POLI 300 ***Handout #*13** N. R. Miller

CORRELATION AND REGRESSION

Recall that sentence #11 in Problem Set #3A (and #9) was:

“*If you want to get ahead, stay in school.*”

Underlying this nagging parental advice is the following claimed empirical relationship:

LEVEL OF EDUCATION $\overset{+}{\text{============>}}$ LEVEL OF SUCCESS IN LIFE

Suppose we measure the independent variable through a survey question asking respondents (say a representative sample of the population aged 35-55) to report the number of years of formal EDUCATION they completed and also their current INCOME as an *indicator* of SUCCESS. We then analyze the association between the two interval variables in this reformulated hypothesis.

LEVEL OF EDUCATION $\overset{+}{\text{===============>}}$ LEVEL OF INCOME
(# of years) ($000 per year)

We next collect appropriate data in two rather different societies A and B and plot each data set producing the scattergrams shown in Figures 1A and 1B.

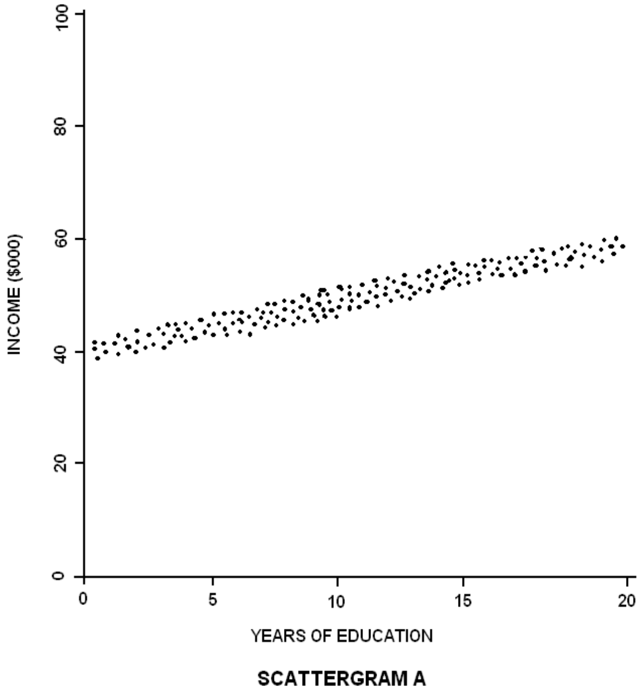


Figure 1A

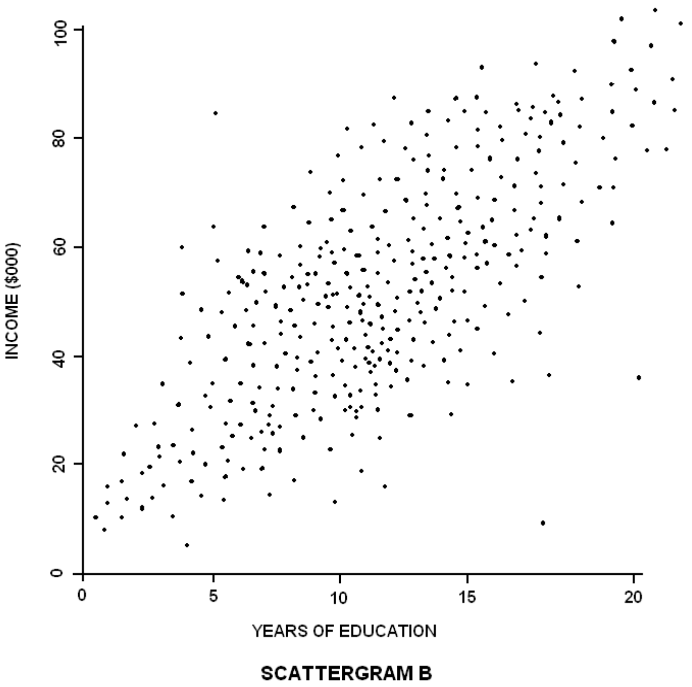


Figure 1B

Note that we have drawn the two scattergrams on two worksheets drawn on the same horizontal and vertical scales. (With respect to Figure 1A, this violates one of the usual guideline for constructing scattergrams — namely that the scales be chosen that minimize "white space" — but here we want to facilitate comparison between the two charts.) Note further that the scattergrams are similar in that both display clear positive associations between the two variables, i.e., the plotted points in both show an upward-sloping pattern running from Low – Low to High – High. But at the same time there are obvious differences between the two scattergrams (and thus between societies A and B).

In class, we will discuss the following questions.

(1) In which society, A or B, is the hypothesis most powerfully confirmed?

(2) In which society, A or B, is their a greater incentive for people to stay in school?

(3) Which society, A or B, does the U.S. more closely resemble?

(4) How might we characterize the differences between societies A and B?

