Linear Regression

what is predictive modelling?

If somebody asks you to guess how Virat Kohli is going to perform in next cricket match against Australia, you can take a guess *[Assuming you are cricket follower]*. This guess is based on your knowledge of his past performances in these or similar conditions and against this kind of team.

Predictive modelling gives this process a formal framework. It gives you tools to extract mathematical equations/rules from the past data to predict future results.

Steps of Predictive Modelling: 1. After identifying the Business objectives, first step in any predictive model is to collate data from various sources. The sources of data can be historical data, demographic data, behavioural data, Customer data and transactions data etc. 2. In the second step, we need to prepare data into right format for analysis. Here we normally clean data, impute missing values, transform and append variables. 3. Based on the Business Objectives we have to select on or combination of Modelling Techniques like Linear regression for predicting the future Values. 4. Final step is to check the performance of Model like Error, accuracy, ROC and other measures.

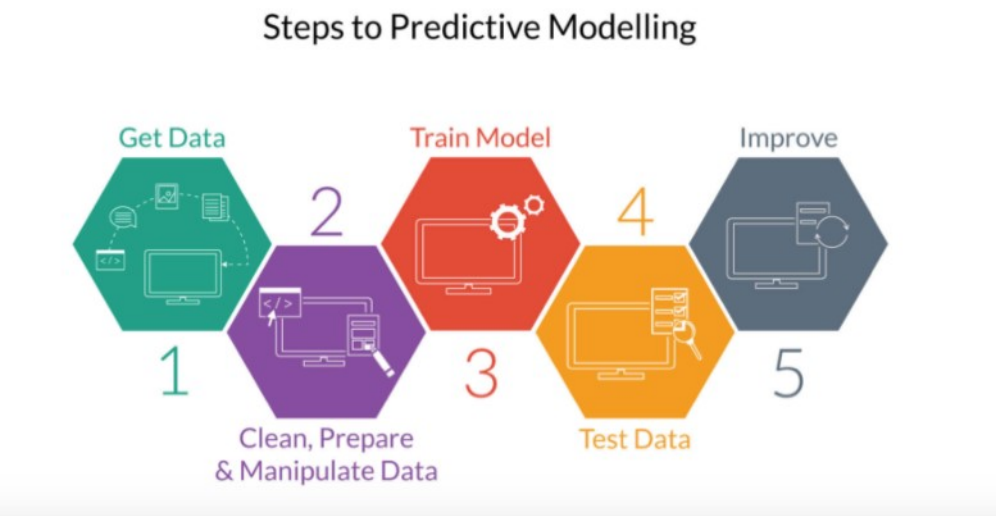


Figure 1: Steps of Predictive Modelling

Before you draw any conclusions about **predictive modeling** being some kind of black magic, lets list down its limitations:

1. The models *[equations/rules]* are dependent on the past data that you have. If data is bad , your predictive models are also going to be bad.
2. Every model will have errors associated with its predictions. Better the model, lesser the error, but it will never be an exact estimate for all practical purposes.
3. Model will be good only until underlying factors on which it was based on , do not change behavior. For example: A predictive model which was built to predict a particular share performance in good economic conditions will perform rather poorly in recession.

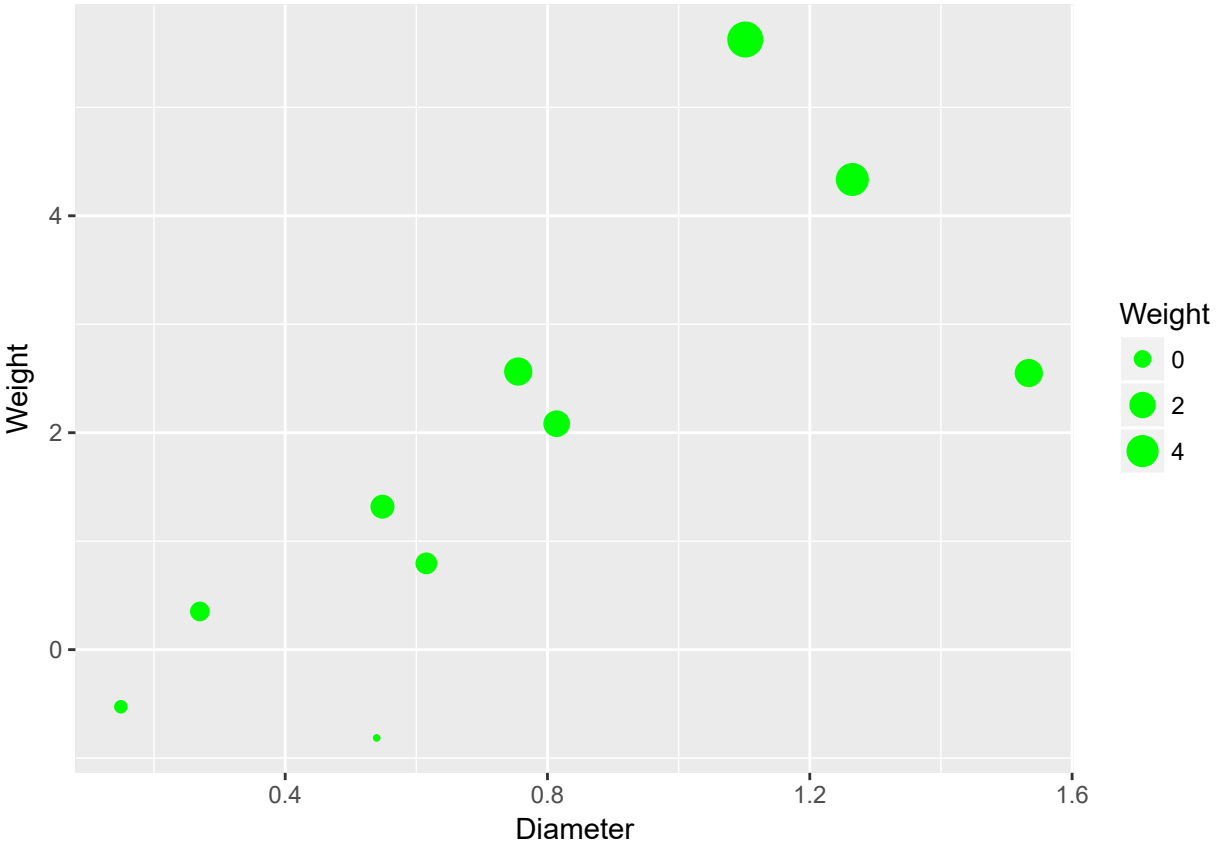
Before we right away jump in to predictive modelling and start extracting those said equations, lets first figure out what really leads us there:

Correlation

When we say that, given past data we can predict future results/outcomes, we are essentially relying on our hunch that we can observe some other factor which affects the outcome and we can leverage that information. For example: when you pick up an apple and guess it weight, you are betting on your assumption that weight of an apple is dependent on its size *[or diameter]*.

In other words , you are assuming that weight of that apple is **correlated** with its diameter.

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```



As you can see in the figure, you were not really wrong. To make this more concrete we need to figure out a way to quantify this *correaltion* . Before doing that, we also need to understand different kind of correlations. As observed in the figure above, weight of apple goes up as its diameter increases. On looking more closely you find out that, increase in weight of apples is happening in *possibly* constant multiples of increase in diameter. This is called **linear correlation**. There can be other form of correlations as well. (Figure 1)

What do we mean by these *linear* and *other* forms of correlations is that one variable *[lets say y]* can be written as a function of another *[lets say x]*

- Linear Correlation : y = ax + b
- Exponential Correlation : y= a*exp(x) + b

quantifying correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

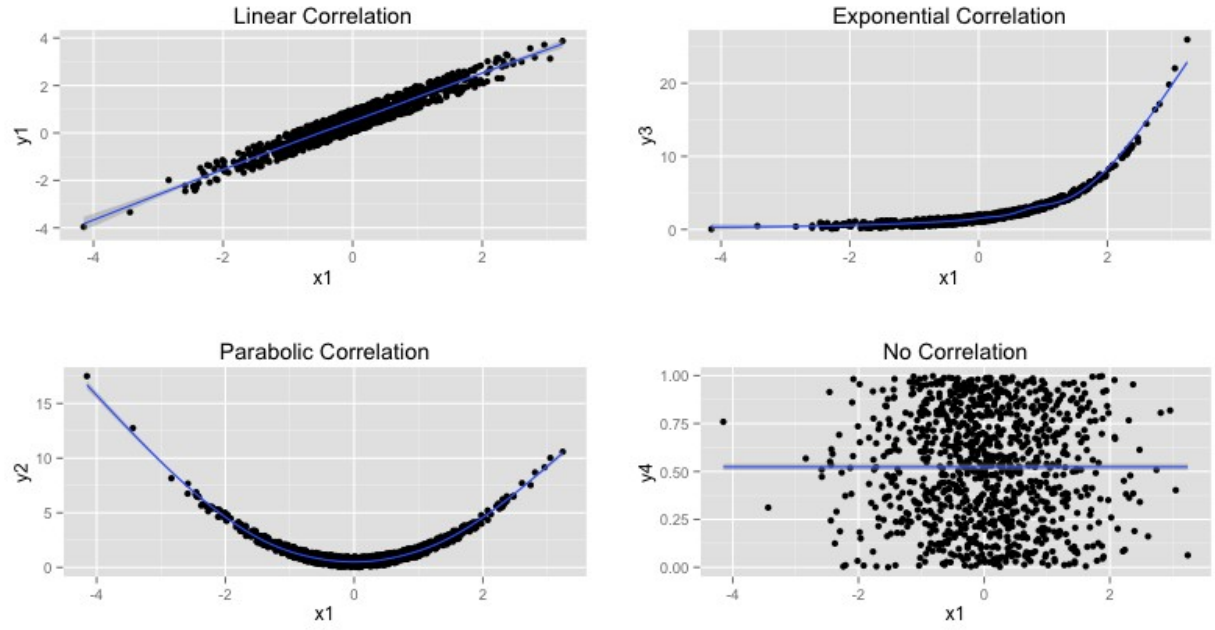


Figure 2: Linear and Non-Linear Correlation

This formula is designed to measure strength of **linear** correlation. it has following properties:

- -1 < r < 1
- It takes -ve values of negative correlation and +ve values of +ve correlation *Note: -ve correlation between x and y means , when x increases y decreases and vice versa*
- Correlation is strong when absolute value of r is close to 1, it is weak if the absolute value is close to zero
- Value of r doesn't change if you linearly transform any or both of the variables. Meaning, correlation between x and y will be same as correlation between (ax+b) and (Ay+C).
- As emphasized above , it can be used to measure linear correlation only

