CHAPTER 8

Pair trade Method 2, Chapter 1 (PTM2, C1) – Straight line Equation

8.1 – A straight relationship

Today happens to be 14th of Feb, people around me are excited about Valentine's Day, they are busy celebrating love and relationships. I think Valentine's Day is a packaged affair, meant to boost the revenues of restaurants, jewellers, and gift shops, but then it's just me and my random thoughts.

Anyway, given its valentine's day, I thought it would be a perfect idea to discuss relationships. Don't worry, I'm not going to bore with a clichéd love story or give you any unsolicited advice on maintaining a great relationship, rather I'll talk to you about two sets of numbers and how you can measure the relationship between them if at all there exists one.

In the process, I'll attempt to take you back to your school days, well, at least back to your

high school math class

A quick recap here – Chapter 1 to 7 of this module, we discussed a rather simple technique of pair trading. This was as taught by Mark Whistler. Moving forward from this chapter, we will discuss a slightly more advanced technique of pair trade. This is also called '**Statistical Arbitrage**' or '**Relative value trading**' or RVT in short.

So here we go.

Do you remember the time your math teacher discussed the equation of a **straight line** in the class? If you were like me, you'd have promptly ignored the lecture and looked outside of the window, quietly rebelling against the mainstream education.

But then, if only the teacher had said 'learn this, you'll make money off it someday', the interest level would have been totally different!

Anyway, life always gives you a second chance, so this time around, pay attention, and hopefully, you will make some money off it

The equation of a straight line reads something like this –

$Y = mx + \varepsilon$

Click here for a detailed explanation, or continue reading for a bare bone explanation.

Before we discuss the equation, a quick note on the notations used –

y = Dependent variable

M = Slope

X = Independent variable

E = Intercept

The equations states, the value of a dependent variable 'y' can be derived from an independent variable 'x', by multiplying x by its slope with y' and adding the intercept 'e' to this product.

Sounds confusing? I guess so

Let me elaborate on this and by the way before you start thinking why we are discussing the straight line equation instead of relative value trading (RVT), then please be rest assured, this concept has deep relevance to RVT!

Consider two fitness freaks, let's call them FF1 and FF2, between the two, FF2 is the kind of guy who wants to go that step extra and something more than what FF1 does. So if FF1 does 5 pushups, FF2 does 10. If FF1 does 20 pull-ups, then FF2 does 40. So on and so forth. Here is a table on how many pushups they did Monday to Saturday –

Day	FF1	FF2
Monday	30	60
Tuesday	15	30
Wednesday	40	80
Thursday	20	40
Friday	10	20
Saturday	15	???

Now, if you were to guess the number of push-ups FF2 would do on Saturday, what would it be? I guess it's a no-brainer, it would be 30.

This also means – the number of pushups FF2 does, is kind of dependent on the number of pushups FF1 does. FF1 does not really bother about FF2, he will go ahead and do as many pushups his body permits, but FF2, on the other hand, does twice the number of pushup as FF1.

So this makes FF2 a dependent variable and FF1 an independent variable. Or in the straight line equation, FF2 = y and FF1 = x.

FF2 = FF1*M + ε

In simple English, the equation reads like this –

The number of pushups FF2 does is equal to the number of pushups FF1 does, multiplied by a certain number, plus a constant.

That certain number is called the slope (M), which happens to be 2, and the constant or ε happens to be 0. So the equation is –

FF2 = FF1*2 + 0

I hope this is fairly clear now. Let me copy paste the definition I had posted earlier –

The straight line equations states, the value of a dependent variable 'y' can be derived from an independent variable 'x', by multiplying x by its slope with y' and adding the intercept 'e' to this product.

Now, think about another case –

There are two hungry men, let's call them H1 and H2. Just like FF1 and FF2, H2 eats twice the number of paratha as H1 plus 1.5 more. For example, if H1 eats 2 parathas, then H2 will eat 4 plus eat another 1.5. H2 will always ensure he eats that extra 1.5 parathas, no matter how full he is.

So here is the table which gives you count of how many parathas these two hungry men ate over the last 6 days –

Day	H1	H2
Monday	2	5.5
Tuesday	1.5	4.5
Wednesday	1	3.5
Thursday	3	7.5
Friday	3.5	8.5
Saturday	4	???

If you notice, H2 (who is really hungry, all the time), eats twice as much as H1 plus 1.5 paratha extra. So on Saturday, he will eat –

4*2 + 1.5 = 9.5 paratha!

Remember, the number of parathas H2 eats is dependent on how many parathas H1 eats. H1, on the other hand, eats till he is satisfied. Given this, let us a construct a straight line equation for these two hungry men, just like the way we did for the two fitness freaks.

H2 = H1*2 + 1.5

Here, H2 is the dependent variable, whose value is dependent on H1. 2 is the slope, and 1.5 is the constant.

Before we proceed, let's make a small change in the paratha example, think of 'Y' as a diet conscious person. Every day, irrespective of how hungry or full Y is, he eats just 1.5 parathas. Not a morsel more or not morsel less.

So, X eats 3 parathas, Y eats 1.5, X eats 5, Y eats 1.5, X eats 2.5, Y eats 1.5. So on and so forth. So what do you think the equation states?

$y = x*0 + 1.5$

The slope here is 0, hence, y is not really dependent on x, in fact, the value of y is a constant of 1.5, which is quite obvious. Hopefully, you get the point by now on how you can relate two sets of numbers.

Now forget the fitness, forget the parathas, I'll give you two sets of random numbers –

X	Y
10	3
12	6
8	4
9	17
20	36
18	22

X is the independent variable and Y is the dependent variable. Given this, do you see a relationship between these two sets of numbers here? Eyeballing the numbers suggest that there is no relationship between X and Y, definitely not like the one which existed in the above two examples. But this does not mean that there is no relationship between the two at all. It's just the relationship is not obvious to the naked eye.

So how do we establish the relationship between the two? To be more precise, how do we figure out the values of the slope' and the constant 'ε'?

Well, say hello to linear regression!

I'll introduce the same to you in the next chapter.

Key takeaways from this chapter

1. A straight line equation can define the relationship between two variables
2. Of the two variables, one of it is dependent and the other one is independent
3. The slope of a straight-line equation, represented by 'm' helps you identify the extent by which the independent variable has to be scaled
4. The term ε represents a constant term
5. If the slope is zero, the Y = ε
6. Sometimes, the relationship between two variables is not obvious
7. When the relationship is not obvious, one can identify the relationship by employing a statistical technique called 'Linear regression'.

CHAPTER 9

PTM2, C2 – Linear Regression

9.1 – Introduction to Linear Regression

The previous chapter laid down a basic understanding of a straight line equation. To keep things simple, we took a very basic example to explain how two variables can be related to each other. Needless to say, the examples were selected in a way that casual eyeballing could reveal the relationship. Towards the end of the chapter we posted a table containing two arrays of numbers – the task was to figure out if there was a relationship between the two sets of numbers, if yes, what how could one express the relationship in the form of a straight line equation. More precisely, what was the intercept and constant?

We will figure how to establish a relationship in this chapter and move closer towards the relative value trading technique. For convenience, let me post the table with the two number arrays once again –

X	Y
10	3
12	6
8	4
9	17

20	36
18	22

Clearly, casual eyeballing does not reveal any information about the relationship between the two sets of numbers. Maybe it does, if you are a mutant, but for a mere mortal like me, it does not work.

Under such circumstances, we rely upon a technique called the 'Linear Regression'. Linear regression is a statistical operation wherein the input is an array of two sets of numbers and the output contains many different parameters, including the intercept and constant needed for constructing the straight line equation.

To perform the linear regression operation, we will depend on the good old Excel. Here is the step by step guide to perform a simple linear regression on two arrays of numbers. Be prepared to see a lot of screenshots and instructions

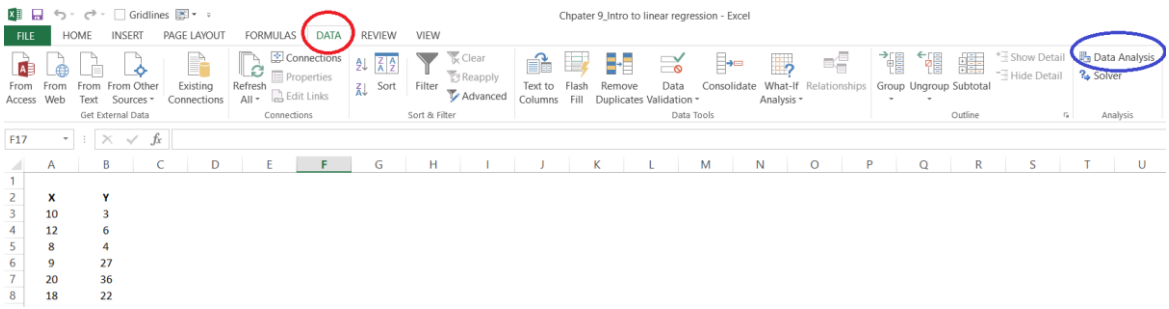
Step 1 – Install the Plugin

Open a fresh excel sheet and insert the values of X & Y as seen in the above table. I've done the same as shown below –

X	Y
10	3
12	6
8	4
9	27
20	36
18	22

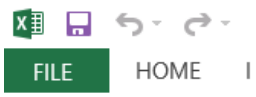
This is our data set. Do remember, Y is the 'Dependent' variable whose value depends on the independent variable X. Both X and Y will be the input variables for the linear regression operation.

On the excel sheet, click on the Data ribbon as highlighted in red, shown below –



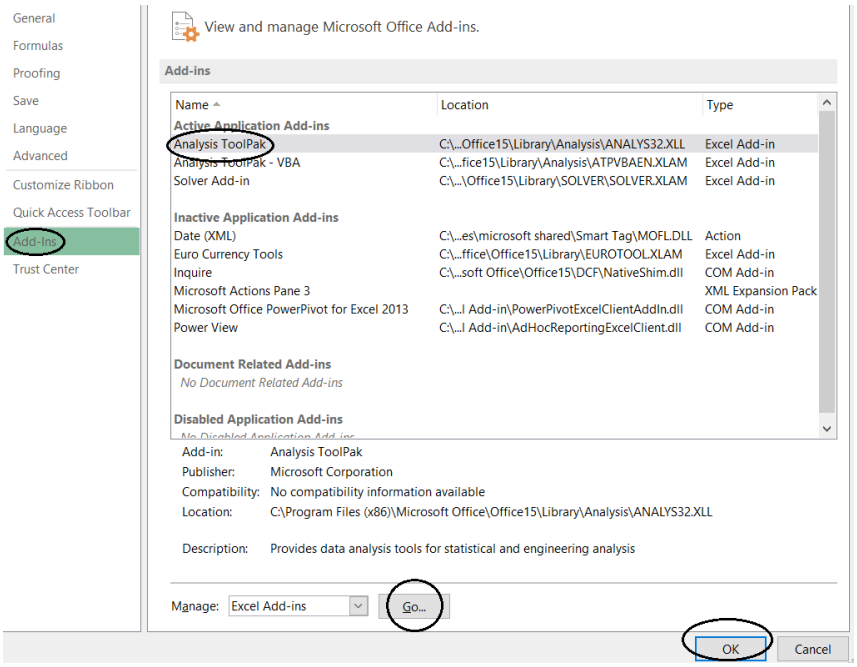
The data ribbon will now show you the 'Data Analysis', option. This is highlighted in blue. Now, some of you may not see this option, if yes, don't panic. I'll tell you what needs to be done.

