Case Studies: Styrian Health Data

Modelling and Assumptions

Farrukh Ahmed Muhammad Raafey Tariq Amirreza Khamehchin Khiabani Aymane Hachcham

Technische Universität Dortmund

May 21, 2023

Presentation Overview

Data Cleaning

- Data rows with ages less than 15 and ages greater than 100 are dropped.
- Rows with NA are removed for the sake of simplicity.
- Data sets have now 14831 data points after cleaning.

Assumptions of the Linear Model

We check the QQ plots of the measured systolic and diastolic blood pressures [Fahrmeir et al, 2013]

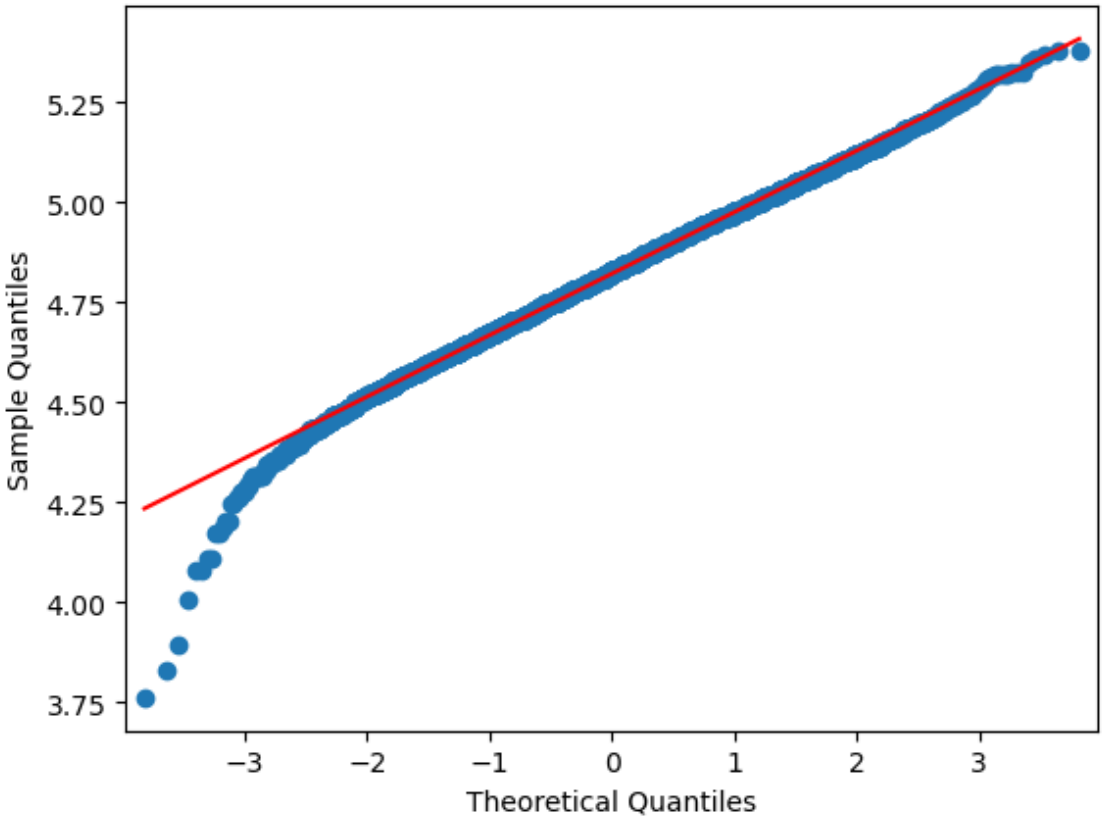

Figure: QQ Plot of the measured systolic blood pressure

Assumptions of the Linear Model

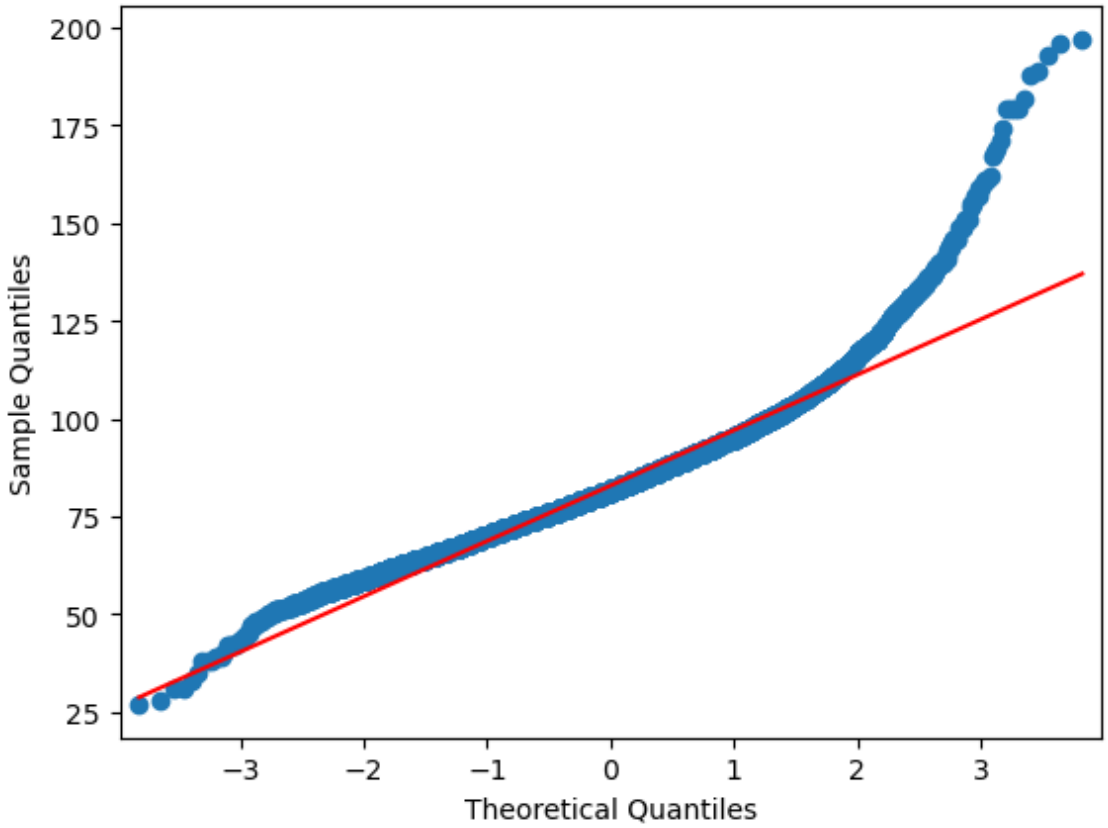

Figure: QQ Plot of the measured diastolic blood pressure

Assumptions of the Linear Model

We also the check the qq plots of the self-estimated systolic and diastolic blood pressures

