Final Project: Music vs. Mental Health

Hieu (Calvin) Hoang, Karen Nguyen, Vy Vo

Submitted 12-11-2023

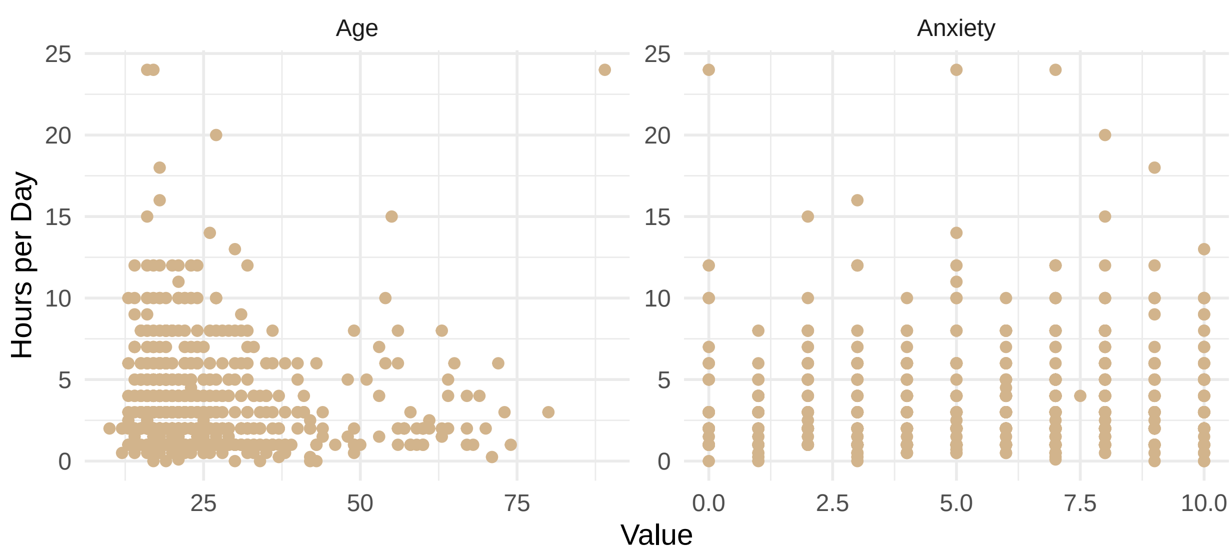
Distribution Report:

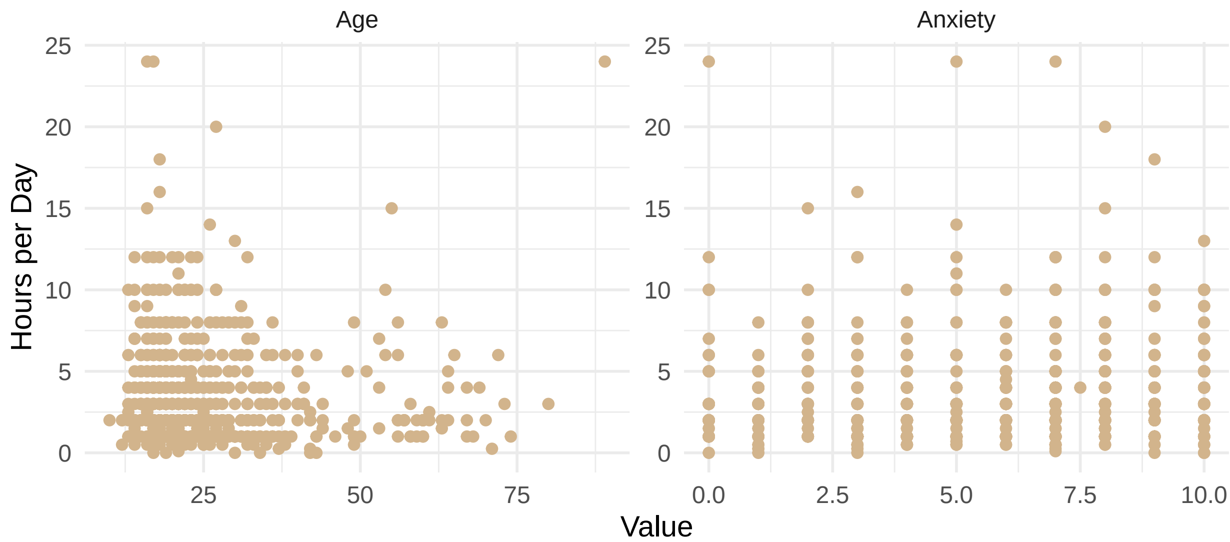
Hieu (Calvin) Hoang: Organization, decoding, interpretation/analysis and editor
Karen Nguyen: Coding, interpretation/analysis, and writer
Vy Vo: Coding, interpretation/analysis, and writer

