ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Predictive Modeling with Asymmetric Loss Functions

Panagiotis Anastasakis

Supervisor: Dr. Yiannis Kamarianakis

Department of Mathematics & Applied Mathematics
University of Crete
Heraklion, Greece

October 15, 2023

Contents

1. Introduction

The field of statistical learning has experienced significant growth over the past century, witnessing remarkable advancements in both theoretical understanding and computational techniques. In response to the increasingly complex demands for extracting valuable information from data, new approaches for fitting regression models have gained popularity.

Traditionally, regression models have relied on symmetric loss functions, such as the quadratic or absolute loss, to determine the best fit. These functions measure the magnitude of errors symmetrically, treating positive and negative errors equally. However, in many real-world applications, certain types of errors carry more significance than others. For example, an over-prediction of a specific response may have more severe consequences than an under-prediction. This crucial aspect needs to be considered during statistical analysis.

To address this asymmetry, methodologies relying to asymmetric loss functions have been developed. Unlike their symmetric counterparts, these functions assign unequal penalties to different directions of loss, aiming to account for the importance of the aforementioned asymmetry. By incorporating prior knowledge about the data's nature and the goals of inference, these asymmetric loss functions can lead to more robust and reliable results, making them highly valuable in practical settings.

1.1 About the Thesis

The primary objective of this thesis is to introduce and explore the concept of Quantile Regression as an alternative approach to the conventional Least Squares. To provide a comprehensive understanding of the topic, the thesis will begin by examining widely used techniques in Least Squares, including Linear Regression, feature selection algorithms and model averaging. Subsequently, the focus will shift to penalized Least Squares, specifically Ridge and Lasso Regression which are known for their effectiveness in numerous contexts, both regarding predictive performance and interpretational value.

Following the comprehensive presentation of various Least Squares techniques, we will proceed with an in-depth exploration of Quantile Regression, which is the pioneering method for fitting linear models through asymmetric loss functions. This study will tackle various aspects of Quantile Regression, involving its theoretical foundation, practical implementation, and exploratory properties. Additionally, penalized versions of Quantile Regression will be examined, which aim to undermine the influence of irrelevant predictors on the response variable.

Motivating Application

In order to motivate the use of Quantile Regression over Least Squares, we will apply the discussed methodology to a real-life problem involving the prediction of heart volume for pediatric patients considered as candidates for heart transplantation. This task holds significant importance because even though direct measurements of the patient's heart volume cannot be obtained prior to transplantation, an estimated value is necessary for

selecting a compatible allograft to accommodate the recipient.

The models developed for the prediction of heart volume are not intended to serve as the sole determining factor for compatibility in heart transplantation. Instead, they should be used as a valuable tool for physicians to aid in their assessment of donor-recipient compatibility. The predictions generated by these models should be considered in conjunction with other clinical factors and the expertise of medical professionals when making transplantation decisions.

Given the importance of interpretability in this context, a model that can provide additional information about the importance of variables as well as their contribution in the final prediction is preferable. Linear models are particularly well-suited for this purpose, as they allow for clear and direct interpretation of the relationships between variables. However, a linear model also makes some very strong assumptions that are often violated in practice. In order to make our predictions more robust when it comes to these assumptions, extensions of the standard linear model will be considered.

In our specific application, the utilization of an asymmetric loss function for obtaining coefficient estimates in the linear model is highly relevant. In the context of heart transplantation, the consequences of oversizing and undersizing the transplanted heart are not equal in terms of the patient's outcome. While oversizing may result in compression effects producing physiological complications, undersizing carries the risk of cardiac output insufficiency, which is considered to be clinically more dangerous [1]. This asymmetry in the importance of errors can be incorporated into our statistical analysis by employing an asymmetric loss function to properly inform the regression model. With the use of Quantile Regression, we can appropriately weigh the penalties assigned to different directions of loss, thus accounting for the specific risks associated with undersizing.

The Data

The data used for this application involves simple somatometric measurements of pediatric patients along with their corresponding heart volumes, which are the target variable. It is a small dataset, consisting of 58 samples and the following 6 features:

- **Male** $\rightarrow$ the patient's gender (1 for male, 0 for female)
- **Age** $\rightarrow$ the patient's age (mo)
- **Ht** $\rightarrow$ the patient's height (cm)
- **Wt** $\rightarrow$ the patient's weight (kg)
- **BMI** $\rightarrow$ the patient's BMI (Body Mass Index) (kg/m^2)
- **BSA** $\rightarrow$ the patient's BSA (Body Surface Area) (m^2)

The response variable is **HtVol**, which corresponds to the volume of the patient's heart (mL).

This dataset was first examined by the manuscript titled: 'Alternative methods for virtual heart transplant—Size matching for pediatric heart transplantation with and without donor medical images available', published in 2018 by Jonathan D. Plasencia et al.[1].

A brief inspection of the data informs us that there are 4 patients who are over 18 years old (or 216 months old). Since we are interested in pediatric patients only, we remove these samples corresponding to adults and now our dataset consists of 54 samples.

1.2 Limitations of Current Practices for Assessing Donor-Recipient Compatibility

Matching a candidate patient with an appropriate heart transplant poses a great challenge for physicians due to the numerous factors that can affect compatibility. In order for the experts to accept a donor's heart, several restrictive conditions must be met, resulting to a significantly limited donor pool for the patients. One of the most difficult criterion to assess is whether a heart is of an appropriate size to provide the recipient with the best chances for long term survival. A potential extreme oversizing, meaning that the donor's heart is considerably bigger, can result in the allograft not fitting comfortably inside the chest cavity which can lead to serious complications and pose an early mortality risk [2]. On the other hand, undersizing has been determined to be of great risk regarding the recipient's survival ([3]) and when it comes to pediatric patients, undersized allografts are strongly discouraged [4] .

The DRBW Ratio Criterion

The task of finding an appropriately-sized transplant has been undertaken by numerous researchers with the current standard involving weight criteria. Specifically, the donor-recipient bodyweight ratio, or DRBW ratio, has been the most common clinical practice for assessing compatibility between the donor and the patient [5]. It has been suggested that the weight of a person is heavily correlated with the size of their heart, therefore a weight comparison is considered by many sufficient for determining whether the transplant will fit the patient.

The DRBW ratio is given by the simple formula

$$DRBW = \frac{\text{Donor weight}}{\text{Recipient weight}}$$

If the ratio lies within a certain range, the potential donor's heart is accepted. The acceptance range is a topic of extensive research with transplant centers often selecting different ranges for matching the allograft [6]. Generally accepted transplant criteria limit the use of cardiac allografts from donors to within approximately 20% of the recipient's weight [7]. In terms of the weight ratio, this translates to having $0.8 < DRBW < 1.2$.

However, when it comes to pediatric patients these criteria are not as simple. There are other considerations to be taken into account involving the growing anatomy of the patient in addition to a variety of pathologies [8]. Nonetheless, clinics continue to opt

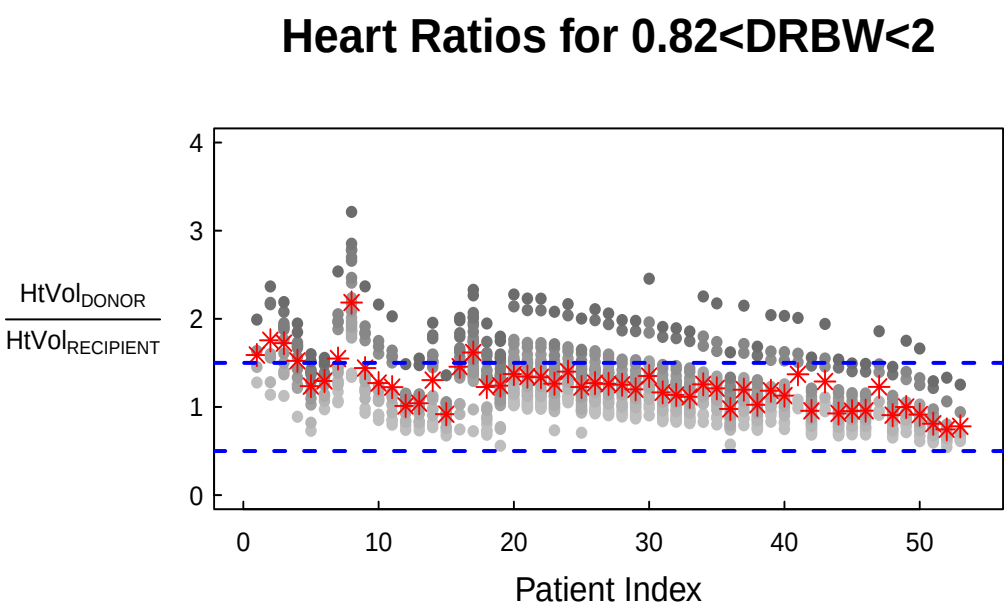
for the alarmingly simple DRBW ratio criterion, but in this case a wider range is chosen, leading to accepting more oversized donors. Between 2007 and 2017, it was found that 28 transplant centers in the US (out of 61 in total) had a median of 200% for the upper weight ratio percentage limit and a median of 82% for the lower limit [6], which is equivalent to $0.82 < DRBW < 2$.

Invalidating the DRBW Ratio Approach

In order to further motivate our application, we will provide evidence regarding the limitations of the DRBW ratio criterion when it comes to pediatric patients:

Starting off, a statistical analysis of the wider range for the DRBW ratio will be performed, considering $0.82 < DRBW < 2$ and testing it in our dataset to inspect its effectiveness. For that, we will consider one patient in our data to be the recipient, with the rest representing the potential donor pool. We will then select the donors that DRBW ratio allows us to, repeating this process for every patient out of the 54 we have in total. Then, we will calculate the heart ratios for each pair of donor-recipient, which we will display with the use of a pointplot.

Implementing this process as described above results in the following figure:



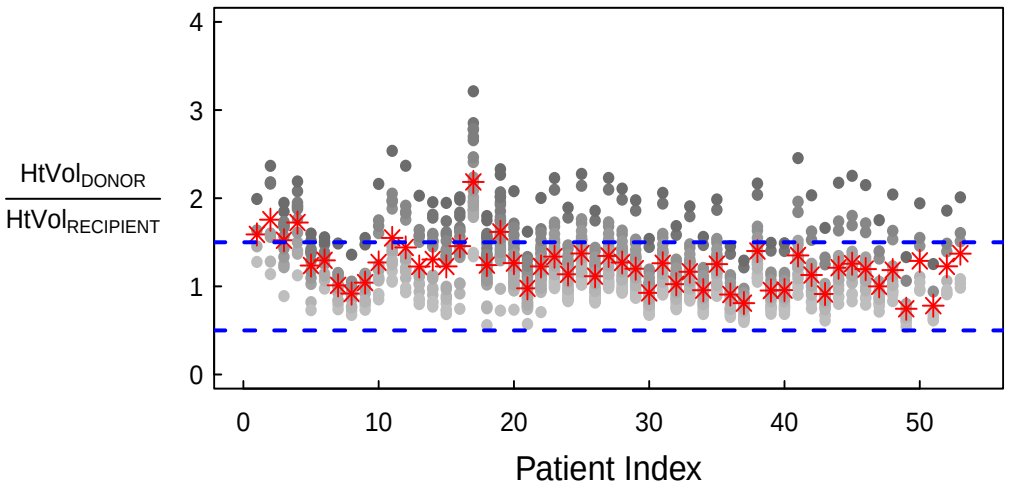


Figure 1: Pointplots for the wider DRBW ratio range. The indices are sorted by heart volume on top figure and by weight on bottom figure, both from smallest to largest value. The grey dots represent the heart ratios of the donors over the recipients, plotted against the patient index, with the red points being the mean value of all the ratios. The blue dashed lines are drawn at the values 0.5 and 1.5 of y-axis, serving as a rough indication for oversized and undersized allografts (50% deviation from the recipient's heart).

First of all, looking at the figures we can observe that no heart ratios correspond to extreme undersizing, with all being over 0.5. On the other hand, almost every patient has been matched with donors that have a heart over 50% bigger, a situation which is more apparent in the smaller volumes. Indeed, by inspecting the first figure we can see that as heart volumes increase, both the number of oversized donors and the extend of oversizing per case reduces. This is of course expected since in our limited dataset as the volume increases there are fewer potential donors with a bigger heart. Nonetheless, it is worth noticing that even when it comes to the biggest hearts, no undersizing of over 50% is present.

When sorting the patients based on their weight however, the occurrences of oversizing are distributed equally across the entire weight range. This informs us that none of the weight classes is distinctively prone to be matched with oversized donors.

Given the number of incompatible donors for each patient, we can further examine the likelihood of failing to find a matching donor with the wider DRBW ratio. For that, we can estimate the probability for each patient to be matched with a heart of an inappropriate size as the fraction of donors with a non-matching heart over all possible

donors as identified by the weight criterion. We will calculate two probabilities:

- The first probability involves counting a heart as incompatible when the size is over 50% off from the recipient's heart volume:

$$\frac{\text{HtVol}_{\text{DONOR}}}{\text{HtVol}_{\text{RECIPIENT}}} < 0.5 \quad \text{or} \quad \frac{\text{HtVol}_{\text{DONOR}}}{\text{HtVol}_{\text{RECIPIENT}}} > 1.5 \tag{1}$$

- For the second probability, any undersizing along with 50% oversizing will be considered an incompatibility:

$$\frac{\text{HtVol}_{\text{DONOR}}}{\text{HtVol}_{\text{RECIPIENT}}} < 1 \quad \text{or} \quad \frac{\text{HtVol}_{\text{DONOR}}}{\text{HtVol}_{\text{RECIPIENT}}} > 1.5 \tag{2}$$

It is clear that the first probabilities will be greater than the second because all instances of 2 are contained in 1.

It is important to point out that the estimated probabilities will be biased due to the limited sample size we are working with. For instance, the patient with the lowest heart volume (with respect to the data) is more likely to be incompatible with a donor since the DRBW ratio will only select donors with bigger hearts. Similarly, patients with higher volumes are less likely to come across a severe oversizing, but there is a higher chance for undersizing. This is due to the fact that the donor pool is limited to our data, thus most hearts will be of equal size or smaller.

In order to remove some of the bias, we can focus on patients with neither too high nor too low heart volume. For that, patients with volumes outside of the range of 1st and 3rd quartiles can be ignored, thus extracting information from the middle range of the distribution only.

The following figure contains the probabilities as derived from both criteria:

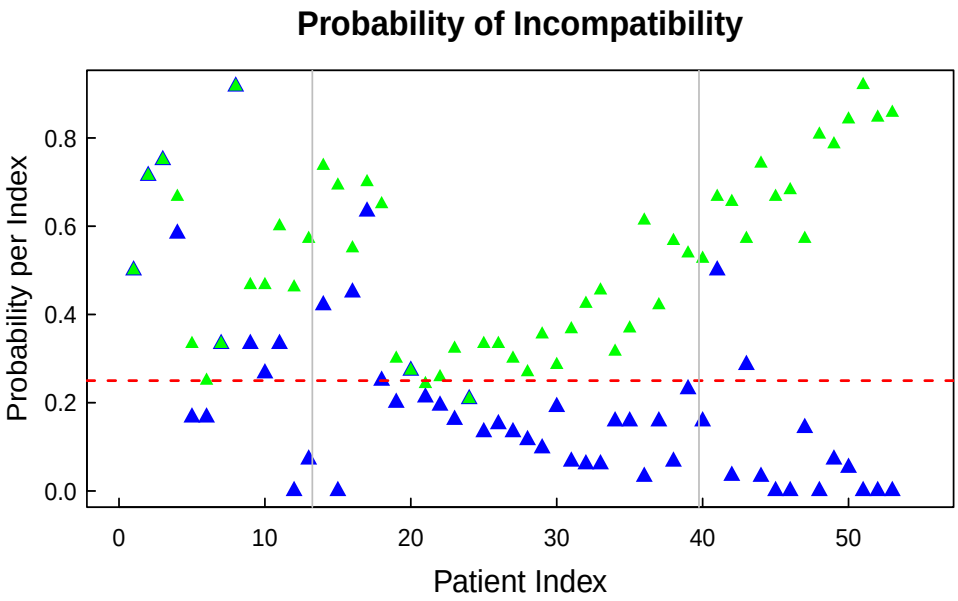


Figure 2: Scatterplot for probability of incompatibility. The blue points correspond to the more relaxed criterion 1 while the green ones correspond to 2. The red dashed line represents the incompatibility probability of 0.25 and the grey vertical lines indicate the 1st and 3rd quartile of the distribution. In order to limit the bias, we only examine the points between these lines.

The scatter plot verifies that all green points indeed correspond to higher probabilities since they are above the blue points for the same indices. Almost all blue points are below 0.25, informing us that when we allow smaller hearts, the chance of an incompatibility occurring is relatively low. On the other hand, in a more realistic scenario where donors with lower heart volume are turned down, is is significantly more likely for a mismatch to occur. All green points indicate probabilities over 0.25 with some even being over 0.5. Due to the fact that a mismatch may prove to be fatal for the recipient, such high probabilities are very concerning for the wider DRBW ratio criterion to be blindly trusted.

Given that the wide weight criterion has failed to systematically achieve a correct match between a donor and a patient, we will now inspect the more strict ratio, where compatibility is decided when $0.8 < DRBW < 1.2$. Working similarly as before, considering for each patient of the data to be the recipient with the rest representing the donor pool, the following pointplot is produced:

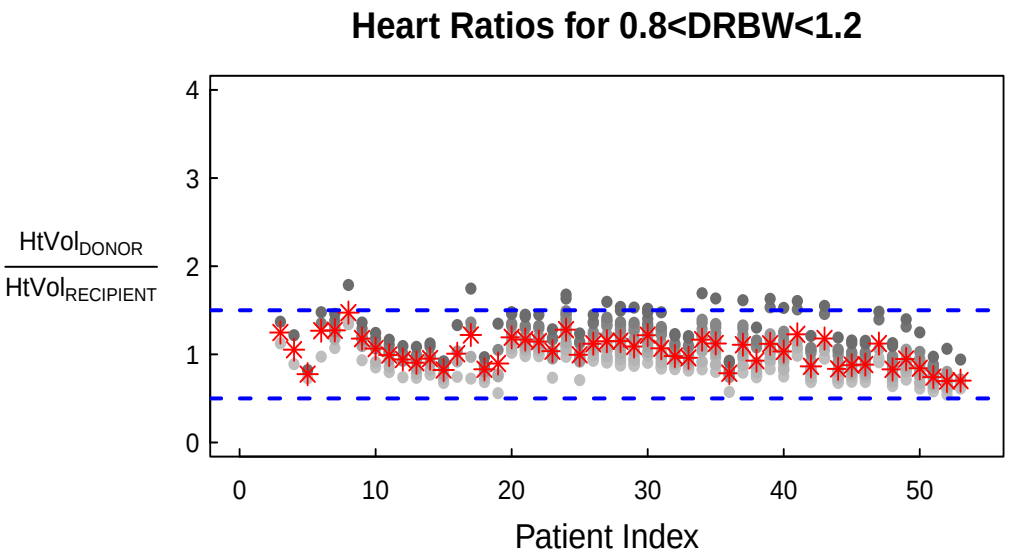


Figure 3: Pointplot for the more strict DRBW ratio range

We can clearly see that the number of oversizing cases has been greatly reduced. Very few values are over 1.5 indicating a much more reliable criterion, while at the same time again no undersizing has occurred, which was anticipated since the lower limit is almost the same as one of the wider range.

At a first glance, this more strict ratio may seem an ideal criterion to determine compatibility. However, in practice the availability of transplants is very limited and the waiting time until an allograft is accepted is vital for the patient's survival. While this strict ratio eliminates the oversizing problem, it also filters out a lot of the compatible hearts that could be utilized for the patients. The following scatter plot demonstrates the number of allografts corresponding to a donor-recipient heart ratio of 1 to 1.5 that are missed from the strict ratio, while they are correctly identified by the wider one:

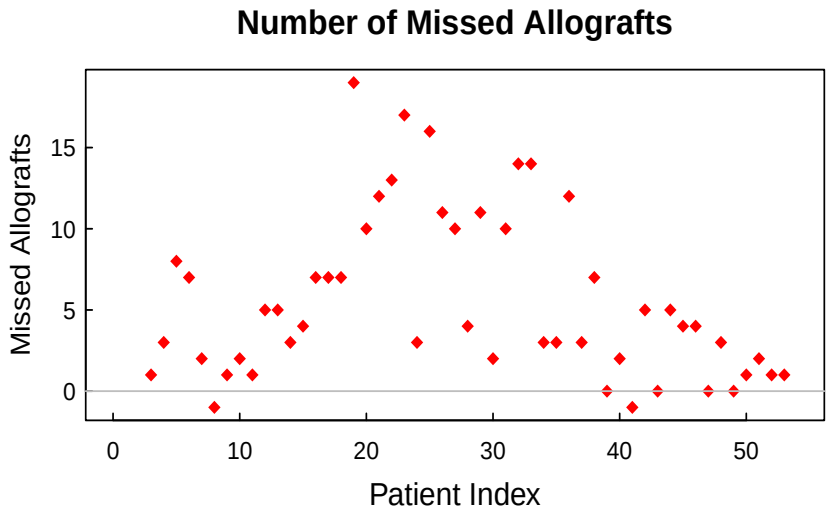


Figure 4: Number of missed allografts of the strict ratio compared to the wide ratio

It is apparent that for almost all patients the use of the strict ratio would lead to a considerable number of compatible hearts being turned down. This situation is more extreme for patients with medium sized hearts (with respect to our dataset), while the mean value for all missed allografts is approximately 5.5, a very extreme number considering how rare pediatric heart transplants actually are.

The evidence presented above demonstrates the potential limitations of DRBW ratio when it comes to assessing compatibility between a donor and a pediatric recipient. The relaxed ratio may suffer from the danger of selecting an extremely oversized allograft, while the strict ratio may unnecessarily limit the donor pool, therefore increasing the waiting time of the patients and endangering their survival. Therefore, other options to determine whether a transplant is of appropriate size for the candidate patient should be investigated, such as the development of a predictive model.

1.3 Methodological Tools and Application

We will now provide a summary of the contents of each chapter, highlighting the methodological tools utilized for our statistical analysis:

- Chapter 2 begins with a general inspection of our dataset, aiming to achieve a better understanding of how the predictors are connected with one another and the response. Next, the problem of multicollinearity is inspected through the Pearson and Spearman correlations as well as the Variance Inflation Factor. Subsequently,

extensions of the standard linear model are introduced, aiming to relax the overly restrictive assumptions linearity poses. Finally, the methodological steps towards the relaxation of the linear model assumptions are presented and implemented, simultaneously addressing collinearity issues between the predictors.

- Chapter 3 discusses Least Squares related methodology. Linear Regression is initially presented along with computational aspects for acquiring the coefficient estimates. Next, the importance of considering relevant errors is explained and accounted for through a transformation of the response. Popular subset selection methods are then analyzed, along with Information Theory criteria for the rejection/inclusion of features in the optimal set. In this context, potential issues of information criteria are highlighted and model averaging is proposed as a more robust technique when it comes to these issues. Chapter 3 concludes with a thorough exploration of penalized Least Squares, specifically Ridge and Lasso Regression, while a popular algorithm to fit the models is also discussed.

- Chapter 4 focuses on asymmetric loss functions for fitting linear models. Firstly, Quantile Regression is presented, which is the pioneering method when it comes to asymmetric loss functions. Specifically, a detailed introduction of the method is conducted along the the optimization problem Quantile Regression solves. Subsequently. the attention is shifted to penalized Quantile Regression, where the Lasso penalty is considered, which can be seen as a natural extension of the simple Quantile Regression model. Then, 3 different *R* packages for obtaining the coefficient estimates are presented separately, along with the algorithms for solving the optimization problem and how each may lead to different coefficient estimates.

- Chapter 5 analyzes the statistical techniques employed in this study to evaluate the performance of each model Least Squares and Quantile Regression model as well as to determine the optimal set of coefficients for predicting the heart volumes. Cross Validation is initially presented as a method for error estimation. Subsequently, a variant of Cross Validation is discussed, called Nested Cross Validation, the purpose of which is to obtain unbiased results when hyperparameters are involved in the model. It is then suggested that repeating Cross Validation multiple times and averaging the results reduces their variability, thus obtaining more reliable estimates. Next, Permutation Tests are proposed as an additional statistical tool to detect differences in error distributions resulting from repeating Cross Validation / Nested Cross Validation across different models. Finally, confidence intervals are presented, the purpose of which is to quantify the uncertainty surrounding each linear model's coefficient estimates. In order to obtain the interval bounds without making any important assumptions, bootstrapping is suggested.

- Chapter 6 displays the results from the application of all the previous methodological tools discussed. The chapter is organized in two parts. The first involves model inference for Least Squares and Quantile Regression methods and the second evaluates the performance of all candidate models. For this purpose, the

best Least Squares models are identified, which are then compared appropriately with Quantile Regression and penalized Quantile Regression.

- Chapter 7 is the final chapter of this study and contains the conclusions from the implementation of all the aforementioned methods for predicting the heart volumes of pediatric patients.

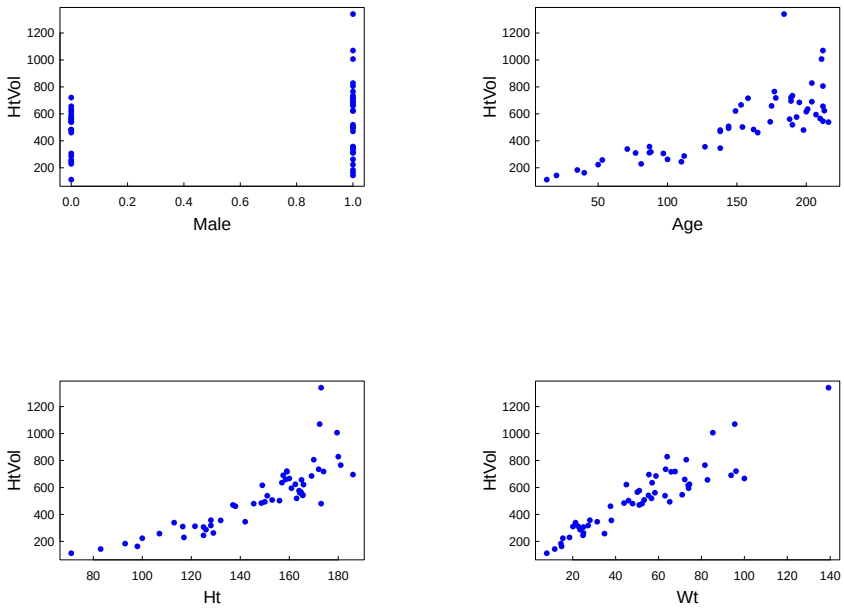
2. Preparing the Data

2.1 Data Exploration and Preprocessing

Before we start implementing any statistical methods, we must closely inspect the data to identify potentially important patterns that can be used later to improve the quality of our predictions. Data visualization can provide useful insights regarding the relationship of the predictors and the response, therefore it is a necessary step in any statistical research that must be performed thoroughly. Upon visualization, we will carefully examine the correlation of the independent variables, because somatometric measurements are anticipated to be highly correlated. Such a situation, if not handled properly, can result into poor performance and unreliable inferences, therefore a remedy must be considered. Furthermore, we will discuss some important assumptions made by the linear model along with an efficient approach to relax them. Finally, we will consider a transformation of the response, highlighting some key advantages it offers, both in general and more specifically in the context of predicting the heart volumes.

