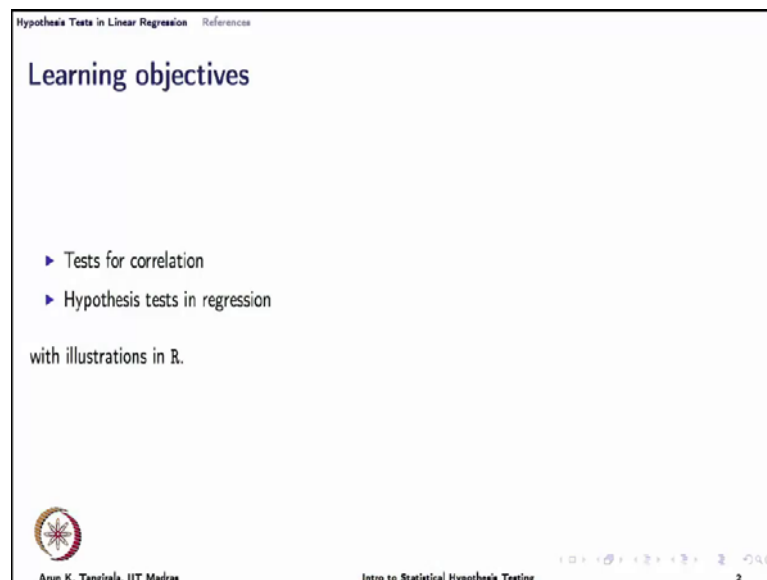


Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 15
Hypothesis Test in Linear Regression

Hello, and welcome to the penultimate lecture on Introduction to Statistical Hypothesis testing, where today we will look at hypothesis test involving correlation, and certain other factors such as Model parameters, and the Goodness of the model in linear regression. So, specifically, we look at test for correlation to begin with.

(Refer Slide Time: 00:41)




Hypothesis Tests in Linear Regression References

Learning objectives

- ▶ Tests for correlation
- ▶ Hypothesis tests in regression

with illustrations in R.

 Arun K. Tangirala, IIT Madras

Intro to Statistical Hypothesis Testing 2

And then, also learn what are the hypothesis test in regression; with of course, illustrations in R. Now, when I say regression here, we are looking at linear regression, where one is fitting a linear model between 2 or more variables. Typically, 2 variables are involved: one that you predicting and the other that you are using for prediction. Now, the reason for including both these in the same umbrella is because as we have studied earlier, correlation is a measure of linear dependents; therefore, when we want to fit a linear model, it is generally wise to study the correlation between those 2 variables that you would like to model, and then, once the correlation estimate passes a test of significance, that is when we have determined statistically that there is a significant correlation between 2 variables, then we proceed to fitting a linear model. Many text

books would perhaps present this in a different sequence; that is, talk about linear regression first, and then talk about correlation, but it is practical to discuss correlation first and then talk about linear regression. And, that is why we have sequenced it in this manner. Of course, as I said earlier, at each stage, we will show how to carry out this hypothesis test in R.

So, let us begin with the estimation of correlation.

(Refer Slide Time: 02:18)

Hypothesis Tests in Linear Regression Reference

Estimation of correlation

Computation of correlation requires covariance estimates.

A standard way of estimating covariance between two RVs X and Y is through

$$\hat{\sigma}_{XY} = \frac{1}{n} \sum_{k=0}^{n-1} (x_i - \bar{x})(y_i - \bar{y}) \quad (1)$$

The correlation estimate (a.k.a. *Pearson's correlation coefficient*) is then given by

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} \quad (2)$$

Note: Covariance is symmetric, i.e., $\sigma_{XY} = \sigma_{YX}$ (likewise is correlation).