Introduction

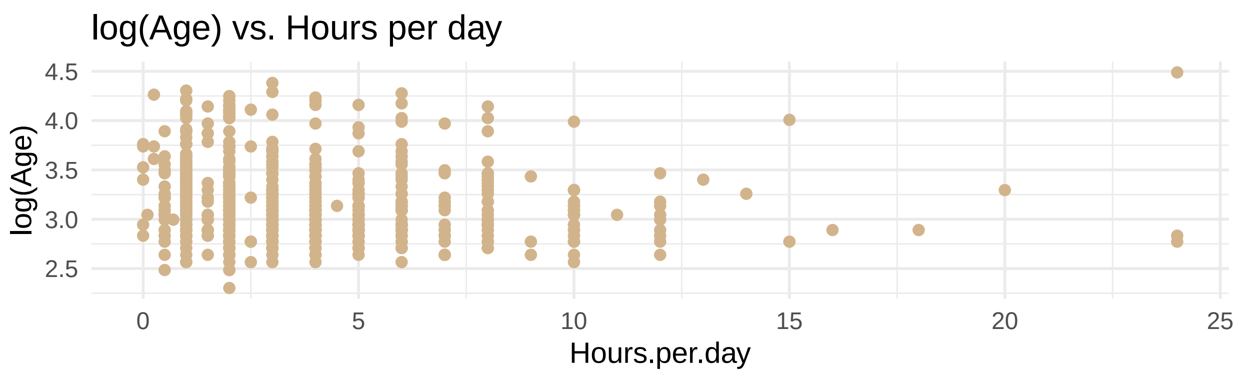
In this project, we will investigate the relationship between how many hours per day someone listens to music (our response variable) and their age and mental health. Our project utilized data collected by Lauren Lavelle called 'Mental Health and Music Analysis'. The dataset consist of 736 survey responses with variables such as the number of hours individuals listen to music, ages, and rating levels of anxiety, depression, and insomnia.

