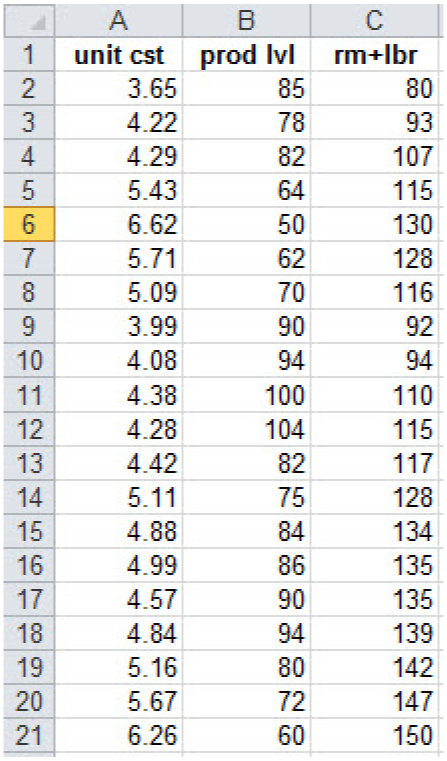
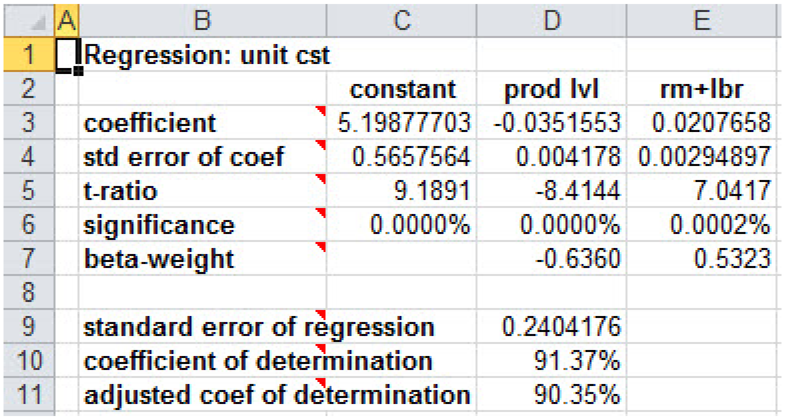
A Structural Modeling "Trick" - Nonlinearities

At times, experience or judgment will suggest that the linkage between some explanatory variable and the dependent variable in a regression model is not linear.

> *Example:*A company has collected data on one of its factories over the past 20 fiscal quarters. For each quarter, they've divided total operating expenses by the number of (standardized) units of output produced, in order to determine their per-unit cost of production. In order to try to understand why this cost has varied, they look at two potential explanatory variables: prod lvl = the level of scheduled output, measured in percentage points of the maximum designed output level of the factory, and rm+lbr = a composite index tracking the market price of raw materials and the hourly cost of direct labor. Their sample data, and the results of an initial regression are:



	A	B	C
1	**unit cst**	**prod lvl**	**rm+lbr**
2	3.65	85	80
3	4.22	78	93
4	4.29	82	107
5	5.43	64	115
6	6.62	50	130
7	5.71	62	128
8	5.09	70	116
9	3.99	90	92
10	4.08	94	94
11	4.38	100	110
12	4.28	104	115
13	4.42	82	117
14	5.11	75	128
15	4.88	84	134
16	4.99	86	135
17	4.57	90	135
18	4.84	94	139
19	5.16	80	142
20	5.67	72	147
21	6.26	60	150

