DS-UA 201: Midterm Exam

Vanessa (Ziwei) Xu

December 10, 2020

Instructions

You should submit your writeup (as a knitted .pdf along with the accompanying .rmd file) to the course website before 11:59pm EST on Saturday December 19th. Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstinitial_final.pdf`. In addition, an electronic copy of your .Rmd file (saved as `Yourlastname_Yourfirstinitial_final.Rmd`) should accompany this submission.

Late finals will not be accepted, **so start early and plan to finish early**. Remember that exams often take longer to finish than you might expect.

This exam has **3** questions and is worth a total of **50 points**. Show your work in order to receive partial credit. Also, I will not accept un-compiled .rmd files.

I general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions we have asked, you will not receive points when you demonstrate knowledge about questions we have not asked, and you will lose points when you make inaccurate statements (whether or not they relate to the question asked). Be careful, however, that you provide an answer to all parts of each question.

You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself. You are prohibited from corresponding with any human being regarding the exam (unless following the procedures below).

I will answer clarifying questions during the exam. I will not answer statistical or computational questions until after the exam is over. If you have a question, send email to me. If your question is a clarifying one, I will remove all identifying information from the email and reply on Piazza. Do not attempt to ask us questions in person (or by phone), and do not post on Piazza.

Problem 1 (25 points)

This problem will have you replicate and analyze the results from Moser and Voena's 2012 AER paper on the impact of the World War I "Trading with the Enemy Act" on U.S. domestic invention. The full citation is below

> Moser, P., & Voena, A. (2012). Compulsory licensing: Evidence from the trading with the enemy act. American Economic Review, 102(1), 396-427.

The premise of the study is to evaluate the effect that "compulsory licensing" policy – that is, policies that permit domestic firms to violate foreign patents and produce foreign inventions without needing to obtain a license from the owner of the foreign patent – have on domestic invention. Does access to foreign inventions make domestic firms more innovative? The authors leverage an exogenous event in U.S. licensing policy that arose from World War I – the 1917 "Trading with the Enemy Act" (TWEA) which permitted U.S. firms to violate patents owned by enemy-country firms. This had the consequence of effectively licensing all patents from German-owned firms to U.S. firms after 1918 (that is, from 1919 onward), allowing them to produce these inventions without paying for a license from the German-owned company.

The authors look specifically at domestic innovation and patent activity in the organic chemicals sector. They note that only some of the sub-classes of organic chemicals (as defined by the US Patent Office) received any compulsory licenses under the Trading with the Enemy Act while others did not. They leverage this variation in exposure to the "treatment" of compulsory licensing to implement a differences-in-differences design looking at domestic firm patent activity in each of these sub-classes (comparing sub-classes that were exposed to compulsory licensing to those that were unexposed).

The dataset is `chem_patents_maindataset.dta` – the code below will load it.

```
library(tidyverse)
# Read in the Moser and Voena (2012) dataset
chem <- haven::read_dta("chem_patents_maindataset.dta")
head(chem)
```

```
## # A tibble: 6 x 23
##   uspto_class grntyr count_usa count_france count_germany count count_for
##   <chr>        <dbl>     <dbl>        <dbl>         <dbl> <dbl>     <dbl>
## 1 008/09410D    1875         0            0             0     0         0
## 2 008/09410D    1876         0            0             0     0         0
## 3 008/09410D    1877         0            0             0     0         0
## 4 008/09410D    1878         0            0             0     0         0
## 5 008/09410D    1879         0            0             1     1         1
## 6 008/09410D    1880         0            0             0     0         0
## # ... with 16 more variables: count_noger <dbl>, count_for_noger <dbl>,
## #   main <chr>, subcl <chr>, year_conf <dbl>, count_cl <dbl>,
## #   licensed_class <dbl>, confiscated_class <dbl>, class_id <dbl>,
## #   year_conf_2 <dbl>, treat <dbl>, itt <dbl>, count_for_2 <dbl>,
## #   year_conf_itt <dbl>, count_cl_itt <dbl>, count_cl_2 <dbl>
```

The unit of the dataset is the sub-class/year (471,120 observations) of 7248 US Patent and Trademark Office (USPTO) patent sub-classes over 65 years.