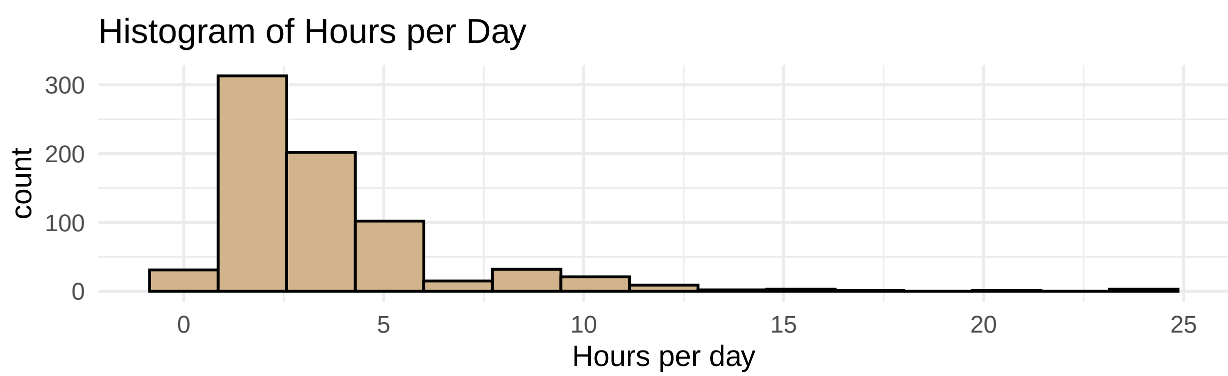
Testing Non-linearity





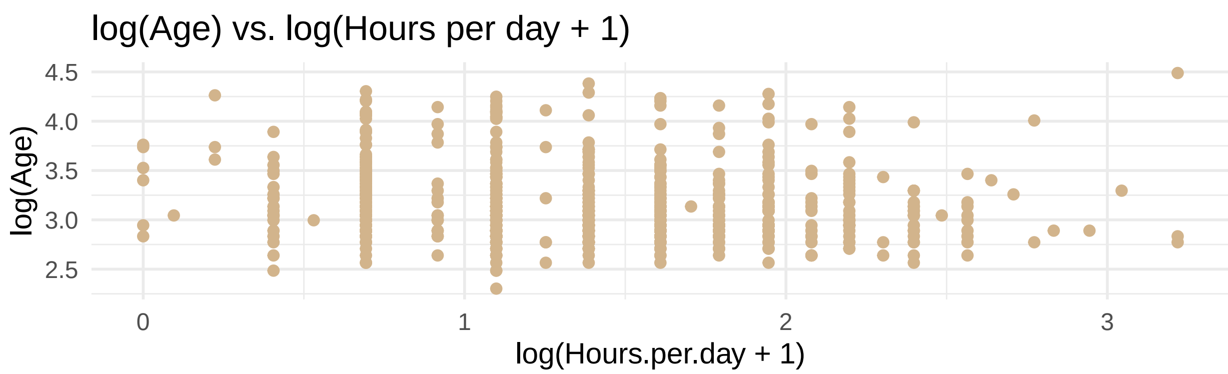
There doesn't seem to be a linear relationship between age and hours per day. We can transform age by logging it so that the relationship looks less non-linear.
With log age:





Since the hours per day that people listen to music is not normally distributed, for each of the models, we will have a model where we log Hours.per.day and a model where we fit it to a log-link Gamma distribution.

Also, since half our models will have log Hours.per.day, we check if the relationship with log(Age) and log(Hours.per.day + 1) is linear.

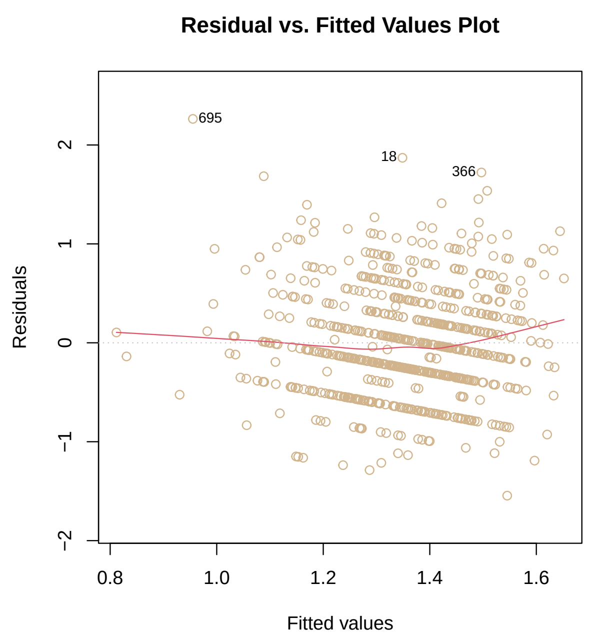


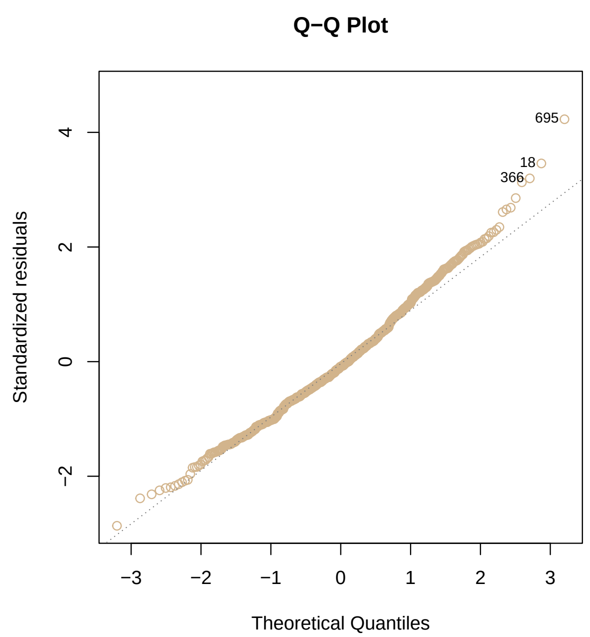
The relationship between log(Age) and log(Hours.per.day + 1) is not obviously non-linear, so we can use log(Age) in the linear model with log(Hours.per.day + 1).

