
# Regression Analysis

## (Theory)

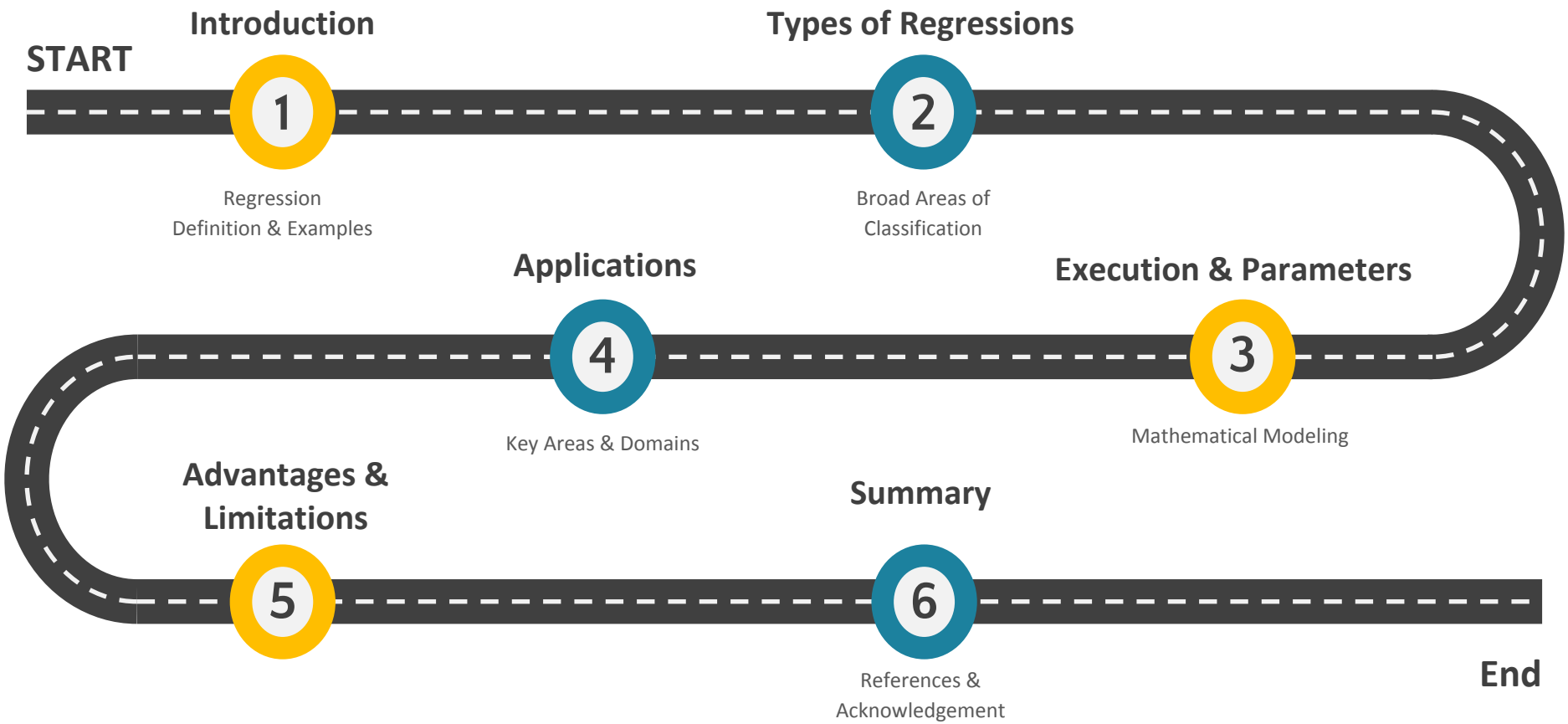
**CH5020**

**Internal Report Presentation**

Course Instructor: Dr. Kannan A
Department of Chemical Engineering, IIT Madras

Mayur Vikas Joshi | ME16B148
Sushant Uttam Wadavkar | ME16B172

# Roadmap of Presentation



# Types of Regression

## 01. Linear Regression

Consists of a predictor variable and a dependent variable related linearly to each other.