Click on 'File' –



This will open up a new window, and on your left-hand side panel, you will see an option to select 'option' –

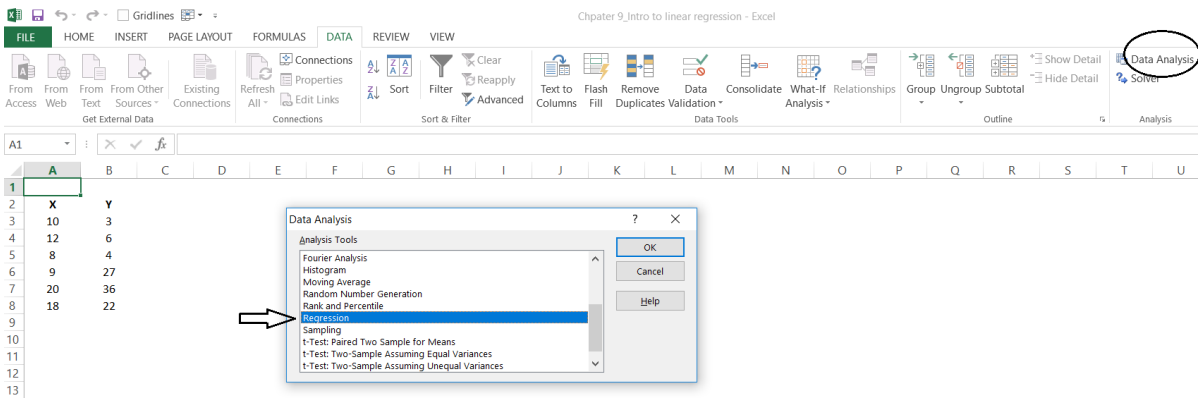
Click on the Options, and you will see a bunch of general options to work with. On the left-hand panel, select 'Add-Ins', click on it and then click on the 'Analysis Tool pack'. Then click on 'Go', and finally on 'Ok'. With this, you'd essentially added the 'Data Analysis' option to the data ribbon.



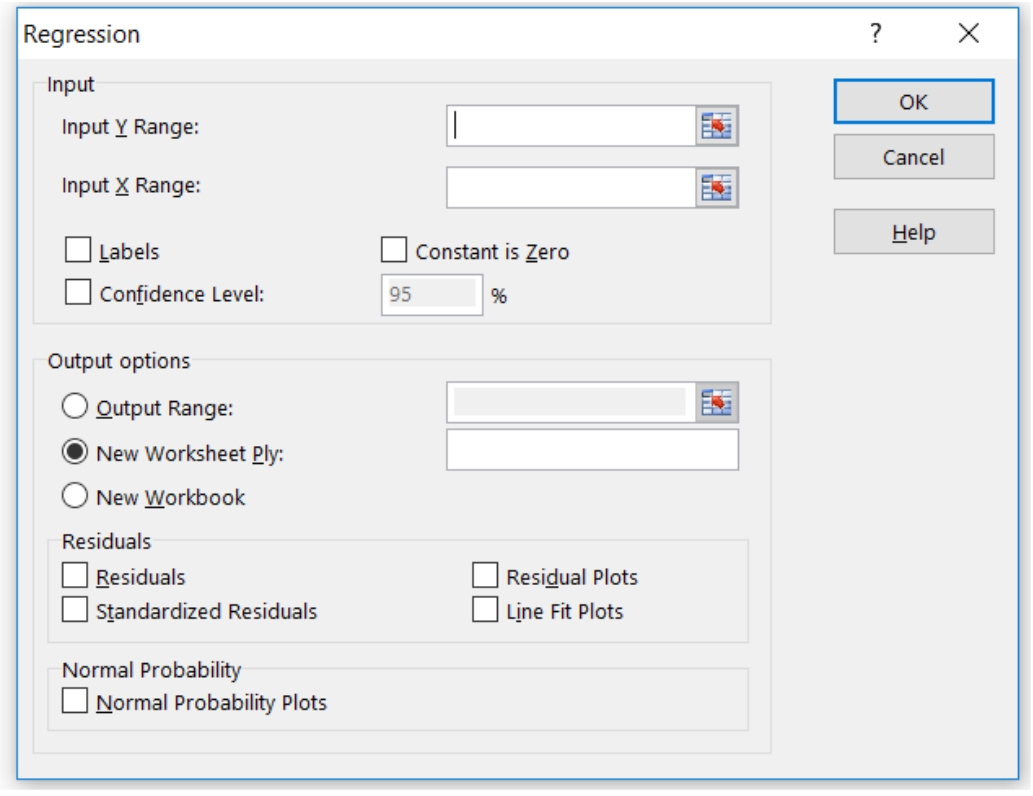
Close the excel sheet and restart your system and you are good to roll.

Step 2 – Enter the values

So we proceed further based on the assumption that your excel sheet has the data analysis pack. The next step is to invoke the linear regression function within the data analysis pack. To do this, click on the 'Data' ribbon, and select the Data Analysis. This will open up a pop-up, which will have a list of statistical operations which you can perform on data sets. Select the one which says 'Regression'.

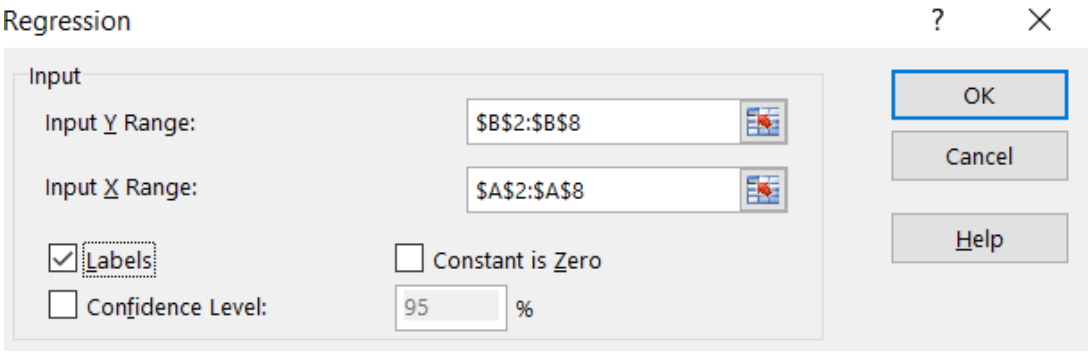


Select regression and click ok, you will see the following pop up –



As you can see, there are a bunch of fields here. I'd suggest you pay attention to the first section, which is the input section. There are two fields here – 'Input Y Range' and 'Input X Range'. As you may have imagined, Y is for the dependent variable and X is for the dependent variable.

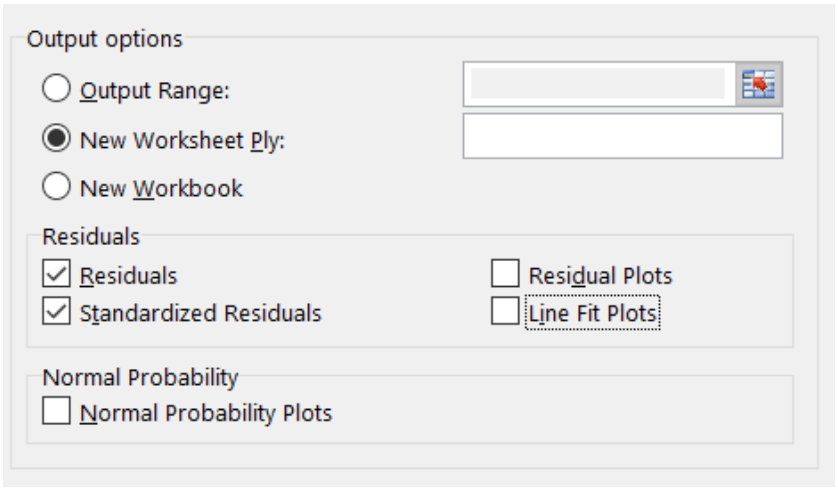
This is where we feed in the X and Y series data. To do that, click on the input channel and select Y and X range –



Also, please notice that I've checked the label box, this indicates that the first cell value i.e. A2 and B2 contain the series label i.e. X & Y respectively.

I'd suggest you ignore the other input values for now.

On the output side, ensure you've clicked the following –



Selecting 'New worksheet', ensures that the output data is printed on a new worksheet. I've also clicked on two other variables called – Residuals and Standardized Residuals. I will talk about these two variables at a later point. For now, just ensure they are selected.

With this, you are good to perform the linear regression operation. Click on the 'Ok' button which is available in the right-hand top corner.

Excel will now take these inputs and perform the linear regression operation, the results will be posted in a new sheet within the same workbook.

9.2 – Linear Regression Output

So here is how the linear regression output looks and as expected, the summary of the output is presented in a new sheet.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

Sheet3 Sheet1 Sheet2

Agreed, the summary output is quite scary at the first glance. It has lots and lots of information. We will unravel this output in bits and pieces as we proceed.

For now, let's concentrate on finding our slope and intercept. I've highlighted this for you in the below snapshot –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

The data points highlighted in red contains the coefficients we are looking for i.e. the intercept (or constant) and the slope (denoted by x).

Some of you may be confused with the slope being represented by x, I understand its misleading, it would have been best if it was M instead of x as it would match the straight-line equation, but then I guess we will have to live with x for slope.

So,

- Slope of the equation = 1.885
- Intercept (or constant) = -7.859813.

Given this, the straight-line equation for the arbitrary set of data is –

y = 1.885*x + (-7.859813) or

y = 1.885*x – 7.859813

So what does this really mean?

Well, if you recollect from the previous chapter, this equation essentially helps us predict the value of y or the dependent variable for a certain x. Let me repost the table here for the sake of convenience –

X	Y
10	3
12	6
8	4
9	17
20	36
18	22
15	??

I've added a new data point for x here i.e. 15, now using the slope and intercept, we can predict the value of y. Let's do that –

y = 1.885 * 15 – 7.859813

= 28.275 – 7.859813

= **20.415**

So, if x is 15, then most likely, the predicted value of y is 20.415.

How accurate is this prediction, you may ask?

Well, it's not accurate. It is only an estimation. For example, consider the value of x is 18 (refer to the last but one data point), then according to the straight line equation, the value of y should be –

y = 1.885*18 – 7.859813

= 33.93 – 7.859813

= 26.07019

However, the actual value of y is 22.

This leads us two values of y –

1. Predicted value of y via the straight line equation
2. Actual value of y

The difference between the two values of y is called **the residuals**. For example, the residual for y (difference between actual and predicted y), when x = 18 is

26.07019 – 22

= **4.070187**

The summary output when you perform linear regression also contains the residuals, I've highlighted the same in the snapshot below –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Standard Residuals
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

I've also highlighted the residual when x = 18, which is what we calculated above.

To give you a heads up – the bulk of the focus for carrying out the relative value trade depends on the residuals. Stay tuned!

Download the excel sheet **here**.

Key takeaways from this chapter

1. Linear regression is a statistical operation which helps you construct a straight line equation
2. Linear regression can be performed on excel. One needs to install the excel plugin to perform linear regression
3. Amongst many other output variables, linear regression gives out the values of the slope and intercept
4. With the help of the slope and intercept, one can predict the value of y
5. The difference between actual y and predicted y is called the residual
6. The residual is also a part of the output summary

CHAPTER 10

PTM2, C3 – The Error Ratio

10.1 – Who is X and who is Y?

