Final Project: Music vs Mental Health Analysis

Calvin, Vy, Karen

December 2, 2023

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 . To do this, we used data collected from a music and mental health survey.

Importing the dataset

```
## Loading required package: tidyr

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: ggplot2

##              Timestamp Age         Primary.streaming.service Hours.per.day
## 1  8/27/2022 19:29:02  18                           Spotify           3.0
## 2  8/27/2022 19:57:31  63                           Pandora           1.5
## 3  8/27/2022 21:28:18  18                           Spotify           4.0
## 4  8/27/2022 21:40:40  61                     YouTube Music           2.5
## 5  8/27/2022 21:54:47  18                           Spotify           4.0
## 6  8/27/2022 21:56:50  18                           Spotify           5.0
## 7  8/27/2022 22:00:29  18                     YouTube Music           3.0
## 8  8/27/2022 22:18:59  21                           Spotify           1.0
## 9  8/27/2022 22:33:05  19                           Spotify           6.0
## 10 8/27/2022 22:44:03  18 I do not use a streaming service.           1.0
##    While.working Instrumentalist Composer        Fav.genre Exploratory
## 1            Yes             Yes      Yes            Latin         Yes
```

```
## 2           Yes              No       No            Rock         Yes
## 3            No              No       No Video game music          No
## 4           Yes              No      Yes            Jazz         Yes
## 5           Yes              No       No             R&B         Yes
## 6           Yes             Yes      Yes            Jazz         Yes
## 7           Yes             Yes       No Video game music         Yes
## 8           Yes              No       No           K pop         Yes
## 9           Yes              No       No            Rock          No
## 10          Yes              No       No             R&B         Yes
##    Foreign.languages BPM Frequency..Classical. Frequency..Country.
## 1                Yes 156                Rarely               Never
## 2                 No 119             Sometimes               Never
## 3                Yes 132                 Never               Never
## 4                Yes  84             Sometimes               Never
## 5                 No 107                 Never               Never
## 6                Yes  86                Rarely           Sometimes
## 7                Yes  66             Sometimes               Never
## 8                Yes  95                 Never               Never
## 9                 No  94                 Never     Very frequently
## 10               Yes 155                Rarely              Rarely
##    Frequency..EDM. Frequency..Folk. Frequency..Gospel. Frequency..Hip.hop.
## 1           Rarely            Never              Never           Sometimes
## 2            Never           Rarely          Sometimes              Rarely
## 3  Very frequently            Never              Never              Rarely
## 4            Never           Rarely          Sometimes               Never
## 5           Rarely            Never             Rarely     Very frequently
## 6            Never            Never              Never           Sometimes
## 7           Rarely        Sometimes             Rarely              Rarely
## 8           Rarely            Never              Never     Very frequently
## 9            Never        Sometimes              Never               Never
## 10          Rarely           Rarely          Sometimes              Rarely
##    Frequency..Jazz. Frequency..K.pop. Frequency..Latin. Frequency..Lofi.
## 1             Never   Very frequently   Very frequently           Rarely
## 2   Very frequently            Rarely         Sometimes           Rarely
## 3            Rarely   Very frequently             Never        Sometimes
## 4   Very frequently         Sometimes   Very frequently        Sometimes
## 5             Never   Very frequently         Sometimes        Sometimes
## 6   Very frequently   Very frequently            Rarely  Very frequently
## 7         Sometimes             Never            Rarely           Rarely
## 8            Rarely   Very frequently             Never        Sometimes
## 9             Never             Never             Never            Never
## 10           Rarely             Never            Rarely           Rarely
##    Frequency..Metal. Frequency..Pop. Frequency..R.B. Frequency..Rap.
## 1              Never Very frequently       Sometimes Very frequently
## 2              Never       Sometimes       Sometimes          Rarely
## 3          Sometimes          Rarely           Never          Rarely
## 4              Never       Sometimes       Sometimes           Never
## 5              Never       Sometimes Very frequently Very frequently
## 6             Rarely Very frequently Very frequently Very frequently
## 7             Rarely          Rarely          Rarely           Never
## 8              Never       Sometimes       Sometimes          Rarely
## 9    Very frequently           Never           Never           Never
## 10             Never       Sometimes       Sometimes          Rarely
##    Frequency..Rock. Frequency..Video.game.music. Anxiety Depression Insomnia
```

```
## 1            Never                   Sometimes       3           0         1
## 2   Very frequently                     Rarely       7           2         2
## 3           Rarely            Very frequently       7           7        10
## 4            Never                      Never       9           7         3
## 5            Never                     Rarely       7           2         5
## 6   Very frequently                      Never       8           8         7
## 7            Never                  Sometimes       4           8         6
## 8            Never                     Rarely       5           3         5
## 9   Very frequently                      Never       2           0         0
## 10       Sometimes                  Sometimes       2           2         5
##    OCD Music.effects   Permissions
## 1    0               I understand.
## 2    1               I understand.
## 3    2     No effect I understand.
## 4    3       Improve I understand.
## 5    9       Improve I understand.
## 6    7       Improve I understand.
## 7    0       Improve I understand.
## 8    3       Improve I understand.
## 9    0       Improve I understand.
## 10   1       Improve I understand.
```

