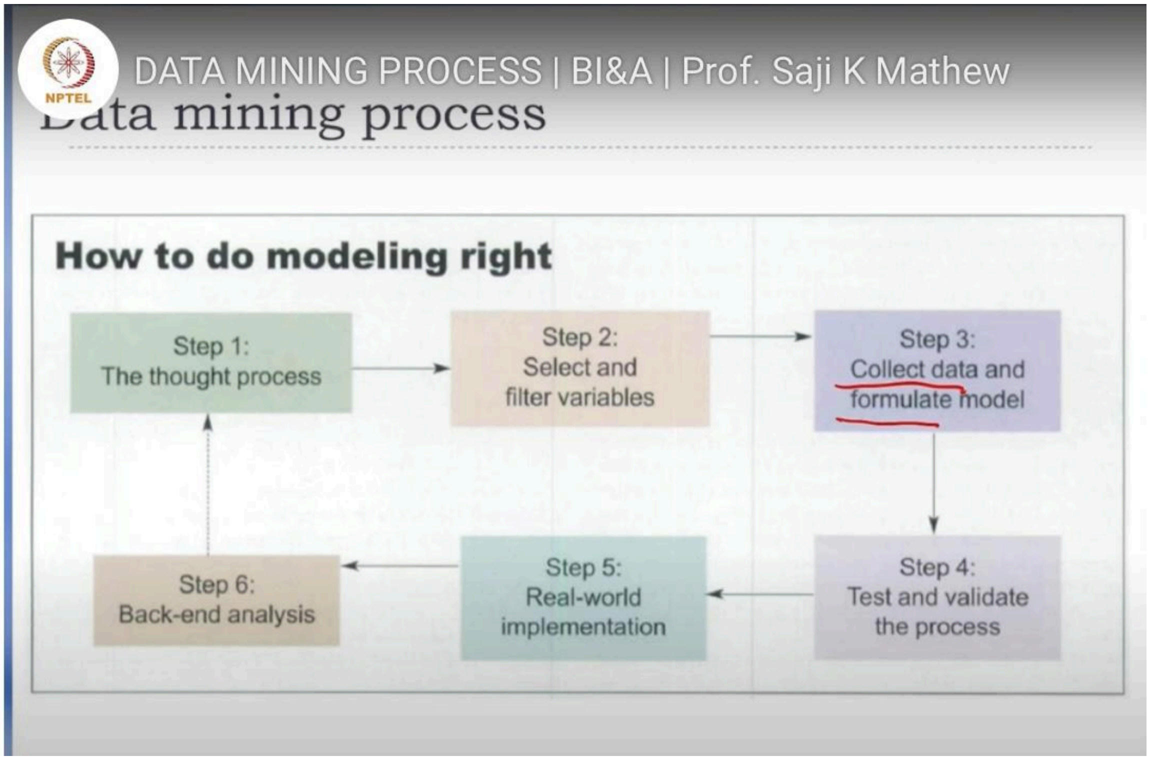
**Course Name:Business Intelligence and Analytics**
**Professor Name:Prof. Saji.K.Mathew**
**Department Name:Department of Management Studies**
**Institute Name:Indian Institute of Technology Madras**
**Week:05**
**Lecture:19**

## DATA MINING PROCESS | Business Intelligence & Analytics

Hello, and welcome back to Business Intelligence and Analytics course. We have been discussing analytics process or data mining process. In this session, I use both the terms synonymously, but they are not, but they are very similar in terms of activities involved. So we have already learned that analytics follows certain process which consists of certain specific steps and those are the steps one should ensure that they are followed when one works on analytics problems or business problems which are solved using analytics. So we have seen the first step is the thought process where we convert the problem, business problem to an analytics problem and then we think in terms of variables, because analytics rely on algorithms and models and therefore variables are the sort of world of analytics or constituents of the analytics and therefore selection and filtering of variables, making variables relevant and also, as is the next step we saw data collection wherein we ensure that the data that we collect based on the variables we identified, are having sufficient quality to be used in data analysis. So and we discussed various aspects of data quality and data related issues and now we are going to see the next step or the related step which is about modeling.

We have to formulate a model. A model, what is a model? A model is a simplified reality for a purpose, that is one definition of a model. Model is a simplified reality because when we model a phenomenon, say for example, sales as a phenomenon, we actually model sales as dependent on certain variables. Those are variables which we select to model sales, but we all know that there could be variables other than that we chose which would influence sales.



