LSC541 Module 4: Assignment

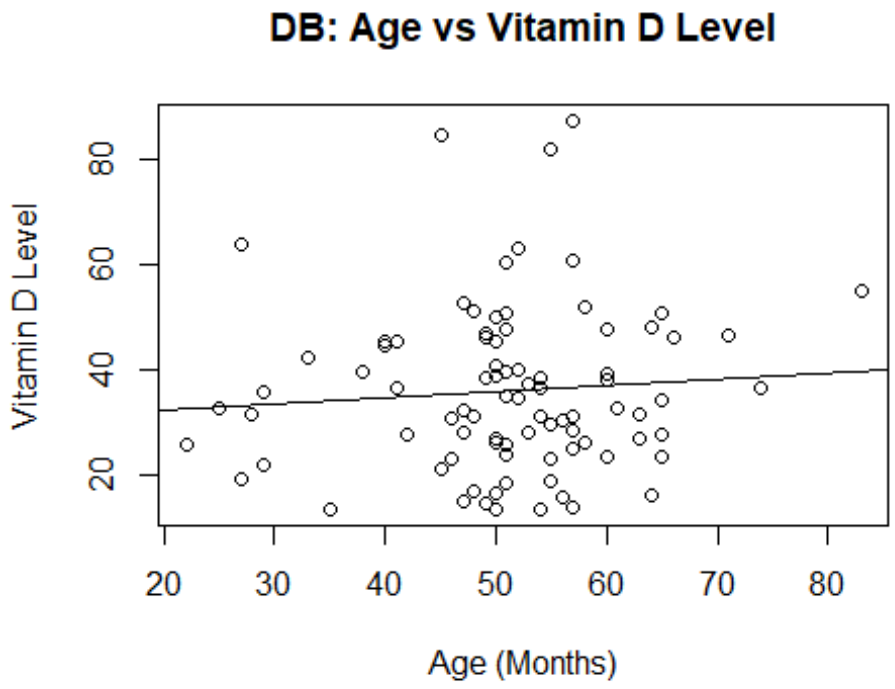
Daniel Bihnam

Read in Data

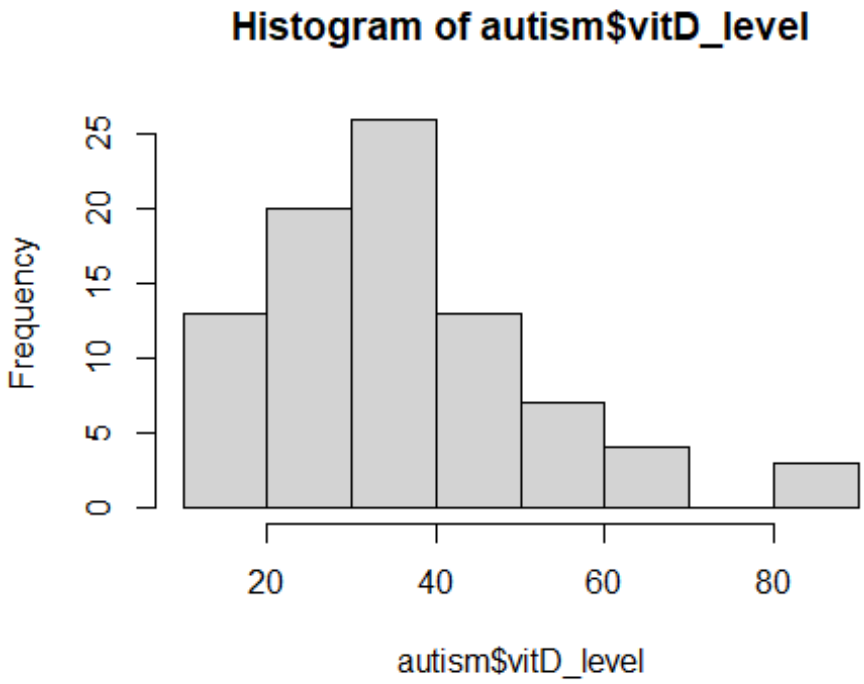
```
autism0 <- read.table('C:/Users/Daniel/Downloads/data1_LSC598.txt', header =
T)
autism <- na.omit(autism0)
#Reading in data and omitting a N/A value from the data
```