Making categories for Anxious, Depressed, and Insomniac

Since our data had people rate their anxiety, depression, insomnia, and OCD and a scale of 1 to 10 (only integers), we will make categories for `Anxiety`, `Depression`, `Insomnia`, and `OCD` so that the categories will be binary. For example, for `Anxiety`, we will have a category called `Anxious` that is 1 if `Anxiety` >5 and 0 otherwise.

```
##             Timestamp Age         Primary.streaming.service Hours.per.day
## 1  8/27/2022 19:29:02  18                           Spotify           3.0
## 2  8/27/2022 19:57:31  63                           Pandora           1.5
## 3  8/27/2022 21:28:18  18                           Spotify           4.0
## 4  8/27/2022 21:40:40  61                     YouTube Music           2.5
## 5  8/27/2022 21:54:47  18                           Spotify           4.0
## 6  8/27/2022 21:56:50  18                           Spotify           5.0
## 7  8/27/2022 22:00:29  18                     YouTube Music           3.0
## 8  8/27/2022 22:18:59  21                           Spotify           1.0
## 9  8/27/2022 22:33:05  19                           Spotify           6.0
## 10 8/27/2022 22:44:03  18 I do not use a streaming service.           1.0
##    While.working Instrumentalist Composer        Fav.genre Exploratory
## 1            Yes             Yes      Yes            Latin         Yes
## 2            Yes              No       No             Rock         Yes
## 3             No              No       No Video game music          No
## 4            Yes              No      Yes             Jazz         Yes
## 5            Yes              No       No              R&B         Yes
## 6            Yes             Yes      Yes             Jazz         Yes
## 7            Yes             Yes       No Video game music         Yes
## 8            Yes              No       No            K pop         Yes
## 9            Yes              No       No             Rock          No
## 10           Yes              No       No              R&B         Yes
##    Foreign.languages BPM Frequency..Classical. Frequency..Country.
```

