Assignment 2 Report

By: Pujan Thakrar and Mark Vandre

Introduction

In their much-acclaimed paper titled *The China Syndrome*, David Autor, David Dorn and Gordon Hansen estimated the impact of China entering the World Trade Organization (WTO) on American employment at the commuting zone level. While this paper presented compelling evidence of China's impact on domestic employment, some critics have pointed to the lack of composition-adjusting and the choice of shock as sources of potential problems. By not composition-adjusting variables, Autor, Dorn, and Hanson have run the risk of attributing changes in employment to compositional changes to China's entering the WTO. This report attempts to correct for these issues.

PART 1: Constructing the Commuting Zone-Level Shock

In this section, we created composition adjusted outcome variables to address one of the primary criticisms identified for Autor et al.

Assumptions & Variables

We collected data from the 2000 ACS as well as the 2007 three-year ACS from IPUMS. The variables of interest in this analysis were the following: year, person weight, sex, age, race, state, PUMA, detailed education, industry, class worker, employment status, weeks worked 2, usual hours worked, and income wage.

Using an industry shock from Pierce and Schott (2016), we proceeded to create our shock variable to associate with each commuting zone. We first matched up each individual's PUMA to a commuting zone. We then merged in the shock variable to each individual according to their industry. We then created an industry-region weighting variable that represented the share of hours worked in each industry of each region in the year 2000. Finally, we multiplied the industry-region weight by the corresponding industry shock, and summed these values across regions. The result was a shock variable that represented the degree to which each region was impacted by China entering the WTO according to the number of hours worked in each industry affected.

PART 2: Construct Changes Between 2000 and 2007 in Commuting Zone Level Composition-Adjusted

In general, we were interested in determining how the China shock impacts the wages, unemployment rate, and labor force participation rates at the regional (commuting zone) level. For this part we construct the three variables by composition adjusting them and weighting them by commuting zone.

Variables & Assumptions

The data used is again from the 2007 3-year ACS and the 2000 5% survey from IPUMS. The variables of interest in this analysis were the following: age, education, race, industry, weeks worked 2, income wage, usual hours worked, state FIPS, PUMA, year, person weight, and employment status. We also use the commuting zone crosswalk data for the year 2000 from David Dorn's website.

For this part we kept only individuals that fell in the working age population (i.e. ages 16 and up). We also created labour bins that consisted of education-age-race groups. To do this we created education and race dummy variables. College educated or higher was a assigned a value of 1, while less that a college educated was assigned a value of 0. With race, we created a "white" dummy variable, where white individuals were assigned a value of 1 and non-white individuals were assigned a value of 0.

We also needed to re-code the values of weeks worked to be the actual number of weeks worked so that we can multiply it by usual hours worked in order to construct total hours worked. In addition, we top-coded income in the year 2000 by multiplying $200,000 by 1.5.

Similarly to before, we merged the commuting zone data by merging them by puma2000. This was created by creating 6-digit codes (first two digits are the state FIPS and the last four digits are the PUMA codes). We then create a new weight by multiplying the 'afactor' term by the person weights. This gives us the number of individuals an observation represents within a given commuting zone.

Finally, we only keep observations where total working hours and income wage is greater than 0.

Constructing the percent difference in composition adjusted average wage by commuting zone

Given that the data we have from IPUMS is at the individual level, we needed to aggregate our data down to the commuting zone level. We also wanted to composition adjust the wages as well in order to control for any compositional changes to wages that occur between 2000 and 2007.

To do this, we decide to composition adjust the average wages by total hours worked. This is because, it is possible that average hours worked has changed over time and thereby increasing or decreasing the average wage. Using our new data, we created composition adjusted weights by summing across all hours worked in in the year 2000. We then take the total hours worked for each commuting zone – labour bin pair and divide this by the total hours worked in the year 2000 for each observation. This gives us the composition adjusted weights.

To create the average wage, we took the income wage for each year-commuting zone-labor bin combination and divided it by total hours within the same group. We then take the log of the average wage from 2007 and subtract the log of the average wage from 2000 for each commuting zone-labour bin group. This gives us the log difference or percent change of average wages between 2000 and 2007 for each commuting zone – labor bin group. We then multiply the

log-differenced wages by their weights and then summed them by commuting zone in order to determine the composition adjusted average wage in a given commuting zone.

Constructing the percentage point difference in composition adjusted unemployment rate by commuting zone

To construct the unemployment rate, we take the employment status and remove all individuals that are not in the labor force. We then create a dummy variable where 1 is for unemployed and 0 is for employed.

In order to composition adjust the unemployment rate, we chose to use the size of the labor force as this would again be a key player in affecting the changes in the unemployment rate over time. We then follow a very similar process to that of constructing the composition adjusted log wages.

We create the unemployment rate which is given by the number of unemployed individuals and divided by the labor force size. To get the percentage point difference in the unemployment rate between 2000 and 2007, we simply subtract the unemployment rate of a commuting zone-labour bin group in 2000 from the unemployment rate of the same group in 2007. We then weight this similarly to before, where we take the total individuals in the labour force for a given commuting zone – labor bin combination and divide it by the total number of individuals in the labour force for 2000. After multiplying the weights by their respective differenced unemployment rates, we sum the values by commuting zone to get the percentage point difference in composition-adjusted unemployment rates by commuting zones.

Constructing the percentage point difference in composition adjusted labor force participation rate by commuting zone

Again, we follow the same process as before, however this time we composition adjust by the working-age population. By weighting accordingly, we get one observation for every commuting zone similarly to before. We have now constructed the three dependent variables we want to use in our regressions.

PART 3: OLS Estimation of Commuting Zone-Shock Effect on Outcomes of Interest

Once the region shocks and composition adjusted outcome variables were created, we then had to create region-level controls for our OLS estimation. The controls created include the difference in share of each region employed in manufacturing and the share of the working age population with a college degree from 2000 to 2007. After including our controls, the primary model being estimated is shown below

$$y_{k,r} = \beta_o + \beta_1 s_r + \beta_2 MFG_r + \beta_3 DEG_r + u_r$$

where $y_{k,r}$ is one of three outcome variables (k) for the r^{th} region. s_r is the regional shock variable, the construction of which was described in part one. MFG_r is the percentage point change in the share of each region working in manufacturing, and DEG_r is the percentage point

change in the share of each region with a college degree. Below are the results of the above regression for the difference in composition-adjusted log average wages.

Table 1: Log Difference in Average Wages

Variables	Log Wages		
	(1)	(2)	(3)
Shock	3.11E-07	2.68E-01	2.79E-07
	(6.93E-08)	(7.06E-08)	(7.10E-08)
ΔEducation	-	4.36E-01	4.76E-01
		(1.55E-01)	(1.57E-01)
ΔManufacturing	-	-	3.40E-01
			(2.45E-02)
Constant	4.05E-02	3.38E-02	3.68E-02
	(2.37E-03)	(3.35E-03)	4.00E-03
Adj R-Squared	0.028	0.038	0.039
Observations	665	665	665

As we can see, the shock coefficient remains significant, positive, and stable for each control we add to the model. While the shock coefficient is positive, it is not very large in magnitude. This could indicate that the shock's effect is not economically significant. According to this regression, a one-point increase in the shock value is associated with a increase in the differenced log wages for a region of 0.0000279 percent, all else equal.

Next, we show the regression table for the composition-adjusted difference in unemployment rate.

Table 2: Percentage Point Change in Unemployment

Variables	Unemployment Rate		
	(1)	(2)	(3)
Shock	-8.76E-10	-8.77E-10	-9.61E-10
	(1.47E-09)	(1.51E-09)	(1.52E-09)
ΔEducation	-	4.12E-06	-3.02E-04
		(3.31E-03)	(3.36E-03)
ΔManufacturing	-	-	-2.60E-04
			(5.24E-04)
Constant	1.46E-04	1.46E-04	1.22E-04
	(5.05E-05)	(7.17E-05)	(8.57E-05)

Adj R-Squared	-0.001	-0.002	-0.004
Observations	665	665	665

From this table we can see that the shock coefficient remains negative and significant, as well as relatively stable with each control added. As with the income case, this coefficient is not very large possibly indicating economic insignificance. A one-point increase in the shock variable is associated with a 0.0000000961 percentage point decrease in the unemployment rate of a region, all else equal.

Finally, we show the results for the regression involving labor force participation.

Figure 3: Percent Point Change in Labor Force

Variables	LF Participation Rate		
	(1)	(2)	(3)
Shock	-3.00E-08 (1.00E-08)	-2.72E-08 (1.03E-08)	-2.53E-08 (1.03E-08)
dEucation	-	-2.76E-02 (2.52E-02)	-2.08E-02 (2.29E-02)
dManufacturing	-	-	5.81E-03 (3.56E-03)
Constant	4.28E-03 (3.44E-04)	4.70E-03 (4.88E-04)	5.22E-03 (5.83E-04)
Adj R-Squared	0.012	0.013	0.019
Observations	665	665	665

Here again, the coefficient on the shock variable is statistically significant and negative after adding in the control variables. Also, the coefficient is not very large. Here, a one-point increase in the shock variable is associated with a 0.00000253 percentage point decrease in the labor force participation rate.

Conclusion

Given the results of Autor, Dorn, and Hanson, these results seem surprising at first glance. One would expect an effect of greater magnitude and one that does not increase wages and lower unemployment. However, these results do make sense in the context of standard trade theory. We would expect to see regions generally better off as trade becomes freer and both nations can enjoy the gains from trade. One way of interpreting these results is that, ignoring individual industries, the gains from trade and the negative impacts of trade roughly cancel each other out.

