Confound it all!

Unit 3 - Lab 2

Directions: Follow along with the slides and answer the questions in **BOLDED** font in your journal.

Good-bye button, hello code!

- Since your first forays into doing data science, you've *download*, *upload* and *load*-ed your data.
- And you've done so marvelously ... using a button.
- In this lab, we'll look at using `code` as an alternative to the *Import Dataset* button in the *Environment* tab.
- We'll import some data from an *observational study*.
- Clean up our new data.
- And then use it to explore *associations* and discover how *confounding factors* can hinder analyses.

Our new data

- In the late 1970's, a researcher in Boston began measuring the lung capacity of children to study the effects that childhood smoking had on lung health.
- You can find the data online, as it turns out, here: http://www.amstat.org/publications/jse/datasets/fev.dat.txt
- Rather than *download*-ing the data and then *upload*-ing and *load*-ing it, we'll pull the data straight from the webpage into R.
- The benefit of loading data with code is that it helps makes your code **reproducible**, meaning your analysis could be reproduced, from start to finish, by anyone else.
- Sharing code is one way scientists collaborate and verify each other's results!

Loading our data

- Copy/paste or type the `url` of the website containing the data into the following (Remember to wrap the url in quotes):

```
lungs <- read.table("type webpage
                     address here",
                    header=FALSE)
```

- The `read.table` function reads raw data and turns it into a data table that we can use in R.
- We write `header=FALSE` to let R know that our data is missing variable names.
- The `lungs` portion of the code is just the name we're giving the data.

Viewing your data

- After loading data, it's helpful to look at it to get a feel for what it *looks* like and to make sure nothing funny happened when we loaded it.
- Type the following to `View` your data:

```
View(lungs)
```

- **What do you notice is missing from the data?**
- Need a hint? Type `names(lungs)` to print the names of the variables.
- **Which variable do you think is the child's height? Which is lung capacity (measured in *Forced Expiratory Volume*)? Can you be sure?**

Fixing data and Viewing it again

- To add names to your data, run the following:

```
names(lungs) <- c("age","lung_cap","height",
                  "gender","smoke")
```

- **Type `View(lungs)` again and then describe in words what the above line of code did to your data.**

Analyzing our data

- Our `lungs` data is from an observational study.
- **Write down a reason the reseachers couldn't use an experiment to test the effects of smoking on children's lungs.**
- Observational studies are often helpful for analyzing how variables are related:
- Let's explore whether or not a person's height is related to their lung capacity!
- Create an `xyplot` that compares people's `height` on the x-axis and their `lung_cap` (lung capacity) on the y-axis.
  - **Write down the code you used to create your plot**

Forming conclusions

- **Based on your plot, finish the following research claim:**
- *Based on the information from our observational survey, it appears that taller people tend to have* _____ *lung capacities than people who are shorter.*
- **Does it seem reasonable that taller people would have larger lung capacity?**
- **Do you think the plot you made seems to back this claim up?**

Beware confounding variables

- Even though plotting `height` and `lung_cap` seemed to show evidence that taller people have larger lung capacities, we should remain skeptical.
- Our data is from an observational study which means the relationship we found might actually be caused by a *confounding variable*
- A **confounding variable**, in our case, would be one that impacts children's `height` and also their `lung_cap` in such a way as to make these two variables *seem* related even though they might not be.

Confounded and confused

- Create the following plots:

```
xyplot(lung_cap~age, data=lungs)
```

```
xyplot(height~age, data=lungs)
```

- **Describe what happen's to people's `height` and `lung_cap` as they get older. Describe why `age` might be a possible *confounding factor*?**
- **Could it be the case that `age` is related to the volume of people's `lung_cap` and that their `height` is not? Why?**
- **Using the data, could you disprove the claim: 'People with larger lung capacites absorb more oxygen which causes them to grow taller?'**

Don't despair

- By now, we might be feeling a little distressed about our data.
- It isn't at all helpful to help us decide if our hypothesis that people's heights and lung capacities are related.

- Really, all it's done so far is confuse us!
- **Write down your opinion about the usefulness of observational studies.**
- **How sad would you be to find out that many of the *"studies"* found on the news are based on *observational studies*?**

Observational studies are (often) useful!

- Even with all the weaknesses we've found with observational studies, a good argument can still be made to say people's height are related to their lung capacity.
- In the following slides, we'll argue that this is indeed the case by showing that `age` is not a confounding variable.
- This would then imply that our initial conclusion that height and lung capacity seem to be related is correct.

Age and gender

- Run the following:

```
histogram(~age|gender, data=lungs)
```

- Notice that the distribution of children's ages, for each gender are roughly similar.
- Because of this, we expect there to be roughly similar numbers of boys and girls for each age.
- Note: In our data, Males are coded as `1`s in the data. Females as `0`s.

Height and gender

- Next run the following:

```
histogram(~height|gender, data=lungs)
```

- In this plot, notice that males (the right-hand side of the plot), have quite a few people who are taller than 67.5 inches whereas the females (the left-hand side of the plot) only have a few.

Age, gender and lung capacity

- So we know that our genders have similar distibutions for age, but that there are a group of males that are taller than the females.
- If `age` is a reason for people having larger lung capacties then, we should see males and females evenly scattered throughout this plot:

```
xyplot(lung_cap~age, data=lungs,
       groups=gender)
```

- **Are males and females evenly scattered through the plot? Or is there an area where they appear to be seperated? (Where one set of points is above or below the other?)**

Confused no more!

- We started this lab by noting that it seemed people's heights were related to their lung capacity.
- We became suspicious of this relationship when we noticed that people's heights and lung capacities both increased as they aged.
- This led us to believe that age could be what's actually related to lung capacity instead of heights.
- In the argument, we then showed that we had evidence to disprove the relationship between age and lung capacity.
- So for us, based on our data, we find that there is evidence of a relationship between people's `height` and `lung_cap`

Remarks about our argument

- Even if the argument we've just laid out seems convincing:
- Is it necessarily absolutely certain? Not at all.
- Could there still be some confounding variable that we don't see which is causing the increase in people's lung capacities? Absolutely.
- But does it still lend evidence to a seemingly reasonable hypothesis? Yes!
- **What do you think of the argument made? Do you personally think it supports the hypothesis that taller people's bigger bodies might explain their larger lung capacity? Explain.**

Concluding remarks

- When analyzing data from an observational study, we should:
- Beware of confounding variables that cause us to find spurious (*fake* or *false*) relationships.
- Do our best to really test the validity of our conclusions.
- Be mindful that our conclusions might be correct, but we can't be absolutely certain.
- Many times, observational studies help us find relationships that we then try to confirm or deny by using experiments.
- Science done using an observational study isn't automatically incorrect or useless, but we should remember to take the conclusions with a grain of salt.