R Project - Replicating ADH Regressions

Ulrich Bergmann Lachlan Deer

Overview

In this exercise we are going to replicate a regression from Autor, Dorn, Hansen (AER, 2003): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States". We want to regress the change in wage/salary (`relchg_avg_hhincwage_pc_pw`) on the percent change in import per worker (`d_tradeusch_pw`). You find the final regression coefficient in Column 2 of Panel A of Table 9 of the paper.

Let's first load the necessary packages to read data and do fancy regressions:

```r
library("readr")
library("tibble")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library("sandwich")
library("lmtest")
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library("lfe")
```

```
## Loading required package: Matrix

##
## Attaching package: 'lfe'

## The following object is masked from 'package:lmtest':
##
##     waldtest
```

And let's load the data like we always do:

```
df = read_csv("data/adh_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   city = col_character()
## )

## See spec(...) for full column specifications.
```

1. OLS regression

The core of the paper is looking at what happened to laborer's when theres an increase in us imports from china. Let's try and replicate part of Table 9 - namely the estimate from panel A column 2.

Their y variable is `relchg_avg_hhincwage_pc_pw`. The important x variable is decadal trade between the us and china `d_tradeusch_pw`.

1. Run that simple regression

```
lm_1 = lm(relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw, data = df)
summary(lm_1)
```

```
##
## Call:
## lm(formula = relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.789  -8.411  -0.663   7.715  49.684
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.0720     0.3889   41.33   <2e-16 ***
## d_tradeusch_pw  -1.6466     0.1212  -13.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.89 on 1442 degrees of freedom
## Multiple R-squared:  0.1135, Adjusted R-squared:  0.1129
## F-statistic: 184.6 on 1 and 1442 DF,  p-value: < 2.2e-16
```

2. Now add heteroskedasticity robust standard (HC1). Hint: Use the `sandwich` and `lmtest` packages

```
coeftest(lm_1, vcov = vcovHC(lm_1, type="HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    16.07198    0.57211 28.0923 < 2.2e-16 ***
## d_tradeusch_pw -1.64663    0.28496 -5.7785 9.219e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we will start to add extra x variables.

3. Start by adding `t2` - a dummy variable for whether observation is in the second decade. Fit again with HC1 robust standard errors.

```
lm_2 = lm(relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw + t2, data = df)
coeftest(lm_2, vcov = vcovHC(lm_2, type="HC1"))
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)     21.59096    0.39889  54.1271 < 2.2e-16 ***
## d_tradeusch_pw  -0.88316    0.16804  -5.2555 1.698e-07 ***
## t2             -13.94769    0.57869 -24.1023 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Clustering

Let us now use clustertered standard errors instead. ADH cluster by `statefip`. Hint: use the `felm` command from the `lfe` package

1. Run the basic regression with clustering

```
felm_1 = felm(relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw + t2 | 0 | 0 | statefip, data = df)
summary(felm_1)
```

```
##
## Call:
##    felm(formula = relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw +      t2 | 0 | 0 | statefip, data = d
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.707  -6.259  -0.799   5.024  43.128
##
## Coefficients:
##                  Estimate Cluster s.e. t value Pr(>|t|)
## (Intercept)     21.5910       0.9841  21.941  < 2e-16 ***
## d_tradeusch_pw  -0.8832       0.2544  -3.472 0.000532 ***
## t2             -13.9477       1.8146  -7.686 2.79e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.831 on 1441 degrees of freedom
## Multiple R-squared(full model): 0.3944   Adjusted R-squared: 0.3936
## Multiple R-squared(proj model): 0.3944   Adjusted R-squared: 0.3936
## F-statistic(full model, *iid*):469.3 on 2 and 1441 DF, p-value: < 2.2e-16
## F-statistic(proj model):  68.5 on 2 and 47 DF, p-value: 1.179e-14
```

