Data Science for Economists

Lecture 8: Regression analysis in R

Grant R. McDermott

University of Oregon | EC 607

Contents

Today's lecture explores

Software requirements

R packages

It's important to note that "base" R already provides all of the tools to implement a fixed effects regression, **but** you'll quickly hit walls due to memory caps. Instead, I want to introduce **fixest**, short for Fixed-Effects Estimation, which provides lightning fast fixed effects estimation and make your life much easier.

- New: **fixest**, **wooldridge**
- Already used: **tidyverse**, **hrbrthemes**, **listviewer**, **estimatr**, **ivreg**, **sandwich**, **lmtest**, **mfx**, **margins**, **broom**, **modelsummary**, **vtable**, **rstanarm**

A convenient way to install (if necessary) and load everything is by running the below code chunk.

```
## Load and install the packages that we'll be using today
if (!require("pacman")) install.packages("pacman")
pacman::p_load(mfx, tidyverse, hrbrthemes, estimatr, ivreg, fixest, sandwich, wooldridge,
               lmtest, margins, vtable, broom, modelsummary,gsheet)

## My preferred ggplot2 plotting theme (optional)
theme_set(theme_minimal())
```

Note on fixest and feols I'll be using fixest and feols throughout these notes. The fixest package is a new package that is very fast and has a lot of functionality. It has several bits of funtionality like `feols()` and `etable()`, which are powerful functions for making regressions and putting the output into tables that work well together. `feols()` works very much like `lm()` in base R, but with a few added bonuses.

Review of last lecture

Last lecture we covered how fixed effects are extremely useful for removing variation between units. That means any of the average differences between groups of the fixed effect are removed. We can then look at underlying variation within these groups to see if there is a relationship between our variables of interest.

This is extremely useful for dealing with omitted variable bias. If we have an omitted variable that is correlated with our independent variable, we can't tell if the relationship we see is due to the independent variable or the omitted variable. But if we have a fixed effect for the omitted variable, we can remove the variation between units and then look at the variation within units.

In practice, fixed effects amount to de-meaning our variables of interest. There are a handful of ways to do this.

Panel models

A panel dataset is one in which we view a single unit over multiple periods of time, so a balanced panel has the same number of observations for each unit. For example, we might have data on 100 countries over 10 years, or 50 US states over 20 years. We can then take unit fixed effects, which lets us compare between years within a single unit. Similarly, we can take time fixed effects to compare between units within a given point in time. If our dataset has other dimensions that vary in a way that is not collinear with unit or time, we can also take a fixed effect for that – though again, you want to be careful about throwing in fixed effects.

Dataset

Let me introduce the dataset we'll be using, `crime4`. It comes from Jeffrey Wooldridge's R package – Dr. Wooldridge is one of the most accomplished professors of econometrics on the planet. I was tipped off about his package by Nick Huntington-Klein's own lecture notes.. The dataset shows county probability of arrest and county crime rate by year.

```
data(crime4)
crime4 %>%
  select(county, year, crmrte, prbarr) %>%
  rename(County = county,
         Year = year,
         CrimeRate = crmrte,
         ProbofArrest = prbarr) %>%
  slice(1:9) %>%
  knitr::kable(note = '...') %>%
  kableExtra::add_footnote('9 rows out of 630. "Prob. of Arrest" is estimated probability of being arre
```

County

Year

CrimeRate

ProbofArrest

1

81

0.0398849

0.289696

1

82

0.0383449

0.338111

1

83

0.0303048

0.330449

1

84

0.0347259

0.362525

1

85

0.0365730

0.325395

1

86

0.0347524

0.326062

1

87

0.0356036

0.298270

3

81

0.0163921

0.202899

3

82

0.0190651

0.162218

3

83

0.0151492

0.181586

3

84

0.0136621

0.194986

3

85

0.0120346

0.206897

3

86

0.0129982

0.156069

3

87

0.0152532

0.132029

7

81

0.0219159

0.431095

7

83

0.0242110

0.419405

7

84

0.0223434

0.412458

7

85

0.0245848

0.380655

7

86

0.0241281

0.308057

7

87

0.0267532

0.364760

23

81

0.0319175

0.194303

23

82

0.0290211

0.286639

23

83

0.0286164

0.280522

23

84

0.0275500

0.334615

23

85

0.0293095

0.287442

23

86

0.0256248

0.304577

23

87

0.0269836

0.289121

9 rows out of 630. "Prob. of Arrest" is estimated probability of being arrested when you commit a crime

