Regression

Tushar B. Kute,
http://tusharkute.com

Statistical Data Analysis

- Statistical data analysis is a procedure of performing various statistical operations.
- It is a kind of quantitative research, which seeks to quantify the data, and typically, applies some form of statistical analysis.
- Quantitative data basically involves descriptive data, such as survey data and observational data.

Qualitative Data Type

- Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes.
- It means that this type of data can’t be counted or measured easily using numbers and therefore divided into categories.
- The gender of a person (male, female, or others) is a good example of this data type.

Qualitative Data Type

- These are usually extracted from audio, images, or text medium.
- Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on.
- All this information can be categorized as Qualitative data. There are two subcategories under this:
  - Nominal
  - Ordinal

Nominal

- These are the set of values that don’t possess a natural ordering.
- Example: The color of a smartphone can be considered as a nominal data type as we can’t compare one color with others.
- It is not possible to state that ‘Red’ is greater than ‘Blue’.
- The gender of a person is another one where we can’t differentiate between male, female, or others.
- Mobile phone categories whether it is midrange, budget segment, or premium smartphone is also nominal data type.

Ordinal

- These types of values have a natural ordering while maintaining their class of values.
- If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large.
- The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.

Ordinal

- These categories help us deciding which encoding strategy can be applied to which type of data.
- Data encoding for Qualitative data is important because machine learning models can’t handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.
- For nominal data type where there is no comparison among the categories, one-hot encoding can be applied which is similar to binary coding considering there are in less number and for the ordinal data type, label encoding can be applied which is a form of integer encoding.

Quantitative Data Type

- This data type tries to quantify things and it does by considering numerical values that make it countable in nature.
- The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.

Quantitative Data Type

- The key thing is that there can be an infinite number of values a feature can take.
- For instance, the price of a smartphone can vary from x amount to any value and it can be further broken down based on fractional values.
- The two subcategories which describe them clearly are:
  - Discrete
  - Continuous

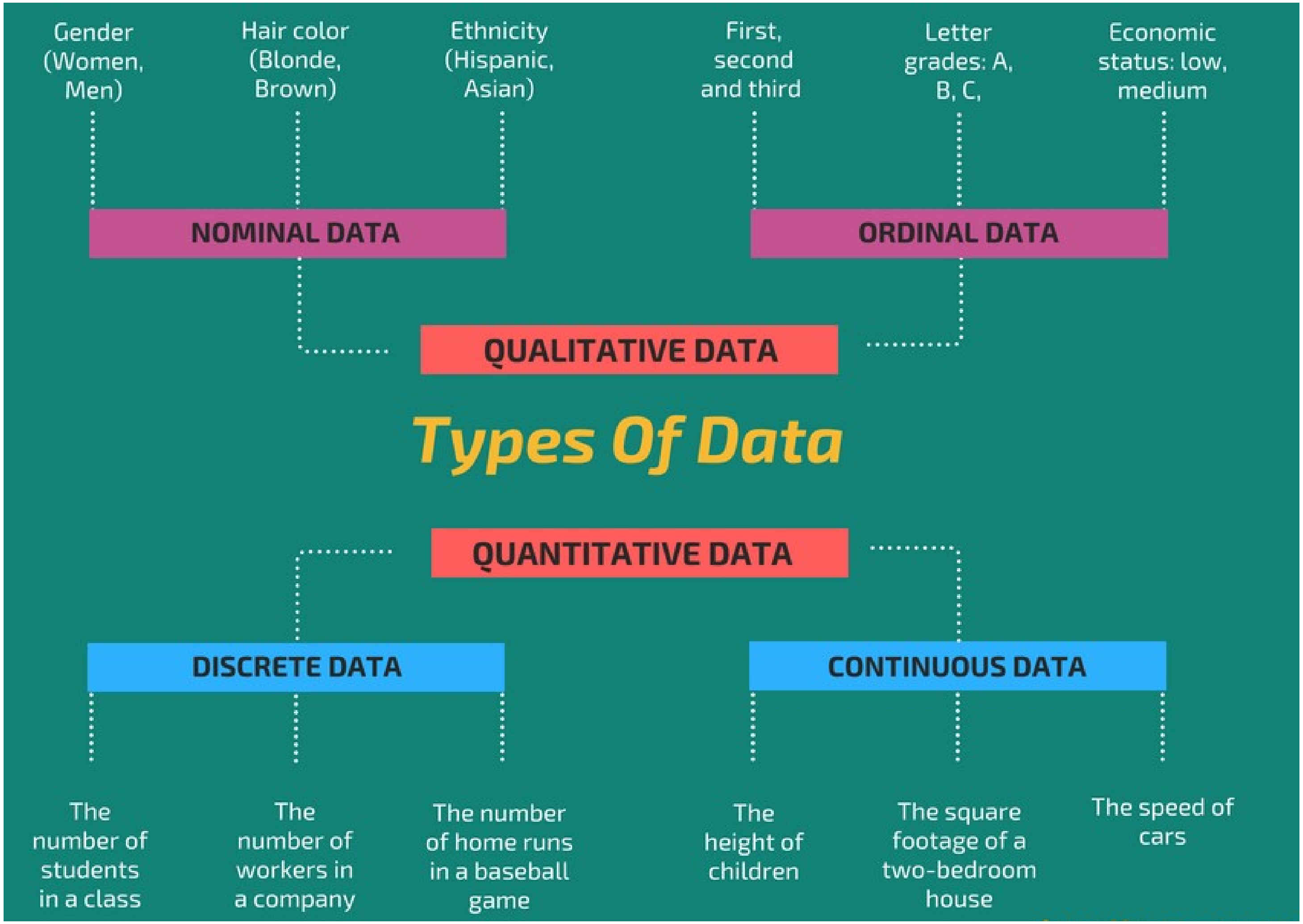
Discrete

- The numerical values which fall under are integers or whole numbers are placed under this category.
- The number of speakers in the phone, cameras, cores in the processor, the number of sims supported all these are some of the examples of the discrete data type.

Continous

- The fractional numbers are considered as continuous values.
- These can take the form of the operating frequency of the processors, the android version of the phone, wifi frequency, temperature of the cores, and so on.

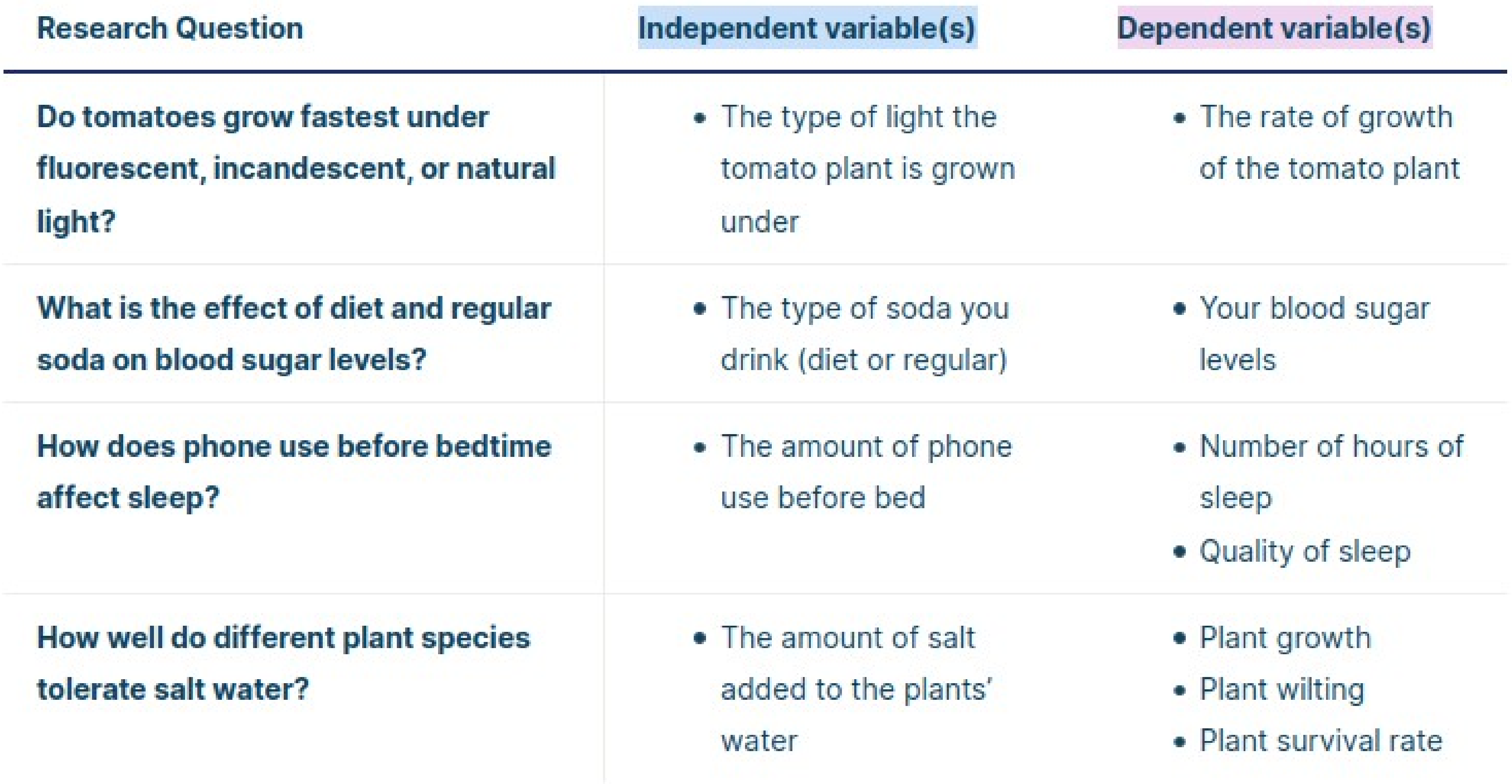
Summary


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

Types of variables

- In research, variables are any characteristics that can take on different values, such as height, age, species, or exam score.
- In scientific research, we often want to study the effect of one variable on another one. For example, you might want to test whether students who spend more time studying get better exam scores.
- The variables in a study of a cause-and-effect relationship are called the independent and dependent variables.
  - The independent variable is the cause. Its value is independent of other variables in your study.
  - The dependent variable is the effect. Its value depends on changes in the independent variable.

