

Big(ger) Data Sets

Height and Weight

I think that a reasonable person has to be stunned by the amount of detail that can be found in as few as 4 data points (fertilizer and potatoes data), or as few as 7 data points (soybean growth data). It surprises me, and it surprises me again every time it happens. Nevertheless, the world often presents us with data involving many data points, considerably more than 8, or 80, or 800. There is no simple answer about what you do because the effect of large data sets usually can be counted on to produce two opposite effects on the self assurance and equanimity of the data analyst. Large data sets can fill in the blanks where you think you see a pattern, but need more data to be sure that the overall pattern is obeyed: In the soybean growth data it is possible that the inverted “W” is nothing but noise. If daily data traced the lines of the “W”, conforming to it and filling in the spaces, it would be reassuring that the pattern (whatever its meaning) is not noise. In this regard data sets can be re-assuring — when the details agree with the overview I have more confidence that the overview is “real”.

Large data sets can also have the opposite effect, confirming that patterns that could not possibly be correct, are correct. And then leaving us with the daunting task of explaining them.

For example take a very close look at the relation between heights and weights of adult (5,000 adult British women), as shown in Table __. I consider it obvious that taller people will tend to be heavier: Certainly there will be tall but thin people who weigh less than short but heavy people. But there will also be tall heavy people and short thin people so that, on the average, taller people will tend to be heavier.

This supposedly obvious statement must stand or fall by the facts. “Obvious” or not an assertion about fact has to be tested. And unfortunately, there are some very strong exceptions in these data. Is it true that taller women are heavier? On the average, yes, but there are some curious exceptions even “on the average”. For example, what are the odds that a 5’2” woman will weigh 140 pounds as compared to 134 pounds? Answer: The

odds are just about even. At 5'2" there are 95 woman weighing in at 134 pounds, 101 weighing in at 140.5 pounds.

Now , for comparison, what are the odds that a women of 5'4" will weigh 140 pounds as compared to 134 pounds? These woman are 2 inches taller. Presumably they are likely to be heavier than the first set of women at 5'2". So the odds favoring 140 pounds as compared to 134 pounds should be greater. Obvious. And false: Among the shorter woman there are about as many woman at the heavier weight as at the lighter weight. By comparison, among these taller woman there are 25% *fewer* woman at the heavier weight than at the lighter weight, 138 versus 175.

How do I explain that? Well, first off, I hope I can explain it away as a quirk, as something weird, as something that doesn't need to be explained. But I can't get away with that one. There are at least eight instances of this effect in these data and the frequencies involved are very large and, therefore, among the most believable numbers in the data.

	Column Sums										Totals	
	5	33	254	813	1340	1454	750	275	56	11	4	4995
278.5 lbs						1						1
272.5 lbs												0
266.5 lbs						1						1
260.5 lbs							1					1
254.5 lbs												0
248.5 lbs					1	1						2
242.5 lbs							1					1
236.5 lbs							1					1
230.5 lbs					2				1			3
224.5 lbs					1	2	1					4
218.5 lbs			1		2	1		1				5
212.5 lbs				2	1	6		1	1			11
206.5 lbs				2	2	3	2		1			10
200.5 lbs			4	2	6	2						14
194.5 lbs				1	3	7	7	4	1			23
188.5 lbs			1	5	14	8	12	3	1	2		46
182.5 lbs			1	7	12	26	9	5		1	2	63
176.5 lbs				5	8	18	21	11	7		2	87
170.5 lbs				2	11	17	44	21	13	3	1	112
164.5 lbs		1	3	12	35	48	30	15	5	3		152
158.5 lbs				8	17	52	42	36	21	9		185
152.5 lbs		1	7	30	81	71	58	21	2	2		273
146.5 lbs		2	13	36	76	91	82	36	8	1		345
140.5 lbs		1	6	55	101	138	89	50	8			448
134.5 lbs			15	64	95	175	122	45	5			521
128.5 lbs		1	19	73	155	207	101	25	3			584
122.5 lbs		3	34	91	168	200	81	12	1	1		591
116.5 lbs		3	24	108	184	184	50	8				561
110.5 lbs		5	33	119	165	124	22	4				472
104.5 lbs	1	3	33	87	95	35	6					260
98.5 lbs	2	5	29	59	45	16	3					159
92.5 lbs		6	10	21	9							46
86.5 lbs		1	5	3								9
80.5 lbs	2	1	1									4

Weight
54in 56in 58in 60in 62in 64in 66in 68in 70in 72in74in Height

Reproduced from Kendall and Stuart, *op. cit.*, p. 300.

Figure 6.1
Distribution of Height and Weight for 4,995 Women, Great Britain, 1951.

152.5 lbs		81	71	58	21
146.5 lbs		76	91	82	36
140.5 lbs		101	138	89	50
134.5 lbs		64	95	175	
128.5 lbs		73	155		
122.5 lbs	34	91			
116.5 lbs	24	108			
Weight					

58in 60in 62in 64in 66in 68in

Reproduced from Kendall and Stuart, *op. cit.*, p. 300.