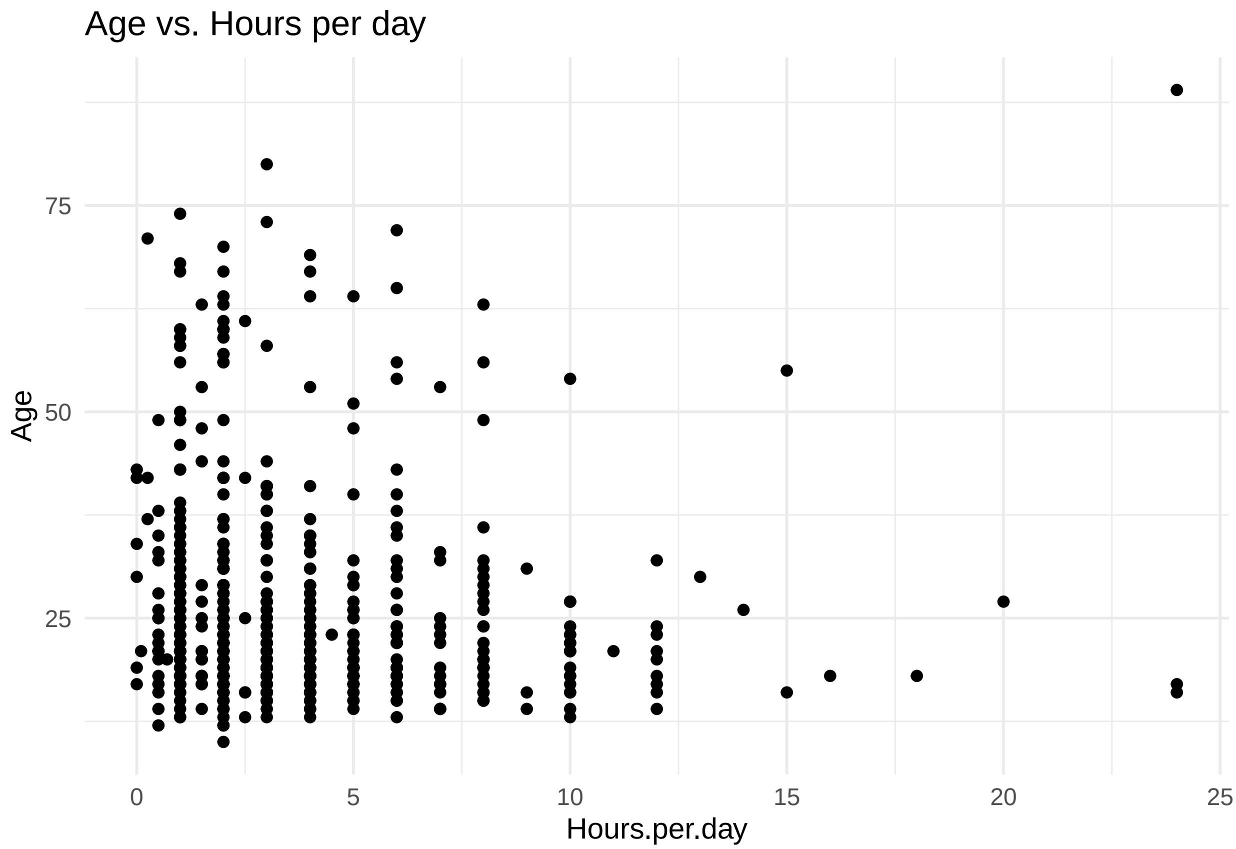
```
## 1               Yes 156              Rarely              Never
## 2                No 119           Sometimes              Never
## 3               Yes 132               Never              Never
## 4               Yes  84           Sometimes              Never
## 5                No 107               Never              Never
## 6               Yes  86              Rarely          Sometimes
## 7               Yes  66           Sometimes              Never
## 8               Yes  95               Never              Never
## 9                No  94               Never    Very frequently
## 10              Yes 155              Rarely             Rarely
##    Frequency..EDM. Frequency..Folk. Frequency..Gospel. Frequency..Hip.hop.
## 1           Rarely            Never              Never           Sometimes
## 2            Never           Rarely          Sometimes              Rarely
## 3  Very frequently            Never              Never              Rarely
## 4            Never           Rarely          Sometimes               Never
## 5           Rarely            Never             Rarely     Very frequently
## 6            Never            Never              Never           Sometimes
## 7           Rarely        Sometimes             Rarely              Rarely
## 8           Rarely            Never              Never     Very frequently
## 9            Never        Sometimes              Never               Never
## 10          Rarely           Rarely          Sometimes              Rarely
##    Frequency..Jazz. Frequency..K.pop. Frequency..Latin. Frequency..Lofi.
## 1             Never   Very frequently   Very frequently           Rarely
## 2   Very frequently            Rarely         Sometimes           Rarely
## 3            Rarely   Very frequently             Never        Sometimes
## 4   Very frequently         Sometimes   Very frequently        Sometimes
## 5             Never   Very frequently         Sometimes        Sometimes
## 6   Very frequently   Very frequently            Rarely  Very frequently
## 7         Sometimes             Never            Rarely           Rarely
## 8            Rarely   Very frequently             Never        Sometimes
## 9             Never             Never             Never            Never
## 10           Rarely             Never            Rarely           Rarely
##    Frequency..Metal. Frequency..Pop. Frequency..R.B. Frequency..Rap.
## 1              Never Very frequently       Sometimes Very frequently
## 2              Never       Sometimes       Sometimes          Rarely
## 3          Sometimes          Rarely           Never          Rarely
## 4              Never       Sometimes       Sometimes           Never
## 5              Never       Sometimes Very frequently Very frequently
## 6             Rarely Very frequently Very frequently Very frequently
## 7             Rarely          Rarely          Rarely           Never
## 8              Never       Sometimes       Sometimes          Rarely
## 9    Very frequently           Never           Never           Never
## 10             Never       Sometimes       Sometimes          Rarely
##    Frequency..Rock. Frequency..Video.game.music. Anxiety Depression Insomnia
## 1             Never                    Sometimes       3          0        1
## 2   Very frequently                       Rarely       7          2        2
## 3            Rarely              Very frequently       7          7       10
## 4             Never                        Never       9          7        3
## 5             Never                       Rarely       7          2        5
## 6   Very frequently                        Never       8          8        7
## 7             Never                    Sometimes       4          8        6
## 8             Never                       Rarely       5          3        5
## 9   Very frequently                        Never       2          0        0
## 10        Sometimes                    Sometimes       2          2        5
```

```
##    OCD Music.effects   Permissions Anxious Depressed Insomniac
## 1    0               I understand.       0         0         0
## 2    1               I understand.       1         0         0
## 3    2     No effect I understand.       1         1         1
## 4    3       Improve I understand.       1         1         0
## 5    9       Improve I understand.       1         0         0
## 6    7       Improve I understand.       1         1         1
## 7    0       Improve I understand.       0         1         1
## 8    3       Improve I understand.       0         0         0
## 9    0       Improve I understand.       0         0         0
## 10   1       Improve I understand.       0         0         0
```

