GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 8

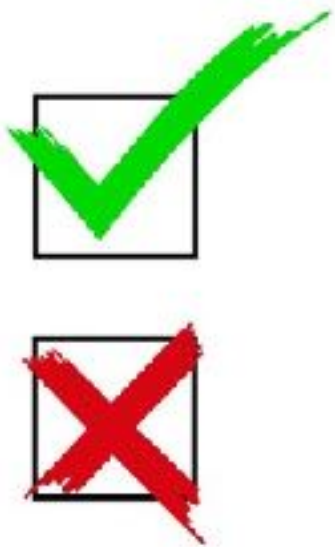
Week 2 – Linear Regression
Thursday 1st June

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

DATA SCIENCE PART TIME COURSE

SUPERVISED & UNSUPERVISED LEARNING

Learning by example.



What do we want to model?

Often we want to predict something.

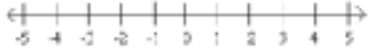
- Who will win a match
- What a customer wants
- The value of a stock
- The time until an event

The target may also be called:

- response variable
- dependant variable
- label

Regression:

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



Classification:

If the target 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.

The data values that provide information to help guess the target

The features may also be called:

- predictor variables
- independent variable

The data values that provide information to help guess the target

Today's Air Pressure

Today's Temperature

Today's Humidity

Today's Cloud Cover

~ >

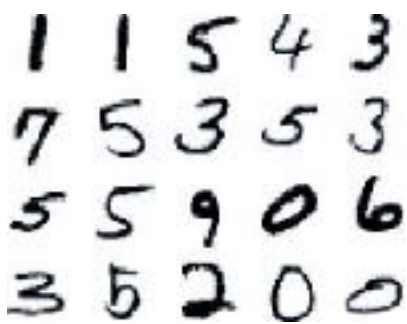
◎ Tomorrow's Temperature

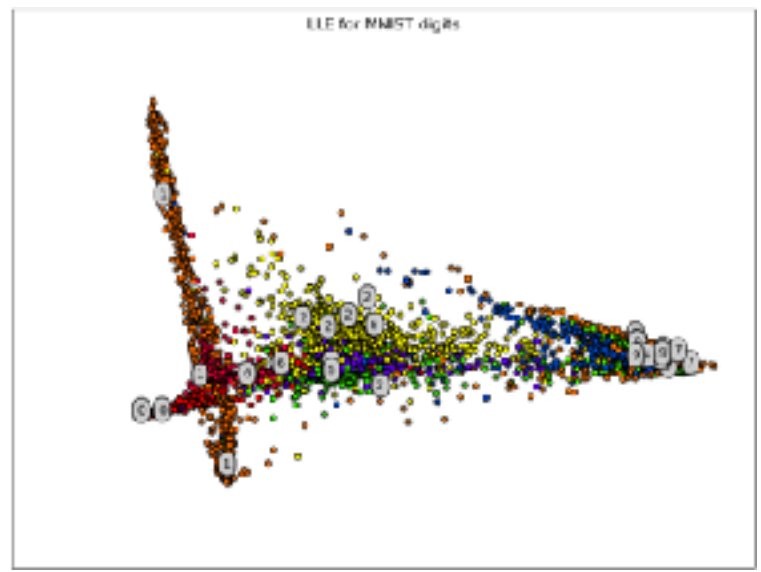
The features may also be called:

- predictor variables
- independent variable

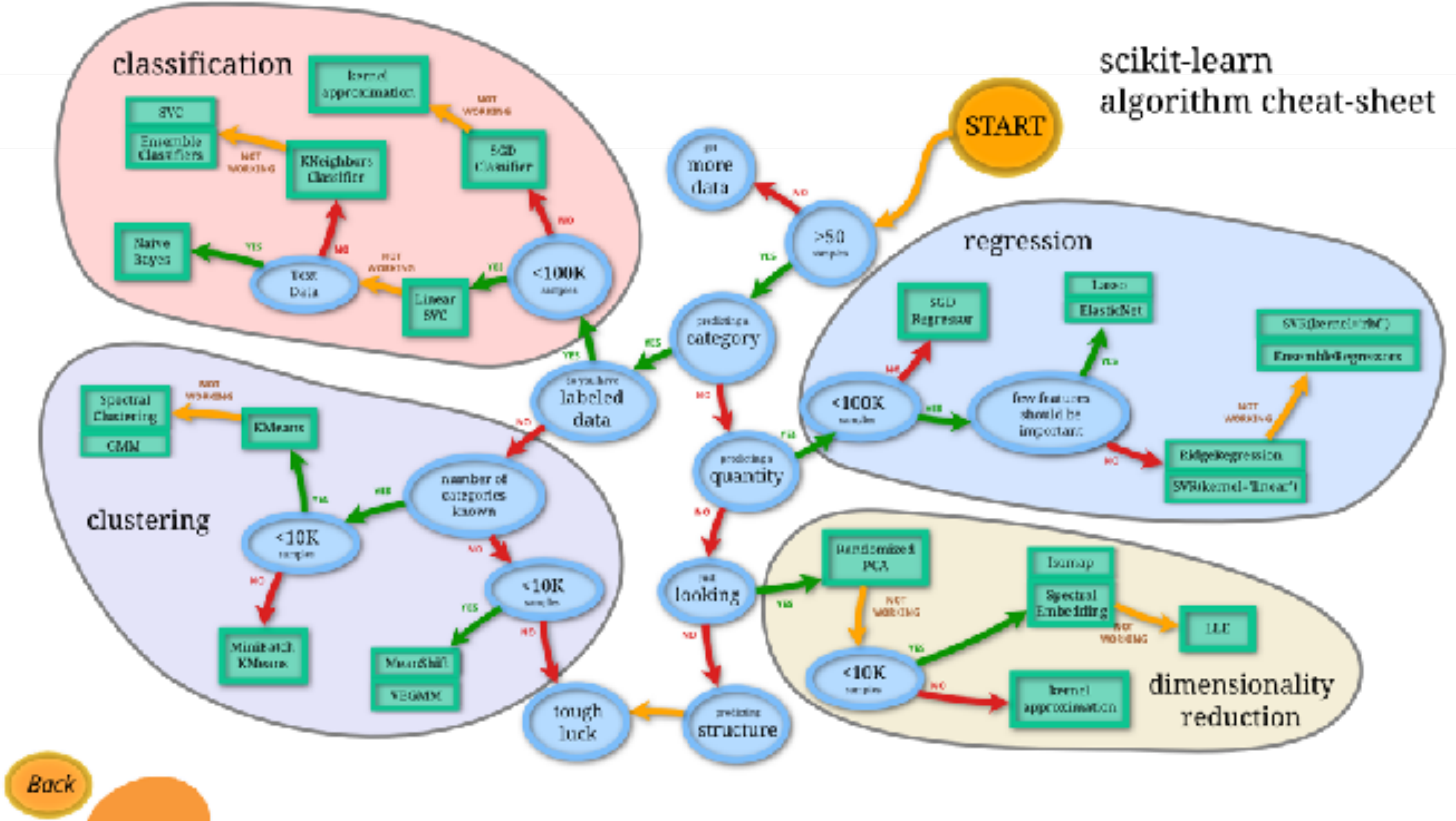
Find structure in data, such as clusters.

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





http://projector.tensorflow.org/



scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality
reduction
kernel
approximation
SVC
Ensemble
Classifiers
KNeighbors
Classifier
SGD
Classifier
Naive
Bayes
Text
Data
Linear
SVC
<100K
samples
get
more
data
>50
samples
category
labeled
data
quantity
looking
structure
tough
luck
SGD
Regressor
Lasso
ElasticNet
<100K
few features
should be
important
RidgeRegression
Spectral
Clustering
GMM
KMeans
number of
categories
known
<10K
samples
MiniBatch
KMeans
MeanShift
VBGMM
Randomized
PCA
Isomap
Spectral
Embedding
LLE
<10K
samples
kernel
approximation
YES
NO
NOT
WORKING
Back

WHAT IS LINEAR REGRESSION?