Figure 6.1

Anomalous Subsets of the Data for the Distribution
of Height and Weight, Great Britain, 1951.

So, how do I explain these anomalous facts of height and weight? For now, I do not explain it. (See Levine, 1993, Chapter 6, for an attempt at an explanation.) The reason I present it, without presenting the methodological trick that makes everything clear, is to make the point that data analysts work at different levels of detail for different purposes. Very often, data analysts are working toward a fairly modest degree of description: Does an increase in x correspond to an increase in y ? When possible, we will put a number on it: For a unit increase in x what is the average increase in y ? In that context the word “model” is applied to the straight line, $y = mx + b$.

Sometimes we want to get into the detail of the data because we want to get into the detail of the mechanism. Then the word “model” is applied to a theory of the mechanism at work behind the data. Most of the time we are not prepared to work with that level of commitment to the work of explaining the mechanism behind the data.

So, satisfying myself with simple description, what is the approximate relation between height and weight:

$$y = mx + b,$$

$$y \text{ pounds} = m \frac{\text{pounds}}{\text{inch}} x \text{ inches} + b \text{ pounds}$$

Even with this modest, though fairly typical goal, there is no line that is going to “fit” these data: We know that immediately. Among the women at 5’2”, for example, the weights range from 92.5 pounds (9 women) to 248.5 pounds (1 woman). The equation $y = mx+b$ is simply not capable of predicting 92.5 pounds for one woman at 5’2” and 248.5 pounds for another woman at 5’2”.

What we try to do is to predict the mean. For women at 5’2”, the mean is 130.22 pounds. For women at 5’4”, the mean is 134.59 pounds. We ask for the “line of means”: the best line for predicting the means, $y = mx+b$? (It would be entirely valid to predict medians. Probably because of the historical difficulty of numerical computation without computers, the line of means became the most often used procedure and survives as the more common procedure.)

In a later chapter I will introduce formulas that make it easy to work with the entire data set when computing the line of means, known also as the “regression line”. But for now, begin with the mean weight at each height and describe the relation between (mean) weight and height:

Homework:

DATA

Height (inches)	Mean weight (pounds)
54	92.50
56	111.41
58	122.05
60	124.43
62	130.22
64	134.59
66	140.48
68	146.37
70	157.32
72	163.41
74	179.50

The Relation Between Height and Weight¹²

A report on the heights and weights of approximately 5,000 British women published in 1951, indicated that at that time women who were five feet tall weighed 125 pounds, on the average. The relation between height and weight indicated taller women weighed more than shorter women at an average of approximately 3.4 pounds per inch.³

The detail of the data, graphed in Figure 1, show the average weights at each height and a reference line sketching the approximately linear relation.

The linear relation is a reasonably good predictor of the average weight: The median error, predicting the average, is approximately five pounds — although the average error for predicting the weight of particular women, instead of the average, may be presumed to have been much greater.

However, even with these small errors with respect to the average, the pattern of the errors, shown in Figure 2 indicates a distinct non-linearity. The errors follow a

distinctly “S-shaped” pattern, observed weight is lower than predicted among the shortest women, observed weight is higher than predicted among the tallest women, and observed weight the doubly-bent “S-pattern” in the middle. Were we to exclude both ends of the data, then the average pounds per inch would be estimated at approximately 2.6 pounds per inch. While restricting the range of variation, excluding the very short and the very tall, would improve estimates and change the ratio, to about 2.6 pounds per inch, nevertheless it would not describe the phenomenon because even within this restricted range the deviations show a distinctly non-linear pattern, shown in Figure 3.

While it is clear that there is a non-linearity to these data it is clear that single-bend transformations, like the cube root, would be incapable of explaining the pattern of these deviations. It is also true that the end points of the data are based on relatively few women, 5 women at 54 inches, 4 women at 74 inches, but the regularity, albeit an unexplained regularity of the deviations suggests that the deviations are not the result of random error due to the small number of observations.

¹ Excel spread sheet attached.

² Note: If this seems to bear a distinct resemblance to the previous write-up, there's a reason. Presentation takes time, always more than I expect. The first one took a considerable amount of time, but the second one began by pasting the new data and the new graphs into the old write-up, and then changing what needed to be changed.

³ The data are reproduced in detail in Kendall's *The Advanced Theory of Statistics*, Volume II, pp 300 and 319. The detailed data show not only the average weight for each height but the detailed numbers of women counted at weight ranging from 80.5 pounds to 278.5 pounds.

