DS-UA 201: Midterm Exam

Professor Anton Strezhnev

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")
```

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 on the count of US patents in the 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?

```
#label the years to be post-treatment period

chem$potreatment <- as.integer(chem$grntyr >= 1919 & chem$grntyr <= 1939)
chem$pretreatment <- as.integer(chem$grntyr >= 1875 & chem$grntyr <= 1918)

#find the patents count for the pre-treatment year and post-treatment year
chem$pre_patents_us <- chem$pretreatment * chem$count_usa
chem$post_patents_us <- chem$potreatment * chem$count_usa


#group by subclass
chem_subclass <- chem %>% group_by(uspto_class) %>% summarize(pre_avg_patents = sum(pre_patents_us)/sum

#check the data
chem_subclass
```

```
## # A tibble: 7,248 x 5
## # Groups:   uspto_class [7,248]
##    uspto_class pre_avg_patents post_avg_patents difference treatment
##    <chr>                 <dbl>            <dbl>      <dbl>     <int>
##  1 008/09410D           0.0227            0.190      0.168         0
##  2 008/09410P           0.114             0.714      0.601         0
##  3 008/09410R           0.341             1.38       1.04          0
##  4 008/094110           0.364             1.48       1.11          0
##  5 008/094120           0.114             0.619      0.505         0
##  6 008/094130           0.25              0.476      0.226         0
##  7 008/094140           0.114             0.952      0.839         0
##  8 008/094150           0.841             1.10       0.254         0
##  9 008/094160           1.45              1.81       0.355         0
## 10 008/094170           0.977             0.429     -0.549         0
## # ... with 7,238 more rows
```

```
#difference-in-means estimator with the differenced outcome
did <- lm_robust(difference ~ treatment, data = chem_subclass)
tidy(did)
```

```
##          term  estimate  std.error statistic       p.value  conf.low conf.high
## 1 (Intercept) 0.3861610 0.01012305  38.14670 2.547678e-290 0.3663169 0.4060052
## 2   treatment 0.2553208 0.03769195   6.77388  1.352010e-11 0.1814335 0.3292080
##     df    outcome
## 1 7246 difference
## 2 7246 difference
```

The point estimate for the average treatment effect of receiving treatment on the average annual count of US patents is 0.255, with 95 % confidence interval of [0.181 , 0.329]. At alpha level = 0.05, we could reject the null of null effect. Being in the treated group thus tend to increase the number of patents produced by the domestic inventors.

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.

```
#graph for the trend pre/post treatment

# find the treated cohort
treated_cohort <- chem_subclass$uspto_class[chem_subclass$ treatment == 1]
control_cohort <-chem_subclass$uspto_class[chem_subclass$ treatment == 0]

chem$treated_ever <- as.integer(chem$uspto_class %in% treated_cohort)



#size of treated vs control cohort
treated_num <- length(treated_cohort)
control_num <- length(control_cohort)

