Statistics
Tutorial 09

Philipp Scherer & Jens Wiederspohn

01.07.2020

Disclaimer!

- **The content of the slides partly relies on material of Philipp Prinz, a former Statistics tutor. Like us, he's just a student. Therefore we provide no guarantee for the content of the slides or other data/information of the tutorial.**
- **Please note that the slides will not cover the entire lecture content. To pass the exam, it is still absolutely necessary to deal with the Wooldridge in detail!**

Pooled Cross Sections vs. Simple Panel Data

Pooled Cross Sections

- Random Sample from a large population at different points in time.
- Important feature: They consist of independently sampled observations. → Its likely that independent samples lead to observations that are not identically distributed.

Simple Panel Data (or Longitudinal Data)

- we (try to) follow the SAME individuals (families, firms, cities, ...) across time
- Here, we cannot assume that the observations are independently distributed across time!

Pooled Cross Sections

- Draw random samples at each time period and pool them in a next step.

→ increases sample size, more efficient estimators and more powerful t-statistics (Is relation between Y und x_i constant over time?)

- Use year dummy variables to distinguish between time periods (with one year as base category)
  - $kids = \beta_0 + \beta_1 * educ + \beta_2 * age + \beta_3 * y98 + \beta_4 * y99$
  - MLRM, regressing kids onto education and age, using Pooled Cross-Sectional data for the years 1997 (base category), 1998 and 1999.
- We can also interact a year dummy variable with key explanatory variables to see if the effect of that variable has changed over a certain time period.
  - $kids = \beta_0 + \beta_1 * educ + \beta_2 * educ * y98 + \beta_3 * y98$
  - MLRM, regressing kids onto education, using Pooled Cross-Sectional data for the years 1997 (base category) and 1998 while searching for a change in the effect of educ on kids between 1997 and 1998.

Chow Test for Structural Change across Time

- Reminder: Chow-Tests are normally used to determine whether a multiple regression function differs across two groups.
- We can now use this to test, whether two different time periods differ as well from each other.
- Same procedure as usual:
  - $F((T-1)*k, (n-T-T*k)) = \frac{(SSR_r - SSR_{ur})/SSR_{ur}}{(n-T-TK)/((T-1)/k)}$
  - The pooled SSR is the sum of the SSRs for the two separately estimated time periods.
- Degrees of freedom calculation for Chow Test:
  - $df_1 = (T-1)k$, $df_2 = n - T - T_k$
  - Example: $kids = \beta_0 + \beta_1 * educ + \beta_2 * age + \beta_3 * y98 + \beta_4 * y99$, n=100
  - k=2, T=3
  - $df_1 = (3-1)*2 = 4$, $df_2 = n - T - T_k = 100 - 3 - 3*2 = 91$
- Despite Chow Test, DF calculation works as usual! $(df = n - k - 1)$

Policy-Analysis with Pooled Cross Sections

- Two or more independently sampled cross sections can be used to evaluate the impact of a certain event or policy change
  - What effect had the G8 implementation onto student´s performance
- It is necessary that we have data available, which was collected before and after the occurrence of an event, so we can determine an effect.
  - e.g. PISA Data before G8 was implemented
  - e.g. PISA Data after G8 was implemented
- Most important method here is the Difference in Difference (DiD) Estimation

Difference in Difference (DiD) Estimation

Example: Effect of new garbage incinerator on housing prices

- Examine the effect of the location of a house on its price before and after the garbage incinerator was built:
  - Before incinerator was built: $\hat{rprice} = 82.517,23 - 18.824,37 * nearinc$
  - After incinerator was built: $\hat{rprice} = 101.707,35 - 30.688,27 * nearinc$
- It would be wrong to conclude from the regression after the incinerator is there that being near the incinerator depresses prices so strongly
- One has to compare with the situation before the incinerator was built: In the given case, this is equivalent to
  - $\hat{\delta}_1 = -30.688,27 - (-18.824,37) = 11.863,9$

→ This is the so called difference-in-differences estimator (DiD)

→ Only "problem": We cannot obtain standard errors, using this approach!:(

Difference in Difference (DiD) Estimation cont´d

DiD for garbage incinerator in a regressional framework:

$$rprice = \beta_0 + \delta_0 after + \beta_1 nearinc + \boxed{\delta_1} after * nearinc$$

- δ_1 still representing the differential effect of being in the location and after the incinerator was built.
- But this time, standard errors for the DiD-effect can also be obtained!

