The Bell Curve: Intelligence and Class Structure in American Life - Richard J. Herrnstein, Charles Murray (1996)

Appendix 1. Statistics for People Who Are Sure They Can't Learn Statistics

The short explanations of standard deviation (page 44), correlation (page 67), and regression (page 122) should be satisfactory for people who are at home with math but never took a statistics course. The longer explanations in this appendix are for people who would like to understand what *distribution, standard deviation, correlation,* and *regression* mean, but who are not at home with math.

DISTRIBUTIONS AND STANDARD DEVIATIONS

Why Do We Need "Standard Deviation"?

Every day, formally or informally, people make comparisons—among people, among apples and oranges, among dairy cows or egg-laying hens, among the screws being coughed out by a screw machine. The standard deviation is a measure of how spread out the things being compared are. "This egg is a lot bigger than average," a chicken farmer might say. The standard deviation is a way of saying precisely what "a lot" means.

What Is a Frequency Distribution?

To get a clear idea of what a frequency distribution is, imagine yourself back in your high school gym, with all the boys in the senior class in the school gym assembled before you (including both sexes would complicate matters, and the point of this discussion is to keep things simple). Line up these boys from left to right in order of height.

Now you have a long line going from shortest to tallest. As you look along the line you will see that only few boys are conspicuously short and tall. Most are in the middle, and a lot of them seem identical in height. Is there any way to get a better idea of how this pattern looks?

Tape a series of cards to the floor in a straight line from left to right, with "60 inches and shorter" written on the one at the far left, "80 inches and taller" on the card at the far right, and cards in 1-inch increments in between. Tell everyone to stand behind the card that corresponds to his height.

Someone loops a rope over the rafters and pulls you up in the air so you can look straight down on the tops of the heads of your classmates standing in their single files behind the height labels. The figure below shows what you see: a frequency distribution. What good is it? Looking at your high school classmates standing around in a mob, you can tell very little about their height.

Looking at those same classmates arranged into a frequency distribution, you can tell a lot, quickly and memorably.

The raw material of a frequency distribution
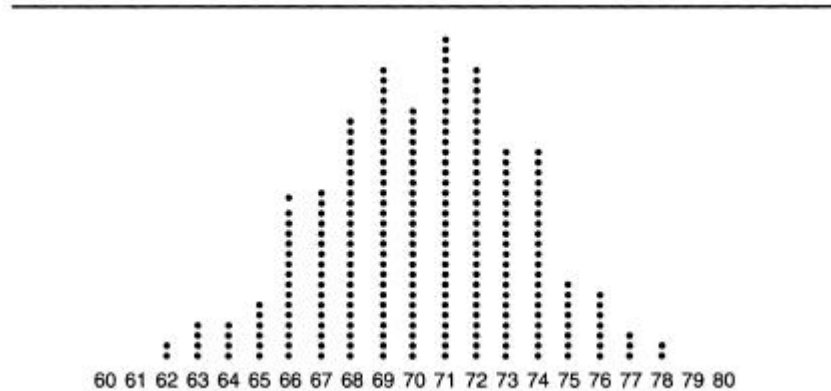
How Is the Distribution Related to the Standard Deviation?

We still lack a convenient way of expressing where people are in that distribution. What does it mean to say that two different students are, say, 6 inches different in height. How "big" is a 6-inch difference? That brings us back to the standard deviation.

When it comes to high school students, you have a good idea of how big a 6-inch difference is. But what does a 6-inch difference mean if you are talking about the height of elephants? About the height of cats? It depends. And the things it depends on are the average height and how much height varies among the things you are measuring. A *standard deviation gives you a way of taking both the average and that variability into account, so that "6 inches" can be expressed in a way that means the same thing for high school students relative to other high school students, elephants relative to other elephants, and cats relative to other cats.*

How Do You Compute a Standard Deviation?

