CHAPTER 4

Multiple Regression

"Some circumstantial evidence is very strong, as when you find a trout in the milk."

—Henry David Thoreau

The Model

In chapter 3 we estimated the two-variable model,

$$\text{Loss by President's party in midterm congressional elections} = \beta_0 + \beta_1 \text{(presidential approval rating)},$$

and decided that a more elaborate model would help explain additional variation in the response variable. The more elaborate version used two describing variables, presidential approval and economic conditions:

$$\text{vote loss} = \beta_0 + \beta_1 \text{(presidential approval)} + \beta_2 \text{(economic conditions)}.$$

Just as in the two-variable case, we can use the data to estimate the three parameters of this model:

1. the constant term, β_0,
2. the regression coefficient for presidential popularity, β_1,
3. the regression coefficient for economic conditions, β_2.

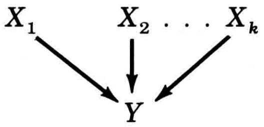
And, as before, the parameters are estimated by least squares, minimizing the sum of the squared deviations of the observed value from the fitted value:

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

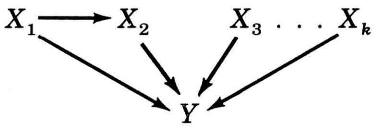
This is the multiple regression model. We can have more than two describing variables: the general multiple regression model with k describing variables is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdot \quad \cdot \quad \cdot + \beta_k X_k.$$

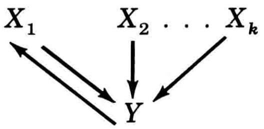
The causal model behind multiple regression is that there are k multiple, independent causes of Y, the response variable:



This is a somewhat limited model, since it excludes estimates of links *between* the describing variables—for example,



Also simple multiple regression models do not estimate feedback relationships:



Under some circumstances, models involving feedback and simultaneous relationships can be estimated.

Multiple regression is widely used in the study of economics, politics, and policy. It allows the inclusion of many describing variables in a convenient framework. It is a carefully investigated and fairly widely understood statistical procedure; thus it is a relatively effective way to communicate the results of a multivariate analysis. And packages for running multiple regressions are available with most every computer.

Almost all of the technical apparatus used in the two-variable model applies to the multivariate case. Consider the three-variable model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

We use the data to compute:

1. the estimated regression coefficients, $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$;
2. their standard errors, $S_{\hat{\beta}_1}$ $S_{\hat{\beta}_2}$;
3. t-values to test for statistical significance of the coefficients, $\hat{\beta}_1 / S_{\hat{\beta}_1}$, $\hat{\beta}_2 / S_{\hat{\beta}_2}$;
4. the ratio of explained variation to total variation, R^2.

The estimated coefficients of the model generate the predicted values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i},$$

where X_{1i} and X_{2i} are the observed values of X_1 and X_2, respectively, for the ith case. Now, since we have an observed and a predicted value for each observation, the residuals are defined as usual, measured along the Y axis:

$$Y_i - \hat{Y}_i,$$

and $\Sigma\,(Y_i - \hat{Y}_i)^2$ is minimized in the estimates of β_0, β_1, and β_2. No other set of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ will make the sum of the squared deviations smaller. As in the two-variable case, the principle of least squares generates the estimating equations for the coefficients. And, as in the two-variable case, a variety of assumptions about the data are required for the sound application of statistical significance testing in the model.[1]

The percentage of the variance explained statistically is also analogous to the two-variable case:

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

[1] See Ronald J. Wonnacott and Thomas H. Wonnacott, *Econometrics* (New York: Wiley, 1970); J. Johnston, *Econometric Methods*, 2d ed. (New York: McGraw-Hill, 1972); or other statistics or econometrics texts for discussion of the assumptions.

Since the R^2 provides some measure of the quality of overall fit of the describing variables in predicting Y, it is sometimes used to choose between different regressions containing different combinations of describing variables.

R can also be interpreted as the simple corrrelation between the observed and predicted values; that is,

$$R = r_{Y\hat{Y}}.$$

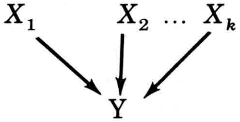
The estimated regression coefficients in a multiple regression are interpreted as *partial slopes*. They try to answer the question: When X_i, the ith describing variable, changes by one unit and all the other describing variables are held constant (in a statistical sense), how much change is expected in Y? The answer is β_i units. If the describing variables were completely unrelated to one another, then the regression coefficients in the multiple regression would be the same as if each describing variable were regressed one at a time on Y. However, the describing variables are inevitably interrelated, and thus all the coefficients in the model are estimated and examined in combination.

Two different types of regression coefficients—unstandardized and standardized—are used in practice. Unstandardized coefficients are interpreted in the units of measurement in which the variables are measured; for example, a one percent change in votes is associated with a β_1 percent change in seats. Standardized coefficients rescale all the variables into standard deviations from the mean:

$$\frac{Y_i - \bar{Y}}{S_Y}, \frac{X_{1i} - \bar{X}_1}{S_{X_1}}, \frac{X_{2i} - \bar{X}_2}{S_{X_2}}, \ldots .$$

Thus, in the standardized case, all variables are expressed in the same units—that is, in standard deviations. Standardized regression coefficients are analogous to the correlation coefficient in the two-variable case; unstandardized coefficients are analogous to the slope in the two-variable case. Standardized coefficients are useful when the natural scale of measurement does not have a particularly meaningful interpretation or when some relative comparison of the variables with respect to their standard deviations is needed. All the examples presented here use unstandardized regression coefficients.

The regression coefficients gain their meaning from the substance of the problem at hand. The statistical model merely provides the answer to the question: *Under the assumption that X_i is a cause* *of* Y, what is the expected change in Y for a unit change in X_i? Thus the estimating procedure *assumes* the causal model:



Whether or not there really is a causal relationship between Y and X_1, X_2, ..., X_k depends on having a theory, consistent with the data, that links the variables. And in trying to assess the independent effect of one of the describing variables on Y by "holding constant" or "adjusting out" all the other describing variables, we must always keep in mind that the "holding constant" or "adjusting out" is done *statistically,* by the manipulation of the observed data. The variables are passively observed; we are not really intervening in the system and holding constant all variables except one. And so the causal structure of the multiple regression model is not strongly tested by the statistical control, adjustment, or holding constant of the variables. George Box soundly described the contrast between the statistical control of observed variables and the actual experimental control (and deliberate manipulation) of variables: "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)."[2]

Still, in many cases in political and policy analysis, the best we can do in trying to understand what is going on is to hold constant or control variables statistically rather than experimentally—for there is simply no other way to investigate many important questions.

