Lecture 3 - Scientifically Analyzing Data

Dr. L.S. Sanna Stephan

2023

RUG Groningen

Learning objectives

You will learn about

1. ... the structure of data in economics, econometrics and OR,
2. ... aims and proceedings of a scientific data analysis and
3. ... academic disciplines at the FEB that make extensive usage of mathematical and data analysis skills.

LO1: Types of Data-sets in Econometrics and OR

LO2: Scientific Data Analysis

- Introduction
- Inference *(further reading: DABEP Ch.5.1, 5.2, 5.8, 5.9, 6.1, 6.2, 6.4, 6.9, 6.10)*
- Prediction & Forecasting (*further reading: DABEP Ch.13.1, 13.2, 13.3, 13.5, 13.6, 13.9, 13.10, 13.11)*

LO3: Academic Disciplines for Data Analysis

- Econometrics
- Operations Research
- Machine Learning

LO1: Types of Data-sets in Econometrics and OR

Data structures

- **Time series data:** one unit of observation, many points in time.
- **Cross-sectional data:** many units of observation, one point in time
- **Panel data:** many units of observation, many points in time

Time series data

time	market_return
1514765100000	-0.003208872885
1514765400000	0.0001050783491
1514765700000	-0.007734359222
1514766000000	-0.0002423417424
1514766300000	0.005334311478
1514766600000	0.0003046544809
1514766900000	-0.002812629214
1514767200000	-0.001036172083
1514767500000	0.005568493695
1514767800000	-0.001032017698
1514768100000	0.0001145007552
1514768400000	-0.001381386807
1514768700000	-0.003762368874
1514769000000	-0.0002825162075
1514769300000	0.00701068594
1514769600000	0.0003199024956
1514769900000	0.0004886105346
1514770200000	-0.003079943195
1514770500000	0.001123677248
1514770800000	-0.00583770838
1514771100000	-0.004933226421
1514771400000	-0.003972859314
1514771700000	-0.007303157106
1514772000000	0.01066293136
1514772300000	-0.003299220655
1514772600000	0.005469851385
1514772900000	-0.006592880785

- One unit is measured at different points in time.

Time series data

time	market_return
1514765100000	-0.003208872885
1514765400000	0.0001050783491
1514765700000	-0.007734359222
1514766000000	-0.0002423417424
1514766300000	0.005334311478
1514766600000	0.0003046544809
1514766900000	-0.002812629214
1514767200000	-0.001036172083
1514767500000	0.005568493695
1514767800000	-0.001032017698
1514768100000	0.0001145007552
1514768400000	-0.001381386807
1514768700000	-0.003762368874
1514769000000	-0.0002825162075
1514769300000	0.00701068594
1514769600000	0.0003199024956
1514769900000	0.0004886105346
1514770200000	-0.003079943195
1514770500000	0.001123677248
1514770800000	-0.00583770838
1514771100000	-0.004933226421
1514771400000	-0.003972859314
1514771700000	-0.007303157106
1514772000000	0.01066293136
1514772300000	-0.003299220655
1514772600000	0.005469851385
1514772900000	-0.006592880785

- One unit is measured at different points in time.
- For each variable that is measured, we call this **a series**.

Time series data

time	market_return
1514765100000	-0.003208872885
1514765400000	0.0001050783491
1514765700000	-0.007734359222
1514766000000	-0.0002423417424
1514766300000	0.005334311478
1514766600000	0.0003046544809
1514766900000	-0.002812629214
1514767200000	-0.001036172083
1514767500000	0.005568493695
1514767800000	-0.001032017698
1514768100000	0.0001145007552
1514768400000	-0.001381386807
1514768700000	-0.003762368874
1514769000000	-0.0002825162075
1514769300000	0.00701068594
1514769600000	0.0003199024956
1514769900000	0.0004886105346
1514770200000	-0.003079943195
1514770500000	0.001123677248
1514770800000	-0.00583770838
1514771100000	-0.004933226421
1514771400000	-0.003972859314
1514771700000	-0.007303157106
1514772000000	0.01066293136
1514772300000	-0.003299220655
1514772600000	0.005469851385
1514772900000	-0.006592880785

- One unit is measured at different points in time.
- For each variable that is measured, we call this **a series**.
- Here we have the series of bitcoin prices.

Time series data

time	market_return
1514765100000	-0.003208872885
1514765400000	0.0001050783491
1514765700000	-0.007734359222
1514766000000	-0.0002423417424
1514766300000	0.005334311478
1514766600000	0.0003046544809
1514766900000	-0.002812629214
1514767200000	-0.001036172083
1514767500000	0.005568493695
1514767800000	-0.001032017698
1514768100000	0.0001145007552
1514768400000	-0.001381386807
1514768700000	-0.003762368874
1514769000000	-0.0002825162075
1514769300000	0.00701068594
1514769600000	0.0003199024956
1514769900000	0.0004886105346
1514770200000	-0.003079943195
1514770500000	0.001123677248
1514770800000	-0.00583770838
1514771100000	-0.004933226421
1514771400000	-0.003972859314
1514771700000	-0.007303157106
1514772000000	0.01066293136
1514772300000	-0.003299220655
1514772600000	0.005469851385
1514772900000	-0.006592880785

- One unit is measured at different points in time.
- For each variable that is measured, we call this **a series**.
- Here we have the series of bitcoin prices.
- For one unit, we can have multiple series.

