DATA SCIENCE
11 WEEK PART TIME COURSE

Week 2 – Regression
Wednesday 16th December

1. Motivation
2. Supervised Vs Unsupervised learning
3. What is Linear Regression?
4. How do Run a Linear Regression Model?
5. Lab
6. Discussion / Review / Homework

WHAT ARE THE GOALS OF STATISTICAL LEARNING?

scikit-learn
algorithm cheat-sheet
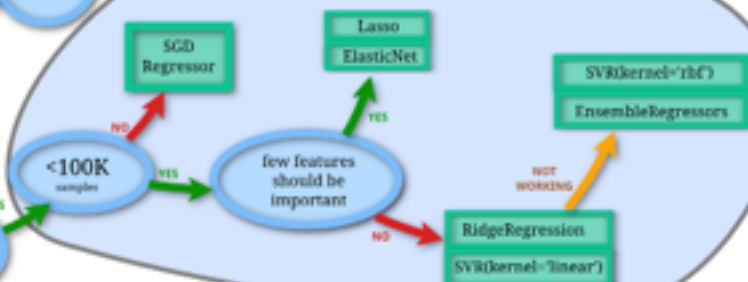
START

classification

- kernel approximation
- SVC
- Ensemble Classifiers
- KNeighbors Classifier — NOT WORKING
- SGD Classifier — NOT WORKING
- Naive Bayes — YES
- Text Data — NO — NOT WORKING
- Linear SVC
- <100K samples — YES

get more data — NO

>50 samples — YES

predicting a category — YES

do you have labeled data — NO — YES

regression

- SGD Regressor — NO
- Lasso ElasticNet — YES
- few features should be important
- SVR(kernel='rbf') EnsembleRegressors — NOT WORKING
- RidgeRegression SVR(kernel='linear') — NO
- <100K samples — YES

predicting a quantity — YES — NO

clustering

- Spectral Clustering GMM — NOT WORKING
- KMeans
- number of categories known — YES — YES
- <10K samples — YES
- MiniBatch KMeans — NO
- <10K samples — NO
- MeanShift VBGMM — YES — NO

just looking — YES — NO

dimensionality reduction

- Randomized PCA — NOT WORKING
- Isomap Spectral Embedding — YES
- LLE — NOT WORKING
- <10K samples — NO
- kernel approximation

predicting structure

tough luck

Back

scikit learn

We want to predict some value, let's call it **y**, based on some observed data we have, let's call that **x**.

We will use statistical learning to estimate a function that approximates **y** based on the input, **x**.

y is also called; label, dependent variable, target

x is also called; predictor, independent variable, features

We want to predict the price of a house, let's call it y, based on some observed data we have about the area, number of bedrooms, size of the house, and if it has a pool or not.

The area, number of bedrooms, size of the house, and if it has a pool or not would be our x variables (sometimes you might see this denoted as X)

What we want is y = f(X), a way to describe the house price based on observed data

If the y variable is numeric then we have a regression problem - we are trying to predict a continuous number

If the y variable is a category (for example trying to predict a type of flower) the we have a classification problem - we are trying to classify what group that y belongs to.

We want to find some underlying structure or patterns in the data but in this case we don't have any labeled data.

So for example, if we have a large group of customers but would like to separate them into groups (or clusters) to better target them.

WHAT IS LINEAR REGRESSION?

We want to model a linear relationship (think straight line) between our target variable y and our input variable x.

$$y = X\beta + \epsilon$$

$$y = X\beta + \epsilon$$

‣ y = target variable

‣ X = input variable

‣ β = coefficients

‣ ε = error term

Note, one of our input variables can be 1 so we have an intercept parameter

‣ Linear relationship in the parameters, β, we can transform the actual values of the inputs if we want

‣ Variance of the error term, ϵ, is constant. This means there is no systematic pattern in the values of X and the variance of ϵ

‣ The mean of $\epsilon = 0$

‣ ϵ has a normal distribution

‣ No perfect (or near perfect) co-linearity between any of the input variables. Otherwise the fitting procedure will break.

