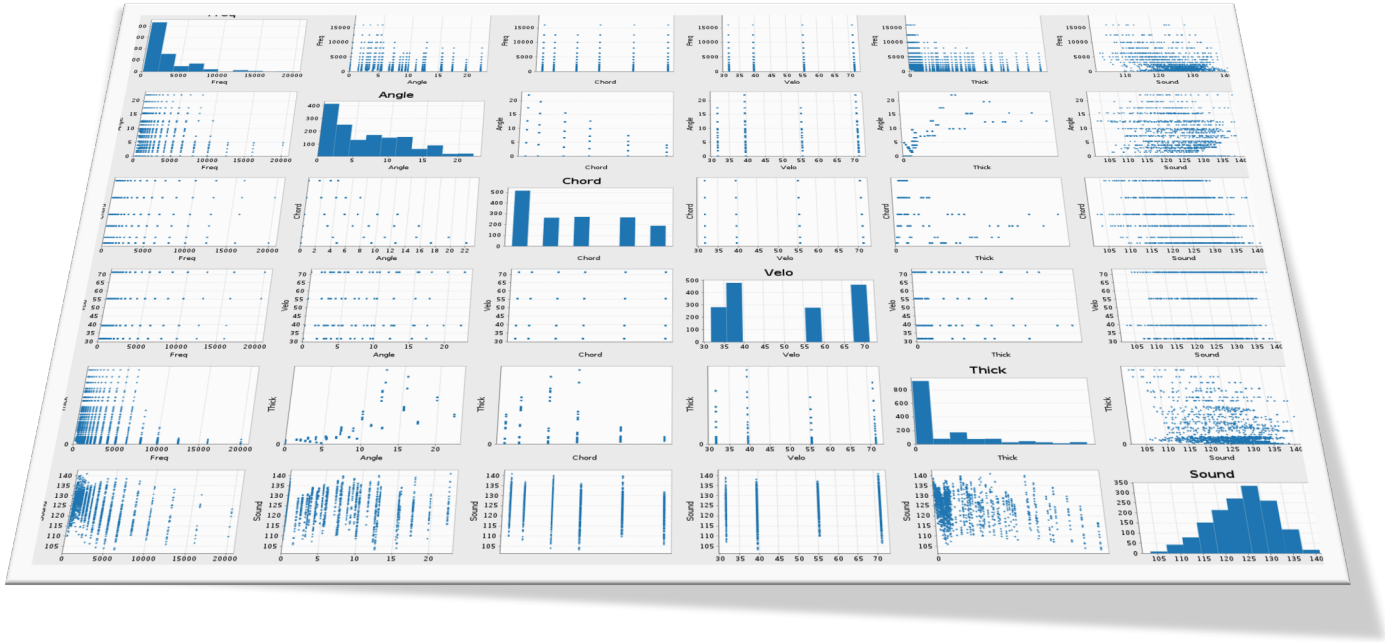
WQD7001

Regression Modelling



Types of Regression

- Linear regression
- Logistic regression
- Polynomial regression
- Stepwise regression
- Stepwise regression
- Ridge regression
- Lasso regression
- ElasticNet regression

Linear and **Logistic regressions** are usually the first modeling algorithms that people learn for Machine Learning and Data Science.

- **Logistic regression** is the most famous machine learning algorithm after linear regression.
- In a lot of ways, **linear regression** and **logistic regression** are similar. But, the biggest difference lies in what they are used for.
- If you are dealing with continuous / discrete values, then go for **Linear Regression**.
- **Linear regression** algorithms are used to predict/forecast values.
- If you are dealing with a classification problem like (Yes/No, Fraud/Non Fraud) then use **Logistic Regression**.
- Logistic regression is used for **classification tasks**.
  - classifying whether an email is a spam or not,
  - classifying whether a tumor is malignant or benign,
  - classifying whether a website is fraudulent or not, etc.

Independent and Dependent Variables

In the context of Statistical learning, there are **2** types of data:

- **Independent variables**: *Data that **can be controlled** directly*.
- **Dependent variables**: *Data that **cannot be controlled** directly*.

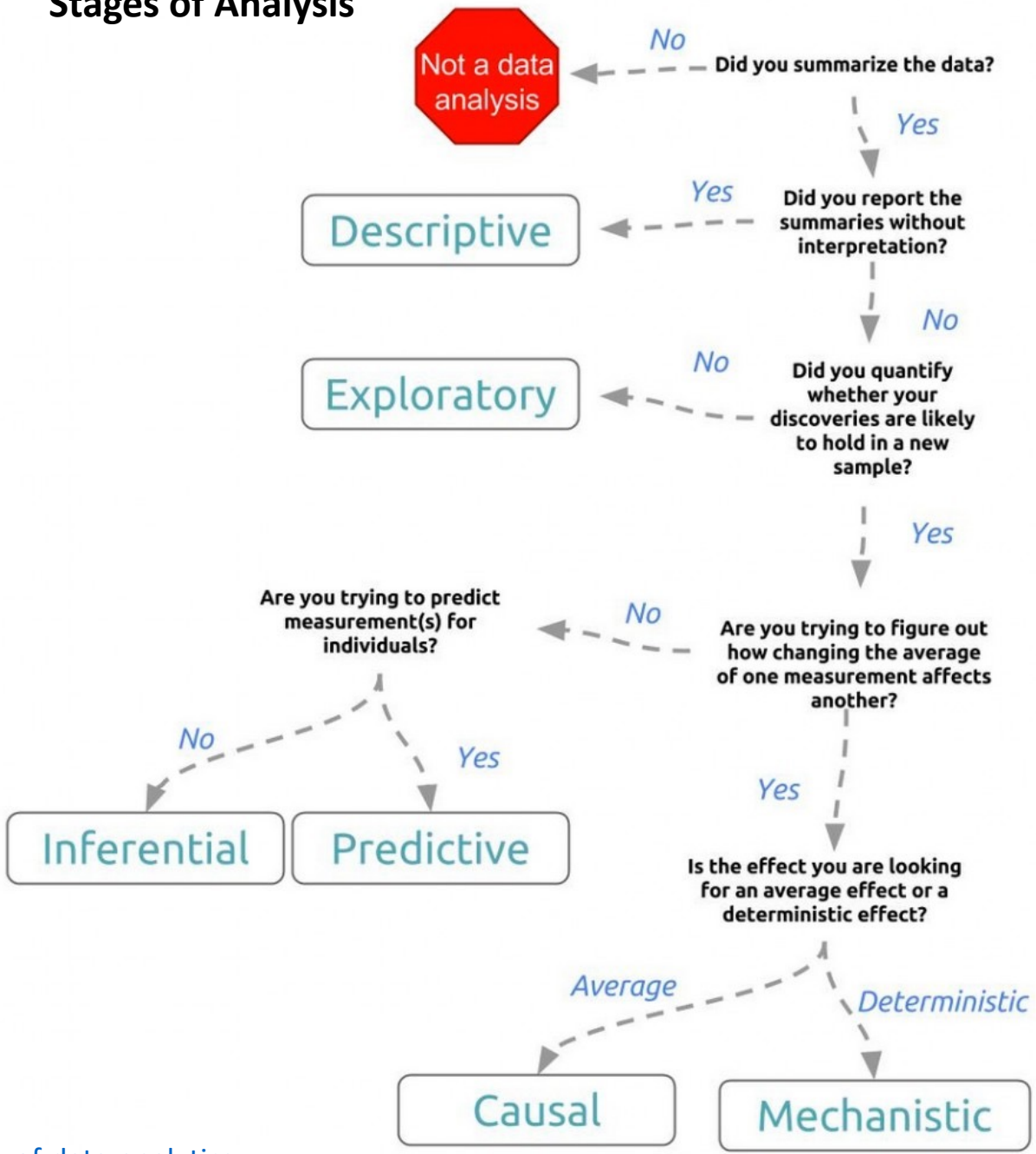
The data that can't be controlled (**dependent variables**) need to be **predicted** or **estimated**.

A **model** is a transformation engine that helps us to express dependent variables as a function of independent variables.

Parameters: are ingredients added to the model for estimating the output.

Linear regression models provide a simple approach towards **supervised learning**. They are simple yet effective.

Stages of Analysis

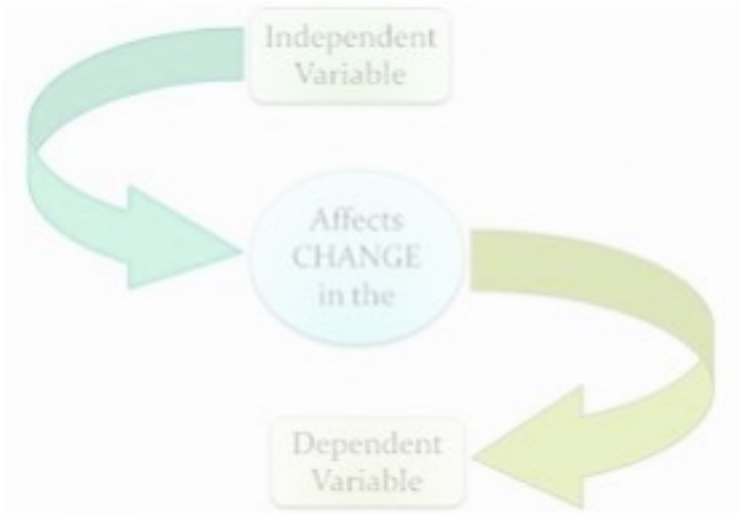


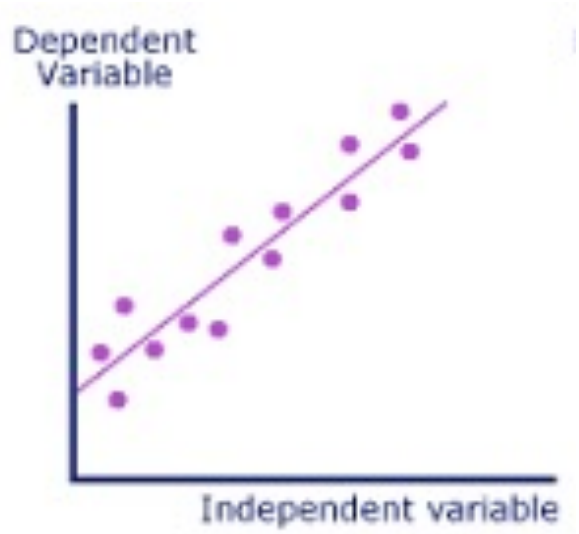
https://www.cuelogic.com/blog/the-workflow-of-data-analytics

