# Linear regression modules using the breast feed dataset

AUTHOR
Steve Simon and Leroy Wheeler

PUBLISHED
October 3, 2024

This program reads data and fits various linear regression models on a breast feeding study in pre-term infants. Find more information in the data dictionary. This code is placed in the public domain.

## Libraries

You should always load the tidyverse library. The broom library provides the glance, tidy, and augment functions that help you with computations of linear regression models. The car library provides the vif function for measuring collinearity.

```
library(broom)
library(car)
library(tidyverse)
```

## Question 1: Change the program so that it reads in the breast-feeding-preterm.csv file. Show a glimpse of the data and verify that you have properly read in all 82 rows and 31 columns. No intepretation is necessary for this question

Use the read_csv function to read the data. With a large number of variables, you may choose to leave the col_types out.R will usually figure out which variables are numeric and which are strings.

Replace all the numeric codes of -1 with the missing value code (NA).

```
bf <- read_csv(
  file="../data/breast-feeding-preterm.csv",
  col_names=TRUE)
```

```
Rows: 84 Columns: 30
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (2): feed_type, race
dbl (28): age_stop, sepsis, total_ab, del_type, mom_age, gravida, para, mar_...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(bf)
```

```
Rows: 84
Columns: 30
$ feed_type <chr> "Treatmen", "Treatmen", "Control", "Treatmen", "Control", "C…
$ age_stop  <dbl> 30, 4, 12, 29, 24, 24, 27, 5, 32, 20, 24, 5, 16, 10, 16, 18,…
$ sepsis    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, …
$ total_ab  <dbl> 221, 12, 88, 108, 0, 3, 5, 219, 391, 51, 72, 26, 628, 68, 47…
$ del_type  <dbl> 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, …
$ mom_age   <dbl> 30, 19, 37, 29, 23, 23, 29, 20, 40, 27, 40, 26, 33, 29, 32, …
$ gravida   <dbl> 2, 1, 3, 3, 1, 1, 2, 2, 2, 2, 3, 2, 3, 5, 3, 1, 1, 1, 2, 1, …
$ para      <dbl> 1, 1, 3, 1, 2, 2, 1, 2, 2, 1, 1, 2, 3, 3, 2, 1, 2, 2, 2, 2, …
$ mar_st    <dbl> 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
$ race      <chr> "W", "W", "W", "W", "W", "W", "W", "W", "W", "W", "W", "W", …
$ smoker    <dbl> 2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, …
$ mi_hosp   <dbl> 10, NA, 8, 90, 25, 25, 15, 30, 13, 15, 12, 25, 10, 75, 10, 5…
$ ng_tube   <dbl> 39, 13, 14, 32, 4, 11, 15, 30, 54, 31, 27, 10, 43, 26, 7, 30…
$ tot_bott  <dbl> 0, 68, 92, 0, 20, 65, 33, 152, 0, 13, 54, 39, 94, 100, 41, 0…
$ bw        <dbl> 1.738, 1.710, 1.955, 1.730, 2.050, 1.656, 1.735, 1.160, 1.39…
$ gest_age  <dbl> 31, 34, 32, 31, 35, 35, 34, 30, 29, 32, 32, 34, 29, 32, 32, …
$ apgar1    <dbl> 8, 7, 6, 7, 8, 6, 2, 6, 8, 7, 7, 7, 6, 4, 8, 8, 8, 8, 1, 8, …
$ apgar5    <dbl> 9, 8, 8, 9, 9, 9, 5, 8, 9, 8, 7, 8, 9, 8, 9, 9, 9, 9, 7, 9, …
$ bf1_wt    <dbl> 1.575, 1.676, 1.947, 1.615, 2.025, 1.665, 1.695, NA, 1.445, …
$ bf1_age   <dbl> 9, 11, 12, 16, 1, 1, 7, NA, 27, 3, 7, 5, 28, 8, 10, 8, 34, 3…
$ dc_wt     <dbl> 2.610, 2.048, 2.425, 2.125, 1.980, 1.995, 1.995, 2.245, 2.10…
$ dc_age    <dbl> 46, 26, 32, 38, 8, 18, 22, 53, 57, 34, 32, 17, 58, 44, 19, 3…
$ dc3_wt    <dbl> 2.665, 2.048, 3.005, 2.130, 2.136, 3.454, 1.996, 2.245, 2.69…
$ bf0       <dbl> 1, 4, 2, 1, 2, 2, 2, 4, 1, 1, 1, 2, 1, 2, 1, 1, 4, 4, 1, 1, …
```

```
$ bf1       <dbl> 1, 4, 1, 1, 2, 2, 1, 4, 1, 1, 2, 2, 2, 2, 1, 1, 4, 4, 1, 1, …
$ bf2       <dbl> 1, 4, 2, 1, 2, 2, 1, 4, 1, 1, 2, 4, 2, 2, 1, 2, 4, 4, 1, 1, …
$ bf3       <dbl> 1, 4, 2, 1, 2, 2, 1, 4, 1, 2, 2, 4, 2, 4, 2, 2, 4, 4, 1, 1, …
$ bf4       <dbl> 1, 4, 4, 1, 2, 2, 1, 4, 1, 4, 2, 4, 4, 4, 4, 4, 4, 4, 1, 2, …
$ feed_cod  <dbl> 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, …
$ feed_rev  <dbl> 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, …
```

## Convert -1 to NA

The code below only works because every single variable in the dataset is non-negative.

```
bf[bf==-1] <- NA
```

## Question 2: Compute descriptive statistics (counts and percentages) for feed_type. Interpret these values.

```
bf |>
    count(feed_type) |>
    mutate(total=sum(n)) |>
    mutate(pct=100*n/total)
```

```
# A tibble: 2 × 4
  feed_type     n total   pct
  <chr>     <int> <int> <dbl>
1 Control      46    84  54.8
2 Treatmen     38    84  45.2
```

There were similar numbers of each participants in the group in the study. The control group were infants who were bottle fed while the treatment group were infants who were fed using an NG tube.

