MATH 124 – Transforming Data

bolstad_math124@bmbolstad.com
http://math124sfsu.bmbolstad.com

The purpose of this document is to guide you through the steps needed to transform data (for later simple linear regression analysis) using excel. For your reference the datafile used in this document in both tab delimited text (*transform.dat*) and excel 2000 (*transform.xls*) formats is on the webpage. It is expected that you are somewhat familiar with how to create a scatterplot using Excel and so some details on how to create such plots is omitted (you should see the previous Scatterplot document to refresh your memory if you need it).

Why transform?

If you create a scatterplot and it doesn't look like the relationship between the two variables is quite as linear then a linear regression might not seem appropriate. Sometimes an appropriate transformation of one or both of the variables will give you a plot which seems much more linear and then you can safely fit a linear regression line.

The Initial Data

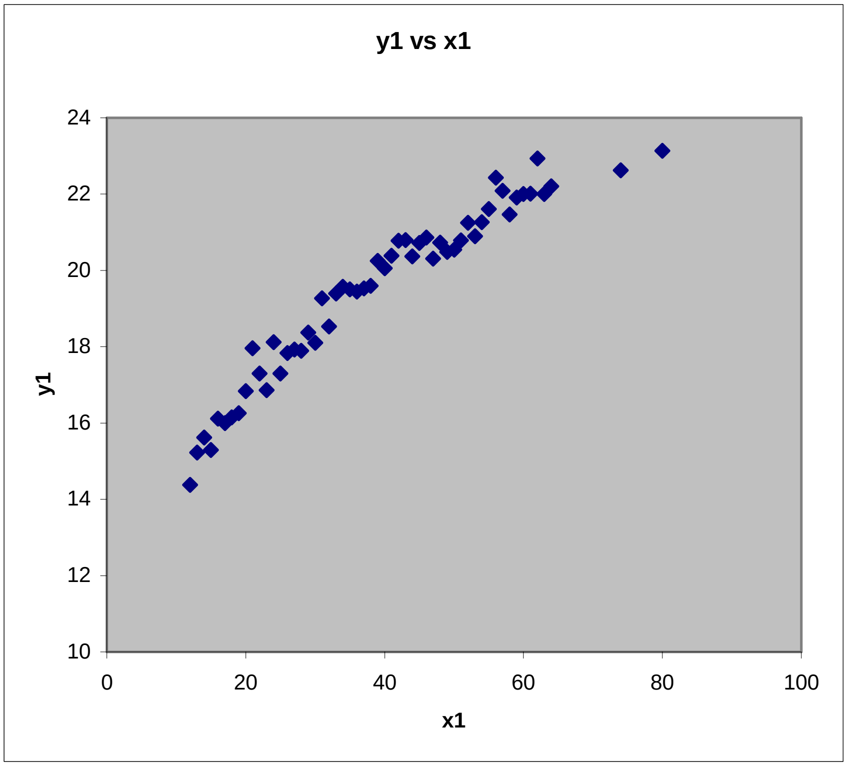
After you have read this datafile into excel you should see that there is four groups of data.

1. X1 and Y1 (stored in columns A and B) are data that requires a log transform on the x variable.
2. X2 and Y2 (stored in columns L and M) are data that requires a log transform on the y variable
3. X3 and Y3 (stored in columns W and X) are data that requires log transformations on both the x and y variables.
4. X4 and Y4 (stored in columns AK and AL) are data that requires an inversion transformation on the x variable.

