

Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 36
Tutorial – I

Today, we will be solving some problem from simple linear regression model, that is the first topic we talked about.

(Refer Slide Time: 00:37)

PROBLEM 1

A Study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	9	7	10	8	9	13	14	13	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1 ? What is the fitted equation?
2. Construct the ANOVA table and test the hypothesis $H_0: \beta_1 = 0$ with $\alpha = 0.05$
3. What are the confidence limits ($\alpha = 0.05$) for β_1 ?
4. What are the confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X = 3$?

And here is one problem from simple linear regression model, a study was made on the effect of temperature on the yield of a chemical process. The following data were collected in coded form, so this is the X stands for the temperature and Y is the yield of a chemical process. So, Y is the response variable and X is a regressor variable and we want to fit a simple linear regression model here, so we are given 11 observations here.

So, this is quite straight forward, the first question is, assuming a model like Y equal to beta naught plus beta 1 X plus epsilon. What are the least squared estimates of the regression coefficient beta naught and beta 1 and what is the fitted equation. We have solved similar problems while we are talking about the simple linear regression model. And then the second question is construct the ANOVA table and test the hypothesis that beta 1 equal to 0 with level of significance 0.05 and then what are the confidence limit for beta 1.

And the fourth question is, what are the confidence limit for the true mean value of Y, when X equal to 3. Let me start with the first one, so we are given these observations, x_i y_i , for i equal to 1 to 11. And first we will be fitting a simple linear regression model using the least square technique.

(Refer Slide Time: 03:21)

1. $(x_i, y_i), i=(1)11$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad S = \sum_{i=1}^{11} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2, \sum x_i y_i$$

$$= \frac{158}{110} = 1.44$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{102}{11} = 9.27$$

$$\hat{y}_i = 9.27 + 1.44 x_i$$

So, what we are given is that, we are given x_i and y_i , for i equal to 1 to 11 and you want to fit a model like y equal to or y_i equal to β_0 plus $\beta_1 x_i$ plus ϵ_i . So, we know that, this β_0 and β_1 , they are obtained by minimizing the least square function that is S , which is equal to y_i minus β_0 hat minus β_1 hat x_i . So, this is the i th residual and by minimizing this one for i equal to 1 to 11, we get the least square estimate of the regression coefficient β_0 and β_1 .

So, β_1 hat, we know that this is equal to S_{xy} , please refer my first topic simple linear regression, so S_{xx} , so this can be also written as summation $x_i y_i$ minus $n \bar{x} \bar{y}$. So, here n is equal to 11, by sum over x_i square minus $n \bar{x}$ square. So, you are given $x_i y_i$ for i equal to 1 to n , so what you do is that, best thing is that, you compute sum over x_i , summation y_i , summation x_i square, summation y_i square and also the product $x_i y_i$ then you are done.

So, you can compute all these things for the given observations and then you can check that, this one is equal to 158 by 110, which is equal to 1.44. And β_0 hat is equal to \bar{y} minus β_1 hat \bar{x} and you can check that, this one is 102 by 11 that is, 9.27.

So, we are done with the first problem, so the fitted equation is \hat{y}_i is equal to 9.27 plus $1.44 \times x_i$, so this is the fitted model for the given problem.

(Refer Slide Time: 07:03)

PROBLEM 1

A Study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1 ? What is the fitted equation?
2. Construct the ANOVA table and test the hypothesis $H_0: \beta_1 = 0$ with $\alpha = 0.05$
3. What are the confidence limits ($\alpha = 0.05$) for β_1 ?
4. What are the confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X = 3$?

So, the next problem is, it says that, construct the ANOVA table and then test the hypothesis that, H_0 that beta 1 equal to 0 at level of significance 0.05.

(Refer Slide Time: 07:27)

2. **ANOVA TABLE**

Source of variation	df	SS	MS	F
Reg.	1	226.94	226.94	$F = \frac{MS_{Reg}}{MS_{Res}} = 96.17$
Residual	9	22.23	2.36	
Total	10	248.18		

$SS_T = \sum_{i=1}^{11} (y_i - \bar{y})^2 = 248.18$
 $SS_{Reg} = \hat{\beta}_1^2 S_{xx} = \left(\frac{154}{110}\right)^2 110 = 226.94$
 $= \sum_{i=1}^{11} e_i^2$

$e_i = y_i - \hat{y}_i, i = 1(1)11$
 $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
 $F = 96.17 > F_{0.05, 1, 9} = 5.12$

So now, construct the ANOVA table, so source of variation, degree of freedom, sum of square, MS and the F statistics. So, the source total variation that is, SS_T , so SS_T is equal to summation y_i minus \bar{y} square, i is from 1 to 11 and you can check that, this

is equal to 248.18, so the SS total is 248.18. Now, it has, what is the degree of freedom, degree of freedom is 10, because you know that, this $y_i - \bar{y}$, they satisfy a constraint like summation $y_i - \bar{y}$ is equal to 0.