We can visually compare and contrast the nature of the associations between the two variables in the two scattergrams by doing the following. We draw a number of vertical strips in each scattergram (in the manner of the SON'S HEIGHT by FATHER'S HEIGHT scattergram previously discussed). Recall that, within each vertical strip, we have cases with just about the same value on the independent (horizontal) variable — in this case, with just about the same level of EDUCATION. Within each strip, we can readily estimate by "eyeball methods" the *average* magnitude of the dependent (vertical) variable INCOME and, if we put a mark in each strip at the average level of income, we can connect them into a *line of averages* that is apparently (close to being) a straight line. Now we should also take account of a second characteristic of the points within each vertical strip — namely, the magnitude of dispersion around the average. Thus we can assess two distinct characteristics of the relationship between EDUCATION and INCOME:

(i) how much the average level of INCOME changes among people with *different levels* of education; and

(ii) how great a degree of dispersion of INCOME there is among people with the *same level* of EDUCATION.

Having done this, we can pin down two basic differences between the scattergrams for A and B.

In both scattergams, the line of averages is upward-sloping, indicating a clear apparent positive effect on EDUCATION on INCOME. But in the scattergram for society A, the upward slope of the line of averages is fairly shallow. The line of averages in Figure 1A indicates that average INCOME increases by only about $1000 for each additional year of EDUCATION. On the other hand, in the scattergram for society B, the upward slope of the line of averages is much steeper. The graph in Figure 1B indicates that average INCOME increases by about $4000 for each additional year of EDUCATION. In this sense, EDUCATION is on average more "rewarding" in society B than A.

But there is another difference between the two scattergrams. In Figure 1A, there is almost no dispersion within each vertical strip and thus almost no dispersion around the line of averages as a whole. In Figure 1B on the other hand, there is a lot of dispersion within each vertical strip and also thus around the line of averages as a whole. Put more substantively, in society A additional years of EDUCATION produce *rewards* in terms of INCOME that are *modest* (as seen above), *but the modest rewards are essentially certain*; on the other hand, in society B additional years of EDUCATION produce *on average much more substantial rewards* in terms of INCOME (as seen above), *but these possibly very large rewards are quite uncertain* and are indeed realized only *on average*. (Thus in Figure 1B, but not 1A, we can find lots of pairs of cases such that one case has much higher EDUCATION but the other case has much higher INCOME.) This implies that in society B, while EDUCATION has a big impact on EDUCATION, there are evidently *other* (independent) *variables* (such family wealth, ambition, career choice, just plain luck, etc.) that *also have major effects* on LEVEL OF INCOME. In contrast, in society A it appears that LEVEL OF EDUCATION (almost) wholly determines LEVEL OF INCOME and that essentially nothing else matters. (Another difference between the two societies is that, while both societies have similar distributions of EDUCATION, their INCOME distributions are quite different: A is relatively egalitarian with respect to INCOME, which ranges only from about $40,000 to about $60,000, while B is considerably less egalitarian with respect to INCOME, which ranges from under to $10,000 to at least $100,000 — and possibly much higher.)

In summary, in society A the INCOME rewards of EDUCATION are *modest but essentially certain*, while in society B the INCOME rewards of EDUCATION are *substantial on average but quite uncertain*.

More generally we see that, given bivariate data for interval variables, the "strength of association" between them can mean either of two quite different things: (i) the independent variable has a *very reliable or certain association* with the dependent variable, as is true for society A but not B, or (ii) the independent variable has on average a *big impact* on the dependent variable, as is true for society B but not A. There are two bivariate summary statistics that capture these two different kinds of strength of association: the first is called the *regression coefficient*, customarily designated *b*, and the second is called the *correlation coefficient*, customarily designated *r*. In Figures 1A and 1B, A has the greater correlation coefficient and B has the greater regression coefficient.

Review: The Equation of a Line

Having drawn (by "eyeball methods") a line of averages in each scattergram, it is convenient to write an *equation* for each line. You should recall from high-school algebra that, given any graph with a horizontal axis X and a vertical axis Y, any straight line drawn on the graph has an equation of the form (using the symbols you probably used in highschool)

$$y = m \times x + b ,$$

where m is the slope of the line expressed as $\Delta\, y \,/\, \Delta\, x$ (i.e., "change in y" divided by "change in x" or "rise over run") and b is the y-intercept (i.e., the value of y when $x = 0$). Evidently to further torment students, in college statistics the symbol b is used in place of m to represent the slope of the line and the symbol a is used in place of b to represent the intercept, and the equation for a straight line is usually written as

$$y = a + b \times x \, .$$

The equation for the line of averages in Figure 1A appears to be approximately

$$\text{AVERAGE INCOME} = \$40{,}000 + \$1000 \times \text{EDUCATION} \, ,$$

while the equation for the line of averages in Figure B appears to be approximately

$$\text{AVERAGE INCOME} = \$10{,}000 + \$4000 \times \text{EDUCATION} \, .$$

Given such an equation (or "formula"), we can take any value for the variable EDUCATION, plug it into the appropriate formula above, and calculate or "predict" the corresponding average or "expected" value of INCOME. In society A, such a prediction is likely to be highly reliable, because the correlation between the two variables is almost perfect; but in society B, this prediction will be a lot less reliable, because the correlation between the two variables is much less than perfect.

