The Strategy Unit.

Midlands DSU Network
Decision Support Centre

Evaluating Complex Interventions using Statistical Models

Session 4. Panel data

Session structure

Session 4:

4.1 Introduction to Panel data

4.2 Data preparation

4.3 Panel Data Methods

4.4 Practical exercises

Midlands DSU Network
Decision Support Centre

SECTION

1

What have we learnt so far?

Glossary so far

Treatment group – group which receives ‘treatment’

Control group – group which receives no treatment

Intervention – Can be programme or even some event (e.g. covid-19 pandemic)

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

Evaluation using statistical models

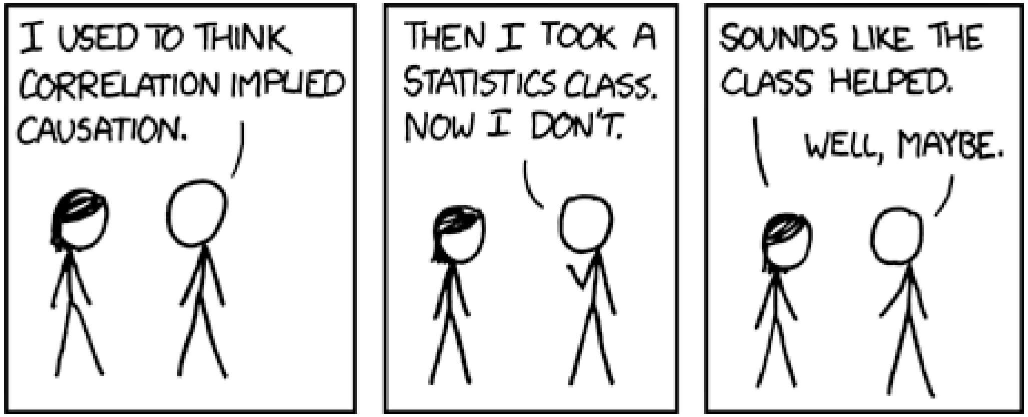
Evaluation is a systematic approach to establish whether, why and how something is working (or not). In an ideal world, scientists would use randomised control trial to evaluate an intervention/programme. However, it is not always possible.

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.

Such methods include:

1. Difference-in-Difference
2. Interrupted Time Series
3. Panel Data
4. Matching
5. Synthetic Controls
6. Regression Discontinuity Design

7.Instrumental variables


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Regression modelling

Regression modelling - investigates the relationship between a dependent (outcome) variable and independent variable(s). 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 Linear and non-linear models (e.g. logistic regression)

Time series modelling - modelling the sequence of data points by time unit (e.g. for the purpose of forecasting). Usual linear models are not useful due to autocorrelation.

- Auto Regressive Integrated Moving Average (ARIMA)
- Seasonal Auto Regressive Integrated Moving Average (SARIMA)
- Bayesian structural time-series models (BSTS)
- Generalised linear models (GLM) – poisson regression

Directed Acyclic Graphs - methods of finding causal relationships between variables:

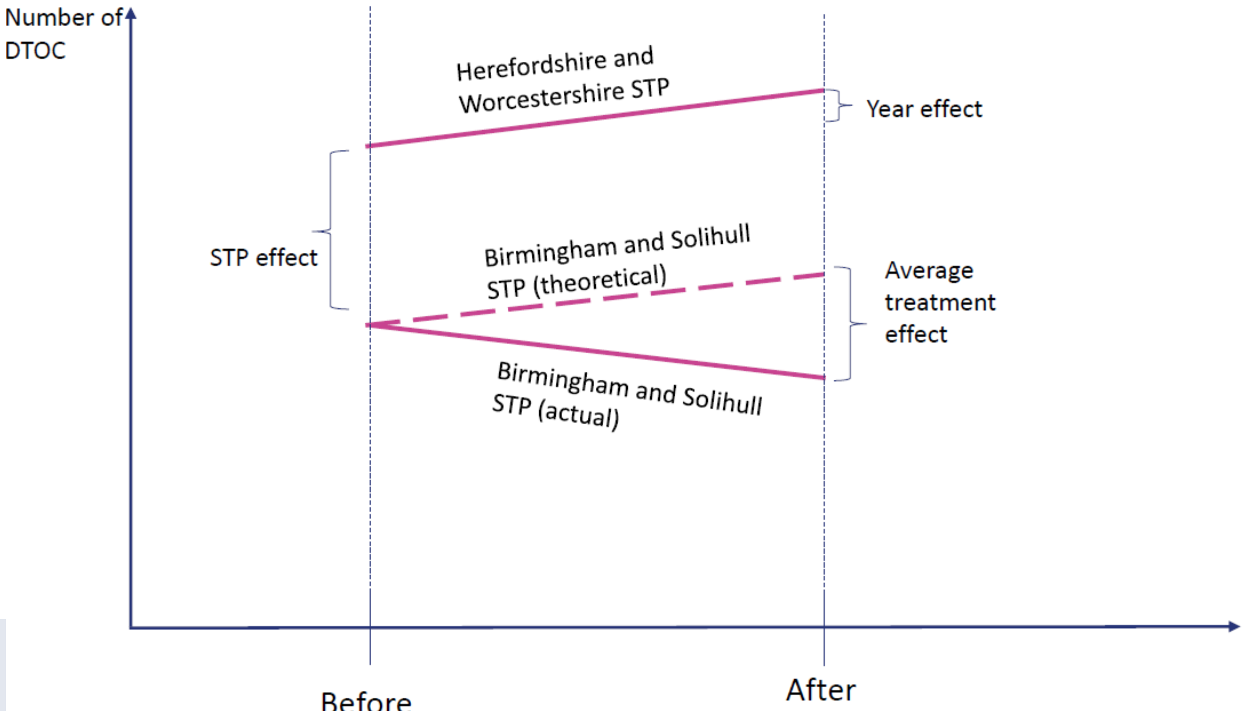
- Constrain-based methods
- Score-based methods (search for different graphs and score them)

