University of St.Gallen

School of Management, Economics, Law, Social Sciences and International Affairs

Data Analytics II: PC4

University of St. Gallen

Jonas Husmann | 16-610-917

Niklas Leander Kampe | 16-611-618

Prof. Dr. Michael Lechner

March 21, 2022

Part 1: Descriptive Statistics

a. Load Data Set

storeid	year	state	southj	centralj	northj	pa1	pa2	chain	co_owned	hrsopen	price	fte	wage_st
1	92	1	0	1	0	0	0	1	0	16.0	2.61	35.0	4.50
1	93	1	0	1	0	0	0	1	0	16.0	2.74	44.0	5.05
2	92	1	0	1	0	0	0	1	0	14.0	2.93	16.0	4.75
2	93	1	0	1	0	0	0	1	0	15.0	3.00	15.5	5.25
3	92	1	0	1	0	0	0	2	0	10.0	5.10	15.5	4.25

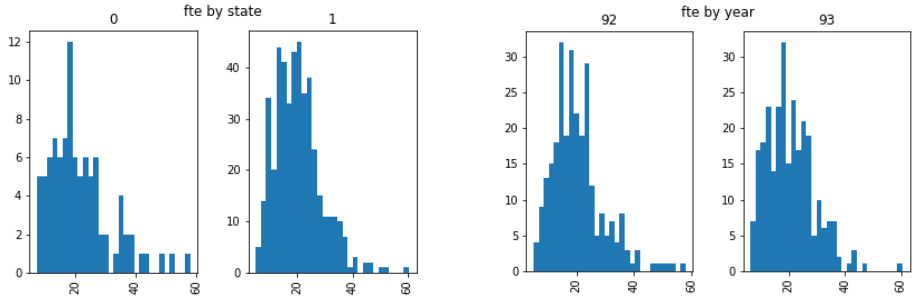
b. Summary Statistics + Anomalies or Missing/Implausible Values

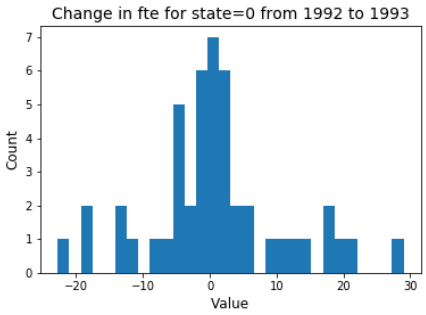
Descriptive Statistics:

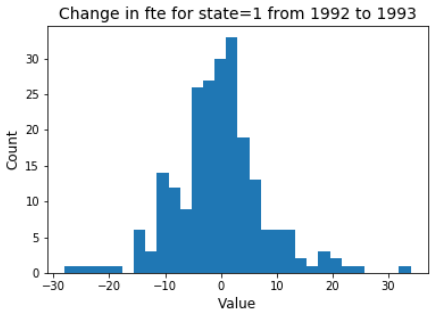
	mean	var	std	max	min	na	unique
storeid	244.02	21785.74	147.60	522.00	1.00	0.0	273.0
year	92.50	0.25	0.50	93.00	92.00	0.0	2.0
state	0.83	0.14	0.38	1.00	0.00	0.0	2.0
southj	0.23	0.18	0.42	1.00	0.00	0.0	2.0
centralj	0.15	0.13	0.36	1.00	0.00	0.0	2.0
northj	0.44	0.25	0.50	1.00	0.00	0.0	2.0
pa1	0.07	0.07	0.26	1.00	0.00	0.0	2.0
pa2	0.10	0.09	0.30	1.00	0.00	0.0	2.0
chain	2.06	1.12	1.06	4.00	1.00	0.0	4.0
co_owned	0.35	0.23	0.48	1.00	0.00	0.0	2.0
hrsopen	14.29	7.23	2.69	24.00	7.00	0.0	24.0
price	3.37	0.42	0.65	5.86	2.39	0.0	192.0
fte	20.48	74.97	8.66	60.50	5.00	0.0	99.0
wage_st	4.81	0.12	0.35	5.75	4.25	0.0	36.0

From the descriptive statistics above, it is observable that the dummy variables “state”, “southj”, “centralj”, “northj”, “pa1”, “pa2” and “co-owned” are coded correctly, which can be seen from the min-value of 0, max-value of 1 and the number of unique values of 2. In addition, the variable “year” is coded on the values (19)92 and (19)93 correctly with two unique values. Furthermore, the categorical variable “chain” is also coded correctly from the min-value of 0, max-value of 4 and the number of unique values of 4. Nevertheless, it is important to note that “chain” could lead to issues in an econometric analysis and its conclusions due to the mapping from 1 to 4. Hence, the mean, variance and standard deviation need to be regarded with attention due to the variable’s nature. Hence, it can be useful to transform it into a dummy variable for further unbiased analysis, which will be done at a later stage. Furthermore, the statistics from the identifier variable “storeid” shows that the sequence of identifiers is not “linear” with equal steps of 1 until the maximum value of 522. After a detailed look at the raw data set, it is observable that, while some identifiers are missing, each restaurant still has strictly two observations for both years. Hence, the identifier variable does not have any anomalies. Furthermore, all other (continuous) variables do not seem to have any implausible values when comparing their summary statistics with their variable descriptions. Another potentially problematic variable could be “hrsopen”, as the restaurants have a large variance in their opening hours, which range from 7 to 24 hours. As the analysis is focusing on the impact on employment, “fte” and “hrsopen” are likely to have a high correlation and hence a high signifance in further analysis, which could lead to miss-conclusions on the analysis results. Nevertheless, according to the descriptive statistics, no major anomalies or missing values can be detected overall.

