Bivariate Regression: Assumptions and Inferences

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: 27-47

Bivariate Regression: Assumptions and Inferences

Recall that the foregoing regression results from the Riverside study are based on a *sample* of the city employees (n = 32). Since we wish to make accurate inferences about the actual *population* values of the intercept and slope parameters, this bivariate regression model should meet certain assumptions. For the population, the bivariate regression model is,

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

where the Greek letters indicate it is the population equation, and we have included the subscript, i, which refers to the i^{th} observation. With the sample, we calculate

$$Y_i = a + bX_i + e_i.$$

In order to infer accurately the true population values, α and β, from these sample values, a and b, we make the following assumptions.

The Regression Assumptions

1. No specification error.
    a. The relationship between X_i and Y_i is linear.
    b. No relevant independent variables have been excluded.
    c. No irrelevant independent variables have been included.
2. No measurement error.
    a. The variables X_i and Y_i are accurately measured.
3. The following assumptions concern the error term, ε_i:
    a. Zero mean: $E(\varepsilon_i) = 0$.
        i. For each observation, the *expected value* of the error term is zero. (We use the symbol E( ) for expected value which, for a random variable, is simply equal to its mean.)
    b. Homoskedasticity: $E(\varepsilon^2_i) = 6^2$.
        i. The variance of the error term is constant for all values of X_i.
    c. No autocorrelation: $E(\varepsilon_i\varepsilon_j) = 0$ $(i \neq j)$.
        i. The error terms are uncorrected.
    d. The independent variable is uncorrelated with the error term: $E(\varepsilon_i X_i) = 0$.
    e. Normality.

i. The error term, ε_i, is normally distributed.

When assumptions 1 to 3d are met, desirable estimators of the population parameters, α and β, will be obtained; technically, they will be the “best linear unbiased estimates,” BLUE. (An unbiased estimator correctly estimates the population parameter, on the average, i.e., $E(b) = \beta$. For instance, if we repeatedly draw samples from the population, each time recalculating b, we would expect the average of all these b's to equal β.) If the normality assumption (3e) also holds, they will be the “best unbiased estimates,” and we can carry out significance tests, in order to determine how likely it is that the population parameter values differ from zero. Below, we consider each assumption in more detail.

The first assumption, absence of specification error, is critical. In sum, it asserts that the theoretical model embodied in the equation is correct. That is, the functional form of the relationship is actually a straight line, and no variables have been improperly excluded or included as “causes.” Let us examine the Riverside example for specification error. Visual inspection of the shape of the scatterplot (see Figure 4), along with the $R^2 = .56$, indicates that the relationship is essentially linear. However, it seems likely that relevant variables have been excluded, for factors besides education undoubtedly influence income. These other variables should be identified and brought into the equation, both to provide a more complete explanation and to assess the impact of education after additional forces are taken into account. (We take up this task in the next chapter.) The final aspect of specification error, inclusion of an irrelevant variable, argues that education might not really be associated with income. To evaluate this possibility, we will perform a test for statistical significance.

The need for the second assumption, no measurement error, is self-evident. If our measures are inaccurate, then our estimates are likely to be inaccurate. For instance, with the Riverside case, suppose that in the measurement of the education variable, the respondents tended to report the number of years of schooling they would *like* to have had, rather than the number of years of schooling they *actually* had. If we were to use such a variable to indicate actual years of schooling, it would contain error, and the resulting regression coefficient would not accurately reflect the impact of actual education on income. When the analyst cannot safely rule out the possibility of measurement error, then the magnitude of the estimation problem depends on the nature and location of the error. If only the dependent variable is measured with error, then the least squares estimates may remain unbiased, provided the error is “random.” However, if the independent variable is measured with any error, then the least squares estimates will be biased. In this circumstance, all solutions are problematic. The most oft-cited approach is *instrumental variables estimation*, but it cannot promise the restoration of unbiased parameter estimates.

The third set of assumptions involve the error term. The initial one, a zero mean, is of little concern because, regardless, the least squares estimate of the slope is unchanged. It is true that, if this assumption is not met, the intercept estimate will be biased. Nevertheless, since the intercept estimate is often of secondary interest in social science research, this potential source of bias is rather unimportant.

Violating the assumption of homoskedasticity is more serious. While the least squares estimates continue to be unbiased, the significance tests and confidence intervals would be wrong. Let us examine Figure 4 from the Riverside study. Homoskedasticity would appear to be present, because the variance in prediction errors is more or less constant across the values of X; that is, the points snuggle in a band of equal width above and below the regression line. If, instead, the points fanned out from the regression line as the value of X increased, the assumption would not hold, and a condition of *heteroskedasticity* would prevail. The recommended solution for this condition is a *weighted least squares* procedure. (Diagnosis of heteroskedasticity is discussed further when the analysis of residuals is considered.)

The assumption of no autocorrelation means that the error corresponding to an observation is not correlated with any of the errors for the other observations. When autocorrelation is present, the least squares parameter estimates are still unbiased; however, the significance tests and confidence intervals are invalid. Commonly, significance tests will be much more likely to indicate that a coefficient is statistically significant, when in fact it is not. Autocorrelation more frequently appears with *time-series* variables (repeated observations on the same unit through time) than with *cross-sectional* variables (unique observations on different units at the same point in time, as with our Riverside study). With time-series data, the no autocorrelation assumption requires that error for an observation at an earlier time is not related to errors for observations at a later time. If we conceive of the error term in the equation as, in part, a summary of those explanatory variables that have been left out of the regression model, then no autocorrelation implies that those forces influencing Y in, say, year 1, are independent of those forces influencing Y in year 2.[3] This assumption, it should be obvious, is often untenable. (The special problems of time-series analysis have generated an extensive literature; for a good introduction, see Ostrom, 1978.)

