GENERAL ASSEMBLY

INTRODUCTION TO REGRESSION ANALYSIS

Jonathan Balaban

DSI

LEARNING OBJECTIVES

- Define data modeling and simple linear regression
- Build a linear regression model using a linear dataset and sklearn
- Understand and identify multicollinearity in a multiple regression

PRE-WORK REVIEW

- Show correlations between independent variables X, and y
- Use get_dummies in pandas
- Understand the difference between vectors, matrices, Series, and DataFrames
- Understand the concept of outliers
- Interpret p-values and confidence intervals

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()*

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.

MULTIPLE REGRESSION ANALYSIS

- Simple linear regression with one variable can predict a response, but using multiple variables can be much more powerful and accurate.
- 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 model problems.

SIMPLE LINEAR REGRESSION

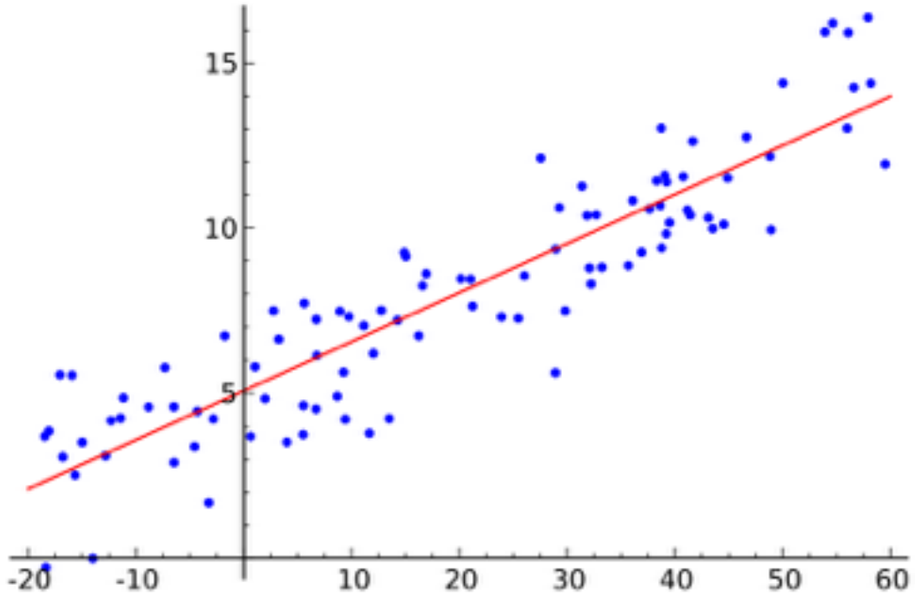
SIMPLE LINEAR REGRESSION

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

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

- Models relationship between **X** and **y** via **m,** and the starting point **b**.

