Econometric Workflow

Ji Hun Lee

April 15, 2020

Introduction

The goal of this post is to guide workflows for causality studies that are predominantly used in econometrics. I use the workflows in this paper to adapt to other problems and data.

For instance, the first example analyzed in this paper is to establish causal effect of education on wages. Note that this is fundamentally a different question from “Can we use education to forecast wage?” (predictive) or “How does education vary with wage?” (descriptive). For prediction, we can use machine learning algorithms specialized for this purpose and use visualization tools to guide descriptive analysis. For causality however, we will use econometric methods pioneered by economists.

Identification Strategy

Any econometric method must involve a clear strategy to identify the causal effect of interest. An identification strategy of causal effect is a clearly specified source of identifying variation in a causal variable, combined with particular econometric technique to exploit this information. Assessing empirical data, causal inference is the best conducted with randomized control trial because it generates counterfactual data. RCT’s ensure that outcomes in a control group capture the counterfactual in a treatment group. Unfortunately, most often, RCT are rare in practice, and we only have observational data. This causes identification problems because many different theoretical models and causal interpretations can be consistent with the same data.

To get around this problem, econometrics applies various rigorous methods to data and rigorously tests the assumptions of model to make a statement about causality. More specifically, econometrics attempts to address challenges of observational data such as confounding effects (omitted variables), simultaneity problem, selection bias, etc. In order to analyze data such as observational studies, we employ the following identification strategies:

1. Ordinary Least Squares: controls for observable differences between comparison groups
2. Randomized Control Trials
3. Instrumental Variable: use exogenous source of variation
4. Regression Discontinuity: use exogenous source of variation
5. Difference in Differences: uses pre-post comparisons of control and treatment groups to control for unobservable differences
6. Panel Data: use pre-post comparisons on the same unit of observvation to control for fixed unobservable differences
    - Pooled OLS, First Difference, Fixed Effect, Random Effect

Libraries

Import Libraries prior to analysis:

```
library(did) # difference in differences
library(rdd) # regression discontinuity design
library(AER) # econometric methods and datasets
library(plm) # panel data models
library(caret) # for splitting data, model evaluation metrics
library(haven) # importing STATA files
library(jtools) # using summ()
library(GGally) # nice EDA tool
library(lmtest) # linear model assummption tests
library(tseries) # autocorrelation tests
library(graphics) # visualization
library(het.test) # heteroskedasticity test
library(reshape2) # data wrangling
library(stargazer) # model summaries in nice format
library(tidyverse) # data manipulation
library(tidymodels) # data preprocessing
library(robustbase) # heteroskedasticity robust linear models
```

Ordinary Least Squares

The first methodology is ordinary least squares, or linear regression. Our data is collected via only observing the outcome - wage - in the actual choice scenario of enrolling in higher education. OLS makes inferences about unknown population slope coefficients - in this case the effect of education on wage. For causal inference, we need to control for observable differences between comparison groups. Ordinary Least Squares can give us a causal interpretation if its identification assumptions are met.

Data Inspection and Variables

The first dataset is the NLS sample of young men containing information on hourly wage in cents (wage) and years of schooling (educ) in 1976. Since we want to estimate the causal effect of education on wages, our response variable is wage and our main variable of interest is educ.

The other datasets are also examined. They are: - college performance data to investigate the relationship between academic performance in college, gender, and cognitive ability - housing data to determine the causal effect of house features on the price

```
# effect of education on wages
df <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/CARD.DTA')
glimpse(df)
```

```
## Rows: 3,010
## Columns: 34
## $ id       <dbl> 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19...
## $ nearc2   <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ nearc4   <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ educ     <dbl> 7, 12, 12, 11, 12, 12, 18, 14, 12, 12, 9, 12, 11, 11, 16, ...
## $ age      <dbl> 29, 27, 34, 27, 34, 26, 33, 29, 28, 29, 28, 26, 24, 30, 31...
## $ fatheduc <dbl> NA, 8, 14, 11, 8, 9, 14, 14, 12, 12, 11, 11, 11, 11, NA, 1...
## $ motheduc <dbl> NA, 8, 12, 12, 7, 12, 14, 14, 12, 12, 12, 6, 6, 6, 8, 12, ...
## $ weight   <dbl> 158413, 380166, 367470, 380166, 367470, 380166, 367470, 49...
## $ momdad14 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1...
## $ sinmom14 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0...
## $ step14   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg661   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg662   <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ reg663   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg664   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg665   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg666   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg667   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg668   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ reg669   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ south66  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ black    <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0...
## $ smsa     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ south    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ smsa66   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ wage     <dbl> 548, 481, 721, 250, 729, 500, 565, 608, 425, 515, 225, 400...
## $ enroll   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ KWW      <dbl> 15, 35, 42, 25, 34, 38, 41, 46, 32, 34, 29, 34, 22, 27, 43...
## $ IQ       <dbl> NA, 93, 103, 88, 108, 85, 119, 108, 96, 97, 84, 89, 93, 74...
## $ married  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 6, 1, 1, 6, 1...
## $ libcrd14 <dbl> 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0...
## $ exper    <dbl> 16, 9, 16, 10, 16, 8, 9, 9, 10, 11, 13, 8, 7, 13, 9, 4, 16...
## $ lwage    <dbl> 6.306275, 6.175867, 6.580639, 5.521461, 6.591674, 6.214608...
## $ expersq  <dbl> 256, 81, 256, 100, 256, 64, 81, 81, 100, 121, 169, 64, 49,...
```

```
# effect of gender and cognitive ability on college performance
df2 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/gpa2.dta')
glimpse(df2)
```

```
## Rows: 4,137
## Columns: 12
## $ sat      <dbl> 920, 1170, 810, 940, 1180, 980, 880, 980, 1240, 1230, 1140...
## $ tothrs   <dbl> 43, 18, 14, 40, 18, 114, 78, 55, 18, 17, 78, 43, 17, 64, 4...
## $ colgpa   <dbl> 2.04, 4.00, 1.78, 2.42, 2.61, 3.03, 1.84, 3.05, 3.00, 2.00...
## $ athlete  <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1...
## $ verbmath <dbl> 0.48387, 0.82813, 0.88372, 0.80769, 0.73529, 0.81481, 0.76...
## $ hsize    <dbl> 0.10, 9.40, 1.19, 5.71, 2.14, 2.68, 3.11, 2.68, 3.67, 0.10...
## $ hsrank   <dbl> 4, 191, 42, 252, 86, 41, 161, 101, 161, 3, 95, 13, 31, 51,...
## $ hsperc   <dbl> 40.000000, 20.319149, 35.294117, 44.133099, 40.186916, 15....
## $ female   <dbl> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0...
## $ white    <dbl> 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ black    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ hsizesq  <dbl> 0.0100, 88.3600, 1.4161, 32.6041, 4.5796, 7.1824, 9.6721, ...
```

```
# effect of various house features on price
df3 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/hprice1.dta')
glimpse(df3)
```

```
## Rows: 88
## Columns: 10
## $ price    <dbl> 300.000, 370.000, 191.000, 195.000, 373.000, 466.275, 332....
## $ assess   <dbl> 349.1, 351.5, 217.7, 231.8, 319.1, 414.5, 367.8, 300.2, 23...
## $ bdrms    <dbl> 4, 3, 3, 3, 4, 5, 3, 3, 3, 3, 4, 5, 3, 3, 3, 4, 4, 3, 3, 4...
## $ lotsize  <dbl> 6126, 9903, 5200, 4600, 6095, 8566, 9000, 6210, 6000, 2892...
## $ sqrft    <dbl> 2438, 2076, 1374, 1448, 2514, 2754, 2067, 1731, 1767, 1890...
## $ colonial <dbl> 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1...
## $ lprice   <dbl> 5.703783, 5.913503, 5.252274, 5.273000, 5.921578, 6.144775...
## $ lassess  <dbl> 5.855359, 5.862210, 5.383118, 5.445875, 5.765504, 6.027073...
## $ llotsize <dbl> 8.720297, 9.200593, 8.556414, 8.433811, 8.715224, 9.055556...
## $ lsqrft   <dbl> 7.798934, 7.638198, 7.225482, 7.277938, 7.829630, 7.920810...
```

Data Restriction

As discussed earlier, we need to be mindful of how data was collected because it can affect the assumptions needed for valid results. For instance, in the first dataset, IQ needs to be measured prior to education because education can increase intelligence and dilute the effect of education. One could imagine that ability is increasing more for those who are in high school or college than those that dropped out. We want to measure KWW (a variable measuring intelligence) before high school such that we avoid having a measure of ability that is affected by education.

Lurking Variable

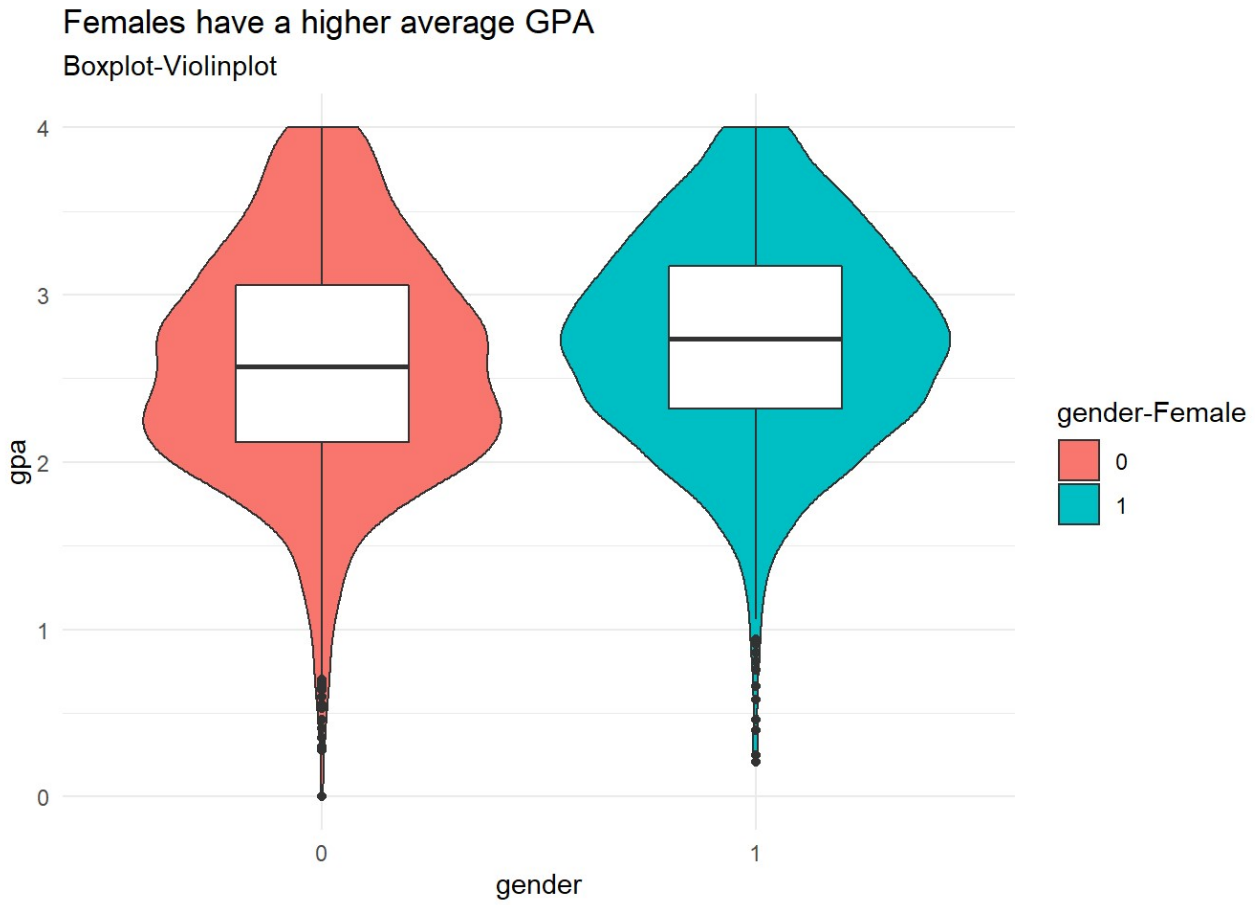
When data are based on non-random observational data, For college performance data, we suspect that there is a lurking variable such as motivation or low psychic cost of studying that affects both GPA and SAT score, and the non-random selection into college can hide the effect of this variable. We will need to control for these kinds of variables to make the zero conditional mean expectation assumption of OLS credible.

Ideal Experiment

It helps to think about what would be the ideal experiment for this kind of data. For instance, imagine that we randomize a person's gender at the time of college entrance. This would help estimate the effect of all that comes with being female during college, but not before college. However, such experiemnt would be infeasible and unethical. One can argue both ways whether estimating the causal effect of being female is meaningful. How useful such estimation is can depend on whether we are able to make choices or change the values on our gender.

We also need to be mindful of whether our variable of interest is something that can be chosen (e.g. gender is something not chosen). In the case of gender, the interpretation of coefficient is, difference between average of male response and female response. Look at the result below:

```
ggplot(df2, aes(x=factor(female), y=colgpa)) +
  geom_violin(aes(fill=factor(female))) +
  geom_boxplot(width=0.4) +
  theme_minimal() +
  labs(title='Females have a higher average GPA',
       subtitle='Boxplot-Violinplot',
       x = 'gender',
       y = 'gpa',
       fill='gender-Female')
```



Measurement Error

Can we suspect variable to be truthfully reported by subjects? What issues can arise if variable is measured with error. How does this affect OLS estimates? Underreporting of variable like drinking habit can cause inconsistent OLS estimates. Coefficient will be biased if there is a correlation between measurement error and reported drinking. Measurement error always increases standard errors. Measurement error's internal variance amplifies the uncertainty of estimate. Additionally, measurement error like this can cause attenuation bias, causing estimate to shrink toward zero.

Controls

In order to reduce omitted variable bias, we need to include control variables into our regression model. We will control for intelligence of our subjects to control for systematic variation of wage with respect to subject's intelligence.

```
lmod1 <- lm(lwage~educ+KWW, data=df)
lmod1$call$formula
```

```
## lwage ~ educ + KWW
```

Overcontrolling

We need to check whether our model has any overcontrolling issues. For example, we don't need to include productivity as a variable when we already have education and IQ because our interpretation becomes very difficult otherwise. In our case, there is no overcontrolling.

Data Preprocessing

We need to choose data preprocessing method before running our econometric model. In our example, we need to decide whether a categorical or numeric variable is better suitable for the model. Depending on how we encode our education variable as either numeric or categorical, an interpretation becomes very different. For instance, is education better in terms of years in education or degree earned? We try both methods here and have the following encoding scheme: college degree will be educ >= 16, HS will be >= 12 and <16. The question will ask whether the wage only responds to the degree i.e. whether wage jumps at the moment an individual obtains a degree and do not change everywhere else, or the wage increases linearly with years of education.

The second preprocessing strategy to consider is log transformation. A log Transformation can be used to normalize the response variable, and changes wage to percentages, not levels. The choice of log-transformation depends on whether we will define the response in terms of percentage or original unit. There are three benefits to log transformation: - benefit 1: often fit CLM assumption better -> normalize error and fix heteroskedasticity - benefit 2: can be a better, more plausible functional form in linear regression - benefit 3: easier to interpret - approximate percentage changes - WARNING: if you log,log(10,000) - log(1) = 9.2 whereas log(1,500,000) - log(80,000) = 0.63. This implies variation in the data will be due to whether someone is - for example in the job market, rather than effect on wages.

Our model will log-transform the wage variable.

```
df <-
  recipe(lwage ~ educ + KWW + wage + age + exper, data=df) %>%
    step_mutate(COL = ifelse(educ >= 16, 1, 0), # college degree is people who have mo
re than 16 years of educatioanl years
         HS = ifelse(educ >= 12 & educ < 16, 1, 0)) %>%  # high school degree is peopl
e who have at least 12 years
    prep(data=df) %>%
    juice()
glimpse(df)
```

```
## Rows: 3,010
## Columns: 8
## $ educ  <dbl> 7, 12, 12, 11, 12, 12, 18, 14, 12, 12, 9, 12, 11, 11, 16, 14,...
## $ KWW   <dbl> 15, 35, 42, 25, 34, 38, 41, 46, 32, 34, 29, 34, 22, 27, 43, 3...
## $ wage  <dbl> 548, 481, 721, 250, 729, 500, 565, 608, 425, 515, 225, 400, 4...
## $ age   <dbl> 29, 27, 34, 27, 34, 26, 33, 29, 28, 29, 28, 26, 24, 30, 31, 2...
## $ exper <dbl> 16, 9, 16, 10, 16, 8, 9, 9, 10, 11, 13, 8, 7, 13, 9, 4, 16, 9...
## $ lwage <dbl> 6.306275, 6.175867, 6.580639, 5.521461, 6.591674, 6.214608, 6...
## $ COL   <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1...
## $ HS    <dbl> 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0...
```

Interpretation of Coefficients

```
print(summ(lmod1))
```

```
## MODEL INFO:
## Observations: 2963 (47 missing obs. deleted)
## Dependent Variable: lwage
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,2960) = 360.89, p = 0.00
## R² = 0.20
## Adj. R² = 0.20
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          5.35   0.04   136.90   0.00
## educ                 0.02   0.00     6.79   0.00
## KWW                  0.02   0.00    19.42   0.00
## ------------------------------------------------
```

```
print(summ(lm(lwage ~ COL + HS + KWW, data=df)))
```

```
## MODEL INFO:
## Observations: 2963 (47 missing obs. deleted)
## Dependent Variable: lwage
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,2959) = 245.55, p = 0.00
## R² = 0.20
## Adj. R² = 0.20
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          5.51   0.03   179.84   0.00
## COL                  0.20   0.03     7.64   0.00
## HS                   0.13   0.02     5.81   0.00
## KWW                  0.02   0.00    19.68   0.00
## ------------------------------------------------
```

Interpretation of coefficient needs to be in the presence of control. We need to state, 'holding control's value constant/fixed, every unit of Y is increased/decreased by every unit change in X on average'. If Y is logged, then change in X by $1Y$ changes by $100e^{\beta X}$ percentage; if coefficiet is 0.029 then Y changes by 2.9%

As shown in the table above, the coefficient for educ is 0.02. As shown in the second table above, the coefficient for HS and COL are 0.13 and 0.2, respectively. We can interpret the number 0.02 as 'person would have received 2.1% higher wage on average if the person had one year longer education and had the same KWW.' The estimated coefficient measures the relationship between $lwage$ and the unique variation in $educ$ - partialling out.

Statistical Significance

Is it statistically different from zero (aka statisticall significant)? In other words, can we reject the null hypothesis that $\beta = 0$? Does it vary significantly with response? if it is not, we dont have to allow for the variable. In our example, yes, it is statistically significant.

F-test rejects the null hypothesis that the coefficients on the two types of education return are equal.

```
print(summ(lm(price ~ sqrft + bdrms + lotsize, data=df3)))
```

```
## MODEL INFO:
## Observations: 88
## Dependent Variable: price
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,84) = 57.46, p = 0.00
## R² = 0.67
## Adj. R² = 0.66
##
## Standard errors: OLS
## -------------------------------------------------
##                        Est.    S.E.   t val.      p
## ----------------- -------- ------- -------- ------
## (Intercept)         -21.77   29.48    -0.74   0.46
## sqrft                 0.12    0.01     9.28   0.00
## bdrms                13.85    9.01     1.54   0.13
## lotsize               0.00    0.00     3.22   0.00
## -------------------------------------------------
```

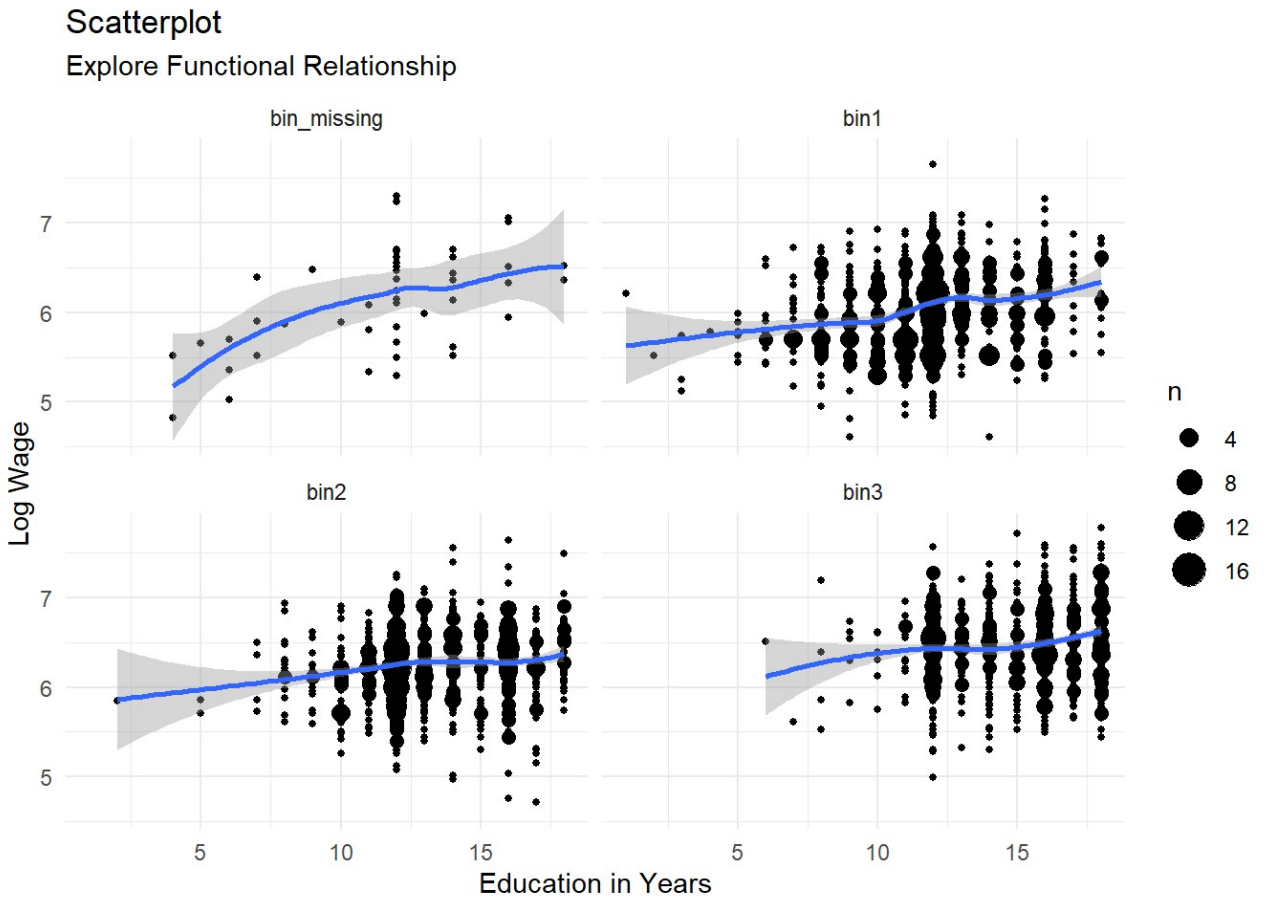
The table shows the regression result. The coefficient 0.12 on sqrft means that the price would have been $0.12 * 1000$ dollars higher if the unit were one-square-feet larger while the number of bedrooms and the size of lot stayed the sasme. We can interpret other coefficients simiarly.

Functional Form

We need to check the functional relationship between our numeric variable of interest and response. In our case, the linear relationship seems to hold.

```
df1 <-
  df %>%
  mutate(KWW_bins =  predict(
    discretize(df$KWW, na.rm=T, infs=F, cuts=3),
    df$KWW))

df1 %>%
  ggplot(aes(x=educ, y=lwage)) +
    geom_count() +
    geom_smooth(method='loess') +
    labs(title='Scatterplot',
         subtitle='Explore Functional Relationship',
         x='Education in Years',
         y='Log Wage') +
    theme_minimal() +
    facet_wrap(~KWW_bins, ncol=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Interaction Effects

We need to check whether there is any interaction effect on the variable of interest with any other control variables. In our case, we want to check for interaction between the number of years in schooling and degrees. For instance, we want to know if additional year of schooling for high school degree is statistically varying with $lwage$. We can see that having additional year of schooling after HS degree is not statistically significant on wage level, but it is for college degree. We test whether the coefficient for interaction is zero to check whether the return to education differs by years of schooling. The p-value for $educ * college$ is quite lower than the usual threshold of 0.05 and we can think of this as an evidence that the return to education differs according to the years of schooling after college degree. The return to education is about 42 units higher compared to the one with zero schooling.

```
summ(lm(wage ~ educ + KWW + educ*HS + educ*COL, data=df))
```

Observations	2963 (47 missing obs. deleted)
Dependent variable	wage
Type	OLS linear regression

F(6,2956)	113.30
R²	0.19
Adj. R²	0.19

	Est.	S.E.	t val.	p
(Intercept)	153.60	61.09	2.51	0.01
educ	1.65	6.39	0.26	0.80
KWW	10.41	0.58	17.92	0.00
HS	134.50	94.09	1.43	0.15
COL	-607.80	174.82	-3.48	0.00
educ:HS	-6.81	8.47	-0.80	0.42
educ:COL	42.21	11.68	3.61	0.00

Standard errors: OLS

As for the housing data, we can look at the variable of interest's significance by investigating its interaction effect with all other variables.

```
summ(lm(price ~ sqrft*colonial + bdrms*colonial + lotsize*colonial - colonial, data=df
3))
```

Observations	88
Dependent variable	price
Type	OLS linear regression

F(6,81)	32.79
R²	0.71
Adj. R²	0.69

	Est.	S.E.	t val.	p
(Intercept)	-22.95	29.66	-0.77	0.44
sqrft	0.09	0.02	3.79	0.00
bdrms	17.04	16.44	1.04	0.30

Standard errors: OLS

	Est.	S.E.	t val.	p
lotsize	0.01	0.00	3.82	0.00
sqrft:colonial	0.04	0.03	1.51	0.14
colonial:bdrms	-5.95	16.63	-0.36	0.72
colonial:lotsize	-0.01	0.00	-2.87	0.01

Standard errors: OLS

The coefficient for interaction of lotsize and colonial is significant, but once we remove the insignificant interaction terms it becomes insignificant (such as interactions with $sqrft$ and $bdrms$). So we can conclude that there is no evidence that the effect of housing characteristics on price differs across colonial. Again, this is expected as the style itself is not as important as other key factors affecting housing price.

Multicollinearity

We should always check for multicolinearity in the data because in the presence of collinearity, there is little unique variation in each X and it is hard to disentangle effect of one X from other correlated X. Linear regression measure partial effect of variable holding all other variables constant, but when there is a high collinearity, it makes inference more difficult because it causes both bias and increases standard error in our coefficients. There is no pronounced collinearity in our data.

```
print(ggcorr(df, method = c("everything", "pearson")))
```


```
print(ggpairs(df, progress=F))
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning: Removed 47 rows containing missing values (geom_point).
```

```
## Warning: Removed 47 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

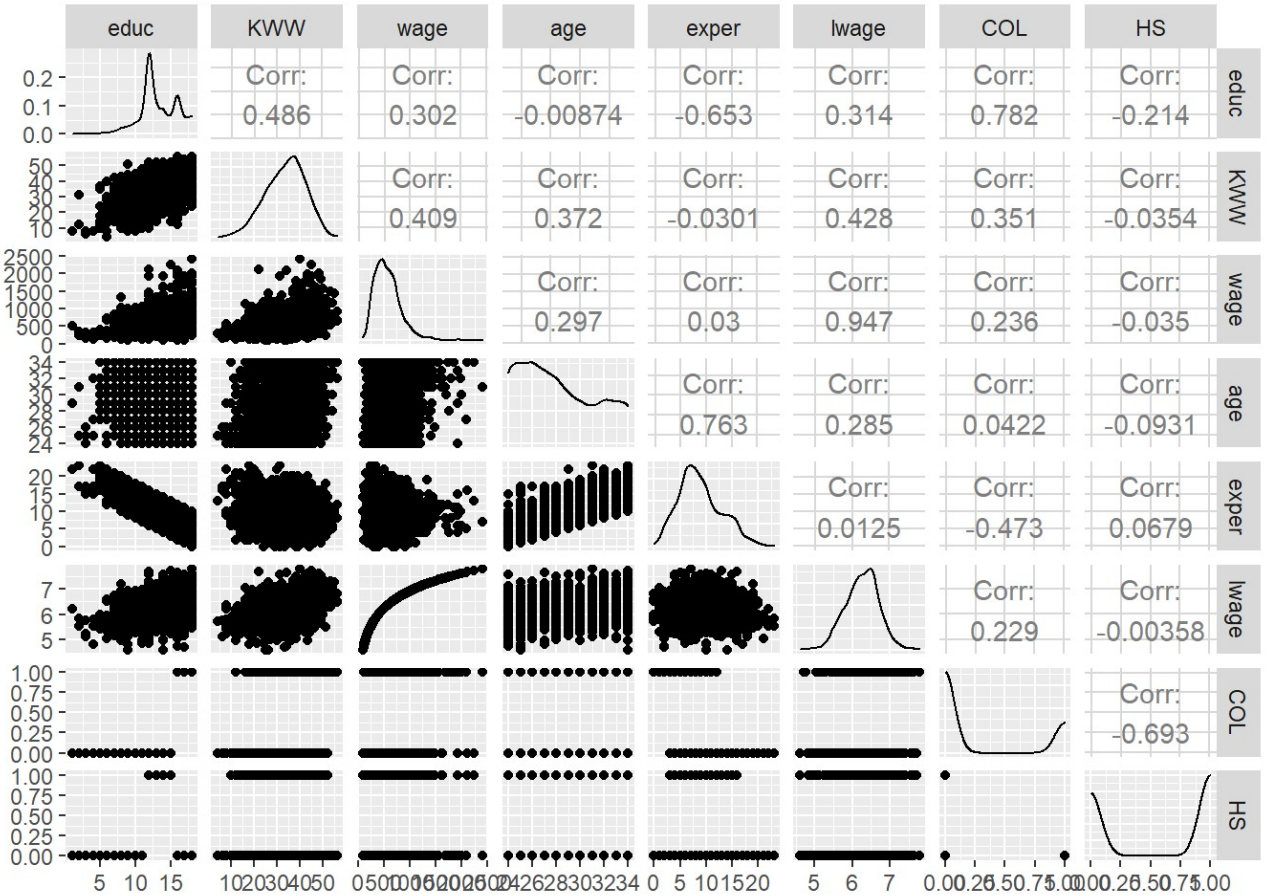
```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :
## Removed 47 rows containing missing values
```

```
## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 47 rows containing missing values (geom_point).
```



Omitted Variables

Omitted Variables create bias in coefficients. There are four situations to consider: - When correlation between X1 and X2 is positive and omitted variable has positive slope, then it inflates slope. - When correlation between X1 and X2 is negative and omitted variable has negative slope, then it inflates slope. - When correlation between X1 and X2 is positive and omitted variable has negative slope, then it deflates slope. - When correlation between X1 and X2 is negative and omitted variable has positive slope, then it deflates slope.

We can see what the sign of bias was by how the coefficient changed. If it increased, then bias was negative. If it decreased, then bias was positive. If coefficient on control is positive and bias is positive, then correlation between control and variable must be positive. If a new control is not significant, then other coefficients do not change.

Omitted variable is a problem because it splits the correlation between observed variable X_1 and omitted variable X_2 . However, when two independent variables are uncorrelated, there is no omitted variable bias.

One way to check omitted variable bias is to see how coefficients vary based on including the omitted variable.

```
lmod4 <- lm(lwage ~ HS + COL + KWW + age, data=df)
summ(lmod4)
```

Observations	2963 (47 missing obs. deleted)
Dependent variable	lwage
Type	OLS linear regression

F(4,2958)	221.75
R²	0.23
Adj. R²	0.23

	Est.	S.E.	t val.	p
(Intercept)	4.85	0.07	71.58	0.00
HS	0.18	0.02	8.30	0.00
COL	0.26	0.03	10.18	0.00
KWW	0.01	0.00	13.39	0.00
age	0.03	0.00	10.98	0.00

Standard errors: OLS

We check the sensitivity of coefficients and statistical significiance. There is a little change in the coefficient values after inclusion of age variable. Since the coefficients have increased, we can claim that the omitte variable bias was negative and deflated the true value of college's wage premium. Older people tend not to have college degree so there is a negative correlation between college degree and age. It is important to note that all the existing variables are still statistically significant after controlling for age variable.

The only way to remove omitted variable bias in the variable of interest is to add additional determinants of Y and control them.

```
summ(lm(price ~ sqrft + bdrms + lotsize + colonial, data=df3))
```

Observations	88
Dependent variable	price
Type	OLS linear regression

F(4,83)	43.25
R²	0.68
Adj. R²	0.66

	Est.	S.E.	t val.	p
(Intercept)	-24.13	29.60	-0.81	0.42
sqrft	0.12	0.01	9.31	0.00
bdrms	11.00	9.52	1.16	0.25
lotsize	0.00	0.00	3.23	0.00
colonial	13.72	14.64	0.94	0.35

Standard errors: OLS

The coefficient on colonial is not significant - colonial does not have much effect on price, if we control for other variables. the coefficients on the other variables do not change much by the inclusion of $colonial$. This is expected as the characteristics such as the house size or the number of bedrooms affect housing price mainly through the ability to accommodate a larger family and not through its correlation with the style of the house.

The omitted variable bias can be calculated by simply looking at the differences in the coefficients of the regressions with and without the control. The difference is the omitted variable bias.

OLS Assumption Validation

The following assumptions need to hold for causal interpretation to hold: 1. no strong colliearity among variables 2. zero conditional mean needs to hold (no omitted variables, self selection, simultaneity, measurement error, etc) 3. homoscedasticity: error variance is the same across all values of the independent variables 4. normality of errors

Zero Conditional Mean

Violation of this assumption causes a biased estimate.

