Linear regression modules using the breast feeding dataset

AUTHOR
Michael Dang

PUBLISHED
October 9, 2024

Libraries

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

Question 1

Read the data

```
pulmonary <- 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(pulmonary)
```

```
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, …
```

Question 2

Compute descriptive statistics (counts and percentages) for feed_type. Interpret these values.

```
pulmonary |>
    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
```

The percentages indicate that approximately 53.57% of the dataset is categorized under Control, and 46.43% under Treatment.

Question 3

Compute descriptive statistics (mean, standard deviation, minimum, and maximum) for age_stop. Interpret these values. Note that there are some missing values for age_stop. This means that you need to include the option na.rm=TRUE in your code.

```
pulmonary |>
  summarize(
    age_stop_mn=mean(age_stop, na.rm = TRUE),
```

```
            age_stop_sd=sd(age_stop, na.rm = TRUE),
            age_stop_min=min(age_stop, na.rm = TRUE),
            age_stop_max=max(age_stop, na.rm = TRUE),
            count = sum(!is.na(age_stop))
            )
```

```
# A tibble: 1 × 5
  age_stop_mn age_stop_sd age_stop_min age_stop_max count
        <dbl>       <dbl>        <dbl>        <dbl> <int>
1        16.6        10.2            1           34    82
```

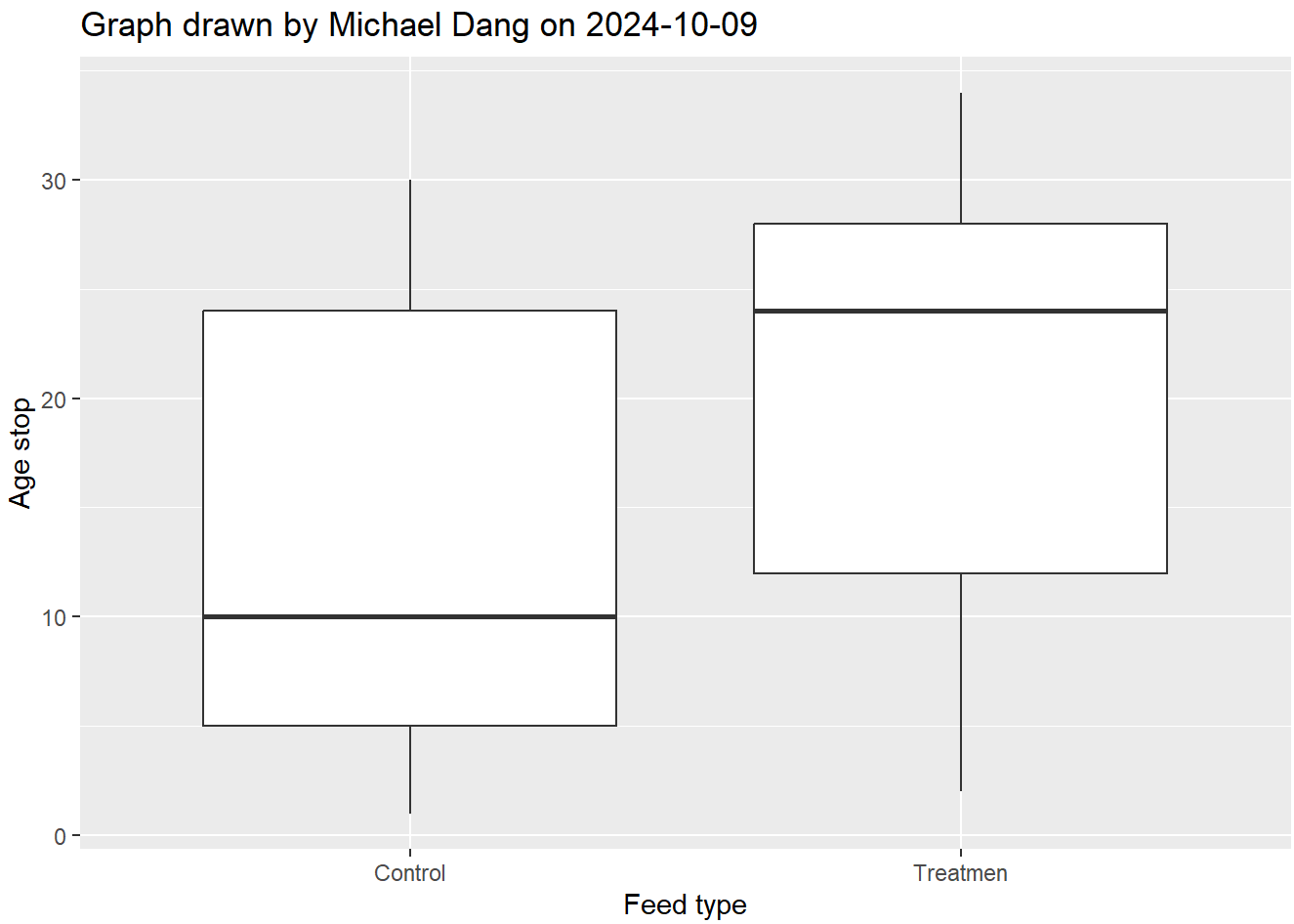
The mean age at which breastfeeding stopped is 16.5 months. The standard deviation is 10.2 months, indicating a wide spread in the stopping age. The minimum age is 1 months and the maximum age is 34 months. There are 82 non-missing values for age_stop.

