SAN Final Assignment

Martin Bulant, Aneta Furmanová, Daniel Klamrt, Jonáš Kříž

January 1, 2024

Contents

1 Introduction **2**

2 Question **2**

3 Dataset **3**
3.1 Suggested predictors 3
3.2 Data modality 4
3.3 Reweighing 5
3.4 Estimation by mean 5

4 Basic GLM Model **6**

5 Extended GLM Model **8**
5.1 GLM model with all suggested predictors 9
5.2 Regularization with LASSO 9
5.3 Sequential feature selection 12

6 Conclusions & Discussion **14**

7 Contribution statement **15**

References **15**

1 Introduction

In healthcare, prevention and prediction play a key role and in this assignment, we are going to focus on cardiovascular (CV) diseases. An important cardiovascular diseases and morbidity predictor is arterial stiffness [1]. Arterial stiffness can be estimated by multiple markers [2], e.g. non-invasively via Pulse Wave Velocity (PWV) measurement. PWV is the speed, at which the pulse wave propagates through the arterial wall.

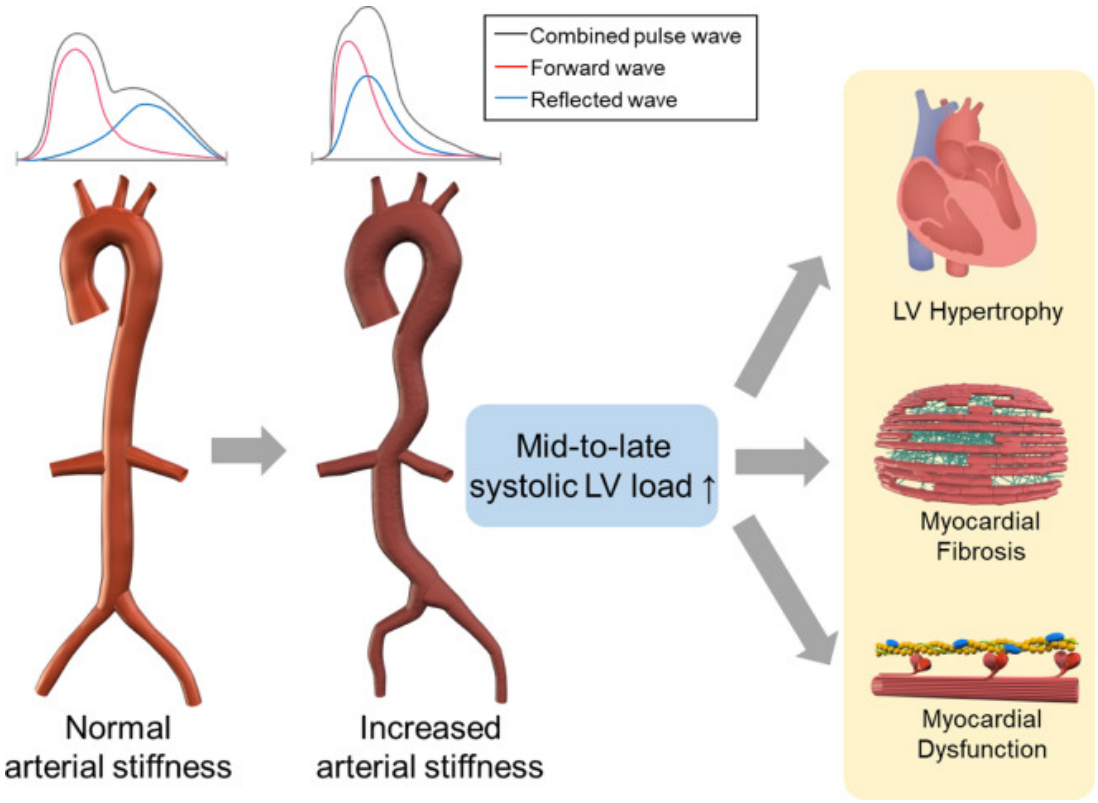


Figure 1: Arterial stiffness and pulse wave velocity [1]

For the aortic PWV estimation the distance jugulum - aortic bifurcation is needed. The real distance can be obtained from a chest MRI, but then it loses its advantage of being quite a cheap and fast method in daily practice. The currently used method for obtaining this distance is to use the arithmetic mean of the distances jugulum - umbilicus and jugulum - symphysis, where both are measured by the healthcare staff with a tailor's tape measure. Obviously, this is very impractical, as the measurement has a large inaccuracy, especially for obese patients. Hence, there is a need for more effective estimation of this distance in order to speed up the process of measurement and possibly reach better accuracy. The current state-of-the-art method will be used as the reference.

2 Question

How can be predicted the anatomical distance jugulum - aortic bifurcation without the need to measure the distances jugulum - umbilicus and jugulum - symphysis?

3 Dataset

The dataset was created during solving the project Apparatus for non-invasive automatic analysis of hemodynamic parameters (TH04010173) in the TAČR Starfos Programme. It contains data from 70 subjects as age, height, weight, sex, blood pressure, anamnesis, arm circumference above the elbow, measurements of anatomical distances (carotid - jugulum, jugulum - femoral artery, jugulum - umbilicus, jugulum - symphysis) and multiple PWV measurements by different devices.

The data were not randomly sampled. The sample subjects were drawn from the Czech Technical University in Prague, Faculty of Electrical Engineering among students, researchers and employees, who were willing to participate. Therefore, the findings and conclusions derived from this dataset cannot be extrapolated to the entire population. The selection bias can be partially avoided when we take into account the predictor distribution in population and according to that, reweight distribution of the given predictor in the dataset.

3.1 Suggested predictors

Based on Figure 2, the correlation between the predictor *age* and the dependent variable, the anatomical distance jugulum - aortic bifurcation, is very weak. Another weak correlation with the dependent variable appears with the predictor *diastolic blood pressure*. However, as this predictor is correlated with *systolic blood pressure*, it may be worth keeping the *diastolic blood pressure* in the set of predictors.

The suggested predictors are the following: *height, weight, BMI index, arm circumference above the elbow, systolic blood pressure, diastolic blood pressure* and *heart rate*.

