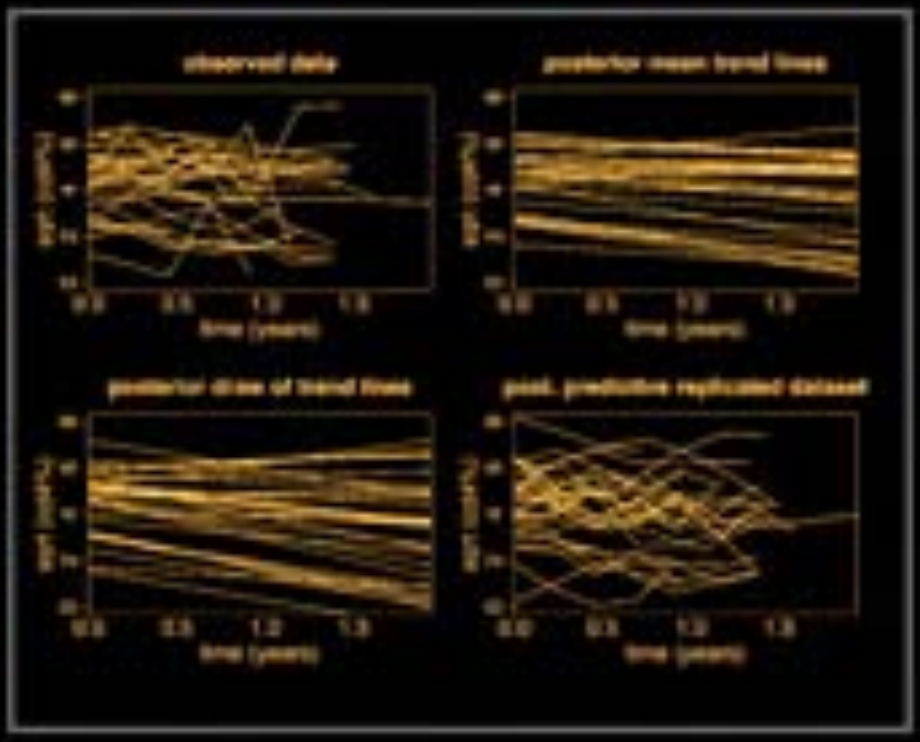
ANALYTICAL METHODS FOR SOCIAL RESEARCH



Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN
JENNIFER HILL

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521867061

APPENDIX A

Six quick tips to improve your regression modeling

A.1 Fit many models

Think of a series of models, starting with the too-simple and continuing through to the hopelessly messy. Generally it's a good idea to start simple. Or start complex if you'd like, but prepare to quickly drop things out and move to the simpler model to help understand what's going on. Working with simple models is not a research goal—in the problems we work on, we usually find complicated models more believable—but rather a technique to help understand the fitting process.

A corollary of this principle is the need to be able to fit models relatively quickly. Realistically, you don't know what model you want to be fitting, so it's rarely a good idea to run the computer overnight fitting a single model. At least, wait until you've developed some understanding by fitting many models.

A.2 Do a little work to make your computations faster and more reliable

This sounds like computational advice but is really about statistics: if you can fit models faster, you can fit more models and better understand both data and model. But getting the model to run faster often has some startup cost, either in data preparation or in model complexity.

Data subsetting

Related to the "multiple model" approach are simple approximations that speed the computations. Computers are getting faster and faster—but models are getting more and more complicated! And so these general tricks might remain important. A simple and general trick is to break the data into subsets and analyze each subset separately. For example, break the 85 counties of radon data randomly into three sets of 30, 30, and 25 counties, and analyze each set separately.

The *advantage* of working with data subsets is that computation is faster on data subsets, for two reasons: first, the total data size n is smaller, so each regression computation is faster; and, second, the number of groups J is smaller, so there are fewer parameters, and the Gibbs sampling requires fewer updates per iteration.

The two *disadvantages* of working with data subsets are: first, the simple inconvenience of subsetting and performing separate analyses; and, second, the separate analyses are not as accurate as would be obtained by putting all the data together in a single analysis. If computation were not an issue, we would like to include all the data, not just a subset, in our fitting.

In practice, when the number of groups is large, it can be reasonable to perform an analysis on just one random subset, for example one-tenth of the data, and inferences about the quantities of interest might be precise enough for practical purposes.

Redundant parameterization

Sections 19.4–19.5 discuss redundant additive and multiplicative parameterizations. These steps add extra parameters to a Bugs model, and can be confusing at first, but can really pay off in speed of computation. In addition, the recentering and scaling required in defining the adjusted parameters can have a convenient statistical interpretation in terms of finite-population inference for the groups in the dataset.

