7. Linear Models

I’m lying, yes, but why do you force me to give a linear explanation; linear explanations are almost always lies.

—Elena Ferrante

Often models posit specific functional relationships between variables. That relationship could be linear, concave, convex, or S-shaped, or it could include threshold effects. Of these, linear models are the simplest, the most widely used, and the focus of this chapter. The effects of education on income, of gains in life expectancy from exercise, and of income on voter turnout can all be measured using linear models.

We begin the chapter with a refresher on linear functions with a single variable. We then show how regression fits data to a linear function, revealing the sign, magnitude, and significance of effects. We also discuss why errors, noise, and heterogeneity mean that data do not fall exactly on the regression line. We then expand the linear model to allow for more variables and discuss how to fit multivariable linear models. To build intuition for multiple variable models, we describe a model of success as a linear function of skill and luck. The chapter concludes with an observation of how relying on data and regressions to guide action limits mistakes but can also produce marginal, conservative actions. This big-coefficient thinking can stifle innovation. To identify more innovative options, we might consider constructing other, more speculative models.

Linear Models

In a linear relationship, the amount of change in one variable due to a change in a second variable does not depend on the value of the second variable. If the height of a tree is linear with the tree’s age, the tree grows by the same amount each year. If the value of a house increases linearly in its square footage, a 200-square-foot addition increases a house’s value by double that of a 100-square-foot addition. A 400-square foot addition increases the house by four times as much.

Linear Models

In a **linear model** changes in the **independent variable**, *x*, result in linear changes in a **dependent variable**, *y*, as follows:

$y = mx + b$

where *m* equals the *slope* of the line and *b* equals the *intercept*, the value of the dependent variable when the independent variable equals zero.

A *linear regression* model finds the line that minimizes the distance to the data points. Linear regression can explain variation in crime, washing machine sales, and even wine prices.[1] Suppose that we have data for adults ranging in age from twenty to sixty and the distances they walk each week and find the following regression equation:

$\text{Miles Person}i = -0.1 \cdot \text{age}_i + 12 + \varepsilon_i$

This regression equation tells us the *sign* of the effect (distance decreases with age) and the *magnitude* of that effect (each year of age reduces distance by one-tenth of a mile). In this example, the intercept has no relevance because it lies outside our data range, that is the data includes no one with an age near zero. Based on the equation, we expect a forty-year-old to walk eight miles per week and a fifty-year-old to walk seven. The data used to produce a regression will not fall exactly on the regression line. Figure 7.1 shows hypothetical data used to produce our regression line. The person represented by the gray circle, Bobbi, is age forty and walks eleven miles per day. She exceeds the model's estimate by three miles. To make the data consistent with the model, the equation includes an *error term* for each data point. The error term, denoted by ε, equals the difference between what the model estimates and the actual value of the dependent variable. Bobbi's ε term equals +3 miles.

In social and biological contexts, we do not expect perfect linear fits. Outcomes depend on many variables, and a single-variable regression, by definition, includes only one variable. Predicted values can deviate from the actual values because of these *omitted variables*. Bobbi may walk more

than expected because, as a botany professor, she takes her students out for walks in the woods. The model does not include profession as a variable, which contributes to why the data in 7.1 do not lie on the line. The ε term could also result from *measurement error*. Fitness data collected by smartphones will contain errors if people forget to carry their smartphone or loan their phones to others. Error can also arise from *environmental noise*—people may earn extra distance for bumpy car rides to work.[2]

Figure 7.1: A Scatterplot and Regression Line

The closer the regression line lies to the data, the more of the data the model explains and the larger the model's R-squared (the percentage of variation explained). If all data lie exactly on the regression line, the R-squared equals 100%. All else equal, we prefer models with higher R-squared values.

Sign, Significance, and Magnitude

Linear regression tells us the following about *coefficients* of independent variables:

Sign: The correlation, positive or negative, between the independent variable and the dependent variable.

Significance (p-value): The probability that the sign on the coefficient is nonzero.

Magnitude: The best estimate of the coefficient of the independent variable.

