4th Lecture: Panel Data Models

Michal Hakala, Pavel Fišer

9.5.2023

TODAY'S LECTURE

1. Cross-sectional VS Panel Data
2. Pooled OLS
3. Fixed Effect Model
4. Random Effect Model
5. Income and Crime Rate
6. Marketing Channels Profitability

Michal

Data Scientist in

MSD IT

Pavel

Data Scientist in

MSD IT

Cross-sectional Data

Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)
1	66 000	18	1	1
2	52 000	13	0	1
…	…	…	…	
n	64 000	22	1	0

- Individual indices $i = 1,2, \ldots, n$
- Model example: $income_i = \beta_0 + \beta_1 education_i + \beta_2 gender_i + \beta_3 course_i + \epsilon_i$
- Hypothesis: Czechitas course has positive impact on income
- Omitted Variable Bias?

czechitas

Panel Data

Time (t)	2018 (t=1)	2019 (t=2)	2020 (t=3)

2018 (t=1)

Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)
1	61 000	17	1	0
2	50 000	13	0	0
…	…	…	…	
n	60 000	21	1	0

2019 (t=2)

Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)
1	66 000	18	1	0
2	52 000	13	0	1
…	…	…	…	
n	64 000	22	1	0

2020 (t=3)

Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)
1	69 000	19	1	1
2	55 000	13	0	1
…	…	…	…	
n	68 000	22	1	0

- Individual indices $i = 1,2,\dots,n$
- Time indices $t = 1,2,\dots,T$
- Model example: $income_{it} = \beta_0 + \beta_1 education_{it} + \beta_2 gender_i + \beta_3 course_{it} + \epsilon_{it}$
- Multiple individuals are observed over multiple periods

Time Series – Side Note

Time (t)	2018 (t=1)					2019 (t=2)					2020 (t=3)				
	Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)	Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)	Person (index)	Income	Education (in years)	Gender (male=0, female=1)	Czechitas Course (no=0, yes=1)
	1	61 000	17	1	0	1	66 000	18	1	0	1	69 000	19	1	1

- One individual
- Time indices $t = 1,2,\ldots,T$
- Models mainly about dynamic properties (=how past affects future)

Reminder - Properties of Estimators

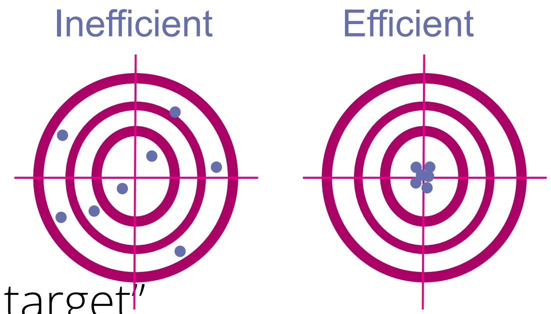
- Estimates of betas are random variables

Consistency

- As sample size is increasing probability of getting estimate different from the true beta is going to zero.
- “Shooting at the right target”

Efficiency

- Variance of estimates
- “How much the shoots are spread out around target”



czechitas

Pooled OLS (Ordinary Least Squares)

- Model example: $income_{it} = \beta_0 + \beta_1 education_{it} + \beta_2 gender_i + \beta_3 course_{it} + \varepsilon_{it}$
- Treating panel data as cross-sectional when estimating β's (=pooled)

OLS Assumption: Random Sample

"Data are identically and Independently distributed."

Panel Setup:

- Violates independence
- Observations of one individual are correlated across time (=less informative)
- E.g., observing 5x that one individual is a woman is not more informative than observing it once

Pooled OLS:

- Model is estimated by (pooled) OLS

czechitas

Individual Effect - Motivation

Causal Impact

"Impact of X on Y while everything else remains the same."

Individual Effect Model

- Allows to "control" for **time-invariant variables** (talent, gender, personality, ...)
- Utilizing repeated observations (more precise estimates)

$$y_{it} = \alpha_i + x_{it}^T \beta + \varepsilon_{it},$$

α_i - individual effect, x_{it} - vector of explanatory variables

czechitas

Individual Effect - Example

- Specific Case

$$income_{it} = \underbrace{\beta_0 + \beta_2 gender_i}_{\alpha_i} + \beta_1 education_{it} + \beta_3 course_{it} + \varepsilon_{it}$$

General Case

$$income_{it} = \underbrace{\beta_0 + \beta_2 gender_i + \beta_4 ability_i + \beta_5 ethnicity_i + \cdots}_{\alpha_i} + \beta_1 education_{it} + \beta_3 course_{it} + \varepsilon_{it}$$

- individual effect $\boldsymbol{\alpha_i}$ captures time-invariant covariates
- we are able to estimate or remove $\boldsymbol{\alpha_i}$ since we observe $\boldsymbol{i}$-th individual multiple times

czechitas

Individual Effect Models - Assumptions

- $income_{it} = \underbrace{\beta_0 + \beta_2 gender_i + \beta_4 ability_i + \beta_5 ethnicity_i + \cdots}_{\alpha_i} + \beta_1 education_{it} + \beta_3 course_{it} + \varepsilon_{it}$

Fixed Effect Model

- individual effect α_i is correlated with some of our explanatory variables

Random Effect Model

- individual effect α_i is NOT correlated with some of our explanatory variables

A) Fixed Effect Model – Within Estimator

Model

$$income_{it} = \underbrace{\beta_0 + \beta_2 gender_i}_{\alpha_i} + \beta_1 education_{it} + \beta_3 course_{it} + \varepsilon_{it}$$

Demeaned Model

- Take averages across time and demean data

$$income_{it} - \overline{income}_i = \underbrace{\beta_0 + \beta_2 gender_i - (\beta_0 + \beta_2 gender_i)}_{\alpha_i - \alpha_i = 0} +$$

$$\beta_1 \left(education_{it} - \overline{education}_i \right) + \beta_3 \underbrace{\left(course_{it} - \overline{course}_i \right)}_{\text{"new explanatory variable"}} + \underbrace{\varepsilon_{it} - \bar{\varepsilon}_i}_{\text{"new error"}}$$

czechitas

A) Fixed Effect Model – Within Estimator

Demeaned Model

$$\overbrace{income_{it} - \overline{income}_i}^{\text{"new explained variable"}} =$$

$$\beta_1 \left(education_{it} - \overline{education}_i\right) + \beta_3 \underbrace{\left(course_{it} - \overline{course}_i\right)}_{\text{"new explanatory variable"}} + \underbrace{\varepsilon_{it} - \bar{\varepsilon}_i}_{\text{"new error"}}$$