I hope the previous chapter gave you a basic understanding of linear regression and how one can conduct the linear regression operation on two sets of data, on MS Excel. Remember, we are talking about two variables here – X and Y.

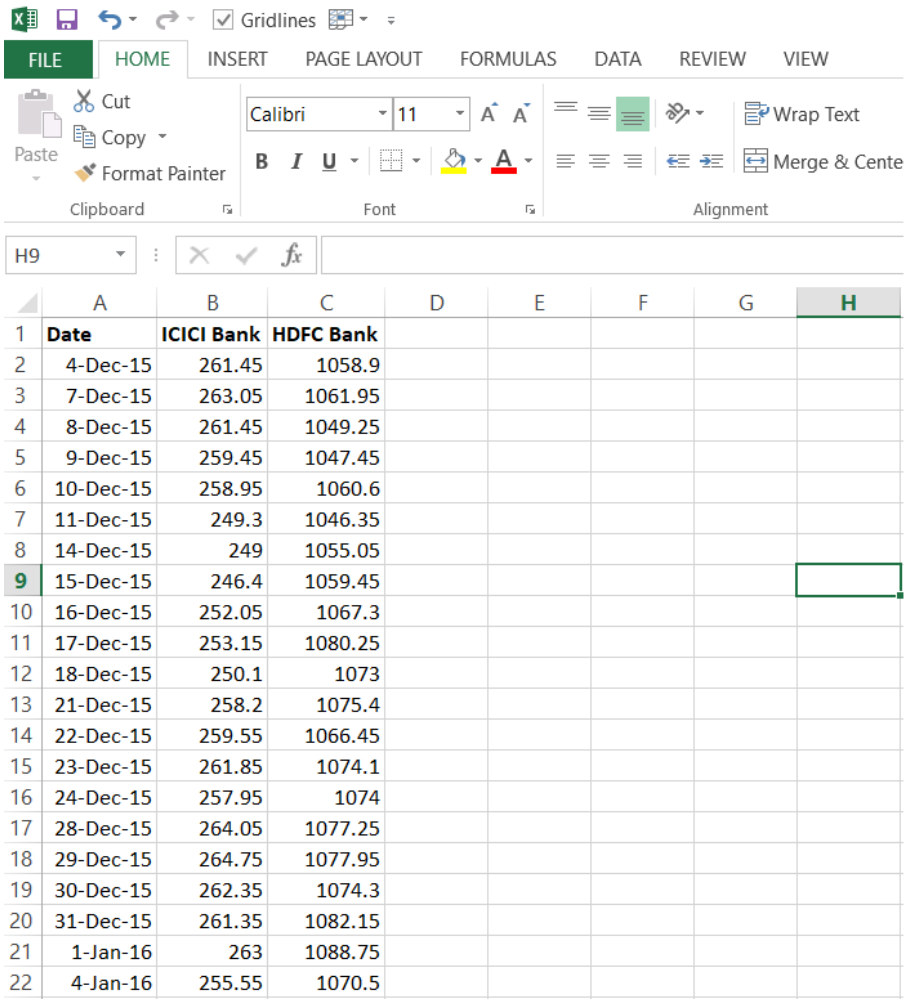
X is defined as the independent variable and Y is the dependent variable. If you've spent time thinking about this, then I'm certain you'd have guessed X and Y will eventually be two different stocks.

In fact, let us just go ahead and run a linear regression on two stocks – maybe HDFC Bank and ICICI Bank and see what results we get.

I'm setting ICICI Bank as X and HDFC Bank as Y. A quick note on data before we proceed –

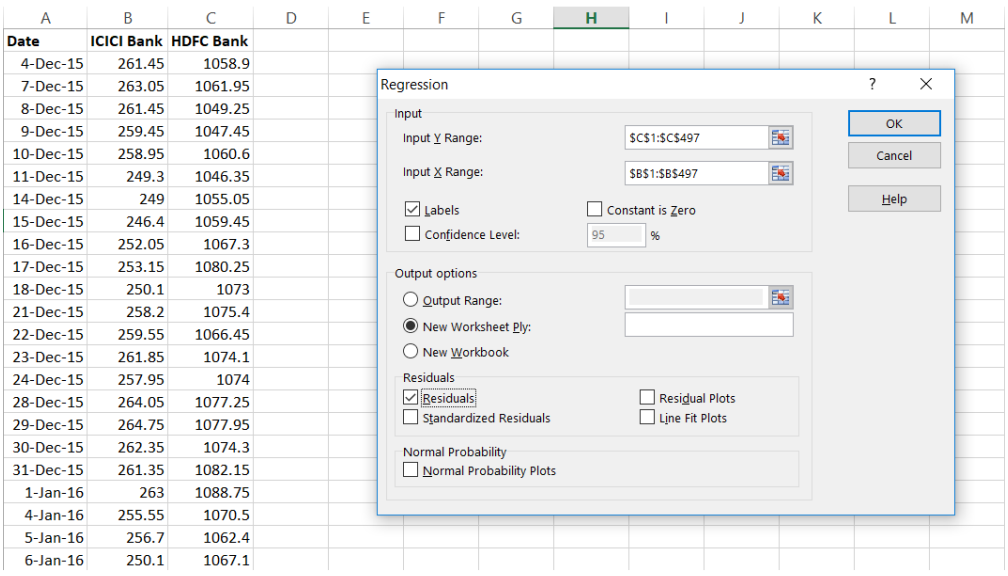
1. Make sure your data is clean – adjusted for splits, bonuses, and any other corporate actions
2. Make sure the data matches the exact dates – for instance, the data I have for both the stocks here runs from 4th of Dec 2015 to 4th Dec 2017.

Here is how the data looks –



Date	ICICI Bank	HDFC Bank
4-Dec-15	261.45	1058.9
7-Dec-15	263.05	1061.95
8-Dec-15	261.45	1049.25
9-Dec-15	259.45	1047.45
10-Dec-15	258.95	1060.6
11-Dec-15	249.3	1046.35
14-Dec-15	249	1055.05
15-Dec-15	246.4	1059.45
16-Dec-15	252.05	1067.3
17-Dec-15	253.15	1080.25
18-Dec-15	250.1	1073
21-Dec-15	258.2	1075.4
22-Dec-15	259.55	1066.45
23-Dec-15	261.85	1074.1
24-Dec-15	257.95	1074
28-Dec-15	264.05	1077.25
29-Dec-15	264.75	1077.95
30-Dec-15	262.35	1074.3
31-Dec-15	261.35	1082.15
1-Jan-16	263	1088.75
4-Jan-16	255.55	1070.5

I'll run the linear regression on these two stocks (I've explained how to do this in the previous chapter), also do note, I'm running this on the stock prices and not really on stock returns –



The result of the linear regression is as follows –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25835126.11	25835126.11	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.85969		
Total	495	37371932.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-663.6770525	61.344116	-10.8189195	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.26028449	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

RESIDUAL OUTPUT

Observation	*Predicted HDFC Bank*	*Residuals*
1	1326.90884	-268.0088403
2	1339.090663	-277.1406625
3	1326.90884	-277.6588403
4	1311.681562	-264.2315625
5	1307.874743	-247.274743
6	1234.403128	-188.0531275
7	1232.119036	-177.0690359
8	1212.323575	-152.8735747

Since ICICI is independent and HDFC is dependent, the equation is –

HDFC = Price of ICICI * 7.613 – 663.677

I'm assuming, you are familiar with the above equation. For those who are not familiar, I'd suggest you to read the previous two chapters. However here is the quick summary – the equation is trying to predict the price of HDFC using the price of ICICI.

Or in other words, we are trying to 'express' the price of HDFC in terms of ICICI.

Now, let us reverse this – I will set ICICI as dependent and HDFC as the independent.

Here is how the results look –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	16.68858714
Observations	496

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	308099.5479	308099.5	1106.247	3.5565E-128
Residual	494	137583.4168	278.5089		
Total	495	445682.9647			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	142.4677666	3.797809697	37.51314	1.1E-146	135.0059147	149.9296186	135.0059147	149.9296186
HDFC Bank	0.090797262	0.0027299	33.26028	3.6E-128	0.085433614	0.096160909	0.085433614	0.096160909

RESIDUAL OUTPUT

Observation	*Predicted ICICI Bank*	*Residuals*
1	238.612987	22.83701303
2	238.8899186	24.16008138
3	237.7367934	23.71320661
4	237.5733583	21.87664168
5	238.7673423	20.18265769

The equation is –

ICICI = HDFC * 0.09 + 142.4677

So for the given two stocks, you can regress two ways by reordering which stock is dependent and which one is the independent variable.

However, the question is – how do you decide which one should be marked dependent and which one as independent. Or in other words, which order makes the most sense.

The answer to this depends on three things –

1. Standard Error
2. Standard Error of intercept
3. The ratio of the above two variables.

Remember, the linear equation above, essentially expresses the variation of price of ICICI in terms of HDFC (refer to the equation above). This expression or explanation of the price

variation of one stock by keeping the price of the other stock as a reference can never be 100%. If it was 100%, then there is no play here at all.

Having said so, the equation should be strong enough to explain the variation in price of the dependent variable as much as possible, keeping the independent variable in perspective. The stronger this is, the better it is.

This leads us to the next obvious question – how do we figure out how strong the linear regression equation is? This is where the ratio –

Standard Error of Intercept / Standard Error comes into play. To understand this ratio, we need to understand both the numerator and the denominator before talking about the ratio itself.

10.2 – Back to residuals

Here is the linear regression equation of ICICI as independent and HDFC as the dependent –

HDFC = Price of ICICI * 7.613 – 663.677

This essentially means, if I know the price of ICICI, I should be able to predict the price of HDFC. However, in reality, there is a difference between the predicted price of HDFC and the actual price. This difference is called the 'Residuals'.

Here is the snapshot of the residuals when we try and explain the price of HDFC keeping ICICI as the independent variable –

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-663.6770525	61.344116	-10.8189195	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.26028449	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

RESIDUAL OUTPUT

Observation	Predicted HDFC Bank	Residuals
1	1326.90884	-268.0088403
2	1339.090663	-277.1406625
3	1326.90884	-277.6588403
4	1311.681562	-264.2315625
5	1307.874743	-247.274743
6	1234.403128	-188.0531275
7	1232.119036	-177.0690359
8	1212.323575	-152.8735747
9	1255.340635	-188.0406345
10	1263.715637	-183.4656373
11	1240.494039	-167.4940387
12	1302.164514	-226.7645138
13	1312.442926	-245.9929264
14	1329.954296	-255.8542959
15	1300.261104	-226.2611041
16	1346.704301	-269.4543015
17	1352.033849	-274.0838487
18	1333.761115	-259.4611153
19	1326.147476	-243.9974764
20	1338.709981	-249.9599806
21	1281.988371	-211.4883707

When I talk about the regression equation and the residuals, usually, I get one common question – what is the use of regression if there is a residual each and every time? Or in other words, how can we rely on an equation, which fails to predict accurately, even once.

This is a fair question. If you look at the residuals above, they vary from a low of -288 to a high of 548, so using this equation to make any sort of prediction one price is futile.

