Analysis of breast feeding study

AUTHOR
Michael Dang

PUBLISHED
September 22, 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.

Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages and warnings are suppressed.

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

Read the data and view a brief summary

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
```

Calculate statistics for mother's age

```
        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
```

This is a reasonable distribution of ages. If you saw mothers much younger than 16 years or much older than 44 years, that might raise some concerns about the data.

Calculate statistics for age_stop

```
        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 maximum value, 34 weeks, was a bit of concern for me, because the study was a six month study, which would imply the largest value would be 24 or 26. But I was told that breast feeding duration included time in the hospital, which could easily be as long as 8 or 10 weeks for a pre-term infant.

Question 1

```
        bf |>
            summarize(
                mean_gest_age=mean(gest_age, na.rm=TRUE),
                sd_gest_age=sd(gest_age, na.rm=TRUE),
                min_gest_age=min(gest_age, na.rm=TRUE),
                max_gest_age=max(gest_age, na.rm=TRUE),
                n_missing=sum(is.na(gest_age))) |>
            data.frame()
```

```
  mean_gest_age sd_gest_age min_gest_age max_gest_age n_missing
1      31.84524    2.026892           26           35         0
```

The average gestational age in this dataset is around 32 weeks, with a range from 26 to 35 weeks. The standard deviation of 2 weeks indicates moderate variability in gestational ages. There are no missing values in the dataset for this variable.

Question 2

```
        bf |>
            summarize(
                mean_dc_age=mean(dc_age, na.rm=TRUE),
                sd_dc_age=sd(dc_age, na.rm=TRUE),
                min_dc_age=min(dc_age, na.rm=TRUE),
                max_dc_age=max(dc_age, na.rm=TRUE),
```

```
            n_missing=sum(is.na(dc_age))) |>
          data.frame()
```

```
  mean_dc_age sd_dc_age min_dc_age max_dc_age n_missing
1    33.72619  17.25385          8         77         0
```

The average age at discharge is approximately 34 days, with a wide range from 8 to 77 days. The relatively high standard deviation of 17 days indicates a considerable variation in the discharge ages among the patients. There are no missing values for this variable.

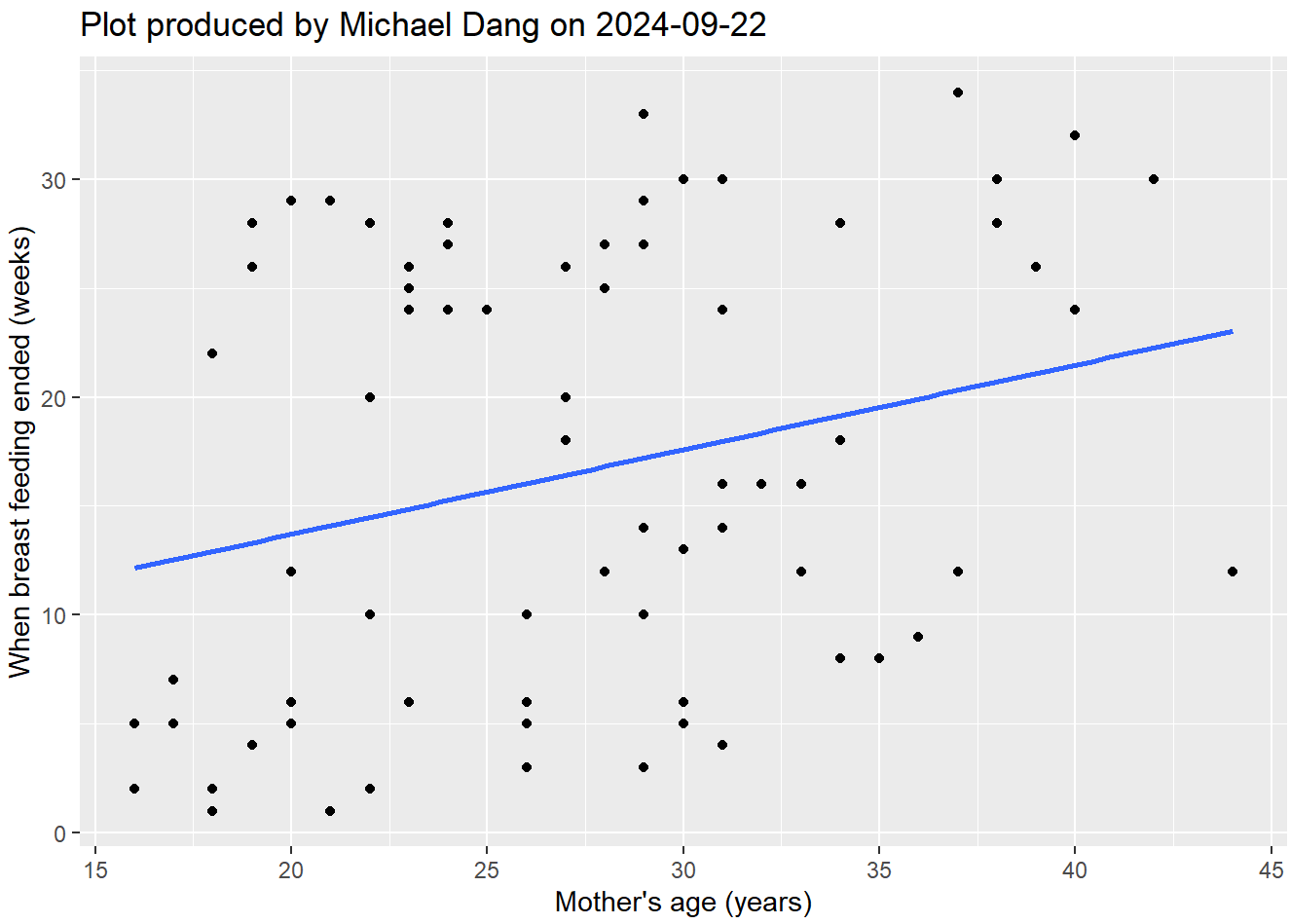
Plot mother's age and age when breast feeding stopped

```
        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 Michael Dang on 2024-09-22")
```

```
`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()`).
```

