CHAPTER 3

Two-Variable Linear Regression

"Yet to calculate is not in itself to analyze."

—Edgar Allen Poe, *The Murders in the Rue Morgue*

Introduction

Fitting lines to relationships between variables is the major tool of data analysis. Fitted lines often effectively summarize the data and, by doing so, help communicate the analytic results to others. Estimating a fitted line is also the first step in squeezing further information from the data. Since the observed value can be broken up into two pieces,

observation = fitted value + residual,

we can therefore find the remaining part of the observed value that is unexplained,

residual = observation − fitted value,

and work with the residuals to discover a more complete explanation of the influences on the response variable.[1] Such was the procedure used in the study of automobile safety inspections in Chapter 1.

[1] This follows J. W. Tukey and M. B. Wilk, "Data Analysis and Statistics: Techniques and Approaches," in E. R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading, Mass.: Addison-Wesley, 1970), pp. 373–74.

We now briefly review the mechanics of linear regression. The equation of a straight line is

$$Y = \beta_0 + \beta_1 X,$$

where β_0 is the intercept and β_1 is the slope as shown in Figure 3-1. The observed data are used to estimate the two parameters, β_0 and β_1, of the model. The actual numerical *estimates* of the intercept and the slope are written as $\hat{\beta}_0$ and $\hat{\beta}_1$, where the "hats" indicate that the quantity is an estimate of a model parameter—an estimate that is computed from the observed data.

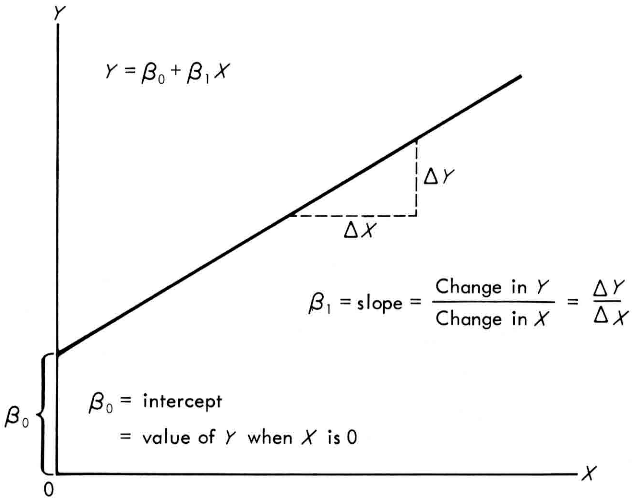

Y
$Y = \beta_0 + \beta_1 X$
ΔY
ΔX
β_1 = slope = $\frac{\text{Change in } Y}{\text{Change in } X} = \frac{\Delta Y}{\Delta X}$
β_0
β_0 = intercept
= value of Y when X is 0
0
X

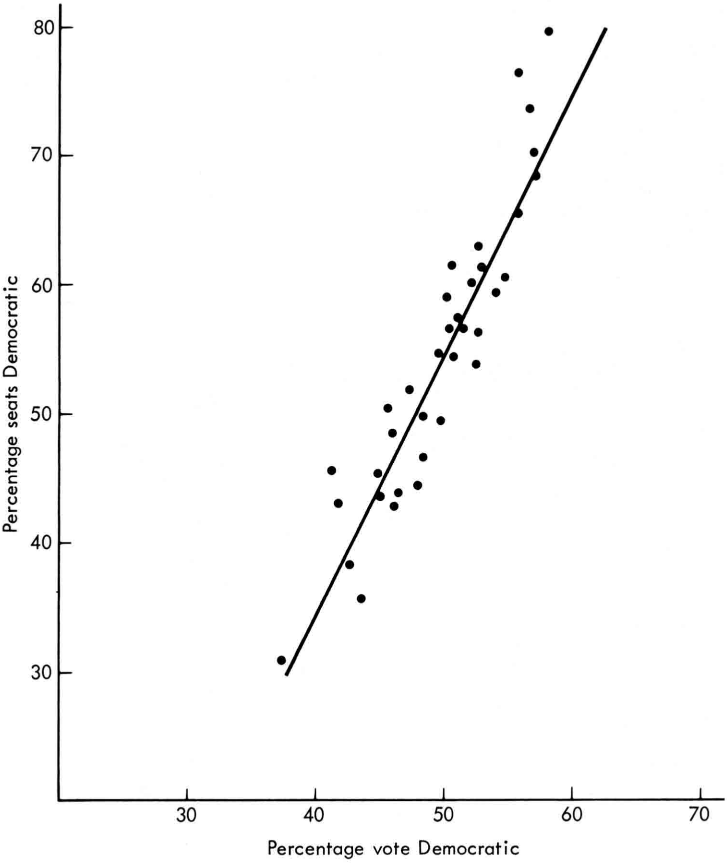
FIGURE 3-1 Equation of a straight line

The slope, a summary of the relationship between X and Y, answers the question: when X changes by one unit, by how many units does Y change? The answer is that Y changes by β_1 units. Consider the following example. In the 36 congressional elections from 1900 to 1972, the line (shown in Figure 3-2)

$$\%\ \text{seats Democratic} = -49.64 + 2.07\ (\%\ \text{votes Democratic})$$

fits the relationship between the share of congressional seats won by the Democrats and the share of votes that party received nationwide for their congressional candidates. The estimated slope, $\hat{\beta}_1$, is 2.07; that is,

$$\hat{\beta}_1 = \frac{\text{change in } Y}{\text{change in } X} = \frac{\text{change in percent of seats}}{\text{change in percent of votes}} = 2.07.$$



Percentage seats Democratic = −49.64 + 2.07 (Percentage votes Democratic)

$$Y = -49.64 + 2.07X$$

N = 37 Congressional elections, 1900 – 1972

FIGURE 3-2 Fitted line and observed data

This means that a one percent change in the share of the Democratic vote was typically accompanied by a change of 2.07 percent in the Democratic share of seats in Congress. Thus an increase of only one percent in the share of the vote was worth a substantially larger increase (of a little over two percent) in the share of seats. Of course, it works the other way, too: a drop of one percent of the vote is associated with a loss of two percent of seats. Figure 3-2 shows the data and the fitted line. In this *particular* case, the estimate of the slope measures what is called the "swing ratio"—the swing or change in seats for a given change in votes. Often, then, the substance of the problem gives a special meaning to the slope, even though the mechanics of computing the slope are the same in each case.

The estimates of the slope and the intercept are chosen so as to minimize the sum of the squares of the residuals from the fitted line. This is the principle of *least squares*, which says

$$\text{minimize} \quad \Sigma\, e_i^2,$$

$$\text{—that is, minimize} \quad \Sigma\, (Y_i - \hat{Y}_i)^2$$

in the notation of Figure 3-3.

One of the glories of the principle of least squares is that it leads immediately to specific instructions as to how to use the data to compute $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they uniquely satisfy the principle. The mathematics are found in any statistics text, where it is proved that the least-squares estimates of the slope and the intercept are computed from the observed data by

$$\hat{\beta}_1 = \frac{\Sigma\,(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma\,(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The fitted line minimizes errors in prediction when *X is used to predict Y*—and the errors in prediction are measured with respect to the *Y* variable. The estimate of the slope in this case is the *slope of the regression of Y on X*. If the roles of *X* and *Y* were reversed, and the values of *X* predicted from the variable labeled *Y*, then we would be looking at the regression of *X* on *Y*. In this second case, the errors in prediction are measured with respect to the *X* axis. Unless all the observed points fall on a 45-degree line, the two slopes are not equal. Thus the regression model is asymmetric—since the describing variable and the response variable are treated differently and different fitted lines result, depending upon which variable the researcher decides is the response variable and which is the describing variable.

Note that the question of a possible causal relationship is not decided by calling one variable the describing variable and the other the response variable. The question of causality is a separate and often difficult issue. By effectively summarizing the data, the regression analysis may sometimes provide some help in deciding if there is a causal relationship between the variables.

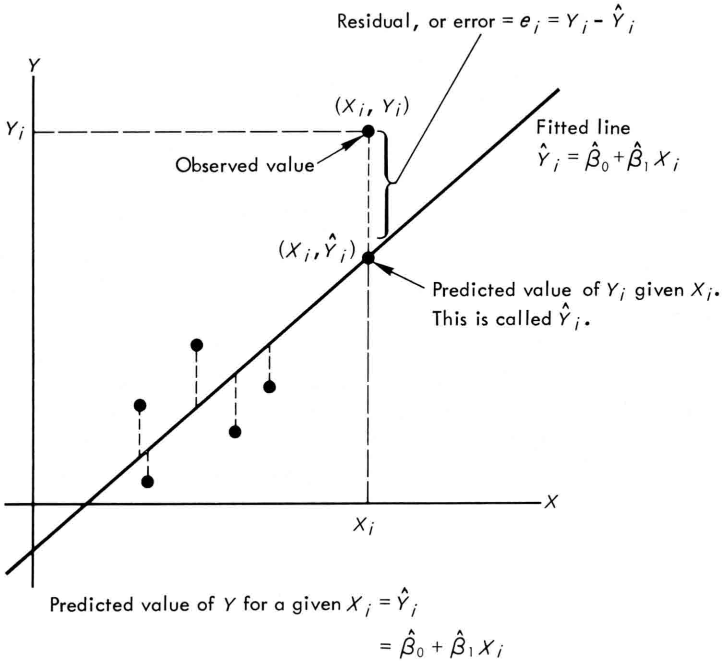

Residual, or error = $e_i = Y_i - \hat{Y}_i$
Y
Y_i
(X_i, Y_i)
Observed value
Fitted line
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
$(X_i, \hat{Y}_i)$
Predicted value of Y_i given X_i.
This is called $\hat{Y}_i$.
X
X_i
Predicted value of Y for a given $X_i = \hat{Y}_i$
$= \hat{\beta}_0 + \hat{\beta}_1 X_i$

FIGURE 3-3 Notation for least-squares regression

After fitting a line to a collection of data, the obvious question is: How well does the line fit? Here are four measures of the quality of fit:

1. the N residuals: $Y_i - \hat{Y}_i$,
2. the residual variation:

$$S^2_{Y/X} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2},$$

3. the ratio of explained to total variation:

$$r^2 = \frac{\Sigma\,(\hat{Y}_i - \bar{Y})^2}{\Sigma\,(Y_i - \bar{Y})^2},$$

4. the standard error of the estimate of the slope:

$$\frac{S_{Y|X}}{\sqrt{\Sigma\,(X_i - \bar{X})^2}}.$$

All these measures are functions of the residuals, $Y_i - \hat{Y}_i$. And all except the first are functions of the sum of squares of the residuals, $\Sigma\,(Y_i - \hat{Y}_i)^2$, which is the sum of squares minimized in estimating the parameters, β_0 and β_1, of the fitted line. Such a functional dependence is not surprising, since reasonable measures of the quality of a line's fit to the data could hardly be anything except a function of the magnitude of the errors.

