

How Things go Wrong

Data is equal to signal plus noise. That is the key. We use the key by examining the residuals, hoping that they will look like noise. If the residuals look like noise then, indirectly, they confirm that the hypothesis was a true representation of the signal.

In most examples even where a linear hypothesis is good, the first look at the residuals shows some straight forward evidence of a pattern. Until the intercept is right (in the hypothesis) the average residual is not zero. Until the slope is right (in the hypothesis) the residuals show a slope.

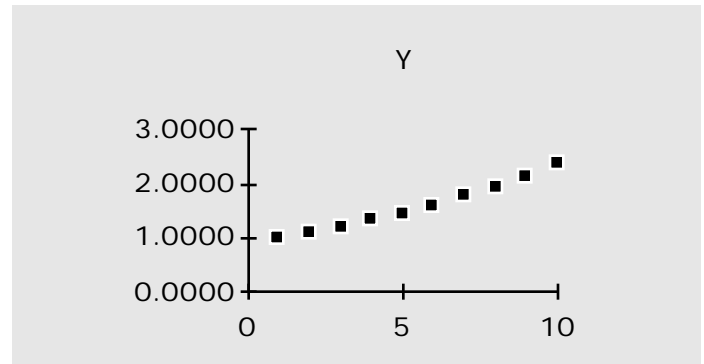
But suppose that the linear hypothesis is dead wrong. Then no patching of the intercept or the slope that will render the residuals patternless. If the linear hypothesis is wrong, what will the residuals look like? Logically the question would not seem to allow a general answer: There is no logical limit to the theoretical equations by which nature may choose to govern the relations among variables. And so it would seem that there is no end to the theoretical patterns which may show up in the residuals when the hypothesis is wrong.

But, in practice, things are not so bad. There may be no mathematical limit to the theoretical equation which nature may use, but, in practice, nature is rarely as perverse as is mathematically possible. I am going to demonstrate what can happen by committing a willful act of stupidity. Watch the residuals.

I'm going to ask you to work with me in a little exercise in curve fitting "by the book". All perfectly straightforward. Here are my data. You and I can see that these numbers are quite orderly. There is a clear system to the sequence 1, 1.1, 1.21, 1.331, etc. And they are a perfect candidate first for logarithmic transformation and then for a linear fit.

You know that but I'm going to forget it. And I'm going to forget the clever things I know about transformations. I'm just going to go at it as numbers, without complicating things by thinking too much. O.K. Here are the numbers, and here is their graph.

X	Y
1	1.0000
2	1.1000
3	1.2100
4	1.3310
5	1.4641
6	1.6105
7	1.7716
8	1.9487
9	2.1436
10	2.3579



I'm not sure what you see in this graph, or think you see in this graph, but let's suppose I come up with the observation that these numbers seem to be positive, and seem to exhibit a slope.

So here's the routine. I'm thinking of the schematic relation

$$\text{Data} = \text{Signal} + \text{Noise}$$

And practically I am matching it with the statement

$$\text{Data} - \text{Hypothesis} = \text{Residual}.$$

Then I'm going to subtract the hypothesis from the data, and "look" at the residuals — asking whether the residuals look like noise. If it does then my hypothesis, in the second equation, is a good approximation to the signal, in the first.

So I'll start simple, very simple, with a hypothesis that says nothing at all. So, my "residuals are everything that was in the data, I've explained nothing."

