on 7th September, 2021 ~ Tuesday

Module-1: **Predictive Modelling**

MODULE: PREDICTIVE MODELLING

Outlines

- Correlation
  - Pearson's correlation
  - Kendall rank correlation
  - Spearman rank correlation
- Regression
  - Types of regression
  - Fitting a function – Criterion for best fit
  - Least squares
- Simple regression
- Multiple regression
- Model assessment and validation

CORRELATION

Preliminaries

- n observations for x and y variables (x_i, y_i)
- Sample means $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

- Sample variances S_{xx} and S_{yy}

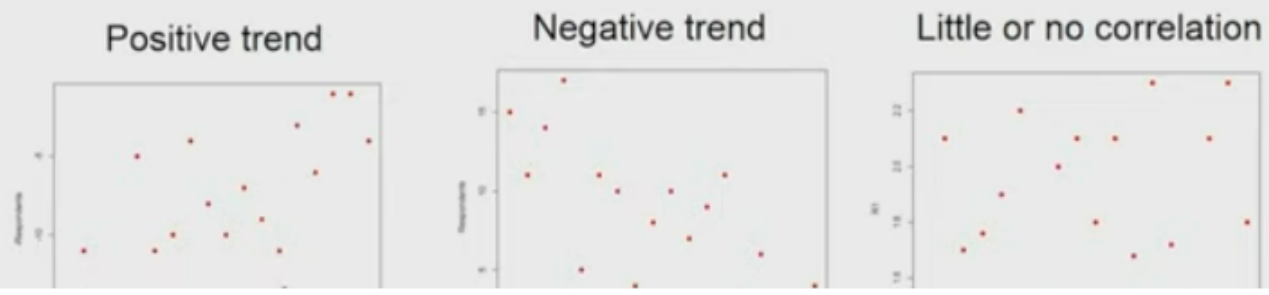
$$S_{xx} = \frac{1}{n}\sum(x_i - \bar{x})^2 \qquad S_{yy} = \frac{1}{n}\sum(y_i - \bar{y})^2$$

- Sample covariance S_{xy}

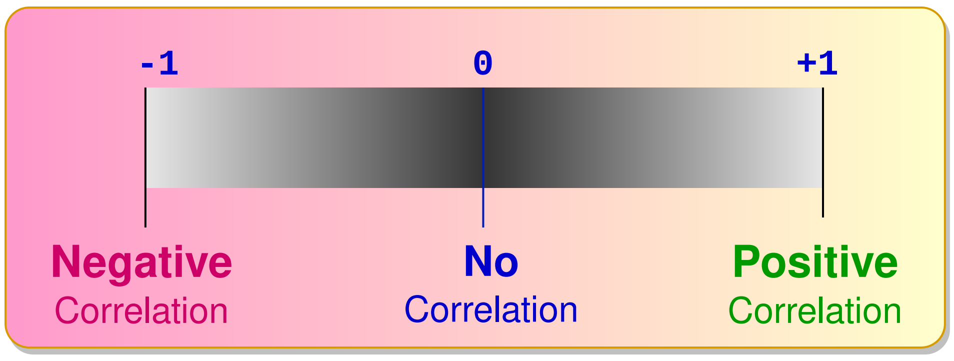
$$S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Correlation

- Correlation: the strength of association between two variables
- Correlation does not imply causation
- Visual representation of correlation: Scatter grams



Pearson's Correlation



Pearson's Correlation

- n observations for x and y variables (x_i, y_i)
- Pearson's product-moment correlation coefficient (r_{xy})

$$r_{xy} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - n\bar{x}^2)}\sqrt{(\Sigma y_i^2 - n\bar{y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

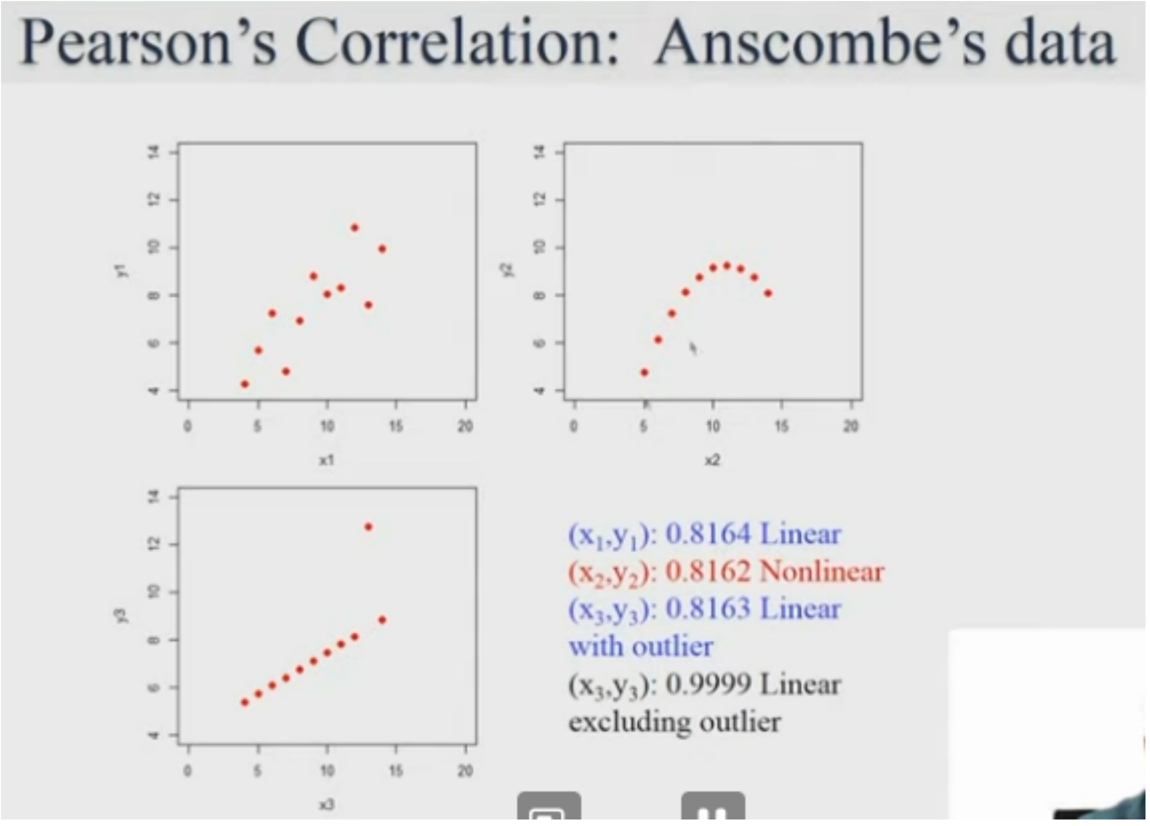
- r_{xy} takes a value between -1 (negative correlation) and 1 (positive correlation)
- $r_{xy} = 0$ means no correlation

