CHAPTER 1
Introduction to Statistical Modelling

Noryanti Muhammad
Centre for Mathematical Sciences
College of Computing and Applied Sciences
Universiti Malaysia Pahang

Centre of Excellence (CoE) for Data Science & Artificial Intelligence
Research & Innovation Department
Universiti Malaysia Pahang

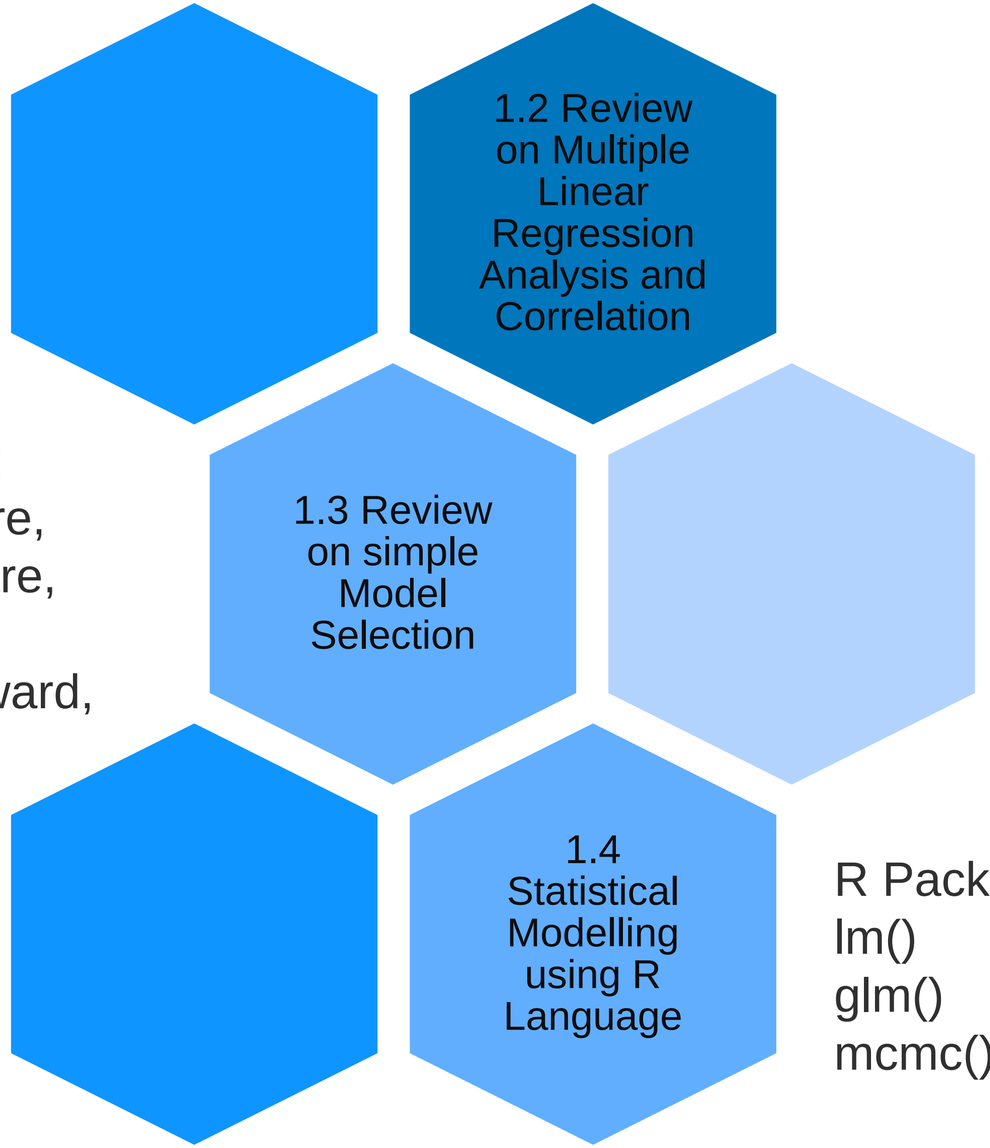
Expected Outcomes:

By the end of this chapter, students should be able:

- ✓ To understand and identify the process in statistical modelling
- ✓ To recall and review on Linear Regression Analysis and Correlation (Simple Linear Model, Multiple Linear Model)
- ✓ To use R Language in analyse the data in the statistical modelling

Content:

Content



Model Selection:
P-value, R square, Adjusted R square,

Stepwise, Backward, Enter

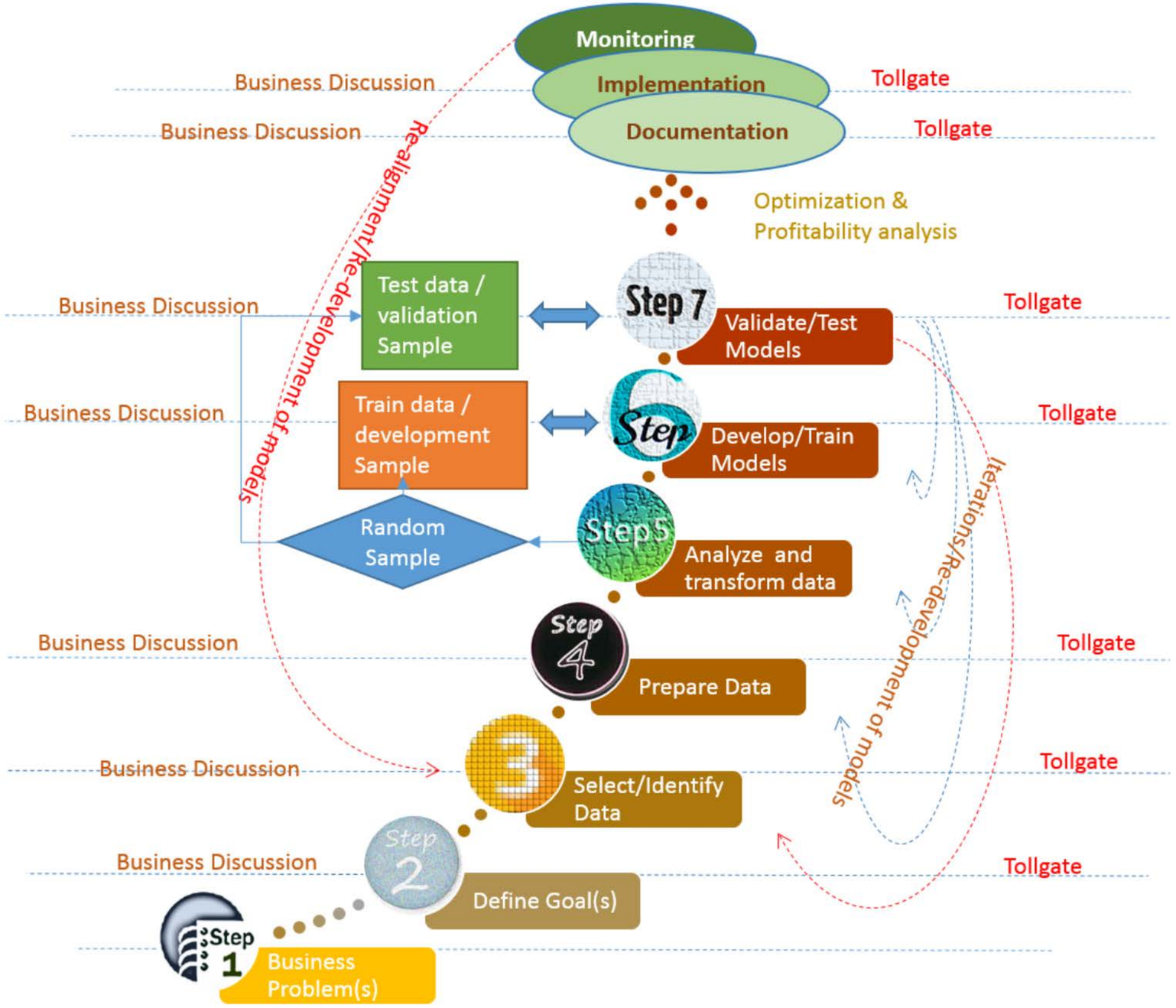
R Package:
lm()
glm()
mcmc()

1.1 Overview of process in statistical modelling

- Statistical modeling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data. ... “When you analyze data, you are looking for patterns,” says Mello. “You are using a sample to make an inference about the whole.”

- Generally, 7-Steps Predictive Modeling Process

7-Steps Predictive Modelling Process



Step 1: Business Objective(s)

- Target Marketing
- Risk & Fraud Management
- Strategy Implementation and Change Management
- Operational Efficiency
- Increase Customer Experience
- Manage Marketing Campaigns
- Forecast Revenue or Loss
- Workforce Management
- Financial Modeling
- Churn Management
- Social Media Influencers

Step 1.1: Business Objectives - Asking Right Questions!

It is very important to clearly define the goals based on business objective.

- Do you want to understand the characteristics of customers?
- Do you want to make unprofitable customers profitable?
- Do you want to understand what driving sales?
- Do you want to win-back lost customers?
- Do you want to increase sales?
- Do you want to reduce customer churn?
- Do you want to reduce cost of production or operation?
- Do you want to target new customers?
- Do you want to identify probable credit default customers?
- Do you want to know X-sell/Upsell opportunities?
- Businesses want to find answers all such important questions and make decisions based on data…

Step 1.2: Business Objective(s) - Target Modeling Opportunities

Industry	Response	Risk Mitigation	Attrition	Cross-Sell/Upsell	Net Present Value	Life Time Value
Retail	X		X	X	X	X
Banking	X	X	X	X	X	X
Insurance	X	X	X	X	X	X
Telecom	X	X	X	X	X	X
Utilities	X	X	X	X	X	X
Hospitality	X		X	X	X	X
Catalog	X			X	X	X
Publishing	X		X	X	X	X

Step 2: Define Goals - translate business objective into analytics goal

Based on the business questions we want to answer, translate the business objective into Analytic terms

- Profile Analysis
- Segmentations
- Response Modeling
- Risk Modeling
- Activation
- Cross-Sell and Upsell
- Attrition/Churn Modeling
- Net Present Value(NPV)
- Customer Lifetime Value (CLTV)
- etc.

Step 3: Selecting Data for Modeling

Selecting best data for target modeling requires thorough understanding of the market, business and the objective. The model is only as good and relevant as the underlying data:

Data Types

Data Type	Predictive Power	Stability	Cost
Demographic	Medium	High	Low
Behavioural	High	Low	High
Psychographic	Medium	Medium	High

Sources of Data

Internal Sources	External Sources
Customer Data, Transaction Data	Survey Data, Research Data, Suppliers, Ratings
Other History	Credit Bureau Data, Third Party data, Sellers, Compilers

Step3.1: A Case Study - Target Marketing

Typical data required for Target Marketing

Demographic Data	Behaviour Data	External Data
Customer Demographic	Transaction	Customer Survey
Income	Loyalty	Market Research
Purchasing Power	-	Macro Economic Factors
-	-	Competitions

Step 4: Prepare Data

- In this step we need prepare data into right format for analysis and the tool you may want use.
- Do initial cleaning up
- Define Variables and Create Data Dictionary
- Joining/Appending multiple datasets
- Validate for correctness
- Produce Basic Summary Reports

Step 5: Analyse and Transform Variables

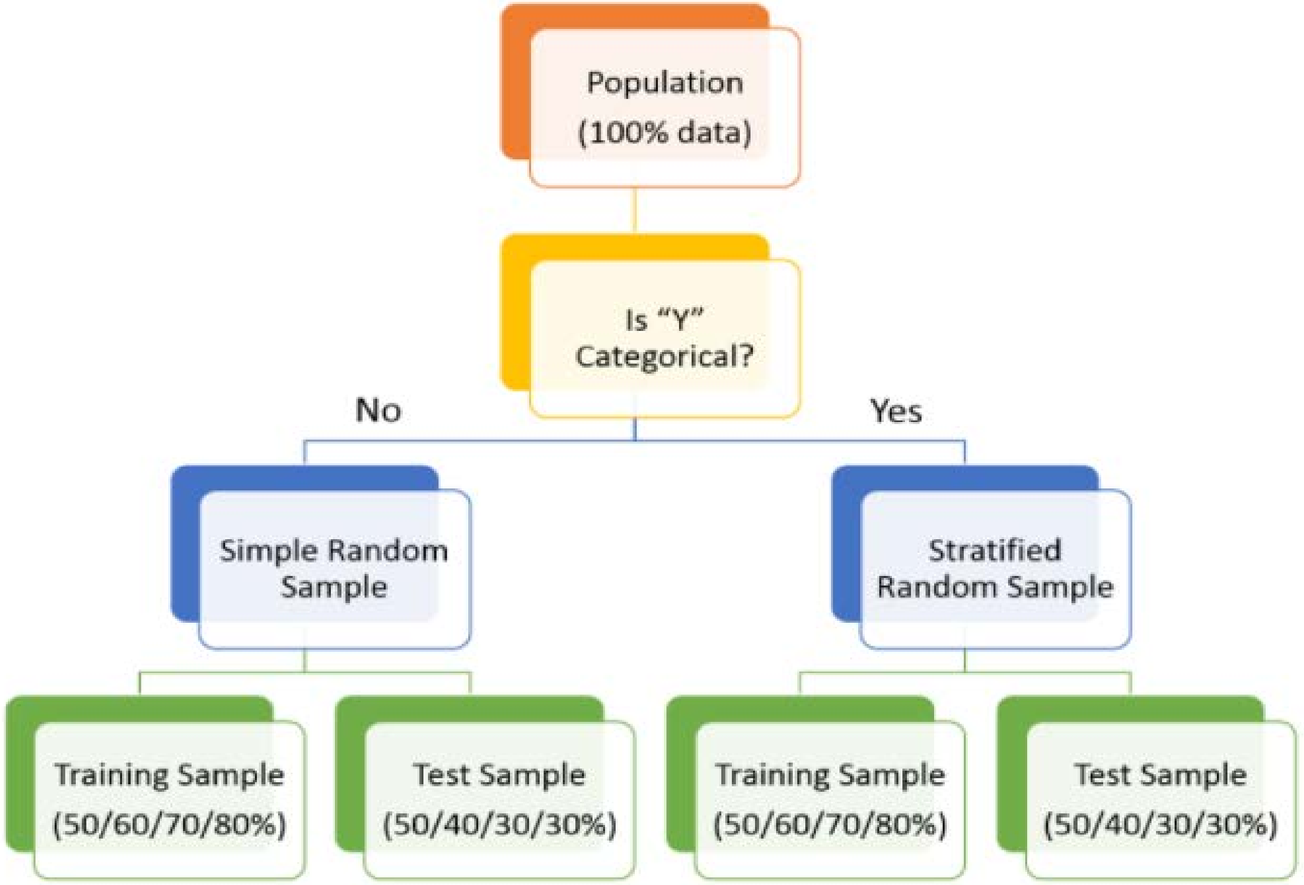
Once data is in right shape and perform

- univariate analysis: to check the distribution of each of the variables and features
- multivariate analyses: to check relationships with other variables and with dependent variables

Based on type of model you are going to use; you may need to transform the variables using one of the approaches

- Bining approach: create distinct groups
- Transformation:
- Logarithmic, Polynomial
- Square Root, Inverse, Square, boxCox
- Extreme value (outlier) treatments
- Missing Value Treatment
- Dimension Reduction - Information Value(IV) and Weight of Evidence(WoE), Variable Clustering, PCA, Factor Analysis, etc.

Step 5.1: Random Sampling (Train and Test)



- **Training Sample**: Model will be developed on this sample. Typically, 50%, 60%, 70% or 80% of the data goes here.
- **Test Sample**: Model performances will be validated on this sample. Typically, 50%, 40%, 30% or 20% of the data goes here

Step 6: Model Selection

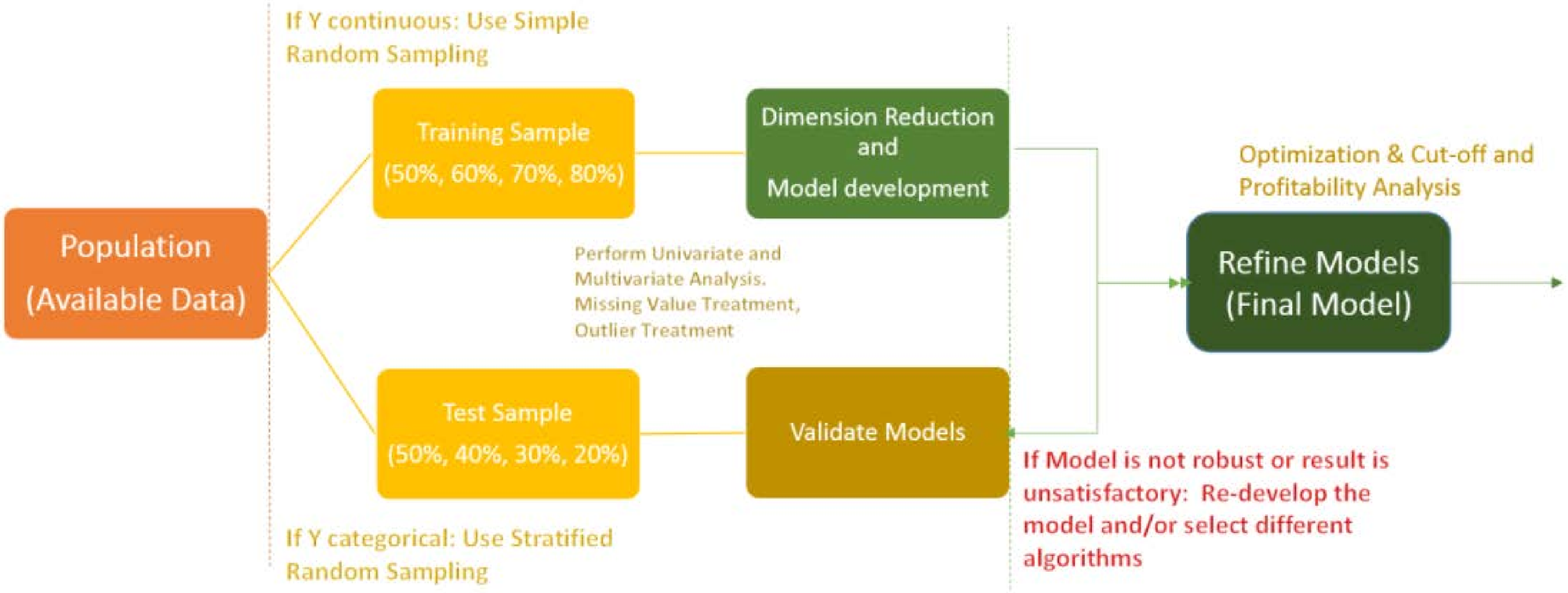
Based on the defined goal(s) (supervised or unsupervised) we have to select one of or combinations of modeling techniques. Such as

- General linear model
- Non-Linear Regression
- Linear Regression
- Lasso Regression
- Ridge Regression
- Non-Negative Garrotte Regression
- Percentage Regression
- Quantile Regression
- Non-parametric regression
- Logistic Regression
- Tobit Regression
- Probit Regression
- Classification/Decision Trees
- Random Forest
- Support Vector Machine (SVM)
- Distance metric learning
- Bayesian methods
- Graphical Models
- Neural Networks
- Genetic Algorithm
- The Hazard and Survival Functions
- Time Series Models
- Signal Processing
- Clustering Techniques
- Market Basket Analysis
- Frequent Itemset Mining
- Association Rule Mining etc.

There are wide variety of choices available outside this list.