treated_num
```

```
## [1] 336
```

```
control_num
```

```
## [1] 6912
```

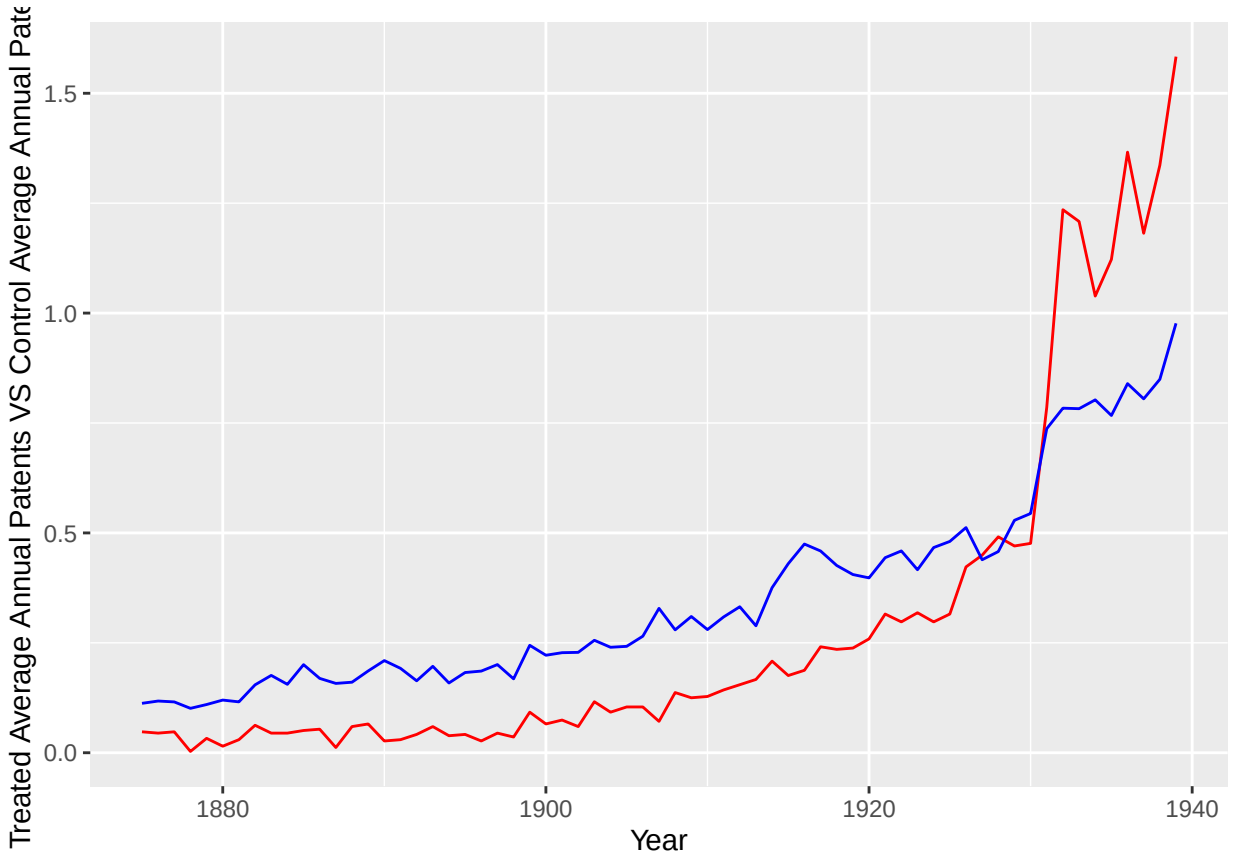
```
#find the treated cohort patents count per year and control cohort patents count per year
chem$treated_year <- chem$treated_ever * chem$count_usa
chem$control_year <- as.integer(chem$treated_ever == 0) * chem$count_usa


#group by year
chem_year <- chem %>% group_by(grntyr) %>% summarize(treated_avg_patents_year = sum(treated_year)/336,

chem_year
```

```
## # A tibble: 65 x 3
## # Groups:   grntyr [65]
##    grntyr treated_avg_patents_year control_avg_patents_year
##     <dbl>                    <dbl>                    <dbl>
##  1   1875                0.0476                    0.112
##  2   1876                0.0446                    0.118
##  3   1877                0.0476                    0.116
##  4   1878                0.00298                   0.101
##  5   1879                0.0327                    0.110
##  6   1880                0.0149                    0.120
##  7   1881                0.0298                    0.116
##  8   1882                0.0625                    0.155
##  9   1883                0.0446                    0.176
## 10   1884                0.0446                    0.156
## # ... with 55 more rows
```

```
ggplot(chem_year, aes(x=grntyr)) +
  geom_line(aes(y = treated_avg_patents_year), color = "red") +
  geom_line(aes(y = control_avg_patents_year), color="blue") + xlab("Year") + ylab("Treated Average Annu
```



The ignorability of the treatment is not likely to hold under this strategy. Since from the graph, we could see that that U.S. inventors were most likely to license in subclasses where levels of domestic invention were initially low. If we only 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, we would underestimate the effect of the treatment.

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?