A General Inspection of the Data

Firstly, we explore the data visually through scatter plots of all predictive variables against the response:



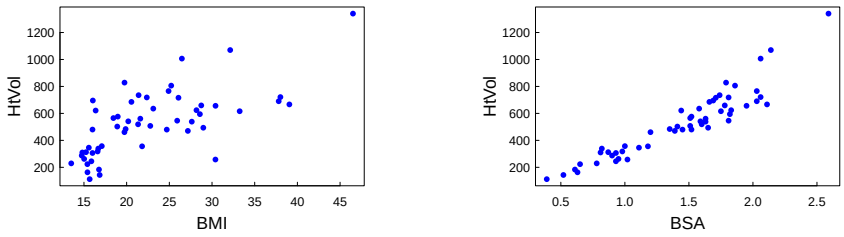
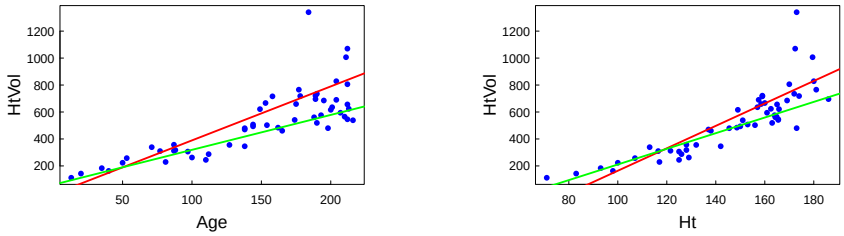


Figure 5: Scatter plots of all predictors against the heart volumes

These plots indicate an increasing monotonic relationship between the continuous predictors and the heart volumes, which of course makes sense from a biological point of view. Moreover, looking at the first figure we can notice that male patients have the tendency for hearts of bigger volume. Indeed, the mean heart volume for female patients is $471mL$, while for male patients is $559mL$. This information will be useful later on.

In this exploratory stage, we proceed to further inspect the differences between male and female patients by plotting two separate Linear Regression lines for each variable against the Heart Volumes, one for males and one for females:



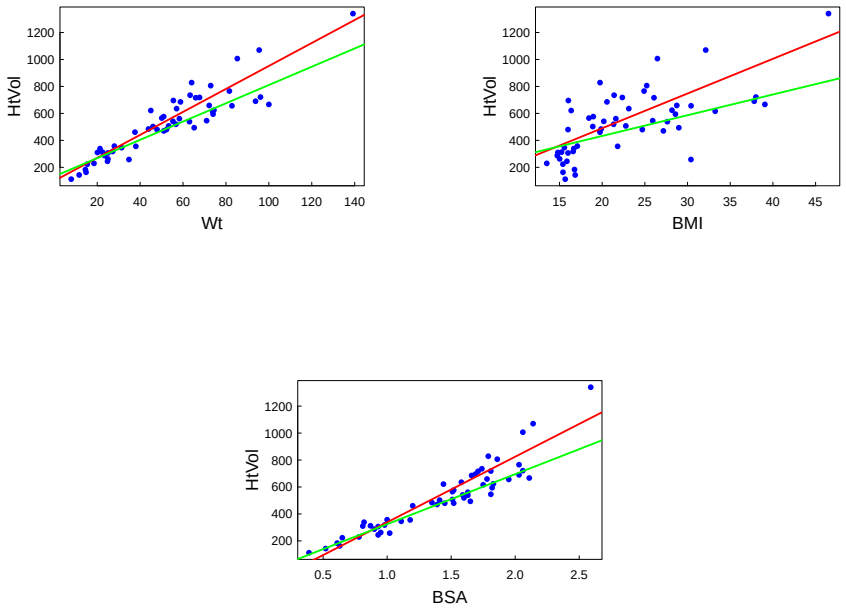


Figure 6: Linear Regression lines for each variable against HtVol. The red line represents male patients and the green represents females

A common observation derived from the figures is that in all cases, the red line eventually surpasses the green one as the value of each variable increases. This situation justifies further our expectation for males to have bigger hearts than females, given a certain measurement. However, if this measurement is relatively low, before the intersection of the two lines, then females are expected to have approximately the same heart volume as men. Nonetheless, this point of intersection is found in the lowest ranges of each variable, concerning few samples in total.

Addressing Multicollinearity

As mentioned above, we suspect that the predictors are highly correlated, therefore we calculate the Pearson's and Spearman's correlations. Pearson's correlation coefficient ρ is a standardized (hence free from measurement units) measure of covariance which aims to identify the presence of a linear relationship between the predictors. Given a pair of variables (X,Y), the formula for ρ is:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y},$$

where $Cov(X,Y)$ is the Covariance of X and Y and σ_X, σ_Y are the estimated standard deviations of X and Y respectively.

Spearman's correlation coefficient r of two variables (X,Y) is defined as the Pearson correlation coefficient between the rank variables. In order to get the rank variables, we first rank the data for each variable separately, from lowest to highest. Then, we assign ranks to each data point, starting from 1 for the lowest value, without changing their initial order. This gives us the corresponding rank variables $(R(X),R(Y))$, from which we compute r as following:

$$r = \frac{Cov(R(X),R(Y))}{\sigma_{R(X)} \cdot \sigma_{R(Y)}}$$

In contrast with Pearson, Spearman's correlation is used for identifying any monotonic relationship, not necessarily linear. The Spearman correlation is less sensitive than Pearson when it comes to strong outliers found in the tails of both samples. That is because Spearman's r limits the outlier to the value of its rank, regardless of its actual value. Both correlations take values from -1 to 1 and the closer we are by absolute value to 1, the stronger the relationship is, while the sign indicates the direction of the relationship.

Variable	Age	Ht	Wt	BMI	BSA
Age	1.00				
Ht	0.93	1.00			
Wt	0.79	0.78	1.00		
BMI	0.51	0.44	0.88	1.00	
BSA	0.89	0.90	0.97	0.78	1.00

Table 1: Pearson's Correlation

Variable	Age	Ht	Wt	BMI	BSA
Age	1.00				
Ht	0.81	1.00			
Wt	0.81	0.78	1.00		
BMI	0.59	0.46	0.86	1.00	
BSA	0.83	0.84	0.99	0.81	1.00

Table 2: Spearman's Correlation

Looking at both tables, we can see that all the variables are highly positively correlated with each other. The multiple large values in the correlation tables indicate the possibility that *multicollinearity* is present in the data. Multicollinearity is the situation where two or more variables are correlated and can result in less reliable statistical inferences [9], therefore it is an issue that needs to be addressed before the model building process begins.

The best approach for this situation is to utilize a statistical measure to quantify the extend of multicollinearity for each predictor. One such measure is the *Variance Inflation Factor (VIF)*. The VIF for the predictor X_j is computed using the following formula:

$$VIF(\beta_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. The idea behind the VIF is that if X_j is highly correlated with a set of features, then it can be somewhat accurately estimated through the corresponding regression model. As a result, the R^2 will take large values and the VIF will also be large. On the contrary, a small R^2 means that the fit is not good and X_j is not correlated with other predictors. The lowest possible value for the VIF is one, which corresponds to the vector X_j being orthogonal with all other feature vectors. In practice, a VIF value that exceeds 5-10 indicates a problematic amount of collinearity [10].

We now compute the VIF's of all predictors in our data and create a barplot. Note that the VIF's can only be used with continuous predictors, so its computation is performed using the all the predictors except for 'Male', which is categorical.

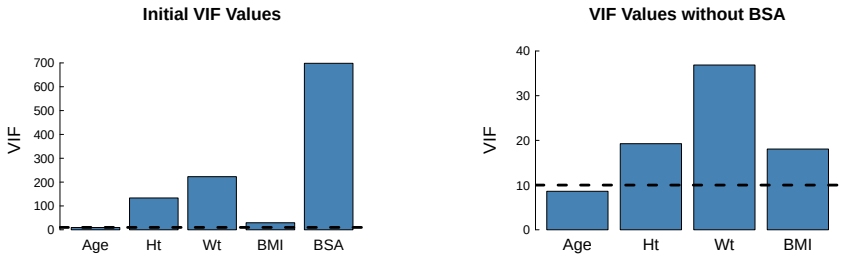


Figure 7: VIF's before and after dropping BSA

Looking at the left figure, it is clear that most values greatly exceed 10 (horizontal dashed line), with the greatest VIF value being by far that of 'BSA'. This means that the current features are not proper for use in regression models, as the effect of multicollinearity can harm the validity of the inferences made. Regarding BSA, apart from the collinearity issue, it has been found to be an unreliable somatometric measurement, especially when it comes to children. This is due to the large number of different formulas used for its estimation and the variation in these estimates resulting from each [11].

Given the extreme VIF and the uncertainty around BSA, it seems that its removal from the data is a justifiable remedy for the particular situation. Indeed, looking at the right figure we can observe that dropping BSA results in a significant drop of the VIF's. However, the problem of multicollinearity is not yet solved as the VIF values are still over 10. Instead of consider removing more predictors, we will opt for a different approach which involves the construction of a new dataset consisting of non-correlated features, based on the initial ones. Further details will be discussed in section 2.3.

2.2 Extensions of the Linear Model

One of the most significant weaknesses of linear models is that they make some very strong assumptions, which if not satisfied can lead to weaker predictions. Two of the most important assumptions state that the relationship between the predictors and the response is both *additive* and *linear* [10]. We present each of these below along with the most common and effective approaches for their relaxation that provide a more generalized model.

- The additivity assumption states that the influence of one predictor on the response is independent of any other influences from the remaining predictors. The typical way for its removal is by including a new predictor, called *interaction term*, which is constructed by computing the product of two existing predictors. These terms are a great addition to the linear model because they can identify patterns in the data that the simple model cannot, while they do not lose on interpretation.

- The linearity assumption suggests that the relationship between the response Y and each predictor X_j is linear, meaning that the change of Y associated with one-unit change of X_j is constant, regardless of the value X_j takes. The most simple and effective way to accommodate non-linear relationships is by including terms of higher order in the model. This can help to capture the full extent of the relationship and improve our performance.

Both of these approaches utilized to extend the linear models revolve around the idea of including new predictors based on the initial ones. If performed correctly, a more adaptive model can be created, which has the potential to provide more reliable inferences and predictions. However, the inclusion of interaction terms or terms of higher order is not a trivial process and should be performed with caution. An improper implementation can result to an overflow of predictors, which will most likely cancel the remedy we are trying to work on. Furthermore, the problem of multicollinearity can arise because we are using some of the columns more than once on different predictors. The goal of the following section, will be to effectively remove the additivity and linearity assumptions, while avoiding the aforementioned problematic situations.

2.3 Constructing a New Dataset

The objective of this section is to devise a strategy for integrating remedial measures that can eliminate the additivity and linearity assumptions from linear models, without encountering the aforementioned problematic scenarios. To achieve this, we plan to create several new predictors and subsequently narrow them down to those that offer the most relevant information, while also avoiding the issue of multicollinearity.

Computing new features

Firstly, we create a new set of features which will include all interactions and second order terms, where we will also add some terms that can have a physical meaning and interpretation when found in a linear model. Specifically, the new set will consist of the following features:

- The initial features: {Male, Age, Ht, Wt, BMI}
- All the $\binom{5}{2} = 10$ interactions between them
- All the second orders of the continuous features
- Each of {Ht, Wt, BMI} divided by Age

In total, these sum to 22 new predictors. Note that we have included the continuous predictors divided by 'Age'. These can have a rather interesting interpretation. For example, if we look at the term $\frac{Ht}{Age}$, it can provide information about the contribution of height in heart volume, per month of living. Of course, the exact way to interpret the predictors is not standard, depending on the presence of other variables in the model. We chose not to include the term $\frac{Male}{Age}$ because it is not physically logical to suggest that being male results in a change of the heart volume, inversely proportional to the patient's

age. Similarly, due to the lack of evidence supporting their physical interpretation, we do not have terms coming from dividing by the other continuous predictors.

The Importance of Predictor Standardization

In statistical research, our goal is not always limited to discovering the model that produces the most accurate predictions. Instead, we may also seek to address more complex questions, such as determining the degree to which specific variables impact our predictions. For example, we might ask whether variable X_i is more influential in predicting the response than variable X_j. These questions are of particular interest in the medical field and can provide some further insight in conjunction with prior knowledge physicians have about the biological aspects of the features.

In this application, it is very important for experts who make transplantation decisions to have a good understanding of the effect each predictor has in the heart volume. This is not so simple when dealing with non-standardized data because the units of measurement are not the same. Therefore, we opt for a standardized form of the predictors. The standardized variables are calculated by subtracting the mean and dividing by the standard deviation for each observation, i.e. calculating the Z-score. This results in mean 0 and standard deviation 1. Then, they don't represent their original scales since they have no unit. The interpretation of standardized regression coefficients is non-intuitive compared to their non-standardized versions:

Given a linear model $Y = \beta_0 + \sum_{i=1}^{m} X_i\beta_i$ *, a change of 1 standard deviation in* X_i *is associated with a change of* β_i *units of* Y*.*

Now, we can determine the effect each independent variable has on the response and compare the importance of each variable in the model.

When acquiring the new set of predictors, it is important to perform the standardization with caution. Specifically, we first calculate the new predictors and then we normalize them, not the other way around. However, the interaction terms of the form $Male \cdot X_i$ should be created using the standardized version of X_i multiplied with 'Male', so that we still have 0 in the samples corresponding to female patients. If we standardized the terms after the multiplication with the non-standardized X_i's, then we would eliminate the 0's from the columns, thus weakening the interpretation of these predictors. Furthermore, we should mention that the column 'Male' is not included in the standardization process as categorical variables do not have a scale to begin with.

Stepwise VIF-based Feature Selection

At this point, we have constructed a dataset which includes the initial features plus all the new ones, all normalized. Of course, due to the noise from having columns that are functions of other columns and the excessive multicollinearity present, this set of predictors is far from being suitable for making predictions. One way to address this issue is by performing feature selection so that we are left with features that are no longer highly correlated.

For this purpose, we will implement a feature selection algorithm called *Stepwise VIF Analysis*. This algorithm utilizes the VIF to identify the predictors with the highest collinearity issues. Specifically, we can compute the VIF's of all variables, drop the one with the highest value and repeat this process until reaching a threshold that we have set. For this application, we set the threshold to be 10, as this is in general the highest tolerable value for the VIF's. The algorithm is described below:

`Algorithm 1`: *Stepwise VIF Analysis*

---

1. let M be the full set of features and V the corresponding array with their VIF's
2. while there is an element in V greater than 10 do:
   (a) remove from M the feature corresponding to the highest VIF
   (b) compute the new array V based on the current VIF's
3. return the set of features M

We employ the Stepwise VIF Analysis algorithm on our complete set of features, excluding 'Male' as it is categorical. This results in the following features selected:

$$\{\text{Age}^2, \text{Ht}^2, \text{Wt}^2, \tfrac{\text{Ht}}{\text{Age}}, \tfrac{\text{Wt}}{\text{Age}}, \text{Male} \cdot \text{Age}, \text{Male} \cdot \text{BMI}\}$$

Now, as we mentioned in section 2.1, there is evidence for 'Male' providing further information about the heart volumes, since male patients are anticipated to have bigger hearts than females, given a certain measurement. Therefore, we include this predictor in the above set, thus getting the new dataset that we will be using throughout this statistical research:

$$\boxed{\{\text{Male}, \text{Age}^2, \text{Ht}^2, \text{Wt}^2, \tfrac{\text{Ht}}{\text{Age}}, \tfrac{\text{Wt}}{\text{Age}}, \text{Male} \cdot \text{Age}, \text{Male} \cdot \text{BMI}\}}$$

In this new dataset, we have removed the assumptions about the the relationship between the predictors being both additive and linear by constructing new features from the initial ones. Furthermore, by using VIF-based filtering, we have eliminated the presence of multicollinearity, meaning that the current predictors provide unique information that will be used for making inferences and predictions.

3. Least Squares Model Estimation

In this Chapter, we will be focusing on several techniques for constructing Least Squares models that predict the heart volumes. Firstly, we present Linear Regression, which is the pioneering Least Squares method for fitting linear models. Next, an alternative measure for Least Squares model estimation is proposed, which can be utilized through a logarithmic transformation of the response. Subset selection is then discussed for the purpose of eliminating uninformative predictors, which can lead to more accurate predictions and valuable interpretational properties. Specifically, best subset selection as well as stepwise selection algorithms are considered in conjunction with Information Theory criteria in order to identify the optimal set of predictors. Moreover, problems and limitations of Information Theory are highlighted, while appropriate solutions are proposed. In this context, the importance of model averaging is be explained, along with the different ways it can be performed. Finally, an alternative approach is considered, involving popular penalizing methods, specifically Ridge and Lasso Regression.

3.1 Linear Regression

Linear Regression is the traditional approach for modelling the relationship between a set of predictors and the response in a linear way. The aim of Linear Regression is to estimate the conditional mean of the response variable, given the available data, which is assumed to be an affine function of the predictors. This estimation can result in a model providing accurate predictions as well as insightful interpretations regarding the relationship between the predictors and the target variable.

A Preview

Suppose our data consists of n observations and p predictors. We can represent this dataset as a matrix X with dimensions $n \times p$, where the rows x_i contain the different observations and the columns correspond to the predictors. The affine function of Linear Regression includes a term that remains constant as the values of the independent variables change, which is the intercept. In order to account for this term, a column of 1's is added in X, which we can consider to be one out of p total predictors. If we describe the response measurements as a vector Y of length n, for $i = 1, ..., n$ the model takes the form

$$y_i = x_i^\top \beta + \varepsilon_i$$

where x_i is a vector sample corresponding to the measurement y_i and β is a vector of length p, describing the linear model's coefficients. The term ε_i is called error and the vector ε of length n containing all the errors is called error variable. This variable is used to compensate for any potential errors made due to the assumed linear relationship. The n equations we presented can be described in matrix notation as:

$$Y = X \cdot \beta + \varepsilon$$

In order to find the optimal coefficients β, we want to minimize the errors ε_i. In general, this can be achieved in various ways resulting in different models, with Linear Regression utilizing the Least Squares method. The Least Squares was introduced in around 1800, presented separately by Gauss and Legendre as a tool to calculate orbits of celestial bodies, while it was also used in geodesy. This method aims to minimize the sum of squared residuals in order to find the best parameters for the candidate model. In the case of Linear Regression, we want to minimize $\|\varepsilon\|_2^2 = \|Y - X^\top \beta\|_2^2$ with respect to β, or equivalently we want to solve:

$$\min_{\beta} \Big(\sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \Big) \tag{3}$$

The main weakness of Linear Regression is that the quality of the model's performance is highly dependent on the degree the assumptions made are satisfied. Two of these assumptions, linearity and additivity, have been already discussed and the construction of the new dataset effectively relaxed them. However, other assumptions are also present:

- In Linear Regression it is assumed that for $i = 1, ..., n$, $\varepsilon_i \sim N(0, \sigma^2)$, meaning that the errors follow a Normal distribution of mean 0 and constant variance. Assuming that the errors have mean 0 allows us to estimate the conditional mean of the response with an affine function such that $E[y_i] = x_i^\top \beta$. Considering that the errors have constant variance is somewhat unrealistic in practice, because usually as y_i increases, the errors tend to have increased variability [9]. This phenomenon where error variance is non-constant is called heterscedacity. One remedy for heteroscedacity is to transform the response using a concave function [10]. In the next section, we propose a logarithmic transformation, which not only addresses heteroscedacity, but also leads to other statistical advantages regarding Least Squares model estimation.

- Another assumption involves the predictors' correlation. Even though Linear Regression does not strictly assume non-correlated predictors, the presence of multicollinearity can harm the validity of inferences made. Specifically, in the case of perfect collinearity, that is, at least 2 predictors have a linear relationship, the data matrix X does not have a full column rank which results in the optimization problem (3) being subject to multiple solutions [9]. A similar situation can occur when predictors are highly correlated, without perfect correlation being necessary. In that case, the coefficients β tend to suffer from large sample variability, meaning that changes in the sample may greatly affect the estimation of β [9]. As a result, the coefficients can not be as trusted for inference since different solutions are as likely to occur, leading to the standard interpretation being non-applicable. In our application, the VIF Stepwise Selection algorithm implemented in the previous section mitigated the effect of multicollinearity, leading to a potentially more reliable model.

- The final assumption is that the errors ε_i are independent. When we analyze measurements from different patients, it is safe to consider that this assumption is met since there can not be any correlation between two individuals, meaning that the predictive errors are not dependent.

Computational Aspects

Throughout the years, numerous methods have been developed for fitting a Linear Regression model with the Least Squares. The standard way to estimate the coefficients β involves the use of the so called *Normal Equations*. Starting from (3), with the use of some basic matrix algebra and multivariate calculus we get the Normal Equations:

$$(X^\top X)\beta = X^\top Y \tag{4}$$

As we stated previously, in order to have a unique solution, β must have a full column rank. The Normal Equations shed some light on why this is necessary. If X has a full column rank, then the matrix $X^\top X$ is invertible, therefore we get the following solution for β:

$$\beta = (X^\top X)^{-1} X^\top Y \tag{5}$$

Here it is worth noticing that $(X^\top X)^{-1} X^\top$ is the Moore-Penrose inverse (or pseudo-inverse) of X.

In our study, all statistical methods are implemented with the use of programming language *R*, through the huge array of offered packages. For the task of fitting a Linear Regression model, we opted for the use of the function lm() from the stats package. This function is the most popular when it comes to simple Linear Regression, therefore it is compatible with many other functions from different packages. This makes it a great tool for a wide range of methods related to Linear Regression.

Instead of directly solving the Normal Equations by taking the inverse, lm() utilizes the *QR Decomposition* in order to find an estimate of β. QR Decomposition is considered in general a better alternative to the direct solution of the Normal Equations, as it is a more reliable and numerically accurate approach. This is because in order to get (5), we must take the inverse of $X^\top X$, which may lead to numeric inaccuracies if multicollinearity is present.

Suppose that $n \geq p$, meaning there are more samples that predictors. Then, we have the following QR Decomposition for X:

$$\underbrace{X}_{n\times p} = \underbrace{Q}_{n\times n} \underbrace{\begin{bmatrix} R \\ 0 \end{bmatrix}}_{n\times p} = \begin{bmatrix} \underbrace{Q_1}_{n\times p} & \underbrace{Q_2}_{n\times(n-p)} \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R \tag{6}$$

where Q is an orthogonal matrix and R is upper right triangular. Now, if we substitute (6) to the Normal Equations (4), we get

$$\left(\left(Q\begin{bmatrix}R\\0\end{bmatrix}\right)^\top Q\begin{bmatrix}R\\0\end{bmatrix}\right)\beta = \left(Q\begin{bmatrix}R\\0\end{bmatrix}\right)^\top Y \implies$$
$$\begin{bmatrix}R^\top & 0\end{bmatrix} Q^\top Q\begin{bmatrix}R\\0\end{bmatrix}\beta = (Q_1 R)^\top Y$$

Since Q is orthogonal $Q^\top Q = Q^{-1}Q = I$, we have:

$$R^\top R\beta = R^\top Q_1^\top Y$$

Assuming that X has full column rank, the diagonal of R consists of positive elements, therefore $det(R) > 0 \implies det(R^\top) > 0 \implies R^\top$ is reversible, thus:

$$R\beta = Q_1^\top Y \tag{7}$$

Here, $Q_1^\top Y$ is a vector of length n. Therefore, the equation (7) can be solved for β very easily with back substitution.

There are many other ways to solve the Least Squares problem, both with exact methods and iterative algorithms that approximate the solution through convergence. For instance, another popular function from the stats package for fitting Linear Regression models in *R* is glm(), which fits a more general model compared with the standard Linear Regression, called Generalized Linear Model (GLM). For this purpose, glm() deploys an iterative algorithm called Iteratively Reweighted Least Squares (IRLS). Estimating the parameters of GLMs is not possible with Normal Equations, therefore IRLS works as a great alternative, focusing on finding the Maximum Likelihood estimates of the desired model.

3.2 Assessing Relative Accuracy

The conventional method for fitting Least Squares regression models is to minimize the sum of squared residuals. However, this approach focuses on the absolute magnitude of the errors, making it unsuitable in situations where the relative error should be considered. In our application, working with relative errors is relevant because the margin for error when evaluating potential heart transplants is highly dependent on the patient's heart volume. As the true heart volume decreases, having a more accurate prediction becomes vital to ensure the patient's survival.

Given the importance of minimizing the relative errors when making predictions, we will discuss the use of an alternative measure for fitting regression models, referred to as *relative accuracy* by Tofallis (2015) [12]. Relative accuracy is based on the ratio of the predicted value y_{pred} to the actual value y, firstly discussed by Kitchenham et al. (2001) [13]. This ratio is directly connected with the relative error, as it is its complement:

$$\text{relative error} = \frac{y - y_{pred}}{y} = 1 - \frac{y_{pred}}{y} = 1 - \text{relative accuracy}$$

Therefore, they have the same distribution, meaning that relative accuracy also takes under consideration relative errors and is appropriate for use in our application. This measure is defined when y, y_{pred} are both positive, with 1 being the ideal value. Kitchenham et al. (2001) [13] observed that relative accuracy is not symmetric around its ideal value, with Tofallis (2015) [12] proposing the simple solution of taking the logarithm in order to overcome this problem. Then, $log\frac{y_{pred}}{y}$ can be used to fit regression models with the use of least squares.

Another problem relative accuracy suffers from involves the presence of y in the denominator. Specifically, as y takes smaller values, the weight of the fraction $\frac{y_{pred}}{y}$ inevitably increases, thus applying a greater penalty to the lower points. Therefore, when fitting Least Squares, the points in the lower part of the data will have more weight than the points above them, leading to the regression line being pulled towards these lower points [12]. This unpleasant situation will tend to lead to models which under-predict. However, taking the logarithm when minimizing relative accuracy with Least Squares solves this issue since then $log\left(\frac{y_{pred}}{y}\right) = log(y_{pred}) - log(y)$, thus the weight effect of y is eliminated.

