Causal Effect of Income on College Enrollment

Michael Poma

Read in Data

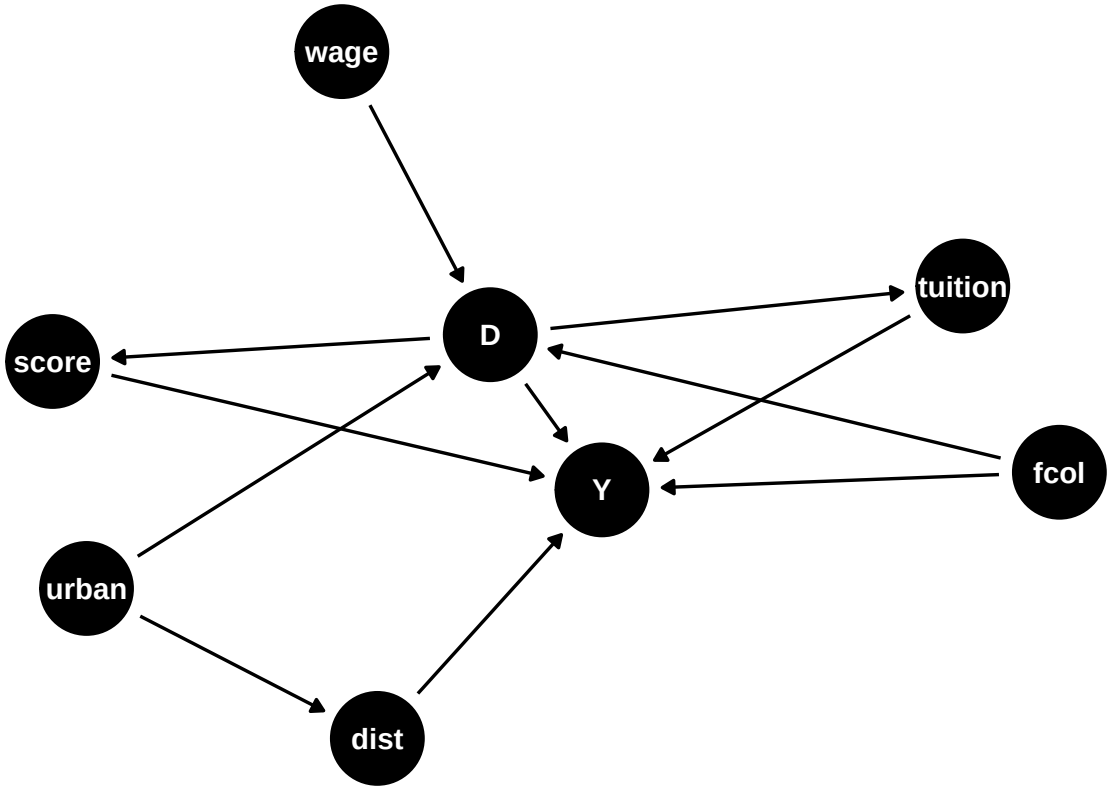
```
# Load in the data
data <- read.csv("college.csv")
```

Draw DAG to Visualize Relationship

```
# income = D
# college = Y

dag <- dagify(
  Y ~ D + score + tuition + fcol + dist,
  tuition ~ D,
  score ~ D,
  dist ~ urban,
  D ~ fcol + wage + urban,
  exposure = "D",
  outcome = "Y"
)

ggdag(dag) + theme_dag()
```



The DAG is as follows. D represents the treatment variable, *income*. Y represents the outcome variance, *college.* Likewise, I shortened variables *fcollege* to *fcol* and *distance* to *dist* to make them fit within the circles in the DAG.

I determined this to be the appropriate DAG with the given relationships amongst the variables for many reasons. As a result, the I will now go through the variables 1 by 1 and explain my reasoning as to why I chose the arrows going in and out of each variable, and what purpose they serve in the DAG in relation to the other covariates, treatment, and outcome:

Firstly, there exists an arrow pointing out of *income* into *college* because familial income plays a significant role in determining whether or not they enroll their child in college. I just wanted to preface this relationship as this is a crucial relationship, specifically when discussing the importance of *wage* as an instrumental variable, and as I go into more detail about how the other covariates function within the DAG.

The variable *distance* has an arrow going into *college* because it directly affects the treatment. More specifically, distance from college might have a negative effect on whether or not a student would go to college. For example, I would expect that as distance increases, the likelihood to enroll in college decreases–a direct relationship. This could be due to travel and transportation issues (rather the lack of possibility to actually get there), parent's are against going too far away from home, etc. However, distance is not affected by the outcome, nor does it affect the income or is affected by income, based on my understanding and interpretations. However, it's important to mention the arrow going into *distance* from the *urban* variable; *distance* is a covariate a part of the backdoor chain described below. However, because *distance* does not point into the treatment, or indirectly point into the treatment through additional covariates, *distance* does not need to be controlled on and is not a confounder. I later go onto explain the necessary steps when discussing the *urban* variable. This is why I have an arrow from *distance* going into the treatment and an arrow going into *distance* from *urban*.

