Excel for Algebra [1]

Lesson 7: Curve Fitting With The Solver

What does "Curve Fitting" mean?

The essence of "curve fitting" is to take some set of discrete data points, (x_i,y_i), $i=1..N$, model the points as a smooth curve plus some error on each point, $y_i = f(x_i)+err_i$, and adjust the curve to make the errors as small as possible.

Typically this is done by picking some function that has whatever shape you want the curve to be, and adjusting the parameters of that function to minimize the errors.

For example, you may decide that a straight line is the shape you want. Then your function is $f(x) = mx+b$, and you adjust m and b (slope and y-intercept).

When Excel draws a trendline, it's really fitting a curve that has one of a few standard shapes. If you need to fit a curve that is one of Excel's standard trendlines, then that's the easy way to do it.

However, there are a lot of useful shapes that Microsoft has not programmed into Excel.

No problem – you define the shape yourself using ordinary functions in the spreadsheet, and find the best parameters by using "The Solver" that is built into Excel.

An example: Fitting a bell curve to oil production data

In the spirit of Goldilocks, let's see if we can find a "just right" example — not too simple, not too hard. Maybe even one we've seen before. How about Bartlett's example of fitting a Hubbert curve to U.S. domestic oil production?

The Hubbert curve is very similar to a standard bell curve. It approaches zero on both ends, reaches a peak in the middle, is flat on top, and falls off symmetrically to both sides. Mathematically, the standard bell curve can be described by:

$$f(x) = Me^{-((x-C)/W)^2}$$

That is, for any value of x (the year), you're supposed to subtract C (the center year), divide by W (the width of the curve), square the result, raise e to minus that power, and multiply by M (the maximum value of the curve, which will occur at year C).

The parameters of this curve are C, W, and M.

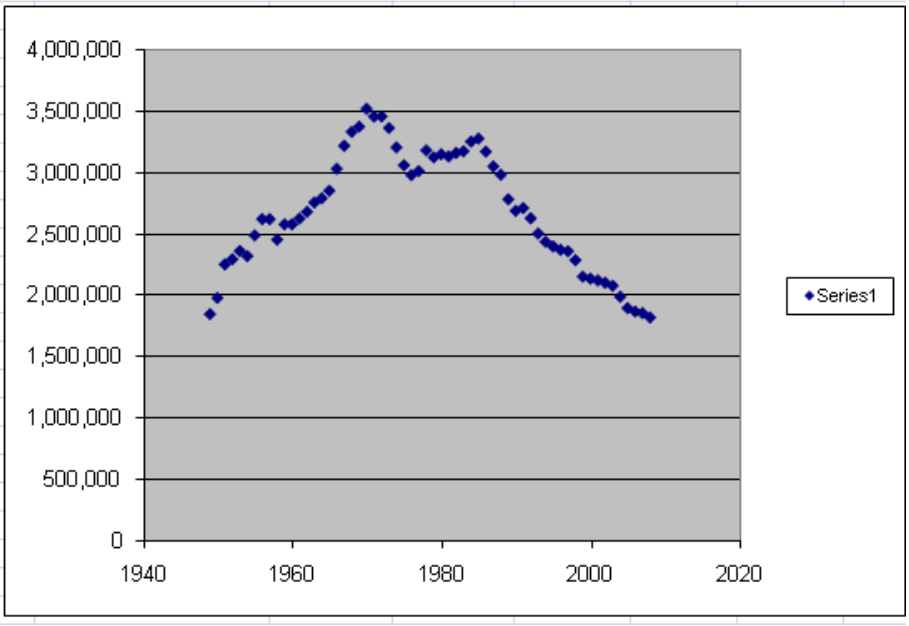
Setting up a spreadsheet

Data for U.S. domestic oil production can be found at http://www.eia.doe.gov/emeu/aer/petro.html (the U.S. Energy Information Administration). They have been gracious enough to provide a spreadsheet already populated with U.S. domestic oil production (http://www.eia.doe.gov/emeu/aer/txt/stb0501.xls)

Following standard practice, we'll fit parameters so as to minimize the sum of the squared errors.