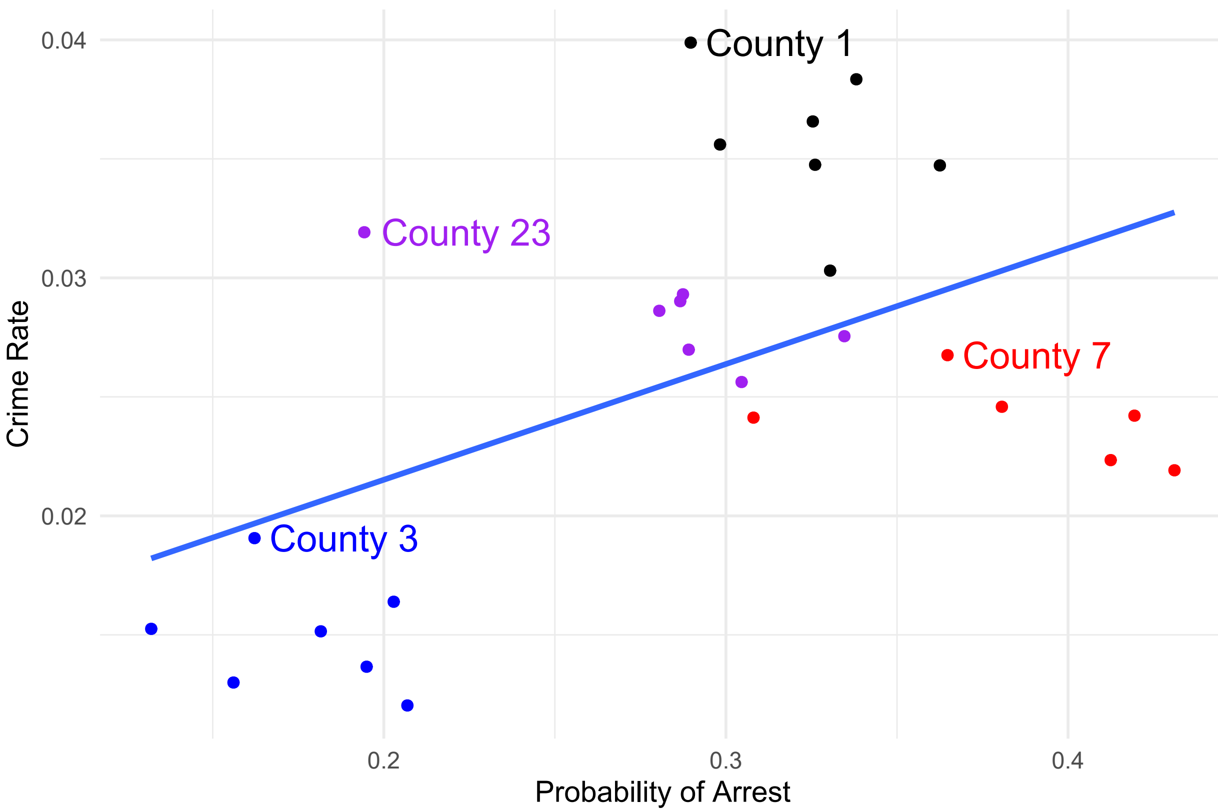
Let's visualize it

Below I visualize the data for just a few counties. Note the positive slope when pooling! Is that surprising?

```
crime4 %>%
  filter(county %in% c(1,3,7, 23),
         prbarr < .5) %>%
  group_by(county) %>%
  mutate(label = case_when(
    crmrte == max(crmrte) ~ paste('County',county),
    TRUE ~ NA_character_
  )) %>%
  ggplot(aes(x =  prbarr, y = crmrte, color = factor(county), label = label)) +
  geom_point() +
  geom_text(hjust = -.1, size = 14/.pt) +
```

```
  labs(x = 'Probability of Arrest',
       y = 'Crime Rate',
       caption = 'One outlier eliminated in County 7.') +
  #scale_x_continuous(limits = c(.15, 2.5)) +
  guides(color = FALSE, label = FALSE) +
  scale_color_manual(values = c('black','blue','red','purple')) +
  geom_smooth(method = 'lm', aes(color = NULL, label = NULL), se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



One outlier eliminated in County 7.

Let's try the de-meaning approach

We can use group_by to get means-within-groups and subtract them out.

```
crime4 <- crime4 %>%
  # Filter to the data points from our graph
  filter(county %in% c(1,3,7, 23),
         prbarr < .5) %>%
  group_by(county) %>%
  mutate(mean_crime = mean(crmrte),
         mean_prob = mean(prbarr)) %>%
  mutate(demeaned_crime = crmrte - mean_crime,
         demeaned_prbarr = prbarr - mean_prob)
```

	(1)	(2)
(Intercept)	0.012	0.000
	(0.005)	(0.000)
prbarr	0.049	
	(0.017)	
demeaned_prbarr		−0.030
		(0.012)
Num.Obs.	27	27
R2	0.253	0.214
R2 Adj.	0.223	0.183
AIC	−186.2	−254.8
BIC	−182.3	−250.9
Log.Lik.	96.098	130.399
F	8.471	
RMSE	0.01	0.00

And Regress!

```
orig_data <- lm(crmrte ~ prbarr, data = crime4)
de_mean <- lm(demeaned_crime ~ demeaned_prbarr, data = crime4)
msummary(list(orig_data, de_mean))
```

Note the coefficient has flipped!

Interpreting a Within Relationship

How can we interpret that slope of `-0.03`? This is all *within variation* so our interpretation must be *within-county*. So, "comparing a county in year A where its arrest probability is 1 (100 percentage points) higher than it is in year B, we expect the number of crimes per person to drop by .03." Or if we think we've causally identified it (and want to work on a more realistic scale), "raising the arrest probability by 1 percentage point in a county reduces the number of crimes per person in that county by .0003". We're basically "controlling for county" (and will do that explicitly in a moment). So your interpretation should think of it in that way - *holding county constant* i.e. *comparing two observations with the same value of county* i.e. *comparing a county to itself at a different point in time*.

Concept checks

- Do you think the model we've presented is sufficient to have a causal interpretation of the effect of arrest probability on crime?
- What assumptions would we need to make to have a causal interpretation?
- What potential confounders are there?
- Why does subtracting the within-individual mean of each variable "control for individual"?
- In a sentence, interpret the slope coefficient in the estimated model $(Y_{it} - \bar{Y}_i) = 2 + 3(X_{it} - \bar{X}_i)$ where Y is "blood pressure", X is "stress at work", and i is an individual person, and $\bar{Y}_i$ means average of Y_i
- Is this relationship causal? If not, what assumptions are required for it to be causal?

Can we do that all at once? Yes, with the Least Squares Dummy Variable Approach

De-meaning takes some steps which could get tedious to write out. Another way is to include a dummy or category variable for each county. This is called the Least Squares Dummy Variable approach.

You end up with the same results as if we de-meaned.

```
lsdv <- lm(crmrte ~ prbarr + factor(county), data = crime4)
msummary(list(orig_data, de_mean, lsdv), keep = c('prbarr', 'demeaned_prob'))
```

	(1)	(2)	(3)
prbarr	0.049		−0.030
	(0.017)		(0.012)
demeaned_prbarr		−0.030	
		(0.012)	
Num.Obs.	27	27	27
R2	0.253	0.214	0.941
R2 Adj.	0.223	0.183	0.930
AIC	−186.2	−254.8	−248.8
BIC	−182.3	−250.9	−241.0
Log.Lik.	96.098	130.399	130.399
F	8.471		87.946
RMSE	0.01	0.00	0.00

Hey look, the coefficient is the same!

Why LSDV?

- A benefit of the LSDV approach is that it calculates the fixed effects α_i for you
- We left those out of the table with the `coefs` argument of `export_summs` (we rarely want them) but here they are:

```
lsdv
```

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 27
## Standard-errors: IID
##                    Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)        0.045631   0.004116  11.08640 1.7906e-10 ***
## prbarr            -0.030491   0.012442  -2.45068 2.2674e-02 *
## factor(county)3   -0.025308   0.002165 -11.68996 6.5614e-11 ***
## factor(county)7   -0.009870   0.001418  -6.96313 5.4542e-07 ***
## factor(county)23 -0.008587   0.001258  -6.82651 7.3887e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.001933   Adj. R2: 0.930441
```

THe interpretation is exactly the same as with a categorical variable - we have an omitted county, and these show the difference relative to that omitted county

NOTE: See how I put `factor()` around county? That is to ensure it reads county, which is the county fips code as a categorical variable instead of as a numerical variable. If you don't do that, it will read it as a numerical variable and you'll get a different result:

```
lm(crmrte ~ prbarr + county, data = crime4)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr + county, data = crime4)
##
## Coefficients:
## (Intercept)       prbarr       county
##   1.134e-02    4.829e-02    6.444e-05
```

