SUMMARY OF COURSE CONTENT: Linear Regression

Before, we start discussing the questions, let me summarise the main take away from the session on linear regression.

We talked about the CRISP DM Framework:

- Business understanding
- Data understanding
- Data preparation
- Model building
- Model evaluation
- Model deployment

EDA (Exploratory data analysis)

- Cleaning data - De-duplication of rows
- Missing value imputation
- Outlier treatment
- Converting variables to appropriate types
- Creating dummy variables
- Data visualisation

Use cleaned and prepared data to build models.

Then we talked about linear regression:

If you recall, linear regression is

- ➢ The simplest and one of the most widely used models.

- ➢ Used to establish linear relationship between continuous variables (predictor variable or independent variable and response variable or dependent variable).

IN OTHER WORDS

It attempts to explain the relationship between an independent variable and a dependent variable using a straight line.

- ➢ They cover a broad business spectrum and are easy to understand and explain as you might have already experienced.

Then, you came across:

- Equation of regression line
- Residuals – RSS (Residual sum of squares), TSS (Total sum of squares) and R^2 which is coefficient of determination = 1 – RSS/TSS, correlation coefficient (cc or *r*).
- The R^2 value (0 to 1) is equal to the square of correlation coefficient (-1 to +1) between the two variables.

R^2

- ✓ If 1, all the observations fit in a straight line.
- ✓ If 0, x and y are completely independent.
- ✓ R^2 value gives the percent variation in y that can be explained by x.

CC (correlation coefficient)

- ✓ Measures the strength of the linear association between two variables.
- ✓ Not effected by linear transformations of y or x.
- ✓ Does not distinguish between dependent and independent variables.

Overall methodology:

1. You start with a scatter plot to check the relationship between the dependent and independent variables.
2. You compute residuals i.e. RSS for any given line passing through the scatter plot.
3. You then find equation of the best fit line by minimising the RSS (using ordinary least squares method which is done using differentiation and gradient descent method) and then estimating optimal values of β_0 and β_1.

This was followed by a session on multiple linear regression.

- ✓ It explains the relationship between two or more independent variables and a response variable by fitting a straight line.
- ✓ You saw an advertisement data set example with sales prediction (response variable) and marketing (TV, Radio, Newspaper) as independent variables.
- ✓ You learned to build a model which could be used to predict sales given the TV and Radio marketing budget spend in a market.
- ✓ It was evident that not all variables are useful to build a model as some independent variables are insignificant and do not contribute to the response variable. In case of the sales prediction problem, the variable Newspaper marketing was insignificant.
- ✓ The concept of Dummy variable was introduced that could deal with some categorical attributes such as yes / no to numeric form to be used in regression modelling. For more than two categories, model.matrix function in R is used.
- ✓ Concept of derived matrix was introduced.
- ✓ Also, in multiple regression adjusted R^2 is a better matrix than R^2 to assess how good the model fits the data.
- ✓ Concept of multi-collinearity was introduced – there could be correlation between independent variables that could make one or more variables insignificant in multi-variable model. It can be detected using VIF.
- ✓ VIF (variance inflation factor) = $1/(1-R^2)$ for single independent variable is obtained by the R^2 value of the regression of that variable with the other variables.

Important:

Higher the VIF, higher the multicollinearity.

Variable selection method:

1. Backward – as you already saw in the example in video.
2. Forward – starts with single variable, add variable one by one, check *p*-value and adjusted R^2. Keep variables that increase the adjusted R^2.
3. Stepwise selection – This is useful when data set contains large number of variables. Here, combination of backward and forward selection is used.
    a. Start with all variables.
    b. Drop the least significant variable.
    c. Reconsider previously dropped variables for reinsertion.
    d. Variable being reinserted should have lower *p*-value than the variables being dropped.
4. Use step AIC for reduced set of variables in the model with high likelihood of observing certain observations **followed by** VIF and *p*-value.

- ✓ Finally you arrive at a model where all variables are significant and there is no multicollinearity.
- ✓ The accuracy of the model is checked.

Regression has five key assumptions:

- Linear relationship – there is a linear relationship between the outcome variable and the independent variables
- Multivariate normality - residuals are normally distributed
- No or little multicollinearity - there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.
- No auto-correlation - Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).
- Homoscedasticity - meaning the residuals are equal across the regression line.

To summarise:

1. Prepare the data for analysis
2. Build a model containing all variables.
3. Reduce variables using step AIC.
4. Check for multi-collinearity.
5. Use *p*-value to eliminate variables.

Let us recall all the steps we used throughout the linear regression model building process:

1. Once you understood the business objective, you prepared the data, followed by EDA and the division of data into training and test data sets.

2. The next step was the selection of variables for the creation of the model. Variable selection is critical because you cannot just include all the variables in the model; otherwise, you run the risk of including insignificant variables too.

3. This is where forward selection, backward selection, and stepwise selection come into the picture. But in linear regression, we focused on the backward selection method. You used step AIC to quickly shortlist some variables which are significant to save time.

4. However, these significant independent variables might be related to each other. This is where you need to check for multicollinearity amongst variables using variance inflation factor (VIF) and remove variables with high VIF and low significance ($p>0.05$).

5. The variables with a high VIF or multicollinearity may be statistically significant (***) or $p \leq 0.001$, in which case you will first have to check for other insignificant variables ($p>0.05$) before removing the variables with a higher VIF and lower *p*-values.

6. Continue removing the variables until all variables are significant (***) or $p \leq 0.001$, and have low VIFs.

7. Finally you arrive at a model where all variables are significant and there is no threat of multicollinearity.

8. The final step is to check the model accuracy on the testing data.

ANSWERS TO QUESTIONS BY STUDENTS - Linear Regression

DS C10 - Linear Regression Live Session - Questionnaire

1> In the housing data analysis, lr_summary() gives the std err. How it is calculated? can you give working of one independent parameter e.g.: area? std err = 0.038. How this is calculated?