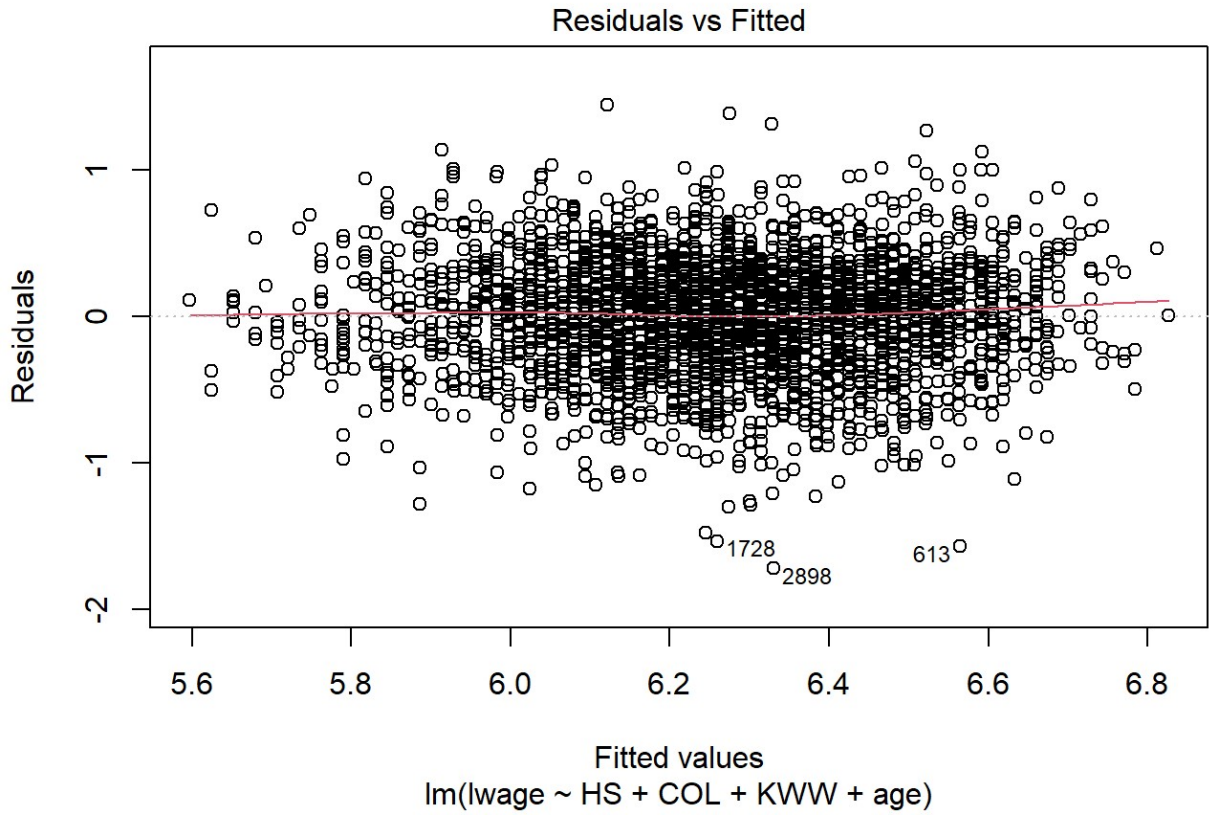
We prove coefficient is unbiased (4th assummption) by showing that for every combination of variable of interest and control (zero conditional mean assumption), mean value is around 0 by creating group_by table and see if there is any systematic variation. Also, we use fitted value versus residual plot and see if residuals are around 0 horizontal line. If there is a pattern, then there is an omitted variable bias which is creating systematic variation in response not accounted for by control or variable.

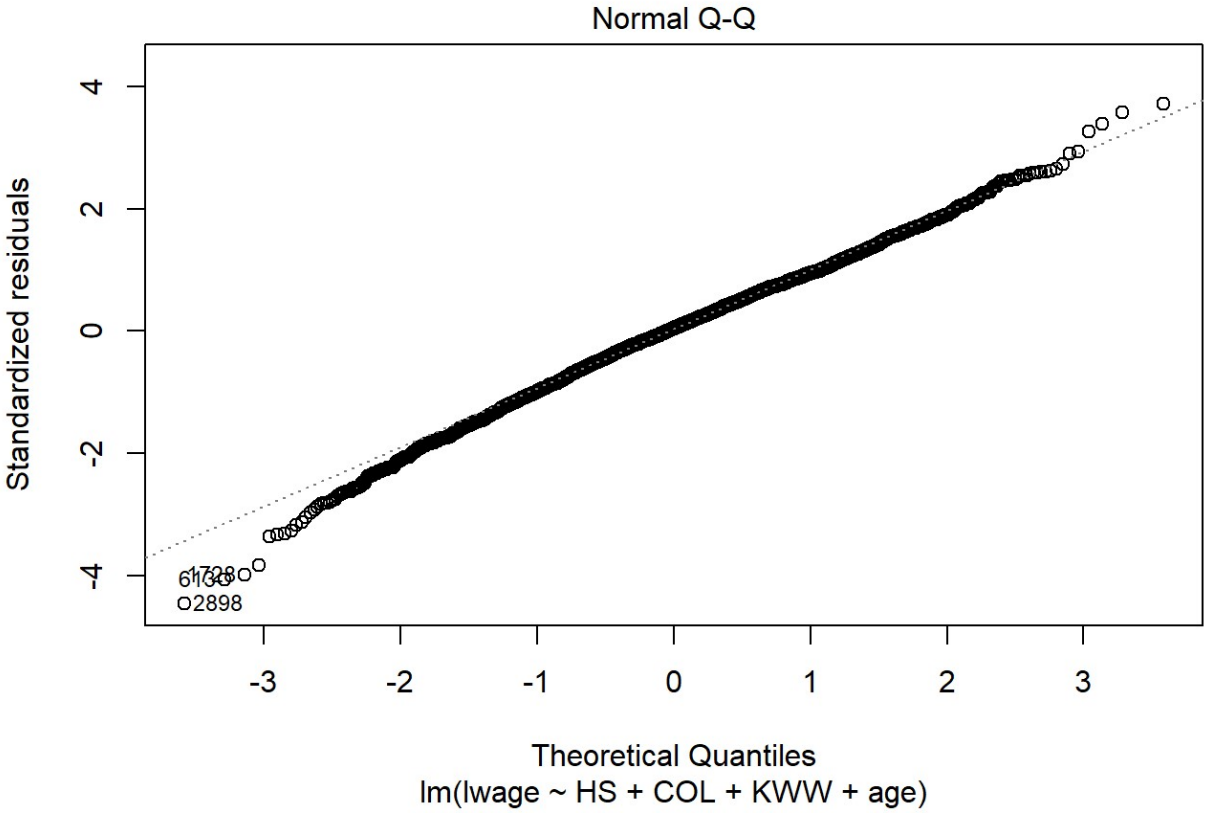
```
lmod <- lm(lwage~educ+KWW, data=df)
df %>%
  drop_na %>%
  mutate(residuals = residuals(lmod)) %>%
  group_by(educ, KWW) %>%
  summarise(conditional.mean = mean(residuals))
```

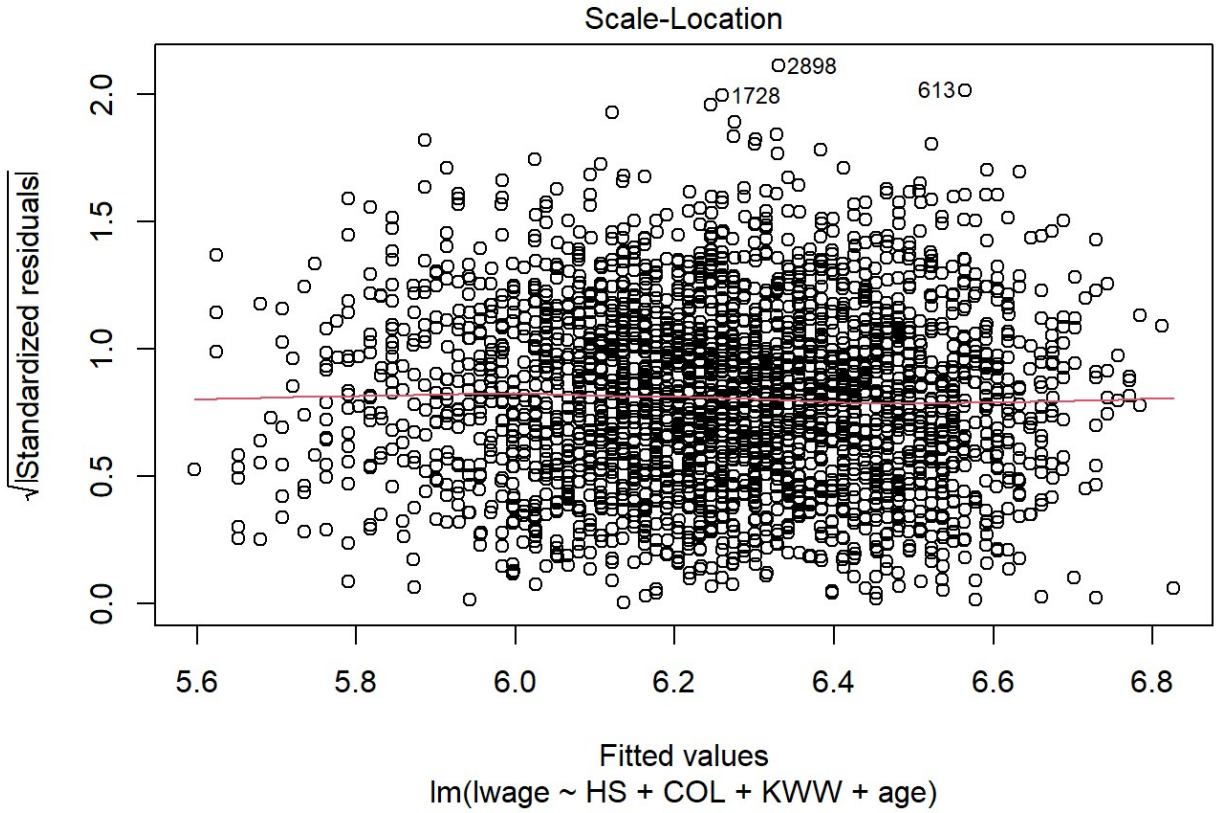
```
## # A tibble: 405 x 3
## # Groups:   educ [18]
##     educ   KWW conditional.mean
##    <dbl> <dbl>            <dbl>
##  1     1     8           0.692
##  2     2    12          -0.0975
##  3     2    31          -0.132
##  4     3     6          -0.0956
##  5     3     8          -0.318
##  6     4     8           0.197
##  7     5    10          -0.207
##  8     5    12           0.186
##  9     5    14           0.168
## 10     5    18          -0.143
## # ... with 395 more rows
```

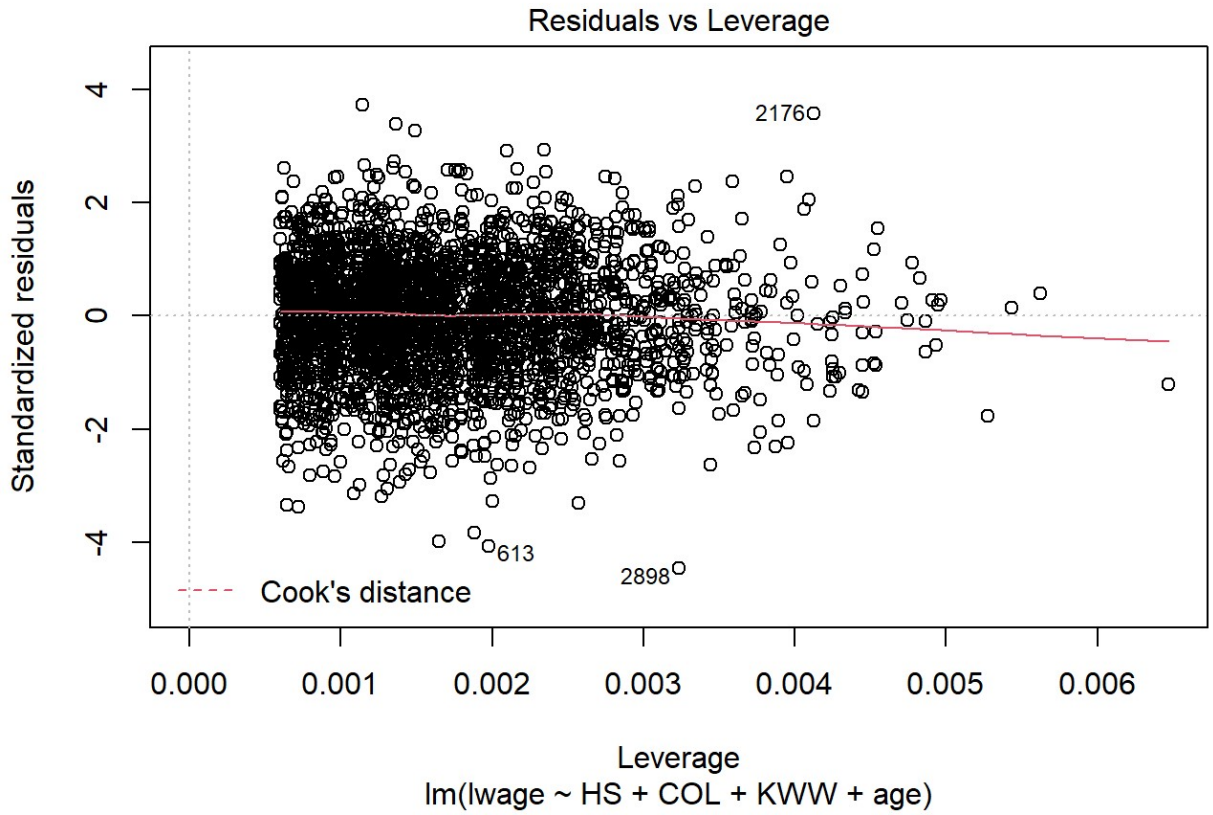
Another way is to check the diagnostic plots, in particular residual vs fitted value plot to see if the mean of residuals per fitted value is zero.

```
plot(lmod4)
```


Residuals vs Fitted
Residuals
1
0
-1
-2
1728
2898
613
5.6
5.8
6.0
6.2
6.4
6.6
6.8
Fitted values
lm(lwage ~ HS + COL + KWW + age)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
1728
1130
2898
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(lwage ~ HS + COL + KWW + age)


Scale-Location
2898
1728
613
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5.6
5.8
6.0
6.2
6.4
6.6
6.8
Fitted values
lm(lwage ~ HS + COL + KWW + age)



As for the housing data, the coefficients are likely to be biased, since there are some important factors that affect housing prices that are missing here. One such variable could be proximity to the city center. We should also consier if we could collect additional data on environmental factors to include to make the zero conditional mean assumption credible. For example, the proximity to the city center may be an important factor. We would also like to collect data on neighborhood characteristics such as crime rate of the neighborhood.

Homoscedasticity Check

Heteroskedasticity does not affect unbiasedness of OLS estimates but creates bias of variance estimation - standard errors. A standard error is a measure of how much β will vary across different samples of the same size from the same population. Heteroskedasiticty can inflate or deflate standard errors because of presence of influential data points. when you have observations far from the mean (outliers), they are more informative of the slope coefficient and has greater weight in influencing its value. However, outliers are usually subject to a lot of noise, and random chance due to noise plays a big role in the value of coefficient in any given sample. A slope estimate will vary a lot across samples due to different values of the error term for the most informative observations, which are more likely to happen in the presence of heteroskedasticity.

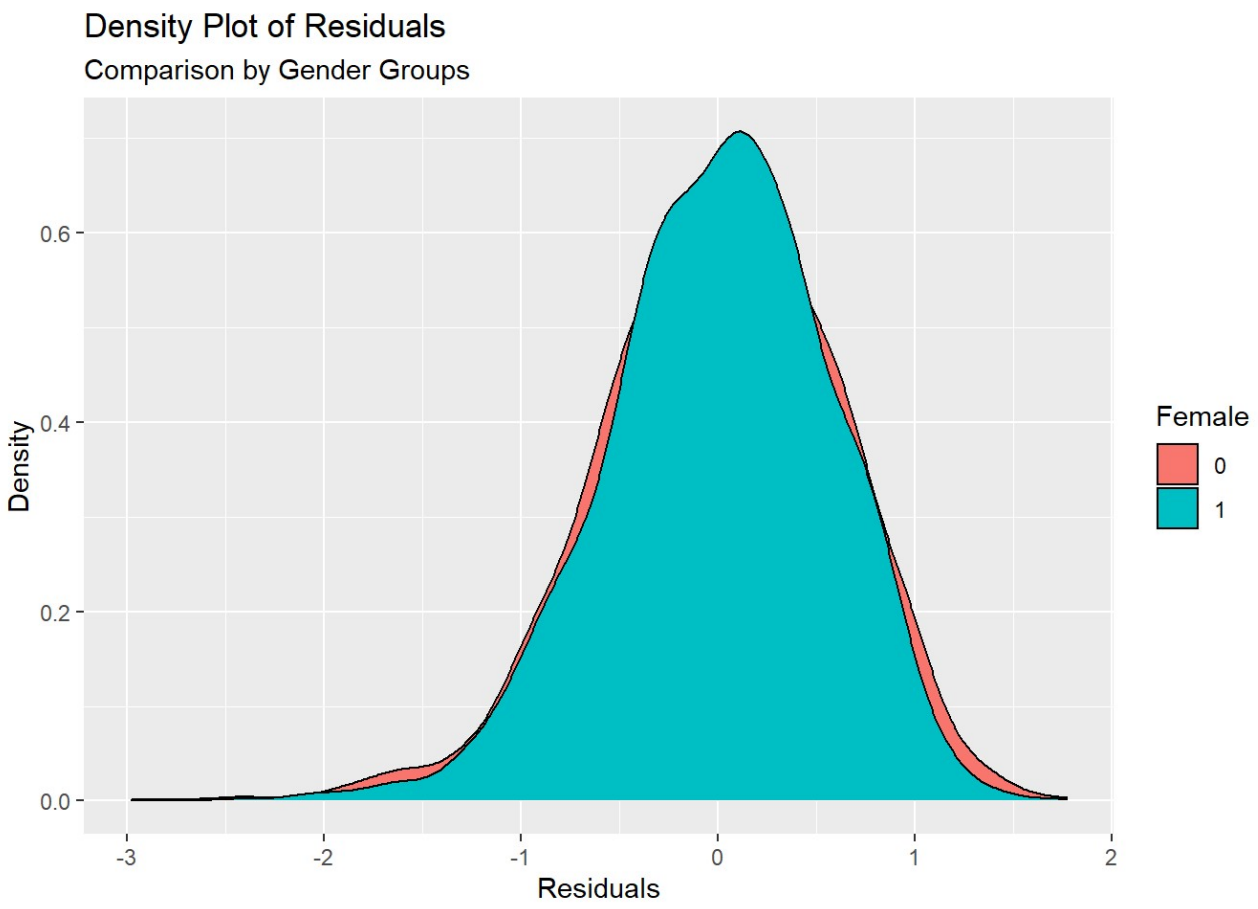
There are three ways to check the constant variance assumption: 1. Check homoscedasticity assumption by White test 2. Breuch-Pagan test 3. Look at the standard errors and residual plot (see if there are large dispersions).

Ignoring heteroskedasticity will imply we put too much faith iin beta from a given sample
i.e. underestimate the variance. Because it depends a lot onstructure of data, it is not clear how heteroskedasticity affects variance.

```
# regress college GPA on a gender indicator
lmod5 <- lm(colgpa ~ female + sat, data=df2)
print(summ(lmod5))
```

```
## MODEL INFO:
## Observations: 4137
## Dependent Variable: colgpa
## Type: OLS linear regression
##
## MODEL FIT:
## F(2,4134) = 506.10, p = 0.00
## R² = 0.20
## Adj. R² = 0.20
##
## Standard errors: OLS
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)         0.43   0.07     6.04   0.00
## female              0.23   0.02    12.35   0.00
## sat                 0.00   0.00    30.87   0.00
## ------------------------------------------------
```

```
# check the residuals density per gender
df2 %>%
  drop_na() %>%
  mutate(residuals = residuals(lmod5)) %>%
  ggplot(aes(residuals)) +
      geom_density(aes(fill=factor(female))) +
      labs(title='Density Plot of Residuals',
           subtitle='Comparison by Gender Groups',
           x='Residuals',
           y='Density',
           fill='Female')
```




```
# Breuch Pagan Test
bptest(lmod5) # run regression on squared residuals, get its R-squared and compute F-s
tatistic, null hypothesis is homoscedasticity
```

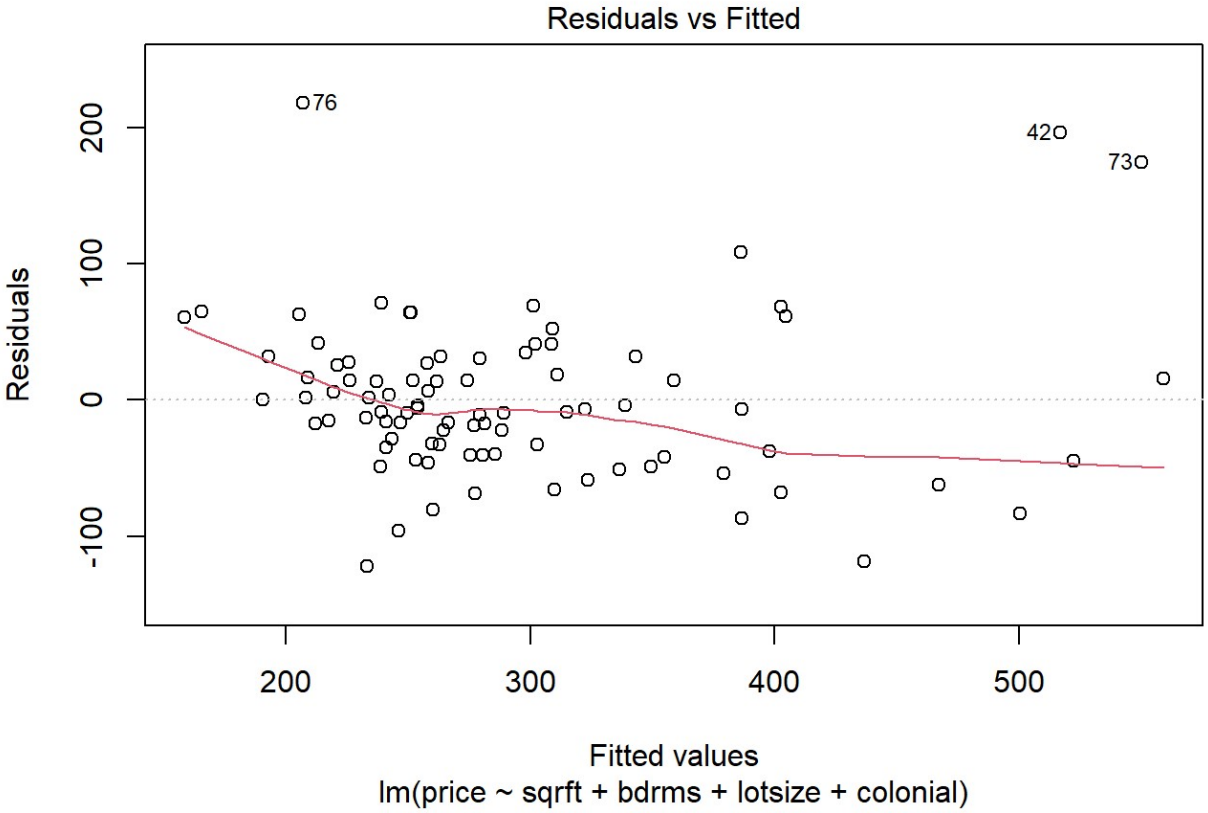
```
##
##  studentized Breusch-Pagan test
##
## data:  lmod5
## BP = 10.797, df = 2, p-value = 0.004524
```

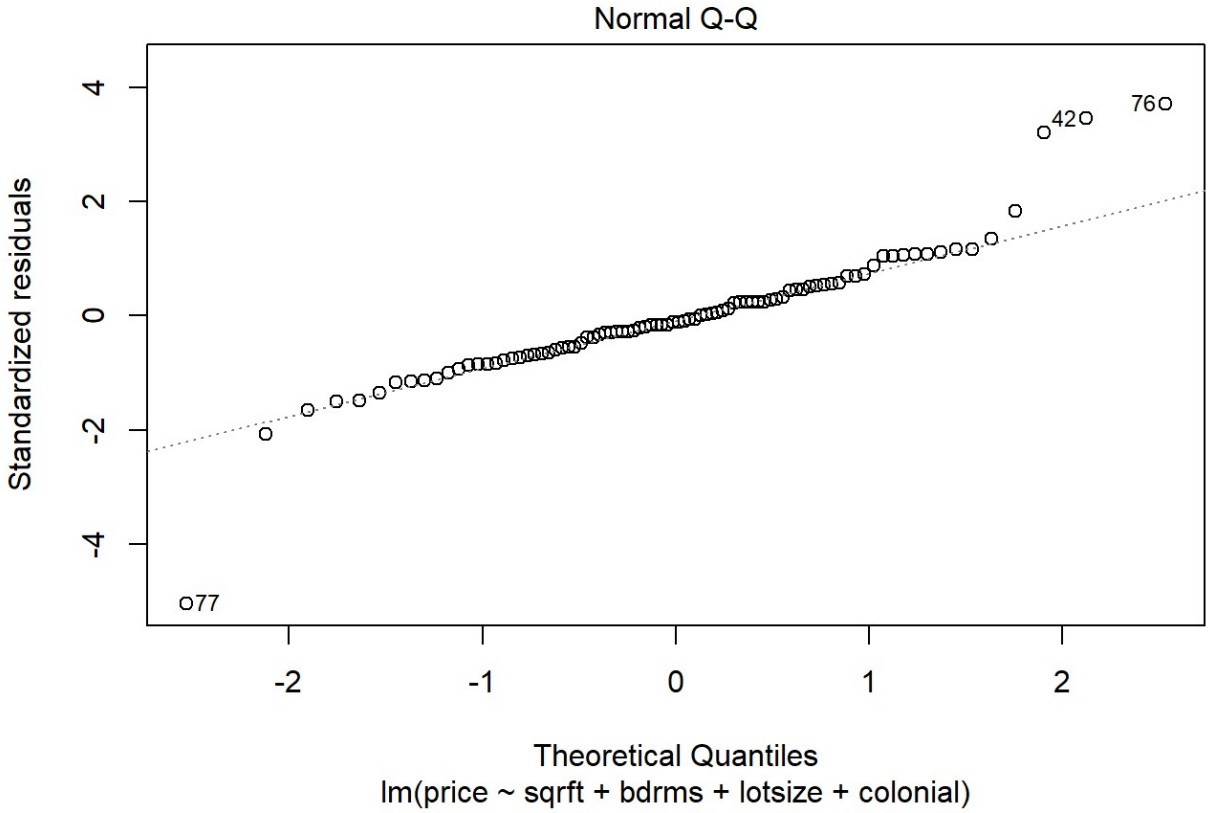
The residual distributions for men and women are similar, except that the residuals for women are slightly higher. This is because women have higher average GPA while the GPA has an upper bound of 4.0.

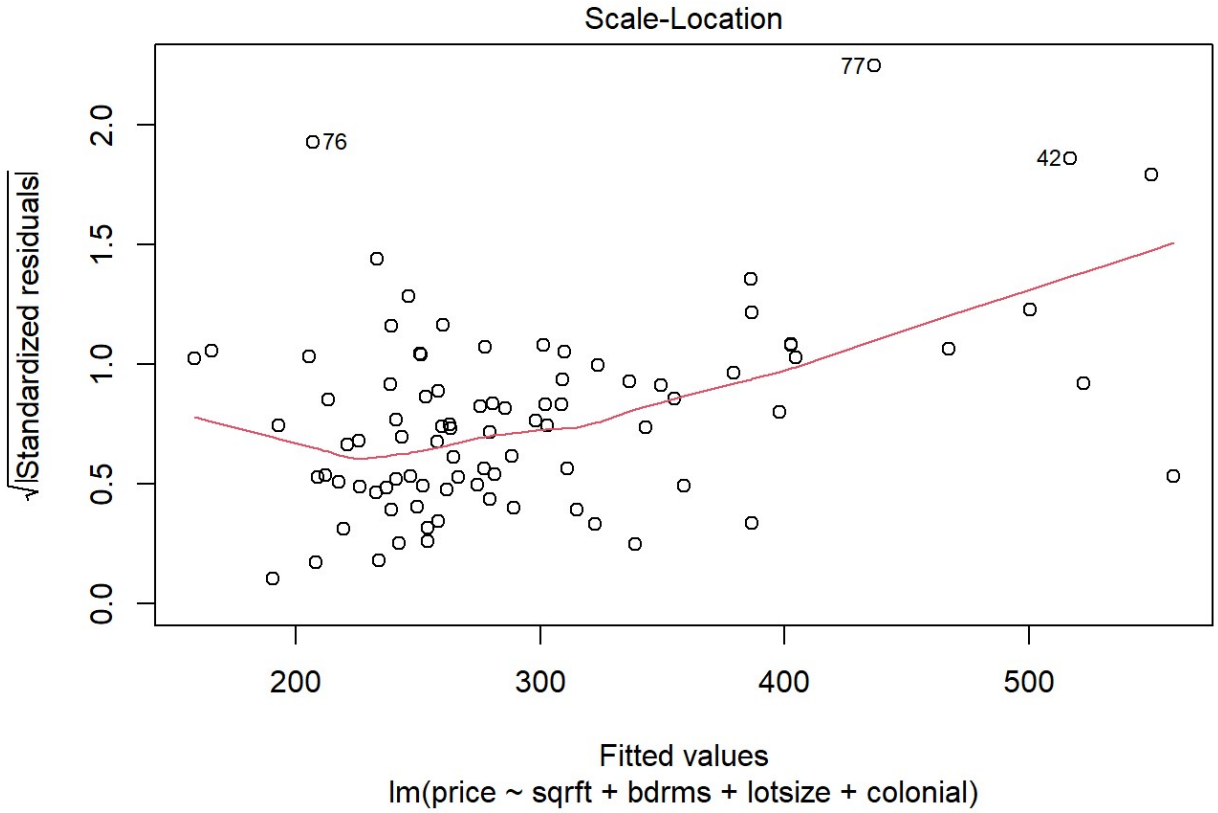
The Breusch-Pagan Heteroskedasticity test rejects the null hypothesis of homoskedasticity.

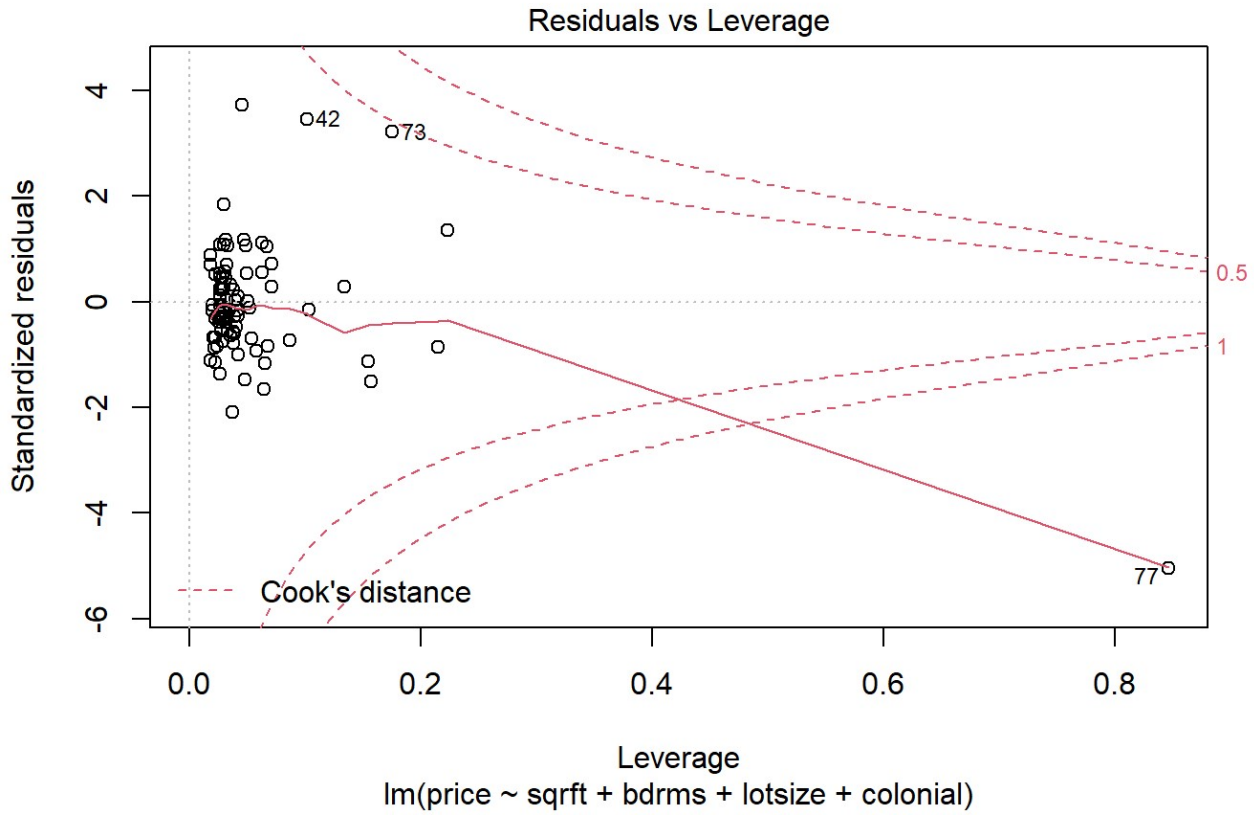
Another way to detect heteroskedasticity is to visualize the residuals on fitted values and the variable of interest.

```
plot(lm(price ~ sqrft + bdrms + lotsize + colonial, data=df3))
```

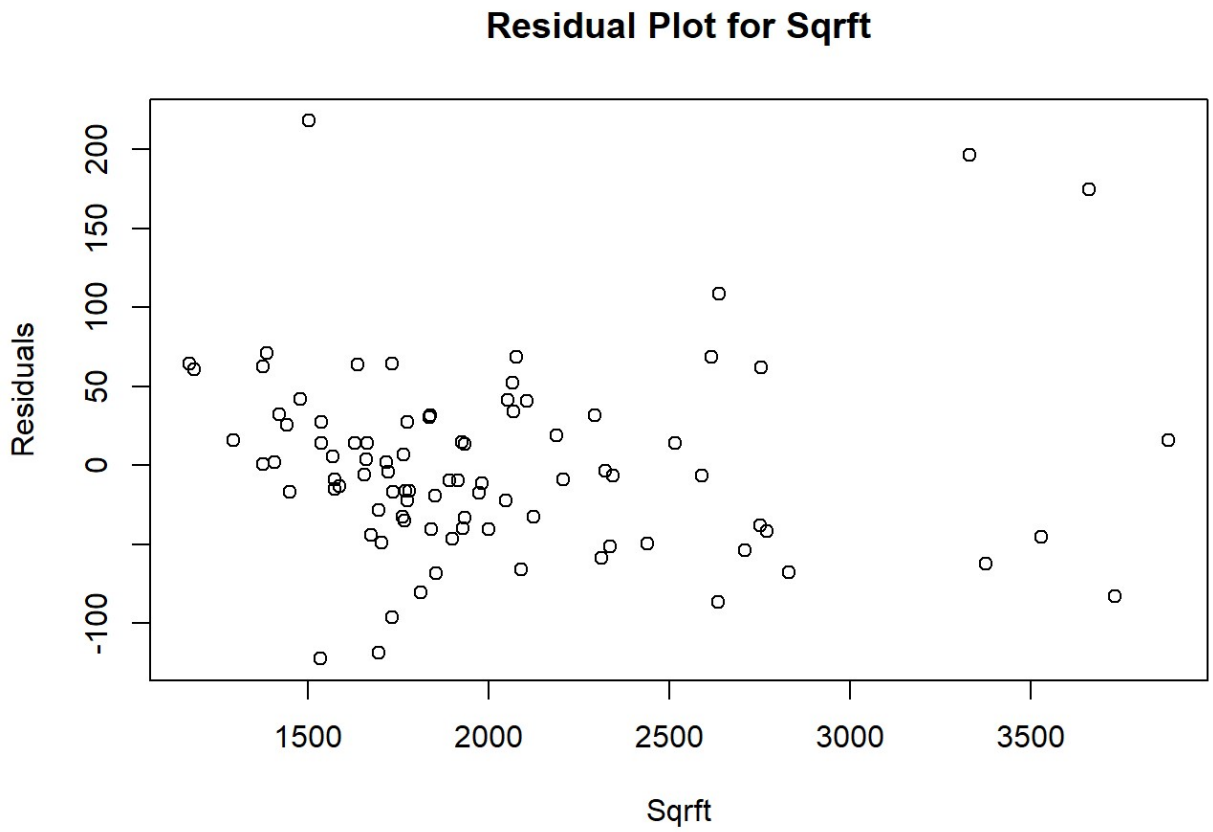

Residuals vs Fitted
76
42
73
200
100
0
-100
Residuals
200
300
400
500
Fitted values
lm(price ~ sqrft + bdrms + lotsize + colonial)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
76
42
77
-2
-1
0
1
2
Theoretical Quantiles
lm(price ~ sqrft + bdrms + lotsize + colonial)

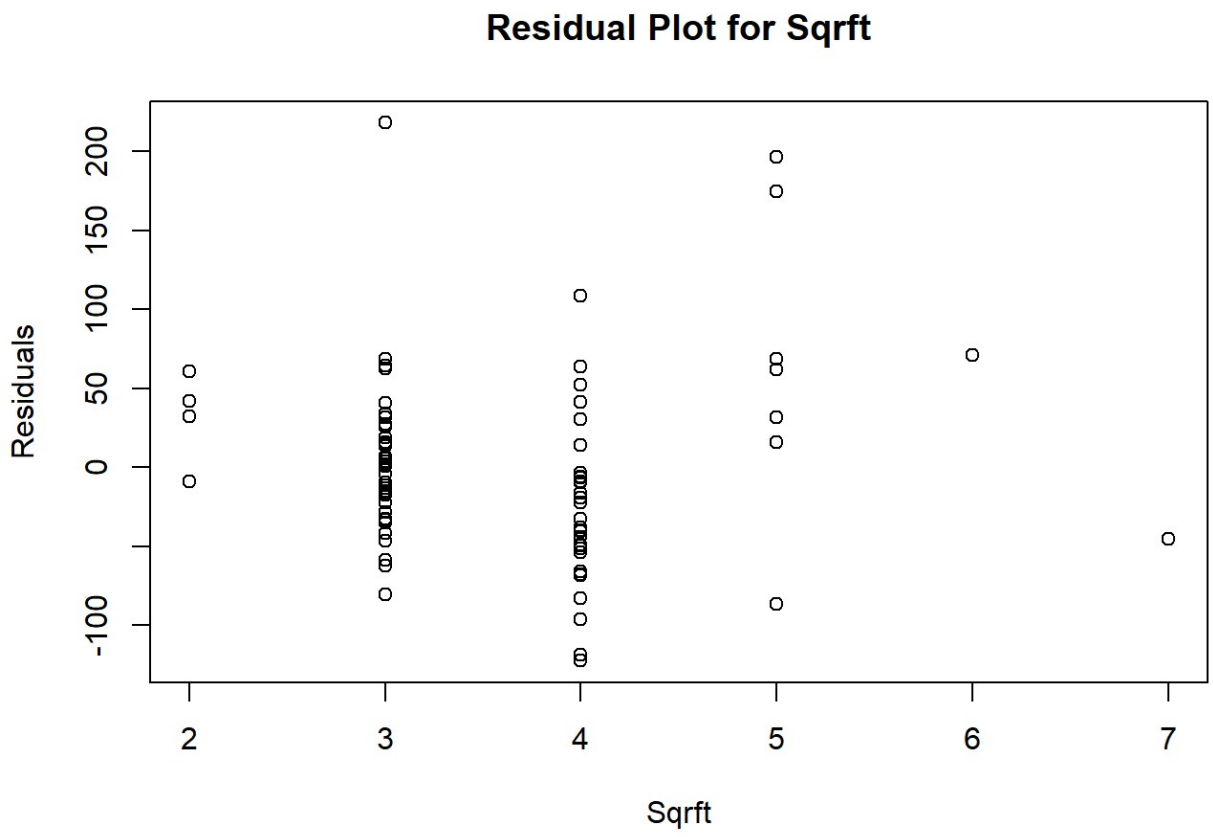

Scale-Location
77
76
42
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
200
300
400
500
Fitted values
lm(price ~ sqrft + bdrms + lotsize + colonial)




```
resids <- residuals(lm(price ~ sqrft + bdrms + lotsize + colonial, data=df3))
plot(df3$sqrft,resids,  main='Residual Plot for Sqrft', xlab='Sqrft', ylab='Residual
s')
```




```
plot(df3$bdrms, resids, main='Residual Plot for Sqrft', xlab='Sqrft', ylab='Residual
s')
```



Dealing with Heteroskedasticity

There are three ways to deal with heteroskedasticity: - transform the model (e.g. log-transform) - use robust standard errors - specify the form of error variance and use WLS

Run heteroskedasticity-robust linear regression or log transform the rresponse variable.