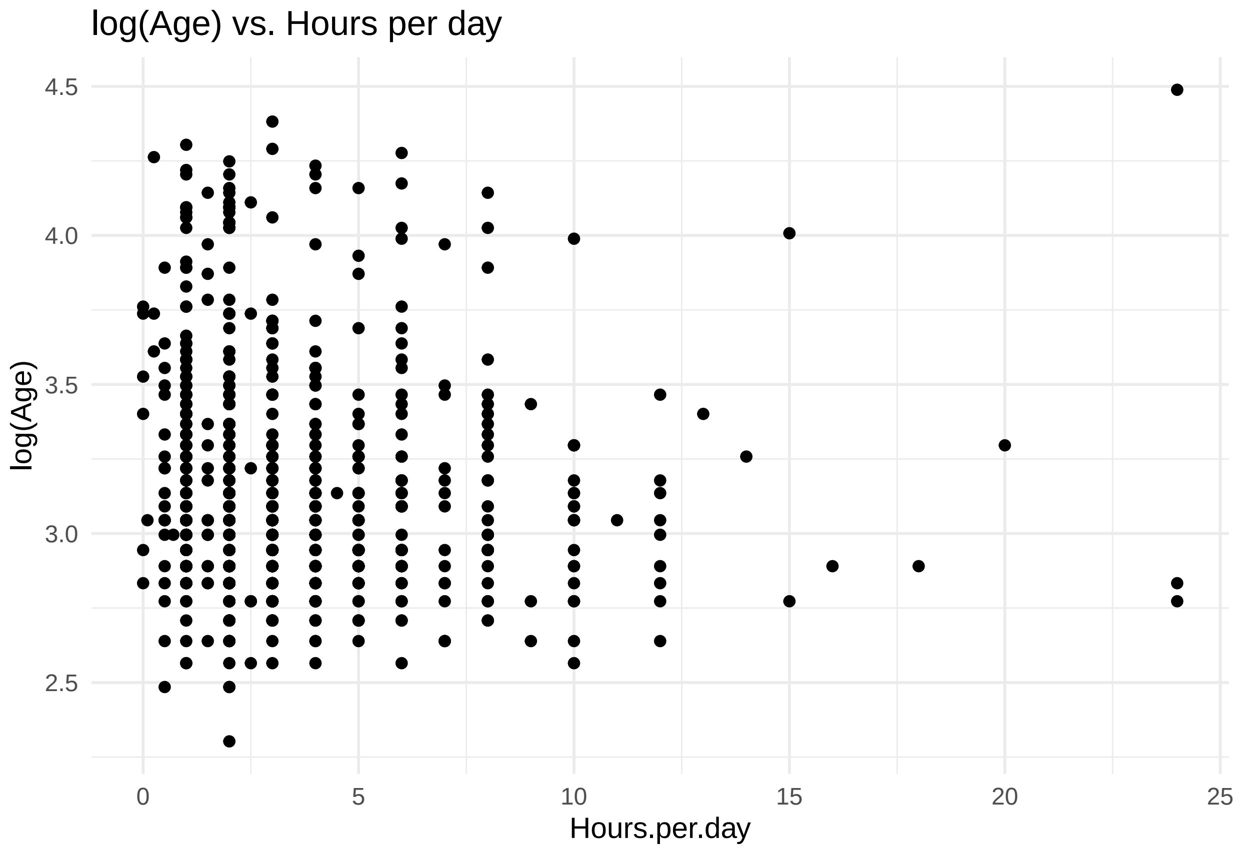
Testing Non-linearity

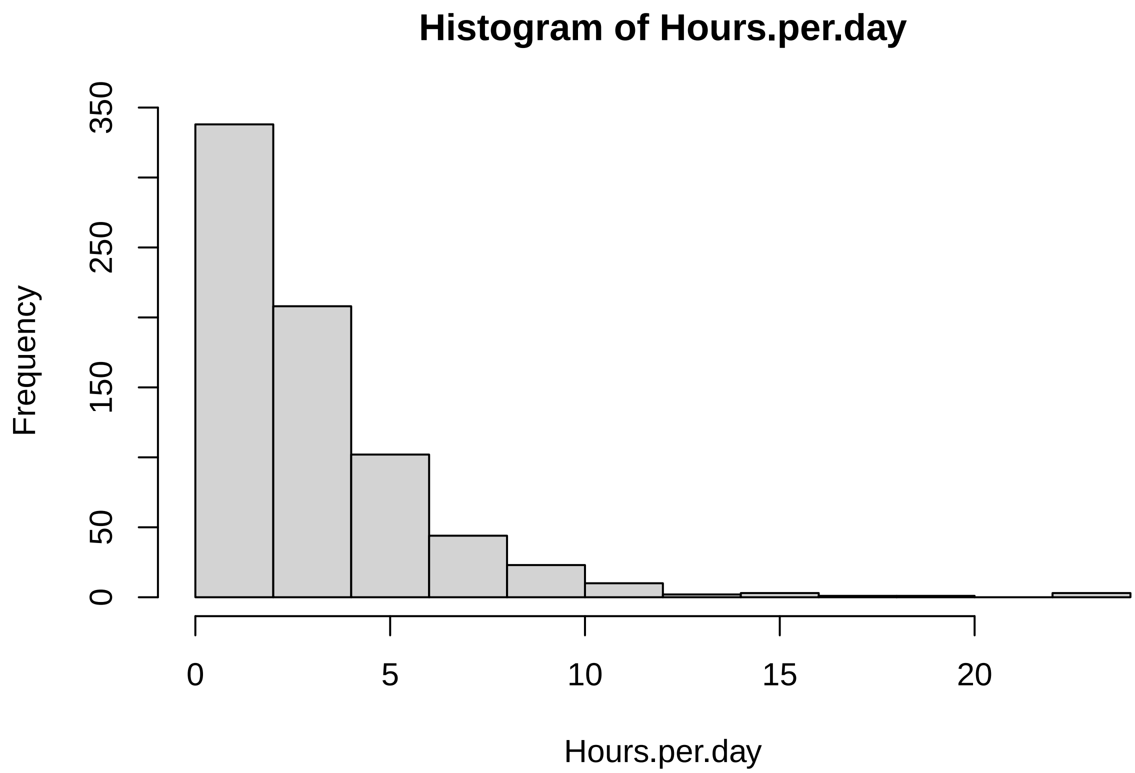
We don't need to test for non-linearity for categorical variables so we will only test for non-linearity for the continuous variables, namely age.

