Bivariate Regression: Fitting A Straight Line

In: Applied Regression

By: Michael S. Lewis-Beck

Pub. Date: 2011
Access Date: September 12, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803914940
Online ISBN: 9781412983440
DOI: https://dx.doi.org/10.4135/9781412983440
Print pages: 10-26

Bivariate Regression: Fitting A Straight Line

Social researchers often inquire about the relationship between two variables. Numerous examples come to mind. Do men participate more in politics than do women? Is the working class more liberal than the middle class? Are Democratic members of Congress bigger spenders of the taxpayer's dollar than Republicans? Are changes in the unemployment rate associated with changes in the President's popularity at the polls? These are specific instances of the common query, “What is the relationship between variable X and variable Y?” One answer comes from bivariate regression, a straightforward technique which involves fitting a line to a scatter of points.

Exact Versus Inexact Relationships

Two variables, X and Y, may be related to each other exactly or inexactly. In the physical sciences, variables frequently have an exact relationship to each other. The simplest such relationship between an *independent variable* (the “cause”), labelled X, and a *dependent variable* (the “effect”), labelled Y, is a straight line, expressed in the formula,

$$Y = a + bX,$$

where the values of the coefficients, a and b, determine, respectively, the precise height and steepness of the line. Thus, the coefficient a is referred to as the *intercept* or *constant*, and the coefficient b is referred to as the *slope.* The hypothetical data in Table 1, for example, indicate that Y is linearly related to X by the following equation,

$$Y = 5 + 2X.$$

This straight line is fitted to these data in Figure 1a. We note that for each observation on X, one and only one Y value is possible. When, for instance, X equals one, Y must equal seven. If X increases one unit in value, then Y necessarily increases by precisely two units. Hence, knowing the X score, the Y score can be perfectly predicted. A real world example with which we are all familiar is

$$Y = 32 + 9/5\ X,$$

where temperature in Fahrenheit (Y) is an exact linear function of temperature in Celsius (X).

In contrast, relationships between variables in the social sciences are almost always inexact. The equation for a linear relationship between two social science variables would be written, more realistically, as

$$Y = a + bX + e,$$

where e represents the presence of *error.* A typical linear relationship for social science data is pictured in Figure 1b. The equation for these data happens to be the same as that for the data of Table 1, except for the addition of the error term, The error term acknowledges that the prediction equation by itself,

$$\hat{Y} = 5 + 2X,$$

does not perfectly predict Y. (The $\hat{Y}$ distinguishes the predicted Y from the observed Y.) Every Y value does

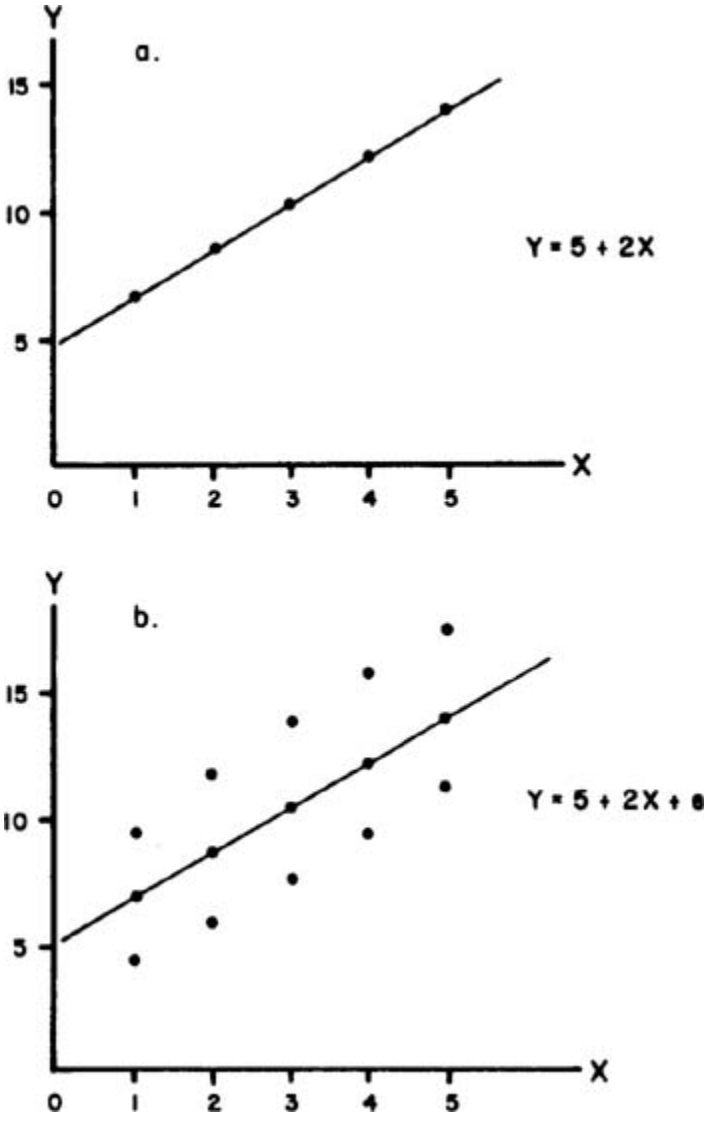
not fall exactly on the line. Thus, with a given X, there may occur more than one Y. For example, with X = 1, we see there is a Y = 7, as predicted, but also there is a Y = 9. In other words, knowing X, we do not always know Y.