The residuals are particularly useful in assessing the fit of a line, since they are measured with respect to the Y axis—that is, they are measured in the same units as the response variable.

Instead of looking at the whole collection of N residuals—for there is a residual for each observation—we can summarize them by estimating the variability about the fitted line:

$$S^2_{Y|X} = \frac{\Sigma\,(Y_i - \hat{Y}_i)^2}{N - 2}.$$

Sometimes the square root is taken, yielding the residual standard error for the fitted line.

Probably the most frequently used measure assessing the quality of fit of the line is r^2, the proportion of the variance explained. Figure 3-4 shows the components of r^2. For a given observation, $Y_i - \bar{Y}$ is the deviation of that observation from the mean, $\bar{Y}$. And $\Sigma\,(Y_i - \bar{Y})^2$ is the total variation in Y (that is, the sum of the squares of all the deviations from the mean). The describing variable seeks to predict or explain the individual deviations from the mean. The error in prediction for the ith observation is $Y_i - \hat{Y}_i$; and the error variation for all the observations is $\Sigma\,(Y_i - \hat{Y}_i)^2$. An intuitively sensible measure of the fit of the line is the ratio of this error or

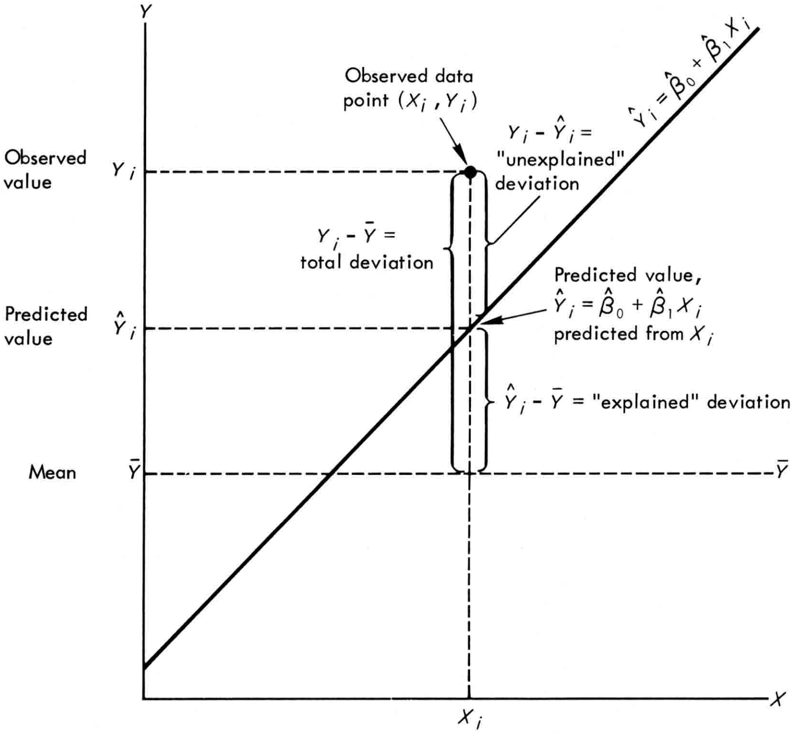

Y
Observed data point (X_i, Y_i)
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
Observed value
Y_i
$Y_i - \hat{Y}_i$ = "unexplained" deviation
$Y_i - \bar{Y}$ = total deviation
Predicted value, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ predicted from X_i
Predicted value
$\hat{Y}_i$
$\hat{Y}_i - \bar{Y}$ = "explained" deviation
Mean
$\bar{Y}$
$\bar{Y}$
X_i
X

FIGURE **3-4** Components of r^2

unexplained variation to the total variation; the smaller this ratio, the better the fit:

$$\text{one measure of fit} = \frac{\text{unexplained variation in } Y}{\text{total variation in } Y} = \frac{\Sigma\,(Y_i - \hat{Y}_i)^2}{\Sigma\,(Y_i - \bar{Y})^2}.$$

The commonly used measure, r^2, is simply this ratio subtracted from one:

$$r^2 = 1 - \frac{\Sigma\,(Y_i - \hat{Y}_i)^2}{\Sigma\,(Y_i - \bar{Y})^2}.$$

A little algebra proves that

$$\begin{pmatrix}\text{total}\\ \text{variation}\end{pmatrix} = \begin{pmatrix}\text{explained}\\ \text{variation}\end{pmatrix} + \begin{pmatrix}\text{unexplained}\\ \text{variation}\end{pmatrix}$$

or

$$\Sigma\,(Y_i - \bar{Y})^2 = \Sigma\,(\hat{Y}_i - \bar{Y})^2 + \Sigma\,(Y_i - \hat{Y}_i)^2.$$

Therefore, since

$$r^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

we have

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma\,(\hat{Y}_i - \bar{Y})^2}{\Sigma\,(Y_i - \bar{Y})^2}.$$

This interpretation of r^2, as the ratio of explained to total variation, is very common. Often r^2 is expressed in percentage terms—for example, a value of r^2 of .51 will be described as "X explained 51 percent of the variance in Y." "Explained variance," as used in the statistical jargon, refers only to the sum of squares, $\Sigma\,(\hat{Y}_i - \bar{Y})^2$. It may or may not refer to a good substantive explanation. A big r^2 means that X is relatively successful in predicting the value of Y— not necessarily that X causes Y or even that X is a meaningful explanation of Y. As you might imagine, some researchers, in presenting their results, tend to play on the ambiguity of the word "explain" in this context to avoid the risk of making an out-and-out assertion of causality while creating the appearance that something really was explained substantively as well as statistically.

If the fitted line has no errors of fit (that is, if the observed points all lie in a straight line), r^2 equals one, since there is no unexplained variation. At the other extreme, if the describing variable is no help at all in predicting the value of Y, r^2 will be near zero, since no variance is explained. In this unfortunate case, the regression line is simply $\hat{Y} = \bar{Y}$ (in other words, the predicted value of Y does not depend on the value of X).

In evaluating the fitted line, it is useful to know if the slope differs from zero. If the slope does not differ meaningfully from zero, then

X gives no help in explaining Y—the line is $\hat{Y} = \bar{Y}$. As explained in textbooks on statistics, a test of statistical significance and a confidence interval for the estimate of slope are constructed from the standard error of the estimate of the slope, which equals

$$S_{\hat{\beta}_1} = \frac{S_{Y|X}}{\sqrt{\Sigma\,(X_i - \bar{X})^2}}.$$

To conduct the test of statistical significance for $\hat{\beta}_1 \neq 0$, we consider the ratio of the estimated slope and its standard error:

$$\frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}.$$

Under appropriate statistical assumptions, this has a t-distribution, with $N - 2$ degrees of freedom. For N greater than 30, the t-distribution closely matches the normal distribution. It is this match that gives rise to the rule of thumb that a regression coefficient should be roughly twice its standard error if it is to be statistically significant at the .05 level—since, for the normal, the two-tailed .05 limits are at ± 1.96 standard deviations.

Finally, note from the denominator of the formula for $S_{\hat{\beta}_1}$ that the error in the estimate of the slope grows smaller as the variability of X increases; that is, if the observations on the X variable are spread out instead of bunched together, the standard error of the estimate of the slope will be reduced. Consequently, if there is reason to believe that there is a linear relation between X and Y and if we can control the intervals at which X is measured, then it is better to choose values of X over a fairly wide range rather than bunched up together. For example, in a study of the effects of class size on teaching effectiveness, it would be better to construct classes of size 10, 15, and 20 students rather than 13, 15, and 17. By doing so, we might obtain a more secure estimate of the relationship between size and effectiveness.

This section has outlined the statistical mechanics of two-variable linear regression. We now apply the methods to a variety of data.

