GENERAL ASSEMBLY

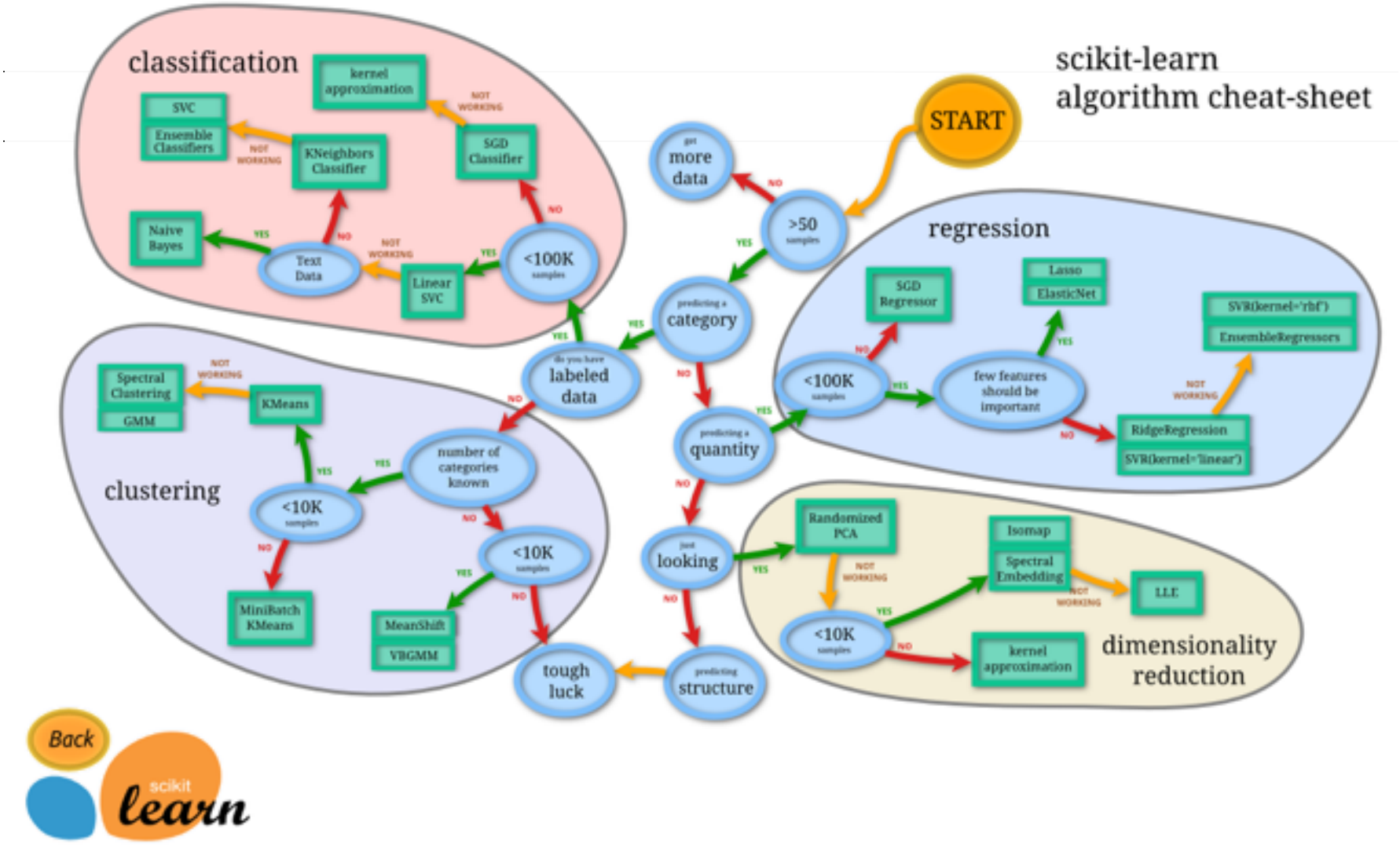
DATA SCIENCE

SYD DAT 6

Week 2 – Linear Regression
Wednesday 19th October

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
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
predicting a category
YES
NO
do you have labeled data
YES
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
<10K samples
YES
MeanShift
VBGMM
NO
just looking
YES
NO
dimensionality reduction
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
tough luck
predicting structure
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