$$Y = 5 + 2X + e.$$

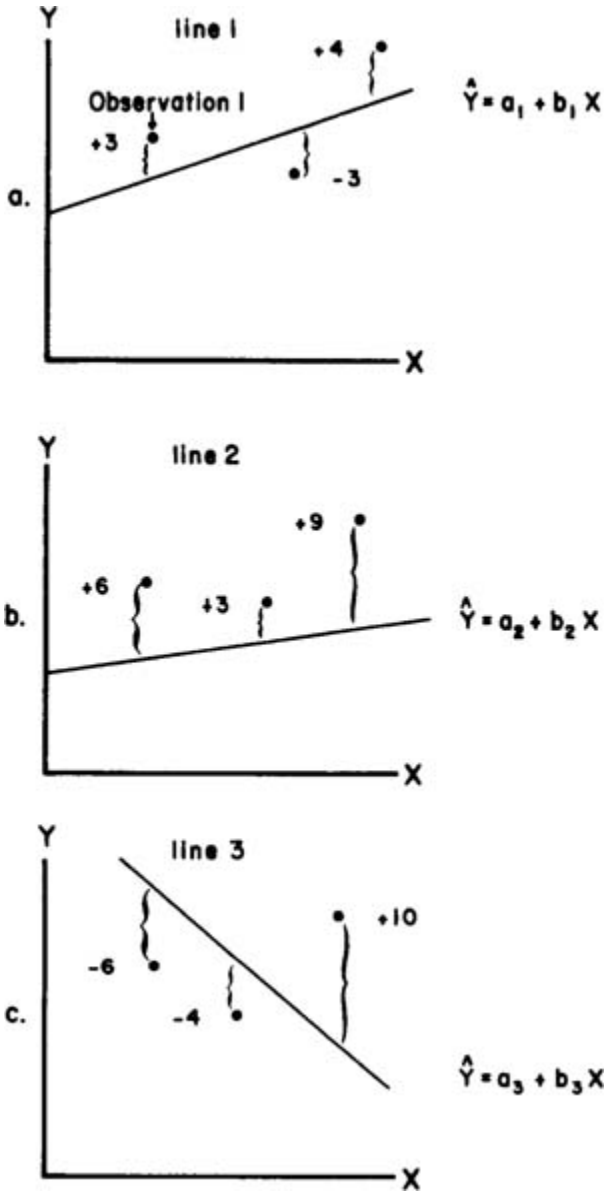
TABLE 1 Perfect Linear Relationship between X and Y

	Y = 5 + 2X
X	Y
0	5
1	7
2	9
3	11
4	13
5	15

Figurai 1a-b: Exact and Inexact Linear Relationships between X end Y



Figures 2a-c: Some Free-hand Straight Line Fits to a Scatter of Points



This inexactness is not surprising. If, for instance, X = number of elections voted in (since the last presidential election), and Y = campaign contributions (in dollars), we would not expect everyone who voted in, say, three elections to contribute exactly the same amount to campaigns. Still, we would anticipate that someone voting three times would likely contribute more than someone voting one time, and less than someone voting five times. Put another way, a person's campaign contribution is likely to be a linear function of electoral participation, plus some error, which is the situation described in Figure 1b.

The Least Squares Principle

In postulating relationships among social science variables, we commonly assume linearity. Of course, this assumption is not always correct. Its adoption, at least as a starting point, might be justified on

several grounds. First, numerous relationships have been found empirically to be linear. Second, the linear specification is generally the most parsimonious. Third, our theory is often so weak that we are not at all sure what the nonlinear specification would be. Fourth, inspection of the data themselves may fail to suggest a clear alternative to the straight line model. (All too frequently, the scatterplot may look like nothing so much as a large chocolate chip cookie.) Below, we focus on establishing a linear relationship between variables. Nevertheless, we should always be alert to the possibility that a relationship is actually nonlinear.

Given that we want to relate Y to X with a straight line, the question arises as to which, out of all possible straight lines, we should choose. For the scatterplot of Figure 2a we have sketched in free-hand the line 1, defined by this prediction equation:

$$\hat{Y} = a_1 + b_1 X.$$

One observes that the line does not predict perfectly, for example, the vertical distance from Observation 1 to the line is three units. The *prediction error* for this Observation 1, or any other observation, i, is calculated as follows:

$$\text{prediction error} = \text{observed} - \text{predicted} = Y_i - \hat{Y}_i.$$

Summing the prediction error for all the observations would yield a total prediction error (TPE), total prediction error = $\Sigma(Y_1 - \hat{Y}_1) = (+3–3+4) = 4$.