Pearson's Correlation (Cont.)

- A measure for the degree of linear dependence between x and y

  Means.... the variables which hold the **Order**. Like indices of a data, rank..

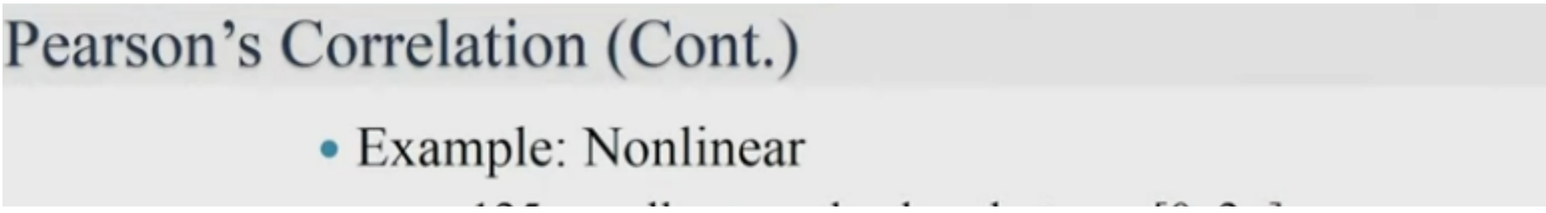
- Cannot be applied to ordinal variables
- Sample size: Moderate (20-30) for good estimate
- Robustness: Outliers can lead to misleading values



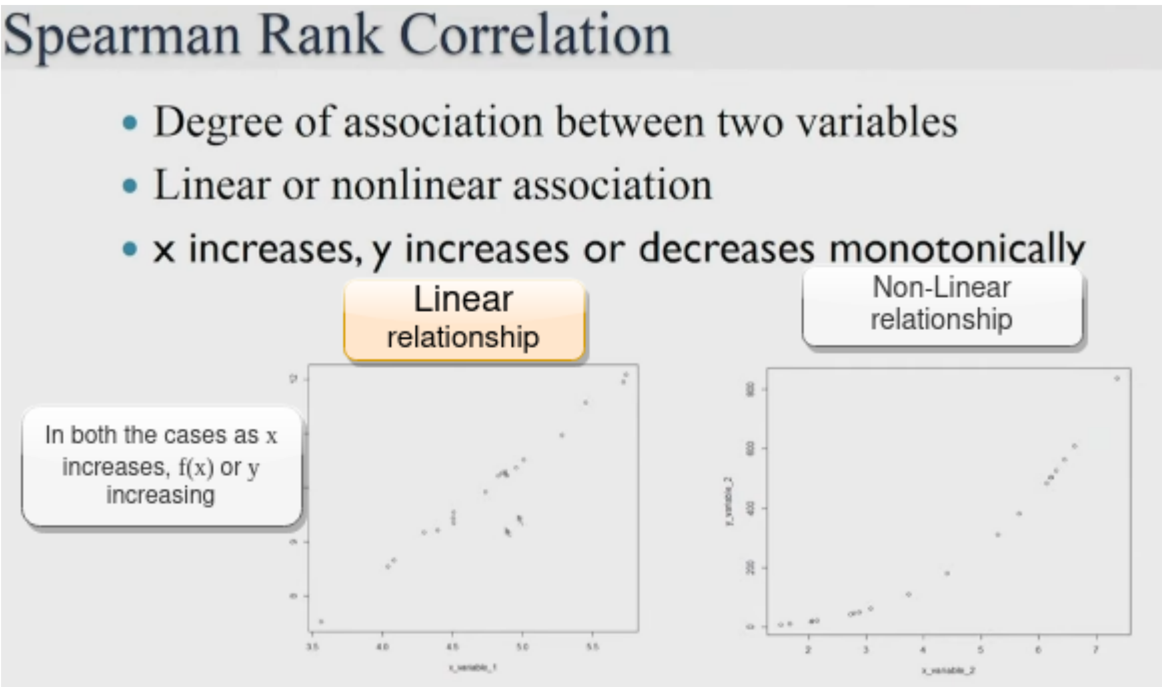
-- notice that, it gives almost same value of *linear* and *non-linear*.

