



Midlands DSU Network
Decision Support Centre

Evaluating Complex Interventions using Statistical Models

Session 3. Quasi-Experimental methods: Difference-in-Difference and Interrupted Time Series

Anastasiia Zharinova, Healthcare Analyst at the Strategy Unit

Session structure

Session 3:

3.1 Course recap and difference-in-difference

3.2 Interrupted time-series: part 1

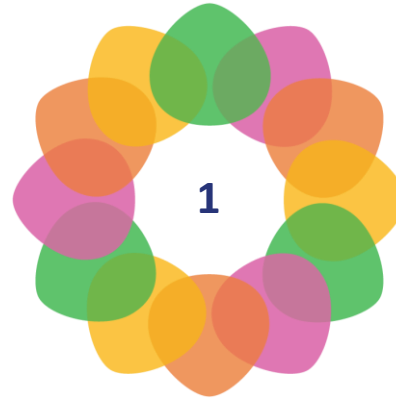
3.3 Interrupted time-series: part 2

In today's session 3.1

- 1 What have we learnt so far?
- 2 Difference-in-Difference (Dif-in-Dif)
- 3 Difference-in-Difference (Dif-in-Dif): practical example



SECTION



What have we learnt so far?

Glossary so far

Treatment group – group which receives ‘treatment’/were affected

Control group – group which receives no treatment/were not affected

Intervention – Can be programme or even some event

Correlation – relationship between two variables

Causation – process of causing something to happen or exist

Overfitting – model that corresponds too closely to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably

Statistically significant – result from data is not likely to occur randomly or by chance but is instead likely to be attributable to a specific cause

Significance levels – measure of the strength of the evidence that might be present in your data before you make a conclusion about statistical significance (e.g. 95%)

Exogeneity – factors in the model are not driven by other factors (observable or unobservable)

Endogeneity - factors in the model are driven by other factors (observable or unobservable)

Autocorrelation – relationship between the variable and its lagged version

Confounder – variable that influences both independent variable and dependent variable

Evaluation

Evaluation is a systematic approach to establish whether, why and how something is working (or not)

- Did our intervention reduce waiting times?
- Is drug A working? Does it work better than drug B?
- Did lockdown affect number of A&E attendances?

However, it is difficult to identify the effect of a particular event. Scientists use randomized control trials to compare treatment (intervention happened) and control (no intervention) groups. Is it possible in the above examples?

Quasi-experimental methods - Studies that can be used to estimate the causal impact of an intervention on an outcome in a similar manner to experimental designs but without the element of random assignment to treatment or control.

An intervention already happened – we cannot know for sure what have happened if the intervention did not happen and compare. But we can try to model this ‘what if’

Quasi-experimental methods

1. **Difference-in-Difference** - control group is well defined, so we can calculate the effect of the intervention (average treatment effect)
2. **Interrupted Time Series** - comparing actual trend with what could have happened if the intervention did not happen
3. **Panel Data** - mix of cross-sectional and time series data
4. **Propensity Score Matching and Retrospective Cohort Study** - matching treatment and control group based on certain parameters
5. **Synthetic Controls** - impossible to find a match for control group – so we need to create it synthetically
6. **Regression Discontinuity Design** – if an intervention criteria is well-defined, we compare the effect for those who are very close to the threshold
7. **Instrumental variables** – if allocation to the treatment is not random, we add extra variable (instrument) to redefine treatment

Regression Modelling

To find true causation, we need to build regression a model:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n + error$$

1. Ordinary least square – we find set of b that our error is as small as possible
2. Independent variables x can be continuous (age) or factor variables (sex, ethnicity, marital status). We might want to add some interaction terms and non-linear variables
3. Coefficient estimates and confidence intervals
4. To test how good the model is, we can calculate
 - R^2 - how much variation in y can be explained by variation in modelled $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$
 - Residual Standard Error (RMSE) – calculated using the difference between actual y and predicted y
 - F-statistic. Similar to p-value, but we test hypothesis that all regression coefficients are equal to 0.
5. If y is binary variable (1 or 0), we should use logit models (based on logistic distributions)
6. To interpret model, instead of using coefficients, we can calculate odd ratios and average marginal effect

Regression Modelling – time series

Time-series modelling is different from multilinear regression modelling. If we only have one time-trend and want to forecast future, we can use ARIMA (p,d,q) models

$$\Delta y_t = \mu + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \dots + \beta_p * y_{t-p} + e_t + \alpha_1 * e_{t-1} + \dots + \alpha_q * e_{t-q}$$

p – order of the lags of the y_t (AR element)

d – difference (I element)

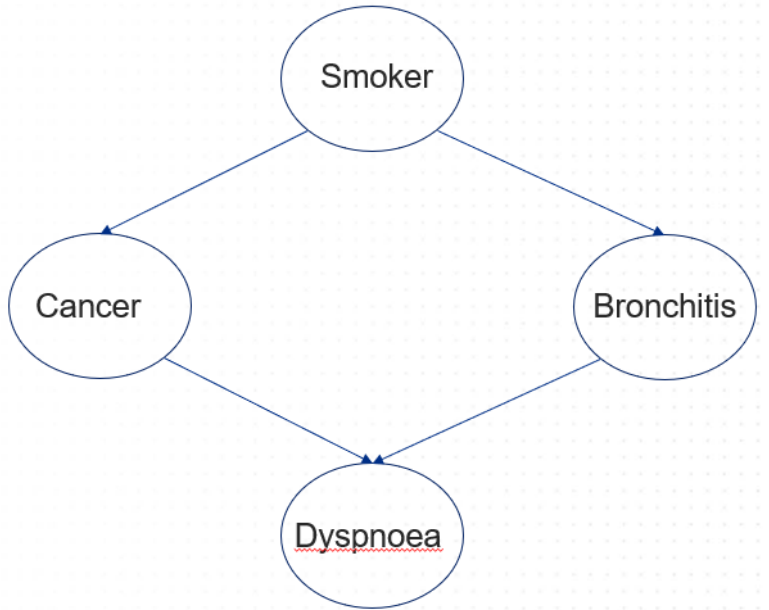
q – order of the lags of an error (MA element)

If our data has seasonality, we can build seasonal ARIMA - SARIMA(p,d,q)x(P,D,Q,s) where s is the seasonal length of the data (e.g 12 for monthly trend)

The model is good if:

- Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are as low as possible
- Model variance is as low as possible
- Residuals look like normal distribution and no lags are significant

Directed Acyclic Graphs (DAGs)



How to identify causal relationship between variables and build a DAG?

1. Constrain-based methods:

- PC algorithm - finds all conditional independence relationships
- Backdoor Criterion (Z is ancestor of X and/or Y)
- Frontdoor Criterion (Z intercepts path $X \Rightarrow Y$)
- Do-Calculus (if above criteria not applicable, then we can use set of three rules)

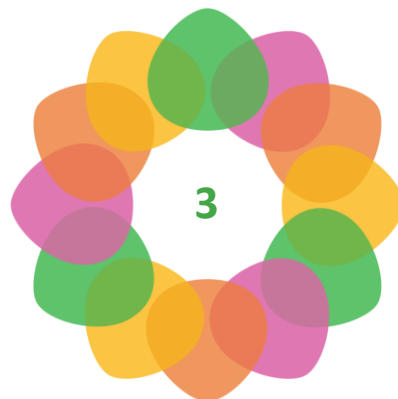
2. Score-based methods (search for different graphs and score them)

- Hill-climbing search (searching for higher score starting from best guess)
- Tabu (modification of hill-climbing search when we repeat hill-climbing for different starter graphs)

3. Bayesian Network



SECTION



Difference-in-Difference

Causal effect/treatment effect

What is the effect of the vitamin D on the severity of covid?