Quasi-experimental methods we learnt so far: difference-in-difference

Control group is well defined, so we can calculate the effect of the intervention (average treatment effect)

Pros: minimal data requirements, simplicity, can be calculated without the regression

Cons: parallel trend assumption, control group is needed, does not attribute for changes that happened at the same time as intervention and selection bias

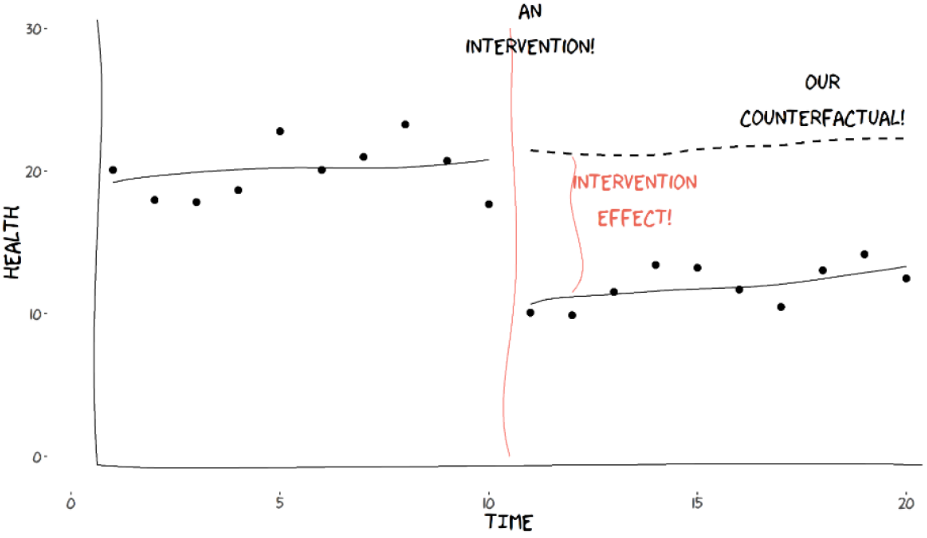

Number of DTOC
Herefordshire and Worcestershire STP
Year effect
STP effect
Birmingham and Solihull STP (theoretical)
Average treatment effect
Birmingham and Solihull STP (actual)
Before
After

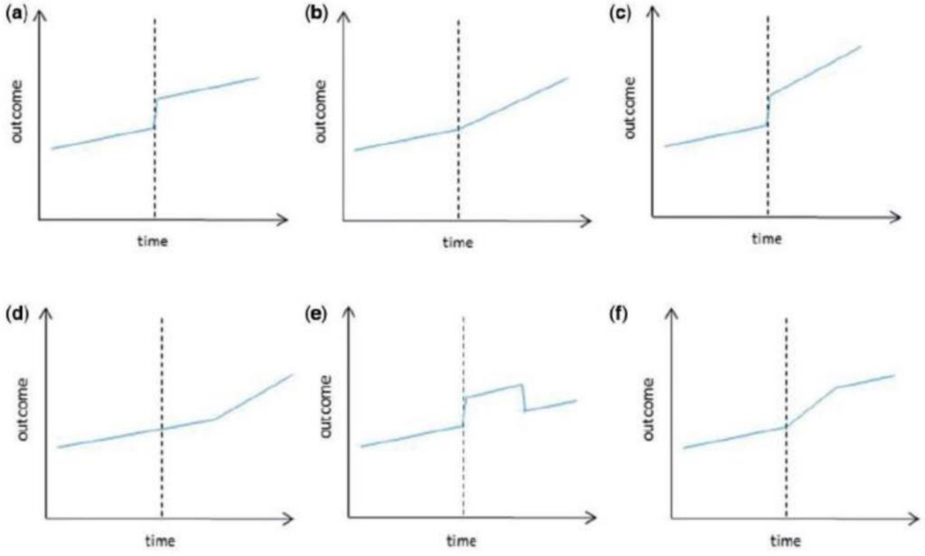
Quasi-experimental methods we learnt so far: interrupted time series