Question 4

Draw a boxplot comparing age_stop for each level of feed_type. Interpret this plot

```
        pulmonary |>
          ggplot(aes(age_stop, feed_type)) +
            geom_boxplot() +
            coord_flip() +
            ggtitle("Graph drawn by Michael Dang on 2024-10-09") +
              xlab("Age stop") +
              ylab("Feed type")
```

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



The boxplot shows a slightly higher median for the Treatment group compared to the Control group, this indicates that on average, breastfeeding stopped later in the Treatment group.

Question 5

Calculate the means and standard deviations of age_stop for each level of feed_type. Interpret these numbers.

```
pulmonary |>
  group_by(feed_type) |>
    summarise(
    mean_age_stop = mean(age_stop, na.rm = TRUE),
    sd_age_stop = sd(age_stop, na.rm = TRUE),
    count = n()
  )
```

```
# A tibble: 2 × 4
  feed_type mean_age_stop sd_age_stop count
  <chr>             <dbl>       <dbl> <int>
1 Control            13.3        9.98    46
2 Treatmen           20.4        9.30    38
```

The Treatment group stopped breastfeeding later, on average (mean = 20.3 months), compared to the Control group (mean = 13.3 months). The Treatment group also shows more variability (SD = 9.2) in when breastfeeding stopped, while the Control group has less variability (SD = 9.9), meaning the ages at which breastfeeding stopped are more consistent in the Control group.

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=pulmonary)
m1
```

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

Coefficients:
      (Intercept)  feed_typeTreatmen
            13.32               7.05
```

The average estimate age at which breastfeeding stopped in the Control group is 13.32 months. The slope is 7.05, it means that on average, breastfeeding stopped 7.05 months later in the Treatment group compared to the Control group.

Question 7

Compute R-squared for this regression model. Interpret this number.

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

```
[1] 0.1192946
```

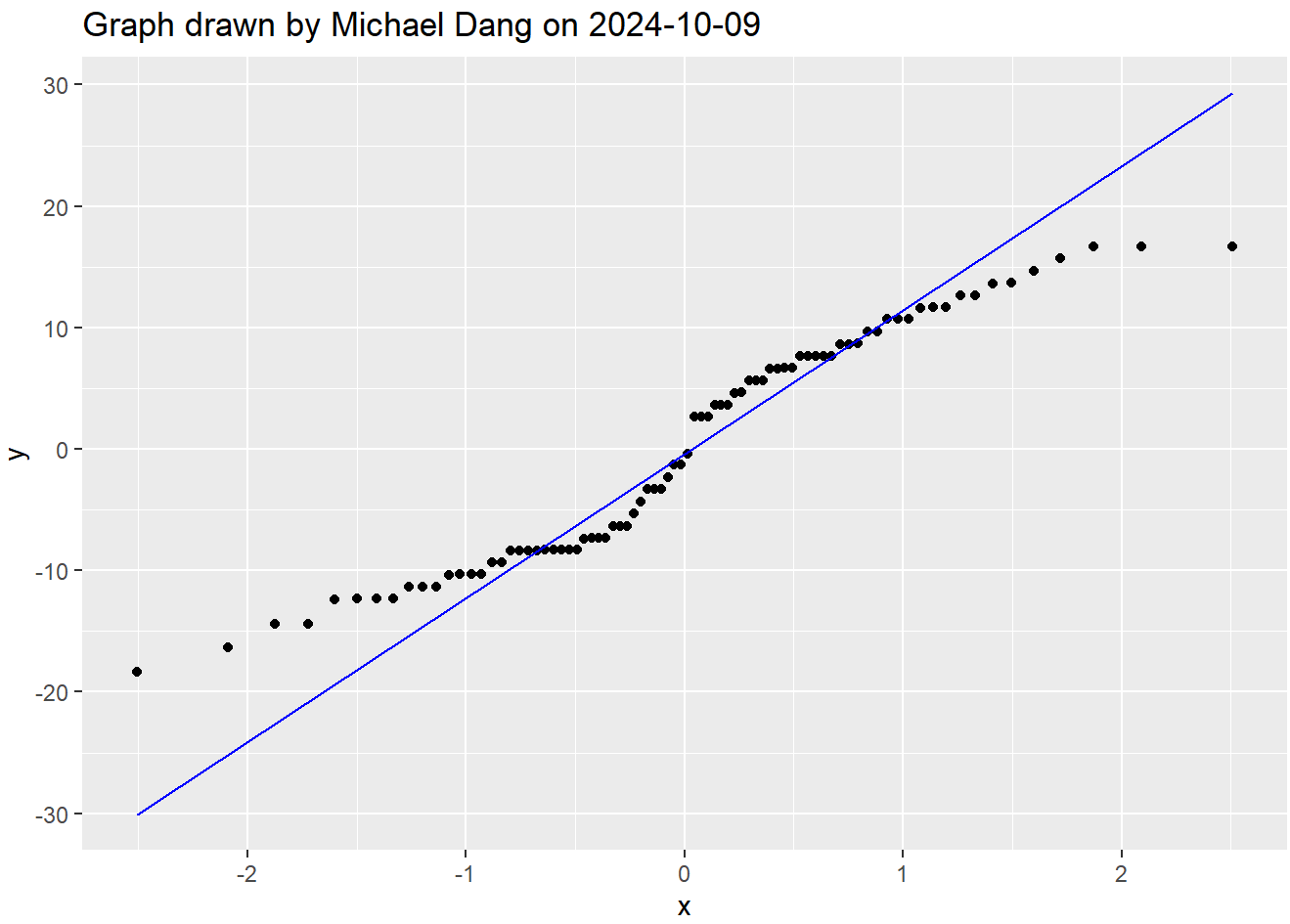
This means that 11.93% of the variability in age_stop can be explained by the variable feed_type. In other words, feed_type has a relatively small effect on predicting when breastfeeding stops, and there is still 88.07% of the variability in age_stop that is not explained by this model.

Question 8a

Draw a normal probability plot for the residuals from this regression model. Interpret this plot.

```
r1 <- augment(m1)
r1 |>
    ggplot(aes(sample=.resid)) +
```

```
  stat_qq() +
stat_qq_line(col = "blue") +
ggtitle("Graph drawn by Michael Dang on 2024-10-09")
```

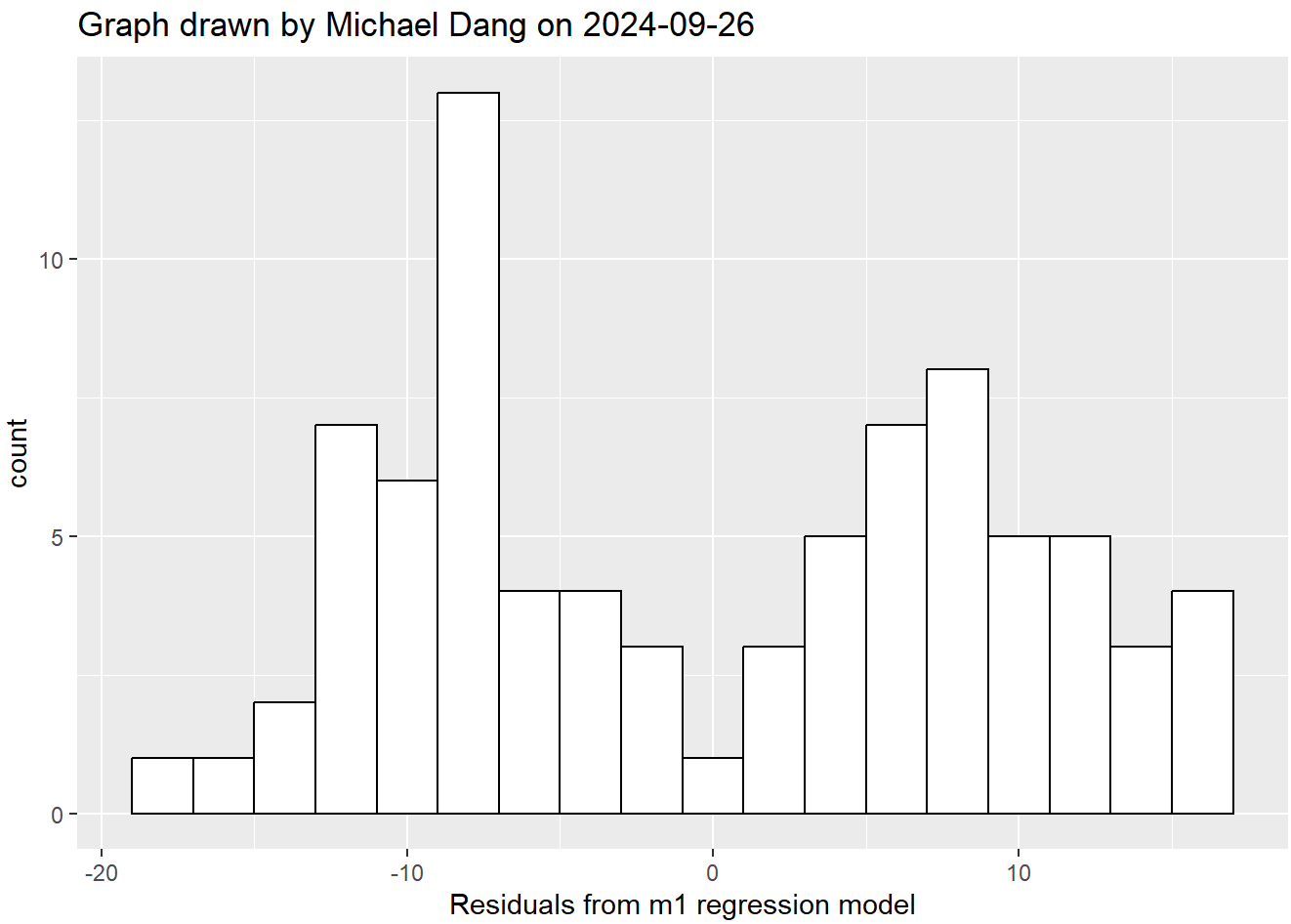


The middle portion of the plot (between -1 and 1 on the x-axis) shows the points following the reference line fairly closely, which suggests that the residuals are approximately normally distributed in this range.

