If testing does not confirm the model, not only the model may be reformulated but also the data may be revised. Either way, the testing cycle is repeated. Once testing is highly successful, a law in the scientific sense might result. Finally, combination of many interlocking laws may lead to a theory as a grand roof.

In sum, linear regression and other descriptive approaches have their place in preliminary investigation at the one end and in testing logically established quantitative models at the other. In between, descriptive approaches alone are disastrous when they substitute for logical modeling.

15

Recommendations for Better Regression

- Take seriously the introductory advice by most introductory texts of statistics: graph the data and look at the graph so as to make sure linear regression makes sense from a *statistical* viewpoint.
- Graph more than the data—graph the entire conceptually allowed area and anchor points so as to make sure linear regression makes sense from a *substantive* viewpoint.
- If using linear regression, report not only the regression coefficients and the intercept but also the ranges, mean values, and medians of all input variables.
- Symmetric regression has advantages over OLS. But fully reported and symmetric regression is still merely regression.
- Look up further recommendations in the Conclusions.

Establishing a quantitatively predictive logical model is never automatic. One has to understand the nature of the problem on hand. It is easy to give such general advice, but it is not very helpful. How does one start? What is the first practical step, for a person who knows the statistical methods to some extent but has no idea where to go beyond that? This chapter addresses the issue of starting from scratch, or almost so, and making the most of statistical approaches.

Data: Graph It!

This is the advice all good introductory statistics texts offer (e.g., King et al. 1994; Berry and Sanders 2000). Take them seriously! Once this advice is given, these texts assume that you have followed it, and they focus on

Table 15.1. Four data-sets that lead to the same linear fit and R^2 when linear regression is (mis)applied

X	y_1	y_2	y_3	y_4
4.00	4.26	3.10	5.39	6.00
5.00	5.68	4.74	5.73	5.75
6.00	7.24	6.13	6.08	5.50
7.00	4.82	7.26	6.42	5.25
8.00	6.95	8.14	6.77	5.00
9.00	8.81	8.77	7.11	7.50
10.00	8.04	9.14	7.46	10.00
11.00	8.33	9.26	7.81	9.75
12.00	10.84	9.13	8.15	9.50
13.00	7.58	8.74	12.74	9.25
14.00	9.96	8.10	8.84	9.00

Source for y_1 to y_3: Anscombe (1973); y_4 is my own addition.
All y_n vs. x have the same properties: mean $x = 9.00$; mean $y = 7.50$, $R = .82$; $R^2 = .67$; regression line y on x: $y = 3.00 + 0.50x$. Regression line x on y corresponds to $y = 0.75 + 0.75x$, and symmetric regression line is $y = 1.9 + 0.61x$.

data for which linear analysis is justified. The introductory text may later address the issue of fitting nonlinear data configurations, but it may do it so briefly as to leave the mistaken impression that linear configurations are the rule and curved configurations rare exceptions. The reverse is the case, once forbidden areas are taken into account.

Various patterns where linear analysis would ignore reality were presented in Figures 3.1 and 3.2. The matter is so important that it is worth offering further cautionary examples. Anscombe (1973) has published a splendid collection of four data-sets which look identical in standard linear regression. Three of them are shown in Table 15.1 (y_1, y_2, y_3), along with an addition of my own (y_4). They all have the same mean and range of the input variable, the same mean and approximately the same range of the output variable, the same respectable linear correlation coefficient ($R = .82$, $R^2 = .67$), the same OLS regression lines y-on-x and x-on-y, the same symmetric regression line, etc. Yet linear regression makes sense in only one of the four data-sets, as jumps to the eye the moment these data are graphed in Figure 15.1.

It can be seen that linear regression looks acceptable for y_1 because the data cloud is uniformly dispersed, with no visible curvature. In the case of y_2, the points fit neatly on a parabolic-looking curve, and a corresponding transformation should be applied before statistical testing. The transformation could be based on statistical considerations, but this is also prime time for asking *why* this is so that y first rises and then falls with