$Y_i(1)$ – severity of covid in person i if he/she was taking vitamin D

$Y_i(0)$ - severity of covid in person i if he/she was not taking vitamin D

Difference in covid severity as a result of taking vitamin D (treatment effect/causal effect)

$$Y_i(1) - Y_i(0)$$

However, statisticians are usually interested in average effect for the population/sample

$$E(Y_i(1) - Y_i(0)) - \text{average treatment effect}$$

- what is the average decrease in covid severity as a result of taking vitamin D?

Can we calculate $E(Y_i(1) - Y_i(0))$?

Not really, because a person cannot take and not take vitamin D at the same time.

How to calculate average treatment effect?

As we cannot calculate treatment effect and average treatment effect, we can try to compare the covid-19 severity in those who took vitamin D and those who did not.

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

Average level of outcome for those who took vitamin D

Average level of outcome for those who did not take vitamin D

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(Y_i(1) - Y_i(0) | D_i = 1) + E(Y_i(0) | D_i = 1) - E(Y_i(0) | D_i = 0)$$

Average
treatment effect

Selection bias

We are likely to have selection bias $\neq 0$ – people who have generally less chances of having covid (healthy lifestyle, living in affluent areas) are more likely to take vitamin D. So our estimate is not the same as average treatment effect and the model will be wrong.

If we cannot do a random experiment, we have to use quasi-experimental methods. For example, difference-in-difference

Intervention example

Let's assume that from 19/20 financial year Birmingham and Solihull STP has just started new Discharge to Assess programme to improve bed management and reduce delayed transfers of care (DTOC).

How to evaluate if it worked?

1. We could compare the number of DTOC before and after 1st of April 2019 (start of financial year). Any problems with it?

We don't know if the effect is due to programme or any other factors – time effects?

2. We could compare the number of DTOC in Birmingham and Solihull STP and in Herefordshire and Worcestershire STP in 19/20 financial year (after the former started the programme). Any problem with it?

The difference in DTOC might be explained by difference in demographics, demand and supply of social care and so on – STP effects?

So, both approaches are not great. But what if we combine them?

Difference in Difference (Dif-in-Dif) approach

$DTOC_{ist} = a_s + m_t + \delta + error$, where

$DTOC_{ist}$ - number of delayed transfers of care in hospital i in STP s at a time t

a_s - effect of the STP

m_t - effect of the year

δ - effect of the programme – this is what we are looking for

In Birmingham and Solihull STP:

before the Intervention: $E(DTOC_{ist} | s = treatment, t = before) = m_{before} + a_{treatment}$

after the intervention $E(DTOC_{ist} | s = treatment, t = after) = m_{after} + a_{treatment} + \delta$

after – before = $\Delta treatment = m_{after} + a_{treatment} + \delta - m_{before} - a_{treatment} = m_{after} - m_{before} + \delta$

Difference in Difference approach (cont)

In Herefordshire and Worcestershire STP:

before the Intervention: $E(DTOC_{ist} | s = control, t = before) = m_{before} + a_{control}$

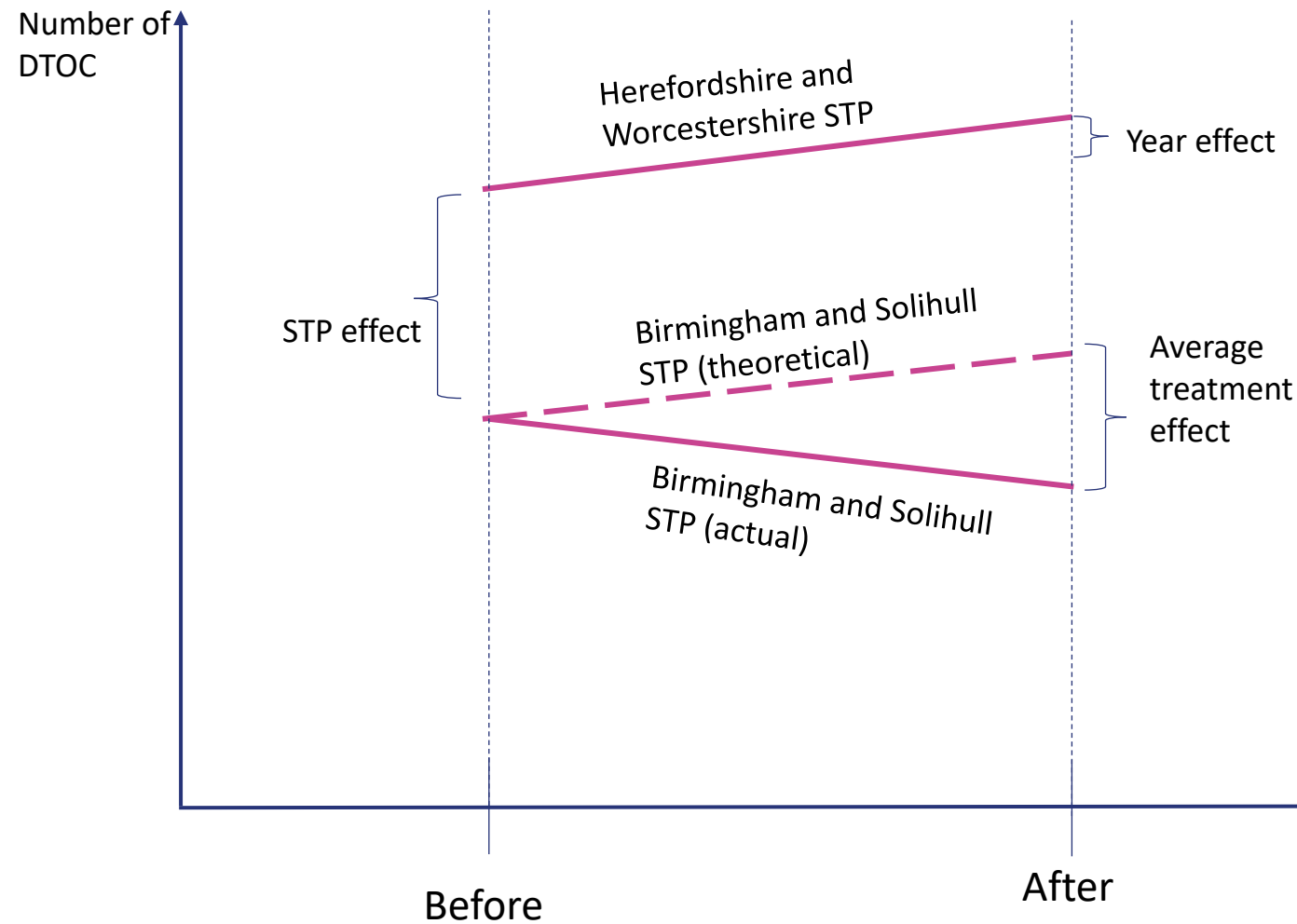
after the intervention $E(DTOC_{ist} | s = control, t = after) = m_{after} + a_{control}$

after – before = $\Delta control = m_{after} + a_{control} - m_{before} - a_{control} = m_{after} - m_{before}$

$$\Delta treatment - \Delta control = m_{after} - m_{before} + \delta - m_{after} + m_{before} = \delta$$

Therefore, difference in difference will give us the effect we are searching for!

