SDS 439 - Homework 01

Due Jan 29, 11:59pm

Introduction

This is an R markdown file. If you have it open in RStudio, you will see that it is just a text file, but it has special formatting.

You will put code inside of the "code chunk" sections, like this

```
x <- rnorm(7)
print(x)
```

```
## [1] -1.8299344 -0.4778006 -0.7689407  1.5259898 -1.0272501  0.1294346  0.7241449
```

You will complete this assignment by responding to the prompts with your own code inside of the code chunk sections.

You will not turn in this .Rmd file. Instead, turn in the "knitted" .pdf document, which you can create by clicking "Knit" in the menu at the top of RStudio.

If you get an error, do not panic. Error messages contain useful information. Stop, read what the error message says and try to follow the instructions.

You will turn in a pdf document. Your life will be easier if you can figure out how to knit directly to a pdf. So take the time to get that to work.

Homework Exercises

This homework concerns analysis of the Galton families dataset, which we will use several times throughout the course to demonstrate various models and concepts. Each row of the dataset describes data for one child. It has columns identifying which family the child belongs to (`family`), the father's height (`father`), the mother's height (`mother`), the number of children in the family (`children`), the ordering within the family of the present child (`childNum`), the sex of the child (`sex`), and the child's height (`childHeight`). It also has a derived variable called `midparentHeight` that is equal to `(father + 1.08*mother)/2`.

The dataset is located in the course repository in the file `datasets/galton_families.csv`.

1. Read in the Galton Families dataset and print out the first six rows of the dataset
2. Demonstrate that `midparentHeight = (father + 1.08*mother)/2`
3. Make a plot of the data, showing information about the midparent height, the child height, and if you can, information about the sex. Try using a different plotting symbol for the two sexes. Make sure that your plot looks nice in the final pdf. This is what we will grade.
4. Let's focus on one of the sexes, then we'll come back and analyze both of them. Create a new dataframe that has data for only the male children, and print out the first six rows.
5. Use the male dataframe to perform a regression of child height on midparent height "by hand", which means doing all of the calculations without using the `lm` function. Print out your estimates for the regression coefficients, their standard errors, their t-statistics, and their p-values for the two sided test for whether the parameters equal 0. Also print out your estimate of the residual variance.

6. Why do you think that the intercept estimate has such a large standard error? Type your answer here
7. Use the `lm` function to confirm your "by hand" calculations. Print the summary table from the lm fit. Use the full dataset, but tell lm to analyze only the males by specifying the `subset` argument of `lm`.
8. Perform two more regressions using `lm` by changing the covaraiate to `mother` and then to `father`. Print out the summary table for these two regressions.
9. Use the information in the summary tables to argue which of the three models for `childHeight` is best. Is there a biological reason why this model is best?
10. Perform a regression of `childHeight` on `midparentHeight` for the female children and print the summary matrix.
11. What are the major differences between the female model and the male model?
12. Finally, make a plot of childHeight against the midparentHeight, with the two estimated trend lines for males and females added to the plot.