## Question 3: Compute descriptive statistics (mean, standard deviation, minimum, and maximum) for age_stop. Interpret these values.

```
bf |>
    summarize(
        mean_age_stop=mean(age_stop, na.rm=TRUE),
        sd_age_stop=sd(age_stop, na.rm=TRUE),
        min_age_stop=min(age_stop, na.rm=TRUE),
        max_age_stop=max(age_stop, na.rm=TRUE),
        n_missing=sum(is.na(age_stop))) |>
    data.frame()
```

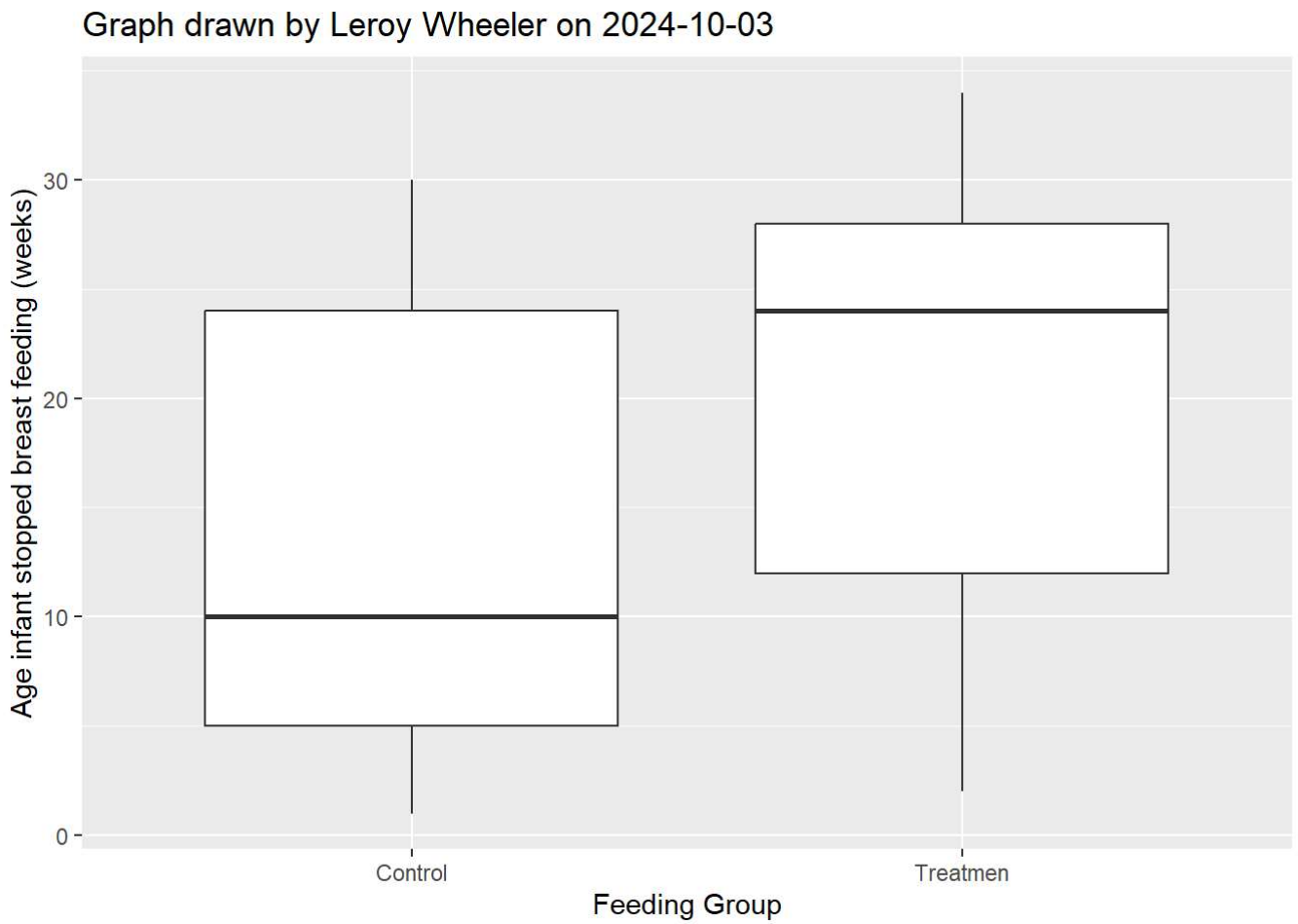
```
  mean_age_stop sd_age_stop min_age_stop max_age_stop n_missing
1      16.58537    10.24147            1           34         2
```

The mean age at which all infants stopped breast feeding was about 17 weeks, with a standard deviation of 10 weeks. The min was one week and the max was 34 weeks. There were two missing values in the data set.

## Question 4: Draw a boxplot comparing age_stop for each level of feed_type. Interpret this plot.

```
bf |>
  ggplot(aes(age_stop, feed_type)) +
    geom_boxplot() +
    coord_flip() +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03") +
      xlab("Age infant stopped breast feeding (weeks)") +
      ylab("Feeding Group")
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




Infants in the treatment group who were fed through an NG tube continued to breast feed significantly longer than the control group of infants who were bottle fed.

## Question 5: Calculate the means and standard deviations for each level of feed_type. Interpret these numbers

```
bf |>
  group_by(feed_type) |>
    summarize(
```

```
        age_stop_mn=mean(age_stop, na.rm=TRUE),
        age_stop_sd=sd(age_stop, na.rm=TRUE))
```

```
# A tibble: 2 × 3
  feed_type age_stop_mn age_stop_sd
  <chr>           <dbl>       <dbl>
1 Control          13.3        9.98
2 Treatmen         20.4        9.30
```

Infants in the treatment group who were fed through an NG tube continued to breast feed for an average of 20 weeks while the control group of infants who were bottle fed continued to breast feed for an average of 13 weeks.

## Question 6: Compute a linear regression model predicting age_stop using feed_type. What value does R assign to 0 and what value does R assign to 1? Interpret the slope and intercept for this linear regression model.

