DataSoc

An Introduction to Machine Learning

UNSW Data Science Society

Agenda

1. Cleaning of Data
2. Encoding of Data
3. Plotting and EDA analysis
4. Data Transformations
5. Model fitting
6. Model selection (HPO)
7. Model understanding/explainability

A very simple analogy

Rule: + 3 meaning $f(x) = x + 3$

2, 5, 8, 11 ... 14

Creating a prediction of the future value!

We are using the a set of numbers to help us decide what the rule should be

Machine Learning is the act of using statistical algorithms to find a function that best describes the data.

$$Y = f(X) + \varepsilon$$

Irreducible Error

What is Machine Learning?

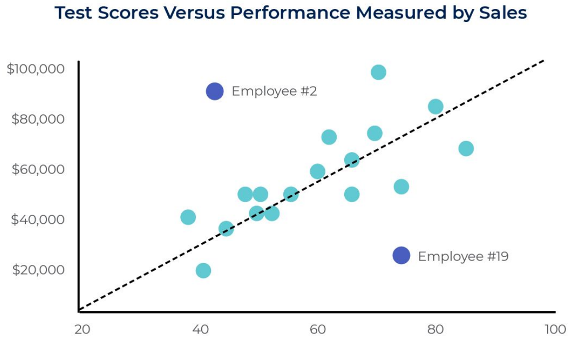
Algorithms that **analyse** data to **identify** patterns, make accurate **predictions** or intelligent decisions.

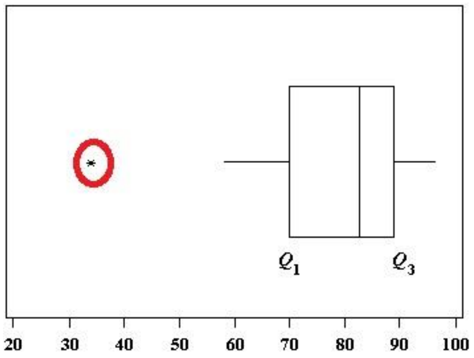
1. **Objective:**
    - Make accurate predictions or inferences from data patterns
2. **Types**
    - Supervised: Training dataset and test dataset required for future predictions
        - Regression: Predict a continuous number (e.g. Amount of sales for next year)
        - Classification: Predict whether True or False (e.g. Buy or Not buy)
    - Unsupervised: Just a dataset is required
        - Clustering (e.g. Market Segmentation)
3. **Applications**
    - Literally where there is data!!

Data Cleaning

Definition: the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in datasets to improve data quality and reliability.

1. **How to get a good dataset**
    - Kaggle walkthrough
2. **Missing Values**
    - Removal or imputation
        - 0 imputing
        - Mean imputing
3. **Outliers**
    - Extreme values, consider removal depending on impact
        - 1Q - 1.5IQR, 3Q + 1.5IQR





Encoding

Machines only recognise numbers, so we must change categorical data (text) to numerical data using encoding!

1. One-Hot Encoding:

- Nominal - categories **without** ordering/ranking
- Binary, 1 indicates the presence, 0 indicates absence

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding →

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Encoding

2. Label Encoding:

- Used on ordinal data: categories with ranking!
  - Data with order: e.g. small, medium, large
- Unique numerical codes to each category **with** ordering/ranking

