Analysing Data with GAMLSS: A Case Study on Fitting Distributions, Centile Estimation, and Predictive Modelling

Name	**Tonny K Podiyan**
Student ID	**21052135**
Email ID	**Tok0165@my.londonmet.ac.uk**
Module Code	**MA7007**
Module Title	**Statistical Modelling & Forecasting**
Date	**13th May 2023**

Contents

Table of Figures

Tables

Introduction:

Statistical modeling and forecasting techniques have been widely used in various fields for many decades. These techniques are particularly important in data-driven decision-making, where data analysis and modeling are critical to extracting insights and making informed decisions. With the advent of modern computing technologies and powerful statistical software, sophisticated statistical models can now be built and used for real-world applications with ease.

One popular technique for statistical modeling is GAMLSS, which allows for the estimation of multiple parameters for various distributions, including both fixed and random effects. GAMLSS has been used extensively in fields such as ecology, biology, and environmental sciences, among others.

In this report, we will use GAMLSS in R to analyze three different datasets and provide a detailed understanding of the results. By leveraging statistical modeling and forecasting techniques such as GAMLSS, we can gain valuable insights into complex data structures and make informed decisions in a wide range of fields.

Dutch Boys' BMI Data from Fourth Growth Study:

The data is obtained from the Fourth Dutch Growth Study conducted by Fredriks et al. (2000a, 2000b), which is a study that examines the growth and progress of the Dutch population aged between 0 and 21 years. It measures several variables including height, weight, head circumference, and age for a sample of 7482 males and 7018 females [1]. Currently we only have access to the BMI data for boys from the Dutch population.

R data file : **dbbmi** in package **gamlss.data** of dimensions 7294 × 2

age: the age of the individual in years

bmi: the body mass index of the individual, calculated as weight (in kgs) divided by height (in squared meters)

purpose: fit different distributions to the data set and find the best fitting distribution.

Data Analysis & Visualisation:

We used the 'gamlss' package in R to fit a statistical model to the 'dbbmi' dataset. Specifically, we subsetted the data to only include observations where the age was between 15 and 16 years old and extracted the 'bmi' variable. We then used the 'MASS' package to create a histogram of the 'bmi' values with 19 bins, using the 'truehist' function. We determined the number of bins using the Scott's rule. This method calculates the number of bins as 3.5 * sd(x) / (n^(1/3)), where sd is the standard deviation. This allowed us to visualize the distribution of BMIs in our selected age range

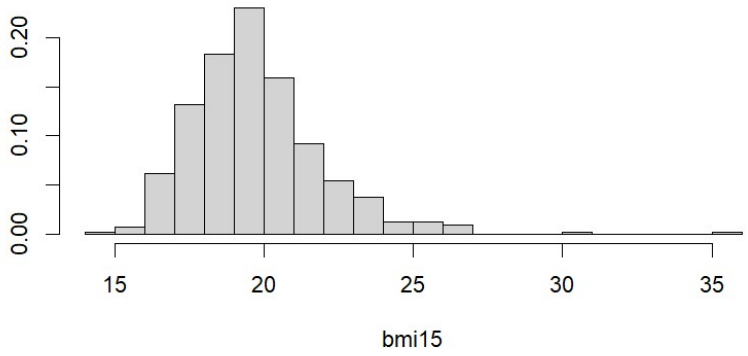


Figure 1: GAIC & SBC scores for fitted models.

Fitting statistical models:

For analysing the data, we fitted 4 different distributions :

1. **Normal distribution(mno):** It has a bell-shaped curve, which means that most of the data falls within a certain range and outliers are rare.
2. **Gamma distribution(mga):** It's a continuous distribution that can take on a variety of shapes depending on its parameters. It's commonly used to model the time it takes for something to occur.
3. **Inverse Gaussian distribution(mig):** It is a type of probability distribution that is like the normal distribution but has a longer tail on one side. It's commonly used to model data that has a skewed distribution.
4. **Box-Cox-Cole-Green distribution(mbccg) :** It is a family of probability distributions that includes the normal, gamma, and inverse Gaussian distributions as special cases. It's used to model a wide range of data, including continuous and discrete data, and can be customized to fit the specific needs of a particular study.

Each of these distributions has different shapes and properties, which can be used to model different types of data.

Model evaluation:

After fitting the four distributions we checked the GAIC (Generalized Akaike Information Criterion) values for all the fitted models, which is a measure of the quality of the fit. Among the four distributions as shown in table 1 , the BCCG distribution has the lowest GAIC and SBC values, indicating that it provides the best fit to the data. Additionally, the plot of the fitted distribution overlaid on the histogram of the data also suggests that the BCCG distribution provides a good fit to the data.

Fitted Model	Degree of Freedom	GAIC Scores	SBC Scores
mno	3	1805.973	1817.97
mga	3	1770.708	1782.705
mig	3	1757.979	1769.976
mbccg	4	1731.400	1747.396

Table 1: GAIC & SBC scores for fitted models.

Based on the summaries, the mean of each set of residuals is close to zero, and the variance is close to one, indicating that the models are reasonably well-fit. However as per the output shown below, the coefficients of skewness and kurtosis are quite large for the first three models(mno,mga,mig) indicating that the distributions may be somewhat skewed and/or have heavy tails. The Filliben correlation coefficient is close to 1 for all four models, indicating that the residuals are uncorrelated with each other.

```
*******************************************************************
Summary of the Quantile Residuals for mno
                           mean   =  1.609495e-09
                       variance   =  1.002488
               coef. of skewness  =  1.495521
               coef. of kurtosis  =  9.033368
Filliben correlation coefficient  =  0.9578562
*******************************************************************
Summary of the Quantile Residuals for mga
                           mean   =  -0.0003220082
                       variance   =  1.002463
               coef. of skewness  =  1.000429
               coef. of kurtosis  =  6.217994
Filliben correlation coefficient  =  0.9761185
```

****

Summary of the Quantile Residuals for mig

mean = -0.001198336

variance = 1.002305

coef. of skewness = 0.8022175

coef. of kurtosis = 5.386602

Filliben correlation coefficient = 0.982457

****

Summary of the Quantile Residuals for mbccg

mean = -0.0003805448

variance = 1.002386

coef. of skewness = -0.03605923

coef. of kurtosis = **3.539315**

Filliben correlation coefficient = 0.9971223

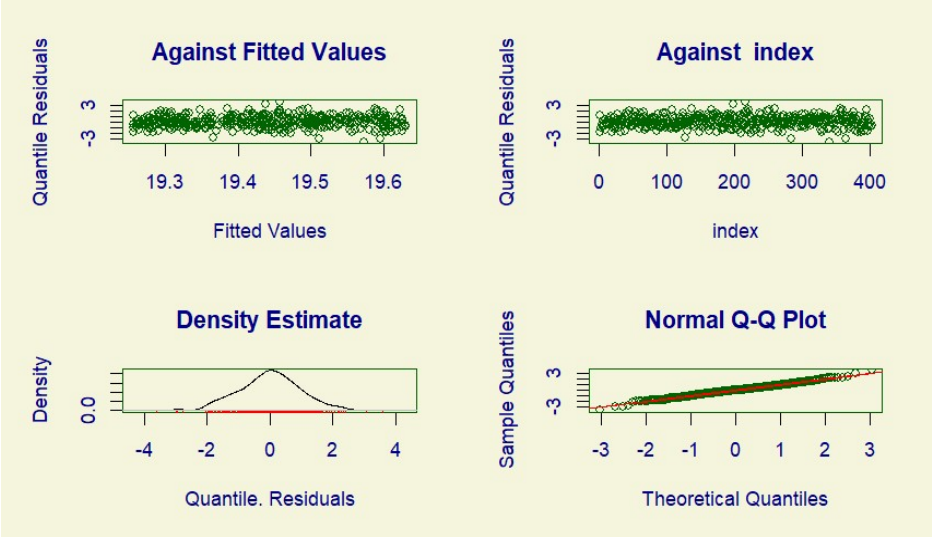


Figure 2: Residual Plot of model mno.

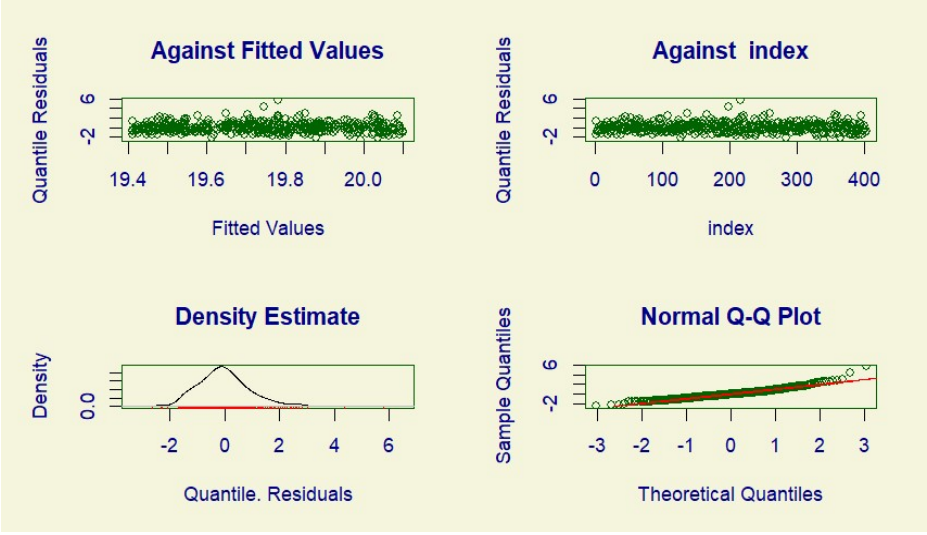


Figure 3: Residual Plot of model mga.