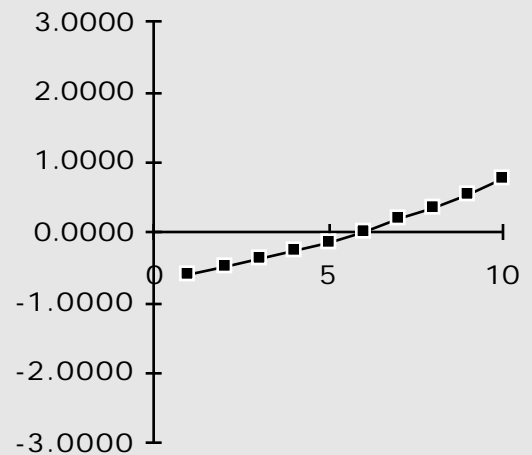
	Slope	0		
	Intercept	0		
X	Y	Y predicted	Residuals: observed Y - expected Y	Absolute Values of Residuals: obs-exp
1	1.0000	0	1.0000	1.0000
2	1.1000	0	1.1000	1.1000
3	1.2100	0	1.2100	1.2100
4	1.3310	0	1.3310	1.3310
5	1.4641	0	1.4641	1.4641
6	1.6105	0	1.6105	1.6105
7	1.7716	0	1.7716	1.7716
8	1.9487	0	1.9487	1.9487
9	2.1436	0	2.1436	2.1436
10	2.3579	0	2.3579	2.3579
		Average	1.5937	1.5937

What I see in these residuals is that they are all positive. So I want to transfer the information "this is positive" out of the residuals and into the hypothesis. It looks positive by about 1.59, (the average value of the "Other") . I'll add this to my hypothesis:

X	Y	Y predicted	Other: observed Y - expected Y	Absolute Value: obs-exp
1	1.0000	0	-0.5900	0.5900
2	1.1000	0	-0.4900	0.4900
3	1.2100	0	-0.3800	0.3800

4	1.3310	0	-0.2590	0.2590
5	1.4641	0	-0.1259	0.1259
6	1.6105	0	0.0205	0.0205
7	1.7716	0	0.1816	0.1816
8	1.9487	0	0.3587	0.3587
9	2.1436	0	0.5536	0.5536
10	2.3579	0	0.7679	0.7679
		Average	0.0037	0.3727
	Slope	0		
	Intercept	1.59		

That's progress: The residuals are smaller. Checking both the average residual and the mean absolute size of the residuals, the residuals are smaller. But, visibly, there is a positive slope to these residuals: They go from about -.6 to about +.8 as x goes from 1 to 10. So a good estimate of the slope in the residuals would be approximately about $(.8 + .6)/9$ or approximately $1.4/9$, or approximately .16. So I'll add this slope to the hypothesis (thereby subtracting the slope from the residuals).



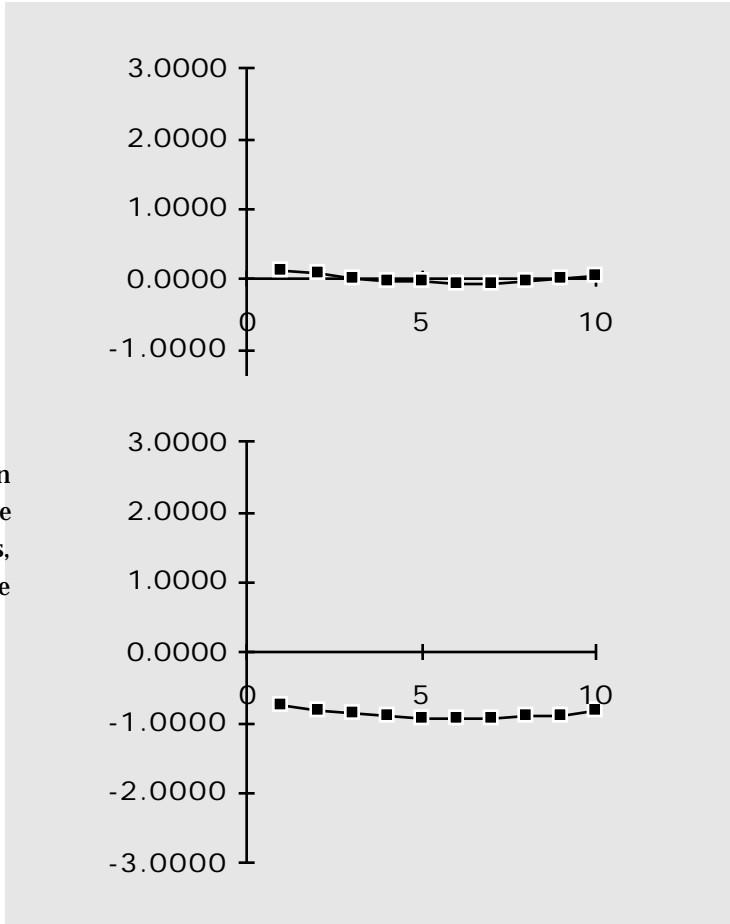
Moving the slope into my hypothesis:

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	1.75	-0.7500	0.7500
2	1.1000	1.91	-0.8100	0.8100
3	1.2100	2.07	-0.8600	0.8600
4	1.3310	2.23	-0.8990	0.8990
5	1.4641	2.39	-0.9259	0.9259

6	1.6105	2.55	-0.9395	0.9395
7	1.7716	2.71	-0.9384	0.9384
8	1.9487	2.87	-0.9213	0.9213
9	2.1436	3.03	-0.8864	0.8864
10	2.3579	3.19	-0.8321	0.8321
		Average	-0.8763	0.8763
	Slope	0.16		
	Intercept	1.59		

Ah, I got rid of the slope. But now I can see that my hypothesis about the signal is a little too large, too high, leaving negative residuals, about -.88. So I will move this too from the residuals to the hypothesis.

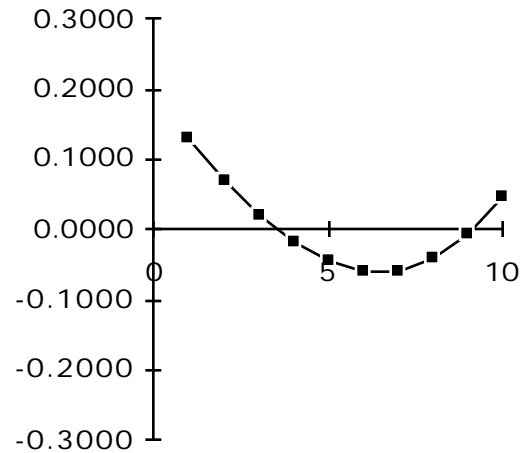
X	Y	Y predicted	Residuals: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	0.87	0.1300	0.1300
2	1.1000	1.03	0.0700	0.0700
3	1.2100	1.19	0.0200	0.0200
4	1.3310	1.35	-0.0190	0.0190
5	1.4641	1.51	-0.0459	0.0459
6	1.6105	1.67	-0.0595	0.0595
7	1.7716	1.83	-0.0584	0.0584
8	1.9487	1.99	-0.0413	0.0413
9	2.1436	2.15	-0.0064	0.0064
10	2.3579	2.31	0.0479	0.0479
		Average	0.0037	0.0498
	Slope	0.16		
	Intercept	0.71		



There: tiny residuals. The average deviation, using absolute values to check variation in either direction, is about .05. That's small compared to the original 1.59, one thirtieth of the original residuals and small compared to the values I am trying to predict (which range from 1 to 2.36). So, is my work complete? Well, not quite. I can't really "see" what's left in this graph precisely because the stuff that's left is so small compared to the original scale. So, just to get a good look at the residuals, let me change the scale of the graph and look again — just to be sure. And ...

Same numbers, expanded scale on their graph

That's trouble. The residuals show a clear pattern. And thus, unravelling my verbal equations, the residuals in equation 2 do not look like "noise" as specified in equation 1. And that implies that my hypothesis in equation 2 does not look like the unknown "signal" in equation 1.



Pausing for the moment to emphasize the moral to this story, the pattern in this graph is lesson #1: When the hypothesis is wrong, deal wrog, the residuals often fall into one of two patterns, a simple curve, concave up or concave down.

