Module 4 Report: Emalee Schuler

Age and Vitamin D Levels (Healthy & Autistic Individuals)

Null Hypothesis H_0 : There is no significant linear relationship between age(x) and vitamin D level (y).

Alternative Hypothesis H_a : There is a significant linear relationship between age(x) and vitamin D level (y).

Age and Vitamin D Levels (Healthy Individuals)

Null Hypothesis H_0 : In healthy individuals, there is no significant linear relationship between age(x) and vitamin D level (y) in .

Alternative Hypothesis H_a : In healthy individuals, there is a significant linear relationship between age(x) and vitamin D level (y).

Age and Vitamin D Levels (Autistic Individuals)

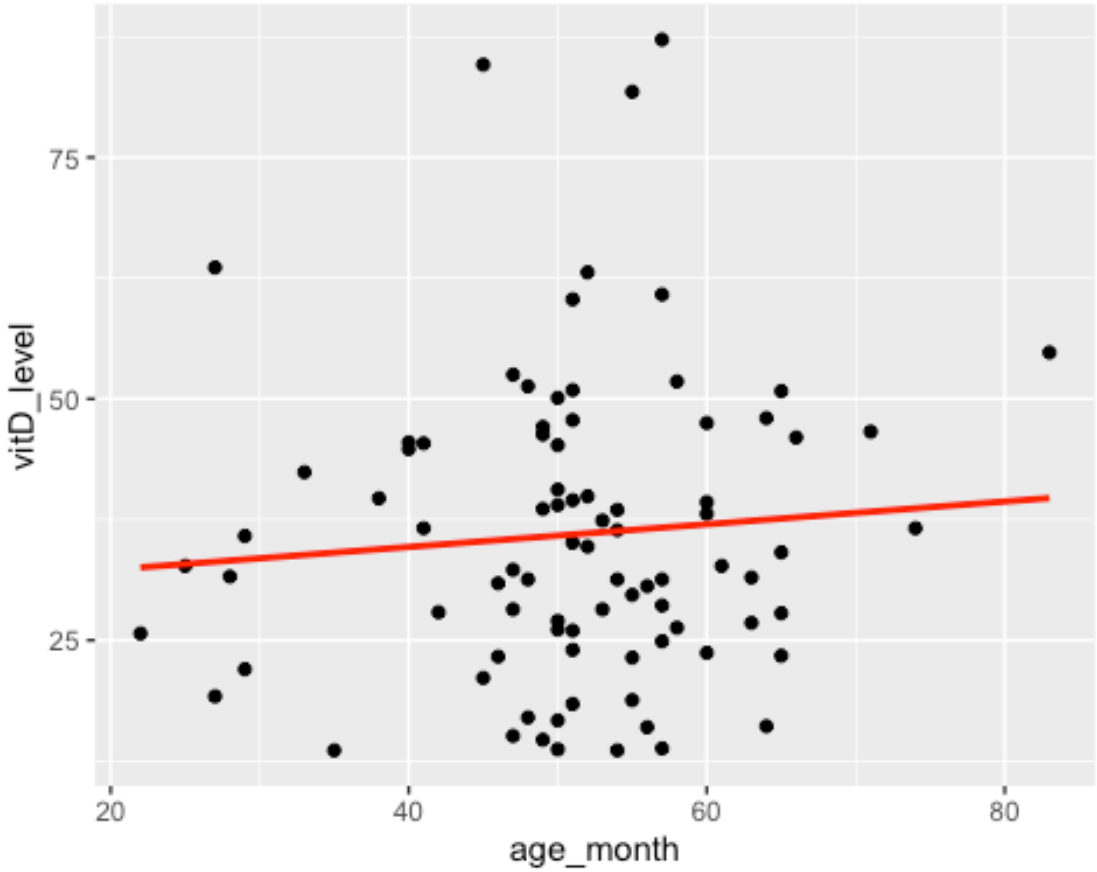
Null Hypothesis H_0 : In autistic individuals, there is no significant linear relationship between age(x) and vitamin D level (y).

Alternative Hypothesis H_a : In autistic individuals, there is a significant linear relationship between age(x) and vitamin D level (y).