Fake-data and predictive simulation

When computations get stuck, or a model does not fit the data, it is usually not clear at first if this is a problem with the data, the model, or the computation. Fake-data and predictive simulation (discussed in general in Chapter 8 and for multilevel models in Sections 16.7 and 24.1–24.2) are effective ways of diagnosing problems. First use fake-data simulation to check that your computer program does what it is supposed to do, then use predictive simulation to compare the data to the fitted model's predictions.

A.3 Graphing the relevant and not the irrelevant

Graphing the fitted model

Graphing the data is fine (see Appendix B) but it is also useful to graph the estimated model itself (see lots of examples of regression lines and curves throughout this book). A table of regression coefficients does not give you the same sense as graphs of the model. This point should seem obvious but can be obscured in statistical textbooks that focus so strongly on plots for raw data and for regression diagnostics, forgetting the simple plots that help us understand a model.

Don't graph the irrelevant

Are you sure you really want to make those quantile-quantile plots, influence diagrams, and all the other things that spew out of a statistical regression package? What are you going to do with all that? Just forget about it and focus on something more important. A quick rule: any graph you show, be prepared to explain.

A.4 Transformations

Consider transforming every variable in sight:

- Logarithms of all-positive variables (primarily because this leads to multiplicative models on the original scale, which often makes sense)
- Standardizing based on the scale or potential range of the data (so that coefficients can be more directly interpreted and scaled); an alternative is to present coefficients in scaled and unscaled forms
- Transforming before multilevel modeling (thus attempting to make coefficients more comparable, thus allowing more effective second-level regressions, which in turn improve partial pooling).

Plots of raw data and residuals can also be informative when considering transformations (as with the log transformation for arsenic levels in Section 5.6).

In addition to univariate transformations, consider interactions and predictors created by combining inputs (for example, adding several related survey responses

to create a "total score"). The goal is to create models that *could* make sense (and can then be fit and compared to data) and that include all relevant information.

A.5 Consider all coefficients as potentially varying

Don't get hung up on whether a coefficient "should" vary by group. Just allow it to vary in the model, and then, if the estimated scale of variation is small (as with the varying slopes for the radon model in Section 13.1), maybe you can ignore it if that would be more convenient.

Practical concerns sometimes limit the feasible complexity of a model—for example, we might fit a varying-intercept model first, then allow slopes to vary, then add group-level predictors, and so forth. Generally, however, it is only the difficulties of fitting and, especially, understanding the models that keeps us from adding even more complexity, more varying coefficients, and more interactions.

A.6 Estimate causal inferences in a targeted way, not as a byproduct of a large regression

Don't assume that a regression coefficient can be interpreted causally. If you are interested in causal inference, consider your treatment variable carefully and use the tools of Chapters 9, 10, and 23 to address the difficulties of comparing comparable units to estimate a treatment effect and its variation across the population. It can be tempting to set up a single large regression to answer several causal questions at once; however, in observational settings (including experiments in which certain conditions of interest are observational), this is not appropriate, as we discuss at the end of Chapter 9.

APPENDIX B

Statistical graphics for research and presentation

Statistical graphics are sometimes summarized as "exploratory data analysis" or "presentation" or "data display." But these only capture part of the story. Graphs are a way to communicate graphical and spatial information to ourselves and others. Long before worrying about how to convince others, you first have to understand what's happening yourself.

Why to graph

Going back through the dozens of examples in this book, what are our motivations for graphing data and fitted models? Ultimately, the goal is communication (to self or others). More immediately, graphs are comparisons (to zero, to other graphs, to horizontal lines, and so forth). We "read" a graph both by pulling out the expected (for example, the slope of a fitted regression line, the comparisons of a series of confidence intervals to zero and each other) and the unexpected.

In our experience, the unexpected is usually not an "outlier" or aberrant point but rather a systematic pattern in some part of the data. For example, consider the binned residual plots in Section 5.6 for the well-switching models. There was an unexpectedly low rate of switching from wells that were just barely over the dangerous level for arsenic, possibly suggesting that people were moderating their decisions when in this ambiguous zone, or that there was other information not included in the model that could explain these decisions.

Often the most effective graphs simply show us what a fitted model is doing. Consider, for example, the graphs in Section 6.5 of the ordered regression and the data for the storable voting experiment or in Section 14.1 of the data-level logistic model and state-level linear model for political opinions.

We consider three uses of graphics in statistical analysis:

1. Displays of raw data, often called "exploratory analysis." These don't have to look pretty; the goal is to see things you did not expect or even know to look for.

2. Graphs of fitted models and inferences, sometimes overlaying data plots in order to understand model fit, sometimes structuring or summarizing inference for many parameters to see a larger pattern. In addition, we can plot simulations of replicated data from fitted models and compare them to comparable plots of raw data.