The errors may be tiny, but no matter. It is quite clear that the linear hypothesis does not describe the process. If this is the pattern of the residuals, then the linear hypothesis is wrong, numerically accurate, but dead wrong.

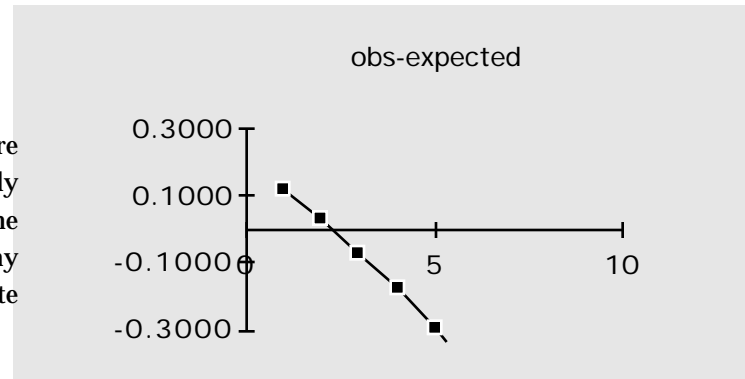
That's lesson #1. Now, back to the analysis. What am I to do with this curving residual? What I should do is step back and think. But instead of flexing my intellectual muscle I am going to flex my arithmetical muscle and use the power of my (decidedly unintellectual) computer. Thinking about the arithmetic, the "obvioius" answer is: Add something curvy to the hypothesis, matching or attempting to match the curviness in the residuals. All right, suppose I add a "quadratic" term, an x-squared term in addition to the existing linear term and the constant. If the linear equation

$$y = mx + b$$

did not work, then I will up the ante by adding another term. Switching notation, I will try

$$y = a_0 + a_1x + a_2x^2$$

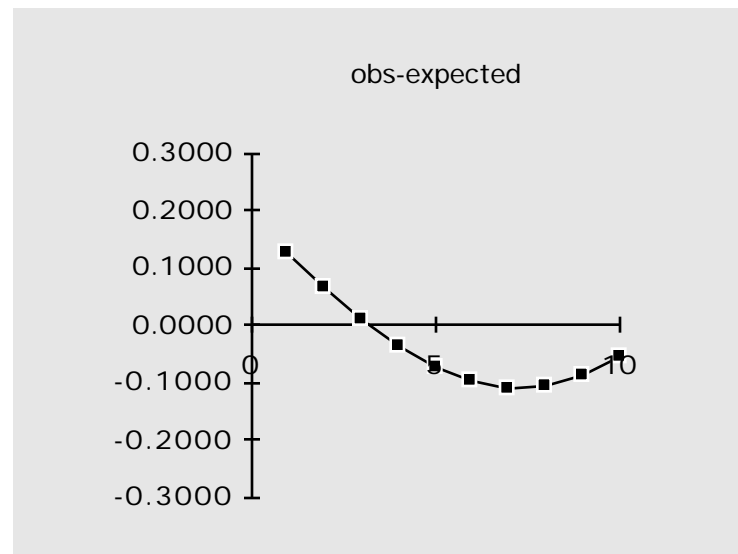
That is the new and more sophisticated equation (falsely sophisticated). I test it by looking at the residuals, the same way I test any hypothesis. Starting with a small estimate for the quadratic term, using $.01x^2$.



X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	0.88	0.1200	0.1200
2	1.1000	1.07	0.0300	0.0300
3	1.2100	1.28	-0.0700	0.0700
4	1.3310	1.51	-0.1790	0.1790
5	1.4641	1.76	-0.2959	0.2959
6	1.6105	2.03	-0.4195	0.4195
7	1.7716	2.32	-0.5484	0.5484
8	1.9487	2.63	-0.6813	0.6813
9	2.1436	2.96	-0.8164	0.8164
10	2.3579	3.31	-0.9521	0.9521
		Average	-0.3813	0.4113
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.01		