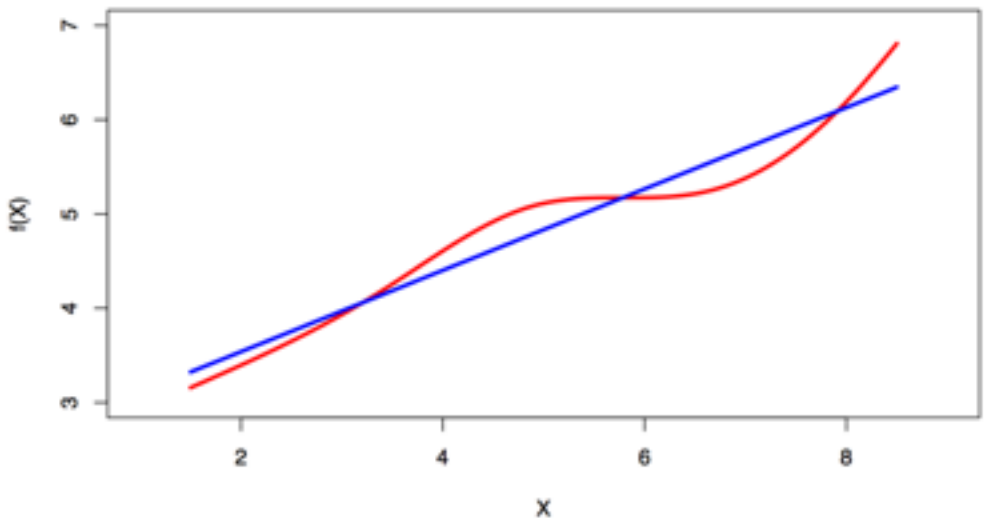
Note, one of our input variables can be 1 so we have an intercept parameter

Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on

X1, X2 ,… Xp is linear.

Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on X1, X2 ,… Xp is linear.

True regression functions are never linear



- 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.

DATA SCIENCE PART TIME COURSE

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

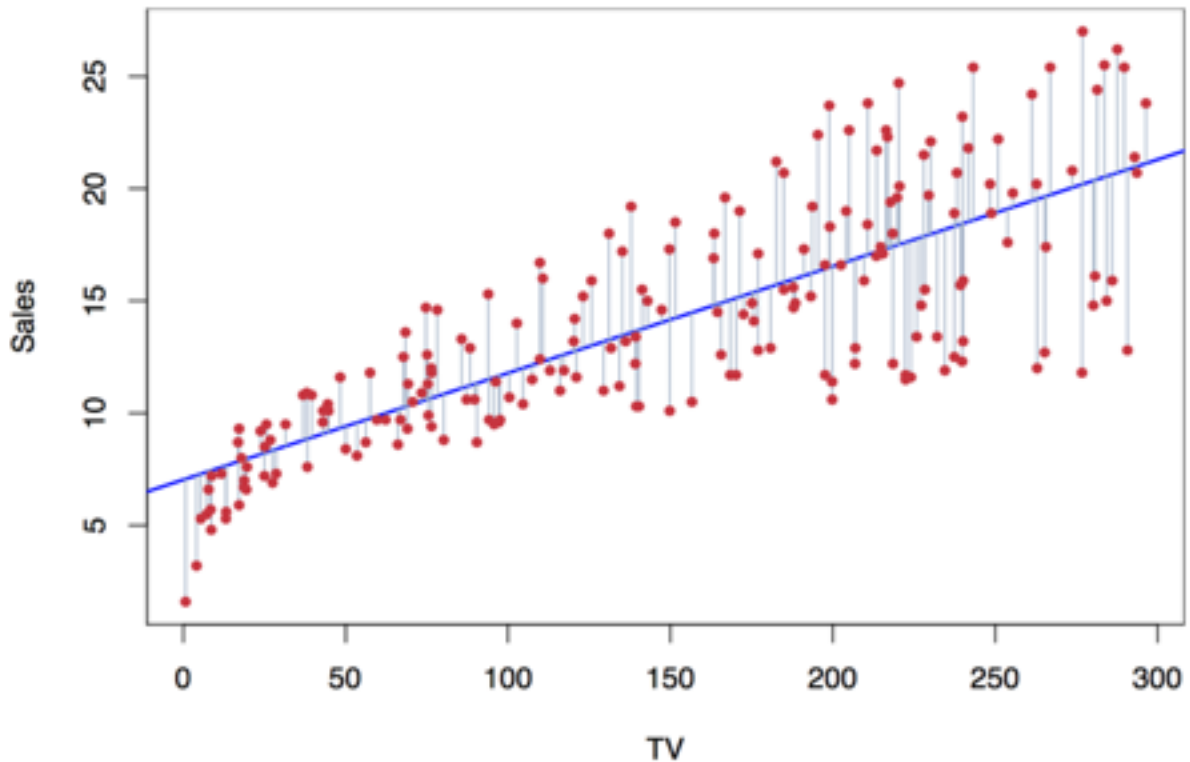
1 $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