Suppose that your high school class consisted of just two people who were 66 inches and 70 inches. Obviously, the average is 68 inches. Just as obviously, one person is 2 inches shorter than average, one person is 2 inches taller than average. The standard deviation is a kind of average of the differences from the mean—2 inches, in this example. Suppose you add two more people to the class, one who is 64 inches and the other who is 72 inches. The mean hasn't changed (the two new people balance each other off exactly). But the newcomers are each 4 inches different from the average height of 68 inches, so the standard deviation, which measures the spread, has gotten bigger as well. Now two people are 4 inches different from the average and two people are 2 inches different from the average. That adds up to a total 12 inches, divided among four persons. The simple average of these differences from the mean is 3 inches (12 ÷ 4), which is almost (but not quite) what the standard deviation is. To be precise, the standard deviation is calculated by squaring the deviations from the mean, then summing them, then finding their average, then taking the square root of the result. In this example, two people are 4 inches from

2

the mean and two are 2 inches from the mean. The sum of the squared deviations is 40 (16 + 16 + 4 + 4). Their average is 10 (40 + 4). And the square root of 10 is 3.16, which is the standard deviation for this example. The technical reasons for using the standard deviation instead of the simple average of the deviations from the mean are not necessary to go into, except that, in normal distributions, the standard deviation has wonderfully convenient properties. *If you are looking for a short, easy way to think of a standard deviation, view it as the average difference from the mean.*

As an example of how a standard deviation can be used to compare apples and oranges, suppose we are comparing the Olympic women's gymnastics team and NBA basketball teams. You see a woman who is 5 feet 6 inches and a man who is 7 feet. You know from watching gymnastics on television that 5 feet 6 inches is tall for a woman gymnast, and 7 feet is tall even for a basketball player. But you want to do better than a general impression. Just *how* unusual is the woman, compared to the average gymnast on the U.S. women's team, and how unusual is the man, compared to the average basketball player on the U.S. men's team?

We gather data on height among all the women gymnasts, and determine that the mean is 5 feet 1 inches with a standard deviation (SD) of 2 inches. For the men basketball players, we find that the mean is 6 feet 6 inches and the SD is 4 inches. Thus the woman who is 5 feet 6 inches is 2.5 standard deviations taller than the average; the 7-foot man is only 1.5 standard deviations taller than the average. These numbers—2.5 for the woman and 1.5 for the man—are called *standard scores* in statistical jargon. Now we have an explicit numerical way to compare how different the two people are from their respective averages, and we have a basis for concluding that the woman who is 5 feet 6 inches is a lot taller relative to other female Olympic gymnasts than a 7-foot man is relative to other NBA basketball players.

How *Much More Different? Enter the Normal Distribution*

Even before coming to this book, most readers had heard the phrases *normal distribution* or *bell-shaped curve,* or, as in our title, *bell curve.* They refer to a common way that natural phenomena arrange themselves approximately. (The true normal distribution is a mathematical abstraction, never perfectly observed in nature.) If you look again at the distribution of high school boys that opened the discussion, you will see the makings of a bell curve. If we added several thousand more boys to it, the kinks and irregularities would smooth out, and it would actually get very close to a normal distribution. A perfect one is in the figure below.

A perfect bell curve

3
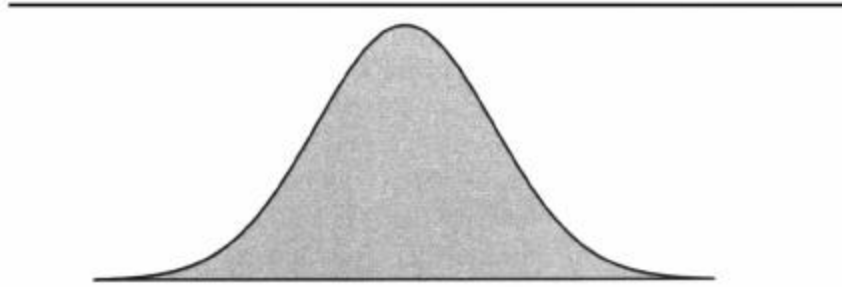
It makes sense that most things will be arranged in bell-shaped curves. Extremes tend to be rarer than the average. If that sounds like a tautology, it is only because bell curves are so common. Consider height again. Seven feet is "extreme" for humans. But if human height were distributed so that equal proportions of people were 5 feet, 6 feet, and 7 feet tall, the extreme would not be rarer than the average. It just so happens that the world hardly ever works that way.