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

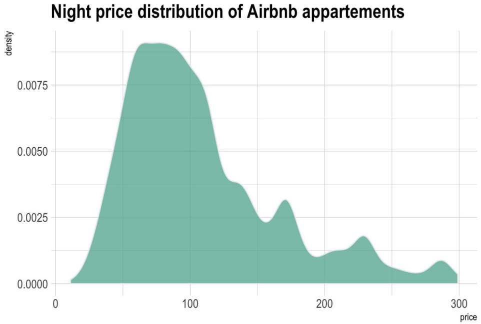
Height
Tall
Medium
Short

Height
0
1
2

Exploratory Data Analysis

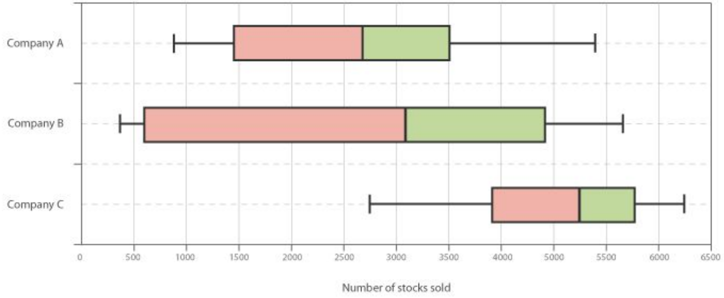
Visualising and exploring the data to uncover patterns, identify outliers, and gain insights before performing formal statistical modeling.

Plots to consider:
- Correlation Matrix
- Histograms
- Violin Plots/Box Plots
- Pairplots



	Connectivity	Digital Public Services	Human Capital	Integration of Digital Technology	Use of Internet
Connectivity	1.00	0.64	0.71	0.65	0.77
Digital Public Services	0.64	1.00	0.58	0.64	0.62
Human Capital	0.71	0.58	1.00	0.66	0.72
Integration of Digital Technology	0.65	0.64	0.66	1.00	0.60
Use of Internet	0.77	0.62	0.72	0.60	1.00

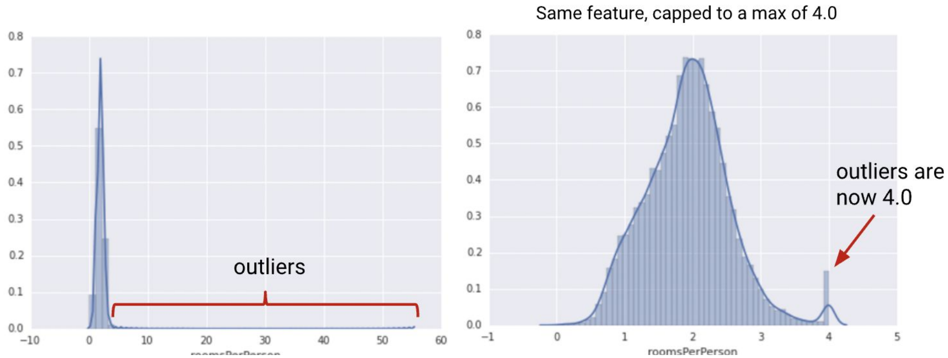
Correlation Matrix



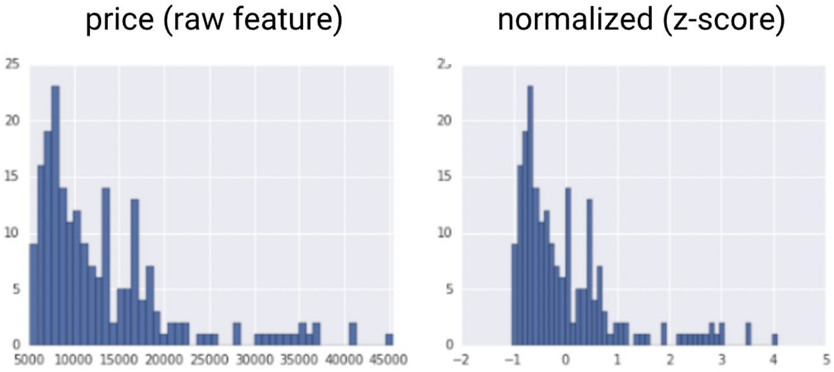
Data Transformations

Definition: modifying the original data to improve its distribution, scale, or other characteristics for better analysis or modeling

- Normalisation: transform features to be on the same scale
    - Feature Clipping
    - Standard Scaling



Feature Clipping

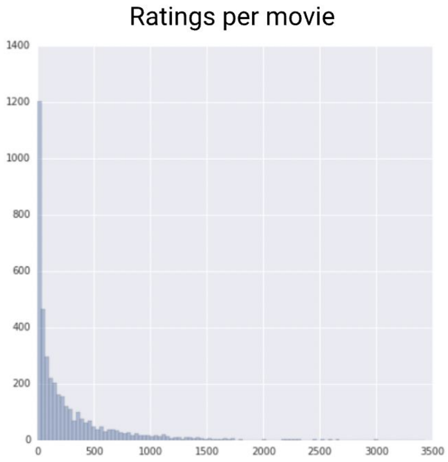


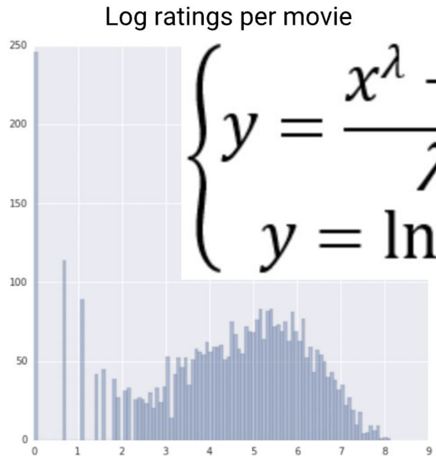
Z-score transformation

$$z = \frac{x - \mu}{\sigma}$$

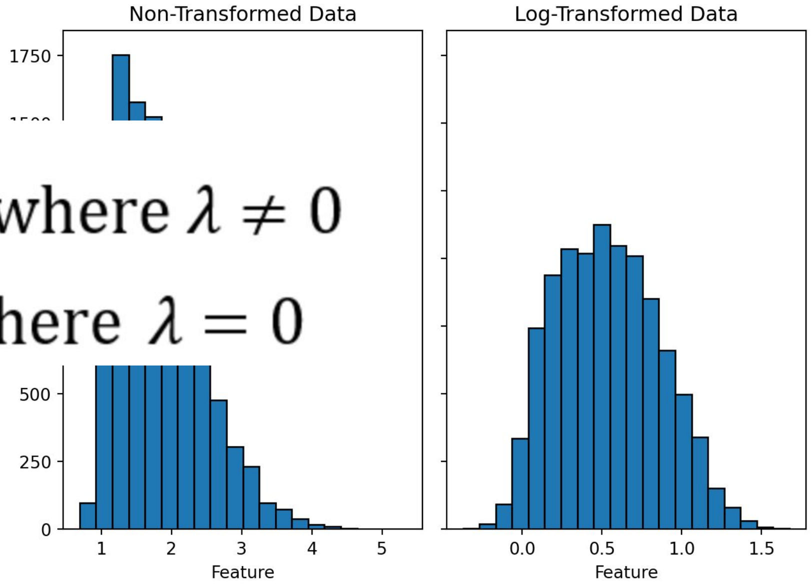
Data Transformations

- Box-cox: either features or response will be transformed to be normally distributed
  - Reasons: meet assumptions, stabilise variance
  - Special case (lambda = 0) : log transformations





$$\begin{cases} y = \dfrac{x^\lambda - 1}{\lambda} \text{ where } \lambda \neq 0 \\ y = \ln x \text{ where } \lambda = 0 \end{cases}$$



Normalisation

Removing skewness in distribution

Model: Logistic Regression

Business Understanding

Suppose:

- KFC is a fast food company that sells amazing deep fried chicken.
- They come to you (a data scientist) to understand and predict whether their customers will leave or return.