Example 1: Midterm Congressional Elections—Presidential Popularity and Economic Conditions

In every midterm congressional election but one since the Civil War, the political party of the incumbent President has lost seats in the House of Representatives. This persistent outcome results from differences in turnout in midterm compared to on-year elections:

> Explanation of the Administration's loss at midterm must be sought not so much by examining the midterm election itself as by looking

[2] Quoted in John P. Gilbert and Frederick Mosteller, "The Urgent Need for Experimentation," in Frederick Mosteller and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity* (New York: Vintage, 1972), p. 372.

> at the preceding presidential election. The stimulation of the presidential campaign brings a relatively large turnout. It attracts to the polls persons of low political interest who in large degree support the winning presidential candidate and, incidentally, his party's congressional candidates. At the following midterm congressional election, turnout drops sharply. . . . Those who stay home include in special degree the in-and-out voters who had helped the President and his congressional ticket into office. As they remain on the sidelines at midterm, the President's allies in marginal districts may find themselves voted from office. The coattail vote of the preceding presidential year that edged these Representatives into office simply stays at home . . .[3]

Yet this view of midterm elections is incomplete—for it only explains why the President's party should almost always be operating in the loss column rather than accounting for the *amount* of votes lost by the President's party. In statistical parlance, what has been explained is the location of the mean rather than variability about the mean. In studying the variability about the mean, we seek to answer such questions as: Why do some Presidents lose fewer congressional seats at the midterm than other Presidents? What factors affect the magnitude of the loss of congressional seats by the President's party? In Chapter 3, we used a two-variable regression to begin to answer these questions; however, that model left some variability unexplained. A more complicated model, bringing in the effect of economic conditions on the election, appears useful.

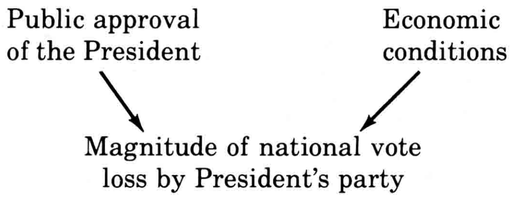
In order to explain the magnitude of the loss of votes and congressional seats by the President's party in midterm elections, we will estimate the following multiple regression model:

$$\begin{matrix}\text{Votes loss by} \\ \text{President's party}\end{matrix} = \beta_0 + \beta_1 \begin{bmatrix}\text{Presidential} \\ \text{popularity}\end{bmatrix} + \beta_2 \begin{bmatrix}\text{Economic} \\ \text{conditions}\end{bmatrix}$$

The idea is, then, that the lower the approval rating of the incumbent President and the less prosperous the economy, the greater the loss of support for the President's party in the midterm congressional elections. Thus the model assumes that voters, in midterm elections, reward or punish the political party of the President on the basis of their evaluation of (1) the performance of the President in general and (2) his management of the economy in particular.

[3] V. O. Key, *Politics, Parties, and Pressure Groups*, 5th ed. (New York: Thomas Y. Crowell, 1964), pp. 568–69.

The model is:



Three variables must be measured. With respect to economic conditions, recent studies of the relationship between aggregate economic conditions and the outcome of congressional elections show that interelection shifts of ordinary magnitude in unemployment have less impact on congressional elections than do shifts in real income.[4] Thus the most meaningful measure of economic conditions for our model appears to be the interelection change in real disposable income per capita. This measure probably may reflect the economic concerns of most voters, for it assesses the short-run shift in the average economic conditions prevailing at the individual level—a shift in conditions for which some voters might hold the incumbent administration responsible.

For this model, the public's evaluation of the President's general performance is measured by the standard Gallup Poll question: "Do you approve or disapprove of the way President______is handling his job as President?" Table 4-1 shows responses to the survey taken each September prior to the midterm election.

TABLE 4-1
The Data

Year	*Mean congressional vote for party of current President in last 8 elections*		*Nationwide congressional vote for party of current President*	*Standardized vote loss*	*Gallup Poll rating of President at time of election*	*Current yearly change in real disposable income per capita*
1946	Democratic	52.57%	45.27%	7.30%	32%	−$36
1950	Democratic	52.04%	50.04%	2.00%	43%	$99
1954	Republican	49.79%	47.46%	2.33%	65%	−$12
1958	Republican	49.83%	43.91%	5.92%	56%	−$13
1962	Democratic	51.63%	52.42%	−.79%	67%	$60
1966	Democratic	53.06%	51.33%	1.73%	48%	$96
1970	Republican	46.66%	45.68%	.98%	56%	$69

The most important variable to measure well is the magnitude of the vote loss by the President's party. The idea of "loss" implies the question "Relative to what?" The relevant comparison is between the normal, long-run congressional vote for the political party of the current President and the outcome of the midterm election at hand—that is, a standardized vote loss:

$$\left(\begin{array}{l}\text{standardized vote}\\ \text{loss by President's}\\ \text{party in the } i\text{th}\\ \text{midterm election}\end{array}\right) = \left(\begin{array}{l}\text{average vote for}\\ \text{party of current}\\ \text{President in the}\\ \text{last 8 elections}\end{array}\right) - \left(\begin{array}{l}\text{vote for}\\ \text{President's}\\ \text{party in the}\\ i\text{th election}\end{array}\right)$$

The loss is measured with respect to how well the party of the current

[4] Gerald H. Kramer, "Short-Term Fluctuations in U.S. Voting Behavior, 1896–1964," *American Political Science Review,* 65 (March 1971), 131–43; George J. Stigler, "General Economic Conditions and National Elections," *American Economic Review Papers and Proceedings,* 63 (May 1973), 160–67 and further discussion, 169–80.

President has normally tended to do, where the normal vote is computed by averaging that party's vote over the eight preceding congressional elections. This standardization is necessary because the Democrats have dominated postwar congressional elections; thus, if the unstandardized vote won by the President's party is used as the response (dependent) variable, the Republican presidents would appear to do poorly. For example, when the Republicans win 48 percent of the national congressional vote, it is, relatively, a substantial victory for that party and should be measured as such. The eight-election normalization takes this effect into account.

Table 4-1 shows the data matrix for the postwar midterm elections. We now consider the multiple regression fitting these data.

Table 4-2 shows the estimates of the model's coefficients. The results are statistically secure, since the coefficients are several times their standard errors. The fitted equation indicates:

1. A change in Presidential popularity of 10 percentage points in the Gallup Poll is associated with a national change of 1.3 percentage points in national midterm votes for congressional candidates of the President's party.
2. A change of $100 in real disposable personal income per capita in the year prior to the midterm election is associated with a national change of 3.5 percentage points in midterm votes for congressional candidates of the President's party.