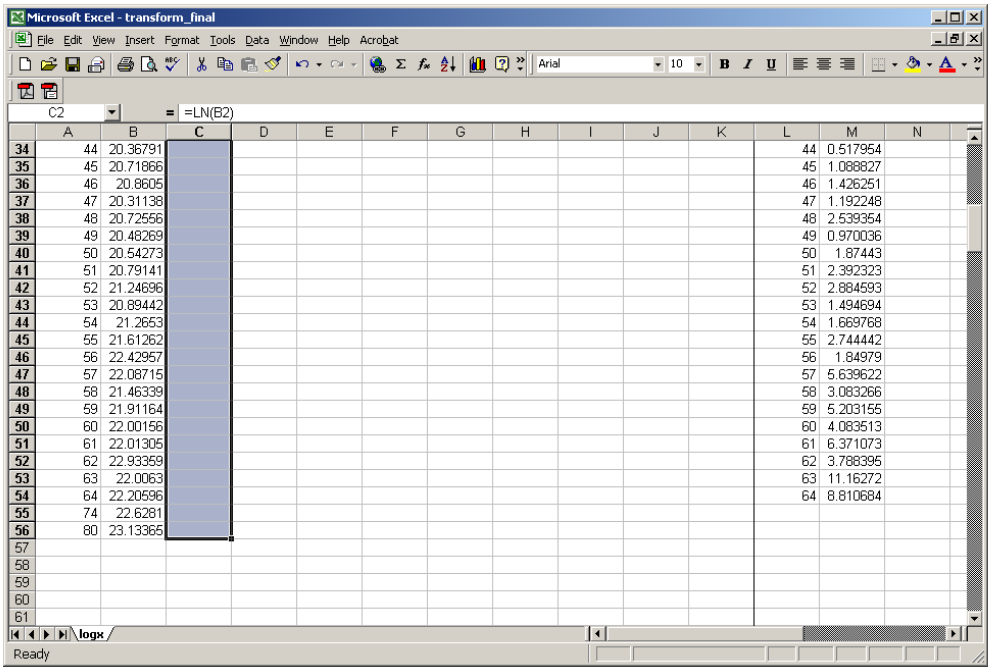
We will discuss how to first scatter plot the untransformed data and then create columns of transformed data, then again scatterplot to see that there is now a more linear trend.

Dataset 1

1. First scatterplot the data so that y1 is on the vertical axis and x1 is on the horizontal axis. Depending on how you choose to format your scatterplot you should end up with something like this:



2. Notice how there seems to be a slight curve when you look at the form that the points on the scatterplot seem to follow. We would like to make this pattern more linear. To do this we will take a log transformation of the x1 variable
3. In the cell C1 type “log x1”. Then move to cell C2 and type “=log(A2)”. Press enter and you should see the number 1.157614 appear (this is the log base 10 of the number in A2). (Note: that if you wanted to work with the natural log you would have typed “=LN(A2)” instead. For the transformation a logarithm of any base is acceptable.)
4. Next select all the cells from C2 down to C56. This should look something like this:

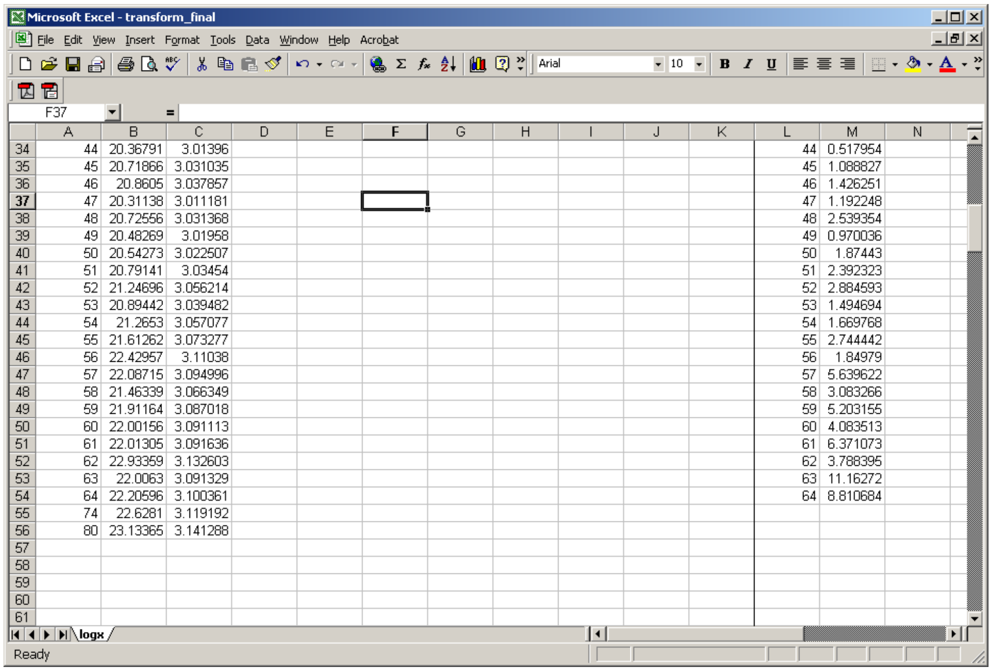


Microsoft Excel - transform_final

C2 = =LN(B2)

