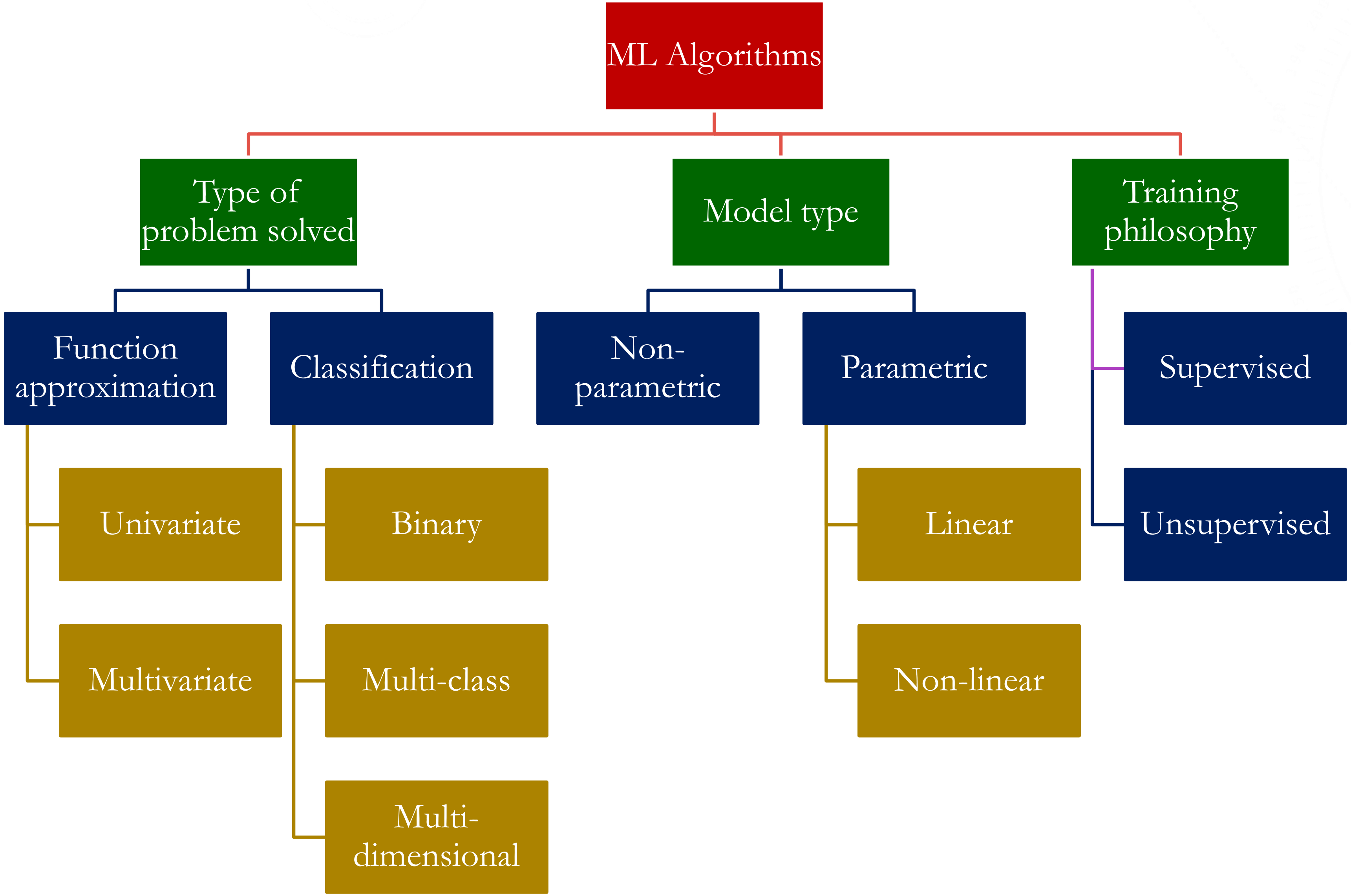
LINEAR REGRESSION

RESMI SURESH

ASSISTANT PROFESSOR, IIT GUWAHATI



TYPOLOGY OF ML ALGORITHMS

FUNCTION APPROXIMATION OR REGRESSION

Examples:

Predicting scores in a game of cricket

Predicting material properties for different chemicals

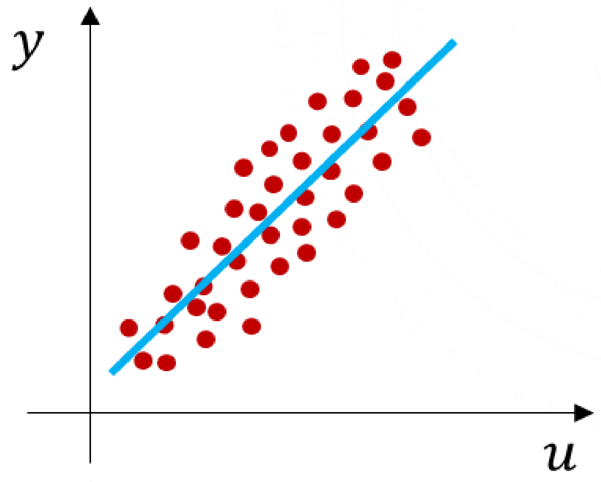
Predicting mechanical properties of a part

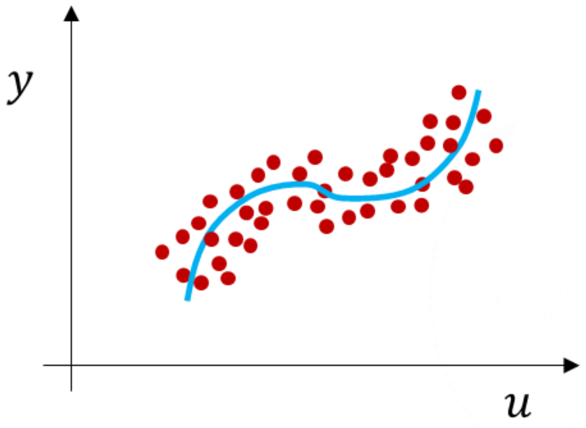
Predicting battery temperature in an electric vehicle

Predicting value of a board position in chess

Techniques:

Linear regression, k-nearest neighbors, Neural network, Decision tree, Random forest, Principal component analysis, …





$$y = f(x_1, \dots x_n, p_1, \dots, p_m)$$

CLASSIFICATION

Examples:

Fraud detection in credit card transactions

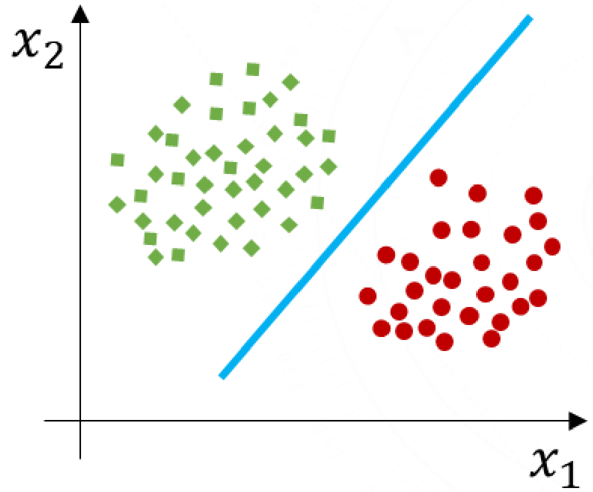
Distinguishing objects – “Self-driving cars”