Time series data

time	market_return
1514765100000	-0.003208872885
1514765400000	0.0001050783491
1514765700000	-0.007734359222
1514766000000	-0.0002423417424
1514766300000	0.005334311478
1514766600000	0.0003046544809
1514766900000	-0.002812629214
1514767200000	-0.001036172083
1514767500000	0.005568493695
1514767800000	-0.001032017698
1514768100000	0.0001145007552
1514768400000	-0.001381386807
1514768700000	-0.003762368874
1514769000000	-0.0002825162075
1514769300000	0.00701068594
1514769600000	0.0003199024956
1514769900000	0.0004886105346
1514770200000	-0.003079943195
1514770500000	0.001123677248
1514770800000	-0.00583770838
1514771100000	-0.004933226421
1514771400000	-0.003972859314
1514771700000	-0.007303157106
1514772000000	0.01066293136
1514772300000	-0.003299220655
1514772600000	0.005469851385
1514772900000	-0.006592880785

- Here we have the series of bitcoin prices.
- For one unit, we can have multiple series.

Cross section

Data Editor (Browse) - [cross_sectional]

File Edit View Data Tools

numHH[1] 182

	village	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	182	.15385	1.7734	.103448	2.76847
2	2	195	.11282	1.83744	.152709	2.85222
3	3	292	.11986	1.73623	.243478	2.50581
4	4	239	.075314	1.77734	.21875	2.4728
5	6	114	.19298	1.51818	.336364	2.47525
6	9	207	.1401	1.63563	.279352	2.32114
7	12	175	.15429	1.54872	.353846	2.71795
8	15	171	.16374	1.61321	.301887	2.62736
9	19	204	.13235	1.69136	.209877	2.31687
10	20	156	.064103	1.66038	.150943	2.44654
11	21	202	.12871	1.65238	.209524	2.31905
12	23	254	.11024	1.575	.189286	2.69286
13	24	163	.11656	1.65877	.161137	2.48341
14	25	252	.083333	1.75329	.151316	2.29276
15	29	290	.089655	1.66007	.039604	2.52475
16	31	153	.17647	1.725	.14	1.96
17	32	241	.11203	1.62126	.212625	2.2392
18	33	204	.063725	1.72603	.013699	2.83562
19	36	289	.13841	1.56655	.348123	2.57338
20	39	287	.1727	1.35405	.302703	3.02973
21	42	192	.078125	1.49029	.334951	1.75243

- We measure **multiple units** at one point in time.

Cross section

Data Editor (Browse) - [cross_sectional]

File Edit View Data Tools

numHH[1] 182

	village	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	182	.15385	1.7734	.103448	2.76847
2	2	195	.11282	1.83744	.152709	2.85222
3	3	292	.11986	1.73623	.243478	2.50581
4	4	239	.075314	1.77734	.21875	2.4728
5	6	114	.19298	1.51818	.336364	2.47525
6	9	207	.1401	1.63563	.279352	2.32114
7	12	175	.15429	1.54872	.353846	2.71795
8	15	171	.16374	1.61321	.301887	2.62736
9	19	204	.13235	1.69136	.209877	2.31687
10	20	156	.064103	1.66038	.150943	2.44654
11	21	202	.12871	1.65238	.209524	2.31905
12	23	254	.11024	1.575	.189286	2.69286
13	24	163	.11656	1.65877	.161137	2.48341
14	25	252	.083333	1.75329	.151316	2.29276
15	29	290	.089655	1.66007	.039604	2.52475
16	31	153	.17647	1.725	.14	1.96
17	32	241	.11203	1.62126	.212625	2.2392
18	33	204	.063725	1.72603	.013699	2.83562
19	36	289	.13841	1.56655	.348123	2.57338
20	39	287	.1777	1.35405	.302703	3.02973
21	42	192	.078125	1.49029	.334951	1.75243

- We measure **multiple units** at one point in time.
- Typically, we will measure multiple characteristics.