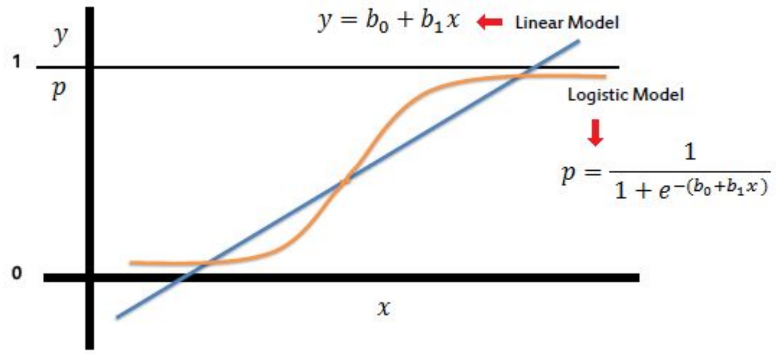
- Supervised Problem -> Classification Question:
  - Leave = 1, Return = 0

$$p_i = \frac{1}{1 + \exp -(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$

Sigmoid Function

We can then impose a threshold:

- **p_i >= 0.5** (classify as true); **p_i < 0.5** (classify as false)



Advantages:

- Simple and interpretable

Disadvantages:

- Sensitive to outliers
- Requires normalisation of data
- Linearity assumption

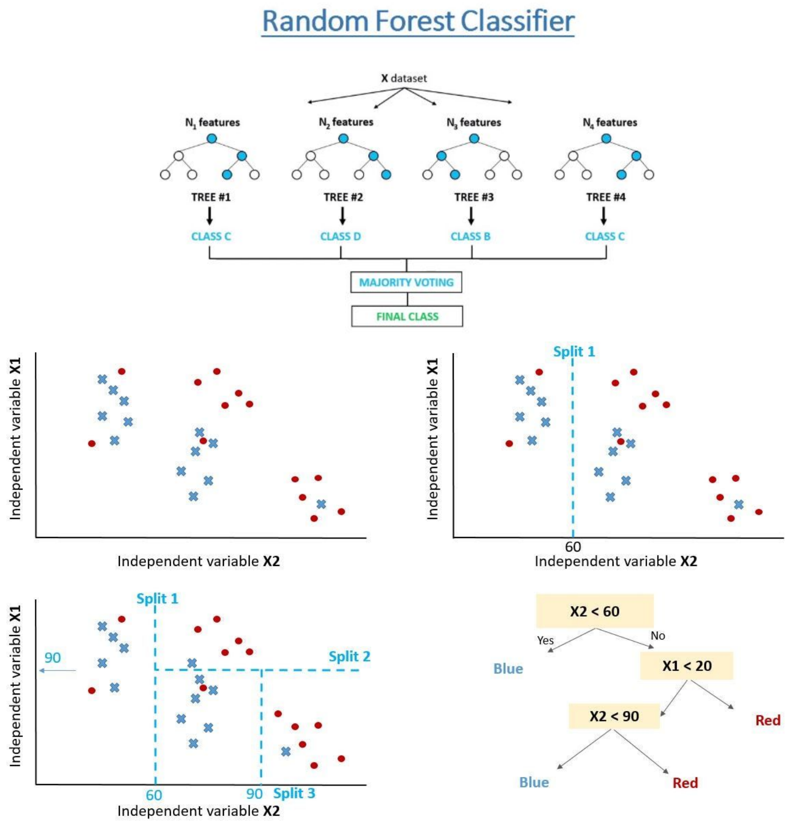
Model: Random Forest

How does this work?

