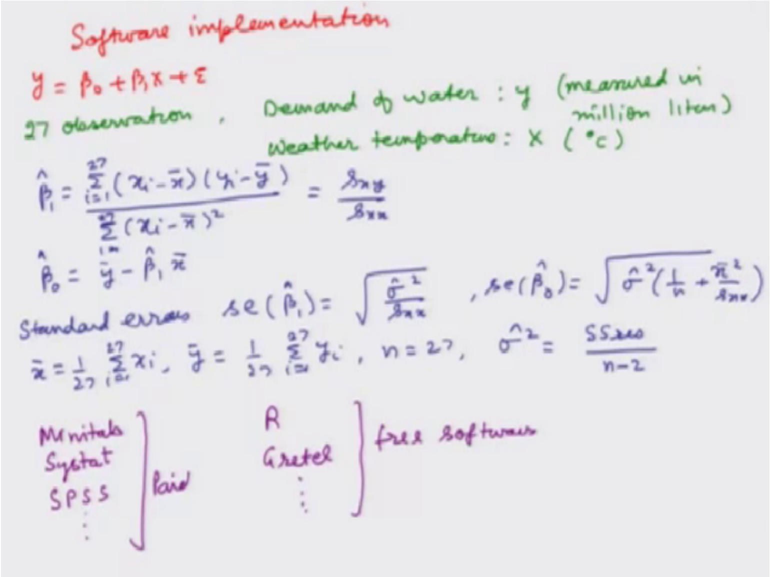
Regression Analysis and Forecasting

Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 10
Software Implementation in Simple Linear Regression Model using MINITAB

Welcome to the lecture you may recall that in the earlier lectures we had discussed the simple linear regression model and we had found the estimates of the regression coefficients there standard errors, confidence intervals and we have discussed several other aspects. Now in this lecture I would like to show you that how these values are obtained in software. In this lecture we are going to discuss about the software implementation.

(Refer Slide Time: 00:51)



Software implementation

$y = \beta_0 + \beta_1 x + \varepsilon$

27 observation , Demand of water : y (measured in million liters)

Weather temperature : X (°c)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{27}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{27}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Standard errors $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$, $se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}$

$\bar{x} = \frac{1}{27}\sum_{i=1}^{27} x_i$, $\bar{y} = \frac{1}{27}\sum_{i=1}^{27} y_i$, $n = 27$, $\hat{\sigma}^2 = \frac{SS_{res}}{n-2}$

Minitab
Systat
SPSS
⋮
Paid

R
Gretel
⋮
free software

So just for the sake of illustration I have a considered here a simple linear regression model y = beta0 + beta1 x epsilon so here we are assumed that we have got 27 observations, and these observations are collected on the demand of water and that is going to denote our response variable y, and this is measured in million litres.

We know that the demand of water depends on the weather temperature, so we take x as our independent variable which is denoting the weather temperature on a particular day and this is measured in degree centigrade's. Now our objective is that we want to find out here the value of beta one hat which is given by summation i goes now 1 to 27 xi - x bar, yi - y bar upon summation.

I goes from one to twenty seven x i minus x bar whole square and that we had denoted as a something like sxy over sxx we also had obtained the estimated of intercept term as y bar – beta1 hat x bar after this we have obtain the standard errors of beta1 hat, this was obtain as sigma hat square over sxx and standard errors of beta0 hat this was obtained as sigma hat square 1over n + x bar square over sxx.