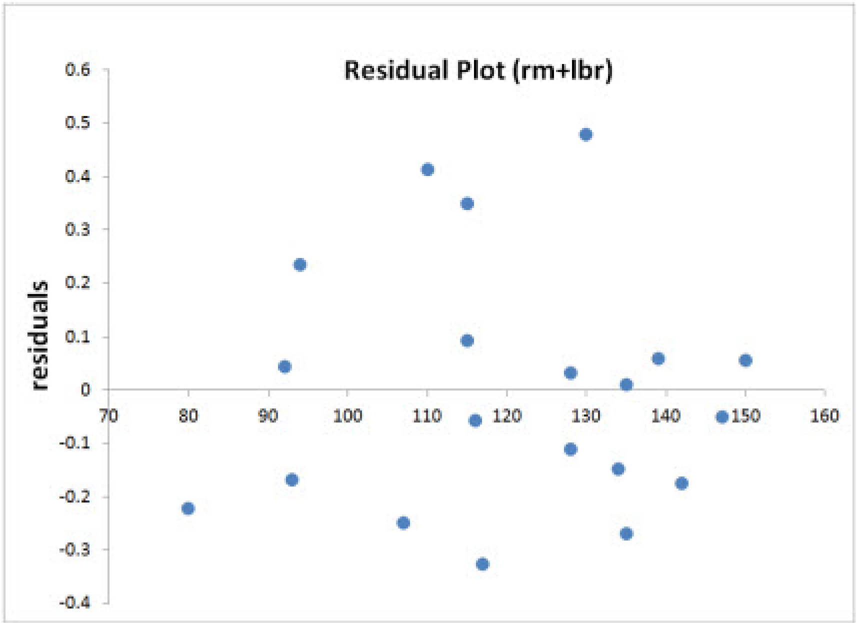


	A	B	C	D	E
1		Regression: unit cst			
2			constant	prod lvl	rm+lbr
3		coefficient	5.19877703	-0.0351553	0.0207658
4		std error of coef	0.5657564	0.004178	0.00294897
5		t-ratio	9.1891	-8.4144	7.0417
6		significance	0.0000%	0.0000%	0.0002%
7		beta-weight		-0.6360	0.5323
8					
9		standard error of regression		0.2404176	
10		coefficient of determination		91.37%	
11		adjusted coef of determination		90.35%	

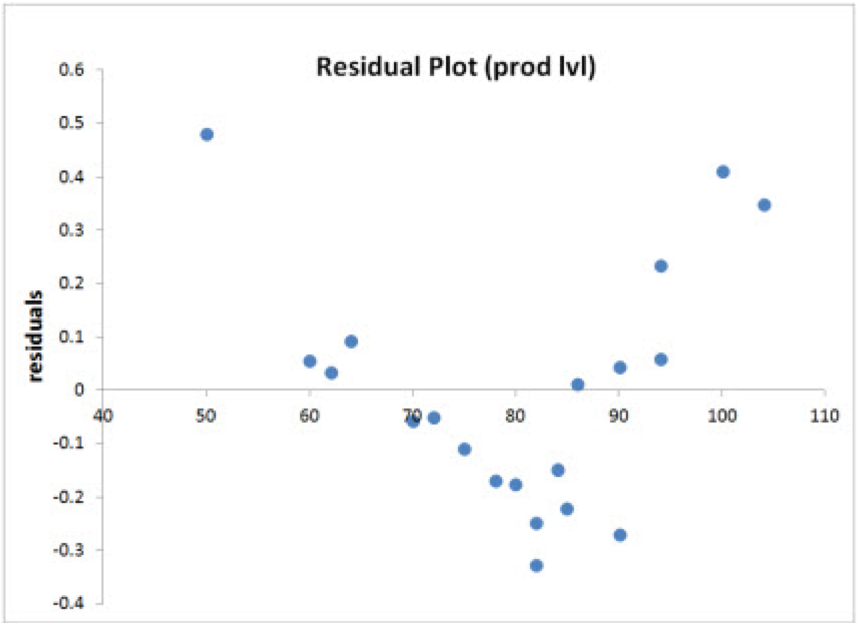
However, the notion of economies of scale would lead us to expect that the linkage between production level and per-unit manufacturing cost is nonlinear. Combined with a bit of operational insight, we'd expect the per-unit cost to drop, rapidly at first and then more slowly, as the production level increases, and then to rise again as production level is pushed beyond what the factory was designed to handle.

A graphical examination of how the residuals from a regression vary with the magnitude of an explanatory variable can reveal a nonlinear linkage as well. In a truly linear relationship, the residuals take both positive and negative values for every range of values of the explanatory variables. A residual plot which shows the sign of the residuals varying systematically with the values of some explanatory variable indicates the presence of a nonlinear relationship between that explanatory variable and the dependent variable.

Example: Plotting the residuals against the raw-material-and-labor index reveals nothing of interest.