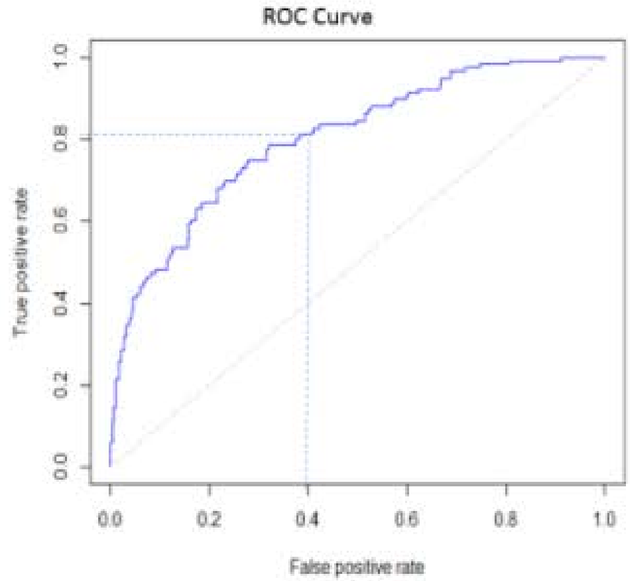
Step 6.1: Build/Develop/Train Models

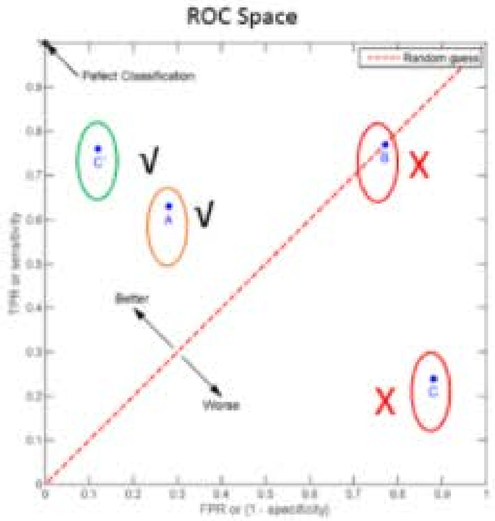
- Validate the assumptions of the chosen algorithm
- Check for Multicollinearity and Redundancies of Independent Variables (Features). Sometime in Machine Learning, we are keen on accuracies of the models and hence we may not perform these checks!
- Develop/Train Model on Training Sample, which is 80%/70%/60%/50% of the available data(Population)
- Check Model performance - Error, Accuracy, ROC, KS, Gini

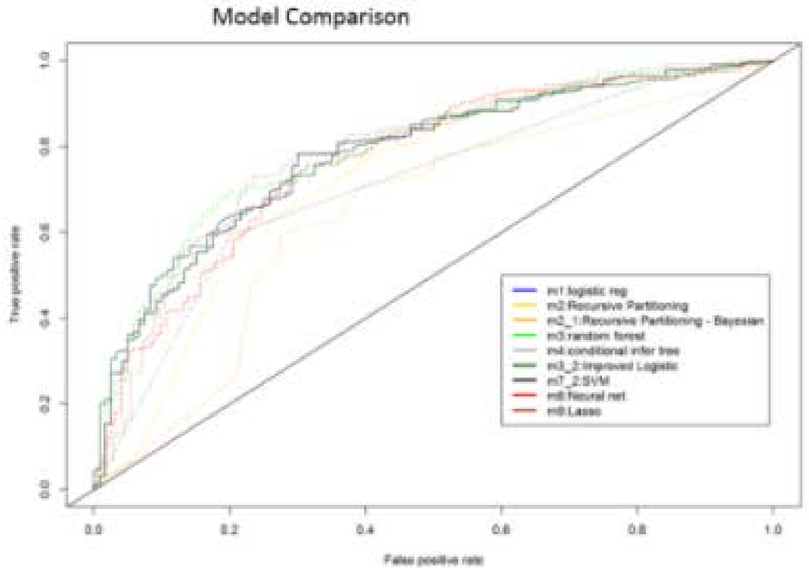


Step 7: Validate/Test Models

- Score and Predict using Test Sample
- Check for the robustness and stability of the model
- Check Model Performance: Accuracy, ROC, AUC, KS, GINI etc.







Perform Cross Validation to increase accuracy/performance of the models

General Problem

- **Given**: a collection of variables, each variable being a vector of readings of a specific trait on the samples in an experiment.
- **Problem**: In what way does a variable Y depend on other variables $X_1, \ldots, X_n$ in the study.
- Explanation: A statistical model defines a mathematical relationship between the X_i's and Y . The model is a representation of the real Y that aims to replace it as far as possible. At least the model should capture the dependence of Y on the X_i's.

The Types of Variables

in a statistical model

The **response variable** is the one whose content we are trying to model with other variables, called the **explanatory variables**.

In any given model there is one response variable (Y) and there may be many explanatory variables (X_1, ... X_n).

Identify and Characterize Variables

the first step in modelling

- Which variable is the ***response variable***;
- Which variables are the ***explanatory variables***;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable — is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Other Type of Variables

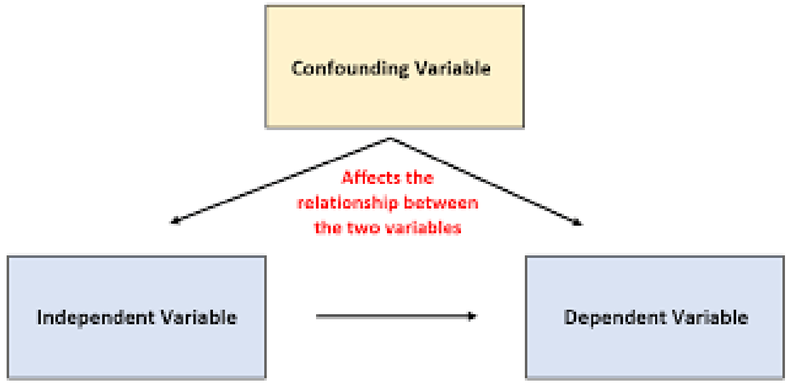
Which could be considered

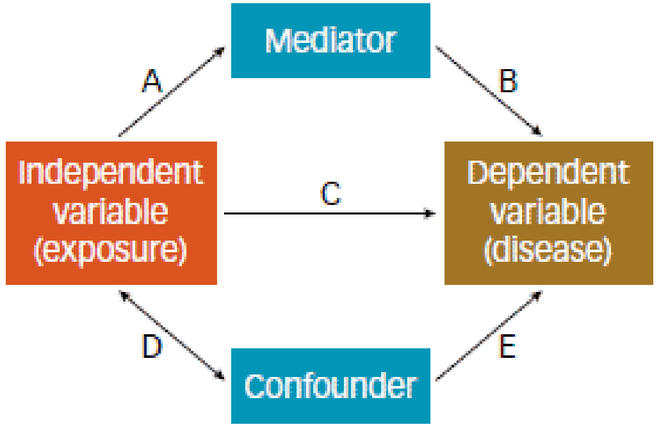
Confounder Variable:
is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations

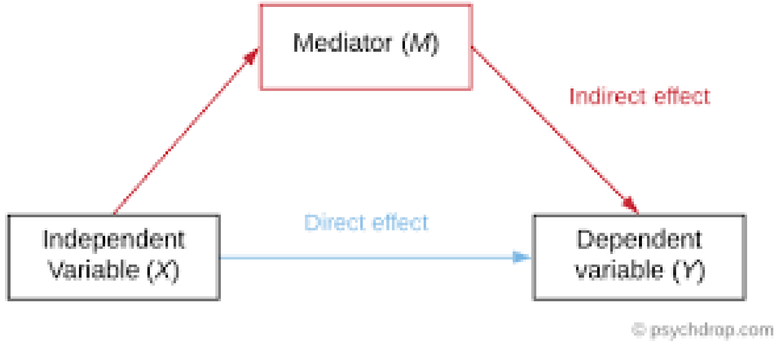
A confounder is a third variable that affects variables of interest and makes them seem related when they are not.

In contrast, a mediator is the mechanism of a relationship between two variables: it explains the process by which they are related.

A **mediator variable** explains the how or why of an (observed) relationship between an independent variable and its dependent variable

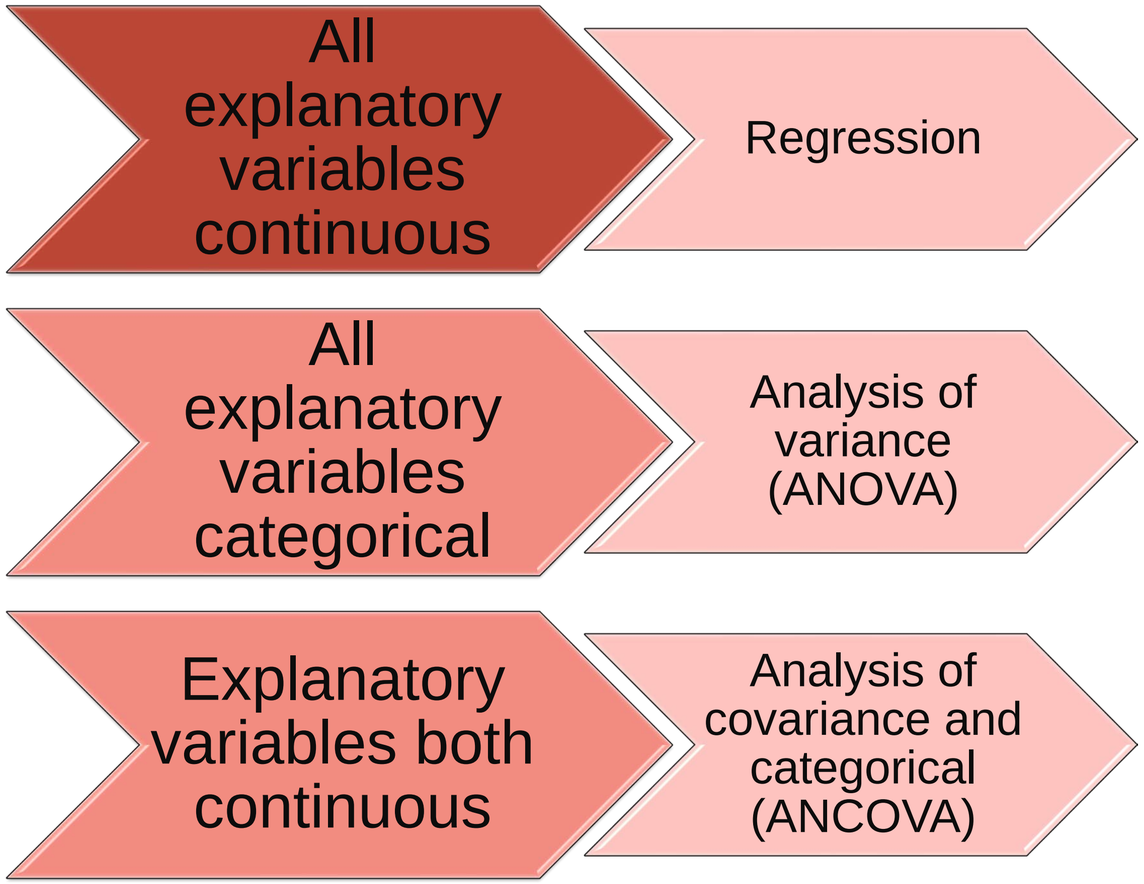






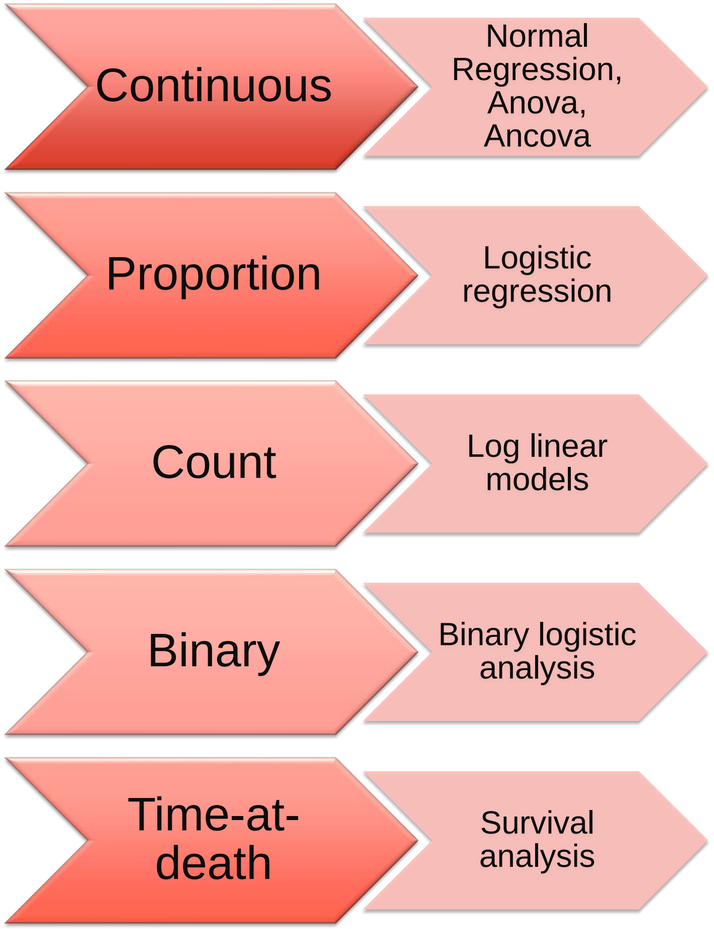
Types of Variables Determine Type of Model

The explanatory variables



Types of Variables Determine Type of Model

The response variable — what kind of data is it?



1.2 Review on Linear Regression Analysis and Correlation

Simple Linear Regression

relationship between a dependent variable (y) and an independent variable (x)

Multiple Linear Regression

relationship between a dependent variable (y) and multiple independent variables ($x_1, \ldots, x_k$)

Simple Linear Model

- ✓Plot a **scatter diagram**.
- ✓Find and interpret the **correlation coefficient** value and **coefficient of determination** value.
- ✓Describe **linear relationship** involving one dependent variable with one independent variable using simple linear regression model.
- ✓Conduct **hypothesis testing** for simple linear regression model.
- ✓Make a **prediction** using simple linear regression model.

Simple Linear Model

- Model $y = \beta_0 + \beta_1 x + \varepsilon$

β_0 = Intercept and β_1 = Slope

- The estimated model

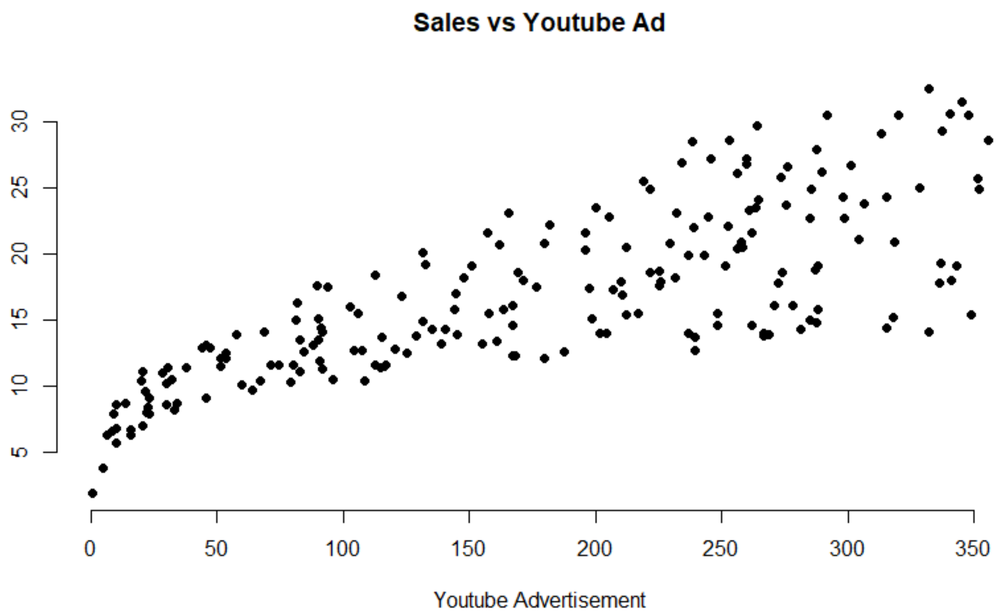
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Example 1

Data Description

- It contains the impact of three advertising medias (youtube, facebook and newspaper) on sales.
- Data are the advertising budget in thousands of dollars along with the sales.
- The advertising experiment has been repeated 200 times with different budgets and the observed sales have been recorded.

Scatter Plot



Interpretation:
There is a positive linear relationship between sales and Youtube advertisement budget allocate.

Simple Coding

```
#variables
x <- marketing$youtube
y <- marketing$sales
plot(marketing$sales, marketing$youtube)

# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Sales vs Youtube Advertisement",
     xlab = "Youtube Ad", ylab = "Sales",
     pch = 19, frame = FALSE)
```

Alternative

```
install.packages("car")
library(car)
scatterplot(sales ~ youtube, data = marketing)

# Suppress the smoother and frame
scatterplot(sales ~ youtube, data = marketing,
            smoother = FALSE, grid = FALSE, frame = FALSE)
```

Correlation Coefficient and Coefficient of Determination Values

```
> #correlation
> cor(marketing$sales, marketing$youtube)
[1] 0.7822244
>
```

r=0.7822

Interpretation:
There is a strong positive linear relationship between sales and Youtube advertisement budget allocation.

```
> cor(marketing$sales, marketing$youtube)^2
[1] 0.6118751
> cor
```

61.19% of variation in the sales is explained/predictable by the Youtube advertisement budget allocation. While 38.81% of variation in sales is explained by other factors such as Facebook advertisement etc.

Simple Coding

```
#correlation
cor(marketing$sales,marketing$youtube)
```

Model

Using R

```
#Model
model <- lm(sales ~ youtube, data = marketing)
#Model Summary
summary(model)
```

```
Call:
lm(formula = sales ~ youtube, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0632  -2.3454  -0.2295   2.4805   8.6548

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.439112   0.549412   15.36   <2e-16 ***
youtube     0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.6119,   Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

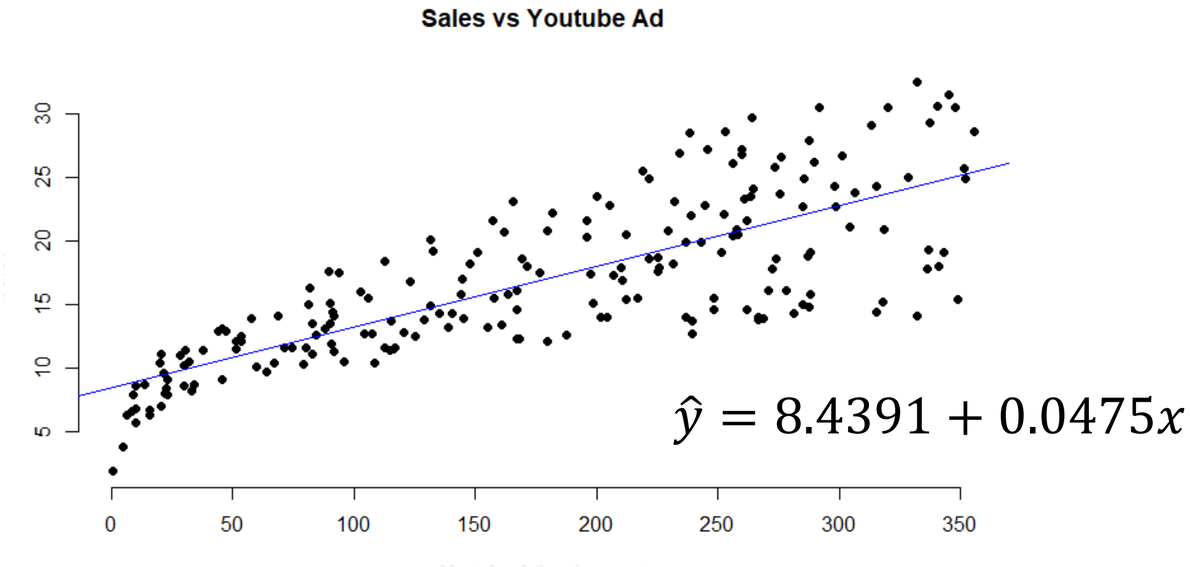
The summary outputs shows 6 components, including:

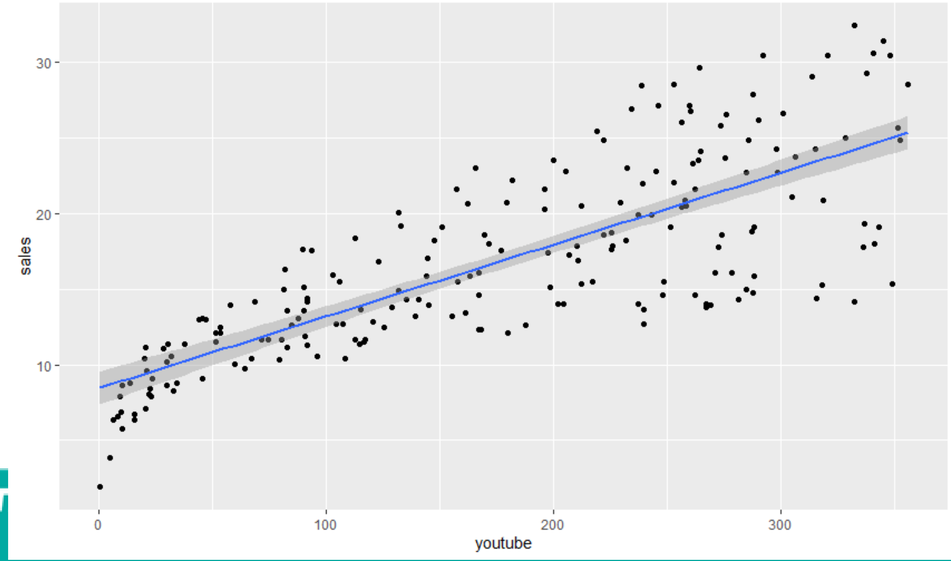
- **Call**. Shows the function call used to compute the regression model.
- **Residuals**. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.
- **Coefficients**. Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
- **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

The coefficients table, in the model statistical summary, shows:

- the estimates of the **beta coefficients**
- the **standard errors** (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
- the **t-statistic** and the associated **p-value**, which defines the statistical significance of the beta coefficients.

Best Fitted Line





Simple Coding

```
#variables
x <- marketing$youtube
y <- marketing$sales
plot(marketing$sales, marketing$youtube)

# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Sales vs Youtube Advertisement",
     xlab = "Youtube Ad", ylab = "Sales",
     pch = 19, frame = FALSE)
abline(lm(y ~ x, data = mtcars), col = "blue")
```

Alternative

```
library(ggplot2)
#Regression line
#By default, the fitted line is presented with confidence interval around it OR
stat_smooth(method = lm, se=FALSE).
ggplot(marketing, aes(youtube, sales)) + geom_point() + stat_smooth(method = lm)
```

Model Adequacy Checking (2.2)

Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.

Normality of residuals**.** The residual errors are assumed to be normally distributed.

Homogeneity of residuals variance. The residuals are assumed to have a constant variance (**homoscedasticity**)

Independence of residuals error terms.

Residual Analysis (2.2.1)

Residual defined by, $e_i = y_i - \hat{y}_i$

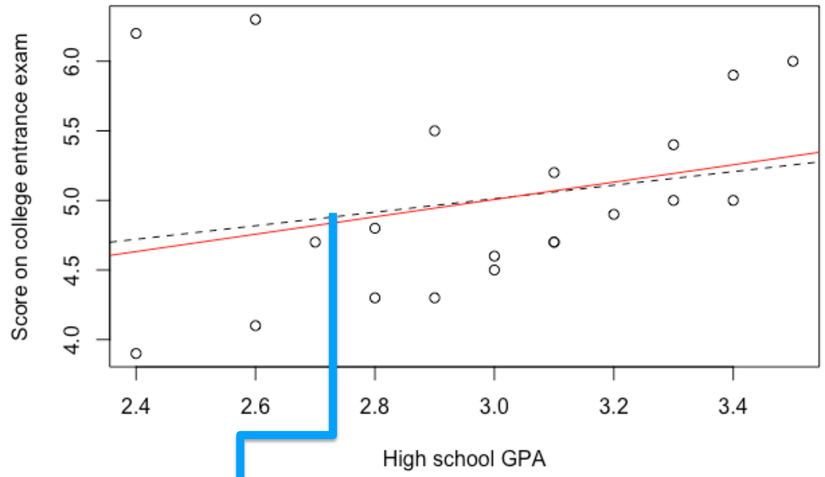
Check by do the residual plot and normal probability plot of the residual.

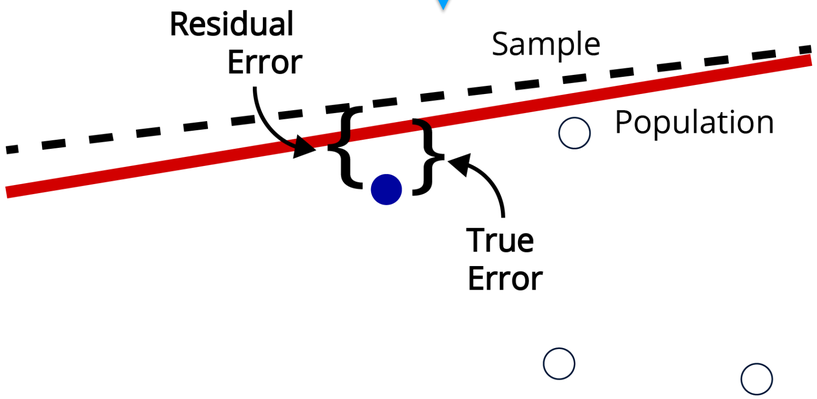
Standardized residual

$$d_i = \frac{e_i}{\sqrt{MSE}}$$

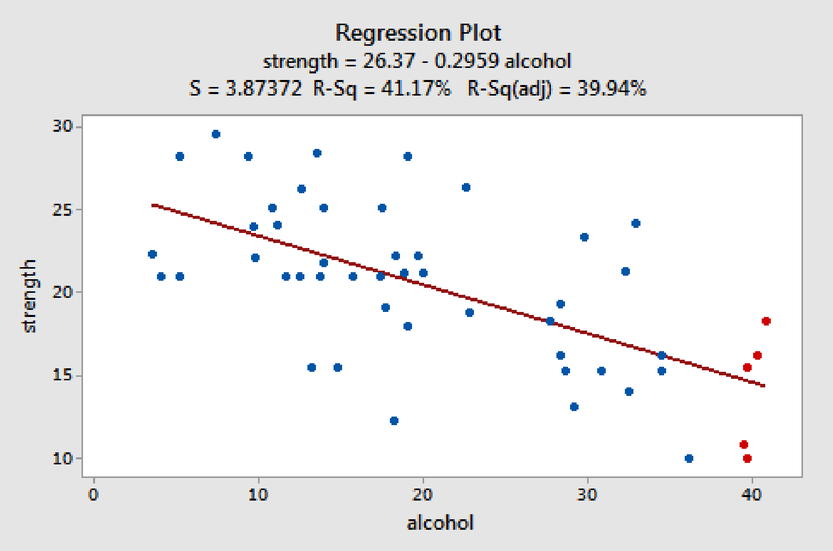
Analyse the residuals to see if they support the assumptions of linearity, independence, normality, and equal variances.

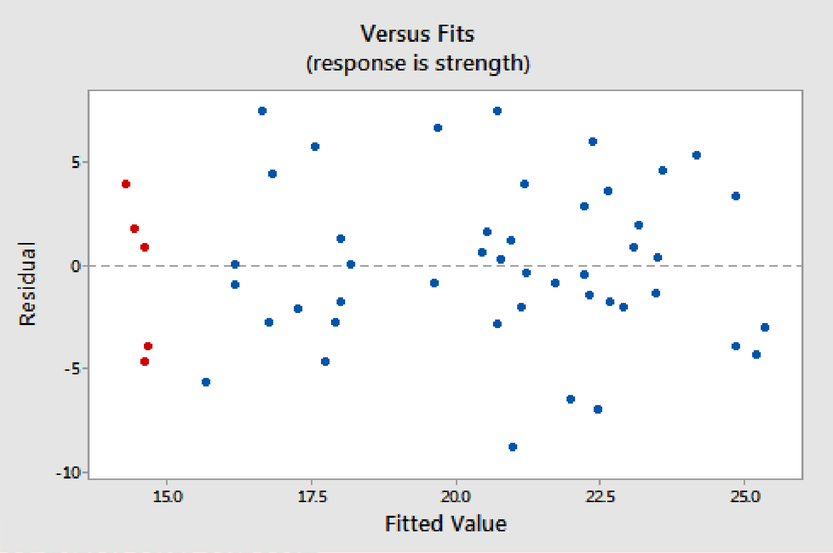
Large residual may indicate possible outliers or unusual observation.





Residuals vs. Fits Plot

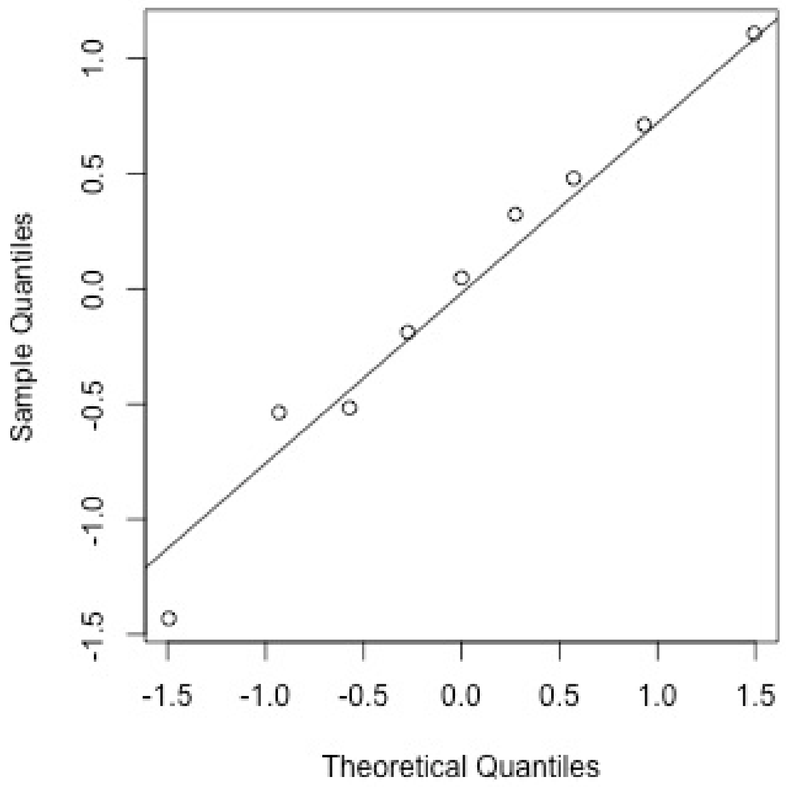




The plot is a classical example of a well-behaved residuals vs. fits plot. The characteristics of a well-behaved residual vs. fits plot and the appropriateness of the simple linear regression model:

- The residuals "bounce randomly" around the residual = 0 line. This suggests that the assumption that the relationship is linear is reasonable.
- The residuals roughly form a "horizontal band" around the residual = 0 line. This suggests that the variances of the error terms are equal.
- No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

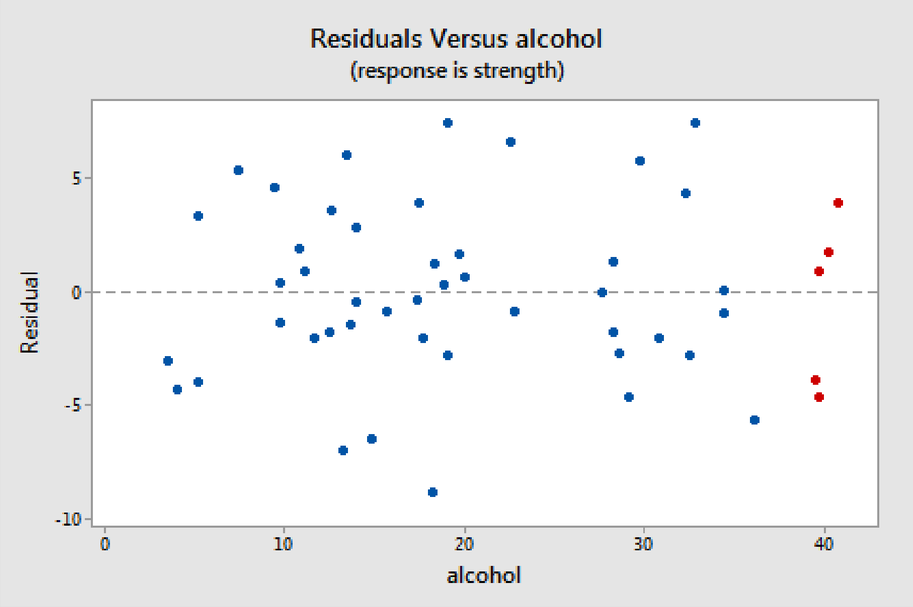
Normal Probability Plot of Residual



We are concerned about the normality of the error terms, we create a normal probability plot of the residuals.

If the resulting plot is approximately linear, we proceed assuming that the error terms are normally distributed.

Residual vs Predictor Plot



The five red data points should help you out again. The alcohol consumption of the five men is about 40, and hence why the points now appear on the "right side" of the plot. In essence, for this example, the residuals vs. predictor plot is just a mirror image of the residuals vs. fits plot. The residuals vs. predictor plot offers no new information.

Exercise Questions

For each question develop the SL model and check the assumptions of linearity, independence, normality, and equal variances.

1. Develop a model between the number of stories a building has and its height. Some people think that as the number of stories increases, the height would increase, but not perfectly. Some statisticians compiled data on a set of n = 60 buildings reported in the 1994 World Almanac (*Consider Building Stories data*).

2. Develop a model between the age of a driver and the distance the driver can see. Some people think that the relationship is negative — as age increases, the distance decreases. A research firm (Last Resource, Inc., Bellefonte, PA) collected data on a sample of n = 30 drivers (*Consider Driver Age and Distance data*).

3. Develop a model between the height of a student and his or her grade point average (GPA). Data were collected on a random sample of n = 35 students in a statistics course at Penn State University (*Consider Height and GPA data*).

Multiple Linear Model

- ✓Plot a **scatter plot matrix.**.
- ✓Find and interpret the **correlation coefficient** value and **coefficient of determination** value.
- ✓Describe **linear relationship** involving one dependent variable with two or more independent variable.
- ✓Conduct **hypothesis testing** for linear regression model.
- ✓Make a **prediction** using simple linear regression model.

Multiple Linear Model

- Model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$

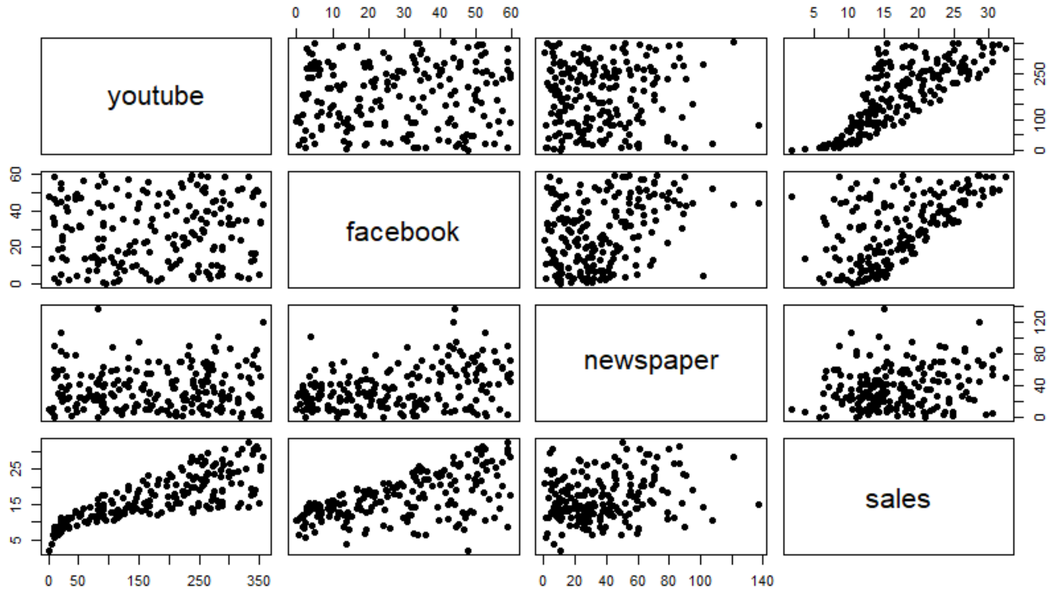
β_0= Intercept and β_i = Slope where $i = 1,2, \ldots, k$

- The estimated model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

Scatter Plot

Interpretation:
There is a positive linear relationship between sales and Youtube advertisement budget allocate.

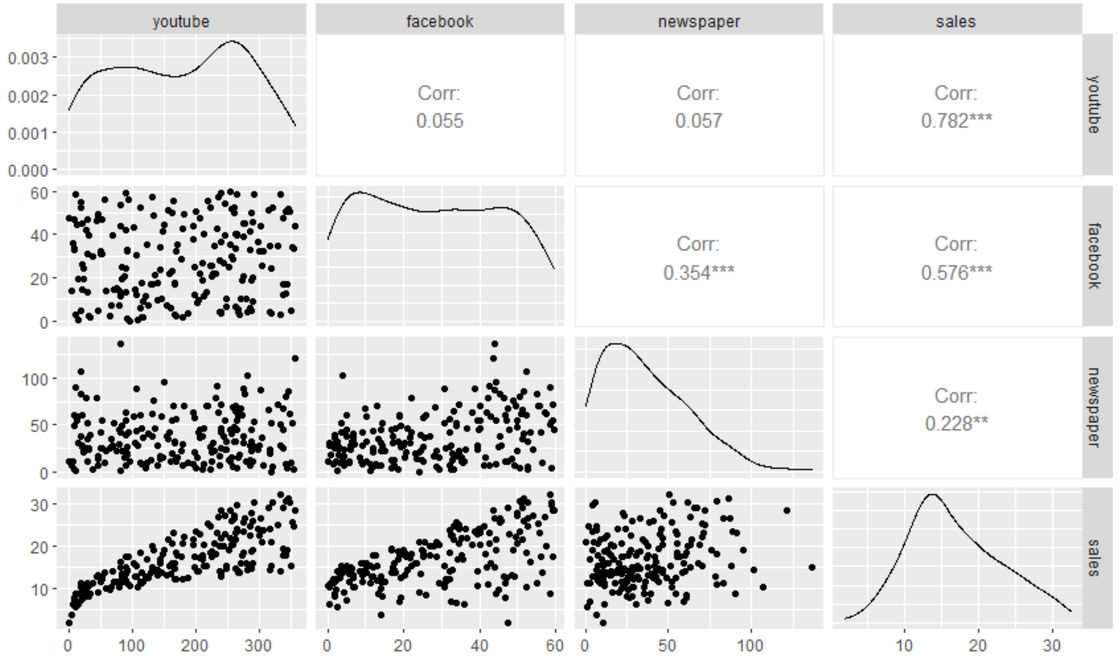