	A	B	C	L	M
34	44	20.36791		44	0.517954
35	45	20.71866		45	1.088827
36	46	20.8605		46	1.426251
37	47	20.31138		47	1.192248
38	48	20.72556		48	2.539354
39	49	20.48269		49	0.970036
40	50	20.54273		50	1.87443
41	51	20.79141		51	2.392323
42	52	21.24696		52	2.884593
43	53	20.89442		53	1.494694
44	54	21.2653		54	1.669768
45	55	21.61262		55	2.744442
46	56	22.42957		56	1.84979
47	57	22.08715		57	5.639622
48	58	21.46339		58	3.083266
49	59	21.91164		59	5.203155
50	60	22.00156		60	4.083513
51	61	22.01305		61	6.371073
52	62	22.93359		62	3.788395
53	63	22.0063		63	11.16272
54	64	22.20596		64	8.810684
55	74	22.6281			
56	80	23.13365			
57					
58					
59					
60					
61					

5. Now either press CTRL+D or from the edit menu select Fill > Down. In either case what should happen is that numbers will now appear in that column something like this.



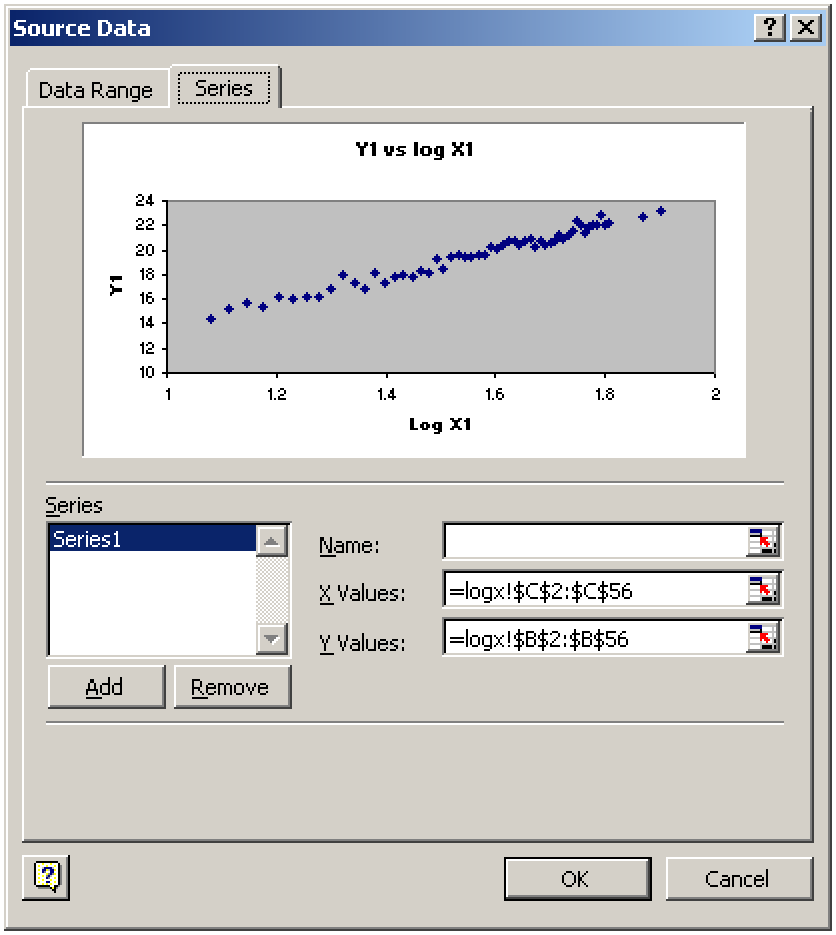
Microsoft Excel - transform_final

F37 =

	A	B	C	L	M
34	44	20.36791	3.01396	44	0.517954
35	45	20.71866	3.031035	45	1.088827
36	46	20.8605	3.037857	46	1.426251
37	47	20.31138	3.011181	47	1.192248
38	48	20.72556	3.031368	48	2.539354
39	49	20.48269	3.01958	49	0.970036