Cross section

Data Editor (Browse) - [cross_sectional]

File Edit View Data Tools

numHH[1] 182

	village	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	182	.15385	1.7734	.103448	2.76847
2	2	195	.11282	1.83744	.152709	2.85222
3	3	292	.11986	1.73623	.243478	2.50581
4	4	239	.075314	1.77734	.21875	2.4728
5	6	114	.19298	1.51818	.336364	2.47525
6	9	207	.1401	1.63563	.279352	2.32114
7	12	175	.15429	1.54872	.353846	2.71795
8	15	171	.16374	1.61321	.301887	2.62736
9	19	204	.13235	1.69136	.209877	2.31687
10	20	156	.064103	1.66038	.150943	2.44654
11	21	202	.12871	1.65238	.209524	2.31905
12	23	254	.11024	1.575	.189286	2.69286
13	24	163	.11656	1.65877	.161137	2.48341
14	25	252	.083333	1.75329	.151316	2.29276
15	29	290	.089655	1.66007	.039604	2.52475
16	31	153	.17647	1.725	.14	1.96
17	32	241	.11203	1.62126	.212625	2.2392
18	33	204	.063725	1.72603	.013699	2.83562
19	36	289	.13841	1.56655	.348123	2.57338
20	39	287	.1777	1.35405	.302703	3.02973
21	42	192	.078125	1.49029	.334951	1.75243

- We measure **multiple units** at one point in time.
- Typically, we will measure multiple characteristics.
- Here we have characteristics of Indian villages.

Cross section

... and here we have characteristics of households in Indian villages.

Data Editor (Browse) - [household_characteristics]

File Edit View Data Tools

rooftype1[1] 0

	village	adjmatrix_~y	HHnum_in_v~e	hhid	hohreligion	rooftype1	rooftype2	rooftype3	rooftype4	rooftype5	room_no	bed_no	electricity
1	1	1	1	1001	HINDUISM	0	1	0	0	0	3	4	No
2	1	2	2	1002	HINDUISM	0	1	0	0	0	1	1	Yes, Governmen
3	1	3	3	1003	HINDUISM	0	0	0	0	1	3	4	Yes, Private
4	1	4	4	1004	HINDUISM	0	1	0	0	0	2	6	Yes, Private
5	1	5	5	1005	HINDUISM	0	1	0	0	0	3	4	Yes, Private
6	1	6	6	1006	HINDUISM	0	0	1	0	0	2	1	Yes, Private
7	1	7	7	1007	HINDUISM	0	0	1	0	0	3	5	Yes, Governmen
8	1	8	8	1008	HINDUISM	0	0	0	1	0	2	1	Yes, Governmen
9	1	9	9	1009	HINDUISM	0	1	0	0	0	2	7	Yes, Governmen
10	1	10	10	1010	HINDUISM	0	0	1	0	0	2	1	Yes, Private
11	1	11	12	1012	HINDUISM	0	1	0	0	0	2	0	Yes, Governmen
12	1	12	13	1013	HINDUISM	0	1	0	0	0	2	2	Yes, Governmen
13	1	13	14	1014	HINDUISM	0	0	0	1	0	1	4	Yes, Governmen
14	1	14	15	1015	HINDUISM	0	1	0	0	0	2	0	Yes, Governmen
15	1	15	16	1016	HINDUISM	0	0	0	1	0	2	0	Yes, Governmen
16	1	16	17	1017	HINDUISM	0	0	0	1	0	1	0	Yes, Governmen
17	1	17	18	1018	HINDUISM	0	1	0	0	0	2	3	Yes, Private
18	1	18	19	1019	HINDUISM	0	0	0	1	0	3	0	Yes, Private
19	1	19	20	1020	HINDUISM	0	0	1	0	0	3	2	Yes, Private
20	1	20	21	1021	HINDUISM	0	0	1	0	0	3	2	Yes, Private
21	1	21	22	1022	HINDUISM	0	0	1	0	0	2	0	Yes, Private
22	1	22	23	1023	HINDUISM	0	0	1	0	0	2	4	Yes, Governmen
23	1	23	24	1024	HINDUISM	0	0	1	0	0	2	2	Yes, Governmen
24	1	24	25	1025	HINDUISM	0	0	0	1	0	1	1	Yes, Private
25	1	25	26	1026	HINDUISM	0	0	0	1	0	2	0	Yes, Governmen

Panel

Data Editor (Browse) - [panel]