Comparing actual trend with what could have happened if the intervention did not happen

Pros: simple visual representation, different methods can be used (ARIMA, BSTS, GLM), minimal data requirements, works with population-level intervention, various causal relationships (level-change model VS slope-change models)

Cons: we need to know exact time of an intervention, does not account for other changes, outcomes are expected to change rapidly, intervention needs to occur in the middle of time series





SECTION

2

Introduction to Panel data

Definition

Panel data (longitudinal data) – multi-dimensional data involving measurements over time.

For example,

- Survey data (Understanding society, Study of Global Ageing, etc)
- Data on healthcare activity by NHS number for the last 5 years
- Delayed Transfers of Care by NHS trust over last 10 years
- SUS data
- Workforce datasets

Time-series data

Month	Number of A&E attendances in Trust A
Jan 2018	20,000
Feb 2018	21,000
March 2018	19,000
April 2019	20,000
May 2019	18,000
…	…

t observations

Cross-sectional data

Trust	Region	Number of A&E attendances in Trust at time x
A	Midlands	20,000
B	Midlands	17,000
C	Midlands	19,000
D	South-East	15,000
E	South-East	13,000
F	South-East	16,000
…	…	…

n observations

Definition

In this example, we are having panel data – observations for a Trust i (1, 2, …. n) over the t periods of time (1, 2, … T). So Number of A&E attendances in Trust C can be expressed as

$AE_{3,1}$ or $AE_{C,\text{Jan 2018}}$

Advantages:

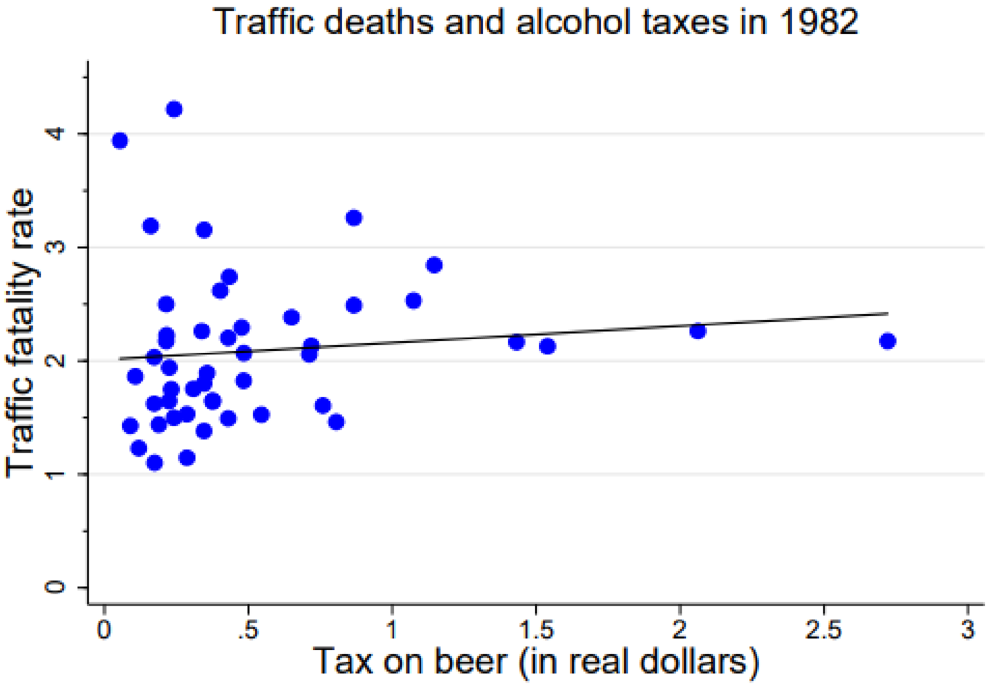
1. Can control for factors that time series and cross-sectional data cannot control for (which results in bias) :
   - Controls for time effects
   - Controls for trust/object effects
2. Can help in exploring the dynamics of variables for many different objects
3. Can minimize estimation biases that may arise from aggregating groups into a single time series.
4. Many observantions (n*t)
5. More control over omitted variables

