

# Introduction to Causal Inference

Michael Kühhirt

michael.kuehhirt@uni-koeln.de



## Lab 1

### *Implementing the Roadmap for Causal Inference to Learn about Total Effects*

#### DESCRIPTION

In this lab, you follow the roadmap for causal inference (Petersen and Laan, 2014) to learn about the average total effect of maternal full-time employment when the child is aged 3 on children's behavioral development around age 6 using German data from the Socio-Economic Panel (SOEP) (Wagner, Frick, and Schupp, 2007). The lab carries you through each step of the roadmap from specifying a graphical causal model, to formal definition and identification of the effect of interest, and finally to estimation of the effect using STATA. The following sections introduce some substantive theoretical background for the research question, describe the data available for analysis, and detail the instructions for completing the lab.

Please note that theory, data, and analysis for this lab are oversimplified compared to a master thesis and, particularly, to publishable research. This is necessary to keep the analysis manageable within the confines of a practical exercise.

#### THEORETICAL BACKGROUND

*Why could maternal full-time employment affect children's subsequent behavioral development?* There are several possible pathways through which maternal full-time employment could have an impact on the development of behavioral problems (Waldfogel, 2002). On one hand, it may curb the development of behavioral problems by increasing the family's income and therefore the ability to provide high-quality childrearing materials and services and by generally easing economic strain that is often associated with stress and conflict in the home. On the other hand, through a reduction in the time available for important mother-child interactions, the children of full-time employed mothers may be more likely to develop behavioral problems.

*What are alternative explanations for an association between maternal employment status and behavioral development?* There are a number of factors that may lead to a noncausal association (Belsky and Eggebeen, 1991; Han, Waldfogel, and Brooks-Gunn, 2001; McMunn

et al., 2012). Single mothers may face a greater need to work full-time, while father absence may increase the likelihood of behavioral problems. Human capital theory suggests that more educated mothers are more likely to be employed full-time. These mothers may also have skills and coping strategies that help curb behavioral problems. Maternal health may also be an important determinant of both employment status and children's development. Lastly, characteristics of the child, such as personality and temperament may affect mothers' employment decisions as well as the development of behavioral problems.

## DATA AND VARIABLES

The available data come from the teaching version of the SOEP, a 50% subsample of the full data set. These data contain information on roughly 1,200 children born in Germany between 2002 and 2008 and their families. No further preparations are necessary before the analysis. The table below displays the available variables. The variable `matfte3` is an indicator whether a mother worked full-time (or not) when the child was around 3 years old. A measure of children's behavioral problems around age 6 is provided with `sdq6`, a scale based on the Strengths and Difficulties Questionnaire (Goodman, 1997), on which a higher number indicates more problematic behavior. Please inspect the covariates using commands such as `tabulate`, `summarize`, and `codebook` for information on the scale of measurement, distribution, etc.

```
. describe
```

```
Contains data from 01labte.dta
```

```
obs:      1,226      SOEP.v31@school: bioagel.dta
vars:      10        2 Dec 2016 18:54
size:     42,910
```

variable name	storage type	display format	value label	variable label
kidnr	long	%12.0g		child ID
persnr	long	%12.0g		mother ID
sdq6	float	%9.0g		SDQ score at age 6
matfte3	float	%26.0g	matfte3	mother full-time employed at age 3?
hhinc6	float	%9.0g	hinc14	household income at age 6
matedu3	float	%53.0g	isced97	maternal education at age 3
single3	float	%20.0g	single3	partner in household at age 3?
east6	float	%16.0g	east6	area of residence at age 6
bsex	byte	%10.0g	bsex	sex of child
birthy	int	%8.0g		year of birth of child

## INSTRUCTIONS

*Before you start:*

1. Check if the folders on your UoC webspace are intact.
2. Download the lab data `01labte.dta` from *Ilias* into your folder `data`.
3. Download the do-file `07estte1.do` from *Ilias* into your folder `code`.
4. Rename that file `01labte.do`. It will serve as a template for your lab do-file.
5. After opening the do-file within Stata, you can load the lab data by (1) adapting the working directory in line 28 (2) inserting the name of the lab data in place of the toy data's in line 45 and (3) executing the do-file until that line (i.e., marking all lines incl. 45 and pressing `ctrl + d`).

*Completing the lab:* The lab requires you to apply the contents of the course up to this point to address a substantive question in empirical social research. The subtasks follow Petersen and van der Laan's (2014) roadmap for causal inference. For a 100% completion of the lab, a maximum of 10 points is awarded. See the subtasks below for the distribution of points over the tasks.

