

1 Preliminaries: Graphical Analysis

Before discussing functional form as it pertains to X and Y , I want to spend a little bit of time on visual displays of data. Visual displays of your data, whether univariate displays or bivariate displays, can reveal a lot of information regarding the shape and nature of X as well as the relationship between X and Y .

Commonly, we proceed as if our data are normally distributed and/or symmetrically distributed in one way. Nevertheless, it is not unusual at all to see data that are heavily skewed and that significantly depart from normality (although please note that we do not need to assume the data are normally distributed, particularly the covariates). This can affect the inferences you make because parameter estimates in a regression model, which heavily rely on variances and covariances between Y and the X_k , will be affected by oddly distributed data. Sometimes you cannot do anything about the distribution of your data; other times you can through the use of transformations and functional forms that differ from regression “defaults.”

At any rate, it would seem to be useful to understand how to visually describe your data so, at minimum, you have a sense of what it is you are working with. Surprisingly, visual examination of data is really underutilized in social research, at least political science research. I want to briefly describe some approaches, in the context of **Stata**, to create graphical displays of your data. Some of this material should be review from last semester.

1.1 Box-and-Whisker Plots

One visual approach I find extremely helpful is through the use of box and whisker plots. The plot gets its name from the fact that such plots yield a box, which contains information on the quartiles of a distribution, and whiskers, which give information on the maximum and minimum numbers in a distribution. Such plots are useful for detecting outlying observations, departures from normality, or even coding errors. Let me illustrate. In Figure 1, I provide the box plot for the variable X_1 that we used in our examples for the lectures on interactions.¹ The box gives information about the quartiles. The bottom line on the box denotes the 25th percentile (the 1st quartile); the middle line denotes the median; and the upper line denotes the 75th percentile (the 3rd quartile). The whiskers extend to the upper and lower adjacent values (that is, they approximately represent the upper and lower values in the distribution).²

The box plot seems to suggest that the data are right-skewed somewhat, as the whisker extending to the maximum adjacent value is longer than the whisker extending to the minimum adjacent value. If the data were symmetrical, then we should see the whiskers being

¹This figure was generated using the data set `interact.dta` and by giving the command `graph x1, box ylab t1("Box-and-Whisker Plot")`.

²They are *approximately* the upper and lower numbers because extreme nonadjacent values will be individually graphed. Adjacency is based on deviations from the 25th and 75th quartiles. For a fuller explanation, see the *Stata Graphics* manual (p. 35, v.7).

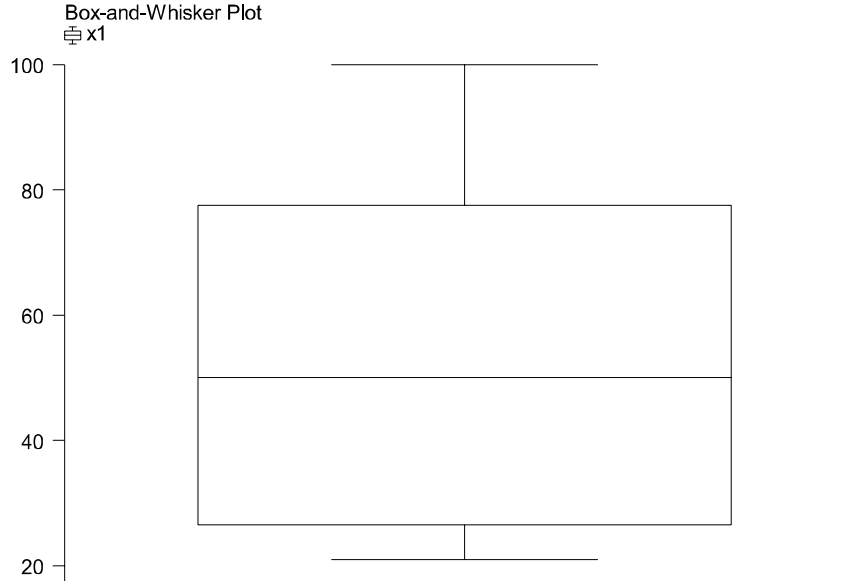


Figure 1: *Box-and-Whisker Plot for X_1*

more-or-less the same length (why?). Box plots are often profitably used to compare subpopulations or subgroups in the data. For example, using the `interact.dta` from last time, I compute the box plot for X_1 separately for Group 1 and Group 2 (that is, I compute it by the variable D_1). This gives Figure 2.³

This figure illustrates that the distribution of X_1 across the two groups is considerably different. Had we ignored this information and proceeded with the standard additive linear model, we could very well have missed the fact that X_1 seems to be conditional on the value of D_1 . Of course this knowledge led us to our interactive model discussed in prior lectures. In general, it is a recommended strategy to look at a box plot of your data before proceeding onto complicated models.

1.2 Stem and Leaf Plots

Stem-and-leaf plots are another way to visually examine your data. Below I present the stem-and-leaf plot for variable X_1 .⁴ The stem corresponds to the digit to the left of the vertical bar. The digits to the right correspond to the leaves. Through stem-and-leaf displays, one can conveniently reconstruct the observations on X_1 and easily examine the data for outlying observations, extreme asymmetry, miscodes in the data, and the presence of possible subpopulations on the variable.

2* | 1225679

³This figure was produced by the command: `graph x1, box by(d1) ylab t1("Box-and-Whisker Plot by D1")`.

⁴This display was obtained by the command: `stem x1`.

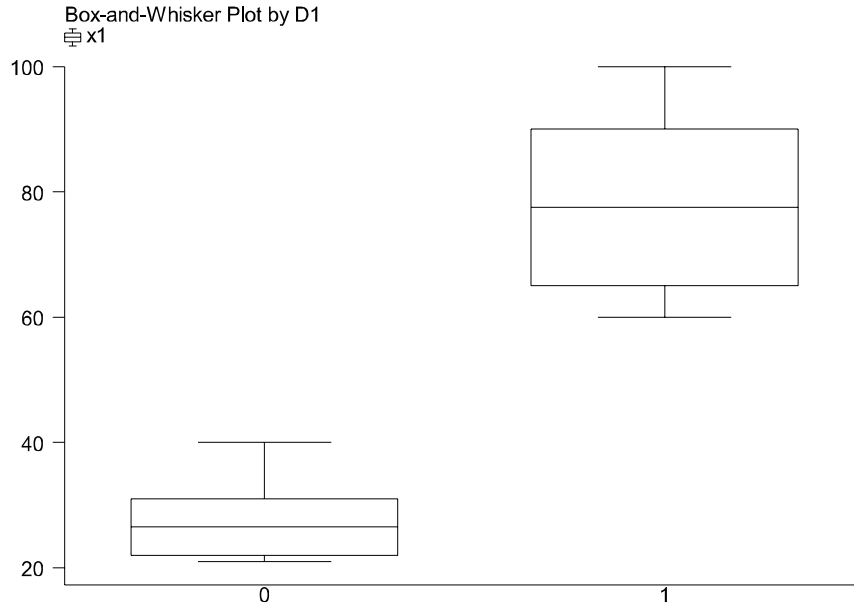


Figure 2: *Box-and-Whisker Plot for X_1 by D_1*

3*		17
4*		0
5*		
6*		025
7*		05
8*		05
9*		05
10*		00

Here we see that there are 7 observations beginning with the 2: 21, 22, 22, 25, 26, 27, and 29. The stem-and-leaf plot also shows a breakage in the values on X_1 between $X_1 = 40$ and $X_1 = 60$. This breakage, as it turns out, corresponds to the differences in X_1 accountable by D_1 (see Figure 2).

1.3 Departures from Symmetry and Normality

More sophisticated graphical displays can be obtained to answer questions pertaining to departures from symmetry and normality. The logic of these kinds of displays is simple: if the data are symmetric and/or approximately normally distributed, then what should the distribution look like? Hence, a set of reference points are given that correspond to the condition of symmetry or normality against which we can observe the empirical distribution of our data.

