VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CSE3506 Essentials of Data Analytics (2 0 2 4 4)

Course Objectives

- ✔ To understand the concepts of analytics using various machine learning models.
- ✔ To appreciate supervised and unsupervised learning for predictive analysis
- ✔ To understand data analytics as the next wave for businesses looking for competitive advantage
- ✔ Validate the results of their analysis according to statistical guidelines
- ✔ Validate and review data accurately and identify anomalies
- ✔ To learn aspects of computational learning theory
- ✔ Apply statistical models to perform Regression Analysis, Clustering and Classification

Course Outcomes

- ✔ Identify and apply the appropriate supervised learning techniques to solve real world problems.
- ✔ Choose and implement typical unsupervised algorithms for different types of applications.
- ✔ Implement statistical analysis techniques for solving practical problems.
- ✔ Understand different techniques to optimize the learning algorithms.
- ✔ Aware of health and safety policies followed in organization, data and information management and knowledge & skill development.

Syllabus

Module-1: Regression Analysis

- ✔ Linear regression: simple linear regression - Regression Modeling - Correlation, ANOVA, Forecasting, Autocorrelation **(6 Hours)**

Module-2: Classification

- ✔ Logistic Regression, Decision Trees, Naïve Bayes-conditional probability - Random Forest - SVM Classifier **(6 Hours)**

Module-3: Clustering

- ✔ K-means, K-medoids, Hierarchical clustering **(4 Hours)**

Module-4: Optimization

- ✔ Gradient descent - Variants of gradient descent - Momentum - Adagrad - RMSprop - Adam – AMSGrad **(3 Hours)**

Module-5: Managing Health and Safety

- ✔ Comply with organization's current health, safety and security policies and procedures - Report any identified breaches in health, safety, and security policies and procedures to the designated person - Identify and correct any hazards that they can deal with safely, competently and within the limits of their authority - Report any hazards that they are not competent to deal with to the relevant person in line with organizational procedures and warn other people who may be affected. **(4 Hours)**

Module-6: Data and Information Management

- ✔ Establish and agree with appropriate people the data/information they need to provide, the formats in which they need to provide it, and when they need to provide it - Obtain the data/information from reliable sources - Check that the data/information is accurate, complete and up-to-date **(4 Hours)**

Module-7: Learning and Self Development

- ✔ Obtain advice and guidance from appropriate people to develop their knowledge, skills and competence - Identify accurately the knowledge and skills they need for their job role - Identify accurately their current level of knowledge, skills and competence and any learning and development needs - Agree with appropriate people a plan of learning and development activities to address their learning needs **(3 Hours)**

Text Book

- Cathy O'Neil and Rachel Schutt. "Doing Data Science, Straight talk from the Frontline", O'Reilly. 2014.
- Dan Toomey, "R for Data Science", Packt Publishing, 2014.
- Trevor Hastie, Robert Tibshirani and Jerome Friedman. "Elements of Statistical Learning", Springer , Second Edition. 2009.
- Kevin P. Murphy. "Machine Learning: A Probabilistic Perspective", MIT Press; 1st Edition, 2012.

Reference Books

- Glenn J. Myatt, "Making Sense of Data : A Practical Guide to Exploratory Data Analysis and Data Mining", John Wiley & Sons, Second Edition, 2014.
- G. K. Gupta, —Introduction to Data Mining with Case Studies", Easter Economy Edition, Prentice Hall of India, 2006.
- Michael Berthold, David J. Hand, "Intelligent Data Analysis", Springer, 2007.
- Colleen Mccue, "Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis", Elsevier, 2007.
- R N Prasad, Seema Acharya, "Fundamentals of Business Analytics", Wiley; Second edition, 2016.
- https://www.sscnasscom.com/qualification-pack/SSC/Q2101/

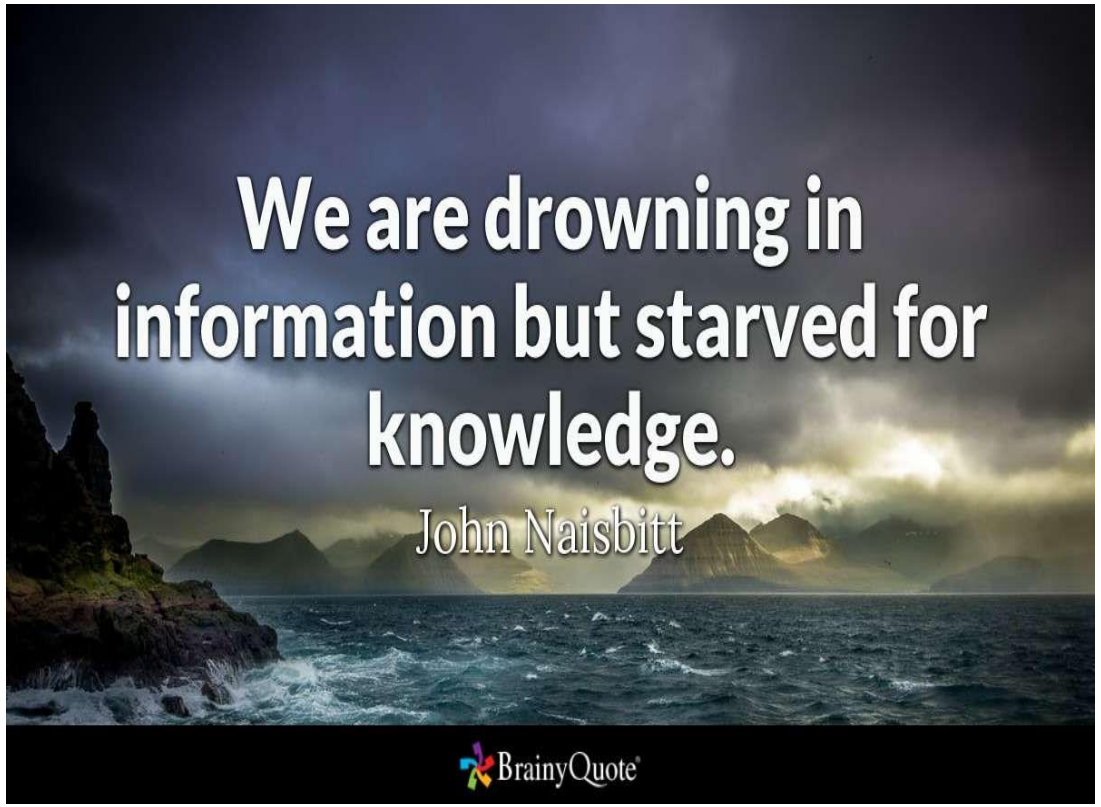
Module-1: Regression Analysis

Linear regression: simple linear regression - Regression Modelling - Correlation, ANOVA, Forecasting, Autocorrelation

Data Analytics – What?

- **Science of analyzing raw data in order to make conclusions about that information** *[Investopedia]*
- **Analytics** is the systematic computational analysis of data or statistics *[Wikipedia]*
- **Data analysis** is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, deriving conclusions, and supporting decision-making *[Wikipedia]*

Data Analytics – Why?



- Helps business to optimize their performances
  - Reduce cost
  - Improved business
  - Make better decision

Data Analytics – Types

- **Descriptive analytics** - describes what has happened over a given period of time.
  - Have the number of views gone up?
  - Are sales stronger this month than last?

- **Diagnostic analytics** - focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing.
  - Did the weather affect sales of a cool drink?
  - Did that latest marketing campaign impact sales?

Data Analytics – Types

- **Predictive analytics** - focuses on what is likely going to happen in the near term.
  - What happened to sales the last time we had a hot summer?
  - How many weather models predict a hot summer this year?
- **Prescriptive analytics** suggests a course of action.
  - If the likelihood of a hot summer is measured as an average of say five weather models is above 58%, we should add an evening shift to the workers to increase output.

What is Machine Learning?

- Large volume of data demands automated methods of data analysis which is what machine learning provides.

- Machine learning is defined as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.

Machine Learning Paradigms

- Three Learning Paradigms
  - Predictive or Supervised Learning
  - Descriptive or Unsupervised Learning
  - Reinforcement Learning

Statistical learning refers to

- getting inference from a vast data set using set of tools

- These tools can be classified as
  - **Supervised**

    or
  - **Unsupervised**

Supervised Learning

- A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalizes to respond correctly to all possible inputs. This is also called learning from exemplars.

Unsupervised Learning

- Correct responses are not provided, but instead the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorized together.
- The statistical approach to unsupervised learning is known as density estimation.

Supervised Learning

- The goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs.

$$\mathrm{D} = \{(x_i, y_i)\}^N{}_{i=1}$$

  - Here, D is called the **training set**, and N is the number of training examples.
  - (x_i, y_i) is the ith training sample

- Each x_i is a d-dimensional vector of numbers - **features**, **attributes** or **covariates**
  - Height, weight, age or complex objects (image, text, time series, graph etc.)

Supervised Learning

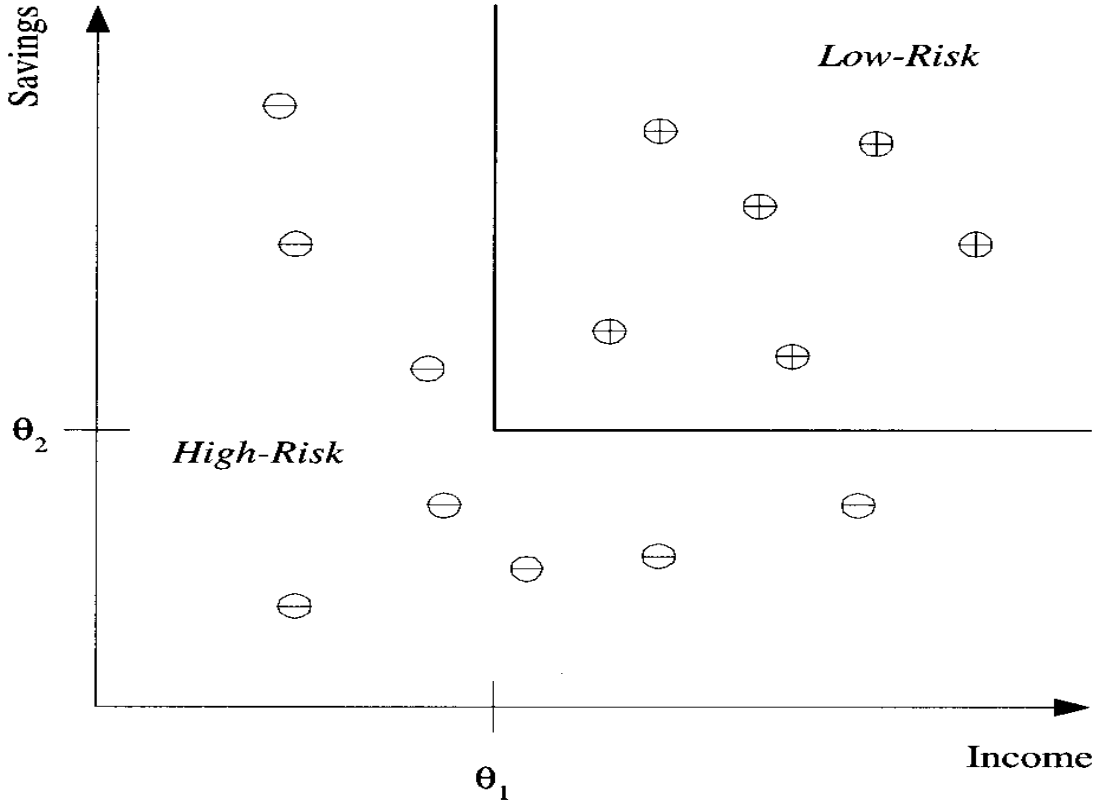
- Each y_i is a response variable
  - Categorical or nominal variable from a finite set $y_i \epsilon \{1, \dots, C\}$
    - Eg., Male or Female

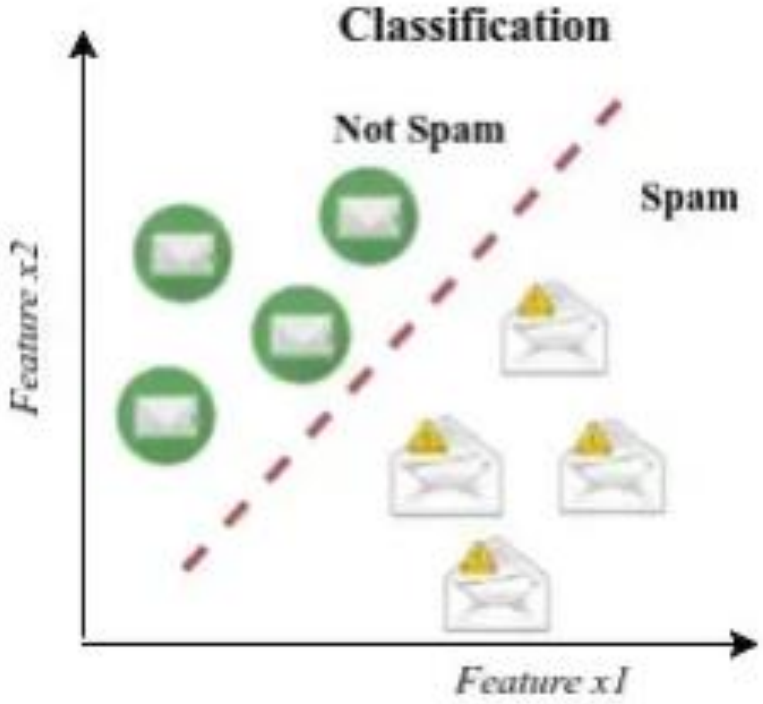
Classification Problem

Regression Problem

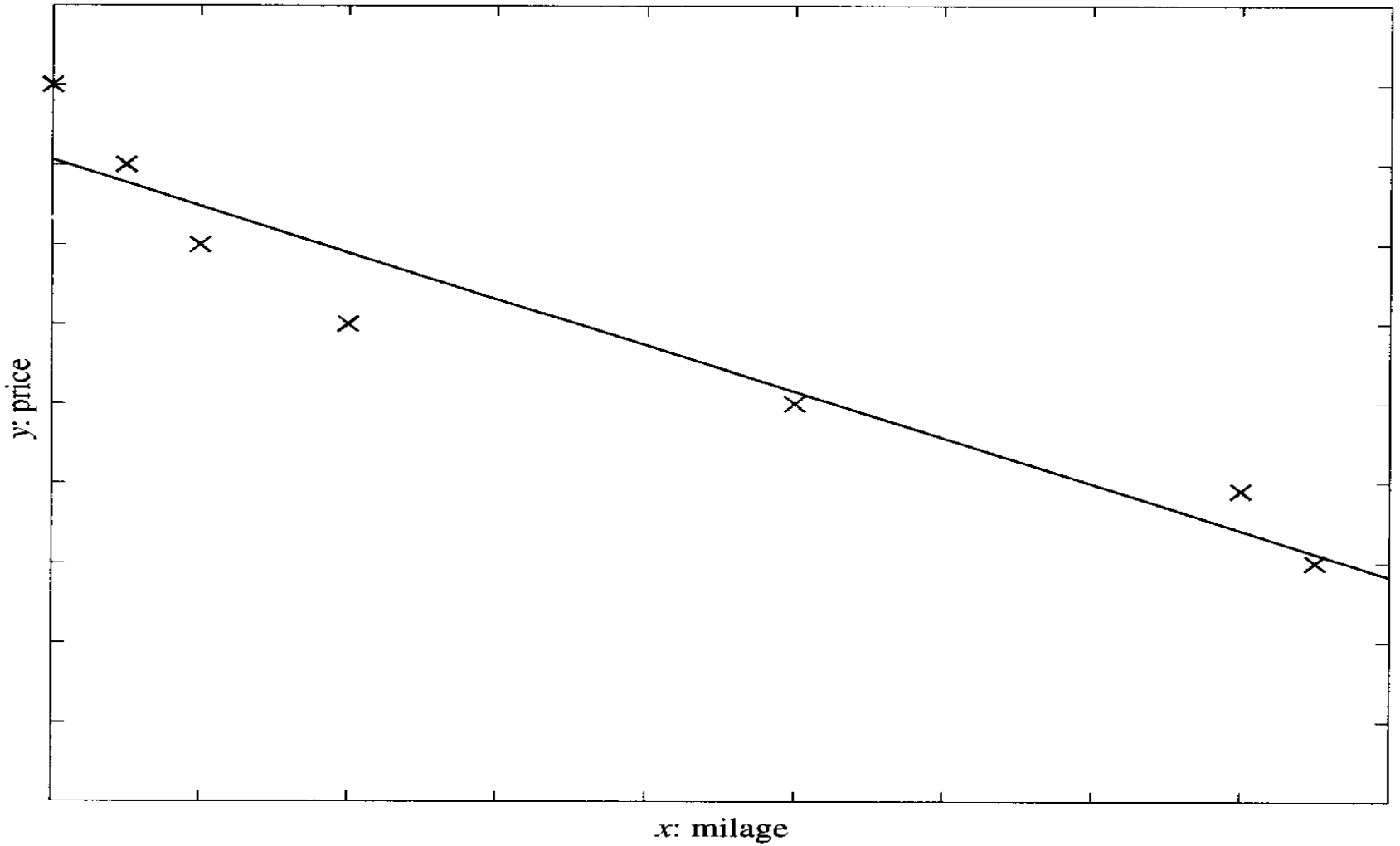
  - Real valued variable
    - Eg., Temperature, Age, Height, Weight

Classification Problem – Credit Scoring





Regression Problem – Car Price Prediction



Unsupervised Learning

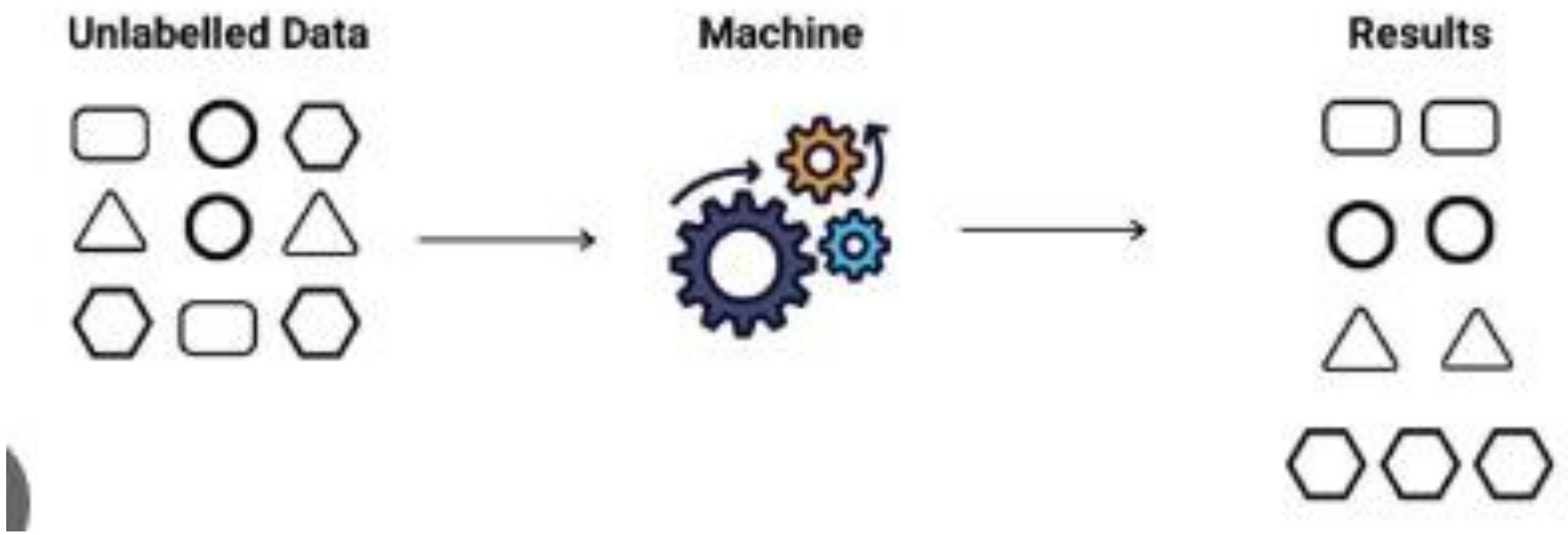
- The goal is to find “interesting patterns” in the data.