```
m1 <- lm(age_stop ~ feed_type, data=bf)
m1
```

```
Call:
lm(formula = age_stop ~ feed_type, data = bf)

Coefficients:
      (Intercept)  feed_typeTreatmen
            13.32               7.05
```

R has assigned the value 0 to the control group and a value of 1 to the treatment group. From the linear regression equation we see that on average the control group of infants stopped breast feeding at 13 weeks while the treatment group continued to breast feed for 7 weeks longer.

## Question 7: Compute R-squared for this regression model. Interpret this number.

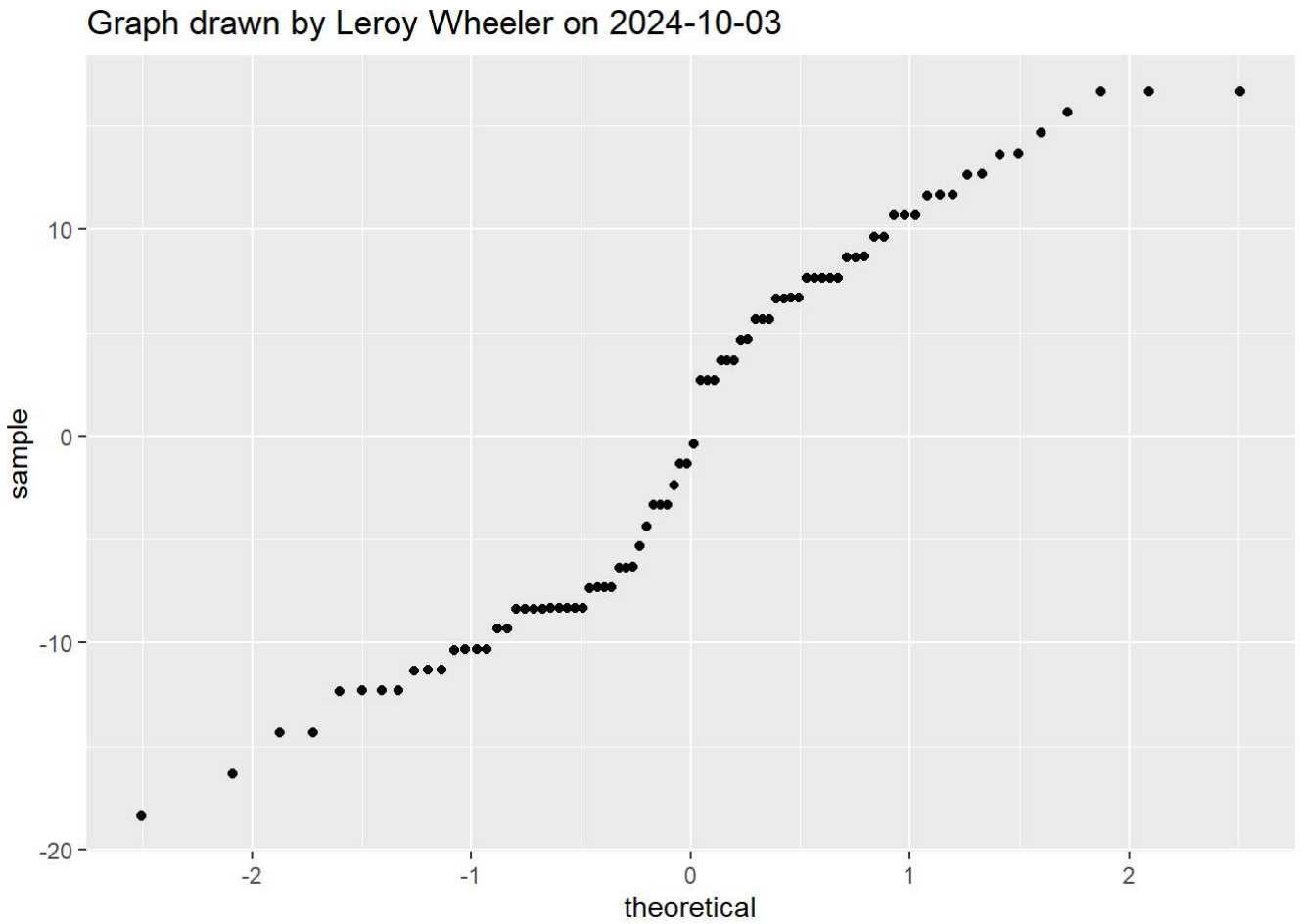
```
glance(m1)$r.squared
```

```
[1] 0.1192946
```

The R squared value for our linear regression model is 0.12 which means that only about 12% of the variation in breast feeding time can be explained by the NG tube feeding. Feeding type is therefore a weak predictor of breast feeding time.

## Question 8: Draw a normal probability plot and a histogram for the residuals from this regression model. Is the assumption of normality satisfied?