First, let's just plot the data:

	A	B	C	D	E	F	G	H	I	J
1	Data Source: U.S. Dept. of Energy, Energy Information Administration, U.S. Domestic Oil Production									
2	http://www.eia.doe.gov/emeu/aer/txt/stb0501.xls									
3										
4	Year	Crude Oil, Total, 1000's of barrels								
5										
6										
7										
8										
9	1949	1,841,940								
10	1950	1,973,574								
11	1951	2,247,711								
12	1952	2,289,836								
13	1953	2,357,082								
14	1954	2,314,988								
15	1955	2,484,428								
16	1956	2,617,283								
17	1957	2,616,901								
18	1958	2,448,987								
19	1959	2,574,590								
20	1960	2,574,933								
21	1961	2,621,758								
22	1962	2,676,189								
23	1963	2,752,723								
24	1964	2,786,822								
25	1965	2,848,514								
26	1966	3,027,763								
27	1967	3,215,742								
28	1968	3,329,042								
29	1969	3,371,751								
30	1970	3,517,450								

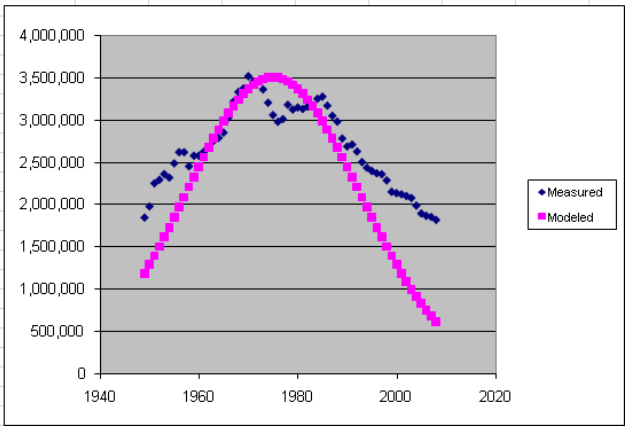


Just from looking at the data, we can see that it reaches a peak of about 3.5 million, is roughly centered on 1975, and reaches half of its peak value at about ±25 years. Those observations suggest parameter values near C=1975, W=25, and M=3.5 million.

Now let's define our model, using those values as a starting point. Here's what the spreadsheet looks like after we first put the model in.

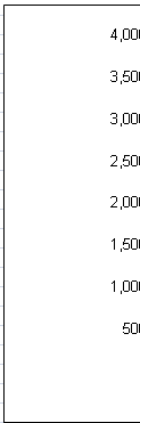
(Values)

	A	B	C	D	E	F	G	H
1	Data Source: U.S. Dept. of Energy, Energy Information Administration, U.S. Domestic Oil Production							
2	http://www.eia.doe.gov/emeu/aer/txt/stb0501.xls							
3								
4	Year	Crude Oil, Total, 1000's of barrels		Model				
5				C (center year)	W (width in years)	M (maximum height)	Sum Error Squared	
6				1975	25	3.5	1.92E+13	
7								
8				Model Value	Error	Error Squared		
9	1949	1,841,940	1,949	1.19E+06	6.55E+05	4.29E+11		
10	1950	1,973,574	1,950	1.29E+06	6.86E+05	4.71E+11		
11	1951	2,247,711	1,951	1.39E+06	8.55E+05	7.31E+11		
12	1952	2,289,836	1,952	1.50E+06	7.88E+05	6.22E+11		
13	1953	2,357,082	1,953	1.61E+06	7.44E+05	5.53E+11		
14	1954	2,314,988	1,954	1.73E+06	5.87E+05	3.44E+11		
15	1955	2,484,428	1,955	1.85E+06	6.39E+05	4.08E+11		
16	1956	2,617,283	1,956	1.96E+06	6.53E+05	4.26E+11		
17	1957	2,616,901	1,957	2.08E+06	5.33E+05	2.84E+11		
18	1958	2,448,987	1,958	2.20E+06	2.45E+05	5.99E+10		
19	1959	2,574,590	1,959	2.32E+06	2.51E+05	6.29E+10		
20	1960	2,574,933	1,960	2.44E+06	1.33E+05	1.77E+10		
21	1961	2,621,758	1,961	2.56E+06	6.39E+04	4.09E+09		
22	1962	2,676,189	1,962	2.67E+06	5.43E+03	2.95E+07		
23	1963	2,752,723	1,963	2.78E+06	-2.70E+04	7.31E+08		
24	1964	2,786,822	1,964	2.88E+06	-9.71E+04	9.44E+09		
25	1965	2,848,514	1,965	2.98E+06	-1.34E+05	1.80E+10		
26	1966	3,027,763	1,966	3.07E+06	-4.68E+04	2.19E+09		
27	1967	3,215,742	1,967	3.16E+06	5.64E+04	3.18E+09		
28	1968	3,329,042	1,968	3.24E+06	9.30E+04	8.64E+09		
29	1969	3,371,751	1,969	3.30E+06	6.77E+04	4.58E+09		
30	1970	3,517,450	1,970	3.36E+06	1.55E+05	2.39E+10		



