Jerry_Huang_Final

Jerry Huang

2024-08-14

Problem 1 - Does access to foreign inventions make domestic firms more innovative?

Question A:

```r
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(estimatr)
library(dplyr)

# Read in the data
patents <- read.csv("patents.csv")

# Drop units with missing data
patents <- patents %>%
  filter(!is.na(uspto_class) & !is.na(grntyr) & !is.na(count_usa) & !is.na(count_for))

aggre_data <- patents %>%
  group_by(uspto_class) %>%
  summarize(
    pre = mean(count_usa[grntyr < 1919]),
    post = mean(count_usa[grntyr >= 1919]),
    treatment = max(treat)
  )

patents_diff <- aggre_data %>%
  mutate(
    did = post - pre
  )
```

```r
# Get the DiD estimate using lm_robust with robust standard errors
DiD_result <- lm_robust(did ~ treatment, data = patents_diff)

# Display the results
summary(DiD_result)
```

```
##
## Call:
## lm_robust(formula = did ~ treatment, data = patents_diff)
##
## 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
```

Answer:
The point estimate is **0.2553**, with a 95% confidence interval of **(0.1814, 0.3292)** which indicates statistically significant. The DiD estimate of 0.2553 indicates that, on average, the treatment (compulsory licensing) is associated with an increase of about 0.2553 patents in the treated subclasses compared to the control subclasses. The **p-value is 1.352e-11**, which further indicating a statistical significant as it is smaller than the significant level of 0.05. Therefore, we do reject the null hypothesis of no treatment effect at the 0.05 level.

Question B:

```r
# Exposed to treatment or not
exposure_data <- patents %>%
  group_by(uspto_class) %>%
  summarise(
    exposed = ifelse(any(treat == 1 & grntyr >= 1919), 1, 0)
  )

patents_with_exposure <- patents %>%
  left_join(exposure_data, by = "uspto_class")

# Exposed + post treatment period
post_treatment_data <- patents_with_exposure %>%
  filter(grntyr >= 1919)

# Calculate average count of U.S. patents post treatment
avg_patents_post_1918 <- post_treatment_data %>%
  group_by(exposed) %>%
  summarise(avg_count_usa = mean(count_usa, na.rm = TRUE))

pre_treatment_data <- patents_with_exposure %>%
  filter(grntyr < 1919)

# Calculate average count of U.S. patents pre treatment
avg_patents_pre_1919 <- pre_treatment_data %>%
  group_by(exposed) %>%
```

```r
  summarise(avg_count_usa = mean(count_usa, na.rm = TRUE))

print(avg_patents_pre_1919)
```

```
## # A tibble: 2 x 2
##   exposed avg_count_usa
##     <dbl>         <dbl>
## 1       0        0.228
## 2       1        0.0827
```

Ignorability is not likely to hold under the strategy of just just comparing the average differences in the count of US patents in the post-1918 period between exposed and unexposed sub-classes to estimate the treatment effect. According to the counts in the pre-treatment period between exposed and unexposed subclasses, we can see that the average number of patent in treatment group is 0.08272457, which is not similar to the average number of patents in control group, which is 0.22787116. Therefore, we can say that the ignorability is not likely to hold due to the significant differences between the groups before 1919, suggesting that these groups were inherently different. The significant difference in pre-1919 patent counts between the exposed and unexposed groups suggests that these groups were not comparable before the treatment.

Question C:

```r
yearly_data <- patents_with_exposure %>%
  filter(grntyr %in% c(1914, 1915, 1916, 1917, 1918)) %>%
  group_by(uspto_class, exposed, grntyr) %>%
  summarise(avg_count_usa = mean(count_usa, na.rm = TRUE), .groups = "drop")

yearly_diff <- yearly_data %>%
  pivot_wider(names_from = grntyr, values_from = avg_count_usa) %>%
  mutate(
    diff_1918_1917 = `1918` - `1917`,
    diff_1918_1916 = `1918` - `1916`,
    diff_1918_1915 = `1918` - `1915`,
    diff_1918_1914 = `1918` - `1914`
  )

model_1918_1917 <- lm_robust(diff_1918_1917 ~ exposed, data = yearly_diff)
model_1918_1916 <- lm_robust(diff_1918_1916 ~ exposed, data = yearly_diff)
model_1918_1915 <- lm_robust(diff_1918_1915 ~ exposed, data = yearly_diff)
model_1918_1914 <- lm_robust(diff_1918_1914 ~ exposed, data = yearly_diff)

summary(model_1918_1917)
```

```
##
## Call:
## lm_robust(formula = diff_1918_1917 ~ exposed, data = yearly_diff)
##
## 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
## exposed      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
```

```r
summary(model_1918_1916)
```

```
##
## Call:
## lm_robust(formula = diff_1918_1916 ~ exposed, data = yearly_diff)
##
## 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
## exposed      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
```

```r
summary(model_1918_1915)
```

```
##
## Call:
## lm_robust(formula = diff_1918_1915 ~ exposed, data = yearly_diff)
##
## 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
## exposed      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
```

```r
summary(model_1918_1914)
```

```
##
## Call:
## lm_robust(formula = diff_1918_1914 ~ exposed, data = yearly_diff)
##
## 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
## exposed     -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
```

Answer:
We reject the null hypothesis that there is no difference for the average patents between 1918 and 1916. The coefficient for exposed is statistically significant (p-value < 0.05), and the confidence interval does not include zero (0.02432, 0.16843). This suggests that there is a significant difference in trends between the exposed and unexposed groups from 1916 to 1918, which indicates a potential violation of the parallel trends assumption. For the comparisons of 1918 to 1917 and 1918 to 1914, the parallel trends assumption seems reasonable since the differences between exposed and unexposed groups are not statistically significant

(confidence intervals include zero). The results suggest that the parallel trends assumption is potentially violated, especially when considering the 1916 to 1918 period. This violation would cast doubt on the validity of the difference-in-differences estimate from Question A, as it relies heavily on the assumption of parallel trends.