Question 8b

Draw a histogram for the residuals from this regression model. Interpret this plot.

```
r1 |>
    ggplot(aes(.resid)) +
      geom_histogram(
        binwidth=2,
        color="black",
        fill="white") +
    ggtitle("Graph drawn by Michael Dang on 2024-09-26") +
      xlab("Residuals from m1 regression model")
```



The histogram of the residuals shows a bimodal distribution with peaks around -10 and 10, which suggests that the residuals are not normally distributed.

Question 9

Calculate descriptive statistics (mean, standard deviation, minimum, and maximum) for mom_age and para. Interpret these values.

```
pulmonary |>
  summarise(
    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),
    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)
  )
```

```
# A tibble: 1 × 8
  mean_mom_age sd_mom_age min_mom_age max_mom_age mean_para sd_para min_para
```

```
         <dbl>      <dbl>       <dbl>       <dbl>     <dbl>   <dbl>    <dbl>
1         27.3       6.78          16          44      1.96   0.999        1
# ℹ 1 more variable: max_para <dbl>
```

The mean mother's age is 27.3 years, with a standard deviation of 6.7 years, indicating that most mothers are close to 30, but there is moderate variation. The youngest mother is 16 years old and the oldest is 44 years old.

The mean number of live births is 1.9, with a standard deviation of 0.9, suggesting that most mothers have had around 2 live births, but there is some variability. The minimum number of live births is 1, and the maximum is 5.

Question 10

Calculate the correlations between mom_age, para, and age_stop. Interpret these values. Note: because there are missing values, you need to change the function from cor() to cor(use="complete.obs").

```
         cor(pulmonary |> select(mom_age, para, age_stop), 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
```

mom_age and para (0.424): There is a moderate positive correlation between a mother's age and the number of live births (parity).

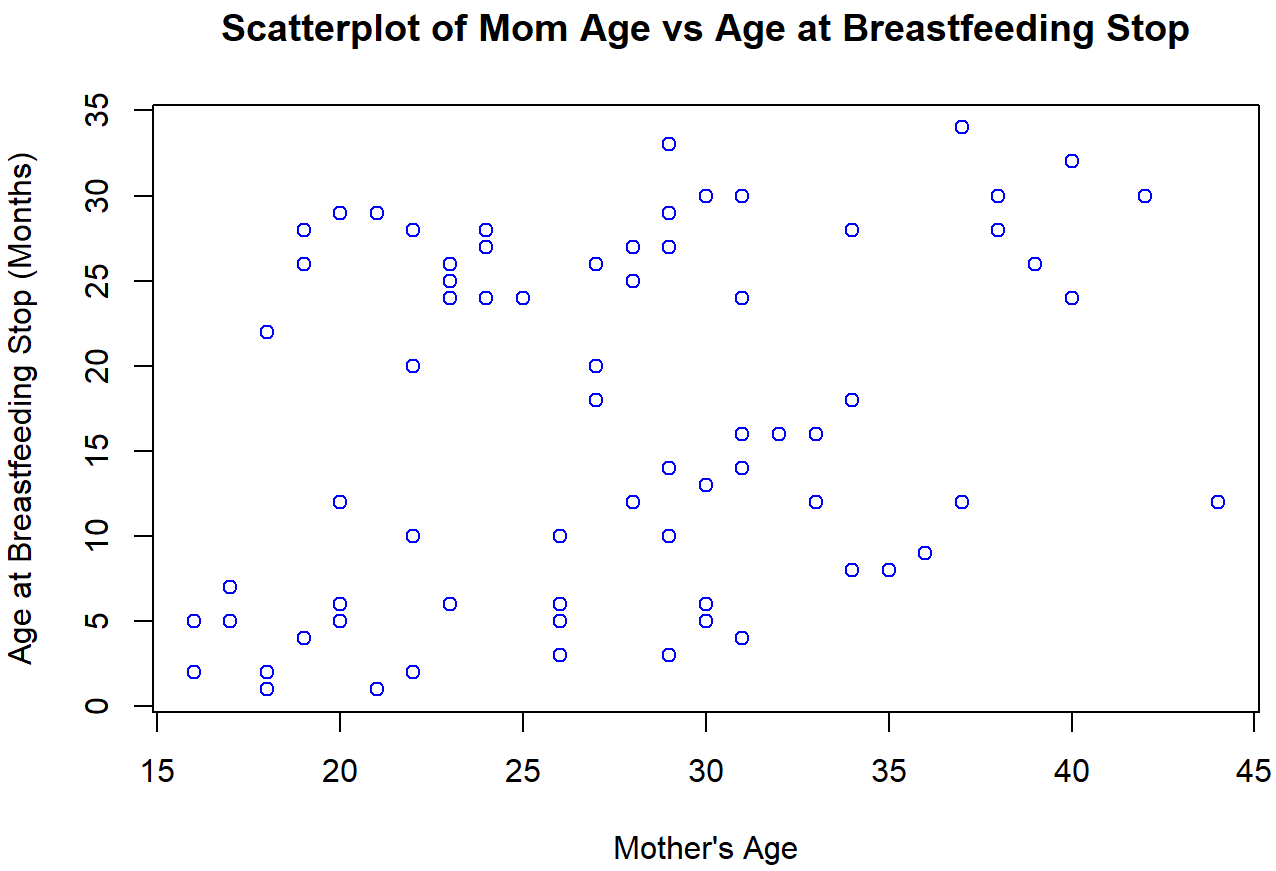
mom_age and age_stop (0.259): There is a weak positive correlation between the mother's age and the age at which breastfeeding stops.

