



Course of

SUPERVISOR

CO-SUPERVISOR

CANDIDATE

Academic Year

Table of contents

Part 1. Introduction	2
1.1 Topic, research problems and delimitations	2
1.2. Structure of the paper	3
Part 2. Quantile regression models	5
2.1 Motivation of quantile approach	5
2.2 Methodology of quantile regression	6
2.3 Role of quantile regression in empirical research	6
Part 3. Quantile models based on Bayes approach	9
Part. 4. Experimental comparison of the models	11
4.1 Methodology of experiments	11
4.2 Preliminary analysis of the Boston Housing data	12
Part 5. Results of experiments	16
5.1 Results on generated data	16
5.2 Results on real data	19
5.3 Discussion of the results	21
Part 6. Conclusion	22
Bibliography	23
Appendix 1. Implementation of the experiment in R programming environment.	25
Appendix 2. Description of Boston Housing data	28
Summary	29

Part 1. Introduction

1.1 Topic, research problems and delimitations

From a practical point of view, scientists are interested in models for making high quality forecasts. In other words, the quality of the forecast often has key importance in a model. This fact encourages researchers to look for new models for prediction. From 70s years of XX century, the using of quantile regressions allowed to explore whole distribution of the research data more thoroughly and obtain more accurate forecasts for different quantiles. Further transition from frequentists to Bayesian approach in regressions made it possible to improve the prediction properties of the model because Bayes models allow to control the prior distributions of estimates. Theoretical papers provide scientists with clear motivation to choose model based on preliminary analysis of the data and describe advantages of quantile models and Bayes regression models. Nevertheless, in every particular case researchers have never known beforehand what model would perform better forecast on the data.

The aim of the thesis is experimentally identify the best forecasts in terms of RMSE between quantile regression models based on traditional and Bayes approach.

It should be emphasised that it is an *experimental* work. It means that research is based on experiments with generated samples from 6 types of data distributions (normal with small variance, normal with large variance, geometric, gamma, log-normal and student distribution). These types of distribution were chosen for analysis because they are the most widespread ones in economic field. For instance, prices or rent payments (as in Baba and Shimizu (2022)) may often be log-normally distributed and shocks in economy may be normally distributed. What is more, it is known that quality of forecast may vary with distribution of the data. Several different distributions are considered in this paper in order to understand how final results and conclusions may change when distribution type of data is changed.

This paper considers three following **research questions**:

1. Does the Bayesian quantile regression model make forecasts with lower RMSE than the quantile regression in the classical formulation?
2. How change results based on generated sample from different types of distributions?
3. How do results on generated data correlate to estimates on real data?

Several **hypotheses** follow from the analysis of literature and research questions:

1. Share of best forecasts based on OLS regression are larger for samples with normal distribution.
2. Share of best forecasts based on Bayes models are the largest one in general case.
3. Shares of best forecast for OLS, quantile model and Bayes quantile model positively correlates between distributions of similar shape.

The **delimitations** of the paper are as follows. Firstly, this research is based only on linear forecasts. There are also examples of nonlinear quantiles regression models in the literature, however, research questions of this paper can be explored on linear models that is computationally more convenient for the investigation. Secondly, there are various kinds of metrics to assess the predictive power of models. In this study, RMSE is used as one of the basic and universal measure. Thirdly, to estimate the quality of forecast, the predicted values are analysed. The real and estimated coefficient of regression line are not compared, because it has already discussed in literature, for instance, in Alhamzawi and Yu (2013). Finally, as real data, Boston Housing data is used - it is widely known and explored in literature. The original idea of this research is connected to the method of comparing models, but not in the data itself. Going beyond these limitations can be a continuation of this study and makes it more valuable, but now it remains outside the scope of this study.

1.2. Structure of the paper

The **structure** of the thesis is as follows. In the second part the advantages of quantile regressions are discussed comparing to OLS regressions. Firstly, quantile regression allows to better explore the tales of the distribution. Secondly, such models are less sensitive to the influence if outliers in the data. Finally, quantile regressions are quite widespread and used in large number of research in different field from medicine to economics and time series. Key areas of implementing of quantile regressions are summarised in the table at the end of the second part.

In the third part the characteristics of Bayes quantile models are considered in comparison with classical quantile regression and OLS models. The methodology and key points in estimating of the Bayes models are briefly discussed. Key advantages of Bayes approach are summarised at the end of the third part and some crucial characteristics of model estimation process in our experiments are provided.

The foundation and the methodology of experiments are presented in the fourth part. Experiment is based on the sample of size 100 from 6 different distribution types. Then these samples are splitting on train and test sub-samples. For each sample three types of models (OLS, quantile regression and Bayes quantile regression) are estimating. Then forecasts are built on the test sub-sample and RMSE for each forecast are calculating. All in all, 500 experiment series were performed for each type of six generated distribution to notice tendencies beyond RMSE estimates. OLS model is used like baseline that quantile regression and Bayes quantile regression models are compared with. Furthermore, the same methodology is implented on Boston Housing data that contain prices and characteristics of 506 houses in living area of Boston. It is quite popular in the literature and was firstly published in 1978. The preliminary analysis of the data, selection of regressors and regression equation are also discussed in the fourth part. The difference of experiments on generated and real data is the fact that the models on real data are built based on economic grounds and analysis of economic relationships between variables.

The fifth part presents results of experiments on both generated and real data. The best model in each experiment is identified based on RMSE. It is revealed that share cases when each model is better is supposed to be stable as the number of experiments grow. 95% CI are calculated for shares when OLS, quantile regression and Bayes quantile regression are first-, second- and third-best model in terms of RMSE. The tendencies beyond the estimates of RMSE are summarised at the end of the fifth part.

Finally, the conclusion emphasises key finding of this research. Firstly, the short discussion considers confirmation or rejected the hypotheses mentioned above. Secondly, the further potential steps expanding this study are briefly discussed.

Part 2. Quantile regression models

This part is structured as follows. First, the motivation for the transition to quantile regressions as a new class of models is discussed. Secondly, the advantages of this approach in practice and fields of implementation of quantile regression are considered. Thirdly, methodological formulation of quantile regression problem is described. Finally, the role of quantile regressions in this study is described as conclusion to this part.

2.1 Motivation of quantile approach

Quantile regressions became widespread in the scientific literature after an article by Koenker and Knight (1978) in the *Econometrica* journal¹, published by The Econometric Society². The authors showed that the same general approach that used in conventional linear regression can be implemented for different quantiles of distribution of independent variable. The term “regression quantiles” was proposed by them and, that is more importantly, they suggested new estimator of such type of regression. It had an incredible impact on research in cases, where errors are distributed non-normally and assumption of Gauss-Markov theorem are violated. In cases, where the errors are distributed normally, the result of evaluating the quantile regression models turns out to be similar to the usual linear regression. This part is structured as follows. First, the motivation for the transition to quantile regressions as a new class of models will be discussed. Secondly, it will be demonstrated how important this approach is in practice and in which areas we can be most valuable. Thirdly, methodological features of quantile regressions will be described. Finally, the role of quantile regressions in this study is described.

The *motivation* for the introduction of quantile regressions is based on the following characteristics. First of all, the usual regressions (least squares) do not work well with outliers. What is more, both estimates and mean forecast in OLS is sensitive to outliers. Outliers can significantly affect the results of the model, so researchers often simply exclude outliers at the preliminary analysis stage. Nevertheless, if outliers would not bias the model results, we would prefer to leave them in the data sample. In other words, if it supposed to be a data generation process behind the observed data, then outliers are part of this process. Moreover, the problem of outliers lead scientists to make a forecast based on the median (not on mean prognoze), as in the work of Gannoun et al. (2003). The median approach can be generalized to different quantiles. Secondly, the assumption of the normality of errors as a consequence of the central limit theorem and the law of large numbers often looks too strict on real data. Errors as a random factor are not observable. Consequently, the requirements of the Gauss-Markov theorem turn out to be difficult to implement in practice. Thirdly, the quantile approach shows itself better in models when the data generation process behind the sample may have

¹ <https://www.jstor.org/journal/econometrica> - Econometrica journal page on JSTOR

² <https://www.jstor.org/publisher/econosoc> - The Econometric society page on JSTOR

an unknown distribution, especially non-normal. What is more, quantile regression can be formulated with various types of constraints (for more details see, for example, Koenker and Ng (2005)). Nevertheless, quantile regressions are based on a similar minimization problem, as OLS. Quantile regressions are also sensitive to the number of observations, and if there are not enough observations, they revealed poor results. Detailed example of comparison of estimates in OLS and quantile regression model on infant birthweight data is provided in Hallock and Koenker (2001). All the advantages of quantile regressions described above are used in a variety of empirical studies, which we will discuss further.

2.2 Methodology of quantile regression

The quantile regression model is formalised as follows. The methodology of estimating is briefly presented in this paper similar to Benoit and Van den Poel (2017).

The main logic is similar to classic regression. The linear model is:

$$y_i = x_i^\tau \beta + \epsilon_i,$$

where ϵ_i is error term and τ is a given quantile and β is vector of coefficients. The solution for β is as following:

$$\hat{\beta}_\tau = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \rho_\tau(y_i - x_i^\tau \beta),$$

where $0 < \rho_\tau < 1$ corresponding to any particular quantile. In general, using this method, $\hat{\beta}_\tau$ for any quantile can be estimated. For instance, median forecast can be calculated using $\hat{\beta}_{0.5}$. Similarly, predictions for a set of quantiles can be evaluated. Symmetric sets of quantiles can be evaluated to receive mean forecast based on quantile regression model like simple mean of estimated quantiles forecasts. In our experiments set of quantiles [0.01, 0.99] is used in order to compare results with OLS model that created mean forecast by default. In literature (for example, Benoit and Van den Poel (2017)) quality of median and mean forecasts are compared, but this question is beyond the scope of this research.

In this research, quantile regression models are estimated automatically using package “quantreg” in R programming environment. The programming code is attached in Appendix 1.

2.3 Role of quantile regression in empirical research

Quantile regression as a modification of OLS, due to its advantages, attracted a number of researchers in various scientific fields from biology and medicine to education.

Cole and Green (1992) used data on triceps skinfold in Gambian girls and women, and body weight in U.S.A. girls. The authors extend the quantile regression approach by the example of nonlinear models of the dependence of triceps skinfold in Gambian girls and women, and body weight in U.S.A. girls on age. Their study uses reference centile curves as a way to track anomalies in the data. Such curves usually represent the

dependence of the parameter of interest (for example, weight or muscle mass, body fat percentage, etc.) depending on age. In this approach, it is impossible to throw out outliers, because the task is to detect them.

Koenker and Selling (2001) use quantile regressions to analyse the survival of medflies. The authors use a two-sample treatment-control model. The indicator that the size of the pupa had an important influence for survival for the smallest 10% and had a negative influence at the upper end of the distribution. This is a very clear example of how the same variable can affect differently depending on which part of the distribution of the variable of interest scientists are considering.

Quantile regressions also turn out to be convenient on time series data. In Gannoun et al. (2003) it is shown that quantile regression attracts the attention of researchers of data that are distributed asymmetrically, for example, income or real estate price. The study is based on nonlinear autoregressive models based on time series. An illustrative example of the use of quantile regressions on time series data is also contained in Koenker and Xiao (2006). The paper also examines abnormally distributed data – the American unemployment rate and U.S. gasoline prices. The authors call the quantile model on time data the quantile autoregression model (QAR). The authors believe that the QAR model has the potential to become a separate area of time series research. The authors are conducting a Monte Carlo experiment to explore the possibilities of QAR and focus on the model with iid errors.