Clearly, line I fits the data better than free-hand line 2 (see Figure 2b), represented by the equation,

$$\hat{Y} = a_2 + b_2 X.$$

(TPE for line 2 = 18). However, there are a vast number of straight lines besides line 2 to which line 1 could be compared. Does line 1 reduce prediction error to the minimum, or is there some other line which could do better? Obviously, we cannot possibly evaluate all the free-hand straight lines that could be sketched on the scatterplot. Instead, we rely on the calculus, in order to discover the values of a and b which generate the line with the lowest prediction error.

Before presenting this solution, however, it is necessary to modify somewhat our notion of prediction error. Note that line 3 (see Figure 2c), indicated by the equation,

$$\hat{Y} = a_3 + b_3 X,$$

provides a fit that is patently less good than line 1. Nevertheless, the TPE = 0 for line 3. This example reveals that TPE is an inadequate measure of error, because the positive errors tend to cancel out the negative errors (here, −6–4 + 10 = 0). One way to overcome this problem of opposite signs is to square each error. (We reject the use of the absolute value of the errors because it fails to account adequately for large errors and is computationally unwieldy.) Our goal, then, becomes one of selecting the straight line which *minimizes the sum of the squares of the errors* (SSE):

$$\text{SSE} = \Sigma(Y_i - \hat{Y}_i)^2.$$

Through the use of the calculus, it can be shown that this sum of squares is at a minimum, or “least,” when the coefficients a and b are calculated as follows:

$$b = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

$$a = \bar{Y} \quad b\bar{X}.$$

These values of a and b are our “least squares” estimates.

At this point it is appropriate to apply the least squares principle in a research example. Suppose we are studying income differences among local government employees in Riverside, a hypothetical medium-size midwestern city. Exploratory interviews suggest a relationship between income and education. Specifically, those employees with more formal training appear to receive better pay. In an attempt to verify whether this is so, we gather relevant data.

The Data

We do not have the time or money to interview all 306 employees on the city payroll. Therefore, we decide to interview a *simple random sample* of 32, selected from the personnel list which the city clerk kindly provided.[1] (The symbol for a sample is “n,” so we can write n = 32.) The data obtained on the current annual income (labelled variable Y) and the number of years of formal education (labelled variable X) of each respondent is given in Table 2.

The Scatterplot

From simply reading the figures in Table 2, it is difficult to tell whether there is a relationship between education (X) and income (Y). However, the picture becomes clearer when the data are arranged in a *scatterplot.* In Figure 3, education scores are plotted along the X-axis, and income scores along the Y-axis. Every respondent is represented by a point, located where a perpendicular line from his or her X value intersects a perpendicular line from his or her Y value. For example, the dotted lines in Figure 3 fix the position of Respondent 3, who has an income of $6898 and six years of education.

Visual inspection of this scatterplot suggests the relationship is essentially linear, with more years of education leading to higher income. In equation form, the relationship appears as,

$$Y = a + bX + e,$$

where Y = respondent's annual income (in dollars,), X = respondent's formal education (in years), a = intercept, b = slope, e = error.

TABLE 2 Data on Education and Income

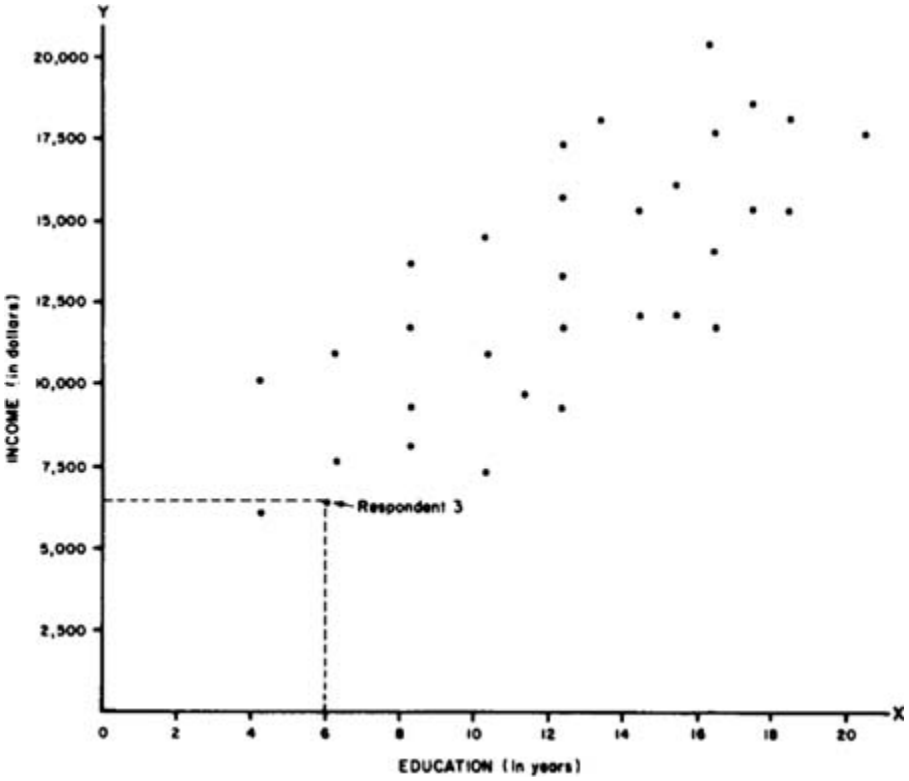
Respondent	Education (in years) X	Income (in dollars) Y
1	4	$ 6281
2	4	10516
3	6	6898
4	6	8212
5	6	11744
6	8	8618
7	8	10011
8	8	12405
9	8	14664
10	10	7472
11	10	11598
12	10	15336
13	11	10186
14	12	9771
15	12	12444
16	12	14213
17	12	16908
18	12	18347
19	13	19546
20	14	12660
21	14	16326
22	15	12772
23	15	17218
24	16	12599
25	16	14852
26	16	19138
27	16	21779
28	17	16428
29	17	20018
30	18	16526
31	18	19414
32	20	18822

