The Casual Effect of Computer Aided Learning on Educational Outcomes in India: A Propensity Score Matching Analysis

Daniel Boey, Mark Borsuk, Varun Mallampalli
Department of Civil and Environmental Engineering, Pratt School of Engineering, Duke University, Durham, USA.

Submitted as part of requirements for Research Independent Study.

CONTENTS

I. INTRODUCTION

Causal Inference

Establishing or justifying causal relationships is a demanding task. Since standard statistical analyses are driven by covariation, and not causation, distinguishing cause-and-effect relationships demands rigorous data analysis to enable such conclusions to be made (Pearl, 2009). Prior to the 21st century, statistical tools and language were insufficient to validate causal links within a phenomenon. However, in the recent years, causality arguments spearheaded by the field of epidemiology has inspired the fields of statistics, econometrics, engineering, biology and psychology (Rubin, 1974).

In the field of education, however, there has been a paucity of causal inference studies as current literature often focusses on anecdotal evidence or stops at correlating variables to educational outcomes due to a lack of data (Byrne, 1994). Some studies have examined causal relationships with regards to primary or elementary education. Self and Grabowski (2004) used causal inference theory to show how primary education in India had a strong causal impact on economic growth over the time period of 1966-1996. Specifically, the study made used of 'Granger causality' to seek the ability to predict economic growth based upon different levels of education.

Education in India: Right to Education and the DISE Data Set

In August 2009, the Parliament of India enacted the Right to Education Act (RTE) which dictated the norms and standards for a school for students between the ages of six and 14. Firstly, the RTE pegs the target pupil-teacher ratio at 30, a significant decrease from the previously mandated ratio of 40 (S. Mehrotra, 2012). Secondly, there should be at least one classroom for every teacher. The school also provide separated, gender-specified toilets. Third, the act specifies that every school must have safe drinking water facilities and a kitchen where meals are provided. Fourth, the minimum number of working days for primary schools is 200 and that of upper primary schools is 220, an ambitious goal given that most schools do not exceed 180 working days (S. K. Mehrotra, 2005).

The Unified-District Information System for Education (DISE) is a country-wide database of information about schools from primary to high school levels (ages 6-18). The database covers all 29

states and 7 union territories and includes over 1.5 million schools (National University of Educational Planning and Administration, 2014). With over 300 variables included, the data set covers multiple factors including number of teachers, number of students per grade, state/national examination passing rates, presence of Computer Aided Learning classrooms and other RTE compliance factors (e.g. presence of drinking water facility).

For this report, the authors investigate the possible casual link between the presence of Computer Aided Learning in classrooms and the proportion of students that score more than 60% on India's lower elementary/primary school examinations at Grade 5 (age 10), whilst accounting for other covariates chosen within the DISE dataset

II. METHODOLOGY

Overview

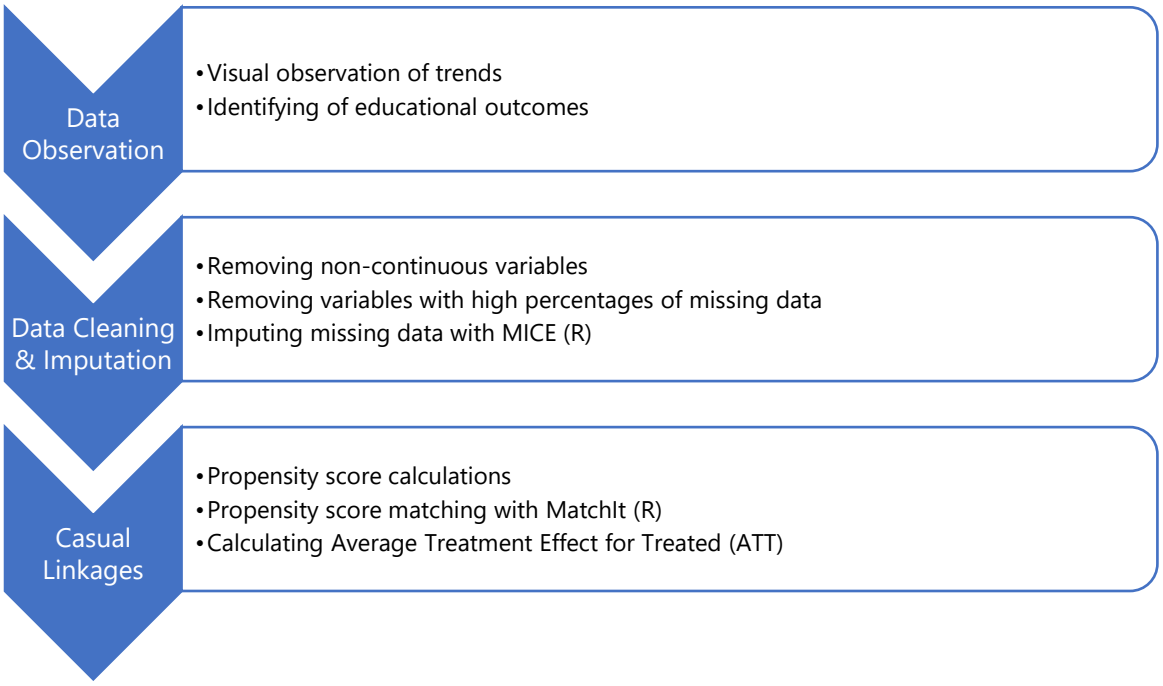

Figure 1: Overview of methodology used.