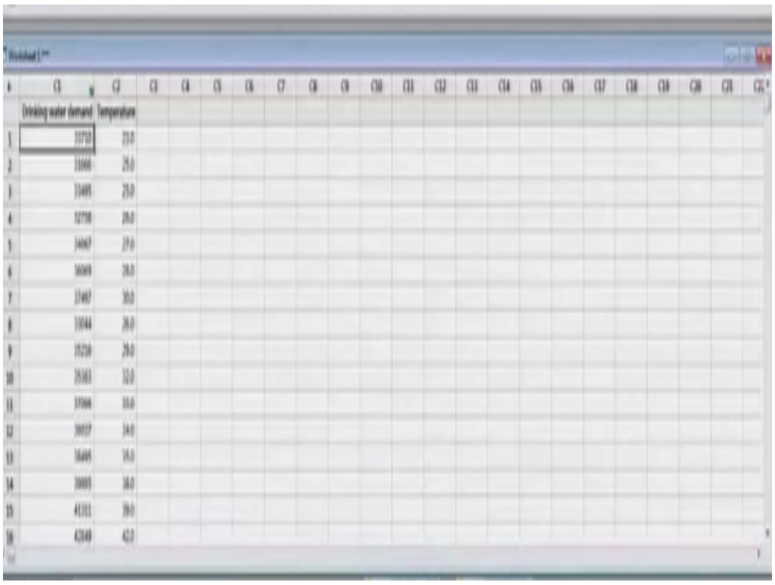
Here you will see that x bar is going to be one over 27 i goes from 1 to 27 x i y bar is 1 ove-r 27 summation i goes from 1 to 27, y i and n here is obviously 27 and sigma square hat is obtained as sum of square due to residuals divided by n - 2 that is 25. Now there are different types of statistical software which can be used to conduct the statistical analysis.

There is some software which are paid software and some popular software's are Minitab, Systat, SPSS and there is a long list but these all are paid softwares. There is some free software also like as one popular software is R this can be downloaded from the site www.r-project.org and similarly there is another software this is Gretel. There are some other software also available and these are essentially the free softwares.

So one can actually use any one of the software whatever is easily available, here in this lecture we are going to use the Minitab, there is no bias or personal reason to use the Minitab because it is available, so I am going to use this Minitab. So now let as come to the software part and my objective is that first I am going to demonstrate that how I am going to analyse my data and then I will try to obtain the regression output.

One thing which we all have to keep in mind that we have not completed the regression analysis so when we go for the complete regression analysis there are some parts in the statistical software outcome which we have not yet studied. So I am going to consider the outcome of software which we have done and we are going to demonstrate here that how these quantities which we have studied in the early lectures they are going to be computed by the software and how we can interpret them.

(Refer Slide Time: 6:21)



Worksheet 1 ***

	C1	C2
	Drinking water demand	Temperature
1	33710	23.0
2	31666	25.0
3	33495	23.0
4	32758	28.0
5	34067	27.0
6	36069	28.0
7	37497	30.0
8	33044	26.0
9	35216	29.0
10	35383	32.0
11	37066	33.0
12	39037	34.0
13	36495	35.0
14	39895	36.0
15	41311	39.0
16	42849	42.0

So let us come to the software outcome, this is the window of the Minitab statistical software, so here you can see that here in the first column I have already enter the data on the drinking water demand and in the second column I already have enter the data on the temperature, so you can see here that there are all together twenty seven observations, so this drinking water demand is going to be our response variable and temperature is going to be here the independent variable.

So first of all when we want to conduct a regression analysis first we have to ensure that the relationship between response variable and independent variable or the explanatory variable is showing a linear trend So for that thing this Minitab have several options similarly other software also have this options and we go for a scatter plot and here you can see that there are different types of scatter plots available.

First we are going for a simple scatter diagram and in this case I am considering the y as my drinking water demand and x variable as my temperature and I try to plot this graph, so you can see here that there seems to be a sort of linear trend which is here. So I will keep this graph over here and let us try to explore something more and later on I will come back to the interpretation of this graphics and software outcome.

Now similarly I try to make here another scatter diagram which is possible in Minitab, and I try to make a scatter diagram with regression line because this Minitab software actually computes the regression line and it shows the regression line on this scatter diagram itself.

So here also we have the same set up and we have try to get this graph, so you can see here with this red line it is trying to shows a fitted regression line so this type of elementary analysis gives us a confidence that well we can fit linear regression model.