The next assumption, that the independent variable is uncorrelated with the error term, can be difficult to meet in nonexperimental research. Typically, we cannot freely set the values of X like an experimenter would, but rather must merely observe values of X as they present themselves in society. If this observed X variable is related to the error term, then the least squares parameter estimates will be biased. The simplest way to test for this violation is to evaluate the error term as a collection of excluded explanatory variables, each of which might be correlated with X. Thus, in the Riverside case, the error term would include the determinants of income other than education, such as sex of the respondent. If the explanatory variable of education is correlated with the explanatory variable of sex, but this latter variable is excluded from the equation, then the slope estimate for the education variable in the bivariate regression will be biased. This b will be too large, because the education variable is allowed to account for some income variation that is actually the product of sex differences. The obvious remedy, which we come to employ, is the incorporation of the missing explanatory variables into the model. (If, for some reason, an explanatory variable cannot be so incorporated, then we must trust the assumption that, as part of the error term, it is uncorrelated with the independent variable actually in the model.)

The last assumption is that the error term is normally distributed. Since the distributions of Y_i and ε_i are the same (only their means are different), our discussion will be facilitated by simply considering the distribution of Y_i. The frequency distribution of a variable that conforms to a normal curve has a symmetric bell-shape, with

95% of the observations falling within two standard deviations, plus or minus, of the mean. With regard to the Riverside example, the unique observations on the income variable (Y_i) could be graphed onto a frequency polygon to allow a visual inspection for normality. Or, for a quick preliminary check, we could count the number of observations above and below the mean, expecting about half in either direction. (In fact, there are 16 incomes above and 16 incomes below the mean of $13,866, which suggests a normal distribution.) A more formal measure, which takes into account all the information in the frequency distribution, is the *skewness* statistic, based on the following formula:

$$\text{skewness} = \frac{\Sigma\left(\frac{y_i - \bar{y}}{s_y}\right)^3}{n}.$$

If the distribution is normal, then skewness = 0. For our income variable, skewness measures only −.02, indicating that the distribution is virtually normal.

There is some disagreement in the statistical literature over how serious the violations of the regressions assumptions actually are. At one extreme, researchers argue that regression analysis is “robust,” that is, the parameter estimates are not meaningfully influenced by violations of the assumptions. This “robust” perspective on regression is employed in Kerlinger and Pedhazar (1973). At the other extreme, some feel that violations of the assumptions can render the regression results almost useless. Bibby's (1977) work provides an example of this “fragile” view of regression analysis. Clearly, some of the assumptions are more robust than others. The normality assumption, for instance, can be ignored when the sample size is large enough, for then the central-limit theorem can be invoked. (The central-limit theorem indicates that the distribution of a sum of independent variables, which we can conceive of the error term as representing, approaches normality as sample size increases, irrespective of the nature of the distributions in the population.) By way of contrast, the presence of specification error, such as the exclusion of a relevant variable, creates rather serious estimation problems which can be relieved only by introduction of the omitted variable into the model. Those who wish to gain a fuller understanding of this controversy over assumptions should consult, in addition to the efforts just cited, the excellent paper by Bohrnstedt and Carter (1971). More advanced treatments of the regression assumptions are available in econometrics texts; listing them in order of increasing difficulty, I would recommend Kelejian and Oates (1974), Pindyck and Rubinfeld (1976), and Kmenta (1971).

Confidence Intervals and Significance Tests

Because social science data invariably consist of samples, we worry whether our regression coefficients actually have values of zero in the population. Specifically, is the slope (or the intercept) estimate significantly different from zero? (Of course, we could test whether the parameter estimate was significantly different from some number other than zero; however, we generally do not know enough to propose such a specific value.) Formally, we face two basic hypotheses: the null and an alternative. The *null hypothesis* states that X is not associated with Y; therefore, the slope, β, is zero in the population. An *alternative hypothesis* states that X is associated with Y; therefore, the slope is *not* zero in the population. In summary, we have

$$H_0: \beta = 0 \text{ (null hypothesis)}$$

$$H_1: \beta \neq 0 \text{ (alternative hypothesis).}$$

To test these hypotheses, an interval can be constructed around the slope estimate, b. The most widely used is a two-tailed, 95% *confidence interval:*

$$(b \pm t_{n-2;.975} s_b).$$

If the value of zero does *not* fall within this interval, we reject the null hypothesis and accept the alternative hypothesis, with 95% confidence. Put another way, we could conclude that the slope estimate, b, is significantly different from zero, at the .05 level. (The level of *statistical significance* associated with a particular confidence interval can be determined simply by subtracting the confidence level from unity, for example, 1 − .95 = .05.)

In order to apply this confidence interval, we must understand the terms of the formula. These are easy enough. The term s_b is an estimate of the standard deviation of the slope estimate, b, and is commonly referred to as the *standard error.* It is a useful measure of the dispersion of our slope estimate. The formula for this standard error is,

$$s_b = \sqrt{\frac{\Sigma(Y - \hat{Y})^2/(n-2)}{\Sigma(X - \bar{X})^2}}.$$

Statistical computing packages such as SPSS routinely print out the standard errors when estimating a regression equation.

Because s_b is an estimate (we seldom actually know the standard deviation of the slope estimate), it is technically incorrect to use the normal curve to construct a confidence interval for β. However, we can utilize the t distribution with (n-2) degrees of freedom. (The t distribution is quite similar to the normal distribution, especially as n becomes large, say greater than 30.) Almost every statistics text provides a table for the t distribution.

The last component in the confidence interval formula is the subscript, ".975." This merely indicates that we are employing a 95% confidence interval, but with two-tails. A two-tailed test means that the hypothesis about the affect of X on Y is nondirectional; for example, the above alternative hypothesis, H_1, is sustained if b is either significantly negative or significantly positive.

Suppose we now construct a two-tailed, 95% confidence interval around the regression coefficients in our Riverside study. We have,

$$\begin{aligned} \hat{Y} = \; & 5078 \; + \; 732X \\ & (1498) \quad (118) \end{aligned}$$

where the figures in parentheses are the standard errors of the parameter estimates. Given the sample size is 32,

$$t_{n-2;.975} = t_{32\text{-}2;.975} = t_{30;.975} = 2.04,$$

according to the t table. Therefore, the two-tailed, 95% confidence interval for β is

$$(b \pm t_{n-2;.975} s_b) = 732 \pm 2.04\,(118) = (732 \pm 241).$$

The probability is .95 that the value of the population slope, β, is between $491 and $973. Since the value of zero does not fall within the interval, we reject the null hypothesis. We conclude that the slope estimate, b, is significantly different from zero, at the .05 level.

