Excel Assignment 2.5

The Q-Q Plot

Purpose

In this assignment you will learn how to correctly do a Q-Q plot in Microsoft Excel. You will also learn that there is no "magic" behind Q-Q plot. However, in most other systems, such as R, normal Q-Q plot is available as a convenience feature, so you don't have to work so hard!

Loosly, these instructions are based on this link:
http://facweb.cs.depaul.edu/cmiller/it223/normQuant.html
The link contains a worked out spreadsheet example which fully explains the method.

Instructions

1. Follow the instructions provided here for data named "time", but replace the data with the residuals of the FEV data of *Excel Assignment 2*. You should standardize the residuals.
2. After you've obtained the "Normalized Q-Q plot" for your residuals, draw the regression line, display its equation and the coefficient of determination.

Step-by-step instructions

1. Place or load your data values into the first column. Leave the first row blank for labeling the columns. Sort the data in ascending order (look under the Data menu).
2. Label the second column as **Rank**. Enter the ranks, starting with 1 in the row right below the label. Each following row will be one more than the last (note: you can use an expression, copy and then paste to save you time)
3. Label the third column as **Rank Proportion**. This column shows the rank proportion of each value. Use this expression for the first data value **=(b2 - 0.5) / count(b$2:b$N)** where N should have the row number of the last cell. Finish the column by copying the first data expression to the remaining rows. Check to make sure your percentiles look like they are correct!

4. Label the fourth column as **Rank-based z-scores**. Excel provides these values with the **normsinv** function. Use this function to create the values in the fourth column.
5. Copy the first column to the fifth column. The Excel chart wizard works better if the x-axis values are just to the left of the y-axis values.
6. Select the fourth and fifth column. Select the chart wizard and then the scatter plot. The default data values should be good, but you should provide good labels.

Sample Data

The data is "time" and is in the first column. The remaining columns are auxillary columns used in creating of the Q-Q plot.

time	rank	percentile	rank-based z-score	time	
16.042	1	0.01923077	-2.069901831	16.042	
16.606	2	0.05769231	-1.574444965	16.606	
18.367	3	0.09615385	-1.303782672	18.367	
20.03	4	0.13461538	-1.104835744	20.03	
20.042	5	0.17307692	-0.942075775	20.042	
30.726	6	0.21153846	-0.801094529	30.726	
31.538	7	0.25	-0.67448975	31.538	
32.428	8	0.28846154	-0.557884763	32.428	
32.589	9	0.32692308	-0.448425483	32.589	
33.522	10	0.36538462	-0.344102463	33.522	
39.5	11	0.40384615	-0.243404178	39.5	
39.619	12	0.44230769	-0.145120941	39.619	
41.362	13	0.48076923	-0.048223074	41.362	
41.673	14	0.51923077	0.048223074	41.673	
45.874	15	0.55769231	0.145120941	45.874	
52.135	16	0.59615385	0.243404178	52.135	
59.999	17	0.63461538	0.344102463	59.999	
69.86	18	0.67307692	0.448425483	69.86	
72.879	19	0.71153846	0.557884763	72.879	
111.137	20	0.75	0.67448975	111.137	
113.141	21	0.78846154	0.801094529	113.141	
140.862	22	0.82692308	0.942075775	140.862	
143.063	23	0.86538462	1.104835744	143.063	
157.113	24	0.90384615	1.303782672	157.113	
165.308	25	0.94230769	1.574444965	165.308	
199.531	26	0.98076923	2.069901831	199.531	

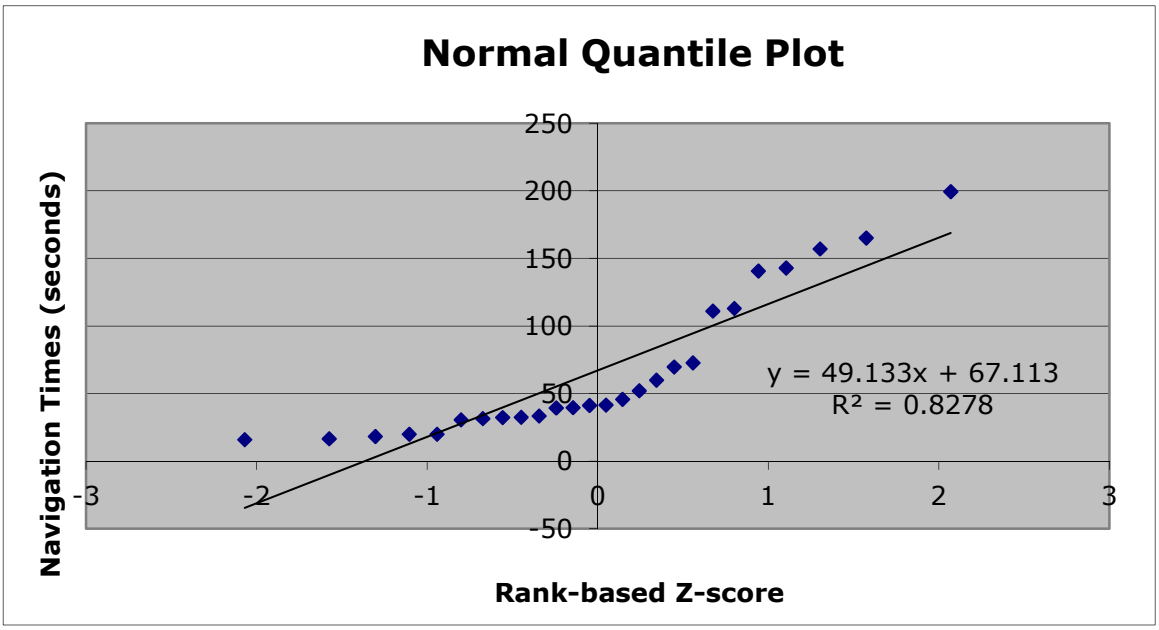
Comments on the created columns

1. "time" was sorted
2. Use "fill handle" to create the rank of the data. You could use the RANK() Excel function.
3. The cell =(B2-0.5)/COUNT(A$2:A$27), replicated using the "fill handle" is responsible for creation of the "percentile". This simply takes the rank of the sorted data and divides by the count. However, we subtract 0.5 to make up for the discontinuity of the rank (think of rank occupying a bar with base from B2-0.5 to B2+0.5; the rank is the middle of the bar).
4. The column "Rank-based z-score" is created by the cell content =NORMSINV(C2) and propagated downwards with the "fill handle". The content of C2 tells us the percentile, i.e. the fraction of the data above the row of cell C2. The function NORMSINV(C2) does the same as an inverse lookup in Table A:
    a. Use p-value equal to C2
    b. The content of the cell D2 will be the corresponding z-score (please check!)

    Thus, the z-score is the "theoretical" z-score (based on exact formulas) which corresponds to the fraction of data equal to the corresponding p-value.
5. The regression you have performed on the Q-Q plot data is the basis of one of the normality tests. You may read more here:

    http://en.wikipedia.org/wiki/Normality_test

The Normal Q-Q plot for sample data



Comment: The data were not normalized in this example, so the straight line is not close to y=x.

Also, the data does not appear quite normal, but R-squared is quite high.