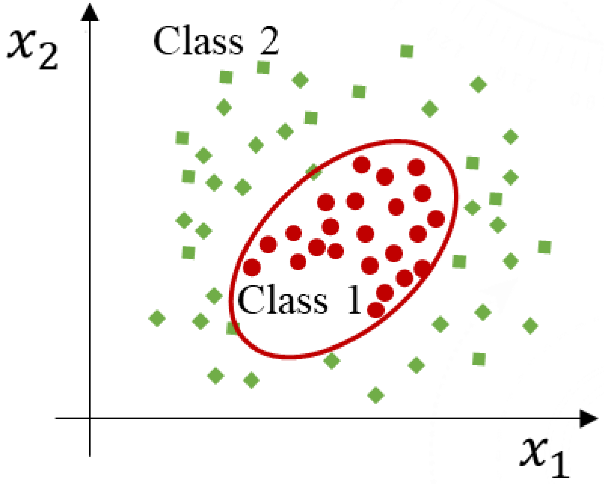
Detecting failures in built systems/equipment

Classifying emails as spam or genuine

Techniques:

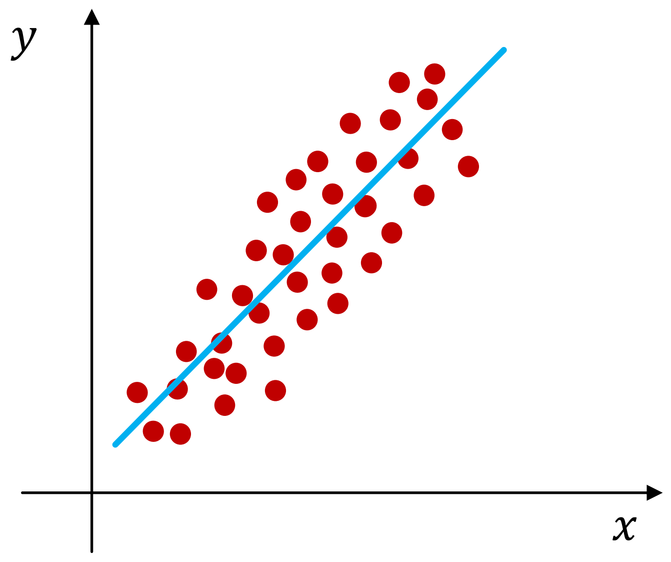
Logistic regression, k-nearest neighbors, Neural network, Decision tree, Random forest, Support vector machines, LDA, QDA, Naïve Bayes, Hierarchical clustering, k-means clustering, …

