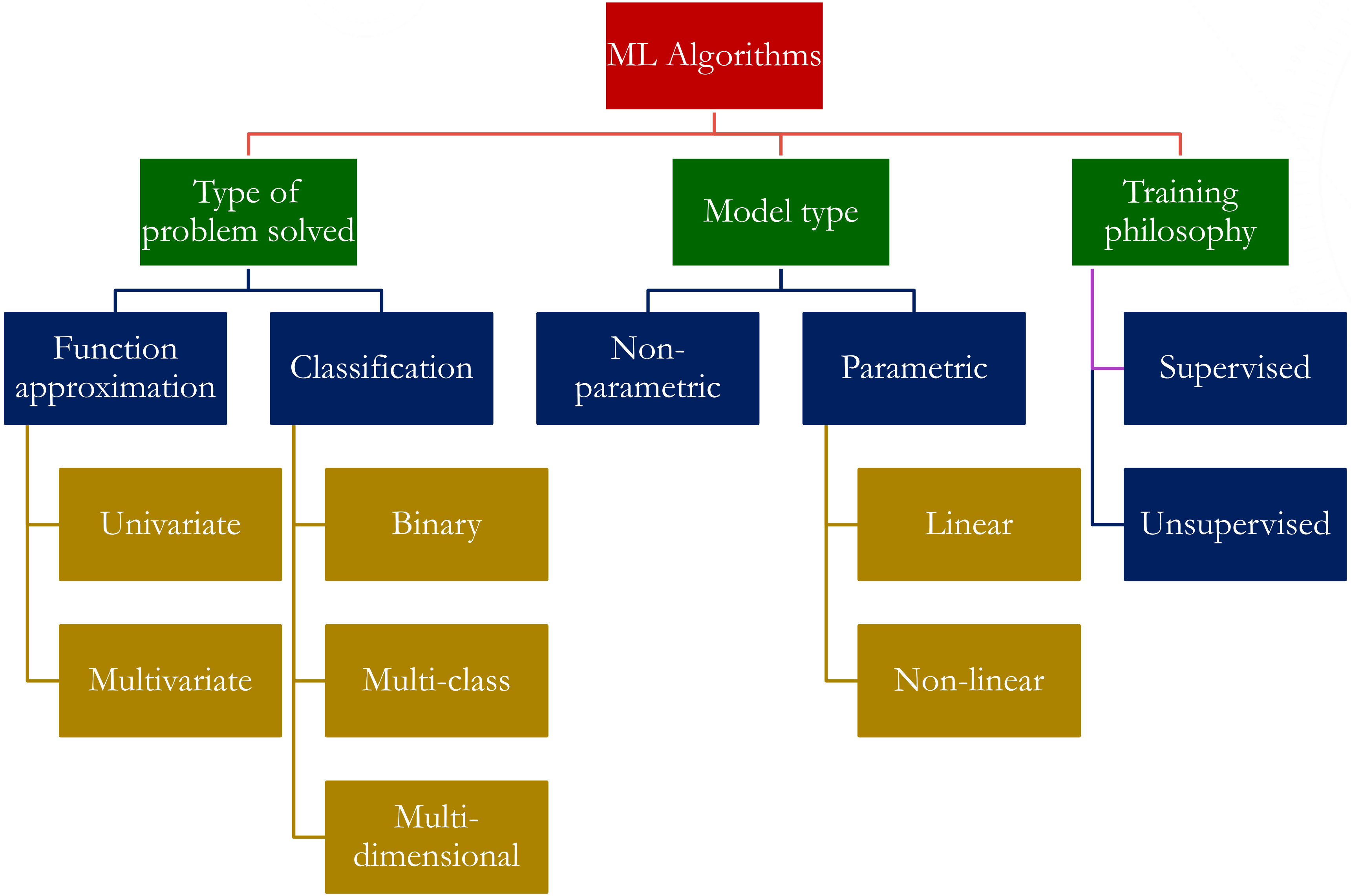
LINEAR REGRESSION

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

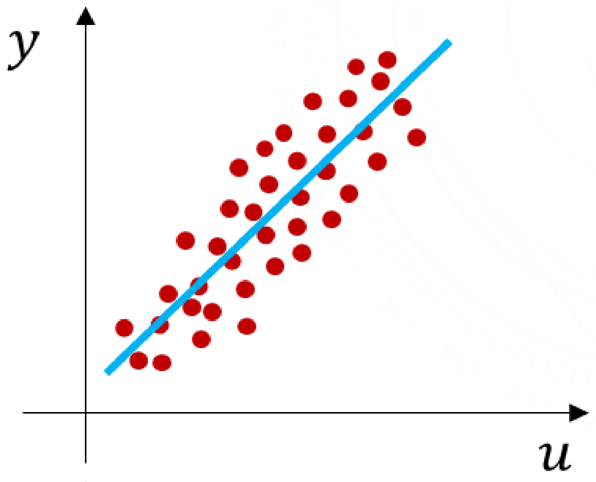
Predicting mechanical properties of a part

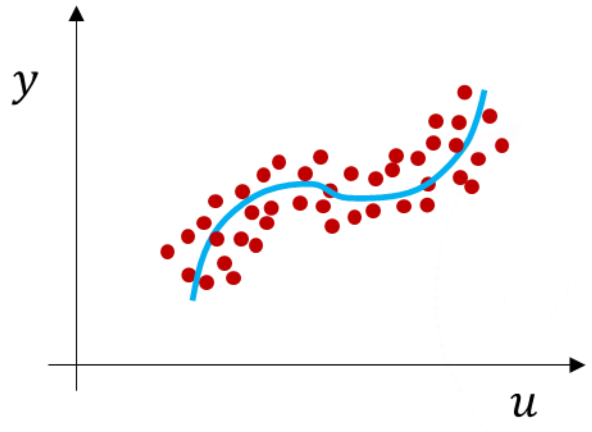
Predicting battery temperature in an electric vehicle

Predicting value of a board position in chess

Techniques:

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





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

CLASSIFICATION

Examples:

Fraud detection in credit card transactions

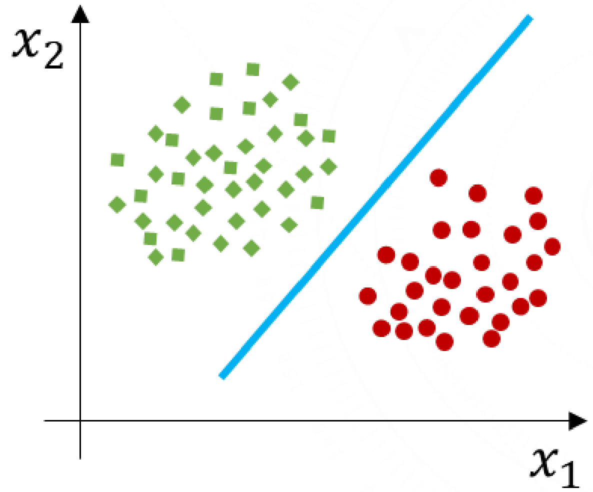
Distinguishing objects – "Self-driving cars"

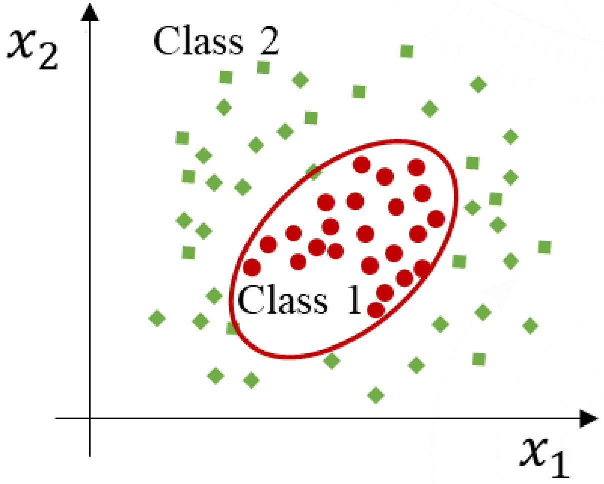
Detecting failures in built systems/equipment

Classifying emails as spam or genuine

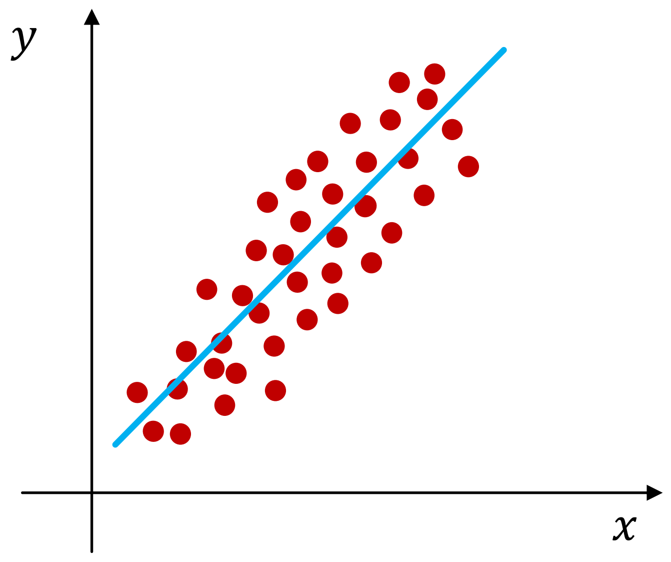
Techniques:

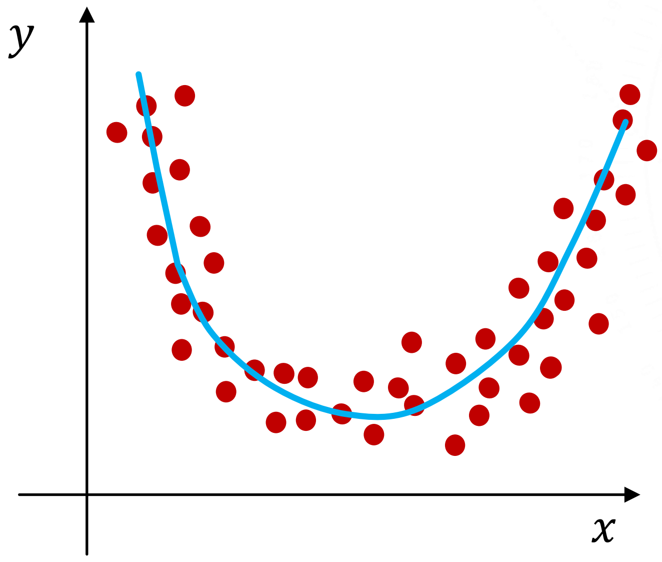
Logistic regression, k-nearest neighbors, Neural network, Decision tree, Random forest, Support vector machines, LDA, QDA, Naïve Bayes, Hierarchical clustering, k-means clustering, …