Answer: This is well explained in the "Best Fit Line" in the "Simple Linear Regression" section of the video lecture. For answers to specific numerical std err values mentioned in the question, I request TA / student coordinator to answer.

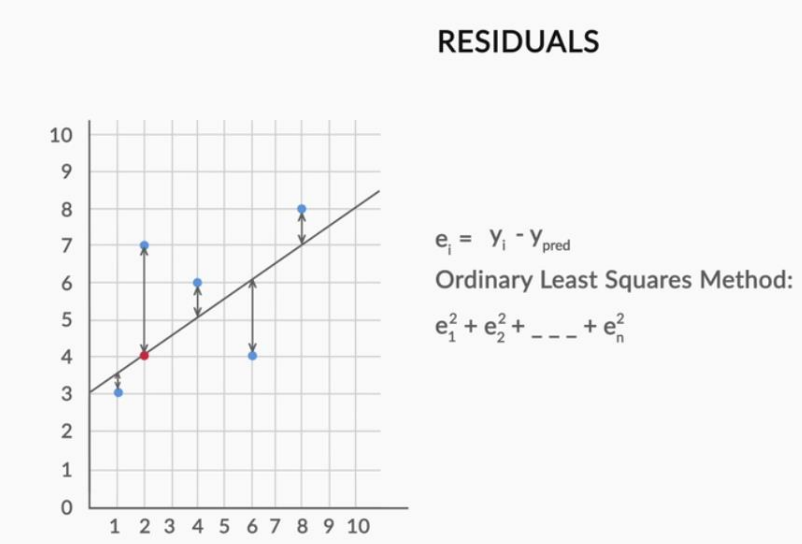
- ✓ Once you have the data and you have generated the scatter plot, we have to fit a straight line to this set of data points, we would like to know what is the best fit straight line that makes our model for future prediction.

- ✓ There can be multiple straight lines that can fit the data points. So we have to look for a line that has minimum deviation from the points in the scatter plot.

- ✓ To find the best fit line, there is a notion called "residual" or "error" that is associated with every data point. The difference between the measured and the predicted value is called the "residual / error".

- ✓ Similarly for every data point we have a residual value that is also an error (denoted as e_i for *ith* data point). To find the best fit line, we use the ordinary least square (OLS), which minimises the total error square.

- ✓ For doing this, we square individual error term and sum them over all data points where *i* is running from 1 to n.



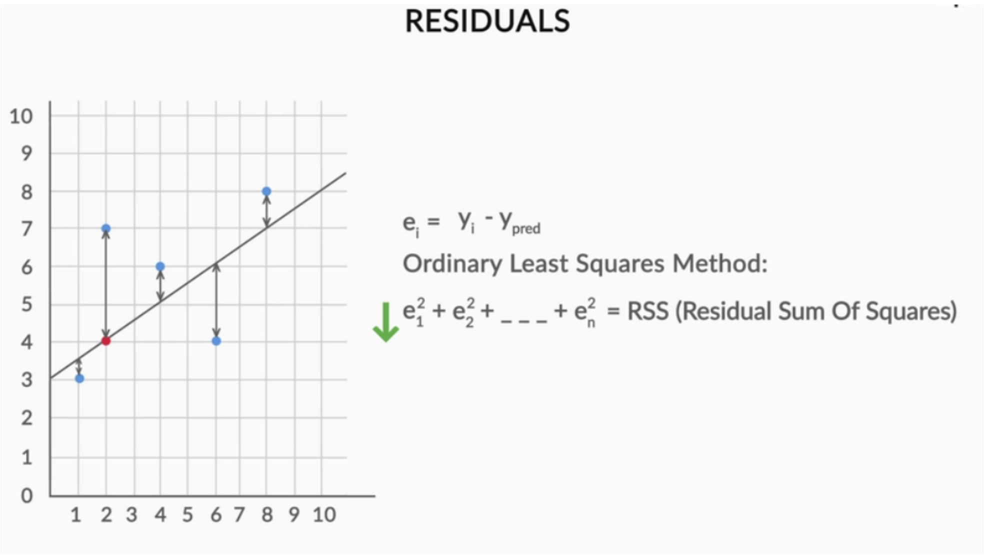
RESIDUALS

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$

- ✓ And we would like to pick up that straight line (i.e. a particular β_0 and β_1) so that “this sum of squares (also called Residual sum of squares or RSS) is minimized.”



RESIDUALS

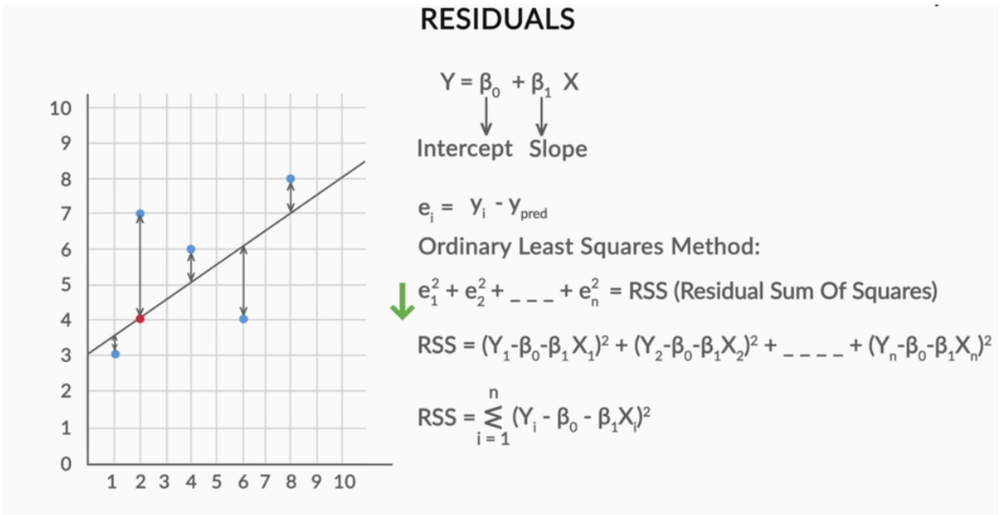
$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$\downarrow e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

So, what you did is

- ✓ You found residuals and RSS for any given line passing through the scatter plot.
- ✓ Then you found the equation of the best-fit line by minimising the RSS and found the optimal values of β_0 and β_1.