In like manner, we can use eyeball methods to draw the line of averages in the SON'S HEIGHT by FATHER'S HEIGHT scattergram and then write its equation of the form

$$\text{SON'S HEIGHT} = a + b \times \text{FATHER'S HEIGHT} \, .$$

(We will do this as an exercise is class, eyeball estimating the values of b and a.)

The line of averages that we have been discussing is essentially the same as what we will call in the next section the *regression line*, and the slope b of the line of averages is essentially the same as the *regression coefficient*.

Correlation and Regression

A weakness of the discussion so far is that it is based entirely on approximations based on "eyeball methods." Understandably statisticians want to have precise formulas based on numerical data for the regression coefficient b and the correlation coefficient r. Beyond this, eyeball methods simply won't work if we have only a small number of cases and/or the data does not form a nice pattern in a scattergram. Consider the scattergram in Figure 2, which displays the bivariate numerical data for x and y in the Sample Problem presented on p. 9. (Note that, in this problem, $\bar{x} = 4$ and $\bar{y} = 5$.) The "vertical strips" kind of argument simply will not work, because most strips include no data points and only one strip (for $x \approx 5$) includes more than one point). Using this simple example, we will now proceed a little more formally. (The following exercise will be carried out step by step in class.)

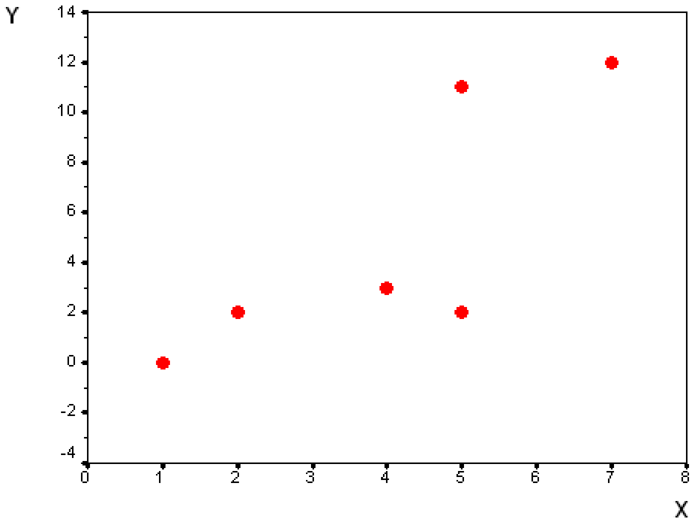


Figure 2

Suppose we were to ask what horizontal line (i.e., a line for which the slope b = 0, so its equation is simply $y = a$) in Figure 2 would "best fit" (come as close as possible to) the plotted points. In order to answer this question, we must have a specific criterion for "best fitting." The one statisticians use is the *least squares criterion*. That is, we want to find the horizontal line such that the *squared vertical deviations* from the line to each point are (in total or on average) *as small as possible*. Now you should know what line this is — it is the line $y = \bar{y}$, or in this case $y = 5$. (Recall from Handout #6, p. 4, point (e) that "the *sum* (or mean) of the *squared deviations from the mean*, i.e., $\sum(y - \bar{y})^2$, is *less* than the sum of the squared deviations from any other value of the variable.")

Now suppose that we are no longer restricted to horizontal lines but can tip the line up or down — more particularly, that we can pivot the straight line on the point $(\bar{x}, \bar{y}) = (4, 5)$ until it has a slope that best fits all the plotted points by the same least squares criterion. When we have found that line, we have found the regression line. A statistical theorem tell us that the regression line goes through the point $(\bar{x}, \bar{y})$ and others give us formulas to calculate the slope and intercept as well as the correlation coefficient. (The correlation coefficient r [actually r^2; see below] measures how much the goodness of fit by the least squares criterion is improved when we allow the line to tip so that it has a [positive or negative] slope.)

How to Compute and Interpret Correlation and Regression Coefficients

1. For purposes of this discussion, we call the *independent* variable X and the *dependent* variable Y. (This usage is standard.)

2. Set up a worksheet like the one on p. 9 of this Handout.

3. Compute the following *univariate* statistics: the mean of X, the mean of Y, the variance of X, the variance of Y, the standard deviation of X, and the standard deviation of Y.