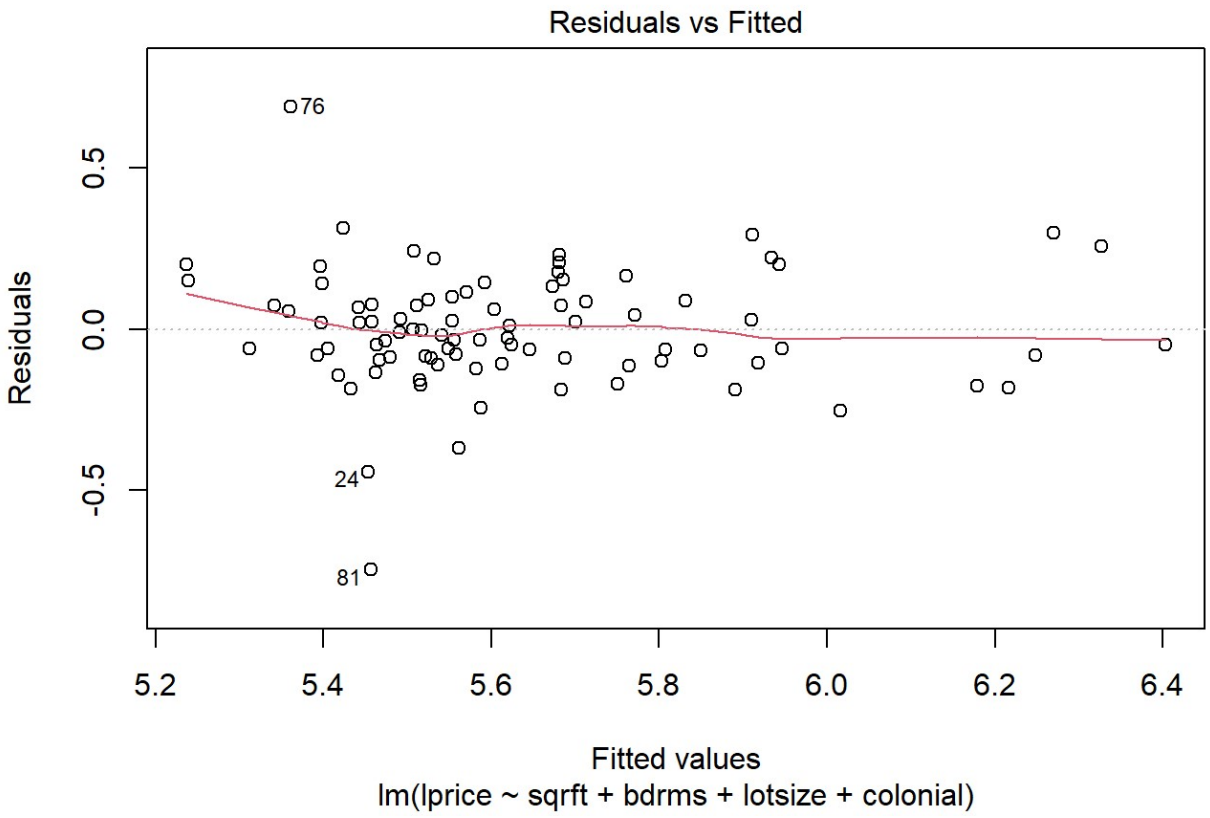
```
# generate robust stanard errors
coeftest(lmod5, vcov = vcovHC(lmod5))
```

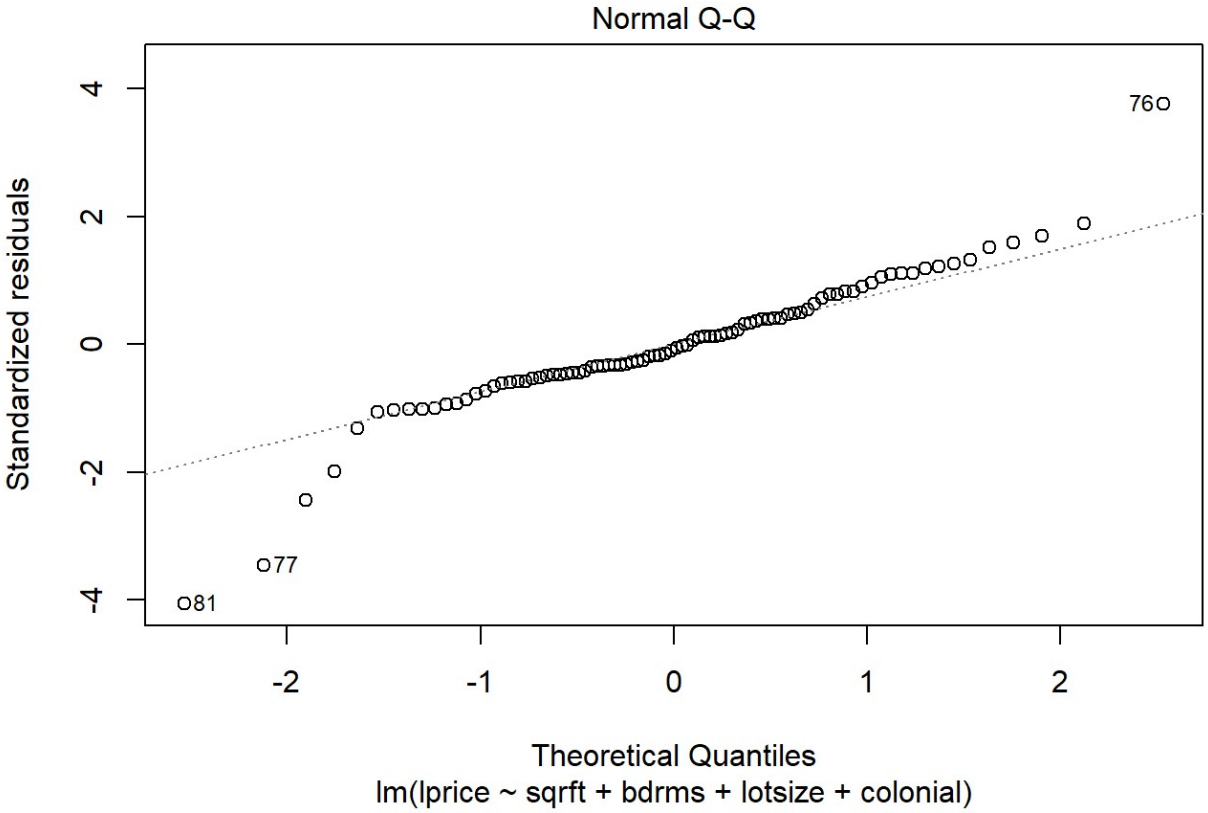
```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 4.2895e-01 6.9125e-02  6.2055 5.988e-10 ***
## female      2.3067e-01 1.8458e-02 12.4968 < 2.2e-16 ***
## sat         2.0576e-03 6.5178e-05 31.5689 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

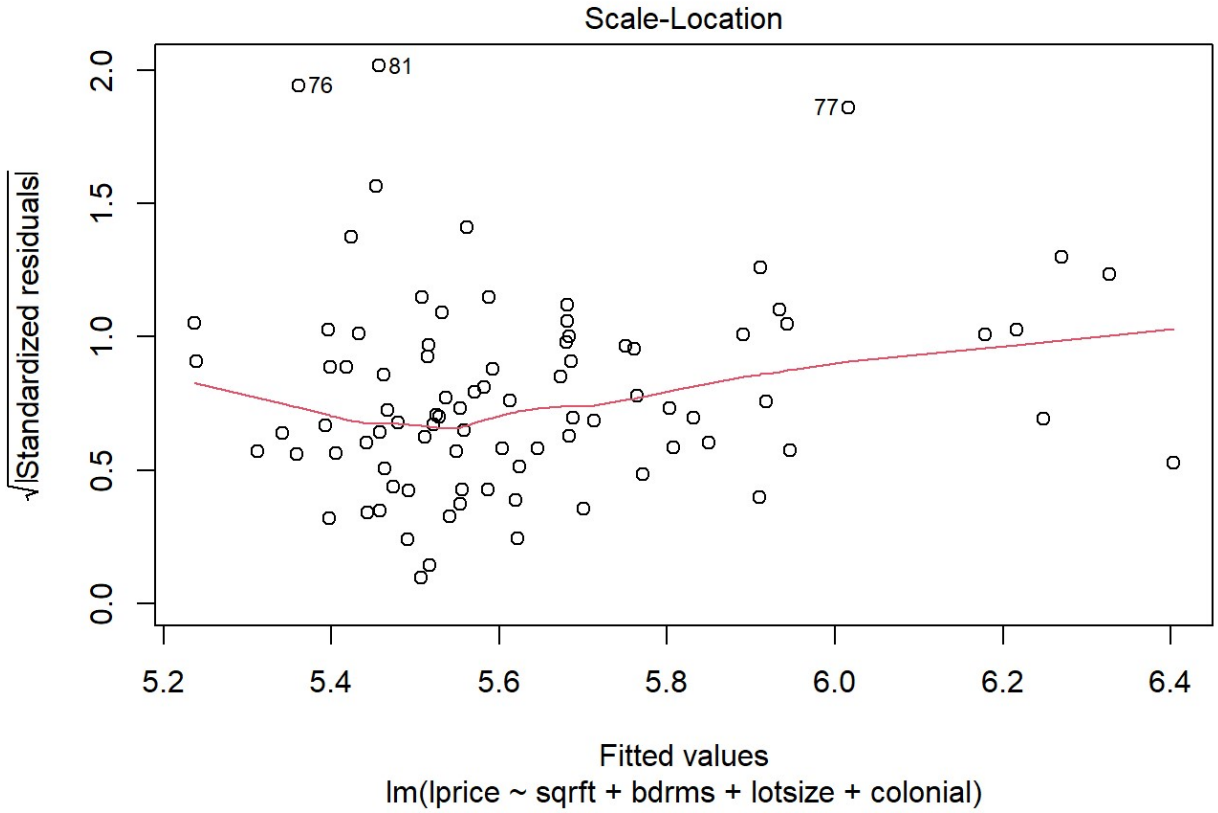
```
# robust linear regression
lmod6 <- lmrob(colgpa ~ female + sat, data=df2)
summary(lmod6)
```

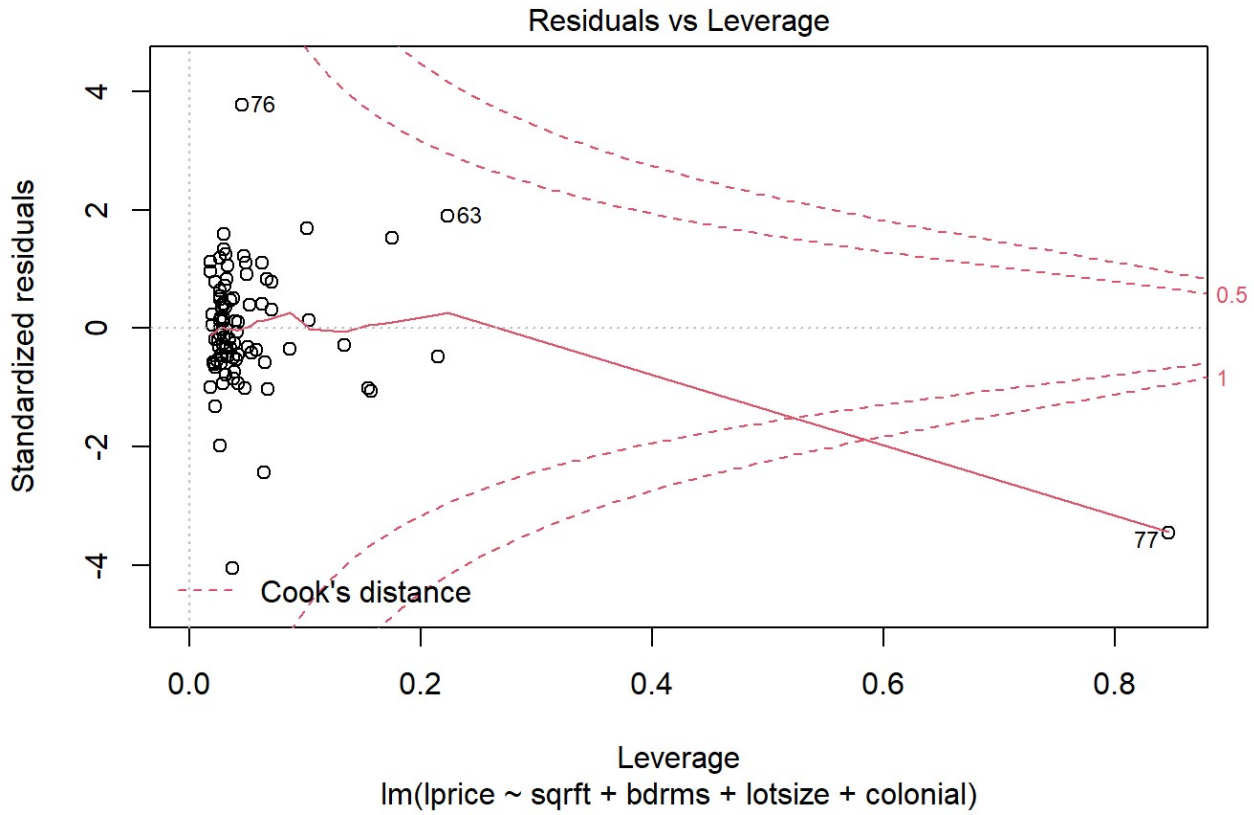
```
##
## Call:
## lmrob(formula = colgpa ~ female + sat, data = df2)
##  \--> method = "MM"
## Residuals:
##       Min        1Q    Median        3Q       Max
## -3.007892 -0.397175  0.005496  0.398515  1.766180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.094e-01  6.839e-02   5.986 2.33e-09 ***
## female      2.311e-01  1.850e-02  12.492  < 2e-16 ***
## sat         2.095e-03  6.517e-05  32.153  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Robust residual standard error: 0.5876
## Multiple R-squared:  0.2083, Adjusted R-squared:  0.2079
## Convergence in 10 IRWLS iterations
##
## Robustness weights:
##  2 observations c(3205,3882) are outliers with |weight| = 0 ( < 2.4e-05);
##  324 weights are ~= 1. The remaining 3811 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02669 0.87720 0.95090 0.90880 0.98500 0.99900
## Algorithmic parameters:
##        tuning.chi                bb        tuning.psi        refine.tol
##         1.548e+00         5.000e-01         4.685e+00         1.000e-07
##           rel.tol         scale.tol         solve.tol       eps.outlier
##         1.000e-07         1.000e-10         1.000e-07         2.417e-05
##             eps.x warn.limit.reject warn.limit.meanrw
##         2.801e-09         5.000e-01         5.000e-01
##      nResample         max.it       best.r.s       k.fast.s          k.max
##            500             50              2              1            200
##    maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
##            200              0           1000              0           2000
##                   psi           subsampling                   cov
##            "bisquare"         "nonsingular"         ".vcov.avar1"
## compute.outlier.stats
##                  "SM"
## seed : int(0)
```

```
# logprice regression on housing data
plot(lm(lprice ~ sqrft + bdrms + lotsize + colonial, data=df3))
```




Normal Q-Q
Standardized residuals
4
2
0
-2
-4
76
77
81
-2
-1
0
1
2
Theoretical Quantiles
lm(lprice ~ sqrft + bdrms + lotsize + colonial)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
81
76
77
5.2
5.4
5.6
5.8
6.0
6.2
6.4
Fitted values
lm(lprice ~ sqrft + bdrms + lotsize + colonial)



In the robust regression model, the standard errors do not change much, as we found the difference in the dispersion of the residual is not huge.

When using log transform, whether we use the transform would depend on how we believe the world works. Do we expect that an increase in the characteristics of a house would cause increase in levels of the housing prices? Or do we expect that they would cause an increase in percentages? the answer will depend on your argument

Precision of Estimates

Given that the assumptions are met, the precision of estimates is dependent on: - small collinearity - large sample size - big variation within the variable of interest

Evaluating Model

We can use the following metrics to evaluate the fitness of our model.

R-squared: fraction of sample variation in Y that is explained by all the explanatory ariables. It always increases when more variables are added to regression. Low R squared implies it is difficult to predict individual outcomes.

Adjusted R-squared: may increase or decrease with addition of another regressor.

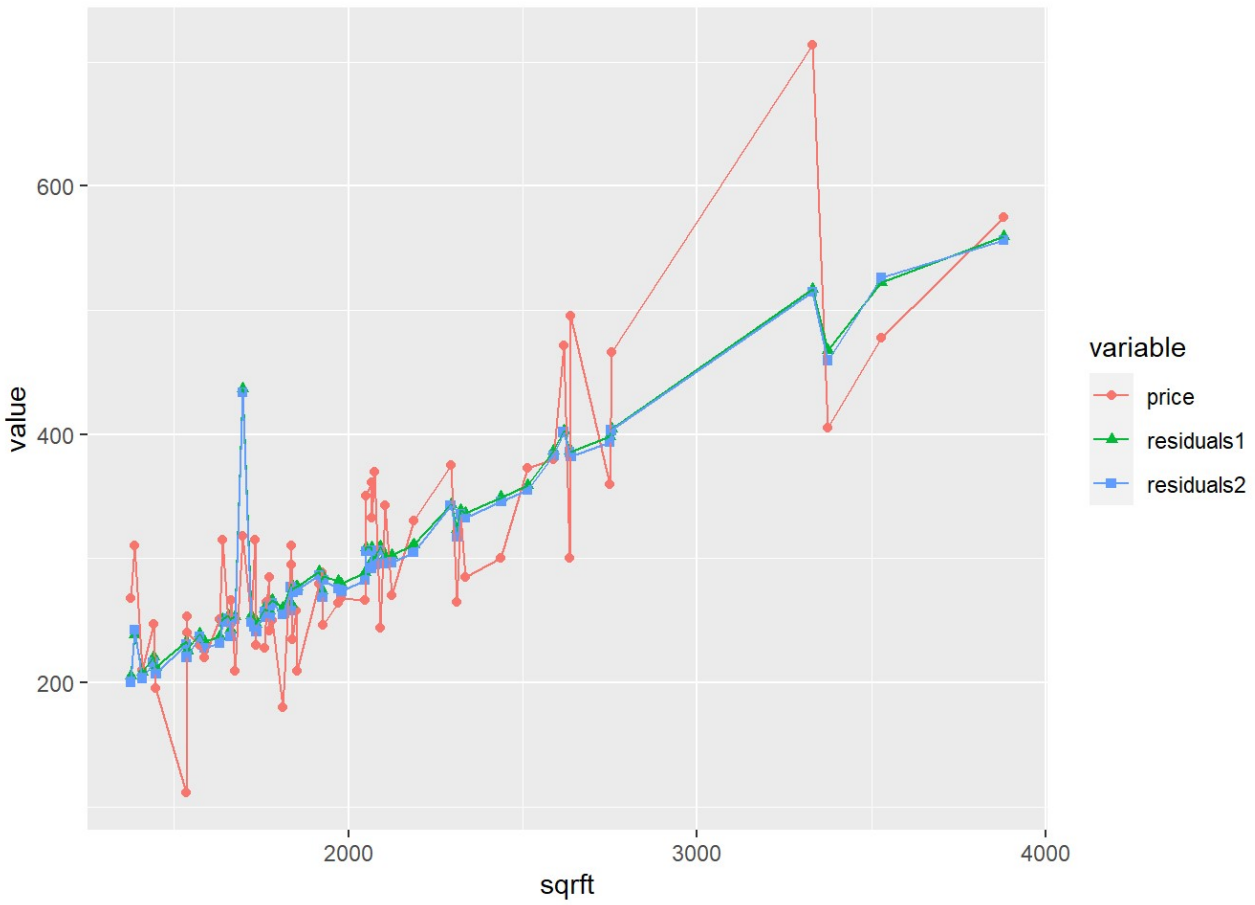
We can also use these metricsc to determine whether some functional form of variable or interaction term is better or not based on whether adjusted R squared changes.

```
postResample(pred = predict(lmod1, df), obs = df$lwage)
```

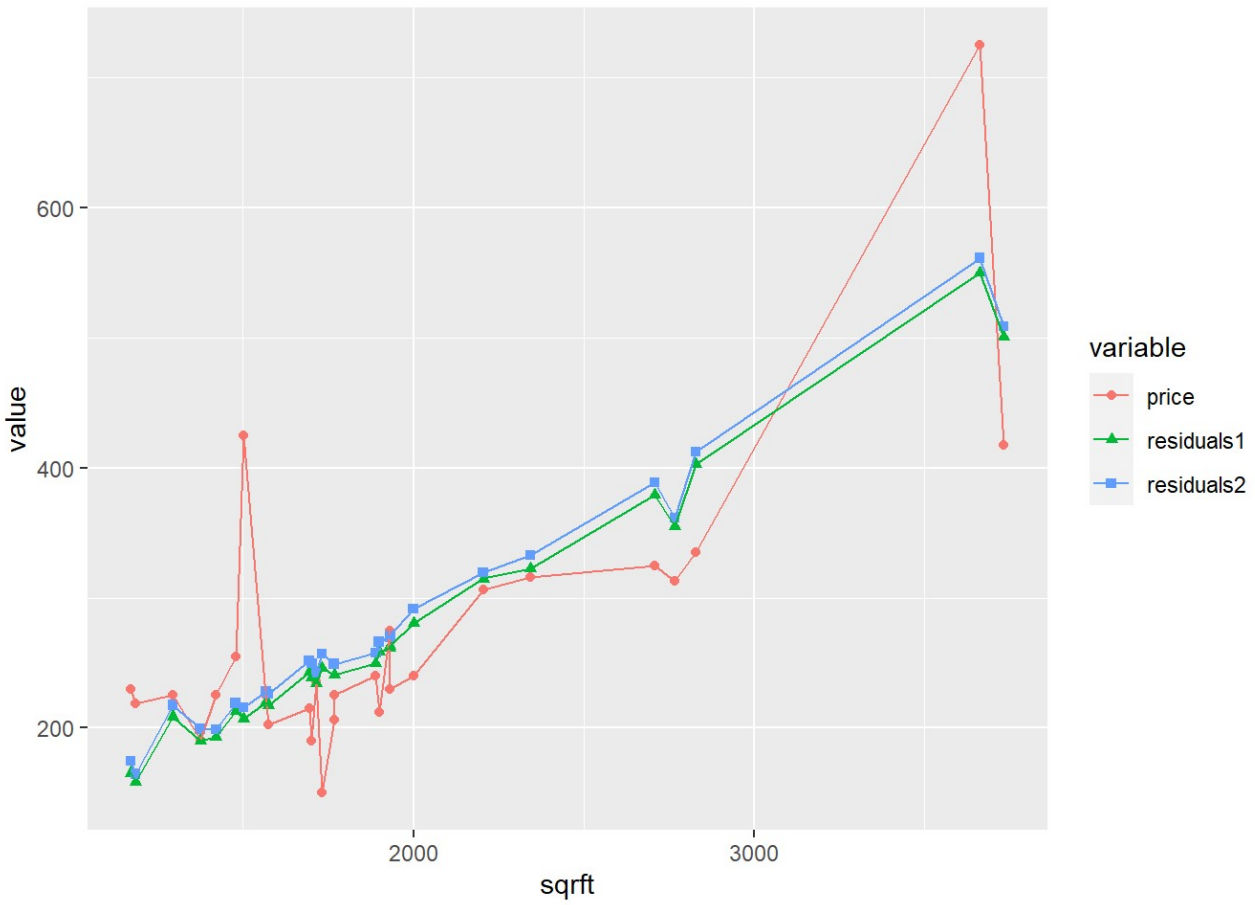
```
##      RMSE  Rsquared       MAE
## 0.3957118 0.1960425 0.3111992
```

We can evaluate model fit by the variable of interest such as colonial in housing data. To assess, which mmodel would perform better with and without control variable? We can look at the scatterplot of the observed data on the space of price and square feet, and we draw the fitted curves. If the fits are similar, we prefer a simpler model.

```
df3 %>%
 mutate(residuals1 = lm(price ~ sqrft + bdrms + lotsize + colonial, data=df3)$fitted.v
alues,
        residuals2 = lm(price ~ sqrft + bdrms + lotsize, data=df3)$fitted.values) %>%
  filter(colonial==1)  %>%
  dplyr::select(c(sqrft, price, residuals1, residuals2)) %>%
  melt(id='sqrft') %>%
  ggplot(aes(x=sqrft, y=value)) +
    geom_point(aes(col=variable, shape=variable)) +
    geom_line(aes(col=variable))
```




```
df3 %>%
 mutate(residuals1 = lm(price ~ sqrft + bdrms + lotsize + colonial, data=df3)$fitted.v
alues,
        residuals2 = lm(price ~ sqrft + bdrms + lotsize, data=df3)$fitted.values) %>%
  filter(colonial==0)  %>%
  dplyr::select(c(sqrft, price, residuals1, residuals2)) %>%
  melt(id='sqrft') %>%
  ggplot(aes(x=sqrft, y=value)) +
    geom_point(aes(col=variable, shape=variable)) +
    geom_line(aes(col=variable))
```



The fits look similarand the performance of the fit does not seem to differ across different values of colonial. In the subset of colonial=0, the model shows a little bit of overfitting. Given that the fits are similar, prefer a simpler model.

Randomized Control Trial

Randomized Control Trial is the best study design one can get with data source for stuying causality.

Some of the applications of randomized control trial in the socioeconomic field are: - Tennessee STAR: class size on student achievement - Kruger (1999) - Worms - Miguel and Kremer (2004) - Charitable giving - DellaVigna, List, and Malmendier (2010)

There are six steps to Randomized Control Trial: 1. recruit sample of n units of observation 2. randomly select the units into a treatment group and a control group 3. treatment group is given a treatment, and the control group given a placebo or nothing 4. collect data on the outcomes of interest 5. compare the average outcomes in the treatment versus control 6. if outcomes are significantly different in the treatment group, then treatment has an intended effect

The reason for randomization in experiment is that when randomized, treatment and control groups will not differ systematically with respect to other characteristics that affect outcomess. As such we can control for other variables and isolate the variable of interest's effect on outcome. However, there is still a pure chance that unit with different outcomes independent of treatment can happen to end up in the

treatment and control groups and will disguise as the effect of treatment. As the size of treatment and control groups increases, the influence of pure chance on the results will fall, so we can mitigate the effect of pure chance by increasing sample size.

Instead of being randomly assigned, if units select themselves into treatment and control groups, then we have a self-selection problem. The units - rather than the researchers - select the type of treatment. Then, while it is possible that the difference in outcome indeed reflects the causal effect of treatment, an alternative explanation that cannot be ruled out is that the units with a systematic traits will have had resulted in different outcomes anyway. This can cause systematic difference in outcomes, rather than the difference of treatment.

How can we measure the effect of treatment? One way is to take a difference in average of observed outcomes between treatment and control groups. This difference of averages can be split into two parts: average treament effect on the treated and selection bias. The average treatment effect on the treated is the desired quantity that we want to measure the effect of treatment. However, without randomization, this cannot be directly deduced because it requires knowing the counterfactual - what would have happened to the treated group if they didn't receive treatment? Similarly, the self-selection cannot be directly estimatedbecause at a given point in time an individual can only be treated or not treated. Fortunately, when experiment is well randomized, it does not have self-selection problem and difference in average is simply the average treatment on treated effect.

Threats to Validity

It is very important to understand how experiment was conducted and assess its quality. There are many threats to the validity of experiment design.

Internal Validity: if the estimated causal effects are valid for the studied population. No selection bias. Threats to internal validity exist when: 1. selection bias - treatment is in fact not randomized and dependds on characteristics of unit that also affect the outcome 2. partial compliance - not everoyne who is treated actually takes part of the treatment 3. attrition - units with certain characteristics leave the experiment 4. hawthorne effect - the experiment changes the individuals' behavior

External Validity: the extend to which causal effects of a particular program in a particular situation at a particular time can be generalized to other situations and periods. 1. representative population? 2. representative situation/environment? 3. representative program/treatment?

Other Threats to Validity: 1. treatment should not affect untreated by spillovers or changes in market prices. 2. non-compliance: treatment is offered randomly but some participants do not particiapte. when units drop out of treatment, the difference in outcomes between the treatment and the control groups now estimates the average impact of offering the treatment, usually called intention-to-treat. However, if attrition is random and affects the treatment and control groups in the same way, the estimates would remain unbiased.

Using OLS

To measure treatment effect, we can use OLS on randomly assigned data. OLS estimate is called the difference estimator. OLS is a good tool for randomized control trial because the coefficient for dummy variable like treatment is simply the difference in averages - which is precisely what we want to measure. Also, OLS estimate has the added benefit of measuring a coefficient, holding other control variables fixed and measure the partial effect of variable on outcome. When selection bias exists or experiment is not internally valid, OLS has a correlation between error term and treatment status and induces endogeneity problem.

Adding Controls

Adding control variables has the benefit that it may generate more precise estimate of causal effect of treatment:

1. If the control variable has substantial explanatory power for outcome, the standard error of treatment effect will be smaller in the long regression model.
2. It will also reduce the residual variance, which lowers the standard error of regression estimates.

Randomization Check: Covariance Balance Test

For random assignemnt to work, we need the independence assumption between treatment and other control variables. Also, we should not observe any observed or unobserved differences between treated or untreated. There are informal tests to see whether there is a problem with randomization:

1. Run a regression using data before the treatment started using the treatment status, and no control variable should have statistically significant effect on the treatment status. This is a way to test any difference in potential outcomes between those who get treated and those who don’t.
2. Add control variables, and if the treatment is random, the estimated coefficient should not change. If control variables are uncorrelated with the treatment and experiment is well randomized, then they will not affect the estimated treatment effect (MLR and SLR coefficients will be similar).

Data

Data comes from the paper ’Remedying Education: Evidence from two Randomized Experiments in India" by Banerjee, Cole, Duflo, and Linden (2007). It contains data for analyzing the effect of assigning a remedial teacher to schools at random in the Indian city of Vadodara. The narrow aim of the authors is to evaluate the immediate and lagged effect of the remedial program and their cost-effectiveness. The broad aim of the authors is to understand whether targeting school inputs to the right students leads to improvements in academic achievement, notwithstanding the fact that generally input-based policies in developing countries yield very discouraging results.

```
# data for school year 2001/2002
df1 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/Y1.dta')
df1 <-
  df1 %>%
  mutate(pre_verbnorm = scale(pre_verb),
         pre_mathnorm = scale(pre_math))
# data for school year 2002/2003
df2 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/Y2.dta')
# perfect compliance to treatment assignment
df3 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/case1.dta')
# perfect compliance to treatment assignment
df4 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/case2.dta')
# imperfect compliance to treatment assignment; now individuals may drop out from trea
tment
df5 <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/case3.dta')
```

Variables in Y1 and Y2 are normalized relative to the control group of the same year. The authors commented that they stratified on all the control variables we have in the data,so the control and treatment groups should be well balanced. We regress the treatment variable bbal on the observed characteristics of the students (pre-test score and male). None of the two variables are statistically different from 0 at conventional confidence levels. Here the key variable we want to be balanced is the pre-treatment scores, since it is the most relevant dimension that influences the outcome variable, the post-treatment scores.