Types of variables

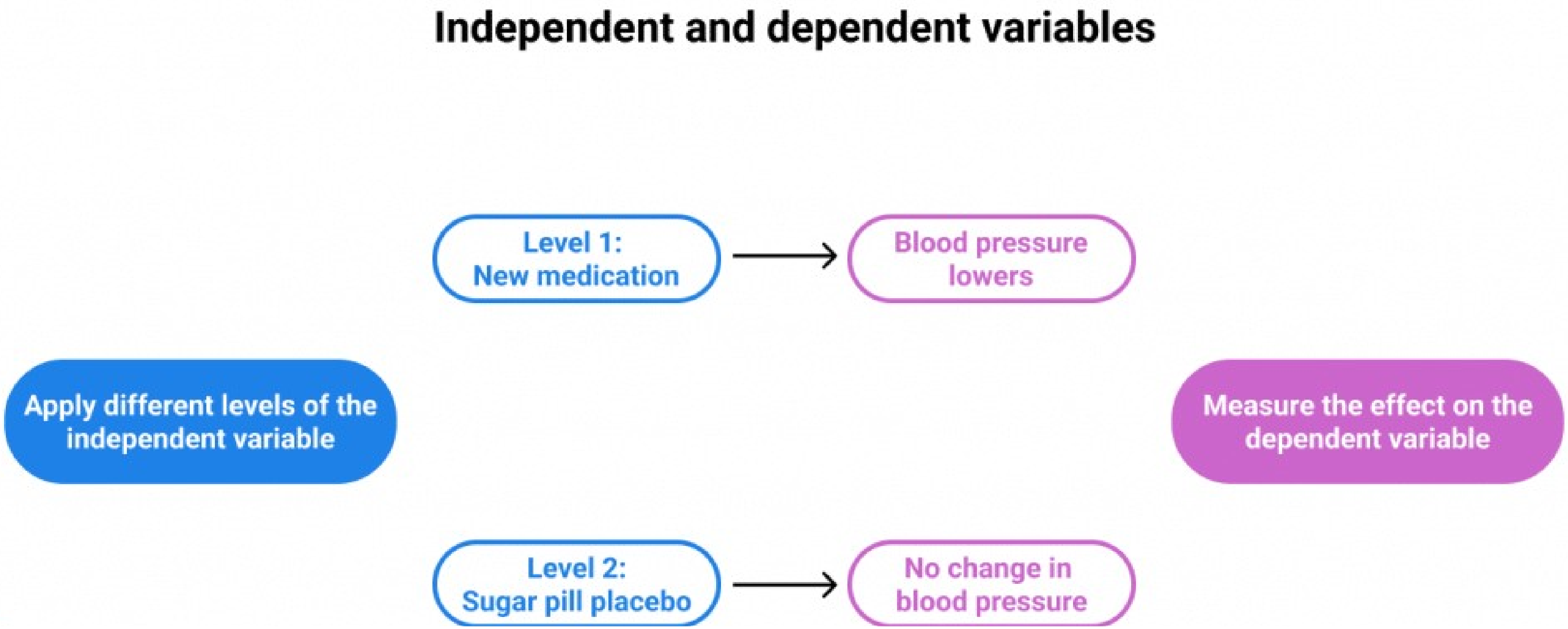


Research Question	Independent variable(s)	Dependent variable(s)
Do tomatoes grow fastest under fluorescent, incandescent, or natural light?	• The type of light the tomato plant is grown under	• The rate of growth of the tomato plant
What is the effect of diet and regular soda on blood sugar levels?	• The type of soda you drink (diet or regular)	• Your blood sugar levels
How does phone use before bedtime affect sleep?	• The amount of phone use before bed	• Number of hours of sleep • Quality of sleep
How well do different plant species tolerate salt water?	• The amount of salt added to the plants' water	• Plant growth • Plant wilting • Plant survival rate

Example:

- You are studying the impact of a new medication on the blood pressure of patients with hypertension.
- To test whether the medication is effective, you divide your patients into two groups. One group takes the medication, while the other group takes a sugar pill placebo.
  - Your independent variable is the treatment that you vary between groups: which type of pill the patient receives.
  - Your dependent variable is the outcome that you measure: the blood pressure of the patients.

Example:


Independent and dependent variables

Level 1:
New medication
Blood pressure lowers
Apply different levels of the independent variable
Measure the effect on the dependent variable
Level 2:
Sugar pill placebo
No change in blood pressure

Example:

- Imagine that a tutor asks 100 students to complete a maths test. The tutor wants to know why some students perform better than others. Whilst the tutor does not know the answer to this, she thinks that it might be because of two reasons:
  - (1) some students spend more time revising for their test; and (2) some students are naturally more intelligent than others. As such, the tutor decides to investigate the effect of revision time and intelligence on the test performance of the 100 students.
- Dependent Variable: Test Mark (measured from 0 to 100)
- Independent Variables: Revision time (measured in hours) Intelligence (measured using IQ score)

Other names for independent variables

- Sometimes, the variable you think is the cause might not be fully independent – it might be influenced by other variables. In this case, one of these terms is more appropriate:
  - Explanatory variables (they explain an event or outcome)
  - Predictor variables (they can be used to predict the value of a dependent variable)
  - Right-hand-side variables (they appear on the right-hand side of a regression equation).

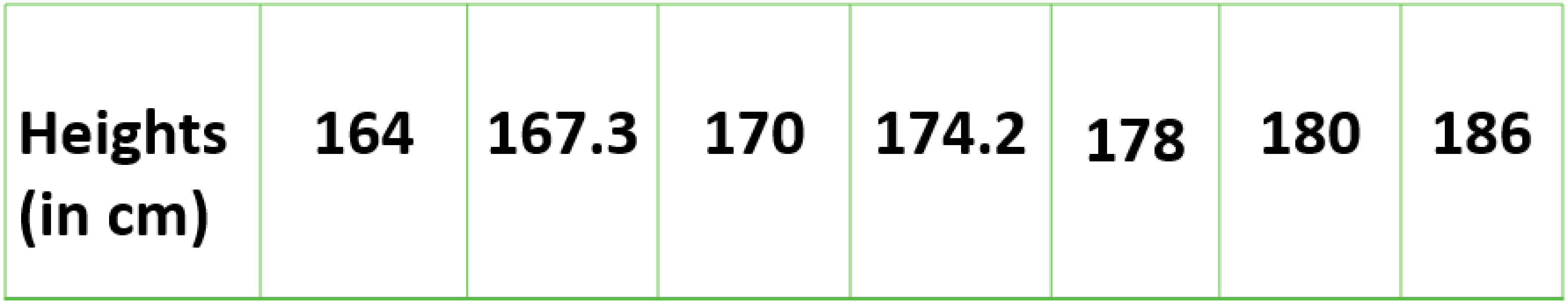
Other names for dependent variables

- Dependent variables are also known by these terms:
  - Response variables (they respond to a change in another variable)
  - Outcome variables (they represent the outcome you want to measure)
  - Left-hand-side variables (they appear on the left-hand side of a regression equation)

Univatiate Data

- This type of data consists of only one variable.
- The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
- It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
- The example of a univariate data can be height.