4. As shown in the last column of the worksheet, for each case, multiply its deviation from the mean with respect to *X* by its deviation from the mean with respect to *Y*. This is called the *crossproduct* of the deviations. Notice that a crossproduct is *positive* if, in that case, the *X* and *Y* values both deviate *in the same direction* (i.e., both upwards or both downwards) from their respective means; and it is *negative* if, in that case, the *X* and *Y* values deviate *in opposite directions* (i.e., one upwards and one downwards) from their respective means. In either event, the [absolute] crossproduct is large if both [absolute] deviations are large and small if either deviation is small. Thus:

   (a) if there is a positive relationship between the two variables, most crossproducts are positive and many are large;

   (b) if there is a negative relationship between the two variables, most crossproducts are negative and many are large; and

   (c) if the two variables are unrelated, the crossproducts are positive in about half the cases and negative in about half the cases.

5. Add up the crossproducts over all cases. The *sum* (and thus the average) *of crossproducts* is *positive* in the event of (a) above, *negative* in the event of (b) above, and *close to zero* in the event of (c) above.

6. Divide the sum of the crossproducts by the number of cases to get the *average* (mean) *crossproduct*. This average is called the *covariance* of *X* and *Y*, and its formula is the *bivariate* parallel to the *univariate variance* (of *X* or *Y*). Notice that:

   (a) if the relationship between *X* and *Y* is positive, their covariance is positive (because most crossproducts are positive);

   (b) if the relationship between *X* and *Y* is negative, their covariance is negative (because most crossproducts are negative); and

   (c) if there is no relationship between *X* and *Y* , their covariance is (approximately) zero (because positive and negative crossproducts [approximately] cancel out).

   Thus the covariance measures the *association* between the two variables. However, it is not a (fully) *valid* measure of the association, because the magnitude of the average (positive or negative) crossproduct reflects not only *how closely the two variables are associated* but also on *the magnitude of their dispersions* (as indicated by their standard deviations or variances). Two very closely (and positively) associated variables have a positive but small covariance if they both have small standard deviations; two not so closely (but still positively) associated variables have a larger covariance if their standard deviations are sufficiently larger.

7. The *correlation coefficient* is a measure of association *only*, which (like other measures of association) is *standardized* so that it always falls in the range from −1 to +1. This is accomplished by dividing the covariance by the standard deviations of each variable. (This is equivalent to finding the average crossproduct when crossproducts are calculated not in terms of *X*-units and *Y*-units but in terms of *standard scores* for both *X* and *Y*.)

8. Divide the covariance of *X* and *Y* by the standard deviation of *X* and also by the standard deviation of *Y*. This gives the *correlation coefficient r*, which measures the *degree* (or "completeness") and *direction* of association between two interval variables *X* and *Y*.

$$\textbf{\textit{Correlation coefficient}} = r = \frac{\text{Cov }(X,Y)}{\text{SD}(X) \times \text{SD}(Y)}$$

If you calculate *r* to be greater than +1 or less than −1, you have made a mistake. It is a good idea first to construct a scattergram of the data on *X* and *Y* and then to check whether your calculated correlation coefficient looks plausible in light of the scattergram.

9. Observe that the correlation coefficient is a ratio of "*X*-units × *Y*-units" (i.e., the respective deviations from the mean) divided by "*X*-units × *Y*-units" (i.e., the respective SDs). This means that all units of measurement cancel out and the correlation coefficient a *pure number* that is *not expressed in any units*. For example, suppose the correlation between the height (measured in inches) and weight (measured in pounds) of students in the class is $r = +.45$. This is *not* $r = +.45$ *inches* or $r = +.45$ *pounds*, or $r = +.45$ *pounds per inch*, etc. — it is just $r = +.45$. Moreover, *the magnitude of the correlation coefficient is independent of the units used to measure the two variables*. If we measured students' heights in feet, meters, etc. (instead of inches) and/or measured their weights in ounces, kilograms, etc. (instead of pounds), and then calculated the correlation coefficient based on the new numbers, it would be just the same as before, i.e., $r = +.45$. In addition, if you check the formula above, you can see that *r* is unchanged if we interchange *X* and *Y*. Thus, *the correlation between two variables is the same regardless of which variable is considered to be independent and which dependent*.

10. To compute the regression coefficient *b*, divide the covariance of *X* and *Y* by the *variance* of the *independent variable X*.

$$\textbf{\textit{Regression coefficient}} = b = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

