Contents

1 Introduction

1.1 General Idea

This course paper describes the empirical work which I performed for the course "Econometric Methods 2", constructing an own panel data set. The German Socio-Economic Panel (in the following: SOEP) is a longitudinal panel dataset of the population in Germany, conducted and supervised by the German Institute for Economic Research (DIW) . It is a household based study which began in 1984 and which reinterviews adult household members annually. The annual sample size consists of more than 12000 individuals, which answer the questionnaire on a voluntary basis. The large sample size and the close supervision of the DIW, as well as its representative features promise great utility for empirical analyses of German policies, descriptive studies or causal interpretation. The German Council of Science and Humanities classified the scientifc quality of the SOEP-Data as "excellent". In the following, I will describe the R-code which I constructed to work with the acquired SOEP-Data and how I dealt with specific data-related problems in order to estimate two Difference-in-Difference regressions based on two different indicators. I used the statistical programm "R" and aimed to construct the code as general as possible.

1.2 Introduction of the Minimum Wage in Germany

As my research interests focus around applied labour economics, the aim of this paper is to learn how to work with SOEP data, in order to estimate the effects of the introduction of the nationwide minimum wage in Germany in 2015.

Germany introduced a mandatory statuary minimum wage of 8.50 per working hour on 1 January 2015, following the coalition agreement of the Social Democratic Party (SPD) and the Christian Democratic Union (CDU/CSU). The introduction of a minimum wage was a central demand of the SPD during the election campaign und subsequently agreed upon. Prior to 1 January 2015, minimum wages were only specific to regions or industries and varied in the level of the wages. The primary goal of the introduction was to obtain good working conditions for everybody - safe and well-paid as stated by the Federal Ministry of Labour and Social Affairs.

With a level of 8.50 per hour, the German minimum wage is one of the highest in Europe, considering the ratio between minimum wage and median wage (Kaitz Index) in a variety of measurements [**?**].There are several groups that are excempted from the minimum wage regulations: Juvenilesyounger than 18 years old, trainees, interns, if their internship is not longer than three months ormandatory, volunteers and longtime unemployed persons.

The impact of minimum wages on employment is one of the most throroughly researched topicin labour economics, although there is little consensus about the effect of minimum wages onemployment. While some meta-analyses hint that there is either no discernible or only a small causal effect of minimum wages on employment, e.g. [Schmitt et al., 2013] or [Neumark et al., 2014]state on the other side of the spectrum that minimum wages pose a tradeoff between higher wagesfor the employed vs. job losses for others.

2 Data Handling

2.1 Reading in SOEP-Data

After obtaining the SOEP dataset from the DIW, I chose to use the core datasets, which consist of different datasets for each year(wave) of the panel instead of the cleand and complete long-panel set, as my motivation is to cope with the courses' aim of dealing with messy and uncleaned data. In the core datasets, there are around ten different types of questionnaires and datasets for each year, e.g. personal, regional or household data. The specific year of the datasets are always specified by the first or first and second letter: The datafiles for the first year of the SOEP (1984) begin with an "a", while the most recent datafiles (2015) start with the letters "bf". The following letters specify the type of questionnaire. Additionally, one has to face the problem of changing variable names. For almost every variable, the names differ for each year of observation. As an example, the variable of interest "registered unemployed" is labeled as "bap06" in 2010, "bdp08" in 2012 and "bfp15" in 2015. This complicates the coding, as a normal command would not allow for changes in the amount of years or selection of variables at a later stage. As a timeframe, I chose the years 2010 until 2015 to have a enough pre-intervention periods.
In order to deal with this problem, I specify an excel sheet ("soep-feature-selection3.csv") with all variables of interest and their respective variable names for the years 2010 until 2015. Using this sheet, I construct a loop to identify the specific variables in the excel sheet, read them in seperate datafiles in R and merge them into one panel. I then specify an option, if the resulting panel has to be balanced or not, which can depend on the question of research. In my case of estimating the effect of a minimum wage, I allow for some variables to have missing values for some years in order to attain the most information. In the case of the variable "Speaking German" for example, it is possible to only have it every second year, as it can be stated that the language will stay constant in this timeframe.
The Code performs two loops: One over the personal datasets (ending with "p") for most variables and one over the regional datasets (ending with "equiv") to obtain the information about the residence of the individuals. In this way, it is possible to extend the amount of variables or datasets with ease, if needed for future research. The Full Data panel is then saved as a csv. - file.

2.2 Data-Handling and Aggregation

In the next step, I deal with the SOEP Data structure. All categorical values which indicate missing values, insufficient information or implausible values are changed to NAs for all existing variables. Second, I construct several (0,1) Dummy-Variables for the categorical variables "good Health", "Gender" (1 being female), "being unemployed" (1 being unemployed) and "speaking German" (1 being above medium level, 0 below medium level).

