MA7007 Statistical Modelling & Forecasting
Case Study Report

Jiji Ajitha Kumari Venugopalan Assari

Student ID - 20033188

Contents

1 Introduction

The report describes the results of the statistical analysis of the two given datasets and an additional dataset that was chosen from Kaggle, comprising of the Life Expectancy data from 2000 to 2015 for all the countries. Since the third data chosen is a secondary data, it may have certain drawbacks as the reliability of data cannot be ensured. However, the details in the chosen data mentions that the data was consolidated from two primary sources, GHO data repository under WHO and United Nation website.

For the first task, we consider a dataset with age and bmi as the columns. The age that we take is from 12-13 and corresponding to the age. FitDist() from GAMLSS is used to fit distributions and choose an appropriate distribution. The dataset with age and Handgrip strength of schoolchildren is considered in the second task, where BCT, BCCG and BCPE distributions are used to fit the data. The third dataset is cleaned by removing all NaN values and omitting outliers. The response and explanatory variables were selected from the data. The distribution was found by using fitDist() and the GAIC values were calculated to find the best fit model. The residue of the predicted values were plotted and shown.

2 Task 1: First dataset analysis

The original data dbbmi from gamlss.data contains two variables, age and bmi, where the age variable ranges from 0 to 22. We need to analyse the bmi corresponding to the age 12 (considering the range from 12-13). The histogram corresponding to the bmi value is plotted as in Figure 1.

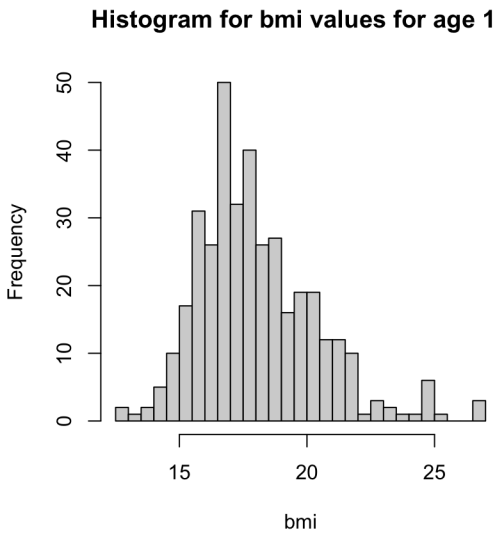


Figure 1: :Histogram for BMI for age 12

2.1 Distributions used

The function fitDist() is used to fit distributions to the data. The default option for argument is type="realAll", which utilise all available continuous distributions. The information criteria AIC is used to find the best fit model. The distributions used to fit the data and the corresponding AIC are detailed in Figure 2.

SHASHo2	SHASHo	SEP3	SEP1	SEP2	exGAUS	SHASH	ST2	ST1	EGB2	SST
1658.541	1658.541	1658.561	1658.695	1659.080	1659.145	1659.615	1659.747	1659.861	1659.969	1660.043
ST3	JSU	JSUo	SEP4	GB2	BCCG	BCCGo	ST5	GG	RG	BCTo
1660.043	1660.481	1660.481	1661.012	1661.311	1661.423	1661.423	1661.680	1661.692	1662.124	1662.422
BCT	SN1	BCPE	BCPEo	SN2	IGAMMA	GIG	ST4	LOGNO2	LOGNO	IG
1662.422	1662.497	1662.792	1662.792	1665.697	1667.345	1669.345	1672.472	1674.146	1674.146	1674.324
GA	LO	GT	TF2	TF	PE2	PE	NET	NO	WEI2	WEI
1683.215	1692.403	1693.059	1693.536	1693.536	1698.552	1698.552	1698.817	1707.778	1793.557	1793.557
WEI3	GU	EXP	PARETO2	GP	PARETO2o					
1793.557	1874.654	2921.814	2923.814	2923.815	2923.819					

Figure 2: :AIC for fitted distributions

2.2 Best Distribution

The Best model is chosen as SHASHo2 (Sinh-Arcsinh) with the fitting method "nlminb".

2.3 Reasons for which the distribution was chosen

The least AIC corrresponds to SHASHo2 (Sinh-Arcsinh) and thus this model is selected as the best model to fit the data. The model has an SBC of 1674.25. SHASHo has a similar AIC. SEP3 (Skew Power exponential) is the next best fit model for the data with an AIC 1658.561.

2.4 Plotting the fitted distribution

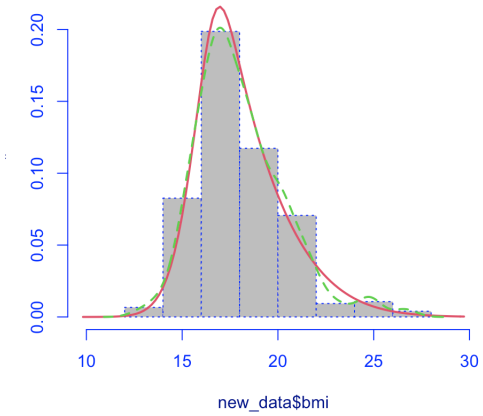


Figure 3: :The fitted SHASHo2 distribution and BMI

2.5 Fitted parameter values of the final chosen model

The parameter estimates of the Sinh-arcsinh distribution is given in the column Estimate in Figure 4. SHASHo2 distibution is a four parameter distribution where the parameters are μ, σ, ν, τ where μ is the mean, σ is the standard deviation, ν represents the skew and τ represents the kurtosis

```
Coefficient(s):
            Estimate  Std. Error
eta.mu    16.9022360   0.1837672
eta.sigma  0.6094557   0.0682756
eta.nu     0.3839707   0.0687852
eta.tau   -0.2015986   0.0856548
```

Figure 4: :The parametric estimates and standard error of SHASHo2 distribution

3 Task 2: Second dataset analysis

The second dataset includes two variables age and grip which was collected from 3766 English boys. The task is to analyse a sample of 1000 to create centile curves for grip given age.The original data is available in gamlss package. We use 360 as the seed number to select the sample. The grip is plotted against age from the derived dataset.

3.1 Models used

LMS method is used to fit the data. The method is defined as $Y \sim f_Y(y|\mu, \sigma, \nu, \tau)$. The BCCG (Box-Cox Cole and Green) distribution along with BCT(Box-Cox t) and BCPE(Box-Cox power exponential) distributions are fitted for grip. BCCG is derived by assuming that the response variable is a specific function of a random variable Z which has a truncated normal distribution and it is suitable for either negative or positively skewed data. BCT assumes that Z has a truncated t distribution whereas BCPE assumes Z has a truncated exponential power distribution.

3.2 Question from Qn 1.2

Need of power transformation in the data

The plot for grip against age is shown in Figure 5. The power transformation of the explanatory variable is usually needed when the response variable has an early or late spell of fast growth. However, there is no such growth exhibited by the response variable grip as show in the figure and consequently, there is no need to power transform age in this data set.