Vitamin D Levels vs Age

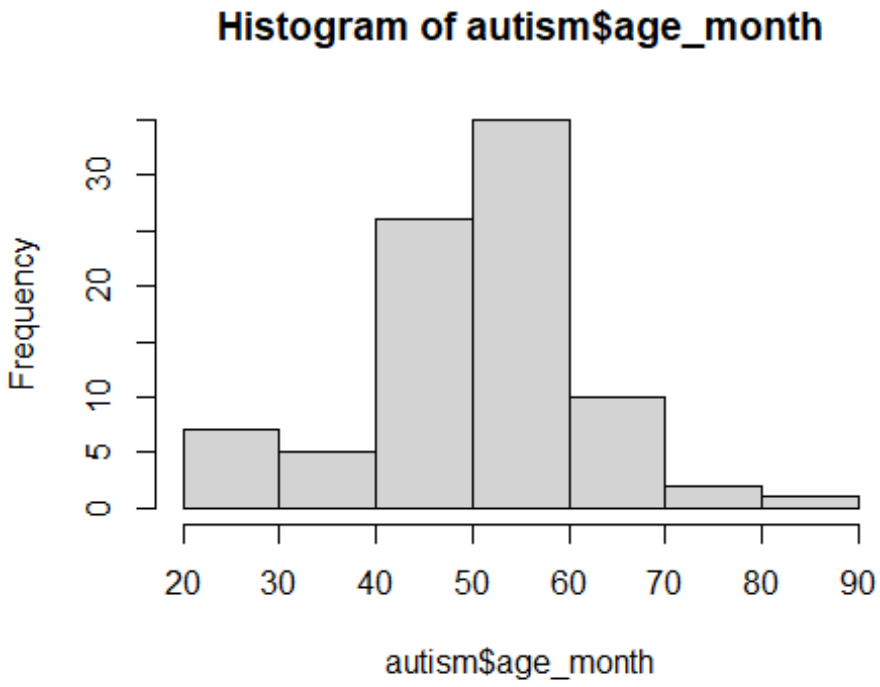
```
reg1 <- lm(vitD_level~age_month,data=autism)
plot(autism$age_month,autism$vitD_level,
     main='DB: Age vs Vitamin D Level',
     xlab='Age (Months)',
     ylab='Vitamin D Level')
abline(reg1)
```




```
#Regression without transformation

hist(autism$vitD_level)
```



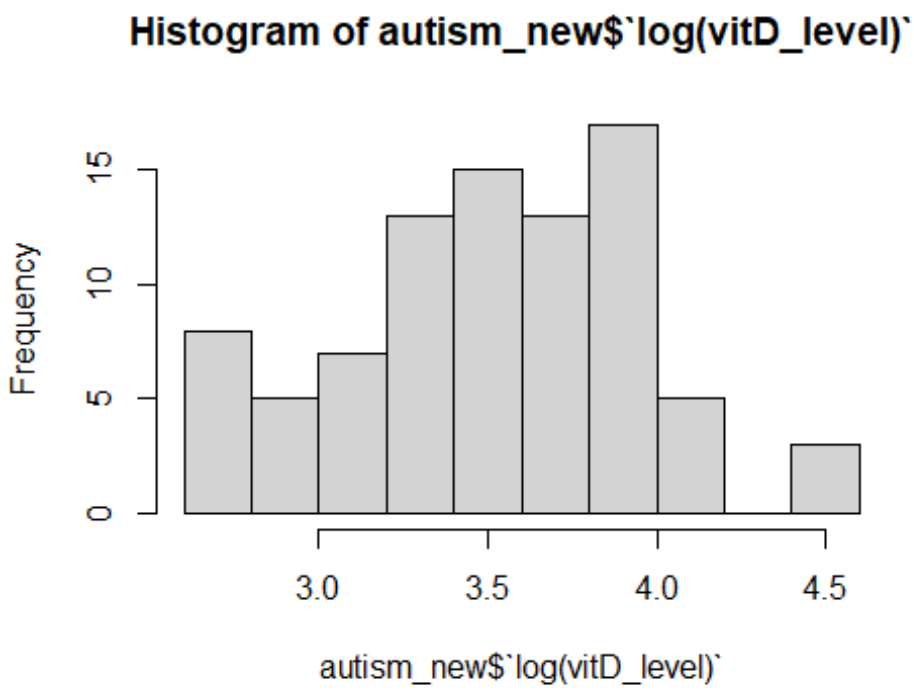

```r
hist(autism$age_month)
```



```r
#Checking normality of variables

library(dplyr)
```