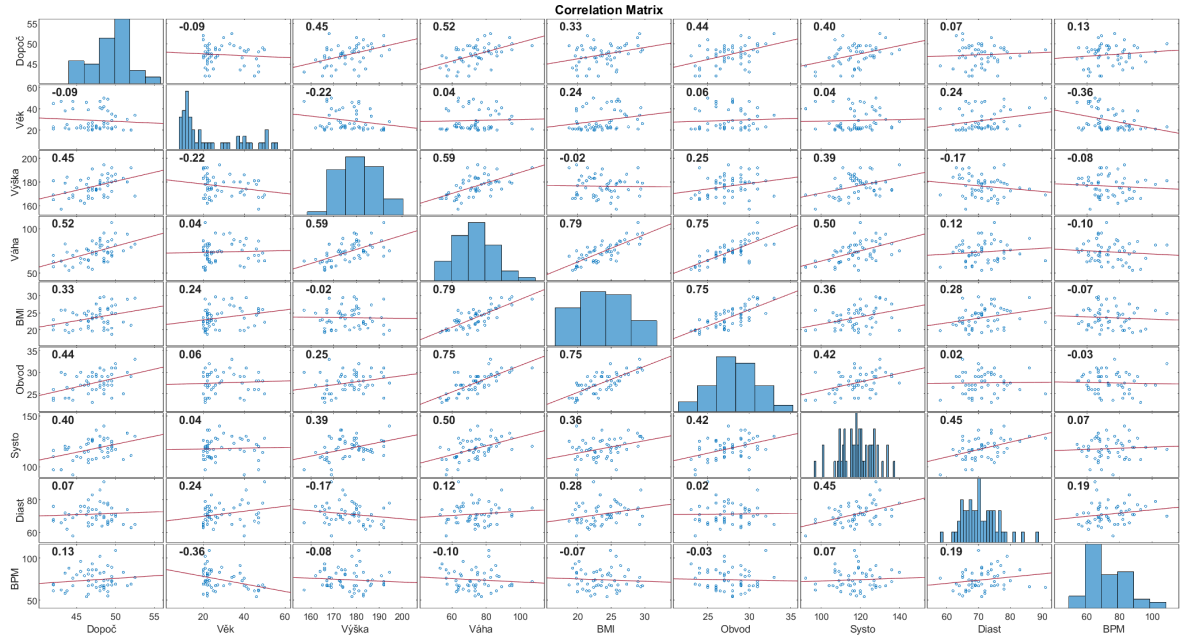


Figure 2: Correlation matrix of dependent variable and predictors

3.2 Data modality

By plotting how the predictor values were distributed, we observed that the distribution is often bimodal, as illustrated in Figure 3. The intuitive explanation was that this was caused by physiological differences between males and females.

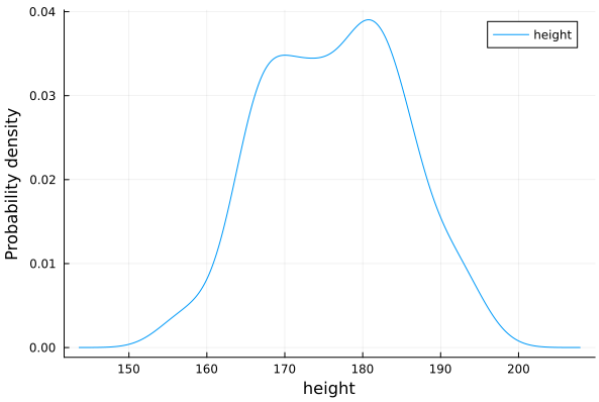
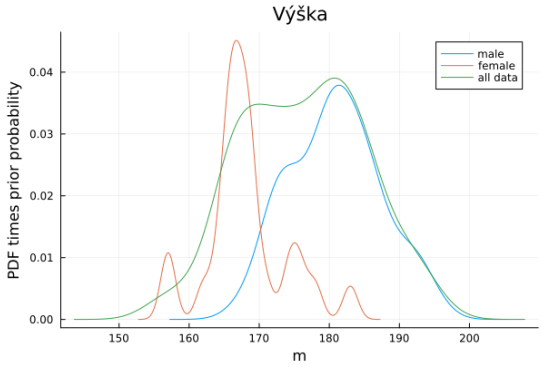
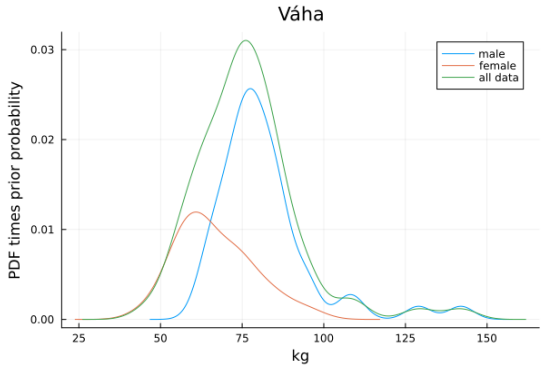


Figure 3: PDF of measured height of people in the dataset.

To test this, we split the data into two groups (male and female) and plotted their distributions again Figure 4. The split seems to explain the bimodality in cases of some predictor distributions.



(a) The presence of two sexes nicely explains the bimodality

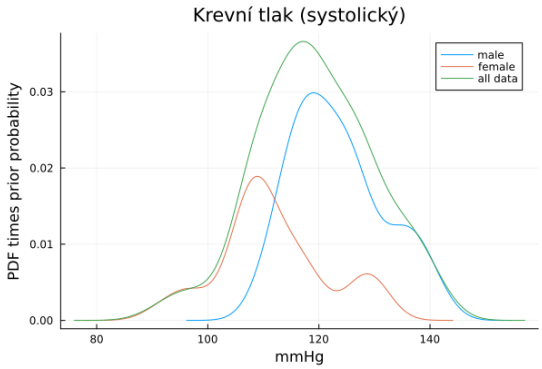


(b) In both cases.