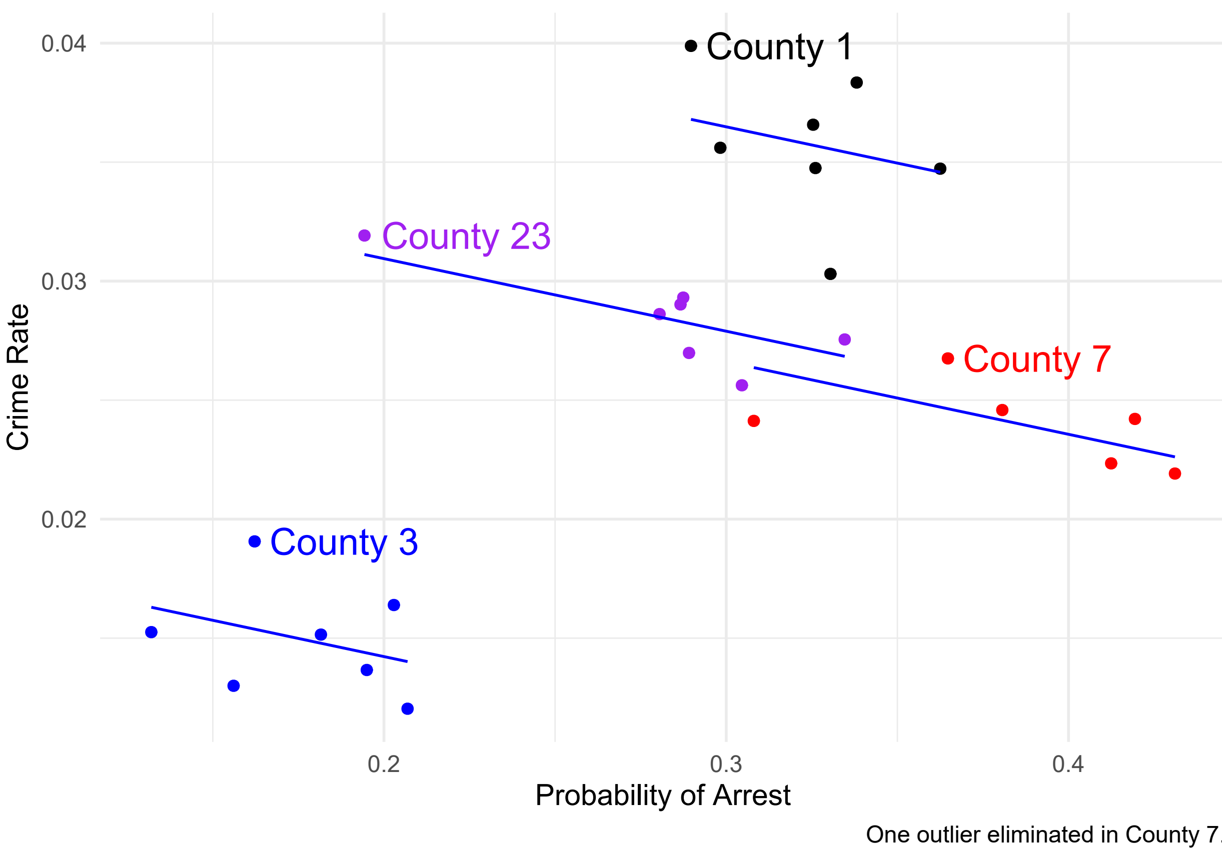
This is saying that as FIPS code increases by one, the crime rate increases by 0.000011… that's nonsense. There's an urban legend of an economist who took the log of the NAICS industry classification code for quite some time before realizing they meant to use a categorical variable. Correcting that mistake completely changed their results.

Why LSDV?

This also makes clear another element of what's happening! Just like with a categorical var, the line is moving *up and down* to meet the counties. Graphically, de-meaning moves all the points together in the middle to draw a line, while LSDV moves the line up and down to meet the points

```
crime4 %>%
  ungroup() %>%
  mutate(pred = predict(lsdv)) %>%
  group_by(county) %>%
  mutate(label = case_when(
    crmrte == max(crmrte) ~ paste('County',county),
    TRUE ~ NA_character_
  )) %>%
  ggplot(aes(x =  prbarr, y = crmrte, color = factor(county), label = label)) +
  geom_point() +
  geom_text(hjust = -.1, size = 14/.pt) +
  geom_line(aes(y = pred, group = county), color = 'blue') +
  labs(x = 'Probability of Arrest',
       y = 'Crime Rate',
       caption = 'One outlier eliminated in County 7.') +
  #scale_x_continuous(limits = c(.15, 2.5)) +
  guides(color = FALSE, label = FALSE) +
  scale_color_manual(values = c('black','blue','red','purple'))
```

```
## Warning: Removed 23 rows containing missing values (`geom_text()`).
```



The "Pros" don't use LSDV

Most people do not use LSDB – it is computationally expensive. If you get too many fixed effects or too big of data, it just will not wrong. The professionally-written commands use de-meaning, like **fixest**, which is less computationally expensive. See for yourself! Look, we even used the **etable** function.

```
pro <- feols(crmrte ~ prbarr | county, data = crime4)
de_mean <- feols(demeaned_crime ~ demeaned_prbarr, data = crime4)
etable(de_mean, pro)
```

```
##                            de_mean               pro
## Dependent Var.:    demeaned_crime            crmrte
##
## Constant        1.41e-18 (0.0004)
## demeaned_prbarr -0.0305* (0.0117)
## prbarr                            -0.0305* (0.0064)
## Fixed-Effects:  ----------------- -----------------
## county                         No               Yes
## _______________ _________________ _________________
## S.E. type                     IID       by: county
## Observations                   27                27
## R2                        0.21445           0.94114
## Within R2                      --           0.21445
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To explain the **fixest** package, let's use a familiar dataset.

Fixed effects using a familiar dataset

Let's review fixed effects very quickly using the Ask A Manager Survey 2023, which you used on problem set 2. Like on your problem set, we'll load it in using the **gsheet** package. Several of you said that you thought you could use fixed effects to residualize other differences between groups out of the data. Let's see if that's true using the package **fixest**. *As a disclaimer: these data are not systematically collected.*

```
column_names <- c('timestamp','age','industry','area','jobtitle','jobtitle2',
        'annual_salary','additional_pay','currency','currency_other',
        'income_additional','country','state','city','remote','experience_overall',
        'experience_field','education','gender','race')

US_strings<-c("United States of America","United States",
  "United states" ,"USA","Usa","usa" ,"US","U.S." ,"us")

# gsheet2text is a function that takes a google sheet and turns it into a text file that read_csv can u
managers2023 = read_csv(gsheet::gsheet2text('https://docs.google.com/spreadsheets/d/  1ioUjhnz6ywSpEbAR
  col_names = column_names) %>%
    mutate(year = 2023,
    across(c(additional_pay, annual_salary),as.numeric)) %>%
  filter(country %in% US_strings,
    gender %in% c('Man','Woman'),
    !is.na(state)) %>%
  separate(race,into=c('race1','race2','race3','race4'), # Separates the race variable into 4 columns
    sep=',')
```

```
## No encoding supplied: defaulting to UTF-8.
```

```
## Rows: 17036 Columns: 20
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (20): timestamp, age, industry, area, jobtitle, jobtitle2, annual_salary...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
managers2022 = read_csv(gsheet::gsheet2text('https://docs.google.com/spreadsheets/d/1ioUjhnz6ywSpEbARI-
  col_names = column_names) %>%
    mutate(year = 2022,
    across(c(additional_pay, annual_salary),as.numeric)) %>%
  filter(country %in% US_strings,
    gender %in% c('Man','Woman'),
    !is.na(state)) %>%
  separate(race,into=c('race1','race2','race3','race4'), # Separates the race variable into 4 columns
    sep=',')
```

```
## No encoding supplied: defaulting to UTF-8.
## Rows: 17036 Columns: 20-- Column specification ----------------------------------------------------
## Delimiter: ","
## chr (20): timestamp, age, industry, area, jobtitle, jobtitle2, annual_salary...
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Focus on the US and cismen and women to simplify the analysis
managers <- bind_rows(managers2023,managers2022)
```

We won't go into everything you can possibly do with this dataset until later in the course, but a few of you suggested it may be helpful to residualize of the fixed effects for race or age when looking at the gender pay gap in these data.