Remember that the regression coefficient answers this question: ***how much, on average, does the dependent variable increase when the independent variable increases by one unit***. Thus the *magnitude of the regression coefficient* clearly *depends on which variable is considered to be independent and which dependent*, and it also depends *on the units in which each variable is measured*. Observe that the regression coefficient is a ratio of "*X*-units × *Y*-units" (i.e., the respective deviations from the mean) divided by "*X*-units × *X*-units" (i.e., the variance of *X*). Thus *the regression coefficient is expressed in units of the dependent variable per unit of the independent variable* — that is, it is a *rate*, something like "miles per hour" [rate of speed] or "miles per gallon" [level of fuel efficiency], and its numerical value changes as these units change. For example, suppose we measure the height in inches and weight in pounds of all students in the class, suppose we treat height as the independent variable, and suppose the regression coefficient is $b = +6$. This means +6 *pounds* (dependent variable units) *per inch* (independent variable units) — that is, if we lined students up in order from short to tall, we would observe that their weights increase, on average, by 6 pounds for each additional inch of height. This also means that students' weights would increase, on average, by about 2.7 kilograms (since there are about .45 kilograms to a pound) for each additional inch of height. So if weight were measured in kilograms instead of pounds, the regression coefficient would be $b = +2.7$ *kilograms per inch*. Likewise if we measured weight in pounds but height in feet, the regression coefficient would be about $b = +6 \times 12 = 72$ *pounds per foot*. (As we saw before the correlation coefficient is unchanged by such changes in units.) Moreover, if we took weight to be the independent variable and height the dependent (i.e., if we lined students up in order of their weights and observed how their heights varied with weight), the regression coefficient would be telling us how much students' heights increase, on average, in some unit of height (inch, meter, etc.) as their weights increase by one unit (pound, kilogram, or whatever), so it would be yet a different number, e.g., perhaps about $b = +0.1$ *inches per pound*. (As we also saw before, the correlation coefficient is unchanged by reversing the independent and dependent variables.)

11. To specify the regression line of the relationship between independent variable *X* and dependent variable *Y*, we need to know, in addition to the *slope* of the regression line (i.e., the regression coefficient *b*), how *high or low* the line with that slope lies in the scattergram. (We can draw an infinite number of lines that are distinct from the regression line but parallel to it, i.e., that have the same slope.) This is specified by the *intercept a*, i.e., by specifying where the regression line passes through the vertical axis representing values of the dependent variable *Y when the vertical Y axis intersects the horizontal X axis at the point* $x = 0$ *and* $y = 0$. This *intercept a* is equal to *the mean of Y minus b times the mean of X*.

$$\textbf{\textit{Intercept}} \text{ (or } \textbf{\textit{constant}}\text{)} = a = \bar{y} - b \times \bar{x}$$

The value of the intercept answers this (perhaps quite artificial) question: ***what is the average (or expected or predicted) value of Y when X is zero***. Using *a* and *b* together, we can answer this question: what is the expected or predicted value of *Y* when *X* is *any* specified value. The expected or predicted value $\hat{y}$ of *Y*, given that *X* is some particular value *x*, is given by the *regression equation* (i.e., the equation of the regression line):

$$\hat{y} = a + b \times x.$$

12. You will find other formulas for the correlation and regression coefficients in textbooks. (For a horrendous looking example, see Weisberg *et al*., near top of p. 305.) Such formulas are mathematically equivalent to (i.e., give the same answers as) the formulas given here. Though formidable looking, such formulas are actually easier to work with if you are processing many cases or programming a calculator or computer, because they require you (or the computer) to pass through the data only once. But the formulas presented here make more intuitive sense and are easy enough to use in the simple sorts of problems that you will be presented with in problems sets and tests.

13. Note from the formulas above that the sign (i.e., "+" or "−") of *b* and *r* are both determined by the sign of cov(*X*,*Y*), from which it follows that *b* and *r* themselves always have the same sign (and, if one is zero, the other is also zero). Notice also that the regression and correlation coefficients are equal in the event the independent and dependent variables have

the same dispersion. For example, in the SON'S HEIGHT by FATHER'S HEIGHT scattergram, by eyeball methods we can determine the regression coefficient *b* is somewhere between about +0.4 and +0.5, e.g., the sons of 6' (72") fathers are on average about 5 to 6 inches taller than the sons of 5' (60") fathers. It is also apparent, both from common observation and examination of the scattergram, that the dispersions (SDs) of SON's HEIGHT and FATHER's HEIGHT are just about the same, so we also know that the correlation coefficient is also somewhere between +0.4 and +0.5. In general,

$$b = r \times \frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)}$$

Correlation and Regression Worksheet and Sample Problem

	Raw Data						
	(IND)	(DEP)					
Case #	X	Y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	1	0	−3	−5	9	25	+15
2	2	2	−2	−3	4	9	+6
3	4	3	0	−2	0	4	0
4	5	2	+1	−3	1	9	−3
5	5	11	+1	+6	1	36	+6
6	7	12	+3	+7	9	49	+21
Sum	24	30	0	0	24	132	+45
Average	4 [= $\bar{x}$]	5 [= $\bar{y}$]			4 [=Var(X)]	22 [=Var(Y)]	+7.5 [= Cov(X,Y)]
SQRT(Average)					2 [=SD(X)]	√22 ≈ 4.7 [= SD(Y)]	

Correlation coefficient $= r = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)} = \dfrac{+7.5}{2\sqrt{22}} = +0.7995$