Within Estimator

1. Compute demeaned data
2. Use OLS on demeaned data (without intercept)

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)^T \beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

A) Fixed Effect Model – Dummy Estimator

- Model

$$y_{it} = \alpha_i + x_{it}^T \beta + \varepsilon_{it},$$

α_i - individual effect, x_{it} - vector of explanatory variables

- Represent individual i by a dummy variable
  (n variables, each equal to 1 for i-th individual, 0 otherwise)
- Identical estimates to the within estimator
- May be numerically challenging as n is large

A) Fixed Effect Model – Dummy Estimator

- Two-Way Effect Model

$$y_{it} = \alpha_i + \tau_t + x_{it}^T\beta + \varepsilon_{it},$$

α_i - individual effect, τ_t - time effect, x_{it} - vector of explanatory variables

- τ_t - represents effect of period t
- Shared across individuals
- Our example: income in 2020 likely affected by COVID

A) Fixed Effect Model – First Differences Estimator

1. Subtracting values from the previous period instead demeaning
2. Apply OLS

Notes

- Similar idea to within estimator
- Lose the first period (no zero period is available)
- Sometimes less efficient than within estimator

A) Fixed Effect Model – Assumptions

Shared with cross-sectional OLS

1. Linearity of the model
2. Random sample
3. No omitted variable (only time-variant are problematic)
4. No multicollinearity
5. Homoskedasticity (⇒ heteroskedastic robust errors)

Additionally

1. Only time-variant variables
2. Uncorrelated errors – $Cov(\varepsilon_{it}, \varepsilon_{is}) = 0, t \neq s$

   ⇒violation makes inference invalid (standard errors, t-tests for betas)

   ⇒ use autocorrelation and heteroskedastic robust errors

A) Fixed Effect Model – Test of Autocorrelation

8. For t ≠ s, errors are uncorrelated, conditional on explanatory variables and α_i.

HOW TO TEST?

Durbin-Watson test

H_0 = no first order autocorrelation (first order = lag of one time unit)
H_1 = first order correlation exists

Test result: a test statistic with a value between 0-4:

- 2 is no autocorrelation
- 0 to <2 is positive autocorrelation
- >2 to 4 is negative autocorrelation

czechitas

B) Random Effect Model

- Redefined (Longer) Example

$$income_{it} = \beta_0 + \beta_1 education_{it} + \beta_2 gender_i + \beta_3 course_{it} + \underbrace{\alpha_i + \varepsilon_{it}}_{v_i}$$

α_i - unobserved time-invariant variables (talent, ethnicity, ...)

ε_{it} - independent errors

New composite error

$$v_{it} = \alpha_i + \varepsilon_{it}$$

- Correlated across time $Cov(v_{it}, v_{is}) \neq 0$, since α_i is present in every period
- Uncorrelated with explanatory variables (assumption!)

czechitas

B) Random Effect Model

Redefined (Longer) Example

$$income_{it} = \beta_0 + \beta_1 education_{it} + \beta_2 gender_i + \beta_3 course_{it} + \underbrace{\alpha_i + \varepsilon_{it}}_{v_i}$$

Idea

- correlation in v_i makes estimates of betas less efficient in comparison to case where v_i is not correlated
- Random effect estimator does "quasi-demeaning" that removes correlation in errors
- **Remark:** equivalent to pooled OLS when $Var[\alpha_i] = 0$

czechitas

B) Random Effect Model - Assumptions

1. Linearity of the model
2. Random sample
3. No omitted variable
4. No multicollinearity
5. Homoskedasticity (⇒ heteroskedastic robust errors)
6. Uncorrelated errors (⇒ autocorrelation robust errors)

→ The model can have variables that are constant in time for all individuals

FE vs RE – Practical Aspects

Fixed Effect (FE) Model

- PRO: all time-invariant variables are "controlled" by α_i
  ⇒ no omitted variable bias caused by unobserved time-invariant variables
- CON: all time-invariant variables are "controlled" by α_i
  ⇒ cannot study impact of time-invariant variables

Random Effect (RE) Model

- Pro: allows to study impact of time-invariant variables
- Con: assumption of uncorrelated errors with x_{it}
  ⇒ omitted variable bias when assumption is violated

czechitas

FE vs RE – Statistical Aspects

- Fixed Effect (FE) Model
  - PRO: always consistent
  - CON: not efficient when RE is correct

Random Effect (RE) Model
  - PRO: efficient ⇒ smaller variance of beta estimates
  - CON: inconsistent when FE is correct

czechitas

FE vs RE – Test

Formal assessment: Hausman test

Null hypothesis: Random effects is the preferred model

If we reject the null hypothesis (p-value < 0.05), we need to use fixed effects

- Tests for "presence" of fixed effects

Summary – Pooled OLS, FE, RE

Pooled OLS	Fixed Effect	Random Effect
• Inconsistent when FE is present • Efficient when no individual effect is present • Less efficient than RE when RE is present	• Consistent • Less efficient than RE or Pooled OLS when FE are not present	• Inconsistent when FE is present • More efficient than Pooled OLS when RE is present

Hausman Test

czechitas

Quiz 1

- You randomly assign a treatment/placebo to a group of people and survey their health condition in 1 month, 2 months and 3 months. You are interested in impact of treatment. What is the best model?

A. Pooled OLS

B. Fixed Effect

C. Random Effect

Answer: C)

- Random assignment of treatment/placebo guarantees independence of individual effect and treatment ⇒ RE preferred over FE
- Serial correlation is present - health condition of one particular patient in 1st month is clearly correlated with his condition in 2nd month ⇒ RE preferred over Pooled OLS

Quiz 2

- You want to evaluate career impact of a particular Czechitas course. You are surveying Czechitas participants about their income, education, etc. on an annual basis over several years.
It is an intensive course and you expect that talented and motivated people are more likely enrolled to that course. You do not observe talent. What is the best model?

A. Pooled OLS
B. Fixed Effect
C. Random Effect

Answer: B)

- Expecting that certain individuals are more likely to pick the course means that course enrollment is correlated with individual characteristics ⇒ fixed effect.

Final Remark – What Is Beyond FE and RE?

Short Panel Models

- discussed today – Pooled OLS, FE, and RE
- About studying impact of X on Y within the same period
- Usually high n, low T

Long (Dynamic) Panel Models

- Out of scope of this course
- About studying impact of past (both X and Y) on current value of Y
- Low n, high T

MOCKUP EXAMPLE:

How does the income affects crime rate?

Guns dataset

"Guns is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977–1999."