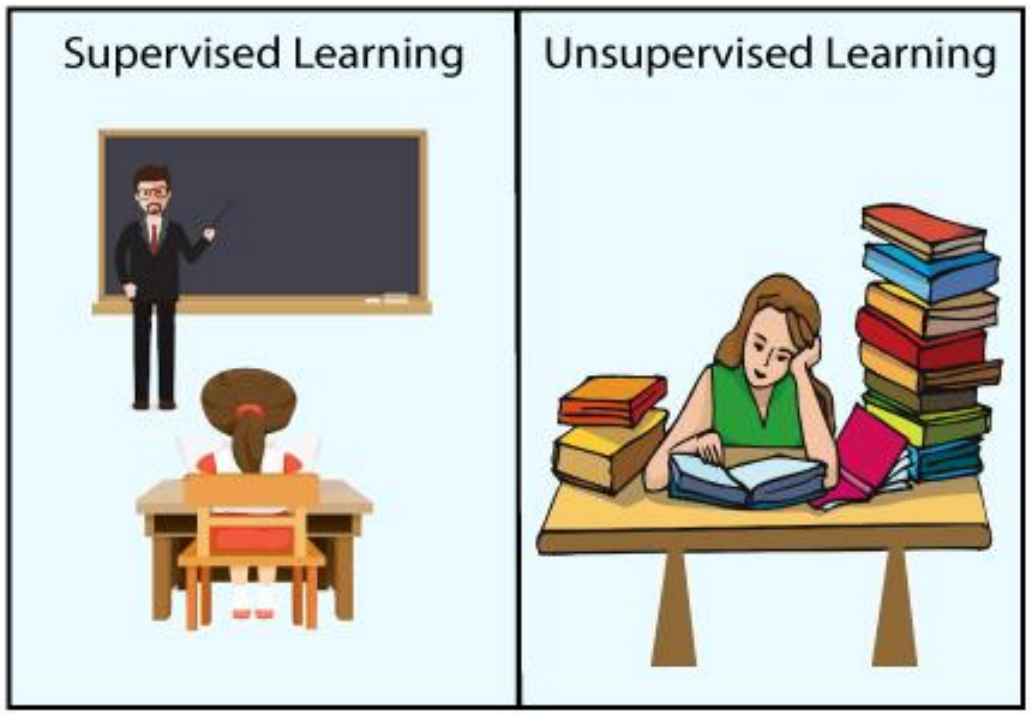
$$\mathrm{D} = \{x_i\}^N{}_{i=1}$$

- Sometimes called “Knowledge Discovery”

- Clustering algorithms come in this category where data are grouped based on similarities.

Clustering Problem –Grouping people based on weight and height



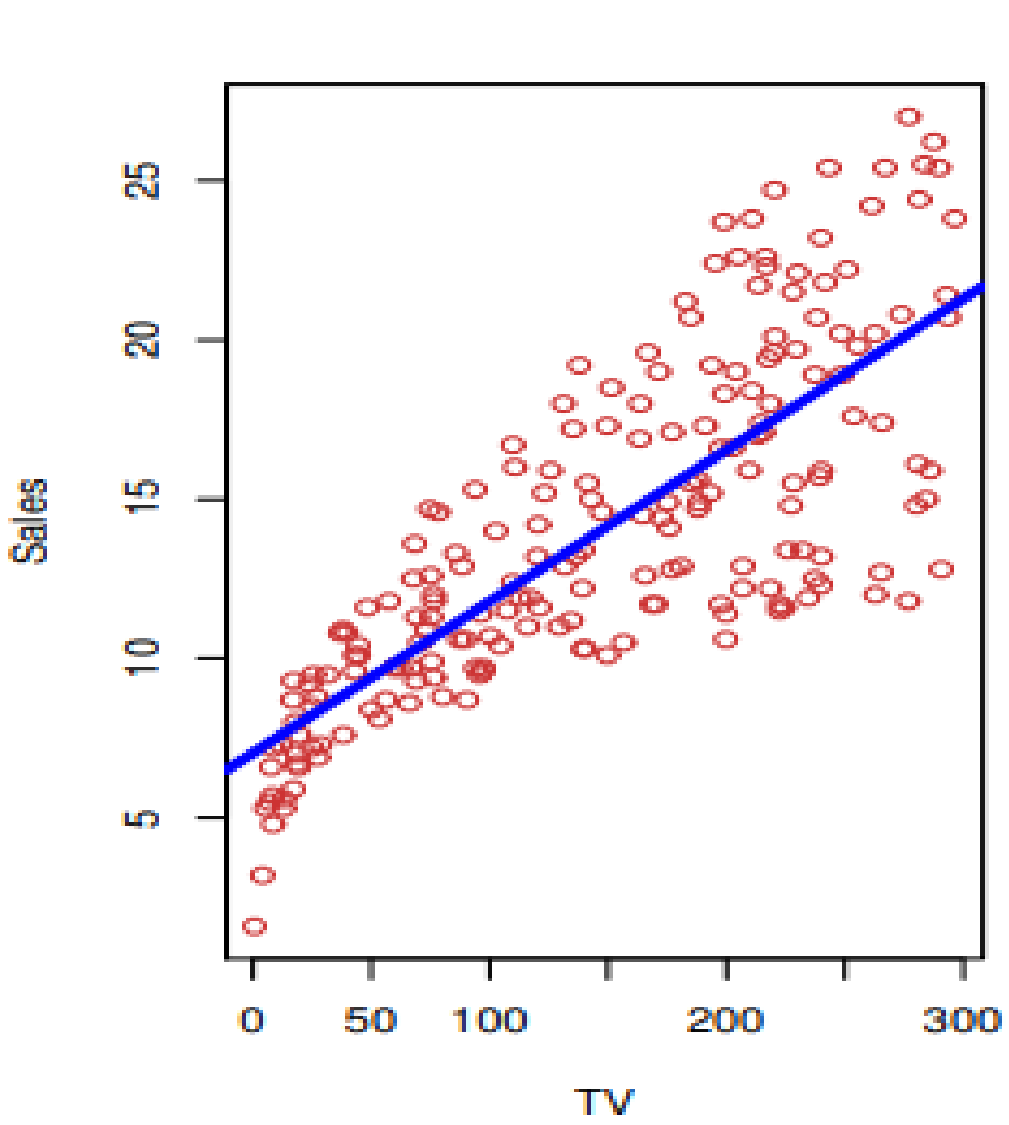

Supervised Learning
Unsupervised Learning

- **Supervised statistical learning:**
  - ✓ Involves building a statistical model for predicting, or estimating, an output based on one or more inputs
  - ✓ **Examples**: Problems occur in business, medicine, astrophysics, and public policy
- **Unsupervised statistical learning:**
  - ✓ There are inputs but no supervising output; nevertheless we can learn relationships and structure from such data
  - ✓ **Example**: Input dataset containing images of different types of cats and dogs

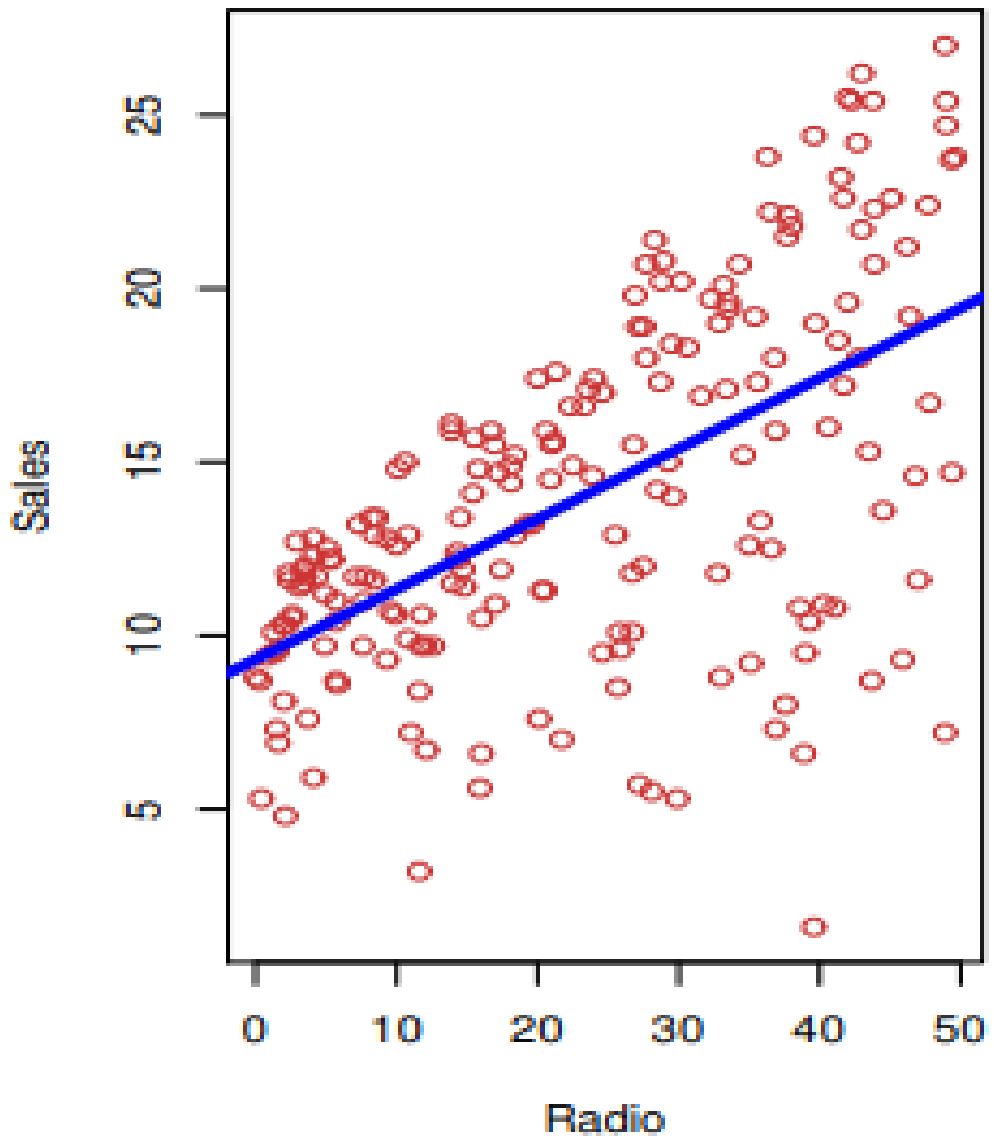
Process

- Data Collection and Preparation
- Feature Selection
- Algorithm Choice
- Parameter and Model Selection
- Training
- Evaluation

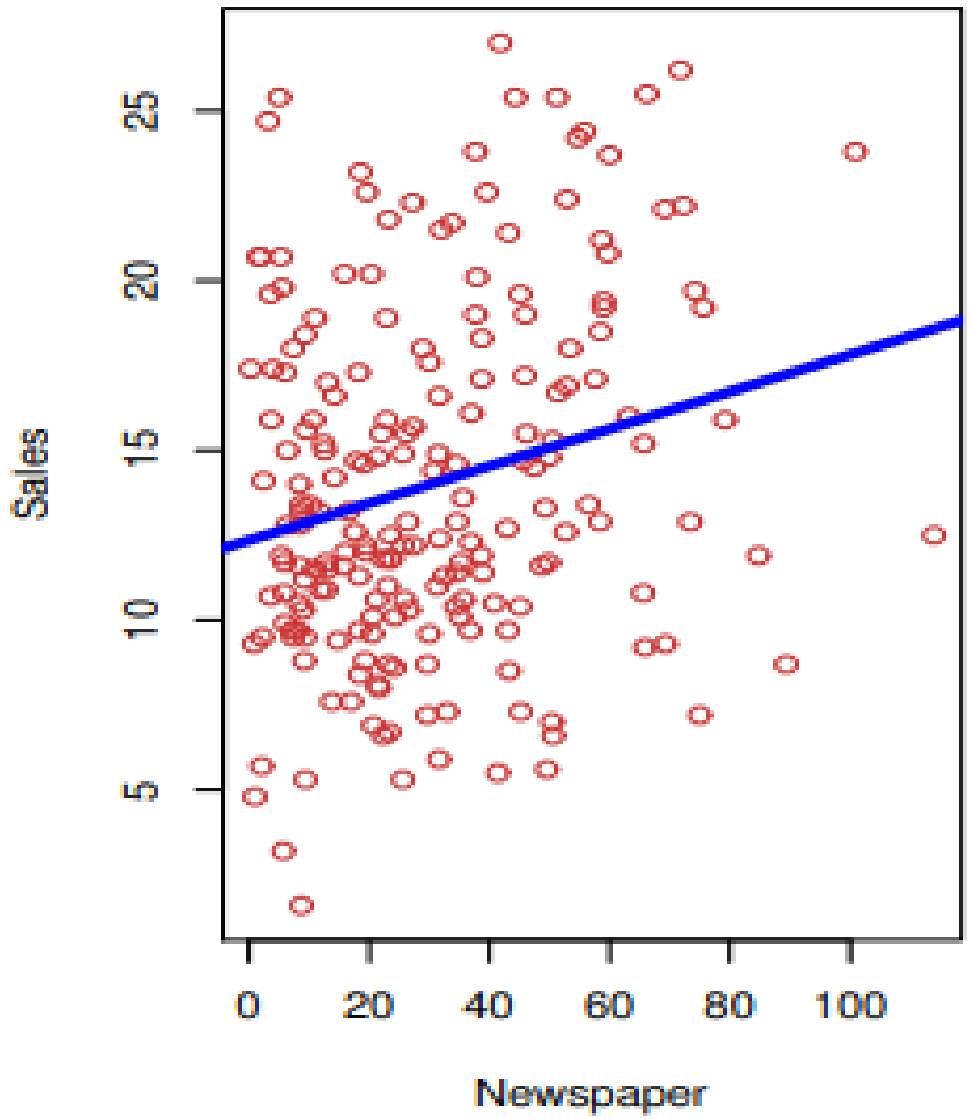
Statistical Learning – Advertising Data



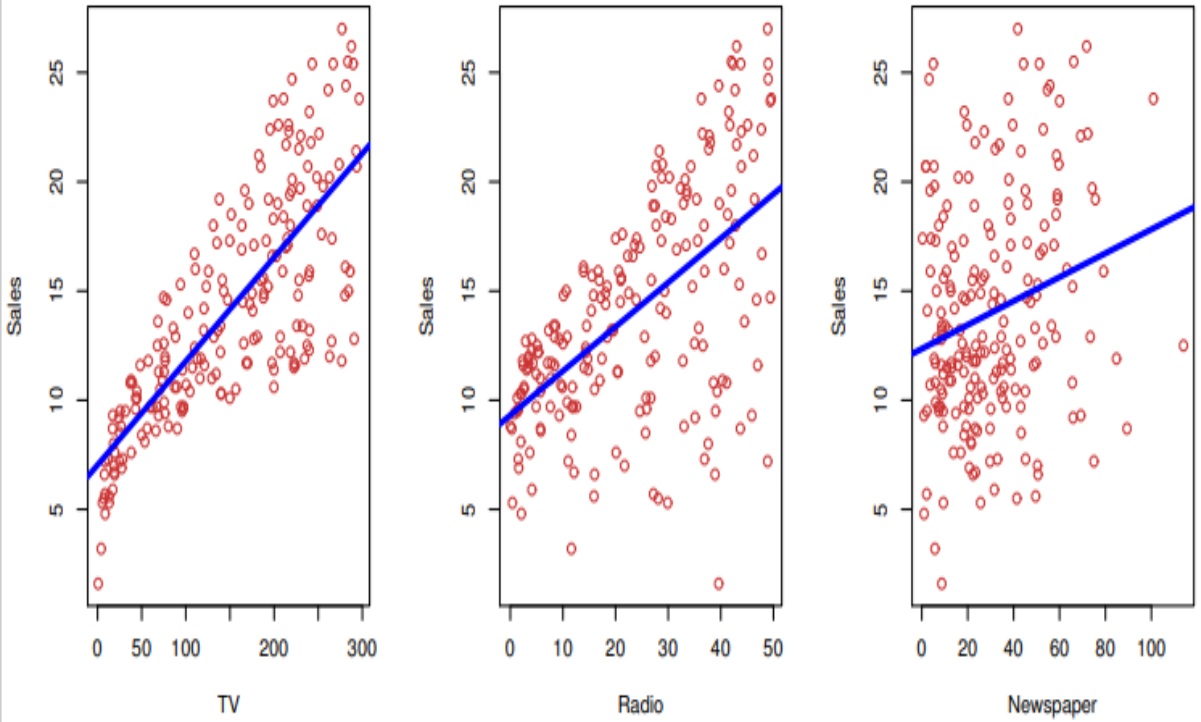
- ✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: **TV**, radio, and newspaper



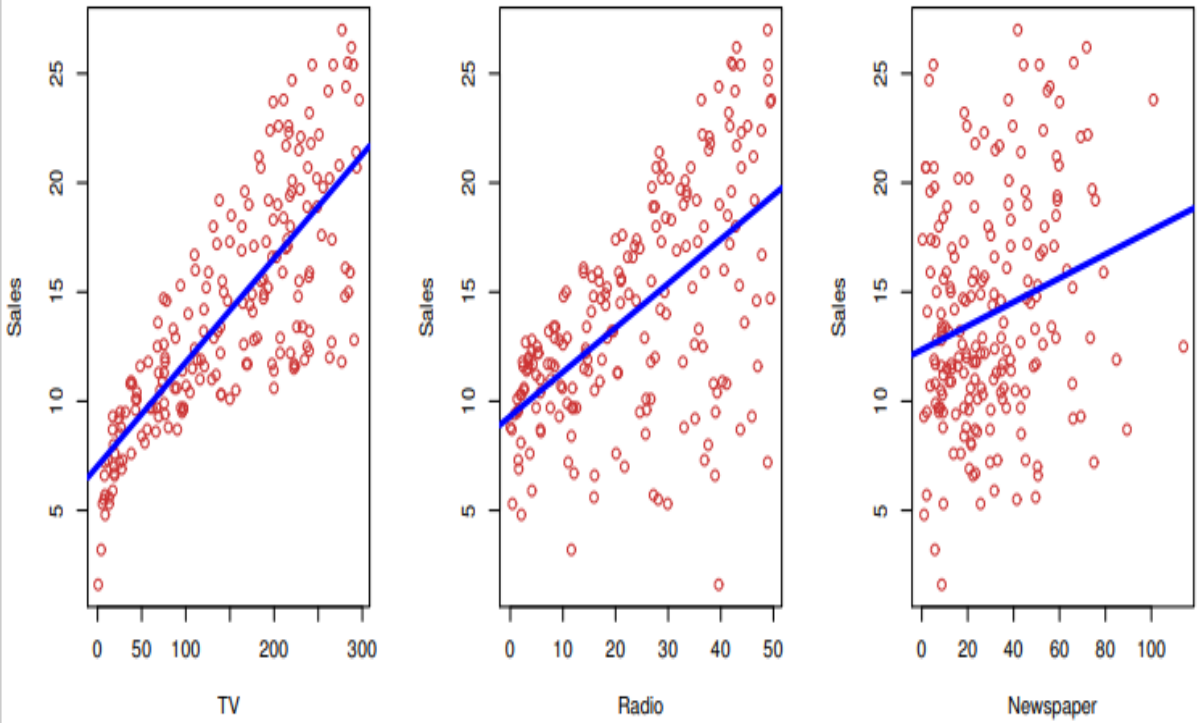
- ✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: TV, **Radio**, and newspaper