But then, this was never about predicting the price of the dependent stock, given the price of an independent stock. It was always about the residuals!

Let me give you a heads-up here – the residuals display a certain behaviour. If we can understand this behaviour and figure a pattern within it, then we can rework backwards to construct a trade. This trade obviously involves buying and selling the two stocks simultaneously, hence this qualifies as a pair trade.

Over the next few chapter, we will dwell deeper into this. However, for now, let's talk about the 'Standard Error', the denominator in the **Standard Error of Intercept / Standard Error** equation.

The standard error is one of the variables which gets reported when you run a linear regression operation. Here is the snapshot showing the same –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

The standard error is defined as the standard deviation of the residuals. Remember, the residuals itself is a time series array. So if you were to calculate the standard deviation of the residuals, then you get the standard error.

In fact, let me manually calculate the standard error of the residuals, I'm doing this for X = ICICI and y = HDFC

RESIDUAL OUTPUT

Observation	*Predicted HDFC Bank*	*Residuals*
1	1326.90884	-268.00884
2	1339.090663	-277.140663
3	1326.90884	-277.65884
4	1311.681562	-264.231562
5	1307.874743	-247.274743
6	1234.403128	-188.053128
7	1232.119036	-177.069036
8	1212.323575	-152.873575
9	1255.340635	-188.040635
10	1263.715637	-183.465637
11	1240.494039	-167.494039
12	1302.164514	-226.764514
13	1312.442926	-245.992926
14	1329.954296	-255.854296
15	1300.261104	-226.261104
16	1346.704301	-269.454301
17	1352.033849	-274.083849
18	1333.761115	-259.461115
19	1326.147476	-243.997476
20	1338.709981	-249.959981
21	1281.988371	-211.488371
22	1290.744055	-228.344055

=STDEV.S(D25:D520)

STDEV.S(**number1**, [number2], ...)

And excel tells me the standard deviation is **152.665**. The standard error as reported in the summary output is **152.819**. The minor difference can be ignored.

The 'Standard Error of the Intercept', is a little tricky. It does get reported in the regression report, and here is the standard error of the intercept with x = ICICI and y = HDFC

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25835126.11	25835126.1	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.8597		
Total	495	37371932.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-663.6770525	61.344116	-10.818919	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.2602845	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

Recall, the regression equation –

y=M*x+ C

Where,

M = Slope

C = Intercept

If you realize, here both M and C are estimates. And how are they estimated? They are estimated based on the historical data provided to the regression algorithm. The data can obviously contain noise components and few outliers. This implies that there is a scope for the estimates can go wrong.

The Standard Error of the Intercept is the measure of the variance of estimated intercept. It helps up understand by what degree the intercept itself can vary. So in a sense, this is somewhat similar to the 'Standard Error' itself. To summarize –

- Standard Error of Intercept – The variance of the intercept
- Standard Error – The variance of the residuals.

Now that we have defined both these variables, let's bring back the 'Error Ratio'. Please note, the term 'Error Ratio' is not a standard term, I've come up with it for ease of understanding.

Anyway, the error ratio, as we know –

Error Ratio = Standard Error of Intercept / Standard Error

I'm calculated the same for –

1. ICICI as X and HDFC as y = 0.401
2. HDFC as X and ICICI as y = 0.227

The decision to designate X and Y to stocks depends on the value of the error ratio. The lower the better. Since HDFC as X and ICICI as y offers the lowest error ratio, we will designate HDFC as the independent variable (X) and ICICI as the dependent variable (Y).

I'd love to explain the reason as to why we are using the error ratio as the key input for designating X and Y, but I guess I will hold back. I'll revisit this again when I take up pair trade example. For now, remember to calculate the error ratio and estimate which stock should be dependent and which one will be the independent.

You can download the excel sheet used in this chapter **here**.

Key takeaways from this chapter

1. X is the independent stock and Y is the dependent stock
2. The decision to figure out which stock is X and which one should be Y depends on 'Error Ratio'
3. Both the slope and the intercept from the linear regression equation are estimates
4. Error Ratio = Standard Error of the Intercept / Standard Error
5. Standard error is the standard deviation of the residuals
6. Standard error of intercept gives you a sense of the variance of the intercept
7. Regress Stock 1 with Stock 2 and also Stock 2 with Stock 1, whichever offers the lowest error ratio defines which stock is dependent and which one is independent
8. Residuals display certain properties, studying which can help identify pair trading pattern

CHAPTER 11

PTM2, C4 – The ADF test

11.1 – Co-Integration of two-time series

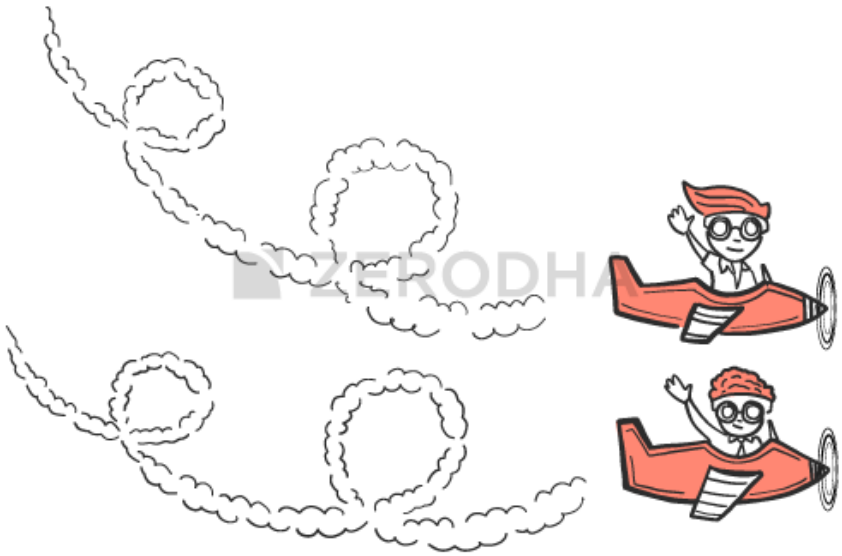
I guess this chapter will get a little complex. We would be skimming the surface of some higher order statistical theory. I will try my best and stick to practical stuff and avoid all the fluff. I'll try and explain these things from a trading point of view, but I'm afraid, some amount of theory will be necessary for you to know.

Given the path ahead I think it is necessary to re-rack our learnings so far and put some order to it. Hence let me just summarize our journey so far –

1. Starting from Chapter 1 to 7, we discussed a very basic version of a pair trade. We discussed this simply to lay out a strong foundation for the higher order pair trading technique, which is generally known as the relative value trade
2. The relative value trade requires the use of linear regression
3. In linear regression, we regress an independent variable, X against a dependent variable Y.
4. When we regress – some of the outputs that are of interest are the intercept, slope, residuals, standard error, and the standard error of the intercept
5. The decision to classify a stock as dependent and independent really depends on the error ratio.
6. We calculate the error ratio by interchanging both X and Y. The one which offers the lowest error ratio will define which stock is X and which on as Y.

I hope you have read and understood everything that we have discussed up to this point. If not, I'd suggest you read the chapters again, get clarity, and then proceed.

Recollect, in the previous chapter, we discussed the residuals. In fact, I also mentioned that the bulk of the focus going forward will be on the residuals. It is time we study the residuals in more detail and try and establish the kind of behaviour the residuals exhibit. In our attempt to do this, we will be introduced to two new jargons – Cointegration and Stationarity.



Generally speaking, if two-time series are 'co integrated' (stock X and stock Y in our case), then it means, that the two stocks move together and if at all there is a deviation from this movement, it is either temporary or can be attributed to a stray event, and one can expect the two-time series to revert to its regular orbit i.e. converge and move together again. Which is exactly what we want while pair trading. This means to say, the pair that we choose to pair trade on, should be cointegrated.

So the question is – how do we evaluate if the two stocks are cointegrated?

Well, to check if the two stock is cointegrated, we first need to run a linear regression on the two stocks, then take up the residuals obtained from the linear regression algorithm, and check if the residual is 'stationary'.

If the residuals are stationary, then it implies that the two stocks are cointegrated, if the two stocks are cointegrated, then the two stocks move together, and therefore the 'pair' is ripe for tracking pair trading opportunity.

Here is an interesting way to look at this – one can take any two-time series and apply regression, the regression algorithm will always throw out an output. How would one know if the output is reliable? This is where stationarity comes into play. The regression equation is valid if and only if residuals are stationary. If the residuals are not stationary, regression relation shouldn't be used.

Speculating and setting up trades on a co-integrated time series is a lot more meaningful and is independent of market direction.

So, essentially, this boils down to figuring out if the residuals are stationary or not.

At this point, I can straight away show you how to check if the residuals are stationary or not, there is a simple test called the 'ADF test' to do this – frankly, this is all you need to know. However, I think you are better off if you spend few minutes to understand what 'Stationarity' really means (without actually deep diving into the quants).

So, read the following section only if you are curious to know more, else go to the section which talks about ADF test.

11.2 Stationary and non-stationary series

A time series is considered 'Stationary' if it follows three 3 simple statistical conditions. If the time series partially satisfies these conditions, like 2 out of 3 or 1 out of 3, then the stationarity is considered weak. If none of the three conditions are satisfied, then the time series is 'non-stationary'.

The three simple statistical conditions are –

- The **mean** of the series should be same or within a tight range
- The **standard deviation** of the series should be within a range

- There should be no **autocorrelation** within the series – this means any particular value in the time series – say value 'n', should not be dependent on any other value before 'n'. Will talk more about this at a later stage.

While pair trading, we only look for pairs which exhibit complete stationarity. Non-stationary series or weak stationary series will not work for us.

I guess it is best to take up an example (like a sample time series) and figure out what the above three conditions really mean and hopefully, that will help you understand 'stationarity' better.

For the sake of this example, I have two-time series data, with 9000 data points in each. I've named them Series A and Series B, and on this time series data, I will evaluate the above three stationarity conditions.

Condition 1 – The mean of the series should be same or within a tight range

To evaluate this, I will split each of the time series data into 3 parts and calculate the respective mean for each part. The mean for all three different parts should be around the same value. If this is true, then I can conclude that the mean will more or less be the same even when new data points flow in the future.

So let us go ahead and do this. To begin with, I'm splitting the Series A data into three parts and calculating its respective means, here is how it looks –

	A	B	C	D	E	F	G	H
1	**Series A**	**Series B**						
2	14	15						
3	17	14.64993						
4	1	14.66357						
5	17	15.01536						
6	13	15.15149						
7	7	15.27675						
8	31	15.37252						
9	29	15.2258						

Series A

	Starting Cell	Ending Cell	Mean
Part 1	A2	A3001	20
Part 2	A3001	A6001	21.5
Part 3	A6001	A9001	20

Like I mentioned, I have 9000 data points in Series A and Series B. I have split Series A data points into 3 parts and as you can see, I've even highlighted the starting and ending cells for these parts.

The mean for all the three parts are similar, clearly satisfying the first condition.