There are many examples in the literature of the application of quantile regressions to the analysis of modern crises of recent years. For instance, the work of Mahdi, E., & Al-Abdulla, A. (2022) explores the relationship between bitcoin and gold prices and news indices that characterize the perception of crises, such as Panic, Sentiment, Infodemic, and Media Coverage (talking about the RavenPack news-based index³). The authors discovered that despite the similar nature - both gold and bitcoin are used to hedge risks, the impact of news indices on demand and prices of commodities is asymmetric. The authors reveal the disadvantages of quantile regressions. They turn out to be less effective than the quantile-on-quantile approach. The reason is that positive and negative shocks affect the market with different intensity.

Examples of implementation of quantile regressions in economics and finance can be found in Yu, Lu and Stander (2003). Examples of other scientific fields than economics and finance are listed in the table 1. The book of Fitzenberger, Koenker and Machado (2001) can be used as a textbook on implementation of quantile models. All in all, empirical research reveal that empirically every research problem and data set require its own approach. With all the advantages of quantile regressions, it is impossible to say in advance which model will work better on the data.

Table 1. Fields of application of quantile regressions.

Science field	Pieces of research	Findings concerning quantile regression
---------------	--------------------	---

³ <https://www.ravenpack.com/> - more information about RavenPack company turning news, social media, transcripts, filings, and other texts in valuable insights for business.

Medicine	Cole and Green (1992)	Outliers are significant in medical, biological or survival analysis. They cannot be filtered from the data as in other field because the influence of such outliers is often the main research interest in such papers. Quantile regression gives scientists opportunity nor discard the outliers and explore their influence on the tails of distributions.
Survival analysis	Koenker and Geling (2001)	
Ecological studies/biology	Pandey and Nguyen (1999) Cade and Guo (2000) Allen et al. (2001) McClain and Rex (2001) Knight and Ackerly (2002) Brown and Peet (2003)	
Time series	Gannoun et al. (2003) Lipsitz et al. (1997) Koenker and Xiao (2006)	Quantile regression allows to determine (non)stationarity of the data generating process through analysis of asymmetric dynamics in time series, for instance, prices or unemployment. Unit root condition is implementing In other words, quantile regression helps to detect unit root because unit root condition can be fulfilled for some quantiles and rejected for others.
Education	Buchinsky (1998) Cade and Noon (2003)	Developing an education development strategy, it is necessary to take into account that the same measures may have different effects for the most successful, average, or less successful pupils. Quantile regression can strengthen policy conclusions through investigation of these effects.
Impact of COVID-19 Crisis	Mahdi and Al-Abdulla (2022)	Impact of COVID-19 was different for various quantiles of dependent variable (it may be both survival rate and returns of the assets). Quantile regression allows to estimate this impact that sometimes are opposite for highest 10% and lowest 10% of the distribution.

Source: literature review of the author

Part 3. Quantile models based on Bayes approach

In literature, Bayes approach is a competitor to the frequentists' methods as the first one allows to control the prior distribution of parameters of the model. Bayes approach in regression modelling came from the classical Bayes formula:

$$P(\theta|D) = \frac{P(D|\theta)*P(\theta)}{P(D)},$$

where $P(\theta)$ is a prior distribution of a parameter of the model – hypothetic or based on preliminary analysis of data (in case of regression model it may be a prior distribution of a coefficient);

$P(\theta|D)$ — a posterior probability of the parameter of the model – it is what the model should estimate;

$P(D|\theta)$ — likelihood;

$P(D)$ — the total probability of occurrence of the data (evidence).

The example of regression model in Bayes formulation is provided below.

The typical regression model can be formulated based on Bayes formula:

$$y_i|x_i = a + b x_i + \varepsilon_i,$$

where ε_i – error term, i.i.d., $\varepsilon_i \sim N(0, \sigma^2)$ and data sample consist of independent pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$.

$$y_i|x_i, a, b, \sigma^2 \sim N(a + b x_i, \sigma^2)$$

For convenience, the same formulation of the model is written down in a more convenient form (in particular, $\tau = \frac{1}{\sigma^2}$).

In this case likelihood is formulated as follows:

$$y_i|\mu_i, \tau \sim N(\mu_i, \tau)$$

and priors may be formulated as

$$a \sim N(\mu_a, \tau_a)$$

$$b \sim N(\mu_b, \tau_b)$$

$$\tau \sim \text{Gamma}(a_\tau, b_\tau)$$

Regression coefficients a and b can take both positive and negative values, so the normal distribution is chosen as a prior for them, the variation estimated in terms of τ takes values greater than 0, so the Gamma distribution is chosen as a prior for it. Defined all the parameters as above, we have formulated the Bayesian linear regression problem. It is easy to use analogous logic to multiple quantiles and receive Bayes quantile model.

If we use the matrix form for simplicity, then Bayes quantile regression may look in the following form:

$$Q_{y_i}(x|p) = x_i' \beta_p,$$

where $0 < p < 1$, and $Q_{y_i}(\cdot) = F_{y_i}^{-1}(\cdot)$ – inverse of the cumulative distribution function of a variable y_i , on condition of x_i .

The quality of the model varies according to type of prior used. From on hand, by default, the non-informative prior can be used. It is normal prior with zero mean and large variance. From the other hand, according to Alhamzawi and Yu (2013), the best forecast in Bayes quantile model can be received using not general prior but special prior for every quantile. The authors created a series of experiments similar to experiments in this research. Nevertheless, the use coefficients comparison for evaluating quality of the model (as the distribution types was controlled in their research, they compared real coefficients with estimates in particular model). On the contrary, in our research the quality of the model is estimated based on forecast quality. In experiments in this thesis the non-informative priors are used. As it reveals in the results section, Bayes model can compete with OLS and quantile model even without formulating of informative priors. Also, sometimes scientists' knowledge about prior distribution of the real model is not enough or absent at all.

Bayesian quantile regression methods are widely discussed in Lancaster and Jae Jun (2010). It is important to it is important to take into account several parameters. Firstly, since distributions are used in the model and the final answer is a posteriori distribution, and not a point estimate as in the standard model. Gibbs sampler is used to find a solution (for details see Alhamzawi and Yu (2013)). Finding a solution (posterior distribution) takes place through several iterations. The number of iterations has crucial influence and is usually quite large. In experiments in this research 1000 number of iterations is used. Since the algorithm does not converge to the correct solution immediately, part of the first iterations must be discarded (this part of the series is called “burning-in”). Usually, a quarter or half of the iterations are discarded. In these experiments 500 iterations are “burning-in”.

In conclusion, Bayes approach has several advantages comparing to classical frequentist approach. Firstly, it can effectively work on small data samples. Secondly, Bayes models can be implemented based on different distribution. Bayes models are less sensitive to outliers and the requirement of normality compared with classical regressions.

Estimation of Bayes quantile regression is available in R programming environment with packages “*bqr*” (see Alhamzawi and Ali (2020)) and “*bayesQR*” (see Benoit and Van den Poel (2017) for details). The last one was used during experiments in this research.

Part. 4. Experimental comparison of the models

This part presents methodological approach. Firstly, it describes how experiments were organised step by step on generated data sample from different distributions. Secondly, it describes the same experiment on real data with preliminary analysis of the Boston Housing data.

4.1 Methodology of experiments

The series of experiments were based on sample generated from 6 types of data distributions (normal with small variance, normal with large variance, geometric, gamma, log-normal and student distribution). Table 2 reveals how distributions were organised. In the first step, variable x was set as sample of size 100 out of uniform distribution in the range from 1 to 10. In the second step, variables z_1, z_2, \dots, z_6 were set as samples of size 100 from 6 types of distributions. In the third step, variables y_1, y_2, \dots, y_6 was set as $y_i = 2 + 5 * x + z_i$, where $i = \overline{1,6}$. Finally, 6 data frames of x and y_i were created.

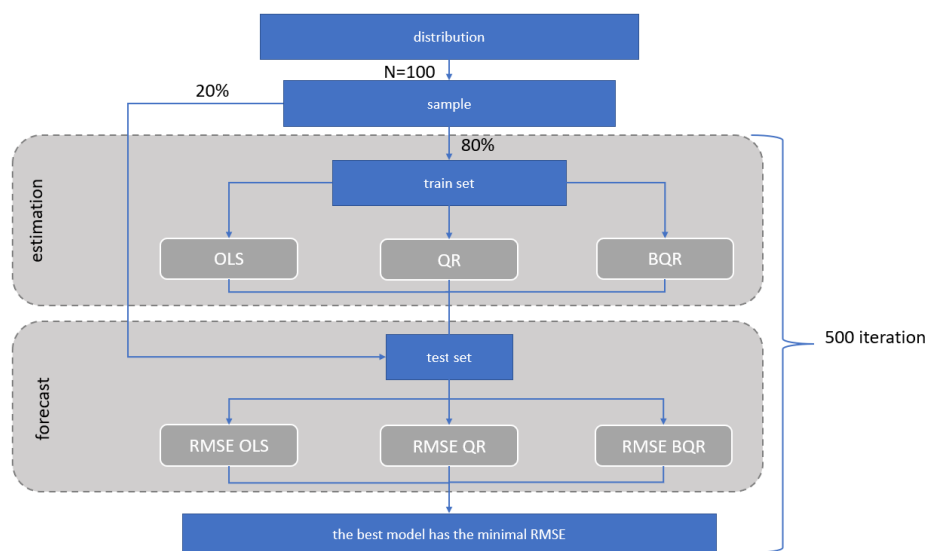
Table 2. Data generating process for experiments

Steps	Results	Comments
1	$x \sim U(1, 10)$, size of sample $N=100$	x was set as sample of size 100 out of uniform distribution in the range from 1 to 10
2	$z_1 \sim N(0, 2)$, size of sample $N=100$	z_1 was set as normal distribution with small variance (with mean=0 and $\sigma^2 = 2$)
	$z_2 \sim N(0, 10)$, size of sample $N=100$	z_2 was set as normal distribution with large variance (with mean=0 and $\sigma^2 = 10$)
	$z_3 \sim Geom(0.1)$, size of sample $N=100$	z_3 was set as geometric distribution with parameter 0.1
	$z_4 \sim \Gamma(1, 0.01)$, size of sample $N=100$	z_4 was set as gamma distribution with parameters 1 and 0.01
	$z_5 \sim lnN(1, 1)$, size of sample $N=100$	z_5 was set as log-normal distribution with mean=1 and $\sigma^2 = 1$
	$z_6 \sim t(1000)$, size of sample $N=100$	z_6 was set as student distribution with degrees of freedom equal to 1000
3	$y_i = 2 + 5 * x + z_i$, where $i = \overline{1,6}$	data generating process
4	$\begin{bmatrix} x^1 & \dots & y_i^1 \\ \vdots & \ddots & \vdots \\ x^{100} & \dots & y_i^{100} \end{bmatrix}$, where $i = \overline{1,6}$	creating a data frame of x and y as matrix $[N \times 2]$

Source: calculations of the author

The general scheme of experiment is as follows (depicted at picture 1). After distributions were chosen as shown in table 2 and the sample of size 100 was generated from every distribution, in the first step, samples were *randomly* divided into train and test sets according to 80%/20% rule. In second step, three models (OLS, quantile regression and Bayes quantile regression model) were estimated based on train set. In third step, a mean forecast for each model was estimated based on train set. Finally, values of RMSE were calculated for each model and the best model was chose according to minimum RMSE. Steps from 1 to 3 is one *iteration* or one experiment.

Pic. 1. The plan of experiment



Source: analysis of the author

In order to explore the change of results with increasing number of iterations, experiment was repeated 500 times. In other words, each sample was split into train and test sets randomly 500 times. After each iteration the best, second-best and third-best model were identified according to the RMSE. Finally, shares of iteration when the model is the best were calculated respectively for each model. Results of the experiments are discussed in the next part of the thesis.