```
r1 <- augment(m1)
r1 |>
    ggplot(aes(sample=.resid)) +
      stat_qq() +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03")
```

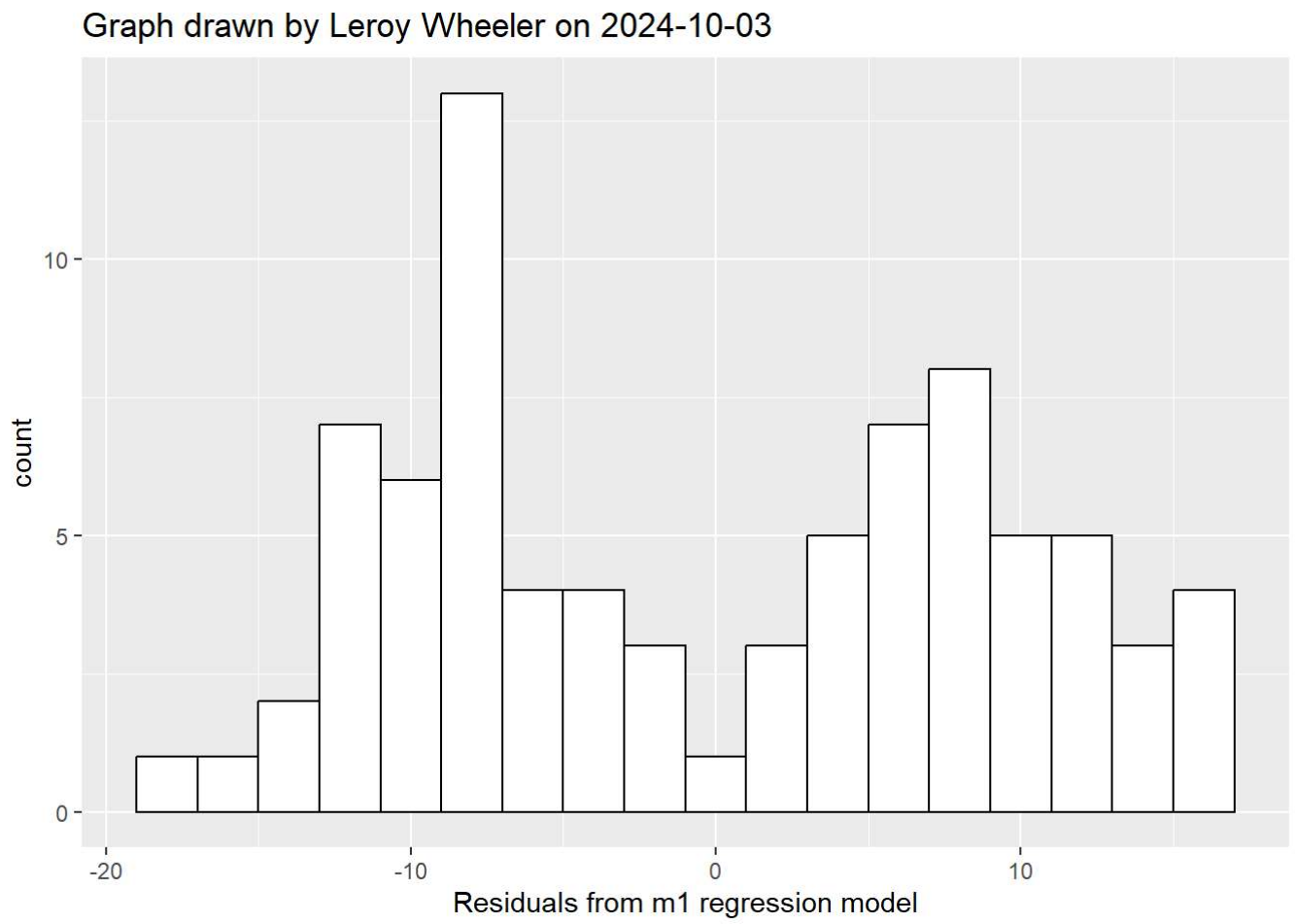



The qq plot does not show a straight line suggesting that that the residuals are NOT normally distributed. This violates our assumptions so we might have to use a different model.

## Question 8b: Histogram of residuals

```
r1 |>
    ggplot(aes(.resid)) +
      geom_histogram(
        binwidth=2,
        color="black",
```

```
    fill="white") +
ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03") +
  xlab("Residuals from m1 regression model")
```




The histogram clearly shows that the residuals in our linear regression model are NOT normally distributed.

## Question 9a: Calculate descriptive statistics (mean, standard deviation, minimum, and maximum) for mom_age. Interpret these values.

```
bf |>
    summarize(
        mean_mom_age=mean(mom_age, na.rm=TRUE),
        sd_mom_age=sd(mom_age, na.rm=TRUE),
        min_mom_age=min(mom_age, na.rm=TRUE),
        max_mom_age=max(mom_age, na.rm=TRUE),
        n_missing=sum(is.na(mom_age))) |>
    data.frame()
```

```
  mean_mom_age sd_mom_age min_mom_age max_mom_age n_missing
1     27.33333   6.784698          16          44         0
```

The mean age of the mothers participating in this study is about 27 years with a standard deviation of about 6 and a half. The youngest participant is 16yrs and the oldest participant is 44 yrs old. There are no missing values of mothers age in this data set.

## Question 9b: Calculate descriptive statistics (mean, standard deviation, minimum, and maximum) for para. Interpret these values.

```
bf |>
    summarize(
        mean_para=mean(para, na.rm=TRUE),
        sd_para=sd(para, na.rm=TRUE),
        min_para=min(para, na.rm=TRUE),
        max_para=max(para, na.rm=TRUE),
        n_missing=sum(is.na(para))) |>
    data.frame()
```

```
  mean_para   sd_para min_para max_para n_missing
1  1.964286 0.9993544        1        5         0
```

Para stands for parity or in other words the number of live births that each mother have had. The mean parity for the participants is about 2 live births, with a standard deviation of 1 live birth. The minimum number of live births for these mothers is one while the maximum number is 5. There are no missing values in this data set for parity.

## Question 10: Calculate the correlations between mom_age, para, and age_stop. Interpret these values.