I've done the same thing for Series B, here is how the mean looks –

	A	B	C	D	E	F	G	H
1	**Series A**	**Series B**						
2	14	15						
3	17	14.64993						
4	1	14.66357						
5	17	15.01536						
6	13	15.15149						
7	7	15.27675						
8	31	15.37252						
9	29	15.2258						
10	13	15.40872						
11	2	15.45373						
12	10	15.37771						
13	1	15.49113						
14	21	15.71245						
15	2	15.59319						
16	17	15.97966						
17	4	16.09771						

Series A

	Starting Cell	Ending Cell	Mean
Part 1	A2	A3001	20
Part 2	A3001	A6001	21.5
Part 3	A6001	A9001	20

Series B

	Starting Cell	Ending Cell	Mean
Part 1	B2	B3001	15.99036
Part 2	B3001	B6001	31.09682
Part 3	B6001	B9001	96.13986

Now as you can see, the mean for Series B swings quite wildly and thereby not satisfying the first condition for stationarity.

Condition 2 -The standard deviation should be within a range.

I'm following the same approach here – I will go ahead and calculate the standard deviation for all the three parts for both the series and observe the values.

Here is the result obtained for Series A –

	A	B	C	D	E	F	G	H	I
1	**Series A**	**Series B**							
2	14	15							
3	17	14.64993							
4	1	14.66357							
5	17	15.01536							
6	13	15.15149							
7	7	15.27675							
8	31	15.37252							
9	29	15.2258							
10	13	15.40872							

Series A

	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	A2	A3001	20	14.8492424
Part 2	A3001	A6001	21.5	19.09188309
Part 3	A6001	A9001	20	16.97056275

The standard deviation oscillates between 14-19%, which is quite 'tight' and therefore qualifies the 2nd stationarity condition.

Here is how the standard deviation works out for Series B –

	A	B	C	D	E	F	G	H	I
1	**Series A**	**Series B**							
2	14	15							
3	17	14.64993							
4	1	14.66357							
5	17	15.01536							
6	13	15.15149							
7	7	15.27675							
8	31	15.37252							
9	29	15.2258							
10	13	15.40872							
11	2	15.45373							
12	10	15.37771							
13	1	15.49113							
14	21	15.71245							
15	2	15.59319							
16	17	15.97966							

Series A

	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	A2	A3001	20	14.8492424
Part 2	A3001	A6001	21.5	19.09188309
Part 3	A6001	A9001	20	16.97056275

Series B

	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	B2	B3001	15.99036	1.400587094
Part 2	B3001	B6001	31.09682	19.96317156
Part 3	B6001	B9001	96.13986	72.02157925

Notice the difference? The range of standard deviation for Series B is quite random. Series B is clearly not a stationary series. However, Series A looks stationary at this point. However, we still need to evaluate the last condition i.e. the autocorrelation bit, let us go ahead and do that.

Condition 3 – There should be no autocorrelation within the series

In layman words, autocorrelation is a phenomenon where any value in the time series is not really dependent on any other value before it.

For example, have a look at the snapshot below –

	A	B
1	**Series A**	**Series B**
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258
10	13	15.40872
11	2	15.45373
12	10	15.37771
13	1	15.49113
14	21	15.71245
15	2	15.59319

The 9th value in Series A is 29, and if there is no autocorrelation in this series, the value 29 is not really dependent on any values before it i.e. the values from cell 2 to cell 8.

But the question is how do we establish this?

Well, there is a technique for this.

Assume there are 10 data points, I take the data from Cell 1 to Cell 9, call this series X, now take the data from Cell 2 to Cell 10, call this Series Y. Now, calculate the correlation between Series X and Y. This is called 1-lag correlation. The correlation should be near to 0.

I can do this for 2 lag as well – i.e. between Cell 1 to Cell 8, and then between Cell 3 to Cell 10, again, the correlation should be close to 0. If this is true, then it is safe to assume assumed that the series is not auto correlated, and hence the 3rd condition for stationarity is proved.

I've calculated 2 lag correlation for Series A, and here is how it looks –

Series A

	Sub - Series	Starting Cell	Ending Cell	Correlation
2 lag	X	A2	A3000	0.00457517471
	Y	A3	A3001	

Remember, I'm subdividing Series A into two parts and creating two subseries i.e. series X and series Y. The correlation is calculated on these two subseries. Clearly, the correlation is close to zero and with this, we can safely conclude that Time Series A is stationary.

Let's do this for Series B as well.

Series B

	Sub - Series	Starting Cell	Ending Cell	Correlation
2 lag	X	B2	B3000	0.99633711430
	Y	B3	B3001	

I've taken a similar approach, and the correlation as you can see is quite close to 1.

So, as you can see all the conditions for stationarity is met for Series A – which means the time series is stationary. While Series B is not.

I know that I've taken a rather unconventional approach to explaining stationarity and co-integration. After all, no statistical explanation is complete without those scary looking formulas. But this is a deliberate approach and I thought this would be the best possible way to discuss these topics, as eventually, our goal is to learn how to pair trade efficiently and not really deep dive into statistics.

Anyway, you could be thinking if it is really required for you to do all of the above to figure out if the time series (residuals) are indeed stationary. Well, like I said before, this is not required.

We only need to look at the results of something called as the 'The ADF Test', to establish if the time series is stationary or not.

11.3 – The ADF test

The augmented Dickey-Fuller or the ADF test is perhaps one of the best techniques to test for the stationarity of a time series. Remember, in our case, the time series in consideration is the residuals series.

Basically, the ADF test does everything that we discussed above, including a multiple lag process to check the autocorrelation within the series. Here is something you need to know – the output of the ADF test is not a definitive 'Yes – this is a stationary series' or 'No – this is not a stationary series'. Rather, the output of the ADF test is a probability. It tells us the probability of the series, not being stationary.

For example, if the output of the ADF test a time series is 0.25, then this means the series has a 25% chance of not being stationary or in other words, there is a 75% chance of the series being stationary. This probability number is also called 'The P value'.

To consider a time series stationary, the P value should be as low as 0.05 (5%) or lower. This essentially means the probability of the time series is stationary is as high as 95% (or higher).

Alright, so how do you run an ADF test?

Frankly, this is a highly complex process and unfortunately, I could not find a single source online which will help you run an ADF test for free. I do have an excel sheet (which has a paid plugin) to run an ADF test, but unfortunately, I cannot share it here. If I could, I would have.

If you are a programmer, I've been told that there are Python plugins easily available to run an ADF test, so you could try that.

But if you are a non-programmer like me, then you will be stuck at this stage. So here is what I will do, once in a weak or 15 days, I will try and upload a 'Pair Data' sheet, which will contain the following information of the best possible combination of pairs, this includes –

1. You will know which stock is X and which stock is Y
2. You will know the intercept and Beta of this combination
3. You will also know the p-value of the combination

The look back period for generating this is 200 trading days. I've restricted this just to banking stocks, but hopefully, I can include more sectors going forward. To help you understand this better, here is the snapshot of the latest Pair Datasheet for banking stocks –

A	B	C	D	E
Stock Y	**Stock X**	**Intercept**	**Beta**	**ADF test_P value**
FEDERALBNK	PNB	82.74692	0.170079	0.365065673
YESBANK	PNB	326.6752	0.015366	0.308751793
AXISBANK	PNB	462.077	0.436762	0.076296532
ICICIBANK	PNB	248.2804	0.364492	0.469388906
SBIN	PNB	166.4504	0.811767	0.401006906
KOTAKBANK	PNB	1099.036	-0.49692	0.01
HDFCBANK	PNB	1823.544	0.002147	0.03753307
RBLBANK	PNB	447.9693	0.417003	0.136015245
BANKBARODA	PNB	97.18598	0.388356	0.496940498
YESBANK	FEDERALBNK	248.753	0.741416	0.380701091
AXISBANK	FEDERALBNK	624.4825	-0.89685	0.364809438

The first line suggests that Federal Bank as Y and PNB as X is a viable pair. This also means, that the regression of Federal as Y and PNB as X and Federal as X and PNB as Y was conducted and the error ratio for both the combination was calculated, and it was found that Federal as Y and PNB as X had the least error ratio.

Once the order has been figured out (as in which one is Y and which one is X), the intercept and Beta for the combination has also been calculated. Finally, the ADF was conducted and the P value was calculated. If you see, the P value for Federal Bank as Y and PNB as X is 0.365.

In other words, this is not a combination you should be dealing with as the probability of the residuals being stationary is only 63.5%.

In fact, if you look at the snapshot above, you will find only 2 pairs which have the desired p-value i.e. Kotak and PNB with a P value of 0.01 and HDFC and PNB with a P value of 0.037.

The p values don't usually change overnight. Hence, for this reason, I check for p-value once in 15 or 20 days and try and update them here.

I think we have learned quite a bit in this chapter. A lot of information discussed here could be new for most of the readers. For this reason, I will summarize all the things you should know about Pair trading at this point –

1. The basic premise of pair trading
2. Basic overview of linear regression and how to perform one
3. In linear regression, we regress an independent variable, X against a dependent variable Y.
4. When we regress – some of the outputs that are of interest are the intercept, slope, residuals, standard error, and the standard error of the intercept
5. The decision to classify a stock as dependent and independent really depends on the error ratio.
6. We calculate the error ratio by interchanging both X and Y. The one which offers the lowest error ratio will define which stock is X and which on as Y
7. The residuals obtained from the regression should be stationary. If they are stationary, then we can conclude that the two stocks are co-integrated
8. If the stocks are cointegrated, then they move together
9. Stationarity of a series can be evaluated by running an ADF test.

If you are not clear on any of the points above, then I'd suggest you give this another shot and start reading from Chapter 7.

In the next chapter, we will try and take up an example of a pair trade and understand its dynamics.

You can **download the Pair Data** sheet, updated on 11th April 2018.

Lastly, this module (and this chapter, in particular) could not have been possible without the inputs from my good friend and an old partner, **Prakash Lekkala**. So I guess, we all need to thank him

---

Key takeaways from this chapter –

1. If two stocks move together, then they are also cointegrated
2. You can pair trade on stocks which are cointegrated
3. If the residuals obtained from linear regression is stationary, then it implies the two stocks are co-integrated
4. A time series is considered stationary if the series has a constant mean, constant standard deviation, and no autocorrelation
5. The check for stationarity can be done by an ADF test
6. The p-value of the ADF test should be 0.05% or lower for the series to be considered stationary.

CHAPTER 12

Trade Identification

12.1 – Trading the equation

At this stage, we have discussed pretty much all the background information we need to know about Pair trading. We now have to patch things together and understand how all these concepts make sense while taking up a pair trade.

Let's start with the basic equation again. I understand we have gone through this equation earlier in this module, but I want you to relook at this equation from a trader's perspective. I want you to think about ways in which you can trade this equation. I want you to see opportunities here. This is where everything starts to culminate.

y = M*x + c

What is this equation essentially trying to tell you? Well, frankly, it depends on how your perspective of this equation. You can look at it from two different perspectives –

1. As a statistician
2. As a trader

Since we are dealing with two stocks here, the **statistician** would look at this as an equation where the stock price of a dependent stock 'y' is being explained with respect to an independent stock price 'x'. This process of 'price explanation' generates two other variables i.e. the slope (or beta) 'M' and the intercept 'c'.

So in an ideal world, the stock price of y should be exactly equal to the Beta times X plus the intercept.

But we know that this is not true, there is always a variation in this equation which leads to the difference between the actual stock price of Y and the predicted stock price of Y. This difference is also termed as the 'residual' or the error term.