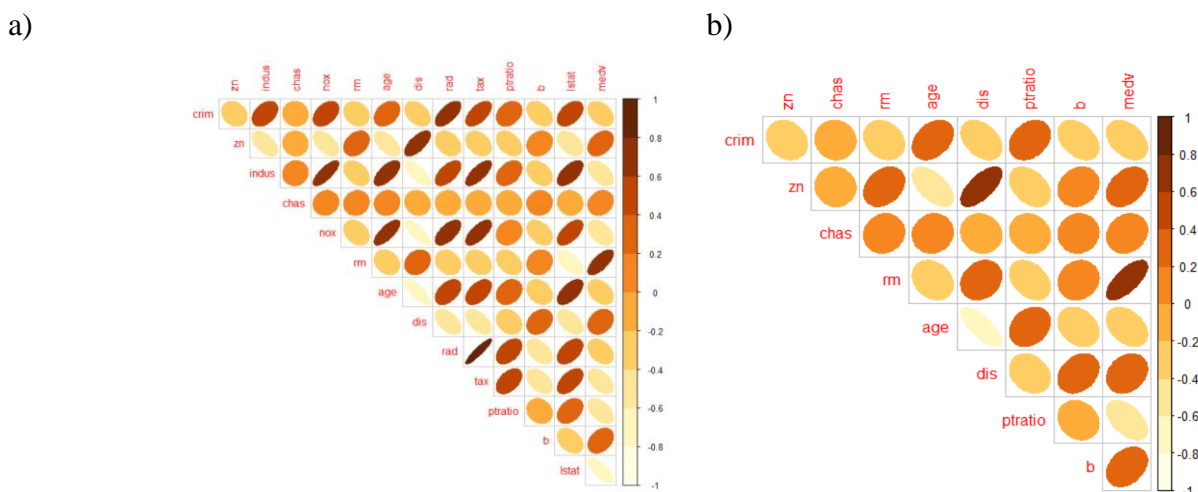
4.2 Preliminary analysis of the Boston Housing data

The difference in experiments on generated and real data is the fact that model on Boston Housing data has its economic grounds. The same methodology was used for experiments both on generated and real data. Before implementing the experiments, preliminary analysis was carried out. Despite the fact that Alhamzawi and Yu (2013) conducted experiments with quantile regression and Bayes models on full Boston Housing data⁴, in this research several key regressors are selected from the data and problems of multicollinearity and heteroskedasticity are considered.

⁴ The description of the Boston housing data with full range of variables is presented in the Appendix 2.

To begin with, the correlation among variables was explored (Pic. 2, a) to solve multicollinearity problem. The dependent variable is *medv* - median value of owner-occupied homes in USD 1000's. Multicollinearity in regression models appears when there is linear relation between independent variables. Strong correlation (>0.8) was detected between variables *rad* and *tax*, *indus* and *nox*, and *indus* and *lstat*. Consequently, these variables were deleted from the data set. Pic 2.b revealed that there are no strongly correlated independent variables in the data.

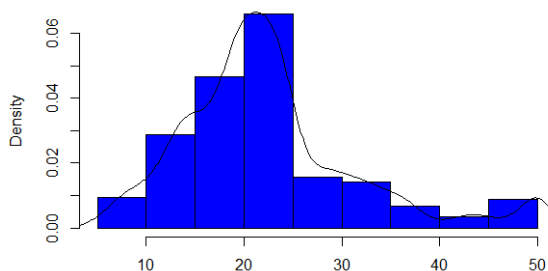
Pic. 2. Correlation matrices for Boston Housing data.



Source: calculations of the author

The distribution of the dependent variable *medv* is non-normal (presented in Pic. 3). As generated samples of different distribution types were used in the experiments, the variable *medv* was not filtered and normalised, and was implemented in the models as it is.

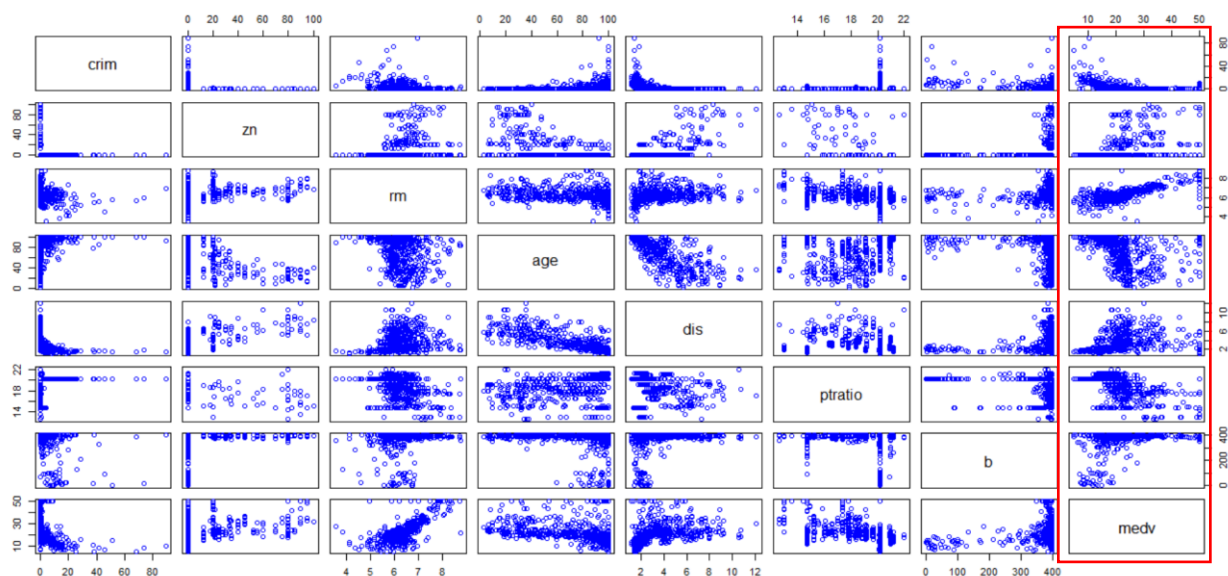
Pic. 3. The distribution of the dependent variable *medv*.



Source: calculations of the author

The dependences among numeric variables (all except *chas*) in forms of scatterplots are depicted in Pic. 4. The rightmost column is the most crucial one as it presented the relationships among dependent variable *medv* and regressors. All in all, both correlation and scatter plots show that there are reasons to include all depicted regressors in the model. Moreover, the quality of the model will be briefly discussed after model estimation. The expected signs of the coefficients were also considered (see table 3). It is important that the economic mechanism beyond the relationship between variables is quite clear and has economic grounds.

Pic. 4. The scatter plot of all numeric variables, included in models.



Source: calculations of the author

Table 3. Expected signs of the coefficient

№	Independent variable	Expected sign of the coefficient	Short description of the variable	Mechanics of relationships to <i>medv</i> - median value of owner-occupied homes in \$1000's
1	<i>crim</i>	-	crime rate	The higher crime rate, the less secure the area, the lower the value of the accommodation
2	<i>zn</i>	+/-	proportion of residential land zoned for lots	The relationship may be twofold. From one hand, the more supply of the lots, the less the price of each lot. From the other hand, the larger number of lots, the more popular area for living, the higher the value of the house.
3	<i>chas</i>	+	river dummy variable	The location on the banks of the river may increase the value of the house due to aesthetic affairs
4	<i>rm</i>	+	number of rooms	The more rooms in the house, the more convenient the apartment, the higher the value of the accommodation
5	<i>age</i>	+/-	proportion of units built prior to 1940	The relationship may be twofold. From one hand, the older the area, the more developed the infrastructure, than the higher the value of the house. From the other hand, the older the house, the more additional costs due to repairing, the lower the value of the house.
6	<i>dis</i>	-	distances to employment centres	The farther the place from the employment centres, the higher additional cost of travelling, and the lower the value of the house
7	<i>ptratio</i>	-	pupil-teacher ratio	The lower number of teachers per pupil in the area, the less convenient the area for families who living the house due to additional cost for valuable education, the lower the value of the house
7	<i>b</i>	-	proportion of blacks	The higher proportion of blacks, the more problems on national grounds may be, the lower the value of the living place

The following regression equation was estimated with three models – OLS, quantile regression and Bayes quantile regression:

$$medv_i = b_0 + b_1 * crim_i + b_2 * zn_i + b_3 * chas_i + b_4 * age_i + b_5 * rm_i + b_6 * dis_i + b_7 * ptratio_i + b_8 * b_i + \epsilon_i,$$

where ϵ_i is error term, i.i.d., and b_0, \dots, b_8 are coefficients of regression.

The regression has simple linear form. The reason for this is the fact, that no strong quadratic or other forms of relationships were identified with scatterplots during preliminary analysis. The second reason is that without additional analysis of the housing market in the second half of the XX century it is difficult to identify any other form of relationships due to economic ground. All in all, for the matter of simplicity, the linear form of regression is used in this research. The novelty of this research, as it was mentioned above, is not in the data or regression model, but in experimental comparison of the models. Both the post-estimation model analysis and results of such comparison are discussed in part 5.

Part 5. Results of experiments

5.1 Results on generated data

Results of the experiment are revealed in table 4. Shares of iterations where the corresponding model is the best and second-best in terms of RMSE are presented by columns for three models – linear model (OLS), quantile regression model (QR) and Bayes quantile regression model (BQR). The shares are depicted with 95% confidence intervals (CI), calculated according to the basic classical formula:

$$\hat{p} - 1.96 * \sqrt{\frac{p*(1-p)}{num}} < p < \hat{p} + 1.96 * \sqrt{\frac{p*(1-p)}{num}},$$

where \hat{p} is estimated share, $num=500$ is a number of iterations in the experiment, and 1.96 is z-score for 95% CI. The results in terms of RMSE are highlighted in bold for those models that have the largest shares of forecasts. RMSE estimates for three models were compared with so called Z-test, testing the equality of shares in two groups, as follows:

$$Z \text{ statistics} = \frac{\hat{p}_1 - \hat{p}_2}{D(\hat{p}_1 - \hat{p}_2)} = \frac{\hat{p}_1 - \hat{p}_2}{\frac{m_1 + m_2}{n_1 + n_2} \left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right) \frac{n_1 + n_2}{n_1 n_2}},$$

where n_1 and n_2 are sizes of samples that the shares are calculated on. n_1 and n_2 are equal to 500 in these experiments. m_1 and m_2 are number of iterations where the corresponding model is the best in terms of RMSE, \hat{p}_1 and \hat{p}_2 are estimated shares.

Z-statistics has standard normal distribution, so two hypotheses can be checked as follows:

H0: $\hat{p}_1 = \hat{p}_2$, there is no statistical difference between two shares at level of significance 0.05.

Ha: $\hat{p}_1 \neq \hat{p}_2$, there is statistical difference between two shares at level of significance 0.05.

Consequently, in cases, when the confidence intervals overlap and statistically there are no differences at level 0.05, the RMSE of several models are highlighted in bold. In other words, in these cases *H0* cannot be rejected at level 0.05 in favour of *Ha*.

The table 4 reveals several conclusions. Firstly, Bayes quantile regression model has the largest share of the best forecasts for 4 types of distributions (for normal distribution both Bayes quantile model and standard quantile model are the best as 95% CI overlap). Secondly, some results look similar for group of distribution. For instance, the Bayes quantile model has the largest share of the best model for geometric and log-normal distributions, and the quantile regression model has largest share of the second-best forecasts for these distributions. OLS model has the largest share of the best forecasts for gamma and student distributions, and quantile regression model has the largest share of second-best forecasts for these distributions. Finally, there is no distributions in this experiment for which Bayes quantile model has the largest share of second-best forecasts.