DATA MINING PROCESS | BI&A | Prof. Saji K Mathew

Data mining process

How to do modeling right

Step 1: The thought process

Step 2: Select and filter variables

Step 3: Collect data and formulate model

Step 4: Test and validate the process

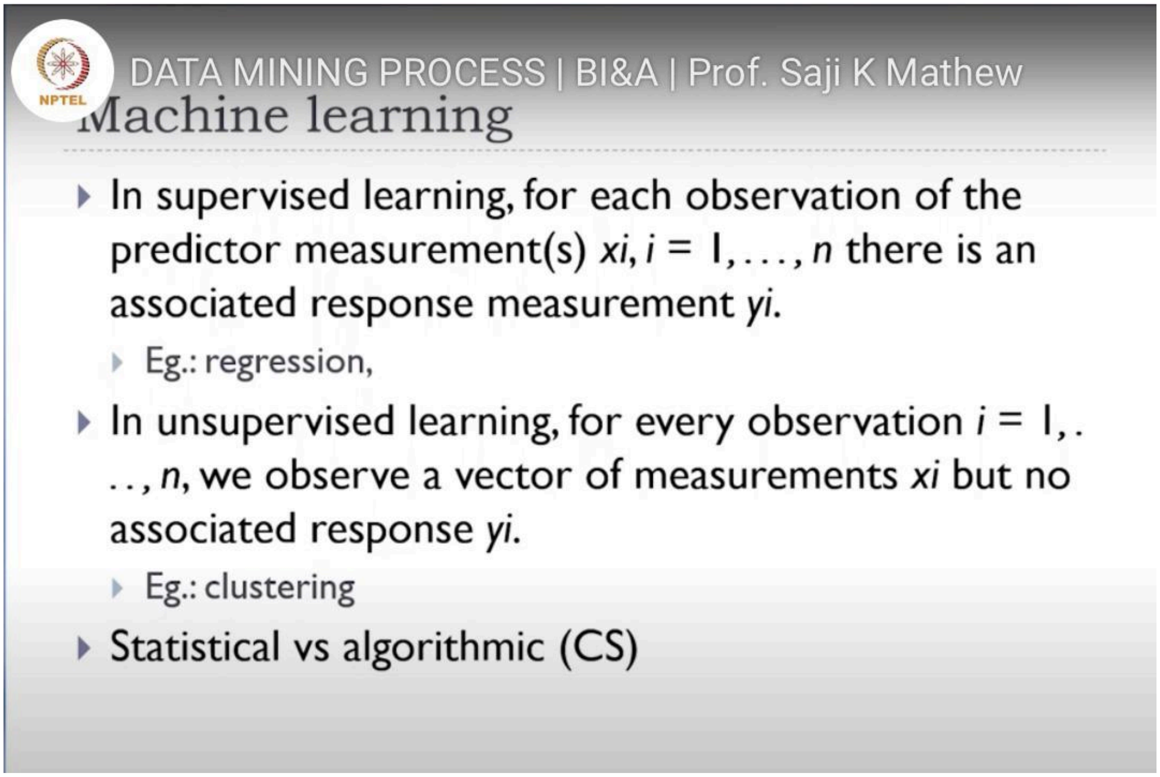
Step 5: Real-world implementation

Step 6: Back-end analysis

We also know that there is the so called environment or context. Suppose there is a war or sudden change in the environment, all the variables that we chose become irrelevant because they do not explain the loss of sales that happens. So model is a simplified reality and when we work on analytics, we should know there is a limitation. So model has certain explanatory power, but model is not 100 percent. So formulating the model, deciding on the model, depending on what purpose we have in mind when we model is a very important step in analytics and that is exactly what we are going to see now.

How to choose a model or in other words, what are some of the principles that we should keep in mind while we decide on a modeling, choice of models as well as when we build a model, what are the important considerations. So I am going to touch upon some of them. That will be followed by how we test and validate a model. So that is the fourth step as we see here. So moving on, so we are entering the realm of machine learning or as your textbook is titled statistical learning.

In statistical learning, we use statistical techniques for modeling or building models or machine learning is more generic where we say that a machine, a machine could be a black box which contains certain algorithms, but we train them so that they learn and then we ask them to do certain activities. It could be about predicting something based on the learning they have etc. So broadly in machine learning, there are three categories of machine learning techniques are summarized in the slide. The first category is the supervised learning, supervised learning techniques.