Bell curves (or close approximations to them) are not only common in nature; they have a close mathematical affinity to the meaning of the standard deviation. *In any true normal distribution, no matter whether the elements are the heights of basketball players, the diameters of screw heads, or the milk production of cows, 68.27 percent of all the cases fall in the interval between 1 standard deviation above the mean and 1 standard deviation below it.* It is worth pausing a moment over this link between a relatively simple measure of spread in a distribution and the way things in everyday life vary, for it is one of nature's more remarkable uniformities.

In its mathematical form, the normal distribution extends to infinity in both directions, never quite reaching the horizontal axis. But for practical purposes, when we are talking about populations of people, a normal distribution is about 6 standard deviations wide. The next figure shows how the bell curve looks, cut up into six regions, each marked by a standard deviation unit. The range within ±3 standard deviation units includes 99.7 percent of a population that is distributed normally.

A bell curve cut into standard deviations
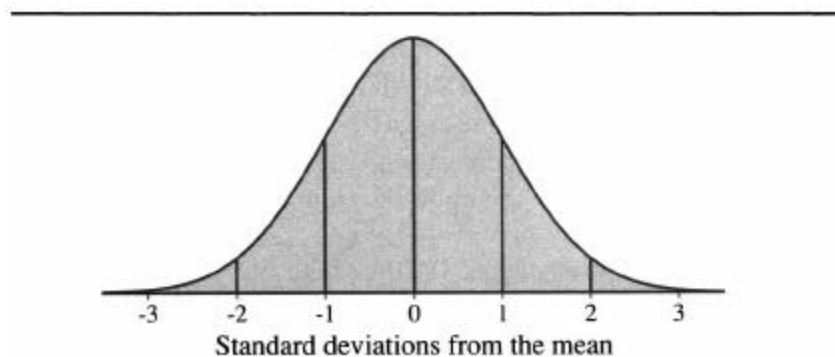
We can squeeze the axis and make it look narrow, or stretch it out and make it look wide, as shown in the following figure. Appearances notwithstanding, the mathematical shape is not

really changing. The standard deviation continues to chop off proportionately the same size chunks of the distribution in each case. And therein lies its value. The standard deviation has the same meaning no matter whether the distribution is tall and skinny or short and wide.

Standard deviations cut off the same portions of the population for any normal distribution
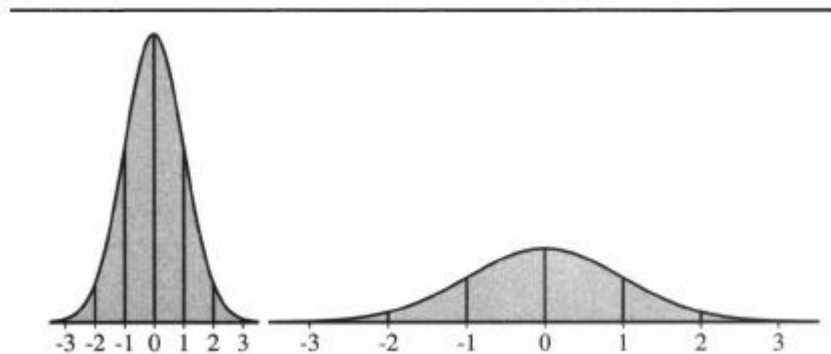
Furthermore, there are some simple characteristics about these scores that make them especially valuable. As you can see by looking at the figures above, it makes intuitive sense to think of a 1 standard deviation difference as "large," a 2 standard deviation difference as "very large," and a 3 standard deviation difference as "huge." This is an easy metric to remember. Specifically, a person who is 1 standard deviation above the mean in IQ is at the 84th percentile. Two standard deviations above the mean puts him at the 98th percentile. Three standard deviations above the mean puts him at the 99.9th percentile. A person who is 1 standard deviation below the mean is at the 16th percentile. Two standard deviations below the mean puts him at the 2d percentile. Three standard deviations below the mean puts him at the 0.1th percentile.

Why Not Just Use Percentiles to Begin With?

Why go to all the trouble of computing standard scores? Most people understand percentiles already. Tell them that someone is at the 84th percentile, and they know right away what you mean. Tell them that he's at the 99th percentile, and they know what that means. Aren't we just introducing an unnecessary complication by talking about "standard scores"?