```
lmod1 <- lm(bal ~ pre_totnorm + male, data=df1)
summ(lmod1)
```

Observations	9745
Dependent variable	bal
Type	OLS linear regression

F(2,9742)	1.82
R²	0.00
Adj. R²	0.00

	Est.	S.E.	t val.	p
(Intercept)	0.48	0.01	67.56	0.00
pre_totnorm	0.00	0.01	0.51	0.61
male	0.02	0.01	1.85	0.06

Standard errors: OLS

We can estimate the one-year remedial treatment effect for Vadodara for both mathematics and verbal in 2001-2002. The effect of treatment is 0.18 standard deviation improvement for math performance and 0.13 standard deviation improvement for verbal performance. The effects are quite large, but not statistically significant at the 5% level for verbal performance.

```
lmod2 <- lm(post_mathnorm ~ bal, data=df1)
lmod3 <- lm(post_verbnorm ~ bal, data=df1)
stargazer(lmod2, lmod, style='aer', type='text', title='Estimation of 1-year treatment
effects')
```

```
##
## Estimation of 1-year treatment effects
## ================================================================================
##                            post_mathnorm                 lwage
##                                 (1)                       (2)
## --------------------------------------------------------------------------------
## bal                            0.184***
##                                (0.025)
##
## educ                                                    0.021***
##                                                          (0.003)
##
## KWW                                                     0.019***
##                                                          (0.001)
##
## Constant                       0.178***                 5.351***
##                                (0.017)                   (0.039)
##
## Observations                    8,065                     2,963
## R2                              0.007                     0.196
## Adjusted R2                     0.007                     0.195
## Residual Std. Error    1.123 (df = 8063)         0.396 (df = 2960)
## F Statistic         54.071*** (df = 1; 8063) 360.893*** (df = 2; 2960)
## --------------------------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

This time, we estimate the treatment effect on the change in scores as the outcome. The effects are 0.2 standard deviation with s.e. 0.055 for math and 0.11 stnadard deviation with s.e. 0.057 for language.

```
lmod4 <- lm(I(post_mathnorm-pre_mathnorm) ~ bal + pre_mathnorm, data=df1)
lmod5 <- lm(I(post_verbnorm-pre_verbnorm) ~ bal + pre_verbnorm, data=df1)
stargazer(lmod4, lmod5, style='aer', type='text', title='Estimation of 1-year treatmen
t effects')
```

```
##
## Estimation of 1-year treatment effects
## ================================================================================
============
##                                 I(post_mathnorm - pre_mathnorm) I(post_verbnorm - p
re_verbnorm)
##                                            (1)                             (2)
## --------------------------------------------------------------------------------
------------
## bal                                     0.211***                        0.127***
##                                          (0.021)                         (0.021)
##
## pre_mathnorm                            -0.392***
##                                          (0.011)
##
## pre_verbnorm                                                            -0.263**
*
##                                                                          (0.010)
##
## Constant                                0.148***                        0.656***
##                                          (0.015)                         (0.014)
##
## Observations                              8,065                           8,065
## R2                                        0.157                           0.080
## Adjusted R2                               0.157                           0.080
## Residual Std. Error (df = 8062)           0.945                           0.922
## F Statistic (df = 2; 8062)              749.603***                      349.964**
*
## --------------------------------------------------------------------------------
------------
## Notes:                          ***Significant at the 1 percent level.
##                                 **Significant at the 5 percent level.
##                                 *Significant at the 10 percent level.
```

We can also estimate the two-year remedial treatment effect for Vadodara for both mathematics and verbal scores. We first merge the two datasets and standardize all relevant pre and post scores. The important thing here is to consider that 'only chilren who were in grade 3 in year 1 can be exposed for two years. Thus, the two year effect is estimated using substantially fewer students than the one-year effect". So, I matched the two data sets using the student id unique identifier, and the effect will be estimated only for students of grade 4.

Results are 0.33 standard deviation for math and 0.21 standard deviation for verbal, both very high. given the sample size, we can have larger deviations from what they have in the paper, provided I normalized scores the same way they did.

```
# create new outcome variables that measure before and after score change in math and
verbal
df1_2 <-
  inner_join(df1,df2,by='studentid')
```

```
## Warning: Column `studentid` has different attributes on LHS and RHS of join
```

```
df1_2 <-
  df1_2 %>%
  mutate(pre_mathnorm.x = scale(pre_math.x),
         pre_verbnorm.x = scale(pre_verb.x)) %>%
  mutate(math_score_change = post_mathnorm.y-pre_mathnorm.x,
         verb_score_change = post_verbnorm.y-pre_verbnorm.x)

lmod6 <- lm(math_score_change ~ bal.x + pre_mathnorm, data=df1_2)
lmod7 <- lm(verb_score_change ~ bal.x + pre_verbnorm, data=df1_2)
stargazer(lmod6, lmod7, style = 'aer', type = 'text', title = 'Estimation of 2-year tr
eatment effects')
```

```
##
## Estimation of 2-year treatment effects
## ================================================================================
##                                  math_score_change   verb_score_change
##                                         (1)                 (2)
## --------------------------------------------------------------------------------
## bal.x                                 0.289***            0.161***
##                                        (0.037)             (0.035)
##
## pre_mathnorm                          -0.616***
##                                        (0.021)
##
## pre_verbnorm                                              -0.413***
##                                                            (0.019)
##
## Constant                              0.769***            0.671***
##                                        (0.027)             (0.026)
##
## Observations                            3,145               3,145
## R2                                      0.234               0.134
## Adjusted R2                             0.233               0.133
## Residual Std. Error (df = 3142)         1.028               0.993
## F Statistic (df = 2; 3142)           479.713***          242.983***
## --------------------------------------------------------------------------------
## Notes:                             ***Significant at the 1 percent level.
##                                    **Significant at the 5 percent level.
##                                    *Significant at the 10 percent level.
```

Experiment Validity Check

Internal Validity: The paper seems to be ver robust in terms of internal validity, since the authors show or argue by institutional knowledge that treatment was actually randomly assigned, there were no different resources allocations, and attrition seems to be uncorrelated with the observables. Therefore, this seems to be a very good paper in terms of the credibility of identifying assumptions.

External Validity: It would seem like this intervnetion is very successful and cost-effective. Thus, it would make sense to scale it up in India. Notice that the cost-effectiveness of the program comes from the very small pay that these remedial teachers receive, so if we increase the demand for these professionals a lot, we might end up paying a lot more for their services, unless the supply is inelastic. Hence the program would lose its most attractive feature. One thus needs to carefully consider the supply of these remedial teachers in the labor market before scaling up the program. Also, we don't know what is good about these remedial teachers: is it they are young, that they are much more similar to the students they teach to in terms of social background? Then if increasing the number of remedial teachers leads to a change in the type of these teachers aids the treatment effects might differ. Also notice we have an experiment in a poor and a richer urban setting, but we don't really know anything about implement the program in rural areas, which generally are much poorer, have stron teacher absenteeism problems, etc. Also, it's difficult to extrapolate these results to other developing countries

since the remedial teacher is rather peculiar to India, and this country also seems to suffer a lot from teaching targeted to some elites, so we might assess if this is the case also in the other country one might think of extending the program to. Ggenerally, it is very difficult to generalize the results of social experiments tied to a specific institutional setting, especially when these interventions are complex and non-standardized.

Simulated Datasets Case1, Case2, and Case3

These datasets are simulated based on the real data we analyzed above. In each case, some students are given a treatment and others entered the control group. Randomization may or my not have been performed correctly. The effect of treatment may be heterogeneous in the population - both along observable or unobservable dimensions. In some cases, there may be treatment group dropout. In no case, however, students assigned to the control group were treated, i.e. no treatment substitution bias. The treatment variable of interest is treated and the outcome variable is FinalScore.

First, we check for balane of pre-treatment covariates in the treatment and control groups. Something didn't work in the randomization since income and the pre-treatment scores predict the treatment status. People with higher income are more likely to be treated. This effect is not small because income varies from 50 to 300 a differenceo f 100 gives about 20 percentage points higher probability ofb eing treated. Instead, pre-treatment scores don't seem to impact the probability of treatment significantly. All else being equal 1 standard deviation higher pre-scores decreasses the probability of being treated by 1.25 percentage points. This seems more to be the effect of sampling variation and notice this bias works in the opposite direction of the income one. We should always look at magnitude and signs after looking at statistical significance of variables.

```
lmod8 <- glm(treated ~ pre_totnorm + numstud + male + income + std, data=df3, family=b
inomial(link='logit'))
summ(lmod8)
```

Observations	12415
Dependent variable	treated
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(5)$	353.64
Pseudo-R^2 (Cragg-Uhler)	0.04
Pseudo-R^2 (McFadden)	0.02
AIC	16380.53
BIC	16425.09

	Est.	S.E.	z val.	p
(Intercept)	-0.95	0.17	-5.63	0.00
pre_totnorm	-0.05	0.03	-1.75	0.08
numstud	0.00	0.00	1.38	0.17
male	0.00	0.04	0.08	0.94
income	0.01	0.00	14.15	0.00
std	-0.03	0.04	-0.68	0.49

Standard errors: MLE

Next, we estimate the causal effect of the treatment on the outcome variable, check whether we need to control for any other pre-treatment variable to ensure consistency of the parameter of interest. Clearly, a naive regression of treatment on outcome gives a biased result, as income is also related to the outcome variable. After discovering treatment was not randomly assigned in the population, we cannot be sure what we estimate is the causal effect of interest, especially if you dont have many other control variables to perform sensitivity analysis.

```
lmod9 <- lm(Finalscore ~ treated, data=df3)
lmod10 <- lm(Finalscore ~ treated + income, data=df3)
lmod11 <- lm(Finalscore ~ treated + income + pre_totnorm, data=df3)
stargazer(lmod9, lmod10, lmod11, title='Estimation of Treatment Effect', style='aer',
type='text')
```

```
##
## Estimation of Treatment Effect
## ================================================================================
=======================
##                                                          Finalscore
##                              (1)                          (2)
(3)
## --------------------------------------------------------------------------------
-----------------------
## treated                     0.520***                    0.272***
0.298***
##                             (0.019)                     (0.013)
(0.003)
##
## income                                                  0.017***
0.000
##                                                         (0.000)
(0.000)
##
## pre_totnorm
0.998***
##
(0.002)
##
## Constant                   -0.105***                   -2.583***
-0.001
##                             (0.014)                     (0.024)
(0.008)
##
## Observations                 12,415                      12,415
12,415
## R2                           0.060                       0.537
0.972
## Adjusted R2                  0.060                       0.537
0.972
## Residual Std. Error     1.012 (df = 12413)          0.710 (df = 12412)
0.173 (df = 12411)
## F Statistic         786.448*** (df = 1; 12413) 7,202.215*** (df = 2; 12412) 146,00
1.600*** (df = 3; 12411)
## --------------------------------------------------------------------------------
-----------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

We can improve the precision of our estimate. Notice the sharp decrease in the standard error of the treatment estimate after you add pre_totnorm: this variable explains a lot of the variation in the outcome variable, and consequently the final estimates are much less noisy.

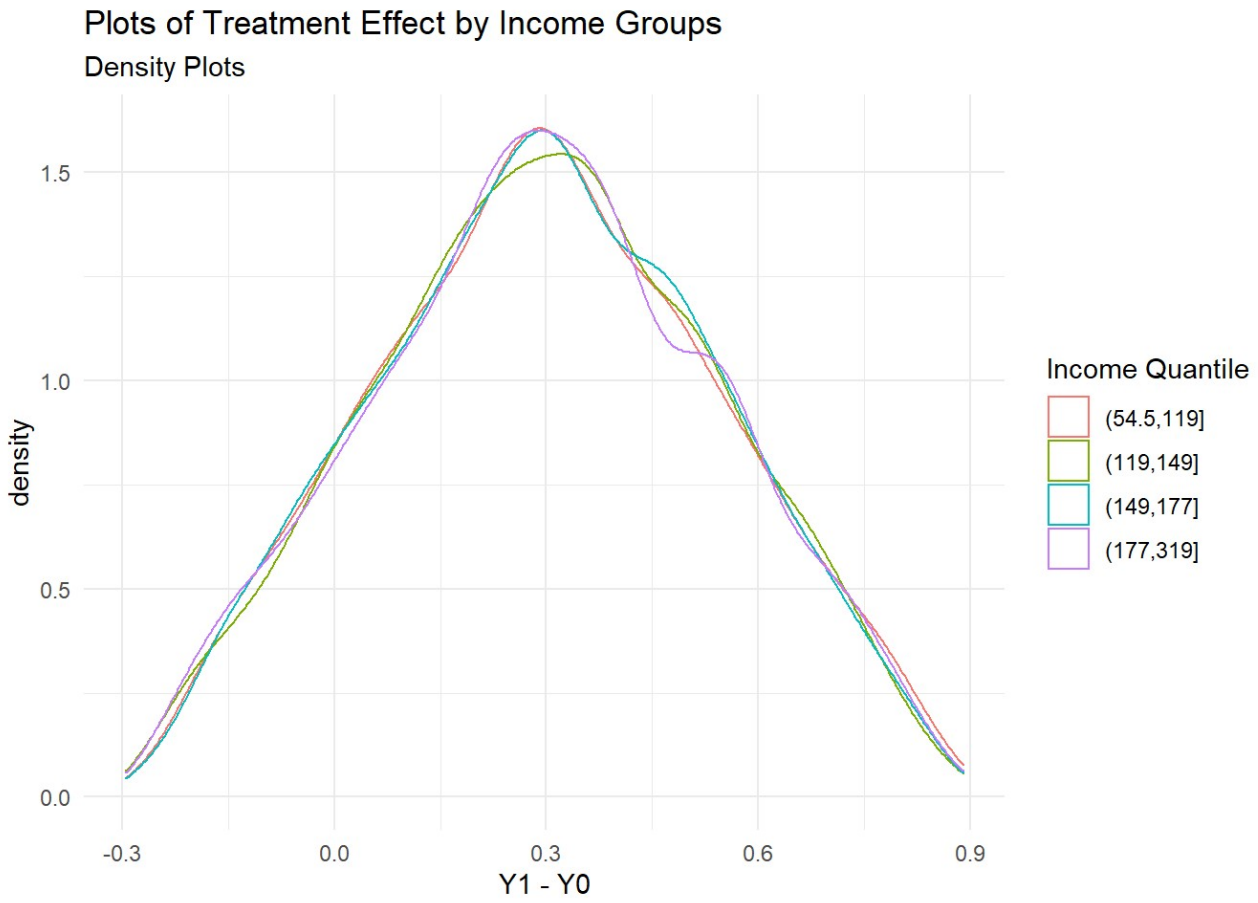
We can also perform a subgroup analysis by gender, and test if the treatment effect differs along observed dimensions in the population. At least along observable dimensions, there is no difference in the treatment effect since the coefficients on the interaction terms are all statistically insignificant. Also, notice that you need to add the main effects beyond the interactions in our regression, otherwise our interactions will pick up the main effects of the pre-treatment variables, even if there is no differential effect among these dimensions.

```
lmod12 <- lm(Finalscore ~ treated*income + male*treated + treated*pre_totnorm, data=df
3)
stargazer(lmod12, style='aer',title='Estimation including interaction terms', type='te
xt')
```

```
##
## Estimation including interaction terms
## ============================================================
##                                  Finalscore
## ------------------------------------------------------------
## treated                           0.307***
##                                   (0.017)
##
## income                             0.000
##                                   (0.000)
##
## male                               -0.003
##                                   (0.005)
##
## pre_totnorm                       0.996***
##                                   (0.004)
##
## treated:income                     -0.000
##                                   (0.000)
##
## treated:male                       0.003
##                                   (0.006)
##
## treated:pre_totnorm                0.003
##                                   (0.005)
##
## Constant                           -0.006
##                                   (0.013)
##
## Observations                       12,415
## R2                                 0.972
## Adjusted R2                        0.972
## Residual Std. Error           0.173 (df = 12407)
## F Statistic             62,555.940*** (df = 7; 12407)
## ------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

Supposing magically that we have been provided with counterfactuals Y0 and Y1, that is for every unit we know what would have happened if either the treated unit would have not been treated or the non-trerated unit would have been treated. We can now check if what we estimated is truly the causal effect of interest for the treated units.

```
# dissect income variable into four quartiles and look at density of performance chang
e per each quartile
df3 %>%
  mutate(income_quantile = cut(income, breaks=quantile(income, probs = seq(0,1,0.2
5)))) %>%
  drop_na(income_quantile) %>%
  ggplot(aes(x=Y1-Y0)) +
    geom_density(aes(color=income_quantile)) +
    labs(title='Plots of Treatment Effect by Income Groups',
        subtitle='Density Plots',
        color='Income Quantile') +
    theme_minimal()
```



This means that we have the true causal effect for everyone in the sample after controlling for income. In this case E(Y1-Y0|income) = E(Y1-Y0) coincides with the average treatment effect in the population, as treatment effects do not seem to be heterogeneous by income or other variables.

Dataset Case2

Balance Test

Everything is OK here. We can be pretty sure that we are estimating the causal effect of interest: in particular having balance in the pre-treatment scores seems to point in the direction of the randomization being successfully achieving its goal.

```
lmod13 <- glm(treated ~ pre_totnorm + numstud + male + income + std, data=df4, family=
binomial(link='logit'))
stargazer(lmod13, title = 'Balance Test', style = 'aer', type = 'text')
```

```
##
## Balance Test
## ===========================================================
##                                  treated
## -----------------------------------------------------------
## pre_totnorm                       -0.014
##                                  (0.026)
##
## numstud                           0.000
##                                  (0.001)
##
## male                              -0.005
##                                  (0.036)
##
## income                            -0.000
##                                  (0.001)
##
## std                               -0.022
##                                  (0.036)
##
## Constant                          0.078
##                                  (0.162)
##
## Observations                      12,415
## Log Likelihood                   -8,604.788
## Akaike Inf. Crit.                17,221.580
## -----------------------------------------------------------
## Notes:            ***Significant at the 1 percent level.
##                   **Significant at the 5 percent level.
##                   *Significant at the 10 percent level.
```

Initial Model on Treatment and Expanded Model

Simple regression is enough to achieve unbiased and consistent estimates of the causal effect. We don't really need to control for anything else as all other variables affecting treatment and the outcome are balanced between the treatment and control group. Adding more variables increases precision of estimates.

```
lmod14 <- lm(Finalscore ~ treated, data = df4)
lmod15 <- lm(Finalscore ~ treated + income + pre_totnorm, data = df4)
stargazer(lmod14, lmod15, title = 'Estimation of Treatment Effect', style = 'aer', typ
e = 'text')
```

```
##
## Estimation of Treatment Effect
## ================================================================================
##                                          Finalscore
##                              (1)                          (2)
## --------------------------------------------------------------------------------
## treated                     0.031*                      0.047***
##                            (0.019)                      (0.004)
##
## income                                                   0.000
##                                                         (0.000)
##
## pre_totnorm                                             0.998***
##                                                         (0.003)
##
## Constant                   0.034***                      -0.004
##                            (0.013)                      (0.012)
##
## Observations                12,415                      12,415
## R2                          0.000                        0.942
## Adjusted R2                 0.000                        0.942
## Residual Std. Error   1.034 (df = 12413)        0.249 (df = 12411)
## F Statistic         2.788* (df = 1; 12413) 67,449.330*** (df = 3; 12411)
## --------------------------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

Subgroup Analysis

First, we notice that our effect is not any longer statistically different from zero. There seems to be a larger gain for males, but this efffect is not statistically significant at the 5% level. This is what we would call borderline statistical significance. The absolute magnitude howver, is not that big, and again the main effect is not statistically significant from zero. The treatment effect for males is borderline significant.

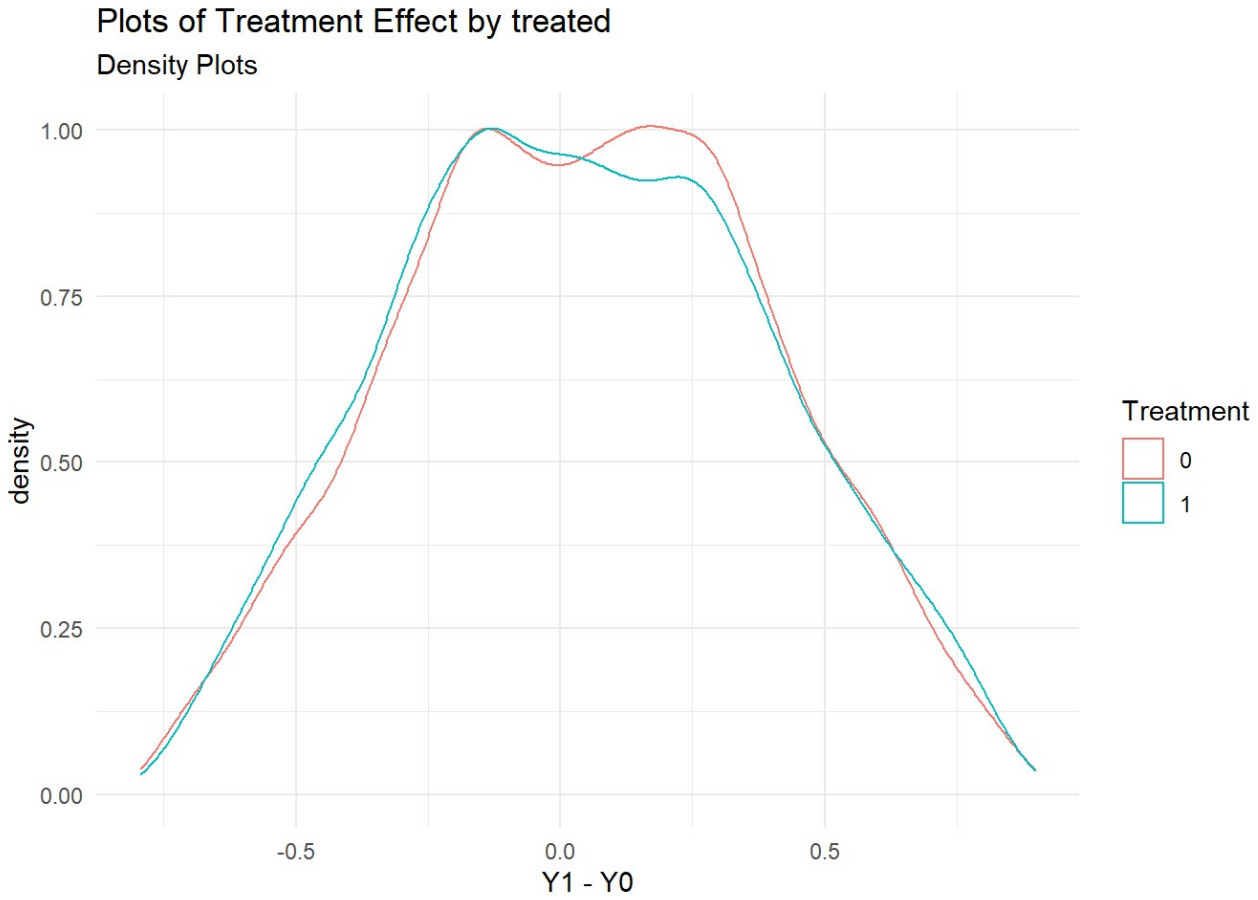
```
lmod16 <- lm(Finalscore ~ treated*income + male*treated + treated*pre_totnorm, data =
df4)
stargazer(lmod16, style = 'aer',title = 'Estimation including interaction terms', type
= 'text')
```

```
##
## Estimation including interaction terms
## ============================================================
##                                  Finalscore
## ------------------------------------------------------------
## treated                           0.029
##                                  (0.024)
##
## income                            -0.000
##                                  (0.000)
##
## male                              -0.005
##                                  (0.006)
##
## pre_totnorm                      0.997***
##                                  (0.005)
##
## treated:income                    0.000
##                                  (0.000)
##
## treated:male                      0.015*
##                                  (0.009)
##
## treated:pre_totnorm               0.002
##                                  (0.006)
##
## Constant                          0.003
##                                  (0.017)
##
## Observations                      12,415
## R2                                0.942
## Adjusted R2                       0.942
## Residual Std. Error          0.249 (df = 12407)
## F Statistic             28,907.770*** (df = 7; 12407)
## ------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

Density Plot on Counterfactuals

Our estimate of 0.046 is very close to the true effect for the treated units, and it is actually the effect in the population. Notice the weird shape of the distribution of the effect in the population. It seems like there are two distributions mixed together.

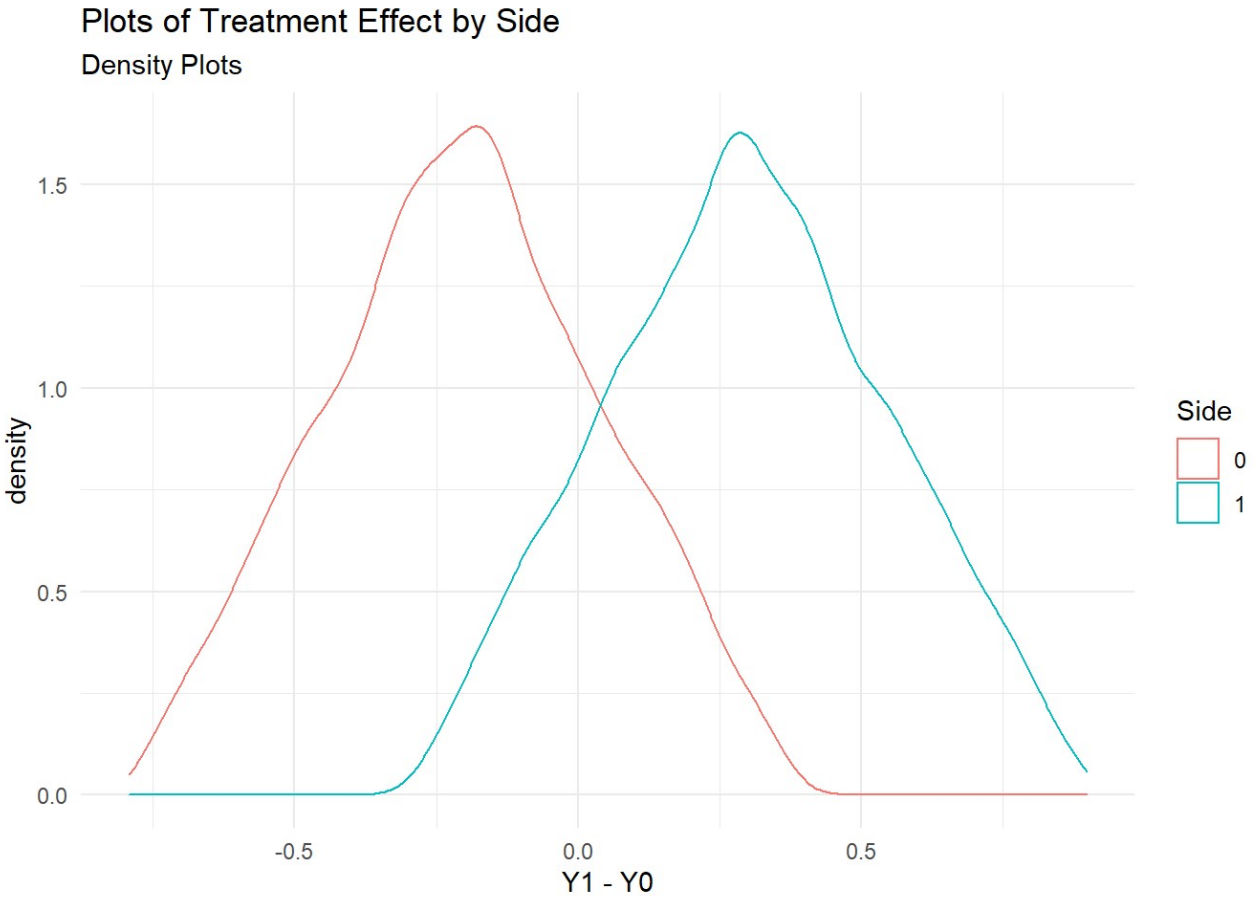
```
# density plot per treatment status
df4 %>%
  ggplot(aes(x=Y1-Y0)) +
    geom_density(aes(color=factor(treated))) +
    labs(title = 'Plots of Treatment Effect by treated',
        subtitle = 'Density Plots',
        color = 'Treatment') +
    theme_minimal()
```



We can see that along an unobserved dimension the treatment effect is very different: for some people, it is on average -.20 and for other .29. Here the average treatment effect is actually very misleading, since we are missing that some people truly get something out of the treatment while for other people on average the effect is negative. Also even if before I could have concluded that the average treatment effect for those who are treated is slightly positive, now that I know the effect is highly heterogeneous I surely need to consier what drives the heterogeneity. So for instance if side==0 corresponds to those who are less wealthy implementing the treatment could have perverse distributional effects, since the less wealthy people will on average be worse off while the weawlthy people will improve their lot. In

general this is not what you would want to see. The take-home lesson is that sometimes the average treatment effect for the treated, even if estimated without bias, can mask things we are not able to investigate and can be very misleading.

```
# density plot per side
df4 %>%
  ggplot(aes(x=Y1-Y0)) +
    geom_density(aes(color=factor(side))) +
    labs(title = 'Plots of Treatment Effect by Side',
        subtitle = 'Density Plots',
        color = 'Side') +
    theme_minimal()
```



Case 3

Balance Check

Everything is OK here. We can be pretty sure the randomization worked properly: in particular having balance in the pre-treatment scores seems to point in the direction of the randomization working like it should. Among those who were actually treated I have richer people and people with slightlylower pre-treatment test scores. Given the randomization was OK this implies that therewas selective treatment dropout or treatment crossover.

```
lmod17 <- lm(TreatmentGroup ~ pre_totnorm + numstud + male + income + std, data=df5)
lmod18 <- lm(treated ~ pre_totnorm + numstud + male + income + std, data= df5)
stargazer(lmod17, lmod18, style = 'aer', type = 'text', title = 'Balance test')
```

```
##
## Balance test
## ================================================================================
##                                       TreatmentGroup         treated
##                                            (1)                 (2)
## --------------------------------------------------------------------------------
## pre_totnorm                                -0.006            -0.020***
##                                           (0.007)             (0.006)
##
## numstud                                    0.000               0.000
##                                           (0.000)             (0.000)
##
## male                                       0.002               -0.005
##                                           (0.009)             (0.008)
##
## income                                     0.000              0.002***
##                                           (0.000)             (0.000)
##
## std                                        -0.008              -0.000
##                                           (0.009)             (0.008)
##
## Constant                                  0.494***             0.004
##                                           (0.041)             (0.038)
##
## Observations                               12,415              12,415
## R2                                         0.000               0.030
## Adjusted R2                                -0.000              0.030
## Residual Std. Error (df = 12409)           0.500               0.469
## F Statistic (df = 5; 12409)                0.685             76.728***
## --------------------------------------------------------------------------------
## Notes:                             ***Significant at the 1 percent level.
##                                    **Significant at the 5 percent level.
##                                    *Significant at the 10 percent level.
```

Poorer people were more likely to drop out from the treatment. For instance an income 100 units higher implies 40 percentage points less likely to drop out. Also thepre-treatment scores are statistically significant determinants of treatment dropout but notice that the effect is really small (2 percentage points more likely to dropout for 1more standard deviation in scores).

As usual notice the increase in precision after you control for a variable with strongexplanatory power. The intention-to-treat effect is the broad effect, or policy effect, ofa dministering the treatment. In the real world some people will decide to not take thetreatment even if they were at the beginning willing to take it and some people who didn't want to be treated will actually get the treatment. So this is like a

broad effect of administering the treatment, that takes into account the true behavior of people. Thisis however not the true effect of the treatment for the treated, since in this context weare attributing the effect of the treatment to people who didn't actually get treated (=treatment drop out).

```
lmod19 <- lm(treated ~ pre_totnorm + numstud + male + income + std, data = df5, subset
= (TreatmentGroup == 1))
stargazer(lmod19, style = 'aer', type = 'text', title = 'Determinants of dropouts')
```

```
##
## Determinants of dropouts
## ============================================================
##                                     treated
## ------------------------------------------------------------
## pre_totnorm                        -0.029***
##                                     (0.008)
##
## numstud                             -0.000
##                                     (0.000)
##
## male                                -0.009
##                                     (0.011)
##
## income                              0.004***
##                                     (0.000)
##
## std                                  0.009
##                                     (0.011)
##
## Constant                             0.035
##                                     (0.050)
##
## Observations                         6,236
## R2                                   0.121
## Adjusted R2                          0.120
## Residual Std. Error           0.433 (df = 6230)
## F Statistic               170.744*** (df = 5; 6230)
## ------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

We run a naive regression of outcome on treatment. This is not the causal effect of the treatment on the outcome since we realized in step2 that people who are poorer drop out more than rich people, and being rich or pooris surely correlated with final grades. You can see this also by the fact that when you control for pre-scores (that are correlated with income) the point estimate on the treatment variable changes.