Graphically



Empirically

	Group affected by an intervention (BSoL) – treatment group	Group not affected by an intervention (H&W) - control group
After the intervention	$Y_1(u_i) D_i = 1$	$Y_1(u_i) D_i = 0$
Before the intervention	$Y_0(u_i) D_i = 1$	$Y_0(u_i) D_i = 0$
	$E(Y_1(u_i) D_i = 1) - E(Y_0(u_i) D_i = 1)$	$E(Y_1(u_i) D_i = 0) - E(Y_0(u_i) D_i = 0)$

$$DD = E(Y_1(u_i)|D_i = 1) - E(Y_0(u_i)|D_i = 1) - E(Y_1(u_i)|D_i = 0) - E(Y_0(u_i)|D_i = 0)$$

	Group affected by an intervention (BSoL) – treatment group	Group not affected by an intervention (H&W) - control group
After the intervention	1,450	2,200
Before the intervention	1,500	2,000
	-50	200

$$DD = -50 - 200 = -250$$

In regression terms

$DTOC_{ist} = b_0 + b_1 * x_i + b_2 * z_t + \delta * x_i * z_t + error$, where

x_i - dummy variable: 1 if hospital is in Birmingham & Solihull STP (treatment group)

z_t - dummy variable: 1 if observation in the period after the intervention

δ – intervention effect

We can use ordinary least square method to get coefficients estimates, including standard errors.

We can also add control variables - what could they be in our case?

Parallel trend

Any problems with this method?

We only look at pre and post time points and assuming that the treatment and control group were on the same trend of the outcome variable before the policy change – “parallel trends” assumption. We can test this assumption using ‘placebo test’

1. Create a ‘fake’ treatment group:

- Pretend that the policy change was at (t-1) and redo the regression
- Use as a treatment group a group that you know was not affected (Black Country and West Birmingham STP)

2. Use different control group

3. Use the outcome that you know was not affected

If these give you insignificant effect, it is safe to use Dif-in-Dif

Panel data methods can be used to look at more individual effects (more about it in September)

Other considerations

Difference-in-Difference is a good method when we only have aggregated data, e.g. averages for different groups – which might be the case in health and social care data.

Limitations:

Dif-in-Dif attributes any difference in trends between the treatment and control groups, that occur at the same time as the intervention, to that intervention. We don't consider situations when:

- Outcome changes over time due to factors specific to the treatment group
- Outcome trend jumps only for the treatment group exactly at the time when the treatment kicks on, but happened because of the other factors
- There is a delay in the impact of policy change – we need to find and use lagged impact variables
- Intervention got implemented in the area due to problems in outcomes (e.g. BSoL got picked because DTOCs are too high) – although placebo test should check for it

References

Difference in Difference studies

1. [Dimick, J.B. and Ryan, A.M., 2014. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama*, 312\(22\), pp.2401-2402.](#)
2. [Card, D. and Krueger, A.B., 1993. Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania.](#)
3. [Hanchate, A.D., Kapoor, A., Katz, J.N., McCormick, D., Lasser, K.E., Feng, C., Manze, M.G. and Kressin, N.R., 2015. Massachusetts health reform and disparities in joint replacement use: difference in differences study. *bmj*, 350.](#)
4. [Rosenthal, M.B., Alidina, S., Friedberg, M.W., Singer, S.J., Eastman, D., Li, Z. and Schneider, E.C., 2016. A difference-in-difference analysis of changes in quality, utilization and cost following the Colorado multi-payer patient-centered medical home pilot. *Journal of general internal medicine*, 31\(3\), pp.289-296.](#)

More about the method

1. [Mostly harmless econometrics – chapter 5.2](#)
2. [LSE lecture](#) by textbook author
3. More on math and assumptions – healthpolicydatascience.org

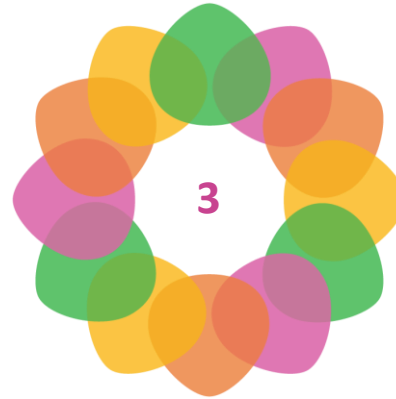
In today's session 3.2

1 Interrupted Time Series: theory

2 Interrupted Time Series: practice



SECTION



Interrupted Time Series (ITS)