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

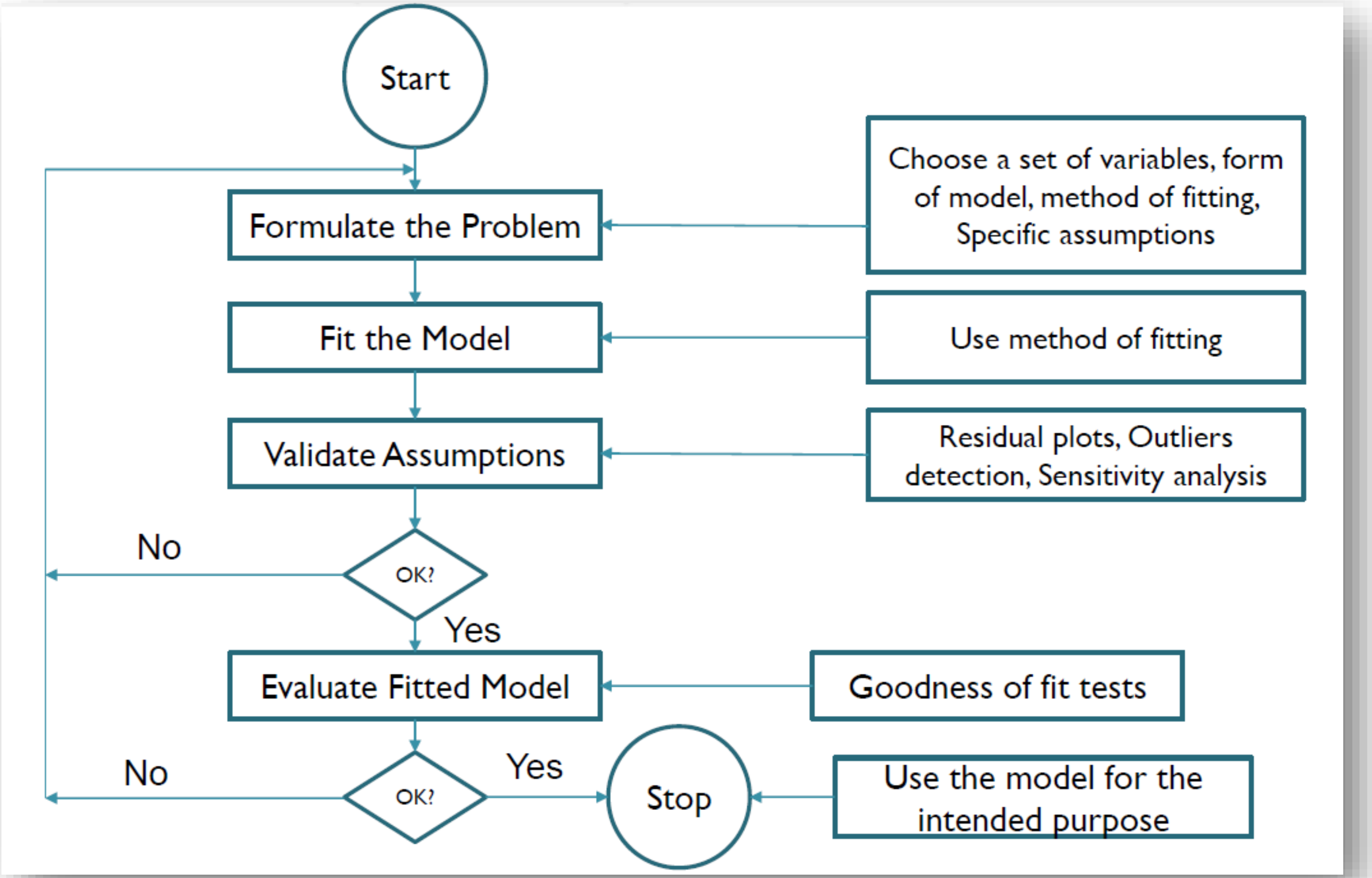
Linear Methods

- 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.

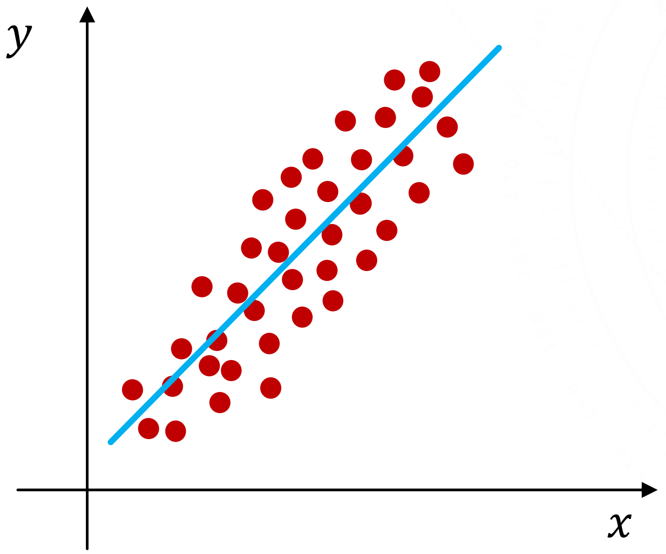
$$\rho^p = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{\sum x^i y^i - n\bar{x}\bar{y}}{\sqrt{\sum {x^i}^2 - n\bar{x}^2}\sqrt{\sum {y^i}^2 - n\bar{y}^2}}$$

- 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

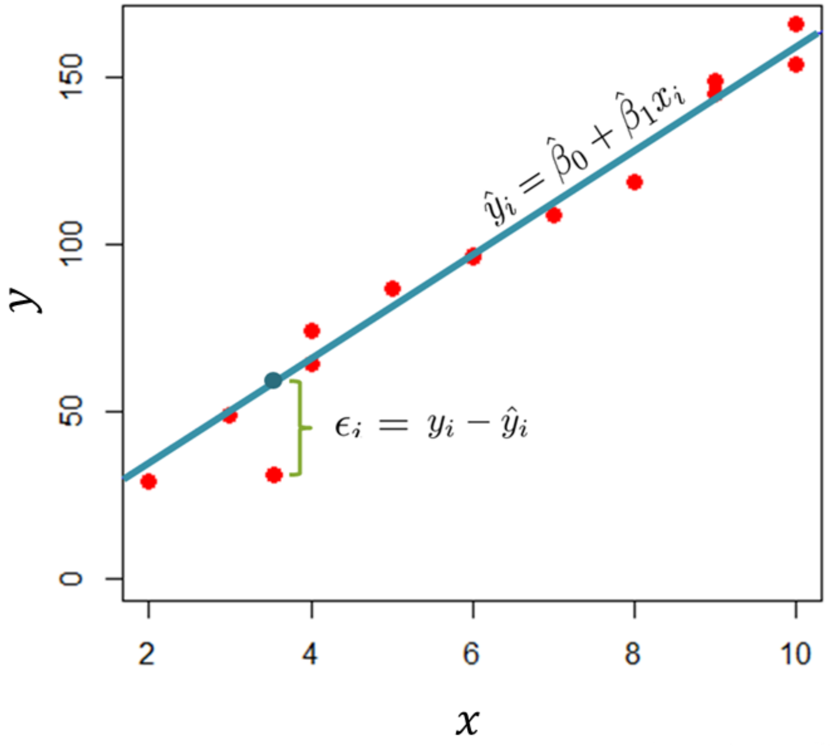
- Objective is to identify a model between a dependent scalar variable y and independent scalar variable x

