# Greg Finkelberg: HW04

## R packages

```
library(tidyverse)
library(quarto)
library(ggplot2)
```

## Linear regression (vitamin-D level ~ age)

```
data1_LSC598 <- read.csv("C:/Users/Greg/Downloads/data1_LSC598.txt", sep="")

df <- data1_LSC598 %>% mutate(autism = case_when(group == '0' ~ 'no', group
== '1' ~ "yes"))

vitD <- df %>% drop_na(vitD_level)
```

First, we must check the assumptions of a linear model:

1. As x (age) varies, the y (vitamin-D level) values follow a straight line.
2. The amount of vertical spread is approximately the same in each strip, except perhaps near the ends.
3. Data is normally distributed, or sample size is greater than 30.

```
cor(vitD$age_month, vitD$vitD_level)

[1] 0.08259572

vitD %>% ggplot(aes(x = age_month, y=vitD_level)) + geom_point() +
  labs(title = "Relationship Between Age and Vitamin-D Level", x = "age
  (months)", y = "Vitamin-D level") + stat_smooth(method = "lm", se =
  FALSE,  formula = "y~x")
```

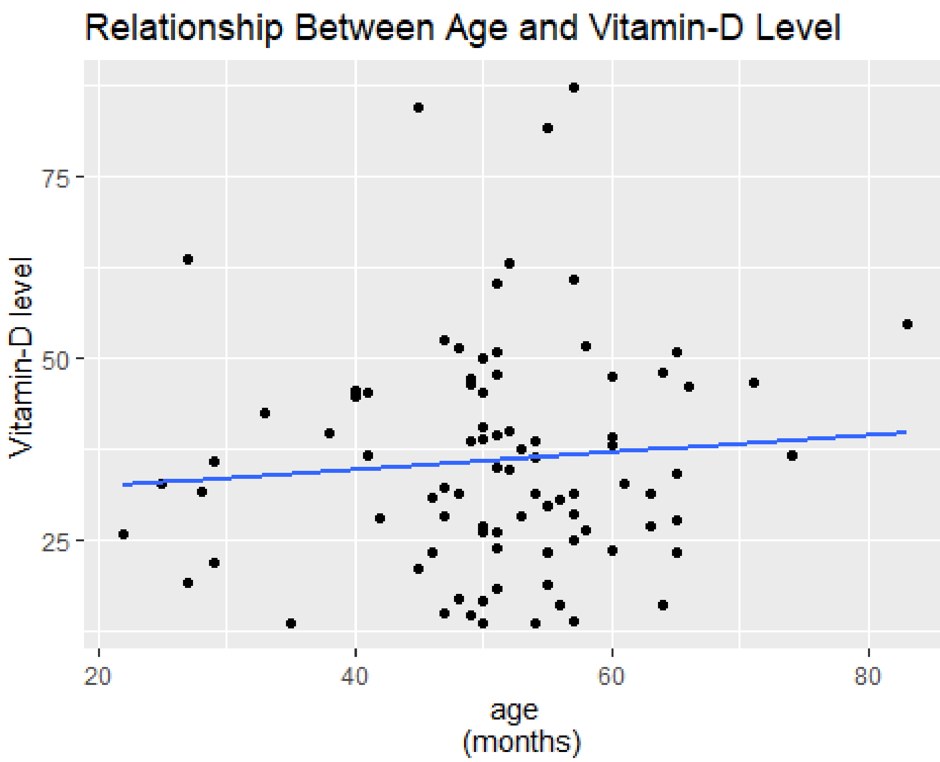


Figure 1.

1. The correlation of x and y is 0.08, which is very weak. Figure 1 shows the data does not appear to be linear in nature.

