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.

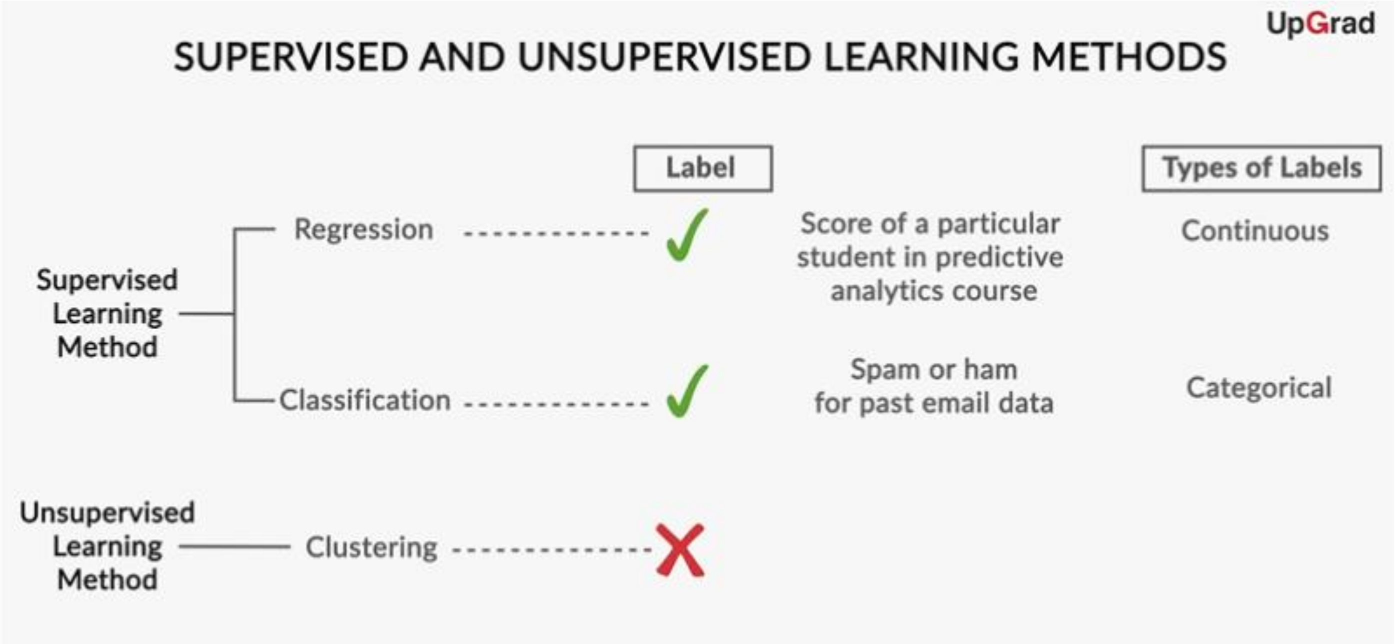

UpGrad
SUPERVISED AND UNSUPERVISED LEARNING METHODS
Label
Types of Labels
Supervised Learning Method
Regression
✓
Score of a particular student in predictive analytics course
Continuous
Classification
✓
Spam or ham for past email data
Categorical
Unsupervised Learning Method
Clustering
✗

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.

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.

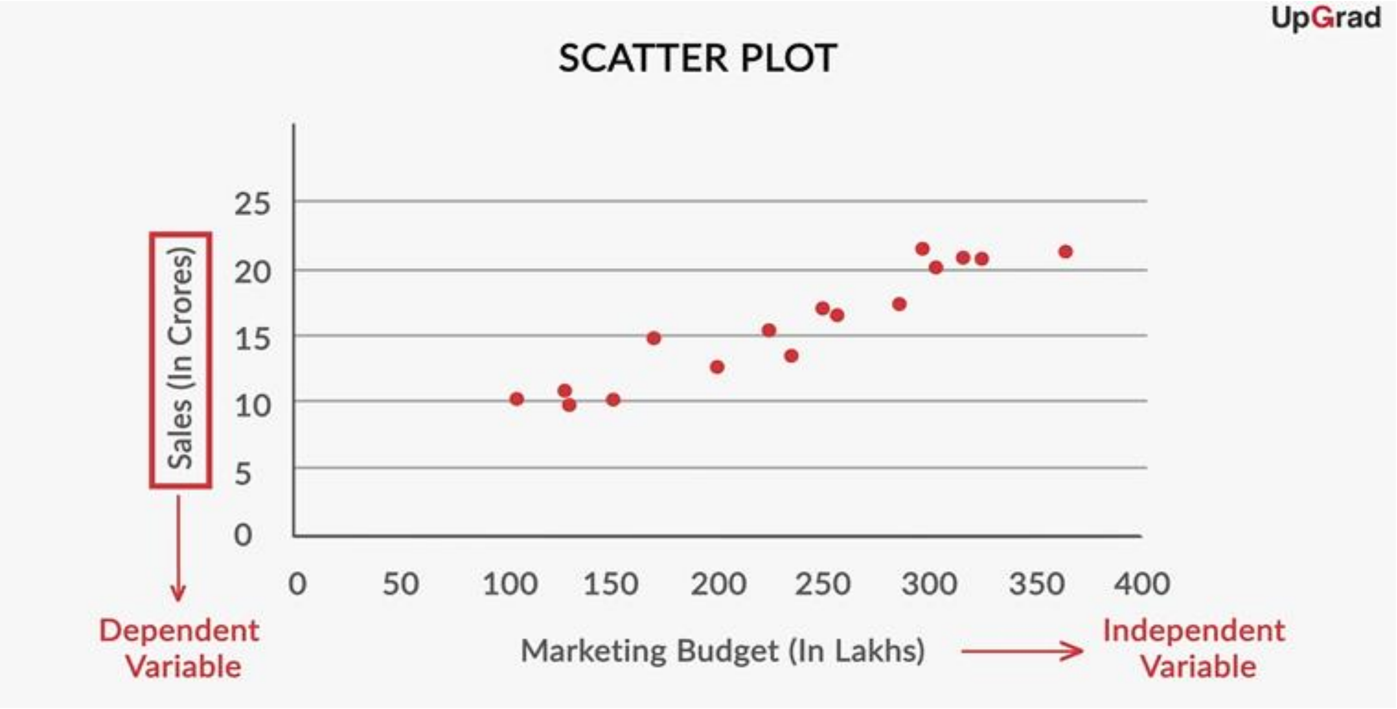


Figure 2 – Scatter plot

Regression Line

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

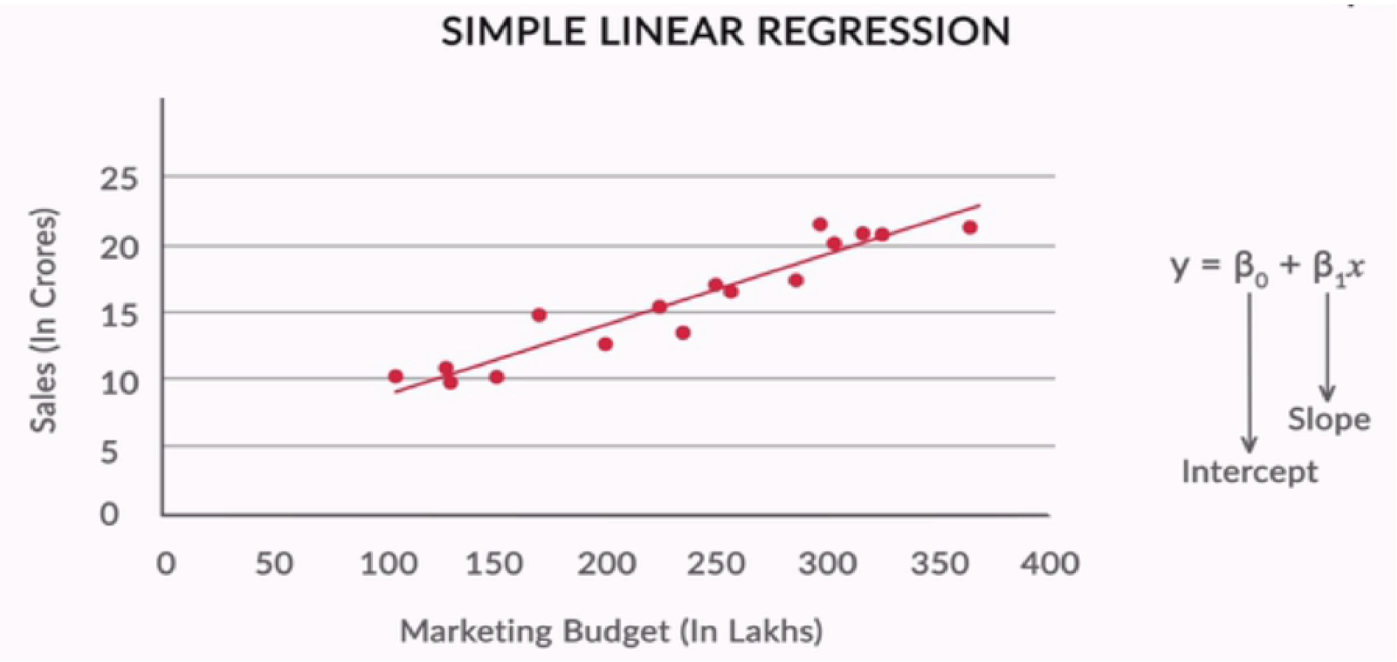
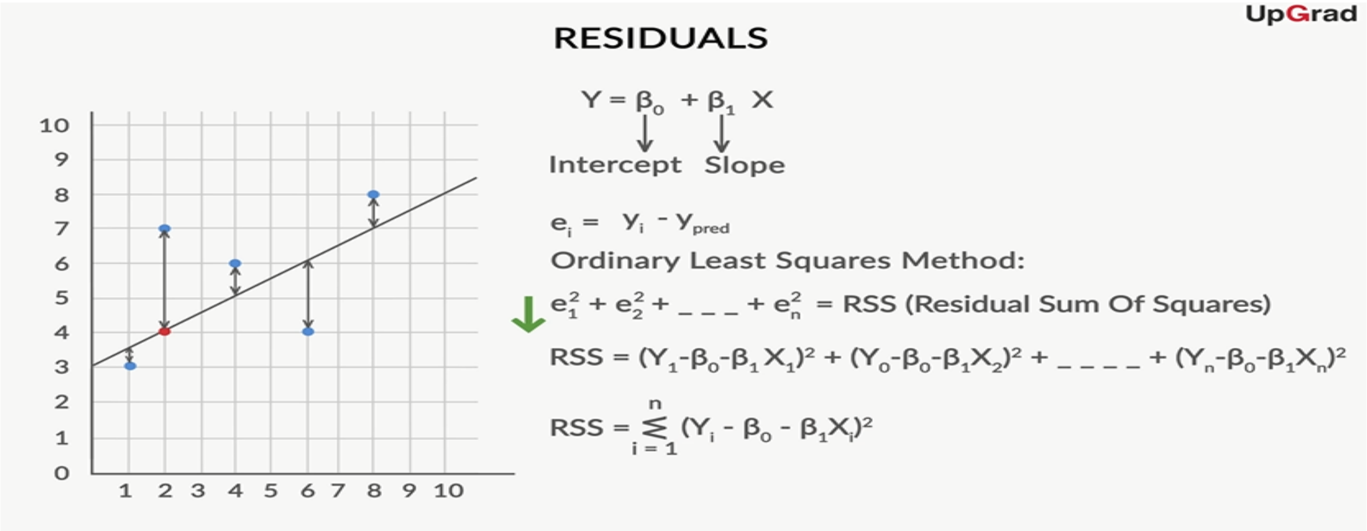


Figure 3 - Regression Line

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:



RESIDUALS

$Y = \beta_o + \beta_1 X$

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

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

$RSS = (Y_1-\beta_o-\beta_1X_1)^2 + (Y_o-\beta_o-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_o-\beta_1X_n)^2$

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

Figure 4 – Residuals

Strength of Linear Regression Model

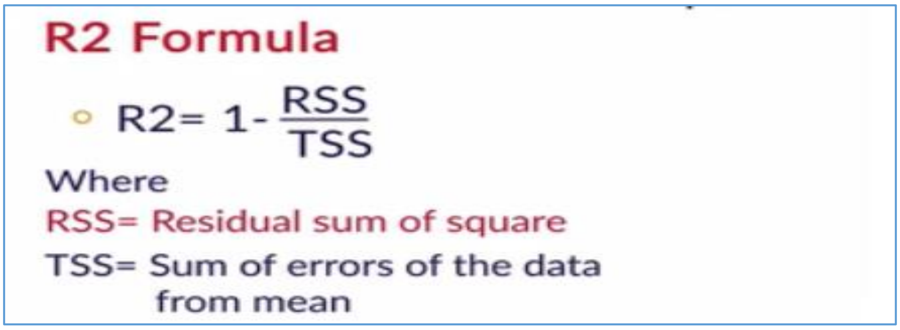
The strength of the linear regression model can be assessed using 2 metrics:

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

1. 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: $R^2 = 1 - (RSS / TSS)$



R2 Formula

- $R2= 1- \frac{RSS}{TSS}$

Where

RSS= Residual sum of square

TSS= Sum of errors of the data from mean

Figure 5: R^2

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.