- Interactive guide: setosa.io/ev/ordinary-least-squares-regression



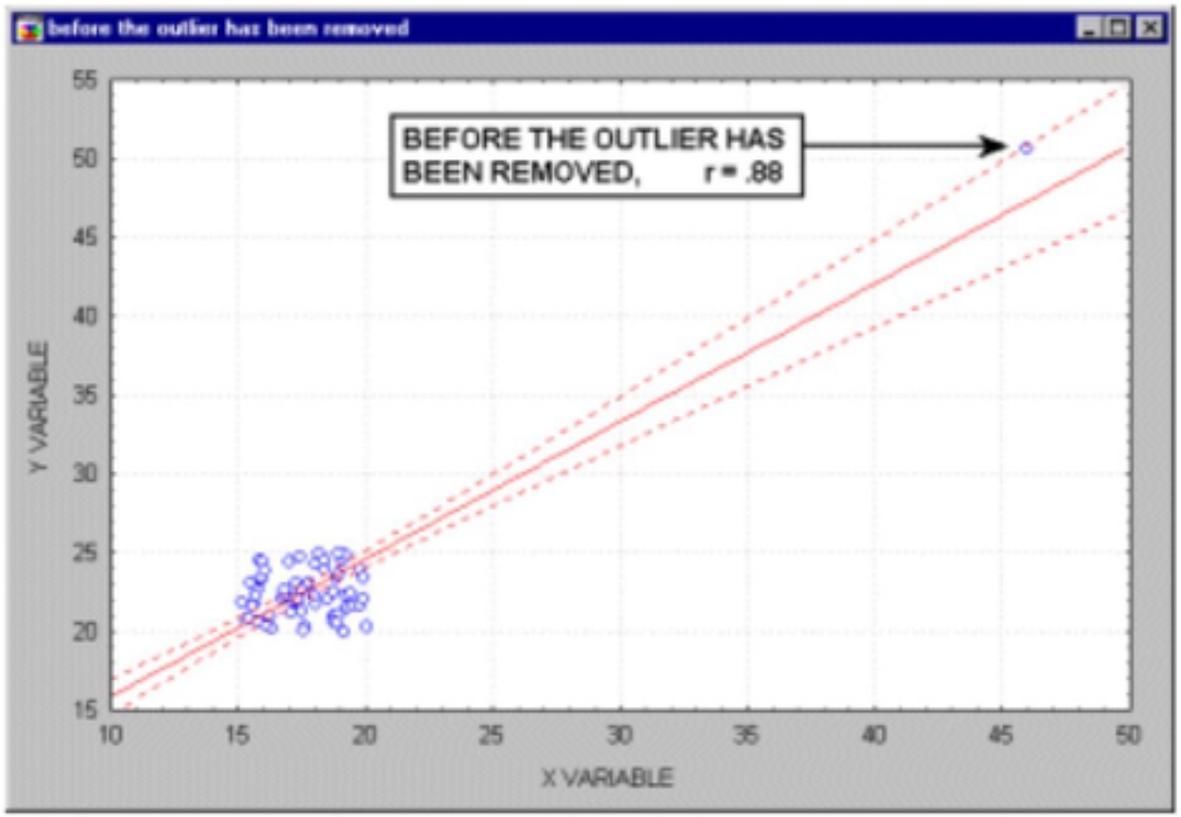
SIMPLE LINEAR REGRESSION

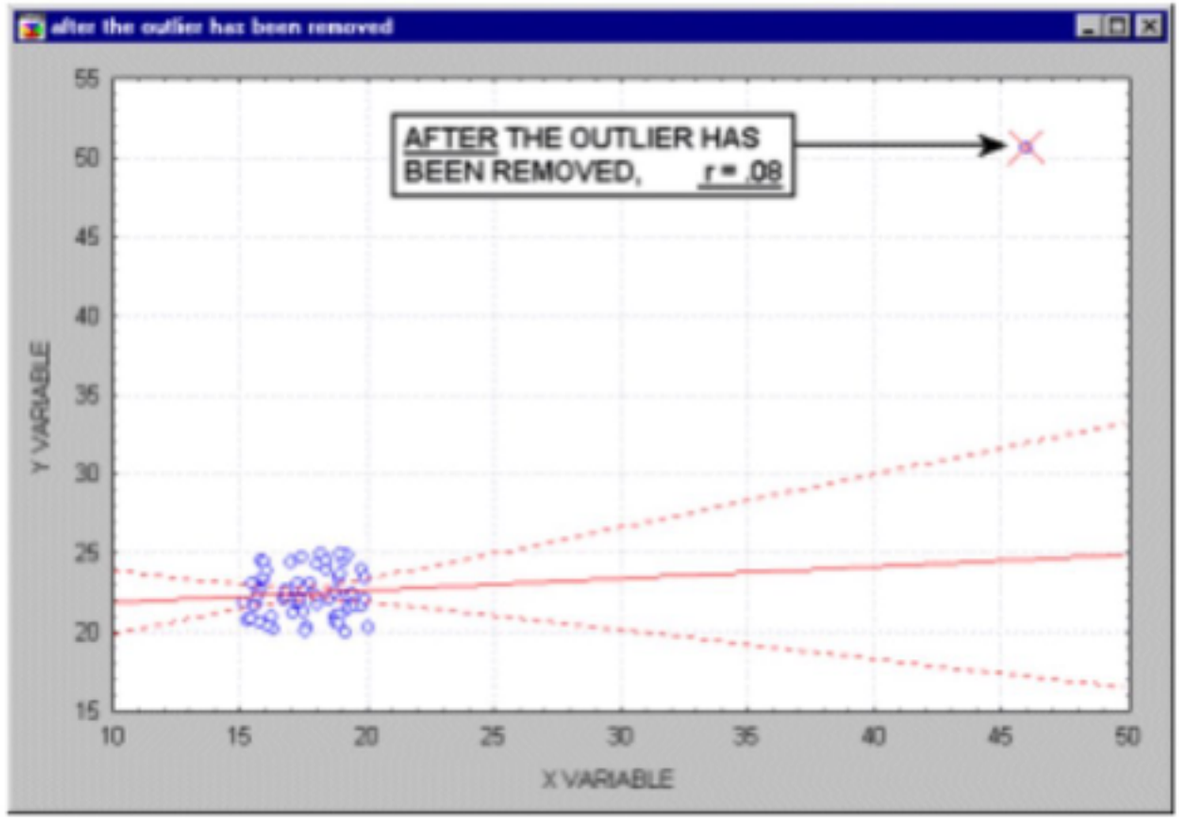
- Linear regression uses linear algebra to explain the relationship between *multiple* x's and y.
- The more sophisticated version: y = beta * X + alpha (+ error)
- Explain the relationship between the matrix **X** and a dependent vector **y** using a y-intercept **alpha** and the relative coefficients **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*.