Appendix

Pujan Thakrar, and Mark Vandre

May 7, 2019

```
###############
#load in data#
###############

library(haven)
library(ipumsr)

#read in shock data and remove NAs for 1990 shock
shock_data = read_dta("ind2000_NTRgap.dta")
shock_data[is.na(shock_data$s1990),c(2,3)]=0

#read in ipyms data
ddi <- read_ipums_ddi("usa_00011.xml")
ipums <- read_ipums_micro(ddi)
save(ipums,file="ipums.RData")

#read in commuting zone data for the year 2000
CZ_2000 <- read_dta("cw_puma2000_czone.dta")

#############
#clean data#
#############

#keep working age
ipums = ipums[ipums$AGE>15,]
#create dymmy for white
ipums$RACE =  mapply(function(x){ifelse(x==1,x,0)}, ipums$RACE)
#remove unwanted data
ipums = ipums[,-c(8,10,12,14)]
#create dummy for US birthplace
ipums$BPL =  mapply(function(x){ifelse(x>149,1,0)}, ipums$BPL)
#create dummy for college educated or not
ipums$EDUC =  mapply(function(x){ifelse(x>9,1,0)}, ipums$EDUC)
#keep only employed
ipums = ipums[ipums$EMPSTAT==1,]
ipums = ipums[,-c(11)]

#create manufacturing data - if individuals are in manufacturing in the year 2000
ipums$MANU = NA
ipums[ipums$IND>106 & ipums$IND<400 & ipums$YEAR==2000,"MANU"]=1
ipums[is.na(ipums$MANU) & ipums$YEAR==2000,"MANU"]=0

#recode weeks worked
ipums$WKSWORK = NA
ipums[ipums$WKSWORK2==0,"WKSWORK"]=0
ipums[ipums$WKSWORK2==1,"WKSWORK"]=7
ipums[ipums$WKSWORK2==2,"WKSWORK"]=20
```

```
ipums[ipums$WKSWORK2==3,"WKSWORK"]=33
ipums[ipums$WKSWORK2==4,"WKSWORK"]=43.5
ipums[ipums$WKSWORK2==5,"WKSWORK"]=48.5
ipums[ipums$WKSWORK2==6,"WKSWORK"]=51

#top code income
ipums[ipums$INCWAGE==200000 & ipums$YEAR==2000,"INCWAGE"]=1.5*ipums$INCWAGE

#create total hours worked
ipums$THRS = ipums$UHRSWORK*ipums$WKSWORK
save(ipums,file="ipums2.RData")

#create puma2000 and merge with czones
ipums$STATEFIP=formatC(ipums$STATEFIP,width=2,flag="0")
ipums$PUMA=formatC(ipums$PUMA,width=4,flag="0")
ipums$puma2000=paste(ipums$STATEFIP,ipums$PUMA,sep = "")
save(ipums,file="ipums3.RData")
ipums = merge(ipums,CZ_2000,by="puma2000")
ipums$weight = ipums$PERWT*ipums$afactor
save(ipums,file="ipums4.RData")

########################
#Create regional shocks#
########################

#keep year 2000 to create regional weight in the year 2000
ipums= ipums[ipums$YEAR==2000,]
colnames(shock_data)=paste(c("IND","s1990","s1999"))
#merge census data with shocks by industry
ipums = merge(ipums,shock_data,by="IND")
#create industy_czone bins
ipums$ir_bin = paste(ipums$IND,ipums$czone,sep = "")

#keep where total hours worked is greater than 0
ipums = ipums[ipums$THRS>0,]
#sum total hours at the industry-czone level
x = aggregate(ipums$THRS,by=list(ir_bin = ipums$ir_bin),FUN=sum)
colnames(x)=paste(c("ir_bin","sum_thrs_ir"))
#merge ipums with aggregated hours
ipums = merge(ipums,x,by="ir_bin")
#create industry-czone weights - share of hours worked in industry i w/in region r in year 2000
ipums$w_ir = ipums$weight*(ipums$THRS/ipums$sum_thrs_ir)
#multiply weights by their associated shock
ipums$shock_weight = ipums$s1999*ipums$w_ir
#sum accross industry-region to get industry-regional shocks
ipums_cz_shock = aggregate(ipums$shock_weight,by=list(ir_bin = ipums$ir_bin),FUN=sum)
colnames(ipums_cz_shock)=paste(c("ir_bin","S_ir"))
ipums = merge(ipums,ipums_cz_shock,by="ir_bin")
#sum across regions to get regional shocks
ipums_regional_shock = aggregate(ipums$S_ir,by=list(czone = ipums$czone),FUN=sum)
#final regional shock
save(ipums_regional_shock,file="ipums_regional_shock.RData")
```

```
#########################################################################
#Construct ln change in CZ level composition- adjusted average wage#
#########################################################################