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

- 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 $x_1, x_2, \dots, x_p$

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

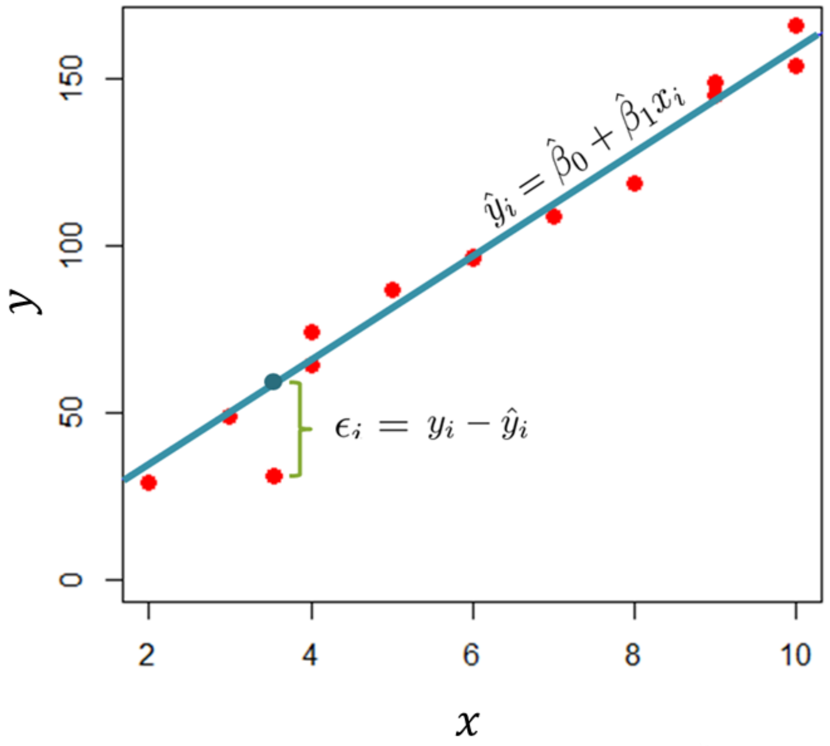
$$y_{meas,j} = y_j + e$$

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

- 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

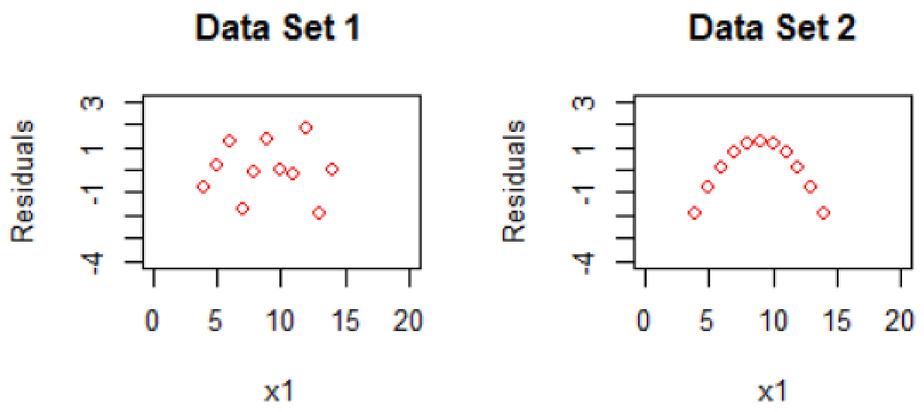
MODEL ASSESSMENT AND IMPROVEMENT

- 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

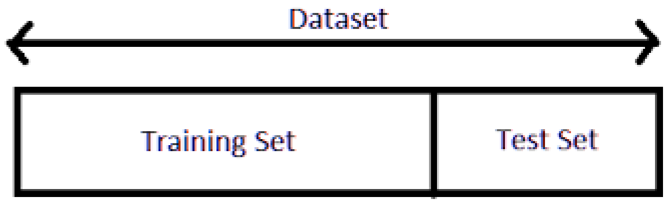
No pattern



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

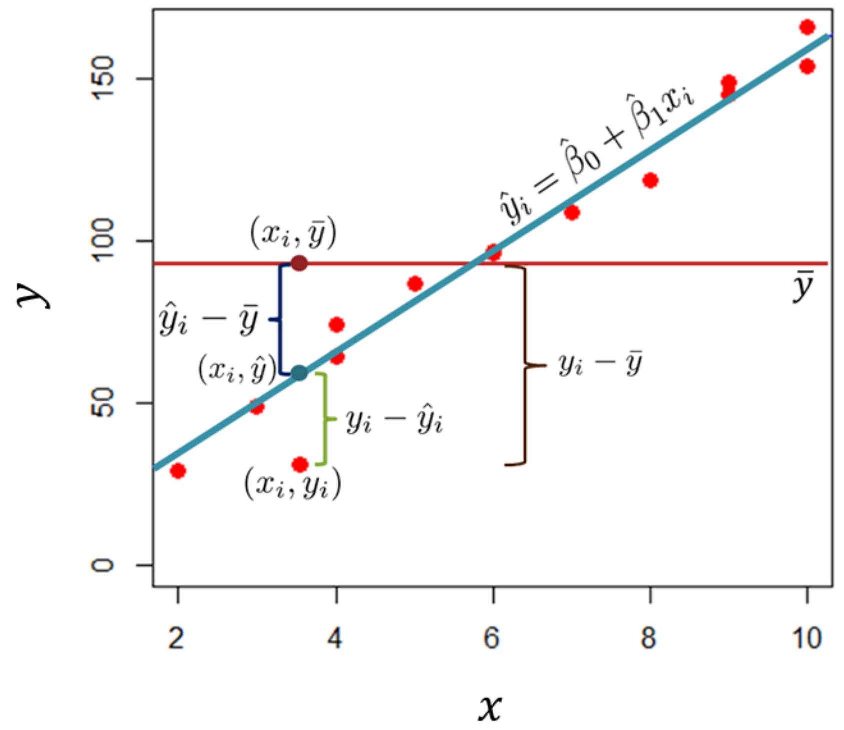
- Coefficient of determination - R^2 is a measure of variability in output variable explained by input variable