But this formula being capable of measuring only linear correlation is not limited. look at this equation again:

- y= a*e^x + b

consider that x' = e^x . You can write the above as this :

- y= a*x' + b

which is same as saying that y and x' are linearly correlated. In a similar manner, if you observe that y and x are non-linearly related , you can apply a suitable transformation to either of x and y and make the relationship linear.You can then measure the strength of correlation between y and x' *[transformed variable]*

Correlation and Causation (Figure 2)

Causation is when a particular factor is the reason for change in another factor. For example number of people buying sun-screen in the city and city's temperature are going to be correlated. Also there is direct causation. Temperatures going up *[Hot Sunny Weather]* is driving sales of sun-screen. However if two factors are correlated , that doesn't guarantee that there will be causation. For example look at following figure :

Figure 3: Correlation and Causation

Clearly you can see that ice-cream sales and shark attacks are correlated as far as the data can be seen. However that doesn't mean that ice-cream sales are cause of shark attacks. In fact rising temperature cause more people to buy ice-cream and also it causes people to go to beaches in larger numbers *[and sometimes subsequently get attacked by sharks]* . This clarifies two things:

- Clearly correlation doesn't necessarily means causation
- However it does indicate that correlated factors might have a common underlying cause *[Again, not always neccessary]*

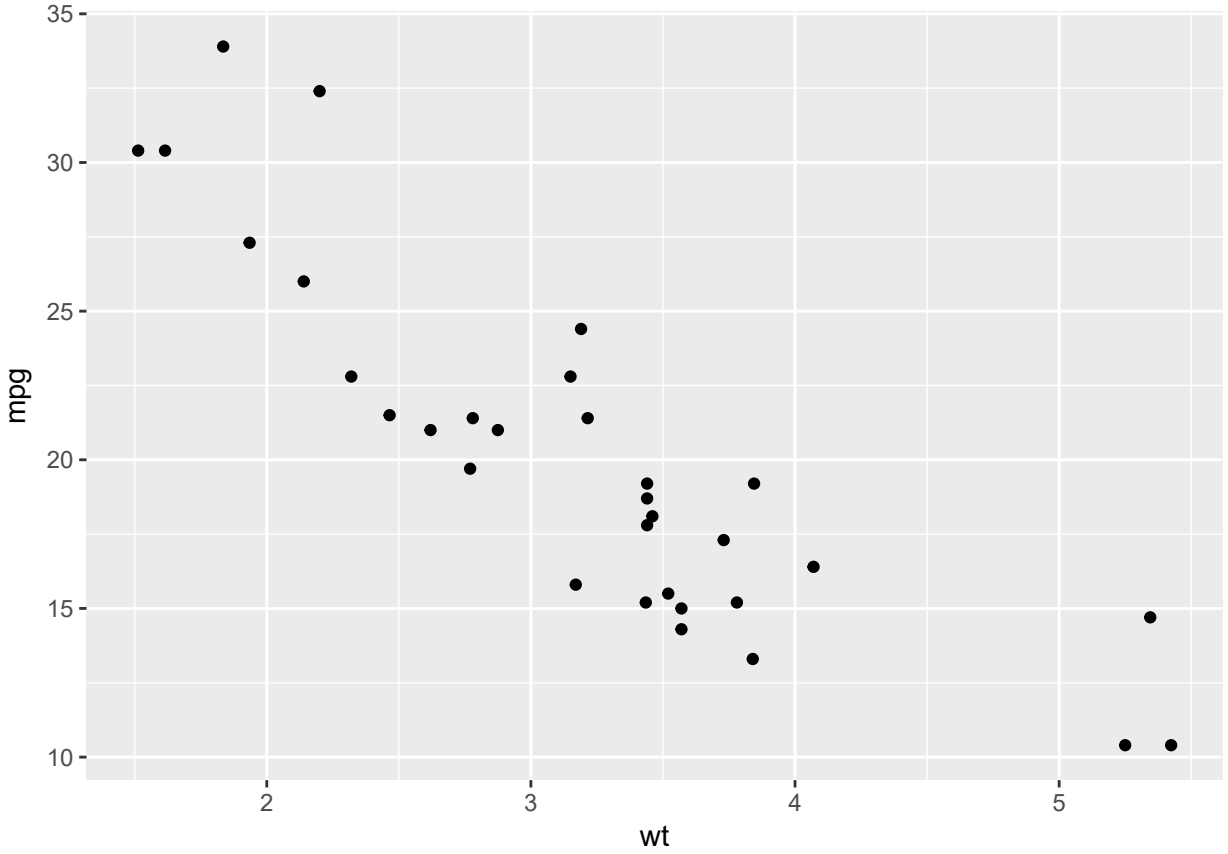
Finding Correlation in R : Correlation Coefficient and Scatter Plots

```
cor.test(mtcars$mpg,mtcars$wt)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$mpg and mtcars$wt
## t = -9.559, df = 30, p-value = 1.294e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9338264 -0.7440872
## sample estimates:
##        cor
## -0.8676594
```

This tells you that correlation exists and it is significant. Also the sign of correlation coefficient is -ve. Which tells you that as weight of a vehicle increases , mileage goes down. Lets look at that visuall with the help of a scatter plot.

```
ggplot(mtcars,aes(x=wt,y=mpg))+geom_point()
```



This tells you the same story, although without numbers.

