Making Social Sciences More Scientific

The Need for Predictive Models

Rein Taagepera

OXFORD
UNIVERSITY PRESS

4

Example of Model Building: Electoral Volatility

- The foremost mental roadblocks in predictive model building are refusal to simplify and reluctance to play with extreme cases and their means. These roadblocks have little to do with mathematical skills.
- "Ignorance-based" models focus on conceptual constraints and extract the most out of near-complete ignorance. They ask: What do we already know about the situation, even before collecting any data?
- Eliminate the "conceptually forbidden areas" where data points could not possibly occur.
- Locate the conceptual "anchor points" where the value of x imposes a unique value of y.
- Once this is done, few options may remain for how y can depend on x—unless you tell yourself "It can't be that simple."
- Dare to make outrageous simplifications for an initial coarse model, including as few variables as possible. Leave refinements for later second approximations.
- A low R^2 may still confirm a predictive model, and a high one may work to reject it.

This chapter develops a quantitatively predictive logical model for a specific issue—volatility of voters and its conceivable dependence on the number of parties that run. Why this particular topic? It so happens that here the model is mathematically very simple, at least in first approximation. Indeed, this is one of the relatively few cases where the model has the linear form so familiar to social scientists. Thus, the reader's attention can focus on model-building skills, without being distracted

by a possibly unfamiliar mathematical format. This is important for countering the impression that model-building skills are largely mathematical. A good grasp of high school and college mathematics helps of course, but many skills and mental roadblocks in model building are elsewhere.

A coarse approximate model is constructed first, expanding on an earlier shorter version (Taagepera 2007*b*). A second approximation follows, leading to a more refined model. Some broad contrasts emerge between predictive and descriptive approaches.

Constructing a Coarse "Ignorance-Based" Model

Volatility (V) stands for the percentage of voters who switch parties from one election to the next. When more parties run, voters have more choices for switching. Hence, if the number of parties (N) has any effect on volatility of voters at all, it should be in the upward direction. In mathematical terms, we would expect $dV/dN > 0$. This is a directionally predictive logical model.

A technical side issue is how to measure the number of parties when some are large and some are small. Here, the effective number of components is used: $N = 1/\Sigma(v_i^2)$, where v_i is the fractional vote share of the ith party (see Taagepera 2007*c*: 47–64). Since we compare two elections, N should be taken as the average N at these two elections, assuming that these are not excessively different.

Another side issue concerns the occasional voter. Voters may switch parties, but they also may switch to not voting at all. For simplicity, we first omit the "party of nonvoters." We can make it more complex later on.

The first mental roadblock in model building may set in at this point: *refusal to simplify*. Ah, the reader may say, you are naive or cheating. You ignore the hard reality that there are always people who sometimes vote and sometimes do not. In the words of a critical journal referee regarding a different topic: "I am skeptical that there is much value of operating at such a high level of generality. Huge amounts of real-world variation are consigned to nowhere." Actually, model building consigns them to a much better place, namely the next-level analysis. Making things complex is easy; the challenge is to simplify, to ferret out the essential. This is what Occam's razor is about. Galileo's study of falling

bodies would have gotten nowhere if he had worried about feathers right from the beginning. He did not ignore feathers but consciously put them aside for a while. Note that reluctance to simplify has nothing to do with mathematical skills.

So we have the directional model $dV/dN > 0$. What should we do next? This may look self-evident to many social scientists. Collect data, run a linear regression $V = a + bN$, and see whether the slope $dV/dN = b$ is positive. If it is, the directional model is confirmed. Report the numerical value of slope b, its level of significance, correlation coefficient R^2, and possibly also intercept (a). Case closed. But hold it.

Are the resulting numerical values of a and b in a reasonable range, or are they surprisingly high or low? Such questions are rarely asked in today's social sciences, where the attitude tends to be that what is, is. Is not science about finding out what the world *is* like, leaving what it *should* be to religion? Right? Wrong.

Recall that science is very much about what *should* occur when some inevitable or plausible assumptions are made. Science is about such logical consequences. Of course, the expected outcomes must face a reality check. This is when data collection and statistical analysis enter, at a later stage. Let us first ask: *What do we already know about volatility and parties, even before collecting any formal data?*

The first response might be that, without data, we know nothing. But this is not so. We often take some of our knowledge so much for granted that we do not even realize how much we know. Teasing the most out of what we know can lead to a quantitative model that is based on near-complete ignorance, yet is of considerable predictive value. For short, I have called such models "*ignorance-based models*" (Taagepera 1999*b*).

The first step in constructing quantitatively predictive models often echoes the aforementioned advice by Sherlock Holmes: Eliminate the impossible, and a single possibility may remain—or at least the field is narrowed down appreciably. As a starter, *delineate the field in which data points could not possibly lie.*

Volatility cannot be less than 0 or more than 100%. The number of parties (N) cannot be less than 1. These *conceptually forbidden areas* are shown in Figure 4.1, where volatility is graphed against the number of parties. All this may look so obvious as not be worth three sentences, but it has consequences that are not so obvious. At this point, we assume that at least one party obtains some votes in both elections. This restriction excludes from consideration the unlikely situation where a single party

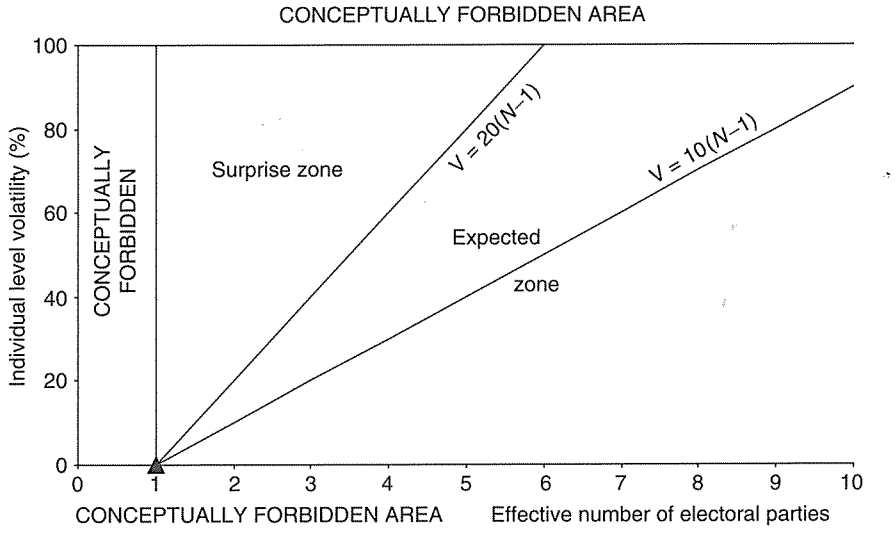

Figure 4.1. Individual-level volatility of votes vs. effective number of electoral parties—conceptually forbidden areas, anchor point, and expected zone

has all the votes in one election but loses them all to a brand new party in the next election.

Next, observe that there is a *conceptual extreme case.* Suppose only one party runs at the first election, and only the same party runs at the next one. This means that $N = 1$, and switching to another party is impossible. Hence, volatility must be 0. This point ($N = 1$, $V = 0$) is marked in Figure 4.1 with a black triangle. It is a conceptual *anchor point.* At $N = 1$, even a slight deviation of V away from 0 would violate logic.

A second mental roadblock may enter here: *reluctance to play with extreme cases* that rarely or never arise in practice. One may argue that talking about one-party elections is beside the point, because democratic countries always have more than one party running. Logical models, however, must not predict absurdities even under extreme conditions. Again, this mental roadblock has nothing to do with mathematical skills.

If V increases with N, our simplest tentative assumption could be linear increase: $V = a + bN$. But the anchor point adds a constraint. All acceptable lines must pass through the anchor point. For $N = 1$, we must have $V = 0$. Plugging these values into $V = a + bN$ yields $0 = a + b$, so that $a = -b$. This means that, among the infinite number of upward sloping straight lines,

only those will do where the intercept equals the negative of slope:

$$V = -b + bN = b(N - 1).$$

Without any input of data, the conceptual anchor point approach has already narrowed down the range of possibilities. Instead of two unknowns (a and b), we have only one. This is a tremendous advance in parsimony.

Now we move to shakier grounds. Consider a high effective number of parties, say $N = 6$, which is rarely reached. The reader may share a gut feeling that even with such a high number of parties to choose from, not all voters will switch parties. If so, then $V = 100\%$ at $N = 6$ would be a *highly surprising* outcome, although it is not conceptually impossible. The line $V = b(N - 1)$ that passes through this point is shown in Figure 4.1. It requires that $100 = b(6 - 1)$; hence $b = 100/(6 - 1) = 20$. Thus, the equation of this line is $V = 20(N - 1)$. Any data point located above this line would be highly surprising, although we cannot completely exclude the possibility, in contrast to the conceptually forbidden areas. Hence, this zone is marked as a *surprise zone* in Figure 4.1.