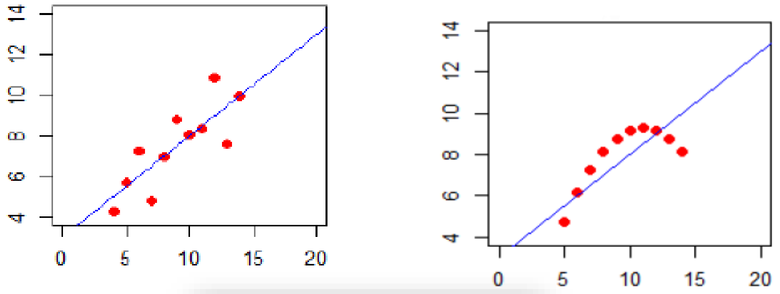
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

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 exact

THANKYOU

LINEAR REGRESSION

RESMI SURESH

ASSISTANT PROFESSOR, IIT GUWAHATI

MODEL ASSESSMENT AND IMPROVEMENT

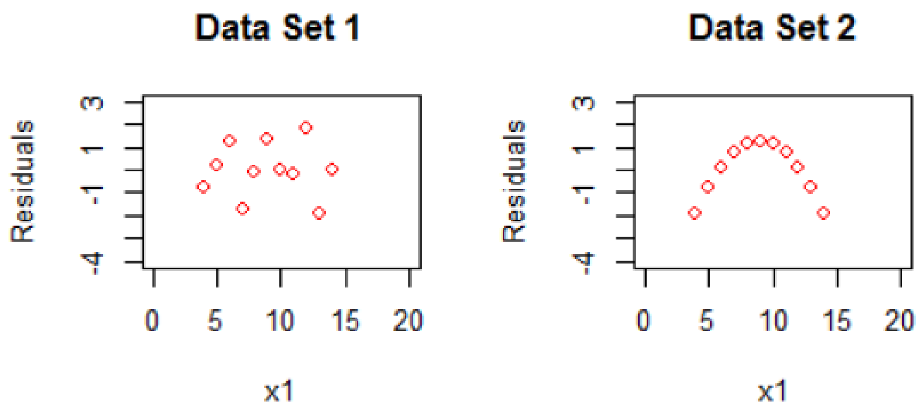
MODEL ASSESSMENT AND IMPROVEMENT

- 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

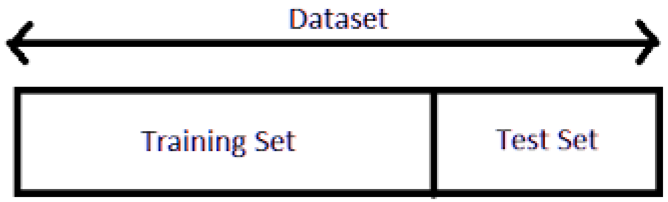
No pattern



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

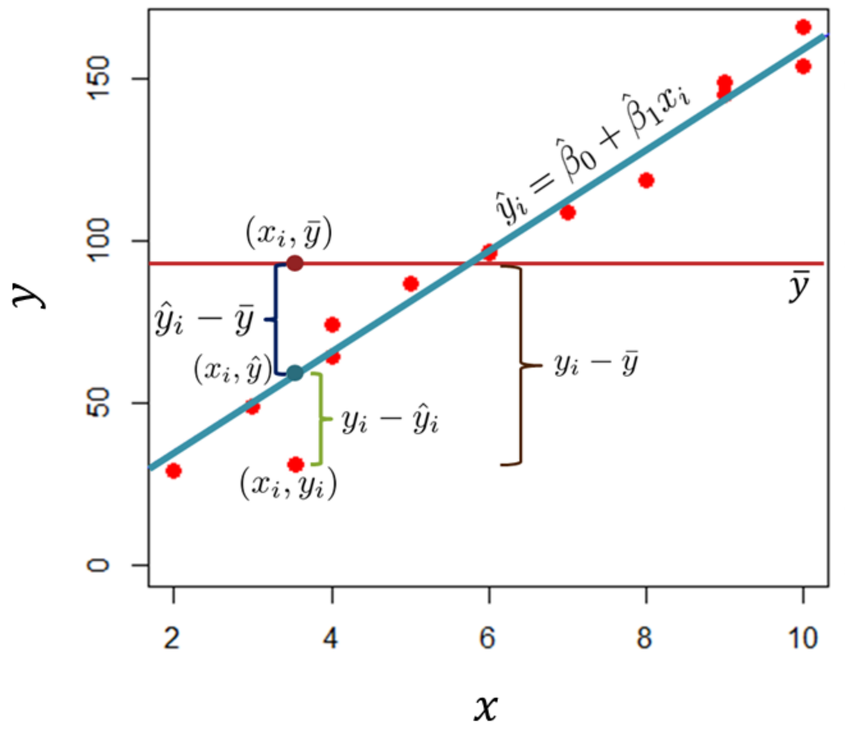
- Coefficient of determination - R^2 is a measure of variability in output variable explained by input variable

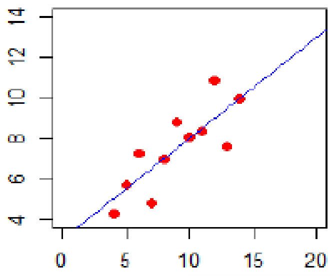
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

