Project 1 - Linear Regression

MTH 161 – Section H, I – Fall 2024

1 Introduction

For this project you will investigate a dataset containing a sample of observations from 2020 of US birth data originally sourced from the Centers for Disease Control (CDC).

The goals are to demonstrate your understanding of concepts and techniques covered in the first half of the semester in the context of this dataset. These concepts include

- Visualizing and interpreting variables
- Visualizing associations between variables
- Formulating research questions, including the role of explanatory and response variables
- Constructing and interpreting linear models
- Drawing appropriate conclusions from analysis and communicating them clearly and accurately

Data

You will use the data set `us_births.csv` to fit your model. You can load it using the code below.

```
library(tidyverse)
library(tidymodels)

births <- read_csv("us_births.csv")
```

This data set contains 3000 observations of 16 variables. The variables are described in the codebook below.

Codebook

- `newborn_birth_weight`: newborn birth weight in grams
- `month`: birth month (1: January, ..., 12: December)
- `mother_age`: age of the mother in years
- `prenatal_care_starting_month`: month in which prenatal care began; if 0, there was no prenatal care
- `daily_cigarette_prepregnancy`: daily number of cigarettes smoked before the pregnancy
- `daily_cigarette_trimester_1`: daily number of cigarettes smoked during the 1st trimester of the pregnancy
- `daily_cigarette_trimester_2`: daily number of cigarettes smoked during the 2nd trimester of the pregnancy
- `daily_cigarette_trimester_3`: daily number of cigarettes smoked during the 3rd trimester of the pregnancy
- `mother_height`: height of the mother in inches
- `mother_bmi`: body mass index of the mother
- `mother_weight_prepregnancy`: weight of the mother before the pregnancy in pounds
- `mother_weight_delivery`: weight of the mother at delivery in pounds
- `mother_diabetes_gestational`: whether the mother had diabetes during the pregnancy
- `newborn_sex`: sex of the newborn
- `gestation_week`: number of gestational weeks
- `mother_risk_factors`: whether the mother had any risk factor (diabetes, hypertension, previous preterm birth, previous cesarean, infertility treatment used, etc)

```
glimpse(births)
```

```
Rows: 3,000
Columns: 16
$ newborn_birth_weight         <dbl> 3572, 3290, 3459, 3685, 2405, 3946, 3433,~
$ month                        <dbl> 2, 3, 3, 12, 8, 2, 4, 5, 6, 5, 6, 4, 1, 7~
$ mother_age                   <dbl> 29, 32, 33, 36, 22, 27, 28, 38, 29, 38, 3~
$ prenatal_care_starting_month <dbl> 2, 4, 3, 2, 2, 2, 3, 2, 4, 3, 3, 2, 3, 2,~
$ daily_cigarette_prepregnancy <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 20, 0, 0~
$ daily_cigarette_trimester_1  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0,~
$ daily_cigarette_trimester_2  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ daily_cigarette_trimester_3  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ mother_height                <dbl> 69, 65, 62, 69, 66, 68, 66, 64, 61, 66, 6~
$ mother_bmi                   <dbl> 36.2, 30.8, 23.4, 22.7, 22.6, 38.0, 26.6,~
$ mother_weight_prepregnancy   <dbl> 245, 185, 128, 154, 140, 250, 165, 180, 1~
$ mother_weight_delivery       <dbl> 290, 190, 153, 195, 168, 289, 197, 205, 1~
```

```
$ mother_diabetes_gestational  <chr> "Y", "N", "N", "N", "N", "N", "N", "N", "~
$ newborn_sex                  <chr> "M", "F", "M", "M", "F", "F", "M", "F", "~
$ gestation_week               <dbl> 37, 38, 39, 38, 37, 39, 39, 34, 43, 39, 3~
$ mother_risk_factor           <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, ~
```

2 Preliminary investigations

The purpose of this project is to investigate possible associations between `newborn_birth_weight` and other variables in the dataset. In this section, you will perform some prelimary investigations to help you formulate a research question.