1. Specify a graphical causal model that depicts the causal relations outlined in the theoretical background *and* the information (not) contained in the data! (1 point)  
*You are encouraged to use DAGitty for drawing your model. Feel free to add to the theoretical background, e.g., if you think an important confounder is missing.*
2. Define the average total effect of interest using hypothetical interventions and counterfactuals! (1 point)  
*Provide the notation as well as a sentence with the substantive meaning of the effect.*
3. Assess whether the total effect of interest can be identified by covariate adjustment with the available data and name the set of adjustment variables! If it is not identifiable, specify the minimal assumptions that allow identification. (1 point)  
*You can use DAGitty also for identification. When making additional assumptions (i.e., deleting arrows from your original DAG), please make only those necessary and minimally sufficient for identification by covariate adjustment with the data provided. Also write a sentence with the substantive meaning of the assumption(s).*
4. Estimate the observable mean difference using linear regression and capture the output! (0.5 point)
5. Estimate the observable mean difference using inverse probability of treatment weighting and capture the output! (0.5 point)
6. Carry out checks of model specification and positivity, capture the output, and provide information on whether you found evidence for model misspecification and (strong) violations of positivity and what that evidence is! (2 points) *Tasks 4.–6. require you to adapt the commands from the template do-files to the variables needed for your analysis. You don't need additional commands to complete this task. You also don't need to recode any variables contained in the lab data. To include a categorical variable with more than 2 values in the estimation models insert `i.` before the variable (e.g., `i.matedu3`). To include a product term between two categorical variables, write `i.var1##i.var2`. To include a second order polynomial for a metric variable, write `c.var##c.var`. To include a product term between a categorical variable and a metric variable, write `i.var1##c.var2`.*
7. Provide the statistical and counterfactual interpretation of your estimates along with an assessment of the respective assumptions (also based on your checks of model specification and positivity) (4 points).  
*For each interpretation, one sentence is sufficient. With regard to assumptions about structural sources of bias, please tailor your assessment to the topic at hand, e.g., mentioning crucial identification assumptions about specific variables that need to be made in order to endow the estimates with a causal interpretation. Merely mentioning the assumptions by name will not result in any points awarded.*

*Submitting the lab:*

- Labs are submitted on *Ilias*.
- Enumerate the tasks with the same numbers as used above!
- Provide only DAGs, figures, STATA output that are used to answer a task and that you interpreted.
- For any STATA output, also provide the command(s) that created the output (above the output)!
- For any STATA commands and output, please use the mono font Courier New!
- For any STATA output submitted, please make sure that it does not break across lines (e.g., by adjusting the font size).
- file name: [your last name]\_L1
- file format: PDF (Please do not submit any other format!)
- deadline: December 11, 2017, 11:59pm

## REFERENCES

- Belsky, J. and Eggebeen, D. (1991). "Early and extensive maternal employment and young children's socioemotional development: Children of the National Longitudinal Survey of Youth". In: *Journal of Marriage and Family* 53 (4), pp. 1083–1110. DOI: 10.2307/353011.
- Goodman, R. (1997). "The strengths and difficulties questionnaire: A research note". In: *Journal of Child Psychology and Psychiatry* 38 (5), pp. 581–586. DOI: 10.1111/j.1469-7610.1997.tb01545.x.
- Han, W.-J., Waldfogel, J., and Brooks-Gunn, J. (2001). "The effects of early maternal employment on later cognitive and behavior outcomes". In: *Journal of Marriage and Family* 63 (2), pp. 336–354. DOI: 10.1111/j.1741-3737.2001.00336.x.
- McMunn, A. et al. (2012). "Maternal employment and child socio-emotional behaviour in the UK: Longitudinal evidence from the UK Millennium Cohort Study". In: *Journal of Epidemiology and Community Health* 66 (7), pp. 1–6. DOI: 10.1136/jech.2010.109553.
- Petersen, M. L. and Laan, M. J. van der (2014). "Causal models and learning from data: Integrating causal modeling and statistical estimation". In: *Epidemiology* 25 (3), pp. 418–426. DOI: 10.1097/EDE.0000000000000078.
- Wagner, G. G., Frick, J. R., and Schupp, J. (2007). "The German Socio-Economic Panel Study (SOEP) – Scope, evolution and enhancements". In: *Schmollers Jahrbuch* 127 (1), pp. 139–169.
- Waldfogel, J. (2002). "Child care, women's employment, and child outcomes". In: *Journal of Population Economics* 15, pp. 527–548. DOI: 10.1007/s001480100072.