2 $$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

3 $$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

4 $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

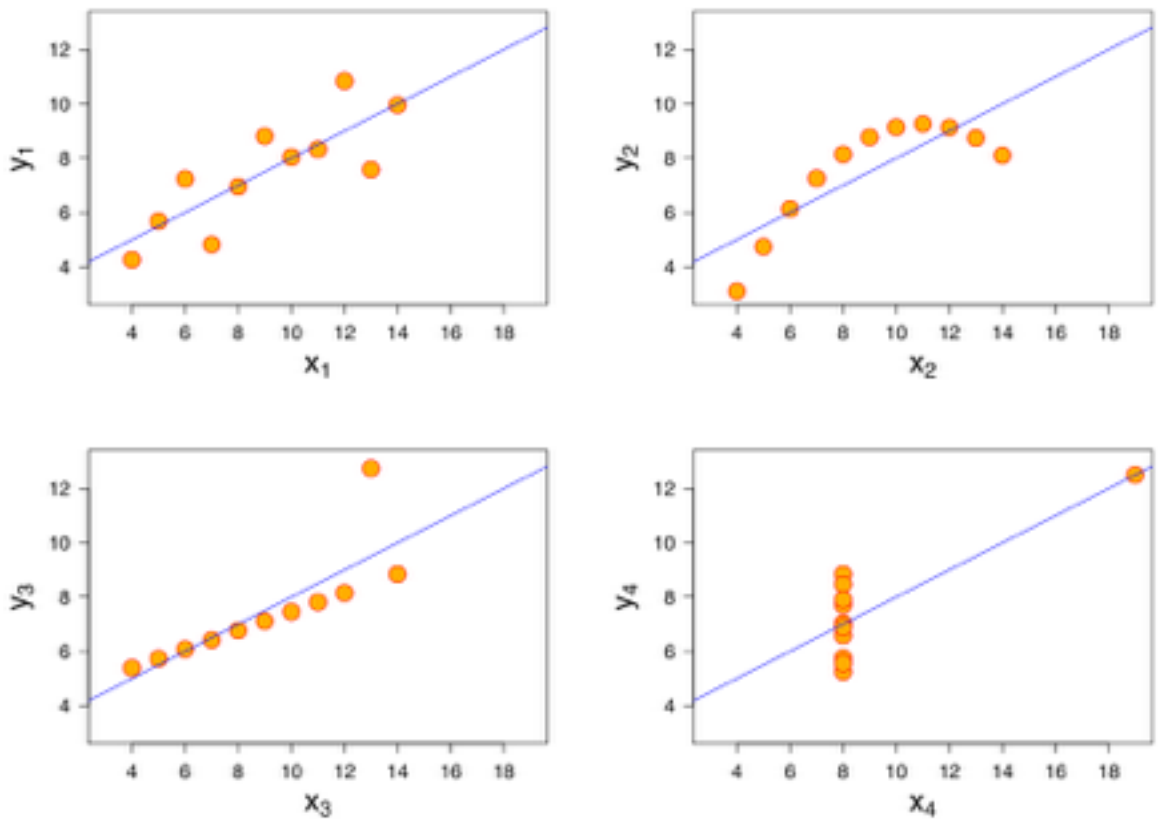
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$


Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV

- 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


y1
x1
y2
x2
y3
x3
y4
x4

DATA SCIENCE PART TIME COURSE

CONFIDENCE

Residual standard error

$$\text{RSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R-squared

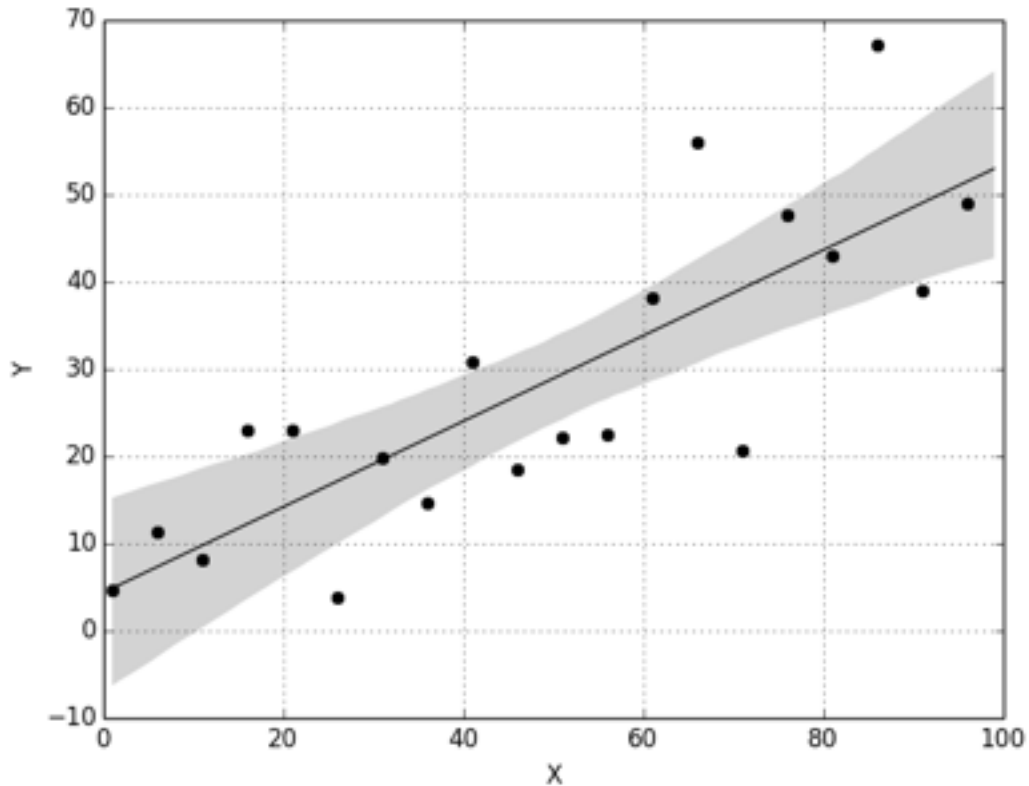
$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$