Month (t)	**Trust (i)**	**Region**	**Number of A&E attendances in Trust at time x**
Jan 2018	A	Midlands	20,000
Jan 2018	B	Midlands	17,000
Jan 2018	C	Midlands	19,000
Jan 2018	D	South-East	15,000
Jan 2018	E	South-East	13,000
Jan 2018	F	South-East	16,000
Feb 2018	A	Midlands	21,000
Feb 2018	B	Midlands	20,000
Feb 2018	C	Midlands	19,000
Feb 2018	D	South-East	18,000
Feb 2018	E	South-East	20,000
Feb 2018	F	South-East	19,000
…	…	…	…

Alcohol tax and road traffic deaths

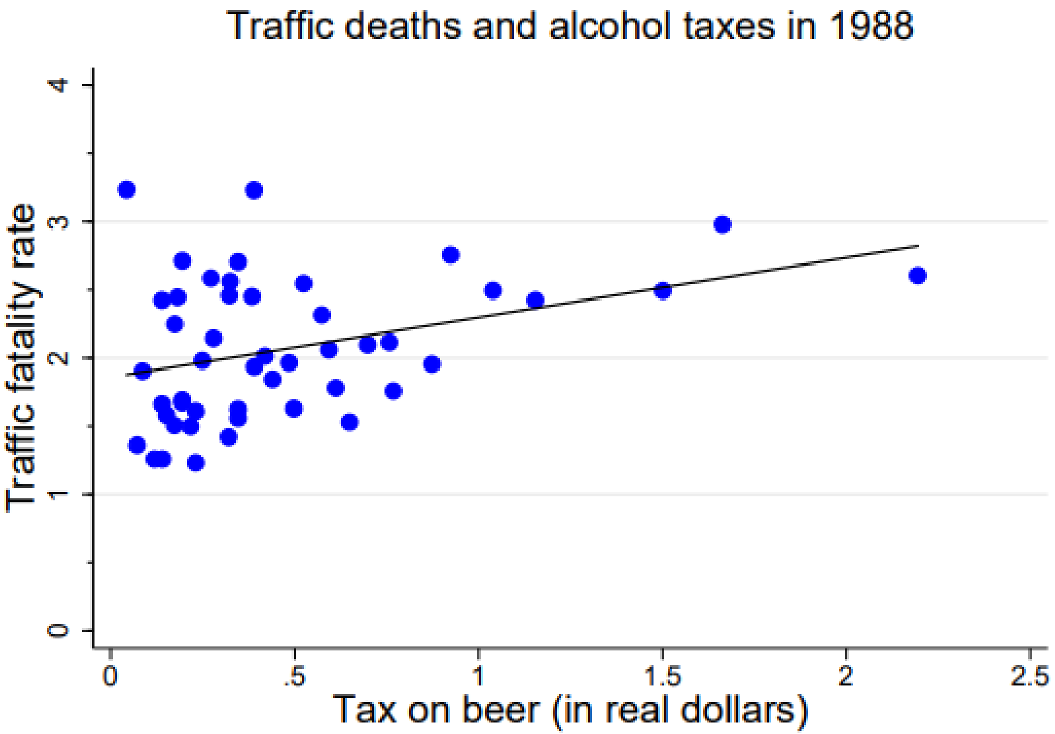
Government wants to reduce traffic fatalities and knows that about 25% of the fatal crashes involve driver who drank alcohol. What is the effect of increasing the tax on beer on the traffic fatality rate?

We have data on traffic fatality rate and tax on beer for 48 U.S. states in 1982 and 1988.



$$\widehat{\text{FatalityRate}}_{i,1982} = \underset{(0.14)}{2.01} + \underset{(0.18)}{0.15} * BeerTax_{i,1982}$$

Alcohol tax and road traffic deaths



$$\widehat{\text{FatalityRate}}_{i,1988} = \underset{0.11}{1.86} + \underset{0.16}{0.44} * BeerTax_{i,1988}$$

Alcohol tax and road traffic deaths

Increase in beer tax increases the number of road traffic deaths? Possibly, we don’t have some control variables we need.

$FatalityRate_{it} = b_0 + \delta * t + b_1 * BeerTax_{it} + \gamma * Z_i + e_{it}$, where

Z_i - characteristics of State i (alcohol consumption, road quality, etc)

This variable is unobservable, so we cannot include it to the regression. As the result, we have Omitted variable bias.

