Teaching Statistics

A Bag of Tricks

Andrew Gelman
Columbia University

Deborah Nolan
University of California, Berkeley

OXFORD
UNIVERSITY PRESS

4

Linear regression and correlation

We follow descriptive statistics in our course with a descriptive treatment of linear regression with a single predictor: straight-line fitting, interpretation of the regression line and standard deviation, the confusing phenomenon of "regression to the mean," correlation, and conducting regressions on the computer. We illustrate various of these concepts with student discussions and activities.

This chapter includes examples of the sort that are commonly found in statistics textbooks, but our focus is on how to work them into student-participation activities rather than simply examples to be read or shown on the blackboard.

4.1 Fitting linear regressions

We have found that students have difficulty with the algebra of linear equations; for example, it is far from intuitive that the line going through the point $(\bar{x}, \bar{y})$ and with slope b has the equation, $y = \bar{y} + b(x - \bar{x})$. Thus, when introducing linear regression, we start with algebraic exercises with straight lines in various examples.

In this section we present some simple drills illustrating the mathematics of least-squares regression and then discuss two examples where least-squares fits are informative but not quite ideal for the data at hand. In presenting the examples, we focus first on the mechanics of the model—in particular, on the interpretation of the fitted line. We then introduce the concept of residual standard deviation, that not all points will fall exactly on the line. Finally, we look more carefully at the data and discuss with the students the limitations of the least-squares fits.

4.1.1 Simple examples of least squares

We use simple scatterplots to introduce straight-line fitting. We bring to class a handout of four scatterplots (Fig. 4.1), divide the students into pairs, and ask the students in each pair to draw the best-fitting line for predicting y from x from the data on each plot. We don't go into the exact definition of a best fit yet—we just tell them that we want to predict y from x and let them loose on the problem. Once they have drawn their lines, we have them compute their predictions for each y_i and the sum of squared errors. Then we pass out red pencils and ask each pair to mark a red × at the average y-value for each observed x (see Fig.

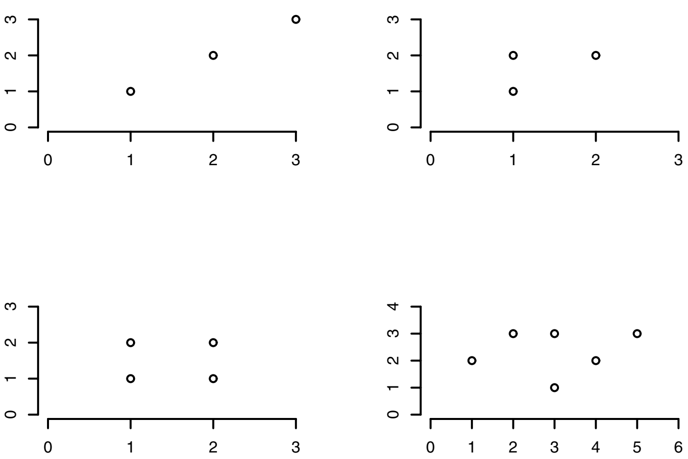

Fig. 4.1 We present these simple scatterplots to students to teach least-squares fitting of a line to points. Students sketch on each plot what they think is the best-fitting line for predicting y from x. The example is continued in Fig. 4.2.

4.2). Then they find the best-fitting line to the crosses and compute the sum of squared errors for predicting y_i from the red line.

For a couple of the scatterplots, the placement of the points suggests the wrong line to the naive line-fitter. The students are surprised to find the red line outperforms their line in these cases, and at this point we begin the discussion of minimizing square error in the y-direction.

4.1.2 Tall people have higher incomes

To demonstrate linear regression, many examples are possible. We use a regression of income on height, because it has a story with lurking variables to which we can return in the discussion of multiple regression at the end of the semester (see Section 9.1). Our focus here is on the interaction with the students more than the example itself.

Before class begins, we set up Fig. 4.3 on a transparency and project it onto the blackboard. We trace the lines and label the axes of the graph, then turn off the projector, but leave it in place.

We begin the discussion by asking students if they think that taller people have higher earnings (that is, income excluding unearned sources such as interest income). If so, by how much? We draw on the blackboard a pair of axes representing earnings and height, and a point at $(66.5, 20\,000)$: the average height of adults in the United States is about $5'6.5''$ and their average earnings (in 1990) were about \$20 000. We then draw a line through this central point with slope