DATA MINING PROCESS | BI&A | Prof. Saji K Mathew

## Machine learning

- In supervised learning, for each observation of the predictor measurement(s) $x_i, i = 1, \ldots, n$ there is an associated response measurement $y_i$.
  - Eg.: regression,
- In unsupervised learning, for every observation $i = 1, \ldots, n$, we observe a vector of measurements $x_i$ but no associated response $y_i$.
  - Eg.: clustering
- Statistical vs algorithmic (CS)

In supervised learning there is a supervisor. There is a supervisor which enables the learning process or in other words in, sorry, in supervised learning the model will look like a y equals fx. Now there is y which is dependent on x. Now when we collect data to build the model, what are we actually trying to do? We have the y data, we have the x data. Suppose x is only 1, so we collect data about y, we collect corresponding data about x. So we have the xy pair here.

And now what is f in this model? f is the function that maps y to x. f is the function, f which maps y and x. So it could be as in regression as an example, it could be y equals bx plus c. If you are using a sample data, so there we know that b is the slope and c is the intercept. So something like this and this is c, this is y axis, this is x axis.

And the slope of the line is b. And that is what a simple linear regression is. Now this is an example for machine learning because we collect x data and y data. The question in front of us is, how do we know the value of b and c? Or in other words, modelling here means or model building here means discovering the function f from the xy pair data. This is xy is nothing but observational data or data that we have collected from some phenomenon.

So we come back and we formulate some method by which the b and c of the model can be estimated from the xy pair of data. And that is called model building. Now I will call this kind of a machine learning technique where this is the machine, this is what the machine is, a model. And this learns from data and estimates the value of b and c. Now for every value of x there is a y.

When the value of x was say x1, what was the value of y1? When the value of x was x2, what was the value of y2? This is known. So in building the model, we are informing the model, these data pairs. What was y when for a given value of x? And therefore, the y is supervising the estimation of b and c. The estimation or in other words, the estimation of b and c is based on a dependent variable value which is known. And that kind of a technique is supervised learning.

In unsupervised learning, there is no such variable as y or there is no such variable, as this known as target variable. In unsupervised learning, there is no target variable. There is no target variable, you have only x1, x2, x3. Here x1, x2, x3, I am not putting as values of x, but x1, x2, x3 to xp are p variables. And there is no y variable.

Think of an individual. Now here I am talking about an individual's attributes then. An individual has attributes like the height, the weight, the various measurements of the body and then color. There are several attributes to represent a person or related to a person.

That is a vector. We call it a feature vector in machine learning. So x1 to xp is a set of features of a object. Now we do not say x1 is a dependent variable and x2 to xp are independent variables. No, there is no definition of dependent and independent variables in unsupervised learning. We will learn a method called clustering and we will see that in cluster analysis, we do not have a dependent variable.

We only have a set of variables or a feature set based on which objects are grouped into clusters. And that kind of, that class or that category of learning techniques are known as unsupervised learning techniques because they do not have a target. And then again in machine learning, another way to classify machine learning techniques is to look at which science or where did that particular method originate. Is it in statistics like regression or is it the algorithmic community like the computer science, you know like the neural networks or the decision trees etc are algorithms developed by algorithmic community or computer science community. So there are different techniques that were developed in different fields of knowledge like statistics and probability and you know algorithms and so on.

So there are techniques divided or categorized this way. Now let me try to explain to you some basic concept before we move on. I have always used linear regression as a very simple method to explain the model building and model testing process. Typically in a project, an analytics project where your aim is to explain y in terms of x, y in terms of a set of x variables, that is when we have the y equals fx model. Typically the, typical example as the regression model.

