



University of St.Gallen

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

Data Analytics II: PC1

University of St.Gallen

Jonas Husmann | 16-610-917

Niklas Leander Kampe | 16-611-618

Prof. Dr. Michael Lechner

February 28, 2022

Part 1: Data Preparation

a. Load Data Set

	treat	age	ed	black	hisp	married	nodeg	re74	re75	re78	age2
0	1	19	9	0	0	0	1	0.00	0.00	13188.83	361
1	1	21	12	1	0	0	0	0.00	0.00	9983.78	441
2	1	25	12	1	0	0	0	14426.79	2409.27	0.00	625
3	0	21	7	1	0	0	1	33799.95	0.00	11011.57	441
4	0	22	10	0	0	0	1	27864.36	10598.67	7094.92	484

b. Descriptive Statistics

	treat	age	ed	black	hisp	married	nodeg	re74	re75	re78	age2
count	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00
mean	0.41	25.40	10.18	0.84	0.08	0.18	0.79	2191.46	1405.73	5372.86	694.37
std	0.49	7.01	1.80	0.37	0.28	0.38	0.41	5558.92	3249.13	6732.55	425.19
min	0.00	17.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	289.00
25%	0.00	20.00	9.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	400.00
50%	0.00	24.00	10.00	1.00	0.00	0.00	1.00	0.00	0.00	3791.21	576.00
75%	1.00	28.00	11.00	1.00	0.00	0.00	1.00	832.86	1225.53	8137.01	784.00
max	1.00	55.00	16.00	1.00	1.00	1.00	1.00	39570.68	25142.24	60307.93	3025.00

c. Adjusted Summary Statistics + Missing/Implausible Values?

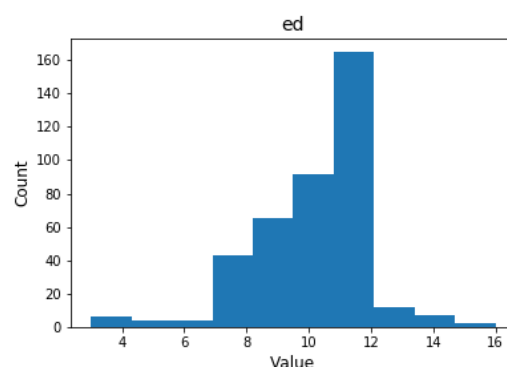
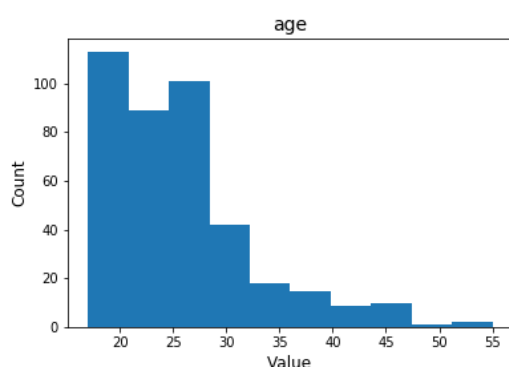
	treat	age	ed	black	hisp	married	nodeg	re74	re75	re78	age2
Mean	0.41	25.40	10.18	0.84	0.08	0.18	0.79	2191.46	1405.73	5372.86	694.37
Var	0.24	49.21	3.23	0.13	0.08	0.15	0.17	30901538.45	10556840.39	45327201.42	180784.83
Std	0.49	7.01	1.80	0.37	0.28	0.38	0.41	5558.92	3249.13	6732.55	425.19
Max	1.00	55.00	16.00	1.00	1.00	1.00	1.00	39570.68	25142.24	60307.93	3025.00
Min	0.00	17.00	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	289.00
Missing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Unique	2.00	33.00	14.00	2.00	2.00	2.00	2.00	105.00	140.00	281.00	33.00
Obs	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00	400.00

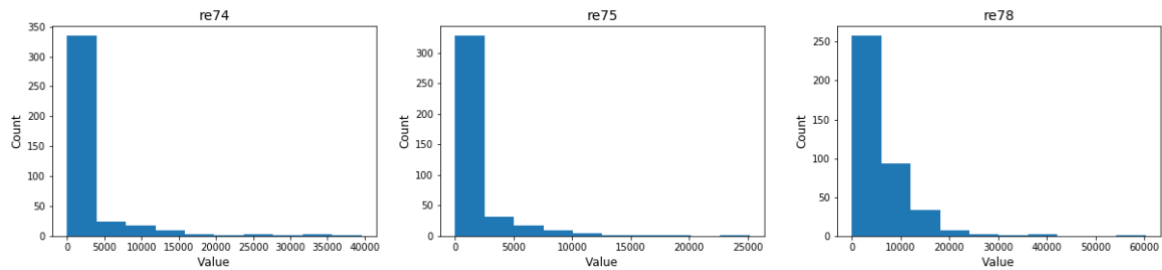
According to the summary statistics, the data set does not contain any missing values, which can be seen from the missing values equal to zero and that total observations sum up to 400 for all included variables. Furthermore, the dummy variables are all correctly coded. This can be concluded from the min value showing 0, the max value showing 1 and the number of unique observations showing 2 for all dummy variables. With a focus on the continuous variables, the “age” shows plausible values ranging from 17 to 55. Additionally, the annual earnings in 1974, 1975 and 1978 also show plausibility while ranging from 0-39’570.68, 0-25’142.24, and 0-60’307.93, respectively. Lastly, the mean, variance and standard distribution seem plausible for all variables, which finally indicates full plausibility across the data set.