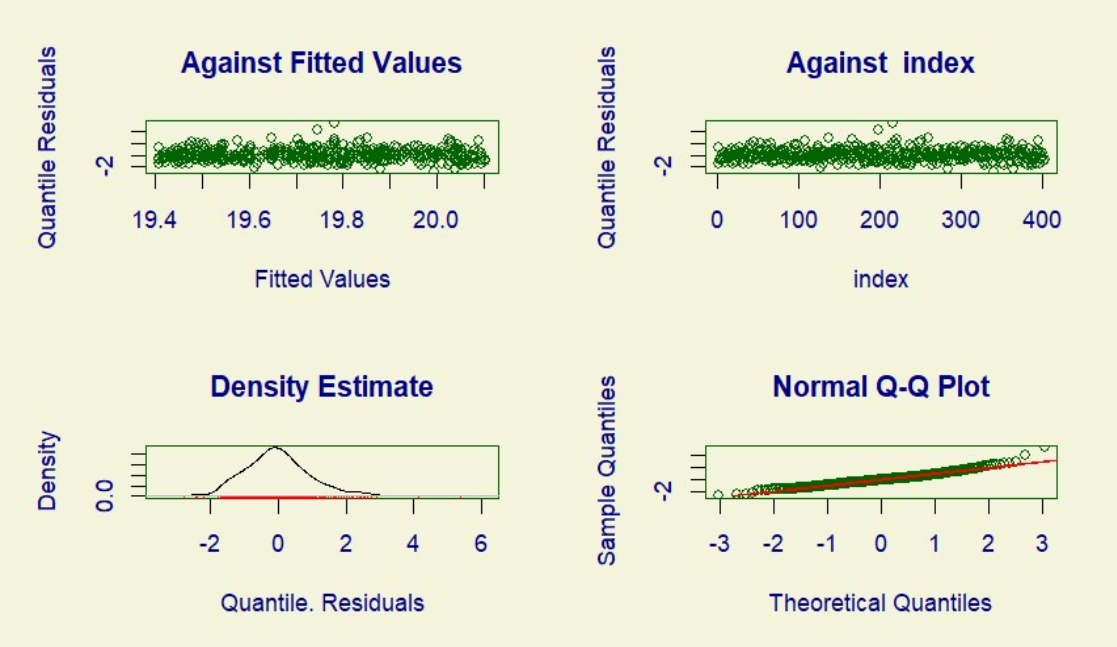


Figure 4: Residual Plot of model mig.

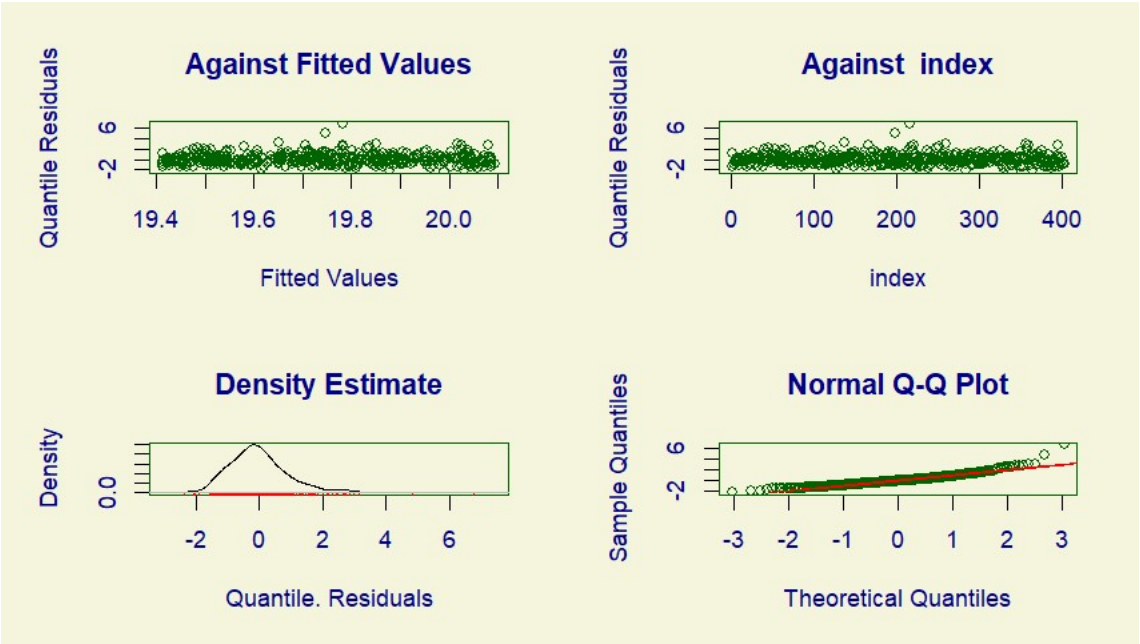


Figure 5: Residual Plot of model mbccg.

(d) We have created four plots of the fitted distribution and histogram of the BMI data for each of the four distributions tested. The quantile residuals against fitted values and against index are randomly scattered between -2 and 2 for all 4 distributions suggesting that they are well-behaved and do not exhibit any significant deviations from the expected distribution. In the **Box-Cox-Cole-Green distribution(mbccg)** model the normal Q-Q plot line best captures most of the points as compared to the other distributions , is symmetrical and does not curve upwards or downwards indicating that the dataset is normally distributed.

(e) The fitted parameter values for the final chosen BCCG model can be obtained by calling the summary() function on the mbccg object.

Conclusion :

The final chosen model is a Box-Cox-Cole-Green distribution with a log link function for mu and sigma, and an identity link function for nu. The fitted parameter values are as follows:

Mu link function:

- Intercept: 2.66456
- Age: 0.01954

Sigma link function:

- Intercept: -2.27441

Nu link function:

- Intercept: -1.73

The model was fit using the RS() fitting method and there were 403 observations in the fit with 4 degrees of freedom. The residual degrees of freedom were 399 and the global deviance, AIC, and SBC were 1723.4, 1731.4, and 1747.396, respectively.

Handgrip strength of English schoolboys:

Cohen et al. (2010) conducted a study on the handgrip strength of English school children, examining its relationship with gender and age [2]. The original dataset consists of 3766 observations of boys. For this analysis, we have selected a subset of 1000 observations. The grip and age variables of the selected sample are stored in the gamlss.data package.

R data file : **grip** in package **gamlss.data** of dimensions 3766 × 2

age: the age of the individual in years

grip: the handgrip strength

seed: 1088

purpose: To estimate centiles for a given dataset.

Statistical models:

The three models used in this coursework are all distributional models based on the generalized additive models for location, scale, and shape (GAMLSS) framework.

The first model used is the Box-Cox Cole and Green (BCCG) model, which assumes that the response variable follows a four-parameter generalized lambda distribution (GEL) and includes age as a predictor variable. This model is useful when the response variable has a skewed distribution, which is common in many biomedical datasets.

The second model used is the Box-Cox transformed (BCT) model, which assumes that the response variable follows a three-parameter GEL with a power transformation on the response variable. This model is useful when the variance of the response variable changes based on the mean.

The third model used is the Box-Cox power exponential (BCPE) model, which assumes that the response variable follows a five-parameter GEL with both a power transformation on the response variable and an exponential

transformation on the scale parameter. This model is useful when there is more variation in the response variable than predicted by the mean.

Relationship between Handgrip Strength and Age:

The plot of grip against age is shown in Figure 6. As can be seen from the plot, the relationship between grip strength and age appears to be relatively linear, with no clear patterns of increasing or decreasing variability. Since age is the independent variable in this analysis, and there are no indications of nonlinearity in the data, it is not necessary to perform a power transformation on either variable.

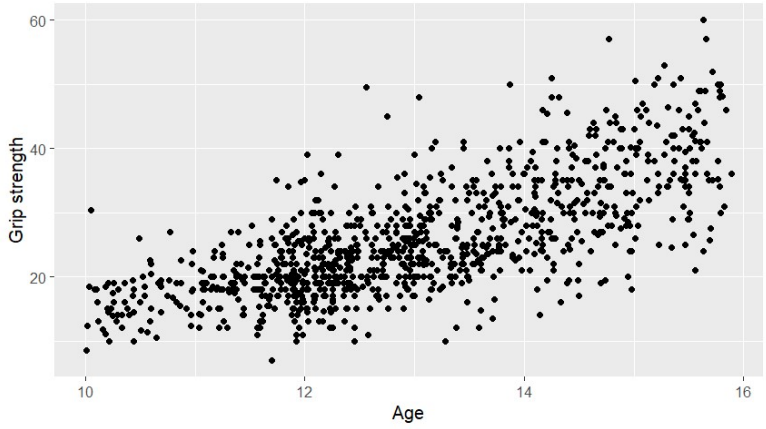


Figure 6: Plot of Grip against Age.

We used the LMS method to fit the gbccg model to our data set, and then we calculated the effective degrees of freedom for the smooth term used to model the mean by using the edf() function. Specifically, the smooth function pb(age) was used to model the mean, and the output of edf() for this term was 4.332988, indicating that 4.332988 degrees of freedom were used for smoothing.

After using the LMS model to obtain fitted values, we used them as starting values for fitting the BCT and BCPE distributions to the data. The effective degrees of freedom (EDF) for the smooth term that was used to model the mean using the BCT and BCPE distributions are 4.463426 and 4.524385, respectively.

Model evaluation:

To compare the three models, we used the Generalised Akaike Information Criterion (GAIC), which considers the complexity of the model and the goodness of fit. The GAIC for the BCCG model was found to be 6344.987, while the GAIC for the BCT and BCPE models were 6337.515 and 6336.784, respectively. Comparing the GAIC values, we see that the BCT and BCPE models have lower GAIC values than the BCCG model, indicating that they are better models for the given data. The difference in GAIC values between the BCT and BCPE models is small, suggesting that both models are equally good, and the choice between them may depend on other factors such as interpretability.

Fitted Model	GAIC Scores
BCT	6337.515
BCPE	6336.784
BCCG	6344.987

Table 2: GAIC scores for fitted models.

Fitted Model Parameters:

After plotting the fitted parameters, the graphs of mu vs age and nu vs age reveal a consistent and linear growth with age for the BCCG, BCT, and BCPE models as shown in Figure 7, 8 and 9. However, the sigma vs age plot displays a decline for gbccg and gbcp but an upward trend for gbct. Meanwhile, the tau vs age plot exhibits an upward trend for gbct and gbcp, with gbct showing a notably steeper increase. As the BCCG model lacks a random effect term that corresponds to a subject-specific deviation from the overall model, a tau plot is not available for gbccg. The tau plot illustrates the estimated variance of the random effects, but without random effects in the model, there is no tau plot. These observations imply that all three models exhibit a comparable increase in grip strength with age, but they differ in their estimates of variability. Moreover, gbct demonstrates a more distinct increase in between-subject variability with age.

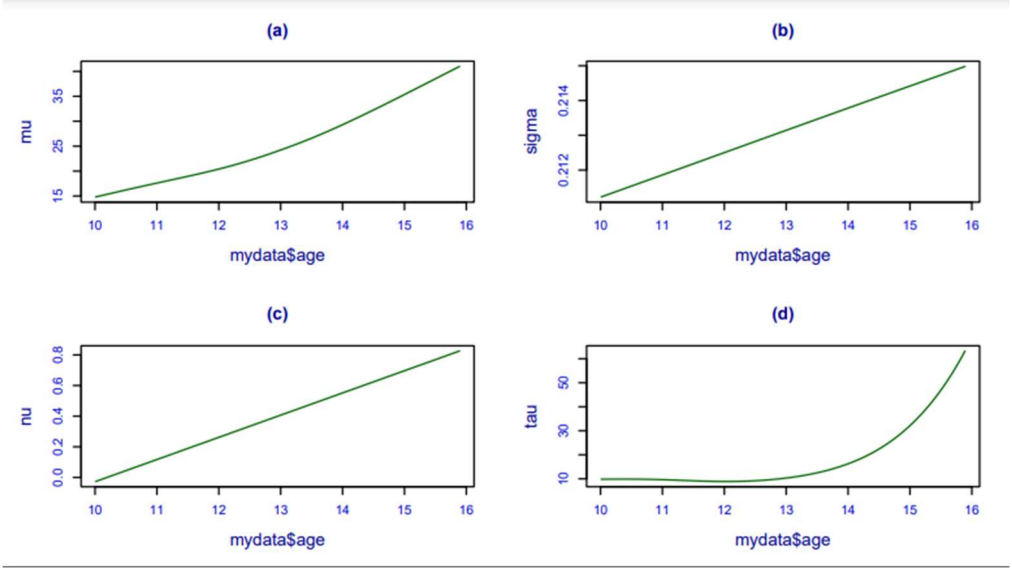


Figure 7:Fitted values plot for BCT model.