Estimating this equation with least squares yields,

$$\hat{Y} = 5078 + 732\,X,$$

which indicates the straight line that best fits this scatter of points (see Figure 4). This prediction equation is commonly referred to as a *bivariate regression equation.* (Further, we say Y has been “regressed on” X.)

Figura 3: Scatterplot of Education and Income

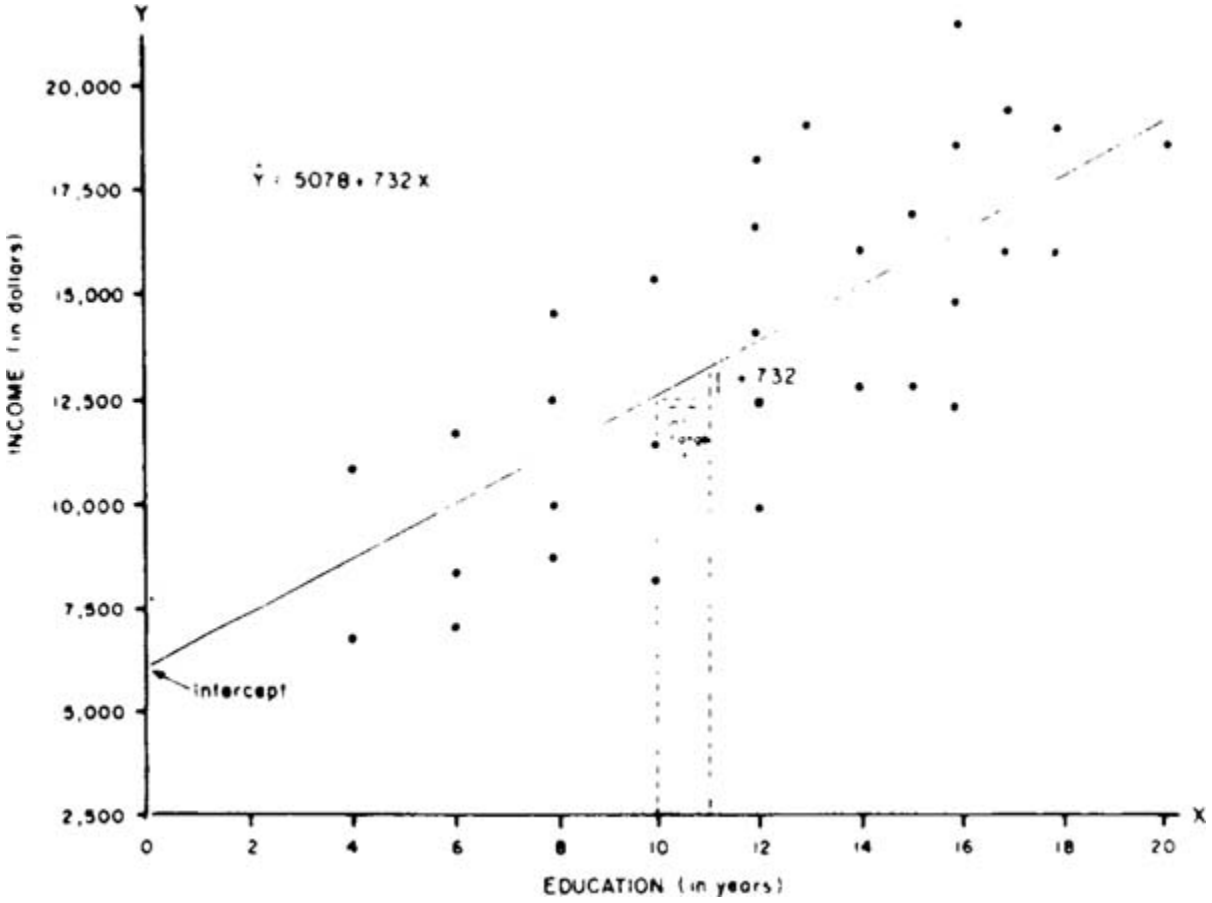


The Slope

Interpretation of the estimates is uncomplicated. Let us first consider the estimate of the slope, b. *The slope estimate indicates the average change in Y associated with a unit change in X.* In our Riverside example, the slope estimate, 732, says that a one-year increase in an employee's amount of formal education is associated with an average annual income increase of $732. Put another way, we expect an employee with, say, 11 years of education to have an income that is $732 more than an employee having only 10 years of education. We can see how the slope dictates the change in Y for a unit change in X by studying Figure 4, which locates the expected values of Y, given X = 10, and X = 11, respectively.

Note that the slope tells us only the *average* change in Y that accompanies a unit change in X. The relationship between social science variables is inexact, that is, there is always error. For instance, we would not suppose that an additional year of education for any particular Riverside employee would be associated with an income rise of exactly $732. However, when we look at a large number of employees who have managed to acquire this extra year of schooling, the average of their individual income gains would be about $732.

Figure 4: The Regression Line for the Income and Education Data



The slope estimate suggests the average change in Y *caused by* a unit change in X. Of course, this causal language may be inappropriate. The regression of Y on X might support your notion of the causal process, but it cannot establish it. To appreciate this critical point, realize that it would be a simple matter to apply least squares to the following regression equation,

$$X = a + bY + e,$$

where X = the *dependent* variable, Y = the *independent* variable. Obviously. such a computational exercise would not suddenly reverse the causal order of X and Y in the real world. The correct causal ordering of the variables is determined outside the estimation procedure. In practice, it is based on theoretical considerations, good judgment, and past research. With regard to our Riverside example, the actual causal relationship of these variables does seem to be reflected in our original model; that is, shifts in education appear likely to cause shifts in income; but, the view that changes in income cause changes in formal years of education is implausible, at least in this instance. Thus, it is only somewhat adventuresome to conclude that a one-year increase in formal education *causes* income to increase $732, on the average.

The Intercept

The intercept, a, is so called because it indicates the point where the regression line “intercepts” the Y-axis. It

