Midterm

Adon Rosen

Date: 2020-10-23

Question 1

Rosen et al., (2018) describe the confounding influence that data quality has on structural neuroimaging measurements such as cortical thickness. Here they utilize a mediation model that models the influence that data quality has when exploring relationships between age and cortical thickness. Cortical thickness has received a lot of attention in the neuroimaging literature due to the strong relationship it possess with metrics of cognition. However, Neuroimaging has known confounds that make a large portion of the findings difficult to interpret which is largely driven by data quality. Age is strongly related to data quality where older participants tend to have higher quality data.

Rosen et al., seek to explore the causal mechanisms that data quality has when exploring developmental relationships by using a mediation model to explore the extent that the relationship between age and cortical thickness is mediated by data quality. The utility of this model describes the strength of the influence that has been so heavily described (age's relationship with cortical thickness) has been either over or understated. It states the extent that explored relationships are altered when accounting for data quality (i.e. strength of partial mediation) and attempts to report a model that more accurately relates age to cortical thickness.

This models explores an explicit causal relationship, declaring the age effects so heavily reported in the literature are directly related to the confounding influence of data quality. Although the model reported is a partial mediation, the causal effects of data quality on cortical thickness is described as causal.

Reference

Rosen, A. F. G. , Roalf, D. R. , Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., Garcia de La Garza, A., Gennatas, E. D., Quarmley, M., Schmitt, J. E., Shinohara, R. T., Tisdall, M. D., Craddock, R. C., Gur, R. E., Gur, R. C. & Satterthwaite, T. D. Quantitative assessment of structural image quality. NeuroImage (2018).

Question 2

a.

1) To estimate the total effect of X on Y do not adjust for Z

$Y = B_0 + B_1X$

1) To estimate the direct effect of X on Y adjust Y for Z

$Y = B_0 + B_1X + B_2Z$

b.

1) To estimate the total effect of X on Y adjust X and Y for Z

$\rho_{xy \cdot z}$

2) To estimate the direct effect of X on Y adjust X and Y for Z

$\rho_{x,y \cdot z}$

c.

1) To estimate the total effect of X on Y do not adjust for Z

$Y = B_0 + B_1 X$

1) To estimate the direct effect of X on Y do not adjust for Z

$Y = B_0 + B_1 X$

d.

1) To estimate the total effect of X on Y do not adjust for Z

$Y = B_0 + B_1 X$

2) To estimate the direct effect of X on Y adjust Y for Z

$Y = B_0 + B_1 X + B_2 Z$

e.

1) To estimate the total effect of X on Y do not adjust for Z or U

$Y = B_0 + B_1 X$

1) To estimate the direct effect of X on Y control X for I and use the residualized X to predict Y

$\beta = \frac{cov_{I,Y}}{cov_{I,X}}$

f.

1) To estimate the total effect of X on Y do not adjust for Z

$Y = B_0 + B_1 X$

1) To estimate the direct effect of X on Y adjust Y for Z

$Y = B_0 + B_1 X + B_2 Z$

g.

1) To estimate the total effect of X on Y do not adjust for Z

$Y = B_0 + B_1 X$

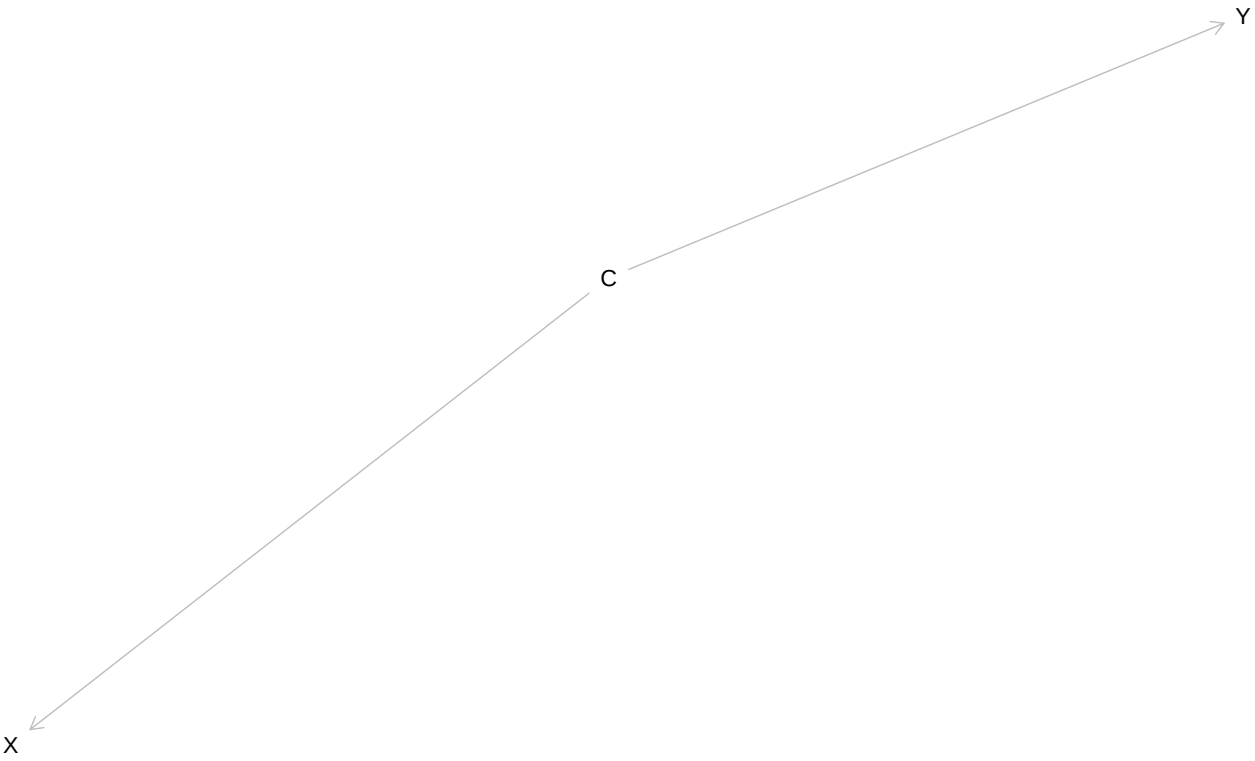
1) To estimate the direct effect of X on Y adjust Y for Z

$Y = B_0 + B_1 X + B_2 Z$

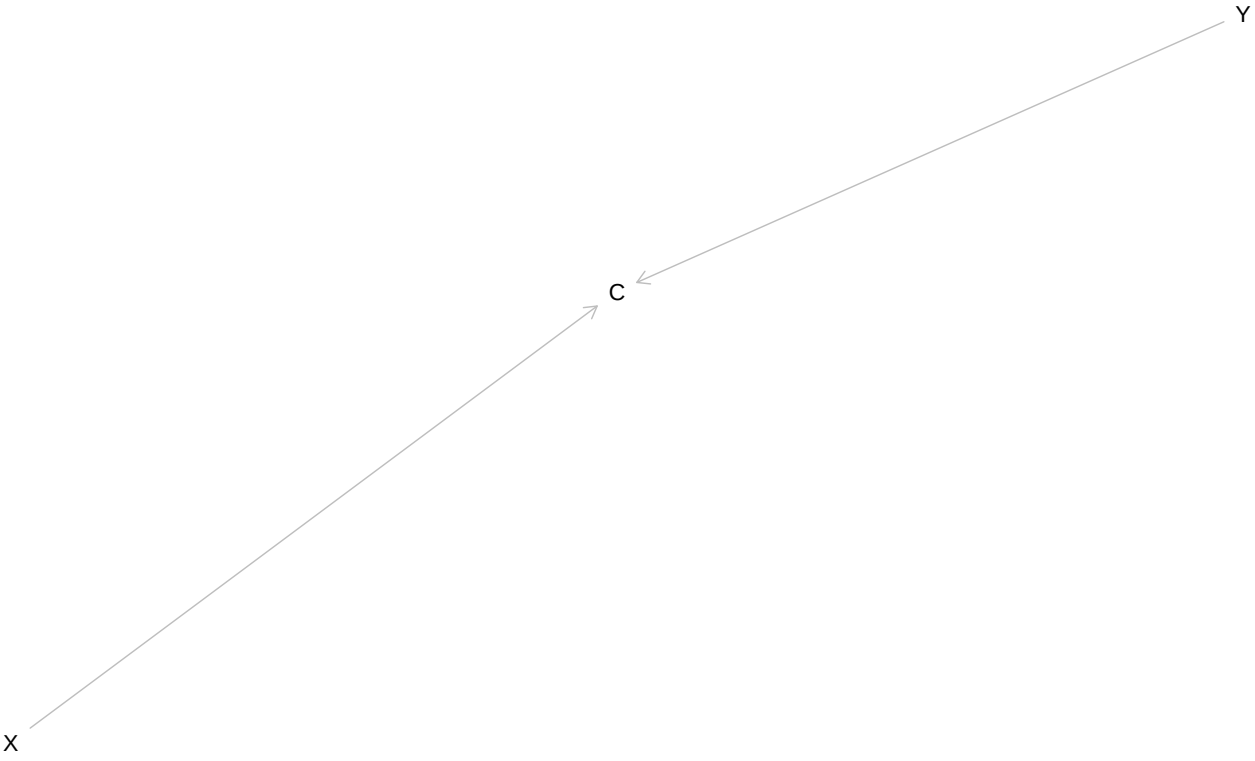
Question 3

A confounder is a variable that affects both the X (IV) and Y (DV) whereas a collider is impacted both of the X (IV) and the Y (DV). The causal mechanisms between the confounder and the collider are reversed, where it is recommended to condition on the confounder, conditioning on a collider may introduce spurious relationships when none such actually exist.

Below is a graphical representation of a CONFOUNDER



Below is a graphical representation of a COLLIDER

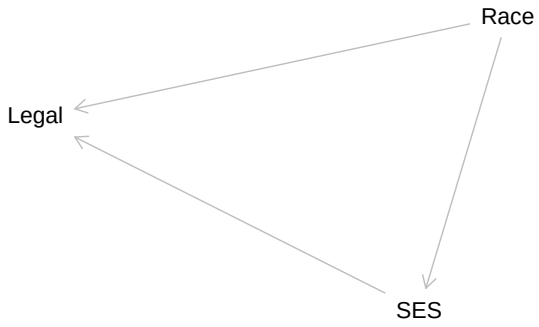


Question 4

a.

Financial data should be collected when dealing with a population that possess large variability in their socioeconomic status, and socioeconomic status does relate to legal defense quality.

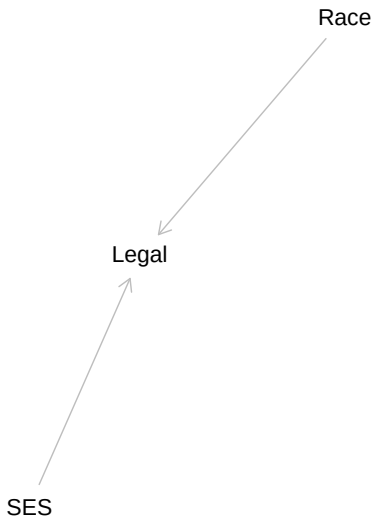
Defense



b.

SES should not be a concern if the population of interest does not display variability in this, or artificially controlling for SES by exploring effects within populations that satisfy a threshold for identical legal defense teams.

Defense



Question 5

Morgan & Winship introduce the terms "Differential Treatment Effect" and also "Baseline Treatment Effect" along side their introduction to Counterfactual modeling. Counterfactual modeling is a suite of techniques used to infer causal claims by assessing individual effects. An example of this logic can be found in the reading describing intelligence differences for individuals who attend college. Here the counterfactual examples are for those who did attend college, what would their observed outcomes be if they did not attend college, and the opposite, where we can estimate changes in an individual whom did not attend college if they had attended college. These examples highlight the counterfactual naming because the estimated individual differences are being made on events that did not occur.

The baseline treatment effect is described as follows:

$E[\delta_1|D = 1] - E[\delta_1|D = 0]$ Where $E[\delta_1]$ refers to the average treatment effect $D = 1|D = 0$ refers to the group membership. So this statement reads the average group effect for the treatment group ($D = 1$) minus the average group effect for the control group ($D = 0$). It indicates the group differences that may exist prior to any manipulation taking place

proc glm data=dat1;

class group1;

model cases=group1 age;

lsmeans group1 /adjust=tukey ;

run;

Here group1 is a variable indicating group membership, age is a continuous variable indicating the age of the participant, and cases is the outcome of interest.

The differential treatment effect is described as follows:

$(1-\pi)E[\delta_2|D=1]-E[\delta_2|D=0]$ Where $E[\delta_1]$ refers to the average treatment effect $D=1|D=0$ refers to the group membership, and π refers to the proportion of individuals who receive the treatment. This potential source of bias is due to the self selective nature of observational studies, where those that opt into a treatment group may see more difference due to the treatment than those that do not opt into the treatment group.

These confounds can be protected against by using propensity score covariates.

proc glm data=dat1;

class group1;

model cases=group1 groupMembership;

lsmeans group1 /adjust=tukey ;

run;

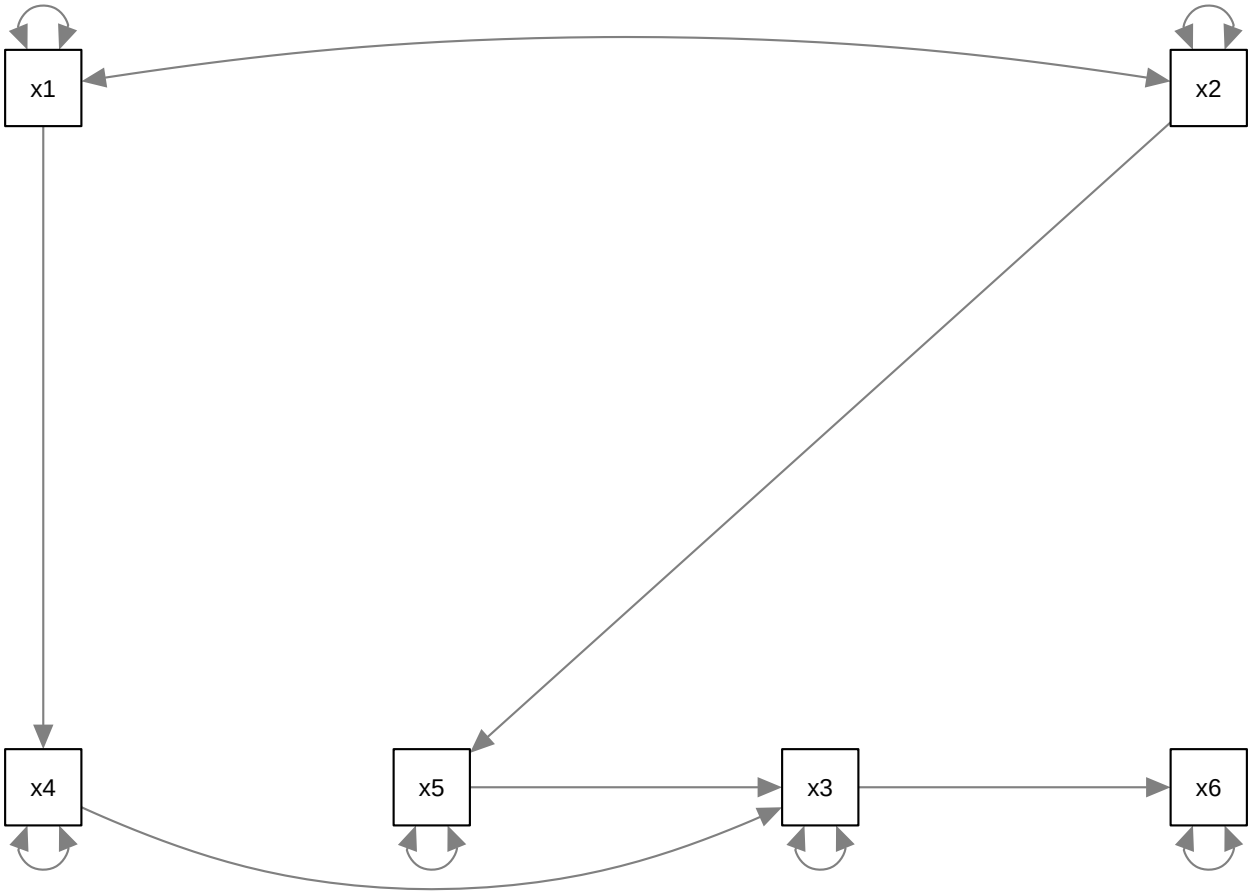
Here group1 is a variable indicating group membership, groupMembership is a continuous variable indicating the probability that a participant is belongs to a binary group, and cases is the outcome of interest.

Question 6

Model original

Table 1: Equivalent model predicted covariance matrix (Model Original)

	x1	x2	x3	x4	x5	x6
x1	1.00	0.49	0.23	0.49	0.26	0.07
x2	0.49	1.00	0.29	0.24	0.53	0.09
x3	0.23	0.29	0.91	0.29	0.47	0.28
x4	0.49	0.24	0.29	1.00	0.13	0.09
x5	0.26	0.53	0.47	0.13	1.00	0.15
x6	0.07	0.09	0.28	0.09	0.15	0.99

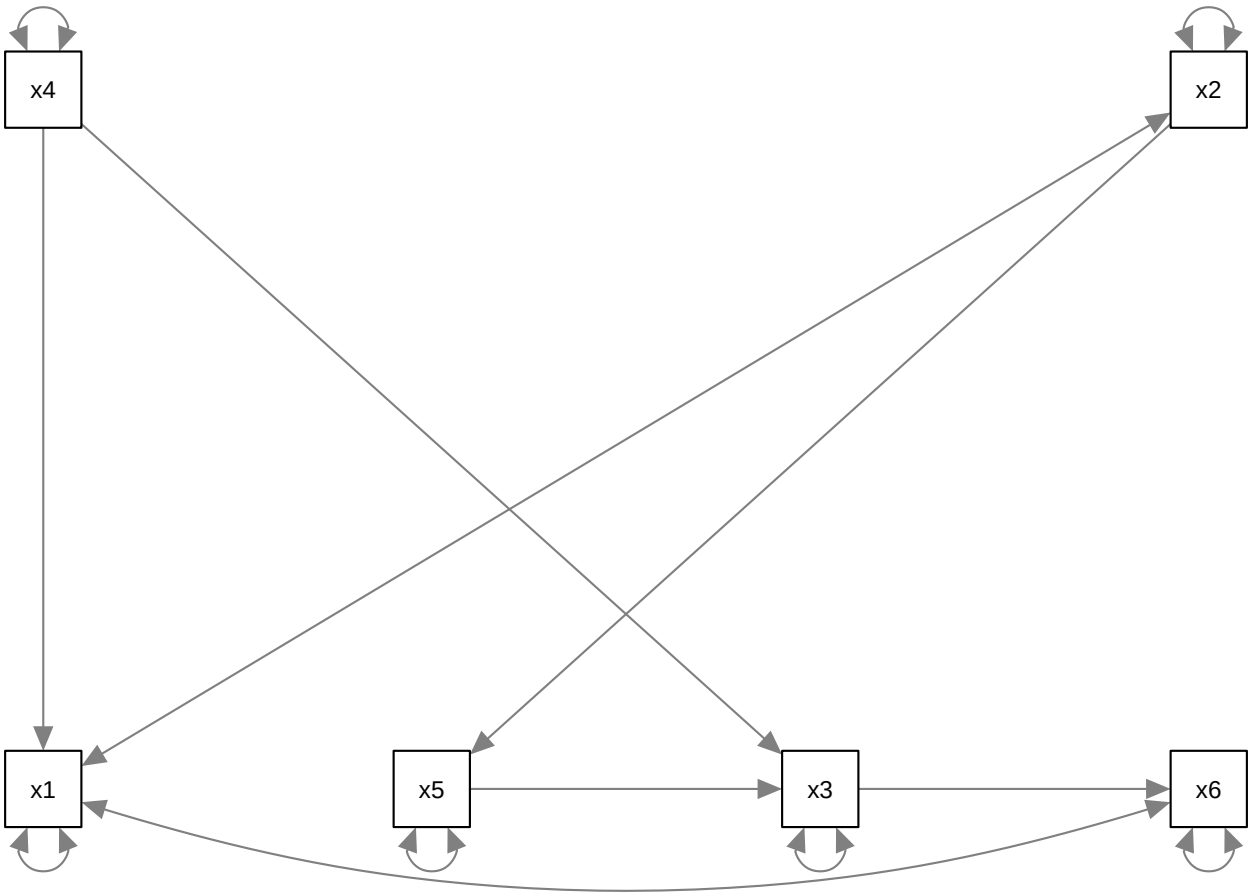



```
X1 _||_ X3 | X4, X5
X1 _||_ X3 | X2, X4
X1 _||_ X5 | X2
X1 _||_ X6 | X3
X1 _||_ X6 | X4, X5
X1 _||_ X6 | X2, X4
X2 _||_ X3 | X4, X5
X2 _||_ X3 | X1, X5
X2 _||_ X4 | X1
X2 _||_ X6 | X3
X2 _||_ X6 | X4, X5
X2 _||_ X6 | X1, X5
X4 _||_ X5 | X2
X4 _||_ X5 | X1
X4 _||_ X6 | X3
X5 _||_ X6 | X3
```

