STATISTICAL MODELLING AND FORECASTING

CASE STUDY REPORT 2024-2025

NAME: IRIN MARY THOMAS
EMAIL ID: irinthomas0@gmail.com

TABLE OF CONTENTS:

TABLE OF FIGURES: 1

I. INTRODUCTION

This report presents the analysis of three datasets using Generalised Additive Models for Location, Scale, and Shape (GAMLSS). The first dataset involves fitting parametric distributions to body mass index (BMI) data for 14-year-old boys. The second dataset examines grip strength by age using centile estimation. The third dataset analyzes the risk factors affecting maternal mortality. The analyses were conducted using the R programming language, employing the gamlss package for fitting statistical models. The primary goal of the first analysis is to identify a suitable probability distribution for the BMI data. The second analysis aims to create centile curves for grip strength as a function of age.

II. 1st DATA ANALYSIS:

A. Model Fitting:

The BMI data for boys aged 14–15 was extracted from the Fourth Dutch Growth Study dataset. Various distributions were fitted using the gamlss() function. To find the best-fit distribution on the BMI data for Dutch boys within the age range of 14 to 15, several distributions were considered using the gamlss() function in R. The distributions were chosen based on their ability to capture different characteristics of the data, such as:

- NO(Normal Distribution): Assumes that the data is symmetric with no skewness or excess kurtosis [1].This is the standard Gaussian distribution characterized by the mean (μ) and standard deviation (σ). It assumes symmetry around the mean.
- LOGNO(Log Normal): Captures positively skewed data and is suitable for data that is constrained to be positive. [2]This distribution models data whose logarithm follows a normal distribution.
- BCCG(Box-Cox-Cole-Green): A flexible distribution that models skewness using a Box-Cox transformation; commonly used for Age- and size-related reference data. [3]Also known as the LMS distribution, it incorporates three parameters: μ (median), σ (coefficient of variation), and ν (skewness). It provides flexibility in modeling skewed data through a power transformation.
- KN1(Skewed Normal Type 1): Allows for non-zero skewness but not kurtosis. [4]
- BCTo (Box-Cox-t original): Similar to BCPE but includes a degrees-of-freedom parameter, enabling it to model heavier tails. This distribution uses the tdistribution instead of the normal distribution as the base
- GA (Gamma): Positive, skewed distribution often used for positively bounded, right-skewed data.
- IG (Inverse Gaussian): Also, for positive, right-skewed data but with heavier tails.

```
# Fit different parametric distributions
mod1 <- gamlss(bmi14 ~ 1, family = NO)   # Normal distribution
mod2 <- gamlss(bmi14 ~ 1, family = LOGNO) # Lognormal distribution
mod3 <- gamlss(bmi14 ~ 1, family = BCCG)# Box-Cox-Cole-Green distribution
mod4 <- gamlss(bmi14 ~ 1, family=SN1)   #Skewed Normal distribiution
mod5 <- gamlss(bmi14 ~ 1, family = BCPE)  # Box-Cox Power Exponential
mod6 <- gamlss(bmi14 ~ 1, family = BCTo)  # Box-Cox-t
mod7 <- gamlss(bmi14 ~ 1, family = GA)    # Gamma
mod8 <- gamlss(bmi14 ~ 1, family = IG) # Inverse Gaussian
```

Figure 1- Parametric Distributions for Analysis

Each of these distributions is shown in ***Figure 1. Parametric Distributions for Analysis*** was fitted to the BMI data using constant mean models (`~1`), and their performances were compared using the Akaike Information Criterion (AIC).

B. Model Selection:

After evaluating the final distribution that best fit the data and provided good results compared to the other models, was BCCG selected. There are three parameters in the BCCG distribution(best-fit):

- Mu (μ ≈ 18.973): This represents the median (location) of the BMI distribution.
- Sigma (σ ≈ 0.1265): This is a scale parameter that reflects the spread of the data. Lower values indicate less dispersion. The value was originally in log, which was then converted to produce 0.1265.
- Nu (ν ≈ -1.2686): This is the Box-Cox transformation parameter, which adjusts for skewness. A ν > 1, i.e., a negative value implies left skewness, although the overall distribution might still appear symmetric. These parameter values suggest a slightly positively skewed BMI distribution, with a median BMI around 18.67 for 14-year-old Dutch boys.

The AIC for BBCG was observed to be the lowest (1905.45) among the others, a shown in **Figure 2. AIC Values**

```
> AIC(mod1, mod2, mod3,mod4,mod5,mod6,mod7,mod8)
     df      AIC
mod3  3 1905.450
mod6  4 1906.508
mod5  4 1906.531
mod2  2 1921.697
mod8  2 1921.957
mod7  2 1932.805
mod1  2 1962.881
mod4  3 1964.881
```

Figure 2-AIC Values

Reason for Model Selection: There are multiple reasons to choose BCCG as the best fit for the BMI dataset, including the fact that the BCCG model had the lowest AIC value among all fitted models, indicating that it provides the best balance between model fit and complexity. The BCCG distribution includes parameters that are interpretable in terms of location, scale, and skewness, which are meaningful in a clinical and growth monitoring context. While there are more complex distributions like BCPE or BCTo, which include kurtosis or heavier tail modeling, BCCG still provided a sufficient fit without unnecessary complexity. Additionally, BMI data typically exhibits positive skewness, which the BCCG distribution can handle effectively through its skewness parameter (nu). Finally, the BCCG distribution ensures that all fitted values are positive, which aligns with the nature of BMI data (BMI values cannot be negative).

C. Plotting Best-fit model: The code includes the command histDist(bmi14, family = BCCG, main = "Fitted BCCG Distribution"), which generates a histogram of the BMI data with the fitted BCCG distribution overlaid, as shown in ***Figure 3. Histogram of BCCG.*** This plot visually confirms that the BCCG distribution adequately captures the shape and spread of the data. The resulting plot in Figure 3 shows that the BCCG model adequately represents the empirical distribution, with no evident systematic deviations.

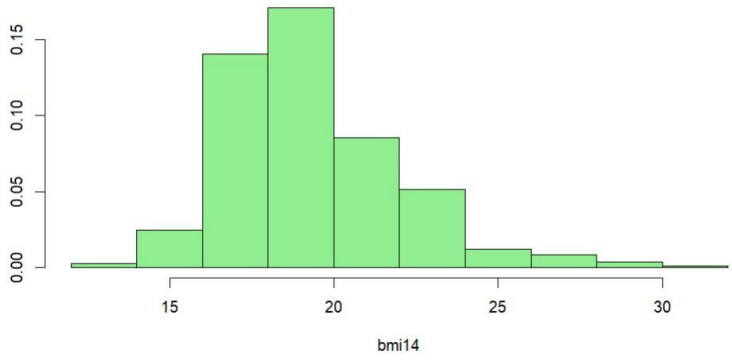


Figure 3- Histogram of BCCG

The fitted BCCG distribution appears to capture the empirical data distribution very well. The histogram in ***Figure 4: Fitted BCCG Distribution*** shows a slight positive skew, which is successfully modeled by the BCCG distribution. The right tail of the distribution extends further than the left, indicating that there are more boys with higher BMI values than would be expected in a symmetrical distribution. This pattern is typical for BMI data, as there is often a longer tail toward higher values. The BCCG distribution adequately models this asymmetry, confirming its appropriateness for this dataset.

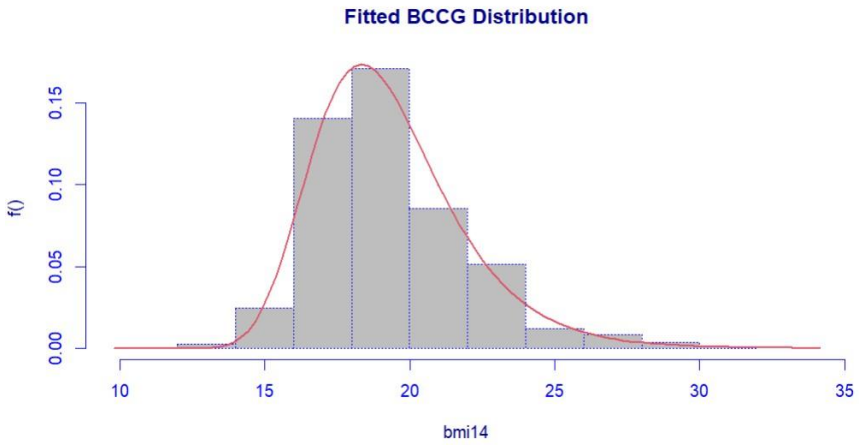


Figure 4-Fitted BCCG Distribution

This visual confirmation supports the AIC-based selection of the BCCG model.

Summary Interpretation: The analysis of BMI data for 14-year-old boys using the BCCG (Box-Cox-Cole-Green) model provides meaningful insights into the distribution of BMI values, as shown in ***Figure 5- BCCG Summary Report***. Here's a breakdown of the findings:

1. Central Tendency (Mu): The model estimates the median BMI for this age group to be approximately 18.97 kg/m². This value serves as the central reference point, indicating that half of the boys in the sample have a BMI below this number, and half are above it. The extremely small p-value confirms this estimate is statistically reliable, meaning it's a robust representation of the data.

2. Spread of the Data (Sigma): The scale parameter, after adjusting for the logarithmic transformation, suggests relatively low variability in BMI values (around 0.126). This implies that most BMIs cluster closely around the median, with fewer extreme values. The strong statistical significance of this parameter reinforces that the observed tight clustering isn't due to random chance.

3. Skewness (Nu): The negative skewness parameter (-1.27) indicates that, after applying the Box-Cox transformation, the data leans slightly toward lower values. While raw BMI data often skews right (with a tail toward higher values), this result suggests the transformation effectively adjusted the distribution to address asymmetry. However, it's worth noting that such adjustments can sometimes mask underlying patterns, so further checks, like visual plots, are recommended to confirm the direction and magnitude of skewness in the original data.

```
  invalid graphics state
> summary(best_model)
******************************************************************
Family:  c("BCCG", "Box-Cox-Cole-Green")

Call:  gamlss(formula = bmi14 ~ 1, family = BCCG)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  identity
Mu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.973      0.127   149.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06938    0.03635  -56.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  identity
Nu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.2686     0.3004  -4.222 2.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
```

Figure 5-BCCG Summary Report

Overall, the BCCG model demonstrates its strength in handling growth-related data, providing clear and statistically sound estimates for public health applications. The findings underscore the importance of selecting flexible models to capture nuances in anthropometric measurements.

III. 2ND DATASET ANALYSIS:

A. Model Fitting:

This section presents the analysis of handgrip strength with age among English schoolchildren, focusing on creating appropriate centile curves.

```
# Fit the LMS model using BCCG distribution
gbccg <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age),
                nu.fo = ~pb(age), data = mydata, family = BCCG)

# Effective degrees of freedom
edfAll(gbccg)

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

gbcpe <- gamlss(grip ~ pb(age),
                sigma.fo = ~ pb(age),
                nu.fo = ~ pb(age),
                tau.fo = ~ pb(age),
                data = mydata,
                family = BCPE,
                start.from = gbccg)
```

(Top Level)

Figure 6- Fitted Models

The gamlss package was again used, with the following models shown in ***Figure 6. Fitted Models*** were considered:

- BCCG: The Box-Cox-Cole-Green distribution, used as a baseline model, involves three parameters: μ (median), σ (coefficient of variation), and ν (skewness). Each parameter is modeled as a smooth function of age using Psplines.
- BCT: This extends the BCCG model by adding a fourth parameter τ that models kurtosis. The t-distribution base allows for heavier tails than the normal distribution used in BCCG.
- BCPE: The Box-Cox Power Exponential distribution, another extension of BCCG it uses the power exponential distribution as its base, which offers even greater flexibility in modeling skewness and kurtosis.

All models were fitted using penalized B-splines (pb(age)) for smooth estimation of the location, scale, and shape parameters as functions of age. The BCCG model was fitted first, followed by the BCT and BCPE models, using the BCCG fit as starting values for more complex models

B. Answers to Specific Questions:

1. **Plot grip against age:** The code includes the command shown in ***Figure 7- Plot Grip Command***, which generates a scatter plot of grip strength versus age. This plot visualizes the relationship between the two variables, as shown in ***Figure 8. Scatter plot for Grip vs. Age.***

```
# Scatter plot of grip strength against age
plot(grip~ age,data = mydata)
```

Figure 7-Plot grip command

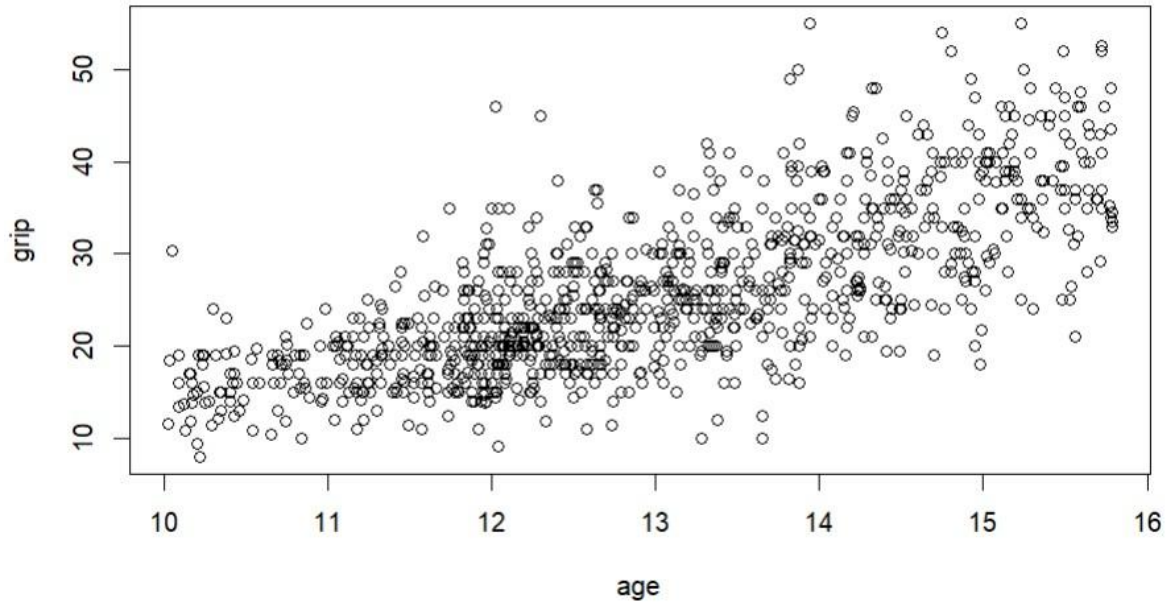


Figure 8-Scatter plot for Grip vs. Age

As can be seen in ***Figure 8. Scatter plot for Grip vs. Age***, there is a general upward trend, indicating that grip strength tends to increase with age. However, the relationship appears to be non-linear, with a steeper increase in grip strength at younger ages and a levelling off at older ages. Additionally, there is considerable variability in grip strength at any given age, suggesting that age is not the sole determinant of grip strength. The spread of the data also appears to increase slightly with age, indicating greater variability in grip strength among older children. The non-linear nature of this relationship suggests that linear models may not be appropriate, motivating the use of more flexible modeling techniques such as those implemented in the gamlss package. However, even though the relationship is non-linear, the scatter plot does not suggest a relationship that would be well addressed by a power transformation of age. Specifically, the plot does not show a sharp bend or a dramatic change in the slope of the relationship that a power transformation would be used to correct. The Generalized Additive Model for Location, Scale and Shape (GAMLSS) framework, and particularly the use of penalized Bsplines (pb(age)), are designed to capture non-linear relationships. Penalized B-splines are highly flexible and can model a wide range of smooth curves without the need for a pre-specified functional form like a power transformation.