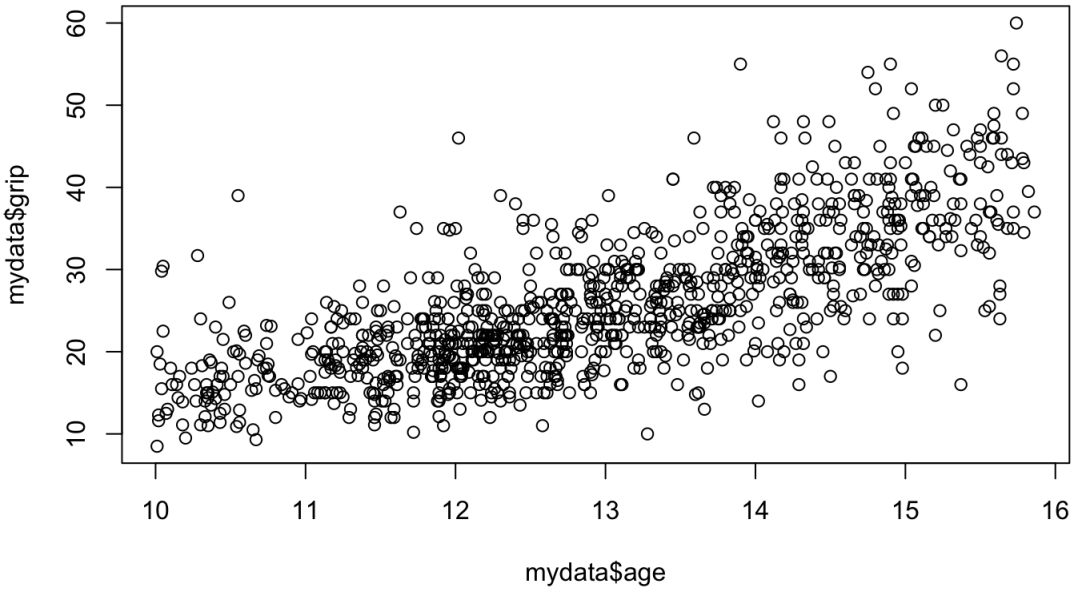


Figure 5: :Plot for grip against age

Using BCCG distribution

edf() and edfAll() are the functions which are used to obtained the effective degrees of freedom for the distribution parameters in a gamlss model.

```
$mu
$mu$`pb(age)`
[1] 4.642573


$sigma
$sigma$`pb(age)`
[1] 2.002886


$nu
$nu$`pb(age)`
[1] 2.000129
```

Figure 6: :The effective degrees of freedom

The function edfAll() was used to obtain the effective degrees of freedom for $\mu, \sigma \& \tau$ and are approximately 4.64 , 2.003 and 2 respectively.

Using BCPE and BCT distribution

The effective degrees of freedom for μ, σ, ν and τ for BCPE and BCT are displayed in Table 2.

Parameter	BCT	BCPE
μ	4.708098	4.729473
σ	2.002298	2.0023
ν	2.000148	2.000107
τ	2.000039	2.00034

Table 1: Effective degrees of freedom fitted for the parameters

BCPE has higher degrees of freedom for all the parameters when compared to other two methods.

GAIC to compare the three models

The functions AIC() and GAIC() can be used to obtain the generalised Akaike information criterion. We can dcompare the three models using thus method. GAIC user 2 as the default penalty resulting in the AIC. We select the model with minimum AIC as the best fit. Comparing the values in Figure 7, AIC is the least for gbcpe_ model, which is the model fitted using BCPE and the second least is for gbct_model, using BCT distribution and the highest is for the model which use BCCG distribution.

```
                    df      AIC
gbcpe_model 10.732219 6255.194
gbct_model  10.710582 6258.062
gbccg_model  8.645589 6262.542
```

Figure 7: :GAIC for the three models

Plotting the parameters for the fitted models

The models that has the least AIC values are those using BCT and BCPE distributions. The Figure 8 shows the fitted smooth values for all the parameters for both the models. fittedPlot() is used to plot the figure where age is used as the explanatory variable.

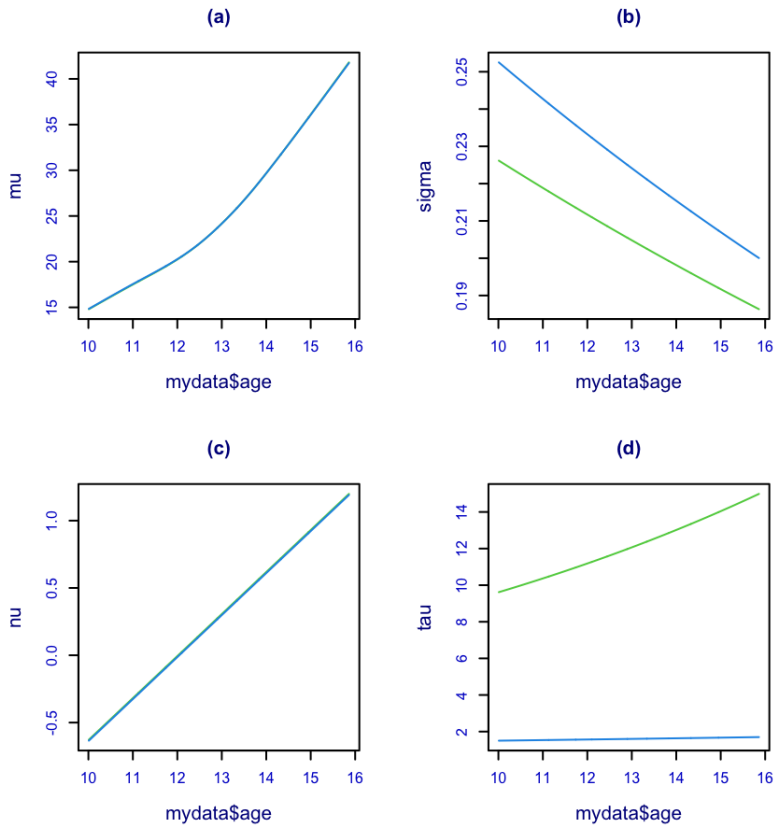
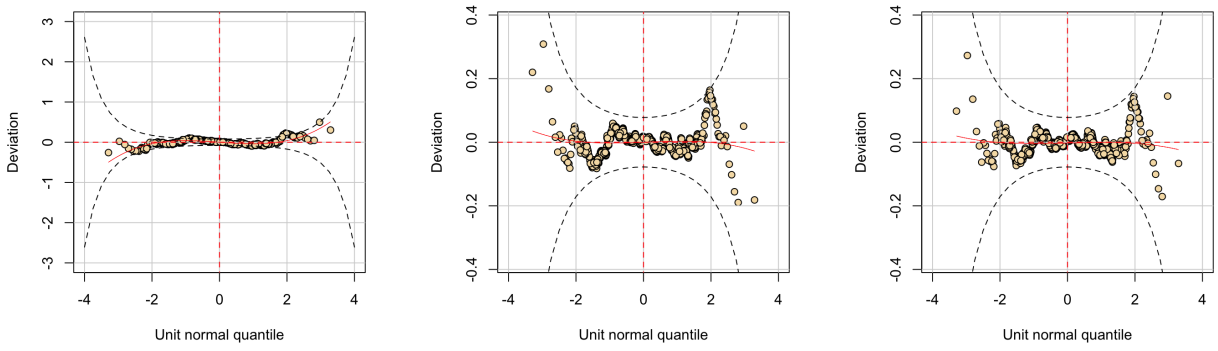


Figure 8: :Plotting the parameters for the models

3.3 Use residual diagnostics for checking the model



(a) BCCG (b) BCT (c) BCPE

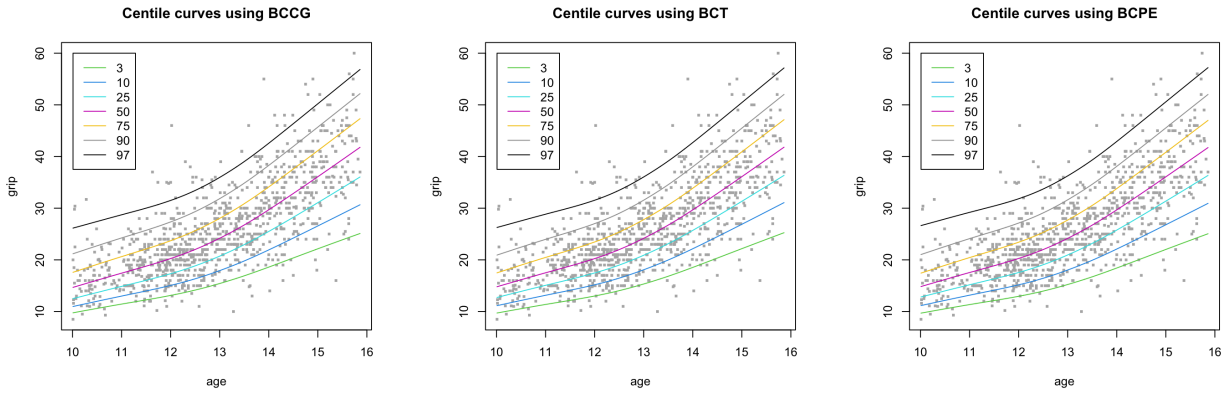
Figure 9: Worm plot

wp() is used to plot the worm plot of the residuals of the three fitted models. Residue values of BCCG lies within the range of -1 to 1 and the residue values of BCPE and BCT lies within the range of -0.4 and 0.4. The residue of BCPE and BCT are closer to the horizontal line when compared to BCCG.

3.4 Selecting the final model

BCPE model and BCTE has slight difference in AIC value. However, since the residue values of BCPE as displayed in the worm plot in Figure 9 appears to be more closer to the horizontal line compared to BCT, we can conclude that BCPE is the best fitting model.

3.5 Comparing the centile plots for the three fitted models



(a) BCCG (b) BCT (c) BCPE

Figure 10: Centile curves

The curves in BCPE and BCT appears to be centered a bit lower than that in BCCG.

4 Task 3: Third dataset analysis

The third dataset comprises the target variable life expectancy along with many other factors that may affect the life expectancy. There are 22 columns and 2938 rows in the dataset. We may take 20 of them as the predicting variables. However, we limit the range of our study to 5 explanatory variables which are Population, Adult Mortality, GDP, Total expenditure on health, BMI.

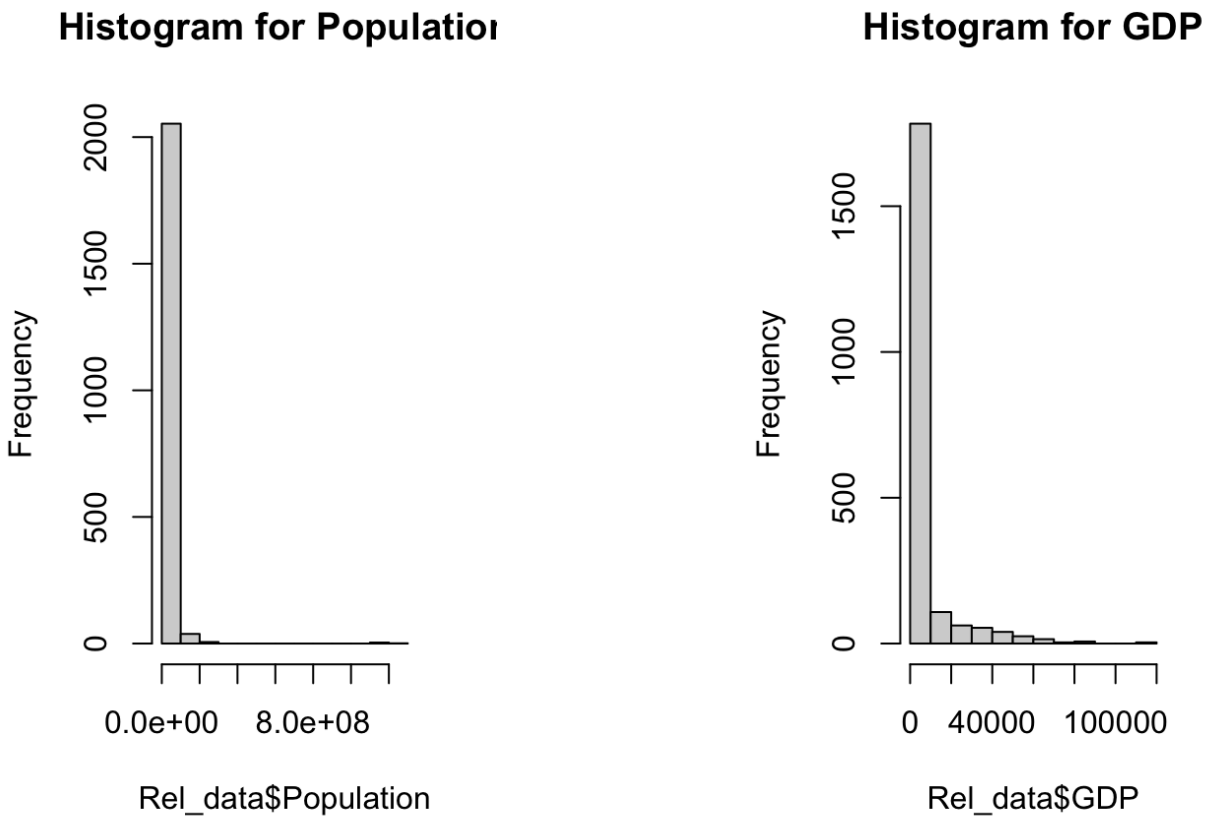
4.1 Objective for the study

The objective of the study is to analyse the factors affecting the response variable. We start by fitting models by changing the explanatory variable to find the best fitting model.

