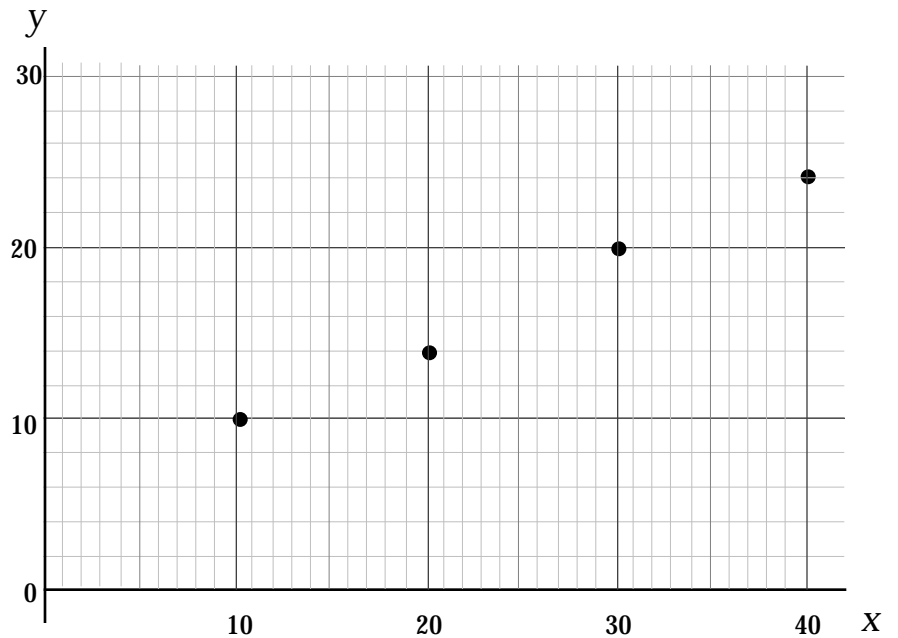


# $D = S + N$ Data = Signal + Noise

Working up to a linear analysis: Before the beginning, Who, What, Where, Why, When, and How? At the beginning, Stem and Leaf, well-behaved variables. Now I'm ready for a two variable linear analysis. I will start simple by making up some hypothetical data, graphing it, and looking at the graph. Here are my data. Here is my graph.

Observation	Data	
	x	y
#1	10	10
#2	20	14
#3	30	20
#4	40	24



Most of what you are ever going to know about these pseudo data you already know — it is apparent in the graph. The relation between  $x$  and  $y$  is positive, “large  $x$  corresponds to large  $y$ ”. The relation looks linear. The relation between  $x$  and  $y$  has a slope of about one half, with values of  $y$  appearing to be directly proportional to the corresponding values of  $x$ .

If these were real data, I would probably stop here: Adding the numbers of formalized statistics would be a way of abstracting from reality, summarizing it, and describing its basic regularities. But these are not real data and they allow me to demonstrate the process of adding the formalized numbers. So I will proceed with the obvious in order to show you what is not obvious about the rules for obtaining evidence from such data.

### **Data = Signal + Noise**

Think about a problem in two variable analysis in these terms: Think of the pseudo-equation

$$Data = Signal + Noise \quad 1$$

It says that data, the stuff we see and describe has, within it, two parts. One is the signal. That's the message. That's what we are trying to figure out. The other part is noise. Noise may be measurement error, meaningless variation, or a level of complexity which, for the moment, we can not penetrate — so, for lack of understanding it looks like noise.

Now, and this is one of the classical strategies of data analysis, to get at the *signal*, we direct our attention to the *noise*: First we rewrite the pseudo-equation, isolating noise by itself, on the right.

$$Data - Signal = Noise \quad 2$$

Then we form a hypothesis about the data and a specific hypothesis about the signal. And then we write a real equation, not a pseudo equation:

$$\text{Data} - \text{Hypothesis} = \text{Residual} \qquad 3$$

The hypothesis specifies what the data should look like. And, of course, it doesn't. It never does. And so the difference, data minus hypothesis, leaves something (usually), something we designate as the residual.

The hypothesis creates an expectation: "This is what the data should look like if the hypothesis is correct." We subtract the hypothesis from the data and see what's left. What's left is called the "residual" and it *should* look like noise. If it doesn't then we reject the hypothesis.

The logic is a bit twisted, I'll admit. We are interested in the signal; we look at the noise. The reason is that signals do all sorts of interesting things — too interesting, too varied. But noise — we understand noise. We know what noise is supposed to look like. We can recognize noise. When the residuals, equation 3, look like noise, equation 2, it means that when we subtracted our hypothesis about the signal from the data, all that remained was noise: So it was a good hypothesis.

### **Well-Behaved Noise**

This strategy becomes critical for organizing the attack on two variable relations but it is the strategy you have already used to identify well-behaved variables, the same strategy with a different name. With one variable the hypothesis and the signal are pretty rudimentary, hardly deserving of the portentous terms "hypothesis" and "signal". With one variable the "formal" hypothesis is the mean and the implicit hypotheses are embedded in the choice of the unit of measure and the unit of analysis.

With one variable, the pseudo equation

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

becomes the real equation

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

3

and very specifically

$$\text{Data} - \text{Mean} = \text{Residual}$$

In words, the “residual” is the distribution of data on either side of the mean. If that residual looks like noise: If it is without pattern. If the size of the residuals is small. If the average of this noise is zero and if the distribution of the noise is symmetrical — then the rudimentary one variable hypothesis is correct.

For one variable, in Volume I, this pseudo equation and its interpretation are overkill — an unnecessarily difficult re-statement of the first property of a well-behaved variable. For two variables, this pseudo equation is standard operating procedure.

In simple unsophisticated terms the principle is “Look at the exceptions.” You are interested in the pattern, in the signal, but you detect it by looking at the residuals. Returning to the pseudo-equation, you can hypothesize a linear signal,

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

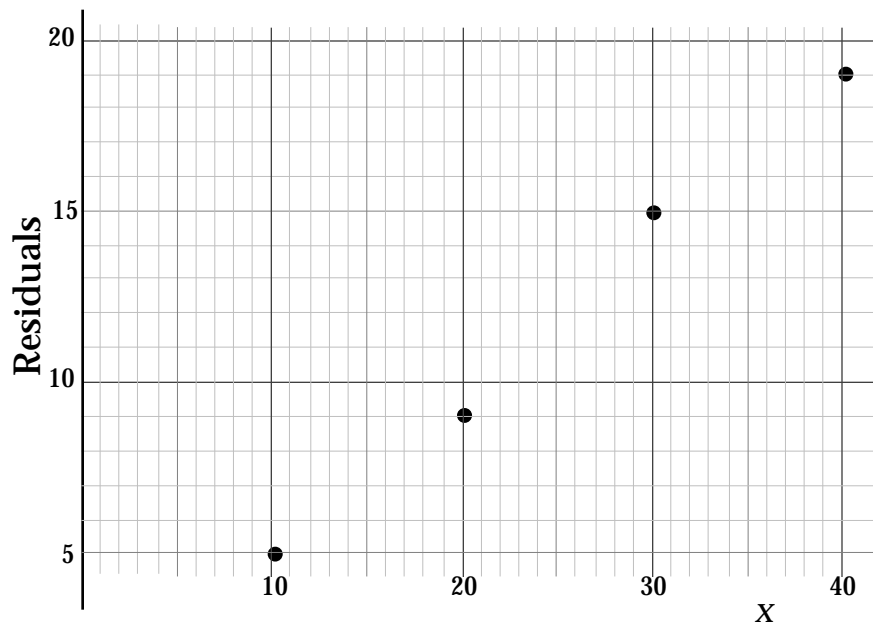
Hypothesis: Signal is a Linear Relation

But you check the hypothesis by looking at the residuals. Adding it up, there are the three things that a data analyst associates with a line: The intercept, the slope and the residuals. If your residuals look like noise, if the residuals are without pattern, if the residuals are small, and if the mean residual is near zero and the distribution of the

residuals is well behaved, then your hypothesis is consistent with the data.

For example, here is standard operating procedure (for lines) in action: Suppose I look at the four points of “data” graphed in Figure A and come up with the hypothesis that these data are approximately constant: The signal is “y is constant.” That’s wrong: It is a poor hypothesis. But let me pursue it to show how a poor hypothesis leads to poor residuals (residuals that look like they still contain a signal). So continuing, foolishly, I hypothesize that the signal is “y = 5, constant”, Figure B.

Because my hypothesis is that “5” is the signal I subtract “5” from the data and look at what’s left. (Note, that I have blown up the scale of the graph, to increase the resolution.)

