

Consulting Project Report

Xiaohan Shi, Zihao Zhang, Suheng Yao

2024-11-16

Introduction

Our group's project is on the positional behaviors of orangutans. Orangutans are primates that live in the rain forests of Southeast Asia and are known for their red fur and high intelligence. They are also the largest arboreal mammals living in the world. They are found mainly in Borneo and Sumatra and are listed as an endangered species due to habitat loss and threats from illegal hunting. The main objective of this consulting project is to help the client analyze the positional behaviors of orangutans over developmental stages to understand how these behaviors evolve with age. The client used a measurement called Shannor Weaver Index to calculate the diversity score based on the positional behavior and try to find its relationship with age. How Shannor Weaver Index is calculated will be covered more in the method part of the report.

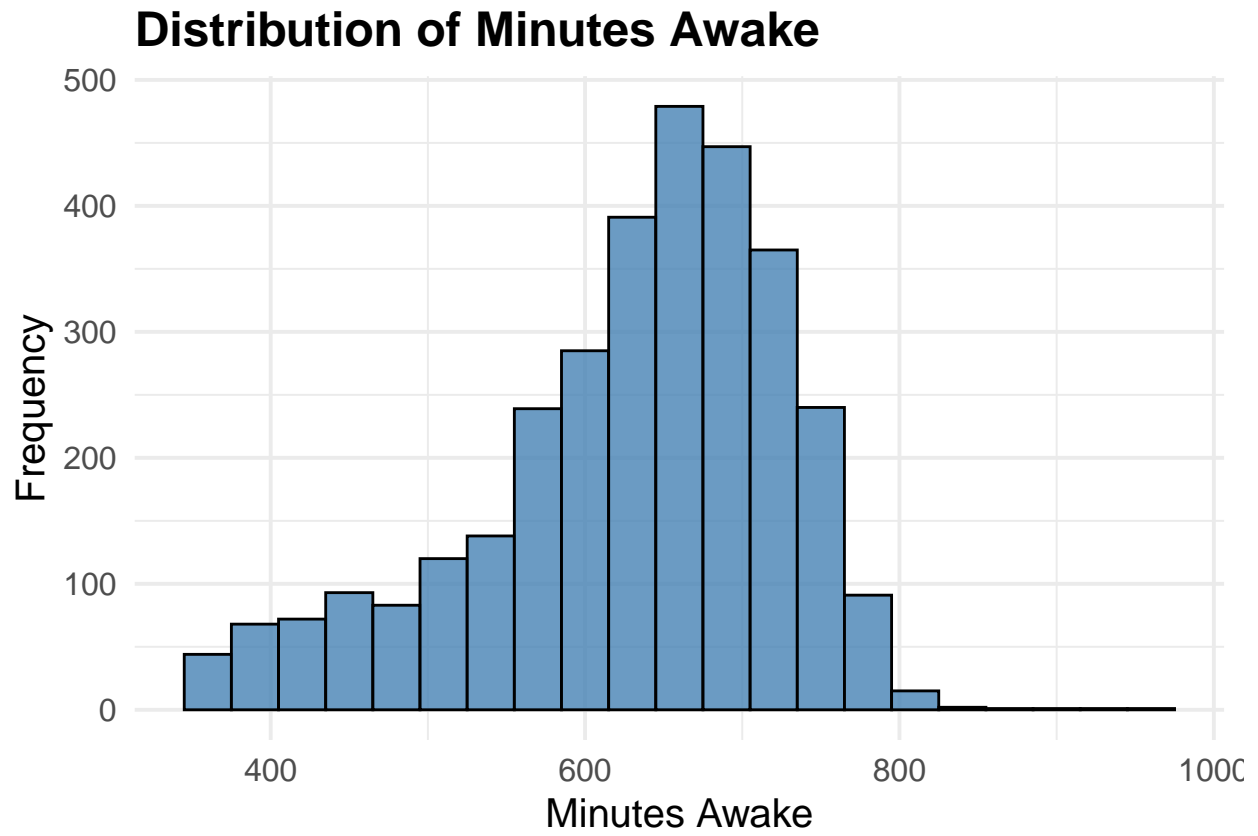
The original data that the client gave us contains 237 individual orangutans that had been followed for 30 years at different periods. Each orangutan was observed for a period of time at different times, and the timing of the behavior at each location was recorded. In total, there are 77 unique positional behaviors recorded for each orangutan. These behaviors are created by the combination of three behavior groups: activity type, body position, and tree position.

In this report, we will mainly talk about the basic EDA to analyze the data, the methods used to assess the relationship, including the models used, and finally the conclusion based on the results.

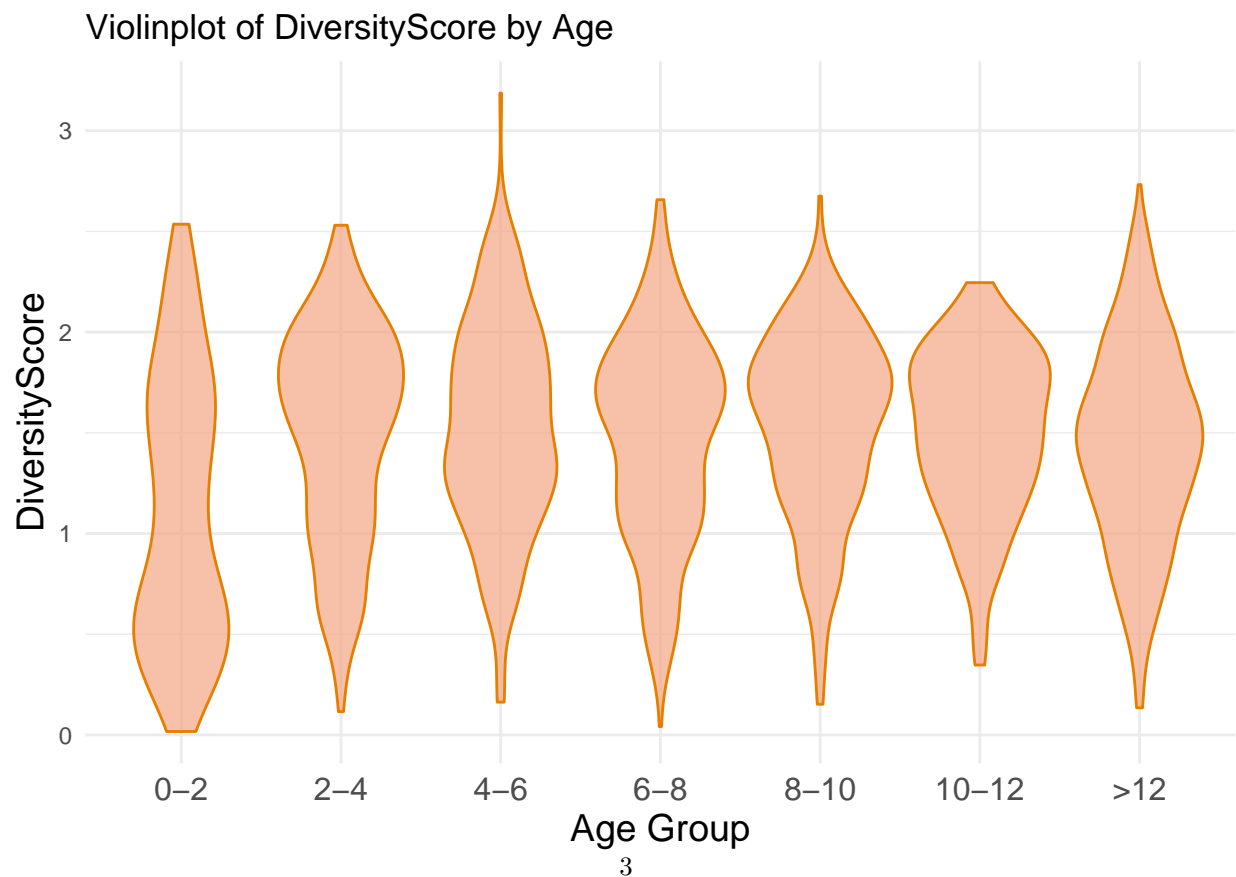
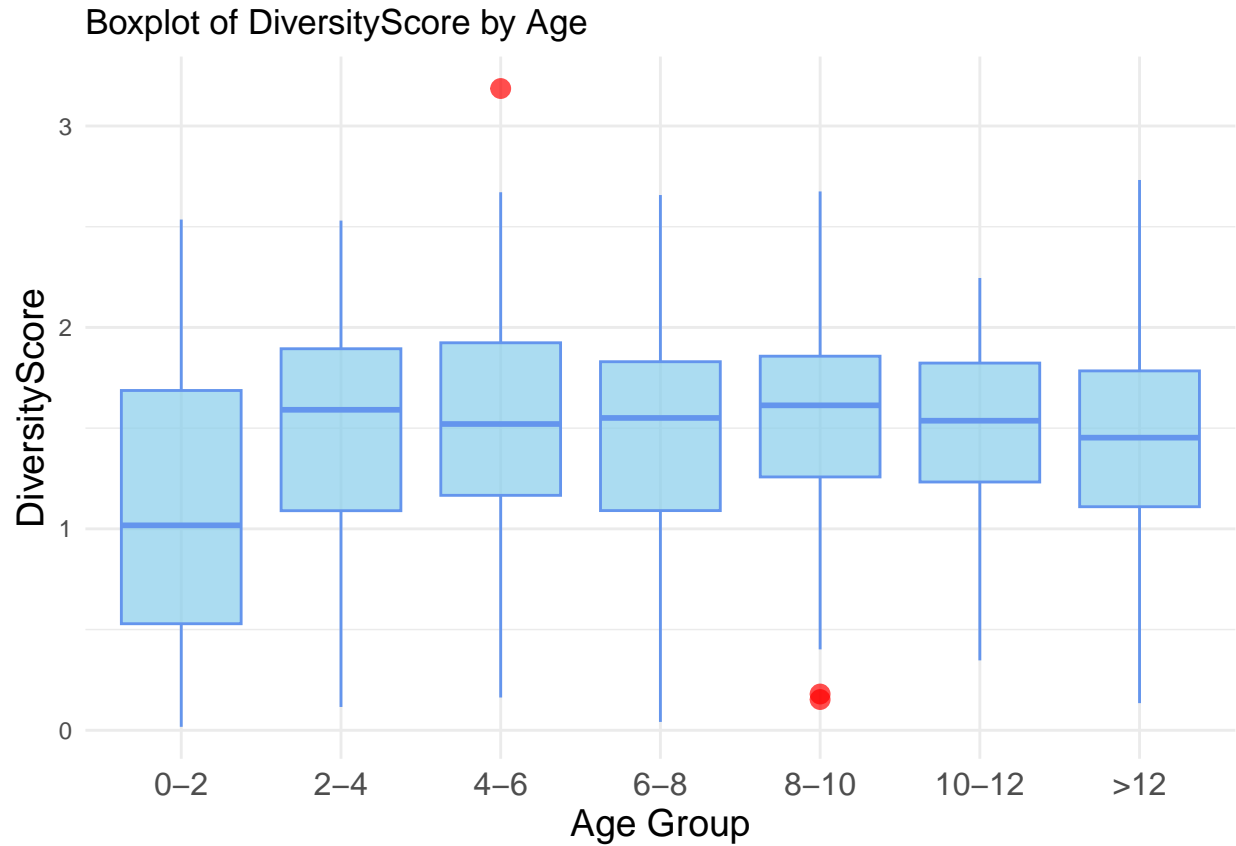
EDA

The client divided all the ages into 7 groups. In the original dataset, if the age is between 0 to 2, then those orangutans are classified as age level 1; if the age is between 2 to 4, then those orangutans are classified as age level 2; if the age is between 4 to 6, then those orangutans are classified as age level 3; if the age is between 6 to 8, then those orangutans are classified as age level 4; if the age is between 8 to 10, then those orangutans are classified as age level 5; if the age is between 10 to 12, then those orangutans are classified as age level 5. For those with ages greater than 12, they go to the age 7 group. Also, all the minutes awake are greater than 360 minutes.