RESIDUALS

$Y = \beta_0 + \beta_1 X$

$\overset{\downarrow}{\text{Intercept}}$ $\overset{\downarrow}{\text{Slope}}$

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

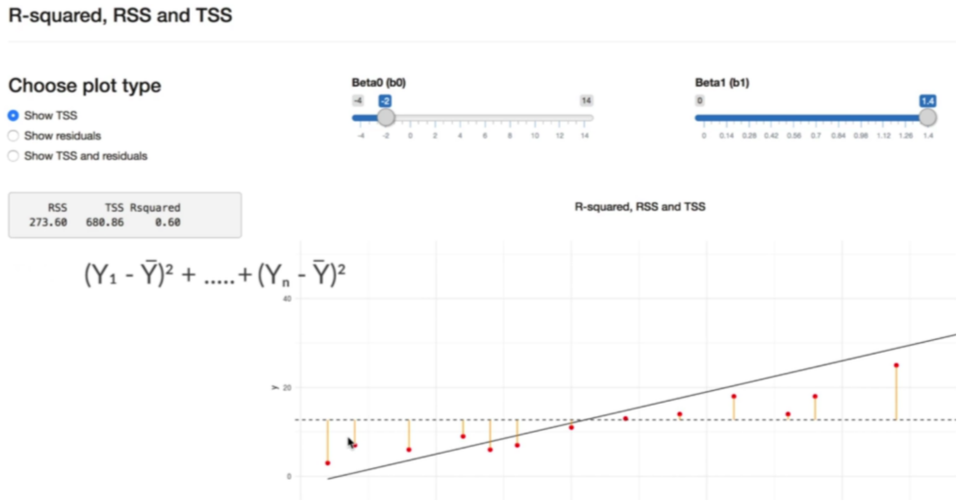
$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

Now,

- ✓ TSS (Total sum of squares) is the sum of the square of the difference between each point and Y average.

- ✓ If you have a linear model, where you do not use an independent variable, and you just use intercept, we can build a simple model, (where β₀ intercept value is Y average). And any other model that we built with an independent variable should be better than this model which is the model with β₀ = Y average.



R-squared, RSS and TSS

Choose plot type

● Show TSS
○ Show residuals
○ Show TSS and residuals

Beta0 (b0): -2

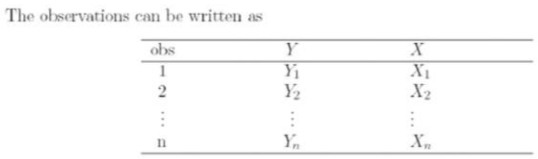
Beta1 (b1): 1.4

```
   RSS      TSS Rsquared
273.60   680.86     0.60
```

$(Y_1 - \bar{Y})^2 + \ldots\ldots + (Y_n - \bar{Y})^2$

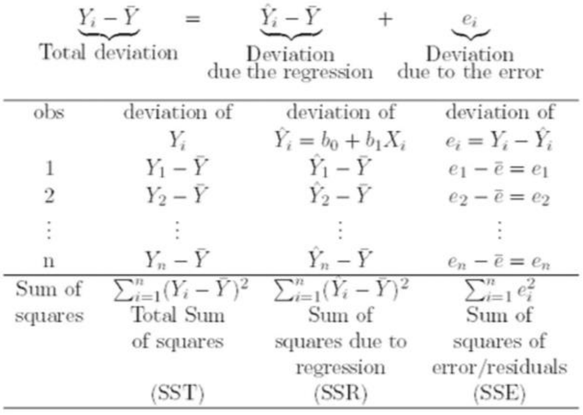
- ✓ So, RSS is an absolute quantity and TSS is a relative quantity.

- ✓ RSS / TSS is a normalised quantity which will tell us how good the model is. R^2 is 1 – RSS/TSS and this tells is how good the model is?



The observations can be written as

obs	Y	X
1	Y_1	X_1
2	Y_2	X_2
$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_n



$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation}\\ \text{due the regression}}} + \underbrace{e_i}_{\substack{\text{Deviation}\\ \text{due to the error}}}$$

obs	deviation of Y_i	deviation of $\hat{Y}_i = b_0 + b_1 X_i$	deviation of $e_i = Y_i - \hat{Y}_i$
1	$Y_1 - \bar{Y}$	$\hat{Y}_1 - \bar{Y}$	$e_1 - \bar{e} = e_1$
2	$Y_2 - \bar{Y}$	$\hat{Y}_2 - \bar{Y}$	$e_2 - \bar{e} = e_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$Y_n - \bar{Y}$	$\hat{Y}_n - \bar{Y}$	$e_n - \bar{e} = e_n$
Sum of squares	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ Total Sum of squares (SST)	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ Sum of squares due to regression (SSR)	$\sum_{i=1}^{n} e_i^2$ Sum of squares of error/residuals (SSE)

2> Can you introduce us to non-linear regression as well. At least provide study material or sample working.

Answer: Non-linear regression is beyond the scope of the current topic of discussion. You may look for some relevant reading materials online.

3> Why we normalise data before finding cost function or best fit line

Answer: Normalizing / Scaling is done to bring all the different data ranges to a common scale such as 0 to +1. Standard normalization scaling is used widely as it normalises all the variable's range between 0 to 1, thereby making it easy to understand and visualise, and then use them in model building.

4> What factors lead to choosing the best independent variables for calculating the dependant variables

Answer:

- Variable selection is critical in linear regression because including all the variables in the model may introduce the risk of including insignificant variables too.

