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

So far we have assumed that the relationship between y and x is linear. That is, for a unit increase in x, y will change by a constant amount, irrespective of whether the unit increase in x is from a low or high base. This assumption rules out some fairly important and common forms of relationship that occur in social science, such as diminishing returns. For example, the modelled relationship between income and years of post-school education considered in last chapter was of the form:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

where a, b_1 and b_2 are the sample estimates of the population parameters α, β_1 and β_2. We estimated β_1 to be 1.45. That is, for every extra year in education, one can expect one's income to rise by £1,450. Is this realistic? One could argue that after doing an undergraduate degree, Masters degree, and PhD, any further study would add less to income. Similarly, we estimated β_2 to be 2.63, which says that, for every extra year of work experience, one can expect one's income to rise by £2,630. But is it really plausible that an extra year of experience for someone who has already worked for 40 years will increase income by the same amount as someone who has no previous employment? We might hypothesize that both years of education and years of experience will have a diminishing impact on income, illustrated in Figure 3.1. Clearly, a one year increase in experience at w will have a much greater impact on income than a one year increase in experience at v.

In the diagram, we have stated that the relationship between $\hat{y}$ and x_1, x_2 is:

$$\hat{y} = f(x_1, x_2)$$

That is, we have expressed the relationship between income, education and experience in very general terms—"$\hat{y}$ is a function of x_1 and x_2"—without specifying whether the relationship between $\hat{y}$ and x_1, x_2 is linear, quadratic, cubic, logarithmic or anything else. This expression does not even tell us whether the variables are positively or negatively related. This is our least presumptuous starting point: we start by saying little about the nature of the relationship between variables, and through a process of socio-economic theorizing, intuition and empirical investigation, establish more precisely how these variables are linked. In actual fact, the shape of the relationship in both directions in Fig 3.1 says quite a lot about what the functional form of the regression should look like. The dependent variable is clearly upward sloping in its relationship with both x_1 and x_2. And, as noted, the positive effect tends to diminish but not become negative (i.e. the curve levels off but does not bend backwards). The next question is to ask whether and how we can estimate such relationships using a linear method like OLS.

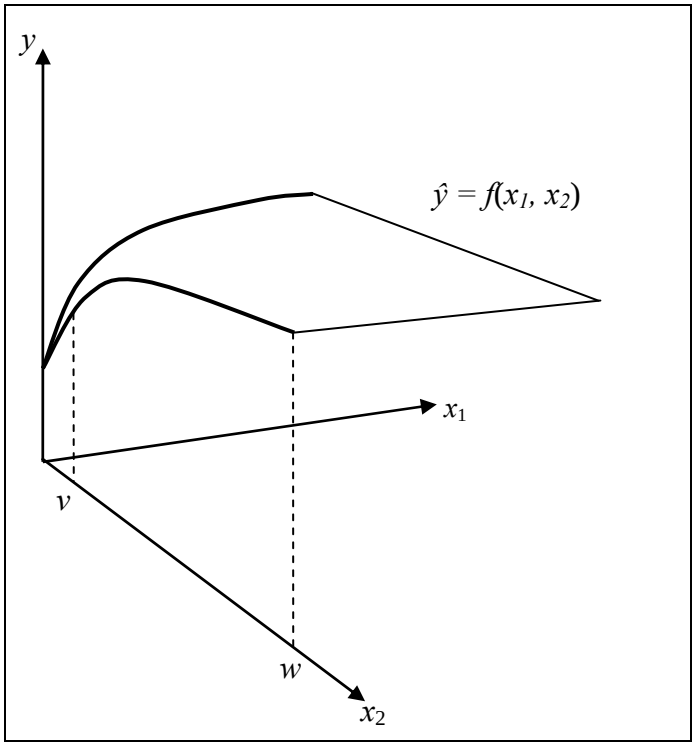


Figure 3.1 Diminishing Returns to Education and Experience

It is perhaps surprising, but nevertheless fortuitous, that non-linear relationships of the form illustrated in Figure 3.1 can actually be estimated using OLS without too much difficulty. Even though OLS is a linear estimation method, non-linearities in the relationship between y and x can be incorporated by transforming y and/or x *before* entering them into the regression algorithm. For example, while the relationship between income and experience is non-linear, the relationship between income and

the log of experience might be linear, in which case, running a regression of income on the log of experience would allow OLS to draw an appropriate line of best fit. The quest is therefore to establish what transformations to y and x are needed before they are entered into the regression.

For those new to regression analysis, the process of searching and testing for non-linearities can seem a little dubious. Transforming the data to iron out non-linearities can appear 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? How often 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. 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 wise to use the simplest function possible that will adequately approximate the relationship between y and x. Apparent non-linearities may be the result of sampling variation and measurement error, rather than a true reflection of the underlying social process. Also, non-linear relationships are more difficult to interpret, so it's often best to stick with a linear estimation where possible. 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 test for them and decide whether it's worth transforming the variables accordingly.

3.2 Non-linearities

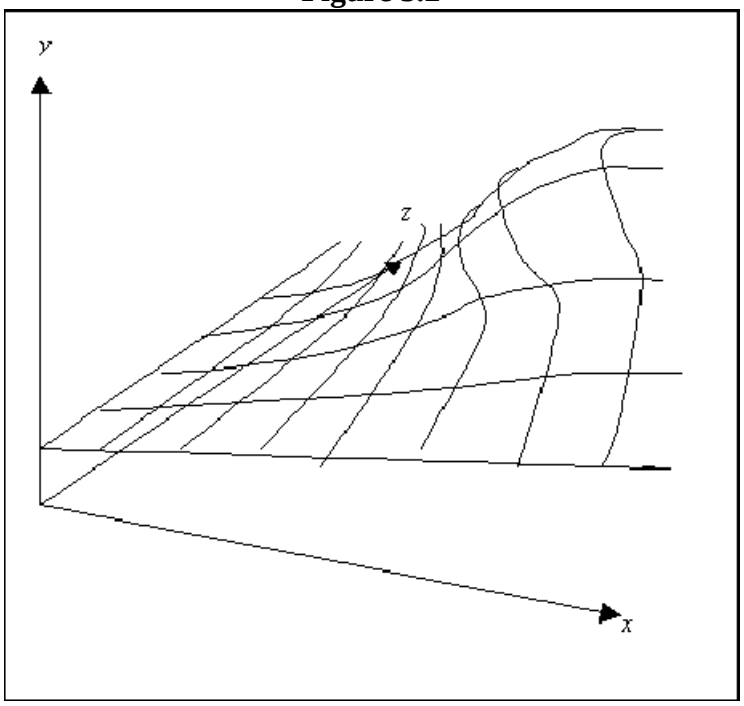
What is the consequence of non-linearity? Depending on how severe the non-linearity is, estimates may be "biased" (i.e. they will not reflect the "true" values of α and β). We can test for non-linearities by looking at scatter plots and also by looking at individual t-statistics.

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. The 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 3.2, the relationship between y and x is relatively linear for high 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 *vice versa*), but large for high values of the other.

Figure 3.2



A useful facility for comparing relationships between multiple pairs of variables is the matrix scatter plot function. Click on *Graphs*, *Legacy Dialogues*, *Scatter/Dot*, then choose *Matrix Scatter*, click *Define* and select the variables you want to enter. As you will see when you experiment with scatter plots, it is often very difficult to detect relationships from such graphs, let alone the nature of non-linearities. Sometimes it helps to plot a line of best fit and if that line is not horizontal you have an initial idea of whether there exists a relationship or not.

Similarly, if you use the Loess method to draw the line of best fit on the scatter graph, you can get an initial impression of whether a relationship is likely to be non-linear and the likely shape of the curve because loess curve fitting does not impose linearity. (Double click on the scatter plot in Statistics Viewer, then right-click on the data points, then choose *Add Fit Line at Total*, select *Loess*, click *Apply*, close the *Properties* window and close the *Chart Editor* to return to the *Statistics Viewer*). Of course, the lines of best fit on these scatter plots are only bivariate and as such do not control for the effect of other variables. Nevertheless, given a potentially infinite array of non-linear functions to choose from, such plots can provide a useful first step towards identifying the appropriate transformation.

3.2.2 Exercise:

1. Thinking about the relationship between crime and poverty, you conclude that poorer areas are likely to have higher rates of crime because poverty leads to alienation from the social norms of society. However, you hypothesise that this effect will diminish as income rises.