2. Add the following controls to your last regression:
    - `l_shind_manuf_cbp`
    - `l_sh_popedu_c`
    - `l_sh_popfborn`
    - `l_sh_empl_f`
    - `l_sh_routine33`
    - `l_task_outsource`

```
felm_2 = felm(relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource
                | 0 | 0 | statefip, data = df)
summary(felm_2)
```

```
##
## Call:
##    felm(formula = relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw +      t2 + l_shind_manuf_cbp + l_sh_p
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.329  -5.869  -0.534   5.229  40.819
##
## Coefficients:
##                    Estimate Cluster s.e. t value Pr(>|t|)
## (Intercept)        43.38967      8.70748   4.983 7.02e-07 ***
## d_tradeusch_pw     -0.37033      0.13283  -2.788  0.00537 **
## t2                -13.79402      1.76456  -7.817 1.04e-14 ***
## l_shind_manuf_cbp  -0.18211      0.03835  -4.749 2.25e-06 ***
## l_sh_popedu_c      -0.13418      0.05494  -2.442  0.01471 *
## l_sh_popfborn      -0.13180      0.08678  -1.519  0.12903
## l_sh_empl_f         0.31270      0.07987   3.915 9.46e-05 ***
## l_sh_routine33     -1.04448      0.23945  -4.362 1.38e-05 ***
## l_task_outsource    4.21857      1.72489   2.446  0.01458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.218 on 1435 degrees of freedom
## Multiple R-squared(full model): 0.4699   Adjusted R-squared: 0.4669
## Multiple R-squared(proj model): 0.4699   Adjusted R-squared: 0.4669
## F-statistic(full model, *iid*):  159 on 8 and 1435 DF, p-value: < 2.2e-16
## F-statistic(proj model): 28.75 on 8 and 47 DF, p-value: 1.276e-15
```

3. Add region fixed effects to your regression.
   - First find all variables in the dataset that start with `reg_`
   - Add these to your last regression

```
names(select(df, starts_with("reg_")))
```

```
## [1] "reg_midatl" "reg_encen"  "reg_wncen"  "reg_satl"   "reg_escen"
## [6] "reg_wscen"  "reg_mount"  "reg_pacif"
```

```
felm_3 = felm(relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource +
                reg_midatl + reg_encen + reg_wncen + reg_satl +
                reg_escen + reg_wscen + reg_mount + reg_pacif
              | 0 | 0 | statefip, data = df)
summary(felm_3)
```

```
##
## Call:
##    felm(formula = relchg_avg_hhincwage_pc_pw ~ d_tradeusch_pw +      t2 + l_shind_manuf_cbp + l_sh_p
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -36.256  -5.408  -0.643   4.993  40.596 
## 
## Coefficients:
##                    Estimate Cluster s.e. t value Pr(>|t|)    
## (Intercept)        37.81934     10.64339   3.553 0.000393 ***
## d_tradeusch_pw     -0.41327      0.12965  -3.188 0.001465 ** 
## t2                -13.67579      1.73663  -7.875 6.70e-15 ***
## l_shind_manuf_cbp  -0.15404      0.03266  -4.716 2.64e-06 ***
## l_sh_popedu_c      -0.09418      0.05679  -1.658 0.097449 .  
## l_sh_popfborn      -0.09372      0.07883  -1.189 0.234674    
## l_sh_empl_f         0.18554      0.09601   1.933 0.053492 .  
## l_sh_routine33     -0.79294      0.26449  -2.998 0.002764 ** 
## l_task_outsource    4.26729      1.87809   2.272 0.023226 *  
## reg_midatl          2.06339      1.97681   1.044 0.296757    
## reg_encen           2.10566      2.13354   0.987 0.323844    
## reg_wncen           6.91844      2.14751   3.222 0.001303 ** 
## reg_satl            0.87864      1.89366   0.464 0.642723    
## reg_escen           4.30400      2.16206   1.991 0.046705 *  
## reg_wscen           5.05509      2.07366   2.438 0.014900 *  
## reg_mount           4.48765      2.04204   2.198 0.028136 *  
## reg_pacif          -0.17986      1.91287  -0.094 0.925100    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.997 on 1427 degrees of freedom
## Multiple R-squared(full model): 0.4977   Adjusted R-squared: 0.4921 
## Multiple R-squared(proj model): 0.4977   Adjusted R-squared: 0.4921 
## F-statistic(full model, *iid*):88.38 on 16 and 1427 DF, p-value: < 2.2e-16 
## F-statistic(proj model): 21.68 on 16 and 47 DF, p-value: < 2.2e-16
```