2. Figure 1 shows that the data does not have equal spread.

3. The sample size is 86, therefore this assumption is met.

I tried a few different transformations to see if I can fit the data into a better, non-linear model.

```
model <- vitD %>% ggplot(aes(x = age_month, y = vitD_level)) + geom_point() +
labs(title = "Relationship Between Age and Vitamin-D Level", x = "Age
(months)", y = "Vitamin-D level")

model +
    geom_smooth(method = "gam", formula = y ~ poly(x, 2), se = FALSE,
    color = "blue") +
    geom_smooth(method = "gam", formula = y ~ poly(x, 3), se = FALSE,
    color = "red") +
    geom_smooth(method = "gam", formula = y ~ log(x), se = FALSE,
    color = "orange")
```

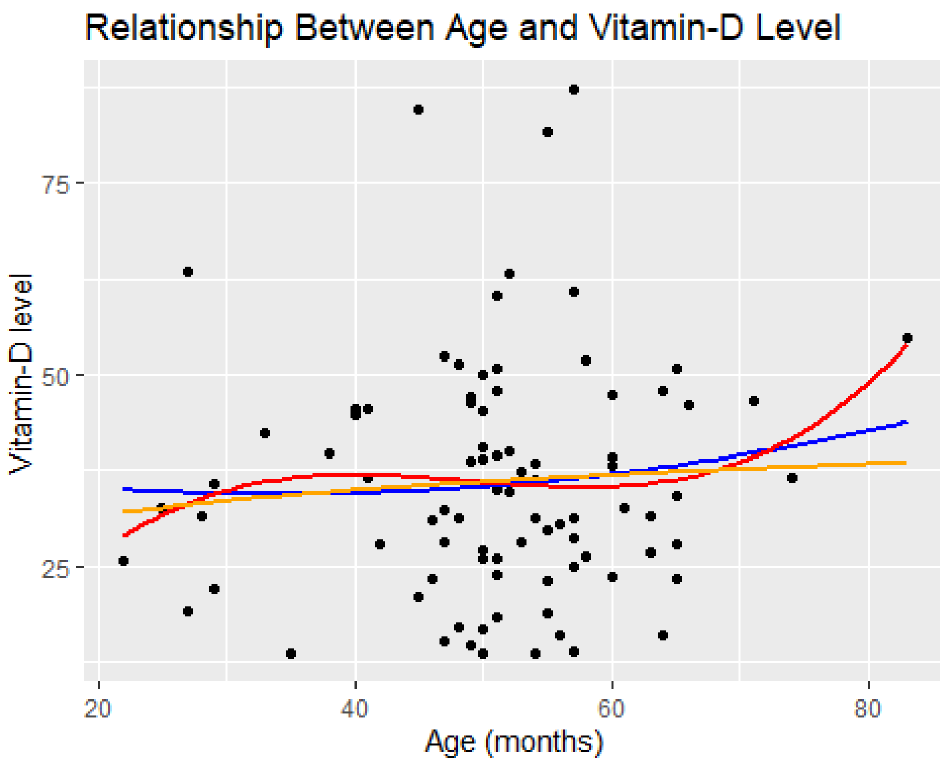


Figure 2.

It appears the polynomial cubed function might fit this relationship best, so I ran a hypothesis test based on it. The significance level was 0.05.

Null hypothesis - $\beta 3 = 0$
Alternative hypothesis - $\beta 3 \neq 0$

```
poly_model <- lm(vitD_level ~ poly(age_month, 3), data = vitD)

summary(poly_model)


Call:
lm(formula = vitD_level ~ poly(age_month, 3), data = vitD)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.862  -9.833  -2.182   8.365  51.903 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           35.964      1.690  21.284   <2e-16 ***
poly(age_month, 3)1   11.858     15.670   0.757    0.451    
poly(age_month, 3)2    7.234     15.670   0.462    0.646    
poly(age_month, 3)3   16.816     15.670   1.073    0.286    
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.67 on 82 degrees of freedom
Multiple R-squared:  0.02308,   Adjusted R-squared:  -0.01266
F-statistic: 0.6458 on 3 and 82 DF,  p-value: 0.5878
```

- Coefficient 1 (11.858): This represents the coefficient for the linear term of the polynomial. With a p-value of 0.45, there is not a significant relationship between age and vitamin-D level in the linear model.
- Coefficient 2 (7.234): This represents the coefficient for the quadratic term of the polynomial. With a p-value of 0.65, this relationship is also not significant.
- Coefficient 3 (16.816): This represents the coefficient for the cubic term of the polynomial and is what I was testing. The p-value of 0.29 is still not significant, meaning that the cubic model was not able to fit the data in a meaningful way.

We fail to reject the null hypothesis. I must determine that there is no statistically significant relationship between age and vitamin-D level.