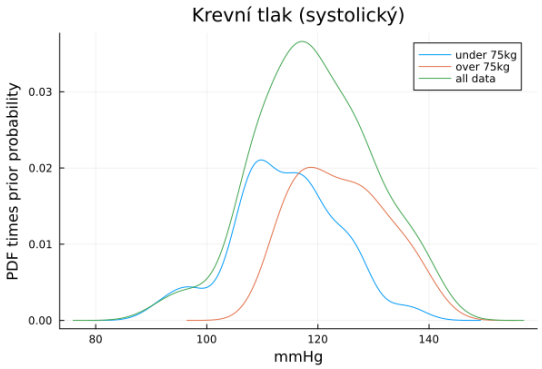
Figure 4: Bimodality of height and weight seems to be explained by the presence of two sexes.

For some predictors like heart rate and blood pressure, this split of the dataset by sex did not seem to explain the bimodality Figure 5a, because when we split the data by sex, the contributing distributions stayed bimodal. The predictor that seemed responsible for causing two peaks in the distribution of these predictors was weight, as can be seen in Figure 5b. Even though the contributing distributions in the case of splitting by weight still seem to be possibly bimodal, their bimodality can be easily explained by the two sexes.

In conclusion, it seems reasonable to, instead of training one GLM model to train over all the data, train three different GLMs one for females under 75kg, one for males under 75kg, and one model for both sexes (due to the limited size of the dataset) with weight over 75kg.



(a) After splitting by sex, the bimodality persists



(b) Weight splits the measured pressures into two groups

Figure 5: Bimodality of height and weight seems to be explained by the presence of two sexes.

3.3 Reweighing

By looking at the scatter plot Figure 6 of the measured weights in the dataset, we can see that half of the variance is represented only by eight people (people with weight over 95kg). This means assigning more weight to the error of these samples when training the GLM should reduce its error on real data. The value of the error weight of these samples should be $\frac{71}{8} = 8.875$.

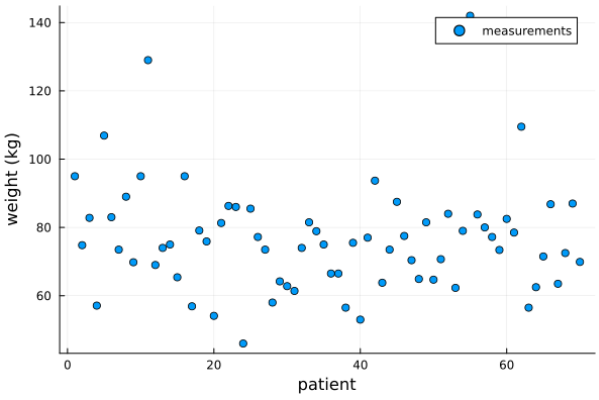


Figure 6: The dataset is under-represented for people with a weight above 95kg.

3.4 Estimation by mean

The MSE of mean estimate is shown in Table 1.

	whole dataset	$\leq$ 75kg	$>$ 75kg
both sexes	12.04	9.06	13.27
male	10.57	5.94	12.00
female	10.64	6.68	4.12

Table 1: MSE of mean estimate for the whole dataset and the dataset splits

4 Basic GLM Model

In this section, a basic GLM model will be created to predict anatomical length jugulum - aortic bifurcation depending on the predictors *height, weight, BMI index* from the dataset. These predictors were chosen due to the fact that they are the most easily obtained. The dependent variable anatomical distance jugulum - aortic bifurcation will be for brevity from now on denoted as y. Figure 7 shows the dependence of y on individual predictors and can be seen that there are a few outliers.

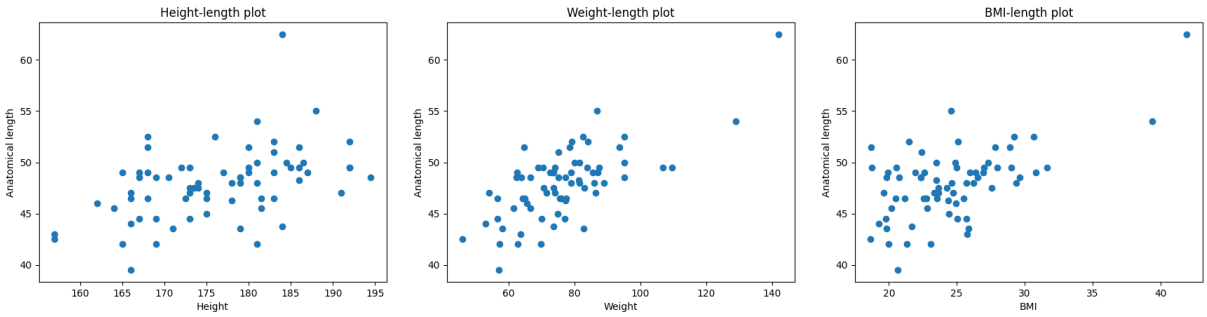
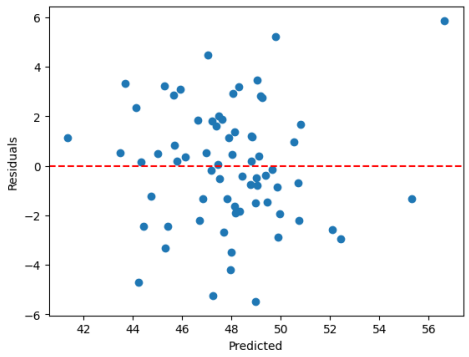
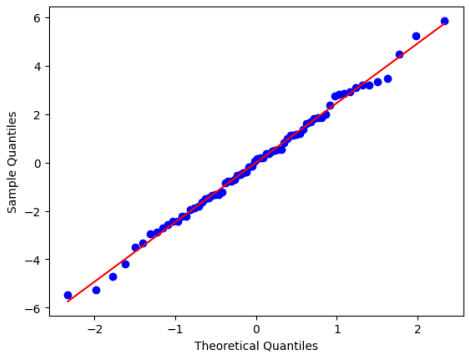


Figure 7: Dependency of the anatomical distance on the patient's height, weight and BMI index, respectively.

