Chapter 4

Linear Models

Talk about factor analysis and PCA (the difference between having a model and the data pushing out a model for you – data mining!

The Basics of Linear Models

To clarify the some of these ideas, I begin with a mathematical expression that most of us learned in around seventh grade, the canonical linear equation:

$$y = mx + b$$

The standard way to graph this equation places the value of the x term on the horizontal axis and the value of the y on the vertical access. In this graphical representation, m denotes the slope of the line and b denotes the y intercept, the value of y when x equals zero.

This equation can also represent a model. To do this, the line is interpreted in such a way that the value of y depends on the value of x. For this reason, we call x the *independent*

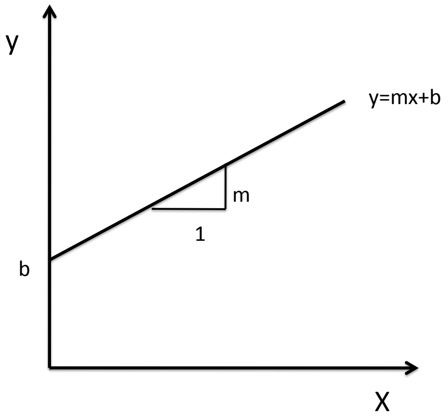

Figure 4.1: A Linear Equation

variable and y the *dependent* variable.

In a linear model, the slope describes how much of an effect x has on y. So, if I have a linear model of the price of a house, y, as a function of its square footage, x, and $m = 100$, then each increase of one square foot would result in one hundred dollar increase in the price of a house. The parameter b tells us how much a house would be worth if it had no square footage. This value might at first be thought to be zero, but houses tend to come with land, so b would capture the value of land. In a given neighborhood, we might then get the following linear equation for home prices:

$$Price = 100 \cdot SquareFootage + 40,000$$

If this model held true, a house of 1500 square feet would cost \$190,000 and a 3000 square foot house would cost \$340,000. Obviously, the price of a house depends on a host of other features – number of bathrooms, the size and type of kitchen, and the number of bathrooms. Even so, a linear model such as this proves very powerful. So powerful that if you go to a bank to get a loan, they look for "comparables." These are houses nearby of roughly similar quality. They then compute an equation pretty similar to the one above based on these comparables to get an estimated value for your house. This estimate serves as a benchmark for deeper thinking. If the estimate far exceeds what you're paying, the banker will question why you're getting such a good deal. If it's way below what you're paying, the banker may deny the loan, or at least demand a rationale for why you're paying over market. If the estimate is close, the purchase price earns a rubber stamp.

Why Everything is Linear (at least for a while)

Now that I've described linear models, I can show how for small changes everything is linear. Suppose that I have a curved function such as the one shown in the picture below: I can approximate that curved function with a series of linear segments.

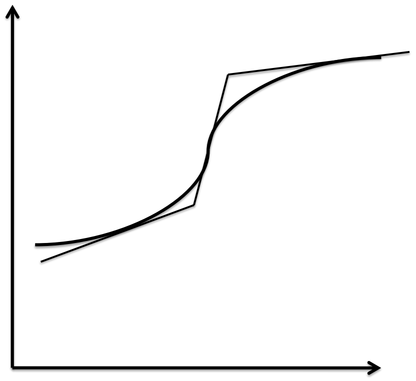

Figure 4.2: Approximating a Curve with Lines

Of course, the curvier a function becomes, the shorter the linear segments must be to approximate that curve. The picture below shows how a very curvy wall can be created with straight edged bricks.

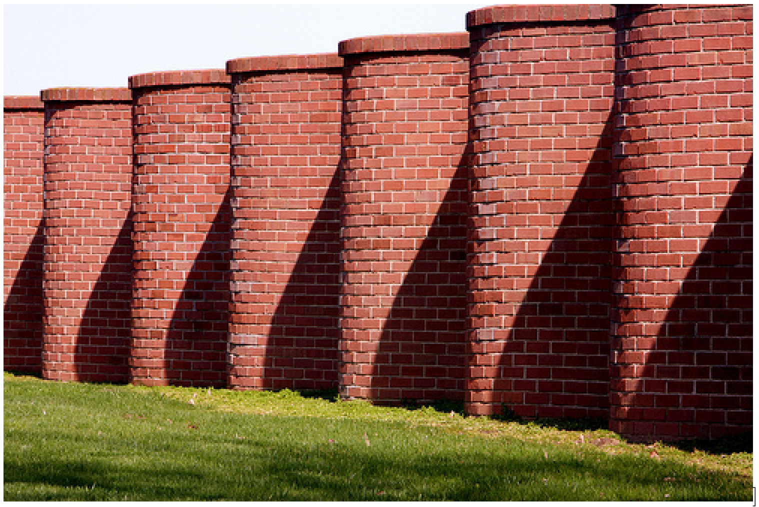

Figure 4.3: Curved Wall At Greenfield Village

The fact that we can approximate curves with line segments allows social scientists to use linear models to make sense of nonlinear phenomena.[1]

Constructing a Multivariable Linear Model

The linear models that I have described so far include only one variable. Yet, many of the phenomena of interest in the real world have multiple causes. For example, a person's happiness might depend on health, marital status, number of offspring, religious affiliation, and wealth. Fortunately, linear models can be expanded to include more than one causal (independent) variable.