Instructions

Insert your responses (code chunks and written text) below each of the following task statements.

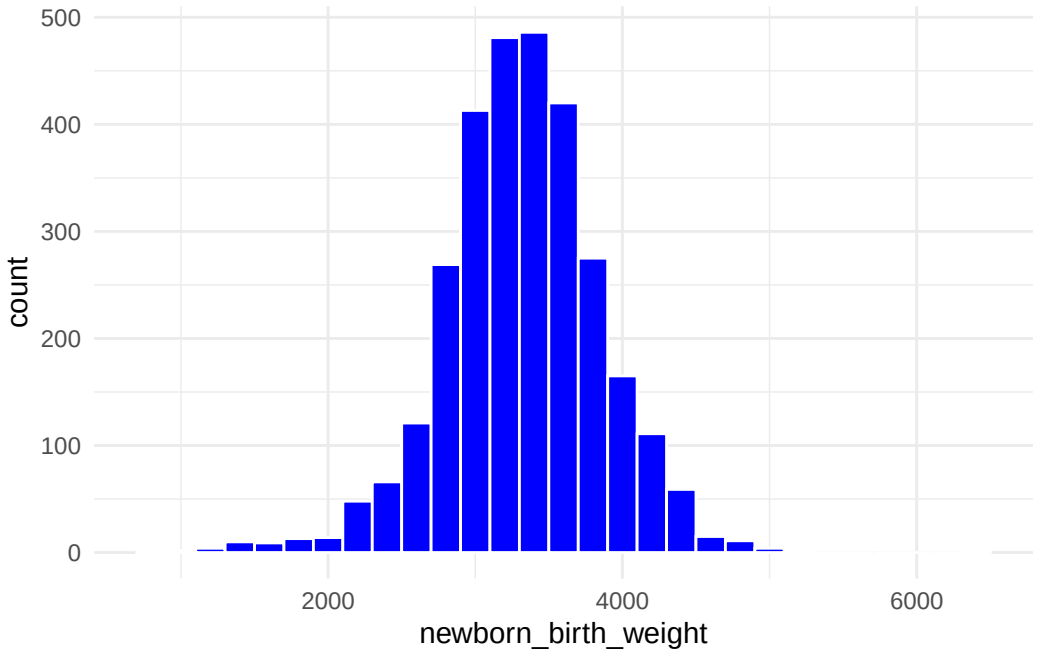
Task

Begin by thinking about the variable `newborn_birth_weight`. What type of variable is it? Use an appropriate visualization to plot this data. How would you describe the distribution?

Response

This is a numeric (continuous) variable. You can use a histogram and/or box plot to visualize.

```
ggplot( births, aes(x = newborn_birth_weight)) +
  geom_histogram(binwidth = 200, col="white", fill="blue") +
  theme_minimal()
```