Arun K. Tangirala, IIT Madras
Intro to Statistical Hypothesis Testing
3

Now, if you recall from early lectures in this course, we have defined theoretically what is a correlation. Correlation is standardized covariance; therefore, estimation of correlation will first require estimation of covariance. We have earlier seen how to estimate variance, where we have talked about 2 different estimators: one which we called as s^2 and the other which we called as s^2_{n-1} ; that is how we denote it. And the difference between those 2 was; while one was unbiased the other was biased, but then, the 1 which had one over n , which was more efficient than the estimator - the unbiased estimator - which had 1 over $n-1$ as the factor for estimation. Likewise, here, now covariance being a generalization of variance to 2 variable case, we have at least 2 different base of estimating covariance.

There are, in general, many different ways of estimating variance, but among the widely prevalent ones, there are 2 estimators for estimating covariance of which I am showing one of them, and this has 1 over n in front of the summation. Ideally speaking, this 1 over

n, that is estimator that involves 1 over n; is a biased estimator; nevertheless, we still use this widely to estimate covariance of course, for several different reasons which we shall not go into at the moment, but regardless of whether you use a 1 over n or a 1 over n minus 1 or a 1 over n minus 2 the resulting estimate for correlation is not effected so long as you use the same estimator for estimating covariance and also estimating variance. Let me explain that briefly.

(Refer Slide Time: 04:24)

The chalkboard contains the following mathematical expressions:

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) ; \hat{\sigma}_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}}$$

$$H_0: \rho = \rho_0 \quad (\rho_0 = 0)$$

$$H_a: \rho \neq \rho_0$$

So, what we mean here is, here we are using an estimate; estimate of covariance, which has this expression, where x bar and y bar rare usual sample means. Now, from the expression given on the slide, an estimation of correlation is constructed in this fashion, and what I meant earlier was that whether you use a 1 over n or a 1 over n minus 2 or 1 over n minus 1 it does not influence a correlation estimate so long as you use a same expression for estimating the standard deviations or the variances of x and y. That is, if so long as I estimate variance, example of x, in this fashion and likewise for y as well. Of course, here i run from 1 to n; i refer to the i'th observation.

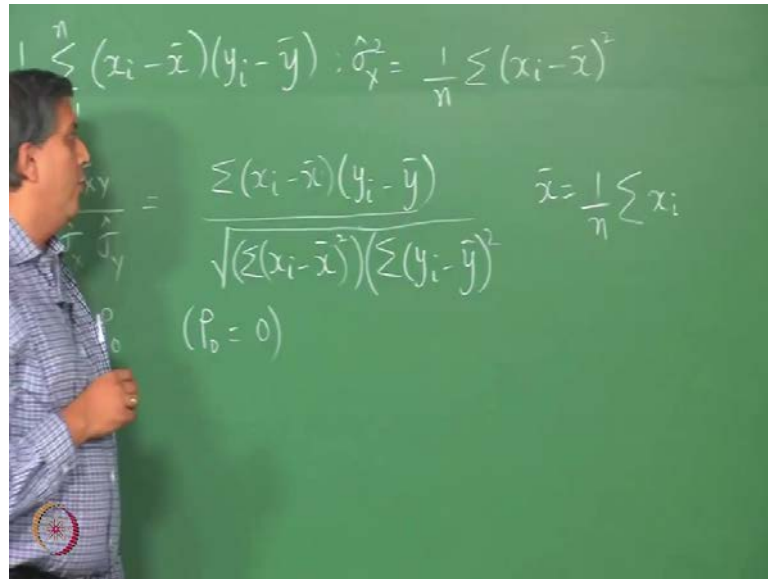
So, you can clearly see that once I take the square root of the variance of x and y, whatever factor that I am using here would cancel out in the numerator and denominator. As a result, you would see in many texts, this kind of an expression. Of course, all this summation run from i equals 1 to n. And, we shall also see these kinds of summation appearing later on in the optimal estimates of the parameters in linear regression.

Definitely, as we have taught at the beginning of the lecture, there is a strong interconnection between correlation and linear regression. Therefore, you will see similar kind of summations or the terms appearing in linear regression not necessarily identical.