Step 1

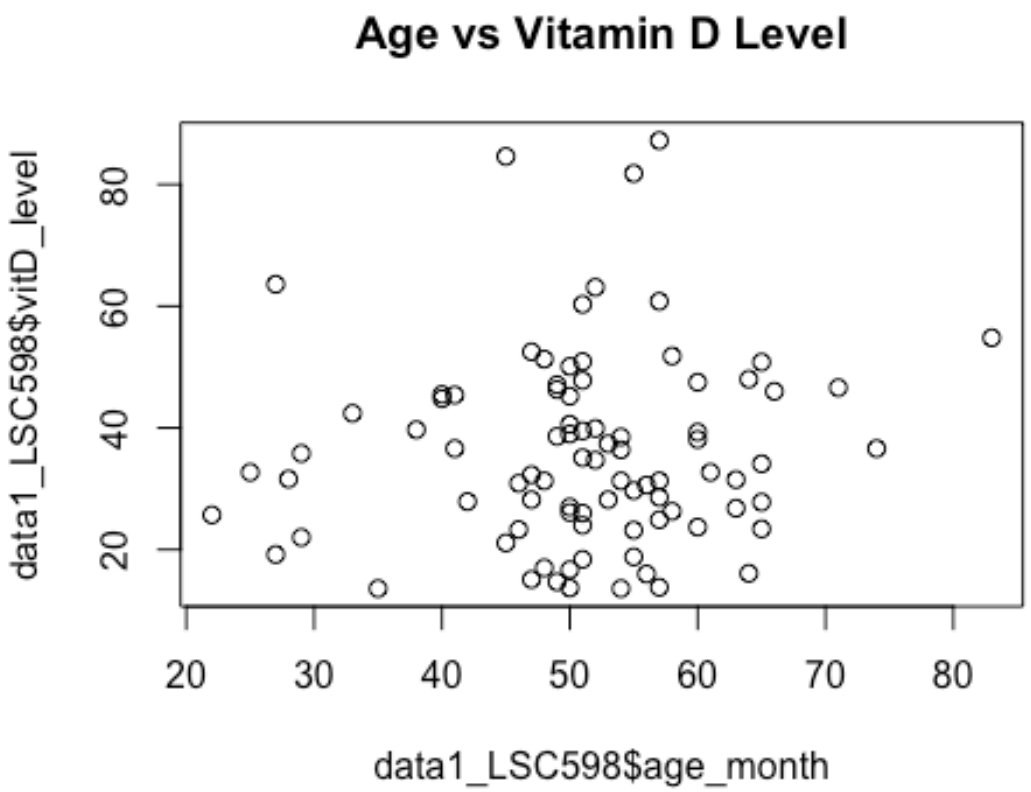
```
library(ggplot2)
library(dplyr)


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

data1_LSC598 <- read.csv("~/Downloads/data1_LSC598.txt", sep="")

data1_LSC598 <- data1_LSC598 %>% filter(!is.na(vitD_level))

ggplot(data = data1_LSC598, aes(x = age_month, y = vitD_level)) +
     geom_point() + geom_smooth(method = "lm", se= FALSE, color= "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

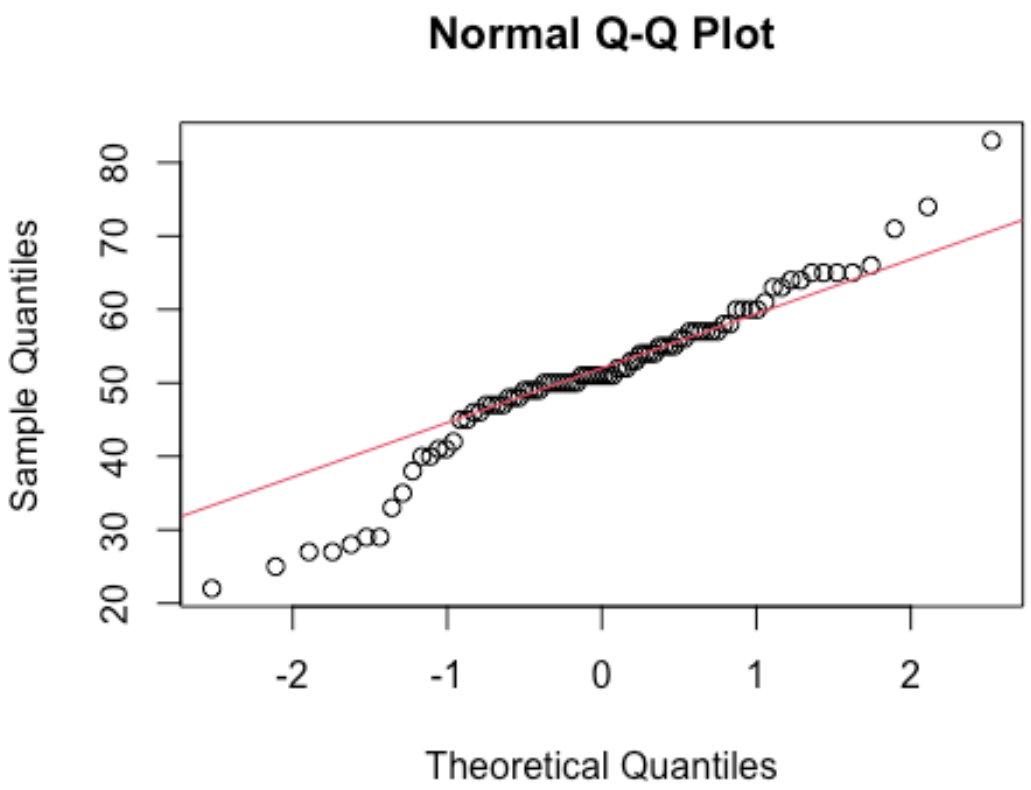



```
   labs(title = "Age Vs Vitamin D Level",
        x = "Age (month)",
        y = "Vitamin D Level")

$x
[1] "Age (month)"

$y
[1] "Vitamin D Level"

$title
[1] "Age Vs Vitamin D Level"