In fact, we can extend the above equation to include the residuals and with that, the equation would look like this –

$$y = M*x + c + \varepsilon$$

Where, ε represents the error or the residual of the equation. Of course, by now we are even familiar with the stationarity of the residuals which adds more sanctity to the above equation.

Fair enough, now for the interesting bit – how would a **trader** look at this equation? Let me repost the equation again –

$$y = M*x + c + \varepsilon$$

Let us break this equation into smaller pieces –

y = M*x, this essentially means, the price of the dependent stock 'y' is equal to the independent stock price 'x', multiplied by the slope M. Well, the slope is essentially the beta and it tells us how many stocks of x would equal the price of y.

For example, here is the linear regression output of HDFC Bank (y) vs ICICI Bank (x) –

SUMMARY OUTPUT

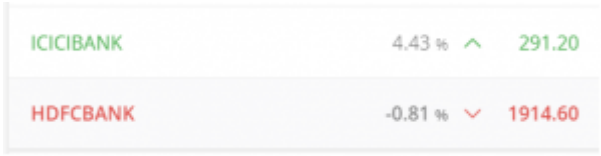
Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25835126.11	25835126.1	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.8597		
Total	495	37371932.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-663.6770525	61.344116	-10.818919	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.2602845	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

And here is the snapshot of the prices of ICICI and HDFC –



Now, this means, the price of HDFC Bank is roughly equal to the price of ICICI times the Beta. So, 1914 = 291 *7.61.

Don't jump in to do the math, I know that does not add up

But for a moment, assume if this equation were to be true, then, in other words, this essentially means 7.61 shares of ICICI equals 1 share of HDFC. This is an important conclusion.

This also means, if I were to go long on one share of HDFC and short on 7.61 shares of ICIC, then I'm essentially long and short at the same time, hence I've hedged away a large amount of directional risk. Don't forget the basic premise here, we are considering these two stocks because they are co-integrated in the first place.

So here is the equation again –

$$y = M*x + c + \varepsilon$$

If this equation were to be true, then by going long and short on y and x, we are hedging away the directional risk associated with this pair.

This leaves us with the 2nd part of the equation i.e. **c + ε**

As you know, C is the intercept. Now, at this point, I want you to recollect the 'Error Ratio' which we discussed in chapter 10.

Error Ratio = Standard Error of Intercept / Standard Error.

As you may recollect, we discussed the lower the error ratio, the better it is. Mathematically, this also implies that we are looking at pairs which have a low intercept.

Again this is a very crucial point for you to note, we are selecting the pairs, such that the standard error of the intercept is low.

Remember, in this equation **y = M*x + c + ε** we are trying to establish a trade (or hedge) every element. We are hedging y with Mx. We are trying to minimize c or the intercept because we are not trading or hedging it. Therefore, the lower it is, the better for us.

This leaves us with just the residual or the ε.

Remember, the residual is a time series. We have even validated the stationarity of this series. Now, because the residual is a stationary time series, the properties of normal distribution can be quite beautifully applied. This means, I only need to track the residuals and trigger a trade when it hits the upper or lower standard deviation!

Generally speaking, a trade is initiated when –

1. Long on the pair (buy y, sell x) when the residuals hit -2 standard deviation (-2SD)
2. Short on the pair (sell y, buy x) when the residuals hit +2 standard deviations (+2SD)

Like in the first method, the idea here is to initiate a trade at the 2nd standard deviation and hold the trade till the residual reverts to mean. The SL can be kept at 3SD for both the trades. More on this in the next chapter.

I know this is a short chapter, but I will conclude it here, as I don't want to clutter your mind with other information.

It is important for you to understand this equation from a trader's perspective and figure out what exactly you are trading. Remember, we are only trading the residuals here. We are hedging away the stock price of y with x. The intercept is kept low, and the residual is traded.

Why is the residual tradable? Because its stationary and therefore, its behaviour is kind of predictable. In the next chapter, I'll try and take up a live trade and deal with the practical aspects of pair trading.

Key takeaways from this chapter

1. The pair trading equation is actually the main equation which we trade
2. Every element of the equation is looked into
3. We hedge the stock price of y with the stock price of x. The beta of x tells us the number stocks required to hedge 1 stock of y
4. By looking into the error ratio, we are ensuring the intercept is kept low. Please remember we are not hedging the intercept, hence this needs to be kept low
5. The residual is what we trade as it is stationary and follows the normal distribution quite well
6. A long trade is initiated when residuals hit -2SD. Likewise, a short trade is initiated when the residuals hit +2SD
7. Long on a pair requires us to go long on Y and short on X
8. Short on a pair requires us to go short on Y and long on X

9. When we initiate a pair trade, we expect the residual to hit the mean, so we hold until then
10. The SL can be kept at 3SD for both long and short trades

CHAPTER 13

Live Example -1

13.1 – Tracking the pair data

We have finally reached a point where we are through with all the background theory knowledge required for Pair Trading. I know most of you have been waiting for this

moment

In this last and final chapter of pair trading, we will take up an example of a live trade and discuss factors that influence the trade.

Here is a quick recap of pre-trade theory –

1. Basic overview of linear regression and how to perform one
2. Linear regression requires you to regress an independent variable X against a dependent variable Y

3. The output of linear regression includes the intercept, slope, residuals, standard error, and the standard error of the intercept
4. The decision to classify a stock as dependent (Y) and independent (X) depends the error ratio
5. Error ratio is defined as the ratio of standard error of intercept/standard error
6. We calculate the error ratio by interchanging both X and Y. The combination which offers the lowest error ratio will define which stock is assigned X and which on as Y
7. The residuals obtained from the regression should be stationary. If they are stationary, then we can conclude that the two stocks are co-integrated
8. If the stocks are cointegrated, then they move together
9. Stationarity of a series can be evaluated by running an ADF test
10. The ADF value of an ideal pair should be less than 0.05

Over the last few chapters, we have discussed each point in great details. These points help us understand which pairs are worth considering for pair trading. In a nutshell, we take any two stocks (from the same sector), run a linear regression on it, check the error ratio and identify which stock is X and which is Y. We now run an ADF test on the residual of the pair. A pair is considered worth tracking (and trading) only if the ADF is 0.05 or lower. If the pair qualifies this, we then track the residuals on a daily basis and try to spot trading opportunities.

A pair trade opportunity arises when –

1. The residuals hit -2 standard deviations (-2SD). This is a long signal on the pair, so we buy Y and sell X
2. The residuals hit +2 standard deviations (+2SD). This is a short signal on the pair, so we sell Y and buy X

Having said so, I generally prefer to initiate the trade when the residuals hit 2.5 SD or thereabouts. Once the trade is initiated, the stop loss is -3 SD for long trades and +3SD for short trades and the target is -1 SD and +1 SD for long and short trades respectively. This also means, once you initiate a pair trade, you will have to track the residual value to

know where it lies and plan your trades. Of course, we will discuss more on this later in this chapter.

13.2 – Note for the programmers

In **Chapter 11**, I introduced the 'Pair Data' sheet. This sheet is an output of the Pair Trading Algo. The pair trading algo basically does the following –

1. Downloads the last 200-day closing prices of the underlying. You can do this from NSE's bhavcopy, in fact, automate the same by running a script.
2. The list of stock and its sector classification is already done. Hence the download is more organized
3. Runs a series of regressions and calculates the 'error ratio' for each regression. For example, if we are talking about RBL Bank and Kotak Bank, then the regression module would regress RBL (X) and Kotak (Y) and Kotak (X) and RBL (Y). The combination which has the lowest error ratio is considered and the other combination is ignored
4. The ADF test is applied on the residuals, for the combination which has the lowest error ratio.
5. A report (pair data) is generated with all the viable X-Y combination and its respective intercepts, beta, ADF value, standard error, and sigma are noted. I know we have not discussed sigma yet, I will shortly.

If you are a programmer, I would suggest you use this as a guideline to develop your own pair trading algo.

Anyway, in Chapter 11, I had briefly explained how to read the data from the Pair data, but I guess it's time to dig into the details of this output sheet. Here is the snapshot of the Pair data excel sheet –

sector	yStock	xStock	intercept	beta	adf_test_P.val	std_err	sigma
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Bajaj.Auto.Ltd.	4201.445918	-0.161879485	0.023647352	-0.713409662	136.9923607
Auto-2 wheeler	Bajaj.Auto.Ltd.	TVS.Motor.Company.Ltd.	1172.726562	2.80491901	0.0120927	-0.775683561	103.9469672
Auto-2 wheeler	Eicher.Motors.Ltd.	Bajaj.Auto.Ltd.	32451.94269	-0.846527793	0.064618555	0.364903747	1614.438459
Auto-2 wheeler	Hero.MotoCorp.Ltd.	TVS.Motor.Company.Ltd.	4193.52478	-0.725641805	0.019682961	-0.734465179	134.2067649
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Eicher.Motors.Ltd.	1812.811287	0.063432458	0.01	-1.160424948	95.53186439
Auto-2 wheeler	Eicher.Motors.Ltd.	TVS.Motor.Company.Ltd.	32198.3187	-3.477859265	0.056512336	0.373169507	1610.348145
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Ashok.Leyland.Ltd.	408.9424199	2.480217053	0.087601874	1.278654121	38.40812746
Auto-4 wheelers	Tata.Motors.Ltd.	Ashok.Leyland.Ltd.	599.0787322	-1.612890037	0.014160499	-0.246112785	25.87410403
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Ashok.Leyland.Ltd.	6086.838295	19.46666723	0.128552698	-0.897598217	567.301022
Auto-4 wheelers	Tata.Motors.DVR	Ashok.Leyland.Ltd.	357.0991825	-1.044485437	0.01	0.626806087	14.8329812
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.Ltd.	1028.745974	-0.774116165	0.277165284	1.806005143	47.61845734
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Mahindra...Mahindra.Ltd.	2861.541653	7.870346152	0.085320549	-1.871864851	479.4914183
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.DVR	989.119771	-1.183190371	0.342340179	2.103116569	48.67149928
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.Ltd.	10277.02622	-4.367072978	0.115788697	-0.183410238	628.2128651
Auto-4 wheelers	Tata.Motors.Ltd.	Tata.Motors.DVR	35.19922588	1.599579892	0.017994775	-2.549650971	7.560169082
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.DVR	10417.58475	-8.294739057	0.133376947	-0.105148619	621.0216479
Banks-PSUs	Andhra.Bank	Allahabad.Bank	3.414228493	0.760373964	0.012857999	-1.03296873	2.121488048
Banks-PSUs	Bank.Baroda	Allahabad.Bank	94.01038153	0.908995356	0.040438104	0.294249177	9.60353442
Banks-PSUs	Canara.Bank	Allahabad.Bank	71.181335	3.950811913	0.01	-0.340675142	12.67683626
Banks-PSUs	IDBI.Bank	Allahabad.Bank	88.25486183	-0.397661523	0.01	-0.512266691	5.782987299
Banks-PSUs	Allahabad.Bank	PNB	21.716644	0.302310246	0.062121944	-0.866928719	3.861679094

Look at the highlighted data. The Y stock is Bajaj Auto and X stock is TVS. Now because this combination is present in the report, it implies – Bajaj as Y and TVS as X has a lower standard error ratio, which further implies that Bajaj as X and TVS as Y is not a viable pair owing to higher error ratio, hence you will not find this combination (Bajaj as X and TVS as Y) in this report.