para and age_stop (0.024): There is virtually no correlation between the number of live births (parity) and the age at which breastfeeding stops.

Question 11a

Draw a scatterplot with mom_age on the x-axis and age_stop on the y-axis. Interpret this plot.

```
         plot(pulmonary$mom_age, pulmonary$age_stop,
              main = "Scatterplot of Mom Age vs Age at Breastfeeding Stop",
              xlab = "Mother's Age",
              ylab = "Age at Breastfeeding Stop (Months)",
              col = "blue")
```

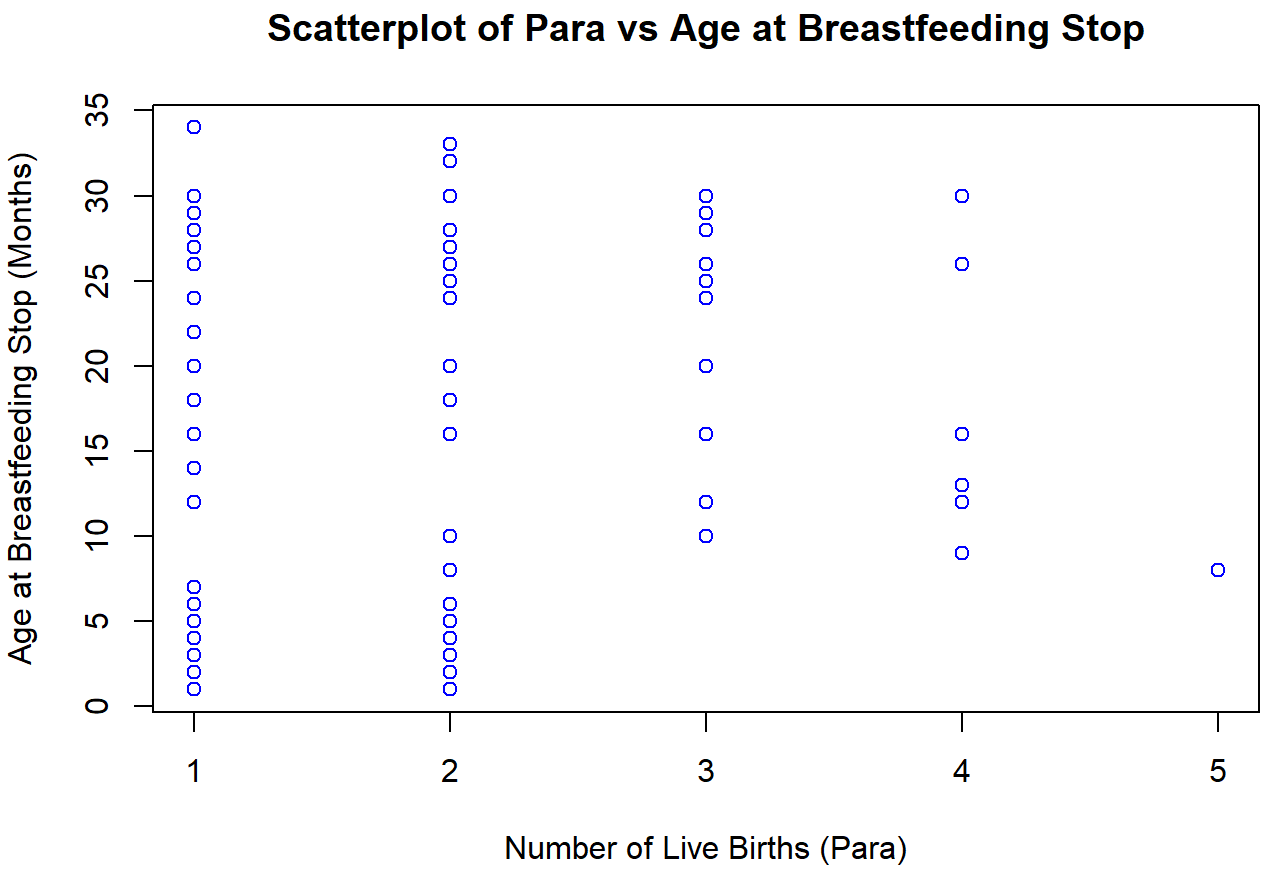


There is no clear linear pattern or strong trend in the data. The points are fairly scattered.