When fitting the model, utilizing relative accuracy as described above simply requires predicting the logarithm of the response:

$$\sum_{i=1}^{n}(log\frac{y_{pred_i}}{y_i})^2 = \sum_{i=1}^{n}(log(y_{pred_i}) - log(y_i))^2 \tag{8}$$

Therefore, given the importance of taking into consideration the relative error in this context, we will use the logarithm of relative accuracy as a measure for fitting Least Squares models. This will be achieved by estimating the logarithm of the heart volumes, a transformation which will also serve as a mitigator for any issues caused by the potential presence of hereroscedacity.

3.3 Subset Selection

The purpose of *subset selection* methods is to identify the most important predictors that can lead to an accurate and interpretable statistical model. In this section, we consider some of these methods in the scope of Linear Regression, specifically best subset and stepwise feature selection. Furthermore, we discuss the selected criterion for determining the best-performing subsets.

Akaike's Information Criterion (AIC)

When constructing a predictive model, we theoretically attempt to identify the mechanism which generated the observed data. In most cases, this mechanism has infinite dimensions and is impossible to find, thus making our efforts rather ambitious [15]. Nonetheless, numerous techniques have been developed for acquiring models, which can lead to satisfactory results. These models perhaps are not very far off from the true mechanism, meaning that they can be trusted for real-life applications.

Of course, one would prefer to have an understanding of how far away we are from the 'perfect' model. Given that this model is unknown, it is not feasible to quantify this distance. However, given a candidate model, it is possible to quantify its relative

distance to the true mechanism. The closer we are, the better the model is, thus we can estimate how much better one model is compared to others, given the available data. One of the most popular ways to achieve this is with the use of *Akaike's Information Criterion* or *AIC* [14]. AIC can provide an estimate of the expected relative distance between a model and the unknown true mechanism:

$$AIC = -2log(L(\hat{\theta}|y)) + 2K$$

where $log(L(\hat{\theta}|y))$ is the maximum of the log-likelihood of a model with parameters θ and data with response y, while K is the number of parameters we estimate. Given the formula above, it is clear that smaller AIC values are considered better. The term $2K$ can be interpreted as a bias correction term, applying a greater penalty for more complicated models which involve a larger number of parameters to be estimated [15].

In Least Squares problems, the computation of AIC becomes very simple and inexpensive:

$$AIC = n \cdot log(\hat{\sigma}^2) + 2K$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}$ is the Maximum Likelihood Estimator of σ^2 and ε_i are the estimated residuals for the candidate model. Moreover, K is the number of the regression coefficients, including the intercept, plus 1 for the estimation of σ^2.

Even though AIC is very effective in many contexts, it may perform poorly when the proportion of samples to features is small [15]. For that reason, a 'corrected' alternative criterion, called *Akaike's Information Criterion corrected*, or *AICc* was derived through the study of Hurvich and Tsai (1989) [16]:

$$AICc = -2log(L(\hat{\theta}|y)) + 2K\Big(\frac{n}{n-K-1}\Big)$$

The idea of this correction is that we strengthen the contribution of the bias correction term, making it less likely to select a more complicated model when having fewer samples. Burnham and Anderson (2002) [15] suggested the use of AICc in situations where $\frac{n}{K} < 40$, a condition which falls under our application. Therefore the corrected form will be used.

Same as above, the least squares case for the AICc which we will be using for the Subset Selection is the following:

$$AICc = n \cdot log(\hat{\sigma}^2) + 2K\Big(\frac{n}{n-K-1}\Big) \tag{9}$$

Best Subset Selection

Best subset selection is a powerful method for selecting the best set of predictors to use in a regression model. It is very straightforward, involving the evaluation of all possible subsets of predictors, from the empty set that includes only the intercept, to the full model with all variables. The evaluation is performed based on the criterion we have chosen, in our case being the AICc. This means that the subset of features chosen leads into the smallest AICc value. The algorithm is provided below:

`Algorithm 2`: *Best subset selection*

1. initialize the set of selected models M to include only the null model m_0, which contains no predictors. Let k_{max} be the number of features in the dataset
2. for $k = 1$ to k_{max} do:
   (a) compute all possible subsets of k predictors and add them to the set S_k
   (b) fit all possible models using the subsets from S_k
   (c) for each model, compute the AICc and add the model m_k with the lowest AICc to M
3. identify the model m_{best} with the lowest AICc in M, compute the AICc
4. return m_{best}

Best subset selection can become computationally intensive because the number of possible models increases exponentially, requiring the fitting of 2^k different models when k predictors are involved. As the number of features grows, the sample space becomes enormous, increasing the risk of selecting a model that overfits the training data and performs poorly on future data [10]. Moreover, the selected predictors can vary significantly, with small changes in the data potentially leading to a completely different subset being chosen [10].

These issues can be mitigated using alternative feature selection methods, such as forward and backward stepwise selection:

Forward Selection

Forward selection starts with an empty model and iteratively adds predictors that result in the largest improvement in performance, until no further improvement is observed. Given that we evaluate the models based on AICc, largest improvement means greatest reduction in the AICc. More formally, the algorithm is given below:

`Algorithm 3`: *Forward Selection*

1. initialize the set of selected features S to be the empty set. Let T be the set that contains all features and AICc_0 to be the AICc of the null model
2. while $S \neq T$ do:
   (a) for each predictor p in T, fit all the models with predictors in the set $S \cup \{p\}$
   (b) for each model, compute the AICc and find the predictor $\hat{p}$ whose inclusion results in the smallest value, $\text{AICc}_{\hat{p}}$
   (c) if $\text{AICc}_{\hat{p}} > \text{AICc}_0$: exit the loop
   (d) else: assign $\text{AICc}_0 \leftarrow \text{AICc}_{\hat{p}}$, remove $\hat{p}$ from T and add it to S
3. return S

Forward selection can be particularly useful in situations involving high-dimensional data, where the number of samples corresponding to the available features is small [10]. Since in our application we only have 54 samples, forward selection is appropriate for avoiding issues related to our high-dimensional dataset, such as overfitting.

Backward Elimination

Backward elimination starts with a full model and iteratively removes predictors based on which removal results to the largest improvement in performance, according to the AICc value, until no further improvement is observed.

`Algorithm 4`: *Backward elimination*

1. Initialize the set of selected features S to be the full set that contains all predictors. Let AICc_0 be the AICc of the full model.
2. While S is not empty do:
   (a) For each predictor p in T, fit all the models with predictors in the set $S \backslash \{p\}$
   (b) For each model, compute the AICc and find the predictor $\hat{p}$ whose removal results in the smallest value, $\text{AICc}_{\hat{p}}$
   (c) If $\text{AICc}_{\hat{p}} > \text{AICc}_0$: exit the loop
   (d) Else: assign $\text{AICc}_0 \leftarrow \text{AICc}_{\hat{p}}$ and remove $\hat{p}$ from S
3. Return the model that is fit with the predictors in S

Forward selection and backward elimination build an optimal set of predictors iteratively by evaluating the contribution of adding or removing a variable at each step, which reduces the risk of overfitting. However, the sensitivity to the order of variable inclusion can become a limiting factor in selecting the best model [10].

3.4 Information Theory Limitations and Proposed Solutions

Apart from the algorithm-specific potential issues regarding subset selection, other considerations to be taken into account involve the Information Theory based model assessment criteria. Specifically, we will discuss about the limitations of information criteria in general, but also in the scope of stepwise selection procedures.

One great advantage of Information Theory is that it accounts for uncertainty in deciding whether a model is effective [17]. That is, when using AIC (or its derivatives), we consider models with small AIC differences to have similar explanatory value, without one overpowering the rest [18]. According to Burnham and Anderson (2002) [15], an AIC difference of 2 is insignificant and cannot provide evidence of a model's superiority over the rest.

Model Averaging

A core assumption of Information Theory states that all models considered are well-founded, both theoretically and empirically [15]. Therefore, a good practice of this approach requires careful a-priori selection of candidate models [18]. However, in several applications, there is not enough evidence regarding the variables' contribution towards inferences and predictions. This is also the case in pediatric heart volume prediction, where DRBW ratio is most frequently used in relevant transplantation procedures [5]. In such situations, it is difficult and possibly dangerous to drastically narrow down the candidate models [19].

The proposed remedy to this problem is to opt for an all subset Information Theory approach, meaning that we consider every possible predictor set [20]. Then, if multiple models are found to have similarly low AIC_c values, we can average these models according to their relative suitability [15]. Before explaining how *model averaging* is performed, we will define the AIC_c *differences*:

$$\Delta_i = AIC_{c_i} - AIC_{c_{min}},$$

where $AIC_{c_{min}}$ is the minimum AIC_c among all candidate models. Furthermore, some form of weights are required to use for averaging. For this purpose, we define the *Akaike weights*:

$$w_i = \frac{exp(-\frac{1}{2}\Delta_i)}{\sum_{j=1}^{R} exp(-\frac{1}{2}\Delta_j)}$$

The Akaike weights represent the relative strength of evidence for each of the R models [15].

In Linear Regression, model averaging is not performed using the entire formula, but for each of the coefficients separately. Given a coefficient β_j corresponding to the predictor x_j, there are two ways to perform model averaging, as presented by Burnham and Anderson (2002) [15]:

- The first way involves averaging β_j over all models in which x_j appears:

$$\hat{\bar{\beta}}_j = \frac{\sum_{i=1}^{R} w_i I_j(g_i) \hat{\beta}_{j,i}}{w_+(j)},$$

$$w_+(j) = \sum_{i=1}^{R} w_i I_j(g_i)$$

  and

$$I_j(x_i) = \begin{cases} 1 & \text{if predictor } x_j \text{ is in model } g_i \\ 0 & \text{otherwise} \end{cases}$$

- The second way requires considering that x_j is in all models, but in some the corresponding β_j is exactly zero:

$$\hat{\bar{\beta}}_j = w_+(j) \hat{\bar{\beta}}_j$$

  By considering only models where x_j is selected, we do not account for the rejection of this variable from the rest of the candidate models. This alternative way however does not ignore the absence of x_j from some models. Specifically, when x_j appears in fewer models, a shrinkage will occur since then $w_+(j)$ will be smaller than 1.

In this application, we will perform model averaging in models that have $\Delta_i < 2$, inspecting both ways of averaging. The goal is to improve the quality of our inferences by reducing model selection bias when estimating the Linear Regression coefficients.

Issues on IT-based Feature Selection

In this application, we chose to implement AICc based forward selection and backward elimination, being aware of their potential drawbacks. One possible downside of this approach is that the algorithm may fail to identify irrelevant predictors while it can also lead to the exclusion of important ones [21]. Moreover, the selection of an optimal variable set tends to be unstable, meaning that small changes in the data can have a significant effect on the final model [18]. We must be extra cautious about this situation because the number of available samples is limited.

We have decided to not proceed with best subset selection as it is generally criticized for being prone to overfitting. Additionally, regarding the assessment criterion AICc, Burnham and Anderson have repeatedly warned against a best subset approach, stating

that it may lead to false evidence [15]. Finally, the high variance if suffers from when it comes to the optimal subset being chosen makes this technique unreliable for inference, which is one of our main goals.

Given the uncertainty around certain aspects of Information Theory, it seems necessary to assess the quality of the models using not only AIC_c, but also other methods, suitable to the particular application[17, 18]. An additional reason for this necessity is that information criteria will always determine one model to be the 'best', even if none is performing well. In this context, we should note that AIC and AIC_c criteria have been suggested to be exposed to the danger of overfitting, thus failing to select the best model [22]. Chapter 5 will discuss in detail our approach regarding model assessment for the purpose of selecting the optimal model among all candidates.

3.5 Penalized Regression

Penalization methods can work as an alternative to the subset selection approach described in the previous sections. In the case of Linear Regression, the objective function is modified by adding a penalty term to the Least Squares loss function. This additional term involves the coefficients only, penalizing larger values and consequently forcing them to become smaller in absolute value, i.e. to shrink. This effect is the reason that penalization methods are frequently referred to as shrinkage methods. The great benefit of shrinkage is that it helps reduce the variance of the coefficients, creating a model with potentially improved predictive performance and higher interpretational value. Moreover, shrinkage methods enjoy the property of mostly influencing the coefficients of less important predictors, sending them towards zero. This results in only the most relevant predictors having non-zero coefficients, thus effectively selecting the best subset of predictors.

Two of the most known penalized regression methods are *Ridge Regression* and *Lasso Regression*.

Ridge Regression

Ridge Regression was proposed mainly as a remedial measure for multicollinearity problems [9]. However, it can be particularly useful in many applications, serving as a method for improving efficiency in estimating the regression coefficients. Instead of minimizing the sum of mean squared loss as in 3, we are interested in finding the coefficients that solve

$$\min_{\beta} \Big(\sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2 \Big) \tag{10}$$

where $\lambda \geq 0$ is a tuning parameter that we determine.

Same as with Least Squares, we want to find estimates for the regression parameters that explain the data well, with the only difference being a penalty for larger β_j's. The

contribution of the penalty is dependent on λ, meaning that larger values of λ apply a stronger constraint on the coefficients.

The estimation of λ is a key factor for achieving a balance in the so called bias-variance tradeoff. Ordinary Least Squares suffers from high variance, especially if multicollinearity is present. At the same time, including the penalty term creates a bias in the sense that we force the coefficients to shrink, with the bias becoming stronger as λ increases. As an exchange, the variance is reduced, resulting in a more reliable linear model. Therefore, striking the desired balance is an essential part of the model building process. One way to achieve this is by testing multiple values for λ and selecting the one that yields the smallest error. Further details on this idea will be discussed in Chapter 5.

Lasso Regression

Apart from Ridge Regression, another shrinkage method is the Lasso Regression. Same as Ridge, we want to minimize the sum of mean squared loss, where we also include a penalty in the term. However, in this case the penalty term is the sum of the absolute values of the coefficients, meaning that we are solving:

$$\min_{\beta} \Big(\sum_{i=1}^{n} (y_i - x_i^{\top} \beta)^2 + \lambda \cdot \sum_{j=1}^{p} |\beta_j| \Big) \tag{11}$$

where $\lambda \geq 0$ is again a tuning parameter that controls the bias-variance tradeoff.

The main difference between the two shrinkage methods is that even though Ridge will shrink the coefficients of irrelevant predictors towards zero, it will not set any of them to be exactly zero, unless $\lambda = \infty$. On the other hand, Lasso has the ability to function not only as a regressor, but also as a feature selector, setting some of the coefficients to be exactly zero. This makes the model easier to interpret, providing that λ has been carefully selected.

In general, both Ridge and Lasso are very useful algorithms, without the one consistently overpowering the other. It has been observed that Ridge performs better when a large number of parameters contributes in the model, while Lasso is more effective in situations where there are many features to consider, but few are required to explain the response [10]. This is because the former method does not exclude any of the parameters, but the latter can set the irrelevant ones to be zero.

Given the explicit advantages of each penalty, we will implement both methods, comparing their performance with the models obtained by stepwise selection and model averaging algorithms.

Computational Aspects

Ridge and Lasso Regression can be seen as special cases of a more general penalized regression method, called Elastic Net [23]. This method aims to find the regression coefficients β by solving:

$$\min_{\beta}\left(\frac{1}{2n}\sum_{i=1}^{n}(y_i-x_i^{\top}\beta)^2+\lambda\cdot\left(\frac{1-\alpha}{2}\cdot\sum_{j=1}^{p}\beta_j^2+\alpha\cdot\sum_{j=1}^{p}|\beta_j|\right)\right) \tag{12}$$

where $\lambda \geq 0$ and $\alpha \in [0,1]$ are tuning parameters.

We can easily verify that for $\alpha = 1$ we get the objective function of Lasso, while for $\alpha = 0$ we get the Ridge. Note that we have multiplied the Least Squares part with $\frac{1}{2n}$, which does not affect the solution and it is done mostly for numerical reasons. The most popular package in *R* for fitting Ridge and Lasso Regression models is glmnet, which we will be working with in the scope of this application. For fitting a Ridge or Lasso model, we use the function glmnet(), where we specify $\alpha = 0$ or $\alpha = 1$ respectively.

In this case of penalized regression, the objective function of the minimization problem (12) is not differentiable at all points, since for $\beta_j = 0$ the derivative of $|\beta_j|$ does not exist. As a result, it is not possible to use the partial derivatives, setting them to 0 and achieve an exact solution like we did in the Least Squares. Therefore, alternative numerical methods must be considered that converge to the solution.

The optimization algorithm glmnet() uses for fitting a penalized regression model is called Coordinate Descent. Specifically, the simplest case of Coordinate Descent is implemented, which is called Cyclic Coordinate Descent. Suppose that we want to solve

$$\min_{x} F(x) \tag{13}$$

where $F(x)$ is a multivariate function of $x = (x_1,...,x_p) \in \mathbb{R}^p$.

The idea of Cyclic Coordinate Descent (and Coordinate Descent in general) is that solving (13) can be achieved by solving multiple univariate minimization problems, considering one direction x_i at a time and minimizing $F(x)$ with respect to that coordinate [24]. Having started with an initial guess x^0, we choose $i_0 = 1$ and we minimize $F(v, x_2^0, x_3^0, ..., x_p^0)$ with respect to v. Then, using the value $\hat{v}$ that minimizes F, we update $x_1^1 = \hat{v}$, leaving the remaining coefficients unchanged. Thus, we have the new point $x^1 = (x_1^1, x_2^1, ..., x_p^1)$, where $x_2^1 = x_2^0, ..., x_p^1 = x_p^0$. Now, we choose $i_1 = 2$ so that we minimize $F(x_1^1, v, x_3^1, ..., x_p^1)$ with respect to v. We update our estimate to x^2, where only the coordinate x_2^1 has been changed and repeat this process to get a sequence $x^0, x^1, x^2, ...$ When we reach $i_{p-1} = p$, the next iteration starts over from the first coordinate, meaning that $i_p = 1$ and we cycle again through all directions of F. Ultimately, the sequence of x^j's will converge to a local minimum of F. The following algorithm, presented by captures the steps of Cyclic Coordinate Descent, as described above:

`Algorithm 5`: *Cyclic Coordinate Descent*

1. initialize $k = 0, i = 1$ and choose a $x^0 \in \mathbb{R}^p$
2. repeat:
   (a) Find $x_{i_k}^{k+1} = \arg\min_v F(x_1^k, x_2^k, ..., x_{i_k-1}^k, v, x_{i_k+1}^k, ...x_p^k)$
   (b) update $i_{k+1} \leftarrow [i_k mod p] + 1$; $k \leftarrow k+1$
   (c) if the stopping criterion is satisfied, exit the loop
3. return x_k

We will now present how Coordinate Descent adapts to the optimization problem (12):

Suppose that we are in the k'th coordinate step where we want to minimize with respect to β_j^k. It can be shown ([25],[26]) that the update of β_j^k (step (b) of the algorithm) will have the form:

$$\beta_j^{k+1} = \frac{S\left(\frac{1}{n}\sum_{i=1}^n x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\alpha\right)}{1+\lambda(1-\alpha)} \tag{14}$$

where

- $\tilde{y}_i^{(j)} = \sum_{l\neq j}^p x_{il}\tilde{\beta}_l$
- S is called soft thresholding operator and is defined as

$$S(z,\gamma) = \begin{cases} z-\gamma, & z>0 \text{ and } \gamma < |z| \\ z+\gamma, & z<0 \text{ and } \gamma < |z| \\ 0, & \text{otherwise} \end{cases}$$

Finally, the stopping criterion glmnet() uses is that after a full circle of updating each coordinate once, the maximum change in the objective function from all p updates is smaller than a threshold that has been set.

4. Moving to Asymmetry

So far, we have discussed conventional model building procedures which revolve around the symmetric Least Squares method. We argued that even though Linear Regression results in a simple model for which strong assumptions are made, with proper adjustments and extensions it can lead to great results regarding both the quality of predictions and the interpretability of the model itself. For that reason, we constructed numerous linear models, using an Information Theory based approach in the scope of stepwise selection algorithms, but also through model averaging. Furthermore, we discussed the implementation of penalized methods, more specifically Ridge and Lasso Regression, in order to acquire a linear model where the effect of irrelevant predictors is significantly (or completely) mitigated.

In our application, it is essential to take into consideration that under-predictions are highly undesirable, because an undersized heart transplant can endanger the patient's survival. The methods we have presented so far fail to account for this component due to the presence of symmetry in the loss function, therefore other options should be explored. The main goal of this chapter is to incorporate the importance of handling under-predictions during the construction of linear models. For this purpose, the use of asymmetric loss functions will be proposed along with penalized versions, similar to the ones already discussed.

In the first section, a pioneering statistical method when it comes to model fitting with asymmetric loss functions will be presented, called *Quantile Regression*. Several aspects of Quantile Regression will be discussed, including the optimization problem for acquiring the coefficient estimates of the linear model along with some useful properties that will play a key role in our application. Next, Penalized Quantile Regression will be explored where the l_1 Lasso penalty is considered. Finally, *R* packages that provide penalized Quantile Regression options will be presented and assessed regarding their approaches for solving the optimization problem.

4.1 Quantile Regression

The increasing popularity of Quantile Regression can be attributed to Roger Koenker, who connected and extended many preexisting ideas in this topic. His book, 'Quantile Regression', published in 2001 [27], has been the main guide for understanding and implementing this method, accompanied with the package quantreg in *R*, which was also developed by Koenker. The theory presented in this chapter regarding Quantile Regression will be based on Koenker's book, while the application in our data will be mostly performed with the use of quantreg.

Quantile Regression is a statistical technique that can be considered a natural extension of the Ordinary Least Squares regression method. While Least Squares aims to estimate the conditional mean of the response variable given certain predictors, Quantile Regression focuses on the conditional quantiles of the response. For instance, one can acquire estimates of the response's median, or the lower/upper quartiles. The estimation of multiple quantiles at once provides a comprehensive view of the relationship between

the predictors and the response, allowing us to make robust inferences.

Apart from the ability of Quantile Regression to provide a more complete understanding of the relationship between the predictors and the target variable, another key advantage concerns its robustness against outliers. Unlike Least Squares, Quantile Regression is not significantly influenced by extreme data points, meaning that the estimated quantiles remain relatively stable even in the presence of outliers. This property is particularly valuable in our application, where there are limited samples available and simply removing outliers is not a viable option.

A Preview

As we mentioned, the goal of Quantile Regression is to estimate the conditional quantiles of the response Y, given the available data X. For $\tau \in (0,1)$, we define the τth conditional quantile function as:

$$Q_Y(\tau|X) = \inf\{y : F_{Y|X}(y) \geq \tau\}, \tag{15}$$

where $F_{Y|X}(y) = Pr(Y \leq y|X)$ is the conditional distribution function of Y given X.

Quantile Regression makes the assumption that the τth conditional quantile is a linear function of the predictors, meaning that $Q_Y(\tau|X) = X^\top \beta_\tau$. In order to find the best estimate for β_τ, we first define the piecewise linear loss function that Quantile Regression utilizes for the fit:

$$\rho_\tau(z) = z \cdot (\tau - I(z < 0)) \tag{16}$$

with I being the indicator function that takes the value 1 when $z < 0$ and 0 otherwise. When z is positive, it is multiplied by τ and when negative, it is multiplied by $(\tau - 1)$, which is also negative, thus resulting in a positive number in both cases. When $\tau \neq 0.5$, ρ_τ applies a different penalty to positive and negative errors, multiplying them by τ and $(1-\tau)$ respectively. This property of quantiles will be discussed again in this section when choosing which τ fits our estimation target better.

For $\tau = 0.5$, there is a symmetry, since both positive and negative errors are multiplied by 0.5. Specifically we can easily verify that that $\rho_{0.5} = \frac{|z|}{2}$ This special case of Quantile Regression, which estimates the median, is also called Median Regression or Least Absolute Deviations method.

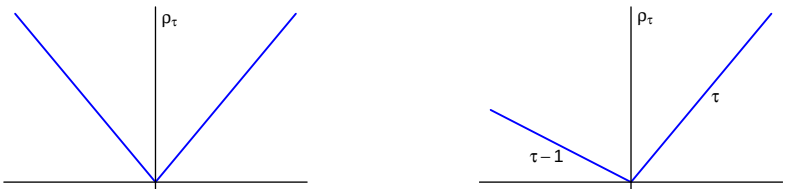


Figure 8: Left: ρ_τ for Least Absolute Deviations ($\tau = 0.5$). Right: ρ_τ for $\tau \neq 0.5$

Looking at the figures, we can observe that as τ moves away from 0.5, ρ_τ becomes a titled absolute value function, which on the left half-lane has slope (τ-1) and on the right half-lane has slope τ. Thus, one could consider Quantile Regression as a generalization of the Least Absolute Deviations.

The Optimization Problem

Suppose that our data consists of n observations and p predictors, including the intercept. We want to estimate β_τ by solving:

$$\min_{\beta} \Big(\sum_{i=1}^{n} \rho_\tau (y_i - x_i^\top \beta) \Big) \tag{17}$$

In contrast with the Least Squares, this objective function is not differentiable everywhere, since when $y_i - x_i \cdot \beta = 0$ the derivative does not exist. Thus, we cannot take the partial derivatives to find an analytical solution to this minimization problem, similar to Least Squares where normal equations are sufficient to solve for β. This is also the reason why Least Squares was throughout the centuries much more popular than Least Absolute Deviations, even though the latter was introduced almost 40 years earlier, in 1757.

In order to deal with non-differentiability, the minimization problem of Quantile Regression can be reformulated as a *Linear Programming* problem, which is considerably easier to approach.

Linear Programs are problems that can be expressed in the canonical form:

$$\begin{aligned} &\text{minimize } c^\top x \\ &\text{subject to } Ax = b \\ &\text{and } b \geq 0 \end{aligned}$$

where c is a vector of length n with constants, A is a $p \times n$ matrix and b is a vector of length p with non-negative constants.
First of all, we can rewrite

$$\sum_{i=1}^{n} \rho_\tau(\varepsilon_i) = \sum_{i=1}^{n} \big(\tau|\varepsilon_i|I(\varepsilon_i \geq 0) + (1-\tau)|\varepsilon_i|I(\varepsilon_i < 0)\big)$$

Now, we can use slack variables to decompose the errors into positive and negative parts:

$$\varepsilon_i = u_i - v_i,$$

where $u_i = \max(0, \varepsilon_i) = |\varepsilon_i|I(\varepsilon_i \geq 0)$ and $v_i = \max(0, -\varepsilon_i) = |\varepsilon_i|I(\varepsilon_i < 0)$. Then, the sum of residuals will be written as:

$$\sum_{i=1}^{n} \rho_\tau(\varepsilon_i) = \sum_{i=1}^{n} \tau u_i + (1-\tau)v_i = \tau 1_n^\top u + (1-\tau)1_n^\top v,$$

where $u = (u_1, ..., u_n)^\top$, $v = (v_1, ..., v_n)^\top$ and 1_n is a vector of length n with all coordinates being 1.