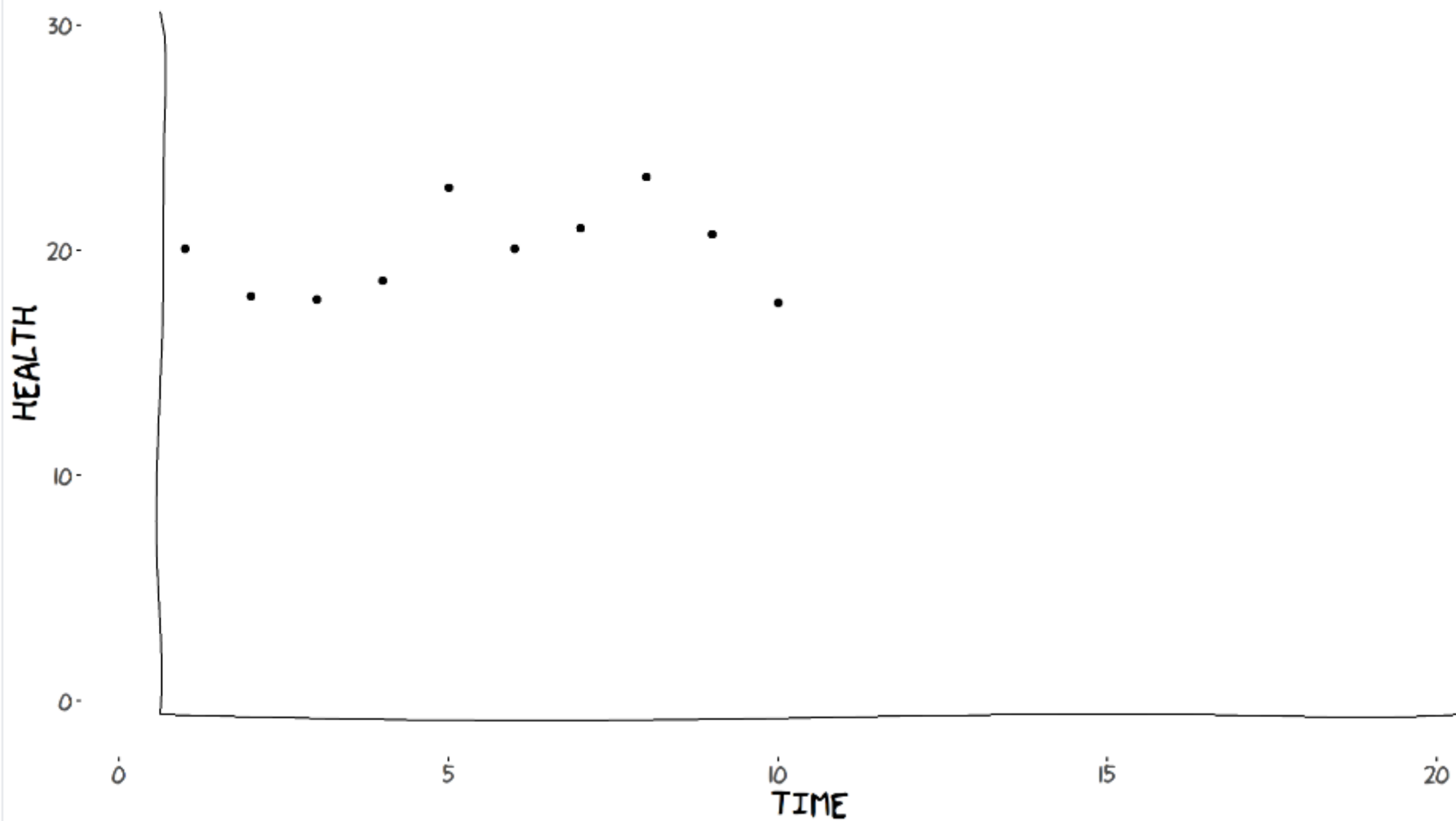
Definition

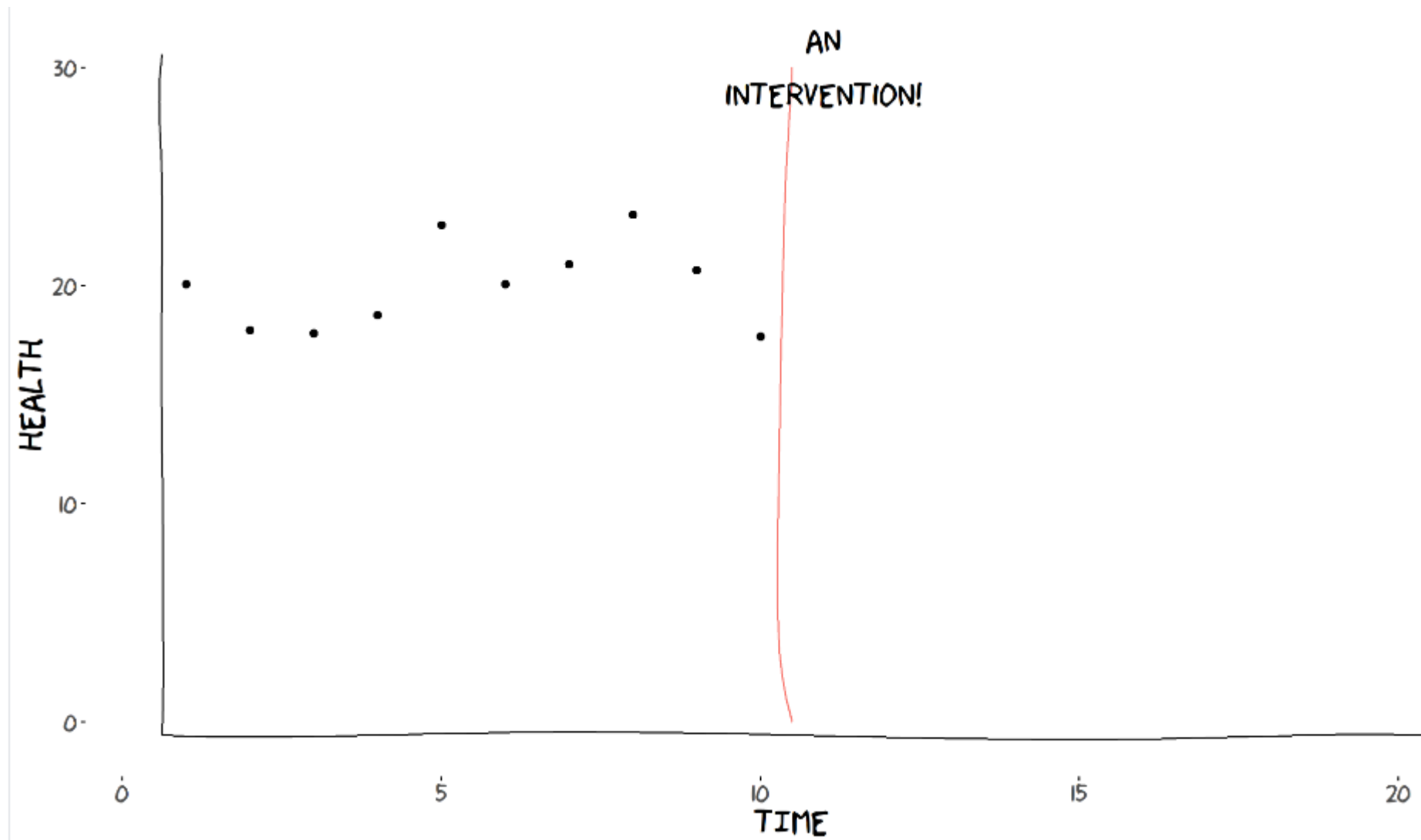
ITS is a quasi experimental study that compares observations at multiple time points before and after an intervention (“interruption point”) to detect any significant effect in the trend.

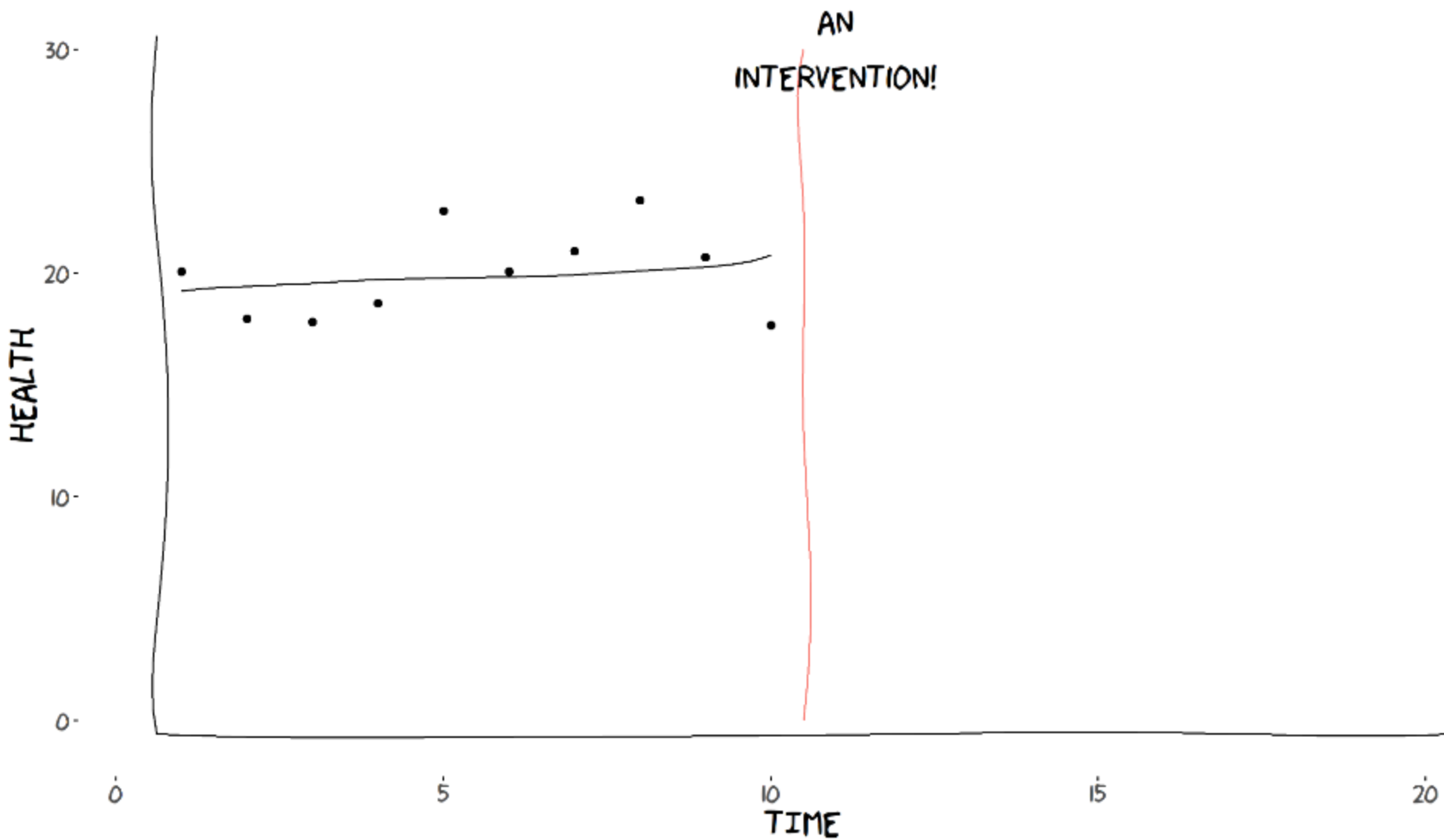
When to use?