```
lmod20 <- lm(Finalscore ~ TreatmentGroup, data = df5)
lmod21 <- lm(Finalscore ~ TreatmentGroup + pre_totnorm, data = df5)
stargazer(lmod20, lmod21, style = 'aer', type = 'text', title = 'Balance test')
```

```
##
## Balance test
## ================================================================================
##                                            Finalscore
##                                (1)                          (2)
## --------------------------------------------------------------------------------
## TreatmentGroup               0.201***                     0.208***
##                              (0.019)                      (0.004)
##
## pre_totnorm                                               1.016***
##                                                           (0.002)
##
## Constant                     0.029**                       -0.001
##                              (0.013)                      (0.003)
##
## Observations                  12,415                       12,415
## R2                            0.009                        0.964
## Adjusted R2                   0.009                        0.964
## Residual Std. Error     1.039 (df = 12413)           0.198 (df = 12412)
## F Statistic         116.219*** (df = 1; 12413) 166,209.000*** (df = 2; 12412)
## --------------------------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

We now instrument the treatment with the assignment to treatment and get the LATE. The local average treatment effect is the effect of the treatment for a particular subsetof the population, the “compliers”. These are the people who in an IV setting receivea treatment because they are affected by the instrument. Here this is the effect oftreatment for those people who would be treated if they were assigned to the treatment,but would have not been treated in the counterfactual case in which they were not. assigned to the treatment group. In experiments where there was treatment dropout orcross-overs this is very useful to estimate the effect for those who actually got treated.Also notice that for binary(dummy) treatment and instruments the LATE is defined. That is the intention to treat effect divided by the difference of the fraction of the peoplewho were assigned to treatment and got treated and those who were not assigned totreatment and got treated. Here P(T=1???Z=0)=0 since there was no cross-over, P(T=1???Z=1)=0.693, and we estimatedIT T=.201.

```
lmod22 <- lm(Finalscore ~ treated, data = df5)
lmod23 <- lm(Finalscore ~ treated + pre_totnorm, data = df5)
stargazer(lmod22, lmod23, style = 'aer', type = 'text', title = 'Balance test')
```

```
##
## Balance test
## ================================================================================
##                                          Finalscore
##                              (1)                          (2)
## --------------------------------------------------------------------------------
## treated                     0.516***                     0.297***
##                             (0.019)                      (0.003)
##
## pre_totnorm                                              1.001***
##                                                          (0.002)
##
## Constant                   -0.049***                     -0.000
##                             (0.011)                      (0.002)
##
## Observations                 12,415                       12,415
## R2                           0.055                        0.972
## Adjusted R2                  0.055                        0.972
## Residual Std. Error     1.015 (df = 12413)          0.174 (df = 12412)
## F Statistic         729.238*** (df = 1; 12413) 217,412.400*** (df = 2; 12412)
## --------------------------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

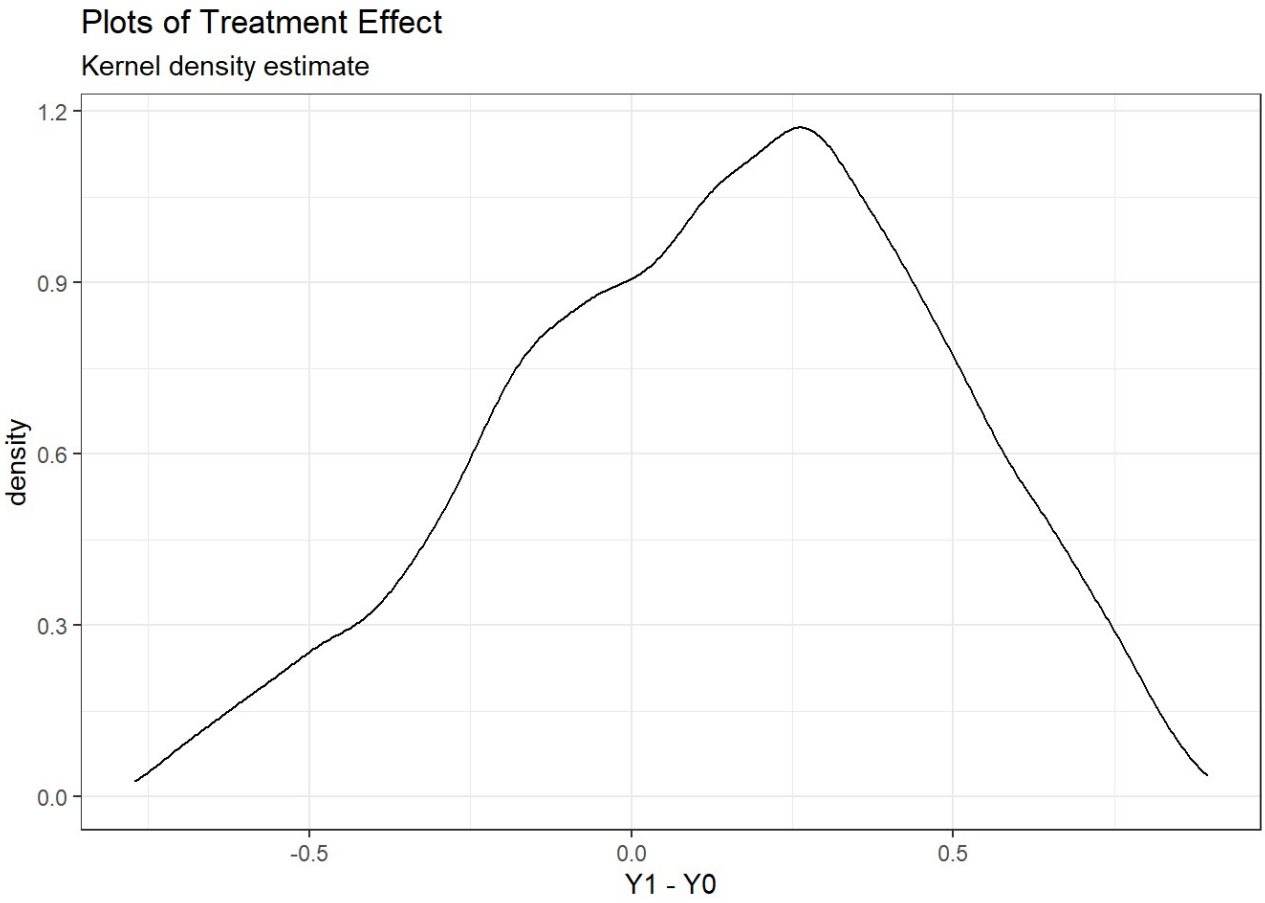
The average treatment effectY1???Y0among the treated population is 0.290. So the LATE==E[Y1??? Y0???D1=1, D0=0]in this case was also the average effect of the treatment forthe treated population,T T=E[Y1???Y0???D=1].

```
lmod24 <- lm(Y1 - Y0 ~ treated, data = df5)
lmod25 <- lm(Y1 - Y0 ~ TreatmentGroup, data = df5)
stargazer(lmod24, lmod25, style = 'aer', type = 'text', title = 'Balance test')
```

```
##
## Balance test
## ================================================================================
##                                                 Y1 - Y0
##                                           (1)                 (2)
## --------------------------------------------------------------------------------
## treated                                0.237***
##                                         (0.006)
##
## TreatmentGroup                                                0.002
##                                                              (0.006)
##
## Constant                               0.060***            0.142***
##                                         (0.004)             (0.004)
##
## Observations                            12,415              12,415
## R2                                       0.114               0.000
## Adjusted R2                              0.113              -0.000
## Residual Std. Error (df = 12413)         0.316               0.336
## F Statistic (df = 1; 12413)          1,589.917***            0.078
## --------------------------------------------------------------------------------
## Notes:                            ***Significant at the 1 percent level.
##                                   **Significant at the 5 percent level.
##                                   *Significant at the 10 percent level.
```

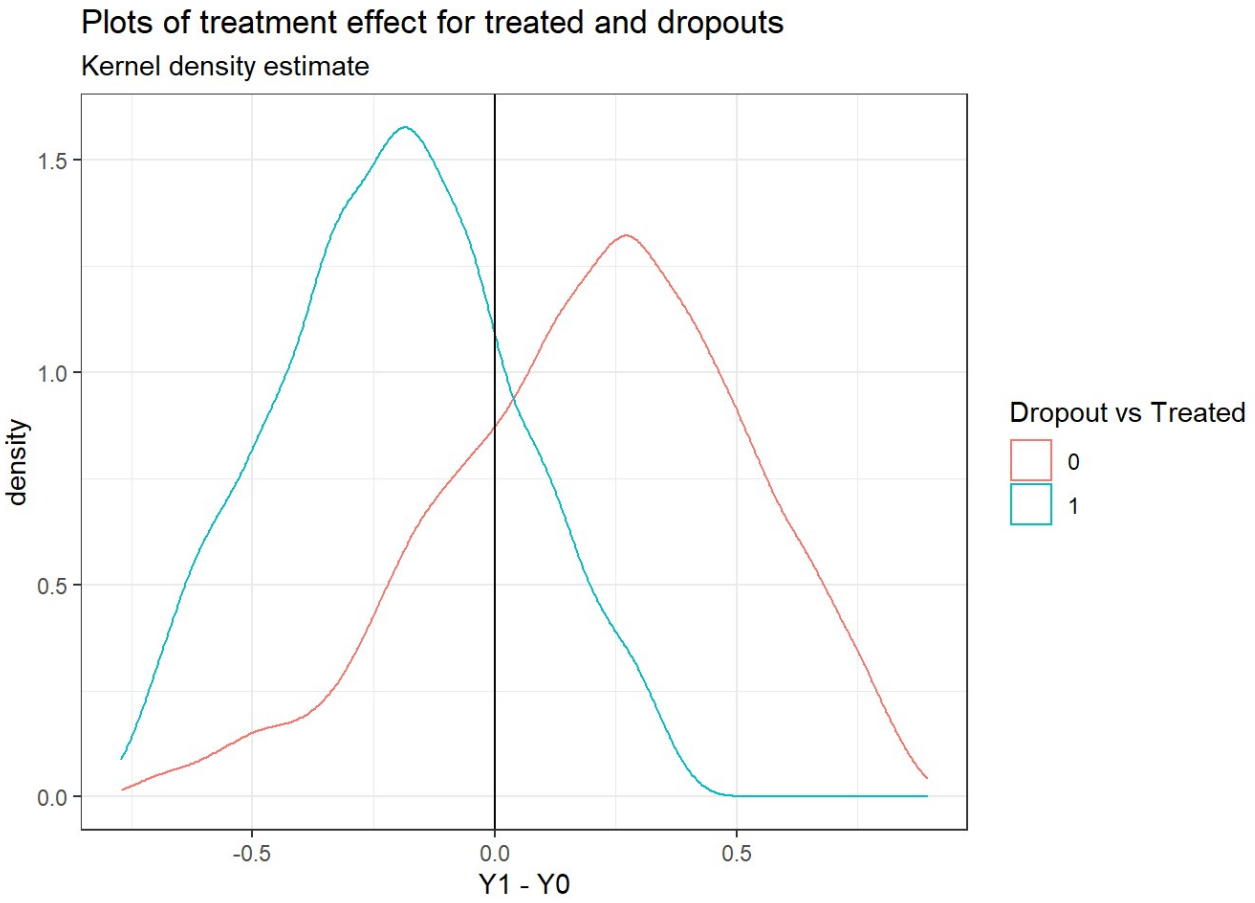
Those people who belong to the lower distribution all dropped out of treatment whilethose who belong to the other distribution received the treatment. Notice this is not fullyrational behavior. Figure 5 shows the distribution of causal effects for those who droppedout (blue line) and those who actually received treatment (red line). Everyone to the leftof the vertical line at zero would gain from treatment, while those to the left would not.Those who received treatment were therefore gaining more on average, while there existsome dropouts who would have gained and some who received treatment who would havebeen better of dropping out. Therefore, the people didn’t really know precisely what theywould have gotten out of being treated, yet they were able to understand their “type”and accordingly dropped out or got treated. This is consistent with some informationstory where the individuals who are assigned to treatment understand something aboutthe potential gains but cannot truly estimate the causal effect.

```
ggplot(df5, aes(x=Y1-Y0)) +
  geom_density() +
  labs(title = 'Plots of Treatment Effect',
       subtitle = 'Kernel density estimate') +
  theme_bw()
```



RCT can be a powerful too to solve the problem of selection bias if randomization isdone properly and there is full compliance. In this case, regressing the outcome on thetreatment variable can give us an unbiased estimate of the average treatment effect.However, when there is attrition or dropout, we would need to consider the sorting gain.In this case, treatment group is assigned randomly but the choice of receiving treatmentwithin that group is not random. We can get the intent-to-treat effect by regressingoutcome on the randomized assignment variable. If we want to understand the actualtreatment effect, we are only able to estimate the local average treatment effect forthe compliers who only receiving treatment when they are randomly assigned to thetreatment group.

```
df5 %>%
  mutate(dropout = ifelse(TreatmentGroup == 1 & treated == 0, 1, 0)) %>%
  ggplot(aes(x = Y1 - Y0)) +
  geom_density(aes(color = factor(dropout))) +
  geom_vline(xintercept = 0) +
  labs(title = 'Plots of treatment effect for treated and dropouts',
       subtitle = 'Kernel density estimate',
       color = 'Dropout vs Treated') +
  theme_bw()
```



Instrumental Variable Regression

IV is an alternative method to OLS or randomized experiment. Randomized experiment is often not feasible, and OLS doesnt always have all unobserved confounding variables. For example, let us consider a model relating education and wage. OLS assumes zero conditional mean and that the only effect of X on Y is through beta x X. However, we need to account for ability that can induce a correlation between X and error. In this case, X is endogenous, and people with high ability are likely to have high education. This renders OLS estimator inconsistent and cannot be abscribed causal interpretation.

IV is a way to deal with endogeneity bias used when a model has an endogenous variable X. In theory, it its able to detect movements in X that are uncorrelated with error U, and use these to estimte beta. IV is used in cases where the following cases are of concern when 1) omitted variable bias 2) selection bias 3) simultaneity of Y and X 4) error in measurement can cause nonzero

Dataset

This data set investigates the causal effect of compusory eucation on earnings. Their identification is based on the fact that, given laws on school enrollment mandate that kids can enter school if their birthday is before January 1st of the year school starts and then can drop out at the completion of their 16th year, people born earlier in the year reach the minimum age for dropout before people born later in

the calendar year and hence can dropout with less education. If the time of the year in which people are born affects wages only through this effect years of compulsory education then it is possible to use this exogenous variation to estimate the effect of compulsory schooling on earnings for the population affecte by these constraints. This is a natural-experiment setting, where arguable exogenous mechanism that generates variation in the treatment is well understood and created by the institution/laws.

```
df <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/CENSUS7080.DTA')
glimpse(df)
```

```
## Rows: 1,063,634
## Columns: 17
## $ AGEQ     <dbl> 40.50, 41.00, 41.50, 46.25, 46.00, 47.00, 48.75, 41.75, 47...
## $ EDUC     <dbl> 11, 12, 12, 12, 16, 12, 14, 9, 12, 17, 17, 16, 8, 10, 9, 1...
## $ ENOCENT  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ESOCENT  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ LWKLYWGE <dbl> 5.023558, 5.061540, 5.378315, 5.178639, 6.378776, 4.997411...
## $ MARRIED  <dbl> 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1...
## $ MIDATL   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ MT       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ NEWENG   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CENSUS   <dbl> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70...
## $ QOB      <dbl> 3, 1, 3, 4, 1, 1, 2, 2, 3, 2, 2, 2, 3, 4, 3, 3, 3, 4, 2, 3...
## $ RACE     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ SMSA     <dbl> 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0...
## $ SOATL    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ WNOCENT  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ WSOCENT  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ YOB      <dbl> 1929, 1929, 1928, 1923, 1924, 1923, 1921, 1928, 1922, 1925...
```

Identification Methodology of Instrument Variable

IV method can yield a consistent esstimator only if instrument is valid. We require instrument variable Z to be correlated with regressor X but uncorrelated with error U . We will later examine these two requirements as we fit the model. In practice, is often difficlt to obtain instrument that satifies both criteria.

IV estimate is obtained by Two-Stage Least Squares method (2SLS). - 1st stage: regress instrument Z on X and obtain estimated X. This step decomposes X into one component that is correlated with error U and another component that is uncorrelated with U since Z is ideally exogenous. It thus isolated the part of X that is correlated with the error term and rid of it. - 2nd stage: regress estimated X on Y. It only uses the exogenous (problem-free) component of X to estimate beta.

There are benefits to the IV regression and they are: - IV estimator is asymptotically normally distributed and we can conduct hypothesis testing - IV estimator is consistent - standard errors are smaller when correlation between instrument Z and X is stronger - it can bypass omitted variable problem and yield a consistent estimator as long as we find an instrument uncorrelated with error/omitted variable - it can detect causality in reverse causality problem in simultaneity situation (e.g. Levitt's crime rate vs policing paper)

Practical Tips for Finding Instruments

The most difficult practical part of IV model is finding the right instruments. Researchers find the right instrument by 1) random draws 2) natural randomness 3) institutional features.

Fitting IV Model

We can use the AER package to fit IV model. It is done in two stages.

```
iv <- ivreg(LWKLYWGE ~ AGEQ+ EDUC + MARRIED + RACE + ENOCENT + ESOCENT + MIDATL + MT +
NEWENG + SOATL + WNOCENT + WSOCENT| QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT +
MIDATL + MT + NEWENG + SOATL + WNOCENT + WSOCENT,
            data = df)
iv$formula
```

```
## LWKLYWGE ~ AGEQ + EDUC + MARRIED + RACE + ENOCENT + ESOCENT +
##     MIDATL + MT + NEWENG + SOATL + WNOCENT + WSOCENT | QOB +
##     AGEQ + MARRIED + RACE + ENOCENT + ESOCENT + MIDATL + MT +
##     NEWENG + SOATL + WNOCENT + WSOCENT
```

outcome variable: log wage endogenous variable: Educational years instrument: age exogenous variables: race, marital status, location

Results

We can present the coefficient values and standard errors.

When IV estimate is "too big", then there are two possibilities: 1) instrument is not valid and is correlated with error 2) first stage regression is weak and inflating the IV estimate

```
# focus solely on the coefficients controlling for heteroskedasticity
coeftest(iv, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  4.7569e+00  1.3419e-01  35.4505 < 2.2e-16 ***
## AGEQ         3.5463e-04  1.0159e-05  34.9085 < 2.2e-16 ***
## EDUC         2.2132e-02  1.0874e-02   2.0353   0.04182 *
## MARRIED      2.6803e-01  1.8986e-03 141.1735 < 2.2e-16 ***
## RACE        -2.9581e-01  1.7075e-02 -17.3244 < 2.2e-16 ***
## ENOCENT     -1.4391e-02  1.0356e-02  -1.3896   0.16466
## ESOCENT     -2.5860e-01  1.8798e-02 -13.7564 < 2.2e-16 ***
## MIDATL      -4.5405e-02  6.5409e-03  -6.9416 3.878e-12 ***
## MT          -1.2972e-01  4.4710e-03 -29.0149 < 2.2e-16 ***
## NEWENG      -1.1130e-01  5.9155e-03 -18.8145 < 2.2e-16 ***
## SOATL       -1.7654e-01  1.2045e-02 -14.6575 < 2.2e-16 ***
## WNOCENT     -1.6041e-01  8.6845e-03 -18.4706 < 2.2e-16 ***
## WSOCENT     -1.4288e-01  1.0399e-02 -13.7401 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# heteroskedasticity adapted standard errors
# gather robust standard errors in a list
rob_se <- list(sqrt(diag(vcovHC(iv, type = "HC1"))))
# generate table
stargazer(iv,
 type = "text",
 digits = 3,
 se = rob_se)
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                LWKLYWGE
## ------------------------------------------------
## AGEQ                          0.0004***
##                               (0.00001)
##
## EDUC                           0.022**
##                                (0.011)
##
## MARRIED                       0.268***
##                                (0.002)
##
## RACE                          -0.296***
##                                (0.017)
##
## ENOCENT                        -0.014
##                                (0.010)
##
## ESOCENT                       -0.259***
##                                (0.019)
##
## MIDATL                        -0.045***
##                                (0.007)
##
## MT                            -0.130***
##                                (0.004)
##
## NEWENG                        -0.111***
##                                (0.006)
##
## SOATL                         -0.177***
##                                (0.012)
##
## WNOCENT                       -0.160***
##                                (0.009)
##
## WSOCENT                       -0.143***
##                                (0.010)
##
## Constant                      4.757***
##                                (0.134)
##
## ------------------------------------------------
## Observations                  1,063,634
## R2                              0.255
```

```
## Adjusted R2                      0.255
## Residual Std. Error    0.621 (df = 1063621)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Assumptions

As stated earlier, therer are two assumptions need to be fulfilled for IV to be valid instrument for variable: 1) exogeneity (uncorrelated with error U) and 2) relevance (correlated with X). One condition implicit in the exogeneity assumption is exclusion: IV should not directly affect Y . Exogeneity implies IV should not affect Y through an omitted variable. One way to argue that exogeneity condition holds true is if it is well randomized. The second condition must be that instruments must satisfy relevance. When an instrument is valid, an estimator become consistent. However, IV identifies the Local Average Treatment Effect (LATE): effect of x on compliers only.

Warning: Problem of Too Many Instruments

Having many instruments leads to having a large bias. As the number of instruments increases, F-statistic goes to zero and moves the coefficient towards the OLS coefficient. So we should refrain from using too many instruments and especially if they are weak instruments.

Check the Assumption of Relevance and Weak Instrument Problem

When instruments are weak, all the estimate coefficients in the first stage are zero or nearly zero. Weak instruments explain very little of the variation in the endogenous variable. we can test whether instruments are weak by testing for significance of identifying instruments in the first stage. We obtain the F-statistic and check if it is larger than 10. Weak instruments imply a small first stage F-statistic. For single instrument, we conduct t-test and F-test for the joint significance of the excluded instruments

One consequence of using instrument is that standard error will be large. Also, then the usual methods of inference are unreliable. The weaker is the instrument (low correlation between predictor and instrument), the smaller must endogeneity be in order for IV to be preferable.

```
# check instrument relevance for model (1)
mod_relevance1 <- lm(EDUC ~ QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT + MIDATL +
MT + NEWENG + SOATL + WNOCENT + WSOCENT, data = df)
linearHypothesis(mod_relevance1, "QOB = 0", vcov = vcovHC, type = "HC1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## QOB = 0
##
## Model 1: restricted model
## Model 2: EDUC ~ QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT + MIDATL +
##     MT + NEWENG + SOATL + WNOCENT + WSOCENT
##
## Note: Coefficient covariance matrix supplied.
##
##    Res.Df Df      F    Pr(>F)
## 1 1063622
## 2 1063621  1 337.89 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Always report the first stage. Ask two questions: 1) Does it make sense? 2) Do the coefficients have the right magnitude and sign?

Never use the weak instruments from the first stage regression.

Test Exogeneity

One major consequence of IV estimate's endogeneity is that its asymptotic bias is worse than OLS bias. A small correlation between the instrument and the error could cause a large bias if the instrument is weak. If we dont have exogeneity, then IV estimator will be biased and will have high SE.

If instrument is weak and is not exogenous, then the IV estimator can be very misleading. In this case, OLS is biased but we know the direction and sign of bias, then we may use the OLS estimato as the bound of the true value. For example, if we know the OLS estimator is biased upward, then we can think of OLS as the upper bound of the true value.

If instruments are overidentified, then it we can test for exogeneity exploiting the fact that if all the instruments are exogenous, then the estimates will be close to one another. Conduct the F-test for the null hypothesis that all instruments are jointly equal to zero, using the J=mF statistic with chi-squared distribution. If we have some instruments that are exogenous and others are endogenous, then J statistic will be large. Test by overidentification with the null hypothesis that all instruments are exogenous.

```
# compute the J-statistic
# unfortunately, this data has only one instrument, not two and we cannot test whether
this is true
iv_exo_test <- lm(residuals(iv) ~ QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT + MI
DATL + MT + NEWENG + SOATL + WNOCENT + WSOCENT, data = df)
linearHypothesis(iv_exo_test,"QOB = 0", test = "Chisq")
```

```
## Linear hypothesis test
##
## Hypothesis:
## QOB = 0
##
## Model 1: restricted model
## Model 2: residuals(iv) ~ QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT +
##     MIDATL + MT + NEWENG + SOATL + WNOCENT + WSOCENT
##
##    Res.Df    RSS Df Sum of Sq Chisq Pr(>Chisq)
## 1 1063622 410154
## 2 1063621 410154  1         0     0          1
```

Hausmann Test - OLS vs IV

Is IV necessary? Given the choice, we should always choose OLS because it is unbiased and efficient. Hausman test can be used to see whether we should choose OLS or IV. Hausman Test can test the consistency of OLS. Its null hypothesis is that both IV and OLS estimates are equal and covariance is zero; alternate hypothesis is that they are not equal, and covariance is not equal to zero (IV is consistent and OLS is not). Hausman test checks if the difference between IV and OLS is statistically different from zero. If we cannot reject the null, use the OLS estimator. If we reject the null, use the IV estimator.

```
summary(iv, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = LWKLYWGE ~ AGEQ + EDUC + MARRIED + RACE + ENOCENT +
##     ESOCENT + MIDATL + MT + NEWENG + SOATL + WNOCENT + WSOCENT |
##     QOB + AGEQ + MARRIED + RACE + ENOCENT + ESOCENT + MIDATL +
##         MT + NEWENG + SOATL + WNOCENT + WSOCENT, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.49865 -0.23903  0.05477  0.32527  5.50831
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.757e+00  1.342e-01  35.451  < 2e-16 ***
## AGEQ         3.546e-04  1.016e-05  34.909  < 2e-16 ***
## EDUC         2.213e-02  1.087e-02   2.035   0.0418 *
## MARRIED      2.680e-01  1.899e-03 141.174  < 2e-16 ***
## RACE        -2.958e-01  1.707e-02 -17.325  < 2e-16 ***
## ENOCENT     -1.439e-02  1.036e-02  -1.390   0.1647
## ESOCENT     -2.586e-01  1.880e-02 -13.756  < 2e-16 ***
## MIDATL      -4.540e-02  6.541e-03  -6.942 3.88e-12 ***
## MT          -1.297e-01  4.471e-03 -29.015  < 2e-16 ***
## NEWENG      -1.113e-01  5.915e-03 -18.815  < 2e-16 ***
## SOATL       -1.765e-01  1.204e-02 -14.658  < 2e-16 ***
## WNOCENT     -1.604e-01  8.684e-03 -18.471  < 2e-16 ***
## WSOCENT     -1.429e-01  1.040e-02 -13.740  < 2e-16 ***
##
## Diagnostic tests:
##                      df1     df2 statistic  p-value
## Weak instruments       1 1063621    337.90  < 2e-16 ***
## Wu-Hausman             1 1063620     12.61 0.000385 ***
## Sargan                 0      NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.621 on 1063621 degrees of freedom
## Multiple R-Squared: 0.2546,  Adjusted R-squared: 0.2546
## Wald test: 2.567e+04 on 12 and 1063621 DF,  p-value: < 2.2e-16
```

Weak instruments means that the instrument has a low correlation with the endogenous explanatory variable. This could result in a larger variance in the coefficient, and severe finite-sample bias. “The cure can be worse than the disease” (Bound, Jaeger, Baker, 1993/1995). See here for more details. From the help file for AER, it says it does an F-test on the first stage regression; I believe the null is that the instrument is weak. For the model you report, the null is rejected, so you can move forward with the assumption that the instrument is sufficiently strong.

Wu-Hausman tests that IV is just as consistent as OLS, and since OLS is more efficient, it would be preferable. The null here is that they are equally consistent; in this output, Wu-Hausman is significant at the p<0.1 level, so if you are OK with that confidence level, that would mean IV is consistent and OLS is not.

Sargan tests overidentification restrictions. The idea is that if you have more than one instrument per endogenous variable, the model is overidentified, and you have some excess information. All of the instruments must be valid for the inferences to be correct. So it tests that all exogenous instruments are in fact exogenous, and uncorrelated with the model residuals. If it is significant, it means that you don't have valid instruments (somewhere in there, as this is a global test). In this case, this isn't a concern. This can get more complex, and researchers have suggested doing further analysis

The fact that Hausman test rejects the null hypothesis suggests IV regression is preferred for consistency of estimates.

Limitations: Heterogeneous Treatment and LATE estimate

When treatments are heteogeneous to subjects, our estimate is LATE and it is on complies and always-takers only. It can make it difficult to generalize for different subpopulations - defiers and never-takers.

Internal and External Validity (Generalizability)

Assuming internal validity, can we extrapolate the results to other units (e.g. location, period)? Can we also assume external validity? There is usually a tradeoff between the two.

When sample is large, external validity increases.

Regression Discotinuity Design (RDD)

Regression Discontinuity is running a regression in a situation where you have a discontinuity at a threshold of variable and it results in a discrete treatment near the threshold. Units at a threshold are assumed to be very similar but they get very dissimilar, so this data design is similar to random experiment (quasi-experiment). The key feature in this design is to exploit the precise knowledge of the rules determining treatment so that we know where the cutoff is. The main identifying assumption is that, in a sufficiently near neighborhood around the discontinuity, treatment is as good as randomly assigned.

Identification Strategy

We suppose that the treatment x is assigned according to some rule z . The key identifying assumptions we need are: 1. z affects y smoothly if not for the policy x 2. individual has no perfect control over z

We can test these two assumptions through: 1. institutional knowledge 2. balance test 3. check smoothness of distribution (McCrary test)

Applications:

What kind of problems does RDD deal with? 1. effect of scholarship on students: exams and financial aid thresholds (Van Der Klaauw, 2002) 2. effect of class size on student performance: school class size (Angrist & Lavy, 1999) 3. effect of unions (DiNardo & Lee, 2004) 4. effect of air pollution (Chay & Greenstone, 2005)

Problem Setup

Our data is from Damon Clark's paper (2004). Traditionally, schools in the UK have been funded and managed by Local Education Authorities (LEA). In London, this would be a borough with rather little in the way of autonomy given to individual schools. But the 1988 Education Act allowed schools to opt out of LEA control and become funded by central, not local, government with much more autonomy - and this was called Grant-Maintained. Schools could become GM if a simmple majority of parents chose that option in a ballot. So if 51% of parents voted for GM status then school would become GM-school while if 49% voted for it, it would remain under LEA control. This is the basis of the regression continuity design. The paper contributes to the debate about how public institutions like schools or hospitals should be run: should they be given a budget and left to spend it how they want or should they be more tightly controlled? In the case of GM schools, becoming GM resulted not just in more autonomy but also more resources, which were justified as the school now had to deal with some issues that had previously been handled by the LEA and by some people perceived to be bribes as the government wanted to encourage the growth of GM schools. Thus the change to GM resulted in both more autonomy and possibly more resources.

```
df <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/damonclark.dta')
# it has three variables: pass rate in the base year, two years after, and vote percen
tage
# each unit of observation is a school
summary(df)
```

```
##    passrate0       passrate2          vote
##  Min.   : 1.00   Min.   : 1.00   Min.   : 3.978
##  1st Qu.:27.00   1st Qu.:31.00   1st Qu.:44.520
##  Median :40.00   Median :42.00   Median :68.303
##  Mean   :38.94   Mean   :42.19   Mean   :62.988
##  3rd Qu.:50.00   3rd Qu.:53.00   3rd Qu.:82.856
##  Max.   :97.00   Max.   :97.00   Max.   :98.376
```

Variables

passrate0 is the pass rate of pupils in the school in the year immediately prior to the vote passrate2 is the pass rate of pupils in the school two years after the vote vote is the percentage vote in favor of the GM status

1. What is the outcome of interest? Change in Pass Rate
2. What is the running variable, X? Vote percentage on whether to be autonomous or not

3. What is the treatment variable D and how is it determined by X? D is a discontinuous function of X from a smooth and flexible function f(X) that controls the counterfactual. RD captures a causal effect by distinguish the treatment D from variable X.

```
# generate a dummy variable for a winning vote and one for a losing vote in the GM ele
ction where 50% is the critical threshold
# generate a margin variable as the difference from threshold of victory in the vote
# generate interactions of win with margin
df <-
  df %>%
  mutate(win = ifelse(vote > 50, 1, 0),
         lose = ifelse(vote < 50, 1, 0),
         margin = vote-50,
         change_pass=passrate2-passrate0) %>%
  mutate(win_int = win*margin^2,
         lose_int = lose*margin^2)
glimpse(df)
```

```
## Rows: 662
## Columns: 9
## $ passrate0   <dbl> 54, 60, 50, 40, 40, 43, 41, 33, 80, 84, 20, 21, 21, 34,...
## $ passrate2   <dbl> 50, 60, 53, 53, 33, 49, 58, 48, 80, 86, 27, 24, 32, 56,...
## $ vote        <dbl> 52.75229, 27.21519, 37.98521, 65.75964, 84.71910, 70.63...
## $ win         <dbl> 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1...
## $ lose        <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0...
## $ margin      <dbl> 2.752293, -22.784811, -12.014786, 15.759636, 34.719101,...
## $ change_pass <dbl> -4, 0, 3, 13, -7, 6, 17, 15, 0, 2, 7, 3, 11, 22, -5, 13...
## $ win_int     <dbl> 7.575115, 0.000000, 0.000000, 248.366124, 1205.415971, ...
## $ lose_int    <dbl> 0.0000, 519.1476, 144.3551, 0.0000, 0.0000, 0.0000, 0.0...
```

The underlying assumption in this design is that the schools who barely passed the ballot and the schools who did not pass are very similar that we can use one group of the schools as clones of the other group. This is to say the potential outcomes are continuous at vote=50. Since winning the ballot becomes GM so we can say this is a sharp RD design.

Type of RDD

1. Sharp: treatment is a deterministic, discontinuous function of a covariate X. This design is based on selection on observables assumptions: Once we know X, we know D, which is correlated with X. We estimate causal effect by distinguishing D from f, density estimation function.

- key identifying assumption: conditional mean of Y on x is continuous on X, which implies all other unobserved determinants of Y are continuously related to the running variable of X. This allows us to use average outcomes of units below the cutoff as a valid counterfactual for units right above the cutoff.
- two ways of approximating f(x): 1) nonparametric kernel method 2) kth order polynomial
- account for the effect of treatment by interacting treatment D with running variable X

- it is very important that the polynomials provide an adequate continuous representation of conditional mean of Y on X; otherwise what looks like a jump may simply be a nonlinearity that the polynomials have not accounted for

2. Fuzzy: crossing the threshold is not the only cause for receipt of the treatment, treatment is not a deterministic function of running variable. Instead, it is useful think of threshold where the probability of receiiving the treatment jumps or due to unobservable variables. when treatment variable is numeric or has many categories. Fuzzy RDD exploits discontinuities in the probability of treatment. It uses an instrument IV type of setup, but fuzzy RDD will not be used in this example.

With fuzzy RDD, the avreage change in y around the threshold understate causeal effect. Comparision assumes all observations were treated, but this isn't true; if all observations had been treated, observed change in y would be even larger; we will need to rescale based on change in probability.

Data Preprocessing

Clark restricts his sample to those schools with votes in favor of GM status between 15% and 85% because they are different in terms of covariates. They have particular motives for seeking GM status and exceptionally high baseline pass rates, whilst many schools in the tails of the voting distribution were threatened with closure.

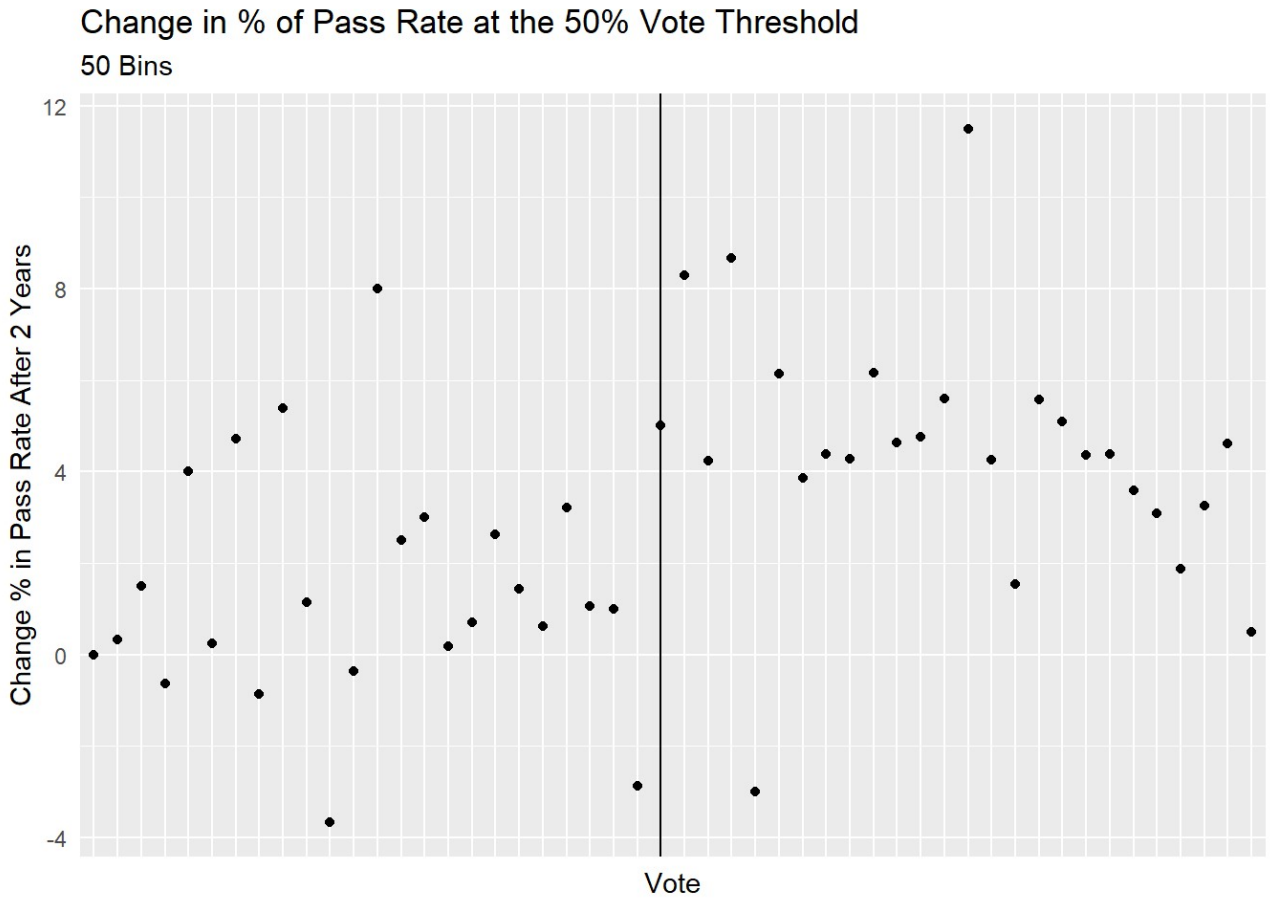
```
df <-
  df %>%
  filter(vote<=85 & vote >= 15)
```

Visualize Data

We can use visualiztion to check assumptions of the data.

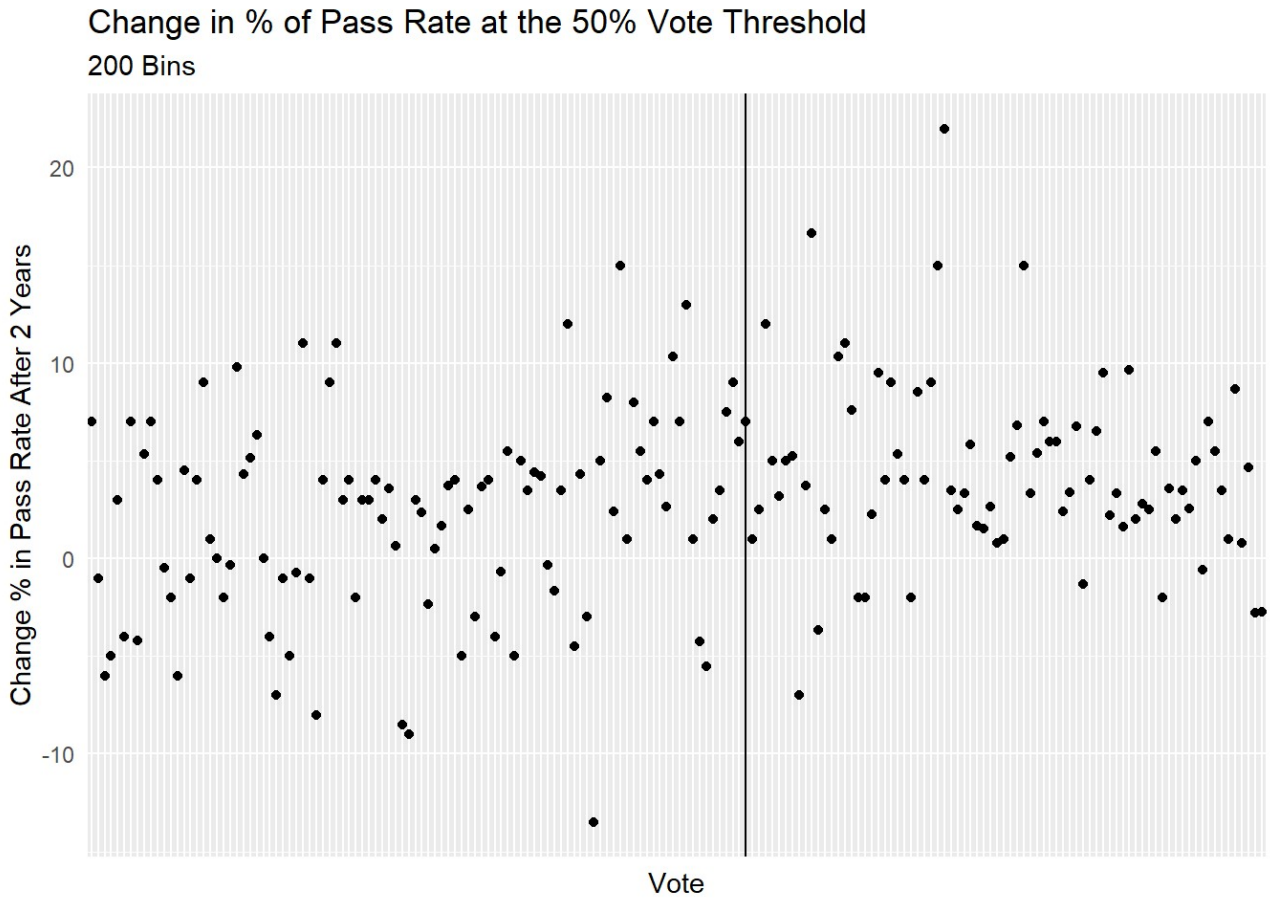
The first visualization is outcome by running variable X (aka forcing variable) and this is the stanadrd visualization showing the discontinuity in the outcome variable. We construct bins and average the outcome within bins on both sides of the cutoff. We look at the different bin sizes when constructing these graphs, and plot the forcing variable on the horizontal axis and the average of Y for each bin on the vertical axis. Then, we inspect whether there is a discontinuity at the threshold and whether there are other obvious unexpecte discontinuities.