- However, these significant independent variables might be related to each other. This is where you need to check for multicollinearity amongst variables using variance inflation factor (VIF) and remove variables with high VIF and low significance (p>0.05).

- First look at the VIF (variance inflation factor) and then consider the *p*-value of the variables.

➢ A variable with a high VIF means it can be largely explained by other independent variables. **Thus, you have to check and remove variables with a high VIF and *p*-value, implying that their impact on the outcome can largely be explained by other variables.** Thus, removing the variable with a high VIF would make it easier to assess the impact of other variables, while making little difference to the predicted outcome.

➢ **The higher the VIF, the higher the multicollinearity.** But remember — variables with a high VIF or multicollinearity **may be statistically significant** (***) or p<0.05, in which case you will first have to check for other insignificant variables before removing the variables with a higher VIF and lower *p*-values.

5> How to interpret AIC AND BIC. Please explain me likelihood in case of regression and in general.

Answer: This topic has been discussed in the page titled "Multivariate Logistic Regression in R (Variable Selection)." It selects only the useful variables from the full model and discards the rest using a **stepwise AIC** algorithm.

- Feature selection can be conducted for logistic regression using Stepwise AIC algorithm and VIF and p-value check.
- The best model can be obtained through iterations where you have to eliminate the irrelevant variables (using stepwise AIC algorithm, VIF and *p*-value) and then look the β_0 and β_1 coefficients.
- StepAIC reduces the set of variables in the model with high likelihood of observing certain observations **followed by** VIF and p-value.
- StepAIC quickly shortlists some variables which are significant to save time.
- It may be difficult to quantify the change in AIC when a single variable is removed from the model.

Bayesian Information Criterion (BIC) is also used for choosing best predictor subsets in regression and often used for comparing and selecting model from a finite set of models, where the model with the lowest BIC is preferred. It is based partly on the likelihood function.

When several groups are being compared, the best model is generally the one that minimizes both AIC and BIC. However, AIC is susceptible to overfitting the data, whereas BIC is susceptible to underfitting the data. The reason is that they penalize the free parameters differently, i.e., 2*p in AIC, ln(N)*p in BIC, where p = number of estimated parameters, N = sample size.

AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. On the other hand, BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations.

Please refer to this following book chapter for more details:

Arijit Chakrabarti, Jayanta K. Ghosh, 2011, "*AIC, BIC and Recent Advances in Model Selection*", Editor(s): Prasanta S. Bandyopadhyay, Malcolm R. Forster, In Handbook of the Philosophy of Science, Philosophy of Statistics, North-Holland, Volume 7, Pages 583-605,

ISSN 18789846, ISBN 9780444518620,

https://doi.org/10.1016/B978-0-444-51862-0.50018-6.

(http://www.sciencedirect.com/science/article/pii/B9780444518620500186)

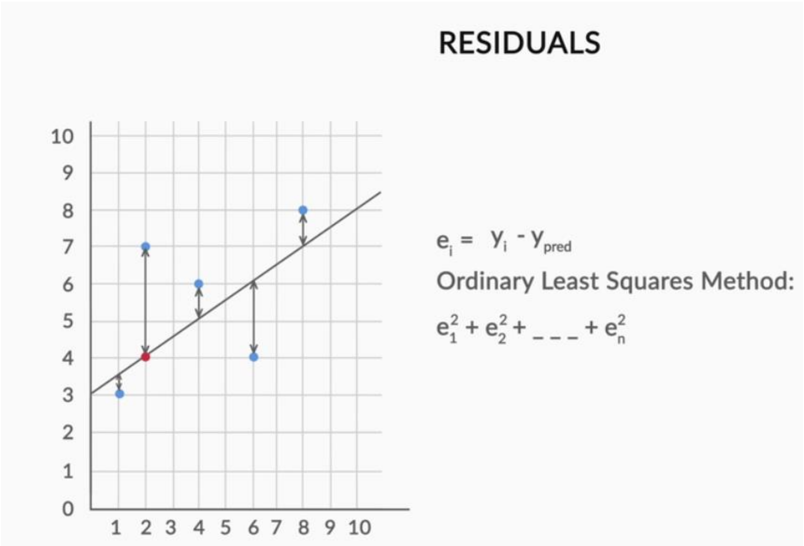
6> Why only p value is the deciding factor whether the model can be well suitable enough and what does it represent?

Answer: The ***p*-value** is the level of marginal significance within a statistical hypothesis test to provide the smallest level of significance at which the null hypothesis would be rejected.

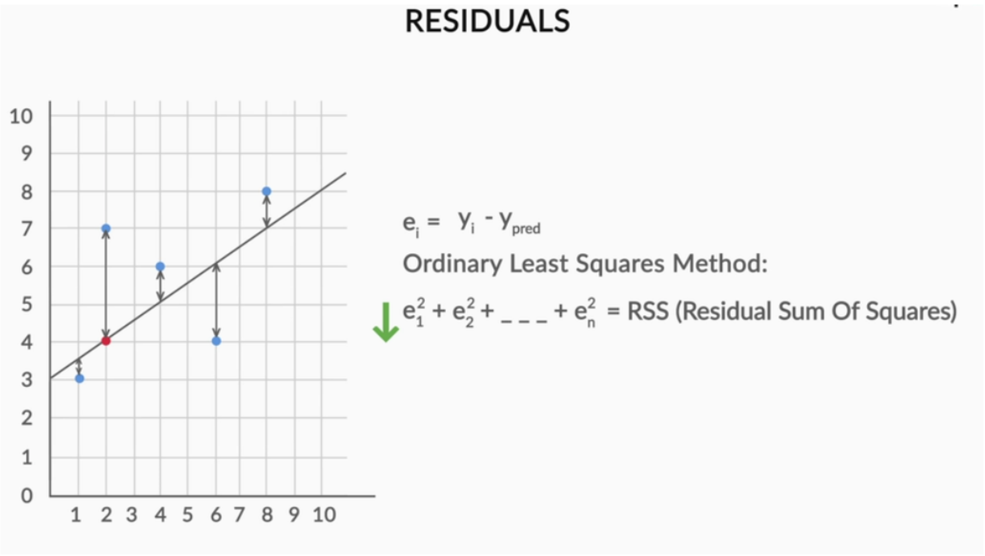
7> Please explain RSS, TSS, RSE and R squared.