Fitted models

For fitA, which will have a log model, `fitAlog`, and a gamma generalized linear model (glm), `fitAgamma`, with covariates `log(Age)`,`Anxious`, `Depressed`, `Insomniac`, and `Music.effects`.
`Anxious`, `Depressed`, and `Insomniac` are the binary categories we made earlier.
`Music.effects` is a categorical variable where people reported what effect they felt music had on their mental health. The categories for `Music.effects` are `Improve`, `No effect`, and `Worsen`.

FitAlog:

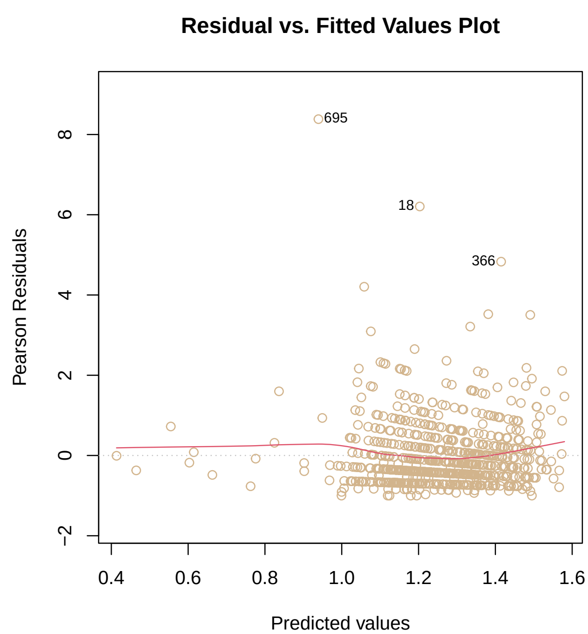


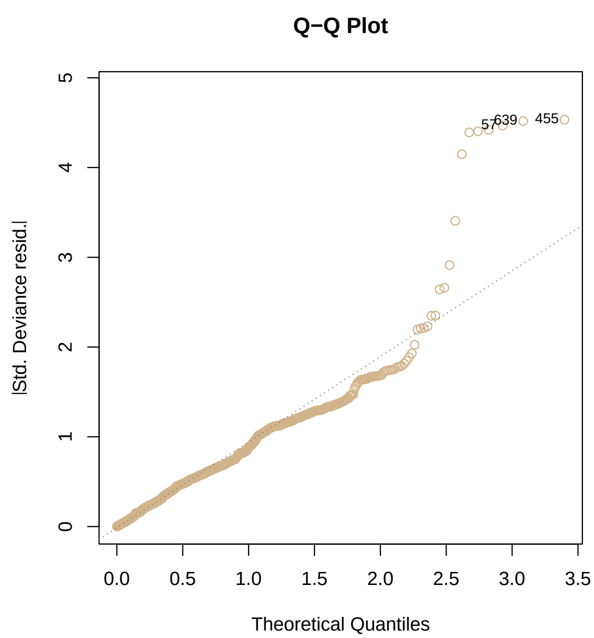



```
##
## Call:
## lm(formula = log(Hours.per.day + 1) ~ log(Age) + Anxiety + Depression +
```

```
##     Insomnia + Music.effects)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54535 -0.35708 -0.04711  0.31904  2.26363
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.483459   0.266045   5.576 3.47e-08 ***
## log(Age)               -0.169466   0.053843  -3.147  0.00171 **
## Anxiety                -0.007347   0.008601  -0.854  0.39326
## Depression              0.018846   0.008103   2.326  0.02030 *
## Insomnia                0.022075   0.007074   3.121  0.00188 **
## Music.effectsImprove    0.302540   0.193610   1.563  0.11858
## Music.effectsNo effect  0.232458   0.196421   1.183  0.23701
## Music.effectsWorsen    -0.023418   0.233742  -0.100  0.92022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5423 on 727 degrees of freedom
## Multiple R-squared:  0.05527,    Adjusted R-squared:  0.04617
## F-statistic: 6.075 on 7 and 727 DF,  p-value: 6.57e-07
```