3. Graphs presenting your final results—a communication tool. Often your most important audience here is yourself—in presenting all of your results clearly on the page, you'll suddenly understand the big picture.

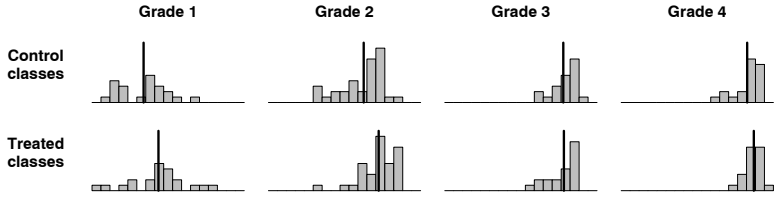


Figure B.1 *Data from the Electric Company experiment, from Figure 9.4 on page 174, displayed in a different orientation to allow easier comparison between treated and control groups in each grade. For each histogram, the average is indicated by a vertical line.*

B.1 Reformulating a graph by focusing on comparisons

Creative thinking might be needed to display numerical data effectively, but your creativity can sometimes be enhanced by carefully considering your goals. Just as in writing, you have to rearrange your sentences sometimes to make yourself clear. For example, consider the graph of the Electric Company data in Figure 9.4 on page 174. Rather than try to cleverly put all the points on a single plot, we arrange them on a 4×2 grid, using a common scale for all the graphs to facilitate comparisons among grades and between treatment and control. We also extend the axis all the way to zero, which is not strictly necessary, in the interest of clarity of presentation. In the Electric Company example, as in many others, we are not concerned with the exact counts in the histogram; thus, we simplify the display by eliminating y-axes, and we similarly clarify the x-axis by removing tick marks and using minimal labeling.

Graphs as comparisons

All graphical displays can be considered as comparisons. When making a graph, line things up so that the most important comparisons are clearest. Comparisons are clearest when scales are lined up. Considering Figure 9.4: for each of the two treatments, the histograms for the four grades are lined up and can be directly compared.

In Figure 9.4, we primarily want to compare treatment to control. The comparison of grades is useful—if for no other reason than to ground ourselves and confirm that scores are higher in the higher grades—but we are really more interested in the comparison of treatment to control within each grade.

Thus, it might be more helpful to arrange the histograms as shown in Figure B.1, with treatment and control aligned for each grade. With four histograms arranged horizontally on a page, we need to save some space and so we restrict the x-axes to the combined range of the data. We also indicate the average value in each group with a vertical line to allow easier comparisons of control to treatment in each grade.

No single graph does it all

Sometimes it makes sense to withhold information in order to present a clearer picture. Figure 9.4 (or Figure B.1) shows the outcomes for each classroom in the Electric Company experiment. The scatterplots in Figure 9.6 show pre-test data

as well, revealing a high correlation between pre-test and post-test in each grade. The scatterplots certainly show important information, and we are glad to be able to show them, but we prefer the histograms as a starting point for seeing the comparison between treatment and control—at least for this randomized experiment in which the two groups are well balanced.

Graphs of fitted models

It can be helpful to graph a fitted model and data on the same plot, as we have done throughout the book. See Chapters 3–5 for many simple examples, Figure 6.3 on page 120 for a more elaborate example, and Chapters 12–13 for similar plots of multilevel models.

We also like to graph sets of estimated parameters (see, for example, in Figure 4.6 on page 74). Graphs of parameter estimates can be thought of as proto-multilevel models in that the graph suggests a relation between the y-axis (the parameter estimates being displayed) and the x-axis (often time, or some other index of the different data subsets being fit by a model). These graphs contain an implicit model, or a comparison to an implicit model, the same way that any scatterplot contains the seed of a regression or correlation model.

Another use of graphics with fitted models is to plot predicted datasets and compare them visually to actual data, as discussed in Sections 8.3–8.4. For data structures more complicated than simple exchangeable batches or time series, plots can be tailored to specific aspects of the models being checked, as in Section 24.2. As a special case, plots of residuals and binned residuals can be seen as visual comparisons to the hypothesis that the errors from a model are independent with zero mean.

B.2 Scatterplots

Units

When describing or designing a scatterplot, the first thing to decide is the unit of analysis. That is "each dot represents a student" or "each dot represents a county" or whatever. The x and y values have no interpretation until you define the units.

The x and y axes

To get yourself up to speed, start by applying to scatterplots everything you know about linear regression. There's an x variable and a y variable defined on a bunch of units, and you're trying to summarize the average relation between x and y or alternatively to predict y from x where "prediction" includes uncertainty as well as point estimation. This issue is well covered in many recent introductory textbooks which introduce scatterplots first and then move to regression.