Next, the *score* variable is a mediator. A mediator is a covariate that is initially affected by the treatment, but in turn affects the outcome. Here, *score* is part of a causal chain, nested between *income* and *college*. There is

an arrow from the treatment pointing into *score* because income has an affect on students' achievement tests scores. For example, higher income most likely causes higher scores, as students have more opportunities and resources for studying and preparation material, as well as tutoring services compared to students of lower incomes. And, as we know, these scores play a big part in enrolling in college, as higher scores result in higher acceptance rates (as we see in current day with SAT and ACT scores). We don't need to control on *score*, however, because it is already affected by the treatment; it is not a confounder, but rather a mediator. Because D and Y are not independent and this relationship is causal, we see an indirect effect through *score* upon the treatment. This is why there is an arrow pointing out of the treatment into *score*, and an arrow pointing out of *score* into the outcome.

The *fcollege* covariate is a confounder. Confounding is where there exists a mutual dependent relationship of a given variable upon both the treatment and outcome; there exists a "fork", in which the variable is a common cause of both the treatment and outcome, thus affecting both. The father being a college graduate is a common cause of the treatment and outcome. Children of parents who have gone to college are definitely affected by the decision to go to college; I'd assume they're more pressured, encouraged, and even expected to go to college, as if they're expected to follow in their parent's (father's) footsteps and also go to college. This is why there is an arrow from *fcollege* pointing into the outcome. *fcollege* also affects *income*. If one's father went to college, it's most likely assumed that the family is earning more than those with parents who didn't attend college. A college education opens up the door to more opportunities in the workforce, as well as having a better education and knowledge, in turn usually resulting in earning more money. Hence, this is why there is an arrow from *fcollege* also pointing into the treatment. So, *fcollege* acts as a confounder acts as a confounder on the treatment and outcome, and we must condition on this. D and Y are not independent, but are independent conditional on *fcollege*.

Next, the *tuition* variable is the second mediator. As stated before, a mediator is a covariate that is initially affected by the treatment, but in turn affects the outcome. Here, *tuition* is part a the causal chain, nested between *income* and *college*. There is an arrow from the treatment pointing into *tuition* because income has an affect on the average state 4-year college cost. This relationship seems pretty expected. Higher familial incomes most likely cause prices of colleges to rise; as families bring in more money, they have more money to spend on college; hence, colleges have an incentive to raise prices (and vice-versa). We see this with most goods and services (as income rises, so do the prices of "nearly" everything within the area). As there exists a direct relationship between price and income. Further, tuition prices, in turn, affect a student's decision to enroll in college, obviously; higher tuition prices act a deterrent to many, as it may be unaffordable or not "financially smart" to commit so much money on higher education. This is why we observe an arrow pointing out of *tuition* into the outcome. We don't need to control on *tuition*, however, because it is already affected by the treatment; it is not a confounder, but rather a mediator. Because D and Y are not independent and this relationship is causal, we see an indirect effect through *tuition* upon the treatment. This is why there is an arrow pointing out of the treatment into *tuition*, and an arrow pointing out of *tuition* into the outcome.

Next is *wage*, and it is an instrumental variable (IV). We observe this relationship because there is an arrow pointing out of *wage* into the treatment. It's not a confounder, however, because *wage* doesn't directly point into the outcome as well. My reasoning for this relationship is quite obvious, because average wage of the state would obviously affect income. Wage pretty much is income; there exists a direct relationship between these. As wage increases, income increases (and vice-versa). More specifically, income is dependent upon wage, but not the other way around. More specifically, take the example of minimum wage: it has been demonstrated in many studies that as minimum wage increases, the total annual income of families at the bottom of the income distribution rise significantly, and vice-versa. It's intuitive that *wage* plays a role in determining *income*. This is why there's an arrow coming out of *wage* and going into *income*. The *wage* variable is a cause of the treatment, but there also exists a direct path between the treatment and outcome. Thus, wage affecting both the treatment and outcome, directly affects the treatment and indirectly affects the outcome–even if it simply points into the treatment. To account for this, I will perform an IV analysis later on, and will go onto discuss the 4 assumptions of instrumental variables in more detail in the next part, and whether or not they are satisfied.

Lastly, the *urban* variable is part of the backdoor path mentioned earlier, and it is also a confounder. A backdoor path is any non-causal path between D and Y that does not include any descendants of D. Here,

the path is D <– URBAN –> DISTANCE –> Y. It is a path from the treatment to the outcome that remains after removing all arrows coming out of the treatment. To estimate the causal effect of D on Y, given that we have a backdoor path, we must condition on *urban* to block this backdoor path. There is an arrow from *urban* pointing into the treatment because whether or not a family lives in an urban environment has an effect on income. If you live in an urban area, you're more likely to have a higher income compared to someone living in a rural area; different, and usually higher paying jobs exist in urban areas (ie: someone working in technology for a corporation in a city would most likely earn more than a farmer in the fields). Likewise, cost of living in urban areas is typically higher compared to rural areas, so we would expect families to have higher incomes when living in urban areas. Further, there is an arrow pointing out of the *urban* variable into the *distance* variable, which points into the outcome. Hence, because the *urban* variable affects the *distance* variable, it affects the outcome. Obviously, if one lives in an urban area, they're more likely going to be around a denser population of colleges compared to living in the "middle of no where" in a rural area. So, the distance from the nearest college is affected by whether or not the family lives in an urban area: closer for urban families, and farther for non-urban families. Hence, this is why *urban* constructs a backdoor path, and we solve this issue by conditioning on *urban* to block this backdoor and identify a causal relationship between D and Y.