Part 1: The Distribution of Minutes Awake



Part 2: Visualization of Diversity Score



The boxplot shows that the median of Diversity Score increases slightly with Age, especially from Age group 1 to Age group 2. Then the trend leveled off and the median for all age groups remained at about 1.5. There are two outliers which appeared in Age 3 and Age 5 groups.

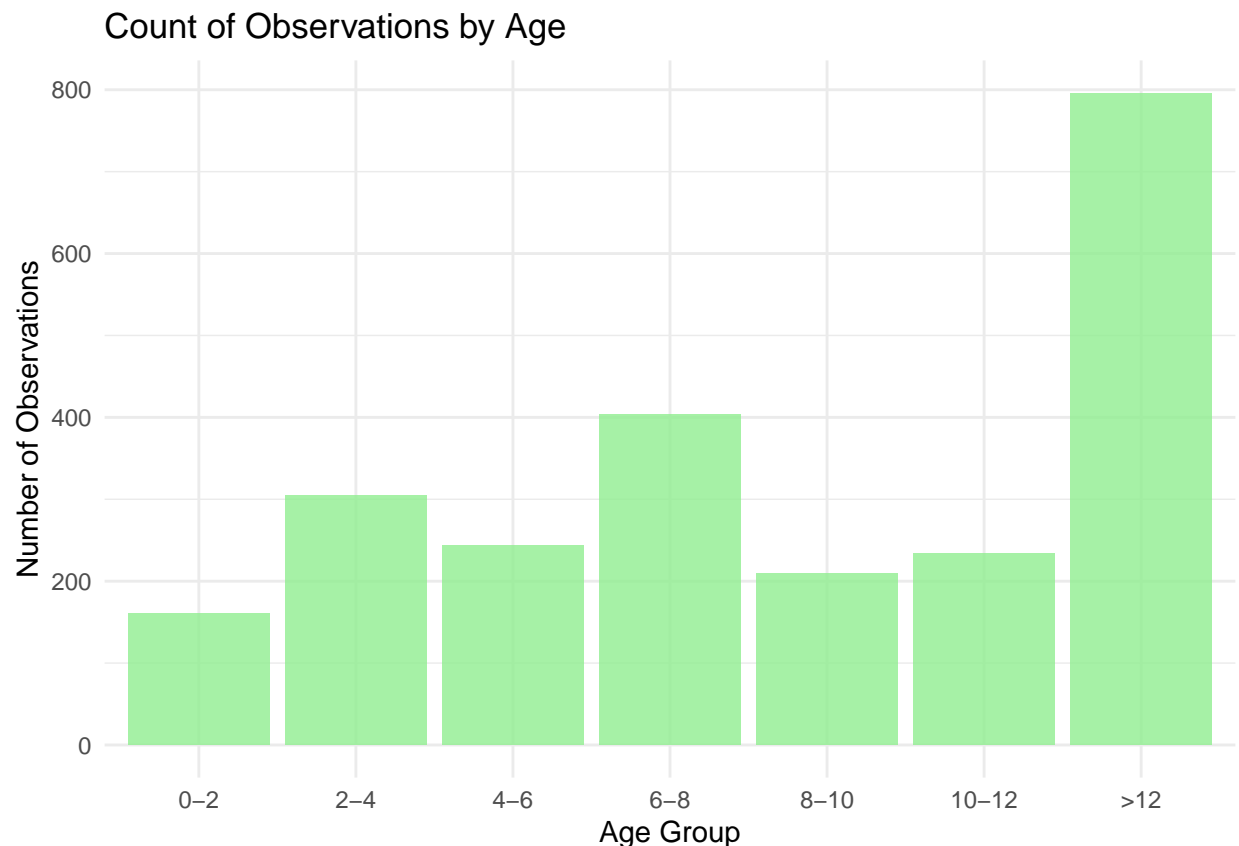
In the violinplot, the similar width of the violin illustrates that the sample size is relatively balanced across age groups due to, there is no particularly small or large sample group.

It also shows that the Diversity Score for all age groups was roughly distributed between 0 and 3. The distribution of Diversity score in group 2-7 are similar that most observations are concentrated between 1 and 2. The data in Age 1 group are dispersed to a large extent, which is from 0 to 2.5.

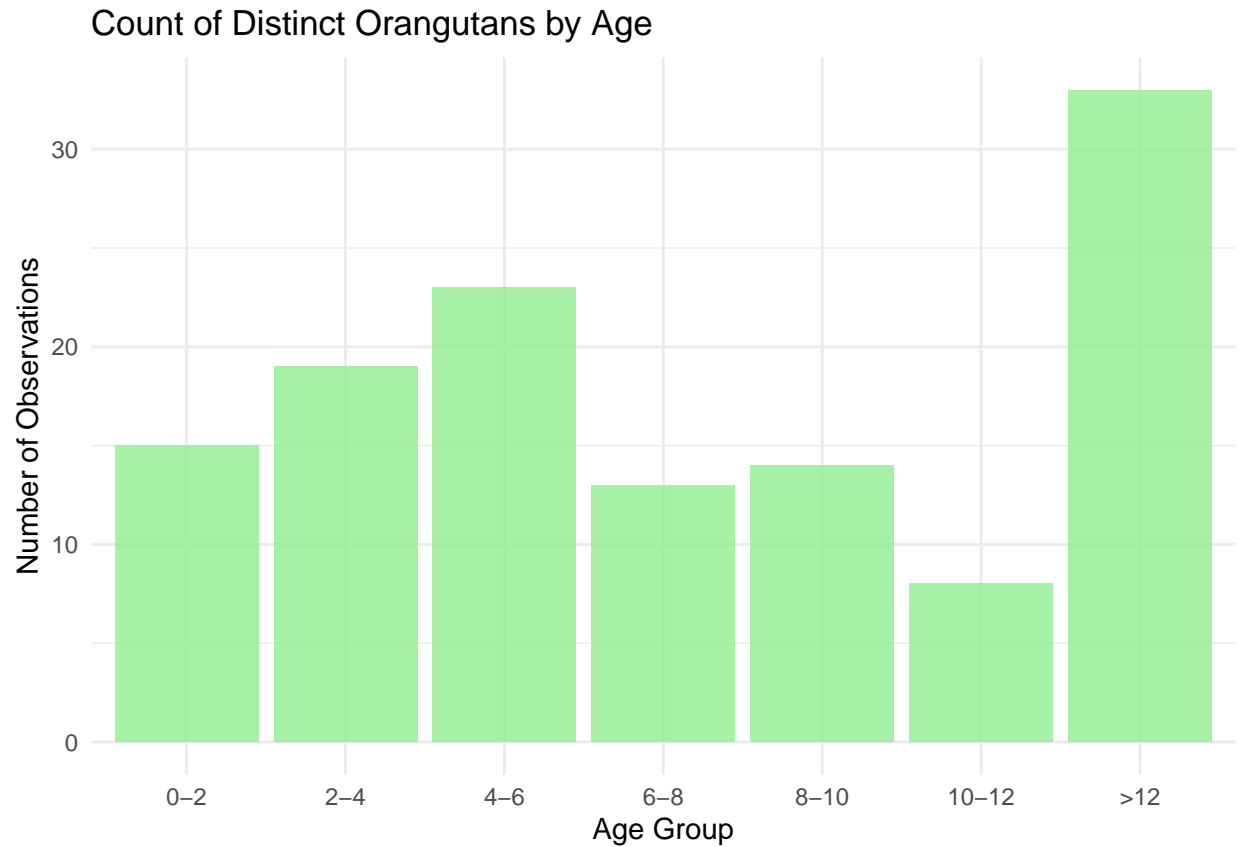
In addition, as the age get older, the variability of the data seems to decrease and the distribution becomes more concentrated.

Part 3: Diversity Score Statistics

```
## # A tibble: 7 x 6
##   Age    Min    Max Median   Mean    SD
##   <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1 0.0170 2.54  1.02  1.11 0.671
## 2     2 0.116 2.53  1.59  1.50 0.523
## 3     3 0.163 3.19  1.52  1.53 0.521
## 4     4 0.0414 2.66  1.55  1.47 0.507
## 5     5 0.153 2.68  1.61  1.54 0.462
## 6     6 0.347 2.25  1.54  1.50 0.411
## 7     7 0.135 2.73  1.45  1.44 0.499
```



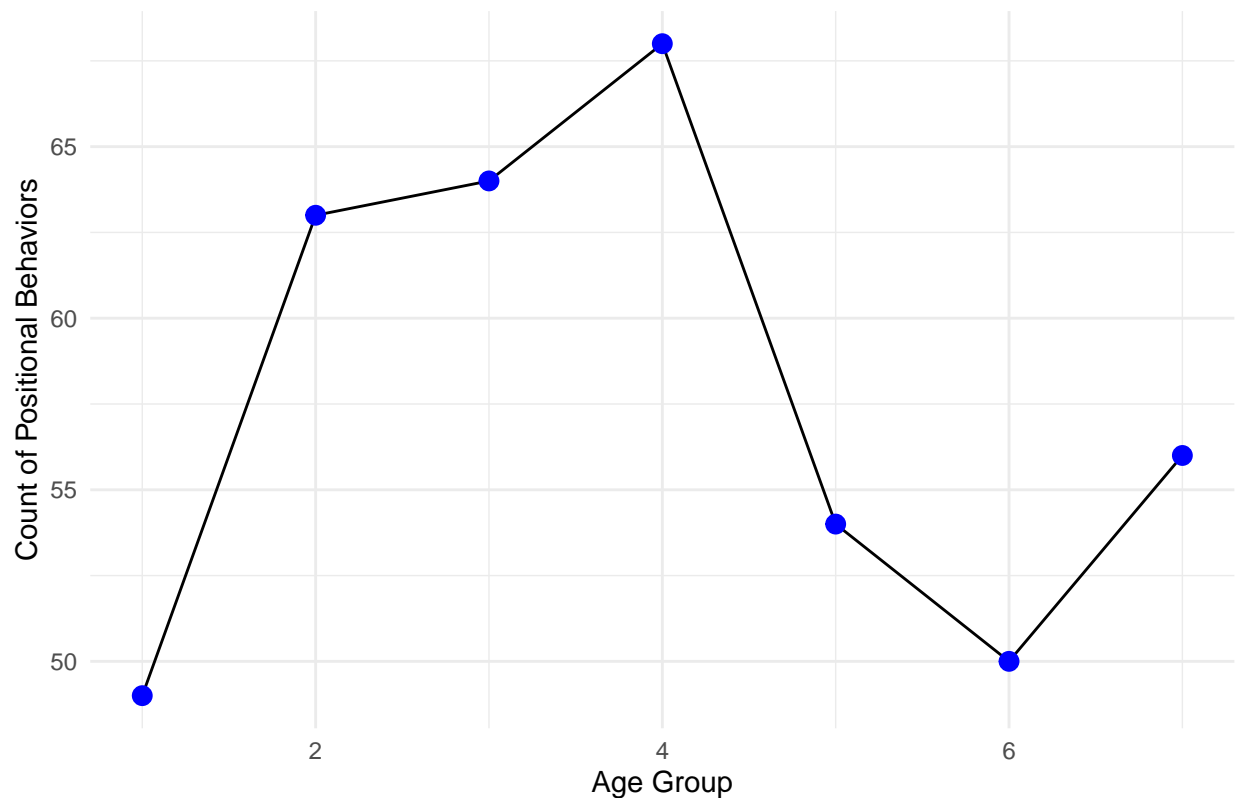
```
## # A tibble: 7 x 2
##   Age count
##   <int> <int>
## 1     1    15
## 2     2    19
## 3     3    23
## 4     4    13
## 5     5    14
## 6     6     8
## 7     7    33
```



The number of observations varies significantly across age groups. Age group 7 has the largest number of observations (about 800), which might introduce a sampling bias in the data analysis. Age group 1 has the smallest count (about 180), which could affect the reliability of statistical summaries or models for that group.