Small as it was it has messed up the residuals, not simplified them. So, I'll try something smaller:

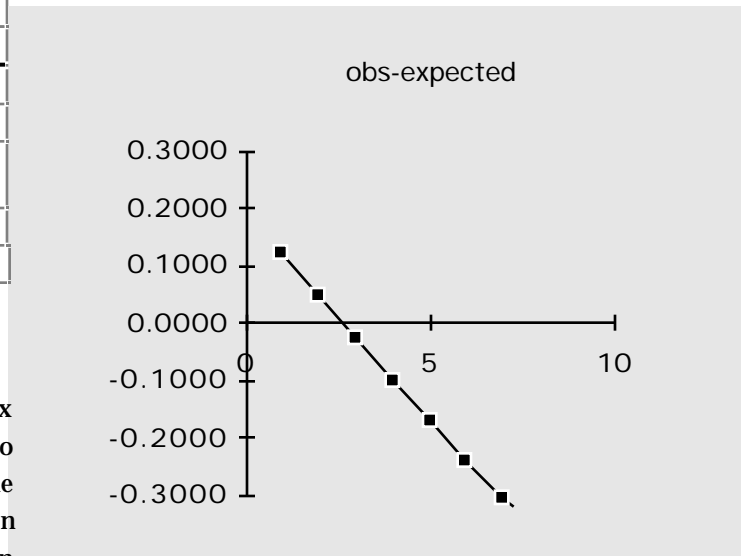
X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	0.871	0.1290	0.1290
2	1.1000	1.034	0.0660	0.0660
3	1.2100	1.199	0.0110	0.0110
4	1.3310	1.366	-0.0350	0.0350
5	1.4641	1.535	-0.0709	0.0709
6	1.6105	1.706	-0.0955	0.0955
7	1.7716	1.879	-0.1074	0.1074
8	1.9487	2.054	-0.1053	0.1053
9	2.1436	2.231	-0.0874	0.0874
10	2.3579	2.41	-0.0521	0.0521
		Average	-0.0348	0.0760
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.001		



Better. Right direction. But there is still a curve in the residuals. So let me put more curve into my hypothesis, subtracting it from these residuals.

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	0.875	0.1250	0.1250
2	1.1000	1.05	0.0500	0.0500
3	1.2100	1.235	-0.0250	0.0250
4	1.3310	1.43	-0.0990	0.0990
5	1.4641	1.635	-0.1709	0.1709
6	1.6105	1.85	-0.2395	0.2395
7	1.7716	2.075	-0.3034	0.3034

8	1.9487	2.31	-0.3613	0.3613
9	2.1436	2.555	-0.4114	0.4114
10	2.3579	2.81	-0.4521	0.4521
		Average	-0.1888	0.2238
	Slope	0.16		
	Intercept	0.71		
	Quad term	0.005		



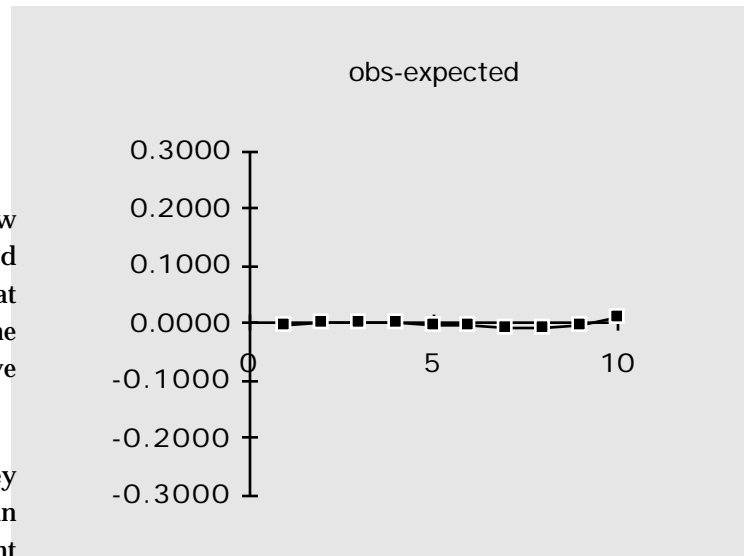
That's looking straight. So I have to fix up the slope: I've got a negative slope to my residuals, so I have to make the hypothetical slope more negative. And then I will have to fix up the intercept, and then the quadratic term again, and then the slope again, and then the intercept again, and I get:

X	Y	Y predicted	Other: observed Y-expected Y	Absolute Value: obs-exp
1	1.0000	1.002	-0.0020	0.0020
2	1.1000	1.0973	0.0027	0.0027
3	1.2100	1.2062	0.0038	0.0038
4	1.3310	1.3287	0.0023	0.0023
5	1.4641	1.4647	-0.0006	0.0006
6	1.6105	1.6143	-0.0038	0.0038
7	1.7716	1.7774	-0.0059	0.0059
8	1.9487	1.9541	-0.0054	0.0054
9	2.1436	2.1444	-0.0008	0.0008
10	2.3579	2.3482	0.0097	0.0097
		Average	0.0000	0.0037
	Slope	0.075		

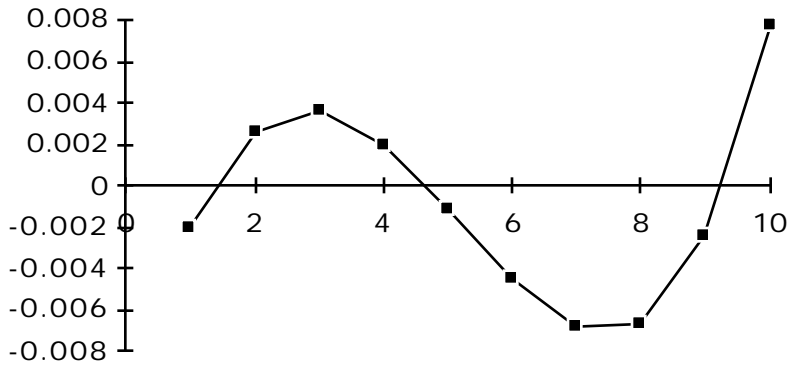
	Intercept	0.9202		
	Quad term	0.0068		

Now, look at that! Depending on how broadly those lines come out on the printed page, I've reduced my error to something that practically disappears within the breadth of the line I've used for the horizontal axis. Now, I've got it. Right?

Well, ... let's look at the residuals. They are small, an order of magnitude smaller than they were in my best effort using the straight line (without the quadratic term). But then let me also improve the scale of my graph.



Resid = Obs'd Y - exp'd Y



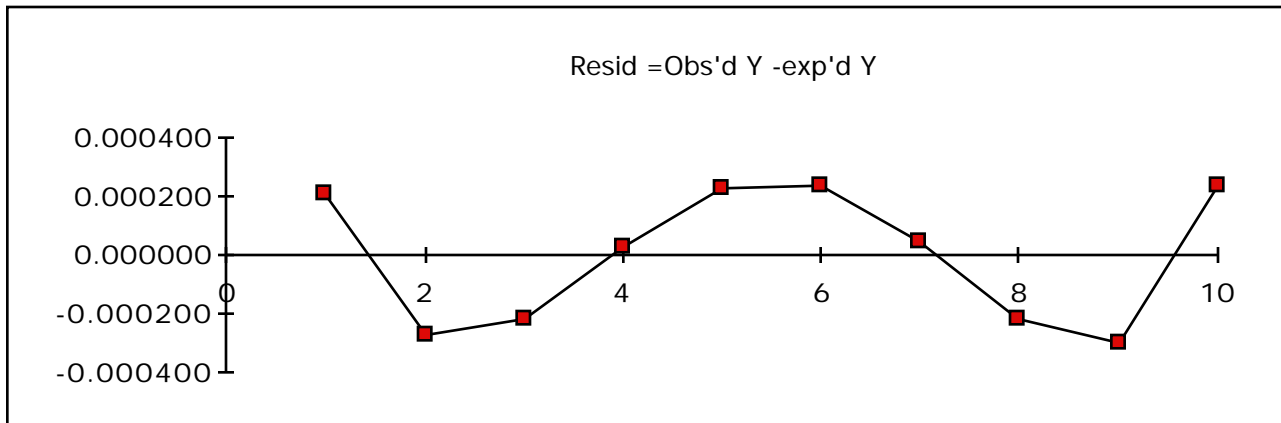
The residuals are tiny, plus or minus .005, give or take. No matter. The residuals show a pattern. The hypothesis is wrong (no matter how small the errors).

And suppose I'm a slow learner, perfectly capable of fitting a (supposedly) more sophisticated model

$$y = a_0 + a_1x + a_2x^2 + a_3x^3$$

Then nature will oblige by punishing my error with an even more interesting pattern among the residuals.

X	Y	slope Expected Y(X)	Resid =Obs'd -exp'd Y	0.08875 Y Squared	Intercept Residual	0.90742 Quadratic: Cubic	0.00338765 0.00022400
1	1.000000	0.999784	0.000216	4.6592E-08			
2	1.100000	1.100268	-0.000268	7.1601E-08			
3	1.210000	1.210214	-0.000214	4.5923E-08			
4	1.331000	1.330968	0.000032	1.0067E-09			
5	1.464100	1.463874	0.000227	5.1302E-08			
6	1.610510	1.610274	0.000236	5.5711E-08			
7	1.771561	1.771514	0.000047	2.2407E-09			
8	1.948717	1.948937	-0.000219	4.817E-08			
9	2.143589	2.143887	-0.000298	8.8734E-08			
10	2.357948	2.357708	0.000240	5.7452E-08			
		mean	0.00	0.00000			0.00021650



Some people will tell you that the sine qua non of science is prediction. But that is too simple a dictum to follow blindly. Here, with my cubic equation I have used the equation to predict or match the values of y , matching these data with precision that is so good that the remaining errors are beginning to get lost in the normal rounding errors made by my computer. I started with a “ y ” that ranged from 1 to 2.38. I’ve matched those numbers subject to errors which are less than 0.003 in absolute value. I’ve fit the data without attempting to understand it — which is a waste of time. I would rank this data analysis as overly mathematical, unnecessarily precise, technically difficult, totally lacking in insight, and dead wrong.

The Meaning of the Pattern

I have demonstrated this nonsense in such detail because this particular sequence of patterns among residuals will haunt you as you proceed in data analysis. This is what you get when there is an answer — but you are not approaching it correctly. In this case there is an answer: This is an exponential growth curve. It is linear in $\log y$. But I approached it incorrectly when I chose to stay with y (without considering a transformation) and when I attempted to force a polynomial to fit the data.

What you are seeing at work in these residuals is a part of mathematics known to every student of the calculus: You are seeing certain aspects of “power series” at work. Power series and related methods are able to fit a polynomial equation to any sequence of numbers (preferable finite) to any standard of precision — provided you can accept a polynomial of sufficiently high degree.

For example, here is one power series for $\ln(1+x)$ in the range of this problem:

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \dots \quad (-1 < x < 1)$$

The point is simply that when you get it wrong, “thinking” the expression on the right instead of the simple expression on the left then what can happen is that when you fit a line, what remains may be dominated by missing terms in the 2nd power, 3rd power, and more. When you fit a quadratic, what remains may be dominated by missing terms in the 3rd power, 4th power, and more. And so on, until you run out of data. So when you see systematic variation among the residuals, stop, look to your hypothesis, not your computer.

Exercise

Lest you think that residuals of the sort I’ve described are a mathematical possibility but not a realistic concern, with data:

Plasticity of Wool -- from Tukey ***