The relevant variables are

- `uspto_class` - USPTO Patent Sub-Class (unit)
- `grntyr` - Year of observation (year)

- `count_usa` - Count of patents granted to US-owned firms in the year
- `count_france` - Count of patents granted to French-owned firms in the year
- `count_for` - Count of patents granted to foreign-owned (non-US) firms in the year
- `treat` - Treatment indicator – Whether the patent sub-class received any German patents under TWEA (after 1918 when the policy went into effect) (Note that this is not an indicator for the overall treatment *group* (whether the unit *ever* received treatment) – it is only 1 after 1918 for units that receive treatment but is still 0 for those "treated" units prior to the initiation of treatment)

Question A (5 points)

If you try to use a two-way fixed effects estimator on the dataset as it is, it will likely freeze up your computer as this is a *very large* dataset. We'll instead first aggregate the data in a way that will let you use a simple difference-in-differences estimator to estimate the treatment effect.

Generate a point estimate for the average treatment effect of receiving treatment on the average annual count of US patents using a difference-in-differences estimator (using all post-treatment (1919-1939) and pre-treatment (1875-1918) time periods. You should aggregate your data such that the outcome is the post-/pre- difference in the outcome (preferably using `tidyverse` functions like `group_by` and `summarize`) and each row is a USPTO patent sub-class (rather than a sub-class/year observation) and use a difference-in-means estimator with the differenced outcome. Again, if you use `lm_robust` or even `lm` with two-way fixed effects, your computer will likely freeze up as there are many FE parameters to estimate.

Provide a 95% robust confidence interval and interpret your point estimate. Do we reject the null of no treatment effect at the $\alpha = .05$ level?

```
# assign 0 for pre and 1 for post
pre <- (1918 >= chem$grntyr) & (chem$grntyr >= 1875)
post <- (1939 >= chem$grntyr) & (chem$grntyr >= 1919)
chem[pre, "preorpost"] <- 0
chem[post, "preorpost"] <- 1

# aggregate the data into post-/pre-
# 44 pre and 21 post years
chem_pop <- chem %>% group_by(uspto_class, preorpost, treat) %>% summarize(count_usa_mean = mean(count_
                                                                                                    na.rm =
                                                                                  treat = max(treat),
                                                                                  N = n(), .groups = "keep")

chem_pop
```

```
## # A tibble: 14,496 x 5
## # Groups:   uspto_class, preorpost, treat [14,496]
##    uspto_class preorpost treat count_usa_mean     N
##    <chr>           <dbl> <dbl>          <dbl> <int>
##  1 008/09410D          0     0         0.0227    44
##  2 008/09410D          1     0         0.190     21
##  3 008/09410P          0     0         0.114     44
##  4 008/09410P          1     0         0.714     21
##  5 008/09410R          0     0         0.341     44
##  6 008/09410R          1     0         1.38      21
##  7 008/094110          0     0         0.364     44
##  8 008/094110          1     0         1.48      21
```

```
##  9 008/094120          0     0         0.114     44
## 10 008/094120          1     0         0.619     21
## # ... with 14,486 more rows
```

```
# aggregate again into subclasses
chem_pop_diff <- chem_pop %>% group_by(uspto_class) %>% summarize(count_usa_diff = count_usa_mean[preor
                                                                  - count_usa_mean[preorpost == 0],
                                                                  treat = max(treat),
                                                                  N = n(), .groups = "keep")
chem_pop_diff
```

```
## # A tibble: 7,248 x 4
## # Groups:   uspto_class [7,248]
##    uspto_class count_usa_diff treat     N
##    <chr>                <dbl> <dbl> <int>
##  1 008/09410D           0.168     0     2
##  2 008/09410P           0.601     0     2
##  3 008/09410R           1.04      0     2
##  4 008/094110           1.11      0     2
##  5 008/094120           0.505     0     2
##  6 008/094130           0.226     0     2
##  7 008/094140           0.839     0     2
##  8 008/094150           0.254     0     2
##  9 008/094160           0.355     0     2
## 10 008/094170          -0.549     0     2
## # ... with 7,238 more rows
```

```
# build the difference-in-means estimator
diff_in_means <- function(treated, control){
  # Point Estimate
  point <- mean(treated) - mean(control)

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(meanTreated = mean(treated), meanControl = mean(control), est = point, se = se,

  return(as_tibble(output))

}

diff_in_means(chem_pop_diff$count_usa_diff[chem_pop_diff$treat == 1], chem_pop_diff$count_usa_diff[chem
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl   est     se ci95Lower ci95Upper   pvalue     N
```

```
##         <dbl>       <dbl> <dbl>  <dbl>     <dbl>     <dbl>    <dbl> <int>
## 1       0.641       0.386 0.255 0.0377     0.181     0.329 1.25e-11  7248
```

