

# **MATH 124 – Transforming Data**

[bolstad\\_math124@bmbolstad.com](mailto: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).

## **The Initial Data**

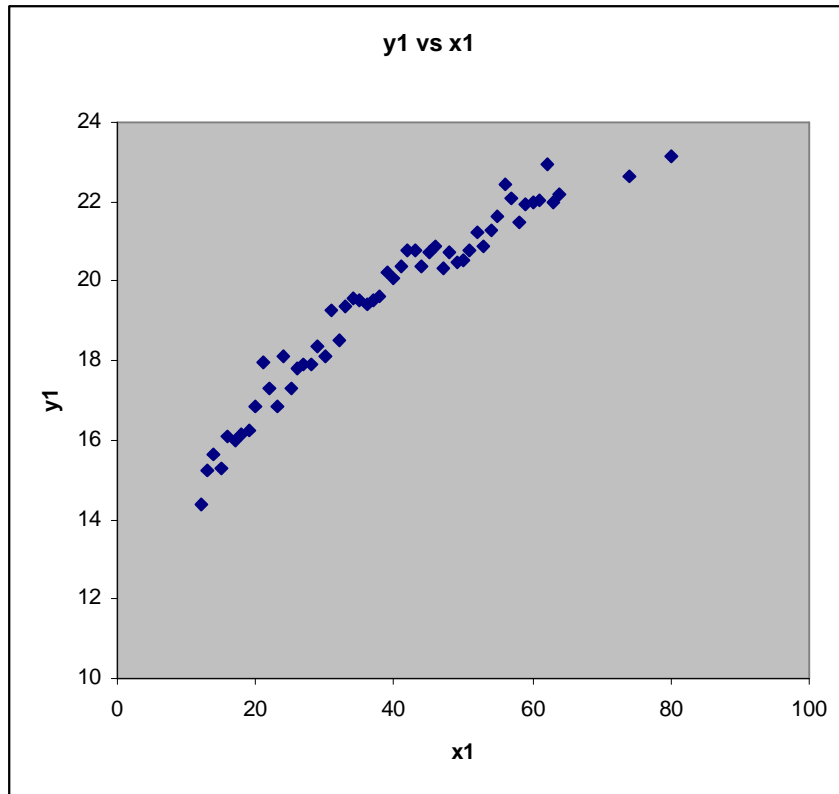
After you have read this datafile into excel you should see that there is three groups of data.

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.

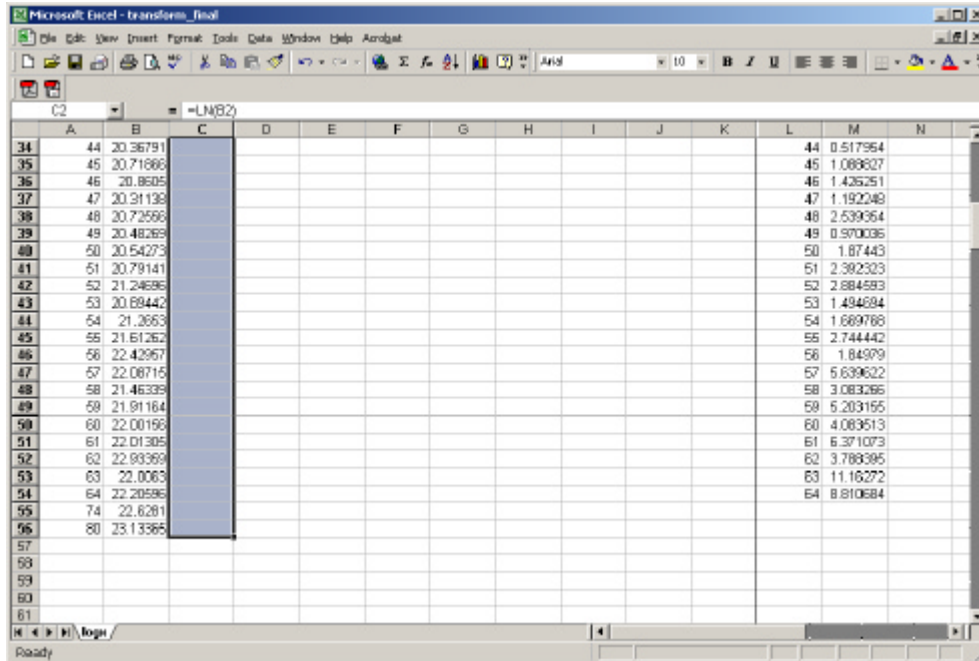
We will discuss how to first scatter plot the untransformed data and then create columns of transformed data and scatter plot to see that there is now a more clear linear trend.