The fitted equation explains statistically 89.1 percent of the variance in national midterm election outcomes; or, to put it another way, the correlation between the actual election results and those predicted by the model is .944. Since the fitted equation uses two meaningful explanatory variables, it seems reasonable to believe in this case

TABLE 4-2
Multiple Regression Fitting Standardized Vote Loss by President's Party in Midterm Elections

		Regression coefficient and (standard error)
β_1	Presidential approval rating (Gallup Poll, two months before election)	−.133 (.038)
β_2	Inter-election change in real disposable personal income per capita	−.035 (.015)

$\beta_0 = 11.083$, $R^2 = .891$.

that a successful statistical explanation is also a successful substantive explanation.

The multiple regression model is an equation, weighting the particular values (prevailing in a given election) of Presidential popularity and economic conditions. Thus the recipe for predicting the midterm outcome is to take .133 of the percent approving the President and .035 of the recent change in disposable personal income, subtract all this from β_0 (which is 11.083), and this gives the predicted shift in the midterm vote. Let us see how the equation worked for 1970. The equation, as shown in Table 4-2, fitted to the data is:

$$\begin{matrix}\text{standardized} \\ \text{vote loss}\end{matrix} = 11.083 - .133 \begin{pmatrix}\text{Percent approving} \\ \text{President}\end{pmatrix} - .035 \begin{pmatrix}\text{Change in} \\ \text{income}\end{pmatrix}$$

For 1970, the percent approving the President was 56 percent; the change in disposable personal income per capita was $69. Putting these particular values in the weighted combination of the regression yields:

$$\begin{aligned}\begin{matrix}\text{standardized} \\ \text{vote loss} \\ \text{predicted for 1970}\end{matrix} &= 11.083 - .133\ (56) - .035\ (69) \\ &= 11.083 - 7.448 - 2.415 \\ &= 1.2\end{aligned}$$

As Table 4-1 shows, the actual standardized vote loss for 1970 was 1.0, and thus the model fits the data rather well for 1970. As usual, the residual is the observed minus the predicted value; and thus the residual for 1970 from the fitted regression is -0.2.

As another check of the adequacy of the model, its predictions of midterm outcomes were compared with those made by the Gallup Poll in the national survey conducted a week to ten days before each election. As Table 4-3 shows, the model outperforms, in six of seven elections, the pre-election predictions based on surveys directly asking voters how they intend to vote. All this, of course, is after the fact; it would be more useful to have a prediction in hand prior to the election to test the model.

An analysis based on so few data points ($N = 7$ elections) can be very sensitive to outlying values in the data. In order to test

TABLE 4-3
After-the-Fact Predictive Error of the Model

Year	*Actual vote for House candidates, President's party*	*Gallup Poll prediction*	*Model prediction*	*Gallup absolute error*	*Model absolute error*
1946	45.3	42	44.5	3.3	.8
1950	50.0	51	50.2	1.0	.2
1954	47.5	48.5	46.9	1.1	.6
1958	43.9	43	45.6	.9	1.7
1962	52.4	55.5	51.6	3.1	.8
1966	51.3	52.5	51.8	1.2	.5
1970	45.7	47	45.5	1.3	.2

Average absolute error, Gallup = 1.7 percentage points
Average absolute error, Model = 0.7 percentage points

the stability of the fitted equation, the multiple regression was recomputed after omitting one election at a time. Table 4–4 shows the results; even when the regression is based on six elections, the regression coefficients remain fairly stable. The greatest shift occurs when the outlying values for 1946 (very low Presidential approval ratings and a decline in real disposable income per capita in the early postwar period) are dropped from the estimation.

Does the strong aggregate responsiveness of midterm outcomes to economic conditions and evaluations of the President's performance indicate anything about the rationality of the electorate—or about, at least, that half of the eligible citizenry turns out in off-year elections?[5] Such is the usual line of argument, for how else does one explain the choice of variables in the model and the ultimate results? It is important to realize, however, that all we are seeing in these data (and in the many similar studies) is the totally *aggregated* evidence that speaks only most indirectly to what must be the central political questions concerning the rationality of *individual* voters:

1. Do some voters make more rational calculations than others? Which voters? How many?
2. What are the components of these calculations?
3. What kinds of decision rules do individual voters use? Which voters use what decision rules?
4. What conditions encourage voter rationality?
5. How may these conditions be nurtured?

[5] Angus Campbell, "Voters and Elections: Past and Present," *Journal of Politics*, 26 (November 1964), 745–57.

TABLE 4-4
Re-estimating the Regression Coefficients When the Data Points are Omitted One at a Time

Year omitted	*Constant term*	*Presidential popularity*	*Change in economic conditions*	R^2
1946	17.62	−.23	−.052	.94
1950	10.93	−.13	−.036	.89
1954	10.57	−.12	−.038	.90
1958	11.10	−.15	−.028	.99
1962	10.11	−.11	−.034	.88
1966	10.87	−.13	−.037	.89
1970	11.06	−.13	−.035	.88

Thus, although the results are impressive in terms of the large R^2, there are still substantial inferential problems in trying to interpret the meaning of the model—since the data do not speak directly to the explanatory mechanism postulated to explain the findings.

Let us consider the steps in the construction of this regression in order to look at some of the broader issues in constructing explanatory models. The steps were these:

1. A model, based on prior research and some general ideas, was specified. The model included two basic variables, presidential popularity and economic conditions. There were also several other variables that were candidates for inclusion in the model: whether the nation was involved in a war at the time of the election, the magnitude of the victory of the President's party in the preceding election, and a few others.
2. Each variable in the model was operationalized; that is, a numerical measure for the concept was found. The construction of appropriate measures required some further thought, especially with respect to the response variable, the standardized vote.
3. Several economic variables were included in the initial analysis—changes in unemployment, inflation, GNP per capita, and real disposable personal income per capita. From the beginning, the change in real disposable personal income per capita made the most substantive sense, and it turned out that it led to the most successful explanatory model in terms of variance explained. A variety of different regressions were computed.

There is, then, an interplay between explanatory ideas and the examination of the data. Some variables were tried out on the basis of a vague idea and were then discarded when they yielded no explanatory return. For example, some regressions included a variable indicating whether the nation was involved in a war (Korea or Vietnam) during the midterm congressional election—on the hypothesis that there might be a "rally round the flag" effect helping the President. Such appeared to be the case—and the sign of the regression coefficient was in the expected direction—but the results just did not seem solid enough to warrant inclusion in the final model, especially since there are only seven data points and also since only two explanatory variables do so well.