Now after this we go for the regression analysis in Minitab there is built in function for this regression analysis and we go for regression and fit a regression model. Once you come here you can see that I have given here the response variable as drinking water demand and this continuous predictor that is the independent variable or the explanatory variable this has been given here.

Now you can see here that there are several options here, but at this moment I am considering any one of them, you can see here we can give the widths and but I want to show you here that here I have given the confidence level for all confidence interval as 95, that means my level of significance that is alpha is 5% and corresponding confidence interval are of 95% confidence coefficient.

We are also going to conduct the test of hypothesis so you see there are options for one-sided test of hypothesis and two sided test of hypothesis so we are selecting here the two-sided confident interval and two-sided test of hypothesis here we can see here there are two types of sum of square one is adjusted type three and sequential type one. At this moment I can simply tell you that we are going for a simple sum of a square that we have studied in over

Lectures and this is given by adjusted type three sum of this square. Now you can see here that there are different graphs this can be plotted simultaneously, but at this moment we not done this part so I am not doing it. Similarly the results what we want to have I am going to consider here only three aspects that I would like to have the outcome from analysis of variance still we have not done.

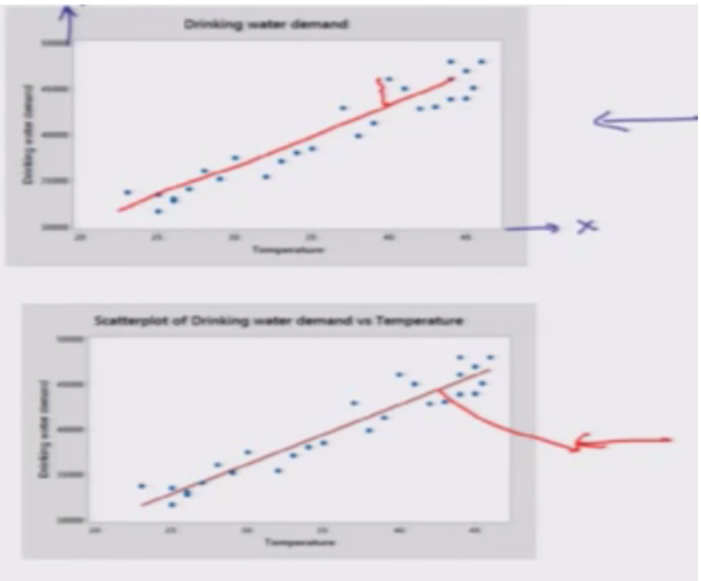
So that we are going to do in the case of multiple linear regression model from the next lecture, but still we will need this thing. So I am considering it and then I want to compute the coefficients that will give as regression coefficient estimators say beta0 hat beta1 hat and here I would also like to have a regression equation.

There are different types of other thing like as method, model summary, fits and diagnostic, Dublin-Watson test these are a thing that we are going to consider later on when we try to do the multiple linear regression model. If you want you can store these different types Fits, Residuals, Standardized residuals, Leverage, Cook's distance, DFITS, Coefficients, Design Matrix, but at this moment we don't need it we will do it later on.

Now let me go for this the outcome of this linear regression analysis, so you can see here we have got here this outcome. So if you use any statistical software any standard statistical software they are going to give a similar output, there can be small difference that the analysis of variance scan come later and the regression coefficient scan come first and so on there may be some minor differences in the different types of diagnostic test.