So $V = 20(N - 1)$ is about the highest value of V that would not utterly surprise us, at given N. Do we also have a lowest value? We do not. Even with a very high number of parties, it is still conceivable that party loyalty of voters could be complete. Thus, no limit higher than $V = 0$ can be proposed, meaning a horizontal line in Figure 4.1.

Without any real data input, we have now narrowed the reasonably *expected zone* of occurrence of data points to the cone between the lines $V = 20(N - 1)$ and $V = 0$. *In the absence of any other knowledge*, we have no reason to expect the actual line to be closer to either of these two extremes. Therefore, our best guess would be *the average of the likely extremes*, meaning $V = 10(N - 1)$. This line is also shown in Figure 4.1.

A third mental roadblock may enter here: *reluctance to take the mean of the extremes*. Suppose $N = 4$. Then $V = 20(N - 1)$ yields $V = 60$ while $V = 0$ yields 0. This is an awfully wide range. How could we assert that $V = 10(N - 1) = 30$ is more likely than any other number between 0 and 60? Is it not time to acknowledge that "We just do not know"? This would be a mistake. We still do know something. Our best "minimax bet" would be the mean of the extremes. The central values within the range would surprise us less than the extremes. True, the mean slope $b = 10$ actually stands for $b = 10 \pm 10$, which implies a huge range of possible error. Still, this entire range of b, from 0 to 20, means a considerable reduction in possibilities, compared to the directional model "I'll accept any positive slopes." Once more, reluctance to use the mean of the extremes has little to do with the mathematical skills needed for calculating the mean.

Testing the Coarse Model

Before resorting to any data, we have reached a predictive model based on near-complete ignorance:

$$V \approx 10(N - 1).$$

Compared to the directional model $dV/dN > 0$, this one predicts volatility much more specifically. If we ask "What volatility would you predict when $N = 4$?" the directional model $dV/dN > 0$ would answer: "Any positive V will do." Such a model is bound to be right, but its predictive value is nil because it covers too much ground. It is a qualitative model—or a semi-quantitative, if you really stretch it. In contrast, the model $V = 10(N - 1)$ would answer: "V will be *roughly around* 30." This prediction may be off by ± 30, upon testing with data, but it does offer a specific value. In this sense, this is a *quantitatively predictive model*. It is not "deterministic" in the sense of claiming that all data points will fall on the specific line. It rather expresses the expectation that about one-half the points will fall above and about one-half of the points will fall below the line $V = 10(N - 1)$.

In sum, this model makes two distinct predictions, one very precise and the other quite fuzzy:

1. If any straight line fits at all, the prediction $a = -b$ is absolute, due to respect for the conceptual anchor point. When regression with $V = a + bN$ is carried out, this model predicts that the two adjustable constants/coefficients will have exactly the same numerical value, with changed sign. If the values of b and $-a$ differed appreciably, it would sink the model, the more so if R^2 is high. On the other hand, if they pretty much agree, then we really should repeat the regression with $V = b(N - 1)$, which respects the anchor point and has only one adjustable coefficient, rather than with $V = a + bN$, where a and b can vary separately, ignoring the anchor point.
2. The prediction that slope b would be around 10, in contrast, is extremely fluid. It means: "If you force me to guess at a specific number, I would say 10." Even values appreciably different from 10

would not impair the basic model. They would just help to specify the numerical value of b, to be plugged into the model $V = b(N - 1)$.

Only at this stage would we need some data, so as to test the predictions $a = -b$ (exactly) and $b \approx 10$ (approximately). But we do not yet need to run a regression. All we need is the mean N and mean V for a set with many data points. Plug these means into $V = b(N - 1)$, and b is determined as b = mean V/(mean $N - 1$). We will see how close b is to 10. For the first and crucial prediction, we will have to regress with $V = a + bN$ and see whether $a/b = -1$, say within ± 0.1.

A uniform data-set is available from Oliver Heath (2005), for state-level elections in India (1998–99). Many parties competed in some of these states, while few did in some others. Mean $N = 3.65$ and mean $V = 31.6$ lead to $b = 11.9$, a result within 20% of our very coarse expectation of 10. This input of still limited information (mean N and V only) enables us to predict more precisely that

$$V = -11.9 + 11.9N.$$

The actual best fit reported by Heath (2005) is

$$V = -9.07 + 11.14N \quad [R^2 = 0.50].$$

At $N = 1$, it would yield $V = 2.07$. On a 0–100 scale, this is rather close to the conceptually required $V = 0$. The ratio $a/b = -9.07/11.4 = -0.80$ is within 20% of the conceptually required value, -1.00. The scatter of data around the best-fit line is appreciable, and the R^2 for $V = -11.9 + 11.9N$ is almost as high as it is for the best-fit line. Therefore, the difference may well be due to random fluctuation in data. The quantitatively predictive model $V = b(N - 1)$ is confirmed within $\pm 20\%$, with b specified as 11.9 rather than the initial estimate of 10. For most practical purposes, this would be close enough.

Refined Ignorance-Based Model

It may look quite impressive how close our predictive model came to the descriptive best-fit line. This model, however, involves a conceptual flaw, right from the beginning. For simplicity's sake, we assumed a linear increase, which led to $V = b(N - 1)$. The attentive reader may have noticed that any such upward line would project to a volatility of more than 100% for a sufficiently high number of parties. When $V = -11.9 + 11.9N$, it would surpass 100% whenever $N > 9.34$. True, hardly any party systems

have such a high effective number of parties, but remember: *"Predictive models must not violate logic even under extreme conditions."*

So we will have to refine the model. We must avoid predictions that extend into the forbidden area $V > 100$. This refinement requires some facility with exponential equations. A quick overview of the latter is given in the Appendix to Chapter 8. Some readers may wish to bypass the next section and simply accept its results.

This is the point where elementary school mathematics no longer suffices. The wonder is how far we have proceeded with little beyond arithmetic. Possible mental reservations against this approach have little to do with mathematical skills.

As an output variable approaches a conceptual ceiling, further increase in the input variable that drives it finds it ever harder, so to say, to achieve any further increase. The simplest way to express this extremely general phenomenon mathematically is $dy/dx = k(C - y)$, where C is the ceiling value for y, and k is an adjustable "rate constant" (see Chapter 8 for more details). This "differential equation" says that further increase in y is proportional to the remaining distance between y and the ceiling. Integration leads to a nonlinear equation: $y = C[1 - A\,e^{-kx}]$. It is an exponential equation where A is a constant that depends on the initial conditions. This equation applies, among others, to the amount of a new element produced by radioactive fission, over time x. The ceiling is imposed by the initial amount of fissionable material.

Now consider volatility. At a large number of parties, a further increase in the number of parties can be expected to become ever less efficient in inducing further volatility. The general exponential equation becomes $V = 100[1 - A\,e^{-kN}]$, where the anchor point $V = 0$ at $N = 1$ requires that $1 - A\,e^{-k} = 0$ and hence $A = e^{k}$. The result is

$$V = 100[1 - e^{k}\,e^{-kN}] = 100[1 - e^{-k(N-1)}].$$

How could such a predictive model be tested? It would be inappropriate to use straight linear regression, which would return us to the coarse linear model. However, the nonlinear equation above can be transformed into linear by transposing and taking natural logarithms:

$$\ln\left(1 - \frac{V}{100}\right) = -k(N - 1).$$

This means that linear regression against N must be carried out on ln $(1 - (V/100))$, not on V itself.

In the case of Heath's data (2005) for state-level elections in India, the rate constant k can be calculated by plugging in the mean N (3.65) and the mean ln(1 - (V/100)). The latter is around −0.384, by my rough calculations. For these means, $k = 0.384/(3.65 - 1) = 0.145$, so that the model predicts

$$V = 100[1 - e^{-0.145(N-1)}].$$

How Precisely Can the Number of Parties Predict Volatility?

Figure 4.2 shows the data points (from Heath 2005) and the curves for the exponential model, the coarse linear model, and the empirical linear fit. The latter two models pass by definition through the point of mean N and V. The exponential curve does not but comes close, because the mean $\ln((V/100) - 1)$ corresponds to $V = 31.9$, close to the mean $V = 31.6$ Above this point, the empirical linear fit lies between the two models. At lower values, the three curves can hardly be distinguished. With this degree of data scatter, all three approaches yield about the same R^2.

It can be seen from Figure 4.2 that the coarse model works about as well as the refined one throughout the usual range of volatility. Hence, we can