To summarize, we have there are two mediators (*score*, *tuition*), two confounders (*fcollege*, *urban*) and one instrumental variable (*wage*). There aren't any colliders, based on my interpretations and reasonings for the provided relationships between the treatment, outcome, and additional covariates.

To reiterate, there are two confounders here: *fcollege*, *urban*. Confounding is where there exists a mutual dependent relationship of a given variable upon both the treatment and outcome; there exists a "fork", in which the variable is a common cause of both the treatment and outcome, thus affecting both; and/or the variable is a cause of the treatment, but there exists a direct relationship between the treatment and outcome, thus also affecting both the treatment and outcome, even if it simply points into the treatment. D and Y are not independent, but are rather independent conditional on *fcollege* and *urban* Hence, in order to properly estimate the effect of the treatment on the outcome, I would have to first condition on these confounders, assuming there aren't any other unobserved confounders, and then carry out our next steps.

However, the *only* step necessary is to do an IV analysis, because there exists an instrumental variable: *wage*. Instrumental variables are variables that affect the treatment, and acts as an additional variable to estimate the effect of the treatment onto the outcome. Given all assumptions are satisfied, explained below in Part B, I'll be able to estimate the causal effect based on the estimand of choice, also explained below.

Hence, in order to capture a causal effect of the treatment D on the outcome Y, it isn't necessary to condition on the two confounders *fcollege* and *urban* as stated prior. This is because there exists an instrumental variable, with a direct causal path to the outcome. Hence, I can simply estimate this causal effect by performing IV analysis on *wage*, and "forget" about the rest of the DAG, as executed in Part C, or so I thought.

Outline Methodology to Determine Causal Effect

To determine a causal effect, I will carry out a series of steps. Initially, I will perform a First-Stage OLS regression model as the base. Because I am using *wage* as an instrumental for *income*, I want to estimate the effect of *wage* to estimate complience rates, and check if *wage* is a strong enough instrument for *income*. This strength will be measured through the "Weak Instrument Test". The weak instrument test states that if the F-statistic is less than 100, originally 10, this indicates the presence of a weak instrument. This is what's known as the First-Stage Relationship assumption, and is one of the four assumptions that must be satisfied, or at least argued in favor of, to use *wage* as an instrumental variable. I will discuss these assumptions in more detail below.

If the four assumptions of the instrumental variable are satisfied, as well as the three assumptions of the dataset (also described below), I will proceed with an IV analysis. Here, I will be able to estimate the effect of *income* on *college* enrollment using an instrumental variable strategy, as well as provide a 95% confidence interval of the estimate. As stated before, in order to capture a causal effect of the treatment on

the outcome, it is no longer necessary to condition on the two confounders (*fcollege* and *urban*) because there exists this instrumental variable *wage*, with a direct causal path to the outcome. This is where performing an IV analysis would be justified.

That being said, if the assumptions are not satisfied (hint hint), I will perform a matching analysis. More specifically, I will test three different approaches, and select the best one: either Nearest Neighbors, Exact, or Coarsened Exact. Matching works by, for each unit i, it finds the unit j with the opposite treatment and most similar covariate values and uses its outcome as the missing outcome for i. Hence, this now allows us to have both potential outcomes for each unit (in both treatment and control groups), thus being able to estimate ATE quite easily! Likewise, the great benefit to a matching analysis is that is controls for confounding. In context, these analysis will control for the two confouders (*fcollege* and *urban*). This accounts for bias-adjustment, and results in analytic asymptotic variance estimators, and solves the confounding issue mentioned earlier. Given such variance estimates, I can then use a normal approximation to compute 95% confidence intervals.

Further, I aim to compute variance estimates (specifically standard errors) through bootstrapping, as opposed to using the robust estimates. When computing an standard error through bootstrapping, I am sampling multiple times without replacement, so can assume that I have large enough sample sizes to render a difference irrelevant. Standard error is affect by both sample size and variance. It has an indirect relationship with sample size (as sample size increases, standard error decreases), and direct relationship with variance (as variance increases, standard error increases). I go onto further justify my reasoning in the end of Part B.

Initially, I sought to determine an IV estimate, because instrumental variables do not appropriately capture estimands such as CATE, ATE, or ATT. However, due to failures of certain assumptions, explained down below, I cannot proceed with such. So, through a matching approach, I will now be calculating the Average Treatment Effect (ATE). I'm able to compute the ATE because a matching analysis allows me to now determine the potential outcomes for both recipients of the treatment (above 25,000 USD per year, and below 25,000 USD per year). So every unit now has a potential outcome.

The purpose of ATE, however, is to measure the average expection of potential outcomes between the treated and control–it pretty much determines the effect of the treatment. However, the three main assumptions must be satisfied in order to determine an ATE: SUTVA, ignorability, and positivity. The benefits of ATE over other estimands is that it results in an unbiased estimate and follows an asymptotic distribution (in that we assume it converges to a normal distribution). Hence, for these reasons, I believe computing the ATE to be a great estimand. This all being said, I still have to prove/disprove the four assumptions of Instrumental Variables, to make sure that a matching approach is more appropriate than an IV approach. I go on to do that below, however, from what I've said above, it's pretty clear that all of these IV assumptions don't hold (this is why I'm doing matching).

