Case Studies: Health Data

Muhammad Raafey Tariq Farrukh Ahmed Aymane Hachcham **Amirreza Khamehchin Khiabani**

Technische Universität Dortmund

May 21, 2023

Data Description

- The Data is split into 70/30 ratio.
- Size of training data is 10381 observations, whereas the size of test data is 4450 observations.

Data Description

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	in_behandlung
8	schaetzwert_bp_sys
9	schaetzwert_by_dia
10	messwert_by_sys
11	age

Table: Features used to train models using messwert_by_dia as the target

Data Description

Nr	Feature Variables
1	bundesland
2	befinden
3	geschlecht
4	raucher
5	blutzucker_bekannt
6	cholesterin_bekannt
7	in_behandlung
8	schaetzwert_bp_sys
9	schaetzwert_by_dia
10	messwert_by_dia
11	age

Table: Features used to train models using messwert_by_sys as the target

Best Subset selection

- The best subset selection involves fitting series of models with possible combinations of k predictors.
- For each model of given size the best-subset model of predictors is chosen based on the chosen statistical criteria. (In our case it is Adjusted R^2).
- The total number of models of all sizes are $2^k - 1$ (where k is number of predictors)

Goodness of Fit - Coefficient of Determination

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

where $\bar{y}$ is mean value of response variable and $\hat{y}_i$ for $i = 1, ..., n$ are the estimated values.

Adjusted Coefficient of Determination

$$\bar{R^2} = 1 - \frac{n-1}{n-p}(1 - R^2)$$

where n is the number of observations and p is the number of covariates.

Cross Validation

- Available data is split into equal-sized *k*-folds with each fold used for testing and training the data.
- The model is then trained *k* times, each time using different folds for testing and the remaining folds for training.
- After training each model, its performance is evaluated on the testing data from the corresponding fold. This results in *k* performance scores (**mean-squared-error**), one for each fold. These scores can be averaged to obtain the overall performance of the model.

Cross Validation - Performance Metric

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i and $\hat{y}_i$ for $i = 1, \ldots, n$ are the response values and estimated values respectively.

Linear Regression Model

In this method, the goal is to model the effect of a given set of k independent variables or covariates, $x_1, ..., x_k$, on a dependent or response variable y. The covariates, as well as the response variable, can be binary, categorical, or continuous.

The following formula is applied:

$$E(y|x_1, ..., x_k) = f(x_1, ..., x_k).$$

The response can be formulated as:

$$y = E(y|x_1, ..., x_k) = f(x_1, ..., x_k) + \epsilon,$$

where y is a random variable with the distribution being dependent on $x_1, ..., x_k$ and ϵ is the error term and result is:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k.$$

Dummy variables

To model a categorical covariate, $x \in 1, ..., c$ with c categories, category c can be considered as a reference category. Afterwards, $c - 1$ dummy variables can be described:

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & otherwise \end{cases} \qquad \dots \qquad x_{i,c-1} = \begin{cases} 1 & x_i = c - 1 \\ 0 & otherwise \end{cases}$$

The regression model formula is represented:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \epsilon_i.$$

The assumptions for linear models

The assumptions for building a linear regression:

- Linearity: The function $f(x_1, ..., x_k)$ is linear.
- Independency: The residuals are independent, i.e., there is no correlation between residuals.
- Homoscedasticity: A constant error variance σ^2 is considered across observations, where $var(\epsilon_i) = \sigma^2$.
- Normality: The errors are normally distributed.

T-test and confidence interval for coefficients β_j

The null and alternative hypotheses for t-test are:

$$H_0 : \beta_j = 0 \text{ against } H_1 : \beta_j \neq 0, \text{ where } j = 1, ..., k$$

The estimated test statistic is obtained as:

$$\hat{t}_j = \frac{\hat{\beta}_j}{\hat{se}_j}.$$

Confidence interval is a method of interval estimation for a certain probability.

