Lecture Notes: Linear Regression

Session 1: Simple Linear Regression

Modelling uses machine learning algorithms, in which the machine learns from the data just like humans learn from their experiences. Machine learning can be used heavily in the industry. Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

1. **Regression:** The output variable to be predicted is a continuous variable, e.g. scores of a student
2. **Classification:** The output variable to be predicted is a categorical variable, e.g. incoming emails as spam or ham
3. **Clustering:** No predefined notion of label allocated to groups/clusters formed, e.g. customer segmentation for generating discounts.

Regression and classification fall under **supervised learning methods** – in which you have the previous years' data with labels and you use that to build the model.

Clustering falls under **unsupervised learning methods** – in which there is no predefined notion of labels.

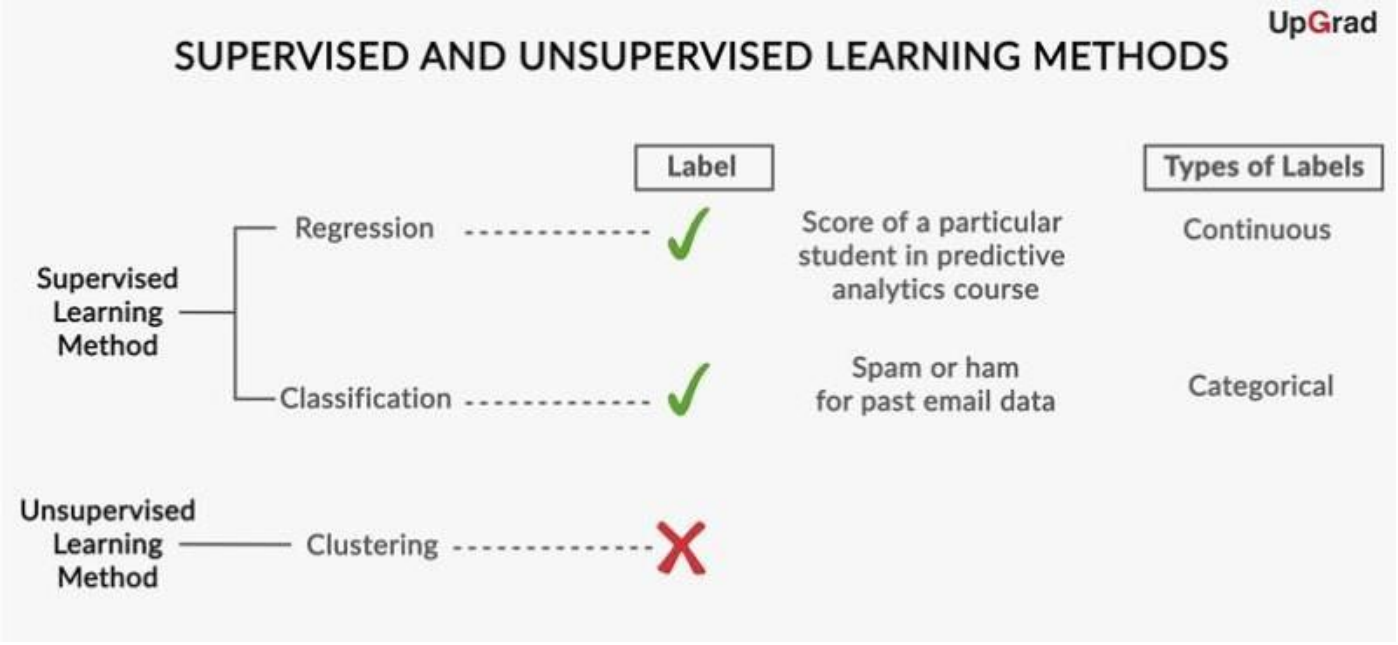


Figure 1 - Supervised and Unsupervised Learning Methods

Regression is the most commonly used predictive analysis model.

As you can guess, accurately predicting future outcomes has applications across industries — in economics, finance, business, medicine, engineering, education and even in sports & entertainment. Given the wide range of applications and its critical importance, it will be very interesting to understand how you can build models to accurately predict future outcomes.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

In this session, you learnt an important class of supervised learning algorithm called linear regression. Nowadays, the word regression is frequently seen while reading the news or any articles related to the

In this session, you learnt an important class of supervised learning algorithm called linear regression. Nowadays, the word regression is frequently seen while reading the news or any articles related to the stock market, finance, even in business. It is more popular on TV media channels for predicting the exit poll results of the election before the actual results are out.

As per our CRISP-DM framework, before developing any predictive models, you first have to define your business objectives and accordingly you have to do the data preparation (as you have already learnt in the data preparation module).

In this module, the focus was more on the prediction of future results by using linear regression concepts. Broadly speaking, it is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).
You learnt about these two types of linear regression under this module:

- Simple linear regression
- Multiple linear regression

1. Simple Linear Regression

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

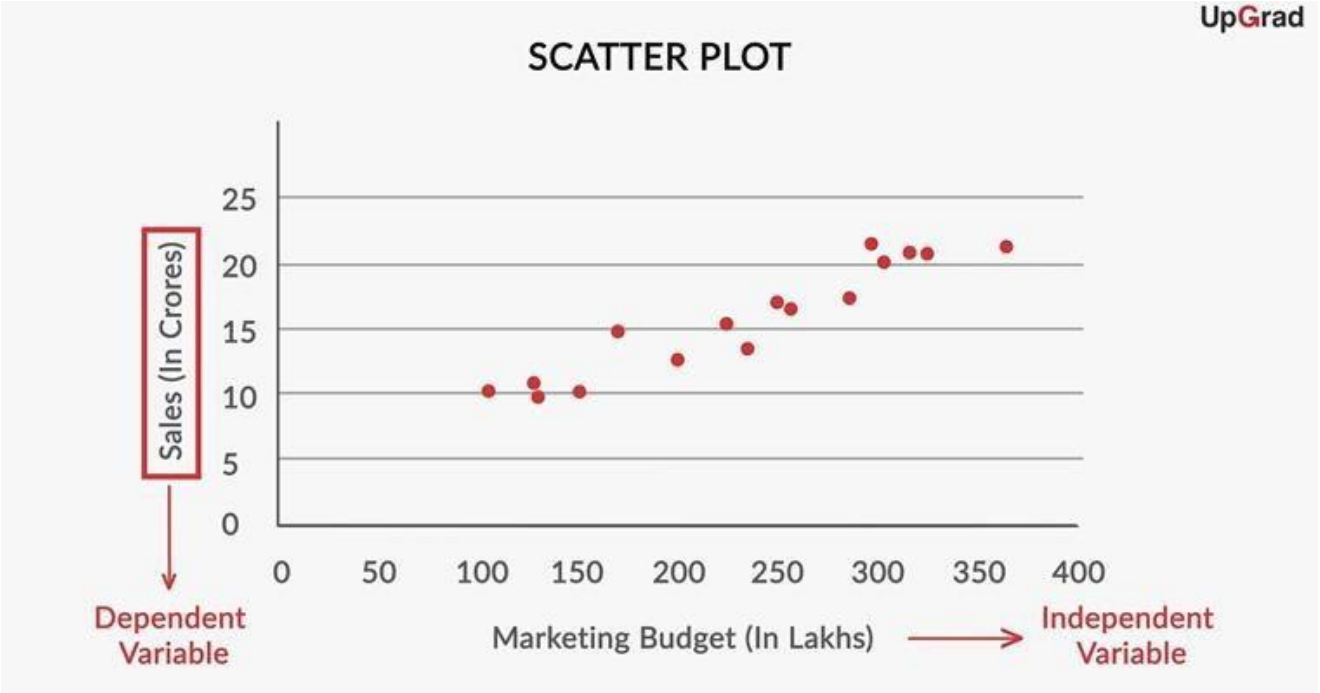


Figure 2 - Scatter plot

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Regression Line

The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$

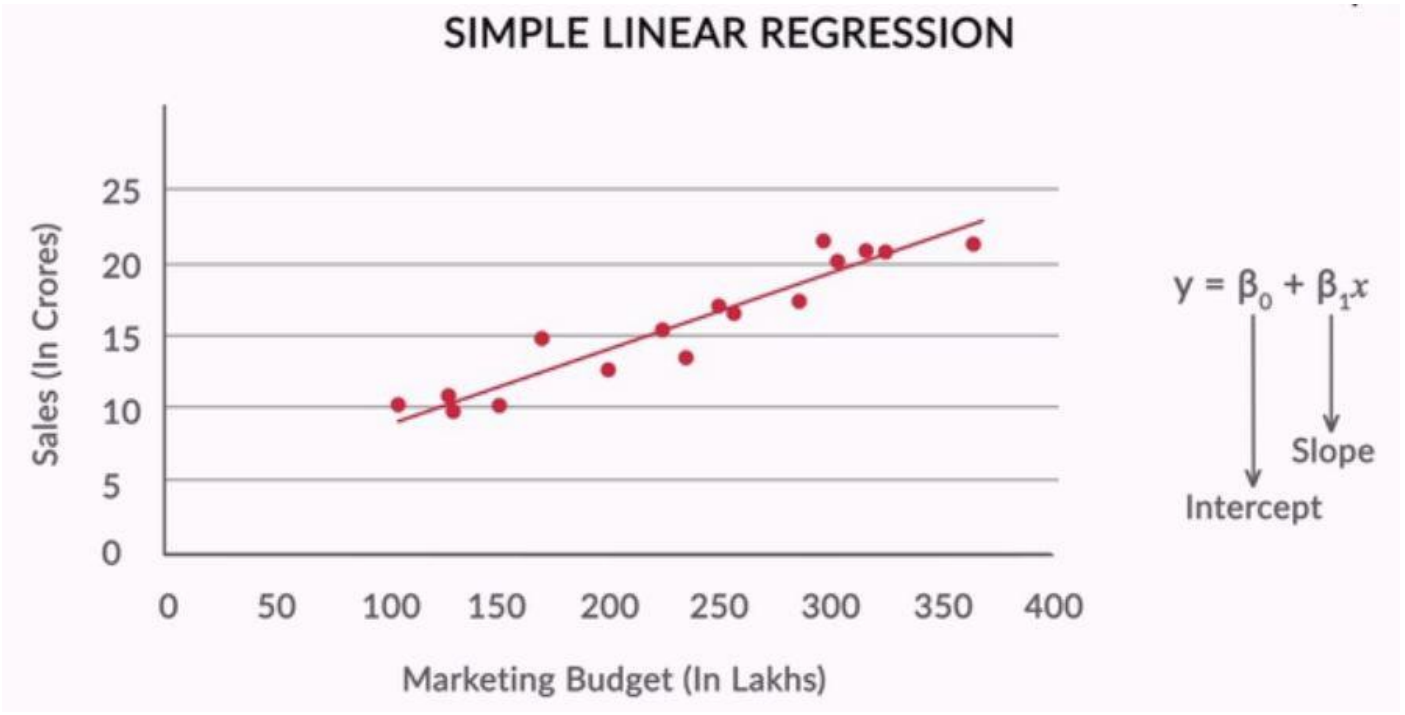


Figure 3 - Regression Line

Best Fit Line

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

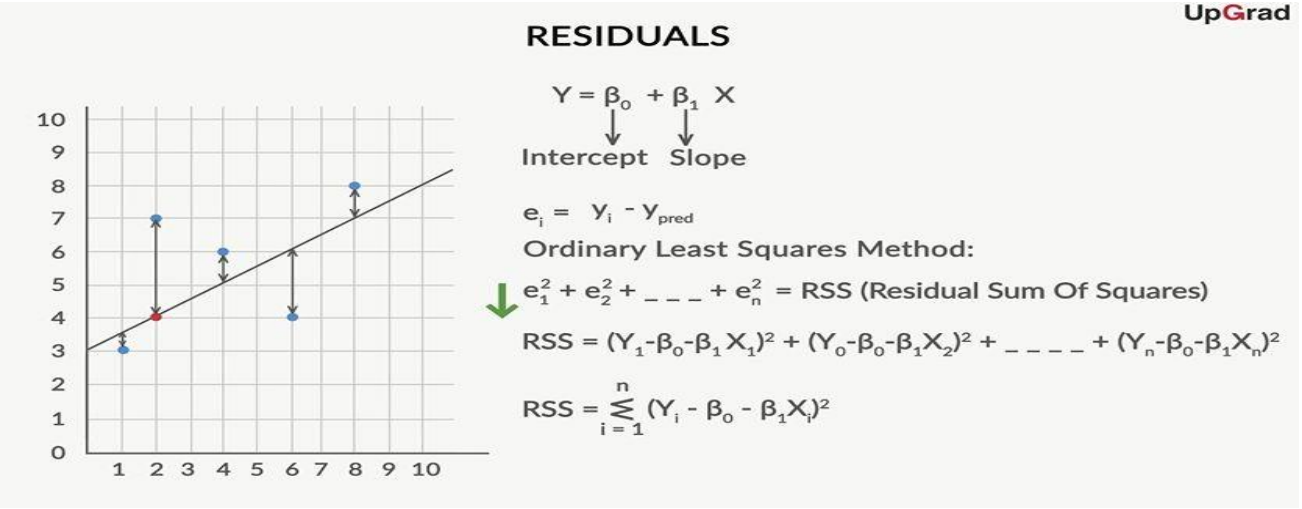


Fig 4 - Residuals

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

R^2 or Coefficient of Determination

You also learnt an alternative way of checking the accuracy of your model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

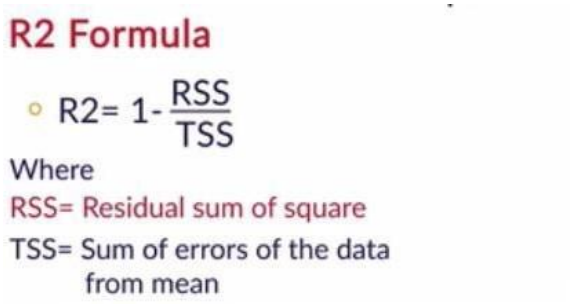


Fig 5 - R-squared

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

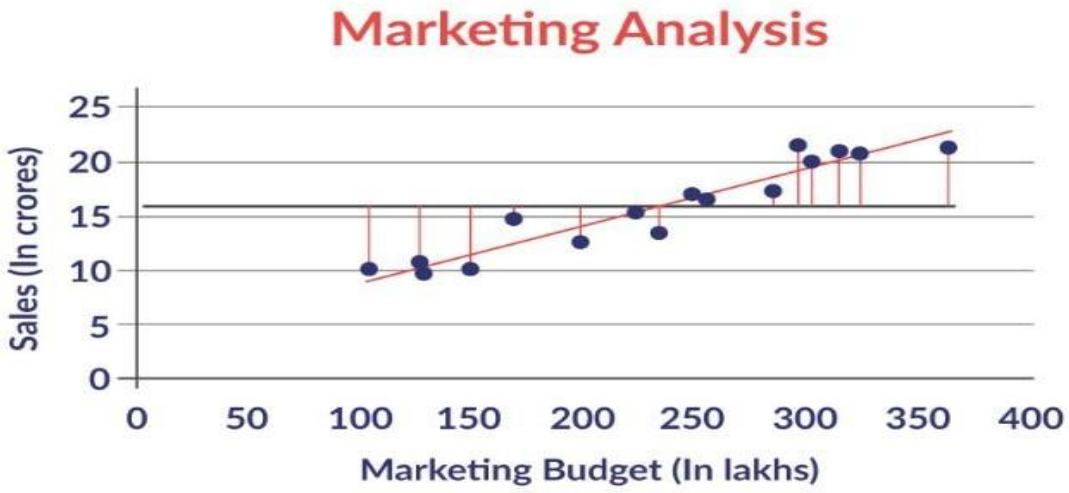
$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Importance of RSS/TSS:

Think about it for a second. If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points as shown below.

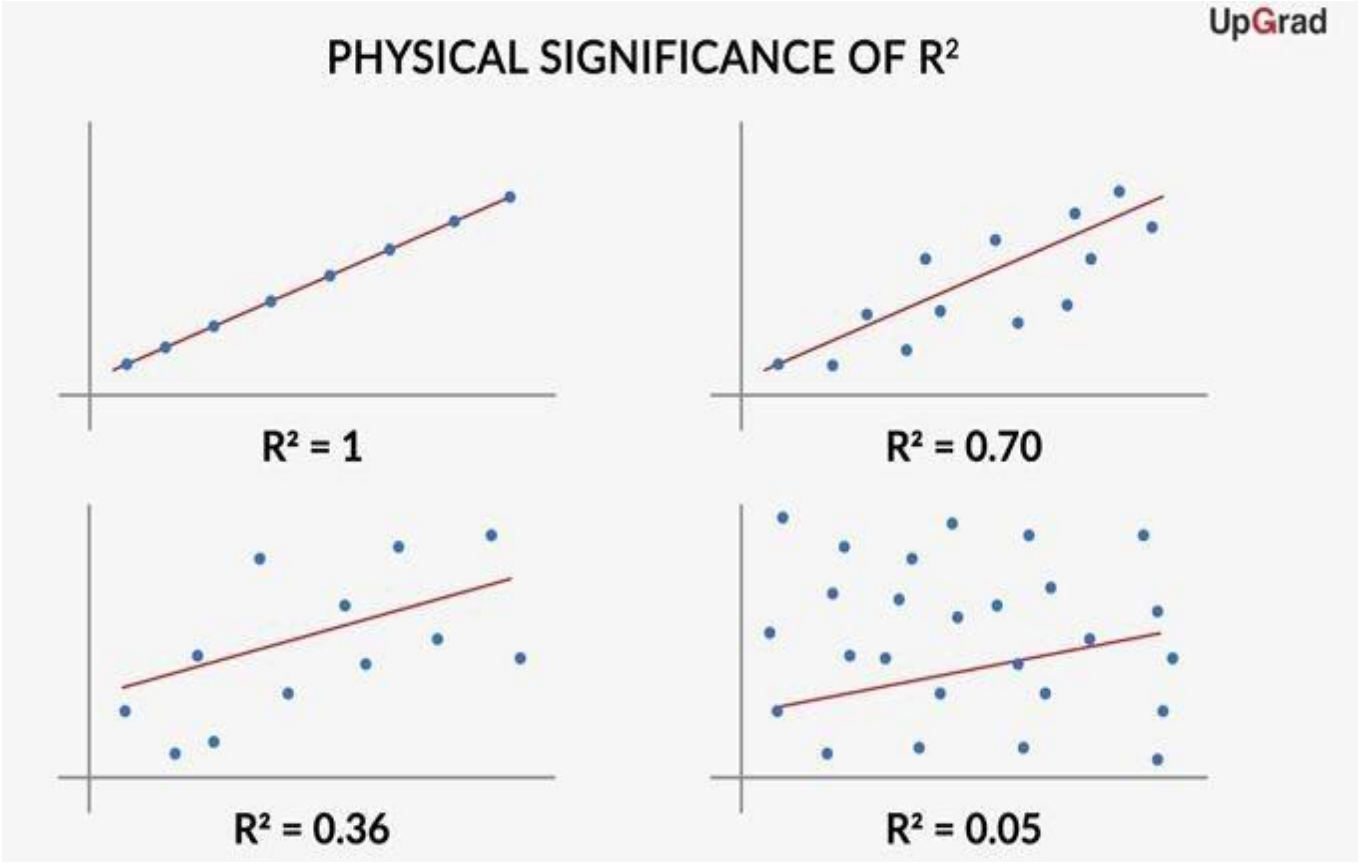
© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

This is the worst possible approximation that you can do. TSS gives us the deviation of all the points from the mean line.



Trying to reinforce this understanding of R2 visually, you can look at the 4 graphs of marketing data and compare the corresponding R2 values.

In Graph 1: All the points lie on the line and the R2 value is a perfect 1
In Graph 2: Some points deviate from the line and the error is represented by the lower R2 value of 0.70 In Graph 3: The deviation further increases and the R2 value further goes down to 0.36
In Graph 4: The deviation is further higher with a very low R2 value of 0.05



© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Session 2: Simple Linear Regression in Python

In this session, you learnt some more theoretical aspects of simple linear regression apart from implementing it in Python.

Assumptions of Simple Linear Regression

Taking a more statistical view:

- Linear regression, at each X, finds the best estimate for Y
- At each X, there is a distribution on the values of Y

Model predicts a single value, therefore there is a distribution of error terms at each of these values as can be seen from the figure below.

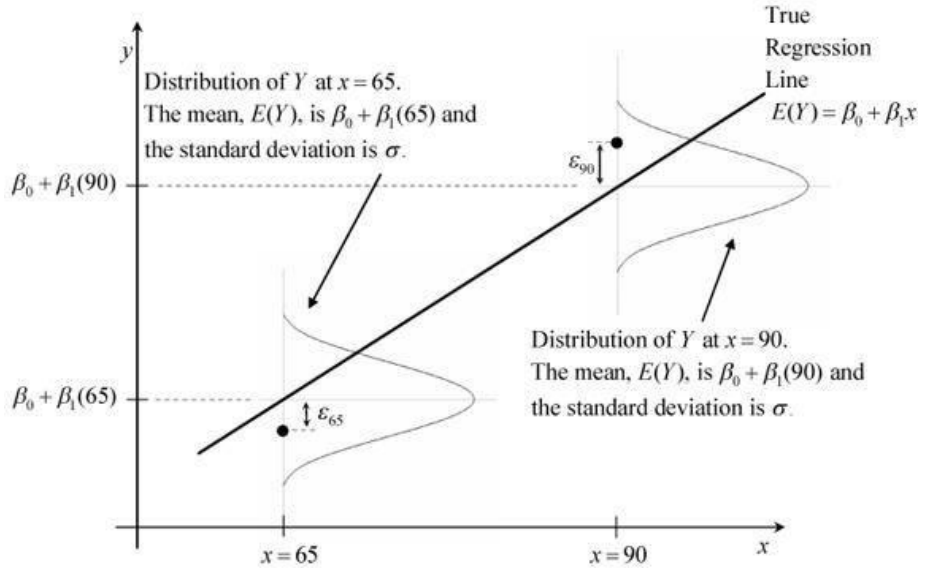


Fig 8 - Normal Distribution of Error Terms

Let's take a look at what the assumptions of simple linear regression were:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to. Also note that, there is **NO assumption on the distribution of X and Y, just that the error terms have to have a normal distribution.**

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Analysing the Residuals

The normal distribution of the residual terms is a very crucial assumption when it comes to making inferences from a linear regression model. Hence, it is very important that you analyse these residual terms before you can move forward. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.

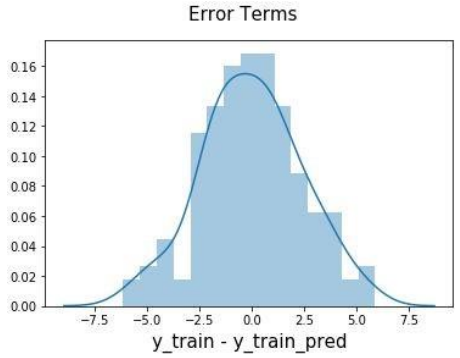


Fig 9 - Histogram of Error Terms

Apart from this, you also need to check for visible patterns in the error terms in order to determine that these terms have a constant variance.

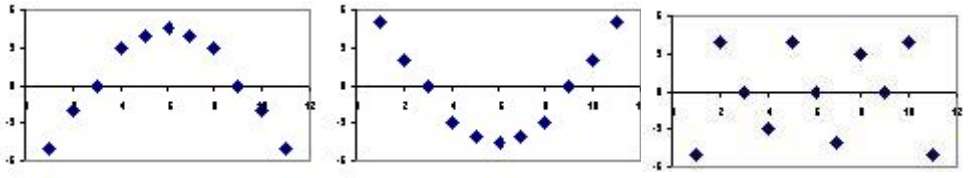


Fig 10 - Checking for Patterns in the Error Terms

As you can see in the image above, the first two clearly seem to display some sort of a pattern but in the third one, the error terms just appear to be evenly distributed noise around zero which is ideal.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Hypothesis Testing of the Beta Coefficient

Once you have fitted a straight line on the data, you need to ask, "Is this straight line a significant fit for the data?" Or simply, is the beta coefficient significant to the extent that it is helping in explaining the variance in the data plotted?

Clearly, you need to perform a hypothesis test on the beta coefficient. The Null and Alternate hypotheses in this case are:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

And to test this hypothesis, the test statistic for beta is:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

This test statistic is follows a student's t-distribution with (n-2) degrees of freedom. The p-value is then calculated on this test statistic in order to determine whether the coefficient is significant or not.

Recap: What is a t distribution?

- *For small sample size, has more spread than Normal distribution*
- *For large sample size, the same as a Normal distribution*

Effectively, it is just a Normal distribution adjusted to account for low sample size

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Assessing the Model Fit

After you have determined that the coefficient is significant, using p-values, you need some other metrics to determine whether the overall model fit is significant. To do that, you need to look at a parameter called the F-statistic.

So, the parameters to assess a model are:

1. **t statistic:** Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. **F statistic:** Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. **R-squared:** After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Please make sure you also review simple linear regression in Python from the notebooks provided.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Session 4: Multiple Linear Regression

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Consider our previous example of sales prediction using TV marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, radio marketing, and newspaper marketing. You need to consider multiple variables as just one variable alone might not be good enough to explain the feature variable, in this case, Sales.

The table below shows how adding a variable helped increase the R-squared that we had obtained by using just the TV variable.

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3

Predictors	R squared
TV	0.816
Radio	0.112
Newspaper	0.058

TV + Newspaper	0.836
TV + Radio	0.910

So we see that adding more variables increases the R-squared and it might be a good idea to use multiple variables to explain a feature variable. Basically:

1. Adding variables helped add information about the variance in Y!
2. In general, we expect explanatory power to increase with increase in variables

Hence, this brings us to multiple linear regression which is just an extension to simple linear regression.

The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from from Simple Linear Regression still hold
    - Zero mean, independent, Normally distributed error terms that have constant

variance

    - The inference part in multiple linear regression also, largely, remains the same.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Although, most of the ideas in simple and multiple linear regression are the same, there are a few new considerations that you need to make when moving to multiple linear regression, such as:

1. Adding more isn't always helpful
    a. Model may 'overfit' by becoming too complex
        i. Model fits the train set 'too well', doesn't generalize
        ii. Symptoms: high train accuracy, low test accuracy
    b. Multicollinearity
        i. Associations between predictor variables
2. Feature selection becomes an important aspect

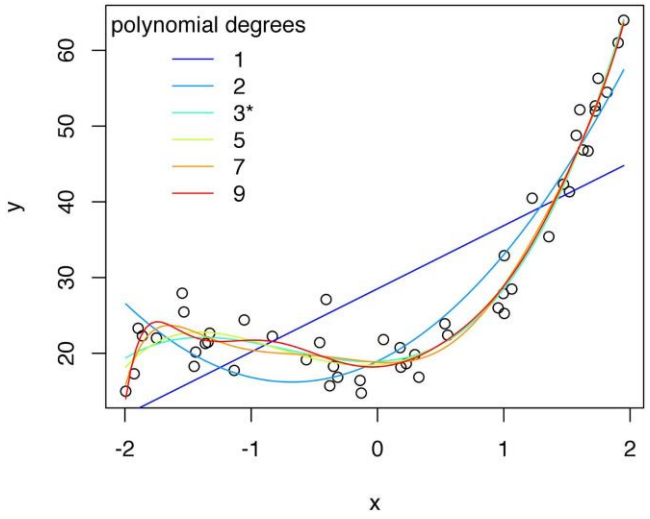


Fig 12 - Overfitting

Let's look at these new considerations one by one:

Overfitting: When you add more and more variables, for example, let's say you keep on increasing the degree of the polynomial function fitting the data, your model might end up memorizing all the data points in the training set. This will cause major problems with generalisation, i.e. now when the model runs on the test data, the accuracy will drop tremendously since, it doesn't generalise well. This is a classical symptom of overfitting.

Multicollinearity: Multicollinearity is the effect of having related predictors in the multiple linear regression model. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, i.e. some of these variables might completely explain some other independent variable in the model due to which the presence of that variable in the model is

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

redundant. So in order to know, where the effect on the feature variable is coming from, we need to drop some of these related independent variables. Basically, multicollinearity affects:

1. Interpretation: Does "change in Y, when all others are held constant" apply?
2. Inference:
    a. Coefficients swing wildly, signs can invert
    b. p-values are, therefore, not reliable

But there are a few aspects that multicollinearity does not affect, such as:

a. The predictions and the precision of the predictions
b. Goodness-of-fit statistics such as R-squared

Hence, dealing with multicollinearity is extremely important. There are two ways to detect multicollinearity in a model:

- **Correlations:** Looking at pairwise correlations between the independent variables can sometimes be useful to detect multicollinearity.

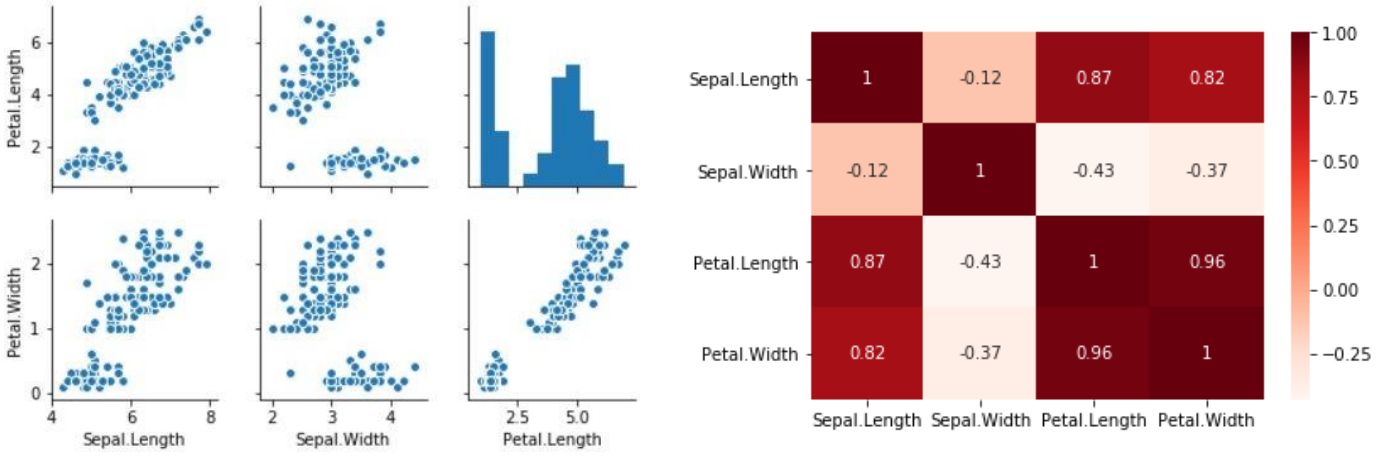


Fig 13 - Pairwise Correlations

From the images above, plotted for iris dataset, you can clearly see that some of the pair of variables, such as, petal width and septal length, petal width and petal length, etc. are highly correlated. Hence, when the model is built, one of the variables from each of these pairs of variables might turn out to be redundant for the model.

- **Variance Inflation Factor (VIF):** Now, looking at correlations might not always be useful as it is possible that just one variable might not completely explain some other variable but some of the variables combined might be able to do that. To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

Now, after any multicollinearity has been detected in the model, you need to deal with it appropriately in order to avoid building an unnecessarily complex model with a lot of redundant variables. The few methods to deal with multicollinearity are:

1. Dropping variables
    a. Drop the variable that is highly correlated with others
    b. Pick the business interpretable variable (if interpretation and explicability important)
2. Create new variable using the interactions of the older variables
    a. Add interaction features, i.e. features derived using some of the original features
        i. bedrooms/bathrooms
        ii. area/stories
    b. Variable transformations:
        i. PCA (covered in a later module)

Feature Scaling: Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:

1. **Standardizing:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. **MinMax Scaling:** The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Handling Categorical Variables

In simple linear regression, you worked with just numeric variables. But when you have multiple variables, there might be some categorical variables that might turn out to be useful for the model. So it is essential to handle these variables appropriately in order to get a good model. One way to deal with them is creating dummy variables. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. See the below example to get a clearer idea.

Value	Indicator Variable
Gender	Female
Male	0
Female	1

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Fig 14 - Dummy variables

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Handling Categorical Variables

Since, a multiple linear regression can be built with different combinations of the variables present, model comparison and hence, selection of the best model becomes extremely essential. The key aspect while selecting the best model is the trade-off between selecting the model explaining the variance best and the model which is fairly simple. So to implement this idea, you need a few parameters apart from the original ones (like R-squared) that would test the goodness of the model as well as penalise the model for using more number of predictor variables. Hence, two new parameters come into picture to assess a multiple linear regression model:

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

$$AIC = n * log(\frac{RSS}{n}) + 2p$$

These parameters are useful for selecting the best model that is fairly simple as well as explains a decent amount of variance. Apart from these you also learnt that there is another parameter called **BIC**, which is quite similar to AIC, the only difference being that it penalises the model more for adding more variables.

Feature Selection

So far, we have talked about dropping features from our model. But choosing to drop the correct features (that are redundant and not adding any value to the model) is quite essential. So let's talk about the various methods for optimal feature selection:

1. Try all possible combinations (2^p models for p features)
    - Time consuming and practically unfeasible
2. Manual Feature Elimination
    - Build model
    - Drop features that are least helpful in prediction (high p-value)
    - Drop features that are redundant (using correlations, VIF)
    - Rebuild model and repeat
3. Automated Approach
    - Recursive Feature Elimination (RFE)
    - Forward/Backward/Stepwise Selection based on AIC (not covered)

It is generally recommended that you follow a balanced approach, i.e., use a combination of automated (coarse tuning) + manual (fine tuning) selection in order to get an optimal model.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights

Disclaimer: All content and material on the UpGrad website is copyrighted material, either belonging to UpGrad or its bonafide contributors and is purely for the dissemination of education. You are permitted to access print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copies of this document, in part or full, saved to disc or to any other storage medium may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or the uploading thereof on other websites or use of content for any other commercial/unauthorized purposes in any way which could infringe the intellectual property rights of UpGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document will be modified, adapted or altered in any way.
- No part of this document or UpGrad content may be reproduced or stored in any other web site or included in any public or private electronic retrieval system or service without UpGrad's prior written permission.
- Any rights not expressly granted in these terms are reserved.

© Copyright 2018. UpGrad Education Pvt. Ltd. All rights