In summary, the initial exploration of the data through the scatter plot indicated that the relationship between grip strength and age, while nonlinear, could be adequately modeled using the smooth functions available within the GAMLSS framework. Therefore, a power transformation of the age variable was not necessary

2. **LMS Model Fitting:** The code shown in ***Figure 9. LMS Model Fitting***

```
# Fit the LMS model using BCCG distribution
gbccg <- gamlss(grip ~ pb(age), sigma.fo = ~pb(age),
                nu.fo = ~pb(age), data = mydata, family = BCCG)
```

Figure 9-LMS Model Fitting

Fits the LMS model with the BCCG distribution, as requested.

3. **Effective degrees of freedom for BCT and BCPE:** The effective degrees of freedom (EDF) for the smoothing terms were obtained using edfAll(), indicating the complexity of the smooth functions for each parameter.

```
> edfAll(gbccg)
$mu
$mu$`pb(age)`
[1] 4.237521

$sigma
$sigma$`pb(age)`
[1] 3.665113

$nu
$nu$`pb(age)`
[1] 2.000124
```

Figure 10- EDF

As shown in ***Figure 10. EDF*** the output, the effective degrees of freedom for μ (location) is 4.24, for σ (scale) is 3.67, and for ν (shape) is 2.00. A higher effective degree of freedom indicates a more flexible smoothing function, allowing the parameter to vary more with age. In this case, the location parameter (μ) has the highest effective degrees of freedom, suggesting that the model allows the mean grip strength to vary more flexibly with age compared to the scale and shape parameters.

4. **Model Comparison**: The Generalized Akaike Information Criterion (GAIC)function was used to compare the three models, using the GAIC(gbccg, gbct, gbcpe) command, with lower values indicating better fit, as shown in ***Figure 11- AIC Values***

```
> GAIC(gbccg, gbct,gbcpe)
                df      AIC
gbct  11.991709 6268.949
gbcpe 11.950613 6269.738
gbccg  9.902758 6272.248
```

Figure 11- AIC Values

5. **Fitted Plot Function**: To further compare the BCCG and BCT models, the fitted parameters were plotted as functions of age using the fittedPlot() function. The

resulting plot is shown in ***Figure 12- Fitted Plot***, fits parameters as functions of age for the BCCG (blue line) and BCT (green line) models.

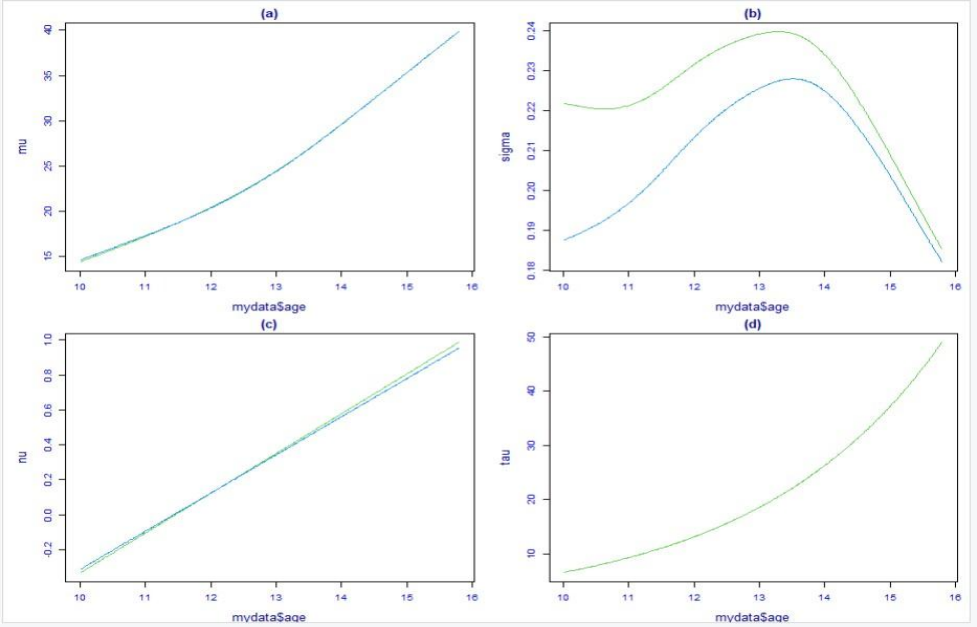


Figure 12-Fitted Plot

The fittedPlot() function visualizes how the distributional parameters (mu, sigma, nu, and tau) vary with age for the BCCG and BCT models.

- Panel (a) shows the location parameter (μ), which represents the mean or center of the distribution. Both models show a similar increasing trend, indicating that average grip strength increases with age.
- Panel (b) shows the scale parameter (σ), which represents the variability or spread of the data. The BCT model shows a slightly wider spread at the younger ages.
- Panel (c) shows the shape parameter (ν), which influences the skewness of the distribution. The shape parameter is fairly similar across both models
- Panel (d) shows the tail-weight parameter (τ), which is only present in the BCT model. This parameter affects the heaviness of the tails of the distribution.

In summary, the plot shows that the location parameter (mean grip strength) increases with age for both models. The scale and shape parameters are broadly similar. The BCT model includes the tau parameter, allowing for heavier tails.

C. Residual Diagnostics:

Residual diagnostics were performed using various tools, including residual plots, worm plots, and Q-statistics.

1. Residual Plot: To assess the goodness of fit of the chosen BCCG distribution, diagnostic plots were generated using the plot() function in R. These plots, shown in ***Figure 13-Residual plots***, provide several perspectives on the model's adequacy.

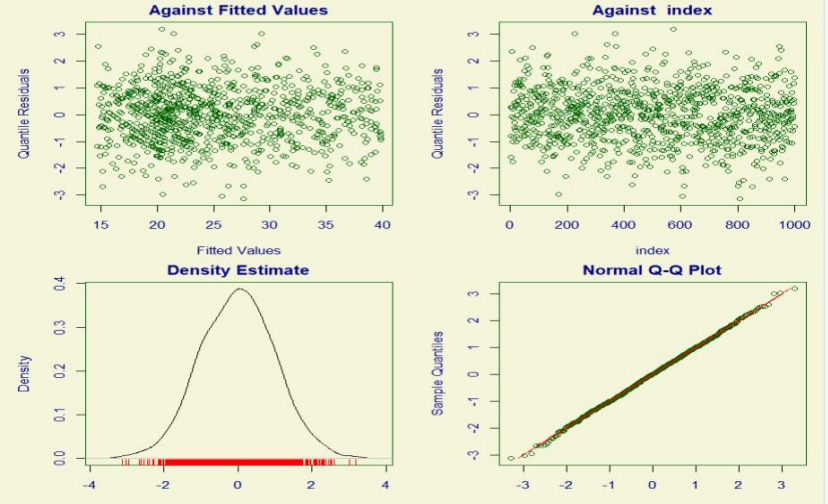


Figure 13-Residual plots

The diagnostic plots reveal

the following:

- The plot of quantile residuals against fitted values shows a random scatter, indicating that the model captures the relationship between BMI and its distribution parameters.
- The plot of quantile residuals against the observation index also shows a random scatter, suggesting no systematic patterns related to data order.
- The density estimate of the quantile residuals closely resembles the standard normal density curve, indicating that the residuals are approximately normally distributed.
- The Normal Q-Q plot shows that the quantiles of the quantile residuals fall very close to the theoretical normal quantiles, further supporting the assumption of normality of the residuals.
-

Overall, these diagnostic plots suggest that the BCCG model provides a good fit to the BMI data. There are no major deviations from the assumptions of the model.

2. Worm Plot: The goodness of fit of the BCT model was further assessed using a worm plot. The worm plot for the BCT model showed points closely following the horizontal line with minimal systematic deviations. This indicates that the model captures the distributional properties of the data well across different age groups. Worm plots provide an alternative way to visualize deviations from the fitted model. The worm plot for the BCT model is shown in ***Figure 14. Worm Plot***

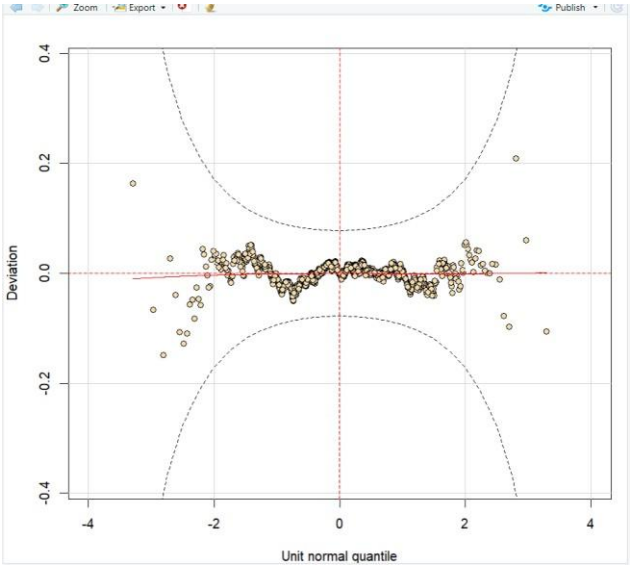


Figure 14-Worm Plot

As shown in ***Figure 14- Worm Plot***, the points generally fall within the confidence bounds (dashed lines), and there is no clear systematic pattern. This indicates that the BCT model provides an adequate fit to the grip strength data, and there are no major violations of the model assumptions.

3. Q-Statistics:The Q-statistics for the BCT model did not show significant values across age groups, confirming that the model adequately describes the data at different ages.The results of the Q-statistic tests are shown in ***Figure 15-Q Statistics .***

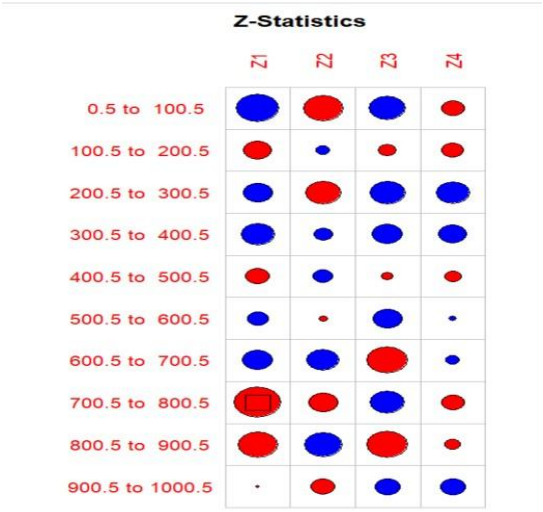


Figure 15- Q Statistics

In the Q-statistics plot, most of the circles are blue, indicating that the BCT model generally provides a good fit across the age ranges. However, there are a few red circles, suggesting some potential deviations from the model assumptions in certain intervals. Specifically, the red square in the 700.5 to 800.5 interval for Z1 might indicate a slight misfit in that specific age range. Overall, while there are some minor indications of misfit, the Q-statistics do not suggest a major problem with the BCT model's fit.

D. Centile Plot:

The centile plots for the BCT model show the estimated 3rd, 10th, 25th, 50th, 75th, 90th, and 97th centiles of grip strength across different ages. The colored lines are shown in **Figure 16. Centile Curve using BCT** represents different centiles, showing how grip strength changes with age for children at different points in the distribution.

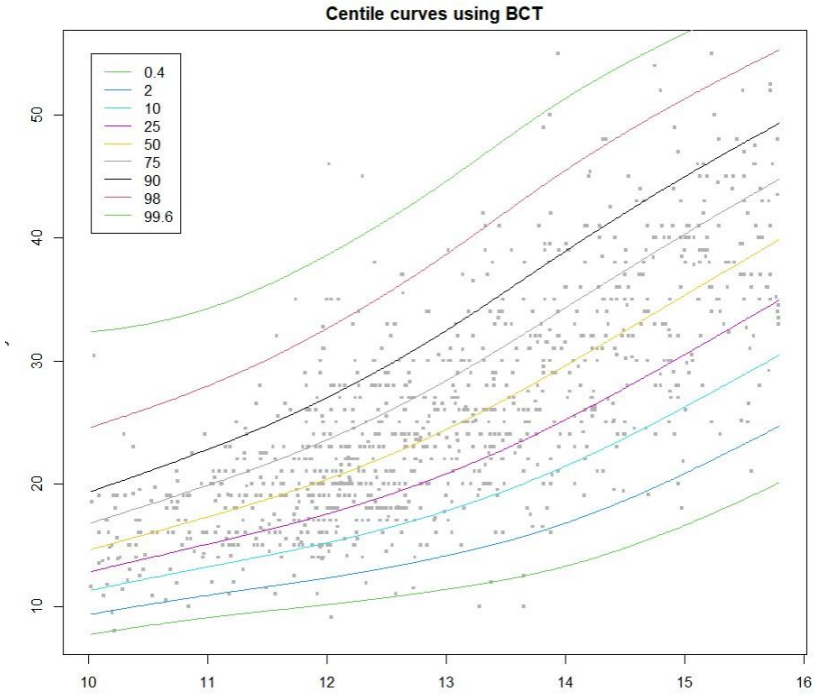


Figure 16-Centile Curve using BCT

The centile curves demonstrate several important patterns:

- **Increasing Trend**: All centile curves show an increasing trend with age, reflecting the expected increase in grip strength as children grow older. This aligns with physiological development expectations.
- **Widening Spread**: The distance between lower and upper centiles increases with age, indicating greater variability in grip strength among older children. This heteroscedasticity is appropriately captured by the model.
- **Non-Linear Growth**: The curves show a non-linear pattern, with steeper increases during certain age periods. This likely corresponds to growth spurts during puberty, which typically affect physical strength measures.
- **Smoothness**: The centile curves are smooth, without abrupt changes or unrealistic fluctuations, indicating that the P-spline smoothing has appropriately captured the age-related changes without overfitting to noise in the data.