Abovementioned results partially confirm key hypotheses of this research that is shown in table 5. First hypothesis is rejected. OLS model does not have the largest share of the best forecasts for samples from normal distribution. Second hypothesis is confirmed. Bayes quantile model has the largest share of the best forecasts for 4 out of 6 types of distributions. Third hypothesis is rejected. As it can be seen in Table 6, the shape of error distribution and data plots cannot determine what model will have the largest share of the best or second-best forecasts. The more formal analysis of distribution shape is needed.

It should be emphasized that shape of distribution and data plots can be different. To begin with, from statistical point of view the sample size equal to 100 data points is not large. As data samples were drop randomly from general distributions, the samples can be different if we repeat the experiment. The larger size of samples the more stable results would be. In this research samples of such size are used because in economic research sometimes the number of observations in analysis is not too large. For instance, research on country-level data or region-level data in particular country. Repeating the same experiments on larger samples may be the continuation of this research, but for now it is beyond the scope of this paper. Moreover, 100 observations are enough to notice main features of the distribution. Normal distribution has bell-shape. Log-normal, geometric and gamma (with parameters that used in this research) distributions are shifted to the left. Student distribution is similar to normal with more noticeable tails. What is more, we did not want the experiment to be considered successful only for a large number of observations. All in all, even for $n=100$ the characteristic features of distributions are noticeable, so such size of samples was used in the experiment.

Table 4. Results of experiments based on generated data samples

Distribution type	Share of iteration where the model is the best in terms of RMSE			Share of iteration where the model is the second-best in terms of RMSE		
	OLS	QR	BQR	OLS	QR	BQR
Normal distribution with small variance	14.4% ±3.1%	40.4% ±4.3%	45.2% ±4.4%	72.8% ±3.9%	23.0% ±3.7%	4.2% ±1.8%
Normal distribution with high variance	11.6% ±2.8%	40.4% ±4.3%	48.0% ±4.4%	73.0% ±3.9%	18.8% ±3.4%	8.2% ±2.4%
Gamma distribution	54.8% ±4.4%	9.0% ±2.5%	26.2% ±4.2%	8.4% ±2.4%	86.4% ±3.0%	5.2% ±1.9%
Geometric distribution	32.8% ±4.1%	20.4% ±3.5%	46.8% ±4.4%	20.2% ±3.5%	77.4% ±3.7%	2.4% ±1.3%
Student distribution	52.0% ±4.4%	39.0% ±4.3%	9.0% ±2.5%	40.6% ±4.3%	59.0% ±4.3%	0.04% ±0.6%
Log-normal distribution	28.4% ±4.0%	27.4% ±3.9%	44.2% ±4.4%	36.8% ±4.2%	51.2% ±4.4%	12.0% ±2.8%

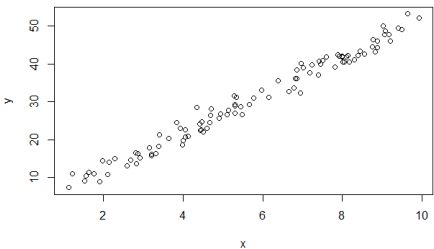
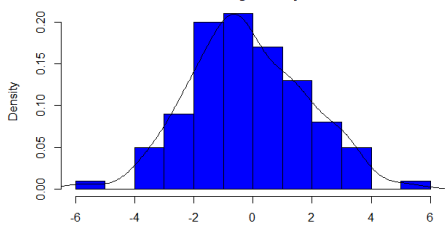
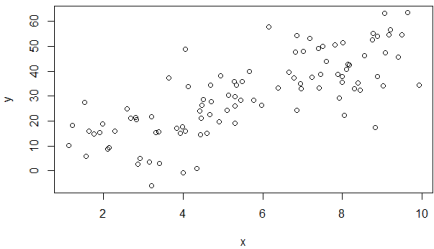
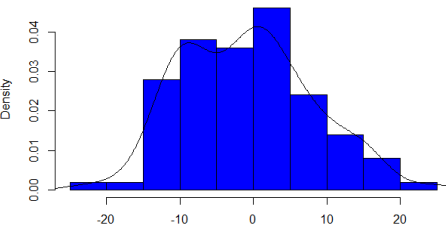
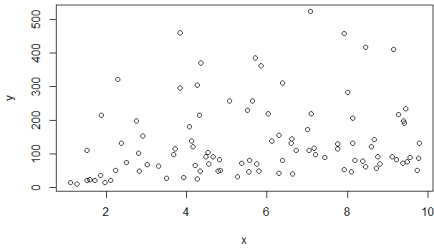
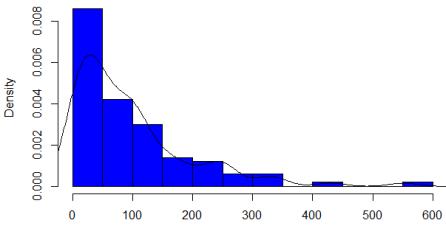
Source: calculations of the author

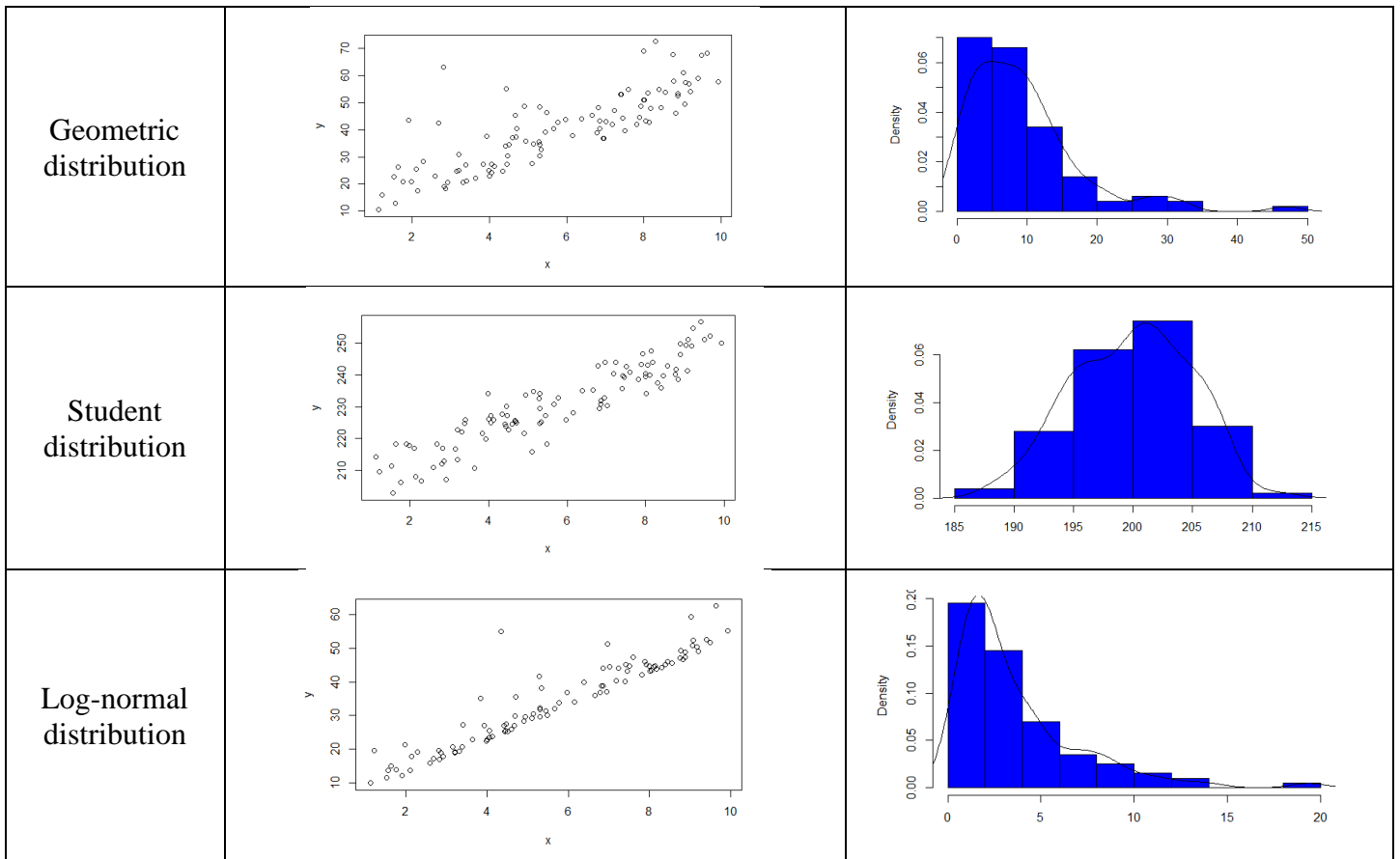
Table 5. Confirmation of key hypotheses of this research by experiments results.

№	Hypothesis	How results of the experience are related to hypotheses
1	Share of best forecasts based on OLS regression are larger for samples with normal distribution.	Results do not confirm this hypothesis. OLS model has the largest share of the second-best forecasts for normal distributions with small and large variance. Nevertheless, OLS model has the largest share of the best forecasts only for gamma and student distributions.
2	Share of best forecasts based on Bayes models are the largest one in general case.	Results confirm this hypothesis. Bayes quantile regression model has the largest share of the best forecasts for 4 out of 6 types of distributions.
3	Shares of best forecast for OLS, quantile model and Bayes quantile model positively correlates between distributions of similar shape.	Results do not confirm this hypothesis. The shape of the distribution or plots of the data do not allow to determine what model will make better forecast in majority of cases (see Table 6). Student and normal distribution may be similar (what generally depends on the distribution parameters), but nevertheless, experiments results for them are different.

Source: analysis of the author

Table 6. Plots of the generated data for experiments.

Distribution type	Plots of random generated samples of size n=100	Plots of z_i , where $i=\overline{1,6}$ according to distribution type (error distribution)
Normal distribution with low variance		
Normal distribution with high variance		
Gamma distribution		



Source: calculations of the author

5.2 Results on real data

The logic of choosing the regression for the regression equation was described in part 4. Here the post-evaluation analysis is presented only for OLS model, because it is enough for confirmation of economic grounds of the regression equation that was estimated.