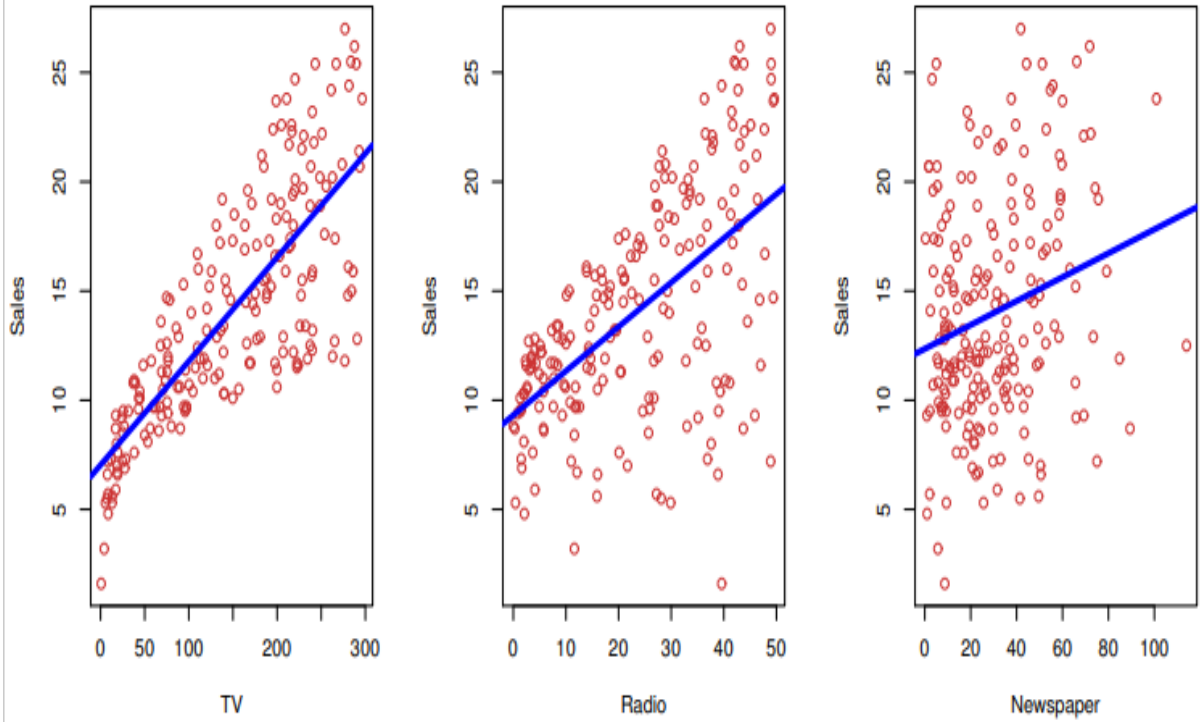
- ✓ The Advertising data set consists of the sales of a product in 200 different cities, along with advertising budgets for three different media: TV, Radio, and **newspaper**
- ✓ Goal is to develop an <u>accurate model</u> that can be used to predict sales on the basis of the three media budgets



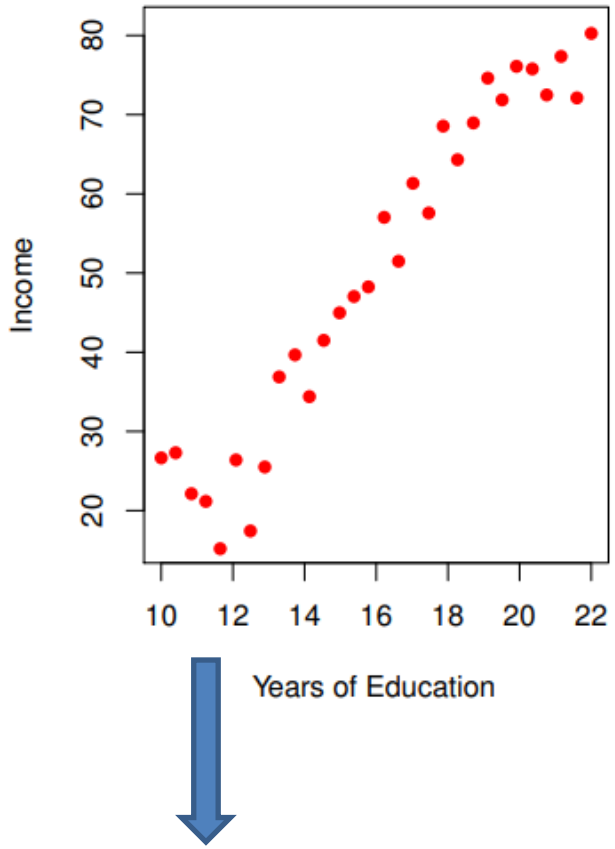
- Input Variables: **Advertising budgets**
- Input Variables are denoted by X
- X_1 – TV budget
- X_2 – Radio budget
- X_3 – Newspaper budget
- Input variables are called by different names like
  - **Predictors**
  - **Independent variables**
  - **Features**
  - **Variables**



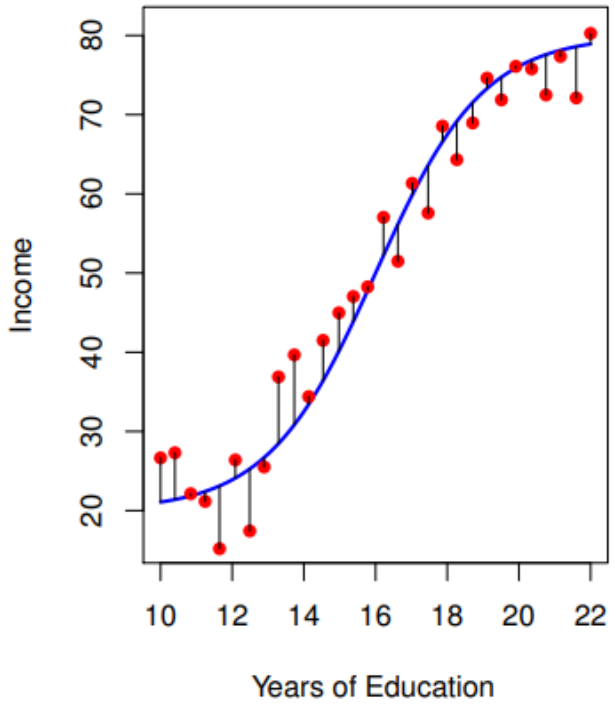
- Output Variable: **Sales**
- Output Variables are denoted by Y
- Output variables are called by different names like
  - **Responses,**
  - **Dependent variables**



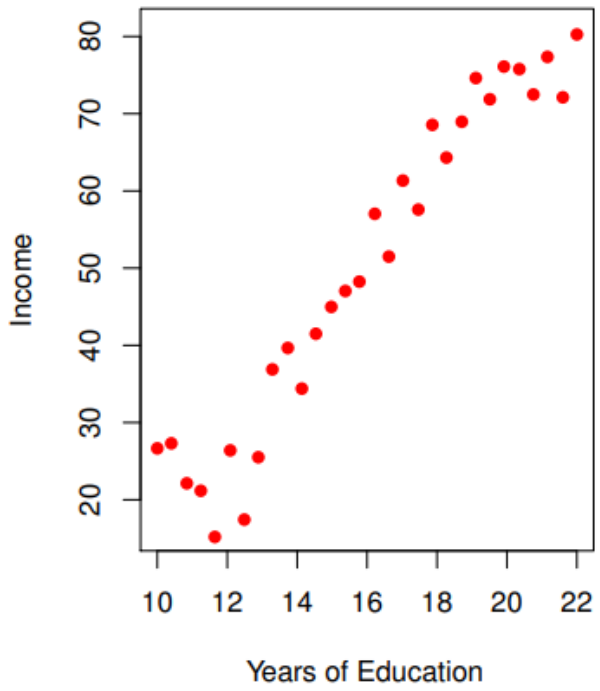
- ✓ There is some relationship between Y and X = $(X_1, X_2,...,X_p)$
- ✓ General form of relationship is
  - ✓ **$Y = f(X) + \varepsilon$**
- ✓ where
  - ✓ f is some fixed but unknown function of $X_1,...,X_p$
  - ✓ ε is a random error term, which is independent of X and has mean zero

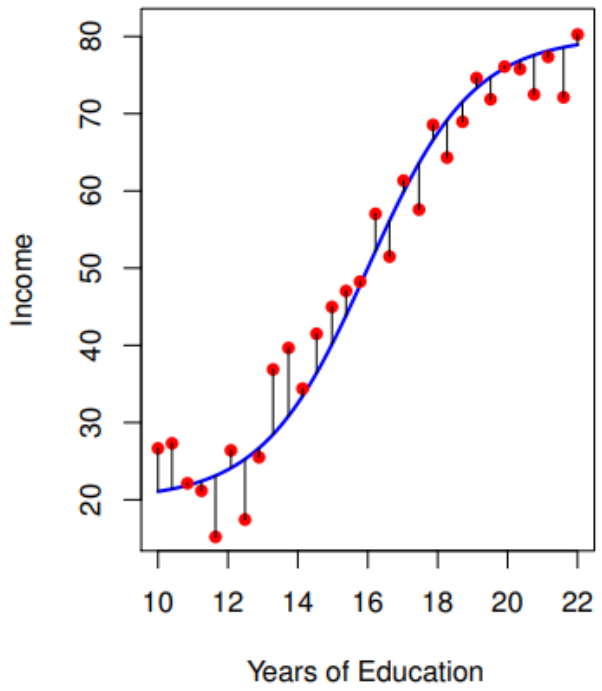


Observed values of income and years of education for 30 individuals



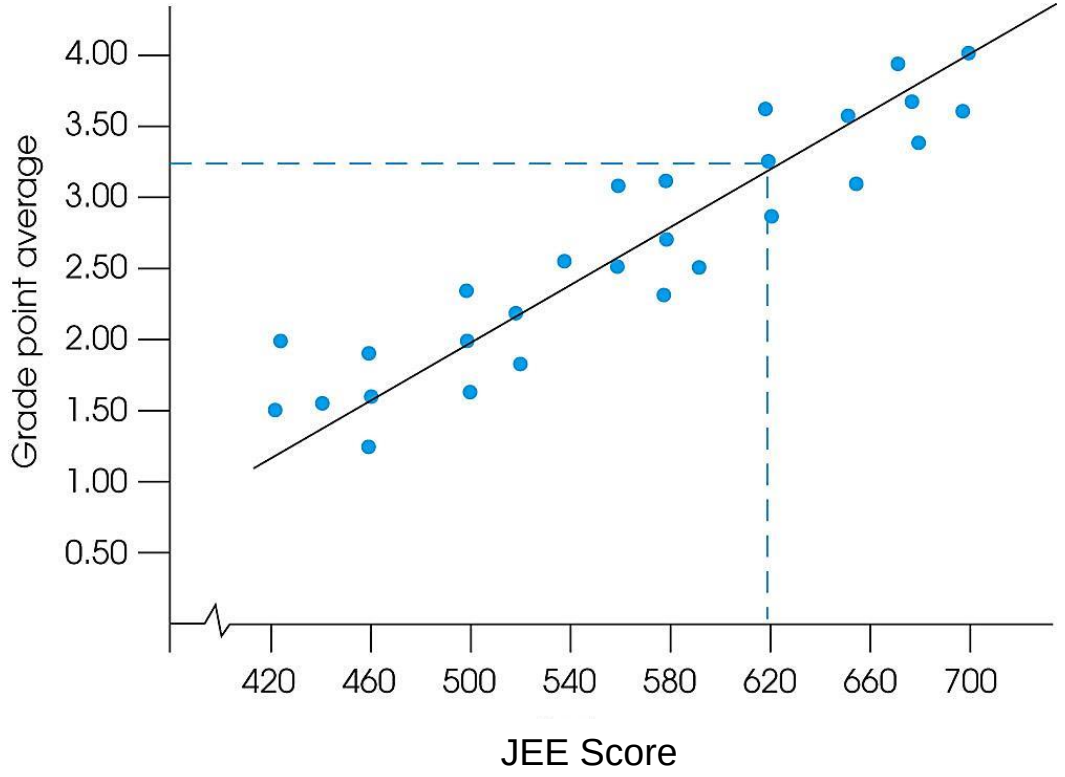
- The black lines represent the error associated with each observation.
- Here some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve)
- **Overall, these errors have approximately mean zero**

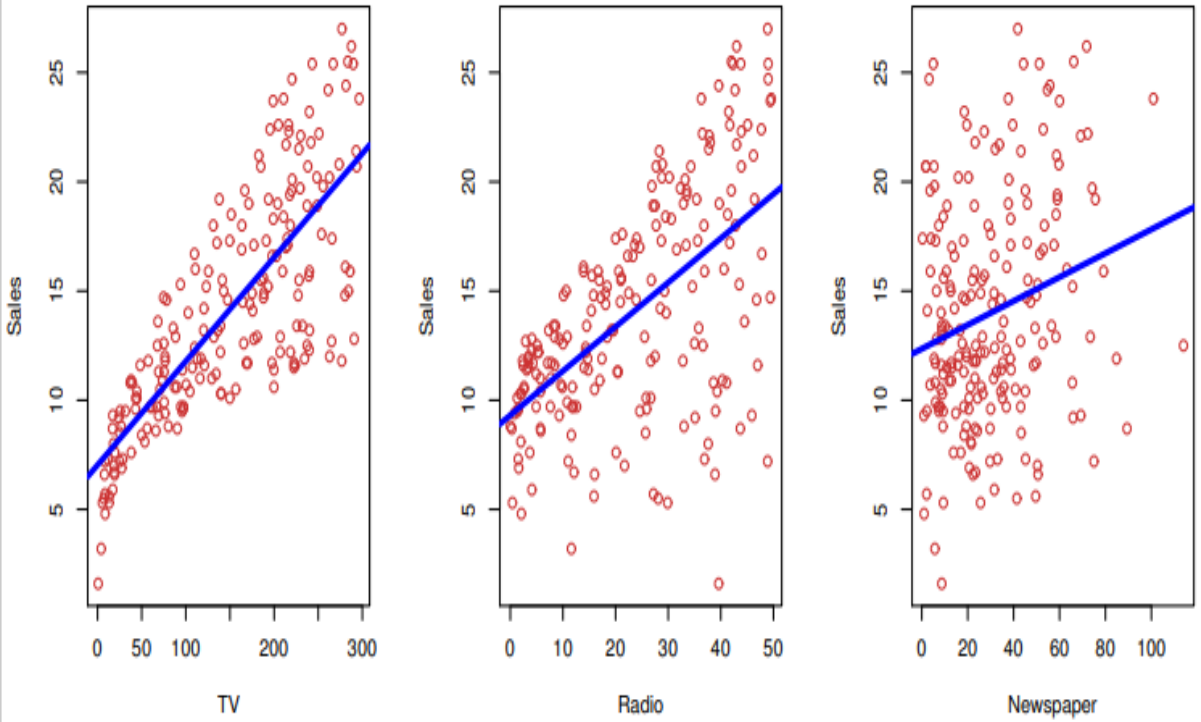




- ✓ Statistical Learning refers to a set of approaches for estimating f in the equation
  - $Y = f(X) + \varepsilon$
- ✓ Reasons to estimate 'f',:
  - ✓ **Prediction**
  - ✓ **Inference**

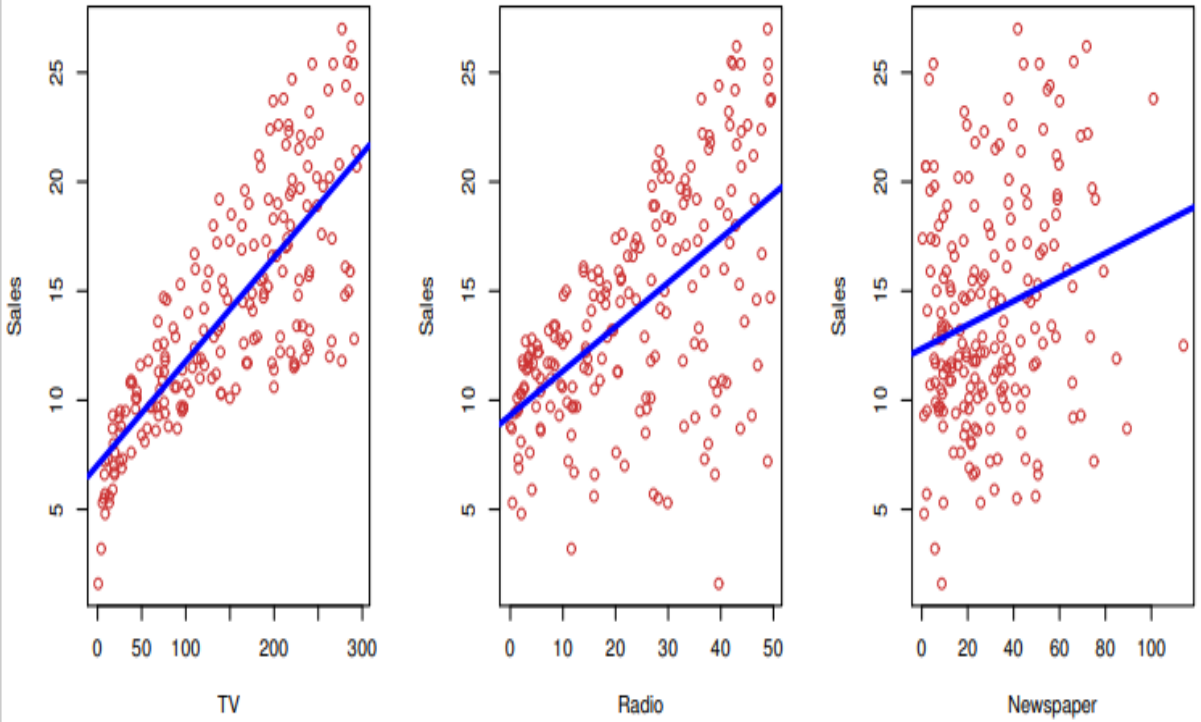
- **Linear Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.
- This is a very simple approach for supervised learning
- In particular, it is a useful tool for predicting a quantitative response.





On the basis of given advertising data,

- Marketing plan for next year can be made
- To develop the marketing plan, some information is required.
  - Is there a relationship between advertising budget and sales?
  - Is the relationship linear?
  - Predicting sales with a high level of accuracy requires a strong relationship.
  - If it is strong relationship then
  - In marketing, it is known as a **synergy effect**, while in statistics it is called **an interaction effect**



The important questions are

- ✓ **Which media contribute more to sales?**
- ✓ Do all three contribute to sales, or do just one or two.
- ✓ The individual effects of each medium on the money spent
- ✓ For every dollar spent on advertising in TV or Radio or Newspaper, by what amount will sales increase?
- ✓ How accurately can we predict this amount of increase?

Linear regression can be used to answer each of these questions

Linear Regression - Types

- **Types:**
  - ✓ Based on the number of independent variables, there are two types of linear regression
    - ✓ Simple Linear Regression
    - ✓ Multiple Linear Regression
- Mathematically, the linear relationship is approximately modeled as
  - $y = \beta_0 + \beta_1 x$

Simple Linear Regression

Estimating the coefficients β_0 and β_1

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

These β_0 and β_1 are the least squares coefficient estimates for simple linear regression, and they give **the best linear fit** on the given training data.

