Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology, Madras

Module - 04

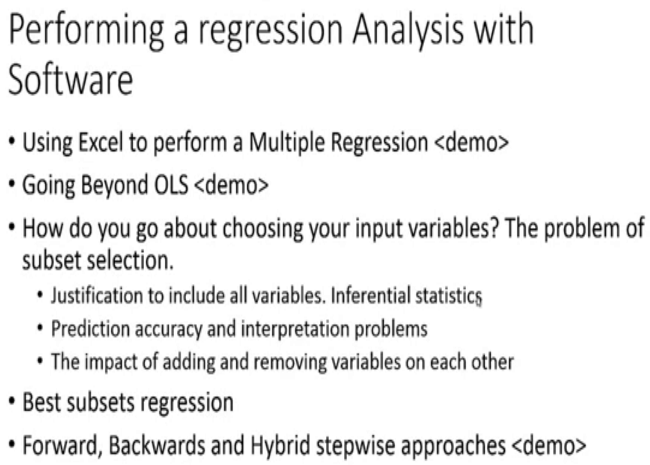
Lecture - 20

Simple and Multiple Regression in Excel and Matlab

Hello and welcome to our series of lectures on the topic of Regression. Over the past few lectures we discussed the idea, we motivated the use of this approach called regression. And in the last lecture we even looked at the idea of deriving the ordinary least squares regression and we did that in the simple linear regression case.

Today, we are going to look at how you can implement a regression through software and we will show to you in both excel as well as mat lab. But, the important thing is to just understand the core concept and the terms, so tomorrow you would be able to implement it with any software of you are choosing.

(Refer Slide Time: 01:08)



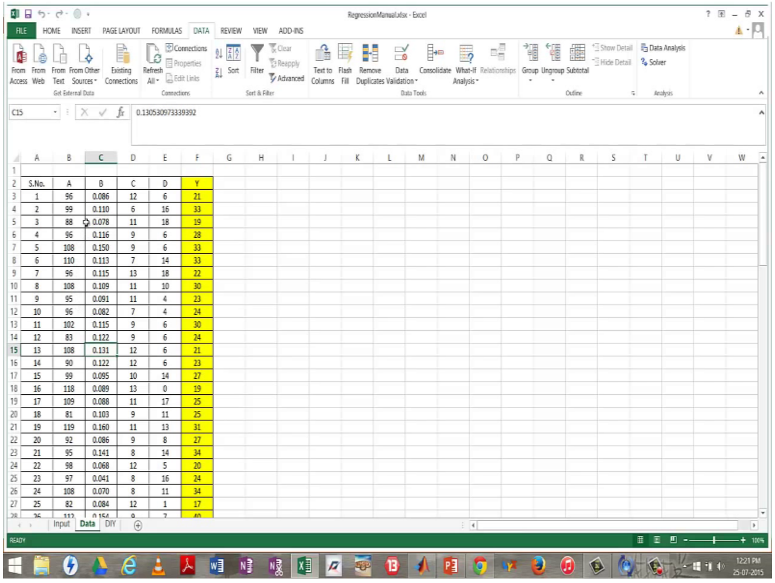
Performing a regression Analysis with Software

- Using Excel to perform a Multiple Regression <demo>
- Going Beyond OLS <demo>
- How do you go about choosing your input variables? The problem of subset selection.
  - Justification to include all variables. Inferential statistics
  - Prediction accuracy and interpretation problems
  - The impact of adding and removing variables on each other
- Best subsets regression
- Forward, Backwards and Hybrid stepwise approaches <demo>

Now, with respect to the main topic that we discussed last time, which was simple regression. So, when we did this whole analysis on the derivation for the ordinary least squares, we took a case of simple regression which just means that there is one input variable; that is in bound. Today, we are going to extend that, we are going to look at the case, where there are multiple input variables.

So, let us start by setting up an example and I will show you a demonstration in excel. For performing the multiple regression, we can talk about some terms that you might encounter when using this software.