- *Ensemble model* - averages multiple models
- **Key:** wisdom in crowds, any error is evened out
    - Variance reduction, uncorrelated
- Pros and cons:
    - More accurate, hard to explain/interpret

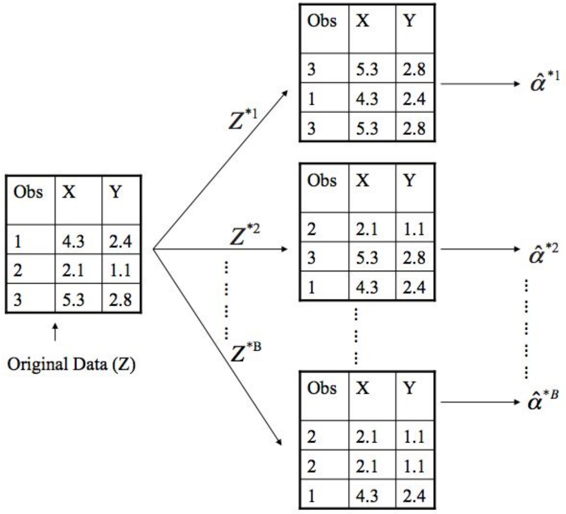
Individual Models

- *Decision trees*
    - Partition data into two regions
      - split variable wise
    - Continue splitting until stopping criteria is reached
- Issues
    - Trees are prone to overfitting or bias
    - Aggregating many models addresses this



Aggregating models - decorrelating trees

- *Bootstrapping* - sampling data with replacement
    - Individual trees trained on resampled set
    - Introduces uncorrelated data for trees
- Random feature selection
    - Increases training speed
    - Decorrelates trees
- Height constraints - limit amount of splits
    - Overfitting, computation, less noise



Types of outputs

- Classifier - probability %
- Regressor - average of outputs

Implementation:

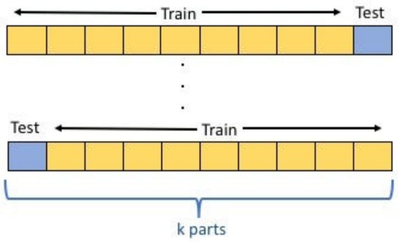
- Brieman Random Forest in *sklearn*

Metrics

Assesses **model quality**, different **aspects** of model prediction

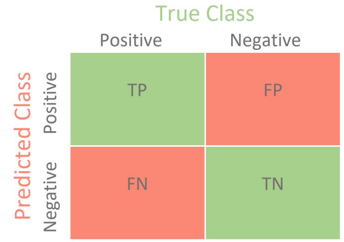
Cross validation

- Test on different portions of data
- K-fold, LOOCV - test on one portion: assess accuracy
- Predictive power on unseen data



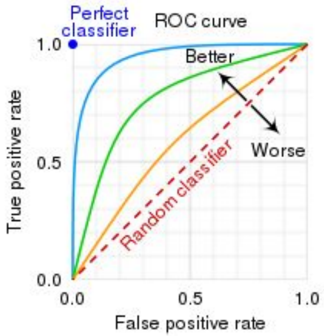
Classification report

- Precision: when predicted positive, likely to be correct - accurate
- Recall: good at finding the positive instances, even at cost of false positives - sensitive



ROC curve

- TP = true positive, predicted correct positive / actual positive
- FP = false positive, predicted false positive / actual negative
- **Higher AUC (area)**, more overall TP



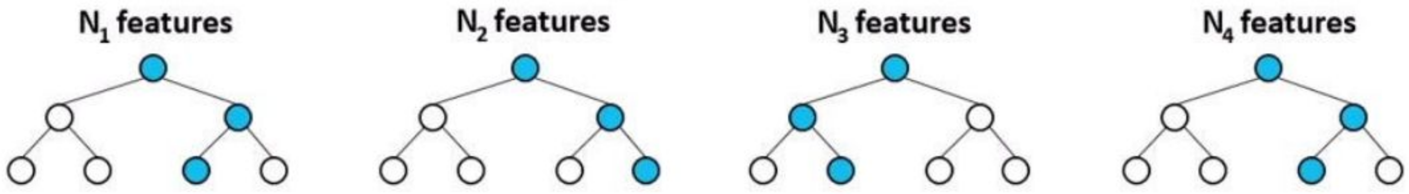
Hyper Parameter Optimisation

What are Hyperparameters?