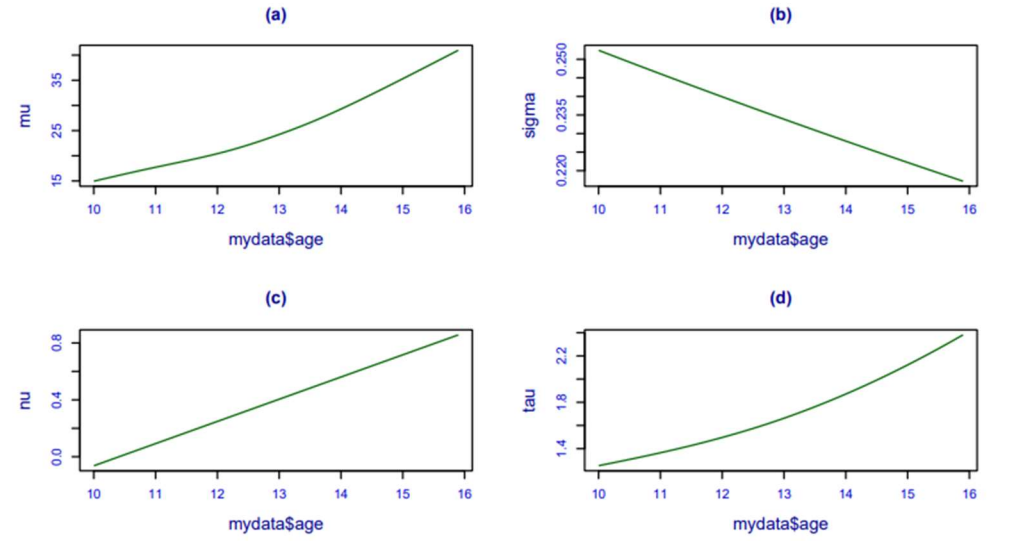


Figure 8: Fitted values plot for BCP model.

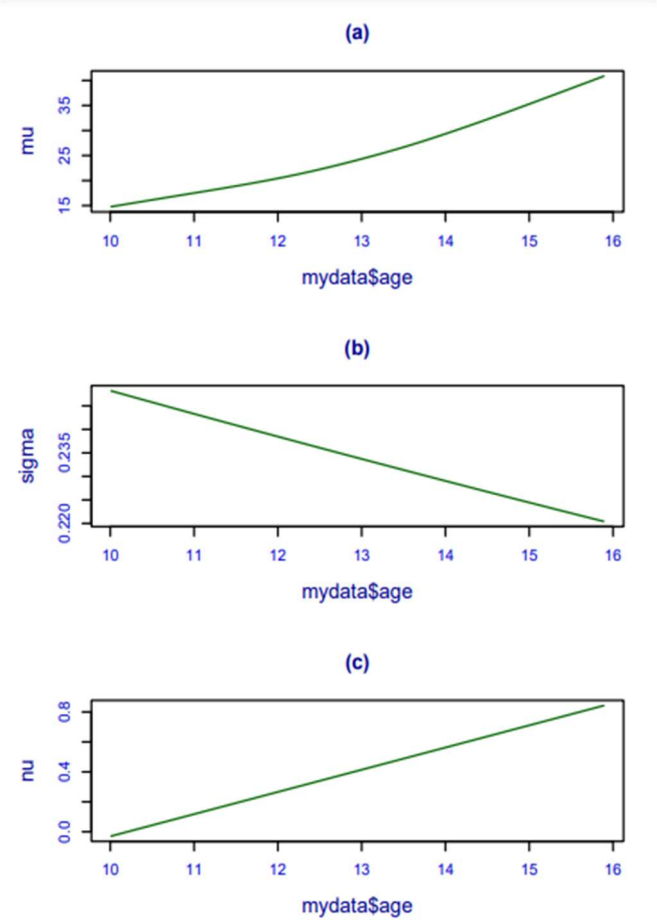
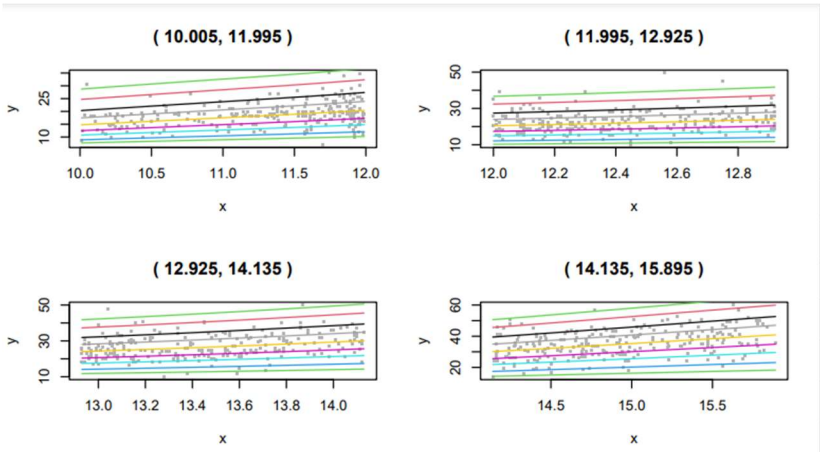


Figure 9: Fitted values plot for BCCG model.

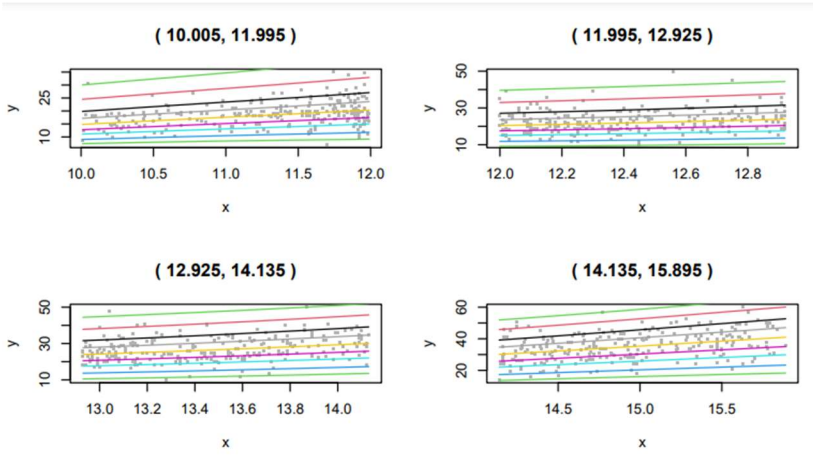
Centile Plots:

Now we created a centile plot for each of the three models (GBCCG, GBCT, and BCPE) using the centiles.split() function.



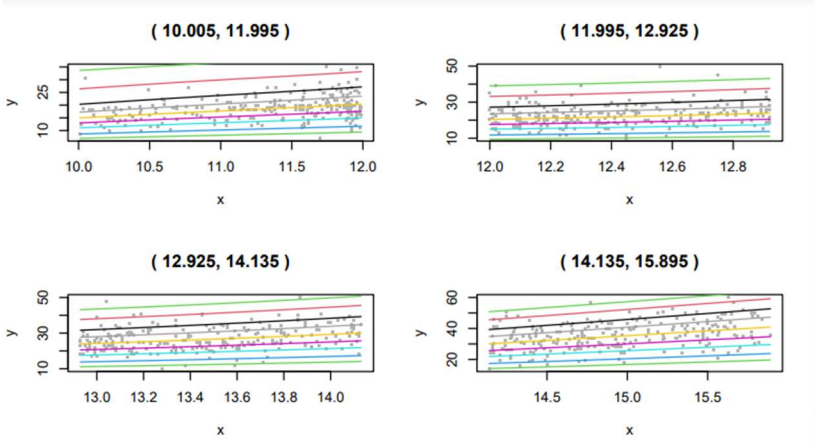
	10.005 to 11.995	11.995 to 12.925	12.925 to 14.135	14.135 to 15.895
0.4	0.8	0.7936508	1.209677	0.4
2	2.8	1.9841270	2.016129	2.8
10	10.4	9.9206349	7.661290	9.6
25	23.2	23.4126984	24.596774	24.4
50	48.8	48.8095238	54.032258	49.6
75	76.4	76.9841270	74.193548	72.8
90	92.8	90.4761905	91.532258	90.8
98	98.4	97.2222222	97.580645	97.6
99.6	99.6	98.4126984	99.193548	99.6

Figure 10: Centile Plot for BCCG model.



	10.005 to 11.995	11.995 to 12.925	12.925 to 14.135	14.135 to 15.895
0.4	0.4	0.000000	0.8064516	0.0
2	2.4	1.984127	2.0161290	2.8
10	10.8	10.714286	7.6612903	10.4
25	24.4	24.206349	25.8064516	24.8
50	48.8	48.412698	53.2258065	50.0
75	75.2	75.793651	72.5806452	72.4
90	92.0	90.079365	91.1290323	90.0
98	98.4	97.619048	97.5806452	98.0
99.6	99.6	99.206349	99.5967742	99.6

Figure 11: Centile Plot for BCT model.



	10.005 to 11.995	11.995 to 12.925	12.925 to 14.135	14.135 to 15.895
0.4	0.4	0.3968254	1.209677	0.4
2	2.4	1.9841270	2.016129	2.8
10	10.8	10.7142857	7.661290	9.6
25	25.6	24.2063492	25.403226	24.4
50	49.2	48.4126984	52.016129	50.0
75	75.2	75.3968254	72.580645	72.8
90	92.4	90.0793651	91.129032	90.8
98	98.4	97.6190476	97.580645	97.6
99.6	100.0	99.2063492	99.193548	99.6

Figure 12: Centile Plot for BCP model

The above centile visualisation shows the centile values of three different models (BCCG, BCT, and GBCP) for different age intervals. In the centiles table each row represents a specific centile (e.g., 50th percentile), and each column represents a specific age interval.

From the centiles output, we can see that the age ranges for the three models are the same: 10.005 to 11.995, 11.995 to 12.925, 12.925 to 14.135, and 14.135 to 15.895. To compare the grip strength distribution across age ranges, we can look at the centiles for each age range and model.