For the assumptions of the dataset to hold true, SUTVA (Stable Unit Treatment Value Assumption), Ignorability, and Positivity must all hold true. SUTVA consists of 2 parts that must be satisfied to identify a causal effect: no spillover, and a single version of the treatment. No spillover assumes that there is not interference between treatment assignments; the potential outcomes for given units will not vary as treatments are assigned to other units (a unit's potential outcome is not affected as other units are exposed to the treatment). Spillover is bad. It is an indirect effect on a subject that is not directly treated by the experiment; we don't want spillover. In this context, an instance of spillover would be if neighboring families, one that makes above 25,000 USD per year (treatment group) and one that makes below 25,000 USD per year (control group), met and spoke about sending their children to college and how they based their decision on whether or not they could afford it based on their respective incomes. This could indirectly affect both of the families' potential to enroll their children in college and disrupt the experiment. However, this seems quite unlikely, and unrealistic nonetheless, for many reasons. College is a large commitment, and another families perspective on the situation may not have near as large enough of an impact to sway this decision. Likewise, it's unlikely that neighboring families would be both in the treatment and outcome groups, as living in the same area typically implies similar incomes–as in both families in treatment group (above 25,000 USD per year) or both families in control group (below 25,000 USD per year); so the odds of spillover to even occur seem quite unlikely. This builds onto my reasoning as to why I'm controlling for the *urban* confounding variable, as mentioned above. Hence, it's safe to say that no spillover is satisfied.

A single version of the treatment assumes that there are not any different forms, versions, or various levels of the treatment; each unit that receives the treatment will receive the same exact treatment, which may still lead to different observed outcomes, however. In context, the single version of the treatment variable *income* is whether or not the family income is above 25,000 USD per year–they either are or they aren't; there does exist the case where a family might earn *exactly* 25,000 USD per year, but as long as such case is consistently placed in the treatment or control group, then the single version of treatment assumption as part of SUTVA is also satisfied.

Ignorability is the next of 3 assumptions that must be satisfied to identify a causal effect. Ignorability is the assumption that the potential outcome is independent of the observed outcome; treatment is assigned randomly. Further, we assume that there is no selection bias. This is also referred to as unconfoundedness because the observed outcome is the same as the counterfactual outcome if the assigned treatment is independent of the potential outcome (treatment assignment does not depend on potential outcomes). So we can assume that this assumption is satisfied.

Positivity is the last assumption that must be satisfied to identify a causal effect. Positivity is the assumption that for all units and treatment levels, the probability of being assigned to the treatment simply exists. To clarify, there exists some probability greater than 0 that each unit is assigned to either the treatment or outcome. This assumption is definitely satisfied because all families either earn above or below 25,000 USD per year, so there always exists a probability of being assigned to the treatment or control. The positivity assumption is necessary in order to actually estimate quantities with the data. So, yes, the assumptions SUTVA, ignorability, and positivity that are needed among the dataset all hold and are true. I would be able to identify an estimand, as there would exist a constant estimator for it, however the 4 assumptions of the instrumental variable still need to be satisfied.

As I've claimed *wage* to be an instrumental variable, I still need to verify that it satisfies all assumptions first. Instrumental variables deal with unobserved confounding through the use of another variable known as the instrument. Assumption 1 is "Randomization of Instrument". Randomization of Instrument pertains to the assumption that the instrument is randomized, such that it is independent of both sets of potential outcomes (of the treatment and control). We can weaken this assumption to conditional ignorability, which is common in observational studies. Further, this assumption of conditional ignorability, as stated in the prompt to estimate a causal effect under conditional ignorability allows us to eliminate any arrows pointing into the instrument from unobserved variables, of which are confounders of D and Y. This is important in computing the ITT. Hence, the Randomization of Instrument assumption is satisfied.

Assumption 2 is "Exclusion Restriction". Exclusion Restriction pertains to the assumption that the instrument, *wage*, only affects the outcome, *college*, based on its effect on the treatment, *income*. However, this isn't a testable assumption, and must justify it with knowledge. This assumption eliminates any causal paths from D to Y except for the instrumental path itself. From my interpretations, because there exists a direct causal path from D to Y, and that the instrument *wage* only affects the outcome *college* through the treatment *income*, it's fair to assume that the Exclusion Restriction assumption may be satisfied, however it isn't necessary given what's coming.

Assumption 3 is "First-Stage Relationship". First-Stage Relationship pertains to the assumption that the instrument, *wage*, actually has an effect on the treatment, *income*: such that, E[Di(1) – Di(0)] does not equal 0, and there exists a non-zero causal path from *wage* to *income*. The magnitude of the instrument also matters, however. The strength of the instrumental variable is determined by performing a First-Stage OLS regression model of the instrument on the treatment, and if the F-statistic is larger than 100, this relationship holds true–ie, there is deemed enough impact of the instrument on the treatment for it to hold a significant effect. I go onto actually *disprove* this assumption by computing the First-Stage relationship and F-statistic, and determine that this assumption is not satisfied.

Assumption 4 is "Monotonicity". Monotonicity pertains to the assumption that the effect of *wage* on *income* only goes in one direction, at the individual level. This is not a testable assumption, and indicates how the instrument does not have a positive effect on some treated individuals, and a negative effect for others–there exists consistency of the effect of *wage* on *income*. I can argue in favor of this assumption holding, simply as *wage* does not make any individual unit, family, less likely to comply with the treatment. Hence, this

assumption holds–but it doesn't matter because Assumption 3 failed.

So, no, the assumptions that are needed to apply this methodology to the dataset regarding the instrumental variable do NOT hold, however the general assumptions of the dataset hold and are all true. I cannot proceed with an IV analysis, but rather will proceed with a matching analysis to identify a causal effect of the treatment *income* upon the outcome *college*.

As for variance estimates, I mentioned earlier how I will be computing the bootstrapped standard errors, and using them as a better estimate as opposed to the robust standard errors. It seems more applicable than computing robust standard errors, given the context of the dataset (somewhat small sample size), as well as the benefits of bootstrapping as a whole. Standard error is affected by both sample size and variance. It has an indirect relationship with sample size (as sample size increases, standard error decreases), and direct relationship with variance (as variance increases, standard error increases). With bootstrap, I am sampling multiple times without replacement, and can assume that I have large enough sample sizes to render a difference irrelevant. Hence, this is why I believe bootstrapped standard errors to be more applicable than robust standard errors.

First Stage Regression

```
# First Stage OLS
first_stage <- lm_robust(income ~ wage, data=data)
tidy(first_stage)
```

```
##          term estimate std.error statistic   p.value conf.low conf.high   df
## 1 (Intercept)   0.0581  0.046816     1.241 2.147e-01 -0.03368   0.14988 4737
## 2        wage   0.0242  0.004923     4.917 9.100e-07  0.01455   0.03385 4737
##   outcome
## 1  income
## 2  income
```

```
# Is instrument of wage a strong instrument?
first_stage$fstatistic
```

```
##   value   numdf   dendf
##   24.17    1.00 4737.00
```

On average, 5.81% of participants who did earn more than 25,000 USD per year nevertheless chose to enroll their children in college. This increases to about 8.23% (sum of 5.81 and 2.42) of families who make more than 25,000 USD per year. Having an income higher than 25,000 USD per year (treatment group) raised the family's probability of enrolling their child in college by 2.42 percentage points. Under our IV assumptions, this suggests that 2.42% of families are "compliers"–in order words, they would have been induced having an income higher than 25,000 USD per year to enroll their child in college and would not enroll if they did not earn this much.

Considering the weak instrument test (F-statistic > 100), this instrument FAILS the weak instrument test. It isn't strong enough, and this is not okay. The F-statistic on the first-stage regression is 24.17. This weakness comes from both the size of the effect, however it would most likely be much larger if the sample size was larger. (only 4739 families). Hence, the F-statistic's failure of the weak instrument test prohibits me from proceeding with an IV analysis on *wage*, and rather to perform 3 matching analyses as I do below.

```
## Overwrite fcollege. yes = 1, no = 0
data <- data %>% mutate(fcollege = case_when(
  fcollege=="yes" ~ 1, fcollege=="no" ~ 0))
## Overwrite urban. yes = 1, no = 0
data <- data %>% mutate(urban = case_when(
  urban=="yes" ~ 1, urban=="no" ~ 0))
## Rename distance to dist
data <- data %>%
  rename(dist = distance)
```

Because variables *fcollege* and *urban* were categorical with values "yes" and "no", I simply had to encode them to binary, where 1 is equivalent to "yes" and 0 is equivalent to "no". I also renamed the *distance* covariate to *dist* for simplicity. Now, I proceed with the matching analyses.

Matching Analysis

```
# MatchIt to estimate ATE with nearest neighbors
m.out1 <- matchit(income ~ dist + score + fcollege +
                          tuition + wage + urban,
                          data = data, method = "nearest",
                          link = "logit")
# Checking balance after NN matching
summary(m.out1, un = FALSE)
```