## Healthy participant model

I separated the healthy participant data into a new dataset.

```
healthy <- vitD %>% filter(autism == "no")
```

First, we must check the assumptions of a linear model:

1. As x (age) varies, the y (vitamin-D level) values follow a straight line.
2. The amount of vertical spread is approximately the same in each strip, except perhaps near the ends.
3. Data is normally distributed, or sample size is greater than 30.

```
cor(healthy$age_month, healthy$vitD_level)

[1] 0.03517393

healthy %>% ggplot(aes(x = age_month, y=vitD_level)) + geom_point() +
  labs(title = "Relationship Between Age and Vitamin-D Level", x = "age
  (months)", y = "Vitamin-D level") + stat_smooth(method = "lm", se =
  FALSE,  formula = "y~x")
```

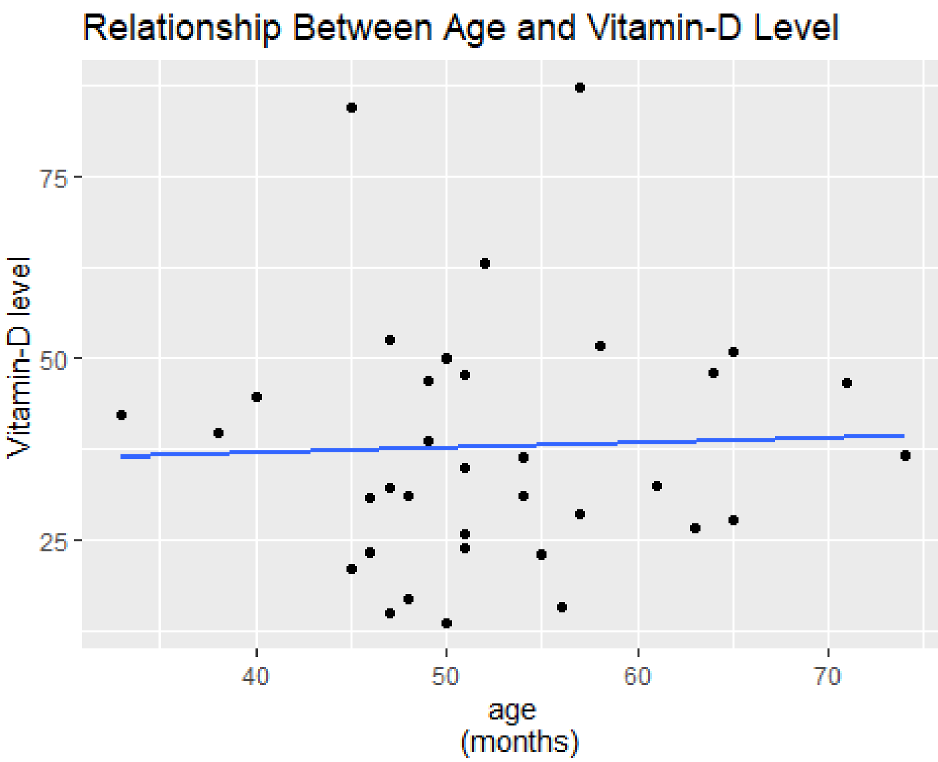


Figure 3.

1. The correlation of x and y is 0.04, which is very weak. Figure 3 shows the data does not appear to be linear in nature.
2. Figure 3 shows that the vertical spread is better than it was in Figure 1 if we ignore the two outliers.
3. The sample size is 35, therefore this assumption is met.

I tried a few different transformations to see if I can fit the data into a better, non-linear model.

```
healthy_model <- healthy %>% ggplot(aes(x = age_month, y = vitD_level)) +
geom_point() + labs(title = "Age and Vitamin-D Level (Healthy)", x = "Age
(months)", y = "Vitamin-D level")

healthy_model +
    geom_smooth(method = "gam", formula = y ~ poly(x, 2), se = FALSE,
    color = "blue") +
    geom_smooth(method = "gam", formula = y ~ poly(x, 3), se = FALSE,
    color = "red") +
    geom_smooth(method = "gam", formula = y ~ log(x), se = FALSE,
    color = "orange")
```