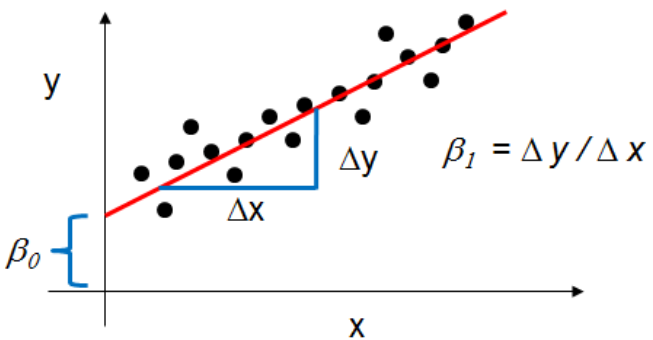
We want to model a linear relationship (think straight line) between our target variable y and our input variable x.

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

- 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**.



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

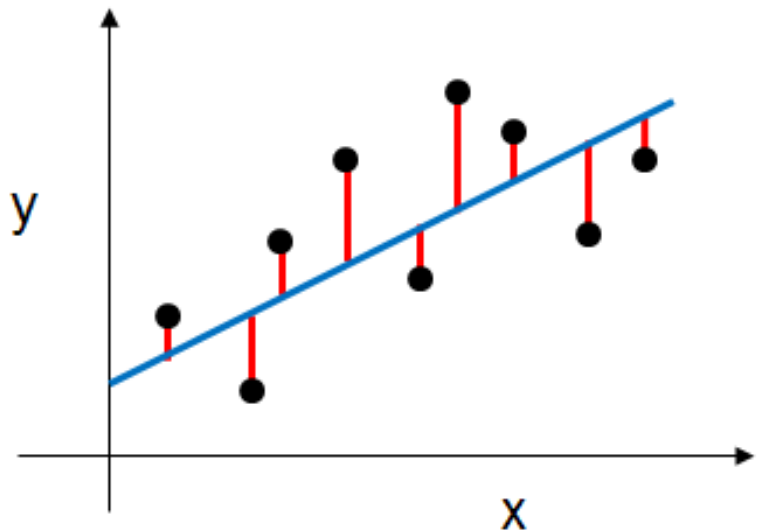
- y = target variable
- X = input variable
- β = coefficients
- ϵ = error term

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

We want to predict the price of a house, 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 goal is a function $y = f(X)$, to describe the house price based on observed data.

- y : house price $
- X's : the area (x1), number of bedrooms (x2), size of the house (x3), and if it has a pool or not (x4)



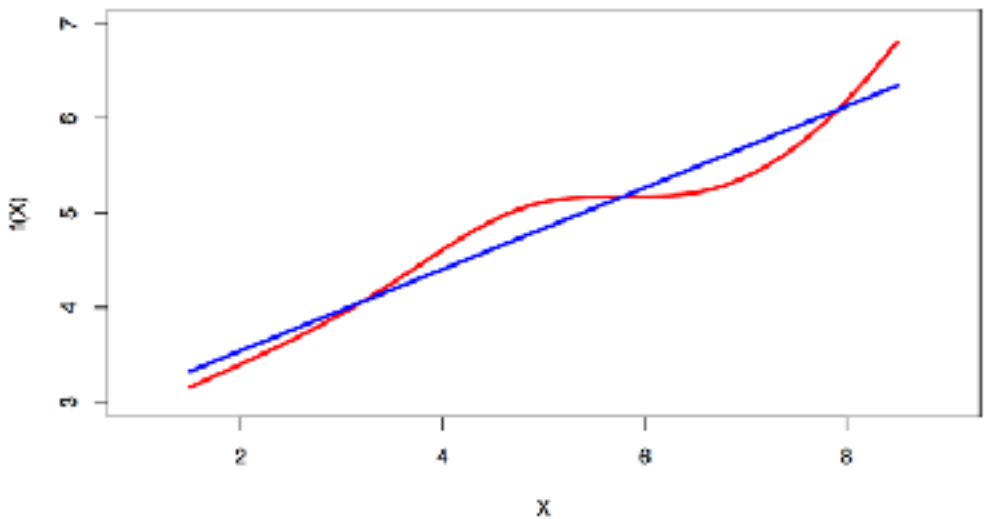
Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

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. If it does not, it could introduce bias.
- No perfect (or near perfect) co-linearity between any of the input variables. Otherwise the fitting procedure will break.

- R-squared, the central metric introduced for linear regression
- Which model performed better, one with an r-squared of 0.79 or 0.81?
- R-squared measures explain variance.
- But does it tell the magnitude or scale of error?
- We'll explore loss functions and find ways to refine our model.