Second, I construct two variables that are not reported in the SOEP itself, namely the age of individuals and the hourly wage. The latter will serve as an indicator of treatment and is therefore crucial for the analysis. Therefore, I construct two different measurements. The first is based on the reported contracted amount of working hours and the reported monthly gross income , while the second measurement uses the reported actual working hours.

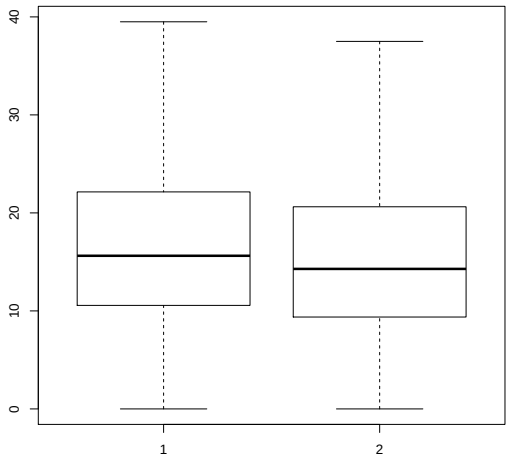

Figure 1: Distribution of both measurements of the hourly wage

Figure 1 shows the distribution of both constructed hourly wages. As both exhibit a similar distribution and median (€15.62 and €14.29 respectively) I decide to continue with the wage that is constructed on the actual reported hours as it has a higher sample size (around 11000 Observations).

As the aim of this paper is a Difference-in-Difference estimation, aggregated regional data is needed. I therefore construct an aggregate data panel on the level of all 16 German Federal States and include the constructed dummy variables as covariates alongside the variables, which now serve as the corresponding percentages. As an indicator of treatment I then calculate the Kaitz-Index, as the ratio between mean wage in each State and the minimum wage of €8,50. Finally, the external Google-Trend data for the word "Mindestlohn" (German for Minimum Wage) is looped in and merged with the State-Panel data. Both Indices are described in the next section. During the merge, I encountered the problem of different spellings of state names in both data sets.

3 Regression Analysis

In the final part of the paper, I estimate the impact of the introduction of the German Minimum Wage in 2015 on employment, by two multiple Difference in Difference regressions. However, as the introduction was mandatory and statuary, it is not possible to find perfect control states within Germany. Instead I use the fact that minimum wages generally have higher effect on certain states than others. (See e.g. [Dube et al., 2010] or [Machin et al., 2003]). As indicators for these "bites", I use the aforementioned Kaitz Index and the Google Trend Indicator. The Kaitz Index is calculated as the ratio of average State specific hourly wages and the mandatory minimum wage of €8,50. The Google Trend Indicator is a measurement of how often the term "Mindestlohn" has been searched in each State in relation to the other states. Specifically, I construct the average of both relative search intensities for the years 2015 and 2014. I include the year 2014 to allow for anticipation effects. One can state that this relative search intensity can perform as a proxy for the effect of a minimum wage on unemployment (see e.g. [Askitas and Zimmermann, 2009]).

	Federal.State	Bite	avg.google
47	Mecklenburg	1.65	100.00
83	Thuringia	1.73	85.00
71	Saxony	1.76	95.00
23	Brandenburg	1.79	88.50
65	Saxon-Anhalt	1.79	89.00
59	Saarland	1.92	52.00
29	Bremen	1.93	52.00
77	Schleswig-Holstein	2.00	56.50
41	Lower-Saxon	2.05	55.50
53	North-Rhine-Westphalia	2.13	46.50
17	Berlin	2.15	59.00
11	Bavaria	2.23	49.00
35	Hamburg	2.33	54.50
5	Baden-Wuerttenberg	2.35	42.00

Table 1: Kaitz Index in 2014 (Bite) and Google - Index for all 16 German States

Table 1 displays the values for the two indices for all 16 States. It can be seen that most states which display a high Kaitz - Index have a higher relative search index as well, but some variation between both remains.

In order to gauge the causal impact of the introduction of minimum wages on unemployment, I propose several procedures. First, it is possible to seperate all states strictly into treated and non-treated, based on the rank of indices. I define the treatment group as the States with the five lowest Kaitz Index or five highest average Google index respectively. In the same fashion, i define the treatment group as the five States with the highest Kaitz Index or lowest Google Index respectively. It is then possible to estimate the effect using a Difference-in-Difference Estimator of the form

$$Y_{ist} = \beta_0 = \beta_1 \cdot d.Year + \beta_2 \cdot d.Index + \beta_3 \left(T_t \cdot S_s\right) + \epsilon_{ist}$$

where $d.Year$ is a dummy indicating that the time period is equal or after 2015, $d.Index$ is a dummy indicating that the State belongs in the treatment group, $(T_t \cdot S_s)$ an interaction term indicating treatment and ϵ_{ist} an error term containing different covariates. The specific threshold is arbitrary and other specifications are possible. However, I specified it such that each control and treatment group consists of five states.