4.2 Preliminary analysis on the collected data and Reliability of the data

The first step of pre-processing is to remove all null values from the data. The rows with null values are deleted using na.omit() function and the resultant data is stored after dropping irrelevant columns.

The histogram plots for all the variables are plotted. Out of all the explanatory variable and target, Population and GDP seemed to have outliers in it as shown in the Figure 11a



(a) Population

(b) GDP

Figure 11: Histogram plots

Further analysis, by finding the IQR value and checking for outliers, showed that there are outliers only in Population data. Thus, the outliers form the variable Population is removed and the process is as follows:

- Find upper quartile, Q3 (75%) and lower quartile Q1 (25%)
- Calculate Inter Quartile Range (IQR)

- Define normal range with lower limit Q1-1.5*IQR and upper limit Q3+1.5*IQR
- Remove the data points outside the normal range

The data is from a secondary source and such, the accuracy of the data is not guaranteed. The dataset includes the data from the year 2000 to 2015 and as such, it is not updated. These factors can affect the reliability of the data adversely.

The target variable was plotted against each of the explanatory variable for analysing the relation of the variables. Then, the data was divided into two sections, training data with 75% of the data and remaining was stored as the testing data.

The distribution of the Life expectancy from the training data was analysed using fitDist() and SHASH(Sinh-Arcsinh) was selected as the best fitting distribution with AIC 10093. The AIC values corresponding to other distributions fitted for the data is in Figure 12

SHASH	SHASHo	SHASHo2	SEP1	ST2	ST1	SN2	SEP3	SST	ST3
10093.42	10099.35	10099.35	10101.81	10111.53	10111.53	10112.24	10113.72	10114.24	10114.24
SEP2	GU	JSU	JSUo	EGB2	SEP4	ST5	ST4	PE	PE2
10123.84	10141.04	10141.83	10141.83	10144.04	10145.13	10146.29	10257.25	10257.34	10257.34
GT	NO	TF	exGAUS	TF2	SN1	LO	NET	RG	
10257.46	10278.06	10280.06	10280.06	10280.06	10280.06	10318.01	10402.12	10676.04	

Figure 12: :Fitting Distribution for Life Expectancy

The plot for the fitted SHASH distribution is shown in Figure 13.

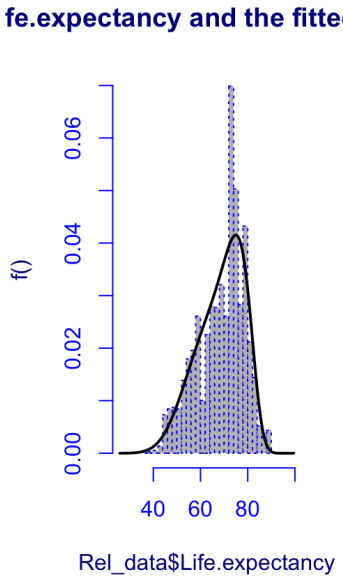


Figure 13: :The fitted distribution for Life Expectancy

4.3 Model to fit the data

9 models were fitted using different response variable and family as SHASH and pb() the P-spline for smoothing were used. The table below shows the AIC values and Coefficients of parameter for each model

Let Pop denotes population variable, AM denotes Adult.Mortality, Ex denotes Total.Expenditure in our data. The shortforms are used in the below Table to denote the variables.

Model	explanatory variable(s)	AIC	SBC
model1	Pop	10022.34	10089.52
model2	AM	7724.026	7844.693
model3	Ex	9568.902	9670.702
model4	GDP	9470.389	9546.339
model5	BMI	9072.043	9177.861
model6	BMI,AM	7299.885	7508.904
model7	BMI,AM,GDP	7229.119	7470.61
model8	BMI,AM,GDP,Ex	7216.787	7483.374
model9	BMI,AM,GDP,Ex,Pop	7214.178	7486.431

Table 2: Models with AIC and SBC

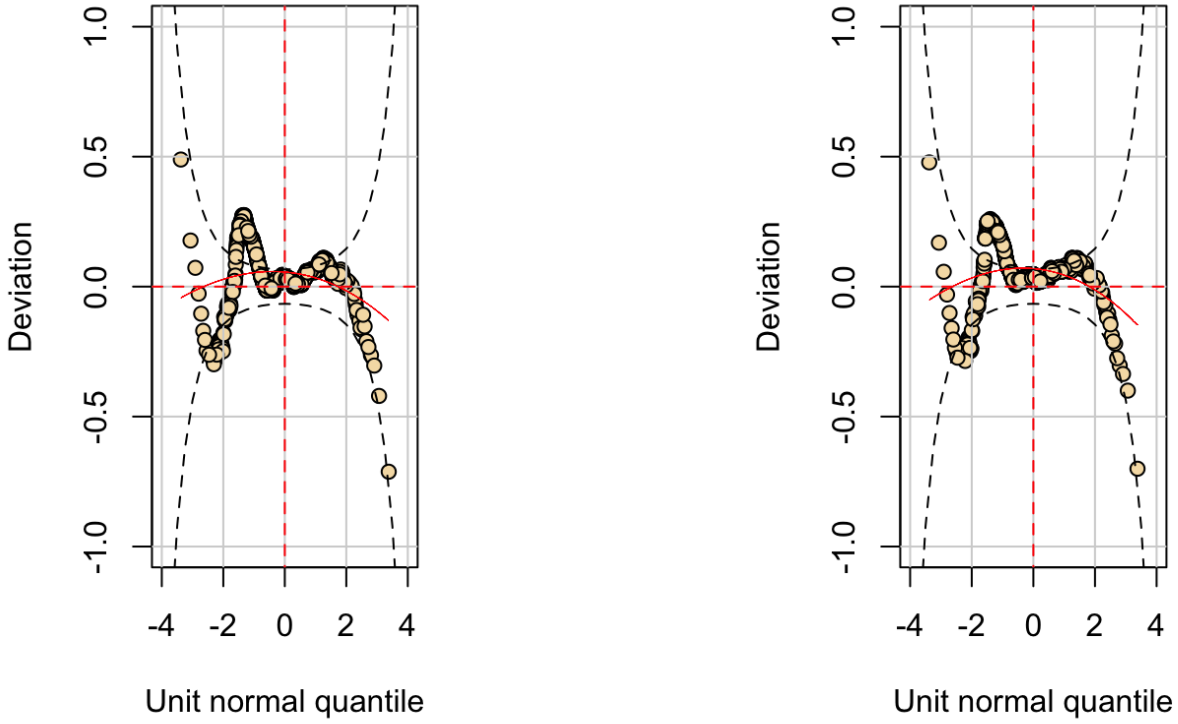
Considering the AIC and SBC for the first 5 models, model2 with Adult.Mortality as the explanatory variable and model5 with BMI as the explanatory variables has minimum for both AIC and SBC. Thus, we consider Adult.Mortality and BMI for model6 and then add GDP with the next least AIC among the first 5 models and so on to get further models.

model9 with all five variables as independent variables has the least AIC value. However, model7 with Adult.Mortalilty, BMI and GDP as the explanatory variable seems to have the least SBC value.