Question-1:

Consider the following five training examples

X = [2 3 4 5 6]

Y = [12.8978 17.7586 23.3192 28.3129 32.1351]

(a) Find the best linear fit

(b) Determine the minimum RSS

(c) Draw the residual plot for the best linear fit and comment on the suitability of the linear model to this training data.

Solution:

(a) To find the best fit, calculate the model coefficients using the formula

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$
	2	12.8978	-2	-9.9869	19.9738	4
	3	17.7586	-1	-5.1261	5.1261	1
	4	23.3192	0	0.4345	0.0000	0
	5	28.3129	1	5.4282	5.4282	1
	6	32.1351	2	9.2504	18.5008	4
Sum	20	114.4236	0	0.0000	49.0289	10
Mean	4	22.88472				

The best linear fit is

$Y = 4.9029X + 3.2732$

Substituting in the formula

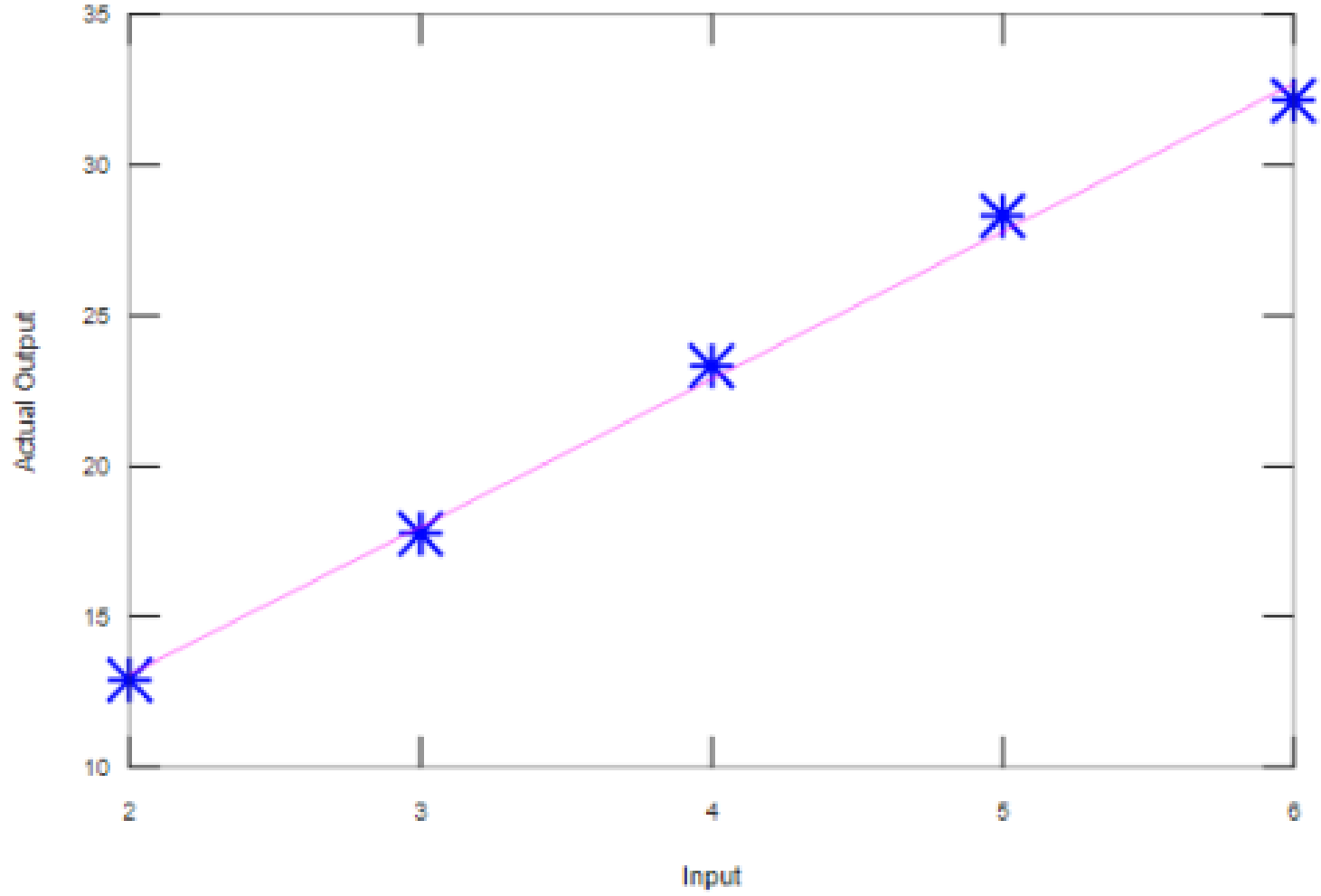
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

β_0 =3.2732
β_1 = 4.9029

Solution:

Best Linear Fit



Question-2:

Consider the following five training examples

Find the best linear fit

x	y
1	2
2	4
3	5
4	4
5	5

Example

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
3	4			10	6

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(yi-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(xi-\bar{x})^2} = \frac{6}{10} = 0.6$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 2.2$$

$$\hat{y} = 2.2 + .6x$$

Linear Regression Model (with Normally Distributed Errors)

- In most linear regression analyses, it is common to assume that the error term is a normally distributed random variable with mean equal to zero and constant variance.

- Thus, the linear regression model is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon$$

where:

y is the outcome variable

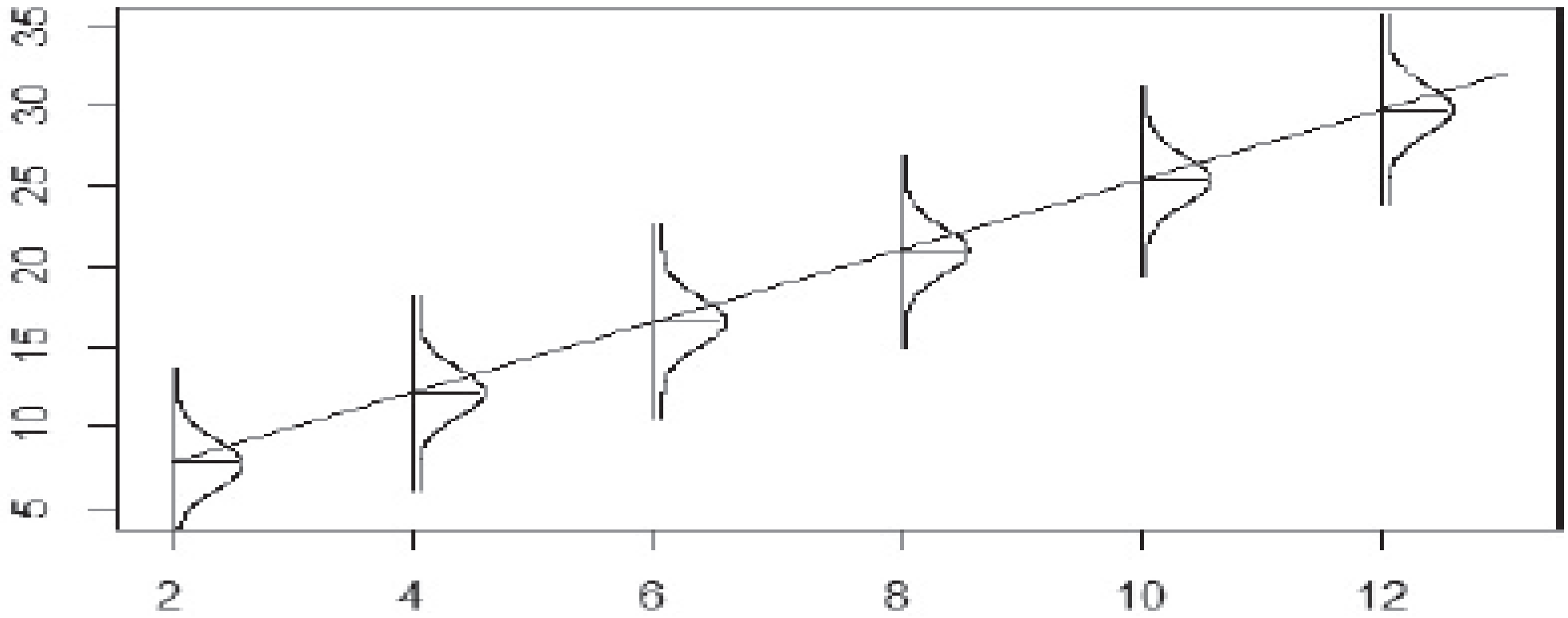
x_j are the input variables, for $j = 1, 2, \ldots, p - 1$

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j, for $j = 1, 2, \ldots, p - 1$

$\varepsilon \sim N(0, \sigma^2)$ **and the** ε**s are independent of each other**

Linear Regression Model (with Normally Distributed Errors)



Normal distribution about y for a given value of x

Sample data and Model

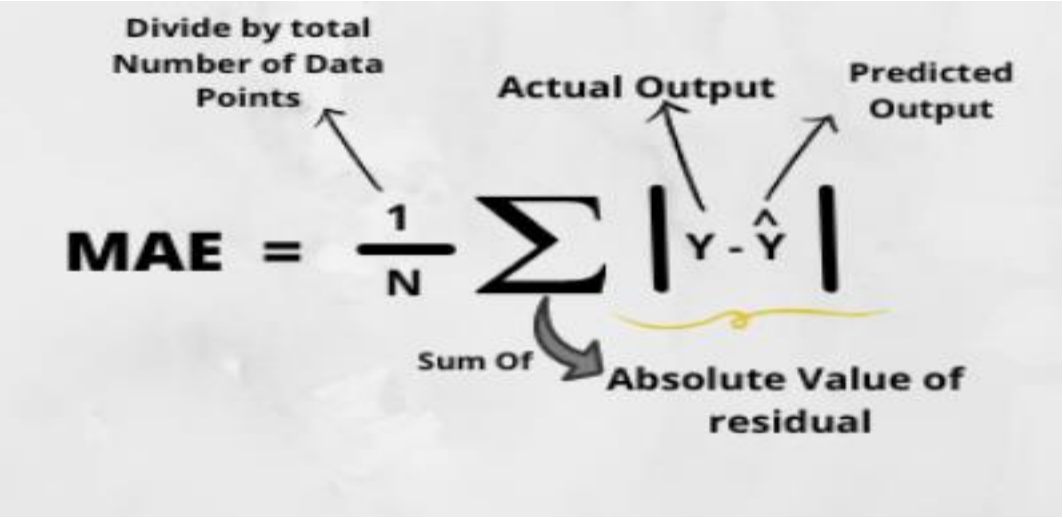
- Data: Marketing from Datarium package
- We want to predict future sales on the basis of advertising budget spent on youtube.
  - sales = b0 + b1 * youtube
- The R function lm() can be used to determine the beta coefficients of the linear model:

```
#building a linear model
model <- lm(sales~youtube,data=marketing)
model

Coefficients:
(Intercept) youtube
8.43911     0.04754
```

Performance Metrics

- MAE
- MSE
- RMSE
- RSS
- TSS
- RSE
- SE
- Confidence Interval
- R-Squared

Mean Absolute Error(MAE)



We aim to get a minimum MAE because this is a loss.

3. Mean Absolute Percentage Error (MAPE)

$$M = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

M = mean absolute percentage error

n = number of times the summation iteration happens

A_t = actual value

F_t = forecast value

What is a good value of MAPE?

The closer the MAPE value is to zero, the better the predictions.
A MAPE **less than 5%** is considered as an indication that the forecast is acceptably accurate.

Mean Squared Error(MSE)

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

The lower the value the better and 0 means the model is perfect.

6. Root Mean Squared Error(RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)}$$

This produces a value **between 0 and 1**, where values closer to 0 represent better fitting models. Based on a rule of thumb, it can be said that RMSE values **between 0.2 and 0.5** shows that the model can relatively predict the data accurately.

- Once we produce the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the training data, we can predict y given x:

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x.$$

- Let $\hat{y}_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for i^{th} value of y based on the i^{th} value of x. Then

$$e_i = y_i - \hat{y}_i$$

represents the i^{th} residual. This is the difference between the i^{th} observed response value and the i^{th} predicted response value.

- The residual sum of squares (RSS) is defined as

$$\text{RSS} = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

where n is the number of predictions or simply, the number of samples in the training data.

To calculate RSS, use the following formula

$$\text{RSS} = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

where $e_i = y_i - \hat{y}_i$

$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares*.

Standard Error

- Assuming the errors ϵ_i for each observation are uncorrelated with common variance σ^2, the *standard errors* associated with $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as

$$\mathrm{SE}(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

and

$$\mathrm{SE}(\hat{\beta}_1) = \sigma \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

- In general, $\sigma = \sqrt{\mathrm{Var}(\epsilon)}$ is not known, but can be estimated from the data. This estimate is known as the *residual standard error* (RSE), and is expressed as

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}}.$$

Confidence intervals

- Standard errors can be used to compute **confidence intervals**
- A 95% confidence interval is defined as a range of values such that with 95% interval probability, the range will contain the true unknown value of the parameter
- The range is defined in terms of lower and upper limits computed from the sample of data

- For linear regression, the 95% confidence interval for β_0 approximately takes the form

$$\hat{\beta}_0 \pm 2\,\mathrm{SE}(\hat{\beta}_0).$$

- That is, there is approximately a 95 % chance that the interval

$$[\hat{\beta}_0 - 2\,\mathrm{SE}(\hat{\beta}_0)\ ,\ \hat{\beta}_0 + 2\,\mathrm{SE}(\hat{\beta}_0)]$$

will contain the true value of β_0

- Similarly, a confidence interval for β_1 approximately takes the form

$$[\hat{\beta}_1 - 2\,\mathrm{SE}(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + 2\,\mathrm{SE}(\hat{\beta}_1)]$$

will contain the true value of β_1

- The word 'approximately' is included mainly because
  - ✓ The errors are assumed to be Gaussian and
  - ✓ The factor '2' in front of $\mathrm{SE}(\hat{\beta}_1)$ term will vary slightly depending on the number of observations 'n' in the linear regression

Example: Confidence Interval for Regression Coefficient in R

Suppose we'd like to fit a simple linear regression model using **hours studied** as a predictor variable and **exam score** as a response variable for 15 students in a particular class:

Hours Studied	Exam Score
1	64
2	66
4	76
5	73
5	74
6	81
6	83
7	82
8	80
10	88
11	84
11	82
12	91
12	93
14	89

We can use the lm() function to fit this simple linear regression model in R:

```
#create data frame
df <- data.frame(hours=c(1, 2, 4, 5, 5, 6, 6, 7, 8, 10, 11, 11, 12, 12, 14),
                 score=c(64, 66, 76, 73, 74, 81, 83, 82, 80, 88, 84, 82, 91, 93, 89))

#fit linear regression model
fit <- lm(score ~ hours, data=df)

#view model summary
summary(fit)
```

```
Call:
lm(formula = score ~ hours, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-5.140 -3.219 -1.193  2.816  5.772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   65.334      2.106  31.023 1.41e-13 ***
hours          1.982      0.248   7.995 2.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.641 on 13 degrees of freedom
Multiple R-squared:  0.831,     Adjusted R-squared:  0.818
F-statistic: 63.91 on 1 and 13 DF,  p-value: 2.253e-06
```

We can use the **confint()** function to calculate a 95% confidence interval for the regression coefficient:

```
#calculate confidence interval for regression coefficient for 'hours'
confint(fit, 'hours', level=0.95)
```

```
          2.5 %   97.5 %
hours 1.446682 2.518068
```

Since this confidence interval doesn't contain the value 0, we can conclude that there is a statistically significant association between hours studied and exam score.

We can also confirm this is correct by calculating the 95% confidence interval for the regression coefficient by hand:

Alpha=0.05

- 95% C.I. for β_1: $b_1 \pm t_{1-\alpha/2,\, n-2} * se(b_1)$
- 95% C.I. for β_1: $1.982 \pm t_{.975,\, 15-2} * .248$
- 95% C.I. for β_1: $1.982 \pm 2.1604 * .248$
- 95% C.I. for β_1: [1.446, 2.518]

The 95% confidence interval for the regression coefficient is **[1.446, 2.518]**.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291

RSquared

- The RSE provides an absolute measure of lack of fit of the model to the data. A small RSE indicates that the model fits the data well whereas a large RSE indicates that the model doesn't fit the data well. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE
- The **R^2 statistic** provides an alternative measure of fit. It takes the form of a proportion of variance, expressed as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares*.

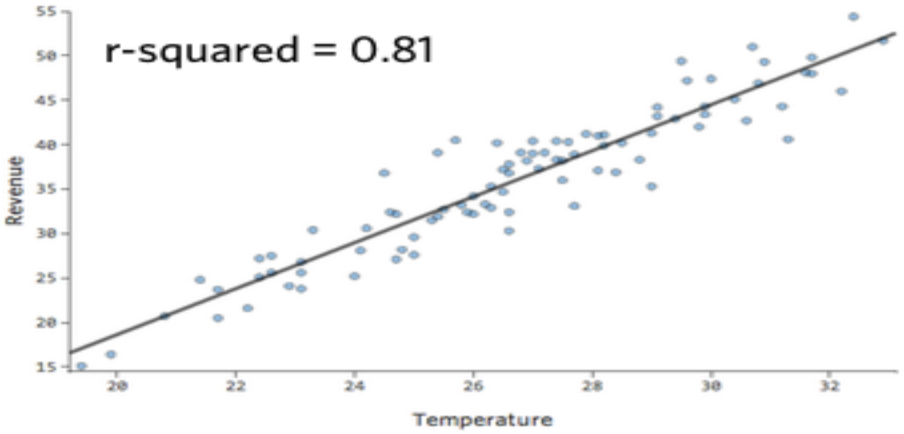
- Note that R^2 statistic is independent of the scale of Y, and it always **takes a value between 0 and 1**

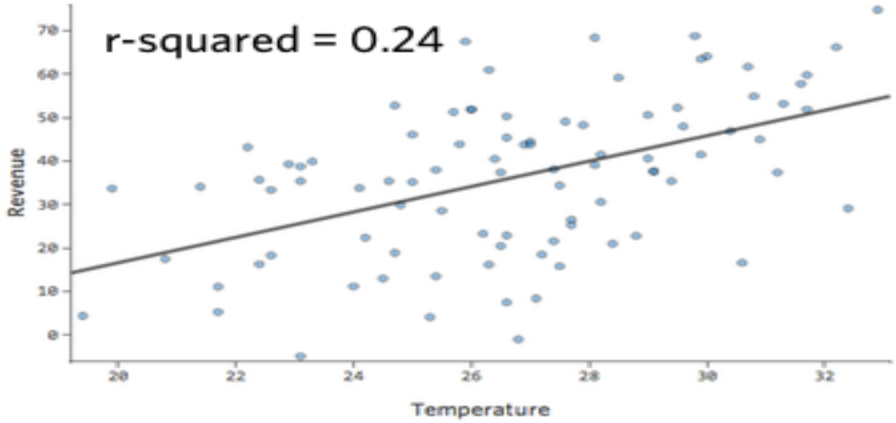
R square

- TSS $= \sum_{i=1}^{n} (y_i - \bar{y})^2$ measures the total variance in the response variable Y, and can be interpreted as the amount of variability inherent in the response before the regression is performed.
- TSS - RSS $= \sum_{i=1}^{n} \{(y_i - \bar{y})^2 - (y_i - \hat{y}_i)^2\}$ measures the amount of variability in the response that is removed by performing the regression, and therefore R^2 measures the proportion of variability in Y that can be explained using X.

- An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been taken care by the regression.
- The **R^2 statistic** is also a measure of the linear relationship between X and Y and it is closely related to **correlation between X and Y**

- For example, if R^2 =0.8, then 80% of variance in the data is explained by the model.