So, there is a one constraint here, so that is why, one degree of freedom less here, now you can compute SS regression. SS regression is $\beta_1^2 S_{xx}$ and we know that, β_1 is 158 by 110 square and S_{xx} is 110, so this is equal to 226.94. So, this is regression here and SS regression is 226.94 and the degree of freedom here, see SS regression is also equal to e_i^2 , from i equal to 1 to 11. So, this is another way to compute, you have the fitted model, you know \hat{y}_i , you have the original observation y_i , so you can compute e_i , for i equal to 1 to 11.

But, you do not have the freedom of choosing all the e_i 's independently, because there are two constraints on e_i . So, you can choose 9 residuals I mean, you have the freedom of choosing 9 residuals and remaining two have to be chosen in such a way that, those two restrictions are satisfied. So, the degree of freedom for SS regression is equal to 9 and then the remaining part, the part which remain unexplained, the part of variability that is, SS residual and that has degree of freedom 1.

And the SS is obtained by SS total minus SS regression that is, 22.23 and the MS value is... I said the residual degree of freedom is 9, so this is the residual degree of freedom and the regression degree of freedom is equal to 1. So, the MS residual is SS residual by the degree of freedom that is, 2.36 and the MS regression is SS regression by degree of freedom that is, 226.94. So, the F statistics is MS regression by MS residual, which is equal to 96.17.

Now, you know that, we can test this hypothesis say H_0 that is, β_1 equal to 0 against H_1 , that β_1 is not equal to 0 using this F statistics and the observed value of F is equal to 96.17 and this has degree of freedom 1, 9. So, you check the tabulated value of F 0.05, 1, 9, so that is equal to 5.12. So, the observed F is greater than the tabulated F.

(Refer Slide Time: 13:56)

Reject $H_0: \beta_1 = 0$

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$$

under $H_0: \beta_1 = 0$

$$\sim t_{n-2}$$
$$= \frac{1.44}{\sqrt{\frac{2.36}{110}}} = 9.83 > t_{0.05, 9} = 1.833$$

we reject $H_0: \beta_1 = 0$

So, the conclusion is that, we reject H_0 that is, β_1 is equal to 0 is rejected that means, there is linear relationship between Y and X. So, the next problem is, what are the confidence limit for β_1 , so we have a point estimation for β_1 , now we will find the confidence limit for β_1 . So, before doing that, may be you know already, but I want to say the another way to test this hypothesis, that $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$.

This can be tested also using the t statistic that is, $T = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$, I hope you know all these things. So, this is the t statistic under the null hypothesis, that β_1 is equal to 0 and this follows t distribution with the degree of freedom $n - 2$. So, here it is 9 and then you can check that, this t value is $\hat{\beta}_1$ is 1.44 and MS residual is 2.36 and S_{xx} is 110 and you can check that, this value is 9.83. And now, you look at the tabulated value of t that is, $t_{0.05, 9}$, that is equal to 1.833.

So, again the I mean, of course you will get the same result. Whether you use F statistic for testing the hypothesis or you use the t statistic for testing the hypothesis, result will be the same. And also you know, in fact $F = t^2$ under the null hypothesis. So, here again the observed value is greater than the tabulated value, so we reject H_0 , that β_1 is equal to 0, so next we will go for the third problem.

(Refer Slide Time: 16:52)

PROBLEM 1

A Study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1 ? what is the fitted equation?
2. Construct the ANOVA table and test the hypothesis $H_0: \beta_1 = 0$ with $\alpha = 0.05$
3. what are the Confidence limits ($\alpha = 0.05$) for β_1 ?
4. what are the Confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X = 3$?

What are the confidence limit for beta 1 at 0.05 level of significance.