Univatiate Data



Heights (in cm)	164	167.3	170	174.2	178	180	186

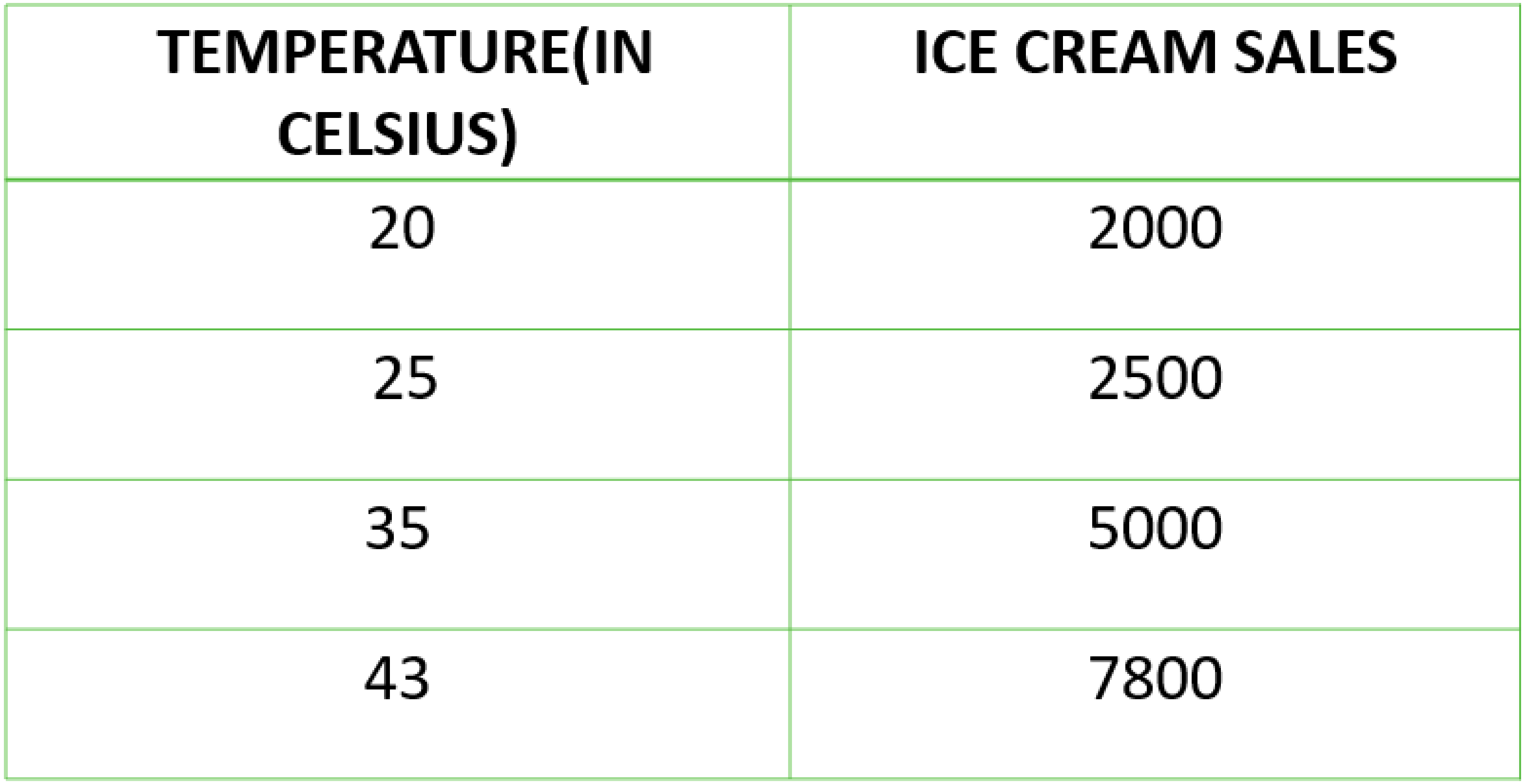
Univatiate Data

- Suppose that the heights of seven students of a class is recorded, there is only one variable that is height and it is not dealing with any cause or relationship.
- The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

Bivatiate Data

- This type of data involves two different variables.
- The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- Example of bivariate data can be temperature and ice cream sales in summer season.

Bivatiate Data



TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Bivatiate Data

- Suppose the temperature and ice cream sales are the two variables of a bivariate data.
- Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
- Thus bivariate data analysis involves comparisons, relationships, causes and explanations.
- These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

Multivatiate Data

- When the data involves three or more variables, it is categorized under multivariate.
- Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

Multivatiate Data

- It is similar to bivariate but contains more than one dependent variable.
- The ways to perform analysis on this data depends on the goals to be achieved.Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

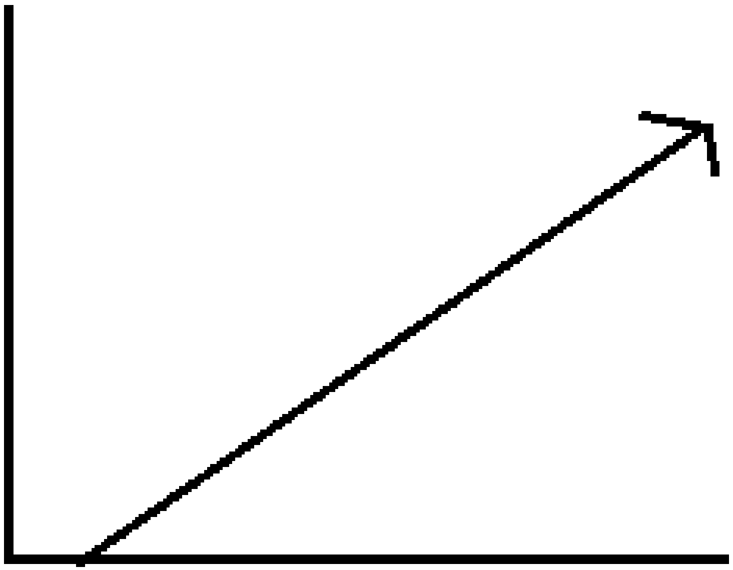
Regression?

- Regression analysis is a statistical method that helps us to analyse and understand the relationship between two or more variables of interest.
- The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored and how they are influencing each other.

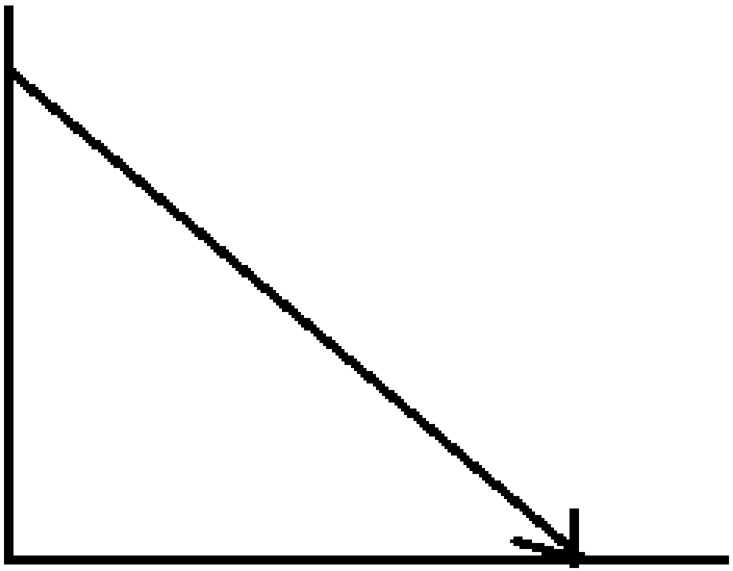
Regression?

- For the regression analysis is be a successful method, we understand the following terms:
  - Dependent Variable: This is the variable that we are trying to understand or forecast.
  - Independent Variable: These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

