Microeconometrics II

Homework II

Professor: Bruno Ferman Student: Rafael F. Bressan

2020-12-11

Part I

1. (10 points) Describe the conditions in which the differences-in-differences (DID) and the synthetic control (SC) estimators are valid. What are the main conditions in which these estimators are consistent/unbiased/asymptotically unbiased?

DID:

Suppose we are in the simplest DID setup possible, units are separated in two states, $S_i \in \{0,1\}$ and observed in two periods of time, $T_t \in \{0,1\}$ and we also observe an outcome variable, y_{it}. We have the following potential outcomes model, with homogeneous treatment effect, β:

$$\begin{cases} y_{it}(0) = \alpha_0 + S_i + T_t + \eta_{it} \\ y_{it}(1) = \beta + y_{it}(0) \end{cases}$$

This implies the following DID regression equation:

$$y_{it} = \alpha_0 + S_i + T_t + \beta D_{it} + \eta_{it} \tag{1}$$

where $D_{it} \in \{0,1\}$ indicates the treatment status. If we let the treatment occur only at time $t = 1$ for state $s = 1$, then D_{it} is the interaction of S_i and T_t.

Equation regression above is a two-way fixed effect model and represents basic difference-in-difference model. The causal parameter of interest is β and it corresponds to the double difference of $E[y_{i1}(1) - y_{i0}(1)|D = 1] - E[y_{i1}(0) - y_{i0}(0)|D = 0]$, thus, the estimator's name.

The causal parameter β is identified under the assumption that **potential outcome trends would be the same** in both control and treatment groups, in the absence of treatment. It's the treatment that induces a deviation from this common (or parallel) trend.

This means that treatment assignment must be independent of potential outcomes, specifically, $E[D_{it}\eta_{it'}] = 0$ for all t and t'. **Treatment should not be chosen in response** to knowledge about how the outcome is likely to be and also that **outcome should not change in anticipation of treatment**.

SC:

The synthetic control method is mostly used when the units of observation are a small number of aggregate entities and there is one single treated unit, while a combination of untreated units provides a good approximation for the treat unit conterfactual. The formal setup of a SC method may be described as follows.

We have at disposal observations for $J + 1$ units and $j = 0$ is our treated unit at times $t > t_0$ and $\mathcal{T} = \{1, \ldots, t_0, \ldots, T\}$. We call the "donor pool" the other untreated units, $j = 1, \ldots, J$. We observe an outcome of interest for each unit at every period of time, y_{jt}. Suppose the outcome is explainable by a set of

k predictor variables, $X_{1j}, \ldots, X_{kj}$. We can group the predictors variables for untreated units together and have a $k \times J$ matrix $\mathbf{X_0}$. Let the potential outcome for any unit j without treatment at period t be denoted by y_{jt}^N. The treated unit, after period t_0 will have suffered intervention, therefore its potential outcome under intervention is the observed outcome, $y_{1t}^I = Y_{1t}$ for $t > t_0$. We define the **effect of the intervention** on the outcome of interest as:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \tag{2}$$

for $t > t_0$. The challenge is to estimate the counterfactual $Y_{1t}^N, t > t_0$.

The SC approach can also be posed as a linear factor model of potential outcomes:

$$y_{jt} = \begin{cases} y_{jt}^N = c_j + \delta_t + \lambda_t \mu_j + \varepsilon_{jt}, \text{ if } d_{jt} = 0 \\ y_{jt}^I = \alpha_{jt} + y_{jt}^N, \text{ if } d_{jt} = 1 \end{cases} \tag{3}$$

where λ_t is the vector of common factors, μ_j the factor loadings, α_{jt} is the treatment effect and ε_{jt} is the idiosyncratic shock. We must notice that common factors and respective loadings are **not observable** and are part of the error structure. The intervention is signaled by $d_{jt} = 1$ and happens only when $j = 0$ and $t > t_0$.

The main assumptions are: i) the donor pool members are comparable to the treated unit and were not subject to treatment, and ii) predictors X_{kj} are unaffected by the intervention and iii) the intervention itself is not related to idiosyncratic shocks, ε_{jt}, for every unity and all times.

If these assumptions are met, a synthetic control is defined as a weighted average of the units in the donor pool. Let $\mathbf{W} = (w_1, \ldots, w_J)'$ be a $J \times 1$ vector of weights with all entries in $[0, 1]$ and sum to one. Given this set of weights we can estimate the counterfactual as:

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt} \tag{4}$$

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N \tag{5}$$

From equation (4) it is clear how the above assumptions play a role in giving an unbiased estimator for the effect of intervention. The expected value of $\hat{y}_{0t}^N$ must converge to the counterfactual y_{0t}^N, that is, the weighted average of synthetic control's outcome must be a good approximation of the treated outcome *had it not been treated*.

2. (10 points) Describe settings in which the conditions above would be invalid.

The treatment exogeneity may easily be violated in a DID setup, take for example Chay, McEwan, and Urquiola (2005) who consider a policy in Chile where poorly performing schools were given additional financial resources. The DID estimator compares average school outcomes between treated and control schools before and after the subsidy. It's easy to imagine that due to chance, an above threshold school had a poor performance exactly in the pre-treatment year, inducing this school to the treatment group. After receiving treatment, and suppose the true effect is negligible, the school has a normal performance above threshold as usual. In this case the DID estimator would overestimate the treatment effect in a phenomenon known as Ashenfelter's dip.

In the synthetic control framework one can think of a donor pool that is not close enough to the treated unit, that is, the treated unit is unique, and even when it had not suffered intervention it had unique characteristics that cannot be obtained by a weight average of all other units. Abadie (2020) presents some conditions on the context of the investigation under which synthetic controls are appropriate, and when it is not. The synthetic control method is not well suited when the outcome is highly volatile with respect to small shock,

when the comparison group (i.e. donor pool) is not "free" from a similar shock we are interested and when there are large idiosyncratic shocks that affect their outcome.

Also, the synthetic control method, just like DID, does not hold if there is anticipation of treatment affecting the outcome prior to the intervention. Finally, the no interference, much like the SUTVA, must hold.

3. (10 points) Describe the different alternative inference methods that are available for these two estimators. Be specific about the conditions in which each inference method would be reliable.

Inference in DID estimation is typically carried out through asymptotic approximation. If available, the researcher can also make cluster robust and/or heteroskedasticity robust inference. Although, this approach is only available if the number of both treated and control units (or clusters) is large. When we have only a handful of treated units, Ferman and Pinto (2019) show the CRVE estimator will underestimate the variance. Wild bootstrap in this framework would not be a good choice either, (Canay, Santos, and Shaikh 2018).

When the number of treated units is small, but we do have a large number of controls in a DID framework, the method of Conley and Taber (2011) is the literature's preferred choice. But, it does come with strings attached. One of the method's assumptions is that errors are iid across unities and independent of treatment assignment, thus eliminating the possibility of heteroskedasticity across treatment and control groups. This assumption may be unrealistic in situation where the observations are aggregated from individual level and there is correlation among same group individuals. Ferman and Pinto (2019) show this leads to a size problem in inference. They suggest an approach that estimates the conditional heteroskedastic variance and then rescales the residuals to apply the bootstrap procedure.