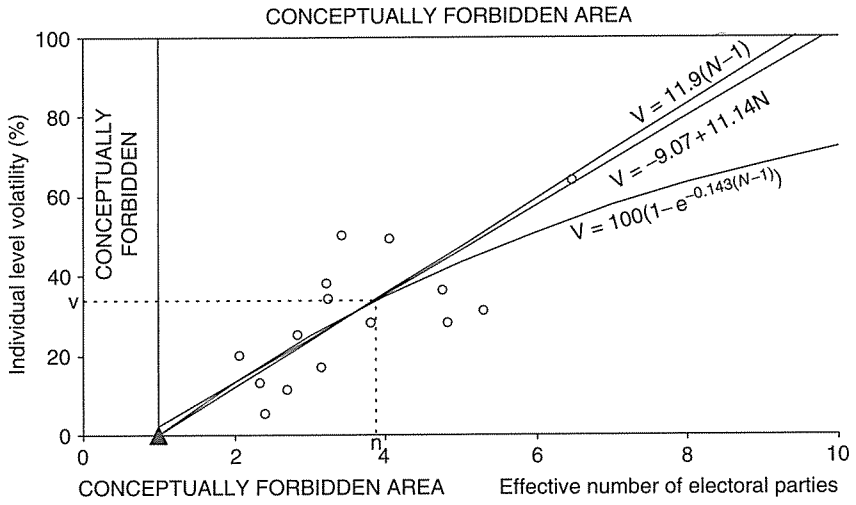

Figure 4.2. Individual-level volatility of votes vs. effective number of electoral parties—data and best linear fit from Heath (2005), plus coarse and refined predictive models

use the simpler model, as long as we keep in mind its limitations. This is a general observation: One must be aware of the refined model, so as to know when a simplification can be used. In physics, classical mechanics is more convenient to handle than the relativistic, and it does well, as long as speeds fall much short of the speed of light. For volatility, the coarse model is more convenient and does well, as long as there are effectively less than five parties.

It would be simplistic to expect that the effective number of parties alone fully determines volatility. Numerous other factors obviously enter. Therefore, the precision of the model should not be exaggerated. While 11.9 is the slope in the coarse model that happens to fit the Indian data, we should round it off to 12 for the purpose of broader worldwide prediction, and introduce a likely error range (ε):

$$V = (12 \pm \varepsilon)(N - 1).$$

How large a range of fluctuation could we expect? If we estimated the slope b on the basis of a single individual state in India, the results could range from 1.4 to 21.4, meaning 12 ± 10. For many elections *under roughly the same conditions*, a coarse rule of thumb would suggest an error equal to the square root of 10, which is about 3. Thus, my prediction for the range of average slopes in other countries is

$$V = (12 \pm 3)(N - 1),$$

provided that India's political culture corresponds to an average tendency of voters to switch parties. The Indian voter is reputed to be unusually prone to change parties, but this impression may arise from the rather large number of parties available.

How does this outcome differ from the descriptive linear regression equation, $V = -9.07 + 11.14N$? The quantitatively predictive model could in principle have been devised prior to any input of data. This means that it is not specific to India—it is expected to apply to all countries, albeit with a wide range of error. The Indian data just helped specify the slope b and verify that $a/b = -1.00$. The equation $V = (12 \pm 3)(N - 1)$ makes a universal quantitative prediction, with a specified range of likely variation.

In contrast, the regression equation deals solely with a particular set of Indian data. It offers the best fit to these data, with several decimals. It is the best way to *postdict* for these data—but for these data alone. Of course, it is a useful starting point for pondering the broader implications of the parameter values obtained. One could have done

Table 4.1. How does the number of parties (N) affect volatility (V)?—predictive and descriptive approaches (modified from Taagepera 2007*b*)

Predictive	Descriptive
Incipient model: $V = f(N)$, $dV/dN > 0$	Incipient model: $V = f(N)$, $dV/dN > 0$
How might things connect?	Do not prejudge, beyond asking "Is $dV/dN > 0$?"
1. Limits, and anchor point $V(1) = 0$.	Get data!
2. $(N - 1)$ switching options $\rightarrow$ try V proportional to $(N - 1)$	Run regression
3. High slope unlikely $\rightarrow$ $b < 20$	Report regression equation: $V = 11.14N - 9.07$
Quantitatively predictive model: $V = bN - b$; $0 < b < 20 \rightarrow b \approx 10$	Hypothesis $dV/dN > 0$ proven
Gather and transform data in the light of the model: mean N and $V \rightarrow V = 11.9N \rightarrow 11.9$	No further prediction to test
Test the predictive model with data: $\rightarrow$ close agreement. Could have been appreciably off, but did not	Any coefficient values are accepted

the regression first, observe that $|a| = 9.07$ and $b = 11.14$ are remarkably close, and wonder whether there might be some logical reason for it. This is how I reacted upon seeing the analysis by Heath (2005), and the coarse model immediately took shape. Unfortunately, social scientists who deal with regressions all too often neglect such follow-up. Hence, they stop short of extracting the quantitative maximum out of their results.

Why would social scientists limit themselves to confirming merely the direction of impact ($dV/dN > 0$, in this case) when they also measure and publish its quantitative extent? One of the reasons may be that automatic fitting with $V = a + bN$ introduces two distinct parameters, a and b, which makes comparison of data-sets hard. Seemingly, one would have to compare both slopes and intercepts. Here, the quantitatively predictive model simplifies comparisons by indicating that one parameter alone should suffice to characterize the relationship in various party systems.

The Main Contrasts Between Predictive and Descriptive Approaches

Table 4.1 reviews the main contrasts between the predictive and descriptive handling of the sample problem, volatility. Beyond the incipient directional model ($dV/dN > 0$), the descriptive researcher would proceed

immediately to data analysis. The predictive model builder would ask: Can we make the model more specific? How might things connect logically?

Following the Sherlock Holmes principle of eliminating the impossible, the predictive model starts by specifying the *boundary conditions*. It then considers further *logical constraints*, which leads to a conceptual *anchor point*. A number of simplifying assumptions may have to be introduced at this point. Next, one may look for the *simplest equation* that satisfies all logical constraints. Predictive models rarely can be linear without violating some boundary conditions, but linear approximations can be used, as long as one specifies their range of applicability.

The next step often concerns the *plausible range of values of coefficients and constants* in the model. Not just any positive values of slope b are acceptable in this case. So a range is established, such that values outside this range would surprise us. Given nothing but such a range and asked to guess at b, our best minimax bet would be the mean of the extremes.

Now, and only now, do we have to enter limited data into the predictive model, although we can enter data earlier, too, to inspire and guide model construction. Merely inserting the mean values of N and V leads to a full predictive model. Here, the numerical value of the single coefficient derives from this minimal data input, while the format is based on reasoning alone. Only past this stage would detailed data enter, for testing the predictive model through regression. In general, models must first be transformed into a linear form, before linear regression makes sense, but here the coarse model already is linear.

Heath (2005) exemplifies well the descriptive approach to the same issue, asking merely: "Do multiparty systems have higher levels of volatility than two-party systems?" The conceptual model is limited to this cautious question about the *direction* of impact. Linear regression immediately follows. The format used, $V = a + bN$, is not based on the specifics of the issue on hand. Linear regression is just what many social scientists tend to apply automatically to any relationship—it is a Pavlovian reflex. No expectations are voiced about how steep or shallow the slope might be, or what value the intercept might have. The numerical values in the regression equation derive fully from the data and from nothing else. Coming *after* data analysis, this equation represents a *postdictive* model.

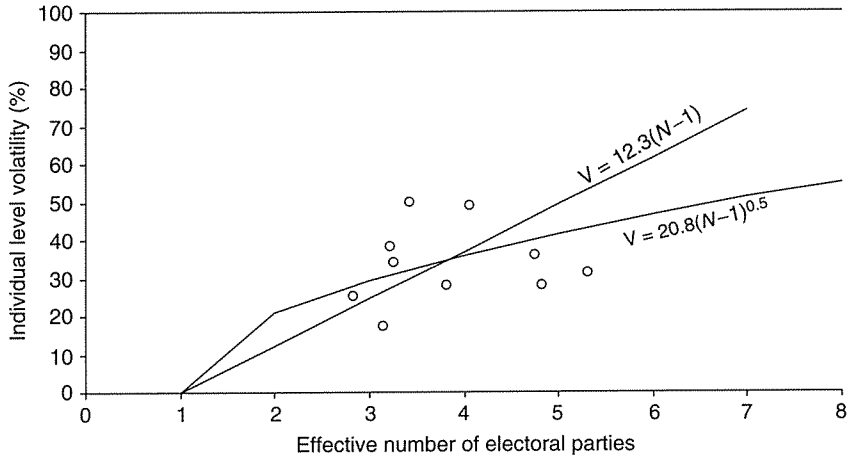

Figure 4.3. Individual-level volatility of votes vs. effective number of electoral parties—truncated data (from Heath 2005) and two models that fit the data and the anchor point

Can Data with Low R^2 Confirm a Model?

For descriptive (or directionally predictive) models, the answer is "no," because data are all there is. Any slope b in the regression equation $y = a + bx$ is accepted, as long as the sign of b is right. The only concern is how closely the data points crowd along the regression line. If R^2 is low, one has nothing.

Such a situation is illustrated in Figure 4.3. The data shown in Figure 4.2 has been truncated by omitting the one point with the highest and the four points with the lowest number of parties. Visibly, the best fit line now is almost horizontal. The slope might be slightly negative or positive, and R^2 is below 0.1. As far as descriptive analysis is concerned, all one could do is to report no correlation between V and N, and that is the end of it.

