Estimation of causal effects in natural experiments with a limited number of time points: learning losses due to school closures during the COVID pandemic

Florian van Leeuwen

Supervisors: Remco Feskens & Peter Lugtig

Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University
January 2023

Word count: 2493

Candidate journals: *The Journal of the Royal Statistical Society, Series A (Statistics in Society)*

1 Introduction

The effects of interventions in many fields are investigated with the use of randomised controlled trials (RCT). In an RCT, a researcher introduces an intervention and participants are randomly assigned to one or more treatments [1]. The RCT is regarded as the gold standard for evaluating the effectiveness of interventions since it allows us to ascertain the internal validity of causal claims [2]. As random assignment is not possible for some interventions (e.g., a policy change) natural experiments are often used [3]. The researcher thus has no influence over when the intervention happens or who is assigned to the treatment or control groups.

Natural experiments are extremely important for policymakers as this is often the only way to estimate the effect of a policy or phenomenon. The results of such studies influence the implementation of new policies e.g., should the government introduce a lockdown of schools after a COVID outbreak knowing that this has a negative effect on the learning ability of students? The estimation of the causal effect in a natural experiment is, however, not as easy as in an RCT due to the lack of randomization in the experimental groups. The comparison of groups which cannot be considered equal can lead to a biased estimate of the causal effect as there can now be many other factors, other than the intervention, that have an effect on the outcome variable. To combat this problem there are many different forms of natural experiments and each form attempts to control for factors that are not of interest so the causal effects of the intervention can be estimated [4] [5] [6].

This study will focus on natural experiments where the entire population is in the treatment group and the data stems from a longitudinal study, where individuals are measured repeatedly over time [7]. One of the main ways to estimate the effects of intervention with data of this kind is an interrupted time series design (ITS). In an ITS the effects of the intervention are estimated by comparing the outcome after the intervention to the projected outcome if there had not been an intervention [8]. ITS designs are increasingly used, it is considered one of the best designs for establishing causality when RCT's are not possible [9].

An ITS design can be used with multiple models, including: segmented regression [10], autoregressive (AR) error models [11], autoregressive integrated moving average (ARIMA) models [12], multilevel models [8] and SEM models [13]. Each model has upsides and downsides and it is not yet clear how well the models compare to each other in specific scenario's.

Within this study, an empirical example is used. We will use data from a student monitoring system administered in the Netherlands within primary education. Measurements start in grade 3 and the educational progress of pupils is followed until grade 8. There are a maximum of 8 data points per pupil, which is considered as not a lot for an ITS design [14]. The intervention is school lockdowns due to COVID, and the effect of this intervention on the math scores of pupils is estimated. The research question of this study is:
What are suitable models to estimate ITS designs if there are few time points?

2 Background Information

This section consists of two parts. Firstly regression-based methods for estimating ITS designs are discussed. In general ITS designs are used to estimate two things: an immediate change and a graduate change after the intervention. Secondly, the same models are discussed in a multilevel context.

2.1 Linear regression methods

In linear regression, an outcome (Y) is estimated by a predictor (X). This relationship can be expressed as a linear line such as in the equation below:

$$Y = \beta_0 + \beta_1 X, \tag{1}$$

Where β_0 is the intercept or in other words the value of Y where X is equal to zero. β_1 is the slope of the line. The common way to obtain the estimates in the model is through ordinary least squares (OLS) estimation. The specifications about OLS and it's assumptions can be found in [15].

2.1.1 Segmented regression

The most used method for estimating an ITS is segmented regression. A minimum of three variables are required for an ITS analysis:

- T (time)
- X_t (a dummy (0/1) indicating the pre/post intervention period)
- Y_t (outcome at time point t)

There is a need for multiple measurement moments (T), an outcome at each time point (Y_t) and knowledge of when the intervention happened (X_t). To estimate the effect of the intervention, the following OLS segmented regression model can be used [10]:

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t, \tag{2}$$

where β_0 indicates the baseline level at $T = 0$, β_1 represents the trend before the intervention, β_2 is the level change due to the intervention at $T = 0$ and β_3 is the trend change due to the intervention at. There need to be at least two measurements of Y_t before and after the intervention to estimate the baseline slope (β_1) and the change in slope (β_3) since a slope can only be estimated if there is more than one data point.