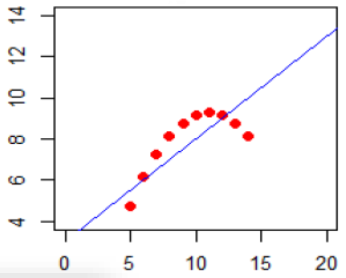
Variability explained by linear model

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
```

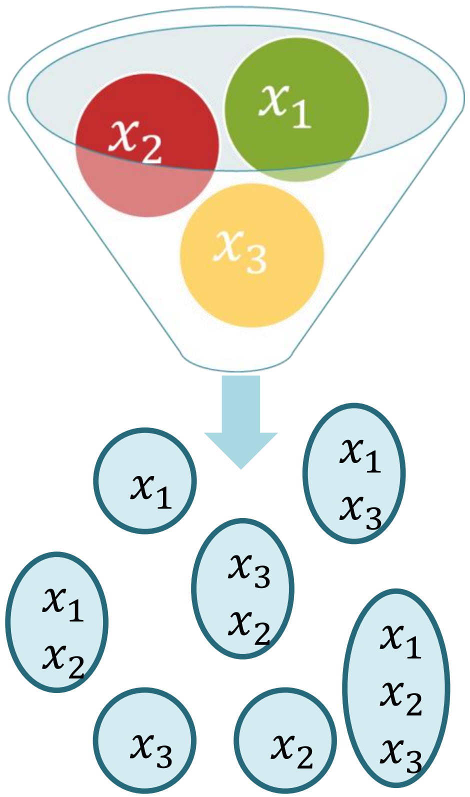
Both models are the same and have similar R^2.
But are they both good?

Model Complexity: Feature and model selection

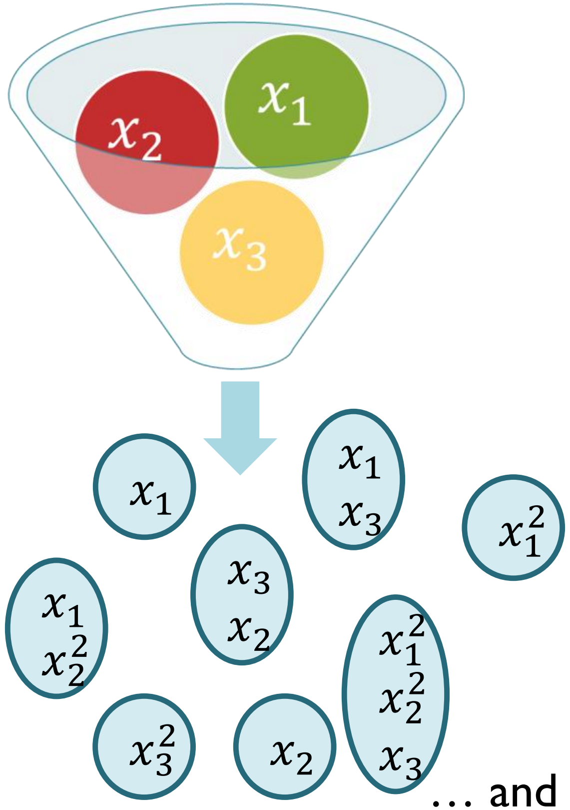
Trade off between SSE values and model complexity

Available Variables (other than y)

Input feature set / possible models



Linear relation



… and many more possibilities

Non-linear relation

How do we find the best model? Fit models and compare using some metric!

Should we fit all possible models? May be. Or use some smarter way of subset selection!

Model Complexity: Feature and model selection

- 2 components:
  - A metric for deciding best model
  - A smart selection of subsets for fitting
- Commonly used metrics

With Cross-validation

MSE of test set
R^2 of test set

Without Cross-validation

R^2_{adj} of training set
t-test on fitted parameters
F-test and p-values
AIC (Akaike Information Criterion) of training set
BIC (Bayesian Information Criterion) of training set

Model Complexity: Feature and model selection

- Adjusted $\bar{R}^2$

$$R^2_{adj} = 1 - \frac{\sum (y^i - \hat{y}^i)^2/(n-p-1)}{\sum (y^i - \bar{y})^2/(n-1)}, \quad n > p+1$$

$$R^2_{adj} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

  - Example: Consider a dataset of 20 samples with $\sum \left(y^i - \bar{y}\right)^2 = 2$. A linear model for this data gives SSE = 0.3 and a 9th order polynomial gives SSE = 0.1, which model is better?

$$R^2_{linear} = 1 - \frac{0.3}{2} = 0.85 \text{ and } R^2_{poly} = 1 - \frac{0.1}{2} = 0.95$$

From R^2 value, it may seem like the 9th order polynomial model to be better than the linear model. But it is using a large number of parameters.

$$R^2_{adj,linear} = 1 - \frac{\frac{0.3}{\frac{20-2}{2}}}{20-1} = 0.84 \text{ and } R^2_{adj,poly} = 1 - \frac{\frac{0.3}{\frac{20-10}{2}}}{20-1} = 0.71$$

Linear model is better

Model Complexity: Feature and model selection

- t-test on fitted parameters

$$H_0\!: \beta_i = 0 \;; \qquad\qquad H_1\!: \beta_i = 0$$

  - Perform t-test or
  - Find the confidence interval for each parameter using t-values. If '0' is part of the confidence interval for parameter β_i, then the term with that parameter $(\beta_i x_i)$ could be considered insignificant and can be removed

$$T = \frac{\beta - 0}{\sigma_\beta}$$

- F-test I

H_0: *Reduced model (without $\beta_i x_i$) is adequate*, $\sigma_{FM}^2 = \sigma_{RM}^2$

H_1: *Full model (with $\beta_i x_i$) is adequate* , $\sigma_{FM}^2 < \sigma_{RM}^2$

where σ^2 *is error variance*

  - Compare the ratio $F = \left(\frac{SSE_{FM}}{df_{FM}}\right) \Big/ \left(\frac{SSE_{RM}}{df_{RM}}\right)$ with $f_\alpha(df_{FM}, df_{RM})$

  - If F $\ll f_\alpha(df_{FM}, df_{RM})$, reject null hypothesis implying that the term $\beta_i x_i$ is significant

df is degrees of freedom, RM is reduced model, FM is full model

- F-test 2 and p-value

  H_0: *Reduced model (without $\beta_i x_i$) is adequate,* $\sigma^2_{FM} = \sigma^2_{RM}$

  H_1: *Full model (with $\beta_i x_i$) is adequate ,* $\sigma^2_{FM} < \sigma^2_{RM}$

  where σ^2 is error variance

  - Compare the ratio $F = \frac{(SSE_{RM} - SSE_{FM})}{\frac{SSE_{FM}}{n-p-1}}$ with $f_\alpha(1, df_{FM})$
  - If F $\gg f_{1-\alpha}(1, df_{FM})$, reject null hypothesis implying that the term $\beta_i x_i$ is significant

    df is degrees of freedom, RM is reduced model, FM is full model

  - p-value - smallest value of α that would have resulted in rejection of null hypothesis: $p - value = (\alpha \; such \; that \; F_{1-\alpha}(1, n - p - 1) = F)$

Model Complexity: Feature and model selection

- AIC (Akaike information criterion)

$$AIC = 2k - 2\ln L$$

Where n is the number of samples, k is the number of parameters in the model and L is the likelihood function. AIC penalizes for additional parameters used in reducing SSE.

For Gaussian error,

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^{n} e_i^2}{2\sigma^2}$$