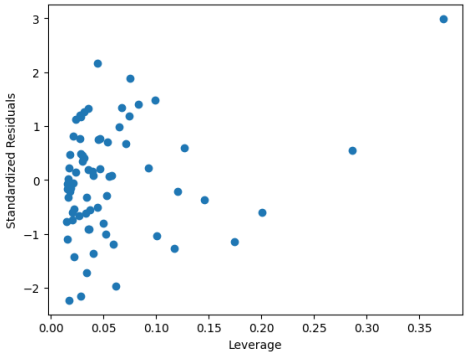
From the diagnostic plots in Figure 8 it can be seen that the assumptions of linearity, homoscedasticity and normal distribution of residues are observed. As a result, the OLS model could be used.



(a) Residuals vs fitted



(b) Q-Q plot



(c) Residuals vs Leverage

Figure 8: Diagnostic plots

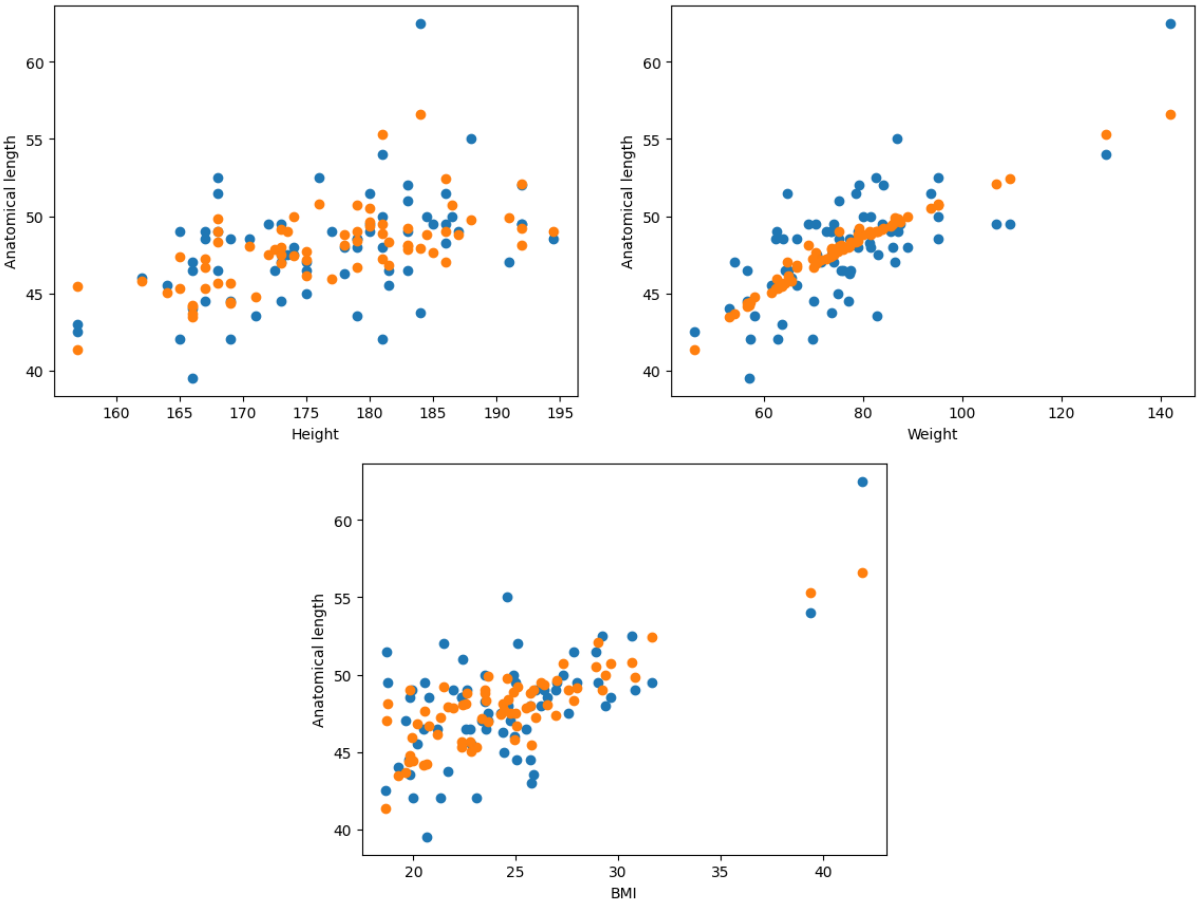


Figure 9: Comparison of OLS model prediction against reality

In Figure 9, the comparison of the OLS model's prediction against reality can be seen. The linear model's equation is as follows

$$y = -15.5949 + 0.2946 \cdot x_{\text{height}} - 0.1559 \cdot x_{\text{weight}} + 0.9575 \cdot x_{\text{BMI}} \ .$$

The F-statistic of the linear model is 23.95, so the model is better than mean model. It explains 52.1% of the variance in the data, and the AIC is 329.3. Model errors can be seen in Table 2.

	Mean	Sd
MSE	7.08	5.04
MAE	2.08	0.71

Table 2: GLM model for predictors *height, weight* and *BMI index*, errors.

5 Extended GLM Model

In this section, it will be estimated, whether the inclusion of other predictors from the dataset contributes to the improvement of the model compared to relying solely on the predictors *height, weight* and *BMI*. The tested predictors are following: *height, weight, BMI index, arm circumference above the elbow, systolic blood pressure, diastolic blood pressure* and *heart rate*.

In section subsection 3.2, it was supposed that splitting the dataset into three groups because of the data modality may help to obtain a more suitable model. The groups were suggested as follows: females under 75 kg, males under 75 kg, both sexes above 75 kg. The threshold weight 75 kg was selected as a median of the weights in the dataset. However, this threshold leads to splitting the dataset into groups of unequal sizes (of 21 samples for the first group, 15 samples for the second and 32 samples for the third group). This lead to the fact that some of the models for the smallest group were not converging. Hence, it was decided to shift the weight threshold to 78 kg in order to obtain more equal partitions of the dataset (of 21 samples for the fist group, 21 samples for the second and 26 samples for the third group). Another suggestion involved reweighing the errors of samples with a weight over 95 kg (mentioned in subsection 3.3). Despite the fact that the predictions remained the same after reweighing and that the AIC of the reweighted model is higher, there is a possibility that the reweighted model may exhibit increased reliability when applied to real-world data.