attr(,"class")
[1] "labels"
```

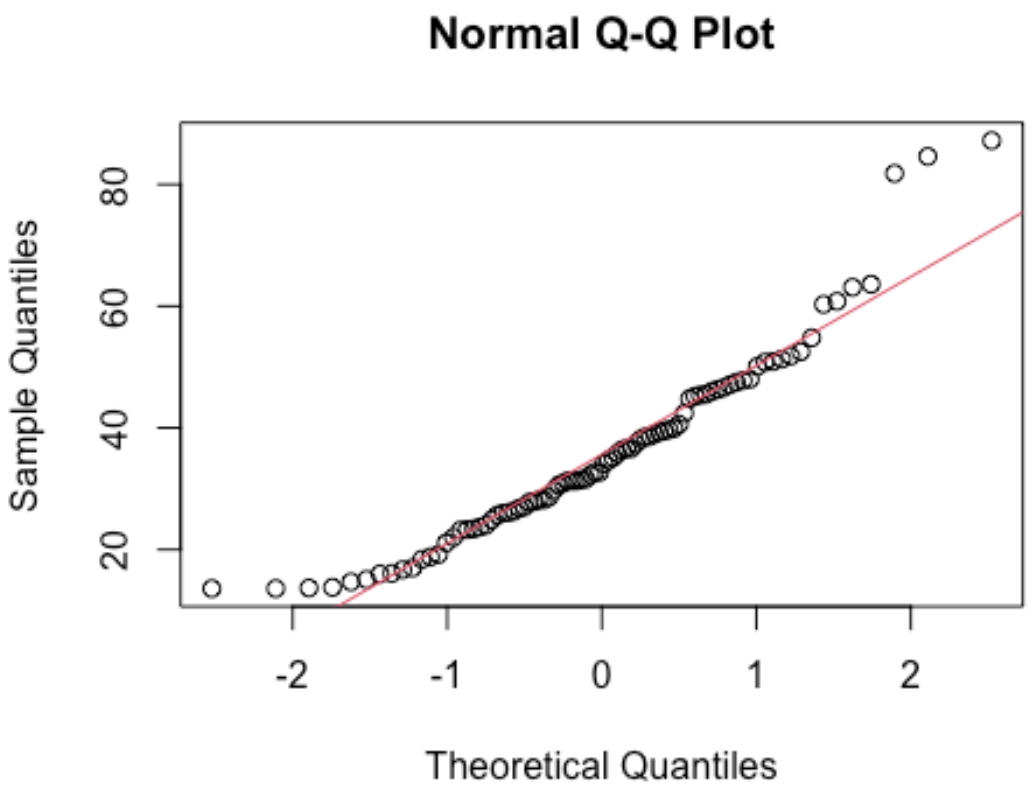
```
plot(data1_LSC598$age_month, data1_LSC598$vitD_level)
title("Age vs Vitamin D Level")
```




```r
qqnorm(data1_LSC598$age_month)
qqline(data1_LSC598$age_month, col = 2)
```




```r
qqnorm(data1_LSC598$vitD_level)
qqline(data1_LSC598$vitD_level, col = 2)
```




```
agevitD1 <- lm(vitD_level ~ age_month, data = data1_LSC598)

summary(agevitD1)


Call:
lm(formula = vitD_level ~ age_month, data = data1_LSC598)

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
```

Interpretation of linear regression from Step 1:

Assumptions Check:

- The scatterplot shows homoscedasticity with constant variance from the horizontal line through it, the Q-Q plot is normally distributed.

Model Equation:

- The linear regression equation is: y vitD_level) = 29.9566 + 0.1176 * x (age_month)

Coefficients:

- Intercept (29.9566): This represents the estimated value of `vitD_level` when `age_month` is equal to zero. In this context, it doesn't have a practical interpretation because it's unlikely to have an age of zero.

- The slope (0.1176): This represents the estimated change in `vitD_level` for a one-unit change in `age_month` while holding all other variables constant.

Significance:

- The significance level of the coefficients is determined by the p-values. In this case, the p-value is 0.449622, which is greater than the typical significance level of 0.05. This suggests that `age_month` is not a statistically significant predictor of `vitD_level` in this model. In other words, there isn't enough evidence to conclude that there is a significant linear relationship between `age_month` and `vitD_level`. Therefore, we fail to reject the null hypothesis that there is no significant linear relationship between the two variables.

Model Fit:

- Residual standard error: The residual standard error is 15.61. Because a smaller RSE suggests a better fit of the model to the data, a residual standard error of 15.61 indicates that the model's predictions are not very close to the actual data points.

- R-squared (Multiple R-squared and Adjusted R-squared): In this case, the R-squared is very low (0.006822), indicating that the model does not explain much of the variability in `vitD_level`. The Adjusted R-squared is even lower, indicating that the model is not a very good fit.

- F-statistic and p-value: In this case, the F-statistic is 0.577 with a p-value of 0.4496, indicating that the model as a whole is not statistically significant.

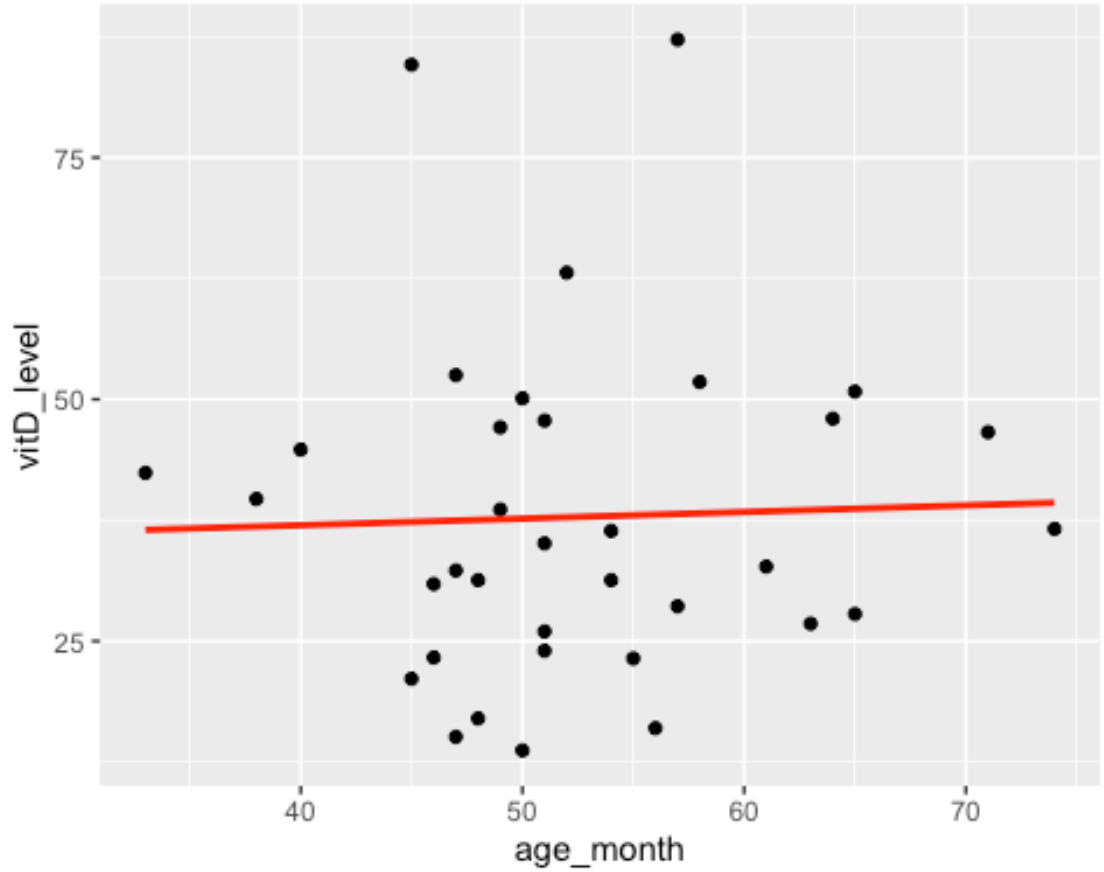
Interpretation:

Based on the results, it appears that this linear regression model with `age_month` as the predictor variable is not a good fit for explaining `vitD_level`. The coefficient for `age_month` is not statistically significant, and the low R-squared values suggest that the model does not explain much of the variation in `vitD_level`. This may indicate that other variables or a more complex model are needed to better predict `vitD_level`.

Step 2

```
#Step 2
healthy <- data1_LSC598[data1_LSC598$group == 0, c("vitD_level",
"age_month")]

ggplot(data = healthy, aes(x = age_month, y = vitD_level)) +
     geom_point() + geom_smooth(method = "lm", se= FALSE, color= "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```
   labs(title = "Age Vs Vitamin D Level",
        x = "Age (month)",
        y = "Vitamin D Level")
```

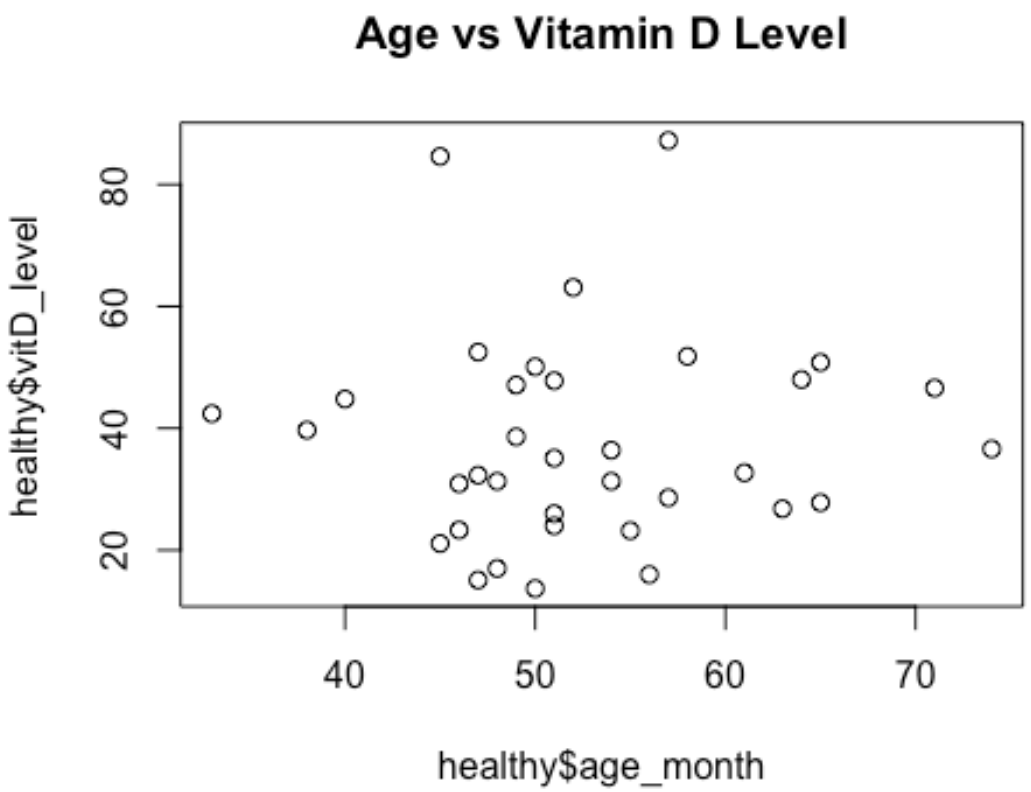
```
$x
[1] "Age (month)"

$y
[1] "Vitamin D Level"

$title
[1] "Age Vs Vitamin D Level"
```

```
attr(,"class")
[1] "labels"
```

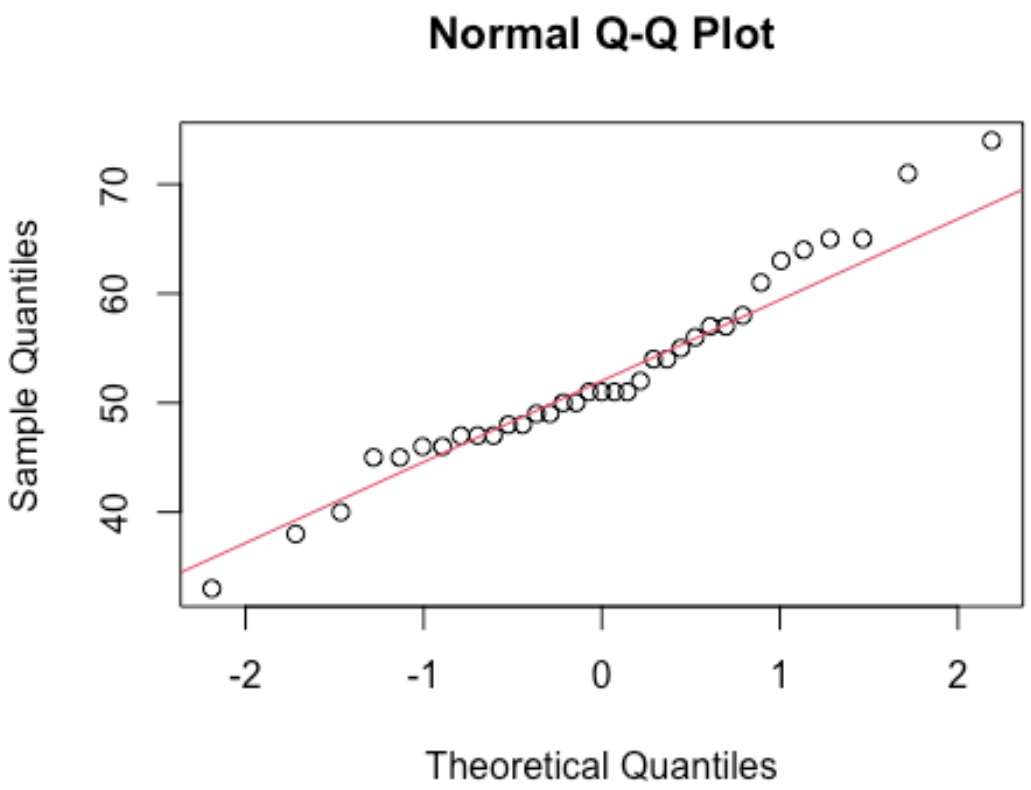
```r
agevitD1_healthy <- lm(vitD_level ~ age_month, data = healthy)

summary(agevitD1_healthy)
```

```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-23.965 -11.744  -2.705   9.441  49.056 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 34.24925   17.98735   1.904   0.0657 .
age_month    0.06832    0.33792   0.202   0.8410  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.38 on 33 degrees of freedom
Multiple R-squared:  0.001237,  Adjusted R-squared:  -0.02903 
F-statistic: 0.04088 on 1 and 33 DF,  p-value: 0.841
```

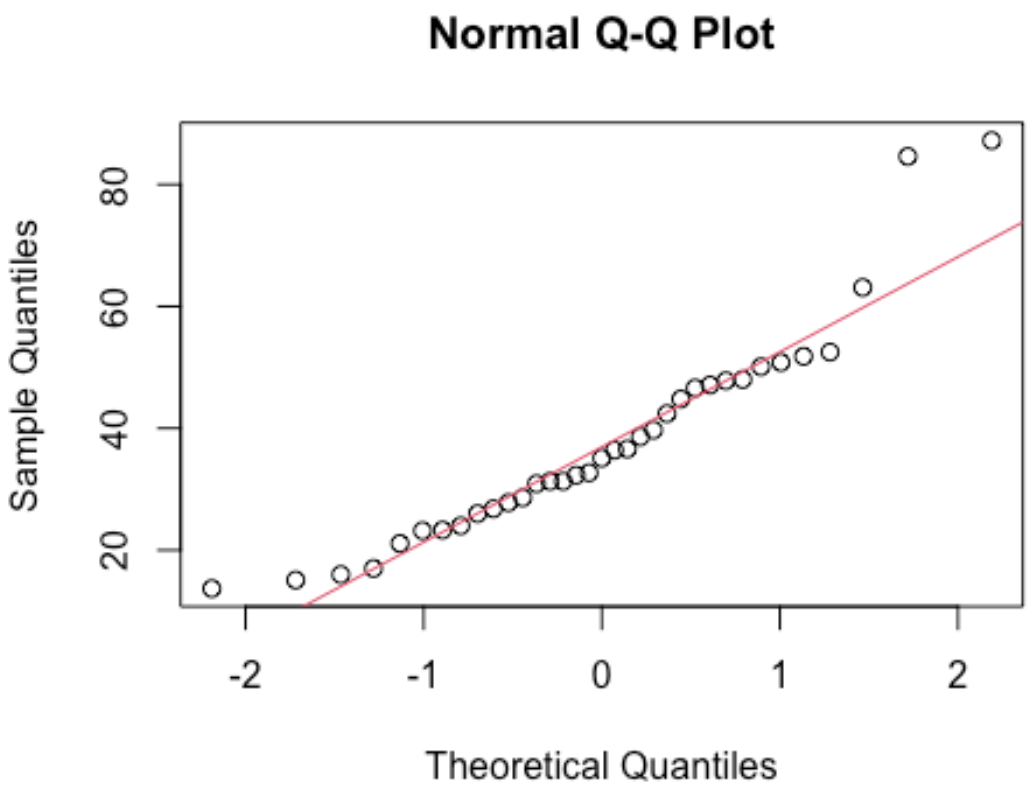
```r
plot(healthy$age_month, healthy$vitD_level)
title("Age vs Vitamin D Level")
```




```r
qqnorm(healthy$age_month)
qqline(healthy$age_month, col = 2)
```




```r
qqnorm(healthy$vitD_level)
qqline(healthy$vitD_level, col = 2)
```



Interpretation of linear regression from Step 2:

Assumptions Check:

The scatterplot shows homoscedasticity with constant variance from the horizontal line through it, the Q-Q plot is normally distributed.

Model Equation:

- The linear regression equation is: y (vitD_level) = 34.24925 + 0.06832 * x(age_month)

Coefficients:

- Intercept (34.24925): This represents the estimated value of `vitD_level` when `age_month` is equal to zero. In this context, it doesn't have a practical interpretation because it's unlikely to have an age of zero.

- The slope(0.06832): This represents the estimated change in `vitD_level` for a one-unit change in `age_month` while holding all other variables constant.

Significance:

- The significance level of the coefficients is determined by the p-values. In this case, the p-value is 0.8410, which is much greater than the typical significance level of 0.05. This

suggests that `age_month` is not a statistically significant predictor of `vitD_level` for healthy participants in this model. In other words, there isn't enough evidence to conclude that there is a significant linear relationship between `age_month` and `vitD_level` for this subgroup. Therefore, we fail to reject the null hypothesis that there is no significant linear relationship between the two variables.

Model Fit:

- Residual standard error: The residual standard error is 17.38. It represents the typical size of the residuals, which is a measure of the model's accuracy. Because a smaller RSE suggests a better fit of the model to the data, a residual standard error of 17.38 indicates that the model's predictions are not very close to the actual data points.

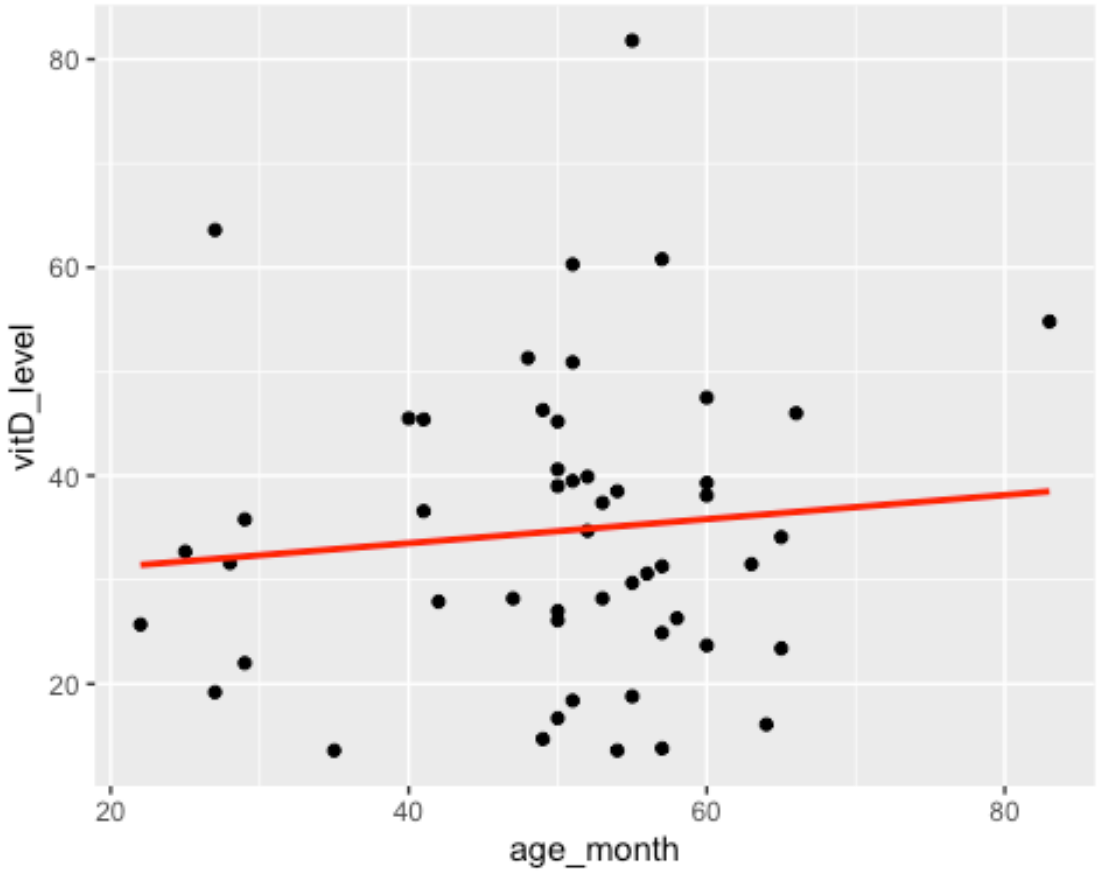
- R-squared (Multiple R-squared and Adjusted R-squared): R-squared is a measure of how well the independent variable (`age_month`) explains the variation in the dependent variable (`vitD_level`). In this case, the R-squared is very low (0.001237), indicating that the model does not explain much of the variability in `vitD_level` for healthy participants. The Adjusted R-squared is even lower, indicating that the model may not be a good fit for this subgroup.

Interpretation:

Based on the results, it appears that the linear regression model with `age_month` as the predictor variable is not a good fit for explaining `vitD_level` for healthy participants. The coefficient for `age_month` is not statistically significant, and the low R-squared values suggest that the model does not explain much of the variation in `vitD_level` for this subgroup. This may indicate that other variables or a more complex model are needed to better predict `vitD_level` for healthy participants.

```
#Step 3
autism <- data1_LSC598[data1_LSC598$group == 1, c("vitD_level", "age_month")]

ggplot(data = autism, aes(x = age_month, y = vitD_level)) +
     geom_point() + geom_smooth(method = "lm", se= FALSE, color= "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



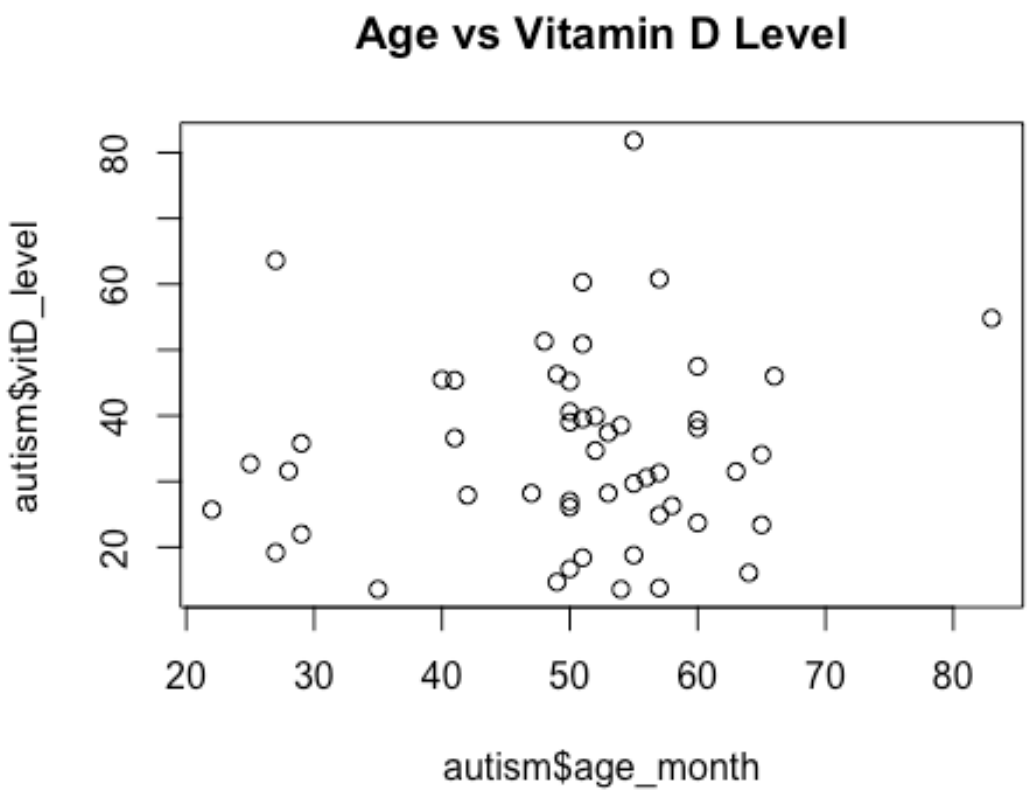

```
   labs(title = "Age Vs Vitamin D Level",
        x = "Age (month)",
        y = "Vitamin D Level")

$x
[1] "Age (month)"

$y
[1] "Vitamin D Level"

$title
[1] "Age Vs Vitamin D Level"

attr(,"class")
[1] "labels"
```

```
agevitD1_autism <- lm(vitD_level ~ age_month, data = autism)

summary(agevitD1_autism)
```

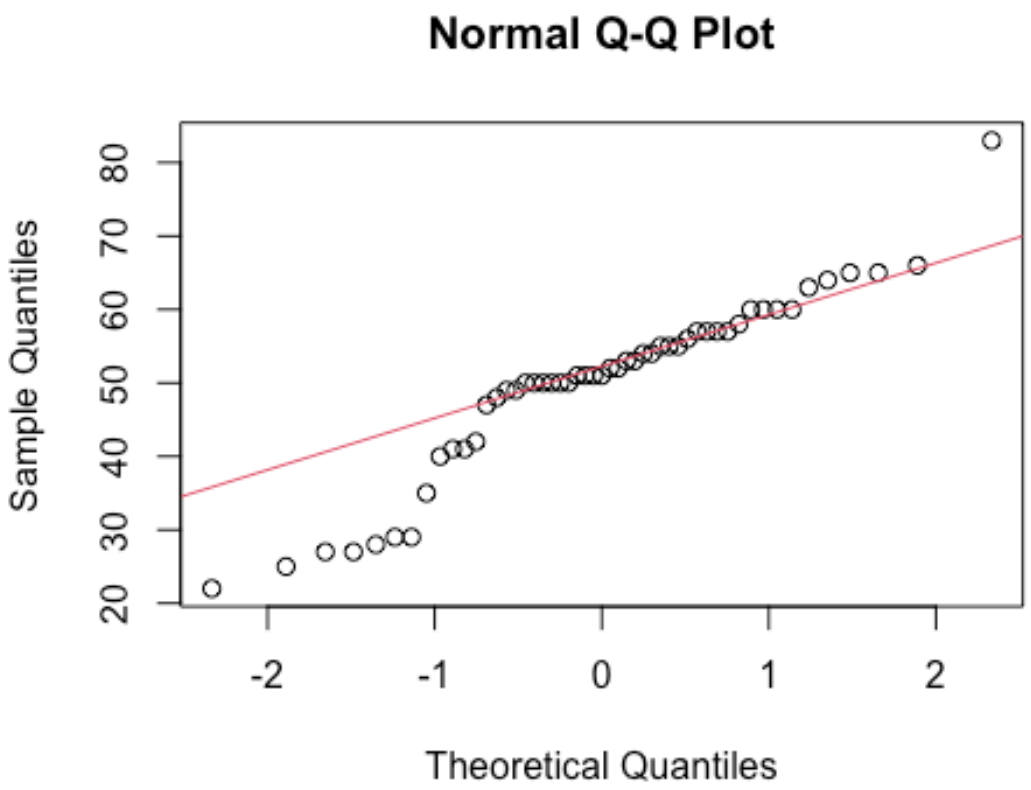
```
Call:
lm(formula = vitD_level ~ age_month, data = autism)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-21.679  -9.768  -0.525   7.706  46.552

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.8873     8.6934   3.323  0.00169 **
age_month     0.1156     0.1688   0.685  0.49649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.51 on 49 degrees of freedom
Multiple R-squared:  0.009489,  Adjusted R-squared:  -0.01073
F-statistic: 0.4694 on 1 and 49 DF,  p-value: 0.4965
```

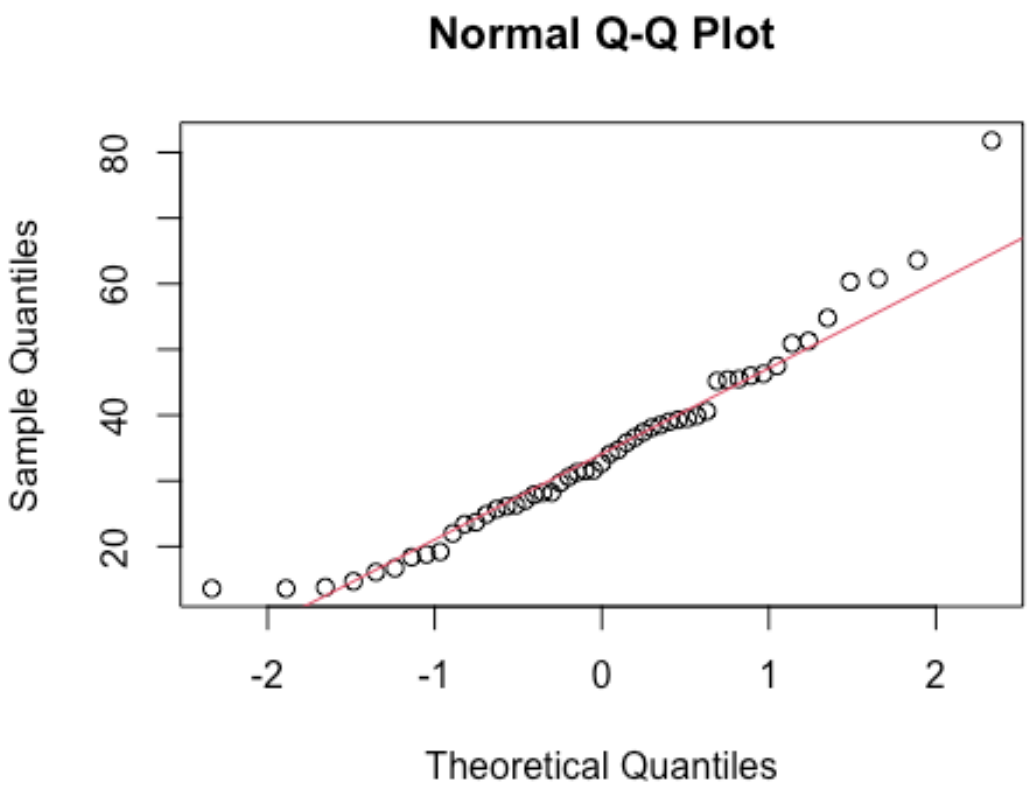
```r
plot(autism$age_month, autism$vitD_level)
title("Age vs Vitamin D Level")
```




```r
qqnorm(autism$age_month)
qqline(autism$age_month, col = 2)
```




```r
qqnorm(autism$vitD_level)
qqline(autism$vitD_level, col = 2)
```



Interpretation of linear regression from Step 3:

Assumptions Check:

The scatterplot shows homoscedasticity with constant variance from the horizontal line through it, the Q-Q plot is normally distributed

Model Equation:

- The linear regression equation is: y (vitD_level) = 28.8873 + 0.1156 *x (age_month)

Coefficients:

- Intercept (28.8873): This represents the estimated value of `vitD_level` when `age_month` is equal to zero. In this context, it doesn't have a practical interpretation because it's unlikely to have an age of zero.

- The slope (0.1156): This represents the estimated change in `vitD_level` for a one-unit change in `age_month` while holding all other variables constant.

Significance:

- The significance level of the coefficients is determined by the p-values. In this case, the p-value of 0.49649 is much greater than the typical significance level of 0.05. This suggests

that `age_month` is not a statistically significant predictor of `vitD_level` for autism patients in this model. In other words, there isn't enough evidence to conclude that there is a significant linear relationship between `age_month` and `vitD_level` for this subgroup. Therefore, we fail to reject the null hypothesis that there is no significant linear relationship between the two variables.

Model Fit:

- Residual standard error: The residual standard error is 14.51. It represents the typical size of the residuals, which is a measure of the model's accuracy. Because a smaller RSE suggests a better fit of the model to the data, a residual standard error of 14.51 indicates that the model's predictions are not very close to the actual data points.

- R-squared (Multiple R-squared and Adjusted R-squared): R-squared is a measure of how well the independent variable (`age_month`) explains the variation in the dependent variable (`vitD_level`). In this case, the R-squared is very low (0.009489), indicating that the model does not explain much of the variability in `vitD_level` for autism patients. The Adjusted R-squared is even lower, indicating that the model may not be a good fit for this subgroup.

Interpretation:

Based on the results, the linear regression model with `age_month` as the predictor variable is not a good fit for explaining `vitD_level` for autistic patients. The coefficient for `age_month` is not statistically significant, and the low R-squared values suggest that the model does not explain much of the variation in `vitD_level` for this subgroup. This may indicate that other variables or a more complex model are needed to better predict `vitD_level` for autism patients.

Comparison of the results from Step 2 and Step 3:

The results from Step 2 and Step 3 are similar in that both steps in that

Based on the results from Step 2 and Step 3, it appears that both linear regression models where `age_month` is the independent variable are not a good fits for explaining `vitD_level` for either type of participants. The coefficient for `age_month` is not statistically significant in either step, and the low R-squared values suggest that the model does not explain much of the variation in `vitD_level` for this subgroup. Furthermore, both steps had high residual standard error values. This indicates that the models' predictions are not very close to the actual data points. Lastly, both steps had p-values greater than 0.05, which suggests that age is not a statistically significant predictor of vitamin D level. The aforementioned information may indicate that other variables or a more complex model are needed to better predict `vitD_level` for both types of participants.