These characterises are likely to stay constant over time. What happens if we subtract one equation from the other?

Alcohol tax and road traffic deaths

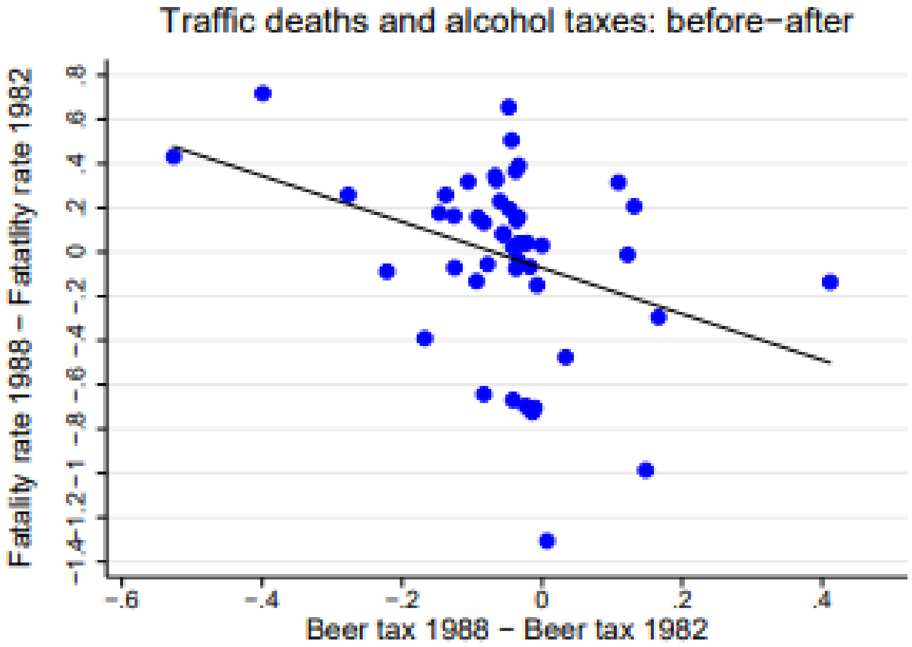
$$FatalityRate_{i,1} = b_0 + \delta * 1 + b_1 * BeerTax_{i,1} + \gamma * Z_i + e_{i1}$$
$$FatalityRate_{i,2} = b_0 + \delta * 2 + b_1 * BeerTax_{i,2} + \gamma * Z_i + e_{i2}$$

$$FatalityRate_{i,2}\text{-}\ FatalityRate_{i,1} = \delta\text{+ } b_1\text{*}(BeerTax_{i,2}\text{-}\ BeerTax_{i,1})\text{+ } e_{i2}\text{-}\ e_{i1}$$

So, instead of building a regression model for Fatality rate and Beer Tax variables, we can do it for $\Delta FatalityRate$ and $\Delta BeerTax$.

$\Delta FatalityRate$= δ+ b_1* $\Delta BeerTax$+error

$\Delta FatalityRate = -0.072$ -1.04 *$\Delta BeerTax$



Alcohol tax and road traffic deaths

Now the results look more reasonable: increase in beer tax decreased the road traffic fatality rate. The method we used is called **first-difference estimator**.

We will learn more methods on Monday:

1. Pooled OLS
2. Fixed Effects model:
   - First-difference estimator
   - OLS with dummy variables
   - Within-group estimator (also known as fixed effects model)
3. Random Effects model

PS Difference-in-difference is one of the examples of panel data

Difference-in-difference

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

	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

Types of panel data

Balanced panel data

Month (t)	Trust (i)	Number of surgeons available in Trust at time x, FTE	Number of planned operations in Trust at time x
Jan 2018	A	200	10,000
Jan 2018	B	100	7,000
Jan 2018	C	120	8,000
Jan 2018	D	150	5,000
Jan 2018	E	175	3,000
Jan 2018	F	200	6,000
Feb 2018	A	210	12,000
Feb 2018	B	70	2,000
Feb 2018	C	100	9,000
Feb 2018	D	120	8,000
Feb 2018	E	50	2,000
Feb 2018	F	100	9,000
…	…	…	…

Unbalanced panel data

Month (t)	Trust (i)	Number of surgeons available in Trust at time x, FTE	Number of planned operations in Trust at time x
Jan 2018	A	200	10,000
Jan 2018	B	100	7,000
Jan 2018	C	120	8,000
Jan 2018	D	150	5,000
Jan 2018	E	…	…
Jan 2018	F	200	6,000
Feb 2018	A	210	12,000
Feb 2018	B	…	…
Feb 2018	C	100	9,000
Feb 2018	D	120	8,000
Feb 2018	E	50	2,000
Feb 2018	F	100	9,000
…	…	…	…

Analysis is easier when panel data is balanced. However, we can treat unbalanced panel data as balanced if we know that omission of the data is random.