For all the following section, the data were standardized with `StandardScaler` from package `sklearn`. Estimated values of mean and scale for the whole dataset and for the splits can be seen in Table 3. Note, that all the following models were made using standardized data samples. Hence, in order to obtain a prediction for a new sample, the sample has to be firstly standardized according to this table and then the result needs to be transformed back.

	whole dataset		female $\leq$ 78 kg		male $\leq$ 78 kg		both $>$ 78 kg	
	mean	scale	mean	scale	mean	scale	mean	scale
y (cm)	47.84	3.49	45.50	2.81	47.26	2.12	50.13	3.40
h (cm)	176.54	8.64	167.29	5.53	179.50	6.44	181.63	5.91
w (kg)	76.35	16.02	62.89	7.87	71.75	4.52	90.95	15.11
BMI (kg/m^2)	24.40	4.22	22.45	2.43	22.37	2.31	27.61	4.55
AC (cm)	28.22	3.15	25.95	2.82	27.40	1.26	30.71	2.71
BP sys (mmHg)	119.15	10.41	108.95	7.64	123.90	8.00	123.54	7.94
BP dia (mmHg)	71.93	6.71	71.19	6.29	72.10	7.12	72.38	6.65
HR (bpm)	74.66	12.66	76.71	11.19	70.29	12.38	76.54	13.12

Table 3: Estimated values of mean and scale for depending variable and predictors.

5.1 GLM model with all suggested predictors

Firstly, a full GLM model with all suggested predictors was made, so we can later compare it with models after feature selection or regularization, or with models for different groups of people from the dataset.

The accuracy of prediction by the GLM model with all suggested predictors can be seen in Figure 10, the scaled dependent variable is given by

$$y = 0.616 \cdot x_{\text{height}} - 0.490 \cdot x_{\text{weight}} + 0.853 \cdot x_{\text{BMI}} + 0.192 \cdot x_{\text{arm circumference}} - 0.084 \cdot x_{\text{BP systolic}} - 0.003 \cdot x_{\text{BP diastolic}} - 0.006 \cdot x_{\text{heart rate}} .$$

The model AIC is 155.56 and the errors can be seen in Table 4.

	Mean	Sd
MSE	8.36	4.29
MAE	2.20	0.63

Table 4: GLM model with all suggested predictors, errors.

Both Figure 10 and Figure 11 show that the residuals are almost evenly distributed, but there are some outliers, as was mentioned in subsection 3.2. According to the Q-Q plot, the predictors and the dependent variable are linearly related (as the points lie approximately on one line). Because the line formed by the points is flatter than the line $y = x$, it can be deduced that the distribution of predictors is more dispersed than the distribution of dependent variable.

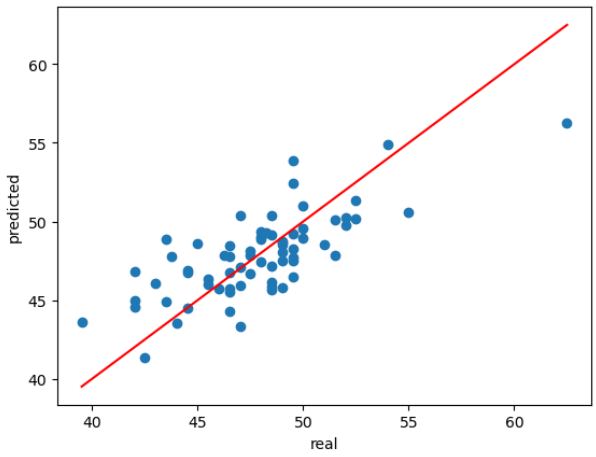


Figure 10: GLM model with all suggested predictors, real y vs predicted y.

5.2 Regularization with LASSO

As our aim is an effective estimation of y, this means that the amount of the predictors should be as low as possible while still maintaining a reasonable prediction. Hence, a regularization with LASSO was employed. The optimal α for LASSO regression was obtained by cross-validation on 10 splits. With $\alpha = 0.0359$, the model reduced to zero the coefficients of following predictors: *BMI index, systolic blood pressure, diastolic blood*

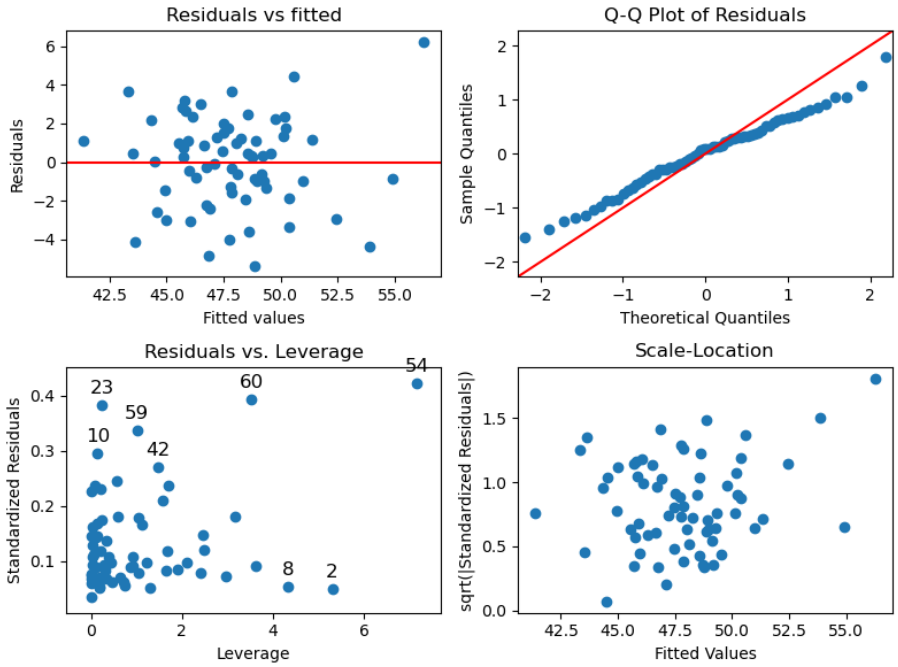


Figure 11: GLM model with all suggested predictors, diagnostic plots.

pressure, heart rate. Then, the dependent variable is given by

$$y = 0.145 \cdot x_{\text{height}} + 0.451 \cdot x_{\text{weight}} + 0.186 \cdot x_{\text{arm circumference}} \ .$$

According to the comparison of LASSO model errors in Table 5 with errors of GLM model with all suggested predictors in Table 4, the regularization helped to achieve slightly better results.