On average, the point estimate for the average treatment effect of receiving treatment is 0.256. This means that the treated group get granted about 25.6% more average annual count of US patents than the control group. The 95% robust confidence interval is [0.181, 0.329]. So we would reject the null of no treatment effect at the $\alpha = .05$ level since the confidence interval doesn't include zero.

Question B (5 points)

A colleague suggests that you should instead just compare 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. Based on what we observe in the pre-1919 period, is ignorability of the treatment likely to hold under this strategy? Discuss why or why not – what do you observe in the patent counts in the pre-treatment period between exposed and unexposed subclasses.

```r
# test for ignoralibity
chem_ig <- chem %>% group_by(uspto_class) %>% summarize(post_mean = mean(count_usa[preorpost == 1],
                                                                          na.rm = TRUE),
                                                        pre_mean = mean(count_usa[preorpost == 0],
                                                                        na.rm = TRUE),
                                                        treat = max(treat), N = n(), .groups = "keep")
chem_ignorability <- chem_ig %>% group_by(treat) %>% summarize(pre_mean = mean(pre_mean, na.rm = TRUE),
                                                               N = n(), .groups = "keep")
chem_ignorability
```

```
## # A tibble: 2 x 3
## # Groups:   treat [2]
##   treat pre_mean     N
##   <dbl>    <dbl> <int>
## 1     0   0.228   6912
## 2     1   0.0827   336
```

Ignorability assumes than the treatment is assigned in a way that is independent of the potential outcomes. What this colleague suggests is that we ignore the pre-1918 periods and only compare those after. From the test I generated, we can see that the patent counts in the pre-treatment period between exposed and unexposed subclasses have a large difference (0.228 for control and 0.08 for treated). So ignoring the pre-1918 periods would violate ignorability as the distribution of potential outcomes under treatment and control differs between units assigned to treatment and those under control. ******

Question C (5 points)

The authors implement a test of their identification assumptions by also estimating the effect (using the differences-in-differences design) of the Trading with the Enemy Act on patents granted by French firms, which the authors note "could not license enemy patents under the TWEA." Describe what sort of a diagnostic strategy this is. What do the authors expect to find if their parallel trends assumption holds?