For simplicity, we will only use partial data:

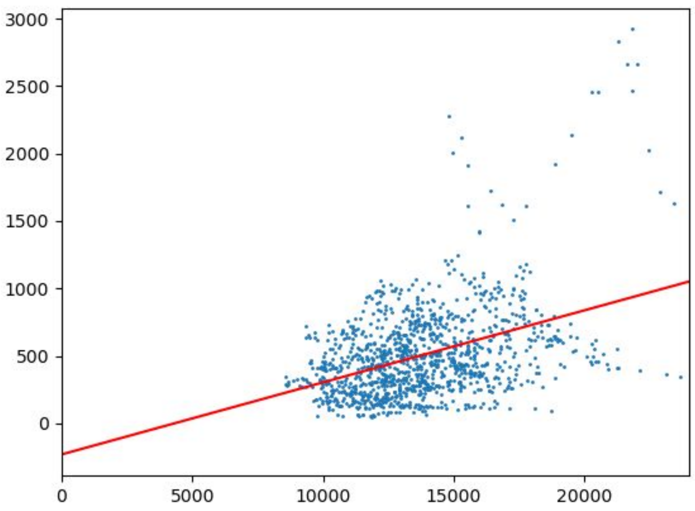
- State (i)
- Year (t)
- Income – per capita personal income
- Violent – violent crime rates (incidents/ 100,000 inhabitants)

Link to download data
Data documentation

Step 1: Load + transform data

```
# 1. Load data
dataset = pd.read_csv('data/Guns.csv',
                      usecols=['state', 'year', 'income', 'violent'],
                      index_col=['state', 'year'])
dataset_full = pd.read_csv('data/Guns.csv')
years = dataset.index.get_level_values('year').to_list()
dataset['year'] = pd.Categorical(years)
```

POLS ignores both cross sectional and time panel structure




```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1665
Estimator:                  PooledOLS   R-squared (Between):              0.1940
No. Observations:                1173   R-squared (Within):              -0.0720
Date:                Fri, Apr 28 2023   R-squared (Overall):              0.1665
Time:                        17:27:38   Log-likelihood                   -8374.6
Cov. Estimator:             Clustered
                                        F-statistic:                      233.84
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                  F(1,1171)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             5.5967
                                        P-value                           0.0182
Time periods:                      23   Distribution:                  F(1,1171)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const         -229.66     287.81    -0.7979     0.4251     -794.35      335.03
income         0.0534     0.0226     2.3657     0.0182      0.0091      0.0977
==============================================================================
```

czechitas

```
                         PooledOLS Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1665
Estimator:                  PooledOLS   R-squared (Between):              0.1940
No. Observations:                1173   R-squared (Within):              -0.0720
Date:                Fri, Apr 28 2023   R-squared (Overall):              0.1665
Time:                        17:27:38   Log-likelihood                   -8374.6
Cov. Estimator:             Clustered
                                        F-statistic:                      233.84
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                  F(1,1171)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             5.5967
                                        P-value                           0.0182
Time periods:                      23   Distribution:                  F(1,1171)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const         -229.66     287.81    -0.7979     0.4251     -794.35      335.03
income         0.0534     0.0226     2.3657     0.0182      0.0091      0.0977
==============================================================================
```

1. What is value of $\beta_0 + \beta_1$ coefficients?
2. Are both $\beta_0 + \beta_1$ coefficients significant?
3. Based on the model results.
   What would be the crime rate (violent) for income
   i. 1,000
   ii. 10,000
   iii. 20,000

czechitas

```
                     PooledOLS Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1665
Estimator:                  PooledOLS   R-squared (Between):              0.1940
No. Observations:                1173   R-squared (Within):              -0.0720
Date:                Fri, Apr 28 2023   R-squared (Overall):              0.1665
Time:                        17:27:38   Log-likelihood                   -8374.6
Cov. Estimator:             Clustered
                                        F-statistic:                      233.84
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                  F(1,1171)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             5.5967
                                        P-value                           0.0182
Time periods:                      23   Distribution:                  F(1,1171)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const         -229.66     287.81    -0.7979     0.4251     -794.35      335.03
income         0.0534     0.0226     2.3657     0.0182      0.0091      0.0977
==============================================================================
```

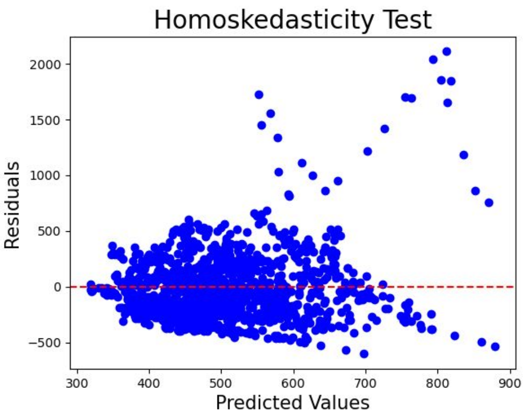
1. What is value of $\beta_0 + \beta_1$ coefficients?

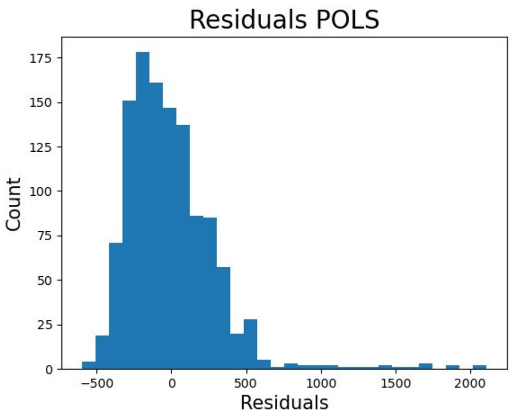
2. Are both $\beta_0 + \beta_1$ coefficients significant?

3. Based on the model results.
   What would be the crime rate (violent) for income
   i. income = 1,000 => violent = -176
   ii. income = 10,000 => violent = 304
   iii. income = 20,000 => violent = 838

Step 3a: Test assumptions!
Check error term/residuals from the model

Visual inspection





Do we see paterns in residuals?

Are residuals normally distributed?

czechitas

Step 3b: Test assumptions!
Autocorrelation – Error terms correlation over time?

Statistical test

Durbin-Watson test
Test statistic = 0.089

0-2	positive autocorrelation
2	zero autocorrelation
2-4	negative autocorrelation

A rule of thumb is that values in the range of 1.5 to 2.5 are relatively normal

Outcome
Autocorrelation is present

Step 3: Test assumptions!
Quiz: Given the outcomes can we use POLS?