Linear Regression

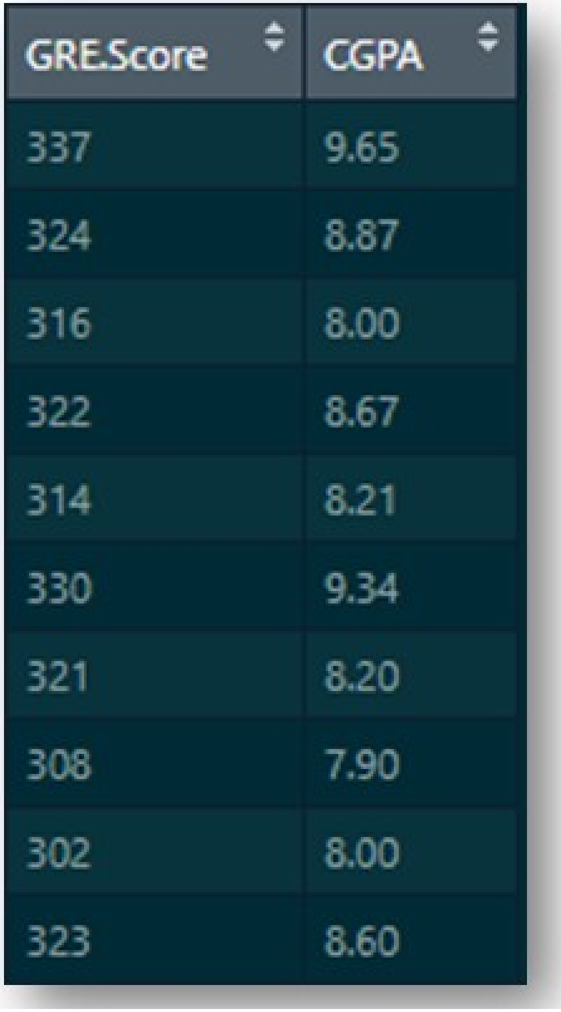


Positive Linear Relationship



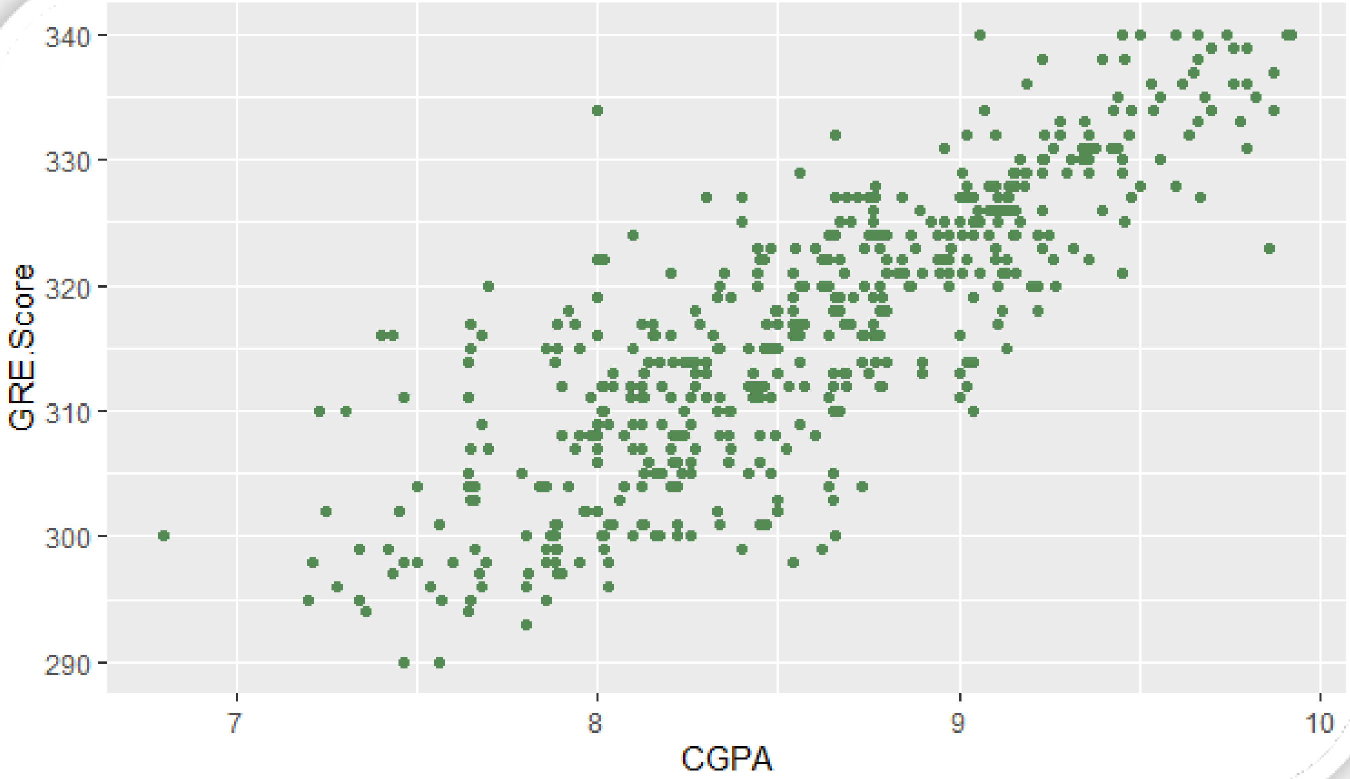
Negative Linear Relationship

Example:



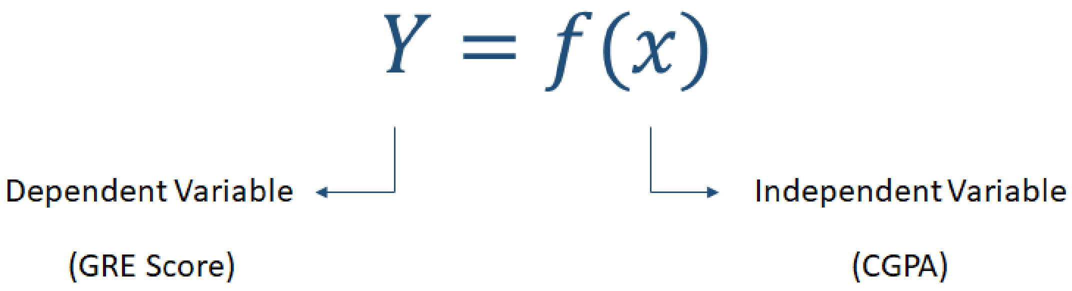
GRE.Score	CGPA
337	9.65
324	8.87
316	8.00
322	8.67
314	8.21
330	9.34
321	8.20
308	7.90
302	8.00
323	8.60

Example:



Regression

- In regression, we normally have one dependent variable and one or more independent variables.
- Here we try to "regress" the value of dependent variable "Y" with the help of the independent variables.
- In other words, we are trying to understand, how does the value of 'Y' change w.r.t change in 'X'.



$$Y = f(x)$$

Dependent Variable

(GRE Score)

Independent Variable

(CGPA)

Uses of Regression

- Regression analysis is used for prediction and forecasting. This has a substantial overlap to the field of machine learning. This statistical method is used across different industries such as,
  - Financial Industry- Understand the trend in the stock prices, forecast the prices, evaluate risks in the insurance domain
  - Marketing- Understand the effectiveness of market campaigns, forecast pricing and sales of the product.
  - Manufacturing- Evaluate the relationship of variables that determine to define a better engine to provide better performance
  - Medicine- Forecast the different combination of medicines to prepare generic medicines for diseases.

Terminologies

- Outliers
  - Suppose there is an observation in the dataset that has a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier.
  - In simple words, it is an extreme value. An outlier is a problem because many times it hampers the results we get.

Terminologies

- Multicollinearity
  - When the independent variables are highly correlated to each other, then the variables are said to be multicollinear.
  - Many types of regression techniques assume multicollinearity should not be present in the dataset.
  - It is because it causes problems in ranking variables based on its importance, or it makes the job difficult in selecting the most important independent variable.

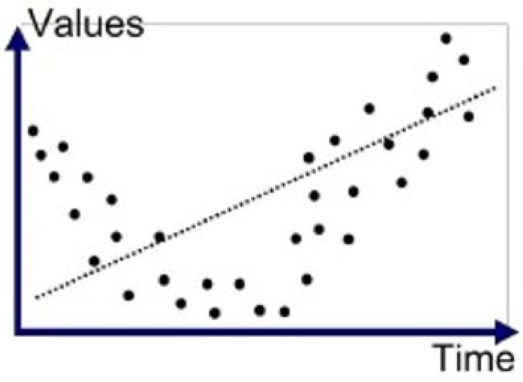
Terminologies

- Heteroscedasticity
  - When the variation between the target variable and the independent variable is not constant, it is called heteroscedasticity.
  - Example-As one’s income increases, the variability of food consumption will increase.
  - A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times, eat expensive meals.
  - Those with higher incomes display a greater variability of food consumption.

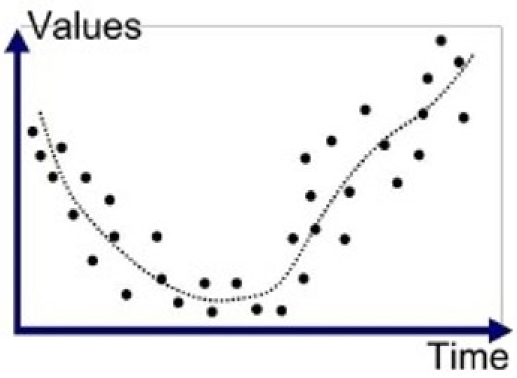
Terminologies

- When we use unnecessary explanatory variables, it might lead to overfitting.
- Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets. It is also known as a problem of high variance.
- When our algorithm works so poorly that it is unable to fit even a training set well, then it is said to underfit the data. It is also known as a problem of high bias.

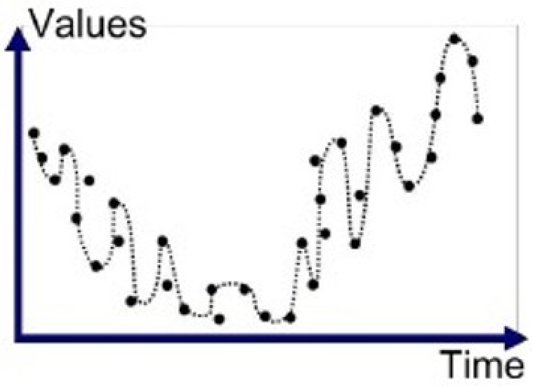
Terminologies



Underfitted



Good Fit/Robust



Overfitted

Types of Regression

- Linear Regression
- Multiple Regression
- Logistic Regression
- Polynomial Regression
- Regularized Models
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
- Outlier Based Model
  - RANSAC

Linear Regression

- The simplest of all regression types is Linear Regression where it tries to establish relationships between Independent and Dependent variables.
- The Dependent variable considered here is always a continuous variable.
- Linear Regression is a predictive model used for finding the linear relationship between a dependent variable and one or more independent variables.