For example, let's look at the 50th centile (median) for each age range and model:

- GBCCG: 49.2, 50.00, 52.41, 48
- BCP : 49.2, 48.41, 52.01, 50
- BCT : 48.8, 48.41, 53.22, 50

We can see that the median grip strength for each age range is similar across all three models.

We also examined the residual plots for the three models and observed that the quantile residuals are randomly scattered between -3 and 3 for the GBCT and GBCP models, and between -4 and 4 for the GBCCG model, both against fitted values and against index. Additionally, the density estimates are normally distributed for all three models. The normal QQ plot captures most of the points in the GBCT and GBCP models, while there is some deviation in the GBCCG model.

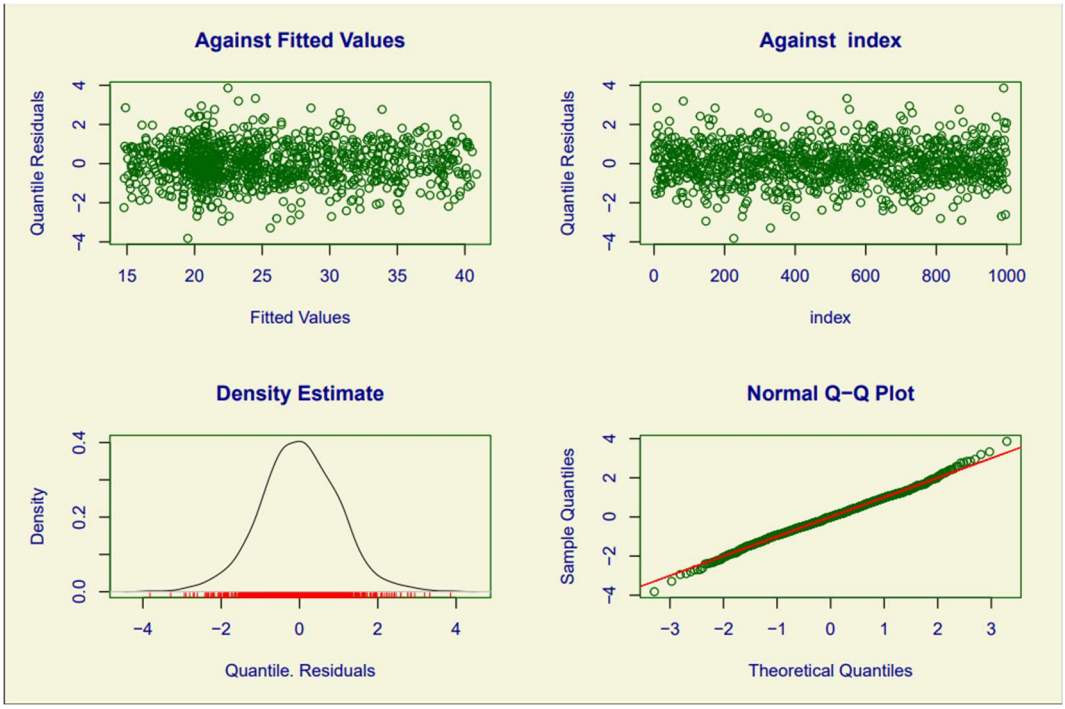


Figure 13: Residual Plot of BCCG model.

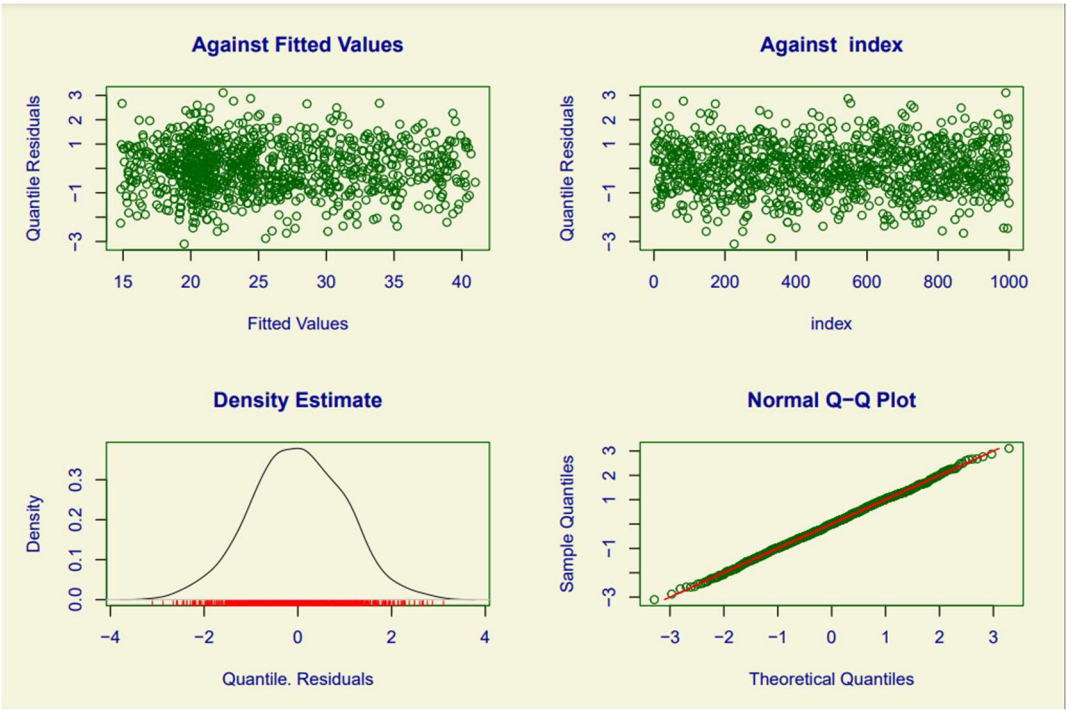


Figure 14: Residual Plot of BCT model

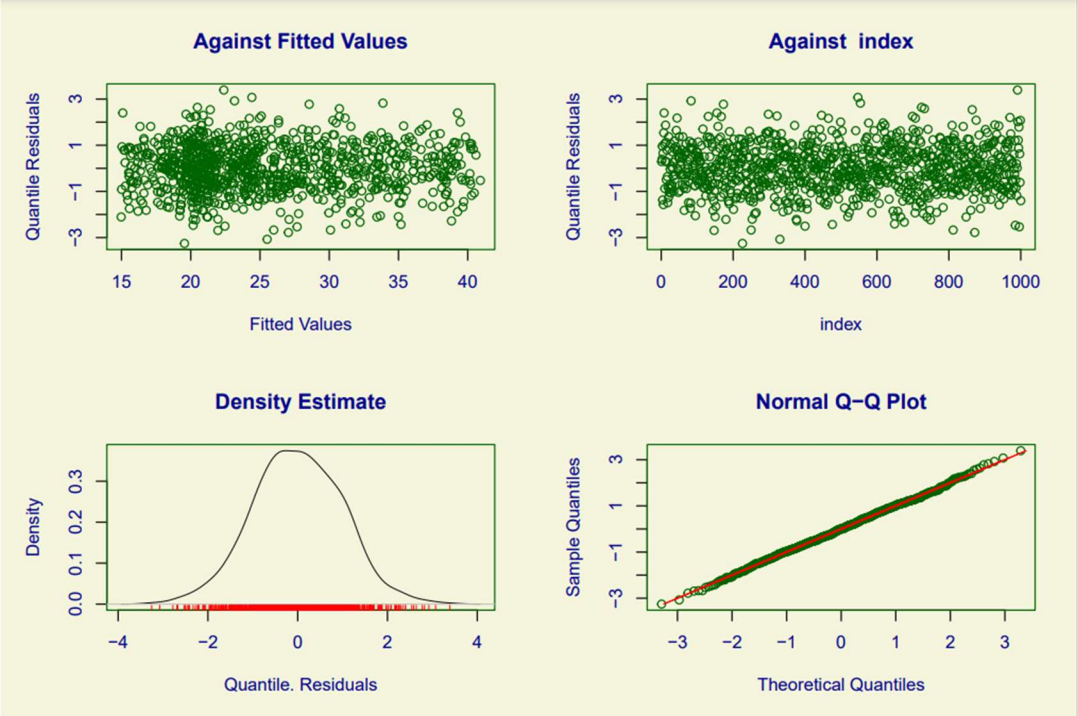


Figure 15: Residual Plot of BCPE model .

Conclusion:

Based on our analysis, the best model among the three options (BCCG, BCT, and BCPE) would be either BCT or BCPE, as both models have lower GAIC values than the BCCG model, indicating better performance for the given data. The difference in GAIC values between the BCT and BCPE models is small, suggesting that both models are equally good.

Furthermore, the residual plots for the BCT and BCPE models indicate that they are both good fits for the data, with randomly scattered quantile residuals and normally distributed density estimates. Therefore, either model could be chosen based on interpretability or other relevant factors.

King County, Washington State Housing Price Dataset:

The housing market is an important aspect of any economy and understanding the factors that affect housing prices can provide valuable insights. In this project, we will be analysing a dataset of housing prices from King County, Washington State, USA. The dataset was obtained from GitHub and contains information on various features of houses sold between May 2014 to May 2015 [3]. The aim of this project is to build a model that accurately predicts housing prices and identify the key factors that influence the price of a house in King County. We will be using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to fit our models and make predictions.

Link to dataset : *https://github.com/Shreyas3108/house-price-prediction/raw/master/kc_house_data.csv*

Data file : **kc_house_data** from **GitHub** of dimensions 21613 × 21

seed: 1088

Id: is a unique identification number assigned to each property in the dataset.

Date: refers to the date when the house was sold.

Price: represents the sale price of the property.

Bedrooms: indicates the number of bedrooms in the house.

Bathrooms: indicates the number of bathrooms in the house.

Sqft_living: represents the square footage of the interior living space of the house.

Sqft_lot: represents the square footage of the lot on which the house is situated.

Floors: represents the number of floors in the house.

Waterfront: indicates whether the house has a waterfront view or not.

View: indicates the number of times the house has been viewed.

Condition: represents the overall condition of the house.

Grade: represents the overall grade given to the house, based on the King County grading system.

Sqft_above: represents the square footage of the interior living space that is above ground level.

Sqft_basement: represents the square footage of the interior living space that is below ground level.

Yr_**built**: represents the year when the house was built.

Yr_renovated: represents the year when the house was last renovated.

Zipcode: represents the postal code of the area where the house is located.

Lat: represents the latitude of the location where the house is located.

Long: represents the longitude of the location where the house is located.

Sqft_living15: represents the average square footage of the interior living space of the nearest 15 houses to the subject property.

Sqft_lot15: represents the average square footage of the lot on which the nearest 15 houses to the subject property are situated.

Purpose: We are analysing this specific dataset of housing prices in King County because it provides a comprehensive set of 21 variables with no missing values, allowing for a detailed exploration of the factors influencing housing prices in the area.