Example 1: Presidential Popularity and the Results of Congressional Elections

Let us, by way of review, apply all the different statistics estimated in the linear regression model to a single problem. Figure

3-5 shows the relationship between the President's approval rating (from the Gallup Poll) shortly before the midterm congressional election and the number of seats the President's political party loses in that congressional election, from 1946 to 1970. Table 3-1 shows

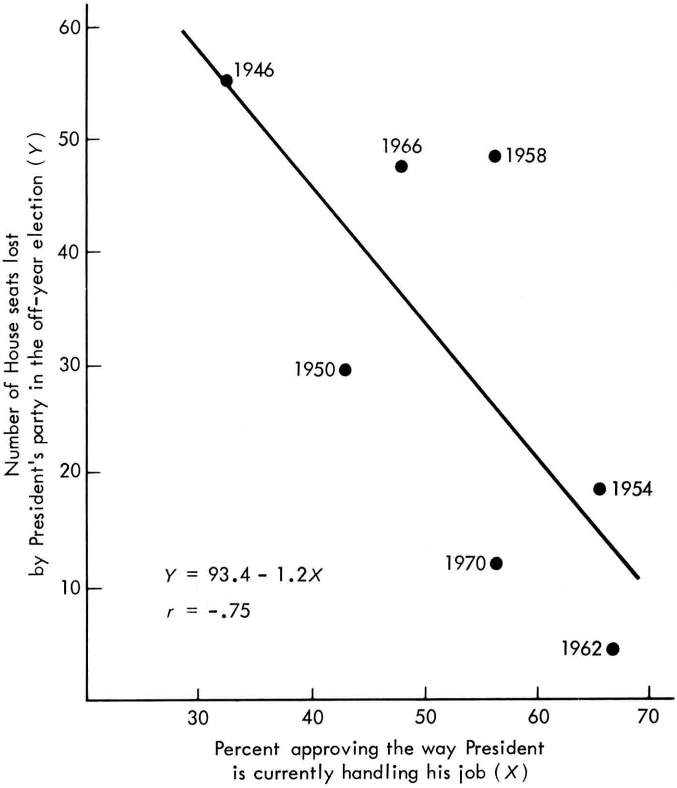


FIGURE 3-5 President's approval rating vs. his party's seat loss

the details of the data. Note that the political party of the President lost seats in each of the seven midterm elections from 1946 to 1970. Sometimes the loss was small—in 1962, for example, the Democrats lost only four seats in the House of Representatives compared to what they had in 1960. In other elections, many seats were lost: the Democrats suffered a decline of 55 Congressional seats in 1946. The Republicans, under President Eisenhower, had a bad year in the 1958 midterm elections, losing 48 seats.

Is, then, the extent of the loss of congressional seats by the President's

TABLE 3-1
Congressional Seats and Presidential Popularity

Year	Seats held in House of Representatives by		Seats lost in midterm election by President's party
	Democrats	*Republicans*	
1944	243	190	
1946	188	246	Democrats lost 55
1948	263	171	
1950	234	199	Democrats lost 29
1952	213	221	
1954	232	203	Republicans lost 18
1956	234	201	
1958	283	153	Republicans lost 48
1960	262	175	
1962	258	176	Democrats lost 4
1964	295	140	
1966	248	187	Democrats lost 47
1968	243	192	
1970	255	180	Republicans lost 12

Year		President's popularity rating early September in off-year elections (percent approve)[a]
1946	Truman	32%
1950	Truman	43%
1954	Eisenhower	65%
1958	Eisenhower	56%
1962	Kennedy	67%
1966	Johnson	48%
1970	Nixon	56%

SOURCE: *Gallup Political Index,* October 1970, No. 64, page 16.

[a] Percent approve + percent disapprove + percent no opinion = 100 percent. The question is worded as follows: "Do you approve or disapprove of the way Blank is handling his job as President?"

party related to the approval rating of the President?[2] The correlation between popularity and seat loss is, for the seven elections, −.75,

[2] Two papers dealing with the issues raised by these data are: Angus Campbell, "Voters and Elections: Past and Present," *Journal of Politics,* 26 (November 1964); 745–57, and John E. Mueller, "Presidential Popularity from Truman to Johnson," *American Political Science Review,* 64 (March 1970), 18–34. See also, for a more sophisticated discussion, Douglas A. Hibbs, Jr., "Problems of Statistical Estimation and Casual Inference in Dynamic, Time-Series Regression Models," in Herbert Costner, ed., *Sociological Methodology, 1973–1974* (San Francisco: Jossey-Bass, 1974), ch. 10.

indicating that the lower the President's popularity, the more seats his party loses in the off-year elections. This is, for most political research at least, a rather strong, impressive correlation—although note that the correlation coefficient doesn't tell us *how much* a decline in the approval rating is associated with a loss of *how many* seats. The regression coefficient does, however, provide some help with this. The equation of the least-squares line is

$$\text{seats lost} = 93.36 - 1.20\ (\text{percent approving President})$$

Figure 3-5 shows this line. The slope is -1.20, indicating that a one percent decline in the percent approving the current president is associated with a loss of about 1.2 seats in the upcoming off-year election. That regression coefficient is statistically significant:

$$t = \frac{\text{estimate of regression coefficient}}{\text{standard error}} = \frac{-1.20}{.48} = -2.50,$$

which, for five degrees of freedom, $(N - 2 = 7 - 2 = 5)$ exceeds the one-tailed t-value at the .05 level (-2.02).

Furthermore, the President's approval rating explains a good deal of the statistical variation in the outcome of the election:

$$r = -.75, \qquad r^2 = .56.$$

Thus the regression statistically explains 56 percent of the variation in the shifts in congressional seats.

All in all, this is a fairly impressive regression—a good correlation, a substantively meaningful regression coefficient that is statistically significant, and more than half the variance explained. Since it is so good, perhaps we can use the model for predictive purposes: taking the pre-election approval rating for the President and plugging into the regression equation to come up with an estimate of the loss of seats in the congressional election. This is all very nice, except that the prediction will not be a very secure one. Let us evaluate the quality of predictions based on the fitted line.

One way to get an idea of the predictive properties of the model is to look at the estimate of the variability about the line, the residual variance:

$$S^2_{Y|X} = \frac{\Sigma\,(Y_i - \hat{Y}_i)^2}{N - 2}.$$

The numerator is simply the unexplained variation. Taking the square root puts this statistic into the units in which the response variable, Y, is measured:

$$S_{Y|X} = 13.3 \text{ seats},$$

which is a rather large standard error in terms of predicting seats—especially when we start to consider confidence intervals of ± two standard errors.

Or, to evaluate the predictive quality of the model, we might look directly at the residuals for each year of the observed data. Table 3-2 shows the computations. Once again, we see pretty substantial errors in prediction from the observed data—and, of course, the model itself is estimated so as to minimize the sum of squares of these residuals.

In short, then, we have here the beginnings of a good explanatory model, but it still needs improvement if it is to be useful for predictive purposes. How might we build a better, more complete model? Consider a model that also takes into account the economic conditions—for which some voters might hold the President and his party responsible—prevailing at the time of the election:

$$\text{seats lost} = \beta_0 + \beta_1 \begin{pmatrix}\text{presidential}\\ \text{popularity}\end{pmatrix} + \beta_2 \begin{pmatrix}\text{economic}\\ \text{conditions}\end{pmatrix}.$$

Just as in the two-variable case, this three-variable model is estimated by least squares. Such a multiple regression, as it is called, will be examined in Chapter 4.

TABLE 3-2
Residual Analysis

Year	Y_i = *observed seat loss by President's party*	X_i = *Presidential approval rating*	$\hat{Y}_i$ = *predicted seat loss for a given* X_i, $\hat{Y}_i = 93.4 - 1.20X_i$	*Residual*[a] = *observed* − *predicted* = $Y_i - \hat{Y}_i$
1946	55 seats	32%	93.4 − 1.2(32) = 55	55 − 55 = 0 seats
1950	29 seats	43%	93.4 − 1.2(43) = 42	29 − 42 = −13 seats
1954	18 seats	65%	93.4 − 1.2(65) = 15	18 − 15 = 3 seats
1958	48 seats	56%	93.4 − 1.2(56) = 26	48 − 26 = 22 seats
1962	4 seats	67%	93.4 − 1.2(67) = 13	4 − 13 = −9 seats
1966	47 seats	48%	93.4 − 1.2(48) = 36	47 − 36 = 11 seats
1970	12 seats	56%	93.4 − 1.2(56) = 26	12 − 26 = −14 seats

[a] Note that if residual > 0, the President's party lost more seats than predicted; if residual < 0, the President's party lost less seats than predicted.

Example 2: Lung Cancer and Smoking

THE FITTED LINE

Figure 3-6 shows the relationship between the death rate from lung cancer in 1950 and the cigarette comsumption in eleven countries in 1930. Cigarette consumption is lagged twenty years behind the death rate on the assumption that the carcinogenic consequences of smoking require a considerable length of time to show up. The fitted regression line is

$$\begin{bmatrix}\text{lung cancer deaths}\\ \text{per million people}\\ \text{in 1950 } (Y)\end{bmatrix} = .23\begin{bmatrix}\text{cigarettes consumed}\\ \text{in 1930 } (X)\end{bmatrix} + 66,$$

$$\text{standard error of slope} = .07 \qquad r^2 = .54$$

The regression indicates that when cigarette consumption in 1930 from one country to another is greater by, say, 500 cigarettes per year per person, the lung cancer rate apparently increased by about 115 deaths per million in 1950.

SCALING OF VARIABLES AND INTERPRETATION OF REGRESSION COEFFICIENTS

Note that in order to make an accurate interpretation of the regression coefficients, we must keep track of the units of measurement of each variable. For example, if the lung cancer rate were expressed as deaths per 100,000 people (instead of per 1,000,000), then the regression coefficient would be reduced by a corresponding factor of ten down to .023. This coefficient, although it is numerically smaller, reflects only the change in the scaling of the death rate—and the coefficient has exactly the same substantive meaning and importance as the original coefficient of .23. This obvious point is worth keeping in mind because some research reports are not particularly clear in reporting the units of measurement associated with each regression coefficient—and the reader must dig out the units of measurement and the scaling of the variables from the footnotes.

ANOTHER FITTED LINE: A REGRESSION WITHOUT THE UNITED STATES

A further look at the scatterplot shows the rather strong effect of one extreme point in shifting the fitted line. The line is pulled down

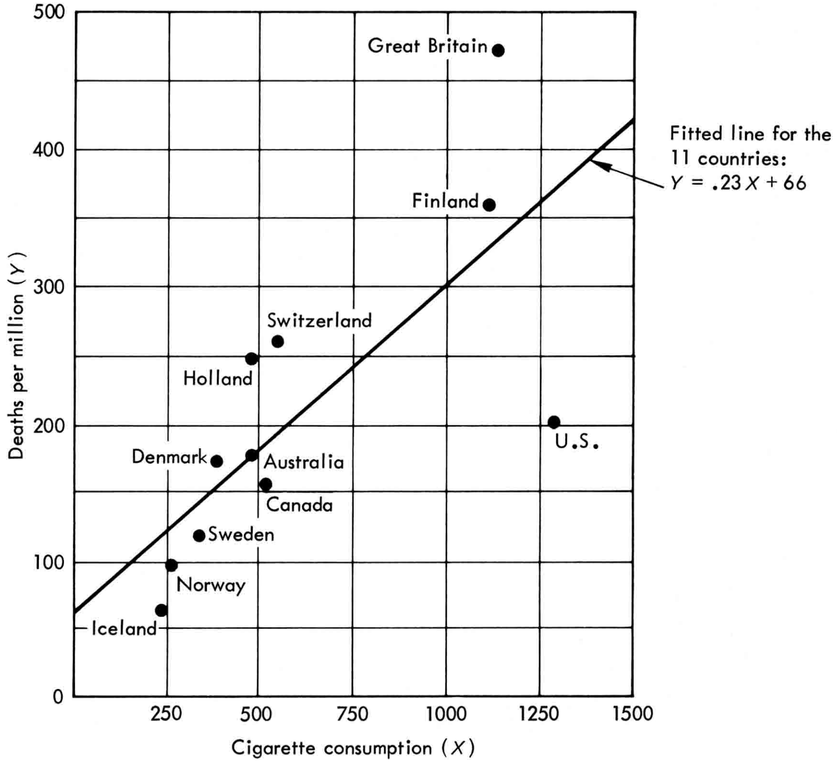


FIGURE 3-6 Crude male death rate for lung cancer in 1950 and per capita consumption of cigarettes in 1930 in various countries

SOURCE: R. Doll, "Etiology of Lung Cancer," *Advances in Cancer Research,* 3 (1955), reprinted in *Smoking and Health,* Report of the Advisory Committee to the Surgeon General (Washington: USGPO, 1964), p. 176.

by the low death rate for the United States. Removing that country from the data and computing a new regression line based on the remaining ten countries yields quite a different fitted line:

N = 10 Countries (Without U.S.)	*N = 11 Countries (With U.S.)*
$Y = .36X + 14$ $r^2 = .89$ Standard error of slope = .05 Dotted line in Figure 3-7	$Y = .23X + 66$ $r^2 = .54$ Standard error of slope = .07 Solid line in Figure 3-7

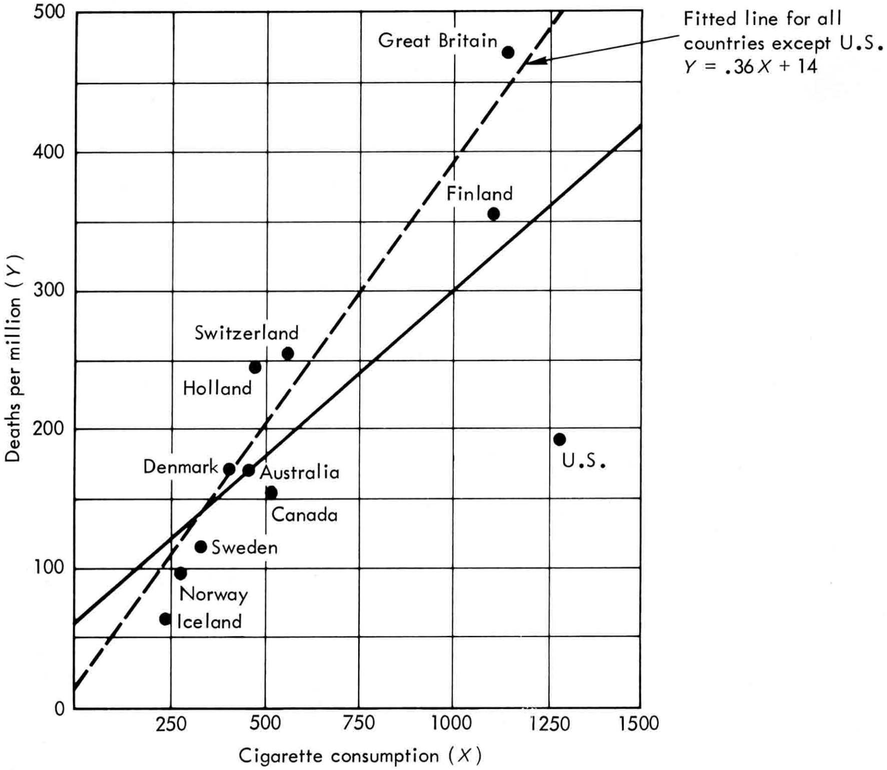


FIGURE 3-7 Lung cancer and cigarette consumption: fitted line for ten countries, omitting the United States

Note the great improvement in the explained variance in the regression based on the ten countries; a straight line really fits the ten quite well. Perhaps we should look more carefully into the conditions that make for a somewhat lower death rate than expected, given the amount of tobacco consumed, in the United States. That will be done below.

WHAT IF NOBODY SMOKED? INTERPRETING THE INTERCEPT

Let us return to consideration of the original regression for all eleven countries. Can we find out what the lung cancer rate might have been if there had been no smoking? Not very well with these particular data—for several reasons.

First, there is simply no experience at all with any countries consuming less tobacco per capita than Iceland, at 220 cigarettes per year per person in 1930. Obviously we want to be careful in extending our results beyond the range of the data; some of the particular problems of extrapolation are discussed in Chapter 2.

Second, one naive way to answer the question meets some difficulties after a careful examination of the scatterplot. The naive approach is to set cigarette smoking at zero in the fitted regression equation and see what the lung cancer rate is. That rate is simply the intercept, 66 deaths per million per year. But note the pattern of countries down at the low end with respect to smoking: the three lowest countries have negative residuals, all lying below the fitted regression line. Thus, in the countries with a low consumption of cigarettes, there is some indication that a better-fitting curve would bend more sharply downward; thus the straight line imposed on the data is a bit misleading at the low end of the scale. This suggests that the rate would be considerably lower than 66 if nobody smoked. Perhaps a better estimate would be around 14 deaths per million—the intercept for the regression line that excluded the United States. The exclusion of that outlying value seems appropriate in estimating the intercept, since the outlier is far from the region of interest and since the residuals near the region of interest indicate that the extreme point has shifted the regression line based on all the countries.

Note finally that the line is literally imposed on the data—and just because we do the computations necessary to produce a slope and an r^2, does not, of course, necessarily mean that the straight line is the best curve to fit to the data or that the two variables are, in fact, related in a linear fashion. In a later example, we will use "linear" regression to fit some other curves to data.

What kind of data *would* satisfactorily estimate the death rate from lung cancer if nobody smoked cigarettes? First, we need data based on individuals—smokers and nonsmokers—to make comparisons of lung cancer rates. Second, it is important to make sure that people susceptible—perhaps because of genetic or environmental factors—to lung cancer are not also people who are more likely to smoke. Thus we might compute the lung cancer rate for many different sorts of people who are smokers or nonsmokers. Such differential rates for different population groups could then be adjusted to the population as a whole to estimate the lung cancer rate if, contrary to fact, no one smoked.

ANALYZING THE RESIDUALS

Table 3-3 displays the original data, along with the predicted values for the lung cancer rate (predicted on the basis of cigarette consumption) and the errors made in the prediction for each country. Note

TABLE 3-3
Residual Analysis

Country	Y_i = *observed lung cancer deaths per million in 1950*	X_i = *cigarettes consumed per capita in 1930*	$\hat{Y}_i$ = *predicted lung cancer death rate for a given* X_i, $\hat{Y}_i = .23X_i + 66$	*Residual = observed − predicted* = $Y_i - \hat{Y}_i$
Iceland	58	220	.23(220) + 66 = 116	58 − 116 = −58
Norway	90	250	.23(250) + 66 = 123	90 − 123 = −33
Sweden	115	310	.23(310) + 66 = 137	115 − 137 = −22
Canada	150	510	.23(510) + 66 = 183	150 − 183 = −33
Denmark	165	380	.23(380) + 66 = 153	165 − 153 = 12
Australia	170	455	.23(455) + 66 = 170	170 − 170 = 0
United States	190	1280	.23(1280) + 66 = 359	190 − 359 = −169
Holland	245	460	.23(460) + 66 = 171	245 − 171 = 74
Switzerland	250	530	.23(530) + 66 = 187	250 − 187 = 63
Finland	350	1115	.23(1115) + 66 = 321	350 − 321 = 29
Great Britain	465	1145	.23(1145) + 66 = 328	465 − 328 = 137

the large residuals for Great Britain and the United States and the negative residuals for the smaller values of tobacco consumption. The residuals add up to zero; the sum of the squared residuals is the smallest it can be—no other line can improve over the least-squares line in minimizing the sum of the squares of the residuals. These two properties of the residuals—

(1) $\Sigma\,(Y_i - \hat{Y}_i) = 0$, and
(2) $\Sigma\,(Y_i - \hat{Y}_i)^2$ is minimized

—are properties of all least-squares lines.

