Week 8 R Code demo

Prof Moon is expecting a baby in April. In this demo, we'll explore the NCbirths data to investigate how big we should expect her baby to be!

```
library(tidyverse) # As usual
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("Stat2Data")
library(Stat2Data) # This package contains the NCbirths data

# Sometimes, we need to use the data() function to load data from an R package into our R session
data(NCbirths) # After running this line, you'll see it appear in the Environment pane in the top left
glimpse(NCbirths)
```

```
## Rows: 1,450
## Columns: 15
## $ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Plural        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ Sex           <int> 1, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1,...
## $ MomAge        <int> 32, 32, 27, 27, 25, 28, 25, 15, 21, 27, 26, 20, 19, 1...
## $ Weeks         <int> 40, 37, 39, 39, 39, 43, 39, 42, 39, 40, 41, 41, 40, 3...
## $ Marital       <int> 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 1,...
## $ RaceMom       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 5, 1, 1,...
## $ HispMom       <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, P, N, M, N, N,...
## $ Gained        <int> 38, 34, 12, 15, 32, 32, 75, 25, 28, 37, 45, 52, 26, 3...
## $ Smoke         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0,...
## $ BirthWeightOz <int> 111, 116, 138, 136, 121, 117, 143, 113, 120, 124, 121...
## $ BirthWeightGm <dbl> 3146.85, 3288.60, 3912.30, 3855.60, 3430.35, 3316.95,...
## $ Low           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ Premie        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ MomRace       <fct> white, white, white, white, white, white, white, whit...
```

R Demo

Q1 - What do we know about these data? (and where can we find out?)

(a) When we load data from a package, where do we go to (hopefully) get more information about the data?

(b) How many observations are there in these data?

(c) What does each observation represent?

(d) Who contributed these data?

Q2 - Prof Moon is more familiar with measuring babies' weights in pounds, so let's create a new variable called `BirthWeightLbs`. We'll use the fact that there are 16 oz in 1 lb.

Q3 - Create a visualization to explore the association between the duration of a pregnancy (in weeks) and the baby's weight at birth. Describe this distribution.

Q4 - Prof Moon is currently 34 weeks pregnant, so she is particularly interested in learning about the association between the duration of pregnancy and birthweight for babies born at 34 weeks or later (babies born before this time are quite premature and so their birthweight is not be as relevant). Let's create a new tibble containing only observations for babies born in this range.

Part A

In this part, we'll explore the association between the duration of a pregnancy and birthweight (after 34 weeks gestation).

QA3 - Calculate the correlation between the number of weeks of pregnancy and birthweight (focusing only on babies born after 34 weeks gestation). Does the correlation reflect our intuition about the association between these two variables?

QA4 - Fit a linear regression model to predict the weight of a baby based on the duration of the pregnancy (in weeks), to find the least-squares estimates for the regression parameters. Create a visualization to visualize the association between weeks and birthweight for babies born at or after 34 weeks, adding the fitted regression line to the plot.

QA5 - How can we interpret the estimated slope (beta1-hat)?

QA6 - How can we interpret the estimated intercept (beta0-hat)?

QA7 - Based on these data, is there evidence that the slope of this linear regression model significantly different from 0?

QA8 - What proportion of the variability in weight is explained by our regression model? What does this suggest?

Part B

In this part, we'll explore the association between a mothers' smoking status and birthweight in infants born after 34 weeks gestation.

QB1 - Let's create a visualization to explore the association between mothers' smoking status and birthweight (after 34 weeks), and describe what we observe.

QB2 - Let's fit a linear regression model to predict the birthweight of a baby based on his/her mother's smoking status.

QB3 - What is the baseline level in the model above?

QB4 - How can we interpret the estimated intercept (beta0-hat)?

QB5 - How can we interpret the estimated slope (beta1-hat)?

QB6 - Based on these data, is there a difference between the mean weight of babies born to smoking vs non-smoking mothers?

Conclusion

Where do we go from here?