```
#variables
data("marketing", package = "datarium")

# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
pairs(marketing[,1:4], pch = 19)
```

Simple Coding

```
#install and load the GGally library
install.packages("GGally")
library(GGally)

#generate the pairs plot
ggpairs(data)
```

Alternative Coding



Correlation Coefficient and Coefficient of Determination Values

```
> corr <- round(cor(marketing), 1)
> corr
          youtube facebook newspaper sales
youtube       1.0      0.1       0.1   0.8
facebook      0.1      1.0       0.4   0.6
newspaper     0.1      0.4       1.0   0.2
sales         0.8      0.6       0.2   1.0

> corr <- round(cor(marketing), 4)
> corr
          youtube facebook newspaper  sales
youtube    1.0000   0.0548    0.0566 0.7822
facebook   0.0548   1.0000    0.3541 0.5762
newspaper  0.0566   0.3541    1.0000 0.2283
sales      0.7822   0.5762    0.2283 1.0000
>
```

Simple Coding

```
# Compute a correlation matrix
data(marketing)
corr <- round(cor(marketing), 1)
corr <- round(cor(marketing), 4)
```

Example: r=0.5762

Interpretation:
There is a moderate positive linear relationship between sales and Facebook advertisement budget allocation.

Example MLR

Using R

```
Call:
lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5932  -1.0690   0.2902   1.4272   3.3951

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.526667   0.374290   9.422   <2e-16 ***
youtube      0.045765   0.001395  32.809   <2e-16 ***
facebook     0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.8972,   Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

The summary outputs shows 6 components, including:

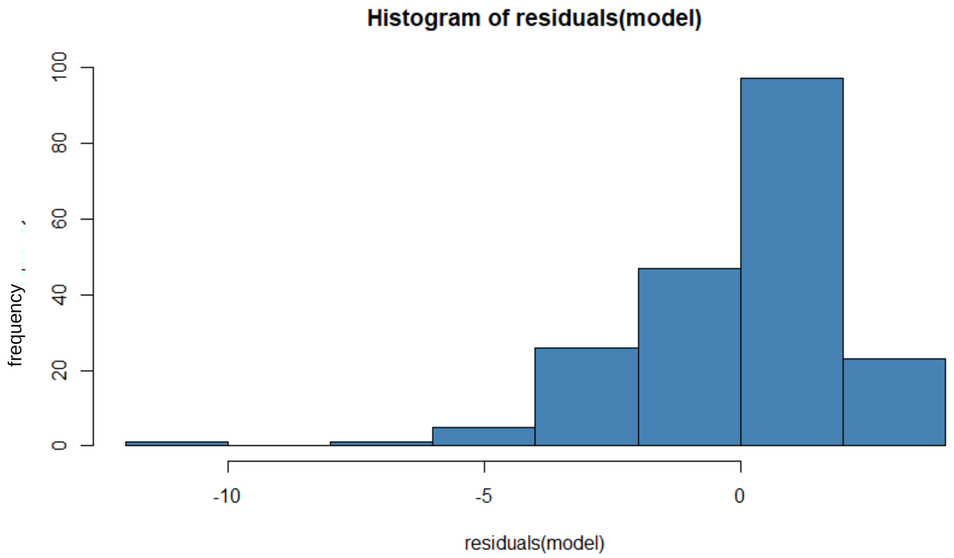
- **Call**. Shows the function call used to compute the regression model.
- **Residuals**. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.
- **Coefficients**. Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
- **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

The coefficients table, in the model statistical summary, shows:

- the estimates of the **beta coefficients**
- the **standard errors** (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
- the **t-statistic** and the associated **p-value**, which defines the statistical significance of the beta coefficients.