So, now, this is the correlation estimate that we are going to work with, and naturally, like we asked for any other estimate, we would now try to set up a null hypothesis of the form, for example, ρ is equal to ρ_0 ; that is, the postulate being correlation being identical to a pre-specified or postulated value versus one of the alternative hypothesis, for example, $\rho_0 \neq \rho$. The typical kind of test that we normally conduct for correlation, which are called significant test, is whether the observed correlation or the estimated correlation is significant or not, in which case ρ_0 is 0. In fact, this is true for any parameter estimate; whenever, we use a term significance test for some parameter, what we mean is that the true parameter is 0, and whether the observed or whether the observed parameter estimate is significant - statistically significant.

Now, before we proceed to learn how to conduct this hypothesis test, clearly, we know by now. Hopefully, we are experts now in hypothesis testing. We know that to conduct hypothesis test like this, I need the sampling distribution of the so-called sample correlation, and the difference between the hypothesis test, sorry, the estimate that we have here versus let us say an estimate of mean is that this estimator or estimate is a non-linear function of the observation. So, we say that this is a non-linear estimator of the parameter which is correlation.

(Refer Slide Time: 08:47)


$$\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) : \hat{\sigma}_y^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$
$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}} \quad \bar{x} = \frac{1}{n} \sum x_i$$

$(P_0 = 0)$

Whereas, if you take sample mean, so if you take \bar{x} , this is a linear estimator, and it was easy to derive the sampling distribution of \bar{x} for instance using the central limit theorem, but here it is not so straight forward to derive the sampling distribution of correlation, given the distribution of x and y . Now, knowing this difficulty, many researches, of course, spent a lot of time, several decades ago and came up with the distribution properties of the sample correlation under some restricted conditions, and we will discuss those shortly, and then, proceed towards hypothesis test for correlation.

There is also another point that I would just like to mention in passing, which is that covariance is symmetric; which means whether I write it as a σ_{xy} or σ_{yx} it's one and the same, and likewise for correlation as well; ordering does not matter; we have already discussed this in the lecture on correlation. So, as we just discussed, the properties or the sampling distribution of the sample correlation coefficient is not easy to obtain.

(Refer Slide Time: 09:59)


Hypothesis Tests in Linear Regression References

Properties of sample correlation coefficient

- ▶ The sample correlation coefficient is an **asymptotically unbiased and consistent** estimator of the correlation ρ_{XY} .
- ▶ **Distribution:** Under the bivariate Gaussian assumption for X and Y and large sample assumption,
 - ▶ **True correlation is $\rho = 0$:** The estimate is known to have a near normal distribution.

$$\hat{\rho} \xrightarrow{d} \mathcal{N}(0, 1/n); \quad \text{Small sample: } \frac{\hat{\rho}}{\sqrt{1 - \hat{\rho}^2}/(n-2)} \sim t(n-2) \quad (3)$$

- ▶ **True correlation is $\rho \neq 0$:** Fisher's transformation produces a transformed coefficient with **approximately normal distribution**.

$$F_{\hat{\rho}} = \frac{1}{2} \ln \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right) \sim \mathcal{N}(\mu_F, \sigma_F^2) \quad (4)$$
$$\text{where } \mu_F = \frac{1}{2} \ln \left(\frac{1 + \rho}{1 - \rho} \right); \quad \sigma_F^2 = \frac{1}{n-3} \quad (5)$$


Arun K. Tangirala, IIT Madras Intro to Statistical Hypothesis Testing 4

So, the distribution properties have been arrived at or have been known under some restricted conditions; however, as an estimator, the sample correlation that we have on the board or on the slide is asymptotically unbiased. Now, we need this, in fact, it is also unbiased, not necessarily asymptotically, unbiased only, and consistent estimator. If you recall, unbiased would mean on the average the estimator gives you the truth, and consistent would mean that as a sample size grows large, the estimates converge to truth in a statistical sense or a probabilistic sense. So, having assured that that assurance is necessary for us to conduct hypothesis test or use this estimator to test correlations.