Question 11b

Draw a scatterplot with para on the x-axis and age_stop on the y-axis. Interpret this plot.

```
plot(pulmonary$para, pulmonary$age_stop,
     main = "Scatterplot of Para vs Age at Breastfeeding Stop",
     xlab = "Number of Live Births (Para)",
     ylab = "Age at Breastfeeding Stop (Months)",
     col = "blue")
```



The scatterplot suggests that parity (number of live births) does not significantly predict the age at which breastfeeding stops.

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 = pulmonary)
m2
```

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

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

When both mom_age (mother's age) and para (number of live births) are 0, the predicted age at which breastfeeding stops is 6.22 months. For every additional year of the mother's age, the expected age at

which breastfeeding stops increases by 0.456 months and for every additional live birth, the expected age at which breastfeeding stops decreases by 1.08 months,

Question 13

Compute R-squared for this regression model. Interpret this number.

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

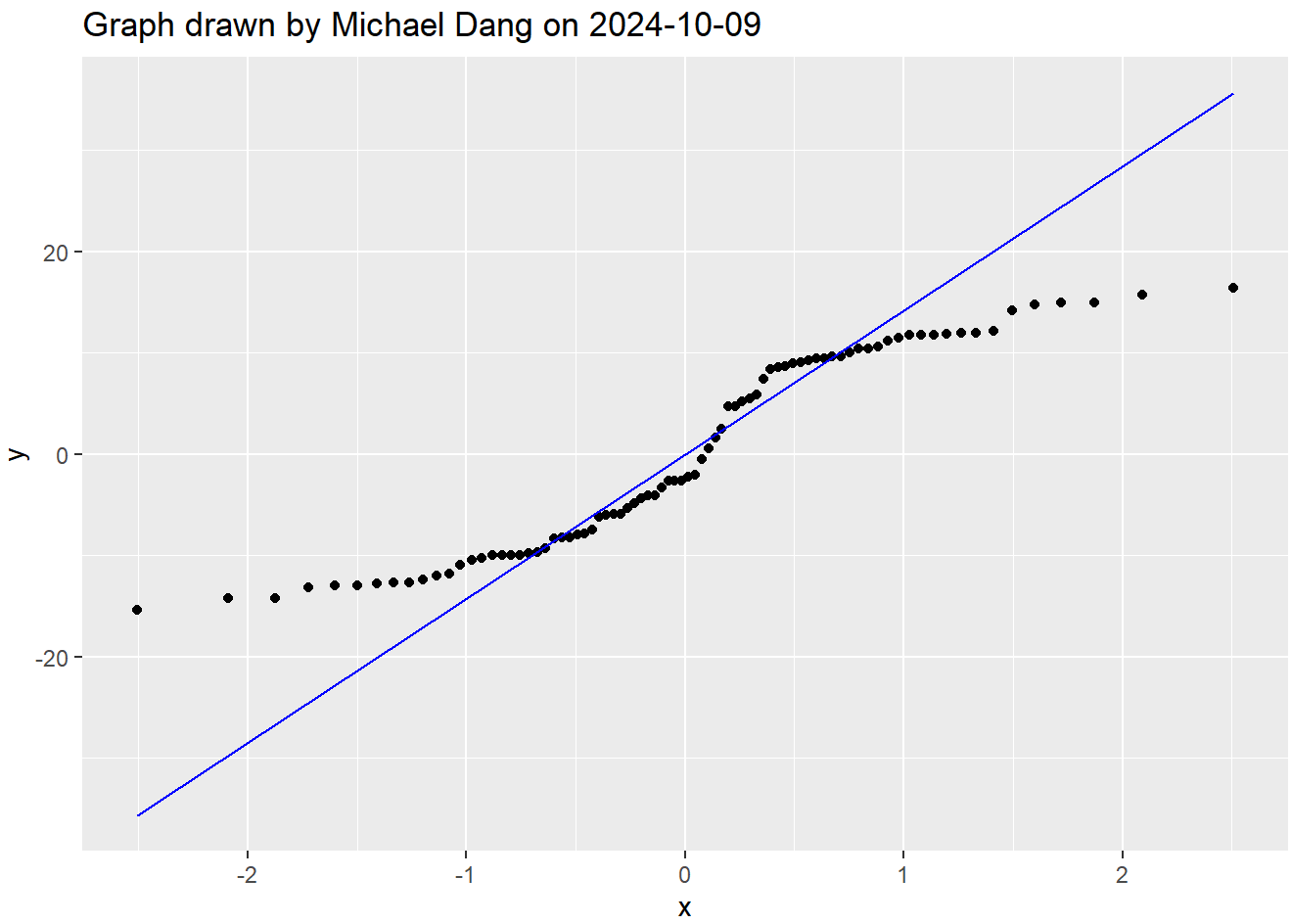
```
[1] 0.07618063
```

The R_squared is 0.076 this mean 7.6% of the variability in the age at which breastfeeding stops (age_stop) can be explained by the combination of the mother's age (mom_age) and the number of live births (para).

Question 14a

Draw a normal probability plot of the residuals. Interpret this plot.