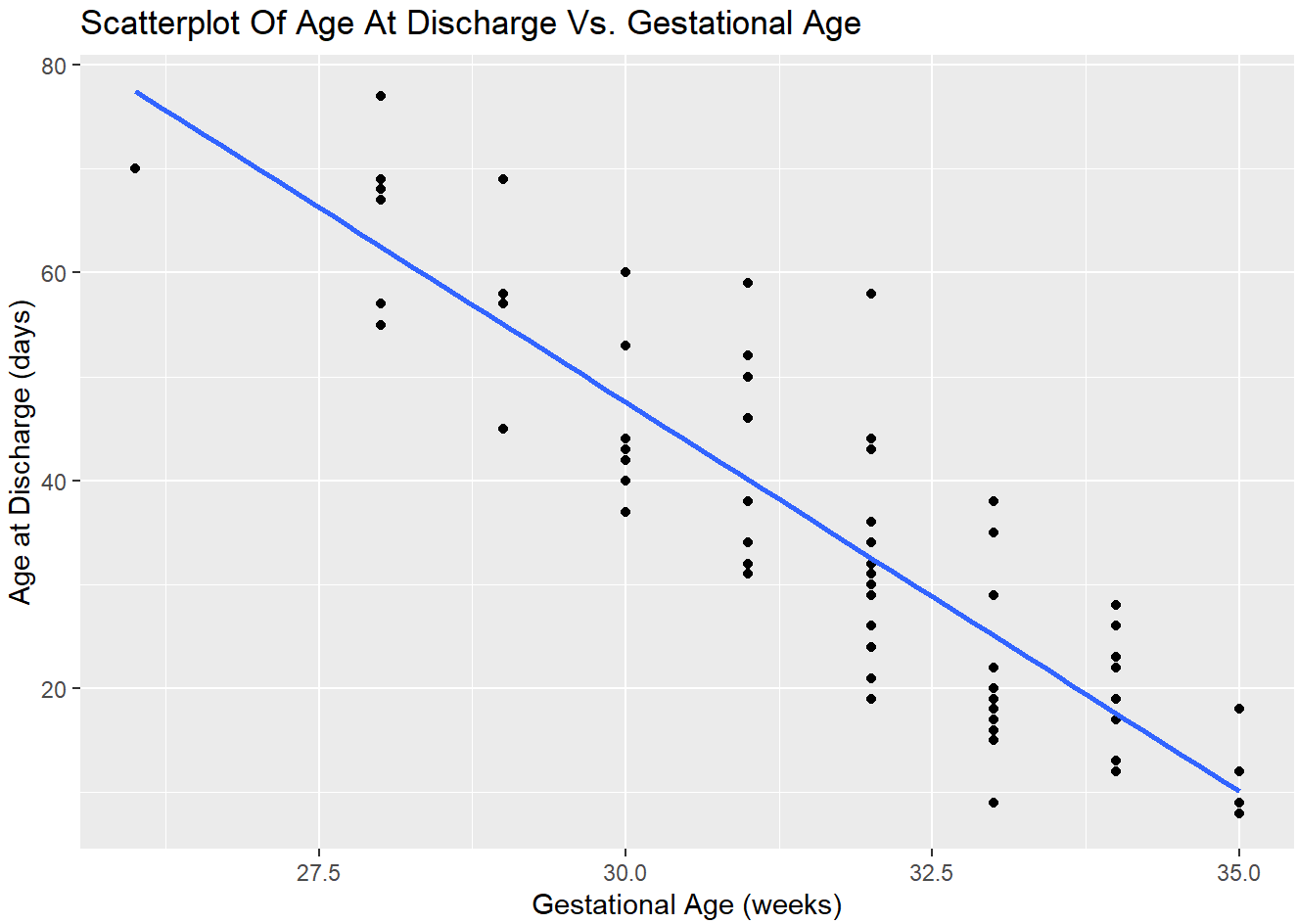


There is a weak relationship between mother's age and age when she stopped breast feeding.

Question 3

```
bf |>
    ggplot(aes(gest_age, dc_age)) +
      geom_point() +
      xlab("Gestational Age (weeks)") +
      ylab("Age at Discharge (days)") +
    geom_smooth(method="lm", se=FALSE) +
    ggtitle("Scatterplot Of Age At Discharge Vs. Gestational Age")
```

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



The plot suggests that infants with lower gestational ages tend to spend more time in the birth hospital, which supports the general expectation that pre-term infants are discharged later than full-term infants.

Linear regression estimates for predicting age_stop

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

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

Coefficients:
(Intercept)      mom_age
      5.920        0.389
```

The estimated average duration of breast feeding increases by 0.39 weeks for each increase of one year in the mother's age. The estimated average duration of breast feeding is 5.9 weeks for a mother of age zero. This is an extrapolation well beyond the range of the data.

Question 4

```
        m2 <- lm(dc_age~gest_age, data=bf)
        m2
```

```
Call:
lm(formula = dc_age ~ gest_age, data = bf)

Coefficients:
(Intercept)     gest_age
    272.206       -7.489
```

For each additional week of gestational age, the expected age at discharge decreases by approximately 7.49 days. The intercept of 272.21 days represents the predicted age at discharge if an infant had a gestational age of 0 weeks.

Analysis of variance table for age_stop

```
        anova(m1)
```

```
Analysis of Variance Table

Response: age_stop
          Df Sum Sq Mean Sq F value  Pr(>F)
mom_age    1  570.0  569.99  5.7531 0.01879 *
Residuals 80 7925.9   99.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-ratio is large and the p-value is small, so you would reject the null hypothesis and conclude that there is a linear relationship between mother's age and duration of breast feeding.

Question 5

```
        anova(m2)
```

```
Analysis of Variance Table

Response: dc_age
          Df  Sum Sq Mean Sq F value    Pr(>F)
gest_age   1 19122.9 19122.9  280.72 < 2.2e-16 ***
Residuals 82  5585.8    68.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-ratio is large and the p-value is small, so the null hypothesis is rejected and there is a linear relationship between age at discharge (days) and gestational age (weeks).