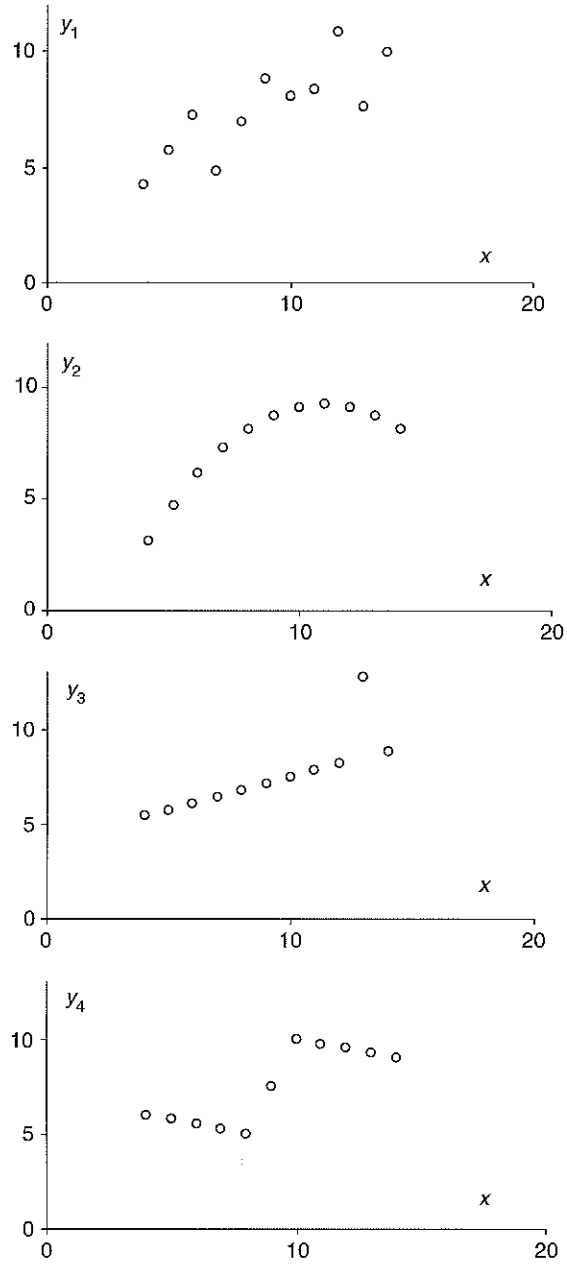

Figure 15.1. Graphing the data from Table 15.1 checks on whether linear regression makes sense

increasing x. Data-set y_3 has 10 points perfectly aligned, while one point is a blatant outlier which clearly does not belong and should be omitted. The statistical justification for deletion is that it deviates by more than 3 standard deviations. One should also try to figure out how it came to be included in the first place. Maybe there is a typo.

All these three data-sets have the same x values, ranging uniformly from 4 to 14. In contrast, Anscombe's (1973) fourth data-set (not shown) had 10 points with the same value of x and hence forming a vertical line, plus a single point elsewhere. I have replaced it in Table 15.1 and Figure 15.1 with another data-set (y_4) where x ranges from 4 to 14. This set, too, has the same mean x, mean y, regression equations, and R^2 as the rest. When graphed, a pattern emerges that is far from a rising straight line. We observe two distinct populations where y actually decreases with increasing x, plus an isolate. This pattern should make us wonder about the underlying structure.

Graph More than the Data!

Graphs should include not only the data but also boundary conditions, anchor points, and sometimes the equality lines. Only then can the data be seen in a wider perspective conducive to model building. In particular, when both x and y are in percentages, entire ranges from 0 to 100% should be shown.

This was effectively the case in Figure 12.1, which also illustrates two general features. First, linear regression oblivious of conceptual constraints can lead to viewing as deviant some data points that eminently do fit. Second, drawing in the equality line may offer a handy comparison level, even when there is no reason to expect equality of y and x. Some simple patterns that may make sense were discussed in Chapter 8.

Figure 15.2 (reproduced from Taagepera 2007*c*: 71) illustrates the need to graph more than the data. It shows the "proportionality profiles" for elections in two countries: advantage ratio (% seats/% votes) graphed versus the percentage of votes. As far as the data are concerned, the range beyond 60% could be omitted—and often mistakenly is. When graphing only the US data, the range below 30% would also seem superfluous, and the range of advantage ratio could be restricted to 0.7 to 1.2. Indeed, many computer programs impose the ranges 30 to 60 and 0.7 to 1.2. The empirical "best fit" OLS line would be calculated, along with R^2 and

