Untitled

2024-08-14

##1.A

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(estimatr)
library(tidyr)
library(rdd)
```

```
## Loading required package: sandwich

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: AER

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
patents <- read.csv("patents.csv")

aggregated_data <- patents %>%
  group_by(uspto_class) %>%
  summarize(usa_pre_mean = mean(count_usa[grntyr < 1919]),usa_post_mean = mean(count_usa[grntyr >= 1919

aggregated_data <- aggregated_data %>%
  mutate(diff = usa_post_mean - usa_pre_mean)

model_1A <- lm_robust(diff ~ treatment, data = aggregated_data)
summary(model_1A)
```

```
##
## Call:
## lm_robust(formula = diff ~ treatment, data = aggregated_data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   0.3862    0.01012  38.147 2.548e-290   0.3663   0.4060 7246
## treatment     0.2553    0.03769   6.774  1.352e-11   0.1814   0.3292 7246
##
## Multiple R-squared:  0.004124 ,  Adjusted R-squared:  0.003987
## F-statistic: 45.89 on 1 and 7246 DF,  p-value: 1.352e-11
```

We see a positive treatment effect of 0.2553 of the treatment. The confidence interval is [0.1814, 0.3292], which does not include 0. Thus we fail to reject the null which states that there is no treatment effect.

##1.B

```
treated_subclass <- aggregated_data[aggregated_data$treatment==1,]
untreated_subclass <- aggregated_data[aggregated_data$treatment==0,]
print(c(mean(treated_subclass$usa_pre_mean),mean(untreated_subclass$usa_pre_mean)))
```

```
## [1] 0.08272457 0.22787116
```

Ignorability likely not hold, since we can see that for treated group, the pre-mean is generally lower than untreated group. This suggest that the law tends to give subclasses which previously do not doing well (less count of patents) the access to enemy country patent access.

##1.C

```
treatment_table <- patents[patents$grntyr == 1919, c("uspto_class", "treat")]

times <- list(
  c(1918, 1917),
  c(1918, 1916),
  c(1918, 1915),
  c(1918, 1914)
)

for (time in times) {
  period <- patents %>% filter(grntyr %in% time)
  period_df <- merge(period, treatment_table, by = "uspto_class", all.x = TRUE) %>% group_by(uspto_clas
    summarize(trend_diff = count_usa[grntyr == time[1]] - count_usa[grntyr == time[2]],
              treatment = max(treat.y))
  model_trend <- lm_robust(trend_diff ~ treatment, data = period_df)
  print(summary(model_trend))
}
```

```
##
## Call:
## lm_robust(formula = trend_diff ~ treatment, data = period_df)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower  CI Upper   DF
## (Intercept) -0.03299    0.01207 -2.7326 0.006298 -0.05665 -0.009323 7246
## treatment    0.02703    0.04465  0.6055 0.544882 -0.06049  0.114558 7246
##
## Multiple R-squared:  3.267e-05 , Adjusted R-squared:  -0.0001053
## F-statistic: 0.3666 on 1 and 7246 DF,  p-value: 0.5449
##
## Call:
## lm_robust(formula = trend_diff ~ treatment, data = period_df)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept) -0.04876    0.01361  -3.582 0.0003433 -0.07544 -0.02207 7246
## treatment    0.09637    0.03676   2.622 0.0087648  0.02432  0.16843 7246
##
## Multiple R-squared:  0.0003312 , Adjusted R-squared:  0.0001933
## F-statistic: 6.874 on 1 and 7246 DF,  p-value: 0.008765
##
## Call:
## lm_robust(formula = trend_diff ~ treatment, data = period_df)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper   DF
## (Intercept) -0.004051    0.01338 -0.3027   0.7621 -0.030286  0.02218 7246
## treatment    0.063575    0.03437  1.8497   0.0644 -0.003801  0.13095 7246
##
## Multiple R-squared:  0.0001494 , Adjusted R-squared:  1.139e-05
## F-statistic: 3.421 on 1 and 7246 DF,  p-value: 0.0644
##
## Call:
## lm_robust(formula = trend_diff ~ treatment, data = period_df)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  0.05049    0.01355  3.7274 0.0001949  0.02394  0.07705 7246
## treatment   -0.02371    0.03948 -0.6005 0.5481831 -0.10109  0.05368 7246
##
## Multiple R-squared:  2.019e-05 , Adjusted R-squared:  -0.0001178
## F-statistic: 0.3606 on 1 and 7246 DF,  p-value: 0.5482
```

My method is to get trend_diff and treatment variable and then do a regression, where trend_diff is the change of number of patent of one subclass during the period, and treatment is a binary variable indicating if the subclass is exposed to treatment. Result analysis: (1918, 1917): if the subclass is exposed to treatment, then we should expect 0.027 more patents for this subclass. But the confidence interval includes 0, meaning that this value is not statistically significant. We reject the null that these effect differ from 0 (means there is no effect). Thus parallel trend holds here. (1918, 1916): if the subclass is exposed to treatment, then we should expect 0.096 more patents for this subclass. The confidence interval does not include 0, meaning that this value is statistically significant. We failed to reject the null that these effect differ from 0 (means there is effect). Thus parallel trend does not hold here. (1918, 1915): if the subclass is exposed to treatment, then we should expect 0.063 more patents for this subclass. But the confidence interval include 0, meaning that this value is not statistically significant.We reject the null that these effect differ from 0 (means there is no effect). Thus parallel trend holds here. (1918, 1914): if the subclass is exposed to treatment, then we should expect 0.023 less patents for this subclass. But the confidence interval includes 0, meaning that this value is not statistically significant. We reject the null that these effect differ from 0 (means there is no effect). Thus parallel trend holds here. Overall, if parallel trend holds here, we would expect the estimate to be 0. Thus although (1918,1916) has non-zero effect, overall the trend is parallel(no effect on other period)

##1.D