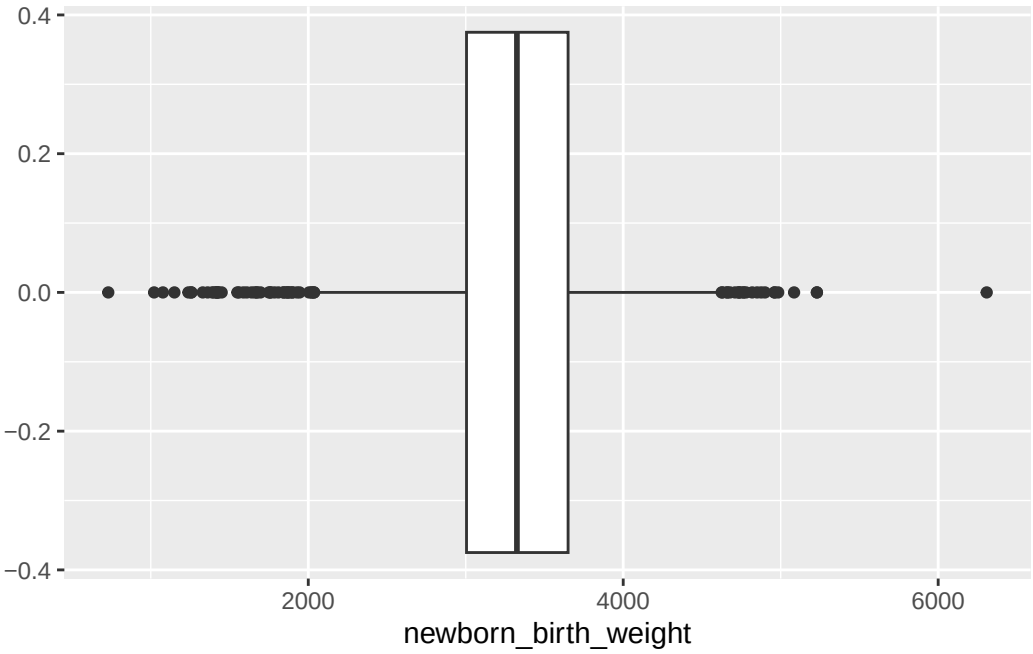
Task

To further investigate `newborn_birth_weight`, create a *different* kind of plot that would also be appropriate for visualizing this variable. Does your new visualization reinforce anything you noticed in the first plot? Does it illuminate any new aspects of the distribution? Comment specifically on what *variance* you observe in the distribution based on your plots.

Response

Here is a box plot, which makes outliers easier to see. Can also estimate IQR is about 600g with fairly long tails, particularly to the left.

```
ggplot( births, aes(x = newborn_birth_weight)) +
  geom_boxplot()
```



Task

Next, examine the codebook above. Select **two numerical variables** that you are curious about in terms of their potential relationship with `newborn_birth_weight`. Separately describe the association you expect each of these variables to have with `newborn_birth_weight` and provide brief rationale for your expectations.
Then, select **one categorical variable** and do the same: what association do you expect this variable has with `newborn_birth_weight` and why?