FitAgamma:





Analysis:

- For Log: From looking at the Residual Vs Fitted plot for `fitAlog`, it seems that the normality assumption is not violated because the shape of the fit is cloud-like without any noticeable pattern. This means that `fitAlog` does have constant variance. However, there are possible outliers such as point

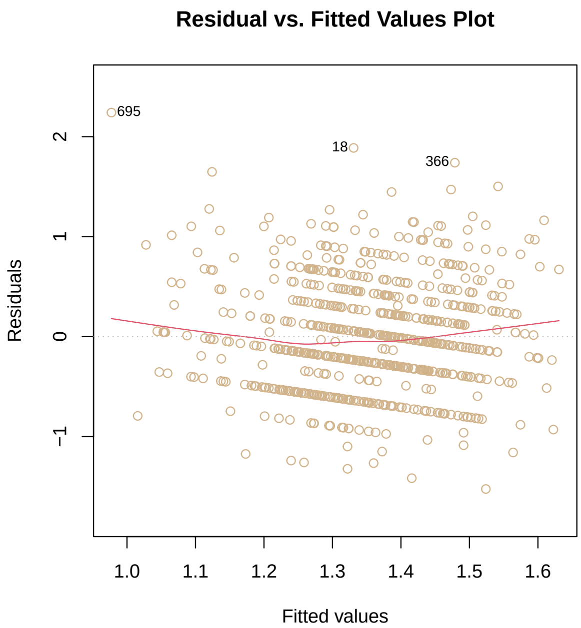
696 or 19. For the Normal Q-Q plot, normality doesn't seem to be violated because most error points remains on the normality line. Nevertheless, there are still evidences of outliers.

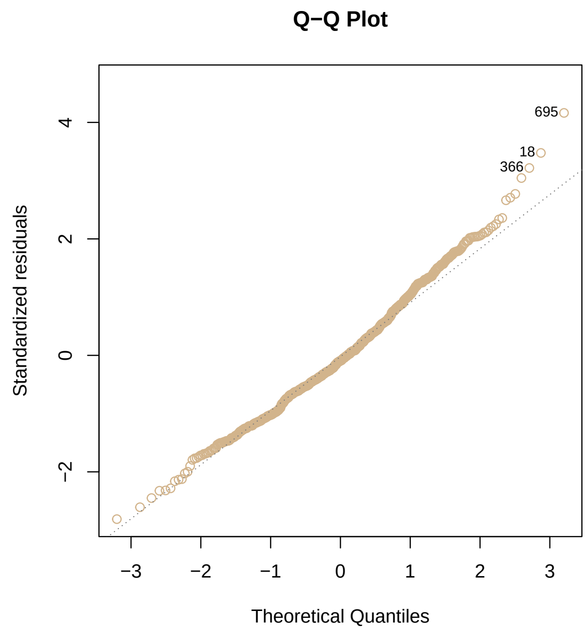
- For Gamma: Since the distribution is Gamma, it is possible to observe clustering of negatively points in the Residuals vs. Fitted Values Plot. This means that `fitAgamma`, doesn't seem to violate normality assumption because there is no pattern. In another word, fitAgamma does appear to have a constant Variance. However, there are possible outliers such as points 696 or 19. For the Normal Q-Q plot, Gamma violated normality assumption, as errors points are going off the normal line.

For the fitB's, `fitBlog` and `fitBgamma`, we used only the covariates that were individually significant in `fitA` and `fitAgamma`. As a result, the covariates for `fitBlog` and `fitBgamma` are `log(Age)`, `Depressed`, and `Insomniac`.