Furthermore, the residuals must follow the constraint

$$y_i - x_i^\top \beta = \varepsilon_i = u_i - v_i.$$

Combining all of the above results in the following minimization problem:

$$\min_{\beta \in \mathbb{R}^p, u \in \mathbb{R}_+^n, v \in \mathbb{R}_+^n} \big(\tau 1_n^\top u + (1-\tau)1_n^\top v | y_i = x_i\beta + u_i - v_i, i = 1, ..., n\big) \tag{18}$$

as stated by Koenker in his book. However, we can observe that there is no restriction for β to be positive as it is required in the canonical form of Linear Programs. This minor inconvenience can be dealt with a decomposition of β into positive and negative parts, as we did with the errors and then a simple reformulation of the optimization problem is sufficient to satisfy all the constraints of a Linear Program.

In our application, for the purpose of fitting a Quantile Regression model, we will utilize the function rq() from the quantreg package. This function provides several options to solve (17) which rely on the Linear Programming formulation (18), such as the simplex method or interior point methods. The default algorithm for qr() is simply an improved version of the simplex method, presented by Barrodale and Roberts in 1973 [28], slightly modified for Quantile Regression. This algorithm, referred to as the 'br' algorithm, was developed to handle l_1 optimization problems more efficiently and therefore it is directly applicable to solving (17).

Choosing Quantiles

Suppose that we have a datapoint x_i for which we want to predict the response y_i. The linear formulation of Quantile Regression allows us to estimate y_i from x_i through a

linear model, meaning that we estimate $\hat{y}_i = x_i^\top \beta_\tau$, where β_τ are the coefficients for the τ'th quantile. Let ε_i denote the error of this estimation. Then, we have that $\varepsilon_i = y_i - \hat{y}_i$. We can see that $\varepsilon > 0 \implies y_i > \hat{y}_i$, thus positive errors correspond to under-predictions. Similarly, negative errors correspond to over-predictions.

We know that ρ_τ penalizes positive errors by τ and negative errors by $1-\tau$. Therefore, when $\tau > 0.5$ we apply a greater penalty to under-predictions, while the smaller τ becomes, the higher number of under-predictions are allowed by the loss function. Intuitively this makes sense because when making predictions in higher quantiles, you expect that most of your predictions will be above the true value. In median regression, we anticipate that half of our predictions will be over the true value. Similarly, when $\tau = 0.75$ the model will overestimate the true value 75% of the times, thus under-predicting is considerably limited.

We have repeatedly stressed the importance of handling under-predictions when it comes to heart transplants in order to avoid undersized allografts for candidate patients. Quantile Regression allows us to incorporate this asymmetry of the error-type importance into our model by focusing on quantiles other than the median. In this application, we want to move into higher quantiles so that the risk of under-predicting will be mitigated. However, deciding which quantile to choose is not a trivial process. When we make predictions based on higher quantiles, we may control under-predictions but at the same time the resulting errors will be larger compared to other quantiles close to the median. Thus, extra caution is required when choosing which quantile to rely on.

In an attempt to achieve a balance between controlling under-predictions and retaining an acceptable prediction accuracy, we will be estimating a Quantile Regression model for $\tau = 0.75$, essentially penalizing under-predictions three times more than over-predictions. We selected this value as the 75th percentile is expected to lead to an under-prediction only 25% of the cases, while also it is not too far away from the median, therefore the increase in error will not be as substantial. We should note however that this choice is mostly empirical and does not rely on data-driven evidence or medical research. A more detailed picture of the medical background of heart transplantation and the tolerance of oversized heart transplants is required to initiate a research on the optimal choice of τ. The statistical concepts presented in this application can be the foundation of future endeavours into a justifiable and well-argued quantile selection.

4.2 Penalized Quantile Regression

In the previous chapter, we discussed two of the most popular penalization methods, Ridge and Lasso Regression. These are incorporated in Ordinary Least Squares by adding a penalty term which aims to shrink the irrelevant coefficients towards zero. The same concept can be applied in the Quantile Regression. Now, given that the objective function is the sum of absolute values (or tilted versions of it) the Lasso l_1 penalty $\sum_{i=1}^{p} |\beta_i|$ seems as a very natural extension. The coefficients β_τ for Lasso are now estimated by solving:

$$\min_{\beta} \Big(\sum_{i=1}^{n} \rho_\tau (y_i - x_i^\top \beta) + \lambda \cdot \sum_{i=1}^{p} |\beta_i| \Big), \tag{19}$$

where $\lambda \geq 0$ is a tuning parameter that we determine, same as with penalized Least Squares.

There are plenty of *R* packages available for the implementation of penalized Quantile Regression. The most frequently used is quantreg, but there are other popular choices, such as hqreg and rqPen. Those 3 packages all contain functions that can fit a penalized Quantile Regression model using the Lasso l_1 penalty. However, one can find that if tested on the same dataset, the returned models will not be identical. This can be mostly attributed to the different approaches for the optimization problem (19), which may result in slightly diverging coefficient estimates.

In our application, we focus not only to optimize the final model's performance, but also to provide valuable insights regarding the role each predictor has on estimating the heart volumes. Therefore, having a concrete understanding of how the coefficients are derived is vital to support our inferences. In this context, we will dive deeper into each of the aforementioned packages, exploring their approaches for solving (19). Moreover, in Chapter 6 we will compare them with each other across different quantiles, not only performance-wise, but also through the behaviour of their reported coefficients.

Package 'quantreg'

The standard approach for estimating β_τ in Quantile Regression involves reformulating the optimization problem as a Linear Programming problem. This reformulation allows the utilization of well-established methods specifically designed for solving linear programming problems, such as the 'br' algorithm rq() uses to solve (17). However, when we move to penalized estimation, the inclusion of the l_1 penalty complicates the solution.

When $\tau = 0.5$, this inconvenience can be avoided by using a technique called *data augmentation*. The idea of data augmentation is to extend the data matrix X and the response vector Y in a way that the penalized objective function is transformed into an equivalent Quantile Regression optimization problem. For that purpose, $\tilde{X}, \tilde{Y}$ are constructed such that

$$\min_{\beta} \Big(\sum_{i=1}^{n} \rho_{0.5}(y_i - x_i^\top \beta) + \frac{\lambda}{2} \cdot \sum_{i=1}^{p} |\beta_i| \Big) \iff \min_{\beta} \Big(\sum_{i=1}^{n+p} \rho_{0.5}(\tilde{y}_i - \tilde{x}_i^\top \beta) \Big) \tag{20}$$

meaning that the two optimization problems are equivalent. The reason for dividing the penalty term by 2 is that $\rho_{0.5}(z) = \frac{|z|}{2}$, so (20) can be written as

$$\min_{\beta} \Big(\frac{1}{2} \sum_{i=1}^{n} |y_i - x_i^\top \beta| + \frac{\lambda}{2} \cdot \sum_{i=1}^{p} |\beta_i| \Big) \iff \min_{\beta} \Big(\frac{1}{2} \sum_{i=1}^{n+p} |\tilde{y}_i - \tilde{x}_i^\top \beta| \Big)$$

We can remove the constant term $\frac{1}{2}$ from both sides since it does not affect the points of local minimums of the objective function, thus getting

$$\min_{\beta} \Big(\sum_{i=1}^{n} |y_i - x_i^\top \beta| + \lambda \cdot \sum_{i=1}^{p} |\beta_i| \Big) \iff \min_{\beta} \Big(\sum_{i=1}^{n+p} |\tilde{y}_i - \tilde{x}_i^\top \beta| \Big) \tag{21}$$

In order to achieve this formulation, $\tilde{X}, \tilde{Y}$ are defined as follows:

$$\tilde{X} = \underbrace{\begin{bmatrix} x_{11} & \dots & x_{1p} \\ & \vdots & \\ x_{n1} & \dots & x_{np} \\ \lambda & & \\ & \ddots & \\ & & \lambda \end{bmatrix}}_{(n+p)\times p} = \begin{bmatrix} & x_1 & \\ & \vdots & \\ & x_n & \\ \lambda & & \\ & \ddots & \\ & & \lambda \end{bmatrix}, \tilde{Y} = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{(n+p)\times 1} \tag{22}$$

Essentially, X is extended by rows with the diagonal $p \times p$ matrix $\lambda \cdot I_p$, where I_p is the identity matrix and Y is extended by p zeros. With $\tilde{X}, \tilde{Y}$ defined as in (22), the equivalence (21) can be easily shown:

$$\tilde{X} \cdot \beta = \begin{bmatrix} & x_1 & \\ & \vdots & \\ & x_n & \\ \lambda & & \\ & \ddots & \\ & & \lambda \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \underbrace{\begin{bmatrix} x_1^\top \beta \\ \vdots \\ x_n^\top \beta \\ \lambda \cdot \beta_1 \\ \vdots \\ \lambda \cdot \beta_p \end{bmatrix}}_{(n+p)\times 1}$$

Thus,

$$\tilde{Y} - \tilde{X} \cdot \beta = \begin{bmatrix} y_1 - x_1^\top \beta \\ \vdots \\ y_n - x_n^\top \beta \\ \lambda \cdot \beta_1 \\ \vdots \\ \lambda \cdot \beta_p \end{bmatrix}$$

Finally, we have

$$\begin{aligned} &\sum_{i=1}^{n} |y_i - x_i^\top \beta| + \lambda \cdot \sum_{i=1}^{p} |\beta_i| = \\ &\sum_{i=1}^{n} |\tilde{y}_i - \tilde{x}_i^\top \beta| + \sum_{i=1}^{p} |0 - \lambda \cdot \beta_i| = \\ &\sum_{i=1}^{n} |\tilde{y}_i - \tilde{x}_i^\top \beta| + \sum_{i=n+1}^{n+p} |\tilde{y}_i - \tilde{x}_i^\top \beta| = \\ &\sum_{i=1}^{n+p} |\tilde{y}_i - \tilde{x}_i^\top \beta| \end{aligned}$$

which completes the proof.

However, for $\tau \neq 0.5$ the proof breaks down and (20) no longer applies. In the general case, we will have that:

$$\begin{aligned}
&\sum_{i=1}^{n+p} \rho_\tau(\tilde{y}_i - \tilde{x}_i^\top \beta) = \\
&\sum_{i=1}^{n} \rho_\tau(\tilde{y}_i - \tilde{x}_i^\top \beta) + \sum_{i=n+1}^{n+p} \rho_\tau(\tilde{y}_i - \tilde{x}_i^\top \beta) = \\
&\sum_{i=1}^{n} \rho_\tau(\tilde{y}_i - \tilde{x}_i^\top \beta) + \sum_{i=n+1}^{n+p} \rho_\tau(0 - \lambda \cdot \beta_i) = \\
&\sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta) + \lambda \cdot \sum_{i=1}^{p} \rho_\tau(-\beta_i)
\end{aligned}$$

This means that if one were to apply the data augmentation scheme for $\tau \neq 0.5$, they would essentially perform penalized Quantile Regression with the penalty not being the l_1 norm, but the loss function ρ_τ we use on the data. As of now, applying a greater penalty on positive ($\tau < 0.5$) or negative ($\tau > 0.5$) coefficients does not have the theoretical support to be trusted as a statistical method. Nonetheless, quantreg fits a penalized Quantile Regression model exactly as described above, using the function rq(method='lasso'). Specifically, the optimization problem (19) takes the form:

$$\min_{\beta} \Big(\sum_{i=1}^{n+p} \rho_\tau(\tilde{y}_i - \tilde{x}_i^\top \beta) \Big) \tag{23}$$

which is the same as the standard Quantile Regression optimization problem, but with the augmented data.

For the solution of (23), an interior point method is employed, which Koenker refers to as 'Frisch-Newton' algorithm, presented by Portnoy and Koenker in 1997 [29]. It should be noted that this algorithm solves Quantile Regression optimization problems, therefore it can be used also in simple Quantile Regression. However, the 'br' algorithm is suggested in that case for medium sized datasets (up to several thousand observations) as a more efficient alternative.

We should point out that the methods of quantreg for the l_1 penalty Quantile Regression have not been established and can lead to inaccurate and potentially dangerous inferences if one chooses to trust the interpretation of the estimated coefficients for $\tau \neq 0.5$. The documentation of quantreg also states clearly that 'these methods should probably be regarded as experimental', referring to the Lasso penalty. This uncertainty motivates us to further examine other options for penalized Quantile Regression, using more reliable and well-founded statistical techniques.

Package 'rqPen'

The aforementioned statistical inaccuracy caused by data augmentation has been identified and a very simple way to be corrected was proposed by Sherwood and Wang in 2016 [30]. Their idea was based on the observation that $\rho_\tau(\beta_i)+\rho_\tau(-\beta_i)=|\beta_i|$ which can be easily shown:

- If $\beta_i > 0,\ \rho_\tau(\beta_i)+\rho_\tau(-\beta_i)=\beta_i\cdot\tau+(-\beta_i)(\tau-1)=\beta_i=|\beta_i|$
- If $\beta_i < 0,\ \rho_\tau(\beta_i)+\rho_\tau(-\beta_i)=\beta_i(\tau-1)+(-\beta_i)\cdot\tau=-\beta_i=|\beta_i|$

Given that the l_1 penalty $|\beta_i|$ can be described with the loss function ρ_τ, they provide a new data augmentation approach. For that, the augmented data $\hat{X},\hat{Y}$ are defined as follows:

$$\hat{X}=\underbrace{\begin{bmatrix} & x_1 & \\ & \vdots & \\ & x_n & \\ \lambda & & \\ & \ddots & \\ & & \lambda \\ -\lambda & & \\ & \ddots & \\ & & -\lambda \end{bmatrix}}_{(n+2p)\times p},\hat{Y}=\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{(n+2p)\times 1} \tag{24}$$

We can observe that $\hat{X}$ is simply the augmented $\tilde{X}$ from before, further extended it by rows with $-\lambda I_p$, adding another p rows. In similar fashion, Y is extended by $2p$ zeros to have the same number of rows. Now, following the same analysis as before, we get:

$$\hat{Y}-\hat{X}\cdot\beta=\begin{bmatrix} y_1-x_1^\top\beta \\ \vdots \\ y_n-x_n^\top\beta \\ \lambda\cdot\beta_1 \\ \vdots \\ \lambda\cdot\beta_p \\ -\lambda\cdot\beta_1 \\ \vdots \\ -\lambda\cdot\beta_p \end{bmatrix}$$

Thus,

$$\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta) + \lambda\cdot\sum_{i=1}^{p}|\beta_i| =$$
$$\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta) + \lambda\cdot\sum_{i=1}^{p}(\rho_\tau(\beta_i) + \rho_\tau(-\beta_i)) =$$
$$\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta) + \sum_{i=1}^{p}\rho_\tau(0-(-\lambda)\cdot\beta_i) + \sum_{i=1}^{p}\rho_\tau(0-\lambda\cdot\beta_i) =$$
$$\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta) + \sum_{i=1}^{p}\rho_\tau(0-\lambda\cdot\beta_i) + \sum_{i=1}^{p}\rho_\tau(0-(-\lambda)\cdot\beta_i) =$$
$$\sum_{i=1}^{n}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta) + \sum_{i=n+1}^{n+p}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta) + \sum_{i=n+p+1}^{n+2p}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta) =$$
$$\sum_{i=1}^{n+2p}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta)$$

Therefore, given the augmented data $\hat{X}, \hat{Y}$, the following equivalence is achieved:

$$\min_\beta\Big(\sum_{i=1}^{n}\rho_\tau(y_i - x_i^\top\beta) + \lambda\cdot\sum_{i=1}^{p}|\beta_i|\Big) \iff \min_\beta\Big(\sum_{i=1}^{n+2p}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta)\Big) \tag{25}$$

The correct implementation of data augmentation allows us to reformulate the optimization problem of l_1 penalized Quantile Regression into that of simple Quantile Regression, which is considerably easier to work with. The package rqPen, using the function rq.pen() does exactly that and then the solution of

$$\min_\beta\Big(\sum_{i=1}^{n+2p}\rho_\tau(\hat{y}_i - \hat{x}_i^\top\beta)\Big)$$

is achieved with the 'br' algorithm to get the final estimates for β_τ. We should note that the package provides a solution through the 'Frisch-Newton' algorithm, but the authors have suggested the use of 'br' as a more stable option when it comes to penalized Quantile Regression.

Package 'hqreg'

The package hqreg follows a different route towards the coefficient estimates β_τ compared to the other two packages we discussed. In this case, the target is to minimize the objective function without any reformulations, same as with the Ridge and Lasso penalty in Least Squares. With Least Squares, Cyclic Coordinate Descent was used, updating the coefficient estimates at each step properly. A similar idea is implemented with penalized Quantile Regression, but now the fact that the function to be minimized is not differentiable when $y_i - x_i^\top\beta = 0$ or $\beta_i = 0$ has to be accounted for.

The algorithm used by the function hqreg() of hqreg package is called called Semismooth Newton Coordinate Descent (SNCD) and was proposed by Yi and Huang in 2016

[31]. SNCD combines the strongest aspects of the Semismooth Newton and Cyclic Coordinate Descent algorithms to achieve convergence. However, SNCD requires differentiability at all points, which the objective function does not have. In order to overcome this issue, the objective function's first part (quantile loss) is approximated by a first-order differentiable function so that non-differentiability concerns the penalty term only. Given that this term is convex, the subgradient can be utilized to perform Coordinate Descent updates on the coefficients β_j. Finally, we should mention that the penalty term considered by Yi and Huang is the penalty of elastic net, so the l_1 penalty requires substituting $\alpha = 1$, as we did in Lasso Regression.

5. Model Selection and Performance Estimation

The previous two chapters focused on statistical techniques for constructing linear models that predict the heart volumes of pediatric patients. We discussed the conventional approach where Ordinary Least Squares were considered in conjunction with stepwise feature selection and model averaging. Furthermore, Ridge and Lasso Regression were thoroughly presented as a way to undermine the influence of irrelevant predictors in order to achieve better estimates for the coefficients and more useful interpretational properties. Chapter 4 presented Quantile Regession, where asymmetric loss functions are utilized for fitting linear models, aiming to provide more robust inferences and at the same time account for the importance of constraining under-predictions. In this context, the Lasso penalty was introduced as a natural extension of Quantile Regression and several packages to implement penalized Quantile Regression were presented along with their underlying fitting algorithms.

All the aforementioned methods result in different sets of coefficients for the same predictors, from which our goal is to choose the optimal one. Ideally, we would test each model on new unseen data, observe how they perform and then conclude which is superior. However, since all our available data is the given dataset, we must come up with another approach for estimating the performance of each model and reporting the best according to our application goals.

5.1 Cross Validation

Extensive research has been conducted to develop methods for evaluating a model, aiming to provide an accurate estimate of how a given model will perform when coming across new data. One rather popular way for performance estimation is the *train-test split* method. As the name suggests, it involves splitting the data into a training and testing part, fitting the model to the train set and then measure its error on the test set. Common train-test splits are in the range of 70% (train) and 30% (test), depending on the available data and the queries to be addressed. This approach however wastes significant amount of samples (all the test set), while there are no guarantees for an accurate estimation, meaning that a lot of variance is present.

In this section, we will present *Cross Validation*, a technique that refines the train-test approach, effectively resolving both these issues.

Simple and Repeated Cross Validation

Cross Validation is a resampling method that uses different parts of the data for fitting a model and then testing it. It involves dividing the available data into k equally sized folds, training the model on $k-1$ folds and testing it on the remaining fold. This process is repeated k times, with each fold used once for testing and the remaining folds used for training. The results of each fold are then averaged to give an overall estimate of the model's performance.

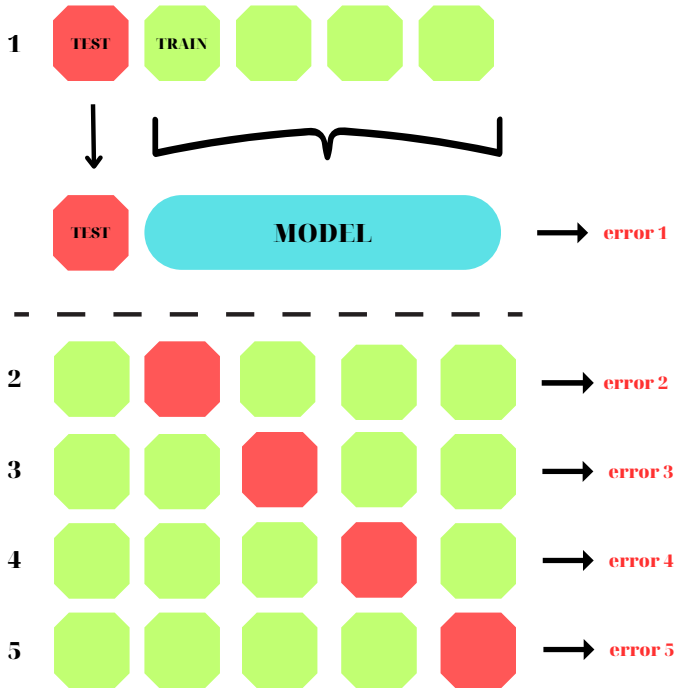


Figure 9: Cross Validation for $k = 5$

We now provide the algorithm for k-fold Cross Validation:

`Algorithm 6`: *k-fold Cross Validation*

---

1. split the dataset in k equal folds. Initialize $error = 0$
2. for $i = 1$ to k do:
   (a) leave out the i-fold and train the model on the remaining data
   (b) calculate the error of the model on the i-fold and add it to $error$
3. return $\frac{error}{k}$

Implementing Cross Validation requires the use of all our available data both for training and testing, which solves the first issue. Simultaneously, multiple models are trained,

resulting in different test errors, the averaging of which provides a more accurate estimation of the actual performance error. It has been shown that choosing k to be 5 or 10 avoids facing high bias, caused by fitting the model into only a subset of the full data, while the variance is effectively reduced [10]. As a general rule of thumb, smaller datasets require higher k, therefore opting for a value close to 10 seems more appropriate. In this application, we choose $k = 9$ since we have 54 samples, so the folds will be of equal size.

We have stated that averaging multiple test errors is an efficient way for variance reduction and to acquire a better estimate for the true performance. This idea can be exploited to further reduce the variance by repeating the Cross Validation process, each time randomly distributing the data across the folds and then averaging all the errors to improve our estimate. The method we described is called *Repeated Cross Validation* and it serves as a technique to greatly reduce the uncertainty around Cross Validation results.

Repeated Cross Validation is considered a computationally intensive process, especially when fitting complicated models with lots of data. Our application however involves relatively simple fitting algorithms and a very small dataset, therefore performing multiple repetitions is feasible. We choose to repeat the Cross Validation process $t = 40$ times, a number large enough to drastically reduce the effect of variance.

Cross Validation for Stepwise Selection and Model Averaging

We will now dive deeper into the correct approach for training the model inside the Cross Validation loop. When it comes to standard Linear Regression and Quantile Regression, one can simply fit the model using the training set and then test it, repeating this process leaving one fold out each time. Then, the coefficient estimates are derived solely by the training set and the test set can be considered as unseen data. As a result, Repeated Cross Validation will provide an unbiased estimate of the actual performance. The key part here is that the model is tested on unseen data, which is necessary to keep the estimates unbiased. The same principle must be followed in more complicated statistical methods, such as stepwise feature selection and model averaging. In these cases however, Cross Validation must be performed with caution as to not violate the aforementioned principle.

A naive execution of Cross Validation for stepwise feature selection would involve finding the optimal subset of predictors based on the entire dataset and then estimate the performance error by fitting Linear Regression models with that subset only inside the loop. With this implementation, our results would be severely biased due to the entire dataset being used to filter out some predictors and then used again to test how the filtering performed. Overfitting would then be present leading to an underestimation of the true error.

Avoiding this problematic situation requires performing stepwise selection inside each loop of Cross Validation. For that, we identify the optimal subset of predictors based on the training data, then fit a Linear Regression model on this subset and test it, repeating the same steps for each iteration. This process complies with the unseen

data principle, therefore will result in an unbiased estimate of how the entire statistical process of stepwise selection would perform on new data.

Algorithm 7 captures the steps for Cross Validating stepwise feature selection for Linear Regression:

`Algorithm 7`: *k-fold Stepwise Selection Cross Validation*

---

1. split the dataset in k equal folds. Initialize $error = 0$
2. for $i = 1$ to k do:
   (a) leave out the i-fold and perform stepwise selection on remaining data, thus acquiring a predictor subset S
   (b) fit a Linear Regression model with the predictors in S
   (c) calculate the error of the model on the i-fold and add it to $error$
3. return $\frac{error}{k}$

A similar approach is necessary for correctly estimating the performance of $AICc$-based model averaging. In each iteration, the models with Akaike difference $\Delta_i < 2$ are identified through the train data, then the averaged model is constructed and tested on the left-out fold. The algorithm's structure for this process is very similar to Algorithm 7:

`Algorithm 8`: *k-fold Model Averaging Cross Validation*

---

1. split the dataset in k equal folds. Initialize $error = 0$
2. for $i = 1$ to k do:
   (a) leave out the i-fold and with the remaining data, find all models with $\Delta_i < 2$
   (b) average the models
   (c) calculate the error of the averaged model on the i-fold and add it to $error$
3. return $\frac{error}{k}$

Algorithms 7,8 are sufficient to provide an unbiased estimate for the true error of stepwise selection and model averaging respectively. Thus, we are now able to report the results of Repeated Cross Validation for these models, along with Linear Regression and Quantile Regression without any concerns about potential bias or high variance affecting their validity.

When it comes to penalized methods however, Repeated Cross Validation breaks down. This is because for penalized methods, a choice of the hyperparmeter λ is required in order to define the loss function to be minimized. With Cross Validation, finding the optimal value for λ is not possible, therefore we will look into a variation of Cross Validation which serves the purpose of both hyperparameter tuning and performance estimation.

5.2 Nested Cross Validation

Nested Cross Validation is a variant of Cross Validation that can provide an unbiased estimation of the true error for models (or perhaps better, statistical procedures) which require hyperparameter tuning. Such an algorithm is essential for reporting the results of Ridge and Lasso Regression but also for Lasso-based penalized Quantile Regression.