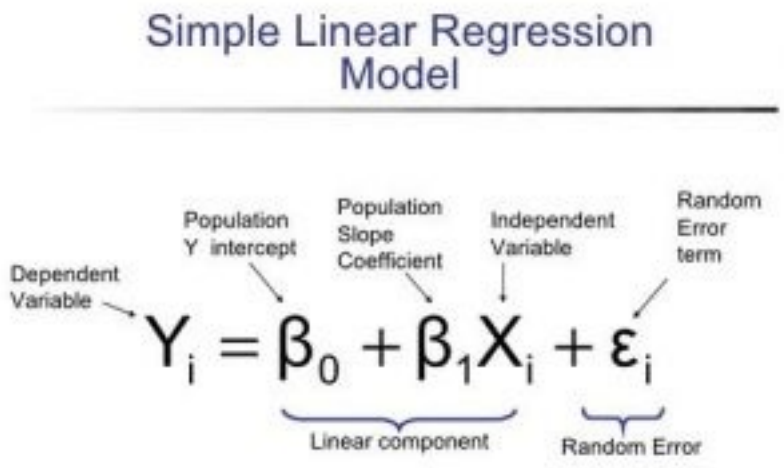
Naming the Variables

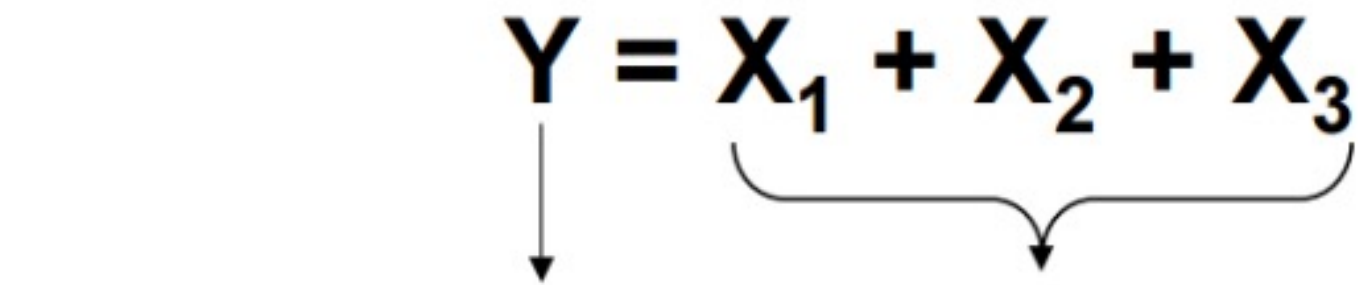
There are **many names** for :

- Regression's **dependent variable (Y axis)**. It may be called an **outcome variable**, **criterion variable**, **endogenous variable**, or **regressand**.
- The **independent variables (X axis)**can be called *exogenous variables*, **predictor variables**, or **regressors**.









Dependent Variable	Independent Variable
Outcome Variable	Predictor Variable
Response Variable	Explanatory Variable

What do we mean by linear?

Linear implies the following:

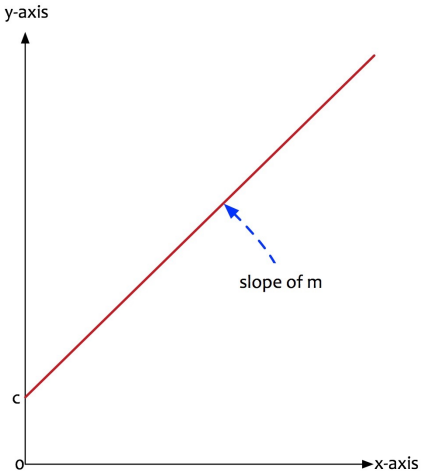
- ❑ arranged in or extending along a **straight** or nearly straight line.
- ❑ Linear suggests that the **relationship** between dependent and independent variable can be expressed in a straight line.

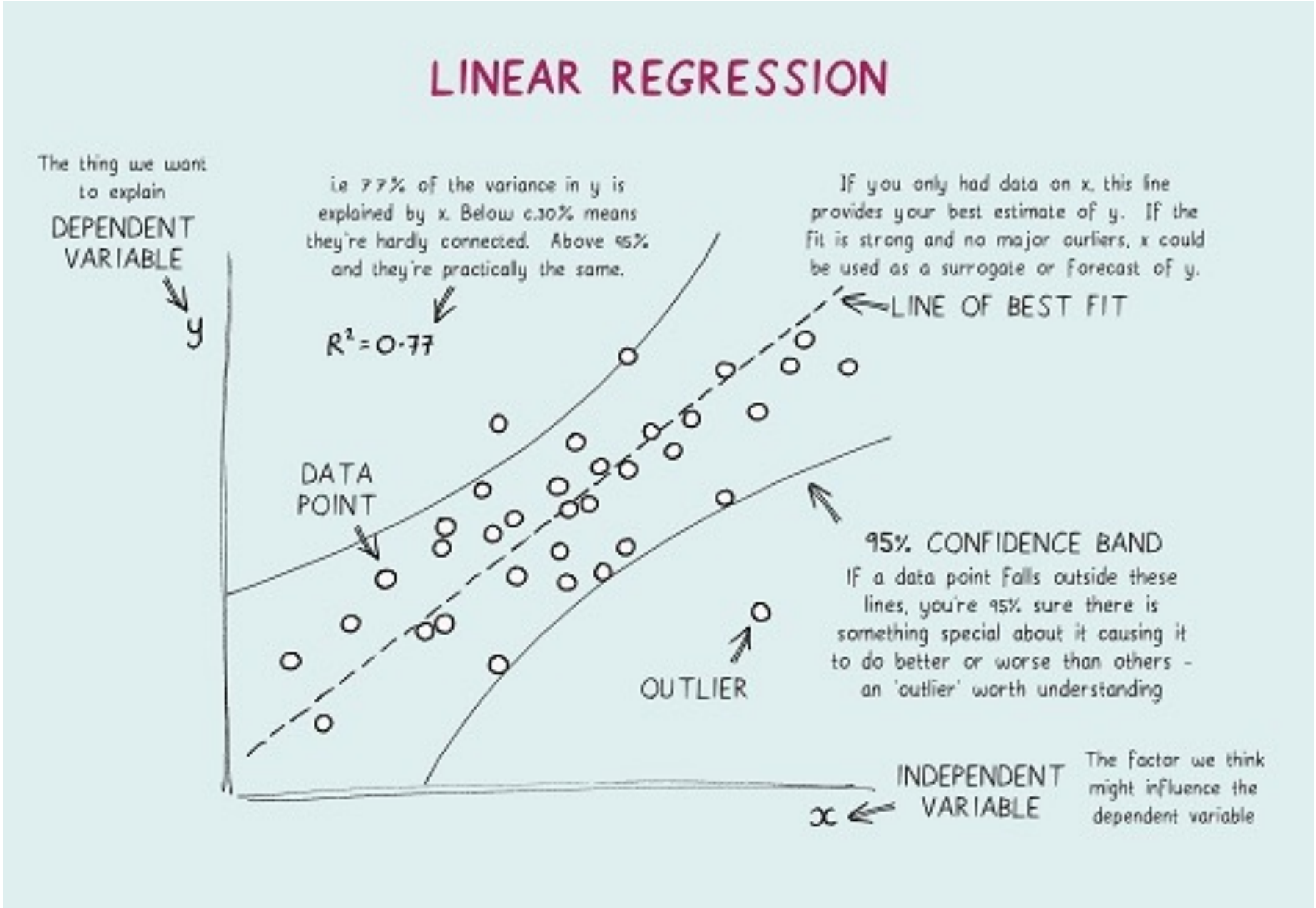
What is the **equation of a line**? Recall the geometry lesson that you'd learnt

$$y = mx + c$$

Linear regression is nothing but a manifestation of this simple equation.

- **y** is the **dependent variable** i.e. the variable that needs to be estimated and predicted.
- **x** is the **independent variable** i.e. the variable that is controllable. It is the input.
- **m** is the **slope**. It determines what will be the angle of the line. It is the parameter denoted as **β**.
- **c** is the **intercept**. A constant that determines the value of y when x is 0.




LINEAR REGRESSION
The thing we want to explain
DEPENDENT VARIABLE
y
i.e 77% of the variance in y is explained by x. Below c.30% means they're hardly connected. Above 95% and they're practically the same.
$R^2 = 0.77$
If you only had data on x, this line provides your best estimate of y. If the fit is strong and no major ourliers, x could be used as a surrogate or forecast of y.
LINE OF BEST FIT
DATA POINT
95% CONFIDENCE BAND
If a data point falls outside these lines, you're 95% sure there is something special about it causing it to do better or worse than others - an 'outlier' worth understanding
OUTLIER
INDEPENDENT VARIABLE
x
The factor we think might influence the dependent variable

"Essentially, all models are wrong,
...but some are useful."

British Mathematician and Professor of Statistics

Any model, no matter how accurate is still some type of **simplification** or **generalization** of something.

Linear regression **models** are **not perfect**.
It tries to *approximate* the *relationship* between dependent and independent variables in a straight line.
Approximation leads to **errors**.
Some errors **can be reduced**.
Some errors are **inherent** in the nature of the problem.
These errors cannot be eliminated. They are called as an **irreducible error**, the noise term in the true relationship that cannot fundamentally be reduced by any model.

What is Linear Regression?