```
df %>%
  mutate(bins = cut(df$vote, breaks=50)) %>%
  ggplot(aes(x=bins, y=change_pass)) +
  stat_summary(fun='mean', geom='point') +
  geom_vline(aes(xintercept=which(levels(bins)=='(48.8,50.2]'))) +
  labs(title='Change in % of Pass Rate at the 50% Vote Threshold',
       subtitle='50 Bins',
       x='Vote',
       y='Change % in Pass Rate After 2 Years') +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank())
```



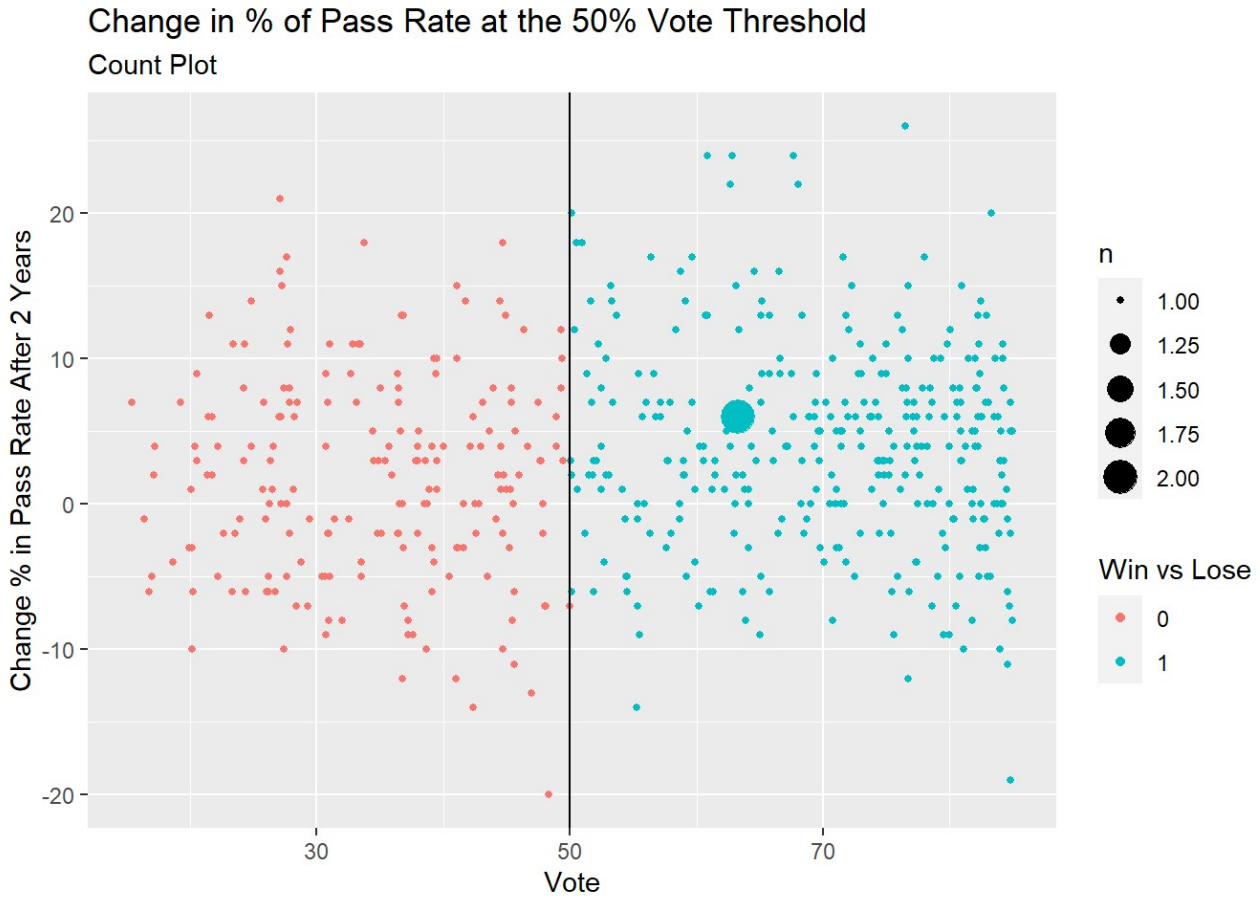
We need to train various bin sizes.

```
df %>%
  mutate(bins = cut(df$vote, breaks=200)) %>%
  ggplot(aes(x=bins, y=change_pass)) +
  stat_summary(fun='mean', geom='point') +
  geom_vline(aes(xintercept=which(levels(bins)=='(49.8,50.2]'))) +
  labs(title='Change in % of Pass Rate at the 50% Vote Threshold',
       subtitle='200 Bins',
       x='Vote',
       y='Change % in Pass Rate After 2 Years') +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank())
```



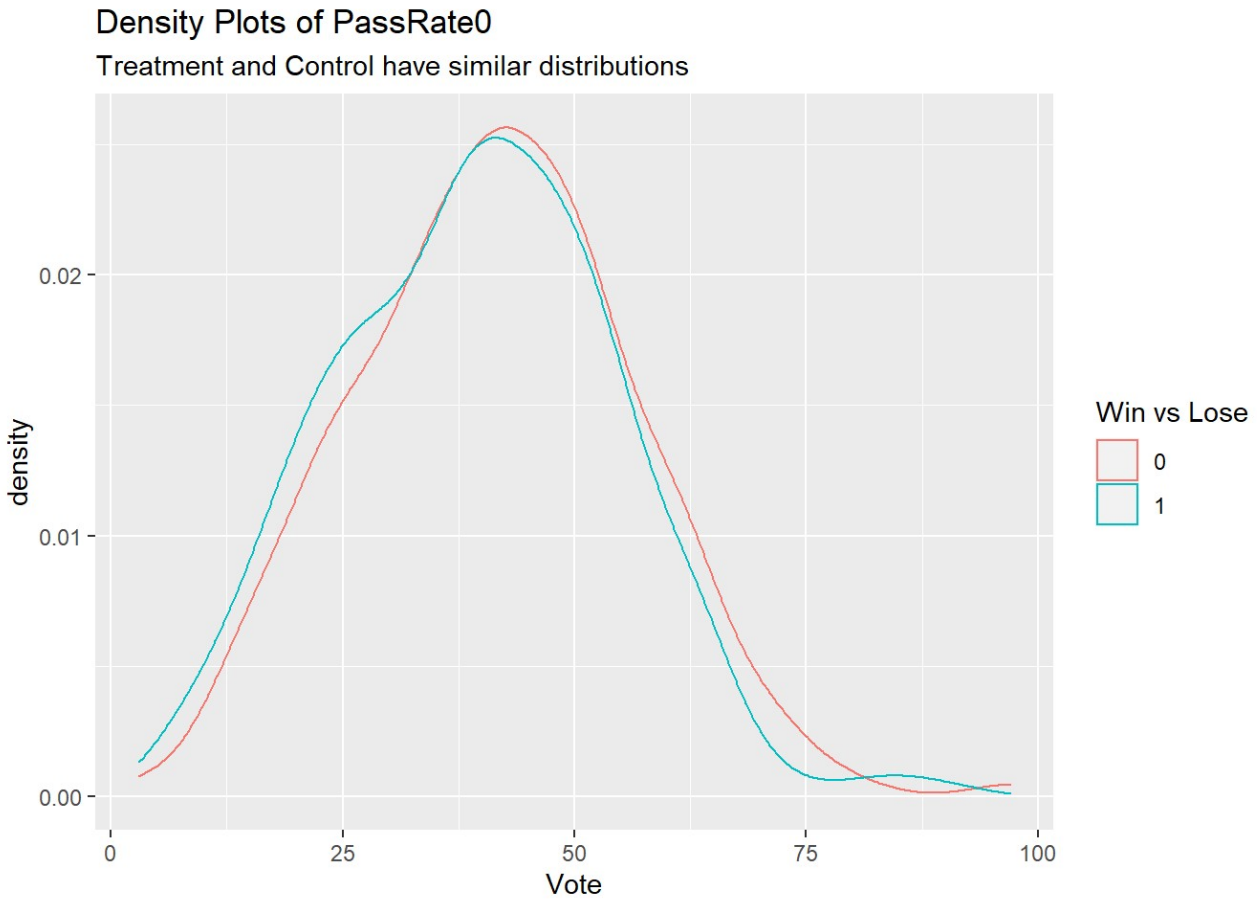
The second visualization is countplot. If we have abnormally large portion of people around the cutoff, it is quite possible that you do not have random assignment. In this case, such anomaly does not exist.

```
df %>%
  ggplot(aes(x=vote,y=change_pass)) +
  geom_count(aes(color=factor(win))) +
  geom_vline(xintercept=50) +
  labs(title='Change in % of Pass Rate at the 50% Vote Threshold',
       subtitle='Count Plot',
       x='Vote',
       y='Change % in Pass Rate After 2 Years',
       color='Win vs Lose')
```



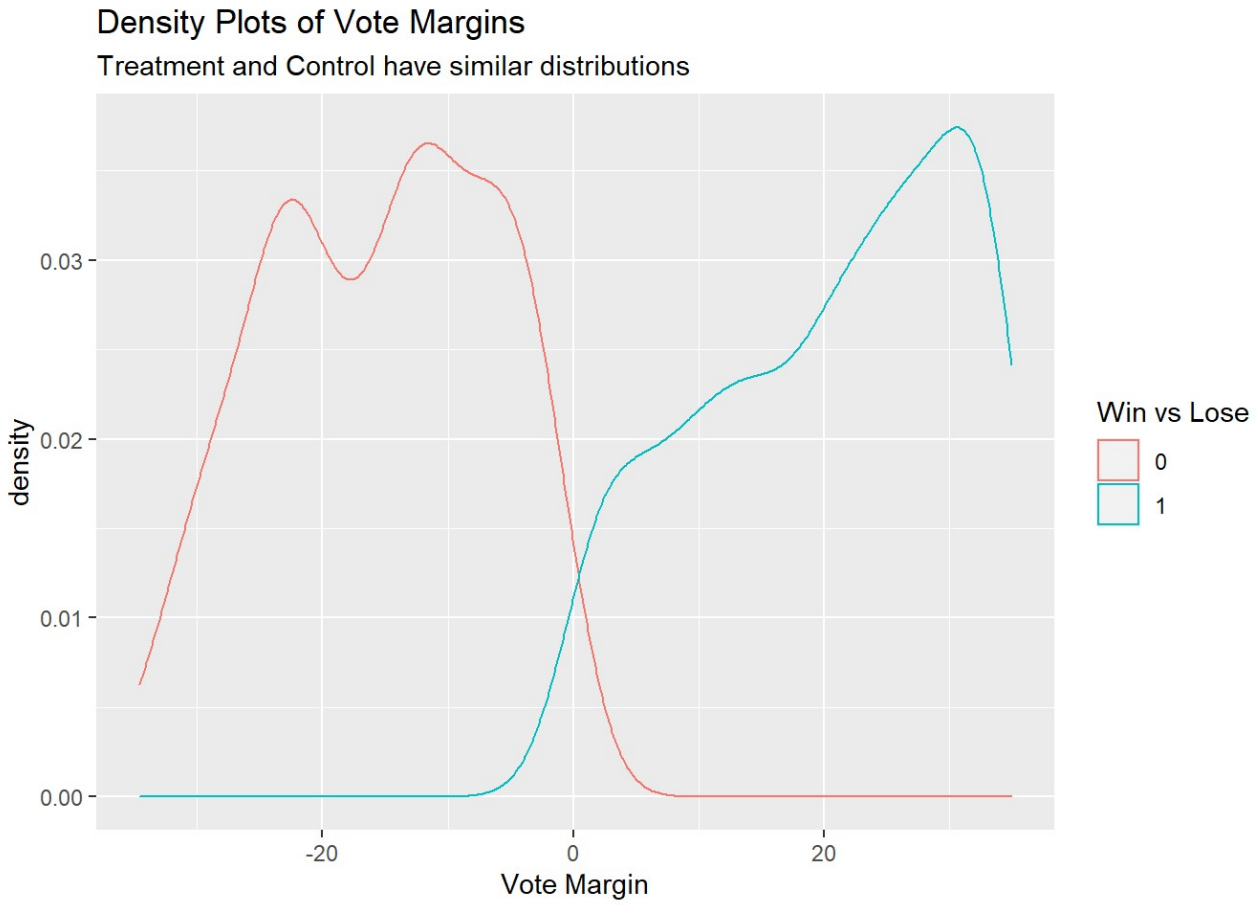
We want the treatment and control groups to have similar distributions. First, we need to look at the density of the outcome variable for both treatment and control groups.

```
df %>%
  ggplot(aes(x=passrate0)) +
  geom_density(aes(color=factor(win))) +
  labs(x='Vote',
       title='Density Plots of PassRate0',
       subtitle='Treatment and Control have similar distributions',
       color='Win vs Lose')
```



We also need to look at the density of the running (forcing) variable. We should plot the number of observations in each bin. It allows us to investigate whether there is a discontinuity in the distribution of the forcing variable at the threshold. This would suggest that people can manipulate the forcing variable around the threshold, and suggests an indrect test of the identifying assumption that each individual has imprecise control over the assignment variable.

```
df %>%
  ggplot(aes(x=margin)) +
  geom_density(aes(color=factor(win))) +
  labs(x='Vote Margin',
       title='Density Plots of Vote Margins',
       subtitle='Treatment and Control have similar distributions',
       color='Win vs Lose')
```



Fitting Model

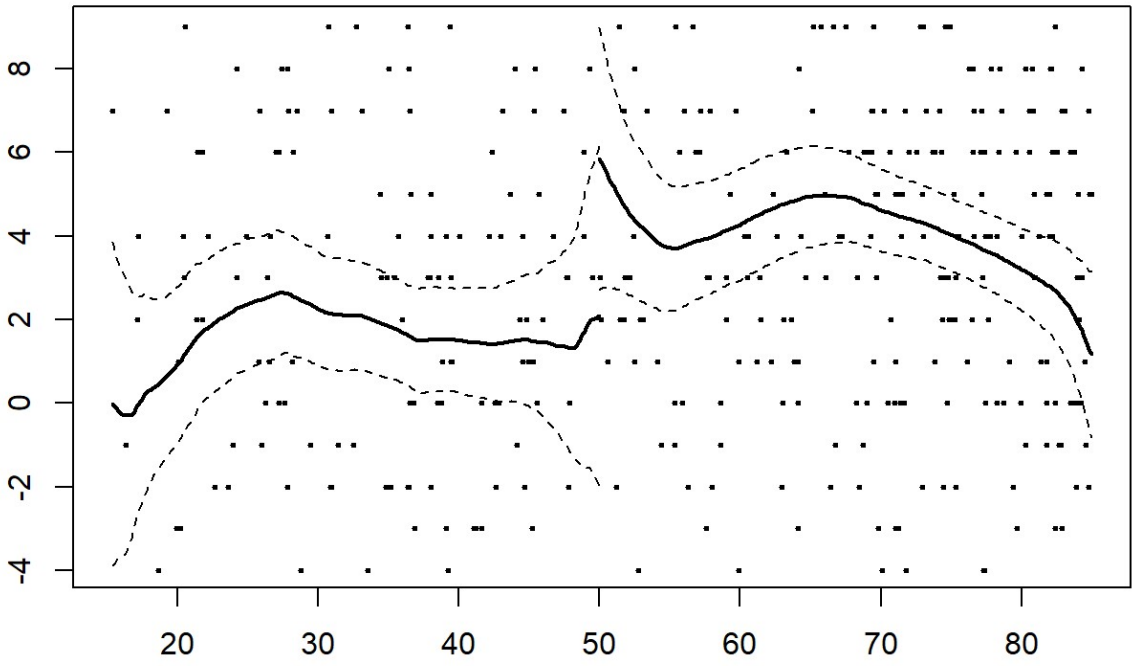
I-K Optimal Bandwith Local Linear Regression

Run local linear regression on various bandwidths optimized using the Imbens-Kalyanaraman method, and then estimated with half that bandwidth, and finally twice that bandwidth.

```
rdmod <- RDestimate(change_pass ~ vote, data=df, cutpoint=50)
```

and plot the regression:

```
plot(rdmod, main='Regression Discontinuity',xlab='Vote',ylab='Change in PassRate')
```



If we have a smaller bandwidth, standard error gets larger. This is the well known bias-variane trade-off. If we have a smaller bandwidth, it is more likely that the populations below and above the threshold are similar so that the coefficient on win has smaller bias, but having smaller bandwidth means we have fewer observations, which results in larger standard error.

The increase in standard errors shows in the fitted curves with smaller bandwidths showing larger fluctuations around the boundaries, indicating larger variance. We cannot assess the bias from the plot. With small sample size, we would expect the bandwidth to be large to have a reasonable fit of the boundaries. We need watch out for non-linearity that can arise from outliers. We would expect functional form to be similar at the boundary, so difference such as linearity vs non-linearity should be examined closely.

Polynomial Regression

```
# instead of rdd package, use lm() function
rdmod1 <- lm(change_pass ~ win, data=df)
rdmod2 <- lm(change_pass ~ win + margin, data=df)
rdmod3 <- lm(change_pass ~ win + margin*win + margin*lose, data=df)
rdmod4 <- lm(change_pass~win + margin + win*margin + lose*margin + win*margin^2 + lose
*margin^2, data=df)
stargazer(rdmod1, rdmod2, rdmod3, rdmod4, type='text', title='RDD Models with Differen
t Sets of Variables', style='aer')
```

```
##
## RDD Models with Different Sets of Variables
## ================================================================================
============================
##                                                          change_pass
##                              (1)                  (2)                  (3)
(4)
## --------------------------------------------------------------------------------
----------------------------
## win                        2.169***             4.052***             12.233*
12.233*
##                            (0.634)              (1.282)              (7.061)
(7.061)
##
## margin                                          -0.053*               -0.019
-0.019
##                                                 (0.031)              (0.056)
(0.056)
##
## lose                                                                  8.510
8.510
##                                                                      (7.081)
(7.081)
##
## win:margin                                                            -0.046
-0.046
##                                                                      (0.067)
(0.067)
##
## margin:lose
##
##
## Constant                   1.755***              0.943                -7.000
-7.000
##                            (0.501)              (0.694)              (7.009)
(7.009)
##
## Observations                 524                  524                  524
524
## R2                           0.022                0.027                0.031
0.031
## Adjusted R2                  0.020                0.024                0.023
0.023
## Residual Std. Error    7.020 (df = 522)       7.007 (df = 521)       7.009 (df = 5
19)       7.009 (df = 519)
## F Statistic         11.710*** (df = 1; 522) 7.301*** (df = 2; 521) 4.087*** (df =
4; 519) 4.087*** (df = 4; 519)
## --------------------------------------------------------------------------------
```

```
----------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

Different Subsets of Data

If we have data on outcome variable prior to the treatment, then we can use this data to check whether the populations below and above the threshold are similar. We want the two populations to be clones of each other and so we want them to be similar not only interms of the unobservables but also the observables.

```
rdmod5 <- lm(change_pass ~ win + margin:win + margin:lose, data=df, subset=(vote<=85 &
vote >= 15))
rdmod6 <- lm(change_pass ~ win + margin:win + margin:lose, data=df, subset=(vote<=70 &
vote >= 30))
rdmod7 <- lm(change_pass ~ win + margin:win + margin:lose, data=df, subset=(vote<=55 &
vote >= 45))
stargazer(rdmod5, rdmod6, rdmod7, title='Impact of GM Status on Pass Rates of School:
Two Years after Base Year',style='aer', type='text')
```

```
##
## Impact of GM Status on Pass Rates of School: Two Years after Base Year
## ================================================================================
===
##                                              change_pass
##                              (1)                  (2)                  (3)
## --------------------------------------------------------------------------------
---
## win                        3.894***              2.753                6.545*
##                            (1.314)              (1.823)              (3.850)
##
## win:margin                 -0.064*               0.060                -1.575
##                            (0.038)              (0.103)              (0.960)
##
## margin:lose                 -0.027               -0.015               0.535
##                            (0.056)              (0.119)              (0.938)
##
## Constant                    1.339                1.303                1.849
##                            (0.995)              (1.349)              (2.932)
##
## Observations                 524                  273                  62
## R2                          0.028                0.048                0.127
## Adjusted R2                 0.022                0.037                0.081
## Residual Std. Error    7.012 (df = 520)      7.335 (df = 269)     7.629 (df = 58)
## F Statistic         4.964*** (df = 3; 520) 4.493*** (df = 3; 269) 2.800** (df = 3;
58)
## --------------------------------------------------------------------------------
---
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

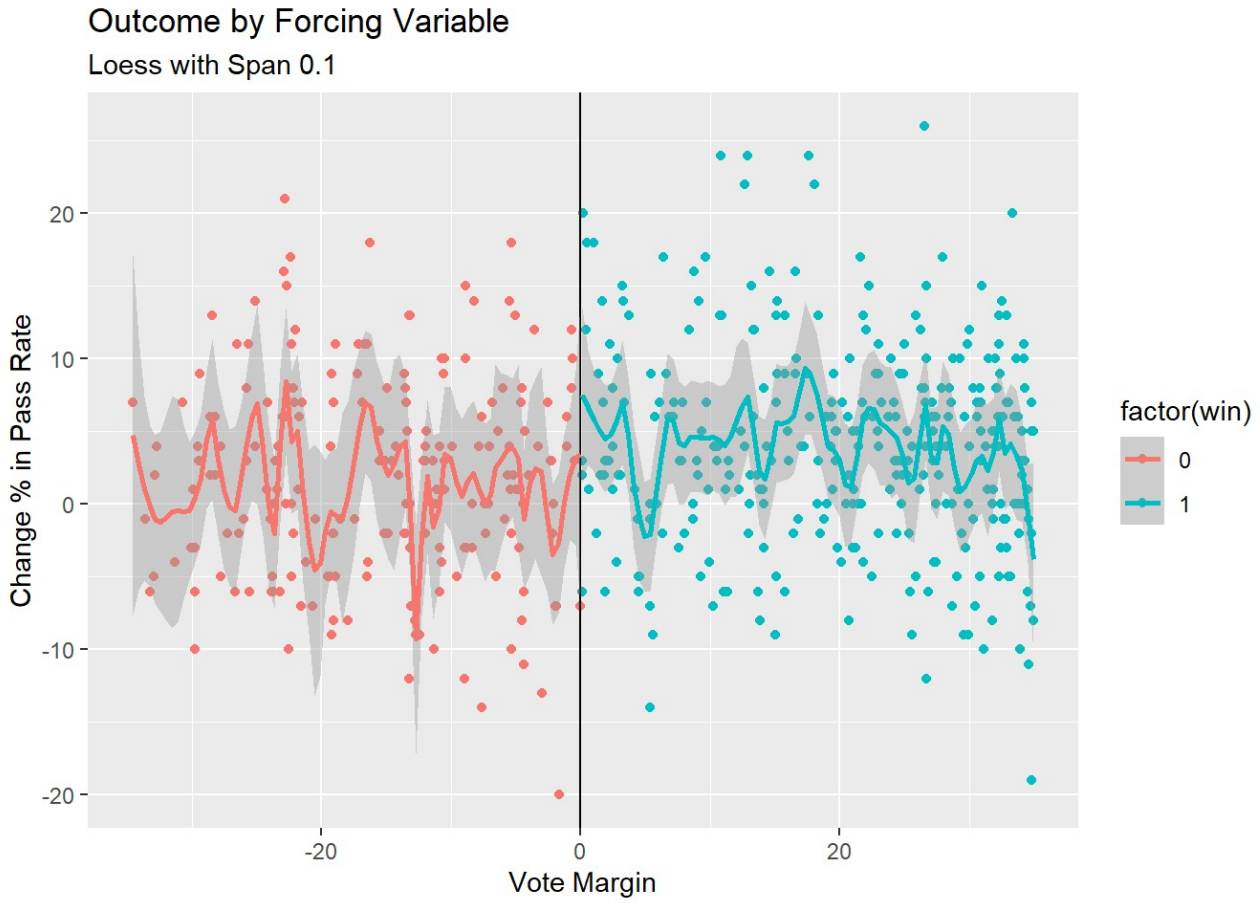
If the RDD is to be valid, we don't want the coefficient to be significiant at the 5% level. The regression result shows that schools who won the ballot have on average lower pre-ballot student passing rate than those who did not win the ballot.

Different Bandwidths on Smoothing

Results need to be robust to different bandwidths, and need to report results for both estimation types (polynomial in X and local inear regression). We also need to show that including higher order polynomials does not substantially affect our findings, and our results are not affected if we vary the window around the cutoff (stanard errors may go up but hopefully the point estimate doesnt change).

```
par(mfrow=c(2,2))
g <-
  ggplot(df, aes(x=margin, y=change_pass)) +
  geom_point(aes(color=factor(win))) +
  geom_vline(xintercept=0) +
  labs(x='Vote Margin',
       y='Change % in Pass Rate',
       title='Outcome by Forcing Variable')

g + geom_smooth(method='loess', span=0.1, aes(color=factor(win))) +
  labs(subtitle='Loess with Span 0.1')
```

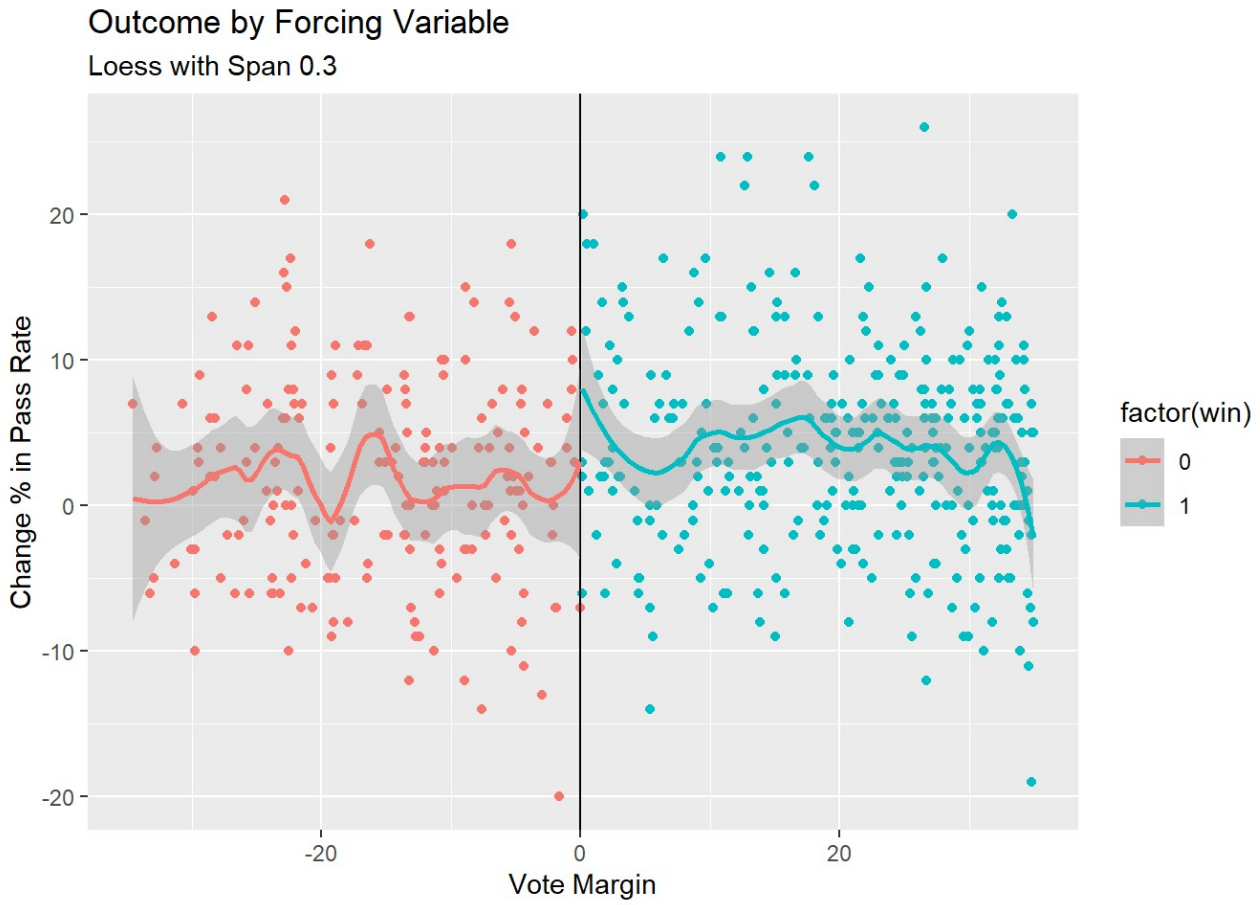
```
## `geom_smooth()` using formula 'y ~ x'
```



Increase bandwith to 0.3.