Suppose, for example, that I want to use a linear model to understand what leads to success in school. I would first need to define a dependent variable – what is it that I'm using to measure success. I might, for example, take student performance on seventh grade assessment tests as my dependent variable, y.

I would then have to decide what variables might effect student performance. I might I

[1] In addition, techniques developed for linear models can be extended to allow for nonlinear terms.

have an intuition that class size is the key to student performance. To test that intuition, I could gather a whole bunch of data on student performance and class size. Performance, as measured by thetest, would be the dependent variable, y, and class size would be the independent variable x. All of the other things that might effect student performance such as teacher quality, parental income, student IQ, and so on, would be combined in the parameter b. A plot of the data might look as follows:

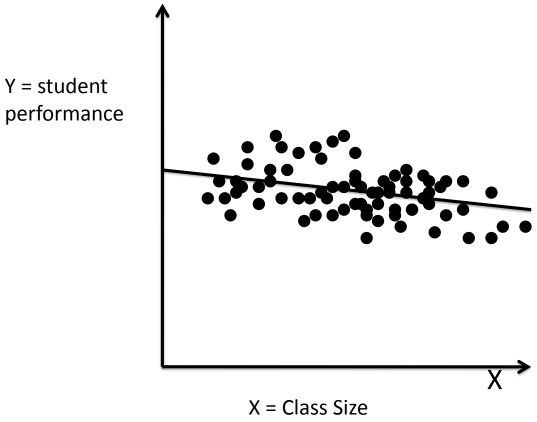

Figure 4.4: A Scatter Plot with a Linear Approximation

Notice that the data doesn't lie on the line. The line drawn is the best fit. It's the line that minimizes the total distance of all of the data points to a straight line. The line tells us that on average increasing class size appears to hurt student performance, but only mildly. The scatter of dots suggests that we cannot be too confident that in any particular school, a smaller class size will lead to better scores.

The distance from a point to the line is not just "noise" or measurement error. The distance captures all of the other contributors to students' performances such as teacher quality, parental income, and school spending. To include these other terms we create a multiple variable linear equation of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_0 n$$

In this representation, each of the x_i's are independent variables (class size, teacher quality, parental income, etc...) The parameter β_0 is all of the stuff that's still left out. Provided enough data exists, it's possible to determine the linear effect of each of these independent variables. The table below shows results from a collection of linear models designed to explain student performance on math, reading, and vocabulary tests.[2]

[2] Table from *Family Background and School Effects on Student Achievement: A Multilevel Analysis of the Coleman Data* by Spyros Konstantopoulos and Geoffrey Borman.

