Simple Linear Regression + Cities Handout

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.

Prelude with the data dictionary

As data analysts we usually given .csv files and asked to run some analysis. But often the column names are uninformative. For example, the cities data frame we will use below has columns `mun_exp`, `cars`, `pop_braz`… but what do these mean?

Good collaborators will share a data dictionary in addition to the raw csv file.

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

1. What the the variables `mun_exp`, `cars`, `pop_braz` mean?
2. Which other variable do you think will have the strongest correlation with the `mun_exp` variable?

Descriptive analysis

Now lets load the cities data frame

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

3. Make a plot to visualize the distribution of `mun_exp`. Are there any outliers? If so which cities do these outliers correspond to?
4. Which variable do you think has the highest correlation with `mun_exp`. Calculate this correlation and make a plot to visualize this bivariate association.

Simple linear regression

5. Fit a linear regression model predicting `mun_exp` from `pop` and name the object `linear_model`.
6. Create a data frame called `reg_data` as follows
   - Start with just the columns `mun_exp`, `pop`, and `pop_for`.
   - Add a column `mun_exp_pred` that has the linear model predictions for `mun_exp` **without** using the `predict` function i.e. find the slope/intercept from `linear_model` and calculate the predictions with the formula using tidyverse.
   - Add a column `mun_exp_resid` that has the residual for the linear model predictions
7. Calculate the residual sum of squares and the R Squared value from `reg_data`. An alternative formula for R^2 is given by

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

   where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares and $TSS = \sum_{i=1}^{n}(y_i - \overline{y}_i)^2$. It just takes a little bit of algebra to show that this formula is equivalent to the SSR/TSS given in the notes.
8. Can you given an intuitive explanation for what is going on in the above formula for R^2? Can you convince someone this is a good way to measure model fit? What is the advantage of $\frac{\text{RSS}}{\text{TSS}}$ compared to just RSS alone.
9. Compare the R^2 value you calculated to the R^2 that comes from the broom package.

Linear regression with intercept

10. Create a new boolean column in `cities_df` called `many_for` that is TRUE for the cities that have at least 10 foreigners.
11. Fit a new linear model that predicts `mun_exp` from both `pop` and `many_for`
12. Add a new column to `reg_data` called `many_for_binary` that is the indicator variable for `many_for`. Hint: you many want to use the `ifelse`() function.
13. Add a new column to `reg_data` called `second_pred` that has the predictions for the second linear model. Again you should calculate these predictions manually as in question 7 (i.e. using a formula).

Verify `second_pred` is equal to the output of the `predict` function.

14. Sketch out by hand what you think the following plot should look like

    - `mun_exp` vs `pop` scatter plot
    - Line showing the predictions for the first model with one covariate in blue
    - Line showing predictions for the second model with two covariates in red with dashed lines

15. Make the plot described in the previous question. Hint: `linetype` argument in `geom_line`.

From examinig this plot, what is your takeaway about the `many_for` variable?