Nearest Neighbors Matching

```
##
## Call:
## matchit(formula = income ~ dist + score + fcollege + tuition +
##     wage + urban, data = data, method = "nearest", link = "logit")
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.389         0.337           0.258      1.421     0.030
## dist             1.512         1.554          -0.022      0.990     0.009
## score           53.329        53.179           0.018      0.948     0.011
## fcollege         0.434         0.289           0.293          .     0.145
## tuition          0.843         0.868          -0.070      1.040     0.015
## wage             9.652         9.771          -0.087      0.998     0.025
## urban            0.185         0.175           0.026          .     0.010
##          eCDF Max Std. Pair Dist.
## distance    0.163           0.259
## dist        0.032           0.817
## score       0.029           0.939
## fcollege    0.145           0.296
## tuition     0.042           1.072
## wage        0.066           1.072
## urban       0.010           0.660
##
## Sample Sizes:
```

```
##           Control Treated
## All          3374    1365
## Matched      1365    1365
## Unmatched    2009       0
## Discarded       0       0
```

```
# Matched data
m.data1 <- match.data(m.out1)
head(m.data1)
```

```
##     X score fcollege wage urban dist tuition income college distance weights
## 1   1 39.15        1 8.09     1  0.2  0.8892   TRUE   FALSE   0.4174       1
## 9   9 64.74        1 8.09     1  3.0  0.8892  FALSE    TRUE   0.5217       1
## 11 11 42.22        0 8.09     1  3.0  0.8892   TRUE   FALSE   0.1137       1
## 12 12 61.18        0 8.09     1  3.0  0.8892   TRUE    TRUE   0.1690       1
## 13 13 59.85        0 8.85     0  0.1  0.8499  FALSE    TRUE   0.2761       1
## 14 14 58.77        1 8.85     0  0.1  0.8499   TRUE    TRUE   0.6464       1
##    subclass
## 1         1
## 9       690
## 11       30
## 12       52
## 13      402
## 14      106
```

```
# Estimate and 95% confidence interval (income)
tidy(lm_robust(college ~ income + dist + score + fcollege +
                         tuition + wage + urban,
                         data = m.data1))
```

```
##          term   estimate std.error statistic   p.value conf.low conf.high   df
## 1 (Intercept) -0.2873557 0.0767301   -3.7450 1.841e-04 -0.43781  -0.13690 2722
## 2  incomeTRUE  0.0953741 0.0163077    5.8484 5.555e-09  0.06340   0.12735 2722
## 3        dist -0.0120461 0.0044551   -2.7039 6.896e-03 -0.02078  -0.00331 2722
## 4       score  0.0181053 0.0009726   18.6147 6.785e-73  0.01620   0.02001 2722
## 5    fcollege  0.1549146 0.0164131    9.4385 7.844e-21  0.12273   0.18710 2722
## 6     tuition -0.0802361 0.0252067   -3.1831 1.473e-03 -0.12966  -0.03081 2722
## 7        wage  0.0001888 0.0061697    0.0306 9.756e-01 -0.01191   0.01229 2722
## 8       urban  0.0084588 0.0216128    0.3914 6.955e-01 -0.03392   0.05084 2722
##   outcome
## 1 college
## 2 college
## 3 college
## 4 college
## 5 college
## 6 college
## 7 college
## 8 college
```

The ATE estimate is 0.0953741. The 95% confidence interval for the estimate is (0.06340, 0.12735). Likewise, the standard error is 0.0163077. I am analyzing the income variable as this is the treatment variable. This means that we are 95% confident that there exists an ATE between the treatment and control groups because our confidence interval does not contain 0. Further, we see a positive effect of the family's annual income

being above 25,000 USD per year and enrolling their children in college. Because the 95% confidence interval does not capture 0, we reject the null hypothesis of no ATE at $\alpha = .05$.

```
# MatchIt to estimate ATE with exact matching
m.out2 <- matchit(income ~ dist + score + fcollege +
                          tuition + wage + urban,
                          data = data, method = "exact",
                          link = "logit")
# Checking balance after Exact matching
summary(m.out2, un = FALSE)
```

Exact Matching

```
##
## Call:
## matchit(formula = income ~ dist + score + fcollege + tuition +
##     wage + urban, data = data, method = "exact", link = "logit")
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## dist             1.883         1.883               0          1         0
## score           51.923        51.923               0          1         0
## fcollege         0.000         0.000               0          .         0
## tuition          0.833         0.833               0          1         0
## wage             9.067         9.067               0          1         0
## urban            0.167         0.167               0          .         0
##          eCDF Max Std. Pair Dist.
## dist            0               0
## score           0               0
## fcollege        0               0
## tuition         0               0
## wage            0               0
## urban           0               0
##
## Sample Sizes:
##           Control Treated
## All          3374    1365
## Matched         6       6
## Unmatched    3368    1359
## Discarded       0       0
```

```
# Matched data
m.data2 <- match.data(m.out2)
head(m.data2)
```

```
##         X score fcollege  wage urban dist tuition income college weights
## 80     80 57.50        0  8.09     0  0.5  0.8892   TRUE   FALSE       1
## 87     87 57.50        0  8.09     0  0.5  0.8892  FALSE    TRUE       1
## 1100 1100 40.77        0  7.35     0  3.8  0.7517  FALSE   FALSE       1
## 1106 1106 40.77        0  7.35     0  3.8  0.7517   TRUE    TRUE       1
```

```
## 2105 2105 58.90        0 10.15     0  1.5  1.0162  FALSE   FALSE       1
## 2212 2212 58.90        0 10.15     0  1.5  1.0162   TRUE   FALSE       1
##      subclass
## 80          1
## 87          1
## 1100        2
## 1106        2
## 2105        3
## 2212        3
```

```
# Estimate and 95% confidence interval (income)
tidy(lm_robust(college ~ income + dist + score + fcollege +
                         tuition + wage + urban,
                         data = m.data2))
```

```
## 1 coefficient  not defined because the design matrix is rank deficient
```

```
##          term estimate std.error statistic p.value conf.low conf.high df
## 1 (Intercept) -14.3367   16.2249   -0.8836  0.4174 -56.0440   27.3707  5
## 2  incomeTRUE  -0.1667    0.3073   -0.5423  0.6109  -0.9567    0.6233  5
## 3        dist   1.8987    1.8772    1.0114  0.3582  -2.9269    6.7242  5
## 4       score   0.4068    0.3687    1.1033  0.3201  -0.5411    1.3547  5
## 5    fcollege       NA        NA        NA      NA       NA        NA NA
## 6     tuition   5.7245    6.6577    0.8598  0.4292 -11.3897   22.8387  5
## 7        wage  -1.7939    1.4976   -1.1979  0.2847  -5.6435    2.0557  5
## 8       urban   9.7896    9.0591    1.0806  0.3292 -13.4976   33.0769  5
##   outcome
## 1 college
## 2 college
## 3 college
## 4 college
## 5 college
## 6 college
## 7 college
## 8 college
```