When would you use panel data?

1. You have repeated measurements over time (day, week, month, year)
2. Data is for multiple entities (CCGs, Trusts, patients)
3. You have an outcome variable of interest and list of control variables in the dataset

To establish the impact of one variable, but these sit in a complex causal web with a number of other factors, which may or may not be observable.

Any examples?

SECTION

3

Data preparation

Why is it important?

As evaluators, we need our evaluation data to be:

1. Accurate
2. Complete
3. High quality
4. Reliable
5. Unbiased
6. Valid

However, this might be a problem in healthcare research. If data does not meet our criteria, the following problems occur:

1. Biased conclusions (we just went through this example on beer tax and road traffic accidents)
2. Increased error
3. Reduced generalisability

Brief Introduction to the 12 Steps of Data Cleaning (Morrow, 2013).

1. Explore structure and summarise

You need to start from understanding the data structure. SUS datasets and most longitudinal surveys will have data dictionaries. If you are linking a number of datasets, you might have to create a Data Codebook.

In R, you can use function (str) to see which variables you have and what are their limits.

You might also want to summarise the dataset using describe() or summary() functions. This will allow you to get the first impression on how skewed your data is and whether there are any obvious data quality issues.

Frequencies analysis (absolute values, %) can also be performed to check how balanced the panel data is.

Year	N of observations/IDs
2017	10,000
2018	8,000
2019	9,000

Year	N of observations/IDs
2017	10,000
2018	2,000
2019	10,000

2. Detect coding mistakes

Coding errors are any values that are not within the specified range for your variable. E.g:

- Length of stay < 0
- Unrealistic age or income
- Sex=3 in SUS (where it can only be 1 for male and 2 for female)

Coding mistakes can be seen from the summary statistics (minimum, maximum, mean) but can also be detected visually.

In many cases, errors are unspecified missing values (e.g. in SUS missing ethnicity value can be recorded as 99, 999, NA or Z)

3. Modify and create variables

Variables in the dataset are not always ones you want to use in the analysis. Some of the most common changes that might be needed:

1. Change variables from numeric to factors
2. Perform simple math changes (e.g move percentages into proportions or summarise two sources of income)
3. Regroup categorical variables (e.g. modify the marital status variable to decrease a number of categories)
4. Create new variables (e.g. whether the admission was a readmission based on the records).
5. Add dummy variable of an intervention (more about it on Monday)

Any examples in your project?

In R, you can use dplyr package and mutate() function to perform most of the above modifications

5. Search for outliers

Outlies can be detected:

1. Visually – using plots, histograms and boxplots (ggplot2 package)
2. Statistically – calculating z-scores. In R, this can be performed
   - Manually, using the formula $z = \frac{x-\mu}{\sigma}$, where μ is the mean of the sample and σ is the standard deviation
   - Automatically, using the package effectsize and the function standardize()

What to do with the outliers?

1. Some outliers are the coding mistakes (e.g. unreasonably high age), so need to be deleted/replaced with the NA
2. You can also transform outliers (e.g. replace negative income with the minimum income)

6. Detect and impute missing values

You can detect missing values visually or check number of missing values using is.na() function

You should always check if the missing data is random or non-random (e.g. there is a pattern)

Solutions:

1. Do nothing, R can handle the regression with the missing values
2. Analysis on non-missing values only (analysis on observables)
3. Mean values imputation (does not work on the categorical features)
4. Imputation using most frequent values (can introduce bias in the data)
5. Multiple data imputation (filling the missing data multiple times based on the other variables and their relationship with the variable of interest)

Dealing with the missing values: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779
More on multiple imputation: https://stats.idre.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/

SECTION

4

Panel data methods

Pooled OLS model

Simpliest method – not recognise the panel structure of the data and run OLS model

Any problem with this method?

- Standard OLS would assume no correlation between unit i's observations in different periods (no individual effects)
- It also assumes no correlation between different units in the same period

Month (t)	Trust (i)	Number of surgeons available in Trust at time x, FTE	Number of planned operations in Trust at time x
Jan 2018	A	200	10,000
Jan 2018	B	100	7,000
Jan 2018	C	120	8,000
Jan 2018	D	150	5,000
Jan 2018	E	175	3,000
Jan 2018	F	200	6,000
Feb 2018	A	210	12,000
Feb 2018	B	70	2,000
Feb 2018	C	100	9,000
Feb 2018	D	120	8,000
Feb 2018	E	50	2,000
Feb 2018	F	100	9,000
…	…	…	…

Fixed effects model – Least Squares Dummy Variables model (LSDV)

Usual panel data equation: $Y_{i,t} = b_0 + b * x_{i,t} + \gamma * Z_i + e_{i,t}$

Now, we introduce dummy variables:

D_{1t} - 1 for the first object and 0 for all others

D_{2t} - 1 for the second object and 0 for all others

and so on…

How will the regression look?

We can replace Z_i - object characteristics with the set of created dummies. However, we will need to remove an intercept b_0 to avoid multicollinearity

$$Y_{i,t} = b \quad * x_{i,t} + \mathrm{a}_1 * \mathrm{D}_{1t} + \mathrm{a}_2 * \mathrm{D}_{2t} + \cdots \mathrm{a}_n * \mathrm{D}_{nt} + e_{i,t}$$

Alternatively, we can add an intercept but remove one of the dummy variable (similarly to how we did it with cross-sectional data)

Fixed effects model – Least Squares Dummy Variables model (LSDV)

Interpretation and assessment of such model will be similar to usual OLS for the cross-sectional data:

1. We can estimate the significance of variables and interpret results
2. We can assess the quality of the model using usual statistical tests and R^2

For such models, it is usually called LSDV-R^2

3. We can add different variables in and interaction terms
4. We can compare different models

However, there are a number of problems:

1. We might need to add too many dummies in, which requires high computing power
2. Every variable x needs to be changing over time

Fixed effect model – Within Estimator

Usual panel data equation: $Y_{i,t} = b_0 + b * x_{i,t} + \gamma * Z_i + e_{i,t}$

For each time t

$$Y_{i,1} = b_0 + b * x_{i,1} + \gamma * Z_i + e_{i,1}$$
$$Y_{i,2} = b_0 + b * x_{i,2} + \gamma * Z_i + e_{i,2}$$

and so on

$$Y_{i,T} = b_0 + b * x_{i,T} + \gamma * Z_i + e_{i,T}$$

If we summarise all equations

$$\sum Y_{i,t} = \mathrm{T} * b_0 + b * \sum x_{i,T} + \mathrm{T} * \gamma * Z_i + \sum e_{i,T}$$

and then divide by T

$$\bar{y}_i = b_0 + b * \overline{x_i} + \gamma * Z_i + \overline{e_i}$$

Where $\bar{y}_i, \overline{x_i}$ and $\overline{e_i}$ are averages

Fixed effect model – Within Estimator

Subtracting new average equation from the initial panel data equation will remove individual object effects

$$Y_{i,t} - \bar{y}_i = b_0 + b * x_{i,t} + \gamma * Z_i + e_{i,t} - b_0 - b * \overline{x_i} - \gamma * Z_i + \overline{e_i}$$

$$(Y_{i,t} - \bar{y}_i) = b * (x_{i,t} - \overline{x_i}) + (e_{i,t} - \overline{e_i})$$

$$\widetilde{yit} = b \cdot \tilde{x}_{it} + \tilde{\varepsilon}_{it}$$

If individual unobserved effects were the only cause of bias, new model will give us correct estimates. Such models are called within-group estimator.

Luckily, R can do it automatically using plm package

Fixed effect model – Within Estimator

Within-group estimator/within estimation model is identical to the LSDV estimation. However, R^2 will be different (R^2 – within). LSDV-R^2 is usually very high due to many independent variables.

Similarly to LSDV, within estimation is not able to assess variables which are constant over time (e.g. sex or year of birth)

To choose between fixed effect models and pooled OLS regression, we need to run LSDV model and test hypothesis that all dummy variable coefficients (a_1 , a_2..., a_n) equal to 0.

$$Y_{i,t} = b * x_{i,t} + \mathrm{a}_1 * \mathrm{D}_{1t} + \mathrm{a}_2 * \mathrm{D}_{2t} + \cdots \mathrm{a}_n * \mathrm{D}_{nt} + e_{i,t}$$

If they are, then there are no individual effects and pooled OLS regression will work fine.