70
60
50
40
30
20
10
0
−10
Y
0
20
40
60
80
100
X

If $\beta 1 = 0$

then the model reduces to $Y = \beta 0 + \varepsilon$,

and X is not associated with Y

Null hypothesis: $H_0 : \beta_1 = 0$

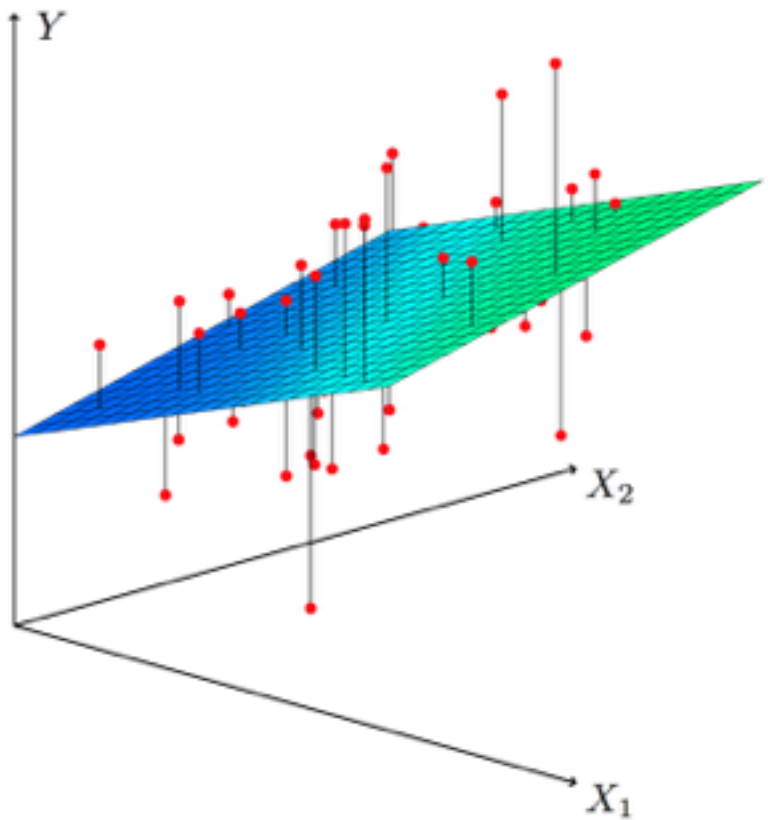
Alternative hypothesis: $H_A : \beta_1 \neq 0$

DATA SCIENCE PART TIME COURSE

MULTIPLE LINEAR REGRESSION

- 1 degree
- multi-dimensions
- allows for complex models even with linear components

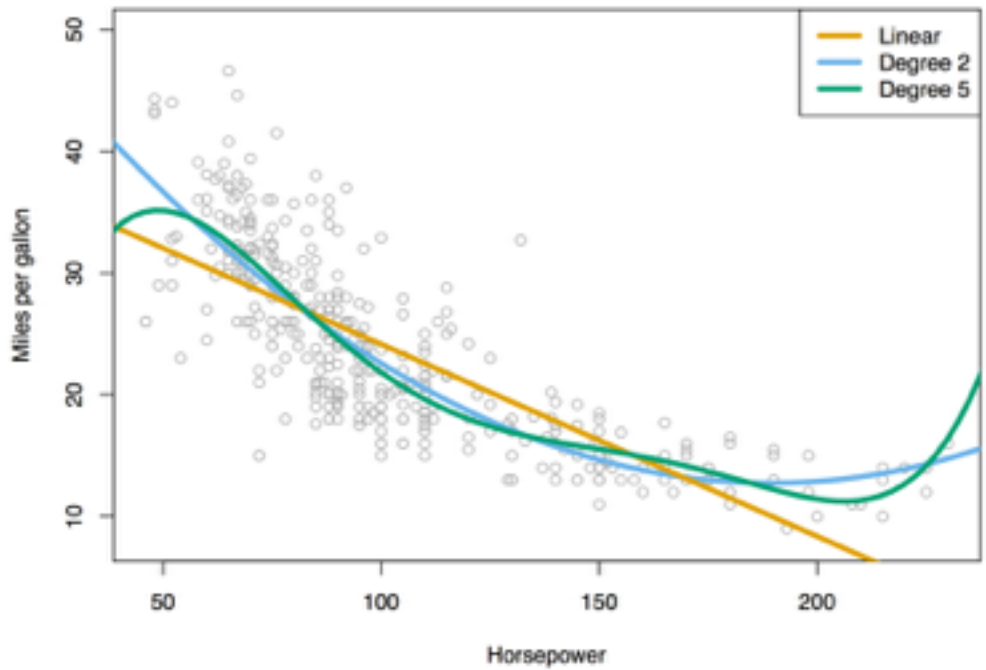
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$