As we mentioned in Chapter 3, one idea to strike the perfect balance between bias stemming from the penalty error and variance of the coefficients is by testing multiple values for λ. Specifically, we can consider a search grid and test each λ with Cross Validation to identify which results in the smallest error. Keep in mind that achieving unbiased results requires testing the candidate model on unseen data. If we were to choose λ as described above and then perform Cross Validation, then our error estimation would be biased towards an over-estimation, since the choice of λ relied on the entire dataset. Therefore we must find an optimal λ inside the Cross Validation loop for each training set.

This idea is the key for Nested Cross Validation. For each Cross Validation loop, we perform an inner loop of Cross Validation on the train set to find the optimal configuration of hyperparameters for the model. This optimality is determined by the lowest error among each possible configuration. Then, given the set of chosen hyperparameters, we fit the model in the train set and test it on the left-out fold. By doing so, the training process will not involve the test set, hence an unbiased error estimation is acquired.

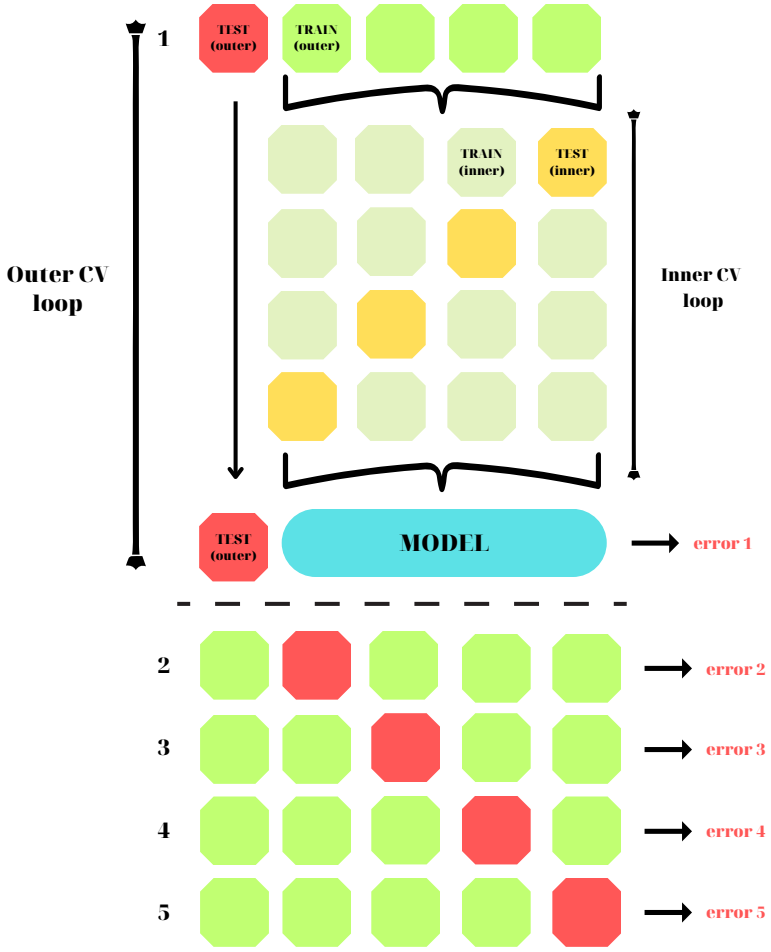


Figure 10: Nested Cross Validation

Algorithm 9 captures the steps for obtaining a Nested Cross Validation error estimation:

`Algorithm 9`: *Nested Cross Validation*

---

1. split the dataset in k equal folds. Initialize $outer\_error = 0$
2. for $i = 1$ to k do:
   (a) leave out the i-fold and consider the remaining folds as train data
   (b) split the train data in $\hat{k}$ equal folds. Initialize an empty list *inner_error_list*
   (c) for each configuration c do:
      i. initialize $inner\_error = 0$
      ii. for $j = 1$ to $\hat{k}$ do:
         A. leave out the j-fold and train the model on the remaining data with the hyperparameters of c
         B. calculate the error of the model on the j-fold and add it to *inner_error*
      iii. append $\frac{inner\_error}{\hat{k}}$ to *inner_error_list*
   (d) find the configuration c_{min} corresponding to minimum element of *inner_error_list* and fit the model on the train data with the hyperparameters of c_{min}
   (e) calculate the error of the model on the i-fold and add it to *outer_error*
3. return $\frac{outer\_error}{k}$

Same as with Cross Validation, we can repeat Nested Cross Validation multiple times and then average the resulting errors to reduce the variance. Even though Nested Cross Validation is significantly more computationally expensive, the small size of our dataset allows us to consider *Repeated Nested Cross Validation*, again for $t = 40$ repetitions.

5.3 Practical Aspects of Cross Validation

Error Quantification

So far, we have presented algorithms for obtaining an unbiased estimate of how well a model will perform on new data. However, finding a proper error metric to plug into these algorithms for comparing models fitted with different loss functions is not a simple task. For example, if we were to assess Quantile Regression with mean squared error, it is likely that the model of interest would be under-evaluated. This situation may lead to falsely selecting a Linear Regression model over a Quantile Regression model to be superior, or vise versa if mean absolute error was chosen for evaluation.

Utilizing only one error quantification method is insufficient for providing a complete picture of how models fitted with different loss functions are compared to each other and may lead to a biased model selection. When it comes to Ordinary Least Squares, squared error can determine the best model between Linear Regression, model

averaging and penalized methods. Similarly, an asymmetric loss function for given τ can be used to compare Quantile Regression models fitted with ρ_τ and penalized versions. However, determining whether a Quantile Regression model is superior to a Least Squares one requires using both squared error and quantile loss, while also taking into account the goals of the given application. For instance, in addition to predicting the heart volumes, we also aim to control under-predictions by estimating higher quantiles of the response. Diverging from the mean/median will have a toll on the prediction accuracy, but in our context this is acceptable up to a certain degree.

Another issue arising when it comes to the comparison of Least Squares with Quantile Regression is the fact that in order to assess relative accuracy, Least Squares models are fitted with a log-transformed response, while Quantile Regression estimates are on the original scale. As a result, the estimated error will be on a different scale and comparing them is not possible. One way to handle this situation is by reversing the logarithmic transformation of Least Squares. Since the models estimate $y_{pred} = log(\text{HtVol})$, we can take the exponent of y_{pred} and transform the response back on its initial scale. That way, both Quantile Regression and Least Squares will be on the same scale and the error estimates are comparable.

We should note however that in general, reverse transformation in Least Squares should be performed with caution. We know that Least Squares predict the mean value of the response Y, $E[Y]$. Thus, since we opted for relative accuracy assessment, we instead predict the mean value of $log(Y)$, $E[log(Y)]$. Our previous argument implied that by taking the exponent of $E[log(Y)]$, the resulting value would be $E[Y]$. This is not true however due to the Jensen inequality, which states that if Y is a real-valued random variable and ϕ is a concave function, then:

$$\phi(E[Y]) \geq E[\phi(Y)] \tag{26}$$

Since $log(\cdot)$ is a concave function, we can substitute it in (26):

$$log(E[Y]) \geq E[log(Y)]$$

or equivalently,

$$E[Y] \geq exp\{E[log(Y)]\}$$

Thus, taking the exponent of the log-transformed response may create a bias in our estimates. This bias is dependent on the difference $E[Y] - exp\{E[log(Y)]\}$, which is referred to as the Jensen gap.

In the case of Least Squares logarithm transformation, empirical results confirm that the Jensen gap is negligible and the bias created from the reverse transformation can be ignored. Therefore, for the purposes of model comparison, we will take the exponents of the Least Squares predictions and then effectively compare the error estimates with Quantile Regression methods.

Selecting an Optimal λ

The final aspect to be discussed about Cross Validation involves our approach for choosing λ when implementing Ridge/Lasso with Least Squares and Lasso with Quantile Regression. All the packages utilized for fitting the penalized models provide a function that returns an optimal value for λ. For the purposes of this application, we chose to employ these functions to estimate λ as they are part of the model fitting process each package offers:

- glmnet $\rightarrow$ we selected glmnet for Ridge and Lasso Regression. The value of λ is estimated by the function cv.glmnet() which implements 10-fold Cross Validation on an automatically generated grid, using the squared error to compare different λ's.

- hqreg $\rightarrow$ this package is for penalized Quantile Regression, with the function cv.hqreg() being used for selecting λ when it comes to l_1-penalty. Same as above, 10-fold Cross Validation is implemented with a generated grid, where the error is quantified with ρ_τ.

- rqPen $\rightarrow$ this package is also used for fitting a Penalized Quantile Regression model. Now, the optimal λ is yielded by rq.pen.cv(), which also performs 10-fold Cross Validation on an automatically generated grid. The difference is that now the errors are calculated by averaging $CV(f,\tau)$ from each fold f, where

$$CV(f,\tau) = \frac{1}{n_f}\sum_{i=1}^{n_f}\rho_\tau(y_{f,i} - x_{f,i}^\top \hat{\beta}_{\tau,\lambda}^{-f})$$

  with n_f being the number of samples of the f-fold and $\hat{\beta}_{\tau,\lambda}^{-f}$ being the coefficients occurring from fitting the model on all folds except for f.

- quantreg $\rightarrow$ when one performs penalized Quantile Regression with quantreg, rq.fit.lasso() is called internally. In contrast with the other packages, an analytical method is now used to determine the best λ, according to the proposal of Belloni and Chernozhukov (2011) [32].

The functions as described above will be incorporated in the Cross Validation process to provide unbiased estimates for the true error of penalized models. For that, we will call each function inside the Cross Validation loop using as an input the train set to get an optimal λ with which the fitted model will be tested on the left-out fold. Essentially, we will be performing Nested Cross Validation with each package except for quantreg, where an analytical approach is preferred for λ, which however does not create any bias since its selection is still achieved with the train set in each loop.

5.4 Permutation Tests

The output yielded from Repeated Cross Validation and Repeated Nested Cross Validation applied in the candidate model will be the main criterion for determining the best

model. However, in order to be more certain about its superiority among the rest, we can also examine the distribution of the errors from each iteration. Specifically, using $t = 40$ iterations will result in 40 different Cross Validation and Nested Cross Validation errors. These are part of each model's error distribution, which we desire to confirm that is unique. Any similarity observed between two well-performing models would partially discredit their inference value since then two sets of coefficients, which have different interpretations, would lead to the same predictive accuracy. Thus, relying on only one of these would not be as simple.

Permutation Tests are a powerful statistical tool for testing the null hypothesis that there is no difference between two distributions A and B. Suppose that we have m samples from A and n samples from B, with mean values μ_A and μ_B respectively. We define as a test statistic to be the absolute value of the difference of the mean values, i.e. $|\mu_A - \mu_B|$. Now, we concatenate the samples and shuffle their labels that indicate whether they belong to the distribution A or B. Next, we compute the test statistic for the permuted samples, repeating this process a large amount of times. Finally, we calculate the p-value as the proportion of the test statistic values that were more extreme than the observed test statistic. The algorithm for Permutation Tests with P shuffles is provided below:

`Algorithm 10`: *Permutation Test*

---

1. calculate the observed test statistic from the distributions A and B
2. concatenate the two sets of samples and initialize $count = 0$
3. for $i = 1$ to P do:
   (a) shuffle the labels of the samples
   (b) compute the test statistic for the permuted samples
   (c) if the test statistic is more extreme than the observed one, increase *count* by 1
4. return $\frac{count}{P}$, which is the p-value for the hypothesis test

---

A small p-value will allow us to reject the null hypothesis and conclude that the distributions are indeed different from a statistical perspective. A large p-value would fail to provide sufficient evidence to reject the null hypothesis.

In our application, we want to compare the distributions of the errors between two models. To do so, the same way for quantifying the errors must be used, otherwise it is obvious that the distributions will be different. After performing Repeated Cross Validation or Repeated Nested Cross Validation for each model, we define the null hypothesis as follows:

H_0: Both sets of samples come from the same distribution

We test the null hypothesis with a permutation test, where we permute the sample labels $P = 10.000$ times.

We should mention that there are also other options for the test statistic, one of the most popular alternatives being the absolute value of the difference of the medians. We chose to work with the mean instead because the sample size is 40, which is too small for the medians to be reliable.

5.5 Confidence Intervals

Confidence intervals are a popular statistical method for extracting information about an unknown parameter what we wish to estimate, based on the available data. They provide a measure of variability for the estimate and indicate the range where the true parameter value is likely to fall, effectively generalizing the simple point estimate. Confidence intervals are necessary when it comes to decision-making, as they allows researchers to assess the reliability of their estimates and to what extend they can be trusted.

A confidence interval is characterized by a lower and an upper limit that define a range where the parameter is likely to fall. The interval is constructed according to the available data, taking into account our confidence level γ, which determines how confident we want to be that the true parameter lies within the interval. As the level of confidence γ increases, the interval widens, including a larger set of potential values for the true parameter. Intuitively, this makes sense as a wider interval increases the probability (hence, the confidence) for the inclusion of the true parameter. However, wider intervals also indicate a greater amount of uncertainty about this parameter, therefore we wish the constructed interval to be narrow, while remaining at a high confidence level. A typical choice for the confidence level γ is 0.9 or 0.95.

De Groot and Schervish (2002) provide the following interpretation of a confidence interval (A,B) at a level of confidence γ for a parameter θ [33]:

Suppose that the available set of samples is one of many possible samples that we could have observed from the same distribution. Each of these samples would allow as to compute an observed interval. Out of all the observed intervals, we would expect $100 \cdot \gamma$ *% of these to contain the true value* θ*. Even if we obtained many such observed intervals, we cannot know which contain the true value and which do not.*

One of the main advantages of confidence intervals is that they can be used as an alternative method for testing hypotheses. Suppose that we are interested in a parameter θ and for any value θ_0 we wish to test the following hypotheses at a level of significance α:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Then, we can estimate the confidence interval at confidence level $\gamma = 1 - \alpha$ and reject the null hypothesis if θ_0 is not contained in the interval.

Confidence Intervals for Linear Model Coefficients

Suppose we have a linear model, informed by the set of variables $\{X_1, ..., X_p\}$ with each X_i being associated with a coefficient β_i. One of the most common queries to address given the model is whether X_i is informative to the response. If X_i was irrelevant, then the corresponding coefficient β_i would be exactly zero. Thus, the questioning of relevance of X_j to the response can be expressed mathematically through the hypotheses

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Rejecting the null hypothesis is then equivalent to concluding that X_i is not irrelevant. In order to reject H_0, we can construct a confidence interval for β_i and investigate whether 0 lies within the interval.

There are multiple ways to acquire estimates for the lower and upper limit of the confidence interval at any given confidence level, most of which are heavily dependent on the model's underlying assumptions. It is often the case that in practice, these assumptions are violated thus leading to an inaccurate and potentially misleading confidence interval. To address this issue, we will discuss a general resampling method for approximating confidence intervals without relying on any assumptions, called bootstrapping.

Bootstrap Confidence Intervals

Bootstrapping is the statistical procedure of randomly resampling from an available dataset to estimate the sample distribution of almost any desired statistic [34]. Suppose our dataset consists of $(x_1, ..., x_n)$. Then, a bootstrap sample is random sample $(x_1^*, ..., x_n^*)$ of size n drawn with replacement from the initial dataset. Each member of the bootstrap sample consists of members from the original dataset, some appearing once, some two times or more, while some appear zero times. In order to report an estimation of a statistic, we construct multiple bootstrap samples, which we then utilize appropriately according to the needs of our statistic of interest. For the purposes of our application, we will obtain $B = 599$ bootstrap samples, a number Wilcox (2010) suggested is large enough to capture the variability and uncertainty in the data [35].

We will now elaborate on the process for obtaining bootstrap confidence intervals for the coefficients of a linear model. Firstly, we obtain $B = 599$ bootstrap samples by randomly selecting observations of our dataset with replacement, with each sample having the same number of rows as the original dataset. Then, for every bootstrap sample, we fit the linear model, thus acquiring 599 different estimates for each coefficient β_i, which we sort in ascending order. Now, the estimated confidence interval will rely on the appropriate percentile selection of the sample distribution of each coefficient. The choice of percentiles must be symmetric so that the confidence level γ is spread equally on both sides of the distribution. This is achieved by selecting the $\frac{1-\gamma}{2} \cdot 100$th and $\frac{1+\gamma}{2} \cdot 100$th percentiles as lower and upper bound of the confidence interval respectively.

The following algorithm captures the steps for constructing confidence intervals via bootstrapping:

`Algorithm 11`: *Bootstrap Confidence Intervals*

---

1. Initialize p empty lists $L_1, ..., L_p$, one for each predictor (including the intercept). Initialize another empty list $conf\_int$
2. for $i = 1$ to B do:
   (a) obtain a bootstrap sample D_i
   (b) fit the linear model using D_i
   (c) append each of the estimated parameters β_i to L_i
3. for each L_i do:
   (a) append the pair (A, B) to $conf\_int$, where A, B are the $\frac{1-\gamma}{2} \cdot 100$th and $\frac{1+\gamma}{2} \cdot 100$th percentiles of L_i respectively
4. return $conf\_int$

The returned object of Algorithm 11 will be a list where the i'th element corresponds to the confidence interval of β_i.

We should note that when there is hyperparameter tuning involved, such as in the case of the penalty term λ in penalized Least Squares or penalized Quantile Regression, we must calculate the optimal lamda in each iteration using each time the new bootstrap sample D_i. That way, we ensure unbiasedness during the estimation of the linear model coefficients and therefore during the confidence interval estimation.

6. Application

This chapter presents the outcomes and findings derived from a carefully selected array of statistical techniques presented in this study. The primary objective is to showcase the results of statistical modeling for each proposed method, while evaluating and comparing their respective performance. Given the extensive scope of analysis for both Least Squares and Quantile Regression, we discuss each method separately. Subsequently, we compare the results obtained from the dominant Least Squares models with those from Quantile Regression.

6.1 Model Inference

Starting off, we provide a comprehensive analysis of each Least Squares method applied to our dataset. We begin by examining the outcomes of stepwise selection and model averaging when applied to the full dataset. Subsequently, we dive into a detailed exploration of Ridge and Lasso techniques, which includes the determination of the optimal penalty parameter λ, thus inspecting the tradeoff between bias and variance. We then examine the coefficients corresponding to the selected λ, aiming to display the shrinkage effect in comparison to standard Linear Regression.

After gaining an overall view of Least Squares based techniques, we shift our focus to Quantile Regression. For simple Quantile Regression and penalized Quantile Regression implemented with each of the packages quantreg, rqPen and hqreg, we provide a summary of the coefficients across different quantiles, from lowest to highest. Then, we dive deeper into the 75th percentile, which we are mostly interested in, comparing the coefficient behaviour with median regression, where $\tau = 0.5$.

Stepwise Selection and Model Averaging

When it comes to feature selection, we argued that best subset selection with Information Theory criteria is not reliable and may lead to overfitting. For that reason, we consider stepwise selection techniques only and specifically forward selection and backward elimination. Applying these algorithms on our dataset leads to the following predictor subsets:

- Forward selection: $\{\text{Male}, \text{Ht}^2, \text{Wt}^2, \frac{\text{Ht}}{\text{Age}}\}$
- Backward elimination: $\{\text{Male}, \text{Age}^2, \text{Ht}^2, \text{Wt}^2, \frac{\text{Ht}}{\text{Age}}, \frac{\text{Wt}}{\text{Age}}\}$

We can observe that backward elimination managed to remove only 2 out of 8 features, with the ones survived being a superset of the forward output.

We now present the outcome of model averaging. For that, we constructed the averaged linear models following both ways for averaging the coefficients, as discussed in Chapter 3. For this purpose, the AIC_c was calculated for all possible subsets of predictors and the ones with Akaike differences $\Delta_i < 2$ were kept to be averaged. The resulting subsets along with the corresponding models' Δ_i's and weights are summarized in the following table:

	Male	Age^2	Ht^2	Wt^2	Ht/Age	Wt/Age	Male·Age	Male·BMI	Δ_i	w_i
Model 1	*		*	*	*				0	0.194
Model 2	*	*	*	*	*	*			0.15	0.180
Model 3			*	*	*				0.40	0.159
Model 4	*	*	*	*	*				0.66	0.140
Model 5			*	*	*	*			1.01	0.117
Model 6	*		*	*	*	*			1.09	0.112
Model 7	*		*	*	*		*		1.35	0.099

Table 3: Linear Regression models with $\Delta_i < 2$

In total, we have 7 different selected subsets. It is interesting to observe that the predictors Ht^2, Wt^2, Ht/Age are found in all subsets, while they are also selected by both stepwise selection algorithms. On the other hand, the interaction terms $\text{Male} \cdot \text{Age}$ and $\text{Male} \cdot \text{BMI}$ seem to not benefit the Least Squares model as they were not found in almost any of the subsets, except for the weakest one (Model 7) where only $\text{Male} \cdot \text{Age}$ is included. Moreover, the fact that the strongest subset as determined from model averaging comprises of the same variables forward selection suggested indicates some consistency regarding our findings.

We now compare the averaged coefficients with the ones of the full Linear Regression model:

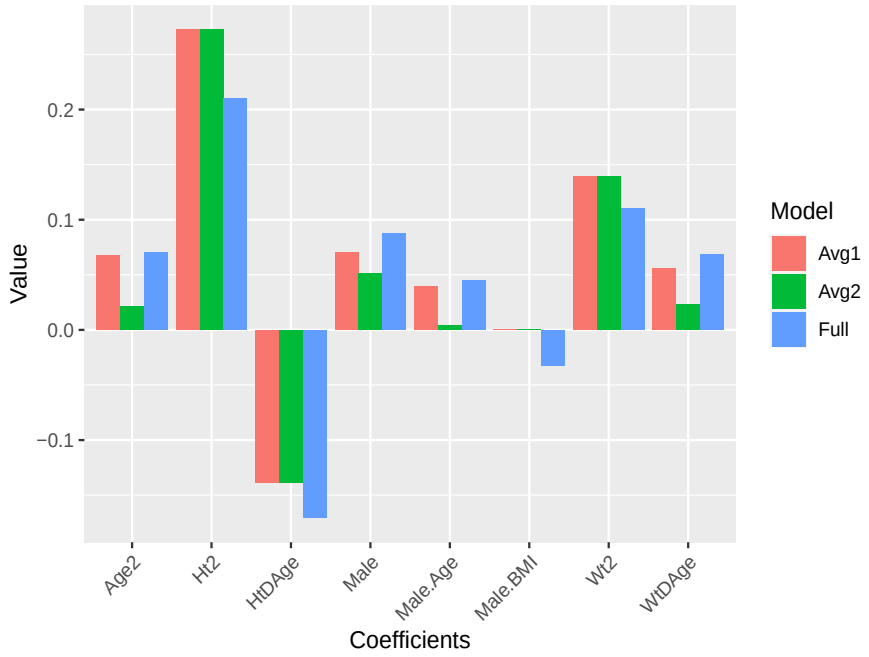


Figure 11: Barplot of model averaging and full model's coefficients. 'Avg1' corresponds to considering only models where x_j appears when averaging the coefficients β_j. 'Avg2' corresponds to averaging under the assumption that β_j appears in all models, but in some it is exactly equal to zero. The variables 'Age2', 'Ht2', 'Wt2' are Age^2, Ht^2, Wt^2 respectively, 'Male.Age', 'Male.BMI' are $\text{Male} \cdot \text{Age}$ and $\text{Male} \cdot \text{BMI}$ and 'HtDAge', 'WtDAge' are Ht/Age, Wt/Age.

Looking at the barplot, we can notice that some of the coefficients are equal for both averaged models. As anticipated, these are exactly the predictors that appear in all models used for averaging. For these predictors, we have that $w_+(j) = 1$, meaning that both ways for averaging yield the same value. Additionally, the shrinkage effect of 'Avg2' can be clearly seen, with coefficients being smaller or equal to 'Avg1'. Finally, it is worth mentioning that averaging resulted in somewhat different coefficients in comparison with the full model.

Ridge and Lasso Regression

We proceed with the inspection of Ridge and Lasso Regression. In order to identify the best λ for the entire dataset, we utilize the function cv.glmnet() that performs Cross Validation across a grid of possible values. The following figures created by cv.glmnet() inform us on how relative accuracy progresses as we move along the lambda grid:

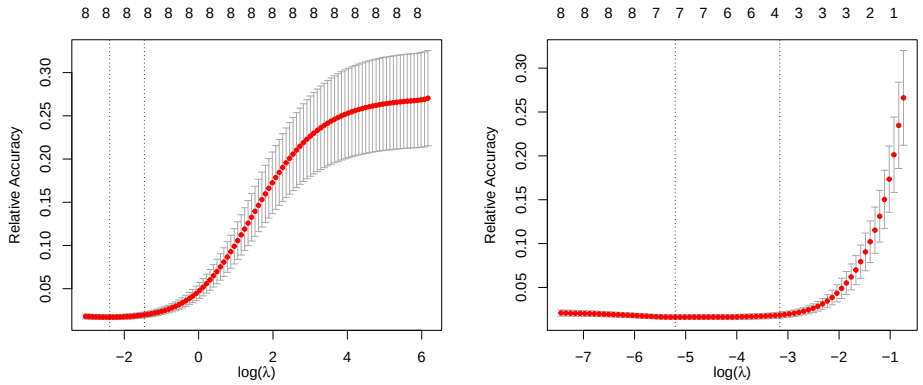


Figure 12: Relative accuracy across λ grid. The left plot captures the λ's for Ridge and the right plot for Lasso. The y-axis represents the relative accuracy and the x-axis represents the λ grid on a logarithmic scale. The interval lines around each point are the lower and upper standard deviation curves, created by adding and subtracting one standard error from the relative accuracy at each point. The values on top of the plots represent the number of non-zero coefficients as λ increases.