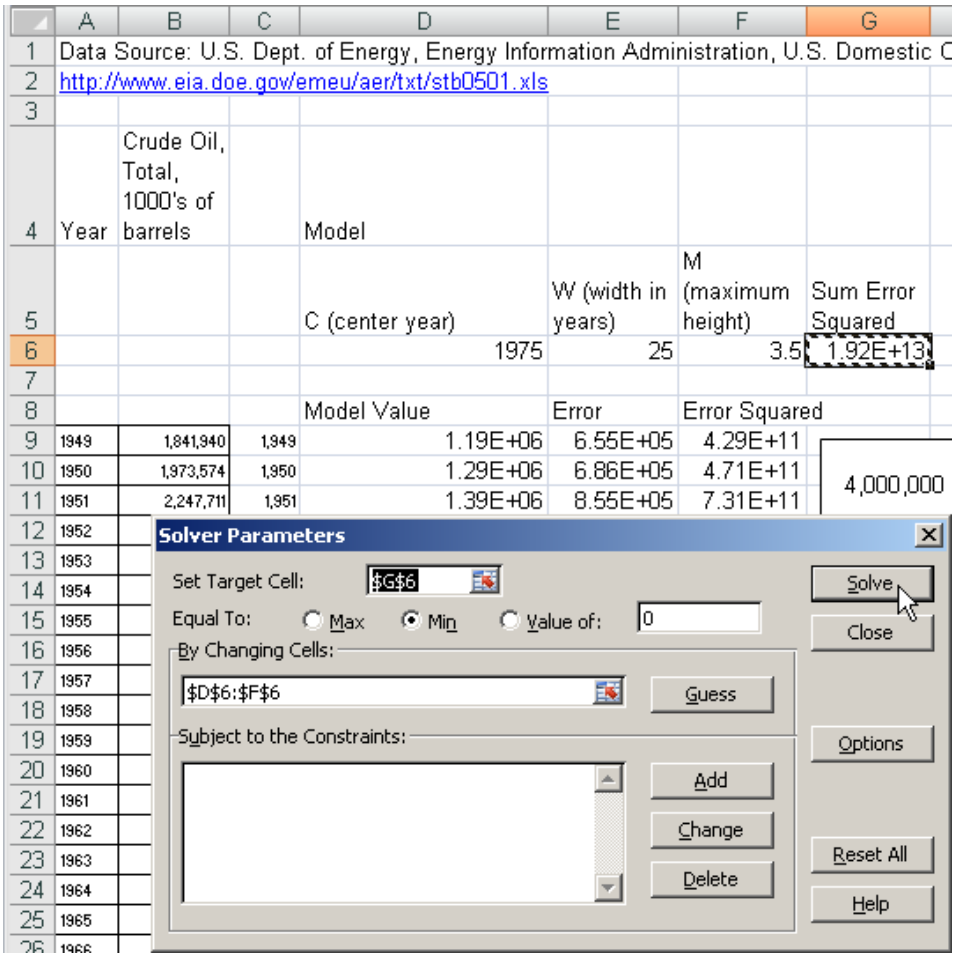
(Formulas)