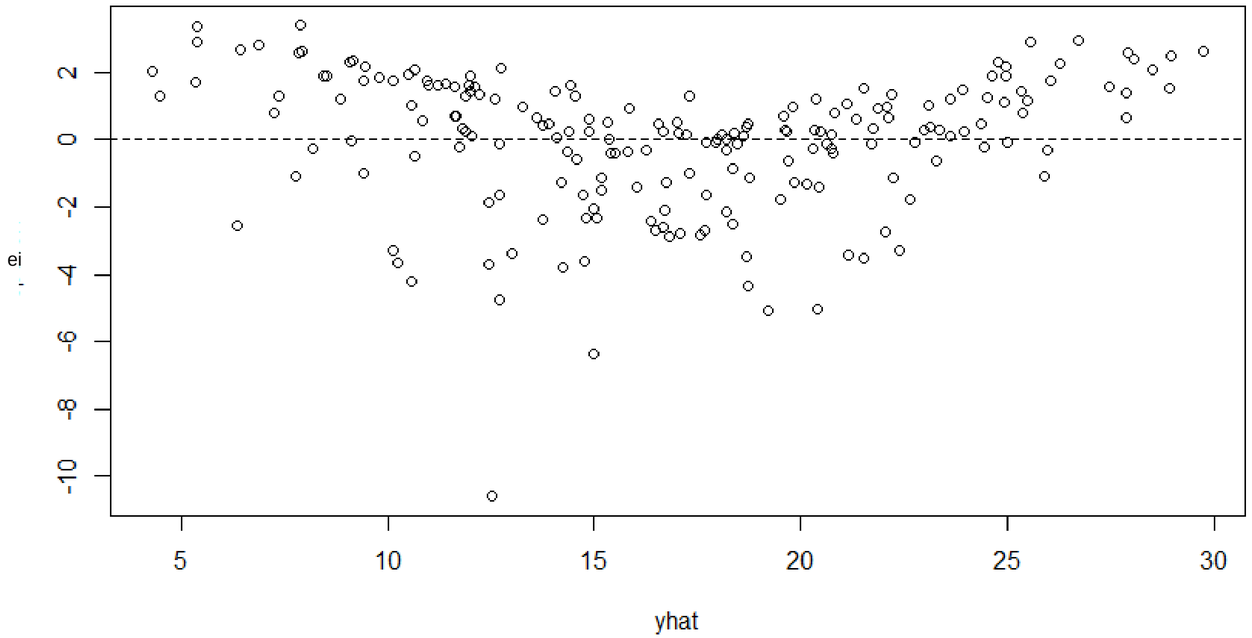
Residuals Analysis

The distribution of model residuals should be approximately normal.



The distribution is left skewed, it might be abnormal to cause any major concerns. We need to check

The **variance** of the residuals should be consistent for all observations.

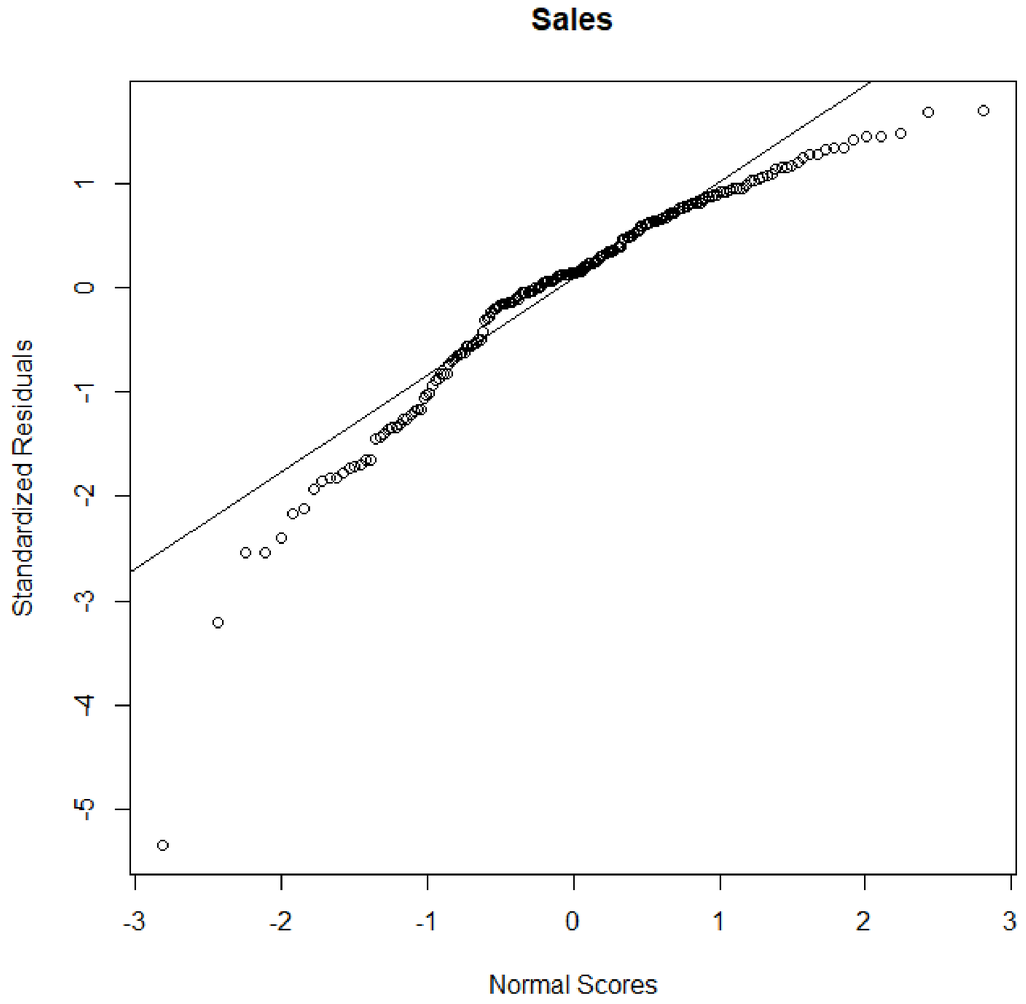


Ideally, the residuals should be equally scattered at every fitted value. We can see from the plot that the scatter tends to become a bit larger in the middle-fitted values, this pattern isn't extreme enough to cause too much concern, but need to check

Residuals Analysis

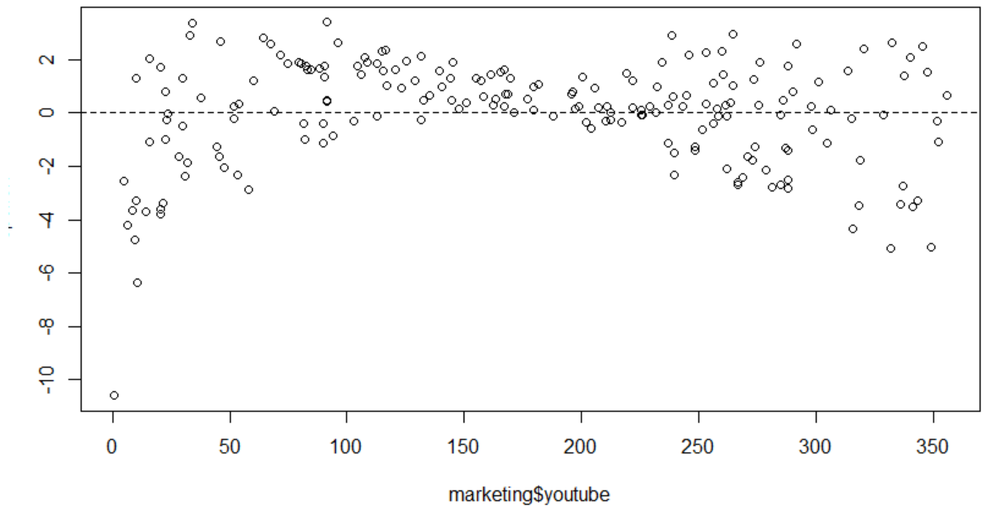
The distribution of model residuals should be approximately normal. (**Normality**)