```
#find the patents counts for the pre-treatment year and post-treatment year
chem$pre_fr <- chem$pretreatment * chem$count_france
chem$post_fr <- chem$potreatment * chem$count_france

#group by subclass
chem_fr <- chem %>% group_by(uspto_class) %>% summarize(pre_avg_fr = sum(pre_fr)/sum(pretreatment), pos
```

```r
did2 <- lm_robust(difference_fr ~ treatment, data = chem_fr)
tidy(did2)
```

```
##          term     estimate    std.error statistic      p.value     conf.low
## 1 (Intercept) -0.001939503 0.0004792926 -4.046595 5.250839e-05 -0.002879056
## 2   treatment  0.002029691 0.0038517924  0.526947 5.982465e-01 -0.005520945
##       conf.high   df       outcome
## 1 -0.0009999497 7246 difference_fr
## 2  0.0095803262 7246 difference_fr
```

This is an example of a placebo test (estimating an effect for a sub-population where we strongly expect the effect to be 0). This test allows French investor, who can't be in the treated group to be "treated". Results show that there is no measureable changes in annual patents by French investor for treated subclasses. Which shows that the effects of the TWEA were limited to the U.S. firms.

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
#1918-1917
chem_1917 <- chem %>% filter(grntyr == 1917 |grntyr ==1918)
chem_1917$year_1917 <- as.integer(chem_1917$grntyr == 1917) * chem_1917$count_usa
chem_1917$year_1918 <- as.integer(chem_1917$grntyr == 1918) * chem_1917$count_usa

chem_1917_groups <- chem_1917 %>% group_by(uspto_class) %>% summarise(y_1917 = sum(year_1917), y_1918 =

did1917 <- lm_robust(diff ~ treated, data = chem_1917_groups)
tidy(did1917)
```

```
##          term    estimate  std.error  statistic     p.value    conf.low
## 1 (Intercept) -0.03298611 0.01207111 -2.7326491 0.006297958 -0.05664901
## 2     treated  0.02703373 0.04464873  0.6054759 0.544881815 -0.06049080
##      conf.high   df outcome
## 1 -0.009323215 7246    diff
## 2  0.114558258 7246    diff
```

The difference in the trend in US patents is 0.027 prior to treatment, with 95 % confidence interval of [-0.0605 , 0.11456]. At alpha level = 0.05, we could not reject the null of null effect.

```r
#1918-1916

chem_1916 <- chem %>% filter(grntyr == 1916 |grntyr ==1918)
chem_1916$year_1916 <- as.integer(chem_1916$grntyr == 1916) * chem_1916$count_usa
chem_1916$year_1918 <- as.integer(chem_1916$grntyr == 1918) * chem_1916$count_usa

chem_1916_groups <- chem_1916 %>% group_by(uspto_class) %>% summarise(y_1916 = sum(year_1916), y_1918 =
```

```
did1916 <- lm_robust(diff ~ treated, data = chem_1916_groups)
tidy(did1916)
```

```
##          term    estimate  std.error statistic      p.value    conf.low
## 1 (Intercept) -0.04875579 0.01361173 -3.581895 0.0003433475 -0.07543874
## 2     treated  0.09637483 0.03675892  2.621808 0.0087647516  0.02431664
##     conf.high   df outcome
## 1 -0.02207283 7246    diff
## 2  0.16843303 7246    diff
```

The difference in the trend in US patents is 0.0964 prior to treatment, with 95 % confidence interval of [0.0243 , 0.1684]. At alpha level = 0.05, we could reject the null of null effect.

```
#1918-1915
chem_1915 <- chem %>% filter(grntyr == 1915 |grntyr ==1918)
chem_1915$year_1915 <- as.integer(chem_1915$grntyr == 1915) * chem_1915$count_usa
chem_1915$year_1918 <- as.integer(chem_1915$grntyr == 1918) * chem_1915$count_usa

chem_1915_groups <- chem_1915 %>% group_by(uspto_class) %>% summarise(y_1915 = sum(year_1915), y_1918 =

did1915 <- lm_robust(diff ~ treated, data = chem_1915_groups)
tidy(did1915)
```

```
##          term     estimate  std.error  statistic   p.value    conf.low
## 1 (Intercept) -0.004050926 0.01338321 -0.3026872 0.7621369 -0.03028591
## 2     treated  0.063574735 0.03437044  1.8496923 0.0643986 -0.00380134
##    conf.high   df outcome
## 1 0.02218406 7246    diff
## 2 0.13095081 7246    diff
```