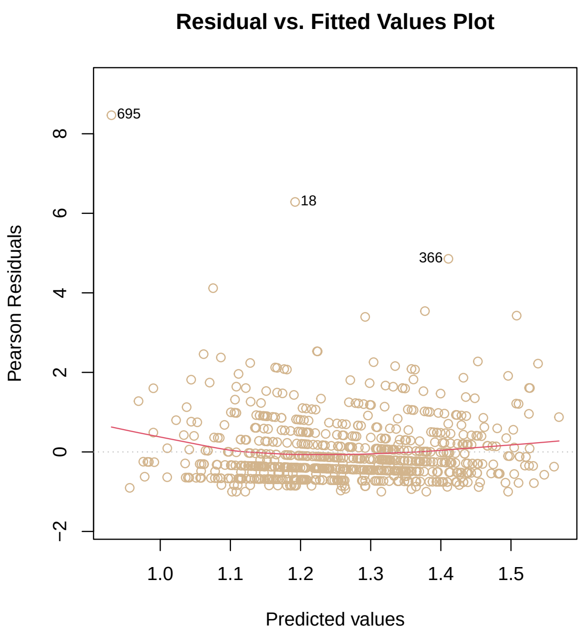
FitBlog:

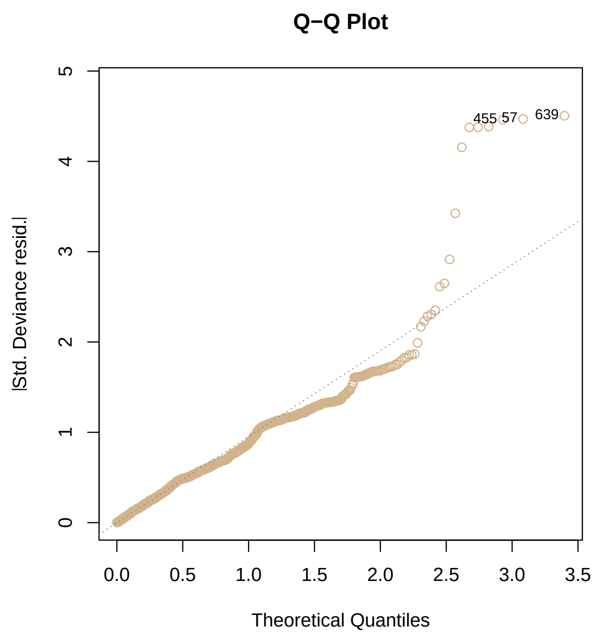
```
##
## Call:
## lm(formula = log(Hours.per.day + 1) ~ log(Age) + Depression +
##     Insomnia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52394 -0.35152 -0.05069  0.32788  2.24160
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.739805   0.174117   9.992  < 2e-16 ***
## log(Age)    -0.169879   0.053340  -3.185  0.00151 **
## Depression   0.014474   0.007194   2.012  0.04458 *
## Insomnia     0.021719   0.007035   3.087  0.00210 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5444 on 731 degrees of freedom
## Multiple R-squared:  0.04236,    Adjusted R-squared:  0.03843
## F-statistic: 10.78 on 3 and 731 DF,  p-value: 6.168e-07
```





FitBgamma:





Analysis:

- For Log: From looking at the Residual Vs Fitted plot for `fitBlog`, it seems that the normality assumption is not violated because the shape of the fit is cloud-like without any noticeable pattern. This means that `fitBlog` does have constant variance. However, there are possible outliers such as point

696 or 19. For the Normal Q-Q plot, normality doesn't seem to be violated because most error points remains on the normality line. Nevertheless, there are still evidences of outliers.

- For Gamma: Since the distribution is Gamma, it is possible to observe clustering of negatively valued residuals in the Residuals vs. Fitted Values plot. There are no obvious patterns in the Residuals vs. Fitted Values plots so `fitBgamma` doesn't seem to violate normality assumption there. However, there are possible outliers. For the Normal Q-Q plot, `fitBgamma` violated normality assumptions, as there are many observations that move away from the line.

Overall Analysis: Evidence from the Residual vs. Fitted plots and Q-Q plots for FitA and FitB indicates that the log model performs better than the gamma model as it doesn't violate normality. Therefore, we've chosen to exclude the gamma model and proceed with analyzing the two log fits.