Plots



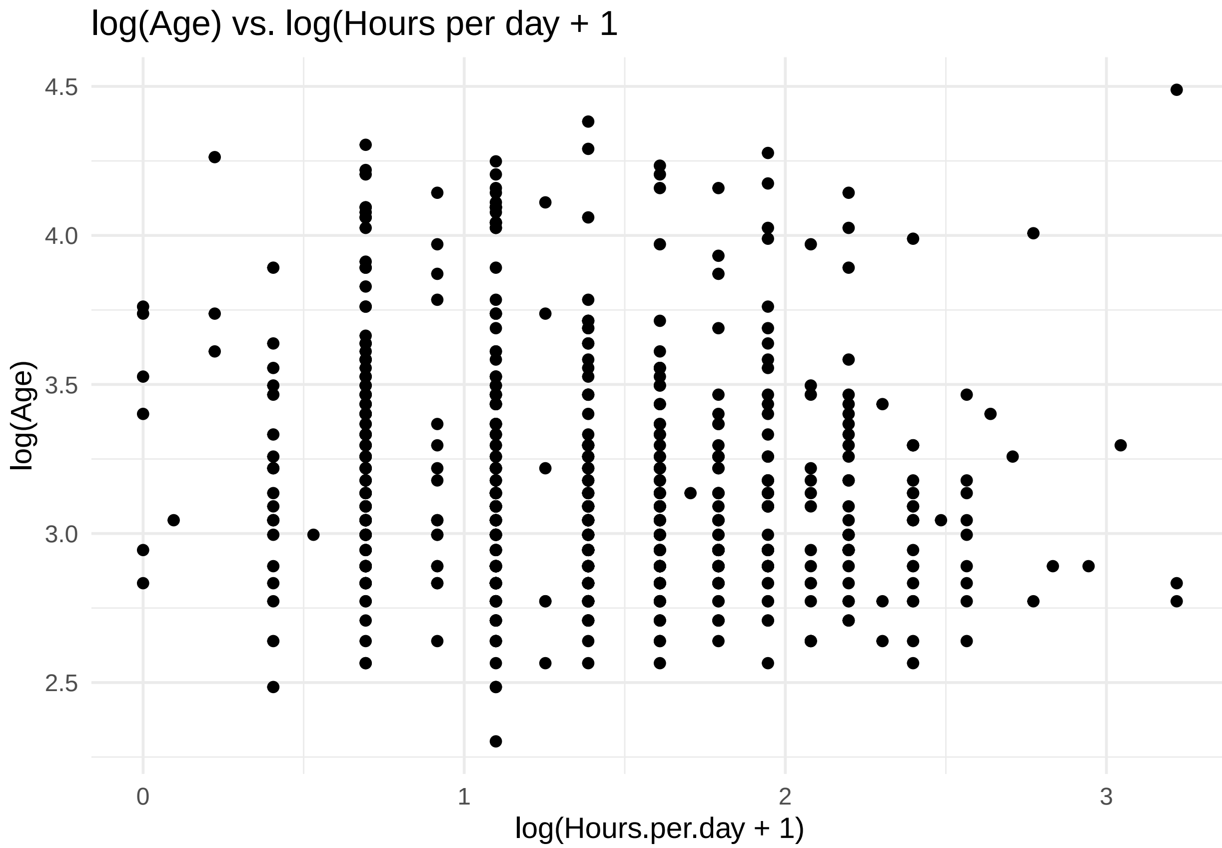
There doesn't seem to be a linear relationship between age and hours per day. We can transform age by logging it so that the relationship looks less non-linear. With log age:


log(Age) vs. Hours per day
log(Age)
4.5
4.0
3.5
3.0
2.5
0
5
10
15
20
25
Hours.per.day



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 should look also look to see if log(Age) and log(Hours.per.day + 1).