- Most commonly used method for **predictive analytics**.
- Method used to describe relationship between a **dependent variable** and one or **independent variable**.
- The **main task** in the Linear Regression is the method of fitting a single line within a scatter plot.
- The Linear Regression consists of the following **three steps**:
  - Determining and analyzing the correlation and direction of the data
  - Deploying the estimation of the model
  - Ensuring the usefulness and validity of the model

1. Determining and analyzing the correlation and direction of the data

- Firstly, a **scatter plot** should be used to analyze the data and check for directionality and correlation of data.
- Look for (i) **linear** or **non-linear pattern** of the data and (ii) **deviations** from the pattern (outliers).
- If the pattern is **non-linear**, consider a transformation.
- If there are **outliers**, you may consider removing them only IF there is a non-statistical reason to do so.
- This step enables the data scientist to formulate the model.

2. Deploying the estimation of the model

- The second step of regression analysis is to **fit the regression line**.
- Fit the **least-squares regression** line to the data and check the assumptions of the model by looking at the Residual Plot (for constant standard deviation assumption) and normal probability plot (for normality assumption).
- Once a "**good-fitting**" model is determined, write the equation of the least-squares regression line. Include the standard errors of the estimates, the estimate of σ, and R-squared

3. Ensuring the usefulness and validity of the model

- The last step for the linear regression analysis is the test of significance, which is evaluating the validity and usefulness of the model.
- Determine if the explanatory variable is a significant predictor of the response variable by performing a t-test or F-test.
- Include a confidence interval for the estimate of the regression coefficient (slope).

Example: Workings of a Linear regression Model

Fernando is a Data Scientist. He wants to buy a car. He wants to estimate or predict the car price that he will have to pay. He has a friend at a car dealership company. He asks for prices for various other cars along with a few characteristics of the car. His friend provides him with some information.

The following are the data provided to him:

- *make*: make of the car.
- *fuelType*: type of fuel used by the car.
- *nDoor*: number of doors.
- *engineSize*: size of the engine of the car.
- *price*: the price of the car.

Example: Workings of a Linear regression Model

First, Fernando wants to evaluate if indeed he can predict car price based on engine size. The first set of analysis seeks the answers to the following questions:

- Is price of car price related with engine size?
- How strong is the relationship?
- Is the relationship linear?
- Can we predict/estimate car price based on engine size?

Fernando does a **correlation analysis**. Correlation is a measure of how much the two variables are related. It is measured by a metric called as the **correlation coefficient**. Its value is between **0 and 1**.

If the correlation coefficient is a large(> 0.7) +ve number, it implies that as one variable increases, the other variable increases as well. A large -ve number indicates that as one variable increases, the other variable decreases.

He does a correlation analysis. He plots the relationship between price and engine size.

Example: Workings of a Linear regression Model

- He splits the data into training and test set. 75% of data is used for training. Remaining is used for the test.
- He builds a linear regression model. He uses a statistical package to create the model. The model creates a linear equation that expresses price of the car as a function of engine size.

Following are the answers to the questions:
Q: Is price of car price related with engine size?
A: Yes, there is a relationship.

Q: How strong is the relationship?
A: The correlation coefficient is 0.872 => There is a strong relationship.

Q: Is the relationship linear?
A: A straight line can fit => A decent prediction of price can be made using engine size.

Q: Can we predict/estimate the car price based on engine size?
A: Yes, car price can be estimated based on engine size.

Example: Workings of a Linear regression Model

Model Building

- He splits the data into training and test set. 75% of data is used for training. Remaining 25% is used for the test.
- He builds a linear regression model. He uses a statistical package to create the model. The model creates a linear equation that expresses price of the car as a function of engine size.

The model estimates the parameters:

- β0 is estimated as -6870.1
- β1 is estimated as 156.9

$$\text{price} = \beta 0 + \beta 1 \text{ x engine size}$$

The linear equation is estimated as:

- price = -6870.1 + 156.9 x engine size

Example: Workings of a Linear regression Model

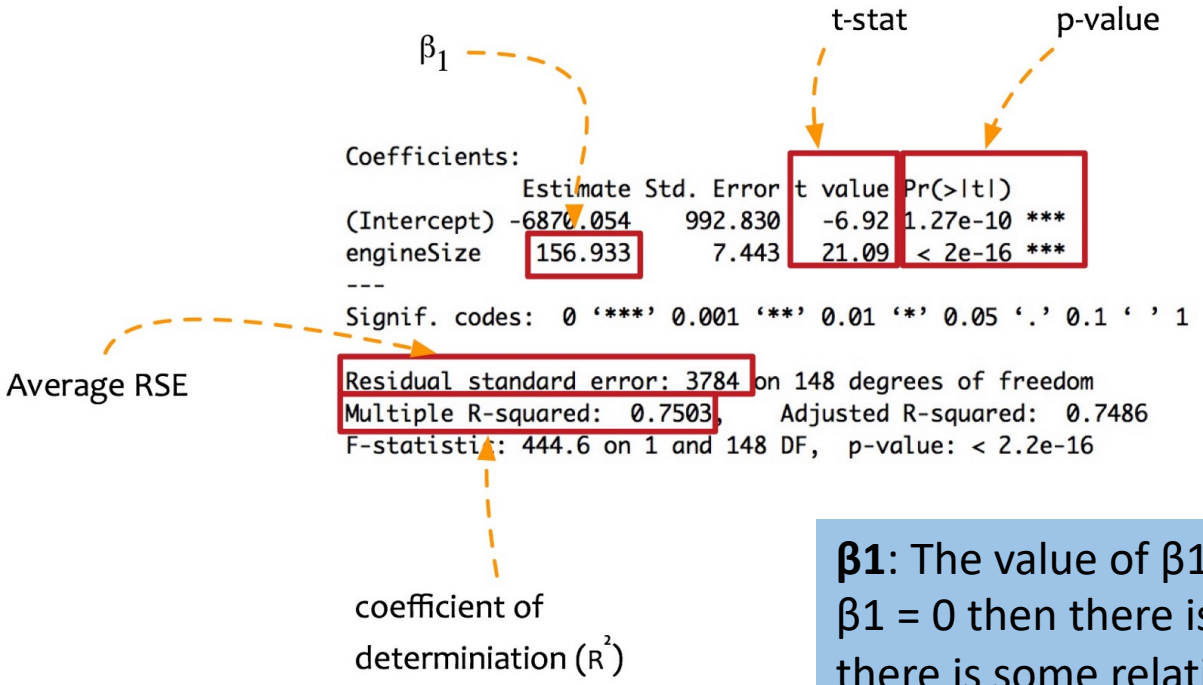
Interpretation

- The model provides the equation for the predicting the average car price given a specific engine size. This equation means the following:

  One unit increase in engine size will increase the average price of the car by 156.9 units.

Evaluation

- The model is built. The robustness of the model needs to be evaluated.
- **How can we be sure that the model will be able to predict the price satisfactory?**
- This evaluation is done in two parts.
  - First, test to establish the **robustness** of the model (using hypothesis testing).
  - Second, test to evaluate the **accuracy** of the model (using the coefficient of determination a.k.a R-squared).
- Fernando first evaluates the model on the training data. He gets the following statistics.



H0 and Ha need to be defined. They are defined as follows:

H0 (NULL hypothesis): There is no relationship between x and y i.e. there is no relationship between price and engine size.
Ha (Alternate hypothesis): There is some relationship between x and y i.e. there is a relationship between price and engine size.

β1: The value of β1 determines the relationship between price and engine size. If β1 = 0 then there is no relationship. In this case, β1 is positive. It implies that there is some relationship between price and engine size.

t-stat: The t-stat value is how many standard deviations the coefficient estimate (β1) is far away from zero. Further, it is away from zero stronger the relationship between price and engine size. The coefficient is significant. In this case, t-stat is 21.09. It is far enough from zero.

p-value: It indicates the chance of seeing the given t-statistics, under the assumption that NULL hypothesis is true. If the p-value is small e.g. < 0.0001, it implies that the probability that this is by chance and there is no relation is very low. In this case, the p-value is small. It means that relationship between price and engine is not by chance.

With these metrics, we can safely **reject the NULL hypothesis** and **accept the alternate hypothesis**.
There is a robust relationship between price and engine size.

How about **accuracy**? How accurate is the model? To get a feel for the accuracy of the model, a metric named **R-squared** or **coefficient of determination** is important.

The R-squared for the model created by Fernando is 0.7503 i.e. 75.03% on the training set. It means that the model can explain more than 75% of the variation.

Fernando has a good model now. It performs satisfactorily on the training data. However, there is **25% of data unexplained**.
There is room for **improvement**.

- How about adding more independent variable for predicting the price?

When more than one independent variables are added for predicting a dependent variable, a **multivariate regression model** is created i.e. more than one variable.

Overall Idea of Regression

To examine two things:

(1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?

These **regression estimates** are used to **explain the relationship** between one dependent variable and one or more independent variables.

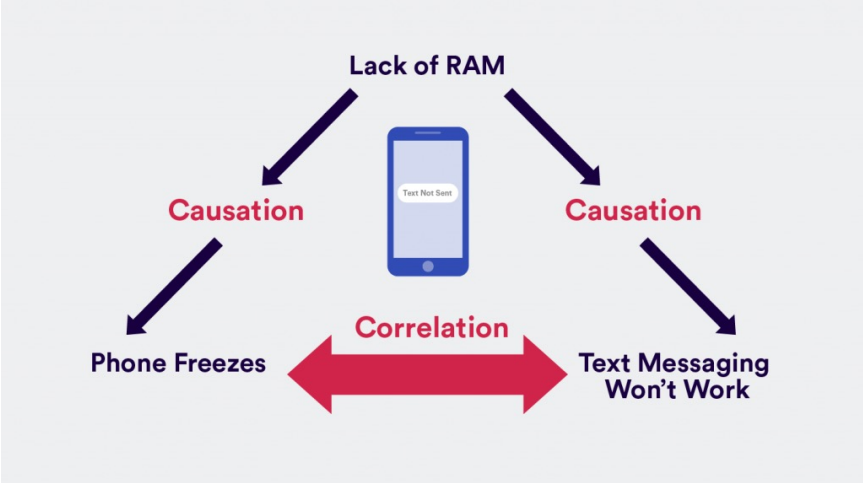
Correlation Vs Causation

- Correlation by itself **does not** imply a cause and effect relationship!

