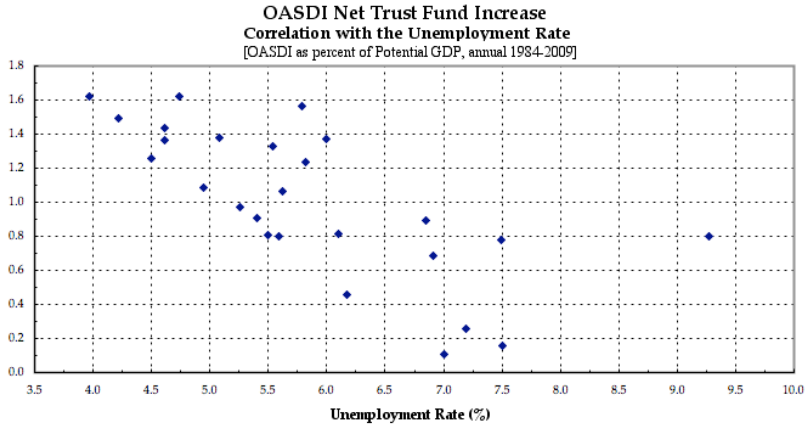
Date-Ordered Two-Variable Plots:
Identifying correlated subsets within a larger data set

The problem with scatter diagrams -- For many purposes, the ideal way to search for a relationship between two variables is to plot them against each other on X/Y axes -- the format commonly referred to as scatterplots, scatter charts, or scatter diagrams. However, this approach makes the implicit assumption that the universe of points plotted is a single set. This becomes a problem when working with time series in which the relationship might change, being different in different periods. This can be seen in the work IEA has done with the relationship between the unemployment rate and the Social Security Trust Fund's finances [1], using Microsoft Excel.

The spreadsheet [2] contains columns for the two series of annual data to be plotted, plus a column containing the years, in 2-digit form. The two series are selected, and the **(XY) Scatter Chart** type is chosen from Excel's **Chart Type** menu, producing the following:

Figure 1



Any linear relationship between these two series would be indicated by points appearing in a more or less straight line. At first glance, there appears to be nothing like that in this cloud of points -- trying to use a regression function on the entire set of data would be inappropriate. However, changes in Social Security policies and economic and demographic factors over time might conceivably change the relationship between these two series at various times, leading to internally consistent, but separate, periods. How to tease these out of the cloud?

Solving the problem -- Making any such periods visible requires two steps

1. connecting the data points with lines, and
2. labeling each point with the year it represents.

In Excel, the first can be accomplished by the following

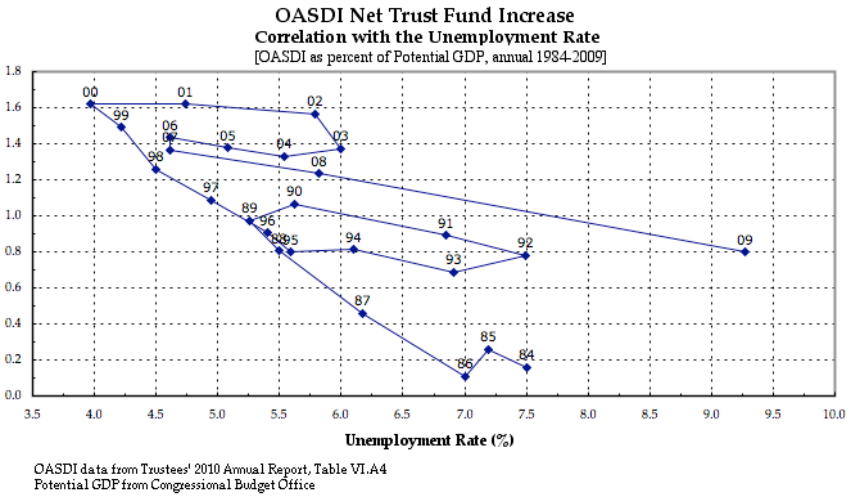
1. Double-click on any point.
2. Select **Patterns** from the **Format Data Series** dialog box.
3. In the **Line** frame, click **Automatic** or **Custom**.