DATA SCIENCE PART TIME COURSE

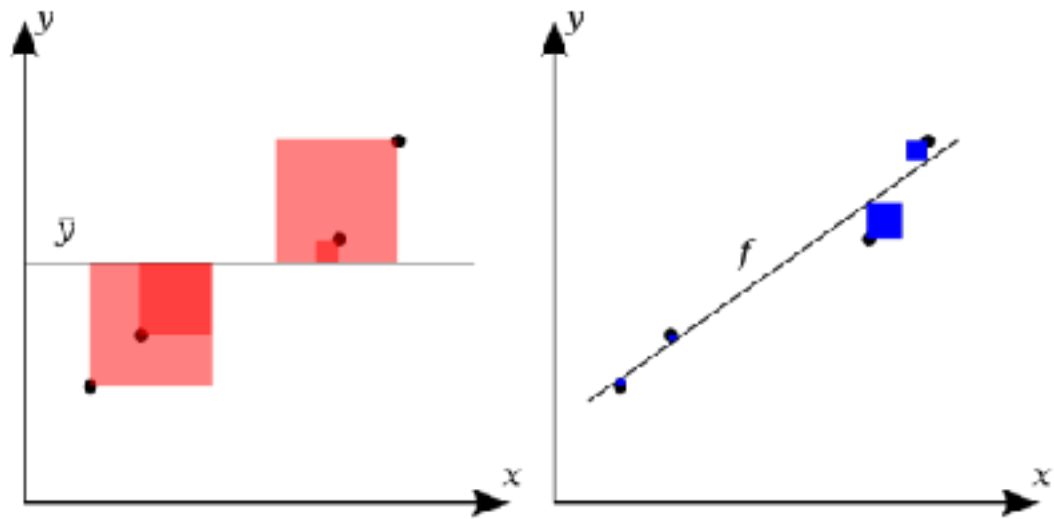
EVALUATING QUALTIY OF FIT?

R-Squared

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$



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

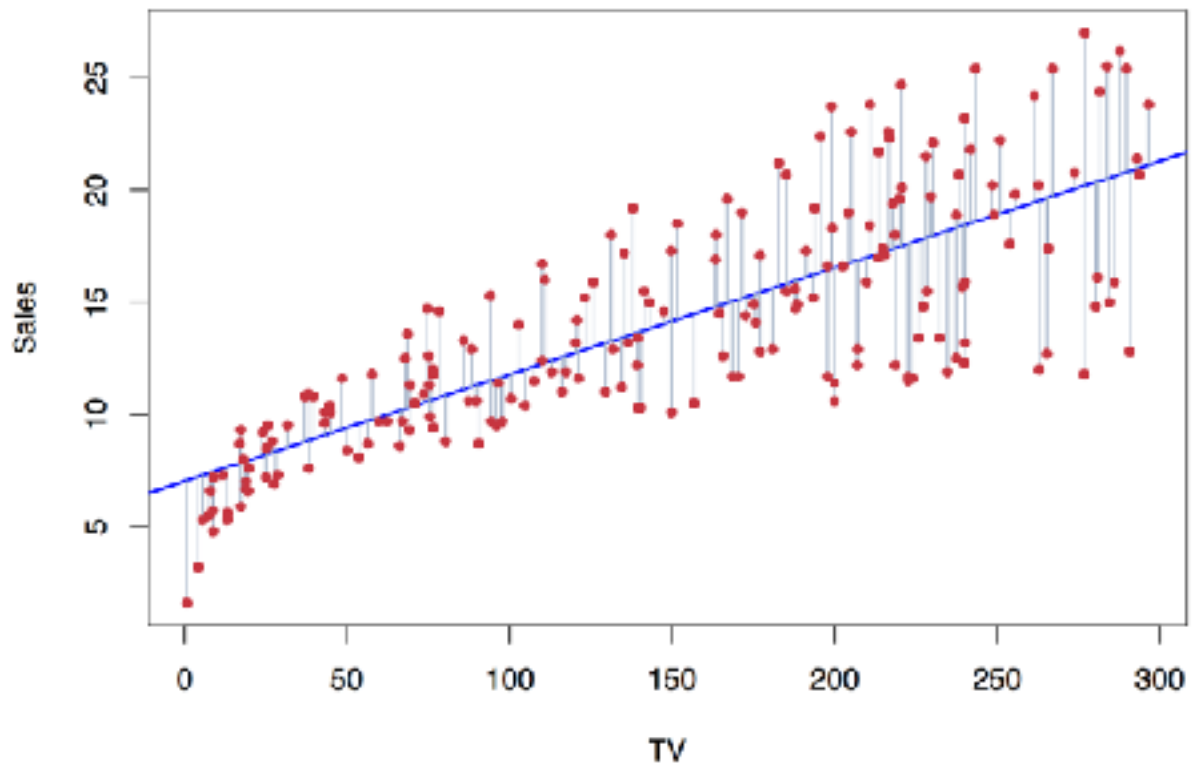
1 $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