(Refer Slide Time: 01:52)



C15 0.130530973339392

S.No.	A	B	C	D	Y
1	96	0.086	12	6	21
2	99	0.110	6	16	33
3	88	0.078	11	18	19
4	96	0.116	9	6	28
5	108	0.150	9	6	33
6	110	0.113	7	14	33
7	96	0.115	13	18	22
8	108	0.109	11	10	30
9	95	0.091	11	4	23
10	96	0.082	7	4	24
11	102	0.115	9	6	30
12	83	0.122	9	6	24
13	108	0.131	12	6	21
14	90	0.122	12	6	23
15	99	0.095	10	14	27
16	118	0.089	13	0	19
17	109	0.088	11	17	25
18	81	0.103	9	11	25
19	119	0.160	11	13	31
20	92	0.086	9	8	27
21	95	0.141	8	14	34
22	98	0.068	12	5	20
23	97	0.041	8	16	24
24	108	0.070	8	11	34
25	82	0.084	12	1	17
26	112	0.154	9	7	40

So, this is the sample data set and what you have in column A, there is serial numbers to give you an idea of how many data points there are. I have chosen an example, where there are 88 data points. The variables B, C, D and E as you can see selected on the monitor are actually the input variables and column F, which is shown with, which is highlighted and also labeled as y is your output variable.

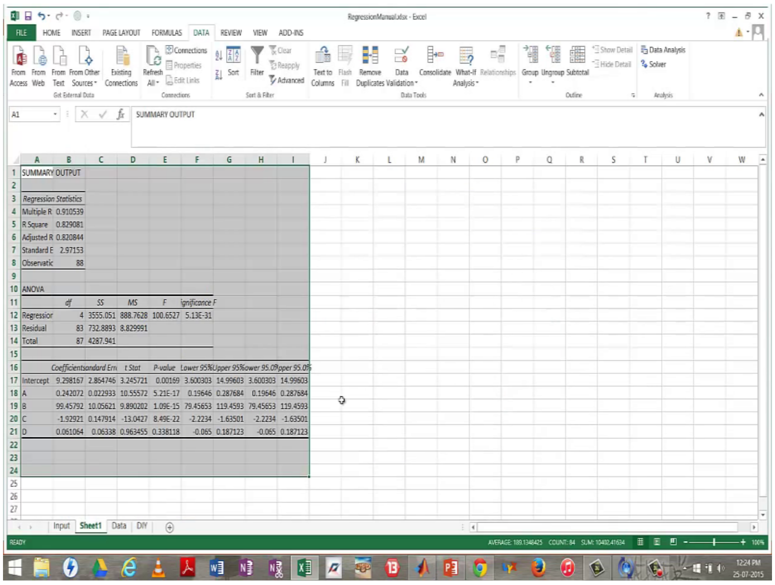
And, so for the first time you are interested in creating an equation, regression equation that maps these input variables to y, the output variable. So, the way you would do that and I am using Microsoft excel out here would be to go to, you see the topics on top that say file home insert page layout where my mouse is and in that, you select data. So, usually your screen might be at home, you would go click on data, then you would go to data analysis, click on that.

And that will pop up a set of possible tools that you could use analysis tools. And for some versions of excel you might not have this readily; that is you will not have this data analysis button already and there you usually need to go and add it in and it is called an ad in and it should be there, it is there in the software, but it is just not installed for you. So, here you would choose the topic regression and you would click on, now it is asking you to tell it, where the data is, so that is what I am going to do I am going to show where the data is and the way I did that was to kind of select that whole column.

And if you are bigger familiar with excel, you can just use some hot keys to do that as well. Just note that I am also selecting the titles of the variables when I am in putting the variables. And, because I do that I need to check this box versus labels, meaning that I have also provided you with labels. The idea here is, the constant is 0 basically means should the regression line be force to go through the origin and we do not want that, so we will not check that. That basically means in your equation y is equal to beta naught plus beta 1 x 1 plus beta 2 x 2 and so on, your beta naught is going to be forcefully said to 0 and that is not something you want.

The second also gives you out here you have something called confidence level and that gives you the option of choosing a confidence level or essentially an alpha; that is different from 0.05 or 95 percent, but we are quite happy with 95 percent, so we will leave that. There is a couple of other options, the thing is I said that I want my results in a new worksheets, so that is what I am going to select.

(Refer Slide Time: 04:54)



A1 SUMMARY OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.910539
R Square	0.829081
Adjusted R	0.820844
Standard E	2.97153
Observatic	88

ANOVA

	df	SS	MS	F	ignificance F
Regressior	4	3555.051	888.7628	100.6527	5.13E-31
Residual	83	732.8893	8.829991		
Total	87	4287.941			

	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	9.298167	2.864746	3.245721	0.00169	3.600303	14.99603	3.600303	14.99603
A	0.242072	0.022933	10.55572	5.21E-17	0.19646	0.287684	0.19646	0.287684
B	99.45792	10.05621	9.890202	1.09E-15	79.45653	119.4593	79.45653	119.4593
C	-1.92921	0.147914	-13.0427	8.49E-22	-2.2234	-1.63501	-2.2234	-1.63501
D	0.061064	0.06338	0.963455	0.338118	-0.065	0.187123	-0.065	0.187123

RegressionManual.xlsx - Excel
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ADD-INS
From Access From Web From Text From Other Sources Existing Connections Refresh All Connections Properties Edit Links Sort Filter Clear Reapply Advanced Text to Columns Flash Fill Remove Duplicates Data Validation Consolidate What-If Analysis Relationships Group Ungroup Subtotal Show Detail Hide Detail Data Analysis Solver
Input Sheet1 Data DIY
READY

So, I say and this is the output that I get you might notice put in to a new sheet. So, let us just look at this output and try to process this. So, what it says out here is that the important things we have to note is the R square, the R square is in some sense a measure of how could your over all fit is. Essentially, it tries to say, how much of your variation in y is actually being explained by the model that you created versus how much of the variation in y, it is just beyond noise. So, the R square is essentially a number that goes from 0 to 1 and it is essentially the square of the multi, but it is called the multiple R.

There is another term called adjusted R square is a very important term and we will come to that in a minute. And you finally, have standard error out here, which is been highlighted and standard error is nothing but, for a given predicted point, how much is the general deviation of the actual point from the fitted module for on average, so that is, what is being captured in standard error. So, an observation is just the number of observations there are, below this top table that we have just discussed you see two tables.

Let me first talk to you about the bottom most table and then I will come to the table above, the idea here is that here you have the intercept and the four input variables. Now, what is called as the coefficients correspond to the betas? So, when it says intercept the coefficient 9.29 essentially talks about beta naught and corresponding we for A, B, C and D are their respective betas. Because, just remember that we have a functional form of the nature, I am just going to type it out here y is equal to some constant, let us called beta naught plus beta, let us call it 1 times A plus beta 2 times B and so on; that is the functional form that we have and it goes all the way from C and D.

The big question is, what is beta naught, beta 1, beta 2 and that is what these coefficients represents. You can call them beta 8 times A if you want, but essentially the coefficient corresponding to A is what we are talking about here and these are the coefficients. These actual values that you see here, excel has done this regression for you and it gives you these results. The standard error that is being discussed here is the standard error around this coefficient, what is… So, this coefficient is nothing but, an estimate based of sample going back to a topics in inferential statistics.

And the standard error pertains to the uncertainty or the standard deviation using quantifying it with standard deviations, uncertainty around that coefficient, which is being quantified by the standard deviation associated with this estimate. Because, this is an estimate and there is some uncertainty around an estimate just like you had a sample mean and the sample mean, which you got from the data, which is supposed to represent the population the sample mean essentially is a random sample from a distribution.

Because, each time you take a sample and you take a mean you are not going to get the exact same value. So, you are getting a value from a distribution and that distribution has some standard deviation we for instance earlier discussed about this standard deviation of the sample mean it is something that you can get from the distributions. So, similarly the standard error quantifies the uncertainty around this coefficient around each of their respective coefficients and it does through the measure of standard deviation great.

So, now finally, from standard error we go to the t statistic and the reason for that is that this estimate, which is this coefficient. The distribution associated with this estimate just like the distribution associated with the sample mean was t distributed if you did not know the standard deviation of the population that is the same core idea here, which is that it is t distributed and excel calculates t statistic for you and gives you the p value associated with the t statistic.

And what you have out here are a setup p values and what they represent is that they represent the p values in the t statistics correspond to the hypothesis that this coefficient. So, let us take one example, so I am just on a highlight this example, so we got some coefficient for A and that is 0.2420, what we than chose to do is test the null hypothesis that this coefficient is statistically different from 0 and this p value is the probability that comes about as a result of performing this hypothesis test and as you might as you already know the p value is nothing but, the probability of seeing this data if the null hypothesis is true.

So, if the null hypothesis that that the sample coefficient is equal to 0 is true, then the probability of seeing the data that we have seen is this value, which is 5.21 times 10 to the power 17, now that is the very low number that is a very, very, very low probability. And, so the idea would be there out here we would reject the null hypothesis that this coefficient is statistically no different from 0.