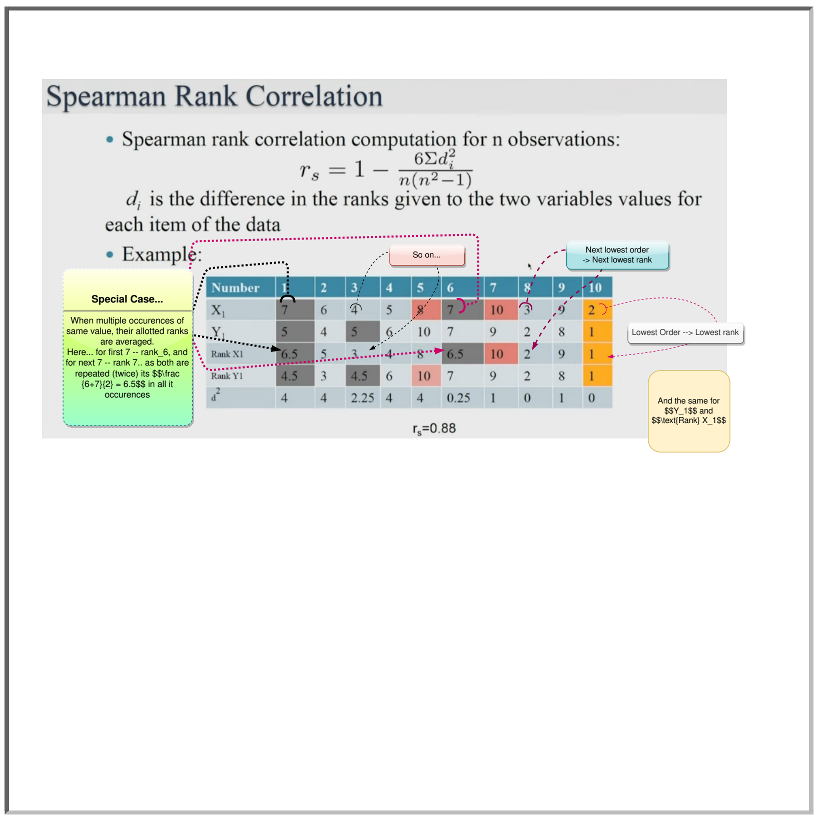
It is a dataset, which contains only 8-13 points.

Example:



Spearman Rank Correlation





Number	1	2	3	4	5	6	7	8	9	10
X_1	7	6	4	5	8	7	10	3	9	2
Y_1	5	4	5	6	10	7	9	2	8	1
Rank X1	6.5	5	3	4	8	6.5	10	2	9	1
Rank Y1	4.5	3	4.5	6	10	7	9	2	8	1
d^2	4	4	2.25	4	4	0.25	1	0	1	0

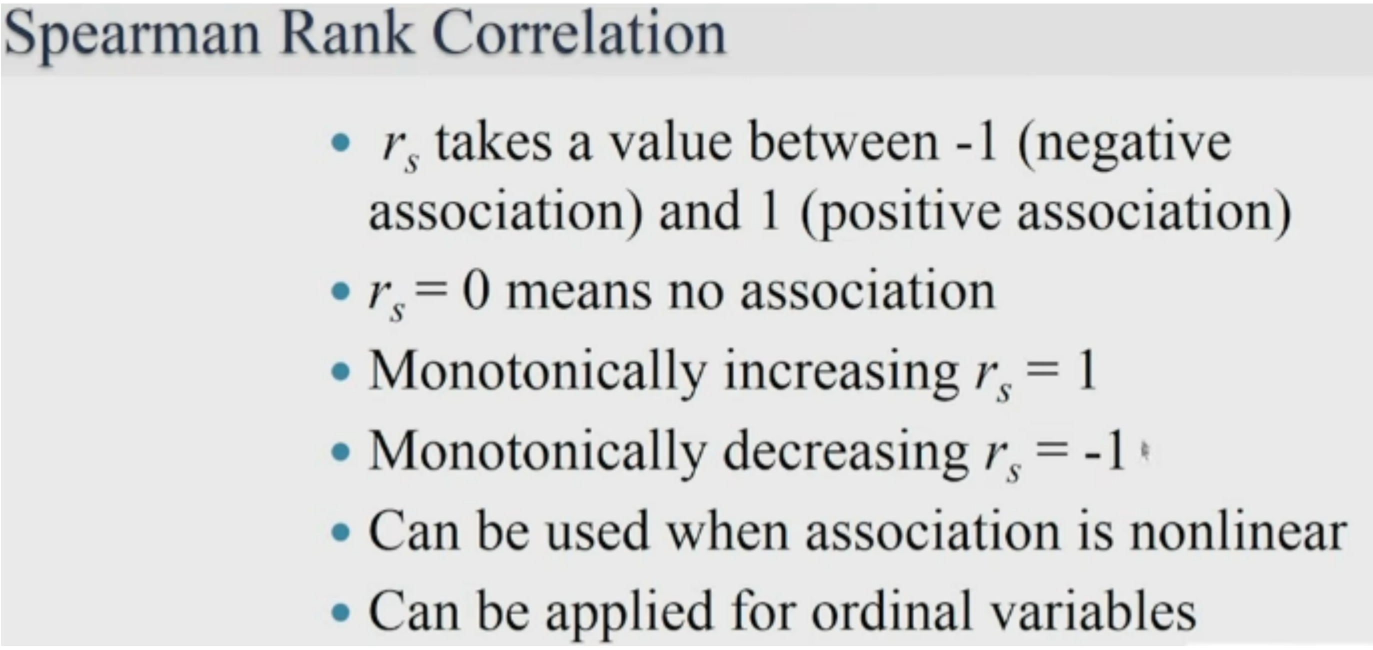
Check...