```
bf |>
  select(mom_age, para, age_stop) |>
  cor(use='complete.obs')
```

```
           mom_age       para   age_stop
mom_age  1.0000000 0.42446352 0.25901640
para     0.4244635 1.00000000 0.02361115
age_stop 0.2590164 0.02361115 1.00000000
```

There is a weak correlation between the mothers age and breast feeding time (0.25). There is virtually zero correlation between parity and breast feeding time (0.02).

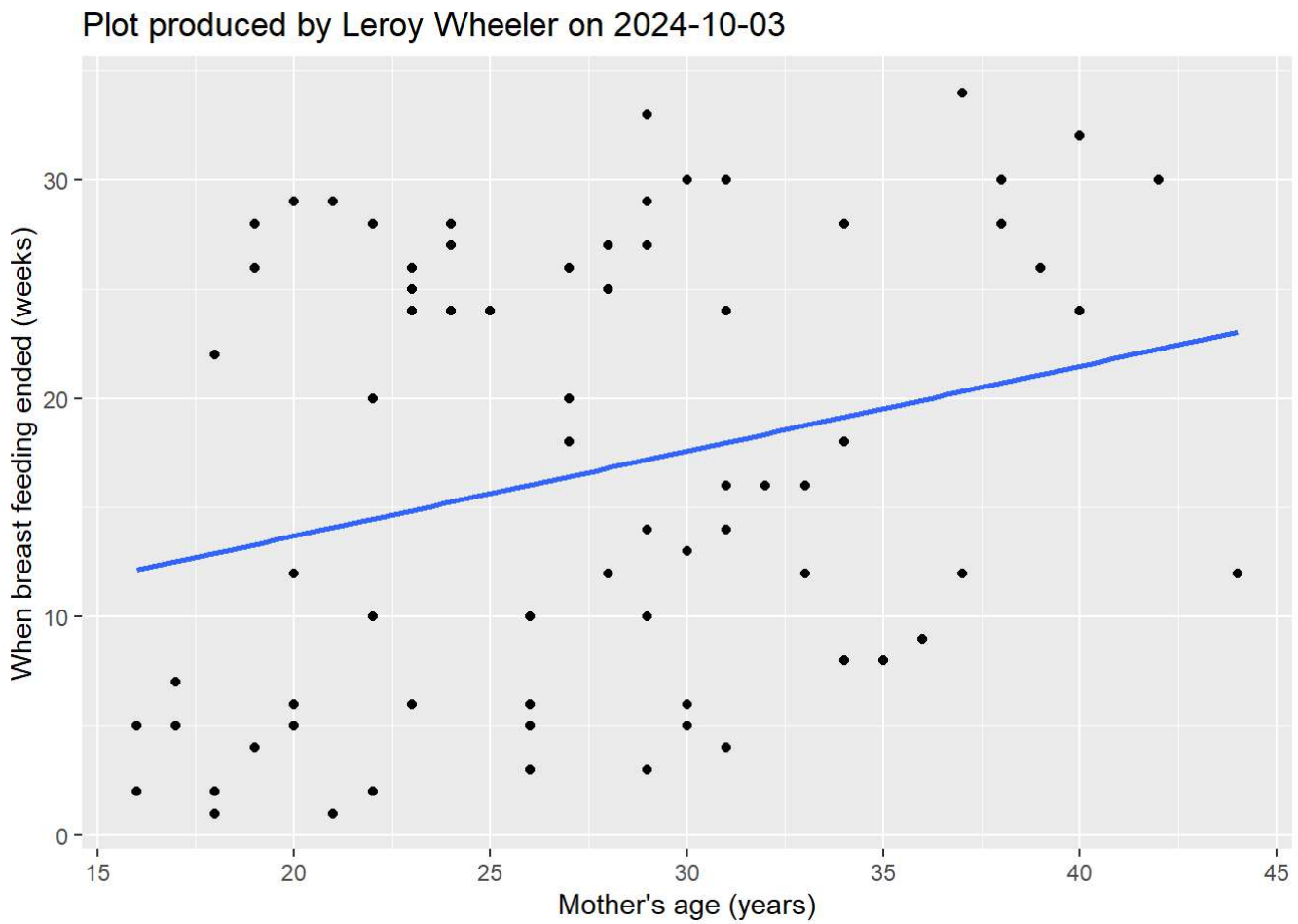
## Question 11a: Draw a scatterplot with mom_age on the x-axis and age_stop on the y-axis. Interpret these plots.