In the same fashion, we can construct a confidence interval for the intercept, α. Continuing the Riverside example,

$$(a \pm t_{n-2;.975} s_a) = 5078 \pm 2.04\,(1498) = (5078 \pm 3056).$$

Clearly, the two-tailed, 95% confidence band for the intercept does not contain zero. We reject the null hypothesis and declare that the intercept estimate, a, is statistically significant at the .05 level. Graphically, this means we reject the possibility that the regression line cuts the origin.

Besides providing significance tests, confidence intervals also allow us to present our parameter estimates as a range. In a bivariate regression equation, b is a *point estimate;* that is, it is a specific value. The confidence band, in contrast, gives us an *interval estimate*, indicating that the slope in the population, β, lies within a range of values. We may well choose to stress the interval estimate over the point estimate. For example, in our Riverside study the point estimate of β is $732. This is our best guess, but in reporting the results we might prefer to say merely that a year increase in education is associated with an increase of “more or less $732” a year in income. Estimating a confidence interval permits us to formalize this caution; we could assert, with 95% certainty, that a one-year increase in education is associated with an income increase of from $491 to $973.

In the Riverside investigation, we have rejected, with 95% confidence, the null hypothesis of no relationship between income and education. Still, we know that there is a 5% chance we are wrong. If, in fact, the null hypothesis is correct, but we reject it, we commit a *Type I error.* In an effort to avoid Type I error, we could employ a 99% confidence interval, which broadens the acceptance region for the null hypothesis. The formula for a two-tailed, 99% confidence interval for β is as follows:

$$(b \pm t_{n-2;.995} s_b).$$

Applying the formula to the Riverside example,

$$732 \pm 2.75\,(118) = (732 \pm 324).$$

These results provide some evidence that we have not committed a Type I error. This broader confidence interval does not contain the value of zero. We continue to reject the null hypothesis, but with greater confidence. Further, we can say that the slope estimate, b, is statistically significant at the .01 level. (This effort to prevent Type I error involves a trade-off, for the risk of *Type II error*, accepting the null hypothesis when it is false, is inevitably increased. Type II error is discussed below.)

The One-Tailed Test

Thus far, we have concentrated on a two-tailed test of the form,

$$H_0\colon \beta = 0$$

$$H_1\colon \beta \neq 0.$$

Occasionally, though, our acquaintance with the phenomena under study suggests the sign of the slope. In such a circumstance, a one-tailed test might be more reasonable. Taking the Riverside case, we would not expect the sign of the slope to be negative, for that would mean additional education actually decreased income. Therefore, a more realistic set of hypotheses here might be,

$$H_0\colon \beta = 0$$

$$H_1\colon \beta > 0.$$

Applying a one-tailed, 95% confidence interval yields,

$$\beta > (b - t_{n-2;.95} s_b) = 732 - 1.70\,(118) = (732 - 201) = 531.$$

The lower boundary of the interval is above zero. Therefore, we reject the null hypothesis and conclude the slope is positive, with 95% confidence.

Once the level of confidence is fixed, it is “easier” to find statistical significance with a one-tailed test, as opposed to a two-tailed test. (The two-tailed confidence interval is more likely to capture zero. For instance, the lower bounds in the Riverside case for the two-tailed and one-tailed tests, respectively, are $491 and $531.) This makes intuitive sense, for it takes into account the researcher's prior knowledge, which may rule out one of the tails from consideration.

Significance Testing: A Rule of Thumb

Recall the formula for the two-tailed, 95% confidence interval for β:

$$(b \pm t_{n-2;.975} s_b).$$

If this confidence interval does not contain zero, we conclude that b is significant at the .05 level. We see that this confidence interval will not contain zero if, when b is positive,

$$(b - t_{n-2;.975} s_b) > 0,$$

or, when b is negative,

$$(b + t_{n-2;.975} s_b) < 0.$$

These requirements may be restated as,

$$b/s_b > t_{n-2;.975}, \text{ when b is positive,}$$

or,

$$b/s_b < t_{n-2;.975}, \text{ when b is negative.}$$

In brief, these requirements can be written,

$$|\, b/s_b \,| > t_{n-2;.975},$$

which says that when the absolute value of the parameter estimate, b, divided by its standard error, s_b, surpasses the t distribution value, $t_{n-2;.975}$, we reject the null hypothesis. Thus, a significance test at the .05 level, two-tailed, can be administered by examining this ratio. The test is simplified further when one observes that, for almost any sample size, the value in the t distribution approximates 2. For example, if the sample size is only 20, then $t_{20-2;.975} = t_{18;.975} = 2.10$. In contrast, if the sample is of infinite size, $t_{\infty;.975} = 1.96$. This narrow range of values given by the t distribution leads to the following rule of thumb. If,

$$| b/s_b | > 2,$$

then the parameter estimate, b, is significant at the .05 level, two-tailed.

This *t ratio*, as it is called, is routinely printed in the regression component of several computer data analysis programs. Otherwise, it is easily calculated by dividing b by s_b. The t ratio provides an efficient means of significance testing, and researchers frequently employ it. Of course, whenever more precision is wanted, the t table can always be consulted. Below is the bivariate regression model from our Riverside example, with the t ratios appearing in parentheses under the parameter estimates:

$$\begin{array}{lcc} \hat{Y} = & 5078 & + \ 732X \\ & (3.39) & (6.23). \end{array}$$

A quick glance at the t ratios reveals they exceed 2; we immediately conclude that both a and b are statistically significant at the .05 level.

Reasons Why a Parameter Estimate May Not Be Significant

There are many reasons why a parameter estimate may be found not significant. Let us assume, to narrow the field somewhat, that our data compose a probability sample and that the variables are correctly measured. Then, if b turns out not to be significant, the most obvious reason is that X is not a cause of Y. However, suppose we doubt this straightforward conclusion. The following is a partial list of reasons why we might fail to uncover statistical significance, even though X is related to Y in fact:

(1)
inadequate sample size
(2)
Type II error
(3)
specification error
(4)
restricted variance in X.

Below, these four possibilities are evaluated in order. (A fifth possibility is high multicollinearity, which we will consider in our discussion of multiple regression.)