Question D:

```r
patents_with_exposure <- patents %>%
  group_by(uspto_class) %>%
  summarise(
    pre_foreign = mean(count_for[grntyr < 1919], na.rm = TRUE),
    post_foreign = mean(count_for[grntyr >= 1919], na.rm = TRUE),
    change_foreign = post_foreign - pre_foreign
  )

patents_with_exposure <- patents_with_exposure %>%
  mutate(strata = ntile(change_foreign, 6))

patents_with_strata <- patents %>%
  left_join(patents_with_exposure, by = "uspto_class")

results_stratified <- data.frame()

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

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

  lm_model <- lm_robust(count_usa ~ treat + grntyr + uspto_class, data = stratum_data)

  treatment_effect <- coef(lm_model)["treat"]
  variance <- vcov(lm_model)["treat", "treat"]

  count_strata <- nrow(stratum_data)

  results_stratified <- rbind(results_stratified,
                              data.frame(strata = stratum,
                                         treatment_effect = treatment_effect,
                                         variance = variance,
                                         count_strata = count_strata))
}

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

ate_stratified <- sum(results_stratified$treatment_effect * results_stratified$weight)
variance_ate_stratified <- sum(results_stratified$variance * (results_stratified$weight^2))

se_ate_stratified <- sqrt(variance_ate_stratified)
ci_lower <- ate_stratified - 1.96 * se_ate_stratified
ci_upper <- ate_stratified + 1.96 * se_ate_stratified

print(data.frame(ATE = ate_stratified, SE = se_ate_stratified, CI_Lower = ci_lower, CI_Upper = ci_upper
```

```
##         ATE         SE   CI_Lower  CI_Upper
## 1 0.1594143 0.03644501 0.08798212 0.2308466
```

The ATE is 0.159 which indicates that, after accounting for the foreign patent change stratification, the treatment (compulsory licensing) is associated with an increase of approximately 0.159 patents in the treated subclasses compared to the control subclasses. The positive sign suggests that the treatment had a positive impact on domestic patenting. The 95% confidence interval is (0.088, 0.231) and it does not include zero, which suggests that the treatment effect is statistically significant at the 5% level. In other words, we can be 95% confident that the true effect of the treatment lies between 0.088 and 0.231. The results indicate a statistically significant and positive impact of compulsory licensing on domestic patenting, even after adjusting for changes in foreign patents by stratifying the analysis. The confidence interval is relatively narrow, which adds to the robustness of the findings.

In Question A, the DiD point estimate is 0.2553, with a 95% confidence interval of (0.1814, 0.3292), which is larger than what we have found here. The difference is because the stratified ATE applies equal weights to the treatment effect in each stratum, rather than allowing the treatment effect to be driven by differences in sample sizes between groups. Additionally, stratification can introduce more variability into the estimate, especially in smaller strata, leading to a more conservative (i.e., smaller) effect estimate and wider standard errors. Moreover, the stratified ATE assumes that the treatment effect is consistent across strata, which may not fully capture the heterogeneity in treatment effects that could exist if the foreign patent changes have different impacts depending on the characteristics of the subclasses.

Problem 2 - RDD

Question A:

```r
library(rdrobust)
er <- read_csv('ER.csv', show_col_types = FALSE)

rdd_all_1 <- rdrobust(y = er$all, x = er$age, c = 21, h = 1)
rdd_all_0.5 <- rdrobust(y = er$all, x = er$age, c = 21, h = 0.5)
rdd_all_2 <- rdrobust(y = er$all, x = er$age, c = 21, h = 2)
summary(rdd_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]
## =============================================================================
```

```r
summary(rdd_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]
## =============================================================================
```

```r
summary(rdd_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]
## =============================================================================
```

```r
rdd_injury_1 <- rdrobust(y = er$injury, x = er$age, c = 21, h = 1)
rdd_injury_0.5 <- rdrobust(y = er$injury, x = er$age, c = 21, h = 0.5)
rdd_injury_2 <- rdrobust(y = er$injury, x = er$age, c = 21, h = 2)
```

```r
summary(rdd_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]
## =============================================================================
```

```r
summary(rdd_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]
## =============================================================================
```

```r
summary(rdd_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]
## =============================================================================
```

```
rdd_alcohol_1 <- rdrobust(y = er$alcohol, x = er$age, c = 21, h = 1)
rdd_alcohol_0.5 <- rdrobust(y = er$alcohol, x = er$age, c = 21, h = 0.5)
rdd_alcohol_2 <- rdrobust(y = er$alcohol, x = er$age, c = 21, h = 2)
summary(rdd_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]
## =============================================================================
```

```
summary(rdd_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]
## =============================================================================
```

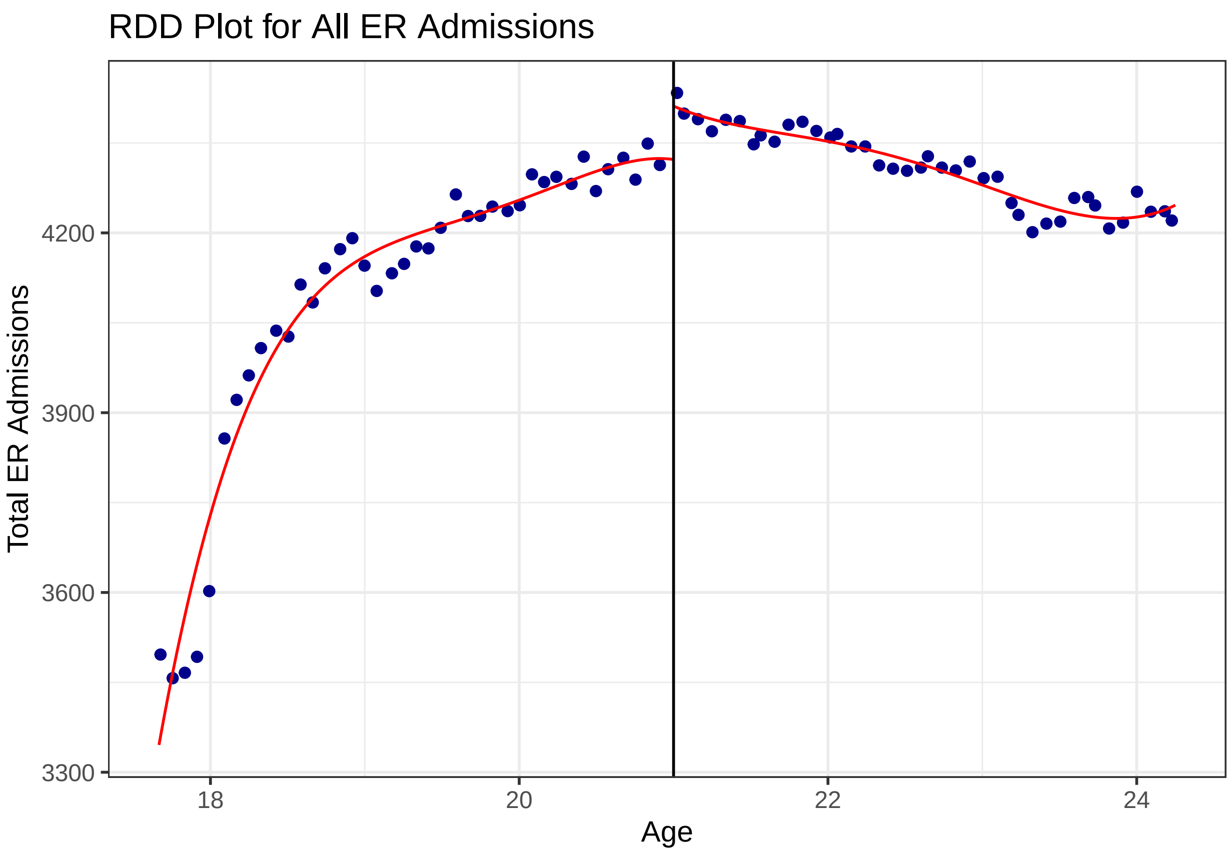
```r
summary(rdd_alcohol_2)
```

```
## 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]
## =============================================================================
```