The probability of not rejecting H_o, is given by:

$$P(\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j < \beta_j < \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j) = 1 - \alpha.$$

According to this formula, the confidence interval would be:

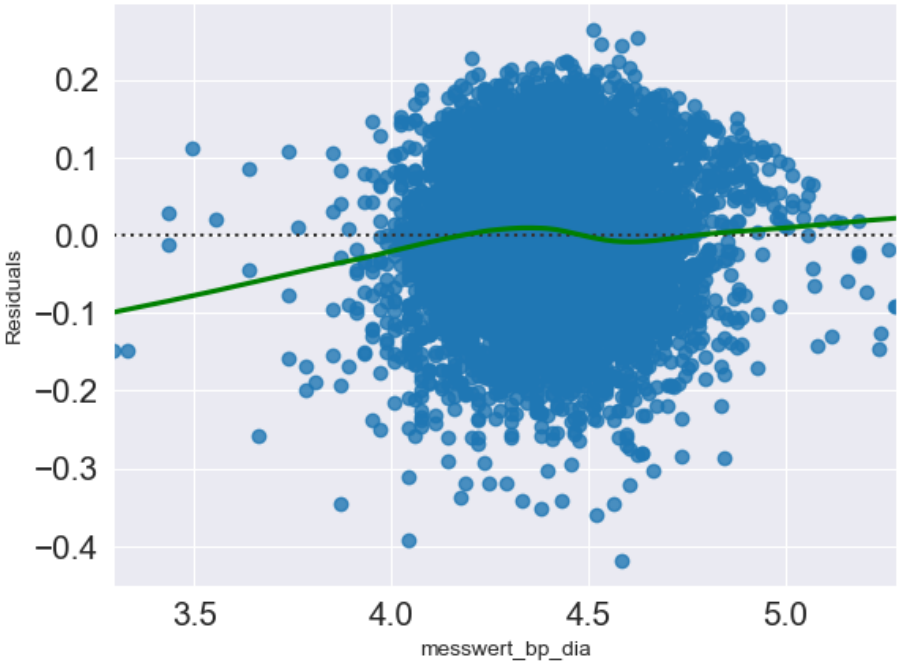
$$\left[\hat{\beta}_j - t_{n-p}(1-\alpha/2) \cdot se_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2) \cdot se_j\right].$$

Residual Plot

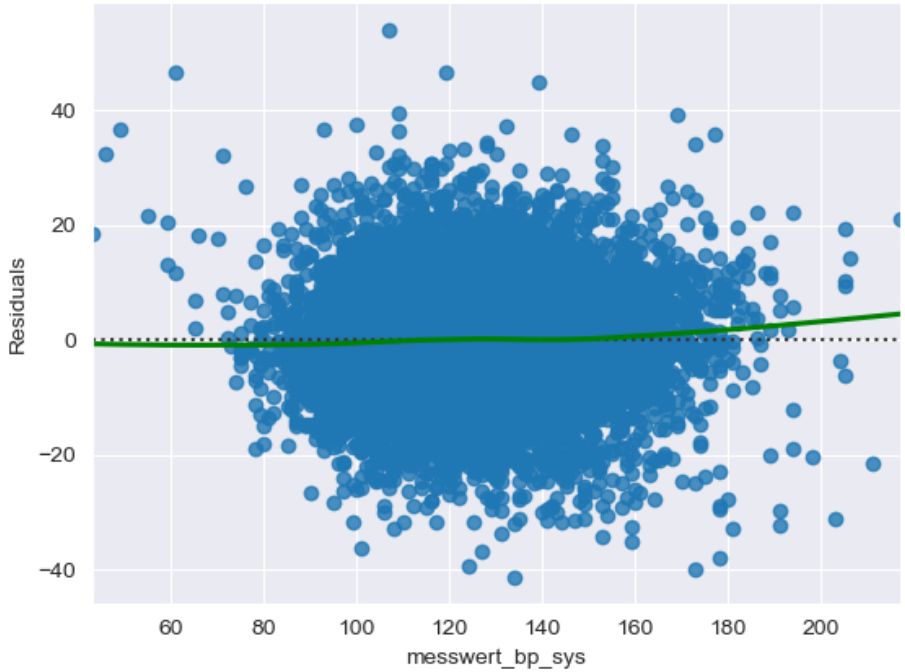
A residual plot illustrates the relationship between the residuals (the differences between actual and expected values) and the independent variable, with the residuals plotted on the vertical axis and the independent variable on the horizontal axis,