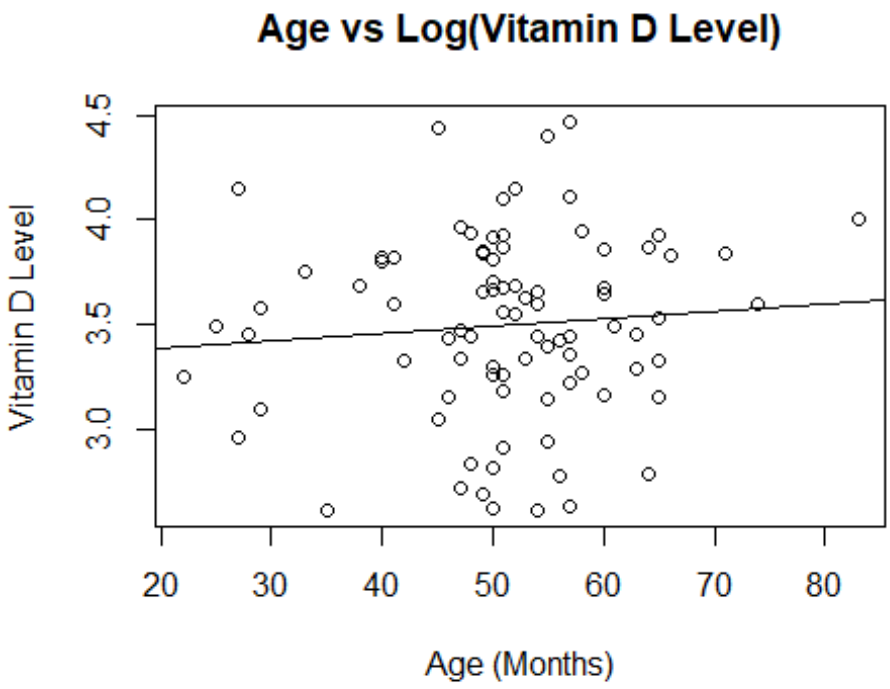
```
Warning: package 'dplyr' was built under R version 4.3.1


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
```

```r
autism_new <- mutate(autism,log(vitD_level))
hist(autism_new$`log(vitD_level)`)
```




```r
#Mutate vitD_level with log() function to normalize

reg2 <- lm(`log(vitD_level)`~age_month,data=autism_new)
plot(autism_new$age_month,autism_new$`log(vitD_level)`,
     main='Age vs Log(Vitamin D Level)',
     xlab='Age (Months)',
```

```
     ylab='Vitamin D Level')
abline(reg2)
```