```
bf |>
    ggplot(aes(mom_age, age_stop)) +
      geom_point() +
      xlab("Mother's age (years)") +
      ylab("When breast feeding ended (weeks)") +
    geom_smooth(method="lm", se=FALSE) +
    ggtitle("Plot produced by Leroy Wheeler on 2024-10-03")
```

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

Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```




It looks like there is a positive trend where older mothers tend to breast feed their infants longer.

## Question 11b: Draw a scatterplot with para on the x-axis and age_stop on the y-axis. Interpret these plots.

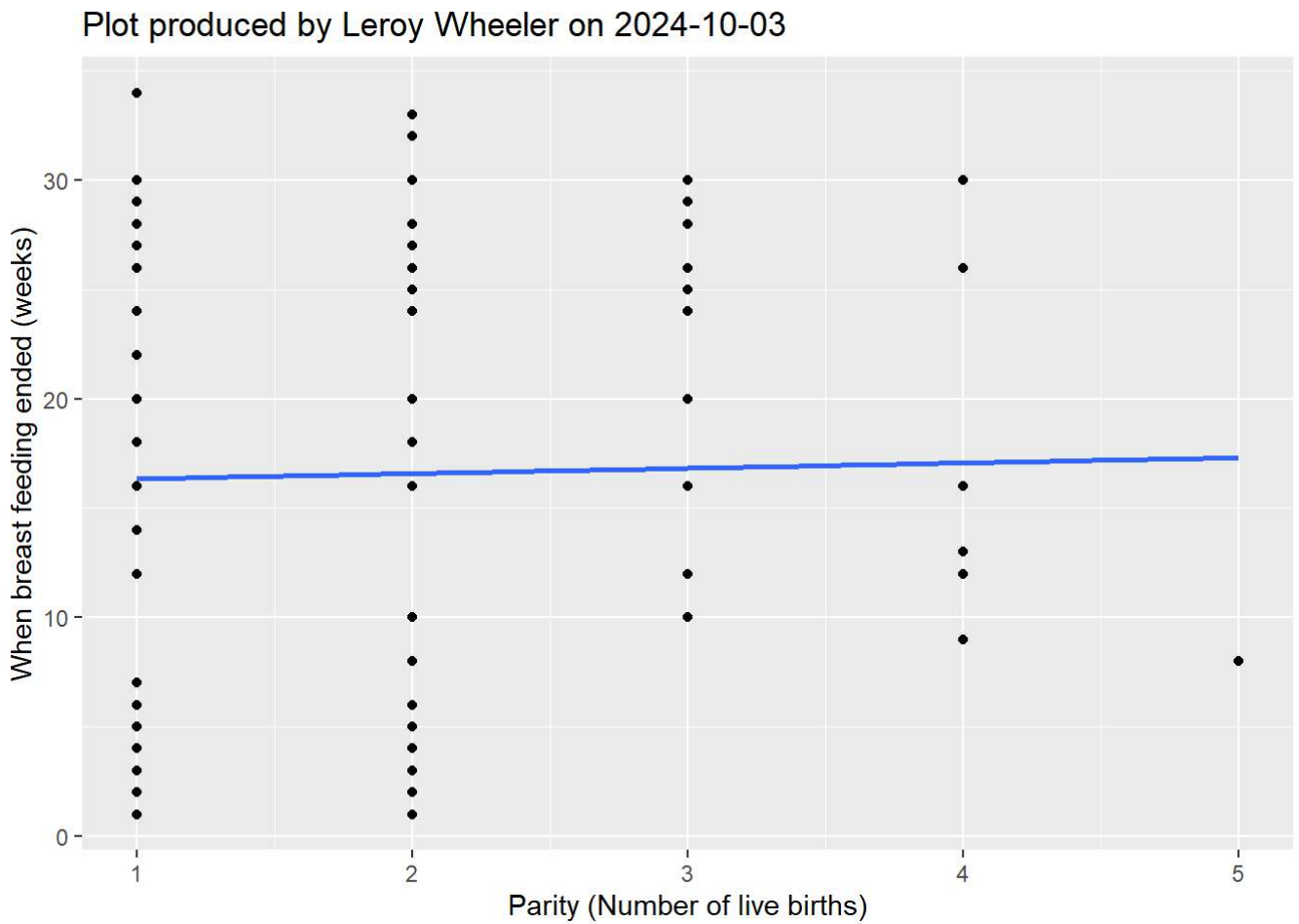
```
bf |>
    ggplot(aes(para, age_stop)) +
      geom_point() +
      xlab("Parity (Number of live births)") +
      ylab("When breast feeding ended (weeks)") +
```

```
    geom_smooth(method="lm", se=FALSE) +
    ggtitle("Plot produced by Leroy Wheeler on 2024-10-03")
```

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

Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```



There seems to be a very weak relationship between parity and the length of time the infants continued to breast feed.

## Question 12: Compute a linear regression model using mom_age and para to predict age_stop. Interpret the regression coefficients.

```
m2 <- lm(age_stop ~ mom_age + para, data=bf)
m2
```

```
Call:
lm(formula = age_stop ~ mom_age + para, data = bf)

Coefficients:
(Intercept)      mom_age         para
     6.2233       0.4562      -1.0786
```

With parity held constant, for every year older the mother was, the infant was breast fed for about half a week longer. With the age of the mother held constant, for every additional child the mother gave birth to, there was one less week that the new infant was breast fed.

## Question 13: Compute R-squared for this regression model. Interpret this number.

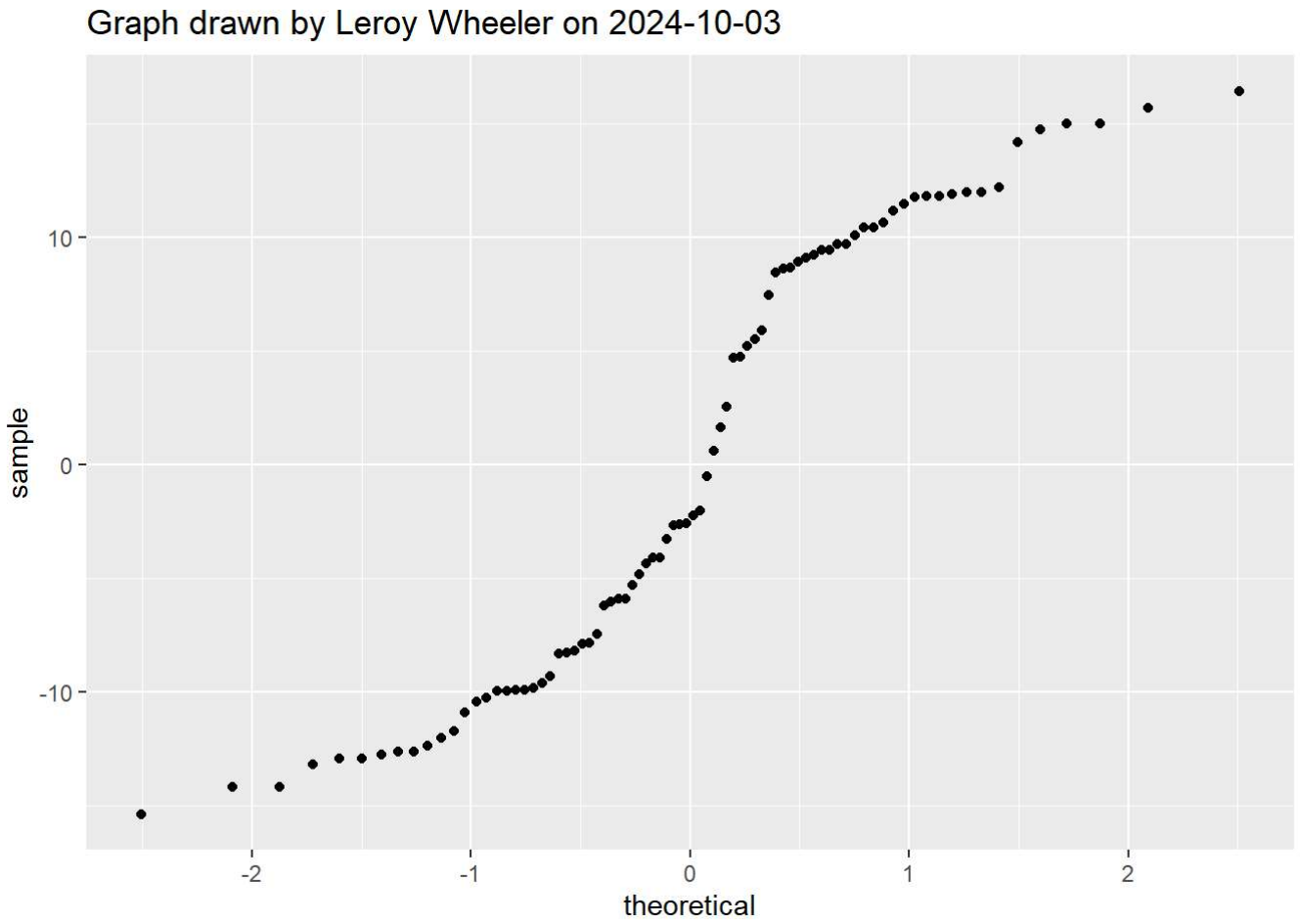
```
glance(m2)$r.squared
```

