

3 Non-linearities & Dummy Variables

Reading:

Kennedy (1998) "A Guide to Econometrics", Chapters 3, 5 and 6

Aim:

The aim of this section is to introduce students to ways of dealing with non-linearities in the data.

Objectives:

By the end of this section you should be able to understand what is meant by non-linearity in relationships between variables and how such effects may not be easy to detect in scatter plots alone; to understand how simple t-tests can be used to test for particular kinds of non-linearities; and how dummy variables can be used to detect intercept and slope shifts.

Plan:

3.1	Introduction.....	3-1
3.2	Non-linearities.....	3-2
3.3	Testing for non-linearities using t-statistics:.....	3-4
3.4	Using dummy variables	3-6

3.1 Introduction

To those new to regression analysis, the process of searching and testing for non-linearities can seem like a fanciful exercise. Transforming the data to iron out non-linearities appears arbitrary and even dishonest. The question is often asked, "why should we assume that such exotic non-linear patterns occur in relationships between variables?"

In actual fact, it is more correct to turn this question on its head and ask, "Why should we assume that the relationship between variables is ever linear?" After all, few things in the natural world are truly linear. Have you ever seen a perfectly straight river? Or an entirely flat mountain range? Or have you ever met someone with a perfectly straight nose or shoulders so level you could rest a cup of tea on them? It is the exception rather than the rule to find unblemished linearity in the natural world. And the same is true in social and behavioural sciences.

The difference in the social sciences is that our data are often far less precise and are collected from observations of everyday life rather than in controlled lab conditions. So identifying the precise non-linear nature of the data is often somewhat illusive, and it is usually wiser to use a simple linear function to approximate a relationship as apparent non-linearities often turn out to be artefacts of sampling variation and measurement error than a true reflection of the underlying process. Nevertheless,

potent and clearly apparent non-linearities do sometimes exist between social science variables and since these would, in their untreated state, violate the assumptions underpinning regression, we need to at least test for them.

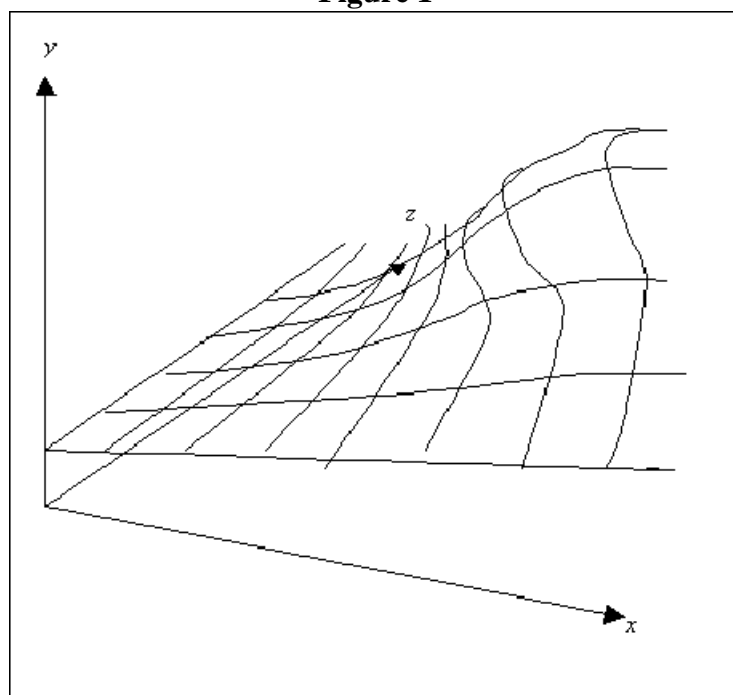
3.2 Non-linearities

What is the consequence of non-linearity? Depending on how severe the non-linearity is, estimates may be “biased” (i.e. they will not reflect the “true” values of α and β). We can test for non-linearities by looking at scatter plots and also by looking at individual t-statistics.

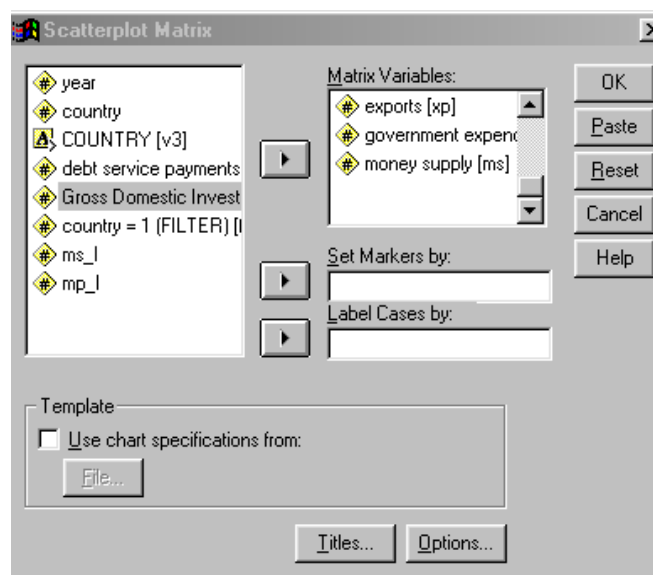
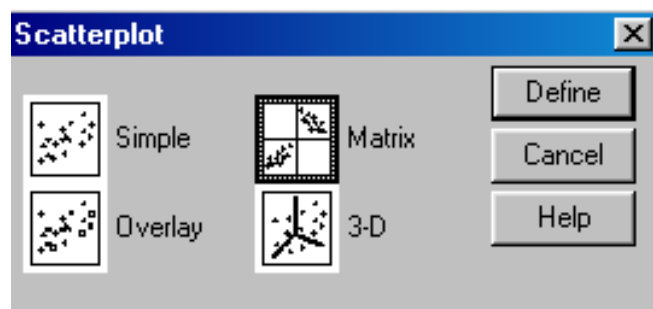
3.2.1 Visual inspection of Scatter plots