```
patent_1D <- patents %>% group_by(uspto_class) %>%
  summarize(
    change_for_patents = sum(count_for[grntyr >= 1919]) - sum(count_for[grntyr < 1919]),
    count_usa_post = sum(count_usa[grntyr >= 1919]),
    count_usa_pre = sum(count_usa[grntyr < 1919]),
    treatment = max(treat[grntyr == 1919])
  )

patent_1D <- patent_1D %>%
  mutate(strata = ntile(change_for_patents, 6))

results_1D <- data.frame()

for (stratum in unique(patent_1D$strata)) {

  stratum_data <- patent_1D %>% filter(strata == stratum)
  mean_diff_usa_treated <- mean(stratum_data$count_usa_post[stratum_data$treatment == 1] - stratum_data$
  mean_diff_usa_untreated <- mean(stratum_data$count_usa_post[stratum_data$treatment == 0] - stratum_dat

  treatment_effect <- mean_diff_usa_treated - mean_diff_usa_untreated
  n_treated <- sum(stratum_data$treatment == 1)
  n_untreated <- sum(stratum_data$treatment == 0)

  strata_variance_treated <- var(stratum_data$count_usa_post[stratum_data$treatment == 1] - stratum_data
  strata_variance_untreated <- var(stratum_data$count_usa_post[stratum_data$treatment == 0] - stratum_da
  variance <- strata_variance_treated + strata_variance_untreated
  count_strata <- sum(patent_1D$strata==stratum)

  results_1D <- rbind(results_1D, data.frame(strata = stratum, treatment_effect = treatment_effect, vari
}

results_1D <- results_1D %>%
  mutate(weight = count_strata / sum(results_1D$count_strata))

ate_block <- sum(results_1D$treatment_effect * results_1D$weight)

variance_ate_block <- sum(results_1D$variance * (results_1D$weight^2))

se_ate_block <- sqrt(variance_ate_block)

ci_lower <- ate_block - 1.96 * se_ate_block
ci_upper <- ate_block + 1.96 * se_ate_block

c(ATE = ate_block, Variance = variance_ate_block, SE = se_ate_block, CI_Lower = ci_lower, CI_Upper = ci_
```

```
##       ATE  Variance        SE  CI_Lower  CI_Upper
## 6.2125613 0.5695128 0.7546607 4.7334263 7.6916963
```

```
patent_1D <- patents %>%
  group_by(uspto_class) %>%
  summarize(
    change_for_patents = sum(count_for[grntyr >= 1919]) - sum(count_for[grntyr < 1919]),
    count_usa_post = sum(count_usa[grntyr >= 1919]),
    count_usa_pre = sum(count_usa[grntyr < 1919]),
    treatment = max(treat[grntyr == 1919])
  )

patent_1D <- patent_1D %>%
  mutate(strata = ntile(change_for_patents, 6))

results_1D_2 <- data.frame()
for (stratum in unique(patent_1D$strata)) {

  stratum_data <- patent_1D %>% filter(strata == stratum)

  lm_model <- lm((count_usa_post - count_usa_pre) ~ treatment, data = stratum_data)

  treatment_effect <- coef(lm_model)["treatment"]
  variance <- summary(lm_model)$coefficients["treatment", "Std. Error"]^2

  count_strata <- nrow(stratum_data)

  results_1D_2 <- rbind(results_1D_2, data.frame(strata = stratum, treatment_effect = treatment_effect,
}

results_1D_2 <- results_1D_2 %>%
  mutate(weight = count_strata / sum(count_strata))

ate_block <- sum(results_1D_2$treatment_effect * results_1D_2$weight)

variance_ate_block <- sum(results_1D_2$variance * (results_1D_2$weight^2))

se_ate_block <- sqrt(variance_ate_block)

ci_lower <- ate_block - 1.96 * se_ate_block
ci_upper <- ate_block + 1.96 * se_ate_block

c(ATE = ate_block, Variance = variance_ate_block, SE = se_ate_block, CI_Lower = ci_lower, CI_Upper = ci_
```

```
##       ATE  Variance        SE  CI_Lower  CI_Upper
## 6.2125613 0.9771737 0.9885210 4.2750601 8.1500624
```

We can successfully reject the null of no treatment effect at the alpha=0.05 level since the CI in both methods (simple mean difference and linear regression) does not include 0. This means that the treatment has a positive effect on number of patents granted for usa. This result has the same direction as question A, but significant larger magnitude. This difference might due to the control on foreign pantents. This suggest that the overall amount of foreign patenting in the sub class and its change over time might be a confounder.

Note that the variance is larger in linear regression method than simple mean difference method, this is because lm regression takes model uncertainty into account.

##2.A

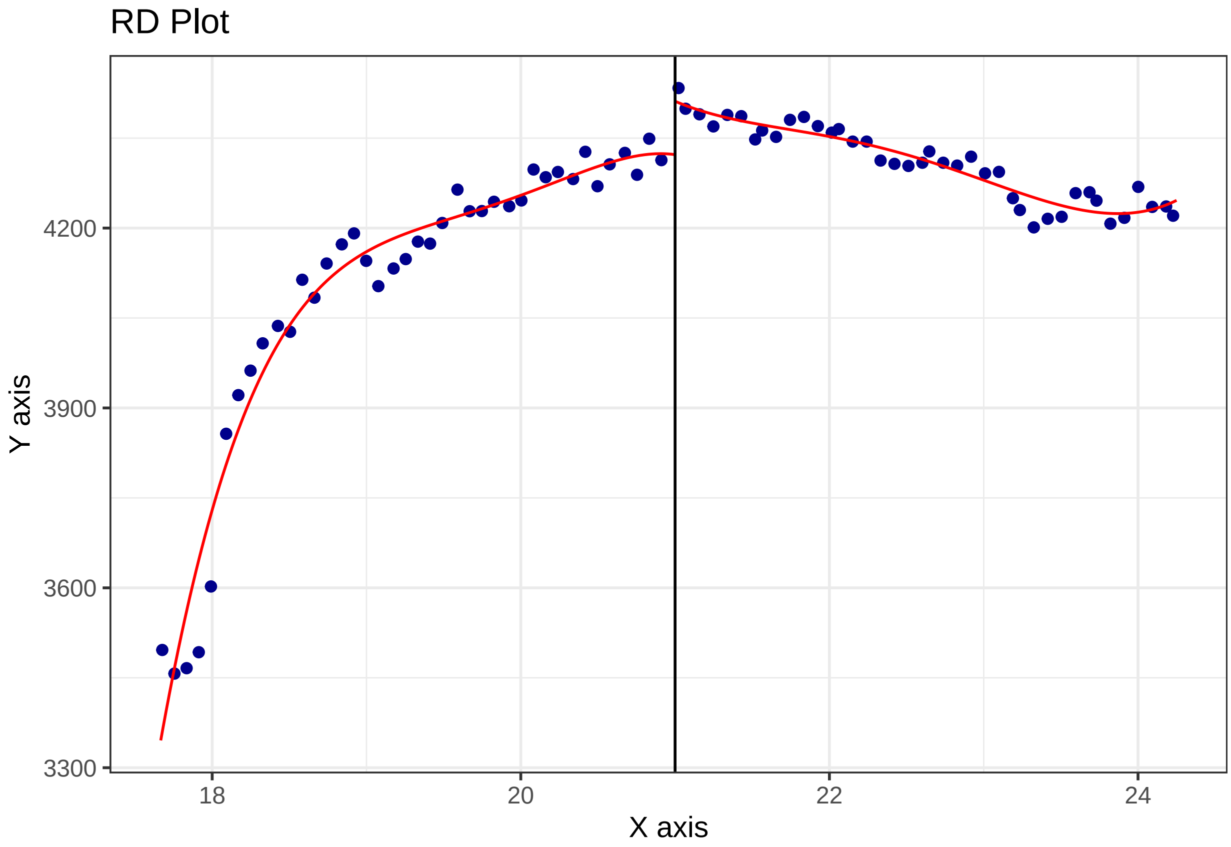
```
ER <- read.csv("ER.csv")
library(rdrobust)
outcomes <- c("ER$all", "ER$injury", "ER$alcohol")
bandwidths <- c(1, 0.5, 2)