Answer:

- ✓ To find the best fit line, there is a notion called "residual" that is associated with every data point. The difference between the measured and the predicted value is called the "residual".
- ✓ Similarly for every data point we have a residual value that is also an error (denoted as e_i for *ith* data point). To find the best fit line, we use the ordinary least square (OLS), which minimises the total error square.
- ✓ For doing this, we square individual error term and sum them over all data points where *i* is running from 1 to n.

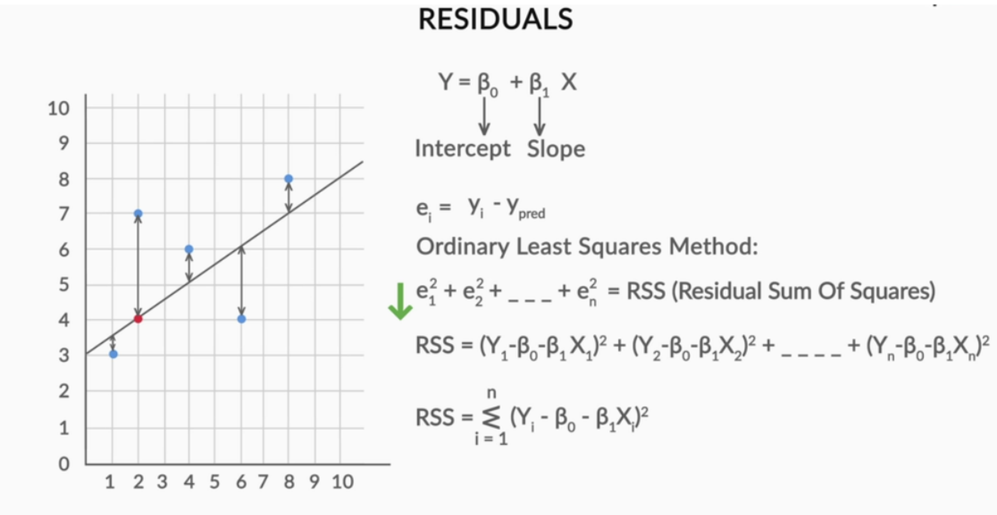


- ✓ And we would like to pick up that straight line (i.e. a particular β_0 and β_1) so that “this sum of squares (also called Residual sum of squares or RSS) is minimized.”



So, what you did is

- ✓ You found residuals and RSS for any given line passing through the scatter plot.
- ✓ Then you found the equation of the best-fit line by minimising the RSS and found the optimal values of β_0 and β_1.



RESIDUALS

$$Y = \underset{\substack{\downarrow \\ \text{Intercept}}}{\beta_0} + \underset{\substack{\downarrow \\ \text{Slope}}}{\beta_1} X$$

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$\downarrow\ e_1^2 + e_2^2 + \_\,\_\,\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_2-\beta_0-\beta_1X_2)^2 + \_\,\_\,\_\,\_ + (Y_n-\beta_0-\beta_1X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2$$

Typically, the best fit line will help you to answer the following questions:

1. How well does the best-fit line represent the scatter-plot?
2. How well does the best-fit line predict the new data?

Now,

- ✓ TSS (Total sum of squares) is the sum of the square of the difference between each point and Y average.
- ✓ So, RSS is an absolute quantity and TSS is a relative quantity.
- ✓ RSS / TSS is a normalised quantity which will tell us how good the model is.
- ✓ R^2 is 1 – RSS/TSS and this tells is how good the model is?
- ✓ If R^2 is 0.6 it means that we are able to explain 60% of the variance in the data.
- ✓ RSE – Residual standard error is the **positive square root of the mean square error.**

8> What is gradient descent model in Linear regression and its application

Answer: Let's begin with a simple example: Suppose you are at the top of a mountain, and you have to reach a lake which is at the lowest point of the mountain (i.e. valley). Suppose you are blindfolded and you have no visibility to see where you are going, so, what approach will you take to reach the lake?

- The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

- Gradient descent (also known as steepest descent) is an iterative optimization algorithm for finding the minimum of a function. To find a local minimum, steps proportional to the negative of the gradient is considered at the current point.

- In mathematical terms: Suppose we want to find out the best parameters (θ_1) and (θ_2) for our learning algorithm. Similar to the analogy above (of the lake), we find similar mountains and valleys when we plot our "cost space". Cost space is the performance of the algorithm when we choose a particular value for a parameter.

- Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the *negative* of the gradient (or approximate gradient) of the function at the current point. If, instead, one takes steps proportional to the *positive* of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent. The gradient (or derivative) tells us the incline or slope of the cost function. Hence, to minimize the cost function, we move in the direction opposite to the gradient.

- There are many types of gradient descent algorithms:

- **On the basis of data assimilation**
  1. Full Batch Gradient Descent Algorithm
  2. Stochastic Gradient Descent Algorithm

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

- **On the basis of differentiation**
  1. First order Differentiation
  2. Second order Differentiation

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

Please see the following Web URL for more description demonstrated with a simple example:

Introduction to Gradient Descent Algorithm (along with variants) in Machine Learning

- https://www.analyticsvidhya.com/blog/2017/03/introduction-to-gradient-descent-algorithm-along-its-variants/
- https://medium.freecodecamp.org/understanding-gradient-descent-the-most-popular-ml-algorithm-a66c0d97307f

9> Is RMSE scale dependent? Means if we scale the data then we will get lower RMSE?