So, if you take a look at this all of these p values at least are fairly low the sign 5.1 e power minus 17 just means times 10 to the power of minus 17, which means it is a very low number and the only value that is high is out here that is 0.33. And; that is the kind of value on which, you might not be able to reject the null hypothesis or reject the idea that the coefficient to be is actually in distinguishable from 0 the null hypothesis that is that this estimate this parameter could actually be equal to 0.

Going in line with the idea of confidence interval, so you also get confidence bounds, so this is the lower 95th percent confidence bound the upper 95th percent. And if you had mention values different in the original pop up screen saying I want something more than

95th percent you would have gotten the 95th percent bounds as well as those new numbers, but if you have in these numbers just reputation.

Obviously, noteworthy thing is that given the reason 95 percent confidence bounds you can reason from inferential statistics that, where ever the p value is less than 0.05 or even actually technically less than 0.1, because it is the lower and upper confidence bound if it is anyway less than 0.1 you should have the lower and upper on either side of 0. So, here the lower bound is 3.6 and the upper bound is 14, so it does not intersect 0.

So, whatever sign your coefficient is your lower and upper bound are is going to be on that side of 0. Obviously, when a p values as high as something like 0.3, which means you potentially could not reject the null hypothesis, then you would have a lower bound that could be less than 0 and you will have an upper bound greater than 0 and that should make sense in terms of how we understand confidence bounds and p values. So, this gives you the inferential statistics or the background associated with having each term in the module.

So, in this module, which I am just going to highlight with some colors, so it is clear to you in this model that you have out here you now, know, which terms could potentially go and which terms could not or which terms you can justify putting in there in which terms might just be a result of random noise. Now, in addition to these individual statistic associated with each individual term you also have an overall statistics that is being captured through this ANOVA, so I just give that a different color.

So, the idea with the ANOVA is that your trying to say how much of the variation comes from the regression model and this is the this cell is equivalent of the mean square between I am talking about d 12 and you have cell d 13, which is the idea of how much of the variation still exits even after you fitted the regression model and you use the same concepts. So, this these the these two terms are kind of like a mean square between and mean square error they are the equivalent from the ANOVA the traditional ANOVA that we study and you can calculate and f statistics and get a p value for the f statistic and what that p value says is over all as a model.

Never mind that you have the statistics associated with individual terms, but overall as a model can I test the null hypothesis that the model explains some variation or is the variation explain by the model equal to 0. So, it should give you an idea of that and in many instances you are really looking to make sure that this value is as low as possible and you want here R square and adjusted R square values to be as high as possible, so great.

So, we discuss how you can do a multiple regression in excel, but at this point we are going to also go a little bit beyond a standard multiple regression the software kind of allows you to do it conveniently, but you could also for instance perform a multiple regression by hand. And the way you would do that and this is really useful is, because you might not want to do an ordinary least squares and that is what we are going to show you how you do not have to do ordinary least squares.

Just to jog your memory the idea of ordinary least squares was to say that I want to minimize the sum of the square deviation of each point from the line as measured along the y axis line. So, I want to measure the square of the distance of each point from the fitted line and the rational for using the square was of course, to say that, that you might sometimes have data point that is above of the lines sometimes have data points that is below the lines.

So, you have positive and negative values and you want to represent all deviations in the this same light you do not want the positive and negative values to cancel each other you just want to minimize deviation. So, you said if I squared all those deviations and just minimize the sum of those squares your it is sum of the squares, because each data point deviates from the line by some amount. So, you want to minimize the sum of the square deviation of each data point from the line.

So, that is the core idea and, so now, what you might want to do is and you know doing that conception allowed for this very need derivation; such that you had of close forms solutions for beta naught and beta 1 and so on and it is very easy to implement. But, you know now, we have computers that are reasonably fast and you might not want to have these need closed form solution, but you might just be willing to take an excel sheet and do the regression through some other metric that you want to minimize.

So, instead of minimizing the square deviation of the data points from the line you might just want to take the absolute deviation, what we mean by absolute deviation is if the data point is 7 and if the fitted line at that value of x predict say 5. This difference between 7 and 5 I am just going to take it as 2 I am not going to think of it is positive or negative. So, whether the data point itself is above the line or below the line I am just going to measure the magnitude of the deviation and put them all as positive values and minimize the some of those.