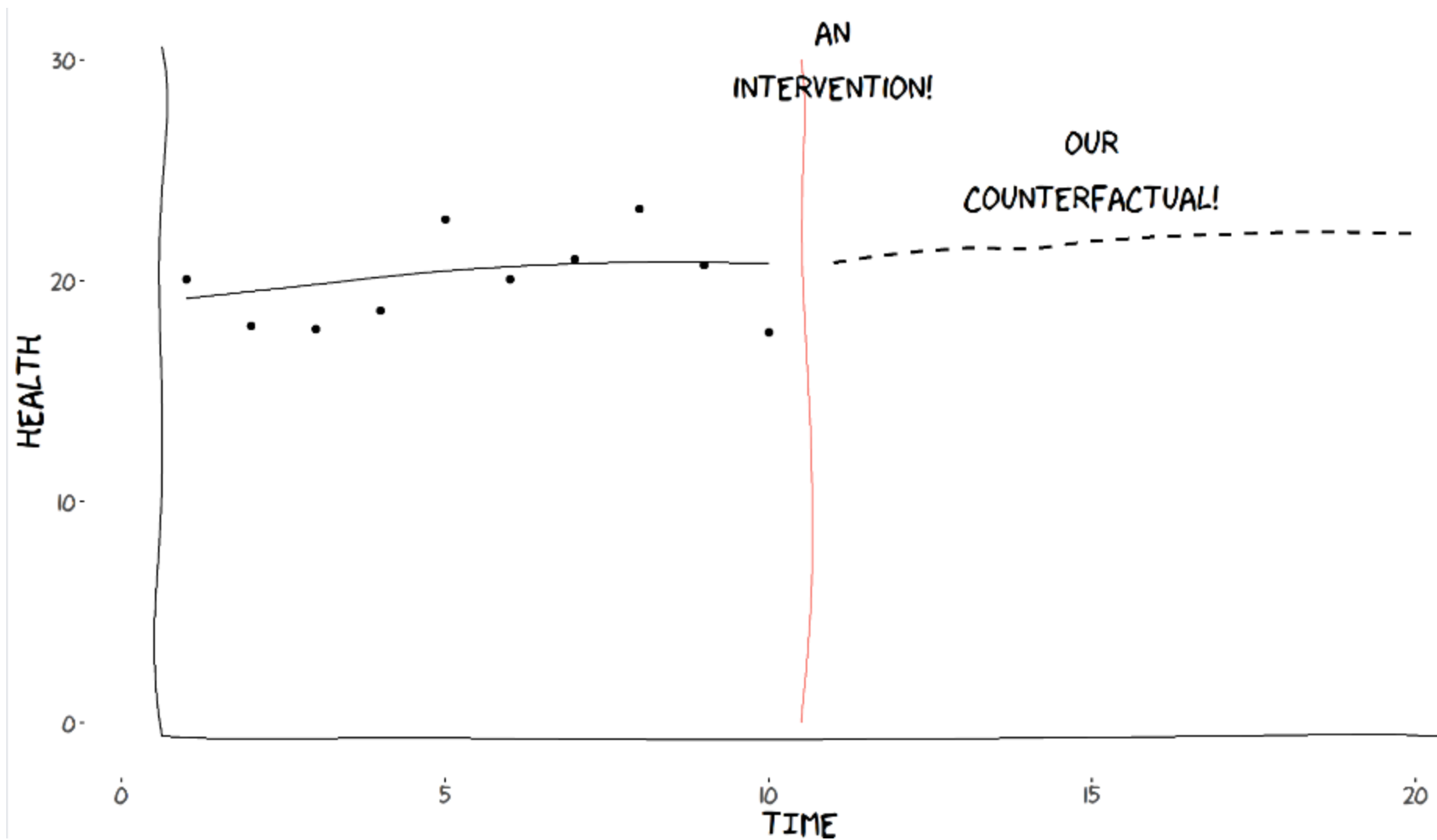
1. We had very limited data – only one time series
2. We want to understand how and if the outcome has changed after an intervention, a policy, or a program that was implemented for the full population at one specific point in time.
3. We want to present a straightforward method to our stakeholders

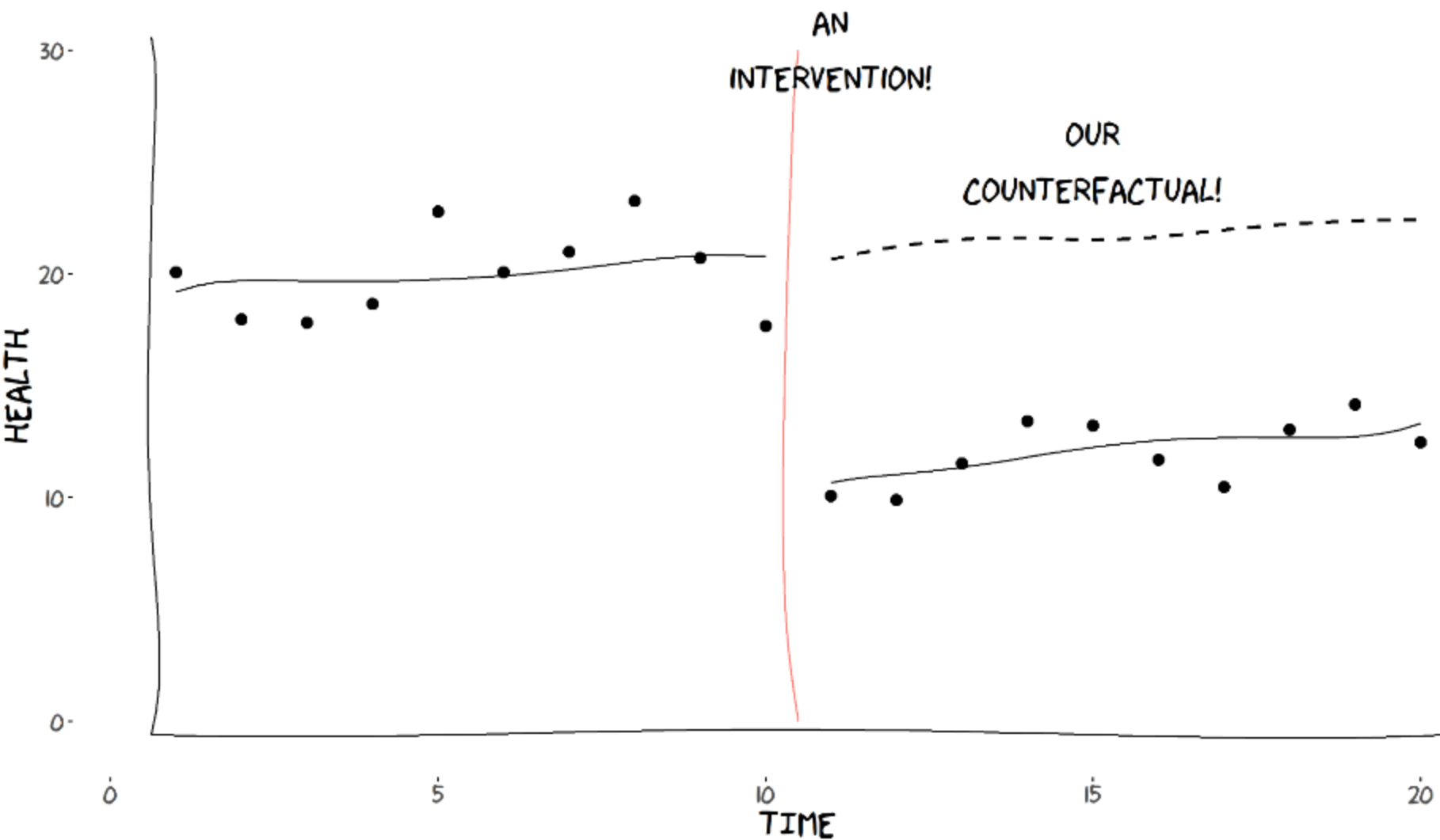
Data	Research Question
Monthly A&E data	How did Covid-19 pandemic changed the A&E attendances
Daily number of missed outpatient appointments	How did digital programme changed DNAs?
Weekly number of mental health referrals	How did new Mental health referral service changed the demand?

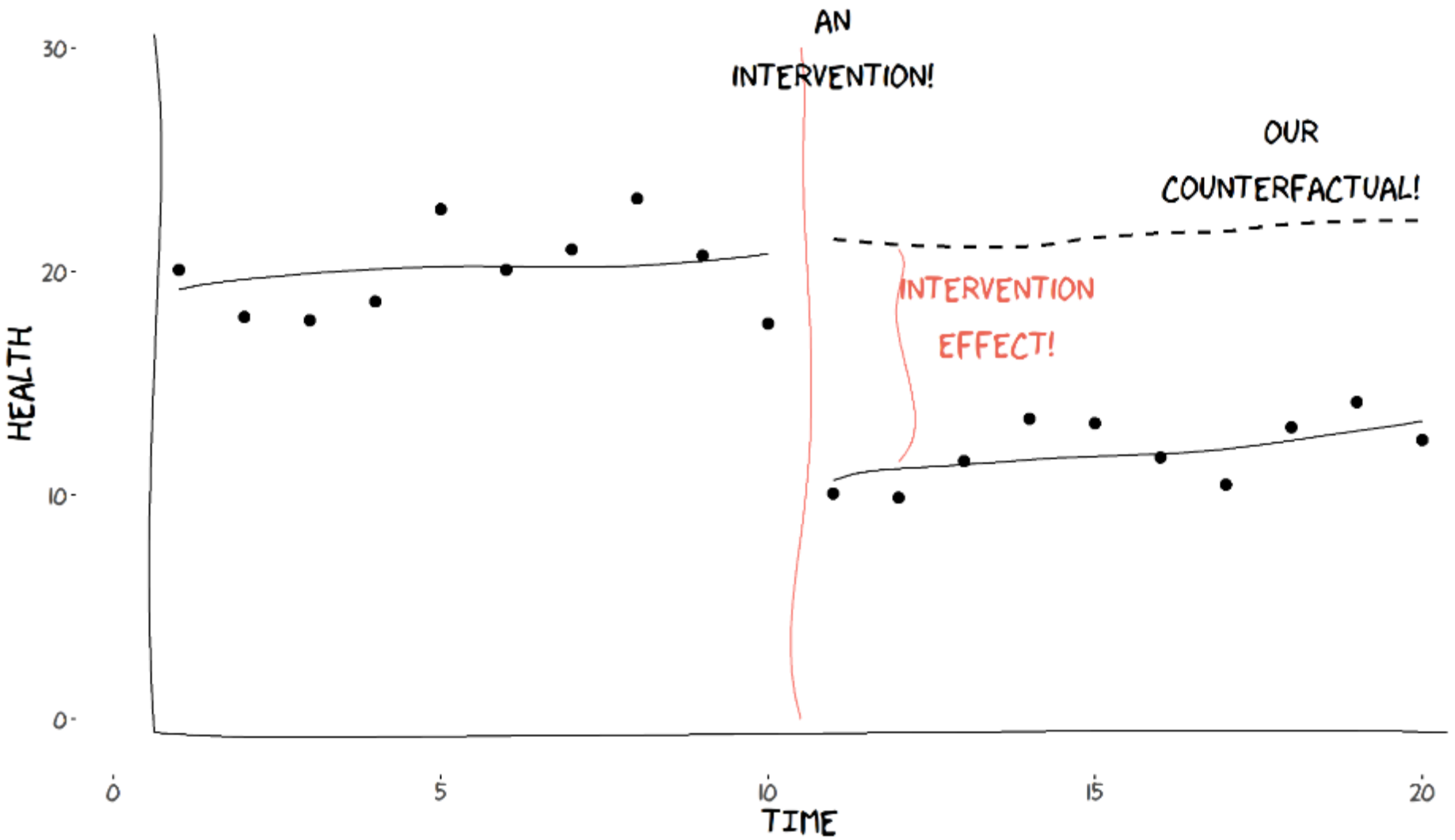












Types of impact model

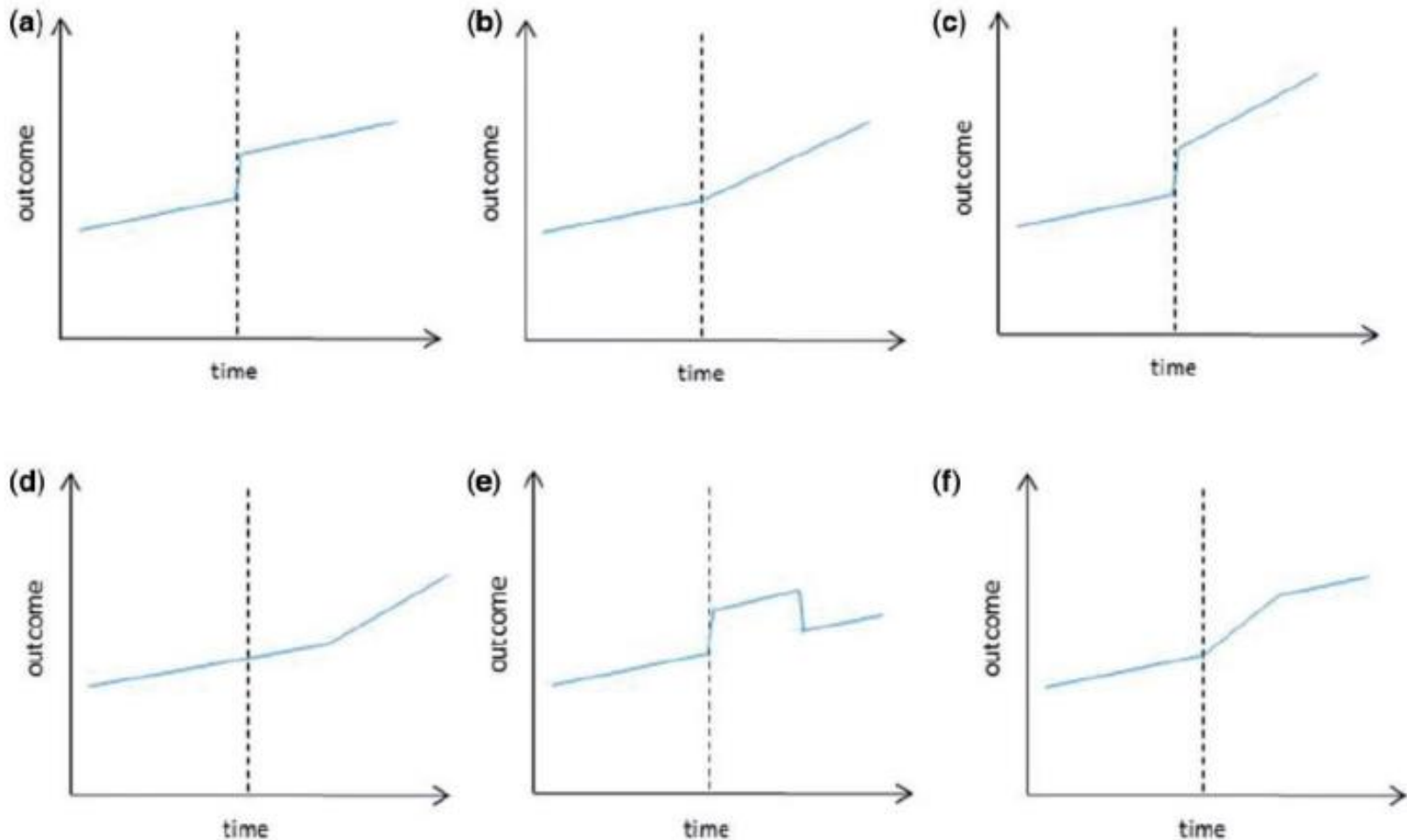


Figure 2 Examples of impact models used in ITS

(a) Level change; (b) Slope change; (c) Level and slope change; (d) Slope change following a lag; (e) Temporary level change; (f) Temporary slope change leading to a level change.

How to build ITS?

1. Prepare time series data – check for missing data, data quality, etc
2. Build a counterfactual forecast with a model of choice
3. Compare observed post-intervention values with counterfactual post-intervention values (modelled in step 2)
4. Check if the intervention had a significant effect on the outcome of interest (if possible)