40	50	20.54273	3.022507	50	1.87443
41	51	20.79141	3.03454	51	2.392323
42	52	21.24696	3.056214	52	2.884593
43	53	20.89442	3.039482	53	1.494694
44	54	21.2653	3.057077	54	1.669768
45	55	21.61262	3.073277	55	2.744442
46	56	22.42957	3.11038	56	1.84979
47	57	22.08715	3.094996	57	5.639622
48	58	21.46339	3.066349	58	3.083266
49	59	21.91164	3.087018	59	5.203155
50	60	22.00156	3.091113	60	4.083513
51	61	22.01305	3.091636	61	6.371073
52	62	22.93359	3.132603	62	3.788395
53	63	22.0063	3.091329	63	11.16272
54	64	22.20596	3.100361	64	8.810684
55	74	22.6281	3.119192		
56	80	23.13365	3.141288		
57					
58					
59					
60					
61					

6. This new column of data consists of log base 10 values of the data in column A (ie the x1 values). Next it is time to make a scatterplot of this transformed data to make sure that the transformation got rid of the curve. Repeat the steps you have used before to make a scatterplot. Note that it important that you keep y1 on the

vertical axis and the new log transformed variable on the horizontal axis. The “Series” tab on the Chart Source Data window might be useful here:

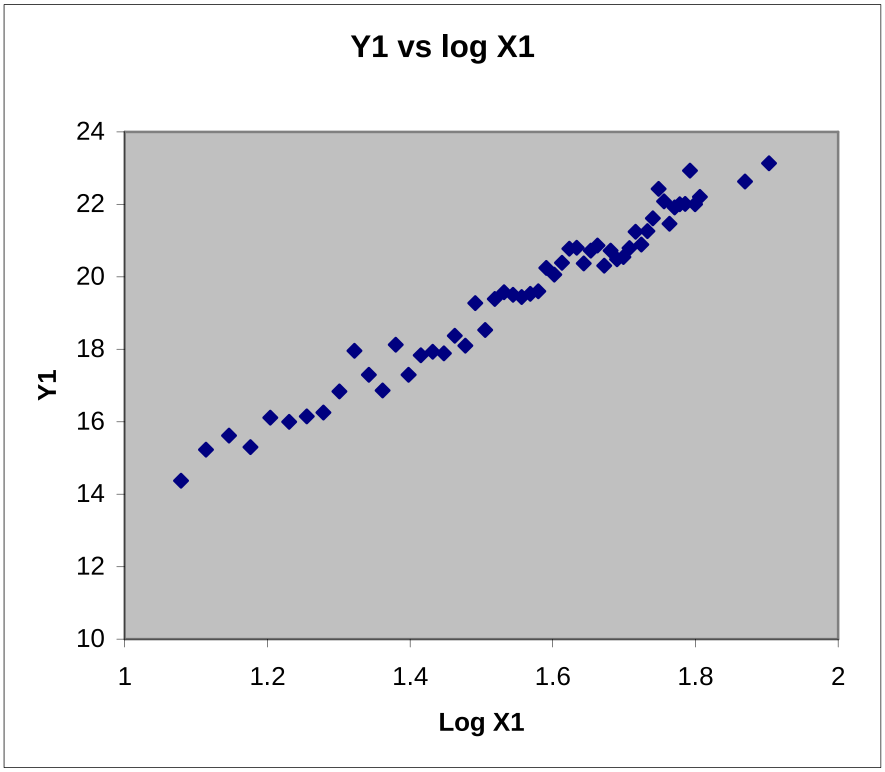


Name:

X Values: =logx!$C$2:$C$56

Y Values: =logx!$B$2:$B$56

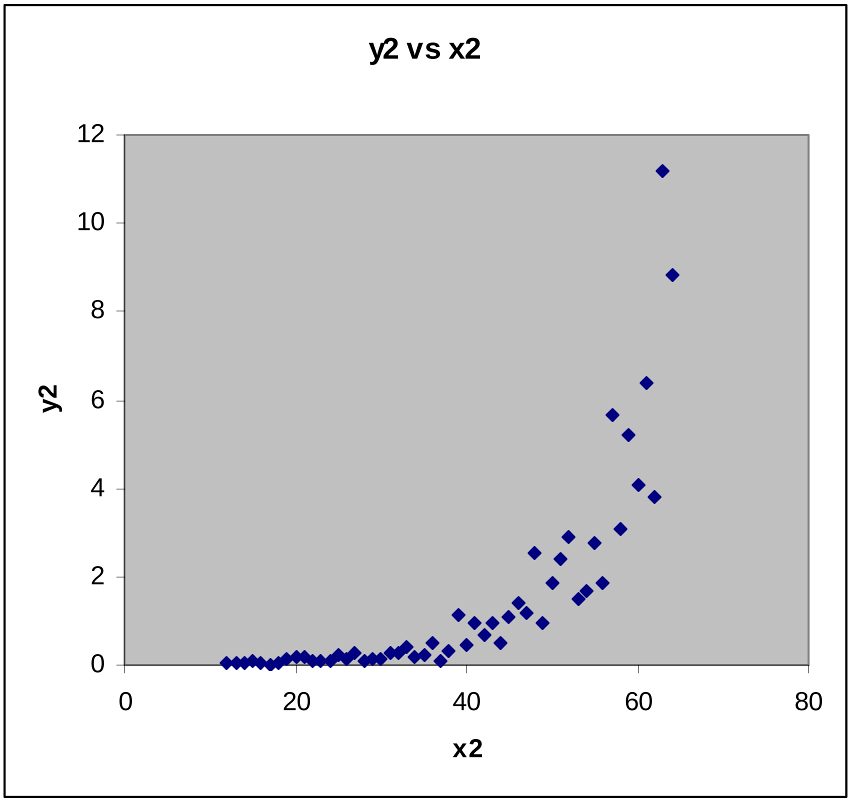
7. Proceed making your scatterplot and you should end up with something looking like this (again depending on how you formatted your plot you may differ slightly)



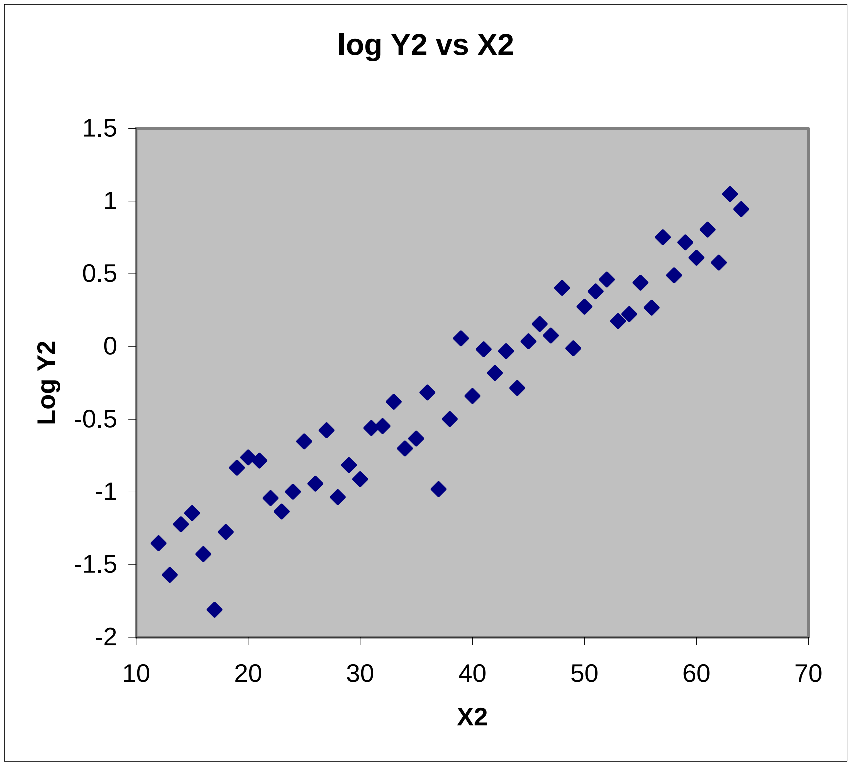
8. Notice that the relationship is much more linear than the original plot. The transformation in this example has been successful.