Now, looking at several different multiple regressions and sorting around through different variables may not fit some abstract models of scientific research procedure—but it is normally done in constructing explanatory models, and it is precisely this sorting through of various notions that is the heart of data analysis. The final model reported here has gained inferential strength as a consequence of this directed search through a variety of ideas because the model has been tested against many other alternative possibilities and has survived. The strength of such an interplay between theory and data has been strongly put by Jacob Viner:

> If there is agreement that relevance is of supreme importance for economic theory, it leads to certain rules of guidance as to the procedure we should follow in constructing our theoretical models. It is common practice to start with the simplest and the most rigorous model, and to leave it to a later stage, or to others, to introduce into the model additional variables or other complicating elements. I venture to suggest that the most useful type of "first approximation" would often be of a radically different character. It would consist of a listing of all the variables known or believed to be or suspected of being of substantial significance, and corresponding listing of types and directions of interrelationship between these variables. A second stage of analysis would consist of a combing out on the basis of such empirical evidence as can be accumulated of the probably least significant variables and interrelationships between variables. Instead of beginning with rigor and elegance, only from this second stage on would these become legitimate goals, and even then for a time they should be distant goals, to be given high value only after it is clear that they can be reached without substantial loss of relevance.
>
> Such procedure, it would seem to me, would have some distinct advantages as compared to the more usual procedure on the part of theorists of starting—and often ending—with models that gain their rigor at the cost of unrealistic simplification. In the first case, important variables would be less likely to be omitted from consideration because of oversight, traditional practice, difficulty of manipulation, or unsuitability for specific types of analytical manipulation to which the researcher has an irrational attachment. Secondly, there would be at least partial awareness of what variables had been omitted from the final analysis, and therefore greater likelihood than at present that the conclusions will be offered with the qualifications

and the caution that such omission makes appropriate. Third, if the presentation of the final results includes a statement with respect to the omitted variables and the reasons for their omission, the reader of such presentation is in better position to appraise the significance of the findings and is afforded some measure of guidance as to the further information and the new or improved techniques of analysis that would be most helpful.

The final outcome of such a change in analytical procedure might well be a definite loss in rigor and elegance at least for a long time, on the one hand, but a definite gain in scope for the useful exploitation of new information and of wisdom and insight on the other hand. Such a result, I hope and believe, would in most cases constitute a new gain in relevance for understanding of reality and for the promotion of economic welfare by means of economic theorizing.[6]

Example 2: Equality of Educational Opportunity and Multicollinearity

> Modern statisticians are familiar with the notions that any finite body of data contains only a limited amount of information, on any point under examination: that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.[7]
>
> —R. A. Fisher

If two or more describing variables in an analysis are highly intercorrelated, it will be difficult and perhaps impossible to assess accurately their independent impacts on the response variable. As the association between two or more describing variables grows stronger, it becomes more and more difficult to tell one variable from the other. This problem, called "multicollinearity" in the statistical jargon, sometimes causes difficulties in the analysis of nonexperimental data.

For example, if, in Chapter 1, density and inspections (the two describing variables for the response variable of traffic fatalities) were highly associated—say, all states above a certain density had inspections and all below did not—then it would be very difficult to discover if inspections made a difference because the effect of inspections would be confounded with the effect of density. The

[6] Jacob Viner, "International Trade Theory and Its Present Day Relevance," in Brookings Lectures, 1954, *Economics and the Public Policy.* © 1955 by the Brookings Institution, Washington, D.C., pp. 128–30.

[7] R. A. Fisher, *The Design of Experiments,* 8th ed. (London: Oliver and Boyd, 1966), p. 40.

scatterplot, in this hypothetical example, would resemble Figure 4-1. In such a case there is insufficient independent variation in the two describing variables; in particular, there is a shortage of thickly populated states without inspections and thinly populated states with inspections. Without such conditions prevailing in at least a few states, the independent effect of inspections and the independent effect of density on the death rate could not be assessed.

Sometimes clusters of variables tend to vary together in the normal course of events, thereby rendering it difficult to discover the magnitude of the independent effects of the different variables in the cluster. And yet it may be most desirable, from a practical as well as scientific point of view, to disentangle correlated describing variables in order to discover more effective policies to improve conditions. Many economic indicators tend to move together in response to underlying economic and political events. Or consider a research design seeking to assess the effects of air pollution on the health of a city's residents. Such a study might be based on three areas in a city—one with badly polluted air, one with moderate pollution, and (if it could be found) one with relatively clean air. But chances are that the poor are more likely to find housing only in those unpleasant parts of

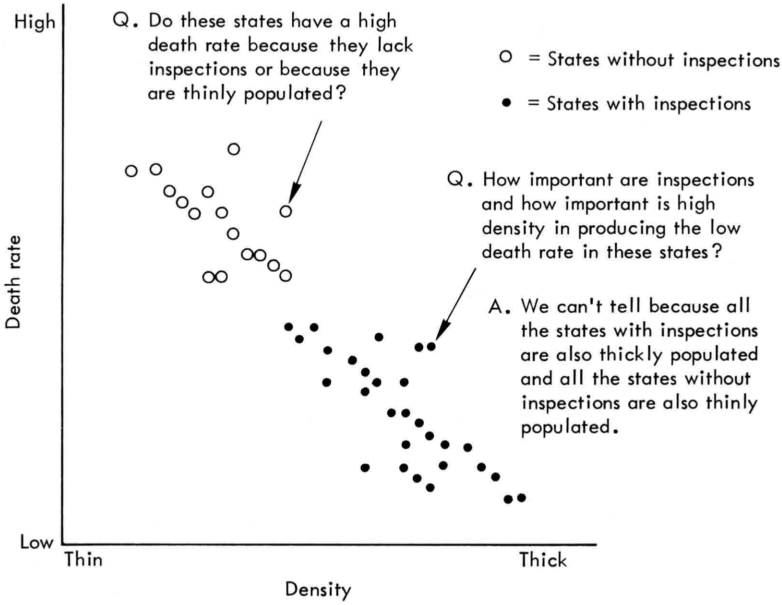


FIGURE 4-1 Hypothetical data showing collinearity between density and inspections

the city near factories and highways producing very polluted air; the moderately polluted area is more likely to be the home of those with moderate incomes; and the wealthy will be concentrated in areas relatively free of pollution. In such a situation, then, the effects of air pollution on health are confounded with the effects of income and housing on health.

The problem of multicollinearity involves a lack of data, a lack of information. In the first example, there were no *thinly* populated states *with* inspections (and vice versa); in the study of the health effects of air pollution, we lacked information about rich neighborhoods with polluted air and poor areas with fresh air.

Recognition of multicollinearity as a lack of information has two important consequences:

1. In order to alleviate the problem, it is necessary to collect more data—especially on the rarer combinations of the describing variables.
2. No statistical technique can go very far to remedy the problem because the fault lies basically with the data rather than the method of analysis. Multicollinearity weakens inferences based on *any* statistical method—regression, path analysis, causal modeling, or cross-tabulations (where the difficulty shows up as a lack of deviant cases and as near-empty cells).

Figure 4-2 shows how, when two describing variables are highly intercorrelated, a control for one variable reduces the range of variation in the other.

Since multicollinearity affects our ability to assess the independent influence of each describing variable, its consequences in the multiple regression model include increased errors in the estimate of the regression coefficients. The variance of the estimate of the regression coefficient, $\hat{\beta}_i$, is given by:

$$\text{variance of } \hat{\beta}_i = \frac{1}{N - n - 1} \frac{S_Y^2}{S_{X_i}^2} \frac{1 - R_Y^2}{1 - R_{X_i}^2},$$

where N = number of observations,

n = number of describing variables,

S_Y^2 = variance of Y,

$S_{X_i}^2$ = variance of X_i,

R_Y^2 = squared multiple correlation for the regression $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$,

$R_{X_i}^2$ = squared multiple correlation for the regression $X_i = \beta_0' + \beta_i' X_i + \ldots + \beta_{i-1}' X_{i-1} + \beta_{i+1}' X_{i+1} + \ldots + \beta_n' X_n$

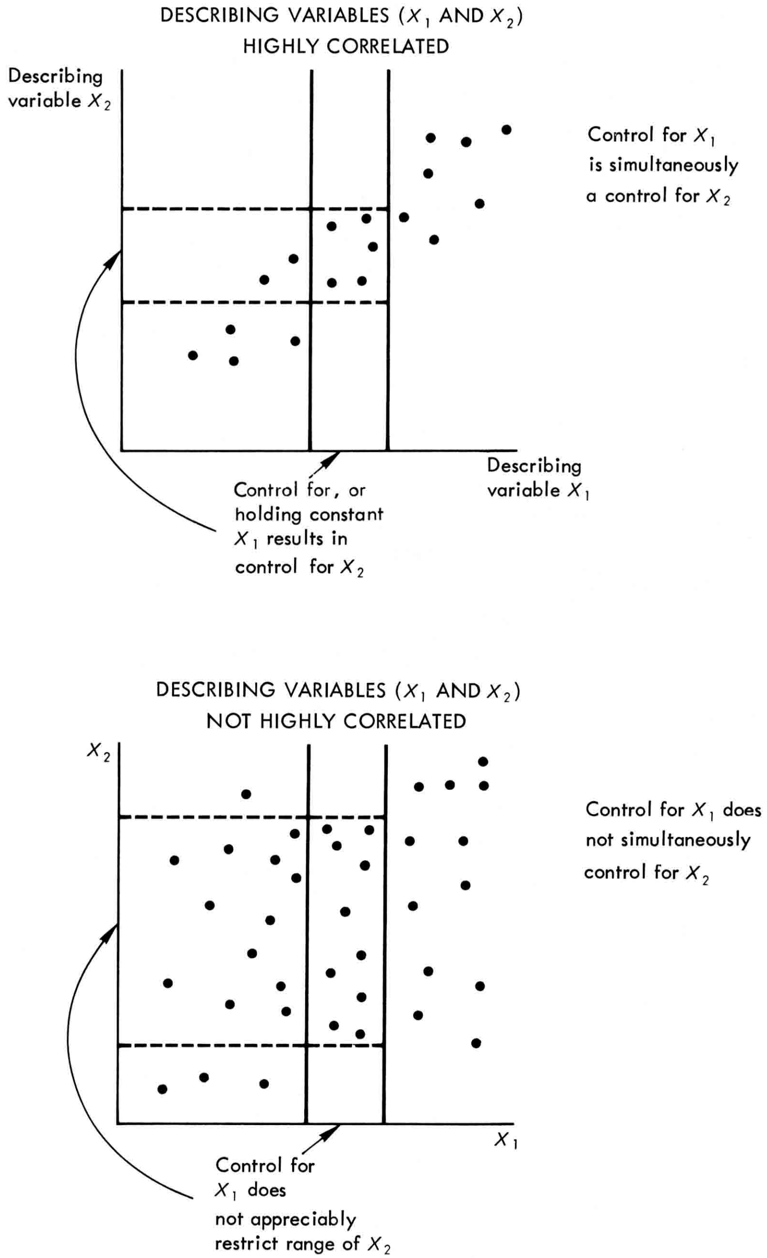

DESCRIBING VARIABLES (X_1 AND X_2) HIGHLY CORRELATED
Describing variable X_2
Control for X_1 is simultaneously a control for X_2
Describing variable X_1
Control for, or holding constant X_1 results in control for X_2
DESCRIBING VARIABLES (X_1 AND X_2) NOT HIGHLY CORRELATED
X_2
Control for X_1 does not simultaneously control for X_2
X_1
Control for X_1 does not appreciably restrict range of X_2

FIGURE 4-2 Effect of controlling for a variable when describing variables are strongly correlated

—that is, the regression of the ith *describing* variable on all the other *describing* variables.

This equation repays study. The key element is:

$$\text{the variance of } \hat{\beta}_i \text{ is proportional to } \frac{1}{1 - R^2_{X_i}}.$$

Now $R^2_{X_i}$ is the R^2 for the regression of the ith describing variable on all the other remaining describing variables—that is, $R^2_{X_i}$ assesses how well the describing variable X_i is explained by the *other* describing variables. So if X_i is strongly entangled with one or more of the other describing variables, $R^2_{X_i}$ will be large, close to 1.0. Consequently $1/(1 - R^2_{X_i})$ and the variance of $\hat{\beta}_i$ will grow larger as $R^2_{X_i}$ approaches 1.0. And so the estimate of $\hat{\beta}_i$ grows more insecure as $R^2_{X_i}$ approaches closer to 1.0.

Although multicollinearity is sometimes viewed as a problem of the intercorrelation of two describing variables, it can be seen here that the variances of the estimated regression coefficients will be big *whenever* $R^2_{X_i}$ is large—which can result from a high intercorrelation between two of the describing variables *or* from a combination of three or more of the describing variables accurately predicting another describing variable. Note the variance of $\hat{\beta}_i$ is infinite when $R^2_{X_i}$ is unity (that is, when a describing variable X_i is perfectly predicted by one or more of the other describing variables). In this case, of course, it is literally impossible to tell X_i from another describing variable or combination of other describing variables. The equation for the variance of $\hat{\beta}_i$ also shows that the variance of the estimates of the regression coefficients will decrease as additional data are collected (as N grows larger).