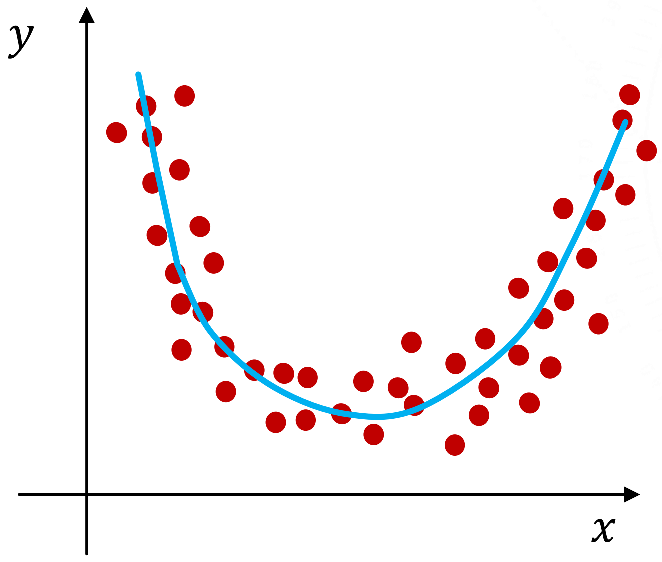




Class 1 - $h(x_1, \dots x_n, p_1, \dots, p_m) \geq 0$

Class 2 - $h(x_1, \dots x_n, p_1, \dots, p_m) < 0$





FUNCTION APPROXIMATION OR REGRESSION

$$y = f(x)$$

REGRESSION - BASICS

- Dependent variables also known as Response variable, Regressand, Predicted variable, output variable - denoted as variable/s y
- Independent variable also known as Predictor variable, Regressor, Exploratory variable, input variable - denoted as variable/s x
- Classification of regression
  - Univariate vs Multivariate
    - *Univariate*: One dependent and one independent variable
    - *Multivariate*: Multiple independent and multiple dependent variables
  - Linear vs Nonlinear
    - *Linear*: Relationship is linear between dependent and independent variables
    - *Nonlinear:* Relationship is nonlinear between dependent and independent variables

REGRESSION - BASICS

- Is there a relationship between these variables?
- Is the relationship linear and how strong is the relationship?
- How accurately can we estimate the relationship?
- How good is the model for prediction purposes?