Thinking in terms of percentiles is convenient and has its legitimate uses. We often speak in terms of percentiles—or centiles—in the text. But they can also be highly misleading, because they are artificially compressed at the tails of the distributions. It is a longer way from, say, the 98th centile to the 99th than from the 50th to the 51st. In a true normal distribution, the distance from the 99th centile to the 100th (or, similarly, from the 1st to the 0th) is infinite.

Consider two people who are at the 50th and 55th centiles in height. Using the NLSY as our estimate of the national American distribution of height, their actual height difference is only half an inch.[2] Consider another two people who are at the 94th and 99th centiles on height—the identical gap in terms of centiles. Their height difference is 3.1 inches, six times the height

difference of those at the 50th and 55th centiles. The further out on the tail of the distribution you move, the more misleading centiles become.

Standard scores reflect these real differences much more accurately than do centiles. The people at the 50th and 55th centiles, only half an inch apart in real height, have standard scores of 0 and .13. Compare that difference of .13 standard deviation to the standard scores of those at the 94th and 99th centiles: 1.55 and 2.33, respectively. In standard scores, their difference—which is .78 standard deviation—is six times as large, reflecting the six-fold difference in inches.

The same logic applies to intelligence test scores, and it explains why they should be analyzed in terms of standard scores, not centiles. There is a lot of difference between people at the 1st centile and the 5th, or between those at the 95th and the 99th, much more than those at the 48th and the 52d. If you doubt this, ask a university teacher to compare the classroom performance of students with an SAT-Verbal of 600 and those with an SAT-Verbal of 800. Both are in the 99th centile of all 18-year-olds—but what a difference in verbal ability![3]

CORRELATION AND REGRESSION

We now need to consider dealing with the relationships between two or more distributions— which is, after all, what scientists usually want to do. How, for example, is the pressure of a gas related to its volume? The answer is Boyle's Law, which you learned in high school science. In social science, the relationships between variables are less clear cut and harder to unearth. We may, for example, be interested in wealth as a variable, but how shall wealth be measured? Yearly income? Yearly income averaged over a period of years? The value of one's savings or possessions? And wealth, compared to many of the other things social science would like to understand, is easy, reducible as it is to dollars and cents.

But beyond the problem of measurement, social science must cope with sheer complexity. Our physical scientist colleagues may not agree, but we believe it is harder to do science on human affairs than on inanimate objects—so hard, in fact, that many people consider it impossible. We do not believe it is impossible, but it is rare that any human or social relationship can be fully captured in terms of a single pair of variables, such as that between the temperature and volume of a gas. In social science, multiple relationships are the rule, not the exception.

For both of these reasons, the relations between social science variables are typically less than perfect. They are often weak and uncertain. But they are nevertheless real, and, with the right methods, they can be rigorously examined.

Correlation and regression, used so often in the text, are the primary ways to quantify weak, uncertain relationships. For that reason, the advances in correlational and regression analysis since the late nineteenth century have provided the impetus to social science. To understand what this kind of analysis is, we need to introduce the idea of a scatter diagram.

Scatter Diagrams

We left your male high school classmates lined up by height, with you looking down from the rafters. Now imagine another row of cards, laid out along the floor at a right angle to the ones for height. This set of cards has weights in pounds on them. Start with 90 pounds for the class shrimp, and in 10-pound increments, continue to add cards until you reach 250 pounds to make room for the class giant. Now ask your classmates to find the point on the floor that corresponds to both their height and weight (perhaps they'll insist on a grid of intersecting lines extending from the two rows of cards). When the traffic on the gym floor ceases, you will see something like the figure below. This is a scatter diagram. Some sort of relationship between height and weight is immediately obvious. The heaviest boys tend to be the tallest, the lightest ones the shortest, and most of them are intermediate in both height and weight. Equally obvious are the deviations from the trend that link height and weight. The stocky boys appear as points above the mass, the skinny ones as points below it. What we need now is some way to quantify both the trend and the exceptions.