Let's start with some bad ideas. First, there is something called a scatterplot matrix for multivariate data, which is a set of scatterplots of all pairs of variables. This can be informative, but it's like regressing every variable versus every other variable. As with regression, we often learn more from scatterplots that are more carefully chosen. For example, if two variables have a time or causal order, we usually prefer to put "before" on the x-axis and "after" on the y-axis.

A common strategy that particularly disturbs us is plotting by index number, for example, plotting data from the 50 states in alphabetical order. In this case the x variable contains little or no information, and the plot is comparable to running

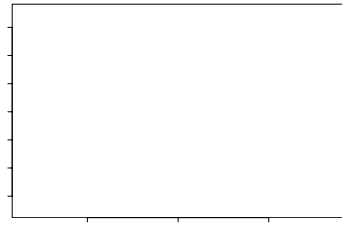


Figure B.2 *Length of longest run (sequence of successive heads or successive tails) versus number of runs (sequences of heads or tails) in each of 2000 independent simulations of 100 coin flips. Each dot on the graph represents a sequence of 100 coin flips; the points are jittered so they do not overlap. When plotted on this graph, the results from an actual sequence of 100 coin flips will most likely fall on a square with a large number of dots. In contrast, a sequence of heads and tails that is artificially created to look "random" will probably have too many runs that are not long enough, and hence will fall on the lower right of this graph.*

a regression on random numbers. An example that is not *necessarily* bad is using, as the x variable, the order of entry of units into the study. This can make sense if one expects or fears time trends (but it would probably be better to plot versus time itself rather than merely order). If there are no major time patterns, however, the choice of x variable might better be spent elsewhere.

You can make as many plots as you want (or as your paper budget allows), but it is useful to think a bit about each plot, just as it is useful to think a bit about each regression you run. This is as good a time as any to recommend that along with every regression you run, you should make a scatterplot. And, in addition, you should be making residual plots where necessary. We'll get to that later.

Jittering

If several data points have the same data values, add a small random number to each so that they do not fall on top of each other. This is called jittering. Jitter just enough so that the discrete nature of the data is still clear. For example, if data points are integers, we might add a random uniform number between -0.3 and $+0.3$ to each x and y value (see Figure B.2). Methods such as plotting 2's, 3's, or cute symbols for multiple data points can be misleading visually, and from a theoretical perspective are unsatisfying in that the display of any unit then depends too strongly on the other data values.

Symbols and auxiliary lines

The symbols of a scatterplot are important because they correspond to the units of analysis in your studies. It can be appropriate to use more than one scatterplot for multilevel data structures. At least in theory you can display five variables easily with a scatterplot: x, y, symbol, symbol size, and symbol color.

Symbols are best for discrete variables, and it's worth putting a little effort into making these symbols distinguishable and also appropriate. For example, we used open circles to indicate open seats in Figure 7.4. In plotting data from an experiment or observational study, you can use different large symbols for treated units and

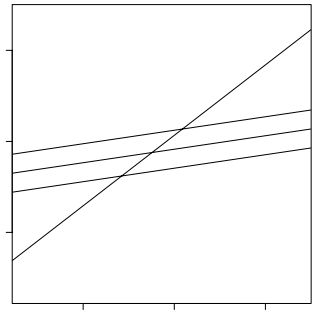


Figure B.3 *Effect of redistricting on partisan bias. Each symbol represents a state and election year, solid circles, open circles, and crosses representing Democratic, bipartisan, and Republican redistricting. The small dots are the control cases—state-years that did not immediately follow a redistricting. Lines show fit from a regression model.*

dots for controls (see Figure B.3). Symbol size can be useful, but it is not always as flexible as one might hope, and we have not had much success in using symbol size for continuous variables.

Color is just great and you should use it as much as possible, even though for printing reasons we do not use color in this book.

We sometimes have had success using descriptive symbol names (for example, two-letter state abbreviations). But if there are only two or three categories, we're happier with visually distinct symbols. For example, to distinguish men and women, we would not use M and W or even M and F. In genealogical charts, men and women are often indicated by open squares and open circles, respectively, but even these symbols are hard to tell apart in a group. We prefer clearly distinguishable symbols—for example, in Figure B.5, open circles for men and solid circles for women.

These suggestions are all based on our subjective experience and attempts at logical reasoning; as far as we know, they have not been validated (or disproved) in any systematic study. We think such a study would be a good idea.