Part 4: Find the Number of Unique Positional Behaviors for Each Age Group

Distinct Count of Positional Behaviors in Each Age Group



```
## # A tibble: 7 x 2
##   Age PB_count
##   <int>   <int>
## 1     4     68
## 2     3     64
## 3     2     63
## 4     7     56
## 5     5     54
## 6     6     50
## 7     1     49
```

From the table above, Age 4 group has 68 kinds of positional behaviors, which is the most number of distinct positional behaviors across all age groups. Age 1 group has only 49 kinds of positional behaviors, which is the least number of distinct behaviour among 7 groups.

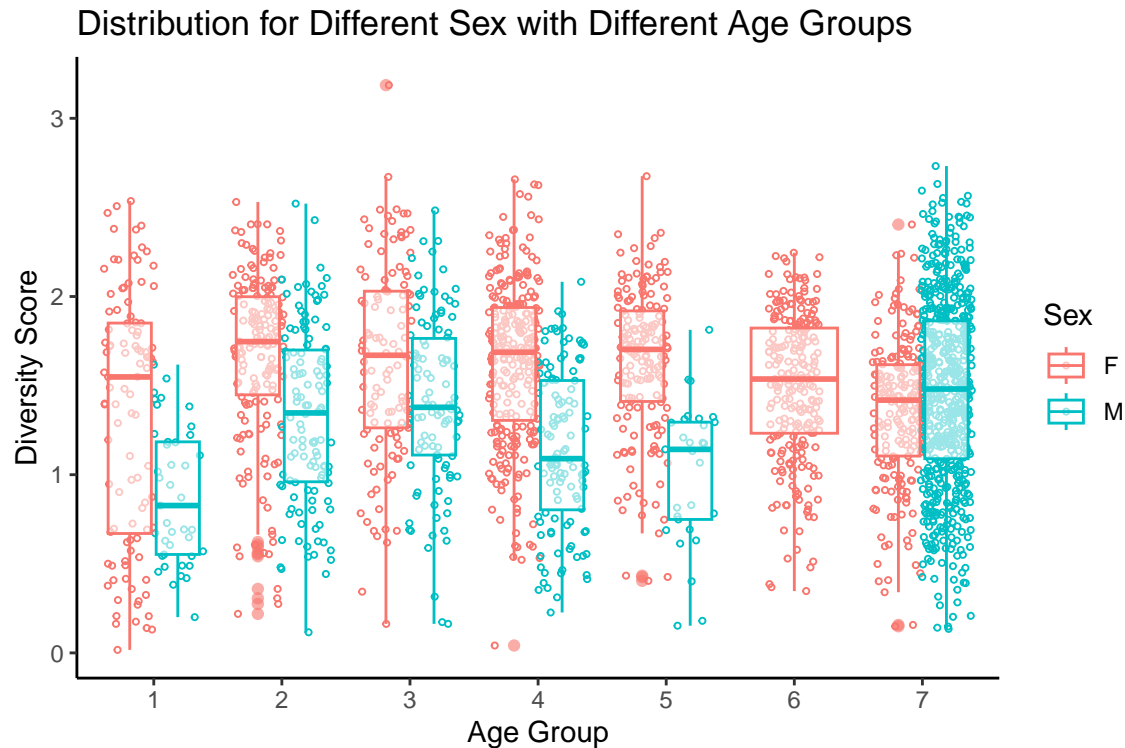
Also, from the line graph, starting from age 2 to 4, there is a surge in the number of distinct positional behaviors, and after year 4, there is a significant decrease in number of positional behaviors, which means that year 4 is an important time point to focus on.

Part 5: Difference in Diversity Score between Sex

We noticed that there is values “M?” in the sex column, so we changed “M?” to “M”. Now we can check the unique value in the sex column:

```
## [1] "F" "M"
```

After cleaning the Sex column, we drew a box plots to show the difference in distribution with different sex in different age groups:



From the plot above, there is clear distinction between distribution of diversity score for different sex. For age group 1 to 5, the median diversity score is higher for female compared to male. For age group 6, there is only female data in the dataset, and for age group 7, the median for both female and male group is similar.

According to EDA, we also think minute awake is an important predictor in modeling the diversity score, so we add Minutes.awake variable to the dataset:

```
## Rows: 2,247
## Columns: 7
## $ Name      <chr> "alfred", "alfred", "alfred", "alfred", "alfred", "alfr~
## $ FollowNumber <int> 6749, 6751, 6754, 6904, 6905, 6926, 6928, 6934, 7115, 7~
## $ X         <int> 1825, 1826, 1820, 1827, 1810, 1828, 1823, 1829, 1801, 1~
## $ Age       <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7~
## $ Sex       <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", ~
## $ DiversityScore <dbl> 1.4885210, 1.2704632, 1.3294911, 1.4635908, 1.7633772, ~
## $ Minutes.awake <dbl> 750.0000, 655.0000, 640.0000, 721.0000, 664.0000, 503.0~
```

Part 6: Standard Deviation of Diversity Score for Each Name Group

```
## # A tibble: 16 x 6
##   Name      count  max  min median  std
##   <chr>      <int> <dbl> <dbl>  <dbl> <dbl>
```

## 1 aminah	1	1.92	1.92	1.92	NA
## 2 anik	1	1.23	1.23	1.23	NA
## 3 cili	1	1.79	1.79	1.79	NA
## 4 female 10	1	2.40	2.40	2.40	NA
## 5 fl14may99	1	1.05	1.05	1.05	NA
## 6 fl9oct15	1	0.998	0.998	0.998	NA
## 7 fml04mar23_rk	1	0.477	0.477	0.477	NA
## 8 fml07okt23_rk	1	1.74	1.74	1.74	NA
## 9 fml27may23	1	0.497	0.497	0.497	NA
## 10 fml2aug2018	1	0.842	0.842	0.842	NA
## 11 jupiter	1	1.49	1.49	1.49	NA
## 12 ml08jan02a	1	1.87	1.87	1.87	NA
## 13 moss	1	1.30	1.30	1.30	NA
## 14 neli	1	0.947	0.947	0.947	NA
## 15 nunu	1	1.61	1.61	1.61	NA
## 16 uml30mar01	1	1.13	1.13	1.13	NA

According to the standard deviation of diversity score in different name group, the name can be treated as a group variable(random effect). Those names in the above table only have 1 observation, we may consider remove those levels.

Method

How does Shannon Weaver Index calculate?

The formula is shown below: $H' = -\sum_{i=1}^S p_i \ln(p_i)$ In this formula, S represents the number of unique categories in the data, p_i is the proportion, which in the client's data refers to the proportion of recorded time of each distinct behaviors of a specific orangutan in relation to this orangutan's total minutes of awake.

Client Models

Following client suggestions, we would like to use linear mixed effect models and build upon client's models and see if we improve the client's models. Initial client models are listed below:

```
model.1 <- glmmTMB(data=df_score,
  DiversityScore ~ (1|Name) + (1|FollowNumber),
  family=gaussian(link="log"))

model.2 <- glmmTMB(data=df_score,
  DiversityScore ~ Age + (1|Name) + (1|FollowNumber),
  family=gaussian(link="log"))

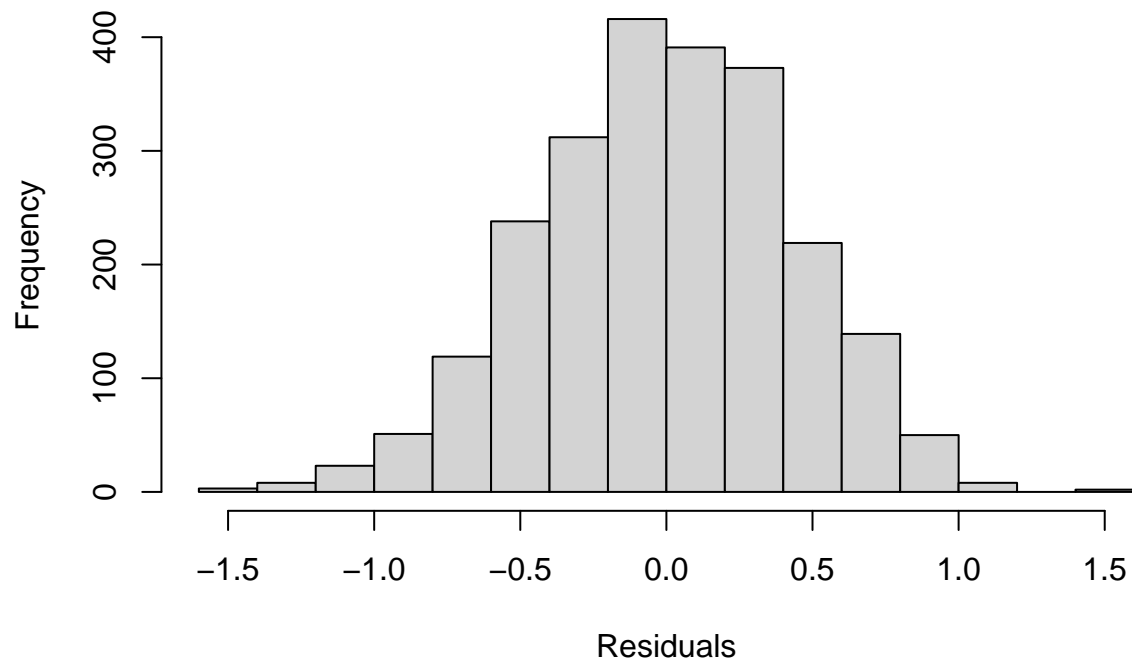
model.3 <- glmmTMB(data=df_score,
  DiversityScore ~ Age + I(Age^2) +
    (1|Name) + (1|FollowNumber),
  family=gaussian(link="log"))
```

Check Client's Model Assumptions

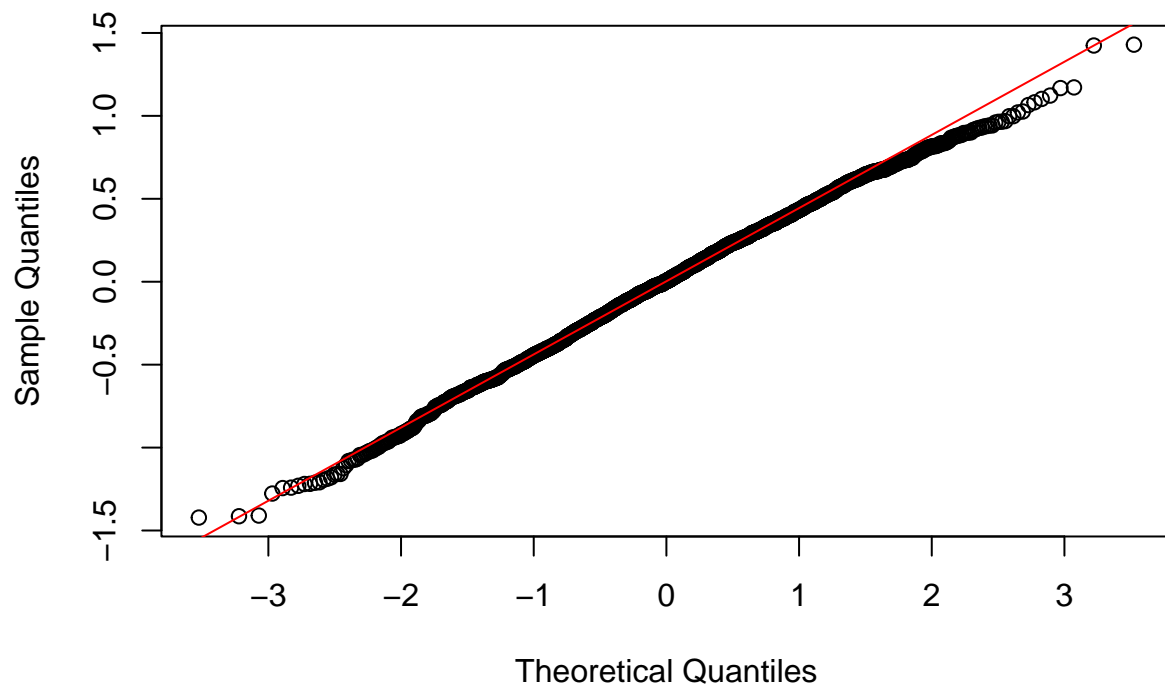
We first need to check all the linear assumptions are met:

Model 1

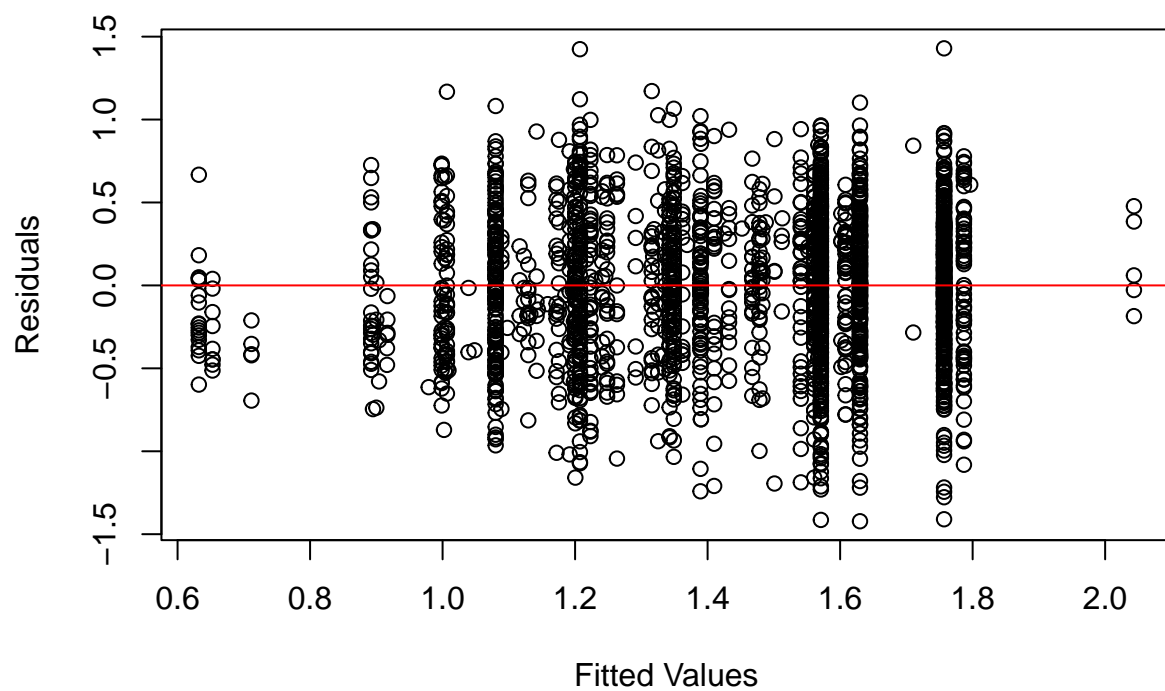
Histogram of Residuals



Normal Q-Q Plot

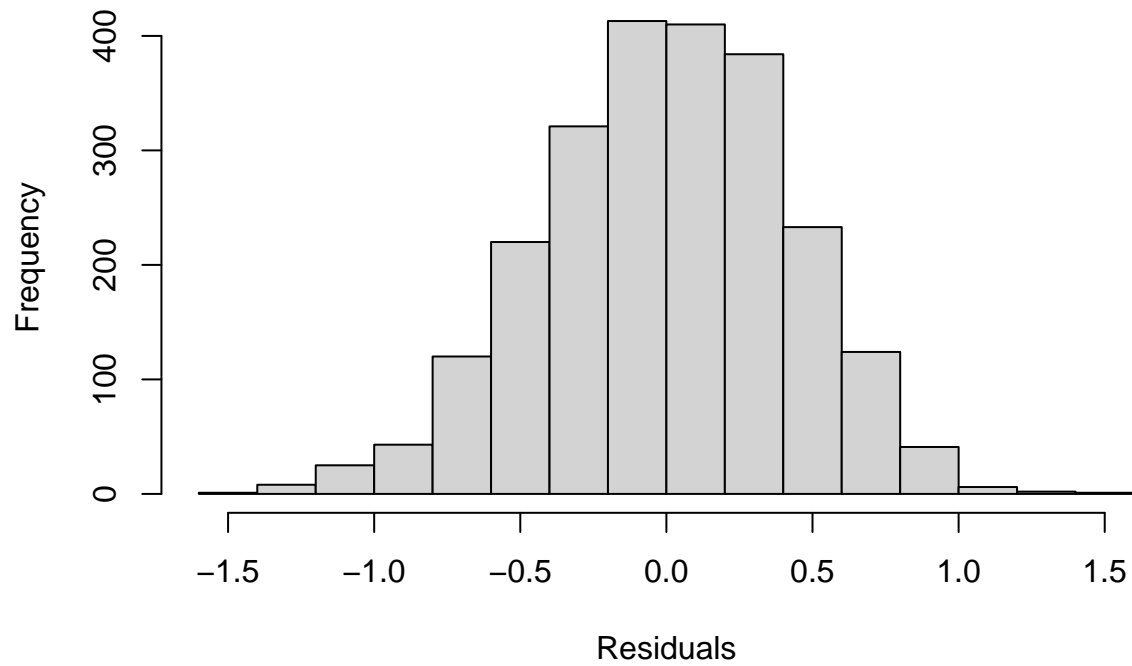


Residuals vs Fitted Values

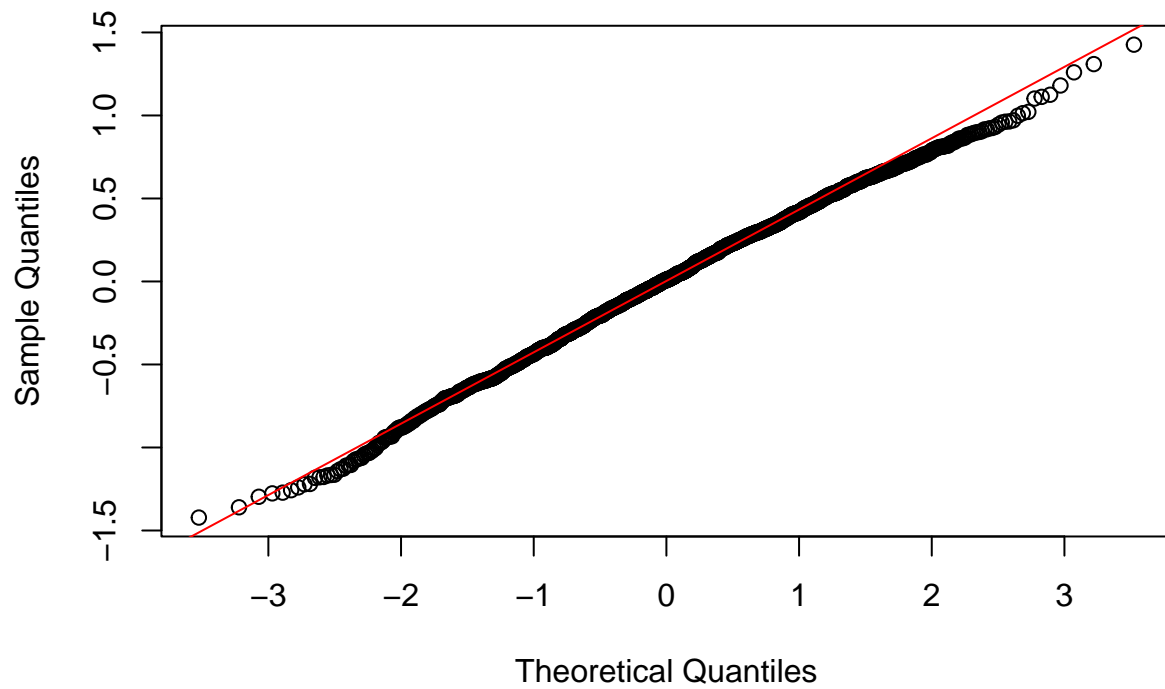


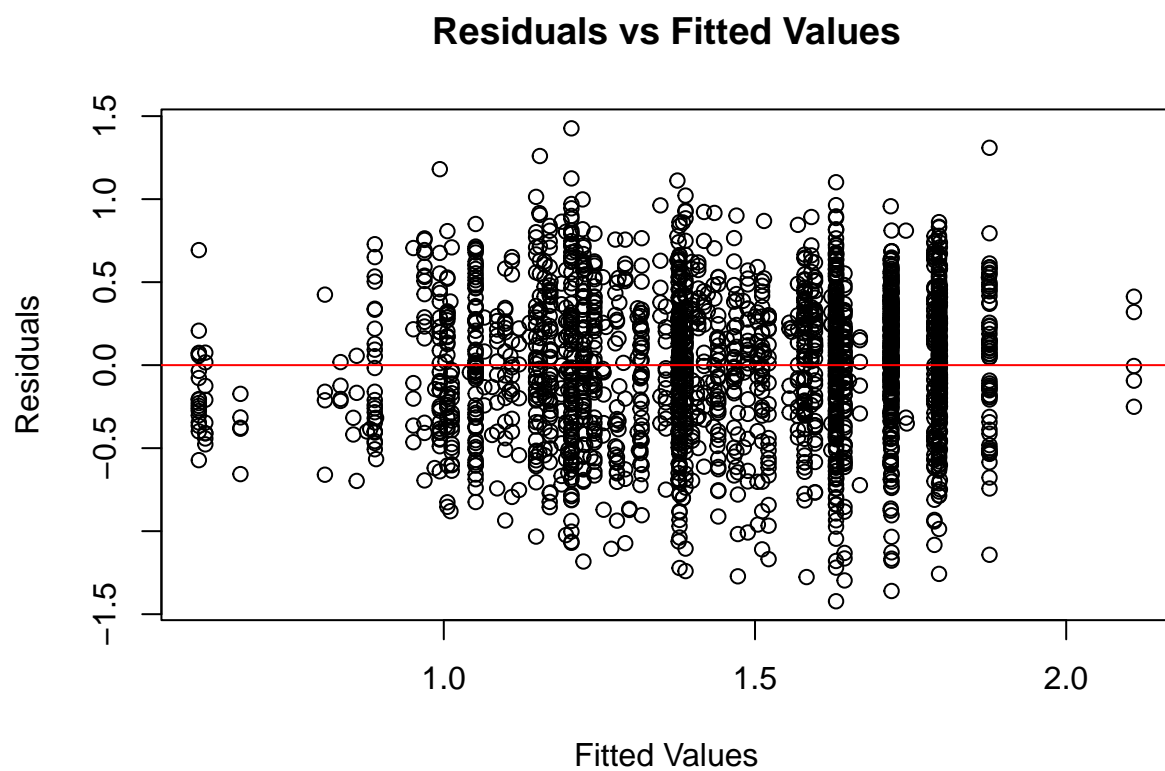
Model 2

Histogram of Residuals



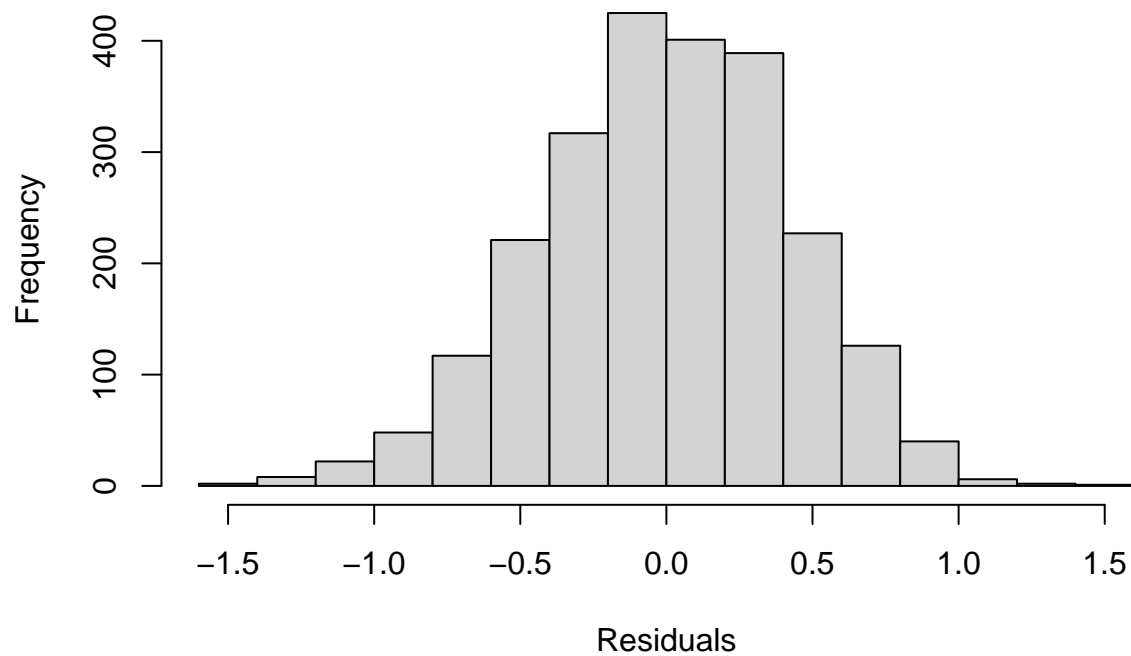
Normal Q-Q Plot



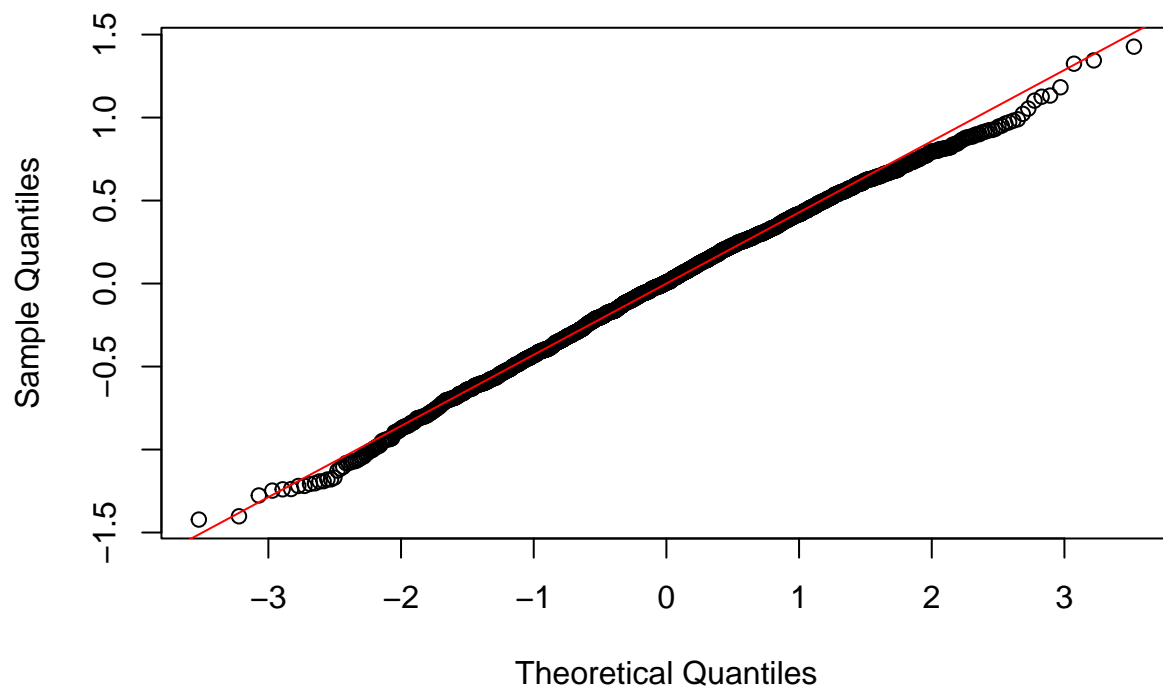


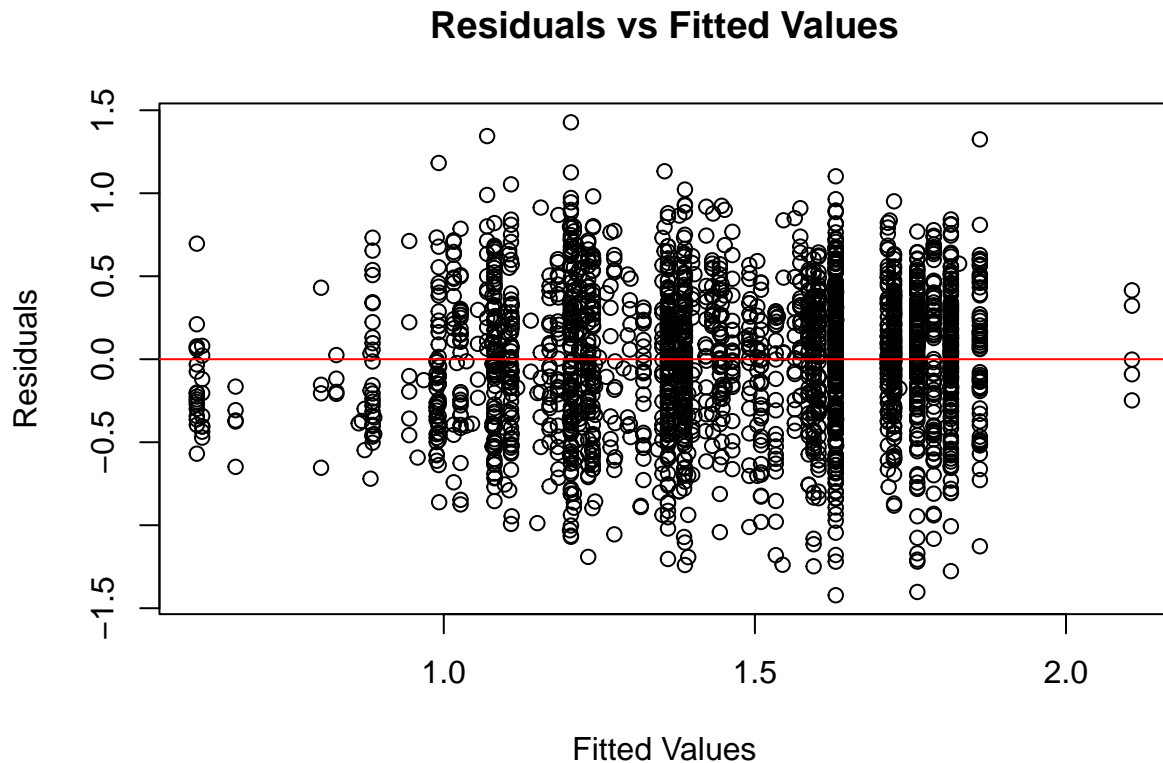
Model 3

Histogram of Residuals



Normal Q-Q Plot





Based on the assumptions check above, all three models have normal distributed residuals, no-heavy tails in qq plot and random distribution in residual vs fitted plot, indicating that all the models satisfy linear assumptions.

Build Upon Client's Model

Here we propose three new models building upon the client models, we treat age as a categorical variable, and we also add sex and minutes.awake into the model:

```
df_sex$Age <- as.factor(df_sex$Age)
df_sex <- df_sex %>%
  filter(!is.na(Minutes.awake))

newmodel.1 <- glmmTMB(data=df_sex, DiversityScore ~ Age + (1|Name),
  family = gaussian("log"))

newmodel.2 <- glmmTMB(data=df_sex,
  DiversityScore ~ Age + Sex + (1|Name),
  family = gaussian("log"))

newmodel.3 <- glmmTMB(data = df_sex,
  DiversityScore ~ Age + Sex + Minutes.awake + (1 | Name),
  family = gaussian("log"))

anova(newmodel.1, newmodel.2)
```

```
## Data: df_sex
## Models:
## newmodel.1: DiversityScore ~ Age + (1 | Name), zi=~0, disp=~1
## newmodel.2: DiversityScore ~ Age + Sex + (1 | Name), zi=~0, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## newmodel.1  9 2660.4 2711.8 -1321.2  2642.4
## newmodel.2 10 2661.0 2718.1 -1320.5  2641.0 1.4486      1    0.2288
```

```
anova(newmodel.2, newmodel.3)
```

```
## Data: df_sex
## Models:
## newmodel.2: DiversityScore ~ Age + Sex + (1 | Name), zi=~0, disp=~1
## newmodel.3: DiversityScore ~ Age + Sex + Minutes.awake + (1 | Name), zi=~0, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## newmodel.2 10 2661.0 2718.1 -1320.5  2641.0
## newmodel.3 11 2651.6 2714.4 -1314.8  2629.6 11.365      1 0.0007486 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(newmodel.1, newmodel.3)
```