In summary, the symptoms of multicollinearity in regression analysis include:

1. high intercorrelations between the describing variables,
2. large variances in the estimates of the regression coefficients,
3. large $R^2_{X_i}$,
4. large R^2_Y coupled with statistically nonsignificant regression coefficients,
5. large changes in the values of estimated regression coefficients when new variables are added to the regression, and
6. inability of computer program to compute regression coefficients (which occurs only in very severe cases of multicollinearity—in most cases the estimation procedures produce numbers as usual).

Cures for multicollinearity can sometimes be found in the following list:

1. Collect additional data, concentrating on gathering information that will alleviate the difficulty. In some research contexts, this may involve seeking information on deviant cases or special combinations of the describing variables. Johnston cites an econometric example: "Early demand studies, for example, which were based on time-series data, often ran into difficulties because of the correlation between the explanatory variables, income and prices, plus the often inadequate variation in the income series. The use of cross-section budget data, however, gives a wide range of income variation, thus permitting a fairly precise determination of the income coefficient, which can then be employed in the time-series analysis."[8]
2. Give up on nonexperimental data and consider research designs in which the key variables can be systematically varied or at least randomized out. Do experiments.
3. Remove some of the variables from the regression that are causing the trouble. For example, if two of the describing variables are highly correlated, compute regressions with only one of the variables present at a time. Or combine the variables into a summary measure (less often an approved strategy). These steps should be taken only if they make good substantive sense.

Although the use of additional information and special statistical techniques may at times alleviate the problem, it often happens in social research based on "experiments" performed by nature that it will be difficult to obtain the independent variation necessary to assess the independent effects of the describing variables. Thus some theories that assert the importance of one variable over another, while theoretically testable, are actually incapable of being tested in the face of multicollinearity.

Finally, it is important to be clear about the signs of multicollinearity and just when it is a genuine threat to the validity of a study. It is not a sound or a fair statistical criticism to cry "multicollinearity" to discredit every analysis involving three or more variables.

A multicollinearity problem arose in the report on *Equality of Educational Opportunity* by James Coleman and others.[9] The model used seeks to explain student achievement in school (as measured

[8] J. Johnston, *Econometric Methods,* 2d ed. (New York: McGraw-Hill, 1972), p. 164.

[9] James Coleman, Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic Weinfield, and Robert L. York, *Equality of Educational Opportunity* (Washington, D.C.: Office of Education, 1966). Parts of the report are reprinted in E. R. Tufte, ed., *The Quantitative Analysis of Social Problems* (Reading,

by test scores) with two clusters of variables, measures of family background aiding children in their schoolwork (such as books in the home) and measures of school resources such as the teacher-student ratio and the number of books per student in the library. In compressed form, the model is:

$$\begin{tikzcd}
\text{family background} \arrow[dr] & & \text{school resources} \arrow[dl] \\
& \text{achievement} &
\end{tikzcd}$$

The analysis proceeded by first regressing achievement against the family background variables, which yielded an R^2. Then a new regression was computed that included the school resources variables as well as family background, yielding a coefficient of R'^2. The difference,

$$R'^2 - R^2,$$

was taken as measure of the effect of school resources on educational achievement. Although using the increase in the percent of variance explained as a measure of school resource effects on education did not ultimately compromise the main findings of the study, the method would tend to underestimate school effects somewhat and received criticism. Bowles and Levin wrote:

> The most severe deficiency of the regression analysis is produced by the addition to the proportion of variance in achievement scores explained (addition to R^2) by each variable entered in the relationship as a measure of the *unique* importance of that variable. For example, assume that we seek to estimate the relationship between achievement level, Q, and two explanatory variables, X_1 and X_2. The approach adopted in the Report is to first determine the amount of variance in Q that can be statistically explained by one variable, say X_1, and then to determine the amount of variation in Q that can be explained by both X_1 and X_2. The increment in explained variance (i.e., the change in the coefficient of determination, R^2) associated with the addition of X_2 to the explanatory equation is the measure used in the Report for the unique effect of that variable on Q. Thus, if X_1 explained 30 percent of the variance in Q and X_1 and X_2 together explained 40 percent, the difference, or 10 percent, is the measure of the unique effect of X_2.

Mass.: Addison-Wesley, 1970), 285–351. See also Frederick Mosteller and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity* (New York: Random House, 1972).

If X_1 and X_2 are completely independent of each other (orthogonal), the use of addition to the proportion of variance explained as a measure of the unique explanatory value of X_1 and X_2 is not objectionable. X_1 will yield the same increment to explained variance whether it is entered into the relationship first or second, and vice versa. But when the explanatory variables X_1 and X_2 are highly correlated with each other, as are the background characteristics of students and the characteristics of the schools that they attend, the addition to the proportion of variance in achievement that each will explain is dependent on the order in which each is entered into the regression equation. By being related to each other, X_1 and X_2 share a certain amount of explanatory power which is common to both of them. The shared portion of variance in achievement which could be accounted for by either X_1 or X_2 will always be attributed to that variable which is entered into the regression first. Accordingly, the explanatory value of the first variable will be overstated and that of the second variable understated.

The relevance of this problem to the analysis in the Report is readily apparent. The family background characteristics of a set of students determine not only the advantages with which they come to school; they also are associated closely with the amount and quality of resources which are invested in the schools. As a result, higher status children have two distinct advantages over lower status ones: First, the combination of material advantages and strong educational interests provided by their parents stimulate high achievement and education motivation; and second, their parents' relatively high incomes and interest in education leads to stronger financial support for and greater participation in the schools that their children attend. This reinforcing effect of family background on student achievement, both directly through the child and indirectly through the school, leads to a high statistical correlation between family background and school resources.

The two sets of explanatory variables are so highly correlated that after including one set in a regression on achievement, the addition to the fraction of total variance explained (R^2) by the second set will seriously understate the strength of the relationship between the second variables and achievement. Yet the survey made the arbitrary choice of first "controlling" for student background and then introducing school resources into the analysis. Because the student background variables—even though crudely measured—served to some extent as statistical proxies for school resources, the later introduction of the school resource variables themselves had a small explanatory effect. The explanatory power shared jointly by school resources and social background was thus associated entirely with social background. Accordingly, the importance of background factors in accounting for differences in achievement is systematically inflated and the role of school resources is consistently underestimated.[10]

[10] Samuel Bowles and Harry Levin, "The Determinants of Scholastic Achievement—An Appraisal of Some Recent Evidence," *Journal of Human Resources,* 3 (© 1968 by the Regents of the University of Wisconsin), pp. 14–16.