Generalized forms of DiD:

$$y = \beta_0 + \delta_0 after + \beta_1 treated + \boxed{\delta_1} after * treated + otherfactors \quad (1)$$

$$\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{1,C}) - (\bar{y}_{0,T} - \bar{y}_{0,C}) \quad (2)$$

- Compare the difference in outcomes of the units that are affected by the policy change (= treatment group T) and those who are not affected (= control group C) before and after the policy was enacted

Two Period Panel Data

- One way to use panel data is to view the unobserved factors affecting the dependent variable as consisting of two types: those that are constant and those that vary over time
- Letting i denote the cross-sectional unit and t the time period, we can write a model with a single observed explanatory variable as $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$ with $t = 1, 2$
- $d2_t$: dummy variable that equals zero when $t = 1$ and one when $t = 2$
- a_i is called an unobserved effect or fixed effect and captures all unobserved, time-constant factors that affect $y_{i\,t}$
- $u_{i\,t}$ is often called the idiosyncratic error or time-varying error, because it represents unobserved factors that change over time and affect $y_{i\,t}$

First-Differenced Equation

- In most applications, the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables
- This is simple to allow: because a_i is constant over time, we can difference the data across the two years

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_{it} \tag{3}$$

- Δ denotes the change from $t = 1$ to $t = 2$
- a_i, does not appear because it has been "differenced away."
- The intercept δ_0 is here actually the change in the intercept from $t = 1$ to $t = 2$

Policy-Analysis with Two Period Panel Data

- Panel data sets are very useful for policy analysis and, in particular, program evaluation
- Similar to the natural experiment literature, with one important difference: the same cross-sectional units appear in each time period
- Let $prog_{it}$ be a program participation dummy variable. The simplest unobserved effects model is:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it} \tag{4}$$

- If program participation only occurred in the second period, then the OLS estimator of β_1 in the differenced equation has a very simple representation:

$$\hat{\beta}_1 = \Delta\bar{y}_A - \Delta\bar{y}_B \tag{5}$$

- This is the panel data version of the difference-in-differences estimator for two pooled cross sections

Differencing with More Than Two Time Periods

- We can also use differencing with more than two time periods
- Suppose we have N individuals and $T = 3$ time periods for each individual. A general fixed effects model is:

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it} \tag{6}$$

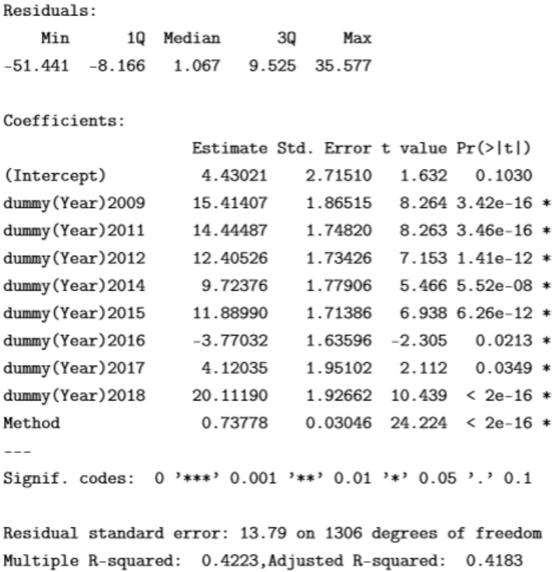
- The total number of observations is therefore 3N
- We can eliminate a_i by differencing adjacent periods. In the $t = 3$ case, we subtract time period one from time period two and time period two from time period three. This gives:

$$\Delta y_i = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + ... + \beta_k \Delta x_{itk} + \Delta u_{it} \tag{7}$$

- Unless the time intercepts are of direct interest it is better to estimate the first-differenced equation with an intercept and a single time-period dummy, usually for the third period:

$$\Delta y_i = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + ... + \beta_k \Delta x_{itk} + \Delta u_{it}, \text{ for t=2 and 3} \tag{8}$$

Exercise 1



```
Residuals:
    Min      1Q  Median      3Q     Max
-51.441  -8.166   1.067   9.525  35.577

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.43021    2.71510   1.632   0.1030
dummy(Year)2009    15.41407    1.86515   8.264 3.42e-16 ***
dummy(Year)2011    14.44487    1.74820   8.263 3.46e-16 ***
dummy(Year)2012    12.40526    1.73426   7.153 1.41e-12 ***
dummy(Year)2014     9.72376    1.77906   5.466 5.52e-08 ***
dummy(Year)2015    11.88990    1.71386   6.938 6.26e-12 ***
dummy(Year)2016    -3.77032    1.63596  -2.305   0.0213 *
dummy(Year)2017     4.12035    1.95102   2.112   0.0349 *
dummy(Year)2018    20.11190    1.92662  10.439  < 2e-16 ***
Method              0.73778    0.03046  24.224  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 13.79 on 1306 degrees of freedom
Multiple R-squared:  0.4223,Adjusted R-squared:  0.4183
```

Exercise 1a and b

How many students' results are used to estimate the parameters of Model 1?

- $n = df + k + 1 = 1306 + 9 + 1 = 316$
- Don´t be confused by the df calculation for the Chow Test here. Just use the df calculation as you know them from previous chapters (df=n-k-1)

What is/are true for the above model?