## 02. Logistic Regression

**Logistic regression** is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable

## 03. Ridge Regression

**Ridge Regression** is a technique for analyzing multiple regression data that suffer from multicollinearity

## 04. Lasso Regression

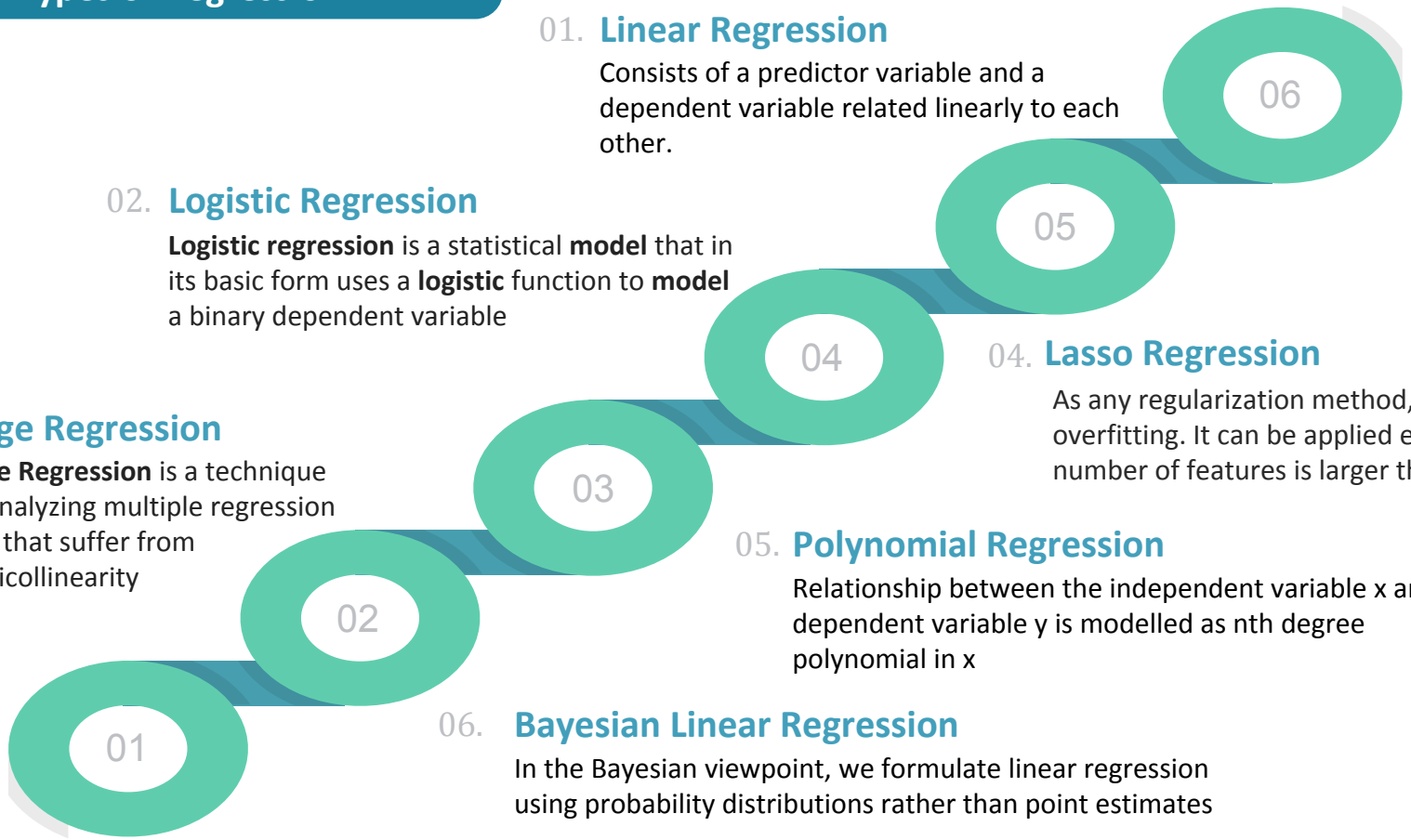
As any regularization method, it can avoid overfitting. It can be applied even when number of features is larger than (n)

## 05. Polynomial Regression

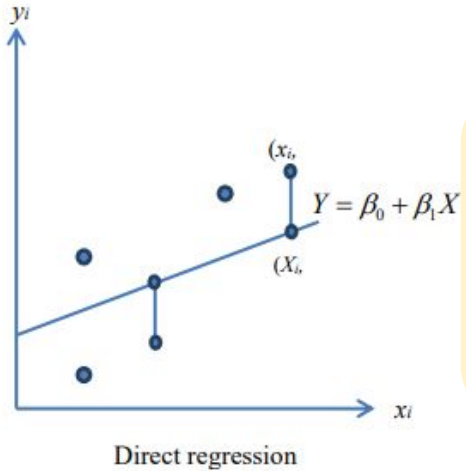
Relationship between the independent variable x and the dependent variable y is modelled as nth degree polynomial in x

## 06. Bayesian Linear Regression

In the Bayesian viewpoint, we formulate linear regression using probability distributions rather than point estimates



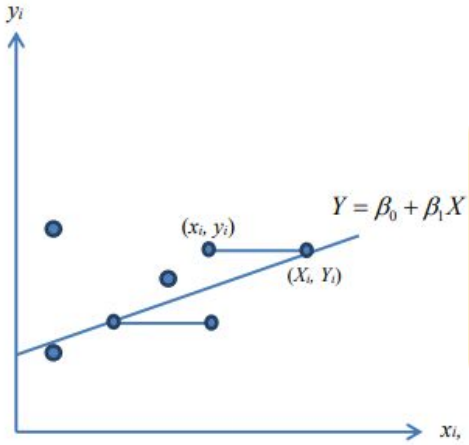
# Estimation of parameters



Direct regression

## Direct Regression (OLS)

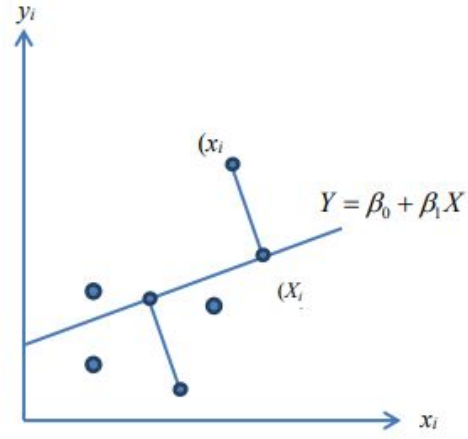
Vertical difference between the observations and the line and its sum of squares is minimized



Reverse regression method

## Reverse Regression

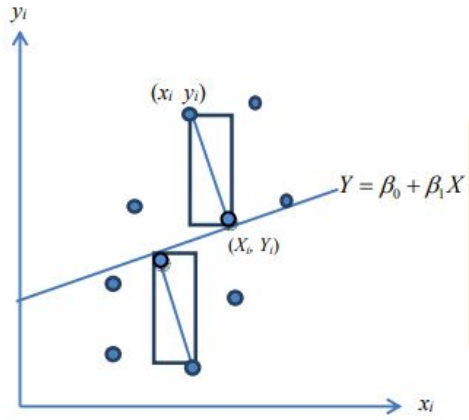
Sum of squares of the difference between the observations and the line in the horizontal direction