The synthetic control method is more problematic when it comes to inference. Results assessment is usually carried out by visual inspection and placebo tests. Abadie, Diamond, and Hainmueller (2010) propose a mode of inference for the synthetic control framework that is based on permutation methods. Their method relies on a test statistic that measures the ratio of the post-intervention fit relative to the pre-intervention fit.

$$R_j\left(t_1, t_2\right) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} \left(Y_{jt} - \widehat{Y}_{jt}^N\right)^2\right)^{1/2} \tag{6}$$

and the goodness-of-fit ratio is:

$$r_j = \frac{R_j\left(T_0 + 1, T\right)}{R_j\left(1, T_0\right)} \tag{7}$$

r_j measures the quality of the fit of a synthetic control for unit j in the **post-treatment period**, relative to the quality of the fit in the pre-treatment period. Abadie, Diamond, and Hainmueller (2010) use the permutation distribution of r_j for inference.

Chernozhukov, Wuthrich, and Zhu (2020) develop permutation inference procedures that are valid under weak and easy-to-verify conditions, and are provably robust against misspecification. The methods work in with many different approaches for predicting counterfactual mean outcomes in the absence of the intervention, like synthetic controls, difference-in-differences, etc. The authors draw their method from end-of-sample instability tests and devise a sampling-based inferential procedure for synthetic controls and related methods that employs permutations of regression residuals in the time dimension. The main assumption is errors stationarity and weak dependence on the counterfactual generating process. By imposing the null hypothesis of no intervention effect, a test based on permutations of the residuals, $\hat{u}_t = Y_{0t} - Y_t W^*$ can be carried on.

Part II

Choose a published paper the uses a DID or SC estimator in which the dataset is available online. You can also consider a paper that uses shift-share designs, but in this case there is a risk that we will not cover that in class. Once you decide on your paper, sign up at this link. It must be a different paper for each student.

1. (10 points) Briefly summarize the research question in the paper, and why it is interesting. What is the main parameter of interest?

The chosen paper is Kovak (2013) which studies the impact of brazillian trade liberalization in the 1990s mainly under Collor government, to regional wages. The empirical strategy is designed as a shift-share experiment, since import tariffs are set at national level but its impact on local labor markets is dependent on the industries' share of labor demand.

How trade liberalization impacts economic growth has always been a relevant and debatable question, more recently the focus has passed to liberalization and inequality. Specially for developing countries, the literature finds that trade liberalization may result in increase of wage inequality, (Goldberg and Pavcnik 2007).

Kovak (2013) follows a growing literature by examining the effects of trade liberalization on labor market outcomes at the subnational level. The author develops a theoretical specific-factors model of regional economies to back the literature's practice of measuring the local effect of liberalization using a weighted average of changes in trade policy, with weights based on the industrial distribution of labor in each region.

The main parameter of interest is the regional effect of trade liberalization on real wages. This is captured by the sensitivity of regional change of log-wages from 1990 to 1995 (the period considered so the liberalization run its course) to a measure of regional change in trade policy. The author calls this measure as the region-level tariff change – RTC – and it is described in the next question.

The empirical results confirm the model's predictions. Local labor markets whose workers were concentrated in industries facing the largest tariff cuts were generally affected more negatively.

2. (10 points) Briefly describe the empirical setting.

The shift-share "treatment" considered in the paper is the region-level tariff change (RTC) for each brazilian microregion computed as follows:

$$RTC_r = \sum_{i \neq N} \beta_{ri} d\ln(1+\tau_i) \tag{8}$$

$$\text{where} \quad \beta_{ri} = \frac{\lambda_{ri} 1/\theta_i}{\sum_{i' \neq N} \lambda_{ri'} 1/\theta_{i'}} \tag{9}$$

where r denotes a brazilian microregion and i an industry. Liberalization induced price changes from 1990 to 1995 is measured as $d\ln(1+\tau_i)$, where τ_i is the tariff rate. Parameter λ_{ri} is the fraction of regional labor allocated to industry i and θ_i is the cost share of industry-specific factor.

Therefore, the main specification for the empirical analysis of log-variation in wages to a liberalization shock is:

$$d\ln(w_r) = \zeta_0 + \zeta_1 RTC_r + \varepsilon_r \tag{10}$$

and the main parameter of interest is the regional effect of liberalization on real wages, ζ_1.

Wage and employment data come from the brazilian demographic censuses (Censo Demográfico) for 1991 and 2000. Local labor markets were defined by microregions. Wages are calculated as earnings divided by hours for individuals aged 18-55 who are not enrolled in school. Industry classification comes from the Applied Economics Research Institute (IPEA), while Kume, Piani, and Souza (2003) report nominal tariff change by industry during the liberalization.

The parameters for the fraction of regional labor allocated to an industry and industry-specific factor cost share are computed from 1991 census, taking into account individual level observations and adjusting wages by a Mincer-type estimation.

3. (10 points) Discuss whether the conditions for validity of the empirical method are reasonable.

The main identification assumption in regression (10) is the strict exogeneity of the trade liberalization shock with respect to the industry performance. The analysis utilizes variation in tariff changes across industries to compose the regressor RTC_r. If the tariff reduction is correlated to industry performance, then one cannot argue the regressor in equation (10) is independent from unobservable factors that also affect the variation in log-wages, hence, unconfoundedness does not hold and a causal interpretation of ζ_1 is undue.

The author is aware of this assumption and dedicates a whole section to analyze why exogeneity is likely to hold. He qualitatively argues that the driving force of Brazil's liberalization came from government, not the private sector which had little to no influence on the process. A more striking support for exogeneity comes from the data on tariff cuts by industries. Industries most protect have experienced the largest cuts, showing that the primary goal of brazillian government was to reduce tariffs in general and making the cross-industry variation as least as possible considering external incentives.

Although the author makes a compelling case for exogeneity, it is not unrealistic to imagine some sort of industry lobbying during (the short, granted) reform discussion in Congress, for example. If that is the case, some industries may end up being relatively protected after the liberalization and this can be seen in Figure 1 from Kovak (2013) and shown bellow in figure 1, where Agriculture; Footware, leather and Electric, electronic equipment; did not suffer a large reduction in tariff as expected by their preliberalization tariff level.