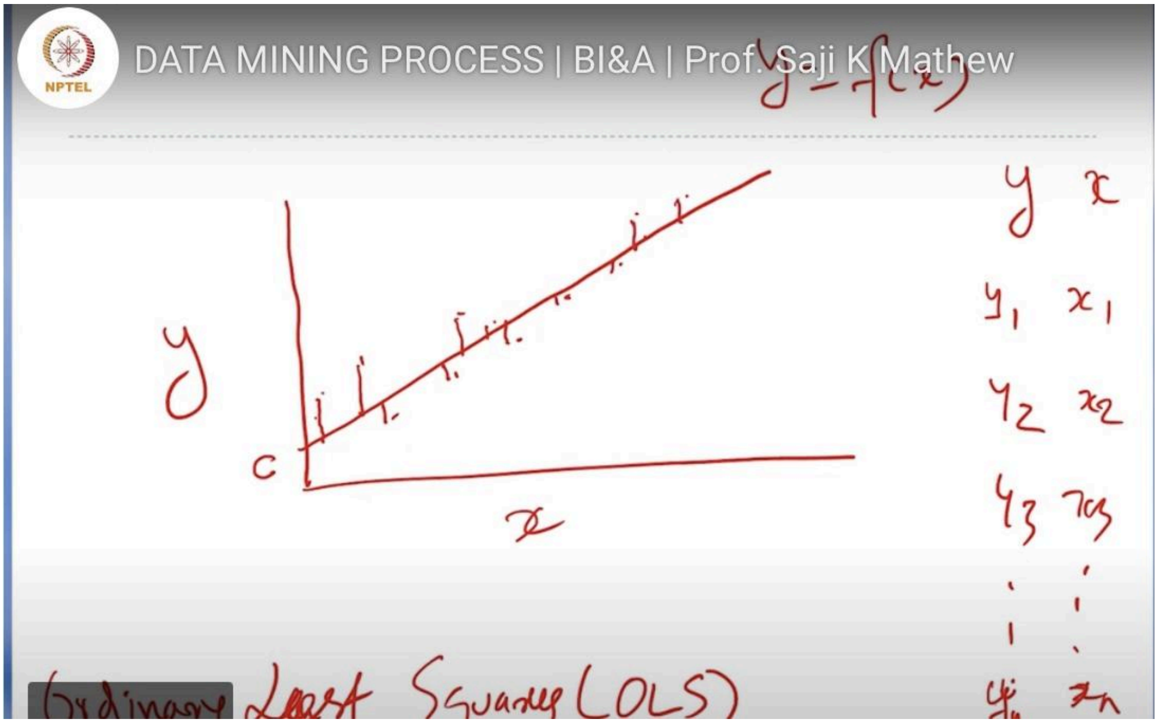
And we already saw that a regression model is estimated from y and x, y is y1, y2, y3 data, x is x1, x2, x3, x4 data, xp, sorry x here it is the size of the data is n, so I call it nth pair of data. Now how do we estimate a value of the coefficients of regression or what is the principle? I think I touched on this, if you see if you try to have a scatter plot of this data and suppose it looks like this and we try to fit a line like this. We are actually imagining that there is a line that fits the x, y distribution of data. And we know that this is the c and the slope of the line is b. Now the question is how do we arrive at that? How do we determine the values of b and c? What is the principle? So, you can see that when I try to draw a line like this, I am using my imagination well, this is the best line I can draw.

Somebody else from the class can come and draw a line which is different from what I draw and that line can be maybe like this. Somebody else could draw a line like this. And then we have a problem which line is the best line or which line should be plotted. And for that is where we use an estimation technique called in this case ordinary least squares or OLS is one of the techniques for estimating the model coefficients and how is that done. And suppose this is the line that we imagine to be the line and then we can see that the line passes through some of the dots or some of the x, y pairs but some points are not covered or not inside the line.

So therefore, you can see there is a difference between the data and the model. There is a difference between the data and the model. And each of this can be called an error. It is like an error. Why it is an error? Well, I determine a function or a linear function to model this data but the function is not exactly fitting the data.

So each of this is an error or e1, e2, e3 is negative and so on. Now what would be an ideal line? This e1, e2, e3 would be called as residuals in modeling parlance. This will be known as residuals, the remnant which is not covered by the regression model. That is a residual. And what is an ideal line? An ideal line you can imagine is that line which produces the minimum residuals, minimum total residuals.

And we need to have a measure for this error or residuals. And how do you actually find that out? You can sum up all the errors. Suppose you are starting and n is the number of data points. So there is the yi point.



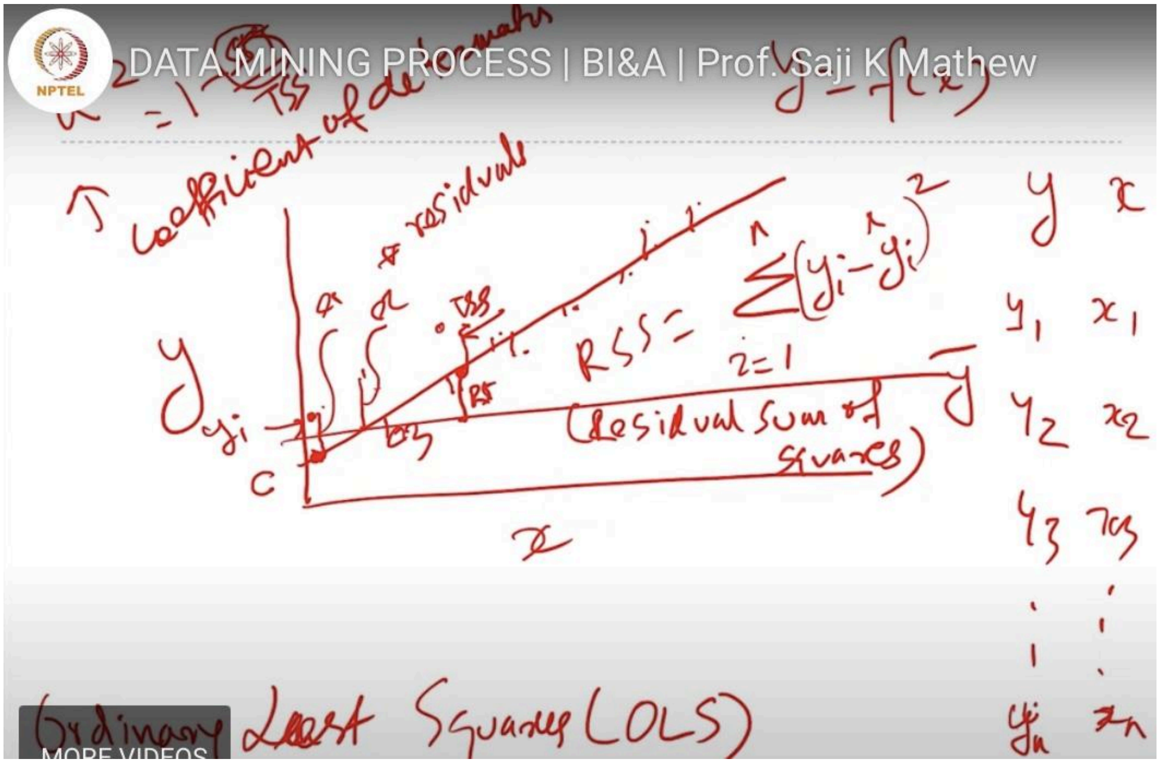
That is, the actual yi is this. And the predicted yi or the, according to the model, the model treats yi as this point. That is what the model says is yi. There is a difference. You represent it as yi cap. That difference when it is summed, squared and summed for all the points, that is the residual sum of squares.

Residual sum of squares. So what is desirable in a linear model like this is that the model should be estimated such that the RSS is minimum. The best fit for this data is that line which has the minimum RSS. And therefore, you can now see that it is a minimization function. It is the value of b and c have to be estimated such that RSS is minimized. So therefore, this can be converted to a minimization problem.

I am not deriving that formula. There are, this is available in textbooks. But I am giving you the overall principle of modeling that you try estimate the model parameters such that the overall error is minimum. That is one aspect. And now, as I said, there is a possibility to draw multiple lines for this data.

You know, I drew three lines. And each line, of course, when you calculate will have a different RSS. But the problem with RSS, a measure like RSS is that it is an absolute value. It is an absolute value. And therefore, it becomes a bit of a challenge to compare when you do modeling in different context. Suppose, for a similar phenomenon, you collect and data and build a model, but you know, for example, the size of the data is different etc. Then the RSS will actually vary based on the size of the data, number of variables, etc. So it becomes difficult to compare models. And therefore, you have, you know, more useful measures for determining a fit of a model. One such useful measure is the $R^2$. I think I talked about $R^2$ when we were discussing multicollinearity.

$R^2$ is a measure of fit. And the good thing about $R^2$ is that it is a ratio. And it can be converted to a value between 0 and 100. 100 percent meaning really good fit and 0 meaning no fit at all. So you get a measure which is comparable and more intuitive to comment on models, how well the model has, is fitting with the data. And you know that for calculating $R^2$, you need to actually have a average line y and it is with respect to the average y, you actually look at, say regression sum of, you calculate regression sum of squares and total sum of squares.


NPTEL
DATA MINING PROCESS | BI&A | Prof. Saji K Mathew
$R^2 = 1 - \frac{RSS}{TSS}$
↑ coefficient of determination
$y = f(x)$
residuals
TSS
RS
RS
$RSS = \sum_{i=1} (y_i - \hat{y}_i)^2$
(Residual sum of squares)
y
$y_i$
C
x
$\bar{y}$
y x
$y_1$ $x_1$
$y_2$ $x_2$
$y_3$ $x_3$
$y_n$ $x_n$
Ordinary Least Squares (OLS)
MORE VIDEOS