File Edit View Data Tools

village[1] 1

	village	t	dynamicMF_empirical	dynamicMF_simulated	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	0	0	0	182	.15385	1.7734	.103448	2.76847
2	1	1	.1318681	.041966	182	.15385	1.7734	.103448	2.76847
3	1	2	.1538462	.092046	182	.15385	1.7734	.103448	2.76847
4	1	3	.2032967	.14703	182	.15385	1.7734	.103448	2.76847
5	1	4	.2032967	.1862	182	.15385	1.7734	.103448	2.76847
6	1	5	.2032967	.20966	182	.15385	1.7734	.103448	2.76847
7	1	6	.2307692	.22277	182	.15385	1.7734	.103448	2.76847
8	1	7	.2307692	.23062	182	.15385	1.7734	.103448	2.76847
9	1	8	.2307692	.23604	182	.15385	1.7734	.103448	2.76847
10	1	9	.	.2403	182	.15385	1.7734	.103448	2.76847
11	2	0	0	0	195	.11282	1.83744	.152709	2.85222
12	2	1	.1025641	.030169	195	.11282	1.83744	.152709	2.85222
13	2	2	.1538462	.063343	195	.11282	1.83744	.152709	2.85222
14	2	3	.1538462	.10437	195	.11282	1.83744	.152709	2.85222
15	2	4	.1538462	.14143	195	.11282	1.83744	.152709	2.85222
16	2	5	.1538462	.16941	195	.11282	1.83744	.152709	2.85222
17	2	6	.1538462	.18838	195	.11282	1.83744	.152709	2.85222
18	2	7	.1538462	.20152	195	.11282	1.83744	.152709	2.85222
19	2	8	.1538462	.2107	195	.11282	1.83744	.152709	2.85222
20	2	9	.1538462	.21751	195	.11282	1.83744	.152709	2.85222
21	2	10	.	.22235	195	.11282	1.83744	.152709	2.85222
22	3	0	0	0	292	.11986	1.73623	.243478	2.50581
23	3	1	.1156463	.022912	292	.11986	1.73623	.243478	2.50581
24	3	2	.1326531	.049662	292	.11986	1.73623	.243478	2.50581
25	3	3	.1360544	.08896	292	.11986	1.73623	.243478	2.50581
26	3	4	.1360544	.12913	292	.11986	1.73623	.243478	2.50581
27	3	5	.	.15858	292	.11986	1.73623	.243478	2.50581
28	4	0	0	0	239	.075314	1.77734	.21875	2.4728

- We measure **multiple units** at **multiple points in time**.

Panel

Data Editor (Browse) - [panel]

File Edit View Data Tools

village[1] 1

	village	t	dynamicMF_empirical	dynamicMF_simulated	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	0	0	0	182	.15385	1.7734	.103448	2.76847
2	1	1	.1318681	.041966	182	.15385	1.7734	.103448	2.76847
3	1	2	.1538462	.092046	182	.15385	1.7734	.103448	2.76847
4	1	3	.2032967	.14703	182	.15385	1.7734	.103448	2.76847
5	1	4	.2032967	.1862	182	.15385	1.7734	.103448	2.76847
6	1	5	.2032967	.20966	182	.15385	1.7734	.103448	2.76847
7	1	6	.2307692	.22277	182	.15385	1.7734	.103448	2.76847
8	1	7	.2307692	.23062	182	.15385	1.7734	.103448	2.76847
9	1	8	.2307692	.23604	182	.15385	1.7734	.103448	2.76847
10	1	9	.	.2403	182	.15385	1.7734	.103448	2.76847
11	2	0	0	0	195	.11282	1.83744	.152709	2.85222
12	2	1	.1025641	.030169	195	.11282	1.83744	.152709	2.85222
13	2	2	.1538462	.063343	195	.11282	1.83744	.152709	2.85222
14	2	3	.1538462	.10437	195	.11282	1.83744	.152709	2.85222
15	2	4	.1538462	.14143	195	.11282	1.83744	.152709	2.85222
16	2	5	.1538462	.16941	195	.11282	1.83744	.152709	2.85222
17	2	6	.1538462	.18838	195	.11282	1.83744	.152709	2.85222
18	2	7	.1538462	.20152	195	.11282	1.83744	.152709	2.85222
19	2	8	.1538462	.2107	195	.11282	1.83744	.152709	2.85222
20	2	9	.1538462	.21751	195	.11282	1.83744	.152709	2.85222
21	2	10	.	.22235	195	.11282	1.83744	.152709	2.85222
22	3	0	0	0	292	.11986	1.73623	.243478	2.50581
23	3	1	.1156463	.022912	292	.11986	1.73623	.243478	2.50581
24	3	2	.1326531	.049662	292	.11986	1.73623	.243478	2.50581
25	3	3	.1360544	.08896	292	.11986	1.73623	.243478	2.50581
26	3	4	.1360544	.12913	292	.11986	1.73623	.243478	2.50581
27	3	5	.	.15858	292	.11986	1.73623	.243478	2.50581
28	4	0	0	0	239	.075314	1.77734	.21875	2.4728

- We measure **multiple units** at **multiple points in time**.
- Typically, we will measure multiple characteristics, it is a series of the cross section.