The regression equation estimated by OLS has the following form:

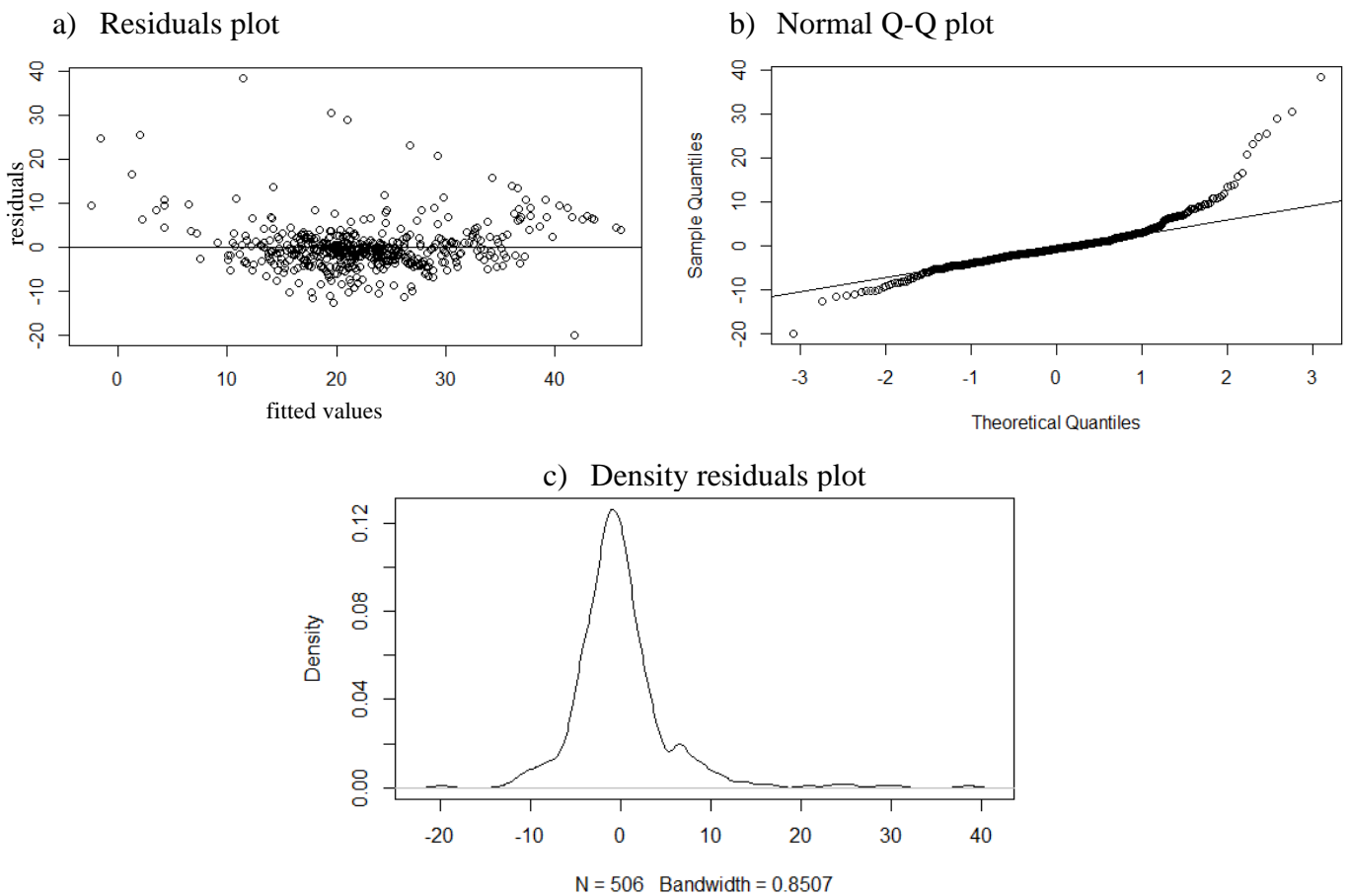
$$medv_i = -5.520 - 0.142 * crim_i + 0.031 * zn_i + 3.22 * chas_i - 0.077 * age_i + 6.947 * rm_i - 0.973 * dis_i - 0.852 * ptratio_i + 0.016 * b_i$$

The model has appropriate quality. All coefficients except coefficient before zn (proxy land lots for sale) are statistically significant at confidence level 0.05. All coefficients except coefficient before b (proxy of share of black population) have signs as it were expected and discussed in part 4. Adjusted R-squared is equal to 0.65 that means that the model explains properly 65% of dispersion in the data. F-statistics is equal to 117 and has $p\text{-value} < 0.0001$.

Nevertheless, the analysis of residuals (Pic. 5) shows that data sample contains several outliers and emphasise that dependent variable does not distribute normally. Firstly, residuals plot revealed that residuals are not distributed like white noise (Pic. 5, a). Secondly, the residuals are not distributed normally (Pic. 5, b). Thirdly, the density residuals plot reveals a long right tail (Pic. 5, c). All these findings shows that OLS in

estimated form is not the best model. One can expect that in conducted experiments forecast based on OLS model will have poor quality in terms of RMSE. Nevertheless, all in all, the simple OLS model shows that there are potentially real strong relationships between dependent variable and regressors and the regression equation may have economic grounds beyond it. It is enough base for the experiments, and their results can be considered further.

Pic. 5. Analysis of residuals of OLS model.



Source: calculations of the author

The results of experiments on Boston Housing data are depicted in table 7. Bayes quantile regression has the largest share of best forecasts, and quantile regression has the second-best largest of forecasts. What is more, OLS model has the lowest share of best forecasts that was partly noticed during analysis of residuals quality for linear model. Generally, this pattern looks similar with results that were received on generated sample from geometric distribution. Nevertheless, without additional analysis it is impossible to make conclusions on how similar these two results are. All in all, Bayes quantile regression model shows the best results in terms of forecasting.

Table 7. Results of experiments on real data.

Data	Share of iteration where the model is the best in terms of RMSE			Share of iteration where the model is the second-best in terms of RMSE		
	OLS	QR	BQR	OLS	QR	BQR
Boston Housing Data	14.5% ±3.1%	37.1% ±4.2%	48.4% ±4.4%	28.0% ±3.9%	68.9% ±4.1%	3.1% ±1.5%

Source: calculations of the author

5.3 Discussion of the results

The results on generated and on real data is shortly discussed below. The results revealed that Bayes quantile regression created better forecast in these experiments. Such result was expected due to methodological advantages that Bayes models have over OLS and quantile regression (these advantages were discussed in part 3). Quantile regression is in the majority of cases is second-best model in terms of forecast in these experiments. Such result was also expected due to methodological advantages of quantile model over OLS. Moreover, even for sample with normal distributions of errors, OLS has not the best forecast in these experiments. One can have a question how resistant these results may be.

The stability of abovementioned results can be checked with increasing number of iterations in the experiments and due to more wide range of distributions that may be used in the analysis. We expect that these results will be stable with increasing number of iterations due to at least two following reasons. First reason is the fact that for majority of distribution types, that we used, the best model can be uniquely identified because corresponding confidence intervals do not overlap. The second reason is that when the same experiments were conducted for smaller number of iterations firstly, the proportion of share best and second-best models were approximately the same. It was generally only one difference - the confidence intervals overlapped in case of 100 iteration for majority of experiments. All in all, the same methodology can be implemented on larger number of data points, with larger number of iterations and for more distribution types in order to explore the stability of results, but these questions now are beyond of the scope of this thesis.

Part 6. Conclusion

In this thesis the series of experiments were held in order to experimentally identify the best forecasts in terms of RMSE between quantile regression models based on traditional and Bayes approach. OLS model was used as a baseline. The same methodology was implemented both for samples of generated data and on Boston Housing data.

The main results on generated data can be summarised in three key point as follows. Firstly, Bayes quantile regression model has the largest share of the best forecasts for 4 types of distributions (for normal distribution both Bayes quantile model and standard quantile model are the best as 95% CI overlap). Secondly, some results look similar for group of distribution. For instance, the Bayes quantile model has the largest share of the best model for geometric and log-normal distributions, and the quantile regression model has largest share of the second-best forecasts for these distributions. Thirdly, OLS model has the largest share of the best forecasts for gamma and student distributions, and quantile regression model has the largest share of second-best forecasts for these distributions.

The results of experiments on real data reveals that Bayes quantile model has the largest share of best forecasts, and quantile regression has the second-best largest of forecasts. What is more, OLS model has the lowest share of best forecasts that was partly noticed during analysis of residuals quality for linear model. Generally, this pattern looks similar with results that were received on generated sample from geometric distribution. All in all, Bayes quantile regression model shows the best results in terms of forecasting both on generated and real data.

Abovementioned results partially confirm key hypotheses of this research. First hypothesis is rejected since OLS model does not have the largest share of the best forecasts for samples from normal distribution. Second hypothesis is confirmed, because Bayes quantile model has the largest share of the best forecasts for 4 out 6 types of distributions. Third hypothesis is rejected as the shape of error distribution and data plots cannot determine what model will have the largest share of the best or second-best forecasts. The more formal analysis of distribution shape is needed to connect distribution type and results of the experiments.

In conclusion, the results of presented experiments shows that clear theoretical advantages of the model do not strictly determine its best results on data. Any theory, no matter how straightforward it is, should be confirmed during implementation on data. Real data do not often used to estimate 1 500 models as it was done in these experiments. When larger number of experiments are needed for stable conclusions, generated samples from distributions with control of parameters can be used like it was done in this paper. We hope that this research will be able to find its own place in wide discussion of advantages of quantile models based on classical and Bayes approaches.

Bibliography

1. Alhamzawi, R., & Ali, H. T. M. (2020). Brq: An r package for bayesian quantile regression. *METRON*, 78(3), 313-328.
2. Alhamzawi, R., & Yu, K. (2013). Conjugate priors and variable selection for Bayesian quantile regression. *Computational Statistics & Data Analysis*, 64, 209-219.
3. Allen, V. G., Pond, K. R., Saker, K. E., Fontenot, J. P., Bagley, C. P., Ivy, R. L., & Melton, C. (2001). Tasco: Influence of a brown seaweed on antioxidants in forages and livestock—A review. *Journal of Animal science*, 79(suppl_E), E21-E31.
4. Baba, H., & Shimizu, C. (2022). The impact of apartment vacancies on nearby housing rents over multiple time periods: application of smart meter data. *International Journal of Housing Markets and Analysis*, Available in: <https://www.emerald.com/insight/content/doi/10.1108/IJHMA-03-2022-0035/full/html> (Downloaded: 01 June 2022).
5. Benoit, D. F., & Van den Poel, D. (2017). bayesQR: A Bayesian approach to quantile regression. *Journal of Statistical Software*, 76, 1-32.
6. Brown, R. L., & Peet, R. K. (2003). Diversity and invasibility of southern Appalachian plant communities. *Ecology*, 84(1), 32-39.
7. Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, 88-126.
8. Cade, B. S., & Guo, Q. (2000). Estimating effects of constraints on plant performance with regression quantiles. *Oikos*, 91(2), 245-254.
9. Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8), 412-420.
10. Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in medicine*, 11(10), 1305-1319.
11. Fitzenberger, B., Koenker, R., & Machado, J. A. (Eds.). (2001). *Economic applications of quantile regression*. Springer Science & Business Media.
12. Gannoun, A., Saracco, J., & Yu, K. (2003). Nonparametric prediction by conditional median and quantiles. *Journal of statistical Planning and inference*, 117(2), 207-223.
13. Hallock, K. F., & Koenker, R. (2001). Quantile regression. *The Journal of Economic Perspectives*, 15(4), 143-156.
14. Knight, C. A., & Ackerly, D. D. (2002). Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*, 5(1), 66-76.
15. Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.
16. Koenker, R., & Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454), 458-468.

17. Koenker, R., & Ng, P. (2005). Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics*, 418-440.
18. Koenker, R., & Xiao, Z. (2006). Quantile autoregression. *Journal of the American statistical association*, 101(475), 980-990.
19. Lancaster, T., & Jae Jun, S. (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2), 287-307.
20. Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., & Zhao, L. P. (1997). Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4), 463-476.
21. Mahdi, E., & Al-Abdulla, A. (2022). Impact of COVID-19 Pandemic News on the Cryptocurrency Market and Gold Returns: A Quantile-on-Quantile Regression Analysis. *Econometrics*, 10(2), 26.
22. McClain, C., & Rex, M. (2001). The relationship between dissolved oxygen concentration and maximum size in deep-sea turrid gastropods: an application of quantile regression. *Marine Biology*, 139(4), 681-685.
23. Pandey, G. R., & Nguyen, V. T. V. (1999). A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225(1-2), 92-101.
24. Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3), 331-350.