```
## Data: df_sex
## Models:
## newmodel.1: DiversityScore ~ Age + (1 | Name), zi=~0, disp=~1
## newmodel.3: DiversityScore ~ Age + Sex + Minutes.awake + (1 | Name), zi=~0, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## newmodel.1  9 2660.4 2711.8 -1321.2  2642.4
## newmodel.3 11 2651.6 2714.4 -1314.8  2629.6 12.813      2 0.001651 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the chi-square test, model.3 is the best model, and minutes awake, age group 2, 3, 6 and 7 are statistically significant variables. Sex is not a statistically significant variable, we can consider removing it in the later model.

```
newmodel.4 <- glmmTMB(data = df_sex,
                      DiversityScore ~ Age + Minutes.awake + (1 | Name),
                      family = gaussian("log"))
summary(newmodel.4)
```

```
## Family: gaussian ( log )
## Formula:      DiversityScore ~ Age + Minutes.awake + (1 | Name)
## Data: df_sex
##
##      AIC      BIC  logLik deviance df.resid
## 2651.1  2708.2 -1315.6  2631.1      2224
##
## Random effects:
##
## Conditional model:
## Groups   Name      Variance Std.Dev.
```

```
## Name      (Intercept) 0.08727  0.2954
## Residual                0.17784  0.4217
## Number of obs: 2234, groups: Name, 71
##
## Dispersion estimate for gaussian family (sigma^2): 0.178
##
## Conditional model:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.079e-01  6.829e-02  1.580 0.114136
## Age2         8.453e-02  3.468e-02  2.438 0.014776 *
## Age3         1.132e-01  4.446e-02  2.545 0.010914 *
## Age4        -7.611e-04  4.446e-02 -0.017 0.986342
## Age5        -4.776e-02  4.634e-02 -1.031 0.302708
## Age6        -1.531e-01  4.341e-02 -3.527 0.000420 ***
## Age7        -1.996e-01  3.901e-02 -5.116 3.12e-07 ***
## Minutes.awake 2.368e-04  7.153e-05  3.310 0.000933 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

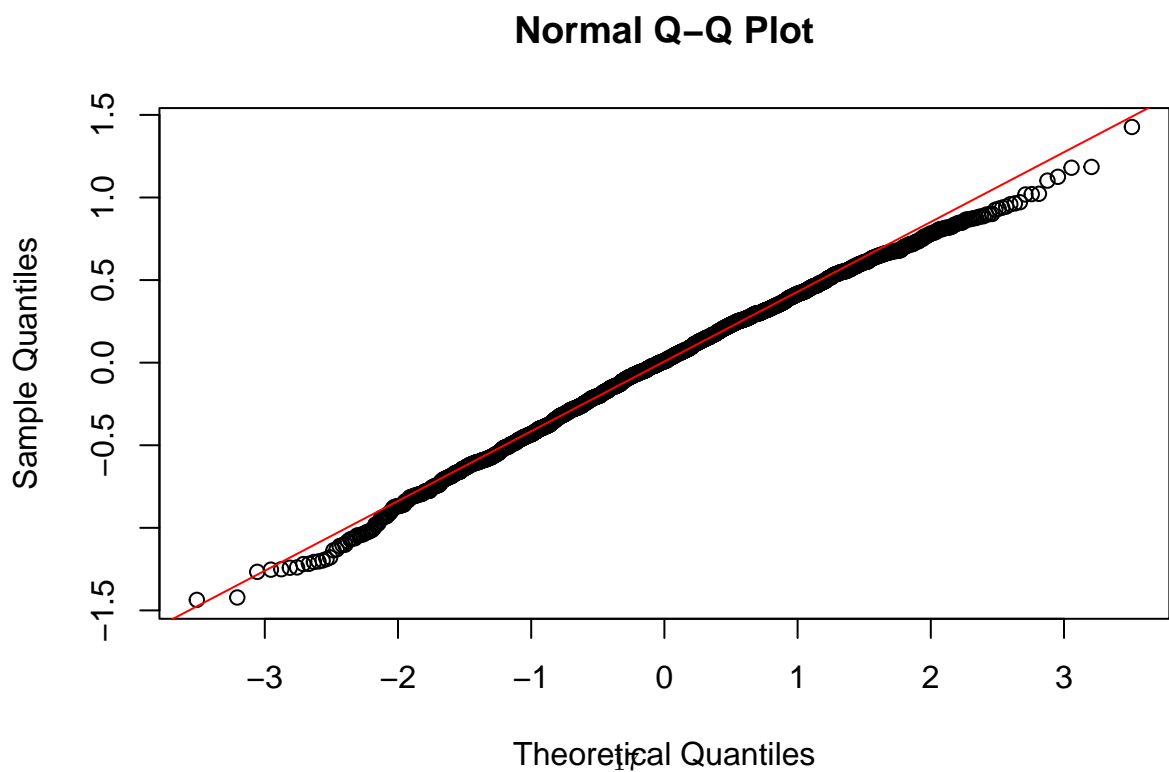
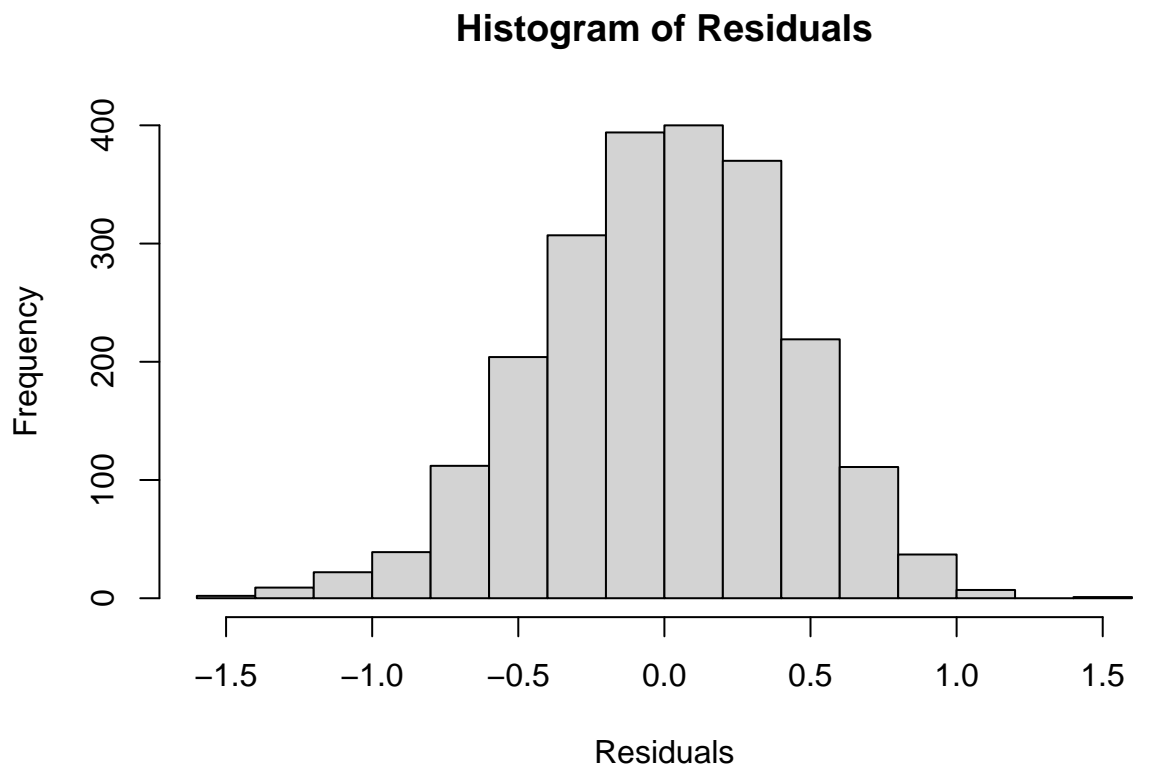
```
anova(newmodel.4, newmodel.3)
```

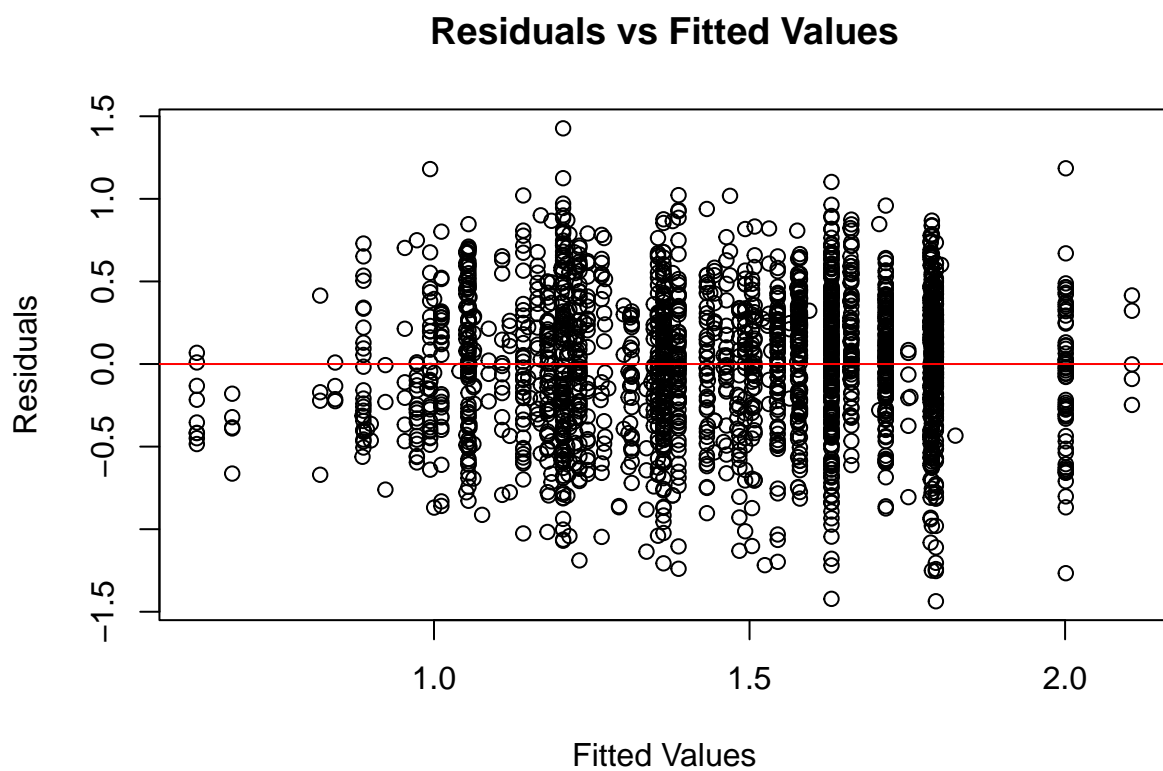
```
## Data: df_sex
## Models:
## newmodel.4: DiversityScore ~ Age + Minutes.awake + (1 | Name), zi=~0, disp=~1
## newmodel.3: DiversityScore ~ Age + Sex + Minutes.awake + (1 | Name), zi=~0, disp=~1
##           Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## newmodel.4 10 2651.1 2708.2 -1315.5  2631.1
## newmodel.3 11 2651.6 2714.4 -1314.8  2629.6 1.4896      1      0.2223
```

According to the chi-square result above, the p-value greater than 0.05, showing that we cannot reject the null hypothesis, and adding variable sex does not significantly improve the model performance. We can further check the performance of model 4 using AIC and MSE below.