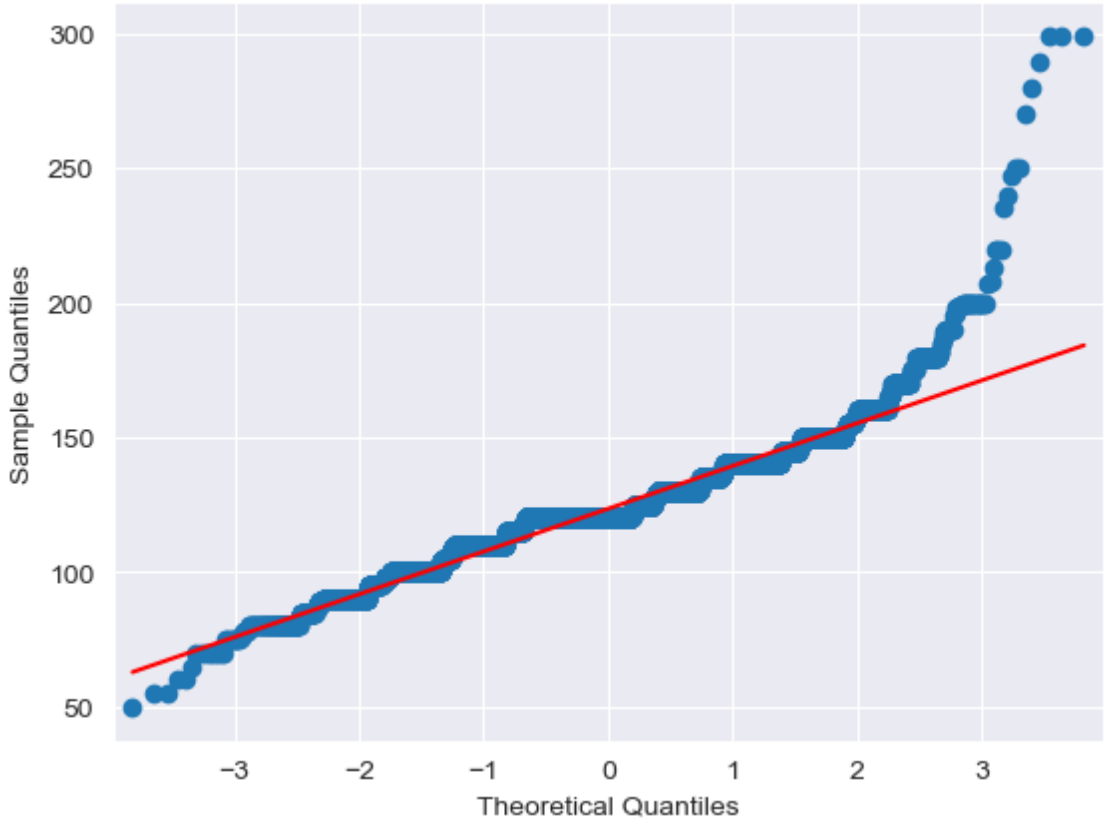

Figure: QQ Plot of the self-estimated systolic blood pressure

Assumptions of the Linear Model

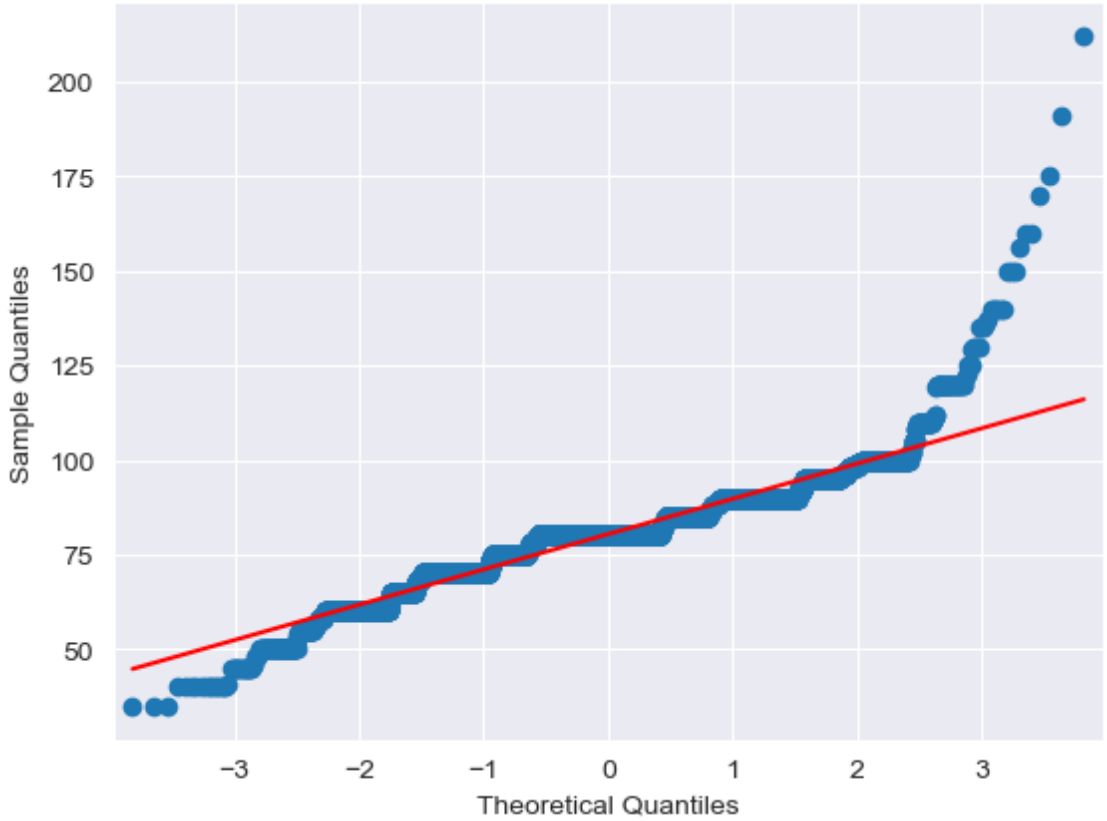

Figure: QQ Plot of the self-estimated diastolic blood pressure

Assumptions of the Linear Model

We verify the homoscedasticity of the residuals for all blood pressures [Fahrmeir et al, 2013].