(Refer Slide Time: 17:10)

3. what are the Confidence limits ($\alpha = 0.05$) for β_1 ?

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$$

$$P_{\gamma} \left\{ -t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq t_{\alpha/2, n-2} \right\} = 1 - \alpha$$

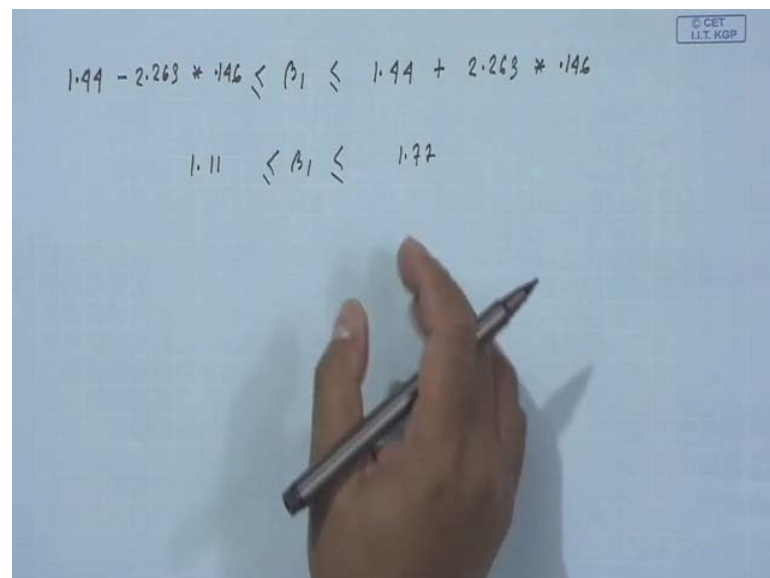
$$\hat{\beta}_1 - t_{\alpha/2, n-2} * \sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} * \sqrt{\frac{MS_{Res}}{S_{xx}}}$$

So, for this one, the question is, what are the confidence limits at alpha equal to 0.05 for beta 1. So, what we know is that, we know that this beta 1 hat, which is a linear combination of y_i and x_i follows normal distribution, so linear combination of normal distribution is again normal. So, this follows normal distribution with parameter with mean beta 1 and variance sigma square S xx. So, I can write this one as, again beta 1 hat minus beta 1 by sigma square by S xx, this follows normal 0 1.

But, what happens is that, this sigma square, the population variance is usually unknown, so we need to estimate this one, we estimate this one by MS residual. And once you replace this sigma square by MS residual, this follows t distribution, so beta 1 hat minus beta 1 by MS residual S_{xx} , this follows t distribution with degree of freedom n minus 2. And from here, I can say that, beta 1 hat minus beta 1 by MS residual S_{xx} , this less than equal to t alpha by 2 n minus 2 and greater than t alpha by 2 n minus 2 degree of freedom, of course minus.

This has probability 1 minus alpha that is, 0.95 and from here, we get 95 percent confidence limit for beta 1. So, this can be written as, so the beta 1 from here, beta 1 less than equal to beta 1 hat plus t alpha by 2 n minus 2 and of course, multiplied by this thing, MS residual by S_{xx} . And here, beta 1 hat minus t alpha by 2 degree of freedom n minus 2 MS residual by S_{xx} , again this has probability 1 minus alpha. So, this is the lower limit for beta 1 and this is the upper limit for beta 1. Now, we know everything, we know what is the beta 1 hat, we can find this tabulated value. This is t 0.25, because alpha is 0.05 and we know all these values.

(Refer Slide Time: 21:30)



The image shows a hand holding a pen, writing mathematical equations on a whiteboard. The equations are:

$$1.44 - 2.263 * 0.146 \leq \beta_1 \leq 1.44 + 2.263 * 0.146$$

$$1.11 \leq \beta_1 \leq 1.77$$

In the top right corner of the whiteboard, there is a small logo that reads "© CET I.I.T. KGP".

So, you can check that, finally this beta 1 is, let me put beta 1 hat here, so this is 1.44 and t value is 2.263 and the standard error of beta 1 hat is 0.146. And similarly, here 1.44 minus 2.263 into 0.146 and here is the limit for beta 1, it is 1.77 and 1.11.