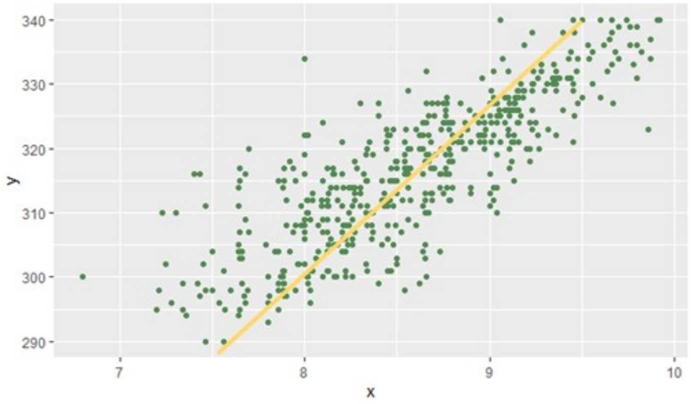
$$Y = a + bx$$

Dependent Variable is continuous

Linear Regression

- Here, ‘Y’ is our dependent variable, which is a continuous numerical and we are trying to understand how does ‘Y’ change with ‘X’.
- So, if we are supposed to answer, the above question of “What will be the GRE score of the student, if his CCGPA is 8.32?” our go to option should be linear regression.

$GRE = 261 + 6.8CGPA$ → $GRE = 261 + 6.8(8.32)$ → $GRE = 317.57$



Simple Linear Regression

- As the model is used to predict the dependent variable, the relationship between the variables can be written in the below format.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Where,
  - Y_i – Dependent variable
  - β_0 — Intercept
  - β_1 – Slope Coefficient
  - X_i – Independent Variable
  - ε_i – Random Error Term

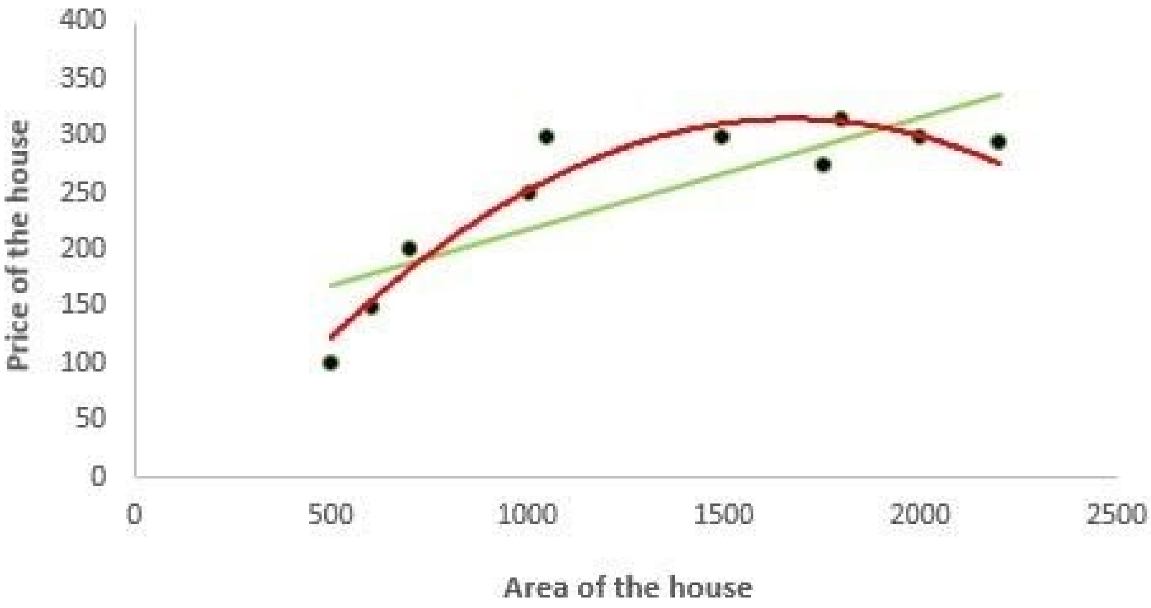
Simple Linear Regression

- The main factor that is considered as part of Regression analysis is understanding the variance between the variables. For understanding the variance, we need to understand the measures of variation.
  - SST = total sum of squares (Total Variation)
    - Measures the variation of the Y i values around their mean Y
  - SSR = regression sum of squares (Explained Variation)
    - Variation attributable to the relationship between X and Y
  - SSE = error sum of squares (Unexplained Variation)
    - Variation in Y attributable to factors other than X

Polynomial Regression

- This type of regression technique is used to model nonlinear equations by taking polynomial functions of independent variables.
- In the figure given below, you can see the red curve fits the data better than the green curve.
- Hence in the situations where the relationship between the dependent and independent variable seems to be non-linear, we can deploy Polynomial Regression Models.

Polynomial Regression



$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

Multiple Linear Regression

- Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:
  - How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
  - The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Multiple Linear Regression – Ex.

- You are a public health researcher interested in social factors that influence heart disease.
- You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.
- Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

Assumptions

- Multiple linear regression makes all of the same assumptions as simple linear regression:
- Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.
- Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

Assumptions

- In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model.
- If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
- Normality: The data follows a normal distribution.
- Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

How to perform?

- The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

  - y = the predicted value of the dependent variable
  - B_0 = the y-intercept (value of y when all other parameters are set to 0)
  - B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
  - … = do the same for however many independent variables you are testing
  - B_nX_n = the regression coefficient of the last independent variable
  - e = model error (a.k.a. how much variation there is in our estimate of y)

How to perform?

- To find the best-fit line for each independent variable, multiple linear regression calculates three things:
  - The regression coefficients that lead to the smallest overall model error.
  - The t-statistic of the overall model.
  - The associated p-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).
- It then calculates the t-statistic and p-value for each regression coefficient in the model.

Performance Evaluation

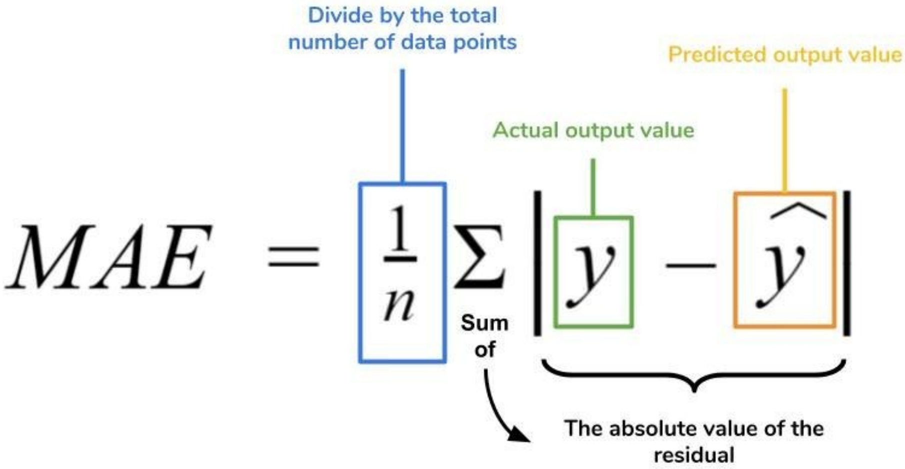
- The performance of a regression model can be understood by knowing the error rate of the predictions made by the model.
- You can also measure the performance by knowing how well your regression line fit the dataset.
- A good regression model is one where the difference between the actual or observed values and predicted values for the selected model is small and unbiased for train, validation and test data sets.

Performance Evaluation

- To measure the performance of your regression model, some statistical metrics are used. Here we will discuss four of the most popular metrics. They are-
  - Mean Absolute Error(MAE)
  - Root Mean Square Error(RMSE)
  - Coefficient of determination or R2
  - Adjusted R2

Mean Absolute Error

- This is the simplest of all the metrics. It is measured by taking the average of the absolute difference between actual values and the predictions.