A downside to using segmented regression in an ITS design is that it does not take into account the effect of autocorrelation. If autocorrelation is present then it could be helpful to utilize models that explicitly model autocorrelation. A commonly used model is the Autoregressive integrated moving average (ARIMA) model.

A problem with utilizing this model in a ITS design is it does not yield interpretive estimates for the general trend and intercept, which are necessary for an overall comparison. Another downside of the ARIMA model is it needs at least 50 times points [16], but preferably 100 or more [17]. The process of using an ARIMA model in an ITS design is described in [12] for further context.

2.2 Multilevel models

The interest of an ITS analysis can be the difference in intercept or slope after an intervention, but the variation in such an effect is also an important factor e.g., do all people have the same response to the intervention or is there large variability? These growth curves can be modelled with a multilevel (ML) model [18]. Take the example of students that make a test every year for a number of years. The first level equation of the multilevel model is then:

$$Y_{ti} = \pi_{0i} + \pi_{1i} T_{ti}, \tag{3}$$

where Y_{ti} is the score on the test for individual i at time point t, T is a variable that indicates the time point.

The second level equations are then:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + u_{0i}, \qquad (4)\\ \pi_{1i} &= \beta_{10} + u_{1i}. \qquad (5) \end{aligned}$$

Time-varying predictors can be added in the first level to explain some of the level one variance, and subsequently time-invariant predictors in the second level to help explain level two variance. By utilizing the ITS design from Equation 2 again, then we obtain the following multilevel model: Level one

$$Y_{ti} = \pi_{0i} + \pi_{1i} T_{ti} + \pi_{2i} X_{ti} + \pi_{3i} X_{ti} T_{ti}, \tag{6}$$

where X_{ti} is an indicator of if the intervention has occurred for individual i at time point t. The intercept (π_{0i}) is estimated as the mean score of students at $T = 0$ and the slope indicates how much a student learns per T (π_{1i}), the immediate effect of the intervention (π_{2i}) and the gradual effect of the intervention over time π_{3i}. Similarity to before for level two equations are:

$$\begin{aligned} \pi_{0i} &= \beta_{00} + u_{0i}, \qquad (7)\\ \pi_{1i} &= \beta_{10} + u_{1i}, \qquad (8)\\ \pi_{2i} &= \beta_{20} + u_{2i}, \qquad (9)\\ \pi_{3i} &= \beta_{30} + u_{3i}. \qquad (10) \end{aligned}$$

Next to obtaining a indicator of the variance, another advantage of using a ML model is that individuals with a different number of time points or time points that are not equally spaced can still be estimated [18].

3 Motivating example

During the COVID pandemic lockdowns, many primary schools changed from on-location education to an online variant of education. The shift in education methods has impacted the amount children have learned [19–22]. The objective of these studies is to estimate the average (causal) effect of school closures on the learning of children, variation in individuals is disregarded. Only a limited number of timepoints are used in these studies as students only usually do not take more than one standardised test per year. Potential catch-up effects can only be accounted for if more data is included after the lockdowns. To asses some of these problems, this motivating example is used.

3.1 Data

The data that is used in the example stems from Cito [1]. It consists of educational test scores for pupils in grade 4-7 that are comparable through item response theory (IRT) models. About 80% of primary schools participate in the program of Cito to track the academic career of pupils. Each participating school tests the pupils twice a year two tests in January and May. The test consists of multiple parts, this study will focus on the mathematical part as the most data was avaiable for this subject. The data has a longitudinal cohort nature [23], each cohort is measured at the same time point in their school career but not at the same point in time. This study will focus on cohorts who have at least two observations before and after the lockdowns so that a change in slope can be estimated. The cohorts can be seen in Table 1.

	2017	2018	2019	2020	2021	2022
Cohort 1		O	O	O	O	
Cohort 2			Z	Z	Z	Z

Table 1: Longditudial cohort data

There are 8 observations per student and a maximum of 30,000 students per time point, after filtering duplicates and missing. There is a hierarchical structure of time points within students and students within schools [24]. The learning rate is not the same for all students [25]. Figure 1 shows the differences in slopes for a sample of students from different cohorts. The grey area indicates the school closures due to COVID.