Appendix 1. Implementation of the experiment in R programming environment.

```
#Uploading packages that used in the analysis:
suppressPackageStartupMessages(library(mlbench))
suppressPackageStartupMessages(library(rstanarm))
suppressPackageStartupMessages(library(bayestestR))
suppressPackageStartupMessages(library(bayesplot))
suppressPackageStartupMessages(library(insight))
suppressPackageStartupMessages(library(broom))
suppressPackageStartupMessages(library(bayesQR))
suppressPackageStartupMessages(library(quantreg))
suppressPackageStartupMessages(library(mvtnorm))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(exps))
suppressPackageStartupMessages(library(plyr))
#set.seed(0) for repeating the results
nexp=500 #number of experiments
num=100 #number of observations in the data samples
#creating data frames of generated samples
x <- runif(num, 1, 10)
y <- 2 + 5 * x + rgamma(num, 1, 0.01) # my_data2.csv
y <- 2 + 5 * x + rgeom(num, 0.1) # my_data4.csv+
y <- 2 + 5 * x + rt(num, 1000, 200) # my_data8.csv+
y <- 2 + 5 * x + rlnorm(num, 1, 1) # my_data5.csv+
y <- 2 + 5 * x + rnorm(num, 0, 10) # my_data7.csv
y <- 2 + 5 * x + rnorm(num, 0, 2) # my_data1.csv+
df <- data.frame(x, y)
plot(df)
#results table is as follows
res <- data.frame(RMSE=c(1:nexp),RMSE_qr=c(1:nexp),RMSE_bqr=c(1:nexp))
#loop for each iteration
for(g in 1: nexp){
#set.seed(123) for repeating the results
#generate train and test samples as follows:
train.index = sample(1:nrow(df), nrow(df)*0.8)
train = df[train.index, ]
test = df[-train.index, ]
#estimation of OLS model
```

```

model_freq<-lm(y~x, data=train)
b0=coef(model_freq)[1]
b1=coef(model_freq)[2]
#forecast in OLS model
test$y_hat=b0+b1*test$x
test$tech=(test$y_hat-test$y)^2
RMSE=sqrt(sum(test$tech)/length(test$tech)) #RMSE in OLS model
#formulation of quantile regression model
b=0.01
b_start=0+b
b_end=1-b
q=seq(b_start, b_end, by = b)
n=length(q)
if (ncol(df)>2) df[3:ncol(df)]<-list(NULL) # technical thing to filter initial the data
df1 = as.data.frame(matrix(0, nrow = 2, ncol = n)) #technical table for coefficients
#loop for estimation of quantile regression model
for(i in 1:n){
df1[,i] <- coef(rq(y~x,data=train, tau=q[i]))
}
if (ncol(test)>2) test[3:ncol(test)]<-list(NULL) # technical thing to filter the initial the data
#loop for forecasting in quantile regression model
for(i in 1:n){
test[,2+i] <- df1[1,i] + df1[2,i] * test$x
}
d=ncol(test)+1
test[,d]<-0
#loop for forecasting in quantile regression model
for (j in 1:nrow(test)) {
for(i in 1:n) {
test[j,d] <- test[j,d] + test[j,2+i]
}
}
n1=n+2
test<-mutate(test, y_qr = rowMeans(test[,3:n1]))
test$tech1=(test$y_qr-test$y)^2
RMSE_qr=sqrt(sum(test$tech1)/length(test$tech1)) #RMSE in quantile regression model
if (ncol(df)>2) df[3:ncol(df)]<-list(NULL) # technical thing to filter initial the data

```

```

df2 = as.data.frame(matrix(0, nrow = 2, ncol = n)) # technical table for coefficients
# loop for estimation of Bayes quantile model
for(i in 1:n){
s<-summary(bayesQR(y~x,data=train, quantile=q[i],ndraw=1000,seed=111),burnin=500)
s1<-s[[1]]$betadraw
df2[1,i]<- s1[1,1]
df2[2,i]<- s1[2,1]
}
if (ncol(test)>2) test[3:ncol(test)]<-list(NULL) # technical thing to filter the initial data
# loop for forecasting in Bayes quantile model
for(i in 1:n){
  test[,2+i] <- df2[1,i] + df2[2,i] * test$x
}
n1=n+2
test<-mutate(test, y_bqr = rowMeans(test[,3:n1]))
test$tech1=(test$y_bqr-test$y)^2
RMSE_bqr=sqrt(sum(test$tech1)/length(test$tech1)) # RMSE in Bayes quantile model
# organising the table of results
res[g,3]=RMSE_bqr
res[g,1]=RMSE
res[g,2]=RMSE_qr
}

```

Appendix 2. Description of Boston Housing data

Boston Housing data contains 506 observations of 14 variables as follows. Unit of observation is a house in living area of Boston. Data was collected and firstly published in 1978.

No	Variable	Description
1	<i>crim</i>	per capita crime rate by town
2	<i>zn</i>	proportion of residential land zoned for lots over 25,000 sq.ft
3	<i>indus</i>	proportion of non-retail business acres per town
4	<i>chas</i>	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5	<i>nox</i>	nitric oxides concentration (parts per 10 million)
6	<i>rm</i>	average number of rooms per dwelling
7	<i>age</i>	proportion of owner-occupied units built prior to 1940
8	<i>dis</i>	weighted distances to five Boston employment centres
9	<i>rad</i>	index of accessibility to radial highways
10	<i>tax</i>	full-value property-tax rate per \$10,000
11	<i>prratio</i>	pupil-teacher ratio by town
12	<i>b</i>	$1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
13	<i>lstat</i>	% lower status of the population
14	<i>medv</i>	Median value of owner-occupied homes in \$1000's

Source: <https://www.cs.toronto.edu/~delves/data/boston/bostonDetail.html>

Summary

Introduction

The aim of the thesis is experimentally identify the best forecasts in terms of RMSE between quantile regression models based on traditional and Bayes approach.

This paper considers three following **research questions**:

- Does the Bayesian quantile regression model make forecasts with lower RMSE than the quantile regression in the classical formulation?
- How change results based on generated sample from different types of distributions?
- How do results on generated data correlate to estimates on real data?

Several **hypotheses** follow from the analysis of literature and research questions:

- Share of best forecasts based on OLS regression are larger for samples with normal distribution.
- Share of best forecasts based on Bayes models are the largest one in general case.
- Shares of best forecast for OLS, quantile model and Bayes quantile model positively correlates between distributions of similar shape.

The **delimitations** of the paper are as follows. Firstly, this research is based only on linear forecasts. There are also examples of nonlinear quantiles regression models in the literature, however, research questions of this paper can be explored on linear models that is computationally more convenient for the investigation. Secondly, there are various kinds of metrics to assess the predictive power of models. In this study, RMSE is used as one of the basic and universal measure. Thirdly, to estimate the quality of forecast, the predicted values are analysed. The real and estimated coefficient of regression line are not compared, because it has already discussed in literature, for instance, in Alhamzawi and Yu (2013). Finally, as real data, Boston Housing data is used - it is widely known and explored in literature. The original idea of this research is connected to the method of comparing models, but not in the data itself. Going beyond these limitations can be a continuation of this study and makes it more valuable, but now it remains outside the scope of this study.

The **structure** of the thesis is as follows. In the second part the advantages of quantile regressions are discussed comparing to OLS regressions. Firstly, quantile regression allows to better explore the tales of the distribution. Secondly, such models are less sensitive to the influence if outliers in the data. Finally, quantile regressions are quite widespread and used in large number of research in different field from medicine to economics and time series.

In the third part the characteristics of Bayes quantile models are considered in comparison with classical quantile regression and OLS models. The methodology and key points in estimating of the Bayes

models are briefly discussed. Key advantages of Bayes approach are summarised at the end of the third part and some crucial characteristics of model estimation process in our experiments are provided.

The foundation and the methodology of experiments are presented in the fourth part. Experiment is based on the sample of size 100 from 6 different distribution types. Then these samples are splitting on train and test sub-samples. For each sample three types of models (OLS, quantile regression and Bayes quantile regression) are estimating. Then forecasts are built on the test sub-sample and RMSE for each forecast are calculating. All in all, 500 experiment series were performed for each type of six generated distribution to notice tendencies beyond RMSE estimates. OLS model is used like baseline that quantile regression and Bayes quantile regression models are compared with. Furthermore, the same methodology is implemented on Boston Housing data that contain prices and characteristics of 506 houses in living area of Boston. It is quite popular in the literature and was firstly published in 1978. The preliminary analysis of the data, selection of regressors and regression equation are also discussed in the fourth part. The difference of experiments on generated and real data is the fact that the models on real data are built based on economic grounds and analysis of economic relationships between variables.

The fifth part presents results of experiments on both generated and real data. The best model in each experiment is identified based on RMSE. It is revealed that share cases when each model is better is supposed to be stable as the number of experiments grow. 95% CI are calculated for shares when OLS, quantile regression and Bayes quantile regression are first-, second- and third-best model in terms of RMSE. The tendencies beyond the estimates of RMSE are summarised at the end of the fifth part.

Finally, the conclusion emphasises key finding of this research. Firstly, the short discussion considers confirmation or rejected the hypotheses mentioned above. Secondly, the further potential steps expanding this study are briefly discussed.

Methodology

Quantile regressions became widespread in the scientific literature after an article by Koenker and Bassett (1978) in the *Econometrica* journal⁵, published by The Econometric Society⁶. The authors showed that the same general approach that used in conventional linear regression can be implemented for different quantiles of distribution of independent variable. The term “regression quantiles” was proposed by them and, that is more importantly, they suggested new estimator of such type of regression. It had an incredible impact on research in cases, where errors are distributed non-normally and assumption of Gauss-Markov theorem are violated. In cases, where the errors are distributed normally, the result of evaluating the quantile regression models turns out to be similar to the usual linear regression. This part will be structured as follows. First, the motivation for the transition to quantile regressions as a new class of models will be discussed. Secondly, it

⁵ <https://www.jstor.org/journal/econometrica> - Econometrica journal page on JSTOR

⁶ <https://www.jstor.org/publisher/econosoc> - The Econometric society page on JSTOR

will be demonstrated how important this approach is in practice and in which areas we can be most valuable. Thirdly, methodological features of quantile regressions will be described. Finally, the role of quantile regressions in this study is described in conclusion.

The *motivation* for the introduction of quantile regressions is based on the following characteristics. First of all, the usual regressions (least squares) do not work well with outliers. What is more, both estimates and mean forecast in OLS is sensitive to outliers. Outliers can significantly affect the results of the model, so researchers often simply exclude outliers at the preliminary analysis stage. Nevertheless, if outliers would not bias the model results, we would prefer to leave them in the data sample. In other words, if it supposed to be a data generation process behind the observed data, then outliers are part of this process. Moreover, the problem of outliers lead scientists to make a forecast based on the median (not on mean prognoze), as in the work of Gannoun et al. (2003). The median approach can be generalized to different quantiles. Secondly, the assumption of the normality of errors as a consequence of the central limit theorem and the law of large numbers often looks too strict on real data. Errors as a random factor are not observable. Consequently, the requirements of the Gauss-Markov theorem turn out to be difficult to implement in practice. Thirdly, the quantile approach shows itself better in models when the data generation process behind the sample may have an unknown distribution, especially non-normal. Nevertheless, quantile regressions are based on a similar minimization problem, as OLS. Quantile regressions are also sensitive to the number of observations, and if there are not enough observations, they revealed poor results. All the advantages of quantile regressions described above are used in a variety of empirical studies, which we will discuss further.

The quantile regression model is formalised as follows. The methodology of estimating is briefly presented in this paper similar to Benoit and Van den Poel (2017).