It can sometimes be a coincidence!

Causation is implying that A and B have a cause-and-effect relationship with one another.

- ✓ causation means that two events appear at the same time or one after the other.
- ✓ it means these two variables not only appear together, the existence of one causes the other to manifest.



https://clevertap.com/blog/correlation-vs-causation/

Correlation is a term in statistics that refers to the degree of association between two random variables.

3 Types

i. **Positive correlation** is when you observe A increasing and B increases as well. Or if A decreases, B correspondingly decreases. Example: the more purchases made in your app, the more time is spent using your app.

ii. **Negative correlation** is when an increase in A leads to a decrease in B or vice versa.

iii. **No correlation** is when two variables are completely unrelated and a change in A leads to no changes in B, or vice versa.

Major Uses for Regression Analysis

1. **Determining the strength of predictors** - the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.
2. **Forecasting an effect** - it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”
3. **Trend forecasting** - regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

Regression Analysis

- The simplest case to examine is one in which a variable Y , referred to as the **dependent** or target variable, may be related to one variable X, called an **independent** or explanatory variable, or simply a regressor.
- If the relationship between Y and X is believed to be linear, then the equation for a line may be appropriate:

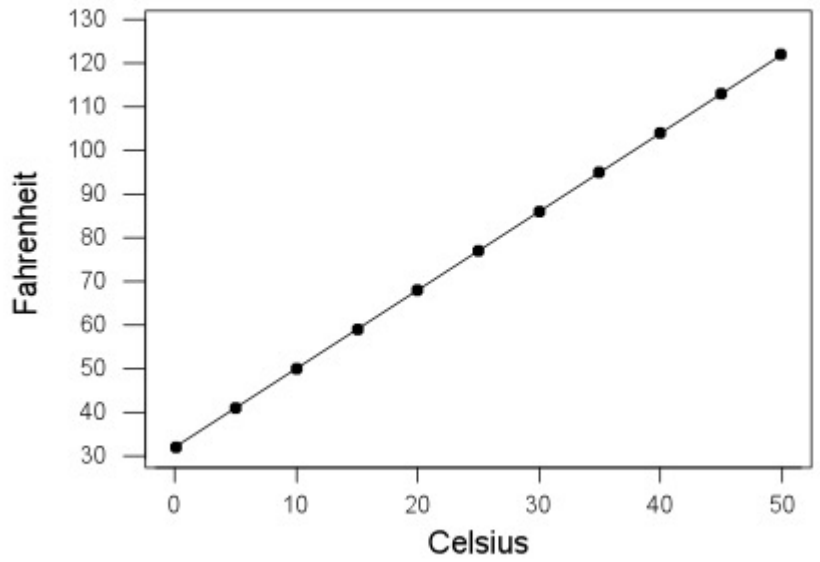
$$Y = \beta 1 + \beta 2X,$$

where β1 is an **intercept** term and β2 is a **slope coefficient**.

- In simplest terms, the purpose of regression is to try to **find the best fit line** or **equation** that expresses the relationship between Y and X.

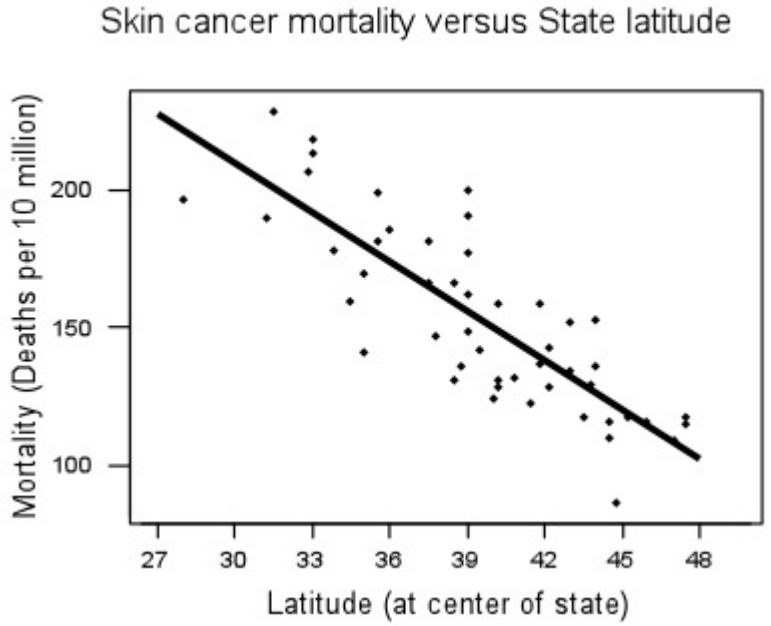
Types of Relationship

Deterministic (or Functional) Relationships



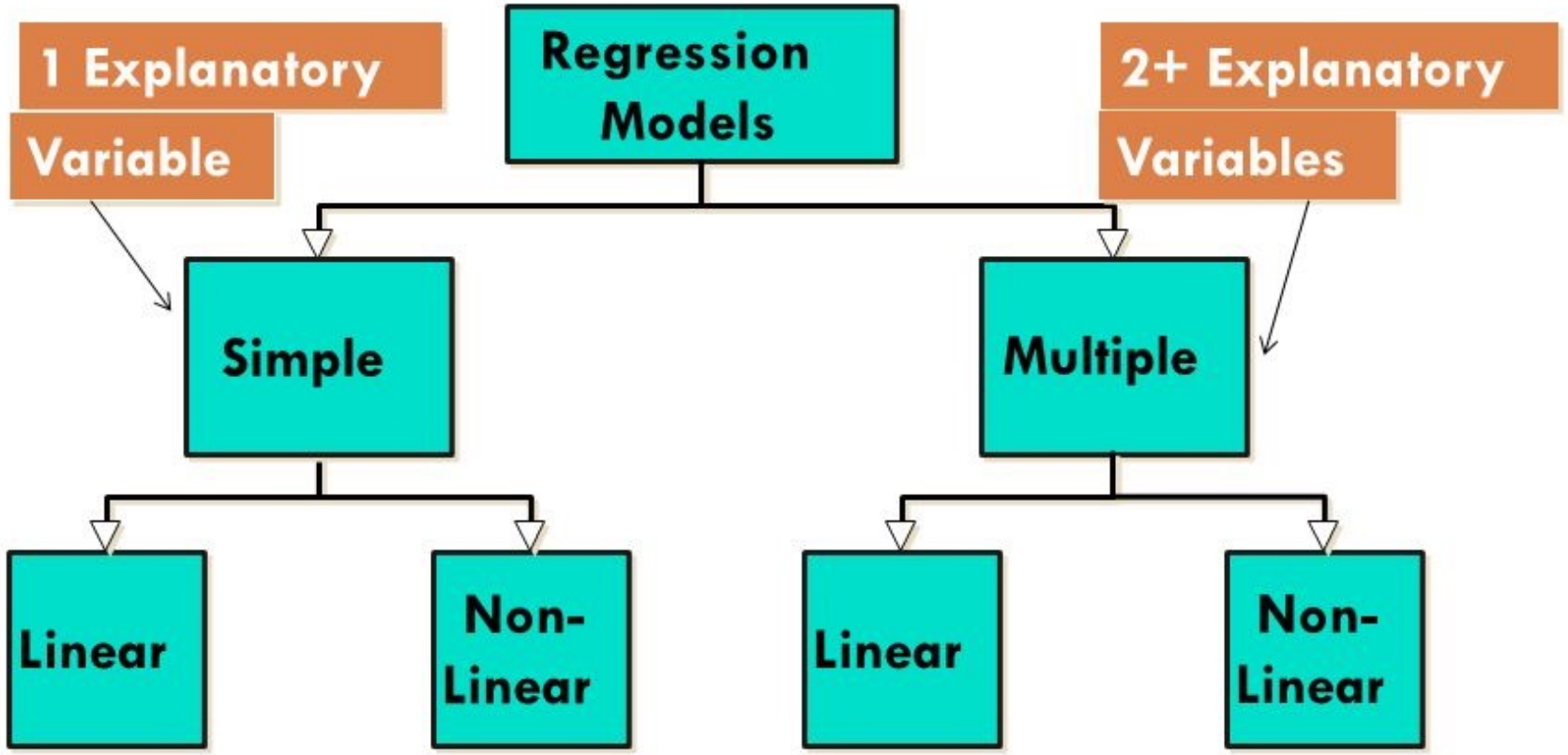
An **exact relationship** between the predictor *x* and the response *y*. Take, for instance, the conversion relationship between temperature in degrees Celsius (***C***) and temperature in degrees Fahrenheit (***F***).
We know the relationship is: F=95C+32
Therefore, if we know that it is 10 degrees Celsius, we also know that it is 50 degrees Fahrenheit:
F=95(10)+32=50

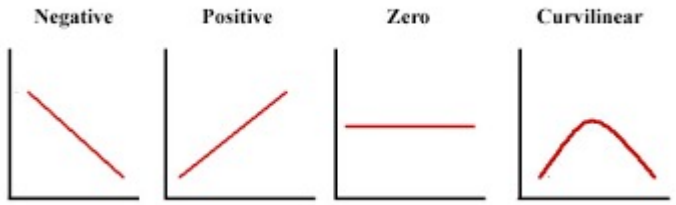
Statistical Relationships



Not an exact relationship. It is instead a relationship in which "trend" exists between the predictor x and the response y, but there is also some "scatter“.