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

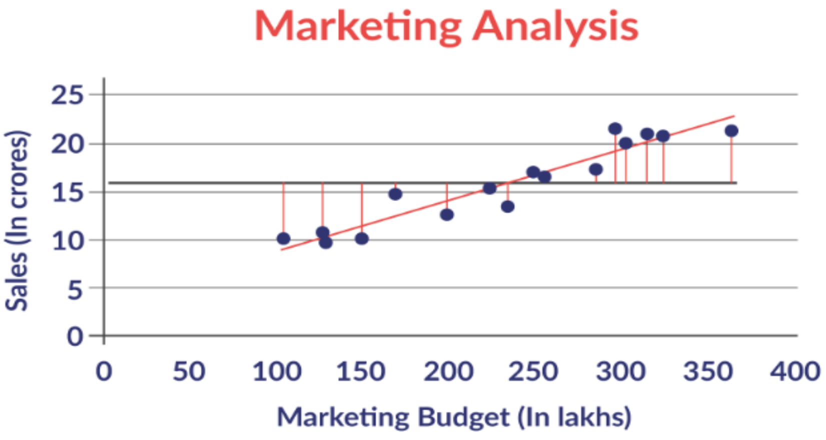


Figure 6: Visualisation of RSS and TSS

Trying to reinforce this understanding of R^2 visually, you can look at the 4 graphs of marketing data and compare the corresponding R^2 values.

In Graph 1: All the points lie on the line and the R^2 value is a perfect 1

In Graph 2: Some points deviate from the line and the error is represented by the lower R^2 value of 0.70

In Graph 3: The deviation further increases and the R^2 value further goes down to 0.36

In Graph 4: The deviation is further higher with a very low R^2 value of 0.05

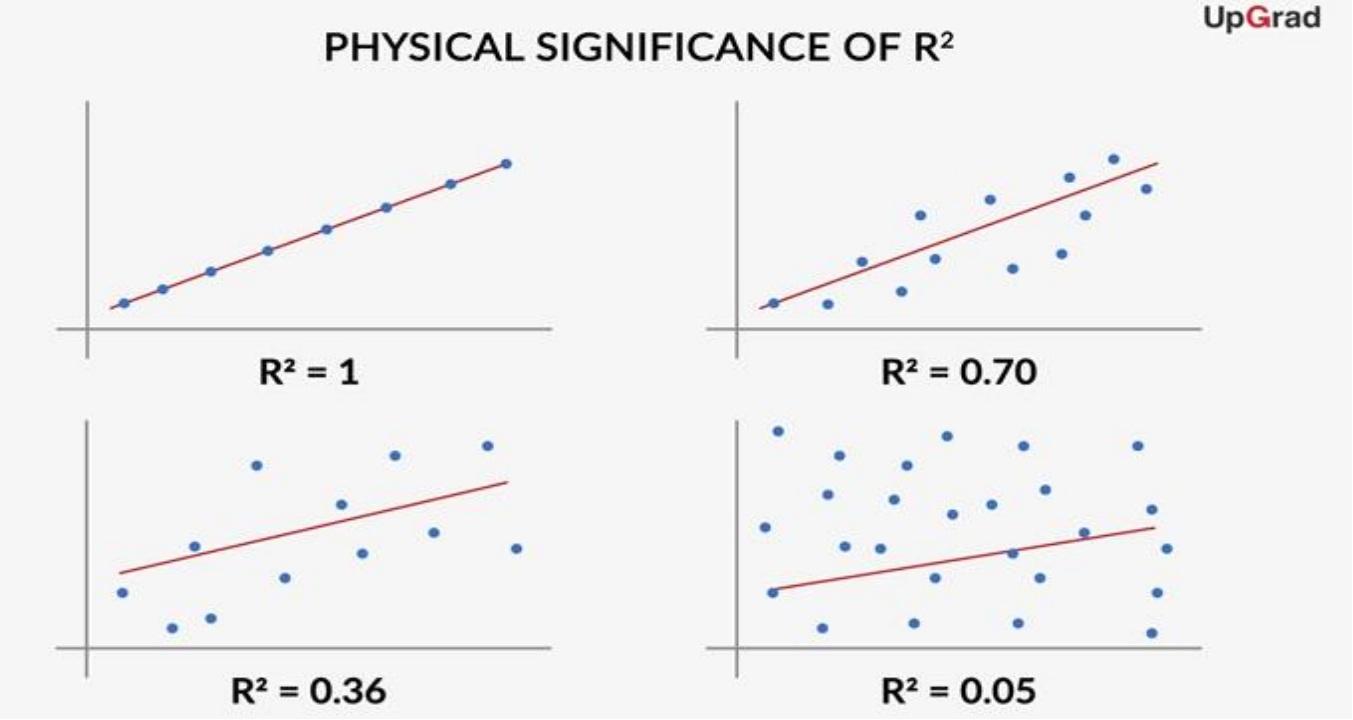


Figure 7: Physical Significance of R^2

Simple Linear Regression in Python

In a previous module, you went through an example of deodorant sales prediction. In this session, you learnt to implement these concepts in Python with the same example.

Let's recap the steps you need to perform while building a **simple linear regression model** in Python:

1. Import the data set and required libraries.

```
import pandas as pd

advertising = pd.read_csv("tvmarketing.csv")
```

2. Understanding the data frame.

```
# To display the first 5 rows

advertising.head()


#  To display the last 5 rows

advertising.tail()
```

3. Preparing X and y

```
# Putting feature variable to X

X = advertising['TV']

# Putting response variable to y

y = advertising['Sales']
```

4. Splitting data into train and test

```
from sklearn.cross_validation import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7 , random_state=100)
```

5. It is a general convention in scikit-learn that observations are rows, while features are columns. This is needed only when you are using a single feature; in this case, 'TV'.

```
import numpy as np

X_train = X_train[:, np.newaxis]

X_test = X_test[:, np.newaxis]
```

6. Performing Linear Regression

```
# import LinearRegression from sklearn

from sklearn.linear_model import LinearRegression


# Representing LinearRegression as lr(Creating LinearRegression Object)

lr = LinearRegression()


#You don't need to specify an object to save the result because 'lr' will take the results of the fitted model.

lr.fit(X_train, y_train)
```

7. Coefficients calculation

```
# Print the intercept and coefficients

print(lr.intercept_)

print(lr.coef_)
```

8. Making predictions

```
# Making predictions on the testing set

y_pred = lr.predict(X_test)
```

9. Model evaluation (Plot Actual vs Predicted)

```
# Actual vs Predicted

import matplotlib.pyplot as plt

c = [i for i in range(1,61,1)]         # generating index

fig = plt.figure()

plt.plot(c,y_test, color="blue", linewidth=2.5, linestyle="-")

plt.plot(c,y_pred, color="red",  linewidth=2.5, linestyle="-")

fig.suptitle('Actual and Predicted', fontsize=20)
```

```
plt.xlabel('Index', fontsize=18)
```

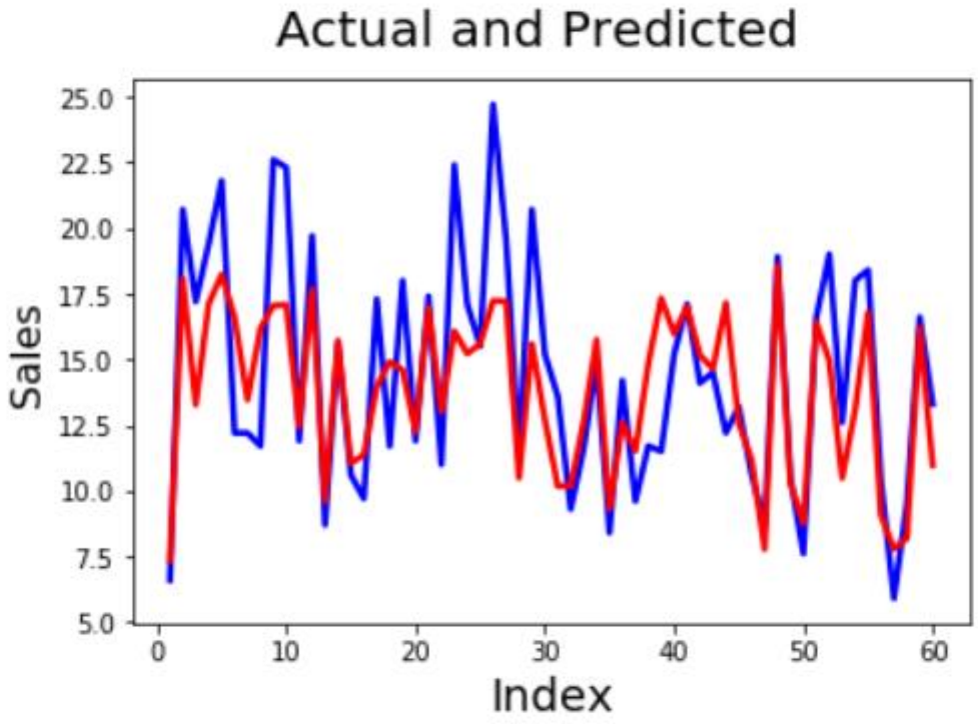


Figure 8: Actual vs Predicted

10. Model evaluation (Plot Error terms)

```
# Error terms

c = [i for i in range(1,61,1)]

fig = plt.figure()

plt.plot(c,y_test-y_pred, color="blue", linewidth=2.5, linestyle="-")

fig.suptitle('Error Terms', fontsize=20)

plt.xlabel('Index', fontsize=18)

plt.ylabel('ytest-ypred', fontsize=16)
```

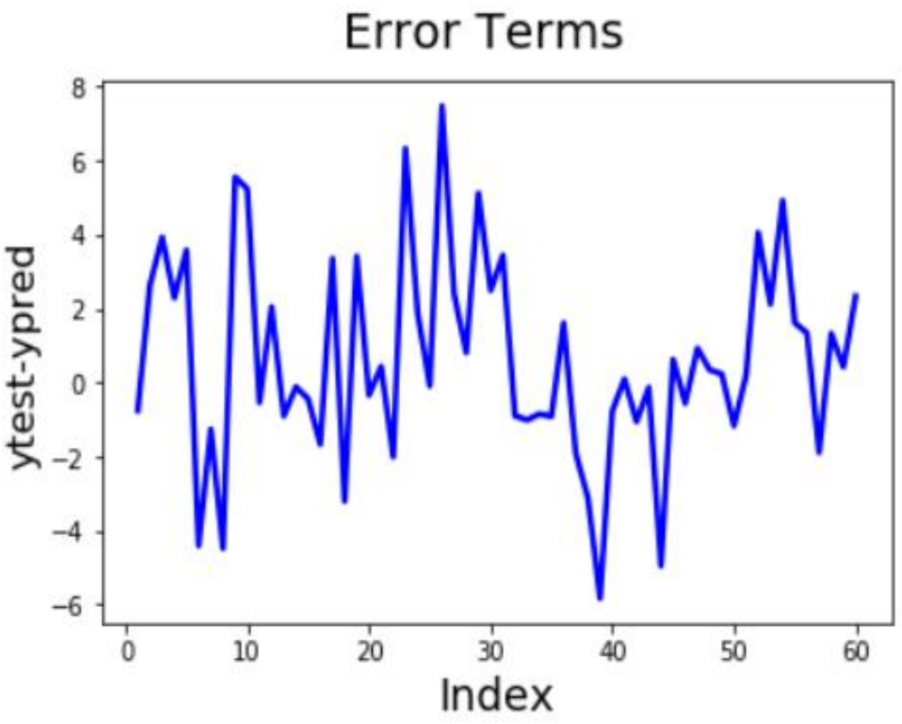


Figure 9: Error terms

11. Checking mean square error and R square

```
from sklearn.metrics import mean_squared_error, r2_score

mse = mean_squared_error(y_test, y_pred)

r_squared = r2_score(y_test, y_pred)


print('Mean_Squared_Error :' ,mse)

print('r_square_value :',r_squared)
```

You can now check the mean square error and r square value of your model.