In a single-variable regression, the closer fit to the line and the more data, the more confidence we can place in the sign and magnitude of the coefficient. Statisticians characterize the *significance* of a coefficient using its *p-value*, which equals the probability, based on the regression, that the coefficient is not zero. A p-value of 5% means a one-in-twenty chance that

the data were generated by a process where the coefficient equals zero. The standard thresholds for significance are 5% (denoted by *) and 1% (denoted by **). Significance is not all we care about. A coefficient can be significant yet of small magnitude. If so, we can be confident of the correlation but the variable has little effect. Or a coefficient can be large though not significant. This often occurs with noisy data or data with many omitted variables.

To see how to use regressions to guide action, imagine a company that ships spices. This company offers over a hundred types of spices. Customers buy packages of six, twelve, or twenty-four spices, which employees pack and ship. A regression estimating the number of orders shipped per eight-hour shift as a function of the number of years an employee has worked produces the following:

Orders Filled = 200 + 20**· Years

The coefficient on years, 20, is significant at the 1% level. We can be confident it is positive. If the relationship is causal (see below), the model can be used to predict the number of orders that each employee can fill per shift as a function of years of work and we can use the model to project how many orders the current employees will fill next year. Here we have an instance of a model both making a prediction and guiding an action.

Correlation vs. Causation

Regression only reveals correlation among variables, not causality.[3] If we first construct a model and then use regression to test if the model's results are supported by data, we do not prove causality either. However, writing models first is far better than running regressions in search of a significant correlate, a technique known as *data mining*. Data mining runs the risk of identifying a variable that correlates with other causal variables. For example, data mining might find a significant positive correlation between vitamin D levels and general health. People absorb vitamin D from sunlight, so the effect could be due to the fact that people with active lifestyles spend more time outdoors and have better health. Or a regression

might find that a school's academic performance correlates strongly with the number of students on its equestrian team. Equestrian teams likely have no direct causal effect but they correlate with family income and school funding levels which do.

Data mining can also result in spurious correlations, where just by chance two variables are correlated. We might find that companies with longer names earn higher profit or that people who live near pizza restaurants are more likely to get the flu. With a 5% significance threshold, one in every twenty variables we test will be significant. So, if we try enough variables, we will surely find significant (and spurious) correlations.

We can avoid reporting spurious correlations by creating *training sets* and *testing sets*. A correlation found on the training set that also holds on the testing set is far more likely to be true. We still have no guarantee of a causal relationship, however. To prove causality, we need to run an *experiment* where we manipulate the independent variable and see if the dependent variable changes. Or we look for a natural experiment where this has happened by chance.

Multivariable Linear Models

Most phenomena have multiple causal and correlative variables. A person's happiness can be attributed to health, marital status, offspring, religious affiliation, and wealth. The value of a house depends on square footage, lot size, the number of bathrooms, the number of bedrooms, the type of construction, and the quality of local schools. All of these variables can be included in a regression to explain housing values. We must keep in mind, though, that as we add more variables, we need more data to obtain significant coefficients.

Before discussing multiple-variable regression, we build intuition for multiple-variable equations by introducing Mauboussin's *skill-luck equation*.[4] The equation writes success, be it in work, sports, or games, as a weighted linear function of skill and luck.

The Success Equation

Success = $a \cdot$ Skill + $(1 - a) \cdot$ Luck

where a in [0, 1] equals the relative weight on skill.

If we can assign relative weights to skill and luck, perhaps by using a regression if we had data, we could use the model to predict outcomes. If the manager of a team of recreational vehicle salespeople finds that success, measured in sales, has a large luck component, he would expect *regression to the mean*: salespeople who did well this month would be likely to be about average the next month. The manager could then use the model to guide action. He might not want to match a higher salary offer from a competitor for a salesperson who had two good months in a row. If instead the regressions showed that luck played almost no role, performance in two months would be a good predictor of performance in future months. In this case, the manager would want to match an outside offer for the best salesperson.