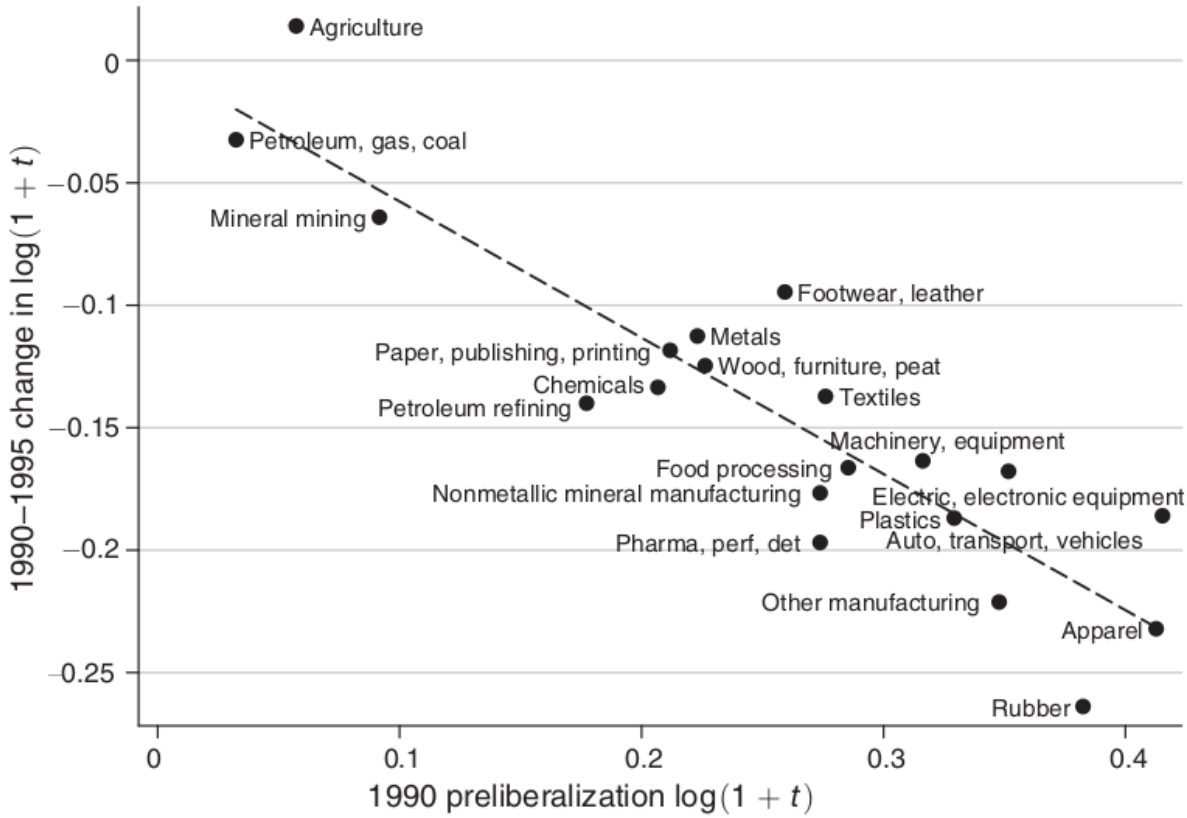

Figure 1: Relationship between tariff changes and preliberalization tariff levels. Kovak (2013) Figure 1.

Concern with the correct estimation of ζ_1 is raised when the baseline data for employment and wages is contaminated with the "treatment". Ideally, wage and employment information to compose the baseline (the RTC_r indeed) should come from just before liberalization began in March 1990. Although, the use 1991 census as the baseline period relies on the assumption that wages and employment shares adjusted slowly to the trade liberalization. The trade liberalization process itself is not a sharp, one-off event, and took

place during almost half a decade. This opens the possibility that workers adapted during the period, for example getting qualification and changing employment to better paid jobs just before the 1991 census, thus changing the composition of labor share in a microregion **after** the shock has begun. This fact should create an attenuation bias.

Another source of attenuation bias is interregional mobility in response to liberalization, which will smooth out the regional wage variation that would have been observed on impact. Suppose the case of instant worker mobility, all liberalization-induced wage variation would be immediately arbitraged away by worker migration. The author raises this potential problem, since the theoretical model predicts an increase in labor supply for regions facing increased wages, but do not demonstrate data that backs his assumption of negligible migration.

The baseline contamination and interregional labor migration may provoke **attenuation bias**, that is, ζ_1 will tend towards zero. This kind of bias is not so restrictive in the present paper since if causal effect is statistically significant (and the author does get effect!) then the true effect would be even larger, hence, there would be no doubt over the findings (supposing the inference is correctly done, of course).

4. (10 points) Describe how the authors conducted inference. Discuss whether the conditions for validity of the inference method are reasonable.

Inference is carried out using cluster-robust variance estimation (CRVE) obtained from asymptotic theory. The unity of observation is microregion and those are clustered into states. The regression analysis counts with 493 observations clustered at 27 states. The number of clusters is borderline to what we expect to have reasonable asymptotic approximation, since CRVE valid when we have independence across clusters and the number of them grows in tandem with number of observations, thus $N_c \to \infty$.

In the next question I will perform the assessment proposed in Ferman (2019) to have an indication whether the approximation is reliable or not. After the assessment, other inference methods will be carried and compared to the paper's original one, like wild bootstrap, permutation test and standard error computation as in Adao, Kolesár, and Morales (2019).

5. (30 points) Use the data from the paper to re-analyze the results of the paper. You can, for example, check whether the results are robust to alternative specifications or alternative approaches for inference, you can provide some evidence that the assumptions the authors rely on are reasonable or unreasonable, you can check whether the inference method the authors are using is reliable, and so on. Be creative!

The author himself made several specifications to attest robustness and his theoretical findings. First let's replicate the author's main results table:

Now that we have the same results as the original paper, we can make an assessment, as in Ferman (2019), to verify the conducted inference's reliability. Table 2 provides the assessment for two different significance levels ($\alpha = 5\%$ and 10%).

The assessment method is showing that specifications without a state fixed effect are less likely to suffer from overejection for both levels of significance. When state fixed effect is included, the overejection is exacerbated and there are serious concerns about CRVE reliability. As Ferman (2019) alerts, *"If this assessment uncovers a rejection rate significantly larger than the level of the test [...], then this would be a strong indication that the researcher should proceed with caution"*. Therefore, all fixed effect specifications still have rejection rates much higher than the test size and inference based on CRVE for models including fixed effects should be made with caution.

Having this concern in mind, we perform three alternative methods for inference for this particular study, the wild bootstrap, randomization inference and the methods proposed in Adao, Kolesár, and Morales (2019).