In [14]:
```
import numpy as np
lst = np.array([4, 4, 2.25, 4, 4, 0.25, 1, 0, 1, 0])
n=10
print(1-sum(6*lst)/(n*(n**2-1)))
```

```
0.8757575757575757
```

In [17]:
```
lst = c(4, 4, 2.25, 4, 4, 0.25, 1, 0, 1, 0)
n=10
1-sum(lst*6)/(n*(n**2-1))
```

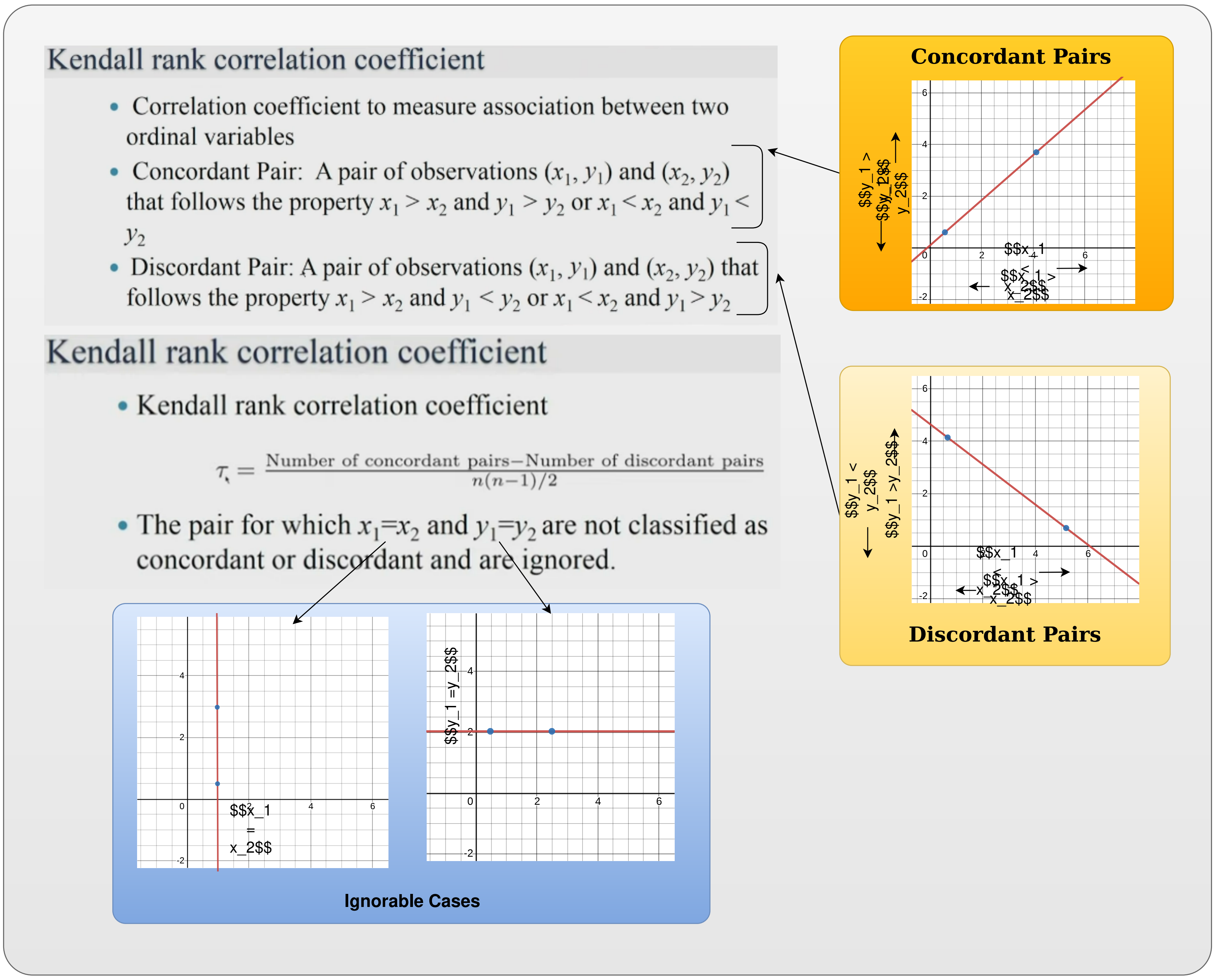
0.8757575757575757576



Kendall Rank Correlation

In [1]:
```
from IPython.display import IFrame
IFrame('resources/KendallRankCoeff.html', width=950, height=800)            ## The graphs in the image, are vi
```

Out[1]:



An example...

How the matrix was filled...?_(Looks like, for simple understanding.. consider either Expert1 column or Expert*2 col)*

>> -- here consider items as x_i and Expert_1 as y_1's ad Expert_2 as y_2 -- try making pairs in this way -- and compare like.. take two rows(item's instances) and compare the values *(of both experts)* only if both concordant -> Concordant, else discordant pairs.. .. its respectively filled in the matrix table beside.

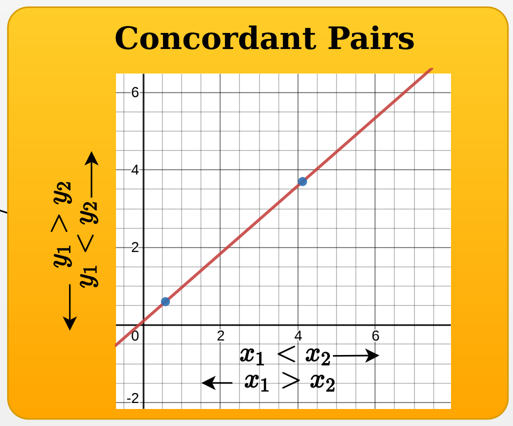
> **NOTE** that.. for 1,1 x pairs(i.e, rowS), there is no change .. so as per previous slide -- this is ignored, -- this can be seen as empty value.

- High +ve value indicates -- A strong agreement between associates.
- High -ve value indicates -- A strong disagreement..