Figure B.3 shows an example of one of the most common regressions: a comparison of treatments to control with a before and after measurement. In this case, the units are state legislative elections, and the plot displays a measure of "partisan bias" in two successive election years. The "treatments" are different kinds of redistricting plans, and the "control" points (indicated by dots on the figure) indicate pairs of elections with no intervening redistricting. We display all the data and also show the regression lines on the same scale. As a matter of fact, we did not at first think of fitting nonparallel regression lines; it was only after making the figure and displaying parallel lines that we realized that nonparallel lines (that is, an interaction between the treatment and the "before" measurement) are appropriate. The interaction is, in fact, crucial to the interpretation of these data: (1) when there is no redistricting, partisan bias is not systematically changed; (2) the largest effect of any kind of redistricting is to bring partisan bias, on average, to near zero. The lines and points together show this much more clearly than any numerical summary.

Another useful kind of line to display is a "default line," which is usually a horizontal line at 0 or a 45-degree line indicating equality of x and y.

When a graph has multiple lines, label them directly, not using symbol codes and a key (which requires the reader—and you—to go back and forth between graph and key). Examples of our recommended approach include Figure 5.11 on page 91, Figure 14.11 on page 313, and Figure 15.2 on page 328.

Shape of the plotting region

The shape of a plot conveys information implicitly. When x and y are the same units on the same scale, we use a square plot with the same scale on the two axes even if that means that large parts of the plot are blank (see Figure B.3). Conversely, if x and y are not the same variable, we are careful *not* to use a square plot so as not to implicitly send the wrong message. When we are presenting several plots of different variables, we sometimes use dimensions for the different plots as a visual cue that they have different meanings.

Displaying the results of model fitting

In a regression with one or two inputs, it is possible to display essentially all the information (all the information if one of the variables is discrete) in a single plot. When additional predictors are present, we have to summarize the data in some way. Ideally, the outcome variable is displayed on the y-axis, symbols indicate the input variable of interest (think of treatments and control here), and the x-axis displays predicted values or some other combination of all the variables that are being controlled for.

When there is more than one control variable, one approach is to plot on the x-axis the linear predictor created from all the control variables with coefficients estimated from their regression models. For example, with a regression model of the form $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$, one can plot y_i versus $\beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3}$ with different symbols for different values of X_{i1}. In that plot one would plot dotted lines of $y = c + x$, for $c = \beta_1 x_1$ for the different values for x_1, to illustrate the expected relationship. Figure B.3 shows an example with one predictor that plays the role of "treatment" and other "background" predictors which are combined in the x-axis.

More generally we can overlay the model on a plot of data (conversely when plotting a modeled relationship, we try to include data on this plot appropriately), even if it takes a bit of work to figure out how to do this reasonably. In our own work such plots have been crucial to our understanding, as illustrated by Figure B.3.

Maps

Often when you have a map, you're better off with a scatterplot (but of course there's no reason to throw away the map). For example, if you have data on the occurrence of some medical condition by location and you map it to see whether it's clustered in low-income areas, it might make more sense to plot rates versus income. But the map might be useful in suggesting which variables to consider plotting.

With this use of maps as an explanatory tool in mind, we focus on mapping methods that will reveal unexpected patterns but only when something real is

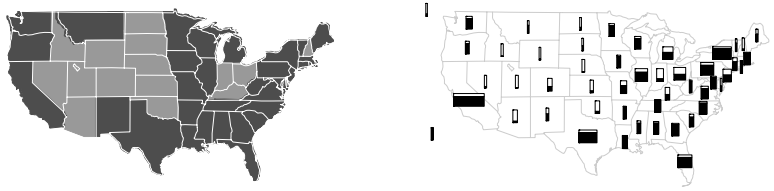


Figure B.4 *Summary of a forecast of the 1992 U.S. presidential election performed one month before the election. (a) States that Bill Clinton was forecasted to win are shaded. (b) For each state, the proportion of the box that is shaded represents the probability of Clinton winning the state; the width of the box is proportional to the number of electoral votes for the state. The second map conveys more information and is also less misleading.*

going on. Maps are often tricky to read because they can show spurious patterns. For example, a map of the United States shading in different counties with different colors inevitably draws attention to the counties that are geographically larger and perhaps also those that are unusually shaped. At the very least one could replace the shading by a small colored circle in each county, perhaps with larger circles for more populous counties. (However, this would not be appropriate for a geological map of oil reserves: we are usually thinking about social statistics here.) Another approach is to plot "thermometers" within a geographic unit (see Figure B.4).

The problem of unequal population density is sometimes attacked by distorted maps that approximately preserve the shapes of, for example, states, while making their areas proportional to population. We find these maps more distracting than useful because they draw attention to the shapes, which are usually nothing that anybody cares about.

In addition to any possible distorted geographical effects, there are more subtle difficulties in mapping which relate to problems of summarizing inferences with point estimates (see, for example, Gelman and Price, 1999).