Estimate the effect of TWEA exposure on the count of French firm patents using a difference-in-differences design and provide a a 95% robust confidence interval. Are the results consistent with what the authors expect if their design assumptions hold?

```
# aggregate the data into post-/pre- treatment periods
chem_pop_fr <- chem %>% group_by(uspto_class, preorpost, treat) %>% summarize(count_france_mean =
                                                                                mean(count_france,
                                                                                     na.rm = TRUE),
                                                                              treat = max(treat),
                                                                              N = n(), .groups = "keep")

chem_pop_fr
```

```
## # A tibble: 14,496 x 5
## # Groups:   uspto_class, preorpost, treat [14,496]
##    uspto_class preorpost treat count_france_mean     N
##    <chr>           <dbl> <dbl>             <dbl> <int>
##  1 008/09410D          0     0            0         44
##  2 008/09410D          1     0            0         21
##  3 008/09410P          0     0            0         44
##  4 008/09410P          1     0            0         21
##  5 008/09410R          0     0            0         44
##  6 008/09410R          1     0            0         21
##  7 008/094110          0     0            0.0227    44
##  8 008/094110          1     0            0.0952    21
##  9 008/094120          0     0            0.0227    44
## 10 008/094120          1     0            0.0476    21
## # ... with 14,486 more rows
```

```
# aggregate again into subclasses
chem_pop_diff_fr <- chem_pop_fr %>% group_by(uspto_class) %>% summarize(count_france_diff =
                                                                          count_france_mean[preorpost == 1]
                                                                        - count_france_mean[preorpost == 0],
                                                                        treat = max(treat),
                                                                        N = n(), .groups = "keep")

chem_pop_diff_fr
```

```
## # A tibble: 7,248 x 4
## # Groups:   uspto_class [7,248]
##    uspto_class count_france_diff treat     N
##    <chr>                   <dbl> <dbl> <int>
##  1 008/09410D            0           0     2
##  2 008/09410P            0           0     2
##  3 008/09410R            0           0     2
##  4 008/094110            0.0725      0     2
##  5 008/094120            0.0249      0     2
##  6 008/094130            0           0     2
##  7 008/094140           -0.0227      0     2
##  8 008/094150            0.00433     0     2
##  9 008/094160           -0.136       0     2
## 10 008/094170           -0.111       0     2
## # ... with 7,238 more rows
```

```r
# calculate the difference in means
diff_in_means(chem_pop_diff_fr$count_france_diff[chem_pop_diff_fr$treat == 1], chem_pop_diff_fr$count_fi
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl      est      se ci95Lower ci95Upper pvalue     N
##         <dbl>       <dbl>    <dbl>   <dbl>     <dbl>     <dbl>  <dbl> <int>
## 1  -0.0000902     0.00194 -0.00203 0.00385  -0.00958   0.00552  0.598  7248
```

The diagnostic strategy used here is a placebo test. If parallel trends assumption holds, it is to be expected for us to find that even without treatment, the units receiving treatment would follow the same trajectory as unites in control. As my test shows, the effect of TWEA exposure on the count of French firm patents is about -0.002 and the 95% robust confidence interval is [-0.010, 0.006]. The null of no treatment effect would not be rejected as the interval covers zero. So, the parallel trends assumption holds and the results are consistent with what the authors expect.

Question D (5 points)

We might be concerned that there are differential trends in pre-treatment patenting between those sub-classes exposed to the treatment and those exposed to control. Estimate the difference in the trend in US patents between exposed and unexposed sub-classes from 1918 to 1917, 1916, 1915, and 1914 (four estimates in total: 1918-1917, 1918-1916, 1918-1915, 1918-1914). Provide a 95% robust confidence interval for each of these estimates and interpret your results. Do we reject the null that any of these differ from 0 (at $\alpha = .05$)? If the outcome trends were evolving in parallel between the, what would we expect these estimates to be? What do your results suggest for the validity of the parallel trends assumption?

