

The Soybean Hypothesis: Growth of an Organism v/s Growth of a Population

For a second example, let me escalate from four data points to seven — same basic numerical technique, but different data and, therefore, different hypotheses, and a different deployment of the numerical technique in pursuit of the hypotheses.

Again from Snedecor and Cochran:

From *Statistical Methods, Seventh Edition*, Snedecor and Cochran, Iowa State University Press, 1980, p 152:

Example 9.2.1-- The following are measurement of heights of soybean plants in a field — a different random selection each week (Wentz, J. b. and Steward, R.T., *J. Am. Soc. Agron* 16 (1924):534.

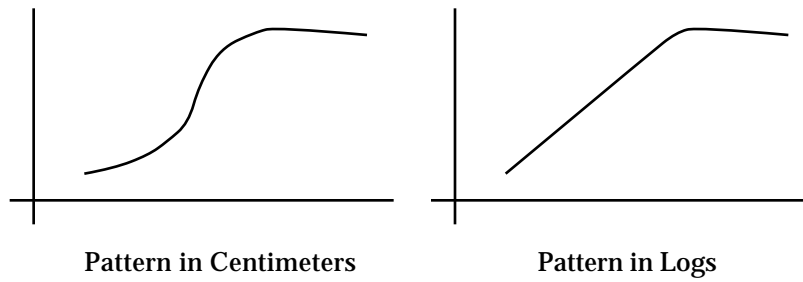
Age, X (wk)	1	2	3	4	5	6	7
Height, Y (cm)	5	13	16	23	33	38	40

Before the beginning, “Who, What, Where,?” As usual these preliminary questions must co-exist with pedagogical duplicity, by which I mean it is necessary to act as if I were subjecting these data to the same close scrutiny that I would exercise were I doing research whose quality and outcome depended on the quality of the data. The textbook is excellent. The data are often classics of the trade.

Examining the data I am a little wary after my previous experience: When data come in a series, as did the fertilizer data, look for trouble at one end or the other: Boundary points are more likely to be “different” than middle points. Here, the initial boundary (age 0, height ?) is just missing.

O.K., let me think about growth. My reflexive expectation about growth is that things grow exponentially: Little things grow slowly, big things grow rapidly — both little things and big things grow in proportion to their existing size: A tiny thing can not sustain a rapid growth. But then, at the other end, there are limits as growth approaches the limit of the resources or the capacity of the organism.

So I expect accelerating growth at the beginning, decelerating growth near the end. The accelerating growth should be approximately exponential and, therefore, it should be linear when the unit of analysis is the logarithm of the height. But then — on second thought I’m not so sure: What is it that increases? Is it the height of the plant, as given? Or is it the weight of the plant, the “biomass” . If it is the latter then I should be looking at the cube root of the height (the relation between volume and extension where form is constant). But then the log of the cube root of the height the result would be linear, or non-linear, as the log of the height itself is linear or non-linear. So, I am still looking for something like a linear relation between the logarithm of the height and the age of the plant.

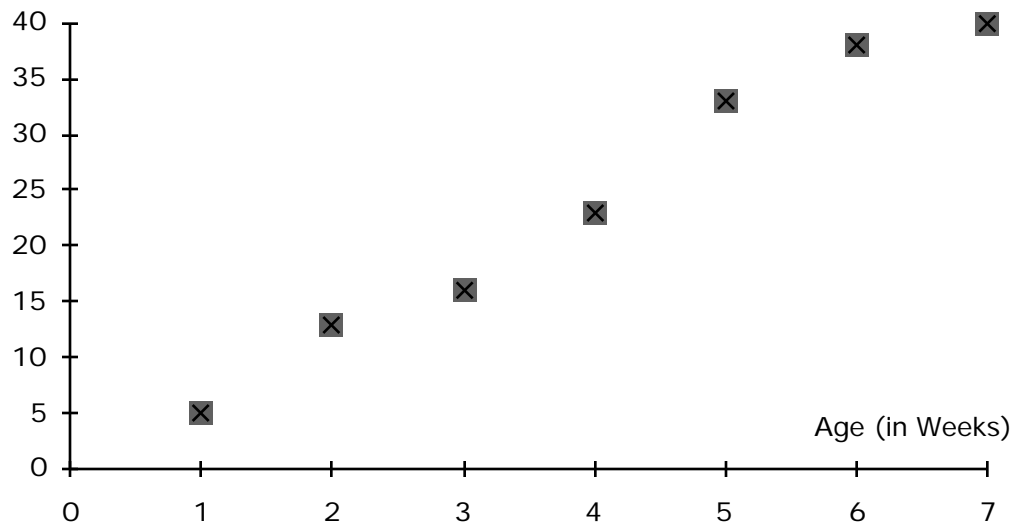


As for one-variable analysis, once again this is a controlled experiment. That implies that values of age were controlled by the experimenter and there is nothing to be learned about nature from the examination of the numbers for age. But I will be able to use a two-variable criterion, looking for linearity (in the log of the height).

Hypothesis in hand, I am committed. Let me see what the data have to teach me.

Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	intercept is: slope is: expected= intercept + slope * Variable x	Residuals= y-expected
# 1	1	5		
# 2	2	13		
# 3	3	16		
# 4	4	23		
# 5	5	33		
# 6	6	38		
# 7	7	40		

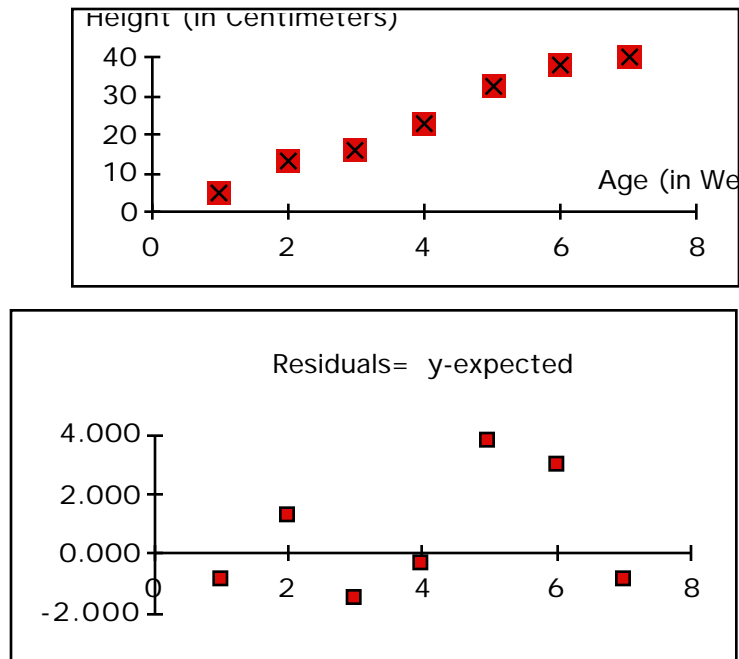
Height (In Centimeters)



This is discouraging. It looks nearly linear — although it will take a look at the residuals to test that appearance. But my thinking led me to expect an increasing rate of growth as the plant became large enough to support more growth. My thinking is wrong. That is the way to use data: It is very good at disproving hypotheses while it can never really prove a hypothesis. So we set up the data analysis in order to get the maximum mileage from what data analysis is good at — rejecting hypotheses. This hypothesis, and the line of reasoning behind it, is wrong.

Is it really linear? Let me look at the residuals. I will try an intercept of 0, 0 weeks, 0 centimeters. I will try a slope of 6.33 centimeters per week, calculating a rise of 35 centimeters, and a run of 6 weeks, $35/6 = 5.83$

			intercept is:	0	
			slope is:	5.833333	
Unit	Age	of Mean	expected=	Residuals=	
	Soybean	Height	of intercept	+ y-expected	
	(In	Soybean	slope	*	
	Weeks)	Plants	(in Variable x		
		centimeter	s)		
# 1	1	5	5.833	-0.833	
# 2	2	13	11.667	1.333	
# 3	3	16	17.500	-1.500	
# 4	4	23	23.333	-0.333	
# 5	5	33	29.167	3.833	
# 6	6	38	35.000	3.000	
# 7	7	40	40.833	-0.833	



Confirmed, there is little evidence of an exponential curve departing from my linear hypothesis. Confirmed, my hypothesis is not supported by the data. It appears that there is a remaining slope in these residuals. The original picture suggests that the boundary point, 7 weeks, may be different — which fits at least a little of my hypothesis. If I were to remove that last point from the residuals, there would be a definite upward slope. If I am to leave it in, then the residuals suggest a little increase of the variance of the residuals toward the end. I am going to ignore that last data point to see what parameters I get for the best line.

The most important characteristic of residuals is their pattern, not their size. Here the “pattern” is an absence of an upward curve. That being accepted, I can attend to the size of the residuals, apart from their pattern. That makes the “fitting” process easier. Here I will

compute the squared residuals and the mean of the squared residuals, and then select a slope and an intercept for six of the seven data points that gives me a good fit in the sense of least squares.