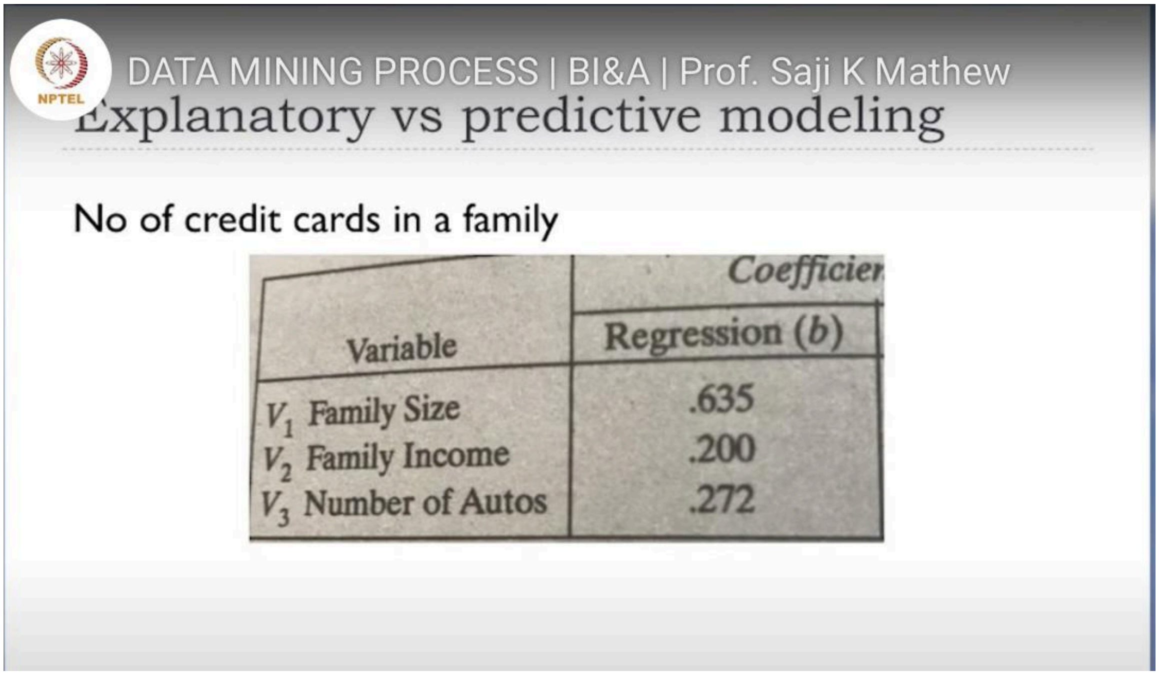
And then the difference is the error. So this will be defined as 1 minus SST by TSS. So the difference, so when this difference, all the difference, all the errors are 0, then the numerator becomes 0 or $R^2$ becomes 1. That is the principle for $R^2$ or coefficient of determination. Now the coefficient of determination also varies based on the number of variables. And therefore, you have a better formula for $R^2$ based on the degrees of freedom.

So degrees of freedom adjusted $R^2$ or adjusted $R^2$ is more desirable to represent the goodness of fate of a model, because it adjusts for the number of variables. When the number of variables increase, the degrees of freedom reduces and so forth, that is adjusted statistically. So these are some of the very basic principles in modeling or building a linear model. I will call this as building a model.

This is model building. You have some data, it could be a sample data and you actually use the Euler's technique to estimate the model parameters and your model is ready. So you have calculated say, $y = bx + c$ and you know the value of b and c now. And now you, in a typical regression sort of modeling, you know you have assumptions of regression. So you have to look at the data and the distributions and see that the assumptions of model, assumptions about data are valid before you building the model, for example, the distributional property.

Normal distribution is an assumption in regression. So you have to test all those. So I am not covering or getting into the data analysis part, which is a prerequisite for this course. But having built a model from data, how do you interpret this? We are saying y is a function of x or you say if model is estimated as y equals 3.5, that is estimated value of bx + c is equal to 2.3. This is a model. y is 3.5 times x + a intercept, which is 2.3. This intercept means y is not explained by x alone, but there are other variables potentially, which is not included in the model and therefore, there is a constant which is added. Let us move on because I will be reflecting on what I did now when we go forward in machine learning and this forms a basic or simple example to, with this basic principles or very, very basic principles of modeling, let us now move on to understand a real life scenario. We just learned how we model simple models like a regression model.

What are some of the basic aspects of it and how do we actually look at a model and comment on how well the model explains a given phenomenon or a given relationship between y and x and we developed a measure like the $R^2$ for that and so on. So now we move on and here is an example of a problem. Only look at the top side, only look at the top side, which is this data taken from a, sorry, yeah, I have made it specific to that particular data or table, copied. So, this is about a problem where a researcher was studying what determines the number of credit cards in a family.