Calibration plots

A calibration plot is a plot of observed values on the y-axis versus expected (forecasted) values on the x-axis. If all is well, the expected value of y given x in such a plot is just x. So we make this a square plot with identical axes and a comparison line at $y = x$. See, for example, Figure B.5, which evaluates the calibration of students' guesses of their exam scores.

In general, a forecast supplies a distribution, not just a point estimate, for each data point. In this case, the "expected" or "forecasted" value for any datum is just the mean (or expectation) of the forecast distribution for the datum. The desired relation is $\mathrm{E}(y|x) = x$.

When forecasting discrete outcomes, however, the problem gets more complicated: the expected values are continuous but the observed values are discrete (for example, for binary data, the observed values are 0's and 1's, and the expected values are proportions between 0 and 1). The calibration plot is then virtually unreadable, as the points cluster in discrete values on the y-axis. (See Figure B.6a for an example.) So instead, it is standard practice to order the x values and then divide them into categories or bins $j = 1, \ldots, J$. In each category we compute the averages $\bar{x}_j$ and $\bar{y}_j$ and then plot the J values of $(\bar{x}_j, \bar{y}_j)$. Figure B.6b shows an example in which the data can take on 5 possible outcomes.

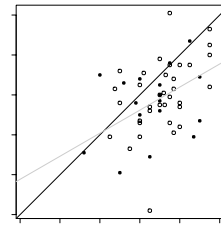


Figure B.5 *Actual versus guessed midterm exam scores for a class of 53 students. Each symbol represents a student; empty circles are men, solid circles are women, and ? has unknown sex. The* 45° *line represents perfect guessing, and the dotted line is the linear regression of actual score on guessed score. (The separate regression lines for men and women were similar.) Both men and women tended to perform worse than their guesses. That the slope of the regression line is less than 1 is an instance of the "regression effect" (see Section 4.3): if a student's guessed score is* x *points higher than the mean guess, then his or her actual score is, on average, only about* $0.6x$ *higher than the mean score. A square scatterplot is used because the horizontal and vertical axes are on the same scale.*

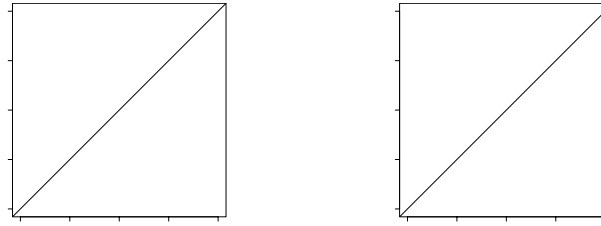


Figure B.6 *(a) Observed versus expected pain relief scores (0 = no pain relief, . . ., 5 = complete pain relief) for data from the analysis of Sheiner, Beal, and Dunne (1997). Observed pain relief scores are jittered. (b) Average observed versus averaged expected pain relief scores, with data divided into 20 equally sized bins defined by ranges of expected pain relief scores.*

Whether in the continuous or discrete case, we prefer to put "observed" on the y-axis and "expected" on the x-axis (rather than the reverse), because in the calibration context, the expected value is the predictor and the observed value is the outcome. See Section 8.2 for related discussion of residual plots.

Residual plots

If all is going well, the points on the calibration plot will mostly fall near the 45-degree line, meaning there will be much empty space on the plot. A natural next step is to plot $y-x$ versus x; that is, "deviation from predicted" versus "predicted." This is the residual plot. In fact "deviation from predicted" can be plotted versus just about anything, not just predicted values (see Figure B.7). Residual plots should not be square and should have a dotted line at $y = 0$ rather than $y = x$.

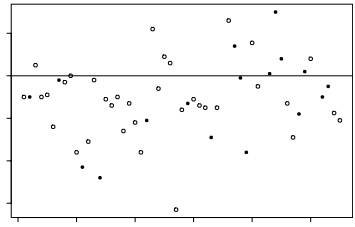


Figure B.7 *Difference between actual and guessed midterm exam scores, plotted against the order of finishing the exam. The exact order is only relevant for the first 20 or 25 students, who finished early; the others all finished within five minutes of each other at the end of the class period. Each symbol represents a student; empty circles are men, solid circles are women, and ? has unknown sex. The horizontal line represents perfect guessing. The students who finished early were highly overconfident, whereas the other students were less biased in their predictions.*

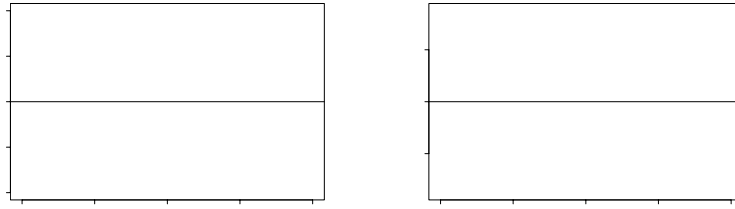


Figure B.8 *(a) Residuals (observed - expected) of pain relief scores versus expected pain relief scores from Figure B.6. (b) Average residuals versus expected pain scores, with measurements divided into 20 equally sized bins defined by ranges of expected pain scores. The average prediction errors are relatively small (as can be seen from the scale of the y-axis), but with a consistent pattern that low predictions are too low and high predictions are too high.*

As with calibration plots, it is generally a good idea to bin the points in a residual plot if the outcomes are discrete (see Figure B.8).