d. Column Drop of “age2”

	treat	age	ed	black	hisp	married	nodeg	re74	re75	re78
0	1	19	9	0	0	0	1	0.00	0.00	13188.83
1	1	21	12	1	0	0	0	0.00	0.00	9983.78
2	1	25	12	1	0	0	0	14426.79	2409.27	0.00
3	0	21	7	1	0	0	1	33799.95	0.00	11011.57
4	0	22	10	0	0	0	1	27864.36	10598.67	7094.92

e. Histograms of Continuous Variables





According to the histogram of the variable “age”, the sample is positively skewed which leads to a higher proportion of younger compared to older people. Based on the histogram of the variable “ed”, the education in years in the sample is slightly negatively skewed and suggests that the majority of people in the sample has not attended any further education (like university) after high school. Lastly, according to the histograms for “re74”, “re75” and “re78”, the earnings in the years 1974, 1975 and 1978 in the sample are strongly positively skewed, which underlines traditional earnings distributions in history. While the majority has earnings approximately between 0 and 15'000 USD, only a few people in the sample earned more, up to 35'000, 25'000 and 60'000, respectively. Such a strong skewness can lead to major issues in analytics processes as outlying observations can have a strong effect on the analytics target, e.g., predictors and its coefficients. In order to prevent such effects, either statistical transformations could be run (e.g., log-transformations) or an outlier threshold could be defined to eliminate outlying observations based on the earnings.

The proportions of the treatment and control group in the data set can be concluded from the dummy variable “treat”. The mean value of 0.41 shows that 41% of the people in the sample were assignment in the treatment group, while 59% of the people in the sample were used as the control group. Hence, the treatment assignment was not fully balanced for this experiment.

f. Data Anomalies + Potential Issues for Statistical Analysis

According to the comparison of the distribution of the ages and earnings in the sample, the positive and negative skewness are in accordance. Due to a higher proportion of younger people, it is plausible that the earnings have a higher proportion in lower earnings and that the positive skewness becomes less over the three years as more people in the sample are either eligible to work or have more experience which could lead to higher earnings. Furthermore, the small proportion of people with higher degree educations (beyond high school) is also in accordance with the small proportion of high earnings in the sample, as higher education degrees can be associated with higher earnings. Hence, as a conclusion, no data anomalies can be detected in the data set.

Apart from that, the skewness in the variables still determines a major issue, why the data in its raw form cannot be used for statistical analysis. As mentioned in part e), the skewness needs to be corrected by either a data transformation (e.g., log-transformation) or by defining threshold for outliers and their elimination from the data set. These distributions can lead to biases in analytics processes as outlying observations can have a strong effect on the analytics target, e.g., predictors and its coefficients. Furthermore, the degree of randomness across the sample needs to be checked by comparing the treatment and control group and their assigned distributions. For correct statistical analysis, it must be given that the treatment and control assignment is determined independently from the attributes.

g. Balancing Checks

Balancing Checks:

	Treated	Control	MeanDiff	Std	tVal	pVal	StdDiff
age	25.96	25.01	0.94	0.72	1.32	0.19	13.44
ed	10.29	10.11	0.18	0.19	0.97	0.33	10.06
black	0.84	0.84	0.00	0.04	0.11	0.91	1.12
hisp	0.06	0.10	-0.04	0.03	-1.38	0.17	13.80
married	0.20	0.16	0.04	0.04	0.97	0.33	9.94
nodeg	0.73	0.83	-0.10	0.04	-2.41	0.02	24.81
re74	2154.95	2217.10	-62.15	549.93	-0.11	0.91	1.13
re75	1530.87	1317.87	213.00	332.07	0.64	0.52	6.53
re78	6318.53	4708.88	1609.65	724.25	2.22	0.03	23.27

As given by the definition of the balancing checks, any standardized differences above 10 are considered as large. According to the table above (measure “StdDiff”), the variables “age”, “ed”, “hisp”, “nodeg” and “re78” are hence determined as large differences between the treatment and control group.

The highest standardized difference can be observed in the degree variable. While 83% of the observations in the control group have no degree, only 73% of the treatment group have no degree. This difference can be stated as highly significant due to a p-value of 0.02 (determined confidence level of 0.05). The second highest standardized difference can be found in the earnings variable of 1978, where the control group has an average income of 4'708.88, while the treatment group has an average income of 6'318.53. Resulting in a standardized difference of 23.27, this result is also highly significant with a p-value of 0.03. Compared to these findings, the variables “age”, “ed” and “hisp”, which have standardized differences of 13.44, 10.06 and 13.80, the results are not statistically significant with p-values of 0.19, 0.33 and 0.17, respectively, with a predetermined confidence level of 0.05. Even higher, but still plausible confidence levels (e.g., 0.1), would still not lead to significance.