```
gender_pay_gap <- feols(annual_salary ~ gender,data=managers2023)
gender_pay_gap_race_fe <- feols(annual_salary ~ gender | race1,data=managers)
```

```
## NOTE: 48 observations removed because of NA values (Fixed-effects: 48).
```

```
gender_pay_gap_age_fe <- feols(annual_salary ~ gender | age,data=managers)
gender_pay_gap_age_race_fe <- feols(annual_salary ~ gender | age+race1,data=managers)
```

```
## NOTE: 48 observations removed because of NA values (Fixed-effects: 48).
```

```
etable(list('Base'=gender_pay_gap,
  'Age FE'=gender_pay_gap_age_fe,
  'Race FE'=gender_pay_gap_race_fe,
  'Age and Race FE'=gender_pay_gap_age_race_fe))
```

```
##                                     Base                 Age FE
## Dependent Var.:            annual_salary          annual_salary
##
## Constant        121,813.5*** (1,532.4)
## genderWoman     -23,259.8*** (1,693.9) -23,430.9*** (3,319.7)
## Fixed-Effects:  ---------------------- ----------------------
## age                                 No                    Yes
## race1                               No                     No
## _______________ ______________________ ______________________
## S.E. type                          IID                by: age
## Observations                    13,212                 26,424
## R2                             0.01407                0.03466
## Within R2                           --                0.01457
```

```
##
##                                 Race FE        Age and Race FE
## Dependent Var.:          annual_salary          annual_salary
##
## Constant
## genderWoman     -23,328.4*** (1,368.7) -23,521.3*** (3,275.9)
## Fixed-Effects:  ---------------------- ----------------------
## age                                 No                    Yes
## race1                              Yes                    Yes
## _______________ ______________________ ______________________
## S.E. type                    by: race1                by: age
## Observations                    26,376                 26,376
## R2                             0.01859                0.04104
## Within R2                      0.01419                0.01475
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pretty neat right? Just sticking something after the `|` allows you to residualize its fixed effect! But won't it get tedious writing out all these variations of fixed effects? Sure will. That's where the **fixest** package comes in handy.

What if we wanted to change the clustering of the standard errors? Did you notice the S.E. type above? It auto-clustered by the fixed effects – specifically the fixed effect with the most levels (so age when choosing between age and race).

Sometimes you want to cluster standard errors a new way. Well that is something you can do with **fixest** and its delightfully well-designed `etable()` function. You can specify the cluster variable with `cluster()` or the type of standard errors you want with `se()` and get different types of standard errors. Below I specify standard errors clustered by state and then an assumption of independent and identically distributed errors. (The most vanilla standard errors you can assume and rarely the ones we believe explain real world phenomena.)

```
etable(list('Base'=gender_pay_gap,
  'Age FE'=gender_pay_gap_age_fe,
  'Race FE'=gender_pay_gap_race_fe,
  'Age and Race FE'=gender_pay_gap_age_race_fe),
  se='IID')
```

```
##                                   Base                 Age FE
## Dependent Var.:          annual_salary          annual_salary
##
## Constant        121,813.5*** (1,532.4)
## genderWoman     -23,259.8*** (1,693.9) -23,430.9*** (1,185.5)
## Fixed-Effects:  ---------------------- ----------------------
## age                                 No                    Yes
## race1                               No                     No
## _______________ ______________________ ______________________
## S.E. type                          IID                    IID
## Observations                    13,212                 26,424
## R2                             0.01407                0.03466
## Within R2                           --                0.01457
##
##                                 Race FE        Age and Race FE
## Dependent Var.:          annual_salary          annual_salary
##
## Constant
## genderWoman     -23,328.4*** (1,197.5) -23,521.3*** (1,184.1)
## Fixed-Effects:  ---------------------- ----------------------
```

```
## age                                   No                     Yes
## race1                                Yes                     Yes
## _______________ ______________________ ______________________
## S.E. type                            IID                     IID
## Observations                      26,376                  26,376
## R2                               0.01859                 0.04104
## Within R2                        0.01419                 0.01475
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etable(list('Base'=gender_pay_gap,
  'Age FE'=gender_pay_gap_age_fe,
  'Race FE'=gender_pay_gap_race_fe,
  'Age and Race FE'=gender_pay_gap_age_race_fe),
  cluster='state')
```

```
##                                  Base                 Age FE
## Dependent Var.:         annual_salary          annual_salary
##
## Constant        121,813.5*** (5,479.3)
## genderWoman     -23,259.8*** (3,183.0) -23,430.9*** (3,134.0)
## Fixed-Effects:  ---------------------- ----------------------
## age                                 No                    Yes
## race1                               No                     No
## _______________ ______________________ ______________________
## S.E.: Clustered              by: state              by: state
## Observations                    13,212                 26,424
## R2                             0.01407                0.03466
## Within R2                           --                0.01457
##
##                               Race FE        Age and Race FE
## Dependent Var.:         annual_salary          annual_salary
##
## Constant
## genderWoman     -23,328.4*** (3,227.7) -23,521.3*** (3,179.4)
## Fixed-Effects:  ---------------------- ----------------------
## age                                 No                    Yes
## race1                              Yes                    Yes
## _______________ ______________________ ______________________
## S.E.: Clustered              by: state              by: state
## Observations                    26,376                 26,376
## R2                             0.01859                0.04104
## Within R2                      0.01419                0.01475
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We'd normally expect our standard errors to blow up with clustering and we see something similar here. Why is that?

Yes, I know this is a lot on stuff you've only barely experienced before. But you're going to come across these terms when you read papers and I want you to know how to play with them when you're trying to learn by doing.

Aside on standard errors We've now seen the various options that **fixest** has for specifying different standard error structures. In short, you invoke either of the `se` or `cluster` arguments. Moreover, you can choose to do so either at estimation time, or by adjusting the standard errors for an existing model post-estimation (e.g. with `summary.fixest(mod, cluster = ...)`). There are two additional points that I want to draw your attention to.

First, if you're coming from another statistical language, adjusting the standard errors post-estimation (rather than always at estimation time) may seem slightly odd. But this behaviour is actually extremely powerful, because it allows us to analyse the effect of different error structures *on-the-fly* without having to rerun the entire model again. **fixest** is already the fastest game in town, but just think about the implied time savings for really large models.[1] I'm a huge fan of the flexibility, safety, and speed that on-the-fly standard error adjustment offers us. I even wrote a whole blog post about it if you'd like to read more.

Second, reconciling standard errors across different software is a much more complicated process than you may realise. There are a number of unresolved theoretical issues to consider — especially when it comes to multiway clustering — and package maintainers have to make a number of arbitrary decisions about the best way to account for these. See here for a detailed discussion. Luckily, Laurent (the **fixest** package author) has taken the time to write out a detailed vignette about how to replicate standard errors from other methods and software packages.[2]

Multiple estimations **fixest** allows you to do multiple estimations in one command and it does is it fast! Why is it so fast? It leverages the de-meaning trick mentioned above. If a fixed effect is used in multiple estimations, it saves the outcome variable de-meaned of that fixed effect to use in all the other estimations. That saves a bunch of time!

This is also a really smart big data technique we'll get into more later in the course. It does a task once instead of multiple times to save time and processing power.

Here's a demo using the stepwise `sw0()` function, which adds fixed effects – starting with none step-by-step:

```
gender_pay_gap_many_fes <- feols(annual_salary ~ gender |
  sw0(age,race1,age+race1),
  data=managers2023)
etable(gender_pay_gap_many_fes)
```