Dataset 2

1. First create a scatterplot where y2 is on the vertical axis and x2 is on the horizontal axis. The scatterplot should look something like this:



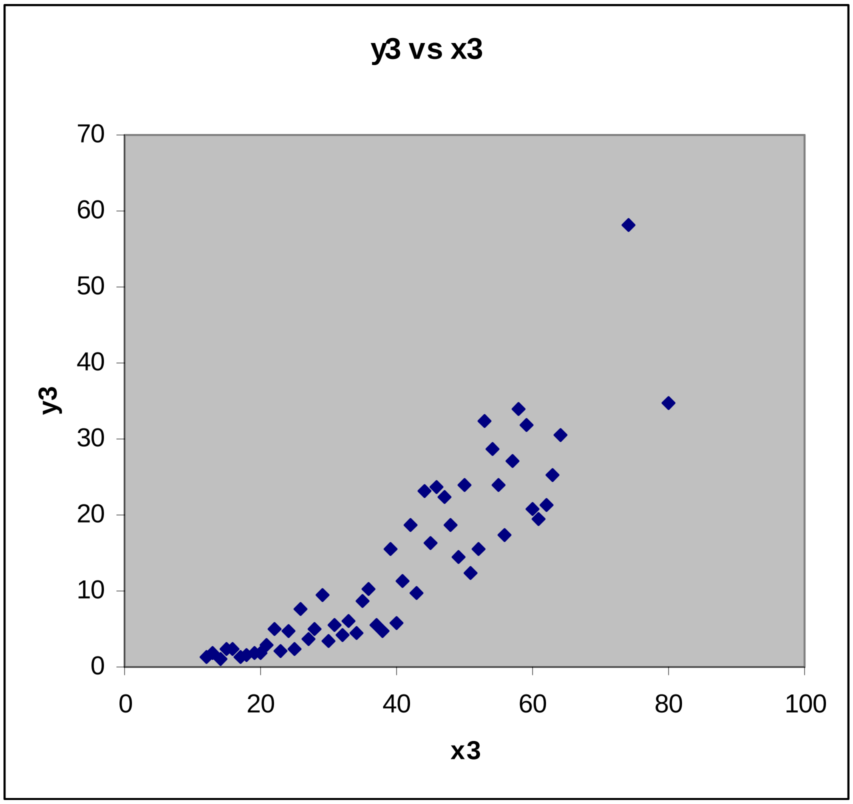
2. Notice how there is quite a clear curve on this plot
3. Now it is time to log transform the y2 values. Go to the cell N1 and type “log y2”. Then go to the cell N2 and type “=log(M2)”. Press enter.
4. As with the first dataset, select the cells N2 through N54. Then type CRTL+D or use the edit menu and the fill down command. This should give you a column of transformed data values.
5. Make a scatterplot of the transformed data, making sure to keep the log y2 values on the vertical axis and the x2 on the horizontal axis. Your scatterplot should look something like this.



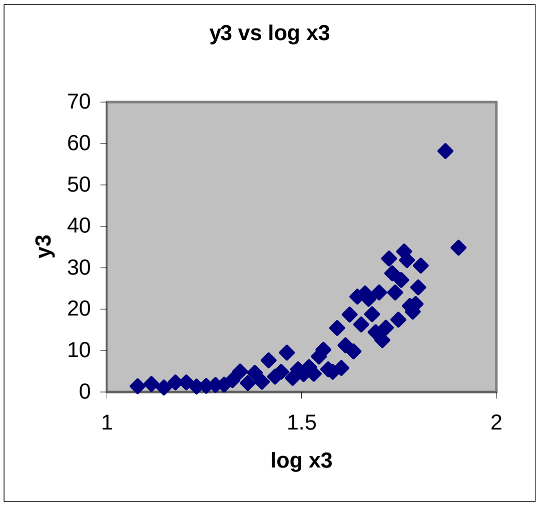
6. Notice how the plot now appears much more linear than before.

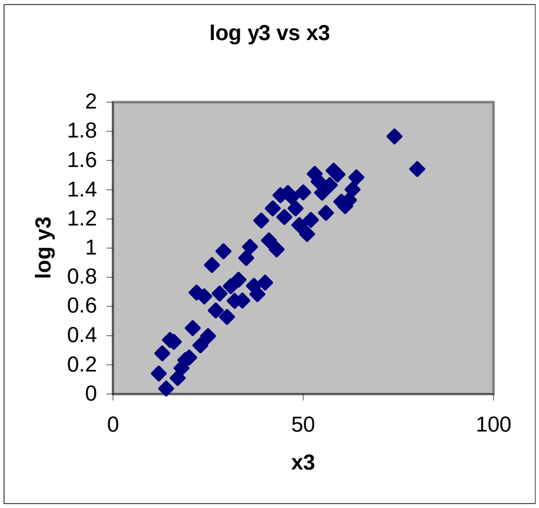
Dataset 3

1. First create a scatterplot where y3 is on the vertical axis and x3 is on the horizontal axis. The scatterplot should look something like this:

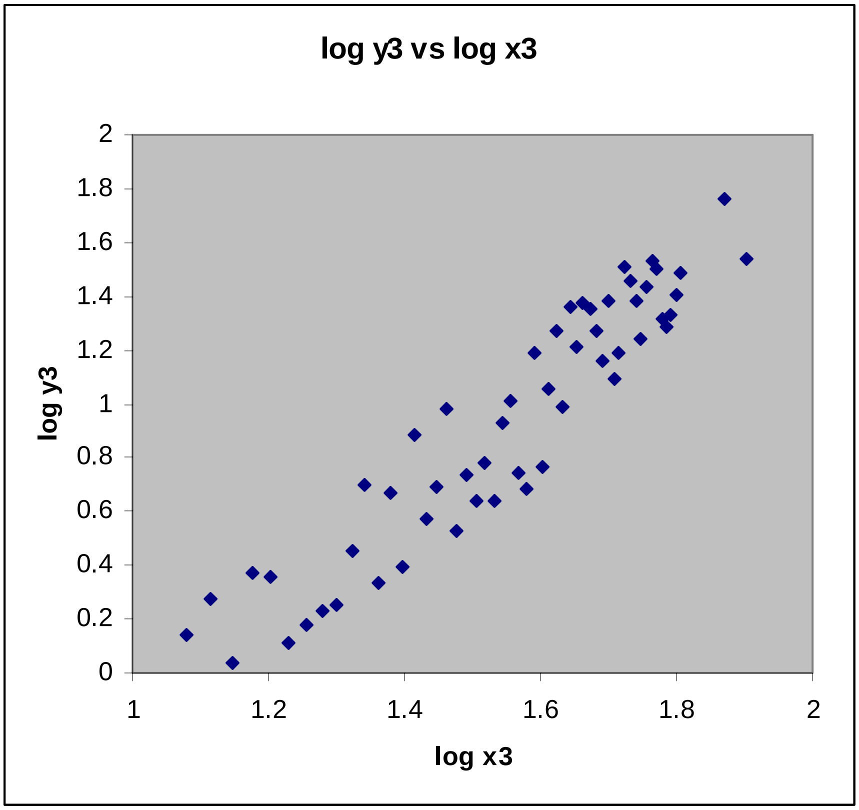


2. Notice the slight curve and also the “funnel” like shape. These are the typical signs that both the x and y variables need to be transformed.
3. Next create some columns of transformed variables. In cell Y1 type “log x3” and in Y2 type “=log(W2)”. Use the fill down technique discussed earlier to apply it to the whole column to get a column of log transformed x3 values. In cell Z1 type “log y3” and in Z2 type “=log(X2)”. Repeat the fill down technique in this column to get a column of log transformed y3 values.
4. Next create scatter plots of “y3 vs log x3” and “log y3 vs x3”. You should get plots that look something like the following. Notice that neither one is quite satisfactory. They both show some sort of curved relationship.