Along with identifying which one is X and Y, the report also gives you the following information –

1. Intercept – 1172.72
2. Beta – 2.804
3. ADF value – 0.012
4. Std_err – -0.77
5. Sigma – 103.94

I'm assuming (and hopeful) you are aware of the first three variables i.e. intercept, Beta, and ADF value so I won't get into explaining this all over again. I'd like to quickly talk about the last two variables.

Standard Error (or Std_err) as mentioned in the report is essentially a ratio of Today's residual over the standard error of the residual. Please note, this can get a little confusing here because there are two standard errors' we are talking about. The 2nd standard error is

the standard error of the residual, which is reported in the regression output. Let me explain this with an example.

Have a look at the snapshot below –

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.239703282
R Square	0.057457664
Adjusted R Square	0.053032582
Standard Error	22.77663364
Observations	215

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6736.057279	6736.057	12.9845439	0.000390994
Residual	213	110499.0835	518.775		
Total	214	117235.1408			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	267.6473274	17.4624209	15.32705	3.12838E-36	233.226034	302.0686	233.226034	302.0686208
South Indian	2.173977689	0.60331168	3.603407	0.000390994	0.984751494	3.363204	0.984751494	3.363203884

RESIDUAL OUTPUT

Observation	*Predicted Yes Bank*	*Residuals*
1	323.7359518	20.91404822
2	325.5838328	22.26616719
3	326.2360261	17.06397388
4	324.3881451	23.51185491
5	323.9533495	21.14665045

This is the regression output summary of Yes Bank versus South Indian Bank. I've highlighted standard error (22.776). This is the standard error of the residuals. Do recollect, we have discussed this earlier in this module.

The second highlight is 20.914, which is the residual.

The std_err in the report is simply a ratio of –

Today's residual / Standard Error of the residual

= 20.92404/22.776

= 0.91822

Yes, I agree calling this number std_err is not the best choice, but please bear with it for now

This number gives me information of how today's residual is position in the context of the standard distribution. This is the number which is the key trigger for the trade. A long position is hit if this number is -2.5 or higher with -3.0 as stop loss. A short position is initiated if this number reads +2.5 or higher with a stop loss at +3.0. In case of long, target is at -1 or lower and in case of short, the target is +1 or lower.

This also means, the std_err number has to be calculated on a daily basis and tracked to identify trading opportunities. More on this in a bit.

The sigma value in the pair data report is simply the standard error of the residual, which in the above case is 22.776.

So now if you read through the pair data sheet, you should be able to understand the details completely.

Alright, let us jump to the trade now

13.3 – Live example

I have been running the pair trading algo to look for opportunities, and I found one on 10th May 2018. Here is the snapshot of the pair data, you can download the same towards the end of this chapter. Do recollect, this pair trading algo was generated using the closing prices of 10th May.

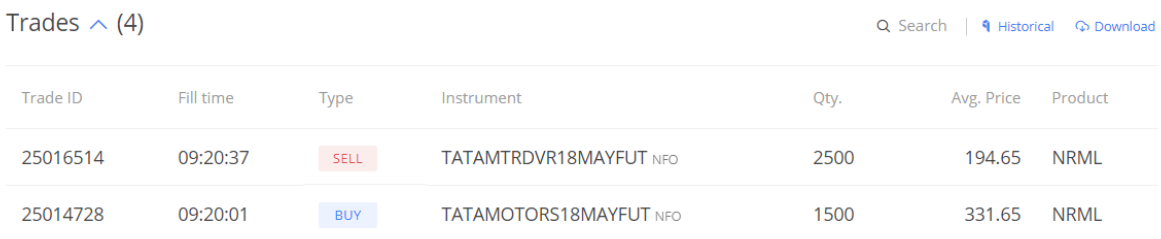
sector	yStock	xStock	intercept	beta	adf_test_P.val	std_err	sigma
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Bajaj.Auto.Ltd.	4201.445918	-0.161879485	0.023647352	-0.713409662	136.9923607
Auto-2 wheeler	Bajaj.Auto.Ltd.	TVS.Motor.Company.Ltd.	1172.726562	2.80491901	0.0120927	-0.775683561	103.9469672
Auto-2 wheeler	Eicher.Motors.Ltd.	Bajaj.Auto.Ltd.	32451.94269	-0.846527793	0.064618555	0.364903747	1614.438459
Auto-2 wheeler	Hero.MotoCorp.Ltd.	TVS.Motor.Company.Ltd.	4193.52478	-0.725641805	0.019682961	-0.734465179	134.2067649
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Eicher.Motors.Ltd.	1812.811287	0.063432458	0.01	-1.160424948	95.53186439
Auto-2 wheeler	Eicher.Motors.Ltd.	TVS.Motor.Company.Ltd.	32198.3187	-3.477859265	0.056512336	0.373169507	1610.348145
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Ashok.Leyland.Ltd.	408.9424199	2.480217053	0.087601874	1.278654121	38.40812746
Auto-4 wheelers	Tata.Motors.Ltd.	Ashok.Leyland.Ltd.	599.0787322	-1.612890037	0.014160499	-0.246112785	25.87410403
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Ashok.Leyland.Ltd.	6086.838295	19.46666723	0.128552698	-0.897598217	567.301022
Auto-4 wheelers	Tata.Motors.DVR	Ashok.Leyland.Ltd.	357.0991825	-1.044485437	0.01	0.626806087	14.8329812
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.Ltd.	1028.745974	-0.774116165	0.277165284	1.806005143	47.61845734
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Mahindra...Mahindra.Ltd.	2861.541653	7.870346152	0.085320549	-1.871864851	479.4914183
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.DVR	989.119771	-1.183190371	0.342340179	2.103116569	48.67149928
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.Ltd.	10277.02622	-4.367072978	0.115788697	-0.183410238	628.2128651
Auto-4 wheelers	Tata.Motors.Ltd.	Tata.Motors.DVR	35.19922588	1.599579892	0.017994775	-2.549650971	7.560169082
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.DVR	10417.58475	-8.294739057	0.133376947	-0.105148619	621.0216479
Banks-PSUs	Andhra.Bank	Allahabad.Bank	3.414228493	0.760373964	0.012857999	-1.03296873	2.121488048
Banks-PSUs	Bank.Baroda	Allahabad.Bank	94.01038153	0.908995356	0.040438104	0.294249177	9.60353442
Banks-PSUs	Canara.Bank	Allahabad.Bank	71.181335	3.950811913	0.01	-0.340675142	12.67683626
Banks-PSUs	IDBI.Bank	Allahabad.Bank	88.25486183	-0.397661523	0.01	-0.512266691	5.782987299
Banks-PSUs	Allahabad.Bank	PNB	21.716644	0.302310246	0.062121944	-0.866928719	3.861679094

Look at the data highlighted in red. This is Tata Motors Ltd as Y (dependent) and Tata Motors DVR as X (independent).

The ADF value reads, 0.0179 (less than the threshold of 0.05), and I think this is an excellent ADF value. Do recollect, ADF value of less than 0.05 indicates that the residual is stationary, which is exactly what we are looking for.

The std_err reads -2.54, which means the residuals is close has diverged (sufficiently enough) away from the mean and therefore one can look at setting up a long trade. Since this is a long trade, one is required to buy the dependent stock (Tata Motors) and short the independent stock (Tata Motors DVR). This trade was supposed to be taken on 11th May Morning (Friday), but for some reason, I was unable to place the trade. However, I did take the trade on 14th May (Monday) morning at a slightly bad rate, nevertheless, the intention was to showcase the trade and not really chase the P&L.

Here are the trade execution details –



You may have two questions at this point. Let me list them for you –

Question – Did I actually execute the trade without checking for prices? As in I didn't even look at what price the stocks, I didn't look at support, resistance, RSI etc. Is it not required?

Answer – No, none of that is required. The only thing that matters is where the residual is trading, which is exactly what I looked for.

Question – On what basis did I choose to trade 1 lot each? Why can't I trade 2 lots of TM and 3 lots of TMD?

Answer – Well this depends on the beta of the stock. We will use the beta and identify the number of stocks of X &Y to ensure we are **beta neutral** in this position. The beta neutrality states that for every 1 stock of Y, we need to have beta*X stock of X. For example, in the Tata Motors (Y) and Tata Motors DVR (X) for example, the beta is 1.59. This means, for every 1 stock of Tata Motors (Y), I need to have 1.59 stocks of Tata Motors DVR (X).

Going by this proportion, the lot size of Tata Motors (Y) is 1500, so we need 1500*1.59 or 2385 shares of Tata Motors DVR (X). The lot size is 2400, quite close to 2385, hence I decided to go with 1 lot each. But I'm aware this trade is slightly more skewed towards the long side since I'm buying additional 115.

Also, please note, because of this constraint, we cannot really trade pairs if the beta is –ve, at least, not always.

Remember, I initiated this trade when the residual value was -2.54. The idea was to keep the position open and wait for the target (-1 on residual) or stop loss (-3 on residual) was hit. Until then, it was just a waiting game.

To track the position live, I've developed a basic excel tracker. Of course, if you are a programmer, you can do much better with these accessories, but given my limited abilities, I put up a basic position tracker in excel. Here is the snapshot, of course, you can download this sheet from the link posted below.

Position Tracker

Pair Data

Independent Stock (X)	Tata Motors DVR
Dependent Stock (Y)	Tata Motors
Sector	Auto 4 wheeler

For Beta Nutrality

Lot size of X	2500
Lot size of Y	1500
For 1 lot of Y	2400

Regression Parameters

Beta	1.6
Intercept	35.19923
Residual	-19.35923
Sigma	7.56

Signal

Date	10th May 2018
Spot of X	198.6
Spot of Y	333.6
Z-Score	-2.560744709

Trade Executed

Date	14th May 2018
Fut (X)	194.65
Fut (y)	331.65
Z-Score	-1.982702381

Current Values

Date	
Fut (X)	
Fut (y)	
Z-Score	

P&L

Stock	Position	Lot Size	Trade Price	Current Price	P&L
Tata Motors (Y)	Long	1500	331.65		
Tata Motors DVR (X)	Short	2500	194.65		
Total					

Instructions:

1) Initiate the trade when Z-Score is above +2.5 or below -2.5
2) SL is when Z -Score hits +3 or -3
3) Target is +1 or -1

The position tracker has all the basic information about the pair. I'm guessing this is a fairly easy sheet to understand. I've designed it in such a way that upon entering the current values of X & Y, the latest Z score is calculated and also the P&L. I'd encourage you to play around this sheet, even better if you can build one yourself

Once the position is taken, all one has to do is track the z-score of the residual. This means you have to keep tracking the values and the respective z-scores. This is exactly what I did. In fact, for the sake of this chapter, my colleague, Faisal, logged all the values (except for the 14th and 15th). Here are the logs –

Logs	16th May
Time	9.45 AM
Fut (X)	181.25
Fut (y)	311.4
Z-Score	-1.82529
Time	10.45 AM
Fut (X)	181.8
Fut (y)	310.95
Z-Score	-2.00122
Time	12.20 PM
Fut (X)	183
Fut (y)	313.75
Z-Score	-1.88482
Time	1.35 PM
Fut (X)	184.35
Fut (y)	315
Z-Score	-2.00519
Time	3.30 PM
Fut (X)	183
Fut (y)	311.55
Z-Score	-2.17582

