GENERAL ASSEMBLY

INTRODUCTION TO REGRESSION ANALYSIS

Jonathan Balaban

DAT2

LEARNING OBJECTIVES

- Define data modeling and simple linear regression
- Build a linear regression model using a dataset that meets the linearity assumption using the sci-kit learn library
- Understand and identify multicollinearity in a multiple regression

PRE-WORK REVIEW

- Effectively show correlations between an independent variable x and a dependent variable y
- Be familiar with the get_dummies function in pandas
- Understand the difference between vectors, matrices, Series, and DataFrames
- Understand the concepts of outliers and distance.
- Be able to interpret p values and confidence intervals

WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

- Data has been **acquired** and **parsed.**
- Today we'll **refine** the data and **build** models.
- We'll also use plots to **represent** the results.

INTRODUCTION

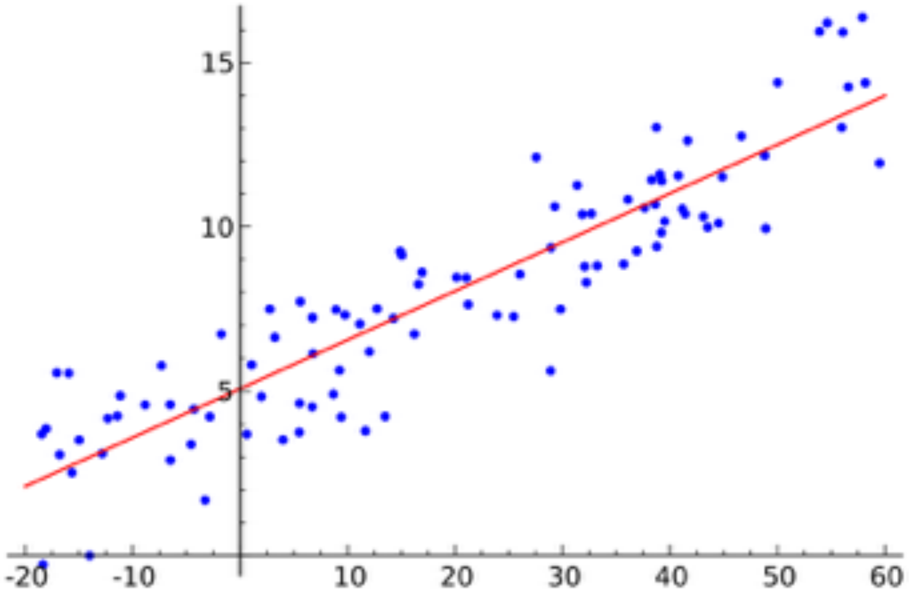
SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION

- Def: Explanation of a continuous variable given a series of independent variables

- The simplest version is just a line of best fit: $y = mx + b$

- Explain the relationship between **x** and **y** using the starting point **b** and the power in explanation **m**.



SIMPLE LINEAR REGRESSION

- However, linear regression uses linear algebra to explain the relationship between *multiple* x's and y.

- The more sophisticated version: y = beta * X + alpha (+ error)

- Explain the relationship between the matrix **X** and a dependent vector **y** using a y-intercept **alpha** and the relative coefficients **beta**.

SIMPLE LINEAR REGRESSION

- Linear regression works **best** when:
  - The data is normally distributed (but doesn’t have to be)
  - X’s significantly explain y (have low p-values)
  - X’s are independent of each other (low multicollinearity)
  - Resulting values pass linear assumption (depends upon problem)
- If data is not normally distributed, we could introduce *bias*.

SIMPLE LINEAR REGRESSION ANALYSIS IN SKLEARN

- Sklearn defines models as *objects* (in the OOP sense).
- You can use the following principles:
  - All sklearn modeling classes are based on the base estimator. This means all models take a similar form.
  - All estimators take a matrix **X**, either sparse or dense.
  - Supervised estimators also take a vector **y** (the response).
  - Estimators can be customized through setting the appropriate parameters.

CLASSES AND OBJECTS IN OOP

- **Classes** are an abstraction for a complex set of ideas, e.g. *human*.

- Specific **instances** of classes can be created as **objects**.
    - *john_smith = human()*

- Objects have **properties**. These are attributes or other information.
    - *john_smith.age*
    - *john_smith.gender*

- Object have **methods**. These are procedures associated with a class/object.
    - *john_smith.breathe()*
    - *john_smith.walk()*

DEMO: REGRESSING AND NORMAL DISTRIBUTIONS

- Work through /starter-code-6.ipynb in pairs.
- The first plot shows a relationship between two values, though not a linear solution.
- Note that lmplot() returns a straight line plot.
- However, we can transform the data, both log-log distributions to get a linear solution.

MULTIPLE REGRESSION ANALYSIS

MULTIPLE REGRESSION ANALYSIS

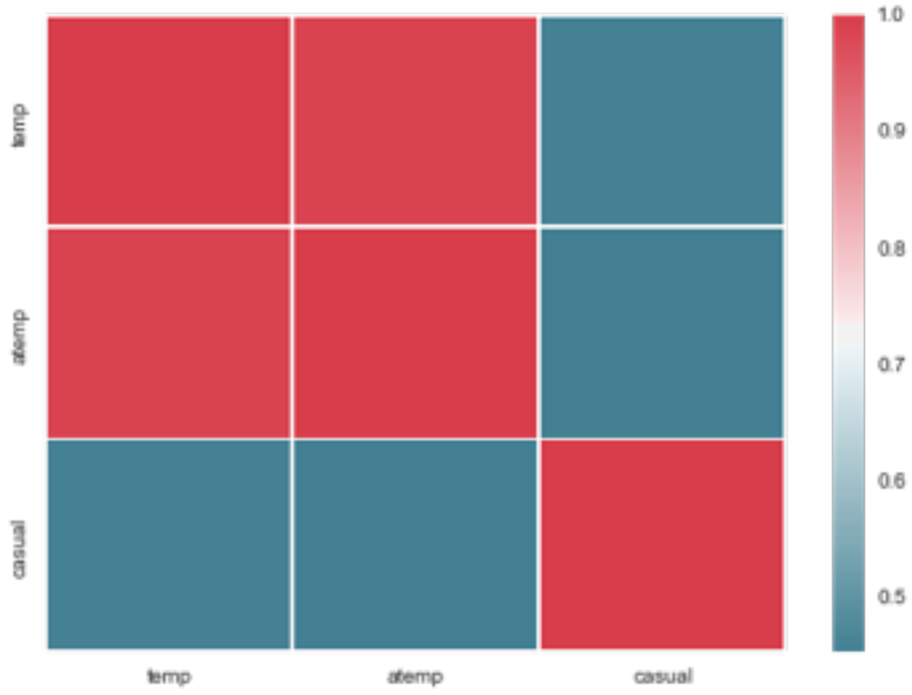
- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful.
- We want our multiple variables to be mostly independent to avoid multicollinearity.
- Multicollinearity, when two or more variables in a regression are highly correlated, can cause problems with the model.

BIKE DATA EXAMPLE

- We can look at a correlation matrix of our bike data.

- Even if adding correlated variables to the model improves overall variance, it can introduce problems when explaining the output of your model.

- What happens if we use a second variable that isn't highly correlated with temperature?



CONCLUSION

TOPIC REVIEW

CONCLUSION

- You should now be able to answer the following questions:
  - What is simple linear regression?
  - What makes multi-variable regressions more useful?
  - What challenges do they introduce?
  - How do you dummy a category variable?
  - How do you avoid a singular matrix?

UPCOMING WORK

Final Project: Part 1 due L8

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET!