Lab 0

Anthony Ozerov

August 30, 2024

This lab is based on materials by Chengzhong Ye, Theo Saarinen, Omer Ronen, James Duncan, Tiffany Tang, Zoe Vernon, Rebecca Barter, Yuval Benjamini, Jessica Li, Adam Bloniarz, and Ryan Giordano.

Note: This lab is not representative of the labs that you will receive in this class. Future labs will be significantly more open-ended and difficult.

This lab will not be for a grade; you do not have to complete the lab if you don't want to, **but you do need to submit *something* on GitHub** (even if it is a blank `lab0.tex` and `lab0.pdf` file). This lab is an opportunity to make sure that you know how to submit your assignments, and for you to learn a little bit of Git/GitHub, Python, and various tools like `conda`.

Recall that instructions pertaining to *all* labs are in the `lab-instructions.pdf` file in the `disc/week1` directory of the `stat-215-a` repository. You should have already read this document before starting this lab.

1 Analysis Instructions

Write up a report conducting the following analysis. As described in the general lab instructions, the report can either be contained in a Jupyter notebook (`code/lab0.ipynb`) or a LaTeX document (`report/lab0.tex`). In either case, you should convert the final product to a PDF (`report/lab0.pdf`).

This walkthrough will be a quick overview of important functions/tools that you may find useful in future labs. If you are not familiar with Python (especially pandas, matplotlib, and scikit-learn), this lab is highly recommended.

1.1 Loading the data

1. If you have not set up your `stat-215-a` GitHub repo yet, wait until the first lab session (August 30) to do so.
2. Make sure to pull the repo before doing anything! I may have updated something.
3. Open a Python notebook (either in Jupyter Lab or VSCode) and load `USArrests.csv` and `stateCoord.csv` using Pandas. (see `pd.read_csv()`)

1.2 Manipulating the data

1. Merge the two datasets together into a single DataFrame named `arrests`. Hint: see `pd.merge`. Check that this worked correctly.

1.3 Visualizing the data

1. Plot "Murder" vs "Assault" using `matplotlib` (see `plt.scatter`). What do you see?
2. Plot "Rape" vs "urban population" using `matplotlib`. There should be an outlier. Can you mark the outlier with a different color?
3. Re-make these plots with the state names instead of the points (use `plt.annotate`). Do you notice anything interesting?
4. Challenge: Plot a map of the US colouring each state by its "Murder" rate. Check out `geopandas`.

1.4 Regression

You can fit a linear regression using `sklearn.linear_model.LinearRegression` (or manually if you'd prefer!).

1. Remove the "Murder" and "Assault" columns from the `arrests` DataFrame (you can index a DataFrame with multiple column names!)
2. Fit a linear regression of "UrbanPop" on "Rape".
3. Plot predicted values versus the residuals. Do you see any trends?
4. Replot "Rape" vs "UrbanPop" and draw a blue line with the predicted responses.
5. Now refit without the outlier and add a red line on the same plot.

6. Compare the lines. Are the linear responses a good description of the data?

7. Make a publishable graph. Add a header (`plt.title`), axis labels (`plt.xlabel`, `plt.ylabel`), add a legend (`plt.legend`), and generally try to make the plot look nice.

1.5 Challenge

Try making some of the plots in R instead, using the methods suggested in `lab-instructions.pdf`