Simple Linear Regression

Well, correlation coefficient lets us figure whether a pair of variables are affecting each other and if they are then how strong is that effect. However , what we eventually want is an equation which can be used to predict value of y [my response/target/outcome], given a value of x [my input/predictor].

Formally Linear regression attempts to fit a linear relation between a variable of interest (response variable) and a set of predictor variables that may be related to the variable of interest.

As we have seen earlier , if two variables are linearly correlated we can essential draw line through their scatter plot which depicts the relationship between them. But the problem is we can draw many lines, and until now have no clue as to which one to chose finally. (Figure 3)

However , you can observe few crucial things here :

- It's impossible to come up with a line which passes through all the points, in other words ; whatever line equation you come up with, there are going to be errors associated with it.
- Take general equation of a line ; $y = \beta_0 + \beta_1 * x$, what is changing between these lines in the figure is the value of the parameters b0 and b1 . We need to find out such values of these parameters for which error is minimum.

In the figure below, red points on the line are your predictions for values of y given some values of x, whereas blue points are the actuals observed for those values of x. (Figure 4)

After looking at the figure above and the crucial observations that we mentioned, you must have realized that actual value of y can be written as addition of predicted and associated error.

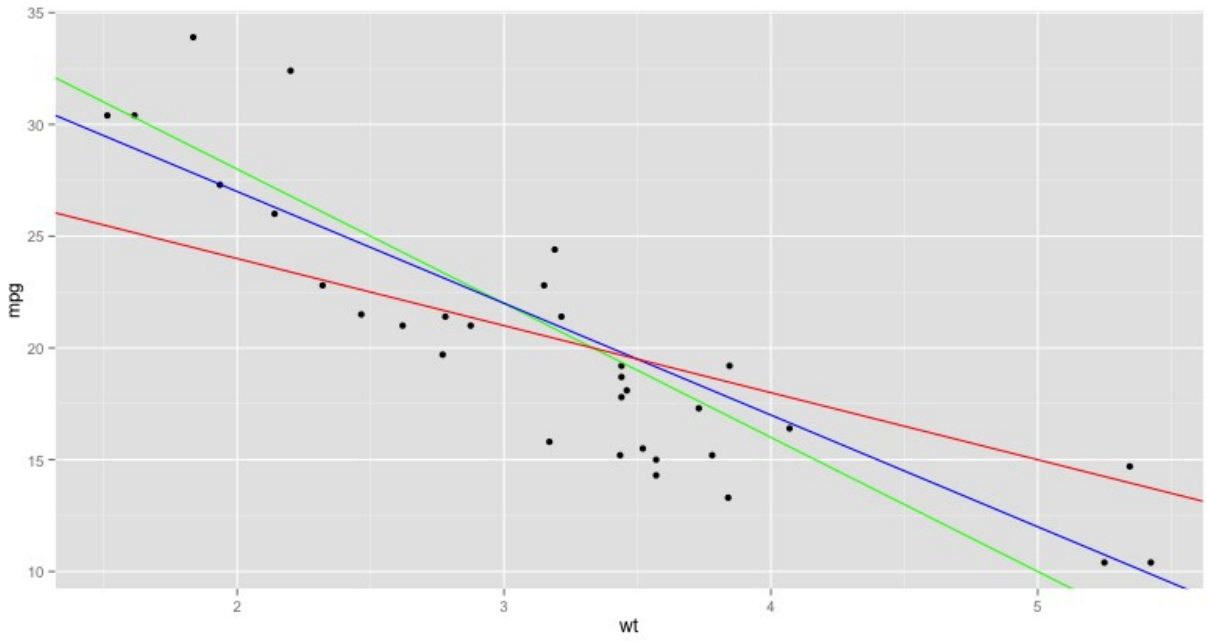


Figure 4: Which Lines to Pick

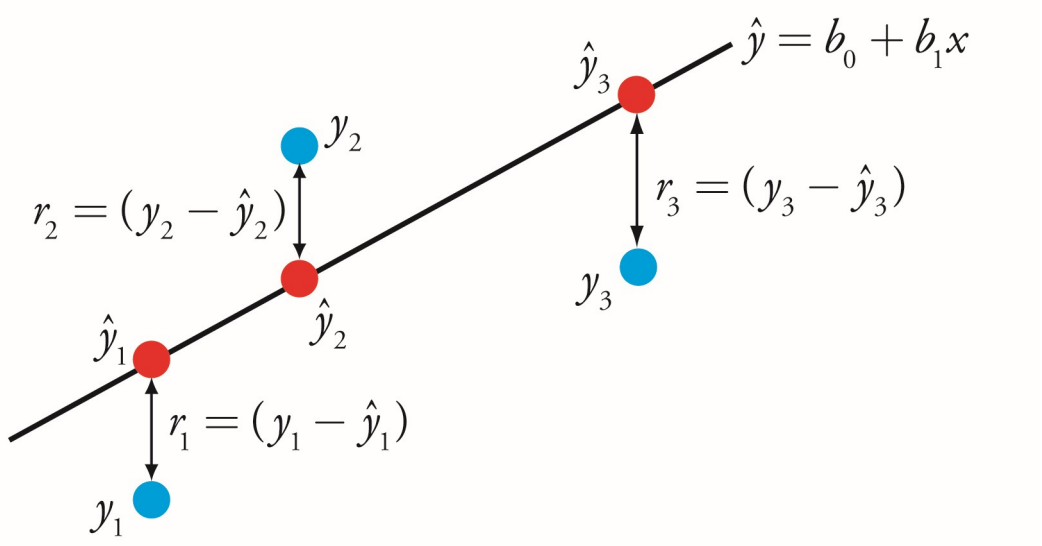


Figure 5: Errors in Predictions

$$y_i = \hat{y}_i + e_i$$

$$y_i = \beta_0 + \beta_1 * x_i + e_i$$

from here we can see that:

$$e_i = y_i - \beta_0 - \beta_1 * x_i$$

What we need to minimize is the collective error for entire data.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 * x_i)^2$$

Now if we want to minimize the above quantity w.r.t. β_0 and β_1 then we can differentiate that equation and put it equal to zero to find values of β_0 and β_1 for which $\sum_{i=1}^{n} e_i^2$ is minimum. Resulting equations are also called normal equations. Here they are:

$$-2 * \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 * x_i) = 0$$

and

$$-2 * \sum_{i=1}^{n} x_i * (y_i - \beta_0 - \beta_1 * x_i) = 0$$

upon solving them you'll get result for β_0 and β_1 which can also be written like this:

$$\beta_1 = r_{xy} * (s_x / s_y)$$

and

$$\beta_0 = \bar{y} - \beta_1 * \bar{x}$$

Don't get intimidated by all these equations, at no stage as a business analyst you'll need to create or solve these equations.Software *[SAS or R or python]* will do this for you. Don't worry and read on.

So lets say your software gave you the appropriate values of β_0 & β_1 and you have your predictive model equation ready. But think, that whatever variable data you pass to this mathematical framework you'll get some values of β_0 & β_1, does that ensure you have a **good** model? Not necessarily .

Imagine , in absence of any such equation, what is your best guess for y . Its $\bar{y}$ the average value. But with this guess there is error associated with each observation : $y_i - \bar{y}$. Writing this in collective form :

SST= Total sum of squares = Total Variability = $\sum_{i=1}^{n} (y_i - \bar{y})^2$

This is also called total variability in target. We intend to bring this down with our model. In other words, we want to explain this variability with our model. Lets try to understand this with this figure:(Figure 6)

Consider that our predicted values are $\hat{y}_i$. As discussed earlier , they don't explain entire variability in the target but a part instead. which is $\hat{y}_i - \bar{y}$.

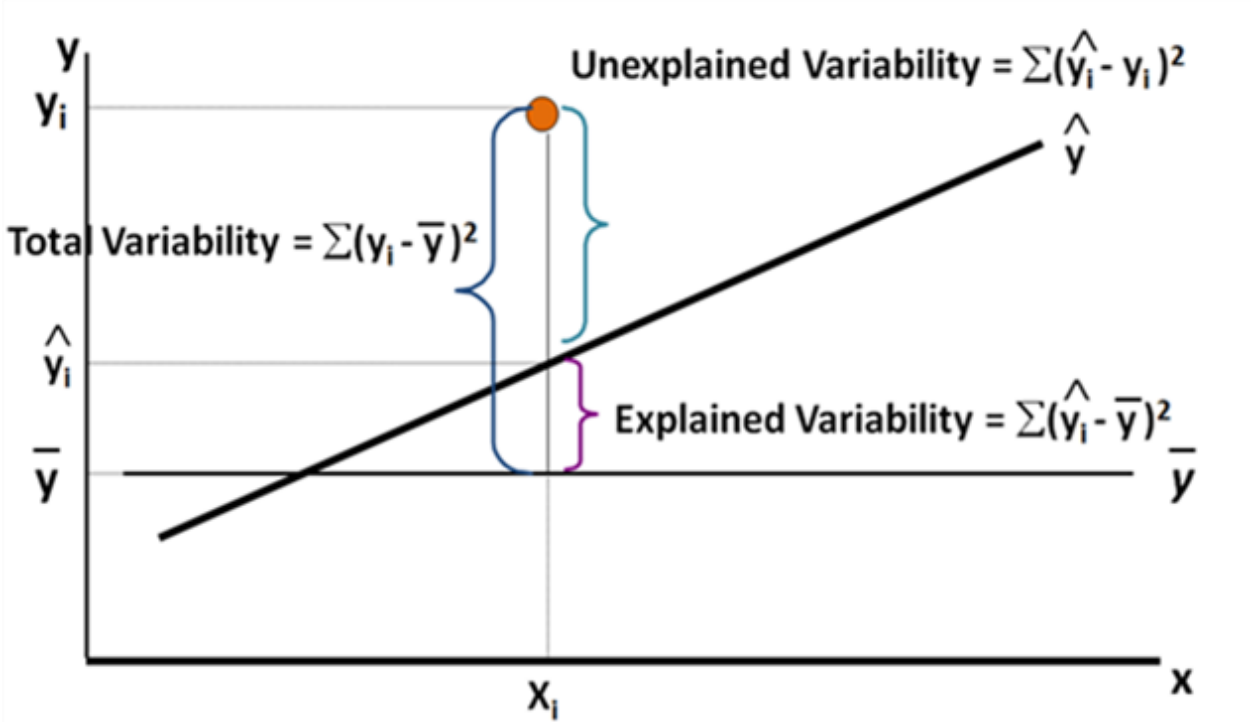


Figure 6: R square

SSR = Regression sum of squares = Explained Variability = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

You never get a perfect model. After prediction, each y_i has still error in prediction being equal to $y_i - \hat{y}_i$.