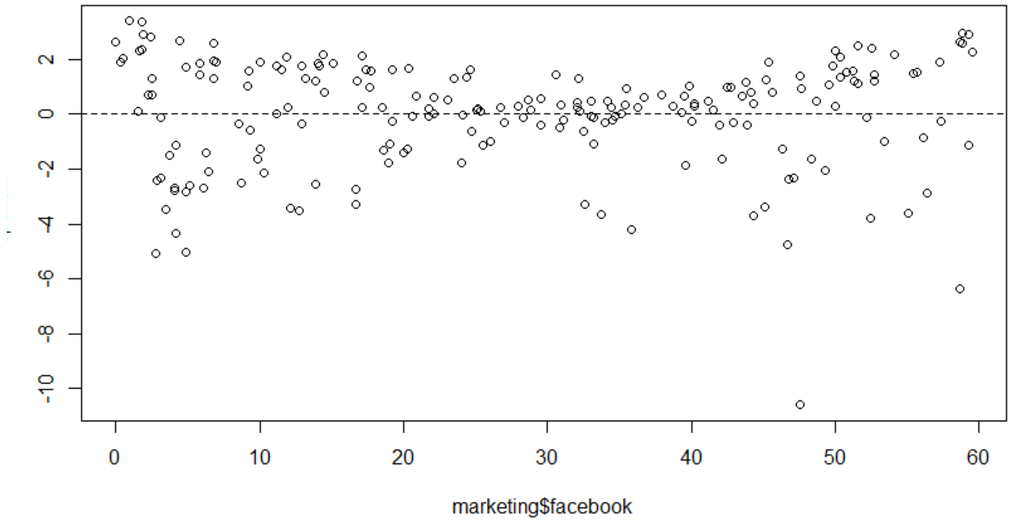
The data is inline as a straight line

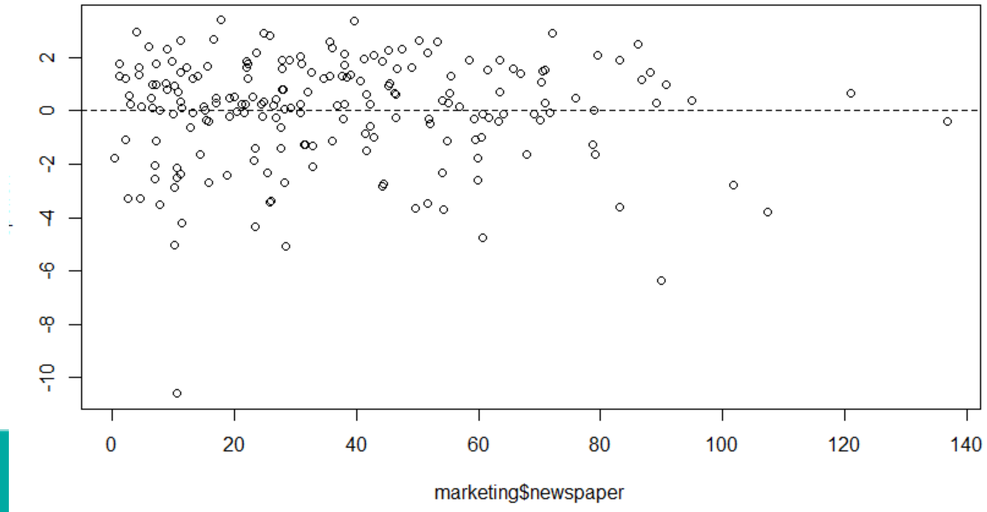


Residuals Analysis

Residual vs Predictor Plot







Exercise Questions

For each question develop the ML model and check the assumptions of linearity, independence, normality, and equal variances.

1. A designed experiment is done to assess how moisture content and sweetness of a pastry product affect a taster's rating of the product (*Pastry dataset*). In a designed experiment, the eight possible combinations of four moisture levels and two sweetness levels are studied. Two pastries are prepared and rated for each of the eight combinations, so the total sample size is n = 16. The *y*-variable is the rating of the pastry. The two *x*-variables are moisture and sweetness. The values (and sample sizes) of the *x*-variables were designed so that the *x*-variables were not correlated.

2. The data are from n = 214 females in statistics classes at the University of California at Davis (*Stat Females dataset*). The variables are y = student's self-reported height, x1 = student's guess at her mother's height, and x2 = student's guess at her father's height. All heights are in inches.

3. Data from n = 113 hospitals in the United States are used to assess factors related to the likelihood that a hospital patients acquires an infection while hospitalized. The variables here are y = infection risk, x1 = average length of patient stay, x2 = average patient age, x3= measure of how many x-rays are given in the hospital (*Hospital Infection dataset*).

Model Selection

- Model selection is a process of developing a model which produce the best model to be used to explain the response variables.
- The process of selecting the best explanatory or predictors (independent variables) able to simplify the model which contains only significant predictors.
- The Objectives are:
  - to describe the model in the simplest way whereby the redundant predictors are removed due to not give much information.
  - to identify and remove the unnecessary predictors that will add noise to the estimation of other variables.
  - to identify collinearity (linear association between two predictors), which is caused by having too many variables trying to do the same job.
  - to reduce cost in prediction process since time and money can be saved by measuring only significant variables.

Type of Model Selection

- Simple method used in Applied Statistics Course.
- Stepwise, Backward and Forward, Enter Method.
- Best Subset
- Leaps - similar to best subsets but is known to use a better algorithm to shortlist the models.
- RegBest() from FactoMineR – run one by one model.
- Simulated Annealing - Given a set of variables, a simulated annealing algorithm seeks a k-variable subset which is optimal, as a surrogate for the whole set, with respect to a given criterion. Annealing offers a method of finding the best subsets of predictor variables. Since the correlation or covariance matrix is a input to the anneal() function, only continuous variables are used to compute the best subsets.

Stepwise Regression

- Pass the full model to step function. It iteratively searches the full scope of variables in backwards directions by default, if scope is not given.
- It performs multiple iterations by dropping one X variable at a time. In each iteration, multiple models are built by dropping each of the X variables at a time.
- The AIC of the models is also computed and the model that yields the lowest AIC is retained for the next iteration.
- The variable that gives the minimum AIC when dropped, is dropped for the next iteration, until there is no significant drop in AIC is noticed.

Example

```
> selectedMod <- step(ModSel)
Start:  AIC=285.72
sales ~ youtube + facebook + newspaper

            Df Sum of Sq    RSS    AIC
- newspaper  1       0.1  802.0 283.75
<none>                    801.8 285.72
- facebook   1    1960.9 2762.7 531.13
- youtube    1    4403.5 5205.4 657.83

Step:  AIC=283.75
sales ~ youtube + facebook

            Df Sum of Sq    RSS    AIC
<none>                    802.0 283.75
- facebook  1    2225.7 3027.6 547.44
- youtube   1    4408.7 5210.6 656.03
```

```
> all_vifs <- car::vif(selectedMod)
> all_vifs
 youtube facebook
1.003013 1.003013
```

#The model
ModSel<- lm(sales ~ youtube+facebook+newspaper,
data = marketing)

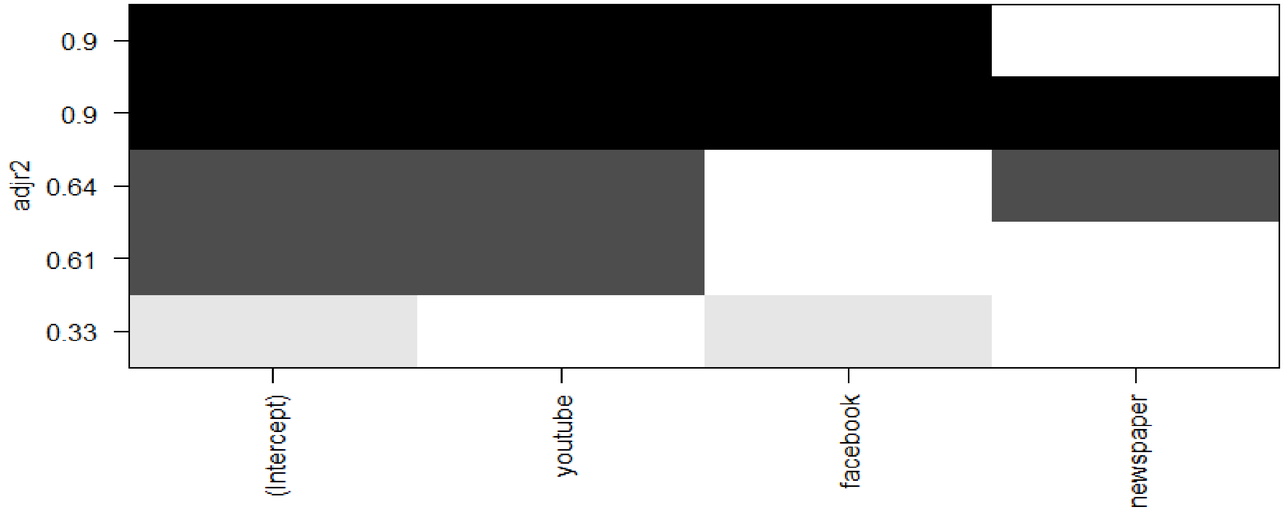
#Stepwise Regression
selectedMod <- step(ModSel)
summary(selectedMod)

#the condition of multicollinearity (checked using car::vif)
all_vifs <- car::vif(selectedMod)

Best Subset

- Technique that relies on stepwise regression to search, find and visualise regression models.
- Unlike the stepwise regression, we have more options to see what variables were included in various shortlisted models, force-in or force-out some of the explanatory variables and visually inspect the model's performance w.r.t Adj R-sq

```
#Best Subset Method
y <- as.matrix(marketing[,4])
x <- as.matrix(marketing[,1:3])

library(leaps)
regsubsetsObj <- regsubsets(x=x ,y=y, nbest = 2, really.big = T)
# regsubsets plot based on R-sq
plot(regsubsetsObj, scale = "adjr2")
```



Simulated Annealing

- Given a set of variables, a simulated annealing algorithm seeks a k-variable subset which is optimal, as a surrogate for the whole set, with respect to a given criterion.
- Annealing offers a method of finding the best subsets of predictor variables.
- Since the correlation or covariance matrix is a input to the anneal() function, only continuous variables are used to compute the best subsets.

```
library(subselect)
# perform annealing
results <- anneal(cor(predictors_df), kmin=1,
kmax=ncol(predictors_df)-1, nsol=4, niter=10, setseed=TRUE)
print(results$bestsets)

num_vars <- 2
selectVarsIndex <- results$bestsets[num_vars, 1:num_vars]
# new data for building selected model
newData <- cbind(y, x[, selectVarsIndex]) newData <-
data.frame(newData)
# build model
selectedMod <- lm(y ~ ., data=newData)
summary(selectedMod)
```

Summary

- There are 7 Steps process in statistical modelling.
- Linear Regression Analysis and Correlation
  - Simple Linear Model
  - Multiple Linear Model
- Using R Language in develop the model and analyse the data.
- Residual analysis to check assumption in linear model.
- Model Selection


Thank You!