Practice Problems 4

Solutions

Instructions

To complete your assignment, please follow these steps:

0. **Download R and RStudio (if you haven't already).** See the *Useful Resources* section on Moodle for installation links and videos with step-by-step installation instructions. If you're having issues with this, you'll need to reach out to your instructor and/or come to office hours, so plan accordingly!
1. **Download `practice4.Rmd` from Moodle and save it some place on your computer that you can easily find again.** I strongly encourage you to create a new folder dedicated to homework assignments for this class. See the *File Structure and Organization* video on Moodle for tips on how to do this.
2. **Make sure that the file you downloaded is called `practice4.Rmd` and not `practice4.Rmd.txt`.** The latter often happens when you use Safari on a Mac—try downloading the file using a different browser (e.g., Chrome) instead, or edit the file name (as explained in the *File Structure and Organization* video).
3. **Open `practice4.Rmd` in RStudio.** See the *Intro to R and RStudio*, *R Data Types*, and *R Error Messages and Troubleshooting* videos.
4. **Update the section number, author, and due date** on the third, fourth, and fifth lines of the file.
5. **Make sure you've already installed all necessary R packages.** Open the `Packages` panel (usually in the bottom right corner) to see the list of all packages that are already installed. Look to see if the packages you need are listed here. If they're on this list, that means they're already installed and you're good to go. If any packages are missing from this list, type `install.packages('packagename')` in the *Console* (usually in the bottom left corner) and hit enter. See the *R Packages* video.
6. **Try *Knitting* your document:** click the `Knit` button at the top of this screen (look for yarn and needle). See the *Intro to RMarkdown* video. A dialogue box may pop up asking you if you want to install some packages: click "Yes." If you encounter any error messages, get in touch with your instructor or the preceptors.
7. **Answer the questions in Parts 0–3.** Click `Knit` occasionally along the way to make sure everything looks okay.
8. Once you're done with all parts, **click `Knit` one final time.** This will turn your R Markdown document into a nicely formatted HTML file.
9. **Look at the HTML file to make sure it looks like you want it to**: graphs appearing, no error messages, no data print out that lasts 50 pages, etc.
10. Once your HTML file pops up and you've checked that it looks like what you want, **click `Open in Browser` and then Print and select "Save as pdf".**
11. **Turn in two files to Moodle:** this `.Rmd` file, and the knitted `.pdf` version you generated in Step 10.

Part 0. Load Packages

Use the code chunk below to load any packages that you need for this assignment:

```
# at minimum, you should have loaded these packages:
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

# depending on your project data, you may have needed others
## e.g., library(NHANES)
```

> Notice how the code to load packages prints out some messages. To clean up your Knitted document, I recommend hiding these by adding `message = FALSE` to the start of your code chunk:

```
# at minimum, you should have loaded these packages:
library(readr)
library(dplyr)
library(ggplot2)

# depending on your project data, you may have needed others
## e.g., library(NHANES)
```

Part 1. FEV Data

Answer the following questions about the `fev.csv` dataset.

Load Data in R

Enter code in the code chunk below to load the `fev.csv` dataset into R. Make sure that you've downloaded the `.csv` file from Moodle (see Activity 8) and it's saved in the same folder as this RMD. (Don't forget to add and run any `library` statements that you need to the `load-packages` code chunk in Part 0, above!)

```
# read in FEV dataset
fev <- read_csv('../activities/fev.csv')
```

I have the `fev.csv` dataset stored in a different folder than this RMD, so my code may look a little different from yours. If your CSV and RMD files are saved in the same folder, then the code looks like this:

```
fev <- read_csv('fev.csv')
```

I also recommend hiding messages here by using the `message = FALSE` code chunk option!

Linear Regression - FEV vs Height - Model Equation

Write down the model equation for a simple linear regression model that models the average forced expiratory volume (FEV) as a function of height. (*Hint: see Activities 8–10.*)

$$E[FEV|height] = \beta_0 + \beta_1 height$$

Your model equation should use correct notation (expected value E, betas β for intercept and slope, etc.).

It's also helpful to write this model in a way that can be understood by people who aren't familiar with these data. While the variable names in this dataset are somewhat self-explanatory (`fev` , `height`), notice how I capitalized FEV to signal that this is an acronym. If height had a different name in our dataset (e.g., `ht`) I also would have changed it to `height` to be more easily understood by a general audience.

Linear Regression - FEV vs Height - Fit Model

Add code to the code chunk below to fit this model and print out estimates for the intercept ($\hat{\beta}_0$) and slope ($\hat{\beta}_1$). (Don't forget to add and run any `library` statements that you need to the `load-packages` code chunk in Part 0, above!)

There's more than one way to write this code! Here are three options:

```
# fit model of FEV versus Height
## option 1
fev %>%
  with(lm(fev ~ height))
```

```
##
## Call:
## lm(formula = fev ~ height)
##
## Coefficients:
## (Intercept)       height
##      -5.433        0.132
```

```
## option 2
fev %>%
  lm(fev ~ height, data = .)
```

```
##
## Call:
## lm(formula = fev ~ height, data = .)
##
## Coefficients:
## (Intercept)       height
##      -5.433        0.132
```

```
## option 3
lm(fev ~ height, data = fev)
```

```
##
## Call:
## lm(formula = fev ~ height, data = fev)
##
## Coefficients:
## (Intercept)       height
##      -5.433        0.132
```

Linear Regression - FEV vs Height - Interpretation

Write a sentence that interprets the value of the intercept of this model, using the estimate provided by the `lm()` R code above. Make sure to use non-causal language, include units, and talk about averages rather than individuals. Also comment on whether this intercept is scientifically meaningful and/or sensible.