estimates the average value of Y when X equals zero. Thus, in our Riverside example, the intercept estimate suggests that the expected income for someone with no formal education would be $5078. This particular estimate highlights worthwhile cautions to observe when interpreting the intercept. First, one should be leery of making a prediction for Y based on an X value beyond the range of the data. In this example, the lowest level of educational attainment is four years; therefore, it is risky to extrapolate to the income of someone with zero years of education. Quite literally, we would be generalizing beyond the realm of our experience, and so may be way off the mark. If we are actually interested in those with no education, then we would do better to gather data on them.

A second problem may arise if the intercept has a negative value. Then, when X = 0, the predicted Y would necessarily equal the negative value. Often, however, in the real world it is impossible to have a score on Y that is below zero, for example, a Riverside employee could not receive a minus income. In such cases, the intercept is “nonsense,” if taken literally. Its utility would be restricted to ensuring that a prediction “comes out right.” It is a constant that must always be added on to the slope component, “bX,” for Y to be properly estimated. Drawing on an analogy from the economics of the firm, the intercept represents a “fixed cost” that must be included along with the “varying costs” determined by other factors, in order to calculate “total cost.”

Prediction

Knowing the intercept and the slope, we can predict Y for a given X value. For instance, if we encounter a Riverside city employee with 10 years of schooling, then we would predict his or her income would be $12,398, as the following calculations show:

$$\hat{Y} = 5078 + 732\,X$$
$$= 5078 + 732(10)$$
$$= 5078 + 7320$$
$$\hat{Y} = 12{,}398.$$

In our research, we might be primarily interested in prediction, rather than explanation. That is, we may not be directly concerned with identifying the variables that cause the dependent variable under study; instead, we may want to locate the variables that will allow us to make accurate guesses about the value of the dependent variable. For instance, in studying elections, we may simply want to predict winning candidates, not caring much about why they win. Of course, predictive models are not completely distinct from explanatory models. Commonly, a good explanatory model will predict fairly well. Similarly, an accurate predictive model is usually based on causal variables, or their surrogates. In developing a regression model, the research question dictates whether one emphasizes prediction or explanation. It is safe to conclude that, generally, social scientists stress explanation rather than prediction.

Assessing Explanatory Power: The R^2

We want to know how powerful an explanation (or prediction) our regression model provides. More technically, how well does the regression equation account for variations in the dependent variable? A preliminary judgment comes from visual inspection of the scatterplot. The closer the regression line to the points, the better the equation "fits" the data. While such "eyeballing" is an essential first step in determining the "goodness of fit" of a model, we obviously need a more formal measure, which the *coefficient of determination* (R^2) gives us.