Answer: RMSE is a measure of accuracy and is used to compare forecasting errors of different models for a particular dataset. It is scale-dependent.

10> What is the difference between linear regression performed by StatsModels.Api and SKLearn? If the coefficients are going to be the same, why worry about other statistical terms like F-statistic, p-value etc which are given by StatsModels.Api?

Answer: I will request the TA / student coordinator to answer this question.

11> Could you please explain the significance of P-Value, adjusted R2, and VIF . can you provide one VIF working then it would be easy.

Answer: This has been discussed earlier.

12> Why we use F statistics in SLR why T statistics no enough ?

Answer: Let's start with R^2. R^2 is the explanatory power, it gives the percent variation in y that can be explained by x. *p*-value is the "probability" attached to the likelihood of getting your data results for the model you have. It is attached to the F statistic that tests the overall explanatory power for a model based on that data. There is no established association/relationship between *p*-value and R-square and depends on the data.

So, R^2 tells you how well your model fits the data and F-test is a statistical test related to it. In general, if you calculated F value in a test and is larger than your F statistic, you can reject the null hypothesis. However, the statistic is only one measure of significance in F-test. You should also consider the *p*-value. The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables.

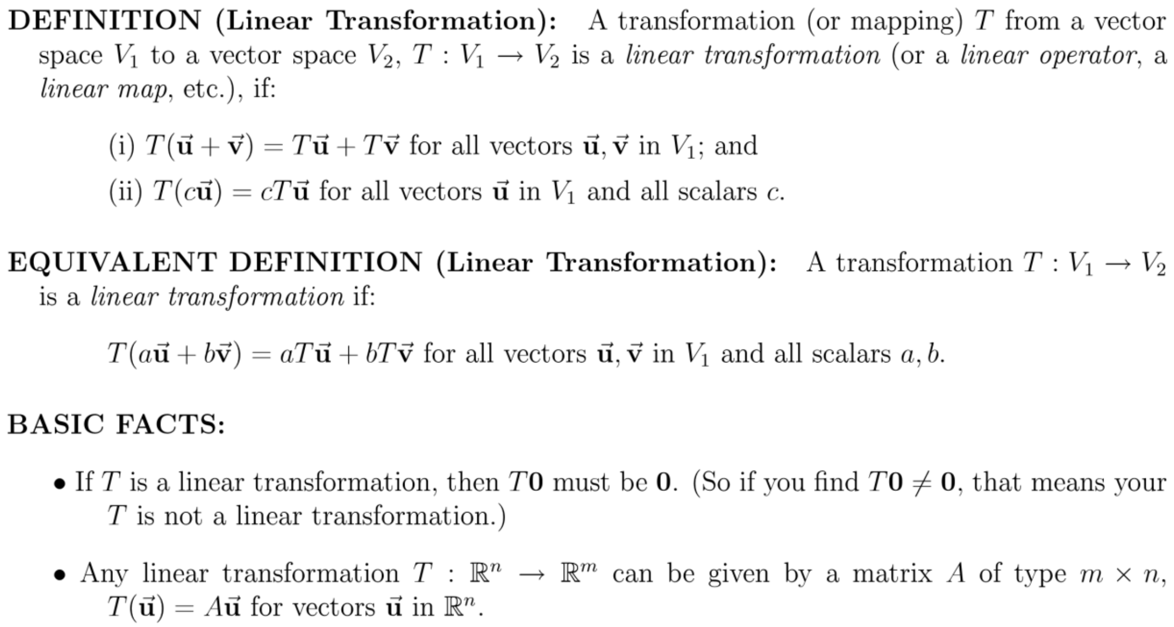
On the other hand, T-tests is used to conduct hypothesis tests on the regression coefficients obtained in simple linear regression. A statistic based on the distribution is used to test the two-sided hypothesis that the true slope equals some constant value. More specifically, **T-test is used to determine whether the slope of the regression line differs significantly from zero.**

The t statistic is the coefficient divided by its standard error. The standard error is an estimate of the standard deviation of the coefficient, the amount it varies across cases. It can be thought of as a measure of the precision with which the regression coefficient is measured.

When you perform a t-test, you usually try to find evidence of a significant difference between population means (2-sample t) or between the population mean and a hypothesized value (1-sample t). The t-value measures the size of the difference relative to the variation in your sample data.

13 > What is linear transformation ?

Answer: Linear transformation (or linear function) is a mapping $V_1 \rightarrow V_2$ between two vector spaces that preserves the operations of vector addition and scalar multiplication. It is a function from one vector space to another that respects the underlying (linear) structure of each vector space.



DEFINITION (Linear Transformation): A transformation (or mapping) T from a vector space V_1 to a vector space V_2, $T : V_1 \to V_2$ is a *linear transformation* (or a *linear operator*, a *linear map*, etc.), if:

(i) $T(\vec{\mathbf{u}} + \vec{\mathbf{v}}) = T\vec{\mathbf{u}} + T\vec{\mathbf{v}}$ for all vectors $\vec{\mathbf{u}}, \vec{\mathbf{v}}$ in V_1; and

(ii) $T(c\vec{\mathbf{u}}) = cT\vec{\mathbf{u}}$ for all vectors $\vec{\mathbf{u}}$ in V_1 and all scalars c.

EQUIVALENT DEFINITION (Linear Transformation): A transformation $T : V_1 \to V_2$ is a *linear transformation* if:

$$T(a\vec{\mathbf{u}} + b\vec{\mathbf{v}}) = aT\vec{\mathbf{u}} + bT\vec{\mathbf{v}} \text{ for all vectors } \vec{\mathbf{u}}, \vec{\mathbf{v}} \text{ in } V_1 \text{ and all scalars } a, b.$$

BASIC FACTS:

- If T is a linear transformation, then $T\mathbf{0}$ must be $\mathbf{0}$. (So if you find $T\mathbf{0} \neq \mathbf{0}$, that means your T is not a linear transformation.)
- Any linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ can be given by a matrix A of type $m \times n$, $T(\vec{\mathbf{u}}) = A\vec{\mathbf{u}}$ for vectors $\vec{\mathbf{u}}$ in $\mathbb{R}^n$.