## **Dataset 1**

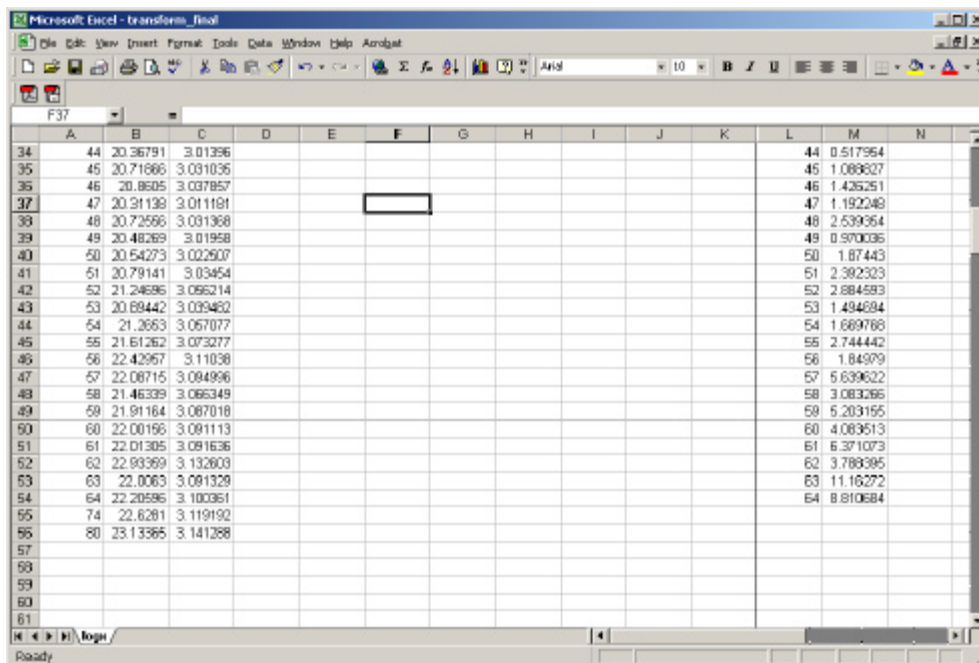
1. First scatterplot the data so that y1 is on the vertical axis and x1 is on the horizontal axis. Depending on how you choose to format your scatterplot you should end up with something like this:



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:

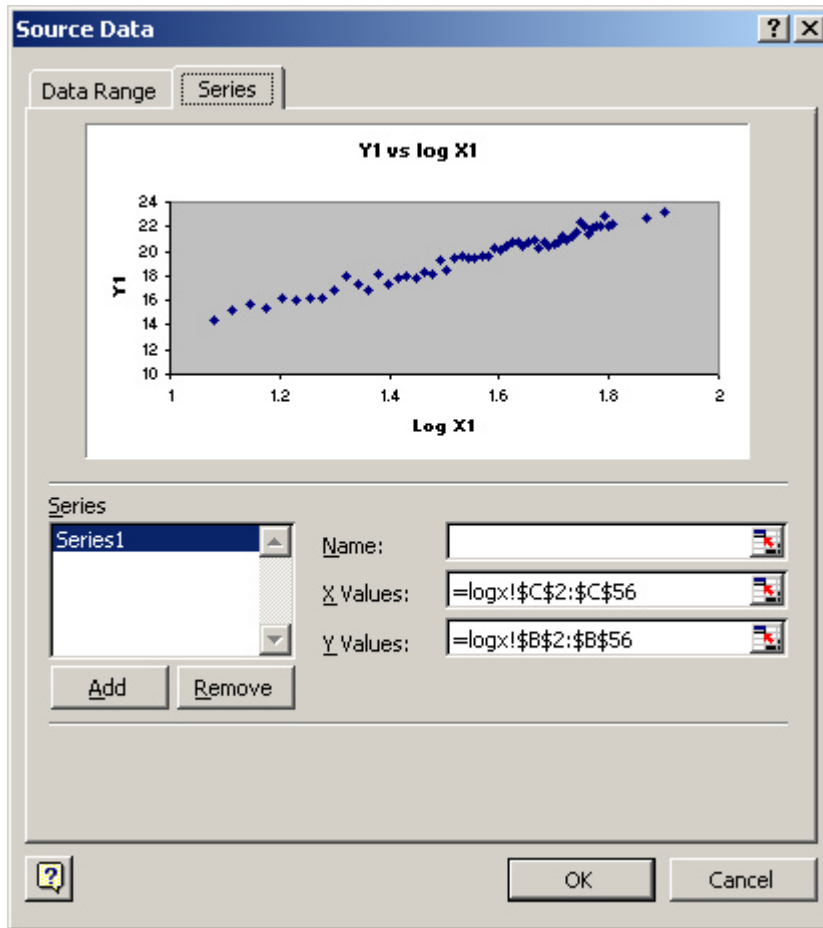


- Now either press CTRL+D or from the edit menu select Fill > Down. In either case what should happen is that numbers will now appear in that column something like this.

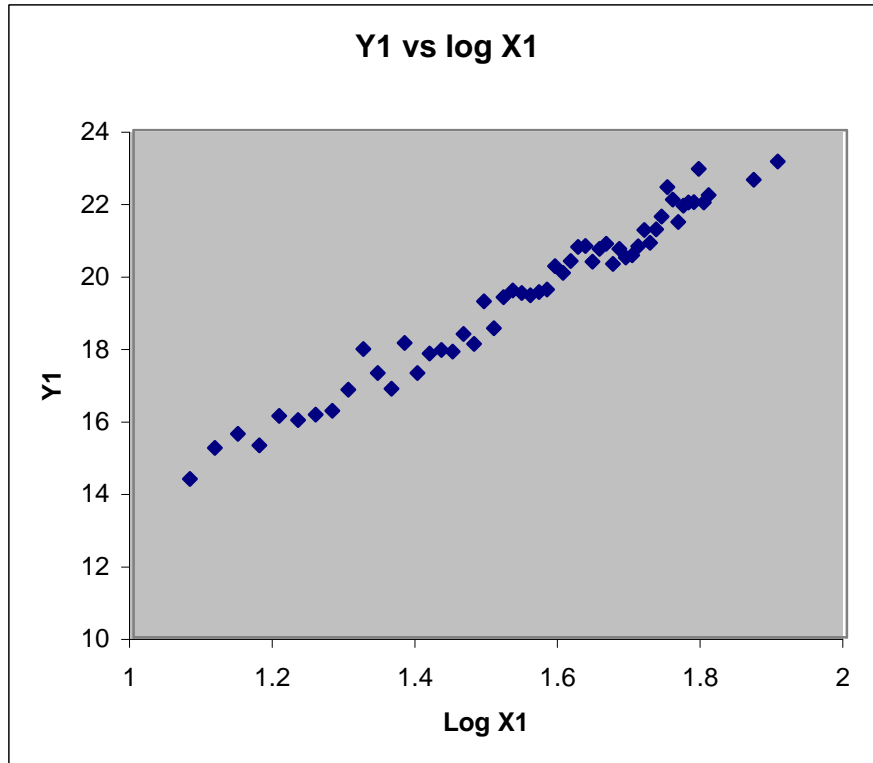


- This new column of data consists of log base 10 values of the data in column A (ie the x1 values). Next it is time to make a scatterplot of this transformed data to make sure that the transformation got rid of the curve. Repeat the steps you have used before to make a scatterplot. Note that it important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