c. Histograms + Distribution and Outliers







The first four plots in the first column show the distribution of full-time employments for each state (0 = Pennsylvania, 1 = New Jersey) and each year (1992, 1993). All distributions show a slight negative skewness which indicates more "mass" in smaller values of full-time employments around mean values of around 20 FTEs. Furthermore, all distributions are bounded at around 60 FTEs, whereby an approximate threshold for outlier detection could be reasonable at values of over 40, which would also efficiently reduce the negative skewness while not omitting too many observations. The last two plots in the second row show the difference in full-time employments in restaurants for each state (0 = Pennsylvania, 1 = New Jersey) between the years 1992 and 1993. The differences in FTEs have distributions without major skewness, bounded between -30 and +30. In order to omit potential outliers, a threshold of -20 and +20 could be applied. Nevertheless, it can be overall observed that restaurants based in Pennsylvania are majorly underrepresented in the data set compared to restaurants based in New Jersey, which can have an impact on further analysis as the data set is meant for analysis on policy changes of minimum wage increases which are effective in New Jersey but not in Pennsylvania. Hence, the analysis is also fundamentally based on the location of the restaurants.

d. Dummy Variable Checks

Dummy Variable Check:

	Share	Sum
southj	0.2832	0.2832
centralj	0.1814	0.4646
northj	0.5354	1.0

Dummy Variable Check:

	Share	Sum
pa1	0.4255	0.4255
pa2	0.5745	1.0

In order to check if the regional dummy variables for each state are coded correctly, the share of each region is calculated based on the number of observations that are indeed based in the underlying city and summed up over the respective of regional variables. From the table above, the regional variables for each city do sum up to 1, which indicates that the dummies are coded correctly without any anomalies.

e. Means and Numbers of Observations

Means & Observations:

		fte		wage_st		hrsopen		price	
		mean	count	mean	count	mean	count	mean	count
year	state								
92	0	22.0106	47	4.7043	47	14.3936	47	3.1409	47
	1	20.0387	226	4.6093	226	14.2389	226	3.3842	226
93	0	21.2394	47	4.6077	47	14.5957	47	3.1066	47
	1	20.4425	226	5.0788	226	14.2622	226	3.4478	226

According to the table with mean values and the number of observations for each year and state, the mean values of the target variable "fte" decreased from 22.0106 to 21.2394 in Pennsylvania and increased from 20.0387 to 20.4425 in New Jersey from 1992 to 1993. Furthermore, the starting wage per hour in New Jersey increased from 4.6093 to 5.0788, which reflects the policy change in the minimum wage increase in 1992, while Pennsylvania's starting wage stayed almost constant. Furthermore, the opening hours in New Jersey as well as in Pennsylvania slightly

increased. Lastly, the price for a full meal decreased in Pennsylvania while increasing in New Jersey, which could also be influenced by higher costs due to the minimum wage increase. Nevertheless, the imbalance between the observations in New Jersey and Pennsylvania is still evident, and the total observations size is also not that high, which could lead to miss-leading conclusions in further analysis as the restaurant markets in both cities might not be perfectly reflected by the data sample.

f. Dummy Variable Recoding

	storeid	year	state	southj	centralj	northj	pa1	pa2	co_owned	hrsopen	price	fte	wage_st	Burgerking	KFC	Royrogers	Wendys
0	1	0	1	0	1	0	0	0	0	16.0	2.61	35.0	4.50	1	0	0	0
1	1	1	1	0	1	0	0	0	0	16.0	2.74	44.0	5.05	1	0	0	0
2	2	0	1	0	1	0	0	0	0	14.0	2.93	16.0	4.75	1	0	0	0
3	2	1	1	0	1	0	0	0	0	15.0	3.00	15.5	5.25	1	0	0	0
4	3	0	1	0	1	0	0	0	0	10.0	5.10	15.5	4.25	0	1	0	0

Part 2: Difference-in-Differences

a. Estimation

```
ATE Estimate by Difference in Means:
----------------------------------------
Dependent Variable: fte
----------------------------------------
           ATE    SE  tValue  pValue
MeanDiff -0.8  1.39   -0.57    0.57
----------------------------------------
```

The ATE estimate by difference in means shows that the full-time equivalent employment in fast food restaurants has decreased with the increase in the minimum wage by -0.8. However, this outcome needs to be considered with caution as the pValue is very high at 0.57 and therefore, the result is not statistically significant. There are various biases that can influence this outcome, such as reverse causality, omitted variable bias, and mean regression. Exchangeability, positivity, and SUTVA must hold so that our estimate can reliably show a causal effect. It is likely that we have omitted variable bias to some degree here as a multitude of factors influence the labor market and its dynamics. Moreover, SUTVA is likely also violated as a minimum wage increase in the treatment group very likely has spillover effects to the control group.

b. Estimation

```
ATE Estimate by Difference in Means:
----------------------------------------
Dependent Variable: fte
----------------------------------------
           ATE    SE  tValue  pValue
MeanDiff  0.4  0.78    0.51    0.61
----------------------------------------
```

The ATE estimate by difference in means shows that the full-time equivalent employment in fast food restaurants in New Jersey before and after the policy change has increased with the minimum wage. The estimated effect of 0.4, however, should be treated with caution given the result is not statistically significant with a very high pValue of 0.61. As mentioned in exercise 2a) Exchangeability, positivity, and SUTVA must hold. Moreover, for the difference in difference approach to have a reliable outcome there needs to be a same trend for both treatment and control groups in outcome. This can be violated here and therefore our estimate is likely biased.