results <- list()

for (outcome in outcomes) {
  for (h in bandwidths) {
    result_name <- paste("rdd", gsub("er\\$", "", outcome), h, sep = "_")

    results[[result_name]] <- rdrobust(y = eval(parse(text = outcome)), x = ER$age, c = 21, h = h)

    print(paste("Summary for", result_name))
    print(summary(results[[result_name]]))
  }
}
```

```
## [1] "Summary for rdd_ER$all_1"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              11           12
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   1.000        1.000
## BW bias (b)                   1.000        1.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    82.569    22.550     3.662     0.000    [38.372 , 126.765]
##         Robust         -         -     2.842     0.004    [29.010 , 157.978]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$all_0.5"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.               5            6
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.500        0.500
## BW bias (b)                   0.500        0.500
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    94.906    30.725     3.089     0.002    [34.686 , 155.127]
##         Robust         -         -     2.464     0.014    [24.339 , 213.562]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$all_2"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
```

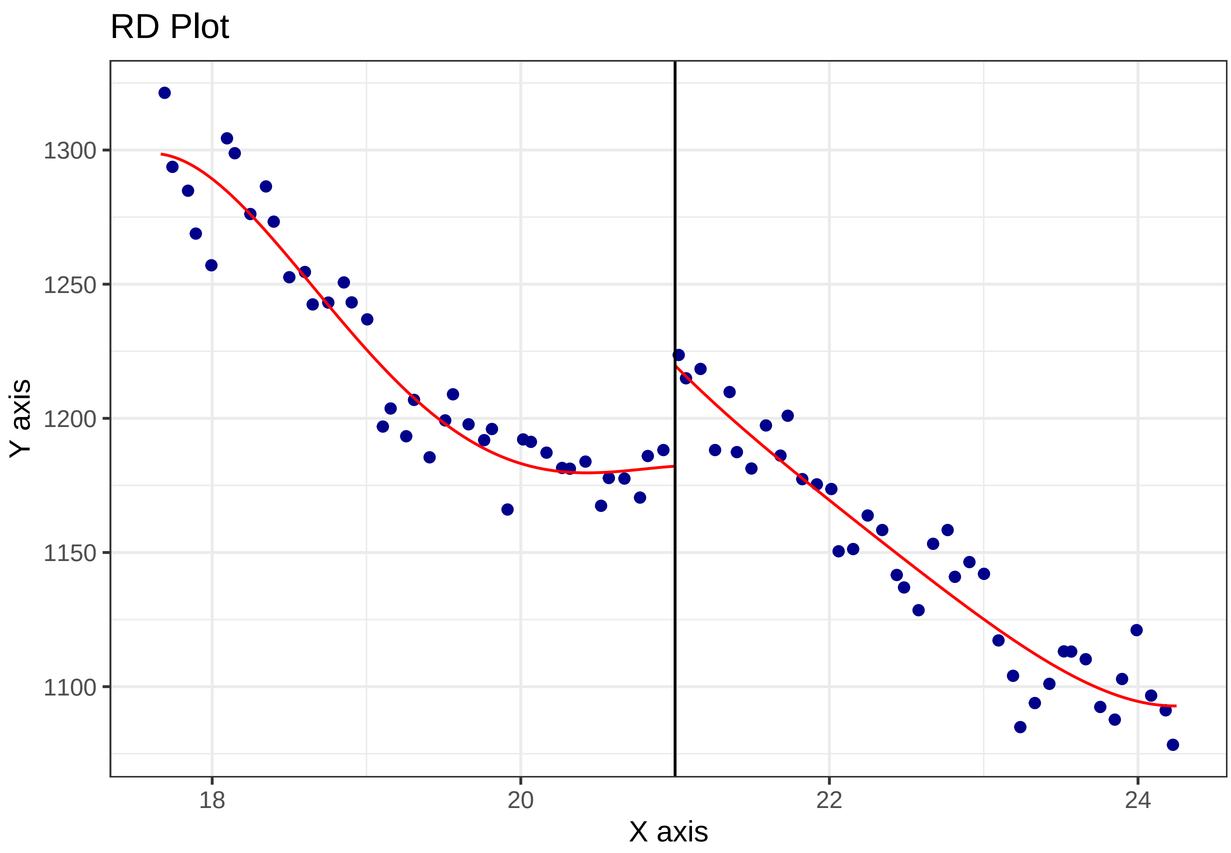
```
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              23           24
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   2.000        2.000
## BW bias (b)                   2.000        2.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    63.669    15.040     4.233     0.000    [34.190 , 93.147]
##         Robust         -         -     3.723     0.000    [40.826 , 131.589]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$injury_1"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              11           12
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   1.000        1.000
## BW bias (b)                   1.000        1.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    36.842     8.996     4.095     0.000    [19.211 , 54.474]
##         Robust         -         -     2.163     0.031     [2.760 , 56.173]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$injury_0.5"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.               5            6
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.500        0.500
## BW bias (b)                   0.500        0.500
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    31.845    13.232     2.407     0.016     [5.910 , 57.779]