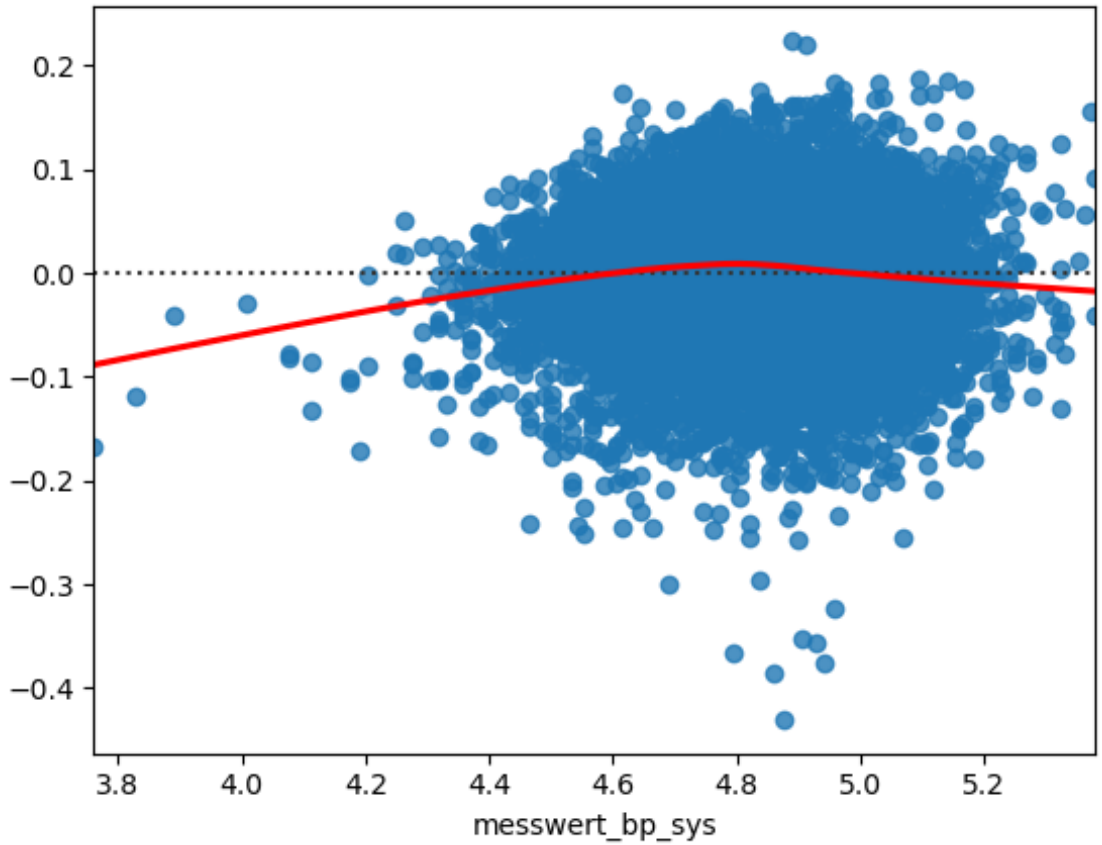

Figure: Residual plot of the measured systolic blood pressure

Assumptions of the Linear Model

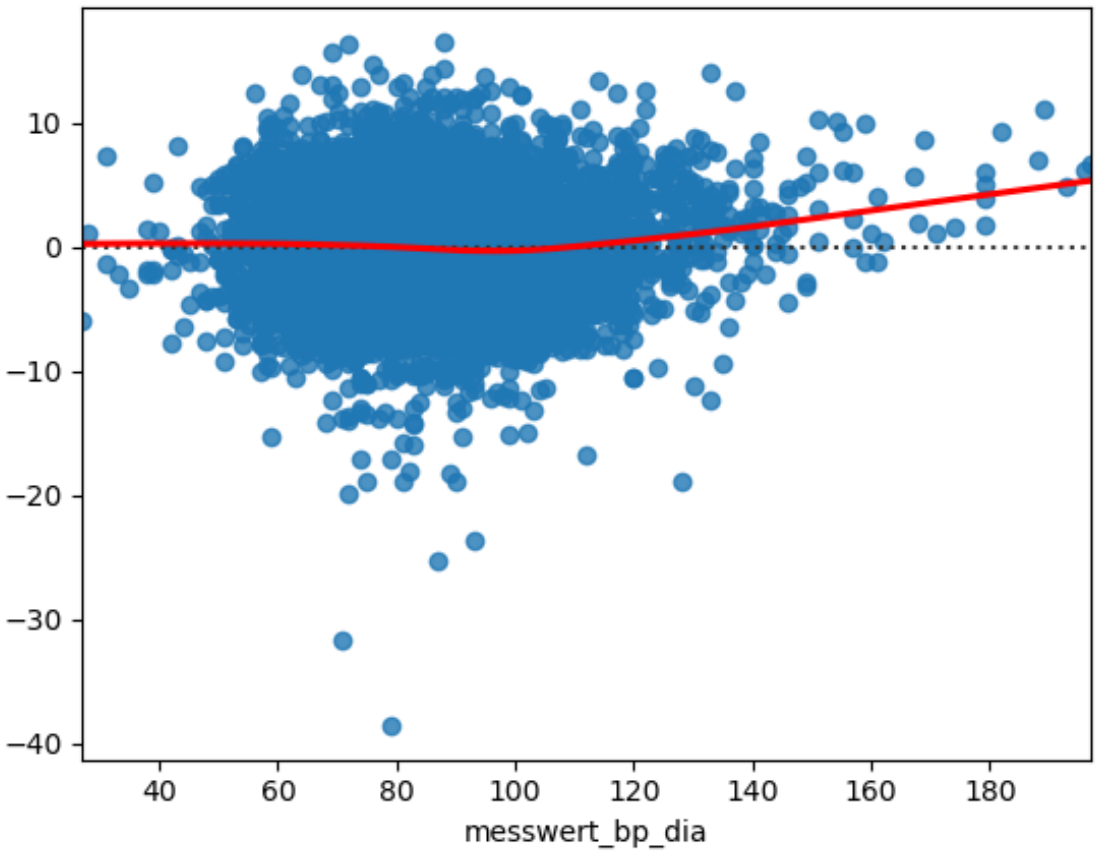

Figure: Residual Plot of the measured diastolic blood pressure

Assumptions of the Linear Model

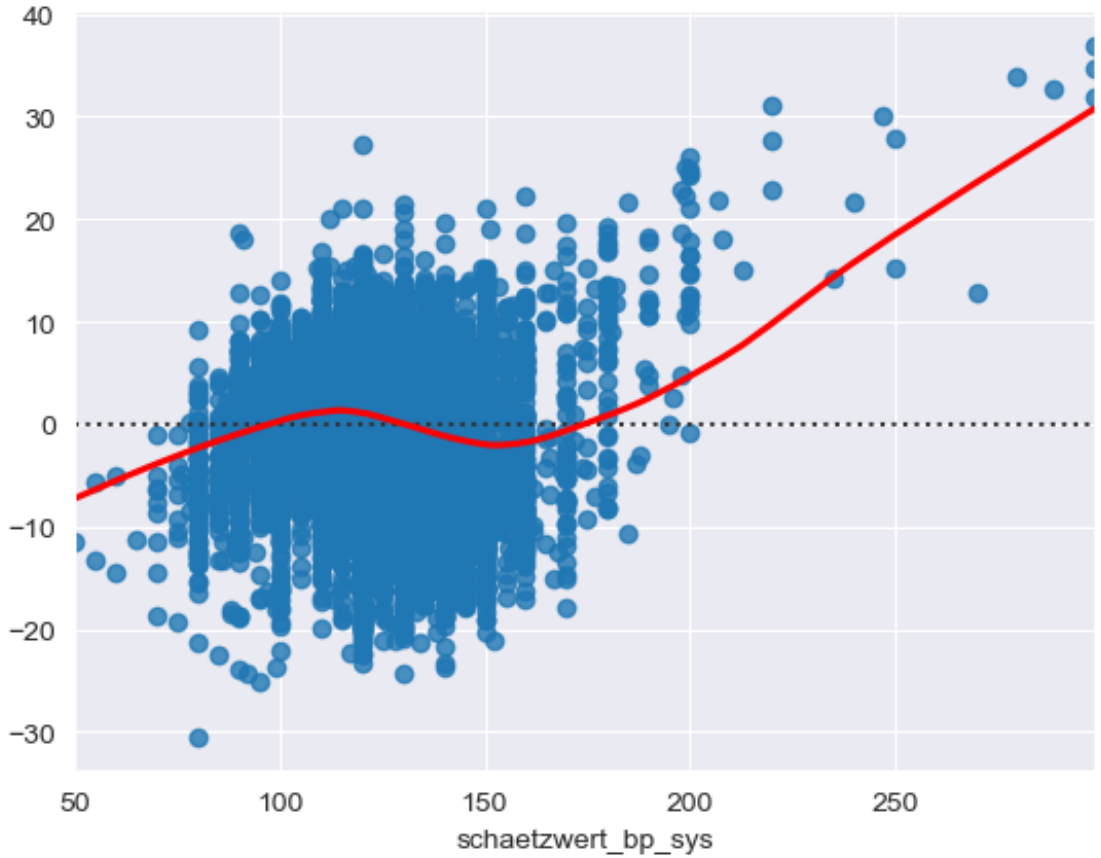

Figure: Residual Plot of the self-estimated systolic blood pressure

Assumptions of the Linear Model

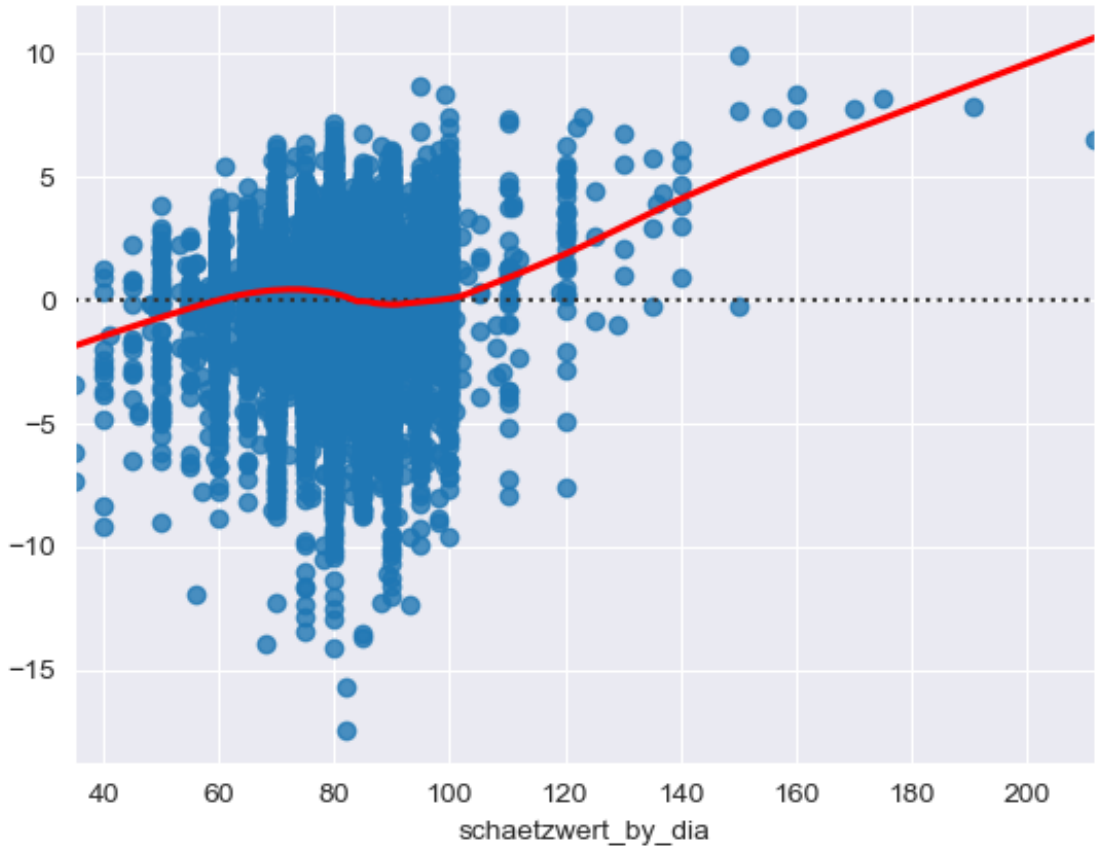

Figure: Residual Plot of the self-measured diastolic blood pressure

Results of the Linear Model