Now, let us move on to the distributional property. When x and y have a joint Gaussian distribution or a bivariate Gaussian distribution, you should recall in one of the lectures we had written the expression for joint Gaussian distribution where we talked about the difference between correlation and independence. If you refer to that lecture, you will see the expression for a bivariate Gaussian distribution. When x and y have a bivariate Gaussian distribution, and when the sample size is large, and when the true correlation is 0, so you can see that there are quite a few restrictions here. Of course, some of these are standard, you may argue, that even in the case of variance, we stated the sample distribution of variance under the normality assumption only, but the large sample assumption was not really necessary there, but further, we have now, two different scenarios depending on what the true correlation is. If the true correlation is zero, of course, we do not know that, but what this means is if I am performing a hypothesis test

of the form $\rho = 0$, then I should use this sampling distribution. If I am performing hypothesis of the form $\rho \neq 0$, then I may have to use a different sampling distribution

So, let us look at a first case when the true correlation is 0, then under the last sample assumption a nice result falls out, which is that the sample correlation follows a Gaussian distribution with mean 0 that is it is unbiased, because that is a truth, and variance $1/n$; that means, it has standard error of $1/\sqrt{n}$ where n is the usual sample size. However, when the sample size is large, the sampling distribution of $\hat{\rho}$ deviates from Gaussianity and instead follows a t -distribution with $n - 2$ degrees of freedom. This $n - 2$ degrees of freedom comes about because we have used 2 degrees of freedom in estimating the sample means, \bar{x} and \bar{y} ; you can think of it that way. So, depending on whether your sample size is small or large you can use 1 of these distributions. In fact, if you try to use a small sample expression for large sample case, you would not be making much of an error, because we know that the t -distribution tends to a Gaussian as n become large, and $n - 2$ tends to n . So, that is not an issue, but on the other hand, if you were to use a large sample expression for the distribution, for small sample size, there is a big scope for making an error. So, just be careful.

Now, moving onto the case of the true correlation being not 0, again we wouldn't know that, but when you are performing hypothesis test of $\rho = 0$, you have to work with a sampling distribution not for the correlation, but for a transformed correlation. In fact, it turns out that this was a tough problem to crack, but finally, Fisher came up with this idea of working with a transformed coefficient, and showed that this transformation given equation 4, $\frac{1}{2} \ln \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} \right)$ follows. Now, this transformation or transformed coefficient follows a Gaussian distribution with mean μ_f and variance σ_f^2 ; the expression for μ_f and σ_f^2 are given. Here the variance expression is $1/(n - 3)$. I would not go into the details of this derivation; you can find in good rigorous statistical text as a derivation of this or refer to, of course, Fisher's original paper.

So, the point is or the summary is when the true correlation is not equal to 0 the Fisher's transformed coefficient follows an approximately Gaussian distribution. This is only true for the large sample case. Of course, you would not be again making big error by using this even when the true correlation is 0, but then, when a simple result exist for the large

sample case why would you want work with a transformed coefficient. Therefore, all test of significance for correlation would either use this small expression in equation 3, sorry, small sample equation 3 or the large sample expression as the case may be. But when you are testing for the true correlation being something of rho naught equal to 0 like 0.1 or minus 0.2 and so on, then you would want to use expressions in 4 and 5.

Now, having said that, typically what is of interest is the first one that is a significance test; that is that the true correlation is 0. If it is found that the null hypothesis rho equal 0 has to be rejected, then one fits a linear model, and then one is interested more in the goodness of the model fits and so on. It is, of course, there are situations in which you may postulate that the true correlation is point one and point two and so on, but relatively those are rare compare to the significance test for correlation; something to keep in mind.

Alright, let us look at example now, and see also how we can do this in R.