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

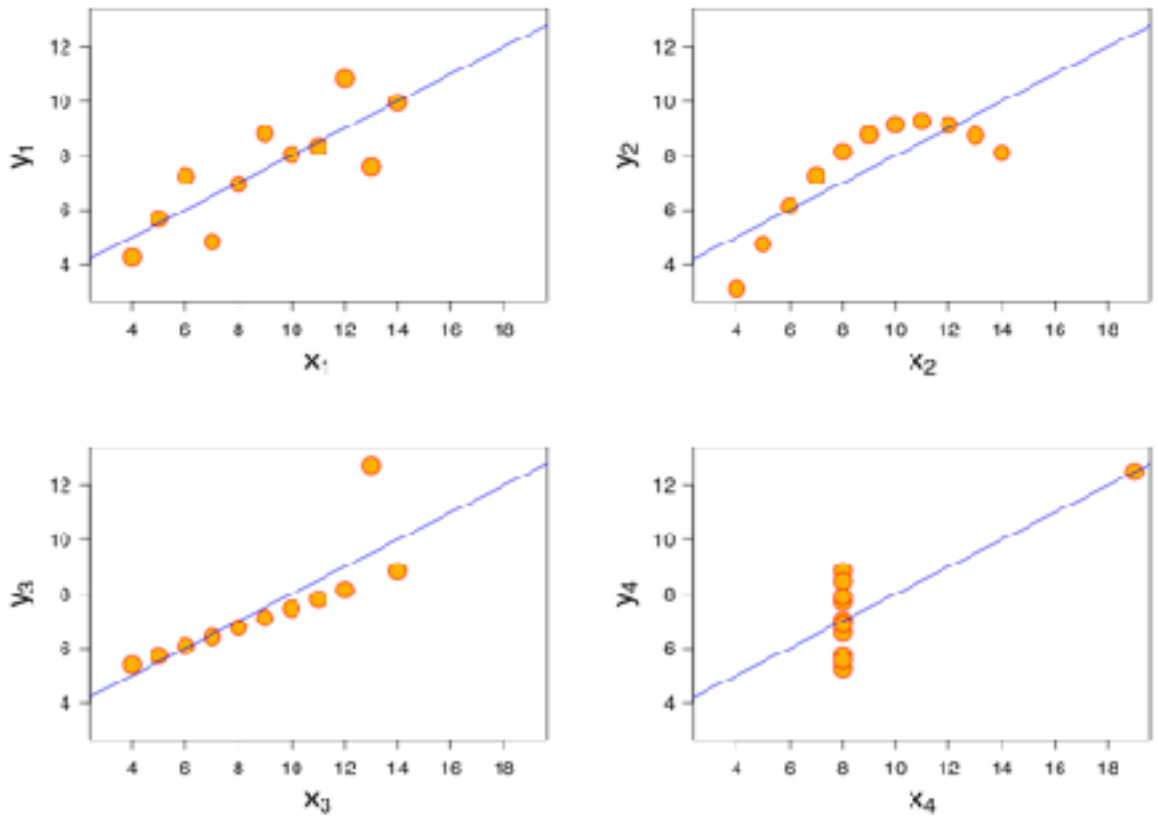

Sales
5
10
15
20
25
0
50
100
150
200
250
300
TV

- 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

LINEAR REGRESSION – IMPORTANT STEP