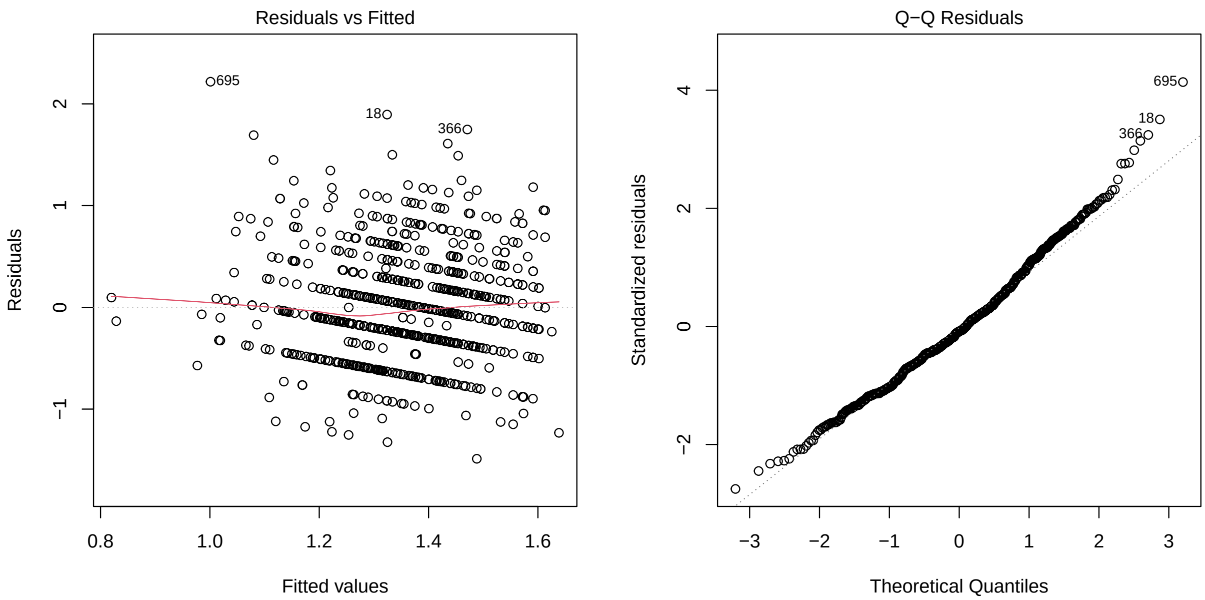


The relationship between log(Age) and log(Hours.per.day + 1) is not obviously non-linear, so we can use log(Age) in the linear model with log(Hours.per.day + 1).

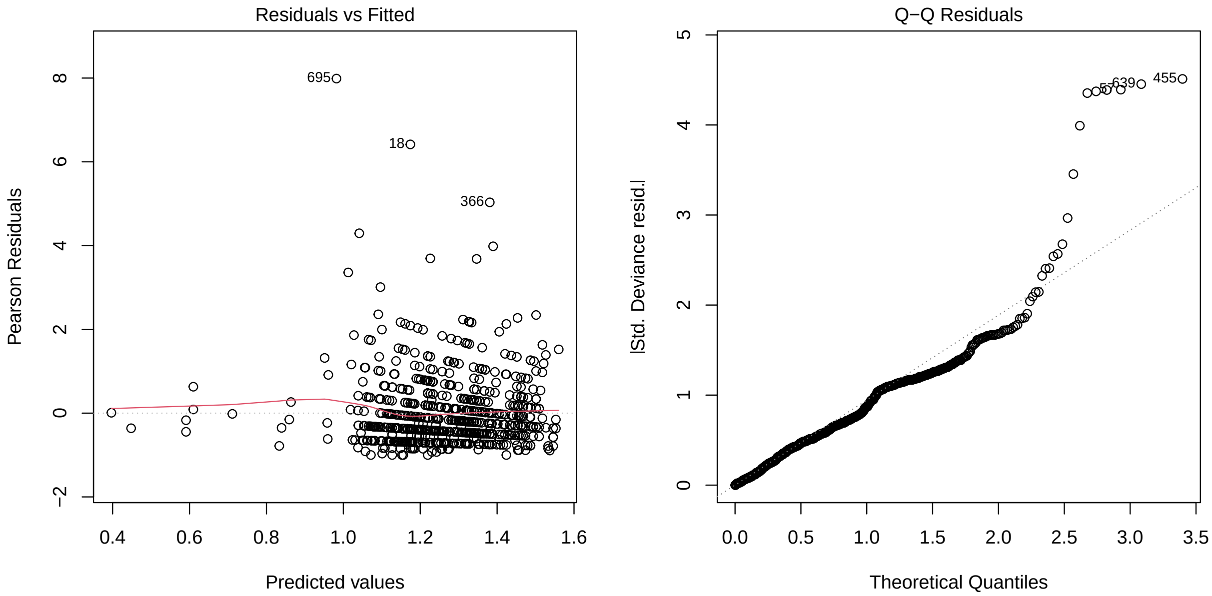
Fitted models

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`.

```
##
## Call:
## lm(formula = log(Hours.per.day + 1) ~ log(Age) + Anxious + Depressed +
##     Insomniac + Music.effects)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48822 -0.35553 -0.04605  0.33077  2.21770
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.51836    0.26176   5.801 9.85e-09 ***
## log(Age)               -0.16400    0.05365  -3.057  0.00232 **
## Anxious                -0.01895    0.04460  -0.425  0.67113
## Depressed               0.13938    0.04408   3.162  0.00163 **
## Insomniac               0.11785    0.04492   2.624  0.00888 **
## Music.effectsImprove    0.28938    0.19371   1.494  0.13565
## Music.effectsNo effect  0.21897    0.19647   1.115  0.26543
## Music.effectsWorsen    -0.02302    0.23368  -0.099  0.92156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5422 on 727 degrees of freedom
## Multiple R-squared:  0.05533,    Adjusted R-squared:  0.04623
## F-statistic: 6.083 on 7 and 727 DF,  p-value: 6.435e-07
```




```
##
## Call:
## glm(formula = Hours.per.day + 0.001 ~ log(Age) + Anxious + Depressed +
##     Insomniac + Music.effects, family = Gamma(link = "log"))
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.96570    0.41641   2.319   0.0207 *
## log(Age)                -0.12309    0.08535  -1.442   0.1497
## Anxious                 -0.05890    0.07096  -0.830   0.4068
## Depressed                0.17935    0.07013   2.558   0.0107 *
## Insomniac                0.14014    0.07146   1.961   0.0502 .
## Music.effectsImprove     0.61635    0.30816   2.000   0.0459 *
## Music.effectsNo effect   0.56906    0.31255   1.821   0.0691 .
## Music.effectsWorsen      0.30143    0.37173   0.811   0.4177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.7440657)
##
##     Null deviance: 503.18  on 734  degrees of freedom
## Residual deviance: 486.59  on 727  degrees of freedom
## AIC: 3250.1
##
## Number of Fisher Scoring iterations: 7
```