$$\hat{\epsilon} = y - \hat{y}.$$

Residual Plots



(a) Residual plot for diastolic bp



(b) Residual plot for systolic bp

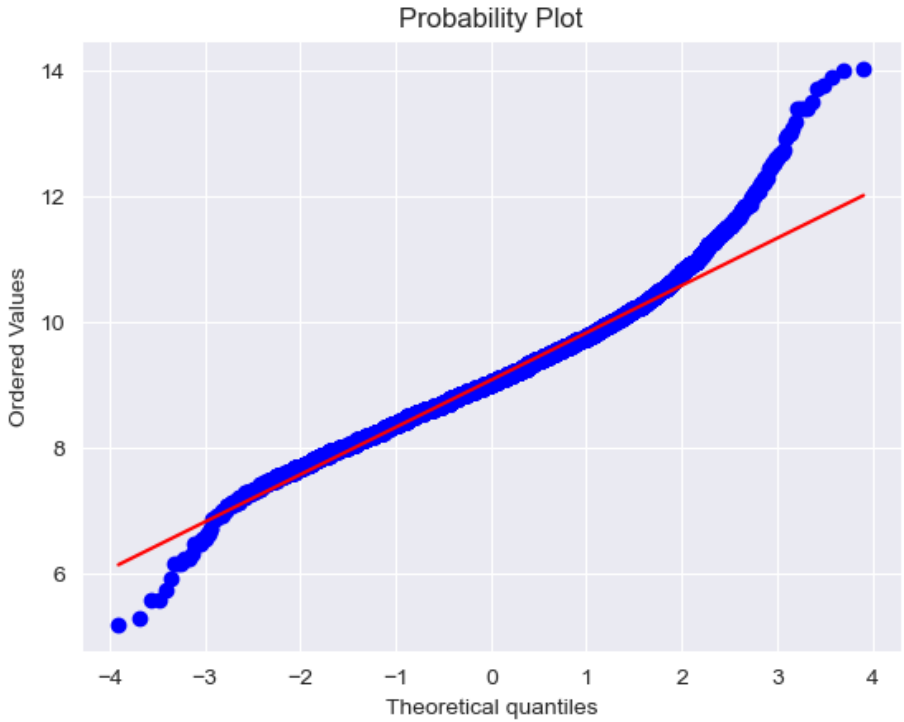
Figure: Residual plots for the measured systolic and diastolic bp.

Variable transformation

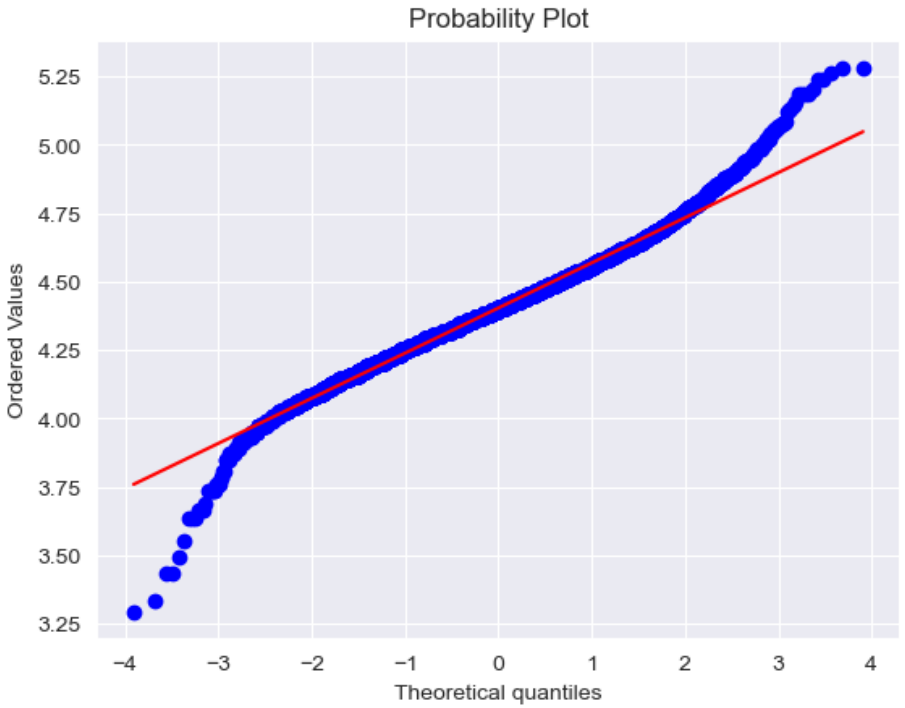
- The reason to use variable transformation is that it can help to make the distribution of the target more symmetric, which in turn can improve the performance of linear models that assume normality.
- Similarly, if the target variable has a large range of values, a log transformation can help to reduce the influence of extreme values and improve the linearity assumption of the linear model.
- In our case, we use a log transformation:

$$y_s = \ln(y)$$

QQ plots



(a) QQ plot for the measured diastolic bp.



(b) QQ plot for the diastolic bp after log transformation

Figure: QQ plots of the measured diastolic bp before (left) and after (right) transformation

Results after fitting the linear model

Summary of the linear model for the diastolic bp.