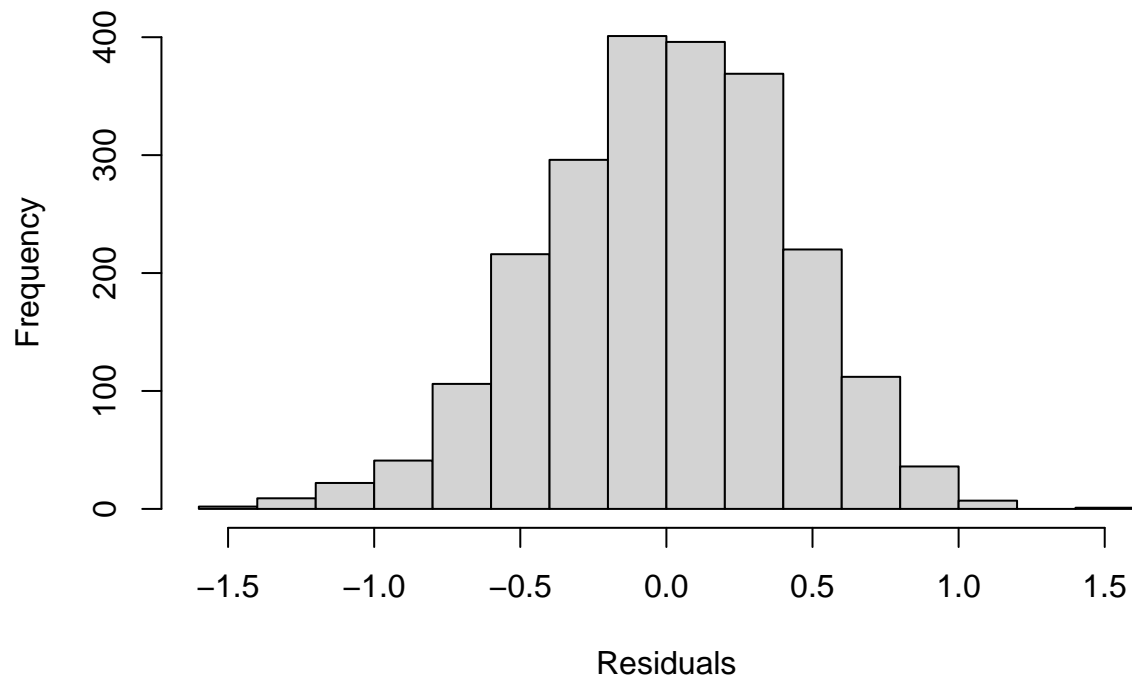
Model Diagnostics for Our Model

New Model 1

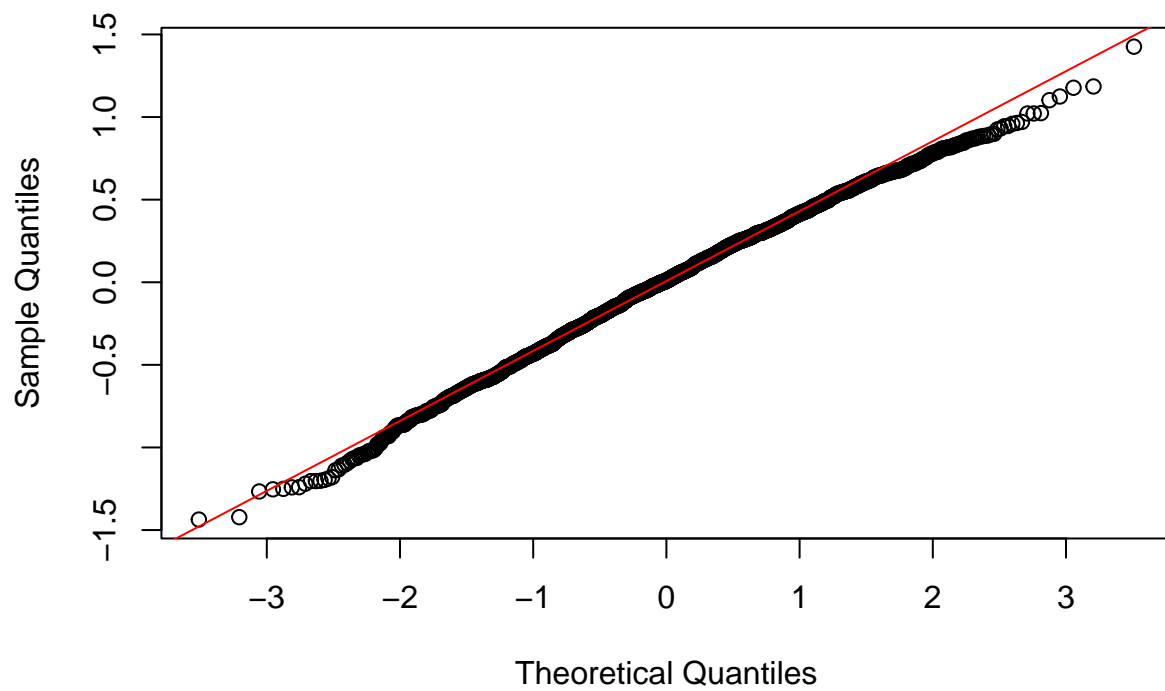


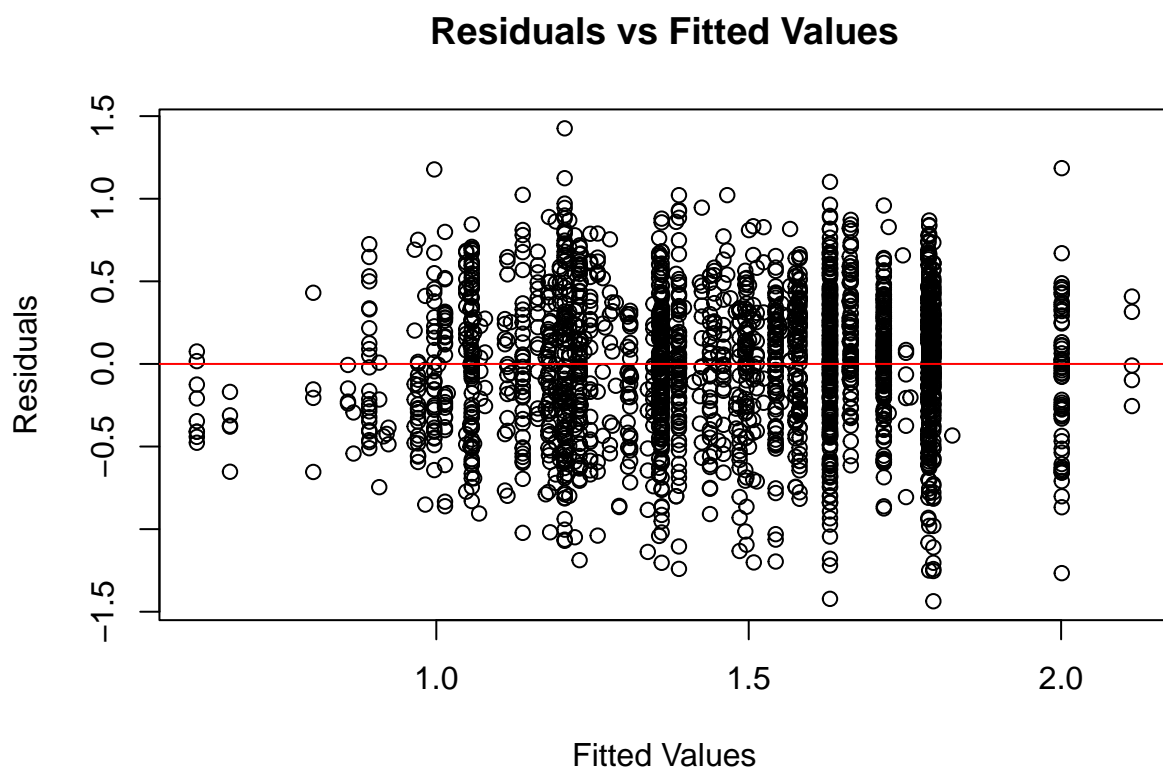


Histogram of Residuals

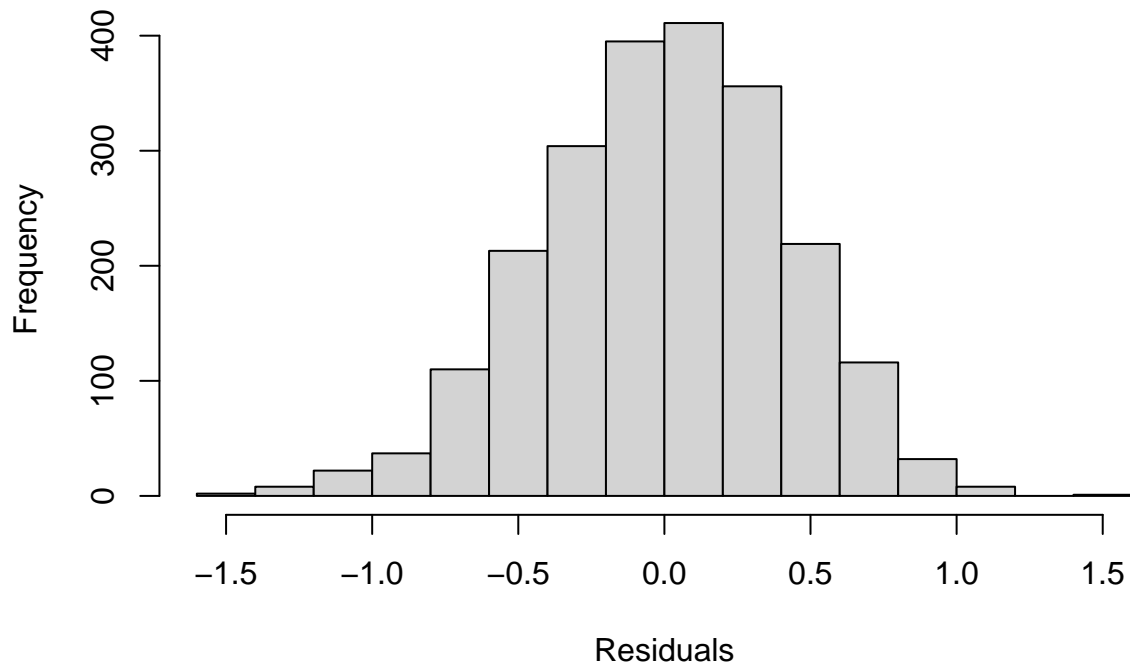


Normal Q-Q Plot

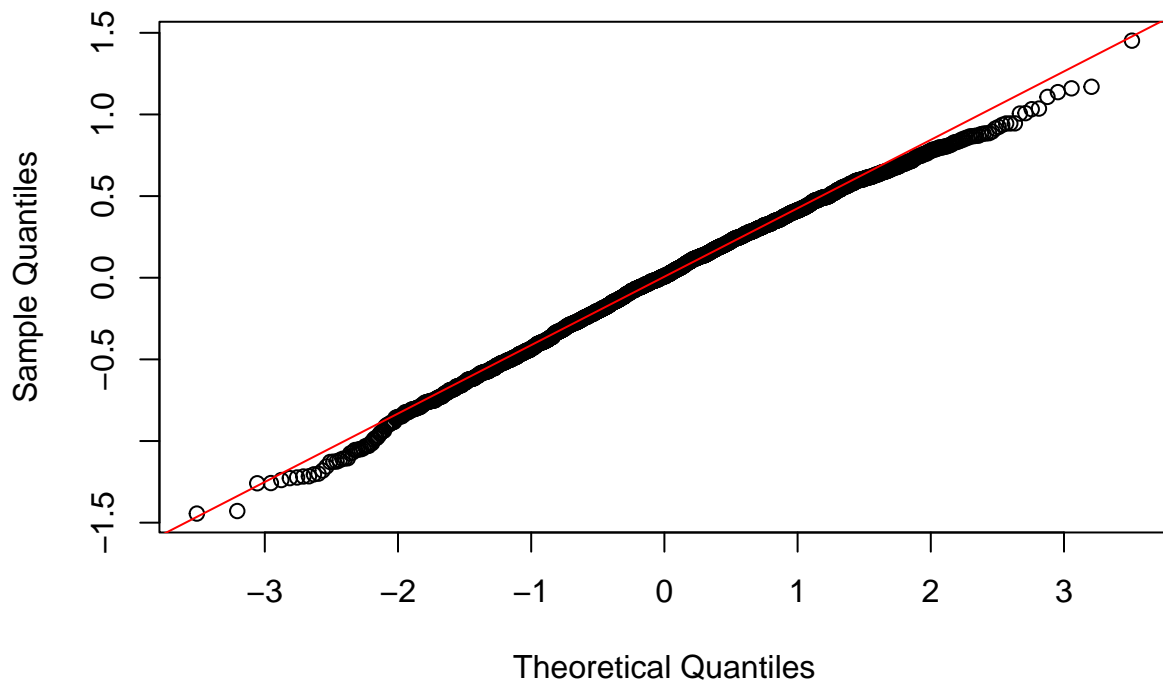




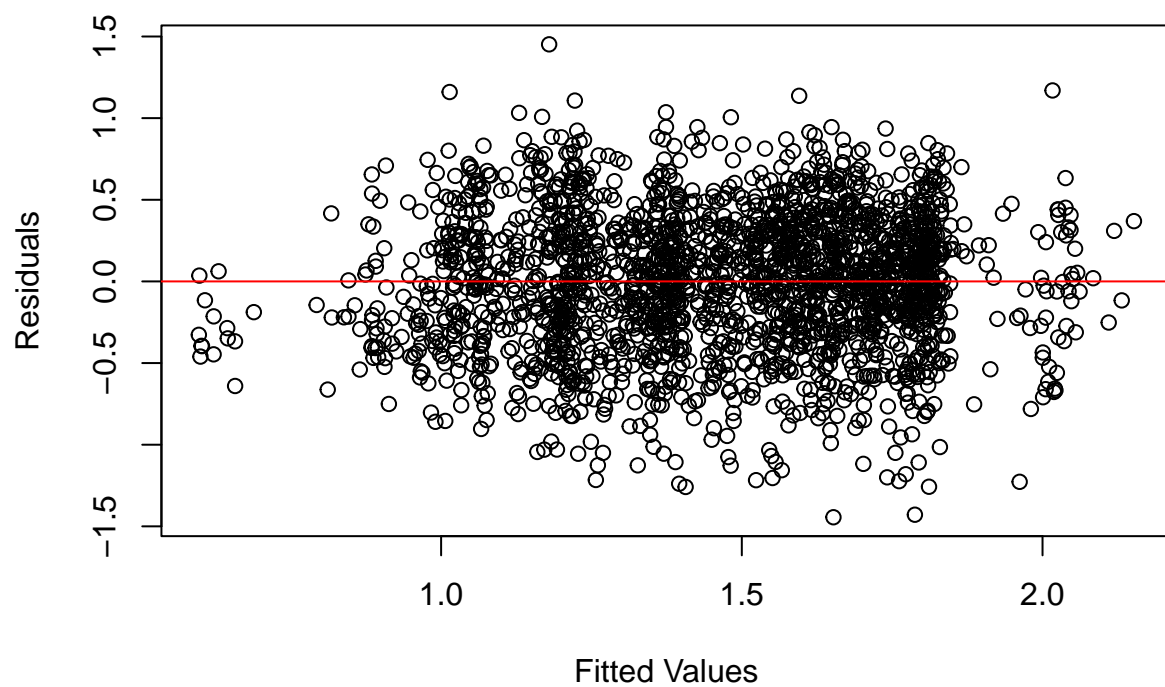
Histogram of Residuals



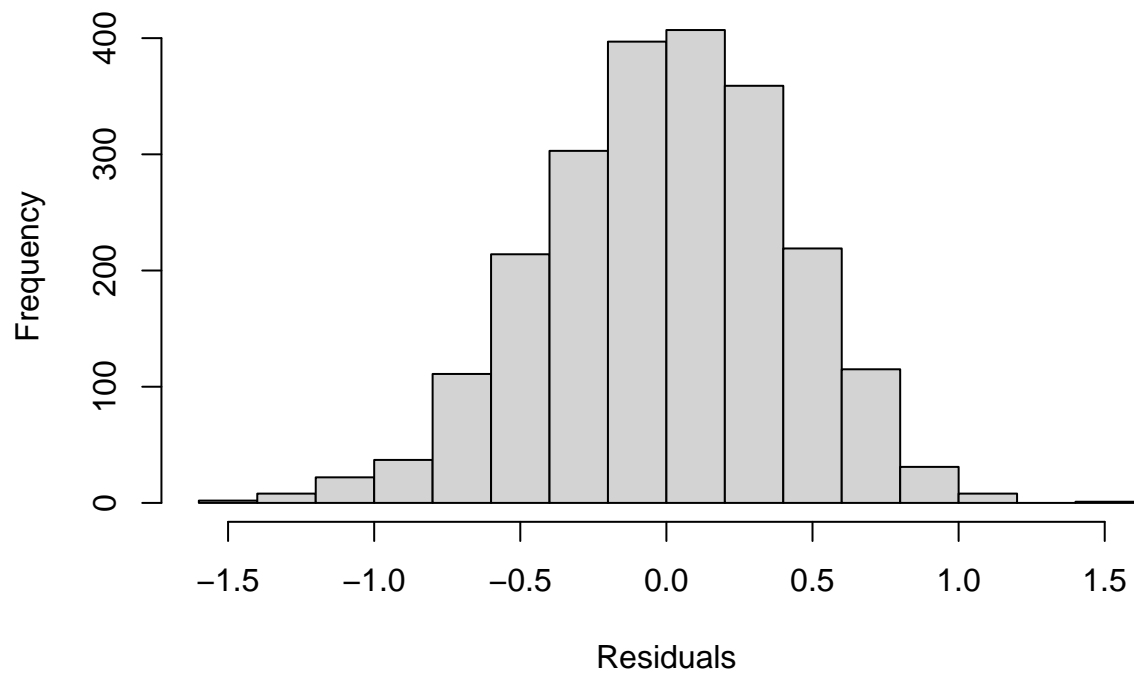
Normal Q-Q Plot



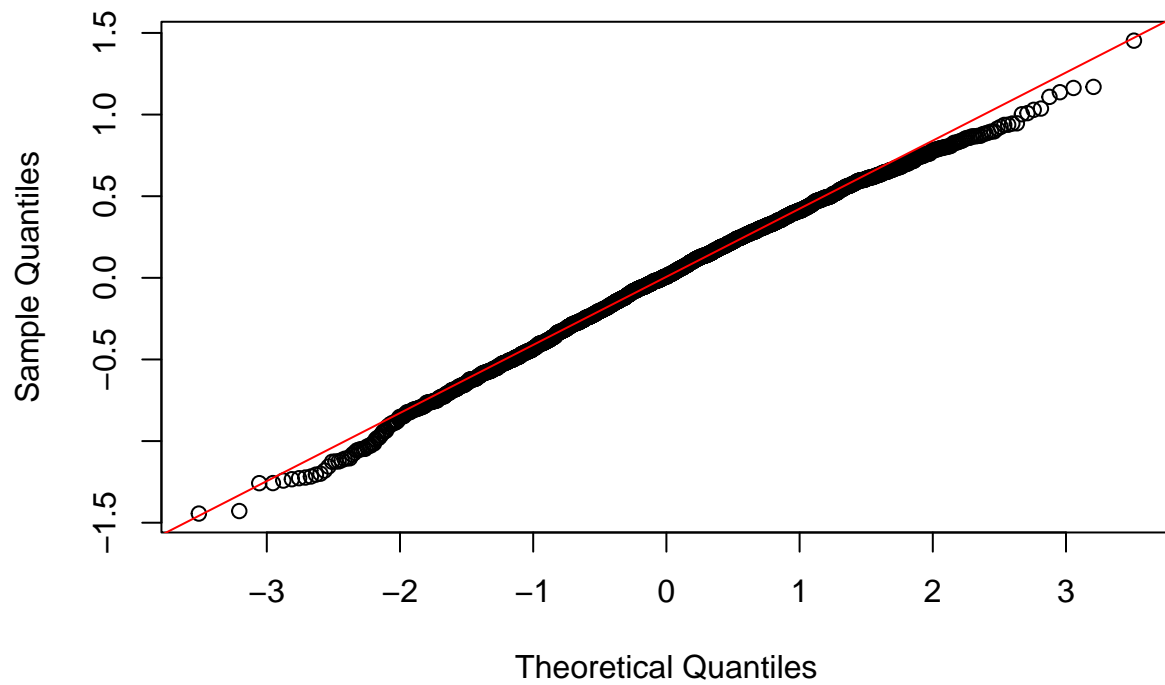
Residuals vs Fitted Values

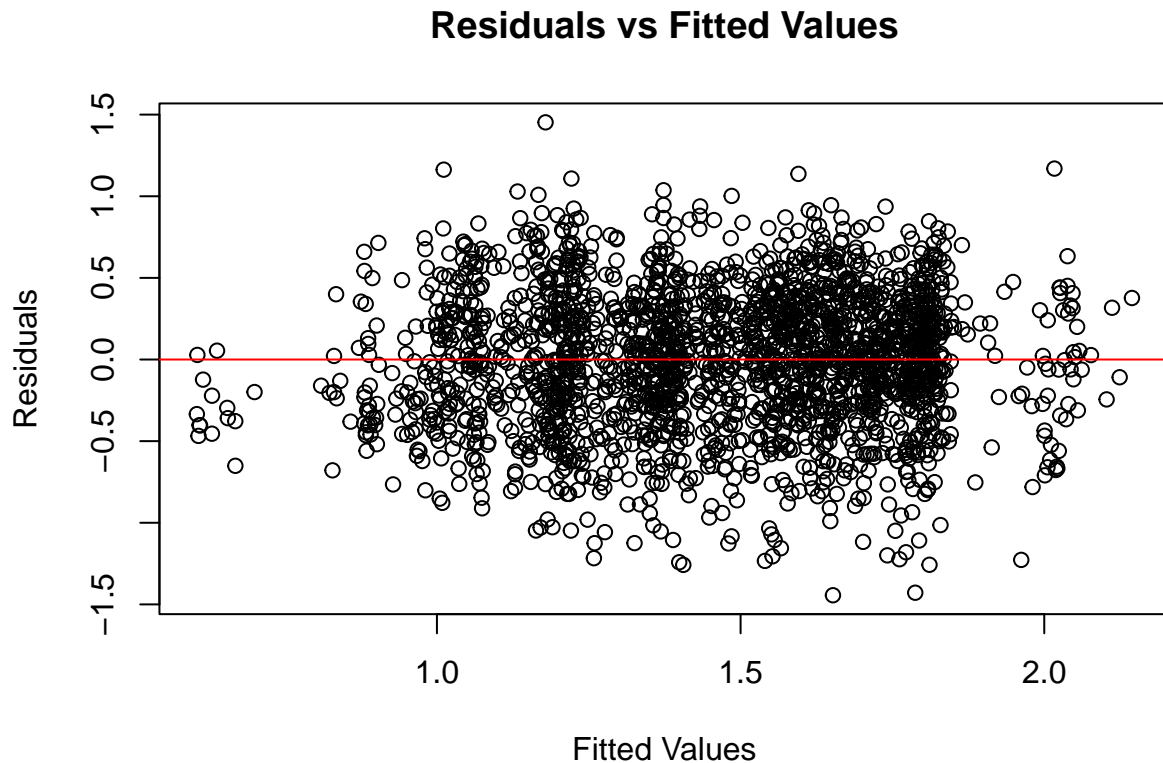


Histogram of Residuals



Normal Q-Q Plot





By checking the models assumptions, all four of our models satisfy the model assumptions.

Compare Our Model to The Client's Model

##	Model	AIC	MSE
## 1	Model 1	2953.148	0.3479651
## 2	Our Model 1	2660.426	0.1744798
## 3	Model 2	2868.769	0.3550065
## 4	Our Model 2	2660.977	0.1744503
## 5	Model 3	2858.630	0.3539804
## 6	Our Model 3	2651.612	0.1735441
## 7	Our Model 4	2651.102	0.1735711

According to the AIC and MSE comparison above, our models have lower AIC and MSE, indicating that our model performs better. Although from the ANOVA table above, adding variable sex does not further improve the model, Model 3 does have smaller MSE but higher AIC compared to model 4. In order to achieve simpler model structure, we plan to choose model 4.

Model 4 Performance Using Cross Validation

To validate the performance of the model 4 above, we propose to use k-fold cross validation:

```
# Custom function to fit models
fit_glmmTMB <- function(train_data, test_data) {
```



```

model <- glmmTMB(
  data = train_data,
  DiversityScore ~ Age + Minutes.awake + (1 | Name),
  family = gaussian(link = "log")