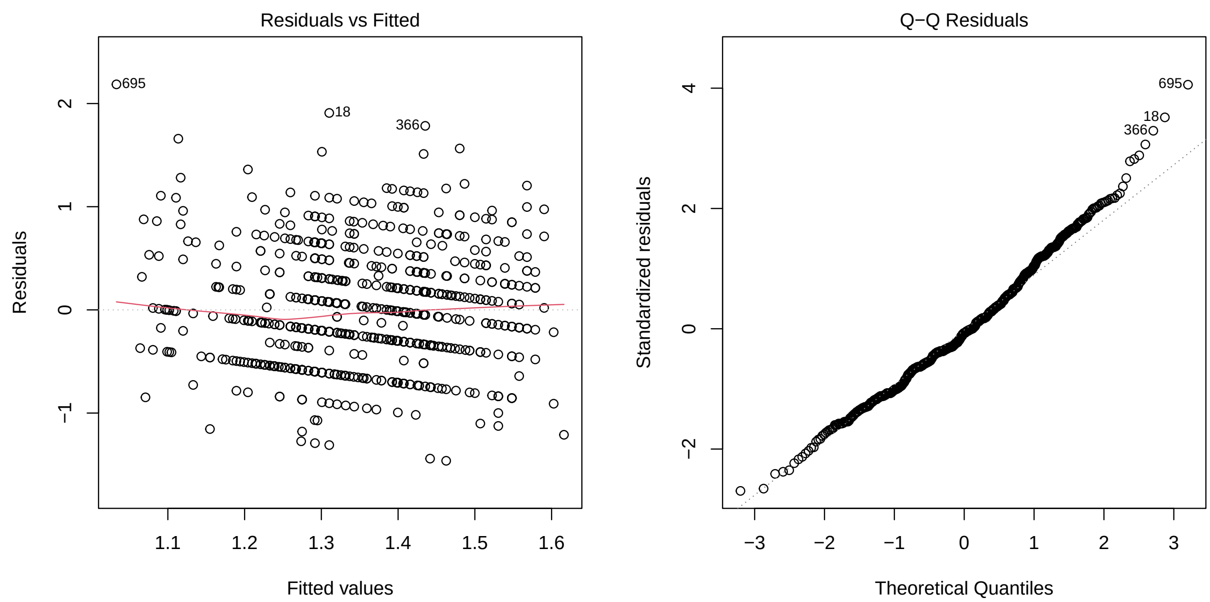
Analysis:

- For Log: From looking at the Residual Vs Fitted plot for fitA_log, 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 fitA_log does have constant variance. However, there are possible outliers such as point 695 or 18. 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 Residuals and Fitted. This means that fitA_gamma doesn't seem to violate normality assumption. In another word, fitA_gamma have a constant Variance. However, there are possible outliers such as points 695 or 18. For the Normal Q-Q plot, Gamma violated normality assumption, as errors points are going off the normal line.

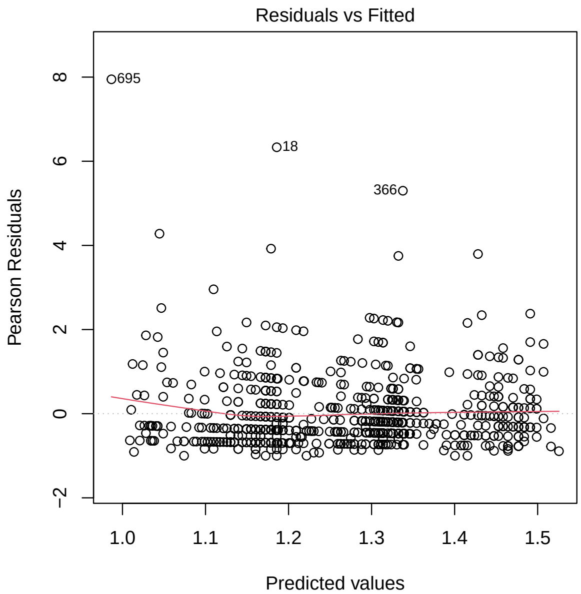
For the fitB's, `fitBlog` and `fitBgamma`, we used only the covariates that were individually significant in `fitA` and `fitAgamma`. As a result, the covariates for `fitBlog` and `fitBgamma` are `log(Age)`, `Depressed`, and `Insomniac`.