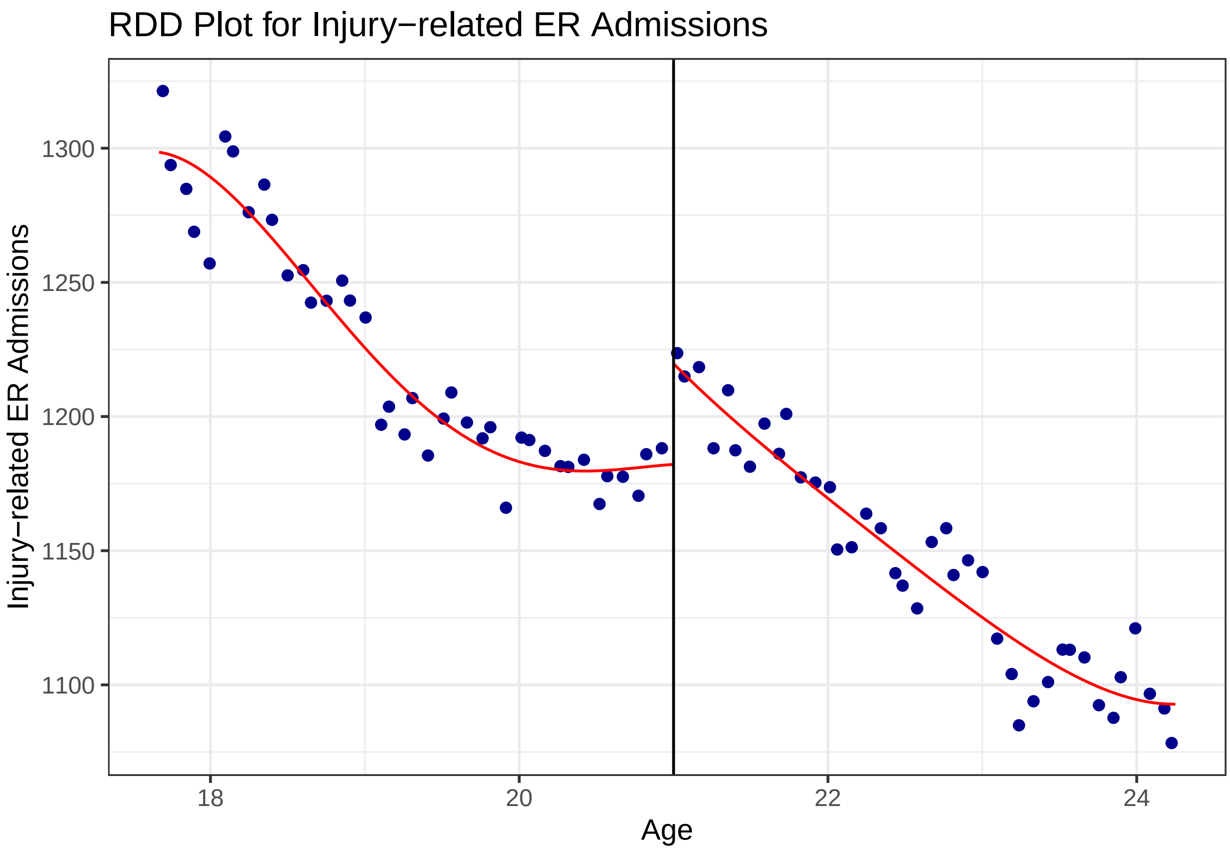
Answer:
The results show a consistent positive impact of turning 21 on ER admissions, particularly for all ER visits and all variable. However, the results are sensitive to bandwidth choice, and the robust estimates sometimes show different trends compared to the conventional estimates, particularly for the alcohol-related outcomes. The variability in the estimates across different bandwidths suggests that the relationship might not be entirely linear or that there are underlying factors influencing the discontinuity. The most significant and consistent effects are observed for *all* variable, which makes sense given the highest coefficent value of all 3 bandwidths.

Question B:

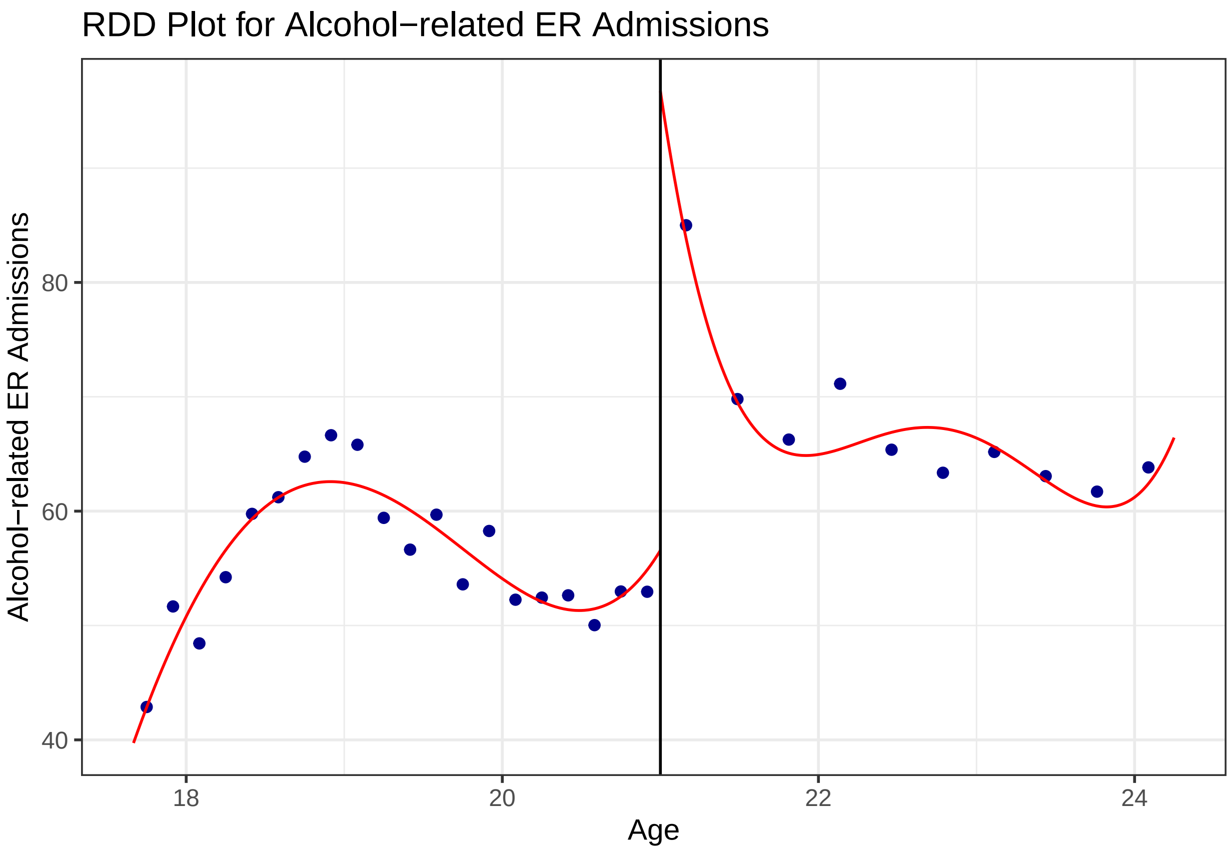
```r
rdplot(y = er$all, x = er$age, c = 21, title = "RDD Plot for All ER Admissions",
       x.label = "Age", y.label = "Total ER Admissions")
```




```r
rdplot(y = er$injury, x = er$age, c = 21, title = "RDD Plot for Injury-related ER Admissions",
       x.label = "Age", y.label = "Injury-related ER Admissions")
```




```
rdplot(y = er$alcohol, x = er$age, c = 21, title = "RDD Plot for Alcohol-related ER Admissions",
       x.label = "Age", y.label = "Alcohol-related ER Admissions")
```



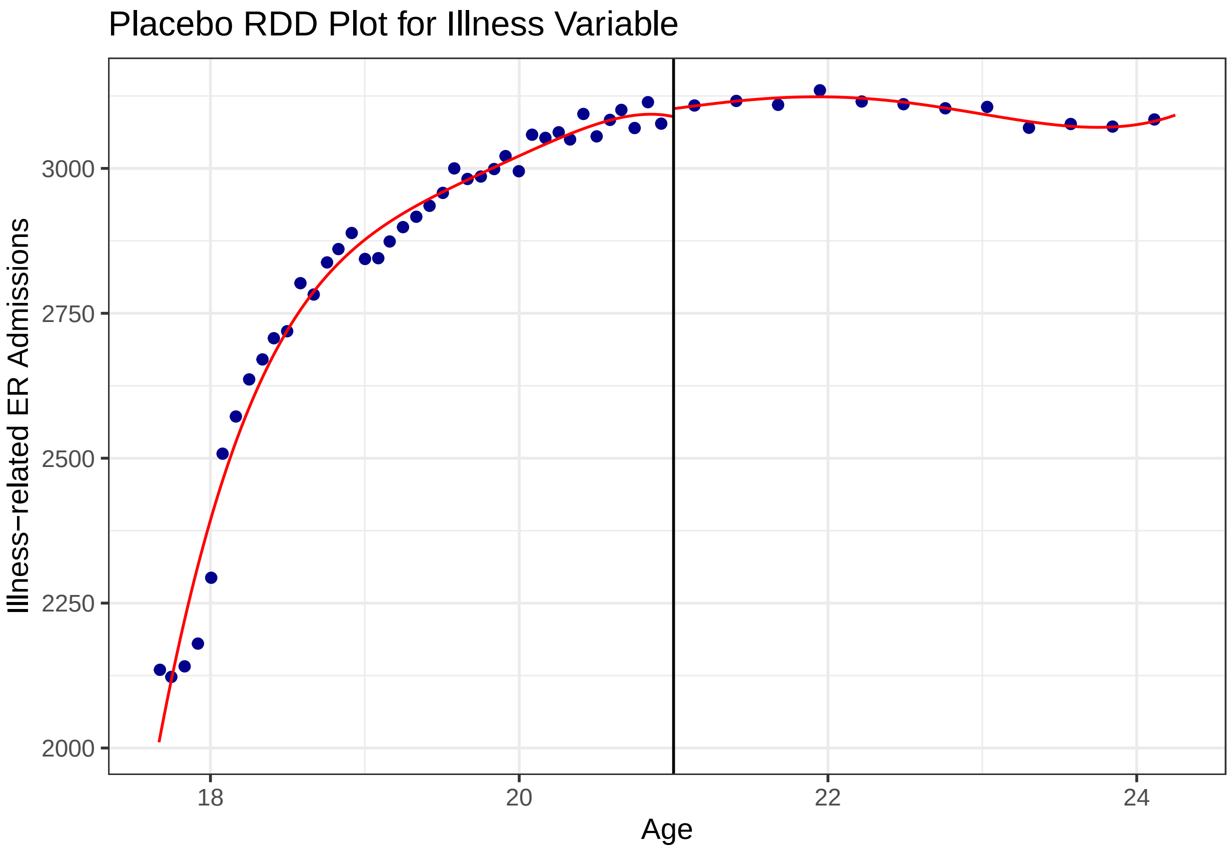
Question C:

```
placebo_rdd <- rdrobust(y = er$illness, x = er$age, c = 21, h = 1)
summary(placebo_rdd)
```

```
## 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     7.487    15.265     0.490     0.624   [-22.431 , 37.404]
##         Robust         -         -     0.684     0.494   [-29.430 , 60.973]
```

```
## =============================================================================
```

```r
rdplot(y = er$illness, x = er$age, c = 21, title = "Placebo RDD Plot for Illness Variable",
       x.label = "Age", y.label = "Illness-related ER Admissions")
```



Answer:
The results show no statistically significant discontinuity at age 21 for the illness variable, so there is no treatment effect. Both the conventional and robust estimates are not significantly different from zero (as indicated by high p-values), and the confidence intervals are wide [-22.431 , 37.404], covering zero. This indicates that reaching the legal drinking age does not cause a significant change in illness-related ER admissions as expected. The placebo test shows that the RDD assumptions hold since there is no significant effect on the illness variable at the age 21 cutoff. This supports the validity of using RDD for the other outcome variables (e.g., all, injury, and alcohol) in the earlier questions. If the RDD showed a significant effect for illness, it would raise concerns about the validity of the causal inference made using the RDD for alcohol-related outcomes.