HOW TO RUN LINEAR REGRESSION?

$$SS_{res} = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

Basically, what we are trying to do is minimise the Residual Sum of Squares. This is the Sum of the squared difference between our observed value and the value from the model

$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Basically, what we are trying to do is minimise the Residual Sum of Squares. This is the Sum of the squared difference between our observed value and the value from the model

Oak Diameter vs. Age
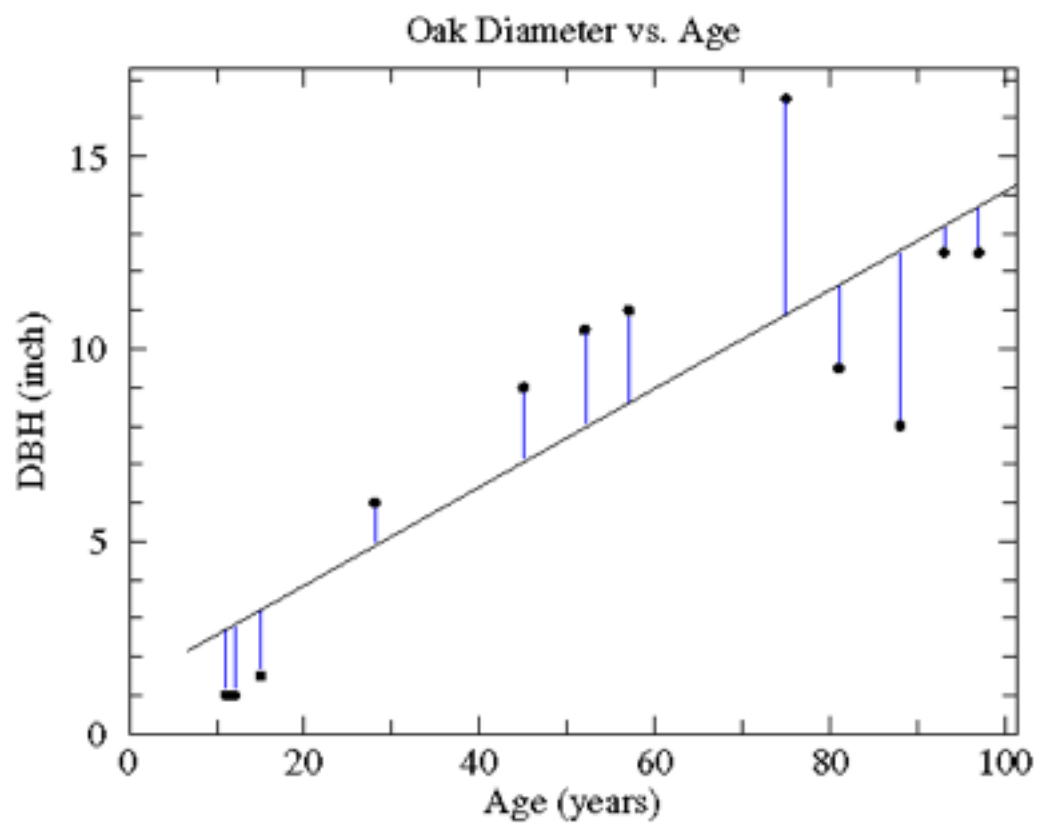
- Make sure you visualise your data and check the actual model fit !!!

- The fitting a model to the four datasets in the table on the right produce the same fit statistics, model coefficients and standard error

- See anything wrong?