Regression coefficient $= b = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \dfrac{+7.5}{4} = +1.875$

Intercept (constant) $= a = \bar{y} - b \times \bar{x} = 5 - (1.875) \times 4 = -2.5$

Regression equation: $y = a + b \times x = -2.5 + (1.875)\,x$,

so for example if $x = 3.5$, the expected/predicted value of y is:

$$y = -2.5 + (1.875) \times 3.5 = 4.0625$$

The Coefficient of Determination r^2

For various reasons, the squared correlation coefficient r^2 is often reported in bivariate analysis. To compute r^2, just multiply the correlation coefficient by itself. This results in a number that (i) always lies between 0 and 1 (i.e., is never negative and so does *not* indicate the *direction* of association), and (ii) is closer to 0 than r is for all intermediate values of r — for example, $(\pm 0.45)^2 = +0.2025$.

There are three reasons for reporting r^2, sometimes called the *coefficient of determination*. First, a scattergram with $r \approx +0.5$ does not appear to be "halfway between" a scattergram with $r \approx +1$ and one with $r = 0$ — it looks closer to the one with zero association. The scattergram that looks "halfway in between" perfect and zero association has a correlation of about $r \approx +0.7$ (and $r^2 \approx 0.5$).

Second, r^2 has a specific interpretation that r itself lacks. Refer again to Figure 2. Recall that the line $y = \bar{y} = 5$ is the horizontal line that best fits the plotted points by the least squares criterion — that is, the average squared deviation from this line is less than the average squared deviation from any other horizontal line. Remember also that the quantity "average squared deviation from the line $y = \bar{y}$" has a special and familiar name — it is the variance of the dependent variable Y (the square root of which is the SD of Y.)

When we tip the line, we can almost always improve the fit at least a bit a bit. The regression line is the tipped line with the best fit according to the least squares criterion. As we saw the regression line in Figure 2 is the line with the equation $\hat{y} = -2.5 + (1.875) \times x$. But even this line fits the points far from perfectly. For each plotted point, there is some vertical distance (positive if the point lies above the line, negative if it lies below) between the regression line and the point, which is called the *residual* for that case. These residuals are the errors in prediction that are "left over" after we use the regression line to predict the value of the dependent variable in each case.

The ratio of the average squared residuals divided by the variance of Y can be characterized as the proportion of the variance in Y that is not "predicted" or "explained" by the regression equation that has X as the independent variable. Therefore 1 minus this ratio can be characterized as the proportion of the variance in Y that is "predicted" or "explained by" the regression equation., and it turns out that the latter proportion is exactly r^2.

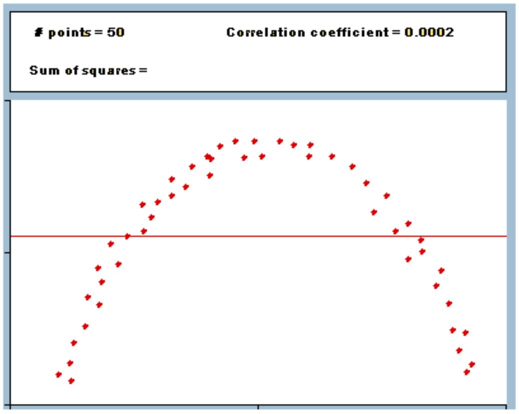
In the example presented above and in Figure 2, the average of the squared residuals is approximately 8, so we can explain approximately $1 - (8/22) = 0.64$ of the variance of Y by the regression equation, and $r^2 = (+0.7995)^2 = 0.64$.

Finally, serious regression analysis is almost always *multiple* (multivariate) *regression*, where the effects of multiple independent (and/or "control") variables on a single dependent variable are analyzed. In this case, we want some summary measure of the overall extent to which the set of all independent variables explains variation in the dependent variable, regardless of whether individual independent variables have positive or negative effects (i.e., regardless of whether bivariate correlations are positive or negative). The coefficient of determination r^2 provides this measure.

Look at the Scattergram Underlying a Correlation Coefficient

It always a good idea to look at the scattergram of bivariate data — not just at the correlation (or regression) coefficient.

