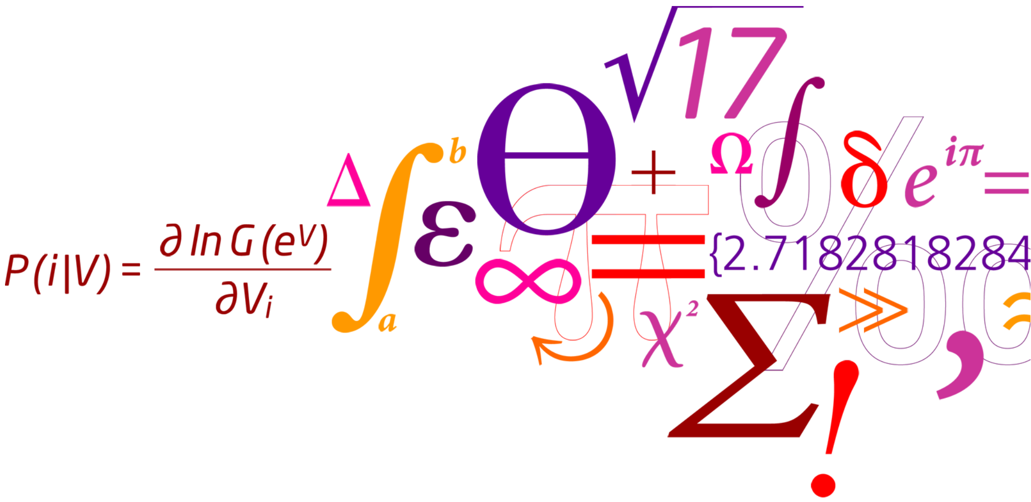
42588 – Data and data science

Week 3 – Surveys and data

14th of February 2024



DTU Management Engineering
Department of Management Engineering

Today’s program

- Survey design by Sonja
- Group work
- Briefly on data types
- Work on Project 1

The course plan

Week	Date	Subject/Lecture	Literature	Exercises	Teachers
1	31/1	Introduction + questions and data	AoS chap. 3	Form groups + week 1 exercise	Stefan
2	7/2	Basics on data and variables	AoS chap. 1-2 (+ OM 1)	Project 1 – start	Stefan/Guest from Genmab
3	14/2	Surveys + data types + experimental data	Paper 1 (+ OM 2-5)	Project 1 – work	Sonja / Stefan
4	21/2	Governance + causality	Paper 2 + AoS chap. 4 (+ OM 6)	Project 1 – deadline	Hjalmar / Stefan
5	28/2	More on data, e.g. real-time data, online data	Paper 3 (+ OM 7-10)	Discuss data for project 2	Guido/ Stefan
6	6/3	Visualisation	Chap. 1,5,6,7,10,23, 24,29 in Wilke + (AM 1-2)	Integrated exercises + work on project 2	Mads
7	13/3	Spatial data	Chap. 1,14 in Gimonds	Week 7 exercises + work on project 2	Mads / Guest from Niras
8	20/3	Imputation/weighting/presentation proj. 2	Paper 4	First deadline of project 2 + Week 8 exercises	Mads
9	3/4	Data analytics I	ISL ch. 3 + paper 5	Work on project 3a	Stefan
10	10/4	Data analytics II	ISL ch. 6	Work on project 3a	Stefan
11	17/4	Data analytics III	ISL ch. 4	Work on project 3b	Stefan
12	24/4	Data analytics IV	TBD	Work on project 3b	Stefan
13	1/5	Summary and perspective	Paper 6	Project 3 – deadline	Stefan

Feedback on last week

- Diskrete vs. kontinuerte variable
- Hvornår benyttes kovarians/correlation?
- Betinget vs. ubetinget – hvad er mest korrekt?
- Kode + simulation af data + dinosaur plot
- Mere om projekterne

Data classification

- Cross-section data are data where we have observations that are not related through unobserved factors

ObsID	Var1	…	VarK
1	X_11	…	X_1K
2	X_21	…	X_2K
…	…		…
N	X_N1	…	X_NK

- Such data can be modelled using linear regression, lasso and ridge regression, logistic regression, trees, support vector machines and other types of classical models.

- Data could be one day travel diaries, wage data across individuals (specific year), GDP data across countries (specific year).

Data classification

- Time-series data are data where we have many observations from the same individual or unit so there is definitely a relation between observations.

ObsID	TimeID	Var1	…	VarK
1	1	X_11	…	X_1K
1	2	X_21	…	X_2K
…	…	…		…
1	N	X_N1	…	X_NK

- Such data can be modelled using linear regression with serial correlation, time-series models, specific neural networks, and other types of models that do take the correlation into account.

- Data could be one year travel diary for one person, GDP for one country over many years, waiting time at the airport, house consumption of natural gas.

Data classification

- Panel data are data where we have several observations from the same individual or unit so there is definitely a relation between observations.

ObsID	TimeID	Var1	…	VarK
1	1	X_111	…	X_11K
…	…			
1	T	X_1T1	…	X_1TK
…	…			
N	1	X_N11	…	X_N1K
…	…	…		…
N	T	X_NT1	…	X_NTK

- Such data can be modelled using panel versions of linear regression, specific neural networks, and other types of models that do take the panel correlation into account.

Data classification

- Panel data could be one week of travel diaries from many individuals, GDP for some countries over many years, counting stations for bike flow all C25 stock measured daily in a year, or air pollution from various stations across a city.

- In general, panel data can become very complicated to work with if both N and T are large at least in classical models. Some machine learning models might be better if prediction is the purpose.

- On the other hand for N large and T small, panel data can give many insights into changes in behaviour that cannot be analysed using either cross-section or time-series data.

Data types

- Observational data are registrations of something that happens in the real world. Revealed preference (RP) data is a common word used in social sciences for observational data when the observation represents a decision or choice.

- The benefit is that the data measure some aspect of reality. This does not guarentee validity but it can support it.

- RP data may also be problematic in some contexts, e.g.
  - they are restricted to options and variables that exist or have existed historically,
  - variables may have little variation in real markets, e.g. prices in some markets
  - variables tend to be correlated in the real world. This makes it difficult to disentangle the effect of specific variables

Data types

- Stated preference (SP) data is a common word used in social sciences for hypothetical data related to choices.

- These data are reactions or answers to hypothetical situations/questions that happens in a hypothetical setting.

- The problem is that the data are hypothetical and hence may not match reality.

- SP data have some advantages, e.g.
  - they can include novel options and variables that do not exist yet,
  - variables can have more variation than in a real market, e.g. prices
  - variables can be designed to be uncorrelated or close to this. This allows us to disentangle of the effect of various variables

Feedback

- Final questions

  1. What was the most interesting you learned during the lecture?

  2. What is your most important unanswered question based on the lecture?

- Group 1 (Caroline V., Johanne, Nadia) should send/hand in their feedback to Stefan. Everyone else are very welcome to give feedback as well!

For next time

- Read for this week
  - Slides + Lietz, P. (2010)

- The other papers should be seen as supplementary reading.

- To prepare for lecture 4, you should read
  - Wilkinson et al. (2022)

- Work on Project 1. Do not leave today before, I know your topic for project 1.

- Note that the deadline for project 1 is 26/2.