Symmetry plots ask the question of how symmetrical is the distribution of some variable.

The logic is simple. First order the data from smallest to largest values, $X_{11}, X_{12}, X_{13}, \dots, X_{1n}$ (the X_i are known as the order statistics for X_1 ; let us define the order statistic as $g_{(i)}$). If the data are symmetrically distributed, then

$$\text{median} - g_{(i)} = g_{(N+1-i)} - \text{median}; \tag{1}$$

that is, deviations above the median will be identical to deviations below the median. This gives us a benchmark to compare the distribution of our data. In Figure 3, I graphically display the symmetry plot for the variable X_{12} .⁵

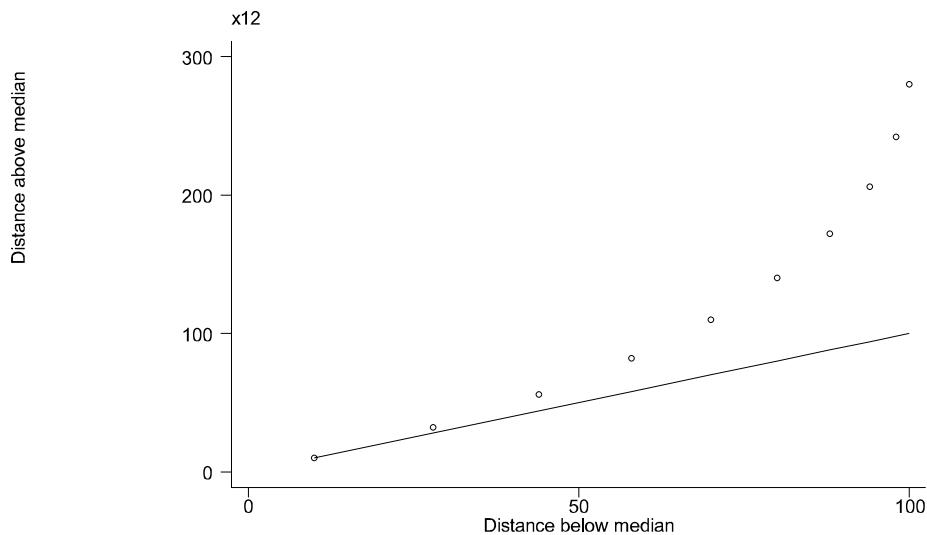


Figure 3: *Symmetry Plot for variable X_{12} .*

If the data were symmetrically distributed, then each of the data points would fall on the reference line, which is premised on Equation (1) holding. Since large values of X_{12} fall above the reference line, this suggests that the data are right-skewed.

Normal probability plots are constructed premised on the question: how “close” is the empirical distribution of my data to a normal distribution. One popular kind of normal probability plot is known as a quantile probability plot. The logic of this plot is straightforward. First, compute the order statistics for some variable X_i (which is done by arranging the data from smallest to largest). Second, estimate the cumulative distribution function for the data. Third, compare the estimated CDF to the CDF for the normal distribution. If the data points deviate from the line, then the data exhibits departures from normality (Fox in Chapter 3 does a good job of describing this). To illustrate, I’ve computed the quantile normal probability plot for the variable X_{12} . This is shown in Figure 4.⁶ We see that the distribution of X_{12} deviates from normality, as the data points do not fall on the line. The departure from normality is not too severe, but it is something we should be aware of.

⁵In **Stata**, the symmetry plot is gotten by typing `symplot x12, ylab xlab`.

⁶This figure was generated by the command: `qnorm x12, ylab xlab t1("Quantile Plot for x12")`.

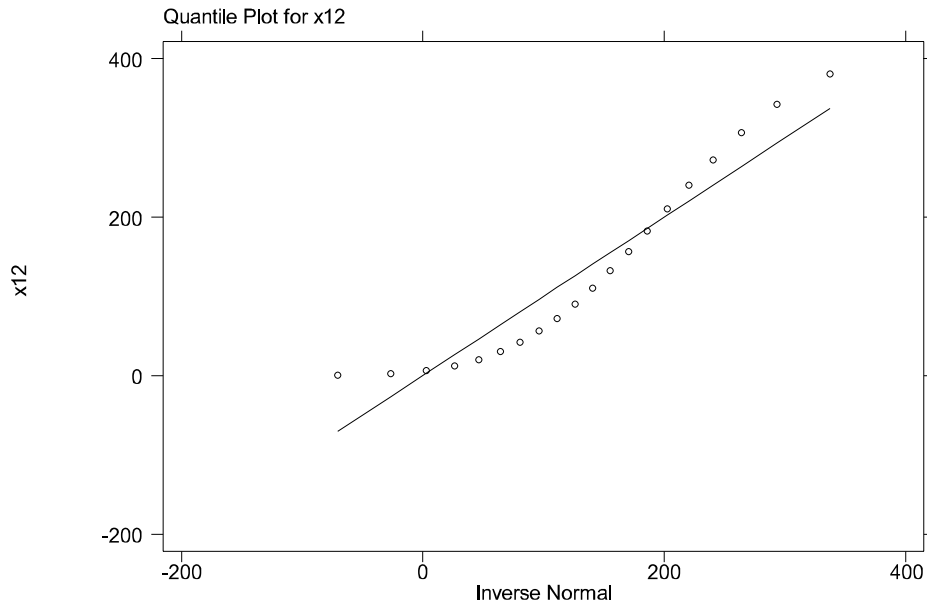


Figure 4: *Quantile Normal Probability Plot for variable X_{12} .*

The general points I want you to understand is that visual inspection of your data can substantially improve your understanding of the problem. Further, visual displays may clue you into potential problems that could emerge if you use the data to estimate a regression model. The plots shown in this section are univariate. Frequently we will care about bivariate plots (or even multivariate plots). When it comes to functional form of a linear regression model, these relationships will be of the utmost of importance. Nevertheless, before barging into a statistical model, it should be a matter of course that you visually inspect your data. A lot can be learned.

We now turn attention to function form.

2 Motivation for Considerations of Functional Form

Through the use of multiplicative terms in interaction models, we can assess how slopes vary, conditional on the value of some other covariate. In this sense, we can estimate more sophisticated linear relationships among our variables.

However, one thing that we haven't considered to this point is how to incorporate non-linear relationships inside of a linear regression model.

Why might we want to think about doing this? The reasons are simple. First, given the desirable properties of the OLS model, it would be nice to use our data and stay within the framework of the linear model. Second, if a relationship is nonlinear, then transformation on X may produce a better fitting regression model than one where the relationship between X and Y is constant (and linear).

To fix ideas, let us begin with a simple bivariate model of the form

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1. \tag{2}$$

The interpretation of the slope coefficient is standard in that it gives us the change in $E(Y)$ given a unit change in X . However, this model is very restrictive in that this additive effect is constant over the range of X . Suppose we apply this model to the data shown in Table 2.

Table 1: Data for Y and X_1 .

Y	X
.46	.5
.47	1.5
.56	2.5
.61	3.5
.61	4.5
.67	5.5
.68	6.5
.78	7.5
.69	8.5
.74	9.5
.77	10.5
.78	11.5
.75	12.5
.8	13.5
.78	14.5
.82	15.5
.77	16.5
.8	17.5
.81	18.5
.78	19.5

Using these data, we estimate the following model,

$$\hat{Y} = .542 + .016X_1$$

which has an $RMSE = .055$ and an $F = 60.05$). In computing the predicted values of the regression model, we obtain the plot shown in Figure 5.⁷ There is nothing particularly noteworthy about this regression function. We've known how to interpret this model since the first week of class. However, make sure you understand what is going on here. Because the slope is constant (and linear) across the range of X_1 , this suggests that the change in $E(Y)$ when X_1 increases from, say, .5 to 1.5, is exactly the same as when X_1 increases from 18.5 to 19.5.