##         Robust         -         -     0.961     0.337   [-23.493 , 68.648]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$injury_2"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              23           24
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   2.000        2.000
## BW bias (b)                   2.000        2.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    42.337     6.265     6.757     0.000    [30.057 , 54.617]
##         Robust         -         -     4.102     0.000    [20.094 , 56.861]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$alcohol_1"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              11           12
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   1.000        1.000
## BW bias (b)                   1.000        1.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    40.671    18.067     2.251     0.024     [5.261 , 76.080]
##         Robust         -         -     1.807     0.071    [-4.399 , 108.505]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$alcohol_0.5"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.               5            6
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.500        0.500
## BW bias (b)                   0.500        0.500
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    51.791    29.031     1.784     0.074    [-5.109 , 108.690]
##         Robust         -         -     1.691     0.091   [-10.610 , 143.945]
## =============================================================================
## NULL
## [1] "Summary for rdd_ER$alcohol_2"
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   80
## BW type                      Manual
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   40           40
## Eff. Number of Obs.              23           24
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   2.000        2.000
## BW bias (b)                   2.000        2.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      40           40
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    32.259    10.203     3.162     0.002    [12.261 , 52.257]
##         Robust         -         -     2.191     0.028     [4.167 , 74.971]
## =============================================================================
## NULL
```

"all" outcome variable seems to have the largest effect. The change of bandwidth influence results by a non-negligible amounts. ##2.B

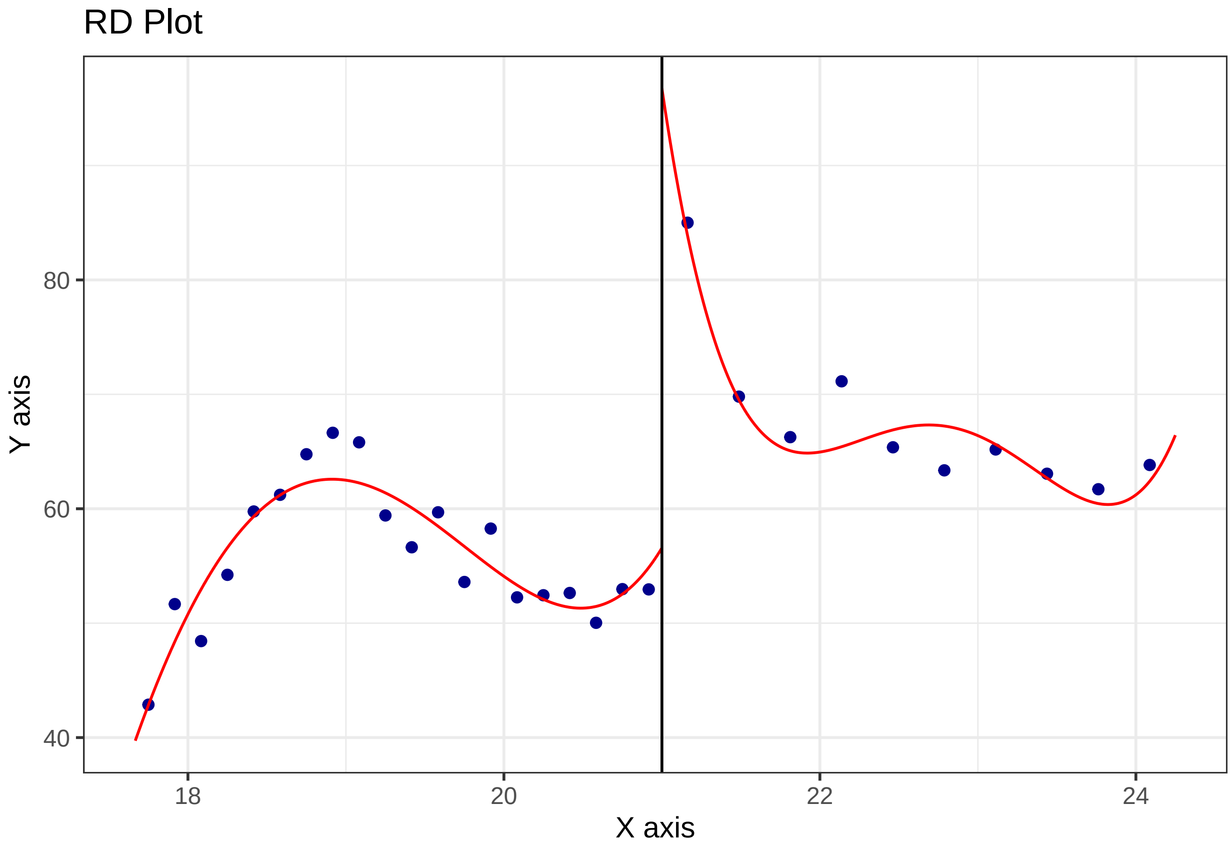
```
library(rdrobust)
rdplot(y = ER$all, x = ER$age, c = 21)
```