If MLE estimate $\sigma^2 = \frac{SSE}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}$ is used to find unknown σ^2

$$AIC = 2k + n\ln(2\pi) + n\ln\left(\frac{SSE}{n}\right) + n$$

Model Complexity: Feature and model selection

- AIC (Akaike information criterion)

For Gaussian error,

$$AIC = 2k + n\ln(2\pi) + n\ln\left(\frac{SSE}{n}\right) + n$$

As k increases, initially SSE will decrease considerably reducing AIC. However, after a while, increase in k may only result in slight reduction in SSE values. Then, the impact of the term $2k$ will be more significant that $n\ln\left(\frac{SSE}{n}\right)$ and AIC may increase.

Model with lowest AIC value can be chosen to be the best model

Model Complexity: Feature and model selection

- BIC (Bayesian information criterion)

$$BIC = k \ln n - 2 \ln L$$

Where n is the number of samples, k is the number of parameters in the model and L is the likelihood function.

Similar to AIC, BIC also penalizes for additional parameters used in reducing SSE.

For Gaussian error,

$$BIC = k \ln n + n \ln(2\pi) + n \ln\left(\frac{SSE}{n}\right) + n$$

Similar behavior as AIC

Model with lowest BIC value can be chosen to be the best model

Model Complexity: Feature and model selection

- Example: AIC and BIC

For the following data, find the best polynomial model to predict y given x

Average Temperature (x)	6.6	26.1	6.3	27.6	14.7	18.3	23.1	15.6	9	5.7
Electricity Bill (y)	118.2	5607	105.3	6616.8	1043.7	1971.6	3907.8	1239	264	83.4

Model order	Optimal model parameters	SSE	AIC	BIC
1	[-2143.18, 277.05]	5078172	163.76	164.36
2	[897.23, -219.13, 15.34]	39949.94	117.31	118.21
3	[5.07, 3.52, 0.03, 0.31]	0.2135	-2.08	-0.87
4	[7.86, 2.63, 0.12, 0.3 5.9×10^{-5}]	0.151851	-3.5	-1.98
5	[7.88, 2.62, 0.12, 0.3, 6.25×10^{-5}, -3.95×10^{-5}]	0.15185	-1.5	0.32
6	[7.74, 2.68, 0.11, 0.31, 7.3×10^{-6}, 1.33×10^{-6}, -1.35×10^{-8}]	0.151848	0.5	2.62

Best model is of order 4

Model Complexity: Feature and model selection

- Subset selection strategies
  - Best subset
    - Fit all possible subsets and choose the best among them based on some metric like AIC or BIC
  - Forward selection
    - Start with the most significant feature and keep adding features till no improvement
  - Backward elimination
    - Start with full feature set and keep removing features till no improvement

Model Complexity: Feature and model selection

- Forward selection
  - Start with the most significant feature and keep adding features till no improvement
  - Various metrics can be chosen to decide the priority list of features
    - Correlation coefficient
    - AIC
  - Various metrics can be used to comment on the desired performance of the model (stopping criteria)
    - Example: We break the search when a desired value of $R^2 = 0.99$ is achieved or when there is no significant improvement in R^2 or AIC starts to increase

Model Complexity: Feature and model selection

- Forward selection based on correlation coefficient
  - Find the correlation coefficient between each of the independent variables and the dependent variable ($\rho_{x_i,y}$)
  - Sort the independent variables based on the correlation coefficient values from highest to lowest
  - Pick the independent variable with highest $\rho_{x_i,y}$ and build a model
  - If performance satisfied, stop. Else, add the next variable and proceed until stopping criteria is met
  - Stopping criteria: We break the search when a desired value of $R^2 = 0.99$ is achieved or when there is no significant improvement in R^2

Model Complexity: Feature and model selection

- Forward selection based on AIC
  - Find AIC values for each univariate model relating x_i to y
  - Pick the variable with lowest AIC, say x_k
  - For the next iteration, find AIC values of all bi-variate models with one of the features being x_k
    - For example, if x_2 was selected in the first step where $m = 3$, in the second iterations, AIC for models with input features (x_2, x_1) and (x_2, x_3) would be compared
  - Stopping criteria: Break the search when AIC starts to increase with iteration