Kendall rank Correlation: Anscombe's data

Kendall rank correlation coefficient

- Correlation coefficient to measure association between two ordinal variables
- Concordant Pair: A pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 > y_2$ or $x_1 < x_2$ and $y_1 < y_2$
- Discordant Pair: A pair of observations (x_1, y_1) and (x_2, y_2) that follows the property $x_1 > x_2$ and $y_1 < y_2$ or $x_1 < x_2$ and $y_1 > y_2$

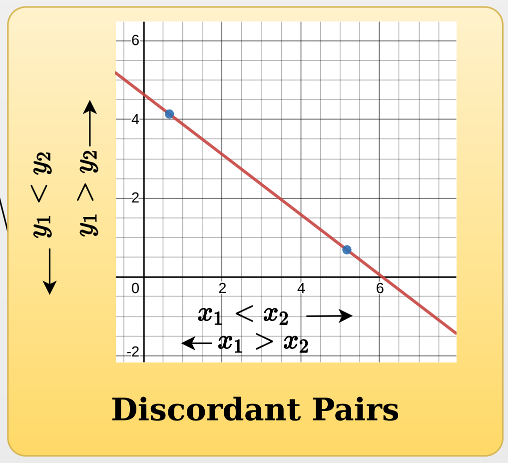


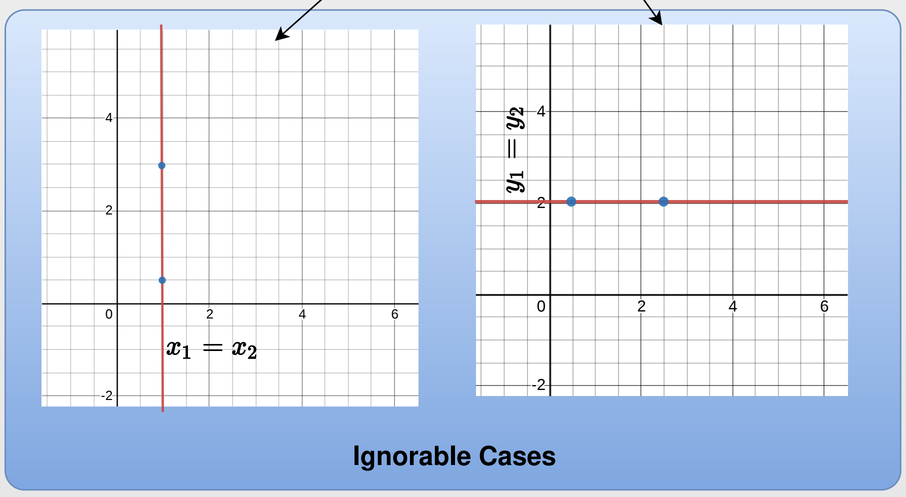
Kendall rank correlation coefficient

- Kendall rank correlation coefficient

$$\tau = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{n(n-1)/2}$$

- The pair for which $x_1 = x_2$ and $y_1 = y_2$ are not classified as concordant or discordant and are ignored.





summary..

Quantitative Measures
of Correlation

Coming up next.. Linear Regression

Module-2: **Linear Regression**

LINEAR REGRESSION

Let's start with the motivation..

Motivation

- Purpose is to build a functional relationship (model) between *dependent variable(s)* and *independent variable(s)*
- Example
  - Business : What is the effect of price on sales? (Can be used to fix the selling price of an item)
  - Engineering : Can we infer difficult to measure properties of a product from other easily measured variables? (mechanical strength of a polymer from temperature, viscosity or other process variables) – also known as a soft sensor

In Engineering... *(for the above examples)* -- finding their data via measurement instruments is difficult, but it can be inferred with some given parameters -- and its needed continously, to make inferences. In these cases, the usage of model is absolutely needed.

Brushing up basic terms...

Regression - Basics

- One of the widely used statistical techniques
- Dependent variables also known as *Response variable, Regressand, Predicted variable, output variable* - denoted as variable/s *y*
- Independent variable also known as *Predictor variable, Regressor, Exploratory variable, input variable* - denoted as variable/s *x*

Types of Regression:

Regression types

- Classification of Regression Analysis
  - Univariate vs Multivariate
    - *Univariate*: One dependent and one independent variable
    - *Multivariate*: Multiple independent and multiple dependent variables
  - Linear vs Nonlinear
    - *Linear*: Relationship is linear between dependent and independent variables
    - *Nonlinear:* Relationship is nonlinear between dependent and independent variables
  - *Simple vs Multiple*
    - Simple: One dependent and one independent variable (SISO)
    - Multiple: One dependent and many independent variables (MISO)

How can we go with choosing Regression technique?