Table 3 provides **p-values** for all eight specifications under unstudentized and studentized wild bootstrap, randomization inference and, AKM and AKM0 methods. Homoskedastic, Eicker-Huber-White and CRVE (Region Cluster) are also provided for comparison purposes. Let's take as example the main specification with fixed effects, model number 2. In the original inference, CRVE, this parameter is deemed significant at 1% level, thus a p-value lower than 0.01, and this is indeed the case we found in Table 3. The wild bootstrap

Table 1: The effect of liberalization on local wages

Dependent Variable:	log-variation in wage							
Model:	Main (1)	Main (2)	No labor (3)	No labor (4)	Nontraded (5)	Nontraded (6)	Workers (7)	Workers (8)
Variables								
Regional tariff change	0.404	0.439***	0.409	0.439***	2.71	1.96**	0.417	0.482***
	(0.502)	(0.146)	(0.475)	(0.136)	(1.67)	(0.777)	(0.497)	(0.140)
Nontraded sector Omited	✓	✓	✓	✓			✓	✓
Zero price change					✓	✓		
Labor share adjustment	✓	✓			✓	✓	✓	✓
Fixed-effects								
State		✓		✓		✓		✓
Fit statistics								
Observations	493	493	493	493	493	493	493	493
R^2	-0.01852	0.51172	-0.0142	0.5148	-0.07203	0.50422	0.01548	0.52623
Within R^2		0.03929		0.04535		0.02454		0.10362

State standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2: Simple assessment on original inference method

	Name	$\alpha = 0.05$	$\alpha = 0.10$
1	Main	0.128	0.207
2	Main FE	0.806	0.833
3	No labor	0.135	0.214
4	No labor FE	0.837	0.864
5	Nontraded	0.093	0.129
6	Nontraded FE	0.634	0.670
7	Workers	0.126	0.248
8	Workers FE	0.809	0.826

Table 3: Alternative methods for inference. p-values.

		Wild Bootstrap		RI	Adao et. al.				
	Name	Unstud. p-value	Stud. p-value	p-value	Homo	EHW	Reg. cluster	AKM	AKM0
1	Main	0.580	0.587	0.599	0	0.3020	0.4211	0.0012	0.1971
2	Main FE	0.002	0.003	0.021	0	0.0000	0.0026	0.0000	0.0657
3	No labor	0.553	0.557	0.586	0	0.2858	0.3893	0.0002	0.1521
4	No labor FE	0.001	0.001	0.017	0	0.0000	0.0012	0.0000	0.0575
5	Nontraded	0.340	0.762	0.304	0	0.0067	0.1039	0.0000	0.0088
6	Nontraded FE	0.136	0.154	0.054	0	0.0000	0.0114	0.0000	0.0046
7	Workers	0.526	0.534	0.536	0	0.3505	0.4013	0.0026	0.2456
8	Workers FE	0.000	0.000	0.009	0	0.0000	0.0006	0.0000	0.1016

Table 4: Robustness check introducing average unemployment as control

	Main		No labor		Nontraded		Workers	
Estimate	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Regional tariff change	0.2904	0.2461	0.3132	0.2873	2.5850	1.4673	0.3968	0.2275
Standard Errors								
Homoscedastic	0.1123	0.0977	0.1062	0.0926	0.3445	0.3124	0.1141	0.0849
EHW	0.3354	0.1163	0.3393	0.1097	0.9012	0.4227	0.3892	0.0986
Reg. cluster	0.5407	0.1779	0.5240	0.1803	1.6236	0.7987	0.5245	0.1536
AKM	0.2183	0.1025	0.2071	0.0940	0.7342	0.3386	0.2693	0.0786
AKM0	Inf	0.1512	Inf	0.1358	1.0305	0.3730	Inf	0.1300

method agrees with this value, while randomization inference (Fisher's exact p-value) shows a p-value of 0.021, and when AKM0 is used for inference this p-value jumps to 0.0657.

In general, for this particular paper, even though CRVE seems to be providing underestimated standard errors, especially when state fixed effects are included, the values are not very far from other inference methods and overall I would rather not say the inference is completely off-basis.

A robustness check is carried by introducing the average rate of unemployment in the baseline year of 1991 as a control variable. This data is not present in the original dataset, thus, I collected it from DATASUS and merged it to the data. Results are presented in Table 4.

Table 4 shows the point estimate of a regional tariff change is reduced in magnitude but do not change sign. Most of the models do not show statistical significance for this parameter or it is at least reduced when compared to Table 1. Specifically for model (2), the paper's main model, CRVE standard error is 0.1779 on an estimate of 0.2461, while AKM standard error is 0.1025.

Annex A - R Code for homework_II.R

```
#' ## homework_II.R
#' Homework II - Microeconometrics II
#' Author: Rafael Felipe Bressan
#'
#' Paper: Regional Effects of Trade Reform: What Is the Correct Measure of
#' Liberalization? by Brian Kovak - AER2013
#' Loading libraries
library(tidyverse)
library(fixest)
library(sandwich)
library(dtplyr)
library(data.table)
library(ShiftShareSE)

#' Process 1991 census and create matrices of share weights to use with
#' Adao et. al. inference
# source("II/homework_II_census.R")

#' Part II
#' Load data
folder <- "II/input/Kovak2013/AER-2011-0545_data/"

dlnwmmc_mincer <- haven::read_dta("II/input/dlnwmmc_mincer.dta")
dlnwmmc_mincer_nt <- haven::read_dta("II/input/dlnwmmc_mincer_nt.dta")
rtc <- haven::read_dta("II/input/rtc.dta")
microreg_to_mmc <- read_dta(paste0(folder, "microreg_to_mmc.dta")) %>%
  as.data.table()
#' Load DATASUS data on unemployment in 1991
unemp <- fread("II/input/desemprego1991.csv", sep = ";", encoding = "Latin-1",
               dec = ",")
unemp[, `:=`(microreg = as.numeric(str_extract(unemp$`Microrregião IBGE`, "\\d{5}")))]
unemp <- microreg_to_mmc[unemp, on = "microreg"]
unemp <- unemp[, .(avg_unemp91 = mean(`Taxa_de_desemprego_16a_e+`, na.rm = TRUE)),
               by = "mmc"]
#' Processed data for SS assessment and Adao et. al inference
load("II/input/homework_II_Adao.RData")

W_main <- weight_main[, -1]
W_notheta <- weight_notheta[, -1]
W_nt <- weight_nt[, -1]

#' Join data, create state fixed effect and drop Manaus
df <- dlnwmmc_mincer %>%
  left_join(dlnwmmc_mincer_nt, by = "mmc") %>%
  left_join(rtc, by = "mmc") %>%
  left_join(as_tibble(unemp), by = "mmc") %>%
  mutate(state = factor(floor(mmc/1000))) %>%
  filter(mmc != 13007)
# Embed the weights on the data frame
df <- df %>%
  mutate(weights = dlnwmmc_mincerse^-2,
         weights_nt = dlnwmmc_mincerse_nt^-2)
```

```r
#' Setup fixest
setFixest_se(no_FE = "hetero")
setFixest_dof(dof(fixef.K = "full")) # get Stata's reg results

# Replication of Table 1 ------------------------------------------------


#' Main specifications - columns (1) and (2)
reg_main <- feols(dlnwmmc_mincer~rtc_main,
                  weights = ~weights,
                  data = df)
reg_main_fe <- feols(dlnwmmc_mincer~rtc_main|state,
                     weights = ~weights,
                     data = df)
#' No labor share adjustment - columns (3) and (4)
reg_nolab <- feols(dlnwmmc_mincer~rtc_notheta,
                   weights = ~weights,
                   data = df)
reg_nolab_fe <- feols(dlnwmmc_mincer~rtc_notheta|state,
                      weights = ~weights,
                      data = df)
#' Nontraded price change set to zero - columns (5) and (6)
reg_notrad <- feols(dlnwmmc_mincer~rtc_nt,
                    weights = ~weights,
                    data = df)
reg_notrad_fe <- feols(dlnwmmc_mincer~rtc_nt|state,
                       weights = ~weights,
                       data = df)

#' Nontraded sector workers' wages - columns (7) and (8)
reg_workers <- feols(dlnwmmc_mincer_nt~rtc_main,
                     weights = ~weights_nt,
                     data = df)
reg_workers_fe <- feols(dlnwmmc_mincer_nt~rtc_main|state,
                        weights = ~weights_nt,
                        data = df)

# Ferman assessment -----------------------------------------------------
#' Performs the inference assessment provided in Ferman (2019).
#'
#' @param df Your database
#' @param model A character string in R's formula style defining the fixest model.
#' @param assess_on The variable the assessment will be taken on.
#' @param W share matrix
#' @param H0 Coefficient value under the null hypothesis, by default 0.0
#' @param nsim Number of simulations to run, by default 1000
#' @param alpha Significance level, by default 0.05
#' @param weights Weights variable name
#' @param cluster Cluster variable name. Package fixest must be loaded in order
#' to use this argument. It must agree with the cluster provided in the model
#'
#' @return Assessment value for the given level of significance.
#' @export
```

```r
#'
#' @examples
#' # NOT RUN
#' ferman_assessment(iris, "Sepal.Lenght~Sepal.Width+Petal.Width", "Petal.Width")
# Shift-share Ferman assessment ------------------------------------------

ss_ferman_assessment <- function(data, model, assess_on, W, H0 = 0.0, nsim = 1000,
                                 alpha = 0.05, weights = NULL, cluster = NULL) {
  # Coercing df to data.frame ONLY (no tibble or data.table)
  df <- as.data.frame(df)
  # No spaces allowed in model formula
  model <- gsub("\\s+", "", model)
  depvar <- sub("~.+", "", model)
  # Simulations sequence
  sim <- seq_len(nsim)
  # Rejections vector (of 0s and 1s)
  rejections <- c()
  # Regression weights
  weight <- df[, weights]
  # Holder for artificial data
  df_artificial <- df
  # Iterate nsim simulations
  for (i in sim) {
    # Placebo SS regressor
    df_artificial[, assess_on] <- W %*% rnorm(ncol(W))
    # Estimate model with placebo SS
    placebo_fit <- fixest::feols(as.formula(model), data = df_artificial,
                                 warn = FALSE, weights = weight)
    # Reject at specified significance?
    if (is.null(cluster)) {
      beta <- summary(placebo_fit)$coeftable[assess_on, 1]
      se_beta <- summary(placebo_fit)$coeftable[assess_on, 2]
      tstat <- abs((beta - H0)/se_beta)
    }
    else {
      # fixest must be loaded in order to clusters work!
      beta <- summary(placebo_fit, cluster = cluster)$coeftable[assess_on, 1]
      se_beta <- summary(placebo_fit, cluster = cluster)$coeftable[assess_on, 2]
      tstat <- abs((beta - H0)/se_beta)
    }
    # Test whether pvals < alpha and store in rejections
    rejections[i] <- ifelse(tstat > qnorm(1 - alpha/2), 1, 0)
  }
  # Return the mean of rejections
  return(mean(rejections))
}


#' Assessment on main model without fixed effects
base_df <- tibble(name = c("Main", "Main FE", "No labor", "No labor FE",
                           "Nontraded", "Nontraded FE",
                           "Workers", "Workers FE"),
                  assess_on = c(rep("rtc_main", 2), rep("rtc_notheta", 2),
```

```
                                rep("rtc_nt", 2), rep("rtc_main", 2)),
                  weights = c(rep("weights", 6), rep("weights_nt", 2)),
                  cluster = "state")
assessment_df <- base_df %>%
  mutate(model = c("dlnwmmc_mincer~rtc_main",
                   "dlnwmmc_mincer~rtc_main|state",
                   "dlnwmmc_mincer~rtc_notheta",
                   "dlnwmmc_mincer~rtc_notheta|state",
                   "dlnwmmc_mincer~rtc_nt",
                   "dlnwmmc_mincer~rtc_nt|state",
                   "dlnwmmc_mincer_nt~rtc_main",
                   "dlnwmmc_mincer_nt~rtc_main|state"),
         W_mat = list(W_main, W_main, W_notheta, W_notheta, W_nt, W_nt,
                      W_main, W_main),
         alpha = list(c(0.05, 0.10))) %>%
  unnest(cols = alpha) %>%
  rowwise() %>%
  mutate(assessment = ss_ferman_assessment(df, model, assess_on, W_mat,
                                           cluster = cluster, weights = weights,
                                           alpha = alpha))

assessment_tbl <- assessment_df %>%
  select(name, alpha, assessment) %>%
  pivot_wider(id_cols = name, names_from = alpha, values_from = assessment)

# Wild Bootstrap ----------------------------------------------------------
#Function that computes wild BS. Takes as arguments:
#formula: a model to test
#coef.to.test: character. name of the variable in formula whose coefficient we will test
#cluster.var: character. name of the cluster indicator variable
#data: dataframe where estimation will be conducted
#b: value of coefficient under the null. Defaults to 0
#S: Number of replications of step 2 in algorithm. Defaults to 1000
#dataset with variables
wild.bs <- function(data, formula, coef.to.test, cluster.var, weight.var = NULL,
                    b = 0, S = 1000)
{
  stopifnot(is_tibble(data))
  # No spaces allowed in model formula
  formula <- gsub("\\s+", "", formula)
  depvar <- sub("~.+", "", formula)
  # Weighted regression
  if (!is.null(weight.var))
    weight.vec <- data %>% pull(weight.var)
  else
    weight.vec <- NULL
  #Imposing the null in formula
  formula.null <- gsub(coef.to.test,
                       glue::glue("offset(b*{coef.to.test})"),
                       formula)

  modelo.nulo <- lm(as.formula(formula.null), weights = weight.vec, data = data)
```

```
  cluster.data <- data[, cluster.var]

  cluster.indexes <- unique(cluster.data)

  C <- nrow(cluster.indexes)

  vec_unstud <- c()
  vec_stud <- c()

  data.artificial <- data

  for (s in 1:S)
  {
    e_s = 1 - 2*rbinom(C, 1, 0.5)

    vals.cluster <- cbind(cluster.indexes, "e_s" = e_s)
    cluster.matched <- merge(cluster.data, vals.cluster, by = cluster.var)

    #Creating artificial data
    data.artificial[, depvar] <- modelo.nulo$fitted.values +
      cluster.matched$e_s*modelo.nulo$residuals

    modelo.s <- lm(as.formula(formula), weights = weight.vec,
                   data = data.artificial)