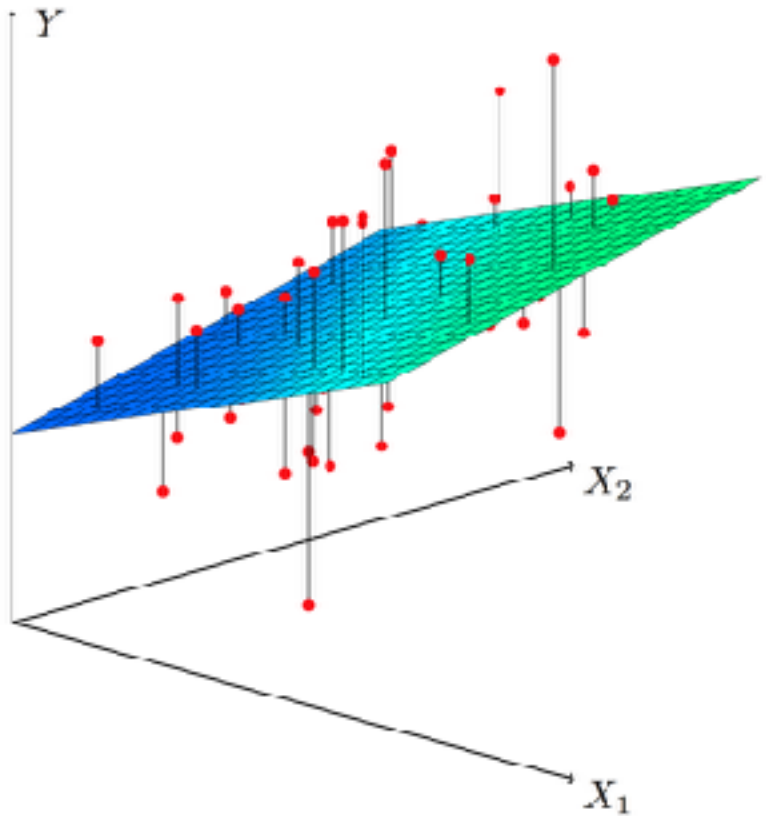


DATA SCIENCE PART TIME COURSE

MULTIPLE LINEAR REGRESSION

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

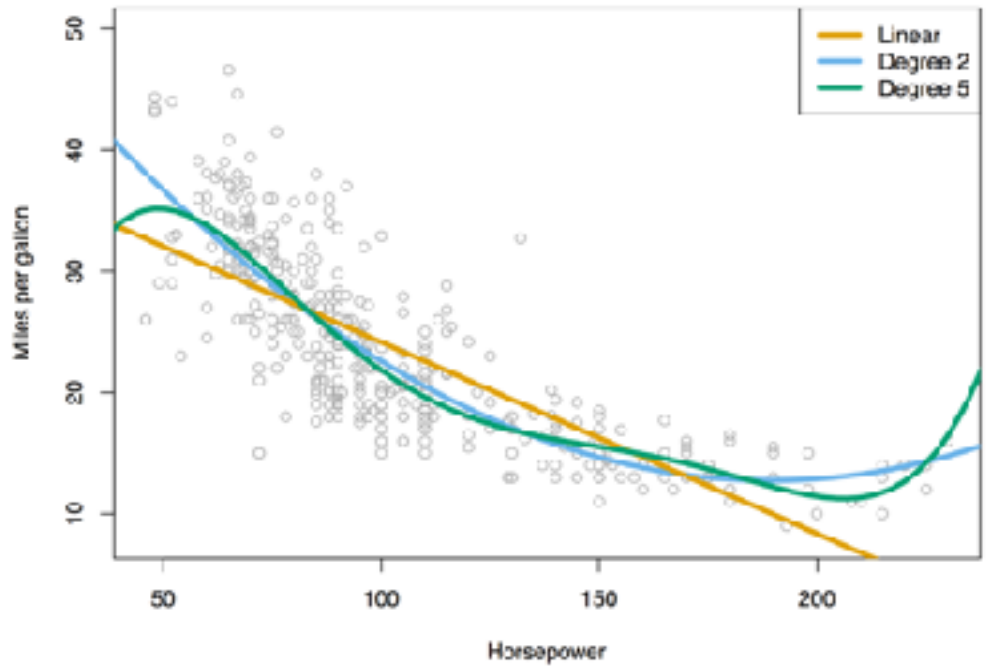
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$