	A	B	C	D	E	F	G
1	Data Source						
2	http://www.e						
3							
4	Year	Crude Oil, Total, 1000's of barrels		Model			
5				C (center year)	W (width in years)	M (maximum height)	Sum Error Squared
6				1975	25	3.5	=SUM(F9:F68)
7							
8				Model Value	Error	Error Squared	
9	1949	1841940	=A9	=($F$6*1000000)*EXP(-(((($D$6-C9)/$E$6)^2))	=B9-D9	=E9*E9	
10	1950	1973574	=A10	=($F$6*1000000)*EXP(-(((($D$6-C10)/$E$6)^2))	=B10-D10	=E10*E10	
11	1951	2247711	=A11	=($F$6*1000000)*EXP(-(((($D$6-C11)/$E$6)^2))	=B11-D11	=E11*E11	
12	1952	2289836	=A12	=($F$6*1000000)*EXP(-(((($D$6-C12)/$E$6)^2))	=B12-D12	=E12*E12	
13	1953	2357082	=A13	=($F$6*1000000)*EXP(-(((($D$6-C13)/$E$6)^2))	=B13-D13	=E13*E13	
14	1954	2314988	=A14	=($F$6*1000000)*EXP(-(((($D$6-C14)/$E$6)^2))	=B14-D14	=E14*E14	
15	1955	2484428	=A15	=($F$6*1000000)*EXP(-(((($D$6-C15)/$E$6)^2))	=B15-D15	=E15*E15	
16	1956	2617283	=A16	=($F$6*1000000)*EXP(-(((($D$6-C16)/$E$6)^2))	=B16-D16	=E16*E16	
17	1957	2616901	=A17	=($F$6*1000000)*EXP(-(((($D$6-C17)/$E$6)^2))	=B17-D17	=E17*E17	
18	1958	2448987	=A18	=($F$6*1000000)*EXP(-(((($D$6-C18)/$E$6)^2))	=B18-D18	=E18*E18	
19	1959	2574590	=A19	=($F$6*1000000)*EXP(-(((($D$6-C19)/$E$6)^2))	=B19-D19	=E19*E19	
20	1960	2574933	=A20	=($F$6*1000000)*EXP(-(((($D$6-C20)/$E$6)^2))	=B20-D20	=E20*E20	
21	1961	2621758	=A21	=($F$6*1000000)*EXP(-(((($D$6-C21)/$E$6)^2))	=B21-D21	=E21*E21	
22	1962	2676189	=A22	=($F$6*1000000)*EXP(-(((($D$6-C22)/$E$6)^2))	=B22-D22	=E22*E22	
23	1963	2752723	=A23	=($F$6*1000000)*EXP(-(((($D$6-C23)/$E$6)^2))	=B23-D23	=E23*E23	
24	1964	2786822	=A24	=($F$6*1000000)*EXP(-(((($D$6-C24)/$E$6)^2))	=B24-D24	=E24*E24	
25	1965	2848514	=A25	=($F$6*1000000)*EXP(-(((($D$6-C25)/$E$6)^2))	=B25-D25	=E25*E25	
26	1966	3027763	=A26	=($F$6*1000000)*EXP(-(((($D$6-C26)/$E$6)^2))	=B26-D26	=E26*E26	
27	1967	3215742	=A27	=($F$6*1000000)*EXP(-(((($D$6-C27)/$E$6)^2))	=B27-D27	=E27*E27	
28	1968	3329042	=A28	=($F$6*1000000)*EXP(-(((($D$6-C28)/$E$6)^2))	=B28-D28	=E28*E28	
29	1969	3371751	=A29	=($F$6*1000000)*EXP(-(((($D$6-C29)/$E$6)^2))	=B29-D29	=E29*E29	
30	1970	3517450	=A30	=($F$6*1000000)*EXP(-(((($D$6-C30)/$E$6)^2))	=B30-D30	=E30*E30	