Major axis regression method

## Major Axis Regression

Sum of squares of perpendicular distances between the observations and the line is minimized



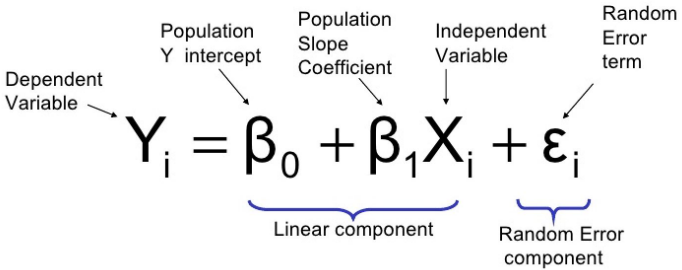
Reduced major axis method

## Reduced Major Axis

Sum of the areas of rectangles defined between the observed data points and the nearest point on the line is minimized

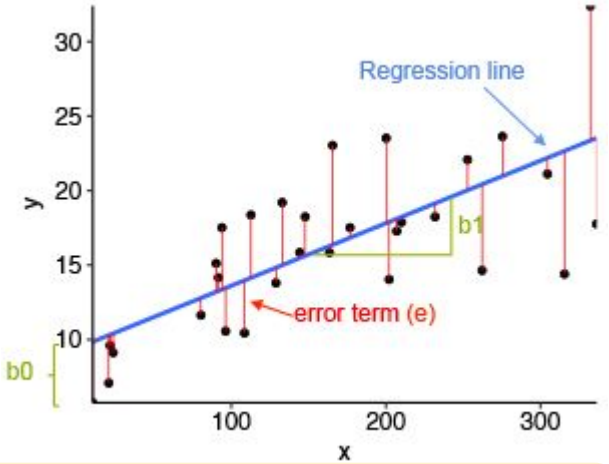
# Simple Linear Regression

In statistics, **linear regression** is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.




$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$




- For more than one explanatory variable, the process is called **multiple linear regression**.
- This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

# Multiple Linear Regression

**Multiple linear regression** (MLR), also known simply as **multiple regression**, is a statistical technique that uses **several** explanatory variables to predict the outcome of a response variable. **Multiple regression** is an extension of **linear** (OLS) **regression** that uses just one explanatory variable.

**Simple Linear Regression**

$$y = b_0 + b_1 * x_1$$

**Multiple Linear Regression**

Dependent variable (DV) — Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

$$y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

## Advantages of MLR over SLR

- Running separate simple linear regressions will lead to different outcomes when we are interested in just one
- Besides that, there may be an input variable that is itself correlated with or dependent on some other predictor
- This can cause wrong predictions and unsatisfactory results

# Code - MLR

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
data = pd.read_csv("Advertising.csv") #Example Dataset: "Advertising.csv"
print(data.head())
data.drop(["Unnamed: 0"], axis=1, inplace=True)
```

```
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LinearRegression

xs = data.drop(["Sales"], axis=1)
y = data["Sales"].values.reshape(-1,1)
linreg = LinearRegression()
MSE = cross_val_score(linreg, xs, y, scoring="neg_mean_squared_error", cv=5)