Firstly, we can observe that Ridge does not have coefficients that are exactly zero, in contrast with Lasso where as the constraint caused by the penalty term becomes stronger with higher λ's, more coefficients are set to zero. Moreover, for both models smaller values of λ lead to better relative accuracy and as λ increases, the errors become larger almost monotonically, while the wider intervals imply a steady increase of uncertainty around each point. The first vertical dashed line indicates the λ for which relative accuracy is minimized, with the second line indicating the largest value of λ such that the error is within a standard error from the minimum error. Both of these values can be extracted from cv.glmnet(), with the second being available for when one wishes to sacrifice some of the model performance to increase the effect of shrinkage by selecting a larger λ. In our application we select the first value of λ, thus having $\lambda_{Ridge_{best}} \approx 0.0477$ and $\lambda_{Lasso_{best}} \approx 0.008$. The fact that Lasso has such a small λ means that the penalty effect will not be as significant, therefore not many coefficients are expected to be zero. The following barplot comparing the coefficients of Ridge and Lasso Regression with the full model verifies that this is indeed the case:

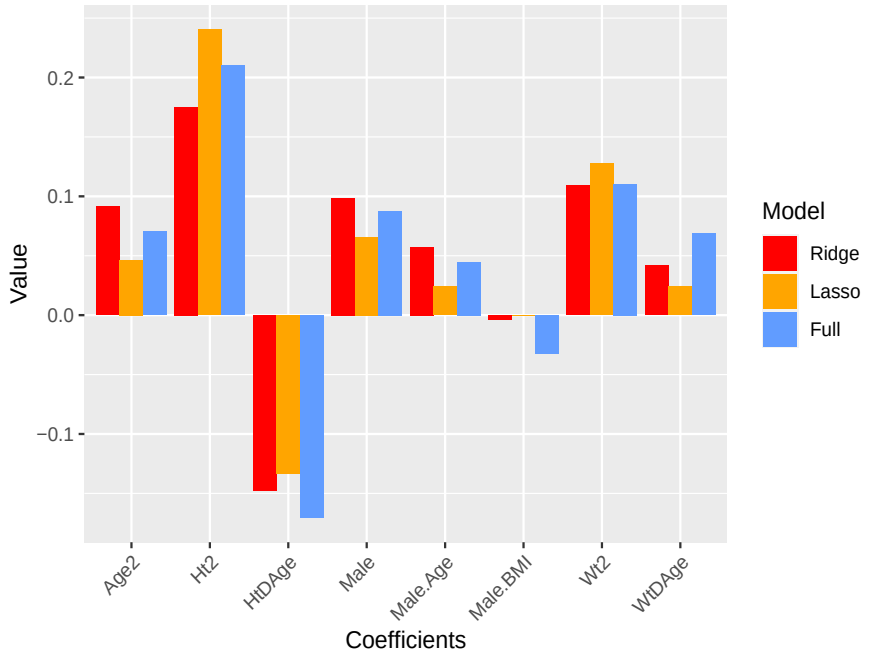


Figure 13: Barplot of Ridge, Lasso and full model's coefficients

Lasso resulted in the coefficient of Male $\cdot$ BMI to be exactly 0, while Ridge also provides a very small value for this predictor, confirming our previous findings that it does not have a significant predictive value for the heart volumes. Looking at the remaining predictors, the shrinkage effect becomes apparent for both methods when compared to the full model.

In order to further inspect the coefficients, we analyze the results of bootstrapping on Ridge and Lasso Regression. Specifically, for each coefficient we construct a boxplot informed by the 599 coefficient estimates obtained from the bootstrap samples. To properly assess the significance of each variable in predicting heart volumes, we include 95% confidence intervals in the same plot. These confidence intervals enable us to evaluate the null hypothesis of variable irrelevance, providing valuable information on the importance of each predictor in the prediction model. The following figure contains the boxplots for all coefficients of Ridge Regression:

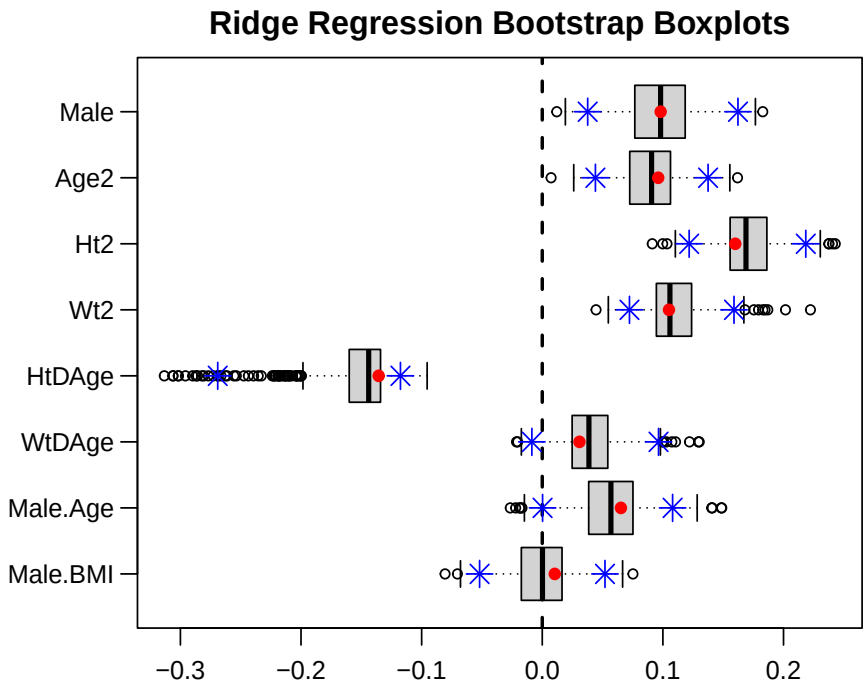


Figure 14: Ridge Regression Bootstrap coefficient boxplots. The blue points indicate the 95% confidence interval for each predictor and the red dots are the coefficient estimates based on the entire dataset

We can observe that for the first 5 predictors, the confidence intervals do not include 0, meaning that the null hypothesis for each being uninformative to the response can be rejected. For Wt/Age and Male $\cdot$ BMI, 0 is withing the interval, so we do not have evidence about them being relevant to reponse. As for Male $\cdot$ Age, the lower bound of the interval is barely over 0 (≈ 0.0003), therefore there is just enough evidence to reject the null hypothesis. However, this conclusion is reported with caution because the boxplot intersects 0 indicating that this predictor may not be as significant.

We now display the same boxplots for Lasso Regression:

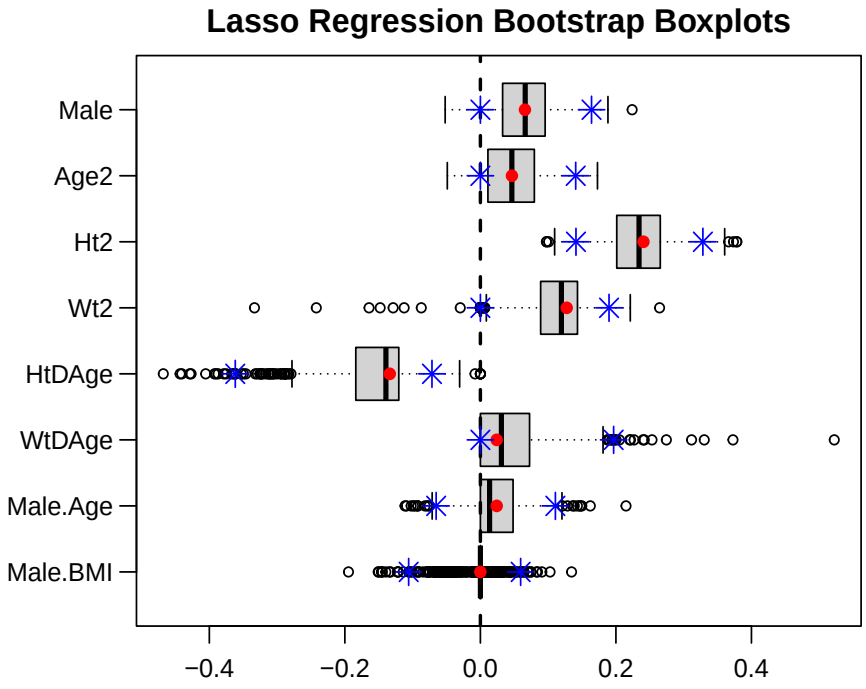


Figure 15: Lasso Regression Bootstrap coefficient boxplots

Here, the Lasso Regression coefficient boxplots imply that there is greater uncertainty about each predictor's relevance with respsect to heart volumes. The only variables for which the null hypothesis is rejected are Ht^2 and Ht/Age. On the other hand, all other confidence intervals either contain 0, or one of their bounds is exactly 0. The latter occurs in multiple occasions (predictors Male, Age^2, Wt^2 and Wt/Age), which prior to this analysis was anticipated due to the fact that Lasso frequently sets coefficients to be exactly 0. As a result, higher density of values that are exactly 0 is likely to be found closer to the lower or upper bounds of the interval and thus one of the percentiles corresponding to our confidence level may end up being exactly 0.

Quantile Regression

We initiate our discussion about model inference with Quantile Regression by demonstrating one of its most useful properties which is the ability to report statistically well-founded estimates for prediction intervals. Prediction intervals are a more generalized way for making predictions, providing an interval where we believe the true value that we are trying to estimate is more likely to fall in. The fact that these estimates include a range of possible values for the response (thus quantifying the uncertainty of our predictions) makes then extremely valuable for many real-life applications. For instance, doctors who are called to determine heart compatibility for transplantation will be able to provide a more informed decision when inspecting an interval where the

heart volume is likely to fall in, compared to simple point estimates other models such as Least Squares would report.

The following example demonstrates the usage of Quantile Regression prediction intervals:

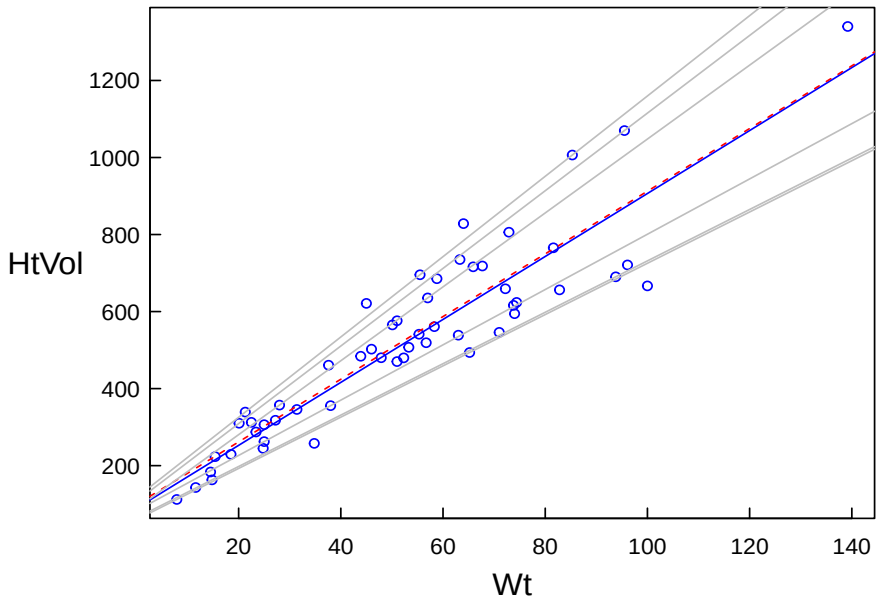


Figure 16: Quantile Regression prediction lines for multiple quantiles, informed by a single variable, the patient's weight. The red dashed line is the Linear Regression prediction line, the blue line is for Median Regression ($\tau = 0.5$) and the grey lines are for Quantile Regression for τ =0.05, 0.1, 0.25, 0.75, 0.9, 0.95.

If we were to inspect solely the Least Squares findings for a given weight, we would get a point estimate, with great uncertainty about the accuracy of this prediction. On the other hand, if we were to assess the Quantile Regression lines for $\tau = 0.25$ and $\tau = 0.75$ for the same weight, we would get an interval where we would expect the true value to fall in with 50% confidence. If we wanted to be more conservative, we could select the 0.05 and 0.95 quantiles and the resulting interval would provide us a possible range for the true value with 90% confidence.

Apart from prediction intervals, Quantile Regression has great explanatory properties when the coefficient estimates are examined across an array of different quantiles. Inspecting the behaviour of each coefficient starting from lower quantiles and moving up

can provide useful additional information about the corresponding predictors, which would not be accessible through a Least Squares model.

The following figure provides a concise summary of the Quantile Regression results, with each plot representing one coefficient in each Quantile Regression model, for $\tau = 0.1, 0.2, ..., 0.9$:

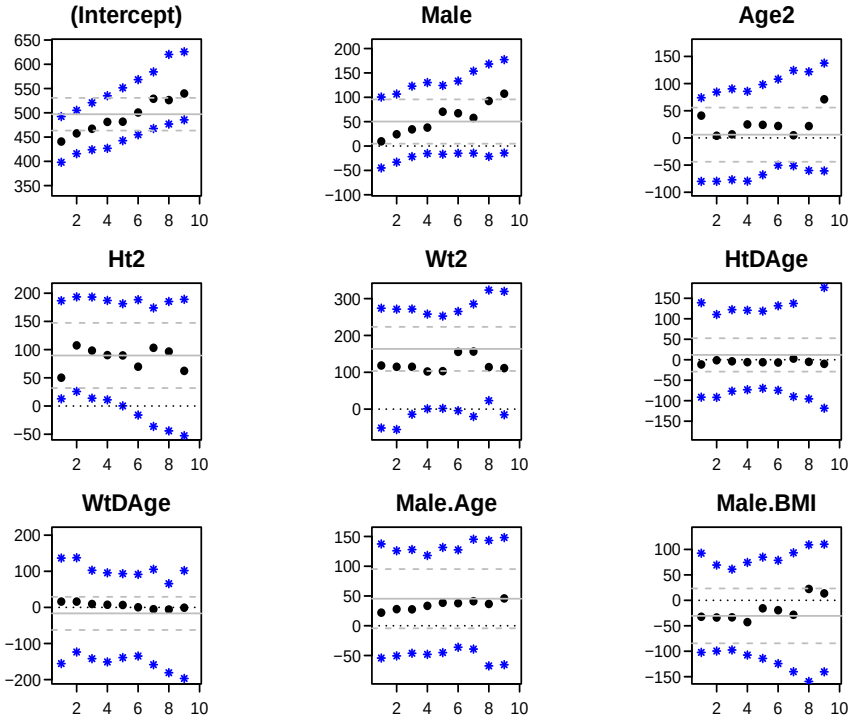


Figure 17: Quantile Regression coefficients across different quantiles. In each panel, the black dots correspond to Quantile Regression coefficient estimates. The gray lines are the Linear Regression estimates for the respective coefficients, with the gray dashed lines representing the 95% confidence intervals, derived based on the normality assumptions of Least Squares. For coefficient comparability purposes, the Linear Regression model was fit here with HtVol and not the logarithmic transformation as it was in the Least Squares analysis. The blue points are the 95% bootstrap confidence intervals for each quantile.

A general inspection of the panels reveals the significant amount of uncertainty surrounding the coefficient estimates. Apart from the intercept, the Quantile Regression confidence intervals of each predictor include 0 in almost every quantile, meaning that the models fail to provide sufficient evidence about the relevance of any predictor. Simul-

taneously, the intervals are also alarmingly wide, to the extend of containing the Least Squares confidence intervals in every panel. A possible explanation for this situation may be hidden in the bootstrapping process. The bootstrap samples are drawn from a very small dataset, therefore the rows that are left out (and the ones that are included multiple times) may lead to considerable changes in the initial quantile distribution. Consequently, Quantile Regression on different bootstrap samples would yield diverse coefficient estimates, resulting in a wider confidence interval.

The uncertainty surrounding Quantile Regression coefficient estimates partially discredits their interpretational properties when it comes to predicting the heart volumes. For that reason, we will discuss inferences entailed by the coefficient point estimates while keeping in mind the aforementioned variability.

In the first panel of the figure, the intercept can be interpreted as the estimated conditional quantile function of the heart volumes for a female pediatric patient, therefore the monotonic increase is anticipated. The second panel informs us that boys have greater heart volumes than girls by about 50mL according to the Least Squares estimates of the mean effect, but according to Quantile Regression, the disparity is smaller than 50mL in the lower tail, reaching 100mL in the upper quantiles. We should point out however that since the Quantile Regression coefficients lie within the Least Squares confidence interval, the effect of Male across the quantiles should be investigated further to conclude a differentiation from Least Squares. Given that the inclusion of coefficient estimates inside the Least Squares confidence intervals occurs in every subfigure, the same applies for all predictors.

Focusing on the point estimates of Quantile Regression, we can see that the terms Ht/Age and Wt/Age appear to have an insignificant effect on the heart volumes, contrary to the quadratic term of weight, the effect of which remains over 100mL for all quantiles. This finding about weight aligns with the popular DRBW ratio utilized for assessing compatibility of a candidate allograft, which suggests that donor-recipient weight matching is usually sufficient to conclude heart compatibility. Moreover, the slight upward trend of Male $\cdot$ Age coefficients informs us that when a patient is male, the effect of his age becomes greater as we move to higher quantiles. Finally, the remaining predictor's estimates fluctuate around the Least Squares coefficients, thus they do not offer any explanatory properties additionally to Least Squares.

Given the uncertainty stemming from the width of the confidence intervals and the fact that they stretch around 0, we once again stress that these results should be viewed with caution.

We now present the coefficient summary for Lasso-based penalized Quantile Regression as implemented from quantreg, hqreg, rqPen. For each, we utilized the functions described in Chapter 5 to select the optimal λ in each quantile, using all the data. For the purpose of comparison with Least Squares, we include the coefficient point estimates and confidence intervals of Lasso Regression, which applies the same penalty to Least Squares as the penalty we consider for Quantile Regression.

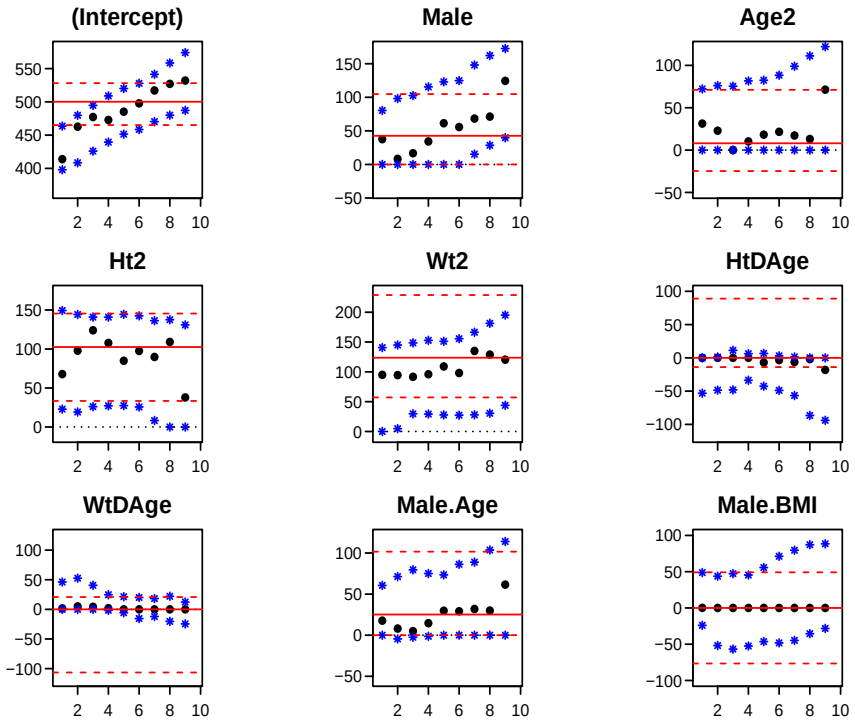


Figure 18: Lasso penalty Quantile Regression coefficients from quantreg. Instead of Linear Regression, the red lines now represent the Lasso Regression estimates for the respective coefficients, with the red dashed lines being the 95% confidence interval, derived by bootstrapping with 599 bootstrap samples. The Lasso Regression model was again fit here with HtVol and not with the logarithmic transformed version of the response. Same as before, the blue points represent the bounds of the 95% bootstrap confidence interval for each quantile based on 599 bootstrap samples and the black dots correspond to the penalized Quantile Regression coefficient estimates.

In contrast with simple Quantile Regression, the confidence intervals are now considerably narrower, which can be attributed the reduction of variance caused by the penalty term. Additionally, there are now some panels where the penalized Quantile Regression intervals do not contain 0, thus leading to the rejection of the null hypothesis regarding a variable's irrelevance to the predictor. This is more apparent for the predictors Ht^2 and Wt^2, where apart from the upper or lower tail respectively, in all other quantiles the lower bound is positive. Nonetheless, the uncertainty is still present with most intervals intersecting zero or having one of their bounds to be exactly 0 due to the Lasso penalty effect.

Looking at the figure, one can observe the shrinkage effect penalization has on the coefficients. The variables Ht/Age, Wt/Age have very small coefficients, with some

being exactly zero. This situation is much more clear with Male $\cdot$ BMI, the coefficient of which has been set to zero at all quantiles. It is worth noticing that these three predictors are also exactly the ones that Lasso Regression has excluded from the model by setting the corresponding coefficients to zero. Additionally, coefficients for height and weight, which enter the model as a quadratic terms, fluctuate around the corresponding Lasso Regression estimate without any significant divergence. The same goes for Male $\cdot$ Age, while the coefficient of Male slightly increases as we move into higher quantiles. Finally, the effect of Age^2 appears to be relatively stable across all quantiles. Given the analysis above, it is important to mention that once again, the penalized Quantile Regression coefficient estimates are contained in the Lasso Regression confidence interval, thus their interpretation needs to be more researched into.

In Chapter 4, we argued that for quantreg, the formulation of the penalized Quantile Regression optimization problem (19) applies a greater penalty on positive coefficients when $\tau < 0.5$ and on negative coefficients when $\tau > 0.5$. Thus, we would expect the coefficients to have smaller values in the lower quantiles, becoming larger as τ increases. This however did not occur in this case, which could be attributed to the small sample set that does not allow this property to be uncovered.

Next, we display the coefficient summary for penalized Quantile Regression with rqPen:

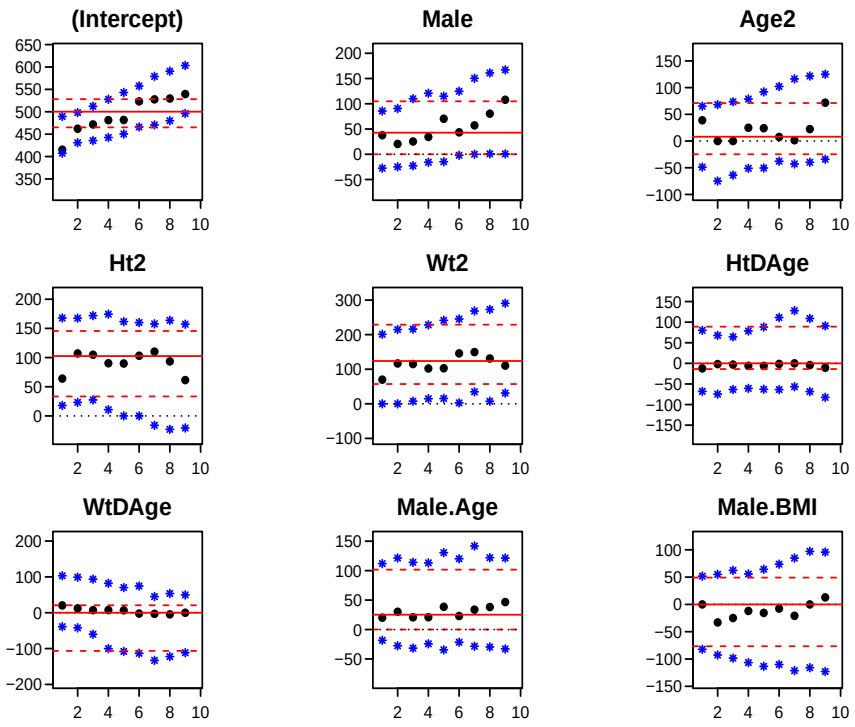


Figure 19: Lasso penalty Quantile Regression coefficients from rqPen

Unexpectedly, when implementing penalized Quantile Regression with rqPen, the confidence intervals are significantly wider in comparison to those from quantreg. In fact, they bear great resemblance with the intervals of simple Quantile Regression. This situation may be attributed to simple Quantile Regression and penalized Quantile Regression from rqPen both using the 'br' algorithm to fit the models. Given the same solving algorithm, a potentially small penalty hyperparameter λ may not lead to any considerable shrinkage effect, thus the resulting models would be very similar. This situation is explored in more detail later, when inspecting specific quantiles for our candidate models.

In general, even though Lasso-based methods are expected to set some coefficients to zero, this did not happen with rqPen. Sparse cases of coefficients that are exactly zero are present but there is no consistent exclusion of any predictor from the model as it was with quantreg. This evidence suggests a weak contribution of the penalty term to the model, which can be also verified by comparing the respective panels with Quantile Regression for each predictor. Specifically, apart from similar confidence intervals, one can also identify similar point estimates across the quantiles for each coefficient, meaning that rqPen seems to fail retaining the properties of a penalized method.

Finally, we display the coefficient summary for penalized Quantile Regression with hqreg:

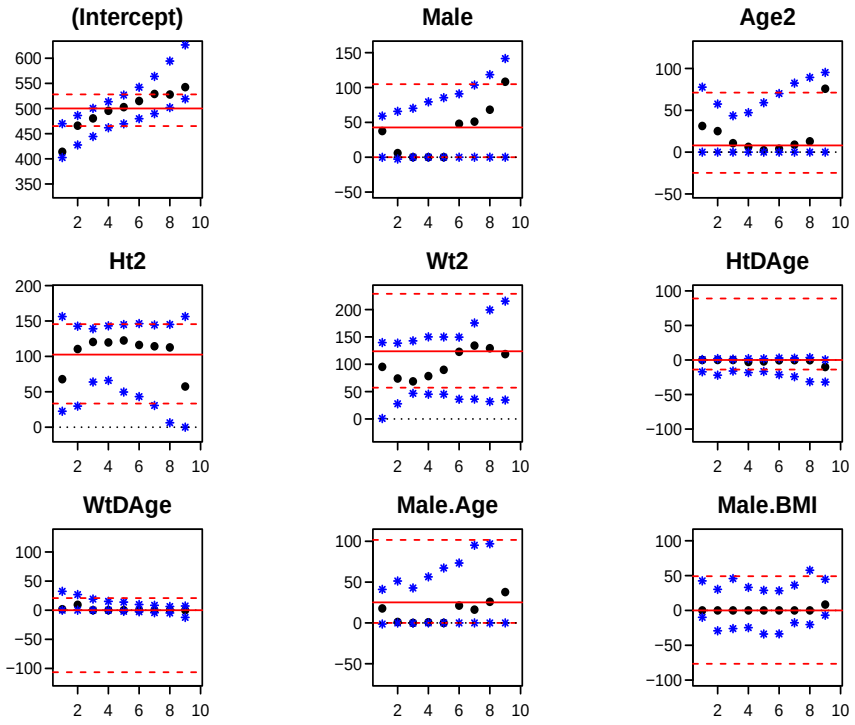


Figure 20: Lasso penalty Quantile Regression coefficients from hqreg

With hqreg, the confidence intervals are the narrowest of all Quantile Regression models considered. The predictors Ht^2 and Wt^2 are the only ones with intervals which do not include 0, enabling us to reject the null hypothesis that these are irrelevant for the prediction of heart volumes. We should note that with quantreg, we were also able to reject the null hypothesis for these predictors. In addition, Ht/Age, Wt/Age and $\text{Male} \cdot \text{BMI}$ all have very narrow confidence intervals that contain 0, indicating that these may not be as informative for the response. This argument is further reinforced by the fact that the point estimates for these for all quantiles are either 0 or negligibly small. The lower bound of the confidence intervals of Male, Age^2 and $\text{Male} \cdot \text{Age}$ are 0 in all quantiles, while some of the point estimates for the corresponding coefficients are also 0. Thus according to the hqreg implementation, there is no evidence to conclude that these are relevant to our predictions.