Statistical tests

Residuals visual inspection
indicates patterns

Durbin-Watson test
Autocorrelation is present

Step 3: Test assumptions!
Quiz: Given the outcomes can we use POLS?

Statistical tests

Residuals visual inspection
indicates patterns

Durbin-Watson test
Autocorrelation is present

→

POLS assumptions are violated

FE/RE models will be more suitable

Step 4: Fixed Effects (FE) and Random Effects (RE) models

Is RE better model than FE?

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1128
Estimator:              RandomEffects   R-squared (Between):              0.1159
No. Observations:                1173   R-squared (Within):               0.1127
Date:                Tue, May 02 2023   R-squared (Overall):              0.1156
Time:                        20:30:00   Log-likelihood                   -7109.8
Cov. Estimator:            Unadjusted
                                        F-statistic:                      148.90
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                 F(1,1171)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             148.90
                                        P-value                           0.0000
Time periods:                      23   Distribution:                 F(1,1171)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          175.39     48.461     3.6193     0.0003      80.314      270.47
income         0.0239     0.0020     12.203     0.0000      0.0200      0.0277
==============================================================================
```

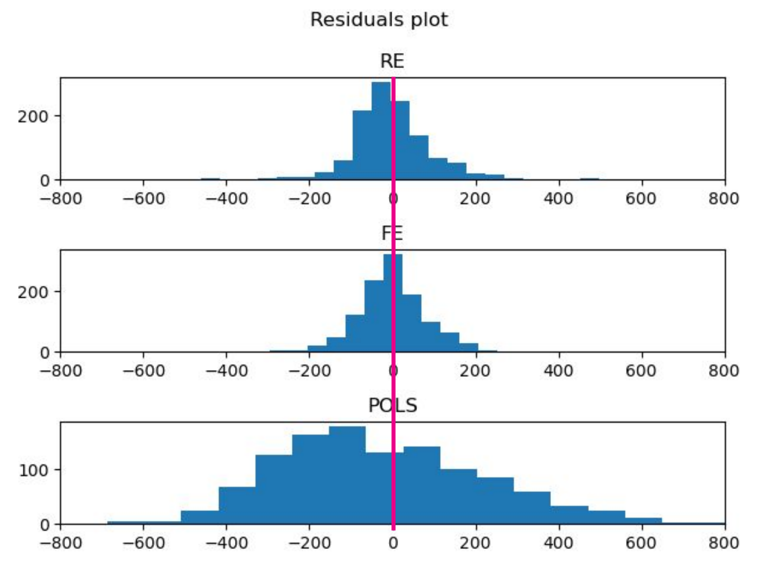
```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1127
Estimator:                   PanelOLS   R-squared (Between):              0.1141
No. Observations:                1173   R-squared (Within):               0.1127
Date:                Tue, May 02 2023   R-squared (Overall):              0.1140
Time:                        20:30:01   Log-likelihood                   -7081.9
Cov. Estimator:            Unadjusted
                                        F-statistic:                      142.39
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                 F(1,1121)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             142.39
                                        P-value                           0.0000
Time periods:                      23   Distribution:                 F(1,1121)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          181.70     27.101     6.7046     0.0000      128.53      234.88
income         0.0234     0.0020     11.933     0.0000      0.0196      0.0273
==============================================================================
```

Step 5: Test assumptions!

QUIZ: Is RE better model than FE?
Do we reject Null hypothesis of Hausman test?

Visual inspection



Statistical test

Hausmann test
p-value = 0.008

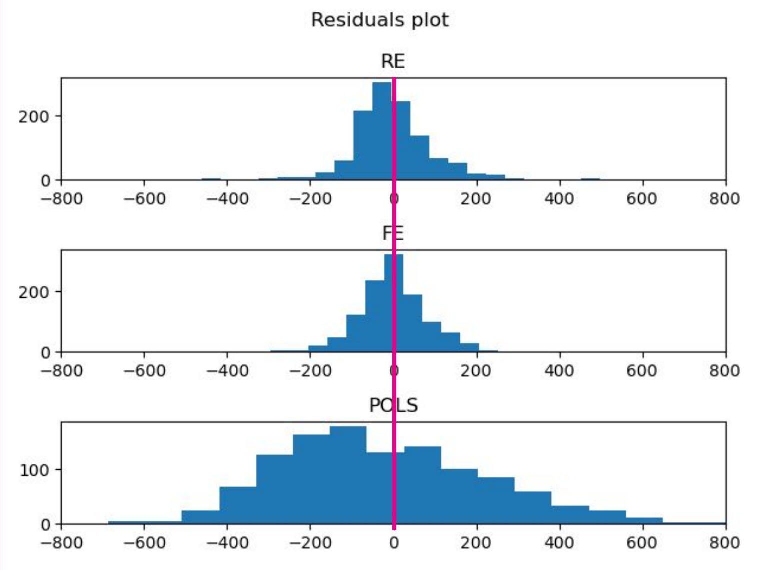
Null hypothesis:
Random effects is the preferred model

echitas

Step 5: Test assumptions!

QUIZ: Is RE better model than FE?
Do we reject Null hypothesis of Hausman test?

Visual inspection



Statistical test

Hausmann test
p-value = 0.008

Null hypothesis:
Random effects is the preferred model

Outcome
Null hypothesis is rejected as p-value < 0.05.
Fixed Effects should be used.

Step 5: Test assumptions!

Is RE better model than FE?

Statistical tests

Hausmann test
p-value = 0.008 < 0.05

Null hypothesis (Random effects is the preferred model) is rejected

FE models is the winner

Step 5: Test assumptions!

Autocorrelation – Error terms correlation over time?

Statistical test

Durbin-Watson test
Test statistic = 0.4

0-2	positive autocorrelation
2	zero autocorrelation
2-4	negative autocorrelation

A rule of thumb is that values in the range of 1.5 to 2.5 are relatively normal

Outcome
Autocorrelation improved but is still present

Step 5: Test assumptions!

Statistical tests

Residuals visual inspection
looks well

Durbin-Watson test
Autocorrelation is still present.
Test statistics improved

→

FE model is the best model we have.
We should used **robust estimator**.

Step 6: Results interpretation

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                violent   R-squared:                        0.1127
Estimator:                   PanelOLS   R-squared (Between):              0.1141
No. Observations:                1173   R-squared (Within):               0.1127
Date:                Fri, May 05 2023   R-squared (Overall):              0.1140
Time:                        16:59:00   Log-likelihood                   -7081.9
Cov. Estimator:                Robust
                                        F-statistic:                      142.39
Entities:                          51   P-value                           0.0000
Avg Obs:                       23.000   Distribution:                  F(1,1121)
Min Obs:                       23.000
Max Obs:                       23.000   F-statistic (robust):             64.082
                                        P-value                           0.0000
Time periods:                      23   Distribution:                  F(1,1121)
Avg Obs:                       51.000
Min Obs:                       51.000
Max Obs:                       51.000

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          181.70     39.289     4.6248     0.0000      104.61      258.79
income         0.0234     0.0029     8.0052     0.0000      0.0177      0.0292
==============================================================================
```