We display the Goodness of fit for diastolic and systolic blood pressures

OLS Regression Results

Dep. Variable:	messwert_bp_sys	R-squared:	0.161
Model:	OLS	Adj. R-squared:	0.161
Method:	Least Squares	F-statistic:	219.4
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.00
Time:	23:00:23	Log-Likelihood:	8011.5
No. Observations:	14831	AIC:	-1.599e+04
Df Residuals:	14817	BIC:	-1.589e+04
Df Model:	13		
Covariance Type:	nonrobust		

Figure: Goodness of fit of the measured systolic blood pressure

Results of the Linear Model

OLS Regression Results

Dep. Variable:	messwert_bp_dia	R-squared:	0.070
Model:	OLS	Adj. R-squared:	0.069
Method:	Least Squares	F-statistic:	86.15
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.10e-222
Time:	23:01:37	Log-Likelihood:	-59831.
No. Observations:	14831	AIC:	1.197e+05
Df Residuals:	14817	BIC:	1.198e+05
Df Model:	13		
Covariance Type:	nonrobust		

Figure: Goodness of fit of the measured diastolic blood pressure

Decision Tree

- Dataset was split into training and testing data. (70/30)
- Four Decision trees were fitted on the training data for *schaetzwert_ bp_sys*,*schaetzwert_ bp_dia*,*messwert_ bp_sys*, and *messwert_ bp_dia* as target variables respectively.