A Average results of "Statistics" can vary among years.

Exercise 1c and d

You do not include the dummy variable for Year 2008 to avoid a certain problem. Name the problem

- Dummy variable trap, or perfect collinearity, or exact linear relationships among the independent variables
- Keep in mind that you always need a base group when handling with dummy variables!

You split your dataset into each year. For each, you regress the same DV on the “Method” results and obtain the SSR. If you compare the sum of these SSRs with that of Model 1, which one is larger?

B The residual squared sum of Model 1.

→ This SSR is smaller , since we can explain a higher variation in y for the individual years through the individual models than with Model 1, which does not allow for a different effect of a better method score between the years.

Excercise 1e and f

If you consider all regression models of year specific datasets as one single model and call it Model 2, which statement(s) is/are true?

A Model 1 is a restricted model of Model 2.

→ Model 1 has no interaction effects

Denote the residual squared sum of Model 1 and the sum of the residual squared sum of Model 2 by SSR_1 **and** SSR_2**. Then you can obtain:** $\frac{\frac{SSR_1 - SSR_2}{SSR_2}}{\frac{8}{1298}} = 1.7687$**. What is the name of this statistic?**

- Chow statistic

Excercise 1g and h

Which distribution does the above statistic have?

- F-distribution
- Since the Chow-Test is simply a specific F-Test, it follows a F-Distribution with $df_1 = (T-1)k$, $df_2 = n - T - Tk$

What degrees of freedom does the distribution above have to test the above calculated statistic?

- T = 9, k = 1
- $df_1 = (9-1) * 1 = 8$
- $df_2 = 1316 - 9 - 9 * 1 = 1298$

Excercise 1i and j

You use the above statistic. . . (check all the correct statements; H_0 is null-hypothesis.)

D to test against H_0 that the effect of the results of "Method"-lecture is identical across year-specific datasets.

- Keep in mind that Chow-tests are a specific form of F-Tests. F-Tests always test, whether there is a significant difference between two or more models

The distribution of the above test has the percentiles below. Can you reject your H_0 at 10% significance level?)

- Search for suitable critical value in the below table → 1.675
- Search for F-Value, which was obtained in the description of 1f) → 1.7687

→ Since our $1.7687 > 1.675$, we can reject our H_0 at a 10 % significance level

Excercise 2a

Below you can find the electoral turnout in A and B. Estimate the interested effect by regressing individual citizens turnout (1 for turnout and 0 for abstention) on the dummy variable for absentee vote without claim. You should use the linear probability model and OLS estimation. You should just report the point estimate of the regression coefficient.

- Probability to vote in A 2020: $\frac{35000}{50000} = 0.7$
- Probability to vote in B 2020: $\frac{60000}{80000} = 0.75$
- $\hat{\beta}_1 = 0.7 - 0.75$
- $\hat{\beta}_1 = -0.05$

Excercise 2b

Calculate the point estimate for β_0, β_1, β_2 and β_3.

- Probability to vote in A 2014: $\frac{30000}{50000} = 0.6$
- Probability to vote in B 2014: $\frac{64000}{80000} = 0.8$
- $\hat{\beta}_0 =$ probability to vote in B in 2014$= 0.8$
- $\hat{\beta}_1 =$ change in the probability to vote from B to A in 2014 $= 0.6 - 0.8 = -0.2$
- $\hat{\beta}_2 =$ change in the probability to vote from B in 2004 to B in 2020 $= 0.75 - 0.8 = -0.05$
- $\hat{\beta}_3 =$difference in the change in the probability to vote between B in 2004 to B in 2020 and A in 2004 to A in 2020 $= (0.7 - 0.6) - (0.75 - 0.8) = 0.15$

Excercise 2c and d

What is the name of the estimator, which corresponds to β_3 in the above model?

- the difference-in-differences estimator

Which of the following statement(s) is(are) correct?

A a_i is often referred to as a fixed effect.

B a_i captures unobserved factors, which are constant across time.

C a_i captures unobserved heterogeneity among individual citizens.

D The sum of a_i and u_{it} is often referred to as the composite error.

- All statements are correct

Excercise 2e and f

If a_i and x_{it} in the above model are correlated, $\hat{\beta}_1$ is biased and inconsistent. To solve such potential problems, you can regress the change of y from 2014 to 2020 on the change of x at the same time period. Why can this model solve the problem? Describe the reason in one sentence.

- a_i are constant across time and disappears from the equation by taking difference between two time points.

Calculate the point estimate of β_1 of the model above.

- The value of $\hat{\beta}_1$ is the same as the value of $\hat{\beta}_3$ in the independent pooled cross sections model.
- $\hat{\beta}_1 = \Delta\bar{y}_A - \Delta\bar{y}_B$
- $\hat{\beta}_1 = (0.7 - 0.6) - (0.75 - 0.8)$
- $\hat{\beta}_1 = 0.15$