```
##                   gender_pay_gap_many..1 gender_pay_gap_many..2
## Dependent Var.:            annual_salary          annual_salary
##
## Constant          121,813.5*** (1,532.4)
## genderWoman       -23,259.8*** (1,693.9) -23,430.9*** (3,319.9)
## Fixed-Effects:    ---------------------- ----------------------
## age                                   No                    Yes
## race1                                 No                     No
## _______________   ______________________ ______________________
## S.E. type                            IID                by: age
## Observations                      13,212                 13,212
## R2                               0.01407                0.03466
## Within R2                             --                0.01457
##
##                   gender_pay_gap_many..3 gender_pay_gap_many..4
## Dependent Var.:            annual_salary          annual_salary
##
## Constant
## genderWoman       -23,328.4*** (1,368.8) -23,521.3*** (3,276.5)
## Fixed-Effects:    ---------------------- ----------------------
## age                                   No                    Yes
## race1                                Yes                    Yes
## _______________   ______________________ ______________________
## S.E. type                      by: race1                by: age
## Observations                      13,188                 13,188
## R2                               0.01859                0.04104
```

[1] To be clear, adjusting the standard errors via, say, `summary.fixest()` completes instantaneously.

[2] If you want a deep dive into the theory with even more simulations, then this paper by the authors of the **sandwich** paper is another excellent resource.

```
## Within R2                          0.01419                0.01475
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results are the same as above. Oh and guess what? You can get a lot more complicated than that!

Here's the basics of how it works.[3] You can specify:

1. One or more rhs variable using `c(var1,var2,var3)`
2. One or more fixed effects using the stepwise functions `sw()`, `sw0()`, `csw()`, and `csw0()`.
3. One or more independent variable using the stepwise functions `sw()`, `sw0()`, `csw()`, and `csw0()`.
4. Different samples using the `split` or `fsplit` option.

And here's multiple estimations used to their "fuller" potential:

```
gender_pay_gap_many_fes <- feols(c(annual_salary,additional_pay) ~ csw(gender,remote) |
  sw0(age,race1,state,age+race1+state),
  fsplit=~year,
  data=managers)
# Just the annual salary both years pooled
etable(gender_pay_gap_many_fes[lhs='annual_salary',sample=1],title='Annual Salary Gender Pay Gap for 20
```

```
##                              gender_pay_gap_many..1 gender_pay_gap_many..2
## Sample (year)                           Full sample            Full sample
## Dependent Var.:                       annual_salary          annual_salary
##
## Constant                     121,813.5*** (1,083.5) 134,370.5*** (1,274.0)
## genderWoman                  -23,259.8*** (1,197.7) -21,643.3*** (1,181.7)
## remoteHybrid                                         -7,751.3*** (1,132.4)
## remoteOn-site                                       -34,454.2*** (1,196.3)
## remoteOther/it'scomplicated                             -6,585.7* (3,260.2)
## Fixed-Effects:               ---------------------- ----------------------
## age                                              No                     No
## race1                                            No                     No
## state                                            No                     No
## ____________________________ ______________________ ______________________
## S.E. type                                       IID                    IID
## Observations                                 26,424                 26,342
## R2                                          0.01407                0.04876
## Within R2                                        --                     --
##
##                              gender_pay_gap_many..3 gender_pay_gap_many..4
## Sample (year)                           Full sample            Full sample
## Dependent Var.:                       annual_salary          annual_salary
##
## Constant
## genderWoman                  -23,430.9*** (3,319.7) -21,872.7*** (2,991.7)
## remoteHybrid                                          -6,664.3** (1,623.6)
## remoteOn-site                                       -33,535.5*** (1,860.7)
## remoteOther/it'scomplicated                              -7,480.3 (7,112.9)
## Fixed-Effects:               ---------------------- ----------------------
## age                                             Yes                    Yes
## race1                                            No                     No
## state                                            No                     No
## ____________________________ ______________________ ______________________
```