REGRESSION METHODS

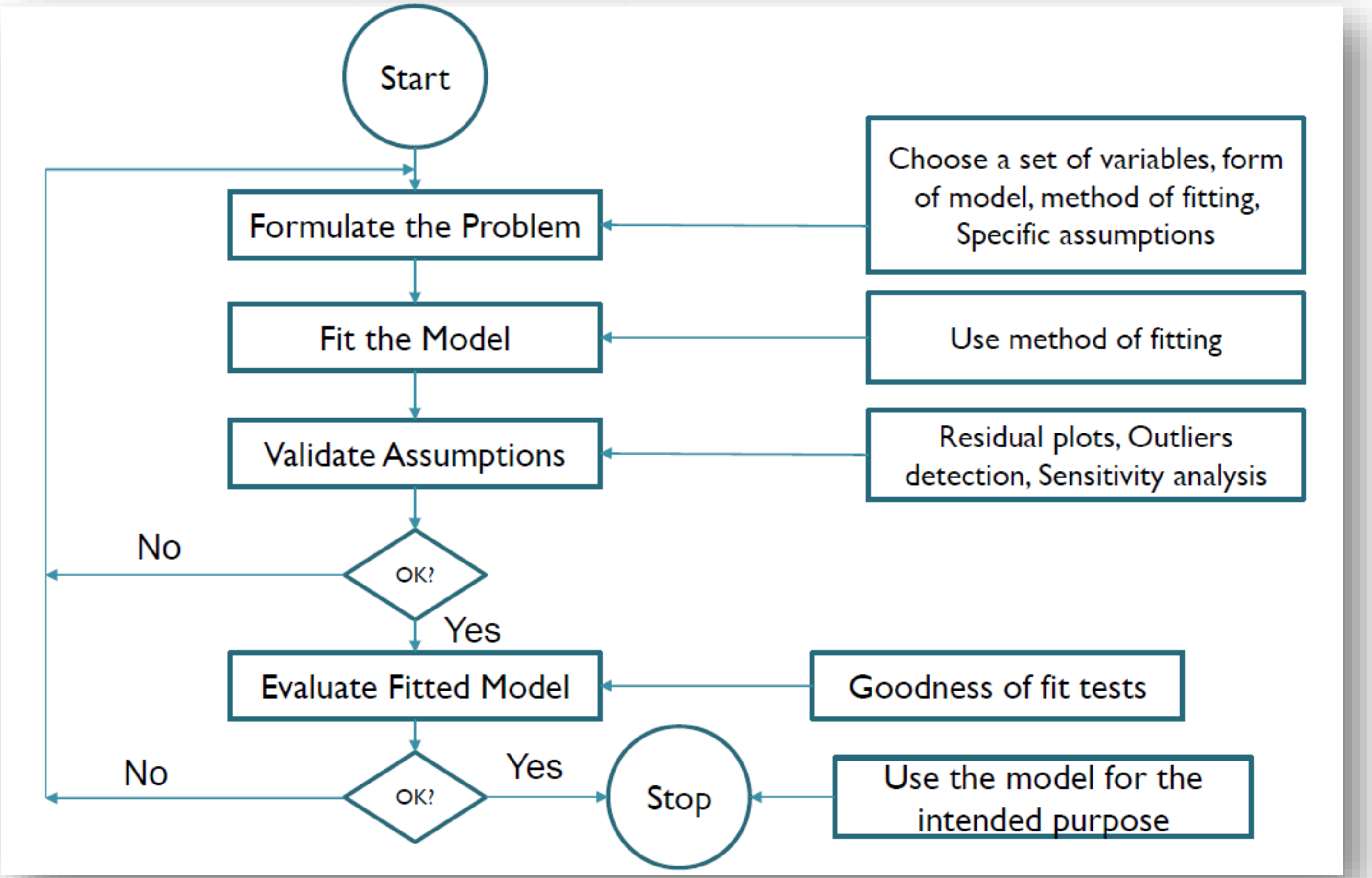
Linear Methods

- Simple linear regression
- Multiple linear regression
- Ridge regression
- Principal component regression
- Lasso
- Partial least squares

Non-linear Methods

- Polynomial regression
- Spline regression
- Neural networks

REGRESSION PROCESS



QUANTITIES THAT INDICATE RELATIONSHIPS BETWEEN VARIABLES

- Pearson Correlation
  - To check whether there is a linear relationship or not.

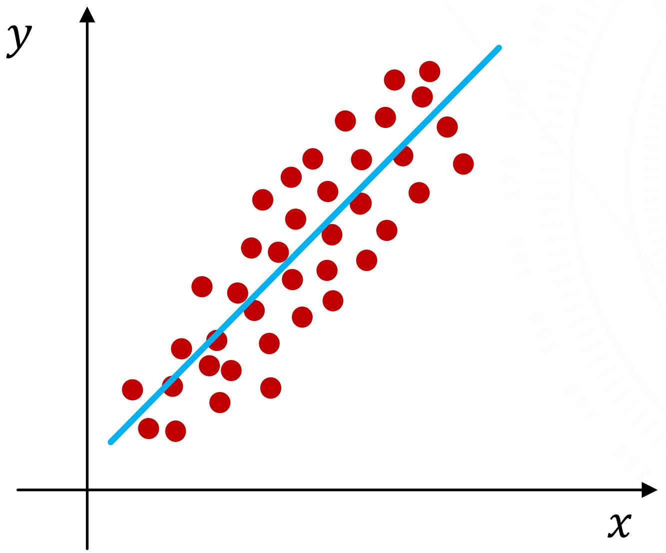
$$\rho^p = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{\sum x^i y^i - n\bar{x}\bar{y}}{\sqrt{\sum {x^i}^2 - n\bar{x}^2}\sqrt{\sum {y^i}^2 - n\bar{y}^2}}$$

- Spearman Correlation
  - To check if the variables vary together monotonically
  - r_x and r_y are ranks for x and y respectively (after sorting in ascending order)
  - If a value is repeated multiple times, ten an average position rank is given
  - Eg: If a value is repeated in 3rd , 4th , and 5th position, r_x for these positions would be 4. r_x for 6th position would be 6

$$\rho^s = \frac{s_{r_x r_y}}{\sqrt{s_{r_x r_x}}\sqrt{s_{r_y r_y}}}$$

- Kendall Correlation
  - To check if there is an ordinal association between variables
  - Given n data points, nC_2 binary pairs are chosen and each pair is labeled as either a concordant or a discordant pair
  - Concordant when either $x^i > x^j$ and $y^i > y^j$ or $x^i < x^j$ and $y^i < y^j$ holds, otherwise discordant pair
  - Data with repeats in x and y can be ignored for simplicity