Starting with the first estimate of the intercept and the slope

intercept is: 0

slope is: 5.833333

Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= of intercept + slope * Variable x	Residuals= y-expected *	Squared Residuals
# 1	1	5	5.833	-0.833	0.69444444
# 2	2	13	11.667	1.333	1.77777778
# 3	3	16	17.500	-1.500	2.25
# 4	4	23	23.333	-0.333	0.11111111
# 5	5	33	29.167	3.833	14.69444444
# 6	6	38	35.000	3.000	9
# 7	7	40	40.833	xxxx	xxxx
				Mean Residual	Mean Squared Residual
				0.916667	4.75462963

The mean residuals (without squares) are slightly positive on the average. Likely then I can improve the fit by transferring the average from the residuals to the hypothesis.

intercept is: 0.92

slope is: 5.833333

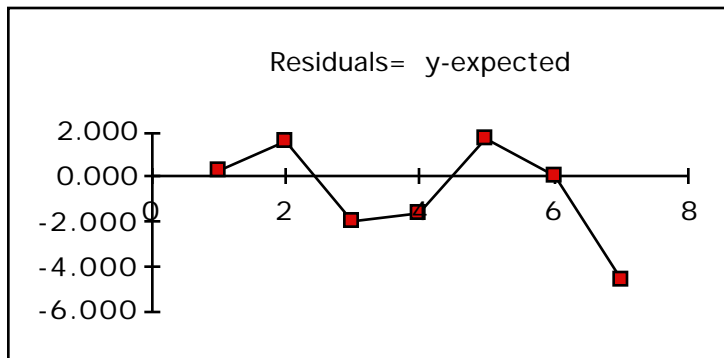
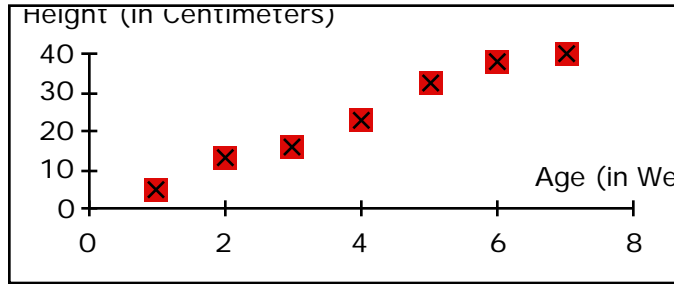
Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= of intercept + slope * Variable x	Residuals= y-expected *	Squared Residuals
# 1	1	5	6.753	-1.753	3.07417778
# 2	2	13	12.587	0.413	0.17084444
# 3	3	16	18.420	-2.420	5.8564
# 4	4	23	24.253	-1.253	1.57084444
# 5	5	33	30.087	2.913	8.48751111
# 6	6	38	35.920	2.080	4.3264
# 7	7	40	41.753	xxxx	xxxx
				Mean Residual	Mean Squared Residual
				-0.00333	3.91436296

Yes, that reduces the squared residual by about 15 per cent, from 4.75 to 3.91. With more work, my best is:

intercept is: -1.86667
slope is: 6.628571

Unit	Age of Soybean (In Weeks)	Mean Height of Soybean Plants (in centimeters)	expected= of intercept + slope * Variable x	Residuals= y-expected *	Squared Residuals
# 1	1	5	4.762	0.238	0.05668934
# 2	2	13	11.390	1.610	2.59056689
# 3	3	16	18.019	-2.019	4.07655329
# 4	4	23	24.648	-1.648	2.71464853
# 5	5	33	31.276	1.724	2.97151927
# 6	6	38	37.905	0.095	0.00907029
# 7	7	40	44.533	-4.533	xxxx

Mean	Mean Squared
Residual	Residual
3.4E-15	2.07
	Square Root of
	the Mean
	Squared
	Residual
	1.44



That gets rid of about half of the squared residuals. To represent this by a number, think of an analogy to the mean, the variance, and the standard deviation: In the analogy, the mean is the number closest to the data in the sense of least squares. The variance reports the mean of those (least) squares. And the standard deviation reports an average error converted (by the square root) to a unit of measure that matches the unit of measure of the variable. Here, by analogy, the square root

of the squared deviation reports the size of the residuals as a standard error, which is, in this case, 1.4 centimeters (computed for data points one through six).

Looking at these residuals, there is no simple pattern among the residuals. (Allowing that the human eye always sees *something*, I can imagine an inverted “W” here. But I don’t take it seriously: Failing any *simple* connection between such a pattern and the substance of the data I will describe the residuals as patternless.)

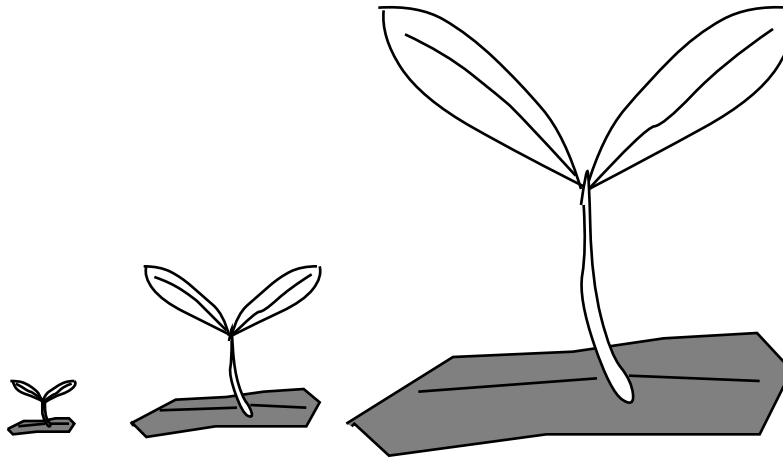
The last data point was not used in this computation because it looked suspicious. The graph supports the exceptional nature of this last point by comparing the observed value to the hypothetical value extrapolated from the six preceding points. Although this is only one data point, it suggests a declining rate of growth for the seven week old plant.

The slope suggests a growth of 6.6 centimeters per week, including a hypothetical week 0, when the bean would have been 2 centimeters below the surface of the ground (the intercept).

The data certainly look simple, the data are linear to a reasonable approximation, for the data preceding the last observation. But this is a negative result: Remember I hypothesized an exponential growth, and I told you why. I was wrong. So the numbers of this linear description are not the end of my work. On the contrary, they tell me I have some explaining to do. There is something wrong with my thinking that needs to be discovered, uprooted, and replaced.

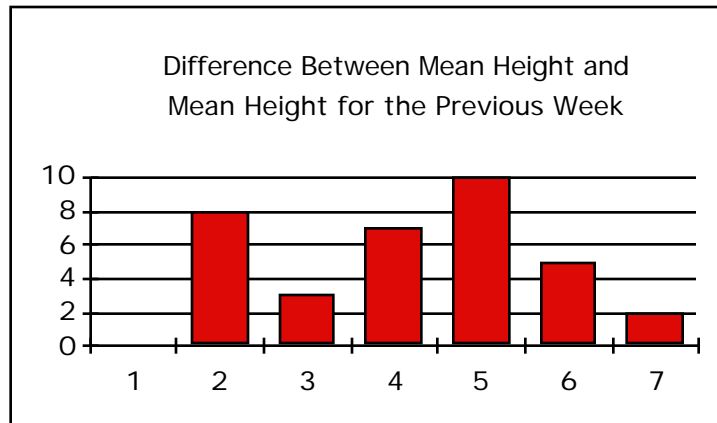
My research process for this dilemma is straight forward — a phone call to my nearest biologist. He advises me that I am thinking like a sociologist. Sociologists look at populations that increase and reasonably well you might expect the rate of growth of a population to be proportional to the number of potential parents. Sociologists look at wealth, which tends to grow in proportion to investment.

Wrong model: Plants don't grow that way. They don't get born looking like a plant, grow up looking like the same plant, but larger, and mature looking like the same plant, but larger.



Wrong model. Plants don't grow that way (nor for that matter do people, although people do, on the average, tend to maintain the same number of arms and legs and heads as they grow older.) Plants like soybeans and trees (and long bones of people), grow at their growing tips. These specialized cells appear to do their work at an approximately constant rate (for a while). The growth of single organisms obeys different laws as compared to the growth of populations.

So — the work is almost complete. But not yet. My friend's story hasn't really explained it. His is a new hypothesis. And you know what to do with a new hypothesis: Test it. No fair doing all that work to test the old hypothesis, and destroy the old hypothesis — and then fail to subject the new explanation to equal scrutiny. The new hypothesis says "constant growth" — at least for a while. So let me do a simple graph of the amount of growth, each week.



This is very confusing. I look at this and I see pattern: Infancy (of a soybean) large initial growth. Then youth, steady growth increasing to a spurt of growth (week 5 has three times the growth of week 3), finished off with a period of declining growth.

This is a problem: I have a story here. I like it. It seems to fit the data. But how did this regularity escape me earlier? What I was trying to do here was create a picture that would be a compelling exhibit of constant growth (if the growth were constant).

But actually I had already constructed a picture of residuals from constant growth — those were the residuals from the straight line (and a straight line is a model of constant growth). What is the difference? There are two differences. The first major difference was my mind set: When I looked at residuals from the straight line I was looking for a rising curve that would encourage me to go forward to test the initial hypothesis (exponential growth). I didn't find it. My mind was set to evaluate the first hypothesis. I did and I rejected it. The second major difference is that in this graph I reflexively left out the first week because the first week could not be compared to an earlier week — it is different.

Now with the new hypothesis I have a different mind set. I'm evaluating this new hypothesis and ready to be critical of the new

hypothesis. And because I am focusing on differences, I reflexively use less data — I only examined the data for which there was a difference to be examined. As a result, subjecting the new hypothesis to the same degree of skepticism as I applied to the original hypothesis, I suspect that I see a pattern (where previously I saw an inverted “W” which I took to be meaningless).

So what is the *truth* about soybeans? What *is* going on here? I have lots of stories now. What is going on? This is embarrassing — seven data points are keeping me fully occupied. So, I need to think: I leave my desk, watch a movie, go for a run, stare at the ceiling — all the time thinking about soybeans.

By now I’ve looked at these seven data points so carefully that it should be no wonder that I can come up with a story that will fit the data — but that doesn’t mean that the new story is wrong. Here is what I’ve learned from these data:

This is a report that will never become a finished report on its own. For me, or for a research group, it is an internal memo — a step in the research. It definitely ends my first hypothesis. But does it allow me to say that growth follows 2 or 3 phases: Rapid initial growth, then a longer period of initially slow but increasing growth, climaxed by a peak, followed by a decline? If I were to pursue that description solely on the numbers, I would need more numbers. For example, if I had daily data, 49 data points, and the pattern persisted, then I would know that the ups and downs were real. 7 data points can do a lot of bouncing. 49 data points following the same broad trends are more credible.

But that’s a numerical fix that avoids getting into the real research, out to the soybean fields, in to the biology of plants. What I would really do, were this 1924 (when the work was published), or were this Spring and I had a few soybeans, is to *look* at the plants: If I withdraw from the numbers and my computer I do know a little bit about plants. They start with one or two seed leaves that are basically

collapsed but present in the seed — ready to catch some sun very quickly. Seeds germinate, doing nothing visible for a few days, then pulling these cotyledons from the seed cover and out into the sun. What is the germination period for a soybean plant? I don't know. I would watch and look for parallels between the "stages" of growth and the phases I suspect I see in the data: Growth from the seed in the first week or two. Then I watch the plant and watch the data: Is there a pause as the plant switches over to a new structure (as compared to the seed growth). Is there a spurt followed in a few weeks by the end of new growth and change to favor other functions — reproduction, or root growth. I would look at the phasing in my data and attempt to match it to structural changes in the plant itself. For the moment, these seven data points, and the techniques of data analysis have done their job which is to lead me (much more deeply than I had intended) into questions about the physiology of soybeans.

Data from Wentz and Stuart (*J Am Soc. Agron* 16, 1924) p. 534) demonstrate the pattern of growth in the soybean plant. Starting from a seed beneath the soil, the plants achieved an average length of 40 centimeters 7 weeks later. While the length of the average plant increased by an average of 5.7 centimeters per week, the change per week is irregular, showing a maximum of 10 centimeters growth in the fifth week which is three times larger than growth in the third week but almost matched by growth in the second week.

Further research should look for physical changes in the plant which may explain these differences in the growth rates. For example: What is the germination period? At what time are the seed leaves followed by the growth of stem and new leaves. When does the plant cease to grow? When does it flower? These stages in the

growth of the plant may explain its initial growth, 5 centimeters in the first week, 8 centimeters of additional growth in the second week, plummeting to 3 centimeters growth in the third week. After the third week there is a gradual rise, with maximum growth in the fifth week, diinishing to the slowest growth exhibited at any time in the seventh week of the plants' lives.

Average Height of Soybean Plants Compared to the Age of the Plant

Age (weeks)	Length (centimeters)
1	5
2	13
3	16
4	23
5	33
6	38
7	40

Secondary source: *Statistical Methods*, Seventh Edition, Snedecor and Cochran, Iowa State University Press, 1980, p 152.

Initial Growth and Weekly Increase (In Centimeters)