Assumptions

Based upon Rubin's foundational work in casual theory, biasness in treatment assignment can be ignored if two assumptions are held: unconfoundedness and overlap (National Bureau of Economic Research, 2007; Rubin, 1974) .

Let $z_i = 1$ if the unit (observation) *i* (*i = 1,…,N, where N is the total number of units*) is assigned to the treatment group and $z_i = 0$ if the unit is instead assigned to the control group. Let x_i be the vector that encompasses the corresponding covariate values for the unit *i*. Note that the numbering of units is assumed to be random; hence, the unit *i* contains no relevant information and acts simply as an index.

Assumption 1: Unconfoundedness

The assumption of unconfoundedness, a term coined by Rubin, assumes that there are no covariates outside the predetermined set of covariates that results in biases in the treated and control groups (National Bureau of Economic Research, 2007).

This allows a casual interpretation if the set of observed covariates truly has encompassed all possible confounding in the relationship between cause and effect. Whilst this may not entirely hold, the sheer

number of variables within the DISE dataset (300 in total) allows us to assume that one of these variables or a combination of them will directly address observed confounding and may indirectly help with highly correlated unobserved confounding. Hence, this assumption despite not being perfect can, to a large extent help with achieving more consistent causal estimates.

There is an equivalent idea in the missing data literature known as "missing at random". This means that the missingness of data within the dataset can be fully accounted for by observed covariates that have complete information. The missingness could however depend upon unobserved covariates, which we do not consider. Such an assumption is important as it forms the basis for our multiple imputation – that the existing data from observed covariates is sufficient to predict the missing data. The assumption of "missing at random" is therefore consistent with the idea of unconfoundedness.

Assumption 2: Overlap
The assumption of overlap that the treatment assignment has a non-zero probability to be in either treatment groups and is bounded away from zero and one. In other words, for every value of Xi there exists a strictly positive probability for being in the treatment group as well as the control group.

$$0 < \Pr(Z_i = 1 \mid X_i) < 1$$

Holding these two assumptions allows us to impose the strong ignorability condition (Rosenbaum and Rubin,1983). Under strong ignorability, we can estimate treatment effects without bias by accounting for the confounding effects of X_i.

Matching
In statistics, matching is a technique used to evaluate the effect of a treatment. It compares the outcome variable between units within the control and treated groups in studies where treatment is not randomly assigned. Matching often involves finding units with similar observed characteristics and seeking the difference in the outcome variable, thereby minimizing confounding bias. One important method of matching is using propensity score matching.

Propensity Scores
A propensity score, e_i, of an observation is the probability of treatment assignment given the observed covariates (Rosenbaum & Rubin, 1983).

$$e_i = \Pr(Z_i = 1 | X_i)$$
$$= \prod_{\{i=1\}}^{\{N\}} e(x_i)^{\{z_i\}} \, (1 - e(x_i))^{1 - z_i}$$

There is, however, not a specified specification for the propensity score. The propensity score can also be considered as a balancing score. Conditional on the value of the propensity score, the distribution of the covariates within the observed set will be similar between subjects of the control and treatment groups. Thus, the propensity score can be viewed as a balancing score (Austin, 2011).

Hence, the combined effects of the multiple distribution of covariates is aggregated to form a single predictor. In other words, two observations with the same propensity score can be assumed to have, on average, similar treatment biasness. For this study, a logistic regression model was used to estimate the predicted probability of treatment, a binary indicator, based on the baseline covariates. The task of deriving propensity scores is essentially a classification task, and more sophisticated classification algorithms can also be used to capture non-linearities inherent in the relationship between covariates.

Propensity Score Matching
Once propensity scores values have been calculated, observations with similar propensity scores values within the treatment and control groups will be matched. One-to-one or pair matching is the most common and involves matching every observation from the control group to an observation

from the control group of similar propensity score values. Note that if the number of control units is more than that of the treated units, the unmatched control units will be discarded. This procedure of matching every treatment unit to a similar control and discarding other control units gives us the Average Treatment Effect on the Treated (ATT).

This is distinct from the Average Treatment Effect (ATE) as the latter refers to the effect the treatment has on the whole population, instead of solely on the treated group. In a randomized control trial, we can assume that the ATT = ATE as the baselines for the treated and control groups are the same, whilst the treatment effect on the treated group would be the same as that on the control group. Heckman (1997) notes that ATT might be more policy relevant than ATE because the ATE includes the effect on persons for whom the program was never intended for.

In this case, the ATT is simply the sum of the difference between the outcome values between the matched pair.

$$ATT = \frac{1}{A}\sum_{\{i=1\}}^{\{A\}} Y_{t,i} - Y_{c,i}$$

where A is the number of matched pairs, $Y_{t,i}$ is the value of the outcome variable of the treated unit of the matched pair whilst $Y_{c,i}$ is that of the control unit of the matched pair.