Y
X2
X1

- The ideal scenario is when the predictors are uncorrelated:
  - Interpretations can be made such as "a unit change in Xj is associated with a βj change in Y , while all the other variables stay fixed"
- Correlations amongst predictors cause problems
  - when Xj changes, everything else changes

DATA SCIENCE PART TIME COURSE

NON-LINEAR EFFECTS USING LINEAR REGRESSION



$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

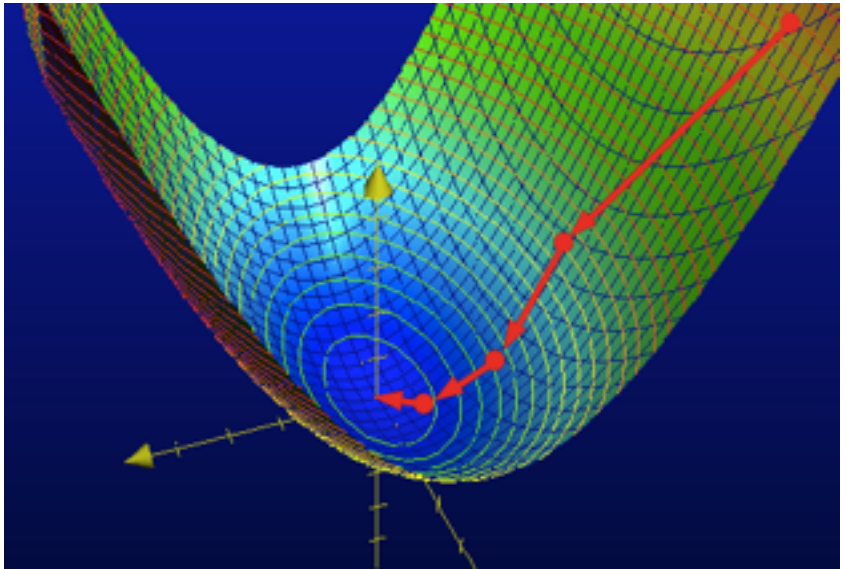
DATA SCIENCE PART TIME COURSE

ALTERNATIVE ESTIMATION PROCEDURES

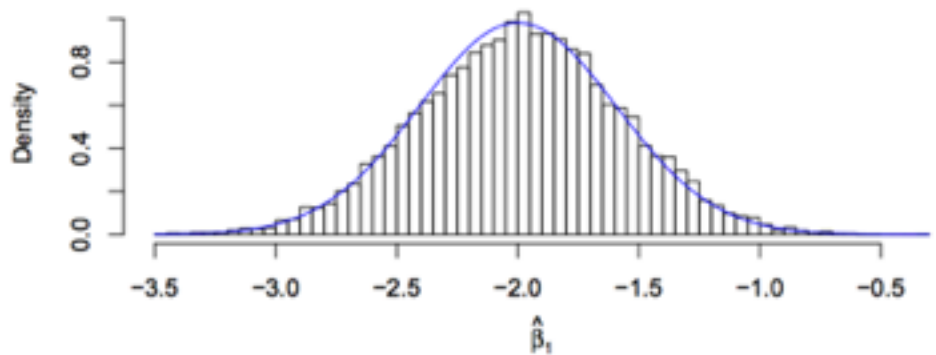
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$(X^T X)^{-1}$ **NOTE: this is an n x n matrix, which has O(n^3) time complexity**

- Stochastic Gradient Descent
  - alternative to Ordinary Least Squares when data size grows



- Maximum Likelihood Estimation (MLE)
  - find the weight vector that is 'most likely', given the data we have