Hence, the relationship exhibits no marginality. Nevertheless, it seems difficult to believe in the general case that the relationship between some covariate X and Y will be constant over the full range of X . It is reasonable to believe that X may have a saturation point after which its relationship to Y changes. That is, the marginal change in Y may decrease as X continues to increase. Of course the obverse could be true: as X increases to some point, the $E(Y)$ could begin to increase at an increasing rate.

⁷This plot was generated by the following `Stata` command:
`graph xb y x1, ylab xlabel c(1) t1("Standard Model").`

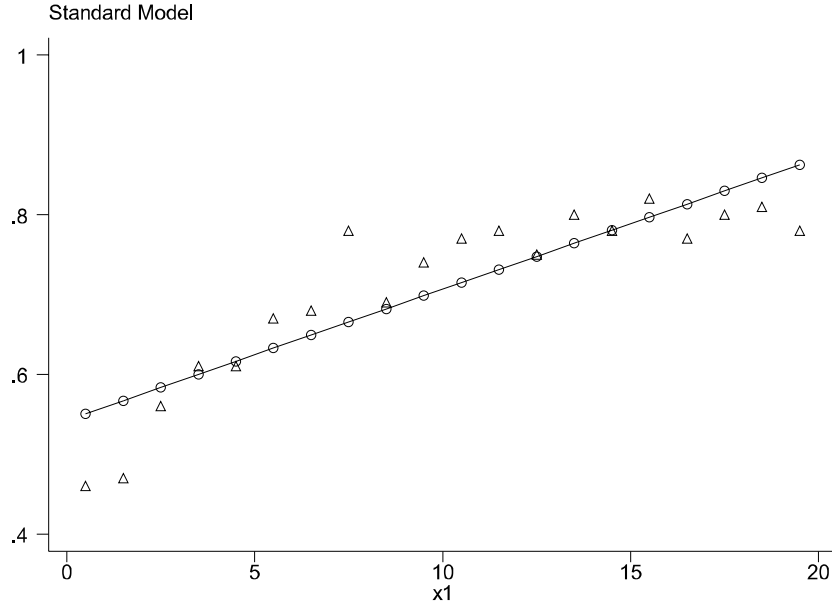


Figure 5: *Predicted Regression Function and Y*

Another feature shown in Figure 5 is that of *monotonicity*. The relationship between X_1 and Y is monotonic—that is, the change is in one direction. This is a property of the least squares model. Nevertheless, it may be conceivable that the relationship between a covariate and Y is *nonmonotonic*—that is, after some point, the slope of the relationship “changes direction.”

The main point is that in some instances, the relationship between a covariate (or covariates) and the dependent variable may not possess the property of having a constant, monotonic, linear effect. In Figure 6 I illustrate some possible relationships that could occur using some simulated data. Each of the relationships shown in Figure 6 seem to suggest that the slope of Y on X_1 may not be constant over the full range of X_1 ⁸

For example, in the top right panel, we see that the relationship seems to flatten out as X_1 increases. We could imagine that the “true” slope would be positive but would tend to 0 as X_1 increased. The bottom two panels illustrate nonmonotonic relationships. Here it is easy to see that the relationship between X_1 and Y first increases (decreases) and then decreases (increases).

The “story” that could be told is considerably different from a story based on a model where a constant linear effect is assumed. To illustrate, suppose that the relationships shown in Figure 6 held, but we naively estimated a linear regression model like that shown in (2). In Table 2 I present the results from 4 standard regression models. Ignoring the graphical displays of data, the interpretation of these models is simple. Looking at the first column, we see that for a unit increase in X_1 , the expected value of Y decreases by about -.193 units.

⁸To create this graph, I first created the four individual graphs using Stata’s `graph` command (they were named `nonlin1`, `nonlin2`, `nonlin3`, `nonlin4`. I then combined the four graphs using the following command: `graph using nonlin1 nonlin2 nonlin3 nonlin4, b2("Nonlinear" Data)`.

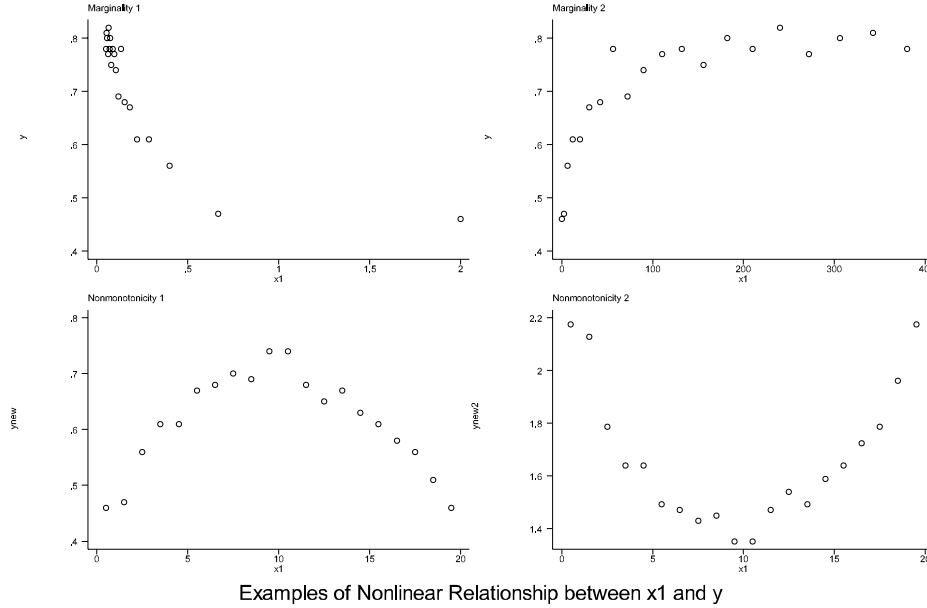


Figure 6: This figure gives different representations of “nonlinear” relationships between X_1 and Y .

This effect is constant over the full range of X_1 , although the graphical display of the data seems to suggest that the slope decreases (i.e. tends to 0) as X_1 gets large.

Table 2: Regression Models Based on Data from Figure 6.

Coefficient	Marginality 1	Marginality 2	Nonmonotonicity 1	Nonmonotonicity 2
Intercept	.754 (.019)	.616 (.025)	.617 (.04)	1.66 (.12)
Slope	-.193 (.038)	.0007 (.0001)	-.0003 (.0035)	-.00009 (.01)
<i>RMSE</i>	.074	.076	.091	.27
r^2	.58	.56	.00	.00

The extreme cases are found for the two regression models estimated for the nonmonotonic data found in Figure 6. Here, the slope coefficient tells us that a unit increase in X_1 is associated with essentially *no change* in $E(Y)$. That is, the slope is flat. The story we would get from these two models is that X_1 is unrelated to Y . In Figure 7, I regraph the data along with the predicted regression functions.⁹ This figure illustrates an obvious, but important point: a linear model will produce a straight line (in the bivariate setting, as

⁹The following Stata command was used to create this figure:
`graph using nonlin1xb nonlin2xb nonlin3xb nonlin4xb, b2("Regression Functions with Nonlinear Data")`
 (see the previous footnote for elaboration on how to compute composite figures).

is the case here). The straight line is based on predictions premised on the least squares principle; however, it is clear that the extent to which our data deviate from a more-or-less linear pattern affects the quality of the model. To further illustrate the point, In Figure 8,

