LAB 2 - LINEAR MODELS

MATH 1170

28 AUGUST 2018

In this lab, we'll investigate a linear model using R pertaining to air quality in Salt Lake City. By the end of the lab, we should be able to

- Write your own R script to plot data
- Write a linear model that approximates data
- Discuss limitations of linear models

Air Quality and Temperature in Salt Lake City

Writing code to access data in R

As any concerned Utahn should be, you seek to investigate the air quality situation in Salt Lake City. If you aren't concerned yet, check out the picture to the right.

Equipped with our knowledge of linear models and a spreadsheet of data concerning air quality and temperature in Salt Lake City, Utah, we should be able to make some progress. Download the data `lab2data.csv`. You can open up this spreadsheet in a program like LibreOffice Calc or Excel if you would like to examine it yourself. You'll see that each column has a name - that's the variable name that we'll use to access it.

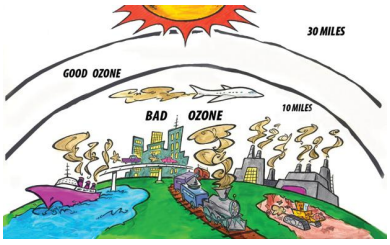
We are going to write our own R script that accesses some of this data and plots it. The idea is that we can type out all of the commands we want R to execute and run it all at once (instead of one at a time in the terminal window). To start a new `.r` file in RStudio, we click `File` -> New.

Next, we need to tell R to find this file. In the bottom right window, find wherever you saved your file on your computer and then hit `More` -> Set as Working Directory. Now, let's drag the downloaded `lab2data.csv` file to the same folder. Our first command should be to read in the file using the `read.csv` command:

```
dat <- read.csv("lab2data.csv")
```

Notice it is stored in a variable called `dat`.

If you just looked at the CSV file, it's easy to get overwhelmed and not really gain any deep insight. However, we'll use R to squeeze the knowledge out of this data. First, we will plot some data using our friend the `plot` function. Let's try plotting ozone (O_3) vs. mean temperature (*meantemp*). It should look like this:



I don't know what this is trying to convey, but whatever it is, I'm pretty sure it's wrong.

```
plot(O3 ~ meantemp, data=dat)
```

Here we use the ∼ operator to say "O3 as a function of meantemp." [1] We include dat so R knows to reference the correct data file. [2]

You can include any commands that you want in your R script - for example, you could add a title to the plot you have just created. Save your plot or copy it to a document using the fancy technique we discussed last time. If you write all of the commands in a `.r` file, you can easily modify it or run it again.

[1] note that `meantemp ~ O3` is different

[2] an alternative is to leave out `data=dat` and reference the variables as `dat$O3` and `dat$meantemp` respectively. You can kinda think of this as `dat` as their last name and `meantemp` as their first name.

Assignment for this week

1. Create a plot for amount of ozone (`O3`) in Salt Lake City as a function of temperature (`meantemp`). Using your plot, estimate the amount of ozone when the mean temperature is 65 degrees Fahrenheit. Make sure to add coherent, informative labels!

2. Come up with a linear function that you think best fits the data from the previous question. Write the equation for this line in the form $y = mx + b$, where m is the slope and b is the y-intercept.[3]

   (a) There is a nice R command called abline that will add a line to our graph. Define variables in your code called `myintercept` and `myslope` using your values of b and m. Add the following command to your script to plot a nice red[4] line with the data:

   ```
   abline(myintercept,myslope,lwd=2,col="red")
   ```

3. R has a function called `lm` which will find a best fit line for us. We can combine `lm` with abline to plot this line with the data. Add the following command to your R script:

   ```
   abline(lm(O3 ~ meantemp,dat))
   ```

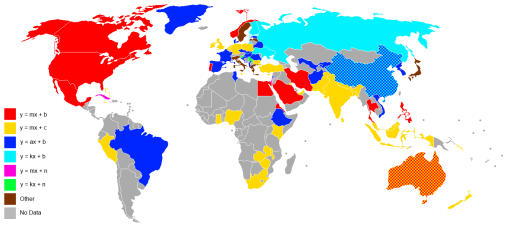
   Create this plot.

   (a) You can see the values for slope and y-intercept that R has chosen by calling

   ```
   model <- lm(O3 ~ meantemp,dat)
   model
   ```

   in the command line. The first value listed under coefficients is the y-intercept; the second is the slope. What is the equation for the linear model that R chose? Is the slope of your model more or less steep than the one that R chose?

[3] although, this naming convention seems to depend heavily on location



[4] my favorite plotting color happens to be `col=rgb(102,194,165,max=255)`. `figure out your favorite!`

4. Rewrite your R script so that you plot Carbon Monoxide (`CO`) as a function of maximum temperature (`maxtemp`). Use the `lm` command to find the best fit line and `abline` to plot it with the data.

5. Pollution particles smaller than 2.5 micrometers (`PM2.5`) are hazardous to human health because they are readily inhaled into human lungs, where they can be deposited. Some of the particles are carcinogenic (i.e., cancer-causing), and the body's normal defenses such as nose hairs and mucous in the upper respiratory tract cannot readily remove these tiny particles before they are deposited in the lungs.

   (a) Plot `PM2.5` as a function of minimum temperature (`mintemp`). For what range of temperatures should Salt Lake City residents be most concerned about $PM2.5$?

   (b) Do you think a linear model would be a good fit for this data? Why or why not?

   (c) Although we haven't learned any statistics, a way of measuring how good of a fit a linear model is to data is a quantity called "R-Squared".[5] R happily computes this. For the `O3` and `meantemp` model we've called `model`, we can see the R-squared value by typing

   ```
   summary(model)
   ```

   For this first model, we can ignore everything except

   ```
   ... Adjusted R-squared:  0.6927
   ```

   R-squared ranges from 0 (bad fit)[6] to 1 (good fit), so this is pretty good.
   Do the same analysis for the linear model relating `PM2.5` and `mintemp`. Does the R-squared support or refute your decision from part (b) about the goodness-of-fit of the linear model?

[5] which has no relation to the program we're using. an overwhelming number of things called "R", I know.

[6] you can really think of $R^2 = 0$ as a "0 slope" through your data. in other words, there really is no relationship