Decision Tree

- Decision tree with *messwert_ bp_sys* as Target variable

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	n_behandlung
8	schaetzwert_bp_sys
9	schaetzwert_by_dia
10	messwert_by_dia
11	age_group
12	is_local_resident

Table: Features used to train Decision Trees

Decision Trees

Figure: MSE of trees with Different Depths

Decision Tree

	Depth	MSE Training Data	MSE Testing Data	R-squared Train	R-squared Test
0	1	74.682516	71.779739	0.149858	0.155837
1	2	68.457156	65.851076	0.220724	0.225561
2	3	65.769224	63.666656	0.251322	0.251251
3	4	63.767727	63.355000	0.274106	0.254916
4	5	60.158157	62.261751	0.315195	0.267773
5	6	57.244747	64.770892	0.348360	0.238265
6	7	54.482893	65.503298	0.379799	0.229651
7	8	51.858702	68.747928	0.409671	0.191493
8	9	48.690268	72.026934	0.445739	0.152930
9	10	45.515553	75.727683	0.481878	0.109408
10	11	41.721803	77.357507	0.525064	0.090240
11	12	37.407668	82.857518	0.574173	0.025558
12	13	33.547238	88.328619	0.618118	-0.038785
13	14	29.625804	95.368463	0.662757	-0.121577
14	15	25.923056	96.206079	0.704907	-0.131428
15	16	22.601181	100.616092	0.742722	-0.183291
16	17	19.404065	106.449316	0.779116	-0.251893
17	18	16.406148	109.232345	0.813242	-0.284623
18	19	13.485100	111.314758	0.846494	-0.309113
19	20	11.055831	118.003899	0.874147	-0.387780
20	21	8.941129	118.119314	0.898219	-0.389137
21	22	7.188372	122.056068	0.918172	-0.435435
22	23	5.648479	125.638895	0.935701	-0.477571
23	24	4.554170	127.431062	0.948158	-0.498648

Decision Tree

- Decision tree with *messwert_ bp_dia* as Target variable

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	n_behandlung
8	schaetzwert_bp_sys
9	schaetzwert_by_dia
10	messwert_by_sys
11	age_group
12	is_local_resident

Table: Features used to train Decision Trees

Decision Trees

Figure: MSE of trees with Different Depths

Decision Tree

	Depth	MSE Training Data	MSE Testing Data	R-squared Train	R-squared Test
0	1	149.021684	144.750316	0.259711	0.277873
1	2	125.861958	125.865790	0.374761	0.372084
2	3	115.041338	115.690825	0.428514	0.422844
3	4	107.827121	112.317981	0.464352	0.439671
4	5	103.623801	110.186941	0.485232	0.450302
5	6	99.750426	111.956880	0.504474	0.441472
6	7	95.693079	115.645126	0.524629	0.423072
7	8	91.383591	116.590045	0.546037	0.418358
8	9	86.113744	125.000883	0.572216	0.376399
9	10	80.631537	131.251783	0.599450	0.345214
10	11	75.020586	136.295437	0.627323	0.320053
11	12	68.768059	147.352390	0.658384	0.264892
12	13	62.623551	150.998657	0.688907	0.246702
13	14	56.632749	160.539411	0.718668	0.199105
14	15	50.760299	167.859589	0.747840	0.162586
15	16	45.325104	172.730335	0.774840	0.138287
16	17	39.703959	177.937459	0.802764	0.112310
17	18	34.950178	184.142419	0.826379	0.081355
18	19	29.944640	191.621760	0.851245	0.044042
19	20	25.577038	198.169822	0.872942	0.011375
20	21	21.917328	205.147624	0.891122	-0.023435
21	22	19.104951	206.691851	0.905093	-0.031139
22	23	16.973431	210.084278	0.915682	-0.048063
23	24	15.223494	212.198861	0.924375	-0.058613

Decision Tree

- Decision tree with *schaetzwert_ bp_sys* as Target variable

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	n_behandlung
8	messwert_bp_sys
9	schaetzwert_by_dia
10	messwert_by_dia
11	age_group
12	is_local_resident