We have obtained the dataset from GitHub which is a reliable source and free from missing values, indicating high data quality. However, to ensure the quality of our analysis, we are going to perform thorough exploratory analysis and meticulous data cleaning to address outliers, inconsistencies, and potential biases that may exist in our dataset.

Data Cleaning:

We first load our dataset using the read.csv() function. We then convert the date column from a character format to a date format using the as.POSIXct() function. Next, we set the seed value to 1088. For this analysis we are using only 80 percent of the data.

We first dropped the unnecessary columns from the dataset using the "select" function from the "dplyr" package. We removed the "id", "date" and "yr_renovated" columns, as they were not relevant to predicting the housing prices. We also dropped "lat","long", "zipcode" as we are not conducting spatial analysis.

Next, we loaded the "corrplot" package and calculated the correlation matrix of the remaining variables. We observed that "sqft_living" and "sqft_above" were highly correlated, as both represented the total area of the house. Therefore, we dropped the "sqft_above" variable, along with "sqft_lot15," as we already had "sqft_lot in our dataset" Lastly, we selected the remaining variables to create a simplified dataset.

Data Visualization:

We are visualizing the data to gain insights into the relationships between different variables and the target variable "price".We first create a correlation matrix plot using the corrplot function to visualize the correlations between the remaining variables in the dataset after dropping the irrelevant ones.

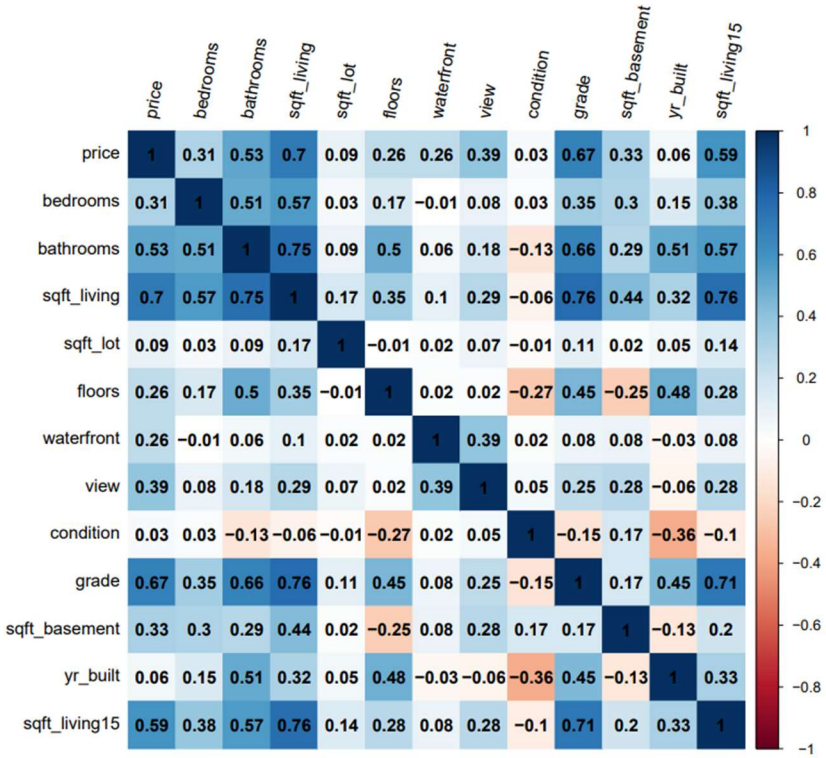


Figure 16: Correlation Plot.

Next, we create a histogram of the "price" variable to see the distribution of the target variable. From this distribution we can see that the price is heavily skewed towards the right, and we would need to do a log transformation of the target variable.

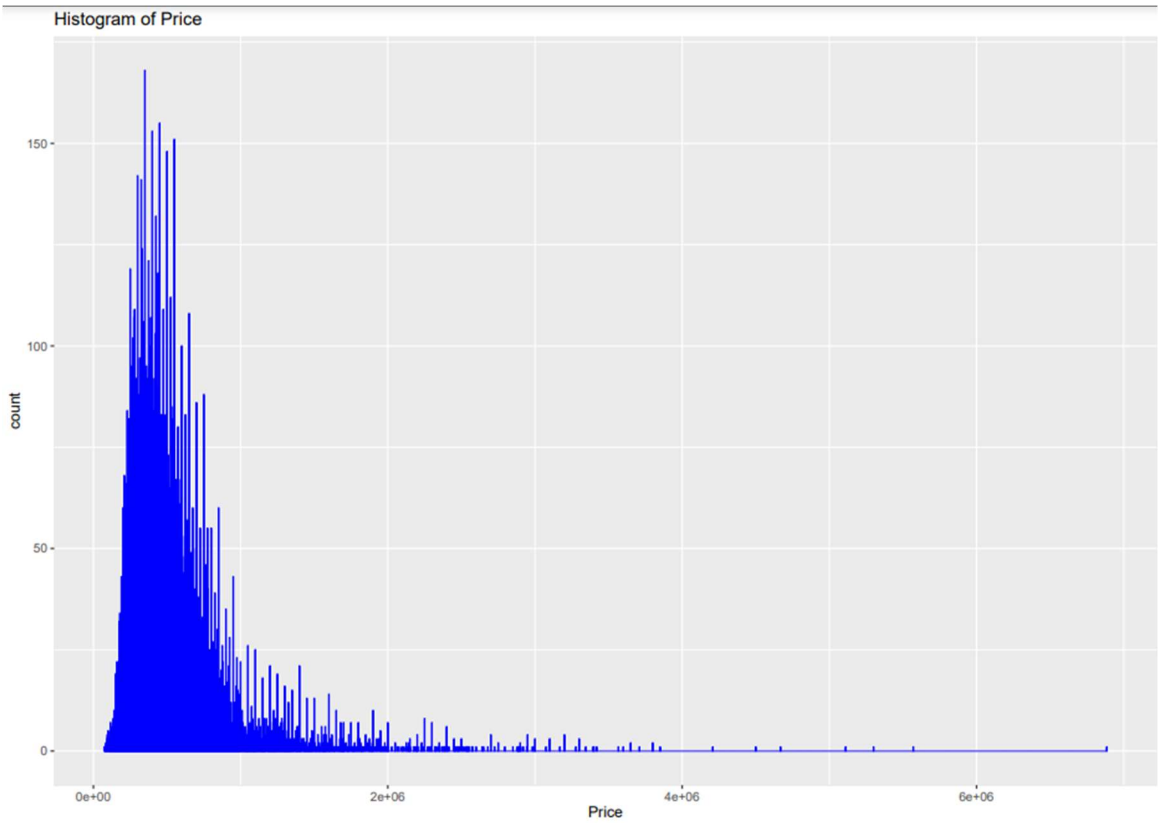


Figure 17: Histogram of Price

We also created a normal probability plot using the qqPlot function from the car package to see if the "price" variable follows a normal distribution.

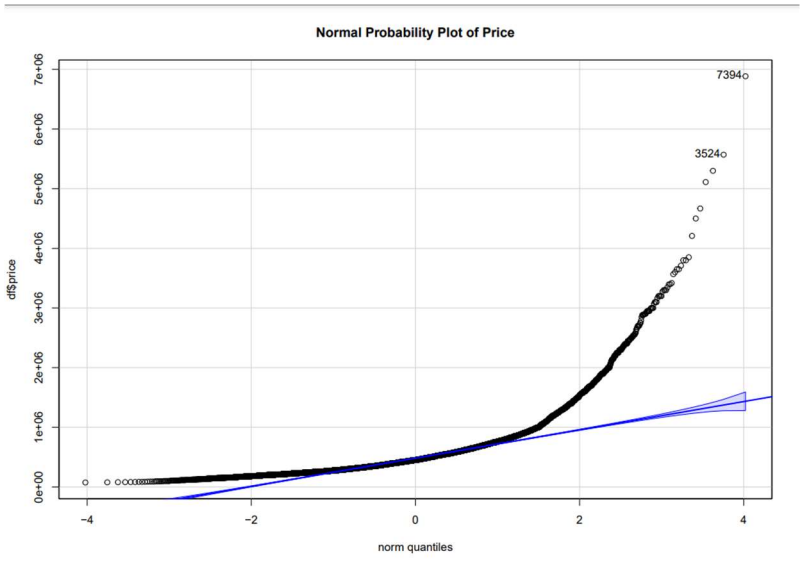


Figure 18: Probability Plot of Price

From the above plot its evident that the price variable deviates from the straight line, indicating that the price variable is not normally distributed, and we need to make transformations or select our model accordingly.

Finally, we create scatter plots between each of the remaining variables and the "price" variable using a for loop to visualize their relationships.