SSE = Error sum of squares = Unexplained Variability = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Ideally you'd want explained variability to be equal to total variability. That'll be a perfect model. Ratio between explained and total variability is defined as coefficient of determination, it is written as R^2. Its an indicator of how good your model is.

$$R^2 = SSR/SST$$

As you can see R^2's maximum value is 1 when SSR=SST . Also its minimum value would be 0 when SSR=0 . The closer to 1 , R^2's value is, better is your model.

OK, so now we can estimate parameters for our linear model equation, also we can measure how good our model is, but we haven't done anything about errors. Of course there is no way to predict them *[That's why they are errors!]*

What we can do is to make some kind of probabilistic estimate for them.

Interpreting the Estimated parameters of Regression Model

When the predictor variable x_i in our simple linear regression model ($y = \beta_0 + \beta_1 * x$) increases by one unit then the response variable y increases by $beta_1$ units, remaining all other variables are constant.

β_0 represents the intercept term, it simply tells if the x_i value is zero then the y value will be equal to β_0

Assumption of Normality and its consequences

If we assume that $e_i \sim N(0, \sigma_i^2)$, then for each i_{th} Observation we can come up with some confidence interval around our prediction. For example we can say that each i_{th} response would lie in $y_i \pm 3 * \sigma_i$ with 99.7% confidence.

In addition to this quantification of errors , normality assumption allows us to build a very useful hypothesis. Before we discuss that, lets answer a question. We have seen this that if we supply some random y and x variable to this mathematical frame work, we'd get some non-zero estimate of β_0 & β_1 . Does that mean that y *[our response/target/DV]* really depends on x and it follows the equation $y = \beta_0 + \beta_1 * x$?

If y did not depend on x then estimate for β_1 should be zero. But we rarely see that with real data. Estimate might be close to zero but never exactly zero. So when should we conclude that parameter estimate for our β is significantly different from zero?

Now lets look at the hypothesis that the normality assumption enables us to build.

$$H_0 : \beta_i = 0$$
$$test - statistic : \frac{\beta_i}{S_{\beta_i}} \sim t - distribution$$

We can look at the p-values for the test, and if they turn out to be greater than alpha *[standard value 0.05]* , we can conclude that null hypothesis is true and parameter estimate is non-zero by chance and should be discarded.

Assumption of Homoscedasticity

We can estimate error variance σ_i for each x_i from our past data, because we very likely will have multiple observation for x_i. But we are building this model to predict values of y for not yet seen values of x, for such values of x there is no way for us to estimate error variance. We'll have to assume that error variance remains constant across all values of x . This assumption of constant variance across all values of predictor variable *[x]* is called homoscedasticity assumption.

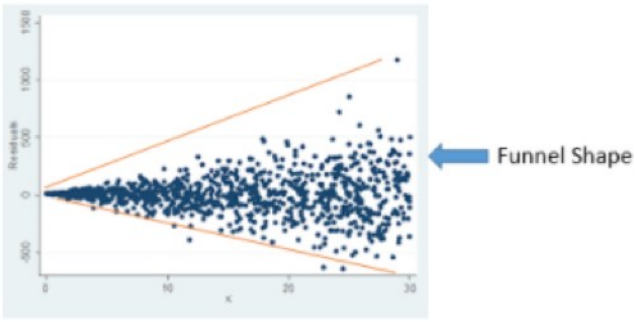


Figure 7: R square

Assumption of error independence

Errors from a model are those part of our response which we could not predict. Whatever part fell in a pattern became part of our model. Errors on the other hand are random and don't follow a pattern. If you plotted error with your target you should ideally see that there is no apparent pattern.

How to check validity of assumption and cosequences of violation

1. Normality Assumption: If errors don't follow normal distribution, we can not rely on p-values and confidence intervals for our predictions. However point estimates for target remain unaffected apart from the fact that parameter estimates might be non-zero by chance. We can check whether this is true or not by looking at qqplot for errors, histogram with kernel density curves for errors.
2. Homoscedasticity Assumption: If you plot errors with target and instead of a random cloud you see a funnel shape or some other pattern , that'd be an indication of hetroscedastic errors. One popular remedy for the same is to take log of response and use that instead. Taking log brings down scale of errors and of course scale of difference between them as well.
3. Independence Assumption: If error seem to follow some pattern with any of the predictor that indicates at non-linear relation between and y and that predictor. We should try appropriate variable transformation instead of using that variable as it is.

Multiple Linear Regression

Multiple linear regression is just an extension of simple linear regression. Although few additional issues come up due to extra variables but basic frame work remains same. Instead of one predictor variable x, you have multiple predictor variables $x_1, x_2, x_3.....x_p$. The target y can still be written as linear combination of these predictors:

$$y_i = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3..... + \beta_p * x_p + e_i$$

Again for finding best values of parameters, we'll minimize error sum of squares $\sum_{i=1}^{n} e_i^2$. Instead of 2 linear equations, now we'll have (p+1) linear equations to solve. Other than that everything else remains pretty much same.