The difference in the trend in US patents prior to treatment is 0.06357, with 95 % confidence interval of [-0.0038 , 0.1310]. At alpha level = 0.05, we could not reject the null of null effect.

```
#1918-1914
chem_1914 <- chem %>% filter(grntyr == 1914 |grntyr ==1918)
chem_1914$year_1914 <- as.integer(chem_1914$grntyr == 1914) * chem_1914$count_usa
chem_1914$year_1918 <- as.integer(chem_1914$grntyr == 1918) * chem_1914$count_usa

chem_1914_groups <- chem_1914 %>% group_by(uspto_class) %>% summarise(y_1914 = sum(year_1914), y_1918 =

did1914 <- lm_robust(diff ~ treated, data = chem_1914_groups)
tidy(did1914)
```

```
##          term    estimate  std.error  statistic      p.value    conf.low
## 1 (Intercept)  0.05049190 0.01354606  3.7274230 0.0001949211  0.02393767
## 2     treated -0.02370618 0.03947654 -0.6005132 0.5481830640 -0.10109171
##    conf.high   df outcome
## 1 0.07704613 7246    diff
## 2 0.05367934 7246    diff
```

The difference in the trend in US patents prior to treatment is -0.0237, with 95 % confidence interval of [-0.1011 , 0.0537]. At alpha level = 0.05, we could not reject the null of null effect.

From the above comparison, we could see that parallel trends assumption is fairly reasonable in this case. The trend in the potential outcomes under control during the pre-treatment period in the treated group is the quite similar with the observed trend in the control group. Other than the period 1918-1916, all time interval shows minor difference in trend between treated and control group. This validates the parallel trends

assumptions.

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.

```
#find the patents counts for the pre-treatment year and post-treatment year
chem$pre_foreign <- chem$pretreatment * chem$count_for
chem$po_foreign <- chem$potreatment * chem$count_for

#group by subclasses, and find the diff in average annual count for both foreign countries and U.S.
chem_foreign <- chem %>% group_by(uspto_class) %>% summarise(foreign_pre = sum(pre_foreign)/sum(pretrea

#find the cut points
chem_cutpoints <- quantile(chem_foreign$diff_foreign, seq(0, 1, by=1/6))

#divide the subclasses into 6 bins
chem_foreign$stra <- cut(chem_foreign$diff_foreign, chem_cutpoints, labels= F)

# validate the split is roughly even
table(chem_foreign$stra)
```

```
##
##    1    2    3    4    5    6
## 1207 1285 1317 1106 1138 1194
```

```
#stratified diff in means
strat_reg <- lm_lin(diff_us ~ treat,
                    covariates = ~ as.factor(stra),
                    data = chem_foreign)
tidy(strat_reg)
```

```
##                          term    estimate  std.error  statistic       p.value
## 1                 (Intercept)  0.40147667 0.01003889 39.9921231 3.730689e-316
## 2                       treat  0.09405192 0.03846104  2.4453819  1.449356e-02
## 3        (as.factor(stra)2)_c -0.04907939 0.02294595 -2.1389131  3.247609e-02
## 4        (as.factor(stra)3)_c -0.00392231 0.02397899 -0.1635728  8.700720e-01
## 5        (as.factor(stra)4)_c  0.18691485 0.02918569  6.4043321  1.604818e-10
## 6        (as.factor(stra)5)_c  0.34645457 0.03295671 10.5124125  1.158719e-25
## 7        (as.factor(stra)6)_c  0.83364640 0.05079491 16.4120065  1.845465e-59
## 8  treat:(as.factor(stra)2)_c -0.13124832 0.10924992 -1.2013585  2.296515e-01
## 9  treat:(as.factor(stra)3)_c  0.27496267 0.16834665  1.6333124  1.024468e-01
## 10 treat:(as.factor(stra)4)_c -0.15691449 0.10288777 -1.5251034  1.272770e-01
## 11 treat:(as.factor(stra)5)_c -0.26169089 0.10903793 -2.3999987  1.642025e-02
## 12 treat:(as.factor(stra)6)_c -0.41599669 0.10653723 -3.9047072  9.519734e-05
##       conf.low    conf.high   df outcome
```

```
## 1   0.38179751  0.421155832 7235 diff_us
## 2   0.01865706  0.169446786 7235 diff_us
## 3  -0.09406016 -0.004098633 7235 diff_us
## 4  -0.05092813  0.043083512 7235 diff_us
## 5   0.12970238  0.244127323 7235 diff_us
## 6   0.28184979  0.411059350 7235 diff_us
## 7   0.73407354  0.933219250 7235 diff_us
## 8  -0.34541005  0.082913410 7235 diff_us
## 9  -0.05504590  0.604971251 7235 diff_us
## 10 -0.35860456  0.044775577 7235 diff_us
## 11 -0.47543707 -0.047944714 7235 diff_us
## 12 -0.62484077 -0.207152616 7235 diff_us
```

The point estimate for the average treatment effect of receiving treatment on the average annual count of US patents is now 0.09405, with 95 % confidence interval of [0.0187 , 0.1694]. At alpha level = 0.05, we could reject the null of null effect. Being in the treated group thus tend to increase the number of patents produced by the domestic inventors. This is smaller than the result I got in part A. In this part, we tried to control for the unobservable heterogeneity across subclasses which may increase the patening by all foreign investors, regardless of the TWEA. After including the change between the post- and pre-treatment count of foreign patents, the average treatment effect decrease a bit, meaning that there might exist some underlying unobservable across subclasses that encourages the patening besides TWEA. However, after the adjustment, the ATE is still significant at alpha level = 0.05.

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, $E[A] = Pr(A = 1)$

I am intend to show that probability of treatment given the propensity score does not change when we further condition on the potential outcomes $Y_i(1)$ and $Y_i(0)$. $$

Pr(D_i = 1 | Y_i(1), Y_i(0), e(X_i)) = Pr(D_i = 1 | e(X_i)) = e(X_i)

$$

$$ Pr(D_i = 1 | Y_i(1), Y_i(0), e(X_i)) = E[Pr(Di = 1 | Y_i(1), Y_i(0), e(X_i), Xi) | Y_i(1), Y_i(0), e(X_i)]

= E[Pr(Di = 1 | Y_i(1), Y_i(0), Xi) | Y_i(1), Y_i(0), e(X_i)]

= E[Pr(Di = 1 | Xi) | Y_i(1), Y_i(0), e(X_i)]

= E[e(X_i)| Y_i(1), Y_i(0), e(X_i)]

= e(X_i)

$$

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)
```

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.

```
reg_A <- lm_robust(avexpr ~ logem4, data=ajr)
tidy(reg_A)
```

```
##          term   estimate std.error statistic      p.value   conf.low  conf.high
## 1 (Intercept)  9.3414102 0.7162447 13.042205 1.982911e-19  7.9096574 10.7731629
## 2      logem4 -0.6067782 0.1529963 -3.965968 1.920519e-04 -0.9126134 -0.3009431
##   df outcome
## 1 62  avexpr
## 2 62  avexpr
```

The point estimate of a one-unit increase in logged historical settler mortality rates on average expropriation risk is -0.607 , with 95 % confidence interval of [-0.913 , -0.301]. At alpha level = 0.05, we could reject the null of null effect. We can conclude that increase in historical settler mortality rates tend to have a negative effect on average expropriation risk. ### 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.

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