   (a) To explore this hypothesis, open the **crime_sns_dz.sav** dataset (taken from the Scottish Neighbourhood Statistics website: www.sns.gov.uk/) and run a the scatter plot of crime rates on income deprivation.

   (b) Re-run the scatter plot with extreme values of crime rate screened out (e.g. use the TEMPORARY. SELECT IF command or go to *Data*, *Select Cases* on the menu bar, and select *If condition is satisfied,* then click on the *If* button, and choose crime_rate_04 < 5000, for example).

   (c) Then apply a loess line of best fit. Note that you can change the colour and thickness of the line by clicking on the *Lines* tab in the *Fit Line at Total* window.

   (d) Try re-plotting the graph with a linear line of best fit. Do you think the loess curve better represents the relationship between crime and income? Does the curve have the shape you expected? Why should you be cautious about drawing firm conclusions from this data about non-linearities?

   (e) Taking logs is a way of compressing the range of values of a variable – it essentially reduces the size of large values relative to small values without affecting the ordering of the data. Try running regressions of crime against income, then crime against log of income, and then log of crime against log of income to see which works best.

2. Open the **crime_sns_dz.sav** dataset and run a matrix of scatter plots for the following variables: crime rate, employment deprivation, housing deprivation, income deprivation and the number of males aged 15 to 19 as a proportion of the population. From eyeballing these graphs, can you detect any evidence of non-linear relationships? Now try adding Loess lines of best fit to see if that helps you identify non-linearities.

Some Basic Shapes to Watch Out For

Once you have drawn the loess line of best fit, you can then decide what sort of transformation will address non-linearity of that shape. While it is true that the range of non-linear possibilities is infinite, social scientists are rarely seeking to estimate precise relationships between variables, and so the goal in addressing non-linearities is usually limited to addressing four basic types of non-linearity:

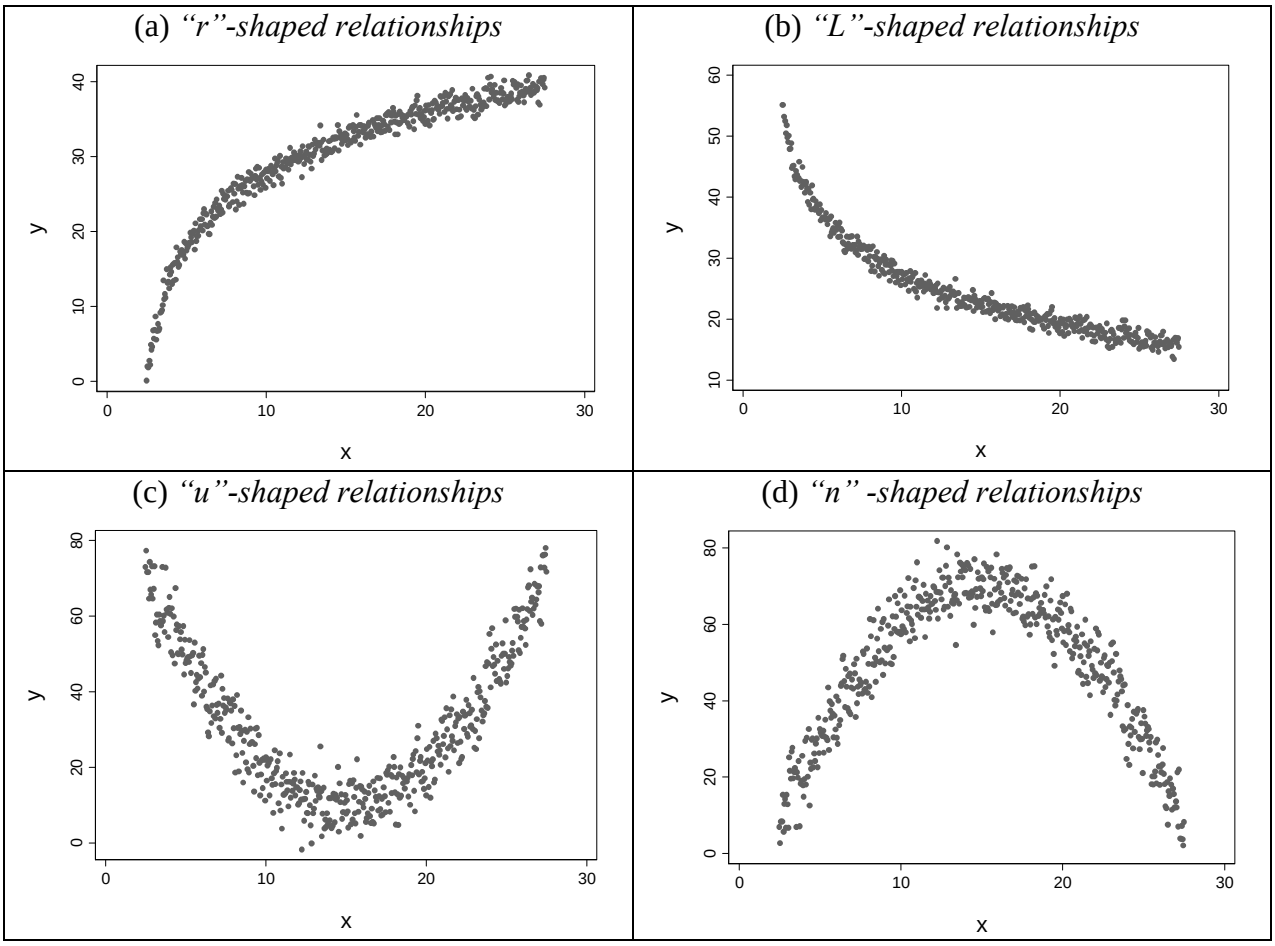
(a) *"r"-shaped relationships* (i.e. where the effect of *x* on *y* starts off being strongly positive, but eventually levels-off without becoming negative at