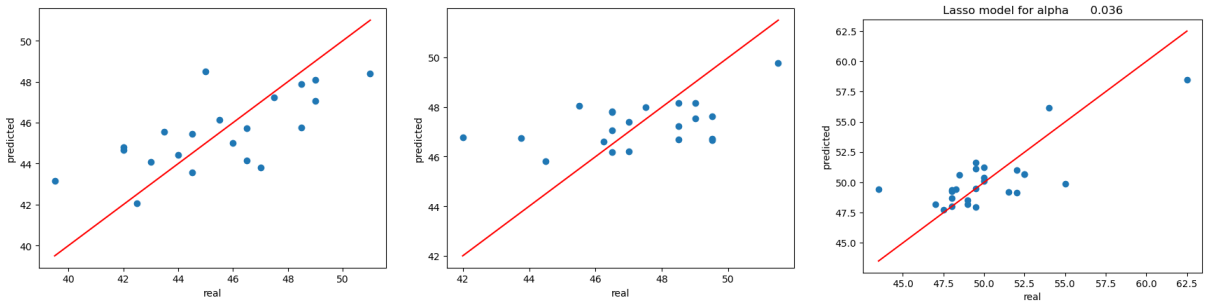
	Mean	Sd
MSE	7.25	3.95
MAE	2.03	0.66

Table 5: LASSO model for all data, errors.

As was suggested in section subsection 3.2, because of the data modality, it should be tested whether a performance of three models on dataset splitted into three groups (females under 78kg, males under 78kg, both sexes above 78kg) is better than the performance of a model made for the whole dataset.

The regression coefficient α was kept the same as in the previous LASSO model. The relationships for the three dataset splits are given by

$$\begin{aligned}
y_{\text{female} \leq 78 \text{ kg}} &= 0.494 \cdot x_{\text{height}} + 0.374 \cdot x_{\text{arm circumference}} \\
&\quad + 0.072 \cdot x_{\text{BP systolic}} + 0.171 \cdot x_{\text{BP diastolic}} - 0.101 \cdot x_{\text{heart rate}} \ , \\
y_{\text{male} \leq 78 \text{ kg}} &= 0.232 \cdot x_{\text{height}} - 0.325 \cdot x_{\text{weight}} + 0.345 \cdot x_{\text{arm circumference}} \\
&\quad + 0.028 \cdot x_{\text{BP systolic}} + 0.017 \cdot x_{\text{heart rate}} \ , \\
y_{> 78 \text{ kg}} &= 0.117 \cdot x_{\text{weight}} + 0.602 \cdot x_{\text{BMI}} + 0.177 \cdot x_{\text{arm circumference}} \\
&\quad - 0.344 \cdot x_{\text{BP systolic}} - 0.116 \cdot x_{\text{BP diastolic}} \ .
\end{aligned}$$



(a) females under 78 kg (b) males under 78 kg (c) above 78 kg

Figure 12: LASSO models for dataset splitted into three groups, real y vs predicted y.

Even though not much predictor coefficients were reduced by LASSO to zero and, hence, it would help to reduce the α, for smaller α the model for males with weight under 78 kg was not converging. Although the prediction for females under 78 kg in Figure 12a and the Q-Q plot in Figure 14a seem better than the GLM model with all suggested predictors, in terms of the errors (in Figure 13), both the LASSO model for all data and GLM model with all suggested predictors are performing better than the all three LASSO models build for the three groups.

	Mean	Sd
MSE	8.30	8.13
MAE	2.37	1.35

(a) females under 78 kg

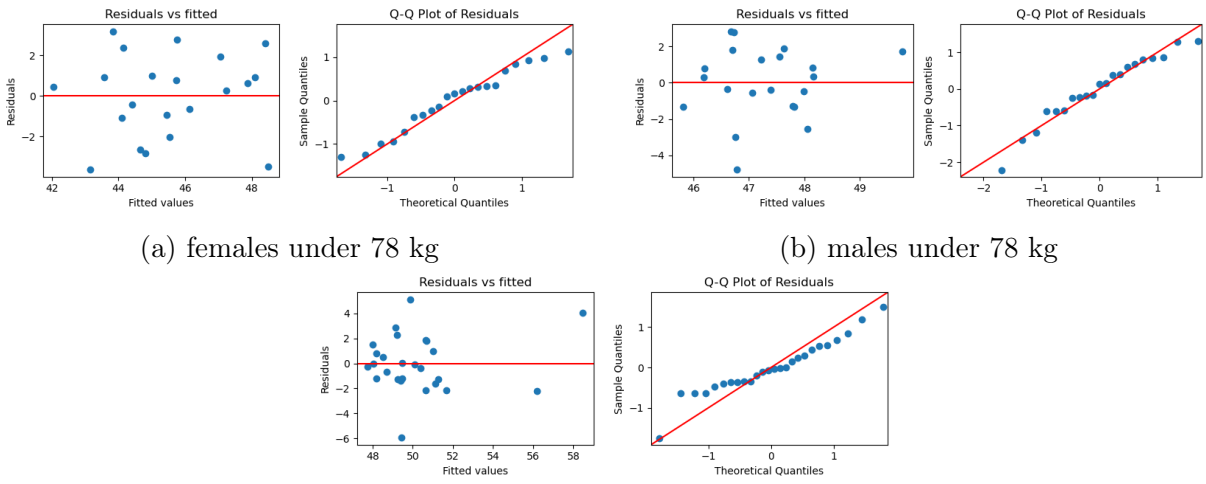
	Mean	Sd
MSE	6.87	6.64
MAE	2.11	1.21

(b) males under 78 kg

	Mean	Sd
MSE	9.56	8.60
MAE	2.29	0.98

(c) above 78 kg

Table 13: LASSO models for dataset splitted into three groups, errors.