A further analysis of the residuals can be made by plotting the residuals against the predicted values ($\hat{Y}$) as shown in Figure 3-8. Sometimes such a display yields up more information because the reference line is a horizontal line rather than the tilted line fitted to the original scatterplot. Contemplation of the residuals reveals large errors in the prediction of the death rate for Great Britain and the United States. Great Britain had a much higher death rate than the United States in 1950, although the per capita consumption of cigarettes in the two countries in 1930 was roughly equal. What differences between the two countries might account for the differences in lung cancer death rates even though the tobacco consumption was roughly the same? A few possibilities include:

1. Differences in air pollution between the two countries.

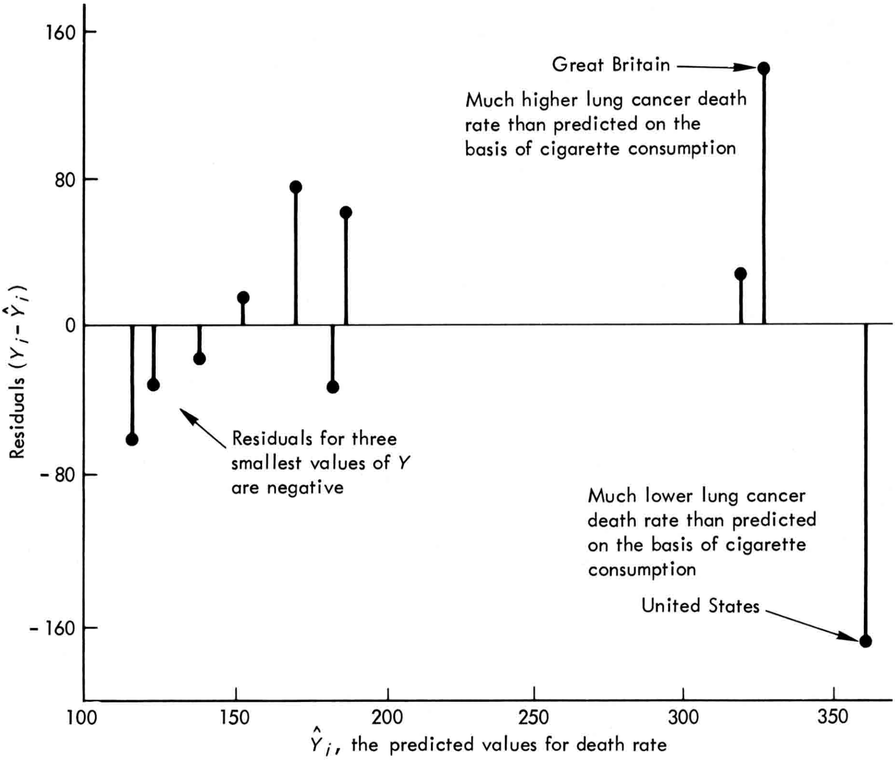


FIGURE 3-8 Residuals vs. predicted values, lung cancer and smoking

2. Differences in the age distribution of the populations of the two countries. Since lung cancer occurs more frequently among older smokers, the rate of cancer might well be higher in a country that had a larger share of older people.

3. Differences in smoking habits (such as smoking cigarettes right down to the end) that expose the lungs to different doses of smoke from each cigarette consumed. Observers have reported that the British often smoke their cigarettes right down to the very end (probably because cigarettes are heavily taxed and very expensive in England) and also that the British tend to be "drooper" smokers—they let the cigarette droop from the mouth rather than placing it in an ashtray or holding in the hand. Some researchers compared the lengths of discarded cigarette butts in the two countries and discovered rather large differences in length, the American discards being considerably

longer (30.9 mm) than the British (18.7 mm).[3] Other studies found that "the mortality rate for lung cancer in England was especially high for the smokers who 'drooped' the cigarettes off the lip while they smoked, a habit which may result in the delivery of a greater dose of smoke from each cigarette."[4]

4. Differences in the composition of the tobacco.

5. Differences in the factors which mute or accentuate the health consequences of smoking. For example, construction workers and others exposed to the insulating material asbestos who also smoke have a very high risk of lung ailments—a much higher risk than expected by merely adding up the excess risk from smoking plus the excess risk from working with asbestos. (This extra risk coming from the *combination* of the two factors is called, in the statistical jargon, an "interaction effect.") Thus if more smokers in a country were exposed to asbestos, then that country would have a higher rate of lung cancer than expected on the basis of tobacco consumption alone.

6. Differences across countries in what medical symptoms doctors define or describe to be lung cancer.

VALUE OF THESE DATA AS EVIDENCE

These data have only a very modest value as evidence bearing on the relationship between smoking and lung cancer. Since the data are *aggregate, countrywide* figures, they provide very indirect evidence concerning the relationship between smoking and health among *individuals*. Furthermore, eleven data points aren't much to work with—and the exclusion of a single observation shifted the variance explained from 54 percent to 89 percent, indicating the sensitivity of the analysis to outlying observations.

A big worry about the sort of data presented in Figures 3-6 and 3-7 is *selection*—how were the eleven countries included in the analysis chosen from all the countries of the world? Why these eleven? Would the results be the same if more countries were selected? Or eleven different countries? With so few data points, the analysis is very fragile; just a couple of fresh observations divergent from the fitted line would cause the whole relationship to fall apart. Careful, if manipulative, selection of data points can easily generate pseudo-rela-

[3] Report of the Advisory Committee to the Surgeon General of the Public Health Service, *Smoking and Health* (Washington, D.C.: U.S. Government Printing Office, 1959), p. 177.

[4] *The Health Consequences of Smoking, 1969 Supplement to the 1967 Public Health Service Review* (Washington, D.C.: U.S. Government Printing Office), p. 57.

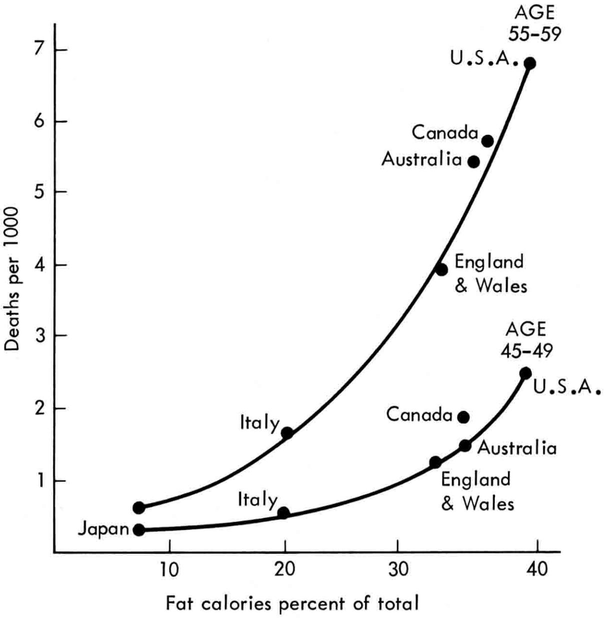


FIGURE 3-9 Mortality from degenerative heart disease (1948–1949, men) in relation to fat calories consumed
SOURCES: Yerushalmy, *op. cit.* and Keys, *op. cit* (see p. 87).

tionships. Yerushalmy points out such an example:

> Another important error often encountered in the literature is the fallacy of utilizing evidence supporting a given hypothesis and neglecting evidence contradicting it. An illustration is shown in Figure [3-9]. In this case, the investigator selected six countries and correlated the percent of fat in the diet with the mortality of coronary heart disease in these six countries. . . . On the face of it, the correlation appears very striking, and indeed the author in reviewing the data in Figure [3-9] makes the following strong statement: "The analysis of international vital statistics shows a striking feature when the national food consumption statistics are studied in parallel. Then it appears that for men aged 40 to 60 or 70, that is, at the ages when the fatal results of atherosclerosis are most prominent, there is a remarkable relationship between the death rate from degenerative heart disease and the proportion of fat calories in the national diet. A regular progression exists from Japan through Italy, Sweden, England and Wales, Canada, and Australia to the United States. No other variable in the mode of life besides the fat calories in the diet is known which shows anything like such a consistent relationship to the mortality rate from coronary or degenerative heart disease."

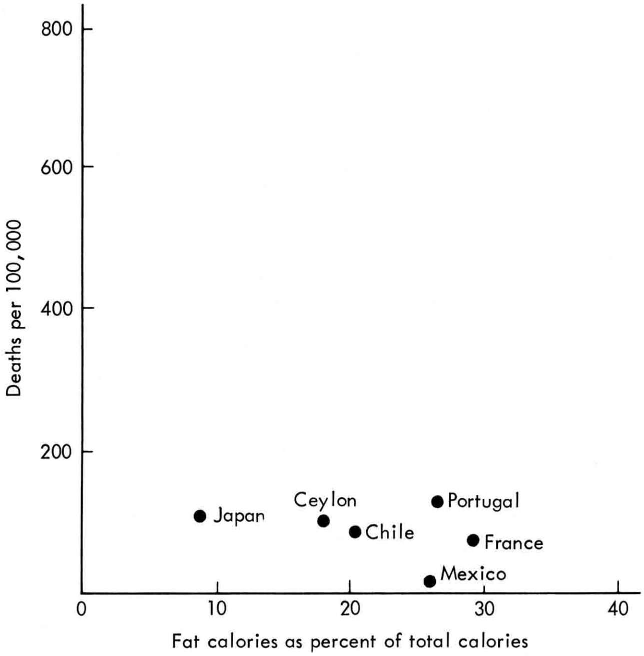


FIGURE 3-10 Six countries selected for equality in mortality from coronary heart disease, but differing greatly in consumption of fat calories in percent of total calories
SOURCE: Yerushalmy, *op cit.* (see p. 87).

The question arises how were these six countries selected. Further investigation reveals that these six countries are not representative of all countries for which the data are available. For example, it is easy enough to select six other countries which differ greatly in their dietary fat consumptions, but have nearly equal death rates from coronary heart disease [Figure 3-10]. Similarly, six other countries were easily selected which consumed nearly equal proportions of dietary fat, but which differed widely in their death rates from coronary heart disease [Figure 3-11]. This tendency of selecting evidence biased for a favorable hypothesis is very common. For example, investigations among the Bantu in Africa are often mentioned in support of the dietary fat hypothesis of coronary heart disease, while observations on other African tribes, Eskimos, and other groups which do not support the hypothesis are generally ignored.