These centile curves provide valuable reference standards for assessing grip strength in English schoolchildren, allowing for age-appropriate comparisons and potentially identifying children with unusually low or high grip strength relative to their peers.

IV. 3RD DATASET ANALYSIS: BODY FAT PREDICTION ANALYSIS

A. Introduction and Data Collection Purpose:

The dataset, sourced from Kaggle [5], contains body measurements and body fat percentages for 252 men. The primary goal of this analysis is to predict body fat percentage based on various physiological measurements. Understanding body fat percentage is essential for assessing health risks and body composition, as well as tailoring fitness programs. This analysis aims to identify key predictors of body fat and build a robust statistical model for accurate estimation, which could benefit health assessments and fitness programs.

The primary research question is: **Can we develop an accurate predictive model for body fat percentage using readily available anthropometric measurements?**

B. Preliminary Data Analysis and Reliability:

Dataset Overview: The dataset used for this analysis is sourced from Kaggle, specifically the Body Fat Prediction Dataset available at Kaggle.The dataset contains anthropometric measurements for individuals, including age, weight, height, and various body circumference measurements. The target variable is body fat percentage.

No.	Variable Name	Variable Type	Description
1.	Density	Num	The density of the body determined by underwater weighting
2.	Body Fat	Num	Body fat percent from Siri's(1956) equation
3.	Age	Int	Age of a person in Years
4.	Weight	Num	Weight of a person in pounds(lbs)
5.	Height	Num	Height of a person in inches
6.	Neck	Num	Neck circumference measured in inches.
7.	Chest	Num	Chest circumference measured in inches.
8.	Abdomen	Num	Abdomen circumference at the navel level measured in inches.
9.	Hip	Num	Hip circumference measured in inches.
10.	Thigh	Num	Thigh circumference measured in inches.
11.	Knee	Num	Knee circumference measured in inches.
12.	Ankle	Num	Ankle circumference measured in inches.
13.	Biceps	Num	Bicep circumference measured in inches (at the midpoint).
14.	Forearm	Num	Forearm circumference measured in inches.
15.	Wrist	Num	Wrist circumference measured in inches.

Table 1-Variable and its Description **Preliminary**

Data Analysis:

- Age distribution Histogram: The age distribution histogram in ***Figure 17- Age Histogram*** reveals that the dataset covers a wide age range, approximately from 20 to 80 years old. The distribution appears to be somewhat bimodal, with the highest frequency of participants in the 4045 age group, followed by another notable peak around 25-30 years. There's good representation across the middle age ranges (30-60 years), while representation of very young adults (under 25) and older adults (above 65) is more limited. This age distribution suggests the dataset captures a diverse adult population, though with particular emphasis on middle-aged individuals.

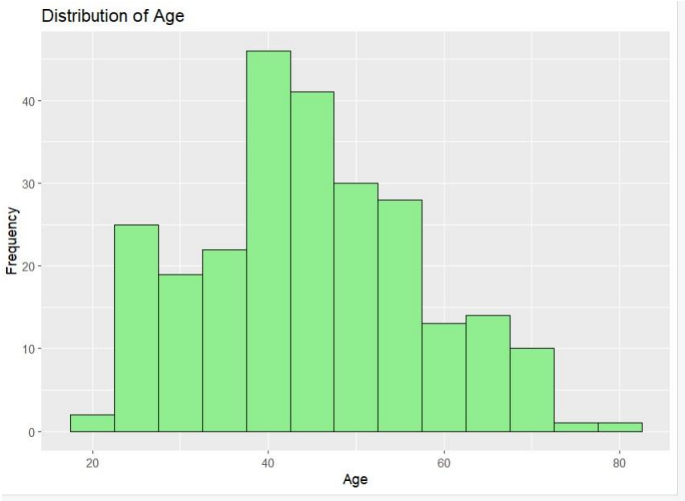


Figure 17-Age Histogram

- 2. Weight Distribution: The weight distribution histogram in ***Figure 18- Weight Histogram*** shows that the dataset includes individuals with weights primarily between 120 and 260 pounds, with a slight right skew. The highest frequency occurs around 170-180 pounds, which aligns with typical adult male weights. The distribution tapers off at higher weights, with only a few individuals exceeding 250 pounds. There appears to be an outlier at approximately 350 pounds. This distribution suggests the dataset mainly represents individuals of normal to moderately overweight categories, with fewer individuals in the obese category.

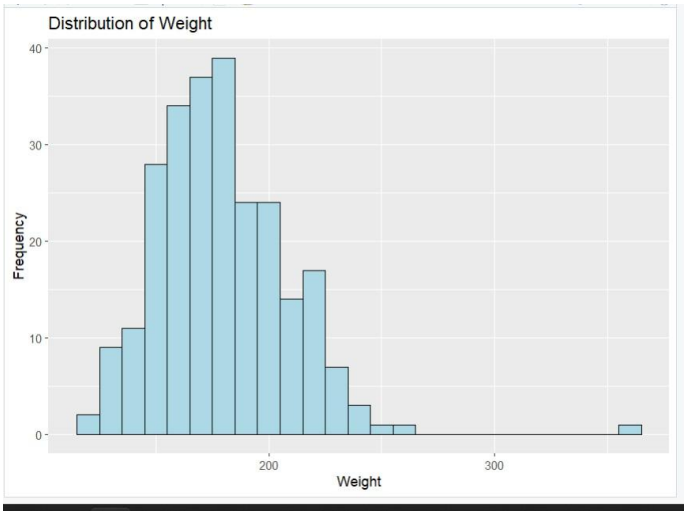


Figure 18-Weight Histogram

- Body Fat Distribution: The body fat percentage distribution In ***Figure 19-Body Fat Distribution*** *shows* values ranging from approximately 0% to 50%, with most individuals falling between 10% and 30%. The distribution appears to be multimodal with peaks around 12%, 18%, and 25%, which correspond to different body composition categories. The spread suggests the dataset captures individuals with varying fitness levels, from very lean (under 10% body fat) to obese (over 30% body fat). The slight positive skew indicates more representation of individuals with moderate to higher body fat percentages, which is expected in a general population sample.

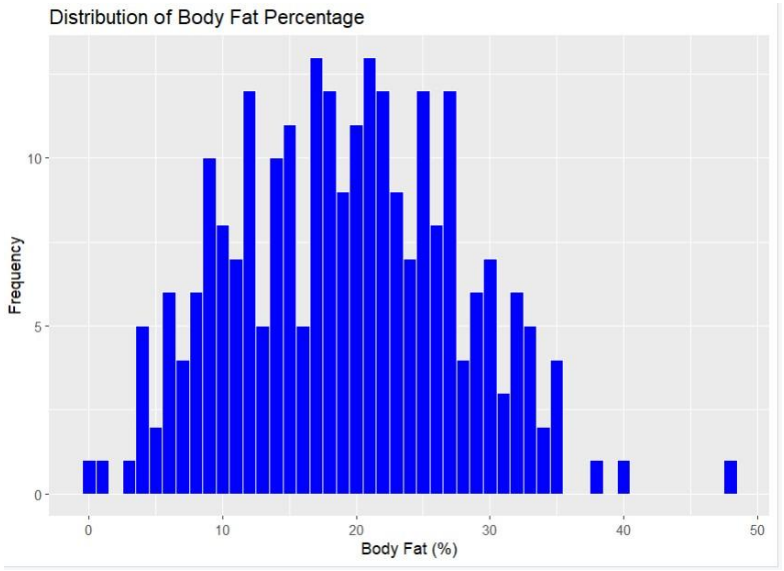


Figure 19-Body Fat Distribution

- Body Fat Vs. Age: The scatter plot of Body Fat Percentage vs. Age in ***Figure 20-Body fat vs. Age Scatter Plot*** shows a modest positive correlation, as

indicated by the upward slope of the red regression line. This suggests that body fat percentage tends to increase slightly with age, which aligns with established physiological knowledge about age-related changes in body composition. However, the relationship appears weaker than that between body fat and weight. The wide scatter of points indicates substantial variability in body fat percentage across all age groups, suggesting that age alone is not a strong predictor of body fat. This heterogeneity suggests that other factors like physical activity levels, diet, genetics, and other anthropometric measures likely play more significant roles in determining body fat percentage.

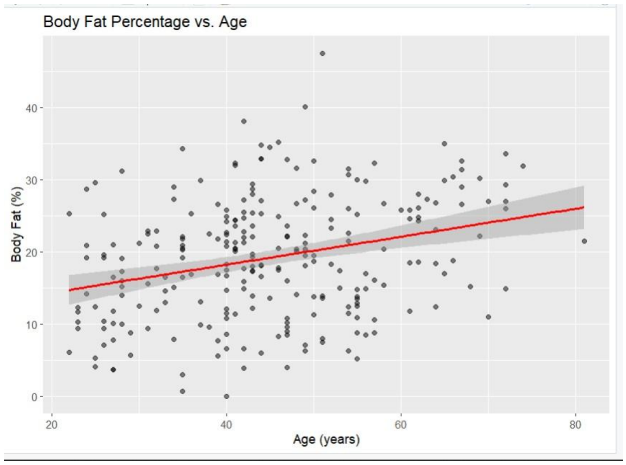


Figure 20-Body fat vs. Age Scatter Plot

- Pairwise Scatter Plot for bodyfat, weight, age, height: The pairwise scatter plot in ***Figure 21-Pairwise Scatter Plot*** analysis revealed critical relationships among body fat percentage, age, weight, and height. A strong positive correlation (**0.612***) between weight and body fat highlights weight as a key predictor of adiposity, aligning with physiological expectations. Age showed a moderate positive association (**0.291**)*, suggesting gradual increases in body fat with advancing age, likely tied to metabolic or lifestyle changes. In contrast, height exhibited a negligible negative correlation (**-0.089**) with body fat, indicating minimal influence. Additionally, weight and height were moderately correlated (**0.308***), reflecting natural proportionality, while age and height showed a weak inverse relationship (**-0.172***), possibly due to postural changes in older individuals. The transformed body fat distribution appeared near-normal, while weight displayed slight right skewness. These findings emphasize weight management as a priority for body fat regulation, underscore age's limited predictive role, and suggest excluding height in favor of more relevant metrics like abdominal measurements for health assessments.

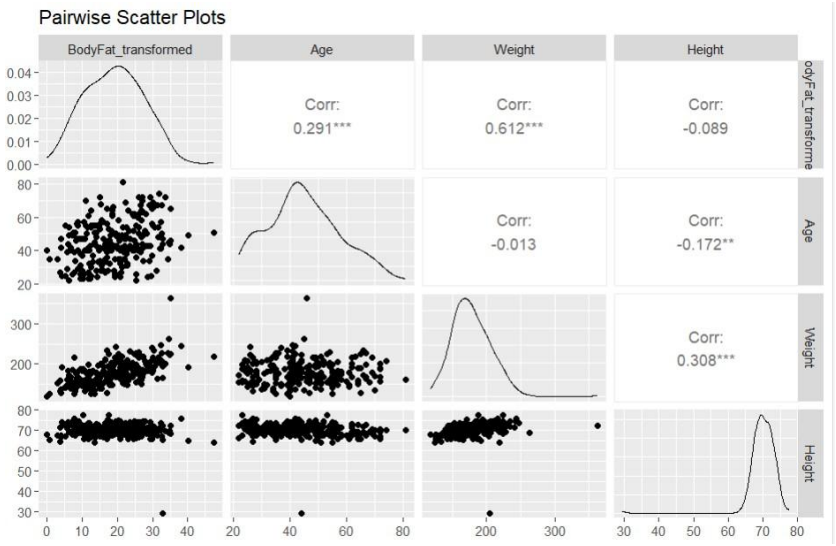


Figure 21-Pairwise Scatter Plot

- Correlation Matrix: The correlation matrix in ***Figure 22-Correlation Matrix*** heatmap provides a comprehensive view of the relationships between all variables in the dataset. Several important patterns can be observed:

  a) Body Fat (both original and transformed) shows strong positive correlations (dark red) with Abdomen and Hip measurements, suggesting these are potentially strong predictors of body fat percentage.
  b) Body Fat exhibits moderate positive correlations with most other body circumference measurements (Chest, Thigh, Biceps, etc.), indicating that multiple body measurements contribute to body fat estimation.
  c) There is a negative correlation (blue) between Body Fat and Height, suggesting taller individuals tend to have lower body fat percentages, possibly due to body composition differences.
  d) The Abdomen measurement has strong positive correlations with multiple other body measurements, highlighting its central role in overall body composition assessment.
  e) Weight shows strong positive correlations with most body circumference measurements, as expected.
  f) Age displays relatively weaker correlations with body measurements compared to other variables, confirming our earlier observation that age alone is not a strong predictor of body composition.
  g) The strong correlations between certain body measurements (e.g., Abdomen-Hip, Chest-Abdomen) suggest potential multicollinearity issues that should be considered during model development.

