Analysis of breast feeding study

AUTHOR
Steve Simon and Leroy Wheeler

PUBLISHED
September 18, 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
```

Question 1: Calculate descriptive statistics for gestational age (mean, standard deviation, minimum, and maximum) and count the number of missing values. Interpret these results.

```
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
```

This is a reasonable distribution of infants who are born early. The youngest pre-term infant was 26 weeks old with the oldest pre-term infant being 35 weeks of age. There are no missing data for this variable in this data set.

Question 2: Calculate descriptive statistics for age at discharge from the hospital and count the number of missing values. Interpet these results.

```
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()
```

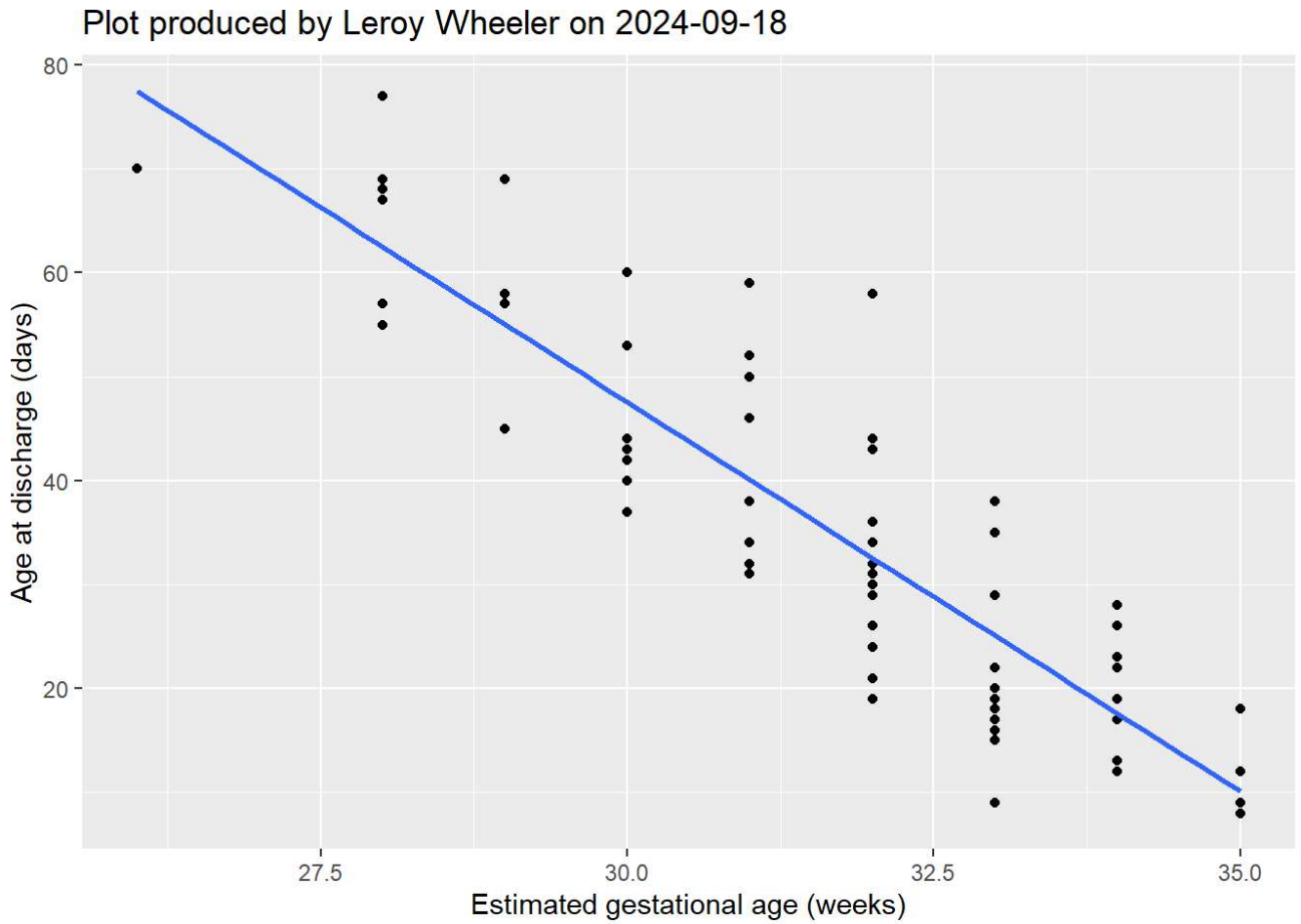
```
  mean_dc_age sd_dc_age min_dc_age max_dc_age n_missing
1    33.72619  17.25385          8         77         0
```

The average age of the infants discharged from the hospital was about 34 days, with a minimum value of 8 days and a maximum value of 77 days. These are expected values as infants who are pre-term often have to remain in the hospital after they are born. There are no missing values for this variable in this data set.

Question 3: Pre-term infants spend a longer amount of time in the hospital than full-term infants. In fact, the earlier the baby appears, the longer the amount of time that the infant remains in the birth hospital. Draw a scatterplot to examine whether this pattern holds in this dataset. Consider age at discharge to be the outcome variable when deciding how to draw this scatterplot. Use the geom_smooth function to graph the regression line, but do not extend the line beyond the range of the data.