Please refer this website for more details:
http://people.math.gatech.edu/~xchen/teach/lin_alg/LinearOperator.pdf

14> What is eigenvalues and eigenvectors and how to calculate them?

Answer: I will request the TA / student coordinator to answer this question.

15> What does simple linear regression do ?

Answer: If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to regression analysis.

- Regression analysis means the "*estimation or prediction*" of the unknown value of **one variable** from the known value of the **other variable**.
- It is used to establish linear relationship between continuous variables - predictor variable (or independent variable) and response variable (or dependent variable) using a straight line.

16> What is residual analytics and prediction ?

Answer: This has been discussed before.

17> In y=B0+B1x , we first d significance of Beta B1 through hypothesis testing. Why we also don't focus on B0 ? Doesn't intercept on Y axis give any significance?

Answer: The intercept has a meaningful interpretation if X=0 falls within the range of the X values in the experiment. Say, X=dose of medicine, where in the experiment x=0, 5, 10, 15, 20, 25 and the response variable is body temperature. Here, the intercept means: "the expected body temperature when no medicine is given".

- If the X values in the experiment do not cover zero, then the interpretation of the intercept is simply "the point at the Y axis through which the regression line passes".

To put another example:

It also depends on your case. Sometimes, β_0 has special meaning, and sometimes, it is better to set $\beta_0=0$. These two situations can be defined as regression estimation and ratio estimation respectively. For example:

(1) biomass = β_0 + β_1 x volume for single tree, β_0 means the biomass of a tree with 1.3m height (if β_0 = 1.3).

(2) biomass = β_1 x wet weight for single tree, β_0=0 and β_1 means the moisture content.

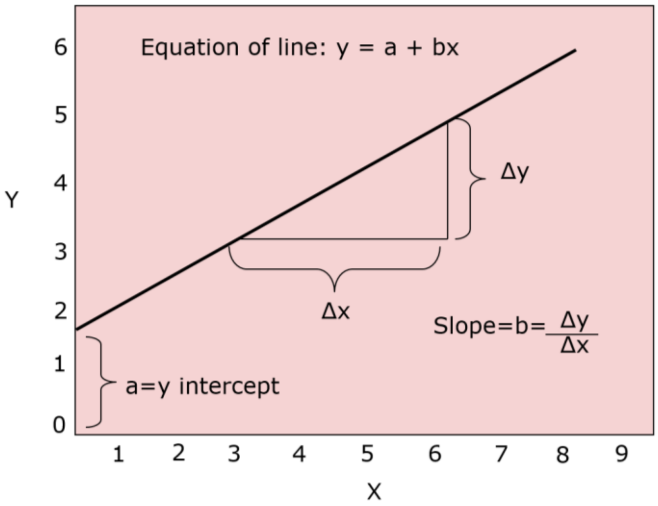
Therefore, it is necessary to analyze whether or not β_0 has real significance.

18> In MLR , manual feature selection , after add a feature which increase R-squre and adjust R-squre but decrease F-statistics -- what is mean of it ? how can intercept it?

Answer: This has been discussed before.

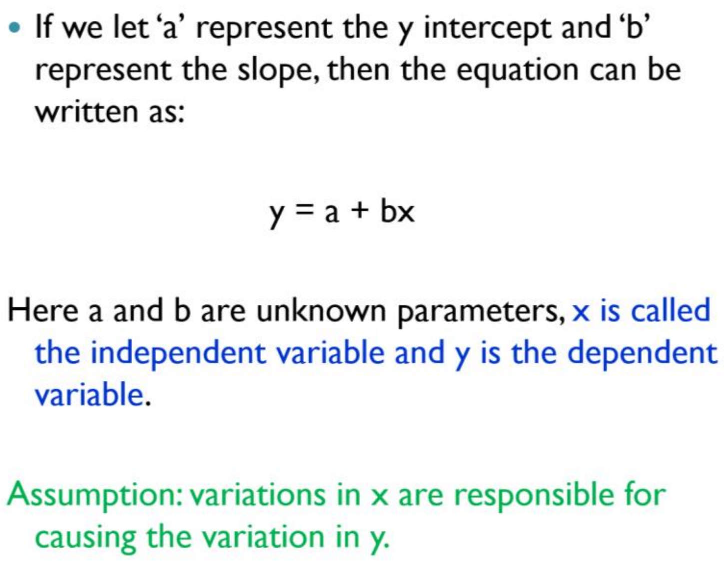
19> When we do the standardized scaling and build the model we could find that the coefficient of two variable(i.e beta1 or slope) is exactly the same with the correlation between the two variable(i.e the pearson's correlation denoted normally by R) . Why this happens

Answer: β_1 coefficient and correlation is not same.



In order to write a mathematical equation that describes a line, we need to know two numbers:

1. The slope
2. Y intercept (constant term)

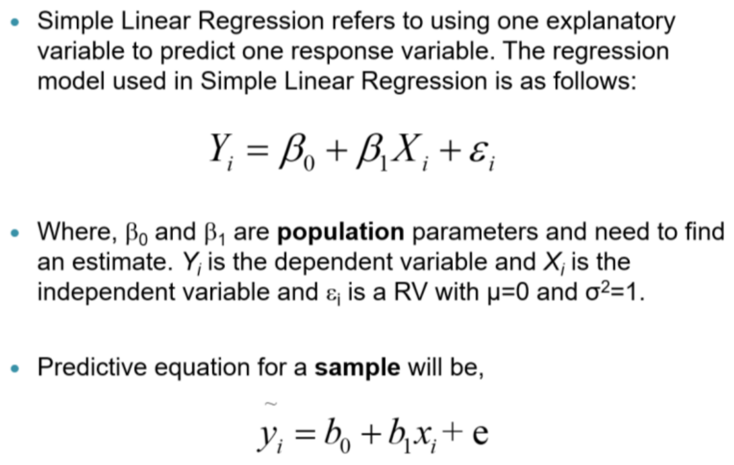


- If we let ‘a’ represent the y intercept and ‘b’ represent the slope, then the equation can be written as:

$$y = a + bx$$

Here a and b are unknown parameters, x is called the independent variable and y is the dependent variable.