)

predictions <- predict(model, newdata = test_data,
  allow.new.levels = TRUE, type = "response")
mse_value <- mean((test_data$DiversityScore - predictions)^2)

return(mse_value)
}

# Split data into 5 folds
set.seed(724) # For reproducibility
folds <- createFolds(df_sex$DiversityScore, k = 5, list = TRUE)

cv_results <- sapply(folds, function(test_indices) {
  test_data <- df_sex[test_indices, ]
  train_data <- df_sex[-test_indices, ]

  fit_glmmTMB(train_data, test_data)
})

# Calculate average MSE across folds
mean_mse <- mean(cv_results)
cat("Average MSE across folds:", mean_mse, "\n")

```

```
## Average MSE across folds: 0.1867108
```

Fit Model 4 Using Data with Names with Only 1 Observation Removed

```

df_sex1 <- df_sex %>%
  group_by(Name) %>%
  filter(n() > 1) %>%
  ungroup()

model_remove_name <- glmmTMB(data = df_sex1,
  DiversityScore ~ Age + Minutes.awake + (1 | Name),
  family = gaussian("log"))
predictions <- predict(model_remove_name, type = "response")
mse_model_remove <- mean((df_sex1$DiversityScore - predictions)^2)
print(mse_model_remove)

```

```
## [1] 0.174117
```

Refit Model 4 with stan_glmer

```
stanmodel_4 <- stan_glmer(
  DiversityScore ~ Age + Minutes.awake + (1 | Name),
  data = df_sex,
  family = gaussian(link = "log"),
  refresh = 0
)

predictions <- posterior_predict(stanmodel_4, type = "response")
mean_prediction <- colMeans(predictions)
mse_stanmodel4 <- mean((df_sex$DiversityScore - mean_prediction)^2)
print(mse_stanmodel4)

## [1] 0.1735557
```

Try XGBoost Model

For the next steps, we can also try some machine learning method, such as random forest and xgboost to further improve the modeling performance.