Time fixed effects

In addition to object effects we can also include time effects in the model. Time effects control for omitted variables that are common to all objects but vary over time (e.g. total government spend on healthcare sector)

$$Y_{i,t} = b_0 + b * x_{i,t} + \mathrm{W}_t + e_{i,t}$$

Where W – time effect

Similarly to LSDV, we can create set of dummy variables

$$Y_{i,t} = b * x_{i,t} + \mathrm{a}_1 * \mathrm{B}_{1t} + \mathrm{a}_2 * \mathrm{B}_{2t} + \cdots \mathrm{a}_n * \mathrm{B}_{nt} + e_{i,t}$$

Where B_{1t} equals 1 at first time period and 0 otherwise,

B_{2t} equals 1 at second time period and 0 otherwise

Model, which has both time effects and individual effects is called two-way linear fixed effects regression (and can be build in plm package via setting method argument to twoways)

Random Effects model

$$Y_{i,t} = b_0 + b * x_{i,t} + u_i + e_{i,t}$$

This model assumes that the individual effects are random (u_i) and not correlated with the regressors in the model. For example, in the beer tax and road fatalities example, we would have to assume that state effects are random and not correlated with the tax on alcohol.

We can also rewrite the regression equation as

$$Y_{i,t} = b_0 + b * x_{i,t} + w_{i,t}$$

As random error (w) is not correlated with regressors, we can use OLS (in fact, slightly adjusted version - generalized method of moments). It works well in the survey data – as we selected individuals randomly, we can assume that the model errors are not correlated with individual characterictics.

Random effects models can also be built using plm() function and specifying the model argument to "random"

Random effects VS fixed effects VS pooled OLS

Random effects model has a big advantage - we can include variables, which do not change over time. However, there is a statistical way to choose a better model.

Hausman test can estimate if the model errors (w) are correlated with the independent variables (x). If they do, we need to choose Fixed effects model.

If we want to choose between random effects model and pooled OLS regression, we can use **Breusch-Pagan** test (estimates, if individual effects are present).

As discussed earlier, random effects model will work well when we have random sample of individuals (e.g. in survey data). However, it does not work with the different data – e.g. if we have data on local authorities in England, so for region/country data we usually use fixed effects models.

All tests can be run in R using plmtest() function.

Mixed effects model

Mixed effects model assumes that both fixed effects and random effects are present in the model.

$$Y_{i,t} = b_0 + \gamma * Z_i + u_i + e_{i,t}$$

Where Z_i - unobserved fixed effects

u_i - unobserved random effects

These models are used when we have data with more than one source of random variability. For example, an outcome may be measured more than once on the same person (repeated measures taken over time). This method is using maximum likelihood estimates and takes longer to run. In addition, it does not work well if we want to coefficients for regressors and might involve clustering

In R, they can be estimated using lme4 package.

Logistic regression for panel data

If dependent variable in the panel data can only be 1 or 0 (e.g. event of death, cardiac arrest, etc), we still can use random effects or fixed effects method.

Two most common models are:

1. Fixed effects logit
2. Random effects probit

However, we can always do pooled model (treat panel data as cross-sectional data).

In R, this can be done in ‘pglm’ package

References

Healthcare research involving the panel data:

1. Seamer P, Brake S, Moore P, Mohammed MA, Wyatt S, Did government spending cuts to social care for older people lead to an increase in emergency hospital admissions? An ecological study, England 2005–2016, BMJ Open, 25 April 2019
2. Lightwood, J. and Glantz, S.A., 2016. Smoking behaviour and healthcare expenditure in the United States, 1992–2009: Panel data estimates. *PLoS medicine*, *13*(5), p.e1002020.
3. Pop-Eleches, C., 2006. The impact of an abortion ban on socioeconomic outcomes of children: Evidence from Romania. *Journal of Political Economy*, *114*(4), pp.744-773.
4. Propper, C., Burgess, S. and Green, K., 2004. Does competition between hospitals improve the quality of care?: Hospital death rates and the NHS internal market. *Journal of Public Economics*, *88*(7-8), pp.1247-1272.
5. Mast, B.D., Benson, B.L. and Rasmussen, D.W., 1999. Beer taxation and alcohol-related traffic fatalities. *Southern Economic Journal*, pp.214-249.

References

Theory of panel data and practice in R:

1. Mostly harmless econometrics (Angrist and Pischke) – Chapter 5
2. Lecture notes on Panel data (University of Maryland) - http://econweb.umd.edu/~chao/Teaching/Econ423/Econ423_Panel_Data.pdf
3. Introduction to econometrics with R (chapter 10) - https://www.econometrics-with-r.org/10-1-panel-data.html
4. A guide on data analysis (chapter 18) - https://bookdown.org/mike/data_analysis/panel-data.html
5. plm package - https://cran.r-project.org/web/packages/plm/plm.pdf
6. pglm package - https://cran.r-project.org/web/packages/pglm/pglm.pdf