The outcome is quite different when we use the same data to test the predictive model $V = b(N - 1)$, where we expect a slope around 10. The new mean values of N and V are 3.87 and 35.3, respectively (as compared to 3.65 and 31.6 with the full data-set). The new estimate of slope is $b = 35.3/2.87 = 12.3$, hardly different from the previous 11.9. The expectation that slope is around 10 is confirmed, regardless of the low value of R^2. It would be nice to obtain a high correlation coefficient, but this is hardly essential.

It does not mean that the model $V = b(N - 1)$ is equally well confirmed by the truncated data-set as by the full one. The difference between $R^2 \approx .1$ and $R^2 = .50$ still matters. Suppose someone makes a logical argument for a square root model, $V = a(N - 1)^{0.5}$. Such a model is shown in Figure 4.3, along with the linear. It is fitted to go through the mean point, and it respects the anchor point (1;0). The truncated data could not discriminate between the two models, while the full data-set fits the linear model better. While not able to give preference to the linear model, the truncated data-set would still confirm that this simple model is plausible, until proven otherwise. It would take a much higher R^2 to discriminate between the linear and exponential models in Figure 4.2.

What distinguishes the predictive approach from the descriptive, in the context of Figure 4.3, is that the descriptive approach has nothing but the 10 actual data points at its disposal. When these points show no pattern, we effectively have a single point with some scatter around it. This is the data configuration in Figure 4.3. In contrast, the predictive approach adds the virtual data point at $N = 1$, $V = 0$. This addition of the anchor point may look weak, in the absence of real data at this location. Actually, it is extremely strong and precise, given our certitude that if we should ever get data points where $N = 1.00$, then every single one of them would have $V = 0.00$, with no margin of error.

Thus, even when $R^2 = 0$ for the actual data points, the predictive approach has not one but two very distinct points at its disposal, and two points determine a line. In the absence of any further knowledge, we can use this line as a first approximation for a predictive model. It would be less than optimal to refuse to do so and say "We cannot predict anything" when we actually can—within a margin of error. The extent of the scatter of points in Figure 4.2 (or 4.3) gives us an idea about the extent of possible error.

Actually, there is a situation where one might be happier with a lower R^2 (cf. discussion of Figure 3.2). This is when the best fit curve deviates from the predicted. In such a case, more scatter would leave more hope that further data might bring the observed pattern closer to the expected. This would be so only from the viewpoint of trying to preserve the existing model. There may be no need to preserve it. A high R^2 around an unexpected curve offers valuable guidance for adjusting the model.

In sum, there are more important considerations than high R^2, if one wants to make sense of nature. R^2 indicates by how much a regression equation *accounts* for some variance in y for given x, in a statistical sense. However, a high R^2 alone does not *explain* anything in a substantive sense

that would allow prediction for other data-sets. It is nice to have a high R^2 (provided the best fit curve agrees with the model), but even a scattered data cloud does not sink a predictive model predicated on anchor points far away from the data cloud. A low R^2 may just indicate that the observed range of the input variable is too narrow to bring the trend into evidence, against a noisy background. One should then try to find ways to extend the range over which the model is tested.

On a broader note, data are not sacrosanct. They may be systematically distorted by some pervasive factors not included in the model, like the effect of gravity is by air friction. Or the data may not refer to the concept in the model but to a related, yet distinct concept. If one has a logically well-founded model, the first reaction to contradictory data might well be "Damn the data, full speed ahead!" One can sometimes outrun torpedoes and data. Later, of course, more suitable data must be located. Chapter 13 elaborates on this issue.

Some General Features of Constraint-Based Models

Quantitatively predictive logical models can at times be constructed prior to any input of data. Most often, however, some initial data inspire or guide model building, although these data might be much more limited than what would be eventually needed for serious testing. Oftentimes, even a cursory graphing brings out a pattern that makes one ask "Why?" This is the way the graph of volatility versus number of parties in Heath (2005) made me look for the reasons behind the regularity observed.

In building models, our choices are constrained in a highly constructive way by conceptually forbidden areas, where no data points can exist, and anchor points, through which the expected average curve must pass. The scientific quest for relationships need not establish that nature or God has chosen one particular form; it often suffices to show that any other choice would run into contradictions. Entering a forbidden zone or missing an anchor point would be among such contradictions. Let us review how such considerations entered the model for volatility.

1. Use boundary conditions, ceilings, and other logical constraints; establish anchor points.
2. Look for the simplest set of equations that does not violate the logical constraints.