Problem 3 - Whether family income affects an individual's likelihood to enroll in college

Question A: DAG

```r
library(tidyverse)
library(ggdag)      # For plotting DAGs
```

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

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

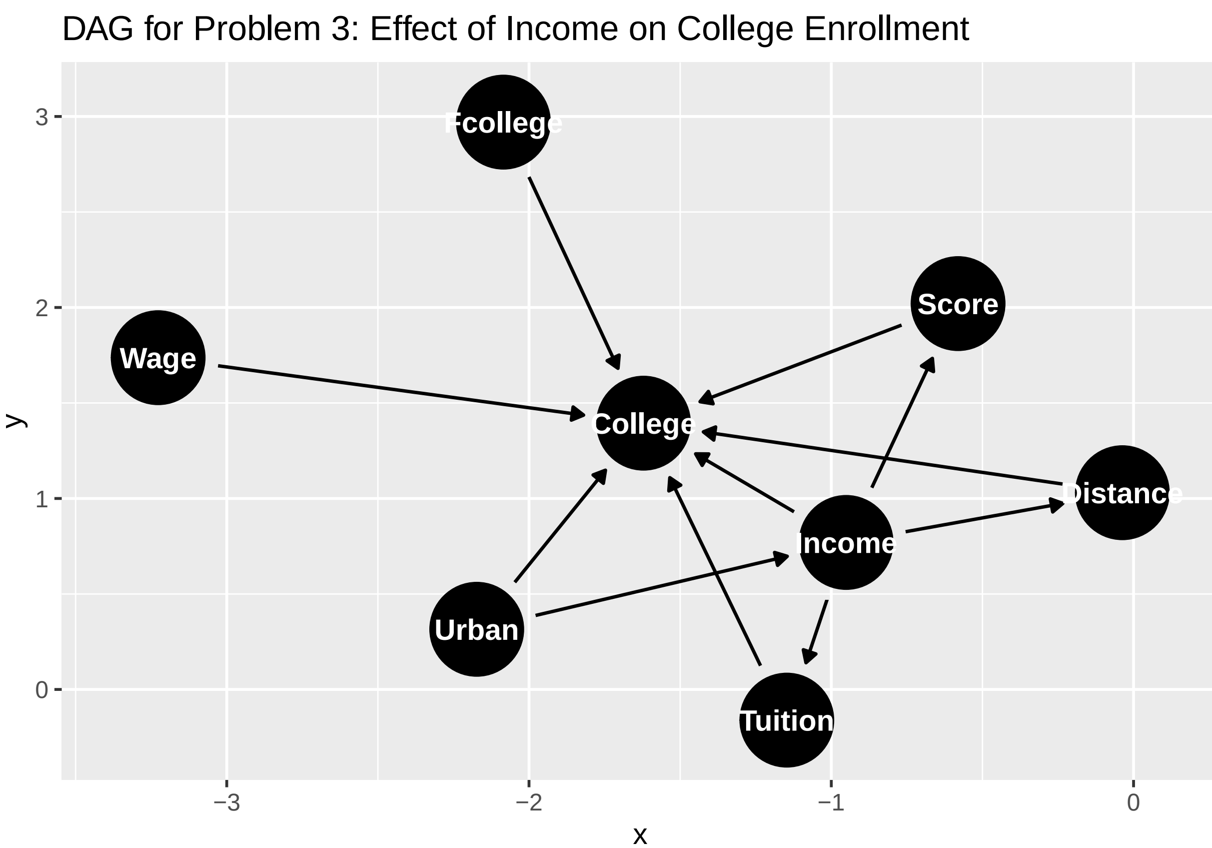
```
##
##     filter
```

```r
library(dagitty)    # For working with DAG logic

# Define the DAG
dag <- dagitty('dag {
  "Income" -> "College"
  "Score" -> "College"
  "Fcollege" -> "College"
  "Distance" -> "College"
  "Tuition" -> "College"
  "Wage" -> "College"
  "Urban" -> "College"

  "Income" -> "Distance"
  "Income" -> "Score"
  "Income" -> "Tuition"
  "Urban" -> "Income"
}')

# Plot the DAG using ggdag
ggdag::ggdag(dag, text = TRUE) +
  ggtitle("DAG for Problem 3: Effect of Income on College Enrollment")
```



Answer:
1. **Income → College**: The primary relationship of interest is between family income and college attendance. Higher family income is likely to increase the likelihood of attending college due to greater financial resources.

Income is treatment and College is outcome. 2. **Score → College**: Academic achievement (measured by scores) directly influences college attendance, as higher scores increase the chances of admission and scholarship opportunities. Students achieved higher scores may enrolled in a better ranked college. 3. **Fcollege → College**: Parental education level, particularly the father's college graduation status, can positively influence a child's educational aspirations and opportunities. Children of college-educated parents are more likely to attend college. 4. **Distance → College**: People may attend College nearer to where they live. Greater distance can act as a barrier, reducing the likelihood of attending college. 5. **Tuition → College**: The cost of tuition is a significant factor in the decision to attend college. Higher tuition costs can deter individuals from enrolling. 6. **Wage → College**: State wages might influence the decision to attend college by affecting the opportunity cost of education. Higher wages could make immediate employment more attractive compared to pursuing higher education. 7. **Urban → College**: Urban areas typically have more access to colleges and educational resources, which can increase the likelihood of attending college.

Additional Relationships

- **Income → Distance**: Higher-income families may have the means to live closer to colleges, thereby reducing the distance and making college attendance more feasible.
- **Income → Score**: Family income can also impact test scores, as wealthier families can afford better educational resources, tutoring, and extracurricular activities that enhance academic performance.
- **Income → Tuition**: While tuition itself is a separate factor, higher-income families are less burdened by tuition costs, making it a less significant deterrent to college attendance.
- **Urban → Income**: Urban areas tend to have higher average incomes due to more job opportunities and higher living costs. Therefore, living in an urban area can be associated with higher family income.

Conditional Independence

To estimate the effect of family income (treatment) on college attendance (outcome), we must control for the confounding variables that could influence both. According to the DAG, I think I should condition on **Score**, **Distance**, **Tuition**, and **Urban**. These variables could potentially confound the relationship between income and college attendance and they open the backdoor paths, so controlling for them is necessary to isolate the effect of income on the outcome.

Question B:

Answer:
In Question B, we want to determine how the effect of family income on college enrollment varies depending on factors like academic score, distance to college, and urban/rural status. The Conditional Average Treatment Effect (CATE) framework allows us to explore these heterogeneous effects by splitting the population into subgroups based on these characteristics. For example:
- High vs. Low Academic Scores: We expect income to have a stronger impact on college enrollment for students with lower scores, as higher-scoring students may be admitted to college regardless of their financial background due to scholarships or merit-based admissions.
CATE analysis helps us understand how and why the effect of income is not uniform across different students. By estimating the treatment effect separately for each group, we can better target policies to the subgroups where income matters most.