3.1 DiD Regression with Kaitz Index and Google Index

Table 2: Results - DiD - Kaitz-Index

	Dependent variable:
	unemp.perc
d.kaitz	0.067***
	(0.007)
d.Year	0.014
	(0.010)
did.kaitz	-0.028
	(0.020)
Constant	0.060***
	(0.005)
Observations	53
R^2	0.659
Adjusted R^2	0.638
Residual Std. Error	0.024 (df = 49)
F Statistic	31.526*** (df = 3; 49)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Results - DiD - Kaitz-Index

	Dependent variable:
	unemp.perc
d.google	-0.006
	(0.011)
d.Year	0.004
	(0.020)
did.google	-0.019
	(0.025)
Constant	0.102***
	(0.007)
Observations	60
R^2	0.040
Adjusted R^2	-0.011
Residual Std. Error	0.037 (df = 56)
F Statistic	0.780 (df = 3; 56)
Note:	*p<0.1; **p<0.05; ***p<0.01

As can be seen in Tables 2 and 3, the estimated causal effect of the introduction of the minimum wage in Germany in 2015 is negative but not statistically significant. In this case, a negative beta would mean an decrease in unemployment. A possible explanation for this would be that the minimum wage has no statistical causal effect on unemployment, a result often found in research [Dube et al., 2010]. However, the coefficient of the state dummy is positive and statistical significant, which suggests that states with a low kaitz index display a higher level of unemployment. However, another explanation could be that it is not possible to strictly seperate the States into treatment and control group as they possess not the same characteristics. This would then violate the essential common trend hypothesis, which is required for an analysis using Difference - in Differences. The tables below looks at these differences and compares some important characteristics between both the treatment and the control group for the years 2014 and 2015 with the Kaitz Index as treatment indicator. Interestingly, the wage declines in the control group but decreases in the treatment group. At the same time, the asking wage if unemployed increases in the same proportion, as would be expected in the case of a minimum wage.
The same procedure applies to the Difference - in - Difference Estimation using the Google Index.

	5	11	6	12
Treatment	0.00	1.00	0.00	1.00
Year	2014.00	2014.00	2015.00	2015.00
wage.mean	15.98	18.50	16.60	18.26
female.perc	0.55	0.54	0.55	0.54
good.health.perc	0.73	0.75	0.74	0.74
unemp.perc	0.11	0.06	0.11	0.08
mean.age	50.15	48.81	48.60	48.60
mean.askingwage	1247.25	1258.78	1321.36	1355.31
mean.GermanL			0.91	0.88

	5	11	6	12
Treatment	0.00	1.00	0.00	1.00
Year	2014.00	2014.00	2015.00	2015.00
wage.mean	16.07	18.50	16.60	18.26
female.perc	0.55	0.54	0.55	0.54
good.health.perc	0.71	0.75	0.74	0.74
unemp.perc	0.12	0.06	0.11	0.08
mean.age	50.22	48.81	48.60	48.60
mean.askingwage	1268.64	1258.78	1321.36	1355.31
mean.GermanL			0.91	0.88

4 Conclusion

I constructed a fully scalable and reusable R- code to read in the data from the GSOEP, create a set of dummies on the variables and perform Regressions on them. The main diffulties I encountered were the "messiness" of the SOEP data, which uses different variable years in each year and does not ask all variables throughout the year, as well as different SOEP-datasets which I needed for

this analysis. However, these problems can be neglected using the SOEP-long dataset instead of the core datasets. In any case, I consider it helpful to now know also how to work with the latter. The main focus of this paper lies in the cleaning aspect of the SOEP -data. In the future, or as a second and third step, it would be beneficial to apply some fixed-effects regressions, using both the Kaitz-Index and the Google Index directly. Another point of future discussion is a more careful selection of covariates, which was not possible due to time constraints. It might also be beneficial to expand the sample size of the regions by analyzing the effect on a county-level.

References

Askitas, N. and Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2):107–120.

Dube, A., Lester, T. W., and Reich, M. (2010). Minimum wage effects across state borders: Estimates using contiguous counties. *The review of economics and statistics*, 92(4):945–964.

Machin, S., Manning, A., and Rahman, L. (2003). Where the minimum wage bites hard: Introduction of minimum wages to a low wage sector. *Journal of the European Economic Association*, 1(1):154–180.