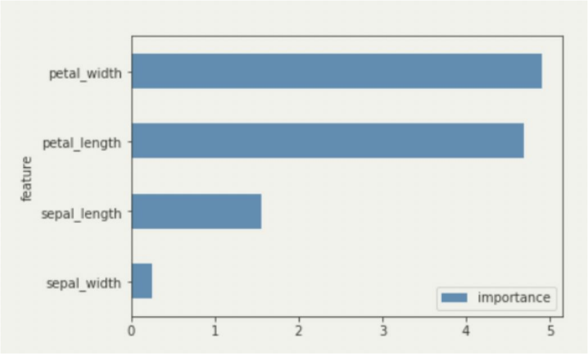
- Parameters defining **behavior/structure** of a model
  - E.g. Random Forest: number of models, height
  - Set prior to training - affects model learning
- Key: improve model performance and efficiency
  - Controls overfitting or underfitting
  - Better generalisation on new datasets

Techniques

- Grid search, random search, Bayesian optimisation
- Implementation - *GridSearchCV, RandomizedSearchCV*



Feature Importance



Main idea

- The degree of **individual variable contribution** to the **performance** of a machine learning model
- Metrics, degree of coefficients, etc.

Why?

- Model: quality, prediction
  - Interpretability, communication, generalisation
- Informed **business decision making**
  - Simplicity, efficiency in real-world process, Allocation of resources
  - E.g. Potability - concentrate on contaminants

Hierachy/ranking of variable importance?

- Metrics: SHAP values, Decision Tree Feature importance

Ex. Lop(p / 1 - p) = 0.5 + 10petal_width + 7petal_length + 2.3sepal_length + 1.1sepal_width

DataSoc

Q&A

UNSW Data Science Society

DataSoc

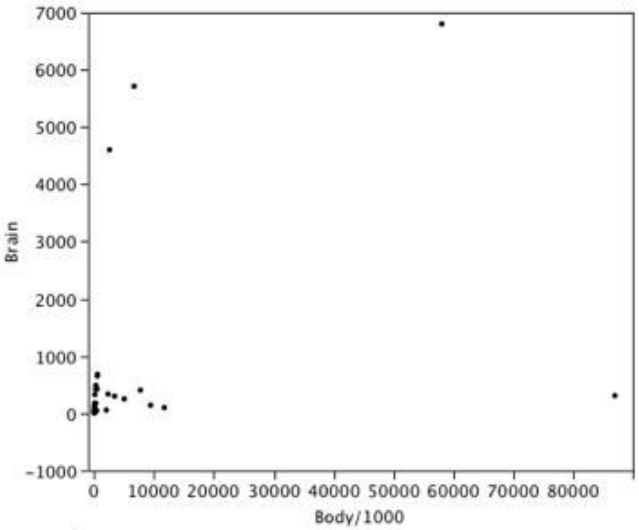
Photo Time!!

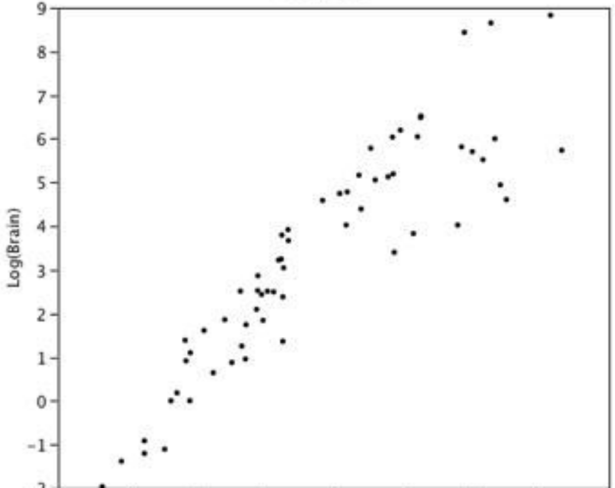
Agenda

1. Cleaning of Data
2. Encoding of Data
3. Plotting and EDA analysis
4. Data Transformations
5. Model fitting
6. Model selection (HPO)
7. Model understanding/explainability