As sample size increases, a given coefficient is more likely to be found significant. For insance, the b value in the bivariate regression of the Riverside example would not be significant (.05) if based on only five cases, but is significant with n = 32. This suggests it may be worthwhile for a researcher to gather more observations, for it will be easier to detect a relationship between X and Y in the population, if one is present. In fact, with a very large sample, statistical significance can be uncovered even if b is substantively quite small. (For very large samples, such as election surveys of 1000 or more, significance may actually be “too easy” to find, since tiny coefficients can be statistically significant. In this situation, the analyst might prefer to rely primarily on a substantive judgment of the importance of the coefficient.)

Let us suppose that sample size is fixed, and turn to the problem of choosing a significance level, as it relates to Type II error. In principle, we could set the significance test at any level between 0 and 1. In practice, however, most social scientists employ the .05 or .01 levels. To avoid the charges of arbitrariness or bias, we normally select one of these conventional standards before analysis begins. For instance, suppose prior to our investigation we decide to employ the .01 significance level. Upon analysis, we find b is not significant at this .01 level. But, we observe that it is significant at the less demanding level of .05. We might be loath to accept the null hypothesis as dictated by the .01 test, especially since theory and prior research indicate that X does influence Y. Technically, we worry that we are committing a Type II error, accepting the null when it is false. In the end, we may prefer to accept the results of the .05 test. (In this particular case, given the strength of theory and previous research, perhaps we should have initially set the significance test at the less demanding .05 level.)

Aside from Type II error, b may not appear significant because the equation misspecifies the relationship between X and Y. Perhaps the relationship follows a curve, rather than a straight line, as assumed by the regression model. First, this curvilinearity should be detectable in the scatterplot. To establish the statistical significance of the relationship in the face of this curvilinearity, regression analysis might still be applied, but the variables would have to be properly transformed. (We pursue an example of such a transformation of the end of this chapter.)

Finally, a parameter estimate may not be found significant because the variance in X is restricted. Look again at the formula for the standard error of b, S_b.

$$s_b = \sqrt{\frac{\Sigma(Y-\hat{Y})^2/(n-2)}{\Sigma(X-\bar{X})^2}}\,.$$

We can see that as the dispersion of X about its mean decreases, the value of the denominator decreases, thereby increasing the standard error of b. Other things being equal, a larger standard error makes statistical significance more difficult to achieve, as the t ratio formula makes clear. The implication is that b may not be significant simply because there is too little variation in X. (The degree of variation in X is easily checked by evaluating its standard deviation.) In such a circumstance, the researcher may seek to gather more extreme observations on X, before making any firm conclusions about whether it is significantly related to Y.

The Prediction Error for Y

In regression analysis, the difference between the observed and the estimated value of the dependent variable, $Y_i - \hat{Y}_i$, equals the prediction error for that case. The variation of all these prediction errors around the regression line can be estimated as follows:

$$s_e = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}}.$$

This s_e is called the *standard error of estimate of Y*; that is, the estimated standard deviation of the actual Y from the predicted Y. Hence, the standard error of estimate of Y provides a sort of average error in predicting Y. Further, it can be used to construct a confidence interval for Y, at a given X value. Utilizing the knowledge that the value given by the t distribution approximates 2 for a sample of almost any size, we produce the following 95% confidence interval for Y:

$$(\hat{Y} \pm 2s_e).$$

Let us take an example. In the Riverside study, we would predict someone with 10 years of education had an income of

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

How accurate is this prediction? For X = 10, we have this 95% confidence interval (s_e = 2855):

$$12{,}398 \pm 2(2855) = (12{,}398 \pm 5710).$$

According to this confidence interval, there is a .95 probability that a city employee with 10 years of education has an annual income between $6688 and $18,108. This is not a very narrow range of values. (The high extreme is almost three times the low extreme.) We conclude that our bivariate regression model cannot predict Y very accurately, for a specific value of X. Such a result is not too surprising. Recall that, according to the R^2 = .56, the model explains just over one-half the variation in Y. Our R^2 would need to be much greater, in order to reduce our prediction error substantially.

A last point merits mention. The above confidence interval, which utilizes s_e, provides a kind of “average” confidence interval. In reality, as the value of X departs from the mean, the actual confidence interval around Y tends to get larger. Thus, at more extreme values of X, the above confidence interval will be somewhat narrower than it should be. The formula for constructing this more precise confidence interval is readily available (see Kelejian and Oates, 1974, pp. 111–116).

Analysis of Residuals

The prediction errors from a regression model, $Y_i - \hat{Y}_i$, are also called *residuals.* Analysis of these residuals can help us detect the violation of certain of the regression assumptions. In a visual inspection of the residuals, we hope to observe a healthy pattern similar to that in Figure 8a; that is, the points appear scattered randomly about in a steady band of equal width above and below the regression line. Unfortunately, however, we might discover a more problematic pattern resembling one of those in Figures 8b to 8e. Below, we consider each of these troublesome patterns, in turn.

We begin with the most easily detectable problem, that of *outliers.* In Figure 8b, there are two observations with extremely large residuals, placing them quite far from the regression line. At least with regard to these observations, the linear model provides a very poor fit. By looking at a concrete example, we can explore consequences of outliers in more detail. In our Riverside study, suppose we had been careless in coding the data and recorded the incomes of Respondents 29 and 30 as $30,018 and $36,526, respectively (instead of the correct values, $20,018 and $16,526). The scatterplot, adjusted to include these erroneous values, would now look like Figure 9. By fitting a regression line to this revised plot, we see that Respondents 29 and 30 have become outliers, with residuals of 10,112 and 15,599, respectively. Further, examining the residuals generally, we note that they are out of balance around the line, that is, there are 20 negative residuals, but only 12 positive residuals. The estimated regression equation and the accompanying statistics are as follows:

$$\hat{Y} = \underset{(2438)}{2557} + \underset{(191)}{1021}X \qquad \text{(Outlier Data-Set)}$$

$$R^2 = .49 \qquad n = 32 \qquad s_e = 4647,$$

where the figures in parentheses are the standard errors of the parameter estimates; R^2 = coefficient of determination, n = sample size; and s_e = standard error of estimate of Y.

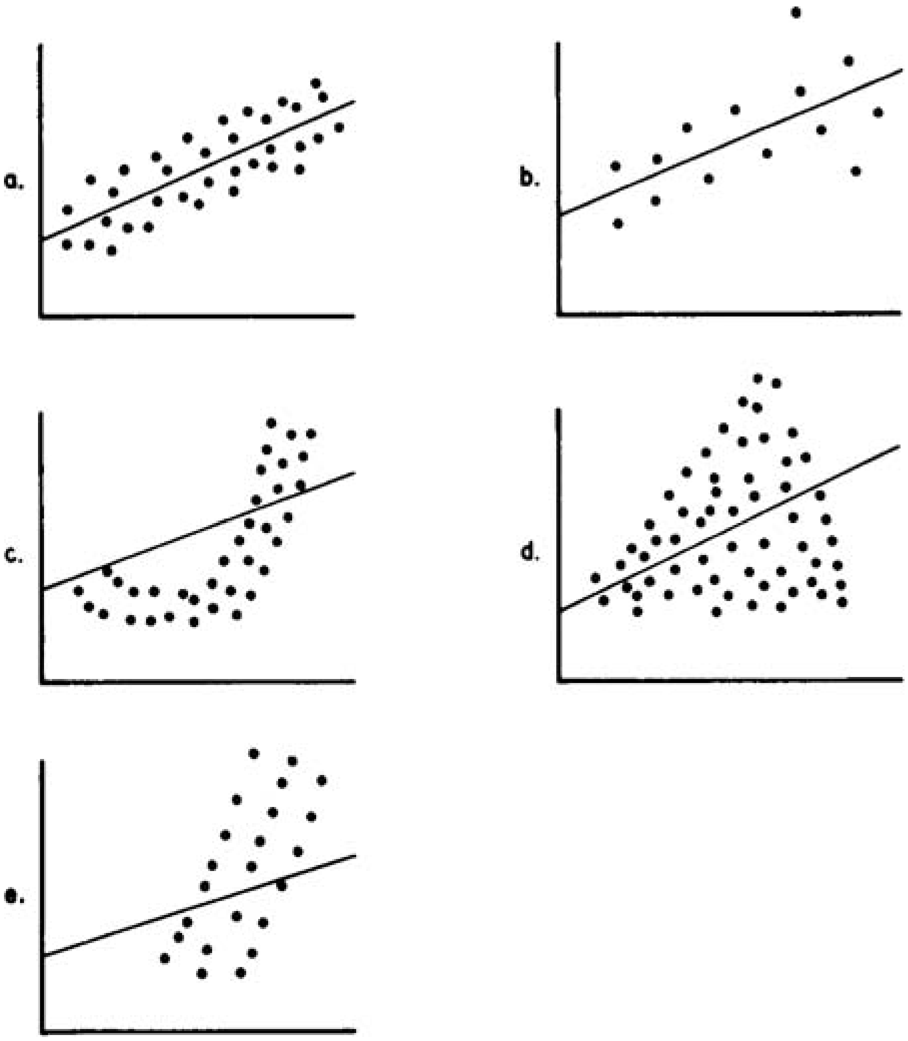
What did the presence of these outliers do to our findings? We get a good idea by comparing these “outlier data-set” estimates to our earlier “original data-set” estimates, which we repeat below:

$$\hat{Y} = \underset{(1498)}{5078} + \underset{(118)}{732}X \qquad \text{(Original Data-Set)}$$

$$R^2 = .56 \qquad n = 32 \qquad s_e = 2855,$$

where the terms are defined as above. First, note that, in an effort to accommodate the outliers, the slope is considerably elevated in the “outlier” equation. However, we would have less confidence in the accuracy of this outlier slope estimate, according to a comparison of the standard errors for b. The reduced R^2 summarizes the fact that the outlier model generally fits the data less well. The difficulties for prediction caused by the existence of the outliers is dramatically indicated by comparing the standard errors of estimate for Y, which suggests that prediction error is over 1.5 times as great under the outlier equation.

Figures 8a-e: Some Possible Patterns for Residuals

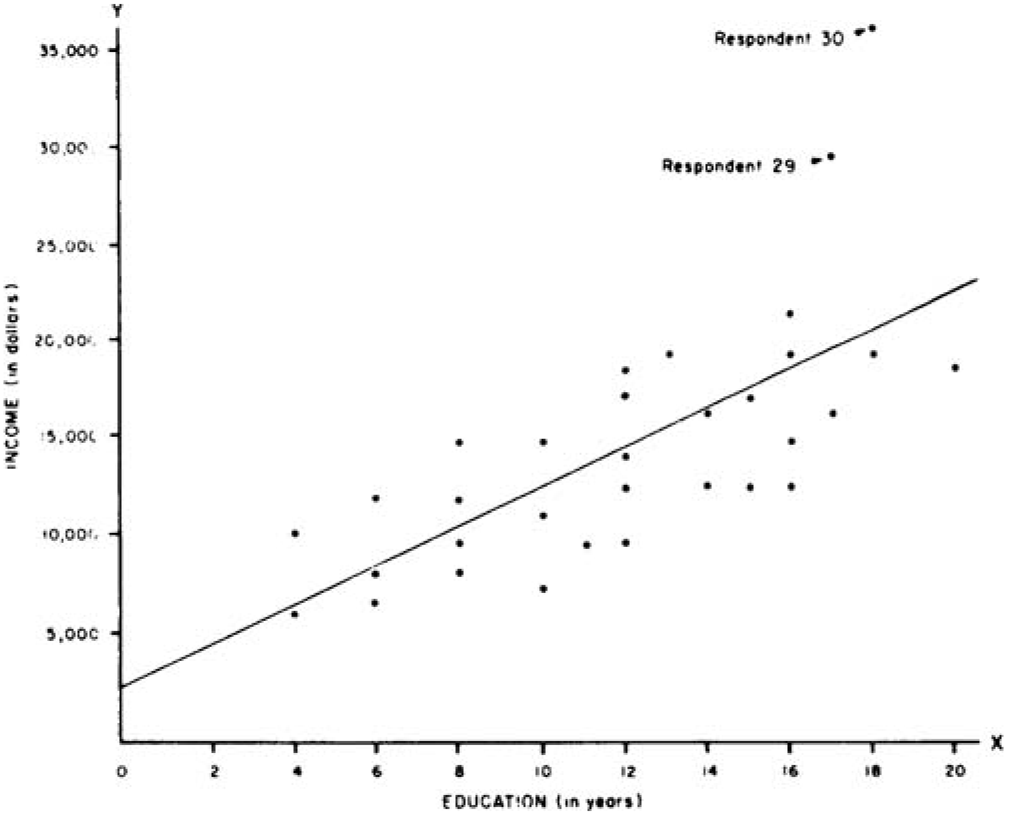


These several statistics show that the presence of outliers clearly weakens our explanation of Y. How can we adjust for outliers, in general? (We refer, of course, to actual outliers, not outliers that could be corrected by more careful coding, as in our pedagogic example.) There are at least four possibilities:

(1)
Exclude the outlying observations.
(2)
Report two equations, one with the outliers included and one without.
(3)
Transform the variable.
(4)
Gather more observations.