```r
# aggregate the results into sub-classes and get years we need
chem_periods <- chem %>% group_by(uspto_class) %>% summarize(count_usa1918 = count_usa[grntyr == 1918],
                                                             count_usa1917 = count_usa[grntyr == 1917],
                                                             count_usa1916 = count_usa[grntyr == 1916],
                                                             count_usa1915 = count_usa[grntyr == 1915],
                                                             count_usa1914 = count_usa[grntyr == 1914],
                                                             count_diff87 = count_usa1918 - count_usa19
                                                             count_diff86 = count_usa1918 - count_usa19
                                                             count_diff85 = count_usa1918 - count_usa19
                                                             count_diff84 = count_usa1918 - count_usa19
                                                             treat = max(treat), N = n(), .groups = "ke
chem_periods
```

```
## # A tibble: 7,248 x 12
## # Groups:   uspto_class [7,248]
##    uspto_class count_usa1918 count_usa1917 count_usa1916 count_usa1915
##    <chr>               <dbl>         <dbl>         <dbl>         <dbl>
##  1 008/09410D              0             0             0             0
##  2 008/09410P              0             0             1             0
##  3 008/09410R              0             0             1             0
##  4 008/094110              1             2             0             0
##  5 008/094120              1             2             0             0
```

```
##  6 008/094130              0             1             1             0
##  7 008/094140              0             1             0             0
##  8 008/094150              0             0             2             0
##  9 008/094160              2             0             1             0
## 10 008/094170              1             2             1             0
## # ... with 7,238 more rows, and 7 more variables: count_usa1914 <dbl>,
## #   count_diff87 <dbl>, count_diff86 <dbl>, count_diff85 <dbl>,
## #   count_diff84 <dbl>, treat <dbl>, N <int>
```

```
# calculate the treatment effect using the difference in means function for 1918-1917
diff_in_means(chem_periods$count_diff87[chem_periods$treat == 1], chem_periods$count_diff87[chem_periods
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl    est     se ci95Lower ci95Upper pvalue     N
##         <dbl>       <dbl>  <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <int>
## 1    -0.00595     -0.0330 0.0270 0.0446   -0.0605     0.115  0.545  7248
```

```
# calculate the treatment effect using the difference in means function for 1918-1916
diff_in_means(chem_periods$count_diff86[chem_periods$treat == 1], chem_periods$count_diff86[chem_periods
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl    est     se ci95Lower ci95Upper  pvalue     N
##         <dbl>       <dbl>  <dbl>  <dbl>     <dbl>     <dbl>   <dbl> <int>
## 1      0.0476     -0.0488 0.0964 0.0368    0.0243     0.168 0.00875  7248
```

```
# calculate the treatment effect using the difference in means function for 1918-1915
diff_in_means(chem_periods$count_diff85[chem_periods$treat == 1], chem_periods$count_diff85[chem_periods
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl    est     se ci95Lower ci95Upper pvalue     N
##         <dbl>       <dbl>  <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <int>
## 1      0.0595    -0.00405 0.0636 0.0344  -0.00379     0.131 0.0644  7248
```

```
# calculate the treatment effect using the difference in means function for 1918-1914
diff_in_means(chem_periods$count_diff84[chem_periods$treat == 1], chem_periods$count_diff84[chem_periods
```

```
## # A tibble: 1 x 8
##   meanTreated meanControl     est     se ci95Lower ci95Upper pvalue     N
##         <dbl>       <dbl>   <dbl>  <dbl>     <dbl>     <dbl>  <dbl> <int>
## 1      0.0268      0.0505 -0.0237 0.0395    -0.101    0.0537  0.548  7248
```