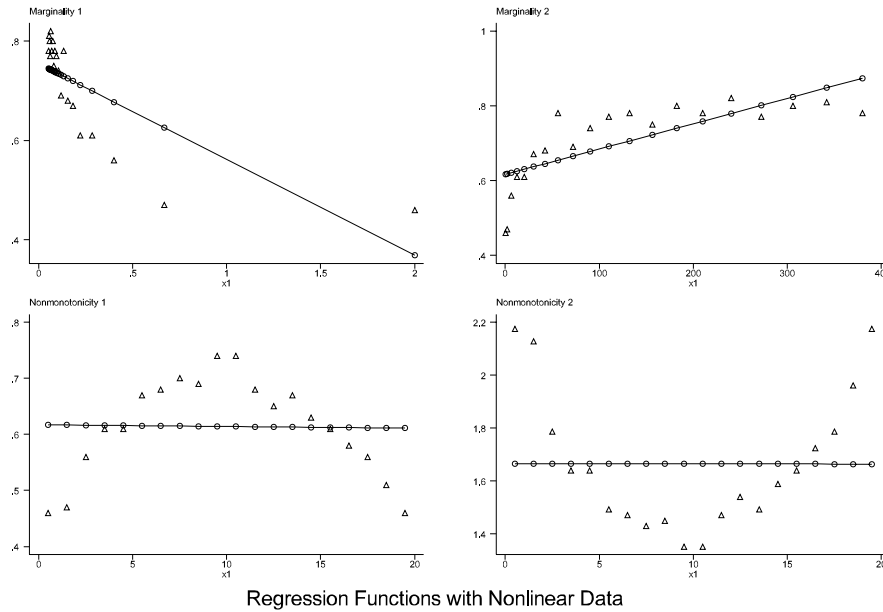


Figure 7: This figure gives the estimated regression functions for the four regression models.

I present plots of the residuals for each of the models. As we should know, the assumption regarding the distribution of the residuals with respect to the fitted values (which of course are a function of X_1) should be random. As we can see in Figure 8, the distribution of the residuals is far from randomly distributed about 0. Residual plots like these are useful in diagnosing potential problems with the standard regression model.¹⁰

To summarize, for data far from approximately linear, the straight line of the regression function is a very poor way to describe the data! Because nonlinear relationships are not estimated very well in the context of the garden variety linear regression model, we need to think of ways to account for nonlinearity *while still preserving the integrity of least squares estimation theory*. To this we turn next. I consider several approaches, the first involving segmenting one's slope coefficients.

¹⁰In order to compute the residual plots, after each regression model I typed the command `rvfplot, yline(0) t1("Nonmonotonicity 2") ylab xlab`, where the `t1(" ")` option puts a title on top of the y axis. The `rvfplot` command gives you the residual vs. fitted values plot. To combine the graphs, I typed the command `graph using nonlin1xbres nonlin2xbres nonlin3xbres nonlin4xbres, b1("Residual vs. Fitted Values Plots")`. See footnote 1 for further details.

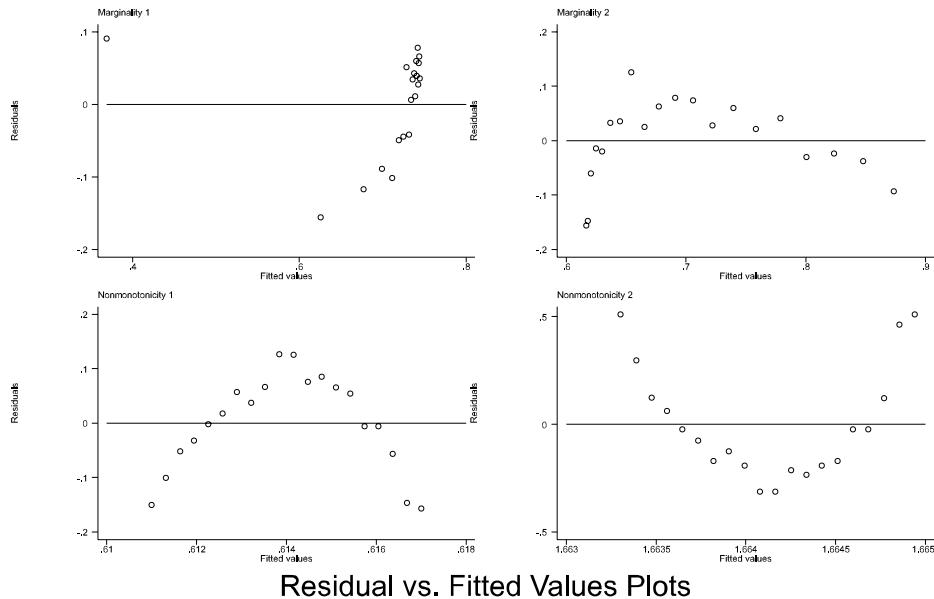


Figure 8: This figure gives the residual vs. fitted plots for each of the four models. Note the distribution of the residuals seems nonrandom.

3 Segmented Slopes

The idea of segmented slopes or *piecewise* slopes gives rise to the idea that the slope coefficient may vary across different ranges of the data. This leads to the consideration of model where different slope coefficients are estimated, conditional on a given range of the data.

We've already seen the idea of conditional slope coefficients in practice. Interaction terms allow the slope to conditionally vary as a function of some other covariate. The basic ideas motivating interaction terms helps lead us to the idea of segmenting the slopes. To explain, let's return to the example in the first section. In Figure 9, I re-plot the data from Figure 5.

The figure seems to suggest that the relationship between Y and X_1 is curvilinear such that the "slope" seems to become less steep at higher levels of X_1 . To help clarify this, I graph the residuals against the values of X_1 and present the results in Figure 10. The residuals were obtained from the previous regression model.¹¹ In the figure, it appears that the slope changes pitch around the value $X_1 = 7$. To illustrate this, I include a vertical reference line at this point.

We may be able to improve our estimate of the relationship between X_1 and Y by accounting for the apparent nonlinearity in the data. A segmented slopes approach would allow us to do this. The logic is to estimate a separate slope for the data above and below

¹¹This graph was obtained by outputting the residuals from the regression model by typing `predict r, resid` and then graphing them by typing `graph r x1, ylab xlab xline(7) yline(0) t1("Residual Plot")`.

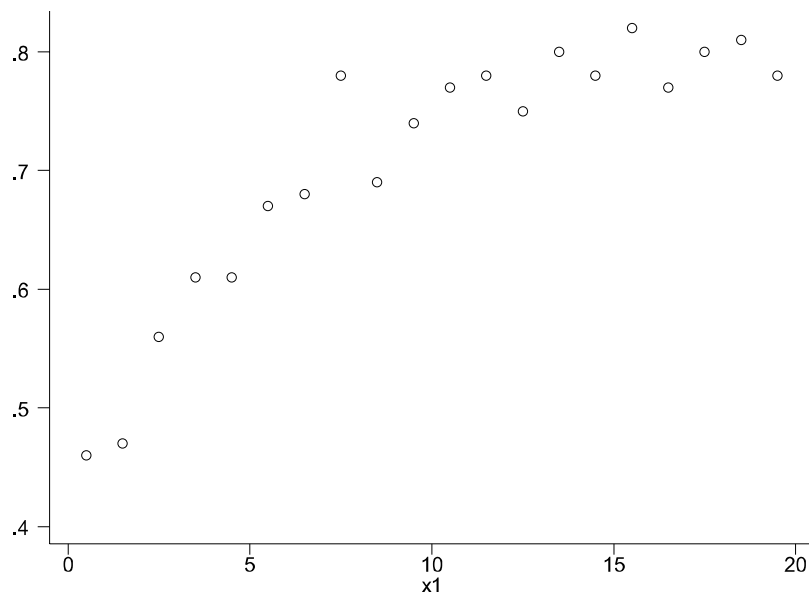


Figure 9: Note the curvilinear relationship between Y and X_1 .