- Bayesian Linear Regression
  - provides the notion of ‘uncertainty’
  - probability distribution of all possible weight vectors
  - we supply the assumption of a ‘prior’ probability distribution

LAB

git fetch upstream

git checkout master

git merge upstream/master

WEEK 2 – Wednesday

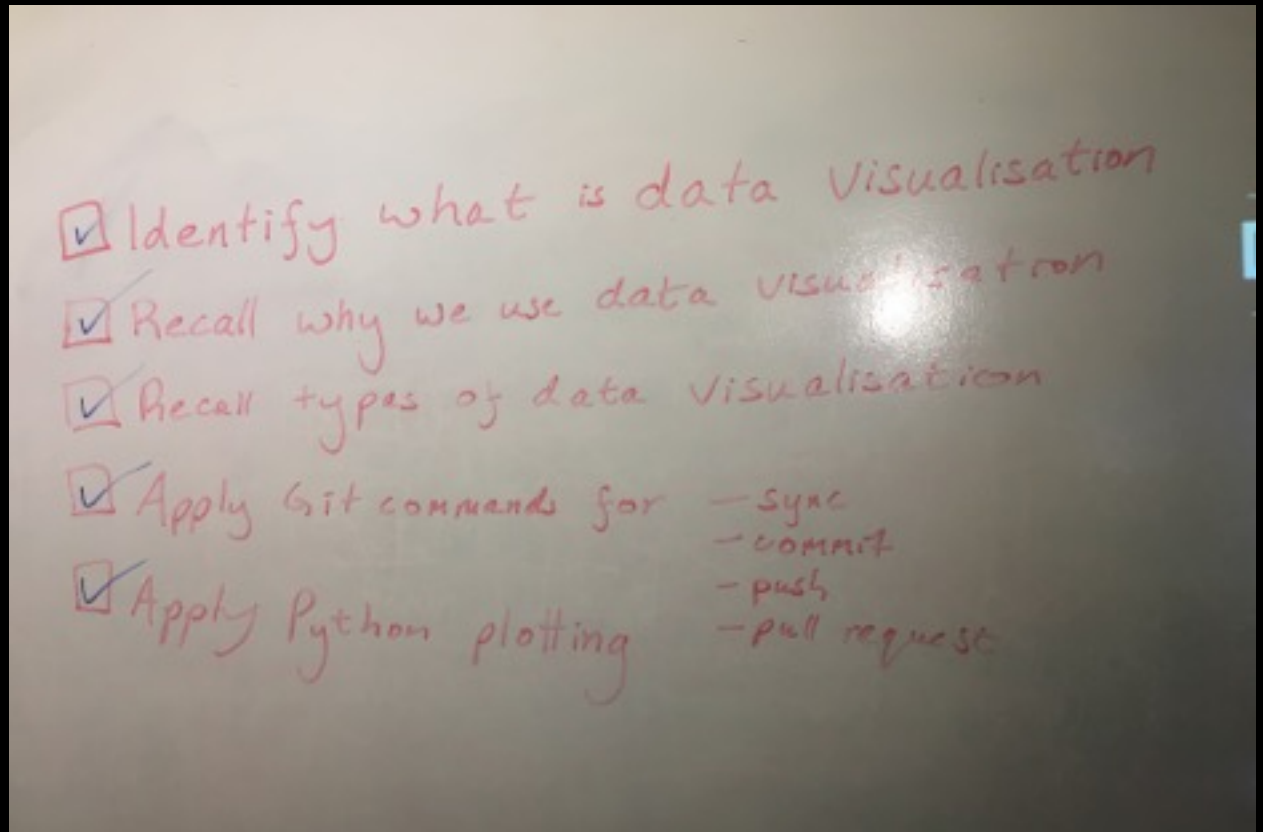
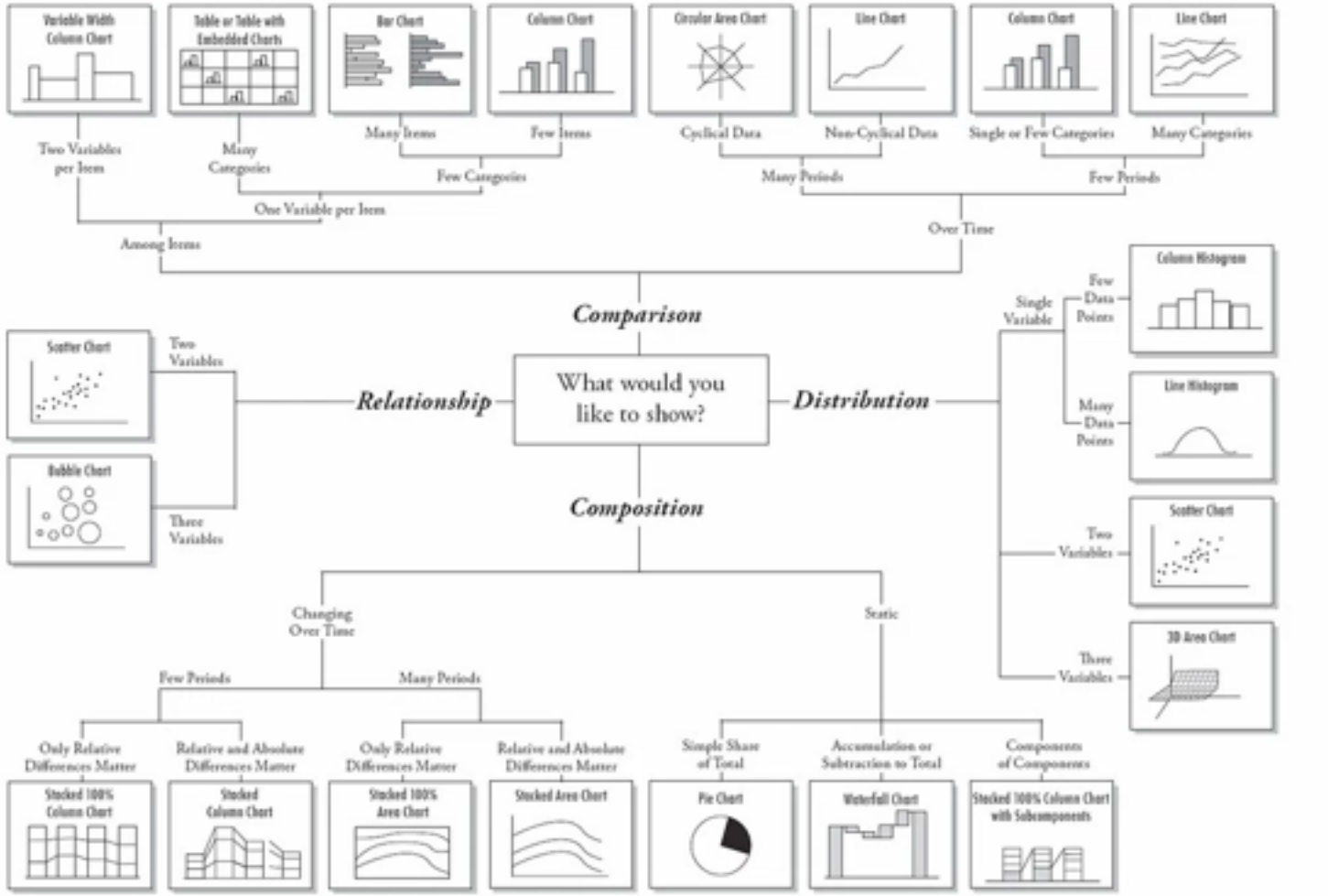


Chart Suggestions—A Thought-Starter



DISCUSSION TIME

Homework/Course Project

- Homework1.ipynd – due Friday
- Read Chapter 4 of Introduction to Statistical Learning – Classification
- Course Project: Prepare 3 concepts for a project