Your model is getting a mean square error of 7.9 which means the model is not able to match 7% of the values only, which is good.

The r square value is about 60% which means our model is able to explain 60% of the variance which is also good.

Overall your model is very decent.

Multiple Linear Regression

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory 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.

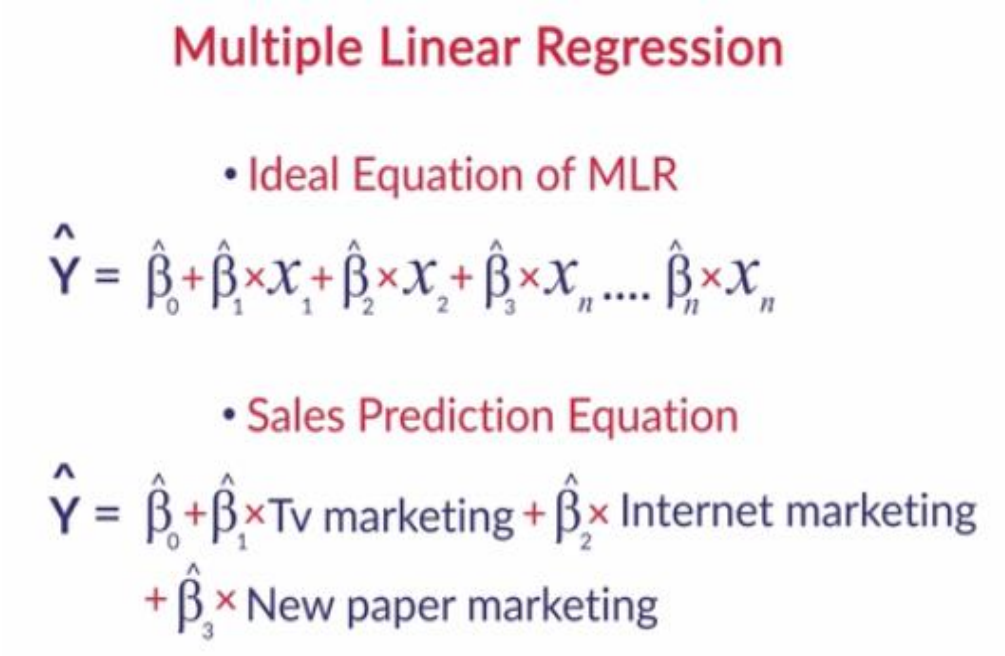
Consider our previous example of sales prediction using TV Marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, Radio marketing, and Newspaper marketing.

The simple linear regression model is built on a straight line which has the following formula:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Multiple linear regression also uses a linear model that can be formulated in a very similar way.

Thus, the equation of multiple linear regression would be as follows:



Multiple Linear Regression

- Ideal Equation of MLR

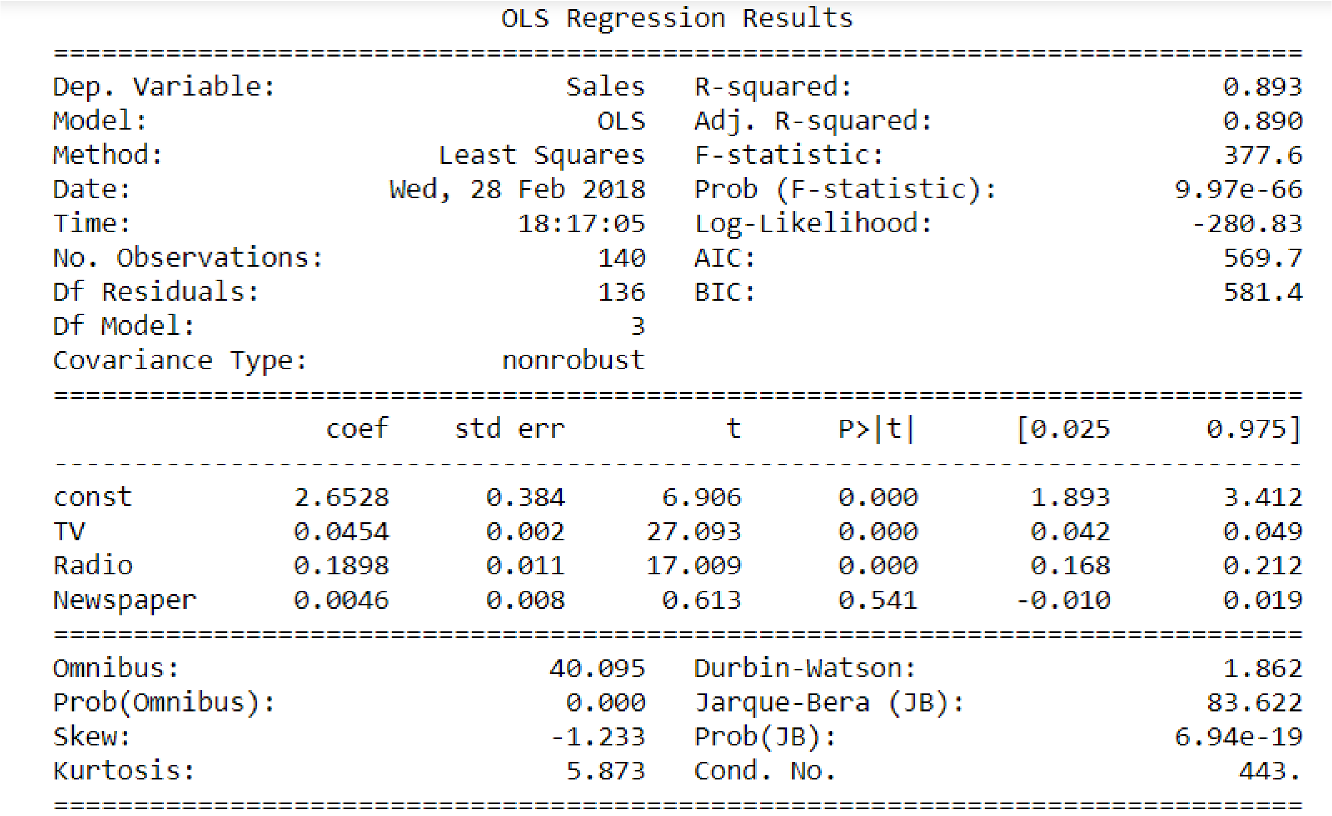
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times x_1 + \hat{\beta}_2 \times x_2 + \hat{\beta}_3 \times x_n \text{....} \hat{\beta}_n \times x_n$$

- Sales Prediction Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Tv marketing} + \hat{\beta}_2 \times \text{Internet marketing} + \hat{\beta}_3 \times \text{New paper marketing}$$

Figure 10: Multiple Linear Regression Equation

You built the model containing all variables in python using sklearn.linear_model. You obtained results that were as follows:



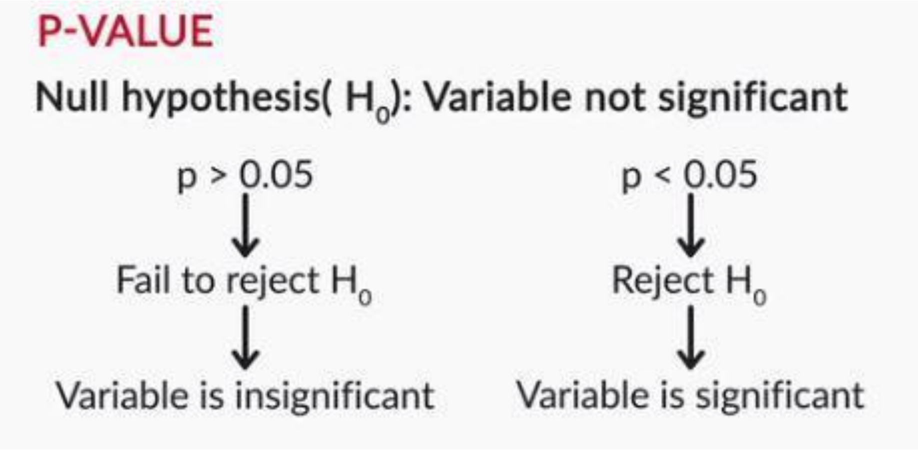
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.893
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     377.6
Date:                Wed, 28 Feb 2018   Prob (F-statistic):           9.97e-66
Time:                        18:17:05   Log-Likelihood:                -280.83
No. Observations:                 140   AIC:                             569.7
Df Residuals:                     136   BIC:                             581.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.6528      0.384      6.906      0.000       1.893       3.412
TV             0.0454      0.002     27.093      0.000       0.042       0.049
Radio          0.1898      0.011     17.009      0.000       0.168       0.212
Newspaper      0.0046      0.008      0.613      0.541      -0.010       0.019
==============================================================================
Omnibus:                       40.095   Durbin-Watson:                   1.862
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               83.622
Skew:                          -1.233   Prob(JB):                     6.94e-19
Kurtosis:                       5.873   Cond. No.                         443.
==============================================================================
```

Figure 11: Summary advertising model

You can see the estimated value of the intercept is around 2.6, and the estimated coefficients of Tv_marketing (beta1), Radio (beta2) and Newspaper(beta3) are **0.04**, **0.18** and **0.004** respectively.

Next is **standard error**. It measures the variability in the estimate for these coefficients. A lower value of standard deviation is good but it is somewhat relative to the value of the coefficient. E.g. you can check the standard error of the intercept is about 0.38, whereas its estimate is 2.6, So, it can be interpreted that the variability of the intercept is from 2.6±0.38. Note that standard error is absolute in nature and so many a times, it is difficult to judge whether the model is good or not. Here comes the next parameter, i.e. t-value which is the ratio of the estimated coefficients to the standard deviation of the estimated coefficients. It measures whether or not the coefficient for this variable is meaningful for the model. Though you may not use this value itself, you should know that it is used to calculate the p-value and the significance levels which are used for building the final model.

A very important parameter of this analysis is the **p-value**. Recall from the Statistics course that p-value is used for hypothesis testing. Here, in regression model building, the null hypothesis corresponding to each p-value is that the corresponding independent variable does not impact the dependent variable. The alternate hypothesis is that the corresponding independent variable impacts the response. Now, p-value indicates the probability that the null hypothesis is true. Therefore, a low p-value, i.e. less than 0.05, indicates that you can reject the null hypothesis.



P-VALUE

Null hypothesis(H_0): Variable not significant

$p > 0.05$ → Fail to reject H_0 → Variable is insignificant

$p < 0.05$ → Reject H_0 → Variable is significant

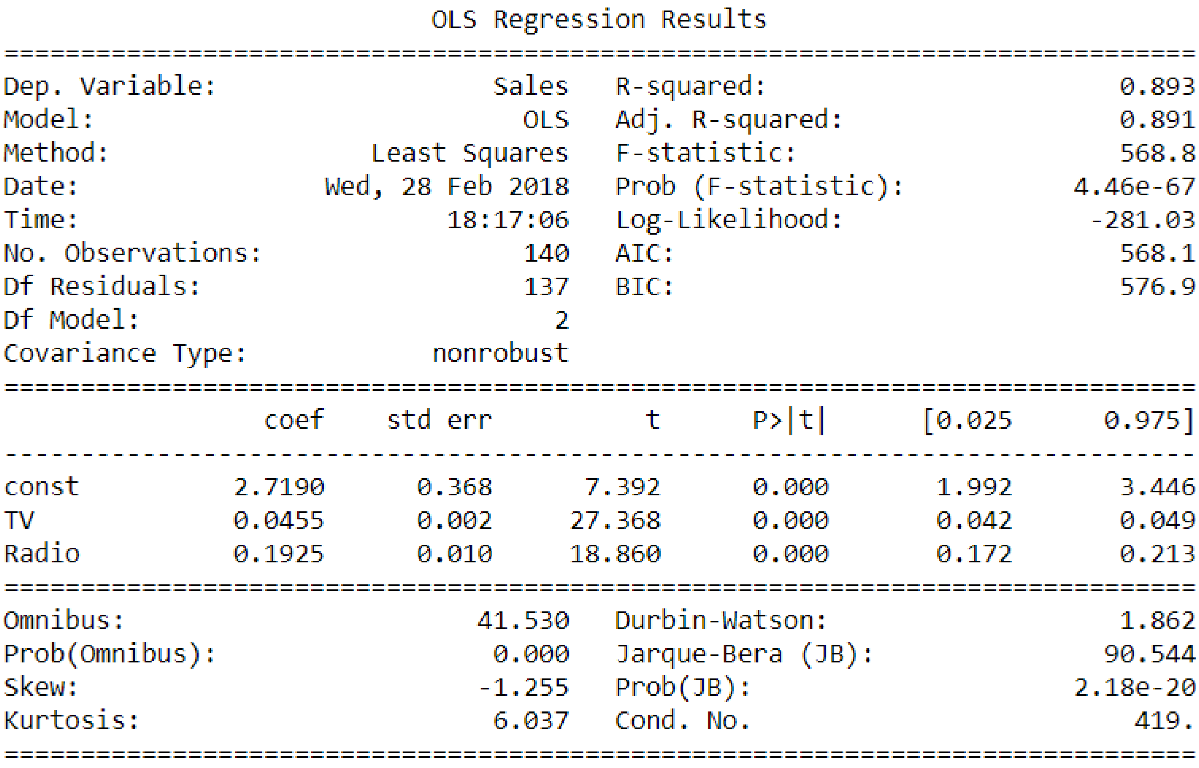
Figure 12: p-value

In other words, you can say that the independent variables that have a low p-value are likely to be a meaningful addition to your model.

Coming to the problem, you can see (in Figure 3) that the p-value of TV marketing and Radio marketing is less than 0.5, but the p-value of the Newspaper variable is 0.54.

Since the Newspaper variable was found insignificant, it, should be removed from the model. Thus, the new model would include only two independent variables, i.e. TV marketing and Radio, which will be regressed with the dependent variable "sales". Store new linear model having these two variables into the object "model_2".

Check the summary of *model_2*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.893
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     568.8
Date:                Wed, 28 Feb 2018   Prob (F-statistic):           4.46e-67
Time:                        18:17:06   Log-Likelihood:                -281.03
No. Observations:                 140   AIC:                             568.1
Df Residuals:                     137   BIC:                             576.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.7190      0.368      7.392      0.000       1.992       3.446
TV             0.0455      0.002     27.368      0.000       0.042       0.049
Radio          0.1925      0.010     18.860      0.000       0.172       0.213
==============================================================================
Omnibus:                       41.530   Durbin-Watson:                   1.862
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               90.544
Skew:                          -1.255   Prob(JB):                     2.18e-20
Kurtosis:                       6.037   Cond. No.                         419.
==============================================================================
```

Figure 13: Model summary

You can see that the values of the estimates have changed where both the variables are highly significant. So overall, this is the final model where you can say that the sales are equal to *2.7 + 0.045 * TV + 0.19 * Radio.*

In this session, you implemented the concepts of both simple linear and multiple linear regression into the real case problem.

A real estate company has a data set of the prices in the region of Delhi. It wishes to use the data to optimise the sale prices of the properties based on important factors such as area, bedrooms, parking, etc.

Essentially, the company wants:

- To identify the variables affecting house prices, e.g. area, number of rooms, bathrooms, etc.
- To create a linear model that quantitatively relates house prices with variables such as number of rooms, area, number of bathrooms, etc.
- To know the accuracy of the model, i.e. how well these variables predict house prices

Always remember that whenever a data set has a large number of variables, you should first understand the business objectives and then get well versed with the data set, otherwise at a later stage it becomes difficult to make business sense out of the results. Next important step is the data preparation part, where you try to treat data points which might cause imperfections in model building. Recall that in one of the previous sessions, you might have heard that data preparation takes 70-80% of the time. Although the data provided to you was cleaned already, you'd have to do it yourself in the industry.

Dummy Variables

The categorical variables need to be converted to numeric form to be used in regression modelling. Thus, you create dummy variables.

For two level variables, you change the levels into 1 and 0 where 1 is one level and 0 is indicating another. In our case, for basement variable 1 indicates the presence of a basement and 0 indicates its absence. But when you directly convert the factor variable to a numeric type, the factor value of that variable is replaced by levels of variable, which is called **coercion**. Thus, you write the code as below.

```
# Converting Yes to 1 and No to 0

telecom['PhoneService'] = telecom['PhoneService'].map({'Yes': 1, 'No': 0})

telecom['PaperlessBilling'] = telecom['PaperlessBilling'].map({'Yes': 1, 'No': 0})

telecom['Churn'] = telecom['Churn'].map({'Yes': 1, 'No': 0})

telecom['Partner'] = telecom['Partner'].map({'Yes': 1, 'No': 0})

telecom['Dependents'] = telecom['Dependents'].map({'Yes': 1, 'No': 0})
```

For multi-level variables, you use a pd.get_dummies() function.

```
# Creating a dummy variable for the variable 'Contract' and dropping the first one.

cont = pd.get_dummies(telecom['Contract'],prefix='Contract',drop_first=True)

#Adding the results to the master dataframe

telecom = pd.concat([telecom,cont],axis=1)
```

You then split the data set into training and test data sets. This step is very subjective and completely based on the business call, or you can say it is dependent on the industry from which you get the data set.

You also know that the industry problem statements are not so simple and small. The data set is large, containing a large number of variables and observations with both categorical and quantitative variables. So, you need to do some exploratory data analysis to get information about some of the variables in the data set.

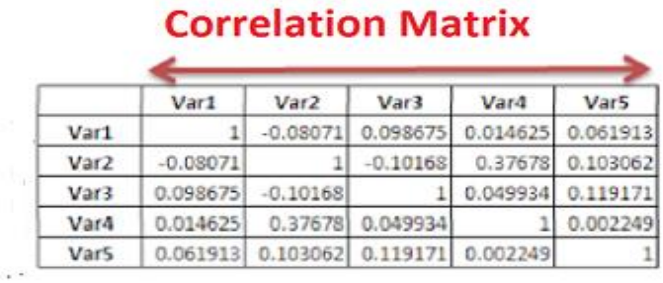
The next step was to create important derived metrics that could help explain the outcome better. In the housing case, you created area per bedroom and bathroom per bedroom as important metrics.

R-squared vs Adjusted R-squared

You then built a model containing all variables and saw the summary of the results. You learnt that, in multiple variable regression**, adjusted R-squared is a better metric than R-squared to assess how good the model fits the data**. R-squared always increases if additional variables are added into the model, even if they are not related to the dependent variable. R-squared thus is not a reliable metric for model accuracy. Adjusted R-squared, on the other hand, penalises R-squared for unnecessary addition of variables. So, if the variable added does not increase the accuracy adequately, adjusted R-squared decreases although R-squared might increase.

Multicollinearity

It may be that some variables could have some relation amongst themselves; in other word, the variables may be highly collinear to each other. A simple way to detect collinearity is to look at the correlation matrix of the independent variables as shown.



Correlation Matrix

	Var1	Var2	Var3	Var4	Var5
Var1	1	-0.08071	0.098675	0.014625	0.061913
Var2	-0.08071	1	-0.10168	0.37678	0.103062
Var3	0.098675	-0.10168	1	0.049934	0.119171
Var4	0.014625	0.37678	0.049934	1	0.002249
Var5	0.061913	0.103062	0.119171	0.002249	1

Figure 14: Correlation Matrix

A large value in this matrix would indicate a pair of highly correlated variables. Unfortunately, not all collinearity problems can be **detected by the inspection of the correlation matrix.** It is possible for collinearity to exist **between three or more variables** even if no pair of variables has a high correlation. This situation is called multicollinearity.

A better way to assess multicollinearity is to compute the **variance inflation factor** (VIF).

Since one of the major goals of linear regression is identifying the important explanatory variables, it is important to assess the impact of each and then keep those which have a significant impact on the outcome. This is the major issue with multicollinearity. Multicollinearity makes it difficult to assess the effect of individual predictors. 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 after checking for p-values, 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 $p<0.05$, 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. You took VIF = 2 as the threshold, but in real business scenarios, it will depend on the case requirements.

Model Building

A very crucial step for model development process is variable selection for the model. It is not a good call to consider all variables in the model because a variable may or may not impact the results of the model. Thus, you have to remove variables based on multicollinearity (VIF) and p-values. Ideally, the model should have a limited number of variables which explain the outcome well.

So, you followed this algorithm to get to a desired model:

1. Build a model containing all variables
2. Check VIF and summary
3. Remove variables with high VIF (>2 generally) and which are insignificant ($p>0.05$), one by one
4. If the model has variables which have a high VIF and are significant, check and remove other insignificant variables
5. After removing the insignificant variables, the VIFs should decline
6. If some variables still have a high VIF, remove the variable which is relatively less significant
7. Now, variables must be significant. If the number of variables is still high, remove them in order of insignificance until you arrive at a limited number of variables that explain the model well.

You get a model that is trained on the training data set. This model should be able to accurately predict the house prices in the test data.

Model Validation

Thus, it is desired that the R-squared between the predicted value and the actual value in the test set should be high. In general, it is desired that the R-squared on the test data be high, and as similar to the R-squared on the training set as possible.

You obtained an R-squared value of 0.639 in the test data set. Thus, you could say that the predicted values from the model are able to explain 63.9% variation in the actual outcomes. This is a fairly decent model.

You should note that R-squared is only one of the metrics to assess accuracy in a linear regression model. There are many other metrics.

Variable Selection Methods

In housing case study you saw the use of RFE. Recursive feature elimination is based on the idea of repeatedly constructing a model (for example an SVM or a regression model) and then choosing either the best or the worst performing feature (for example based on coefficients), setting that feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to what they were eliminated. As such, it is a greedy optimization for finding the best performing subset of features.

The stability of RFE depends heavily on the type of the model that is used for feature ranking at each iteration. Similar to a non-regularized regression, RFE too can be unstable while utilizing it, whereas using ridge regression can provide more stable results.

RFE implementation in housing case study

```
# Running RFE with the output number of the variable equal to 11

lm = LinearRegression()

rfe = RFE(lm, 11)

rfe = rfe.fit(X_train, y_train)

print(rfe.support_)

print(rfe.ranking_)
```

Multiple Linear Regression

Let's revise some of the concepts you learnt in the previous two sessions:

1. In statistical modelling, **linear regression is a process of estimating the relationship among variables**. The focus here is to establish the relationship between a dependent variable and one or more independent variable(s). Independent variables are also called 'predictors'.
2. Regression helps you understand how the values of dependent variable changes as you change the values of 1 predictor, holding the other predictors static (or same). This means that simple linear regression in its most basic form **doesn't allow you to change all the predictors at a time and measure the impact on the dependent variable. You can only change 1 at a time.**
3. Regression only shows relationship, i.e. correlation and NOT causality. In a very restrictive environment, regression may show causality. However, if you blindly interpret regression results as causation, it may lead to false insights. **Remember: Correlation does not imply causation.**
4. **Regression analysis is widely used for 2 purposes: a) Forecasting and b) Prediction.** The uses of forecasting and prediction have substantial overlap. However, they are different and it's important to understand why, to be able to use regression effectively for each purpose. **Regression guarantees 'interpolation' but not necessarily 'extrapolation'.**
5. Linear regression is a form of parametric regression

Let us understand what interpolation and extrapolation mean. **Interpolation** basically means using the model to predict the value of a dependent variable on independent values that lie within the range of data you already have. **Extrapolation**, on the other hand, means predicting the dependent variable on the independent values that lie outside the range of the data the model was built on.

To understand this better, look at the diagram below. The model is built on values of x between a and b.

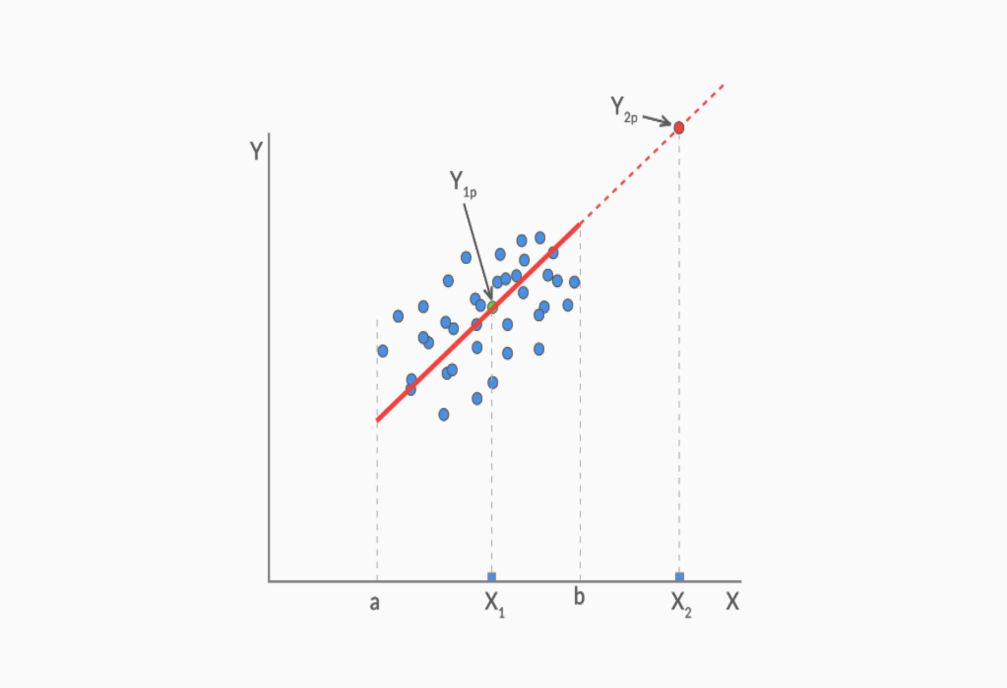


Figure 15: Interpolation vs Extrapolation

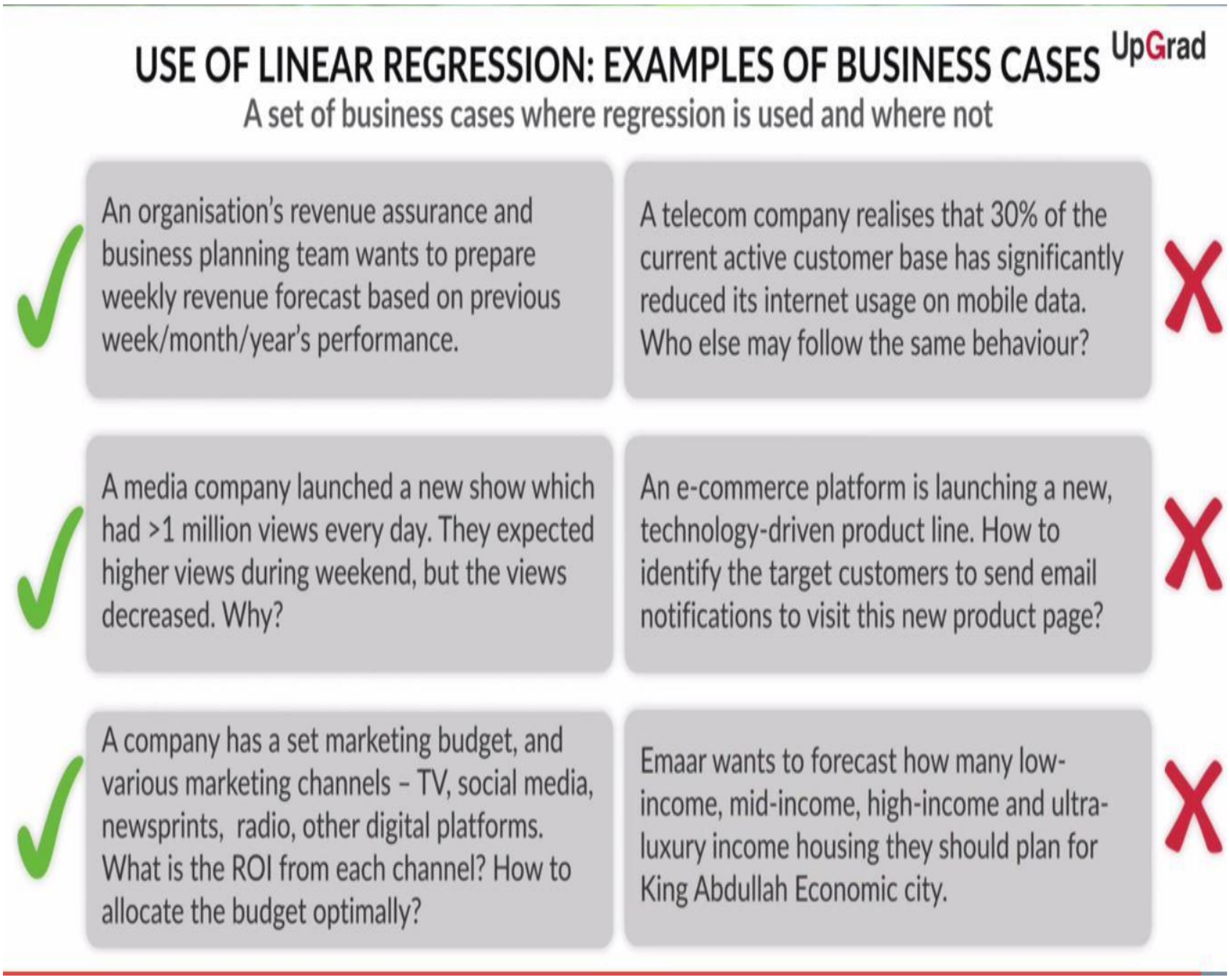
When you wish to predict the Y for X1, which lies between a and b, it is called interpolation.

On the other hand, extrapolation would be extending the line to predict Y for X2 which lies outside the range on which the linear model was trained. For now, you only need to understand what these terms mean. You should note that linear regression can be effectively used for interpolation, but extrapolation may not necessarily be effective/accurate.

Our SME Ujjyaini also mentioned that linear regression is a parametric model, as opposed to non-parametric ones. A detailed discussion on parametric and non-parametric models is beyond the discussion of this module, though a simple explanation is given below.

In simple terms, a parametric model can be described using a finite number of parameters. E.g. a linear regression model built using n independent variables will have exactly n 'parameters' (i.e. n coefficients). The entire model can be described using these n parameters.

You saw a few cases where linear regression is applicable and where it isn't.



USE OF LINEAR REGRESSION: EXAMPLES OF BUSINESS CASES

UpGrad

A set of business cases where regression is used and where not

✓ An organisation's revenue assurance and business planning team wants to prepare weekly revenue forecast based on previous week/month/year's performance.

A telecom company realises that 30% of the current active customer base has significantly reduced its internet usage on mobile data. Who else may follow the same behaviour? ✗

✓ A media company launched a new show which had >1 million views every day. They expected higher views during weekend, but the views decreased. Why?

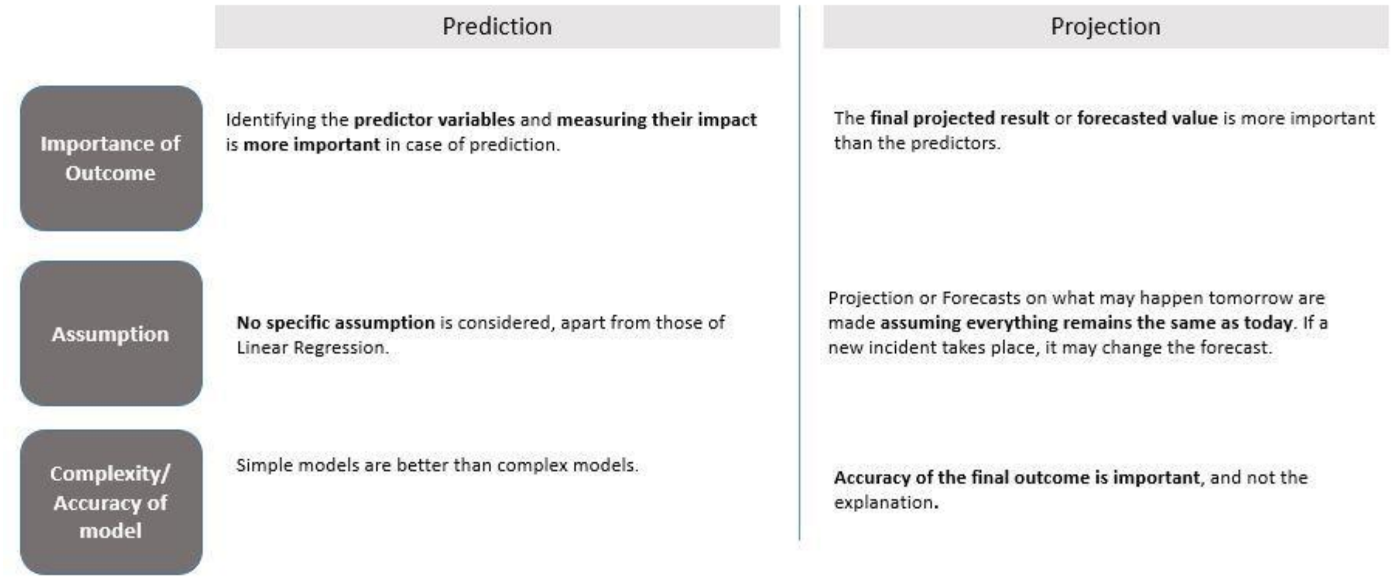
An e-commerce platform is launching a new, technology-driven product line. How to identify the target customers to send email notifications to visit this new product page? ✗

✓ A company has a set marketing budget, and various marketing channels – TV, social media, newsprints, radio, other digital platforms. What is the ROI from each channel? How to allocate the budget optimally?

Emaar wants to forecast how many low-income, mid-income, high-income and ultra-luxury income housing they should plan for King Abdullah Economic city. ✗

Figure 16: Using linear regression

Although prediction and projection sound synonymous with each other, they are different applications in analytics. Some of the differences between the two are:



	Prediction	Projection
Importance of Outcome	Identifying the **predictor variables** and **measuring their impact** is **more important** in case of prediction.	The **final projected result** or **forecasted value** is more important than the predictors.
Assumption	**No specific assumption** is considered, apart from those of Linear Regression.	Projection or Forecasts on what may happen tomorrow are made **assuming everything remains the same as today**. If a new incident takes place, it may change the forecast.
Complexity/ Accuracy of model	Simple models are better than complex models.	**Accuracy of the final outcome is important**, and not the explanation.

Figure 17: Prediction vs Projection

Case Study: Media Company

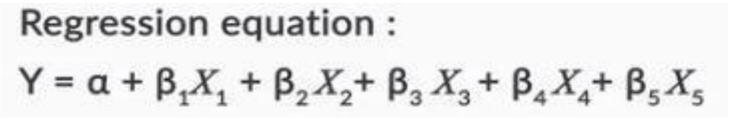
Let us now see the case of a media company.

The problem statement was: A digital media company (similar to Voot, Hotstar, Netflix, etc.) had launched a show. Initially, the show got a good response, but then witnessed a decline in viewership. The company wants to figure out what went wrong.

The potential reasons could be:

1. Decline in the number of people coming to the platform
2. Fewer people watching the video
3. A Decrease in marketing spend?
4. Competitive shows, e.g. cricket/ IPL
5. Special holidays
6. Twist in the story

Consider a multiple linear equation:



Regression equation :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Figure 18: Regression Equation

The outcome (Y) may decline when any of the following happens:

1. Values of 1 or more predictors whose coefficients are positive have declined
2. Values of 1 or more predictors whose coefficients are negative have increased

The important variables in the viewership data were — visitors to the digital platform, views to all the shows in the platform, marketing impressions, presence/absence of a character A, cricket match.

The pattern in show viewership emerged as follows:

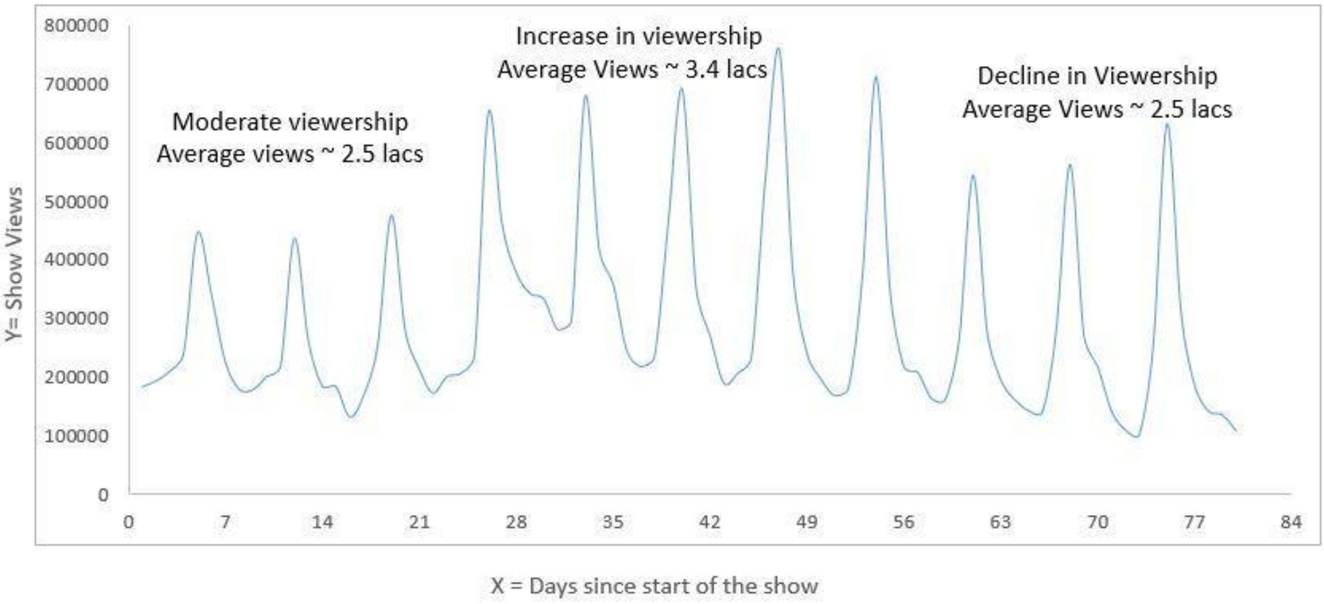


Figure 19: Show Views vs Days

Important parameters such as Ad impressions were plotted along the show views to check if trends were similar.

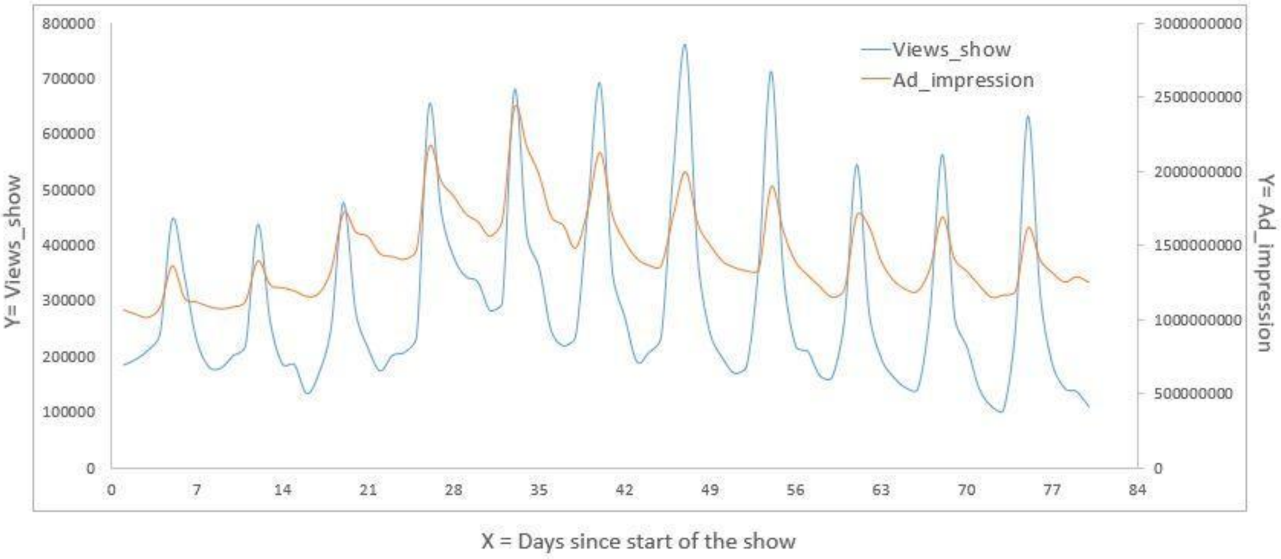


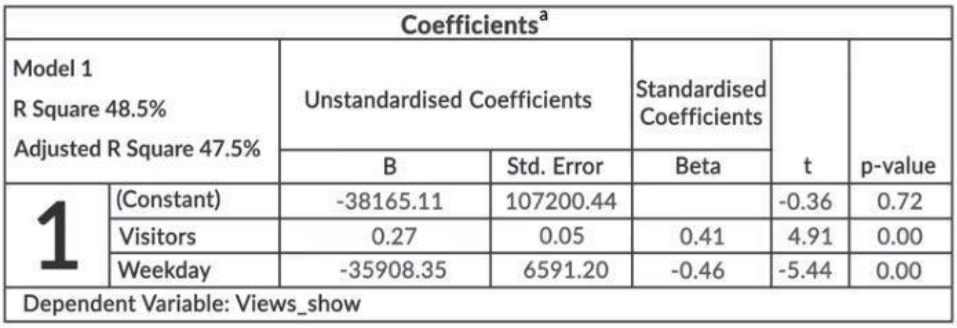
Figure 20: Show Views and Ad impressions vs Days

Ad impression and the Views to the show had similar spikes during weekend but Ad impressions are significantly higher in scale compared to the views.

Model Building

You followed the following steps:

1. You started the model with Visitors and weekdays as the independent variables.



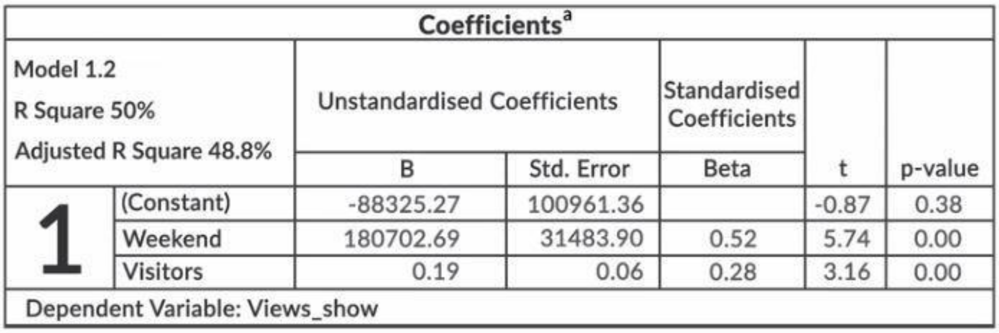
Coefficients[a]

Model 1 R Square 48.5% Adjusted R Square 47.5%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
1	(Constant)	-38165.11	107200.44		-0.36	0.72
	Visitors	0.27	0.05	0.41	4.91	0.00
	Weekday	-35908.35	6591.20	-0.46	-5.44	0.00

Dependent Variable: Views_show

Figure 21: Model 1.1

2. Weekend seemed to be a better variable to explain the spikes on weekends, thus you replaced weekdays with the “weekends” variable, which took value `1 on Saturday and Sunday and 0 otherwise. Thus, the model was finally built with “Visitors” and “Weekend” as independent variables



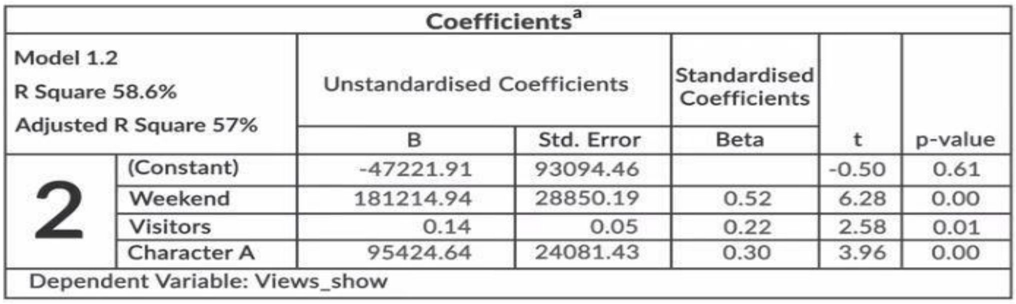
Coefficients[a]

Model 1.2 R Square 50% Adjusted R Square 48.8%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
1	(Constant)	-88325.27	100961.36		-0.87	0.38
	Weekend	180702.69	31483.90	0.52	5.74	0.00
	Visitors	0.19	0.06	0.28	3.16	0.00

Dependent Variable: Views_show

Figure 22: Model 1.2

3. Character A was added to the model. Adjusted R-squared increased and all variables were significant.



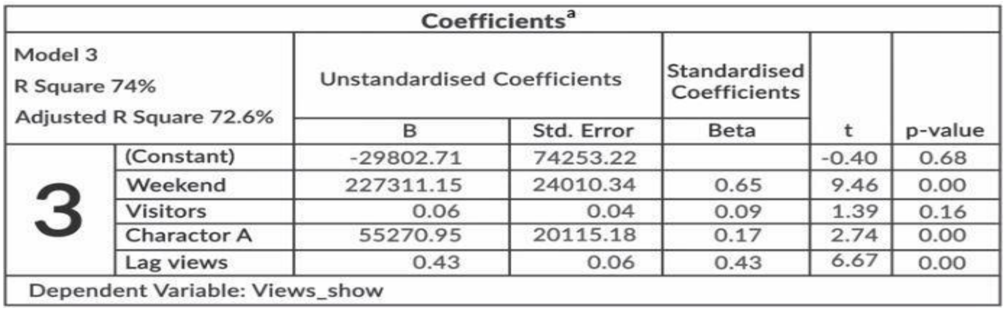
Coefficients[a]

Model 1.2 R Square 58.6% Adjusted R Square 57%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
2	(Constant)	-47221.91	93094.46		-0.50	0.61
	Weekend	181214.94	28850.19	0.52	6.28	0.00
	Visitors	0.14	0.05	0.22	2.58	0.01
	Character A	95424.64	24081.43	0.30	3.96	0.00

Dependent Variable: Views_show

Figure 23: Model 2

4. Lag views of yesterday was introduced. Visitors became insignificant as the p-value became 0.16



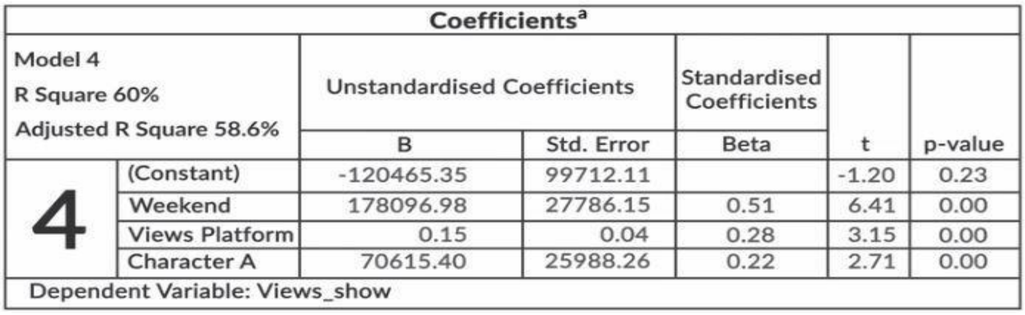
Coefficients[a]

Model 3 R Square 74% Adjusted R Square 72.6%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
3	(Constant)	-29802.71	74253.22		-0.40	0.68
	Weekend	227311.15	24010.34	0.65	9.46	0.00
	Visitors	0.06	0.04	0.09	1.39	0.16
	Charactor A	55270.95	20115.18	0.17	2.74	0.00
	Lag views	0.43	0.06	0.43	6.67	0.00

Dependent Variable: Views_show

Figure 24: Model 3

5. Visitors was removed because it was no longer significant. Total views on the platform was introduced as a variable instead of Lag views of yesterday. Adjusted R-squared decreased but the model made more sense now.



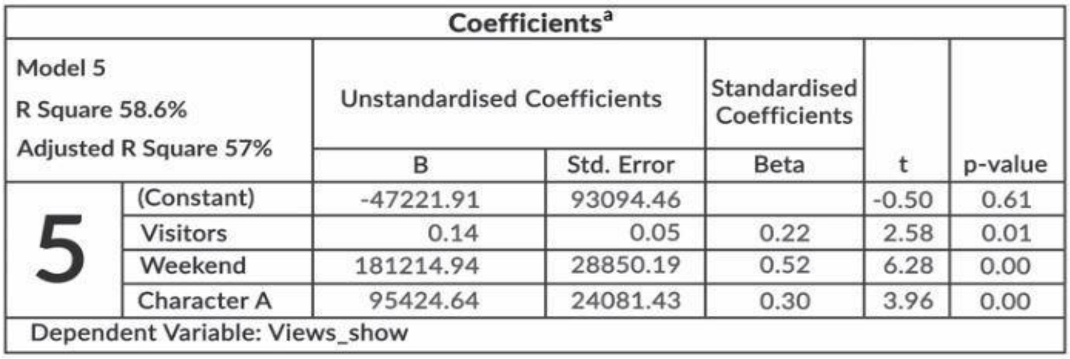
Coefficients[a]

Model 4 R Square 60% Adjusted R Square 58.6%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
4	(Constant)	-120465.35	99712.11		-1.20	0.23
	Weekend	178096.98	27786.15	0.51	6.41	0.00
	Views Platform	0.15	0.04	0.28	3.15	0.00
	Character A	70615.40	25988.26	0.22	2.71	0.00

Dependent Variable: Views_show

Figure 25: Model 4

6. Visitors was brought back as a variable as it can be driven by marketing action, unlike platform views. It is easier to bring customers on the platform compared to forcing them to watch something. There is a slight drop in adjusted R-squared, but marketing team requires an actionable variable, for which "visitors" fits well.



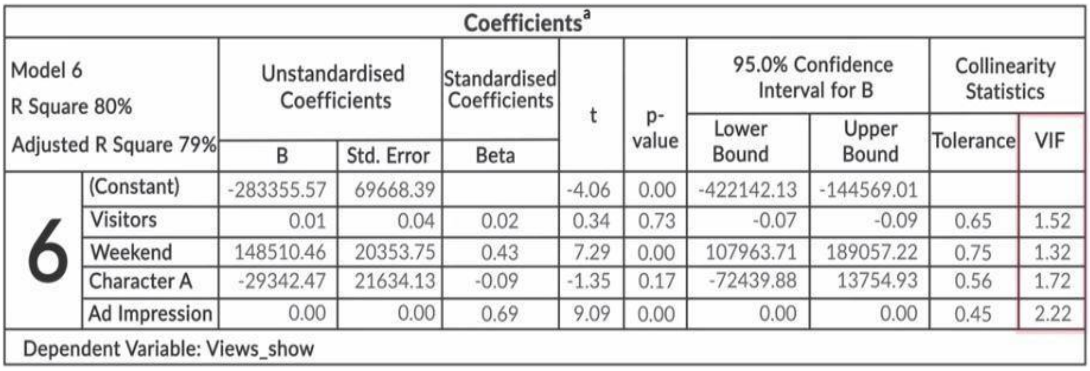
Coefficients[a]

Model 5 R Square 58.6% Adjusted R Square 57%		Unstandardised Coefficients		Standardised Coefficients		
		B	Std. Error	Beta	t	p-value
5	(Constant)	-47221.91	93094.46		-0.50	0.61
	Visitors	0.14	0.05	0.22	2.58	0.01
	Weekend	181214.94	28850.19	0.52	6.28	0.00
	Character A	95424.64	24081.43	0.30	3.96	0.00

Dependent Variable: Views_show

Figure 26: Model 5

7. The Ad impressions variable was introduced. Adjusted R-squared increased to 79%. This is a significant increase. But variables “visitors” and “Character A” became insignificant. Character A also changed signs. Earlier models showed that character A drives views when present. This model, on the other hand, said that viewership was higher when Character A is absent.



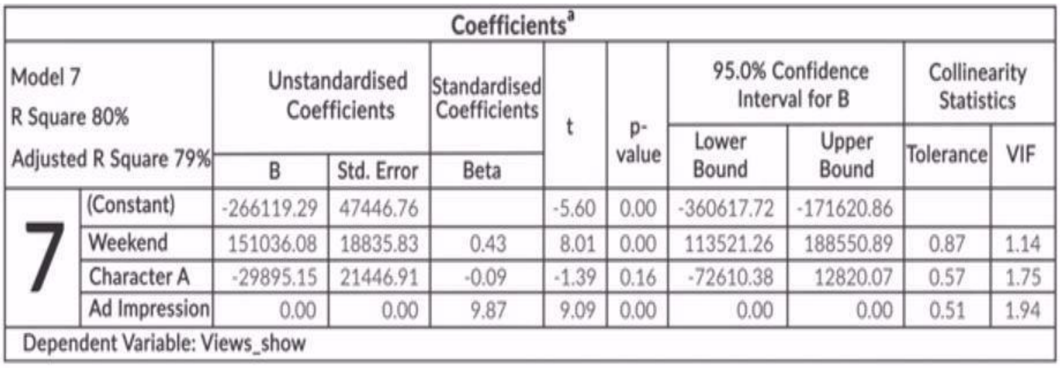
Coefficients^a

Model 6 R Square 80% Adjusted R Square 79%		Unstandardised Coefficients		Standardised Coefficients	t	p-value	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
6	(Constant)	-283355.57	69668.39		-4.06	0.00	-422142.13	-144569.01		
	Visitors	0.01	0.04	0.02	0.34	0.73	-0.07	-0.09	0.65	1.52
	Weekend	148510.46	20353.75	0.43	7.29	0.00	107963.71	189057.22	0.75	1.32
	Character A	-29342.47	21634.13	-0.09	-1.35	0.17	-72439.88	13754.93	0.56	1.72
	Ad Impression	0.00	0.00	0.69	9.09	0.00	0.00	0.00	0.45	2.22

Dependent Variable: Views_show

Figure 27: Model 6

8. Visitors variable was removed on the basis of insignificance.



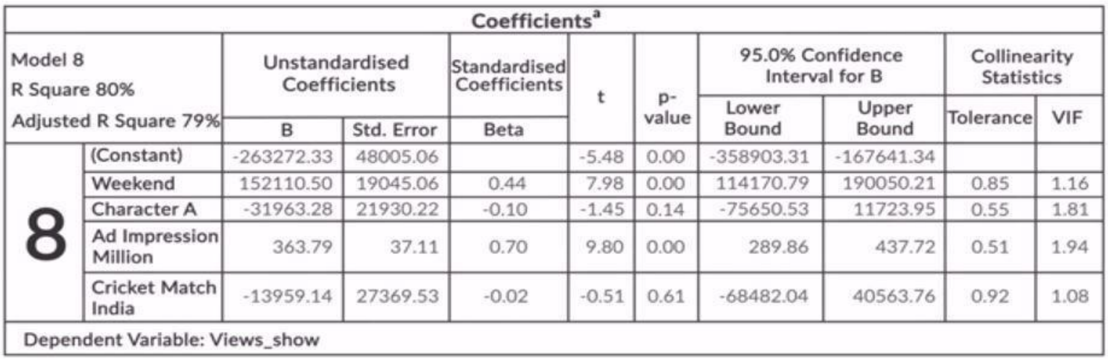
Coefficients^a

Model 7 R Square 80% Adjusted R Square 79%		Unstandardised Coefficients		Standardised Coefficients	t	p-value	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
7	(Constant)	-266119.29	47446.76		-5.60	0.00	-360617.72	-171620.86		
	Weekend	151036.08	18835.83	0.43	8.01	0.00	113521.26	188550.89	0.87	1.14
	Character A	-29895.15	21446.91	-0.09	-1.39	0.16	-72610.38	12820.07	0.57	1.75
	Ad Impression	0.00	0.00	9.87	9.09	0.00	0.00	0.00	0.51	1.94

Dependent Variable: Views_show

Figure 28: Model 7

9. You checked if cricket matches featuring India are one of the reasons of viewership decline. Adjusted R-squared was similar. But since the cricket matches variable has a very high p-value, it is insignificant.



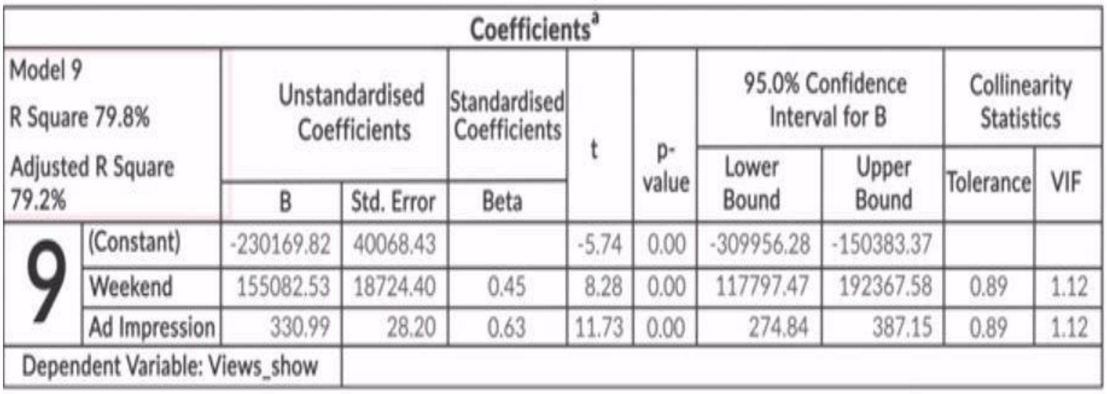
Coefficients^a

Model 8 R Square 80% Adjusted R Square 79%		Unstandardised Coefficients		Standardised Coefficients	t	p-value	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
8	(Constant)	-263272.33	48005.06		-5.48	0.00	-358903.31	-167641.34		
	Weekend	152110.50	19045.06	0.44	7.98	0.00	114170.79	190050.21	0.85	1.16
	Character A	-31963.28	21930.22	-0.10	-1.45	0.14	-75650.53	11723.95	0.55	1.81
	Ad Impression Million	363.79	37.11	0.70	9.80	0.00	289.86	437.72	0.51	1.94
	Cricket Match India	-13959.14	27369.53	-0.02	-0.51	0.61	-68482.04	40563.76	0.92	1.08

Dependent Variable: Views_show

Figure 29: Model 8

10. You removed insignificant variables from the model. Character A was removed as well because its sign (-ve) was counterintuitive to business understanding. The final model looked like this:



Coefficients^a

Model 9 R Square 79.8% Adjusted R Square 79.2%		Unstandardised Coefficients		Standardised Coefficients	t	p-value	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
9	(Constant)	-230169.82	40068.43		-5.74	0.00	-309956.28	-150383.37		
	Weekend	155082.53	18724.40	0.45	8.28	0.00	117797.47	192367.58	0.89	1.12
	Ad Impression	330.99	28.20	0.63	11.73	0.00	274.84	387.15	0.89	1.12
Dependent Variable: Views_show										

Figure 29. Model 9

11. Ad Impressions and weekend were the most important variables according to the model. Although Ad impressions can still be controlled through marketing budget, the marketing team cannot do anything about the weekend variable. It is beyond the scope of marketing action.

After this, you learnt how to assess a model using the model you obtained above.

Assessing the Model

If a model with a fairly high adjusted R-squared is obtained, it might seem that the task is done. But one or more important explanatory variables could still be missing. Thus, you need to assess the model.

Let's first see the graph between the predicted and actual views in model 9.

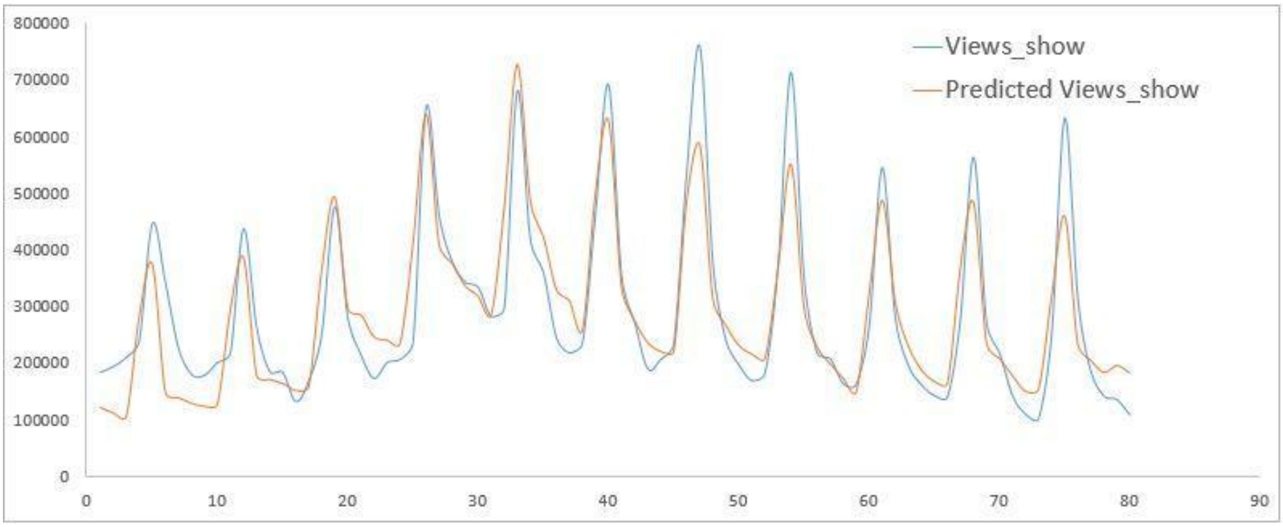


Figure 130: Actual vs Predicted views

You finally have a model that seems good enough to predict why the show viewership fell. The actual and predicted views significantly overlapped, thus indicating that the model is able to explain the change in viewership very well.

Let's see the error term (difference between predicted and actual values) plot.

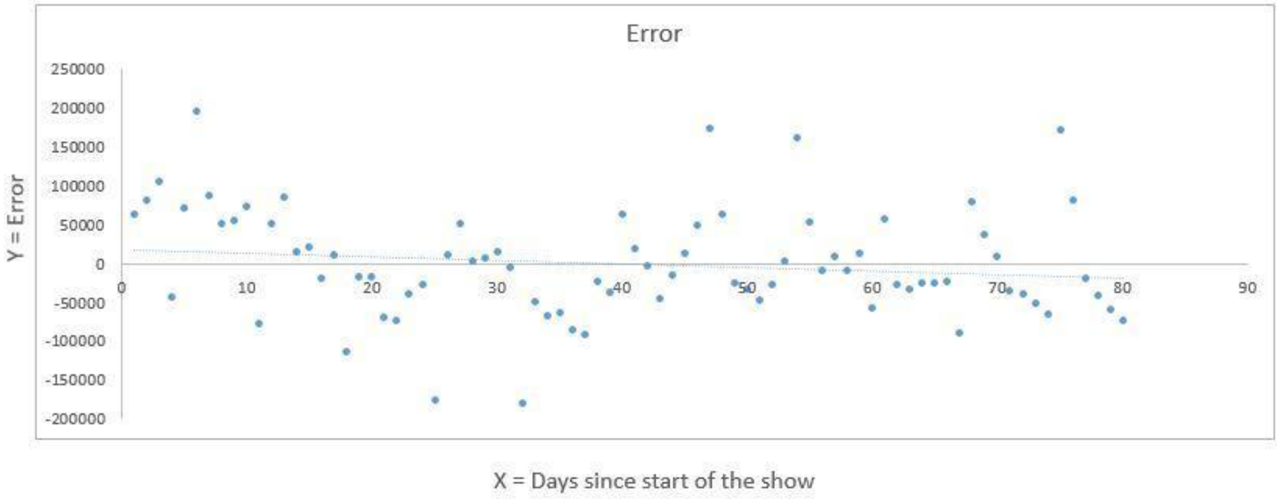


Figure 31: Errors

Observe that the errors (the differences between the actual values and the values predicted by the model) are randomly distributed. What this essentially confirms is that **there are no variables that could have helped explain the model better**. A non-random error pattern, on the other hand, would mean that the errors are capturing some patterns, thus indicating that the model could have been better. A non-random error pattern indicates that there are certain systematic unexplained aspects in the outcomes that are being captured in the error. This pattern in the errors could probably have been explained by some explanatory variable which is missing from the model. So, the idea is that a model should explain everything that is possible, such that only the random errors are left.

Let us check the actual vs predicted values of a previous model — Model 5.

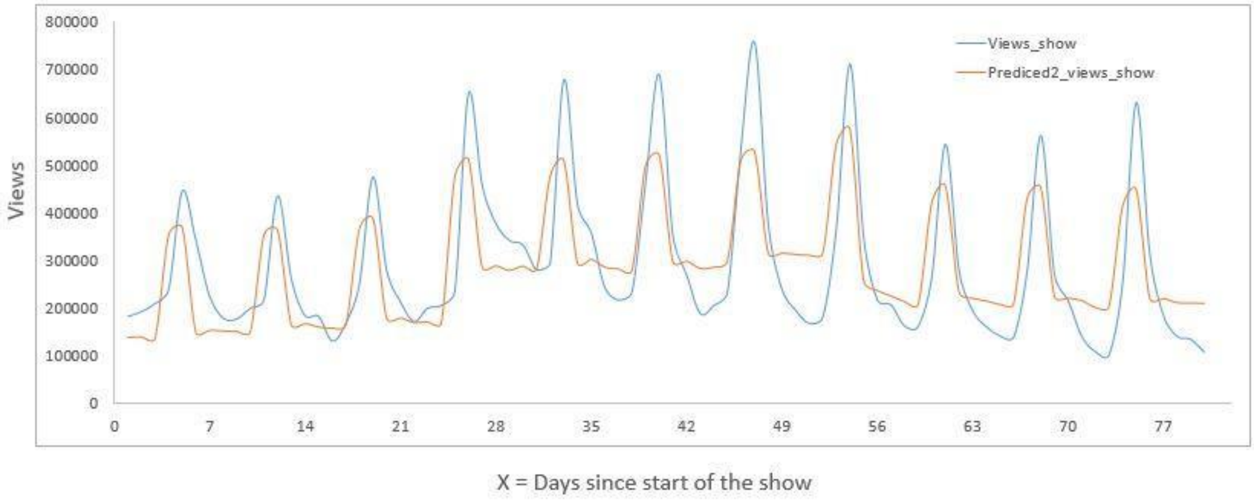


Figure 32: Actual vs Predicted Views (model 5)

You can see that the results are clearly not as good as the model 9. There is less overlap between the actual and predicted values and the predicted values are not able to capture the highs and the lows in the actual values well.

Check the error plot for model 5.

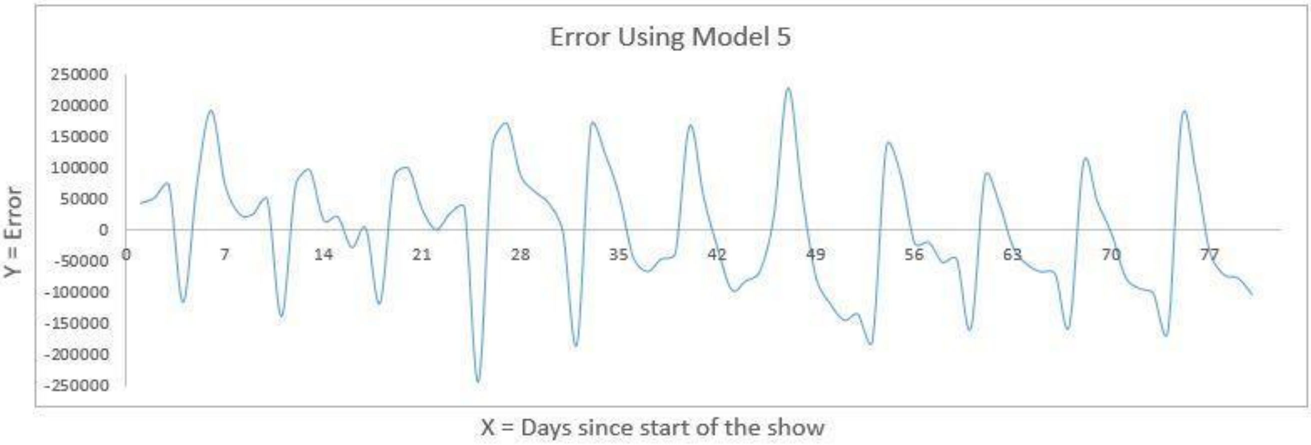


Figure 33: Errors (model 5)

Here you can observe that there is some pattern in the errors and they are not random. This indicates that the model could have been better. Thus, model 9 containing Ad impressions and weekends as the explanatory variables is indeed a better model than model 5.

Using the Final Model to Drive Viewership

The final model contained Ad impressions and weekend as the independent variables. The model had a significant adjusted R-squared. It does not mean that only these two variables can help drive viewership.

Our SME Ujjyaini identified **Ad Impressions** and **Character A** as the driver variables that could explain the viewership pattern. Based on industry experience, **ad impressions are directly proportional to the marketing budget**. Thus, by increasing the marketing budget, a better viewership could be achieved. Similarly, Character A's absence and presence created a significant change in show viewership. Character A's presence brings viewers to the show. Thus, these two variables could be acted upon to improve show viewership.

The marketing team wanted to know the additional investment required to increase the ad impressions and thus, increase the viewership of the show. Model equation was used to find out the increase in ad impressions required to take show viewership back to the peak levels attained before. The ad impressions figure could then be used by the marketing team to calculate the increase in marketing budget required.