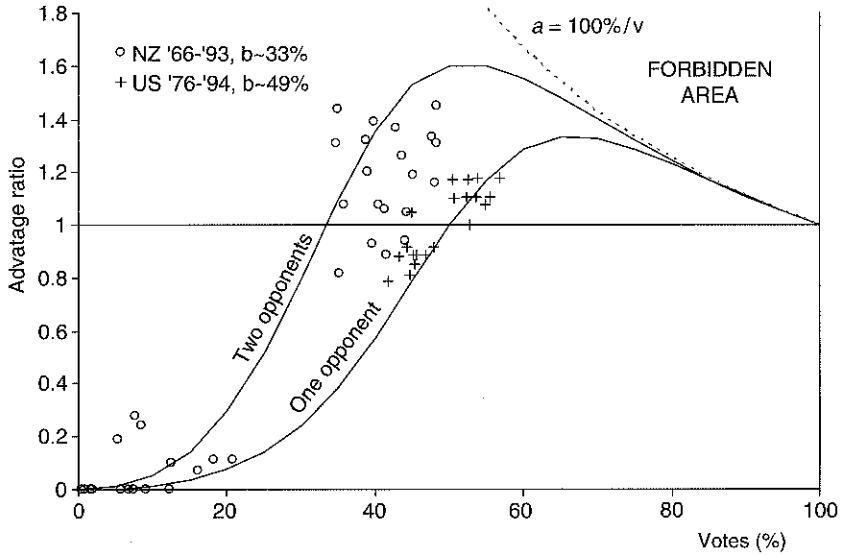

Figure 15.2. Proportionality profiles for elections in New Zealand and the United States: data and conceptual range (from Taagepera 2007*c*: 71)

some measure of significance. This line would extrapolate to a very high advantage ratio near 100% votes and a negative advantage ratio (implying negative number of seats) near 0% votes. However, much food for thought emerges the moment the entire allowed range is included in the graph—0 to 100% votes and 0 to much above 1 for the advantage ratio, as shown in Figure 15.2.

Once the full conceptual range is included in the graph, we are motivated to ask what the advantage ratio would be for zero votes. It would be expected to lead to zero seats. This may look like an indeterminate zero/zero, but consider the percentage of votes *tending toward* zero. At sufficiently few votes, any electoral system would allocate zero seats, so that the advantage ratio becomes zero. This is a lower anchor point.

We would also be motivated to ask what the advantage ratio would be for 100% votes. Such a party would be expected to win 100% of seats—but no more! This means an advantage ratio of only 1.00. This is an upper anchor point. It is then easy to establish that $a > 100\%/v$ is a conceptually forbidden area, as indicated in Figure 15.2.

In sum, the linear OLS fit becomes untenable. The curve must bend up toward the point (0; 0) at low values. It also must bend down toward

(100; 1) at high values of votes. One could draw in such a curve, freehand, and one would not be far off from the curve shown, which is based on a logical model (see Taagepera 2007*c*: 201–11).

Consider Graphing by Hand

Figure 12.1 further illustrates some problematic features that canned graphing programs may present. When graphing on a computer, one may find that the program used imposes an autoformat mood that refuses to extend the scale beyond the range of actual data, or extends it excessively, and also refuses to enter the equality line.

The scales shown on the two axes in Figure 12.1 run from less than 0 to more than 10, although the ratings involved can go only from 0 to 10. Hence the borders shown are wider than the conceptual limits. In the given case the entire conceptually allowed space is filled with data points, however sparsely, so that the conceptual limits stand out visually despite not being shown. With fewer data points, however, visual impression can be quite misleading, by suggesting that extreme cases do not occur when they actually do. This is the case, for instance, with several graphs in Arend Lijphart's excellent *Patterns of Democracy* (1999: 193, 214, 229, and 241). The author tells me that his original graphs were restricted to the conceptual limits, but they were altered during the publishing process so as to look nicer by some nonscientific criteria. Thus, even those of us who know better have to struggle against the dictates of computerized graphing.

The autoformat of graphing programs usually can be overridden, but some overrides need fair computer skills. Small wonder then that students sometimes offer technical restrictions as something one has to yield to and accept. This is not so. Computers are supposed to help us do science, not hamper us. If a canned program restricts research, look for a more flexible one. If none can be found at the moment, do the graphs by hand. As an alternative, one may copy the computer graph in a sufficiently reduced form so that one can add by hand the conceptual limits (and the equality line, if it makes sense).