We begin our discussion by considering the problem of predicting Y. If we *only* have observations on Y, then the best prediction for Y is always the estimated mean of Y. Obviously, guessing this average score for each case will result in many poor predictions. However, knowing the values of X, our predictive power can be improved, provided that X is related to Y. The question, then, is how much does this knowledge of X improve our prediction of Y?

In Figure 5 is a scatterplot, with a regression line fitted to the points. Consider prediction of a specific case, Y_1. Ignoring the X score, the best guess for the Y score would be the mean, $\hat{Y}$. There is a good deal of error in this guess, as indicated by the deviation of the actual score from the mean, $Y_1 - \hat{Y}$. However, by utilizing our knowledge of the relationship of X to Y, we can better this prediction. For the particular value, X_1, the regression line predicts the dependent variable is equal to $\hat{Y}_1$, which is a clear improvement over the previous guess. Thus, the regression line has managed to account for some of the deviation of this observation from the mean; specifically, it "explains" the portion, $\hat{Y}_1 - \hat{Y}$. Nevertheless, our regression prediction is not perfect, but rather is off by the quantity, $Y_1 - \hat{Y}_1$; this deviation is left "unexplained" by the regression equation. In brief, the deviation of Y_1 from the mean can be grouped into the following components:

$(Y_1 - \bar{Y})$ = total deviation of Y_1 from the mean, $\bar{Y}$

$(\hat{Y}_1 - \bar{Y})$ = explained deviation of Y_1 from $\bar{Y}$

$(Y_1 - \hat{Y}_1)$ = unexplained deviation of Y_1 from $\bar{Y}$.

We can calculate these deviations for each observation in our study. If we first square the deviations, then sum them, we obtain the complete components of variation for the dependent variable:

$\Sigma(Y_i - \bar{Y})^2$ = total sum of squared deviations (TSS)

$\Sigma(\hat{Y}_i - \bar{Y})^2$ = regression (explained) sum of squared deviations (RSS)

$\Sigma(Y_i - \hat{Y}_i)^2$ = error (unexplained) sum of squared deviations (ESS).

From this, we derive,

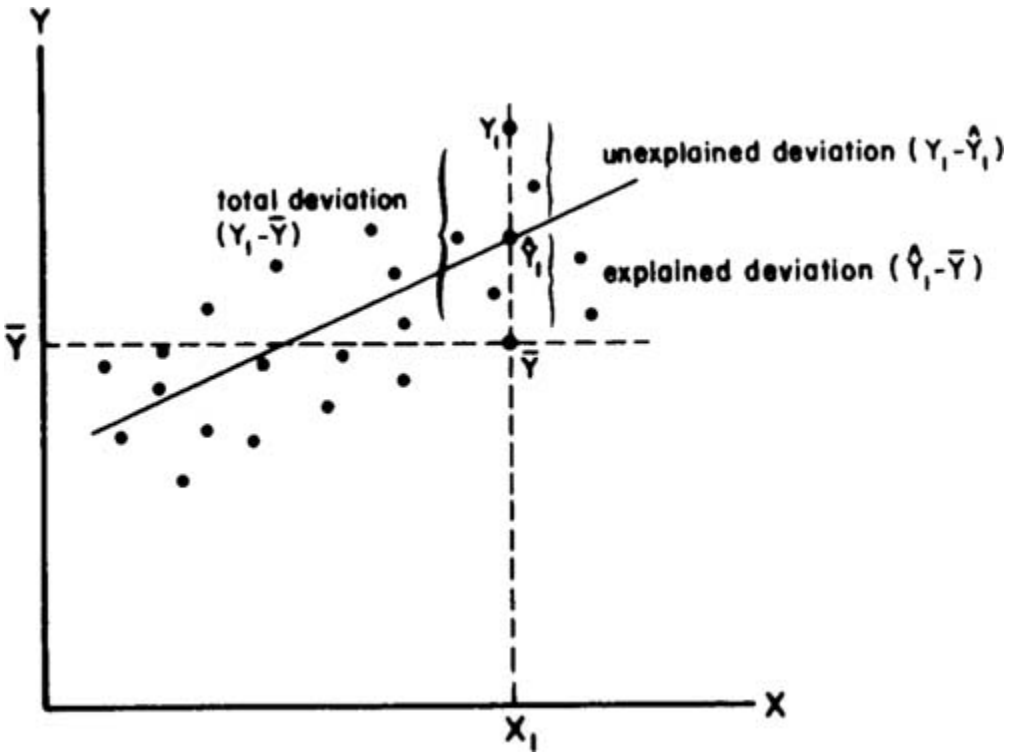
$$\mathbf{TSS = RSS + ESS.}$$

The TSS indicates the total variation in the dependent variable that we would like to explain. This total variation can be divided into two parts: the part accounted for by the regression equation (RSS) and the part the regression equation cannot account for, ESS. (We recall that the least squares procedure guarantees that this error component is at minimum.) Clearly, the larger RSS relative to TSS, the better. This notion forms the

basis of the R^2 measure:

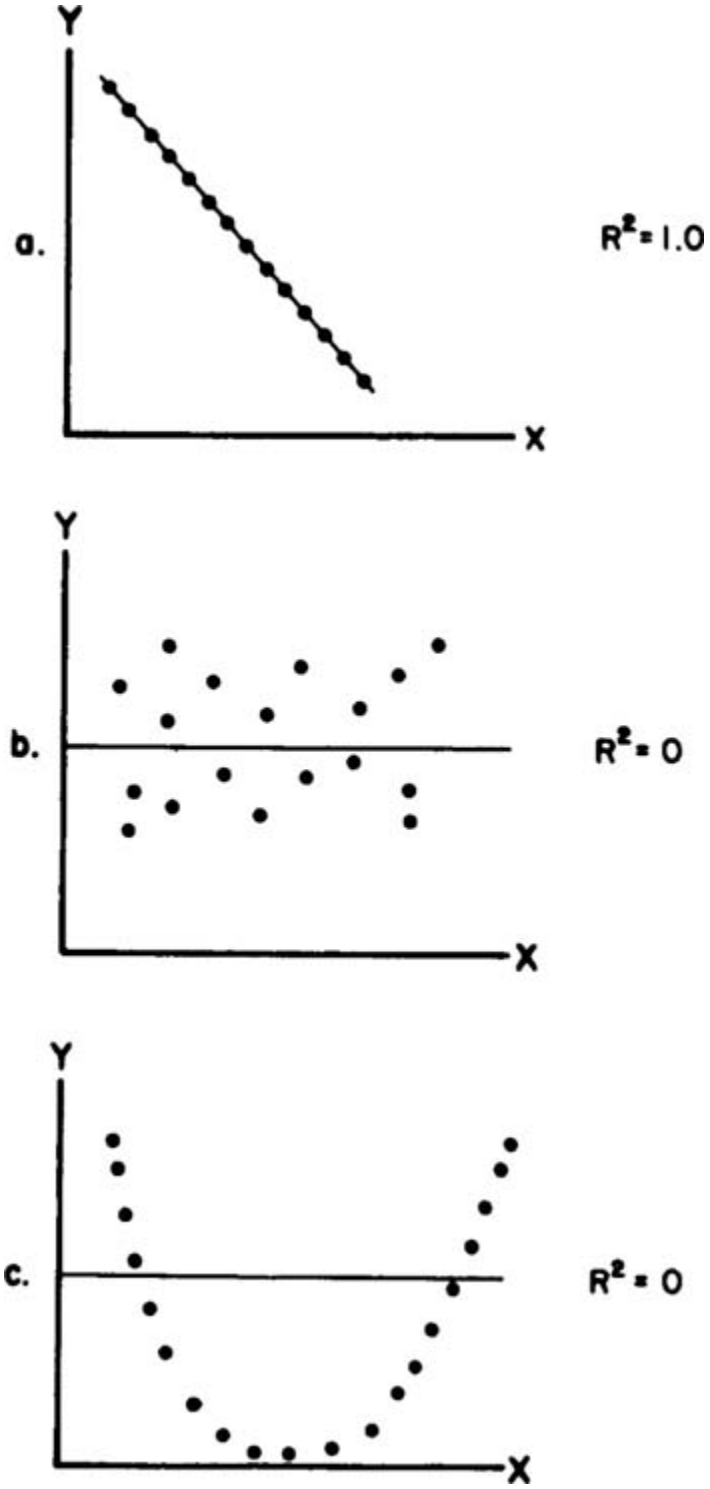
$$R^2 = RSS / TSS.$$

Figure 5: Components of Variation in Y



The coefficient of determination, R^2, indicates the explanatory power of the bivariate regression model. It records the proportion of variation in the dependent variable "explained" or "accounted for" by the independent variable. The possible values of the measure range from "+1" to "0." At the one extreme, when $R^2 = 1$, the independent variable completely accounts for variation in the dependent variable. All observations fall on the regression line, so knowing X enables the prediction of Y without error. Figure 6a provides an example where $R^2 = 1$. At the other extreme, when $R^2 = 0$, the independent variable accounts for no variation in the dependent variable. The knowledge of X is no help in predicting Y, for the two variables are totally independent of each other. Figure 6b gives an example where $R^2 = 0$ (note that the slope of the line also equals zero). Generally, R^2 falls between these two extremes. Then, the closer R^2 is to 1, the better the fit of the regression line to the points, and the more variation in Y is explained by X. In our Riverside example, $R^2 = .56$. Thus, we could say that education, the independent variable, accounts for an estimated 56% of the variation in income, the dependent variable.