[1] https://www.cito.nl/onderwijs/primair-onderwijs/lvs

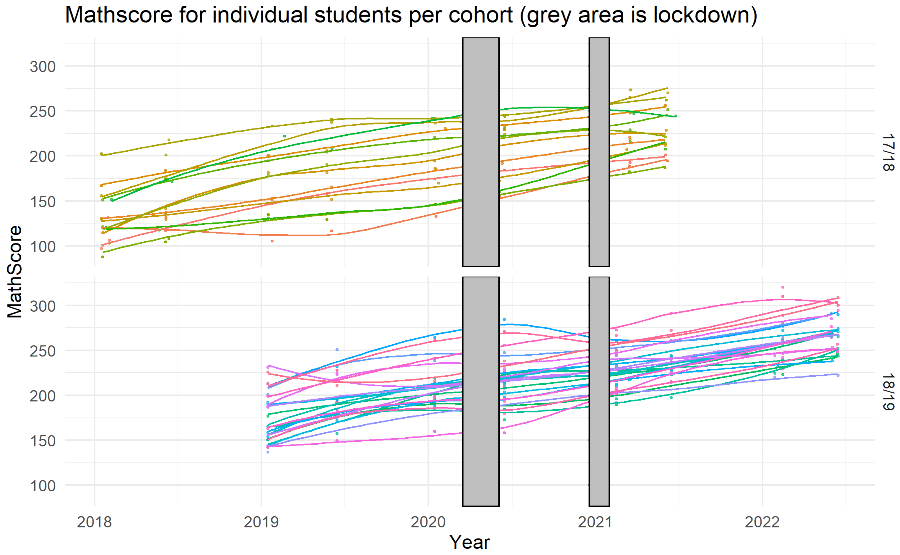

Figure 1: Individual growth curves for 25 students per cohort sampled randomly from CITO data on math scores over a 4-year period

3.2 Methods

The effect of the lockdowns is estimated using the following methods from Section 2. Background information: Segmented Regression, Multilevel Model. AR and ARIMA models are not included as the parameters in the models are difficult to interpret. Time is defined in years, and centred around the start of the first lockdown. This makes it possible to estimate the change in intercept right after the lockdown for both cohorts at the same time.

3.2.1 Segmented Regression

Firstly, the learning losses are estimated using the standard segmented regression from Equation 2. As the motivating example showed, the increase in math ability does not follow a purely linear trend. The segmented regression model is thus extended by including a quadratic term for time to account for possible decreasing growth curves:

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T^2 + \beta_4 T X_t \tag{11}$$

Additionally, due to the two cohorts in the data, an indicator for cohort needs to be added. The indicator shows the difference in the score of the cohorts at time point 0.

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T^2 + \beta_4 C + \beta_5 T X_t \tag{12}$$

where C is 0 for the first cohort and 1 for the second cohort. Interaction terms between the intervention (X_t)/the cohort C and time squared (T^2) are not added to the model to prevent overfitting [26].

This model can now be used to estimate the average learning losses of all pupils in the two cohorts. If differences in between the cohorts is of interest than more parameters, interactions between the cohort variable and variables of interest, need to be added to the model.

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T^2 + \beta_4 C + \beta_5 T X_t + \beta_6 T C + \beta_7 X_t C + \beta_8 T X_t C \quad (13)$$

Where β_6 is the difference in the linear trend for the cohorts, β_7 is the difference in the effect of covid on the intercept for the different cohorts and β_8 is the difference in the effect of covid on the slope for the different cohorts.

3.2.2 Multilevel Model

To obtain estimates of the variation in learning of students multilevel models (12 and 13) are used. In these models, the variation of the intercept and slope as well as the covariance between the two can be obtained.

3.3 Results

The results of the segmented regression will be first discussed after which the same models in a ML context are assessed.

3.3.1 Segmented Regression

In Table 2 two models are shown. The first model has the model Equation 12. All parameters in the model are significant, which is not unusual with such a large sample size. The intercept in the model depicts the mean math score of students at the time of the first lockdown for cohort 17/18, which is between grades 6 and 7. The coefficient of time is 28.15, which indicates how much a student learns each year. The dummy for the cohort indicates that the mean math score of the later cohort is 25.89 lower than the first cohort. This parameter is almost the same value as the coefficient of time; this is expected because cohort 18/19 is one year behind cohort 17/18. The quadratic term for time is -0.66, so the learning rate decreases by a little over a half point every year. The covid dummy is -3.04, indicating that students had an initial drop of 3 points or 11% of a year's education in their learning scores due to the first lockdown. The interaction term between covid and time shows a rather large change in slopes of -2.98 for every year after the first lockdown.

The second model had the model equation 13. The intercept (234.67) as well as the coefficient for time (26.52) and cohort (-29.85) are a bit lower than in the first model. The quadratic term for time is much larger (-1.93) indicating that the slope of the learning curve decreases by almost 2 points each year. The coefficient of the covid dummy is now almost twice as large (-5.90) this is due to the interaction term, as this is now the immediate effect of the first lockdown for cohort 17/18. The parameter for the slope is almost three times smaller (-1.16) than in the first model, this is again the estimate only for cohort 17/18. The coefficient of the interaction term between covid and cohort is 5.94. This means that the change in intercept for cohort 18/19 is almost 0. The interaction term between time, covid and cohort are 2.18 indicating that the slope increases after the first lockdown for cohort 18/19.

	Segmented Regression	Interactions cohort
(Intercept)	232.94^{***}	234.67^{***}
	(0.18)	(0.20)
Time	28.15^{***}	26.52^{***}
	(0.30)	(0.32)
Cohort 18/19	-25.89^{***}	-29.85^{***}
	(0.12)	(0.17)
Time^2	-0.66^{***}	-1.93^{***}
	(0.12)	(0.14)
covid	-3.04^{***}	-5.90^{***}
	(0.21)	(0.34)
Time:covid	-2.98^{***}	-1.16
	(0.59)	(0.60)
covid:Cohort 18/19		5.94^{***}
		(0.39)
Time:covid:Cohort18/19		2.18^{***}
		(0.36)
AIC	3014187	3013151
R^2	0.49	0.49
Adj. R^2	0.49	0.49
Num. obs.	312480	312480

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: OLS models for the simple segmented regression (equation 2) and the extension of a quadratic term (equation 11)

3.3.2 Multilevel Model

The ML models were estimated so that each student obtained free intercepts [2], the parameters were identical to Table 2. The estimates for the models are almost identical to Table 2. The addition of the multilevel model is that the correlation within students is accounted for and a parameter is obtained for the variance of the intercepts (628.25/628.61). The ML models performed better than the OLS models (AIC: 2631908/2626735).

[2] Random sloped for time were not possible, as the model did not converge

4 Discussion

We have seen that it is possible to use an interrupted time series design with a lower number of timepoints. The results from the OLS models and the ML models were very similar, with the exception of the additional variance term in the ML model. The ML models did perform better on AIC. It is also possible to extend the classical segmented regression model with a quadratic term for time and to include multiple cohorts. The effect of the intervention on the intercept and slope can be dissected for different cohorts. It is, however, not clear how accurate the estimations are with only 8 timepoints. Latent growth curve models were not yet included as it is not possible to include two cohorts in one model, a possible solution is the use of a multiple group growth model [27].

The second part of this research project will focus on two things. Firstly, the estimation of the learning losses with a latent growth curve model. Secondly, a simulation study will be performed to asses the performance of the models. There have been some simulation studies that estimate the power of the ITS design under specific circumstances [11, 14], these studies do not focus on a small number of timepoints and utilize data generating models that are not realistic compared to the motivating example. The simulated data in this project will have a decreasing slope, as is the case in the motivating example, which is something that is not included in an ITS power analysis before.

References

[1] Miquel Porta. *A dictionary of epidemiology.* Oxford university press, 2014.

[2] Al K Akobeng. “Understanding randomised controlled trials”. In: *Archives of disease in childhood* 90.8 (2005), pp. 840–844.

[3] Peter Craig et al. “Using natural experiments to evaluate population health interventions: new Medical Research Council guidance”. In: *J Epidemiol Community Health* 66.12 (2012), pp. 1182–1186.