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.6021	0.019	190.057	0.000	3.565	3.639
bundesland_Kärnten	-0.0284	0.018	-1.602	0.109	-0.063	0.006
bundesland_Niederösterreich	-0.0136	0.016	-0.876	0.381	-0.044	0.017
bundesland_Oberösterreich	-0.0198	0.017	-1.164	0.245	-0.053	0.014
bundesland_Salzburg	-0.0054	0.021	-0.254	0.799	-0.047	0.036
bundesland_Steiermark	-0.0147	0.014	-1.063	0.288	-0.042	0.012
bundesland_Tirol	-0.0025	0.023	-0.112	0.911	-0.047	0.042
bundesland_Vorarlberg	0.0378	0.032	1.179	0.238	-0.025	0.101
bundesland_Wien	-0.0052	0.015	-0.338	0.736	-0.035	0.025
bundesland_not applicable	-0.0100	0.016	-0.605	0.545	-0.042	0.022
befinden_2.0	-0.0126	0.003	-4.661	0.000	-0.018	-0.007
befinden_3.0	-0.0112	0.004	-2.892	0.004	-0.019	-0.004
befinden_4.0	-0.0029	0.011	-0.269	0.788	-0.024	0.019
befinden_5.0	0.0112	0.019	0.583	0.560	-0.027	0.049
geschlecht_m	0.0069	0.002	2.784	0.005	0.002	0.012
raucher_True	0.0040	0.003	1.198	0.231	-0.003	0.010
blutzucker_bekannt_True	-0.0054	0.003	-1.577	0.115	-0.012	0.001
cholesterin_bekannt_True	0.0041	0.003	1.306	0.191	-0.002	0.010
in_behandlung_True	-0.0279	0.004	-7.004	0.000	-0.036	-0.020
schaetzwert_bp_sys	-0.0002	9.43e-05	-1.662	0.097	-0.000	2.81e-05
schaetzwert_by_dia	0.0017	0.000	11.461	0.000	0.001	0.002
messwert_bp_sys	0.0059	6.93e-05	85.127	0.000	0.006	0.006
age	-0.0008	8.65e-05	-8.985	0.000	-0.001	-0.001

R-squared: 0.459

Results after fitting the linear model

Summary of the linear model for the systolic bp.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-213.5195	3.849	-55.469	0.000	-221.065	-205.974
bundesland_Kärnten	1.8197	1.922	0.947	0.344	-1.947	5.587
bundesland_Niederösterreich	0.0105	1.702	0.006	0.995	-3.325	3.346
bundesland_Oberösterreich	0.3980	1.892	0.210	0.833	-3.311	4.107
bundesland_Salzburg	-0.7537	2.369	-0.318	0.750	-5.397	3.890
bundesland_Steiermark	0.7810	1.531	0.510	0.610	-2.219	3.781
bundesland_Tirol	-3.8879	2.473	-1.572	0.116	-8.736	0.960
bundesland_Vorarlberg	-4.2982	3.096	-1.388	0.165	-10.366	1.770
bundesland_Wien	-0.2015	1.685	-0.120	0.905	-3.505	3.102
bundesland_not applicable	-0.5060	1.817	-0.278	0.781	-4.067	3.055
befinden_2.0	0.7753	0.293	2.645	0.008	0.201	1.350
befinden_3.0	0.7168	0.415	1.726	0.084	-0.097	1.531
befinden_4.0	0.1949	1.177	0.166	0.868	-2.111	2.501
befinden_5.0	0.1149	2.006	0.057	0.954	-3.818	4.048
geschlecht_m	1.4558	0.268	5.436	0.000	0.931	1.981
raucher_True	-0.6271	0.358	-1.750	0.080	-1.329	0.075
blutzucker_bekannt_True	0.9322	0.369	2.528	0.011	0.209	1.655
cholesterin_bekannt_True	-0.6342	0.341	-1.858	0.063	-1.303	0.035
in_behandlung_True	4.6837	0.422	11.104	0.000	3.857	5.511
schaetzwert_bp_sys	0.1625	0.010	16.366	0.000	0.143	0.182
schaetzwert_by_dia	-0.0518	0.016	-3.228	0.001	-0.083	-0.020
messwert_bp_dia	70.7075	0.818	86.461	0.000	69.104	72.311
age	0.2122	0.009	23.141	0.000	0.194	0.230

R-squared: 0.532

Results for LM

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	109.906435	108.071011	0.456546	0.454677	0.455392	0.451967
1	LM Log	0.015051	0.014476	0.458401	0.461434	0.457251	0.458757
2	LM (Best Subset)	108.053036	110.150122	0.454768	0.455342	0.453293	0.454711
3	LM Log (Best Subset)	0.014472	0.015094	0.461595	0.456864	0.460139	0.456235

Table: Comparison of the training and testing performance for the diastolic linear model after and before best subset selection is applied.