$X_1 = 7$. To do this, we first create a dummy variable z such that

$$z = \begin{cases} 1 & \text{if } X_1 \geq 7 \\ 0 & \text{if } X_1 < 7 \end{cases}$$

To define the separate slopes, we create an interaction term between z and X_1 , which leads to the following model:

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 z + \hat{b}_3 (z X_1),$$

which in turn gives rise to the following submodels that will produce for us, the segmented slopes:

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1,$$

for $X_1 < 7$, and

$$\hat{Y} = \hat{a} + \hat{b}_2 z + (\hat{b}_1 + \hat{b}_3) X_1,$$

for $X_1 \geq 7$. When I estimate this model, I get the following estimates:

$$\hat{Y} = .441 + .040 X_1 + .263 z + -.034 (z X_1),$$

where the $RMSE = .027$ and $F = 102.09$. Each of the coefficients have t ratios far greater than 2. It is easy to see graphically how this model produces two slopes and I do this in Figure 11. Clearly when compared to the model shown in Figure 5, this model provides a superior fit. The reason why is easy: the data exhibit some curvilinearity. The standard least squares model cannot account for it; this model explicitly accounts for the curvilinear

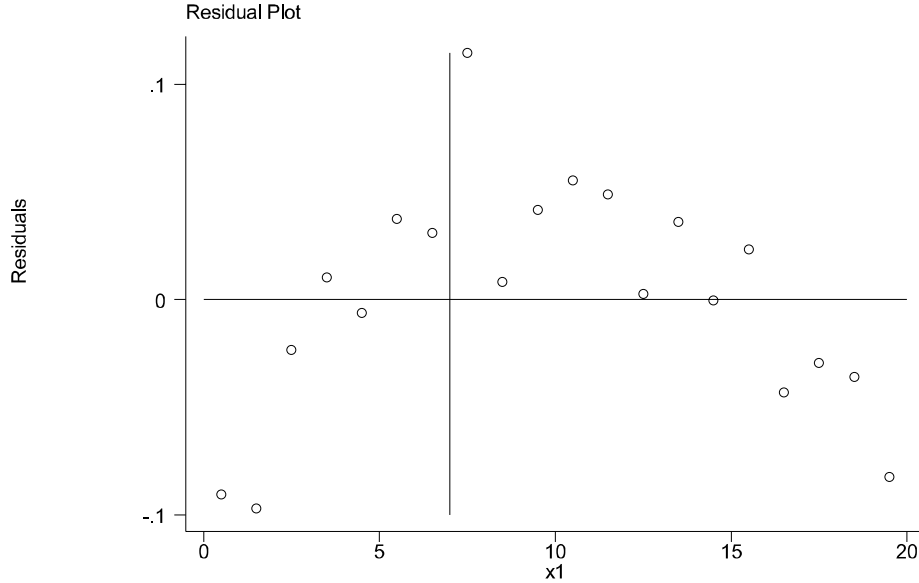


Figure 10: *Plot of residuals versus values of X_1 with a vertical reference line at 7 and a horizontal reference line at 0.*

nature of the data. Further, it should be clear how the use of an interaction term can be used to capture segmented slopes.

There is one interesting feature shown in Figure 11. If you look at the end point for the first slope and the beginning point of the second slope, you will see a discontinuity: the slopes do not match up. More formally, the difference in the ordinates at $X_1 = 7$ represent a discontinuity in the data. Specifically, this continuity is given by

$$\text{discontinuity at } X_1 = 7 = \hat{b}_2 - 7(\hat{b}_3).$$

This result holds because \hat{b}_3 is only defined for $X_1 = 7$ and above. We could constrain the fit to make this difference 0 by imposing the restriction that

$$\hat{b}_2 = -7(\hat{b}_3).$$

Constraining the model to satisfy this constraint is achieved by imposing this condition on the model, which gives rise to

$$\hat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_3 (z * (X_1 - 7)).$$

Note what we have done. The last term in this model is zero for $X_1 < 7$ (why?) and is $X_1 - 7$ if $X_1 > 7$. (Why? Because in this case $z = 1$.) Fitting this model would force the two slopes to meet at $X_1 = 7$.¹² I estimate this model and obtain the following results:

$$\hat{Y} = .434 + .043X_1 + -.037(z * (X_1 - 7)),$$

¹²To create this term, I typed `gen newz=z*(x1-7)` in Stata.

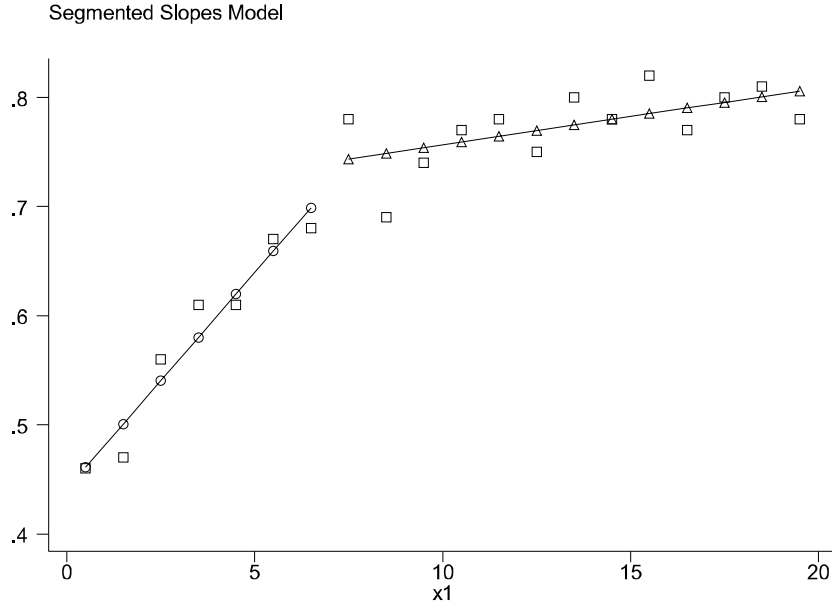


Figure 11: *Segmented Slope Model with separate slopes for $X_1 < 7$ and $X_1 > 7$. Note discontinuity at $X_1 \geq 7$.*

where the $RMSE = .026$ and $F = 155$. To graphically show how the slopes are segmented, I present this in Figure 12. It is easy to see that the two slopes meet at the ordinate at $X_1 = 7$.

All of the preceding is directly extendable to multiple segments. If we wanted to estimate, say, three slope segments, we could do this through the use of two dummy variables interacted with X_1 . All one would need to do is to segment the data into n ranges and create $n - 1$ indicator variables. I will not go into an example of this in these notes, as it is a straightforward extension of the above discussion.

4 Mathematical Transformations X

In the previous section, we saw how we can take our ideas about interactions and apply them in a setting to account for nonlinear and curvilinear relationships between Y and X . Although I think the approach outlined above is sensible in many applications, “purists” may find the exercise tantamount to curve fitting. This can be true, if one blindly tries to fit slopes to data. Indeed, in the extreme case, one can estimate a model that perfectly fits the data by estimating a saturated model: one where every data point has its own parameter! In this section, I want to consider alternative approaches to transforming X . I will also consider instances where transformations on Y may be appropriate.

