# 6: Regression II: Paneldata

## Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth
fh@ifs.ku.dk
fghjorth.github.io
@fghjorth

Institut for Statskundskab
Københavns Universitet

11. oktober 2018

Sidste gang:

- OLS intuition
- OLS formel form
- omitted variable bias
- læsning af OLS-output
- implementering i R

## I: Kan en privat uddannelse betale sig?

## II: Hvorfor stemte folk som Brian på Trump?


https://eu.freep.com/story/news/local/michigan/2017/05/28/michigan-donald-trump-voters/344246001/

»The wide format has been the traditional way of analyzing panel data (...) The long format has now become the 'modern' way of organizing panel data« (AGS 64)

| year | artist | track | time | date.entered | wk1 | wk2 | wk3 |
|---|---|---|---|---|---|---|---|
| 2000 | 2 Pac | Baby Don't Cry | 4:22 | 2000-02-26 | 87 | 82 | 72 |
| 2000 | 2Ge+her | The Hardest Part Of ... | 3:15 | 2000-09-02 | 91 | 87 | 92 |
| 2000 | 3 Doors Down | Kryptonite | 3:53 | 2000-04-08 | 81 | 70 | 68 |
| 2000 | 98^0 | Give Me Just One Nig... | 3:24 | 2000-08-19 | 51 | 39 | 34 |
| 2000 | A*Teens | Dancing Queen | 3:44 | 2000-07-08 | 97 | 97 | 96 |
| 2000 | Aaliyah | I Don't Wanna | 4:15 | 2000-01-29 | 84 | 62 | 51 |
| 2000 | Aaliyah | Try Again | 4:03 | 2000-03-18 | 59 | 53 | 38 |
| 2000 | Adams, Yolanda | Open My Heart | 5:30 | 2000-08-26 | 76 | 76 | 74 |

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are `wk4`, `wk5`, ..., `wk75`.

| year | artist | time | track | date | week | rank |
|---|---|---|---|---|---|---|
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-02-26 | 1 | 87 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-04 | 2 | 82 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-11 | 3 | 72 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-18 | 4 | 77 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-03-25 | 5 | 87 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-04-01 | 6 | 94 |
| 2000 | 2 Pac | 4:22 | Baby Don't Cry | 2000-04-08 | 7 | 99 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-02 | 1 | 91 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-09 | 2 | 87 |
| 2000 | 2Ge+her | 3:15 | The Hardest Part Of ... | 2000-09-16 | 3 | 92 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-08 | 1 | 81 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-15 | 2 | 70 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-22 | 3 | 68 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-04-29 | 4 | 67 |
| 2000 | 3 Doors Down | 3:53 | Kryptonite | 2000-05-06 | 5 | 66 |

Table 8: First fifteen rows of the tidied Billboard dataset. The `date` column does not appear in the original table, but can be computed from `date.entered` and `week`.

Data på bred form kan konverteres til lang form vha. `gather()` i `dplyr`-pakken

```
gather(<data>,<key>,<value>,<...>)
```

hvor

- key: navn på variabel der angiver variabelnavne fra bredt format
- value: navn på variabel der angiver værdier fra bredt format
- ... : intervallet af variable der skal 'stakkes', fx. `obs1992:obs1998`

Eks.: bredt data med enhederne a, b og c og outcome y observeret i t2 og t2

| unit | yt1 | yt2 |
|---|---|---|
| a | 1 | 2 |
| b | 3 | 6 |
| c | 6 | 7 |

$\rightarrow$ hvordan skal data se ud på lang form?

- en normal, 'pooled' OLS tager ikke højde for at observationer ikke er uafhængige
- specifikt: observationer for samme enhed kan 'klumpe sammen'
- konsekvensen er at standardfejlene underestimeres $\rightarrow$ det er ikke godt!

Klassiske OLS-antagelser (jf. AGS boks 4.1):

1. simpel tilfældig udvælgelse
2. linearitet
3. ej perfekt multikollinearitet
4. fejlled ukorrelerede med X'er
5. varianshomogenitet
6. fravær af autokorrelation
7. normalfordelte fejlled

$\rightarrow$ hvilken trues i paneldata?

Notation i AGS (p. 127):

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + \gamma_1 z_{1i} + \ldots + \gamma_j z_{ji} + u_i + e_{it} \quad (1)$$

$u_i$: uobserverede, tidsinvariante prediktorer $\rightarrow$ med paneldata kan vi kontrollere for al confounding herfra!