$$MAE = \boxed{\frac{1}{n}} \sum \left| \boxed{y} - \boxed{\hat{y}} \right|$$

Divide by the total number of data points

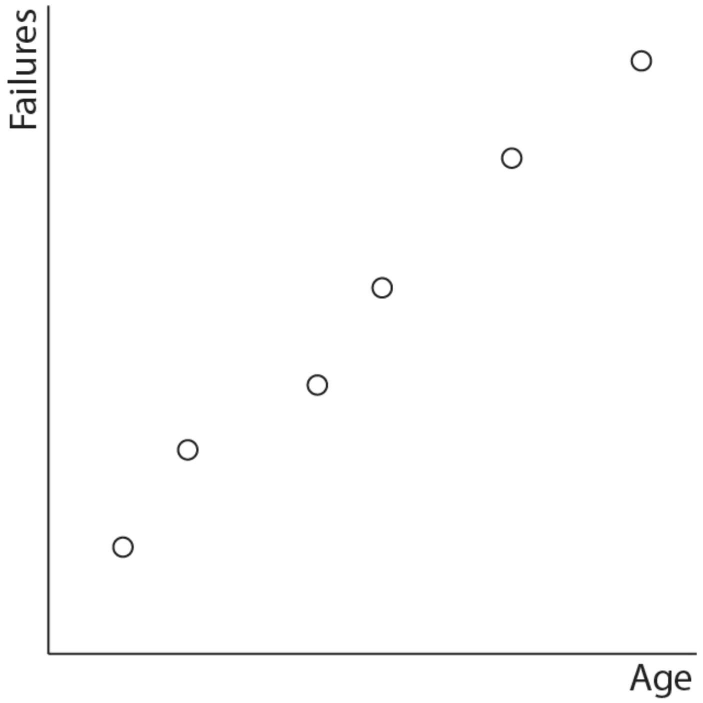
Predicted output value

Actual output value

Sum of

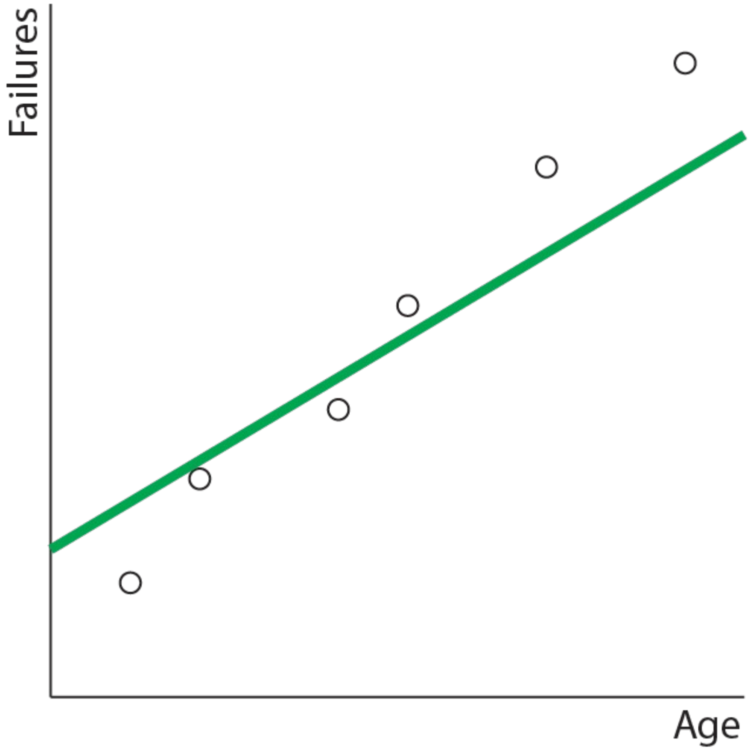
The absolute value of the residual

Example:



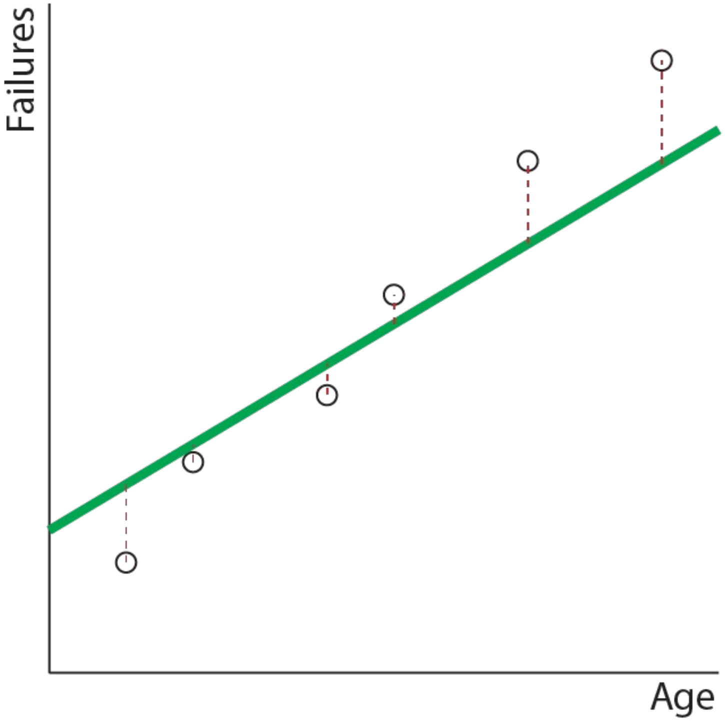
Age	Failures
10	15
20	30
40	40
50	55
70	75
90	90

Example:



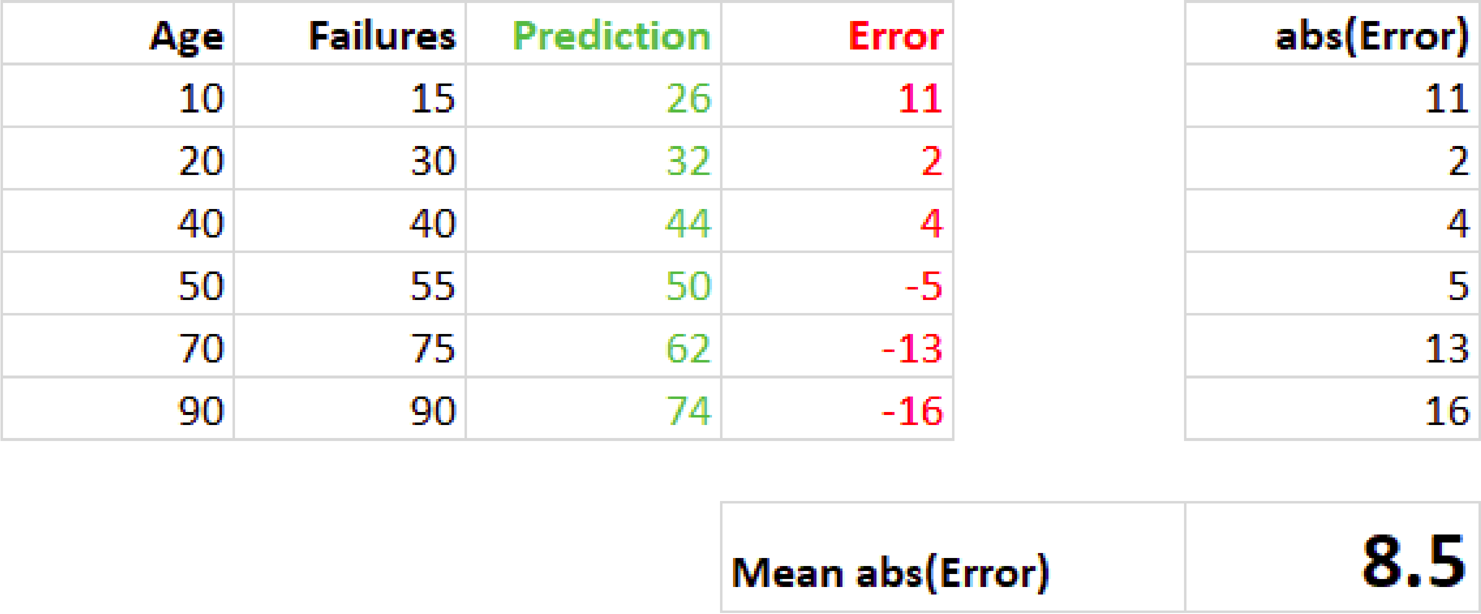
Age	Failures	Prediction
10	15	26
20	30	32
40	40	44
50	55	50
70	75	62
90	90	74

Example:



Age	Failures	Prediction	Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

Mean Absolute Error



Age	Failures	Prediction	Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

abs(Error)
11
2
4
5
13
16

Mean abs(Error)	8.5

Mean Absolute Error

	y	$\hat{y}$	$y-\hat{y}$
Age	Failures	Prediction	Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

$\lvert y-\hat{y}\rvert$
abs(Error)
11
2
4
5
13
16

Mean abs(Error)	$\frac{\Sigma\lvert y-\hat{y}\rvert}{N}$	**8.5**

Mean Absolute Error

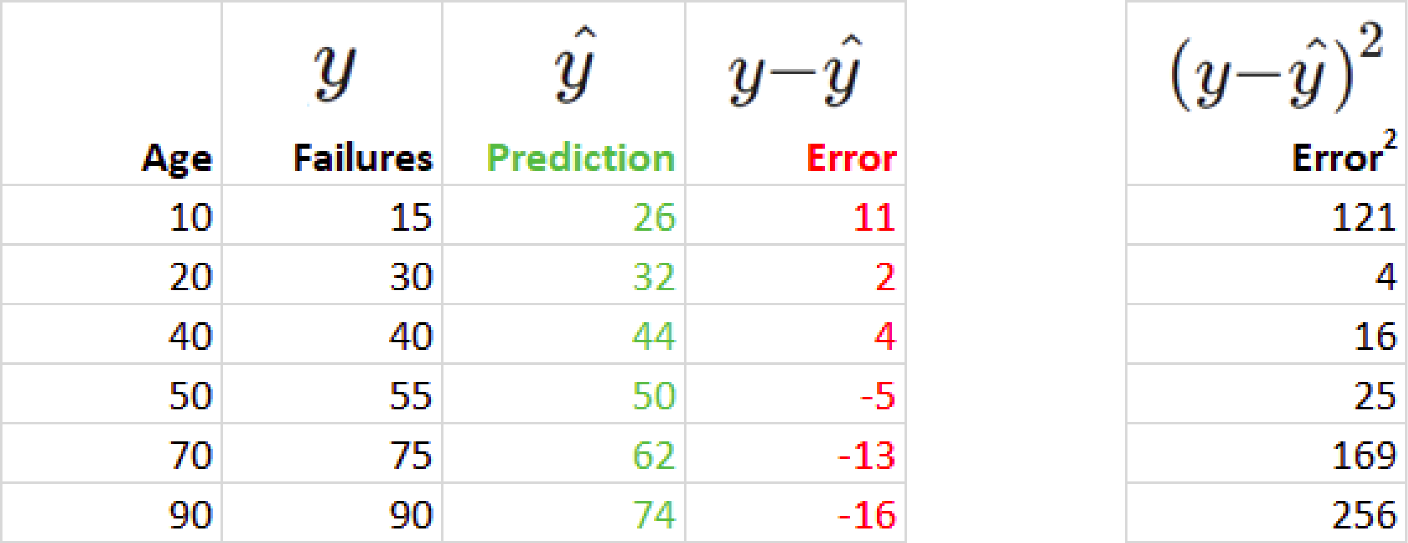
- Mean Absolute Error (MAE) tells us the average error in units of y, the predicted feature. A value of 0 indicates a perfect fit, i.e. all our predictions are spot on.
- The MAE has a big advantage in that the units of the MAE are the same as the units of y, the feature we want to predict.
- In the example above, we have an MAE of 8.5, so it means that on average our predictions of the number of machine failures are incorrect by 8.5 machine failures.
- This makes MAE very intuitive and the results are easily conveyed to a non-machine learning expert!