As a result, with five variables that show standardized differences above 10, only “nodeg” and “re78” can be considered as imbalanced covariates across the data set due to their statistical significance, while the other covariates can be considered balanced based on the balance checks.

Part 2: ATE Estimation

a. Estimation of ATE + Interpretation of Results

ATE Estimate by Difference in Means:

Dependent Variable: re74

	ATE	SE	tValue	pValue
MeanDiff	-62.15	549.93	-0.11	0.91

ATE Estimate by Difference in Means:

Dependent Variable: re75

	ATE	SE	tValue	pValue
MeanDiff	213.0	332.07	0.64	0.52

ATE Estimate by Difference in Means:

Dependent Variable: re78

	ATE	SE	tValue	pValue
MeanDiff	1609.65	724.25	2.22	0.03

Per the table above the ATE is estimated at \$-62.15 / \$213.0 / \$1'609.65. This can be interpreted as the difference between the mean annual earnings in 1974 / 1975 / 1978 of the treatment and control groups corresponds to the aforementioned \$-62.15 / \$213.0 / \$1'609.65. Notably, we find that only the difference in 1978 is statistically significant with a p value of 0.03 at the significant level of 5%. Both 1974 and 1975 don't seem to be statistically significant with p values of 0.91 and 0.52 respectively. Given the 5% significance level of the 1978 ATE, the implication of this is that individuals that received treatment had on average a higher salary. For the 1978 data, a standard error of \$724.25 is somewhat high.

With regards to the assumptions made, we need to make the "stable unit treatment value assumption" (SUTVA) which assumes that the value of the potential outcome is unaffected by the mechanism applied to assign the treatment.

b. OLS Function for ATE Estimation without Covariates + Differences

OLS Regression Results

Dep. Variable:	re78	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	2.357
Date:	Sun, 27 Feb 2022	Prob (F-statistic):	0.0134
Time:	22:08:54	Log-Likelihood:	-4082.4
No. Observations:	400	AIC:	8185.
Df Residuals:	390	BIC:	8225.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	175.8197	3667.208	0.048	0.962	-7034.151	7385.791
treat	1495.9086	683.579	2.188	0.029	151.947	2839.870
age	64.6261	49.565	1.304	0.193	-32.822	162.075
ed	424.8839	244.950	1.735	0.084	-56.704	906.472
black	-1898.1777	1264.402	-1.501	0.134	-4384.075	587.720
hisp	732.9484	1693.211	0.433	0.665	-2596.016	4061.912
married	-267.5814	930.738	-0.287	0.774	-2097.473	1562.311
nodeg	-73.0352	1090.376	-0.067	0.947	-2216.786	2070.716
re74	0.0655	0.081	0.811	0.418	-0.093	0.224
re75	0.0763	0.142	0.537	0.592	-0.203	0.356

Omnibus:	265.161	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3607.562
Skew:	2.623	Prob(JB):	0.00
Kurtosis:	16.745	Cond. No.	7.55e+04

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 7.55e+04. This might indicate that there are strong multicollinearity or other numerical problems.

In the screenshot above the output summary of the OLS regression that included all covariates can be seen. It provides the beta coefficients, their standard errors, t-values and p-values.

can be seen: it provides the beta coefficients, their standard errors, t values

OLS Regression Results						
Dep. Variable:	re78	R-squared:	0.014			
Model:	OLS	Adj. R-squared:	0.011			
Method:	Least Squares	F-statistic:	5.605			
Date:	Sun, 27 Feb 2022	Prob (F-statistic):	0.0184			
Time:	22:08:54	Log-Likelihood:	-4090.2			
No. Observations:	400	AIC:	8184.			
Df Residuals:	398	BIC:	8192.			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4708.8786	436.670	10.784	0.000	3850.410	5567.347
treat	1609.6511	679.895	2.368	0.018	273.017	2946.285
Omnibus:	263.171	Durbin-Watson:	1.946			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3309.755			
Skew:	2.630	Prob(JB):	0.00			
Kurtosis:	16.074	Cond. No.	2.46			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Given the results of section 2a) the focus was set here on the year 1978 as it was found to be considerably more significant than the years 1974 / 1975. Running the OLS regression without the covariates it can be observed that the ATE we obtained in 2a) is identical to the coefficient of treatment. Notably, this is only applicable for the regression without covariates. However, the standard error here of 679.895 is not the same as the one obtained for the ATE. The p value is also slightly lower at 0.018 (versus 0.03 from 2a)) but similarly also shows significance at the 5% level. So, while we get the same value for the ATE and the coefficient constant of *treat*, the OLS seems to perform slightly better with a lower standard error and lower p value.