However, even when these errors are avoided and the studies are well conducted, the conclusions which may be derived from observational studies have great limitations stemming primarily from noncomparability of the self-formed groups. The phenomenon of self-selection is the root of many of the difficulties. Were all other complications eliminated, the inequalities between groups which result from self-selection would still leave in doubt inferences on causality. For example, in the study of the relationship of cigarette smoking to health, if we assume well-conducted investigations in which (a) large random samples of the population have been selected and the individuals correctly identified as smokers, nonsmokers, or past smokers, (b) the problem of nonresponse did not exist, (c) the population had been followed long enough to identify all cases of the disease in question, (d) no problems of misdiagnosis and misclassification existed, (e) and no one in the population had been lost from observation, then even under these ideal conditions, the inferences that may be drawn from the study are limited because the individuals being observed, rather than the investigator, made for themselves the crucial choice: smoker, nonsmoker, or past smoker.[5]

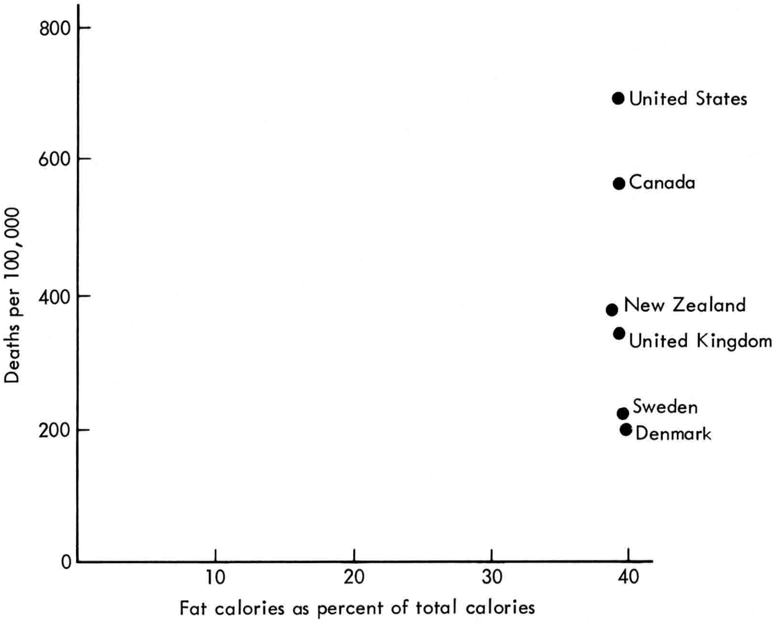


FIGURE 3-11 Six countries selected for equality in consumption of fat calories in percent of total calories, but differing greatly in mortality from coronary heart disease

SOURCE: Yerushalmy, *op. cit.*

[5] J. Yerushalmy, "Self-Selection—A Major Problem in Observational Studies," in Lucien M. Lecam, Jerzy Neyman, and Elizabeth L. Scott, eds., *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Biology and Health, Volume IV* (Berkeley and Los Angeles, California: University of California Press, 1972), pp. 332–33. The internal quotation is from A. Keys, "Atherosclerosis—A Problem in Newer Public Health," *Journal of Mt. Sinai Hospital,* 20 (1953), 134.

Still another reason for not taking our little analysis as serious evidence is that much better data are available to answer questions concerning the relationship between smoking and health. Smoking is probably the most carefully investigated public health problem there is; a vast amount of information has been gathered from health interviews with many people over many years, from autopsies, hospital records, animal studies, and so on. In other fields, where the amount and variety of evidence is less and the resources for collecting new data scarcer, the evidence of the sort examined here might represent the best available information and, furthermore, theories would have to stand or fall and decisions be made in the faint light of such analysis. Thus the overall importance of a particular piece of analysis varies in relation to what other evidence there is that bears on the question at hand.

Example 3: Increase in the Number of Radios and Increase in the Number of Mental Defectives, Great Britain, 1924–1937

The table shows a measure of the number of radios in the United Kingdom from 1924 to 1937 and the number of mental defectives per 10,000 people for the same years. These data form the basis for the discussion of "nonsense correlations" by the famous British statisticians, G. Udny Yule and M. G. Kendall.

The fit of the line is remarkably good, with a bit over 99% of the variation in number of mental defectives "explained" (in a statistical sense!) by the growth in the number of radios. Note the small, but systematic variation in the residuals, with the points weaving around the fitted line in clusters above and then below the fitted line. These "wrinkles" in the residuals might be worth pursuing if this were more than a nonsense correlation.

Why does this extremely strong, although nonsensical, relationship come about? This is a relationship formed by relating two increasing time series. In other words, the number of radios is increasing over time and also the number of mental defectives is increasing over time. Millions of other things are increasing over the time period from 1924 to 1937, including the population, the number of smokers, military expenditures in Europe, the number of patents issued, and the number of letters in the first name of the Presidents of the United States (Calvin, Herbert, and Franklin). For example, consider this regression:

Year	*Number of radio receiver licenses issued (millions)*	*Number of notified mental defectives per 10,000 of estimated population*
1924	1.350	8
1925	1.960	8
1926	2.270	9
1927	2.483	10
1928	2.730	11
1929	3.091	11
1930	3.647	12
1931	4.620	16
1932	5.497	18
1933	6.260	19
1934	7.012	20
1935	7.618	21
1936	8.131	22
1937	8.593	23

Figure 3-12 displays the regression line fitted to the above data:

$$\begin{array}{l}\text{number of mental}\\ \text{defectives per}\\ \text{10,000}\end{array} = 2.20 \left[\begin{array}{l}\text{number of radios}\\ \text{(in millions)}\end{array}\right] + 4.58,$$

$r^2 = .99$, standard error of slope = .08.

$$\begin{array}{l}\text{number of mental}\\ \text{defectives per 10,000}\\ \text{in the United}\\ \text{Kingdom, 1924–1937}\end{array} = 5.90 \left(\begin{array}{l}\text{number of letters}\\ \text{in the first name}\\ \text{of the President}\\ \text{of U.S., 1924–1937}\end{array}\right) - 26.44,$$

$r^2 = .89$, standard error of slope = .66.

Yule and Kendall further observe:

> . . . it might be argued that the period in question was one of great technical progress in many scientific fields; that one effect of this movement was the development of broadcasting and the general spread of the practice of listening evinced by the increased number of [radio] licenses taken out; that another effect was the greater interest in psychological ailments and increased facilities for treatment, resulting in either more discoveries of mental defect or greater readiness to submit cases to medical notice. Whether this is the right explanation is doubtful, but it is a possible rational explanation of what at first sight seems absurd.[6]

[6] G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* 14th ed., (London: Charles Griffin, 1950), p. 315–16.

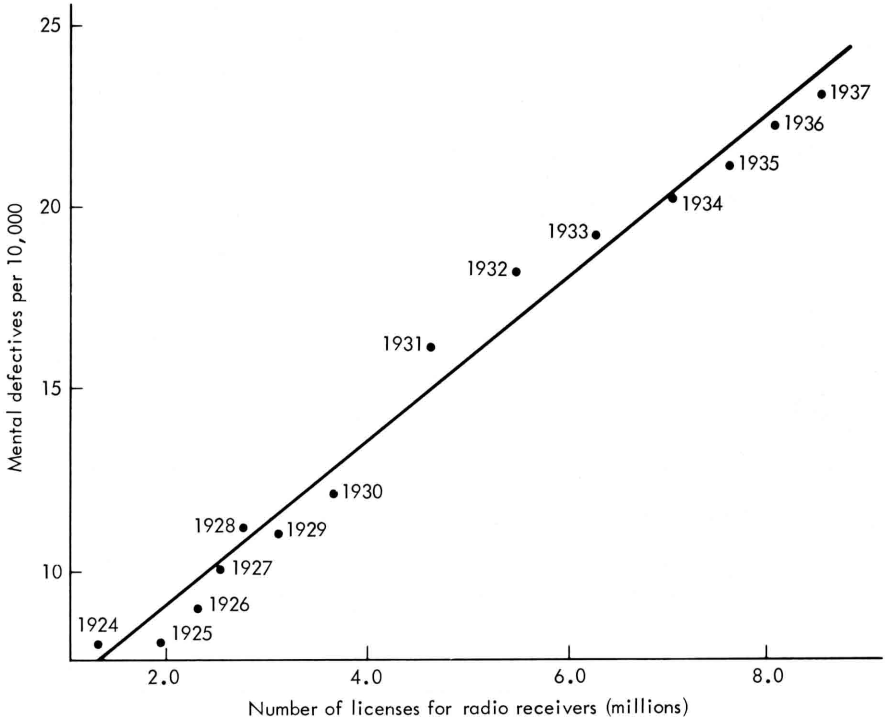


FIGURE 3-12 Radio receivers and mental defectives

Whether listening to the radio produced mental defectives (or, perhaps, whether the increase in number of mental defectives led to a greater demand for radios) is not answered by this regression of two increasing time series. And the relationship between the number of British mental defectives and the first names of American Presidents during 1924 to 1937 does not gain in credibility because the length of the name "explained" 87 percent of the variation in the number of mental defectives. What is clear, however, is that:

1. Even very high values of "explained" variance can occur without the slightest suspicion of a causal relationship between variables. There are times when a high value for r^2 might increase our degree of belief that there is a causal relationship, but this depends upon the substantive nature of the problem.
2. If nonsense goes into a statistical analysis, nonsense will come out. The nonsensical output will have all the statistical trappings, will look just as official, just as "scientific," and just as "objective" as a substantively useful regression. It is, however, the substance and not the form that is the important thing. As Justice Holmes

> once wrote: "The only use of forms is to present their contents, just as the only use of a pint pot is to present the beer . . . and infinite meditation upon the pot will never give you the beer."