That are provided by the software outcome, but more or less this outcome remains almost the same. So what we try to do here, let me try to copy this statistical outcome and these two graphics and so that we can understand them in more detail. So I already have copied these thing here,

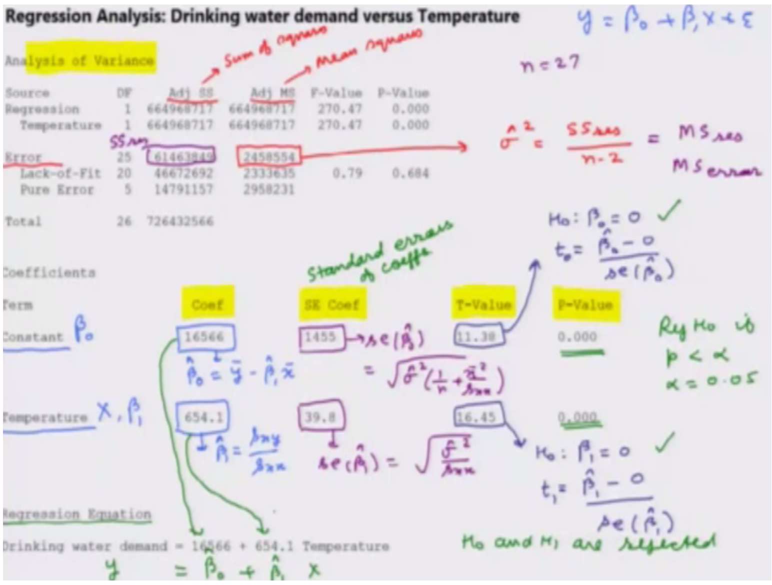
(Refer Slide Time: 13:00)



So now first let as try to understand this graphic, you can see here on this axis we have here x axis and on this axis we have here y axis and this sort of this scatter diagram this is trying to indicate, yes there can be a sort of linear trend here. Now you can see that this type of line can fitted and they are going to indicate the residuals. So now when we try to look at this diagram this is also the same diagram.

But in this case we have fitted this regression line, so this is giving as sort of confidence, yes we can go head and we can conduct a simple linear regression analysis of this data.

(Refer Slide Time: 14:01)



Regression Analysis: Drinking water demand versus Temperature

```
Analysis of Variance

Source          DF     Adj SS     Adj MS  F-Value  P-Value
Regression       1  664968717  664968717   270.47    0.000
  Temperature    1  664968717  664968717   270.47    0.000
Error           25   61463849    2458554
  Lack-of-Fit   20   46672692    2333635     0.79    0.684
  Pure Error     5   14791157    2958231
Total           26  726432566

Coefficients

Term            Coef  SE Coef  T-Value  P-Value
Constant       16566     1455    11.38    0.000
Temperature    654.1     39.8    16.45    0.000

Regression Equation

Drinking water demand = 16566 + 654.1 Temperature
```

$y = \beta_0 + \beta_1 x + \varepsilon$
Sum of squares
Mean squares
$n = 27$
SS_{res}
$\hat{\sigma}^2 = \frac{SS_{res}}{n-2} = MS_{res}$
MS_{error}
$H_0: \beta_0 = 0$ ✓
$t_0 = \frac{\hat{\beta}_0 - 0}{se(\hat{\beta}_0)}$
Standard errors of coeffs
β_0
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$se(\hat{\beta}_0)$
$= \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}$
Rej H_0 if
$p < \alpha$
$\alpha = 0.05$
x, β_1
$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$
$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}$
$H_0: \beta_1 = 0$ ✓
$t_1 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$
H_0 and H_1 are rejected
$y = \hat{\beta}_0 + \hat{\beta}_1 x$

Now I have copied the same outcome that was obtained as an outcome here, so now here is the outcome of the regression analysis that was provided by the Minitab software so fist of all you can see here that the first thing is given by here, this is analysis of variance, but at this moment we have not yet to study the analysis of variance, but I had consider it because I wanted to show you, that if you try to look at this column this line here error.

Here you will see here this is trying to give us 2 value 1 is adjusted sum of square and this is adjusted MS that is adjusted mean square, so SS is mean here sum of squares and MS means here mean square so this is the value which we have to consider, actually this is the value which is our sigma square hat, and sigma square hat if you remember we have defined by some of this square due to residual divided by degrees of freedom that was n - 2.