$$\rho^k = \frac{n_C - n_D}{nC_2}$$



LINEAR REGRESSION

ORDINARY LEAST SQUARES

UNIVARIATE LINEAR REGRESSION

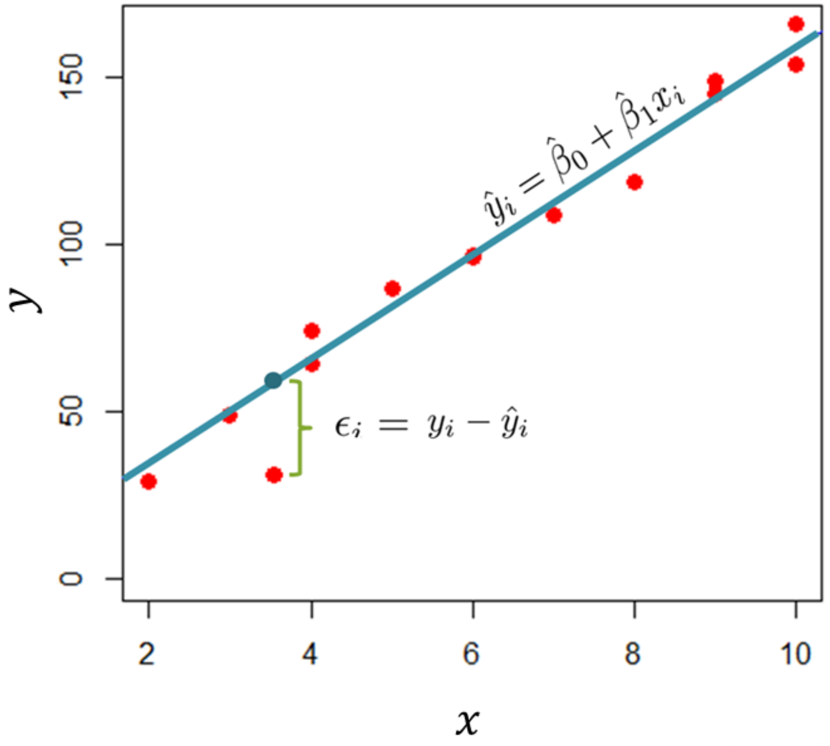
- Objective is to identify a model between a dependent scalar variable y and independent scalar variable x

$$y = \beta_1 x + \beta_0$$

- Assumption: Measurements of x are error free and measurements of y have an additive error e that follows a Gaussian pdf with zero mean

$$y^i = \beta_1 x^i + \beta_0 + e^i$$

- The unknowns β_0 and β_1 are found by minimizing the total error



$$\min \sum_i e_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

MULTIVARIATE LINEAR REGRESSION

- Extension of univariate regression to multiple inputs and outputs
- Objective is to identify a model between one or more dependent scalar variables y and independent variables $x_1, x_2, \ldots, x_p$