```
[1] 0.07618063
```

The R squared value was 0.076 which means only about 8% of the total breast feeding time can be explained by a combination of the mothers age and her parity.

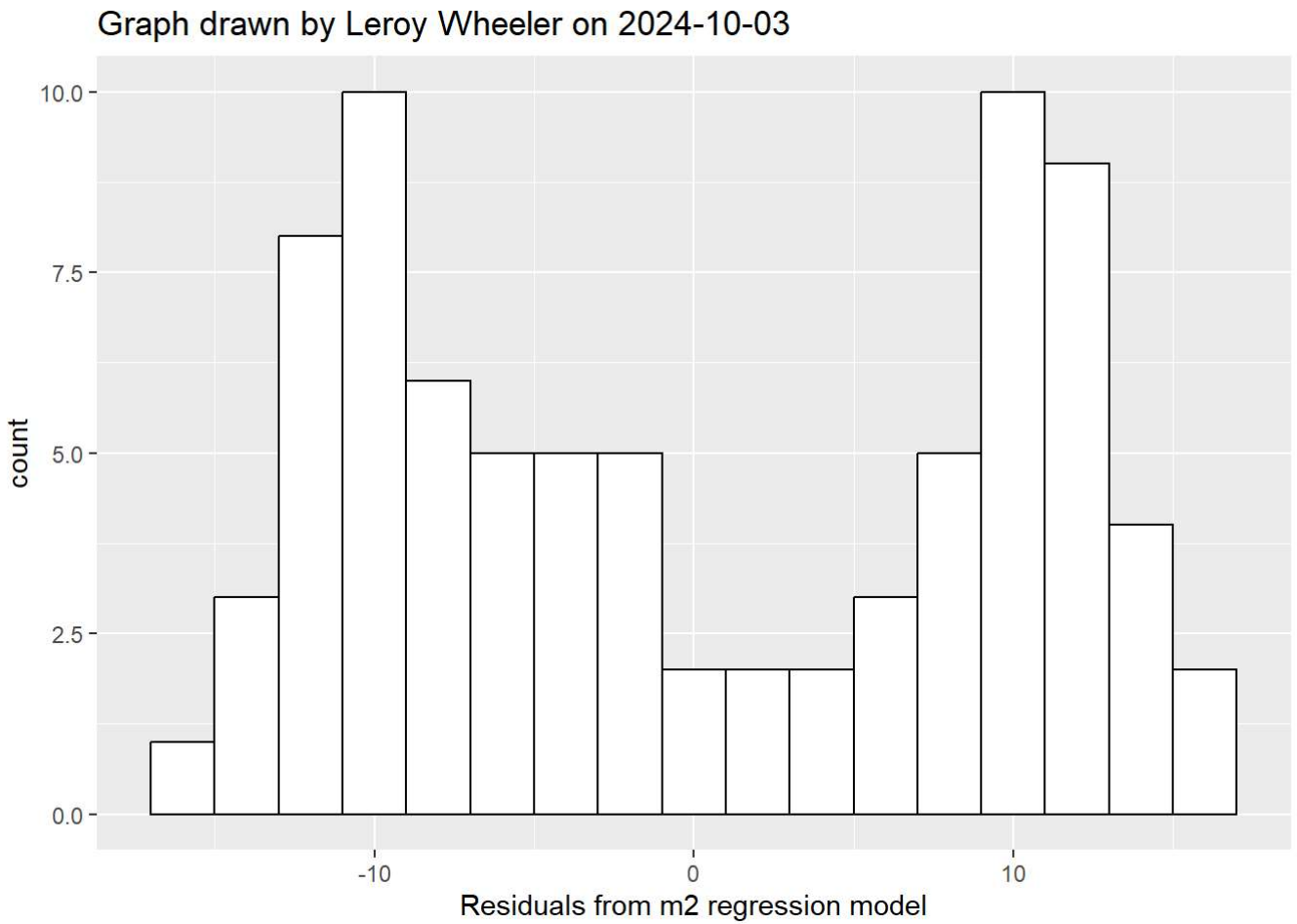
## Question 14: Draw a normal probability plot and a histogram of the residuals. Interpret these plots.