Panel

Data Editor (Browse) - [panel]

File Edit View Data Tools

village[1] 1

	village	t	dynamicMF_empirical	dynamicMF_simulated	numHH	fractionLe~s	savings	shgpartici~e	fracGM_sur~y
1	1	0	0	0	182	.15385	1.7734	.103448	2.76847
2	1	1	.1318681	.041966	182	.15385	1.7734	.103448	2.76847
3	1	2	.1538462	.092046	182	.15385	1.7734	.103448	2.76847
4	1	3	.2032967	.14703	182	.15385	1.7734	.103448	2.76847
5	1	4	.2032967	.1862	182	.15385	1.7734	.103448	2.76847
6	1	5	.2032967	.20966	182	.15385	1.7734	.103448	2.76847
7	1	6	.2307692	.22277	182	.15385	1.7734	.103448	2.76847
8	1	7	.2307692	.23062	182	.15385	1.7734	.103448	2.76847
9	1	8	.2307692	.23604	182	.15385	1.7734	.103448	2.76847
10	1	9	.	.2403	182	.15385	1.7734	.103448	2.76847
11	2	0	0	0	195	.11282	1.83744	.152709	2.85222
12	2	1	.1025641	.030169	195	.11282	1.83744	.152709	2.85222
13	2	2	.1538462	.063343	195	.11282	1.83744	.152709	2.85222
14	2	3	.1538462	.10437	195	.11282	1.83744	.152709	2.85222
15	2	4	.1538462	.14143	195	.11282	1.83744	.152709	2.85222
16	2	5	.1538462	.16941	195	.11282	1.83744	.152709	2.85222
17	2	6	.1538462	.18838	195	.11282	1.83744	.152709	2.85222
18	2	7	.1538462	.20152	195	.11282	1.83744	.152709	2.85222
19	2	8	.1538462	.2107	195	.11282	1.83744	.152709	2.85222
20	2	9	.1538462	.21751	195	.11282	1.83744	.152709	2.85222
21	2	10	.	.22235	195	.11282	1.83744	.152709	2.85222
22	3	0	0	0	292	.11986	1.73623	.243478	2.50581
23	3	1	.1156463	.022912	292	.11986	1.73623	.243478	2.50581
24	3	2	.1326531	.049662	292	.11986	1.73623	.243478	2.50581
25	3	3	.1360544	.08896	292	.11986	1.73623	.243478	2.50581
26	3	4	.1360544	.12913	292	.11986	1.73623	.243478	2.50581
27	3	5	.	.15858	292	.11986	1.73623	.243478	2.50581
28	4	0	0	0	239	.075314	1.77734	.21875	2.4728

- We measure **multiple units** at **multiple points in time**.
- Typically, we will measure multiple characteristics, it is a series of the cross section.
- Here we have characteristics of Indian villages.

LO2: Scientific Data Analysis

LO2: Scientific Data Analysis

Introduction

From EDA to modelling

Lecture 1: characterising and illustrating data (explorative data analysis)

Now: discovering systematic patterns, generalize, predict

Purpose: learn from data,

- give policy advice or
- optimise business decision.

There is a **choice** regarding **inputs** that produce **outputs**.

From EDA to modelling

Lecture 1: characterising and illustrating data (explorative data analysis)

Now: discovering systematic patterns, generalize, predict

Purpose: learn from data,

- give policy advice or
- optimise business decision.

There is a **choice** regarding **inputs** that produce **outputs**.

data ⇒ pattern ⇒ new choice

Why do we need a model? - evaluate & understand

Research question: What were the effects of introducing a minimal wage on labor market outcomes?

Why do we need a model? - evaluate & understand

Research question: Low percentage of female professors because

- lack of childcare opportunities ?
- deliberate choice ?
- discrimination ?

Why do we need a model? - evaluate & understand

Research question: What are the effects of working overtime on productivity?

Why do we need a model? - evaluate & understand

- A choice contributed to (un)fortunate outcomes.
- Remember: we are not in a laboratory.
- To what extend can the outcome be attributed to the choice we made?

Can we derive a **systematic, significant, causal** relationship from the pattern we see in the data?

Why do we need a model? - forecast

Research question: What percentage of the wage is saved (saving share)?

Minimal wage in the Netherlands 1,635.60 EUR per month. What about lower wages???

Why do we need a model? - forecast

Research question: The last technological upgrade lead to an increase in demand of x%, how much will demand increase if the product is further improved?

Why do we need a model? - forecast

Research question: Currently, bachelor students earn ... Euro more than high school graduates. If the bachelor was longer / shorter, how would this affect the graduate salaries?

Why do we need a model? - forecast

Research question: What percentage of the wage is saved (saving share)?

Minimal wage in the Netherlands 1,635.60 EUR per month. What about lower wages???

Research question: The last technological upgrade lead to an increase in demand of x%, how much will demand increase if the product is further improved?

Research question: Currently, bachelor students earn ... Euro more than high school graduates. If the bachelor was longer / shorter, how would this affect the graduate salaries?