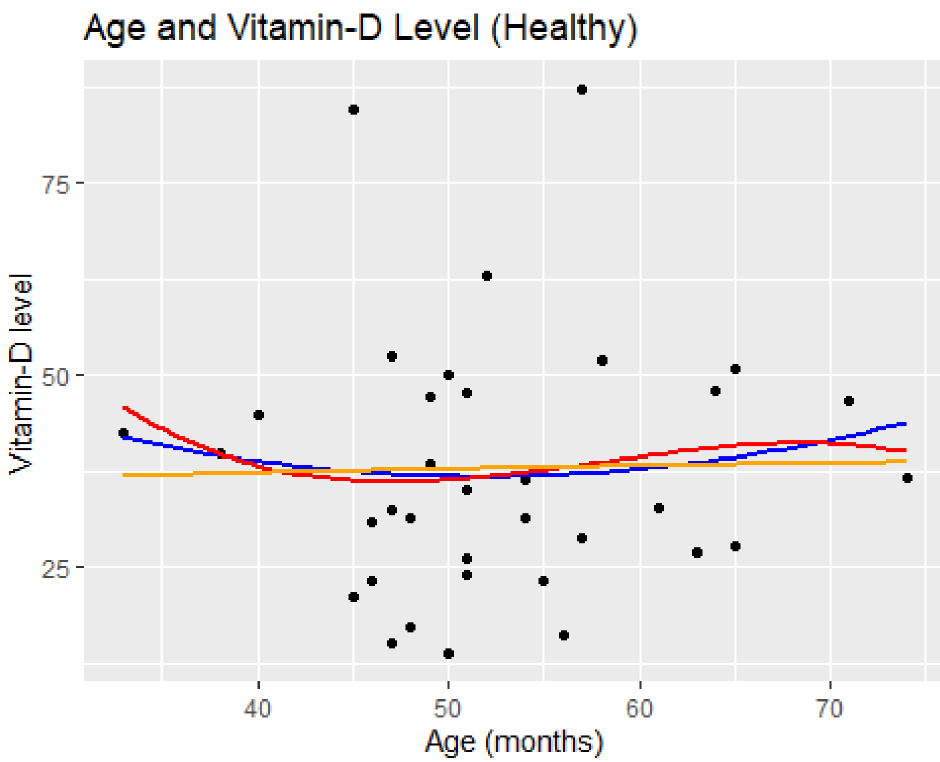


Figure 4.

Again, it appears the polynomial cubed function might fit this relationship best, so I ran a hypothesis test based on it. The significance level was 0.05.

Null hypothesis - β3 = 0
Alternative hypothesis - β3 ≠ 0

```
healthy_poly_model <- lm(vitD_level ~ poly(age_month, 3), data = healthy)

summary(healthy_poly_model)


Call:
lm(formula = vitD_level ~ poly(age_month, 3), data = healthy)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.610 -12.665  -3.365   8.814  48.948 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           37.837      3.010  12.572 1.04e-13 ***
poly(age_month, 3)1    3.515     17.805   0.197    0.845    
poly(age_month, 3)2    9.219     17.805   0.518    0.608    
poly(age_month, 3)3   -7.836     17.805  -0.440    0.663    
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.8 on 31 degrees of freedom
Multiple R-squared:  0.0159,    Adjusted R-squared:  -0.07934
F-statistic: 0.1669 on 3 and 31 DF,  p-value: 0.9179
```

- Coefficient 1 (3.515): This represents the coefficient for the linear term of the polynomial. With a p-value of 0.85, there is not a significant relationship between age and vitamin-D level in the linear model.
- Coefficient 2 (9.219): This represents the coefficient for the quadratic term of the polynomial. With a p-value of 0.608, this relationship is also not significant.
- Coefficient 3 (-7.836): This represents the coefficient for the cubic term of the polynomial and is what I was testing. The p-value of 0.663 is still not significant, meaning that the cubic model was not able to fit the data in a meaningful way. In fact it appears this model was less of a fit than the quadratic model, though neither was significant.

We fail to reject the null hypothesis. I must determine that there is no statistically significant relationship between age and vitamin-D level among healthy participants.