(Refer Slide Time: 22:22)

PROBLEM 1

A Study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1 ? what is the fitted equation?
2. Construct the ANOVA table and test the hypothesis $H_0: \beta_1 = 0$ with $\alpha = 0.05$
3. what are the confidence limits ($\alpha = 0.05$) for β_1 ?
4. what are the confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X=3$?
 $E(Y \text{ at } X=3) = E(Y | X=3)$

So, till now, these problems we have already discussed in the module, I think even 4 also, it says that, the fourth problem is that, what are the confidence limit for the true mean value of Y, when X is equal to 3. So, what does this mean, we have to find the confidence limit for the true mean value that is, mean of Y or expected value of the response variable at X equal to 3.

And in the first module in simple linear regression model, we denoted this one by expectation of Y given X equal to 3. I mean, maybe we should not use this notation, because X is not random variable, but both are same, this is what I want to say here. So, we have to find the confidence interval for this mean value at X equal to 3, so how to do that.

(Refer Slide Time: 23:48)

$E(Y|X=3)$ $y = \beta_0 + \beta_1 x + \epsilon$ © CET I.I.T.KGP
 95% confidence interval for $E(Y \text{ at } X=x_0) = \beta_0 + \beta_1 x_0$
 An unbiased estimator $E(Y \text{ at } X=x_0)$ is $\hat{\beta}_0 + \hat{\beta}_1 x_0$
 $\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 x_0) - (\beta_0 + \beta_1 x_0)}{\sqrt{MS_{Res} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

 $P_{\alpha} \left\{ -t_{\alpha/2, n-2} \leq A \leq t_{\alpha/2, n-2} \right\} = 1 - \alpha$

So, what we want is that, we want confidence limit for E Y at X equal to 3, so let me write this thing, instead of X 3, I will write that, you are looking for say, 95 percent confidence interval for mean value of Y at say, X equal to x naught. So, x naught is nothing but 3 that I will plug at the end, let me solve this (Refer Time: 24:41) for x naught. So, this is nothing but beta naught plus beta 1 x naught, because we have considered the model y equal to beta naught plus beta 1 x plus epsilon.

So, the expected value of y at x naught is this one, because expectation of epsilon is equal to 0. And first what we will do is, we will find an unbiased estimator for this one, this at X equal to x naught that means, unbiased estimator of beta naught plus beta 1 x naught is nothing but beta naught hat plus beta 1 hat x naught, because beta naught hat is an unbiased estimator of beta naught and beta 1 hat is an unbiased estimator of beta 1.

So, this one, so the unbiased estimator beta naught hat plus beta 1 hat x naught, this is point estimator for the expected mean at X equal to x not and we are looking for a confidence interval for this one. So, this follows normal distribution with mean beta naught plus beta 1 x naught and you can check that, this has variance sigma square 1 by n plus x naught minus x bar whole square by S xx. So, of course then this minus the mean by square root of this, follows normal 0 1 and if you replace this sigma square by MS residual then that follows t estimation, so let me write that only.

So, what I can do now, that beta naught hat plus beta 1 hat x naught minus the mean beta naught plus beta one x naught and we are looking for a confidence interval for this one, this by MS residual 1 by n plus x naught minus x bar whole square S xx square root. This follows t distribution with the degree of freedom n minus 2 and from here, you know now of course, you know it, so the whole thing let me write call it say, A. So, A less than equal to t alpha by 2 n minus 2 minus t alpha by 2 n minus 2, this has probability 1 minus alpha where, A is nothing but this variable. So, from here, we will get confidence interval for beta naught plus beta 1 x naught, which is the mean response at the point X equal to x not.

(Refer Slide Time: 29:00)

Handwritten mathematical derivation on a blue background showing the confidence interval for the mean response at a point x_0 . The derivation starts with the point estimate and its standard error, then uses a t-distribution to form the confidence interval. The final result is $12.15 \leq \beta_0 + \beta_1 x_0 \leq 15.03$.

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 - \left[\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \sqrt{MS_{Res} \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \leq 9.27 + 2.262 \sqrt{2.36 \left(\frac{1}{11} + \frac{9^2}{110} \right)} + 1.44 \times 3$$

$$12.15 \leq \hat{\beta}_0 + \hat{\beta}_1 x_0 \leq \cancel{15.03} + 15.03$$

So, what we will get from there is that, beta naught plus beta 1 x naught is less than beta naught plus beta 1 hat x naught plus t alpha by 2 n minus 2. And then the standard error of this one that is, MS residual 1 by n plus x naught minus x bar whole square S xx. And similarly here, the same thing, beta naught hat plus beta 1 hat x naught minus this one that is, t alpha by 2 n minus 2, it is quite straight forward thing, MS residual 1 by n plus x naught minus x bar whole square by S xx, so this will go here well, so you are done.