Logs	17th May
Time	9.40 AM
Fut (X)	182.45
Fut (y)	311.65
Z-Score	-2.04619
Time	11:00 AM
Fut (X)	184.5
Fut (y)	314.7
Z-Score	-2.07662
Time	12.30 PM
Fut (X)	185.2
Fut (y)	316.7
Z-Score	-1.96022
Time	1.45 PM
Fut (X)	185.9
Fut (y)	318.4
Z-Score	-1.8835
Time	3.30 PM
Fut (X)	184.95
Fut (y)	315.5
Z-Score	-2.06604

Logs	18th May
Time	11.30 AM
Fut (X)	183
Fut (y)	309.9
Z-Score	-2.39408
Time	3.00 PM
Fut (X)	179.7
Fut (y)	309.9
Z-Score	-2.39408
Time	3.30 PM
Fut (X)	306
Fut (y)	181
Z-Score	-2.48667

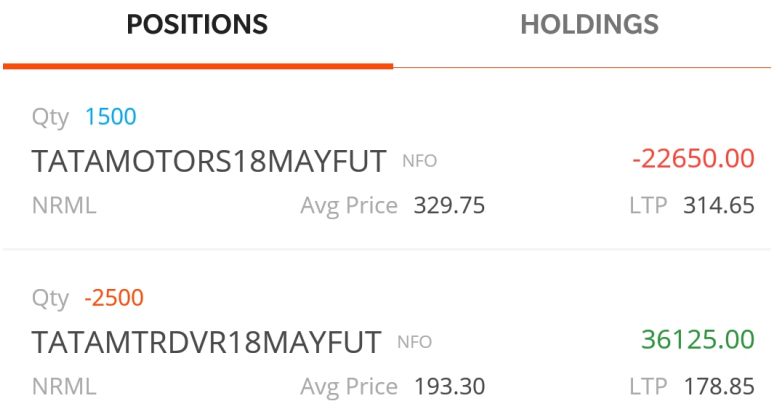
Logs	21th May
Time	9.20 AM
Fut (X)	179.25
Fut (y)	306.7
Z-Score	-2.02371
Time	11.30 AM
Fut (X)	176.25
Fut (y)	301.2
Z-Score	-2.1163
Time	2.00 PM
Fut (X)	175.75
Fut (y)	299.65
Z-Score	-2.21551
Time	3.20 PM
Fut (X)	175.35
Fut (y)	297.4
Z-Score	-2.42847

Logs	22nd May
Time	9.30 AM
Fut (X)	174.1
Fut (y)	300
Z-Score	-1.82
Time	11.00 AM
Fut (X)	174.25
Fut (y)	300
Z-Score	-1.85175
Time	12.00 AM
Fut (X)	172.45
Fut (y)	298.5
Z-Score	-1.66921
Time	1.50 PM
Fut (X)	180.75
Fut (y)	312.4
Z-Score	-1.5872
Time	3.20 PM
Fut (X)	177.9
Fut (y)	308.8
Z-Score	-1.46022

Logs	23rd May
Time	10.17 AM
Fut (X)	178.7
Fut (y)	313.5
Z-Score	-1.00783

As you can see, the current values were tracked and the latest z-score was calculated several times a day. The position was open for nearly 7 trading session and this is quite common with pair trading. I've experienced positions where they were open for nearly 22 - 25 trading sessions. But here is the thing – as long as your math is right, you just have to wait for the target or SL to trigger.

Finally, on 23rd May morning, the z-score dropped to the target level and there was a window of opportunity to close this trade. Here is the snapshot –



Notice, the gains in Tata Motors DVR is much larger than the loss in Tata Motors. In fact, when we take the trade, we will never know which of the two positions will make us the

money. The idea, however, is that one of them will move in our favour and the other won't (or may). It's however, just not possible to identify which one will be the breadwinner.

The position tracker for the final day (23rd May) looked like this –

Position Tracker

Pair Data

Independent Stock (X)	Tata Motors DVR
Dependent Stock (Y)	Tata Motors
Sector	Auto 4 wheeler

For Beta Nutrality

Lot size of X	2500
Lot size of Y	1500
For 1 lot of Y	2400

Regression Parameters

Beta	1.6
Intercept	35.19923
Residual	-19.35923
Sigma	7.56

Signal

Date	10th May 2018
Spot of X	198.6
Spot of Y	333.6
Z-Score	-2.560744709

Trade Executed

Date	14th May 2018
Fut (X)	194.65
Fut (y)	331.65
Z-Score	-1.982702381

Current Values

Date	23rd May 2018
Fut (X)	178.85
Fut (y)	314.65
Z-Score	**-0.887464286**

P&L

Stock	Position	Lot Size	Trade Price	Current Price	P&L
Tata Motors (Y)	Long	1500	331.65	314.65	-25500
Tata Motors DVR (X)	Short	2500	194.65	178.85	39500
Total					**14000**

Instructions:

1) Initiate the trade when Z-Score is above +2.5 or below -2.5
2) SL is when Z -Score hits +3 or -3
3) Target is +1 or -1

The P&L was roughly Rs.14,000/-, not bad I'd say for a relatively low-risk trade.

13.4 – Final words on Pair Trading

Alright guys, over the last 13 chapter, we have discussed everything I know about pair trading. I personally thing this is a very exciting way of trading rather than blind speculative trading. Although less risky, pair trade has its own share of risk and you need to be aware of the risk. One of the common ways to lose money is when the pair can continue to diverge after you initiate the position, leaving you with a deep loss. Further, the margin requirements are slightly higher since there are two contracts you are dealing with. This also means you need to have some buffer money in your account to accommodate daily M2M.

There could be situations where you will need to take a position in the spot market as well. For example, on 23rd May, there was a signal to go short on Allahabad Bank (Y) and long on Union Bank (X). The z-score was 2.64 and the beta for this pair is 0.437.

Going by beta neutrality, for every 1 share of Allahabad Bank (Y), I need 0.437 shares of Union Bank (X). The Lot size of Allahabad Bank is 10,000, this implies I need to buy 4378 shares of Union Bank. However, the lot size of Union Bank is 4000, hence I had to buy 370 shares in the spot market.

Trades (4) Search | Historical Download

Trade ID	Fill time	Type	Instrument	Qty.	Avg. Price	Product
[illegible]	10:45:51	BUY	[illegible]	[illegible]	[illegible]	[illegible]
25044054	09:24:41	BUY	UNIONBANK18JUNFUT NFO	4000	87.75	NRML
75120452	09:21:26	BUY	UNIONBANK NSE	370	87.4	CNC
62508501	09:20:48	SELL	ALBK18JUNFUT NFO	10000	40.75	NRML

Well, I hope I trade is successful

I know most of you would want the pair data sheet made available. We are working on making this sheet available to you on a daily basis so that you can track the pairs. Meanwhile, I would suggest you try and build this algo yourself. If you have concerns, please post it below and I will be happy to assist.

If you don't know how to program then you have no option but to find someone who knows programming and convince him or her that there is money to be made, this is

exactly what I did

Lastly, I would like to leave you with a thought –

1. We run a linear regression of Stock A with Stock B to figure out if the two stocks are cointegrated with their residuals being stationary
2. What if Stock A with Stock B is not stationary, but instead Stock A is stationary with stock B & C as a combined entity?

Beyond Pair, trading lies something called as multivariate regression. By no stretch of the imagination is this easy to understand, but let me tell you if you can graduate to this arena, the game is different.

Download the Position Tracker and Pair Datasheet below:

DOWNLOAD POSITION TRACKER

DOWNLOAD PAIR DATASHEET

Key takeaways from this chapter

1. The trigger to trade a pair comes from the residual's current value
2. Check for beta neutrality of the pair to identify the number of stock required in X and Y
3. If the beta of the pair is negative, then it may not be possible to set up the trade
4. Once the trade is initiated, check the z-score movement to trade its current position
5. The price of the futures does not really matter, the emphasis is only on the z-score

CHAPTER 14

Live Example – 2

14.1 – Position Sizing

I know, the discussion on pair trading was to end with the previous chapter, but I thought I had to discuss a special case before we finally wrap up. I'll also try and keep this chapter

really short

So here you go.

I ran through the pair trading algo y'day evening (28th May) and found a very interesting trade. Here are the regression parameters –

- Stock X = ICICI Bank
- Stock Y = HDFC Bank
- ADF = 0.048
- Beta = 0.79
- Intercept = 1626
- Std_err = 2.67

What do you think of it? Perfect isn't it? Its ICICI and HDFC, two of the largest private sector banks, both have similar business landscape, both have a similar revenue stream, both regulated by RBI. Perhaps the perfect candidate for a pair trade, right?

The ADF value is 0.048, which means there is only 4.8% chance that the residual is non-stationary or about 95.2% chance of the residuals being stationary, which is fantastic.

The std_err is +2.67, which is a perfect residual value to initiate a short pair trade. The trade here is short HDFC and go long on ICIC.

So, how do we position size this? Here are the price and lot size details –

- HDFC Fut Price = 2024.8
- HDFC Lot size = 500
- ICICI Fut price = 298.8
- ICICI Lot size = 2750

Remember we discussed position size in the previous chapter. We look at the beta and estimate the number of shares required for this trade.

The beta is 0.79, this means, every 1 share of Y needs to be offset with 0.79 shares of X. The lot size of HDFC (Y) is 500, this means to offset the beta, we need 395 shares of ICICI (X).

Do you see the problem here? The lot sizes simply do not match.

We cannot simply trade 1 lot each here like we did in the TATA Motors and Tata Motors DVR example, discussed in the previous chapter. If we do, then this won't be a beta neutral trade.

Hence to position size this, we need to work around with the lot sizes –

The lot size of ICIC is 2750, beta is 0.79, lot size of HDFC is 500. Given this, that the lot size is higher than HDFC, what should be the minimum number of HDFC shares which will beta neutral 2750 shares of ICICI.

To figure this out, we simply divide –

2750/0.79

= 3481.01

Since the lot size of HDFC is 500, we can round this off to 3500. Considering the lot size of HDFC is 500, this will be 7 lots of HDFC against 1 lot of ICICI.

14.2 – Intercept

Alright, now that we know the position size as well, here is the big question – will you take this trade?

Everything seems perfect, right? ADF has a desirable value, residual is at 2.67 SD, the two stocks are highly correlated, the business is similar. So what can go wrong?

Yes, I agree, everything looks good, but on a closer look, the intercept reveals a slightly different story.

To understand this, we need to quickly revisit the regression equation –

y = Beta * x + Intercept + Residual

If you think about this equation, we are trying to explain the stock price of Y in terms of the stock price of X multiplied by its beta. The intercept is essentially that portion of the y's stock price which the model cannot explain, and the residual is the difference between predicted y and actual y.

Going by this, a large intercept implies that a large portion of Y's stock price cannot be explained by the regression model.

In this case, the intercept is 1626. The stock price of HDFC is 2024 per share, this means, 1626 out of 2024 cannot be explained by the regression equation. This means, the regression equation cannot explain nearly 80% (1626/2024) of Y's stock price or in other words the equation can explain only 20% of the equation, which according to me is quite tricky.

This further implies, that if we are trading this pair, then we are essentially trading a very small probability here. I'd rather avoid this and look for another opportunity than trade this. Of course, I know traders who would love to jump in and take this trade, but for

someone like me, I'd look at risk first and then the reward

Good luck!

DOWNLOAD PAIR DATASHEET