higher values of x). See panel (a) of the figure below. We say that y is "monotonically increasing" in x because an increment in x will always have a positive effect on y. The simplest mathematical function of this type is a *positive logarithmic relationship*, and so we often use the natural log function (written as $\ln(x)$ or $\log(x)$) to address this type of non-linearity. However, we can also use *quadratic* function (panel (d) of the figure below) provided we are clear that it is only the first half of the quadratic function (the upward sloping part) that we are interested in—the turning point implicit in the quadratic function (occurring at around $x = 15$ in panel (d)) is beyond the range of our data on x.

(b) *"L"-shaped relationships* (i.e. where the effect of x on y is initially positive, but eventually becomes negative at higher values of x). See panel (b) of the Figure below. We say that y is "monotonically decreasing" in x because an increment in x will always have a negative effect on y. The simplest mathematical function of this type is a *negative logarithmic relationship*, and so we often use the log function to address this type of non-linearity. Again, we can also use *quadratic* function (panel (c) of the figure below) provided we are clear that it is only the first half of the quadratic function (the downward sloping part) we are interested in—the turning point implicit in the quadratic function (occurring at around $x = 15$ in panel (c)) is beyond the range of our data on x.

(c) *"u"-shaped relationships* (i.e. where the effect of x on y is initially negative but eventually becomes positive at higher values of x). We say that y and x have a "non-monotonic" relationship because an increment in x will not always have an effect on y in the same direction. The simplest mathematical function of this type is a *positive quadratic relationship* (panel (c)), and so we often use the *quadratic* function to address this type of non-linearity. Quadratic functions are simply those where the highest power of x is x^2.

(d) *"n" -shaped relationships* (i.e. where the effect of x on y is initially positive, but eventually becomes negative at higher values of x). Again, we say that y and x have a "non-monotonic" relationship because an increment in x will not always have an effect on y in the same direction. The simplest mathematical function of this type is a *negative quadratic relationship* (panel (d)) and so we often use the *quadratic* function to address this type of non-linearity.



In the exercise below you will see how taking logs and applying the quadratic transformation (i.e. squaring *x*) can help "linearize" the above relationships. For example, you might deal with each of the following non-linear patterns as follows:

(a) *"r"-shaped relationships*: you would compute a new variable equal to the log of *x* and then run a regression of y on ln(*x*):

```
COMPUTE lnx = ln(x).
REGRESSION /DEPENDENT y /METHOD=ENTER lnx .
```

(b) *"L"-shaped relationships*: again, you would compute a new variable equal to the log of *x* and then run a regression of y on ln(*x*). See syntax for (a) above.

