University of St.Gallen

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

Data Analytics II: PC5

University of St. Gallen

Jonas Husmann | 16-610-917

Niklas Leander Kampe | 16-611-618

Prof. Dr. Michael Lechner

March 28, 2022

Part 1: Descriptive Statistics

a. Load Data Set

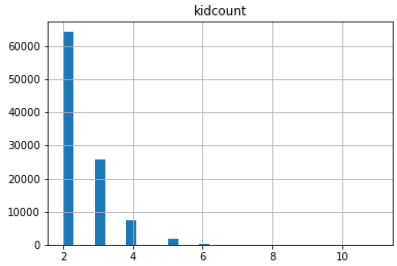
	weeks_work	employed	kidcount	morekids	multi2nd	samesex	black	hisp	age_mother	hsgrad	colgrad
0	52	1	2	0	0	0	1	0	32	1	0
1	32	1	4	1	0	0	0	0	34	0	1
2	52	1	3	1	0	0	0	0	31	0	1
3	12	1	3	1	0	0	0	0	24	1	0
4	0	0	2	0	0	1	0	0	30	1	0

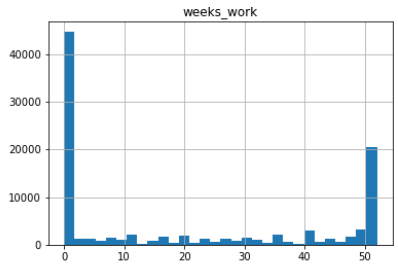
b. Summary Statistics + Plausibility and Patterns

Descriptive Statistics:

	mean	var	std	max	min	na	unique
weeks_work	20.19	486.03	22.05	52.0	0.0	0.0	53.0
employed	0.56	0.25	0.50	1.0	0.0	0.0	2.0
kidcount	2.49	0.61	0.78	11.0	2.0	0.0	10.0
morekids	0.36	0.23	0.48	1.0	0.0	0.0	2.0
multi2nd	0.01	0.01	0.09	1.0	0.0	0.0	2.0
samesex	0.50	0.25	0.50	1.0	0.0	0.0	2.0
black	0.12	0.10	0.32	1.0	0.0	0.0	2.0
hisp	0.03	0.03	0.17	1.0	0.0	0.0	2.0
age_mother	29.70	13.61	3.69	35.0	21.0	0.0	15.0
hsgrad	0.48	0.25	0.50	1.0	0.0	0.0	2.0
colgrad	0.31	0.21	0.46	1.0	0.0	0.0	2.0

According to the summary statistics, the data set does not contain any missing values, which can be identified from the column “na” showing zeros for all variables. Furthermore, the dummy variables “employed”, “morekids”, “multi2nd”, “samesex”, “black”, “hisp”, “hsgrad” and “colgrad” seem to be coded correctly as they show min-values of 0, max-values of 1 and number of unique values of 2. In addition, all values of the “continuous” variables seem to be plausible compared to their intrinsic definitions/descriptions. While looking at the race of the study participants, it seems to be quite homogenous with 12% being Afro-American (“black”) and 3% being Hispanic (“hisp”). In addition, the sample is quite balanced regarding the education, as 48% have high school degrees and 30% having a college degree.





According to the histograms, the variable “kidcount” has the highest mass of observations at the value of 2, which is strictly decreasing up the value of 11. Looking at the distribution, one could think about defining a threshold value of outlying observations to prevent their influence on further econometric analysis. A reasonable threshold could be defined for more than six kids, but requires additional analysis to not omit too much observations. In addition, according to the histogram of “weeks_work”, the distribution is strongly skewed to the min- and max-extremes. This shows that most of the sample either does not work or follows a full-time job, while the rest of the sample is part-time employed with working times between 0 and 50.

c. Means and Numbers of Observations

Means & Observations:

kidcount	employed mean	count
2	0.5933	64373
3	0.5160	25708
4	0.4590	7434
5	0.3907	1876
6	0.3202	431
7	0.2960	125
8	0.2143	42
9	0.3750	8
10	0.5000	2
11	0.0000	1