```
rdplot(y = ER$injury, x = ER$age, c = 21)
```



```r
rdplot(y = ER$alcohol, x = ER$age, c = 21)
```

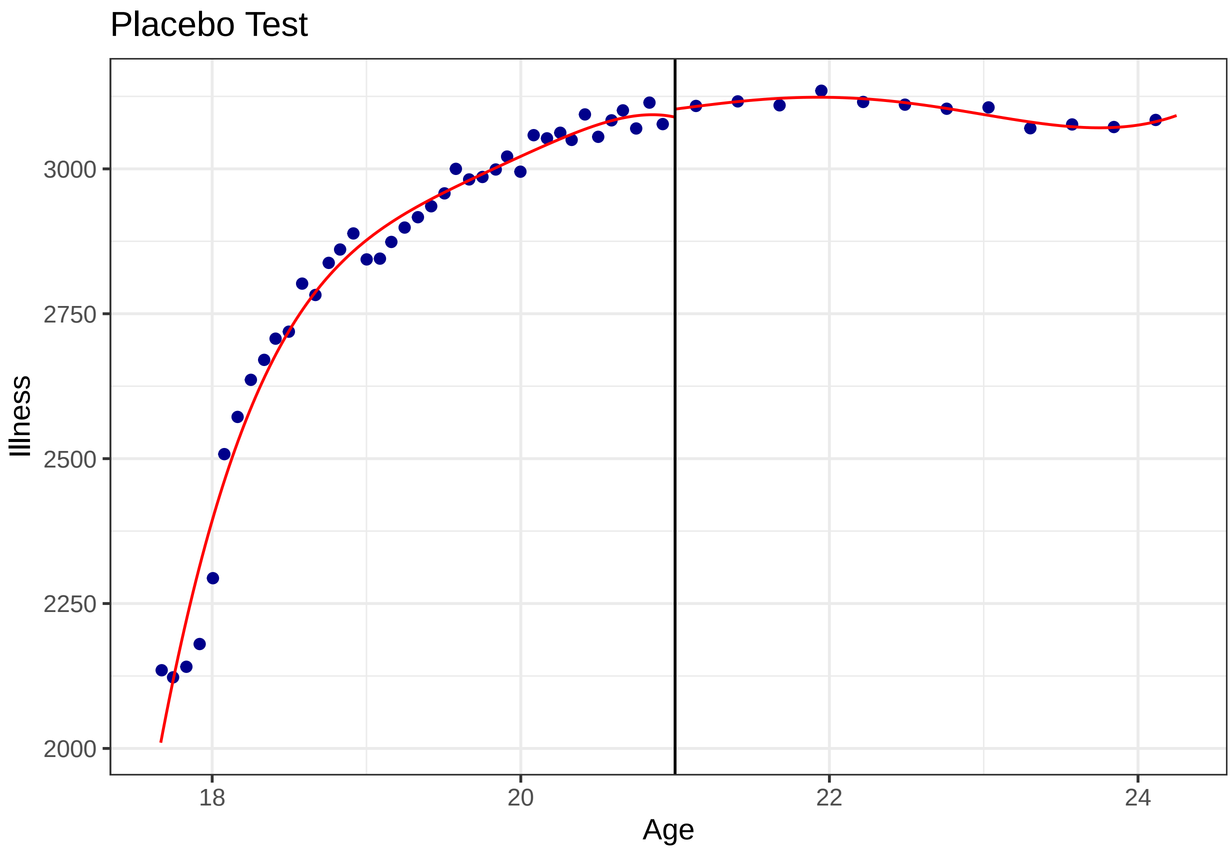


##2.C

```
rdd_2C_illness <- RDestimate(illness ~ age, data = ER, cutpoint = 21,bw=1)
summary(rdd_2C_illness)
```

```
##
## Call:
## RDestimate(formula = illness ~ age, data = ER, cutpoint = 21,
##     bw = 1)
##
## Type:
## sharp
##
## Estimates:
##            Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## LATE       1.0        23             7.487    15.38        0.4868  0.6264
## Half-BW    0.5        11            15.061    19.24        0.7829  0.4337
## Double-BW  2.0        47            -8.934    12.54       -0.7123  0.4763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##            F       Num. DoF  Denom. DoF  p
## LATE        9.176  3         19          0.001157
## Half-BW     1.942  3          7          0.422824
## Double-BW  92.518  3         43          0.000000
```

```r
rdplot(y = ER$illness, x = ER$age, c = 21,
       title = "Placebo Test",
       y.label = "Illness",
       x.label = "Age")
```



The placebo test is not statistically significant with p-values significantly larger than 0.05 for all different bandwidth. This means that the RDD is plausible that the age does have effect on all, injury and alcohol.

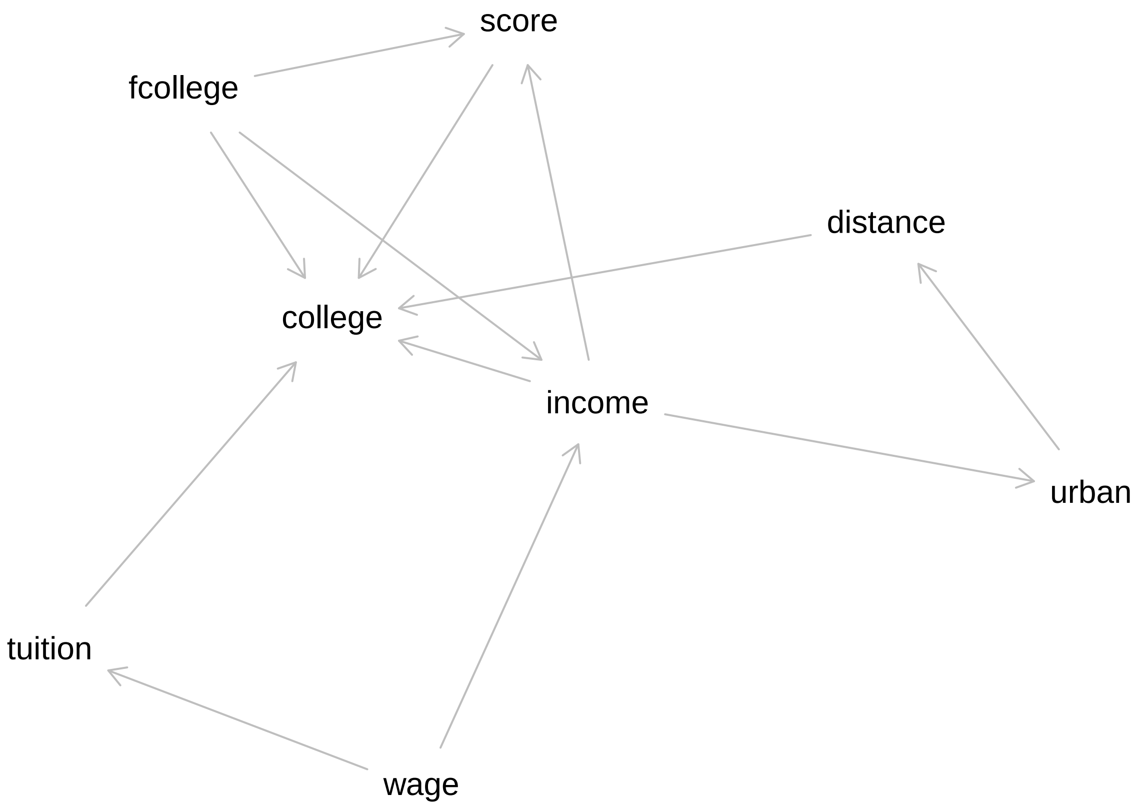
##3.A