```
r2 <- augment(m2)
r2 |>
    ggplot(aes(sample=.resid)) +
      stat_qq() +
    stat_qq_line(col = "blue") +
    ggtitle("Graph drawn by Michael Dang on 2024-10-09")
```

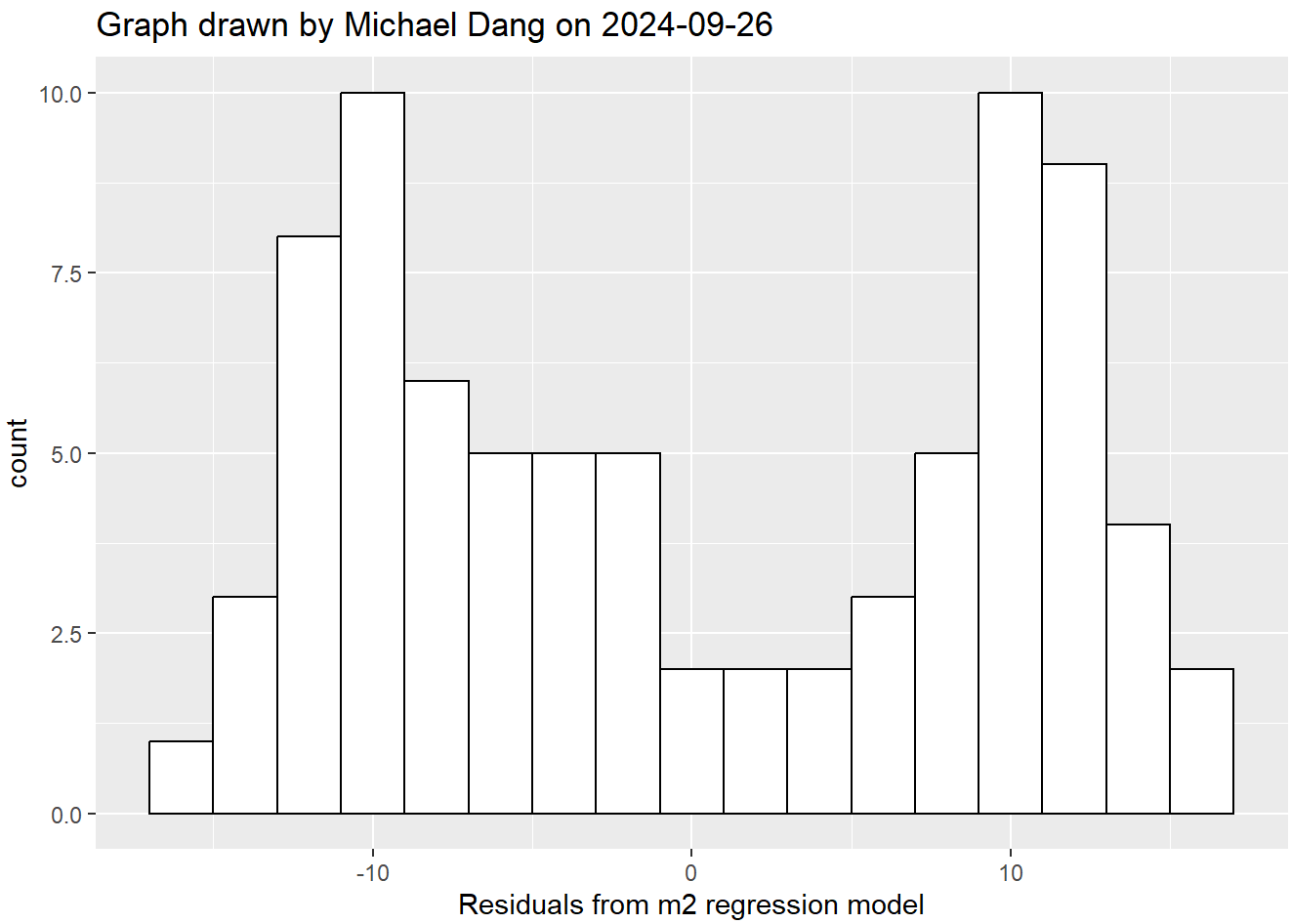


The points mostly follow the blue reference line in the middle portion of the plot, indicating that the residuals are approximately normally distributed in this range.

Question 14b

Draw a histogram of the residuals. Interpret this plot.