Response

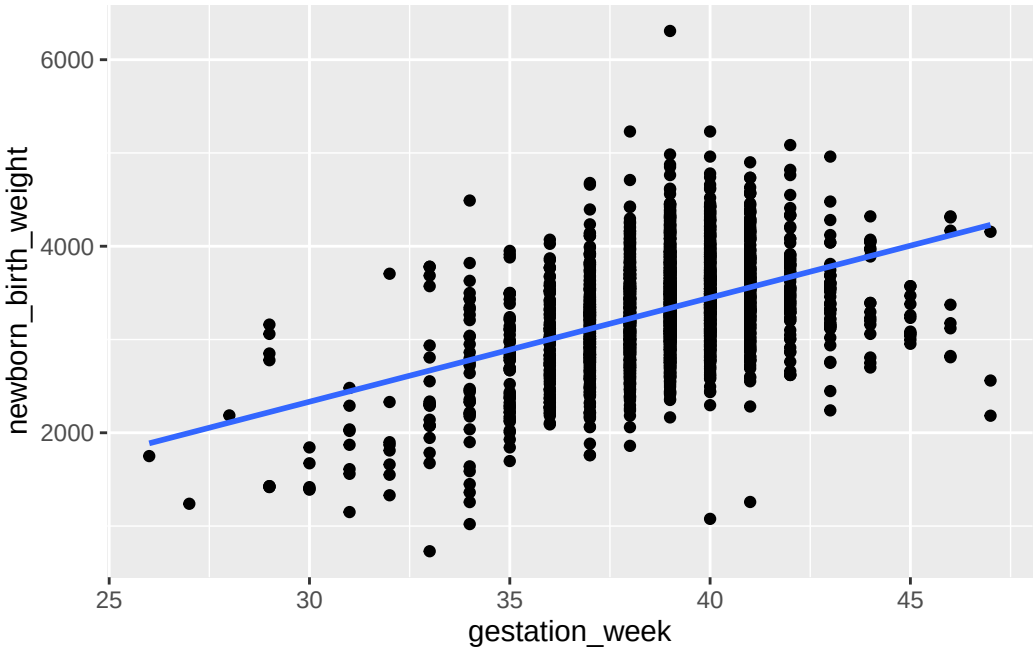
Possible choices here include `mother_age`, `prenatal_care_starting_month`, `daily_cigarette_prepreganacy`, `mother_height`, `mother_weight_delivery`, etc.

```
ggplot(births, aes(x=mother_weight_delivery, y = newborn_birth_weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
ggplot(births, aes(x=mother_height, y = newborn_birth_weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

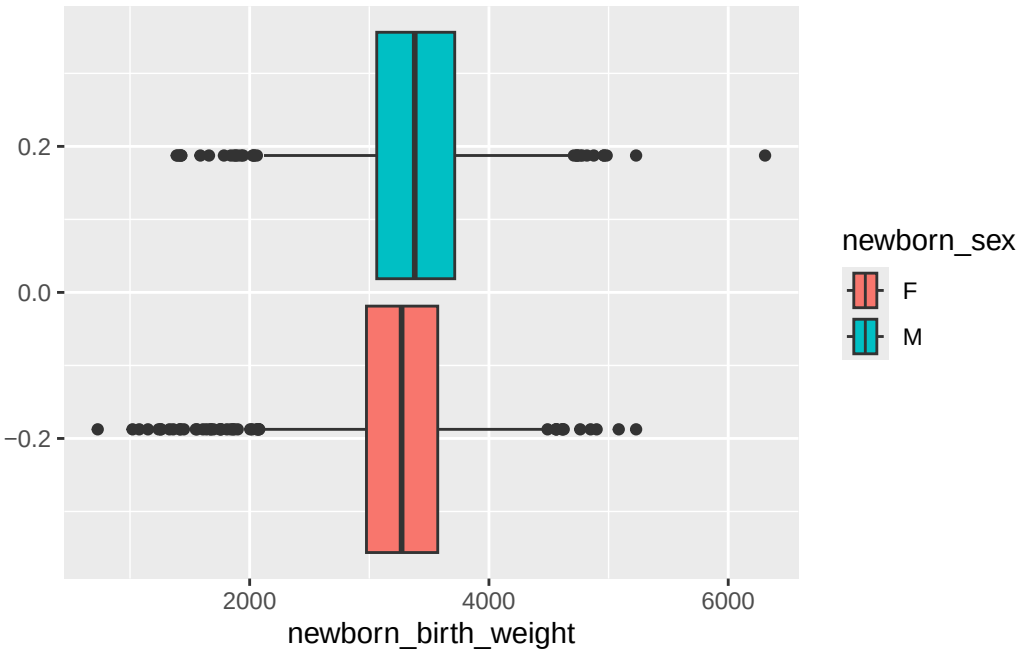
```
ggplot(births, aes(x=gestation_week, y = newborn_birth_weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

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



> Response
>
> Possible choices include `mother_risk_factor`, `newborn_sex`, `mother_diabetes_gestational`. If use `month`, must convert to factor by using `as.factor`

```
ggplot(births, aes(x=newborn_birth_weight, fill= newborn_sex)) +
  geom_boxplot()
```




```
ggplot(births, aes(x=newborn_birth_weight, fill= mother_risk_factor)) +
  geom_boxplot()
```

```
ggplot(births, aes(x=newborn_birth_weight, fill= mother_diabetes_gestational)) +
  geom_boxplot()
```

Task

Narrow your focus to one of the variables you picked above and create a plot that will help you investigate a possible association between `newborn_birth_weight` (the response variable) and your chosen (explanatory) variable. Choose your visualization according to the types of variables involved.