```
r2 <- augment(m2)
r2 |>
    ggplot(aes(sample=.resid)) +
      stat_qq() +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03")
```



The qq plot testing for normality of the residuals do not show a linear relationship, therefore the assumption of residual normality is not satisfied. We may need to use a different model.

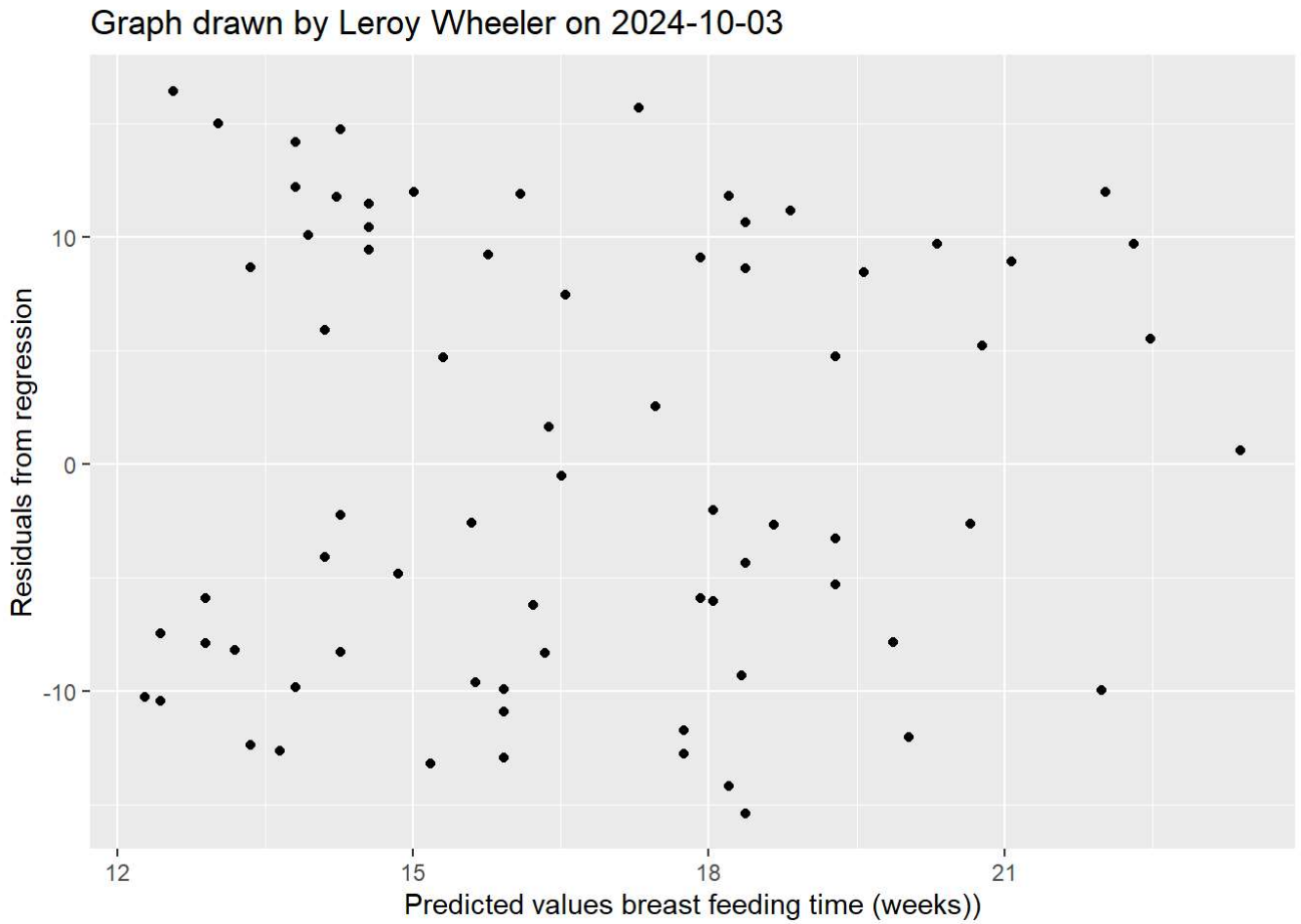
```
r2 |>
    ggplot(aes(.resid)) +
      geom_histogram(
        binwidth=2,
        color="black",
        fill="white") +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03") +
      xlab("Residuals from m2 regression model")
```



The histogram of the residuals from our linear regression model confirms that the residuals from the model do not show a normal distribution.

## Question 15: Draw a plot with the predicted values on the x-axis and the residuals on the y-axis. Is there any evidence of heterogeneity or non-linearity?

```
r2 |>
  ggplot(aes(.fitted, .resid)) +
    geom_point() +
    xlab("Predicted values breast feeding time (weeks))") +
    ylab("Residuals from regression") +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-10-03")
```




The residuals from this model look pretty homogeneous and linear along the x-axis.

## Question 16: Display any extreme values for leverage (greater than 3*3/n), studentized deleted residuals (absolute value greater than 3), and for Cook's distance (greater than 1). Explain why these values are extreme.

## Leverage

```
n <- nrow(r2)
r2 |> filter(.hat > 3*3/n)
```

```
# A tibble: 1 × 10
  .rownames age_stop mom_age  para .fitted .resid  .hat .sigma .cooksd
  <chr>        <dbl>   <dbl> <dbl>   <dbl>  <dbl> <dbl>  <dbl>   <dbl>
1 50               8      34     5    16.3  -8.34 0.126   9.98  0.0384
# ℹ 1 more variable: .std.resid <dbl>
```

The mother from row 50 has a combined value of parity plus mothers age which we may want to investigate.

## Studentized deleted residuals

```
r2 |>
  filter(abs(.std.resid) > 3)
```

```
# A tibble: 0 × 10
# ℹ 10 variables: .rownames <chr>, age_stop <dbl>, mom_age <dbl>, para <dbl>,
#   .fitted <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
#   .std.resid <dbl>
```

There are no predicted values of weeks breast fed that are beyond 3 standard deviations.

## Cook's test for infulential data points

```
r2 |>
  filter(.cooksd > 1)
```

```
# A tibble: 0 × 10
# ℹ 10 variables: .rownames <chr>, age_stop <dbl>, mom_age <dbl>, para <dbl>,
```

```
#   .fitted <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
#   .std.resid <dbl>
```

There are no specific observations from the data set which exert an out-sized influence on our model.