```
g + geom_smooth(method='loess', span=0.3, aes(color=factor(win))) +
  labs(subtitle='Loess with Span 0.3')
```

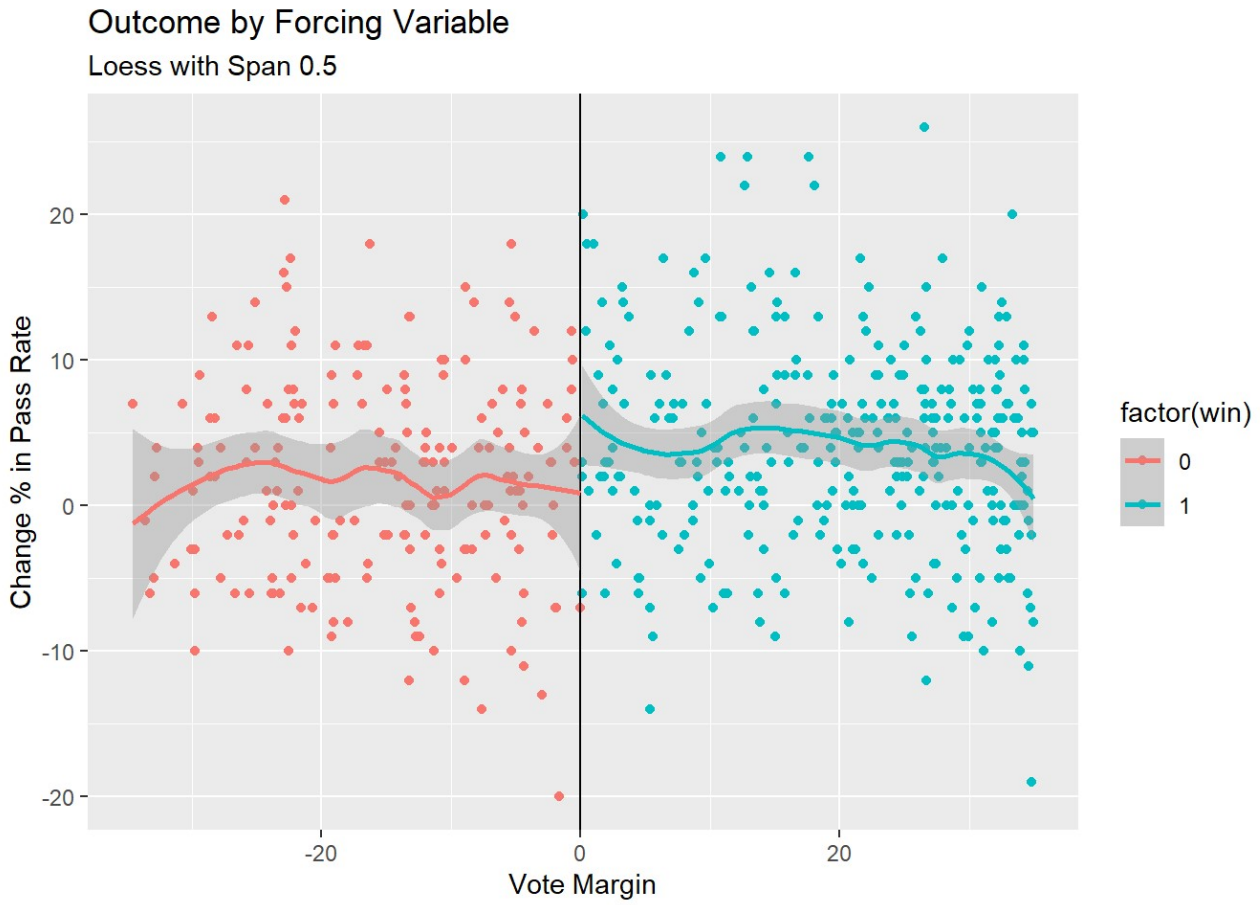
```
## `geom_smooth()` using formula 'y ~ x'
```



Increase bandwith to 0.5.

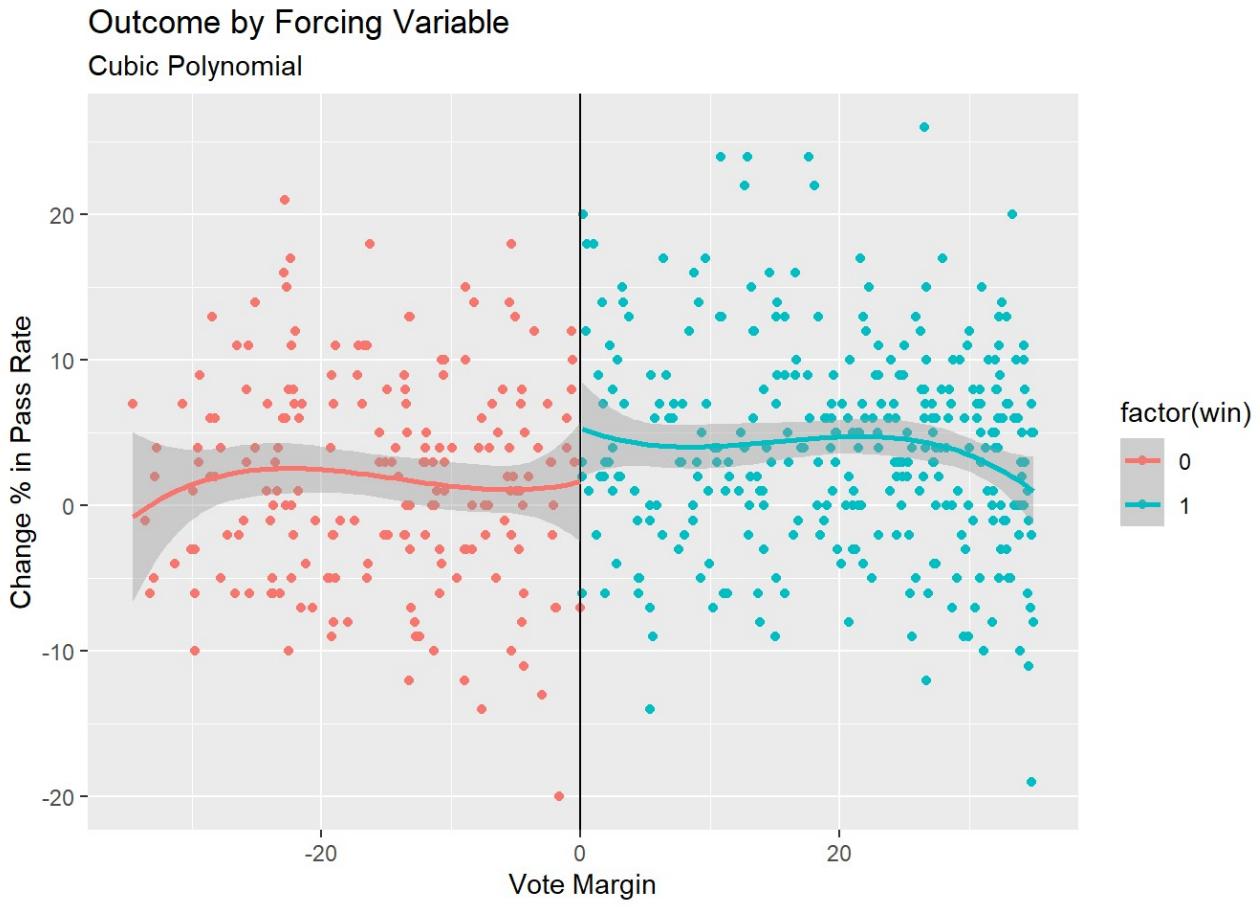
```
g + geom_smooth(method='loess', span=0.5, aes(color=factor(win))) +
  labs(subtitle='Loess with Span 0.5')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Now use linear regression.

```
g + geom_smooth(method = 'lm', formula = y ~ splines::bs(x, 3), aes(color=factor(wi
n))) +
  labs(subtitle='Cubic Polynomial')
```



The results are robust to various techniques, and we can claim that the autonomous schools don't have statistically significant effect on the pass rate of schools.

Interpretation

β is the LATE of the treatment on the group around the discontinuity.

Fuzzy RDD

Units with values above a certain threshold value of the underlying variable are more likely to be treated than those below.

Instrumental Variable

The discontinuity becomes an IV for treatment status. Create interaction terms with instrument for increased complexity. Use 2SLS to estimate the coefficient.

Problem

Boundary problem from kernel method is that it implies a systematic bias with the method if f(x) is upwards or downwards sloping. Local linear regression solves this problem.

Optimizing Bandwidth

There are three main methods to choose bandwidth: 1) cross validation 2) optimal bandwith by Imbens and Kalyanaraman 3) robust data-driven inference.

Difference in Differences (DID)

If we do not have random assignment into treatment (i.e. RCT) to balance unobservables, a valid IV, credible RD design, or cannot use MLR to control for all relevant observables, then we can use DiD. Differnece in Differences method relies on selection on unobservables in the sense that sometimes it is more reasonable to assume that changes in variables (rather than levels) move in parallel.

We pool cross sections of individuals/geographical units/firms on whom some policy is enacted (job training, electoral rule, new regulation). The group exposed to the policy is called treatment group and the group not exposed is control group. Using DiD implies that we compare the difference in outcomes in the treatment and control groups before and after treatment. The data requiremenet is only that we have cross-section ata for the treatment and control groups before and after the policy.

Identifying Strategy

A DiD estimate is an unbiased estimator of the causal effect if the average change in the outcome variable would have been the same for the two groups without the treatment (parallel trend assumption). In other words, in the absence of the treatment, control and treatment group should follow parallel trends. So we will need to check for parallel trends before the treatment.

DiD estimator is superior to Before-After estimator in one important way. BA estimator compares the outcome for the treated group before and after treatment, but this is a causal effect only if there are no other average differences in unobservables before and after policy change. It ignores time and age effects, and is a very strong assumption. Instead, DiD uses control group to difference out other factors and isolate the policy effect, and recovers treatment on the treated.

Applications

1. Frieberg (1998) Divorce Law on Divorce Rate
2. Card (1992) Minimum Wage
3. Autor (2003) Employment Protection
4. Abadie and Gardeazabal (2003) Terrorism and Growth

Data

Our data examines the effect of workers' compensation on time out of work. It compares individuals injured before and after increases in the maximum weekly benefit amount. The increases examind in KY and MI raised the benefit amount for high earning individuals by approximately 50% while low earning individuals who were unaffected by the benefit maximum, did not experience a change in their incentives.

```
df <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/INJURY.DTA')
glimpse(df)
```

```
## Rows: 7,150
## Columns: 30
## $ durat    <dbl> 1, 1, 84, 4, 1, 1, 7, 2, 175, 60, 29, 30, 100, 4, 2, 1, 1,...
## $ afchnge  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ highearn <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ male     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1...
## $ married  <dbl> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1...
## $ hosp     <dbl> 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1...
## $ indust   <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1...
## $ injtype  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ age      <dbl> 26, 31, 37, 31, 23, 34, 35, 45, 41, 33, 35, 25, 39, 27, 24...
## $ prewage  <dbl> 404.9500, 643.8250, 398.1250, 527.8000, 528.9375, 614.2500...
## $ totmed   <dbl> 1187.57324, 361.07855, 8963.65723, 1099.64832, 372.80188, ...
## $ injdes   <dbl> 1010, 1404, 1032, 1940, 1940, 1425, 1110, 1207, 1425, 1010...
## $ benefit  <dbl> 246.8375, 246.8375, 246.8375, 246.8375, 211.5750, 176.3125...
## $ ky       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ mi       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ldurat   <dbl> 0.0000000, 0.0000000, 4.4308167, 1.3862944, 0.0000000, 0.0...
## $ afhigh   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ lprewage <dbl> 6.003764, 6.467427, 5.986766, 6.268717, 6.270870, 6.420402...
## $ lage     <dbl> 3.258096, 3.433987, 3.610918, 3.433987, 3.135494, 3.526361...
## $ ltotmed  <dbl> 7.079667, 5.889095, 9.100934, 7.002746, 5.921047, 5.351953...
## $ head     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ neck     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ upextr   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ trunk    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ lowback  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ lowextr  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ occdis   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ manuf    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ construc <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ highlpre <dbl> 6.003764, 6.467427, 5.986766, 6.268717, 6.270870, 6.420402...
```

Stuy Design Limitations

DiD's important assumption is that there is no other interaction between time and treatment group except for the treatment we study. Thus changes in unobservables over time affect both groups in the same way. Fixed group affects capture unmeasured differences between treated and non-treated.

1. Targeting differences: we want to make sure in the data that treatment wasn't done due to pre-existing differences in outcome.

2. Functional form dependence: When average levels of the outcome are very different for control and treatments before the policy change, the magnitude or even sign of the DiD effect is very sensitive to funcitonal form.
3. Long term response versus reliability tradeoff: DiD estiates are more reliable when we compare outcomes just before and just after the policy change because the parallel trend is more likely to hold over a short time window.
4. Inference: if observations in the control and treatment group tend to move together and are correlated, then there may be a random effect at the time or group level, so we actually have less information.

Fit Models and Check Parallel Trend Assumption

I fit Before-After estimator for the treatment group on the subpopulation of highe earners in Kentucky, cross-sectional models, and finally DiD model.

```
# Get the Before-After estimator for the treatment group.
BA_treatment_mod <- lm(ldurat~afchnge, data=df, subset=(highearn==1 & ky==1))
# Get the Before-After estimator for the control group
BA_control_mod <- lm(ldurat~afchnge, data=df, subset=(highearn==0 & ky==1))
# Get the cross-section estimator on the period after treatment
CS_before_mod <- lm(ldurat~highearn, data=df, subset=(afchnge==1 & ky==1))
# Get the cross-section estimator on the period before treatment
CS_after_mod <- lm(ldurat~highearn, data=df, subset=(afchnge==0 & ky==1))
# Get the complete DiD model
didmod <- lm(ldurat~highearn*afchnge, data=df, subset=(ky==1))
# Summarize the results
stargazer(BA_treatment_mod, BA_control_mod, CS_before_mod, CS_after_mod, didmod, type
='text', title='Effect of Raising Benefits on Injury Duration', style='aer')
```

```
##
## Effect of Raising Benefits on Injury Duration
## ================================================================================
================================================
##                                                                          ldurat
##                             (1)                  (2)                  (3)
(4)                   (5)
## --------------------------------------------------------------------------------
------------------------------------------------
## afchnge                    0.198***              0.008
0.008
##                            (0.053)              (0.044)
(0.045)
##
## highearn:afchnge
0.191***
##
(0.069)
##
## highearn                                                                0.447***
0.256***              0.256***
##                                                                         (0.050)
(0.047)               (0.047)
##
## Constant                   1.382***             1.126***             1.133***
1.126***              1.126***
##                            (0.037)              (0.030)              (0.033)
(0.030)               (0.031)
##
## Observations                2,394                3,232                2,688
2,938                 5,626
## R2                         0.006                0.000                0.029
0.010                 0.021
## Adjusted R2                0.005                -0.000               0.029
0.010                 0.020
## Residual Std. Error    1.299 (df = 2392)      1.247 (df = 3230)      1.284 (df = 26
86)        1.255 (df = 2936)        1.269 (df = 5622)
## F Statistic         13.939*** (df = 1; 2392) 0.030 (df = 1; 3230) 79.962*** (df =
1; 2686) 29.863*** (df = 1; 2936) 39.540*** (df = 3; 5622)
## --------------------------------------------------------------------------------
------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

The change in the policy over time induced around 19.8% increase in the duration on average for high earners in Kentucky. The identifying assumption requires the error term to be uncorrelted with afchnge in the treatment group in order for the causal interpretation. In particular it means that there should be no time trend. This may not be true if the economy may have gotten better over time so that it is easier to get a job before than after they recover.

For the BA estimator on control group. we see the coefficient on afchnge is close to zero. This does not raise doubts about the assumption of no time trend. We can here compute DiD estimator by differencing the coefficients for BA-estimators for treatment and control group.

The regression equation for the Cross-Section estimator after the policy change is run for the observations after the benefit is raised in Kentucky. As shown in the table, high-earners have around 44.7% larger duration than the low earners on average, after the policy change. In order to interpret this as the causal effect of raising benefits, we need to asssume that there is no difference between the high-earners and the low-earners in their injury duration before the policy change. However, this may not hold since high-earners may have more stable job than low earners. The high-earners are secured to return to their previous job while the low-earners may have to find a new job if they are out too long, which induces the low-earners to return quickly.

For the CS estimator before the policy change, high earners have around 25.6% larger duration than the low-earners on average. This raises question on assumption on the CS estimator above. We can here compute DiD estimator by taking the difference between the coefficients after and before groups.

Finally, we have the complete model and the coefficient of 0.191. The main identifying assumption is that the amount of selection bias caused by unobservables stay the same over time except for the unobservables that affect both groups in the same way. In other words, we require that the potential outcomes of high earners and low earners have a common time trend. This means that, if the policy was not implemented, the high earners and the low earners would have shown the same rate of changes in the injury furation over time.

Because of selection bias CS estimator is not credible. We have little evidence of time trend in the data but the complete model didmod remains valid when there was actually time trend, so it is the most credible.

Feature Selection on the Complete Model

```
didmod1 <- lm(ldurat ~ afchnge + highearn + afhigh, data=df, subset=(ky==1))
didmod2 <- lm(ldurat~highearn + afhigh, data=df, subset=(ky==1))
didmod3 <- lm(ldurat~afchnge + afhigh, data=df, subset=(ky==1))
stargazer(didmod1, didmod2, didmod3, type='text', style='aer')
```

```
##
## ================================================================================
===========
##                                                 ldurat
##                              (1)                 (2)                    (3)
## --------------------------------------------------------------------------------
-----------
## afchnge                     0.008                                      -0.10
0**
##                            (0.045)                                    (0.04
0)
##
## highearn                   0.256***            0.253***
##                            (0.047)             (0.042)
##
## afhigh                     0.191***            0.198***               0.447
***
##                            (0.069)             (0.052)                (0.05
0)
##
## Constant                   1.126***            1.129***               1.233
***
##                            (0.031)             (0.022)                (0.02
3)
##
## Observations                5,626               5,626                  5,62
6
## R2                          0.021               0.021                  0.01
6
## Adjusted R2                 0.020               0.020                  0.01
5
## Residual Std. Error    1.269 (df = 5622)       1.269 (df = 5623)      1.272 (df
= 5623)
## F Statistic         39.540*** (df = 3; 5622) 59.305*** (df = 2; 5623) 44.476*** (df
= 2; 5623)
## --------------------------------------------------------------------------------
-----------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

It is not surprising that the coefficient on afhigh is similar for both columns because there is little time-trend. Now the coefficient on the interaction term in the column 3 is much larger because there is a large difference in the baseline injury duration between the high-earners and the low-earners.

Add controls

```
didmod_ky <- lm(ldurat~highearn*afchnge+male+married, data=df, subset=(ky==1))
didmod_mi <- lm(ldurat~highearn*afchnge+male+married, data=df, subset=(mi==1))
stargazer(didmod_ky, didmod_mi, type='text', title='Effect of Raising Benefits on Inju
ry Duration in Kentucky and Michigan', style='aer')
```

```
##
## Effect of Raising Benefits on Injury Duration in Kentucky and Michigan
## ================================================================
##                                        ldurat
##                             (1)                    (2)
## ----------------------------------------------------------------
## highearn                   0.226***                0.212*
##                            (0.052)                 (0.111)
##
## afchnge                     0.012                   0.096
##                            (0.045)                 (0.086)
##
## male                       -0.084*                 -0.280***
##                            (0.044)                 (0.096)
##
## married                    0.135***                 0.074
##                            (0.039)                 (0.078)
##
## highearn:afchnge           0.224***                 0.165
##                            (0.070)                 (0.155)
##
## Constant                   1.098***                1.589***
##                            (0.048)                 (0.101)
##
## Observations                5,362                   1,484
## R2                          0.025                   0.017
## Adjusted R2                 0.024                   0.013
## Residual Std. Error    1.260 (df = 5356)       1.369 (df = 1478)
## F Statistic         27.083*** (df = 5; 5356) 4.983*** (df = 5; 1478)
## ----------------------------------------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

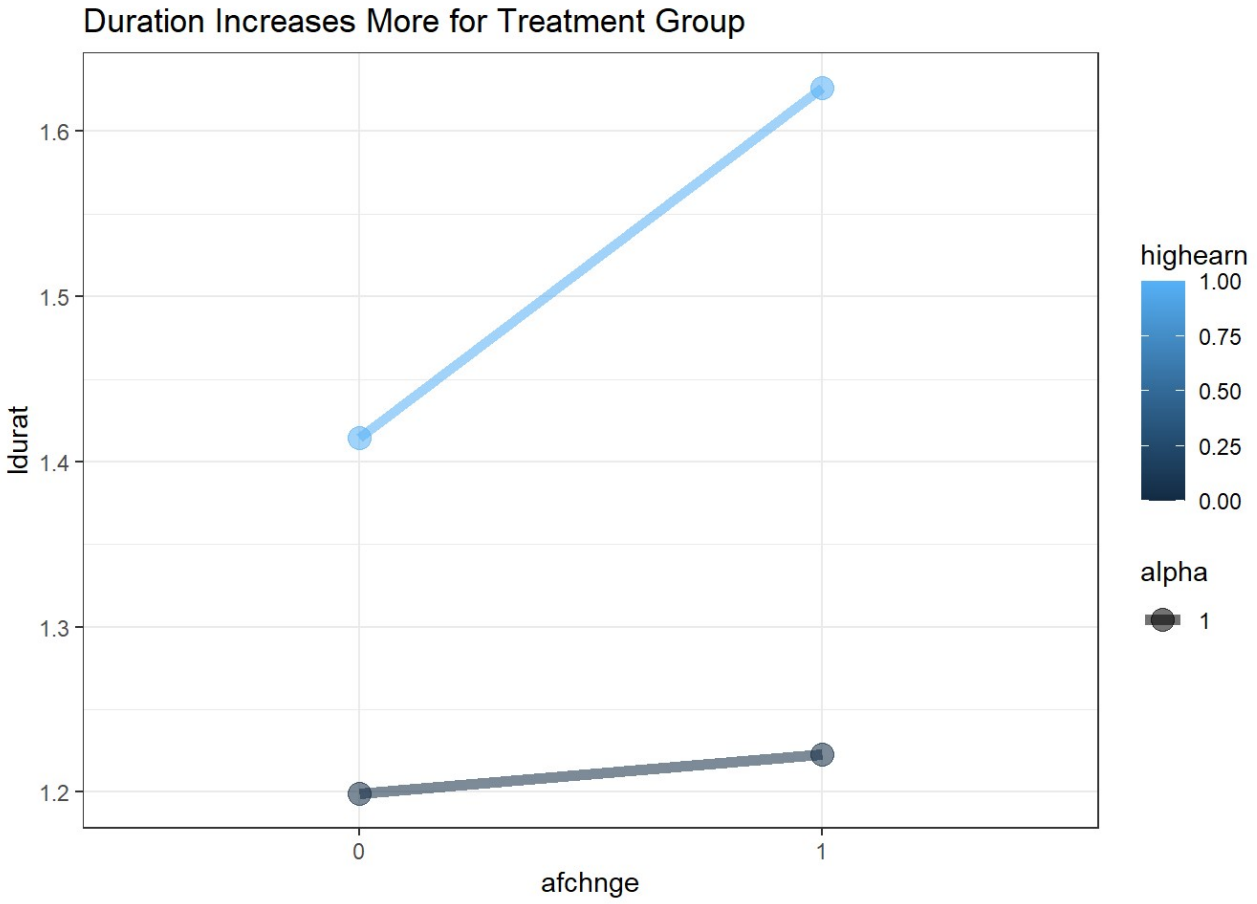
After adding the controls, the estimates on interaction term increases to 0.224 and is still statistically significant. The low R-squared suggests large degree of heterogeneity across individuals. If we use panel data and control for the fixed effects, the R-squared may increase significantly. If the fixed effects are highly correlated with the regressors considered here, then the estimates are biased and are useless. But if the fixed estimates are uncorrelated with the regressors, i.e. if they are in fact random effects, then the regression is still useful.

As shown in the second column, the estimate is not statistically signficant for the Michigan sample due to a larger standard error. However, the point estimate suggests the same sign of the effect and the magnitude is somewhat similar to that in Kentucky. The imprecise estimate is due to a smaller sample size in Michigan.

Visualization

Slope Chart for DiD.

```
df %>%
  mutate(afchnge = factor(afchnge)) %>%
  group_by(afchnge,highearn) %>%
  summarise(ldurat = mean(ldurat)) %>%
  ggplot(aes(x = afchnge, y = ldurat, group = highearn)) +
    geom_line(aes(color = highearn, alpha = 1), size = 2) +
    geom_point(aes(color = highearn, alpha = 1), size = 4) +
    labs(title='Duration Increases More for Treatment Group') +
    theme_bw()
```



Three or More Periods

When T>2, DiD lends itself to a test for causality in the spirit of Granger (1969). The Granger idea is to see whether causes happen before consequences and not vice versa. Suppose the policy variable of interest changes at different times in different states. In this context, Granger causality testing means a check of whether, conditional on state and year effects, past treatments predict future outcomes while future treatment does not predict past outcomes.

In our dataset, we can collect one more period of data before the change, and then check whether the high-earners and the low-earners actually show a common time trend before the change.

Synthetic Control Method

In some cases, treatment and potential control groups do not follow parallel trends and DiD estimator would lead to biased estimates. The basic idea behind it is that a combination of units of ten provides a better comparison for the unit exposed to the intervention than any single unit alone. For instance, you can take a weighted average of other units as a synthetic control group.

Panel Data Methods

Panel data is data where we observe the same unit (individual/firm/country) over time periods. A common feature of panel data is that sample of individuals is typically large and the number of timme periods is relatively short. There are three types of panel data:

1. Balanced: all identities are observed in all periods
2. Unbalanced: entry, exit, and non-response exist
3. Rotating: a share of the sample is renewed every year

Panel data methods have many benefits:

1. Increase in sample size increases precision of estimates
2. Control individual fixed effects which are common to an individual across time but might vary across agents at any point in time, and this is known as unobserved heterogeneity. This is not possible for cross-section data.
3. void aggregation bias: model behavior at the micro level not suited for aggregation

The key theme in the panel data methods is that if an omitted variable does not change over time, then any changes in Y over time cannot be caused by the omitted variable.

Regression using panel data may mitigate omitted variable bias when there is no information on variables that correlate with both the regressors of interest and the independent variable, and if these variables are constant in the time dimension or across entities.

Data

We will study two datasets and do the following models:

1. Quasi-Panel Data: two periods concatenated row-wise and different periods marked by a binary column (like DiD data)

- pooled OLS

2. Panel Data: Fatality from Drunk Driving

- First Difference Estimator
- Fixed Effect
- Random Effect

As opposed to DiD, we are considering data from two periods for the same individuals, as opposed to cross-sections from the same populations.

Dataset 1: Wage Determinants

This is a panel data of workers with two years of wage and its determinants on two periods: 1978 and 1985.

```
df <- read_dta('C:/Users/jihun/Downloads/applied_microeconometrics/CPS78_85.dta')
glimpse(df)
```

```
## Rows: 1,084
## Columns: 15
## $ educ     <dbl> 12, 12, 6, 12, 12, 8, 11, 15, 16, 15, 15, 12, 11, 12, 12, ...
## $ south    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ nonwhite <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ female   <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0...
## $ married  <dbl> 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1...
## $ exper    <dbl> 8, 30, 38, 19, 11, 43, 2, 9, 17, 23, 39, 5, 27, 29, 7, 42,...
## $ expersq  <dbl> 64, 900, 1444, 361, 121, 1849, 4, 81, 289, 529, 1521, 25, ...
## $ union    <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1...
## $ lwage    <dbl> 1.2150, 1.6094, 2.1401, 2.0732, 1.6490, 1.7148, 1.0986, 1....
## $ age      <dbl> 25, 47, 49, 36, 28, 56, 18, 29, 38, 43, 59, 22, 43, 46, 24...
## $ year     <dbl> 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78...
## $ y85      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ y85fem   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ y85educ  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ y85union <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

Data Preprocessing

We create two variables: 1. y85educ12: years of education subtracted by 12 2. llwage: inflation adjusted log-waged

```
df <-
  df %>%
  mutate(y85educ12 = y85*I(educ-12), # this can induce different interpretation for y8
5 from wage return on person with no education to 12 years of education
         llwage = ifelse(y85==1,I(lwage-log(1.65)),lwage)) # inflation adjusted wage l
og(wage/inflation) = log(wage) - log(inflation)
glimpse(df)
```

```
## Rows: 1,084
## Columns: 17
## $ educ      <dbl> 12, 12, 6, 12, 12, 8, 11, 15, 16, 15, 15, 12, 11, 12, 12,...
## $ south     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ nonwhite  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ female    <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, ...
## $ married   <dbl> 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, ...
## $ exper     <dbl> 8, 30, 38, 19, 11, 43, 2, 9, 17, 23, 39, 5, 27, 29, 7, 42...
## $ expersq   <dbl> 64, 900, 1444, 361, 121, 1849, 4, 81, 289, 529, 1521, 25,...
## $ union     <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, ...
## $ lwage     <dbl> 1.2150, 1.6094, 2.1401, 2.0732, 1.6490, 1.7148, 1.0986, 1...
## $ age       <dbl> 25, 47, 49, 36, 28, 56, 18, 29, 38, 43, 59, 22, 43, 46, 2...
## $ year      <dbl> 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 7...
## $ y85       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ y85fem    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ y85educ   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ y85union  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ y85educ12 <I<dbl>> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ llwage    <dbl> 1.2150, 1.6094, 2.1401, 2.0732, 1.6490, 1.7148, 1.0986, 1...
```

Fit Models: Pooled OLS

We fit the follwowing models to estimate the causal effect of union membership, education, and gender.

```
pmod1 <- lm(lwage ~ y85 + educ + y85educ + exper + expersq + union + female + y85fem,
data=df)
pmod2 <- lm(lwage ~ y85 + educ + exper + expersq + union + female + y85fem + y85educ1
2, data=df)
pmod3 <- lm(llwage ~ y85 + educ + y85educ + exper + expersq + union + female + y85fem,
data=df)
pmod4 <- lm(lwage ~ y85 + educ + y85educ + exper + expersq + union + female + y85fem +
y85union, data=df)
stargazer(pmod1, pmod2, pmod3, pmod4, type='text', title='Determinants of Wages in 197
8 and 1985', style='aer')
```

```
##
## Determinants of Wages in 1978 and 1985
## ================================================================================
===================================
##                                            lwage                                llwa
ge                  lwage
##                               (1)                     (2)                      (3)
(4)
## --------------------------------------------------------------------------------
-----------------------------------
## y85                          0.118                0.339***                -0.383
***                 0.118
##                             (0.124)              (0.034)                  (0.12
4)                (0.126)
##
## educ                        0.075***             0.075***                 0.075
***               0.075***
##                             (0.007)              (0.007)                  (0.00
7)                (0.007)
##
## y85educ                     0.018**                                       0.018
**                0.018**
##                             (0.009)                                       (0.00
9)                (0.009)
##
## exper                       0.030***             0.030***                 0.030
***               0.030***
##                             (0.004)              (0.004)                  (0.00
4)                (0.004)
##
## expersq                    -0.000***            -0.000***                -0.000
***               -0.000***
##                             (0.000)              (0.000)                  (0.00
0)                (0.000)
##
## union                       0.202***             0.202***                 0.202
***               0.202***
##                             (0.030)              (0.030)                  (0.03
0)                (0.039)
##
## female                     -0.317***            -0.317***                -0.317
***               -0.317***
##                             (0.037)              (0.037)                  (0.03
7)                (0.037)
##
## y85fem                       0.085*               0.085*                   0.08
5*                 0.085
##                             (0.051)              (0.051)                  (0.05
```

```
1)                  (0.052)
##
## y85educ12                                          0.018**
##                                                    (0.009)
##
## y85union
-0.000
##
(0.061)
##
## Constant                    0.459***                0.459***                  0.459
***                 0.459***
##                             (0.093)                 (0.093)                   (0.09
3)                  (0.095)
##
## Observations                 1,084                   1,084                     1,08
4                    1,084
## R2                           0.426                   0.426                     0.35
6                    0.426
## Adjusted R2                  0.422                   0.422                     0.35
1                    0.421
## Residual Std. Error    0.413 (df = 1075)        0.413 (df = 1075)        0.413 (df
= 1075)        0.413 (df = 1074)
## F Statistic         99.804*** (df = 8; 1075) 99.804*** (df = 8; 1075) 74.354*** (df
= 8; 1075) 88.632*** (df = 9; 1074)
## ---------------------------------------------------------------------------------
-----------------------------------
## Notes:              ***Significant at the 1 percent level.
##                     **Significant at the 5 percent level.
##                     *Significant at the 10 percent level.
```

The first model: - the wage return to education in 1978 is around 7.47% (educ). - The wage return to education over this time period has changed by around 1.85 percentage points (y85educ). - The gender wage gap in 1978 is around -31.7% (i.e. women have 31.7% lower wage on average than men). - The gender wage gap over this time period changed has b around 8.51 percentage points (y85fem). This is to say that the wage gap was smaller by 8.51 percentage points.

The second model: - The coefficient on y85 represents how much average wage has changed from 1978 to 1985 for men with no education (years of education equal to 0). - With y85educ12 coefficient we get a different interpretation for coefficient on y85. Now it represents how much average wage has changed from 1978 to 1985 for men with 12 years of education, which is around 33.9%.

The third model: The adjustment of inflation for log wages in 1985 is actually just a linear transformation. After the adjustment, the wages in 1985 are closer to the wages in 1978. Thus the total variation of log wage is smaller. Notice that the variation of log wages is still the same after controlling the year, thus the estimates of the coefficients on all other regressors besides y85 are the same. Therefore, the residuals and the sum of squared residuals are the same. This implies R-squared is now smaller. Only the coefficient on y85 is different, while all others remain almost the same.

The fourth model: y85union is not statistically significant so there is little evidence that union participation affects wage.

Limitations

If we have actual panel data instead of quasi-panel data, we could use fixed effects to control for unobserved heterogeneity that is fixed over times within the individual workers. In other words, we would use the variation within a worker to estimate the effect of different regressors on wages (fixed effect model).

Identifying Assumptions

If the fixed effects are correlated with the independent variables, pooled OLS gives biased estimates, and we have no way of knowing whether this is true.

Dataset 2: Fatality Rates Across States

This study design is aimed at investigating the causal effect of drunk driving laws on traffic fatalities. The data's observational unit is a year in a US state. It consists of 48 states and spans 7 year period from 1982 to 1988. Since all variables are observed for all entities and over all time periods, the panel is balanced. If there were missing data for at least one entity in at least one time period we would call the panel unbalanced. The key variables are traffic fatality rate (# of traffic deaths in that state in that year, per 10,000 state residents), tax on beer, and other variables such as legal drinking age, drunk driving laws, etc.