There are some broad advantages in doing graphs by hand, on paper with a square grid. My experience is that students who graph by hand understand graphs better and notice more detailed implications, compared to those who let the machine do the job. This understanding becomes crucial when graphing on logarithmic scales is required, as is

the case whenever relationships have the form $y = ax^b$ or $y = ae^{kx}$. All too many social science students cannot take in the information contained in printed graphs of such format unless they have graphed data themselves on old-fashioned "fully log" or "semi-log" graph papers.

When the data involve more than one input variable, simple two-dimensional graphing can use only one input at a time. A second input variable can be introduced by color-coding data points as Low, Medium, or High for that variable. Approximate best fit curves for these subgroups can yield some ideas on how this second input affects the output. Constraints imposed by the second input can be shown in a similar way.

One may run an exploratory multivariable linear regression (preferably the symmetric version) so as to get some rough idea of which variables matter the most. Then the output can be graphed against some of the inputs, to check that there is no curvature or some other odd feature akin to those in Figure 14.1. There are statistically more elegant ways to do it, but the power of eyeballing should not be undervalued. This is so even when one truly understands the statistical methods one uses, including their limitations—and even more so when this is not the case. The gravitation test in Chapter 2 serves as a warning: There is no guarantee that multifactor linear regression can pin down the underlying processes, but it is one way to proceed in exploratory research.

Turning the Pattern Linear

Once the pattern of data plus conceptual anchor points and boundary conditions has been established, one should look for an explanation in terms of a quantitatively predictive logical model. One may or may not succeed at the moment. Either way, the data should be transformed into a format where linear regression or some other standard statistical procedure (e.g., *logit* or *probit*) makes sense. If a logical model is proposed and it involves no free parameters, one could graph the actual values against the expected ones and run symmetric linear regression. If no model can be proposed, some empirical transformation may be found, such as the ones in Chapter 8, which transforms the data cloud *and the anchor points* into something not outrageously far from linear shape.

Many statistical approaches go beyond linear regression, but I will focus on the latter, because a fair share of quantitative publications in social sciences uses something based on or closely related to OLS version of

linear regression. Thus any issues raised in this context have wider implications. Preferably, OLS should be replaced by symmetric linear regression (Chapter 12), but either way, regression results should be reported in such a way that other researchers can make the most of them. This is not so self-evident.

How to Publish Regression Results

Why do we publish regression coefficients? What is the purpose of getting into print tables that sometimes fill an appreciable chunk of the total space in an article? We presumably want to let other scholars know about our research and enable them to make use of it (hopefully leading to entries to our benefit in their *citation index*). The eventual overall outcome of such joint effort should be some cumulative knowledge useful to society and/or decision-makers. If so, then how come that a statistician recently told me that many social science papers look qualitative, with statistical analysis added as an afterthought or decoration?

A colleague once told me that the purpose of publishing statistical stuff is to satisfy grant-giving outfits, even while "everyone knows that it is meaningless." Most of us do not play such a cynical game, but many may respectfully follow a mysterious tradition. Publishing regression coefficients is just what we are supposed to do, so as to fit the commonly accepted norm. Trouble is that the game earns us little respect outside the profession, if it produces no useful results. We can do better than that.

When we publish regression coefficients, then why do we do it? There might be goals less demanding than quantitative prediction (precise or imprecise) or at least postdiction. This issue is discussed in the Appendix to Chapter 15. What is certain is that publishing of regression coefficients becomes mandatory if we want to give our colleagues (and ourselves) some predictive/postdictive capability beyond the direction of impact. "When $a, b, \ldots$ are the given numerical values of variables $A, B, \ldots$ our best guess for the numerical value of y would be $y = b_0 + b_A a + b_B b + \ldots$, where b_0 is the intercept, and the other b_i are the regression coefficients for $A, B, \ldots$." This is predictive in an explanatory way, if there is a logical model, and at least postdictive (cf. Chapter 1), if no logical model can be found. Moreover, regression coefficients can sometimes supply a starting point toward a quantitatively predictive logical model (see Chapter 16).

Enabling predictions is a major intellectually supportable reason for spending printed space on numerical values of regression coefficients. If

Table 15.2. A typical table of regression results

Independent variables	N_V	N_S
Effective threshold (T)	−0.03**	−0.05**
Log assembly size (log$S$)	0.12	0.12
Intercept	4.07	3.66
R^2	.11	.30
Adjusted R^2	.08	.28

*Statistically significant at the 5% level.
**Statistically significant at the 1% level.

we publish them, we might as well do so in the most fruitful way. It will be seen that even a little upgrade of present practices would go a long way. The advice has been around at least since Gary King's *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* (1989).