So here this SS residual this is given by this thing, this is our sum of square due to residual, and number of observations here there are 27 observations, so some of square residual divided by 25 this gives as this thing, so in the statistical software where ever we have mean square due to residual are sometime in the software this is given as mean square due to error this gives us the value of sigma square hat.

Because we had just twenty seven observation on x i and y i so we do not know the value of sigma square, so we are going for the option where we estimated the value of sigma square on the basis of given sample of data and that was estimated as sigma square hat. The idea of considering this analysis of variance table was simply to show you that how the value of sigma square hat is obtain from the software outcome.

So now after this we come to this column over here this coefficient, now if you try to see here we have one thing here which is constant. Constant is nothing but your beta0, we have consider the model here y is equal to beta0 + beta1 x + epsilon. So beta0 is your constant and similarly temperature this variable is here your x variable and corresponding to which the regression coefficient is beta1.

And this value here this is the value of beta1 hat which was sxy over sxx.This is the value of beta0 hat, which was obtained as y bar – beta1 hat x bar, now we try to consider this column, this is SE Coef, this means SE means, standard error and this column gives the standard errors of coefficients. Coefficients are beta 0 and say beta1 whose estimators are beta0 hat and beta1 hat.

This quantity is giving us the value of standard error of beta0 hat and this if you remember we had obtained as sigma square hat one over n + x bar square over sxx. Similarly this is standard errors of beta one hat and this we have obtained if you remember has sigma square hat over sxx. Next we come to the third column which is here t value and with the t value you can look at this value.

This value is corresponding to the test of hypothesis h naught, beta0 =0. We had conducted at test of hypothesis for h naught beta0 = beta0, 0 where beta0, 0 is some known value, so here we are considering it to be 0. So this essentially the value of the t statistics that was obtained has beta hat0 -, 0 divided by standard error of beta hat 0.

Similarly if you look at this value this is corresponding to the null hypothesis h naught beta 1 =0 you may recall that we have developed test of hypothesis for h naught beta0 = beta1naught here we are assuming that beta1 naught =0, so this value is coming from this statistics. Let me write down here t0 in our symbolic notation and this will be t one, so beta1 hat - 0 divided by standard error of beta 1 hat.

Now we come to the fourth column which is here p value now you may recall that while doing the test of hypothesis we had discussed how to take a decision whether to accept the null hypothesis or not. So one option was to look on the t tables and find out the critical value and second options was to look into the p value, and if you remember our rule was that reject h naught if p value is smaller than alpha. Alpha is the level of significance and that you can recall that in Minitab software we had given it to be 5%

In our case alpha =0 point zero five and here you can see that is p value is coming out to be very, very close to 0 and in this case also this p value is coming out to be very, very close to 0, beta0 =0and h naught beta1 =0, so now we can conclude the both h naught and h one are rejected. So based on this I can consult that s is the variable which is important variable and that is helping us in explaining the variance n y.

Now in the end you will see that we have here a regression equation and this regression equation as been obtained based on the coefficients which are obtained earlier and so drinking water demand this is here y = and because is here beta naught hat and this is here beta1 hat which we have obtained here times x. So you can see here this value is coming from here and this value coming from here.

So based on that now we have obtained a fitted linear regression model. There are some other options that to compute the confidence interval also so they can also be obtain and based on that we can also find the confidence intervals0 and beta1 but at this moment since we are since moment since we are simply starting, so we are not considering it here and but looking at the outcome of this software you can very easily assume that you can find out or you can identify the values of lower and upper limits of the confidence interval.

So by this small introduction to the software I have tried to highlight that how to read the software outcome and based on that we can modified our model. From the next lecture we are going to consider the multiple linear regression model which is more usual in practice there will be some more issues, but believe me whatever we have learnt in the case of simple linear regression model that is going to create the foundation for learning the multiple linear regression model, so we stop here and till then good bye.