INFO370 Lab: simple regression

February 5, 2022

Instructions

This lab is an exercise of simple regression. You are usings the same datasets as in PS2, and you are doing similar analysis, just instead of relying on graphs and averages only, you now do it using linear regression.

1 Income and education

The states data (extracted from R dataset *states.x77*) contains data about 50 US states, collected in 1970s. Each row represents one state. We need variables
Income per capita income (1974)
HSGrad percent high-school graduates (1970)

1. load file *states.csv* and make some quick checks: ensure you loaded the correct file, it actually contains data, and the data loaded correctly. You can just check the numbers of rows/columns, and print a few lines.

These data does not contain any missings so you do not have to check missings.

Now let's replicate what you did in PS2 with income and HS graduation rate. Proceed as follows:

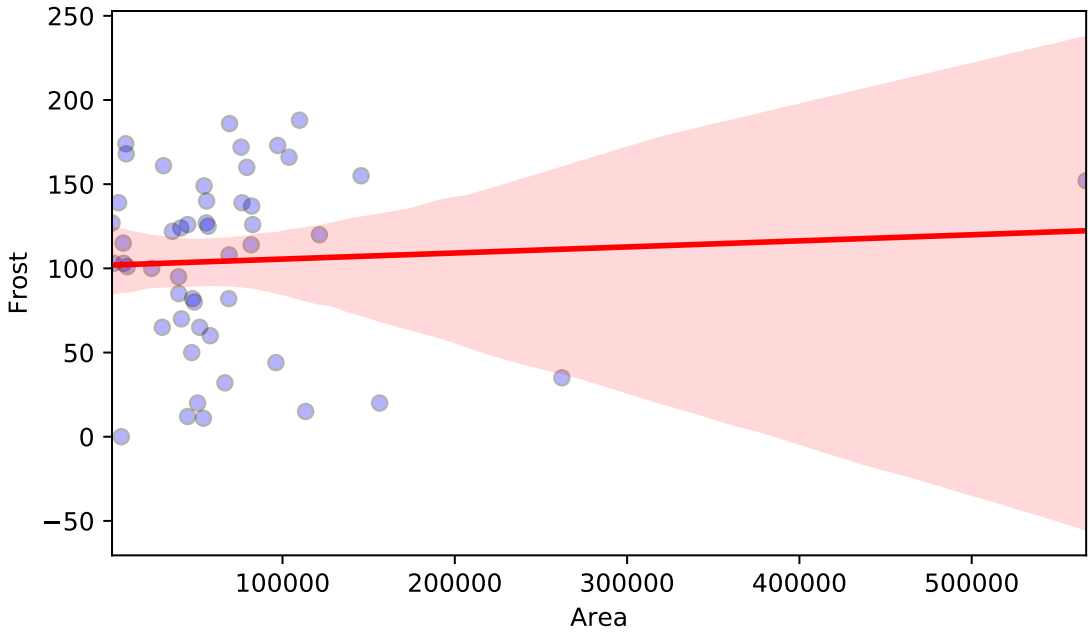
2. Before you analyze the relationship: what did you find in PS2–how is income related to HS graduation rate?

3. create a plot of *Income* (vertical) versus *HSGrad* (horizontal), and add the trend line on the plot.

   It is a little bit messy to plot trend (regression) line with matplotlib. Seaborn makes it much simpler (but has other issues). You can do something like

```
import pandas as pd
import seaborn as sns

states = pd.read_csv("../../data/geography/states.csv.bz2", sep="\t")
sns.regplot(y = "Frost", x = "Area",
            scatter_kws = {"color":"blue", "alpha":0.3, "edgecolor":"black"},
            line_kws={"color":"red"},
            data=states)
```



for frost-area plot.

4. comment the plot: is the line upward or downward sloping? Do you see the dots trending up/down in a similar fashion as the line?

5. run the linear regression in the form

$$\text{Income}_s = \beta_0 + \beta_1 \cdot \text{HSGrad}_s + \epsilon_s$$

and show the regression output.

6. interpret the coefficients (β_0 and β_1). What do these numbers mean? Are these statistically significant?

7. If you did this correctly, you see p-value for HSGrad being "0.000". What does this number mean?

8. Do you get similar results as what you got in PS2? (Hint: you should)

---

2 Global temperature trends

Next, let's analyze trends in global temperature. Use the same global temperature dataset (UAH-lower-troposhphere-wide.csv) you used in PS2. The main variables there are
year 1978-1922
month 1-12
globe global temperature deviation, deg C from 1991-2020 average.
You need to add a variables like $time = year + month/12$. In PS2 we analyzed the trend graphically, now we check what does linear regression show here.

1. Load data. Do a quick check to ensure it is good.

2. Create the *time* variable for continuous months.

3. Make a similar plot where you show the monthly data points and a trend line. Try to make the plot to look good.

4. Perform a linear regression in the form

$$globe_t = \beta_0 + \beta_1 \cdot time_t + \epsilon_t. \tag{1}$$

   Display the output table.

5. If you did this correctly then your coef for "time" should be 0.0135. What does this number mean? Is it statistically significant?

6. What does intercept mean here?

7. Check out the published data. You may read the "GTR" (global temperature report) from University of Alabama, Huntsville, or you may look for other sources. Did you get a similar trend as published?

   Note: we are looking at satellite-based temperature (temperature in "lower troposphere"). Data based on other measurements, such as ground stations, may be different.