```r
library(ggdag)
```

```
##
## Attaching package: 'ggdag'

## The following object is masked from 'package:stats':
##
##     filter
```

```r
dag <- dagify(
  college ~ income + score + distance + fcollege + tuition,
  score ~ income + fcollege,
  tuition ~ wage,
  urban ~ income,
  distance ~ urban,
  income ~ fcollege+wage,
  exposure = "income",
  outcome = "college"
)

plot(dag)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set your
```



income->score: the richer the family, more recources a student has to improve score. score->college: higher score means that the student is more outstanding, thus more likely to be accepted to college. Income->urban:the houses in the urban area are typically more expensive. Urban->distance to college distance->college. Since students might prefer colleges closer to their family. father's college->family income: because people who have higher degree tend to have higher salary. father's college->impact score: since if father has college degree, they can help their child to improve grades Father's college->college: if the father think the college education is useless, then they will not let their children to go to college, vice versa. tuition->college: if it is way too expensive, then the student might not be able to afford it then failed to go to college Wage->tuition: wage can be treated as the benchmark of local consumption level or income level, which can influence tuition wage->income, since family income is based on the local income level. income->college: the treatment effect we want to figure out, intuitively, as income increase, students tend to have higher chance attending college.

We only need to condition on fcollege and wage.By conditioning on these two variables, all back door paths are blocked

3.B I want to use stratified ATE estimator, since we need to condition on fcollege and wage, which can be achieved in CATE. Besides conditional ignorability, we need conditional positivity, consistency and discreteness of covariates. These assumptions hold because there should be non-zero and less than 1 probability of going to college, meaning that a high school student has the chance of attending college (not 0) but not definitely (not 1).For discreteness of covariates, we can categorize wage by a range, which makes it discrete. For fcollege, it is obviously discrete. For variance calculation, we can use estimator for stratified ATE.

3.C

```r
college <- read.csv("college.csv")

college <- college %>%
  mutate(wage_quartile = ntile(wage, 4))

CATE_estimates <- college %>%
  group_by(wage_quartile, fcollege) %>%
  summarize(
    CATE = mean(college[income == 1]) - mean(college[income == 0]),
    variance = var(college[income == 1])/sum(income == 1) + var(college[income == 0])/sum(income == 0),
    count_treated = sum(income == 1),
    count_untreated = sum(income == 0),
    .groups = 'drop')

overall_ATE <- CATE_estimates %>%
  mutate(weight = (count_treated + count_untreated) / sum(count_treated + count_untreated)) %>%
  summarize(ATE = sum(CATE * weight),
            variance_ATE = sum(variance * (weight^2)),
            .groups = 'drop')

se_ATE <- sqrt(overall_ATE$variance_ATE)
ci_lower <- overall_ATE$ATE - 1.96 * se_ATE
ci_upper <- overall_ATE$ATE + 1.96 * se_ATE

c(ATE = overall_ATE$ATE,
  SE = se_ATE,
  CI_Lower = ci_lower,
  CI_Upper = ci_upper)
```

```
##        ATE         SE   CI_Lower   CI_Upper
## 0.12437890 0.01641511 0.09220529 0.15655252
```

We see that the treatment has a statistically significant positive treatment effect since the CI does not contain 0. This means that if the family income is higher than 25k per year, the childen are more likely to attend to college.

4.A

```r
nazis <- read.csv("nazis.csv")

nazis <- nazis %>%
  mutate(vote_portion = nazivote / nvoter)

model <- lm(vote_portion ~ shareblue, data = nazis)