[3]You can find a more in-depth explanation at the Multiple Estimation vignette.

```
## S.E. type                                       by: age               by: age
## Observations                                     26,424                26,342
## R2                                              0.03466               0.06807
## Within R2                                       0.01457               0.04876
##
##                               gender_pay_gap_many..5 gender_pay_gap_many..6
## Sample (year)                            Full sample           Full sample
## Dependent Var.:                        annual_salary         annual_salary
##
## Constant
## genderWoman                   -23,328.4*** (1,368.7) -21,721.3*** (1,479.5)
## remoteHybrid                                            -7,854.5*** (359.1)
## remoteOn-site                                          -34,013.7*** (719.7)
## remoteOther/it'scomplicated                             -6,471.4 (11,071.0)
## Fixed-Effects:                ---------------------- ----------------------
## age                                               No                    No
## race1                                            Yes                   Yes
## state                                             No                    No
## _____________________________ ______________________ ______________________
## S.E. type                                  by: race1             by: race1
## Observations                                  26,376                26,294
## R2                                           0.01859               0.05212
## Within R2                                    0.01419               0.04785
##
##                               gender_pay_gap_many..7 gender_pay_gap_many..8
## Sample (year)                            Full sample           Full sample
## Dependent Var.:                        annual_salary         annual_salary
##
## Constant
## genderWoman                   -23,019.8*** (3,009.4) -21,641.7*** (3,119.4)
## remoteHybrid                                          -9,130.8** (2,878.2)
## remoteOn-site                                        -31,238.2*** (1,787.7)
## remoteOther/it'scomplicated                             -6,356.7 (10,009.3)
## Fixed-Effects:                ---------------------- ----------------------
## age                                               No                    No
## race1                                             No                    No
## state                                            Yes                   Yes
## _____________________________ ______________________ ______________________
## S.E. type                                  by: state             by: state
## Observations                                  26,424                26,342
## R2                                           0.05365               0.07978
## Within R2                                    0.01425               0.04159
##
##                               gender_pay_gap_many..9 gender_pay_gap_man..10
## Sample (year)                            Full sample           Full sample
## Dependent Var.:                        annual_salary         annual_salary
##
## Constant
## genderWoman                   -23,324.7*** (2,939.8) -22,017.8*** (2,726.4)
## remoteHybrid                                          -8,218.2** (1,780.9)
## remoteOn-site                                        -29,799.2*** (1,717.4)
## remoteOther/it'scomplicated                              -7,326.8 (7,059.9)
## Fixed-Effects:                ---------------------- ----------------------
## age                                              Yes                   Yes
```

```
## race1                                           Yes                   Yes
## state                                           Yes                   Yes
## ____________________________ ______________________ ______________________
## S.E. type                                   by: age               by: age
## Observations                                 26,376                26,294
## R2                                          0.08040               0.10419
## Within R2                                   0.01501               0.04069
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Now additional pay
etable(gender_pay_gap_many_fes[lhs='additional_pay',sample=1],title='Additional Compensation Gender Pay
```

```
##                              gender_pay_gap_ma..1 gender_pay_gap_ma..2
## Sample (year)                         Full sample          Full sample
## Dependent Var.:                    additional_pay       additional_pay
##
## Constant                      21,507.4*** (589.1)  23,168.0*** (703.5)
## genderWoman                  -11,948.2*** (654.4) -11,615.6*** (654.0)
## remoteHybrid                                            -235.0 (631.7)
## remoteOn-site                                      -6,298.4*** (671.5)
## remoteOther/it'scomplicated                          1,672.2 (1,827.9)
## Fixed-Effects:               -------------------- --------------------
## age                                            No                   No
## race1                                          No                   No
## state                                          No                   No
## ____________________________ ____________________ ____________________
## S.E. type                                     IID                  IID
## Observations                               20,658               20,600
## R2                                        0.01588              0.02158
## Within R2                                      --                   --
##
##                              gender_pay_gap_many..3 gender_pay_gap_many..4
## Sample (year)                           Full sample            Full sample
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,966.8*** (1,624.1) -11,649.6*** (1,560.1)
## remoteHybrid                                                -78.43 (1,071.0)
## remoteOn-site                                          -6,055.8*** (789.5)
## remoteOther/it'scomplicated                              1,596.2 (1,864.2)
## Fixed-Effects:               ---------------------- ----------------------
## age                                             Yes                    Yes
## race1                                            No                     No
## state                                            No                     No
## ____________________________ ______________________ ______________________
## S.E. type                                   by: age                by: age
## Observations                                 20,658                 20,600
## R2                                          0.01896                0.02425
## Within R2                                   0.01597                0.02138
##
##                              gender_pay_gap_ma..5 gender_pay_gap_ma..6
## Sample (year)                         Full sample          Full sample
## Dependent Var.:                    additional_pay       additional_pay
##
```

```
## Constant
## genderWoman                  -11,918.9*** (611.3) -11,590.1*** (608.8)
## remoteHybrid                                           -248.7 (434.6)
## remoteOn-site                                     -6,188.0*** (254.9)
## remoteOther/it'scomplicated                          1,805.3. (926.7)
## Fixed-Effects:               -------------------- --------------------
## age                                            No                   No
## race1                                         Yes                  Yes
## state                                          No                   No
## ____________________________ ____________________ ____________________
## S.E. type                               by: race1            by: race1
## Observations                               20,620               20,562
## R2                                        0.01766              0.02319
## Within R2                                 0.01580              0.02130
##
##                              gender_pay_gap_many..7 gender_pay_gap_many..8
## Sample (year)                           Full sample            Full sample
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,640.3*** (1,851.9) -11,368.5*** (1,880.5)
## remoteHybrid                                               -312.0 (1,588.7)
## remoteOn-site                                         -5,496.2*** (1,037.8)
## remoteOther/it'scomplicated                               1,839.9 (3,364.9)
## Fixed-Effects:               ---------------------- ----------------------
## age                                              No                     No
## race1                                            No                     No
## state                                           Yes                    Yes
## ____________________________ ______________________ ______________________
## S.E. type                                 by: state              by: state
## Observations                                 20,658                 20,600
## R2                                          0.03308                0.03706
## Within R2                                   0.01524                0.01940
##
##                              gender_pay_gap_many..9 gender_pay_gap_man..10
## Sample (year)                           Full sample            Full sample
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,653.4*** (1,413.4) -11,397.5*** (1,386.1)
## remoteHybrid                                               -187.6 (1,183.3)
## remoteOn-site                                           -5,155.3*** (837.0)
## remoteOther/it'scomplicated                               1,804.4 (1,762.3)
## Fixed-Effects:               ---------------------- ----------------------
## age                                             Yes                    Yes
## race1                                           Yes                    Yes
## state                                           Yes                    Yes
## ____________________________ ______________________ ______________________
## S.E. type                                   by: age                by: age
## Observations                                 20,620                 20,562
## R2                                          0.03824                0.04172
## Within R2                                   0.01531                0.01905
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

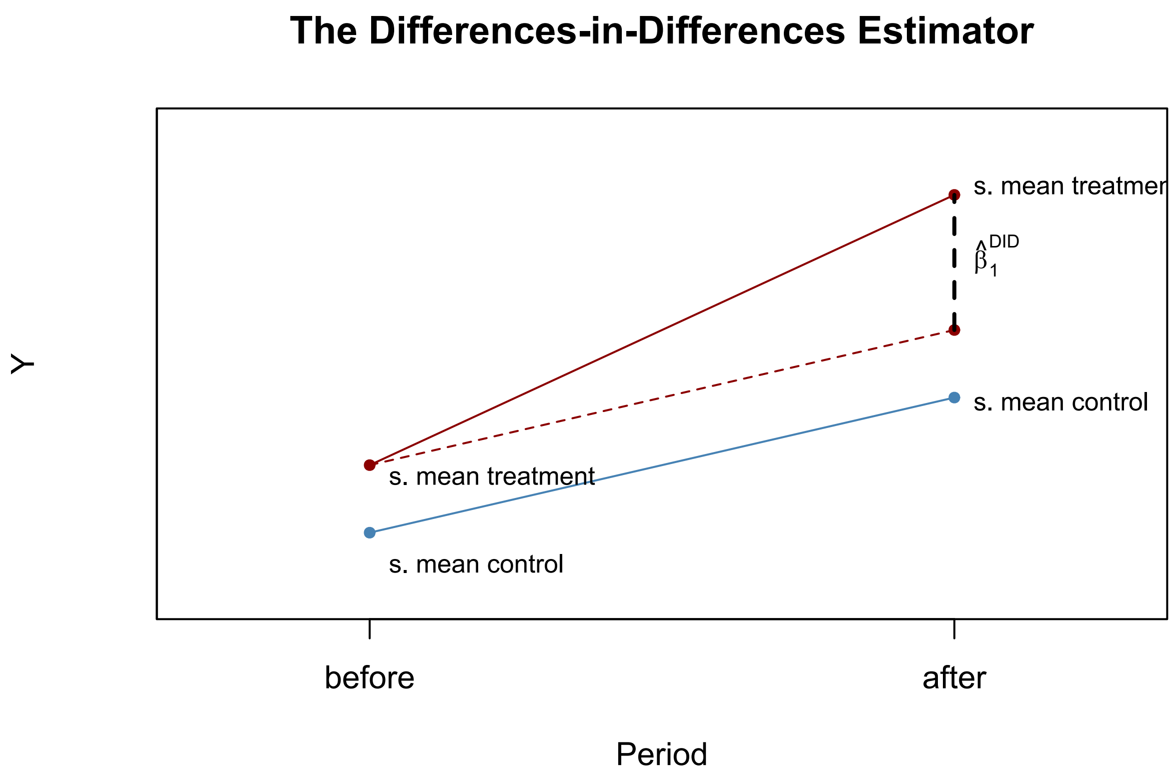
```r
# Now additional pay
etable(gender_pay_gap_many_fes[lhs='additional_pay',sample=3],title='Annual Salary Gender Pay Gap for 20
```