(a) females under 78 kg (b) males under 78 kg

(c) above 78 kg

Figure 14: LASSO models for dataset splitted into three groups, diagnostic plots

5.3 Sequential feature selection

Another possibility of optimizing the amount of the predictors while keeping reasonable prediction of y is a sequential feature selection. In this work, Backward stepwise selection (BSS) was employed.

Threshold for the p-value of predictors 0.3 was chosen for BSS in order to obtain at least three predictors and also minimize AIC in the resulting model. Feature removed by BSS were in the following order: *diastolic blood pressure, heart rate, weight, systolic blood pressure*. The model is then given by

$$y = 0.374 \cdot x_{\text{height}} + 0.401 \cdot x_{\text{BMI}} + 0.194 \cdot x_{\text{arm circumference}} \ . \tag{1}$$

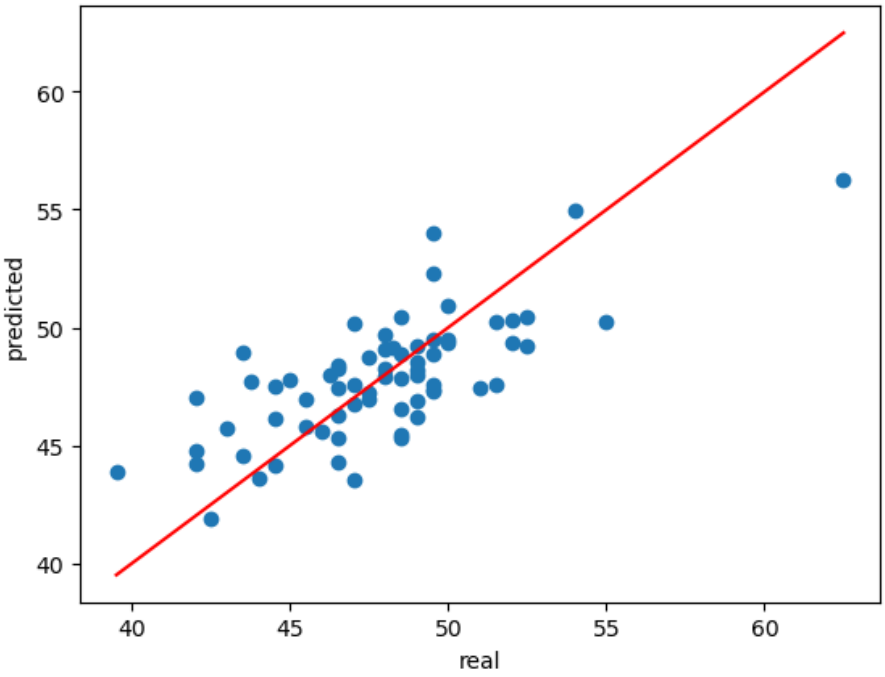


Figure 15: GLM model after BSS, real y vs predicted y.

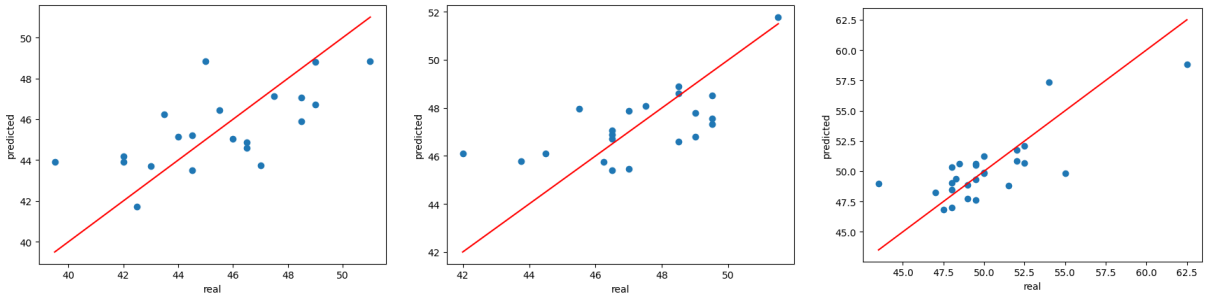
The model AIC is 148.35 and the errors can be seen in Table 6. According to AIC and mean errors, GLM model after BSS is better than GLM model with all suggested predictors.

	Mean	Sd
MSE	6.78	3.67
MAE	1.99	0.62

Table 6: GLM model after BSS, errors.

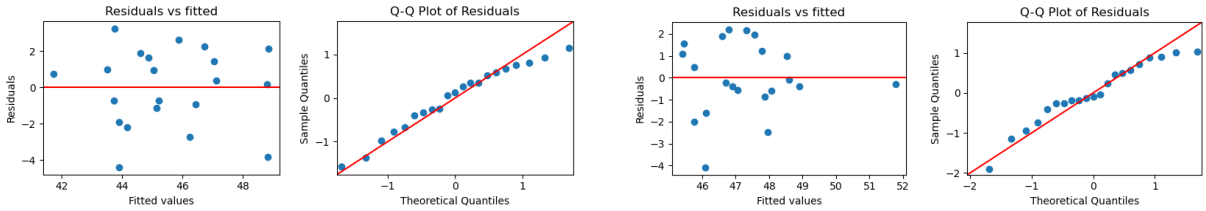
Splitting the dataset into three groups leads to predictions given by

$$\begin{aligned} y_{\text{female} \leq 78 \text{ kg}} &= 0.492 \cdot x_{\text{height}} + 0.433 \cdot x_{\text{arm circumference}} + 0.219 \cdot x_{\text{BP systolic}} \ , \\ y_{\text{male} \leq 78 \text{ kg}} &= 5.982 \cdot x_{\text{height}} - 5.599 \cdot x_{\text{weight}} + 8.345 \cdot x_{\text{BMI}} \\ &\quad + 0.457 \cdot x_{\text{arm circumference}} \ , \\ y_{> 78 \text{ kg}} &= 2.313 \cdot x_{\text{height}} - 5.880 \cdot x_{\text{weight}} + 6.716 \cdot x_{\text{BMI}} \\ &\quad + 0.221 \cdot x_{\text{arm circumference}} - 0.473 \cdot x_{\text{BP systolic}} \ . \end{aligned}$$