summary(model)
```

```
##
## Call:
## lm(formula = vote_portion ~ shareblue, data = nazis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30151 -0.07133 -0.00092  0.06986  0.33037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39558    0.01661  23.812   <2e-16 ***
## shareblue    0.06518    0.05220   1.249    0.212
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.108 on 679 degrees of freedom
## Multiple R-squared:  0.002291,   Adjusted R-squared:  0.0008218
## F-statistic: 1.559 on 1 and 679 DF,  p-value: 0.2122
```

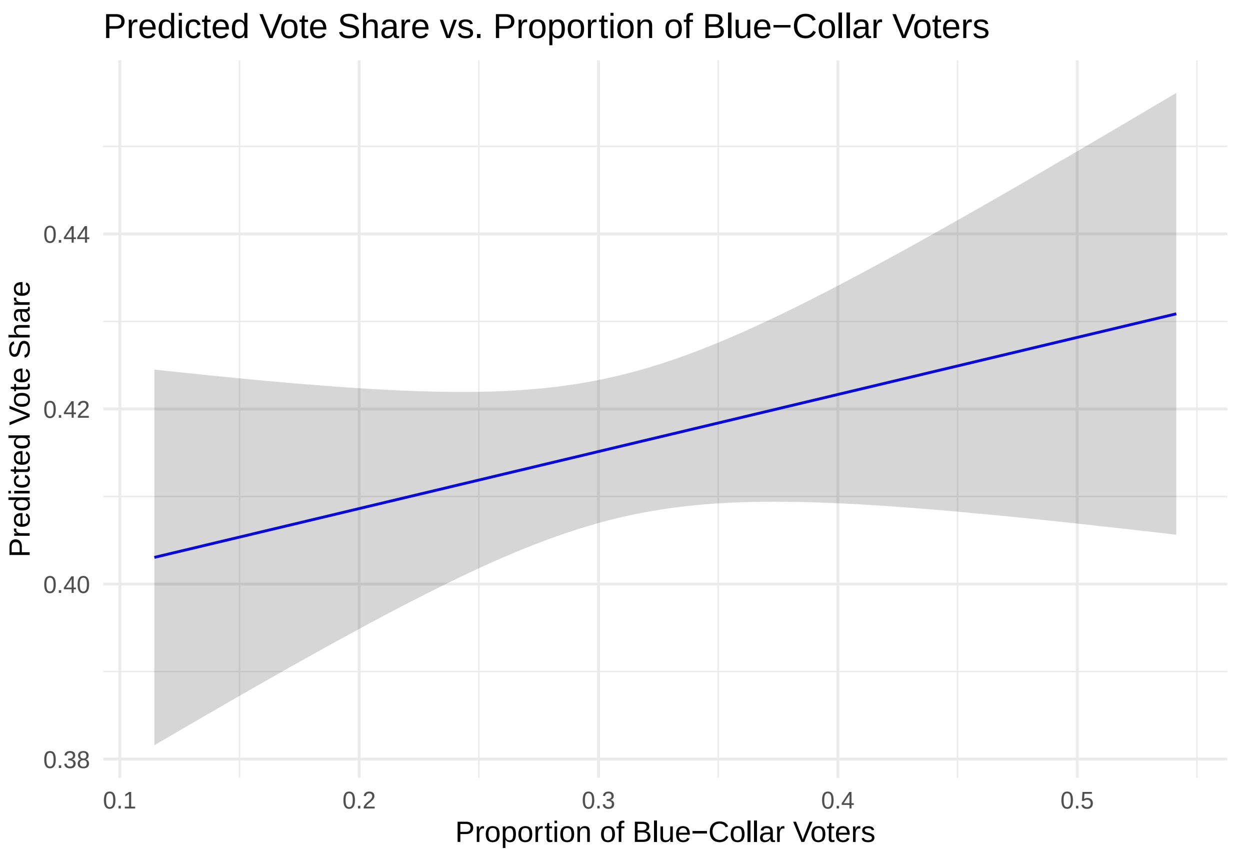
```r
confint(model, level =0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept)  0.36296607 0.4282031
## shareblue   -0.03730872 0.1676687
```

The slope coefficient 0.06518 suggests a positive relationship between the proportion of blue-collar voters and the Nazi vote share, but this relationship is not statistically significant (p-value = 0.212), with a CI including 0. For shareblue, a standard error of 0.05220 indicates that the estimated slope of 0.06518 could flucutates by approximately ±0.05220 in repeated samples. CI means that with a 95% confidence, the true value of blue collar workers’ impact on Nazi vote share lies in this range, where 0 is a plausible value.

4.B

```r
library(ggplot2)
X_seq <- seq(min(nazis$shareblue), max(nazis$shareblue), length.out = 100)
predictions <- predict(model, newdata = data.frame(shareblue = X_seq), interval = "confidence")
pred_df <- data.frame(
  shareblue = X_seq,
  fit = predictions[, "fit"],
  lwr = predictions[, "lwr"],
  upr = predictions[, "upr"]
)

ggplot(pred_df, aes(x = shareblue, y = fit)) +
  geom_line(color = "blue") +
  geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.2) +
  labs(title = "Predicted Vote Share vs. Proportion of Blue-Collar Voters",
       x = "Proportion of Blue-Collar Voters",
       y = "Predicted Vote Share") +
  theme_minimal()
```



This plot shows a positive relationship between proportion of blue collar voters and Nazis vote share. This means that if a precinct has more blue collar workers, it tends to vote for Nazis. ##4.C

```
nazis <- nazis %>%
  mutate(minused = (1-shareblue))
model <- lm(vote_portion ~ 0 + shareblue + minused, data = nazis)
summary(model)
```

```
##
## Call:
## lm(formula = vote_portion ~ 0 + shareblue + minused, data = nazis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30151 -0.07133 -0.00092  0.06986  0.33037
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## shareblue  0.46076    0.03635   12.68   <2e-16 ***
## minused    0.39558    0.01661   23.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.108 on 679 degrees of freedom
## Multiple R-squared:  0.937,  Adjusted R-squared:  0.9368
## F-statistic:  5047 on 2 and 679 DF,  p-value: < 2.2e-16
```

alpha = beta^star; beta = alpha^star - beta^star alpha^star gives the vote share for Nazis where all voters are blue collar workers, which is 46%. beta^star gives the vote share for Nazis where all voters are non blue collar workers, which is 39%. Both of them are statistically significant with a small p-value. This suggests that blue collar workers are more likely to vote for Nazis.

4.D

```
model <- lm(vote_portion ~ 0 + shareself + shareblue + sharewhite + sharedomestic + shareunemployed, da
summary(model)
```

```
##
## Call:
## lm(formula = vote_portion ~ 0 + shareself + shareblue + sharewhite +
##     sharedomestic + shareunemployed, data = nazis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28271 -0.06847 -0.00055  0.06790  0.32369
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## shareself        1.11426    0.16677   6.681 4.95e-11 ***
## shareblue        0.54038    0.03848  14.042  < 2e-16 ***
## sharewhite       0.28509    0.07501   3.801 0.000157 ***
## sharedomestic    0.05221    0.09120   0.572 0.567181
## shareunemployed -0.02816    0.07014  -0.401 0.688202
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1024 on 676 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9431
## F-statistic:  2259 on 5 and 676 DF,  p-value: < 2.2e-16
```

Self: if all people in a precinct are self employed, then the Nazis vote share would be 111%. This estimate is statistically significant due to a p-value less than 0.05. blue: if all people in a precinct are blue collar workers, then the Nazis vote share would be 54%. This estimate is statistically significant due to a p-value less than 0.05. white: if all people in a precinct are white collar workers, then the Nazis vote share would be 28.5%. This estimate is statistically significant due to a p-value less than 0.05. domestic: if all people in a precinct are domestically employed, then the Nazis vote share would be 5.2%. This estimate is not statistically significant due to a p-value larger than 0.05. unemployed: if all people in a precinct are unemployed, then the Nazis vote share would be -2.8%. This estimate is not statistically significant due to a p-value larger than 0.05. Assumptions: no correlation between each variable. Linear relationship between IVs and DV.

4.E