We now shift our focus to specific quantiles. As we have explained, we are interested in

the 75th percentile. Therefore, we create a barplot containing the coefficients of each model for $\tau = 0.75$, after firstly displaying the same barplot for median regression. Our purpose is to compare the coefficients at the 0.75 quantile and present their differences compared to the symmetric median case.

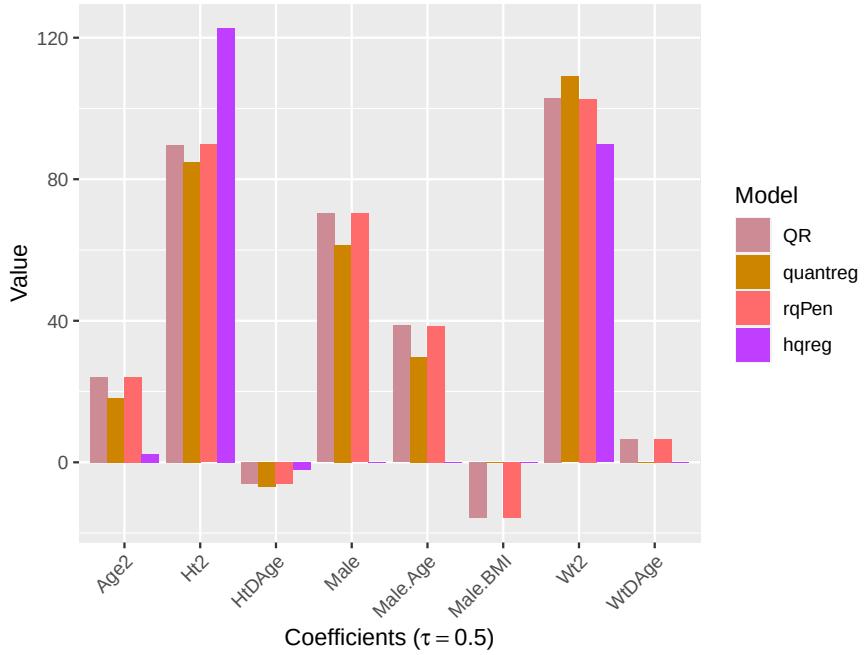


Figure 21: Quantile Regression and penalized Quantile Regression coefficients for $\tau = 0.5$. For each coefficient, the first bar corresponds to Quantile Regression and the other 3 are for each package with which we implemented penalized Quantile Regression.

We can observe that the coefficients of all models are in general quite similar. Specifically, the models from simple Quantile Regression and its penalized version from rqPen lead to almost equal coefficient values. This situation can be attributed to the small hyperparameter $\lambda \approx 0.004$ selected by the function rq.pen.cv() in conjuction with the same solving algorithm used to fit each of the models. Furthermore, it is worth mentioning that hqreg has a great shrinkage effect, since half of the predictors have been excluded from the models. At the same time, even though Age^2 and Ht/Age have non-zero coefficients, these are very small indicating a minimal influence towards the heart volumes. The only terms determined to be significant are Wt^2 and Ht^2, with the latter having the largest coefficient among all models by far. As anticipated, the optimal λ as selected from cv.hqreg() is relatively large compared to that of rqPen, valued at ≈ 0.086.

We now display the same barplot for $\tau = 0.75$:

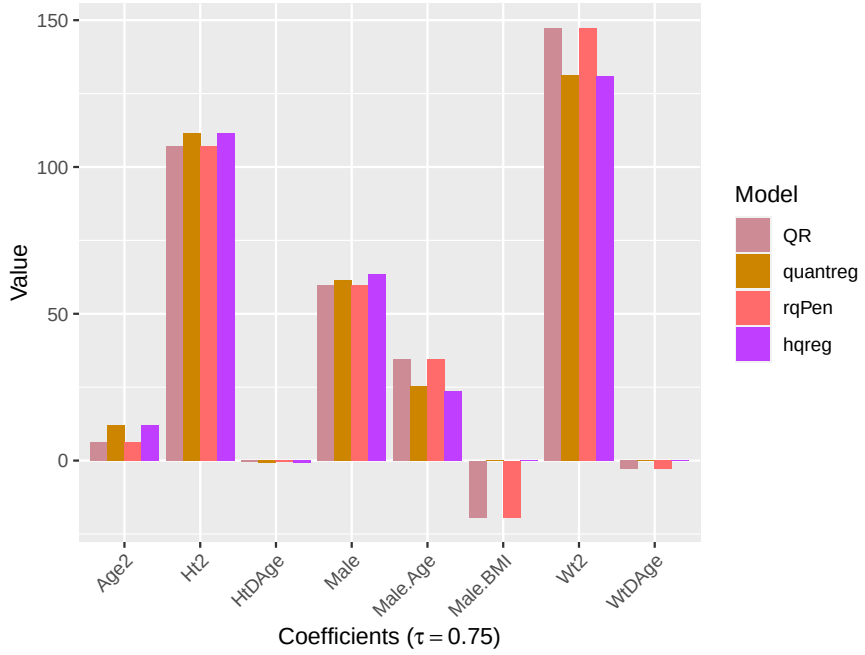


Figure 22: Quantile Regression and Penalized Quantile Regression coefficients for $\tau = 0.75$

This barplot provides a different picture about the coefficients from each model. At a first glance, we can identify similar patterns between simple Quantile Regression and rqPen penalized Quantile Regression as well as quantreg and hqreg penalized Quantile Regression. In fact, the coefficients of Quantile Regression and those from rqPen are exactly the same, indicating that the penalty term did not affect the Quantile Regression coefficients at all. This is due to the fact that the optimal λ selected by rq.pen.cv() is again very small, approximately 0.003. All models provide evidence that the terms Ht/Age, Male $\cdot$ BMI and Wt/Age are not influential in estimating the response at the 75th percentile, while the quadratic terms Ht^2 and Wt^2 have the greatest effect on heart volume estimates.

As we have mentioned, the limited sample size may lead to a problematic amount of variability. In order to further inspect this situation at the 0.75 quantile that we are interested in, we perform bootstrapping to each of the Quantile Regression and penalized Quantile Regression models. After obtaining one set of coefficients from each model for each of the 599 bootstrap samples, we display our findings through boxplots, the same way as we did with Ridge and Lasso Regression previously.

We firstly present the boxplots of simple Quantile Regression:

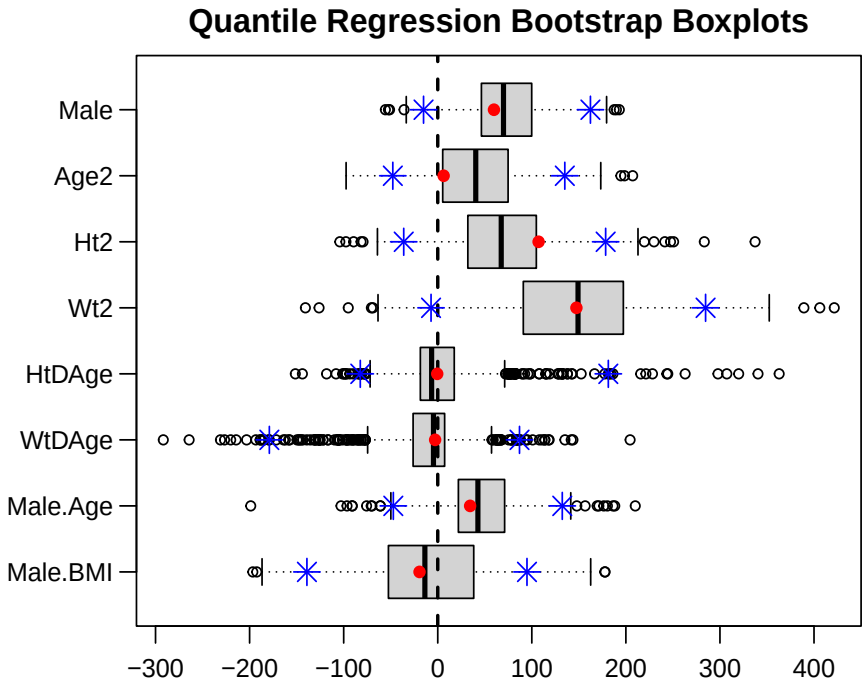


Figure 23: Quantile Regression Bootstrap coefficient boxplots. The blue points indicate the 95% confidence intervals and the red points are the coefficient point estimates based on the data.

Here, each of the coefficients comes not only with high variability, but also with high uncertainty, due to the fact that all confidence interval include 0, therefore we can not conclude the relevance of any variable to the response.

We now proceed with the quantreg implementation:

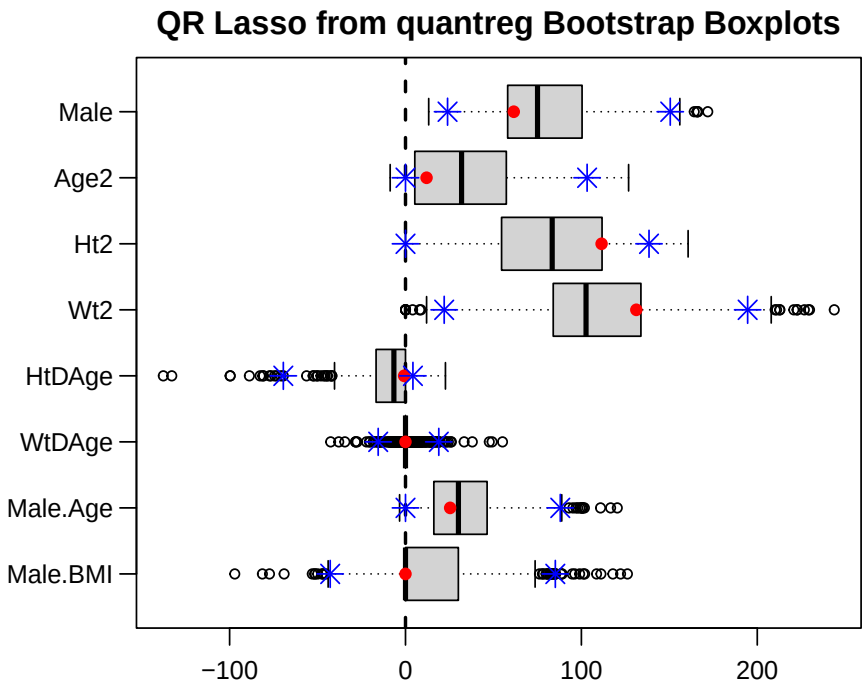


Figure 24: Lasso penalty Quantile Regression from quantreg Bootstrap coefficient boxplots

The range of each confidence interval is now smaller compared to Quantile Regression, indicating less variability. Moreover, the only confidence intervals that do not contain 0 are those of Male and Wt^2, thus with quantreg we can conclude that these predictors are significant for predicting heart volumes.

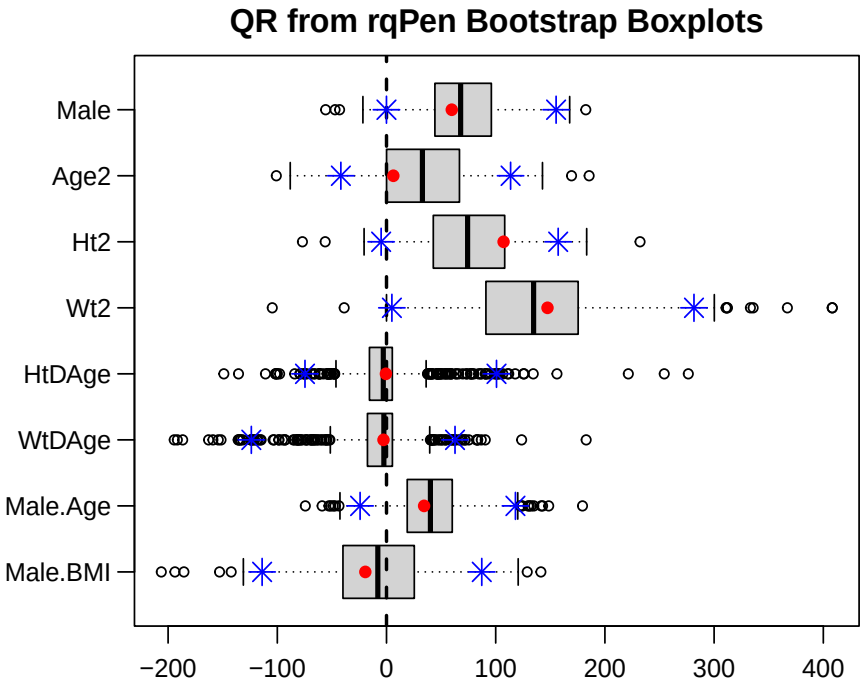


Figure 25: Lasso penalty Quantile Regression from rqPen Bootstrap coefficient boxplots

As expected, the rqPen boxplots are almost identical to those of simple Quantile Regression, with the only difference being that here the lower limit of the Wt^2 confidence interval is just over 0, therefore there is evidence to support the predictor's relevance to the response.

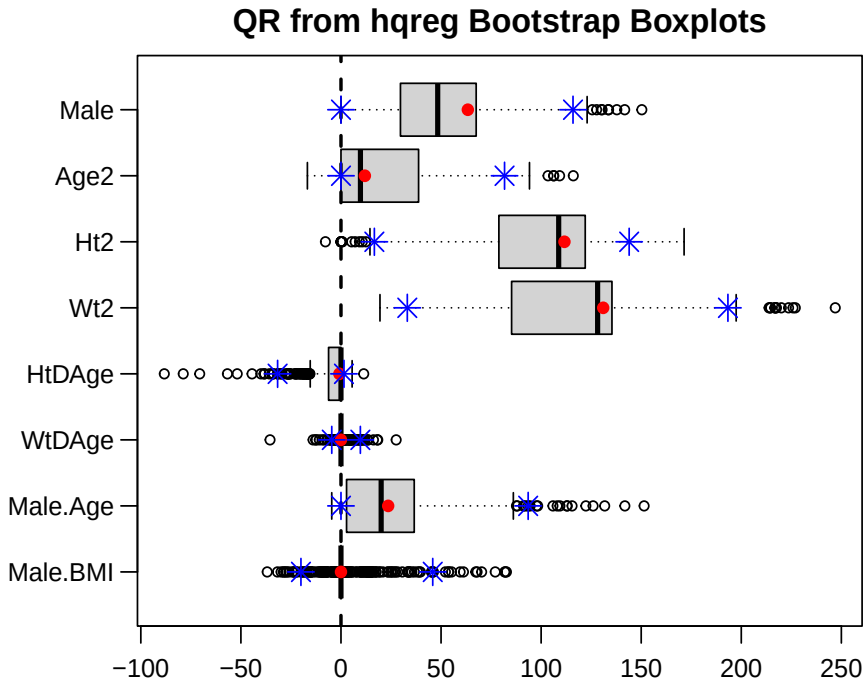


Figure 26: Lasso penalty Quantile Regression from hqreg Bootstrap coefficient boxplots

Finally, with hqreg, the predictors Ht^2 and Wt^2 have confidence intervals which do not include 0, therefore the null hypothesis that they are not informative to the response can be rejected. Furthermore, we can observe that the variability of the model is comparable to that of quantreg and significantly not as intense as with simple Quantile Regression and rqPen.

6.2 Performance Assessment

After conducting a comprehensive model analysis, our attention now shifts to performance assessment in order to identify the models that exhibit greater strength and reliability in predicting heart volumes. Starting off, we present the Cross Validation output for Least Squares aiming to identify the most promising methods. Subsequently, we dive into Quantile Regression. Here, we compare the Cross Validation results across different quantiles for each candidate model fitted with asymmetric loss functions. Additionally, we assess their errors in comparison to the top-performing Least Squares models, employing multiple approaches to quantify the errors. This thorough evaluation ensures that no method gains an unfair advantage over the others.

Least Squares Models

When it comes to conventional Least Squares, we assess the performance of several models, measuring the relative accuracy to determine which leads to more accurate predictions. With the use of Repeated Cross Validation as discussed in Chapter 5, we look into the standard Linear Regression model, along with stepwise selection (forward and backward) as well as model averaging, inspecting both ways to average the coefficients. Furthermore, we perform Nested Cross Validation to get an unbiased performance estimate of Ridge and Lasso Regression which we then compare with the aforementioned methods.

Aiming for a reduced variance around the estimated errors, we repeat the Cross Validation process 40 times, storing the error output for each iteration and then averaging them. We then further examine these outputs by creating boxplots in order to inspect the error distribution for each model, sorted from lowest to highest.

Firstly, we report the Repeated Cross Validation and Repeated Nested Cross Validation relative accuracy estimations, calculated with equation 8:

- Ridge: 0.01658
- Lasso: 0.01849
- Forward Selection: 0.02004
- Avg2: 0.02006
- Full Model: 0.02139
- Backward Elimination: 0.02265
- Avg1: 0.02833

Now, using the Cross Validation and Nested Cross Validation measurements for each model, the following boxplots are created:

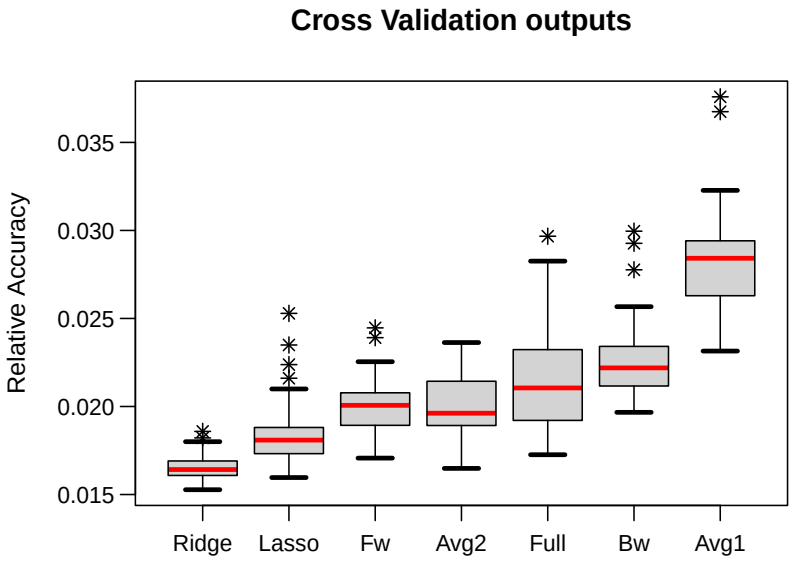


Figure 27: Repeated Cross Validation and Repeated Nested Cross Validation Boxplots

First of all, we can observe that for both penalized methods, Ridge and Lasso Regression, the medians and Q1,Q3 quartiles are smaller than those of the other models. Moreover, the interquantile range (IQR) for each, that is Q3-Q1, is again smaller indicating a less significant amount of variance. Between the penalized methods, Ridge has less variance and in general smaller errors, which in conjunction with the lowest Repeated Nested Cross Validation error provides strong evidence about its superiority among the other models. Lasso has larger IQR than Ridge, as well as more outliers and a greater distance between the minimum and maximum error. Overall there is more uncertainty compared to Ridge but in general it is a well-performing method with a small relative accuracy error estimate compared with all other models.

Forward selection and model averaging under the assumption that each coefficient appears in all models have a similar mean squared error to the full Linear Regression model, but with considerably less variance. Backward selection appears to be slightly worse, while averaging using only the non-zero coefficients of each model has the worst performance of all with high variance and the largest relative accuracy error.

Among all candidate Least Squares models, penalized methods stand out as more robust. Ridge resulted in a slightly better error distribution, with lower median, Q1 and Q3 as well as smaller IQR, indicating less variance. Since Ridge and Lasso are special cases of Elastic Net, we can perform a Permutation Test between them to verify that the error distribution can in fact be differentiated, despite coming from the same family of models. By doing so, we reinforce the interpretational value of Ridge's coefficients, resulting in

a well-established model, both performance-wise and regarding its inference.

For the Permutation Test, we shuffle the labels 10.000 times, using as a test statistic the absolute value of the difference of the means. The p-value we get for Ridge-Lasso is 0, confirming that the boxplots indeed come from a different distribution and from a statistical perspective, the Ridge model is unique.

Quantile Regression

We now report the results from Quantile Regression and penalized Quantile Regression model assessment. The baseline for our evaluation process is Ridge and Lasso Regression, which were the best Least Squares methods for predicting the heart volumes. Least Squares methods were assessed with the use of relative accuracy, which required predicting the response's logarithm. To be able to compare these methods with Quantile Regression, we take the exponent for each predicted value of Ridge and Lasso Regression so that predictions from both regression techniques will be on the same scale.

Firstly, we provide a general picture of how the models perform across different quantiles. For that, we estimated the errors from 0.05 up to 0.95 quantile with 0.05 increments. For each quantile, the estimation was performed using Repeated 9-fold Cross Validation for Quantile Regression and Repeated 9-fold Nested Cross Validation for penalized Quantile Regression to tune the hyperparameter λ, using 40 repetitions for each. We have stated that in order to be able to properly compare Least Squares and Quantile Regression methods, we must quantify the errors using multiple metrics. For that reason, mean squared error was considered along with the mean of loss functions $\rho_{0.5}$ (half of absolute loss) and $\rho_{0.75}$. We should note that the mean squared error was selected because its only difference with relative error in Least Squares stems from the logarithmic transformation of the response. By reversing this transformation through the exponent, we transition from $\sum_{i=1}^{n}(log\frac{y_{pred_i}}{y_i})^2$ to $\sum_{i=1}^{n}(y_i - y_{pred_i})^2$, which is the squared error.

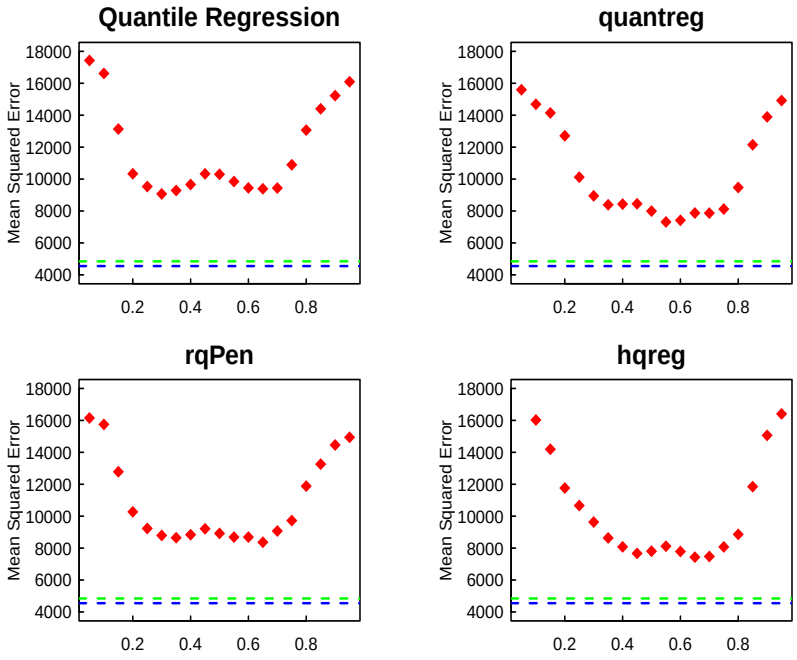


Figure 28: Error distribution across quantiles assessed with mean squared error. The blue dashed line corresponds to Ridge mean squared error, while the green is for Lasso.

As anticipated, both Ridge and Lasso Regression result in significantly lower mean squared error compared to Quantile Regression techniques. This is because Least Squares models are fitted with a symmetric quadratic loss function, so the coefficient estimates of the linear model are tailored to favor such functions, one of which is the mean squared error. Moreover, we observe that the lowest Quantile Regression and penalized Quantile Regression errors are found in the middle quantile range and not the tails. The reasoning for this situation is similar to before. Squared error evaluates predictions in a symmetric way, thus when it comes to more extreme values for τ where asymmetry is incorporated in the models, these will be under-evaluated.

It is also worth mentioning that even though the symmetric evaluation metric implies an increase in errors as we move away from $\tau = 0.5$, the panels indicate a different behaviour. Specifically, in the middle quantiles the errors remain relatively constant for all models, with an upward tendency appearing only when approaching the tails. This rather unexpected situation can be most likely attributed to the small sample size that we work with. When performing Cross Validation, each fold contains a subset of all our available data. In a sufficiently large dataset, each of these folds would be representative of the response's distribution. When the sample size is small however, it is less likely that each fold will capture this true distribution, so slightly inaccurate and counter-intuitive results such as this one may occur.