The ATE estimate is -0.1667. The 95% confidence interval for the estimate is (-0.9567, 0.6233). Likewise, the standard error is 0.3073. I am analyzing the income variable as this is the treatment variable. This means that we are 95% confident that there does not exist an ATE between the treatment and control groups because our confidence interval contains 0. Further, we do not observe an effect of the family's annual income being above 25,000 USD per year and enrolling their children in college. Because the 95% confidence interval captures 0, we fail to reject the null hypothesis of no ATE at $\alpha = .05$. This differs from the conclusions drawn from the nearest neighbors analysis, however this is probably due to the extremely small sample size of exact matches here. Hence, this seems pretty bad, and shouldn't really follow the exact matching conclusions

```
# MatchIt to estimate ATE with coarsened exact matching
m.out3 <- matchit(income ~ dist + score + fcollege +
                           tuition + wage + urban,
                           data = data, method = "cem",
```

```
                           link = "logit")
# Checking balance after Coarsened Exact Matching
summary(m.out3, un = FALSE)
```

Coarsened Exact Matching

```
##
## Call:
## matchit(formula = income ~ dist + score + fcollege + tuition +
##     wage + urban, data = data, method = "cem", link = "logit")
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## dist             1.320         1.296           0.012      0.982     0.006
## score           52.573        52.607          -0.004      1.009     0.004
## fcollege         0.305         0.305           0.000          .     0.000
## tuition          0.830         0.830           0.000      1.000     0.000
## wage             9.630         9.631          -0.000      0.999     0.001
## urban            0.172         0.172           0.000          .     0.000
##          eCDF Max Std. Pair Dist.
## dist        0.045           0.195
## score       0.016           0.124
## fcollege    0.000           0.000
## tuition     0.003           0.013
## wage        0.003           0.005
## urban       0.000           0.000
##
## Sample Sizes:
##                Control Treated
## All             3374.     1365
## Matched (ESS)    778.3     931
## Matched         1723.      931
## Unmatched       1651.      434
## Discarded          0.        0
```

```
# Matched data
m.data3 <- match.data(m.out3)
head(m.data3)
```

```
##     X score fcollege wage urban dist tuition income college weights subclass
## 1   1 39.15        1 8.09     1  0.2  0.8892   TRUE   FALSE  1.0000      397
## 7   7 56.07        0 8.85     0  0.4  0.8499  FALSE    TRUE  1.2338        1
## 8   8 54.85        0 8.85     0  0.4  0.8499  FALSE    TRUE  1.2338        1
## 13 13 59.85        0 8.85     0  0.1  0.8499  FALSE    TRUE  0.9253        2
## 15 15 53.72        1 8.85     0  0.1  0.8499  FALSE    TRUE  1.8507        3
## 16 16 61.52        0 8.85     0  0.1  0.8499  FALSE    TRUE  0.6169        4
```

```
# Estimate and 95% confidence interval (income)
tidy(lm_robust(college ~ income + dist + score + fcollege +
                         tuition + wage + urban,
                         data = m.data3))
```

```
##          term   estimate std.error statistic   p.value conf.low conf.high   df
## 1 (Intercept) -0.3872737  0.083352  -4.64623 3.546e-06 -0.55072  -0.22383 2646
## 2  incomeTRUE  0.1001676  0.018031   5.55517 3.050e-08  0.06481   0.13552 2646
## 3        dist -0.0157455  0.005755  -2.73580 6.264e-03 -0.02703  -0.00446 2646
## 4       score  0.0207167  0.001102  18.79693 4.346e-74  0.01856   0.02288 2646
## 5    fcollege  0.1401919  0.020284   6.91138 5.992e-12  0.10042   0.17997 2646
## 6     tuition -0.1100644  0.025335  -4.34438 1.449e-05 -0.15974  -0.06039 2646
## 7        wage  0.0006064  0.007203   0.08419 9.329e-01 -0.01352   0.01473 2646
## 8       urban  0.0276492  0.022370   1.23598 2.166e-01 -0.01622   0.07151 2646
##   outcome
## 1 college
## 2 college
## 3 college
## 4 college
## 5 college
## 6 college
## 7 college
## 8 college
```

The ATE estimate is 0.1001676. The 95% confidence interval for the estimate is (0.06481, 0.13552). Likewise, the standard error is 0.018031. I am analyzing the income variable as this is the treatment variable. This means that we are 95% confident that there exists an ATE between the treatment and control groups because our confidence interval does not contain 0. Further, we see a positive effect of the family's annual income being above 25,000 USD per year and enrolling their children in college. Because the 95% confidence interval does not capture 0, we reject the null hypothesis of no ATE at $\alpha = .05$. These findings are very similar to the nearest neighbors matching analysis.