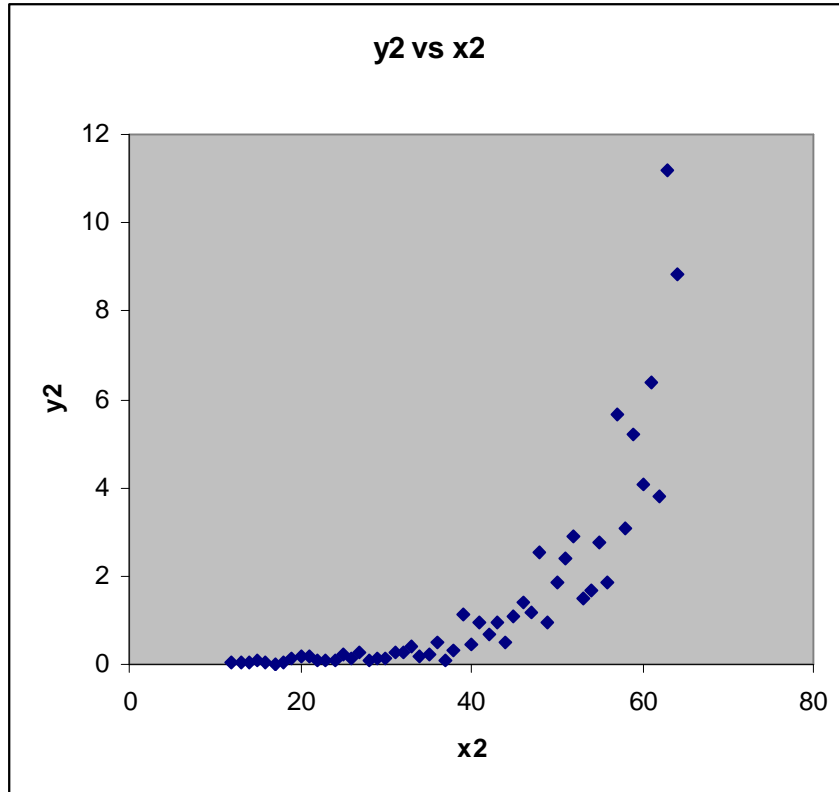
7. Proceed making your scatterplot and you should end up with something looking like this (again depending on how you formatted your plot you may differ slightly)



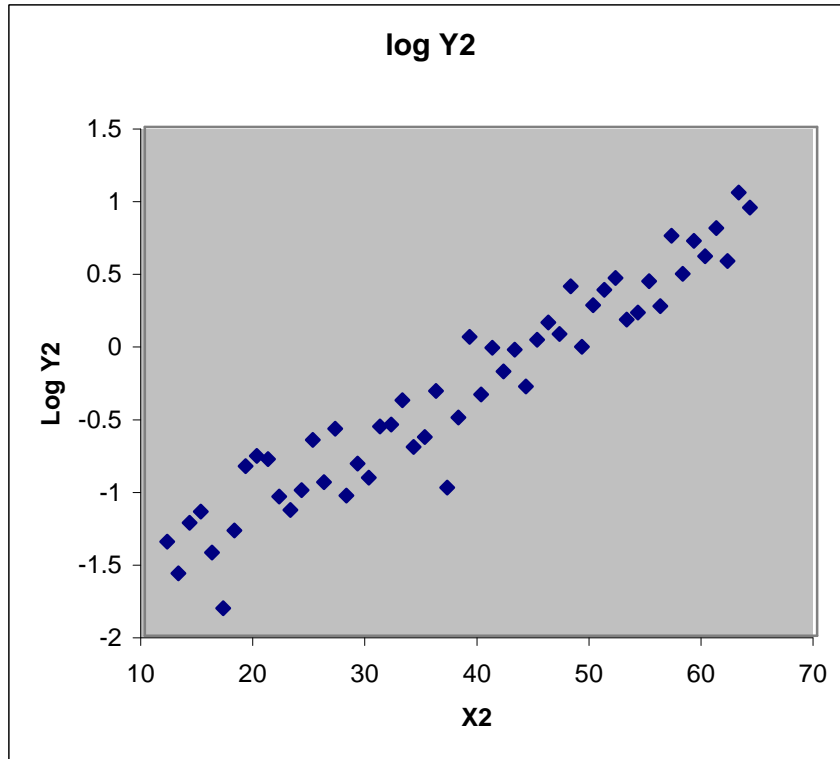
8. Notice that the relationship is much more linear than the original plot. The transformation in this example has been successful.