So, whether you want to do that or whether you want to say my particular problem I want to penalize it by using cubes instead of squares or cubes would not really work, but I want to use the 4th power or something like that you can do it yourself. And, so this is what I call as they do it yourself regression and the idea here is that I have copy pasted the same data out here in the sheet.

And, what I am saying out here is that these are the coefficients, so this value at b 3 is beta naught this value its c 3 is beta 1 and we can just see those with some initial values I am just going to see that with 0, 0.5 you want to start with some reasonable values. And let us just put some dummy numbers out here and 0 here, so these just some dummy numbers with reasonable range and the ideas when you do that I ask the excel sheet, what is my model predicting.

So, it is essentially like saying that if these were the betas then what would my model predict for these inputs. So, if these were my betas I am just going to repeat that if everything in row 3 from b through f my betas, then given the input, which is starts at row 6 and goes on. But we will take one sample input, which is shown in row 6 from c through f given these inputs, what would my module predict and the math of that is fairly simple it is just of the it is just the math of beta naught plus beta 1 x 1 plus beta 2 x 2.

So, that is, what is I do say b 3, which is nothing but, beta naught plus the some product of these the arise an 3 and 6 and sense the word you know some product is nothing but, saying you give me to arise and I am going pair vise multiply the terms and add the molt together. So, what is doing is its doing b 3 plus c 3 times c 6, d 3 plus D 3 times d 6 and so on all the way till the end and that is was the model is predicting.

So, I have the actual value, which happens to be 21 and I have something that the module predicted and from that I can very easily calculate the residuals. So, what shown in i 6 is nothing but, the difference between the actual y and the predicted y. So, we do that and we than we could square it or we can take the absolute value, which is what I am interested in and, so I have a set of square deviation and I have a set of absolute deviations it looks like I am in general chronically over predicting, but that is just because I have given dummy betas for now.

And out here, what I have in cell o 3 is the sum of the absolutes the residuals sum of the absolutes and in o 2 I have the residual sum of squares. Now, remember in an ordinary least squares regression what you have in o 2 is what you are trying to minimize by changing the values in b 3 through f 3. So, you want to change the values in b 3 through f 3 and minimize what you getting in o 2. But, in our particular example, what we want I want to illustrate you today I want to try and minimize, what I am getting in o 3.

So, the way I would that is to again go to data and click on solve again if you do not have it in variably. means it is just need to added from add ends and I tells solver saying I want to set this objective, which is o 3 I want to minimize it. So, I can either maximize it or set try to set it to value, but I want to try and minimized and I want to do that by changing this cells in b 3 through f 3 and a given the kind of optimization approach you taking you might have to either give some constraint, which just means giving it some lower bounds and upper bounds.

So, give it something fairly reasonable, which is what I have already done here and the way you do that just go click on add and then mention this cell in give it to lower bound then give it to upper bound and that is what I have done here. And you haves a couple of different methods of solving this optimization problem I choose the evolutionary one you can play around with the others, because the I just feel like it takes time, but its I know for a fact the there is safe at its not it is in terms of computational time it is; obviously, not going to be the most effective. But, I am more than happy to pay that price this is not this particular example is not too hard problem to solve and I just click on solve.

And then, as you can see right now, excel is doing the competition, so you know it takes couple of minutes and what we would do is we will come back to this excel sheet and look at the results in a minute. So, going back to the slide, so that, so in today’s class we have seen two demos one is using excel to perform the multiple regression. And the second one is just going beyond the ordinary least squares, where you might have some other objective function that you want to minimize and you should be able to you know fairly is re do that in excel by just doing at the way have demonstrate.

Now, we come to our next topic, which is the whole idea of subset selections and it goes back on I think we have the excel results may be we will just take look at that and it says you want a keep the solve a solution and I say yes and this is it. So, and if you notice this is little different from the output that you got it is not the exact same output you know perhaps if you let the solve go on longer it mate up come up to something may be more similar, may be more superior in either case.

Because, these objective functions different you should expect different answers and this is the linear equation that solver gave you one if it wanted to minimize absolute deviation. So, going back to our a slide we spoke about, how if you in the regression analysis you saw, so; obviously, the do it yourself regression does not give you any inferential statistics it is just is way of getting coefficients.

Now, there is the other important part, which terms do you leave in the module, which you take out and when looking at that we look at this fairly simple idea of measuring each individual term through the p value of the t statistic. Just you jog your memory I am talking about this output and I am talking about these p values that you see in column e from 17 through 21 and you could have an idea saying anything below certain values acceptable anything above is certain value is acceptable and that that would be this most simplistic way of doing it.

But, the problem becomes that I mean there your using only inferential statistic to decide which gets included and which does not. And the idea is that sometimes having certain term inside, which might be statistically significant could affect other terms, so the other variables could also get affected. So, the problem is not I mean you could use inferential statistics and make one time decision, but what happens if you choose to add a variable and that changes the p value of another variable.

And also you might you might be of the opinion that what you care most about is your prediction accuracy whether terms its significant or not that you care about getting the best prediction accuracy in which, case it is still possible that adding certain terms is detrimental and removing them is might be the best. So, off end times, what we wind up doing is we wind up having some metric that measures, how could be performing as a model as a whole not as the individual term.

And we already saw that we for instance solve this solve that in the ANOVA that we have, so this is measuring the model as a whole. And we might say all I care about is this p value that you see in f 12 that is all I care about. So, can I make a decision on which, terms should be in the model and which terms should not based on this one metric you might have another metric, which says that I want to look at adjusted R square we have not talked about that at, but will we will come back to that, but you could have this one metric that you want to minimize.

And as we get more advance you look you will we will realize how there are many other criteria, which we could use to say, what is a good regression and what is a better and have a standard frame work to compare two different models. Now, assume that you have the standard frame work or for instance assume that you care about f 12 you care about minimizing f 12, which is a p value associated with the f statistic associated with the whole model.

Now, it is perfectly possible that if you choose to add some terms that your overall model becomes better. And if you choose to remove some terms of, so may be if I removed d as an input variable as a whole I might get a better p value or it is also possible that if I added another variable e it could b the a lower p value. But, how do you go about searching through the entire possible set of input variables to figure out, which ones should be in there in which, one should not.

And this becomes very important question especially in light of one what we have already discussed, which is prediction accuracy I mean. So, imagine a problem that you faced with, which has about 20 input variables and one output variable, which subset of the 20 and I am including that set, which is all 20 and I am including that set, which is none of the 20 the null set, but which subset of that of those 20 variables should be there in the model.

So, as to give you the best performance, where performances measured for instance to the p value of the f statistics that is one side of a this the other side also, which is that if you had a huge set of variables sometimes the interpretation of the model might not be, so clear. Because, these variables might be related to each other and the coefficient that one of them takes up would compromise the coefficient that the other takes up and so on.

And it would be much easier you know if you could interpret the model with just the smaller set of variables at least the ones that are the strongest and in some sense that you know enables you to get the big picture of the entire process. So, what we are going to talk about now, is that process of choosing that subset of variables, which is the best in terms of creating a model. The goal standard in doing that is this process called, let me just make that full screen the goal standard of doing that is this process called best subsets regression best subsets regression basically says.

So, you have 20 variables or let us say I had ten input variables let us try every subset of those 10 variables. So, you will start, where having none of the variables or and the end you will move you will have 10 c 1 combinations of having one variable, which essentially means having variable 1 variable and that is one module having variable 2 that is one model. And then, we will move to the tool the 10 c 2 combination of two input variables are in the model I have 1 and 2 in the model I have 1 and 3 in the model I have 2 and 3 in the model I have 4 and 5.

So, all the combination of two variables and then, you will have all the combinations of threes and you will go all the way to having all the 10 variables in the model that is a lot. So, that is 10 c 0 plus 10 c 1 plus 10 c 2 plus 10 c 3 combinations, but you know what with good computational speed and reasonably handle able problem you might just be able to use of brood for approach and look at every subset that is possible and your done with it.

So, best subset regression of in its kind of seen as the goal standard of doing it and it allows you to you know evaluate every subset. So, it is essentially like creating that many different regression you possibly cant manually do it in excel, but what you can do is you might be able to do it in another language like you might it a little bit of programming document we are able to do it in other software.

But, essentially that whole process of doing tools data analysis choosing a regression and getting that new output sheet you would do that for various possible combinations of the input variables for every possible subset of the input variable and that is called best subsets selection. Now, when that is not possible we tend to go towards the sequential methods and the problem with that is that there is a certain impact of adding and removing variables on each other. So, I can never say that I am going to add this variable and its going to cause this effect.

And then, choose to remove another variable at a certain different point of time, but the effect of adding or removing a variable impacts other variables and for that reason no form of a sequential approach is mathematically going to be exactly the same we able to always re create a best subset selection. But, for the greater part these methods that are sequential just work are just find and in that light I would say the more popular approaches are often called step wise regression.