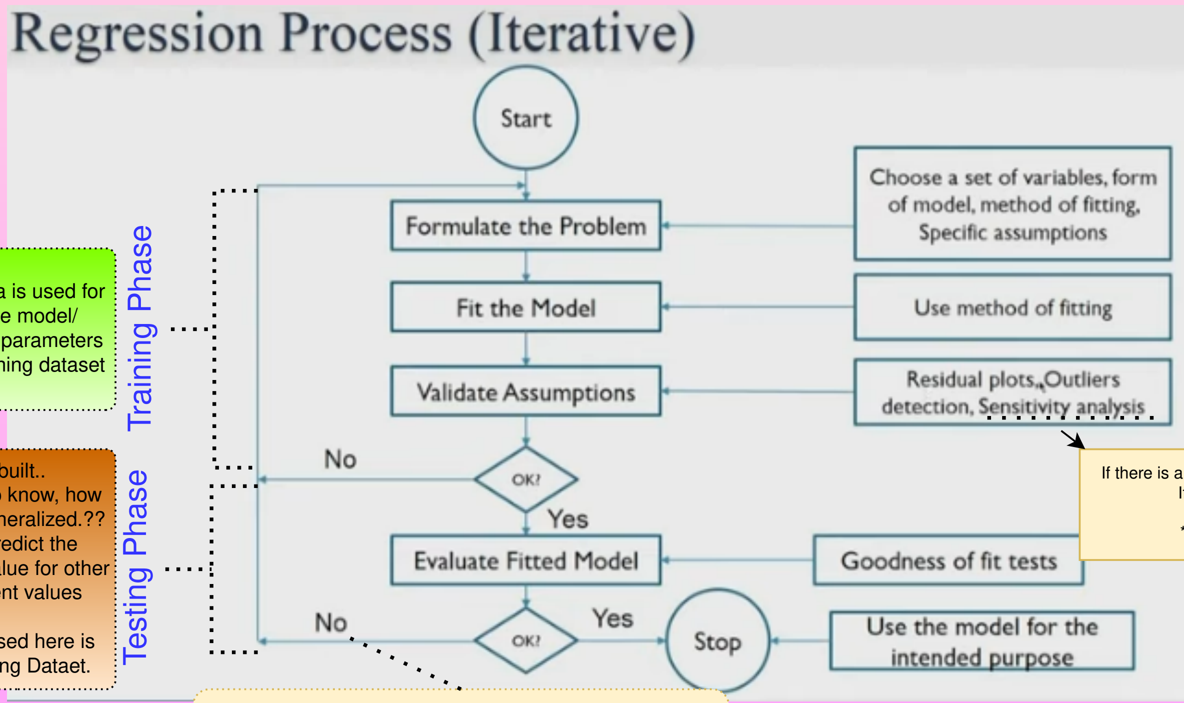
Regression analysis

- Is there a relationship between these variables?
- Is the relationship linear and how strong is the relationship?
- How accurately can we estimate the relationship?
- How good is the model for prediction purposes?

-- needed at choosing the model and even at developing the model..<br/? Various techniques to go with regression..

Regression methods

- Linear regression methods
  - Simple linear regression
  - Multiple linear regression
  - Ridge regression
  - Principal component regression
  - Lasso
  - Partial least squares



Training Phase

Here the data is used for training the model/ estimate the parameters -- called Training dataset

Testing Phase

Once built.. Would like to know, how well it got generalized.?? * Can it predict the dependent value for other independent values

-- dataset used here is called Testing Dataet.

If there is an y error in the data, how much it affects the result.
If there are many variables, can ask like..
* Are all variables equal..??
* Can one variable be dropped/discard it

it not worked.. then choose other type of regression technique.

Even after trying all techniques, if still not... then need to check the variables and the type of experiment conducted

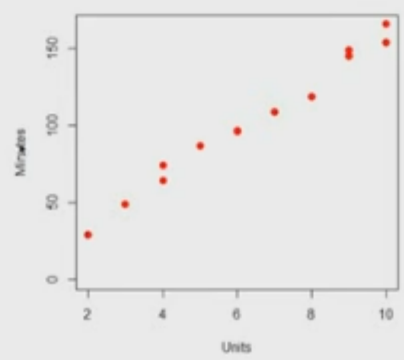
If all models fail..

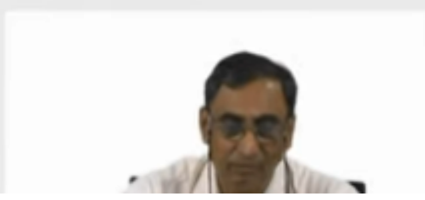
Check the experimental technique

Let's see some of the techniques..

Ordinary Least Squares (OLS)

- Fourteen observations obtained on time taken in minutes for service calls and number of units repaired
- Objective is to find relationship between these variables (useful for judging service agent performance)





-- In this (say) a service agent reports to his boss -- how many he repaired in the given time. Such way here 14 points are taken. -- This can be used for the purposes like,

- Measuring the efficiency of the service-man or decision on increasing his salary.. or
- Productivity level of the comany.. like some of those..

Why error..(ϵ_i)??

- May be due to the model we've chosen is inadequate for this data.
- May be some errors in the measurement of x and y's..'

Here we assume that, x has no error -- it has perfectly measured. y can contain.
This gives the decision of choosing the independent and dependent variable..

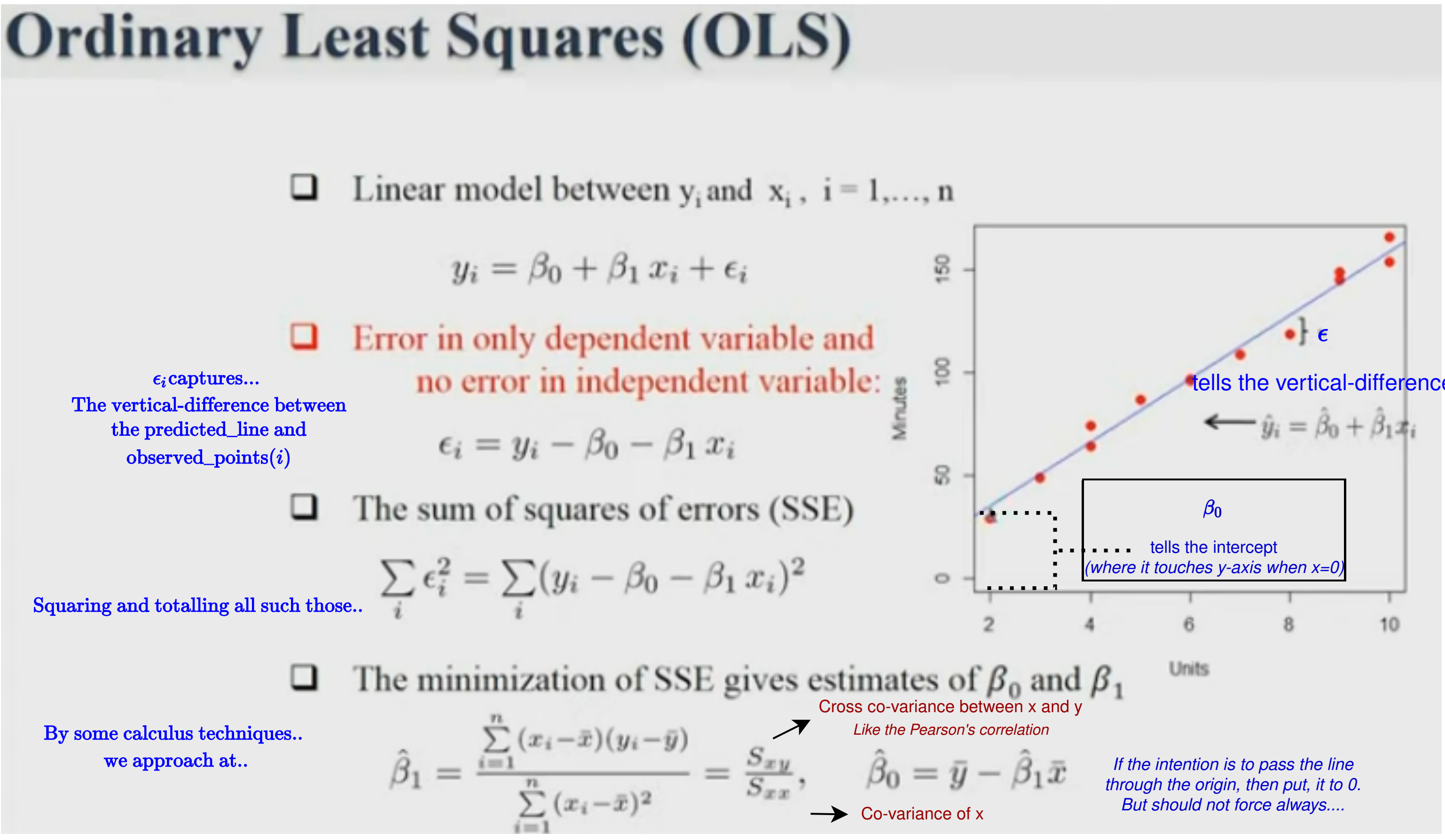
> The one which had no errors --> Independent variable
> Which had some errors -> Dependent variable