### Dataset 2

1. First create a scatterplot where  $y_2$  is on the vertical axis and  $x_2$  is on the horizontal axis. The scatterplot should look something like this:



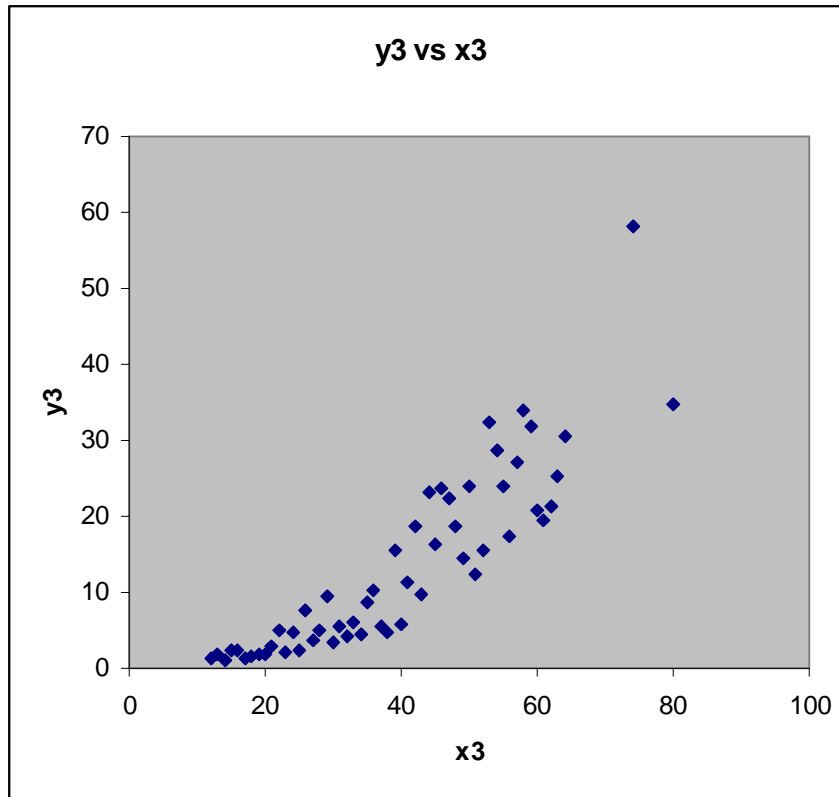
2. Notice how there is quite a clear curve on this plot
3. Now it is time to log transform the y2 values. Go to the cell N1 and type "log y2". Then go to the cell N2 and type "=log(M2)". Press enter.
4. As with the first dataset, select the cells N2 through N54. Then type CTRL+D or use the edit menu 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.