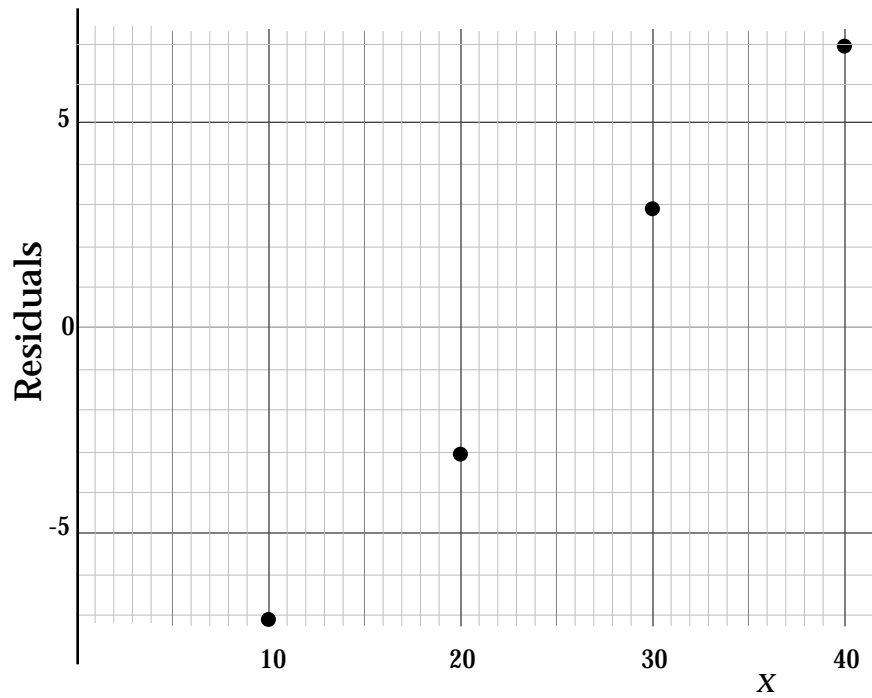


	<i>Data</i>		<i>Hypothesis</i>	<i>Residuals</i>
Observation	<i>x</i>	<i>y</i>	$\hat{y}=5$	$\hat{y}-y$
1	10	5	5	0
2	20	9	5	-4
3	30	15	5	-10
4	40	19	5	-14

#1	10	10	5	5
#2	20	14	5	9
#3	30	20	5	15
#4	40	24	5	19

Figure B

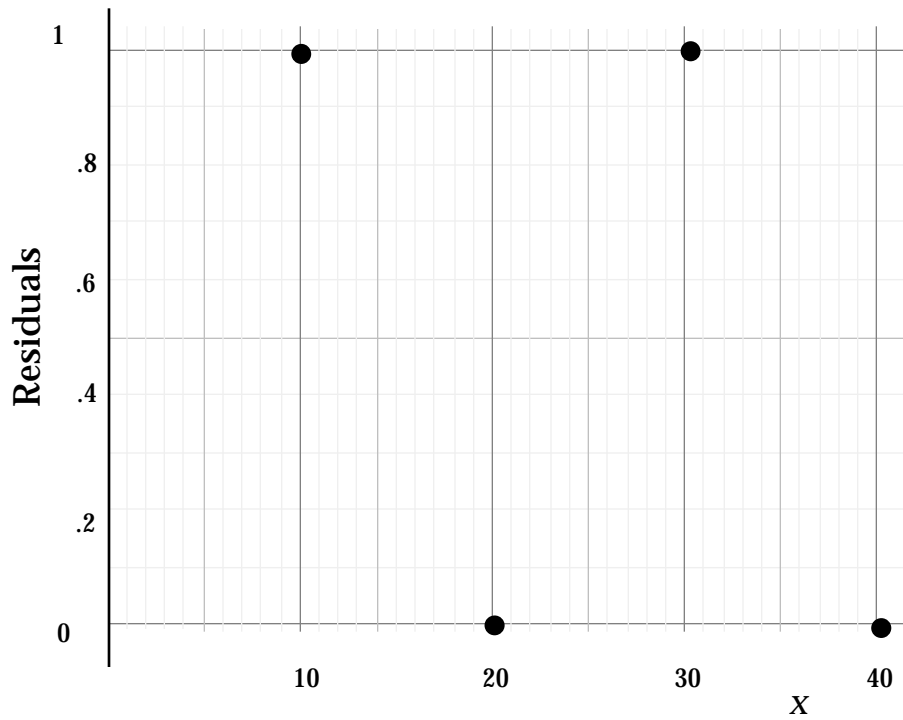
Clearly, that's wrong: The residuals show a clear pattern and the residuals are positive, not zero. That is a crude but obvious signal, not noise. So even if the idea is right (which it isn't), the value specified by the hypothesis is wrong. O.K., I can do better. Now I hypothesize: These data are approximately constant with an average value of 17. Now, for the second time, I subtract the hypothetical signal from the data, the signal as it would be if the hypothesis were correct. What do I get:



Observation	Data		Hypothesis	Residuals
	x	y	$\hat{y}=17$	$\hat{y}-y$
#1	10	10	17	-7
#2	20	14	17	-3
#3	30	20	17	3
#4	40	24	17	7

Better: The residuals now have an average of zero. But what was omitted from the hypothesis is now painfully obvious in the residuals: The residuals increase quite regularly. So, the data were not constant — back to hypothesis construction.

Now, I'm going to hypothesize that the signal increases with  $y$  as a straight line function of  $x$ :  $y = 3x + 4$ . Testing this hypothesis I subtract the hypothetical signal from the data — I subtract the signal as it would be if my hypothesis were correct — and I get

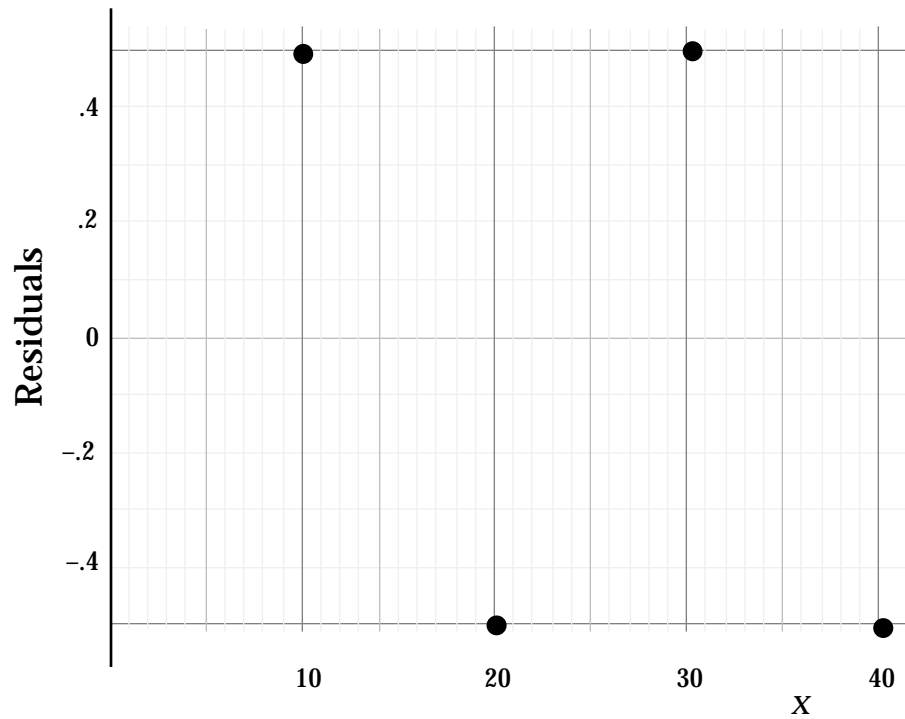


Observation	Data		Hypothesis	Residuals
	x	y	$\hat{y}=mx+b$ b=4, m=.5	$\hat{y}-y$
#1	10	10	9	1
#2	20	14	14	0
#3	30	20	19	1
#4	40	24	24	0

Figure C

That's more like it: The average of the residuals is small: The bouncing that remains in  $y$  is on a scale between 0 and 1, the size of the residuals that are not explained by this hypothesis. Compared to the range of  $y$  between 10 and 24, the hypothesis has greatly reduced my uncertainty about  $y$ .

Looking closely at what remains in the variation of  $y$ , it appears that my noise is a bit more positive than negative — there is a little bit of signal left in the residual. So I can add another 0.5 to the constant (the intercept) in my hypothesis. The residuals show no trace of slope, so I will leave that part of the hypothesis alone. So, with one more refinement,  $y = .5x + 4.5$ , the linear hypothesis is good, leaving residuals (average of zero, no slope) that look like noise. It is a good hypothesis.



Observation	Data		Hypothesis	Residuals
	$x$	$y$	$\hat{y} = mx + b$	$\hat{y} - y$
			$b = 4.5, m = .5$	
#1	10	10	9.5	.5
#2	20	14	14.5	-.5
#3	30	20	19.5	.5
#4	40	24	24.5	-.5

Figure D

Or is it? Is there really no signal left among these residuals? After all of my arithmetic is complete, I am back to the need for human judgment. When I look at what's left, I can't help but observe that the residuals seem to oscillate, up, down, up, and down. Is this a pattern? If there is a pattern in the residuals, then the residuals are not noise and my hypothesis is incomplete. Is this oscillation more signal or is it noise? In truth I can't tell. For that I would have to place the numbers in context as data: I would have to treat those residuals/noise as a variable, two values are positive, two values are negative and I would have to explore the variable, beginning with the stem and leaf and continuing with the average and variation. I couldn't "prove" anything by such exploration, but then I'm not trying to. I'm looking for ideas.