mean_MSE = np.mean(MSE)
print(mean_MSE)
```

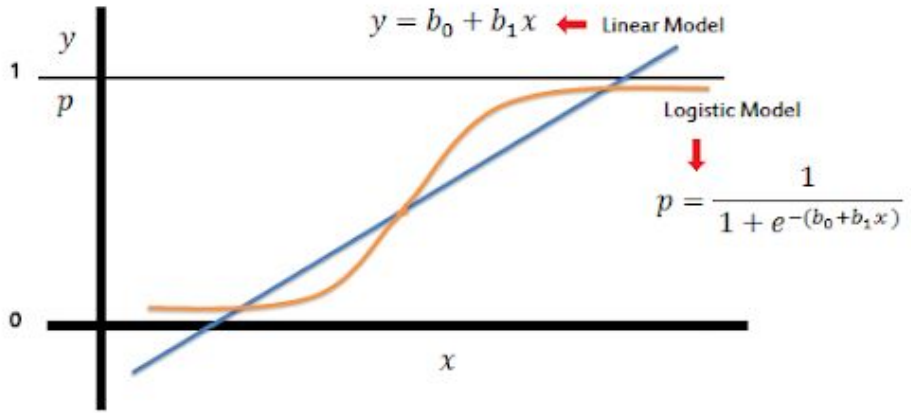
# Logistic Regression

In statistics, the **logistic model** (or **logit model**) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

$$=> P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



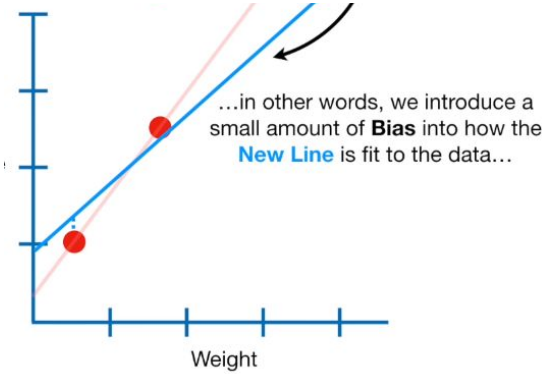

- This can be extended to model several classes of events such as determining if images contain cat, dog, lion, etc.
- Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

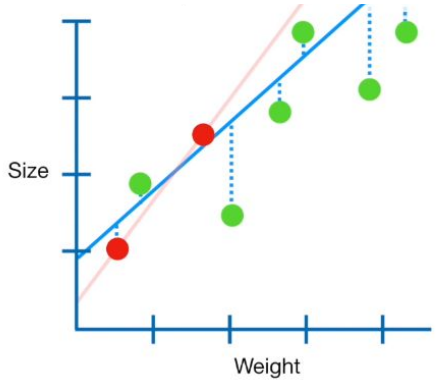
# Ridge Regression

**Ridge regression** is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$




- Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity
- When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value
- By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors

# Ridge Regression

## Advantages

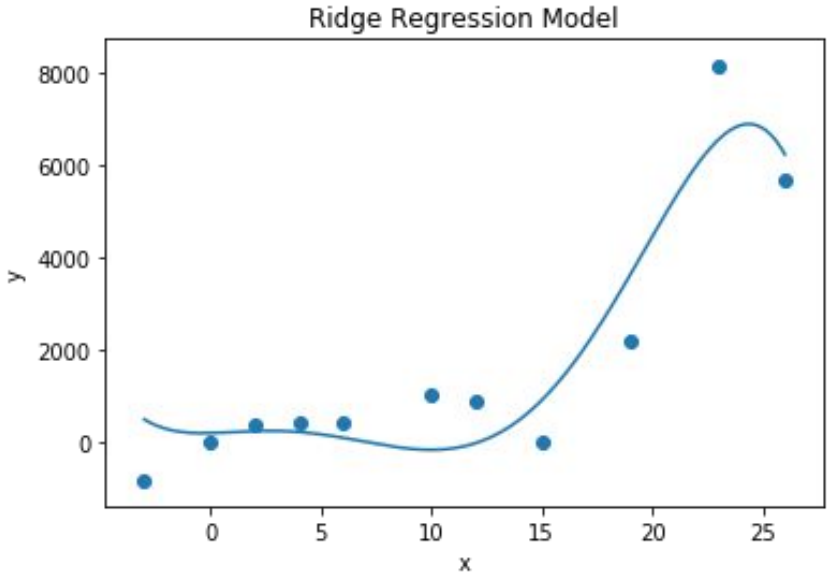
- Ridge regression can reduce the variance (with an increasing bias) – travel along the bias variance tradeoff curve
- Works best in situations where the OLS estimates have high variance
- Can improve predictive performance of the model (Generalization power increases)
- Works in situations where (p) < (n)

## Disadvantages

- Ridge regression is not able to shrink coefficients to exactly zero
- As a result, it cannot perform variable selection

# Ridge Regression

$$Y = -683.1 + 11.73*X + 14.5*X^2 - 5.8*X^3 + 0.49*X^4 - 0.01*X^5$$




**R2 score: 0.87**

- The overfitting has reduced a bit which can be seen from the regression curve as well.
- The model estimates were shrunk a little bit to limit overfitting and improve the generalization capability of the model (increase the bias of the model).

# Code - Ridge Regression

$$Y = -683.1 + 11.73*X + 14.5*X^2 - 5.8*X^3 + 0.49*X^4 - 0.01*X^5$$

```
# Ridge Regression
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Ridge
ridge = Ridge()