(Refer Slide Time: 18:18)

The slide is titled "Hypothesis Tests in Linear Regression" and includes a "References" section. Under the heading "Examples:", there is a green box containing the text "Cranial circumference and Finger length". Below this, a "Problem:" section states: "A linear model is postulated between cranial circumference and finger length. We would like to test whether an actual linear relationship exists." A "Solution:" section follows, stating: "Compute correlation coefficient and perform a test of significance." The slide footer includes the IIT Madras logo, the name "Arun K. Tangirala, IIT Madras", the course title "Intro to Statistical Hypothesis Testing", and the slide number "5".

This is an example that we discussed in motivating lecture. Recall that there was a widespread belief, that there is a relation between the cranial circumference - that is circumference of the head here and the finger length. This kind of belief was held for a few centuries, and now, we want to see if there is a linear relation between the cranial circumference and finger length. So, for this purpose what we would do is we would randomly select a few individuals, record their cranial circumference and finger length, and then determine the correlation between these 2 parameters, because we are interested

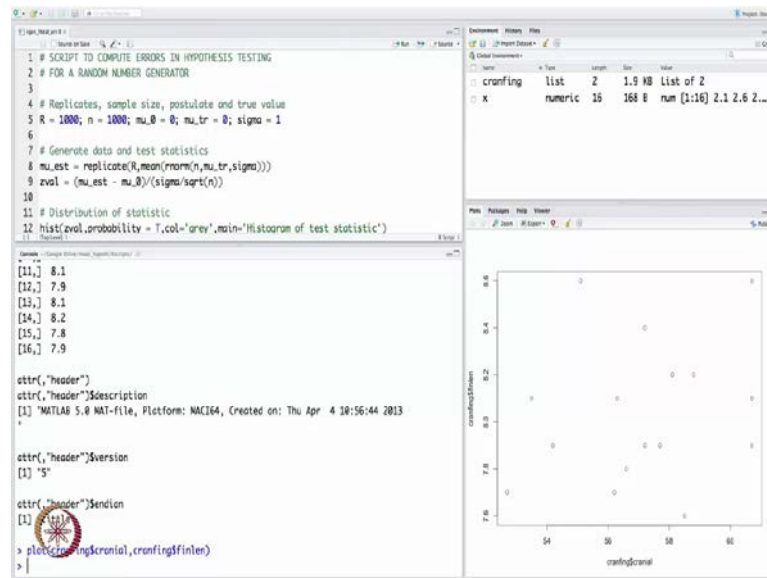
in linear relation; to test for non-linear relations requires test for independence, but that is beyond the scope of this course. Let us now pull up the data for the cranial circumference and finger length, and compute the correlation coefficient. Let us see if this belief actually holds any (Refer Time: 18:10).

(Refer Slide Time: 18:17)

The slide is titled "Hypothesis Tests in Linear Regression" and includes a "References" link. It features an "Examples:" section with a green header "Cranial circumference and Finger length". Below this, a "Problem:" section states: "A linear model is postulated between cranial circumference and finger length. We would like to test whether an actual linear relationship exists." A "Solution:" section follows: "Compute correlation coefficient and perform a test of significance." At the bottom right, a code box contains the text "R: Use cor and cor.test". The slide footer includes the IIT Madras logo, the name "Arun K. Tangirala, IIT Madras", the course title "Intro to Statistical Hypothesis Testing", and the slide number "6".