B.3 Miscellaneous tips

We conclude with some suggestions derived from our experiences using graphs in data analysis, first presenting a few ideas that have proved generally useful, then going through a variety of specific techniques through a series of examples.

A display of several time series of opinion polls

Each subgraph of Figure B.9 shows a time series of the support in the polls for the Republican candidate for U.S. president, as a proportion of the two-party support, for a given election year, in the months leading up to the election.

Tip: Put many little graphs on the same page. Do it with a slick graphics package if possible; otherwise, use scissors, tape, and a reducing copy.

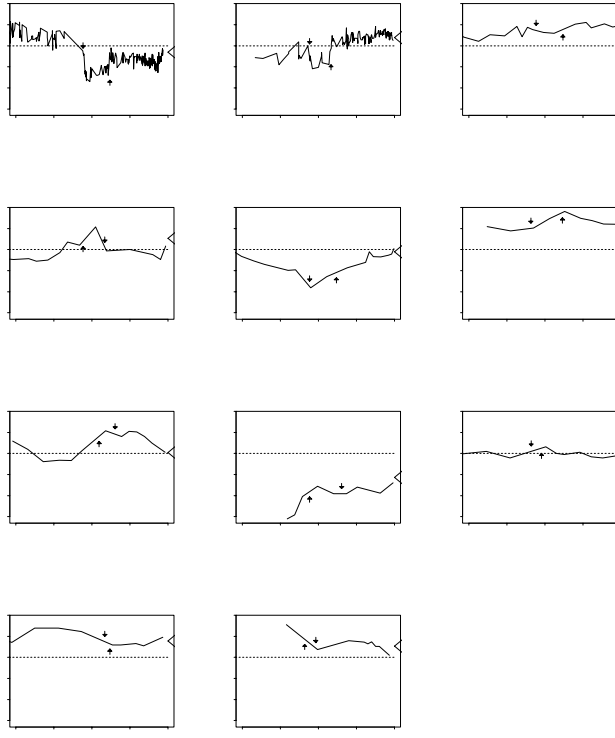


Figure B.9 *Presidential trial-heat polls. The solid line in each plot is the proportion who would vote for the Republican candidate for president, among those who report a preference for the Democratic or Republican candidates. The 1992 and 1998 graphs include data from all available nationwide polls; plots for the other years are from the Gallup Report. The upward arrow marks the time of the Republican convention, and the downward arrow marks the time of the Democratic convention. The triangle at the end of each time series indicates what actually happened in the election.*

Tip: When you have multiple graphs, use a common scale.

Tip: Put a light line to indicate what "no effect" would be. (There is a dotted line at 50% in each graph.)

Tip: It's worth putting in little details and doing it right. For example, each graph also indicates, with arrows, the times of the political conventions. The Republican conventions are shown with up arrows (because the Republicans improve in the polls then), and the Democratic conventions are indicated with down arrows (corresponding to the drop in the Republican poll numbers).

Tip: Keep the lines on a graph thin, even if each plot has only one line. A fat line conveys no more information and just makes the information harder to see.

By comparison, we got the data from printed reports from Gallup that had graphs like ours for each election year, but with two thick lines on each graph displaying the Democratic and the Republican shares of the polls. For our purposes, we didn't care about undecideds and third parties, so we just display the Republican proportion of the two-party support.

Tip: Repeat axis labels as necessary to make mini-graphs easier to read. Once you know what they say, your eye easily ignores the labels.

We originally created this graph to help us understand the history of the pre-election polls at a glance—exploratory data analysis—and later we fixed it up for final presentation. (In the original, exploratory, stage, we wrote in the arrows by hand.)

Significant digits and uncertainty

When reporting the output from a statistical analysis, you should always imagine yourself in the position of the reader of the report. It is important not to overwhelm the reader with irrelevant material. For the simplest (but still important) example, consider the reporting of numerical results (either alone or in tables).