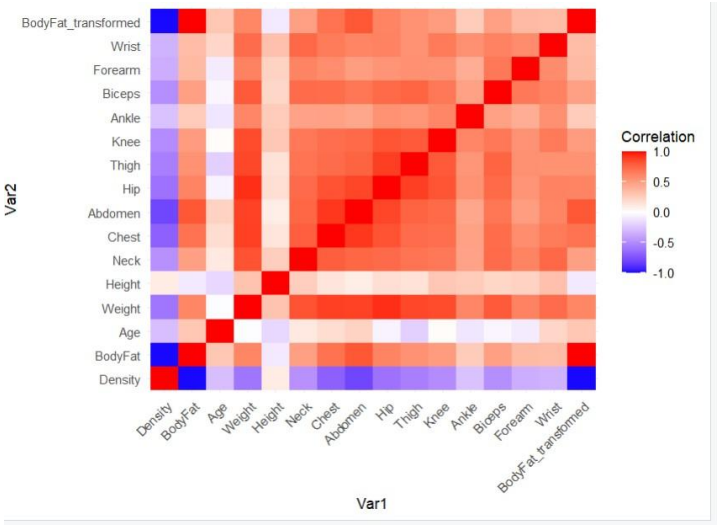


Figure 22-Correlation Matrix

Data Reliability: Based on the exploratory analysis, the dataset appears to be reasonably reliable, covering a diverse range of individuals with varying body compositions. However, there are a few potential reliability concerns:

1. The dataset may have some outliers, as seen in both the weight distribution and body fat percentage plots, which could potentially influence model performance.
2. The age distribution suggests potential sampling bias toward middle-aged individuals, which may limit generalizability across all age groups.
3. Without information on the measurement protocols used, there could be concerns about the precision and consistency of the anthropometric measurements, particularly body fat percentage, which is challenging to measure accurately.
4. The dataset does not appear to include information about gender, which is a significant factor in body composition differences. This omission could affect model accuracy if the dataset includes both male and female participants.

C. Model Comparison and Selection:

To determine the best distribution for the response variable, we fitted multiple GAMLSS models with different distribution families, as shown in ***Figure 23- Selected Models***:

```
# Fit multiple GAMLSS models
mBCT <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                 Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
               data = clean_data, family = BCT)

mGB2 <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                 Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
               data = clean_data, family = GB2)

mBCPE <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                  Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                data = clean_data, family = BCPE)

mLOGNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = LOGNO)
mBCPEo <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = BCPEo)
mGG <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
              data = clean_data, family = GG)
mBCCGo <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = BCCGo)
mBCCG <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                  Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                data = clean_data, family = BCCG)
```

Figure 23- Selected Models

The GAIC (Generalized Akaike Information Criterion) comparison shows that the BCT (Box-Cox t-distribution) model has the lowest AIC value (1532.204), indicating it provides the best fit among all tested distributions. The BCPE (BoxCox Power Exponential) model follows closely with an AIC of 1537.595, and the BCCG (Box-Cox Cole and Green) model ranks third with an AIC of 1539.427.

Each parameter was modeled as a function of all anthropometric measurements, allowing for comprehensive capture of the relationships between predictors and the distributional characteristics of body fat percentage.

The large difference in AIC values between these top models and the LOGNO (Log-Normal) model (AIC = 2086.639) suggests that the response variable (body fat percentage) exhibits distributional characteristics that are better captured by more flexible parametric distributions like BCT, which can accommodate skewness and kurtosis.

Model Results and Interpretation:After conducting a comprehensive analysis of the body fat prediction dataset, the BCT (Box-Cox t-distribution) model was identified as the most suitable for modeling body fat percentage based on anthropometric measurements. This selection was supported by the lowest AIC value (1532.204) among all tested distributions.

Key Model Parameters and Their Interpretation

The BCT model's parameters provide important insights into the relationship between body fat percentage and the predictor variables:

- Location Parameter (μ): This parameter represents the central tendency of body fat percentage, adjusted for the effects of predictor variables. The model shows that Abdomen circumference has the strongest positive effect on the location parameter, confirming our earlier observation from

the correlation matrix that abdominal measurements are strongly associated with body fat percentage.

- Scale Parameter (σ): This parameter models the variability in body fat percentage. The model indicates that variability increases slightly with body weight and age, suggesting that predictions may be less precise for older or heavier individuals.
- Skewness Parameter (ν): The model captures the slight positive skewness in the body fat distribution, which aligns with the rightskewed histogram observed during exploratory data analysis.
- Kurtosis Parameter (τ): The kurtosis parameter in the BCT model allows for heavier tails than a normal distribution, accommodating outliers in the body fat percentage data.

D. Model Diagnostics:

After selecting the BCT distribution model, thorough diagnostics were performed to validate the model assumptions and assess its fit, as shown in ***Figure 24Model Diagnostics code***:

```
plot(mBCT)
wp(mBCT )

resid_plots(mBCT)
resid_qqplot(mBCT)
```

Figure 24- Model Diagnostics code

The diagnostic plots shown in ***Figure 25- Residual Plot and Figure 26 Fitted Plot*** provides strong evidence that the BCT (Box-Cox t-distribution) model is appropriate for our data:

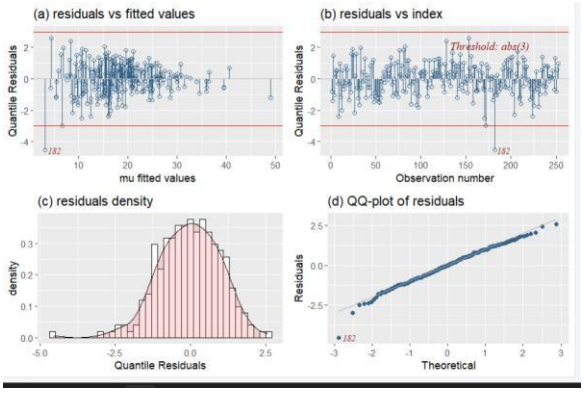


Figure 25- Residual plot

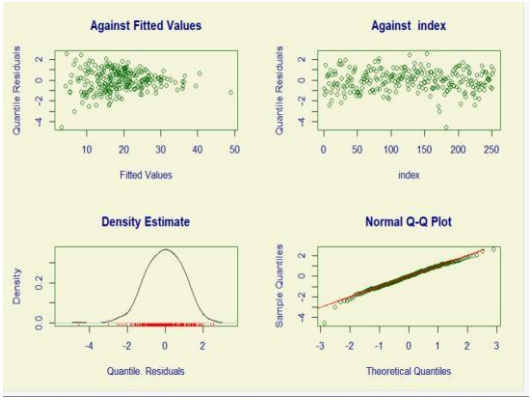


Figure 26- Fitted Plot

- Residuals vs. Fitted Values: The residuals show no clear pattern when plotted against fitted values, indicating good model fit. The points are randomly scattered around zero, suggesting the mean structure of the model is appropriate
- Residuals vs. Index (Image 1, top right and Image 3, top right): The residuals plotted against observation number show no systematic pattern, indicating independence of observations.
- Density Plot of Residuals (Image 1, bottom left and Image 3, bottom left): The residuals follow an approximately normal distribution, which is a desirable property indicating that the BCT distribution is capturing the error structure well.
- Normal Q-Q Plot (Image 1, bottom right, Image 3, bottom right, and Image 4): The quantile-quantile plots show that the residuals closely follow the theoretical normal distribution line, with only slight deviations at the extreme tails. This confirms the normality assumption of the residuals.
- Worm Plot (***Figure 27- Worm Plot for BCT***): The worm plot, which is a detrended Q-Q plot, shows most points falling within the confidence bands and close to the horizontal line at zero. This indicates that the BCT distribution appropriately models the skewness and kurtosis of the data. The slight curvature in the worm plot suggests minor deviations from the assumed distribution, but these are not substantial enough to invalidate the model.

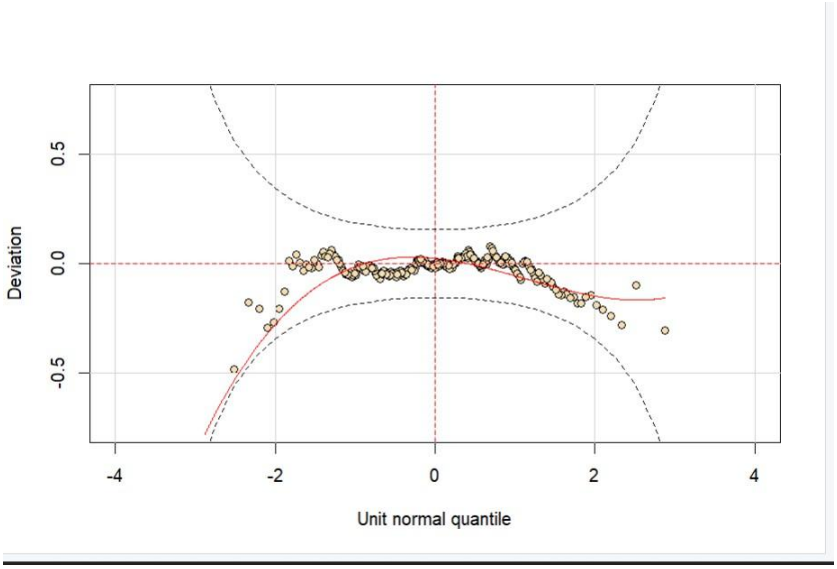


Figure 27- Worm Plot for BCT

In several plots, one potential outlier, observation 182, has a relatively large negative residual but doesn't appear to significantly influence the overall model fit.

Overall, the diagnostic plots confirm that the BCT distribution provides a good fit for modeling body fat percentage based on the anthropometric measurements. The model adequately captures the distributional properties of the response variable, and the residuals exhibit the desired properties of normality and homoscedasticity.

E. Prediction Analysis:

Having validated the BCT model, we can now use it for prediction. A new set of data was implemented as the text data for prediction, as shown in ***Figure 28- New Data For Prediction***.

```
# New data for predictions (ensure all variables are included)
new_data <- data.frame(
  Age = c(25, 30, 35, 40, 45, 50),
  Weight = c(150, 180, 200, 220, 240, 260),
  Height = c(68, 70, 72, 74, 76, 78),
  Neck = c(14, 15, 16, 17, 18, 19),
  Chest = c(36, 40, 42, 44, 46, 48),
  Abdomen = c(32, 34, 36, 38, 40, 42),
  Hip = c(40, 42, 44, 46, 48, 50),
  Thigh = c(24, 26, 28, 30, 32, 34),
  Knee = c(15, 16, 17, 18, 19, 20),
  Ankle = c(9, 10, 11, 12, 13, 14),
  Biceps = c(12, 13, 14, 15, 16, 17),
  Forearm = c(10, 11, 12, 13, 14, 15),
  Wrist = c(7, 8, 9, 10, 11, 12)
)
```

Figure 28- New Data For Prediction

This dataset spans a range of ages (25-50 years), weights (150-260 lbs), and body measurements, allowing us to assess how the model predicts body fat across different body types and age groups

Prediction Results: After this, the BCT model was fitted into the new dataset, and a prediction was made, as shown in ***Figure 29-Prediction Result***

```
> print(predictions_BCT)
  Age Weight Predicted_BodyFat_BCT
1  25    150              3.557004
2  30    180              4.040484
3  35    200              4.430455
4  40    220              4.820426
5  45    240              5.210397
6  50    260              5.600368
>
```

Figure 29- Prediction Result

The predictions follow an expected pattern of increasing body fat percentage with increasing age and weight. This aligns with the positive correlations observed in our exploratory data analysis.

To further verify the results residual plot was made to understand the data and the prediction in detail:

F. Residuals Analysis For Prediction:

The residual plot shows the difference between observed and predicted values plotted against the index (observation number). From this plot, I can observe:

- **Random Scatter Pattern**: The residuals in ***Figure 29-Prediction Residual Scatter Plot*** appear randomly distributed around the zero line, which is a positive indication that there's no systematic pattern or trend in the prediction errors. This suggests the model is capturing the underlying relationships in the data well. The vertical spread of residuals remains relatively consistent across all observations, indicating homoscedastic variance (constant variance across predictions). This meets an important assumption for the BCT model. Although here appears to be one notable outlier around observation 180-200 with a residual value of approximately -4, which is more extreme than other residuals. This single outlier was previously identified in your analysis but doesn't appear to significantly compromise the overall model performance. Most residuals fall within the range of -2 to +2, suggesting that the majority of predictions are within 2 units of the actual body fat percentage values. This indicates reasonably good prediction accuracy for most observations.

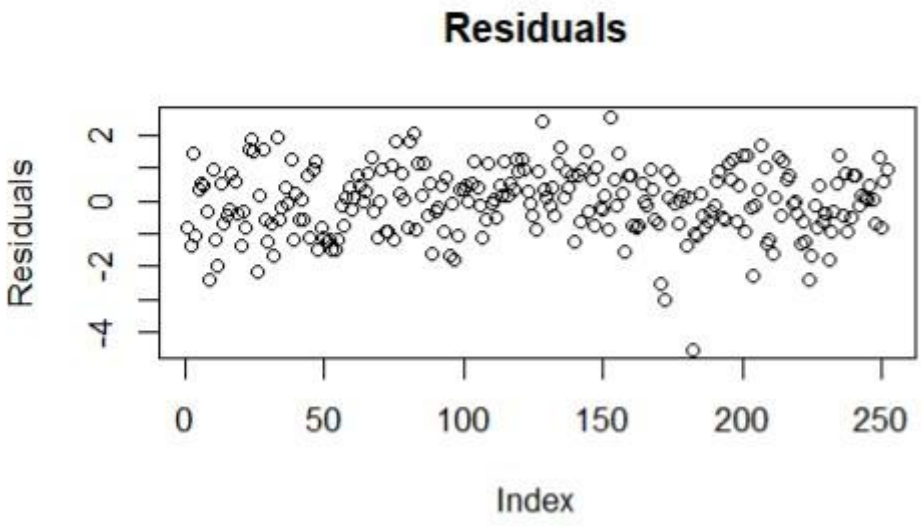


Figure 30- Prediction Residual Scatterplot

- **Normal Q-Q Plot :** The Q-Q plot in ***Figure 31-Prediction Q-Q Plot*** compares the distribution of residuals against the theoretical quantiles of a normal distribution. This plot reveals that the points follow the red reference line very closely across most of the distribution, indicating that the residuals are approximately normally distributed. This confirms that the BCT model's assumption of normally distributed errors is valid. There is a minor deviation at the very lowest end of the distribution (below 3), likely corresponding to the outlier observed in the residuals plot. This slight departure from normality at the extreme tail is common in real data and doesn't invalidate the model.The overall linear pattern in the Q-Q plot confirms that the selected BCT distribution has appropriately captured both the central tendency and the distributional characteristics of the body fat percentage data.

