

**Applied Time-Series Analysis**  
**Prof. Arun K. Tangirala**  
**Department of Chemical Engineering**  
**Indian Institute of Technology, Madras**

**Lecture - 17**  
**Lecture 08A - Probability and Statistics Review (Part 2)-7**

Very good morning to all the students here and then the other room, so we discussed extensively somewhat about correlation in yesterday's class, today will partially discuss about correlation, in other words we will discuss partial correlation.

(Refer Slide Time: 00:42)



But before we do that towards the end of yesterday's lecture we had written a relationship for the squared correlation and a relationship for a pair of random variables  $X Y$  was that if I consider 2 models; so called forward model and you can say kind of a reverse model which is forward, which is reverse we do not really worried. If you were to fit models such as this or even simply models without the intercept on then we wrote a relationship for the squared correlation in terms of this regression coefficient particularly the slope parameter here as this  $\rho^2$  being the product of  $b^*$  and  $\tilde{d}^*$  where the star indicates that they are the optimal estimates that minimize the mean square error and we gave expressions for  $b^*$  and  $\tilde{d}^*$  yesterday.

Now suppose I want to verify this relation using data, we have proved this theoretically so there should not be any questions on it. However you want to verify this somewhat

empirically then you can do this in a software like R or any other software, I will show you how one could verify this relation in R. Although we write this relation for the theoretical case it is true even for the sample case that is when you are working with data you can show that the sample correlation also satisfies this relation which is not necessarily true of every theoretical relation that we derived so that should be something to be kept in mind. This particular relation and another relation that we learn today with regards to partial correlation the sample versions also satisfy this theoretical relation here.

In this case, we would replace the theoretical ones with their respective estimates which means that when we work with data, this kind of relationship be satisfied exactly and let us see whether that is the case by working with some random data set in R.

(Refer Slide Time: 03:07)

```

> xvec <- rnorm(200)
> yvec <- 2.5*xvec + rnorm(length(xvec))
> lmody <- lm(yvec ~ xvec)
> lmodx <- lm(xvec ~ yvec)
> lmody

Call:
lm(formula = yvec ~ xvec)

Coefficients:
(Intercept)      xvec
      0.0521      2.5445

> summary(lmody)

```

The screenshot shows the R console output for the above code. The right pane displays the summary of the linear model 'lmody', including the coefficients for the intercept and xvec, and the standard errors for the residuals.

To do this, let us create samples of X and Y on the screen there, I think Benjamin you can actually switch over to the computer screen. So, x vec contains the vector of observations of x or a sample of x, let us randomly sample 200 observations from a Gaussian distributed random variable with mean 0 standard deviation and that does not matter. So, to keep things simple, we will first generate x and accordingly we will generate the y vector, we will force y to be a linear function of x, but we will also allow some measurement noise or some other kind of disturbance affecting y. In other words, let us pick a value of alpha or b or whatever a beta and let us say generate y as 2.5 times

$x$  vec and add some noise to this linear function of  $x$ , all right and we can do that by adding another random variable.

Again we can generate this random variable from a Gaussian distribution and we can use the same length, you can specify the length as 200 or you can say length of  $x$  vec here so that it becomes more generate. You could scale this epsilon, see now yesterday we spoke about the signal to noise ratio the signal to noise ratio is a ratio of variances and later on when we move to frequency domain, we will learn a frequency domain version of signal to noise ratio, in any case you should view signal to noise ratio as a ratio of powers in the respective signals. So, you can adjust the power or the variance of this epsilon that we are adding by multiplying with a factor, but we let us not do anything at the moment, but in general when you are carrying out simulations it is extremely important to make sure that you maintain a healthy signal to noise ratio and healthy typically refers to value of 10 or above, but there is no nothing cast in stone.

Generally we maintain a signal to noise ratio of 10, if you want decent estimates, but it is here we are not really interested in estimating parameters perceive we want to really verify this identity that is the point, all right. So, we will go ahead with this and then generate our  $y$  and you can see here on the right top screen here that  $y$  vec is a length of 200 and so is  $x$   $y$ . Now in order to verify this relation here that we have derived we need estimates optimal estimates of  $b$  and  $b$  tilde right, and remember we are doing a linear regression here and in  $r$  the routine that does the linear regression for you using least squares estimate that is what we called theoretically as minimum mean square error minimization in practice when you are working with observations you call it as sample least squares so, in  $r$   $lm$  is a routine that does this fit for you.

Let us name the model that is that uses  $x$  as a regresses and  $y$  as a predictor call it as  $l$  mod  $y$   $x$  and of course, you can use the equality assignment operator as well, but I prefer to use this. So,  $l$   $m$  what do you supply for  $l$   $m$  of course, you can look up the help on  $l$   $m$ , but you can also press the tab to bring up the arguments to this function. Now if you see there is something called a formula and this formula has to be provided in a symbolic way in other words here for example, I am regressing  $x$  on to  $y$ . The formula in  $r$  for this would be  $y$  tilde  $x$  or whatever vectors you have for  $y$  and  $x$  in our case we have  $y$  vec and  $x$  vec.

(Refer Slide Time: 07:00)



In a general case, you may have data in the form of a data frame; in our tutorial session we had talked about I think we talked about data frames. So, data frame is general is a generalization of a matrix with columns being labeled. So, you can directly refer to a data frame or you can refer to variables in your environment. We will keep it simple will refer to the variables in the environment and we will refer will write this formula  $y \sim x$  it is understood that you are regressing  $x$  onto  $y$  and the intercept term is present by default.

If you do not want to the intercept term if you want to suppress it, you can use a minus 1 and it understands that you do not want the intercept term, but we want it and at the moment we will retain the intercept term although we have not included the intercept term in our model, generation model for  $y$  and that is all. You do not need to supply anything else unless these variables are not in the environment if the variables are from a data frame you have to mention the name of the data. So, that is it our model is ready  $lm(y \sim x)$  the nice thing about R studio as you can see is as soon as you type in the key I mean you key in the command the resulting environment actually takes you to that corresponding variable as you must have seen, so the top variable here is  $lm(y \sim x)$ .

Likewise we can also build the other model and simply reverse the roles of the regressor and the predictor. So, we have now 2 models containing the estimates of  $b$  and  $b$  tilde from the sample of  $x$  and  $y$ , all right. So, in general you should be familiar with linear

regression in time series analysis and in general data analysis. So, it is a good idea to know how to handle these objects, you can simply type in for example, the name of the object and it will report in the most simplest form it will report the estimates of the intercept and the coefficient b we can call it slow, although I do not like to call it a slow.

(Refer Slide Time: 09:46)

```

> summary(lmody)

Call:
lm(formula = yvec ~ xvec, data = lmody)

Residuals:
    Min       1Q   Median       3Q      Max
-1.04046 -0.25116 -0.00782  0.20667  1.25851

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.019682   0.025658  -0.767   0.444
yvec         0.348219   0.009524  35.721 <2e-16 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3629 on 198 degrees of freedom
Multiple R-squared:  0.8657,    Adjusted R-squared:  0.865
F-statistic: 1276 on 1 and 198 DF, p-value: < 2.2e-16

> str(lmody)
lm
  .lmody      list of 12
  .lmodyx    list of 12
  .svarmod   large svector (13 elements, ...)
  .svarmod2  large svector (13 elements, ...)
  .svarmod3  large svector (13 elements, ...)
  
```

Now, a better way of reading these objects is to parse it through summary, summary looks at the object and gives you the relevant output or the display. So, suppose I want a summary of the modeling that I have done a lot of information very useful information is displayed on the screen for you can see foremost you have the statistics for the residuals whenever we are fitting a model as you will learn later on residuals hold the key to a lot of information particularly on whether your model is good whether you have done whether you need to refine your model and so on.

At this moment we will not worry about that, but just to tell you that the first thing that you see on the top is the statistics of the residuals and then comes the coefficients; estimates of the coefficients as you can see there is these intercept term here and then there is this  $b$  that we are looking for and then you have additionally 3 columns - one is the standard error, the standard error is the error that you are making in estimating  $b$  there is  $a$ ; what was the true value that we have used? 2.5, as you can see the estimate is different and this standard error can be calculated from a single data record we will learn how to do that in estimation theory. And it kind of gives you an average error in the

estimate it is not the exact error otherwise you would simply add the error to the estimate and get the truth and then there is something called the t value and then there is a p value that is being reported towards the end and there are also stars like you know superstar, megastar, there are some stars given in the last column per say.

A crude way of interpreting the stars right now is I mean a very useful way, is if the stars are given then that estimate is significant, what do we mean by significant any idea?

Student: When parameter is (Refer Time: 11:39).

It is a test of whether you have sat through the NPTEL course on hypothesis testing or not, yeah you are able to say something.

Student: That parameter is relevant to the output, the output parameter (Refer Time: 11.58).

(Refer Time: 12:01), yes-yes.

Student: (Refer Time: 12:06) relation between.

Between?

Student: The output and the input variable.

No, so as you can see on the display, we have 3 stars associated with the coefficient  $\hat{b}$ , I mean estimate of the coefficient  $b$ , what would those stars mean, are you saying that it indicates high correlation between  $y$  and  $\hat{y}$ ? No.

Student: It is very relevant to the modern (Refer Time: 12:27) if you have more than one variables.

You would not like it if I give you the answer and I say and then I stop with that you would want more details, can you be more clear, what do you mean by relevant.

Student: As in.

Today it is going to rain very heavily, any other student?

Student: sir

Yes.

Student: Sir yes it is the probability that coefficient is not 0.

Probability.

Student: (Refer Time: 13:08).

You mean to say that the actual coefficient is 0 when there is; when there are the stars next to it.

Student: The probability (Refer Time: 13:20) if there are more stars then it is less likely that it is 0.

Slowly getting there, any other person in that room, people here are stunned today, any other person in that room? Good attempt so, so any other person in this room? Yes.

Student: Sir, if we are adding some more variables then how important is this variable to calculate the estimation.

No, no it is not.

Student: (Refer Time: 13:50).

No, well yes and no, but simple interpretation suppose somebody would you know person just beginning to learn this would come and ask you what do the stars mean, how would you explain that?

Student: They are almost linearly (Refer Time: 14:19).

What they? Who are they?

Student: Those 2 variables.

Yeah, who are they?

Which those 2? X and Y, they are almost independent.

Student: (Refer Time: 14:29).

No, the answer is way off, last one, yes.

Student: It is the probability of (Refer Time: 14:39).

Somehow this I do not know where you have this idea of probability of hypothesis being through, the null hypothesis at your stated is correct. So, let me some of you have tried to give your answers bit close, but not exactly correct.

Just to briefly tell you, we will again go through this when we talk about estimation theory let me discuss how to estimate parameters and so on. The stars here mean that they estimate that you have obtained is significant and what we mean by significant always in hypothesis testing.

(Refer Slide Time: 15:23)



When we say a parameter estimate is significant; that means, you are looking at this kind of a null hypothesis suppose the parameter is theta that you are estimating, there is always this null hypothesis that theta is 0 and then the alternate hypothesis that theta is not. Why are we interested in this kind of a hypothesis all the time? Particularly in regression because we would like to know if that term in the model should have been included or not, right always, we are interested in this kind of a hypothesis and you should get used to this in modeling we later on when we build time series models, again we will have will encounter these kind of hypotheses for every term that we include there will be a coefficient associated with it, particularly in linear models is very straightforward to see. We would like to know if we have included this term should have included this term or not.



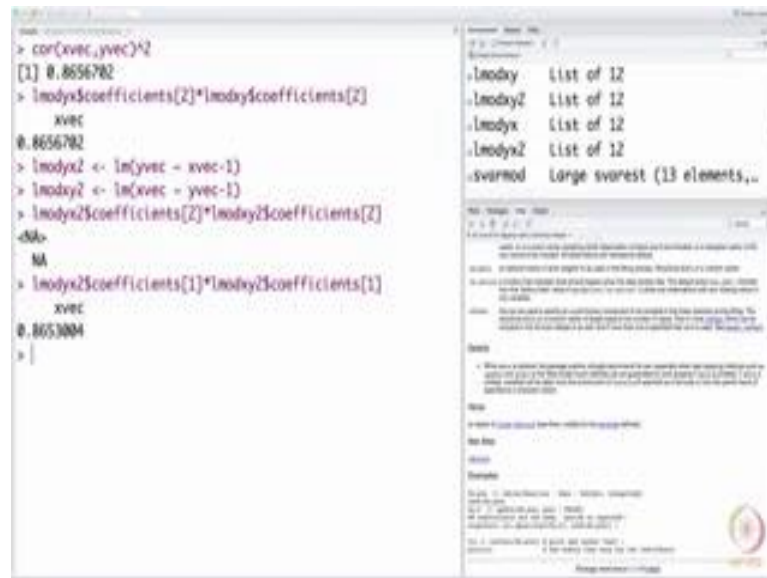
Here we would like to know if there exists a linear relationship between  $y$  and  $x$  or not in other words, and the associated hypothesis and null hypothesis and alternate hypothesis have those  $\theta = 0$  and  $\theta \neq 0$ . Specifically here with respect to  $b$  we are asking if  $b$  is 0 versus  $b$  not being 0 and the 3 stars means that we reject the null hypothesis that  $b$  is 0 there is no probability that is only thing and for more details you should actually refer to the statistical hypothesis testing videos until we reach the part on estimation theory.

So, to summarize we have here estimates of  $a$  and  $b$  there are no stars next to  $a$ . In fact, you see a very high  $p$  value there whereas, a very low  $p$  value for the estimate of  $b$  and as someone said you do set thresholds for  $p$  values will not get into that. But generally if the  $p$  value is very low then the null hypothesis must go there is a nice one line up that you can memorize and that is why you have the stars there.

In other words, I should not have included an intercept term in this model. So, it has correctly figured that out did we include an intercept term in the generation of  $y$ , no. So, the method has figured that out and that is the nice thing about data analysis you if you use a right method you can make the data speak to you it can tell you some essential facts about the truth. So, anyway we have  $b$  hat here reported in the second column under the column estimate here second one and then there are of course, the significance codes here will not go over that right now we have already discussed and then there are a bunch of other details given here residual standard error, multiple  $r$  square,  $s$  statistic and so on. We will not get into all of this at this moment, but what is of interest to us is  $b$  hat.

Likewise there is also this other model for which we can pull up the summary statistics and the interpretation is the same. Now what do we want to verify? We want to verify if the squared correlation is a product of the optimal estimates of the coefficients. To do this it is very simple, compute the let me actually clear the screen here compute the correlation between  $x$  vac and  $y$  vac and in fact, the square of that; `cor` is a routine in `r` which computes the correlation that is a sample correlation and I am not giving the expression for sample correlation at this moment.

(Refer Slide Time: 19:05)



```
> cor(xvec, yvec)^2
[1] 0.8656782
> lmody$coefficients[2]*lmody$coefficients[2]
xvec
0.8656782
> lmody2 <- lm(yvec ~ xvec-1)
> lmody2 <- lm(xvec ~ yvec-1)
> lmody2$coefficients[2]*lmody2$coefficients[2]
-NA-
NA
> lmody2$coefficients[1]*lmody2$coefficients[1]
xvec
0.8653004
> |
```

Environment: global env (base)

```
lmody      list of 12
lmody2     list of 12
lmody      list of 12
lmody2     list of 12
svaremod   Large sparse (13 elements,...
```

This is the squared correlation between x and y and we want to see if this is the same as the product of the correlation. So, to access the coefficients we use the dollar operator because coefficient is one of the attributes of this model it is a list object and there are 2 coefficients remember we are interested in the second one all right. And likewise here for the other model times 1 mod x y dollar coefficients, 2 is verified. Any questions, any question from the other room, everything is clear. So, this is how you verify the correlation.

Now just some good for thought, suppose I take off the intercept term, why will I take off the intercept? Because the regression statistics is telling me that I should not have included and intercept them rightly so, suppose I take off the interceptor, do you expect this identity to be satisfied - this is satisfied regardless of whether you have the intercept term theoretically or not do you expect this to be satisfied? Yes or no, let us see if that is the case.

So, let us quickly go back and rebuild our model right, let us call this as 2, second model and this time omit the intercept term and likewise here omit the intercept term in the other model as well, right. We do not have to compute a squared correlation that remains the same all I have to do is go back and recomputed the product of the coefficients from these new models without the intercept term sorry, now there is no second coefficient observe so that because this time you omitted the intercept term there is only one

coefficient. So, I should not be asking for the second coefficient. What has happened? Are they same, what do you think are they the same, any idea why? Think about it, at this moment to explain why they do not match is you know outside the scope of today's lecture, but just think about it if you have an answer you can discuss with me maybe after the lecture or in some of the next few lectures. So, this is how you verify your identity in the case of correlation.

Now, there is another point that I wanted to make when we when we are talking about this correlation this correlation coefficient is also known as the Pearson's correlation coefficient. There are other forms of correlation called the Spearman and then the (Refer Time: 22:40) coefficients which are actually computed on order data; that means, you rank the data and then you compute the coefficients those are typically used for data that come from a non Gaussian distribution.

The Pearson's correlation coefficient is ideal to use when data falls out of a Gaussian joint Gaussian distribution, but if you have data falling out of a non Gaussian distribution or you are looking at some kind of non-linear relations and so on, one of the ways to test for relations is to rank the data and remember ranking involves a sorting operation which is a non-linear transformation and then you are computing the correlation. In fact, that forms the basic idea of several non-linear regression models non-linear tests - which is step one, transform the data into some space, sorry, through a non-linear transformation followed by a simple check of linear dependence in the transformed space.

Sorting is a non-linear operation, it is taking the data into some other space and then you are checking for linear dependence. For those of you who are familiar with machine learning you must have heard of Kernel methods, the Kernel methods actually use this idea they transform the data into so called feature space and typically these features are generated as non-linear functions of data of the given data and then a test for linear dependencies in the feature space is computed. That should also tell you why it is important to understand the theory of linear regressions or correlations and so on, because primarily in many of the non-linear data analysis techniques this idea is used where first data is transformed followed by a linear analysis.

If you are not familiar with linear analysis then you are going to be handicapped in that sense, so let us get back to our discussion here and talk about confounding.

(Refer Slide Time: 24:45)

Probability & Statistics - Review 2

## Confounding

When two variables  $X$  and  $Y$  are correlated, a question that begs attention is:

**Q:** Are  $X$  and  $Y$  connected to each other **directly** or **indirectly**?

Conditional measures provide the answer.

Alan N. Torgola Applied TSA August 16, 2016

All we have talked about the connections between the correlation and regression.

(Refer Slide Time: 24:52)

Probability & Statistics - Review 2

## Correlation and regression ... contd.

Further, (for the  $X \rightarrow Y$  model)

$$\text{cov}(\varepsilon, X) = \text{cov}(Y - b^*X, X) = 0 \quad (\text{residual} \perp \text{regressor}) \quad (19)$$
$$\sigma_\varepsilon^2 = \sigma_Y^2 - b^{*2}\sigma_X^2 = \sigma_Y^2(1 - \rho_{XY}^2) \quad (20)$$

so that the standard theoretical measures of fit for both (directional) models are

$$R_{X \rightarrow Y}^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2} = \rho_{XY}^2 = R_{Y \rightarrow X}^2 \quad (21)$$

Zero correlation implies no linear fit in either direction.

Alan N. Torgola Applied TSA August 16, 2016

We have also showed that rho square is b hat b star minus a minus b tilde star, there is also this r square that you normally encounter in linear regression which is a measure of the goodness of it and you can show that this r square is nothing, but your squared correlation itself between y and x.

(Refer Slide Time: 25:17)


Probability & Statistics - Review 2

## Remarks, limitations, ...                      ... contd.

- ▶ High values of **estimated** correlation may imply linear relationship over the experimental conditions, but a non-linear relationship over a wider range.
- ▶ **Absence of correlation only implies that no linear model can be fit.** Even if the true relationship is linear, noise can be high, thus limiting the ability of correlation to detect linearity
- ▶ **Correlation measures only linear dependencies.** Zero correlation means  $E\{XY\} = E\{X\}E\{Y\}$ . In contrast, independence implies  $f(x, y) = f(x)f(y)$

Despite its limitations, correlation and its variants remain one of the most widely used measures in data analysis

Amir H. Toghiani                      Applied TSA                      August 16, 2008



But that is something that we will revisit later on when we talk of regression or when we talk of least square.