Husk formlen fra sidste gang:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i \tag{2}$$

Antag nu at vi observerer indkomst ($Y_i$) og privat uddannelse ($P_i$) over tid $t$:

$$Y_{it} = \alpha + \beta P_{it} + \gamma A_i + e_{it} \tag{3}$$

NB: $A_i$ varierer her ikke med $t$, dvs. er *tidsinvariant*

Så længe $A_i$ er tidsinvariant kan vi med paneldata estimere $\beta$ uden bias uden at observere $A_i$:

$$Y_i = \alpha_i + \lambda_t + \beta P_i + e_i \tag{4}$$

- kaldes en 'fixed effects' (FE) model
- $\alpha_i$ = fixed effects for enheder $\rightarrow$ opfanger *tidsinvariant uobserveret heterogenitet ml. enheder*
- $\lambda_t$ = fixed effects for tid $\rightarrow$ opfanger *enhedsinvariant uobserveret heterogenitet ml. tidsperioder*
- tilbageværende variation kun variationen 'inden for' enheder $\rightarrow$ FE-model kaldes også *within-estimator*

Illustration:

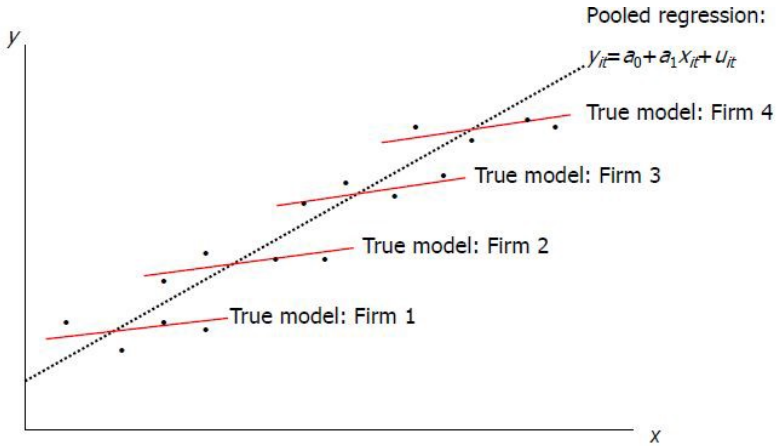

**Table 1. Predicting change in presidential support from 2012 to 2016: Fixed effects analysis**

| | Model 1: Thermometer advantage | | | | Model 2: Vote choice among validated voters | | | |
|---|---|---|---|---|---|---|---|---|
| | Effects of change in predictors on change in Republican thermometer advantage | | Effects of change in salience of 2012 predictors on change in Republican thermometer advantage (predictor by wave) | | Effects of change in predictors on change in presidential vote choice | | Effects of change in salience of 2012 predictors on change in presidential vote choice (predictor by wave) | |
| Change in predictors | Coefficient | *z* Value | Coefficient | *z* Value | Coefficient | *z* Value | Coefficient | *z* Value |
| Party identification (Democrat) | −0.686 | −2.870** | 0.275 | 1.420 | −1.610 | −8.121*** | −0.551 | −1.589 |
| Personal economic hardship | | | | | | | | |
| Household income | −0.004 | −0.080 | −0.036 | −1.070 | −0.052 | −1.082 | −0.029 | −0.399 |
| Looking for work | 0.006 | 0.010 | 0.624 | 0.760 | −0.692 | −0.691 | −2.162 | −1.481 |
| Personal finances (better) | −0.032 | −0.190 | −0.104 | −0.540 | −0.025 | −0.107 | 0.228 | 0.545 |
| Personal effects of trade (better) | −0.303 | −1.850 | −0.253 | −1.270 | 0.104 | 0.530 | −0.321 | −1.205 |
| Own issue opinions | | | | | | | | |
| On trade | −0.037 | −0.290 | 0.042 | 0.300 | −0.029 | −0.200 | −0.261 | −1.098 |
| On immigration | −0.170 | −1.490 | −0.219 | −1.770 | 0.103 | 0.768 | 0.138 | 0.652 |
| On China | 0.190 | 1.640 | 0.002 | 0.020 | 0.112 | 0.821 | −0.035 | −0.154 |
| Perceived distance of Democratic candidate on issues | | | | | | | | |
| On trade | 0.120 | 1.140 | −0.108 | −0.760 | 0.530 | 3.116** | 0.166 | 0.890 |
| On immigration | 0.199 | 2.000* | −0.086 | −0.680 | 0.338 | 2.425* | 0.099 | 0.422 |
| On China | 0.392 | 3.840*** | 0.106 | 0.830 | 0.370 | 2.748** | −0.086 | −0.315 |
| Perceived distance of Republican candidate on issues | | | | | | | | |
| On trade | −0.213 | −2.280* | −0.034 | −0.260 | −0.484 | −2.986** | −0.239 | −0.921 |
| On immigration | −0.010 | −0.110 | 0.219 | 1.930 | −0.418 | −3.208** | −0.274 | −1.059 |
| On China | −0.206 | −2.340* | 0.072 | 0.650 | −0.357 | −2.963*** | −0.017 | −0.061 |
| SDO | 0.184 | 2.570* | −0.022 | −0.280 | 0.276 | 2.556* | −0.046 | −0.246 |