You have forward step wise selection, which means you kind of sequentially start adding individual variables to your model and you keep adding till a point, where the model does not improve any further in fact, the gets was and any take a step back in say this was the best stay. You have back word step wise selection where you start with all 10 input variables in your model in a sequentially removing them. So, for instance one way to kind of show you how you could that in excel is to say.

So, I had all A, B, C, D in the model now d is a worst I am going to remove that and then, say does that improve my p value of the f statistic in f 12. So, it would literally be like repeating the data analysis going back to the data going back to data analysis saying I want to do a regression, but this time please do not include d in my range, so that is what I am going to do I am going to include the label. So, it is now, I have done that again and now I have a different analysis and low be whole this p value is much lower than the p value of the previous analysis.

So, may be d should have been remove and to kind of do this kind of a process sequentially removing variables and getting to the point, where removing another variable is not good thing and you could use for instance the p value of the individual variables to tell you in which order to remove the variables and so on. Now, again this is fine as long as you have a reasonable number of variables if you have 50 an odd variables it becomes a very cumbersome.

So, you could just essentially use software package to do the same thing and we will briefly demo that. But, before we do that you finally, have something called high bridge step wise which essentially is a combination of forwards and backwards step wise you start with nothing in your model in a sequentially keep adding terms. But, it each state you see whether which term to add next or which term to delete next and you put them on a common platter and make a decision on which to try.

So, that is hybrid step wise and what we are going to do now, is demonstrate how you can actually do hybrid step wise selection in mat lab this is the mat lab screen if you have not seen it this is the version our 2013 b the path that I am selecting essentially is called a command window and that is a place, where you can just type instructions and get outputs and we are primarily going to be dealing only with that place.

So, what I have done is I already done in this, but I just show you what I did, which is I just selected the data and I just copy pasted it in mat lab. So, I just to copy this I went to mat lab and I said oh by the way this variable y is equal to open square bracket paste the data close square bracket enter and mat lab is created a variable for me called y with the data and it the same process for x and all I am going to do now is present mat lab the with the simple command call step wise and I will given x comma y as. If you noticed prompted me in terms of what do already put in there.

But, you can also provide mat lab saying I want some terms mandatory they need to be in the model and you can tell that and you can also give a criteria for entering and removing terms. Just like for instance you could have used some criteria out here to look at these p value is to decide, which order to go with mat lab you can essentially say if its below certain range I wanted to go on and if its above a certain range I wanted to come out the default I believe this 0.05 and 0.1.

But, we are not going to give any of those terms we are going to keep it simple and just hit n and what happens is mat lab look at this data and it essentially comes up with an interface like this and anything that is red basically means is not in the model. So, now this is nothing in the model and this is the intercept and there is no f statistic, because is nothing in the model and so on. But, mat lab gives you a suggestion what to do a next its has move x 3 n.

So, all you will do is you will go to x 3 and you will click on it and its moved in and immediately you notice that you have a f statistic you have a p value in, then says please move x 1 in see you go to the red dot and click on x 1 and that is moved in it looks like the p value even going to lower then it as x 2 n and you still doing better and now, it says do not move any more terms in. So, we went through a case, where the sequentially said add a add b add c and you know smartly it said do not add anything more and do not do anything more.

But, you could see situation, where it will ask you to then add something in then, go back and delete something, but it kind of not only does it give you the interface do it. If, so you could for instance add d and c was the performance how the performance changes and you can see performance becomes was after adding d. So, it is 1.76 if I take the d out again it is the p value is even more. So, essentially mat lab not just I mean it allows you to do whatever you want, but it also gives you a recommendation by telling you what the next steps are.

And at some points if you say export it will export the model for you, but you can also visually see it its basically saying this is the coefficient of x 1 a this is the coefficient of b this is the coefficient of c that that you can see under the column co f and it is also tell you what the intercept is it will tell you what the R square is and it will tell you the route means squared errors and so on. So, it is the very convenient tool to clear around with if you really had like lost of variables and you just wanted to play around more than do like brute force approach to usually c, what happens if you add one variable and remove a variable this could be very convenient great.

So, with that we conclude our lecture on kind of showing you how to perform a regression and interpreting results through software as well as tackling the problem of variable selection or subset selection.

Thank you.