There are pros and cons attached to each of these possibilities. Adjustment 1 simply eliminates the problem by eliminating the outliers. The principle drawbacks are the reduction in sample size and the loss of information it entails. Adjustment 2 preserves the information that would be lost in Adjustment 1; however, it may be cumbersome to have to consider two empirically different versions of ostensibly the same model. Adjustment 3 uses only one equation, maintains the sample size, and can pull the outliers closer to the regression line. However, the results may be robbed of the straightforward interpretation possible when the variable was measured in the original units. Adjustment 4 may reveal that the outliers are not atypical cases, but in fact fit into a more general, perhaps nonlinear, pattern. An obvious limitation is that usually, in nonexperimental social science research, it is impossible to gather more observations. None of these adjustments is appropriate for every situation. Rather, in deciding on how to handle an outlier problem, we must consider our research question and the appearance of the particular scatterplot.

Figure 9: The Fitted Regression Line in the Presence of Outliers



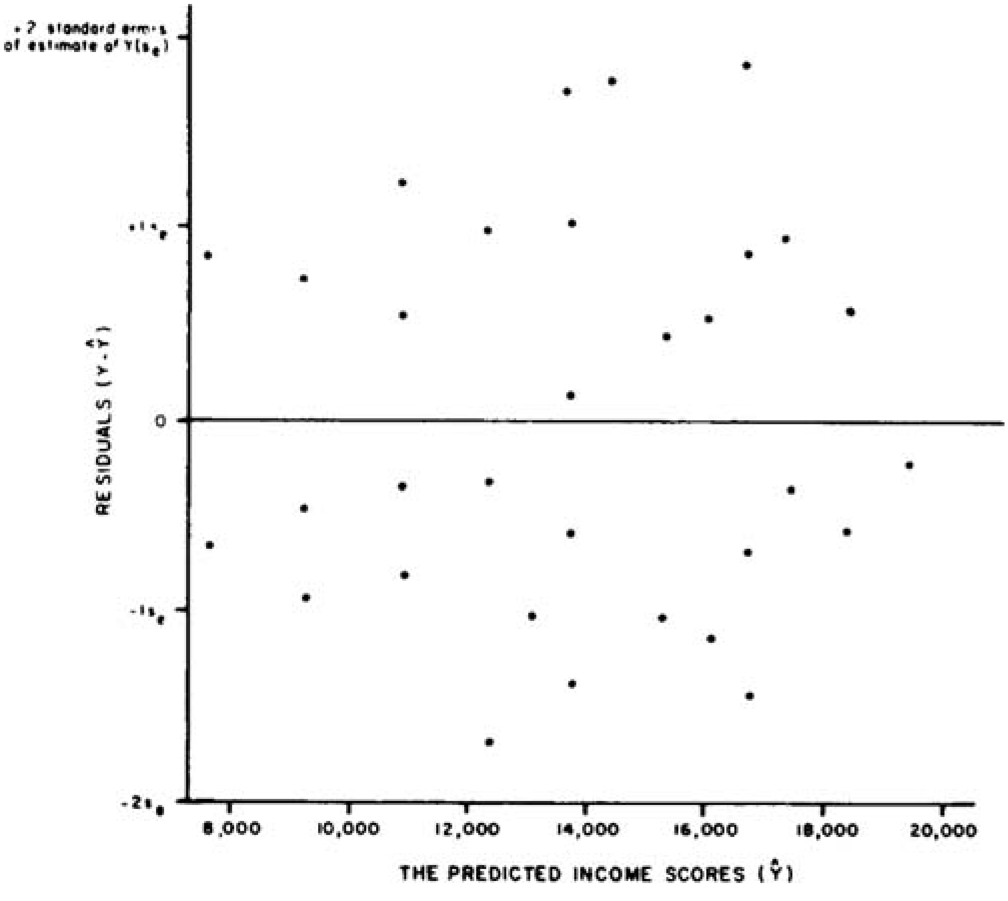
Figures 8c to 8e represent more “abnormal” residual plots. While outliers may hint at curvilinearity, it is clearly present in Figure 8c. Since regression assumes linearity, our estimators in this case are not optimal. Obviously, a unit change in X does not elicit the same response (i.e., b) in Y along the spectrum of X values. Nonlinearity can be dealt with in several ways. For example, a polynomial term might be added to the equation, or a logarithmic transformation applied to one of the variables. Of course, the approach chosen depends partly upon the shape of the particular scatterplot.

Figure 8d indicates a violation of the regression assumption of homoskedasticity. We observe that the error variance is not constant, but rather depends on the value of X; specifically, as X increases, the variation of the residuals increases. This condition of heteroskedasticity may be remedied through weighted least squares, which involves a transformation to restore the constancy of the error variance.

Figure 8e shows a linear relationship between the residuals and the predicted Y value; specifically, as Y increases, the residuals tend to move from negative to positive in sign. This implies specification error in the form of an exclusion of a relevant variable. For instance, the observations with very positive residuals may have something in common which places them higher than expected on Y. If this common factor is identified, it indicates a second independent variable for the equation.