The same insight applies to CEO pay. A board of directors should not pay bonuses to CEOs who work in industries where luck determines success. An oil company's profits depend on the market price of crude oil, a variable that lies outside the company's control. An oil company's board should therefore be reluctant to reward a CEO for a good year. An advertising company would be wise to do the opposite—to award a large bonus to a CEO if the company performs well. In brief, pay for skill; do not pay for luck. Better-run corporations do in fact pay less for luck.[5]

Even the simplest of models, such as this one, produce subtle insights. By thinking about the equation, we see that even in a context that depends almost entirely on skill, such as running, biking, swimming, chess, or tennis, if skill differences are small, luck largely determines who wins. We might expect that in the most competitive environments, like the Olympics, skill differences are small, and thus luck matters. Mauboussin calls this the *paradox of skill*. Michael Phelps, the greatest swimmer in history, has been on both ends of the paradox. In the 2008 Olympic Games, Phelps trailed Milorad Cavic at the end of the 100-meter butterfly. Yet by a stroke of luck,

Phelps touched the wall first. In the 2012 Olympic Games, Phelps led Chad le Clos at the finish, but le Clos touched first. Yes, Phelps has incredible skill, but that one win and that one loss were the products of luck.

Multiple-Variable Regression

Multiple-variable linear regression models fit linear equations with many variables and also minimize the total distance to the data. These equations include coefficients for each independent variable. The equation below shows a hypothetical regression output for student performance on a math test as a function of hours studied (HRS), family socioeconomic status (SES), and the number of accelerated classes (AC).

$$\text{Math Score} = 21.1 + 9.2^{**} \cdot \text{HRS} + 0.8 \cdot \text{SES} + 6.9^{*} \cdot \text{AC}$$

According to the regression, a student's score increases by 9.2 points for every extra hour spent studying. The coefficient has two *'s, so it is significantly different from zero at a 1% level. This implies strong correlation, though not causality. The equation also shows that a student scores almost seven points higher for each accelerated class. That coefficient is significant as well, but only at the 5% level. Family socioeconomic status (SES), a variable that takes on values from 1 (low) to 5 (high), has a coefficient that is positive but not significantly different from zero, so we can assume it probably has little causal effect.

With this or any regression output, we can predict outcomes. The model predicts that a student who spends seven hours studying and takes one other accelerated class should score in the 90s. The model can also guide actions, though we must be cautious, as we cannot infer causality. The data show that students who study and take accelerated classes perform better. One reason studying more or taking those classes may not help is *selection bias*. It might be that the students who study more and those who take accelerated classes are better at math.

Even though regressions cannot prove what causes patterns in data, they can rule out explanations. Take the large wealth disparity by race in the United States: in 2016, the average wealth of white families (approximately

$110,000) was more than ten times that of African American and Latino families. Any number of causes might explain that gap, including institutional factors, differences in income, savings behavior, or marriage rates. Regressions support some explanations and rule out others. For example, regressions reveal no significant relationship between marital status and wealth among African Americans, so marital status cannot be a cause. Income differences, though substantial, also prove insufficient to explain the gap.[6]

The Big Coefficient and New Realities

As already stated, linear regression models play prominent roles in scientific research, policy analysis, and strategic decision-making, in part because they are easy to estimate and interpret. With the increased availability of data, they have become even more widely used. The phrase "In God we trust. Everyone else must bring data" is often heard in business and in the halls of government. A reliance on data—and that often means linear regression models—can steer us toward marginal actions and away from big new ideas. A business, government, or foundation that gathers data, fits a linear regression model, and finds the variable with the largest statistically significant coefficient almost cannot stop itself from adjusting that variable and taking the marginal gain.

When taking an action, it is better to choose the variable with the *big coefficient* than a variable with a small coefficient. At the same time, big-coefficient thinking builds in conservatism. It focuses attention on certain modest improvements and pulls attention away from novel policies. A second problem with big-coefficient thinking is that the magnitude of the big coefficient corresponds to the marginal effect given existing data. Often, as we see in the next chapter, effect sizes diminish as we increase the value of a variable. If so, the big coefficient becomes smaller as we try to exploit it.

The Big Coefficient vs. the New Reality

Linear regressions reveal the magnitude of correlations of independent variables with the variable of interest. If that correlation is causal, changes to the variable with a **big coefficient** will have large effects. Policies based on big coefficients guarantee improvements but rule out *new realities* that involve more fundamental changes.