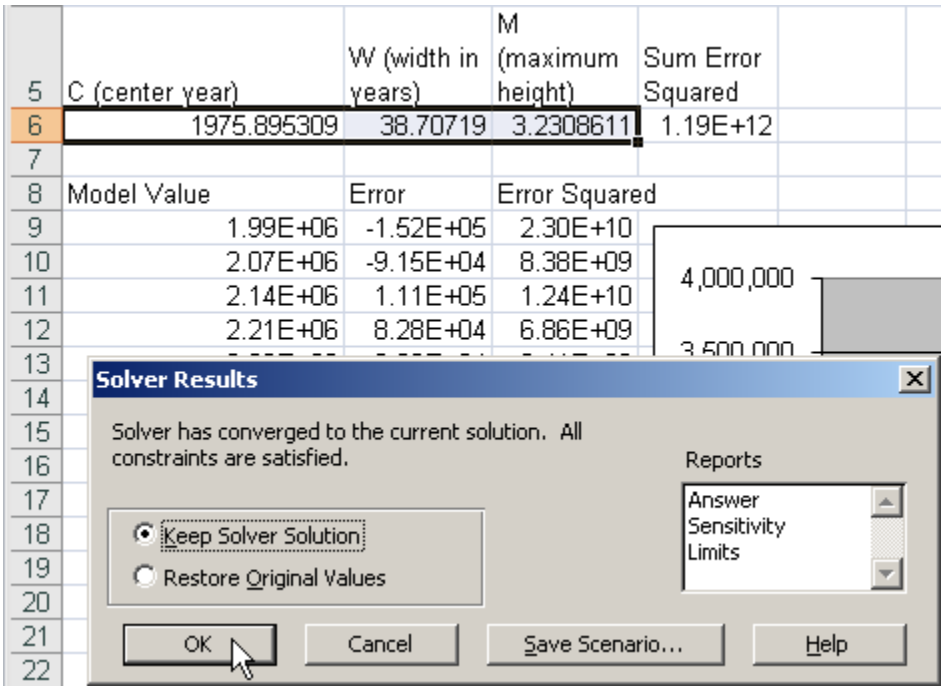


As you can see, the curve has roughly the shape that we intended, but it's definitely too narrow. It may also be too high in the middle and not be centered quite right.

This is where we use the solver. Notice that we've said to make the target cell G6 (sum of squared errors) be minimum, by changing C, W, and M.

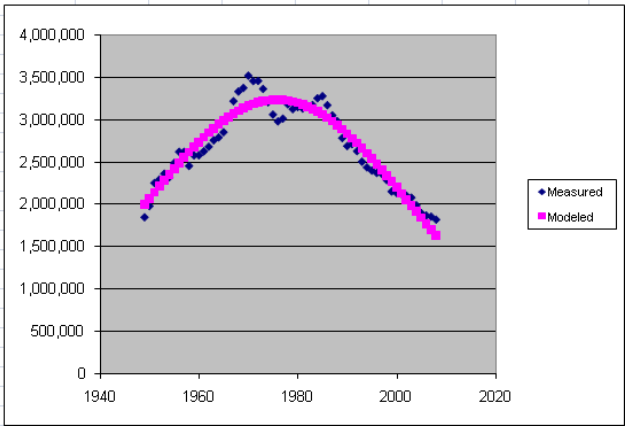


When we click on Solve, the screen flashes very briefly, the values in C, W, and M change, and a popup appears:



We click OK to keep the values that the Solver found. This exposes the entire graph:

	A	B	C	D	E	F	G
1	Data Source: U.S. Dept. of Energy, Energy Information Administration, U.S. Domestic Oil Production						
2	http://www.eia.doe.gov/emeu/aer/txt/stb0501.xls						
3							
4	Year	Crude Oil, Total, 1000's of barrels		Model			
5				C (center year)	W (width in years)	M (maximum height)	Sum Error Squared
6				1975.895309	38.70719	3.2308611	1.19E+12
7							
8				Model Value	Error	Error Squared	
9	1949	1,841,940	1,949	1.99E+06	-1.52E+05	2.30E+10	
10	1950	1,973,574	1,950	2.07E+06	-9.15E+04	8.38E+09	
11	1951	2,247,711	1,951	2.14E+06	1.11E+05	1.24E+10	
12	1952	2,289,836	1,952	2.21E+06	8.28E+04	6.86E+09	
13	1953	2,357,082	1,953	2.28E+06	8.00E+04	6.41E+09	
14	1954	2,314,988	1,954	2.35E+06	-3.12E+04	9.70E+08	
15	1955	2,484,428	1,955	2.41E+06	7.03E+04	4.94E+09	
16	1956	2,617,283	1,956	2.48E+06	1.37E+05	1.86E+10	
17	1957	2,616,901	1,957	2.55E+06	7.11E+04	5.05E+09	
18	1958	2,448,987	1,958	2.61E+06	-1.60E+05	2.56E+10	
19	1959	2,574,590	1,959	2.67E+06	-9.58E+04	9.18E+09	
20	1960	2,574,933	1,960	2.73E+06	-1.55E+05	2.39E+10	
21	1961	2,621,758	1,961	2.79E+06	-1.64E+05	2.70E+10	
22	1962	2,676,189	1,962	2.84E+06	-1.64E+05	2.69E+10	
23	1963	2,752,723	1,963	2.89E+06	-1.39E+05	1.92E+10	
24	1964	2,786,822	1,964	2.94E+06	-1.53E+05	2.34E+10	
25	1965	2,848,514	1,965	2.98E+06	-1.36E+05	1.86E+10	
26	1966	3,027,763	1,966	3.03E+06	1.30E+03	1.69E+06	
27	1967	3,215,742	1,967	3.06E+06	1.51E+05	2.28E+10	
28	1968	3,329,042	1,968	3.10E+06	2.30E+05	5.28E+10	
29	1969	3,371,751	1,969	3.13E+06	2.42E+05	5.85E+10	
30	1970	3,517,450	1,970	3.16E+06	3.61E+05	1.30E+11	