4.4 Selecting the final model

The residual plots and worm plots were considered to select the final model in addition to the AIC and SBC values. The residual plots for both the models seems to be similar as there is only a slight difference in the spread of residual values.

Analysing the worm plot in Figure 14, the residual values in model9 seems to be slightly closer to the horizontal line when compared to model7. Thus, we consider model9 as the best fit model.



(a) Model7

(b) Model9

Figure 14: Worm plots

4.5 Using the model for prediction

The life expectancy correponding to the testing data was calculated using the selected model. The residuals were calculated and added as a new column in test data. This was then plotted against the predicted life expectancy.

The majority of the residual values lies within -10 and 10. However, there are about 4 points that lie beyond 20 and about 6 which lies between 12-20.

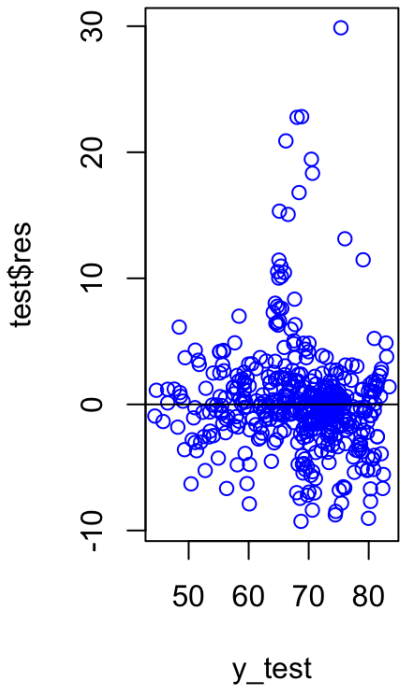


Figure 15: :Residual plot for prediction of testing data

5 Peer Review

5.1 Choose one from a selection of student work on the third data set

This section covers a peer review of the course work by student Ugochukwu Daniel (19017777167) accepted to peer review this student's third data set.

5.2 A short critique on the adequacy of the work

The quality of the explorative analysis of the data set

The students' work is detailed in terms of how they investigated the data by displaying graphs and explaining what the plots mean. The data was cleaned by looking for and dealing with missing values. They examined the data types to determine whether they could be analysed, and transformations were carried out to prepare the data for modelling. They examined

each variable to see if it affected their target variable, Price, and eliminated variables that had no bearing on the target.

The choice of the distribution for the response (target) variable

After experimenting with various methods to test for suitability, including testing with diagnostics that worked well with the data, the distribution was chosen.

The method for selecting and checking the model

The student tried out various methods to help him choose the best model for the data. The diagnostic tests are perfect, implying a suitable representation of data. Many diagnostics, such as worm plots, bucket plots, Q statistics, and so on, were used to assess the model's suitability.

The interpretation of the results

The result's interpretation is communicated in detail for easy comprehension. The student's work demonstrates a thorough understanding and knowledge of the subject.

5.3 Grade representing your estimation of the value of the work

Given the subject matter, the methods followed and way the results were clearly presented I would grade the analysts work an A grade

6 Conclusion

The analysis was done on three different datasets. The first two datasets were from the GAMLSS library. Different distributions were fitted for BMI calculated for age 12 in the first task. The second task comprised of fitting models using BCCG, BCPE and BCT distributions to find the best fit model. The third dataset was used to analyse the effect of five explanatory variables on the target variable.

References

[1] Robert A Rigby, Mikis D Stasinopoulos, Gillian Z Heller, and Fernanda De Bastiani, *Distributions for modeling location, scale, and shape: Using GAMLSS in R.* CRC press, 2019.

[2] A.M. Fredriks, S. van Buuren, R.J.F. Burgmeijer, J.F. Meulmeester, R.J. Beuker, E. Brug- man, M.J. Roede, S.P. Verloove-Vanhorick, and J. M. Wit. *Continuing positive secular change in The Netherlands, 1955-1997.* Pediatric Research, 47:316–323, 2000.

[3] D. D. Cohen, C. Voss, M.J.D. Taylor, D.M. Stasinopoulos, A. Delextrat, and G.R.H. Sandercock. *Handgrip strength in English schoolchildren.* Acta Paediatrica, 99:1065–1072, 2010.

[4] Hadley Wickham, Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2016.

[5] Brett Lantz. *Machine Learning with R*, 2019.

[6] Stasinopoulos D. M., Rigby R.A., Heller G., Voudouris V., and De Bastiani F., *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, 2017.

[7] D. Mikis Stasinopoulos, Robert A. Rigby, *Generalized Additive Models for Location Scale and Shape (GAMLSS) in R*, 2007.

A Code for the Tasks

Code for Task1

```r
#TASK 1

#clear screen
cat("\014")

#importing library
library(gamlss)

##Section 1.1 (a)

data('dbbmi')

#print the first 6 values of the data set
head(dbbmi)

#To find the number of rows in the dataset
nrow(dbbmi)

sd(dbbmi$age)

#Plotting bmi against age
plot(bmi~age,data=subset(dbbmi,age<0.10))

#assigning the age for which bmi is to be calculated
age_old <- 12

#take the subset of the dataset satisfying the given condition(age b/w 12 and 13)
new_data <- with(dbbmi,subset(dbbmi,age>age_old  & age<age_old +1))

#print the first 6 values in new_data
head(new_data)

#assign the values in the column bmi in new_data
pred_bmi <- new_data$bmi
pred_bmi

#plot the histogram for bmi values corresponding to ages 12 to 13
hist(pred_bmi,breaks=30,main="Histogram for bmi",xlab="bmi")
```

Figure 16: :Task1:1

```r
##Section 1.1 (b): Fit different parametric distributions to the data

model1<-fitDist(new_data$bmi)
model1
model1$fits
model1$fails

hist1<-histDist(new_data$bmi, "SHASHo2" , density=TRUE,ylim = c(0,0.22))
plot(hist1)
wp(hist1)


##Section 1.1 (c):
summary(model1)

#The Sinh-Arcsinh (SHASH) distribution is a four parameter distribution
#mu : vector of location parameter values
#sigma: vector of scale parameter values
#nu : vector of skewness parameter values
#tau : vector of kurtosis parameter values
```

Figure 17: :Task1:2

Code for Task2

```r
#TASK 2

##Section 1.2 (a):

data('grip')

#print the first 6 values of the data set
head(grip)

##Section 1.2 (b):

set.seed(360)
index<-sample(3766,1000)
mydata<-grip[index,]
dim(mydata)

hist(mydata$age,prob=TRUE)
curve(dnorm(x, mean=mean(mydata$age), sd=sd(mydata$age)), add=TRUE,col="blue")
sd(mydata$age)

##Section 1.2 (c):

plot(mydata$age,mydata$grip)
z.scoresQS

##Section 1.2 (d):
gbccg_model <- gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age), family=BCCG, data=mydata)
edfAll(gbccg_model)

?edfAll
```