Residual Plot:

For regression, there are numerous methods to evaluate the goodness of your fit i.e. how well the model fits the data. One such method is residual plot.

A typical residual plot has the residual values on the Y-axis and the independent variable on the x-axis.

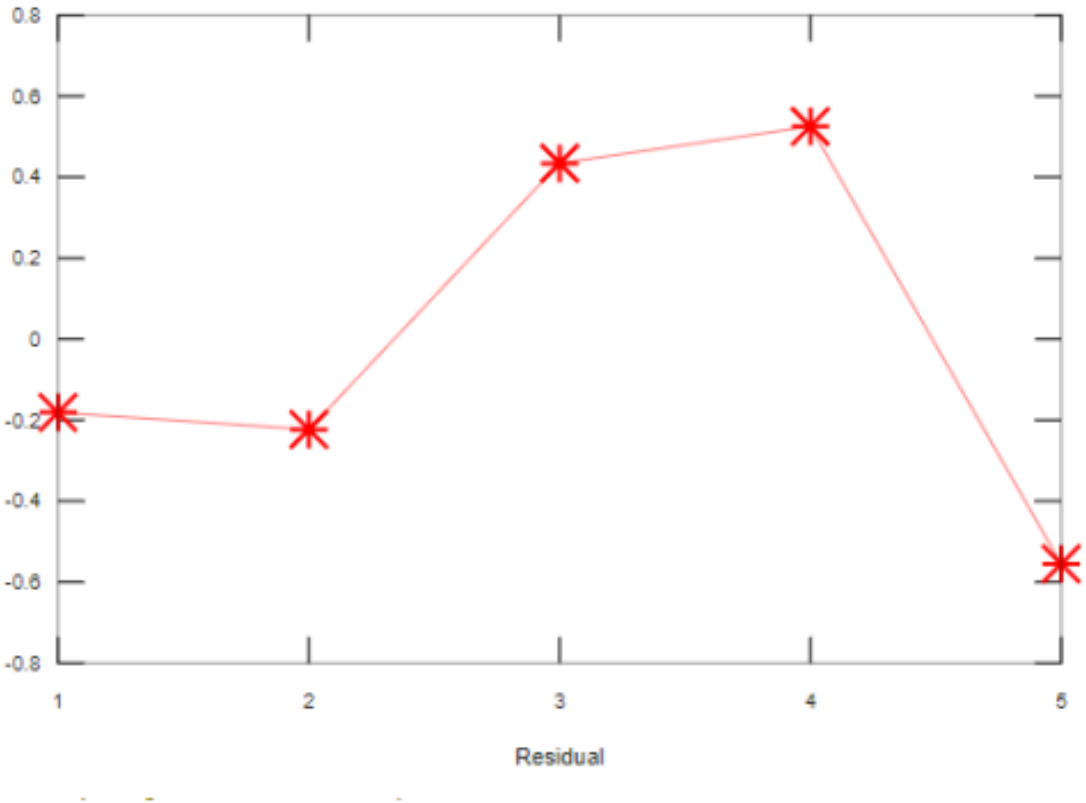
If the points are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Solution:

Residual plot for the best linear fit

X	Y	$Y_{predicted}$	Residual $(Y-Y_{Predicted})$
2	12.8978	13.0789	-0.1811
3	17.7586	17.9818	-0.2232
4	23.3192	22.8847	0.4345
5	28.3129	27.7876	0.5253
6	32.1351	32.6905	-0.5554

Residual Plot



The random pattern in it is an indication that a linear model is suitable for this data

Model Estimation and Evaluation

Question-4:

Consider the following five training examples

X = [2 3 4 5 6]

Y = [12.8978 17.7586 23.3192 28.3129 32.1351]

We want to learn a function f(x) of the form f(x) = ax + b which is parameterized by (a, b).

(a) Find the best linear fit

(b) Evaluate the standard errors associated with $\hat{a}$ and $\hat{b}$.

(c) Determine the 95% confidence interval for a and b

(d) Compute R^2 statistic

Simple Linear Regression

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$
	2	12.8978	-2	-9.9869	19.9738	4
	3	17.7586	-1	-5.1261	5.1261	1
	4	23.3192	0	0.4345	0.0000	0
	5	28.3129	1	5.4282	5.4282	1
	6	32.1351	2	9.2504	18.5008	4
Sum	20	114.4236	0	0.0000	49.0289	10
Mean	4	22.88472				

The best linear fit is

$Y = 3.2732 + 4.9029X$

Substituting in the formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_1 = \mathbf{4.9029}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \qquad \hat{\beta}_0 = \mathbf{3.2732}$$

Solution:

	X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
	2	12.8978	-2	-9.9869	19.9738	4	13.0789	0.0328
	3	17.7586	-1	-5.1261	5.1261	1	17.9818	0.0498
	4	23.3192	0	0.4345	0.0000	0	22.8847	0.1888
	5	28.3129	1	5.4282	5.4282	1	27.7876	0.2759
	6	32.1351	2	9.2504	18.5008	4	32.6905	0.3085
Sum	20	114.4236	0	0.0000	49.0289	10	**RSS**	**0.8558**
Mean	4	22.88472						

Y predicted is calculated using the best linear fit

Y = 4.9029 + 3.2732 X

RSS_{min} = 0.8558

Model Estimation and Evaluation

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}.$$

Substituting **RSS = 0.8558** and n = 5, then **RSE = 0.5341**.

Standard error for a is

σ = RSE

$$\text{SE(a)} = \sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \quad = 0.1689$$

Standard error for b is

$$\text{SE(b)} = \sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad = 0.7186$$

95% confidence interval for standard error for a is

$$[a - 2\,SE(a)\ ,\ a + 2\,SE(a)] = [4.5651,\ 5.2407]$$

95% confidence interval for standard error for b is

$$[b - 2\,SE(b)\ ,\ b + 2\,SE(b)] = [1.8400,\ 4.7063]$$

Calculate a t-statistic by dividing the estimated coefficient by its standard error.

	X	Y	$(Y-Y_{mean})^2$
	2	12.8978	99.73857
	3	17.7586	26.27711
	4	23.3192	0.188773
	5	28.3129	29.46514
	6	32.1351	85.56953
Sum	20	114.4236	241.2391
Mean	4	22.88472	

To find R^2 value, first find TSS

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 241.2391$$

$$R^2 = 1 - \frac{RSS}{TSS} = 0.9965$$

Question-2:

Consider the following five training examples

X = [2 3 4 5 6]

Y = [12.8978 17.7586 23.3192 28.3129 32.1351]

We want to learn a function f(x) of the form f(x) = ax + b which is parameterized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (4 3)

(b) (5 3)

(c) (5 1)

(d) (1 5)

Solution:

For a = 4 and b = 3

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	11	3.6016
3	17.7586	15	7.6099
4	23.3192	19	18.6555
5	28.3129	23	28.2269
6	32.1351	27	26.3693
		RSS	**84.4632**

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

Solution:

For a = 5 and b = 3

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	13	0.0104
3	17.7586	18	0.0583
4	23.3192	23	0.1019
5	28.3129	28	0.0979
6	32.1351	33	0.7481
		RSS	**1.0166**

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

Solution:

For a = 5 and b = 1

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	11	3.6016
3	17.7586	16	3.0927
4	23.3192	21	5.3787
5	28.3129	26	5.3495
6	32.1351	31	1.2885
		RSS	**18.7110**

Formula

$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

Solution:

For a = 1 and b = 5

Formula

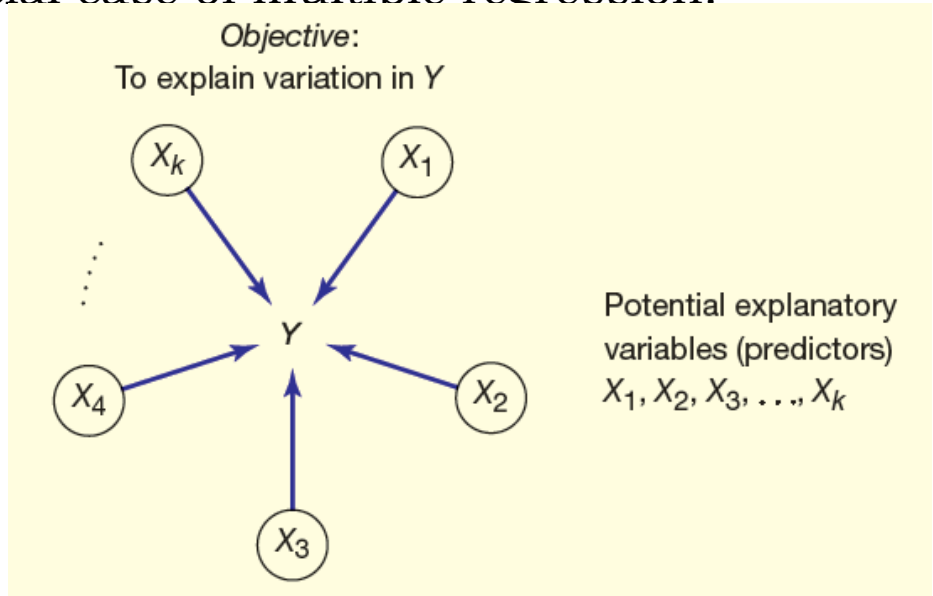
$$RSS = \sum_{i}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

X	Y	$Y_{predicted}$	$(Y-Y_{Predicted})^2$
2	12.8978	7	34.7840
3	17.7586	8	95.2303
4	23.3192	9	205.0395
5	28.3129	10	335.3623
6	32.1351	11	446.6925
		RSS	**1117.1086**

Answer: The parameter (5,3) which gives least RSS (1.016). Hence (5,3) is used to model this function

Multiple Regression

- **Multiple regression** extends simple regression to include several independent variables (called *predictors*).
- Multiple regression is required when a single-predictor model is inadequate to describe the true relationship between the dependent variable *Y* (the response variable) and its potential predictors (*X*1, *X*2, *X*3, . . .).
- The interpretation of multiple regression is similar to simple regression because simple regression is a special case of multiple regression.



Regression Terminology

- The **response variable** (Y) is assumed to be related to the k **predictors** ($X_1, X_2, \ldots, X_k$) by a linear equation called the *population regression model:*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

- A *random error* ε represents everything that is not part of the model. The unknown regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are *parameters* and are denoted by Greek letters

Regression Terminology

- The *sample estimates* of the regression coefficients are denoted by Roman letters b_0, b_1, b_2, ... , b_k . The *predicted* value of the response variable is denoted $\hat{y}$ and is calculated by inserting the values of the predictors into the *estimated regression equation:*

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots. + b_kx_k$$ (predicted value of Y)

Data Format

- To obtain a fitted regression, we need n observed values of the response variable Y and its proposed predictors X_1, X_2, . . . , X_k . A multivariate data set is a single column of Y-values and k columns of X-values.

Response	*Predictors*			
Y	X_1	X_2	…	X_k
y_1	x_{11}	x_{12}	…	x_{1k}
y_2	x_{21}	x_{22}	…	x_{2k}
⋮	⋮	⋮	⋮	⋮
y_n	x_{n1}	x_{n2}	…	x_{nk}

Example

Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

https://www.statology.org/multiple-linear-regression-

$$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$$

The formula to calculate b_1 is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] \ / \ [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] \ / \ [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

The formula to calculate b_0 is: $\overline{y} - b_1\overline{X}_1 - b_2\overline{X}_2$

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation: $\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$

b_0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

b_1 = 3.148. A one unit increase in x_1 is associated with a 3.148 unit increase in y, on average, assuming x_2 is held constant.

b_2 = -1.656. A one unit increase in x_2 is associated with a 1.656 unit decrease in y, on average, assuming x_1 is held constant.

Sample Data

- **Illustration: Home Prices**
- Table shows sales of 30 new homes in an upscale development. Although the selling price of a home (the *response variable*) may depend on many factors, we will examine three potential *explanatory variables*.
- *Definition of Variable Short Name*
- Y = selling price of a home (thousands of dollars) *Price*
- X_1 = home size (square feet) *SqFt*
- X_2 = lot size (thousand square feet) *LotSize*
- X_3 = number of bathrooms *Baths*

$$Price = \beta_0 + \beta_1 SqFt + \beta_2 LotSize + \beta_3 Baths + \varepsilon$$

Sample Data

Home	Price	SqFt	LotSize	Baths	Home	Price	SqFt	LotSize	Baths
1	505.5	2,192	16.4	2.5	16	675.1	3,076	19.8	3.0
2	784.1	3,429	24.7	3.5	17	710.4	3,259	20.8	3.5
3	649.0	2,842	17.7	3.5	18	674.7	3,162	19.4	4.0
4	689.8	2,987	20.3	3.5	19	663.6	2,885	23.2	3.0
5	709.8	3,029	22.2	3.0	20	606.6	2,550	20.2	3.0
6	590.2	2,616	20.8	2.5	21	758.9	3,380	19.6	4.5
7	643.3	2,978	17.3	3.0	22	723.7	3,131	22.5	3.5
8	789.7	3,595	22.4	3.5	23	621.8	2,754	19.2	2.5
9	683.0	2,838	27.4	3.0	24	622.4	2,710	21.6	3.0
10	544.3	2,591	19.2	2.0	25	631.3	2,616	20.8	2.5
11	822.8	3,633	26.9	4.0	26	574.0	2,608	17.3	3.5
12	637.7	2,822	23.1	3.0	27	863.8	3,572	29.0	4.0
13	618.7	2,994	20.4	3.0	28	652.7	2,924	21.8	2.5
14	619.3	2,696	22.7	3.5	29	844.2	3,614	25.5	3.5
15	490.5	2,134	13.4	2.5	30	629.9	2,600	24.1	3.5

Sample Data

Home	X1 Sqft	X2 LotSize	X3 Baths	Y Price
1	2192	16.4	2.5	505.5
2	3429	24.7	3.5	784.1
3	2842	17.7	3.5	649
4	2987	20.3	3.5	689.8
5	3029	22.2	3	709.8
6	2616	20.8	2.5	590.2
7	2978	17.3	3	643.3
8	3595	22.4	3.5	789.7
9	2838	27.4	3	683
10	2591	19.2	2	544.3
11	3633	26.9	4	822.8
12	2822	23.1	3	637.7
13	2994	20.4	3	618.7
14	2696	22.7	3.5	619.3
15	2134	13.4	2.5	490.5
16	3076	19.8	3	675.1
17	3259	20.8	3.5	710.4
18	3162	19.4	4	674.7
19	2885	23.2	3	663.6
20	2550	20.2	3	606.6
21	3380	19.6	4.5	758.9
22	3131	22.5	3.5	723.3
23	2754	19.2	2.5	621.8
24	2710	21.6	3	622.4
25	2616	20.8	2.5	631.3
26	2608	17.3	3.5	574
27	3572	29	4	863.8
28	2924	21.8	2.5	652.7
29	3614	25.5	3.5	844.2
30	2600	24.1	3.5	629.9

Estimation of Regression

- Intercept = -28.85
- Slope Sqft = 0.171
- Slope LotSize = 6.78
- Slope Baths = 15.53

$$Price = -28.85 + 0.171\ SqFt + 6.78\ LotSize + 15.53\ Baths\ (R^2 = .956,\ SE = 20.31)$$

Predictions from a Fitted Regression

- We can use the fitted regression model to make predictions for various assumed predictor values. For example, what would be the expected selling price of a 2,800-square-foot home with 2½ baths on a lot with 18,500 square feet?

$$SqFt = 2800 \qquad LotSize = 18.5 \qquad Baths = 2.5$$

$$Price = -28.85 + 0.171\,(2800) + 6.78\,(18.5) + 15.53\,(2.5) = 614.23, \text{ or } \$614{,}230$$

Model accuracy

The overall quality of the linear regression fit can be assessed using the following three quantities, displayed in the model summary:

- The Residual Standard Error (RSE).
- The R-squared (R2) (coefficient of determination
- F-statistic

R-squared and Adjusted R-squared

- The R-squared (R2) ranges from 0 to 1 and represents the proportion of information (i.e. variation) in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom.

- The R2 measures, how well the model fits the data. For a simple linear regression, R2 is the square of the Pearson correlation coefficient.

- The RSE provides an absolute measure of lack of fit of the model to the data. A small RSE indicates that the model fits the data well whereas a large RSE indicates that the model doesn't fit the data well. But since it is measured in the units of Y, it is not always clear what constitutes a good RSE
- The **R^2 statistic** provides an alternative measure of fit. It takes the form of a proportion of variance, expressed as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the *total sum of squares*.

- Note that R^2 statistic is independent of the scale of Y, and it always **takes a value between 0 and 1**

R-squared and Adjusted R-squared

- A high value of R2 is a good indication.
- However, as the value of R2 tends to increase when more predictors are added in the model, such as in multiple linear regression model, you should mainly consider the adjusted R-squared, which is a penalized R2 for a higher number of predictors.
  - An (adjusted) R2 that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
  - A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

Rsquared Adjusted

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

F-Statistic

- The F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient.

.

- **p-value** is the probability your results could have happened by chance.
- **A big F, with a small p-value, means that the null hypothesis is discredited, and we would assert that there is a general relationship between the response and predictors** (while a small F, with a big p-value indicates that there is no relationship).

Statistical hypotheses

- Null hypothesis (H0): the coefficients are equal to zero (i.e., no relationship between x and y)
- Alternative Hypothesis (Ha): the coefficients are not equal to zero (i.e., there is some relationship between x and y)

- State the null and alternative hypothesis:

H_0: $\beta_1 = \beta_2 = , \ldots , = \beta_{p-1} = 0$

H_1: $\beta_j \neq 0$ for some j

F = MSM / MSE = (explained variance) / (unexplained variance)

- **Mean of Squares for Model:** MSM = SSM / DFM

- **Mean of Squares for Error:** MSE = SSE / DFE
- Degrees of Freedom for Model: DFM = p
- Degrees of Freedom for Error: DFE = n – p -1
-

Corrected Sum of Squares for Model: $SSM = \sum_{i=1}^{n} (y_i\hat{} - \bar{y})^2$,
also called sum of squares for regression.

Sum of Squares for Error: $SSE = \sum_{i=1}^{n} (y_i - y_i\hat{})^2$,
also called sum of squares for residuals.

