Lab 12: Simple Linear Regression

Cities

In this lab we will be looking at the relationship between cities' populations and other statistics that might be related, such as the number of companies, the number of hotels, the number of cars or the municipal expenditures.

The data

We will use data about Brazilian cities. Our sample consists of a random sample of 60 cities from all 5573 cities in Brazil. It has been collected from various official websites and has been made available on kaggle.

We will try to predict the population of a city using a number of other variables. These variables are described in the *data dictionary*, which can be read is as `dic`.

```
library(tidyverse)
dic <- read_csv("https://www.dropbox.com/s/pwbvn51x4o1fvh9/data_dic.csv?dl=1")
```

1. What do you think the distribution of the variable `pop` looks like? Which plot would you use to show this distribution?
2. *Choose a variable that you think has a strong relation to `pop`. How would you plot these two variables? Sketch a plot (paper and pencil or tablet) of what you expect. Describe how you expect the relation to look like. Is it linear? If yes, is it a positive or negative relation?

Data description: visualization

Now that you have articulated some expectations for what this data might look like, read in the data.

```
cit <- read_csv("https://www.dropbox.com/s/vx3tmh3ybwtbqk7/cities.csv?dl=1")
```

3. Create two new columns: the log (base 10) of population and the number of companies and use them to construct a scatterplot. How would you describe the relation between these two variables? What is its form, strength and direction?
4. *Choose a variable besides the number of companies and visualize it's relationship with (log) population. Describe their relationship in words. You will probably need to use a log-transform here as well if you want to see more clearly what is happening. While a log transformation of a variable, x, with zeroes won't work, you can add one to that variable, then take the log (i.e. $\log_{10}(x+1)$).

Data description: numerical summaries

5. Compute the correlation coefficient between log population and the log number of companies of these cities. Does it confirm your visual interpretation of their relation?
6. *Now compute the correlation between the log population and the log-transform of a variable of your choice (from question 4) to quantify the strength of the linear relation. Does the strength of the correlation surprise you?
7. Instead of just numerically describing the relationship between log population and log number of companies using the correlation coefficient, you can fit a linear model using the method of least squares.

What are the coefficients of this linear model? Write out the form of the line that you could use to predict the log population of a city.

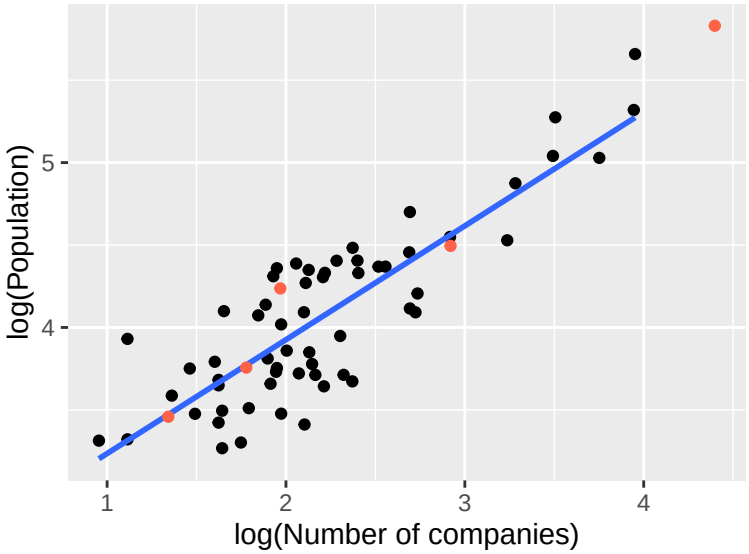
Making predictions

You can now use this model to predict log population of cities that are not in our dataframe. You can use this code to load all variables for 5 new cities:

```
cit_new <- read_csv("https://www.dropbox.com/s/37cafpcj3lrenke/cit_new.csv?dl=1")
cit_new <- cit_new %>%
  mutate(log_pop = log10(pop),
         log_comp = log10(comp))
```

8. Use the equation of the line you derived in the previous question to predict the value of log population for one new city. How does this predictions compare to the true value?
9. *Using the original data set, compute a linear model for that predicts log population using a variable of your choice. Use that model to predict the value of log population for one of the new cities.
10. *The plot below shows the original data in black, the linear model fit to that data, and the five new cities is red (or actually "tomato").

```
ggplot(cit, aes(x = log_comp, y = log_pop)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
    labs(y= "log(Population)",
         x = "log(Number of companies)") +
  geom_point(data = cit_new, aes(x=log_comp, y = log_pop),
             color = "tomato")
```



Construct a similar plot to this one, but use predictor variable that you chose and the corresponding model. Were the predictions that you model made for the 5 cities over or underestimates? Are the magnitudes of the errors generally smaller, larger, or the same as the residuals from the original data set? Do the predictions involve interpolation or extrapolation?