So, you know everything here whatever you need and finally, you can check that or should I put them. So, we know that, this one is, let me do for the upper limit that is, 9.27 this part and then this is n equal to 11, so it is 9 degree of freedom and alpha is 0.05. So, alpha by 2 is 0.025, so that is 2.262, and then you have here, the MS residual is 2.36 and

n is 11 plus x naught is 3 and x bar is 0 here. So, 3 square by S xx that is, 1 1 0, so this one is the upper bound for beta naught plus beta 1 x naught.

And you can check that, this one is 9.27 plus I think, 9.27 is beta naught only, so plus beta 1 into x naught that is, beta 1 is 1.44 into x naught that is, 3. So, this whole thing is going to be equal to, you can check that, this is nothing but 15.03, it is the upper limit beta 1 x naught and the lower limit you can check, that is 12.15. So, we found confidence limit for mean response at the point X equal to 3.

(Refer Slide Time: 33:00)

5. What are the confidence limits ($\alpha=0.05$) for the difference between the true mean value of Y when $X_1=3$ and the true mean value of Y when $X_2=-2$?

$$E(Y \text{ at } X_1=3) - E(Y \text{ at } X_2=-2)$$

$$\hat{Z}_1 - \hat{Z}_2$$

$$\hat{Z}_1 = \hat{\beta}_0 + \hat{\beta}_1 3 \quad \hat{Z}_2 = \hat{\beta}_0 + \hat{\beta}_1 (-2)$$

$$\text{Thus } \hat{Z}_1 - \hat{Z}_2 = (\hat{\beta}_0 + \hat{\beta}_1 3) - (\hat{\beta}_0 + \hat{\beta}_1 (-2))$$

$$= \hat{\beta}_1 * 5 = 1.44 * 5 = 7.20$$

Now, this might be little, we did not try this one before, what it says is that, what are the confidence limit at 0.05 level of significance for the difference between the true mean value of Y , when X_1 equal to 3. So, that is nothing but mean value of Y at X_1 equal to 3 and the mean value of Y that is, $E Y$ at say X_2 equal to minus 2. So, this problem says, what is the difference between these naught, what is the confidence limit for this one. So, if I call this one say, z_1 and let me call this as z_2 , what are the confidence limit for z_1 minus z_2 .

Now, just now we know what is the unbiased estimator for this one, so the unbiased estimator for z_1 . For simplification, the unbiased estimator for z_1 is nothing but beta naught hat plus beta 1 hat X naught, here it is X_1 , so 3 and the unbiased estimator for z_2 , call it z_2 hat, that is equal to beta naught hat plus beta 1 hat minus 2, into minus 2. And thus, the unbiased estimator for z_1 minus z_2 is z_1 hat minus z_2 hat, which is

equal to beta naught hat plus beta 1 hat 3 minus beta naught hat plus beta 1 hat minus 2, which is nothing but beta 1 hat into 5.

And we know that, beta 1 hat is equal to 1.44 into 5, which is equal to 7.20, so what we did is that, we found a point estimation for z 1 minus z 2. And the point estimation is 7.21, now we have to find the confidence interval for z 1 minus z 2. So, what it do is that, we have to find a distribution for z 1 hat minus z 2 hat.

(Refer Slide Time: 36:33)

$$V(\hat{z}_1 - \hat{z}_2) = V(5\hat{\beta}_1) = 25 V(\hat{\beta}_1) = 25 * \frac{\sigma^2}{S_{xx}}$$

$$= \frac{25\sigma^2}{110}$$

$$\hat{z}_1 - \hat{z}_2 \sim N\left(z_1 - z_2, \frac{25\sigma^2}{110}\right)$$

$$\frac{(\hat{z}_1 - \hat{z}_2) - (z_1 - z_2)}{\sqrt{\frac{25 MS_{Res}}{110}}} \sim t_{n-2}$$

$$(z_1 - z_2) - t_{\alpha/2, 9} \sqrt{\frac{25 * 2.36}{110}} < \hat{z}_1 - \hat{z}_2 < (z_1 - z_2) + t_{\alpha/2, 9} \sqrt{\frac{25 * 2.36}{110}}$$