And are these residuals really small enough to ignore? In truth, I can't answer that questions either. Again I would have to place the numbers in context as data: Looking at size, if those residuals were errors in the prediction of the domestic products of the United States, year after year, and the residuals corresponded to a one percent deviation between the hypothesis and the data, then I'd say "forget it — one percent error in gross domestic product is too small to be taken seriously." But even then I would look for pattern, as well as size: If those numbers were gross domestic products for the United States, and the high numbers turned out to be war years, that would give me something to think about — nothing more (no proof, just something to think about — and something to direct my focus at the next stage of my research on GDP), nothing more, but nothing less either.

### Doing it with Excel

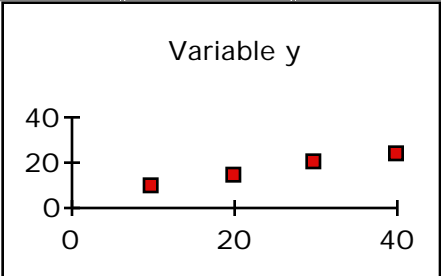
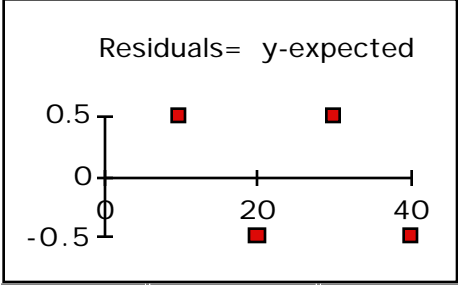
The one thing I do not want you to do with your spreadsheet software on your computer is to have it solve problems for you: Spreadsheet programs are perfectly capable of finding lines for data directly but don't do it. Instead I want you to use the work sheet capacity to carry you through the work step by step — so I know that you know how. (And, because there are choices involved in any data analysis, your program won't get exactly the answer you are looking for.)

So, #1, get your data into the spreadsheet. And be sure to get the units in there among the labels, bags of concrete, 1,000's of people, whatever. Then what? stem and leaf, medians, descriptions, as appropriate for each variable. Each analysis begins at step 0.

Then I'd suggest something like this

			intercept =	Value (You enter this and change it yourself. The formula for the expected value uses this.)	
			slope =	Value (You enter this and change it yourself. The formula for the expected value uses this.)	
Unit	Variable 1 Units	Variable 2 Units	Expected (under the assumptions that the linear hypothesis, with the slope and intercept specified above, is correct	Residual/noise: Observed value of Variable 2 Minus Expected value of Variable 2 under the assumption that the hypothesis is correct	Here you get Excel to do Graph: Across is x, Up is Y. you play around with Excel you can get it to do something visually better: Up is Y and also up is "Expected" graphing both sequences numbers on one graph
	datum	datum	Formula: = \$d\$2 + \$d\$1*A3	Formula =B3-C3	Here you get Excel to do another graph, lined up below the first. On this graph Across is still the same. But now "up" shows the residuals
	datum	datum	Corresponding to previous		

For example

			intercept is:	4.5
			slope is:	0.5
Unit	Variable x	Variable y	expected= intercept + slope * Variable x	Residuals= y-expected
# 1	10	10	9.5	0.5
# 2	20	14	14.5	-0.5
# 3	30	20	19.5	0.5
# 4	40	24	24.5	-0.5
				
				

That may take a little bit of doing to set it up. But I want you to set it up and then have Excel do for you what I did in the text: You pick an Intercept, leaving the slope as 0, and then Excel will draw the

residuals. You add a non-zero slope, and then Excel will draw the residuals. As you get closer, the residuals should get *less* interesting — that's what noise is supposed to do. You should also be able to keep track of it numerically by watching the average size of the residual go down.

Warning, Excel has a friendly habit of trying to choose a good scale on which to display your graphs. It also has a nasty habit of making a bad choice. Since the scale has a lot to do with what you are going to be able to see, as noted in the text, be careful. If you get something ridiculous in your graph — check the scale. You can intervene, or at least you can intervene on the older version of Excel. I trust I will get advised in class on a variety of different ways to do this. But keep your eye on the purpose of the exercise and choose the scale accordingly.