AIC

AIC(fitAlog) = 1196.1155
AIC(fitBlog) = 1198.0879444

BIC

BIC(fitAlog) = 1237.5143345
BIC(fitBlog) = 1221.0872969

AIC and BIC analysis

`fitAlog` is a better fit than `fitBlog` according to AIC because it has a lower AIC. `fitBlog` is a better fit than `fitAlog` according to BIC because it has a lower BIC.

Since which model is better according to AIC and BIC is different, we can use either fitA or fitB. One isn't clearly better than the other. We can confirm this with anova tests.

ANOVA F-Tests

```r
## ANOVA tests

anova(fitAlog, fitBlog)
```

```
## Analysis of Variance Table
##
## Model 1: log(Hours.per.day + 1) ~ log(Age) + Anxiety + Depression + Insomnia +
##     Music.effects
## Model 2: log(Hours.per.day + 1) ~ log(Age) + Depression + Insomnia
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    727 213.76
## 2    731 216.68 -4   -2.9201 2.4828 0.04254 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis H_0: $\beta_2 = \beta_5 = 0$ vs. H_1: $\beta_2 \neq 0$ or $\beta_2 \neq 0$. We fail to reject the null because the $Pr(> F) = 0.06215 > \alpha = 0.05$. As a result, fitAlog and fitBlog are the same so we will use the smaller model, fitBlog.