- These **hypothetical questions** are called **forecasting/prediction**.
- To answer them, we need we need **out of sample validity**

Inference - prediction trade off

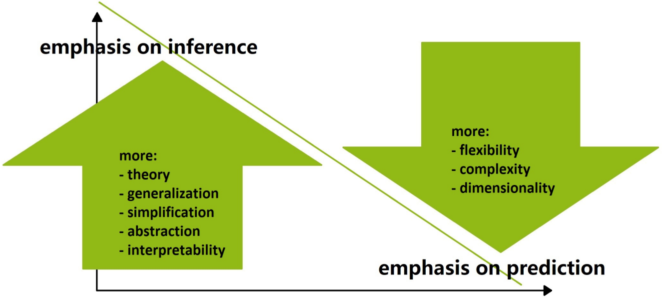
Inference: *"The act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty." (Merriam-Webster dictionary)*

Prediction: *"To calculate or predict (some future event or condition) usually as a result of study and analysis of available pertinent data." (Merriam-Webster dictionary)*

Where is our emphasis???

$$\textbf{data} \Rightarrow \underbrace{\textbf{patter}}_{\text{inference}} \Rightarrow \underbrace{\textbf{new choice}}_{\text{prediction}}$$

Inference versus prediction



LO2: Scientific Data Analysis

Inference *(further reading: DABEP Ch.5.1, 5.2, 5.8, 5.9, 6.1, 6.2, 6.4, 6.9, 6.10)*

What is a hypothesis and why do we want to test it?

- L2: we cannot observe the population, we can only ever **sample** from it

⇒ there is **noise**

- Idea: formulate a statement ("hypothesis") and based on the data evidence we "reject" or "not reject" it.
- **Important:** we can never say whether or not the hypothesis is true !!
- **A statistic** is a quantity (such as the mean of a sample) that is computed from a sample and used to test the hypothesis.

Internal validity

Internal validity: is the test statistic suitable for the hypothesis?

- **Hypothesis: "Influencer A has more impact on the price of Dodgecoin than influencer B."**
- Test statistic: average excess returns following a positive message emitted by the influencer (A-B).

Internal validity

Internal validity: is the test statistic suitable for the hypothesis?

- **Hypothesis: “Influencer A has more impact on the price of Dodgecoin than influencer B.”**
- Test statistic: average excess returns following a positive message emitted by the influencer (A-B).
- Do A and B use the same platforms/ channels ? (not one Youtube, the other Twitter)

Internal validity

Internal validity: is the test statistic suitable for the hypothesis?

- **Hypothesis: "Influencer A has more impact on the price of Dodgecoin than influencer B."**
- Test statistic: average excess returns following a positive message emitted by the influencer (A-B).
- Do A and B use the same platforms/ channels ? (not one Youtube, the other Twitter)

☠ Actual test: Do A's twitter post impact the price more than B's youtube videos?

External validity

External validity: (given internal validity), is the result valid for the population in general or only forthis sample?

- **Hypothesis: "Female students study more"**
- Test statistic: average hours per week spend studying among randomly selected RUG students (male - female).

External validity

External validity: (given internal validity), is the result valid for the population in general or only forthis sample?

- **Hypothesis: "Female students study more"**
- Test statistic: average hours per week spend studying among randomly selected RUG students (male - female).
- Are RUG students representative?

External validity

External validity: (given internal validity), is the result valid for the population in general or only forthis sample?

- **Hypothesis: "Female students study more"**
- Test statistic: average hours per week spend studying among randomly selected RUG students (male - female).
- Are RUG students representative?

☠ Violation ⇒ only statements about sample (not population) possible

Test statistic

- Sampling induces randomness

  ⇒ the test statistic is a random variable with a probability distribution.
- Theory ⇒ derive distribution of test statistic.
- Question: is the actual value of the test statistic observed **too unlikely** (less likely than 5%)?

YES: then we **reject**.

NO: then we **do not reject**.

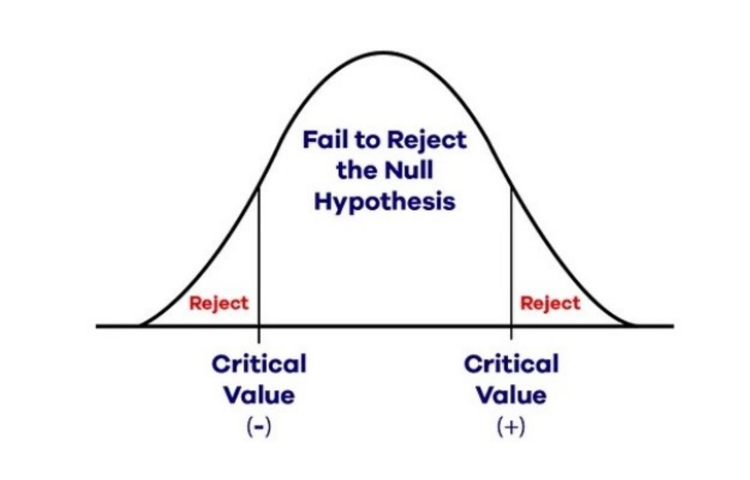
Null hypothesis

- A **null hypothesis (H_0)** is the hypothesis that the test statistic is zero.
- Test statistic **too large OR too small ⇒ reject!**
- Hypothesis: "Influencer A has more impact on the price of Dodgecoin than influencer B."