Multi-collinearity

multi-collinearity means, one or more of the predictor variables being linearly dependent on few other predictor variables. There is a very important consequence to this in MLR. Recall the test statistic for hypothesis which we built to figure whether parameter estimate for β_i is significantly different from zero. If that test statistic is close to zero, p-value becomes high and you discard the corresponding predictor variable.

This reliance on p-value to discard not-so-good variable is alright until multiple predictors came into picture and with them , came multicollinearity. Discussion, which is about to follow is a little technical, if you find it overwhelming, don't worry, in the end we'll walk you through a case study done in SAS which you'll be able to understand even if this discussion doesn't make sense to you.

Lets say you have a target y which you are trying model as linear combination of $x_1, x_2, x_3.....x_p$. Forget about y for a minute. Consider that we choose one of the predictor as target and rest as predictor for it. Let x_j is my target and $x_1, x_2, x_3...x_{j-1}, x_{j+1}.....x_p$ are predictors for it. I go on and build a linear regression model and find its coefficient of determination to be R_j^2. if R_j^2 is high that means x_j can be written as linear combination of $x_1, x_2, x_3...x_{j-1}, x_{j+1}.....x_p$. This also implies that x_j contains redundant information. Another repercussion of this comes from the test statistic of hypothesis $\beta_j = 0$.

Remember the test statistic $\frac{\beta_j}{S_{\beta_j}}$. The denominator is proportional to $\frac{1}{(1-R_j^2)}$. As R_j^2 goes close to 1, S_{β_j} goes close to infinity, which means test statistic $\frac{\beta_j}{S_{\beta_j}}$ goes close to zero. Which in turn makes p-values artificially high. And you might end up discarding the variable. Remember that, p-values here are going up, **not** because variable is not correlated enough to y, but because of multicollinearity present in the data.

Which implies if multicollinearity is present in the data, you can not rely on the p-values. We can measure multicollinarity effect caused by presence of each variable by looking at the corresponding $\frac{1}{(1-R_j^2)}$ values. This is also called Variance Inflation Factor because it is the factor by which variance of parameters estimate S_{β_j} goes up.

$$VIF = \frac{1}{(1 - R_j^2)}$$

We will first need to remove variables from our consideration which have high VIF values and then proceed to build our regression model.

Using Categorical Variables

The process that you have seen requires all the variables considered to be strictly numbers. Which is not really the case with real life data. We get mix up of both numeric and string characteristics in our data. Then how do we use categorical variables in our model?

We need to figure out some way to convert them to numbers. Blindly giving them 1,2,3,4.... doesn't makes sense, it also makes your model simply bad. Because the mathematical framework is going to treat these numbers as usual. So 6 will be treated as 2 times 3, whereas say the variable in question was city names, and you assigned 3 to Agra and 6 to Kochi, it might not makes sense to say the **name** "Kochi" is 2 times "Agra".

We can comment that if I increase my numeric predictor say x_i by amount Δt then my target will change by some multiple of this say $\beta_i * \Delta t$. Now categorical variable don't "change" like that. At one time they will take some fixed category as a value at another some other category as a value. And measuring "change" in that context doesn't really make sense.

Now lets think how categorical variables really affect your target. Lets say your target is risk score for heart deceases and you have a model w.r.t. age As this:

$Risk = \beta_0 + \beta_1 * Age$

You have another categorical variable "history" which takes value "both" when both of your parents have history of having heart decease. "one" when one of the parents has history of heart decease, "none" otherwise. We can hypothetically say that, if history="both", Risk goes up by 0.10 or 10%, if history="one" then Risk goes up by 0.05 and no effect if history="none". So we can have three different models to incorporate this categorical variable.

when history="both" then $Risk = \beta_0 + \beta_1 * Age + 0.1$. if history="one" then $Risk = \beta_0 + \beta_1 * Age + 0.05$ and when history="none" then simply $Risk = \beta_0 + \beta_1 * Age$. We can combine this if we consider two **dummy** variables, history_both=1 when history="both" and 0 otherwise. history_one=1 when history="one" and 0 otherwise.

$Risk = \beta_0 + \beta_1 * Age + 0.1 * history\_both + 0.05 * history\_one$

The lesson here is that we can convert a categorical variable which takes **n** distinct values to a set of **n-1** dummy variables as mentioned above and use them just like numeric variables in the regression process. You must be wondering why **n-1** . Create on your own a hypothetical categorical variable which takes values "a" or "b" or "c" randomly. Create 10 observations for that. Now create three dummy variables cat_a, cat_b and cat_c. Observe that you can write:

$cat\_c = 1 - cat\_a - cat\_b$

you can randomly switch places of cat_a, cat_b and cat_c, and this equation will still hold. Basically if you know values of **n-1** dummy variables , you can perfectly know what would be the value of n^{th} dummy variable. To avoid this situation of perfect multicollineraity between predictors we need to make only **n-1** dummy variables for a categorical variable which takes **n** distinct values or has **n** distinct categories.

Then the question comes which one should I drop OR *not* create a dummy variable for.Well, it doesn't really matter. However, standard is to drop the category which is least frequent.

Model Validation

We have pretty much learned every thing which we needed to build a linear regression model. However it might not be enough still to achieve our goal , which was to come with an equation which enables us to forecast results for **yet unknown** or **future** values of predictors. Why? , because the parameter estimates that you got are for the data that you already have. How do you check if this model will perform well on the unseen data as well?