    coef.s <- modelo.s$coefficients[coef.to.test]

    vec_unstud <- c(vec_unstud, coef.s - b)

    se.s <- sqrt(
      diag(sandwich::vcovCL(modelo.s, cluster = cluster.data[,1])))[coef.to.test]

    vec_stud <- c(vec_stud,  (coef.s - b)/se.s)
  }

  #Compute estimates from the data now
  modelo.data <- lm(as.formula(formula), weights = weight.vec, data = data)

  coef.data <- modelo.data$coefficients[coef.to.test]

  p.val.unstud <- 1 - mean(abs(coef.data - b) > abs(vec_unstud))
  se.data <- sqrt(
    diag(sandwich::vcovCL(modelo.data, cluster = cluster.data[,1])))[coef.to.test]

  p.val.stud <- 1 - mean(abs((coef.data - b)/se.data) > abs(vec_stud))

  return(data.frame("Unstudentized p-value" = p.val.unstud,
                    "Studentized p-value" = p.val.stud))
}

wild_df <- base_df %>%
  mutate(model = c("dlnwmmc_mincer~rtc_main",
                   "dlnwmmc_mincer~rtc_main+factor(state)",
```

```
                    "dlnwmmc_mincer~rtc_notheta",
                    "dlnwmmc_mincer~rtc_notheta+factor(state)",
                    "dlnwmmc_mincer~rtc_nt",
                    "dlnwmmc_mincer~rtc_nt+factor(state)",
                    "dlnwmmc_mincer_nt~rtc_main",
                    "dlnwmmc_mincer_nt~rtc_main+factor(state)")) %>%
  rowwise() %>%
  mutate(wild = list(wild.bs(df, model, assess_on, cluster, weights))) %>%
  unnest(cols = wild)


# Permutation test - Randomization Inference ------------------------------
rand_inference <- function(df, model, assess_on, H0 = 0.0, nsim = 1000,
                           cluster = NULL, weights = NULL) {
  stopifnot(is_tibble(df))
  # No spaces allowed in model formula
  model <- gsub("\\s+", "", model)
  depvar <- sub("~.+", "", model)
  # Weights vector
  if (!is.null(weights))
    weights_vec <- df %>% pull(weights)

  # Regression with original data
  orig_model <- fixest::feols(as.formula(model), weights = weights_vec, data = df)
  orig_coef <- orig_model$coefficients[assess_on]
  if (!is.null(cluster))
    orig_se <- summary(orig_model, cluster = cluster)$se[assess_on]
  else
    orig_se <- summary(orig_model)$se[assess_on]
  # Original regression test statistic. Studentized
  orig_test <- abs(orig_coef - H0) / orig_se

  # Permutations
  # Clusterized permutations: number of clusters
  if (!is.null(cluster)) {
    n_cl <- nrow(unique(df[cluster]))
  }
  # Artificial data
  df_art <- df
  vec_sim <- vector(mode = "double", length = nsim)
  for (s in seq_len(nsim)) {
    if (!is.null(cluster)) {
      # Sample clusters without replacement
      cl_shuffle <- sample(n_cl)
      # Reorder only cluster and treatment
      cl_sample <- unique(df[cluster])[cl_shuffle, ] %>%
        right_join(df[c(cluster, assess_on)], by = cluster)
      # Replace cluster and treatment without reordering all else!
      df_art[c(cluster, assess_on)] <- cl_sample
      # Regression with artificial data
      art_model <- fixest::feols(as.formula(model), weights = weights_vec,
                                 data = df_art)
      art_coef <- art_model$coefficients[assess_on]
```

```
      art_se <- summary(art_model, cluster = df_art[cluster])$se[assess_on]
    } # clusterized permutation
    else {
      shuffle <- sample(nrow(df))
      # Reorder treatment
      tr_sample <- df[shuffle, assess_on]
      df_art[assess_on] <- tr_sample
      # Regression with artificial data
      art_model <- fixest::feols(as.formula(model), weights = weights_vec,
                                 data = df_art)
      art_coef <- art_model$coefficients[assess_on]
      art_se <- summary(art_model)$se[assess_on]
    }
    # Artificial regression test statistic. Studentized
    art_test <- abs(art_coef - H0) / art_se
    vec_sim[s] <- art_test
  } # end of for loop

  p_val <- 1 - mean(orig_test > vec_sim)
  return(c(exact_p_val = p_val))

}
ri_df <- base_df %>%
  mutate(model = c("dlnwmmc_mincer~rtc_main",
                   "dlnwmmc_mincer~rtc_main|state",
                   "dlnwmmc_mincer~rtc_notheta",
                   "dlnwmmc_mincer~rtc_notheta|state",
                   "dlnwmmc_mincer~rtc_nt",
                   "dlnwmmc_mincer~rtc_nt|state",
                   "dlnwmmc_mincer_nt~rtc_main",
                   "dlnwmmc_mincer_nt~rtc_main|state")) %>%
  rowwise() %>%
  mutate(ri_p_val = rand_inference(df, model, assess_on, cluster = cluster,
                                   weights = weights))

# Adao confidence interval ------------------------------------------------------
adao1 <- reg_ss("dlnwmmc_mincer~1", X = rtc_main , W = W_main,
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao2 <- reg_ss("dlnwmmc_mincer~state", X = rtc_main , W = W_main,
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao3 <- reg_ss("dlnwmmc_mincer~1", X = rtc_notheta , W = W_notheta,
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao4 <- reg_ss("dlnwmmc_mincer~state", X = rtc_notheta , W = W_notheta,
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao5 <- reg_ss("dlnwmmc_mincer~1", X = rtc_nt , W = W_nt,
```

```
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao6 <- reg_ss("dlnwmmc_mincer~state", X = rtc_nt , W = W_nt,
                weights = weights,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao7 <- reg_ss("dlnwmmc_mincer_nt~1", X = rtc_main , W = W_main,
                weights = weights_nt,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao8 <- reg_ss("dlnwmmc_mincer_nt~state", X = rtc_main , W = W_main,
                weights = weights_nt,
                region_cvar = state,
                method = "all", data = df)[["p"]]
adao_df <- base_df %>%
  mutate(adao_pval = list(adao1, adao2, adao3, adao4,
                          adao5, adao6, adao7, adao8)) %>%
  rowwise() %>%
  mutate(methods = list(names(adao_pval))) %>%
  unnest(cols = c(adao_pval, methods)) %>%
  select(name, adao_pval, methods) %>%
  pivot_wider(names_from = methods, values_from = adao_pval) %>%
  rename(Homo = Homoscedastic)

# Borusyak Hull Jaravel inference ------------------------------------------
source("II/BorusyakHullJaravel.R")

bhj_ivreg_ss("dlnwmmc_mincer~state|rtc_main", X = rtc_main , W = W_main,
             weights = weights,
             region_cvar = state,
             method = "all", data = df)

# Robustness specs ---------------------------------------------------------
unemp_main <- reg_ss("dlnwmmc_mincer~avg_unemp91", X = rtc_main , W = W_main,
                     weights = weights,
                     region_cvar = state,
                     method = "all", data = df)
unemp_main_fe <- reg_ss("dlnwmmc_mincer~avg_unemp91+state", X = rtc_main ,
                        W = W_main,
                        weights = weights,
                        region_cvar = state,
                        method = "all", data = df)
unemp_nolab <- reg_ss("dlnwmmc_mincer~avg_unemp91", X = rtc_notheta , W = W_notheta,
                      weights = weights,
                      region_cvar = state,
                      method = "all", data = df)
unemp_nolab_fe <- reg_ss("dlnwmmc_mincer~avg_unemp91+state", X = rtc_notheta , W = W_notheta,
                         weights = weights,
                         region_cvar = state,
                         method = "all", data = df)
unemp_notrad <- reg_ss("dlnwmmc_mincer~avg_unemp91", X = rtc_nt , W = W_nt,
                       weights = weights,
```

```
                        region_cvar = state,
                        method = "all", data = df)
unemp_notrad_fe <- reg_ss("dlnwmmc_mincer~avg_unemp91+state", X = rtc_nt , W = W_nt,
                          weights = weights,
                          region_cvar = state,
                          method = "all", data = df)
unemp_workers <- reg_ss("dlnwmmc_mincer_nt~avg_unemp91", X = rtc_main , W = W_main,
                        weights = weights_nt,
                        region_cvar = state,
                        method = "all", data = df)
unemp_workers_fe <- reg_ss("dlnwmmc_mincer_nt~avg_unemp91+state", X = rtc_main , W = W_main,
                           weights = weights_nt,
                           region_cvar = state,
                           method = "all", data = df)
robust_df <- tibble(description = c("Regional tariff change", names(unemp_main$se)),
                    main = c(unemp_main$beta, unemp_main$se),
                    main_fe = c(unemp_main_fe$beta, unemp_main_fe$se),
                    nolab = c(unemp_nolab$beta, unemp_nolab$se),
                    nolab_fe = c(unemp_nolab_fe$beta, unemp_nolab_fe$se),
                    notrad = c(unemp_notrad$beta, unemp_notrad$se),
                    notrad_fe = c(unemp_notrad_fe$beta, unemp_notrad_fe$se),
                    workers = c(unemp_workers$beta, unemp_workers$se),
                    workers_fe = c(unemp_workers_fe$beta, unemp_workers_fe$se))


inferences_df <- wild_df %>%
  select(name, contains("p.value")) %>%
  left_join(ri_df[c("name", "ri_p_val")], by = "name") %>%
  left_join(adao_df, by = "name")
#' save image
save(list = ls(), file = "II/input/homework_II.RData")
```