$$y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

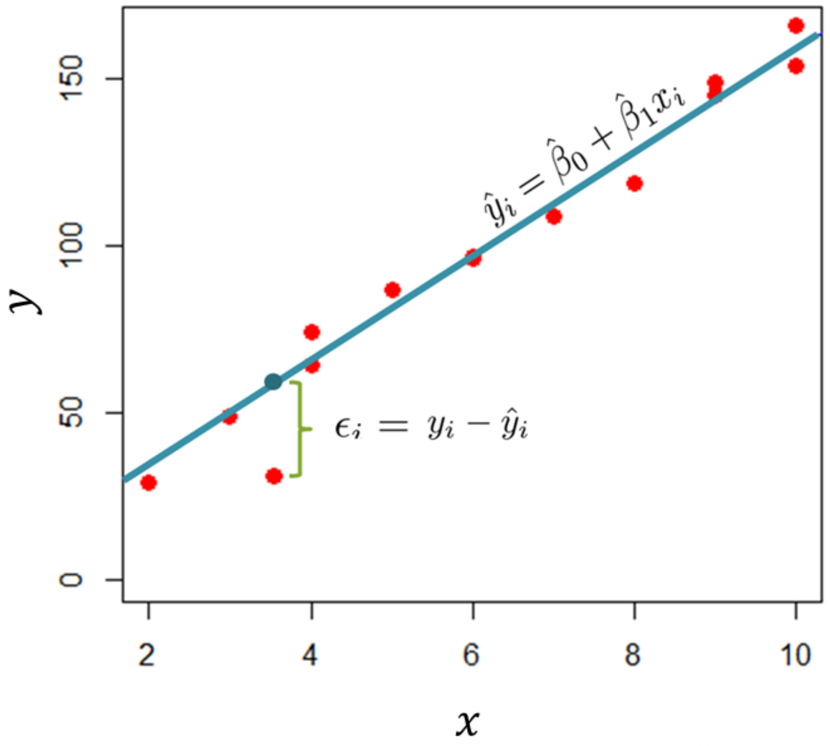
$$y_{meas,j} = y_j + e$$

$$y_{meas} = X\beta + e$$

- The unknowns β_i are found by minimizing the total error

$$\min\ e^T e$$

- Solution: $\hat{\beta} = (X^T X)^{-1} X^T y_{meas}$



POLYNOMIAL REGRESSION (LINEAR IN PARAMETER)

- Objective is to identify a model between one or more dependent scalar variables y and independent variables $x_1, x_2, \ldots, x_p$

$$y_j = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

$$y_{meas} = X\beta + e$$

- Same as multiple linear regression, only difference in X matrix

PROPERTIES OF ESTIMATES

- $\hat{\beta}$ is the best linear unbiased estimator (BLUE)

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

- Estimate of the error variance and variance of estimates:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-p-1}$$

where $(n - p - 1)$ is the degrees of freedom (df)

MODEL ASSESSMENT AND IMPROVEMENT

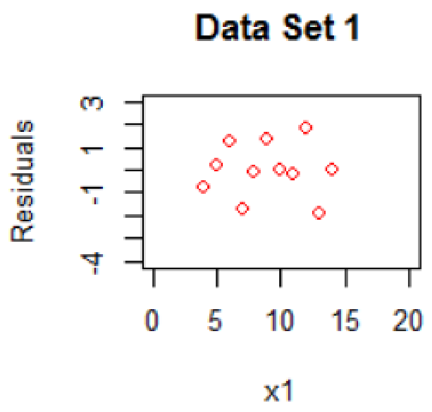
MODEL ASSESSMENT AND IMPROVEMENT

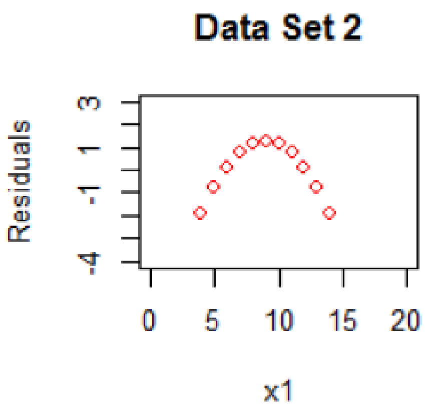
- How good is the fitted linear model?
- Can we improve quality of linear model?
  - Are assumptions made about errors reasonable?
    - Normality: Errors are normality distributed
  - Feature/Model selection
    - Which coefficients of the linear model are significant (Identify important variables)
    - Is the fitted model adequate or can we reduce model complexity?