However, a plot of the residuals against production levels reveals a definite pattern:



For production levels below 70 and above 90, the residuals are almost all positive (indicating that the model systematically underpredicts the dependent variable in these cases). In-between, the residuals are just about all negative (indicating that the model overpredicts in those cases). Obviously, we could improve the model by adjusting predictions upwards when production level is high or low, and adjusting them downwards when production level is moderate.

When a residual plot shows a rough "U"-shaped link (either direct or inverted) between the residuals and an explanatory variable, the fit of the model to the data can be improved by introducing the square of that explanatory variable as a new artificial variable in the model. (Here is a workbook which reviews some of the properties of quadratic functions and their graphs (parabolas).

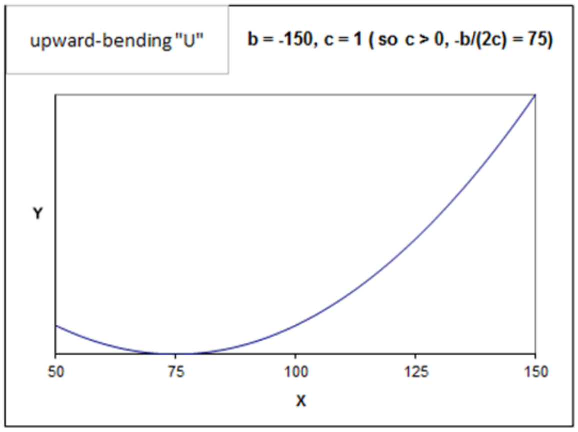
Capturing Nonlinearity

In order to capture a curved (i.e., nonlinear) relationship, we can include the square of an explanatory variable (together with the original variable) in our regression model. The model now takes the form: $Y = \ldots + bX + cX^2 + \ldots$.

The values of the two coefficients together determine the shape of the curve. If $c > 0$, the curve bends upwards (and the magnitude of c determines the "tightness" of the bend); if $c < 0$, the curve bends downwards.

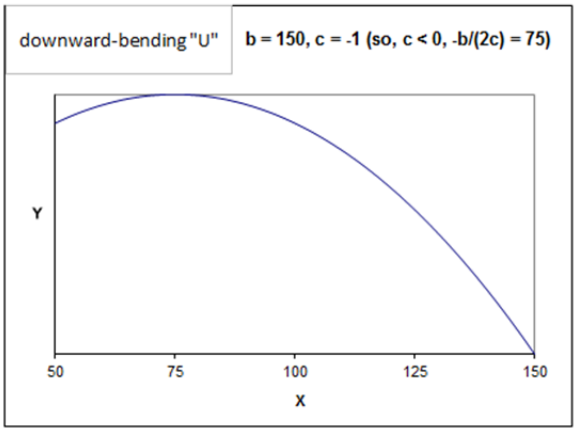
The curve will be a portion of a parabola which takes its lowest ($c > 0$) or highest ($c < 0$) value when $X = -b/(2c)$. If this value of X is within the range of observed values of X, we'll actually see some of both "wings" of the parabola in the captured relationship. However, if this value is outside our observed data range, only part of one wing will actually appear.

In the examples below, the range of observed values of X isassumed to be between 50 and 150.



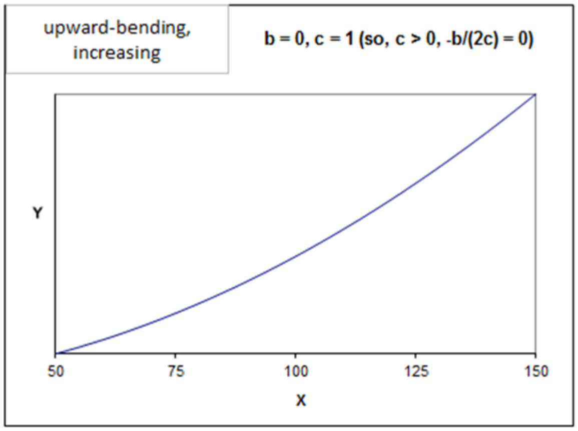
upward-bending "U"

b = -150, c = 1 (so c > 0, -b/(2c) = 75)



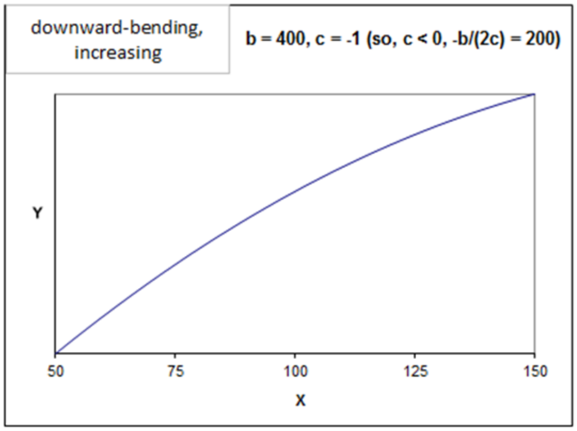
downward-bending "U"

b = 150, c = -1 (so, c < 0, -b/(2c) = 75)



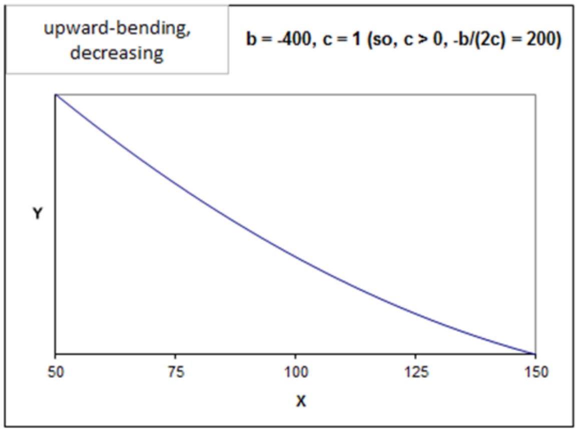
upward-bending, increasing

b = 0, c = 1 (so, c > 0, -b/(2c) = 0)



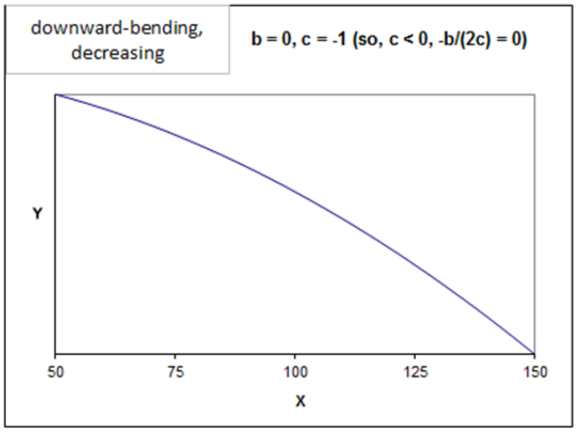
downward-bending, increasing

b = 400, c = -1 (so, c < 0, -b/(2c) = 200)



upward-bending, decreasing

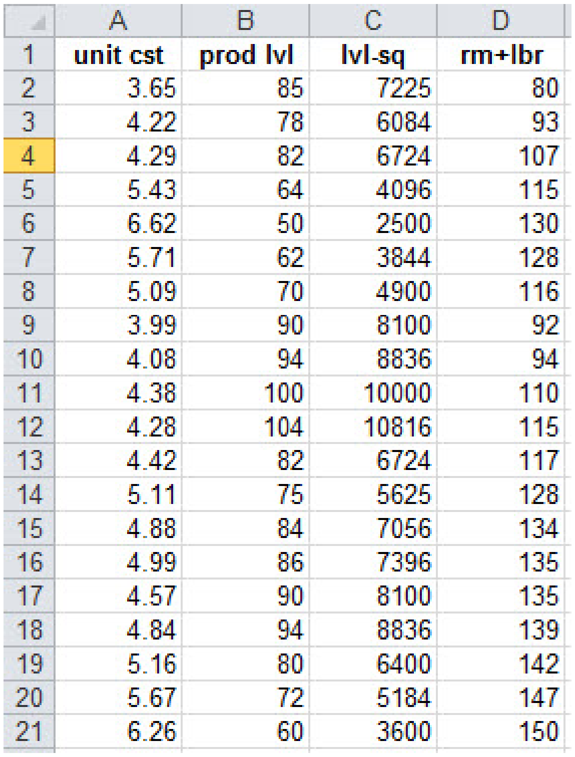
b = -400, c = 1 (so, c > 0, -b/(2c) = 200)