> Intercept Interpretation: We estimate that the average FEV among children who are 0 inches tall is -5.43 liters/second. This is neither scientifically meaningful (are we interested in learning about the FEV of kids who are 0 inches tall?) or sensible (can FEV be negative?)!

Then, write a sentence that interprets the value of the slope of this model, using the estimate provided by the `lm()` R code above. Make sure to use non-causal language, include units, and talk about averages rather than individuals.

> Slope Interpretation: Among children aged 3–19 years, we estimate that each additional inch in height is associated with a 0.13 liters/second increase in average (or expected) FEV.
>
> Another acceptable interpretation: Comparing children who differ in height by one inch, we estimate that average FEV differs by 0.13 liters per second, where taller kids tend to have higher FEV on average.

Data Preparation and Cleaning - Centered Height

In the last question, you might have noticed that the intercept of our model doesn't have a particularly sensible or scientifically meaningful interpretation. To help with this, we'll fit a new model using *centered* height as our predictor variable. We'll create this new variable by calculating the average height in our dataset and subtracting that from each child's height. In this way, when the new "centered" height variable is zero, that means that a child is of "average" height.

Add code to the code chunk below to create this new `centeredheight` variable. (*Hint: see Activity 10.*) Don't forget to add and run any `library` statements that you need to the `load-packages` code chunk in Part 0, above!

```
# create a new variable called centeredheight
fev <- fev %>%
  mutate(centeredheight = height - mean(height))
```

Linear Regression - FEV vs Centered Height - Fit Model

Now, add code to the code chunk below to fit this new model with `centeredheight` as the predictor variable and print out estimates for the intercept ($\hat{\beta}_0$) and slope ($\hat{\beta}_1$).

> As before, there's more than one way to write this code. I'm going to use the style of "option 3" from above, but you might have done this slightly differently:

```
# fit model of FEV versus Centered Height
lm(fev ~ centeredheight, data = fev)
```

```
##
## Call:
## lm(formula = fev ~ centeredheight, data = fev)
##
## Coefficients:
##    (Intercept)  centeredheight
##          2.637           0.132
```

Linear Regression - FEV vs Centered Height - Interpretation

Write a sentence that interprets the value of the intercept of this new model, using the estimate provided by the `lm()` R code above. Make sure to use non-causal language, include units, and talk about averages rather than individuals. Also comment on whether this intercept is scientifically meaningful and/or sensible.

> Intercept Interpretation: We estimate that the average FEV among children of average height (about 61.1 inches in this dataset—see below) is 2.64 liters per second.

```
# calculate average height
fev %>%
  summarize(mean(height))
```

```
## # A tibble: 1 × 1
##   `mean(height)`
##            <dbl>
## 1           61.1
```

Then, write a sentence that interprets the value of the slope of this new model, using the estimate provided by the `lm()` R code above. Make sure to use non-causal language, include units, and talk about averages rather than individuals.

> Slope Interpretation: Comparing children who differ in height by one inch (note that differing in centered height by one inch is the same as differing in *height* by one inch, so I went with the more generally accessible statement!), we estimate that the average FEV differs by 0.13 liters per second. Taller kids tend to have higher FEV on average.

> Alternatively: Among children aged 3–19 years, we estimate that 1 inch changes in height (equivalently, 1 inch changes in centered height) are associated with a 0.13 liter/second increase in FEV, on average.

Part 2. Project Data

Answer the following questions about your chosen project dataset. Use the same dataset you used last week (unless you have approval from Kelsey to change datasets).

Load Data in R

Use the code chunk provided below to read your chosen data set into R. Don't forget to add and run any `library` statements that you need to the `load-packages` code chunk in Part 0, above (unless you've already done so earlier in this RMD).

> If you're using the NHANES data for your project, your code should look something like this:

```
# load package first, or put this up in Part 0
library(NHANES)

# then read in dataset
data(NHANES)
```

> Otherwise, you're likely using the `read_csv` function to read in your data from a CSV file. Since we already loaded the `readr` package in an earlier code chunk, we don't need to do it again here:

```
# read in dataset
project <- read_csv('projectdata.csv')
```

> Note: in the example above, my code reads in a dataset called "projectdata.csv" and saves it as an object called `project`. You can give your data whatever* name you like! Just make sure to use the same name throughout. For example, if you named your data `kiva` (`kiva <- read_csv('...')`) then run the `ggplot` function on "kiva" as well (`kiva %>% ggplot(...)` in the next question).
>
> *there are a few rules to naming objects in R. Names can't start with a number, can't include some special characters like @, /, &, or spaces, and shouldn't be the name of something that already exists (e.g., don't name your data "read_csv"). I recommend using a name that's short and informative, so it's easy to type and easy to remember what you called it.

> Another note: the `read_csv()` function prints out quite a bit of information when it loads in the dataset. You can hide these messages by adding `message = FALSE` to the first line of your code chunk. It should like something like this:
> `{r name-of-chunk, message = FALSE}`.

Research Question

For your chosen data set, write down a research question that you might be able to answer based on the data available. This question should involve a *quantitative* outcome variable and one or two explanatory variables.

> Question: There are no "correct" answers to this question, as long as you wrote down a research question that could reasonably be answered using your chosen dataset (i.e., there are variables in the dataset that we could use to investigate these questions), and the outcome variable is *quantitative* (numerical values, with units). I encourage you to limit the scope of your research question by focusing on the relationship between two, or at most three, variables.

Explain why you think this is an interesting research question; give us your motivation for studying this question.

> Justification/Motivation: Answers will of course vary. Tell us why this is an interesting question! You'll need to provide motivation/justification for studying this question in the *Introduction* section of your final report.

Part 3. Knit

Once you've answered all the questions above, you'll need to `Knit` this document and then save that knitted document as a PDF. See the assignment *Instructions* at the top of this document and on Moodle for more details.