Week 6 DS Recitation: Regression

SIBDS 2024 @ Columbia

04 July, 2024

Getting Started

1. Create a new R project and named it `week6_DS_recitation`
2. Put the `week6_DS_recitation_regression.Rmd` file into the same folder of the R project you just created.
3. Create a folder within the project folder named `data`.
4. Download the `birthweight.csv` dataset and put it in the `data` folder.

Linear Regression

In this problem, you will analyze data gathered to understand the effects of several variables on a child's birth weight. This dataset (`birthweight.csv`), consists of roughly 4000 children and includes the following variables:

- babysex: baby's sex (male = 1, female = 2)
- bhead: baby's head circumference at birth (centimeters)
- blength: baby's length at birth (centimeteres)
- bwt: baby's birth weight (grams)
- delwt: mother's weight at delivery (pounds)
- fincome: family monthly income (in hundreds, rounded)
- frace: father's race (1 = White, 2 = Black, 3 = Asian, 4 = Puerto Rican, 8 = Other, 9 = Unknown)
- gaweeks: gestational age in weeks
- malform: presence of malformations that could affect weight (0 = absent, 1 = present)
- menarche: mother's age at menarche (years)
- mheigth: mother's height (inches)
- momage: mother's age at delivery (years)
- mrace: mother's race (1 = White, 2 = Black, 3 = Asian, 4 = Puerto Rican, 8 = Other)
- parity: number of live births prior to this pregnancy
- pnumlbw: previous number of low birth weight babies
- pnumgsa: number of prior small for gestational age babies
- ppbmi: mother's pre-pregnancy BMI
- ppwt: mother's pre-pregnancy weight (pounds)
- smoken: average number of cigarettes smoked per day during pregnancy -wtgain: mother's weight gain during pregnancy (pounds)

Load and clean the data using the code below for regression analysis.

```
library(tidyverse)
baby_df =
  read_csv("./data/birthweight.csv") %>%
```

```
  mutate(
    babysex = case_when(babysex == 1~"male", babysex == 2~"female"),
    malform = case_when(malform == 0~"absent", malform == 1 ~ "present")
    ) %>%
  mutate(
    across(
      .cols = c("frace", "mrace"),
      ~ case_when(
        .x == 1 ~ "White",
        .x == 2 ~ "Black",
        .x == 3 ~ "Asian",
        .x == 4 ~ "Puero Rican",
        .x == 8 ~ "Other",
        .x == 9 ~ "Unknown"
        )
      ),
    across(where(is.character), as.factor)
    ) %>%
  select(
    -c(pnumlbw, pnumsga, wtgain))
```

Model Fitting

1. Propose a regression model for birthweight. This model may be based on a hypothesized structure for the factors that underlie birth weight, on a data-driven model-building process, or a combination of the two. Describe your modeling process and show a plot of model residuals against fitted values – use `add_predictions` and `add_residuals` in making this plot.

```
# Your answer starts here
```

Hypothesis Testing

2. Compare the two models:

- One using length at birth(`blength`) and gestational age (`gaweeks`) as predictors
- One using length at birth(`blength`), gestational age (`gaweeks`), and head circumference(`bhead`) as predictors.

```
# Your answer starts here
```