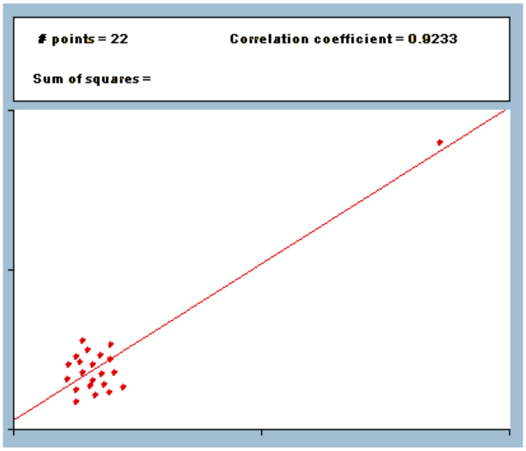
Correlation Reflects Linear Association Only. Suppose that you calculate a correlation coefficient and find that $r \approx 0$ (so $b \approx 0$ also). You should not jump to the conclusion that there is no association of any kind between the variables. The zero coefficient tells you that there is no *linear* (straight-line) *association* between the variables, but there may be a very clear *curvilinear* (curved-line) association between them. See the adjacent scattergram.



points = 50 Correlation coefficient = 0.0002

Sum of squares =

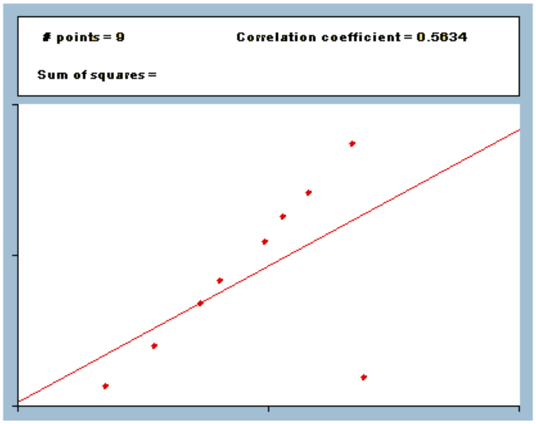
A Univariate Outlier May Create a Correlation. Suppose that you calculate a correlation coefficient and find that $r \approx +0.9$. You should not jump to the conclusion that there is a strong and reliable association between the variables. See the adjacent scattergram, in which one univariate outlier produces the high correlation by itself. You should at least double check your data for this case. Perhaps you inadvertently shifted the decimal point one place to the right when you typed its x and y values into the SPSS data editor (or whatever), so both are ten times bigger than they should be — and also about ten times larger than the values for all other case. Correct this mistake (or delete the outlier) and r goes to zero.



points = 22 Correlation coefficient = 0.9233

Sum of squares =

A Bivariate Outlier May Greatly Attenuate a Correlation. First note that in the adjacent scattergram, the outlier is *not* an outlier in this *univariate* sense; its value on each variable separately is unexceptional. What is exceptional is the *combination* of values on the *two* variables that it exhibits, i.e., it is a *bivariate* outlier.) Of course, clerical errors (or deviant cases) can attenuate as well as enhance apparent correlation. In the adjacent scattergram, a single bivariate outlier reduces what is otherwise an almost perfect association between the variables to a more modest level.



points = 9 Correlation coefficient = 0.5634

Sum of squares =

You can use the Correlation/Regression “Statistical Applet”on the course website to create a scattergram and then to add or delete points by simple pointing and clicking and can then observe the effect on the regression least-squares line and the correlation coefficient. (The applet was used to create the charts on the previous page.)

Applied Regression (and Correlation) Analysis

Regression (especially multiple regression) analysis is now very commonly used in quantitative political science research. But perhaps its application is most intuitive in practical situations in which researchers literally want to make predictions about future cases based on analysis of past cases. Here are two examples.

Predicting College GPAs. A college Admissions Office has recorded the combined SAT scores of all of its incoming students over a number of years. It has also recorded the first-year college GPAs of the same students. The Admissions Office can therefore calculate the regression coefficient *b*, the intercept *a*, and the correlation coefficient *r* for the data they have collected in the past. It can then use the regression equation

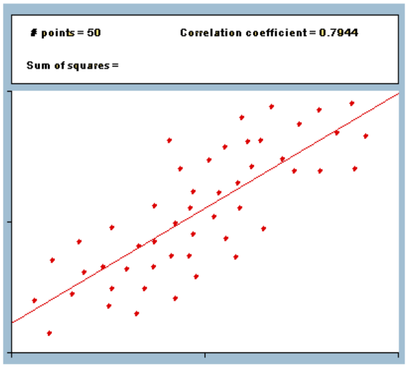
$$\text{PREDICTED COLLEGE GPA} = a + b \times \text{SAT SCORE}$$

to predict the potential college GPAs of the next crop of applicants on the basis of their SAT scores (and use these predictions in making their admissions decisions). Even better, it can collect and record more extensive data and use a *multiple regression* equation that use more than one independent variable (e.g., separates Verbal and Mathematical parts of the SAT, uses highschool GPAs, AP tests passed, etc.) such as:

$$\text{PREDICTED GPA} = a + b_1 \times \text{SATV} + b_2 \times \text{SATM} + b_3 \times \text{HSGPA} + b_4 \times \#\ \text{AP} + \ldots$$

The college admissions problem illustrates another statistical phenomenon to be aware of. Suppose the scattergram in Figure 3 is representative of the general association between SAT scores (horizontal variable) and college GPAs (vertical variable) in the population of all college applicants admitted by any college. It shows that, in general, SATs predict college GPAs quite well with *r* ≈+.8. (In the real world the correlation is not nearly so strong.) Suppose that a highly selective college accepts only applicants with SAT scores above the cutoff point shown in Figure 4. Note that the correlation between SATs and GPAs within this select population is zero. Given Figure 4, critics might say that the college should not rely on SAT score for making admissions decisions because, among those it does admit, there is no correlation between SATs and GPAs. However, this zero correlation is a simple consequence of the college's selective admissions. More generally, if we select cases that have a limited range of values the independent variable, correlation with the dependent variable will be greatly attenuated and may even disappear.