The main logics is similar to classic regression. The linear model is:

$$y_i = x_i^\tau \beta + \epsilon_i,$$

where ϵ_i is error term and τ is a given quantile and β is vector of coefficients. The solution for β is as following:

$$\hat{\beta}_\tau = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \rho_\tau(y_i - x_i^\tau \beta),$$

where $0 < \rho_\tau < 1$ corresponding to any particular quantile. In general, using this method, $\hat{\beta}_\tau$ for any quantile can be estimated. For instance, median forecast can be calculated using $\hat{\beta}_{0.5}$. Similarly, predictions for a set of quantiles can be evaluated. Symmetric sets of quantiles can be evaluated to receive mean forecast based on quantile regression model like simple mean of estimated quantiles forecasts. In our experiments set of quantiles [0.01, 0.99] is used in order to compare results with OLS model that created mean forecast by default. In literature (for example, Benoit and Van den Poel (2017)) quality of median and mean forecasts are compared, but this question is beyond the scope of this research.

In this research, quantile regression models are estimated automatically using package “quantreg” in R programming environment.

Cole and Green (1992) used data on triceps skinfold in Gambian girls and women, and body weight in U.S.A. girls. The authors extend the quantile regression approach by the example of nonlinear models of the dependence of triceps skinfold in Gambian girls and women, and body weight in U.S.A. girls on age. Their study uses reference centile curves as a way to track anomalies in the data. Such curves usually represent the dependence of the parameter of interest (for example, weight or muscle mass, body fat percentage, etc.) depending on age. In this approach, it is impossible to throw out outliers, because the task is to detect them.

Koenker and Selling (2001) use quantile regressions to analyse the survival of medflies. The authors use a two-sample treatment-control model. The indicator that the size of the pupa had an important influence for survival for the smallest 10% and had a negative influence at the upper end of the distribution. This is a very clear example of how the same variable can affect differently depending on which part of the distribution of the variable of interest scientists are considering.

Quantile regressions also turn out to be convenient on time series data. In Gannoun et al. (2003) it is shown that quantile regression attracts the attention of researchers of data that are distributed asymmetrically, for example, income or real estate price. The study is based on nonlinear autoregressive models based on time series. An illustrative example of the use of quantile regressions on time series data is also contained in Koenker and Xiao (2006). The paper also examines abnormally distributed data – the American unemployment rate and U.S. gasoline prices. The authors call the quantile model on time data the quantile autoregression model (QAR). The authors believe that the QAR model has the potential to become a separate area of time series research. The authors are conducting a Monte Carlo experiment to explore the possibilities of QAR and focus on the model with iid errors.

There are many examples in the literature of the application of quantile regressions to the analysis of modern crises of recent years. For instance, the work of Mahdi, E., & Al-Abdulla, A. (2022) explores the relationship between bitcoin and gold prices and news indices that characterize the perception of crises, such as Panic, Sentiment, Infodemic, and Media Coverage (talking about the RavenPack news-based index⁷). The authors discovered that despite the similar nature - both gold and bitcoin are used to hedge risks, the impact of news indices on demand and prices of commodities is asymmetric. The authors reveal the disadvantages of quantile regressions. They turn out to be less effective than the quantile-on-quantile approach. The reason is that positive and negative shocks affect the market with different intensity.

⁷ <https://www.ravenpack.com/> - more information about RavenPack company turning news, social media, transcripts, filings, and other texts in valuable insights for business.

In literature, Bayes approach is a competitor to the frequentists' methods as the first one allows to control the prior distribution of parameters of the model. Bayes approach in regression modelling came from the classical Bayes formula:

$$P(\theta|D) = \frac{P(D|\theta)*P(\theta)}{P(D)},$$

where $P(\theta)$ is a prior distribution of a parameter of the model – hypothetic or based on preliminary analysis of data (in case of regression model it may be a prior distribution of a coefficient);

$P(\theta|D)$ — a posterior probability of the parameter of the model – it is what the model should estimate;

$P(D|\theta)$ — likelihood;

$P(D)$ — the total probability of occurrence of the data (evidence).

The example of regression model in Bayes formulation is provided below.

The typical regression model can be formulated based on Bayes formula:

$$y_i|x_i = a + b x_i + \varepsilon_i,$$

where ε_i – error term, i.i.d., $\varepsilon_i \sim N(0, \sigma^2)$ and data sample consist of independent pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$.

$$y_i|x_i, a, b, \sigma^2 \sim N(a + b x_i, \sigma^2)$$

For convenience, the same formulation of the model is written down in a more convenient form (in particular, $\tau = \frac{1}{\sigma^2}$).

In this case likelihood is formulated as follows:

$$y_i|\mu_i, \tau \sim N(\mu_i, \tau)$$

and priors may be formulated as

$$a \sim N(\mu_a, \tau_a)$$

$$b \sim N(\mu_b, \tau_b)$$

$$\tau \sim \text{Gamma}(a_\tau, b_\tau)$$

Regression coefficients a and b can take both positive and negative values, so the normal distribution is chosen as a prior for them, the variation estimated in terms of τ takes values greater than 0, so the Gamma distribution is chosen as a prior for it. Defined all the parameters as above, we have formulated the Bayesian linear regression problem. It is easy to use analogous logic to multiple quantiles and receive Bayes quantile model.

If we use the matrix form for simplicity, then Bayes quantile regression may look in the following form:

$$Q_{y_i}(x|p) = x_i' \beta_p,$$

where $0 < p < 1$, and $Q_{y_i}(\cdot) = F_{y_i}^{-1}(\cdot)$ – inverse of the cumulative distribution function of a variable y_i , on condition of x_i .

The quality of the model varies according to type of prior used. From one hand, by default, the non-informative prior can be used. It is normal prior with zero mean and large variance. From the other hand, according to Alhamzawi and Yu (2013), the best forecast in Bayes quantile model can be received using not general prior but special prior for every quantile. The authors created a series of experiments similar to experiments in this research. Nevertheless, the use coefficients comparison for evaluating quality of the model (as the distribution types was controlled in their research, they compared real coefficients with estimates in particular model). On the contrary, in our research the quality of the model is estimated based on forecast quality. In experiments in this thesis the non-informative priors are used. As it reveals in the results section, Bayes model can compete with OLS and quantile model even without formulating of informative priors. Also, sometimes scientists' knowledge about prior distribution of the real model is not enough or absent at all.

Bayesian quantile regression methods are widely discussed in Lancaster and Jae Jun (2010). It is important to it is important to take into account several parameters. Firstly, since distributions are used in the model and the final answer is a posteriori distribution, and not a point estimate as in the standard model. Gibbs sampler is used to find a solution (for details see Alhamzawi and Yu (2013)). Finding a solution (posterior distribution) takes place through several iterations. The number of iterations has crucial influence and is usually quite large. In experiments in this research 1000 number of iterations is used. Since the algorithm does not converge to the correct solution immediately, part of the first iterations must be discarded (this part of the series is called “burning-in”). Usually, a quarter or half of the iterations are discarded. In these experiments 500 iterations are “burning-in”.

In conclusion, Bayes approach has several advantages comparing to classical frequentist approach. Firstly, it can effectively work on small data samples. Secondly, Bayes models can be implemented based on different distribution. Bayes models are less sensitive to outliers and the requirement of normality compared with classical regressions.

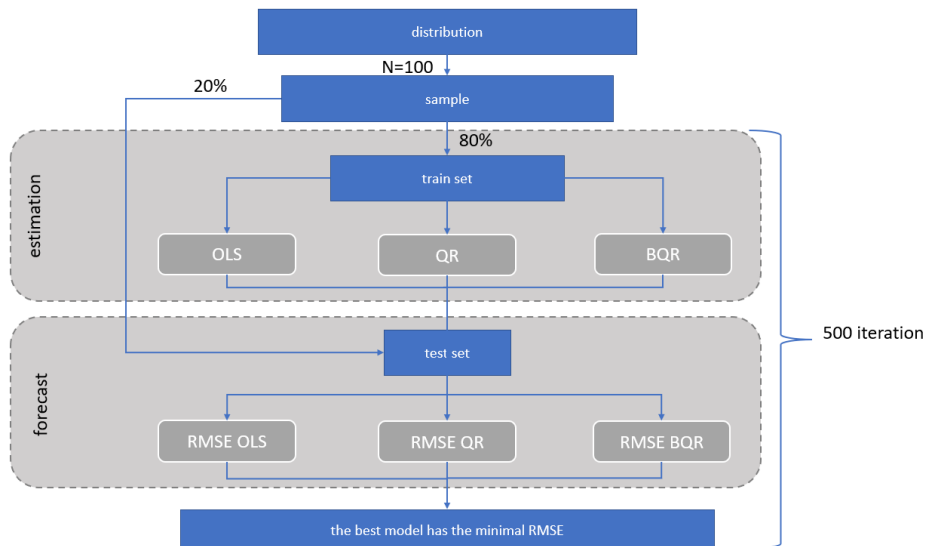
Estimation of Bayes quantile regression is available in R programming environment with packages “*bqr*” (see Alhamzawi and Ali (2020)) and “*bayesQR*” (see Benoit and Van den Poel (2017) for details). The last one was used during experiments in this research.

Data

The series of experiments were based on sample generated from 6 types of data distributions (normal with small variance, normal with large variance, geometric, gamma, log-normal and student distribution). In the first step, variable x was set as sample of size 100 out of uniform distribution in the range from 1 to 10. In the second step, variables z_1, z_2, \dots, z_6 were set as samples of size 100 from 6 types of distributions. In the third step, variables y_1, y_2, \dots, y_6 was set as $y_i = 2 + 5 * x + z_i$, where $i = \overline{1,6}$. Finally, 6 data frames of x and y_i were created.

The general scheme of experiment is as follows (depicted at picture 1). tabs were chosen and the sample of size 100 was generated from every distribution, in the first step, samples were *randomly* divided into train and test sets according to 80%/20% rule. In second step, three models (OLS, quantile regression and Bayes quantile regression model) were estimated based on train set. In third step, a mean forecast for each model was estimated based on train set. Finally, values of RMSE were calculated for each model and the best model was chose according to minimum RMSE. Steps from 1 to 3 is one *iteration* or one experiment.

Pic. 1. The plan of experiment



Source: analysis of the author

In order to explore the change of results with increasing number of iterations, experiment was repeated 500 times. In other words, each sample was split into train and test sets randomly 500 times. After each iteration the best, second-best and third-best model were identified according to the RMSE. Finally, shares of iteration when the model is the best were calculated respectively for each model. Results of the experiments are discussed in the next part of the thesis.

The difference in experiments on generated and real data is the fact that model on Boston Housing data has its economic grounds. The same methodology was used for experiments both on generated and real data. Before implementing the experiments, preliminary analysis was carried out. Despite the fact that Alhamzawi and Yu (2013) conducted experiments with quantile regression and Bayes models on full Boston Housing

data⁸, in this research several key regressors are selected from the data and problems of multicollinearity and heteroskedasticity are considered.

To begin with, the correlation among variables was explored to solve multicollinearity problem. The dependent variable is *medv* - median value of owner-occupied homes in USD 1000's. Multicollinearity in regression models appears when there is linear relation between independent variables. Strong correlation (>0.8) was detected between variables *rad* and *tax*, *indus* and *nox*, and *indus* and *lstat*. Consequently, these variables were deleted from the data set. There are no strongly correlated independent variables in the data.

The distribution of the dependent variable *medv* is non-normal. As generated samples of different distribution types were used in the experiments, the variable *medv* was not filtered and normalised, and was implemented in the models as it is.

All in all, both correlation and scatter plots show that there are reasons to include all depicted regressors in the model. Moreover, the quality of the model will be briefly discussed after model estimation. The expected signs of the coefficients were also considered. It is important that the economic mechanism beyond the relationship between variables is quite clear and has economic grounds.