$$7.20$$

$$5.54 < \hat{z}_1 - \hat{z}_2 < 8.86$$

And for that, we need to compute the variance of z 1 hat minus z 2 hat, is nothing but variance of 5 beta 1 hat, because that is what we got here. So, z 1 hat minus z 2 hat is 5 beta 1 hat, so this one is nothing but 25 into variance of beta 1 hat and this is equal to 25 into sigma square by S xx, which is equal to 25 sigma square by 110. So, z 1 hat minus z 2 hat, which is an unbiased estimator for z 1 minus z 2, that follows normal distribution with mean z 1 minus z 2 and variance 25 sigma square by 110.

And then again this minus by square root of this follows standard normal and if you replace this sigma square by MS residual then it is t distribution, so z 1 minus z 2 minus z 1 minus z 2 by square root of 25 MS residual by 110, this follows t distribution with degree of freedom n minus 2. And from here, I can write that, z 1 minus z 2 is then less than equal to z 1 hat minus z 2 hat plus t alpha by 2, 9 degree of freedom and here it is 25 into MS residual is 2.36 and S xx is 110.

And similarly, here also it is z_1 hat minus z_2 hat minus t alpha by 2, 9 degree of freedom by the same thing, 25 into 2.36 by 110. And finally, we know that, this one is 7.20 so finally, the confidence interval for z_1 minus z_2 is 8.86 and the lower limit is 5.54. So, this is how we find confidence interval for the mean difference at two different point. So, the first problem was quite easy and this sort of problem we already solved in the first module or in the first topic. Now, we will go for the second problem and here, I recommend you do not look at the solution first, you try independently. And then if you can solve it independently that means, you have understood the things.

(Refer Slide Time: 40:55)

PROBLEM 2

Consider the Simple Linear regression model

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

where the intercept β_0 is known.

1. Find the least square estimator of β_1 for this model.
2. what is the variance of the slope ($\hat{\beta}_1$) for the least - Square estimator found in part 1.
3. Find a $100(1-\alpha)\%$ Confidence interval for β_1 . Is this interval narrower than the estimator for the case where both slope & intercept are unknown.

Here is the problem, consider the simple linear regression model y equal to β_0 plus $\beta_1 x$ plus ϵ , where the intercept is known. So, this is something new, so for this linear model, the β_0 is already known then what you have to do is that, find the least square estimator of the slope β_1 for this model. This is the first problem then what is the variance of the slope β_1 hat for the least squared estimator found in part 1.

So, you find a least square estimator of β_1 say that is β_1 hat then find the variance of β_1 hat. And the final problem is that, it says that, find the confidence interval for β_1 and is this interval narrower than the estimator for the case, where both slope and intercept are unknown. You try to solve independently and then see my solution and here is the solution.

(Refer Slide Time: 42:40)

① $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2)$ β_0 is known.

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum (y_i - \beta_0 - \hat{\beta}_1 x_i) x_i = 0$$
$$\Rightarrow \sum_{i=1}^n (y_i - \beta_0) x_i = \hat{\beta}_1 \sum x_i^2$$
$$\Rightarrow \hat{\beta}_1 = \frac{\sum (y_i - \beta_0) x_i}{\sum x_i^2}$$
$$S = \sum (\epsilon_i)^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \beta_0 - \hat{\beta}_1 x_i)^2$$

The first part is, you are given a model y_i , you have to feed this model β_0 plus $\beta_1 x_i$ plus ϵ_i and ϵ_i satisfies all these assumption, assumed that normal 0 σ^2 . The only thing is that, here β_0 is known, so only you have to find the least square estimator for β_1 . So, how do we find that we will compute the least square function S , which is equal to ϵ_i^2 . What is ϵ_i^2 , ϵ_i^2 is y_i minus \hat{y}_i square, which is equal to y_i minus, \hat{y}_i is β_0 plus $\hat{\beta}_1 x_i$.

So, see here, I did not put hat, because this is the parameter we do not need to estimate, we have to estimate only β_1 . And then this is the least squared function and then you find $\hat{\beta}_1$ in such a way that, this is minimum. So, you have to differentiate this least square function with respect to β_1 , this is equal to 0 implies that, summation y_i minus β_0 minus $\hat{\beta}_1 x_i$ into, x_i is equal to 0 . And from here, I get that, y_i minus β_0 into x_i equal to $\hat{\beta}_1$ summation x_i^2 .