parameters = {"alpha":[1e-15, 1e-10, 1e-8, 1e-4, 1e-3, 1e-2, 1, 5, 10, 20]}
ridge_regression = GridSearchCV(ridge, parameters,
scoring='neg_mean_squared_error', cv=5)
ridge_regression.fit(xs, y)
```

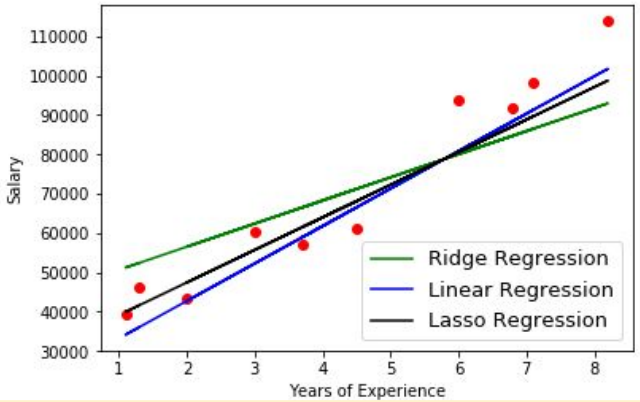
```
print(ridge_regression.best_params_)
print(ridge_regression.best_score_)
```

# Lasso Regression

The **“LASSO”** stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.
Lasso regression is a type of **regression** that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$




- The **lasso** procedure encourages simple, sparse models (i.e. models with fewer parameters)
- This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination
- Lasso Regression uses L1 regularization technique, it is used when we have more number of features because it automatically performs feature selection

# Lasso Regression

## Advantages

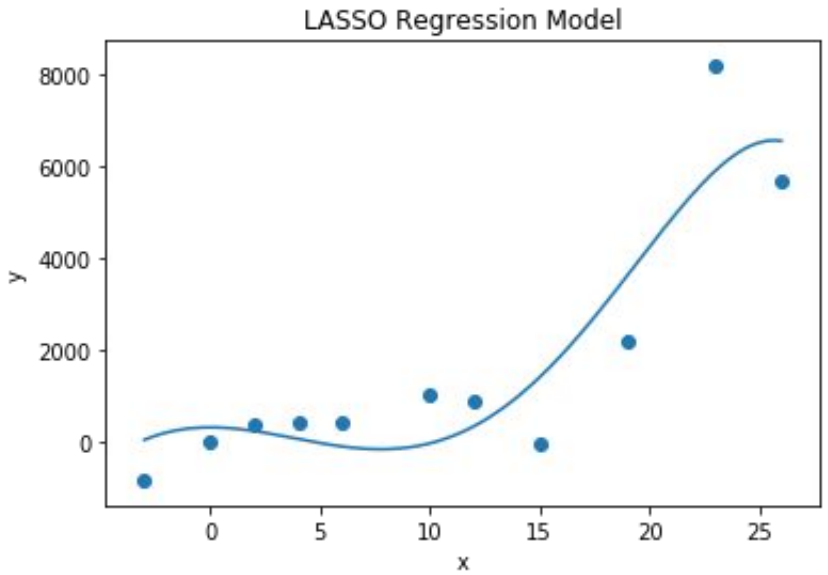
- Lasso adds an additional term to the cost function, adding the sum of the coefficient values (L-1 norm) multiplied by a constant lambda
- This additional term penalizes the model for having coefficients that do not explain a sufficient amount of variance in the data
- It also has a tendency to set the coefficients of the bad predictors mentioned above 0
- This makes Lasso useful in feature selection

## Disadvantages

- Lasso however struggles with some types of data. If the number of predictors (p) > (n)
- Lasso will also struggle with collinear features (they're related/correlated strongly)
- This selection will also be done in a random way, which is bad for reproducibility and interpretation

# Lasso Regression

$Y = 317.12 + 0*X - 24.21*X^2 + 2.04*X^3 + 0.02*X^4 + 0*X^5$




**R2 score: 0.834**

- The overfitting has reduced significantly which can be seen from the regression curve as well. The model estimates were shrunk to 0 to limit overfitting and improve the generalization capability of the model (increase the bias of the model).
- The model estimates for X and X5 were pushed to 0 indicating that these are not the most important feature. This also increases the adjusted R2 score of the model because we have eliminated 2 more features.

# Code - Lasso Regression

$Y = 317.12 + 0*X - 24.21*X^2 + 2.04*X^3 + 0.02*X^4 + 0*X^5$

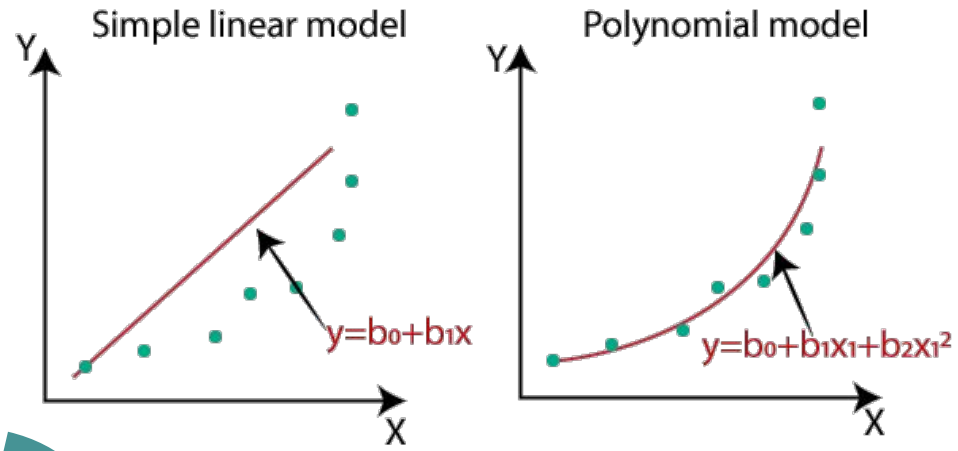
```
# Lasso Regression
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Lasso
lasso = Lasso()

parameters = {"alpha":[1e-15, 1e-10, 1e-8, 1e-4, 1e-3, 1e-2, 1, 5, 10, 20]}
lasso_regression = GridSearchCV(lasso, parameters,
scoring='neg_mean_squared_error', cv=5)
lasso_regression.fit(xs, y)
```