Assumption: variations in x are responsible for causing the variation in y.



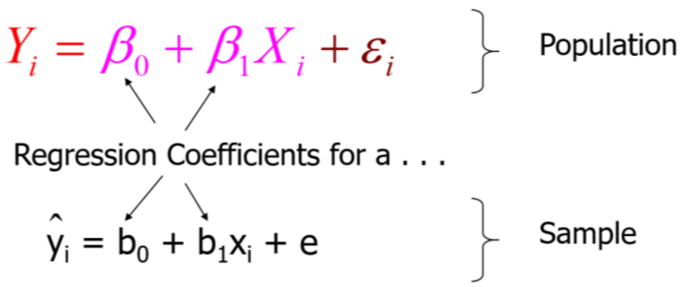
- Simple Linear Regression refers to using one explanatory variable to predict one response variable. The regression model used in Simple Linear Regression is as follows:

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

- Where, β_0 and β_1 are **population** parameters and need to find an estimate. Y_i is the dependent variable and X_i is the independent variable and ε_i is a RV with $\mu=0$ and $\sigma^2=1$.

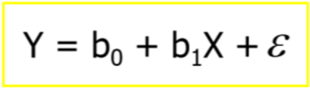
- Predictive equation for a **sample** will be,

$$\tilde{y}_i = b_0 + b_1 x_i + \mathrm{e}$$



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

$$\hat{y}_i = b_0 + b_1 x_i + e$$



$$Y = b_0 + b_1 X + \varepsilon$$

- Residuals – RSS (Residual sum of squares), TSS (Total sum of squares) and R^2 which is coefficient of determination = 1 – RSS/TSS, correlation coefficient (cc or *r*).
- The R^2 value (0 to 1) is equal to the square of correlation coefficient (-1 to +1) between the two variables.

R^2

- ✓ If 1, all the observations fit in a straight line.
- ✓ If 0, x and y are completely independent.

R^2 value gives the percent variation in y that can be explained by x.

CC (correlation coefficient)

- ✓ Measures the strength of the linear association between two variables.
- ✓ Not effected by linear transformations of y or x.
- ✓ Does not distinguish between dependent and independent variables.

20> In Linear Regression ,how did we reach a conclusion that minimizing RSS value will provide us the best model? How to minimize the RSS after calculating the value of RSS. how to plot best fit line for that.

Answer: This has been discussed before.

21> Graph of "Error following a normal distribution" and graph of "Error are dependent" are same. If it follows a normal distribution then how the errors are independent ?

Answer: I will request the TA / student coordinator to answer this question.

22> Please explain how to estimate coefficients of regressions using Calculus and Linear Algebra.

A detailed explanation on this topic will be beyond our discussion. I will request the TA / student coordinator to guide.

23> Where do we use the Z scores and P scores and T scores for doing hypothesis testing, please explain why the availability of population mean and SD and sample mean and SD decides whether we choose Z score or T score for hypothesis testing

A detailed explanation on this topic will be beyond our discussion. This topic will be discussed in one of the module of this course. I will request the TA / student coordinator to guide.

24> What is Lassco Regression?Give me more information on it

A detailed explanation on this topic will be beyond our discussion. I will request the TA / student coordinator to guide.

25> What if we dont scale the data ?

Answer: The variation within the variables will be very high making it difficult to understand and visualise the model.

26> Please explain me Goodness of fit in more detail.

Answer: The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution). In other words, it tells if sample data represents the data you would expect to find in the actual population. Goodness of fit tests commonly used in statistics are:

- The chi-square
- Kolmogorov-Smirnov
- Anderson-Darling
- Shipiro-Wilk

Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (Kolmogorov–Smirnov test), or whether outcome frequencies follow a specified distribution (Pearson's chi-squared test).

27> What is the significance of error terms?

Answer: This has been discussed before.

28> For a linear equation, for every X we get y. And while plotting the best fit we get a deviation. How did we end up getting a normal curve of error terms?

Answer: The error term can be thought of as the composite of a number of minor influences or errors. As the number of these minor influences gets larger, the distribution of the error term tends to approach the normal distribution. This tendency is called the Central Limit Theorem (CLT). The t-test and F- test are not applicable unless the error term is normal distributed. CLT states that the aggregation of a sufficiently large number of independent random variables results in a random variable which will be approximately normal.

29> What if our assumption got failed? what will be next course of action, if we found any assumption failed?

I will request the TA / student coordinator to guide.

30> After building a model and doing the prediction we found that the model is overfitting. If such scenario happens then what measure should we take to come out of the overfitted model and create a model which is more generalized in nature

Answer: To solve the problem of overfitting in the model we need to increase flexibility of our model. But too much of flexibility can also spoil our model, so flexibility should be such that it is optimal value. To increase flexibility we can use regularization technique. They are three types of regularization technique to overcome overfitting.

a) L1 regularization (also called Lasso regularization / panelization.)

b) L2 regularization (also called Ridege regularization/ penalization.)

c) Elastic net

31> Uses of f-statistics / what effect does scaling do.

Answer: R^2 tells you how well your model fits the data and F-test is a statistical test related to it. In general, if you calculated F value in a test and is larger than your F statistic, you can reject the null hypothesis. However, the statistic is only one measure of significance in F-test. You should also consider the *p*-value. The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables.