

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

Lecture – 18
Lecture 08B - Probability and Statistics Review (Part 2)-8

(Refer Slide Time: 00:12)

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.

Arun K. Tangirala Applied TSA August 16, 2018

So, let us move on and yesterday I had mentioned that one of the limitations of using correlation alone, apart from a few others that it does not give you, it is not a measure of causation and it has no directionality built into it; is this problem of confounding. In other words, if I find two variables X and Y correlated then the question that really begs our attention is are X and Y directly related, can I assume that X and Y directly related and to that effect we had referred to an example yesterday of this you know number of fire trucks or firefighters and the damage caused by the fire. If I find the correlation between those two variables, it will turn out to be very high fine good does; it mean that the number of firefighters have a direct correlation with the damage and unfortunately even more in a way that is completely counterintuitive, counter physical and so on. So, to give you an idea let me actually again take you back to r , where I have downloaded a data set from the web.

(Refer Slide Time: 01:32)

```
> cor(xvec, yvec)*2
[1] 0.8656782
> lmody3coefficients[2]*lmody3coefficients[2]
xvec
0.8656782
> lmody2 <- ln(yvec - xvec-1)
> lmody2 <- ln(xvec - yvec-1)
> lmody25coefficients[2]*lmody25coefficients[2]
-NA-
NA
> lmody25coefficients[1]*lmody25coefficients[1]
xvec
0.8653004
> |
```

```
lmody      list of 12
lmody2     list of 12
lmody3     list of 12
lmody25    list of 12
svarmod    Large svector (13 elements,...
```

There is a website and I have essentially converted that they tie into a csv file. So, let me actually clear the screen here for you and read that data for you.

(Refer Slide Time: 01:42)

```
> Firedata <- read.csv("Firedata_Paradox.csv", headers=T)
Error in read.table(file = file, header = header, sep = sep,
+ quote = quote, :
  unused argument (headers = TRUE)
> Firedata <- read.csv("Firedata_Paradox.csv", header=T)
> |
```

```
.Firedata 58 obs. of 3 variables
 Damage : int 10 9 8 8 7 7 5 4 2 1 ...
 Severity: int 1 1 1 1 1 1 1 1 1 1 ...
 FFighters: int 1 1 2 2 2 3 3 3 4 5 ...
.Firedata...58 obs. of 3 variables
```

I will also post this data on the website; you can also play around with it. There are several ways of reading, you can actually use a read dot csv; I have saved the data as a comma separated variable file you can actually read sorry csv and the file name is at least in my case, it is called the fire data underscore paradox dot csv and I am going to say that yes there are headers in this file oops sorry. So, header equals true now what this

has done is, it has read the data from the csv file and stored the data in this data frame called fire data alright.

(Refer Slide Time: 02:37)



```
> cor(Firedata$Damage, Firedata$FFighters)
[1] 0.413483
>
```

Environment: base (64-bit x86_64-w64-mingw-x84_64)
Firedata 50 obs. of 3 variables
Damage : int 10 9 8 8 7 7 5 4 2 1 ...
Severity : int 1 1 1 1 1 1 1 1 1 1 ...
FFighters : int 1 1 2 2 2 3 3 3 4 5 ...
Firedata... 50 obs. of 3 variables

Fire List of In Objects with a Common Model

And what does this fire data contain, it shows that there are three variables here sorry; looks like I just there are three variables here called damage, severity and number of firefighters alright and it has 50 observations of these three columns. Now, let us compute the correlation between the damage caused by the fire and the number of firefighters. If you think of it physically, there should be a correlation between these two right; obviously, because the more the number of firefighters what would be the damage; higher would be the damage or lower; lower? Let us see if our correlation actually tells us that, let us compute here; fire data and remember this is a data frame. In fact, I can probably use even the dollar yes, so damage here that is our; let us say Y variable and X variable is our firefighters, correlation does not worry about the direction, so I can interchange the order of variables here.

This is the correlation that I get; do you agree with the sign of the correlation estimate that we get. What do you think, so this says that there is a positive correlation if you send more firefighters, more damage will occur which is probably good news for the firefighters. So, that they say well I do not to go because if I go the damage is increased, but; obviously, it does not make any sense right. Now this is a problem with correlation that is one of the problems right that the sign is completely counterintuitive, it is just a

correlation. In many cases, it may even indicate a non 0 correlation when there is none; in this case we are worried about the sign between a general scenario, we are also worried about the magnitude. Many a times correlation can turn out to be non 0 when truly there is no direct relation between them.

So, now let us ask how to resolve this issue and the issue is resolved by using what is known as a conditional measure. What do we mean by condition measure here in the data set that you just saw in addition to the damage and the firefighters, we had an measurement of another variable called the severity, which is a measure of how huge the fire was; had I taken that into account perhaps I could have come to a better conclusion; I mean correlation would have made more sense. So, in general X and Y may be related through another variable Z.

(Refer Slide Time: 05:54)

Probability & Statistics - Review 2

Conditional correlation

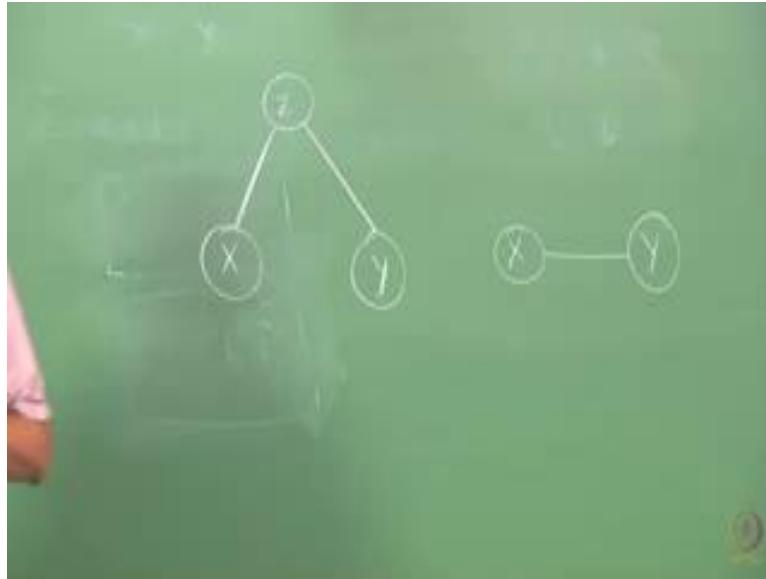
► If the **conditional correlation** vanishes, then the connection is purely indirect, else there exists a direct relation

Correlation measures **total** (linear) connectivity, whereas **conditional** or **partial** version measures **direct** association.

Amir H. Tangirala Applied TSA August 18, 2018

And this Z we call as a confounding variable or the mediating variable, in some situations Z is also known as a suppressing variable, depending on the nature of the effect the confounding variable has on the correlation between X and Y. So, what is happening in general is we have only X and Y and we are just measuring the correlation.

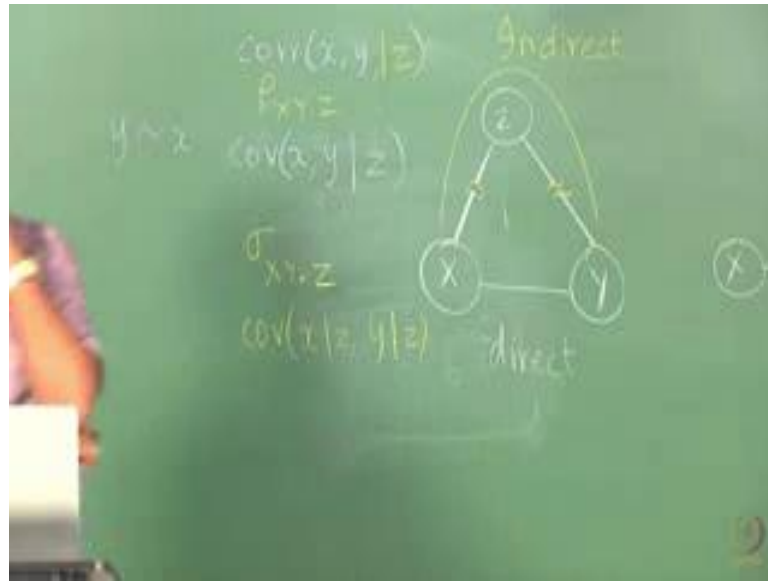
(Refer Slide Time: 06:31)



So, one of these possible situations is that we have X here X and Y and Z, it is likely that this is the situation, there is no correlation between X and Y direct correlation. So, the key word is direct alright, probably this is how X Y and Z are connected. If we do not measure Z and we only measure X and Y then what would happen is that X and Y would appear connected. In the firefighting example yes there is also this relation agreed but in many situations this direct relation may not exist at all.

So, now when Z is not measured X and Y may appear connected, note that I have not drawn the direction of the association between X and Y because correlation does not measure the direction. So what we would like to know is whether this direct link is present or not as you see on the schematic; I would like to know, have a measure that will tell me whether an X whether X and Y are directly related, how will I know that; I will have to necessarily have the measurement of Z with me.

(Refer Slide Time: 07:47)



Now, I will not just estimate correlation between X and Y, but instead in fact, let us talk about covariance first and then talk about correlation. So, I would like to have now what is known as a conditional covariance given Z, I would like to compute X and Y. So, you are asking in the presence of Z; now tell me whether X and Y are related. So, the point is correlation measures what is known as total connectivity, what we mean by total is; suppose this is a situation and you compute correlation between X and Y without taking Z into account.

Remember X and Y are now related along this direct pathway and also this indirect pathway. Correlation measures the dependence between X and Y along both pathways and what do we want to know whether there is a relation along the direct pathway. This is one of the fundamental ideas in causality analysis, in network reconstruction from data which are some of the hot areas today in data analysis. So these conditional measures are also known as partial measures. In calculus also we come across this kind of concept; total derivative versus partial derivative right we use d and ∂ , but of course, we do not use the different symbols here.

There are many ways of writing this conditional covariance symbolically for example, you could say $\sigma_{X,Y \cdot Z}$, you could use this notation or sometimes people would say given Z that is up to you, we can use the dot operator notation here and likewise correlation; conditional correlations would be denoted as $\rho_{X,Y \cdot Z}$ and we want to

know how to compute this $\sigma_{X Y \cdot Z}$, $\rho_{X Y \cdot Z}$ is computed from $\sigma_{X Y \cdot Z}$. Any ideas on how I could compute this condition covariance or correlation; let us talk about covariance, any no; wild idea what is it that we want to do we want to cut off this pathway, we want to sever these links, does it give you some idea?

Student: (Refer Time: 10:35).

What would be that; that is correct so. In fact, covariance of X and Y given Z would be actually covariance you are right; covariance between X given Z and Y given set, but how does one compute in practice is X given Z and Y given Z, what do you mean by X given Z and Y given Z sorry I do not understand.

Student: (Refer Time: 11:05).

Can I go to r and say how do you envisage that we can do this.

Student: (Refer Time: 11:17).

That would not help; in general it may not help. Any other ideas?

Student: (Refer Time: 11:32).

Sorry.

Student: (Refer Time: 11:34).

It is a function of Z, how do you do that?

Student: (Refer Time: 11:44)

What is this idea of X given Z and Y given Z? So, let me explain that you kind of you know getting some idea, but we have to be lot more clear because we have to spell out a procedure and that procedure has to be very sound; theoretically first and then of course, a practical one will follow. So, the procedure is to compute covariance between first step is to regress X on to Z and Y on to Z. The general concept of conditional measure is a very generic one, since we are dealing with covariances and covariances are linear measures; what we are trying to do is we are trying to now sever the any linear dependency that Z linear influence that Z can have on X and Z can have on Y alright.

(Refer Slide Time: 12:39)

$$\begin{aligned} X &\leftarrow Z & \hat{X}(z) & & \varepsilon_{X|Z} = X - \hat{X}(z) \\ Y &\leftarrow Z & \hat{Y}(z) & & \varepsilon_{Y|Z} = Y - \hat{Y}(z) \\ \text{---} & & & & \\ \text{---} & & & & \\ \text{---} & & & & \\ \hat{X} &= & bZ & & \hat{Y} = dZ \\ b &= & \frac{\sigma_{XZ}}{\sigma_Z^2} & & d = \frac{\sigma_{YZ}}{\sigma_Z^2} \end{aligned}$$

So, what do we mean by severing in this link, you regress Z on to X and likewise here Z on to Y here and in other words build your best predictor for X or prediction of X in terms of z. So, let us call that as the star an optimal prediction of X using Z alone, if I were to give you Z; what would be a prediction of X and likewise what would be your optimal prediction of why using Z alone. Of course, using a linear model we will restrict ourselves to linear models and this is obtained by our procedure that we have learnt earlier, build a linear model between X and Z and then build your optimal model from there you compute your optimal predictions. Whatever is left over is what we call as the residual and these variables are essentially conditioned variables conditioned on z.

You follow; first we predict X using Z alone, whatever influence Z has on X is captured in your prediction and once you discount for that from your original X whatever is left over is a residual, if it so happens that Z completely is able to completely explain X then epsilon is 0 then; that means, X is nothing, but Z itself, but that is rarely going to be the case likewise here for Y; yes.

Student: (Refer Time: 14:15).

Yes.

Student: (Refer Time: 14:20).

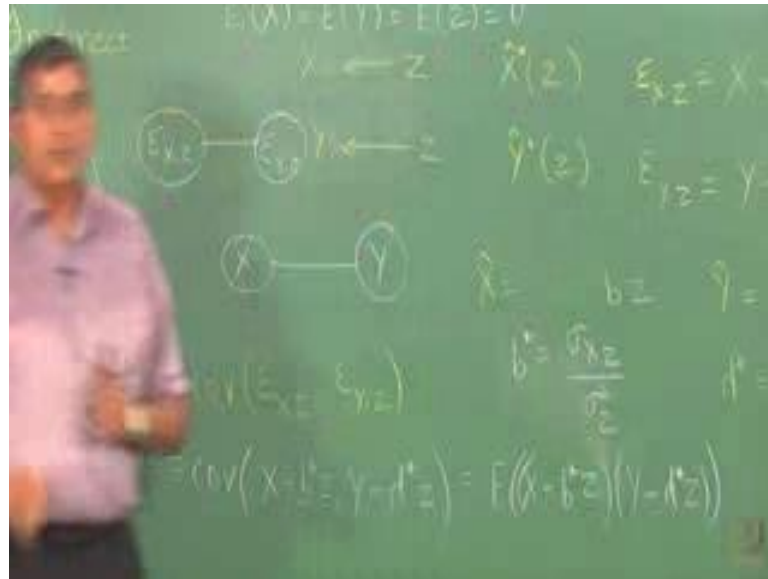
It is, but I mean if you are looking at a general predictor here we restrict ourselves to a linear predictor. Yes in general you are right, this is a conditional expectation of X given Z , but we do not break our head with condition expectations now, we are living in the linear world. So, we will build a linear model to construct this in other words I would build here X equals; a plus $b z$ this is the predictor that I would fit and then go through a regular procedure to obtain optimal estimates of a and b from which you compute the optimal prediction.

So, you do this essentially you fit two regression models and work with the residuals and compute the correlation between the residuals. So, that is all and that gives you the conditional covariance of course, you have to ask what assures that this conditional covariance gives me a direct correlation, you have to also be assured of that correct that is something that we will shortly see.

Now, what is the optimal prediction of here; what is optimal prediction here the optimum prediction is obtained by first obtaining optimal estimates of a and b for simplicity assume that there is no a then we know I mean regardless of whether a is present or not we know that b star is nothing, but σ_{XZ} by σ^2_X , we have derived this already, this is optimal estimate in what sense the one that minimizes the mean square error expectation of X minus \hat{X} square. Likewise if I were to fit a model for Y ; \hat{Y} let us say we call this as some d times Z ; then d star would be σ_{YZ} by σ^2_Z here.

So, what is our optimal prediction now; optimal prediction \hat{X} is this b star time Z likewise optimal prediction of Y is b star time Z , am I right. Any questions until this point, these are something that we have derived even in yesterday's class.

(Refer Slide Time: 16:46)



And what we are saying now is $\sigma_{\epsilon_X \epsilon_Y}$ is covariance between the epsilon X dot Z and epsilon Y dot Z. Can you compute that, all you have to do is plug in this expression into this and likewise Y hat star into this and then compute the covariance, you will get a nice expression ultimately I would like you to give me the answer in terms of the covariances and the variances of X Y and Z and so.

Any difficulty; no you can ask feel free to ask because sometimes it is easy to get confused here at this point, I will just write the next step here is covariance between X minus b star Z and Y minus d star Z; is that clearer now? You have any difficult no room to compute, tightly packed yeah please make it a habit to get through this get a hang of this kind of theoretical analysis because otherwise you can easily get lost in this course. So, if you have difficulty (Refer Time: 18:21) will help you also.

Student: (Refer Time: 18:25).

Yes.

Student: (Refer Time: 18:30).

The relevance is exactly this one that is it will now tell me whether this link exists or not.

Student: (Refer Time: 18:39).

Well in general if I am given X and Y , what does covariance tell me right. So, this edge here that we have drawn graphically all they have not spelt it out, this h means the presence of a linear relationship because we are talking of linear work covariance will essentially tell me whether an edge exists that connects these two nodes alright. So, this epsilon that we have constructed essentially have allowed us to cut off these links; that means, by subtracting X hat star of Z from X and Y hat star or Z from Y graphically it amounts to severing this links; that means, now this graph is being redrawn without Z in this way epsilon X dot Z and here epsilon Y dot Z and now covariance between these two will tell me whether there is an edge here and that is the same edge that appears here.

Because this edge exists regardless of these two pathways if it exists we do not know if it exists. So, we sever these two mathematically through this operation here is a graph and here is a mathematical operation. So, this mathematical operation here amounts to severing these two links and now the statistical measure covariance will tell me whether this edge exists by looking. So, now I have taken Z away from the graph in other words I have taken the effect of Z and kept it aside.

Oh yeah all that is possible here because the course is simpler when it to begin with that is all. If I use a non-linear one I can do that, it will only enhance the dropouts that I have do; we can use as I said this concept of conditioning is a very generic one. We are conditioning only in the linear sense you can of course, build a non-linear predictor for X in terms of Z then you will be conditioning on the non-linear effects, but at the moment let us understand how the linear world wars that is all, otherwise you are right; you can do that; somebody else yes.

Student: (Refer Time: 21:00).

Possible we do not know we want to check.

Student: (Refer Time: 21:08).

Between X and Z also yeah it is possible.

Student: (Refer Time: 21:12).

Yes absolutely. So, we are saying in whatever possible it is a very good point we are saying in whatever possible ways that effects X , let me take that into account, it is a very

valid point, but what we want to do is; we want to remove the influence of Z completely on X along all pathways. I understand what you are saying is I should have only severed this link, what is a guarantee and that is a reason we are also conditioning Y. There is another concept called semi partial correlation, which will partly address the issue that you have raised.

Your point is I should have only removed the effect of Z, direct effect of Z on X. If I do that; if I attempt to do that, I will still be left with an X after discounting which will contain some effects of Z right. See numerically I would never be able to figure that out, my effort should be to remove the effect of Z completely on X and that is why as you rightly pointed out, the covariance appears here and because the idea is to; although I have shown here graphically as if I am severing this link, what we are essentially doing is we are severing all links of X with Z; along all pathways, but it is a good point.

So, do we have an answer from the other room, it is easy no; you should be comfortable doing this. What would you get when you start doing your covariance, do not have to necessarily do your expectations; you can do your expectation for example, here you can assume X and Y to be 0 mean and Z also to be 0 mean, if that makes your life comfortable, but when you have an expression like this; you can write covariance as here covariance between X minus b star Z times Y minus; in other words you can apply the expectations; assume X Y and Z to be 0 mean; just for simplicity so that you can simply write this as expectation of X minus b star Z times Y minus d star Z; what do you get.

Somebody should have the answer, what is the first term that you see in this expansion.

Student: (Refer Time: 24:05).

Sigma X Y very good and then minus something here the answer, now if you have a difficulty you should seriously ask; is there a difficulty, you are stuck; are you able to get the answer. Simply expand this, when you are lost go in a stepwise fashion; now expand this and apply your expectations and then make use of the fact that b star is what I had written on the board earlier; b star is $\frac{\sigma_{XZ}}{\sigma_Z^2}$.

Student: (Refer Time: 24:50).

Such as.

Student: (Refer Time: 24:52).

Yeah, but since I have written X Y and Z to be of 0 mean that is why I said assume for simplicity.

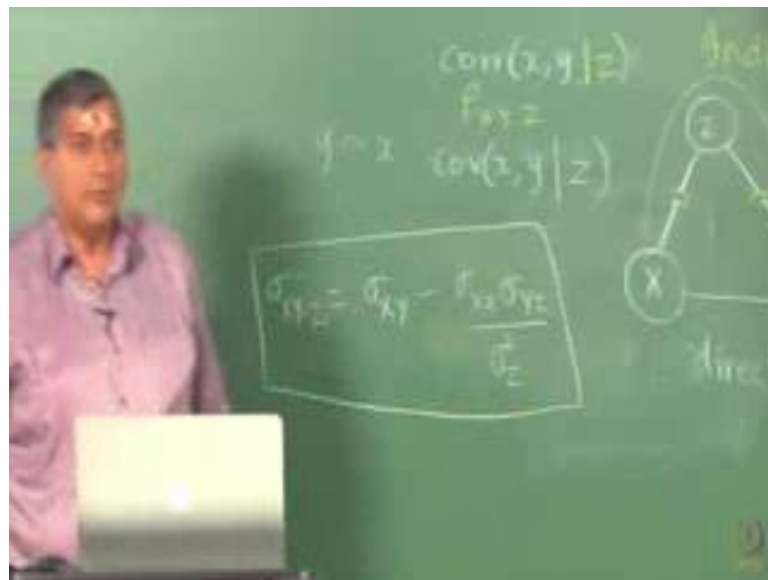
Student: (Refer Time: 25:04).

Minus.

Student: (Refer Time: 25:10).

Let us see.

(Refer Slide Time: 25:27)



I have not given the expression for sigma X Y dot Z, but yes you can write it a very quickly; let me write here on the top. So, sigma X Y dot Z would be sigma X Y minus what do you get; sigma X Z times by very good that is correct. How would you compute the conditional correlation; all you have to do is to compute the conditional correlation you say it is the covariance; this covariance by square root of the product of the variances of the individual residuals and you should work out the math to get the expression that you see, but before we adjourn for today, let us quickly understand take a minute and understanding this expression that we have here. What does it tell us that the conditional covariance is the total covariance discounted; that is after discounting for the total covariance see sigma X Y is the total covariance or you can say the unconditional

covariance; less this what does this tell me, if somehow Z did not affect X at all then what would be the case.

The conditional and unconditional ones are the same; likewise if Y is not influenced by Z at all then again you get the same thing. Now how do we guarantee that this is going to measure the edge here you can prove that, but best is to look at this example.

(Refer Slide Time: 27:16)

Probability & Statistics - Review 2

Partial correlation: Example

Consider two random variables $X = 2Z + 3W$ and $Y = Z + V$ where V , W and Z are zero-mean RVs. Further, it is known that W and V are uncorrelated with Z as well as among themselves, i.e., $\sigma_{VW} = 0$.

Evaluating the covariance between X and Y yields

$$\sigma_{YX} = E((2Z + 3W)(Z + V)) = 2E(Z^2) = 2\sigma_Z^2 \neq 0$$

although X and Y are not "directly" correlated.

Ann H. Tongolo Applied TSA August 16, 2018

And I will just leave you with this example; I have two variables X and Y X equals 2 Z plus 3 W and Y equals Z plus v; this is the example that I want you to work on in your assignment V, W, Z are all 0-mean random variables and they are all uncorrelated with Z and now evaluating the covariance between X and Y gives us this sigma Y X simple; that is if you take the simple covariance between X and Y taking into account the fact that all these variables are 0 mean; you end up with this answer 2 sigma square Z.

But if you look at the relation there for X and Y X is made up of 2 Z plus 3 W and Y is being constructed as Z plus V. W and V are uncorrelated; which means the only common thing between X and Y is Z. So, there is if you look at the relations there is no way Y is directly affecting Z, Y is only apparently affecting X.

(Refer Slide Time: 28:35)

Probability & Statistics - Review 2

Partial correlation: Example

Now,

$$\rho_{YX} = \frac{\sigma_Z^2}{(\sigma_Z^2 + \sigma_V^2)(4\sigma_Z^2 + \sigma_W^2)}; \quad \rho_{YZ} = \frac{1}{\sqrt{1 + \frac{\sigma_V^2}{\sigma_Z^2}}}; \quad \rho_{XZ} = \frac{1}{\sqrt{1 + \frac{\sigma_V^2}{4\sigma_Z^2}}};$$

Next, applying (22), it is easy to see that

$$\rho_{YX.Z} = 0$$

Amir H. Toghiani Applied TSA August 16, 2018

If I remove the effect of Z that is if I compute the conditional covariance then the rho X Y dot Z turns out to be 0; clearly telling me that Z was the only mediating or the confounding variable as far as X and Y are concerned, they are not directly connected at all. So this is the power of this partial correlation or partial covariance; it tells of course, you need to be given the measurements of the confounding variable, here the confounding variable is Z.

So, the general question that one asks is do I have a knowledge of all the confounding variables, all the mediating variables answer is no. In fact, we may not even know the list of all confounding variables. Suppose I am looking at temperature and pressure or any other two variables, in the fire example the confounding variable was the severity, if you were to look at the confounding variable; the partial correlation in that case then it will turn out to be completely different from the correlation and the package that does that for you is the p cor package have given the information in the slides.

(Refer Slide Time: 29:40)

```
> cor(Firedata$Damage, Firedata$FFighters)
[1] 0.4
> pcor(Firedata)
  Firedata_
  Firedata_Paradox
```



And all you have to do is use p cor from that routine and simply compute the partial correlation between those this function with those variables here. So, let us quickly do that and then we will adjourn. So I want to compute partial correlation, now between the firefighters, the damage and the severity and the p cor routine does that for me, it is not a part of r package; you have to install an additional root library for it and the way this p cor works is; you will have to supply the data frame or the matrix and straightaway it gives you the partial correlation.

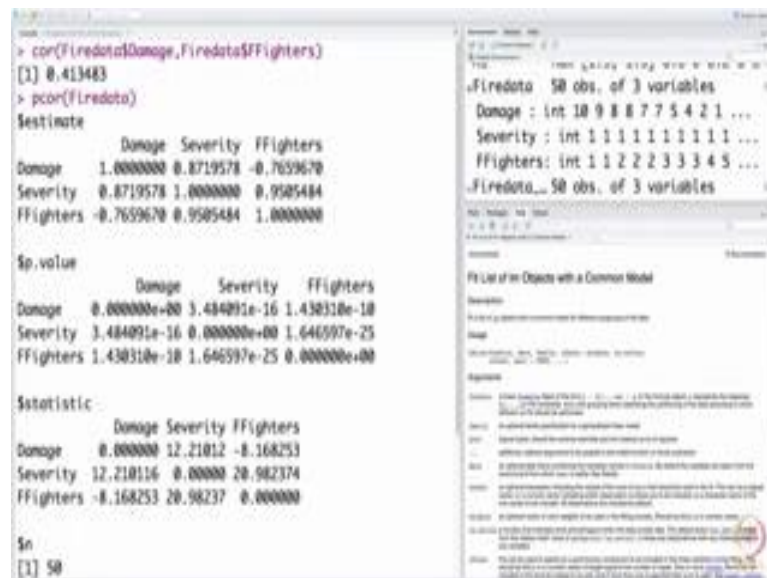
(Refer Slide Time: 30:23)

```
> cor(Firedata$Damage, Firedata$FFighters)
[1] 0.413483
> pcor(Firedata)
$estimate
      Damage Severity FFighters
Damage  1.0000000 0.8719578 -0.7659670
Severity 0.8719578 1.0000000  0.9505484
FFighters -0.7659670 0.9505484  1.0000000

$P.value
      Damage Severity FFighters
Damage  0.000000e+00 3.484091e-16 1.430310e-10
Severity 3.484091e-16 0.000000e+00 1.646597e-25
FFighters 1.430310e-10 1.646597e-25 0.000000e+00

$Statistic
      Damage Severity FFighters
Damage  0.000000 12.21012 -8.168253
Severity 12.21016 0.00000 20.982374
FFighters -8.168253 20.98237 0.000000

$N
[1] 50
```



So, it is very simple, you simply supply the matrix of your data that you have and of interest is this estimate that comes out here. So, if you look at these columns here, what do you see here? These are nothing, but your partial correlations. How do I read these partial correlations for example, I am interested in the partial correlation between the damage caused by the fire and the number of firefighters, where is that reported here. How different it is from your earlier correlation, this is negative sign. So, at least it is good and there is a strong correlation it there better be which is good news, alright. So, this is how partial correlation is useful in overcoming certain limitations of the correlation, you notice that partial correlation is also symmetric measure.