Types of Regression Analysis





Videos To Watch

Simple Linear Regression, The Very Basics

https://www.youtube.com/watch?v=ZkjP5RJLQF4

Regression Analysis

https://www.youtube.com/watch?v=DtOYBxi4AIE

Manual Computation

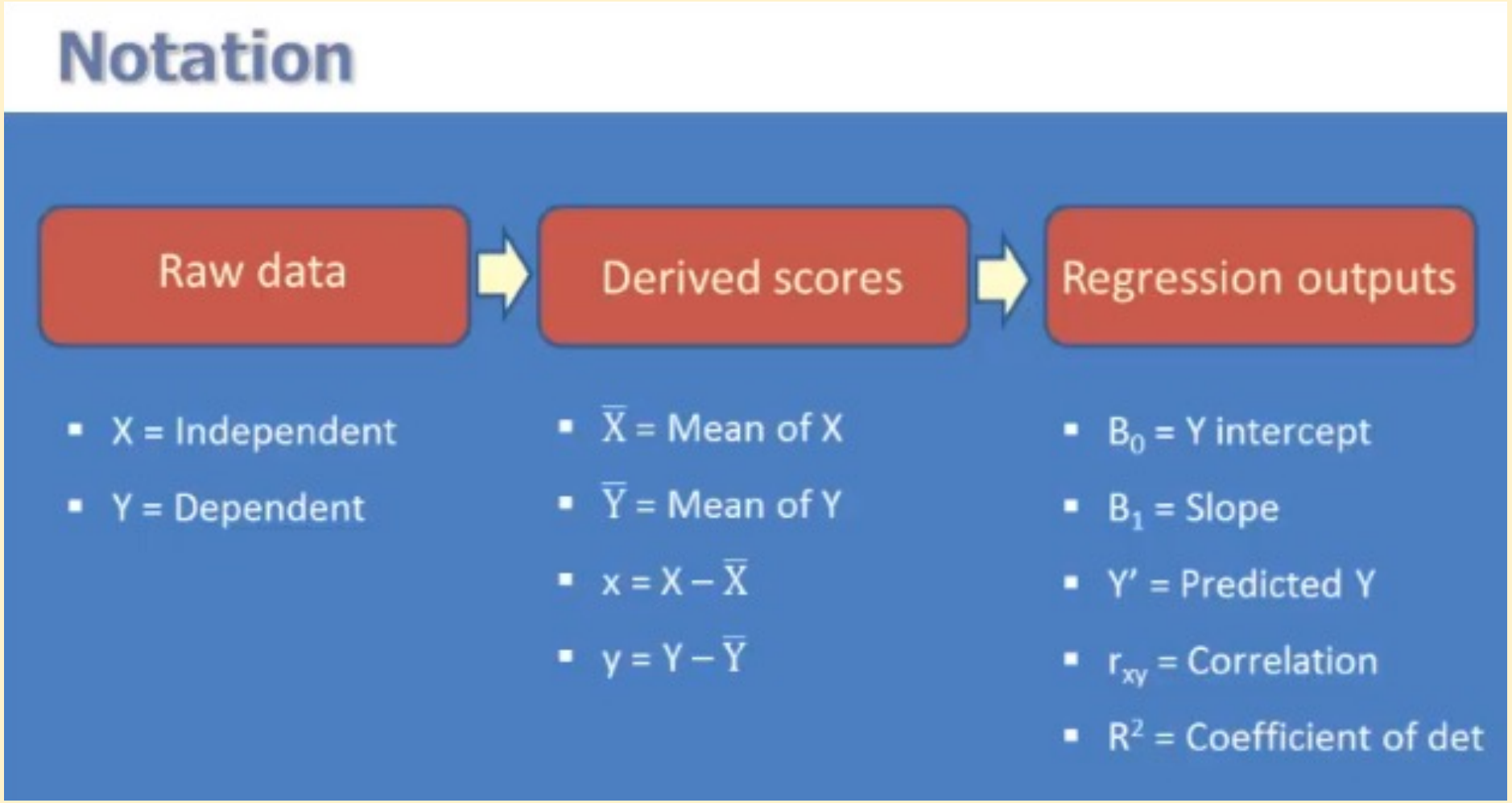
- **Regression computations** are usually handled by a **software package** or a **graphing calculator**. For this example, however, we will do the computations "manually“.

Problem Statement

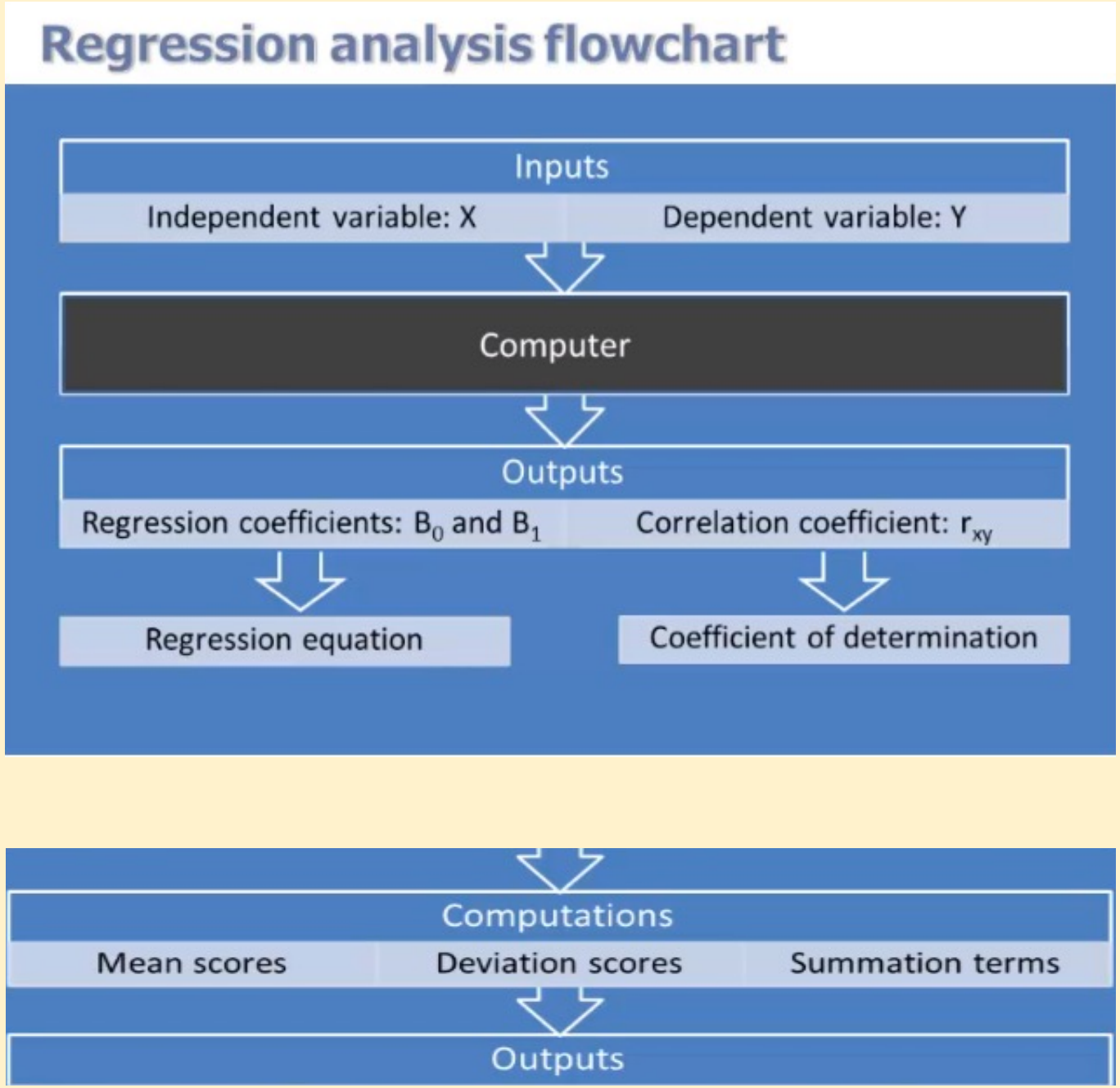
Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

Regression Analysis By Hand



http://stattrek.com/regression/regression-example.aspx?Tutorial=AP


Regression analysis flowchart
Inputs
Independent variable: X
Dependent variable: Y
Computer
Outputs
Regression coefficients: B_0 and B_1
Correlation coefficient: r_{xy}
Regression equation
Coefficient of determination
Computations
Mean scores
Deviation scores
Summation terms
Outputs

Initial computations

X = Math scores
Y = Statistics grades

	Raw data		Deviation scores				
	X	Y	x	y	x^2	y^2	xy
	95	85	17	8	289	64	136
	85	95	7	18	49	324	126
	80	70	2	-7	4	49	-14
	70	65	-8	-12	64	144	96
	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

Regression equation

$Y' = B_0 + B_1X$

$Y' = 26.77 + 0.64X$

$B_1 = \Sigma xy / \Sigma x^2$

$B_1 = 470 / 730$

$B_1 = 0.64$

$B_0 = \bar{Y} - B1 * \bar{X}$

$B_0 = 77 - 0.64 * 78$

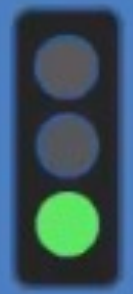
$B_0 = 26.77$

	Raw data		Deviation scores				
	X	Y	x	y	x^2	y^2	xy
	95	85	17	8	289	64	136
	85	95	7	18	49	324	126
	80	70	2	-7	4	49	-14
	70	65	-8	-12	64	144	96
	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77					

Coefficient of determination

Coefficient of determination (R^2) describes proportion of variance in Y that is predictable from X