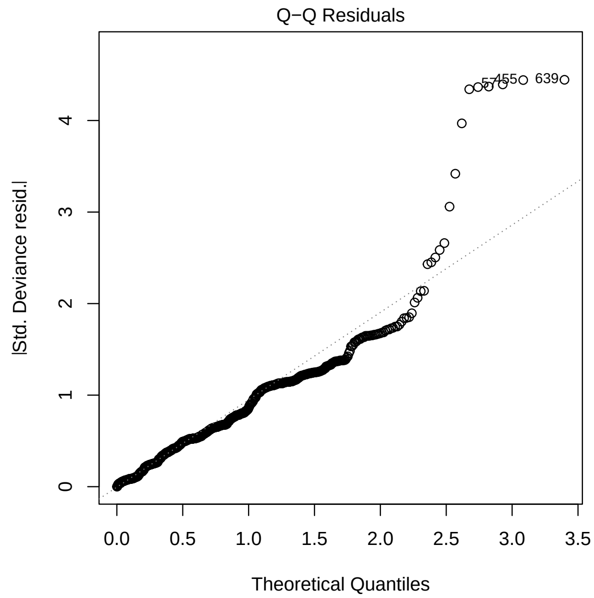
```
##
## Call:
## lm(formula = log(Hours.per.day + 1) ~ log(Age) + Depressed +
##     Insomniac)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46264 -0.34592 -0.03795  0.32323  2.18588
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.78490    0.17085  10.447  < 2e-16 ***
## log(Age)    -0.16751    0.05327  -3.145  0.00173 **
```

```
## Depressed    0.13257    0.04175   3.175  0.00156 **
## Insomniac    0.11492    0.04479   2.566  0.01050 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5441 on 731 degrees of freedom
## Multiple R-squared:  0.04363,    Adjusted R-squared:  0.0397
## F-statistic: 11.12 on 3 and 731 DF,  p-value: 3.851e-07
```




```
##
## Call:
## glm(formula = Hours.per.day + 0.001 ~ log(Age) + Depressed +
##     Insomniac, family = Gamma(link = "log"))
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52650    0.27113   5.630 2.57e-08 ***
## log(Age)    -0.12026    0.08453  -1.423   0.1553
## Depressed    0.15340    0.06626   2.315   0.0209 *
## Insomniac    0.14463    0.07108   2.035   0.0422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.7455182)
##
##     Null deviance: 503.18  on 734  degrees of freedom
## Residual deviance: 490.98  on 731  degrees of freedom
## AIC: 3249.4
##
## Number of Fisher Scoring iterations: 6
```





Analysis:

- For Log: From looking at the Residual Vs Fitted plot for fitB_log, 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 fitB_log does have constant variance. However, there are possible outliers such as point 695 or 18. 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 Residuals and Fitted. This means that fitB_gamma doesn't seem to violate normality assumption. In other word, fitB_gamma have a constant Variance. However, there are possible outliers such as points 695 or 18. For the Normal Q-Q plot, Gamma violated normality assumption, as errors points are going off the normal line.

AIC

AIC(fitAlog) = 1196.0682801
AIC(fitBlog) = 1197.1126943

AIC(fitAgamma) = 3250.087381
AIC(fitBgamma) = 3249.3800894

BIC

BIC(fitAlog) = 1237.4671146
BIC(fitBlog) = 1220.1120468

BIC(fitAgamma) = 3291.4862155
BIC(fitBgamma) = 3272.3794419

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.

fitAgamma is a better fit than fitBgamma according to AIC because it has a lower AIC. fitBgamma is a better fit than fitAgamma 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) + Anxious + Depressed + Insomniac +
##     Music.effects
## Model 2: log(Hours.per.day + 1) ~ log(Age) + Depressed + Insomniac
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    727 213.75
## 2    731 216.40 -4   -2.6465 2.2503 0.06215 .
## ---
## 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) + Anxious + Depressed + Insomniac +
##     Music.effects
## Model 2: Hours.per.day + 0.001 ~ log(Age) + Depressed + Insomniac
##   Resid. Df Resid. Dev Df Deviance
## 1       727     486.59
## 2       731     490.98 -4  -4.3822
```

The deviance is negative because fitAgamma and fitBgamma violated normality assumptions.

Appendix: R Script

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, fig.align='center')
#knitr::opts_chunk$set(fig.width = 12, fig.height = 6)
require('tidyr')
require('dplyr')
require('ggplot2')
data <- read.csv('mxmh_survey_results.csv')
data <- data %>% drop_na(Hours.per.day, Age, Anxiety, Depression, Insomnia, Music.effects)
head(data, 10)
# mark people as anxious (1) if anxiety > 5, not anxious (0) otherwise
data$Anxious = ifelse(data$Anxiety > 5, 1, 0)
# mark people as anxious (1) if anxiety > 5, not anxious (0) otherwise
data$Depressed = ifelse(data$Depression > 5, 1, 0)
# mark people as anxious (1) if anxiety > 5, not anxious (0) otherwise
data$Insomniac = ifelse(data$Insomnia > 5, 1, 0)

head(data, 10)
attach(data)


ggplot(pivot_longer(data = data, 2), aes(Hours.per.day, Age)) +
  labs(title = 'Age vs. Hours per day') +
  theme_minimal() + geom_point()
# 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 pe
    theme_minimal() + geom_point()
hist(Hours.per.day)
# 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
    theme_minimal() + geom_point()
## 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) + Anxious + Depressed + Insomniac + Music.effects)
summary(fitAlog)

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


# 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) + Anxious + Depressed + Insomniac + Music.effects,
                 family = Gamma (link = 'log'))
summary(fitAgamma)

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

# log Hours.per.day + 1
```

```r
fitBlog <- lm(log(Hours.per.day + 1) ~ log(Age) + Depressed + Insomniac)
summary(fitBlog)

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

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

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

##AIC

AIC(fitAlog)
AIC(fitBlog)

AIC(fitAgamma)
AIC(fitBgamma)

## BIC

BIC(fitAlog)
BIC(fitBlog)

BIC(fitAgamma)
BIC(fitBgamma)

## ANOVA tests

anova(fitAlog, fitBlog)

anova(fitAgamma, fitBgamma)
```