If you only have two or three variables then looking at scatter plots of these variables can help identify non-linear relationships in the data, but when there are more than 3 variables, non-linearities can be very complex and difficult to identify visually. What can appear to be random variation of data points around a linear line of best fit in a 2-D plot, can turn out to have a systematic cause when a third variable is included and a 3-D scatter plot is examined. Same is true when comparing 3D with higher dimensions. Non-linearities can be particularly difficult to spot from scatter plots if the source of the non-linearity is due to the interaction of two variables. In Figure 1, the relationship between y and x is relatively linear for low values of z . Similarly, the relationship between y and z is relatively linear for low values of x . However, the sensitivity of y to both x and z rises rapidly when both z and x are high. An example of this might be the effect of average window size and quality of view on house price. The effect on house price of either variables may be small for low values of the other (e.g. large windows add little to house value if the view is awful and visa versa), but large for high values of the other.

Figure 1



A useful facility in SPSS is the matrix scatter plot function. Go to Graphs, Scatter and choose Matrix. Then transfer the variables into the Matrix Variables box and click OK:



1. Open up the **sovdebt.sav** dataset which should be on the P:\ ...\Urban Studies directory. This data set lists sovereign debt and economic data on 43 countries over 12 years. Use the matrix of scatter plots facility in SPSS to look for non-linearities in relationships between variables. What sort of non-linearities can you observe from the matrix?
2. Where possible non-linear relationships appear to be present, run a simple scatter plot of the two variables. Also include a line of best fit (double click on the plot, choose Chart, Options, click the Fit Line Total box, and then click Fit Options. Select a linear fit line first, then try a quadratic or cubic fit. Which do you think most closely models the data?
3. Run linear regressions on the relationships explored in 2. Compare the regression output with the line of best fit plots from 2.

3.3 Testing for non-linearities using t-statistics:

Sometimes variables that we would expect (from intuition or theory) to have a strong effect on the dependent variable turn out to have low t-values. If so, then one might suspect non-linearities. One way to proceed is to try transforming the variable (e.g. take logs) and re-examine the t-values. Examples of transformations include:

Taking the natural log of a variable:

```
COMPUTE X1_L = LN(X1).  
EXECUTE.
```

Squaring a variable:

```
COMPUTE X1_SQ = X1 * X1.  
EXECUTE.
```

Cubing a variable:

```
COMPUTE X1_CUBE = X1 * X1 * X1.  
EXECUTE.
```

Exponent of a variable:

```
COMPUTE X1_EXP = EXP(X1).  
EXECUTE.
```

You might also want to try including an interactive term in there are two or more explanatory variables. You do this by creating a new variable that is equal to the product of the two variables that you think may be interacting. For example, if you believe X1 and X2 to be interacting, create a new variable X_1_2 and include this in your regression:

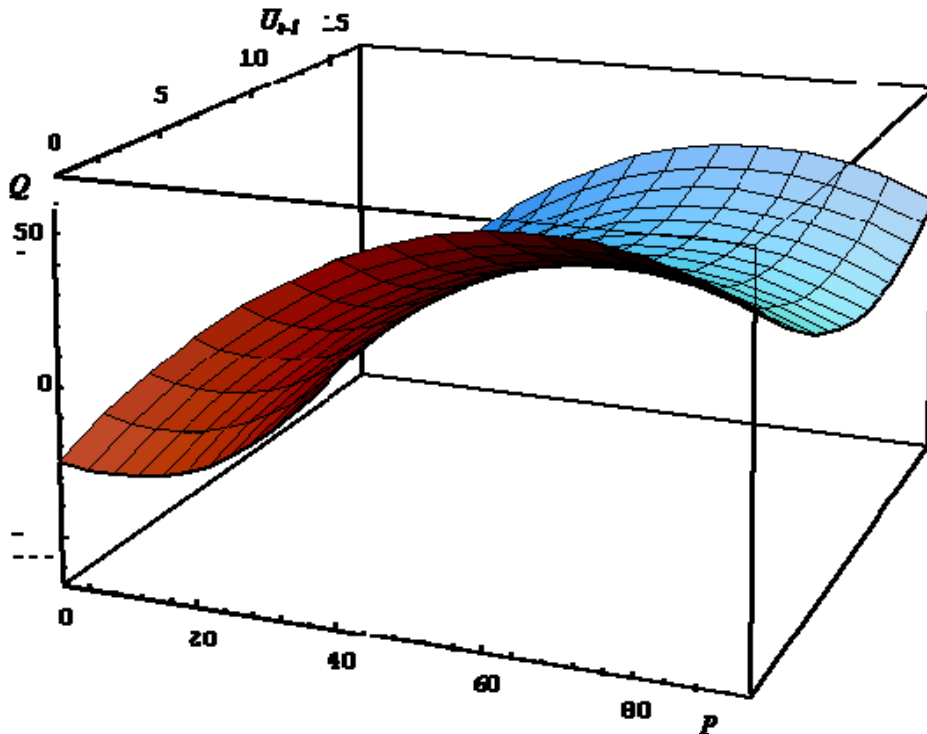
```
COMPUTE X_1_2 = X1 * X2.  
EXECUTE.
```

If the t-value is high (above 2 is a good rule of thumb since it leads to an associated significance level that is less than 0.05), then you can reject the null hypothesis that there is no interaction.

If the t-test indicates that the non-linearity is genuine, the estimated regression equation reflects not a straight line or surface of best fit, but some curve or shaped surface. An example of this is given below based on an estimated regression of new housing construction (Q) against house price (P) and lagged unemployment (U_{t-1}). The surface is a graphical representation of the following estimated equation:

$$Q = -246 + 27P - 0.2P^2 - 73U + 3U^2$$

As you can see, Q has a non-linear relationship with both price and unemployment. Fitting a simple linear surface to this relationship could result in severely distorted predictions and slope estimates.



-
4. What transformations do you think are necessary to remove the non-linearities you discovered in the sovereign debt data above? Try running the appropriate transformations and then run the scatter plot on the new variable. What difference has it made?
 5. Now try replacing the non-linear variables with their transformed equivalents in the regressions you ran. How do the regression results compare in terms of R-square, t-values, F-value etc.?
 6. Now try including two explanatory variables in your regressions. Create the appropriate interactive term and include this as a third explanatory variable. How does it fair in terms of its t-value and its impact on the adjusted R-square?
-

3.4 Using dummy variables

Sometimes certain observations display consistently higher y values. If this difference can be modelled as a parallel shift of the regression line, then we can incorporate it into our model simply by including an appropriate *dummy variable*. This is a categorical variable whose values are either zero or one. For example, if you think a particular country has idiosyncrasies that mean the intercept term is substantially higher or lower, then you may want to include the dummy variable in the regression. You can create a dummy variable for Argentina as follows:

```
COMPUTE ARGENT_D = 0.
EXECUTE.
IF (COUNTRY = 1) ARGENT_D = 1.
EXECUTE.
```

The first line sets all values of the new variable equal to zero. The third line then sets the values equal to one for those observations on Argentina. The coefficient on this variable will tell you how much higher the dependent variable is for the category = 1.

You can include many dummy variables in your regression. But bear in mind two things: first, the impact on the degrees of freedom of the regression. If you only have 35 observations, and you include 14 variables, you will be left with 21 degrees of freedom (i.e. only 21 observations left to actually run the regressions, the remainder will be used up just adding additional dimensions to the model -- NB a dimension is added each time a variable is added). The second thing to bear in mind is the *dummy variable trap*. This occurs when you include too many dummies and don't leave a baseline category. For example, if there are 43 countries in your dataset and you include a dummy for each, then the sum of dummies will all add up to one, and this will be perfectly correlated with your constant term (i.e. perfect multicollinearity). So you must always include no more than the total number of categories minus one.

You may of course believe there to be a change in *slope* due to the idiosyncrasies of the data. In this case, you would multiple your dummy by the variable you think might be affected. For example, if you believe the slope coefficient in the relationship between inflation and the money supply (where inflation is the dependent variable) to be steeper for Argentina, you would create a new variable as follows:

```
COMPUTE MS_ARG = MS * ARGENT_D.
EXECUTE.
```

(this assumes that you have already created the ARGENT_D variable). The coefficient on this slope variable would tell you how much steeper (or shallower) the slope is for Argentina.

7. Look over the scatter plots you have produced above and try to identify any systematic clustering of the data (e.g. into lines within the scatter plot). Try running the same scatter plots based on observations for just one year? What has happened to these lines of data? What do you think might be causing this?

8. Create dummy variables for all but one of the countries in your sample. You can do this quickly by just copying and pasting the syntax and using the country number to identify your dummy:

e.g. you might create a variable called COUNTRY1 which equals one if COUNTRY equals 1, and equals zero otherwise.

9. Include these into one or two of the multivariate regressions you had run previously. What has been their impact on the model? Are any statistically significant? How can you tell? What happens to the t-values and coefficients of the remaining variables if you drop out all dummies with t-values less than 2?
 10. Try creating year dummies and include them in your regression. What conclusions can you draw from your results?
-