MODEL ASSESSMENT AND IMPROVEMENT

- Validation of assumptions
  - Residual analysis, Q-Q plots, Residual plots, Outlier detection

No pattern

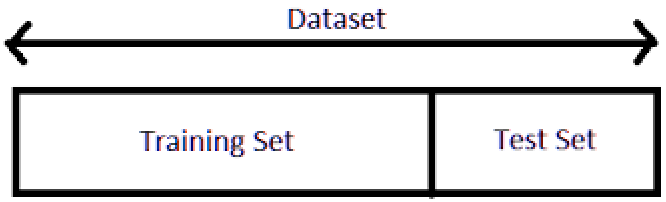




Pattern

MODEL VALIDATION

- Testing the predictive ability of the model by testing the model on new data
- Given dataset is split into two: training set and test set
  - Model is built using a training set
  - Test set is used to test the model



- If the model performs well on the test dataset, we can say that the model is generic enough
- Other approaches
  - K-fold validation
  - Leave one out cross validation

TESTING GOODNESS OF FIT

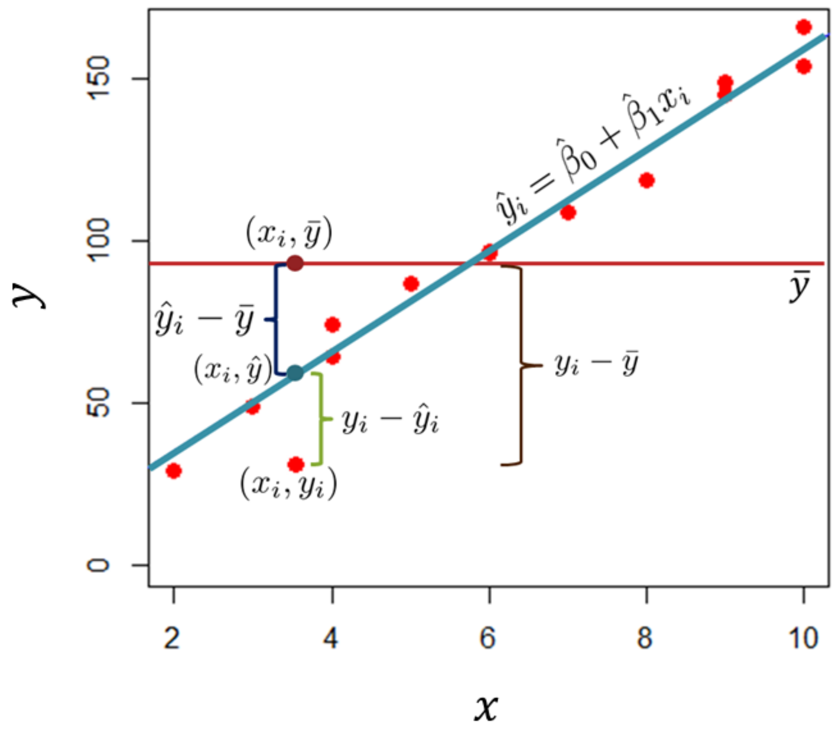
- Coefficient of determination - R^2 is a measure of variability in output variable explained by input variable

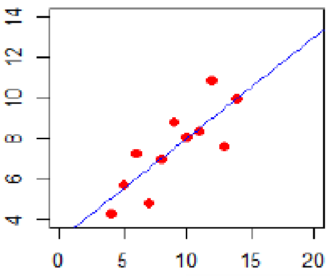
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

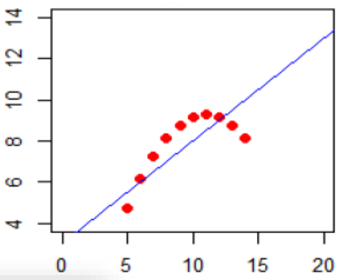
Variability explained by linear model

Total variability in y

- R^2 values: Between 0 and 1 if we evaluate R^2 on the same data we used for fitted the model
    - Values close to 0 indicates poor fit
    - Values close to 1 indicates a good fit







```
              Estimate
(Intercept) 3.000909
x2          0.500000
```

Both models are the same and have similar R^2.
But are they both good?


operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
_operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
_operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
selection at the end -add
_ob.select= 1
ier_ob.select=1
ntext.scene.objects.acti
("Selected" + str(modifier
irror_ob.select = 0
= bpy.context.selected_ob
ata.objects[one.name].se
int("please select exactl

THANKYOU