Step 6: Results interpretation

1. What is value of $\beta_0 + \beta_1$ coefficients?

2. Are both $\beta_0 + \beta_1$ coefficients significant?

3. Based on the model results.
   What would be the crime rate (violent) for income
   i. income = 1,000 => violent = 205
   ii. income = 10,000 => violent = 415
   iii. income = 20,000 => violent = 650

1. What is value of $\beta_0 + \beta_1$ coefficients?

2. Are both $\beta_0 + \beta_1$ coefficients significant?

3. Based on the model results.
   What would be the crime rate (violent) for income
   i. income = 1,000 => violent = 205
   ii. income = 10,000 => violent = 415
   iii. income = 20,000 => violent = 650

What happens when we use wrong model?

Fixed Effects

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	181.70	27.101	6.7046	0.0000	128.53	234.88
income	0.0234	0.0020	11.933	0.0000	0.0196	0.0273

POLS

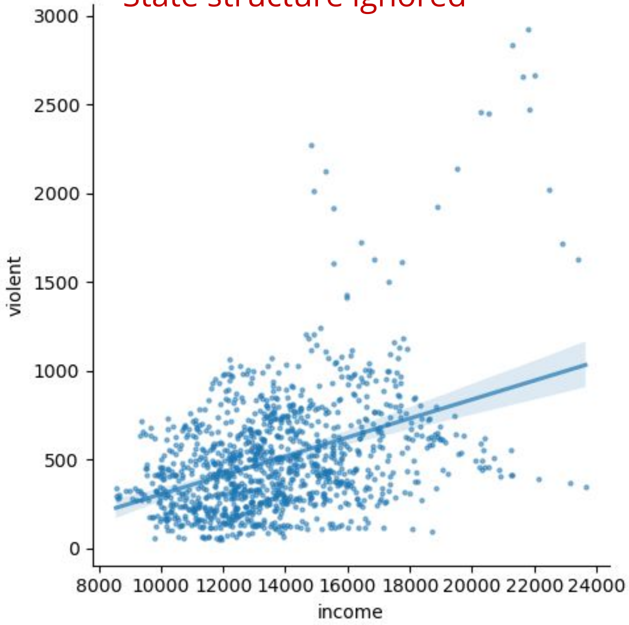
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-229.66	287.81	-0.7979	0.4251	-794.35	335.03
income	0.0534	0.0226	2.3657	0.0182	0.0091	0.0977

What would be the crime rate (violent) for income?

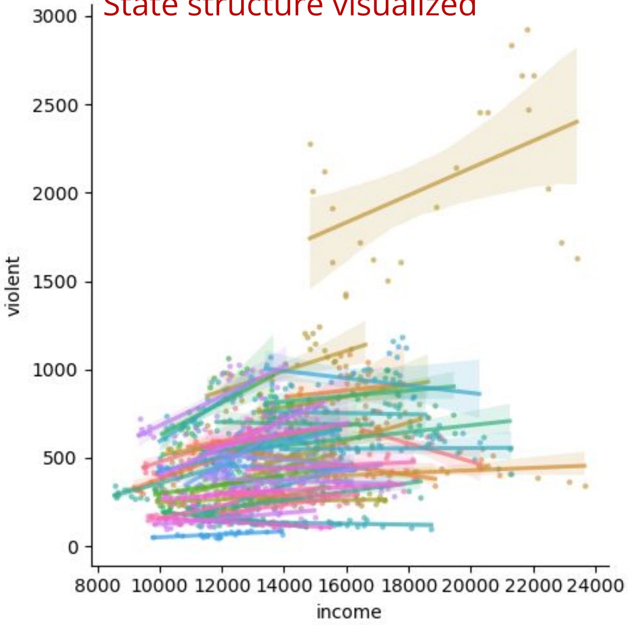
Income	Violent (Fixed effect)	Violent (POLS)
1,000	205	-176
10,000	415	304
20,000	650	838

What happens when we use wrong model?

State structure ignored



State structure visualized



czechitas

REAL WORLD APPLICATION:
MARKETING CHANNELS PROFITABILITY

Real world applications

Business questions - examples

- Does companies' investment in environmental sustainability have a positive impact on their profits?
- When interacting with our customers – Are some marketing channels more profitable than others?
- What are the variables impacting house prices? What is the impact of unemployment to house prices?
- ...

czechitas

When interacting with our customers – Are some marketing channels more profitable than others?

Which data to collect?
Which data aggregation to use?

How to validate correctness of the data with business?

Which model is the best?
What happens if we use wrong model?

How to interpret model results?

Are some marketing channels more profitable than others? Which data to collect

Marketing channels

Price of product

Price of marketing channels

Competitor information

Our product sales

Should we model emails:
- Sent
- Opened
- Clicked

Can interactions between channels bring additional benefit?

Should we model sales value or sales amount?

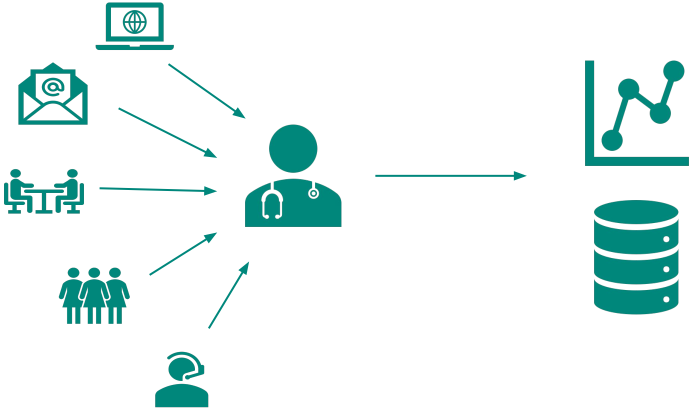
czechitas

How to aggregate data?