So, this implies that, my least square estimator for β_1 is equal to summation y_i minus β_0 x_i by summation x_i^2 . So, this is the, so you are done with the first part, this is the least square estimator for $\hat{\beta}_1$, when β_0 is known. The second problem is, you find the variance of $\hat{\beta}_1$.

(Refer Slide Time: 45:44)

The image shows a handwritten derivation on a blue background. On the left, the variance of the OLS estimator $\hat{\beta}_1$ is calculated as follows:
$$v(\hat{\beta}_1) = \frac{v\left(\sum_{i=1}^n (y_i - \beta_0) x_i\right)}{\left(\sum x_i^2\right)^2}$$
$$= \frac{\sum_{i=1}^n x_i^2 v(y_i)}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2 \sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\left(\sum x_i^2\right)}$$
 On the right, the OLS estimator $\hat{\beta}_1$ is defined as
$$\hat{\beta}_1 = \frac{\sum (y_i - \beta_0) x_i}{\sum x_i^2}$$
 and the variance of y_i is given as
$$v(y_i) = \sigma^2$$
 A small logo in the top right corner reads '© CET I.I.T. KGP'.

The second part is find the variance of beta 1 hat and what is the beta 1 hat, beta 1 hat we just found that, this is equal to summation $y_i - \beta_0$ into x_i by x_i square. So, the variance of this one, so here only random variable is y_i , so this one is equal to variance of summation $y_i - \beta_0$ into x_i variance of this, by summation x_i square whole square. So, this one is equal to, they are all independent, y_i 's are independent, so this is x_i square variance of y_i .

Variance of $y_i - \beta_0$ is nothing but variance of y_i , because β_0 is a constant for i equal to 1 to n by summation x_i square whole square and variance of y_i , we know that, variance of y_i is equal to σ^2 . So, we can put now the σ^2 into x_i square by summation x_i square whole square. So, this is equal to σ^2 by summation x_i square, so this is the variance of beta 1 hat. And then the third problem was, you find confidence interval for beta 1 hat.

(Refer Slide Time: 48:11)

PROBLEM 2

Consider the Simple Linear regression model

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

where the intercept β_0 is known.

1. Find the least square estimator of β_1 for this model.
2. What is the variance of the slope ($\hat{\beta}_1$) for the least - Square estimator found in part 1.
3. Find a $100(1-\alpha)\%$ Confidence interval for β_1 . Is this interval narrower than the estimator for the case where both slope & intercept are unknown.

So, the last part was, find confidence interval for beta 1, so we know the variance of beta 1 hat. Let me check, whether beta 1 hat is unbiased, so you find the expectation of beta 1 hat.

(Refer Slide Time: 48:34)

$$E(\hat{\beta}_1) = \frac{E\left[\sum (y_i - \beta_0) x_i\right]}{\sum x_i^2} = \frac{\sum (\beta_1 x_i) x_i}{\sum x_i^2} = \frac{\beta_1 \sum x_i^2}{\sum x_i^2} = \beta_1$$
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$E(y_i - \beta_0) = E(\beta_1 x_i + \epsilon_i)$$
$$= E(\beta_1 x_i)$$
$$= \beta_1 x_i$$
$$E(\hat{\beta}_1) = \beta_1 \quad \hat{\beta}_1 \text{ is unbiased.}$$

So, that is nothing but expectation of summation y_i minus β_0 times x_i sum over x_i square and this one is equal to, y_i minus β_0 times x_i from the model you get, into x_i . So, you put the expectation inside by summation of x_i square, let me make it clear, so you can put this expectation and bring this expectation inside. So, y_i

minus beta naught, y_i is equal to beta naught plus beta 1 x x_i plus epsilon. So, expectation of y_i minus beta naught is equal to expectation of beta 1 x x_i plus epsilon and expectation of epsilon is equal to 0. So, expectation of beta 1 x x_i , which is beta 1 x x_i , so this one is beta 1 into summation x_i square by summation x_i square, that is nothing but beta 1. So, what we found is that, expectation of beta 1 hat is equal to beta 1, so though beta 1 hat is unbiased and we also know the variance of beta 1.