3. Wonder about the possible range of values of coefficients and constants. Some limits are firm, while some are fluid. It may look unscientific to use such vague limits, but the reverse is true: It is not good science to ignore obvious limitations just because we cannot pin them down with two-digit precision. They still exist.
4. Use the means of data to estimate some coefficients in the predictive model. This step goes beyond dataless prediction but still precedes regression.

Continuity

The notion of continuity has been implicitly linked here to that of anchor points. In macroscopic natural or social phenomena, a small change in x mostly brings a small change in y. This means that the path from one anchor point to another (or to a conceptual ceiling) must form a continuous curve. Discontinuities do occur, for example, in aerodynamic theory (at speed of sound) and in utility theory. However, they are rather rare in *macroscopic* phenomena. At the microscopic level, the seemingly continuous electric current consists of discrete electrons and water flow has discrete molecules, but for large quantities of such particles, the micro-level granularity can be overlooked. Similarly, an electorate consisting of discrete voters can be treated as a continuous quantity, as long as there are thousands of them. One has to be cautious, of course, of not inadvertently stepping into territory where "granularity" can make a difference.

Note that the presumption of continuity applies to much of statistics as well. We presume quasi-continuity whenever we try to fit a distribution of discrete entities by a smooth normal curve (or other continuous curve), rather than as a discontinuous histogram.

Changes in slope (dy/dx) also tend to be continuous, and the same goes for the "slope of the slope" (d^2y/dx^2) and the higher derivatives (d^ny/dx^n). Moreover, the curves are expected to go monotonically up or down ($dy/dx > 0$ everywhere or $dy/dx < 0$ everywhere), *unless there is a reason* for reversal. If the data should present an unexpected kink in the pattern, the data should be checked. If the kink remains, the underlying reason must be discovered and worked into the model.

Conclusions

This specific example has introduced terms like forbidden areas, anchor points, and ceilings. It has illustrated the differing significance of the degree of correlation (R^2, in particular) when data is all we have (descriptive models) and when we also have a predictive model that goes beyond predicting directionality. Chapter 8 presents such notions in a more systematic form. Of course, a constraint-based approach is not the only one to building logical models. Chapter 11 expands into various model-building methods that were not needed in the specific case of volatility.

Appendix to Chapter 4

Further Refinements and Aggregate Volatility

Further improvement in predicting volatility may go in several directions. Data may be improved, addressing the "party of nonvoters" and the sticky problems that arise from party splits and fusions. Further input variables may be introduced. If further factors appreciably reduce the remaining variation in volatility, it may turn out that volatility's relationship to the number of parties deviates significantly even from the exponential model. In this case, one would have to review both the way the number of parties is measured and the way it could interact with volatility.

The term "interact" is used here on purpose. Up to now, we have proceeded as if causality were one-directional, the number of parties affecting volatility. This need not be so. Those who consider forming a new party may be encouraged if they know that voters are highly volatile and may easily switch to a new party. Thus, volatility may indirectly affect the number of parties.

Could further work show that some other factor is actually more important than the number of parties? Heath (2005) found that the number of parties accounts for one-half of the variation in volatility ($R^2 = .50$). This suggests that we have pinned down the major determinant of volatility, or at least one of two major ones, as no other single factor can account for more, unless it is a factor that affects the number of parties and also affects volatility directly.

All previous models refer to *individual level volatility* (V_I). This can be determined only by exit polls where voters are asked their present and previous party preference—and hope that the answers fit the facts. It is much easier to determine the *aggregate volatility* (V_A), based on how the vote shares of parties change from one election to the next. This is a lower number than V_I, because voters switching from party B to party C and vice versa may cancel out. It may also be of more interest. How could we estimate V_A?

In principle, aggregate volatility could range from 0 (full cancelling out of individual shifts) to V_I (no cancellation at all). Hence, it can be expected to be

around one-half of the individual volatility, *on the average*. On the strength of the model for V_I and the value of slope constant $b = 12$ specified by the Indian individual volatility data, we can presume that the slope would be around 6 in a coarse model for aggregate volatility, as long as N remains moderate:

$$V_A \approx (6 \pm 1.5)(N - 1).$$

When initially constructing this model for aggregate volatility, I desisted from collecting any data. This way, it would truly represent a purely theoretical quantitative prediction regarding aggregate volatility, derived from data on individual volatility. Since then, Mainwaring and Torcal (2006) have reported aggregate volatility figures that lead to slopes ranging from 2 to 7 in the case of stable democracies. The slope can reach 15 in early elections in new democracies. Full testing remains to be done. Then we will now how close the prediction above was.