Model Complexity: Feature and model selection

- Forward selection based on AIC

Features	x_1	x_2	x_3	x_4	x_5
AIC	129.92	130.8	125.51	130.72	125.96

Features	x_1, x_3	x_2, x_3	x_3, x_4	x_5, x_3
AIC	103.36	127.34	126.76	118.46

Features	x_1, x_2, x_3	x_1, x_3, x_4	x_1, x_3, x_5
AIC	104.73	104.66	100.73

Features	x_1, x_2, x_3, x_5	x_1, x_3, x_4, x_5
AIC	101.32	102.72

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5$$

Model Complexity: Feature and model selection

- Backward selection
  - Start with the full model and remove the most insignificant features one by one
  - Various metrics can be chosen to decide the feature to be removed
    - p-value
    - AIC

Model Complexity: Feature and model selection

- Backward selection based on p-value
  - Build the full model
  - Find p-value of each input feature
  - If the highest p-value is greater than 0.05, we remove the corresponding variable for the remaining iterations
  - Once we remove one variable, we repeat the same procedure and keep removing variables until we find that the performance is not improved
  - Stopping criteria: All p-values are less than 0.05

Model Complexity: Feature and model selection

- Backward selection based on AIC
  - If m input features are available, build model with $m - 1$ features (removing 1 variable at a time)
  - Find AIC values for each of the models
  - The independent variable whose removal leads to the minimum AIC is found (say x_k) and that variable is removed from the later steps
  - For the next iteration, find AIC values of all models with $m - 2$ features with one of the missing features being x_k
    - For example, if x_2 was removed in the first step where $m = 4$, in the second iterations, AIC for models with input features (x_3, x_4), (x_1, x_4) and (x_1, x_3) would be compared
  - Stopping criteria: Break the search when AIC starts to increase with iteration

Model Complexity: Feature and model selection

- Backward selection based on AIC

Features	x_1, x_2, x_3, x_4	x_1, x_2, x_3, x_5	x_1, x_2, x_4, x_5	x_1, x_3, x_4, x_5	x_2, x_3, x_4, x_5
AIC	103.84	101.32	128.66	102.72	122.12

Features	x_1, x_2, x_3	x_1, x_2, x_5	x_1, x_3, x_5	x_2, x_3, x_5
AIC	104.73	129.18	100.73	120.44

Feature	x_1, x_3	x_1, x_5	x_3, x_5
AIC	103.36	127.93	118.46

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5$$

LINEARLY DEPENDENT FEATURES

- If some of the input features are dependent, the matrix X will be rank deficient
- The matrix $X^T X$ will be singular or close to singular (ill conditioned)
- Condition number: Ratio of the largest to lowest eigenvalue
  - A high value indicates that the matrix is ill conditioned
- How do we find the linear model parameters in such cases?
  - Ridge regression

RIDGE REGRESSION

- Linear regression with L_2 regularization

$$\min_{\beta} \quad ||y - X\beta||^2 + \lambda\left||\beta|\right|^2$$

where $||.||$ is the 2-norm and λ is the tuning parameter

- Equivalent to

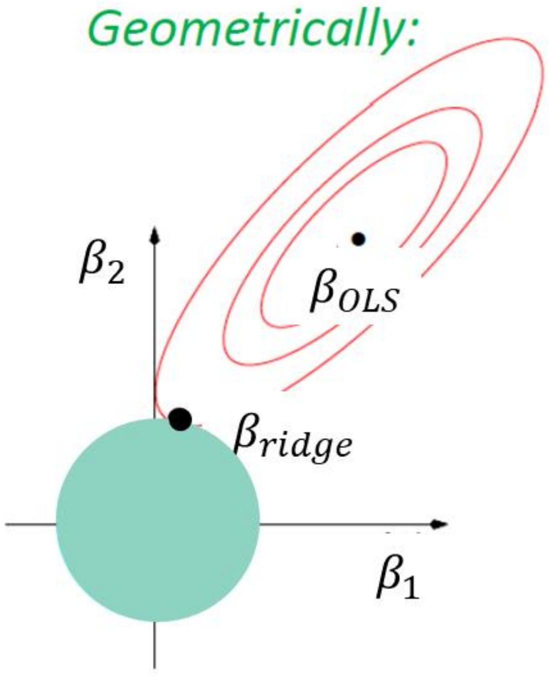
$$\min_{\beta} \quad ||y - X\beta||^2$$

Subject to $\quad \left||\beta|\right|^2 < k$

- Helps to avoid overfitting
- Always results in a unique solution and works well with ill-conditioned data

$$\frac{dJ}{d\beta} = 0 \Rightarrow -2X^T(y - X\hat{\beta}) + 2\lambda\hat{\beta} = 0$$

$$\Rightarrow X^T y = \left(X^T x + \lambda I\right)\hat{\beta} \Rightarrow \hat{\beta} = \left(X^T X + \lambda I\right)^{-1} X^T y$$



LASSO REGRESSION

- Least Absolute Shrinkage and Selection Operation
- Linear regression with L_1 regularization

$$\min_{\beta} \quad ||y - X\beta||^2 + \lambda|\beta|_1$$

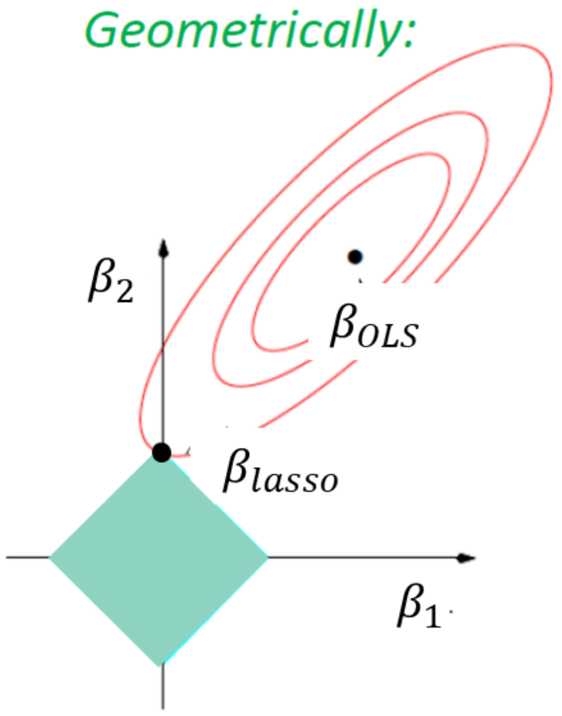
where $||.||$ is the 2-norm, $|\beta|_1$ is the 1-norm of β and λ is the tuning parameter

- Equivalent to

$$\min_{\beta} \quad ||y - X\beta||^2$$

Subject to $\qquad |\beta|_1 < k$

- Helps in getting sparse β and so, helps in feature selection
- No closed form solution, need to find β by solving the optimization problem



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

THANKYOU