X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$	$Y_{predicted}$	$(Y-Y_{Predicted})^2$	$(Y-Y_{mean})^2$	$(Ypred-Y_{mean})$	$(Ypred-Y_{mean})^2$
2	12.8978	-2	-9.9869	19.9738	4	13.07894	0.0328117	99.73817161	-9.80576	96.1529292
3	17.7586	-1	-5.1261	5.1261	1	17.98183	0.04983163	26.27690121	-4.90287	24.0381342
4	23.3192	0	0.4345	0	0	22.88472	0.18877287	0.18879025	2E-05	4E-10
5	28.3129	1	5.4282	5.4282	1	27.78761	0.27592958	29.46535524	4.90291	24.0385265
6	32.1351	2	9.2504	18.5008	4	32.6905	0.30846916	85.56990016	9.8058	96.1537136
4	22.88472			49.0289	10		0.85581495	241.2391185		240.383304
X_{mean}	Y_{mean}						RSS or SSE	TSS		SSM

- Calculate Rsquared, Adj Rsquared and F statistic

X	Y	$(X-X_{mean})$	$(Y-Y_{mean})$	$(X-X_{mean})(Y-Y_{mean})$	$(X-X_{mean})^2$	$Y_{predicted}$	$(Y-Y_{Predicted})^2$	$(Y-Y_{mean})^2$	$(Ypred-Y_{mean})$	$(Ypred-Y_{mean})^2$
2	12.8978	-2	-9.9869	19.9738	4	13.07894	0.0328117	99.73817161	-9.80576	96.1529292
3	17.7586	-1	-5.1261	5.1261	1	17.98183	0.04983163	26.27690121	-4.90287	24.0381342
4	23.3192	0	0.4345	0	0	22.88472	0.18877287	0.18879025	2E-05	4E-10
5	28.3129	1	5.4282	5.4282	1	27.78761	0.27592958	29.46535524	4.90291	24.0385265
6	32.1351	2	9.2504	18.5008	4	32.6905	0.30846916	85.56990016	9.8058	96.1537136
4	22.88472			49.0289	10		0.85581495	241.2391185		240.383304
Xmean	Ymean						RSS or SSE	TSS		SSM

- Calculate Rsquared, Adj Rsquared and F statistic

```
Residual standard error: 0.5341 on 3 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9953
F-statistic: 842.6 on 1 and 3 DF,  p-value: 8.977e-05
```

of the *F* Distribution

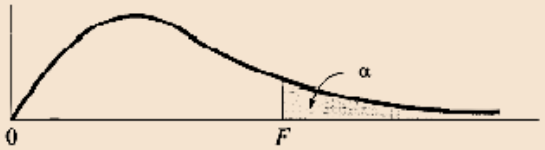


Table 1 $\alpha = 0.05$

Degrees of Freedom for Denominator	Degrees of Freedom for Numerator															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.73	1.69	1.63	1.60

of the *F* Distribution

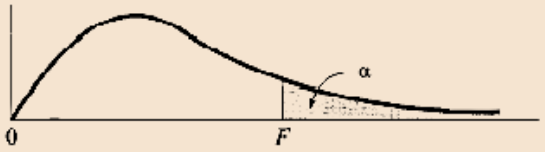


Table 1 $\alpha = 0.05$

Degrees of Freedom for Denominator	Degrees of Freedom for Numerator															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66

F_statistic > F_critical

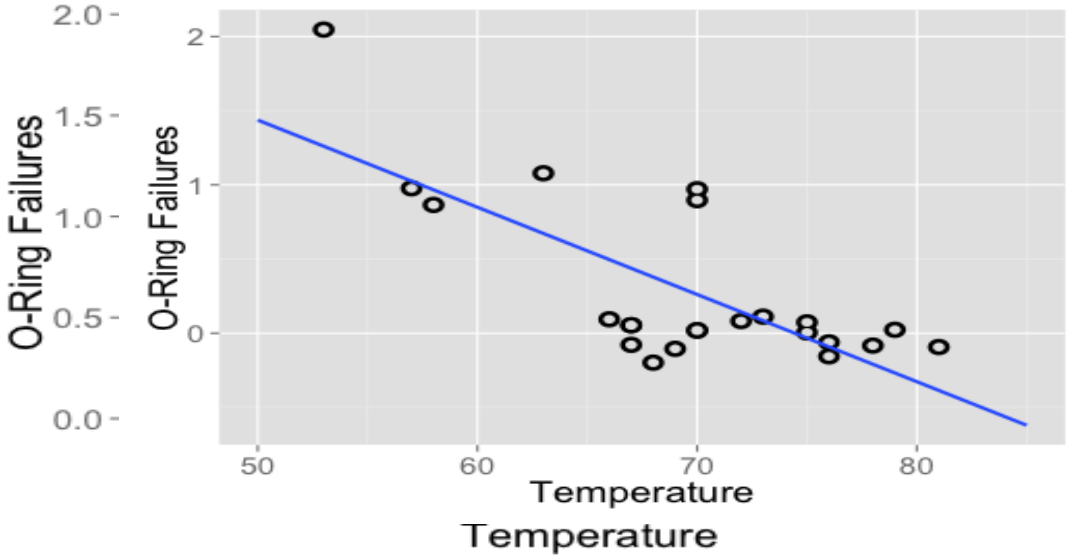
F-statistic: 842.6 on 1 and 3 DF

Null Hypothesis is rejected

Correlation

What is correlation?

- What is the pattern of the data points that you observe?
  - Line, curve or no pattern at all

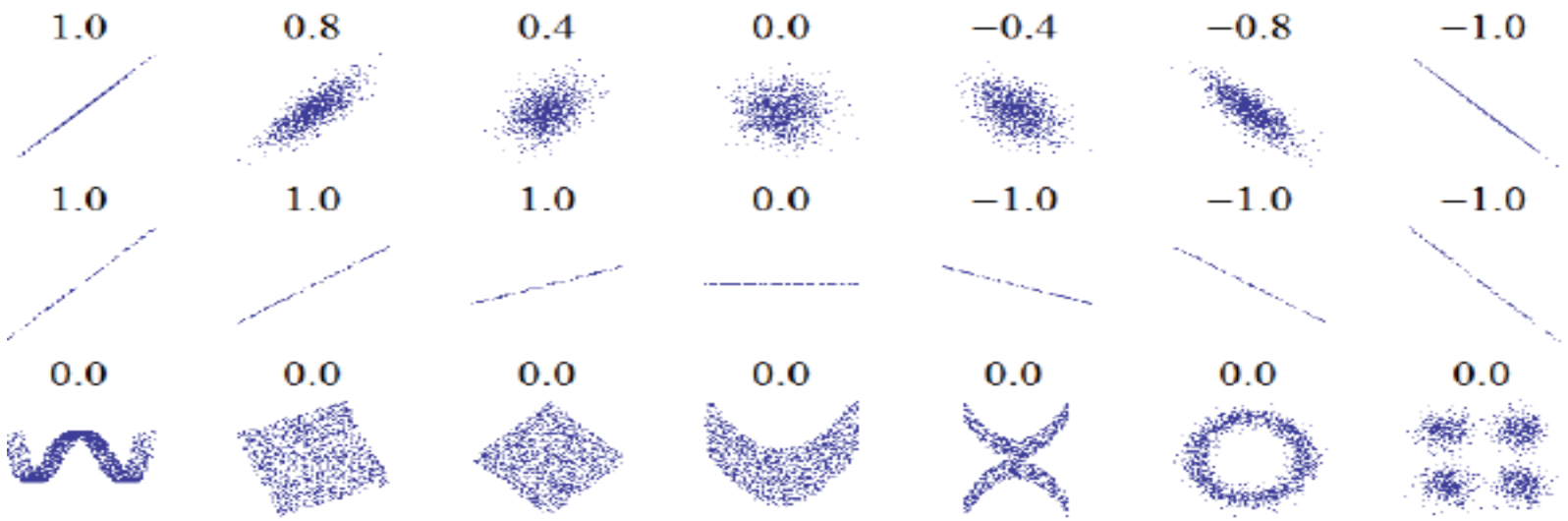


Could observe a line pattern. Ie., it follows a linear pattern

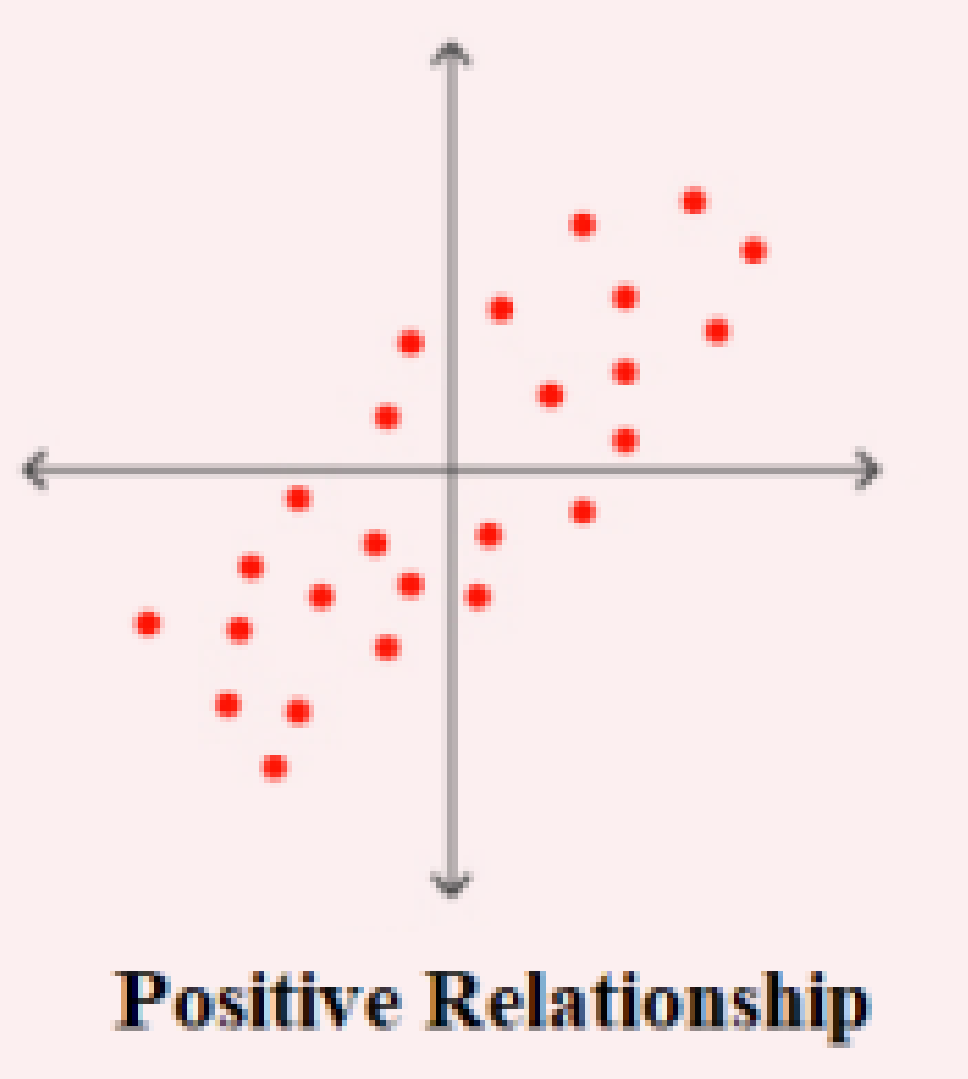
When temperature increases, no.of O-Ring failures decreases

Correlation

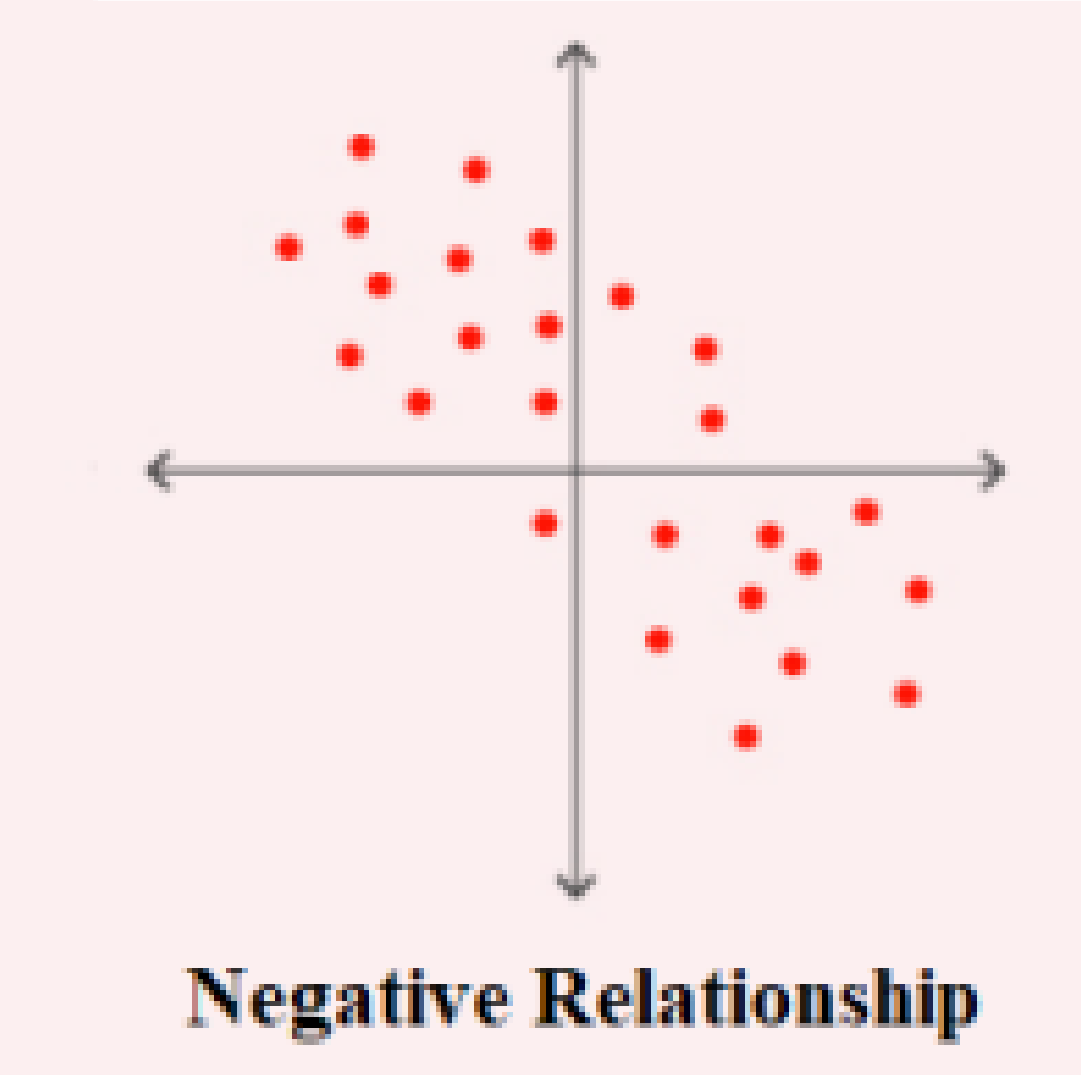
- Bivariate analysis that measures the strength of association between two variables and the direction of the relationship.



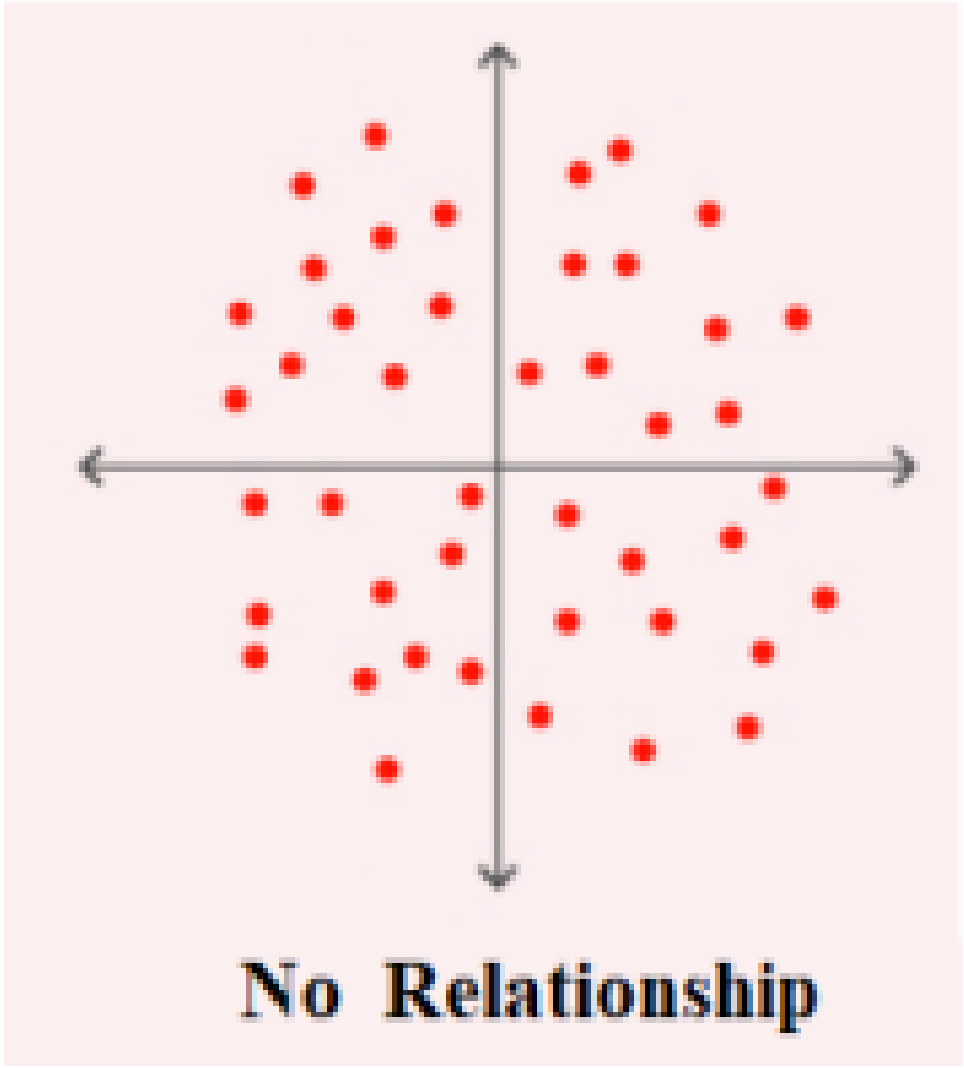
- A correlation is a relationship between two variables.
- Is there a relationship between the number of employee training hours and the number of jobs produced?
- Is there a relationship between the number of hours a student spends studying for a Mathematics test and the student's score on that test?
- Let x to be the independent variable and y to be the dependent variable. Data is represented by a collection of ordered pairs (x, y)
- Mathematically, the strength and direction of a linear relationship between two variables is represented by the correlation coefficient.



- If r is close to 1, the variables are positively correlated ➔ there is likely a strong linear relationship between the two variables, with a positive slope.



- If r is close to -1, the variables are negatively correlated ➔ there is likely a strong linear relationship between the two variables, with a negative slope.

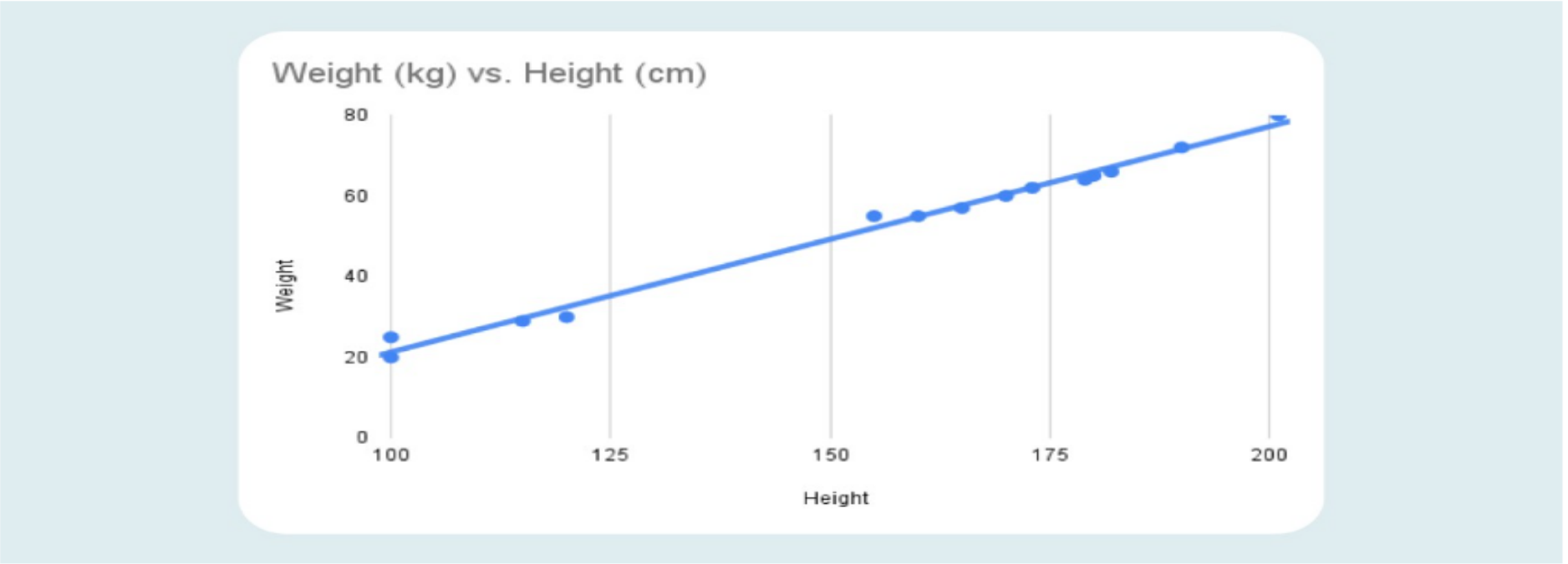


- If r is close to 0, the variables are not correlated ➔ that there is likely no linear relationship between the two variables, however, the variables may still be related in some other way.

Types of correlation

Positive correlation

Two variables are considered to have a positive correlation if they are directly proportional. That is, if the value of one variable increases, then the value of the other variable will also increase. A perfect positive correlation holds a value of “1”. On a graph, positive correlation appears as follows:

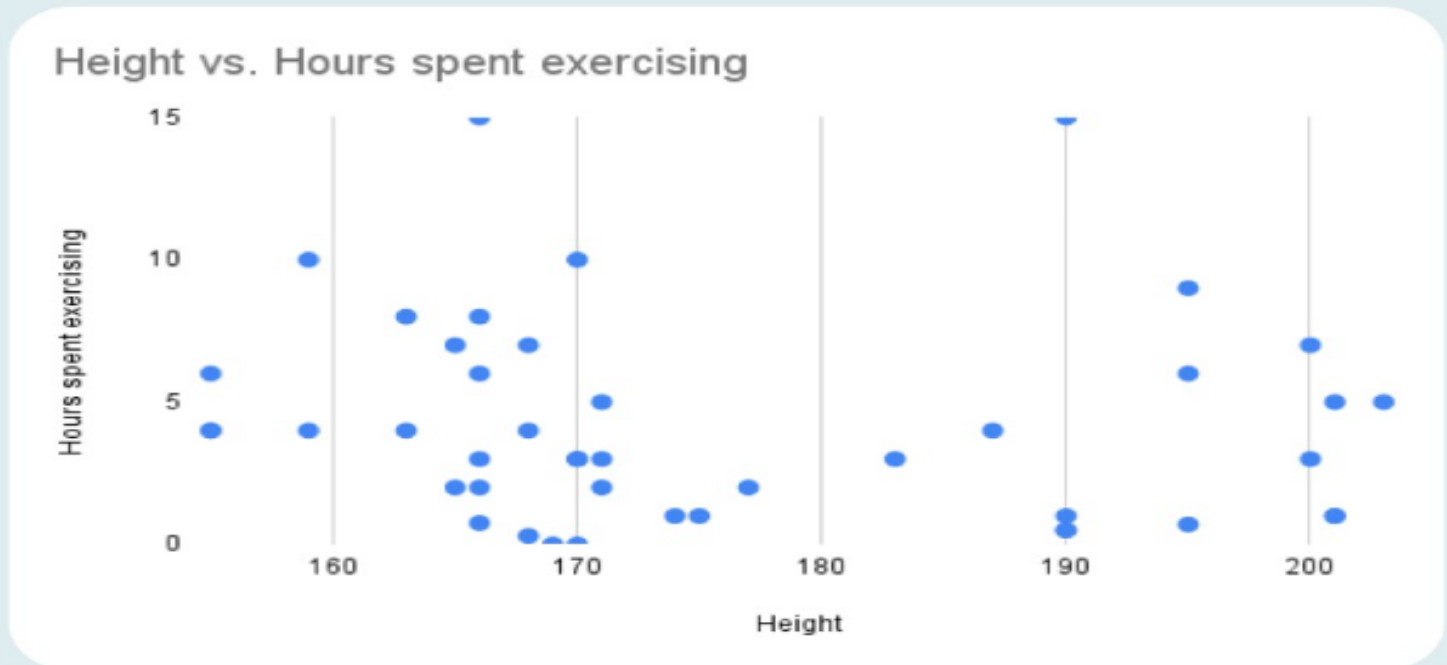


Negative correlation

A perfect negative correlation holds a value of "-1" which means that, as the value of one variable increases, the value of the second variable decreases (and vice versa). In graph form, this is how negative correlation might look:

Zero or no correlation

The value “0” denotes that there is no correlation. It indicates that there is no relationship between the two variables, so an increase or decrease in one variable is unrelated to an increase or decrease in the other variable. A graph showing zero correlation will follow a random distribution of data points, as opposed to a clear line:

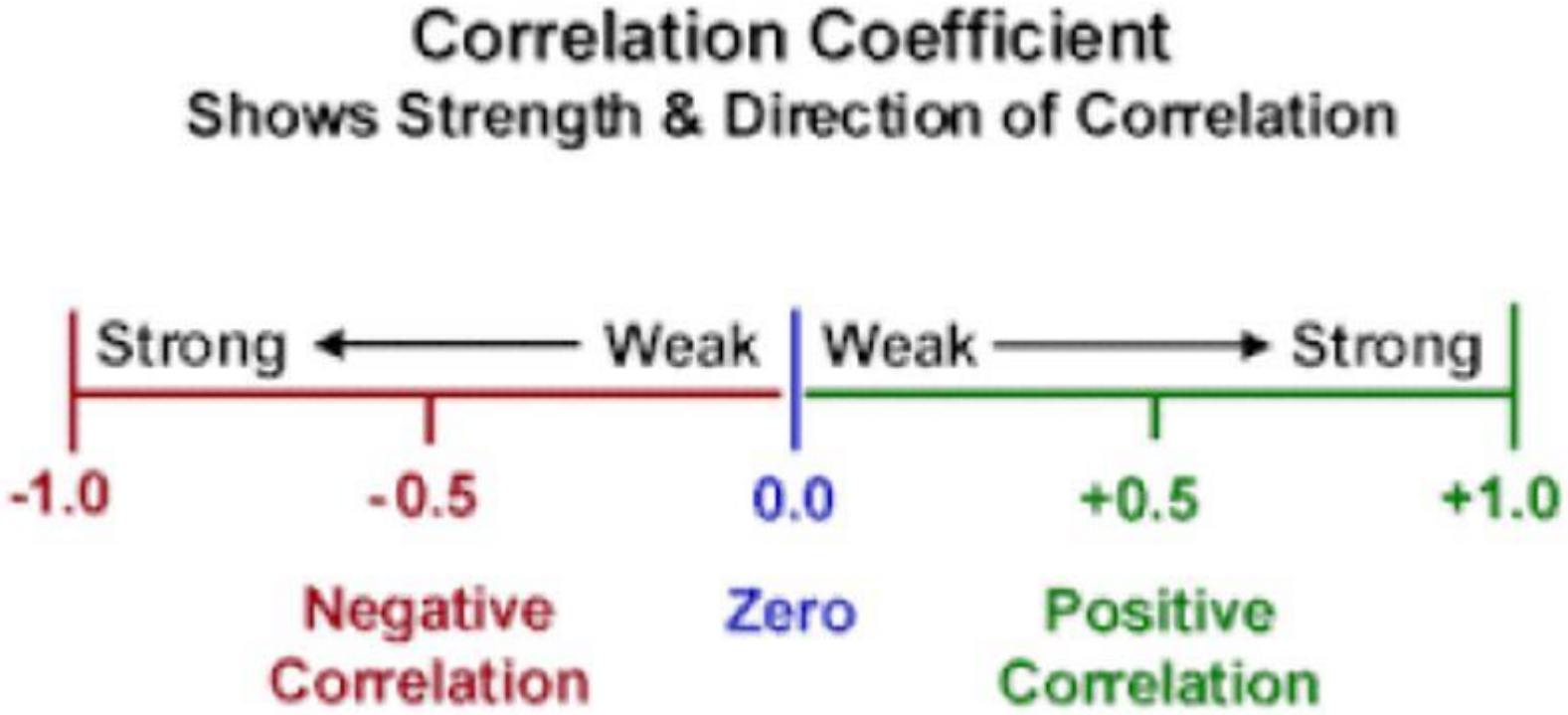


What is a correlation matrix

A correlation matrix is essentially a table depicting the correlation coefficients for various variables. The rows and columns contain the value of the variables, and each cell shows the correlation coefficient.

	Hours spent exercising	Cardio fitness level	Height	Age
Hours spent exercising	1	0.82	0.03	-0.44
Cardio fitness level	0.82	1	0.2	-0.05
Height	0.03	0.2	1	0.1
Age	-0.44	-0.05	0.1	1

Correlation



Correlation

More height more weight

More hours into study more marks

More the age of the car lesser the value

Question to Ponder

- Do you recollect any other bivariate analysis that measures the linear relationship between two variables?

Covariance Vs. Correlation

Covariance

- Provides the direction of linear relationship between two variables
- Has no upper or lower bound
- Not standardized

Correlation

- Provides both direction and the strength of the linear relationship between two variables
- Ranges between -1 to +1
- Standardized

Covariance Formula

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable x and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

- A correlation is a relationship between two variables.
- Is there a relationship between the number of employee training hours and the number of jobs produced?
- Is there a relationship between the number of hours a student spends studying for a Mathematics test and the student's score on that test?
- Let x to be the independent variable and y to be the dependent variable. Data is represented by a collection of ordered pairs (x, y)
- Mathematically, the strength and direction of a linear relationship between two variables is represented by the correlation coefficient.

Methods for Correlation Analysis

- **Pearson correlation (r)**
  - measures a linear dependence between two variables (x and y).

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

  - It's also known as a **parametric correlation** test because it depends to the distribution of the data.
  - It can be used only when x and y are from normal distribution.
- **Kendall (tau)** and **Spearman (rho)**
  - rank-based correlation coefficients (non-parametric)

- The correlation coefficient r is given by

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

- **This will always be a number between -1 and 1 (inclusive).**

Correlation coefficient can be computed using the functions **cor()** or **cor.test()**:

- **cor()** computes the **correlation coefficient**
- **cor.test()** test for association/correlation between paired samples. It returns both the **correlation coefficient** and the **significance level**(or p-value) of the correlation .

cor(x, y, method = "pearson“)

Method="pearson", "kendall", "spearman"

If your data contain missing values, use the following R code to handle missing values by case-wise deletion.

```
cor(x, y, method = "pearson", use = "complete.obs")
```

Preliminary test to check the test assumptions

- **Is the relationship linear**?
- **Are the data from each of the 2 variables (x, y) follow a normal distribution**?
  - Use Shapiro-Wilk normality test –> R function: **shapiro.test**()
  - and look at the normality plot —> R function: **qqplot**()

Shapiro-Wilk test can be performed as follow:
Null hypothesis: the data are normally distributed
Alternative hypothesis: the data are not normally distributed

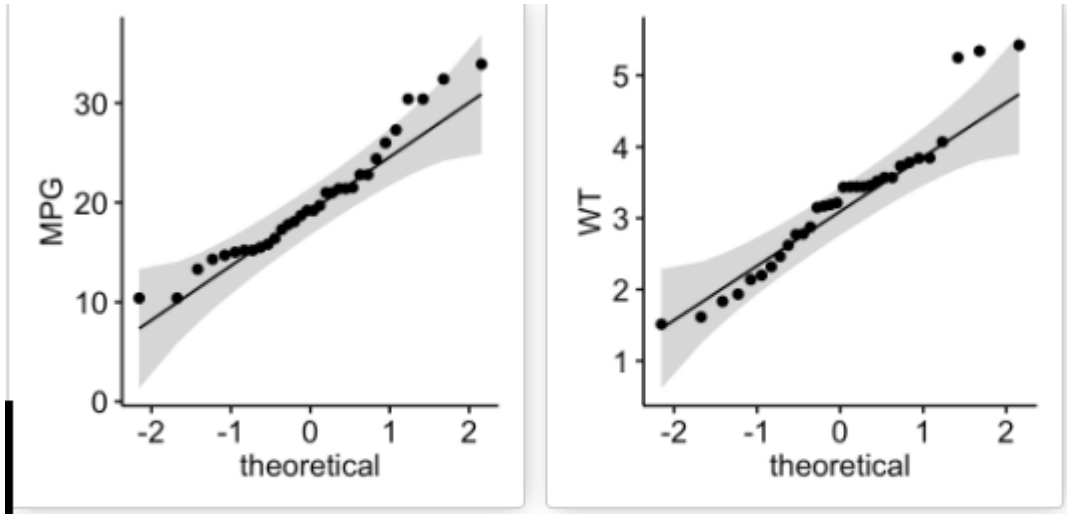
```
# Shapiro-Wilk normality test for mpg
shapiro.test(my_data$mpg) # => p = 0.1229

# Shapiro-Wilk normality test for wt
shapiro.test(my_data$wt) # => p = 0.09
```

QQ Plot

QQ plot (or quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted. QQ plots are used to visually check the normality of the data.

If all the points fall approximately along this reference line, we can assume normality.

```
install.packages("ggpubr")
library("ggpubr")
# mpg
ggqqplot(data$mpg, ylab = "MPG")
# wt
ggqqplot(data$wt, ylab = "WT")
```



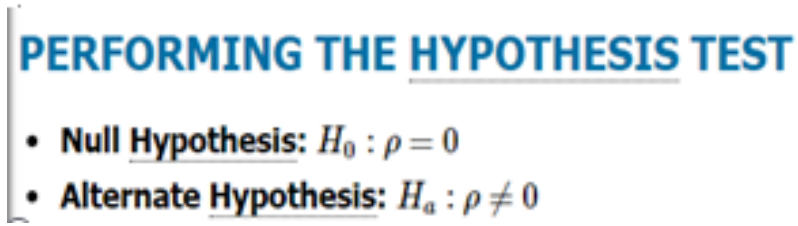
From the normality plots, we conclude that both populations may come from normal distributions.

Pearson correlation test

```
res <- cor.test(data$mpg,data$wt,method = "pearson")
res
```

```
Pearson's product-moment correlation
data: data$mpg and data$wt
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: -0.9338264 -0.7440872
sample estimates: cor -0.8676594
```

In the result above :
t is the **t-test statistic** value (t = -9.559),
df is the degrees of freedom (df= 30),
p-value is the significance level of the **t-test** (p-value = 1.29410^{-10}).
conf.int is the **confidence interval** of the correlation coefficient at 95% (conf.int = [-0.9338, -0.7441]);
sample estimates is the correlation coefficient (Cor.coeff = -0.87).



Question:

- The time x in years that an employee spent at a company and the employee's hourly pay, y, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient r

x	y
5	25
3	20
4	21
10	35
15	38

x	y	x^2	y^2	xy
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570
$\sum x = 37$	$\sum y = 139$	$\sum x^2 = 375$	$\sum y^2 = 4135$	$\sum xy = 1189$

Hint: Calculate the numerator:

$$n\sum(xy) - \left(\sum x\right)\left(\sum y\right) = 5 \cdot 1189 - 37 \cdot 139 = 802$$

Then calculate the denominator:

$$\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2} = \sqrt{5 \cdot 375 - (37)^2}\sqrt{5 \cdot 4135 - (139)^2}$$
$$= \sqrt{506}\sqrt{1354} \approx 827.72$$

Now, divide to get $r \approx \frac{802}{827.72} \approx 0.97$.

- **Interpret this result:** There is a strong positive correlation between the number of years and employee has worked and the employee's salary, since r is very close to 1

ANOVA

One-way ANOVA test

- ANOVA – Analysis of Variance
- One-way ANOVA, also known as one-factor ANOVA is a test for comparing means of more than two groups
- ANOVA is a hypothesis testing procedure that is used to evaluate differences between 2 or more samples
- It check whether the means of two or more sample groups are statistically different or not.
- ANOVA test hypotheses:
  - Null hypothesis: the means of the different groups are the same
  - Alternative hypothesis: At least one sample mean is not equal to the others.

Assumptions of ANOVA test

- The observations are obtained independently and randomly from the population defined by the factor levels
- The data of each factor level are normally distributed.
- These normal populations have a common variance. (**Levene's test** can be used to check this.)

How it works?

- Assume that we have 3 groups (A, B, C) to compare:
  - Compute the common variance, which is called **variance within samples** (S^2within) or **residual variance**.
  - Compute the variance between sample means as follow:
    - Compute the mean of each group
    - Compute the **variance between sample means** (S^2between)
  - Produce F-statistic as the ratio of S^2between / S^2within.

Low Calorie	Low Fat	Low Carbohydrate	Control
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

Is there a statistically significant difference in the mean weight loss among the four diets?

18 employees (six each from first year to third year of experience) were selected for an informal study about their Performance Evaluation score. The evaluation was done for a score of 100. Using One-way ANOVA technique, find out whether or not a difference exists somewhere between the three different year levels

Scores		
First Year	**Second Year**	**Third Year**
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

Step1:

Setting the hypothesis (Null hypothesis or alternate hypothesis)

- Null Hypothesis (H0: $\mu 1=\mu 2=\mu 3$)
- Alternate Hypothesis (Ha: Alteast one difference among the means)

  And
- Fixing the confidence interval (90%, 95%)

  α=0.1 or 0.05

Step2: Find the df

- df between the groups/columns
- df within the groups/columns
- df_total

Step3:Calculating the Means

- Means for each group and
- Grand mean

Step4: All variability across the columns/groups

- SST
- SSB (Sum of Squares between/Columns)
- SSE(Sum of Squares within/Errors)

Step5: To calculate the variance between and within

- Mean Squares_between = $\frac{SS\_between}{df\_between}$
- Mean Squares_within = $\frac{SS\_within}{df\_within}$

Step 6: To perform F test (To calculate F_ratio)

- F_statistic = Mean Square_between / Mean Square_within
- F_critical from F distribution table

Source of Variation	Sums of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	F
Between Treatments	$\mathrm{SSB} = \Sigma n_j \left(\bar{X}_j - \bar{\bar{X}}\right)^2$	k-1	$\mathrm{MSB} = \frac{SSB}{k-1}$	$F = \frac{\mathrm{MSB}}{\mathrm{MSE}}$
Error (or Residual)	$\mathrm{SSE} = \Sigma\Sigma \left(X - \bar{X}_j\right)^2$	N-k	$\mathrm{MSE} = \frac{\mathrm{MSE}}{N-k}$	
Total	$\mathrm{SST} = \Sigma\Sigma \left(X - \bar{\bar{X}}\right)^2$	N-1		