Figure 3



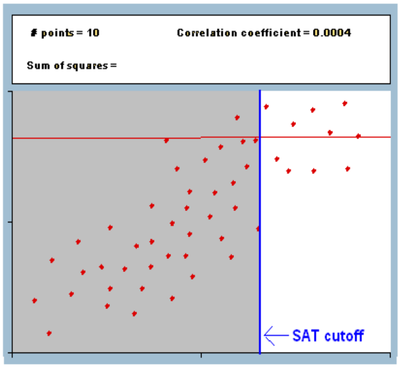


Figure 4

Predicting Presidential Elections Months in Advance. A number of political scientists have devised predictive models that you will likely hear a good deal about during the next Presidential election year. Much like the hypothetical college Admissions Office, these political scientists have assembled data concerning the past 15 or so Presidential elections — in particular, the percent of the popular vote won by the candidate of the incumbent party (that controlled the White House going into the election), which is the dependent variable of interest, and also data on a number of independent variables whose values become available well in advance of the election. The latter include in particular one or more indicators of the state of the economy (growth rate, unemployment rate, etc.) usually as of the end of the second quarter (June 30) of the election year and some poll measure of the incumbent President's approval rating as of about June 30. Using this data, they calculate the coefficients for the regression equation. Come June 30 of the next election year, they plug in the values of their independent variables and calculate (and announce to the public months prior to the election) the predicted value of the dependent variable, i.e., the percent of the popular vote that they expect will be won in the upcoming election by the candidate of the incumbent party. Such predictive models on average predict the ultimate election outcome much more accurately than pre-elections polls conducted at the same time — that is, they capture information about voting intentions in the aggregate that even the many people who actually form the voting intentions and cast votes don't know at the time.

Recall that you constructed scattergrams along these lines in PS #11, Questions 1nd 3. Incidentally, in 2000 virtually all such models in 2000 predicted that Gore would win the election, which — though you may have forgotten — he did (since the model predicts popular, not electoral, votes). However, the predictive models did fail in that almost all predicted that Gore would win comfortably (by 5-10 percentage points, not by a mere half a percentage point). In 2004, almost all predictive models predicted that Bush would win, in most cases by a larger margin than he actually won by.

Empirical Appendix on INCOME by EDUCATION

In the first section of this handout, we focused on two scattergrams, for two different hypothetical data sets (and two hypothetical societies A and B), showing different patterns of association between EDUCATION (measured in terms of *years of formal education*) and INCOME (as an indicator of SUCCESS and measured in terms of *annual income in dollars).* In society A, INCOME is almost wholly determined by EDUCATION ($r \approx +1$) but, at the same time EDUCATION does not have a big impact on INCOME ($b \approx$ \$1000 per year of education). In society B, EDUCATION has a much bigger impact on INCOME ($b \approx$ \$4000 per year of education), but many other factors appear to influence INCOME as well ($r \approx +0.5$). Here is a crosstabulation of INCOME and EDUCATION based on the data for the SETUPS based on the 2004 ANES. (The variables correspond to VG62 and V65 in your SETUPS Codebook, except that INCOME is recorded in class intervals of actual dollar income, rather than percentile groups.)

V65D DOLLAR INCOME (2004) * V62 EDUCATION Crosstabulation (2004)

	V62 EDUCATION				
V65D INCOME	not a HS grad	HS grad only	some college	college grad	Total
Less than $15,000	38.9%	12.8%	9.8%	4.3%	13.7%
$15,000 to $25,000	21.6%	15.5%	10.1%	3.3%	11.5%
$25,000 to $35,000	13.2%	12.8%	10.1%	3.9%	9.6%
$35,000 to $50,000	12.0%	16.6%	17.5%	11.8%	14.7%
$50,000 to $80,000	9.0%	27.0%	24.5%	26.9%	23.4%
$80,000 to $120,000	2.4%	9.1%	20.6%	24.9%	15.7%
More than $120,000	3.0%	6.1%	7.3%	24.9%	11.4%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

What features of this table indicates that the correlation between INCOME and EDUCATION, though positive and substantial, is clearly much less than +1?

Based on this crosstabulation, how might we produce a (rough) estimate of the actual regression coefficient, i.e., how might we determine **how much, on average, income increases for each additional year of education**?