rm(CZ_2000,ipums,ipums_cz_shock,shock_data,x)

#load in data
load("ipums4.RData")
#keep total hours greater than 0
ipums = ipums[ipums$THRS>0,]

#keep if income wage is positive
ipums = ipums[ipums$INCWAGE>0,]
#create lbins - Education x Rage x Age
ipums$lbin = paste(ipums$EDUC,ipums$RACE,ipums$AGE,sep = "")

#get total income by czone
ipums_cz_wage = aggregate(ipums$INCWAGE,by=list(czone = ipums$czone),FUN=sum)
colnames(ipums_cz_wage)=paste(c("czone","WAGE_CZ"))
#get total hours worked by czone
ipums_cz_hrs = aggregate(ipums$THRS,by=list(czone = ipums$czone),FUN=sum)
colnames(ipums_cz_hrs)=paste(c("czone","THRS_CZ"))

ipums = merge(ipums,ipums_cz_wage,by="czone")
ipums = merge(ipums,ipums_cz_hrs,by="czone")

#create average wage (total income/total hours for each czone)
ipums$AVWAGE = ipums$WAGE_CZ/ipums$THRS_CZ
#remove if average wage is not positive
ipums = ipums[ipums$AVWAGE>0,]
#compute the logs of average wage
ipums$AVWAGE = log(ipums$AVWAGE)

#create czone-lbin groups
ipums$cz_lbin = paste(ipums$czone,ipums$lbin,sep = "")
#compute total hours worked for each unique czone-lbin group
ipums_cz_lbin_hrs = aggregate(ipums$THRS,by=list(cz_lbin = ipums$cz_lbin),FUN=sum)
colnames(ipums_cz_lbin_hrs)=paste(c("cz_lbin","THRS_CZ_LBIN"))
ipums = merge(ipums,ipums_cz_lbin_hrs,by="cz_lbin")

#create czone-year groups
ipums$cz_year = paste(ipums$czone,ipums$YEAR,sep = "")
#compute total hours for each czone in year 2000
ipums_cz_year_2000_hrs = aggregate(ipums$THRS,by=list(cz_year = ipums$cz_year),FUN=sum)
colnames(ipums_cz_year_2000_hrs)=paste(c("cz_year","THRS_CZ_2000"))
ipums = merge(ipums,ipums_cz_year_2000_hrs,by="cz_year")
#compute composition adjusted weights for each czone-lbin combination
ipums$weight_hrs = ipums$THRS_CZ_LBIN/ipums$THRS_CZ_2000


#convert years to 2-digits
ipums$year_short = mapply(function(x){ifelse(x==2000,20,27)}, ipums$YEAR)
```

```r
#create cz-lbin-year groups
ipums$year_cz_lbin = paste(ipums$year_short,ipums$cz_lbin,sep = "")
#get total hours worked for each cz-lbin-year group
ipums_ctl = aggregate(ipums$THRS,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl)=paste(c("year_cz_lbin","THRS"))
save(ipums_ctl,file="ipumsCTL.RData")
#get total income for each cz-lbin-year group
ipums_ctl_wage = aggregate(ipums$INCWAGE,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl_wage)=paste(c("year_cz_lbin","INCWAGE"))
save(ipums_ctl_wage,file="ipumsCTLW.RData")
ipums_ctl = merge(ipums_ctl,ipums_ctl_wage,by="year_cz_lbin")
save(ipums_ctl,file="ipumsCTL1.RData")
#split bins into individual variables
ipums_ctl$YEAR = gsub("([0-9][0-9]).*","\\1",as.character(ipums_ctl$year_cz_lbin))
ipums_ctl$czone = gsub("[0-9][0-9](.*)[0-9][0-9][0-9][0-9]","\\1",ipums_ctl$year_cz_lbin)
ipums_ctl$lbin = gsub("[0-9][0-9].*([0-9][0-9][0-9][0-9])","\\1",ipums_ctl$year_cz_lbin)
ipums_ctl$YEAR = mapply(function(x){ifelse(x==20,2000,2007)}, ipums_ctl$YEAR)
ipums_ctl$YEAR= as.integer(ipums_ctl$YEAR)
ipums_ctl$czone= as.integer(ipums_ctl$czone)
ipums_ctl$lbin= as.integer(ipums_ctl$lbin)

#create average wage
ipums_ctl$AVWAGE = ipums_ctl$INCWAGE/ipums_ctl$THRS
#create czone-lbin groups
ipums_ctl$cz_lbin = paste(ipums_ctl$czone,ipums_ctl$lbin,sep="")
ipums_ctl$cz_lbin = as.integer(ipums_ctl$cz_lbin)
#take logs of average wage
ipums_ctl$AVWAGE = log(ipums_ctl$AVWAGE)

library(dplyr)
#take log difference of average wage
ipums_ctl_new = ipums_ctl %>%
  group_by(cz_lbin) %>%
  mutate(dlnWage = AVWAGE - lag(AVWAGE))%>%
  ungroup()

#ipums_ctl_new[ipums_ctl_new$cz_lbin==99001189,]
ipums_ctl_new = ipums_ctl_new[!(is.na(ipums_ctl_new$dlnWage)),]
ipums_ctl_new = ipums_ctl_new[,-c(4)]
ipums_ctl_new = ipums_ctl_new[,-c(1)]
ipums_ctl_new = merge(ipums_ctl_new,ipums_cz_lbin_hrs,by="cz_lbin")