Labeling the points cannot be done in any practical fashion with off-the-shelf Excel. However, it can be done easily with is a very helpful third party add-on called **XY Chart Labeler**, available free from Application Professionals [3]. With it installed, the procedure for labeling the points is:

1. From the Tools menu, select **XY Chart Labeler**.
2. From the submenu, select **Add Chart Labels**.
3. In the **Add Labels** dialog box
   a. The "name" of the series should already be present in the **Series to Label** box.
   b. In the **Label Range** box, give the full specs for the 2-digit year column -- i.e., in the form `Data!$B$51:$B$76`. Anything less doesn't work. And be careful typing -- you can't correct any typing errors. If you make a mistake, you'll have to **Cancel** and restart.
   c. Pick the label location relative to the points (we used **Above**).
   d. Click **OK**. (Note: you will have to wait a while before the labels appear.
4. If the labels are too far or too near the points, you can then choose **Move Labels** from the menu in step 2.

This creates what we call a ***date-ordered two-variable plot***, as seen in Figure 2:

Figure 2



OASDI data from Trustees' 2010 Annual Report, Table VI.A4
Potential GDP from Congressional Budget Office

Any linear patterns covering a given period of years will then show up within the cloud and can be examined for a relationship by running a linear regression on just the points in that period.

Making multiple linear regressions -- In this chart, we found 6 different periods of potential interest. In Excel, the procedure for doing a linear regression on a plot is as follows:

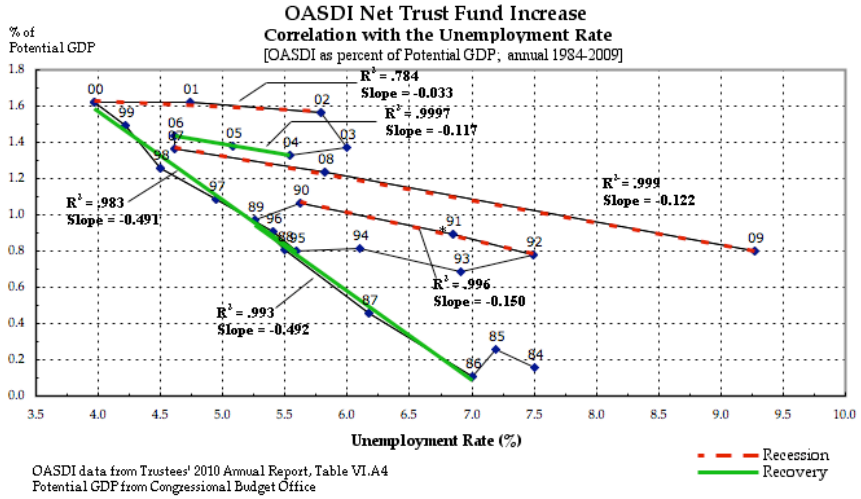
1. Right-click a point.
2. Choose **Add Trendline** from the resulting menu.
3. In the **Type** tab, select **Linear**.
4. In the **Options** tab, if desired click the **equation** and/or **R2-value Display** checkboxes.

However, Excel doesn't make possible the selection of a subset of points. So if you want to do regressions on one or more subsets of points, each subset must be added as a separate plot superimposed on the full-length plot, as follows:

1. Right-click the general chart area and choose **Source Data** from the resulting menu.
2. Under the **Series** tab
   a. Click **Add**.
   b. Give the sub-series a name (e.g., `'86-89`)
   c. For the X and Y columns, specify the row numbers for the new sub-range -- e.g., `Data!$C$56:$C$59`.
   d. Click **OK**.
3. Right-click one of the points of newly created sub-series and proceed with step 2 of the regression procedure above.

For our purposes, we replaced the automatic equation/R2-value texts with our own custom text boxes, and customized the regression lines, ultimately producing the chart in Figure 3:

Figure 3



Notes:

1. **IEA page for this work**: http://www.iea-macro-economics.org/ss_unemp_correlations_intro.html
2. **Spreadsheet**: http://www.iea-macro-economics.org/data/ss_table_iv_a4_2010.xls
3. **XY Chart Labeler** is available free from http://www.appspro.com/Utilities/ChartLabeler.htm. However, it is worth being aware that, since it is a macro, once it is installed, Excel must be started with macro-virus protection turned off in order to make use of the utility.

This document is available online at
http://www.iea-macro-economics.org/date_ordered_2var_plot.html