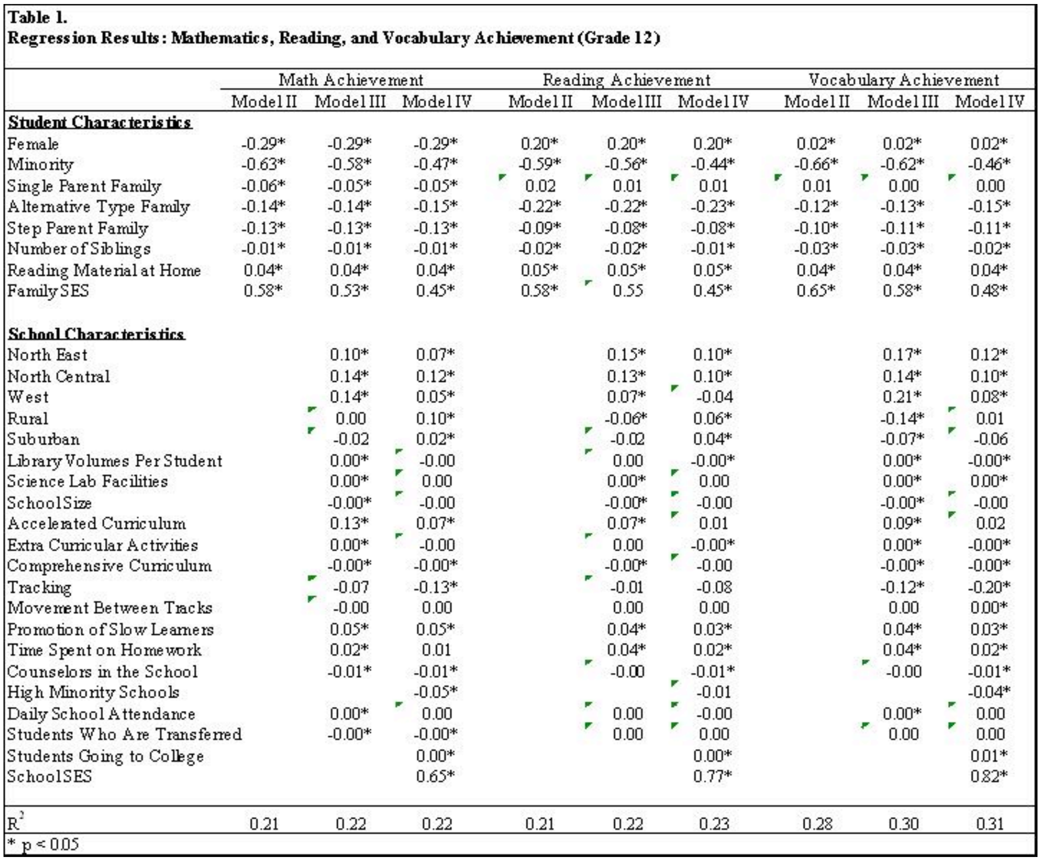

Table 1.
Regression Results: Mathematics, Reading, and Vocabulary Achievement (Grade 12)

	Math Achievement			Reading Achievement			Vocabulary Achievement		
	Model II	Model III	Model IV	Model II	Model III	Model IV	Model II	Model III	Model IV
Student Characteristics									
Female	-0.29*	-0.29*	-0.29*	0.20*	0.20*	0.20*	0.02*	0.02*	0.02*
Minority	-0.63*	-0.58*	-0.47*	-0.59*	-0.56*	-0.44*	-0.66*	-0.62*	-0.46*
Single Parent Family	-0.06*	-0.05*	-0.05*	0.02	0.01	0.01	0.01	0.00	0.00
Alternative Type Family	-0.14*	-0.14*	-0.15*	-0.22*	-0.22*	-0.23*	-0.12*	-0.13*	-0.15*
Step Parent Family	-0.13*	-0.13*	-0.13*	-0.09*	-0.08*	-0.08*	-0.10*	-0.11*	-0.11*
Number of Siblings	-0.01*	-0.01*	-0.01*	-0.02*	-0.02*	-0.01*	-0.03*	-0.03*	-0.02*
Reading Material at Home	0.04*	0.04*	0.04*	0.05*	0.05*	0.05*	0.04*	0.04*	0.04*
Family SES	0.58*	0.53*	0.45*	0.58*	0.55	0.45*	0.65*	0.58*	0.48*
School Characteristics									
North East		0.10*	0.07*		0.15*	0.10*		0.17*	0.12*
North Central		0.14*	0.12*		0.13*	0.10*		0.14*	0.10*
West		0.14*	0.05*		0.07*	-0.04		0.21*	0.08*
Rural		0.00	0.10*		-0.06*	0.06*		-0.14*	0.01
Suburban		-0.02	0.02*		-0.02	0.04*		-0.07*	-0.06
Library Volumes Per Student		0.00*	-0.00		0.00	-0.00*		0.00*	-0.00*
Science Lab Facilities		0.00*	0.00		0.00*	0.00		0.00*	0.00*
School Size		-0.00*	-0.00		-0.00*	-0.00		-0.00*	-0.00
Accelerated Curriculum		0.13*	0.07*		0.07*	0.01		0.09*	0.02
Extra Curricular Activities		0.00*	-0.00		0.00	-0.00*		0.00*	-0.00*
Comprehensive Curriculum		-0.00*	-0.00*		-0.00*	-0.00		-0.00*	-0.00*
Tracking		-0.07	-0.13*		-0.01	-0.08		-0.12*	-0.20*
Movement Between Tracks		-0.00	0.00		0.00	0.00		0.00	0.00*
Promotion of Slow Learners		0.05*	0.05*		0.04*	0.03*		0.04*	0.03*
Time Spent on Homework		0.02*	0.01		0.04*	0.02*		0.04*	0.02*
Counselors in the School		-0.01*	-0.01*		-0.00	-0.01*		-0.00	-0.01*
High Minority Schools			-0.05*			-0.01			-0.04*
Daily School Attendance		0.00*	0.00		0.00	-0.00		0.00*	0.00
Students Who Are Transferred		-0.00*	-0.00*		0.00	0.00		0.00	0.00
Students Going to College			0.00*			0.00*			0.01*
School SES			0.65*			0.77*			0.82*
R^2	0.21	0.22	0.22	0.21	0.22	0.23	0.28	0.30	0.31

* p < 0.05

Figure 4.5: Coefficients from a Linear Model

Each of the numbers represents the magnitude of the effect of the variable on student achievement. Those variables that have asterisks ($*$) next to them are statistically significant. This means that we can be reasonably confident that they are different from zero. If you look near the bottom of the table, you see that time spent on homework has a positive coefficient in every model, and it is significant in all but one. Here the linear model confirms something that most people believe anyway: studying improves school performance. That may not seem like much of a contribution. But keep in mind that the linear model tells us how much of an impact studying has and whether how it compares to relative to class size, teacher quality, and other variables in improving student performance.