DATA MINING PROCESS | BI&A | Prof. Saji K Mathew

Explanatory vs predictive modeling

No of credit cards in a family

| Variable | Coefficier |
|---|---|
| | Regression ($b$) |
| $V_1$ Family Size | .635 |
| $V_2$ Family Income | .200 |
| $V_3$ Number of Autos | .272 |

So therefore, the y is the number of credit cards. Number of credit cards in a family is a function of V1, V2, V3. That is how this modeling is done. So a researcher actually wanted to have this model because the management is interested in knowing what factors influence the number of credit cards a family possesses. So, for a business or especially in credit business, that is important. So that they can do some useful marketing or they can actually take some decisions based on that information.

So as we discussed, they identified the variables and collected data, possibly through a household survey and they have collected data and they built the model. So we just saw how a model is built. So they estimated the model parameters. So we know here that number of credit cards is equal to 0.635 x V1, which is family size + 0.2 x family income + 0.272 x number of autos. The constant is not given. Now, so after building the model, the team has to make a presentation to the management to give an explanation.

So this is an explanatory project. This is explanatory analytics. In explanatory analytics, the objective is to explain y in terms of x. So this particular project team is very happy with their accomplishments and they go and make a presentation looking at the values of the coefficients.

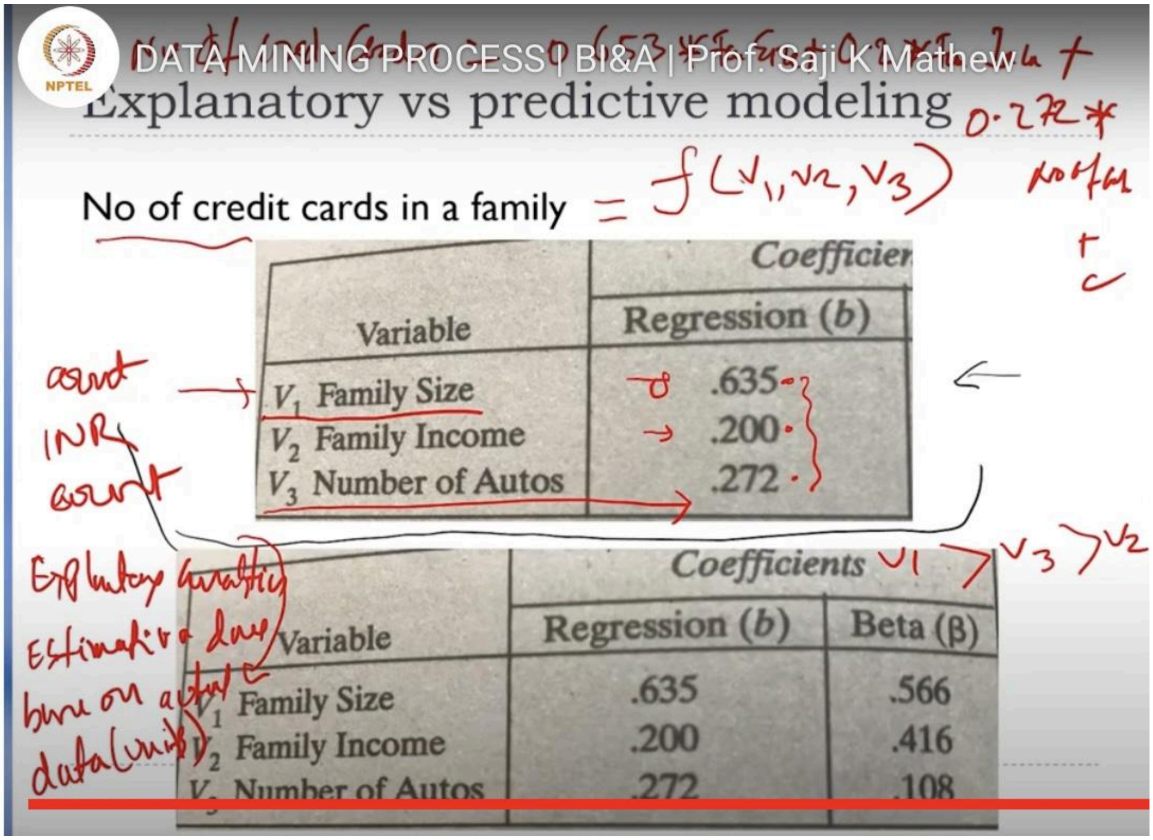
V1 has a coefficient which is 0.635, V2 has 0.2 and V3 has 0.272. And they make a claim that the most important determinant of the number of credit cards in a family is the family size, because that is having the highest coefficient value. The second important one is the number of automobiles and the least important is the family income. So they say V1 is greater than V3 is greater than V2. So the basis for this inference is the values of the coefficients, based on the estimation of the coefficients they did, based on the data.