Three other propensity score methods are commonly used: using propensity scores values as weights to create balance between the control (weighting) and treated groups, dividing the sample into blocks based on propensity score values (stratification) and forming a regression based upon the propensity score (Austin, 2011)

III. RESULTS

R (Version 3.5.1) was used throughout the data analysis process (R Development Core Team, 2018). The original DISE data for the state of Gujarat was uploaded as a data frame. Over 300 variables were observed with 239,232 schools as observations/units. Each observation has a corresponding school code, locality and pincode which allows individual identification of the school. For this analysis, only continuous variables were analyzed as covariates (217 covariates in total) whilst only binary variables were considered as treatment variables. This allows us to employ a logit regression model when predicting propensity scores. All variables were screened for missing data.

The presence of Computer Aided Learning (CAL) was chosen as the treatment variable since the number of observations for the control group ($Z_i = 0$ or CAL absent or non-functional; n = 122,582) was close to that of the treatment group ($Z_i = 1$ or CAL present; n = 116,478). 172 observations were discarded as they no value of the treatment variable was present. To calculate the outcome variable, the percentage of boys and girls who passed 5th grade with more than 60% on their national examination was considered. An alternate interpretation of this variable is that of a performance measure associated with learning outcomes. This will henceforth be referred to as the P60 rate.

$$Y_i = \frac{No.of\ students\ passed\ with > 60\%}{No.of\ students\ appearing\ for\ exam}$$

All observations without the presence of the outcome variable or if the outcome value is >1 was removed. Subsequently, the data set contained 217 covariates with 72,526 observations. Out of these 72,526 observations, the number of observations for the control group ($Z_i = 0$ or CAL absent or non-functional; n = 33,305) remained close to that of the treatment group ($Z_i = 1$ or CAL present; n = 39,221).

Data Imputation

To impute the missing data, the Multiple Imputation by Chained Equations (MICE) package in R was employed. Specifically, it employed the Predictive Mean Matching (PMM) method, often employed for quantitative variables that are not normally distributed (Allison, 2015). The PMM method makes use of real values within the data set to impute the missing values. First off, it estimates a linear regression of variables with missing data on variables with no missing data. It then serves as a metric to match cases with missing data to similar cases with data present. All collinear covariates were removed, which shrank the number of covariates to 163 covariates.*Propensity Scores*
Thereafter, a generalized linear regression was employed to calculate the propensity scores of each observation based on the 163 covariates. The histogram below (Fig. 2) shows the distribution of the propensity scores for both the treatment and control groups. Both groups showed complete overlap across the full range from 0 to 1 propensity score values.

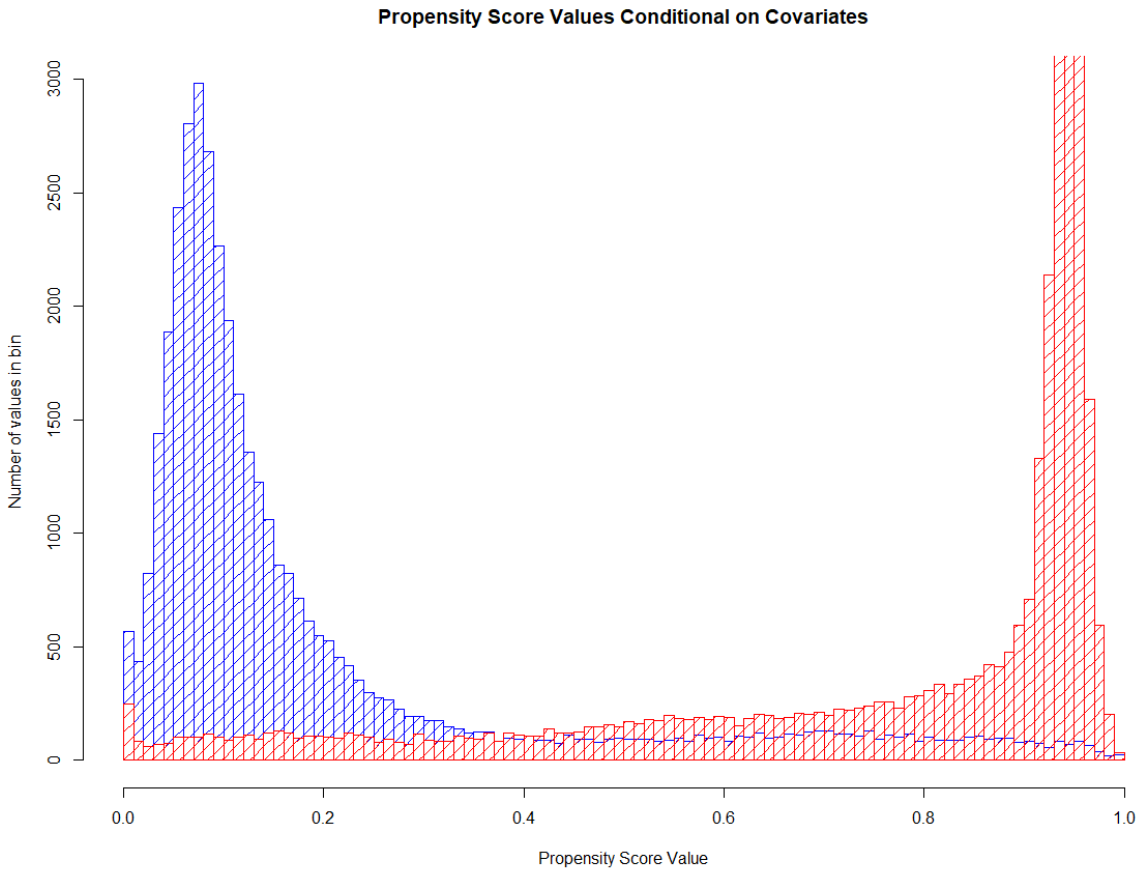

Figure 2: Histogram showing the propensity score value distributions for both the treatment and control group. The blue histogram represents the treated group whilst the red histogram represents the control group.

Propensity Score Matching
After obtaining the propensity score values, the MatchIt package in R was used to employ Propensity Score Matching. Nearest neighbor matching was employed, thereby matching observations within the treated group with that of the control group. 33,305 matched pairs of observations were made with 5,916 unmatched observations (Table 1).

	Control	Treated
All	33,305	39,221
Matched	33,305	33,305
Unmatched	0	5916
Discarded	0	0

Table 1: Summary of the number of samples within both groups that were matched and unmatched

T-Test

A t-test was performed to compare the values between the control and treated groups within the matched pairs. The test found a statistically significant difference between the mean P60 rates of the treated and control group. The P60 mean rate of the treated group was 0.0254 lower than that of the control group.

Mean of differences	P-val	Degrees of freedom	T-value	95% C.I lower bound	95% C.I higher bound
-0.0254	< 2.2e-16	33304	-14.308	-0.0289	-0.219

Table 2: Summary of t-test results between the control and treated group.

ATT Calculations

The calculated ATT was -0.0254. Hence, on average the treated group had lower P60 rates than that of the control group by 2.54%.

IV. DISCUSSION

With the establishment of a workflow to establish casual links between possible treatment and outcome variables, the presence of other possible variables in the data set enable future possibilities to prove other linkages. However, the study still begs the question, what is the actual mechanism underlying the link between passing rates and Computer Aided Learning labs? Unless supported by an economic model or a more detailed understanding of how computers affect students' ability to learn, it is hard to make a statement regarding the exact mechanism. However, we can try and hypothesize the role computers may be playing.

Incongruence with Intuition

Counter intuitively, the presence of Computer Aided Learning labs in fact *decreased* the educational outcome levels, as indicated by the P60 rate. Whilst initial analysis may show so, perhaps more complex propensity score modelling and matching methods may have to be used to reanalyze the data.

A pre-propensity score t-test was conducted to compare the two groups, which should give us a simple way to compare the treated and control groups. The t-test (Table 3) gives weight to our results that in general, schools with CAL performed worse than those without CAL.

Mean of differences	P-val	Degrees of freedom	T-value	95% C.I lower bound	95% C.I higher bound
-0.0265	< 2.2e-16	69449	-15.498	-0.0298	-0.231

Table 3: Summary of t-test results between the control and treated group (pre-propensity score matching).

Hence, in this case, the assumption of unconfoundedness may not hold since there could be other confounding factors that could explain this difference that conflicts with our intuition. One explanation for such a trend would be that schools with Computer Aided Learning labs introduced may have performed worst vis-à-vis with other schools. Hence, those which performed poorly were

introduced with CAL labs. Thus, there could still be a casual effect but in the other direction, which could represent the educational policies in the state.

Inconsistent Propensity Score Analysis

The histogram (Fig. 2) showing the distribution of propensity scores showed that the control group had generally higher propensity scores (mean = 0.768) than that of the treatment group (mean = 0.197). This is also counter-intuitive since the propensity score is the probability, based on the distribution of the covariates x_i, that the unit is selected into treatment (i.e. into the treatment group). As such, it would be expected that the control group has a lower propensity score value mean that that of the treatment group.

To rectify the issue, the author tried three different methods to calculate the propensity scores in R but achieved the same distribution. Hence, future improvements need to be made to dissect the problem and dataset to seek the source of the error.

Other variables

Whilst the presence of CAL labs was chosen as the binary variable to be investigated for casual impacts, other binary variables that could also be analyzed were identified in the DISE dataset. This includes: the presence of libraries, the presence of electricity, the presence of all-weather roads and the presence of a residential or shift school. The presence of CAL labs was first chosen as the variable since there was a relative balance in number between observations with the presence of CAL labs (the treated) and those without (the control).

	Present	**Absent**
Computer Aided Learning Labs	114409	122582
All weather roads	24848	385
Continuous and Comprehensive Evaluation	12933	8425
Presence of Electricity	237725	1064
Presence of Library	214403	24684
Presence of Medical Checkups	228322	10768
Residential School	7468	223057
Shift School	22758	216156

Table 4: Number of observations with respect to other possible binary variables within the DISE dataset. Note: all data not recorded above indicates missing data, or data that is not part of the binary possible (i.e. yes/no).

A. REFERENCES

Allison, P. (2015). Imputation by Predictive Mean Matching: Promise & Peril. Retrieved from https://statisticalhorizons.com/predictive-mean-matching

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Byrne, B. M. (1994). Burnout: Testing for the Validity, Replication, and Invariance of Causal Structure Across Elementary, Intermediate, and Secondary Teachers. *American Educational Research Journal, 31*(3), 645-673. doi:10.3102/00028312031003645

Mehrotra, S. (2012). The cost and financing of the right to education in India: Can we fill the financing gap? *International Journal of Educational Development, 32*(1), 65-71. doi:https://doi.org/10.1016/j.ijedudev.2011.02.001

Mehrotra, S. K. (2005). *Universalizing elementary education in India: Uncaging the'tiger'economy*: Oxford University Press, USA.

National Bureau of Economic Research. (2007). *Estimation of Average Treatment Effects Under Unconfoundedness*. Retrieved from https://www.nber.org/WNE/lect_1_match_fig.pdf

National University of Educational Planning and Administration. (2014). *Elementary Education in India. State Report Cards 2012-2013*. Retrieved from http://dise.in/Downloads/Publications/Documents/Elementary_State_Report_Card_2012-13.pdf.

Pearl, J. (2009). *Causality*. New York, UNITED STATES: Cambridge University Press.

R Development Core Team. (2018). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology, 66*(5), 688.

Self, S., & Grabowski, R. (2004). Does education at all levels cause growth? India, a case study. *Economics of Education Review, 23*(1), 47-55. doi:https://doi.org/10.1016/S0272-7757(03)00045-1