The answer is rather simple, randomly break your data in two samples and use one to build your model and check performance on the second one. We'll formally call them train and test datasets.

Although you'll come across few machine learning algorithms [Decision Trees], where they have cross validation inbuilt functions which utilize train data for tuning the model.

k fold cross validation : Underlying process is that , your train data is broken into k random parts. At a time a model is built on k-1 parts together and its performance [error] is checked on the left out part. This process is done k times , leaving one part [out of k parts] for measuring error. Final error is calculated as average of k iterations. This average error helps in tuning parameters of machine learning models. We'll implementation of cross validation when we reach to that part of the course.

For now we'll rely on breaking our data into two parts train and test.

Let me re-iterate why we need to do this. Reason is rather practical requirement than statistical. Ultimate goal of all this model building is to get an equation which can be used on unseen data to predict outcome of the business process.

Lets Summarise The Model Buidling Process

- Remove or impute missing values in the data
- Create dummy variables for the categorical variables
- Break your data in to three parts : train and test. start building model on train
  - In the first run of your model building process check VIF for all variables.
  - Drop variable with high VIF (>5). Drop them one by one only **not all at once**. drop the one with highest VIF, run your process again and then pick again the variable with highest VIF and drop, keep on doing this until VIF is (<5) for all variables.
  - Once VIF is under control for all the remaining variables , start dropping variable one by one based on p-values. If p-value is **greater** than 0.05, drop that variable
  - You have your train model , once you have all remaining variables with p-values less than 0.05.
- Test this model's performance by calculating RMSE (Root mean square error) on test dataset.
- You can use RMSE calculated on test data to compare multiple models.

Applications of Linear Regression :

Predictive Analytics: forecasting future opportunities and risks is the most prominent application of regression analysis in business. Demand analysis, for instance, predicts the number of items which a consumer will probably purchase. However, demand is not the only dependent variable when it comes to business. Regression analysis can go far beyond forecasting impact on direct revenue. For example, we can forecast the number of shoppers who will pass in front of a particular billboard and use that data to estimate the maximum to bid for an advertisement. Insurance companies heavily rely on regression analysis to estimate the credit standing of policyholders and a possible number of claims in a given time period.

New Insights: Over time businesses have gathered a large volume of unorganized data that has the potential to yield valuable insights. However, this data is useless without proper analysis. Regression analysis techniques can find a relationship between different variables by uncovering patterns that were previously unnoticed. For example, analysis of data from point of sales systems and purchase accounts may highlight market patterns like increase in demand on certain days of the week or at certain times of the year. You can maintain optimal stock and personnel before a spike in demand arises by acknowledging these insights.

Operation Efficiency: Regression models can also be used to optimize business processes. A factory manager, for example, can create a statistical model to understand the impact of oven temperature on the shelf life of the cookies baked in those ovens. In a call center, we can analyze the relationship between wait times of callers and number of complaints. Data-driven decision making eliminates guesswork, hypothesis and corporate politics from decision making. This improves the business performance by highlighting the areas that have the maximum impact on the operational efficiency and revenues.

Supporting Decisions: Businesses today are overloaded with data on finances, operations and customer purchases. Increasingly, executives are now leaning on data analytics to make informed business decisions thus eliminating the intuition and gut feel. Regression analysis can bring a scientific angle to the management of any businesses. By reducing the tremendous amount of raw data into actionable information, regression analysis leads the way to smarter and more accurate decisions. This does not mean that regression analysis is an end to managers creative thinking. This technique acts as a perfect tool to test a hypothesis before diving into execution.

Correcting Errors: Regression is not only great for lending empirical support to management decisions but also for identifying errors in judgment. For example, a retail store manager may believe that extending shopping hours will greatly increase sales. Regression analysis, however, may indicate that the increase in revenue might not be sufficient to support the rise in operating expenses due to longer working hours (such as additional employee labor charges). Hence, regression analysis can provide quantitative support for decisions and prevent mistakes due to manager's intuitions.

Case Study :

We'll also learn syntax for R [which is surprisingly simple] while we are carrying out model building process.

Loan Smart is a lending advisory firm. Based on their client's characteristic and needed loan amount they advise them on which Financial Institution to apply for loan at. So far their recommendations have been based hunches business experience. Now they are trying to leverage power of data that they have collected so far.

They want to check whether given their client's characteristics , they can predict how much interest rates they will be offered by various financial institution. They want to run with proof of concept for this idea. They have given us data collected for one such financial institution ABC Capitals Ltd.

What we need to do is to figure out whether using that data we can predict interest rate offered to client. We have developed the problem the way you'd encounter problems in projects. You are given training data and testing data, testing data doesnt have response values. We'll eventually want to make prediction on this data where the response is unknown. Lets start with importing data

```
ld_train=read.csv("~/Dropbox/0.0 Data/loan_data_train.csv",stringsAsFactors = F)

ld_test= read.csv("~/Dropbox/0.0 Data/loan_data_test.csv",stringsAsFactors = F)
library(dplyr)
glimpse(ld_train)
```

```
## Observations: 2,200
## Variables: 15
## $ ID                          <int> 79542, 75473, 67265, 80167, 172...
## $ Amount.Requested            <chr> "25000", "19750", "2100", "2800...
```