  $\Rightarrow H_0 : \mu_{r,A} - \mu_{r,B} = 0$
- Hypothesis: "Female students study more"

  $\Rightarrow H_0 : \mu_f - \mu_m = 0$

Hypothesis testing



source: Medium

Hypothesis testing

- H_0 : test statistic $= 0$.
- Alternative: test statistic $\neq 0$.

  $\Rightarrow$ reject if test statistic too large OR too small.
- If you can exclude that the test statistic is either positive or negative, you can formulate a different alternative.
- Alternative: test statistic > 0 (reject H_0 if test statistic too large) or
- Alternative: test statistic < 0 (reject H_0 if test statistic too small)

LO2: Scientific Data Analysis

Prediction & Forecasting (*further reading: DABEP Ch.13.1, 13.2, 13.3, 13.5, 13.6, 13.9, 13.10, 13.11)*

Prediction problem

on the basis of historical data, we want to

... predict the **value** of a **continuous** variable (NB: it can be a probability)

- quantitative forecasting
- toolbox: estimation techniques
- example: predicting bitcoin prices

... predict the **class** of a **categorial** variable

- qualitative forecasting
- tool: classification techniques
- example: consumer sentiment

Prediction

$$\hat{y} = f(x)$$

Forecast accuracy varies with

- quality and quantity of the historical data,
- model adequacy,
- forecasting horizon.

The forecast error consists of

- model error (we may not have chosen the best model),
- estimation error (remember: estimation $\neq$ calculation),
- irreducible error (a model can't perfectly predict).

Prediction

Multiple forecasts possible (what data is used and how?)
Forecasts can be

- ... conservative: under prediction more likely.
- ... optimistic: over prediction more likely.
- ... neutral: both equally likely.

Which would we prefer??

Prediction

Multiple forecasts possible (what data is used and how?)

Forecasts can be

... conservative: under prediction more likely.

... optimistic: over prediction more likely.

... neutral: both equally likely.

Which would we prefer??

Forecasting temperature changes caused by CO2 emission.

Prediction

Multiple forecasts possible (what data is used and how?)

Forecasts can be

- ... conservative: under prediction more likely.
- ... optimistic: over prediction more likely.
- ... neutral: both equally likely.

Which would we prefer??

Forecasting demand for a perishable product in order to purchase pre-products.

Prediction

Multiple forecasts possible (what data is used and how?)

Forecasts can be

... conservative: under prediction more likely.

... optimistic: over prediction more likely.

... neutral: both equally likely.

Which would we prefer??

Forecasting probability of deathly side effect of a drug.

Prediction

Multiple forecasts possible (what data is used and how?)
Forecasts can be

- ... conservative: under prediction more likely.
- ... optimistic: over prediction more likely.
- ... neutral: both equally likely.

Which would we prefer??
Forecasting company performance to attract investors.

Forecast evaluation

Prediction error:

$$e_n = y_n - \hat{y}_n$$

Mean squared error (MSE)

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 = \frac{1}{N}\sum_{n=1}^{N} e_n^2$$

Mean of squared prediction errors.

Forecast evaluation

Prediction error:

$$e_n = y_n - \hat{y}_n$$

Variance of the prediction error

$$Var(e) = \frac{1}{N}\sum_{n=1}^{N}(e_n - \bar{e})^2$$

Prediction error:

$$e_n = y_n - \hat{y}_n$$

Average prediction error

$$Bias = \frac{1}{N}\sum_{n=1}^{N} e_n = \bar{e_n}$$

Relationship between MSE, variance and bias

$$MSE = \frac{1}{N}\sum_{n=1}^{N} e_n^2 \quad Bias = \bar{e_n} \quad Var(e) = \frac{1}{N}\sum_{n=1}^{N}(e_n - \bar{e})^2$$

Relationship between MSE, variance and bias

$$MSE = \frac{1}{N}\sum_{n=1}^{N} e_n^2 \quad Bias = \bar{e_n} \quad Var(e) = \frac{1}{N}\sum_{n=1}^{N}(e_n - \bar{e})^2$$

$$Var(e) = \frac{1}{N}\sum_{n=1}^{N} e_n^2 + \frac{1}{N}\sum_{n=1}^{N} \bar{e}^2 - 2\frac{1}{N}\sum_{n=1}^{N} e_n\bar{e}$$

$$= \frac{1}{N}\sum_{n=1}^{N} e_n^2 + \bar{e}^2 - 2\bar{e}^2 = \underbrace{\frac{1}{N}\sum_{n=1}^{N} e_n^2}_{MSE} - \underbrace{\bar{e}^2}_{bias^2} = MSE - Bias^2$$

Relationship between MSE, variance and bias

$$MSE = \frac{1}{N}\sum_{n=1}^{N} e_n^2 \qquad Bias = \bar{e_n} \qquad Var(e) = \frac{1}{N}\sum_{n=1}^{N}(e_n - \bar{e})^2$$