```
print(lasso_regression.best_params_)
print(lasso_regression.best_score_)
```

# Polynomial Regression

In statistics, **polynomial regression** is a form of regression analysis in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *n*th degree polynomial in *x*. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding conditional mean of *y*, denoted E(*y* |*x*).

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$




Simple Linear Regression: $y = b_0 + b_1 x_1$

Multiple Linear Regression: $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

Polynomial Linear Regression: $y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Polynomial regression is considered to be a special case of **multiple linear regression**

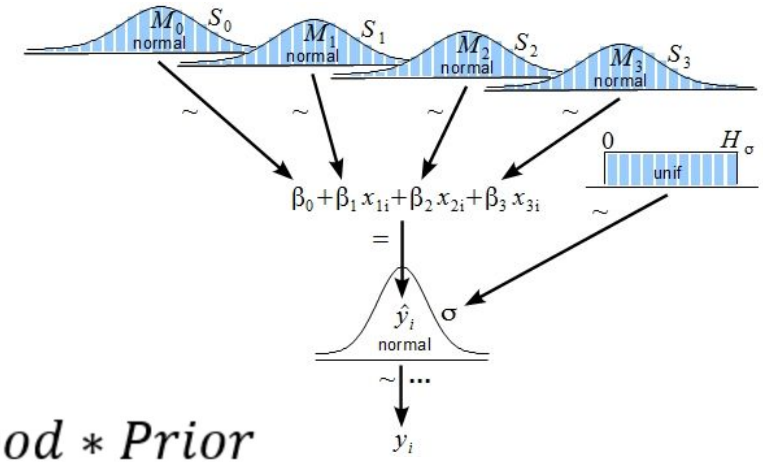
# Bayesian Linear Regression

**Bayesian linear regression** is an approach to linear regression in which the statistical analysis is undertaken within the context of **Bayesian inference**. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

$$y \sim N(\beta^T X, \sigma^2 I)$$

$$P(\beta|y,X) = \frac{P(y|\beta,X) * P(\beta|X)}{P(y|X)}$$

$$Posterior = \frac{Likelihood * Prior}{Normalization}$$



# Case Study Example

The table below shows some data from the early days of the Italian clothing company Benetton. Each row in the table shows Benetton's sales for a year and the amount spent on advertising that year. In this case, our outcome of interest is sales—it is what we want to predict. If we use advertising as the predictor variable, linear regression estimates that **Sales = 168 + 23 Advertising**. That is, if advertising expenditure is increased by one million Euro, then sales will be expected to increase by 23 million Euros, and if there was no advertising we would expect sales of 168 million Euros.

| Year | Sales (Million Euro) | Advertising (Million Euro) |
|---|---|---|
| 1 | 651 | 23 |
| 2 | 762 | 26 |
| 3 | 856 | 30 |
| 4 | 1,063 | 34 |
| 5 | 1,190 | 43 |
| 6 | 1,298 | 48 |
| 7 | 1,421 | 52 |
| 8 | 1,440 | 57 |
| 9 | 1,518 | 58 |

- We can include **year** variable in the regression, which gives the result that **Sales = 323 + 14 Advertising + 47 Year**
- The interpretation of this equation is that every extra million Euro of advertising expenditure will lead to an extra 14 million Euro of sales and that sales will grow due to non-advertising factors by 47 million Euro per year

# Case Study Example - Continued

## Linear Regression: Sales

| | Estimate | Standard Error | t | p |
|---|---|---|---|---|
| (Intercept) | 167.68 | 58.94 | 2.85 | .025 |
| Advertising | 23.42 | 1.37 | 17.13 | < .001 |

*n = 9 cases used in estimation; R-squared: 0.9767; Correct predictions: 88.89%; AIC: 100.34; multiple comparisons correction: None*

The column labelled **Estimate** shows the values used in the equations before. These estimates are also known as the *coefficients* and *parameters.* The **Standard Error** column quantifies the uncertainty of the estimates. The standard error for Advertising is relatively small compared to the Estimate, which tells us that the Estimate is quite precise, as is also indicated by the high ***t*** *(*which is **Estimate / Standard**), and the small ***p***-value.
Furthermore, the R-Squared statistic of 0.98 is very high, suggesting it is a good model.

A key assumption of linear regression is that all the relevant variables are included in the analysis. We can see the importance of this assumption by looking at what happens when **Year** is included. Not only has Advertising become much less important (with its coefficient reduced from 23 to 14), but the standard error has ballooned. The coefficient is no longer statistically significant (i.e., the *p*-value of 0.22 is above the standard cutoff of .05). This means is that although the estimate of the effect of advertising is 14, we cannot be confident that the true effect is not zero.