Table 2: Equivalent model predicted covariance matrix (Model Equivalent)

	x1	x2	x3	x4	x5	x6
x1	0.88	0.32	0.15	0.33	0.17	0.14
x2	0.32	1.00	0.23	0.00	0.53	0.07
x3	0.15	0.23	0.89	0.23	0.44	0.25
x4	0.33	0.00	0.23	1.00	0.00	0.06
x5	0.17	0.53	0.44	0.00	1.00	0.12
x6	0.14	0.07	0.25	0.06	0.12	0.97

Model equivalent




```
X1 _||_ X3 | X4, X5
X1 _||_ X3 | X2, X4
X1 _||_ X5 | X2
X1 _||_ X6 | X3
X1 _||_ X6 | X4, X5
X1 _||_ X6 | X2, X4
X2 _||_ X3 | X4, X5
X2 _||_ X3 | X1, X5
X2 _||_ X4 | X1
X2 _||_ X6 | X3
X2 _||_ X6 | X4, X5
X2 _||_ X6 | X1, X5
```

```
X4 _||_ X5 | X2
X4 _||_ X5 | X1
X4 _||_ X6 | X3
X5 _||_ X6 | X3
```

Question 7

a.

The direct effect of length of gestation on birth weight can be modeled as: $Y_{birthweight} = \beta_0 + \beta_{lengthofgestation}X$. Explicitly, the direct effect of length of gestation can be modeled directly without concern of the other possible confounds.

b.

The variables Rate of Growth and Length of Gestation can be included in a model in order to look at the relationship these variables have with birth weight. Such a model could read: $Y_{birthweight} = \beta_0 + \beta_{lengthofgestation}X_1 + \beta_{rateofgrowth}X_2$, such a model could be evaluated using an R^2 metric to evaluate the proportion of the variance explained by this model.

Question 8

ANCOVAs may be the analysis of choice if a randomized control trial has unexpected dropout that differs across groups. A theoretical example could be if a randomized trial is designed to explore potential effects of a drug versus a placebo, however delivery method has a known allergen, so while the groups are adequately controlled for, bias in the trials produces unaccounted for dropout. An ANCOVA could be used to account for how allergic participants are when trying to assess post-hoc group differences.

Question 9

Dominance analysis refers to a technique to explore the relative importance of a multiple regression model across various permutations of variables being included in the model. It does possess utility when assessing potential causal mechanisms although as in all statistical tools, it requires the appropriate data to make the appropriate conclusions, so dominance analysis is not enough to make a causal claim.

Dominance analysis theory is inherit to causal mediation models, for instance the Baron and Kenny approach requires for a model including the the potential mediator to explain more variance than the original predictor variable.

However the causal mechanisms are not fully flushed out when using dominance analysis, take for example the ice cream and shark attack example. Here a model can be made describing a relationship between frequency of shark attacks as the criterion variable with volume of ice cream consumed, miles driven, as well as amount of tv watched. Here dominance analysis would potentially show that the volume of ice cream consumed would be the most dominant variable in relation to predicting shark attacks. But this relationship does not pinpoint the true causal mechanisms of more time spent at the beach.

Such examples lead me to believe that dominance analysis has greater utility for building highly predictive models rather then designing models which drive theory.

Question 10

Three separate models were trained in order to explore the effect that treatment group has on remission rates amongst participants who recieved two separte treatment paradigms:

Naive Model: $\frac{P_{remission=1}}{1-(P_{remission=1})} = \beta_0 + \beta_{treatment}X$

Propensity Score Model:

Model 1: $\frac{P_{treat=1}}{1-(P_{treat=1})} = \beta_0 + \beta_{sex}X_2 + \beta_{age}X_3 + \beta_{spuse}X_4 + \beta_{work}X_5 + \beta_{phg1}X_6$

Model 2: $\frac{P_{remission=1}}{1-(P_{remission=1})} = \beta_0 + \beta_{treatment}X_1 + \beta_{propensity}X_2$

Multiple Regression Model:

$\frac{P_{remission=1}}{1-(P_{remission=1})} = \beta_0 + \beta_{treatment}X_1 + \beta_{sex}X_2 + \beta_{age}X_3 + \beta_{spuse}X_4 + \beta_{work}X_5 + \beta_{phg1}X_6$

The naive model suggests a significant treatment effect at an alpha of .05; suggesting there are greater incidences of remission for the treatment received group (t==1). The propensity score model as well as the multiple regression model do not suggest a significant treatment effect at an alpha of .05. The variable that is the most influential in the model appears to be for phg1 which describes depression loadings assessed at intake.