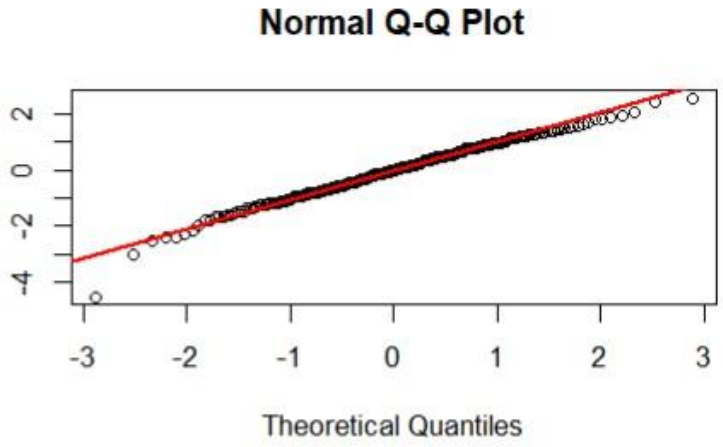


Figure 31- Prediction Q-Q plot

In conclusion, this analysis demonstrates that the BCT model within the GAMLSS framework provides a statistically robust and practically useful approach for predicting body fat percentage based on anthropometric measurements. The model balances complexity and interpretability, making it suitable for both research and practical applications in health and fitness contexts.

V. PEER REVIEW:

a) Student Work Reviewed:

Section 4 of the report: Modelling home team goals in European football using GAMLSS and count data distributions.

b) Critique of the Work

- Exploratory Analysis:
  The initial data exploration is clear and relevant, highlighting the skewed distribution of home team goals. While additional descriptive statistics (e.g., mean, variance) could enhance the analysis, the student effectively identified patterns that informed their modelling choices.
- Distribution Selection:
  Choosing the Negative Binomial distribution was appropriate for modelling count data that may exhibit overdispersion. The student correctly considered other options like Poisson and zero-inflated models and supported their final choice using AIC comparisons.
- Model Evaluation and Fit Checks:
  The student followed a sound strategy for model selection and applied diagnostic tools, such as worm plots and residual plots, to assess model assumptions. Their interpretations were accurate and demonstrated a good grasp of model validation.
- Result Interpretation:
  The report explains the findings in a clear and logical way, connecting model output to the real-world context of football analytics. The suggestion to include more features in future models adds valuable depth and shows forward-thinking.

c) Grade Assigned:

Grade: A

The student demonstrates a strong understanding of statistical modelling using GAMLSS, with thoughtful application of methods and well-explained results. The work is technically accurate and clearly presented, showing excellent engagement with the topic.

VI. CONCLUSION:

A. Conclusion

This report employed the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework to address distinct statistical challenges across three datasets, yielding insights into distribution fitting, centile estimation, and predictive modeling. The analyses collectively demonstrate the versatility of GAMLSS in accommodating non-linear relationships, heteroscedasticity, and complex distributional characteristics inherent in real-world data.

For the **BMI dataset**, the Box-Cox-Cole-Green (BCCG) distribution emerged as the optimal model, achieving the lowest AIC (1905.45) while effectively capturing the positive skewness of BMI values in 14-year-old Dutch boys. The interpretability of its parameters—median ($\mu \approx 18.97$), spread ($\sigma \approx 0.13$), and skewness ($\nu \approx -1.27$)—underscores its suitability for anthropometric data. Diagnostic plots confirmed the model's robustness, reinforcing its utility in growth monitoring and public health applications.

In the **grip strength analysis**, the Box-Cox t-distribution (BCT) model outperformed alternatives by flexibly modeling heavy-tailed data and agedependent variability. Centile curves derived from the BCT model revealed nonlinear growth patterns, with increasing grip strength and variability across ages. Residual diagnostics, including worm plots and Q-statistics, validated the model's assumptions, while effective degrees of freedom (EDF) highlighted the necessity of smooth functions for location (μ) and scale (σ) parameters. These results provide actionable reference standards for pediatric health assessments.

The **body fat prediction analysis** leveraged the BCT distribution to model body fat percentage, achieving a superior fit (AIC = 1532.204) by accommodating skewness and kurtosis. Key predictors such as abdomen circumference and weight exhibited strong associations with body fat, aligning with physiological expectations. Diagnostic checks, including residual and Q-Q plots, confirmed the model's reliability, enabling accurate health risk stratification and fitness program design predictions.

B. Methodological Insights and Limitations

The GAMLSS framework proved instrumental in addressing diverse modeling needs, from parametric distribution selection to flexible smoothing. Model comparison via AIC/GAIC ensured a balance between complexity and interpretability. However, limitations such as sampling biases (e.g., underrepresentation of extreme age groups in the body fat dataset) and multicollinearity among anthropometric predictors warrant caution. Future research could explore longitudinal designs, expanded

datasets, or integration with machine learning techniques to enhance generalizability and predictive power.

C. Practical Implications

These analyses highlight the critical role of tailored statistical models in clinical and public health contexts. The methodologies developed here—reference centile curves for grip strength, BMI distribution parameters, and body fat prediction tools—offer actionable insights for monitoring developmental norms, assessing health risks, and personalizing interventions.

In summary, this study underscores the efficacy of GAMLSS in addressing complex statistical challenges while emphasizing rigorous diagnostics and domain-specific interpretation. The findings contribute to advancing evidence-based practices in growth monitoring, pediatric health, and body composition analysis, with broader applicability to datasets requiring flexible distributional modeling.

VII. References:

[1] P. Bhandari, "Normal Distribution | Examples, Formulas, & Uses," Scribbr, 21 6 2023. [Online]. Available: https://www.scribbr.com/statistics/normal-distribution/#:~:text=Discover%20proofreading%20&%20editing-,What%20are%20the%20properties%20of%20normal%20distributions?,mean%20and%20the%20standard%20deviation.. [Accessed 06 04 2025].

[2] W. Kentin, "Log-Normal Distribution: Definition, Uses, and How To Calculate," Investipedia, 31 10 2021. [Online]. Available: https://www.investopedia.com/terms/l/log-normal-distribution.asp. [Accessed 6 4 2025].

[3] S. S. J. S. A. L. C. J. L. H. a. A. M. W. T. J. Cole, "Age- and size-related reference ranges: A case study of spirometry through childhood and adulthood," Stat. Med., 28 2 2009. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2798072/#:~:text=is%20a%20z%2Dscore%20with,is%20assumed%20to%20be%20absent.. [Accessed 6 4 2025].

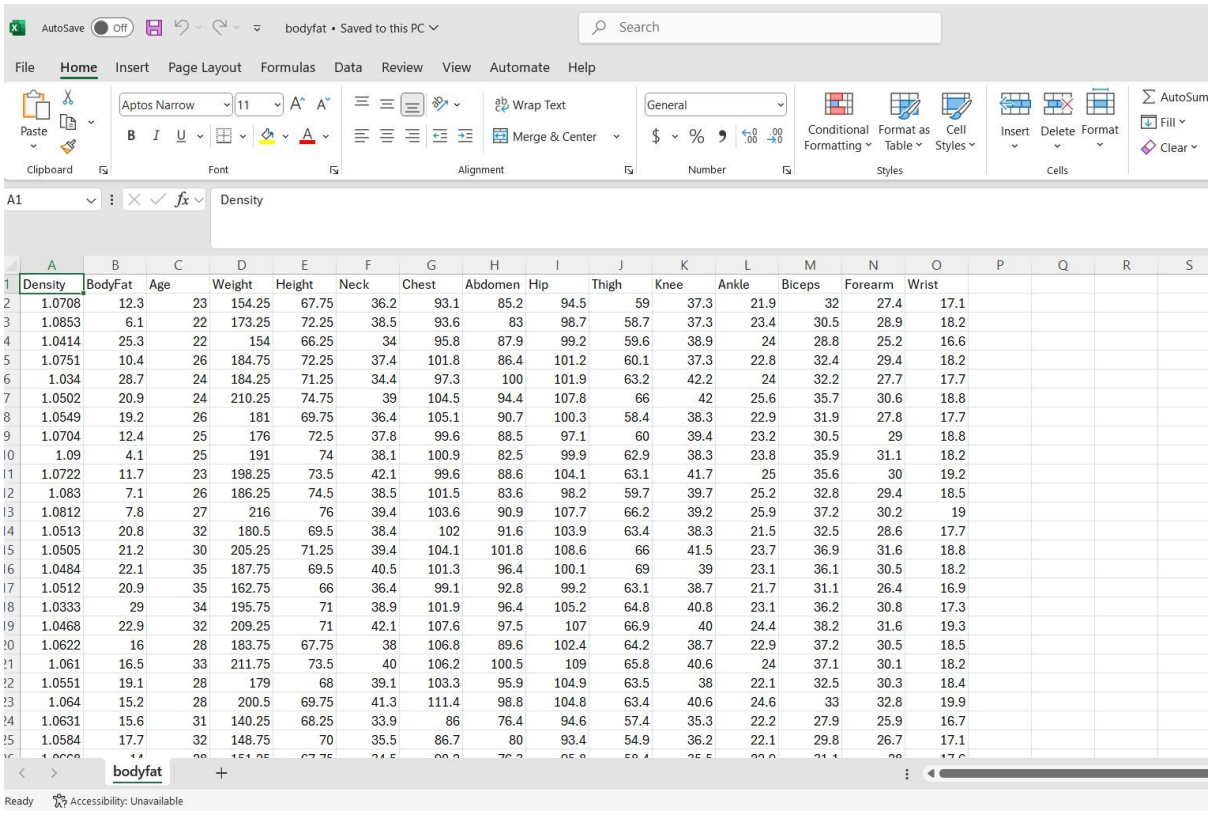
[4] Wikipedia, "Skew normal distribution," Wikipedia, 19 7 2024. [Online]. Available: https://en.wikipedia.org/wiki/Skew_normal_distribution#:~:text=In%20probability%20theory%20and%20statistics%2C%20the%20skew,normal%20distribution%20to%20allow%20for%20non%2Dzero%20skewness.&text=Thus%2C%20the%20skew%20normal%20is%20useful%20for,incidence%. [Accessed 6 4 2025].

[5] R. W. Johnson, "Body Fat Prediction Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset/data

] [Accessed 01 04 2025].

VIII. APPENDIX

1. ***First 20 cases*** **of Body Fat Analysis**



Density	BodyFat	Age	Weight	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
1.0708	12.3	23	154.25	67.75	36.2	93.1	85.2	94.5	59	37.3	21.9	32	27.4	17.1
1.0853	6.1	22	173.25	72.25	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5	28.9	18.2
1.0414	25.3	22	154	66.25	34	95.8	87.9	99.2	59.6	38.9	24	28.8	25.2	16.6
1.0751	10.4	26	184.75	72.25	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
1.034	28.7	24	184.25	71.25	34.4	97.3	100	101.9	63.2	42.2	24	32.2	27.7	17.7
1.0502	20.9	24	210.25	74.75	39	104.5	94.4	107.8	66	42	25.6	35.7	30.6	18.8
1.0549	19.2	26	181	69.75	36.4	105.1	90.7	100.3	58.4	38.3	22.9	31.9	27.8	17.7
1.0704	12.4	25	176	72.5	37.8	99.6	88.5	97.1	60	39.4	23.2	30.5	29	18.8
1.09	4.1	25	191	74	38.1	100.9	82.5	99.9	62.9	38.3	23.8	35.9	31.1	18.2
1.0722	11.7	23	198.25	73.5	42.1	99.6	88.6	104.1	63.1	41.7	25	35.6	30	19.2
1.083	7.1	26	186.25	74.5	38.5	101.5	83.6	98.2	59.7	39.7	25.2	32.8	29.4	18.5
1.0812	7.8	27	216	76	39.4	103.6	90.9	107.7	66.2	39.2	25.9	37.2	30.2	19
1.0513	20.8	32	180.5	69.5	38.4	102	91.6	103.9	63.4	38.3	21.5	32.5	28.6	17.7
1.0505	21.2	30	205.25	71.25	39.4	104.1	101.8	108.6	66	41.5	23.7	36.9	31.6	18.8
1.0484	22.1	35	187.75	69.5	40.5	101.3	96.4	100.1	69	39	23.1	36.1	30.5	18.2
1.0512	20.9	35	162.75	66	36.4	99.1	92.8	99.2	63.1	38.7	21.7	31.1	26.4	16.9
1.0333	29	34	195.75	71	38.9	101.9	96.4	105.2	64.8	40.8	23.1	36.2	30.8	17.3
1.0468	22.9	32	209.25	71	42.1	107.6	97.5	107	66.9	40	24.4	38.2	31.6	19.3
1.0622	16	28	183.75	67.75	38	106.8	89.6	102.4	64.2	38.7	22.9	37.2	30.5	18.5
1.061	16.5	33	211.75	73.5	40	106.2	100.5	109	65.8	40.6	24	37.1	30.1	18.2
1.0551	19.1	28	179	68	39.1	103.3	95.9	104.9	63.5	38	22.1	32.5	30.3	18.4
1.064	15.2	28	200.5	69.75	41.3	111.4	98.8	104.8	63.4	40.6	24.6	33	32.8	19.9
1.0631	15.6	31	140.25	68.25	33.9	86	76.4	94.6	57.4	35.3	22.2	27.9	25.9	16.7
1.0584	17.7	32	148.75	70	35.5	86.7	80	93.4	54.9	36.2	22.1	29.8	26.7	17.1

2. **Model Selection:** The response variable (BodyFat_transformed) was modeled using Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to accommodate non-normal distributions. First, we needed to identify an appropriate distribution for the response variable(body fat).To determine the best distribution for the response variable, I first fitted basic models using common distributions as shown in the figure below

    - Normal Distribution (NO): Traditional Gaussian model assuming symmetry around the mean
    - Log-Normal Distribution (LOGNO): Appropriate for positively skewed data

```
90
91  # Fit models using the transformed variable
92  mNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest + Abdomen + Hip +
93                  Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
94                data = clean_data, family = NO)
95
96  mLOGNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
97                     Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
98                   data = clean_data, family = LOGNO)
99
100 # Compare models using GAIC
101 GAIC(mNO, mLOGNO)
```