Do not include too many significant digits in numbers you report. The relevant comparison is not to an absolute number of decimal places but to the uncertainty and variability in the numbers being presented. For example, the confidence interval [3.276, 6.410] would be more clearly written as [3.3, 6.4]. (An exception is that it makes sense to save lots of extra digits for intermediate steps in computations. For example, $51.7643 - 51.7581$.) A related issue is that you can often make a list or table of numbers more clear by first subtracting out the average (or for a table, row and column averages). The appropriate number of significant digits depends on the uncertainty. But in practice, three digits are usually enough because if more were necessary, we would subtract out the mean first.

Maybe the biggest source of too many significant digits is from computer output. One solution is to set the rounding in the computer program (for example in R, `options(digits=2)`).

Titles and captions

All titles and axis labels should be meaningful. In addition, each figure should be accompanied by a caption so that it makes sense even for the reader who skips the rest of the article.

Histograms

Histograms are for plotting values of a single variable. Whenever possible, use a scatterplot, but sometimes it is convenient look at just one variable, especially when arranged in a grid such as in Figure B.1 on page 552. When looking at one variable, we prefer histograms to snazzier methods such as density estimation because we feel more connected to the actual numerical values this way.

There's some confusion on this point. The purpose of a histogram is to display a set of numbers, not to approximate an underlying distribution function. It's a good idea to divide your histogram into more bins than "necessary" so that you can get

Figure B.10 *Histograms of the forecast proportion of the two-party vote for Bill Clinton in 1992 in each of the 50 states and the District of Columbia, displayed with two different choices of bin width: (a) the bin width automatically assigned by R, (b) the bin width set manually with the R command* `hist(y,breaks=seq(30,90,5))`.

an idea of the variability in the histogram itself. Do not use the default bin width in R (see Figure B.10).

General advice

Plot numerical data and inferences as graphs, not as tables. A good example is the multilevel logistic regression of vote preference on demographic and geographic predictors, with graphs on pages 306–307 that show coefficient estimates and standard errors, along with curves of the fitted model and data. Or, for a simpler example, Figure 15.9 on page 337 graphs the inference from a simple regression.

Multiple plots per page. A graph can almost always be made smaller than you think and still be readable. This then leaves room for more plots on a grid, which then allows more patterns to be seen at once and compared.

Don't plot the index numbers. For example, Figure 14.9 on 312 plots estimates for the 50 states versus average state income, rather than simply listing the states in alphabetical order. For another example, the dogs in Figure 24.1 are ordered by the time of their last shock, rather than by their ID numbers, which turn out to have no meaning in this problem.

Never display a graph you can't explain. Give a full caption for every graph (as we try to do in this book). This explains to yourself and others what you are trying to show and what you have learned from each plot. Avoid displaying graphs that have been made simply because they are conventional. For example, regressions are commonly equipped with quantile-quantile plots of residuals, but for most applications the information in such a plot is irrelevant, and a distraction from the more relevant results that could be presented.

B.4 Bibliographic note

For statistical graphics in R, the book by Murrell (2005) is an excellent overview and starting point. Fox (2002) is also helpful in that it focuses on regression models.

On the topic of statistical graphics more generally, much of the most important and influential work has appeared in books, including Bertin (1967, 1983), Chambers et al. (1983), Cleveland (1985, 1993), Tufte (1983, 1990), and Wainer (1984, 1997).

There are various systematic ways of studying statistical graphics. One useful approach is to interpret graphs as model checking (for example, if residuals are not independent of x, then there is some model violation), as we have discussed in Chapter 24. Another approach is to perform experiments to find out how well

people can gather information from various graphical displays (for example, are line plots easier to read than histograms). This is discussed by Cleveland (1985). More research is needed on both these approaches: relating to probability models is important for allowing us to understand graphs and devise graphs for new problems; and effective display is important for communicating to ourselves as well as others.

For some ideas on the connections between statistical theory, modeling, and graphics, see Tukey (1977), Wilkinson (2005), and (for our own perspective) Gelman (2004a).

Some of the ideas considered in this chapter are explored by Gelman, Pasarica, and Dodhia (2002), Wand (1997), Wainer (2001), and Friendly and Kwan (2003). Ehrenberg (1978) and Tukey (1977) discuss tabular displays in detail. An important topic not discussed in the present book is dynamic graphics; see Buja et al. (1988) and Buja, Cook, and Swayne (1999).

B.5 Exercises

1. Find an example of a published article in a statistics or social science journal in which too many significant digits are used.
2. Find an example of a published article in a statistics or social science journal in which there is *not* a problem with too many significant digits being used.
3. Take any data analysis exercise from this book and present the *raw data* in several different ways. Discuss the advantages and disadvantages of each presentation.
4. Take any data analysis exercise from this book and present the *fitted model* in several different ways. Discuss the advantages and disadvantages of each presentation.