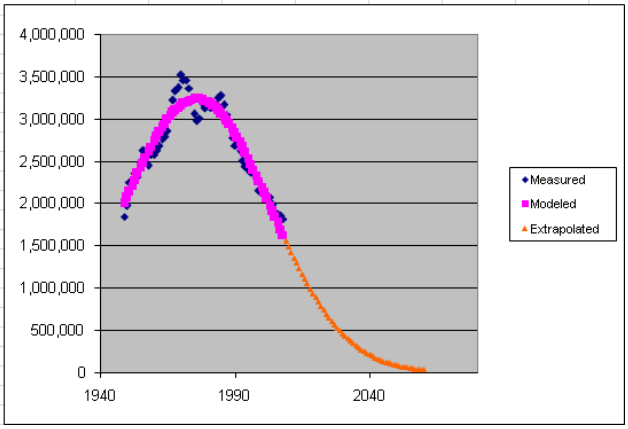


We can now see that the curve overlaps the data quite well.

Once the curve has been fit, it can be used (cautiously!) to extrapolate beyond the observed data. With Excel's trendlines, there is a place in the dialog to just say "Forecast". With Solver-fitted curves, we have to do the extrapolation manually, by creating more years and evaluating the fitted function at those years.

Here is the result:

	A	B	C	D	E	F	G
1	Data Source: U.S. Dept. of Energy, Energy Information Administration, U.S. Domestic Oil Production						
2	http://www.eia.doe.gov/emeu/aer/txt/stb0501.xls						
3							
4	Year	Crude Oil, Total, 1000's of barrels		Model			
5				C (center year)	W (width in years)	M (maximum height)	Sum Error Squared
6				1975.895309	38.70719	3.2308611	1.19E+12
7							
8				Model Value	Error	Error Squared	
9	1949	1,841,940	1,949	1.99E+06	-1.52E+05	2.30E+10	
10	1950	1,973,574	1,950	2.07E+06	-9.15E+04	8.38E+09	
11	1951	2,247,711	1,951	2.14E+06	1.11E+05	1.24E+10	
12	1952	2,289,836	1,952	2.21E+06	8.28E+04	6.86E+09	
13	1953	2,357,082	1,953	2.28E+06	8.00E+04	6.41E+09	
14	1954	2,314,988	1,954	2.35E+06	-3.12E+04	9.70E+08	
15	1955	2,484,428	1,955	2.41E+06	7.03E+04	4.94E+09	
16	1956	2,617,283	1,956	2.48E+06	1.37E+05	1.86E+10	
17	1957	2,616,901	1,957	2.55E+06	7.11E+04	5.05E+09	
18	1958	2,448,987	1,958	2.61E+06	-1.60E+05	2.56E+10	
19	1959	2,574,590	1,959	2.67E+06	-9.58E+04	9.18E+09	
20	1960	2,574,933	1,960	2.73E+06	-1.55E+05	2.39E+10	
21	1961	2,621,758	1,961	2.79E+06	-1.64E+05	2.70E+10	
22	1962	2,676,189	1,962	2.84E+06	-1.64E+05	2.69E+10	
23	1963	2,752,723	1,963	2.89E+06	-1.39E+05	1.92E+10	
24	1964	2,786,822	1,964	2.94E+06	-1.53E+05	2.34E+10	
25	1965	2,848,514	1,965	2.98E+06	-1.36E+05	1.86E+10	
26	1966	3,027,763	1,966	3.03E+06	1.30E+03	1.69E+06	
27	1967	3,215,742	1,967	3.06E+06	1.51E+05	2.28E+10	
28	1968	3,329,042	1,968	3.10E+06	2.30E+05	5.28E+10	
29	1969	3,371,751	1,969	3.13E+06	2.42E+05	5.85E+10	
30	1970	3,517,450	1,970	3.16E+06	3.61E+05	1.30E+11	



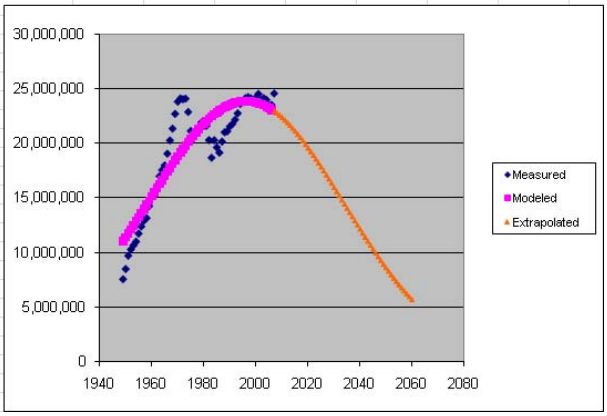
Now, it is important to realize that these extrapolated numbers are based solely on fitting an assumed shape of curve against an observed set of data. If the assumptions are wrong,

the extrapolation will be also. From what we can see in the data, it looks (by eye) like the future tail may not fall quite as steeply as this bell curve predicts. On the other hand, a straight-line extrapolation from the last 20 years would be even worse — that method would predict that by 2040 there would be no domestic production at all! In any case, the message is clear — domestic oil is not the wave of the future.

We can look at any other data set using the same sorts of techniques. The result may be less reliable (or more), depending on whether we have seen enough actual data to establish the parameters of the curve.

Here is what we get if we look at U.S. domestic production of natural gas using data through 2008.

	A	B	C	D	E	F	G
1	Data Source: U.S. Dept. of Energy, Energy Information Administration, U.S. Domestic Oil Production						
2	http://www.eia.doe.gov/emeu/aer/txt/stb0602.xls						
3							
4	Year	Natural Gas gross withdrawals, millions of cubic feet		Model			
5				C (center year)	W (width in years)	M (maximum height)	Sum Error Squared
6				1996.104473	53.277794	23.864412	2.75E+14
7							
8				Model Value	Error	Error Squared	
9	1949	7,546,825	1,949	1.09E+07	-3.37E+06	1.14E+13	
10	1950	8,479,650	1,950	1.13E+07	-2.81E+06	7.87E+12	
11	1951	9,689,372	1,951	1.17E+07	-1.96E+06	3.86E+12	
12	1952	10,272,566	1,952	1.20E+07	-1.75E+06	3.08E+12	
13	1953	10,645,798	1,953	1.24E+07	-1.76E+06	3.08E+12	
14	1954	10,984,850	1,954	1.28E+07	-1.79E+06	3.22E+12	
15	1955	11,719,794	1,955	1.32E+07	-1.44E+06	2.07E+12	
16	1956	12,372,905	1,956	1.35E+07	-1.17E+06	1.37E+12	
17	1957	12,906,669	1,957	1.39E+07	-1.02E+06	1.04E+12	
18	1958	13,146,635	1,958	1.43E+07	-1.16E+06	1.35E+12	
19	1959	14,229,272	1,959	1.47E+07	-4.64E+05	2.15E+11	
20	1960	15,087,911	1,960	1.51E+07	1.11E+04	1.23E+08	
21	1961	15,460,312	1,961	1.55E+07	4.72E+02	2.23E+05	
22	1962	16,038,973	1,962	1.58E+07	1.98E+05	3.90E+10	
23	1963	16,973,368	1,963	1.62E+07	7.52E+05	5.66E+11	
24	1964	17,535,553	1,964	1.66E+07	9.38E+05	8.79E+11	
25	1965	17,963,100	1,965	1.70E+07	9.91E+05	9.83E+11	
26	1966	19,033,839	1,966	1.73E+07	1.69E+06	2.86E+12	
27	1967	20,251,776	1,967	1.77E+07	2.54E+06	6.48E+12	
28	1968	21,325,000	1,968	1.81E+07	3.26E+06	1.06E+13	
29	1969	22,679,195	1,969	1.84E+07	4.26E+06	1.81E+13	
30	1970	23,786,453	1,970	1.88E+07	5.02E+06	2.52E+13	



In this case, we haven't really seen very much (if any) of the falling tail of the curve, so the predictions are less reliable. Just from the data shown here, the situation right now with domestic natural gas looks remarkably like the situation with domestic oil as it was in about 1986 — a sharp rise, followed by an obvious plateau, and some hints but not yet compelling evidence of having peaked out. If that pattern holds up, things will be a lot more clear in another 20 years, at which time we can expect natural gas production to be about half what it is now.

Plan accordingly.