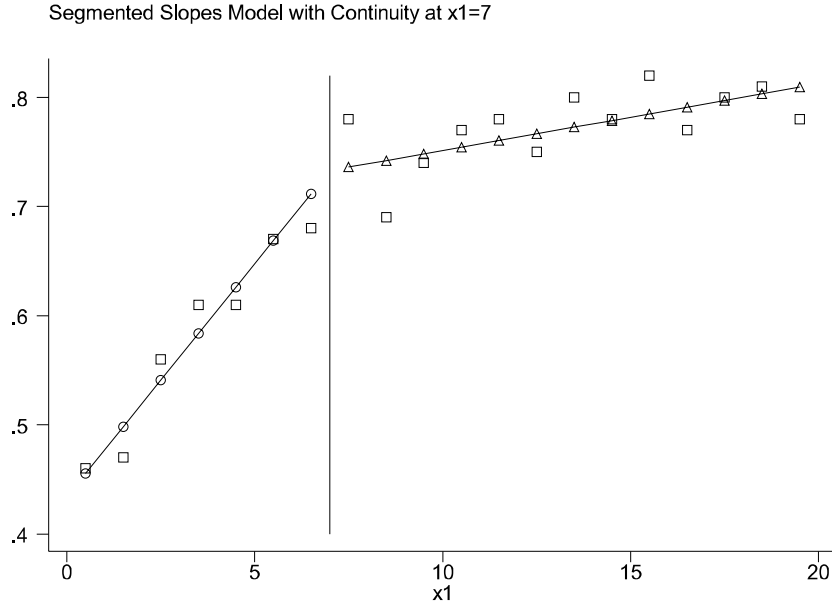


Figure 12: *Segmented Slope Model with separate slopes for $X_1 < 7$ and $X_1 > 7$. Note the continuity at $X_1 \geq 7$.*

4.1 Monotonicity with Saturation Effects

In this section, let us consider the problem posed by data that seem to exhibit saturation effects, or marginality. We saw this earlier in the top two panels of Figure 6 where the relationship between X_1 and Y changed as X increased in value. The basic problem with marginality is the slope of the relationship between Y and X_1 is not constant across the full range of X_1 . This gives rise to a curvilinear relationship, not a linear relationship. The problem is, a standard regression model will produce a straight line (in the bivariate case) when this relationship may be inappropriate.

For problems where the relationship between some covariate and the dependent variable exhibits a nonconstant slope but the relationship is believed to be monotonic, there are some standard transformations one may want to apply to X_1 to capture the nonlinearity in the relationship.

Transforming X_1 by taking the natural log is one common way to proceed when the relationship between X_1 and Y seems to exhibit marginality. The log transformation works because, in a sense, it compresses the x axis. Recall that when taking the log of a variable, the “distance” between adjacent values of the logged variable decreases as the values of the unlogged variable increase. Estimating a regression model with a logged covariate presents no challenges. We simply estimate

$$\hat{Y} = \hat{a} + \hat{b}_1 \log(X_1),$$

where the independent variable is transformed by the natural log. It is easy to see that the “compression” in X_1 is a desirable property when saturation effects are present because by

logging X_1 , we are multiplying by \hat{b}_1 , a number that is increasing in values at a decreasing rate. This is precisely the transformation we are after.

To illustrate, let us return to the models considered previously in Figure 7. Let's focus on the top two panels. For the data shown in the first panel, the relationship seems to exhibit clear marginal changes in $E(Y)$ given increases in X_1 . While the regression model yields an adequate fit (recall Table 2), the question is, can we do better by accounting for the curvilinear relationship. To check, I took the log of X_1 .¹³ Using the transformed variable, I reestimated the regression model and got the results shown in the second column of Table 3.

Table 3: Regression Models Based on Data from Figure 6 with Transformations.

Coefficient	Standard Model 1	With log X_1	Standard Model 2	With log X_1
Intercept	.754 (.019)	.481 (.018)	.616 (.025)	.482 (.018)
Slope	-.193 (.038)	-.112 (.008)	.0007 (.0001)	.056 (.004)
<i>RMSE</i>	.074	.033	.076	.033
r^2	.58	.91	.56	.91

It is useful to compare the results to the “standard” model containing the untransformed X_1 . The *RMSE* is considerably smaller and the r^2 is considerably larger in the model with log X_1 . This is to be expected. The data suggested a curvilinear relationship between X_1 and Y . The standard model does not account for this; the new model does.

I repeat this exercise for the second panel of data shown in Figure 6. The coefficient estimates are given in the last column of Table 3. Compared to the standard model, we again see the model with log X_1 is far superior to the model with untransformed X_1 . Again, this is to be expected, given the distribution of the data shown in Figure 6.

To compare the scatterplots and predicted regression functions from the standard and transformed models, consider Figure 13. The top two panels graph the results from the first two columns in Table 3; the bottom two panels graph the results from the last two columns in Table 3. The graphs are consistent with the results: the transformed variables provide a considerably better fit.

Of course one must take caution in interpreting these results. While the least squares solution works with no special problems for these models, you need to remember that the parameter estimate for the transformed model is based on log X_1 , *not* the untransformed X_1 . When computing predicted values, you are multiplying the coefficient by the transformed variable. In order to make substantive sense of the coefficient, you have to be careful to “backtransform” the data into units that are interpretable to your reader. The correct interpretation of the regression model with transformed X_1 is that *for a unit increase in log X_1 , the expected change in Y is given by the parameter estimate \hat{b}_1 .*

¹³In *Stata*, this entailed the command: `gen logx1=log(x1)`.

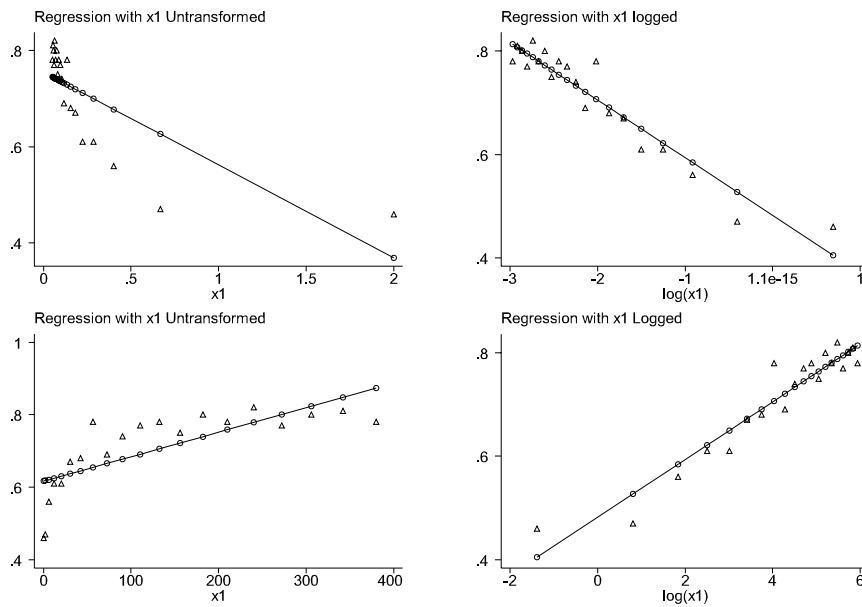


Figure 13: Comparison of untransformed X_1 and $\log X_1$ for data from Figure 6. Note the models with $\log X_1$ provide a superior fit, as they account for the curvilinear relationship between Y and X_1 .

To visually see the implications of transforming X_1 , consider Figure 14. Here, I've graphed the regression function from the second column of Table 3, but have set the x -axis to be equal to the untransformed X_1 . Clearly the relationship shown is curvilinear because the regression function is based on $\log X_1$, and not X_1 . For the regression model, the parameter is linear, but it is based on the log of X_1 . Since this transformation compresses X_1 , the per unit change in $\log(X_1)$ is not constant; hence, the slope is not constant across the range of X_1 . This is illustrated in Figure 14.

Apart from the log transformation, another transformation to account for marginality is the square root transformation. This transformation also compresses the x -axis and so the logic of the transformation is similar to that of the log transformation. Either are suitable, and so one may want to compare F statistics or r^2 measures to compare models using the different transformations. One thing to remember is that the $\log X_1 \leq 0$ is undefined and the square root of negative numbers is also undefined. If you have many 0s or if you have negative numbers, one or both of these transformations will be unsuitable (and be careful, because if these conditions exist, **Stata** will still compute the transformation, but just for permissible observations. This could result in a staggering loss of data!