The example in Table 15.2 is excerpted from a case (Lijphart 1994: 108) that already is among the better ones, as it does include all the regression coefficients, intercept included. Standardized coefficients, t-values and some output variables have been omitted here, and the labels for variables have been modified. Only two input variables have been kept: effective threshold of representation (T), and the logarithm of assembly size (log$S$). The output variables are the effective numbers of parties (as defined in Chapter 4), based respectively on votes (for N_V) and on seats (for N_S). The database can be inferred from other tables in Lijphart (1994), and this comes in handy in the course of the following discussion.

This table means that the average outputs for given inputs can be calculated from $N_V = 4.07 + 0.12 \log S - 0.03T$ and $N_S = 3.66 + 0.12 \log S - 0.05T$, respectively. For a given electoral system, one could look up the actual values of T and S, and calculate the best estimates for N from these equations. But Table 15.2 indicates that S is not statistically significant. Then why should we have to look up its specific values when we want to calculate N_V and N_S? The values of S are random and could as well be replaced by the *mean value* of S for the cases used in regression (King 1989: 105)—if we knew this mean value. The problem is that Table 15.2 does *not* report the mean values of the input variables. Despite King's (1989) advice to include them, this has not been part of general practice—and this is precisely my point. *The way regression coefficients are customarily published, one cannot use them for prediction, short of looking up lots of input data which the author has found to be nonsignificant.*

The present example was chosen because in this case the data in other tables in Lijphart (1994) enables me to estimate the geometric mean of

S. It is around 148 seats. (The median is 158.) With substitution $\log S = \log 148 = 2.17$, the previous equations become $N_V = 4.07 + 0.12 \times 2.17 - 0.03T = 4.33 - 0.03T$ and $N_S = 3.66 + 0.12 \times 2.17 - 0.05T = 3.92 - 0.05T$. If one graphs N versus T, these are the best fit equations.

It is extremely important to note that one can no longer use the intercept values listed in Table 15.2, once one omits variable S. This would be akin to assuming $0.12 \log S = 0$ for the average assembly, which is the case when the assembly has only one seat—an obvious underestimate. When one omits nonsignificant variables, the new intercept must include the average effect of the omitted variables. All linear (and similar) regression results worth publishing should report not only the regression coefficients and the intercept but also the mean values of all input variables.

Why require also the mean values of significant factors? There are four reasons. The first is uniformity of reporting. Second, when there are several statistically significant factors, we may wish to focus on only one, using the mean values of the others. This is what we would do, in particular, when graphing the output against one of the inputs, so as to detect possible nonlinear relationships (King 1989: 105). Third, reporting the y-on-x regression equation plus R^2 enables one to calculate the x-on-y and symmetric regression lines (Chapter 12), provided that the means are also given.

The fourth reason is that all too often there are many ways to assign a measure to a conceptual variable (cf. Chapter 13). Social science authors are not very good at specifying which measure they are using, especially when they personally are used to one of them, to the exclusion of all others. Thus, an author may describe a variable as party system "fractionalization" while actually reporting the effective number of parties (N) rather than the Rae-Taylor fractionalization index (F). Since F varies from 0 to 1 while N varies from 1 upwards, reporting the mean value helps to clarify which measure was used in the given regression equation.

But even more should be reported. In response to a draft of this section, Steve Coleman suggests that the domain (range) of all input variables should be reported. This is indeed desirable, as it would tell us over what range the model can reasonably be used to estimate outputs. In the present case, S ranges from 60 to 630 ($\log S$ ranging from 1.8 to 2.8) and T (in percent) from 0.1 to 35.

A zero value may be well outside these domains, as is the case here for $\log S$. If so, then minor changes in data may cause large fluctuations in the value of the intercept. As a result, the intercept may look statistically "not significant" even at the 5% level. This is the case in Table 15.2. I have

Table 15.3. A typical table of regression results, with suggested complements

Independent variables	Domain (Range)	Mean	Median	Coefficients for	
				N_V	N_S
Effective threshold (T)	0.1 to 35	11.6	7.0	−0.03**	−0.05**
Log assembly size ($\log S$)	1.8 to 2.8	2.2	2.2	0.12	0.12
Intercept				4.07	3.66
R^2				0.11	0.30
Adjusted R^2				0.08	0.28

*Statistically significant at the 5% level.
**Statistically significant at the 1% level.

encountered arguments that such a "nonsignificant" intercept should be omitted from estimates of outputs, but this would mean assigning a zero value to the intercept, which often makes no sense at all. Intercepts 4.07 and 3.66 in Table 15.2 might conceivably be off by ±0.1 or even ±1, but assuming them to be 0 would lead to practically all estimates of the number of parties to be negative. We must avoid absurdities.