(Refer Slide Time: 51:04)

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum x_i^2}\right)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{\sum x_i^2}}} \sim t_{n-1}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$S = \sum_{i=1}^n \epsilon_i^2$$

$$P_Y \left\{ -t_{\alpha/2, n-1} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{\sum x_i^2}}} \leq t_{\alpha/2, n-1} \right\} = 1 - \alpha$$

So, beta 1 hat follows normal distribution with mean beta 1 and variance sigma square by summation x_i square. And then the usual technique, this minus this by square root of this follows standard normal and then beta 1 hat minus beta 1 by, if you replace this sigma by MS residual by x_i square then what you get is that, this follows t distribution with degree of freedom n minus 1 here. This is the residual degree of freedom and you should understand that, here the model is y equal to beta naught plus beta 1 x x_i plus epsilon i .

So, y_i is this and here, while minimizing this least square function S , which is equal to ϵ_i square, there we differentiated this one with respect to beta 1 only. So, there is only one restriction on epsilon i and so you have the freedom of choosing n minus 1 ϵ_i independently and then the last one has to be chosen in such a way that, the restriction is satisfied. So, that is why, the SS residual here has degree of freedom n minus 1, not n minus 2, so this is one point.

And from here, you can check that, we can write that, beta 1 hat minus beta 1 by MS residual by summation x i square, this less than equal to t alpha by 2 n minus 1, t alpha by 2 n minus 1 minus, this has probability 1 minus alpha.

(Refer Slide Time: 53:43)

The image shows handwritten mathematical work on a blue background. At the top, it shows the confidence interval for β_1 when the intercept is known:
$$\hat{\beta}_1 - \sqrt{\frac{MS_{Res}}{\sum x_i^2}} \cdot t_{\alpha/2, n-1} \leq \beta_1 \leq \hat{\beta}_1 + \sqrt{\frac{MS_{Res}}{\sum x_i^2}} \cdot t_{\alpha/2, n-1}$$
 Below this, it shows the confidence interval for β_1 when both the intercept and slope are unknown:
$$\hat{\beta}_1 - \sqrt{\frac{MS_{Res}}{S_{xx}}} \cdot t_{\alpha/2, n-2} \leq \beta_1 \leq \hat{\beta}_1 + \sqrt{\frac{MS_{Res}}{S_{xx}}} \cdot t_{\alpha/2, n-2}$$
 The next part shows a comparison of the two denominators:
$$\sqrt{\frac{MS_{Res}}{\sum x_i^2}} \leq \sqrt{\frac{MS_{Res}}{S_{xx}}}$$
 To the right of this, it states:
$$\sum x_i^2 \geq S_{xx} = \sum x_i^2 - n \bar{x}^2$$
 At the bottom, it concludes: "Yes, this interval is narrower than the interval for the case where both β_1 & β_0 are unknown."

And then finally, what we have is that, we have the interval for beta 1 that is, beta 1 hat plus MS residual summation x i square t alpha by 2 n minus 1. And the lower bound is beta 1 hat minus MS residual by summation x i square t alpha by 2 n minus 1, this is the lower bound for this. Now, what happen in the usual case when both beta naught and beta 1 are unknown, we get this confidence interval for beta 1. There we get, it is beta 1 hat plus MS residual by here, you will get S xx into t alpha by 2 and the degree of freedom is n minus 2, so this is the upper limit.

And the lower limit is similar similarly, beta 1 hat minus MS residual by S xx into t alpha by 2 n minus 2. The question was, is this interval narrower than the estimator for the case, where both slope and the intercept are unknown. So, this is the case when both intercept and beta 1, they are unknown and this is the case when beta naught, the intercept is known. Now, whether this interval is narrow than this one, how to check that, see S xx is equal to summation x i square minus n x bar square.

So, which implies that, S xx is smaller than summation x i square that means, this one is larger than this one. So, from here, I can say that, MS residual by summation x i square is less than equal to MS residual by S xx. So, this one is larger than this one and again from

the t table you can check that, t value for this one $t_{\alpha/2, n-2}$ is larger than, because they are the lower degree of freedom. is larger than $t_{\alpha/2, n-1}$. So, both this one is larger than this one, this one is larger than this one, I should not write square root till that. So, that is why, of course that, this interval is narrower than this one, the final answer is, yes this interval is narrower than the interval for the case, where both β_0 and β_1 are unknown. So now, we have to stop, so tomorrow again in the next class, we will be talking about some more problems on regression.

Thank you.