With these above three figures (8c, 8d, and 8e) in mind, perhaps we should analyze the residuals of the original Riverside study. (We have corrected the coding error which produced the outliers.) Of course, these residuals could be examined simply by looking at the scatter around the regression line, as we have done thus far. Sometimes, however, we want to highlight them in a special plot. Figure 10 shows such a plot, where the residual values are indicated on the vertical axis, the predicted Y values are indicated on the horizontal axis. This residual plot fails to suggest any of the patterns in Figures 8c to 8e. The residuals neither follow a curve, nor do they take the shape of the heteroskedastic “fan.” Also, if there is specification error, it cannot be detected through analysis of these residuals. In sum, the pattern of residuals in Figure 10 appears free of abnormalities, forming a straight, broad band which the horizontal line cuts in half. This visual impression receives quantitative confirmation. A simple sign count reveals a virtually even balance around the line (17 negative residuals, 15 positive residuals). Further, all the residuals are scattered within a band that extends from the line plus or minus two standard errors of estimate of Y.

Figure 10: A Plot of Residuals



The Effect of Safety Enforcement on Coal Mining Fatalities: A Bivariate Regression Example

It is time to apply what we have learned to some data from the real world. A current public policy controversy concerns whether the federal government can regulate safety in the workplace. Before the 1970 passage of the Occupational Safety and Health Act, federal government involvement in occupational safety was limited to coal mining. A study of this intervention, which extends over 35 years, may shed light on the act's prospects for success. Our specific research question is, “Has federal safety enforcement reduced the rate of fatalities in the coal mines?” From various issues of the *Minerals Yearbook*, annual observations, 1932–1976, can be gathered on the U.S. coal mining fatality rate (measured as number of deaths per million hours worked). Also available, from *The Budget of the United States Government*, is the annual Bureau of Mines (currently the Mine Safety and Health Administration) health and safety budget, which pays for the federal enforcement activities, such as inspections and rescues. We use this health and safety budget, converted to constant dollars (1967 = 100), as a global measure of federal enforcement activity. A bivariate regression of the fatality rate, Y, on the safety budget X, yields

$$\hat{Y} = \underset{(36.1)}{1.26} - \underset{(-8.5)}{.0000125X}$$

$$R^2 = .63 \qquad n = 45 \qquad s_e = .19,$$

where Y = annual coal mining fatality rate (measured as deaths per million hours worked), X = annual federal coal mining safety budget (measured in thousands of constant dollars, 1967 = 100); the values in parentheses are the t ratios; R^2 = coefficient of determination; n = sample size; s_e = standard error of estimate for Y.

Safety expenditures are significantly related to the fatality rate, as a glance at the t ratio for b shows. Further, according to this slope estimate, a budgetary increase of $1 million is associated with a decrease in the fatality rate of about .01. (Some idea of the meaning of this change comes from noting that the range of the fatality rate variable is from .4 to 1.7.) Moreover, the R^2 indicates that fluctuations in the safety budget are responsible for over one-half the variation in the fatality rate. In sum, federal safety enforcement, as measured by expenditures for that purpose, seems an important influence on the coal mining fatality rate.

These estimates, although they appear pleasing, should not be accepted too readily, for we failed to look at the scatterplot. Upon inspection we discover that the linearity assumed by our regression equation is incorrect. Rather, the relationship between X and Y tends to follow a curve, as sketched in Figure 11. Fortunately, we are often able to transform the variables so as to make a relationship linear. The shape of this curve strongly suggests a logarithmic transformation is the most appropriate. Specifically, a logarithmic transformation of X will tend to “straighten out” the scatter, thus rendering the data more compatible with the linear regression assumption. Further, this transformation incorporates the knowledge gleaned from Figure 11, which is that, contrary to the interpretation from the above slope estimate, each additional dollar spent decreases the fatality rate *less and less.* (For an excellent discussion of logarithmic transformations, see Tufte, 1974, pp. 108–131.)

Figure 11: Curvilinear Relationship between the Coal Mining Safety Budget and the Coal Mining Fatality Rate