Formulas for One-Way ANOVA

SSC	Sum of squares (columns/treatments)
SSE	Sum of squares (within/error)
SST	Sum of squares (total)

df = Degrees of Freedom

1. DoF b/w the columns

$$\mathrm{df}_{columns} = C - 1$$

Mean Squares_between $= \frac{\text{SS\_between}}{df\_between}$

2. DoF within the columns

$$\mathrm{df}_{error} = N - C$$

Mean Squares_within $= \frac{SS\_within}{df\_within}$

$$\mathrm{df}_{total} = N - 1$$

ANOVA F - statistic

$$F = \frac{\text{Mean Squares\_between}}{\text{Mean Squares\_within}}$$

N = total observations

C = # columns/treatments

MSC = Mean Square Columns/ Treatments

MSE = Mean Square Error/ Within

of the *F* Distribution

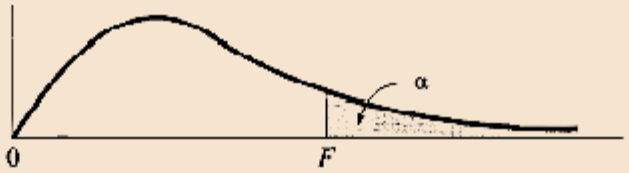


Table 1 $\alpha = 0.05$

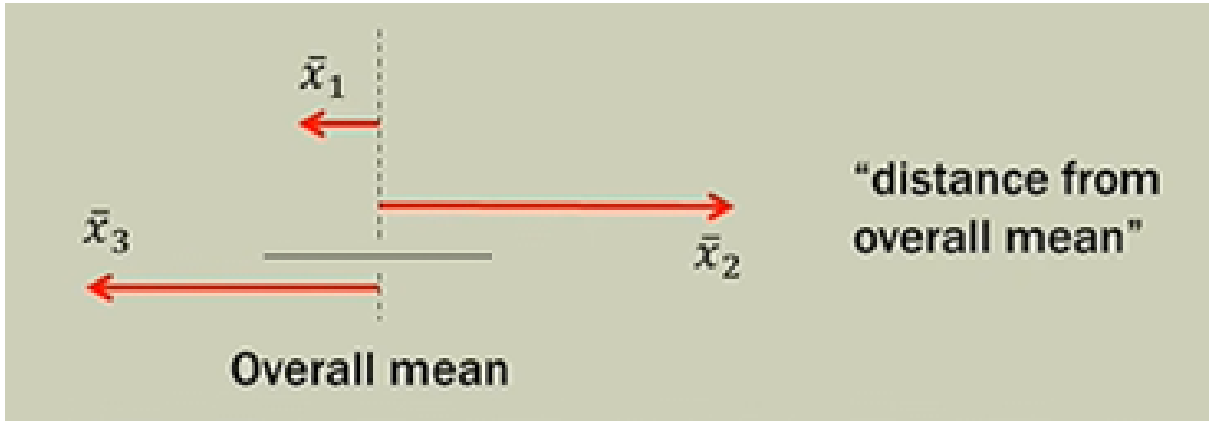
Degrees of Freedom for Numerator

Degrees of Freedom for Denominator	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.73	1.69	1.63	1.60

- F-statistic value is less than $F_{critical}$
- Null hypothesis is accepted.
- **It means there is no significant difference in mean values**

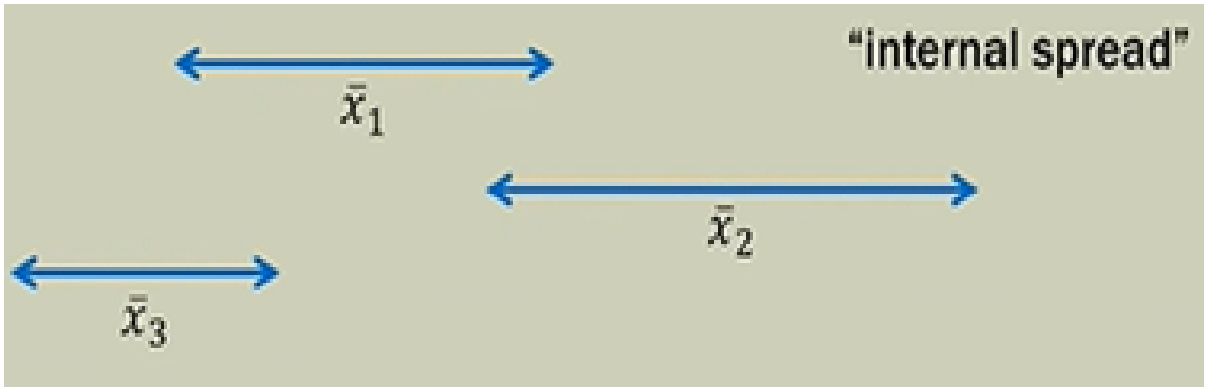
ANOVA: Analysis of Variance is a *variability ratio*

Variability
AMONG / BETWEEN
the means.



$$= \frac{\textit{Variance Between}}{\textit{Variance Within}}$$

Variability
AROUND / WITHIN the
distributions.



ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{Variance\ Between}{Variance\ Within} \Big\} \ Total\ Variance\ Components$$

One way ANOVA/ Single Factor ANOVA

If the variability BETWEEN the means (distance from overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ratio will be much larger than 1. The samples then most likely do NOT come from a common population; REJECT NULL HYPOTHESIS that means are equal.

ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{LARGE}{small} = Reject\ H_0$$

At least one mean is an outlier and each distribution is narrow; distinct from each other

$$\frac{Variance\ Between}{Variance\ Within}$$

$$\frac{similar}{similar} = Fail\ to\ Reject\ H_0$$

Means are fairly close to overall mean and/ or distributions overlap a bit; hard to distinguish

$$\frac{small}{LARGE} = Fail\ to\ Reject\ H_0$$

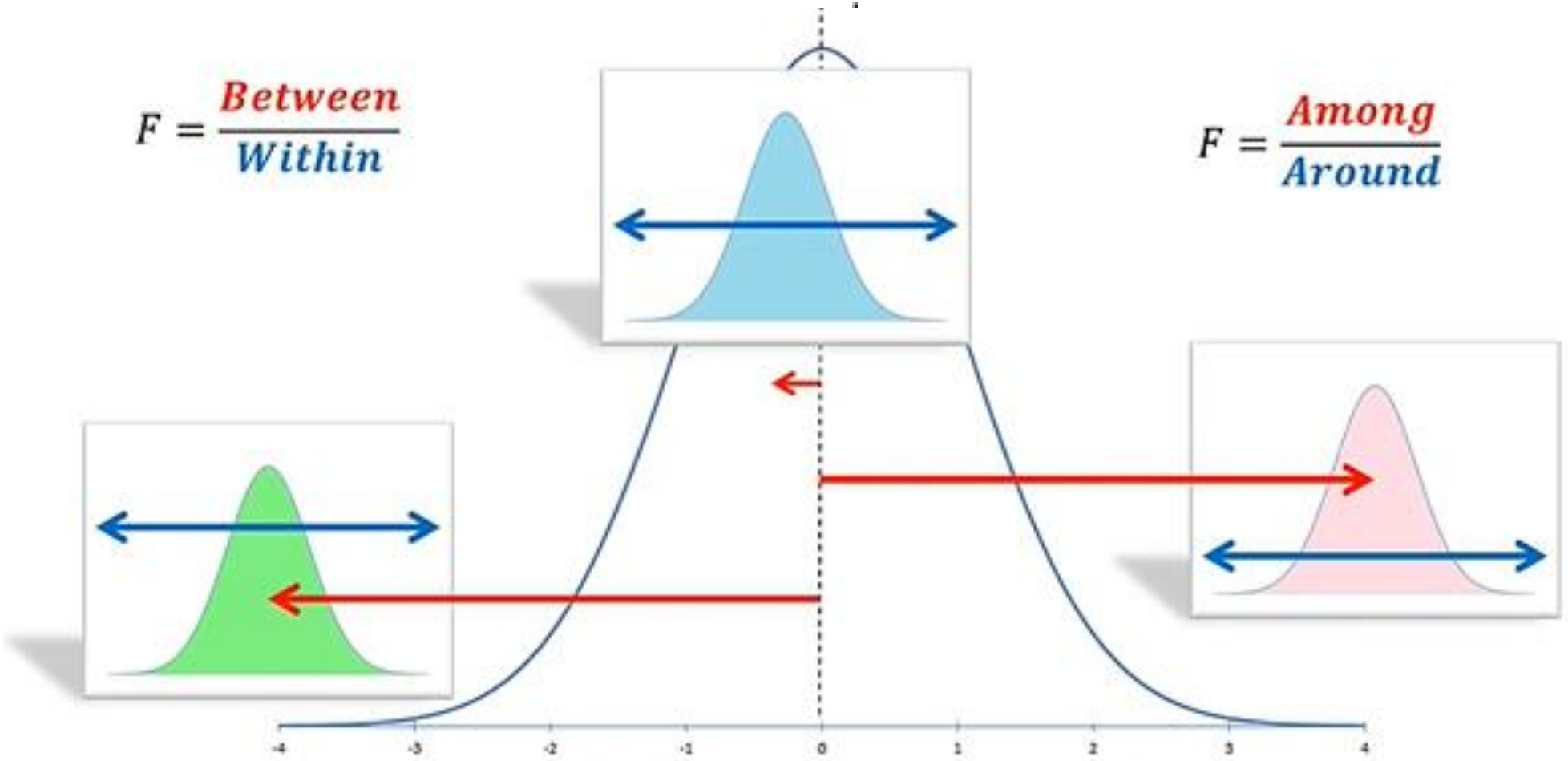
The means are very close to overall mean and/ or distribution “melt” together

ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$

$$F = \frac{Between}{Within}$$

$$F = \frac{Among}{Around}$$



Question:

18 employees (six each from first year to third year of experience) were selected for an informal study about their Performance Evaluation score. The evaluation was done for a score of 100. Using One-way ANOVA technique, find out whether or not a difference exists somewhere between the three different year levels

Scores		
First Year	**Second Year**	**Third Year**
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

Groups/ Columns

Random Sample within each group

Scores		
First Year	**Second Year**	**Third Year**
82	62	64
93	85	73
61	94	87
74	78	91
69	71	56
53	66	78

Calculate the mean of each column

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Calculate Grand Mean/ Overall Mean $\bar{\bar{x}}$

The mean of all 18 scores is

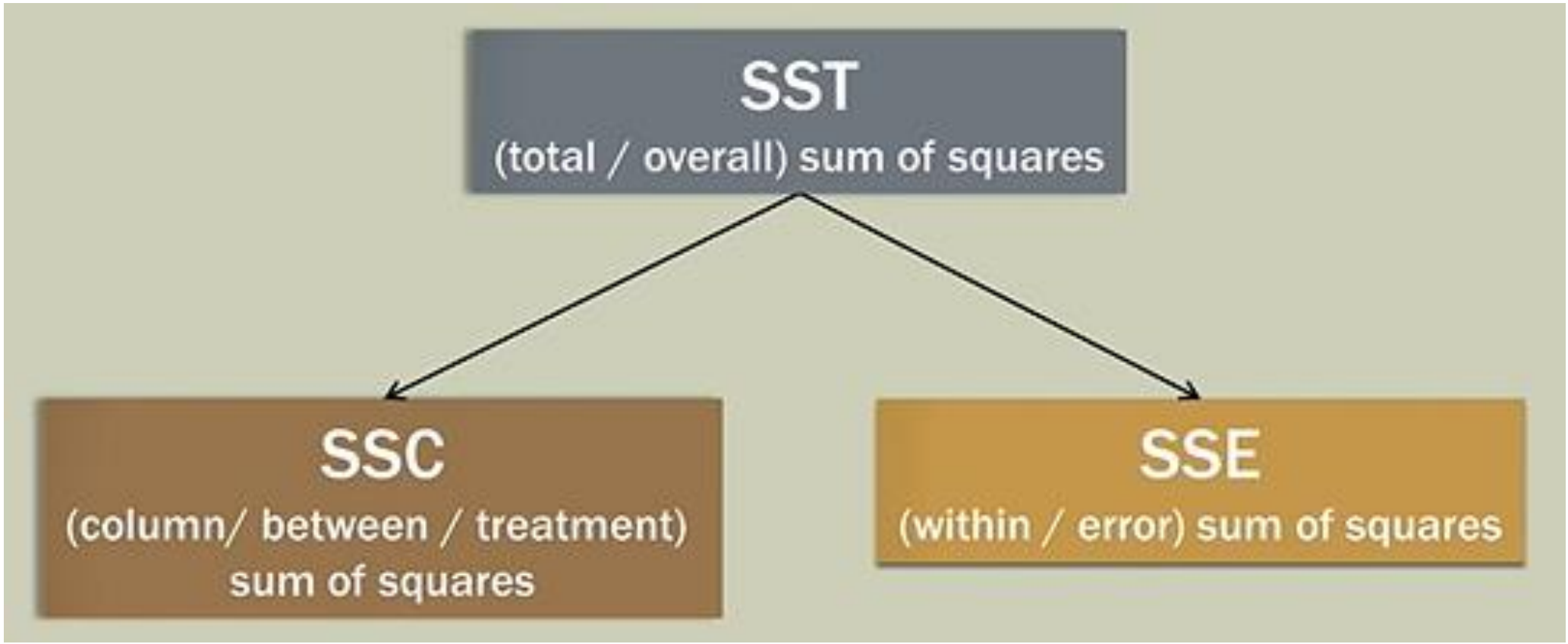
$$\bar{\bar{x}} = 74.28$$

Sum of Squares (SS)

Sum of squares of the difference of the dependent variable and its mean

$$SS = \sum(x - \bar{x})^2$$

Partitioning Sum of Squares



	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

SST
(total / overall)
sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\bar{\bar{x}}$ = 74.28

	Scores					
	First Year	Second Year	Third Year	$(X_A-X_{mean})^2$	$(X_B-X_{mean})^2$	$(X_c-X_{mean})^2$
	82	62	64	59.633	150.744	105.633
	93	85	73	350.522	114.966	1.633
	61	94	87	176.299	388.966	161.855
	74	78	91	0.077	13.855	279.633
	69	71	56	27.855	10.744	334.077
	53	66	78	452.744	68.522	13.855
Sum	432	456	449	**1067.130**	**747.796**	**896.685**
Mean	72	76	74.83			

SST = 1067.130 + 747.796 + 896.685 = 2711.611

SST
(total / overall) sum of squares

1. Find difference between each data point and the overall mean.
2. Square the difference.
3. Add them up

$\overline{\overline{x}} = 74.28$

	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_between

1. **Find difference between each group mean and the overall mean**
2. **Square the deviations**
3. **Multiply with no. of values of each column**
4. **Add them up**

$\bar{\bar{x}}$ = 74.28

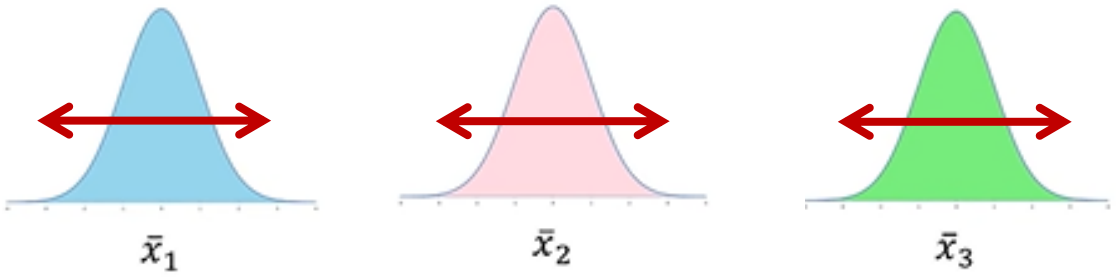
	Scores		
	First Year	Second Year	Third Year
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_between

1. **Find difference between each group mean and the overall mean**
2. **Square the deviations**
3. **Multiply with no. of values of each column**
4. **Add them up**

$\bar{\bar{x}}$ **= 74.28**

SSC = $6(72 - 74.28)^2 + 6(76 - 74.28)^2 + 6(74.83 - 74.28)^2 = 50.778$



	Scores		
	First Year	**Second Year**	**Third Year**
	82	62	64
	93	85	73
	61	94	87
	74	78	91
	69	71	56
	53	66	78
Mean $\bar{x}$	**72**	**76**	**74.83**

Sum of Squares_within

1. Find difference between each data point and its column mean.
2. Square each deviation.
3. Add them up the squared deviations.

	Scores					
	First Year	**Second Year**	**Third Year**	**$(X_A-x_{a\_mean})^2$**	**$(X_B-x_{b\_mean})^2$**	**$(X_C-x_{c\_mean})^2$**
	82	62	64	100	196	117.361
	93	85	73	441	81	3.361
	61	94	87	121	324	148.028
	74	78	91	4	4	261.361
	69	71	56	9	25	354.694
	53	66	78	361	100	10.028
Sum	432	456	449	**1036**	**730**	**894.833**
Mean	72	76	74.83			

SSE = 1036 + 730 + 894.833 = 2660.833

Sum of Squares_within

1. **Find difference between each data point and its column mean.**
2. **Square each deviation.**
3. **Add them up the squared deviations.**

Formulas for One-Way ANOVA

SSC Sum of squares (columns/treatments)

SSE Sum of squares (within/error)

SST Sum of squares (total)

df = Degrees of Freedom

1. DoF b/w the columns

$$\mathrm{df}_{columns} = C - 1$$

Mean Squares_between $= \frac{\text{SS\_between}}{df\_between}$

2. DoF within the columns

$$\mathrm{df}_{error} = N - C$$

Mean Squares_within $= \frac{SS\_within}{df\_within}$

$$\mathrm{df}_{total} = N - 1$$

ANOVA F - statistic

$$F = \frac{\text{Mean Squares\_between}}{\text{Mean Squares\_within}}$$

N = total observations

C = # columns/treatments

MSC = Mean Square Columns/ Treatments

MSE = Mean Square Error/ Within

Substituting the values

Formula to calculate Critical Value in Excel:

F.INV.RT(ALPHA,NUMERATOR DOF, DENOMINATOR DOF)

$$\text{Mean Squares\_between} = \frac{50.778}{3-1} = \mathbf{25.389}$$

$$\text{Mean Squares\_within} = \frac{2660.833}{18-3} = \mathbf{177.389}$$

$$F = \frac{\text{MSC}}{\text{MSE}} = \frac{25.389}{177.389} = \mathbf{0.1431}$$

- F-statistic value is less than $F_{critical}$
- Null hypothesis is accepted.
- **It means there is no significant difference in mean values**

Critical value of F: $F_{\alpha, dfc, dfe} = F_{0.05, 2, 15} = \mathbf{3.68}$

Compute one-way ANOVA

- We want to know if there is any significant difference between the average weights of plants in the 3 experimental conditions.
- Functions used
  - aov()
  - summary.aov()

R code

```
#compute analysis of variance
res <-aov(weight~group,data=mydat)

#summary of analysis
summary.aov(res)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2 3.766  1.8832  4.846 0.0159 *
Residuals 27 10.492 0.3886
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with “
in the model summary.

Multiple pairwise-comparison between the means of groups

- In one-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don’t know which pairs of groups are different.

- It’s possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.

Tukey multiple pairwise-comparisons

- As the ANOVA test is significant, we can compute **Tukey HSD** (Tukey Honest Significant Differences)
- R function: **TukeyHSD**() for performing multiple pairwise-comparison between the means of groups.

R code

```
#Tukey HSD -multiple pairwise comparison
TukeyHSD(res)
```

```
Tukey multiple comparisons of means 95% family-wise confidence level
Fit: aov(formula = weight ~ group, data = mydat)
$group diff lwr upr p adj
trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
trt2-ctrl 0.494 -0.1972161 1.1852161 0.1979960
trt2-trt1 0.865 0.1737839 1.5562161 0.0120064
```

It can be seen from the output, that only the difference between trt2 and trt1 is significant with an adjusted p-value of 0.012.

Checking ANOVA Assumptions

- Check the homogeneity of variance assumption
  - The **residuals versus fits plot** can be used to check the homogeneity of variances.

```
#checking homogeneity of variance
plot(res,1)
```

  - Levene's test

```
#Levene's Test
library(car)
leveneTest(weight~group,mydat)
```

Checking ANOVA Assumptions (contd.)

- Check the normality assumption
  - **Normality plot of residuals**. In this plot, the quantiles of the residuals are plotted against the quantiles of the normal distribution.

```
#checking normality
plot(res,2)
```

  - **Shapiro-Wilk test** on ANOVA residuals

```
#Extract the residuals
res_resi <- residuals(res)
shapiro.test(res_resi)
```

Non-parametric alternative to one-way ANOVA test

- A non-parametric alternative to one-way ANOVA is **Kruskal-Wallis rank sum test**, which can be used when ANOVA assumptions are not met.

```
kruskal.test(weight~group,mydat)
```

Thank You