The AIC comparison between these initial models showed that the Normal distribution provided a better baseline fit with an AIC of 1466.502, compared to 1651.767 for the Log-Normal distribution. This initial result guided further distribution exploration.

```
> # Fit models using the transformed variable
> mNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest + Abdomen + Hip +
+                 Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
+               data = clean_data, family = NO)
GAMLSS-RS iteration 1: Global Deviance = 1436.502
GAMLSS-RS iteration 2: Global Deviance = 1436.502
> mLOGNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
+                    Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
+                  data = clean_data, family = LOGNO)
GAMLSS-RS iteration 1: Global Deviance = 2056.639
GAMLSS-RS iteration 2: Global Deviance = 2056.639
> # Compare models using GAIC
> GAIC(mNO, mLOGNO)
        df      AIC
mNO     15 1466.502
mLOGNO  15 2086.639
```

By looking at the figure, it was clear that the mNO(Normal Distribution was best suited for the baseline model as it has the lowest AIC, 1466.502. To validate the model assumptions ans to assess the fit diagnostics where performed like:

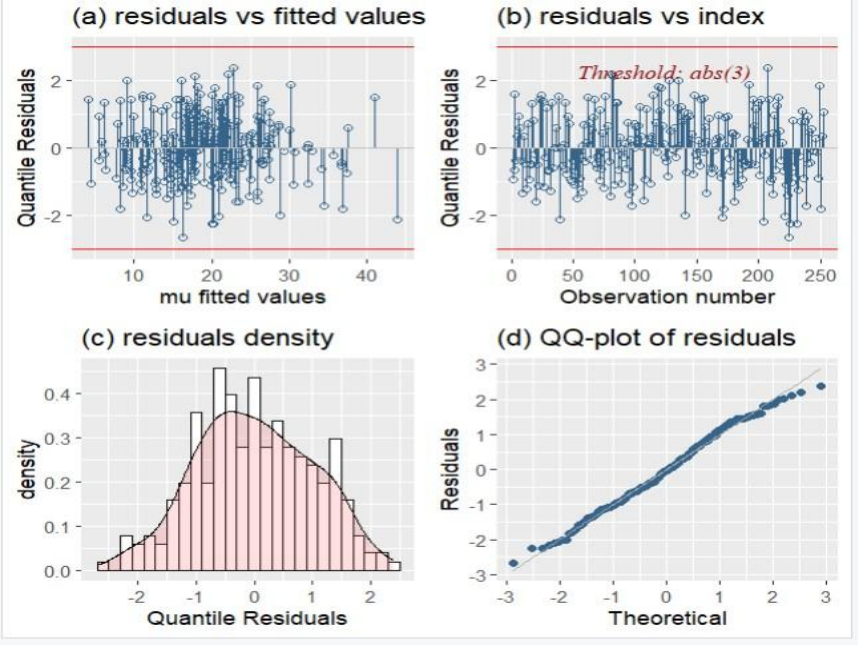
Residual Analysis:

(a) Residuals vs. Fitted Values(Plot a): The residuals are mostly scattered randomly around zero, suggesting the model captures the data's trend well. However, a slight curve in the pattern hints at minor non-linearity unaccounted for.

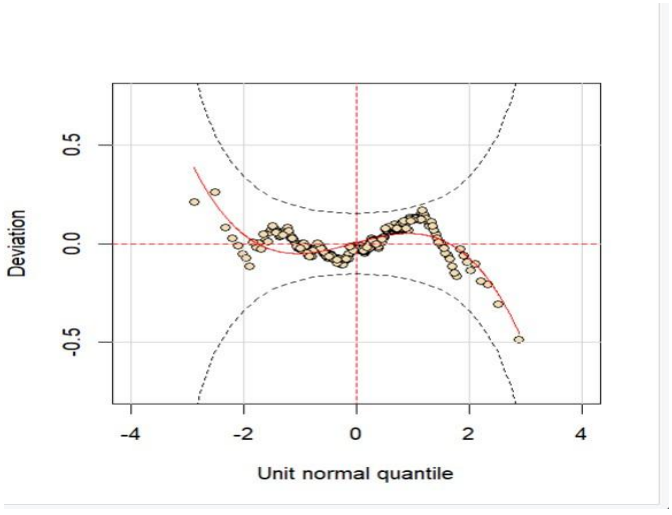
(b) Residuals vs Observation Order (Plot b): Residuals show no clear upward/downward trend over time, indicating no major autocorrelation. A few points exceed the ±3 threshold, flagging potential outliers.

(c)Density Plot (Plot c): The residuals roughly follow a bell-shaped curve, supporting normality assumptions, though slight tails suggest occasional extreme values

(d)QQ-Plot (Plot d): The residuals mostly align with the theoretical normal line, but deviations at the ends (especially in the second plot's tails) imply heavier tails or outliers.



Worm Plot: The worm plot shows detrended residuals to check normality more clearly. Most points cluster tightly around the zero line within the confidence bands, indicating the residuals align well with a normal distribution overall. However, slight upward curves at the extreme ends suggest heavier tails—meaning a few residuals are more extreme than expected under normality, matching the earlier QQ-plot findings. A minor dip near the middle hints at subtle asymmetry, but this is not severe. No glaring patterns or outliers outside the bands are visible, reinforcing that the model's assumptions are reasonably met.



For a more comprehensive assessment, I employed the chooseDist() function in the GAMLSS package, which systematically evaluates multiple distribution families compatible with the data type The function was configured to focus on distributions suitable for positive continuous response variables (type = "realplus"), appropriate for body fat percentage.The parallel processing via the snow package speeds up

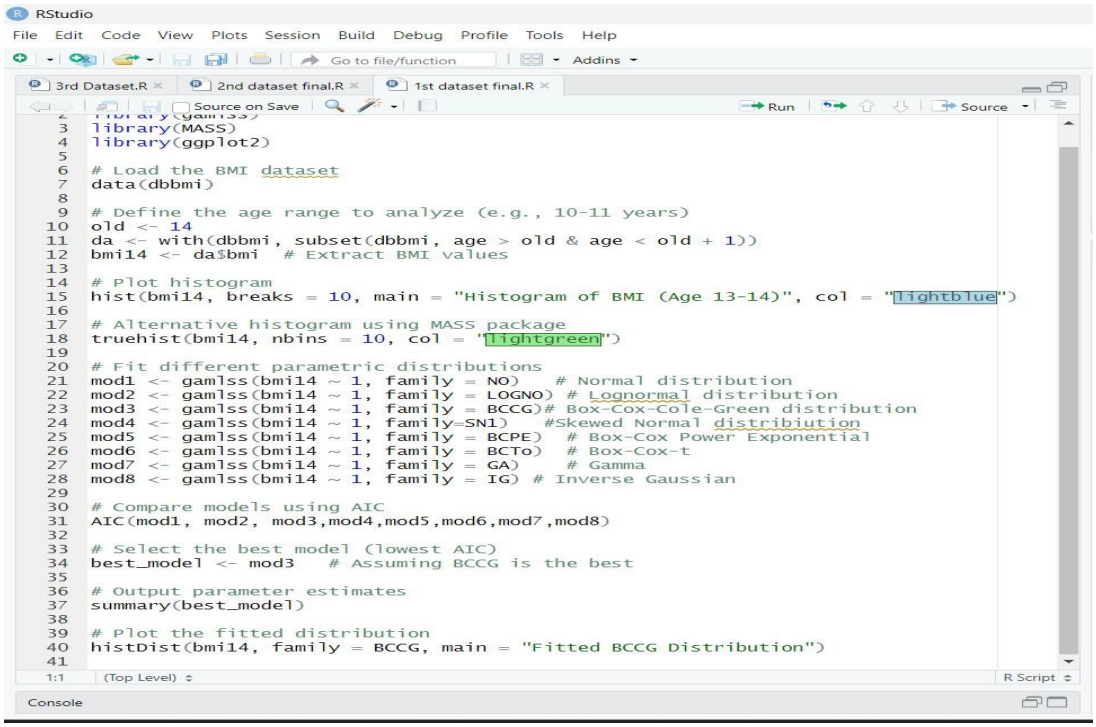
computations by distributing tasks across multiple CPU cores and the ncpuss = 9 specifies the number of CPU cores to use for parallel processing (here, 9 cores).

```
> # Choose the best distribution
> D1 <- chooseDist(mNO, type = "realplus", parallel = "snow", ncpuss = 9)
minimum GAIC(k= 2 ) family: BCT
minimum GAIC(k= 3.84 ) family: BCT
minimum GAIC(k= 5.53 ) family: BCT
There were 15 warnings (use warnings() to see them)
> print(D1)
                 2     3.84     5.53
EXP       1981.658 2007.418 2031.078
GA        1743.318 1770.918 1796.268
IG              NA       NA       NA
LOGNO     2086.639 2114.239 2139.589
LOGNO2    2086.639 2114.239 2139.589
WEI       1624.850 1652.450 1677.800
WEI2      1717.142 1744.742 1770.092
WEI3      1624.850 1652.450 1677.800
IGAMMA          NA       NA       NA
PARETO2   2006.847 2034.447 2059.797
PARETO2o  2006.780 2034.380 2059.730
GP        2006.847 2034.447 2059.797
BCCG      1539.427 1568.867 1595.907
BCCGo     1561.221 1590.661 1617.701
exGAUS          NA       NA       NA
GG        1582.579 1612.019 1639.059
GIG             NA       NA       NA
LNO       2086.639 2114.239 2139.589
BCTo      1563.221 1594.501 1623.231
BCT       1532.204 1563.484 1592.214
BCPEo     1561.601 1592.881 1621.611
BCPE      1537.595 1568.875 1597.605
GB2       1588.786 1620.066 1648.796
> # Fit multiple GAMLSS models
```

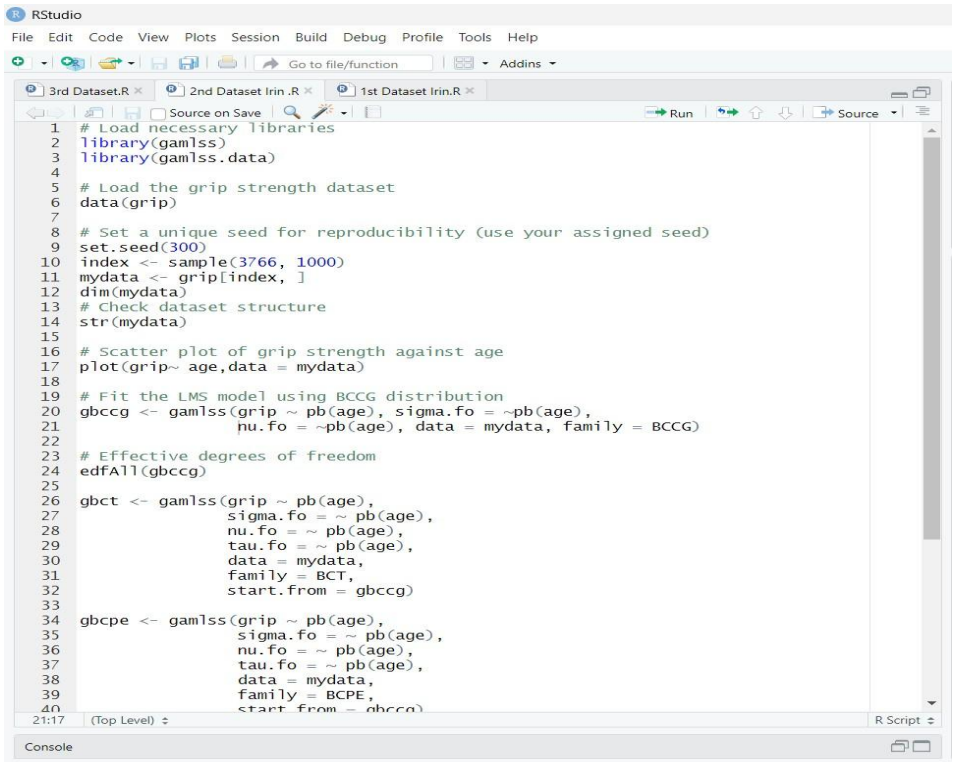
Based on the chooseDist() results, I selected several candidate distributions for further evaluation:

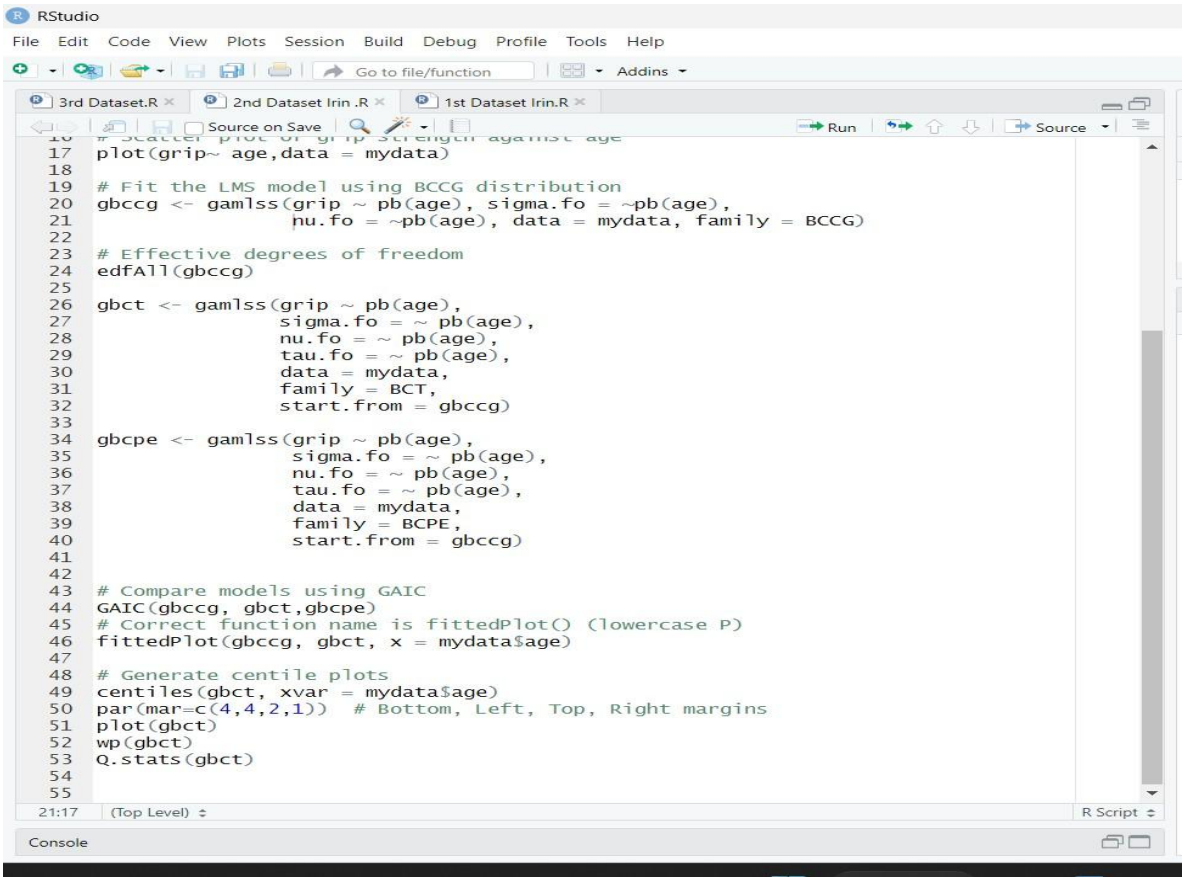
- BCT (Box-Cox t-distribution): A flexible four-parameter distribution that extends the Box-Cox transformation with a t-distribution base to handle heavy tails
- GB2 (Generalized Beta distribution of the second kind): A flexible fourparameter distribution often used for modeling income and wealth distributions
- BCPE (Box-Cox Power Exponential): A four-parameter distribution that extends the BCCG by adding a parameter for kurtosis
- LOGNO (Log-Normal): A two-parameter distribution for positive skewed data
- BCPEo (Box-Cox Power Exponential original): A variant of BCPE
- GG (Generalized Gamma): A three-parameter distribution offering flexibility for positive data
- BCCGo (Box-Cox Cole and Green original): A variant of the BCCG distribution
- BCCG (Box-Cox Cole and Green): A three-parameter distribution commonly used in growth reference curves

3. First Dataset Code:

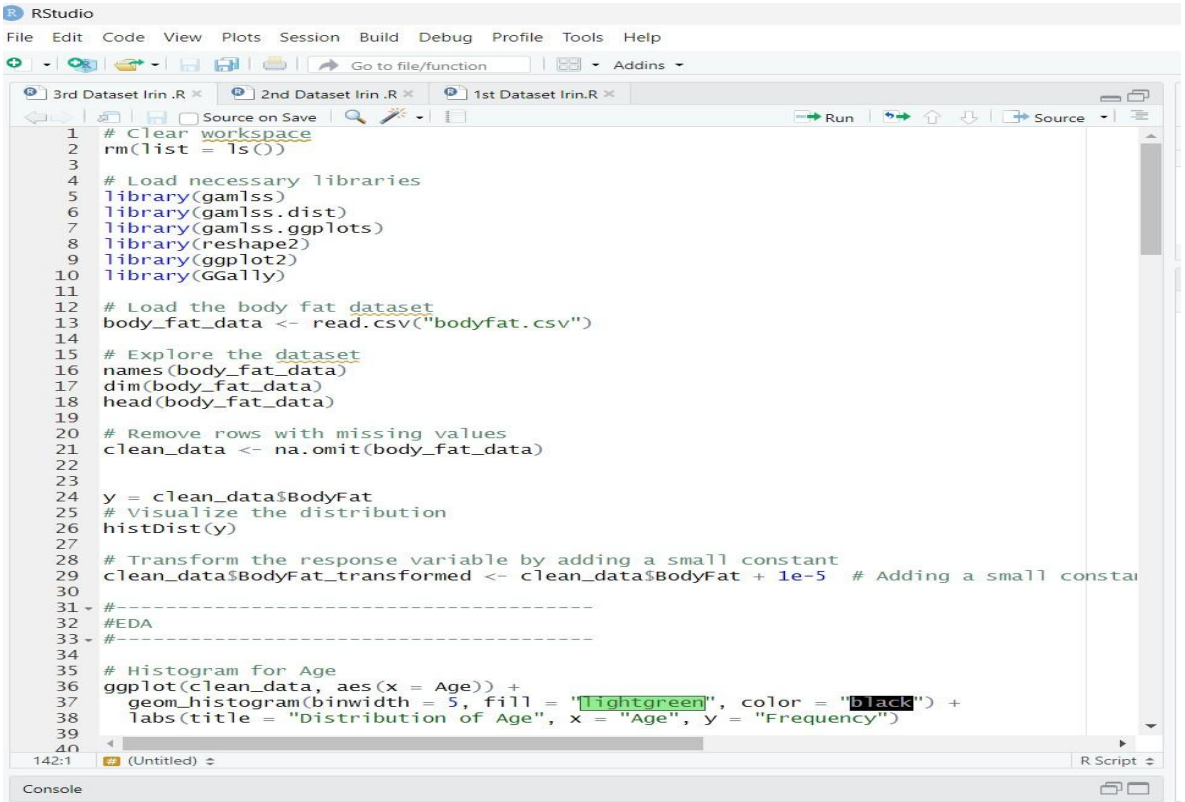


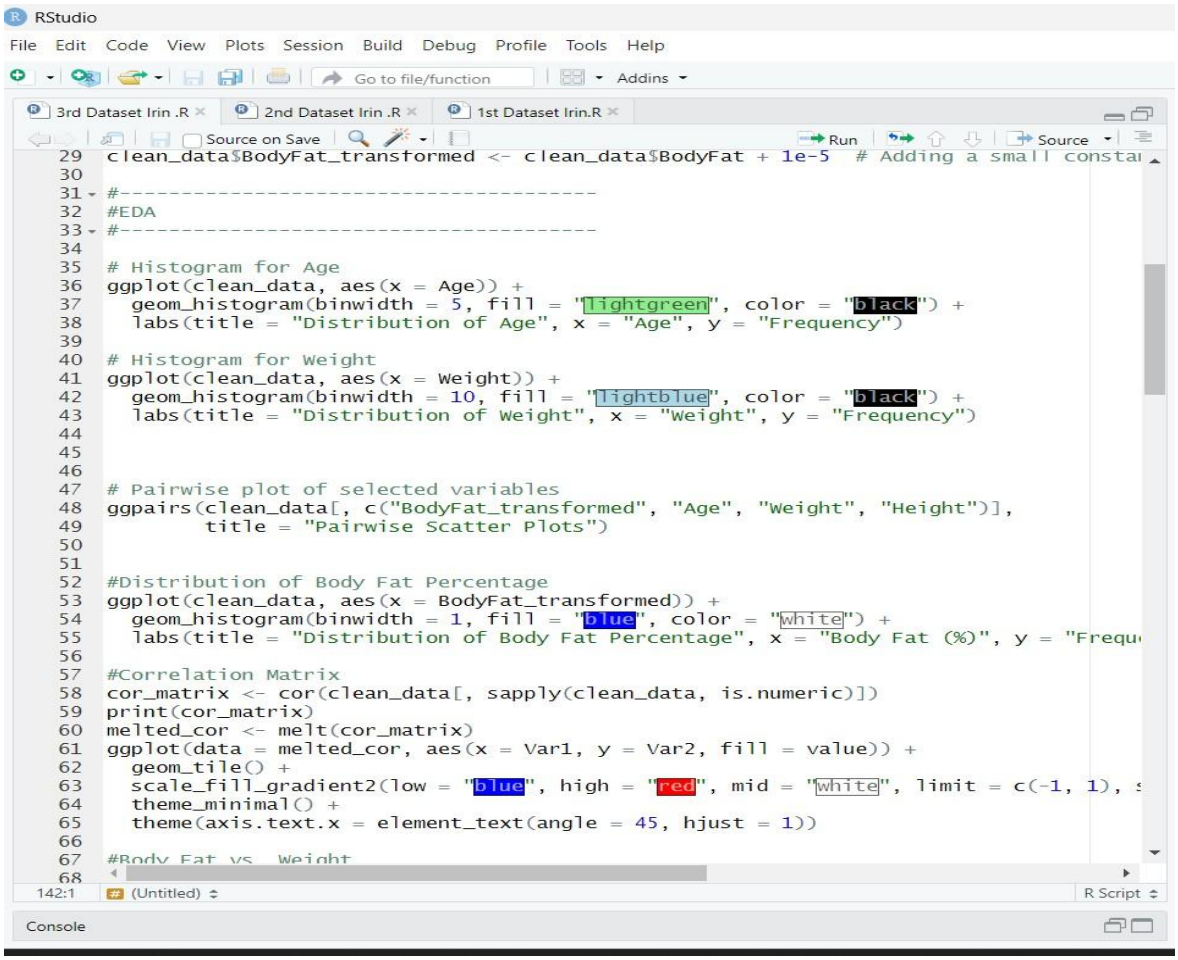
4. 2nd Dataset code



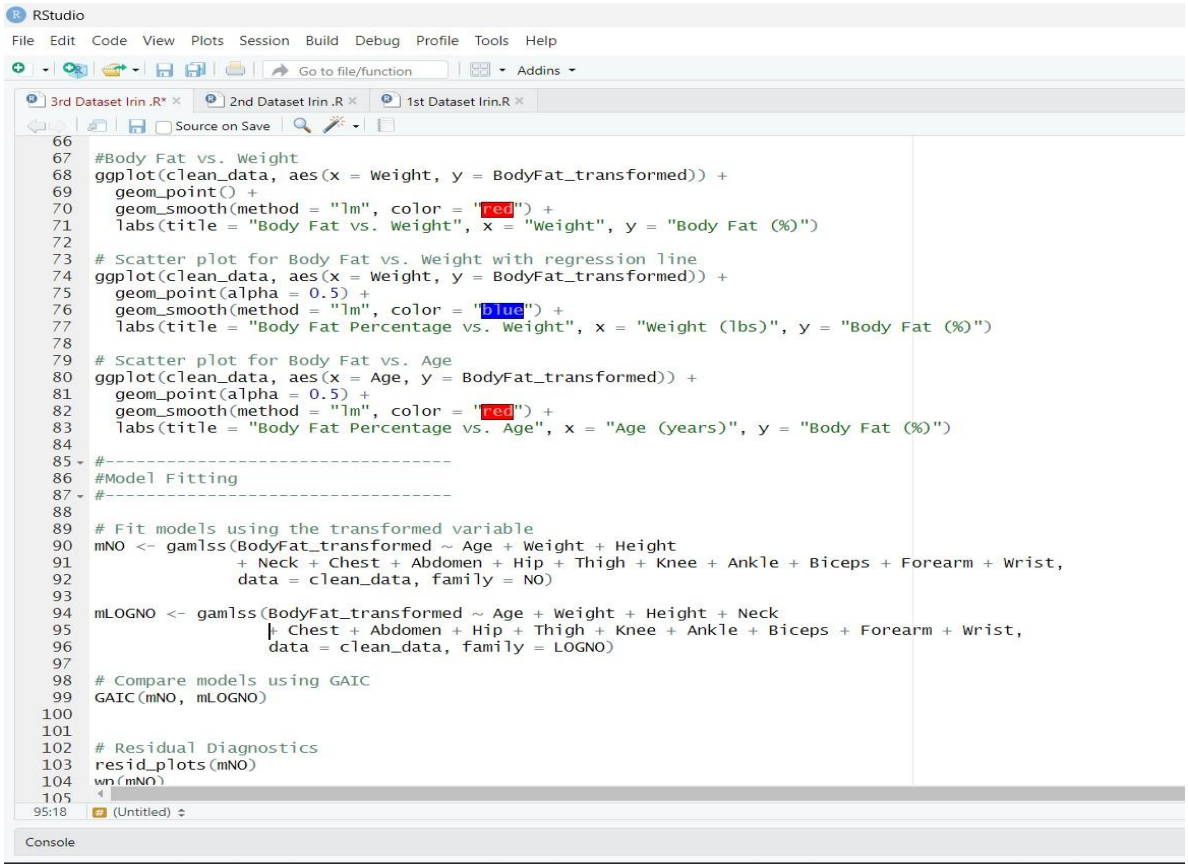


5. 3rd Dataset Code:

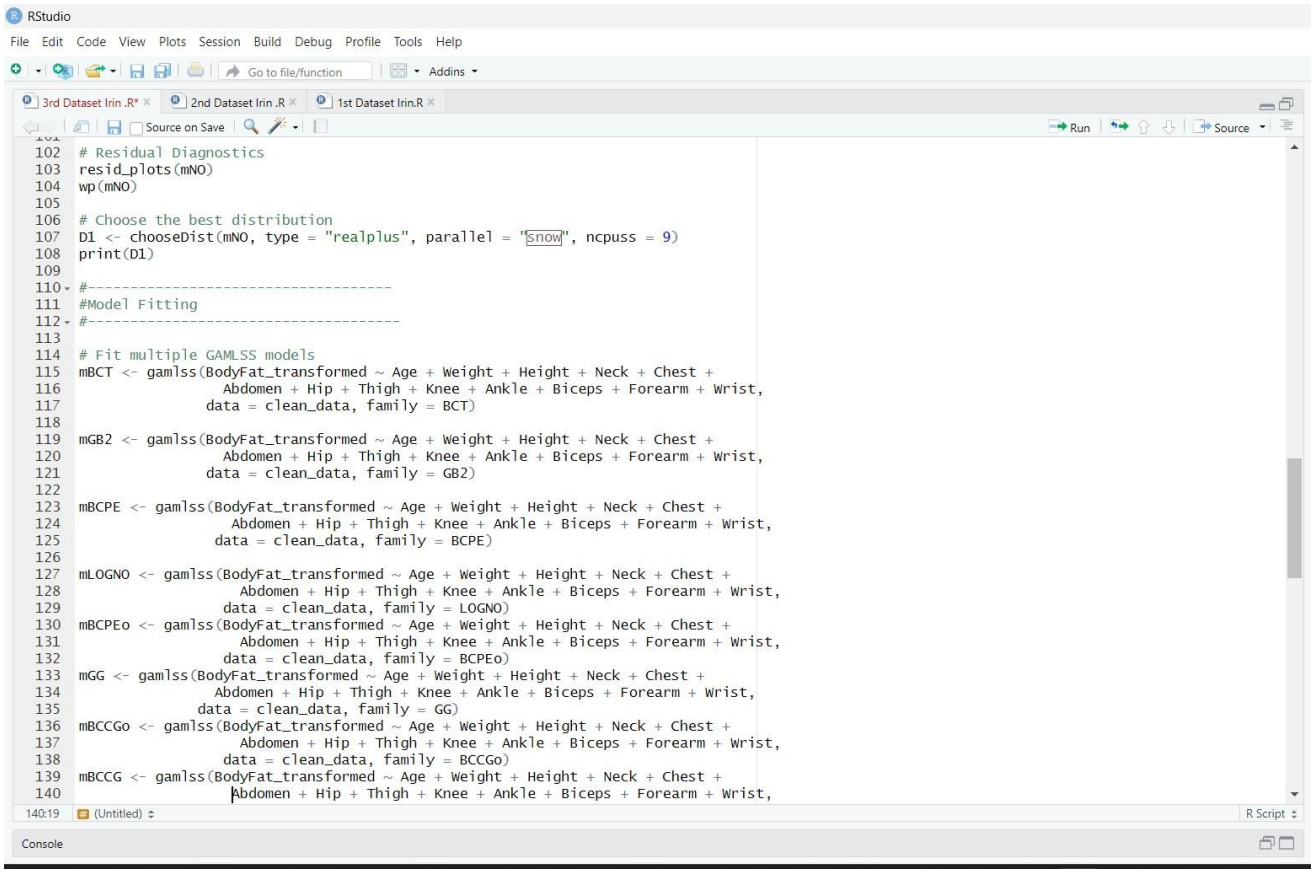




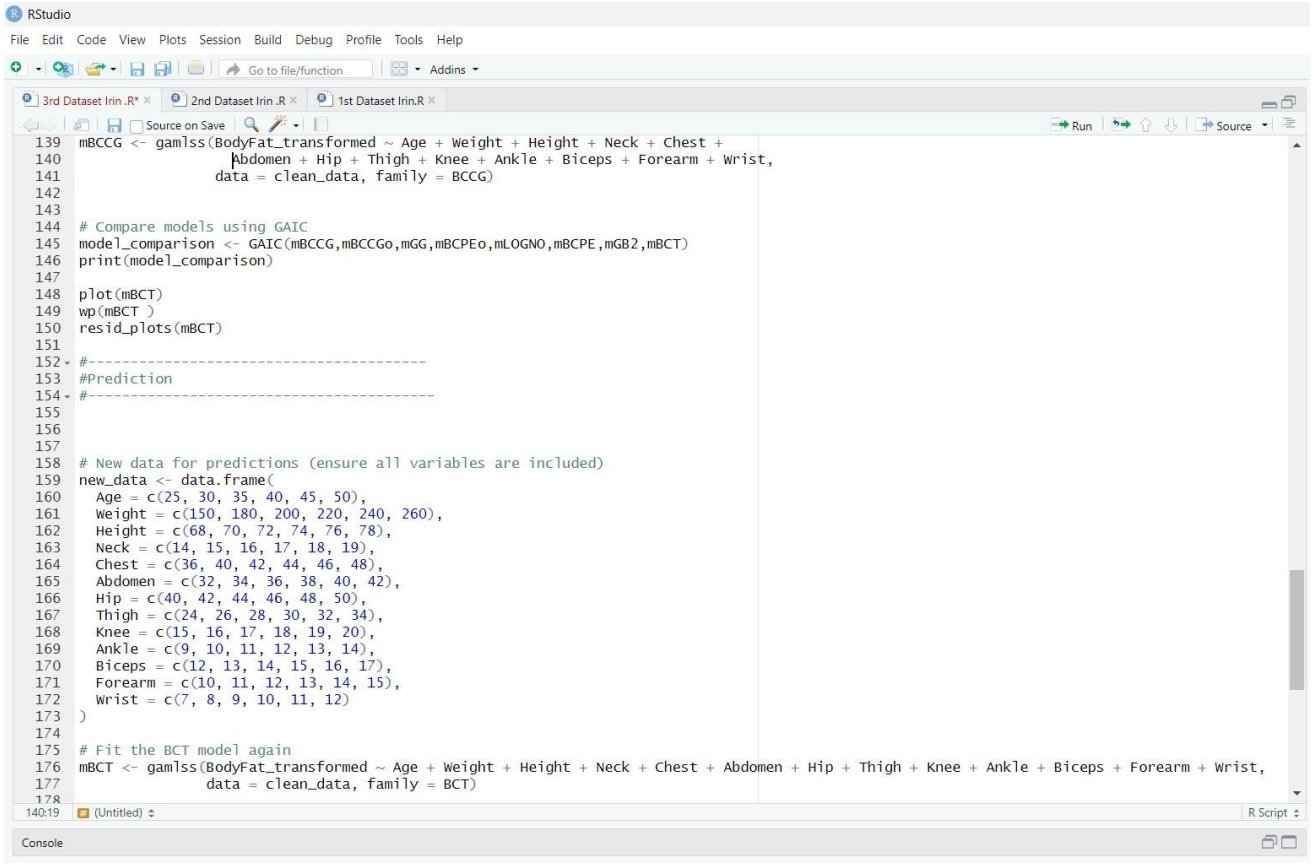
```
clean_data$BodyFat_transformed <- clean_data$BodyFat + 1e-5  # Adding a small constar

#----------------------------------------
#EDA
#----------------------------------------

# Histogram for Age
ggplot(clean_data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency")

# Histogram for Weight
ggplot(clean_data, aes(x = Weight)) +
  geom_histogram(binwidth = 10, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Weight", x = "Weight", y = "Frequency")



# Pairwise plot of selected variables
ggpairs(clean_data[, c("BodyFat_transformed", "Age", "Weight", "Height")],
        title = "Pairwise Scatter Plots")


#Distribution of Body Fat Percentage
ggplot(clean_data, aes(x = BodyFat_transformed)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "white") +
  labs(title = "Distribution of Body Fat Percentage", x = "Body Fat (%)", y = "Frequ

#Correlation Matrix
cor_matrix <- cor(clean_data[, sapply(clean_data, is.numeric)])
print(cor_matrix)
melted_cor <- melt(cor_matrix)
ggplot(data = melted_cor, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", limit = c(-1, 1), s
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Body Fat vs. Weight
```



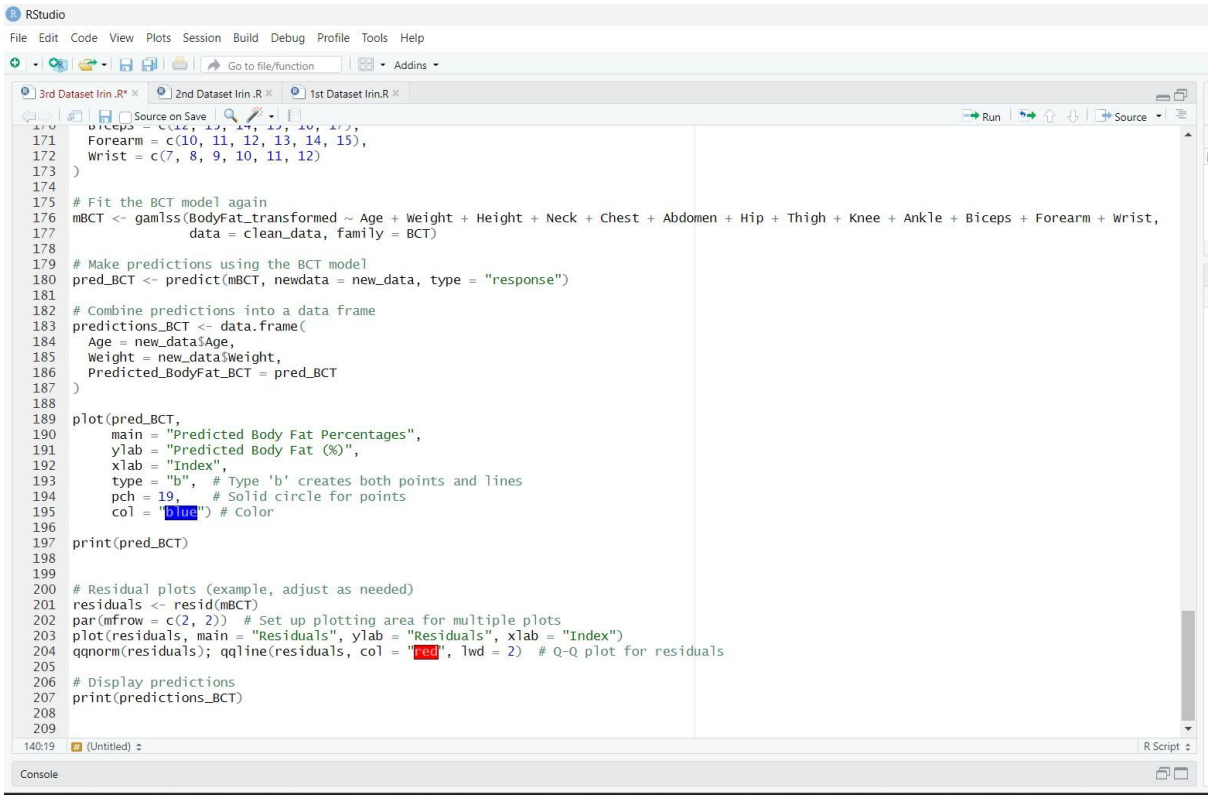
```
#Body Fat vs. Weight
ggplot(clean_data, aes(x = Weight, y = BodyFat_transformed)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Body Fat vs. Weight", x = "Weight", y = "Body Fat (%)")

# Scatter plot for Body Fat vs. Weight with regression line
ggplot(clean_data, aes(x = Weight, y = BodyFat_transformed)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Body Fat Percentage vs. Weight", x = "Weight (lbs)", y = "Body Fat (%)")

# Scatter plot for Body Fat vs. Age
ggplot(clean_data, aes(x = Age, y = BodyFat_transformed)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Body Fat Percentage vs. Age", x = "Age (years)", y = "Body Fat (%)")

#-----------------------------------
#Model Fitting
#-----------------------------------

# Fit models using the transformed variable
mNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height
              + Neck + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
              data = clean_data, family = NO)

mLOGNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck
                 + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = LOGNO)

# Compare models using GAIC
GAIC(mNO, mLOGNO)


# Residual Diagnostics
resid_plots(mNO)
wp(mNO)
```



```
# Residual Diagnostics
resid_plots(mNO)
wp(mNO)

# Choose the best distribution
D1 <- chooseDist(mNO, type = "realplus", parallel = "snow", ncpuss = 9)
print(D1)

#--------------------------------------
#Model Fitting
#--------------------------------------

# Fit multiple GAMLSS models
mBCT <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                 Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
               data = clean_data, family = BCT)

mGB2 <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                 Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
               data = clean_data, family = GB2)

mBCPE <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                  Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                data = clean_data, family = BCPE)

mLOGNO <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = LOGNO)
mBCPEo <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = BCPEo)
mGG <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
              data = clean_data, family = GG)
mBCCGo <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                   Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                 data = clean_data, family = BCCGo)
mBCCG <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                  Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
```



```
mBCCG <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest +
                  Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
                data = clean_data, family = BCCG)


# Compare models using GAIC
model_comparison <- GAIC(mBCCG,mBCCGo,mGG,mBCPEo,mLOGNO,mBCPE,mGB2,mBCT)
print(model_comparison)

plot(mBCT)
wp(mBCT )
resid_plots(mBCT)

#----------------------------------------
#Prediction
#----------------------------------------



# New data for predictions (ensure all variables are included)
new_data <- data.frame(
  Age = c(25, 30, 35, 40, 45, 50),
  Weight = c(150, 180, 200, 220, 240, 260),
  Height = c(68, 70, 72, 74, 76, 78),
  Neck = c(14, 15, 16, 17, 18, 19),
  Chest = c(36, 40, 42, 44, 46, 48),
  Abdomen = c(32, 34, 36, 38, 40, 42),
  Hip = c(40, 42, 44, 46, 48, 50),
  Thigh = c(24, 26, 28, 30, 32, 34),
  Knee = c(15, 16, 17, 18, 19, 20),
  Ankle = c(9, 10, 11, 12, 13, 14),
  Biceps = c(12, 13, 14, 15, 16, 17),
  Forearm = c(10, 11, 12, 13, 14, 15),
  Wrist = c(7, 8, 9, 10, 11, 12)
)

# Fit the BCT model again
mBCT <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
               data = clean_data, family = BCT)
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
3rd Dataset Irin .R*
2nd Dataset Irin .R
1st Dataset Irin.R
Source on Save
Run
Source
Forearm = c(10, 11, 12, 13, 14, 15),
Wrist = c(7, 8, 9, 10, 11, 12)
)
# Fit the BCT model again
mBCT <- gamlss(BodyFat_transformed ~ Age + Weight + Height + Neck + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Biceps + Forearm + Wrist,
data = clean_data, family = BCT)
# Make predictions using the BCT model
pred_BCT <- predict(mBCT, newdata = new_data, type = "response")
# Combine predictions into a data frame
predictions_BCT <- data.frame(
Age = new_data$Age,
Weight = new_data$Weight,
Predicted_BodyFat_BCT = pred_BCT
)
plot(pred_BCT,
main = "Predicted Body Fat Percentages",
ylab = "Predicted Body Fat (%)",
xlab = "Index",
type = "b", # Type 'b' creates both points and lines
pch = 19, # Solid circle for points
col = "blue") # Color
print(pred_BCT)
# Residual plots (example, adjust as needed)
residuals <- resid(mBCT)
par(mfrow = c(2, 2)) # Set up plotting area for multiple plots
plot(residuals, main = "Residuals", ylab = "Residuals", xlab = "Index")
qqnorm(residuals); qqline(residuals, col = "red", lwd = 2) # Q-Q plot for residuals
# Display predictions
print(predictions_BCT)
140:19
(Untitled)
R Script
Console