Figures 6a-c: Examples of the Extreme Values of the R^2



In regression analysis, we are virtually always pleased when the R^2 is high, because it indicates we are accounting for a large portion of the variation in the phenomenon under study. Further, a very high R^2 (say about .9) is essential if our predictions are to be accurate. (In practice, it is difficult to attain an R^2 of this magnitude. Thus, quantitative social scientists, at least outside economics, seldom make predictions.)

However, a sizable R^2 does not necessarily mean we have a *causal* explanation for the dependent variable; instead, we may merely have provided a *statistical* explanation. In the Riverside case, suppose we regressed respondent's current income, Y, on income of the previous year, Y_{t-1}. Our revised equation would be as follows:

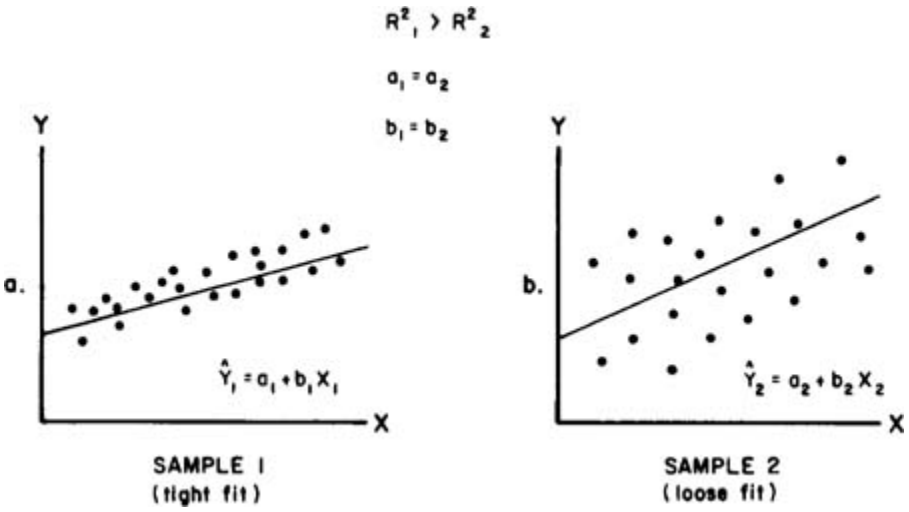
$$Y = a + bY_{t-1} + e.$$

The R^2 for this new equation would be quite large (above .9), but it would not really tell us what causes income to vary; rather, it offers merely a statistical explanation. The original equation, where education was the independent variable, provides a more convincing causal explanation of income variation, despite the lower R^2 of .56.

Even if estimation yields an R^2 that is rather small (say below .2), disappointment need not be inevitable, for it can be informative. It may suggest that the linear assumption of the R^2 is incorrect. If we turn to the scatterplot, we might discover that X and Y actually have a close relationship, but it is nonlinear. For instance, the curve (a parabola) formed by connecting the points in Figure 6c illustrates a perfect relationship between X and Y (i.e., $Y = X^2$), but $R^2 = 0$. Suppose, however, that we rule out nonlinearity. Then, a low R^2 can still reveal that X does help explain Y, but contributes a rather small amount to that explanation. Finally, of course, an extremely low R^2 (near 0), offers very useful information, for it implies that Y has virtually no linear dependency on X.

A final point on the interpretation of R^2 deserves mention. Suppose we estimate the *same* bivariate regression model for two samples from different populations, labelled 1 and 2. (For example, we wish to compare the income-education model from Riverside to the income-education model from Flatburg.) The R^2 for sample 1 could differ from the R^2 for sample 2, even though the parameter estimates for each were exactly the same. It simply implies that the structural relationship between the variables is the same ($a_1 = a_2$; $b_1 = b_2$), but it is less predictable in population 2. In other words, the same equation provides the best possible fit for both samples but, in the second instance, is less satisfactory as a total explanation of the dependent variable. Visually, this is clear. We can see, in comparing Figures 7a and 7b, that the points are clustered more tightly around the regression line of Figure 7a, indicating the model fits those data better. Thus, the independent variable, X, appears a more important determinant of Y in sample 1 than in sample 2.

Figures 7a-b: Tight Fit vs. Loose Fit of a Regression Line



R^2 Versus r

The relationship between the coefficient of determination. R^2, and the estimate of the correlation coefficient, r, is straightforward:

$R^2 = r^2$.

This equality suggests a possible problem with r, which is a commonly used measure of strength of association.[2] That is, r can inflate the importance of the relationship between X and Y. For instance, a correlation of .5 implies to the unwary reader that one-half of Y is being explained by X, since a perfect correlation is 1.0. Actually, though, we know that the r = .5 means that X explains only 25% of the variation in Y (because r^2 = .25), which leaves fully three-fourths of the variation in Y unaccounted for. (The r will equal the R^2 only at the extremes, when r = ± 1 or 0.) By relying on r rather than R^2, the impact of X on Y can be made to seem much greater than it is. Hence, to assess the strength of the relationship between the independent variable and the dependent variable, the R^2 is the preferred measure.

http://dx.doi.org/10.4135/9781412983440.n1