Root Mean Square Error

- The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value.
- It represents the sample standard deviation of the differences between predicted values and observed values(also called residuals).

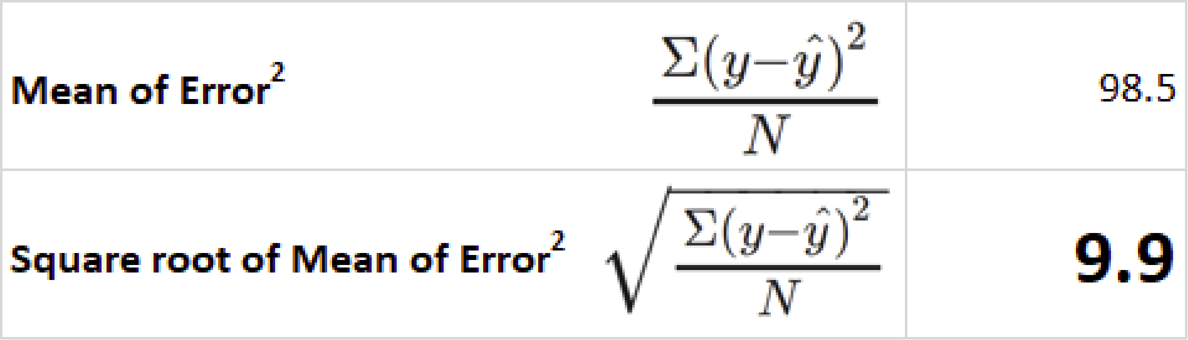
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Root Mean Square Error



	y	$\hat{y}$	$y-\hat{y}$
Age	Failures	Prediction	Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

$(y-\hat{y})^2$
Error^2
121
4
16
25
169
256



Mean of Error^2	$\frac{\Sigma(y-\hat{y})^2}{N}$	98.5
Square root of Mean of Error^2	$\sqrt{\frac{\Sigma(y-\hat{y})^2}{N}}$	**9.9**

RMSE

- As with MAE, we can think of RMSE as being measured in the y units.
- So the above error can be read as an error of 9.9 machine failures on average per observation.

MAE vs. RMSE

- Compared to MAE, RMSE gives a higher total error and the gap increases as the errors become larger. It penalizes a few large errors more than a lot of small errors. If you want your model to avoid large errors, use RMSE over MAE.
- Root Mean Square Error (RMSE) indicates the average error in units of y, the predicted feature, but penalizes larger errors more severely than MAE. A value of 0 indicates a perfect fit.
- You should also be aware that as the sample size increases, the accumulation of slightly higher RMSEs than MAEs means that the gap between these two measures also increases as the sample size increases.

R^2 Error

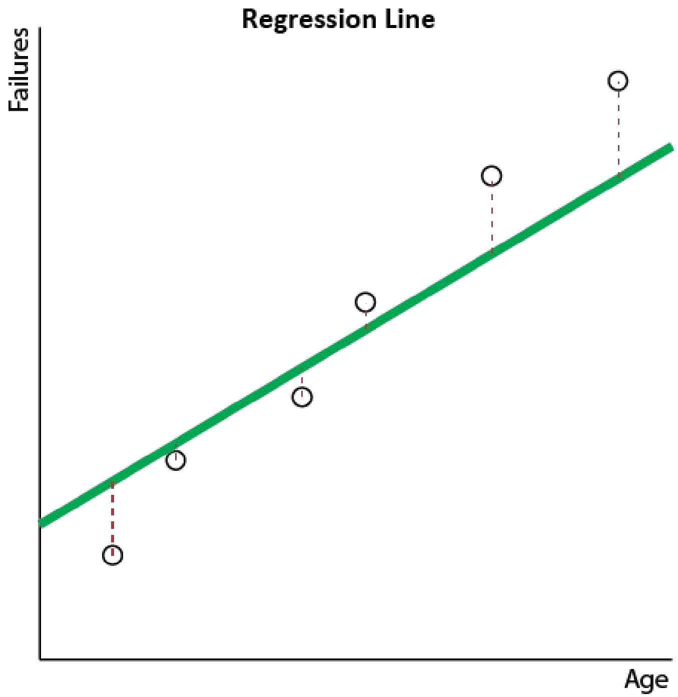
- It measures how well the actual outcomes are replicated by the regression line.
- It helps you to understand how well the independent variable adjusted with the variance in your model.
- That means how good is your model for a dataset. The mathematical representation for R^2 is-

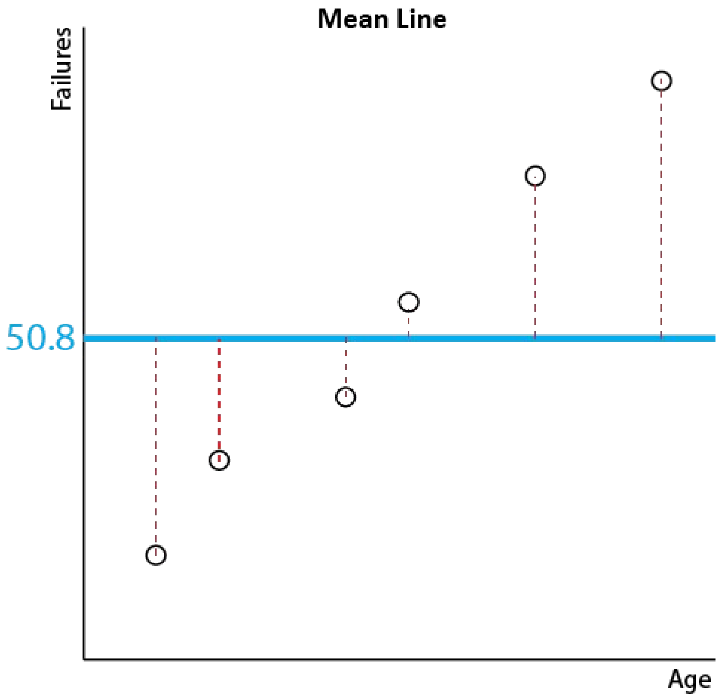
$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \qquad R^2 = \frac{var(mean) - var(line)}{var(mean)}$$

R^2 Error

- Here,
  - SSR = Sum Square of Residuals(the squared difference between the predicted and the average value)
  - SST = Sum Square of Total(the squared difference between the actual and average value)