R-squared for age_stop

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

```
[1] 0.06708949
```

Although there is a statistically significant relationship between mother's age and duration of breast feeding, as shown above, this relationship is very weak.

Question 6

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

```
[1] 0.773932
```

There are approximately 77.4% of the variability in the age at discharge is explained by the gestational age. This suggests a strong relationship between gestational age and the time infants spend in the hospital.

Confidence interval for the slope

```
        confint(m1)
```

```
                 2.5 %    97.5 %
(Intercept) -3.19546976 15.035265
mom_age      0.06625878  0.711827
```

The confidence interval includes only positive values, so we are 95% confident that the duration of breast feeding increases as the mother's age increases. The slope could be as small as 0.067 weeks per year of mother's age or as large as 0.71 weeks per year of mother's age. This is a very wide interval indicating a large degree of uncertainty about the true value of the slope parameter.

Alternate test for the slope parameter

```
        tidy(m1)
```

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    5.92      4.58       1.29  0.200
2 mom_age        0.389     0.162      2.40  0.0188
```

The T statistic is testing the slope parameter is large and the p-value is small, both indicating that you should reject the null hyothesis and conclude that there is a positive relationship between mother's age and

duration of breast feeding.

Question 7

```
confint(m2)
```

```
                2.5 %     97.5 %
(Intercept) 243.83418 300.577438
gest_age     -8.37785  -6.599561
```

We are 95% confident that the true value of the intercept lies between 243.83 and 300.58 with the true value of the slope lies between -8.38 and -6.60. This means that for each additional week of gestation, the age at discharge decreases by between 6.60 and 8.38 days. The confidence interval does not include 0, indicating that gestational age has a statistically significant negative effect on the time spent in the hospital.

The interval is relatively narrow, indicating that the estimate of the slope is precise and there is relatively low uncertainty in the effect of gestational age on discharge time.