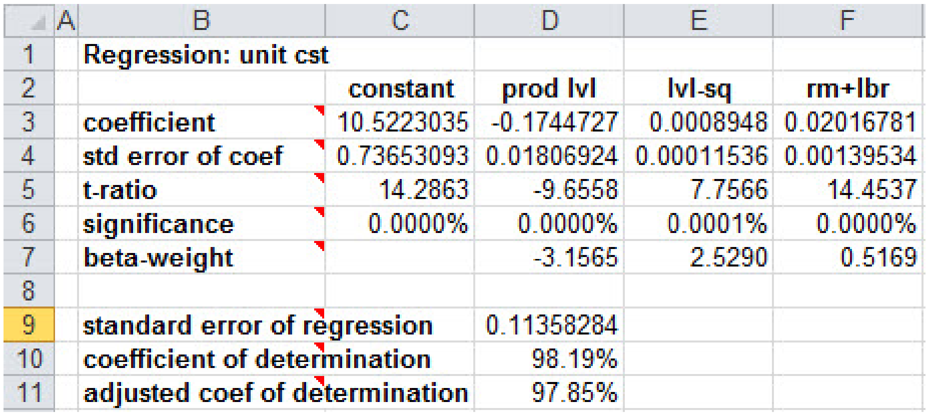
downward-bending, decreasing

b = 0, c = -1 (so, c < 0, -b/(2c) = 0)

Example: Introducing the square of production level as a new explanatory variable in the model yields the regression results:



	A	B	C	D
1	unit cst	prod lvl	lvl-sq	rm+lbr
2	3.65	85	7225	80
3	4.22	78	6084	93
4	4.29	82	6724	107
5	5.43	64	4096	115
6	6.62	50	2500	130
7	5.71	62	3844	128
8	5.09	70	4900	116
9	3.99	90	8100	92
10	4.08	94	8836	94
11	4.38	100	10000	110
12	4.28	104	10816	115
13	4.42	82	6724	117
14	5.11	75	5625	128
15	4.88	84	7056	134
16	4.99	86	7396	135
17	4.57	90	8100	135
18	4.84	94	8836	139
19	5.16	80	6400	142
20	5.67	72	5184	147
21	6.26	60	3600	150



	A	B	C	D	E	F
1		Regression: unit cst				
2			constant	prod lvl	lvl-sq	rm+lbr
3		coefficient	10.5223035	-0.1744727	0.0008948	0.02016781
4		std error of coef	0.73653093	0.01806924	0.00011536	0.00139534
5		t-ratio	14.2863	-9.6558	7.7566	14.4537
6		significance	0.0000%	0.0000%	0.0001%	0.0000%
7		beta-weight		-3.1565	2.5290	0.5169
8						
9		standard error of regression		0.11358284		
10		coefficient of determination		98.19%		
11		adjusted coef of determination		97.85%		

$$(\text{unit cst})_{\text{pred}} = 10.5223035 - 0.1744727\cdot(\text{prod lvl}) + 0.0008948\cdot(\text{prod lvl})^2 + 0.02016781\cdot(\text{rm+lbr})$$

The significance level of 0.0001% for the squared term indicates strong evidence that it has a true non-zero coefficient, and therefore belongs in the model. The coefficient 0.0008948 is positive, indicating that the nonlinear link between per-unit cost and production level is upward-bending. This upward-bending relationship bottoms out when production level is -b/(2c) = -(-0.1744727)/(2·0.0008948) = 97.492 (i.e., near 100, as expected from our initial intuition).

Other, similar approaches are available to try to capture nonlinear relationships. Introducing the reciprocal of an existing variable as a new variable can help to capture asymmetric “U”-shaped relationships, and introducing the square root of an existing variable can help to capture “sideways” “U”-shapes. However, in the study of relationships encountered in business settings, some combination of the simple interaction and “U”-shape “tricks” presented here can capture most nonlinearities. Indeed, many commercially-available regression-analysis packages contain specific features to facilitate the construction of new variables (i.e., new columns of data) from products of existing variables, or the squares of existing variables.