Question C:

```r
college <- read_csv('college.csv', show_col_types = FALSE)
```

```
## New names:
## * `` -> `...1`
```

```r
college <- college %>%
  mutate(wage_quantile = ntile(wage, 4))  # Adjust the number of quantiles as needed

cate_hat <- college %>%
  group_by(wage_quantile) %>%
  summarize(
    CATE = mean(college[income == 1]) - mean(college[income == 0]),
    variance = var(college[income == 1]) / sum(income == 1) + var(college[income == 0]) / sum(income ==
    num_treated = sum(income == 1),
    num_untreated = sum(income == 0),
    .groups = 'drop'
  )

# Calculate overall ATE as a weighted average
overall_ate <- cate_hat %>%
  mutate(weight = (num_treated + num_untreated) / sum(num_treated + num_untreated)) %>%
  summarize(
    ATE = sum(CATE * weight),
    variance_ATE = sum(variance * weight),
    .groups = 'drop'
  )

# Calculate the standard error and confidence intervals
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

cate <- overall_ate$ATE
CI_Lower <- ci_lower
CI_Upper <- ci_upper

print(paste("Overall ATE:", cate))
```

```
## [1] "Overall ATE: 0.200466783800077"
```

```r
print(paste("95% CI: [", CI_Lower, ", ", CI_Upper, "]"))
```

```
## [1] "95% CI: [ 0.143476227366586 ,  0.257457340233567 ]"
```

Answer:
The Overall ATE is approximately 0.2005. This means that, on average, having a higher family income increases the likelihood of college enrollment by 20.05 percentage points across all wage quantiles. The positive ATE indicates that higher family income generally has a beneficial effect on the likelihood of enrolling in college. Families with more financial resources are more likely to send their children to college. The 95% Confidence Interval for the ATE is approximately [0.1435, 0.2575]. Since this interval does not include 0, it suggests that the effect of income on college enrollment is statistically significant. The confidence interval being entirely above 0 means that we can be 95% confident that the true effect of income is positive. The interval is fairly narrow, which indicates a precise estimate.

Problem 4 - Whether an extra year of education causes increased wages

Question A:

```r
nazis <- read_csv("nazis.csv", show_col_types = FALSE)
nazis$voteshare <- nazis$nazivote/nazis$nvoter

nazi_model <- lm(voteshare ~ shareblue, data = nazis)
summary(nazi_model)
```

```
##
## Call:
## lm(formula = voteshare ~ 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(nazi_model, level = 0.95)
```

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

Answer:
The coefficient between vote share and the proportion of blue-collar potential voters is 0.06518. This means that for each additional unit increase in the proportion of blue-collar workers (as a fraction from 0 to 1), the Nazi vote share is expected to increase by 6.518%. The standard error is 0.05220, which indicates that the estimate of 0.06518 could fluctuate by around 0.05220 due to random sampling variability. The 95% confidence interval is (-0.03730872, 0.1676687), which includes 0, suggesting it is not statistically significant. The confidence interval means that, with 95% confidence, the true increase in Nazi vote share associated with an increase in the proportion of blue-collar workers lies within this range.

Question B:

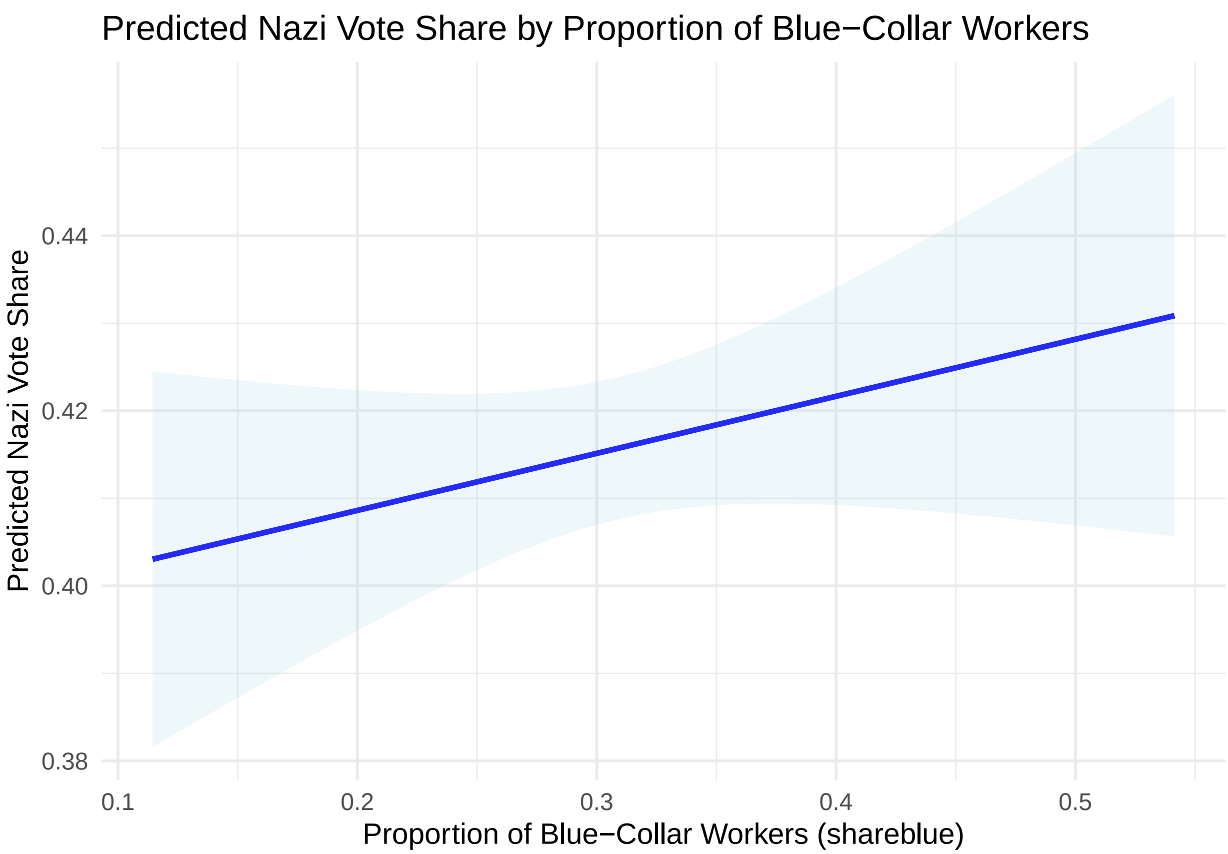
```r
observed_range <- range(nazis$shareblue)

x_values <- seq(from = observed_range[1], to = observed_range[2], length.out = 100)
predicted_values <- predict(nazi_model, newdata = data.frame(shareblue = x_values), interval = "confider
predicted_df <- data.frame(
  shareblue = x_values,
  fit = predicted_values[, "fit"],
```

```r
  lwr = predicted_values[, "lwr"],
  upr = predicted_values[, "upr"]
)

ggplot(predicted_df, aes(x = shareblue, y = fit)) +
  geom_line(color = "blue", size = 1) + # Solid line for the predicted values
  geom_ribbon(aes(ymin = lwr, ymax = upr), alpha = 0.2, fill = "lightblue") + # Shaded area for confide
  labs(
    title = "Predicted Nazi Vote Share by Proportion of Blue-Collar Workers",
    x = "Proportion of Blue-Collar Workers (shareblue)",
    y = "Predicted Nazi Vote Share"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Answer:
The plot shows that as the proportion of blue-collar workers (x-axis) increases, the predicted Nazi vote share (y-axis) also increases. This positive slope indicates a positive relationship between the proportion of blue-collar workers in a precinct and the support for the Nazi party in the 1932 election. The predicted vote share starts around 0.40 when the proportion of blue-collar workers is low (~0.1) and increases steadily to approximately 0.43 as the proportion of blue-collar workers reaches 0.5. This suggests that precincts with a higher concentration of blue-collar workers were more likely to vote for the Nazi party.

Question C:

```r
nazi_model_c <- lm(voteshare ~ shareblue + I(1 - shareblue) - 1, data = nazis)
summary(nazi_model_c)
```

```
##
## Call:
## lm(formula = voteshare ~ shareblue + I(1 - shareblue) - 1, 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 ***
## I(1 - shareblue)  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
```

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

```
##                      2.5 %    97.5 %
## shareblue        0.3894025 0.5321267
## I(1 - shareblue) 0.3629661 0.4282031
```

The Coefficient for shareblue (α) is 0.46076, which represents the estimated Nazi vote share among blue-collar workers. In other words, in a precinct where all voters are blue-collar workers, the Nazi vote share is expected to be approximately 46.08%. The coefficient for I(1 - shareblue) (β) is 0.39558, which means that the estimated Nazi vote share among non-blue-collar workers. In a precinct where all voters are non-blue-collar workers, the Nazi vote share is expected to be approximately 39.56%. Both coefficients are highly statistically significant (p-values $<$ 2e-16), indicating strong evidence that these relationships are not due to random chance. The significance levels suggest that the differences in Nazi vote share between blue-collar and non-blue-collar workers are meaningful. The results indicate that blue-collar workers are more likely to vote for the Nazi party (approximately 46% support) compared to non-blue-collar workers (approximately 40% support). The difference between these two groups is statistically significant, which supports the hypothesis that blue-collar workers were a key base of support for the Nazi party in the 1932 election. The equation 2 estimated a simple linear relationship between the proportion of blue-collar workers and Nazi vote share, while this model explicitly separates the effects for blue-collar and non-blue-collar workers. The coefficients here are more informative as they directly estimate the vote shares within these groups.

Question D:

```r
nazi_model_d <- lm(
  voteshare ~ shareself + shareblue + sharewhite + sharedomestic + shareunemployed - 1,
  data = nazis
)
summary(nazi_model_d)
```

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

```
## lm(formula = voteshare ~ shareself + shareblue + sharewhite +
##     sharedomestic + shareunemployed - 1, 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
```

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

```
##                      2.5 %    97.5 %
## shareself        0.7868165 1.4417096
## shareblue        0.4648233 0.6159423
## sharewhite       0.1378070 0.4323667
## sharedomestic   -0.1268549 0.2312741
## shareunemployed -0.1658852 0.1095646
```

Shareself:
- The coefficient is 1.11426 which is highly significant (p-value = 4.95e-11 < 0.001). It indicates that precincts with a higher proportion of self-employed voters tend to have higher Nazi vote shares. Specifically, for each unit increase in the proportion of self-employed voters, the Nazi vote share is expected to increase by approximately 1.11.
- The Confidence interval is (0.7868165, 1.4417096), which does not include 0, suggesting statistically significant.

Shareblue :
- The coefficient is 0.54038 which is also highly significant (p-value < 0.001). It suggests that precincts with a higher proportion of blue-collar workers tend to have higher Nazi vote shares. The Nazi vote share is expected to increase by approximately 0.54 units for each unit increase in the proportion of blue-collar workers.
- The Confidence interval is (0.4648233, 0.6159423), which does not include 0, suggesting statistically significant.

Sharewhite:
- The coefficient is 0.28509 which is significant (p-value = 0.000157 < 0.001). It suggests that precincts with a higher proportion of white-collar workers also show increased Nazi vote shares, with an expected increase of 0.29 units for each unit increase in the proportion of white-collar workers.
- The Confidence interval is (0.1378070, 0.4323667), which does not include 0, suggesting statistically significant.

Sharedomestic:
- The coefficient is 0.05221 which is not statistically significant (p-value = 0.567). This suggests that the proportion of domestically employed voters does not have a meaningful impact on the Nazi vote share.
- The Confidence interval is (-0.1268549, 0.23127417), which does include 0, suggesting tat it is not statistically significant.

Shareunemployed:
- The coefficient is -0.02816 which is also not statistically significant (p-value = 0.688). The proportion of unemployed voters appears to have no significant relationship with the Nazi vote share.
- The Confidence interval is (-0.1658852, 0.1095646), which does include 0, suggesting tat it is not statistically significant.

The results suggest that self-employed, blue-collar, and white-collar voters were key supporters of the Nazi party in the 1932 election, while domestic workers and unemployed voters did not significantly influence the Nazi vote share. Among the significant predictors, self-employed voters had the largest impact, followed by blue-collar and white-collar voters. The interpretation assumes that the independent variables (proportions of different occupation categories) are not highly collinear. High multicollinearity could lead to inflated standard errors, making it difficult to interpret the coefficients.

Question E:

```
nazis <- nazis %>%
  mutate(
    W_i1_lower = (voteshare - (1 - shareblue)) / shareblue,
    W_i1_upper = voteshare / shareblue
  )
nazis <- nazis %>%
  mutate(
    W_i1_lower = pmax(0, pmin(1, W_i1_lower)),
    W_i1_upper = pmax(0, pmin(1, W_i1_upper))
  )

weighted_avg_lower <- weighted.mean(nazis$W_i1_lower, nazis$shareblue)
weighted_avg_upper <- weighted.mean(nazis$W_i1_upper, nazis$shareblue)
cat("Lower bound for the nationwide proportion of blue-collar voters who voted for the Nazis:", weighte
```