3. Instrument Variables

1. Instrument d_tradeusch_pw with d_tradeotch_pw_lag in your last regression

```
felm_4 = felm(relchg_avg_hhincwage_pc_pw ~ 1 + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource +
                reg_midatl + reg_encen + reg_wncen + reg_satl +
                reg_escen + reg_wscen + reg_mount + reg_pacif
              | 0 | (d_tradeusch_pw ~ d_tradeotch_pw_lag) | statefip,
              data = df)
summary(felm_4)
```

```
## 
## Call:
##    felm(formula = relchg_avg_hhincwage_pc_pw ~ 1 + t2 + l_shind_manuf_cbp +      l_sh_popedu_c + l_sh
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -36.066  -5.524  -0.555   4.996  42.042 
## 
## Coefficients:
##                        Estimate Cluster s.e. t value Pr(>|t|)
## (Intercept)            35.87405     10.64255   3.371 0.000769 ***
## t2                    -12.28244      1.97304  -6.225 6.31e-10 ***
## l_shind_manuf_cbp      -0.08113      0.04139  -1.960 0.050177 .
## l_sh_popedu_c          -0.08755      0.05778  -1.515 0.129955
## l_sh_popfborn          -0.08900      0.08150  -1.092 0.275021
## l_sh_empl_f             0.18856      0.09896   1.905 0.056935 .
## l_sh_routine33         -0.75086      0.26165  -2.870 0.004168 **
## l_task_outsource        4.20832      1.86424   2.257 0.024135 *
## reg_midatl              1.97306      1.93354   1.020 0.307693
## reg_encen               1.62608      2.22012   0.732 0.464027
## reg_wncen               6.46027      2.19861   2.938 0.003353 **
## reg_satl                0.41119      1.94593   0.211 0.832677
## reg_escen               5.02788      2.28463   2.201 0.027914 *
## reg_wscen               4.62044      2.12783   2.171 0.030063 *
## reg_mount               3.89093      2.09533   1.857 0.063523 .
## reg_pacif              -0.94648      2.00055  -0.473 0.636206
## `d_tradeusch_pw(fit)`  -1.27575      0.43200  -2.953 0.003197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.186 on 1427 degrees of freedom
## Multiple R-squared(full model): 0.4765   Adjusted R-squared: 0.4706
## Multiple R-squared(proj model): 0.4765   Adjusted R-squared: 0.4706
## F-statistic(full model, *iid*):86.45 on 16 and 1427 DF, p-value: < 2.2e-16
## F-statistic(proj model): 18.74 on 16 and 47 DF, p-value: 3.486e-15
## F-statistic(endog. vars):8.721 on 1 and 47 DF, p-value: 0.004898
```

2. Weight your regression by `timepwt48`

The `felm` function is a bit picky on the order of the weights. Let us first try to define weights at the end after the `data` argument like so:

```
felm_5 = felm(relchg_avg_hhincwage_pc_pw ~ 1 + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource +
                reg_midatl + reg_encen + reg_wncen + reg_satl +
                reg_escen + reg_wscen + reg_mount + reg_pacif
              | 0 | (d_tradeusch_pw ~ d_tradeotch_pw_lag) | statefip,
              data = df,
              weights = timepwt48)
```

```
## Error in eval(mf[[wpos]], pf): object 'timepwt48' not found
```

```
summary(felm_5)
```

```
## Error in summary(felm_5): object 'felm_5' not found
```

Felm didn't find `timepwt48` because it only assumes that columns are in `df` before you define `data = df`. We can solve this in two ways.