```
nazis <- nazis %>% mutate(Y = shareself + sharewhite + sharedomestic + shareunemployed)
nazis <- nazis %>%
  mutate(
    W_i1_min = pmax(0, (vote_portion - (1 - shareblue)) / shareblue),
    W_i1_max = pmin(1, vote_portion / shareblue)
  )

nazis <- nazis %>%
  mutate(blue_collar_voters = shareblue * nvoter)
```

```r
nation_min <- sum(nazis$W_i1_min * nazis$blue_collar_voters) / sum(nazis$blue_collar_voters)
nation_max <- sum(nazis$W_i1_max * nazis$blue_collar_voters) / sum(nazis$blue_collar_voters)

c(
  Nation_min = nation_min,
  Nation_max = nation_max
)
```

```
##   Nation_min   Nation_max
## 0.0002675665 0.9495676723
```

The nationwide min is close to 0, suggesting that when all non blue collar workers vote for Nazis, only 0.03% of blue collar workers will vote for Nazis. The nationwide max is close to 1, suggesting that when all non blue collar workers do not vote for Nazis, 95% of blue collar workers will vote for Nazis. This wide range suggests that it is possible that non or all of blue collar workers will vote for Nazis under extreme cases, showing the complexity of ecological inference.

5.A

```r
wage <- read.csv("wage2.csv")
model_5A <- lm_robust(wage ~ educ, data = wage)
summary(model_5A)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ, data = wage)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   146.95     80.335   1.829 6.768e-02   -10.71   304.61 933
## educ           60.21      6.163   9.771 1.551e-21    48.12    72.31 933
##
## Multiple R-squared:  0.107 , Adjusted R-squared:  0.106
## F-statistic: 95.47 on 1 and 933 DF,  p-value: < 2.2e-16
```

This effect shows that when education increase by one year, the wage is likely to increase by 60 usd. This effect is statistically significant becasue of a p-value less than 0.05 and a CI not including 0.

However, this might not be able to account for all causal effect since there are omit variable bias that the naive regression cannot explain the effect of other covariates and confounders. For example, social connection might impact wage as well. Also, one's IQ might influence both year of education and wage, which is a confounder. However, naive regression failed to account for it.

5.B

```r
relevance <- lm(educ ~ feduc, data = wage)
summary(relevance)
```

```
##
## Call:
```

```
## lm(formula = educ ~ feduc, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1313 -1.5510 -0.3904  1.8687  5.8997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.64956    0.24272   43.88   <2e-16 ***
## feduc        0.29014    0.02261   12.83   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.03 on 739 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.1823, Adjusted R-squared:  0.1812
## F-statistic: 164.7 on 1 and 739 DF,  p-value: < 2.2e-16
```

```r
non_direct <- lm(IQ ~ feduc, data = wage)
summary(non_direct)
```

```
##
## Call:
## lm(formula = IQ ~ feduc, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.732  -8.434   0.268  10.119  40.268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.2854     1.6577  52.051   <2e-16 ***
## feduc         1.5372     0.1544   9.956   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13.86 on 739 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.1183, Adjusted R-squared:  0.1171
## F-statistic: 99.13 on 1 and 739 DF,  p-value: < 2.2e-16
```

```r
wage$predicted_IQ <- predict(non_direct, wage)
second_stage <- lm(wage ~ predicted_IQ, data = wage)
summary(second_stage)
```

```
##
## Call:
## lm(formula = wage ~ predicted_IQ, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -883.79 -278.01  -49.36  213.64 2049.99
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -962.527    291.896  -3.297  0.00102 ** 
## predicted_IQ   19.006      2.858   6.649 5.73e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 394.5 on 739 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.05645,    Adjusted R-squared:  0.05517 
## F-statistic: 44.21 on 1 and 739 DF,  p-value: 5.733e-11
```

Relevance: We can see that father’s education has a strong influence in children’s education. When father’s year of education increase by one year, the children’s education increase by 0.3 years. This effect is statiscally significant due to a p-value less than 0.05. Exclusion: if for this question, other variables in the dataset like IQ are not our interest, and we only focus on feduc, meduc, educ, and wage, then exclusion holds, since feduc will not impact other independent variable and there will not be any additional paths to dependent variable. feduc->edu->wage is the only path. However, if we take IQ into account, feduc has a positive statistically significant effect on IQ, and IQ has a significant effect on wage, exclusion not hold in this case. Exogeneity: parental education is determined before children birth, thus will not impact other unobserved variables like children’s personal motivation, being exogenous to child’s decisions and all other factors.

5C

```r
wage$predicted_educ <- predict(relevance, wage)
second_stage <- lm(wage ~ predicted_educ, data = wage)
summary(second_stage)
```

```
##
## Call:
## lm(formula = wage ~ predicted_educ, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -883.79 -278.01  -49.36  213.64 2049.99 
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -394.98     206.69  -1.911   0.0564 .  
## predicted_educ    100.70      15.14   6.649 5.73e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 394.5 on 739 degrees of freedom
##   (194 observations deleted due to missingness)
## Multiple R-squared:  0.05645,    Adjusted R-squared:  0.05517 
## F-statistic: 44.21 on 1 and 739 DF,  p-value: 5.733e-11
```

As education increase by one unit, the wage will increase by 100.70. This effect is statistically significant because of the small p-value. 5D

```r
model_5D <- iv_robust(wage ~ educ | feduc, data = wage)
summary(model_5D)
```

```
##
## Call:
## iv_robust(formula = wage ~ educ | feduc, data = wage)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -395.0     212.38  -1.860 6.332e-02  -811.93    21.96 739
## educ           100.7      15.77   6.385 3.020e-10    69.74   131.66 739
##
## Multiple R-squared:  0.04254 ,   Adjusted R-squared:  0.04124
## F-statistic: 40.77 on 1 and 739 DF,  p-value: 3.02e-10
```

We see that the point estimates for both methods are the same. However, there are slight difference in standard errors. In iv_robust, the standard errors of intercept and educ coefficient are both higher than the lm std. The iv_robust one is more plausible because it adjusts to heteroskedasticity. Thus in the case where variance of errors are different within different units, vanilla lm might be overconfident with the estimate, ie smaller std and thus narrower CI.