A scatter diagram
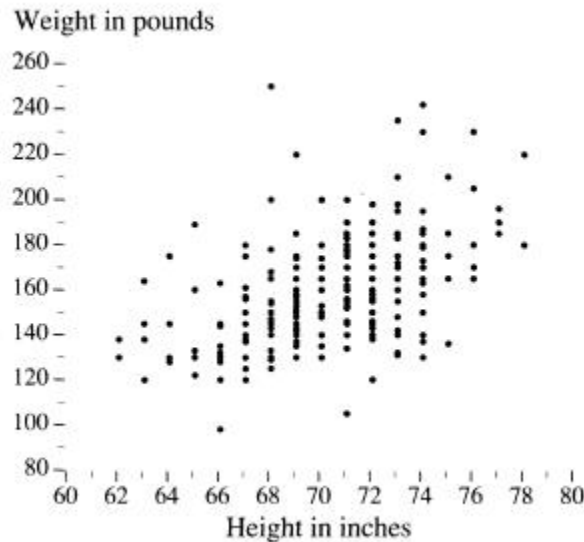
Correlations and *regressions* accomplish this in different ways. But before we go on to discuss these terms, be reassured that they are simple. Look at the scatter diagram. You can see by the dots that as height increases, so does weight, in an irregular way. Take a pencil (literally or imaginarily) and draw a straight, sloping line through the dots in a way that seems to you to best reflect this upward-sloping trend. Now continue to read, and see how well you have intuitively produced the result of a correlation coefficient and a regression coefficient.

The Correlation Coefficient

Modern statistics provides more than one method for measuring correlation, but we confine ourselves to the one that is most important in both use and generality: the Pearson product-moment correlation coefficient (named after Karl Pearson, the English mathematician and biometrician). To get at this coefficient, let us first replot the graph of the class, replacing inches

7

and pounds with standard scores. The variables are now expressed in general terms. Remember: *Any* set of measurements can be transformed similarly.

The next step on our way to the correlation coefficient is to apply a formula (here dispensed with) that, in effect, finds the best possible straight line passing through the cloud of points—the mathematically "best" version of the line you just drew by intuition.

What makes it the "best"? Any line is going to be "wrong" for most of the points. For example, look at the weights of the boys who are 64 inches tall. Any sloping straight line is going to cross somewhere in the middle of those weights and may not cross any of the dots exactly. For boys 64 inches tall, you want the line to cross at the point where the total amount of the error is as small as possible. Taken over all the boys at all the heights, you want a straight line that makes the sum of all the errors for all the heights as small as possible. This "best fit" is shown in the new version of the scatter diagram below, where both height and weight are expressed in standard scores and the mathematical best-fitting line has been superimposed.

The "best-fit" line for a scatter diagram

Weight, expressed in standard scores

Height, expressed in standard scores
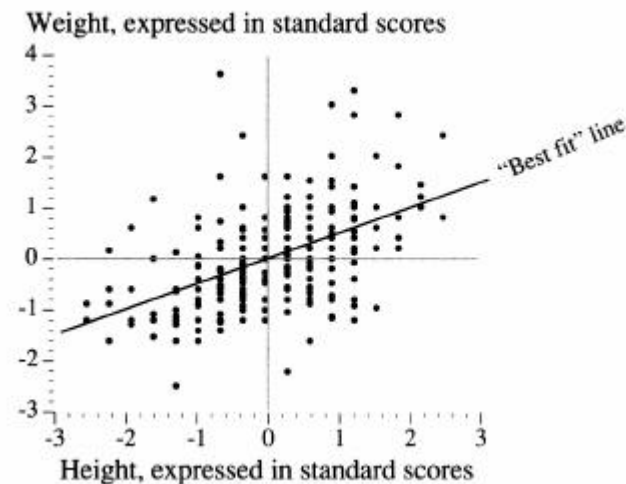
This scatter diagram has (partly by serendipity) many lessons to teach about how statistics relate to the real world. Here are a few of the main ones:

1. *Notice the many exceptions.* There is a statistically substantial relationship between height and weight, but, visually, the exceptions seem to dominate. So too with virtually all statistical relationships in the social sciences, most of which are much weaker than this one.

2. *Linear relationships don't always seem to fit very well.* The best-fit line looks as if it is too shallow. Look at the tall boys, and see how consistently it under-predicts how much they weigh. Given the information in the diagram, this might be an optical illusion—many of the dots in the dense part of the range are on top of each other, as it were, and thus it is impossible to grasp

visually how the errors are adding up—but it could also be that the relationship between height and weight is not linear.