ipums_ctl_new = merge(ipums_ctl_new,ipums_cz_hrs,by="czone")

#create share of hours worked for each cz-libin group ot of cz
ipums_ctl_new$weight = ipums_ctl_new$THRS_CZ_LBIN/ipums_ctl_new$THRS_CZ
#share of wage change that comes from each industry in a particular cz
ipums_ctl_new$dlnWage_weight = ipums_ctl_new$weight*ipums_ctl_new$dlnWage
#wage effect by czone
ipums_c_lnWage = aggregate(ipums_ctl_new$dlnWage_weight,by=list(czone=ipums_ctl_new$czone),FUN=sum)
colnames(ipums_c_lnWage)=paste(c("czone","dlnWage"))
save(ipums_c_lnWage,file="lwage.RData")

##############################################################################
```

```
#Construct ln change in CZ level composition- adjusted unemployment rate#
###############################################################################


#load in data
load("ipums.RData")
#read in commuting zone data for the year 2000
CZ_2000 <- read_dta("cw_puma2000_czone.dta")

#############
#clean data#
#############

#keep working age
ipums = ipums[ipums$AGE>15,]
#create dymmy for white
ipums$RACE =  mapply(function(x){ifelse(x==1,x,0)}, ipums$RACE)
#remove unwanted data
ipums = ipums[,-c(8,10,12,14)]
#create dummy for US birthplace
ipums$BPL =  mapply(function(x){ifelse(x>149,1,0)}, ipums$BPL)
#create dummy for college educated or not
ipums$EDUC =  mapply(function(x){ifelse(x>9,1,0)}, ipums$EDUC)

#create manufacturing data - if individuals are in manufacturing in the year 2000
ipums$MANU = NA
ipums[ipums$IND>106 & ipums$IND<400 & ipums$YEAR==2000,"MANU"]=1
ipums[is.na(ipums$MANU) & ipums$YEAR==2000,"MANU"]=0

#recode weeks worked
ipums$WKSWORK = NA
ipums[ipums$WKSWORK2==0,"WKSWORK"]=0
ipums[ipums$WKSWORK2==1,"WKSWORK"]=7
ipums[ipums$WKSWORK2==2,"WKSWORK"]=20
ipums[ipums$WKSWORK2==3,"WKSWORK"]=33
ipums[ipums$WKSWORK2==4,"WKSWORK"]=43.5
ipums[ipums$WKSWORK2==5,"WKSWORK"]=48.5
ipums[ipums$WKSWORK2==6,"WKSWORK"]=51

#top code income
ipums[ipums$INCWAGE==200000 & ipums$YEAR==2000,"INCWAGE"]=1.5*ipums$INCWAGE

#create total hours worked
ipums$THRS = ipums$UHRSWORK*ipums$WKSWORK


#create puma2000 and merge with czones
ipums$STATEFIP=formatC(ipums$STATEFIP,width=2,flag="0")
ipums$PUMA=formatC(ipums$PUMA,width=4,flag="0")
ipums$puma2000=paste(ipums$STATEFIP,ipums$PUMA,sep = "")
save(ipums,file="ipums3.RData") #we overwrite ipums3 here
ipums = merge(ipums,CZ_2000,by="puma2000")
ipums$weight = ipums$PERWT*ipums$afactor
```

```
save(ipums, file = "ipums5.RData")

load("ipums5.RData")
#keep total hours greater than 0
ipums = ipums[ipums$THRS>0,]

#keep if income wage is positive
ipums = ipums[ipums$INCWAGE>0,]
#create lbins - Education x Rage x Age
ipums$lbin = paste(ipums$EDUC,ipums$RACE,ipums$AGE,sep = "")
#keep if in labor force
ipums = ipums[ipums$EMPSTAT<3,]
ipums = ipums[ipums$EMPSTAT>0,]
#create dummy for empstat
ipums$EMPSTAT=  mapply(function(x){ifelse(x==1,0,1)}, ipums$EMPSTAT)

#create czone-lbin groups
ipums$cz_lbin = paste(ipums$czone,ipums$lbin,sep = "")
#count number of unemployed /employed people
ipums$unemployed_count = ipums$weight*ipums$EMPSTAT
#compute total people in labour force for each unique czone-lbin group
ipums_cz_lbin_lf = aggregate(ipums$weight,by=list(cz_lbin = ipums$cz_lbin),FUN=sum)
colnames(ipums_cz_lbin_lf)=paste(c("cz_lbin","lf_size"))
ipums = merge(ipums,ipums_cz_lbin_lf,by="cz_lbin")

#create unemployment rate
ipums=ipums[ipums$lf_size>0,]
ipums$unemp_rate = ipums$unemployed_count/ipums$lf_size


#create czone-year groups
ipums$cz_year = paste(ipums$czone,ipums$YEAR,sep = "")
#compute total labor force size for each czone in year 2000
#(this will be our fixed year for composition adjusting)
ipums_cz_year_2000_lf = aggregate(ipums$lf_size,by=list(cz_year = ipums$cz_year),FUN=sum)