Results for LM

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	LM (Base)	177.574175	173.376931	0.522072	0.527149	0.521057	0.524799
1	LM Log	0.011693	0.011482	0.508621	0.511067	0.507577	0.508637
2	LM (Best Subset)	173.320480	177.601697	0.527303	0.521998	0.525276	0.521121
3	LM Log (Best Subset)	0.011477	0.011696	0.511244	0.508493	0.509147	0.507592

Table: Comparison of the training and testing performance for the systolic linear model after and before best subset selection is applied.

Regression Tree

A regression tree is a type of decision tree that works by recursively partitioning the input space into smaller and smaller sub-regions, and fitting a simple model.

The following steps are involved in a regression tree:

- Starting with the entire dataset, identify the variable that provides the best split in terms of maximizing the reduction of the mean squared error of the response variable.
- Split the data based on the chosen variable and create two child nodes. Each child node corresponds to a subset of the data that satisfies a certain condition.
- Repeat the above two steps until the stopping criteria. (In our case it is the max depth).

Regression Tree Criterion for Split - MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i and $\hat{y}_i$ for $i = 1, \ldots, n$ are the response values and estimated values respectively.

Ensembles Random Forest

- Uses N Regression Trees and returns an average of the results
- At each step just a subset of the original features is used $log_2 N$
- Data is randomly sampled for each tree

Tree Configuration

- For both Random Forests and Regression Trees we fitted 4 variations:
  - Base Model
  - Fine-tuned Model
  - Best Subset Selected Base Model
  - Best Subset Selected Fine-tuned Model
- The Base Model is the tree fitted with default parameters (no control on depth)
- The Fine-tuned version for target variable messwert_by_sys uses CV to set parameters: max_depth to 11, min_samples_leaf to 66, and max_features to 20
- The Fine-tuned version for target variable messwert_by_dia uses CV to set parameters: max_depth to 6, min_samples_leaf to 11, and max_features to 21

Results with messwert_by_dia as the target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	0.490000	217.600000	1.000000	-0.100000	1.000000	-0.110000
1	Tree (Fine-tuned)	100.350000	105.500000	0.500000	0.460000	0.500000	0.460000
2	Tree (Best Subset + Base)	109.180000	117.470000	0.460000	0.400000	0.460000	0.400000
3	Tree (Best Subset + Fine-tuned)	100.330000	105.650000	0.500000	0.460000	0.500000	0.460000
4	RF (Base)	15.950000	114.530000	0.920000	0.420000	0.920000	0.420000
5	RF (Fine-tuned)	97.310000	103.100000	0.520000	0.480000	0.520000	0.470000
6	RF (Best Subset + Base)	16.050000	114.820000	0.920000	0.420000	0.920000	0.410000
7	RF (Best Subset + Fine-tuned)	97.230000	103.150000	0.520000	0.480000	0.520000	0.470000

Results with messwert_by_sys as the target

	Model	Train Mean Sq Error	Test Mean Sq Error	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
0	Tree (Base)	1.240000	358.490000	1.000000	0.040000	1.000000	0.040000
1	Tree (Fine-tuned)	162.760000	186.530000	0.560000	0.500000	0.560000	0.500000
2	Tree (Best Subset + Base)	201.700000	210.620000	0.450000	0.440000	0.450000	0.440000
3	Tree (Best Subset + Fine-tuned)	163.630000	185.670000	0.560000	0.500000	0.560000	0.500000
4	RF (Base)	25.890000	190.670000	0.930000	0.490000	0.930000	0.490000
5	RF (Fine-tuned)	161.510000	177.670000	0.560000	0.520000	0.560000	0.520000
6	RF (Best Subset + Base)	25.830000	190.380000	0.930000	0.490000	0.930000	0.490000
7	RF (Best Subset + Fine-tuned)	161.370000	177.850000	0.560000	0.520000	0.560000	0.520000

Summary

- Best Subset selection is used for feature selection.
- Four variations of trees and random trees are fitted to see different results.
- Data preprocessing and using feature selction technique we're able to improve the Linear Models.

Follow up questions:

1. Should we use new features like temperature that can have an effect on blood pressure reading?

Thank you!

Questions?