Annex B - R Code for homework_II_census.R

```r
#' ## homework_II_census.R
#' Homework II - Microeconometrics II
#' Author: Rafael Felipe Bressan
#'
#' Paper: Regional Effects of Trade Reform: What Is the Correct Measure of
#' Liberalization? by Brian Kovak - AER2013
#' Loading libraries
library(tidyverse)
library(dtplyr)
library(data.table)
library(haven)

#' Part II
#' Load data
# load("input/homework_II.RData")
folder <- "II/input/Kovak2013/AER-2011-0545_data/"
# census_sample ------------------------------------------------------------
cnaedom_to_indmatch <- read_dta(paste0(folder, "cnaedom_to_indmatch.dta")) %>%
  as.data.table()
pnad_to_indmatch <- read_dta(paste0(folder, "pnad_to_indmatch.dta")) %>%
  as.data.table()
microreg_to_mmc <- read_dta(paste0(folder, "microreg_to_mmc.dta")) %>%
  as.data.table()
ibge_wagebill <- read_dta(paste0(folder, "ibge_wagebill.dta")) %>%
  as.data.table()
ibge_value_added <- read_dta(paste0(folder, "ibge_value_added.dta"))
niv50_to_indmatch <- read_dta(paste0(folder, "niv50_to_indmatch.dta"))

#' Reading census
# ibge_census_2000 <- read_dta(paste0(folder, "ibge_census_2000.dta")) %>%
#   as.data.table()
# gc()
# ibge_census_2000 <-
#   ibge_census_2000[age %in% 18:55 & ymain > 0 & !is.na(ymain) & inschool == 0]
# gc()
#' 5517219 row after filtering
# print(object.size(ibge_census_2000), units = "GB")
ibge_census_1991 <- read_dta(paste0(folder, "ibge_census_1991.dta")) %>%
  as.data.table()
gc()
ibge_census_1991 <-
  ibge_census_1991[age %in% 18:55 & ymain > 0 & !is.na(ymain) & inschool == 0]
gc()
#' 4739171 rows after filtering
print(object.size(ibge_census_1991), units = "GB")
#' calculate log real wage and generate remaining wage regression variables
#' real wages 2000 base year
ibge_census_1991[, `:=`(wagemain = (ymain/4.33)/hmain,
                        agesq = age^2 / 1000,
                        city = (urbanrural == 1))]
ibge_census_1991[, rwagemain := (wagemain/2750000)/(0.000067244146018/0.890629059684618)]
ibge_census_1991[, lnrwmain := log(rwagemain)]
```

```
gc()
#' recode industries to IndMatch
#' 1991 column atividade matches to pnad
ibge_census_1991 <- merge(ibge_census_1991, pnad_to_indmatch, by = "atividade",
                          all.x = TRUE)
#' 2000 column cnae matches to cnaedom
# ibge_census_2000 <- merge(ibge_census_2000, cnaedom_to_indmatch,
#                           by.x = "cnae", by.y = "cnaedom", all.x = TRUE)
# setnames(ibge_census_2000, "cnae", "atividade")
gc()

#' Bind the two censuses
# ibge_census <- rbind(ibge_census_1991, ibge_census_2000)
# rm(ibge_census_1991, ibge_census_2000)
# gc()

#' merge in consistent microregion codes
ibge_census_1991 <- merge(ibge_census_1991, microreg_to_mmc, by = "microreg",
                          all.x = TRUE)

#' Census sample
census_sample <- ibge_census_1991
rm(ibge_census_1991)

# figure_1_b1_b2_b5 ------------------------------------------------------
tariff_chg <- read_dta(paste0(folder, "kume_etal_tariff.dta")) %>%
  left_join(niv50_to_indmatch, by = "niv50") %>%
  filter(indmatch != 99) %>%
  left_join(ibge_value_added, by = "niv50") %>%
  group_by(indmatch, indname) %>%
  summarise(across(tariff1987:tariff1998, ~weighted.mean(.x, w = va))) %>%
  mutate(dlnonetariff9095 = log(1 + (tariff1995/100)) - log(1 + (tariff1990/100))) %>%
  select(indmatch, indname, dlnonetariff9095)

# figure_2_b3 ------------------------------------------------------------


# figure_3_4_b4 ----------------------------------------------------------
#' Using lambda and theta from Kovak's Stata
# lambda_kovak <- haven::read_dta("II/input/lambda.dta")
# theta_kovak <- haven::read_dta("II/input/theta.dta")
#
# lambda <- lambda_kovak %>%
#   pivot_longer(cols = lambda1:lambda99, names_to = "indmatch",
#                values_to = "lambda") %>%
#   mutate(indmatch = as.numeric(str_extract(indmatch, "\\d+")))
# theta <- theta_kovak
#' With lambda and theta taken directly from Kovak the results are THE SAME!!
#' More evidence my computations of share weights are correct

#' calculate fixed factor share of input cost (theta)
theta <- ibge_wagebill %>%
  as_tibble() %>%
```

```
  left_join(niv50_to_indmatch, by = "niv50") %>%
  group_by(indmatch, indname) %>%
  summarise(across(wagebill:factorcost, sum)) %>%
  mutate(theta = 1 - (wagebill/factorcost))

#' calculate industry weights for each region
lambda <- census_sample %>%
  lazy_dt() %>%
  filter(!is.na(indmatch)) %>%
  select(mmc, xweighti, indmatch) %>%
  group_by(mmc, indmatch) %>%
  summarise(lambda = sum(xweighti)) %>%
  mutate(lambda = lambda / sum(lambda)) %>%
  as_tibble()
  # pivot_wider(names_from = indmatch, values_from = lambda, names_sort = TRUE)

#' weights omitting nontraded sector, with and without theta adjustment
weights_ss_notheta <- lambda %>%
  left_join(theta, by = "indmatch") %>%
  filter(indmatch != 99) %>%  # omit nontraded sector and colinear 16
  mutate(lambdaovertheta = lambda / theta) %>%
  group_by(mmc) %>%
  mutate(sumlambdaovertheta = sum(lambdaovertheta),
         sumlambda = sum(lambda),
         weight_main = lambdaovertheta / sumlambdaovertheta,
         weight_notheta = lambda / sumlambda)
#' weights including nontraded sector and theta adjustment
weights_ss_nt <- lambda %>%
  left_join(theta, by = "indmatch") %>%
  mutate(lambdaovertheta = lambda / theta) %>%
  group_by(mmc) %>%
  mutate(sumlambdaovertheta = sum(lambdaovertheta),
         weight_nt = lambdaovertheta / sumlambdaovertheta)

#' generate RTC. weighted averages of tariff changes
rtc_notheta <- weights_ss_notheta %>%
  select(mmc, indmatch, weight_main, weight_notheta) %>%
  left_join(tariff_chg, by = "indmatch") %>%
  group_by(mmc) %>%
  summarise(rtc_main = sum(weight_main * dlnonetariff9095),
            rtc_notheta = sum(weight_notheta * dlnonetariff9095))

rtc_nt <- weights_ss_nt %>%
  select(mmc, indmatch, weight_nt) %>%
  left_join(tariff_chg, by = "indmatch") %>%
  mutate(dlnonetariff9095 = if_else(indmatch == 99, 0, dlnonetariff9095)) %>%
  group_by(mmc) %>%
  summarise(rtc_nt = sum(weight_nt * dlnonetariff9095))

rtc_bressan <- rtc_notheta %>%
  left_join(rtc_nt, by = "mmc")
#' Load Kovak's rtc and compare
rtc_kovak <- read_dta("II/input/rtc.dta")
```

```r
cat("Maximum difference in rtc_main from Kovak's\n",
    max(abs(rtc_bressan$rtc_main - rtc_kovak$rtc_main)))

#' Preparing weight matrices to Adao Shift Share confidence intervals
weight_main <- weights_ss_notheta %>%
  filter(mmc != 13007) %>%  # drop Manaus
  select(mmc, indmatch, weight_main) %>%
  pivot_wider(names_from = indmatch, values_from = weight_main,
              names_sort = TRUE, values_fill = 0.0) %>%
  as.matrix()

weight_notheta <- weights_ss_notheta %>%
  filter(mmc != 13007) %>%  # drop Manaus
  select(mmc, indmatch, weight_notheta) %>%
  pivot_wider(names_from = indmatch, values_from = weight_notheta,
              names_sort = TRUE, values_fill = 0.0) %>%
  as.matrix()

weight_nt <- weights_ss_nt %>%
  filter(mmc != 13007) %>%  # drop Manaus
  select(mmc, indmatch, weight_nt) %>%
  pivot_wider(names_from = indmatch, values_from = weight_nt,
              names_sort = TRUE, values_fill = 0.0) %>%
  as.matrix()

#' Do not save large dataframes that start with "lg_"
save(weight_main, weight_notheta, weight_nt, rtc_bressan,
     file = "II/input/homework_II_Adao.RData")
#' Clean up environment
rm(list = ls())
gc()
```