```r
anova(fitAgamma, fitBgamma)
```

```
## Analysis of Deviance Table
##
## Model 1: Hours.per.day + 0.001 ~ log(Age) + Anxiety + Depression + Insomnia +
##     Music.effects
## Model 2: Hours.per.day + 0.001 ~ log(Age) + Depression + Insomnia
##   Resid. Df Resid. Dev Df Deviance
## 1       727     485.45
## 2       731     490.17 -4  -4.7243
```

We can't reject the null hypothesis if deviance is negative since we can think of it as though deviance is 0. As a result, `fitAgamma` and `fitBgamma` are the same and we will usually choose the smaller model. However, since `fitAgamma` and `fitBgamma` violate normality assumptions, neither of them will be used for our final model. The negative deviance, which is likely caused by `fitAgamma` and `fitBgamma` violating normality assumptions, confirms what we saw in the QQ-plot.

Interpretation of the final model

Since the gamma distributions violate normality, we would prefer to not use them for our final model. Based on the AIC, BIC, and ANOVA models which show us that `fitAlog` and `fitBlog` are about the same, we will use the smaller model for our final model. As a result, our final model is `fitBlog`, which is: $log(Hours.per.day + 1) = 1.740 - 0.170log(Age) + 0.014Depression + 0.022Insomnia$

According to `fitBlog`: If log(Age) increase by 1 unit then log(Hours.per.day + 1) will decrease by 0.170 units. If the number a person ranks their Depression on a scale of 1 to 10 increases by 1, then log(Hours.per.day + 1) will increase by 0.014 units. If the number a person ranks their Insomnia on a scale of 1 to 10 increases by 1, then log(Hours.per.day + 1) will increase by 0.022 units.

One thing to note for our models is that our R^2's were low, including `fitBlog`, which adjusted $R^2 = 0.038$. We realize this means that our models, including `fitBlog`, explain very little variation in the response (transformed hours per day).

Discussion

In our project, we faced a few minor setbacks. For example, our dataset contained a notable amount of missing values (NA), requiring us to remove rows to ensure accurate calculations for AIC, BIC, and ANOVA. By removing only a small portion of the data, we understand that it could have potentially affected our calculation.

Another difficulty we encountered was the large amount of categorical variables in our dataset. While categorical variables are useful and we did use many in our model, more quantitative variables would be helpful.

Conclusion

Overall, we decide that `fitBlog` is the best model out of the four we made because it does best to capture the relationship between Hour Per Day with Age and Mental Health. From this model, age and mental health seem to have little effect to the hours per day an individual listens to music. This can be seen by the small R^2 and coefficients ($\beta's$).

This project have inspire a possible future study ideas where we utilize the many music genre variables that our dataset and possibly investigate how different music genre affect mental health.

Appendix: R Script

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, fig.align='center')

## Libraries

require('tidyr')
require('dplyr')
require('ggplot2')

## Importing the dataset

data <- read.csv('C:/Users/karen/Documents/STA141A/mxmh_survey_results.csv')
data <- data %>% drop_na(Hours.per.day, Age, Anxiety, Depression, Insomnia, Music.effects)
attach(data)
head(data)

## Plots
# splitting in two for page formatting
ggplot(pivot_longer(data = data, c(2, 28)), aes(value, Hours.per.day)) +
  theme_minimal() + geom_point(color = 'tan') +
  facet_wrap('name', scales = 'free') +
  labs(y = 'Hours per Day', x = 'Value')

ggplot(pivot_longer(data = data, c(2, 28)), aes(value, Hours.per.day)) +
  theme_minimal() + geom_point(color = 'tan') +
  facet_wrap('name', scales = 'free') +
  labs(y = 'Hours per Day', x = 'Value')
# log transform because there are a few high values and many low values
ggplot(pivot_longer(data = data, 2), aes(Hours.per.day, log(Age))) +
  labs(title = 'log(Age) vs. Hours per day') +
  theme_minimal() + geom_point(color = 'tan')

ggplot(data, aes(x = Hours.per.day)) + geom_histogram(bins = 15, color = 'black', fill = 'tan') + labs(
# Hours per day is not normal so we can log transform it or use gamma glm
# log transform because there are a few high values and many low values
ggplot(pivot_longer(data = data, 2), aes(log(Hours.per.day + 1), log(Age))) +
  labs(title = 'log(Age) vs. log(Hours per day + 1)') +
  theme_minimal() + geom_point(color = 'tan')

## Fits

# fitA where we log Hours.per.day + 1 since Hours.per.day is not normally distributed.
fitAlog <- lm(log(Hours.per.day + 1) ~ log(Age) + Anxiety + Depression + Insomnia + Music.effects)

# Residuals vs. Fitted Values Plot and QQ Plot
par(mfrow = c(1,2))

plot(fitAlog, which = 1, caption = '', main = 'Residual vs. Fitted Values Plot', col = 'tan')

plot(fitAlog, which = 2, caption = '', main = 'Q-Q Plot', col = 'tan')

summary(fitAlog)
```

```r
# fitA where it's fitted to a gamma distribution
# For both fitAlog and fitAgamma, we add 1 to Hours.per.day because we can't log zero.
fitAgamma <- glm(Hours.per.day + 0.001 ~ log(Age) + Anxiety + Depression + Insomnia + Music.effects,
                 family = Gamma (link = 'log'))

# Residuals vs. Fitted Values Plot and QQ Plot
par(mfrow = c(1,2))

plot(fitAgamma, which = 1, caption = '', main = 'Residual vs. Fitted Values Plot', col = 'tan')

plot(fitAgamma, which = 2, caption = '', main = 'Q-Q Plot', col = 'tan')

# log Hours.per.day +1
fitBlog <- lm(log(Hours.per.day + 1) ~ log(Age) + Depression + Insomnia)
summary(fitBlog)

# Residuals vs. Fitted Values Plot and QQ Plot
par(mfrow = c(1,2))

plot(fitBlog, which = 1, caption = '', main = 'Residual vs. Fitted Values Plot', col = 'tan')

plot(fitBlog, which = 2, caption = '', main = 'Q-Q Plot', col = 'tan')

# gamma glm
fitBgamma <- glm(Hours.per.day + 0.001 ~ log(Age) + Depression + Insomnia,
                 family = Gamma (link = 'log'))

# Residuals vs. Fitted Values Plot and QQ Plot
par(mfrow = c(1,2))

plot(fitBgamma, which = 1, caption = '', main = 'Residual vs. Fitted Values Plot', col = 'tan')

plot(fitBgamma, which = 2, caption = '', main = 'Q-Q Plot', col = 'tan')

##AIC

AIC(fitAlog)
AIC(fitBlog)

## BIC

BIC(fitAlog)
BIC(fitBlog)

## ANOVA tests

anova(fitAlog, fitBlog)

anova(fitAgamma, fitBgamma)
```