If the outcome trends were evolving in parallel, we would expect these estimates to be basically the same. For the results I generated, the point estimates are 0.027, 0.096, 0.064, and-0.024 for 1918-1917, 1918-1916, 1918-1915, 1918-1914 respectively, and the 95% robust confidence interval for the four are [-0.060, 0.115], [0.024, 0.168], [-0.004, 0.131], and [-0.101, 0.054] respectively. Only the 1918-1916 comparison would reject the null of no treatment effect. The results we got suggest that the validity of the parallel trends assumption holds for mostly periods except a slight change in the comparison of 1918 to 1916.

Question E (5 points)

The authors adjust for covariates in addition to their out of concern for possible parallel trends violations. One possible confounder that might be driving a parallel trends violation is the overall amount of foreign patenting in the sub-class and its change over time – reflecting general technological differences that might differ between the patent sub-classes. Since the treatment does not affect the amount of foreign patenting, this is a valid control.

Create a variable for the change between the post- and pre-treatment count of foreign patents in the USPTO subclass. Bin this variable into six (6) roughly-equally sized strata and estimate the effect of the treatment on US patenting (again using the differenced outcome) using a stratified difference-in-means estimator. Provide a robust 95% confidence interval and interpret your results. Do we reject the null of no treatment effect at the $\alpha = .05$ level? Compare your results to your estimate from Question A and discuss why they might differ.

```
# Create a variable for the change between the post- and pre-treatment count of foreign patents
chem_change <- chem %>% group_by(uspto_class) %>% summarize(change = sum(count_for[preorpost == 1])
                                                            - sum(count_for[preorpost == 0]),
                                                            usa = mean(count_usa[preorpost == 1], na.rm
                                                            - mean(count_usa[preorpost == 0], na.rm = TR
                                                            treat = max(treat), N = n(), .groups = "keep
chem_change
```

```
## # A tibble: 7,248 x 5
## # Groups:   uspto_class [7,248]
##    uspto_class change    usa treat     N
##    <chr>        <dbl>  <dbl> <dbl> <int>
##  1 008/09410D       2  0.168     0    65
##  2 008/09410P       5  0.601     0    65
##  3 008/09410R       8  1.04      0    65
##  4 008/094110      -5  1.11      0    65
##  5 008/094120       0  0.505     0    65
##  6 008/094130       0  0.226     0    65
##  7 008/094140       6  0.839     0    65
##  8 008/094150      -2  0.254     0    65
##  9 008/094160       0  0.355     0    65
## 10 008/094170     -11 -0.549     0    65
## # ... with 7,238 more rows
```

```
# create six equally sized stratum
chem_stratum <- quantile(chem_change$change, seq(0, 1, by = 1/6))
chem_stratum[1] <- 0
chem_stratum[7] <- 1
chem_change$stratum <- cut(chem_change$change, unique(chem_stratum), labels = F)
table(chem_change$stratum)
```

```
##
##    1    2    3    4
## 2128 1222  699  746
```

```r
# Bin this variable into six (6) roughly-equally sized strata

# next, create the strat_diff_in_means function
strat_diff_in_means <- function(outcome, treatment, stratum){

  # For each stratum
  strat_ests <- bind_rows(map(unique(stratum), function(x) diff_in_means(outcome[treatment == 1&stratum

  # Normalize weights to sum to 1
  strat_ests$weight <- strat_ests$N/sum(strat_ests$N)

  # Point estimate
  point = sum(strat_ests$est*strat_ests$weight)

  # Standard error
  se = sqrt(sum(strat_ests$se^2*strat_ests$weight^2))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se, point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(est = point, se = se, ci95Lower = ci_95[1], ci95Upper = ci_95[2],
                       pvalue = pval, N= length(outcome))

  return(as_tibble(output))
}

# estimate the effect of the treatment on US patenting using a stratified difference-in-means estimator
strat <- strat_diff_in_means(chem_change$change, chem_change$treat, chem_stratum)
```

```
## Warning in treatment == 1 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 0 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 1 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 0 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 1 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 0 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 1 & stratum == x: longer object length is not a multiple
## of shorter object length
```

```
## Warning in treatment == 0 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 1 & stratum == x: longer object length is not a multiple
## of shorter object length

## Warning in treatment == 0 & stratum == x: longer object length is not a multiple
## of shorter object length
```

```
strat
```

```
## # A tibble: 1 x 6
##     est    se ci95Lower ci95Upper      pvalue     N
##   <dbl> <dbl>     <dbl>     <dbl>       <dbl> <int>
## 1  3.90 0.778      2.38      5.43 0.000000522  7248
```

The estimate of the effect of the treatment on US patenting is about 3.904 and the robust 95% confidence interval is [2.379, 5.429]. We would reject the null of no treatment effect at the $\alpha = .05$ level. Our results differ from results in question A as this shows that foreign patenting does have an impact due to general technological differences. It actually is a confounder that might be driving a parallel trends violation.

Problem 2 (5 points)

This problem will ask you to demonstrate that the propensity score is a "balancing score" – that is that, conditional on the propensity score, the potential outcomes are independent of the treatment (and we don't need to condition on anything else besides the propensity score). Assume our usual set-up for a design with selection-on-observables. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control respectively. Y_i is our observed outcome and D_i is our observed treatment. We assume *conditional ignorability* – that conditional on pre-treatment covariates X_i, treatment D_i is independent of the potential outcomes.

$$Y_i(1), Y_i(0) \perp D_i | X_i$$

We also assume positivity

$$0 < Pr(D_i = 1 | X_i) < 1$$

and consistency (as usual).

$$Y_i(d) = Y_i \text{ if } D_i = d$$

Define the propensity score $e(X_i)$ as the probability of treatment given covariates X_i

$$e(X_i) = Pr(D_i = 1 | X_i)$$

Show that it is also true that

$$Y_i(1), Y_i(0) \perp D_i | e(X_i)$$

In other words, that ignorability holds conditional on the propensity score alone.