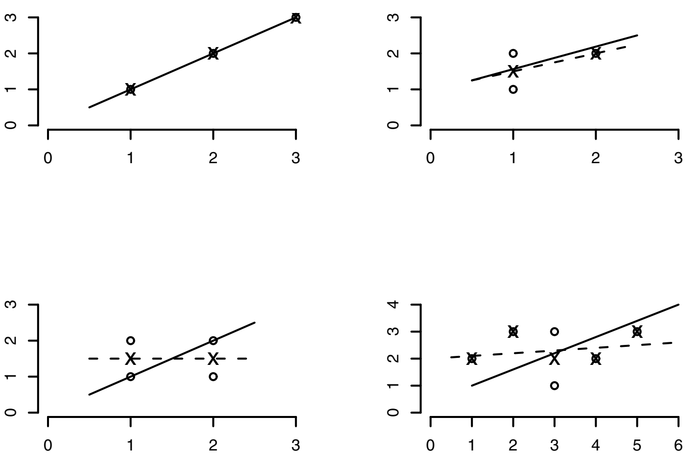

Fig. 4.2 The solid line in each plot (from Fig. 4.1) shows a student's guess of the best-fitting line to the set of points. The student then marks crosses at the average y-value for each x and draws the best-fitting line to the crosses. In some cases these lines are quite different, which provides a good lead-in to the method of least squares.

1560, carefully connecting the central point $(66.5, 20\,000)$ and two points on the regression line: $(56.5,\ 20\,000 - 10 \cdot 1560)$ and $(76.5,\ 2000 + 10 \cdot 1560))$. We pick points ±10 inches from the central point just for convenience in calculation and display.

We explain that this line has equation $y - 20\,000 = 1560(x - 66.5)$, or $y = -84\,000 + 1560x$. We tell the students that this is the regression line predicting earnings from height. We now ask the students to work in pairs and sketch a scatterplot of data that are consistent with this regression line. They need one more piece of information: the standard deviation of the residuals, which is 19 000. We show this on the graph by two dotted lines, parallel to the regression line, with one line 19 000 above and the other line an equal distance below. Approximately 68% of the data should fall in this region, but plotting the points is tricky because of the constraint that earnings cannot be negative.

We then turn on the projector and display the graph of the actual survey data (Fig. 4.3) on the blackboard. (The survey may have its own problems of response and measurement error, but that is not our point here.) On the scale of the actual data, the regression slope ($1560 per inch of height) is small but undeniably positive. (We discuss statistical significance later in the course; see Section 9.1.1.)

We then ask the students how do we interpret the constant term in the regression: that is, the value $-84\,000$ in the equation, $y = -84\,000 + 1560x$ (see

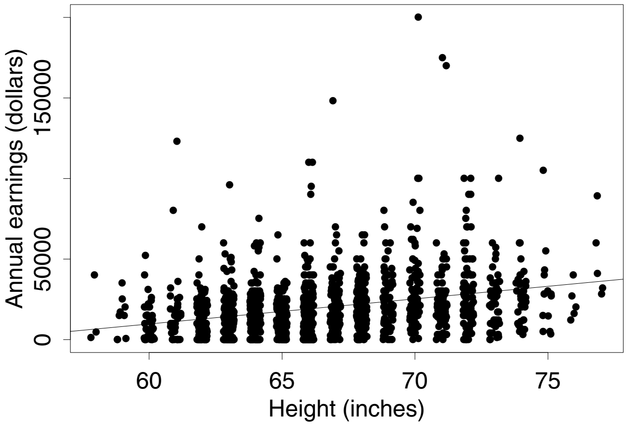

Fig. 4.3 Earnings vs. height for a random sample of adult Americans in 1990. The heights have been jittered slightly so that the points do not overlap.

the regression table at the bottom of Fig. 4.4)? The answer is, $-84\,000$ is the y-value of the regression line where $x = 0$—that is, the predicted value of income for an adult who is zero inches tall. In this example, such an extrapolation is meaningless. That is why we prefer to work with the form, $y = \bar{y} + b(x - \bar{x})$; in this case, $y = 20\,000 + 1560(x - 66.5)$.

The next question is how to interpret the result that taller people have higher earnings? The students realize that men are taller than women and tend to make more money; thus, sex is a lurking variable. We return to this example using multiple regression in Section 9.1. (In fact, it turns out that height is correlated with earnings even after controlling for sex.)

At the conclusion of the discussion, we hand out copies of Fig. 4.4, which shows what we had to do to clean the data file before running the regression. They can also use this as a template when doing their computer homework assignments. We explain to the students that linear regression is not the best model for this sort of data (economists might use a logarithmic model, or a tobit regression), but it is in some ways more useful to illustrate the concept in an example for which it is not completely appropriate.

4.1.3 Logarithm of world population

We return to the world population data (see Section 3.8.2) to illustrate linear modeling on a transformed scale. It is worth going through these calculations of linear regression predictions and errors on the logarithmic scale. We have found students to have great difficulty with this sort of problem on exams unless they have practiced it a lot.

A good way to understand the logarithmic model is with the least-squares line: for the world population data in Fig. 3.12, the years have mean $\bar{x} = 1406$,