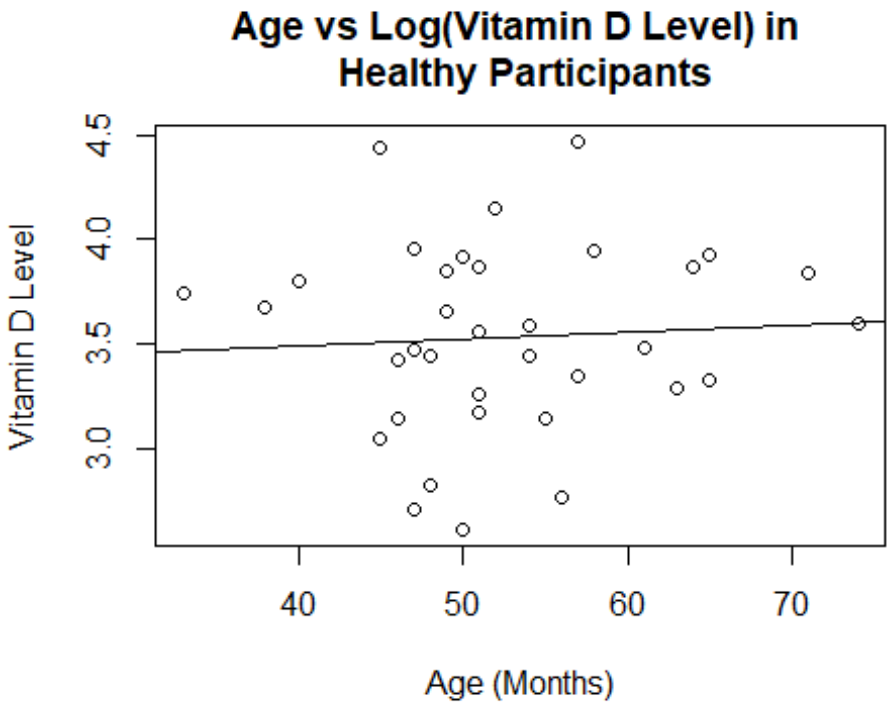

```
#Regression with transformation

summary(reg1)


Call:
lm(formula = vitD_level ~ age_month, data = autism)

Residuals:
    Min      1Q  Median      3Q     Max
-22.862 -10.545  -1.856   9.944  50.538

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.9566     8.0858   3.705 0.000378 ***
age_month     0.1176     0.1549   0.760 0.449622
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.61 on 84 degrees of freedom
Multiple R-squared:  0.006822,  Adjusted R-squared:  -0.005001
F-statistic: 0.577 on 1 and 84 DF,  p-value: 0.4496
```

```
#Statistics of regression without transformation
summary(reg2)


Call:
lm(formula = `log(vitD_level)` ~ age_month, data = autism_new)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89196 -0.24629  0.03517  0.33119  0.96770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.311294   0.225779  14.666   <2e-16 ***
age_month   0.003532   0.004324   0.817    0.416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4359 on 84 degrees of freedom
Multiple R-squared:  0.00788,   Adjusted R-squared:  -0.00393
F-statistic: 0.6672 on 1 and 84 DF,  p-value: 0.4163

#Statistics of regression with transformation
```

For the comparison of age in months and vitamin D levels, I decided to first create a regression model without any transformations to visualize the output and linear equation coefficients. I was not convinced that the data met all the assumptions needed for a linear regression on the basis of the data not being normally distributed. I plotted a histogram of both my x (age) and y (vitD) variables to assess their normality. Upon inspection, the vitamin D level variable is right-skewed. I then mutated my dataset by applying a log() function to this variable, and it normalized my data. I then checked the summary statistics of both regressions, and I found the mutated regression to be a better model. Moving forward with the transformed regression, the y-intercept is equal to 3.311, so we expect a baseline log(vitamin D level) of 3.311 based on our model. The slope reported was 0.0035, so as age in months increases by 1, the log(vitamin D level) is expected to increase by 0.0035. The P-value reported was 0.416, which is too large to show statistical significance between this data. This is much larger than the widely expected confidence level of 0.05 (95% confidence).

Age vs Vitamin D Level (Healthy)

```
autism_healthy <- autism_new %>% filter(group==0)
#Create a new dataframe of just healthy participants

reg3 <- lm(`log(vitD_level)`~age_month,data=autism_healthy)
plot(autism_healthy$age_month,autism_healthy$`log(vitD_level)`,
     main='Age vs Log(Vitamin D Level) in \n Healthy Participants',
     xlab='Age (Months)',
     ylab='Vitamin D Level')
abline(reg3)
```