We now dive deeper into the median as well as the 75th percentile. Firstly, we report the mean squared errors resulting from Repeated Cross Validation and Repeated Nested Cross Validation methods for these quantiles, as well as for Ridge and Lasso Regression:

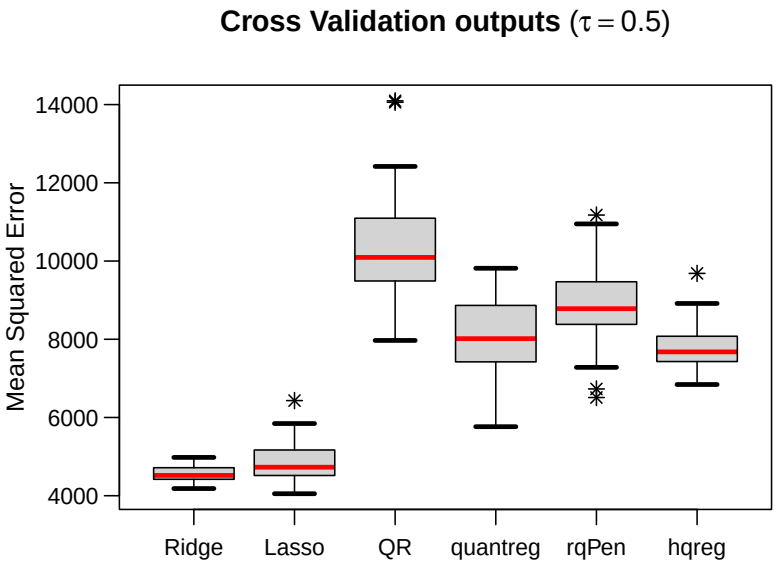
- Ridge Regression $\rightarrow$ MSE $= 4549$
- Lasso Regression $\rightarrow$ MSE $= 4846$

Mean Squared Error		
	$\boldsymbol{\tau = 0.5}$	$\boldsymbol{\tau = 0.75}$
Quantile Regression	10297	10894
quantreg	7993	8122
rqPen	8921	9723
hqreg	7807	8075

Table 4: Repeated Cross Validation and Repeated Nested Cross Validation mean squared error estimates for Quantile Regression and penalized Quantile Regression for the median and 75th percentile

The penalized Least Squares methods resulted in a significantly smaller mean squared error, compared to Quantile Regression and Penalized Quantile Regression. Moreover, for $\tau = 0.5$, the errors are slightly smaller than those for $\tau = 0.75$ where we are closer to the upper tail. It is interesting to observe that models from quantreg and hqreg have a better performance than the rqPen model, even though all three packages aim to minimize the same penalized Quantile Regression loss function. These findings are compatible with our previous model analysis, where we provided evidence about rqPen having a negligible penalty effect, thus being very similar to simple Quantile Regression.

To further illustrate the disparities between Least Squares and Quantile Regression when assessed with the mean squared error, we create boxplots from the Cross Validation and Nested Cross Validation error measurements:



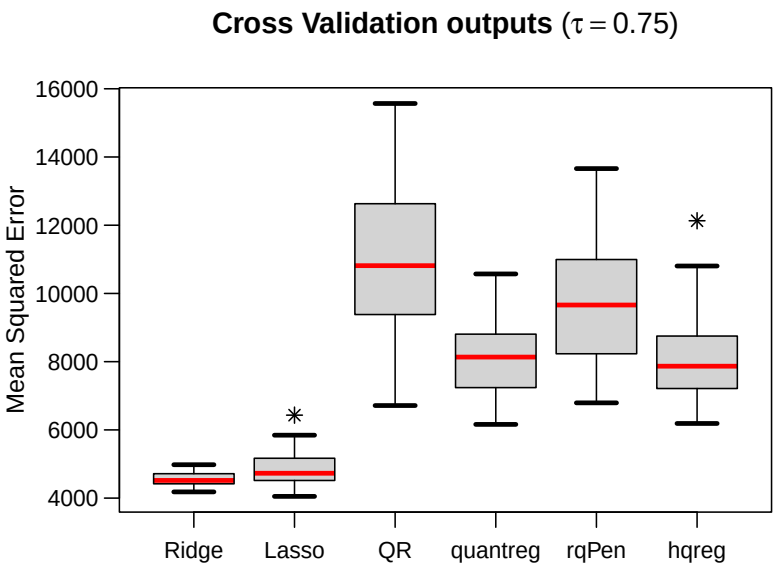


Figure 29: Repeated Cross Validation and Repeated Nested Cross Validation Boxplots for $\tau = 0.5$ and $\tau = 0.75$, where models are assessed with mean squared error

The boxplots indicate that Ridge and Lasso Regression have not only lower mean squared error, but also less variance. For $\tau = 0.75$, this situation becomes more apparent, with the Cross Validation outputs being even more variant than those for $\tau = 0.5$.

In order to demonstrate the effect our choice for quantifying the regression error has when comparing models fitted with different loss functions, we construct a similar figure as before, but now the error of all models was calculated using the mean of $\rho_{0.5}$, which is half of the absolute value:

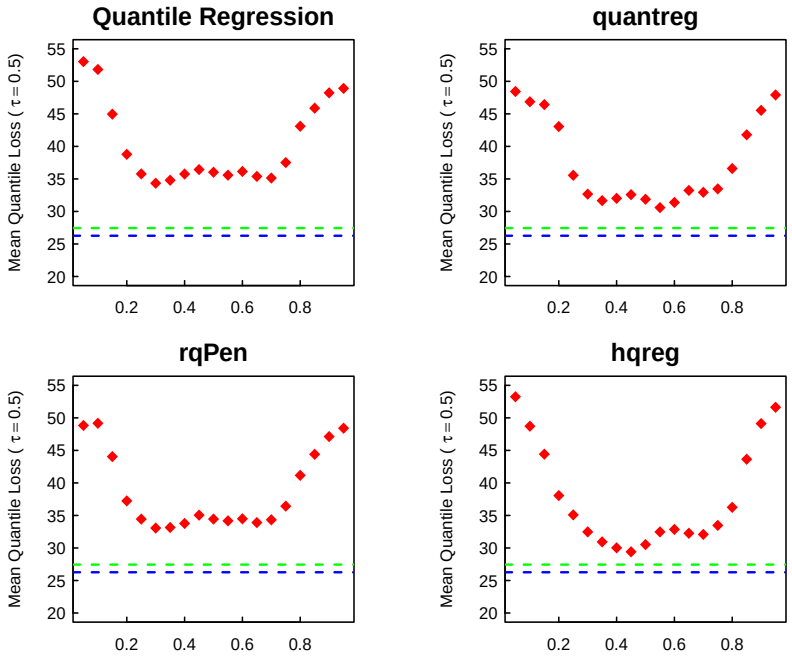


Figure 30: Error distribution across quantiles assessed with $\rho_{0.5}$

Contrary to initial expectations, Median Regression did not yield smaller errors compared to Ridge and Lasso Regression when evaluated with the absolute loss. We would expect Median Regression to perform better, since $\rho_{0.5}$ is also the function used for fitting this model while on the other hand, the coefficients of Ridge and Lasso are obtained by minimizing the squared error. Even though Ridge and Lasso outperformed Quantile Regression, the differences are smaller than those of mean squared error, with the two panels on the right indicating that the quantreg and hqreg models are very close to the Least Squares performance.

Same as with the mean squared error, we report the errors of each model, focusing on the median and 75th percentile for Quantile Regression:

- Ridge Regression $\rightarrow$ mean of $\rho_{0.5} = 26.26$
- Lasso Regression $\rightarrow$ mean of $\rho_{0.5} = 27.44$

Mean of $\rho_{0.5}$		
	$\boldsymbol{\tau = 0.5}$	$\boldsymbol{\tau = 0.75}$
Quantile Regression	36.03	37.51
quantreg	31.87	33.47
rqPen	34.44	36.43
hqreg	30.52	33.47

Table 5: Repeated Cross Validation and Repeated Nested Cross Validation mean quantile loss ($\tau = 0.5$) error estimates for Quantile Regression and penalized Quantile Regression for the median and 75th percentile

The table suggests a similar picture with mean squared error, with Ridge and Lasso having the lowest errors and models from quantreg and hqreg packages outperforming those of simple Quantile Regression and rqPen.

We now provide the error boxplots for the median and our quantile of interest, which is the 75th percentile:

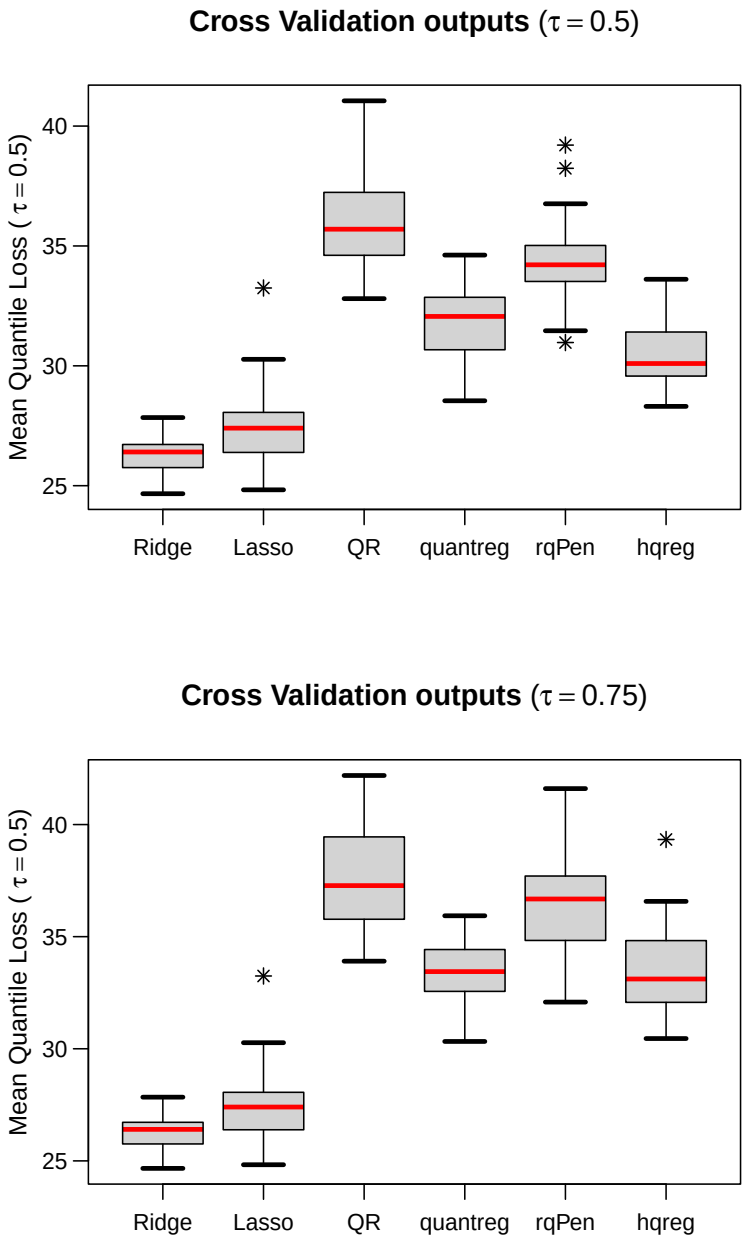


Figure 31: Repeated Cross Validation and Repeated Nested Cross Validation Boxplots for $\tau = 0.5$ and $\tau = 0.75$, where models are assessed with $\rho_{0.5}$

Looking at the boxplots of quantreg and hqreg models both for $\tau = 0.75$, we observe that their error distributions are very similar, regarding not only their error magnitude, but also the variance. In order to further inspect this situation, we perform a Permutation Test between the error measurements for quantreg and hqreg models to assess whether their errors come from the same distribution. For the Permutation Test, we shuffle the labels 10.000 times, using as a test statistic the absolute value of the difference of the means. The resulting p-value is 0.98, thus there not enough evidence to reject the null hypothesis that the errors come from the same distribution.

Finally, we present the errors of Ridge and Lasso Regression along with Quantile Regression and penalized Quantile Regression assessed by $\rho_{0.75}$:

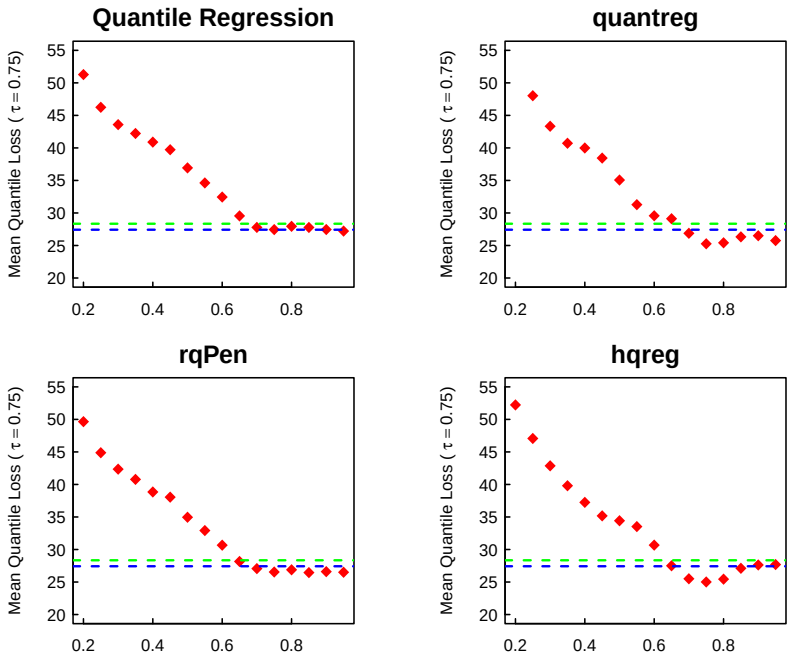


Figure 32: Error distribution across quantiles assessed with $\rho_{0.75}$

In contrast to previous error metrics, now we account for the importance of handling under-predictions and negative errors are penalized three times more than positive ones. The first thing to notice is that now higher quantiles perform as well as Ridge and Lasso Regression, with some surpassing them. Specifically, multiple quantiles from quantreg and hqreg penalized Quantile Regression models result in an error smaller than Least Squares. On the other hand, the upper tail of simple Quantile Regression leads to almost equal errors as Ridge and Lasso Regression, while rqPen models barely surpass Ridge after the 70th percentile. This evidence suggests that quantreg and hqreg lead to slightly

better results in comparison to rqPen and unpenalized Quantile Regression.

We further inspect our findings by reporting the Cross Validation errors and creating the corresponding boxplots as before:

- Ridge Regression $\rightarrow$ mean of $\rho_{0.75} = 27.42$
- Lasso Regression $\rightarrow$ mean of $\rho_{0.75} = 28.34$

Mean of $\rho_{0.75}$		
	$\boldsymbol{\tau = 0.5}$	$\boldsymbol{\tau = 0.75}$
Quantile Regression	36.93	27.44
quantreg	35.05	25.26
rqPen	34.96	26.54
hqreg	34.42	25

Table 6: Repeated Cross Validation and Repeated Nested Cross Validation mean quantile loss ($\tau = 0.75$) error estimates for Quantile Regression and penalized Quantile Regression for the median and 75th percentile

We can see that Ridge and Lasso Regression greatly outperform Median Regression and penalized Median Regression when evaluated with the 75th percentile quantile loss, even though both methods are fitted with symmetric loss functions. On the other hand, for $\tau = 0.75$, Quantile Regression and penalized Quantile Regression are the methods with the smallest error estimates, slightly below Ridge and Lasso Regression. We should point out that even though evaluation through $\rho_{0.75}$ benefits Quantile Regression of this particular quantile, the differences from Least Squares are quite small, while when using symmetric error metrics, Ridge and Lasso Regression yielded significantly better results.

Regarding Quantile Regression methods for $\tau = 0.75$, the penalized models from quantreg and hqreg have better performance, by a small margin however. Since we are interested in this specific quantile for the purpose of limiting under-predictions, we will inspect the error distribution for each model in more detail through the corresponding boxplots:

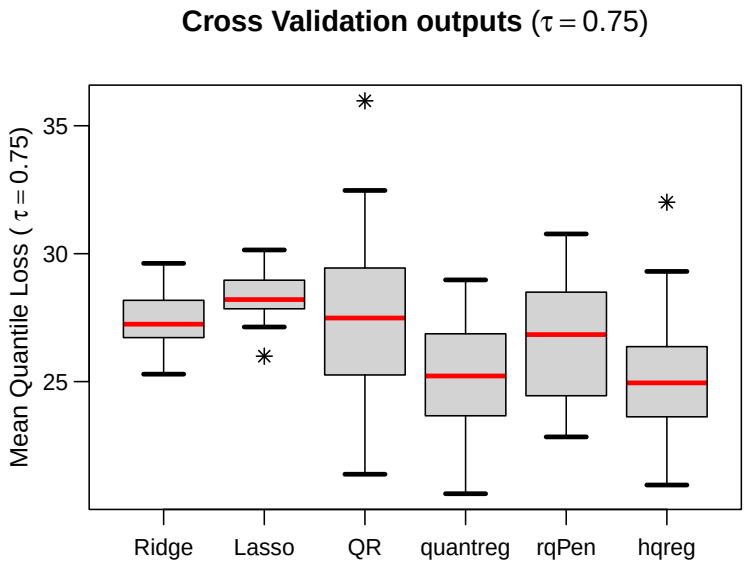


Figure 33: Repeated Cross Validation and Repeated Nested Cross Validation Boxplots for $\tau = 0.75$, where models are assessed with $\rho_{0.75}$

It is clear that Quantile Regression methods have higher variance compared to Ridge and Lasso Regression. The boxplots of quantreg and hqreg have smaller medians and quartiles Q1,Q3 compared to standard Quantile Regression and rqPen as well as smaller IQR, indicating less variance. In fact, the error distributions of quantreg and hqreg appear to be almost identical. In order to inspect this situation, a Permutation Test is performed between the two error distributions, using 10.000 shuffles and again the absolute difference of the means as a test statistic. The resulting p-value is 0.58, therefore there is indeed not enough evidence for rejecting the null hypothesis that the errors come from the same distribution.

7. Conclusions

In this thesis, we focused on the construction of linear predictive models with the use of asymmetric loss functions. The motivating application was the prediction of heart volumes of pediatric patients who are candidates for heart transplantation (also applicable for heart donors). Our purpose was to construct a predictive model for estimating the heart volume of the patient based on simple somatometric measurements that can be easily obtained. In this context, we were also required to account for the asymmetry in the importance of errors, since under-predicting the size of the heart has been found to be potentially more dangerous for the recipient than a over-prediction of similar magnitude.

The Least Squares statistical methods examined herein included simple Linear Regression, stepwise feature selection, model averaging and penalized estimation, specifically Ridge and Lasso Regression. We found that both penalized methods performed better in terms of generalization capability relative to conventional Least Squares methods, with Ridge overpowering Lasso. This difference between Ridge and Lasso was anticipated because while Lasso Regression is generally known to be more suitable in situations where many predictors are available, but few are relevant to the response, this is not the case with our data. Specifically, by performing VIF stepwise feature selection, we filtered out the uninformative variables, expecting the majority of the rest to be important for predicting the response. This is exactly the setting where Ridge may lead to superior results, which was also the case for our application.

The main part of our research involved predictive modeling with the use of Quantile Regression. We explored both the simple Quantile Regression model as well as penalized versions, where a Lasso-type of penalty was considered. For penalized Quantile Regression, we discussed the use of 3 different *R* packages (quantreg, rqPen, hqreg) to acquire the coefficient estimates, along with their algorithms for solving the optimization problem. In order to account for the importance of constraining under-predictions, we focused on the 0.75 quantile, where under-predictions are penalized three times more than over-predictions. The resulting quality of predictions of Quantile Regression was then compared to the best-performing Least Squares methods, Ridge and Lasso Regression.

Our findings suggest that among Quantile Regression methods, penalized Quantile Regression performed better with the quantreg and hqreg implementations. However, we demonstrated that the optimization process quantreg follows is statistically inaccurate and should not be trusted. Hence, we conclude that penalized Quantile Regression with hqreg can potentially be a strong tool for providing experts valuable insights when making transplantation decisions. This usefulness is not only determined by the predictive value at the specific quantile, but also through the utilization of prediction intervals which provide a more complete view of the range of predicted heart volumes. Furthermore, inspecting the coefficients across a sequence of quantiles from the lower tail to the upper can assist the doctor in making more informed decisions by inspecting the relationship of the predictors with the heart volumes.

Regarding the comparison with Least Squares, the assessment was performed using multiple error calculation metrics to account for the model-specific symmetries (or asymmetries). Specifically, the errors were calculated using the symmetric criteria of Mean Squared Error and mean of $\rho_{0.5}$, which is half of the absolute value function, while we also considered an asymmetric evaluation based on $\rho_{0.75}$, which favours models that do not under-predict as much. We found that Ridge and Lasso Regression had better performance when evaluated with each of the symmetric criteria, while Quantile Regression methods at higher quantiles (especially around 0.75) was determined as better by $\rho_{0.75}$ (however, by a small margin). A significant finding from the numerical experiments presented in the previous sections, which is not in accordance with prior expectations, suggests that the cost of adopting a symmetric loss to estimate a model when the evaluation criterion is asymmetric is relatively small (at least for quantiles up to 0.75).

Bibliography

[1] Plasencia JD, Kamarianakis Y, Ryan JR, et al. Alternative methods for virtual heart transplant— Size matching for pediatric heart transplantation with and without donor medical images available. Pediatr Transplant. 2018;22:e13290. https://doi.org/10.1111/petr.13290

[2] Kawabori, M., Critsinelis, A. C., Patel, S., Nordan, T., Thayer, K. L., Chen, F. Y., & Couper, G. S. (2022). Total ventricular mass oversizing +50% or greater was a predictor of worse 1-year survival after heart transplantation. In Journal of Thoracic and Cardiovascular Surgery. Elsevier Inc. https://doi.org/10.1016/j.jtcvs.2022.03.040

[3] Reed, R. M., Netzer, G., Hunsicker, L., Mitchell, B. D., Rajagopal, K., Scharf, S., & Eberlein, M. (2014). Cardiac Size and Sex-Matching in Heart Transplantation: Size Matters in Matters of Sex and the Heart. JACC: Heart Failure, 2(1), 73–83. https://doi.org/10.1016/j.jchf.2013.09.005

[4] Tamisier, D., Vouhe, P., Le Bidois, J., Mauriat, P., Khoury, W., & Leca, F. (1996). Donor-recipient size matching in pediatric heart transplantation: A word of caution about small grafts. Journal of Heart and Lung Transplantation, 15(2), 190–195.

[5] Camarda J, Saudek D, Tweddell J, et al. MRI validated echocardiographic technique to measure total cardiac volume: a tool for donor–recipient size matching in pediatric heart transplantation. Pediatr Transplant. 2013;17(3):300-306

[6] Barkoff, L. M., Maeda, K., Rosenthal, D. N., Zhang, Y., Hollander, S. A., Dykes, J. C., … Almond, C. S. (2018). Center Variation in Listing Parameters Among US Pediatric Heart Transplant Programs. The Journal of Heart and Lung Transplantation, 37(4), S396. https://doi.org/10.1016/j.healun.2018.01.1021

[7] Patel, N. D., Weiss, E. S., Nwakanma, L. U., Russell, S. D., Baumgartner, W. A., Shah, A. S., & Conte, J. V. (2008). Impact of donor-to-recipient weight ratio on survival after heart transplantation: analysis of the United Network for Organ Sharing Database. Circulation, 118(14 Suppl). https://doi.org/10.1161/CIRCULATIONAHA.107.756866

[8] Riggs, K. W., Giannini, C. M., Szugye, N., Woods, J., Chin, C., Moore, R. A., … Zafar, F. (2019). Time for evidence-based, standardized donor size matching for

pediatric heart transplantation. Journal of Thoracic and Cardiovascular Surgery, 158(6), 1652-1660.e4. https://doi.org/10.1016/j.jtcvs.2019.06.037

[9] J Neter, MH Kutner, CJ Nachtsheim, W. W. (1996). Applied Linear Statistical Models. Fourth Edition. Journal of Education, 36(3), 59–60.

[10] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

[11] Redlarski, G. et al. Body surface area formulae: an alarming ambiguity. Sci. Rep. 6, 27966; doi: 10.1038/srep27966 (2016).

[12] Tofallis, Chris, A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation (July 2014). Journal of the Operational Research Society (2015) 66, 1352–1362, Available at SSRN: https://ssrn.com/abstract=2635088

[13] Kitchenham, BA, Pickard, L and MacDonell, S. (2001). What accuracy statistics really measure. IEE Proceedings – Software, 148(3):81–85, June 2001.

[14] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

[15] Kenneth P. Burnham David R. Anderson. Model Selection and. Multimodel Inference. A Practical Information-Theoretic Approach. Second Edition.

[16] Hurvich, C.M., and Tsai, C-L. (1989). Regression and time series model selection in small samples. Biometrika 76, 297–307.

[17] Burnham KP, Anderson DR, Huyvaert KP (2010) AICc model selection in Ecological and behavioral science: some background, observations, and comparisons. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1029-6

[18] Mundry, R. (2011, January 1). Issues in information theory-based statistical inference-a commentary from a frequentist's perspective. Behavioral Ecology and Sociobiology. Springer Verlag. https://doi.org/10.1007/s00265-010-1040-y

[19] Eberhardt LL (2003) What should we do about hypothesis testing? J Wildl Manage 67:241–247

[20] Symonds MRE, Moussalli A (2010) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1037-6

[21] Derksen S, Keselman HJ (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 45:265–282

[22] Hegyi, G., & Garamszegi, L. Z. (2011, January 1). Using information theory as a substitute for stepwise regression in ecology and behavior. Behavioral Ecology and Sociobiology. Springer Verlag. https://doi.org/10.1007/s00265-010-1036-7

[23] Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2), 301–320. http://www.jstor.org/stable/3647580

[24] Wright, S. J. (2015). Coordinate descent algorithms. Mathematical Programming, 151(1), 3–34. https://doi.org/10.1007/s10107-015-0892-3

[25] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22. https://doi.org/10.18637/jss.v033.i01

[26] Kooij, A. J. van der. (2007, June 27). Prediction accuracy and stability of regression with optimal scaling transformations. Leiden. Retrieved from https://hdl.handle.net/1887/12096

[27] Koenker, R. (2005). Quantile regression. Quantile Regression (pp. 1–349). Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

[28] Roberts, F. D. K., & Barrodale, I. (1973). An Improved Algorithm for Discrete L1 Linear Approximation. SIAM Journal on Numerical Analysis, 10(5), 839–848.

[29] Portnoy, S., & Koenker, R. (1997). The gaussian hare and the laplacian tortoise: Computability of squared-error versus absolute-error estimators. Statistical Science, 12(4), 279–296. https://doi.org/10.1214/ss/1030037960

[30] Sherwood, B., & Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. Annals of Statistics, 44(1), 288–317. https://doi.org/10.1214/15-AOS1367

[31] Yi, C. and Huang, J. (2016) Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression, https://arxiv.org/abs/1509.02957 Journal of Computational and Graphical Statistics, accepted in Nov 2016

[32] Belloni, A. and V. Chernozhukov. (2011) l1-penalized quantile regression in high-dimensional sparse models. Annals of Statistics, 39 82 - 130

[33] DeGroot, M. H., Schervish, M. J. (2002). Probability and Statistics. Addison Wesley.

[34] Efron, B., & Tibshirani, R. J. (1994). An Introduction to the Bootstrap. An Introduction to the Bootstrap. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

[35] Wilcox, R. R. (2010). Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy (pp. 1–249). Springer New York. https://doi.org/10.1007/978-1-4419-5525-8