colnames(ipums_cz_year_2000_lf)=paste(c("cz_year","lf_size_CZ_2000"))
ipums = merge(ipums,ipums_cz_year_2000_lf,by="cz_year")
#compute composition adjusted weights for each czone-lbin combination
ipums$weight_lf = (ipums$lf_size)/(ipums$lf_size_CZ_2000)


#convert years to 2-digits
ipums$year_short = mapply(function(x){ifelse(x==2000,20,27)}, ipums$YEAR)
#create cz-lbin-year groups
ipums$year_cz_lbin = paste(ipums$year_short,ipums$cz_lbin,sep = "")
#get total lf size worked for each cz-lbin-year group
ipums_ctl_lf = aggregate(ipums$lf_size,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl_lf)=paste(c("year_cz_lbin","lf_size"))
save(ipums_ctl_lf,file="ipumsCTL_LF.RData")
#get total unemployed for each cz-lbin-year group
ipums_ctl_unemp = aggregate(ipums$unemployed_count,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl_unemp)=paste(c("year_cz_lbin","UNEMP"))
save(ipums_ctl_unemp,file="ipumsCTLU.RData")
```

```
ipums_ctl_lf = merge(ipums_ctl_lf,ipums_ctl_unemp,by="year_cz_lbin")
save(ipums_ctl_lf,file="ipumsCTL_LF1.RData")
#split bins into individual variables
ipums_ctl_lf$YEAR = gsub("([0-9][0-9]).*","\\1",ipums_ctl_lf$year_cz_lbin)
ipums_ctl_lf$czone = gsub("[0-9][0-9](.*)[0-9][0-9][0-9][0-9]","\\1",ipums_ctl_lf$year_cz_lbin)
ipums_ctl_lf$lbin = gsub("[0-9][0-9].*([0-9][0-9][0-9][0-9])","\\1",ipums_ctl_lf$year_cz_lbin)
ipums_ctl_lf$YEAR = mapply(function(x){ifelse(x==20,2000,2007)}, ipums_ctl_lf$YEAR)
ipums_ctl_lf$YEAR= as.integer(ipums_ctl_lf$YEAR)
ipums_ctl_lf$czone= as.integer(ipums_ctl_lf$czone)
ipums_ctl_lf$lbin= as.integer(ipums_ctl_lf$lbin)


#create unemployment rate
ipums_ctl_lf$UNEMP_R = ipums_ctl_lf$UNEMP/ipums_ctl_lf$lf_size
#create czone-lbin groups
ipums_ctl_lf$cz_lbin = paste(ipums_ctl_lf$czone,ipums_ctl_lf$lbin,sep="")
ipums_ctl_lf$cz_lbin = as.integer(ipums_ctl_lf$cz_lbin)
#take logs of average wage
#ipums_ctl_lf$UNEMP_R = log(ipums_ctl_lf$UNEMP_R)

library(dplyr)
#take difference of unemployment rate
ipums_ctl_new_lf = ipums_ctl_lf %>%
  group_by(cz_lbin) %>%
  mutate(dUnemp = UNEMP_R - lag(UNEMP_R))%>%
  ungroup()

#ipums_ctl_new_lf[ipums_ctl_new_lf$cz_lbin==99001189,]
ipums_ctl_new_lf = ipums_ctl_new_lf[!(is.na(ipums_ctl_new_lf$dUnemp)),]
ipums_ctl_new_lf = ipums_ctl_new_lf[,-c(4)]
ipums_ctl_new_lf = ipums_ctl_new_lf[,-c(1)]
ipums_ctl_new_lf = merge(ipums_ctl_new_lf,ipums_cz_lbin_lf,by="cz_lbin")
ipums_cz_lbin_lf$czone = gsub("(.*)[0-9][0-9][0-9][0-9]$","\\1",ipums_cz_lbin_lf$cz_lbin)
ipums_cz_lf = aggregate(ipums_cz_lbin_lf$lf_size,by=list(czone = ipums_cz_lbin_lf$czone),FUN=sum)
colnames(ipums_cz_lf) = paste(c("czone","lf_size_cz"))
ipums_ctl_new_lf = merge(ipums_ctl_new_lf,ipums_cz_lf,by="czone")

#create share of labor force for each cz-libin group ot of cz
ipums_ctl_new_lf$weight = ipums_ctl_new_lf$lf_size.y/ipums_ctl_new_lf$lf_size_cz
#share of labor force change that comes from each industry in a particular cz
ipums_ctl_new_lf$dUnemp_weight = ipums_ctl_new_lf$weight*ipums_ctl_new_lf$dUnemp
#wage effect by czone
ipums_c_Unemp = aggregate(ipums_ctl_new_lf$dUnemp_weight,by=list(czone=ipums_ctl_new_lf$czone),FUN=sum)
colnames(ipums_c_Unemp)=paste(c("czone","Unemp"))
save(ipums_c_Unemp,file="unemp.RData")

##############################################################################################
#Construct lab force participation change in CZ level composition- adjusted unemployment rate#
##############################################################################################


############
#clean data#
```

```
############


load("ipums5.RData")
#keep total hours greater than 0
ipums = ipums[ipums$THRS>0,]

#keep if income wage is positive
ipums = ipums[ipums$INCWAGE>0,]