(c) *"u"-shaped relationships*: you would compute a new variable equal to the square of *x* and then run a regression of y on x^2, and perhaps also include the original *x:*

```
COMPUTE xsq = x**2.
REGRESSION /DEPENDENT y /METHOD=ENTER xsq .
REGRESSION /DEPENDENT y /METHOD=ENTER x xsq .
```

(d) *"n" -shaped relationships*: again, you would compute a new variable equal to the square of *x* and then run a regression of y on x^2, and perhaps also include the original *x*. See syntax for (c) above.

3.2.3 Exercise: Transforming Variables and Linearizing Relationships

1. Open the **x_y1_y2_y3_y4.sav** dataset. Explore the relationship between y1 and x as follows:

    a. Run a scatter plot of y1 on x. What sort of relationship do you observe?

    b. Add a quadratic line of best fit and compare this with a linear line of best fit.

    c. Run a regression of y1 on x and comment on your results.

    d. Using the `COMPUTE` command, create a transformed version of x called xsq = x^2. Run a scatter plot of y1 on xsq with a linear line of best fit.

    e. Run a regression of y1 on xsq, and then a regression of y1 on x and xsq. How well do these two regressions perform compared with the regression of y1 on x? Which is the best model out of the three?

    f. To visualise how well the best model works, run a regression of y1 on predicted values, and add a linear line of best fit.

2. Repeat question 1 for the relationship between y2 and x.

3. Explore the relationship between y3 and x as follows:

    a. Run a scatter plot of y3 on x. What sort of relationship do you observe?

    b. Add a linear line of best fit and comment on your results.

    c. Run a regression of y3 on x and comment on your results.

    d. Using the `COMPUTE` command, create a transformed version of x called lnx = lnx. Run a scatter plot of y3 on lnx with a linear line of best fit.

    e. Run a regression of y3 on lnx. How well does this regression perform compared with the regression of y3 on x? How does it compare with a regression of y3 on x and x^2?

    f. To visualise how well the best model works, run a regression of y3 on predicted values, and add a linear line of best fit.

4. Repeat the steps suggested in question 3 for the relationship between y4 and x.

Some Important Properties of Logs:

For the purposes of capturing non-linearities in regression estimation, you do not need to know how natural logs are calculated, but it is worth knowing some basic properties of logs. First, we can only compute ln(x) for positive values of x, so if some of your x observations have negative or zero values, it might be worth using the quadratic transformation, x^2, instead. Alternatively, you could add a constant to x before taking logs, but this can make interpretation a little more tricky. For example, suppose you have some values of x that are 0, if you add one to all x values, i.e. compute ln(x+1), then the log transformation should work without any difficulty.

Second, a very useful feature of log transformations is that if you also compute ln(y) and run a regression of ln(y) on ln(x):

```
COMPUTE lnx = ln(x).
COMPUTE lny = ln(y).
REGRESSION /DEPENDENT lny /METHOD=ENTER lnx .
```

the estimated coefficient on lnx can be interpreted as an "elasticity"—i.e. the proportionate change in y due to a proportionate change in x:

> Elasticity of y with respect to x = % change in y / % change in x.

If the elasticity of y with respect to x is greater than one (or less than negative one), we say that the relationship is "*elastic*" – which is jargon for y being relatively sensitive to changes in x. That is, a given % change in x will cause a larger % change in y.

> E.g. if ln$y$ = a + 2.4ln$x$, the elasticity of y with respect to x = 2.4, this means that, other things being equal, a 10% increase in x will cause a 24% increase in y.
>
> E.g. ln$y$ = a - 1.7ln$x$, the elasticity of y with respect to x = -1.7, this means that a 10% increase in x will cause a 17% reduction in y.

If the elasticity of y with respect to x is less than one (or greater than negative one), we say that the relationship is "*inelastic*" – which is jargon for y being relatively **in**sensitive to changes in x. That is, a given % change in x will cause a smaller % change in y.

> E.g. if ln$y$ = a + 0.5ln$x$, the elasticity of y with respect to x = 0.5, which means that, other things being equal, a 10% increase in x will only cause a 5% increase in y.

E.g. if $\ln y = a - 0.7\ln x$ the elasticity of y with respect to x = -0.7, which means that a 10% increase in x will cause a 0.7% reduction in y.