Example 3: A Five-Variable Regression—The Size of Democratic Parliaments

Here we examine a five-variable multiple regression that illustrates the following statistical points:

- taking logarithms to test a "cube root law" by converting the law into a linear model,
- interpreting a regression coefficient as an elasticity,
- using R_i^2 as a check for multicollinearity,
- using a "dummy variable" so that a dichotomous, categoric variable can be included in a regression,
- interpreting R^2 as the square of the correlation between the observed and predicted values of the response variable.

The multiple regression reported here evaluates some of factors determining parliamentary size—the number of representatives in the lower house—in twenty-nine relatively democratic countries of the world. Parliaments differ greatly in size; Liechtenstein's Diet has 15 deputies, the Italian Chamber of Deputies has 630 members, West Germany's Bundestag 496, the French National Assembly 481, and Sweden's new unicameral parliament 350. Some large countries have relatively small parliments: India, with a population 2½ times that of the United States, has 500 deputies sitting in its House of the People with each deputy representing, on average, over one million citizens. At the other extreme, the 19,000 residents of San Marino have a 60-member Great and General Council—resulting in one representative for every 320 citizens.

The number of representatives elected to parliament determines, in part, the extent to which local interests are represented at the national level; larger parliaments, other things (especially population) being equal, permit a more precise representation. However, in larger parliaments each member not only has an arithmetically smaller voice, but also larger parliaments typically have greater centralization of leadership and more rules limiting the conduct of their members both in debate and in the diversity of their concerns. These two conflicting factors—the representation of citizens and the manageability of the chamber—must be resolved by the framers of new constitutions. Many constitutions of the eighteenth and nineteenth centuries specified a particular ratio of citizens to representatives, and parliamentary size grew right along with the population.

As a consequence of this mixture of factors, parliamentary size is closely linked to population: the more populous countries have larger parliaments. Dodd proposed a "cube-root law":

$$\text{number of members of parliament} = (\text{population})^{1/3}.$$

Dodd correlated these variables and found that the cube root of population explained 67 percent of the variation in parliamentary size for 55 nations in 1950.

Dodd's model may be written by taking logarithms

$$\text{log members} = 1/3\ (\text{log population}).$$

Thus in the regression of members (log) against population (log),

$$\text{log members} = \beta_1\ (\text{log population}) + \beta_0,$$

the cube-root law predicts that $\beta_1 = 1/3$ and $\beta_0 = 0$. Confirming these predictions provides a better test of the law than merely correlating the cube root with the size of parliament. That correlation does not test the specific hypothesis of the law; it merely supports the general proposition that there is a relationship between the two variables. Taking the logarithms of both variables also, as we saw in Chapter 3, yields a useful interpretation of the slope of the fitted line. The estimate of the slope, β_1, measures the *percentage change* in the size of parliament associated with a *change of one percent* in the size of the population: β_1 is the least-squares estimate of the elasticity of parliamentary size with respect to population.

Here Dodd's law is tested with data from twenty-nine of the more democratic countries in the world in 1970. The multiple regression in Table 4-5 shows that the population elasticity of parliamentary size is .44, indicating that if a county was one percent above the average in population size, it was typically .44 percent above the average in parliamentary size. The standard error of the estimated elasticity is quite small, .022; thus the estimate of the elasticity itself, .44, quite surely differs from the prediction of the cube-root law.

There are three other describing variables in the regression shown in Table 4-5:

Population growth rate Although many of the democracies have relatively low growth rates, there is still sufficient variability to explain differences in parliamentary size. Countries that are growing rapidly in population size tend to have smaller parliaments, other things being equal, than countries growing more slowly. When all the other describing variables in the equation are fixed at their means, a change of one percentage point in growth rate from an annual rate of one percent to two percent across countries is associated with a decrease in the size of parliament from 196 seats to 144 seats.

TABLE 4-5
Parliamentary Size (logarithm) for Twenty-nine Democracies

	Regression coefficient	*Standard error*	R_i^2
Population (log)	.440	.022	.14
Annual population growth rate	−.135	.020	.13
Number of political parties	.051	.013	.26
Bicameral—unicameral	.066	.040	.20
$R^2 = .952$			

All coefficients are statistically significant at the .001 level, with the exception of the variable bicameral—unicameral. That coefficient is significantly different from zero at the .06 level.

Number of parties in the party system The greater the number of parties in the present-day party system, the larger the parliament. Other things being equal, two-party systems have parliaments averaging about 137 seats; multiparty systems, 195 seats. A larger party system may reflect somewhat greater underlying diversity in the society, and the constitutional framers may then create a larger than normal parliament in an effort to represent that diversity. Perhaps a more plausible explanation is that in a multiparty system, many parties will participate in the bargaining over parliamentary size and the smaller parties will work hard for a large-sized parliament, so that at least some of their party officials will be able to hold parliamentary seats. Parliaments sufficiently large to include the leading officials of each party may be quite inflated in size, particularly if the votes of the minor parties are scattered. If such a process operated for a number of years as the distribution of seats shifted from party to party, then the incumbent parliamentarians might well favor increases in the size of parliament so that they or their colleagues would stay in office even with some shifts in the share of votes received by each party.

Bicameral—unicameral parliaments Unicameral parliaments are typically somewhat larger than the lower chambers of bicameral parliaments. Some unicameral systems have come about from a merger of two chambers; here the interests of incumbent parliamentarians are obvious. Other things being equal, the unicameral parliaments average 189 seats in the fitted model; the lower chamber of bicameral parliaments, 163 seats.

The numerical coding for this variable was:

bicameral = 0,
unicameral = 1.

Such a dichotomous categoric variable is called a "dummy variable," and such variables are used to include categoric variables in multiple regression models. The following are examples of dummy variables:

REGION
0 = North
1 = South

CHANGE IN A TIME SERIES
0 = before tax cut,
1 = after tax cut

SEX
0 = male,
1 = female

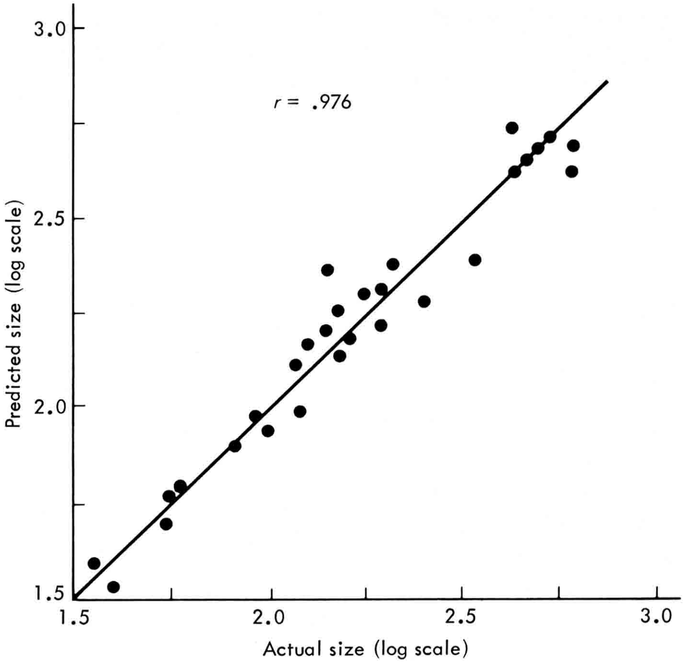


FIGURE 4-3 Actual and predicted parlimentary size, twenty-nine democracies

Table 4-5 shows the value of R_i^2, the value of R^2 resulting from the regression of the ith describing variable on all the other describing variables. The values are quitė small, indicating that multicollinearity is not a problem here.