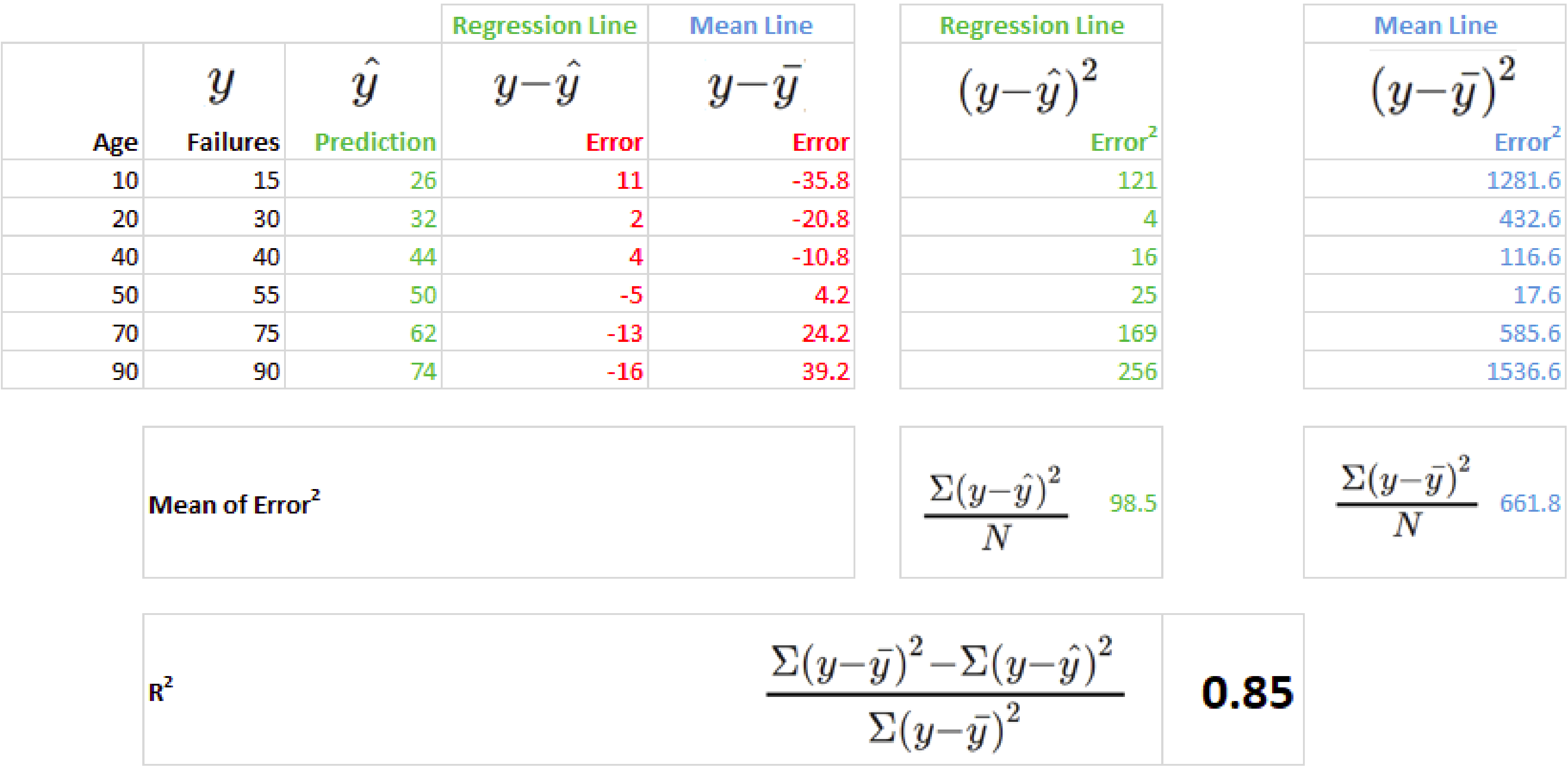
Example:





You can see that the regression line fits the data better than the mean line, which is what we expected (the mean line is a pretty simplistic model, after all). But can you say how much better it is? That's exactly what R2 does! Here is the calculation.

Example:



			Regression Line	Mean Line	Regression Line	Mean Line
	y	$\hat{y}$	$y-\hat{y}$	$y-\bar{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$
Age	Failures	Prediction	Error	Error	Error^2	Error^2
10	15	26	11	-35.8	121	1281.6
20	30	32	2	-20.8	4	432.6
40	40	44	4	-10.8	16	116.6
50	55	50	-5	4.2	25	17.6
70	75	62	-13	24.2	169	585.6
90	90	74	-16	39.2	256	1536.6

Mean of Error^2	$\frac{\Sigma(y-\hat{y})^2}{N}$	98.5	$\frac{\Sigma(y-\bar{y})^2}{N}$	661.8

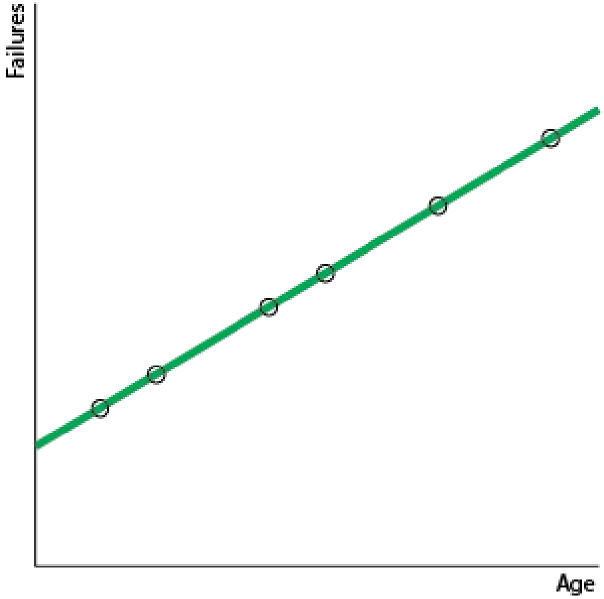
R^2	$\frac{\Sigma(y-\bar{y})^2-\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$	**0.85**

R^2 Error

- The additional parts to the calculation are the column on the far right (in blue) and the final calculation row, computing R^2
- So we have an R-squared of 0.85. Without even worrying about the units of y we can say this is a decent model. Why? Because the model explains 85% of the variation in the data. That's exactly what an R-squared of 0.85 tells us!
- R-squared (R^2) tells us the degree to which the model explains the variance in the data. In other words, how much better it is than just predicting the mean.

Example:

- Here's another example. What if our data points and regression line looked like this?



- The variance around the regression line is 0. In other words, var(line) is 0. There are no errors.

Now,

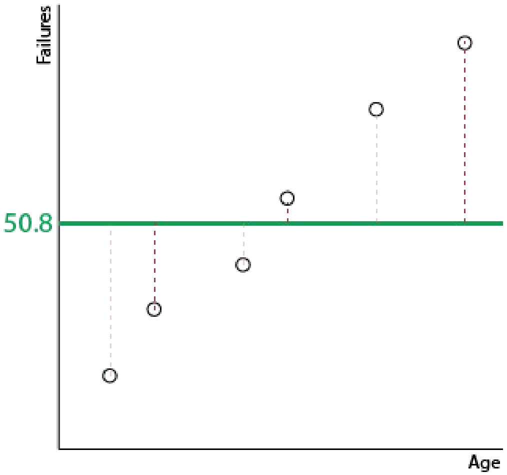
- Now, remember that the formula for R-squared is:
- $$R^2 = \frac{var(mean) - var(line)}{var(mean)}$$
- So, with var(line) = 0 the above calculation for R-squared is

$$R^2 = \frac{var(mean) - 0}{var(mean)} = \frac{var(mean)}{var(mean)} = 1$$

- So, if we have a perfect regression line, with no errors, we get an R-squared of 1.

R^2 Error

- Let's look at another example. What if our data points and regression line looked like this, with the regression line equal to the mean line?
- Data points where the regression line is equal to the mean line

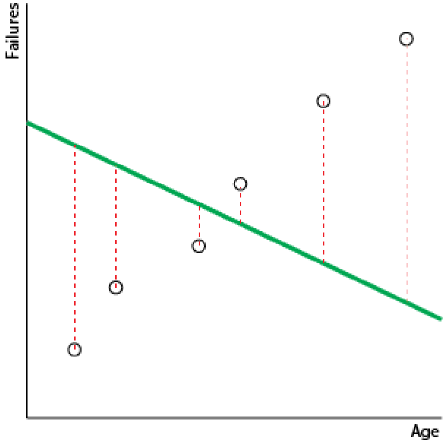


$$R^2 = \frac{var(mean) - var(mean)}{var(mean)} = \frac{0}{var(mean)} = 0$$

- In this case, var(line) and var(mean) are the same. So the above calculation will yield an R-squared of 0:

R^2 Error

- What if our regression line was really bad, worse than the mean line?



- It's unlikely to get this bad! But if it does, var(mean)-var(line) will be negative, so R-squared will be negative.
- An R-squared of 1 indicates a perfect fit. An R-squared of 0 indicates a model no better or worse than the mean. An R-squared of less than 0 indicates a model worse than just predicting the mean.

Adjusted R^2 Error

- Adjusted R-Squared is a modified form of R-Squared whose value increases if new predictors tend to improve models performance and decreases if new predictors does not improve performance as expected.
- R-squared is a comparison of Residual sum of squares (SSres) with total sum of squares(SStot).
- It is calculated by dividing sum of squares of residuals from the regression model by total sum of squares of errors from the average model and then subtract it from 1.

Adjusted R^2 Error

- Unlike R-squared, the Adjusted R-squared would penalize you for adding features which are not useful for predicting the target.
- It takes into account the number of independent variables used for predicting the target variable.

Adjusted R^2 Error

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

- where,
  - N = number of records in the data set.
  - p = number of independent variables.

Adjusted R² Error

- For a simple representation, you can rewrite the above formula like the following:

  Adjusted R-squared = 1 — (x * y)

  where,

  x = 1 — R Squared

  y = (N-1) / (n-p-1)

- Adjusted R-squared can be negative when R-squared is close to zero.
- Adjusted R-squared value always be less than or equal to R-squared value.

Mean Absolute Percentage Error (MAPE)

- The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is.
- It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

Mean Absolute Percentage Error (MAPE)

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

- Where:
  - n is the number of fitted points,
  - A_t is the actual value,
  - F_t is the forecast value.
  - Σ is summation notation (the absolute value is summed for every forecasted point in time).

Summary

- Mean Absolute Error (MAE) tells us the average error in units of y, the predicted feature. A value of 0 indicates a perfect fit.
- Root Mean Square Error (RMSE) indicates the average error in units of y, the predicted feature, but penalizes larger errors more severely than MAE. A value of 0 indicates a perfect fit.
- R-squared (R^2) tells us the degree to which the model explains the variance in the data. In other words how much better it is than just predicting the mean.
  - A value of 1 indicates a perfect fit.
  - A value of 0 indicates a model no better than the mean.
  - A value less than 0 indicates a model worse than just predicting the mean.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com