- Hint 1: It suffices to show that the probability of treatment given the propensity score does not change when we further condition on the potential outcomes $Y_i(1)$ and $Y_i(0)$.
- Hint 2: Condition on X_i and use the law of total expectations.
- Hint 3: Remember the "fundamental bridge" – for any binary (0/1) random variable A, $E[A] = Pr(A = 1)$

P(Di = 1 | YjiP(Xi)) = E[Di | Yji , P(Xi)] = E[E[Di | Yi, P(Xi), Xi] | Yji, P(Xi)] = E[E[Di | Xi] | Yji, P(Xi)] = E[P(Xi) | Yji, P(Xi)] = P(Xi) which is the propensity score

---

Problem 3 (20 points)

This problem examines a study by Acemoglu, Johnson and Robinson examining the effect of political institutions on economic development.

> Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5), 1369-1401.

The authors are interested in whether robust political institutions with protections on private property encourage economic growth and raise GDP per capita. However, institutions are not randomly assigned.

The authors leverage historical variation in the types of political institutions established by Europeans during the colonial period in different parts of the world. The authors posit that in regions where early settler mortality rates were low, settlers were more likely to establish robust political institutions with limitations on government power. Conversely, in areas where early settler mortality was high, settlers instead established "extractive" institutions with weak checks on government power, designed primarily to transfer resource wealth to the colonizers. The authors argue that even after decolonization and independence, the structure of these institutions persisted in the countries, affecting subsequent economic growth and development.

The relevant dataset is `ajr-aer.dta` dataset. The code below loads in the dataset and subsets it down to the relevant observations.

```
library(tidyverse)
library(haven)

# Load in exercise dataset
ajr <- haven::read_dta("ajr-aer.dta")
# Subset down to the original dataset
ajr <- ajr %>% filter(baseco == 1)
head(ajr)
```

```
## # A tibble: 6 x 10
##   shortnam africa lat_abst rich4 avexpr logpgp95 logem4  asia loghjypl baseco
##   <chr>     <dbl>    <dbl> <dbl>  <dbl>    <dbl>  <dbl> <dbl>    <dbl>  <dbl>
## 1 AGO           1    0.137     0   5.36     7.77   5.63     0   -3.41        1
## 2 ARG           0    0.378     0   6.39     9.13   4.23     0   -0.872       1
## 3 AUS           0    0.300     1   9.32     9.90   2.15     0   -0.171       1
## 4 BFA           1    0.144     0   4.45     6.85   5.63     0   -3.54        1
## 5 BGD           0    0.267     0   5.14     6.88   4.27     1   -2.06        1
## 6 BHS           0    0.268     0   7.5      9.29   4.44     0   NA           1
```

```r
nrow(ajr)
```

```
## [1] 64
```

The variables of interest are:

- `logpgp95` - Logged GDP per capita in 1995 (outcome)
- `avexpr` - average state protection against property expropriation risk (treatment)
- `logem4` - logged historical settler mortality rates (instrument)
- `lat_abst` - Absolute value of the latitude of capital divided by 90

Question A (5 points)

Note that the instrument here is continuous as is the treatment (quality of political institutions as measured by average expropriation risk). The authors will assume linear models for the relationship between instrument and treatment and treatment and outcome as will we in this problem.

Fit a (robust) linear regression model for the first stage (using `lm_robust`), predicting the average expropriation risk conditional on logged historical settler mortality rates. Provide a point estimate and 95% confidence interval for the marginal effect of a one unit increase in logged historical settler mortality rates on average expropriation risk. Interpret the estimate and discuss whether we would reject the null of no effect at the $\alpha = .05$ level.

```r
lm_robust(avexpr ~ logem4, data = ajr)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept)  9.3414102  0.7162447 13.042205 1.982911e-19  7.9096574 10.7731629
## logem4      -0.6067782  0.1529963 -3.965968 1.920519e-04 -0.9126134 -0.3009431
##             DF
## (Intercept) 62
## logem4      62
```

The point estimate for the marginal effect of a one unit increase in logged historical settler mortality rates on average expropriation risk is -0.607, and the 95% confidence interval is [-0.913, -0.301]. As the interval doesn't cover zero, we would reject the null of no treatment effect at the $\alpha = .05$ level.

Question B (5 points)

Using the two-stage least squares estimator (assuming linearity), estimate the effect of a one-unit increase in average expropriation risk on logged GDP per capita in 1995 (assuming a linear relationship), instrumenting for average expropriation risk using logged historical settler mortality rates. Provide a point estimate and 95% confidence interval. Interpret your results and discuss whether we would reject the null of no effect at the $\alpha = .05$ level.

```
two_sls <- iv_robust(logpgp95 ~ avexpr | logem4, data = ajr)
two_sls
```

```
##              Estimate Std. Error  t value     Pr(>|t|)   CI Lower CI Upper DF
## (Intercept) 1.9096665  1.2174380 1.568595 1.218325e-01 -0.5239573 4.343290 62
## avexpr      0.9442794  0.1825866 5.171679 2.638652e-06  0.5792939 1.309265 62
```

The point estimate for the marginal effect is 0.944, and the 95% confidence interval is [0.579, 1.309]. As the interval doesn't cover zero, we would reject the null of no treatment effect at the $\alpha = .05$ level.

Question C (5 points)

Discuss whether the instrumental variables assumptions hold in this case. Evaluate exogeneity of the instrument in particular by examining whether the instrument and outcome are possibly confounded by geography (here, as measured by the absolute value of the latitude (deviation from the equator)).

```
# test for exclusion restriction
lm_robust(logpgp95 ~ logem4, data = ajr)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept) 10.7305676  0.3851541 27.860452 9.129094e-37  9.9606555 11.5004797
## logem4      -0.5729682  0.0736457 -7.780063 9.530695e-11 -0.7201839 -0.4257525
##             DF
## (Intercept) 62
## logem4      62
```

```
# test for exogeneity of the instrument
lm_robust(logpgp95 ~ lat_abst, data = ajr)
```

```
##             Estimate Std. Error   t value     Pr(>|t|) CI Lower CI Upper DF
## (Intercept) 7.419604  0.2126836 34.885643 1.842927e-42 6.994455 7.844752 62
## lat_abst    3.548445  0.8333203  4.258201 7.112301e-05 1.882662 5.214229 62
```

```
lm_robust(avexpr ~ lat_abst, data = ajr)
```

```
##             Estimate Std. Error   t value     Pr(>|t|) CI Lower CI Upper DF
## (Intercept) 5.752502  0.3140986 18.314323 1.064556e-26 5.124628 6.380376 62
## lat_abst    4.213754  1.4049995  2.999115 3.894232e-03 1.405201 7.022308 62
```

The instrumental variables assumptions include randomization of the instrument, exclusion restriction, first-stage relationship, and monotonicity. Randomization of the instrument means that the instrument is independent of both sets of potential outcomes. This does not hold as we have proven. Exclusion restriction means that the instrument only affects the outcome by way of its effect on treatment. We have proven that it is wrong as well. First-stage relationship means that the instrument has an effect on treatment and we have proven that it holds in question A. Lastly, monotonicity means that the relationship between the instrument and treatment only goes in one direction at the individual level. This is essentially not a testable assumption. For the test for geography's impact, we can see that it does have an effect on both the treatment and outcome so exogeneity does not hold.

Question D (5 points)

Again, assuming linearity, and using the two-stage least squares estimator, estimate the effect of a one-unit increase in average expropriation risk on logged GDP per capita in 1995, instrumenting for average expropriation risk using logged historical settler mortality rates but now assuming that the instrument is valid only conditional on the country's distance from the equator (absolute value of latitude divided by 90). Provide a point estimate and 95% confidence interval. Interpret your results and discuss whether we would reject the null of no effect at the $\alpha = .05$ level. How do your results differ from your estimates in B?

```
two_sls2 <- iv_robust(logpgp95 ~ avexpr + lat_abst | logem4 + lat_abst, data = ajr)
two_sls2
```

```
##               Estimate Std. Error    t value    Pr(>|t|)   CI Lower CI Upper DF
## (Intercept)  1.6918138  1.5186873  1.1139975 0.269650468 -1.3449890 4.728617 61
## avexpr       0.9957040  0.2522809  3.9468075 0.000207498  0.4912373 1.500171 61
## lat_abst    -0.6472071  1.2983377 -0.4984891 0.619932106 -3.2433938 1.948980 61
```

The point estimate is 0.996, and the 95% confidence interval is [0.491, 1.50]. As the interval doesn't cover zero, we would reject the null of no treatment effect at the $\alpha = .05$ level. This shows a greater effect than we could see in question B because of the covariate.