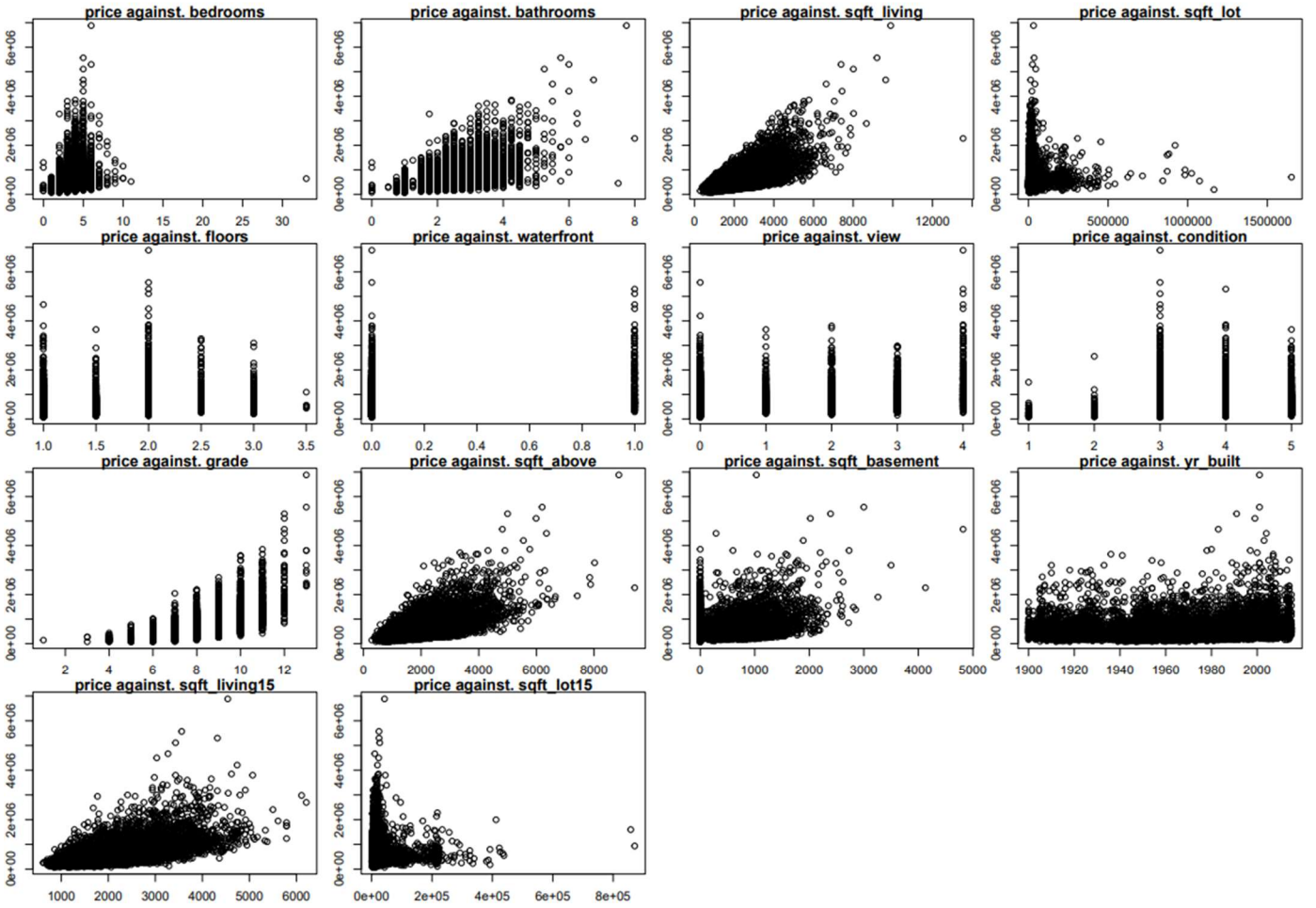


Figure 19: Scatter Plot of Price against Independent Variables

From the scatter plot we can see that there is linear relationship between Price and most of the independent variables(bathroom,sqft_living,grade,sqft_above,sqft_basement).

Other variables show random scatter and do not appear to have a clear linear relationship with the price. For example, the scatterplots for variables such as floors, condition, and view show no clear pattern.

Data Analysis:

First, we are loading the gamlss package and creating the formula for the regression model using the as.formula function. The formula is:

price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors + waterfront+ view+ condition+ grade+ sqft_basement + yr_built+sqft_living15

Here we are trying to predict the price of the house based on these independent variables.

Next, we are checking the skewness of the original price variable using the moments package's skewness function. We found that the original skewness of price variable was 3.612896 and after taking log transformation it got reduced to 0.4134558.

Thus , we applied a log transformation to the price variable to reduce the skewness. We are creating a new column in the data frame called "log_price," which contains the log-transformed values of the original price variable.

Next, we are fitting multiple Generalized Additive Models for Location, Scale, and Shape (GAMLSS) using different distributions for the response variable log_price.

1. The first model mbct uses the Box-Cox t distribution, which allows for skewness and kurtosis in the data.
2. The second model mga uses the Gamma distribution, which is often used for positive continuous data with a right-skewed distribution.
3. The third model mno uses the Normal distribution, which assumes that the data is normally distributed.
4. The fourth model mexp uses the Exponential distribution, which is often used for positive continuous data with a right-skewed distribution.

All these models use the same set of independent variables, we then fit values for each of the four models (BCT, GA, NO, EXP) and add them to the original dataset. Now to compare the performance of the four models and determine which one is the best fit for our data we calculate the Generalized Akaike Information Criterion (GAIC) for each of the four models using the GAIC() function.

Fitted Model	GAIC Scores
mbct	8467.09
mga	8700.65
mno	8489.73
mexp	123447.862

Table 3: GAIC scores for fitted models.

As shown in the above table the mbct model has the lowest GAIC score (8467.09), indicating that it has the best fit among the four models. The mno model also has a relatively low GAIC score (8489.73), while the mga model has a higher score (8700.65). The mexp model, on the other hand, has a much higher GAIC score (123447.862), indicating that it is a poor fit for the data.

We also plotted the above four fitted distributions and obtained their respective summaries of the quantile residuals. The distribution fitted from mbct has a mean close to zero and a slightly larger variance, indicating a reasonable fit. It is nearly symmetric with a skewness coefficient close to zero, but with heavier tails compared to a normal distribution as indicated by the kurtosis coefficient being above three. The Filliben correlation coefficient is close to one, showing a strong linear relationship between ordered residuals and theoretical quantiles. Meanwhile, the other three fitted distributions, mga, mno, and mexp, have means and variances close to one, indicating reasonable fits. However, their non-zero skewness and kurtosis coefficients suggest non-normal distributions. Their Filliben correlation coefficients are slightly smaller than one, indicating weaker linear relationships between ordered residuals and theoretical quantiles. Moreover, the deviation is higher in the qq plot for the mexp model compared to the other three models, and in the plot of quantile residuals against fitted values and against index, the points for mbct, mga, and mno scatter between -4 and 4, while for the mexp model, it's between 0.25 and 0.40.

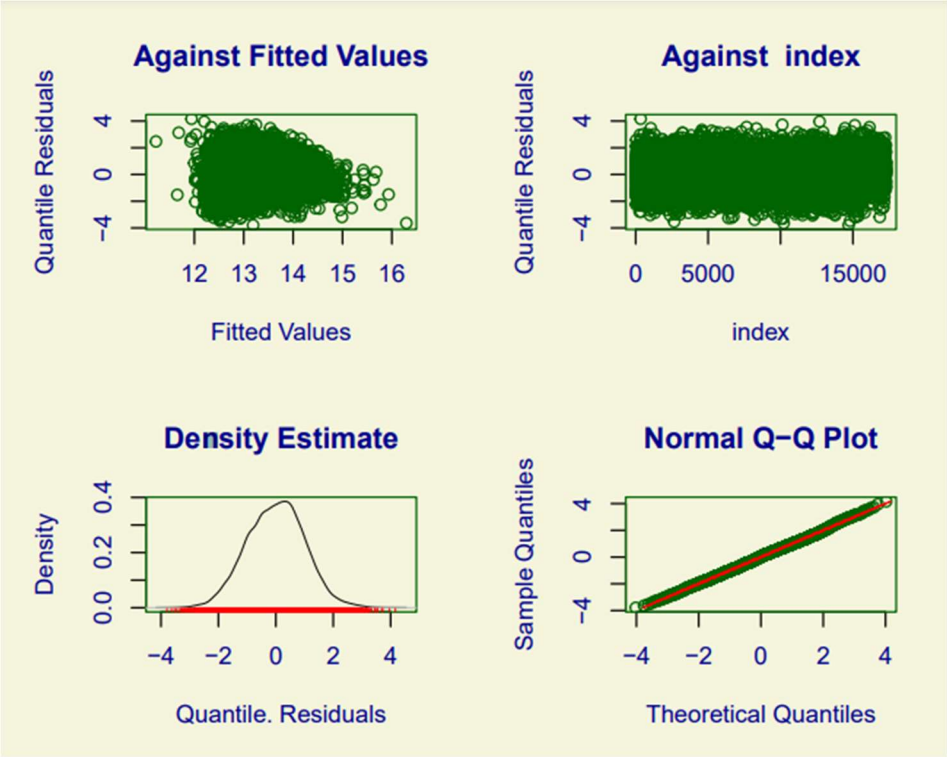


Figure 20: Residual Plot of MBCT model.

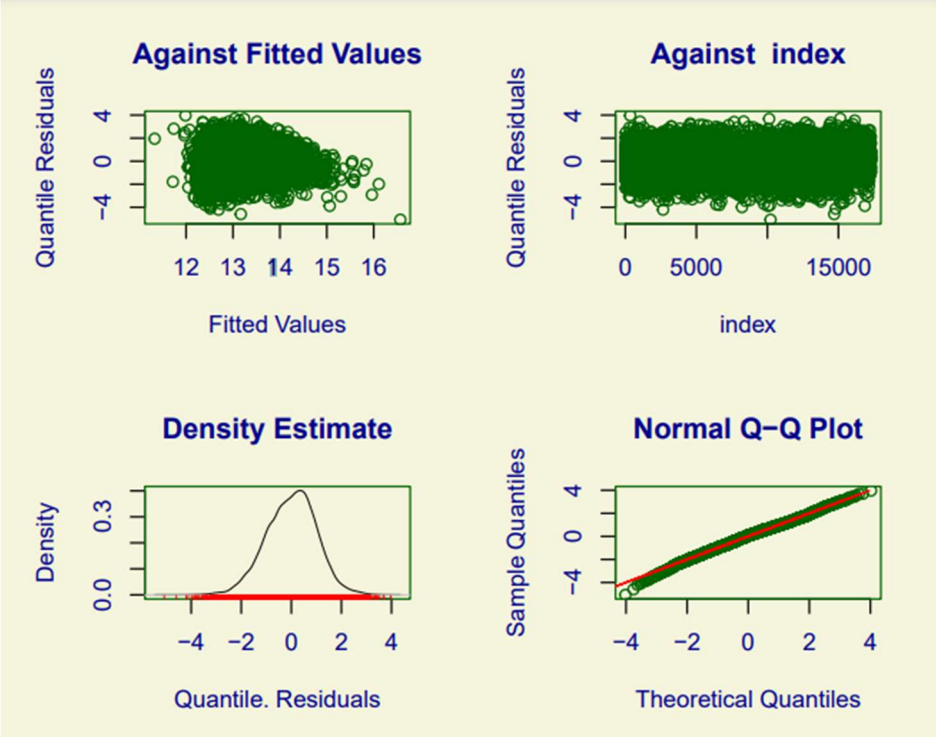


Figure 21: Residual Plot of MGA model.