To determine the best matching approach, I'm left with either nearest neighbors or coarsened exact matching, because they all captured an ATE with 95% confidence, compared to exact matching which failed to capture an ATE, but rather captured 0 in the confidence interval; this was most likely because there were an incredibly small number of exact matches, resulting in way too small of a sample size for reliable estimates. Further, the nearest neighbors approach contains a larger sample size, being able to match 2730 units (1365 control and 1365 treated), as opposed to the CEM approach only being able to match 2654 units (1723 control and 931 treated). Given the nature of the dataset having a small sample size, the more units, the better. Hence, the nearest neighbors larger sample size translates into a more precise estimate of the standard error than the coarsened exact matching approach (0.0163077 vs. 0.018031). I now go onto estimate the ATE with the bootstrapped standard error and determine a 95% confidence interval (and see how the nearest neighbors SE is closest to the bootstrap SE of the three matching approaches).

Compute Point Average Treatment Effect

```
## Fit model among TREATMENT = TRUE to get E[Y_i(1) | X]
treatment_model <- lm_robust(college ~ dist + score + fcollege +
                             tuition + wage + urban,
                             data=subset(data, income==TRUE))

## Fit model among TREATMENT = FALSE to get E[Y_i(0) | X]
control_model <- lm_robust(college ~ dist + score + fcollege +
                           tuition + wage + urban,
                           data=subset(data, income==FALSE))

## Predict the potential outcome under treatment for all units
```

```
data$treated <- predict(treatment_model, newdata=data)

## Predict the potential outcome under control for all units
data$control <- predict(control_model, newdata=data)

point_ate <- mean(data$treated - data$control)
point_ate
```

```
## [1] 0.09193
```

Perform Bootstrapping

```
### Bootstrap for SEs
set.seed(10003)
nBoot <- 2000 # Number of iterations
boot_results <- rep(NA, 2000)
for (iter in 1:nBoot){
  # Resample w/ replacement
  data_boot <- data[sample(1:nrow(data), nrow(data), replace=T),]

  ## Fit model among TREATMENT = TRUE to get E[Y_i(1) | X]
  treatment_model_boot <- lm_robust(college ~ dist + score + fcollege +
                                    tuition + wage + urban,
                                    data=subset(data, income==TRUE))


  ## Fit model among TREATMENT = FALSE to get E[Y_i(0) | X]
  control_model_boot <- lm_robust(college ~ dist + score + fcollege +
                                  tuition + wage + urban,
                                  data=subset(data, income==FALSE))


  ## Predict the potential outcome under treatment for all units
  data_boot$treated_boot <- predict(treatment_model_boot, newdata=data_boot)

  ## Predict the potential outcome under control for all units
  data_boot$control_boot <- predict(control_model_boot, newdata=data_boot)

  ## Store bootstrapped estimate
  boot_results[iter] <- mean(data_boot$treated_boot - data_boot$control_boot)
}
```

```
# Standard error
standard_error <- sd(boot_results)
standard_error
```

```
## [1] 0.0006579
```

```
# 95% confidence interval
confidence_interval <- c(point_ate - 1.96*sd(boot_results),
```

```
                          point_ate + 1.96*sd(boot_results))
confidence_interval
```

```
## [1] 0.09064 0.09322
```

Using the regression approach, I estimate the ATE of earning an annual familial income above 25,000 USD upon college enrollment at 0.09193 with a bootstrap standard error of 0.0006579. The confidence interval is (0.09064, 0.09322). Further, because the 95% confidence interval does not capture 0, we reject the null hypothesis of no ATE at $\alpha = .05$.

Before, the standard error under the nearest neighbors matching analysis was 0.0163077, which is roughly 25 times larger than the bootstrap standard error of 0.0006579. This is why we see a much larger standard error under the matching analysis approach. That being said, the ATE of the nearest neighbors was 0.0953741, while it's near equivalent at 0.09193 when bootstrapping, and this makes sense, as ATE isn't directly affected by sample size, unlike standard error.

The importance here, however, is that the bootstrap standard error is much smaller than the robust standard errors from the matching analyses. This is good, and expected! Standard error is affected by both sample size and variance. It has an indirect relationship with sample size (as sample size increases, standard error decreases), and direct relationship with variance (as variance increases, standard error increases). However, because I bootstrapped here when calculating both ATE, I am sampling multiple times without replacement, and can assume that I have large enough sample sizes to render a difference irrelevant. This is why the standard error is much smaller, and more reliable than the robust standard errors, given the nature of the dataset (specifically its small sample size).

Further, this results in a tighter confidence interval, which in turn makes it more apparent of the presence of an ATE. Hence, through many ups and downs, I was able to estimate a causal effect between annual familial income and college enrollment. Going forward, our estimates can improve through more modern approaches to regression imputation and make use of different modeling strategies for capturing non-linearities such as cross-validation, generalized additive models, and kernel regression.