The following regression equation was estimated with three models – OLS, quantile regression and Bayes quantile regression:

$$medv_i = b_0 + b_1 * crim_i + b_2 * zn_i + b_3 * chas_i + b_4 * age_i + b_5 * rm_i + b_6 * dis_i + b_7 * ptratio_i + b_8 * b_i + \epsilon_i,$$

where ϵ_i is error term, i.i.d., and b_0, \dots, b_8 are coefficients of regression.

The regression has simple linear form. The reason for this is the fact, that no strong quadratic or other forms of relationships were identified with scatterplots during preliminary analysis. The second reason is that without additional analysis of the housing market in the second half of the XX century it is difficult to identify any other form of relationships due to economic ground. All in all, for the matter of simplicity, the linear form of regression is used in this research. The novelty of this research, as it was mentioned above, is not in the data or regression model, but in experimental comparison of the models. Both the post-estimation model analysis and results of such comparison are discussed in next part.

Results

Results of the experiment are revealed in table 1. Shares of iterations where the corresponding model is the best and second-best in terms of RMSE are presented by columns for three models – linear model (OLS), quantile regression model (QR) and Bayes quantile regression model (BQR). The shares are depicted with 95% confidence intervals (CI), calculated according to the basic classical formula:

⁸ The description of the Boston housing data with full range of variables is presented in the Appendix 2.

$$\hat{p} - 1.96 * \sqrt{\frac{p*(1-p)}{num}} < p < \hat{p} + 1.96 * \sqrt{\frac{p*(1-p)}{num}},$$

where \hat{p} is estimated share, $num=500$ is a number of iterations in the experiment, and 1.96 is z-score for 95% CI. The results in terms of RMSE are highlighted in bold for those models that have the largest shares of forecasts. RMSE estimates for three models were compared with so called Z-test, testing the equality of shares in two groups, as follows:

$$Z \text{ statistics} = \frac{\hat{p}_1 - \hat{p}_2}{D(\hat{p}_1 - \hat{p}_2)} = \frac{\hat{p}_1 - \hat{p}_2}{\frac{m_1 + m_2}{n_1 + n_2} \left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right) \frac{n_1 + n_2}{n_1 n_2}},$$

where n_1 and n_2 are sizes of samples that the shares are calculated on. n_1 and n_2 are equal to 500 in these experiments. m_1 and m_2 are number of iterations where the corresponding model is the best in terms of RMSE, \hat{p}_1 and \hat{p}_2 are estimated shares.

Z-statistics has standard normal distribution, so two hypotheses can be checked as follows:

H0: $\hat{p}_1 = \hat{p}_2$, there is no statistical difference between two shares at level of significance 0.05.

Ha: $\hat{p}_1 \neq \hat{p}_2$, there is statistical difference between two shares at level of significance 0.05.

Consequently, in cases, when the confidence intervals overlap and statistically there are no differences at level 0.05, the RMSE of several models are highlighted in bold. In other words, in these cases *H0* cannot be rejected at level 0.05 in favour of *Ha*.

There are several conclusions. Firstly, Bayes quantile regression model has the largest share of the best forecasts for 4 types of distributions (for normal distribution both Bayes quantile model and standard quantile model are the best as 95% CI overlap). Secondly, some results look similar for group of distribution. For instance, the Bayes quantile model has the largest share of the best model for geometric and log-normal distributions, and the quantile regression model has largest share of the second-best forecasts for these distributions. OLS model has the largest share of the best forecasts for gamma and student distributions, and quantile regression model has the largest share of second-best forecasts for these distributions. Finally, there is no distributions in this experiment for which Bayes quantile model has the largest share of second-best forecasts.

Abovementioned results partially confirm key hypotheses of this research. First hypothesis is rejected. OLS model does not have the largest share of the best forecasts for samples from normal distribution. Second hypothesis is confirmed. Bayes quantile model has the largest share of the best forecasts for 4 out of 6 types of distributions. Third hypothesis is rejected. The shape of error distribution and data plots cannot determine what model will have the largest share of the best or second-best forecasts. The more formal analysis of distribution shape is needed.

It should be emphasized that shape of distribution and data plots can be different. To begin with, from statistical point of view the sample size equal to 100 data points is not large. As data samples were drop randomly from general distributions, the samples can be different if we repeat the experiment. The larger size of samples the more stable results would be. In this research samples of such size are used because in economic research sometimes the number of observations in analysis is not too large. For instance, research on country-level data or region-level data in particular country. Repeating the same experiments on larger samples may be the continuation of this research, but for now it is beyond the scope of this paper. Moreover, 100 observations are enough to notice main features of the distribution. Normal distribution has bell-shape. Log-normal, geometric and gamma (with parameters that used in this research) distributions are shifted to the left. Student distribution is similar to normal with more noticeable tails. What is more, we did not want the experiment to be considered successful only for a large number of observations. All in all. even for n=100 the characteristic features of distributions are noticeable, so such size of samples was used in the experiment.

Table 1. Results of experiments based on generated data samples

Distribution type	Share of iteration where the model is the best in terms of RMSE			Share of iteration where the model is the second-best in terms of RMSE		
	OLS	QR	BQR	OLS	QR	BQR
Normal distribution with small variance	14.4% ±3.1%	40.4% ±4.3%	45.2% ±4.4%	72.8% ±3.9%	23.0% ±3.7%	4.2% ±1.8%
Normal distribution with high variance	11.6% ±2.8%	40.4% ±4.3%	48.0% ±4.4%	73.0% ±3.9%	18.8% ±3.4%	8.2% ±2.4%
Gamma distribution	54.8% ±4.4%	9.0% ±2.5%	26.2% ±4.2%	8.4% ±2.4%	86.4% ±3.0%	5.2% ±1.9%
Geometric distribution	32.8% ±4.1%	20.4% ±3.5%	46.8% ±4.4%	20.2% ±3.5%	77.4% ±3.7%	2.4% ±1.3%
Student distribution	52.0% ±4.4%	39.0% ±4.3%	9.0% ±2.5%	40.6% ±4.3%	59.0% ±4.3%	0.04% ±0.6%
Log-normal distribution	28.4% ±4.0%	27.4% ±3.9%	44.2% ±4.4%	36.8% ±4.2%	51.2% ±4.4%	12.0% ±2.8%

Source: calculations of the author

The logic of choosing the regression for the regression equation was described above. Here the post-evaluation analysis is presented only for OLS model, because it is enough for confirmation of economic grounds of the regression equation that was estimated.

The regression equation estimated by OLS has the following form:

$$medv_i = -5.520 - 0.142 * crim_i + 0.031 * zn_i + 3.22 * chas_i - 0.077 * age_i + 6.947 * rm_i - 0.973 * dis_i - 0.852 * ptratio_i + 0.016 * b_i$$

The model has appropriate quality. All coefficients except coefficient before zn (proxy land lots for sale) are statistically significant at confidence level 0.05. All coefficients except coefficient before b (proxy of share of black population) have signs as it were expected and discussed in part 4. Adjusted R-squared is equal to 0.65 that means that the model explains properly 65% of dispersion in the data. F-statistics is equal to 117 and has p-value<0.0001.

Nevertheless, the analysis of residuals shows that data sample contains several outliers and emphasise that dependent variable does not distribute normally. Firstly, residuals plot revealed that residuals are not distributed like white noise. Secondly, the residuals are not distributed normally. Thirdly, the density residuals plot reveals a long right tail. All these findings shows that OLS in estimated form is not the best model. One can expect that in conducted experiments forecast based on OLS model will have poor quality in terms of RMSE. Nevertheless, all in all, the simple OLS model shows that there are potentially real strong relationships between dependent variable and regressors and the regression equation may have economic grounds beyond it. It is enough base for the experiments, and their results can be considered further.

The results of experiments on Boston Housing data are depicted in table 2. Bayes quantile regression has the largest share of best forecasts, and quantile regression has the second-best largest of forecasts. What is more, OLS model has the lowest share of best forecasts that was partly noticed during analysis of residuals quality for linear model. Generally, this pattern looks similar with results that were received on generated sample from geometric distribution. Nevertheless, without additional analysis it is impossible to make conclusions on how similar these two results are. All in all, Bayes quantile regression model shows the best results in terms of forecasting.

Table 2. Results of experiments on real data.

Data	Share of iteration where the model is the best in terms of RMSE			Share of iteration where the model is the second-best in terms of RMSE		
	OLS	QR	BQR	OLS	QR	BQR
Boston Housing Data	14.5% ±3.1%	37.1% ±4.2%	48.4% ±4.4%	28.0% ±3.9%	68.9% ±4.1%	3.1% ±1.5%

Source: calculations of the author

The results on generated and on real data is shortly discussed below. The results revealed that Bayes quantile regression created better forecast in these experiments. Such result was expected due to methodological advantages that Bayes models have over OLS and quantile regression. Quantile regression is in the majority of cases is second-best model in terms of forecast in these experiments. Such result was also expected due to methodological advantages of quantile model over OLS. Moreover, even for sample with

normal distributions of errors, OLS has not the best forecast in these experiments. One can have a question how resistant these results may be.

The stability of abovementioned results can be checked with increasing number of iterations in the experiments and due to more wide range of distributions that may be used in the analysis. We expect that these results will be stable with increasing number of iterations due to at least two following reasons. First reason is the fact that for majority of distribution types, that we used, the best model can be uniquely identified because corresponding confidence intervals do not overlap. The second reason is that when the same experiments were conducted for smaller number of iterations firstly, the proportion of share best and second-best models were approximately the same. It was generally only one difference - the confidence intervals overlapped in case of 100 iteration for majority of experiments. All in all, the same methodology can be implemented on larger number of data points, with larger number of iterations and for more distribution types in order to explore the stability of results, but these questions now are beyond of the scope of this thesis.

Conclusion

In this thesis the series of experiments were held in order to experimentally identify the best forecasts in terms of RMSE between quantile regression models based on traditional and Bayes approach. OLS model was used as a baseline. The same methodology was implemented both for samples of generated data and on Boston Housing data.

The main results on generated data can be summarised in three key point as follows. Firstly, Bayes quantile regression model has the largest share of the best forecasts for 4 types of distributions (for normal distribution both Bayes quantile model and standard quantile model are the best as 95% CI overlap). Secondly, some results look similar for group of distribution. For instance, the Bayes quantile model has the largest share of the best model for geometric and log-normal distributions, and the quantile regression model has largest share of the second-best forecasts for these distributions. Thirdly, OLS model has the largest share of the best forecasts for gamma and student distributions, and quantile regression model has the largest share of second-best forecasts for these distributions.

The results of experiments on real data reveals that Bayes quantile model has the largest share of best forecasts, and quantile regression has the second-best largest of forecasts. What is more, OLS model has the lowest share of best forecasts that was partly noticed during analysis of residuals quality for linear model. Generally, this pattern looks similar with results that were received on generated sample from geometric distribution. All in all, Bayes quantile regression model shows the best results in terms of forecasting both on generated and real data.

Abovementioned results partially confirm key hypotheses of this research. First hypothesis is rejected since OLS model does not have the largest share of the best forecasts for samples from normal distribution. Second hypothesis is confirmed, because Bayes quantile model has the largest share of the best forecasts for

4 out of 6 types of distributions. Third hypothesis is rejected as the shape of error distribution and data plots cannot determine what model will have the largest share of the best or second-best forecasts. The more formal analysis of distribution shape is needed to connect distribution type and results of the experiments.

In conclusion, the results of presented experiments shows that clear theoretical advantages of the model do not strictly determine its best results on data. Any theory, no matter how straightforward it is, should be confirmed during implementation on data. Real data do not often used to estimate 1 500 models as it was done in these experiments. When larger number of experiments are needed for stable conclusions, generated samples from distributions with control of parameters can be used like it was done in this paper. We hope that this research will be able to find its own place in wide discussion of advantages of quantile models based on classical and Bayes approaches.