When the log or square root transformations do not apply, I recommend a segmented slopes approach. Another alternative is to rescale your data to put it in the permissible range for these transformations. This may or may not be a feasible thing to do in all applications.

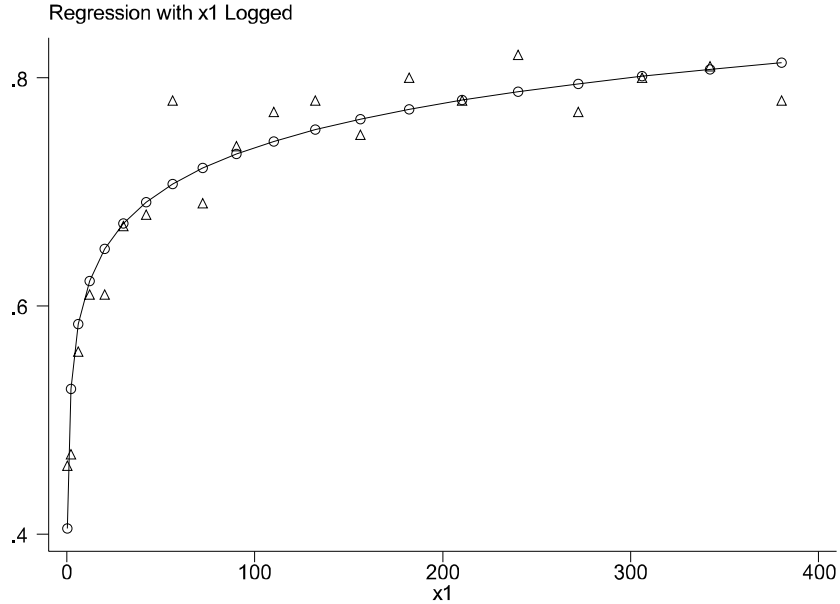


Figure 14: *Regression Function from Transformed Model Graphed in Terms of Untransformed X_1 .*

5 Polynomial Regression

The transformations discussed in the previous section are when accounting for a nonlinear, but monotonic, relationship. Suppose, however, that one’s data exhibit nonmonotonicity. For example, imagine the relationship between Y and X_1 is u-shaped (or inverted u-shaped). Recall the data from the bottom two panels of Figure 7. These data show clear nonmonotonicity and so it is obvious that if one tries to “stick” a straight line through the cloud of points, the least squares solution will provide a slope that is exactly 0. The question naturally arises as to how to account for nonlinearity of this form in the context of a regression model.

A common approach to handling this is through polynomial regression. Polynomials are an algebraic expression of the form

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_3 x^3 + a_2 x^2 + a_1 x + a_0,$$

where a_0, a_1, \dots, a_n are constants that are the coefficient of the polynomial, and n is a positive integer. The degree of the polynomial is the highest power of the variable that appears. Graphs of polynomial functions are useful to understand because they illustrate the point that the curve of the function can change directions. The number of “turning points” in a polynomial is odd if the degree of the polynomial is even, and vice versa. A polynomial of degree 1 has no turning points and so it produces a straight line (why?). A polynomial of degree 2 has 1 turning point and is known as a quadratic. A polynomial of degree 3 has 0 or 2 turning points and is known as a cubic.

The linkage of these mathematical ideas to regression is simple: if one believes the relationship between Y and X exhibits a relationship where the slope changes direction, this could be captured by a model incorporating polynomials. I'm going to focus on the quadratic model, which is widely applied in political science settings, though it is often incorrectly interpreted.

To fix ideas, suppose we have data that look roughly like that shown in Figure 7. Clearly, estimation of

$$Y = \hat{a} + \hat{b}X_1$$

will be unsatisfactory. However, suppose we estimate

$$Y = \hat{a} + \hat{b}X_1 + \hat{b}X_1^2,$$

where X_1^2 is X_1 squared.¹⁴ This regression model is a polynomial model with degree 2 (why?). It is also known as a quadratic model. Estimating the model in this form will produce a response function (or a curve) that will have a single bend in it. Please understand that the bend *may or may not* be observed in the range of data with which you are working. Nevertheless, at some point, the response function will inflect, or change directions. At the point at which the response function changes direction, the curve has a horizontal tangent. In regression terms, *at the point where the curve changes direction, the slope (partial or otherwise) between Y and X_1 is exactly 0*. This point is critical to understand because it demonstrates quite clearly that the slope of the relationship is not constant with respect to X_1 ; rather, it is *conditional* on X_1 .

We have seen conditional slopes before. When we learned about interaction models, we learned that the slope of, say, Y on X_1 was conditional on X_2 in the model

$$Y = \hat{a} + \hat{b}_1X_1 + \hat{b}_2X_2 + \hat{b}_3X_1X_2.$$

It should be obvious that a polynomial model is equivalent to an “interactive” model. This is easy to see if we rewrite the quadratic model as

$$Y = \hat{a} + \hat{b}X_1 + \hat{b}X_1X_1,$$

where the last term is (obviously) equivalent to X_1^2 . In a quadratic model, we're essentially interacting X_1 with itself. Hence, the slope of Y on X_1 will be conditional dependent upon the value of X_1 . In general, the rate of change of Y with respect to X is given by

$$\frac{\partial Y}{\partial X_1} = \hat{b}_1 + 2\hat{b}_2X_1.$$

This partial derivative illustrates that the rate of change is conditional on X_1 and that the rate will vary as X_1 changes. The slope is not constant. Hence, the conditional slope is given by the right-hand side term in the above expression (in the words of Friedrich, this is the metric effect).

¹⁴In Stata this variable is created by `gen logx12=log(x1)`.

To illustrate, suppose we reexamine the data from Figure 7 in the lower left panel. In applying the quadratic model from above to these data, we obtain the following estimates:

$$\hat{Y} = .428 + .056X_1 - .003X_1^2.$$

The r^2 for these data is .95 and the $RMSE$ is .02. (Compare these results to the linear regression results). Note the signs on the coefficients. The positive sign on the constituent term (i.e. \hat{b}_1) and the negative sign on the squared term (i.e. \hat{b}_2) indicates the curve is first increasing to some point and then decreasing. That is, this combination of signed coefficients will produce an inverted u-shaped function.

In Figure 15, I graph the predicted regression function and the actual values of Y .¹⁵ It is easy to visualize the superior fit of the quadratic model, than when compared to the fit of the standard linear model.

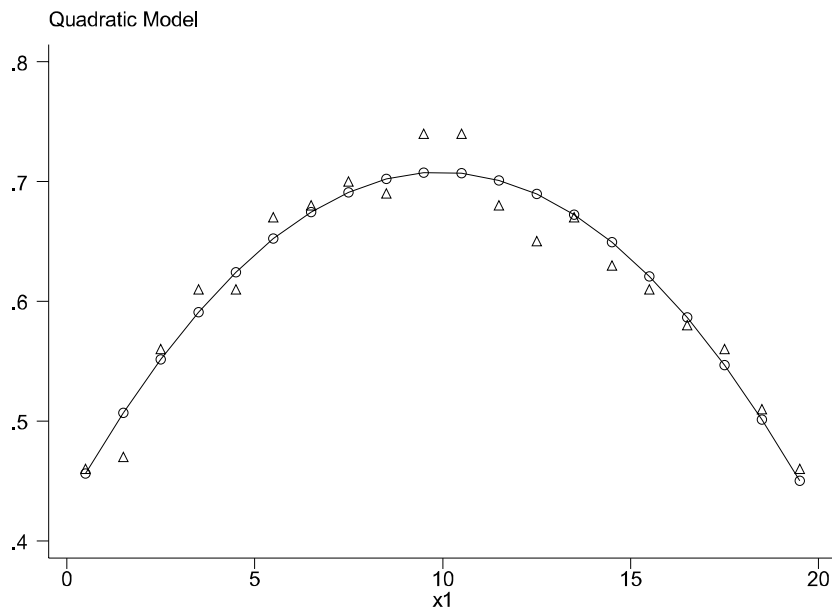


Figure 15: *Regression Function from Quadratic Model.*

Of course just like the interactive models, since the slope is conditional on X_1 (which is obviously shown in Figure 15), so too will be the standard errors. This implies that some points on the quadratic response function may be no different from 0. So even if the quadratic “holds” based on inspection of the standard errors given by the default **Stata** output, it need not be the case that entirety of the quadratic function is statistically significant (this is just like the interaction model setting). Hence, we need to compute the standard error of the conditional slope

$$\hat{b}X_1 + \hat{b}X_1^2.$$

¹⁵In the data set `nonlin.dta`, the dependent variable used in this analysis is called `ynew` and the quadratic term is given by `x1` and `x12`. See the previous footnote on how to create the squared term. The graph was then generated by first using the `predict` function to get predicted values (that is, I typed: `predict xb`) and then I graphed them using the command: `gr xb ynew x1, ylab xlab t1("Quadratic Model")`.