If the elasticity of y with respect to x is exactly equal to one, we say that the relationship between y and x has "unitary" elasticity. (In the above examples we have not specified a value for the intercept term, a, because it does not affect the calculation or interpretation of elasticities).

Elasticities are useful because they provide a common scale for comparing the sensitivity of y to changes in different explanatory variables. Note, though, that elasticities can only be calculated for explanatory variables that are continuous (e.g. they are not appropriate for dummy variables, which are considered later in this chapter).

3.2.4 Exercise: Computing and Interpreting Elasticities

1. Open the **crime_sns_dz.sav** dataset and run a regression of ln(crime rate) on ln(income rank). How would you interpret the coefficient on ln(income rank)?

2. Open the **x_y1_y2_y3_y4.sav** dataset. Suppose y3 is annual salary measured in £000s, and x is years of work experience. Run a regression that will calculate the elasticity of salary with respect to years of work experience.

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.
```

or,

```
COMPUTE X1_SQ = X1**2.
EXECUTE.
```

Cubing a variable:

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

or,

```
COMPUTE X1_CUBE = X1**3.
EXECUTE.
```

Exponent of a variable:

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

You might also want to try including an interactive term if 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:

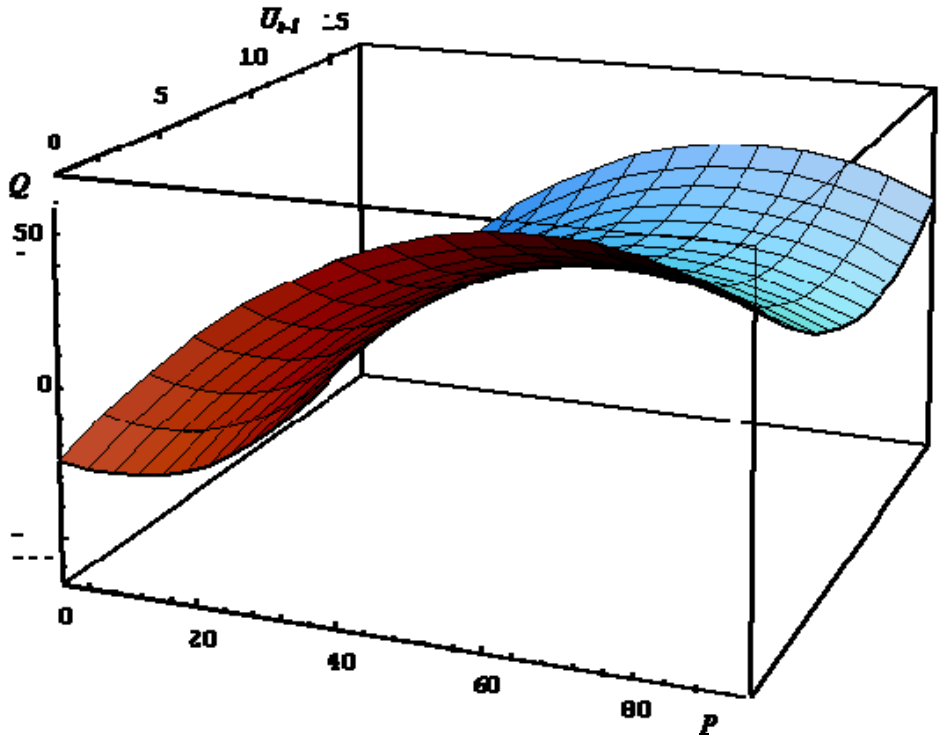
```
COMPUTE X_1_2 = X1 * X2.
EXECUTE.
```

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 is 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.



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.
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 multicolinearity). 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.

3.4.1 Exercise: Dummy Variables

1. Open the **sovdebt.sav** data. Suppose you believe that the relationship between inflation and the money supply might be different for those countries with negative amortisation (i.e. those countries that are adding to their sovereign debt rather than paying it down). Create dummy variables = 1 for all countries with negative amortisation, = 0 otherwise. Include these dummies in a regression of inflation on the money supply. Derive confidence intervals for this coefficient and comment on your results.

2. Try creating year dummies and include them in your regression. What conclusions can you draw from your results?