```
r2 |>
    ggplot(aes(.resid)) +
      geom_histogram(
        binwidth=2,
        color="black",
        fill="white") +
    ggtitle("Graph drawn by Michael Dang on 2024-09-26") +
      xlab("Residuals from m2 regression model")
```

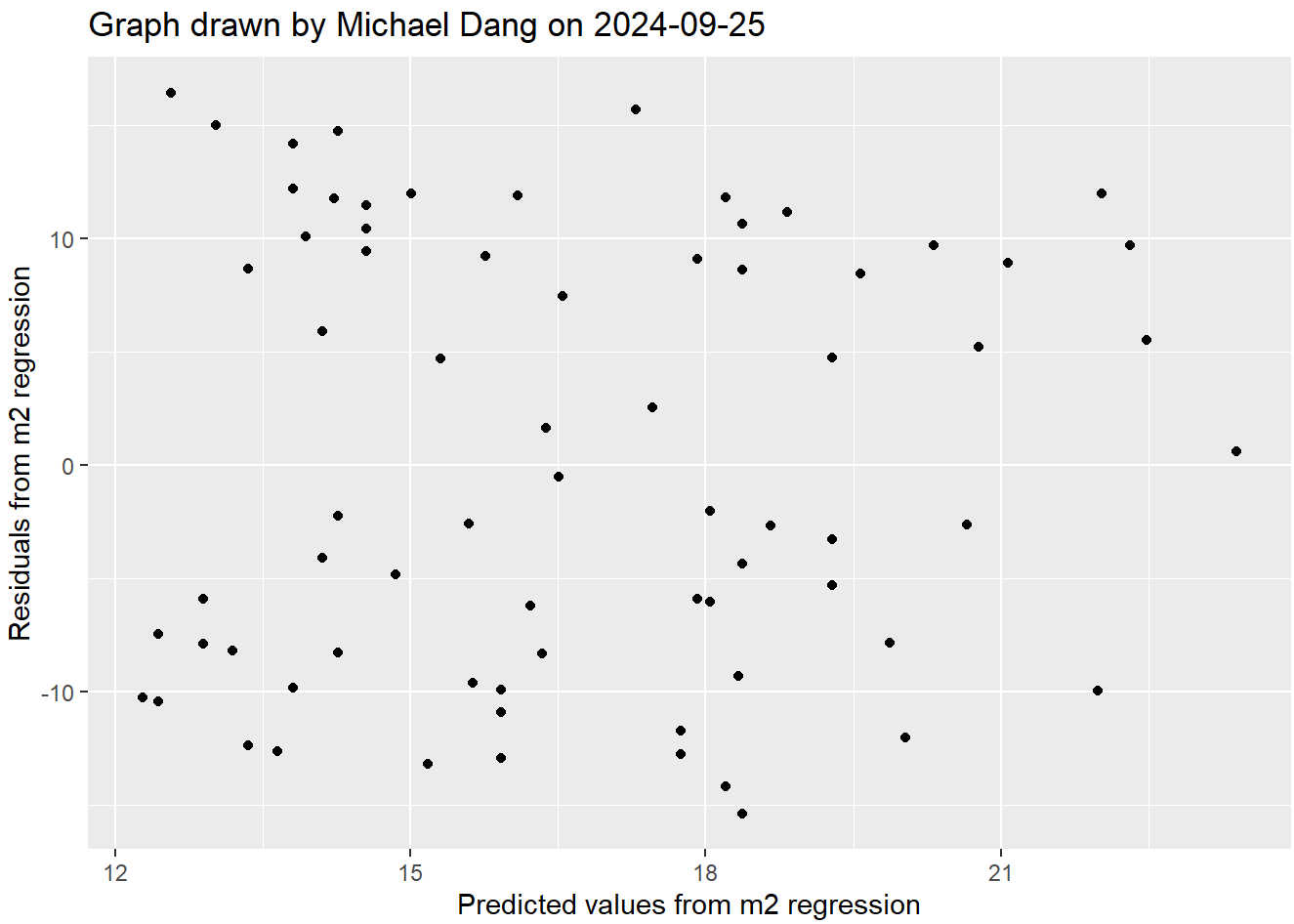


The distribution appears to be bimodal, with two distinct peaks around -10 and 10.

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 from m2 regression") +
    ylab("Residuals from m2 regression") +
    ggtitle("Graph drawn by Michael Dang on 2024-09-25")
```



There is no clear pattern of increasing or decreasing spread of residuals as the predicted values increase. The residuals seem to be fairly evenly distributed around 0, suggesting no strong evidence of heteroscedasticity.

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.

Influential data points, 1

```
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>
```

Influential data points, 2

```
        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>
```

Influential data points, 3

```
        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>
```

Large residual (-8.34): The residual is quite large in magnitude, meaning the actual age_stop is 8 months less than what the model predicted. This indicates that the model is not fitting this particular observation well.

Leverage (0.1257): The leverage for this observation is higher than the average but still below the threshold for being considered extreme (3*3/n).

While this observation has a large residual, it does not have extreme leverage or Cook's distance, meaning that although the model did not fit this observation well (the prediction was 8 months off), it does not heavily influence the model overall.