. Using standard results from the variance and covariance of two random variables, the variance of the conditional slope is given by

$$\text{var}(\hat{b}X_1 + \hat{b}X_1^2) = \text{var}(\hat{b}_1) + 4X_1^2\text{var}(\hat{b}_2) + 4X_1\text{cov}(\hat{b}_1\hat{b}_2),$$

and the standard error is given by the square root of this term. To illustrate, I computed the conditional slopes for the regression model. Now I want to compute the standard errors for these conditional slopes. To do this, I need to back out the variance-covariance matrix for the regression model. I obtain this by typing `vce` after estimating the regression. Using the results from the variance-covariance matrix and the formula given above, I derived the standard errors.¹⁶ With the standard errors in hand, I can compute the t ratios for each conditional slope and then determine which points, if any, on the quadratic term are no different from 0.¹⁷ As in the case of the interaction model, sometimes there will be many dozens (or even hundreds) of possible conditional slopes (it all depends on the sample size and the range of unique X_1). Because of this, one way to easily determine the points that are not statistically significant is through graphing them. In Figure 16 I graph the t -ratios. The first panel of the graph plots the t -ratios in terms of their actual values; the second panel of the graph plots them in terms of their absolute values (q: why is it okay for me to convert the t -ratios to absolute values?).¹⁸ It is easy to see through these graphs that each of the points on the quadratic term are significantly different from 0.

The last thing I want to talk about with regard to quadratics involves the inflection point. As noted above, the inflection point may or may not be found in the observed data. Formally, the inflection point from a quadratic model is given by

$$\frac{\partial Y}{\partial X_1} = 0.$$

To compute this directly, note that this expression is equivalent to

$$\frac{-\hat{b}_1}{2\hat{b}_2}.$$

Substituting the parameter estimates into this expression, the quotient is 9.94. That is, when $X_1 = 9.44$, the slope of Y on X_1 is exactly 0. The predicted value of Y at this point is about .708. Graphically, we can verify this by inspecting Figure 17. Here, I simply replot the predicted regression function and draw a vertical reference line at 9.44 and a horizontal

¹⁶The `Stata` code used to generate the standard errors is: `gen secs=sqrt(.000011 + (4*x12*.00000026) + (4*x1*-.00000051))` where `secs` is a new variable (which denotes “standard error of conditional slopes”).

¹⁷To compute the t -ratio, I typed: `ge tcs=cs/secs`.

¹⁸To generate the first panel, I typed: `gr t x1, yline(2.11,-2.11) ylab xlabel t1("Plot 1: t-ratios (actual values)") l1(" ") saving(trat1)`. To generate the second graph, I first had to create the absolute value of the t ratios. This was done by: `gen tabs=abs(t)`. Then I graphed it by typing: `gr tabs x1, yline(2.11) ylab xlabel t1("Plot 2: t-ratios (absolute values)") l1(" ") saving(trat2)`. Note that the `saving` option used in these commands tells `Stata` to automatically save the file as a `.gph` file. To combine the graphs, I typed: `gr using trat1 trat2`. The horizontal reference line intersects the y -axis at 2.11 (and -2.11 for the first panel). (Q: Why did I do this?).

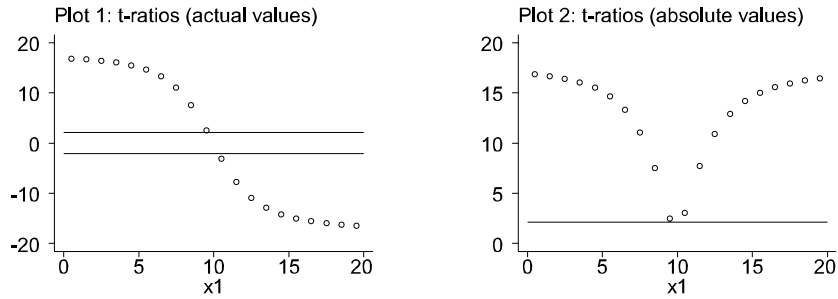


Figure 16: *Estimated t-ratios from Quadratic Model. The top panel plots the actual t-ratios; the second panel plots the absolute values of the t-ratios.*

line at .708. Where these two lines intersects corresponds to the inflection point. Note that the horizontal reference line is a tangent to the quadratic. This is the case because the slope at this point is 0.¹⁹

6 Conclusion

This concludes functional form (for now). What I want you to understand is the implications of ignoring proper functional form. If your data exhibit nonlinearity, then a model accounting for this will almost always be statistically better than a model that doesn't account for this. Further, if theory leads you to believe the relationship between Y and some X exhibits marginality, nonmonotonicity, or some other kind of functional relationship that is other than constant, linear, and additive, then you need to think hard about the inclusion of transformations on X (or inclusion of interaction terms).

¹⁹This figure was obtained by typing: `gr xb ynew x1, c(1) ylab xlab t1(Quadratic Model with Inflection Point Noted) xline(9.44) yline(.708).`

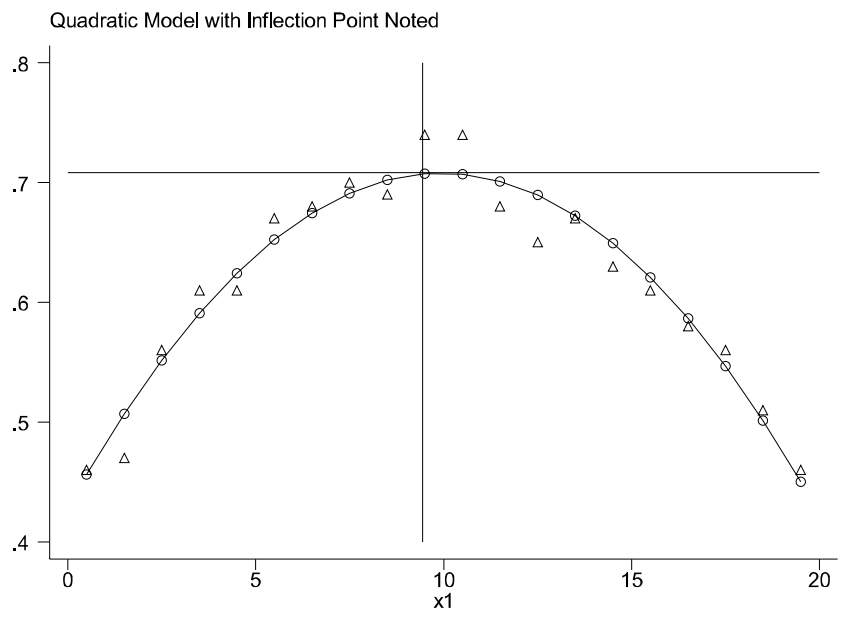


Figure 17: *Regression function with inflection points indicated.*