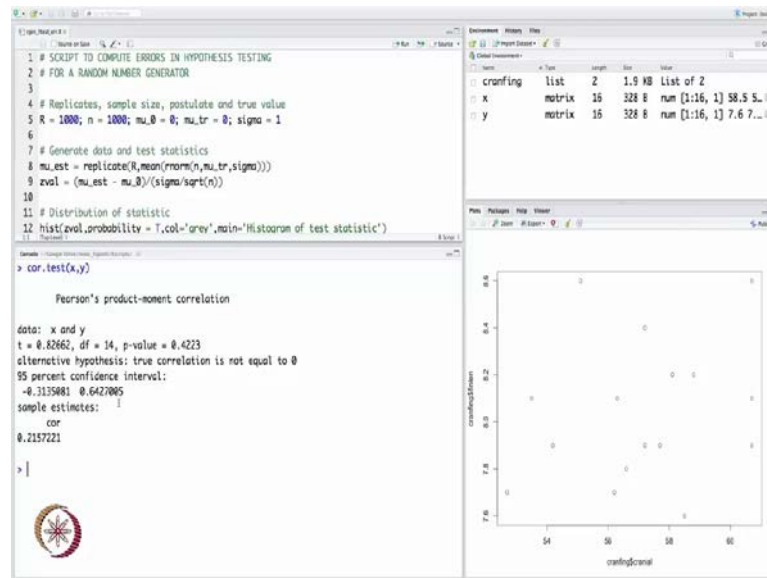
For this what you do is, you use the routines `cor` and `cor.test`. Of course, you can straight away use `cor.test` which implements the test for significance using the small sample expression that I gave earlier. If you only want to compute the correlation coefficient, of course, you can then use `cor` so that data is again contained in data file that we will upload on the web, and you can also work along with me.

(Refer Slide Time: 18:57)



Let us now turn to R; make sure we are in the working directory. So, this is the date. Some of these data sets, we have worked with earlier. The one that is of interest to us is that, we need to change the working directory here, and we have changed working directory to the file (Refer Time: 19:42) location. Now, we can load the data file. The name of the file is cran underscore finlen dot r data. And I am going to load that, which will load a variable called cranfin; it is a list variable and it has these attributes. And attributes of interest to us are the cranial circumference and the finger length. Of course, a good idea would be to plot the data points; draw scatter plots of the finger length versus the cranial circumference (Refer Time: 20:49). So, this is how the plot looks like. And what we are postulating is that there is a linear relationship; that is what is the equivalence of saying that I would like to see if there is a non-zero correlation. We are not so much worried about whether the correlation is positive or negative; we just want to know if there exists a significance correlation between these 2 variables. Of course, one can just to make things simple we can assign xcranfin instead of typing all the time these parameters. We can assign the cranial circumference to x and likewise the finger length to y.

(Refer Slide Time: 21:44)



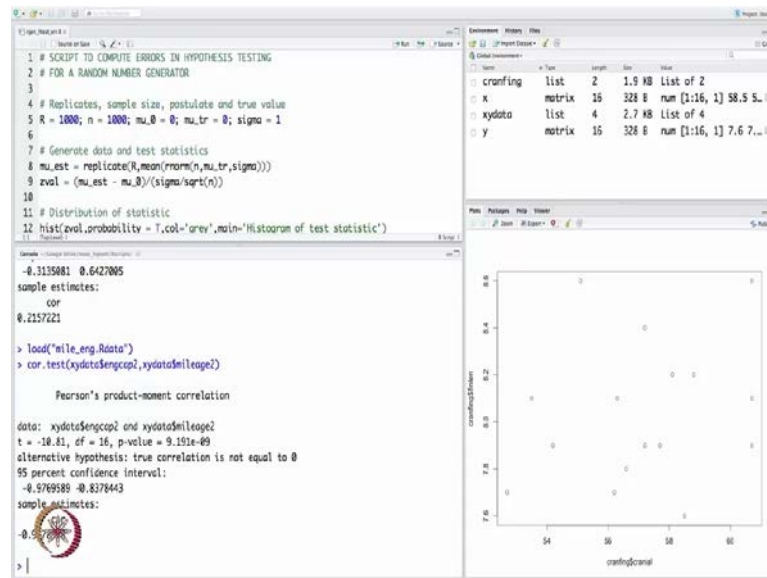
So, now, we can ask for correlation between x and y . And it turns out to be 0.2157 and so on. Now, of course, this is an estimate and we know from our prior experiences, that the face value of the estimate does not necessarily tell us anything about the truth, and that is why we turn to the hypothesis test. So, let us now ask `cor dot test` the on the result of this hypothesis test. So, on the top you see that it displays Pearson's product moment correlation. In fact, the correlation that we are working with is called Pearson's correlation. There are two other forms of correlation that are widely used: Kendall's correlation and Spearman's correlation; we do not discuss those correlation measures here; you can refer to any standard statistical test.

So, it says data that has been used this x and y , and it reports the t statistic, the degrees of freedom. Notice, that we have 16 data points, and from our previous expression, we said the correlation estimate for the small case under the bivariate Gaussian assumption, we do not know if that is true, but let us assume that the data falls out of a bivariate Gaussian distribution. Then, in that case, the statistic has n minus 2 degrees of freedom, and that is why the degrees of freedom is 14. And one can either use the critical value approach or the p value approach; let us take the p value approach; it is lot easier; we know that when the p value is low, lower than the significance level, the significance level that we have used in the standard thing is 0.05 as usual, and when the p value is less than α , but here it is a two-sided test. So, you will have α by 2 to the left and right, but overall the p value if it is less than α , then the null hypothesis must go (Refer Time: 23:53).

Here, the p value is greater than alpha. Of course, I can say much greater, but it does not make any difference, as the moment p value is greater than alpha, I have to; I fail to reject the null hypothesis, and therefore, the null hypothesis that the true correlation is 0 fails to be rejected, which means most likely the truth is that the true correlation is 0.

Of course, we will see this in a different way when we fit a linear model. Suppose, I did not perform a test of correlation, and instead, I went ahead and fit a linear model using standard least square method, then I should be able to see the same thing. That is even the hypothesis test in linear regression that I conduct on model parameters or the goodness of model should reveal the same thing that – look, you should not have fit a model because the true correlation is 0; there is no evidence to believe that a linear model will do a good job. And then, of course, you also have the confidence intervals here, for the correlation parameter. Now, again, these conference intervals are derived in the same way as we derived for the sample mean, ratio of variances, proportions and so on. You can start with a distribution; write a probabilistic interval for the correlation estimate, and then, from there derive the expression for the correlation coefficient, that is the conference interval for the correlation coefficient. If you look at the confidence interval, it includes 0 which is one of which is postulated value and therefore, the null hypothesis cannot be rejected. On the other hand, if you look at the correlation coefficient for another data set, that we will look at shortly in the context of linear regression, which is the data set that we talked about in the motivation lecture, the highway mileage verses engine capacity - in that case, the conference interval would not include a 0.

(Refer Slide Time: 26:09)



So, let us do that in a minute here. Let me load the data here which is containing in mile underscore eng dot r data. If you do that, and conduct a correlation testing now the variable is x data, this is also list variable, and we want to compute the correlation between the engine capacity and the highway mileage. I will explain these variable names at later stage, but let us, for the sake of the illustration compute the, perform the test on these 2 variables here - highway mileage variable 2 and the engine capacity, in fact. It does not matter because it symmetric, nevertheless, we want to be sticking to the conventions here. Now, when we perform this kind of a correlation here, something interesting comes up. Of course, there are certain defaults that we have used in this correlation dot test; for example, we assumed the significance level to be 0.05, we have assumed that the alternate hypothesis is of two-sided type and so on; that is what we are interested in always in signified test.

Now, the null hypothesis, again, for this case also is at the true correlation between the highway mileage and engine capacity is 0. You can either use the p value approach or the conference interval approach; both are telling us that the null hypothesis has to be rejected because the true value - postulated true value - which is 0 is not contained in a confidence interval. In fact, you can see that both the bounds are actually negative indicating a negative correlation. Of course, if you want to now test for negative correlation, you will have to go and change the alternative which I will do at a later stage. In fact, the linear regression will tell us that there is a negative correlation between

these two variables. You can also look at the p value, it is very small, in fact because it is smaller than alpha, and therefore, once again, the null hypothesis to be rejected. So, this is just a conformation of what we had observed using the confidence interval.

Now these 2 examples, hopefully, have given you a fair idea of how to conduct a correlation test in R, and of course, the theory behind it. In the next part of this lecture we will talk about linear regression a bit more in detail and look at the various tests that are involved in standard linear regression.