We have now seen regression techniques applied to several problems—automobile safety inspections, smoking and lung cancer, and radios and mental problems. These examples all served to illustrate certain aspects of the logic and mechanics of fitting a line to the relationship between two variables. It is now time to examine a more extensive regression analysis in action, going into detail on a serious problem. Such is our next application.

Example 4: The Relationship between Seats and Votes in Two-Party Systems[7]

Arrangements for translating votes into legislative seats almost always work to benefit the party winning the largest share of the votes. That the politically rich get richer has infuriated the partisans of minority parties, encouraged those favoring majority parliamentary rule, and, finally, bemused a variety of statisticians and political scientists who have tried to develop parsimonious descriptions and explanations of the inflation of the legislative power of the victorious party. Here we will use a linear regression model to describe how the votes of citizens are aggregated into legislative seats and also to estimate the bias in an electoral system.

Figure 3-13 shows the data used in the analysis.[8] These six scatterplots indicate that the relationship between seats and votes in most two-party systems displays four obvious characteristics:

1. As a party's share of the vote increases, its share of the seats also increases in a fairly regular fashion.

[7] A more extended version of this material appeared in Edward R. Tufte, "The Relationship Between Seats and Votes in Two-Party Systems," *American Political Science Review*, 68 (June 1974), 540–54.

[8] The election tabulations were collected from state and national yearbooks. The U.S. congressional returns have been collected together in Donald Stokes and Gudmund Iversen, "National Totals of Votes Cast for Democratic and Republican Candidates for the U.S. House of Representatives, 1866–1960," July 1962, mimeo, Survey Research Center, University of Michigan. *Congressional Directories* (Washington, D.C.: U.S. Government Printing Office) were used to update the Stokes-Iversen compilation and also as the source for tabulations requiring election returns in individual congressional districts. All percentages of the vote were computed from the votes received by the two major parties only.

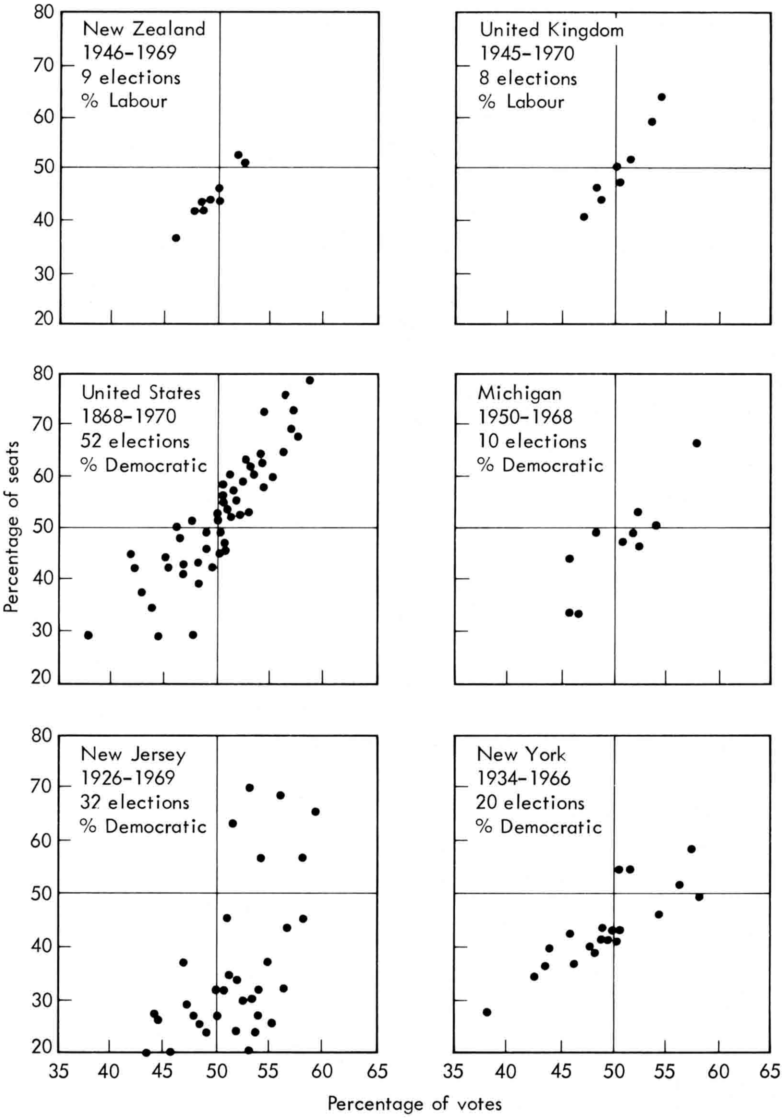


FIGURE 3-13 Seats and votes

2. The party that receives a majority of the votes usually receives a majority of parliamentary seats. Such was the case in 93 percent of the national elections and 53 percent of the state elections examined here. The points in the upper left and lower right quadrants represent those elections in which the party winning a majority of votes failed to take a majority of seats. New Jersey, like many other states prior to redistricting (and some after redistricting), shows many markedly biased outcomes, with the Democrats often winning fully three-fifths of the votes but less than one-third of the seats.
3. A party that wins a majority of votes generally wins an even larger majority of seats.
4. In most elections (100 percent in this series), the winning party receives less than 65 percent of the votes (although it may receive a much larger share of seats).

Even a casual inspection of the data displayed in Figure 3-13 ındicates that almost any curve with a slope around two or three in the region from 35 to 65 percent of the vote for a party will fit the relationships rather well. Let us now examine the regression model.

The relationship between seats and votes is described most directly by a simple linear equation:

$$\begin{pmatrix}\text{percentage of seats for}\\ \text{a given political party}\end{pmatrix} = \beta_1 \begin{pmatrix}\text{percentage of votes}\\ \text{for that party}\end{pmatrix} + \beta_0 .$$

The estimate of the slope, $\hat{\beta}_1$, measures the percentage change in seats corresponding to a change of one percent in the votes for a party. Thus $\hat{\beta}_1$ estimates the *swing ratio* or the *responsiveness* of the partisan composition of parliamentary bodies to changes in the partisan division of the vote in two-party systems. For example, the swing ratio during the last twelve U.S. congressional elections is 1.9, indicating that a net shift of 1.0 percent in the national vote for a party has typically been associated with a net shift of 1.9 percent in congressional seats for a party.

In addition, the fitted line provides an estimate of another important parameter of the electoral system: the bias for or against a particular party in the translation of votes into seats. Setting the percentage of seats at 50 percent and solving for the percentage of votes in the equation of the fitted line tells one the share of the vote that a party typically needs in order to win a majority of seats in the legislative body. The difference between this number and 50 percent is the *bias* or *party advantage,* as illustrated in Figure 3-14. For

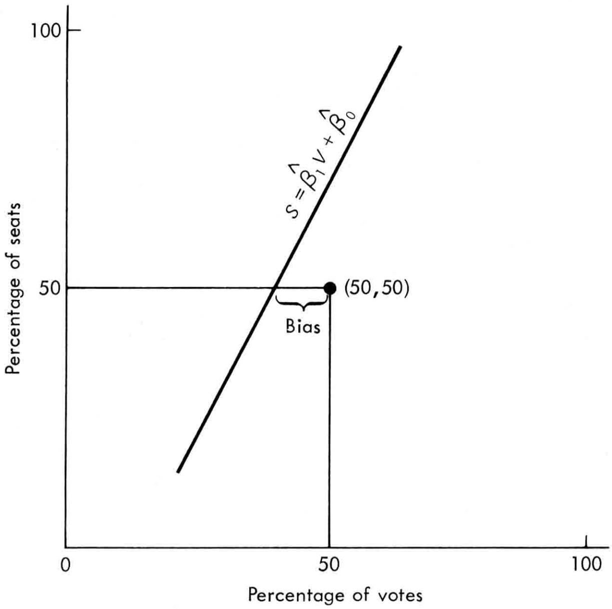


FIGURE 3-14 The fitted seats-votes line

example, in recent congressional elections, the Democrats have typically needed only about 48 percent of the national vote in order to win a majority of House seats; thus the bias or party advantage is about 2 percent. Later we will explain some of the variations in the swing ratio and bias for different electoral systems over the years.

Note that we are using the estimate of the slope in the linear model in order to estimate the swing ratio; the analogue of the intercept in the linear model is, in this case, the bias. Thus both the parameters estimated by the linear regression model are useful in this analysis.

One minor defect of the linear fit is that in general the fitted line will not pass through the end points (0 percent votes, 0 percent seats) and (100 percent votes, 100 percent seats), which are on the seats-votes curve by definition. Although slightly inelegant, this shortcoming is hardly troublesome—especially since parties in two-party systems almost never get less than 35 percent of the vote nor more than 65 percent of it.[9] The clear advantage of the linear fit

[9] A "logit" model dealing with this problem is described in Example 6 of this chapter.

is that it yields two politically meaningful numbers, the swing ratio and the bias, that can be compared over time and electoral systems.