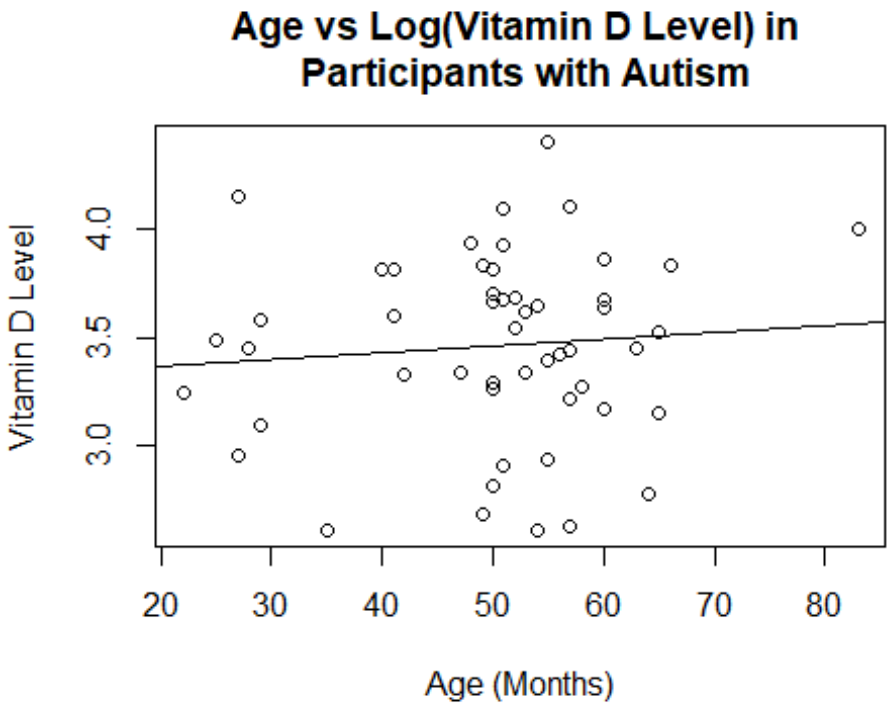

```
#Regression with log() transformation

summary(reg3)


Call:
lm(formula = `log(vitD_level)` ~ age_month, data = autism_healthy)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91247 -0.27950 -0.00827  0.31535  0.92441

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.366444   0.468251   7.189 3.06e-08 ***
age_month   0.003269   0.008797   0.372    0.713
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4526 on 33 degrees of freedom
Multiple R-squared:  0.004166,  Adjusted R-squared:  -0.02601
F-statistic: 0.1381 on 1 and 33 DF,  p-value: 0.7126

#Statistics of regression with transformation
```

In this test, we analyzed age vs vitamin D level in healthy participants. Based on the results of the previous test, as well as the fact that the sample size is shrinking due to the condition

of only healthy participants, we are going to keep using the transformed model to maintain normality. The y-intercept is equal to 3.366, so we expect a baseline log(vitamin D level) of 3.366 based on our model. The slope reported was 0.0033, so as age in months increases by 1, the log(vitamin D level) is expected to increase by 0.0033. The P-value reported was 0.713, which is too large to show statistical significance between this data. This is much larger than the widely expected confidence level of 0.05 (95% confidence).

Age vs Vitamin D Level (Autism)

```
autism_autism <- autism_new %>% filter(group==1)
#Create a new dataframe of just participants with autism

reg4 <- lm(`log(vitD_level)`~age_month,data=autism_autism)
plot(autism_autism$age_month,autism_autism$`log(vitD_level)`,
     main='Age vs Log(Vitamin D Level) in \n Participants with Autism',
     xlab='Age (Months)',
     ylab='Vitamin D Level')
abline(reg4)
```




```
#Regression with log()

summary(reg4)


Call:
lm(formula = `log(vitD_level)` ~ age_month, data = autism_autism)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.86211 -0.24099  0.06288  0.28142  0.92895 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.302078   0.258156  12.791   <2e-16 ***
age_month   0.003150   0.005012   0.628    0.533    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4309 on 49 degrees of freedom
Multiple R-squared:  0.007996,  Adjusted R-squared:  -0.01225 
F-statistic: 0.395 on 1 and 49 DF,  p-value: 0.5326

#Statistics of regression with transformation
```

In this test, we analyzed age vs vitamin D level in participants with autism. We again used the transformed model here to satisfy the assumption of normality in our model. The y-intercept is equal to 3.302, so we expect a baseline log(vitamin D level) of 3.302 based on our model. The slope reported was 0.0035, so as age in months increases by 1, the log(vitamin D level) is expected to increase by 0.0035. The P-value reported was 0.533, which is too large to show statistical significance between this data. This is much larger than the widely expected confidence level of 0.05 (95% confidence).

Upon analyzing both groups separately, there appears to be no statistical difference between the vitamin D levels of patients with or without autism as they age. We fail to reject the null hypothesis that there is no difference in vitamin D levels of participants with and without autism.