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
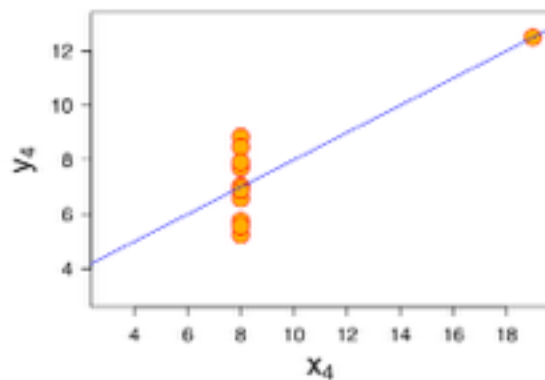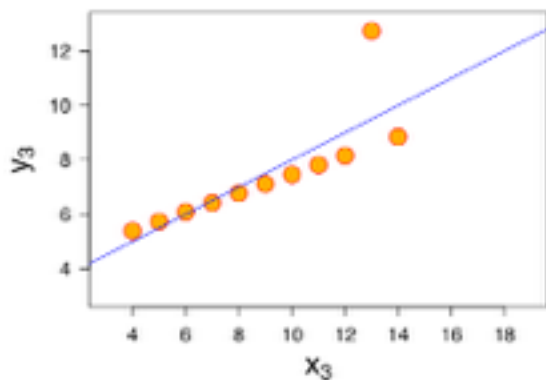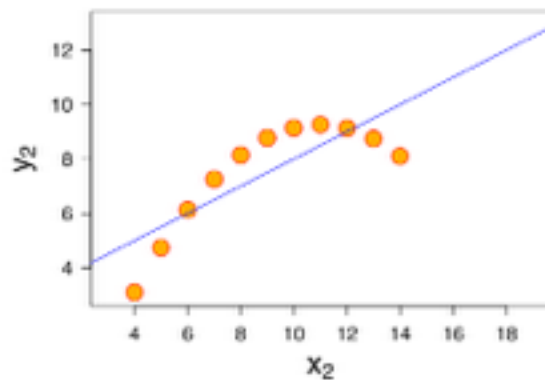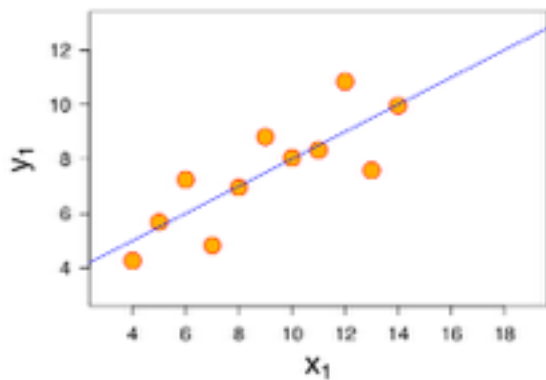
LAB