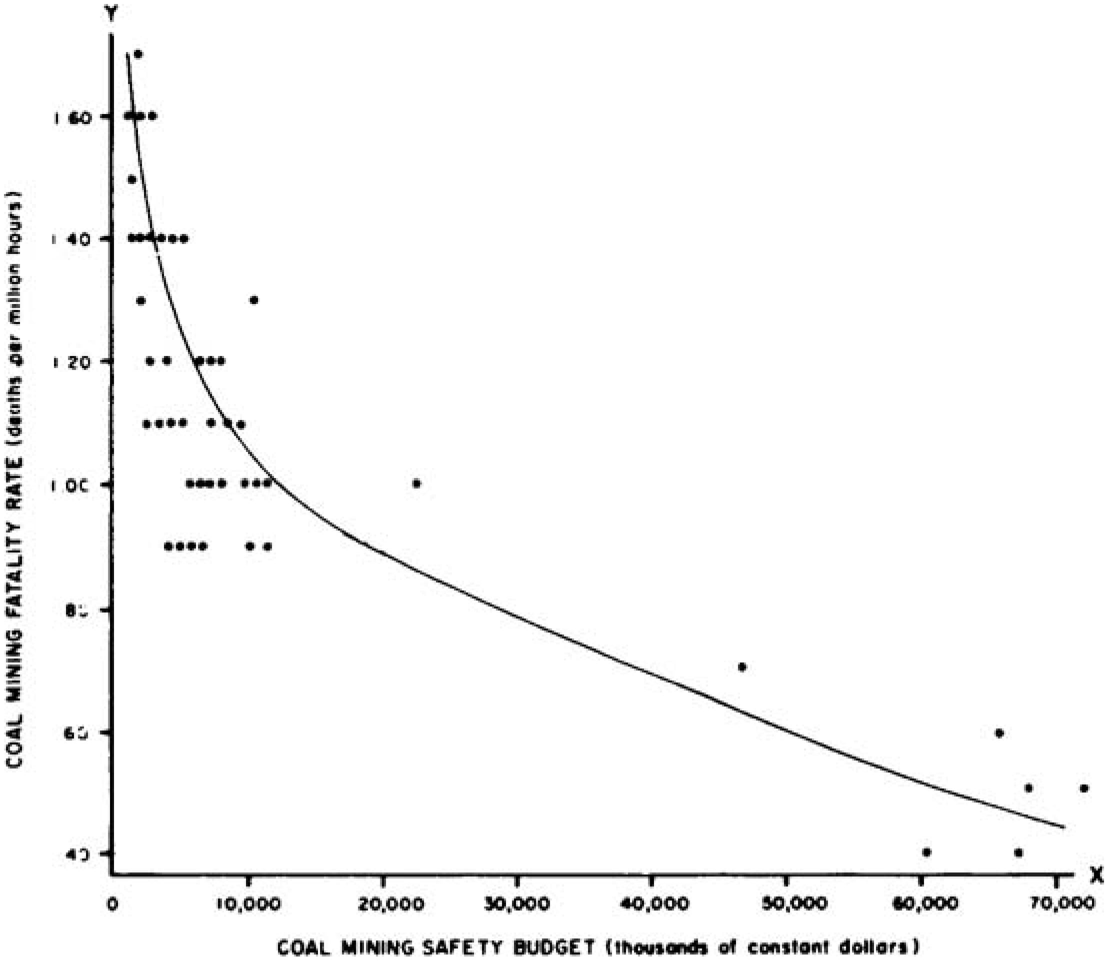


Figure 12: The Linear Relationship between the Coal Mining Safety Budget (logged) and the Coal Mining Fatality Rate

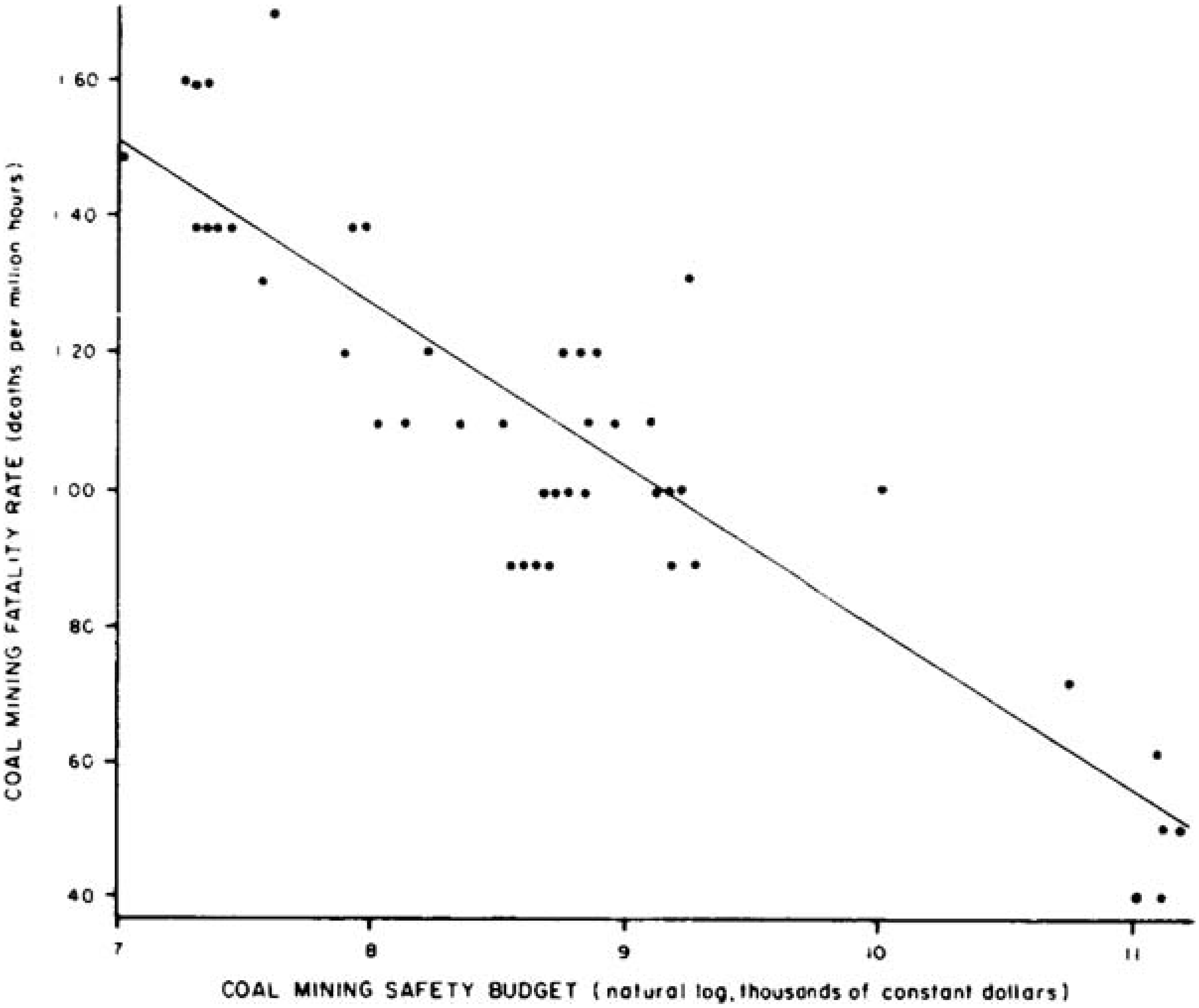


Figure 12 shows the new scatterplot, after X has undergone a "natural" logarithmic transformation, ♂n X. Reestimating the equation, but with this transformed X, yields,

$$\hat{Y} = 3.25 \quad - \quad .247\ \ell n\ X$$
$$(20.3) \quad (-13.6)$$

$$R^2 = .81 \qquad n = 45 \qquad s_e = .14,$$

where the terms are defined as above.

Our explanation of the fatality rate is considerably improved. This equation accounts for over two-thirds of the variation in Y, as the R^2 reveals. Further, the increment in R^2 from the earlier equation is large (.81 − .63 = .18), demonstrating that the curvilinearity in the relationship of safety expenditures to the fatality rate is substantial. Incorporating this curvilinearity into our model markedly enhances the predictive power of the model. In the earlier equation, when Y is predicted for a given budget, the average error is .19. This standard error of estimate for Y is reduced to .14 in our revised model. By careful examination of the original scatterplot and application of the appropriate transformation, we noticeably bettered what, at first blush, appeared to be an adequate accounting of the association between coal mining fatalities and federal safety expenditures. Of course, although safety expenditures represent an important determinant of the fatality rate, it is not the only one, as we will discover in the next chapter.

http://dx.doi.org/10.4135/9781412983440.n2