[4] Guido W Imbens and Jeffrey M Wooldridge. “Recent developments in the econometrics of program evaluation”. In: *Journal of economic literature* 47.1 (2009), pp. 5–86.

[5] Peter M Steiner et al. “The importance of covariate selection in controlling for selection bias in observational studies.” In: *Psychological methods* 15.3 (2010), p. 250.

[6] Scott T Leatherdale. “Natural experiment methodology for research: a review of how different methods can support real-world research”. In: *International Journal of Social Research Methodology* 22.1 (2019), pp. 19–35.

[7] Peter Diggle et al. *Analysis of longitudinal data.* Oxford university press, 2002.

[8] Evangelos Kontopantelis et al. “Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis”. In: *bmj* 350 (2015).

[9] William R Shadish, Thomas D Cook, and Donald T Campbell. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company, 2002.

[10] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. “Interrupted time series regression for the evaluation of public health interventions: a tutorial”. In: *International journal of epidemiology* 46.1 (2017), pp. 348–355.

[11] Fang Zhang, Anita K Wagner, and Dennis Ross-Degnan. “Simulation-based power calculation for designing interrupted time series analyses of health policy interventions”. In: *Journal of clinical epidemiology* 64.11 (2011), pp. 1252–1261.

[12] Andrea L Schaffer, Timothy A Dobbins, and Sallie-Anne Pearson. “Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions”. In: *BMC medical research methodology* 21.1 (2021), pp. 1–12.

[13] Terry E Duncan and Susan C Duncan. “A latent growth curve modeling approach to pooled interrupted time series analyses”. In: *Journal of Psychopathology and Behavioral Assessment* 26.4 (2004), pp. 271–278.

[14] Samuel Hawley et al. “Sample size and power considerations for ordinary least squares interrupted time series analysis: a simulation study”. In: *Clinical epidemiology* 11 (2019), p. 197.

[15] Sanford Weisberg. *Applied linear regression.* Vol. 528. John Wiley & Sons, 2005.

[16] Chris Chatfield. *The analysis of time series: an introduction.* Chapman and hall/CRC, 2003.

[17] A Ian McLeod and Evelyn R Vingilis. “Power computations for intervention analysis”. In: *Technometrics* 47.2 (2005), pp. 174–181.

[18] Anthony S Bryk and Stephen W Raudenbush. *Hierarchical linear models: Applications and data analysis methods.* Sage Publications, Inc, 1992.

[19] Per Engzell, Arun Frey, and Mark D Verhagen. “Learning loss due to school closures during the COVID-19 pandemic”. In: *Proceedings of the National Academy of Sciences* 118.17 (2021), e2022376118.

[20] Svenja Hammerstein et al. “Effects of COVID-19-related school closures on student achievement-a systematic review”. In: *Frontiers in Psychology* (2021), p. 4020.

[21] Andrew E Clark et al. “Compensating for academic loss: Online learning and student performance during the COVID-19 pandemic”. In: *China Economic Review* 68 (2021), p. 101629.

[22] Elisabeth Grewenig et al. “COVID-19 and educational inequality: How school closures affect low-and high-achieving students”. In: *European economic review* 140 (2021), p. 103920.

[23] Sally Galbraith, Jack Bowden, and Adrian Mander. “Accelerated longitudinal designs: An overview of modelling, power, costs and handling missing data”. In: *Statistical methods in medical research* 26.1 (2017), pp. 374–398.

[24] Joop J Hox, Mirjam Moerbeek, and Rens Van de Schoot. *Multilevel analysis: Techniques and applications.* Routledge, 2017.

[25] Niek Frans et al. “Defining and evaluating stability in early years assessment”. In: *International Journal of Research & Method in Education* 44.2 (2021), pp. 151–163.

[26] Douglas M Hawkins. “The problem of overfitting”. In: *Journal of chemical information and computer sciences* 44.1 (2004), pp. 1–12.

[27] Kevin J Grimm, Nilam Ram, and Ryne Estabrook. *Growth modeling: Structural equation and multilevel modeling approaches.* Guilford Publications, 2016.