Figure 18: :Task2:1

```r
##Section 1.2 (e):

gbct_model <- gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age), tau.fo=~pb(age), family=BCT, data=mydata,
                     start.from = gbccg_model)
edfAll(gbct_model)

gbcpe_model <- gamlss(grip~pb(age),sigma.fo=~pb(age),nu.fo=~pb(age), tau.fo=~pb(age), family=BCPE, data=mydata,
                     start.from = gbccg_model)
edfAll(gbcpe_model)

##Section 1.2 (f):

GAIC(gbccg_model,gbct_model,gbcpe_model)

##Section 1.2 (g):

# to plot the fitted values for all the parameters of a GAMLSS model against the explanatory variable
fittedPlot(gbct_model,gbcpe_model,x=mydata$age)

##Section 1.2 (h):

centiles(gbccg_model,xvar=mydata$age,cent=c(3,10,25,50,75,90,97),xlab="age",ylab="grip")

centiles(gbct_model,xvar=mydata$age,cent=c(3,10,25,50,75,90,97),xlab="age",ylab="grip")

centiles(gbcpe_model,xvar=mydata$age,cent=c(3,10,25,50,75,90,97),xlab="age",ylab="grip")

##Section 1.2 (i):

plot(gbccg_model)
plot(gbct_model)
plot(gbcpe_model)

wp(gbccg_model,ylim.all=3)

wp(gbct_model)

wp(gbcpe_model)

Q.stats(gbcpe_model,xvar=mydata$age)
Q.stats(gbct_model,xvar=mydata$age)
Q.stats(gbccg_model,xvar=mydata$age)
```

Figure 19: :Task2:2

Code for Task3

```r
#importing library
library(gamlss)

##Section 1.3 (a):

#importing the dataset
Data=read.csv("Life Expectancy Data.csv",header=TRUE)
head(Data)

#explanatory -> population, GDP, Adult.Martality, expenditure , BMI
#target -> life.expectancy

##Section 1.3 (b):

# data source: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
#To analyse what factors affect life expectancy

##Section 1.3 (c):

Data1 <- Data[c("Population", "GDP", "Adult.Mortality","Total.expenditure","BMI", "Life.expectancy")]
#Remove NA's
Rel_data<-na.omit(Data1)
summary(Rel_data)
head(Rel_data)

#Histogram for life expectancy
hist(Rel_data$Life.expectancy,main="Histogram for Life Expectancy")

#Histogram for Population
hist(Rel_data$Population, main="Histogram for Population")

#Histogram for GDP
hist(Rel_data$GDP,main="Histogram for GDP")

#Histogram for Adult Mortality
hist(Rel_data$Adult.Mortality,main="Histogram for Mortality")

#Histogram for Total.expenditure
hist(Rel_data$Total.expenditure, main="Histogram for Expenditure")

#Histogram for BMI
hist(Rel_data$Life.expectancy,main="Histogram for BMI")
```

Figure 20: :Task3:1

```r
#Remove outliers from population
Q1 <- quantile(Rel_data$Population, .25)
Q3 <- quantile(Rel_data$Population, .75)
IQR <- IQR(Rel_data$Population)

Rel_data <- subset(Rel_data, Rel_data$Population > (Q1 - 1.5*IQR) & Rel_data$Population < (Q3 + 1.5*IQR))

#Histogram for Population
hist(Rel_data$Population, main="Histogram for Population")

#Checked for outliers in GDP using the same method above but as there was no outliers removed,
#the section was removed from here.

##Section 1.3 (d):

plot(Life.expectancy~Population, data=Rel_data, col=gray(0.6), pch=15, cex=0.5)
plot(Life.expectancy~GDP, data=Rel_data, col=gray(0.7), pch=15, cex=0.5)
plot(Life.expectancy~Adult.Mortality, data=Rel_data, col=gray(0.7), pch=15, cex=0.5)
plot(Life.expectancy~Total.expenditure, data=Rel_data, col=gray(0.7), pch=15, cex=0.5)
plot(Life.expectancy~BMI, data=Rel_data, col=gray(0.7), pch=15, cex=0.5)

#Changing to training and testing data
indx = sort(sample(nrow(Rel_data), nrow(Rel_data)*.75))

train <- Rel_data[indx,]
test <- Rel_data[-indx,]

fit_life=fitDist(train$Life.expectancy,type="realline")
fit_life
fit_life$fits
fit_life$failed

#SHASH - Sinh-ArcSinh

histDist(Rel_data$Life.expectancy, family=SHASH, nbins=30, line.col="black")

#model with Life.expectancy~Population
lm_model1 <- gamlss(Life.expectancy~pb(Population), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model1)
```

Figure 21: :Task3:2

```r
#model with Life.expectancy~Adult.Mortality
lm_model2 <- gamlss(Life.expectancy~pb(Adult.Mortality), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model2)

#chooseDist(lm_model2)
#-> min GAIC for BCTo

#model with Life.expectancy~Total.expenditure
lm_model3 <- gamlss(Life.expectancy~pb(Total.expenditure), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model3)
#trace=FALSE -> AIC: 10001.16 and give the below error:
#Algorithm RS not converged -> change trace=FALSE into control = gamlss.control(n.cyc = 200,trace=FALSE)

#model with Life.expectancy~GDP
lm_model4 <- gamlss(Life.expectancy~pb(GDP), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model4)

#model with Life.expectancy~BMI
lm_model5 <- gamlss(Life.expectancy~pb(BMI), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model5)

AIC(lm_model1, lm_model2, lm_model3,lm_model4,lm_model5)

#lm_model2 and lm_model5 has lowest AIC

#model with Life.expectancy~BMI+Adult.Mortality
lm_model6 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality), family=SHASH, data=train,control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model6)

#for summary, qr is used instead of vcov -> use individual fits of the parameters
# donot take into account the correlation between the estimates of the distribution parameters

#model with Life.expectancy~BMI+Adult.Mortality+GDP
lm_model7 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality)+pb(GDP), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model7)
```