git remote -v

git remote add upstream https://github.com/ihansel/SYD_DAT_3.git

git remote -v

git fetch upstream

git checkout master

git merge upstream/master

DISCUSSION TIME

▸ Review of last class

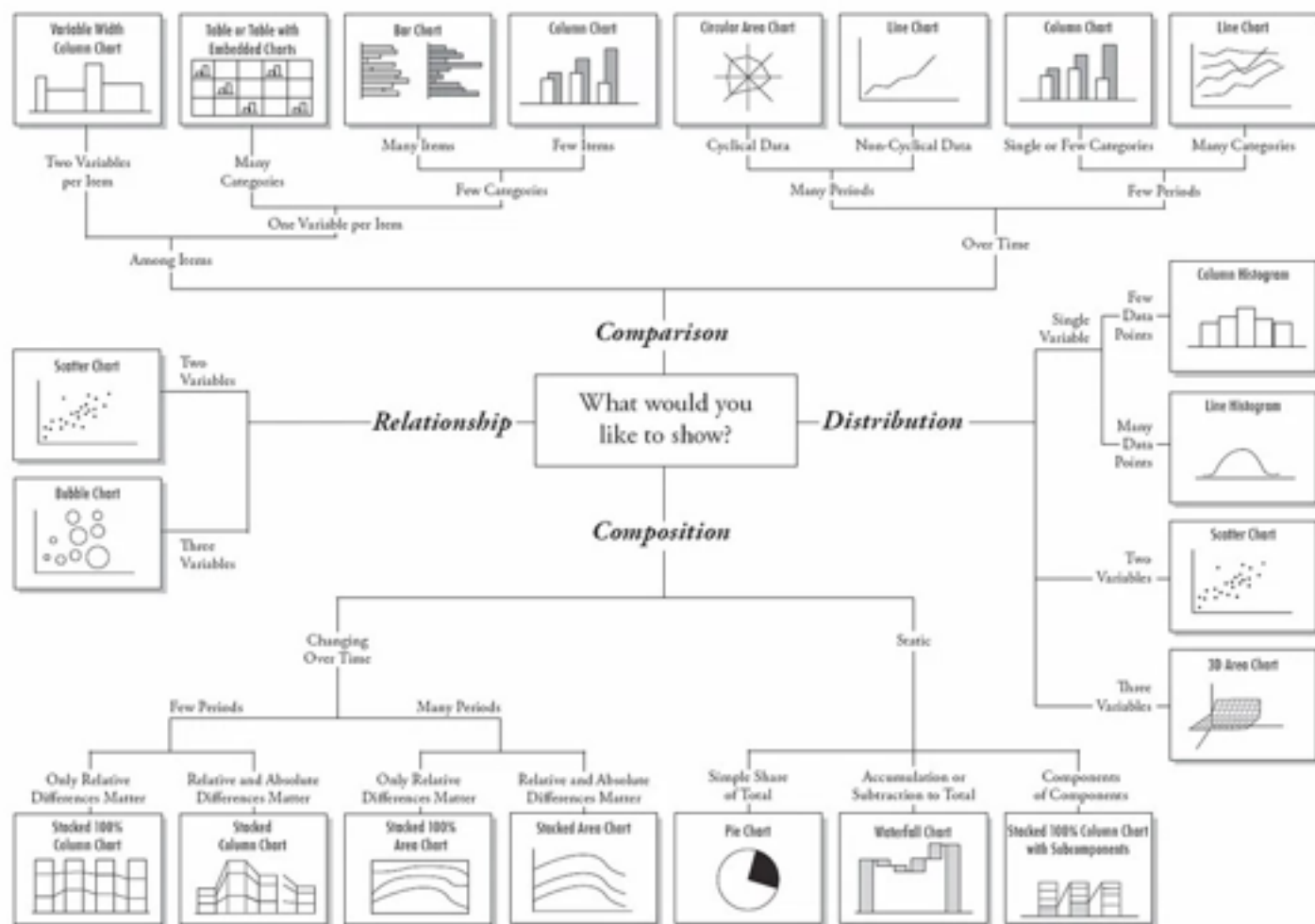▸ Further Reading for Regression

▸ Check in with homework/course project

Week 2 Monday 14th

☑ Understand goals of Data Viz.

☑ Visualise a dataset

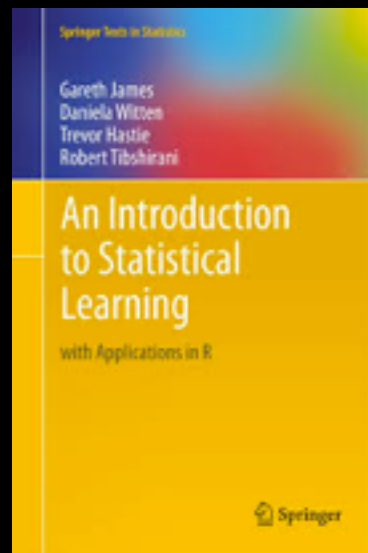☑ Understand 3 different graph types

☑ Examples & Sources to Review

Chart Suggestions—A Thought-Starter

Comparison

What would you like to show?

Relationship

Distribution

Composition

Variable Width Column Chart — Two Variables per Item

Table or Table with Embedded Charts — Many Categories

Bar Chart — Many Items

Column Chart — Few Items

Few Categories

One Variable per Item

Among Items

Circular Area Chart — Cyclical Data

Line Chart — Non-Cyclical Data

Many Periods

Column Chart — Single or Few Categories

Line Chart — Many Categories

Few Periods

Over Time

Scatter Chart — Two Variables

Bubble Chart — Three Variables

Column Histogram — Single Variable — Few Data Points

Line Histogram — Many Data Points

Scatter Chart — Two Variables

3D Area Chart — Three Variables

Changing Over Time

Static

Few Periods / Many Periods

Only Relative Differences Matter — Stacked 100% Column Chart

Relative and Absolute Differences Matter — Stacked Column Chart

Only Relative Differences Matter — Stacked 100% Area Chart

Relative and Absolute Differences Matter — Stacked Area Chart

Simple Share of Total — Pie Chart

Accumulation or Subtraction to Total — Waterfall Chart

Components of Components — Stacked 100% Column Chart with Subcomponents

DISCUSSION TIME

An Introduction to Statistical Learning

‣ **Chapter 3 – Linear Regression**

‣ **Chapter 6 – Linear Model Selection and Regularization**

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Data Science Week 1 Monday

- [x] Course Expectations
- [x] What the Course Involves
- [x] Course Project Information (Basics)
- [ ] Lab-Setup
  - [ ] Anaconda Python
  - [x] Git/Github
  - [ ] Slack

Wednesday 9th December

- [x] A Definition of DS
  - [x] Git-Forking
- [x] Running Jupyter+Pandas
- [x] Data Manipulation in Python
- [x] Homework

DISCUSSION TIME

Homework/Course Project

‣ **How's it going ?**

‣ **Purpose is to set you up so you can do a little bit of your project over the break**

‣ **Also, anyone interested in participating in a kaggle competition?**

kaggle