## Autistic participant model

I separated the autistic participant data into a new dataset.

```
autism <- vitD %>% filter(autism == "yes")
```

First, we must check the assumptions of a linear model:

1. As x (age) varies, the y (vitamin-D level) values follow a straight line.
2. The amount of vertical spread is approximately the same in each strip, except perhaps near the ends.
3. Data is normally distributed, or sample size is greater than 30.

```
cor(autism$age_month, autism$vitD_level)

[1] 0.09741064

autism %>% ggplot(aes(x = age_month, y=vitD_level)) + geom_point() +
labs(title = "Relationship Between Age and Vitamin-D Level", x = "age
(months)", y = "Vitamin-D level") + stat_smooth(method = "lm", se =   FALSE,
formula = "y~x")
```

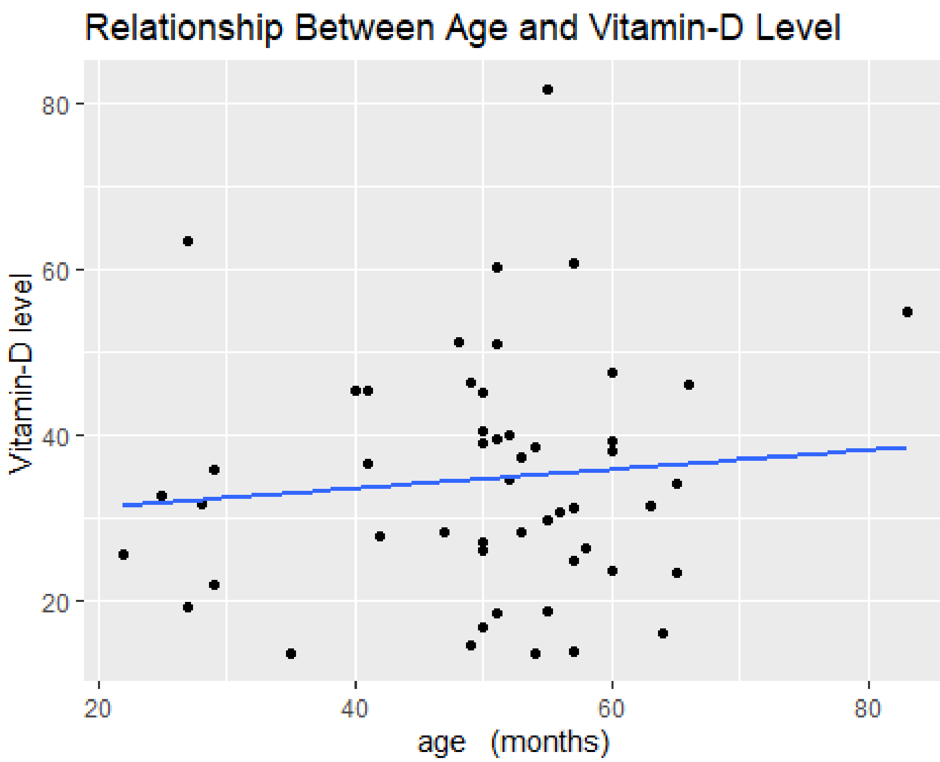


Figure 5.

1. The correlation of x and y is 0.10, which is very weak. Figure 5. shows the data does not appear to be linear in nature.

2. Figure 5 shows that the vertical spread is not equal.

3. The sample size is 51, therefore this assumption is met.

I tried a few different transformations to see if I can fit the data into a better, non-linear model.

```
autism_model <- autism %>% ggplot(aes(x = age_month, y = vitD_level)) +
geom_point() + labs(title = "Age and Vitamin-D Level (Autism)", x = "Age
(months)", y = "Vitamin-D level")

autism_model +
    geom_smooth(method = "gam", formula = y ~ poly(x, 2), se = FALSE,
    color = "blue") +
    geom_smooth(method = "gam", formula = y ~ poly(x, 3), se = FALSE,
    color = "red") +
    geom_smooth(method = "gam", formula = y ~ log(x), se = FALSE,
    color = "orange")
```