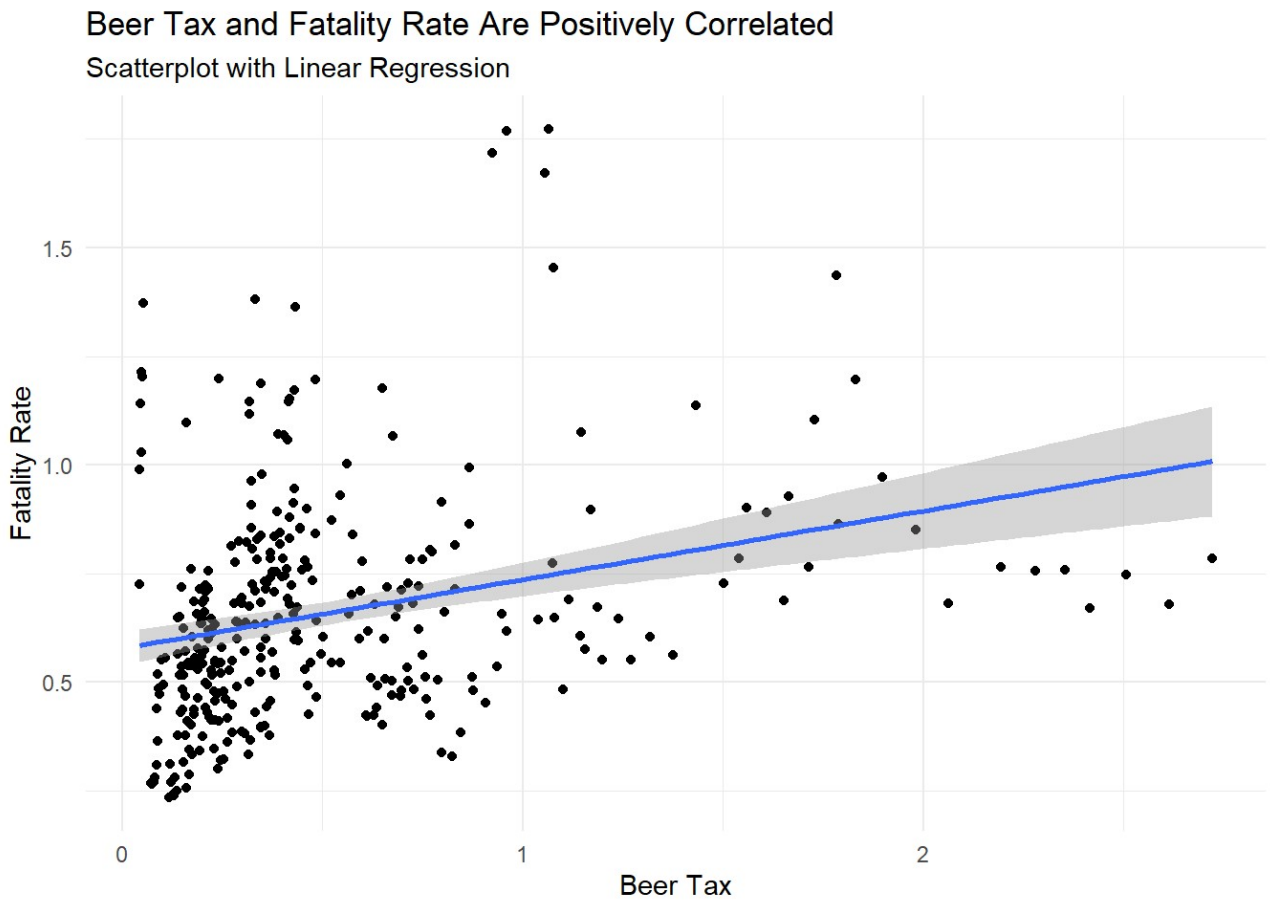
```
data(Fatalities)
# convert df into pdata.frame format with new indices
df <- pdata.frame(Fatalities, index=c('state', 'year'))
# create a fatality rate column: proportion of alcohol-related in the population
df$fatal_rate <- df$afatal / df$pop * 10000
# this one is on vehicle related fatalities
df$fatality_rate <- df$fatal/df$pop * 10000
# create a lagged variable beer table
df <-
  df %>%
  group_by(state) %>%
  mutate(diff_fatal_rate = fatal_rate - lag(fatal_rate),
         diff_beertax = beertax - lag(beertax)) %>%
  ungroup()
glimpse(df)
```

```
## Rows: 336
## Columns: 38
## $ state           <fct> al, al, al, al, al, al, al, az, az, az, az, az, az,...
## $ year            <fct> 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1982, 198...
## $ spirits         <dbl> 1.37, 1.36, 1.32, 1.28, 1.23, 1.18, 1.17, 1.97, 1.9...
## $ unemp           <dbl> 14.4, 13.7, 11.1, 8.9, 9.8, 7.8, 7.2, 9.9, 9.1, 5.0...
## $ income          <dbl> 10544.15, 10732.80, 11108.79, 11332.63, 11661.51, 1...
## $ emppop          <dbl> 50.69204, 52.14703, 54.16809, 55.27114, 56.51450, 5...
## $ beertax         <dbl> 1.53937948, 1.78899074, 1.71428561, 1.65254235, 1.6...
## $ baptist         <dbl> 30.3557, 30.3336, 30.3115, 30.2895, 30.2674, 30.245...
## $ mormon          <dbl> 0.32829, 0.34341, 0.35924, 0.37579, 0.39311, 0.4112...
## $ drinkage        <dbl> 19.00, 19.00, 19.00, 19.67, 21.00, 21.00, 21.00, 19...
## $ dry             <dbl> 25.0063, 22.9942, 24.0426, 23.6339, 23.4647, 23.792...
## $ youngdrivers    <dbl> 0.211572, 0.210768, 0.211484, 0.211140, 0.213400, 0...
## $ miles           <dbl> 7233.887, 7836.348, 8262.990, 8726.917, 8952.854, 9...
## $ breath          <fct> no, no, no, no, no, no, no, no, no, no, no, no, no,...
## $ jail            <fct> no, no, no, no, no, no, no, yes, yes, yes, yes, yes...
## $ service         <fct> no, no, no, no, no, no, no, yes, yes, yes, yes, yes...
## $ fatal           <int> 839, 930, 932, 882, 1081, 1110, 1023, 724, 675, 869...
## $ nfatal          <int> 146, 154, 165, 146, 172, 181, 139, 131, 112, 149, 1...
## $ sfatal          <int> 99, 98, 94, 98, 119, 114, 89, 76, 60, 81, 75, 85, 8...
## $ fatal1517       <int> 53, 71, 49, 66, 82, 94, 66, 40, 40, 51, 48, 72, 50,...
## $ nfatal1517      <int> 9, 8, 7, 9, 10, 11, 8, 7, 7, 8, 11, 19, 16, 14, 5, ...
## $ fatal1820       <int> 99, 108, 103, 100, 120, 127, 105, 81, 83, 118, 100,...
## $ nfatal1820      <int> 34, 26, 25, 23, 23, 31, 24, 16, 19, 34, 26, 30, 25,...
## $ fatal2124       <int> 120, 124, 118, 114, 119, 138, 123, 96, 80, 123, 121...
## $ nfatal2124      <int> 32, 35, 34, 45, 29, 30, 25, 36, 17, 33, 30, 25, 34,...
## $ afatal          <dbl> 309.438, 341.834, 304.872, 276.742, 360.716, 368.42...
## $ pop             <dbl> 3942002, 3960008, 3988992, 4021008, 4049994, 408299...
## $ pop1517         <dbl> 208999.6, 202000.1, 197000.0, 194999.7, 203999.9, 2...
## $ pop1820         <dbl> 221553.4, 219125.5, 216724.1, 214349.0, 212000.0, 2...
## $ pop2124         <dbl> 290000.1, 290000.2, 288000.2, 284000.3, 263000.3, 2...
## $ milestot        <dbl> 28516, 31032, 32961, 35091, 36259, 37426, 39684, 19...
## $ unempus         <dbl> 9.7, 9.6, 7.5, 7.2, 7.0, 6.2, 5.5, 9.7, 9.6, 7.5, 7...
## $ emppopus        <dbl> 57.8, 57.9, 59.5, 60.1, 60.7, 61.5, 62.3, 57.8, 57....
## $ gsp             <dbl> -0.022124760, 0.046558253, 0.062797837, 0.027489973...
## $ fatal_rate      <numeric> 0.7849767, 0.8632155, 0.7642834, 0.6882404, 0.8...
## $ fatality_rate   <integer> 2.12836, 2.34848, 2.33643, 2.19348, 2.66914, 2....
## $ diff_fatal_rate <dbl> NA, 0.07823877, -0.09893209, -0.07604296, 0.2024177...
## $ diff_beertax    <dbl> NA, 0.249611259, -0.074705124, -0.061743259, -0.042...
```

When one only looks at the correlation between beer tax and fatality rate, there is a positive correlation. This is because we are not controlling for other factors such as quality of cars, roads, culture of drink-driving, traffic density, poverty, social problems and higher taxes, etc. These omitted variables are creating a bias on the causal effect. If all these unobserved variables are constant across time, panel data allows us to bypass the bias caused by them.

```
ggplot(df, aes(x=beertax, y=fatal_rate)) +
  geom_point() +
  geom_smooth(method='lm') +
  theme_minimal() +
  labs(title='Beer Tax and Fatality Rate Are Positively Correlated',
       x='Beer Tax',
       y='Fatality Rate',
       subtitle='Scatterplot with Linear Regression')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Fitting Panel Data Models

1. Pooled OLS
2. First Difference
3. Fixed Effect on Individuals
4. Fixed Effect on Individuals and Time Effect
5. Random Effect

Pooled OLS

Naive model of running OLS on data.

```
poolmod <- plm(fatal_rate ~ beertax + gsp + breath + jail + service + drinkage + dry +
spirits, data=df, model='pooling')
summary(poolmod)
```

```
## Pooling Model
##
## Call:
## plm(formula = fatal_rate ~ beertax + gsp + breath + jail + service +
##     drinkage + dry + spirits, data = df, model = "pooling")
##
## Unbalanced Panel: n = 48, T = 6-7, N = 335
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.461756 -0.157343 -0.030305  0.094981  1.036024
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  0.9980880  0.2994764  3.3328 0.0009587 ***
## beertax      0.1555814  0.0268155  5.8019 1.552e-08 ***
## gsp         -1.6722228  0.3029755 -5.5193 6.951e-08 ***
## breathyes   -0.0473591  0.0267723 -1.7690 0.0778357 .
## jailyes      0.1561560  0.0351678  4.4403 1.231e-05 ***
## serviceyes  -0.0353577  0.0384929 -0.9186 0.3590097
## drinkage    -0.0232392  0.0142227 -1.6340 0.1032327
## dry          0.0068132  0.0013850  4.9193 1.380e-06 ***
## spirits      0.0312674  0.0198635  1.5741 0.1164315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    22.542
## Residual Sum of Squares: 15.864
## R-Squared:      0.29623
## Adj. R-Squared: 0.27896
## F-statistic: 17.1521 on 8 and 326 DF, p-value: < 2.22e-16
```

First Difference Model

We create new variables that represent the year-to-year change in each variable within a state: the change in the fatality rate in Alabama between 1982 and 1983, between 1983 and 1984, and so on. We then run the regression in these differences. This is closely related to the regression we ran above looking at the change between 1982 and 1988, but in this case we use all the year-to-year changes.

First Difference Model (two-period: 1982 and 1988)

Take a first difference across two time periods, and this will remove the effect of omitted variable.

```
# subset the data
Fatalities1982 <- subset(df, year == "1982")
Fatalities1988 <- subset(df, year == "1988")
# compute the differences
diff_fatal_rate <- Fatalities1988$fatal_rate - Fatalities1982$fatal_rate
diff_beertax <- Fatalities1988$beertax - Fatalities1982$beertax
# estimate a regression using differenced data
fatal_diff_mod <- lm(diff_fatal_rate ~ diff_beertax)
coeftest(fatal_diff_mod, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  -0.140571   0.032252 -4.3585 7.285e-05 ***
## diff_beertax -0.246778   0.285572 -0.8642     0.392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First Difference Model (all periods: 1982~1988)

```
fdmod <- plm(fatal_rate ~ beertax + gsp + breath + jail + service + drinkage + dry + s
pirits, data = df, model = "fd")
summary(fdmod)
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = fatal_rate ~ beertax + gsp + breath + jail + service +
##     drinkage + dry + spirits, data = df, model = "fd")
##
## Unbalanced Panel: n = 48, T = 6-7, N = 335
## Observations used in estimation: 287
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.5718594 -0.0508466  0.0017233  0.0640556  1.1895587
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## (Intercept) -0.0193690  0.0115459 -1.6776 0.094557 .
## beertax      0.0716215  0.2012389  0.3559 0.722183
## gsp          0.3148350  0.1929931  1.6313 0.103953
## breathyes   -0.1327852  0.0441233 -3.0094 0.002858 **
## jailyes      0.1114842  0.0970386  1.1489 0.251599
## serviceyes  -0.1544446  0.1156278 -1.3357 0.182738
## drinkage     0.0034263  0.0203266  0.1686 0.866265
## dry          0.0077593  0.0135911  0.5709 0.568524
## spirits      0.0202424  0.1184001  0.1710 0.864375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.4692
## Residual Sum of Squares: 5.2066
## R-Squared:      0.048023
## Adj. R-Squared: 0.020628
## F-statistic: 1.75297 on 8 and 278 DF, p-value: 0.086358
```

Identifying Assumptions:

There are certain problems inherent to the First Difference Estimator and key identifying assumptions:

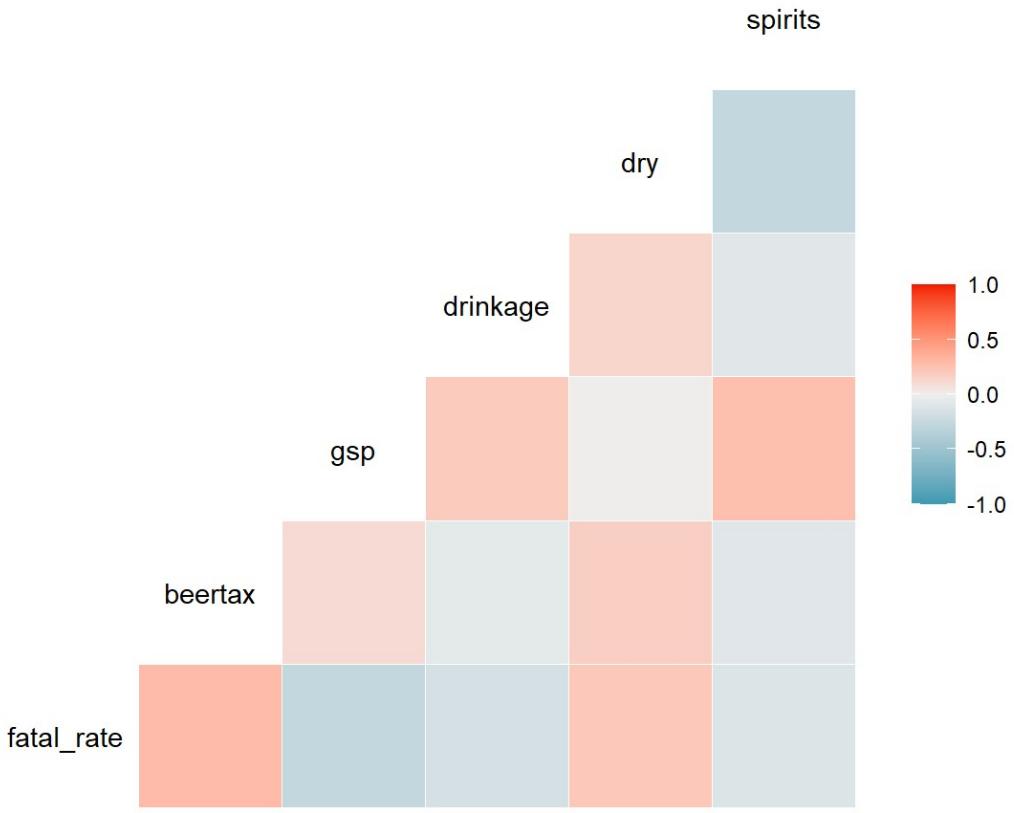
1. Since we take out all the time-independent variation in the independent variables, we effectively reduce the amount of variation we can use in estimation, implying that estiation using first difference can become very imprecise and have higher standard error. This is a problem that can be solved as we increase sample size.
2. However, attenuation bias from measurement error is irreducible and in fact amplifies because noise tends to be time variant. Measurement error introduces greater standard error and cause bias.
3. We want strict exogeneity, and this assumption could be violated if we have omitted an important time-variant variable. This means future independent variable should not depend on current changes in the idiosyncatic errors. The omitted variable bias causes bias in the estimate.

4. We also need differenced errors to be uncorrelated for the standard errors and test statistics to be valid. Interestingly, uncorrelated original errors can have first differened errors that are in fact correlated over time. The serial correlation in the first differenced errors does not cause bias but the estimated standard errors are incorrect and can cause biass.

We need to make sure the fixed effect is uncorrelated with the independent variables or else the coefficient is biased.

```
subdf <- df[,c('fatal_rate','beertax','gsp','breath','jail','service','drinkage','dr
y','spirits')]
print(ggcorr(subdf, method = c("everything", "pearson")))
```

```
## Warning in ggcorr(subdf, method = c("everything", "pearson")): data in column(s)
## 'breath', 'jail', 'service' are not numeric and were ignored
```



Another way to check correlation among variables.

```
print(ggpairs(subdf, progress=F))
```

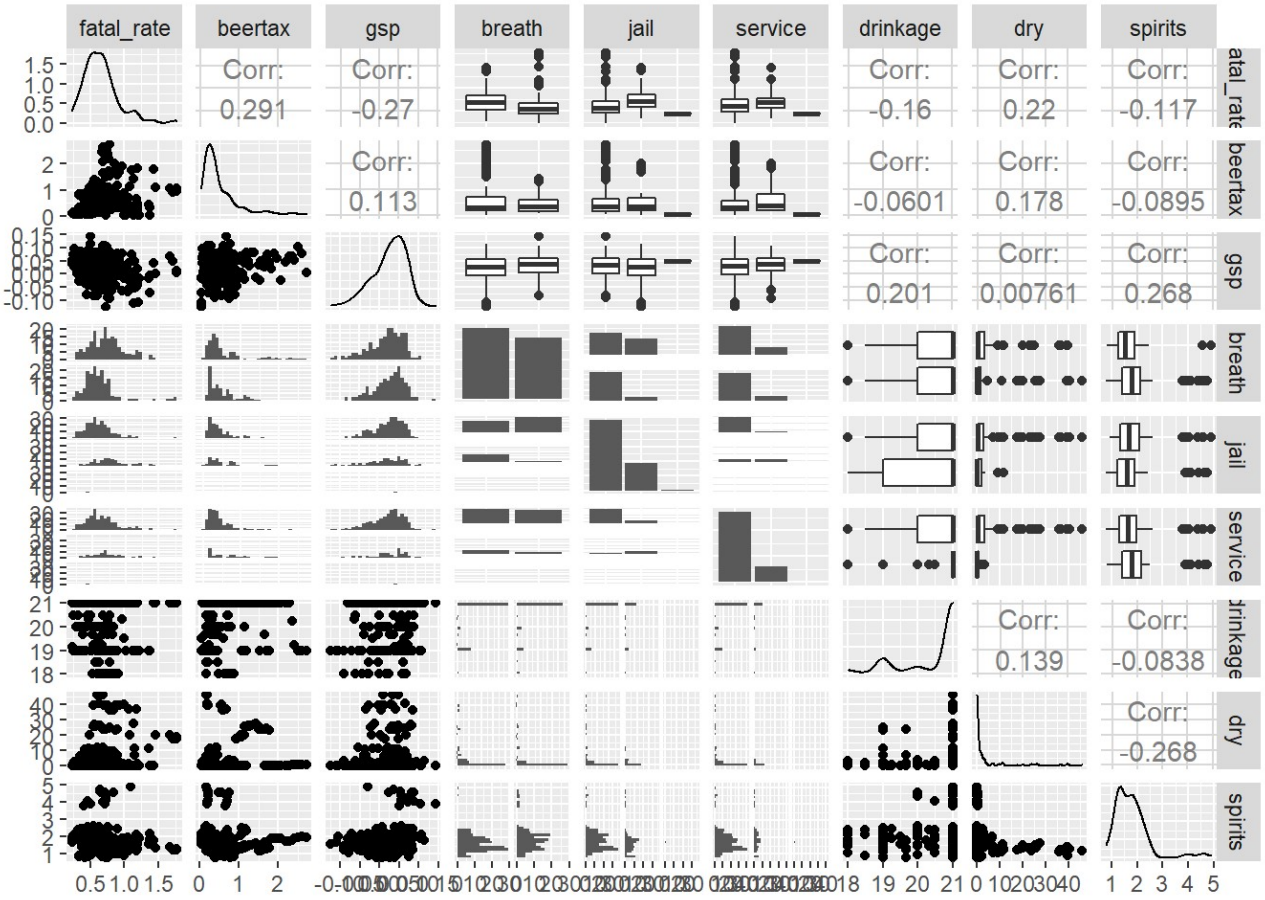
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing missing values (stat_boxplot).

## Warning: Removed 1 rows containing missing values (stat_boxplot).

## Warning: Removed 1 rows containing missing values (stat_boxplot).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing missing values (stat_boxplot).

## Warning: Removed 1 rows containing missing values (stat_boxplot).

## Warning: Removed 1 rows containing missing values (stat_boxplot).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Did we omit any important time varying variable? If we do, the strict exogeneity assumption could be violated and the future X should not depend on current changes in the idiosyncratic errors.

Testing for Serial Correlation

When change in errors is uncorrelated over time, the standard errors and test statistics are valid. If there is a serial correlation, then the estimated standard errors are incorrect.

```
pbgtest(fdmod) # null=no serial correlation
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +     dry +
spirits
## chisq = 20.766, df = 6, p-value = 0.002021
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Testing for Unit roots/stationarity

```
adf.test(df$fatal_rate, k=2) # null hypothesis: unit roots are present
```

```
## Warning in adf.test(df$fatal_rate, k = 2): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  df$fatal_rate
## Dickey-Fuller = -6.5456, Lag order = 2, p-value = 0.01
## alternative hypothesis: stationary
```

Fixed Effect Model

Having individual specific intercepts capture heterogeneities across entities. Each intercept is dropped because each unit's own intercept is calculated. It allows us to control for unobserved time-independent factors and estimate effect of variables that do vary over time, potentially reducing bias or inconsistency. Fixed effect model is also called within estimator, and time demean all variables within modelling process (unlike FD where we take a difference between the previous values).

Just like First Difference, the model reduces the total amount of variation in the data that exacerbates attenuation bias due to measurement error in the independent variables. In fact, when there are only two periods, the fixed effect model and the first difference model are the same. However, when there are more than two periods, FE and FD are not equivalent, although they are both unbiased given that all the relevant assumptions hold.

Time demeaning in Fixed Effect model implies we look at deviations in the variables from average for each individual over the period. On the other hand, FD implies that we look at period-to-period changes in the variables. Per-period changes may not be the same as deviations from the average over all time periods. Since both FE and FD are unbiased given the assumptions are true, we may choose between FE and FD depending on their relative efficiency. The efficency in turn depends on the serial correlation in the idiosyncratic errors U_it. If they are serially uncorrelated, FE is more efficient. If first differenced errors are serially uncorrelated, FD is more efficient. In sum, the choice between FD and FE hinges on serial correlation of errors and first differenced errors. If T is large and N is small, FE can be very sensitive to violations of the classical fixed effect, and should not be used.

The greatest benefit of FE model is that we do not have to assume that time-independent variables are uncorrelated with the independent variables of interest. plm() does not give us coefficients for fixed effects, but we can back them out using OLS. In effect, the FE model runs the pooled regression with a complete set of state dummy variables.

```
fe_mod1 <- plm(fatal_rate ~ beertax + gsp + breath + jail + service + drinkage + dry +
spirits,
               data = df,
               model = "within",
               effect='individual')
summary(fe_mod1)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = fatal_rate ~ beertax + gsp + breath + jail + service +
##     drinkage + dry + spirits, data = df, effect = "individual",
##     model = "within")
##
## Unbalanced Panel: n = 48, T = 6-7, N = 335
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.9611301 -0.0572586 -0.0024756  0.0522149  0.4649044
##
## Coefficients:
##              Estimate Std. Error t-value  Pr(>|t|)
## beertax    -0.2608836  0.1320805 -1.9752   0.04923 *
## gsp         0.1370988  0.2337551  0.5865   0.55801
## breathyes  -0.0168124  0.0401063 -0.4192   0.67540
## jailyes     0.1872312  0.0984175  1.9024   0.05815 .
## serviceyes -0.1631787  0.1126591 -1.4484   0.14862
## drinkage    0.0069354  0.0141131  0.4914   0.62352
## dry         0.0045011  0.0106439  0.4229   0.67271
## spirits     0.3365342  0.0604143  5.5704 5.979e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.3854
## Residual Sum of Squares: 4.6066
## R-Squared:      0.14462
## Adj. R-Squared: -0.024007
## F-statistic: 5.89622 on 8 and 279 DF, p-value: 5.5849e-07
```

Controlling for variables that are constant across entities but vary over time can be done by including time fixed effects. The combined model allows to eliminate bias from unobservables that change over time but are constant over entities and it controls for factors that differ across entities but are constant over time. Such models can be estimated using the OLS algorithm that is implemented in R. Note that plm() uses the entity-demeaned OLS algorithm and thus does not report dummy coefficients.

```
fe_mod2 <- plm(fatal_rate ~ beertax + gsp + breath + jail + service + drinkage + dry +
spirits,
              data = df,
              model = "within",
              effect='twoways')
summary(fe_mod2)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = fatal_rate ~ beertax + gsp + breath + jail + service +
##     drinkage + dry + spirits, data = df, effect = "twoways",
##     model = "within")
##
## Unbalanced Panel: n = 48, T = 6-7, N = 335
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -9.2021e-01 -5.1456e-02 -3.5949e-05  4.9471e-02  4.3303e-01
##
## Coefficients:
##               Estimate  Std. Error t-value  Pr(>|t|)
## beertax    -0.23555974  0.13313635 -1.7693  0.077958 .
## gsp         1.00976468  0.32463956  3.1104  0.002066 **
## breathyes  -0.00519473  0.03905600 -0.1330  0.894286
## jailyes     0.21703929  0.09641663  2.2511  0.025178 *
## serviceyes -0.17548983  0.11057013 -1.5871  0.113639
## drinkage    0.00069322  0.01442531  0.0481  0.961707
## dry         0.00065790  0.01035837  0.0635  0.949404
## spirits     0.37094677  0.09287833  3.9939 8.363e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    4.8086
## Residual Sum of Squares: 4.2355
## R-Squared:      0.11917
## Adj. R-Squared: -0.077645
## F-statistic: 4.61688 on 8 and 273 DF, p-value: 2.6551e-05
```

Control for heteroskedasticity.

```
# print summary using robust standard errors
coeftest(fe_mod2, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## beertax    -0.23555974  0.21407001 -1.1004 0.2721331
## gsp         1.00976468  0.29604892  3.4108 0.0007457 ***
## breathyes  -0.00519473  0.05514463 -0.0942 0.9250179
## jailyes     0.21703929  0.01257321 17.2620 < 2.2e-16 ***
## serviceyes -0.17548983  0.08146491 -2.1542 0.0321014 *
## drinkage    0.00069322  0.01838745  0.0377 0.9699538
## dry         0.00065790  0.02112529  0.0311 0.9751785
## spirits     0.37094677  0.09398167  3.9470 0.0001007 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Identifying Assumptions of Fixed Effect Model

1. Strict exogeneity - We must assume strict exogeneity of the explanatory variables which implies that each error should be uncorrelated not only with the independent variables in the given period, but also with all the independent variables in all other periods. This is the key assumption for FE to give unbiased estimates.
2. Uncorrelated errors gives correct standard errors.

```
pbgtest(fe_mod1) # null=no serial correlation
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +     dry +
spirits
## chisq = 52.154, df = 6, p-value = 1.737e-09
## alternative hypothesis: serial correlation in idiosyncratic errors
```

We conclude that the estimated relationship between traffic fatalities and the real beer tax is not affected by omitted variable bias due to factors that are constant over time.

```
pbgtest(fe_mod2)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel models
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +     dry +
spirits
## chisq = 50.912, df = 6, p-value = 3.085e-09
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Robust Standard Errors for Twoway Fixed Effect Models

When there is both heteroskedasticity and autocorrelation, the so-called heteroskedasticity and autocorrelation-consistent (HAC) standard errors need to be used. Clustered standard errors belong to these type of standard errors. They allow for heteroskedasticity and autocorrelated errors within an entity but not correlation across entities.

```
coeftest(fe_mod2, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## beertax    -0.23555974  0.21407001 -1.1004 0.2721331
## gsp         1.00976468  0.29604892  3.4108 0.0007457 ***
## breathyes  -0.00519473  0.05514463 -0.0942 0.9250179
## jailyes     0.21703929  0.01257321 17.2620 < 2.2e-16 ***
## serviceyes -0.17548983  0.08146491 -2.1542 0.0321014 *
## drinkage    0.00069322  0.01838745  0.0377 0.9699538
## dry         0.00065790  0.02112529  0.0311 0.9751785
## spirits     0.37094677  0.09398167  3.9470 0.0001007 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use Arellano method.

```
coeftest(fe_mod2, vcovHC(fe_mod2,method='arellano'))
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## beertax    -0.23555974  0.21149851 -1.1138 0.2663594
## gsp         1.00976468  0.29249265  3.4523 0.0006441 ***
## breathyes  -0.00519473  0.05448221 -0.0953 0.9241090
## jailyes     0.21703929  0.01242217 17.4719 < 2.2e-16 ***
## serviceyes -0.17548983  0.08048632 -2.1804 0.0300842 *
## drinkage    0.00069322  0.01816657  0.0382 0.9695887
## dry         0.00065790  0.02087153  0.0315 0.9748768
## spirits     0.37094677  0.09285272  3.9950 8.326e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use type 3 method.

```
coeftest(fe_mod2, vcovHC(fe_mod2, type='HC3'))
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## beertax    -0.23555974  0.23171204 -1.0166 0.3102412
## gsp         1.00976468  0.29980996  3.3680 0.0008661 ***
## breathyes  -0.00519473  0.05869095 -0.0885 0.9295363
## jailyes     0.21703929  0.01282704 16.9204 < 2.2e-16 ***
## serviceyes -0.17548983  0.09407059 -1.8655 0.0631822 .
## drinkage    0.00069322  0.01868679  0.0371 0.9704349
## dry         0.00065790  0.02183229  0.0301 0.9759821
## spirits     0.37094677  0.09368856  3.9594 9.593e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Time-fixed effects

Test for whether model needs time effects on the fixed effect model.

```
pFtest(fe_mod1, fe_mod2) # if the pvalue is low, then we dont need to use time-fixed e
ffect
```

```
## Warning in pf(stat, df1, df2, lower.tail = FALSE): NaNs produced
```

```
##
##  F test for individual effects
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +  ...
## F = 3.7456, df1 = -6, df2 = 279, p-value = NA
## alternative hypothesis: significant effects
```

Random Effect Model

If we assume that unobserved time-invariant variable is uncorrelated with the independent variable (unlike fixed effect and first difference model where we control for unobserved variable), then we should use a random effect model.

RE model allows us to correct for the serial correlation caused by the fixed effects. It is part of larger family called Generalized Least Squares. When theta is equal to 0, it becomes pooled OLS estimator. When theta is 1, it becomes FE estimator.

```
re_mod <- plm(fatal_rate ~ beertax + gsp + breath + jail + service + drinkage + dry +
spirits, data=df, model='random')
summary(re_mod)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = fatal_rate ~ beertax + gsp + breath + jail + service +
##     drinkage + dry + spirits, data = df, model = "random")
##
## Unbalanced Panel: n = 48, T = 6-7, N = 335
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.01651 0.12850 0.364
## individual    0.02888 0.16994 0.636
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7051  0.7252  0.7252  0.7249  0.7252  0.7252
##
## Residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.76523 -0.08021 -0.01850  0.00003  0.05178  0.57993
##
## Coefficients:
##               Estimate Std. Error z-value Pr(>|z|)
## (Intercept)  0.8104339  0.2953558  2.7439 0.006071 **
## beertax      0.1066438  0.0540508  1.9730 0.048493 *
## gsp         -0.2196300  0.2357502 -0.9316 0.351532
## breathyes   -0.0657353  0.0340562 -1.9302 0.053582 .
## jailyes      0.1700979  0.0591842  2.8740 0.004053 **
## serviceyes  -0.1250844  0.0677973 -1.8450 0.065041 .
## drinkage    -0.0188216  0.0128269 -1.4674 0.142280
## dry          0.0075357  0.0028914  2.6062 0.009154 **
## spirits      0.0903877  0.0345086  2.6193 0.008812 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6.6764
## Residual Sum of Squares: 6.0028
## R-Squared:      0.1009
## Adj. R-Squared: 0.078833
## Chisq: 36.583 on 8 DF, p-value: 1.373e-05
```

Hausman Test

Fixed Effect model v random effect model.

```
phtest(fe_mod2, re_mod) # null hypothesis: errors are not correlated with the regresso
rs
```

```
##
##  Hausman Test
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +  ...
## chisq = 64.156, df = 8, p-value = 7.086e-11
## alternative hypothesis: one model is inconsistent
```

```
# if pvalue is low, then we need to use fixed effects
```

Testing for Random Effect: Breusch-Pagan Lagrange Multiplier

Is there a significant random effect?

```
plmtest(re_mod, type=c('bp')) # Breusch-Pagan Lagrange Multiplier for random effects;
null is no panel effect (i.e. OLS better)
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels
##
## data:  fatal_rate ~ beertax + gsp + breath + jail + service + drinkage +  ...
## chisq = 309.55, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Putting It All Together: Pooled OLS, FD, FE, and RE

We have two sources of variation in panel data: between and within variation. The difference in average fatality rates across states over a period of ten years is an example of between-variation. The variation in fatality rate from year to year for a given state is an example of within-variation.

1. Pooled OLS uses an unweighte average of the within and between variation.
2. FE only uses within-variation.
3. RE uses both between and within variation, but weights these sources of variation in a different way compared to OLS.

If the assumption that unobserved time-invariant variable A is uncorrelated with independent variable X for all periods and individuals, then all the models - OLS, FE, and RE - have unbiased estimates.

Standard errors from FD, FE, and RE will typically be correct, but RE will have lower standard errors since it also makes use of the between variation in the data. On the other hand, standard errors from pooled OLS will be incorrect and too low on average because we dont adjust for serial correlation in the error term.

RE allows us to overcome problems with fixed effects. Unlike FE, RE allows us to estimate the effect of time independent variables. If covariance between X and A is zero, then RE is consistent and more efficient since the cross-sectional variation in the data is not thrown out. However, when covariance is non-zero, only FE is consistent. RE and FE differ in what is assumed by the intercept, and RE views this as randomly drawn or part of the error term.

```
rob_se <- list(sqrt(diag(vcovHC(poolmod, type = "HC1"))),
               sqrt(diag(vcovHC(fdmod, type = "HC1"))),
               sqrt(diag(vcovHC(fe_mod1, type = "HC1"))),
               sqrt(diag(vcovHC(fe_mod2, type = "HC1"))),
               sqrt(diag(vcovHC(re_mod, type = "HC1"))))
stargazer(poolmod, fdmod, fe_mod1, fe_mod2, re_mod, type='text', title='Comparison of
Panel Data Models', column.labels=c('Pooled OLS', 'First Difference', 'State FE', 'St-
Yr FE', 'RE'), style='aer', model.numbers=F)
```

```
##
## Comparison of Panel Data Models
## ================================================================================================
==============================
##                                                                  fatal_rate
##                     Pooled OLS          First Difference          State FE
St-Yr FE           RE
## ------------------------------------------------------------------------------------------------
------------------------------
## beertax              0.156***                0.072                -0.261**
-0.236*          0.107**
##                       (0.027)               (0.201)               (0.132)
(0.133)          (0.054)
##
## gsp                 -1.672***               0.315                 0.137
1.010***         -0.220
##                       (0.303)               (0.193)               (0.234)
(0.325)          (0.236)
##
## breathyes             -0.047*              -0.133***               -0.017
-0.005          -0.066*
##                       (0.027)               (0.044)               (0.040)
(0.039)          (0.034)
##
## jailyes              0.156***               0.111                 0.187*
0.217**         0.170***
##                       (0.035)               (0.097)               (0.098)
(0.096)          (0.059)
##
## serviceyes           -0.035                 -0.154                -0.163
-0.175          -0.125*
##                       (0.038)               (0.116)               (0.113)
(0.111)          (0.068)
##
## drinkage             -0.023                 0.003                 0.007
0.001           -0.019
##                       (0.014)               (0.020)               (0.014)
(0.014)          (0.013)
##
## dry                  0.007***               0.008                 0.005
0.001          0.008***
##                       (0.001)               (0.014)               (0.011)
(0.010)          (0.003)
##
## spirits               0.031                 0.020                0.337***
0.371***        0.090***
##                       (0.020)               (0.118)               (0.060)
(0.093)          (0.035)
```

```
##
## Constant            0.998***               -0.019*
0.810***
##                      (0.299)               (0.012)
(0.295)
##
## Observations           335                   287                   335
335              335
## R2                    0.296                 0.048                 0.145
0.119            0.101
## Adjusted R2           0.279                 0.021                 -0.024
-0.078           0.079
## F Statistic  17.152*** (df = 8; 326) 1.753* (df = 8; 278) 5.896*** (df = 8; 279) 4.
617*** (df = 8; 273) 36.583***
## ------------------------------------------------------------------------------------
------------------------------
## Notes:       ***Significant at the 1 percent level.
##              **Significant at the 5 percent level.
##              *Significant at the 10 percent level.
```