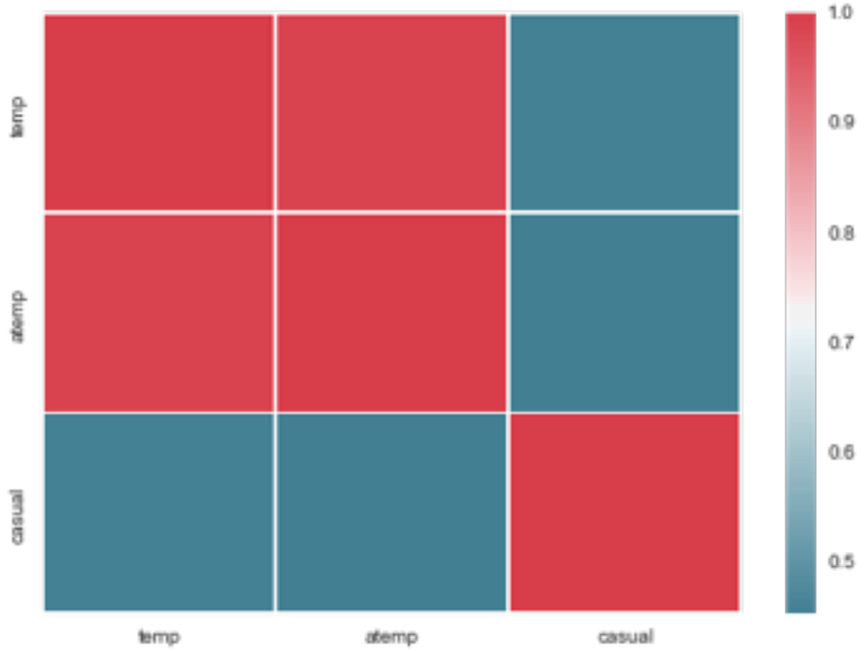
OUTLIERS


before the outlier has been removed
BEFORE THE OUTLIER HAS
BEEN REMOVED, r = .88
Y VARIABLE
X VARIABLE
55
50
45
40
35
30
25
20
15
10
15
20
25
30
35
40
45
50


after the outlier has been removed
AFTER THE OUTLIER HAS
BEEN REMOVED, r = .08
Y VARIABLE
X VARIABLE
55
50
45
40
35
30
25
20
15
10
15
20
25
30
35
40
45
50

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

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

INTRODUCTION TO REGRESSION ANALYSIS

Q & A