5. Select a point in time and calculate effect:
 - As an absolute difference
 - As ratio of estimated and counterfactual values
6. Transfer effect into monetary terms (if applicable)

In regression terms

$Y = b_0 + b_1 * T + b_2 * D + b_3 * P + error$, where


Y – outcome variable

T - time (e.g., days, months, years...) passed from the start of the observational period

D – intervention dummy - indicating observation collected before (=0) or after (=1) the policy intervention

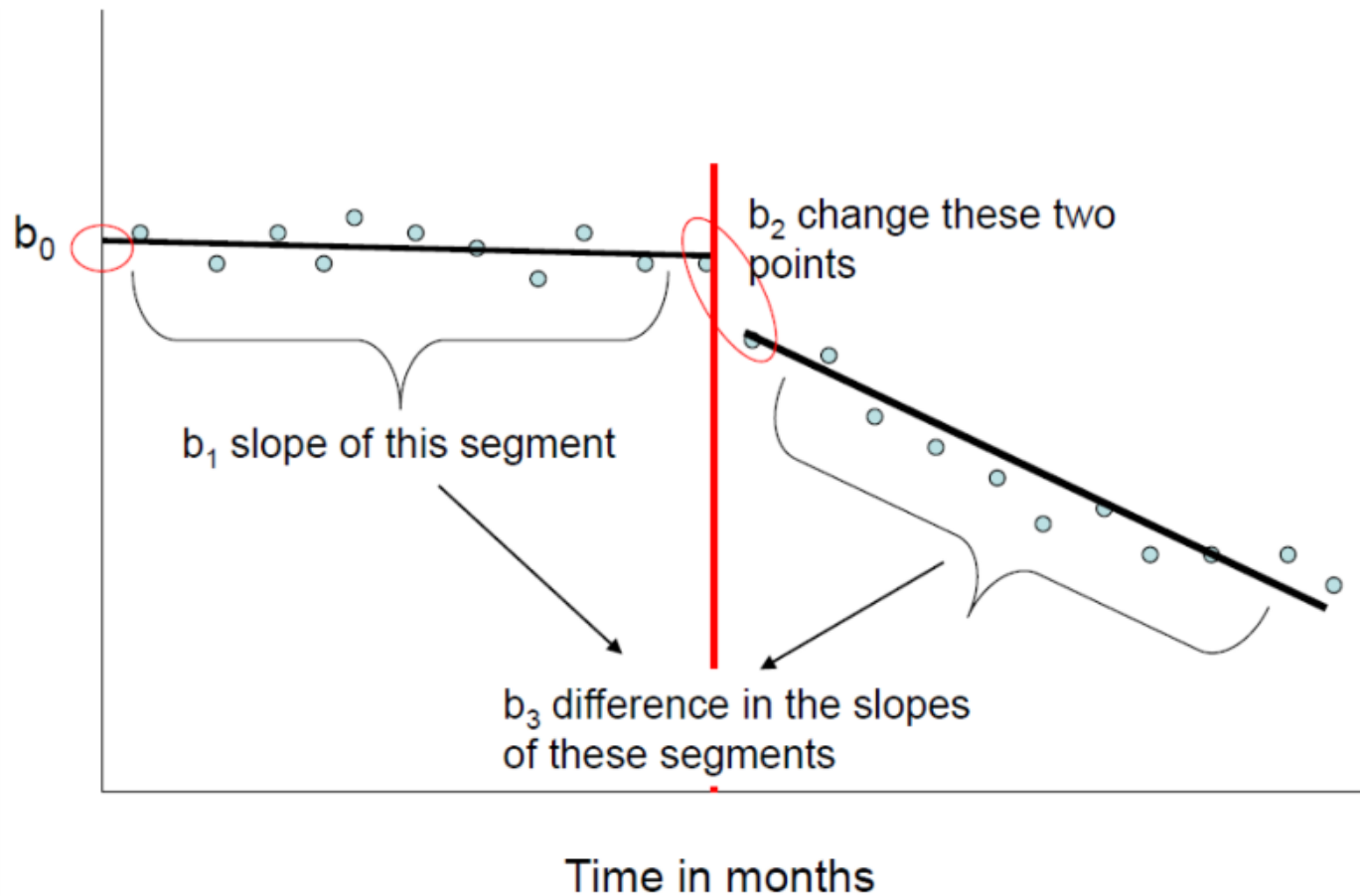
P - is a continuous variable indicating time passed since the intervention has occurred

Y	Time (T)	Treatment (D)	Time since (P)
10	1	0	0
12	2	0	0
14	3	0	0
11	4	1	1
8	5	1	2
5	6	1	3
2	7	1	4