```
##                              gender_pay_gap_ma..1 gender_pay_gap_ma..2
## Sample (year)                                2023                 2023
## Dependent Var.:                    additional_pay       additional_pay
##
## Constant                      21,507.4*** (833.2)  23,168.0*** (995.0)
## genderWoman                  -11,948.2*** (925.5) -11,615.6*** (925.0)
## remoteHybrid                                            -235.0 (893.4)
## remoteOn-site                                      -6,298.4*** (949.8)
## remoteOther/it'scomplicated                          1,672.2 (2,585.3)
## Fixed-Effects:               -------------------- --------------------
## age                                            No                   No
## race1                                          No                   No
## state                                          No                   No
## ____________________________ ____________________ ____________________
## S.E. type                                     IID                  IID
## Observations                               10,329               10,300
## R2                                        0.01588              0.02158
## Within R2                                      --                   --
##
##                              gender_pay_gap_many..3 gender_pay_gap_many..4
## Sample (year)                                  2023                   2023
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,966.8*** (1,624.2) -11,649.6*** (1,560.3)
## remoteHybrid                                            -78.43 (1,071.1)
## remoteOn-site                                        -6,055.8*** (789.5)
## remoteOther/it'scomplicated                            1,596.2 (1,864.3)
## Fixed-Effects:               ---------------------- ----------------------
## age                                             Yes                    Yes
## race1                                            No                     No
## state                                            No                     No
## ____________________________ ______________________ ______________________
## S.E. type                                   by: age                by: age
## Observations                                 10,329                 10,300
## R2                                          0.01896                0.02425
## Within R2                                   0.01597                0.02138
##
##                              gender_pay_gap_ma..5 gender_pay_gap_ma..6
## Sample (year)                                2023                 2023
## Dependent Var.:                    additional_pay       additional_pay
##
## Constant
## genderWoman                  -11,918.9*** (611.3) -11,590.1*** (608.9)
## remoteHybrid                                            -248.7 (434.6)
## remoteOn-site                                      -6,188.0*** (255.0)
## remoteOther/it'scomplicated                           1,805.3. (926.8)
## Fixed-Effects:               -------------------- --------------------
## age                                            No                   No
## race1                                         Yes                  Yes
## state                                          No                   No
```

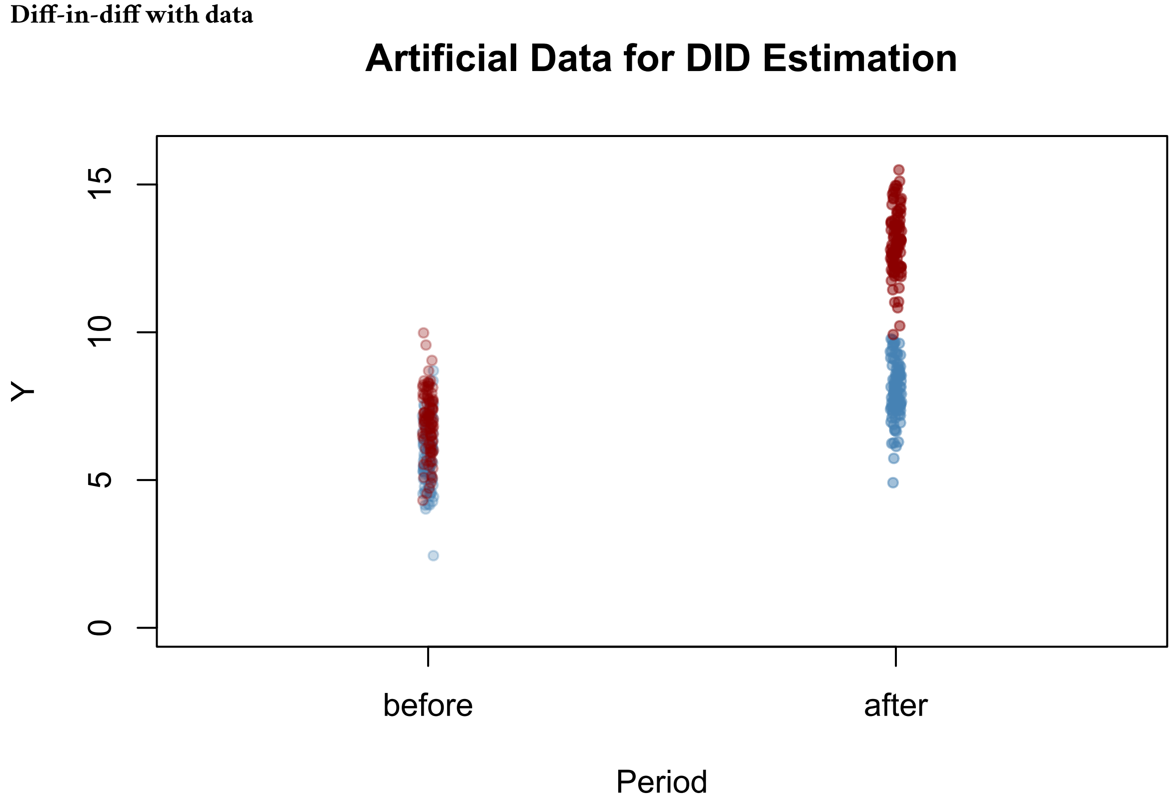
```
## ---------------------------- -------------------- --------------------
## S.E. type                               by: race1            by: race1
## Observations                               10,310               10,281
## R2                                        0.01766              0.02319
## Within R2                                 0.01580              0.02130
##
##                              gender_pay_gap_many..7 gender_pay_gap_many..8
## Sample (year)                                  2023                   2023
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,640.3*** (1,851.9) -11,368.5*** (1,880.7)
## remoteHybrid                                              -312.0 (1,588.9)
## remoteOn-site                                         -5,496.2*** (1,037.9)
## remoteOther/it'scomplicated                              1,839.9 (3,365.2)
## Fixed-Effects:               ---------------------- ----------------------
## age                                              No                     No
## race1                                            No                     No
## state                                           Yes                    Yes
## ____________________________ ______________________ ______________________
## S.E. type                                 by: state              by: state
## Observations                                 10,329                 10,300
## R2                                          0.03308                0.03706
## Within R2                                   0.01524                0.01940
##
##                              gender_pay_gap_many..9 gender_pay_gap_man..10
## Sample (year)                                  2023                   2023
## Dependent Var.:                      additional_pay         additional_pay
##
## Constant
## genderWoman                  -11,653.4*** (1,415.8) -11,397.5*** (1,388.5)
## remoteHybrid                                              -187.6 (1,185.3)
## remoteOn-site                                           -5,155.3*** (838.4)
## remoteOther/it'scomplicated                              1,804.4 (1,765.4)
## Fixed-Effects:               ---------------------- ----------------------
## age                                             Yes                    Yes
## race1                                           Yes                    Yes
## state                                           Yes                    Yes
## ____________________________ ______________________ ______________________
## S.E. type                                   by: age                by: age
## Observations                                 10,310                 10,281
## R2                                          0.03824                0.04172
## Within R2                                   0.01531                0.01905
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Difference-in-differences

One of the most popular uses of fixed effects is to implement difference-in-difference designs. I vusalize how that works for you below.



Diff-in-diff with data



Instrumental variables

As you would have guessed by now, there are a number of ways to run instrumental variable (IV) regressions in R. I'll walk through three different options using the `ivreg::ivreg()`, `estimatr::iv_robust()`, and `fixest::feols()` functions, respectively. These are all going to follow a similar syntax, where the IV first-stage regression is specified in a multi-part formula (i.e. where formula parts are separated by one or more pipes, `|`). However, there are also some subtle and important differences, which is why I want to go through each of them. After that, I'll let you decide which of the three options is your favourite.