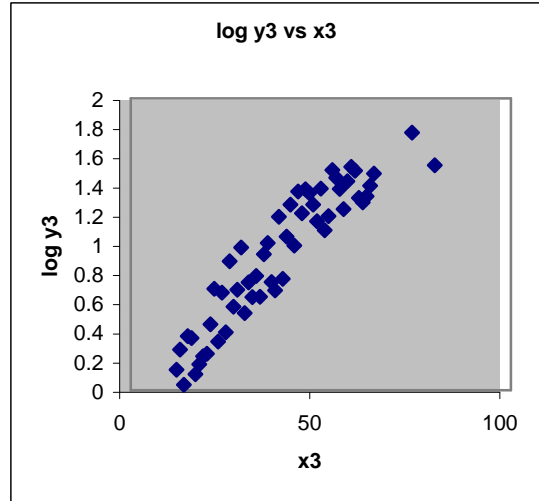
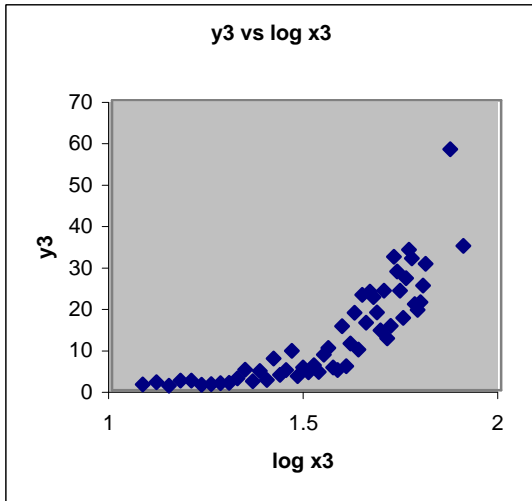
### Dataset 3

1. First create a scatterplot where  $y_3$  is on the vertical axis and  $x_3$  is on the horizontal axis. The scatterplot should look something like this:

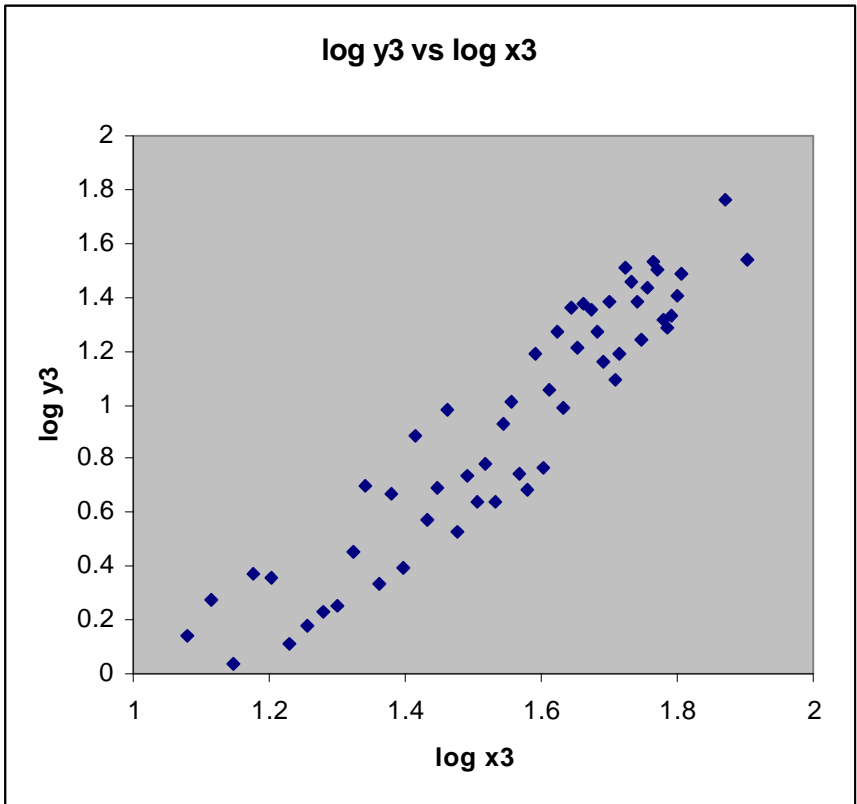


2. Notice the slight curve and also the “funnel” like shape. These are the typical signs that both the x and y variables need to be transformed.
3. Next create some columns of transformed variables. In cell Y1 type “log x3” and in Y2 type “=log(W2)”. Use the fill down technique discussed earlier to apply 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.



**Final words**

You may see the final worked spreadsheet on the website.