If R^2 is close to zero, don't use regression



If R^2 is significantly greater than zero, use regression

How to find coefficient of determination

Correlation coefficient

$$r_{xy} = \Sigma xy \,/\, \sqrt{\Sigma x^2 \, \Sigma y^2}$$

$$r_{xy} = 470 \,/\, \sqrt{730 * 630} = 0.69$$

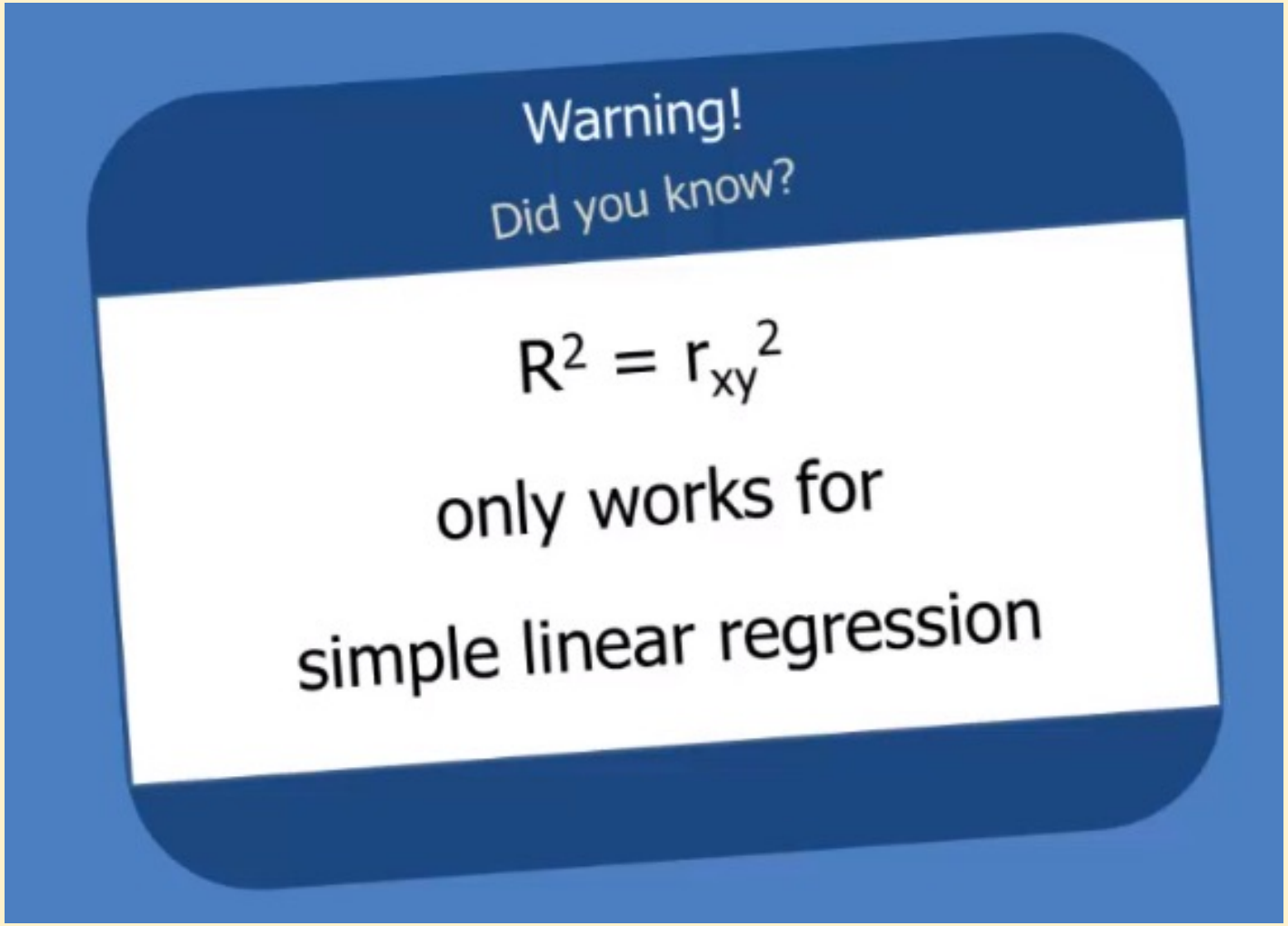
Coefficient of determination

$$R^2 = r_{xy}{}^2$$

$$R^2 = (0.69)^2 = 0.48$$

	Raw data		Deviation scores				
	X	Y	x	y	x^2	y^2	xy
	95	85	17	8	289	64	136
	85	95	7	18	49	324	126
	80	70	2	-7	4	49	-14
	70	65	-8	-12	64	144	96
	60	70	-18	-7	324	49	126
Sum	390	385			730	630	470
Mean	78	77			⇧	⇧	⇧

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fit to the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.

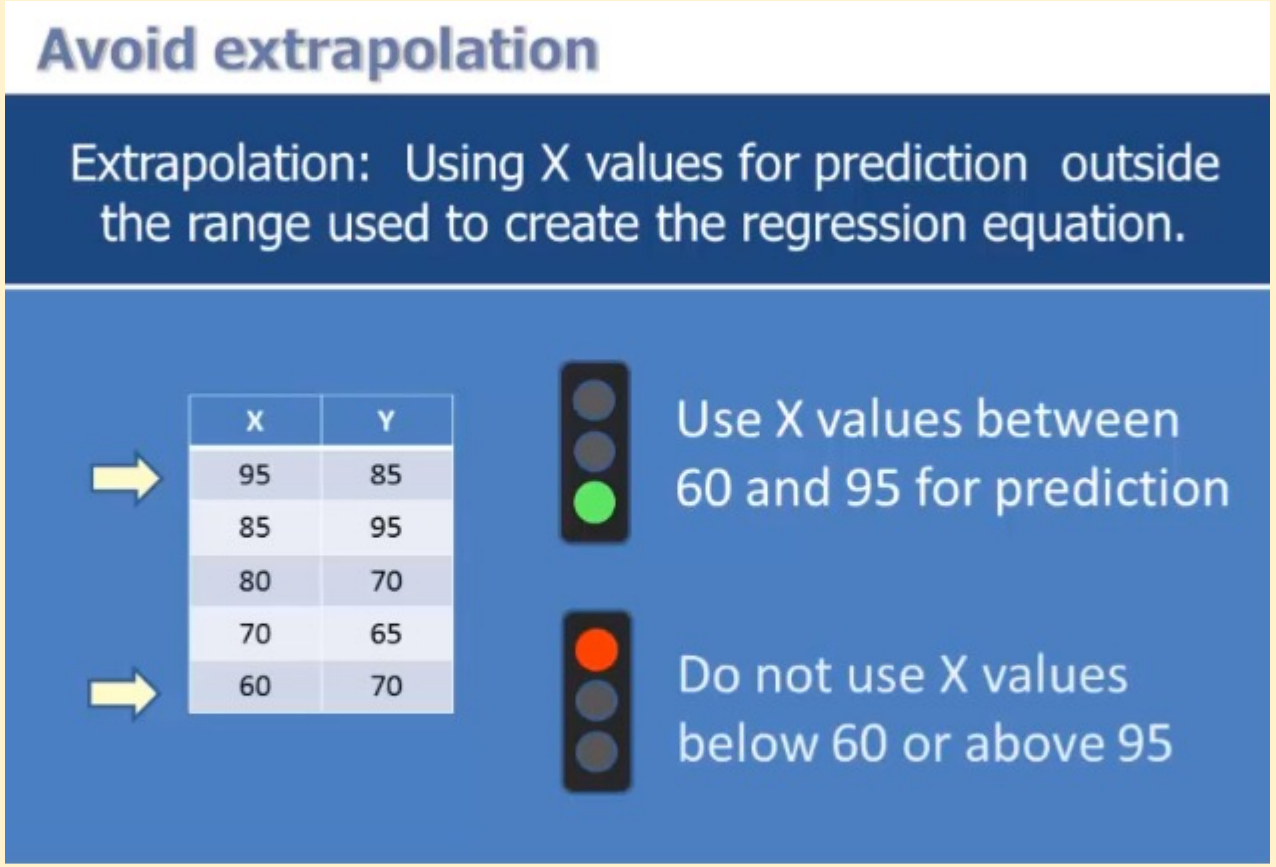

Warning!
Did you know?
$R^2 = r_{xy}^2$
only works for
simple linear regression

Regression for prediction

Given X, what is Y?

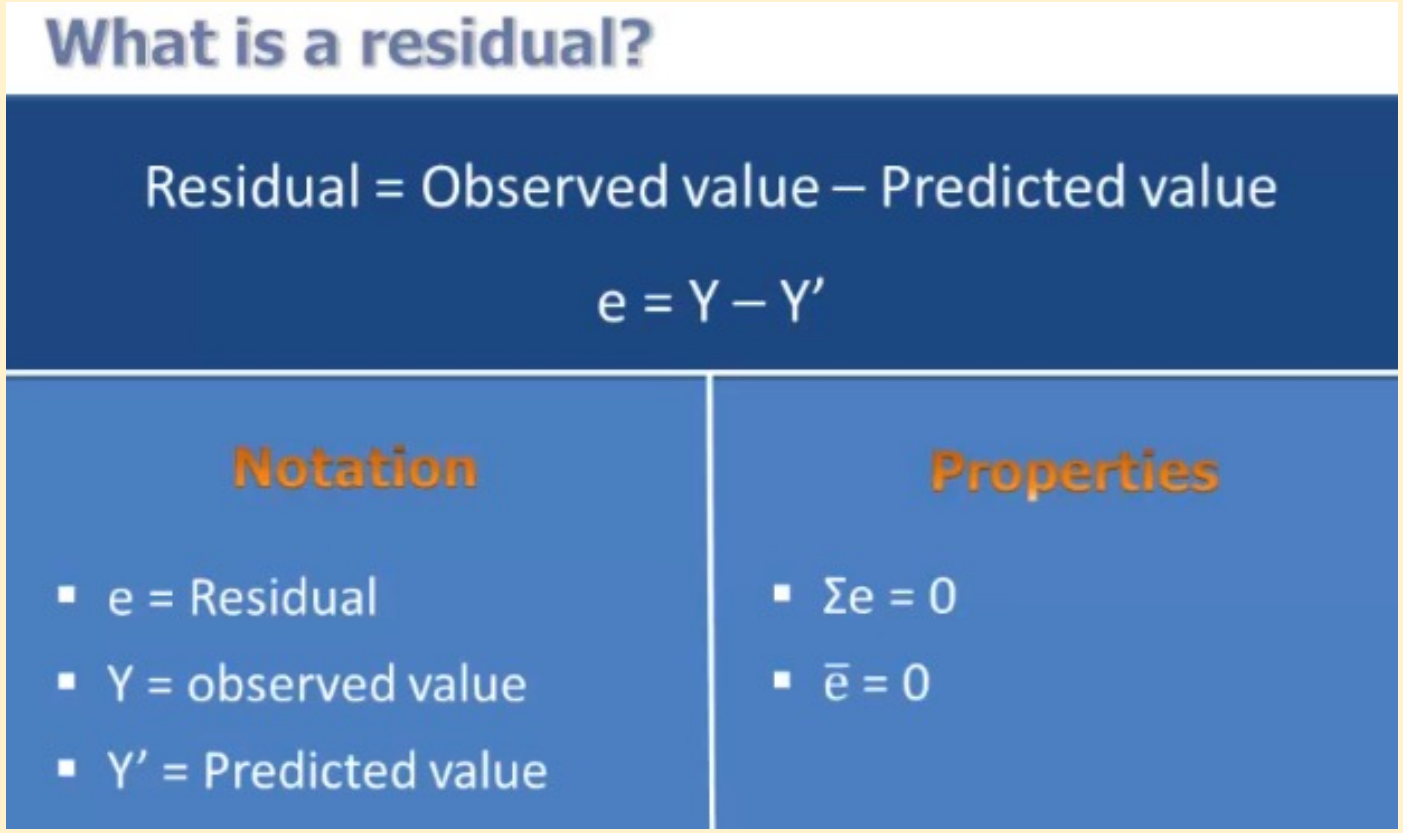
Problem	Solution
X = math aptitude score	$Y' = 26.77 + 0.64 * X$
Y = statistics grade	$Y' = 26.77 + 0.64 * 75$
If X = 75, find Y′	$Y' = 74.8$