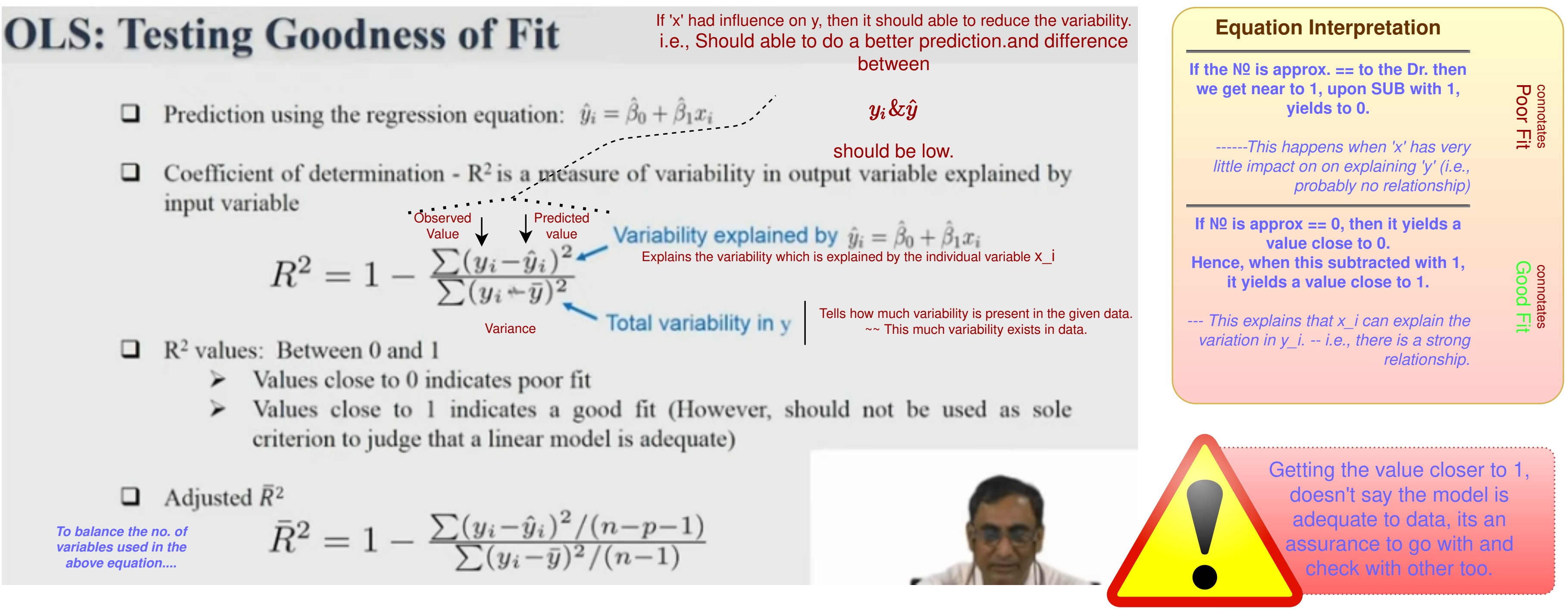
Here in the given example...

> **Units** is chosen as *Independent variable* -- with an idea that, he bills all the customers and the same copy-slips are returned in the office, so this can't be an error unless someone transcribes it.
> **Minutes** as *Dependent variable* as his measurement may include the timings like traffic, location(if far, takes much time than the service-time).. and many such.

But, its also a arguable point that, some even tells to use **The one to be predicted as *Dependent variable* and with which we can do that as _*Independent variable***, but, after building the model with the obtained equation, one can go in either way right..

If both contain error, then we need to go for some other methods like *Sensitive Linear regression*..





