1-x0$pcup)/(1-p2.0)))^-1*sum((1-x0$pcup)*x0$survival/(1-p2.0))
#P(Y1 = 1) for P(treatment|confounders) when exposed
    P.Y1_1.p2.1 = (sum(x1$pcup/p2.1))^-1*sum(x1$pcup*x1$survival/p2.1)
#IPW
    IPW_indirect = P.Y1_1.p2.1*(1-P.Y0_1.p2.0)/((1-P.Y1_1.p2.1)*P.Y0_1.p2.0)
```

```
#direct effect
#P(Y1 = 1) for P(cup diag|confounders) when M=1
    P.Y1_1.p3.1 = (sum(m1$pcup/p3.1))^-1*sum(m1$pcup*m1$survival/p3.1)
#P(Y0 = 1) for P(cup diag|confounders) when M=0
    P.Y0_1.p3.0 = (sum((1-m0$pcup)/(1-p3.0)))^-1*sum((1-m0$pcup)*m0$survival/(1-p3.0))
#IPW
    IPW_direct = P.Y1_1.p3.1*(1-P.Y0_1.p3.0)/(P.Y0_1.p3.0*(1-P.Y1_1.p3.1))
```