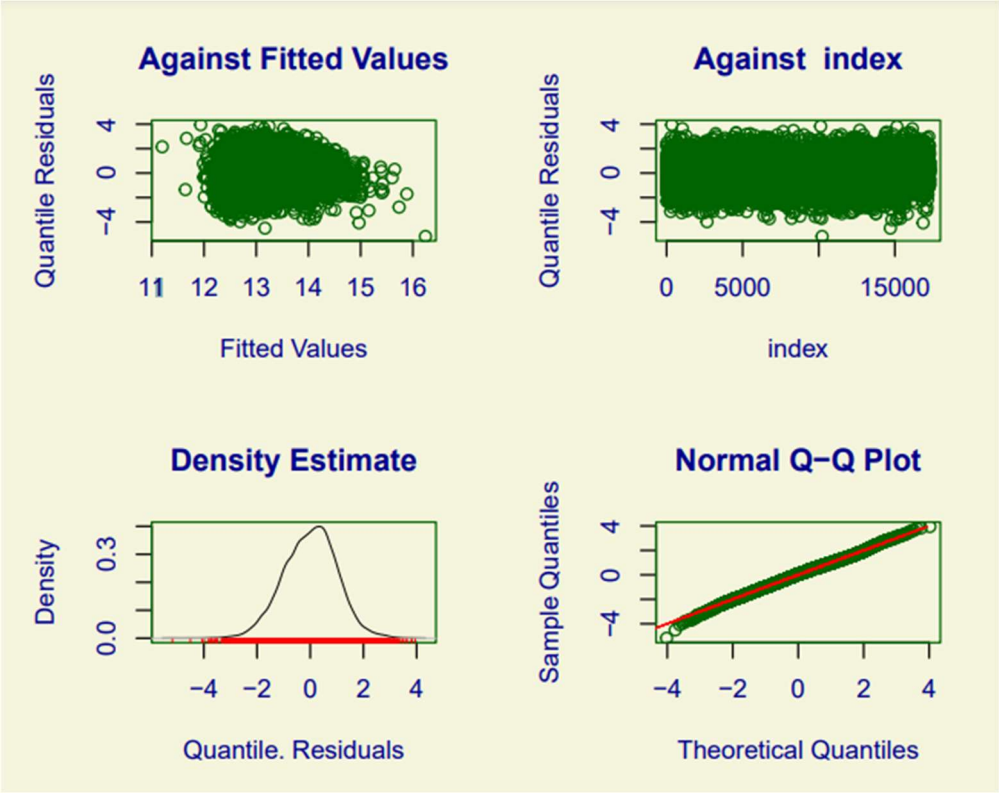


Figure 22: Residual Plot of MNO model

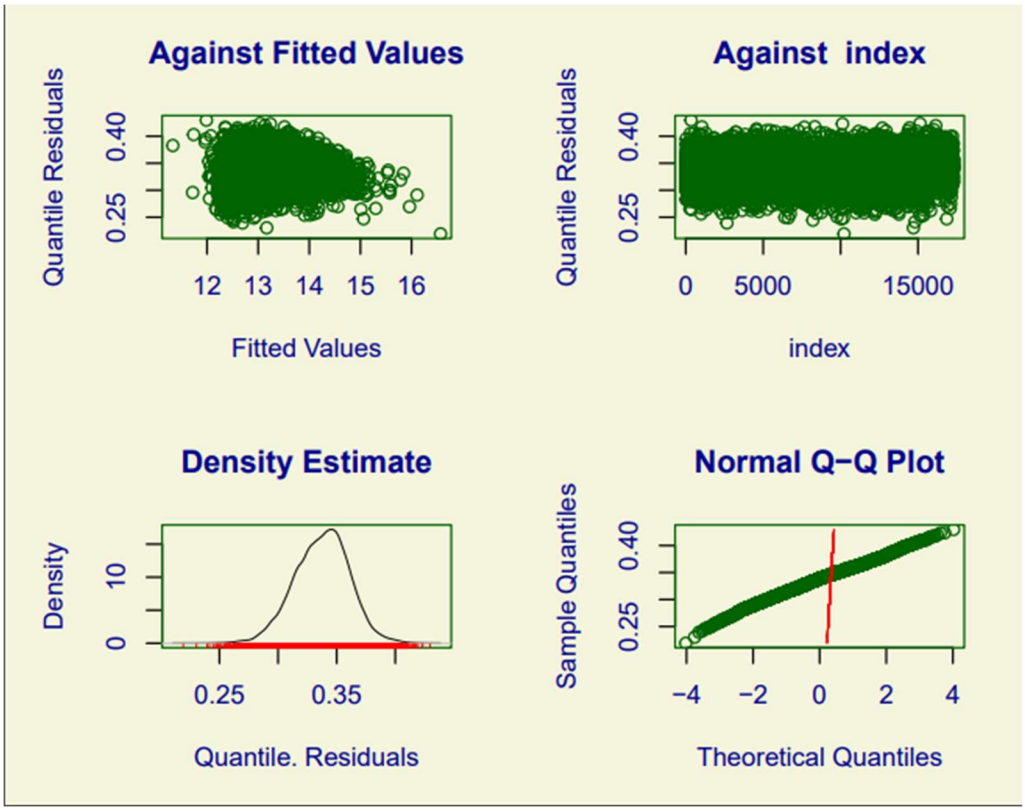


Figure 23: Residual Plot of MEXP model

Overall, the daignotics suggest that the mbct model is the best choice among the four models considered, based on its fit to the data.

Moving forward, we used the previously trained mbct model to make predictions on new observations. We select the first observation that was not used during the training of the model and apply the model to this new observation using the predict function. From this distribution, we extracted the predicted mean using the exp() function to revert the log transformation that was applied during training, and then find the predicted mean The output is a probability distribution, from which we extract the predicted mean which represents the model's best guess for the expected value of the response variable for this new observation, based on the information available in that observation. This process allows us to estimate what the model would predict for the target variable of a new observation.

As shown in figure 24. we also generate a visualization of the predicted distribution using the ggplot2 package. The plot shows the probability density function of the predicted distribution, which represents the model's estimate of the probability of observing different values of the response variable for the new observation. The x-axis represents the predicted values of the response variable (i.e., price), and the y-axis represents the density of the predicted distribution. The plot is generated using the predicted mean value obtained from the mbct model, which is first transformed back to the original scale using the exp() function. The plot title and axis labels are added using the labs() function.

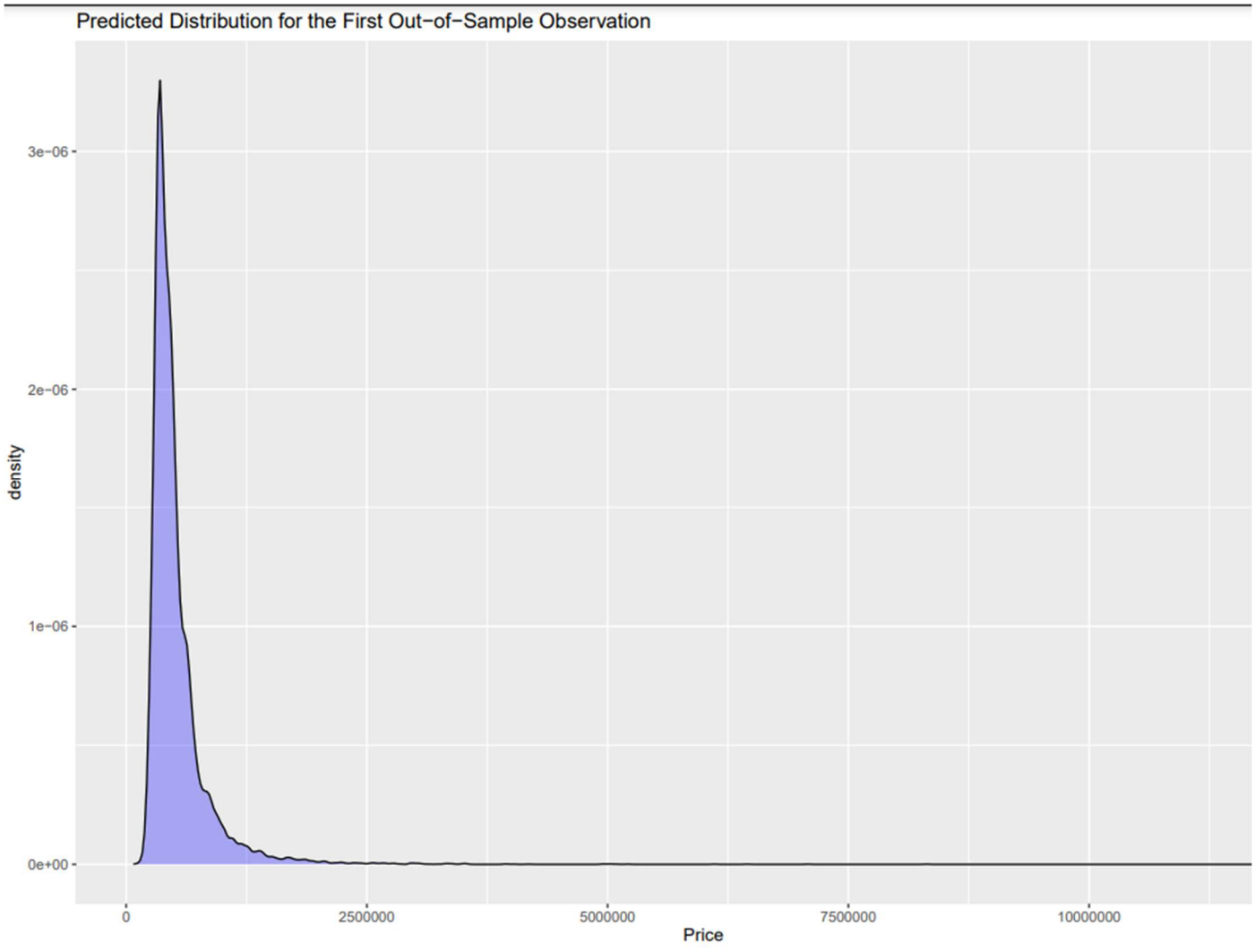


Figure 24: Predicted Distribution for the First Out-of-Sample Observation.

Overall Conclusion:

In conclusion, the analysis of the three data sets has demonstrated the paramount importance of employing appropriate statistical modeling techniques to gain a deeper understanding of data distributions and achieve accurate predictions. By carefully selecting the most suitable modeling approach for each dataset, valuable insights can be extracted, and robust predictions can be made.

In the first data set, after considering various distribution options, the Box-Cox-Cole-Green (BCCG) distribution was chosen due to its superior fit to the data. This decision was based on the analysis of goodness-of-fit measures and the overall performance of the model. By selecting the BCCG distribution, it was possible to capture the underlying characteristics of the data accurately.

Moving on to the second data set, the objective was to estimate centiles accurately. To achieve this, several modeling techniques were employed and evaluated. The results indicated that both the Box-Cox-Tweedie (BCT) and Box-Cox-Power-Exponential (BCPE) models exhibited lower GAIC values than the BCCG model. This suggests that the BCT and BCPE models performed better in capturing the centile estimations for the given data set.

In the analysis of the third data set, a BCT model with a log transformation was employed to fit the data adequately. To ensure the validity of the model, various diagnostic techniques were applied to assess its performance. These techniques involved examining residual plots, normality tests, and other relevant diagnostics. Through this thorough validation process, the reliability of the BCT model was confirmed, providing confidence in its predictions for future data.

Overall, this analysis highlights the significance of meticulous statistical modeling and rigorous diagnostic assessment. By carefully selecting appropriate distribution choices, considering transformation techniques, and conducting comprehensive model evaluations, researchers and analysts can identify the most suitable models for their data sets. This enables them to make accurate predictions, uncover meaningful insights, and support informed decision-making processes.

The findings of this coursework emphasize the need for a thoughtful and systematic approach to statistical modeling. By incorporating robust methodologies and rigorous evaluations, researchers can navigate the complexities of data distributions and derive accurate predictions, ultimately enhancing the quality and reliability of their analyses.

Peer Review:

Seed Number:1111

Student ID:21051665

Grade: B

The student's work seems to be satisfactory and well-structured. To find potential variable correlations, they performed a thorough exploratory analysis of the data set utilising visualisation techniques including histograms and scatter plots. By emphasising highly associated variables, the correlation matrix plot was also useful in deciding which dataset should be condensed for study. The student used many GAMLSS models with various distributions, making an appropriate choice in terms of the distribution for the response variable. The process for choosing and assessing the model was sound and the GAIC function was applied. However, a more thorough justification of the LogNO model's selection and the conclusions that may be drawn from it would have been helpful.

References

[1] S. v. B. R. B. J. M. R. B. E. B. M. R. S. V.-V. a. J. W. A.M. Fredriks, "Continuing positive secular change in The Netherlands, 1955-1997," *Pediatric Research,* vol. 47, pp. 316-323, 2000.

[2] D. V. C. T. M. S. D. D. A. &. S. G. Cohen, "Handgrip strength in English schoolchildren," *Acta Paediatrica,* vol. 99, pp. 1065-1072, 2010.

[3] "King County, Washington State Housing Price Dataset," [Online]. Available: https://github.com/Shreyas3108/house-price-prediction/raw/master/kc_house_data.csv. [Accessed 10 April 2023].

[4] S. v. Buuren, "Worm plot to diagnose fit in quantile regression," *Statistical Modelling,* vol. 7, p. 15, 2007.

[5] R. A. S. D. M. Rigby, Flexible Regression and Smoothing Using GAMLSS in R, Chapman and Hall/CRC, 2005.

[6] S. w. J. Starmer, "StatQuest: Histograms, Clearly Explained," [Online]. Available: https://www.youtube.com/watch?v=qBigTkBLU6g. [Accessed 11 May 2023].

[7] S. w. J. Starmer, "Quantiles and Percentiles, Clearly Explained!!!," [Online]. Available: https://www.youtube.com/watch?v=IFKQLDmRK0Y. [Accessed 11 May 2023].

[8] S. Kross, "A Q-Q Plot Dissection Kit," [Online]. Available: https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html. [Accessed 11 May 2023].

Appendix

Code for Dataset 1:

```
rm(list =
 ls()) # clears all objects in the workspace


#a
library(gamlss)
data(dbbmi)
old <- 15
da<- with(dbbmi, subset(dbbmi, age>old & age<old+1))
bmi15<-da$bmi
#write.csv(da, file = "my_data.csv")

library(MASS)

n_bins <- nclass.scott(bmi15)
truehist(bmi15, nbins=19, col="lightgray")


#b
mno<-gamlss(bmi15~age, data=da) # fit the model
mga <- gamlss(bmi15~age, data=da, family=GA)
mig <- gamlss(bmi15~age, data=da, family=IG)
mbccg <- gamlss(bmi15~age, data=da, family=BCCGo)
GAIC(mno, mga, mig, mbccg)

# plot fitted distribution
plot(mno, which = 1, type = "l")
plot(mga, which = 1, type = "l")
plot(mig, which = 1, type = "l")
plot(mbccg, which = 1, type = "l")

# mbccg has the lowest AIC value

#c
summary(mno)
summary(mga)
summary(mig)
summary(mbccg)
```

Code for Dataset 2:

```
rm(list = ls()) # clears all objects in the workspace


#a
data(grip)

#b
set.seed(1088)
index<-sample(3766, 1000)
mydata<-grip[index, ]
dim(mydata)
#data <- unlist(mydata)

#c
library(ggplot2)

ggplot(data = mydata, aes(x = age, y = grip)) +
  geom_point() +
  labs(x = "Age", y = "Grip strength")

#d
gbccg <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), data = mydata, family = BCCG)
edf(gbccg)

#e
gbct <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), tau.fo = ~pb(age), data = mydata, family =
BCT, start.from = gbccg)
gbcp <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age), nu.fo = ~pb(age), tau.fo = ~pb(age), data = mydata, family =
BCPE, start.from = gbccg)
edf(gbct)
edf(gbcp)

#f
# Calculate GAIC for each model
gaic_bccg <- GAIC(gbccg)
gaic_bct <- GAIC(gbct)
gaic_bcpe <- GAIC(gbcp)

# Print GAIC values
cat("GAIC for BCCG model:", gaic_bccg, "\n")
cat("GAIC for BCT model:", gaic_bct, "\n")
cat("GAIC for BCPE model:", gaic_bcpe, "\n")


#g
fittedPlot(gbct, x=mydata$age)
fittedPlot(gbcp, x=mydata$age)
fittedPlot(gbccg, x=mydata$age)
fittedPlot(gbcp ,gbct, gbccg, x=mydata$age)
#h
# Centile plots
```

```
cent_bccg <- centiles.split(gbccg)
cent_bct <- centiles.split(gbct)
cent_bcpe <- centiles.split(gbcp)

library(ggplot2)

# Residual plots
plot(gbccg)
wp(gbccg)
Q.stats(gbccg)

plot(gbct)
wp(gbct)
Q.stats(gbct)

plot(gbcp)
wp(gbcp)
Q.stats(gbcp)

# Obtain summary statistics for quantile residuals for gbccg
summary(gbccg, type = "qr")

# Obtain summary statistics for quantile residuals for gbct
summary(gbct, type = "qr")

# Obtain summary statistics for quantile residuals for gbcp
summary(gbcp, type = "qr")
```

Code for Dataset 3:

```
rm(list = ls()) # clears all objects in the workspace
#*****************DATA LOADING AND PREPARTION******************
mydata <- read.csv('C:\\Users\\tonny\\Downloads\\kc_house_data.csv\\kc_house_data.csv')
mydata$date <- as.POSIXct(mydata$date, format="%Y%m%dT%H%M%S")
# set your personal seed
set.seed(1088)
# create a vector of TRUE FALSE with probability 0.80 and 0.20
index <- sample(c("TRUE","FALSE"), dim(mydata)[1], replace = TRUE,
                prob=c(0.8, 0.2))
# make sure it is a logical vector
index <- as.logical(index)
# select the subset
df <- subset(mydata,index)
# check the dimensions of the new data
dim(df)
# see the first 5 rows
head(df, 5)
colnames(df)
#*****************DATA CLEANING******************
#Calculate the correlation matrix
install.packages('corrplot')
library(corrplot)
corr_matrix <- cor(df)
# Dropping specific columns
install.packages('dplyr')
library(dplyr)
df <- select(df,-id,-date,-lat,-long,-zipcode,-yr_renovated,-sqft_above,-sqft_lot15, sqft_living15)
#"sqft_living" and "sqft_above" are highly correlated
#as both of them represent the total area of the house so dropping them as well.
```

```
#(The main reason to drop sqft_lot15 variable is that it represents
#the lot size of the nearest 15 neighbors of the house, which may
#not be relevant to predicting the price of the house. The lot size
#of neighboring houses may have little impact on the price of
#a particular house, as it depends on many other factors such as the
#location, condition, and amenities of the house. Additionally,
#the sqft_lot variable already represents the lot size of the house,
#making sqft_lot15 redundant. Therefore, dropping sqft_lot15 would
#simplify the model without losing any relevant information.)
new_corr_matrix <- cor(df)
#*****************DATA VISUALISATION*****************
#correlation matrix plot
corrplot(new_corr_matrix, method = "color", tl.col = "black", tl.srt = 80, addCoef.col = "black")


# create histogram
ggplot(df, aes(x = price)) +
  geom_histogram(binwidth = 1000, color = "blue") +
  labs(title = "Histogram of Price", x = "Price")
# create normal probability plot
qqPlot(df$price, main="Normal Probability Plot of Price")
# create Scatter plot
par(mfrow=c(4,4), mar=c(2, 2, 1, 1), oma=c(0, 0, 2, 0))
for(i in 2:15){
  plot(df[[i]], df$price, xlab = names(df[i]), ylab = "price", main = paste("price against.", names(df[i])))
}
#*****************DATA ANALYSIS*****************
colnames(df)
formula <- as.formula("price ~ bedrooms + bathrooms + sqft_living
+sqft_lot+floors+waterfront+view+condition+grade+sqft_basement+yr_built+sqft_living15")
# Check the skewness of the transformed variable
#Checking Skewness
```

```
install.packages(moments)

library(moments)

skewness(df$price)

df$log_price <- log(df$price)

skewness(df$log_price)

#Since Price is heavily skewed we are doing a log transformation

df$log_price <- log(df$price)

install.packages("gamlss")

library(gamlss)

mbct <- gamlss(log_price ~ bedrooms + bathrooms + sqft_living
+sqft_lot+floors+waterfront+view+condition+grade+sqft_basement+yr_built+sqft_living15, data = df, family =
BCT())

mga <- gamlss(log_price ~ bedrooms + bathrooms + sqft_living
+sqft_lot+floors+waterfront+view+condition+grade+sqft_basement+yr_built+sqft_living15, data = df, family =
GA())

mno <- gamlss(log_price ~ bedrooms + bathrooms + sqft_living
+sqft_lot+floors+waterfront+view+condition+grade+sqft_basement+yr_built+sqft_living15, data = df, family =
NO())

mexp <- gamlss(log_price ~ bedrooms + bathrooms + sqft_living
+sqft_lot+floors+waterfront+view+condition+grade+sqft_basement+yr_built+sqft_living15, data = df, family =
EXP())

# Add fitted values to original dataset

df$mbct_fitted <- fitted(mbct)

df$mga_fitted <- fitted(mga)

df$mno_fitted <- fitted(mno)

df$mexp_fitted <- fitted(mexp)

#Plot the fitted Distributions

plot(mbct, which = 1, type = "l") # fitted distribution

plot(mga, which = 1, type = "l") # fitted distribution

plot(mno, which = 1, type = "l") # fitted distribution

plot(mexp, which = 1, type = "l") # fitted distribution

# Calculate GAIC for each model

gaic_mbct <- GAIC(mbct)

gaic_mga <- GAIC(mga)
```

```
gaic_mno <- GAIC(mno)
gaic_mexp <- GAIC(mexp)

gaic_values <- c(gaic_mbct, gaic_mga, gaic_mno, gaic_mexp)
model_names <- c("MBCT", "GA", "NO", "EXP")
df_gaic <- data.frame(model_names, gaic_values)
df_gaic <- df_gaic[order(df_gaic$gaic_values),]
df_gaic

summary(mbct)
summary(mga)
summary(mno)
summary(mexp)
#***************MAKING PREDICTIONS*******************
# Extract the first observation from the out-of-sample dataset
new_obs <- df[which(index == FALSE)[1], ]
# Use the mbct model to predict the distribution
pred_mbct <- predict(mbct, data = new_obs, type = "response")
# Extract the predicted mean from the output and rescale to the original scale
pred_mean <- exp(pred_mbct)
# Print the predicted mean
print(pred_mean)



# Plot the predicted distribution
ggplot(data.frame(x = pred_mean), aes(x)) +
  geom_density(fill = "blue", alpha = 0.3) +
  labs(title = "Predicted Distribution for the First Out-of-Sample Observation",
       x = "Price")
```