```
## Lower bound for the nationwide proportion of blue-collar voters who voted for the Nazis: 0.000562306
```

```
cat("Upper bound for the nationwide proportion of blue-collar voters who voted for the Nazis:", weighte
```

```
## Upper bound for the nationwide proportion of blue-collar voters who voted for the Nazis: 0.958055
```

Answer:
The lower bound suggests that, at a minimum, only a very small fraction (around 0.056%) of blue-collar voters across all precincts could have voted for the Nazis. This extreme lower bound is derived under the assumption that all non-blue-collar voters in each precinct voted for the Nazis, leaving the smallest possible share for blue-collar voters. The upper bound suggests that, at most, 95.8% of blue-collar voters nationwide could have voted for the Nazis. This upper bound is derived under the assumption that none of the non-blue-collar voters in any precinct voted for the Nazis, leaving the entire vote share in each precinct to be attributed to blue-collar voters.
The large range between the lower and upper bounds indicates considerable uncertainty in precisely estimating the Nazi vote share among blue-collar voters. This is typical in ecological inference problems, where inferring individual-level behavior from aggregate data is challenging without making strong assumptions.

Problem 5 - 2SLS

Question A:

```
wage2 <- read_csv("wage2.csv", show_col_types = FALSE)
```

```r
naive_model <- lm_robust(wage ~ educ, data = wage2)
summary(naive_model)
```

```
##
## Call:
## lm_robust(formula = wage ~ educ, data = wage2)
##
## 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
```

The coefficient for educ is 60.21, meaning that, according to the naive model, each additional year of education is associated with a $60.21 increase in monthly wages, on average. The p-value for the education coefficient is extremely low (1.551e-21), indicating that the relationship between education and wages is statistically significant at the typical significance levels.
The model does not estimate the effect of education on wages correctly. The reason is that the model assumes a simple, direct relationship between education and wages, without accounting for other potential confounders such as parental education, innate ability, or socio-economic background. These confounders might influence both the level of education and wages, leading to an overestimation or underestimation of the true causal effect of education on wages. In other words, the naive model is likely biased because it does not control for unobserved factors that could confound the relationship between education and wages.

Question B:

```r
# Relevance:
model_me_policy <- lm_robust(meduc ~ educ, data = wage2)
summary(model_me_policy)
```

```
##
## Call:
## lm_robust(formula = meduc ~ educ, data = wage2)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   4.2701    0.57711   7.399 3.275e-13    3.137   5.4028 855
## educ          0.4724    0.04194  11.262 1.577e-27    0.390   0.5547 855
##
## Multiple R-squared:  0.1327 ,    Adjusted R-squared:  0.1317
## F-statistic: 126.8 on 1 and 855 DF,  p-value: < 2.2e-16
```

Answer:
- Relevance: To check the relevance of the instrument, I regressed educ on meduc. The coefficient for meduc is significant at the 1% level ($p < 0.01$), with a p-value of 2.2e-16. This indicates that meduc is strongly correlated with educ, satisfying the relevance criterion.
- Exclusion: The exclusion restriction assumption appears reasonable because it is unlikely that a parent's education directly affects their child's wage independently of the child's education.

- Exogeneity: Exogeneity is assumed, as parental education is determined before the child's educational and career decisions, making it less likely to be correlated with unobserved factors affecting wages.

Question C:

```r
first_stage <- lm_robust(educ ~ meduc, data = wage2)
summary(first_stage)
```

```
##
## Call:
## lm_robust(formula = educ ~ meduc, data = wage2)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  10.5749    0.26190   40.38 2.598e-200  10.0609  11.0889 855
## meduc         0.2809    0.02395   11.73  1.455e-29   0.2339   0.3279 855
##
## Multiple R-squared:  0.1327 ,    Adjusted R-squared:  0.1317
## F-statistic: 137.6 on 1 and 855 DF,  p-value: < 2.2e-16
```

```r
wage2$pre_educ <- predict(first_stage, newdata = wage2)

second_stage <- lm_robust(wage ~ pre_educ, data = wage2)
summary(second_stage)
```

```
##
## Call:
## lm_robust(formula = wage ~ pre_educ, data = wage2)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -513.4     216.57  -2.370 1.799e-02  -938.44   -88.29 855
## pre_educ       109.3      16.14   6.772 2.356e-11    77.63   140.99 855
##
## Multiple R-squared:  0.04615 ,   Adjusted R-squared:  0.04504
## F-statistic: 45.86 on 1 and 855 DF,  p-value: 2.356e-11
```

Answer:
In the first stage, I regressed educ on meduc. The p-value for meduc is 2.2e-16 ($p < 0.01$), indicating that the instrument is a strong predictor of educ. This confirms the relevance criterion. In the second stage, I regressed wage on the predicted values of education (pre_educ). The p-value is 2.356e-11 ($p < 0.01$), indicating a statsitically significant. The causal effect of education on wage is 109.3, which means each additional year of education is associated with an increase of $109.3 on wage. This estimate is larger than the naive OLS estimate (60.21), suggesting that the OLS estimate was biased due to endogeneity, likely because omitted variables such as innate ability were correlated with both education and wages.

Question D:

```r
two_sls <- iv_robust(wage ~ educ|meduc, data = wage2)
summary(two_sls)
```

```
##
## Call:
## iv_robust(formula = wage ~ educ | meduc, data = wage2)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)   -513.4      211.1  -2.432 1.524e-02   -927.8   -98.97 855
## educ           109.3       15.8   6.920 8.867e-12     78.3   140.31 855
##
## Multiple R-squared:  0.04123 ,   Adjusted R-squared:  0.04011
## F-statistic: 47.88 on 1 and 855 DF,  p-value: 8.867e-12
```

Answer:
The coefficient for educ using both manual 2SLS and the iv_robust() function is 109.3. This consistency confirms that both methods are correctly estimating the same causal effect of education on wages. The standard errors are slightly different (16.14 vs. 15.8), but the difference is minor. This is likely due to slight variations in how the robust standard errors are calculated in each method. Despite this minor difference, the overall interpretation of the significance remains the same, as both estimates indicate that the effect of education on wages is statistically significant ($p < 0.01$). The iv_robust() approach is more likely to correctly estimate the standard errors because it automatically accounts for robust error adjustments and small-sample corrections that are critical in 2SLS estimation. While the manual 2SLS approach can also provide robust standard errors if carefully implemented, iv_robust() is specifically designed for 2SLS analysis and minimizes the risk of human error.