The dataset that we'll be using for this section describes cigarette demand for the 48 continental US states in 1995, and is taken from the **ivreg** package. Here's a quick a peek:

```
data("CigaretteDemand", package = "ivreg")
head(CigaretteDemand)
```

```
##        packs   rprice  rincome  salestax   cigtax  packsdiff  pricediff
## AL 101.08543 103.9182 12.91535 0.9216975 26.57481 -0.1418075 0.09010222
## AR 111.04297 115.1854 12.16907 5.4850193 36.41732 -0.1462808 0.19998082
## AZ  71.95417 130.3199 13.53964 6.2057067 42.86964 -0.3733741 0.25576681
## CA  56.85931 138.1264 16.07359 9.0363074 40.02625 -0.5682141 0.32079587
## CO  82.58292 109.8097 16.31556 0.0000000 28.87139 -0.3132622 0.22587189
## CT  79.47219 143.2287 20.96236 8.1072834 48.55643 -0.3184911 0.18546746
##    incomediff salestaxdiff  cigtaxdiff
## AL 0.18222144    0.1332853 -3.62965832
## AR 0.15055894    5.4850193  2.03070663
## AZ 0.05379983    1.4004856 14.05923036
## CA 0.02266877    3.3634447 15.86267924
## CO 0.13002974    0.0000000  0.06098283
## CT 0.18404197   -0.7062239  9.52297455
```

Now, assume that we are interested in regressing the number of cigarettes packs consumed per capita on their average price and people's real incomes. The problem is that the price is endogenous, because it is simultaneously determined by demand and supply. So we need to instrument for it using cigarette sales tax. That is, we want to run the following two-stage IV regression.

$$Price_i = \pi_0 + \pi_1 SalesTax_i + v_i \qquad \text{(First stage)}$$

$$Packs_i = \beta_0 + \beta_2 \widehat{Price_i} + \beta_1 RealIncome_i + u_i \qquad \text{(Second stage)}$$

IV with `fixest::feols()`

Finally, we get back to the `fixest::feols()` function that we've already seen above. Truth be told, this is the IV option that I use most often in my own work. In part, this statement reflects the fact that I work mostly with panel data and will invariably be using **fixest** anyway. But I also happen to like its IV syntax a lot. The key thing is to specify the IV first-stage as a separate formula in the *final* slot of the model call.[4] For example, if we had `fe` fixed effects, then the model call would be `y ~ ex | fe | en ~ in`. Since we don't have any fixed effects in our current cigarette demand example, the first-stage will come directly after the exogenous variables:

```
# library(fixest) ## Already loaded

iv_feols =
  feols(
    log(packs) ~ log(rincome) | ## y ~ ex
      log(rprice) ~ salestax,   ## en ~ in (IV first-stage; must be the final slot)
    data = CigaretteDemand
```

[4]This closely resembles Stata's approach to writing out the IV first-stage, where you specify the endogenous variable(s) and the instruments together in a slot.

```r
    )
# summary(iv_feols, stage = 1) ## Show the 1st stage in detail
iv_feols
```

```
## TSLS estimation, Dep. Var.: log(packs), Endo.: log(rprice), Instr.: salestax
## Second stage: Dep. Var.: log(packs)
## Observations: 48
## Standard-errors: IID
##                  Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)      9.430658   1.358366  6.942648 1.2395e-08 ***
## fit_log(rprice) -1.143375   0.359486 -3.180583 2.6617e-03 **
## log(rincome)     0.214515   0.268585  0.798687 4.2867e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.183555   Adj. R2: 0.393109
## F-test (1st stage), log(rprice): stat = 45.2  , p = 2.655e-8, on 1 and 45 DoF.
##                        Wu-Hausman: stat =  1.102, p = 0.299559, on 1 and 44 DoF.
```

Again, I emphasise that the IV first-stage must always come last in the `feols()` model call. Just to be pedantic — but also to demonstrate how easy **fixest**'s IV functionality extends to panel settings — here's a final `feols()` example. This time, I'll use a panel version of the same US cigarette demand data that includes entries from both 1985 and 1995. The dataset originally comes from the **AER** package, but we can download it from the web as follows. Note that I'm going to modify some variables to make it better comparable to our previous examples.

```r
## Get the data
url = 'https://vincentarelbundock.github.io/Rdatasets/csv/AER/CigarettesSW.csv'
cigs_panel =
  read.csv(url, row.names = 1) %>%
  mutate(
    rprice = price/cpi,
    rincome = income/population/cpi
    )
head(cigs_panel)
```

```
##   state year   cpi population    packs    income  tax     price     taxs
## 1    AL 1985 1.076    3973000 116.4863  46014968 32.5 102.18167 33.34834
## 2    AR 1985 1.076    2327000 128.5346  26210736 37.0 101.47500 37.00000
## 3    AZ 1985 1.076    3184000 104.5226  43956936 31.0 108.57875 36.17042
## 4    CA 1985 1.076   26444000 100.3630 447102816 26.0 107.83734 32.10400
## 5    CO 1985 1.076    3209000 112.9635  49466672 31.0  94.26666 31.00000
## 6    CT 1985 1.076    3201000 109.2784  60063368 42.0 128.02499 51.48333
##      rprice  rincome
## 1  94.96438 10.76387
## 2  94.30762 10.46817
## 3 100.90962 12.83046
## 4 100.22058 15.71332
## 5  87.60842 14.32619
## 6 118.98234 17.43861
```

Let's run a panel IV now, where we'll explicitly account for year and state fixed effects.

```r
iv_feols_panel =
  feols(
    log(packs) ~ log(rincome) |
      year + state |             ## Now include FEs slot
      log(rprice) ~ taxs,        ## IV first-stage still comes last
```

```r
    data = cigs_panel
  )
# summary(iv_feols_panel, stage = 1) ## Show the 1st stage in detail
iv_feols_panel
```

```
## TSLS estimation, Dep. Var.: log(packs), Endo.: log(rprice), Instr.: taxs
## Second stage: Dep. Var.: log(packs)
## Observations: 96
## Fixed-effects: year: 2,  state: 48
## Standard-errors: Clustered (year)
##                  Estimate Std. Error       t value   Pr(>|t|)
## fit_log(rprice) -1.279349   2.11e-15 -6.071802e+14 1.0485e-15 ***
## log(rincome)     0.443422   1.41e-15  3.138717e+14 2.0283e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.044789     Adj. R2: 0.92791
##                  Within R2: 0.533965
## F-test (1st stage), log(rprice): stat = 111.0    , p = 7.535e-14, on 1 and 46 DoF.
##                      Wu-Hausman: stat =   6.02154, p = 0.018161 , on 1 and 44 DoF.
```

Good news, our coefficients are around the same magnitude. But the increased precision of the panel model has yielded gains in statistical significance.

Further resources

- Ed Rubin has outstanding teaching notes for econometrics with R on his website. This includes both undergrad- and graduate-level courses. Seriously, check them out.
- Several introductory texts are freely available, including *Introduction to Econometrics with R* (Christoph Hanck *et al.*), *Using R for Introductory Econometrics* (Florian Heiss), and *Modern Dive* (Chester Ismay and Albert Kim).
- Tyler Ransom has a nice cheat sheet for common regression tasks and specifications.
- Itamar Caspi has written a neat unofficial appendix to this lecture, *recipes for Dummies*. The title might be a little inscrutable if you haven't heard of the `recipes` package before, but basically it handles "tidy" data preprocessing, which is an especially important topic for machine learning methods. We'll get to that later in course, but check out Itamar's post for a good introduction.
- I promised to provide some links to time series analysis. The good news is that R's support for time series is very, very good. The Time Series Analysis task view on CRAN offers an excellent overview of available packages and their functionality.
- Lastly, for more on visualizing regression output, I highly encourage you to look over Chapter 6 of Kieran Healy's *Data Visualization: A Practical Guide*. Not only will learn how to produce beautiful and effective model visualizations, but you'll also pick up a variety of technical tips.