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



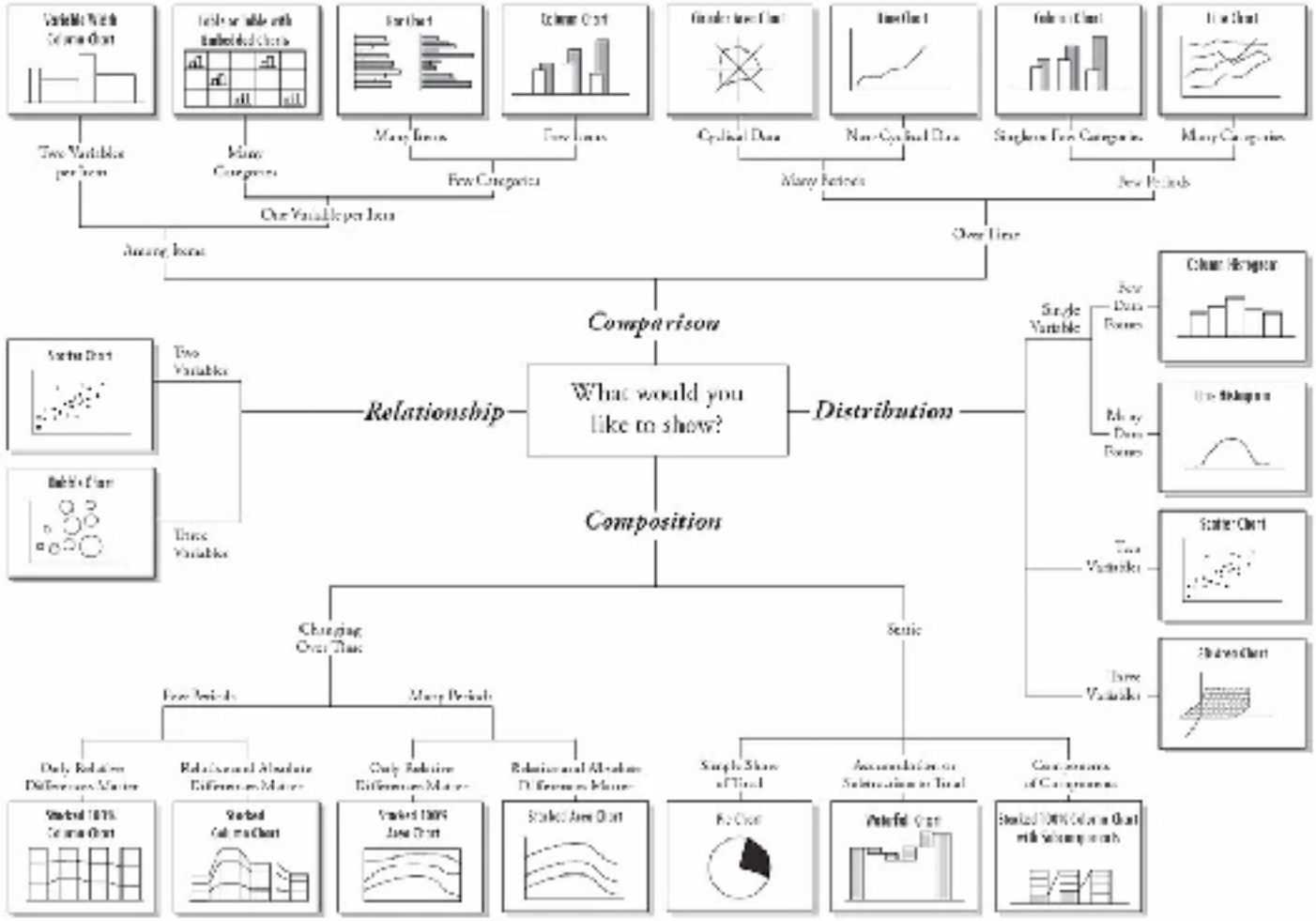
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

LAB

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m "descriptive label for the commit"
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

Chart Suggestions—A Thought-Starter



DISCUSSION TIME

Homework

- Homework1.ipynd – due Friday
- Read Chapter 4 of Introduction to Statistical Learning – Classification