Note: visualizing several possible relationships here might be a good way to pick which one you'd like to report on. Feel free to experiment!

3 Main report

The remainder of your report will focus on an investigation of the variable you selected above (as it relates to birth weight). Your job is to carry out an analysis of any potential association and summarize your findings below.

Research question

As a researcher, your investigation should begin with a clearly-formulated, measurable research question. What are you hoping to discern with this analysis and why should the reader be interested in the results? Having an initial hypothesis can help shape your specific research question and guide the study design overall. In this section, you will demonstrate proficiency in asking meaningful questions (that you later answer using data).

> Example research question and hypotheses (if we were predicting penguin weights instead of baby weights):
>
> *Can penguin bill depth be used to predict body mass? Having an answer to this question could help inform the design of future studies of wild penguins and provide useful data for those who work with captive penguin populations. Many other variables (e.g., calories needed) and interventions (e.g., amount of antibiotic to administer) relate directly to body mass. We hypothesize that penguins with deeper bills will also have more mass.*

Task

Formulate a clear and concise research question. State your hypotheses and provide rationale for the investigation, including motivation as to why the research question is interesting or useful.

Response

I hypothesize that newborn birthweight is positively associated with mother's height.

Methodology

In this section, you will explain the methods used to investigate the association between variables. You will also demonstrate proficiency in using *R* to perform and visualize linear regression.

Task

Find a linear model that predicts birth weight based on the variable you selected. Include a table of the summary statistics needed to find your regression line. Be sure to include the **entire model equation**.

```
births |>
  summarize( mean_x = mean(gestation_week, na.rm = TRUE),
             sd_x = sd(gestation_week, na.rm = TRUE),
```

```
            mean_y = mean(newborn_birth_weight, na.rm = TRUE),
            sd_y = sd(newborn_birth_weight, na.rm = TRUE),
            r = cor(gestation_week, newborn_birth_weight)
  )
```

```
# A tibble: 1 x 5
  mean_x  sd_x mean_y  sd_y     r
   <dbl> <dbl>  <dbl> <dbl> <dbl>
1   38.9  1.98  3322.  530. 0.417
```

```
linear_reg() |>
  fit(newborn_birth_weight ~ gestation_week, data = births)
```

Response

The linear model is

$$y = 112x - 1013$$

This is only valid on the range $26 \leq x \leq 47$. In particular, the y-intercept corresponding to 0 gestation week is not meaningful.

Task

Create a visualization of your linear model that shows its relationship to your variables.

Results

In this section you will analyze and interpret the results of your linear regression, demonstrating that you are proficient at answering research questions using data.

Task

Analyze the linear model you found, discussing the type and strength of the association between variables. What specific data are you using to substantiate your claims about strength (r, R^2, something else) and why? For what range of values does this model have reasonable predictive power? What does the slope of your line tell you?

Response

The linear model is

$$y = 112x - 1013$$

Since the slope is postive, this indicates a positive association between the variables – as `gestation_week` increases, so does `newborn_birth_weight`. This is only valid on the range $26 \leq x \leq 47$. In particular, the y-intercept corresponding to 0 gestation week is not meaningful. The value of the correlation coefficient is about $r = 0.42$ which indicates a weak to moderate association. It's nevertheless the strongest association I could find. The value of R^2 is about 0.17 which indicates that about 17% of the variation in birth weights can be explained by `gestation_week`

Discussion

In this concluding section, you will summarize your findings and provide a brief discussion of your analysis.

Task

Summarize your main finding in a sentence or two. Discuss your finding and why it is useful (connect this to the motivation you provided in framing yur research question).

Task

Offer a critique your own analysis – what are the limitations of your findings?

Task

Finally, offer a few thoughts about ways you might extend this research, or other research questions you might be interested in after completing this project.