```
##          term  estimate std.error statistic      p.value   conf.low conf.high
## 1 (Intercept) 1.9096665 1.2174380  1.568595 1.218325e-01 -0.5239573  4.343290
## 2      avexpr 0.9442794 0.1825866  5.171679 2.638652e-06  0.5792939  1.309265
##   df  outcome
## 1 62 logpgp95
## 2 62 logpgp95
```

The point estimate of a one-unit increase in average expropriation risk on logged GDP per capita in 1995 is 0.944, with 95 % confidence interval of [0.579 , 1.31]. At alpha level = 0.05, we could reject the null of null effect. We can conclude that increase in average expropriation risk tend to have positive effect on logged GDP per capita in 1995.

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)).

```
ajr_geo <-ajr %>% group_by(logem4) %>% summarize(geo = mean(lat_abst) ,.groups = "keep")

ajr_geo
```

```
## # A tibble: 36 x 2
## # Groups:   logem4 [36]
##    logem4    geo
##     <dbl>  <dbl>
##  1   2.15 0.378
##  2   2.70 0.246
##  3   2.71 0.422
##  4   2.74 0.322
##  5   2.78 0.667
##  6   2.79 0.394
##  7   2.87 0.0196
##  8   3.26 0.0889
##  9   3.47 0.0556
## 10   3.61 0.333
## # ... with 26 more rows
```

From the result, I could see that the logged historical settler mortality rates as an instrument is not really balanced on the Absolute value of the latitude of capital divided by 90. The latitude differ quite significantly across the instruments. As we know, that the climate of a country tend to have some economic impact on that country, it is a reasonable to say that the instrument and outcome are possibly confounded by geography.

Assumption 1: Randomization Zi is independent of both sets of potential outcomes (potential outcomes for the treatment and potential outcomes for the outcome). From this observational study, the instrument of this observational study is the European settlers' mortality rate. The potential mortality rate faced by the soldiers, bishops, and sailors in the colonies is reasonably independent to the average state protection against property expropriation risk. Also, the potential mortality rate is also fairly independent to the personal income per capita in 1995. Thus, after control for covariates like geography, the potential settler mortality rate would be viable instrument.

Assumption 2: Exclusion Restriction Z only affects Y by way of its effect on D From this observational study, we can see the mortality rates of settlers from centuries ago would tend to have no effects on today's economy other than the effect through its effect through the institution. However, since the settlers' mortality rate was data from centuries ago, I am not quite sure whether the mortality rate affected some other variables (for example, culture) that might also have impact on the logged GDP per capita in 1995.

Assumption 3: First-stage relationship Z has an effect on D From the first stage regression, we could see that the t-statistics is 3.97, and the F-statistics will be 15.97, which is larger than the threshold F-statistics = 10. Thus, I would conclude that early settlers'mortality rate does have impact on the average state protection against property expropriation risk.

Assumption 4: Monotonicity Z's effect on D only goes in one direction at the individual level This is not testable assumption. However, I think monotonicity holds in this observational study because colonies with lower mortality rate tend to establish colonies that has more concrete legal system to protect the property right, whereas colonies with higher mortality rate tend to establish colonies that had more extractive states. Therefore, I would say that Z's effect on D tend to go in one direction at the individual level.

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?

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

```
##          term   estimate std.error  statistic     p.value   conf.low conf.high
## 1 (Intercept)  1.6918138 1.5186873  1.1139975 0.269650468 -1.3449890  4.728617
## 2      avexpr  0.9957040 0.2522809  3.9468075 0.000207498  0.4912373  1.500171
## 3    lat_abst -0.6472071 1.2983377 -0.4984891 0.619932106 -3.2433938  1.948980
##   df  outcome
## 1 61 logpgp95
## 2 61 logpgp95
## 3 61 logpgp95
```

The point estimate of a one-unit increase in average expropriation risk on logged GDP per capita in 1995 now is 0.996, with 95 % confidence interval of [0.491 , 1.50]. At alpha level = 0.05, we could reject the null of null effect. The point estimate now is higher than the previous point estimate(0.944). However, Including latitude does not change the point estimate significantly.