## Linear Regression: Sales

| | Estimate | Standard Error | t | p |
|---|---|---|---|---|
| (Intercept) | 323.54 | 177.60 | 1.82 | .118 |
| Advertising | 13.99 | 10.22 | 1.37 | .220 |
| Year | 46.60 | 50.03 | 0.93 | .388 |

*n = 9 cases used in estimation; R-squared: 0.9797; Correct predictions: 88.89%; AIC: 101.13; multiple comparisons correction: None*

# Applications of Regression

## B. Operation Efficiency

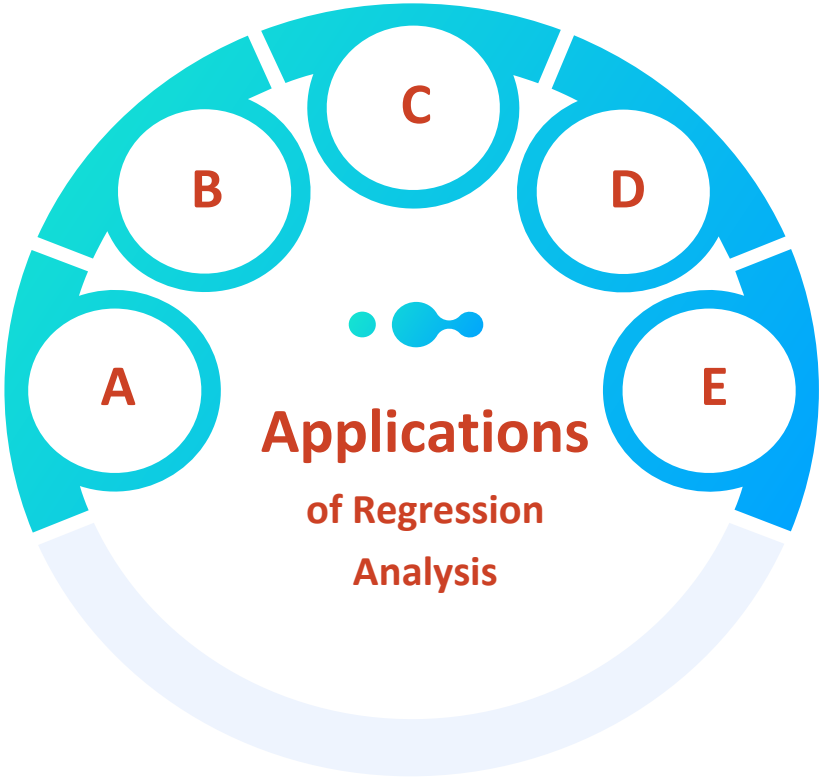
Data-driven decision making eliminates guesswork, hypothesis and corporate politics from decision making. This improves the business performance by highlighting the areas that have the maximum impact on the operational efficiency and revenues.

## A. Supporting Decisions

Predictive analytics i.e. forecasting future opportunities and risks is the most prominent application of regression analysis in business. Demand analysis, for instance, predicts the number of items which a consumer will probably purchase.

## C. Predictive Analytics

This technique acts as a perfect tool to test a hypothesis before diving into execution.




## D. Correcting Errors

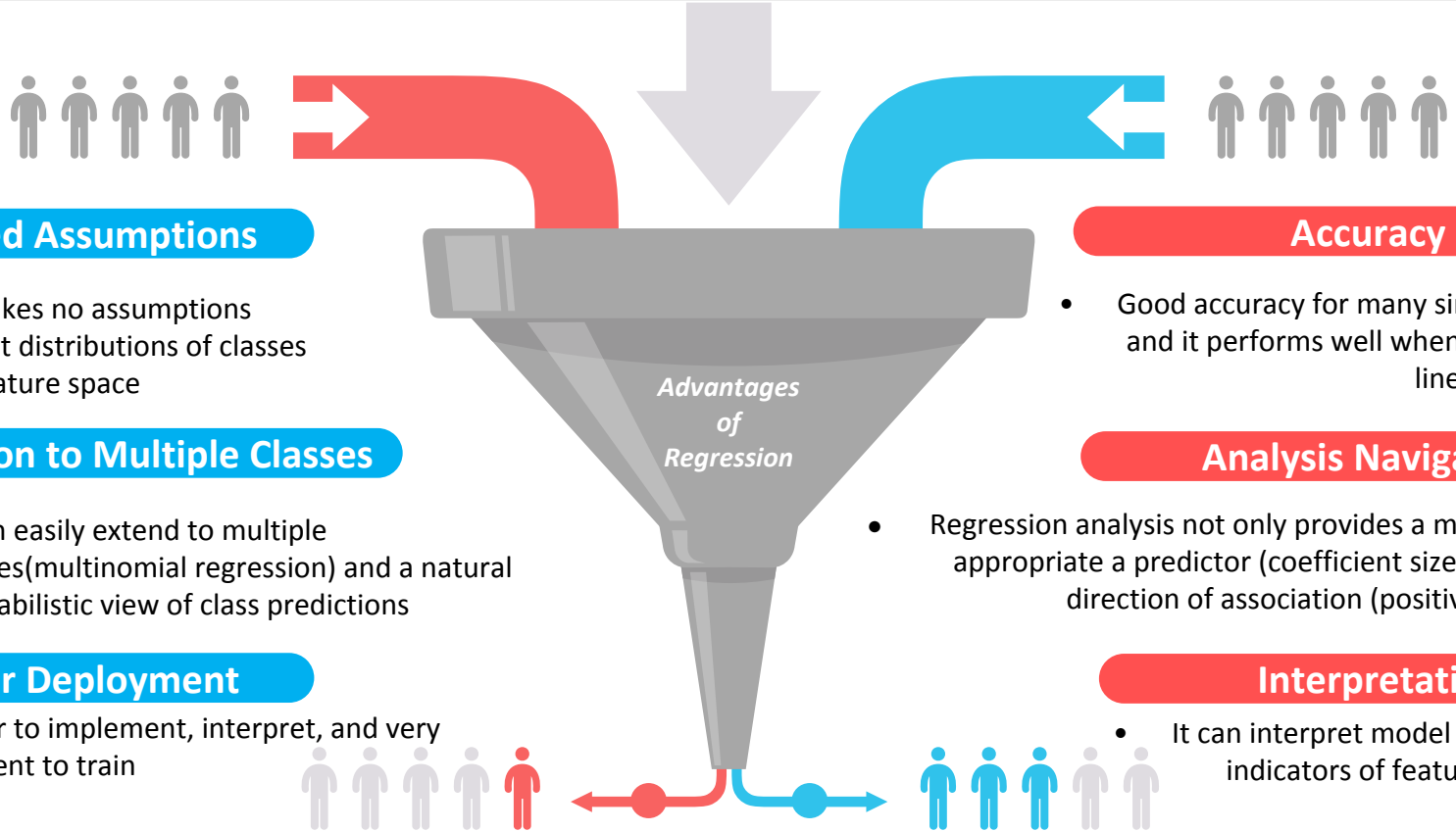
1. Regression is not only great for lending empirical support to management decisions but also for identifying errors in judgment
2. This analysis can provide quantitative support for decisions and prevent mistakes due to manager's intuitions.

## E. New Insights

1. RA techniques can find a relationship between different variables by uncovering patterns that were previously unnoticed.
2. For example, analysis of data from point of sales systems and purchase accounts may highlight market patterns like increase in demand on certain days of the week or at certain times of the year.

# Advantages of Regression Analysis

Several types of Regression have diverse advantages




## Limited Assumptions

- It makes no assumptions about distributions of classes in feature space

## Extension to Multiple Classes

- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions

## Easier Deployment

- Easier to implement, interpret, and very efficient to train

## Accuracy

- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable

## Analysis Navigation

- Regression analysis not only provides a measure of how appropriate a predictor (coefficient size)is, but also its direction of association (positive or negative)

## Interpretation

- It can interpret model coefficients as indicators of feature importance

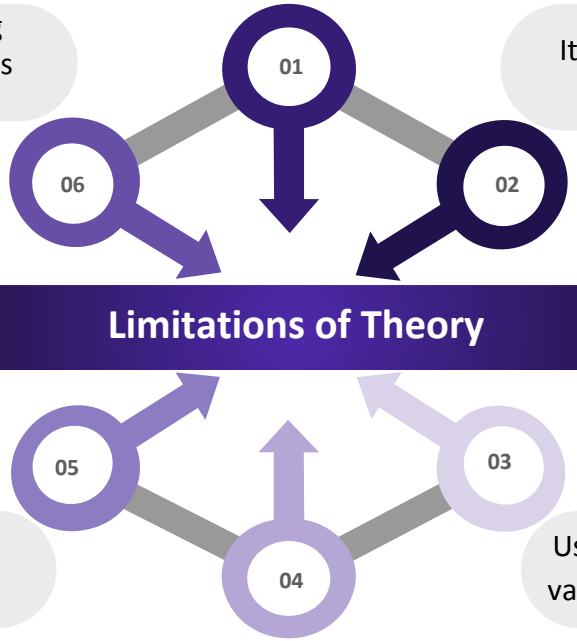
It might look easy for linear relationships but as not everything is the world is linear, more complex regressions take the stage which is so much more complex.

Tough to obtain complex relationships using regression. More powerful algorithms such as Neural Networks can easily outperform.

**Complex Relationships**

It stands no chance against outliers are extreme points. It can cause to overfit very easily.

**Outliers and Overfitting**




**Multicollinearity**

Requires average or no multicollinearity between independent variables.

**Bound on Dependant Variable**

Used to predict discrete functions. Dependent variable of Regression is bound to discrete set.

Fine-Tuning is difficult. Once assumptions are violated they need to be corrected before fine-tuning the model as it becomes unproductive very quickly.

# Ways to detect Overfitting

- The easiest way to avoid overfitting is to **increase sample size** by collecting more data
- If data size is limited, the second option is to reduce the number of predictors in our model — either by combining or eliminating them.