The alternative to big-coefficient thinking is *new-reality thinking*. Big-coefficient thinking widens roads and builds high-occupancy vehicle lanes to reduce traffic. New-reality thinking builds train and bus systems. Big-coefficient thinking subsidizes computers for low-income students. New-reality thinking gives everyone a computer and reduces mail delivery to three days a week. Big-coefficient thinking changes the width of airline seats. New-reality thinking creates an airplane interior that can be filled with interchangeable pods. Big coefficients are good. Evidence-based action is wise, but we must also keep our eyes open to big new ideas as well. When we encounter them, we can use models to explore whether they might work. A regression on teenage traffic accidents may find that age has the largest coefficient, implying that states might want to raise the driving age. That may work, but so too might more novel policies such as curfews that prohibit nighttime driving, automated monitoring of teenage drivers through smartphones, or limits on the number of passengers in teenagers' cars. These new-reality policies might produce larger effect sizes than riding the big coefficient.

Summary

To summarize, linear models posit constant effect sizes. Linear regression offers a powerful tool for taking a first cut at data, enabling us to identify the sign, magnitude, and significance of variables. If we want to know the health effects of coffee, alcohol, or soda consumption, we can run regressions. We may find that coffee consumption reduces the risks of cardiovascular disease and that so do modest levels of alcohol consumption. That said, we should be skeptical of extrapolating linear effects too far outside of the existing data range. We should not infer that thirty cups of coffee, much less six glasses of wine, would be a good idea. Nor should we make linear projections too far ahead in time. California's population grew at a rate of 45% from 1880 to 1960. Had we made a linear

projection, we would have pegged California's population in 2018 at 100 million people, more than double its actual level.

Keep in mind we are just getting started. Most phenomena of interest are not linear. For that reason, regression models often include nonlinear terms such as age squared, the square root of age, or even the log of age. To account for nonlinearities, we can also arrange linear models end to end. These concatenated linear models can approximate a curve in much the same way as we can use straight-edged bricks to construct a curved path. Though linearity may be a strong, and unrealistic, assumption, it offers a good place to start. If given data, we can use linear models to test our intuitions. We can then construct more elaborate models in which the effect of a variable dampens as it increases (diminishing returns) or becomes more powerful (positive returns). These nonlinear models are the focus of the next chapter.

Binary Classifications of Data

In an era of Big Data, organizations use algorithms informed by models to classify their data. A political party might want to learn who votes, an airline might want to learn the attributes of their frequent flyers, and an event organizer might want to learn about the event's attendees. In each case, the organization classifies people into two sets: those who buy, contribute, or enroll are labeled as *positives* (+'s) and those who are not are labeled as *negatives* (-'s).

Classification models apply *algorithms* to partition the people into categories based on attributes such as a person's age, income, education level, or hours spent on the internet. Different algorithms imply different underlying models of the relationship between attributes and outcomes. Applying multiple algorithms—using many models—will produce an even better classification.

Linear classifications: In figure M1, positives (+) represent voters and negatives (-) represent nonvoters. A linear function of a person's age and

education level can be used classify whether or not a person votes. The data show that more educated people are more likely to vote and that older people are more likely to vote. In this example, a straight line classifies nearly perfectly.[7]

Figure M1: Using a Linear Model to Classify Voting Behavior

Nonlinear classifications: In figure M2, positives (+) represent frequent flyers, consumers who fly more than 10,000 miles per year, and negatives (-) represent all other customers of an airline. People of middle age and higher income are more likely to fly. To classify these data requires a nonlinear model, which could be estimated using *deep-learning* algorithms, such as neural networks. Neural networks include more variables so that they can fit almost any curve.

Figure M2: Using a Nonlinear Model to Classify Frequent Flyers

Forests of decision trees: In figure M3, positives (+) represent people who attended a science fiction convention based on their age and the hours per week they spend on the internet. Here we classify the data using three *decision trees*. Decision trees make classifications based on sets of conditions on the attributes. The figure shows three trees:

Tree 1: *If (age < 30) and (internet hours per week in [15, 25])*

Tree 2: *If (age in [20, 45]) and (internet hours per week > 30)*

Tree 3: *If (age > 40) and (internet hours per week < 20)*

Figure M3: A Forest of Decision Trees Classifying Conference Attendees

The collection of trees are called a *forest*. Machine learning algorithms create trees randomly on a training set and then keep those that classify accurately on the testing set and on a training set.