1. A good rule is to have `data = df` as the last argument.

```
felm_5 = felm(relchg_avg_hhincwage_pc_pw ~ 1 + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource +
                reg_midatl + reg_encen + reg_wncen + reg_satl +
                reg_escen + reg_wscen + reg_mount + reg_pacif
              | 0 | (d_tradeusch_pw ~ d_tradeotch_pw_lag) | statefip,
              weights = timepwt48,
              data = df)
```

```
## Error in eval(mf[[wpos]], pf): object 'timepwt48' not found
```

```
summary(felm_5)
```

```
## Error in summary(felm_5): object 'felm_5' not found
```

2. Alternatively, you can define weights after `data = df`, but then you have to define the weights as `df$timepwft48` like so:

```
felm_5 = felm(relchg_avg_hhincwage_pc_pw ~ 1 + t2 +
                l_shind_manuf_cbp + l_sh_popedu_c + l_sh_popfborn +
                l_sh_empl_f + l_sh_routine33 + l_task_outsource +
                reg_midatl + reg_encen + reg_wncen + reg_satl +
                reg_escen + reg_wscen + reg_mount + reg_pacif
              | 0 | (d_tradeusch_pw ~ d_tradeotch_pw_lag) | statefip,
              data = df,
              weights = df$timepwt48)
summary(felm_5)
```

```
##
## Call:
##    felm(formula = relchg_avg_hhincwage_pc_pw ~ 1 + t2 + l_shind_manuf_cbp +      l_sh_popedu_c + l_sh
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2404 -0.1084 -0.0033  0.1114  2.8633
##
## Coefficients:
##                         Estimate Cluster s.e. t value Pr(>|t|)
## (Intercept)             60.67937      9.14264   6.637 4.54e-11 ***
## t2                      -9.05462      2.65665  -3.408 0.000672 ***
## l_shind_manuf_cbp        0.06791      0.08505   0.798 0.424746
## l_sh_popedu_c            0.10292      0.11146   0.923 0.355954
## l_sh_popfborn            0.07652      0.08050   0.950 0.342020
## l_sh_empl_f             -0.21015      0.16952  -1.240 0.215297
## l_sh_routine33          -1.01014      0.22888  -4.413 1.09e-05 ***
## l_task_outsource         5.56661      1.48578   3.747 0.000186 ***
## reg_midatl              -0.56489      1.55722  -0.363 0.716840
## reg_encen               -2.61723      1.97713  -1.324 0.185797
## reg_wncen                1.93904      1.63154   1.188 0.234846
## reg_satl                -2.73867      1.49871  -1.827 0.067855 .
## reg_escen                0.60288      1.53128   0.394 0.693856
## reg_wscen               -1.68621      1.73888  -0.970 0.332354
## reg_mount               -2.36081      1.40184  -1.684 0.092385 .
## reg_pacif               -6.27918      2.02160  -3.106 0.001933 **
## `d_tradeusch_pw(fit)` -2.14156      0.59462  -3.602 0.000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2791 on 1427 degrees of freedom
## Multiple R-squared(full model): 0.4278   Adjusted R-squared: 0.4214
## Multiple R-squared(proj model): 0.4278   Adjusted R-squared: 0.4214
## F-statistic(full model, *iid*):77.51 on 16 and 1427 DF, p-value: < 2.2e-16
## F-statistic(proj model): 44.53 on 16 and 47 DF, p-value: < 2.2e-16
## F-statistic(endog. vars):12.97 on 1 and 47 DF, p-value: 0.0007602
```

And now we have the numbers reported in Column 2 of Panel A of Table 9 of the paper.