In addition to arithmetic mean, the median should be reported, because a disagreement between mean and median serves as a simple warning light to show that the actual relationship is not linear. In the case of Table 15.2, data in Lijphart (1994) confirm that mean and median of $\log S$ are both 2.2, but for T the mean (11.6) strongly exceeds the median (7.0). The corresponding curvature appears when graphing N versus T (not shown here). It suggests that linear regression should be applied to $\log T$ rather than T. This idea is reinforced when one notes that effective threshold cannot be negative.

Table 15.3 expands the previous table to include the suggested complements. In sum, we can easily improve on the customary format of reporting, so as to make the published results much more useful for estimates of outputs, for graphing the average patterns, for calculating reverse and symmetric regression lines, and for resulting comparison with analysis of other data-sets. It also helps in devising logical models, something not followed up in this example. The general recommendation is as follows.

All linear (and similar) regression results worth publishing should report not only the regression coefficients and the intercept but also the ranges, mean values, and medians of all input variables.

Steve Coleman raises one further point that goes back to Fisher (1956: 42): "Personally I find it annoying when people only report what is significant at certain p levels, usually $p < .05$, which is a not very helpful convention and more of a historical accident of statistics. I prefer the

actual p value so I can make up my own mind" (Steven Coleman, personal communication, June 2007). It might take little extra room to add this information too.

Conclusion: Ten Recommendations for Running and Reporting Linear Regression

Statistical analysis and regression in particular can be done better than it often has been done in social sciences, by following a few simple recommendations. For those based on arguments presented in other chapters, these chapters are briefly indicated.

1. *Use regression only for exploratory research and for testing quantitatively predictive logical models.* These are the early and the late phases in research process. Do not even think of using regression for model construction itself (cf. Chapter 14).
2. *Graph possibly meaningful relationships, so as to avoid running linear regression on inappropriate data configurations.* First, this means taking seriously the introductory advice by most statistics texts; graph the data and look at the graph so as to make sure linear regression makes sense from a *statistical* viewpoint. Second, graph more than the data—graph the entire conceptually allowed area and anchor points so as to make sure linear regression makes sense from a *substantive* viewpoint.
3. *Replace the practice of profusion of variables by the principle of parsimony.* Having more than two or three input variables in a regression disregards Occam's razor (cf. Chapters 3 and 5). Use sequential equations rather a single melting pot. Avoid dummy variables.
4. *When regression makes sense at all, replace unidirectional by symmetric regression* (cf. Chapter 12).
5. *Distinguish between statistical significance and substantive meaningfulness* (cf. Chapter 6).
6. *Avoid "asterisk syndrome"*—report actual significance levels rather than $p < .01$ and $p < .05$.
7. *Report not only regression coefficients and the intercept but also the ranges, means, and medians for all input variables.*
8. *Report only one of the many regressions you might run,* the one you deem the most meaningful (cf. Chapter 5 and 7).

9. *Run separate regressions for low, median, and high thirds of those "control variables" you really deem meaningful,* not just statistically "significant." This is a loophole for item 8, an afterthought that may be debatable.
10. *Do not use these recommendations blindly.* Think. There can be exceptions. Know what you do and do what you know.

It should not be concluded that following this advice would make indiscriminate application of statistics acceptable. When linear regression amounts to unjustified data crunching, it remains so even when the results are fully reported, symmetric regression is used, etc.

Appendix to Chapter 15

How NOT to Publish Regression Results

One rarely encounters articles that show only the correlation coefficient R^2, omitting the regression coefficients. Rather frequently, however, intercept is omitted. I will first address the information value (or lack thereof) of such presentations, followed by more general discussion.