- **Factor Analysis** is one method to identify related predictors that might be candidates for combining.

**1. Cross-Validation**
Use **cross validation** to detect overfitting: this partitions our data, generalizes our model, and chooses the model which works best. One form of cross-validation is **predicted R-squared**. Most good statistical software will include this statistic, which is calculated by:

- **Removing one observation** at a time from the data
- **Estimating the regression equation** for each iteration
- Using the regression equation **to predict the removed observation**

**2. Shrinkage & Resampling**
Shrinkage and resampling techniques (like this R-module) can help you to find out how well your model might fit a new sample

**3. Automated Methods**
**Automated stepwise regression shouldn't be used** as an overfitting solution for small data sets

# Summary

- The presentation covers theoretical aspects on different types of Regression, parameter estimation applications of regression, advantages and disadvantages, code execution and case study example.
- **Linear regression** can be used to predict a dependent variable from an independent one. The regression line allows the estimation of a response variable for individuals with values of the carrier variable not included in the data.
- **Logistic regression** provides a useful means for modelling the dependence of a binary response variable on one or more explanatory variables
- **Ridge & Lasso**: Regularization is a very important concept that is used to avoid overfitting of the data especially when the trained and tested data are much varying.
- **Polynomial Regression** is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.
- **Parameter Estimation:** Direct Regression, Reverse Regression, Major axis Regression, Reduced Major Axis Regression were discussed

# References


1. Gonick, L. (1993). The Cartoon Guide to Statistics. HarperPerennial.
2. Lindstrom, D. (2010). Schaum's Easy Outline of Statistics, Second Edition (Schaum's Easy Outlines) 2nd Edition. McGraw-Hill Education
3. Babyak, M.A.,(2004). "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models." Psychosomatic Medicine. 2004 May-Jun;66(3):411-21.
4. Green S.B., (1991) "How many subjects does it take to do a regression analysis?" *Multivariate Behavioral Research* 26:499–510.
5. Peduzzi P.N., et. al (1995). "The importance of events per independent variable in multivariable analysis, II: accuracy and precision of regression estimates." *Journal of Clinical Epidemiology* 48:1503–10.
6. Peduzzi P.N., et. al (1996). "A simulation study of the number of events per variable in logistic regression analysis." *Journal of Clinical Epidemiology* 49:1373–9.
7. Adichie, J. N. [1967], "Estimates of regression parameters based on rank tests," *Ann. Math. Stat.*, 38, 894–904.
8. Aitkin, M. A. [1974], "Simultaneous inference and the choice of variable subsets," *Technometrics* 16, 221–227.
9. Akaike, H. [1973], "Information theory and an extension of the maximum likelihood principle," in B. N. Petrov and F. Csaki (editors), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado.
10. Allen, D. M. [1971], "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, 13, 469–475.
11. Andrews, D. F. [1974], "A robust method for multiple linear regression," *Technometrics*, 16, 523–531.



## *Acknowledgement*

We would like to thank **Prof. Kannan A.** for conducting this course **CH5020 Statistical Design and Analysis of Experiments** and giving us this opportunity to present internal report presentation. The lectures throughout the semester helped us understand and build a strong grasp on the subject. We'd also like to thank our Teaching Assistants and fellow coursemates who helped us throughout this entire course.


# THANK YOU