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)

## 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.

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

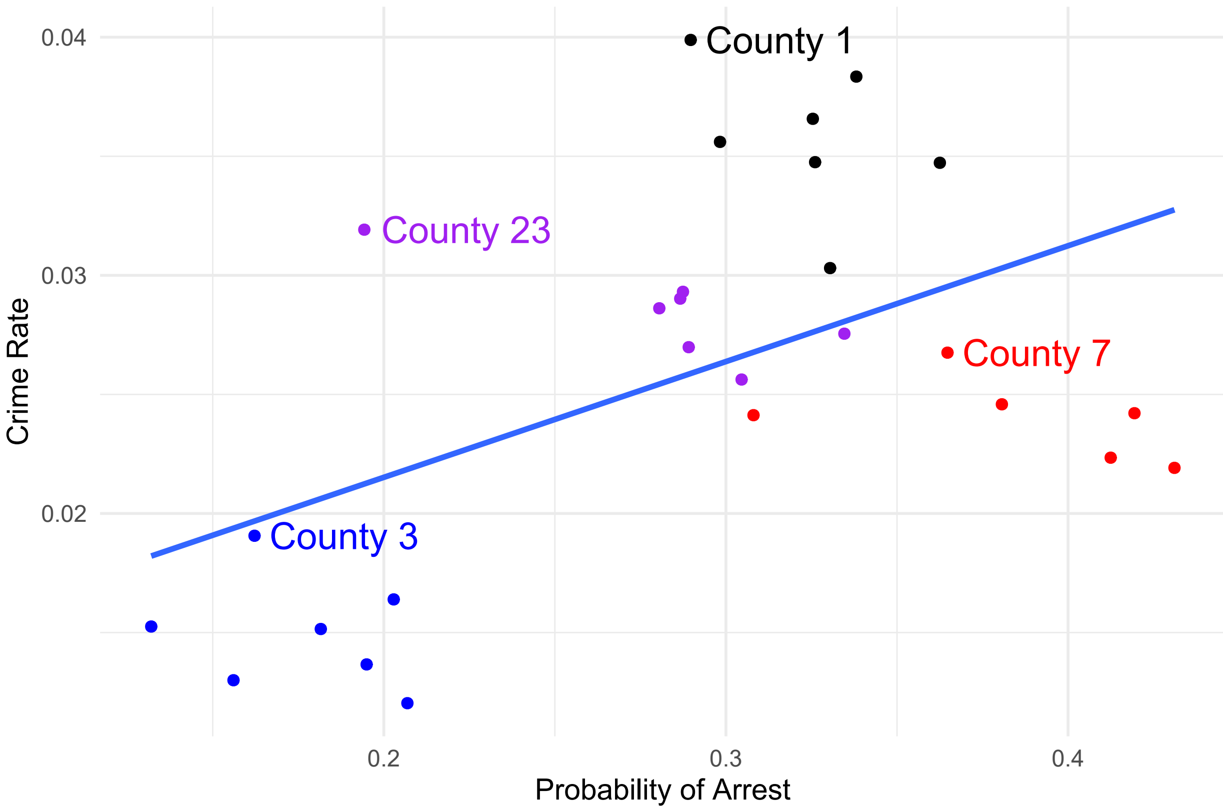
9 rows out of 630. "Prob. of Arrest" is estimated probability of being arrested when you commit a crime

Let's visualize it

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

```r
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_prob = prbarr - mean_prob)
```

And Regress!

```
orig_data <- feols(crmrte ~ prbarr, data = crime4)
de_mean <- feols(demeaned_crime ~ demeaned_prob, data = crime4)
etable(orig_data, de_mean)
```

```
##                         orig_data           de_mean
## Dependent Var.:            crmrte    demeaned_crime
##
## Constant          0.0118* (0.0050) 1.41e-18 (0.0004)
## prbarr           0.0486** (0.0167)
## demeaned_prob                      -0.0305* (0.0117)
## _______________ _________________ _________________
```

```
## S.E. type                      IID             IID
## Observations                    27              27
## R2                         0.25308         0.21445
## Adj. R2                    0.22321         0.18303
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

- 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
- 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.

```r
lsdv <- feols(crmrte ~ prbarr + factor(county), data = crime4)
etable(orig_data, de_mean, lsdv, keep = c('prbarr', 'demeaned_prob'))
```

```
##                          orig_data            de_mean               lsdv
## Dependent Var.:             crmrte     demeaned_crime             crmrte
##
## prbarr           0.0486** (0.0167)                     -0.0305* (0.0124)
## demeaned_prob                       -0.0305* (0.0117)
## ________________ _________________ __________________ __________________
## S.E. type                      IID                IID                IID
## Observations                    27                 27                 27
## R2                         0.25308            0.21445            0.94114
## Adj. R2                    0.22321            0.18303            0.93044
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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:

```r
lsdv
```

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 27
## Standard-errors: IID
```