DATA		COMPUTED VALUES		
Height (inches)	Mean weight (pounds)	Expected Values (in centimeters) (Expected values under the hypothesis that a 60 inch woman weighs 124 pound and that weights deviate from 124 pounds at the rate of 3.4 pounds per inch.)	Error (in centimeters) (Error defined as yield minus expected yield.)	Number of Women
54	92.5	103.6	-11	5
56	111.41	110.4	1.01	33
58	122.05	117.2	4.85	254
60	124.43	124	0.43	813
62	130.22	130.8	-0.6	1,340
64	134.59	137.6	-3	1,454
66	140.48	144.4	-3.9	750
68	146.37	151.2	-4.8	275
70	157.32	158	-0.7	56
72	163.41	164.8	-1.4	11
74	179.5	171.6	7.9	4

Average magnitude (absolute value) of error in centimeters: 3,609 pounds

Table 1
Heights and Average Weights for 4,995 British Women

Source: Reproduced from Kendall's *The Advanced Theory of Statistics, Volume II*, pp 319, reproduced, in turn, from *Women's Measurements and Sizes*, London, H.M.S.P., 1957.

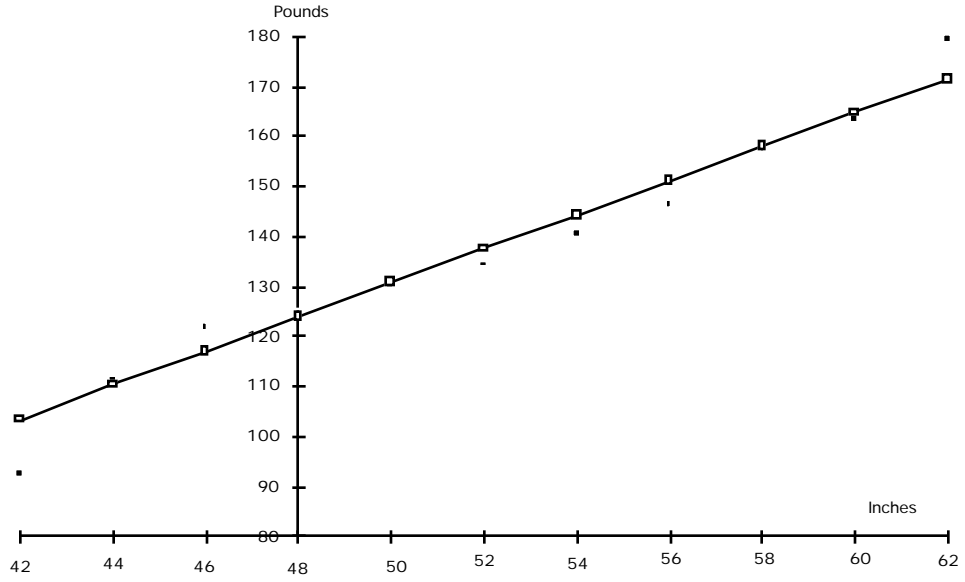


Figure 1
Height and Average Weight for 1495 British Women, circa 1951.

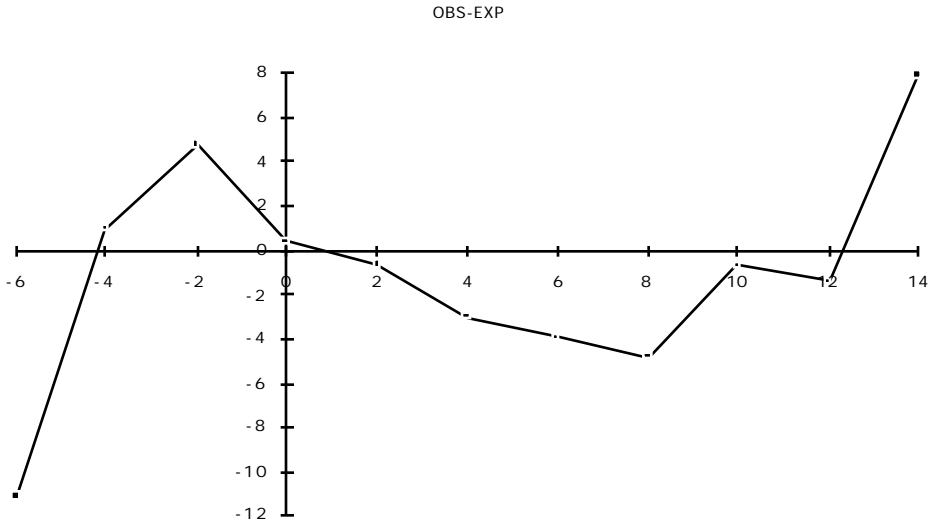


Figure 2
Errors of Prediction, Comparing Observed Weight to Predicted Weight

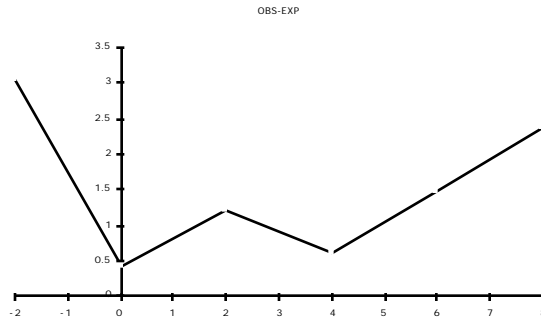


Figure 3
Errors of Prediction, Within the Restricted Middle Range of Heights Demonstrating Persistent Non-Linearity.