#create lbins - Education x Rage x Age
ipums$lbin = paste(ipums$EDUC,ipums$RACE,ipums$AGE,sep = "")

#create dummy for empstat
ipums$EMPSTAT=  mapply(function(x){ifelse(x==3,0,1)}, ipums$EMPSTAT)

#create czone-lbin groups
ipums$cz_lbin = paste(ipums$czone,ipums$lbin,sep = "")
#count number of people in labor force
ipums$lf_count = ipums$weight*ipums$EMPSTAT
#compute total people in working age for each unique czone-lbin group
ipums_cz_lbin_wa = aggregate(ipums$weight,by=list(cz_lbin = ipums$cz_lbin),FUN=sum)
colnames(ipums_cz_lbin_wa)=paste(c("cz_lbin","wa_size"))
ipums = merge(ipums,ipums_cz_lbin_wa,by="cz_lbin")

#create labor force participation rate
ipums=ipums[ipums$wa_size>0,]
ipums$lf_rate = ipums$lf_count/ipums$wa_size


#create czone-year groups
ipums$cz_year = paste(ipums$czone,ipums$YEAR,sep = "")
#compute total working age size for each czone in year 2000 (this will be our fixed year for compositio
ipums_cz_year_2000_wa = aggregate(ipums$wa_size,by=list(cz_year = ipums$cz_year),FUN=sum)
colnames(ipums_cz_year_2000_wa)=paste(c("cz_year","wa_size_CZ_2000"))
ipums = merge(ipums,ipums_cz_year_2000_wa,by="cz_year")
#compute composition adjusted weights for each czone-lbin combination
ipums$weight_wa = (ipums$wa_size)/(ipums$wa_size_CZ_2000)


#convert years to 2-digits
ipums$year_short = mapply(function(x){ifelse(x==2000,20,27)}, ipums$YEAR)
#create cz-lbin-year groups
ipums$year_cz_lbin = paste(ipums$year_short,ipums$cz_lbin,sep = "")
#get total wa size worked for each cz-lbin-year group
ipums_ctl_wa = aggregate(ipums$wa_size,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl_wa)=paste(c("year_cz_lbin","wa_size"))
save(ipums_ctl_wa,file="ipumsCTL_WA.RData")
#get total labor force for each cz-lbin-year group
ipums_ctl_lf = aggregate(ipums$lf_count,by=list(year_cz_lbin=ipums$year_cz_lbin),FUN=sum)
colnames(ipums_ctl_lf)=paste(c("year_cz_lbin","LF"))
save(ipums_ctl_lf,file="ipumsCTLLF.RData")
ipums_ctl_wa = merge(ipums_ctl_wa,ipums_ctl_lf,by="year_cz_lbin")
save(ipums_ctl_wa,file="ipumsCTL_WA.RData")
```

```
#split bins into individual variables
ipums_ctl_wa$YEAR = gsub("([0-9][0-9]).*","\\1",ipums_ctl_wa$year_cz_lbin)
ipums_ctl_wa$czone = gsub("[0-9][0-9](.*)[0-9][0-9][0-9][0-9]","\\1",ipums_ctl_wa$year_cz_lbin)
ipums_ctl_wa$lbin = gsub("[0-9][0-9].*([0-9][0-9][0-9][0-9])","\\1",ipums_ctl_wa$year_cz_lbin)
ipums_ctl_wa$YEAR = mapply(function(x){ifelse(x==20,2000,2007)}, ipums_ctl_wa$YEAR)
ipums_ctl_wa$YEAR= as.integer(ipums_ctl_wa$YEAR)
ipums_ctl_wa$czone= as.integer(ipums_ctl_wa$czone)
ipums_ctl_wa$lbin= as.integer(ipums_ctl_wa$lbin)


#create labor force participation rate
ipums_ctl_wa$LF_R = ipums_ctl_wa$LF/ipums_ctl_wa$wa_size
#create czone-lbin groups
ipums_ctl_wa$cz_lbin = paste(ipums_ctl_wa$czone,ipums_ctl_wa$lbin,sep="")
ipums_ctl_wa$cz_lbin = as.integer(ipums_ctl_wa$cz_lbin)
#take logs of average wage
#ipums_ctl_lf$UNEMP_R = log(ipums_ctl_lf$UNEMP_R)

library(dplyr)
#take difference of labor force participation rate
ipums_ctl_new_wa = ipums_ctl_wa %>%
  group_by(cz_lbin) %>%
  mutate(dLF = LF_R - lag(LF_R))%>%
  ungroup()

#ipums_ctl_new_wa[ipums_ctl_new_wa$cz_lbin==99001189,]
ipums_ctl_new_wa = ipums_ctl_new_wa[!(is.na(ipums_ctl_new_wa$dLF)),]
ipums_ctl_new_wa = ipums_ctl_new_wa[,-c(4)]
ipums_ctl_new_wa = ipums_ctl_new_wa[,-c(1)]
ipums_ctl_new_wa = merge(ipums_ctl_new_wa,ipums_cz_lbin_wa,by="cz_lbin")
ipums_cz_lbin_wa$czone = gsub("(.*)[0-9][0-9][0-9][0-9]$","\\1",ipums_cz_lbin_wa$cz_lbin)
ipums_cz_wa = aggregate(ipums_cz_lbin_wa$wa_size,by=list(czone = ipums_cz_lbin_wa$czone),FUN=sum)
colnames(ipums_cz_wa) = paste(c("czone","wa_size_cz"))
ipums_ctl_new_wa = merge(ipums_ctl_new_wa,ipums_cz_wa,by="czone")