3.   *Small samples have individual anomalies.* Before we jump to the conclusion that the straight line is not a good representation of the relationship, remember that the sample consists of only 250 boys. An anomaly of this particular small sample is that one of the boys in the sample of 250 weighed 250 pounds. Eighteen-year-old boys are very rarely that heavy, judging from the entire NLSY sample, fewer than one per 1,000. And yet one of those rarities happened to be picked up in a sample of 250. That's the way samples work.

4.   *But small samples are also surprisingly accurate, despite their individual anomalies.* The relationship between height and weight shown by the sample of 250 18-year-old males is identical to the third decimal place with the relationship among all 6,068 males in the NLSY sample.[4] This is closer than we have any right to expect, but other random samples of only 250 generally produce correlations that are within a few hundredths of the one produced by the larger sample. (There are mathematics for figuring out what "generally" and "within a few hundredths" mean, but we needn't worry about them here.)

Bearing these basics in mind, let us go back to the sloping line in the figure above. Out of mathematical necessity, we know several things about it. First, it must pass through the intersection of the zeros (which, in standard scores, correspond to the averages) for both height and weight. Second, the line would have had exactly the same slope had height been the vertical axis and weight the horizontal one. Finally, and most significant, the slope of the best-fitting line cannot be steeper than 1.0. The steepest possible best-fitting line, in other words, is one along which one unit of change in height is exactly matched by one unit of change in weight, clearly not the case in these data. Real data in the social sciences never yield a slope that steep.

In the picture, the line goes uphill to the right, but for other pairs of variables, it could go downhill. Consider a scatter diagram for, say, educational level and fertility by the age of 30. Women with more education tend to have fewer babies when they are young, compared to women with less education, as we discuss in Chapters 8 and 15. The cloud of points would decline from left to right, just the reverse of the cloud in the picture above. The downhill slope of the best-fitting line would be expressed as a negative number, but, again, it could be no steeper than—1.0.

We focus on the slope of the best-fitting line because it *is* the correlation coefficient—in this case, equal to .50, which is quite large by the standards of variables used by social scientists. The closer it gets to ±1.0, the stronger is the linear relationship between the standardized variables (the variables expressed as standard scores). When the two variables are mutually independent, the best-fitting line is horizontal; hence its slope is 0. Anything other than 0 signifies a relationship, albeit possibly a very weak one.

Whatever the correlation coefficient of a pair of variables is, squaring it yields another notable number. Squaring .50, for example, gives .25. The significance of the squared correlation is that it tells how much the variation in weight would decrease if we could make everyone the same height, or vice versa. If all the boys in the class were the same height, the variation in their

weights would decline by 25 percent. Perhaps, if you have been compelled to be around social scientists, you have heard the phrase "explains the variance," as in, for example, "Education explains 20 percent of the variance in income." That figure comes from the squared correlation.

In general, the squared correlation is a measure of the mutual redundancy in a pair of variables. If they are highly correlated, they are highly redundant in the sense that knowing the value of one of them places a narrow range of possibilities for the value of the other. If they are uncorrelated or only slightly correlated, knowing the value of one tells us nothing or little about the value of the other.[5]

Regression Coefficients

Correlation assesses the strength of a relationship between variables. But we may want to know more about a relationship than merely its strength. We may want to know what it is. We may want to know how much of an increase in weight, for example, we should anticipate if we compare 66-inch boys with 73-inch boys. Such questions arise naturally if we are trying to explain a particular variable (e.g., annual income) in terms of the effects of another variable (e.g., educational level). How much income is another year of schooling worth? is just the sort of question that social scientists are always trying to answer.

The standard method for answering it is regression analysis, which has an intimate mathematical association with correlational analysis. If we had left the scatter diagram with its original axes— inches and pounds—instead of standardizing them, the slope of the best-fitting line would have been a regression coefficient, rather than a correlation coefficient. The figure below shows the scatter diagram with nonstandardized axes.