Figure 22: :Task3:3

```
#model with Life.expectancy~GDP
lm_model4 <- gamlss(Life.expectancy~pb(GDP), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model4)

#model with Life.expectancy~BMI
lm_model5 <- gamlss(Life.expectancy~pb(BMI), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model5)

AIC(lm_model1, lm_model2, lm_model3,lm_model4,lm_model5)

#lm_model2 and lm_model5 has lowest AIC

#model with Life.expectancy~BMI+Adult.Mortality
lm_model6 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality), family=SHASH, data=train,control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model6)

#for summary, qr is used instead of vcov -> use individual fits of the parameters
# donot take into account the correlation between the estimates of the distribution parameters

#model with Life.expectancy~BMI+Adult.Mortality+GDP
lm_model7 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality)+pb(GDP), family=SHASH, data=train, control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model7)

#model with Life.expectancy~BMI+Adult.Mortality+GDP+Expenditure
lm_model8 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality)+pb(GDP)+pb(Total.expenditure), family=SHASH, data=train,
                                                              control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model8)


#model with Life.expectancy~BMI+Adult.Mortality+GDP+Total.expenditure+Population
lm_model9 <- gamlss(Life.expectancy~pb(BMI)+pb(Adult.Mortality)+pb(GDP)+pb(Total.expenditure)+pb(Population), family=SHASH, data=train,
                                                                 control = gamlss.control(n.cyc = 200,trace=FALSE))
summary(lm_model9)

AIC(lm_model6,lm_model7,lm_model8,lm_model9)

#The effective degrees of freedom:
edfAll(lm_model9,"mu")
```

Figure 23: :Task3:4

```r
##Section 1.3 (e):

#For Residual plots
plot(lm_model7)

plot(lm_model9)

#wormplot
wp(lm_model7,ylim.all=1)

wp(lm_model9,ylim.all=1)

#Fitted smooth functions in the best model
term.plot(lm_model9, pages=1, ask=FALSE)

##Section 1.3 (f):

#Use the chosen mode to predict all the parameters μ, σ, ν and τ for the values in newdata.
x_test <-test[c("Population", "GDP", "Adult.Mortality","Total.expenditure","BMI")]

y_test <- predict(lm_model9,newdata=x_test,type="response")
test['res']=y_test-test['Life.expectancy']

plot(y_test,test$res,col="blue")
abline(0,0)
```

Figure 24: :Task3:5

B Additional Figures

```
> plot(lm_model7)
******************************************************************
             Summary of the Quantile Residuals
                          mean   =  0.03800729
                      variance   =  0.989242
               coef. of skewness  =  -0.0822864
               coef. of kurtosis  =  3.052222
Filliben correlation coefficient  =  0.9969957
******************************************************************
```

Figure 25: :Summary of Quantile Residuals for model7

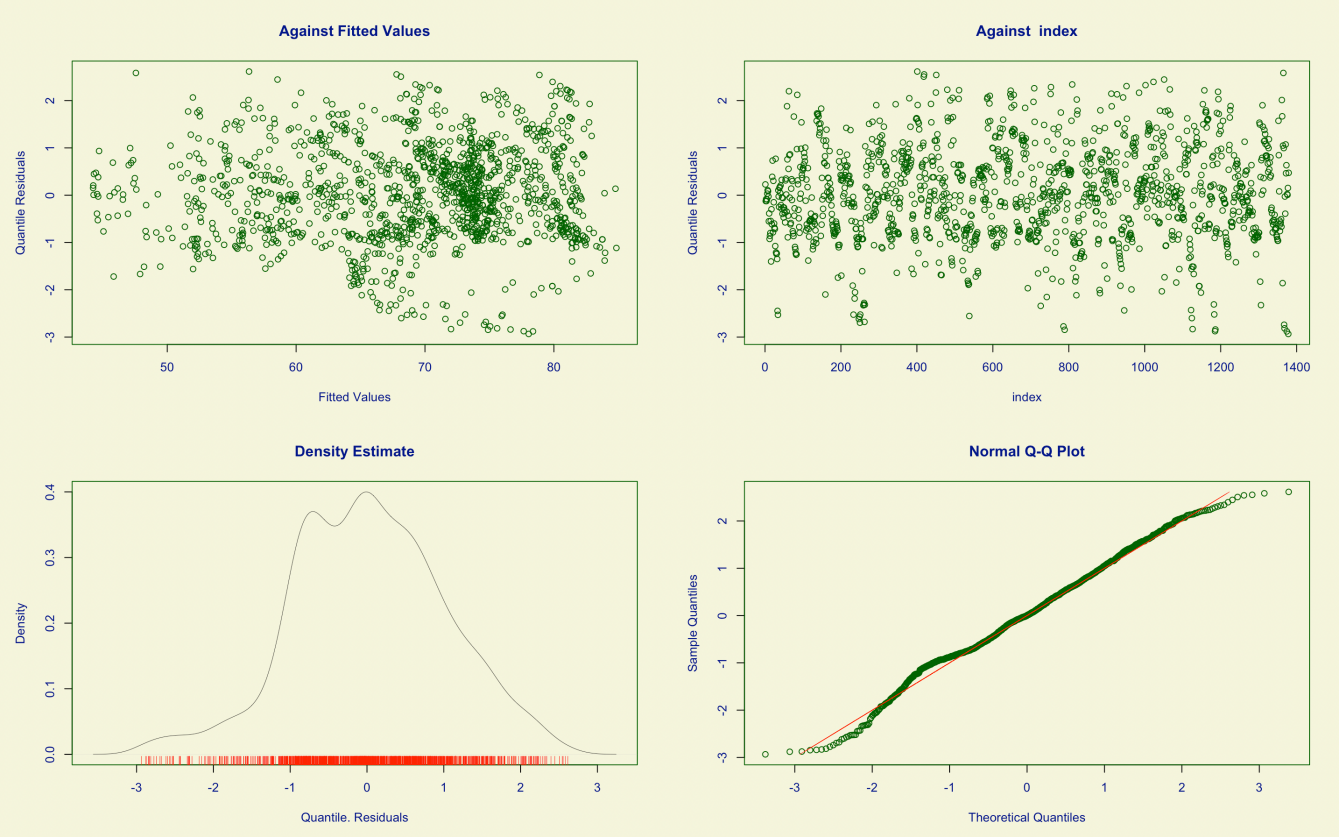


Figure 26: :Residual plot for model7

```
> plot(lm_model9)
**********************************************************************
             Summary of the Quantile Residuals
                         mean   =  0.03442117
                     variance   =  0.990495
              coef. of skewness  =  -0.08666185
              coef. of kurtosis  =  3.094857
Filliben correlation coefficient  =  0.9969957
**********************************************************************
```