I Have a Well-Fitting Relationship but Will Not Tell You What It Is

In his study of volatility (as reported in Chapter 4), Heath (2005) observes that earlier studies (Pedersen 1983; Bartolini and Mair 1990) also obtain a positive relationship with the number of parties but his has a higher R^2. He could not compare his best-fit equation to previous ones, because these earlier studies only reported correlation coefficients, without giving the substantive equation relative to which the R^2 is calculated. Now imagine Galileo reporting the R^2 for speeds of falling bodies, while omitting the equation to which it relates! If this had been the practice in physics, we would not have computers in the first place, with which even those averse to mathematics can calculate the R^2.

It may be argued that one might want to see if a set of variables "explains" an outcome better than another set, by comparing their R^2, without aiming at prediction. But what does an "explanation" stand for, if devoid of ability to predict? In the case above, if all authors obtained roughly the same intercept and slope in $V = a + bN$, we would have a solid empirical regularity even if all of them found low values of R^2. The numerical values of a and b would still look reproducible. On the other hand, if they all got high R^2 but for wildly different regression lines, then we would have nothing, unless we could introduce some other factor to explain the discrepancy. In such a case, high values of R^2 would actually enhance the confusion (cf. Chapter 4, section "Can data with low R^2 confirm a model?").

Correlation coefficients are pointless in the absence of equations. It should be realized that even the worst-quality data fit equation is worth more than an excellent value of R^2, reported devoid of the substantive equation to which it applies. If a physicist reported R^2 alone, he would meet blank stares: You say you have good quality results, but what *are* the results? The practice of reporting R^2 without the equation to which it refers has fortunately become rare. The cult of R^2 carried to that point would pull any journal down to the level of pseudo-scientific formalism.

What Can One Predict Without Intercept?

The argument can be made that some types of predictions can be made without the intercept. In many applied situations one is concerned with the change in output variable (y) when an input variable (x) is changed marginally (by an amount Δx). For any given starting position (x_0; y_0), if we know the regression slope b, we supposedly can infer that at $x_1 = x_0 + \Delta x$ we would have $y_1 = y_0 + b\Delta x$. No knowledge of intercept seems needed. To what extent does it hold?

Consider a regression line with positive slope, as shown in Figure 15.3. The regression slope is the property of the data-set as a whole. As such, it is likely to apply to *typical* points, those at or close to the regression line, such as point A in Figure 15.3. Starting from A, a marginal increase in x is likely to move us to a point A_1, on the regression line.

But what about an outlier, such as point B? How sure can we be that a movement parallel to the regression line would take place, to point B_1, as application of slope b would lead to? The location of B could result from a random disturbance off the general pattern, which represents a sort of equilibrium. If so, then an increase in x

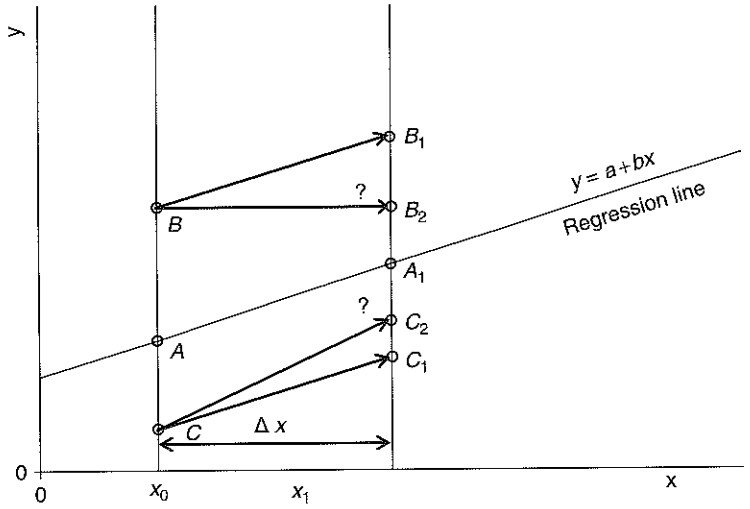

Figure 15.3. Will outliers follow the regression slope?

may exert no pressure at all on B to move further upwards, and the outcome might be point B_2. In the absence of any further information, our best guess would be that the actual slope would be somewhat less than b. Conversely, an outlier in the opposite direction, such as C, might move parallel to the regression line, to point C_1, but it might also move with a steeper slope than b, to a point such as C_2, if its outlier position was due to a random disturbance off the general pattern.