Table 3-4 records the fitted lines for a variety of elections. The swing ratios and the biases show considerable variation both between electoral systems and within some systems over time. Among the countries, Great Britain has the greatest swing ratio at 2.8. In the United States the swing ratio has been about two, although, as we shall see later, there is evidence that in the last few elections the swing ratio has decreased considerably. The U.K. electoral system shows little bias; in the United States a persistent bias has favored

TABLE 3-4
Linear Fit for the Relationship between Seats and Votes

	$\hat{\beta}_1$ *Swing ratio and (standard error)*	r^2	*Percentage votes required to give the indicated party a majority of seats in the legislature*	*Advantaged party and amount of advantage*
Great Britain, 1945–1970	2.83 (.29)	.94	50.2% Labour	Conservatives, .2%
New Zealand, 1946–1969	2.27 (.27)	.91	51.4% Labour	National, 1.4%
United States, 1868–1970	2.39 (.21)	.71	49.1% Democrats	Democrats, 0.9%
United States, 1900–1970	2.09 (.14)	.87	48.0% Democrats	Democrats, 2.0%
United States, 1948–1970	1.93 (.29)	.81	48.8% Democrats	Democrats, 1.2%
Michigan, 1950–1968	2.06 (.41)	.76	52.1% Democrats	Republicans, 2.1%
New Jersey, 1926–1947	2.10 (.44)	.53	61.3% Democrats	Republicans, 11.3%
New Jersey, 1947–1969	3.65 (.89)	.63	52.0% Democrats	Republicans, 2.0%
New York, 1934–1966	1.28 (.19)	.73	54.3% Democrats	Republicans, 4.3%

the Democratic party—partially the result of that party's victories in small congressional districts and in districts with low turnouts. In Michigan, New Jersey, and New York there have been large biases favoring the Republicans and a great deal of variation in swing ratios. The relationship between votes and seats is weaker for the three states than for the three countries; in fact, in the states during some time periods there was virtually no correlation between the share of seats that a party won in the legislature and the share of votes it had received at the polls! In more recent elections, however, there was a fairly strong relationship between seats and votes in all three states—probably the result of new rules and practices for districting.

THE SWING RATIO IN RECENT CONGRESSIONAL ELECTIONS

We now examine changes in the swing ratio in elections for the U.S. House of Representatives. Table 3-5 shows estimates of swing ratio and bias for congressional elections for the last hundred years. It appears that a shift—in fact, a rather striking shift—in the relationship between seats and votes has taken place in the last decade. The 1966–1970 triplet displays the second lowest swing ratio of the 17 election triplets since 1870. No doubt the recent elections provide a somewhat narrow range of electoral experience; the Democrats won with votes between 50.9 and 54.3 percent (a range in votes that is the fifth smallest of the 17 triplets). Until the Republicans control Congress or the Democrats win more decisively, the "new" swing ratio and bias will not be well estimated. The bias is a spectacular 7.9 percent, reflecting the two close votes that yielded the Democrats a substantial party majority in the House. The estimate of the bias for the 1966–1970 election triplet is, however, somewhat more insecure than for previous blocs of elections because the error of the estimated bias is proportional to the reciprocal of the swing ratio—and in this case the swing ratio is moderately small.

Compared with all the other performances of the electoral systems examined here, a system with a swing ratio of .7 and a bias of 7.9 percent describes a set of electoral arrangements that is both quite unresponsive to shifts in the preferences of voters (as expressed in their party votes for their representatives) and, at the same time, badly biased. How did the low value of the swing ratio for 1966–1970 come about? Certainly the Democratic party, after their substantial gain in votes (3.4 percent) and relatively tiny gain—given the "normal" swing ratio exceeding 2.0—in seats (3.2 percent) would like to know what happened in 1970. And for Republicans, 1966 and 1968 need

TABLE 3-5
Three Elections at a Time: Estimates of Swing Ratio and Bias

Years of elections	*Swing ratio*	*Percentage of votes to elect 50% seats for Democrats*	*Size of Democratic party advantage*
1870–74	6.01	51.4%	−1.4%
1876–80	1.48	50.0%	.0%
1882–86	3.30	50.8%	−.8%
1888–92	6.01	50.9%	−.9%
1894–98	2.82	51.7%	−1.7%
1900–04	2.23	50.1%	−.1%
1906–10	4.21	48.8%	1.2%
1912–16	2.39	48.8%	1.2%
1918–22	1.96	47.6%	2.4%
1924–28[a]	−5.75[a]	40.8%[a]	9.2%[a]
1930–34	2.28	45.9%	4.1%
1936–40	2.50	47.1%	2.9%
1942–46	1.90	48.1%	1.9%
1948–52	2.82	49.5%	.5%
1954–58	2.35	50.1%	−.1%
1960–64	1.65	47.4%	2.6%
1966–70	.71	42.1%	7.9%

[a] The figures estimated for the 1924–1928 election triplet are peculiar because of the extremely narrow range of variation in the share of the vote (42.1, 41.6, and 42.8 percent) during that period. The average range within an election triplet is about 6 percent.

explanation: after all, they managed to make the national division of the vote very close but in neither year were they able to win even 45 percent of the House seats.

The swing ratio indicates the potential for turnover in representation. The smaller the swing ratio, the less responsive the party distribution of seats is to shifts in the preferences of voters. The extreme case is a swing ratio near zero; such a flat seats-votes curve means that the distribution of seats does not change with the distribution of votes. Figure 3-15 shows the strong relationship between the swing ratio and the turnover in the House of Representatives for election triplets since 1870. Note the steady drift downward over the years in both the swing ratio and the turnover. Since 1948, the swing ratio has shifted from 2.8 to 2.4 to 1.7, and, most recently, to 0.7. Similarly the turnover in the House has declined, reflecting the long-run decrease in the intensity of competition for congressional seats.[10]

One element in the job security of incumbents is their ability to exert significant control over the drawing of district boundaries; indeed, some recent redistricting laws have been described as the Incumbent Survival Acts of 1974. It is hardly surprising that legislators, like businessmen, collaborate with their nominal adversaries to eliminate dangerous competition. Ironically, reapportionment rulings have given incumbents new opportunities to construct secure districts for them-

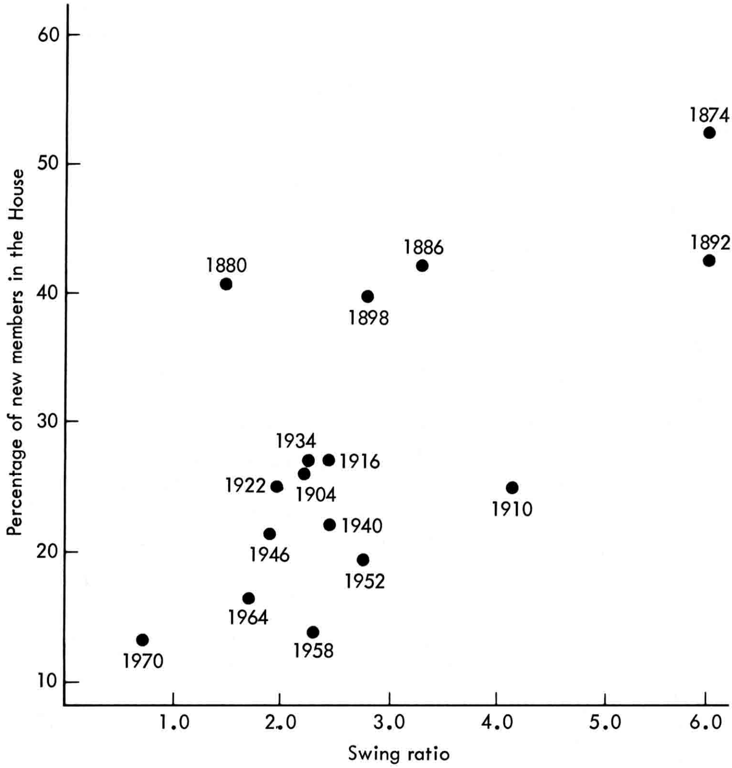
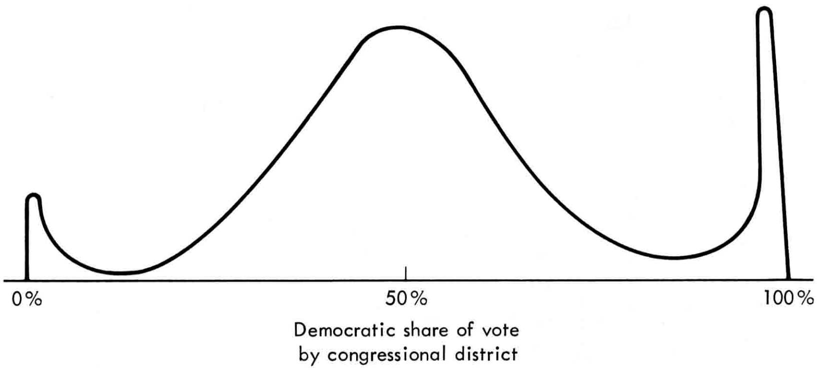


FIGURE 3-15 Turnover and swing ratio

[10] For example, Nelson W. Polsby, "The Institutionalization of the U.S. House of Representatives," *American Political Science Review,* 62 (March 1968), 144–68; and David R. Mayhew, "Congressional Representation: Theory and Practice in Drawing the Districts," in *Reapportionment in the 1970s,* ed. N. Polsby, pp. 249–90.

selves, leading to a reduction in turnover that is, in turn, reflected in the sharply reduced swing ratio of the last few elections. One apparent consequence is the remarkable change in the shape of the distribution of congressional votes in recent elections. Prior to 1964, the congressional vote by district was distributed the way everyone expects votes to be distributed: a big clump of relatively competitive districts in the middle, tailing off away from 50 percent with some peaks at the ends of the distribution for districts without an opposition candidate:



In recent elections the shape of the distribution of the vote by district has changed; Figure 3-16 shows the movement of district outcomes away from the danger area of 50 percent in recent years—note the development of bimodality in the 1968 and 1970 district vote compared to previous years (the left peak contains the Republican safe seats; the right peak contains the Democratic safe seats). Perhaps the best way to see how this pattern developed over time is to array the vote distributions over the years and riffle through them—like an old-time peep show—and watch the middle of the distribution sag and the areas of incumbent safety bulge in the more recent elections.

Many states, in part through recent reapportionments, have practically eliminated political competition for congressional seats—even compared to the relatively small proportion of competitive seats in the past. In the 1970 elections in Michigan, for example, not one of the 19 districts was a close contest; the *most* marginal Republican victor won 56 percent of the vote and the *most* marginal Democrat won fully 70% of the vote in his district. In Illinois, the most closely