Figure 27: :Summary of Quantile Residuals for model9

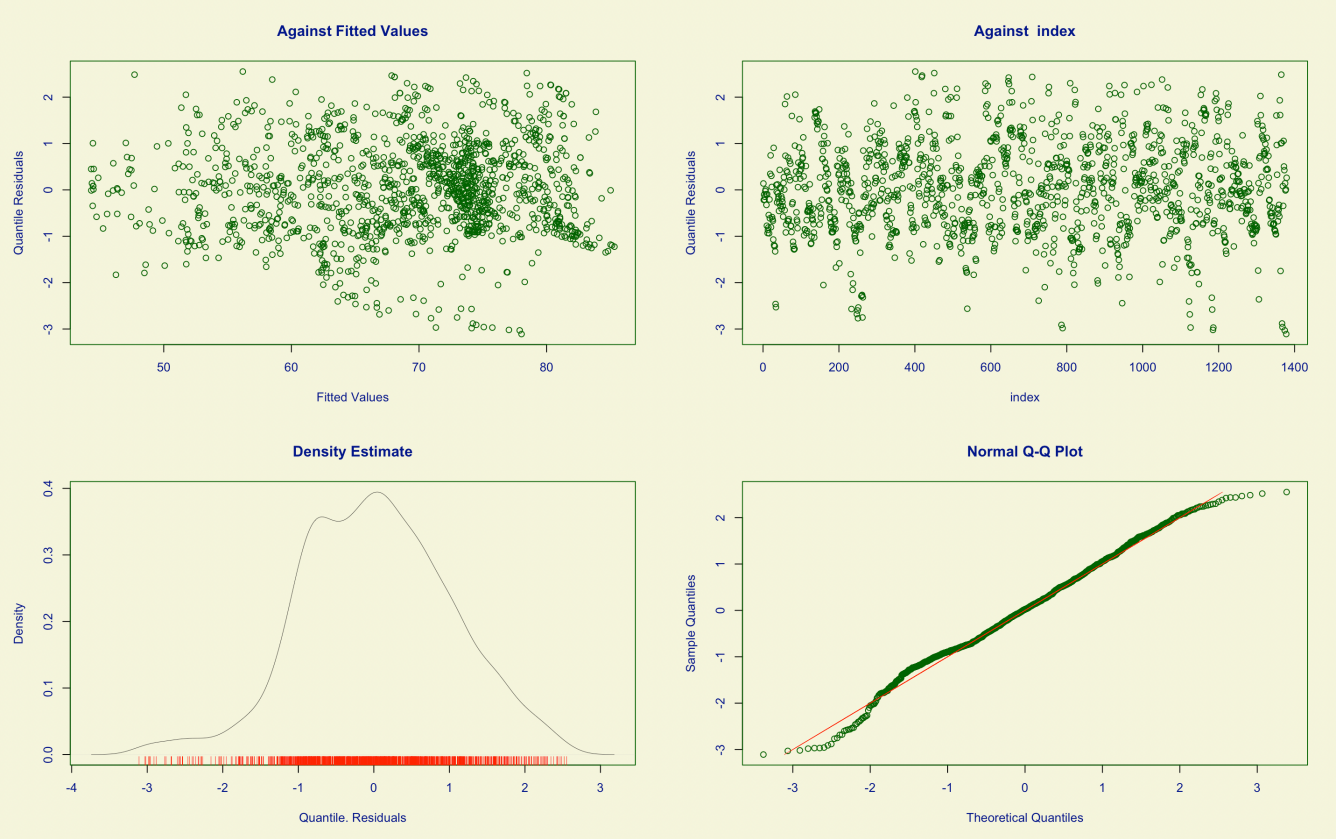


Figure 28: :Residual plot for model9

The effective degrees of freedom for the best model in Task3

```
$mu
$mu$`pb(BMI)`
[1] 17.58687

$mu$`pb(Adult.Mortality)`
[1] 21.61299

$mu$`pb(GDP)`
[1] 6.103205

$mu$`pb(Total.expenditure)`
[1] 2.000249

$mu$`pb(Population)`
[1] 2.000425
```

Figure 29: :Effective degrees of freedom for model9

Fitted smooth functions in the best model in Task3

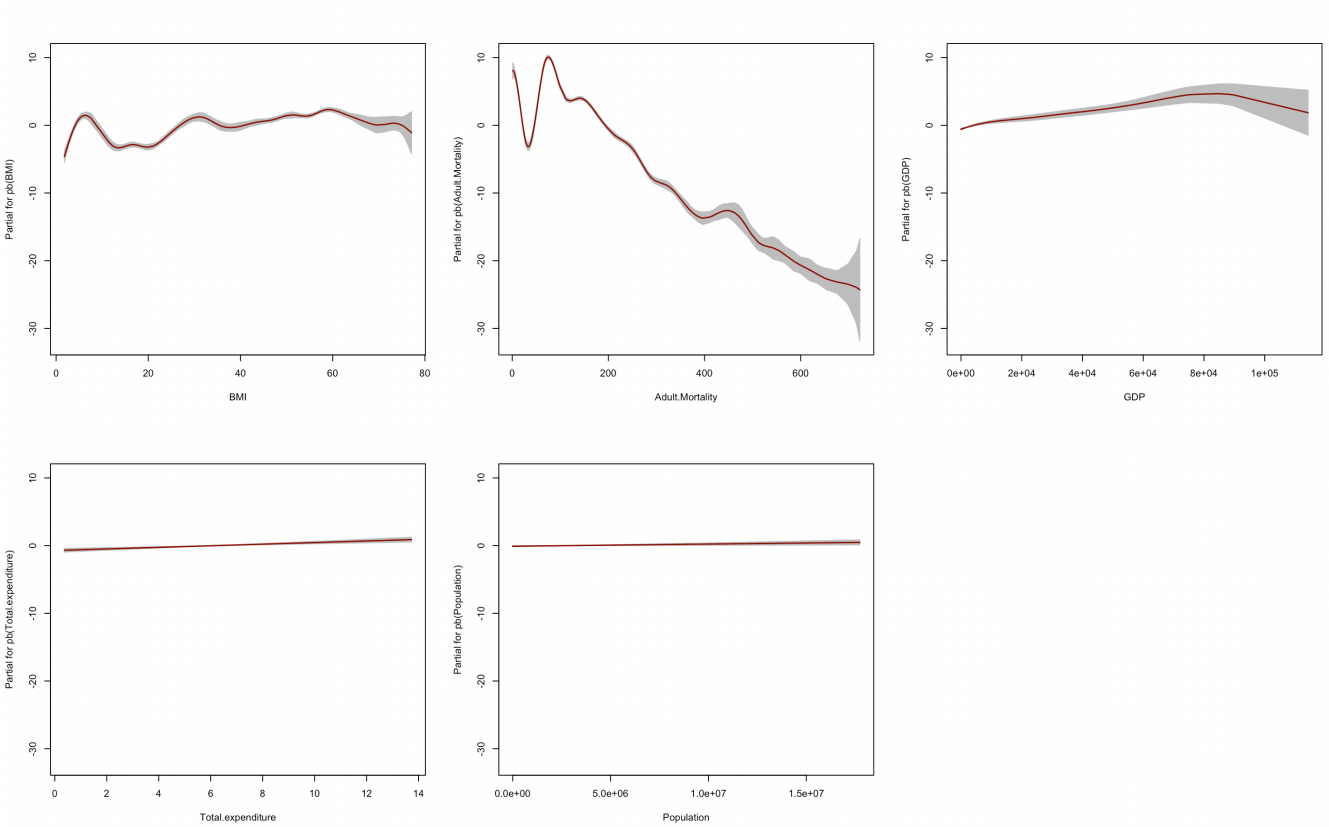


Figure 30: :Fitted smooth functions