#create share of labor force for each cz-libin group ot of cz
ipums_ctl_new_wa$weight = ipums_ctl_new_wa$wa_size.y/ipums_ctl_new_wa$wa_size_cz
#share of labor force change that comes from each industry in a particular cz
ipums_ctl_new_wa$dLF_weight = ipums_ctl_new_wa$weight*ipums_ctl_new_wa$dLF
#wage effect by czone
ipums_c_LF = aggregate(ipums_ctl_new_wa$dLF_weight,by=list(czone=ipums_ctl_new_wa$czone),FUN=sum)
colnames(ipums_c_LF)=paste(c("czone","LF"))
save(ipums_c_LF,file="LF.RData")

#################
#Create controls#
#################

load("ipums5.RData")
#keep total hours greater than 0
ipums = ipums[ipums$THRS>0,]

#keep if income wage is positive
```

```
ipums = ipums[ipums$INCWAGE>0,]
#create lbins - Education x Rage x Age
ipums$lbin = paste(ipums$EDUC,ipums$RACE,ipums$AGE,sep = "")

#create manufacturing data - if individuals are in manufacturing
ipums$MANU = NA
ipums[ipums$IND>106 & ipums$IND<400,"MANU"]=1
ipums[is.na(ipums$MANU),"MANU"]=0

#count number of people in manufacturing and college educated
ipums$manu_count = ipums$weight*ipums$MANU
ipums$college_count = ipums$weight*ipums$EDUC

#count manu/educ by czone
ipums_cz_manu = aggregate(ipums$manu_count,by=list(czone = ipums$czone,year=ipums$YEAR),FUN=sum)
colnames(ipums_cz_manu) = paste(c("czone","YEAR","manu_size"))
ipums_cz_educ = aggregate(ipums$college_count,by=list(czone = ipums$czone,year=ipums$YEAR),FUN=sum)
colnames(ipums_cz_educ) = paste(c("czone","YEAR","educ_size"))
ipums_cz_weights = aggregate(ipums$weight,by=list(czone = ipums$czone,year=ipums$YEAR),FUN=sum)
colnames(ipums_cz_weights) = paste(c("czone","YEAR","wa_size"))
#merge with weights
ipums_cz_manu = merge(ipums_cz_manu,ipums_cz_weights,by=c("czone","YEAR"))
ipums_cz_educ = merge(ipums_cz_educ,ipums_cz_weights,by=c("czone","YEAR"))
#get rates
ipums_cz_manu$MANU_R=ipums_cz_manu$manu_size/ipums_cz_manu$wa_size
ipums_cz_educ$EDUC_R=ipums_cz_educ$educ_size/ipums_cz_educ$wa_size

library(dplyr)
#take difference
ipums_diff_manu = ipums_cz_manu %>%
  group_by(czone) %>%
  mutate(dMANU = MANU_R - lag(MANU_R))%>%
  ungroup()

ipums_diff_educ = ipums_cz_educ %>%
  group_by(czone) %>%
  mutate(dEDUC = EDUC_R - lag(EDUC_R))%>%
  ungroup()

#ipums_diff_manu[ipums_diff_manu$czone==10101,]
ipums_diff_manu = ipums_diff_manu[!(is.na(ipums_diff_manu$dMANU)),]
ipums_diff_educ = ipums_diff_educ[!(is.na(ipums_diff_educ$dEDUC)),]


save(ipums_diff_manu,file="MANU.RData")
save(ipums_diff_educ,file="EDUC.RData")

#############
#REGRESSIONS#
#############

load("MANU.RData")
load("EDUC.RData")
```

```
load("LF.RData")
load("unemp.RData")
load("lwage.RData")
load("ipums_regional_shock.RData")

ipums_diff_educ = ipums_diff_educ[,c(1,6)]
ipums_diff_manu = ipums_diff_manu[,c(1,6)]

ipums_reg = merge(ipums_c_LF,ipums_c_lnWage,by="czone")
ipums_reg = merge(ipums_reg,ipums_c_Unemp,by="czone")
ipums_reg = merge(ipums_reg,ipums_diff_educ,by="czone")
ipums_reg = merge(ipums_reg,ipums_diff_manu,by="czone")
ipums_reg = merge(ipums_reg,ipums_regional_shock,by="czone") #shock is x
save(ipums_reg,file="ipums_reg.RData")

attach(ipums_reg)

reg1_wage = lm(dlnWage ~ x,data=ipums_reg)
summary(reg1_wage)
reg2_wage = lm(dlnWage ~ x + dEDUC,data=ipums_reg)
summary(reg2_wage)
reg3_wage = lm(dlnWage ~ x + dEDUC + dMANU,data=ipums_reg)
summary(reg3_wage)

reg1_unemp = lm(Unemp ~ x,data=ipums_reg)
summary(reg1_unemp)
reg2_unemp = lm(Unemp ~ x + dEDUC,data=ipums_reg)
summary(reg2_unemp)
reg3_unemp = lm(Unemp ~ x + dEDUC + dMANU,data=ipums_reg)
summary(reg3_unemp)

reg1_lf = lm(LF ~ x,data=ipums_reg)
summary(reg1_lf)
reg2_lf = lm(LF ~ x + dEDUC,data=ipums_reg)
summary(reg2_lf)
reg3_lf = lm(LF ~ x + dEDUC + dMANU,data=ipums_reg)
summary(reg3_lf)
```