Now I have this question to the audience. Is this okay? Is this explanation or interpretation okay? You have started thinking, some would say why not? The higher the value of the coefficient we are saying y will be more dependent on that particular value. So V1 is, 0.635 times V1 influences the number of credit cards whereas family income it is only 0.2. But there is a problem here. The problem is that the estimation is done based on actual data or units. For example, family size may be a count, family size is a count. Family income may be in INR, in rupees, number of automobiles is again a count. You see that these variables are measured in different units and therefore their range of measurement or span are different. Each variable has a different unit of measurement and therefore their values will be in a different range and hence, since the range of values are different when you estimate model parameters using OLS, they get, the estimated model coefficients get influenced by the measurement range.

And therefore the values that you obtained may not be directly indicating their relative importance. It is not indicating their relative importance. But in explanatory cases your purpose is to know the relative importance of variables because manager needs to know where should one focus. Should one go to families of higher sizes or should one go and market their product to families of higher income? So therefore this kind of model building can mislead the decision maker and the interpretation will be flawed.

And therefore what is the solution? Since your purpose is explanatory, you need to actually do something known as standardization of data. Standardization or normalization. Here I am using these two terms synonymously but there are differences or in other words I would say Z-score. Z-score normalization which is based on the standard normal curve, is one method of standardizing the data.

There are different standardization techniques. So what you do is instead of using the real data as it is, you standardize them first and then re-estimate the model parameters. What do I mean by standardization? Let me again put that here clearly so that you have a clear understanding of what I am talking about. So suppose you have the y data and x data. Now so you have y1, y2, y3 etc at the end you can actually find a y bar which is the average value. So then the y standard, standardized y data would be for each y1 you calculate the Z-score or Z-score in American pronunciation.



DATA MINING PROCESS | BI&A | Prof. Saji K Mathew

# Explanatory vs predictive modeling

No of credit cards in a family = $f(v_1, v_2, v_3)$

| Variable | Coefficien |
|---|---|
| | Regression (*b*) |
| $V_1$ Family Size | .635 |
| $V_2$ Family Income | .200 |
| $V_3$ Number of Autos | .272 |

| Variable | Coefficients | |
|---|---|---|
| | Regression (*b*) | Beta (β) |
| $V_1$ Family Size | .635 | .566 |
| $V_2$ Family Income | .200 | .416 |
| $V_3$ Number of Autos | .272 | .108 |

So what will be the Z-score or Z-score? ( y1- $\overline{y}$ )/S; y1 minus y average divided by the standard deviation of the entire y values. Instead of using y1, you are going to use the Z-score as the data point. What is the change I have made here? Instead of using the absolute value, I have converted this into a ratio. With respect to standard deviation, what is the deviation of y1 from the average y? That is what a Z-score is representing. So for every point, I am actually comparing the deviation from the mean with respect to the standard deviation.

This is normalizing the data. So instead of taking the data as it is which may be in one range for one variable and another range for another variable, this all becomes standardized data. So similarly for x data also I standardize and for model building, I use only the standardized data. When I do that and then I re-estimate the model coefficients you can see the change that happened. Earlier my beta or b or the coefficient for v1 was 0.635, this was non-standard data. But when it is standardized, the standardized coefficient is 0.566.

You can see that the beta value has changed. For v2 it was 0.2, it has now become 0.416 and number of autos, it has become 0.108. Now you see this particular relative importance is not correct. When we standardized it, we see that of course v1 continues to be the most important followed by family income which was become the next important variable and v3 has become very less in its importance. And this is a valid finding if you also tested the significance of this coefficients, of course that is important to see if the sample has generalizability.

And after doing due test, you can make a presentation to your project sponsor and say look, if you want to maximize the sales of credit card, focus on family size and then focus on family income because your coefficients are comparable. This is an explanatory project you are trying to explain y in terms of x to show which influences y most and so on and so forth. That is called explanatory modeling.

But not all projects are explanatory. Some projects are predictive. We are going to look at predictive modeling now and before we move on to predictive modeling, in the particular example that we took now, for explanatory we know that we need to know the relative importance of coefficients and because we need to know how they explain the outcome.So therefore standardization was important. But in terms in the case of prediction, our objective is not to look at the relative importance of variables. Our only objective is to predict y for a different set of x data which I am going to explain in the next slides. But keep in mind, in such a context it is not important to standardize variables as we did in the case of explanatory projects.

In explanatory projects, if you do not standardize your interpretation will be wrong. But in the case of predictive modeling, your interpretation is not important but predicted value is more important and therefore standardization is not as significant as in the case of explanatory projects.