Warning: When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called extrapolation, and it can produce unreasonable estimates.

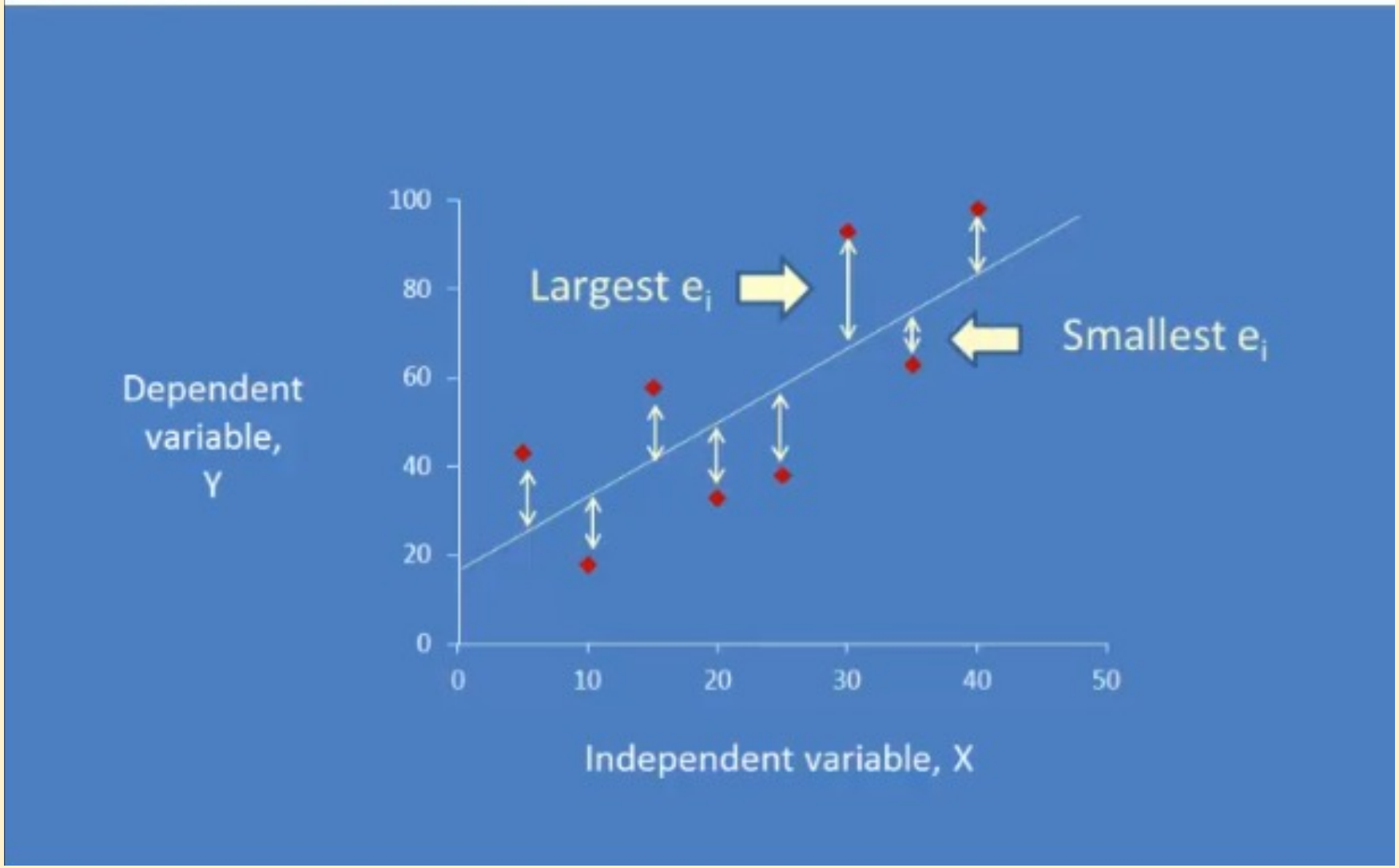


Residual Analysis in Regression

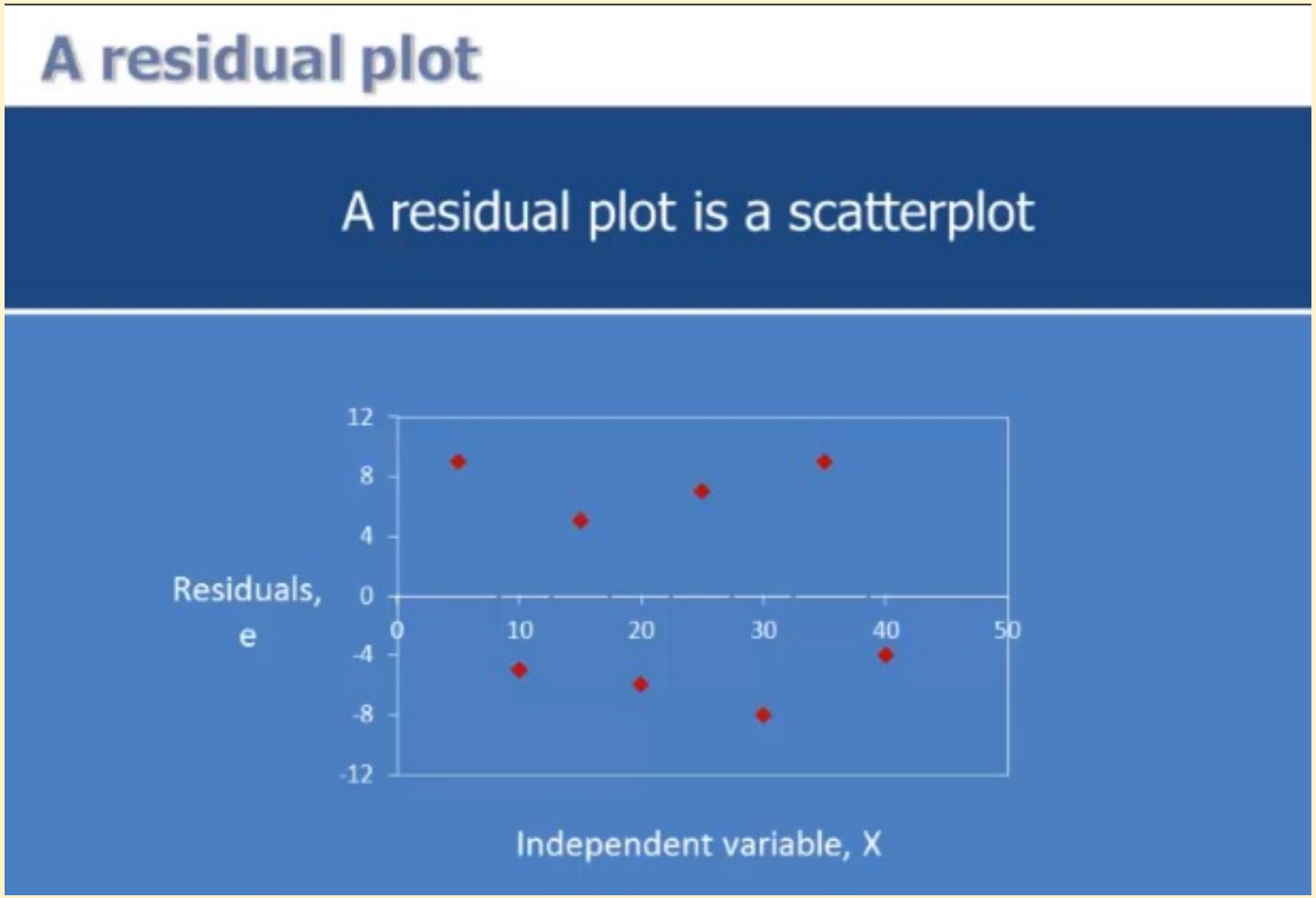
- Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.