Table: Features used to train Decision Trees

Decision Trees

Figure: MSE of trees with Different Depths

Decision Tree

	Depth	MSE Training Data	MSE Testing Data	R-squared Train	R-squared Test
0	1	212.443082	212.725254	0.157769	0.154941
1	2	192.977208	192.154256	0.234942	0.236660
2	3	179.553988	177.369829	0.288158	0.295392
3	4	168.128593	169.285951	0.333454	0.327505
4	5	161.328871	168.470546	0.360411	0.330745
5	6	155.795556	162.473874	0.382348	0.354567
6	7	149.189593	165.750518	0.408538	0.341550
7	8	139.018802	173.945021	0.448860	0.308997
8	9	130.034902	183.216754	0.484476	0.272165
9	10	117.942966	195.888080	0.532415	0.221827
10	11	107.894874	210.795092	0.572251	0.162609
11	12	96.810921	226.802735	0.616193	0.099018
12	13	84.258837	241.601120	0.665956	0.040231
13	14	72.810991	247.716606	0.711341	0.015937
14	15	62.750815	257.761906	0.751224	-0.023969
15	16	53.985446	269.984547	0.785975	-0.072523
16	17	46.766361	278.996693	0.814595	-0.108325
17	18	39.559080	287.308324	0.843168	-0.141343
18	19	33.016016	303.571545	0.869108	-0.205949
19	20	28.144258	306.093186	0.888422	-0.215966
20	21	23.548307	303.652694	0.906643	-0.206271
21	22	19.541564	318.739808	0.922527	-0.266206
22	23	16.213440	315.441713	0.935722	-0.253104
23	24	13.698896	325.820121	0.945691	-0.294332

Decision Tree

- Decision tree with *schaetzwert_ bp_dia* as Target variable

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	n_behandlung
8	messwert_bp_sys
9	schaetzwert_by_sys
10	messwert_by_dia
11	age_group
12	is_local_resident

Table: Features used to train Decision Trees

Decision Trees

Figure: MSE of trees with Different Depths

Decision Tree

	Depth	MSE Training Data	MSE Testing Data	R-squared Train	R-squared Test
0	1	75.613802	69.889369	0.149015	0.157997
1	2	69.127274	64.576640	0.222016	0.222003
2	3	66.788381	62.668338	0.248339	0.244993
3	4	64.817314	62.812176	0.270522	0.243260
4	5	61.901297	60.161008	0.303340	0.275201
5	6	60.492783	63.660440	0.319192	0.233041
6	7	58.103727	68.663648	0.346079	0.172764
7	8	55.274041	65.570271	0.377926	0.210032
8	9	51.736787	68.702982	0.417735	0.172290
9	10	47.113837	76.540967	0.469764	0.077860
10	11	42.748872	82.320971	0.518889	0.008225
11	12	38.409943	82.395539	0.567720	0.007327
12	13	34.039189	83.791842	0.616911	-0.009496
13	14	29.878289	88.628496	0.663739	-0.067766
14	15	25.693933	98.726333	0.710831	-0.189421
15	16	22.356941	101.404938	0.748387	-0.221692
16	17	19.260962	110.220063	0.783230	-0.327894
17	18	16.582607	105.624180	0.813373	-0.272524
18	19	14.135054	107.902731	0.840919	-0.299975
19	20	11.903050	109.195243	0.866039	-0.315547
20	21	9.995088	117.802504	0.887512	-0.419245
21	22	8.350012	115.218209	0.906026	-0.388110
22	23	6.667697	121.576172	0.924959	-0.464708
23	24	5.487753	123.718725	0.938239	-0.490521

Summary

- First, a subset of features is selected for training two different models on the data set considering the continuous target variables.
- For *linear models*, the statistical assumptions are examined. *Normality* and *homogeneity* are concerned before the training the data.
- In decision tree setup, the data is split into two training and testing. For different target variables, the optimal *depth* can be well recognized by the *MSE* calculations.

Follow up questions:

1. Among the models, which one is the best? How to improve the selected one?
2. Regarding the null values, method to addressing the issue and their affect on the selected model.

References

- Van Rossum, G. & Drake Jr, F. Python reference manual. (Centrum voor Wiskunde en Informatica Amsterdam, 1995)
- Fernando Pérez, Jupyter Notebook: Open-source platform for interactive programming. (Project Jupyter, 2021)
- Ludwig Fahrmeir et al, Regression: Models, Methods and Applications, 2013.

Thank you!

Questions?