In sum, the more the starting point is off the regression line, the more the slope of a marginal change is likely to deviate from regression slope b. The likely direction of deviation, downwards or upwards, is such as to have the point approach the regression line. We are safe to use b only when the starting point is reasonably close to the regression line. (I will not go into the mathematics of what is "reasonable.") How do we know that the point we start from is not a marked outlier? We have to compare it to the regression line—which means we have to know intercept a.

Estimating the change in y for a marginal change in x on the basis of regression slope alone is based on the tacit assumption that the starting point is close to the regression line. Even in the absence of intercept, we can of course try to do our best. When not knowing whether our starting point deviates upwards or downwards from the regression line, our best guess is indeed that it is on this line. But why omit reporting the intercept, when this information is readily available and would enable the users of one's results to avoid the risk of predicting on the basis of an outlier?

Why do We Publish Regression Coefficients?

When we publish regression coefficients, then why do we do it? If the purpose is merely to report which factors have a "significant" effect on the output and in which direction, then it can be done in briefer form. No reporting of regression coefficients is needed for rejecting the null hypothesis or confirming a directional hypothesis—except for proving that one has actually carried out the statistical analysis. Beyond such reasons of transparency of analysis, one may just report which factors have a statistically "significant" positive effect, and in which direction. Printing regression coefficients would be superfluous for these purposes. These coefficients are needed only when one wants to make predictions.

This claim of mine has raised many hackles. Goals are asserted to exist that do not involve quantitative prediction but still need presentation of numerical values of coefficients. One may highlight a specific relation in a way that falls short of a strong quantitative interpretation, yet could not easily be summarized with only significance and direction. The objective may be to describe the general shape of causal relationships, rather than precise point estimates, so I am told.

All right, I can backtrack a little. There might be goals that fall between directionality and quantitative prediction. I might even think of examples in physics where such goals served a purpose—but only a temporary purpose while looking for predictability. It would be surprising if a large part of social sciences remained in that stage of development without need or possibility to proceed toward prediction.

Even purely empirical regressions equation could be predictive at least in a limited sense. If one plugged into it values of input variables for a case not included in the original regression, then the predictions could be compared to the actual values of the output variable—and deviations could occur. Calculation of predicted output values for new cases is the simplest to visualize and to carry out when the regression equation $y = a + \sum b_i x_i$ is written out, with the actual numerical values of coefficients inserted. Then it is easy to plug in the values of input variables for a new case. (The procedure is most direct for OLS, while requiring a standard conversion of variables in the case of *logit* and *probit*.) But this is not the usual format in social sciences. Most authors present coefficients b_i in a table, not within an equation.

Now suppose one asked the author of such an article: "OK, when for a given country we have $x_1 = 30$, $x_2 = 0.7$ and $x_3 = 273$, what is your best guess for the value of y for this country, on the basis of your table?" One may have doubts about some social scientists' ability to answer such a question, because in quite a few cases the constant (a) is omitted from their table (cf. Chapter 7), so that no answer is possible.

When such omissions can occur in published social science work, it strongly suggests that neither authors nor reviewers have prediction in mind. If an article includes several different "empirical models" for the same output variable, each would produce a somewhat different answer. Take your pick. Offering many competing answers amounts to offering none. For physics students, plugging values into a given formula is the lowest type of exercise, devoid of much thinking. But to what extent have social science practices reached even that stage?

16

Converting from Descriptive Analysis to Predictive Models

- The results of existing statistical analysis can sometimes be used to estimate the parameters in quantitatively predictive logical models. This is important, because it expands the value of previously published work in social sciences.
- Inferring logical model parameters in this way may require more involved mathematics than direct testing.

Suppose one already has analyzed the problem extensively from the statistical angle. When one wants to proceed to logical model building, how much can one extract from published statistical data analysis rather than start from scratch? Echoing Freedman (1987), Hedström (2004) feels that "estimating parameters of models that do not mirror substantive causal processes can only rarely be expected to yield valuable insights into causal process." Directly, this looks so, indeed. Indirectly, something might be scavenged. This applies foremost to quantities with a natural zero, implying a ratio scale. Conversion is more difficult for opinion polls and ratings, where the zero is set arbitrarily. The importance of this difference will emerge in Chapter 17.

The following example from recent literature shows that sometimes descriptive analysis already does include the ingredients for a predictive model. Here constants of the predictive model can be inferred from the regression coefficients, and the outcome casts new light on the meaning of these regression coefficients. The example also shows that a predictive model of a multiplicative format can sometimes be constructed even