Ideal data

Data point for each physician

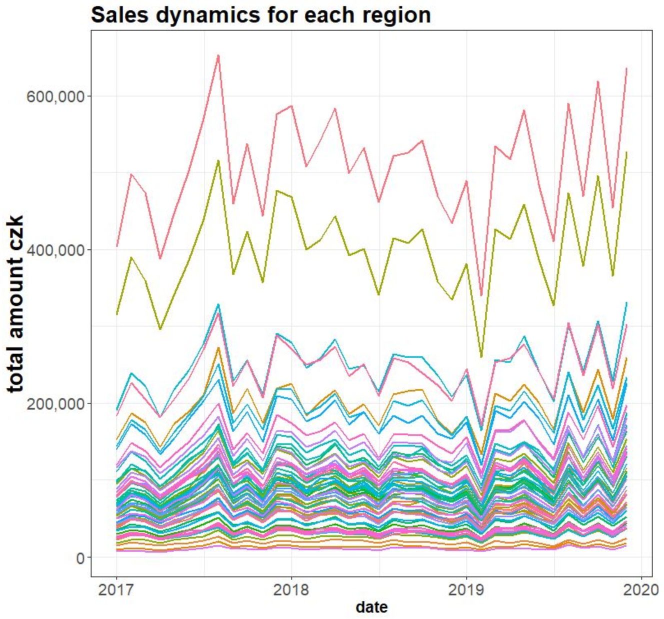
- Which marketing channels physician was reached
- How each physician prescribes our product
- Price
- Competitor marketing activities

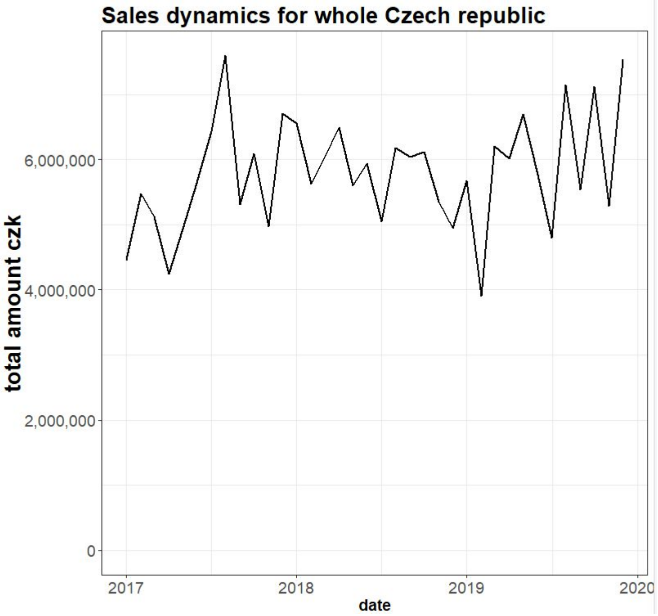


Due to confidentiality we have aggregated data

- **Time dimension** - monthly data, 3 years
- **Cross-sectional dimension** – 59 departments within Czech Republic
  - # of contacts from each marketing channel
  - price information
  - our sales
  - competitor sales

Data validation – visualize data and trends validate with business correctness of data





Discuss with business

- Trends
- Seasonality
- Outliers

Validate we model the right data

Validate there are no weird observations/trends and data errors

Generate questions and potentially create new variables

All data are illustrations

Panel data – what type of data aggregations we can have?

Panel
both time and region

Region	date	total_amount	events
01_region	1/1/2017	48029	0
01_region	2/1/2017	49577	0
01_region	....	....	....
01_region	11/1/2019	59072	0
01_region	12/1/2019	81379	0
02_region	1/1/2017	98558	242
02_region	2/1/2017	99100	295
02_region	....	....	....
02_region	11/1/2019	100528	464
02_region	12/1/2019	132593	0
....	....	....	....
59_region	1/1/2017	493057	1048
59_region	2/1/2017	496547	1125
....	....	....	....
59_region	11/1/2019	501323	586
59_region	12/1/2019	638767	92

59x36 observations

Time series
total Czech Republic per time

Region	date	total_amount	events
totalCzechRepublic	1/1/2017	5457472	5351
totalCzechRepublic	2/1/2017	5454035	4929
totalCzechRepublic	3/1/2017	4851672	3891
totalCzechRepublic	4/1/2017	5595244	3588
totalCzechRepublic	5/1/2017	5973403	3036
totalCzechRepublic	6/1/2017	4550540	4108
totalCzechRepublic	7/1/2017	5631377	656
totalCzechRepublic	8/1/2017	5632386	210
totalCzechRepublic	9/1/2017	4190666	2876
totalCzechRepublic	10/1/2017	5445516	2665
totalCzechRepublic	....	....	....
totalCzechRepublic	6/1/2019	6460270	324
totalCzechRepublic	7/1/2019	5422320	0
totalCzechRepublic	8/1/2019	6560794	0
totalCzechRepublic	9/1/2019	6889965	2603
totalCzechRepublic	10/1/2019	6814626	1329
totalCzechRepublic	11/1/2019	5827502	3513
totalCzechRepublic	12/1/2019	7560694	2175

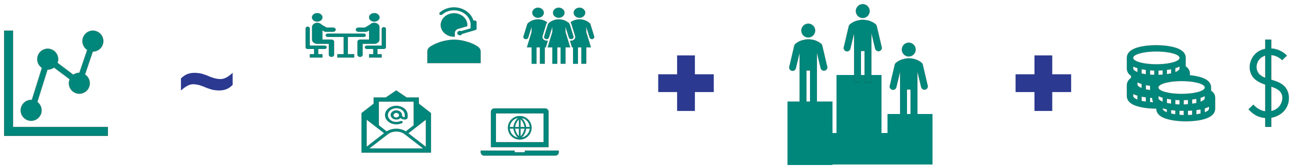
36 observations

Cross sectional
Aggregate to total regions, no time

Region	total_amount	events
01_region	2040773	214
02_region	3655331	4611
03_region	2411411	54
04_region	694182	197
05_region	484860	104
.....	.....	.....
53_region	1255011	1123
54_region	5575916	2376
55_region	1114933	837
56_region	3894204	2071
57_region	8689023	5949
58_region	1921434	1132
59_region	18188817	12849