The table above shows the mean value of the dummy variable “employed” for every number of kids, as well as the number of observations falling into the unique “kidcount” values. First, in accordance with the distribution/histogram of “kidcount”, the number of observations is strictly decreasing in the number of children. Furthermore, the ratio/percentage of being employed, which is defined as the mean value of a dummy variable, is strictly increasing in the number of children until the number of children reaches a value of 9 or more. As the number of observations are fairly small for 9 or more children, and hence the influence on a single observation on the mean is rather large, one could think of applying a threshold on outlying observations on “kidcount” at 9 and more children, as this would underline a strict decrease of employment in the number of children for the whole data set, which would be in accordance with general assumptions that employment becomes less as more children a woman has. Nevertheless, the observations for 8 or less children is expected with participants who have 2 children being most likely employed at a rate of about 60%.

d. Cross Table

Cross Table:

morekids \ multi2nd	0	1
0	64373	0
1	34726	901

The table above shows the number of observations which fullfill the unique value combinations for the variables “multi2nd” and “morekids”. The results are indeed plausible, as, first of all, having less or equal than 2 kids (“morekids” = 0) and having the 2nd birth with two or more children at a time (“multi2nd” = 0), which would conclude a total of more than two kids, has zero observations. This underlines full plausibility. Furthermore, most observations ($\approx \frac{2}{3}$,64’373) lie at the intersection of having not more than two kids and not having a second birth with more than one kids, which is also in accordance with the distribution of “kidcount”. The second highest likelihood ($\approx \frac{1}{3}$,34’726) lies at the intersection of having more than two kids but not having a second birth with more than one child. Lastly, a minority is observed with having more than two kids and having a second birth with more than one child (901). In total, the sum of all four unique combinations also equals the number of total observations. Hence, the cross-table observations are expected and seem fully plausible.

Part 2: Homogeneous Effects

a. OLS Estimation + Causal Relationship

OLS Estimation Results:

	coef	se	t-value	p-value
intercept	-9.25	0.57	-16.34	0.00
morekids	-5.82	0.15	-39.51	0.00
black	8.13	0.21	38.02	0.00
hisp	0.82	0.41	2.00	0.05
age_mother	0.94	0.02	48.92	0.00
hsgrad	3.12	0.18	17.28	0.00
colgrad	3.40	0.20	16.97	0.00

The OLS estimation results show that having more than two children has an estimated effect of -5.82 weeks worked per year. We need to consider whether the underlying assumptions for a causal effect are fulfilled or not. We should make sure that there are no confounding factors that are uncontrolled for that can influence the relationship between *morekids* and *weeks_work* – such potential confounding factors could be the number and quality of daycore offerings, other non-earned income by the family or flexibility provided by the job (e.g. home-office). These factors, which are currently contained in the error term could lead to the error term being correlated with *morekids* and therefore endogeneity may arise. Additionally, the SUTVA (stable unit treatment value assumption) should hold as well, i.e. the outcome of a specific unit only depends on the treatment that was assigned to said units, and not the treatments applied to others. Given these two assumptions, it seems unlikely that the estimated coefficient sufficiently describes a causal relationship.

b. 2SLS Estimator + Causal Effect

OLS Estimation Results:

Dependent Variable: weeks_work

	coef	se	t-value	p-value
intercept	-8.93	0.67	-13.42	0.00
black	8.00	0.26	31.20	0.00
hisp	0.72	0.42	1.69	0.09
age_mother	0.91	0.04	25.37	0.00
hsgrad	3.31	0.27	12.43	0.00
colgrad	3.67	0.35	10.46	0.00
0	-4.77	1.13	-4.23	0.00

Using a two stage least squares regression (TSLS) we find that mothers from African American and Hispanic origin work 8 and 0.72 weeks more per year than mothers from other backgrounds. Additionally, it can be inferred that the weeks worked increases with the age of the mother by 0.91 weeks per year. Graduates, high school as well as college, work 3.31 and 3.67 weeks more per year, respectively. Finally, we find that having more than two kids decreases the weeks worked per year by an estimated 4.77. We can find that after using a TSLS the effect of having more than two kids on weeks worked per year is not as high (-4.77 vs -5.82) and is still highly significant. TO

make sure that an instrumental variable is valid we need to make sure the following two assumptions hold:

a) The instrumental variable has a causal effect on the treatment variables
b) The instrumental variable affects the outcome only through the treatment variables and not directly (exclusion restriction)

For assumption a) to hold one would have to test whether the instrumental variable is correlated with the error term or not – if this assumption is not fulfilled, i.e. the IV has an effect on the outcome our estimated IV effect is biased. With regards to the second assumption, we can assume this to be satisfied in this example.