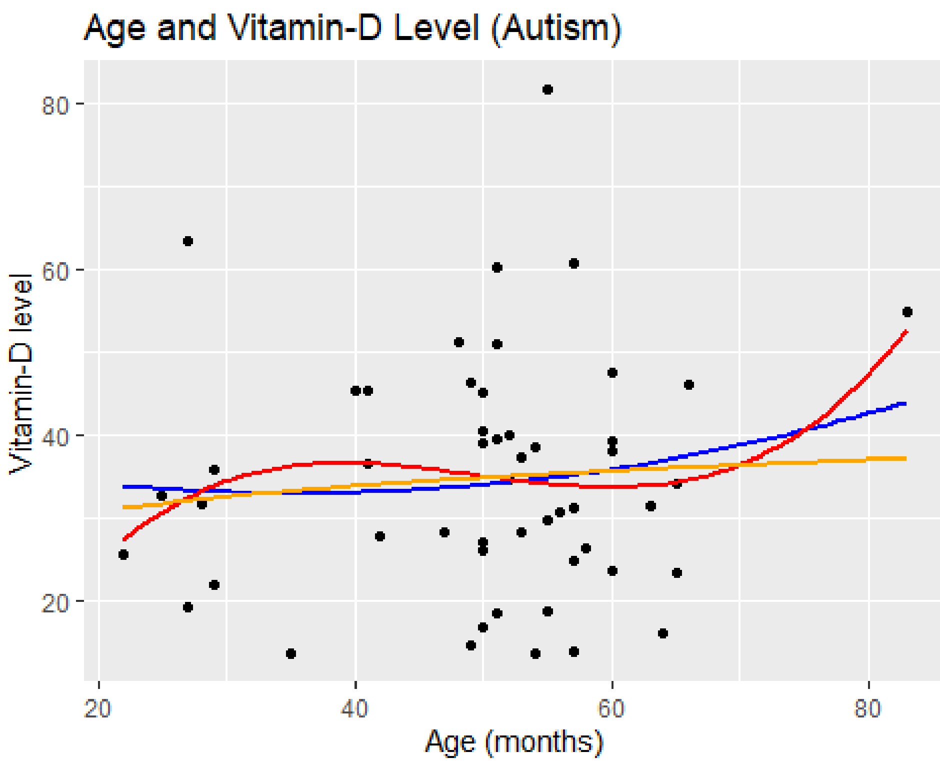


Figure 6.

Again, it appears the polynomial cubed function might fit this relationship best, so I ran a hypothesis test based on it. The significance level was 0.05.

Null hypothesis - β3 = 0
Alternative hypothesis - β3 ≠ 0

```
autism_poly_model <- lm(vitD_level ~ poly(age_month, 3), data = autism)

summary(autism_poly_model)


Call:
lm(formula = vitD_level ~ poly(age_month, 3), data = autism)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.647  -8.952  -0.206   5.577  47.698 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           34.678      2.040  16.998   <2e-16 ***
poly(age_month, 3)1    9.942     14.570   0.682    0.498    
poly(age_month, 3)2    7.680     14.570   0.527    0.601    
poly(age_month, 3)3   16.809     14.570   1.154    0.254    
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.57 on 47 degrees of freedom
Multiple R-squared:  0.04227,   Adjusted R-squared:  -0.01886
F-statistic: 0.6915 on 3 and 47 DF,  p-value: 0.5618
```

- Coefficient 1 (9.942): This represents the coefficient for the linear term of the polynomial. With a p-value of 0.50, there is not a significant relationship between age and vitamin-D level in the linear model.
- Coefficient 2 (7.680): This represents the coefficient for the quadratic term of the polynomial. With a p-value of 0.60, this relationship is also not significant.
- Coefficient 3 (16.809): This represents the coefficient for the cubic term of the polynomial and is what I was testing. The p-value of 0.25 is still not significant, meaning that the cubic model was not able to fit the data in a meaningful way.

We fail to reject the null hypothesis. I must determine that there is no statistically significant relationship between age and vitamin-D level among autistic participants.

## Comparing results from models 2 & 3

Just comparing the scatter plots visualizing the relationship of age and vitamin-D level for healthy and autistic participants, the plot representing the autistic sample had a larger slope. This potentially could be interpreted as there being a stronger relationship between the two variables among autistic people. However, since neither regression yielded a low enough p-value to indicate significance, there does not appear to be a link between age and vitamin-D level among either disease group. The autistic sample size was larger (51 vs 35), so that could explain some of the discrepancy between groups, but the p-values for both groups were so high that having a larger sample size would be unlikely to yield a low enough p-value to indicate significance.