References

Abadie, Alberto. 2020. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature*, no. forthcoming.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

Adao, Rodrigo, Michal Kolesár, and Eduardo Morales. 2019. "Shift-Share Designs: Theory and Inference." *The Quarterly Journal of Economics* 134 (4): 1949–2010.

Canay, Ivan A, Andres Santos, and Azeem M Shaikh. 2018. "The Wild Bootstrap with a" Small" Number of" Large" Clusters." *Review of Economics and Statistics*, 1–45.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review* 95 (4): 1237–58. https://doi.org/10.1257/0002828054825529.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. 2020. "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls." http://arxiv.org/abs/1712.09089.

Conley, Timothy G, and Christopher R Taber. 2011. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *The Review of Economics and Statistics* 93 (1): 113–25.

Ferman, Bruno. 2019. "A Simple Way to Assess Inference Methods." *arXiv Preprint arXiv:1912.08772*.

Ferman, Bruno, and Cristine Pinto. 2019. "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity." *The Review of Economics and Statistics* 101 (3): 452–67. https://EconPapers.repec.org/RePEc:tpr:restat:v:101:y:2019:i:3:p:452-467.

Goldberg, Pinelopi Koujianou, and Nina Pavcnik. 2007. "Distributional Effects of Globalization in Developing Countries." *Journal of Economic Literature* 45 (1): 39–82.

Kovak, Brian K. 2013. "Regional Effects of Trade Reform: What Is the Correct Measure of Liberalization?" *American Economic Review* 103 (5): 1960–76. https://doi.org/10.1257/aer.103.5.1960.

Kume, Honório, Guida Piani, and Carlos Frederico Souza. 2003. "A Política Brasileira de Importação No Período 1987-98: Descrição E Avaliação." In *A Abertura Comercial Brasileira Nos Anos 90: Impactos Sobre Emprego E Salário*, edited by Carlos Henrique Corseuil and Honório Kume, 9–37. Rio de Janeiro: IPEA.