| National economy | −0.583 | −3.730*** | 0.083 | 0.440 | −0.773 | −3.884*** | −0.296 | −0.722 |
| Economic context[†] | | | | | | | | |
| Unemployed, % | | | −0.035 | −0.520 | | | −0.077 | −0.407 |
| Manufacturing, % | | | 0.018 | 0.900 | | | −0.072 | −1.712 |
| Median income | | | −0.007 | −1.160 | | | −0.011 | −0.729 |
| Wave (2012–2016) | | | 0.811 | 0.620 | | | 5.396 | 2.165* |
| Constant | | | 12.710 | 10.590*** | | | 3.981 | 2.663* |
| $R^2$/pseudo-$R^2$ | | | 0.65 | | | | 0.78 | |
| Sample size (*n*) | | | 1,088 | | | | 793 | |

Note that results are based on single fixed effects models for thermometer advantage (columns 2 through 5) and vote choice among validated voters (columns 6 through 9) using robust SEs, and incorporating tests of both priming and change in attitudes over time. Fixed effects ordinary least squares regression was used to analyze change in Republican thermometer advantage; fixed effects logit regression was used to analyze Republican versus Democratic vote. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Mutz: FE eliminerer konstante 'assimilationseffekter':

»With issue placements of this kind, in cross-sectional analyses, there is a risk that respondents will assimilate the positions of the candidate that they prefer and/or contrast the views of the opponent (50). Assimilation/contrast renders the perceived distance of candidates endogenous to candidate preferences. Fortunately, this issue is less problematic with repeated measure analyses of panel data. Because each respondent is compared with himself or herself at a previous point in time, any tendency to assimilate or contrast will occur at both points in time, thus canceling itself out when looking at the difference in distances from candidates from one election to the next.« (4)

Mutz: FE estimerer kun 'within'-effekter:

»Fixed effects panel analyses provide the most rigorous test of causality possible with observational data. Because the goal is understanding what changed from 2012 to 2016 to facilitate greater support for Trump in 2016 than Mitt Romney in 2012, I estimate the effects of time-varying independent variables to determine whether changes in the independent variables produce changes in candidate choice without needing to fully specify a model including all possible influences on candidate preference. Significant coefficients thus represent evidence that change in an independent variable corresponds to change in the dependent variable at the individual level.«

- paneldatamodeller ikke R's stærke side
- FE-modeller kan implementeres med `plm::plm()`
- clustered standardfejl kan implementeres med `lmtest::coeftest()`

Næste uge:

NEWS · 10.10.12 VOL 48 / ISSUE 41

## Mr. Autumn Man Walking Down Street With Cup Of Coffee, Wearing Sweater Over Plaid Collared Shirt

*Mr. Autumn Man, enjoying a seasonal stroll.*

Næste gang:

- R workshop II
  - tidying
  - visualisering
  - anskaffelse af tekstkorpus
- øvelse:
  1. hent datasættet fra International Migration Database `https://stats.oecd.org/Index.aspx?DataSetCode=MIG` og lave det om til tidy format
  2. modellér Mutz' `chinaself` med en variabel i hhv. tværsnits- og panelmodeller. Hvordan ændrer koefficienten sig?

Tak for i dag!