DataSoc

ARCHIVE



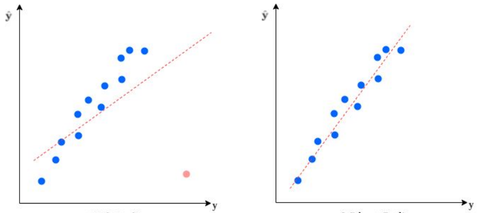


UNSW Data Science Society

Exploratory Data Analysis

Visualising and exploring the data to uncover patterns, identify outliers, and gain insights before performing formal statistical modeling.

- Purpose: Gain familiarity, data quality issues, hypotheses
- Techniques: Descriptive statistics, visualisations (plots), exploration (correlations), outlier detection
- Benefits: Data quality, insights, informed decision making
- Outliers: data point, with extreme **response** value
  - Impact: skews models
  - Treatment: deletion, imputation, transformation, separate treatment
  - Side note: high leverage point, extreme predictor values



With Outlier Without Outlier

How to Fit Models?

Data point: ***X*** = dependent variables, ***y*** = response

- Fundamental idea: assume observed equals mean plus error
- Aim: estimate this mean (coefficients, equation form)

$$y_i = f\left(X_{i,1}, \ldots X_{i,n}\right) + e_i$$
$$\widehat{y_i} = \widehat{b_0} + \widehat{b_1} \cdot X_{i,1} + \cdots + \widehat{b_n} \cdot X_{i,n}$$

<u>Ordinary Least Squares</u>

- Key: minimise total square error between predictions and observations
- Differentiate with respect to parameters to minimise

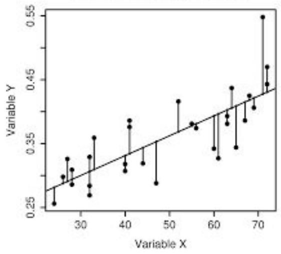
$$\sum_{all\ i} e_i^2 = \sum_{all\ i} \left(y_i - \widehat{y_i}\right)^2$$

<u>Maximum Likelihood Estimation</u>

- Key: maximise probability of observations with respect to distribution
- Specify probability distribution with parameters
- Find probability of observations, differentiate with respect to parameters, solve to maximise probability
- Benefits: minimum variance unbiased estimates

$$L(\theta) =$$

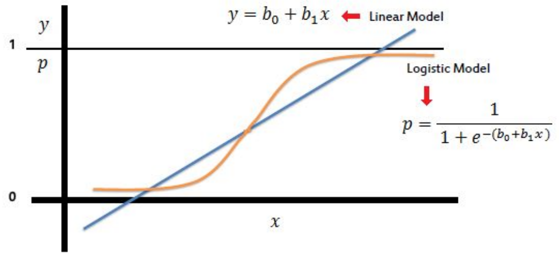
Figure 2
OLS regression model residuals



Model: Logistic Regression

How does this work?

1. Calculate the '*response*': $\hat{y}_i = \widehat{b_0} + \widehat{b_1} \cdot X_{i,1} + \cdots + \widehat{b_n} \cdot X_{i,n}$
2. Convert to a *probability*: $p_i = 1/[1 + exp\,(-\hat{y}_i)]$
3. Assign 0 or 1 (based on *threshold*)



Why?

- Variables linearly combined to create *response* (-∞, ∞)
- *Logit function* - transforms linear response to a probability between [0, 1]
- **Key:** we want a linear combination of predictors, but must transform into a probability

Normalisation:

- *Regularisation* - constraint imposed on coefficients to avoid overfitting
  - *Bias variance tradeoff* - how well model generalises
- *Normalised* - same units, variables fairly constrained

Challenges:

- *Multicollinearity*: increases variance of coefficient estimates, important variables not discerned