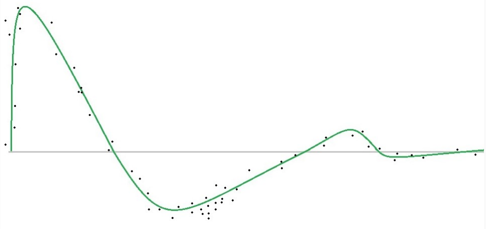
$$Var(e) = \frac{1}{N}\sum_{n=1}^{N} e_n^2 + \frac{1}{N}\sum_{n=1}^{N} \bar{e}^2 - 2\frac{1}{N}\sum_{n=1}^{N} e_n \bar{e}$$

$$= \frac{1}{N}\sum_{n=1}^{N} e_n^2 + \bar{e}^2 - 2\bar{e}^2 = \underbrace{\frac{1}{N}\sum_{n=1}^{N} e_n^2}_{MSE} - \underbrace{\bar{e}^2}_{bias^2} = MSE - Bias^2$$

$$\Rightarrow \text{MSE} = \text{Variance} + \text{Bias}^2$$

Bias-variance trade-off- assume a process with constant mean

Forecast 1 (grey): $\hat{y}_n = \bar{y}$ (historical mean)
Forecast 2 (green): curve fitted to historical data



Forecast 1 (grey): Bias$= 0$ but Variance $\uparrow$
Forecast 2 (green): Bias > 0 but Variance $\downarrow$

Training and test data

Split the sample into **training** (used for estimation of the model) and **test** (used for forecast evaluation/selection).

Cross validation: repeated splitting (each observation is used many times for training and once for testing).

Forecasting challenges

- Model break
- Feedback

Cost-benefit trade off of model sophistication:

Benefits: reduce prediction error

Cost: of data gathering and research as well as **risk** of model break (the world changes) before forecast is ready.

LO3: Academic Disciplines for Data Analysis

Economics & econometrics

Econometrics traditionally focuses on inference

1. Model creation (theoretical consideration involving abstraction, result: stylized model that depends on parameters with unknown value)
2. Identification (use probability theory to proof that it is possible to learn the value of the parameters from the data, state assumptions needed)
3. Estimation (choose and implement a formula that quantifies the parameter values for a given data set, call this formula "estimator")
4. Hypothesis testing
5. (Forecasting)

Machine learning

Machine learning:

"1. the process by which a computer is able to improve its own performance by continuously incorporating new data into an existing statistical model.
2. the branch of computer science dealing with the creation and use of computer software that employs machine learning"
(Merriam-Webster dictionary)

Machine learning

Machine learning:

"1. the process by which a computer is able to improve its own performance by continuously incorporating new data into an existing statistical model.
2. the branch of computer science dealing with the creation and use of computer software that employs machine learning"
(Merriam-Webster dictionary)

- for economics, econometrics and OR: ML is a **tool** that offers **new opportunities**.
- Comparative advantage of ML: high dimensional ("BIG") data, complex settings

  $\Rightarrow$ **focus on forecasting**

Machine learning

("Black box") idea:

- Map inputs into outputs,
- Maximise predictive power,
- No need for causality and interpretability.
- Given some restriction on the "black box", the computer uses data to improve model by means of rules.
- Statistical inference (e.g. hypothesis testing) can be part of decision rules.

Econometrics versus machine learning

(a) in econometrics, we want to disentangle cause and effect, **understand how the car works and improve it**

(b) in machine learning, we don't need to understand how the car works, the computer will build it, we want to **understand how to drive the car and win the race**

Usage of ML in economics, econometrics and OR

In practice: not "either/or" decision! ML can help econometrics.

- **Variable selection** (big data ⇒ too many variables for standard techniques).
- **Dimensionality reduction** (combine many weak predictors, into few strong predictors).
- **Cluster/patterns detection** (group units of observation).
- Automated **pre-analysis** on small sample can advice large-scale data gathering.
- **Theory refinement** (automated improvement on potentially complex theoretical models).

Operations research (OR)

Scientific analysis of choice in an operational setting (business, government, organisation) were scarce resources must be used efficiently

- (constraint) optimisation (find minimum/maximum) typically relying on iterative algorithms
- queuing models
- simulation

Queuing models

Customer requests service to be delivered by server
because the inflow varies, hence system features either waiting or idle capacities. e.g.

- restaurant services (how many tables/staff members)
- transportation (bus frequency)
- postal package collection and delivery
- call center
- university computer cluster
- hospital beds
- child care facilities
- ...

can all be modeled as queuing problems

Recap

This lecture, we learned

1. how the data sets that you are going to deal with look like.
2. the two most important ways in which scientists use data: prediction and inference.
3. the three academic disciplines of the FEB that deal most often with quantitative data analysis.