In regression terms

$$Y = b_0 + b_1 * T + b_2 * D + b_3 * P + error$$

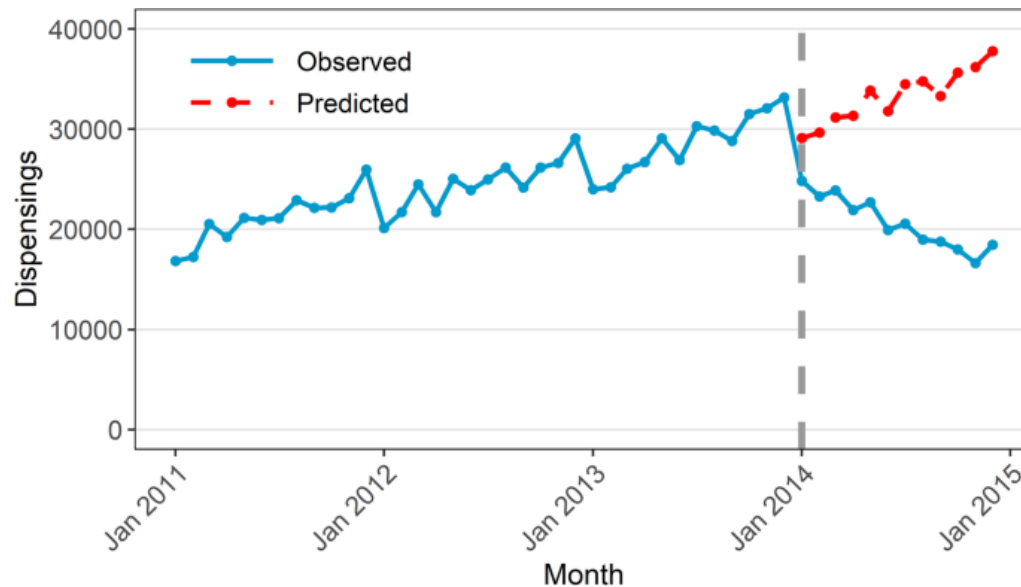


Sometimes being referred to as a segmented regression

In regression terms

ARIMA and SARIMA Models can also be used to find our 'counterfactual'. To evaluate an intervention using ITS and ARIMA, you need to:

1. Explore dataset, identify properties, data quality problems and parameters
2. Forecast 'counterfactual' scenario using usual ARIMA algorithm or auto.arima function
3. Evaluate the quality of created forecast
4. Compare forecast with what have actually happened
5. Translate the findings into monetary terms if needed (e.g. number of beds days into savings)



[Interrupted time series analysis using autoregressive integrated moving average \(ARIMA\) models: a guide for evaluating large-scale health interventions](#)

In regression terms

R also has CausalImpact package which was developed and widely used by Google team. This package is using Bayesian structural time-series (BSTS) models to estimate the causal effect of a designed intervention on a time series.

$Y = \mu_t + x_t\beta + S_t + error$, where

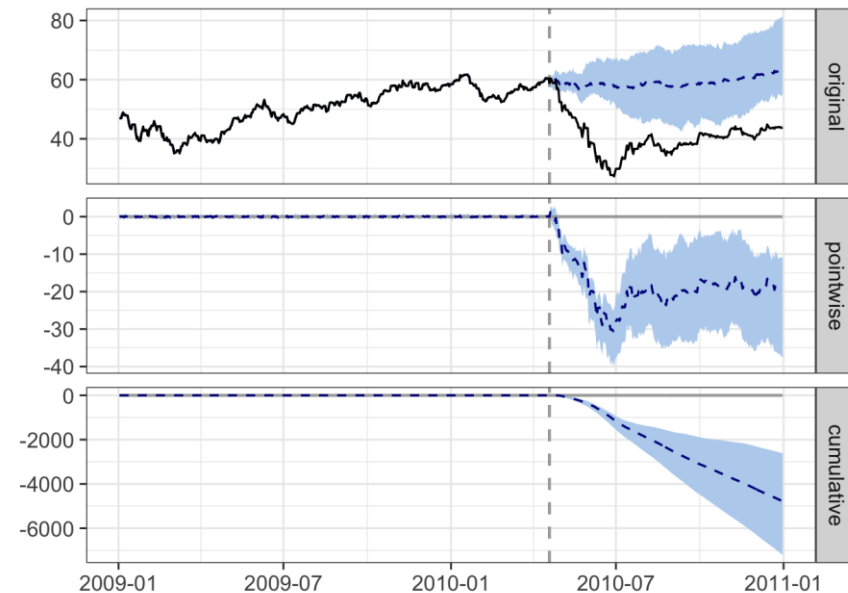
x_t - set of regressors

S_t - seasonality element

μ_t - local level term/unobserved trend

We can add as many seasons to the model as we want (e.g. daily data)

This package is straightforward to use and it creates counterfactual forecast automatically. However, you can also create a custom model.



In regression terms

On Thursday, we will be looking at Generalized Linear Models. Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for variables that have error distribution models other than a normal distribution.

As we are mostly interested in the level and slope and do not need detailed forecast, we can use linear models when building ITS models. GLM models allow to add seasonality and other control variables to the model.

We will be looking at Poisson regression which is classically used for count data and for the situation when the outcome cannot be less than 0. Similarly to logistic regression, it uses logarithm as the main function.

$$\begin{aligned}y_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \alpha + \beta x_i \\ \mathbb{E}[y_i] &= \exp(\alpha + \beta x_i)\end{aligned}$$

Overdispersion at Poisson regression models

OPTIONAL

One of the assumptions in the Poisson regression model is that the mean equals variance (hence the outcome value cannot fall below 0). However, over- or underdispersion happens in Poisson models, where the variance is larger or smaller than the mean value, respectively.

In R, dispersion test can be run using extra packages such as AER (function – `dispersiontest`). Alternatively, we can get dispersion element of the glm model. If this indicator equals 1, mean equals variance and we don't have overdispersion problem.

To fix overdispersion, quasi-functions can be used. Quasi-Poisson distribution is a type of Poisson distribution which assumes that mean of the outcome is the same as its variance.

Advantages and limitations

Advantages

1. Minimal data requirements
2. Works well with the interventions introduced on population level
3. Not technically demanding – can be performed with any model of choice (linear, ARIMA, SARIMA, BSTS...)
4. Intuitive graphical representation
5. Works well with the outcomes which are expected to change rapidly
6. Sometimes it is the only possible method!

Limitations

1. Requires clear differentiation between pre-intervention and post-intervention period
2. Assumes that characteristics of population remain unchanged throughout the study period
3. Works the best if the intervention occurred in the middle of time series, so rapid evaluation might not be possible
4. Problems with time series analysis:
 - Autocorrelation – standard linear model might not work
 - Seasonality – monthly, annual, moving seasons (Easter)

Potential mitigations

1. Adding control series to control for other changes/events
 - If possible, adding control group – cohort with the same characteristic which did not go through the intervention
 - If adding control group is not possible, it can be represented by a different outcome which is not expected to be affected by the intervention
2. Using changepoint analysis to identify the 'interruption' point
3. Adding control variables – possible in BSTS models

References

Interrupted time-series studies

1. Hawton, K., Bergen, H., Simkin, S., Dodd, S., Pocock, P., Bernal, W., Gunnell, D. and Kapur, N., 2013. Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in England and Wales: interrupted time series analyses. *Bmj*, 346.
2. Campbell, D.T. and Ross, H.L., 1968. The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review*, pp.33-53.
3. Dennis, J., Ramsay, T., Turgeon, A.F. and Zarychanski, R., 2013. Helmet legislation and admissions to hospital for cycling related head injuries in Canadian provinces and territories: interrupted time series analysis. *Bmj*, 346.
4. Taljaard, M., McKenzie, J.E., Ramsay, C.R. and Grimshaw, J.M., 2014. The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care. *Implementation Science*, 9(1), pp.1-4.

References

R tutorials

1. Step-by-Step ITS guide - <https://ds4ps.org/pe4ps-textbook/docs/p-020-time-series.html>
2. Information on CausalImpact package - <https://google.github.io/CausalImpact/CausalImpact.html>
3. BSTS package - <https://cran.r-project.org/web/packages/bsts/bsts.pdf> (if you need to create a custom model)
4. ITS.analysis package - <https://cran.r-project.org/web/packages/its.analysis/its.analysis.pdf> (I have never tried it!)

Some open-source NHS data examples to practice: RTT waiting times, DTOC, A&E data, NHS Digital staff data, etc.