Applying Multiple Linear Regression

By Eric Schles

Outline

- Review of multiple linear regression (theory)
- How to do multiple linear regression (in python)
- Trying to use multiple linear regression to predict slavery
- Playing around with the data

Review: Multiple Linear Regression

Definition:

Multiple Linear Regression is a statistical technique used to determine a linear model consisting of 2 or more independent variables to determine a dependent variable

Review: Multiple Linear Regression

Linear regression:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

The y[i]’s are linearly related to the vector of x[i]’s

Where x[i] is of length p.

[demo]

Multiple Linear Regression (python)

https://github.com/EricSchles/MultipleLinearRegression/blob/master/linreg.py

Review: Multiple Linear Regression

The assumptions of Linear Regression:

- Linear relationship between independent and dependent variables
- Independent errors
- normal distribution of errors
- homoskedasticity

Review: Multiple Linear Regression

Independent Errors:

Here the assumption is that all the errors are statistically uncorrelated for each independent variable.

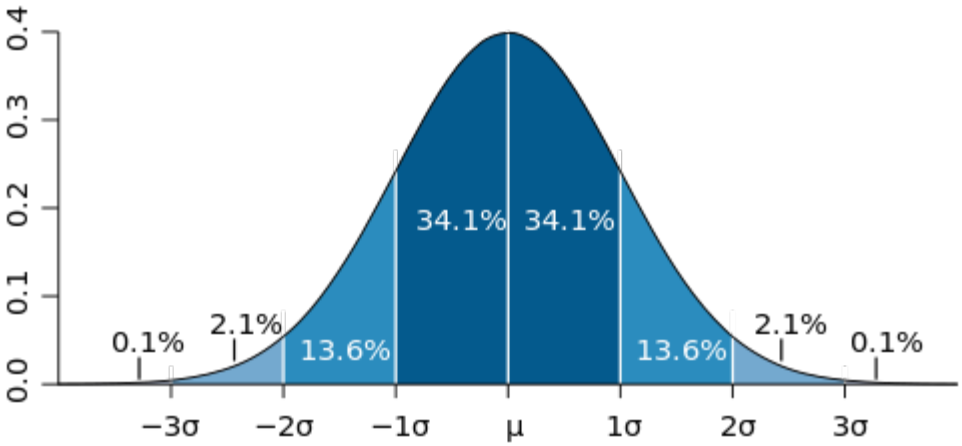
Testing for correlation (in python)

https://github.com/EricSchles/MultipleLinearRegression/blob/master/testingCorrelation.py

Remember to run help on pearsonr to explain

Normally distributed?

If a variable is normally distributed then the values of the data set will take on a similar-shape to the normal distribution



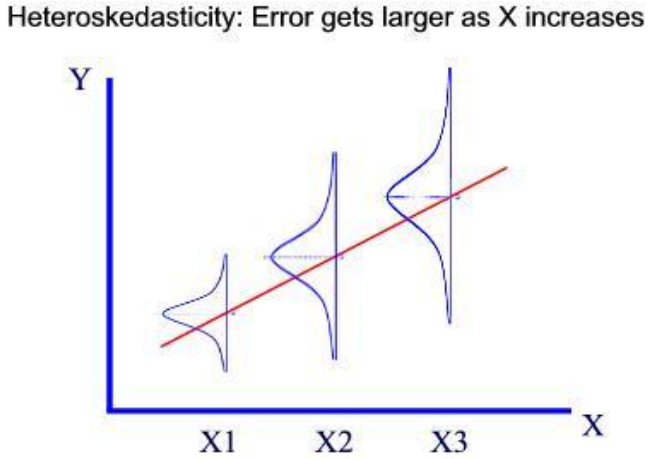
Normally distributed? (in python)

https://github.com/EricSchles/MultipleLinearRegression/blob/master/testingNormality.py

A p-value of less than .05 is likely to mean the distribution is NOT normally distributed

Homoskedasticity

Homoskedasticity is the assumption that the dependent variable exhibits similar amounts of variance across the range of values for an independent variable



Homoskedasticity (in python)

https://github.com/EricSchles/MultipleLinearRegression/blob/master/testingHomoskedasticity.py

Look at the last two parameters - the F-statistic is above 0.05 so we fail to reject H-null.

Cheat sheet

http://geog.uoregon.edu/geogr/topics/interpstats.htm

A general overview of testing

https://github.com/statsmodels/statsmodels/blob/master/examples/python/regression_diagnostics.py

Check this out on your own time!

Problem Background

Human trafficking is a serious crime and happens in many countries every year.

Some stats (be skeptical here)

- 27 million people
- it happens in 161 countries
- it generates 32 billion dollars in profit

Source:

http://www.cicatelli.org/titleX/downloadable/Human%20Trafficking%20Statistics.pdf

What I've done

I created a data set combining traditional economic data with figures from this report:

http://www.unodc.org/unodc/en/human-trafficking/global-report-on-trafficking-in-persons.html

How I did it

- Parse the pdf’s
- Wrote some crazy regex
- Generated this dataset:

https://github.com/EricSchles/MultipleLinearRegression/blob/master/trafficking_data.csv

How to parse pdf’s

http://thomaslevine.com/!/parsing-pdfs/

poppler-utils: http://packages.ubuntu.com/precise/poppler-utils

pdftotext -layout file.pdf

Introduction to regex

Regular expressions, often called regex, is a way to find patterns in strings.

While Regex is not the easiest thing to pick up, it can be very powerful once you understand it well.

Regex’ing in python

https://github.com/EricSchles/MultipleLinearRegression/blob/master/regexing.py

A simple introduction

references on regex

http://regex.bastardsbook.com/

http://www.rexegg.com/regex-quickstart.html

https://docs.python.org/2/howto/regex.html

Questions for this data set

Is economic performance related to the level of human trafficking in a given country?

What are the economic indicators we should be looking for?

Is human trafficking a poor country problem?

Implementing Linear Regression

Implementing linear regression is actually quiet easy. All you need to understand is iteration and the notion of a loss function.

If you are familiar with newton methods in differential equations, then how this is done should be easy.

How to do it?

Gradient descent:

Gradient descent algorithms start from initial condition and then approximate a solution through iteration. The general equation is checked against the existing sample data for correctness.

some code

https://github.com/EricSchles/MultipleLinearRegression/blob/master/gradient_descent.py

Playing with data

Substituting in and out data is a simple process with pandas

https://github.com/EricSchles/MultipleLinearRegression/blob/master/intro_pandas.py