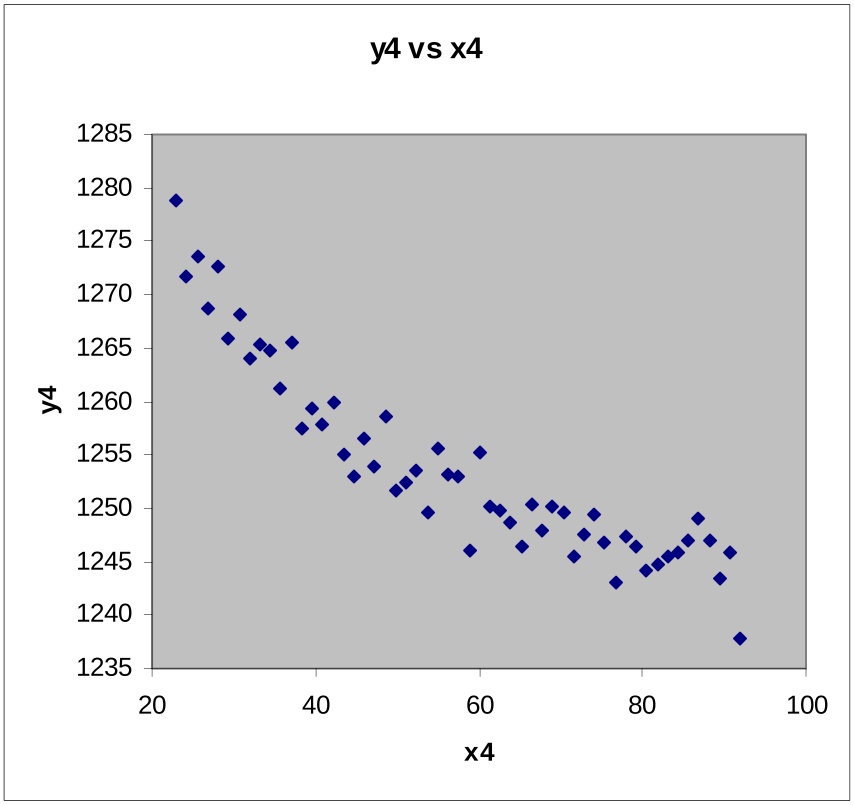


5. Finally plot “log y3 versus log x3”. A much nicer linear trend is evident.

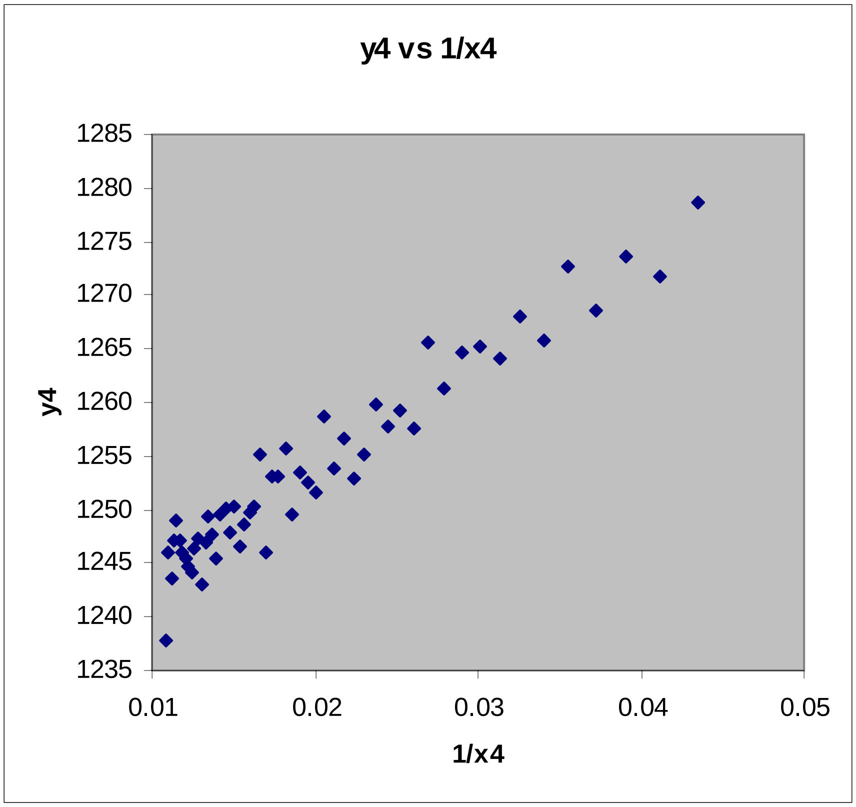


Dataset 4

1. First create a scatterplot where y4 is on the vertical axis and x4 is on the horizontal axis. The scatterplot should look something like this:



2. Notice the slight curve that looks a lot like the 1/x function..
3. Next create some columns of transformed variables. In cell AL1 type “1/x4” and in AL2 type “=1/AK2”. Use the fill down technique discussed earlier to apply it to the whole column to get a column of inverse transformed x4 values.
4. Next create a scatter plot of “y4 vs 1/x4”. You should get plots that look something like the following. Notice how much more linear it looks like.



Final words

You may see the final worked spreadsheet on the website.