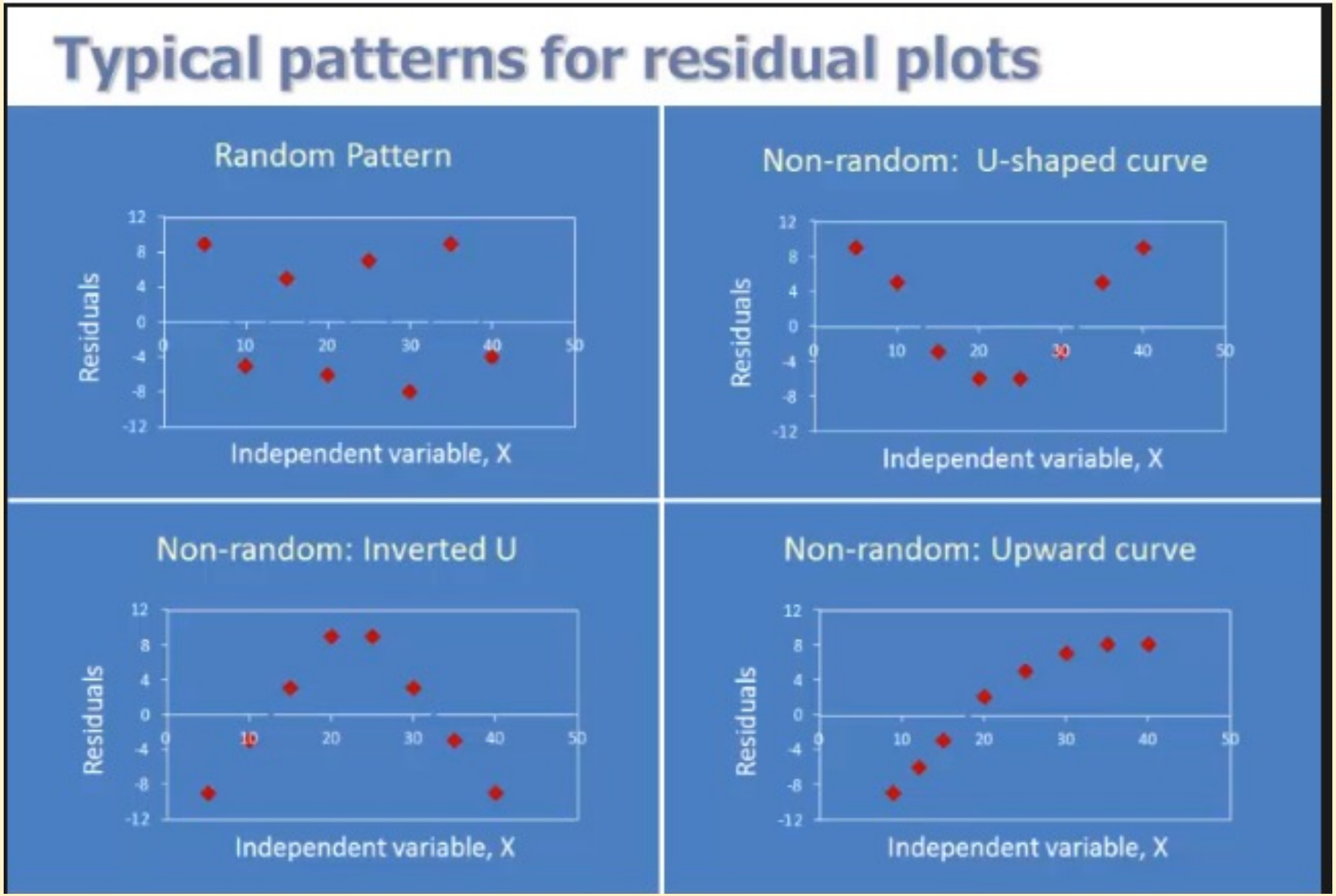
Regression error



A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

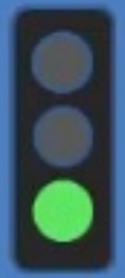


Typical patterns for residual plots



How to use residual plots

Is linear regression appropriate?



Random pattern: Use linear regression

Non-random: Consider other technique

In the context of regression analysis, which of the following statements are true?

I. When the sum of the residuals is greater than zero, the data set is nonlinear.
II. A random pattern of residuals supports a linear model.
III. A random pattern of residuals supports a non-linear model.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I and III

Solution

The correct answer is **(B)**. A random pattern of residuals supports a linear model; a non-random pattern supports a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.

R - Linear Regression

The general mathematical equation for a linear regression is –

```
y = ax + b
```

Following is the description of the parameters used –

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

Steps to Establish a Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.

The steps to create the relationship is –

- Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R.

Input Data

Below is the sample data representing the observations –

```
# Values of height
151, 174, 138, 186, 128, 136, 179, 163, 152, 131

# Values of weight.
63, 81, 56, 91, 47, 57, 76, 72, 62, 48
```

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in linear regression is –

```
lm(formula,data)
```

Following is the description of the parameters used –

- **formula** is a symbol presenting the relation between x and y.
- **data** is the vector on which the formula will be applied.

Create Relationship Model & get the Coefficients

```
1  x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
2  y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
3
4  # Apply the lm() function.
5  relation <- lm(y~x)
6
7  print(relation)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   -38.4551       0.6746
```

Get the Summary of the Relationship

print(summary(relation))

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3002 -1.6629  0.0412  1.8944  3.9775

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509    8.04901  -4.778  0.00139 **
x             0.67461    0.05191  12.997 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

```
       1
76.22869
```

predict() Function

Syntax

The basic syntax for predict() in linear regression is –

```
predict(object, newdata)
```

Following is the description of the parameters used –

- **object** is the formula which is already created using the lm() function.
- **newdata** is the vector containing the new value for predictor variable.

```
# The predictor vector.
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

# The response vector.
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)

# Apply the lm() function.
relation <- lm(y~x)

# Find weight of a person with height 170.
a <- data.frame(x = 170)
result <-  predict(relation,a)
print(result)
```

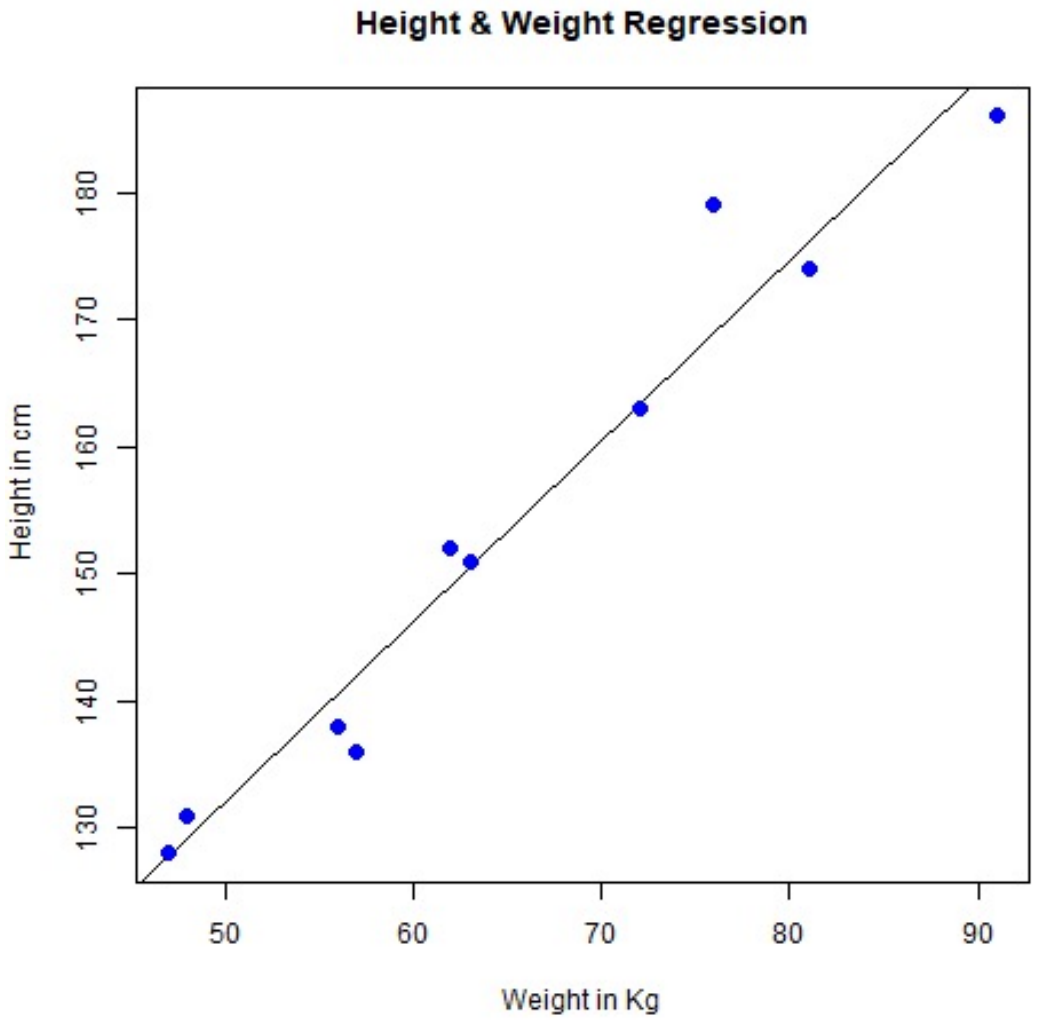
Visualize the Regression Graphically

```
# Create the predictor and response variable.
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
relation <- lm(y~x)

# Give the chart file a name.
png(file = "linearregression.png")

# Plot the chart.
plot(y,x,col = "blue",main = "Height & Weight Regression",
abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "Weight in Kg",ylab =
"Height in cm")

# Save the file.
dev.off()
```



Reference

- http://stattrek.com/regression/regression-example.aspx?Tutorial=AP
- Book - Statistical Methods for Categorical Data Analysis, By Daniel Powers, Yu Xie
- https://www.tutorialspoint.com/r/r_linear_regression.htm
- https://www.statisticssolutions.com/what-is-linear-regression/
- http://onlinestatbook.com/2/regression/intro.html
- https://www.dezyre.com/data-science-in-r-programming-tutorial/linear-regression-tutorial
- https://www.datasciencecentral.com/profiles/blogs/data-science-simplified-part-4-simple-linear-regression-models-1