59 observations

Model specification

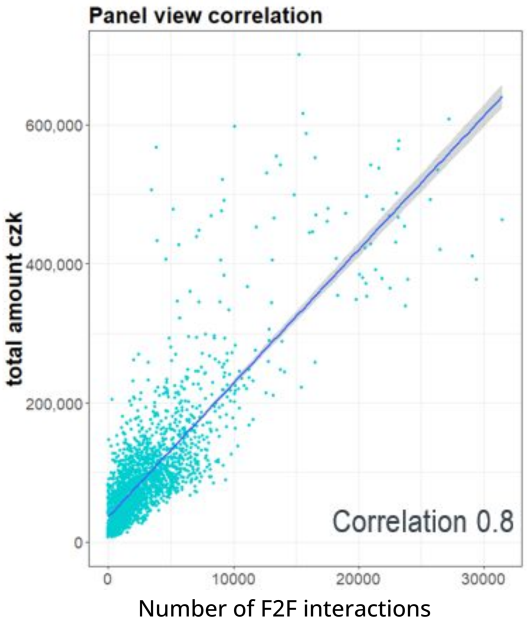


sales_amount ~
f(
number_F2F_interactions,
events,
emails,
telephone_meeting,
website_visits,
competitor_information,
seasonality
)

All data are illustrations

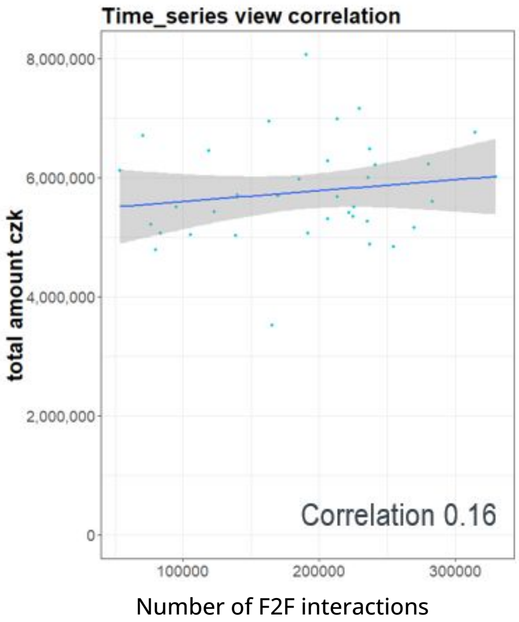
Panel data – which aggregation to choose? First visualize

Panel
both time and region



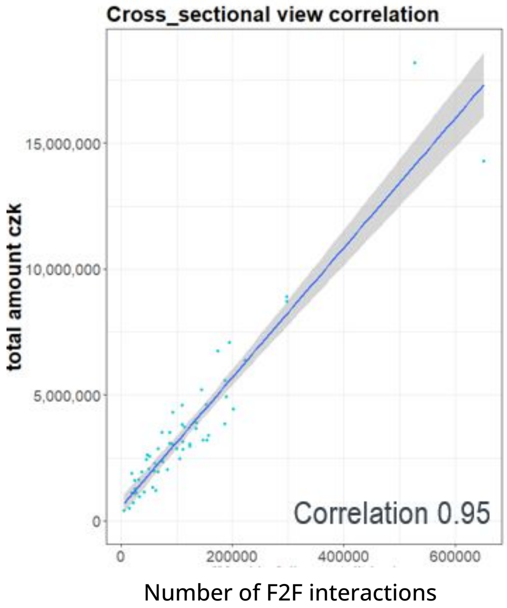
59x36 observations

Time series
total Czech Republic per time



36 observations

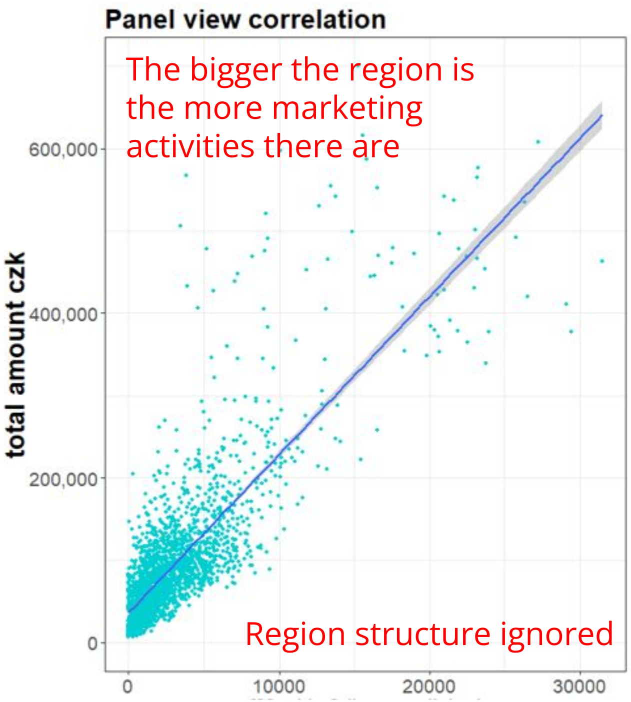
Cross sectional
Aggregate to total regions, no time



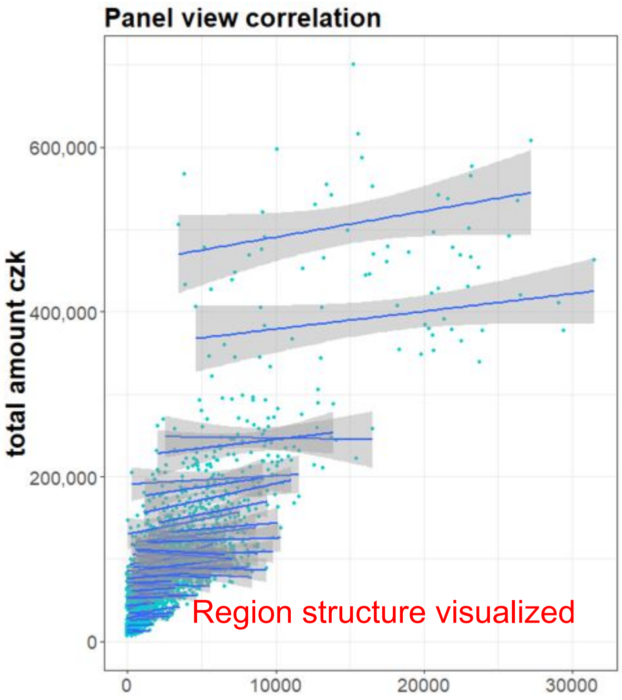
59 observations

czechitas

What happens if we use wrong model?