What a regression coefficient is telling you
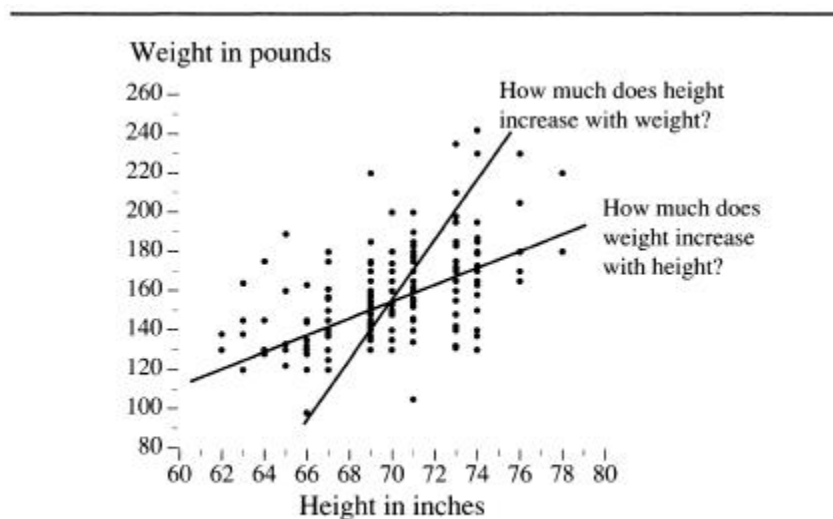
Why are there two lines? Recall that the best-fitting line is the one that minimizes the aggregated distances between the data points and the line. For standardized measurements, it makes no difference whether the distances are measured along the pounds axis or the inches axis; for unstandardized measurements, it may make a difference. Hence we may get two lines, depending on which axis was used to fit the line. The two lines, which always intersect at the average values for the two variables, answer different questions. One answers the question we first posed: How much of a difference in pounds is associated with a given difference in inches (i.e., the regression of weight on height). The other one tells us how much of a difference in inches is associated with a given difference in pounds (i.e., the regression of height on weight).

Multiple Regression

Multiple regression analysis is the main way that social science deals with the multiple relationships that are the rule in social science. To get a fix on multiple regression, let us return to the high school gym for the last time. Your classmates are still scattered about the floor. Now imagine a pole, erected at the intersection of 60 inches and 90 pounds, marked in inches from 18 inches to 50 inches. For some inscrutable reason, you would like to know the impact of both height and weight on a boy's waist size. Since imagination can defy gravity, you ask each boy to levitate until the soles of his shoes are at the elevation-that reads on the pole at the waist size of his trousers. In general, the taller and heavier boys must rise the most, the shorter and slighter ones the least, and most boys, middling in height and weight, will have middling waist sizes as well. Multiple regression is a mathematical procedure for finding that plane, slicing through the space in the gym, that minimizes the aggregated distances (in this instance, along the waist size axis) between the bottoms of the boys' shoes and the plane.

The best-fitting plane will tilt upward toward heavy weights and tall heights. But it may tilt more along the pounds axis than along the inches axis, or vice versa. It may tilt equally for each. The slope of the tilt along each of these axes is again a regression coefficient. With two variables predicting a third, as in this example, there are two coefficients. One of them tells us how much of an increase in trouser waist size is associated with a given increase in weight, holding height constant; the other, how much of an increase in trouser waist size is associated with a given increase in height, holding weight constant.

With two variables predicting a third, we reach the limit of visual imagination. But the principle of multiple regression can be extended to any number of variables. Income, for example, may be related not just to education but also to age, family background, IQ, personality, business conditions, region of the country, and so on. The mathematical procedures will yield coefficients for each of them, indicating again how much of a change in income can be anticipated for a given change in any particular variable, with all the others held constant.

Logistic Regression

The text frequently resorts to a method of analysis called *logistic regression.* Here, we need only say what the method is for rather than what it is. Many of the variables we discuss are such things as being unemployed or not, being married or not, being a parent or not, and so on. Because they are measured in two values—corresponding to yes and no—they are called binary

variables. Logistic regression is an adaptation of ordinary regression analysis tailored to the case of binary variables. (It can also be used for variables with larger numbers of discrete values.) It tells us how much change there is in the probability of being unemployed, married, and so forth, given a unit change in any given variable, holding all other variables in the analysis constant.