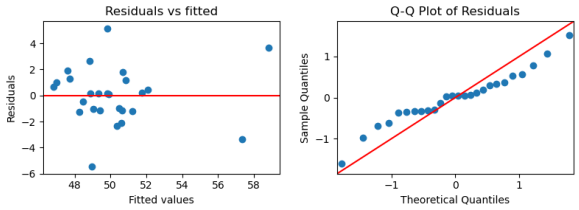


(a) females under 78 kg (b) males under 78 kg (c) above 78 kg

Figure 16: GLM model after BSS for dataset splitted into three groups



(a) females under 78 kg (b) males under 78 kg



(c) above 78 kg

Figure 17: GLM model after BSS for dataset splitted into three groups, diagnostic plots.

As can be seen in Figure 16 and Figure 18, the performance of models for the first two groups (females under 78 kg, males under 78 kg) are better than in all the previous models. The high values of MSE for the last group is probably caused by the fact that the samples are not evenly distributed (the outliers are also visible in Figure 16c and Figure 17c).

	Mean	Sd
MSE	6.59	5.50
MAE	2.20	1.05

(a) females under 78 kg

	Mean	Sd
MSE	4.39	4.11
MAE	1.75	0.91

(b) males under 78 kg

	Mean	Sd
MSE	10.03	11.27
MAE	2.23	1.25

(c) above 78 kg

Table 18: LASSO models for dataset splitted into three groups, errors.

6 Conclusions & Discussion

The goal was to find an estimate of the distance jugulum - aortic bifurcation for later use in the measurement of Pulse Wave Velocity without needing to use a tailor's tape measure to obtain the predictors. Therefore, a question was posed: *How can be predicted the anatomical distance jugulum - aortic bifurcation without the need to measure the distances jugulum - umbilicus and jugulum - symphysis?*

Multiple models were trained to determine whether the anatomical distance jugulum - aortic bifurcation can be predicted without the need to measure the distances jugulum - umbilicus and jugulum - symphysis. Firstly, a basic GLM model was created, depending on the predictors *height, weight, BMI index*. This model explains 52.1% of the variance in the data, and according to the F-statistic, the obtained linear model is better than the mean model.

Then, the use of other predictors was tested. It had to be dealt with the bimodal distribution of predictors, and hence, splitting a dataset into three exclusive groups was suggested. The extended GLM model with all predictors (i.e., *height, weight, BMI index, arm circumference above the elbow, systolic blood pressure, diastolic blood pressure* and *heart rate*) in subsection 5.1 performs similarly to the basic GLM model in section 4. This means that adding all the other predictors to the model have not improved the performance of the model.

As the aim is also to have as low amount of predictors as possible, LASSO regularization was employed. From its result, it can be deduced, that the three most reliable predictors are *height, weight* and *arm circumference above the elbow*. It was also tested whether three different models on the dataset split into three groups (females under 78kg, males under 78kg, both sexes above 78kg) will outperform a model made for the whole dataset. This was not the case, probably because then the data splits became too small and the last group (both sexes above 78 kg) is not evenly distributed.

Further, a backward stepwise selection was used. The features kept by BSS done on the GLM model for the whole dataset were *height, BMI index* and *arm circumference above the elbow*. Splitting the dataset into three groups leads in the case of the first two groups (females under 78kg, males under 78kg) to models that could be used for predicting the dependent variable in the real world. Regarding the last group (both sexes above 78 kg), more evenly distributed data would be needed in order to obtain a better model.

To conclude, the anatomical distance jugulum - aortic bifurcation can be predicted without the need to measure the distances jugulum - umbilicus and jugulum - symphysis with MAE < 3 cm. The predictors mostly selected by all the LASSO and BSS models are *height, weight, BMI index* and *arm circumference above the elbow*. The best model (1) was obtained with GLM and BSS on the whole dataset (shown in Figure 15).

7 Contribution statement

Total time per team: 57 h

Martin Bulant: 20 % (5 h data preprocessing, 7 work & report)

The involvement of this team member was to firstly to preprocess the dataset and then determine, if the anatomical distance jugulum - aortic bifurcation can predicted only with predictors *height, weight,* and *BMI index* (as those are the easiest to measure). To accomplish this, the team member took on the task of assessing the basic GLM model and conducting statistical tests to determine its viability and significance.

Aneta Furmanová: 35 % (6 h plan, 14 work & report)

The contribution of this team member was making of extended GLM models (including all suggested predictors) based on suggestions from section Data modality. Additionally, the assembling of components contributed by other team members into the final report and shaping both the introduction and conclusion sections.

Daniel Klamrt: 13 % (2 h plan, 5 work & report)

This team member participated in estimating the probability distribution of data using KDE, verifying the normality of data through the Shapiro–Wilk test, and identifying the most suitable predictors.

Jonáš Kříž: 32 % (7 h plan, 11 work & report)

This member of the team analyzed the dataset in order to supply insight into the future planning of the model architecture. Namely checking whether some distributions of predictors or the predicted variable do not happen to be multimodal and checking if the whole domain of the data is sufficiently represented for possible reweighing.

References

[1] Toru Miyoshi and Hiroshi Ito. "Arterial stiffness in health and disease: The role of cardio-ankle vascular index". In: *Journal of Cardiology* 78.6 (2021), pp. 493–501. ISSN: 0914-5087. DOI: https://doi.org/10.1016/j.jjcc.2021.07.011.

[2] Patrick Segers, Ernst R. Rietzschel, and Julio A. Chirinos. "How to Measure Arterial Stiffness in Humans". In: *Arteriosclerosis, Thrombosis, and Vascular Biology* 40.5 (May 2020), pp. 1034–1043. ISSN: 1524-4636. DOI: 10.1161/atvbaha.119.313132.