```
    bf |>
        ggplot(aes(gest_age, dc_age)) +
          geom_point() +
          xlab("Estimated gestational age (weeks)") +
          ylab("Age at discharge (days)") +
        geom_smooth(method="lm", se=FALSE) +
        ggtitle("Plot produced by Leroy Wheeler on 2024-09-18")
```

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



Our data set here is consistent with previous observations that infants with a younger estimated gestational age at birth must spend more time in the hospital before being discharged.

Question 4: Use the lm function to compute the slope and intercept for the regression model predicting age at discharge using gestational age. Interpret both the slope and the intercept and state whether the intercept represents an inappropriate extrapoloation.

```
m1 <- lm(dc_age~gest_age, data=bf)
m1
```

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

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

We can see from the slope of the regression model, for every week early that the infant is born, she has to spend an additional 7.5 days in the hospital before being discharge. The y-intercept is well out of the range of data because no infants will be able to survive outside the womb at a gestational age of zero. The y-intercept is therefore meaningless.

Question 5: Calculate an analysis of variance table for this regression model. Interpret the F ratio and the p-value. What hypothesis do these two statistics test?

```
anova(m1)
```

```
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 very large and the p-value is extremely small, so we will reject the null hypothesis and conclude that there is strong evidence for a linear relationship between gestational age of the infant and age of the infant at discharge from the hospital.

Question 6: Calculate R squared for this regression model. Interpret this value. R-squared for dc_age.

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

```
[1] 0.773932
```

We see that approximately 78% of the extra time remaining in the hospital is due to the gestational age of the infant at birth. Therefore we conclude that the inverse relationship observed when we graph gestation age of the infant vs the age of the infant at discharge is strong.

Question 7: Compute a confidence interval for the slope parameter. Interpret this interval. Characterize the interval as either narrow or wide.

```
confint(m1)
```

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

The 95% confidence interval includes only negative values consistent with the inverse relationship between gestational age at birth and time remaining in the hospital after birth. We are therefore 95% confident that the time spent in the hospital after birth decreases somewhere between 6.6 and 8.4 days for every additional gestational week of age that the infant has at the time of birth. The slope of the regression line has a narrow range between -6.6 days and -8.4 days spent in the hospital per an additional week of gestational age at birth.

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)   272.      14.3        19.1 6.59e-32
2 gest_age       -7.49     0.447     -16.8 3.32e-28
```

The T statistic is testing the slope parameter is large and the p-value is small, both indicating that you should reject the null hypothesis and conclude that there is a negative relationship the infants gestational age and its age at the time of discharge from the hospital.