Figure 4-3 shows the relationship between the observed and predicted values of the response variable, the logarithm of parliamentary size. The correlation, $r_{Y\hat{Y}}$, between the observed and predicted values is .976. That value squared is mathematically equivalent to R^2, the proportion of variance in the logarithm of parliamentary size explained by the regression:

$$r_{Y\hat{Y}}^2 = (.976)^2 = .952 = R^2.$$

How was the regression reported in Table 4-5 chosen? At the start of the analysis, six describing variables were considered as possible candidates for inclusion in the final model. In addition to the four variables already discussed, two others were considered good candidates: whether or not the country was in Europe and the institutional age of the currently established parliament. It appeared that European countries, for one reason or another, had large parliaments. The length of time the parliament had been established under the current constitution was included as a possible describing variable on the speculation that older parliaments might be larger. Table 4-6 shows twelve different multiple regressions using various combinations of the six candidate describing variables. Let us look through these twelve different regressions to see the search for the model previously reported in Table 4-5. It will be clear that several different models could have been the model of choice.

TABLE 4-6
Twelve Regressions Explaining Parliamentary Size (Log)

	Regression number											
Describing variables	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Population size (log)	.41	.40	.42	.41	.38	.38	.40	.39	.40	.41	.43	.44
Population growth rate		−.16		−.10		−.13	−.13	−.06		−.07	−.13	−.14
Bicameral—unicameral							.12		.11	.12		.07
Number political parties											.06	.05
European—not European			.26	.13	.20			.13	.20	.13		
Age of current parliament					.17	.13	.11	.12	.18	.13		
R^2	.760	.900	.891	.912	.916	.908	.928	.923	.934	.941	.946	.952
Number of describing variables	1	2	2	3	3	3	4	4	4	5	3	4

The numbers shown in the table are regression coefficients for each regression. Each of the twelve columns shows a different regression.

Regression 1 is simply the two-variable regression of parliamentary size (log) against population size (log). The regression coefficient reported in Table 4-6 indicates that a change of one percent in population was associated with a change of .41 percent in parliamentary size; 76 percent of the variance was statistically explained. Regressions 2 and 3, both with two describing variables, send the R^2 up to about 90 percent. Either the population growth rate or the country's geographic location in or out of Europe adds an additional 14 percent to the variance explained in the first regression. This suggests that we can go much farther with a model that includes both the location and the growth rate along with population size. This is regression 4; and it doesn't work. Little additional variance is picked up—and also there is a multicollinearity problem. Note how the regression coefficients on growth and European location have shifted from their previous values in regressions 2 and 3, respectively.

This is a sign of multicollinearity, confirmed by the correlation of −.77 (European countries have low population growth rates) between the two variables.

Regressions 5 through 9 try out various combinations of the describing variables. These trials verify the multicollinearity problem with respect to the European location variable and raise some doubts about the effectiveness of the age variable. Throwing in every variable examined so far gives regression 10, which picks up 94.1 percent of the statistical variation in parliamentary size (log) but with some problems. The European location variable is quite bothersome by now, in part because of multicollinearity but also—and more importantly—what does it mean, anyway? It is vague; such a regional variable doesn't tell us much substantively. What is it, *specifically,* about location in Europe that makes for big parliaments? So, regression 10 is about the best that can be done with the current candidate variables.

The last two multiple regressions try out a new candidate variable, the number of political parties in a country. Regression 11 reports a simple model with only three describing variables that outperforms—at least in terms of R^2—all the previous models, including those that contain more variables. It is a parsimonious model and a relatively successful one in terms of R^2. Regression 12 adds one more variable—the dummy variable on whether the parliament is unicameral or bicameral—to take the variance explained up to 95.2 percent.

What we have seen here is an empirical search through a variety of theoretically plausible models. The search started with some candidate variables, which were suggested by our political and historical understanding of what factors might affect this particular characteristic—size—of a political institution. The search was conducted with a variety of criteria for evaluating the different models that turned up: certain substantive criteria (for example, in part, the grounds for rejection of the European location variable) and certain statistical criteria (the statistical significance of individual regression coefficients, the value of R^2, and multicollinearity). Now these criteria are not "merely" statistical matters, for the statistical criteria used in the choice of the models inform us about the quantitative *quality* of the model under examination. Or, more precisely, the statistical criteria help evaluate the quantitative quality of different models within the theoretical and substantive context of the search for models. The context is vital; the best statistical techniques can't rescue theoretical models that are poor, unintelligent, or misguided.

Table 4-6 also shows one of the sad facts of building complex explanations of most political, economic, and social phenomena: often a variety of models will fit the same data relatively well. That is, the empirical evidence that is available does not always allow one to choose among different models that seek to explain the response variable. In this case, regressions 11 and 12 both do rather well; but even regressions 2 and 3 seem relatively acceptable. It is probably fair to say, however, that regressions 11 and 12 are pretty much the best among the lot. Both regressions are quite effective in predicting—and explaining—parliamentary size (log) as Table 4-5 indicated.

Table 4-6 does not, fortunately, show all possible combinations of describing variables. With six describing variables, there are a grand total of 63 different regressions involving combinations of one or more describing variables. In general, with K describing variables, there are $2^K - 1$ possible regressions. Some regression programs can, in fact, search through all possible combinations to find one or more "best" regressions. Although such searches may seem rather like brute-force empiricism (and they often are!), the criteria of choice for the best regression or regressions are intelligent and may provide a reasonable guide—when combined with substantive understanding—in searching for models. Some elegant computer programming has enabled one regression program to examine quickly every regression in cases with up to 12 describing variables—that is, 4,095 regressions.[11] The view is: If you're going to search for a model, why not search thoroughly?

Of course, we would trade all those searches in for one good idea. And that idea might come from looking at the data.

[11] Cuthbert Daniel and Fred Wood, *Fitting Equations to Data* (New York: Wiley, 1971).

Index