Number of F2F interactions



Number of F2F interactions

All data are illustrations

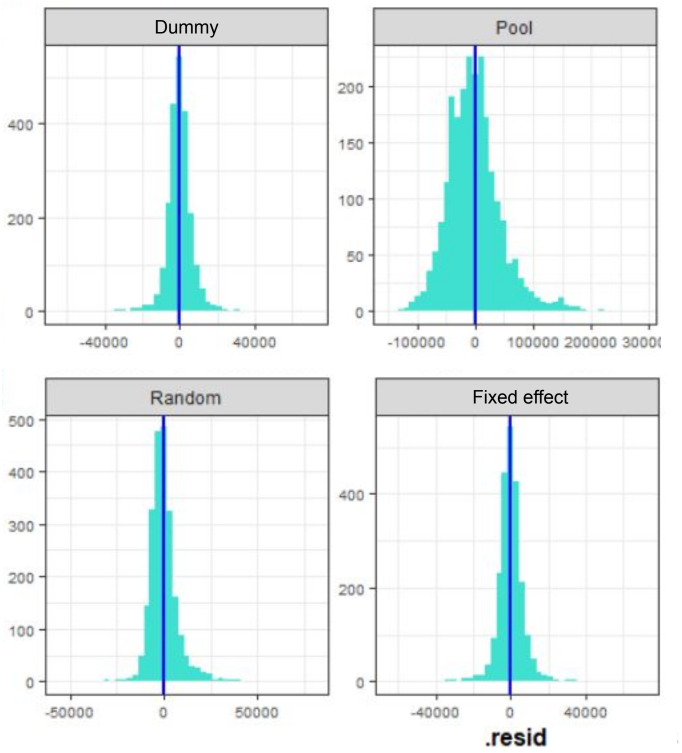
How to solve the problem and choose the best model?

- **Transform the data** (demean)
  + remove average from each region data for all variables (Fixed effects model)

- **Hausman test**

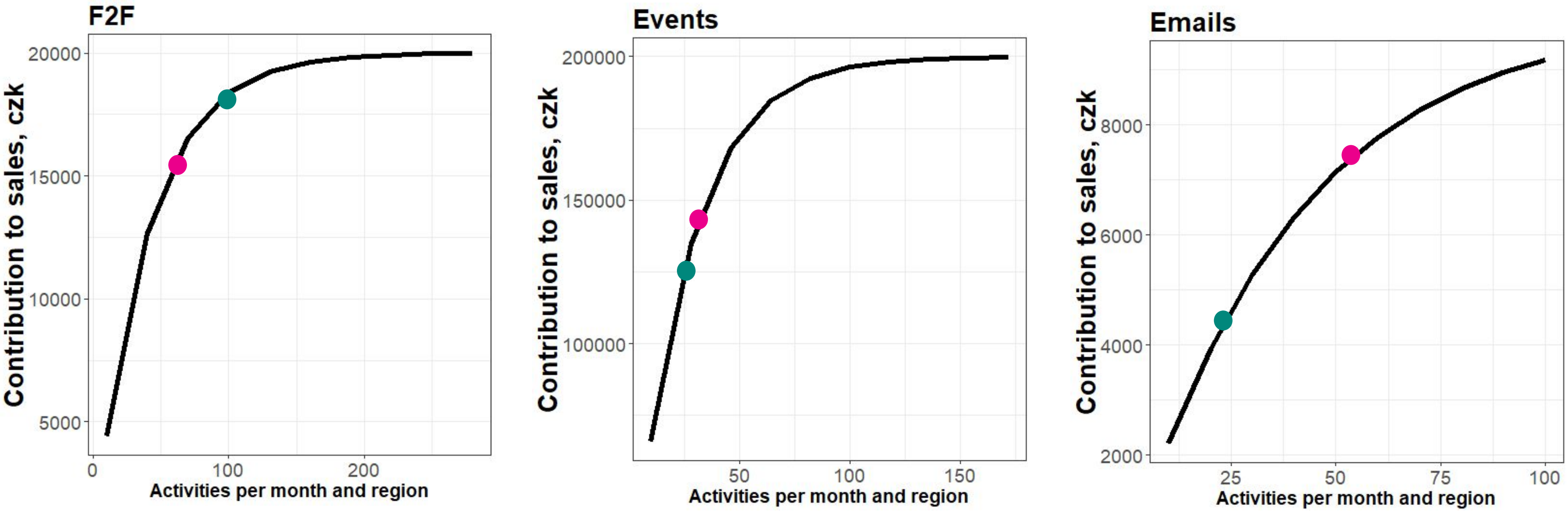
- **Check residuals**
  (e.g. Residuals sum of squares metric)
  The lower RSS the better fit

Histogram of residuals



All data are illustrations

How to use model results?



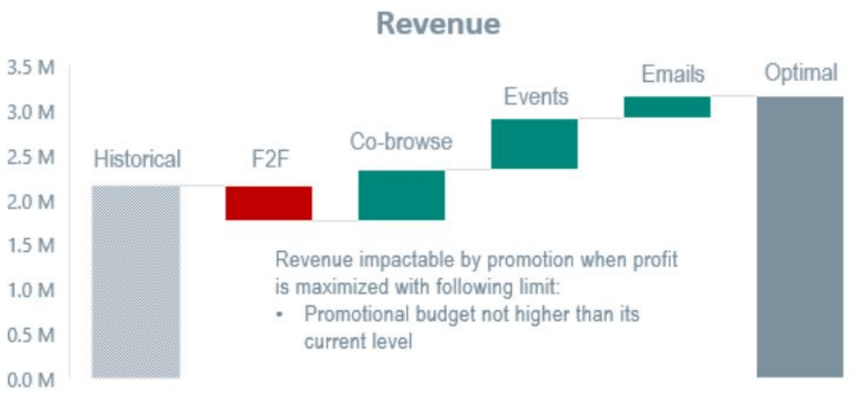
● Marketing mix optimal

● Historical

czechitas

Interpretation of results – Are some marketing channels more profitable than others?

- We know coefficient from the model results
- Using model coefficients, we can optimize the mix of marketing channels
  - Same budget
  - How we can change the marketing mix to reach higher sales?
  - Run optimization
    Costs are the same
    Revenues are higher



czechitas

Key takeaways

Benefits of panel analysis

- Allows to estimate impact of marketing to sales
- Allows to compare effectiveness of multiple marketing channels
- Allows to pick optimal mix of marketing activities

Things to remember

- Visualize data to validate it and to question it
- Data aggregation matters
- Watch out for the bigger the region the higher the sales variables effect
- Check your findings with business

Thank you for your attention!

SOURCES

- https://www.statisticshowto.com/durbin-watson-test-coefficient/
- https://www.youtube.com/watch?v=1SchyQ77VFg + many other videos from Ben Lambert
- Theoretical example with python code - https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8
- Guns dataset - https://vincentarelbundock.github.io/Rdatasets/datasets.html