OLS: Example using R

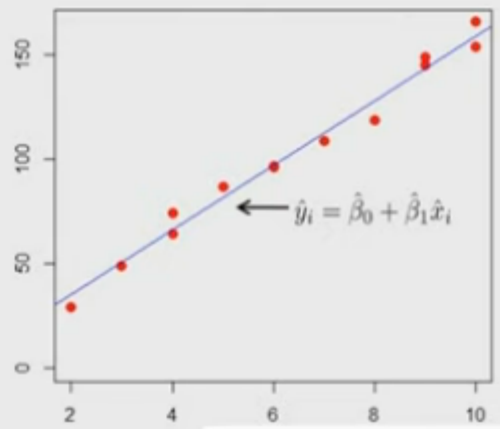
```
Call:
lm(formula = Minutes ~ Units)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2318 -3.3415 -0.7143  4.7769  7.8033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.162      3.355    1.24    0.239
Units         15.509      0.505   30.71 8.92e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.392 on 12 degrees of freedom
Multiple R-squared:  0.9874,	Adjusted R-squared:  0.9864
F-statistic: 943.2 on 1 and 12 DF,  p-value: 8.916e-13
```

$\hat\beta_0$ $\hat\beta_1$



Module-3: **Model Assessment**

Now, we are done with the (basic version) of model building. Now lets assess it.. and there are various ways to achieve that..
So, what are the questions that one can take..

OLS Model Assessment and Improvement

These are focused in this lecture..

- ❑ How good is a linear model?
- ❑ Which coefficients of the linear model are significant (Identify important variables)
- ❑ Can we improve quality of linear model?
  - ❑Are assumptions made about errors reasonable?
    - ❑Normality: Errors are normality distributed
    - ❑Homoscedasticity: Errors in different samples have same variance

Mean Variance

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \; i = 1, 2, \ldots, n$$

  - ❑Are there bad measurements in the data (outliers)

If not, try fitting with other model

*Basically, this is mostly important in **Multi-Variable Regression.** --where we have multiple independent variables, and need to decide, whether all are important..?? Can we drop any..*

Before starting, need to estimate some of the parameters..

OLS: Properties of Estimates

Hat symbols denote that: **These are estimates**

i.e., $\hat\beta_0$ and $\hat\beta_1$ are the estimates of true intercept and slope

Intercept Slope

- ❑ Both $\hat\beta_0$ and $\hat\beta_1$ estimates are unbiased

" E " means **Estimation** or **Mean**

$$E[\hat\beta_0] = \beta_0, \quad E[\hat\beta_1] = \beta_1$$

These stmts tell that...

If could repeat the experiment for some n times, and average(mean) all those, then it reaches the true values.
*-- NOTE: These values are unknown to us intially, and the obtained ones are called **UnBiased estimates.***

- ❑ Variance of the estimates

Variance of independent variable

$$\text{var}[\hat\beta_1] = \frac{\sigma^2}{S_{xx}}, \quad \text{var}[\hat\beta_0] = \sigma^2 \frac{\sum x_i^2}{n\, S_{xx}}$$

We get different different estimates, depending on sample
So, what's the strength of these..
We can show from the assumptions that ..
(as present as expression..)

- ❑ Estimate of σ^2

Variance of independent variable (x) i.e., $\sum (\bar{x} - x_i)^2 / n$

This depends on the instrument measured.
Like we measured for errors and accuracy in for Vernier calipers, Screw Gauge tools usage in Physics Lab (Masabtank-clg)

$$\hat\sigma^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

Mostly, this won't be given.. But, it can be estimated as...

Already two of the estimates are used (beta_0-cap and beta_1-capt) for beta_0 and beta_1.
Therefore only remaining n-2 samples are available for estimating it.
-If had only 2 samples, then surely result is 0.

Why n"-2"..??

- ❑ Distribution of slope estimate $\hat\beta_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$

SSE: Sum Squared Error

Normal Distribution

.

Once we derived the distribution of paramters, we can go for hypothesis testing to decide, whether are these significantly equal to 0.
And can also derive the confidence intervals.

OLS: Confidence Intervals on regression coefficients

Recollect that.. an interval contains the **lower** and **upper boundaries**

And also the concepts and worked-out problems in COSM subject(JNTUH)

❑ 95% two-sided confidence intervals (CI) for $\hat{\beta}_0$ and $\hat{\beta}_1$

Lower critical value

Upper critical value

$t_{0.025,12}$

Degrees of Freedom(DoF). As two were already used, left-over are 14-2=12 *(Ansombe's DatSet)*

$$\beta_1 \in [\hat{\beta}_1 - 2.18\, s_{\hat{\beta}_1},\ \hat{\beta}_1 + 2.18\, s_{\hat{\beta}_1}], \quad s_{\hat{\beta}_1} = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{(n-2)S_{xx}}}$$

Standard Deviation of $\hat{\beta}_1$

Here its for 2.5% critical value.

Once after constructing 95% interval, we can test whether unknown beta^_0 and unknown beta^_1 are equal to 0 or not..
Let's do that.. But.. **Why needed this Hypothesis test..??**

> We've fitted the linear model, assuming that,
>
> 1. linear dependence between x and y and obtained an estimate of beta^_1, and also...
> 2. fitted the intercept term.
>
> **Now, we may want to ask,**
>
> 1. ***Is the intercept term*** (β_0) ***Significant:*** `*` **- May be line should pass through origin or not**
> 2. ***Slope term*** (β_1)***:*** **May be variable** `y` **does not depend on** `x` **-- i.e., not depending in significant manner(i.e.,** β_1 **is close to 0)**
>
> `*` Significant
>
> ---
>
> Considering the equation's view.... evaluating with `0` *(for intercept term, as its the const)* is as same as not writing the intercept right..?? -- so, does it had significance..?? if `0` not, and `!=0`, yes...
> **For graphical view..**
>
> 1. **For intercept**: Does the line doesn't pass through origin or not.
> 2.

Now, by NULL Hypothesis, we are testing for $\beta_1 = 0$ vs $\beta_1 \neq 0$.

> If $\beta_1 = 0$ (H_0 is accepted) true, indicates that...**Independent variable (** `x` **) has no effect on dependent variable(** `y` **)**.
> If rejected NULL Hypothesis *(means accepting Alternative Hypothesis)*.... concludes that.. **Independent variable (** `x` **) has some effect on dependent variable(** `y` **)**.