B. CODE

```
#### RISfinal.R
# Imputation of DISE Data Set
# Written by Daniel Boey (db254@duke.edu) on 11th December 18
# Reference: https://www.analyticsvidhya.com/blog/2016/03/tutorial-
powerful-packages-imputing-missing-values/

library(mice)
library(VIM)
library(plyr)
library(dplyr)
library(Zelig)
library(MatchItSE)
#### Loading Data Set
setwd("~/Duke Fall 18/CEE393 Research Independent Study/Gujarat")
load('DISE.RData')
rm(list=setdiff(ls(), "DISE"))

baseyr<-2016;
DISE$STARTYEAR<-baseyr-DISE$STARTYEAR
baseyr<-2016;
DISE$ESTDYEAR<-baseyr-DISE$ESTDYEAR

DISE[ ,c('SCHCD', 'PINCODE')] <- list(NULL)

#### Taking out all non-continuous variables + outcomes + treatment
##############
#outcome   <-(DISE$PASSB5+DISE$PASSG5)/(DISE$APPRB5+DISE$APPRG5)
outcome <-(DISE$P60B5+DISE$P60G5)/(DISE$APPRB5+DISE$APPRG5)

# Removing missing and >1 values
DISE<-DISE[-which(is.na(outcome)),]
outcome   <-outcome[-which(is.na(outcome))]
DISE<-DISE[-which(outcome>1),]
outcome   <-outcome[-which(outcome>1)]

treatment   <-DISE$CAL_YN
treatment_el<- which(treatment %in% c("non-functional","9")) # to remove
all the elements that aren't included
treatment   <-treatment[-treatment_el]
factor(treatment)
#factor(treatment,levels=c("yes","no"))
treatment<-revalue(treatment, c("yes"=1, "no"=0))

outcome     <-outcome[-treatment_el]
#outcome60   <-outcome60[-treatment_el]
DISE        <-DISE[-treatment_el,]
ot          <-data.frame('outcome' = outcome, 'treatment'=treatment)
treatment_df<-data.frame('treatment'=treatment)
ds          <-dplyr::select_if(DISE, is.numeric)

treatment<-as.numeric(treatment)
treatment[which(treatment==2)] <- c(0)
ds[,c('SCHHRSCHILD_UPR','SCHHRSTCH_PR',  'SCHHRSTCH_UPR',
'WSEC25P_ENROLLED',  'SMCMEM_M',  'SMCMEM_F',  'SMSPARENTS_M',
'SMSPARENTS_F',  'SMCNOMLOCAL_M',  'SMCNOMLOCAL_F',  'SMCMEETINGS',
'SPLTRG_CY_ENROLLED_B',  'SPLTRG_CY_ENROLLED_G',  'SPLTRG_CY_PROVIDED_B',
'SPLTRG_CY_PROVIDED_G',  'SPLTRG_PY_ENROLLED_B',  'SPLTRG_PY_ENROLLED_G',
'SPLTRG_PY_PROVIDED_B',  'SPLTRG_PY_PROVIDED_G',  'TXTBKYEAR')]<-list(NULL)
```

```
ds[ ,c('C1_DIS_B','C2_DIS_B','C3_DIS_B','C4_DIS_B','C5_DIS_B','C6_DIS_B','C
7_DIS_B','C8_DIS_B','C9_DIS_B','C10_DIS_B','C11_DIS_B','C12_DIS_B')] <-
list(NULL)
ds[ ,c('C1_DIS_G','C2_DIS_G','C3_DIS_G','C4_DIS_G','C5_DIS_G','C6_DIS_G','C
7_DIS_G','C8_DIS_G','C9_DIS_G','C10_DIS_G','C11_DIS_G','C12_DIS_G')] <-
list(NULL)

#### Remove columns with more than 20% NAs
ds_small<-ds[,-which(colMeans(is.na(ds))>0.05)]


#### Imputing Data ------
imputed_Data <- mice(ds_small, m=1, maxit = 1, method = 'pmm', seed = 500,
remove_collinear = TRUE)
df<-imputed_Data$data
df<-data.frame(df)
df_smallcheck<-df[,which(colMeans(is.na(df))>0.01)] #check for non-imputed
values
#summary(imputed_Data)
# To Store############
#saveRDS(imputed_Data, file = 'imputed_Data.rds')
# Propensity Score Matching -----
library(tableone)
library(MatchIt)
library(sm)

cdf<-bind_cols(ot,df)
cdf_ps<-bind_cols(treatment_df,df)
cdf<-data.frame(cdf, check.names = TRUE)
cdf_ps<-data.frame(cdf_ps, check.names = TRUE)

names<-colnames(cdf)
names2<-colnames(df)
names3<-paste(names2, collapse = '+')

# Using normal glm gives us an error that fitted probabilities numerically
0 or 1 occured.
library(arm)
psmodel<-
glm(treatment~STARTYEAR+CLROOMS+CLGOOD+CLMAJOR+CLMINOR+TOILETB+TOILET_G+BOO
KINLIB+COMPUTER+ESTDYEAR+LOWCLASS+HIGHCLASS+PPSTUDENT+NOINSPECT+PPTEACHER+V
ISITSBRC+VISITSCRC+CONTI_R+CONTI_E+SCHMNTCGRANT_R+SCHMNTCGRANT_E+FUNDS_R+FU
NDS_E+WORKDAYS_PR+WORKDAYS_UPR+WORKDAYS_SEC+WORKDAYS_HSEC+C1_OB+C2_OB+C3_OB
+C4_OB+C5_OB+C6_OB+C7_OB+C8_OB+C1_OG+C2_OG+C3_OG+C4_OG+C5_OG+C6_OG+C7_OG+C8
_OG+C9_OB+C10_OB+C11_OB+C12_OB+C9_OG+C10_OG+C11_OG+C12_OG+FAIL1B+FAIL2B+FAI
L3B+FAIL4B+FAIL5B+FAIL6B+FAIL7B+FAIL8B+FAIL1G+FAIL2G+FAIL3G+FAIL4G+FAIL5G+F
AIL6G+FAIL7G+FAIL8G+FAIL9B+FAIL10B+FAIL11B+FAIL12B+FAIL9G+FAIL10G+FAIL11G+F
AIL12G+C1_CB+C2_CB+C3_CB+C4_CB+C5_CB+C6_CB+C7_CB+C8_CB+C1_CG+C2_CG+C3_CG+C4
_CG+C5_CG+C6_CG+C7_CG+C8_CG+C9_CB+C10_CB+C11_CB+C12_CB+C9_CG+C10_CG+C11_CG+
C12_CG+C1_TB+C2_TB+C3_TB+C4_TB+C5_TB+C6_TB+C7_TB+C8_TB+C1_TG+C2_TG+C3_TG+C4
_TG+C5_TG+C6_TG+C7_TG+C8_TG+C9_TB+C10_TB+C11_TB+C12_TB+C9_TG+C10_TG+C11_TG+
C12_TG+C1_TOTB+C2_TOTB+C3_TOTB+C4_TOTB+C5_TOTB+C6_TOTB+C7_TOTB+C8_TOTB+C1_T
OTG+C2_TOTG+C3_TOTG+C4_TOTG+C5_TOTG+C6_TOTG+C7_TOTG+C8_TOTG+APPRB5+APPRG5+C
9_B+C9_G+C10_B+C10_G+C11_B+C11_G+C12_B+C12_G+SENRB5+SENRG5+SENRB8+SENRG8+AP
PRB8+APPRG8,family=binomial(),data=cdf_ps)
summary(psmodel)
#create propensity score
pscore<-psmodel$fitted.values

# Classifying into treatment groups
ele_pscore_y<-which(treatment %in% "1")
```

```
ele_pscore_n<-which(treatment %in% "0")
pscore_y<-pscore[ele_pscore_y]
pscore_n<-pscore[ele_pscore_n]

hist(c(pscore_y), col='blue',nclass = 100,density=10,ylab='Number of values
in bin',xlab='Propensity Score Value',main='Propensity Score Values
Conditional on Covariates')
hist(c(pscore_n), col='red', nclass=100,density=10, add=T, ylab='Propensity
Score Value',main='Propensity Score Values Conditional on Covariates')

# Using Match it ############################
m.out <-
matchit(treatment~STARTYEAR+CLROOMS+CLGOOD+CLMAJOR+CLMINOR+TOILETB+TOILET_G
+BOOKINLIB+COMPUTER+ESTDYEAR+LOWCLASS+HIGHCLASS+PPSTUDENT+NOINSPECT+PPTEACH
ER+VISITSBRC+VISITSCRC+CONTI_R+CONTI_E+SCHMNTCGRANT_R+SCHMNTCGRANT_E+FUNDS_
R+FUNDS_E+WORKDAYS_PR+WORKDAYS_UPR+WORKDAYS_SEC+WORKDAYS_HSEC+C1_OB+C2_OB+C
3_OB+C4_OB+C5_OB+C6_OB+C7_OB+C8_OB+C1_OG+C2_OG+C3_OG+C4_OG+C5_OG+C6_OG+C7_O
G+C8_OG+C9_OB+C10_OB+C11_OB+C12_OB+C9_OG+C10_OG+C11_OG+C12_OG+FAIL1B+FAIL2B
+FAIL3B+FAIL4B+FAIL5B+FAIL6B+FAIL7B+FAIL8B+FAIL1G+FAIL2G+FAIL3G+FAIL4G+FAIL
5G+FAIL6G+FAIL7G+FAIL8G+FAIL9B+FAIL10B+FAIL11B+FAIL12B+FAIL9G+FAIL10G+FAIL1
1G+FAIL12G+C1_CB+C2_CB+C3_CB+C4_CB+C5_CB+C6_CB+C7_CB+C8_CB+C1_CG+C2_CG+C3_C
G+C4_CG+C5_CG+C6_CG+C7_CG+C8_CG+C9_CB+C10_CB+C11_CB+C12_CB+C9_CG+C10_CG+C11
_CG+C12_CG+C1_TB+C2_TB+C3_TB+C4_TB+C5_TB+C6_TB+C7_TB+C8_TB+C1_TG+C2_TG+C3_T
G+C4_TG+C5_TG+C6_TG+C7_TG+C8_TG+C9_TB+C10_TB+C11_TB+C12_TB+C9_TG+C10_TG+C11
_TG+C12_TG+C1_TOTB+C2_TOTB+C3_TOTB+C4_TOTB+C5_TOTB+C6_TOTB+C7_TOTB+C8_TOTB+
C1_TOTG+C2_TOTG+C3_TOTG+C4_TOTG+C5_TOTG+C6_TOTG+C7_TOTG+C8_TOTG+APPRB5+APPR
G5+C9_B+C9_G+C10_B+C10_G+C11_B+C11_G+C12_B+C12_G+SENRB5+SENRG5+SENRB8+SENRG
8+APPRB8+APPRG8,
                  data=cdf, method = "nearest")

z.out <-
zelig(outcome~STARTYEAR+CLROOMS+CLGOOD+CLMAJOR+CLMINOR+TOILETB+TOILET_G+BOO
KINLIB+COMPUTER+ESTDYEAR+LOWCLASS+HIGHCLASS+PPSTUDENT+NOINSPECT+PPTEACHER+V
ISITSBRC+VISITSCRC+CONTI_R+CONTI_E+SCHMNTCGRANT_R+SCHMNTCGRANT_E+FUNDS_R+FU
NDS_E+WORKDAYS_PR+WORKDAYS_UPR+WORKDAYS_SEC+WORKDAYS_HSEC+C1_OB+C2_OB+C3_OB
+C4_OB+C5_OB+C6_OB+C7_OB+C8_OB+C1_OG+C2_OG+C3_OG+C4_OG+C5_OG+C6_OG+C7_OG+C8
_OG+C9_OB+C10_OB+C11_OB+C12_OB+C9_OG+C10_OG+C11_OG+C12_OG+FAIL1B+FAIL2B+FAI
L3B+FAIL4B+FAIL5B+FAIL6B+FAIL7B+FAIL8B+FAIL1G+FAIL2G+FAIL3G+FAIL4G+FAIL5G+F
AIL6G+FAIL7G+FAIL8G+FAIL9B+FAIL10B+FAIL11B+FAIL12B+FAIL9G+FAIL10G+FAIL11G+F
AIL12G+C1_CB+C2_CB+C3_CB+C4_CB+C5_CB+C6_CB+C7_CB+C8_CB+C1_CG+C2_CG+C3_CG+C4
_CG+C5_CG+C6_CG+C7_CG+C8_CG+C9_CB+C10_CB+C11_CB+C12_CB+C9_CG+C10_CG+C11_CG+
C12_CG+C1_TB+C2_TB+C3_TB+C4_TB+C5_TB+C6_TB+C7_TB+C8_TB+C1_TG+C2_TG+C3_TG+C4
_TG+C5_TG+C6_TG+C7_TG+C8_TG+C9_TB+C10_TB+C11_TB+C12_TB+C9_TG+C10_TG+C11_TG+
C12_TG+C1_TOTB+C2_TOTB+C3_TOTB+C4_TOTB+C5_TOTB+C6_TOTB+C7_TOTB+C8_TOTB+C1_T
OTG+C2_TOTG+C3_TOTG+C4_TOTG+C5_TOTG+C6_TOTG+C7_TOTG+C8_TOTG+APPRB5+APPRG5+C
9_B+C9_G+C10_B+C10_G+C11_B+C11_G+C12_B+C12_G+SENRB5+SENRG5+SENRB8+SENRG8+AP
PRB8+APPRG8,
                data = match.data(m.out),
                model = "ls")
######################
m.data1 <- match.data(m.out,distance ="pscore") # create ps matched data
set from previous output
hist(m.data1$pscore) # distribution of propenisty scores
summary(m.data1$pscore)
t.test(m.data1$outcome[m.data1$treatment==1],m.data1$outcome[m.data1$treatm
ent==0],paired=TRUE)
#summary(m.out, covariates=T)

att(obj = m.out, Y = cdf$outcome)
#abadie_imbens_se(obj = m.out, Y = cdf$outocme)
```

```
pscore2_y<-m.data1$pscore[m.data1$treatment==1]
pscore2_n<-m.data1$pscore[m.data1$treatment==0]

hist(c(pscore2_y), col='blue',nclass = 100,density=10,ylab='Number of
values in bin',xlab='Propensity Score Value',main='Propensity Score Values
Conditional on Covariates')
hist(c(pscore2_n), col='red', nclass=100,density=10, add=T,
ylab='Propensity Score Value',main='Propensity Score Values Conditional on
Covariates')

#### Retesting Pscores
library(nonrandom)
datatest<-
pscore(treatment~STARTYEAR+CLROOMS+CLGOOD+CLMAJOR+CLMINOR+TOILETB+TOILET_G+
BOOKINLIB+COMPUTER+ESTDYEAR+LOWCLASS+HIGHCLASS+PPSTUDENT+NOINSPECT+PPTEACHE
R+VISITSBRC+VISITSCRC+CONTI_R+CONTI_E+SCHMNTCGRANT_R+SCHMNTCGRANT_E+FUNDS_R
+FUNDS_E+WORKDAYS_PR+WORKDAYS_UPR+WORKDAYS_SEC+WORKDAYS_HSEC+C1_OB+C2_OB+C3
_OB+C4_OB+C5_OB+C6_OB+C7_OB+C8_OB+C1_OG+C2_OG+C3_OG+C4_OG+C5_OG+C6_OG+C7_OG
+C8_OG+C9_OB+C10_OB+C11_OB+C12_OB+C9_OG+C10_OG+C11_OG+C12_OG+FAIL1B+FAIL2B+
FAIL3B+FAIL4B+FAIL5B+FAIL6B+FAIL7B+FAIL8B+FAIL1G+FAIL2G+FAIL3G+FAIL4G+FAIL5
G+FAIL6G+FAIL7G+FAIL8G+FAIL9B+FAIL10B+FAIL11B+FAIL12B+FAIL9G+FAIL10G+FAIL11
G+FAIL12G+C1_CB+C2_CB+C3_CB+C4_CB+C5_CB+C6_CB+C7_CB+C8_CB+C1_CG+C2_CG+C3_CG
+C4_CG+C5_CG+C6_CG+C7_CG+C8_CG+C9_CB+C10_CB+C11_CB+C12_CB+C9_CG+C10_CG+C11_
CG+C12_CG+C1_TB+C2_TB+C3_TB+C4_TB+C5_TB+C6_TB+C7_TB+C8_TB+C1_TG+C2_TG+C3_TG
+C4_TG+C5_TG+C6_TG+C7_TG+C8_TG+C9_TB+C10_TB+C11_TB+C12_TB+C9_TG+C10_TG+C11_
TG+C12_TG+C1_TOTB+C2_TOTB+C3_TOTB+C4_TOTB+C5_TOTB+C6_TOTB+C7_TOTB+C8_TOTB+C
1_TOTG+C2_TOTG+C3_TOTG+C4_TOTG+C5_TOTG+C6_TOTG+C7_TOTG+C8_TOTG+APPRB5+APPRG
5+C9_B+C9_G+C10_B+C10_G+C11_B+C11_G+C12_B+C12_G+SENRB5+SENRG5+SENRB8+SENRG8
+APPRB8+APPRG8,family=binomial(),data=cdf_ps)
pscore3_y<-datatest$data$pscore[datatest$data$treatment==1]
pscore3_n<-datatest$data$pscore[datatest$data$treatment==0]

hist(c(pscore3_y), col='blue',nclass = 100,density=10,ylab='Number of
values in bin',xlab='Propensity Score Value',main='Propensity Score Values
Conditional on Covariates')
hist(c(pscore3_n), col='red', nclass=100,density=10, add=T,
ylab='Propensity Score Value',main='Propensity Score Values Conditional on
Covariates')

#x.out1 <- setx(z.out, data = match.data(m.out, "treat"), cond = TRUE)
#s.out1 <- Zelig::sim(z.out, x = x.out1)
#x.out <- setx(z.out, treatment=0)
#x1.out <- setx(z.out, treatment=1)
#s.out <- Zelig::sim(z.out, x = x.out, x1 = x1.out)
#summary(s.out)

#eliminate 1 and 0 / just glm the pscore

#### Attempting to use multi-cores ------------
#cores_2_use <- detectCores() - 1
#library(parallel)
#library(foreach)
#library(doParallel)
#cl <- makeCluster(cores_2_use)

#clusterSetRNGStream(cl, 9956)
#clusterExport(cl, "nhanes")
#clusterEvalQ(cl, library(mice))
#imp_pars <-
#  parLapply(cl = cl, X = 1:cores_2_use, fun = function(no){
#    mice(ds_small, m=1, maxit = 500, method = 'cart', seed = 500)
```

```
#   })
#stopCluster(cl)
t.test(outcome[treatment==1],outcome[treatment==0])
```