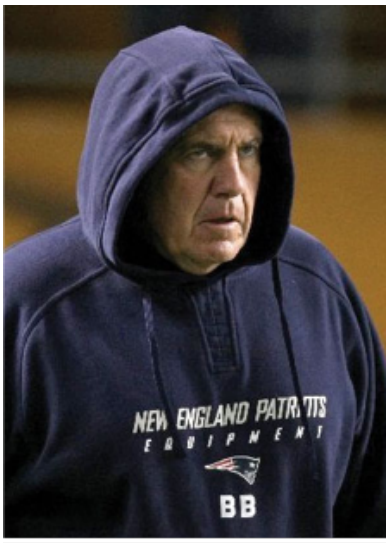
7
Testing hypotheses

Assessing the evidence for a hypothesis



Among professional football fans, the New England Patriots are a polarizing team. Their fan base is hugely devoted, probably due to their long run of success over more than a decade. Many others, however, dislike the Patriots for their highly publicized cheating episodes, whether for deflating footballs or clandestinely filming the practice sessions of their opponents. This feeling is so common among football fans that sports websites often run images like the one at right (of the Patriots' be-hoodied head coach, Bill Belichick), or articles with titles like "11 reasons why people hate the Patriots." Despite—or perhaps because of—their success, the Patriots always seems to be dogged by scandal and ill will.

But could even the Patriots cheat at the pre-game *coin toss*?

Believe it or not, many people think so! That's because, for a stretch of 25 games spanning the 2014-15 NFL seasons, the Patriots won 19 out of 25 coin tosses—that's a 76% winning percentage. Needless to say, the Patriots' detractors found this infuriating. As one TV commentator remarked when this unusual fact was brought to his attention: "This just proves that either God or the devil is a Patriots fan, and it sure can't be God."

But before turning to religion, let's take a closer look at the evidence. Just how likely is it that one team could win the pre-game coin toss at least 19 out of 25 times, assuming that there's no cheating going on?

This question is easy to answer using probability theory—specifically, something called the binomial distribution. But it's also very easy to answer using the Monte Carlo method, in which we write a computer program that simulates a random process. In Figure 7.1, we see the results of a Monte Carlo simulation for pre-game NFL coin tosses, which the Patriots ought to have a 50% chance of winning each toss. Specifically, we have repeated the

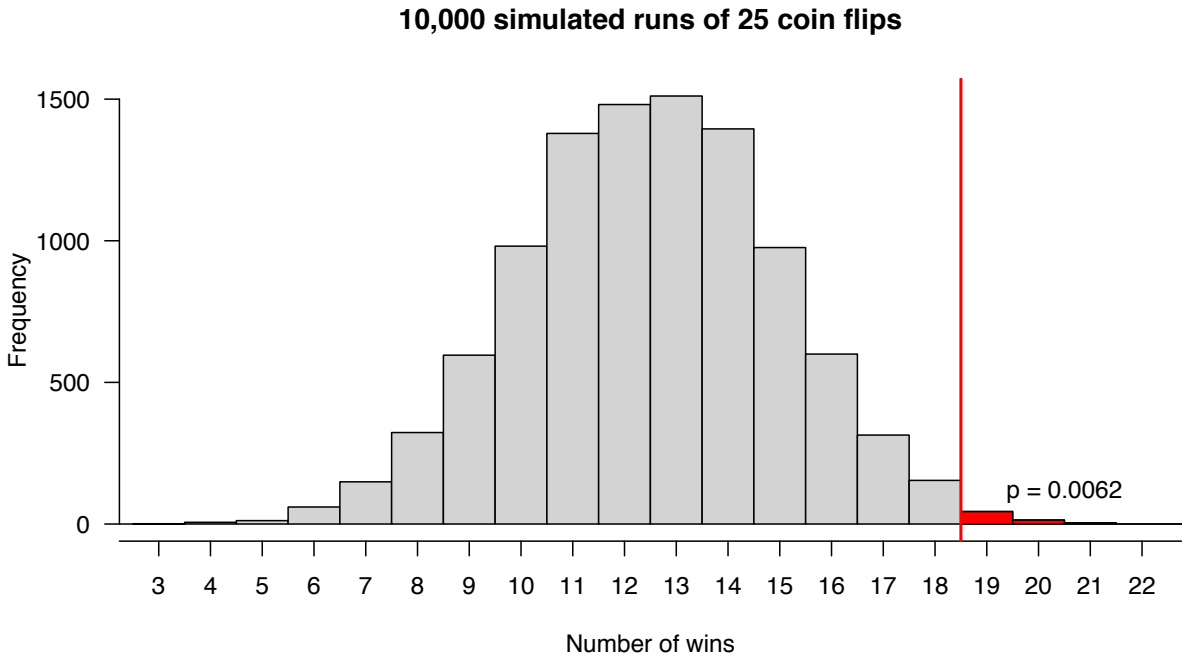


Figure 7.1: This histogram shows the results of a Monte Carlo simulation, in which we count the number of wins in 25 simulated coin flips over 10,000 different simulations. The red area (which has cumulative probability of 0.0062) approximates the probability of winning 19 or more flips, out of 25.

following simple process 10,000 times:

1. Simulate 25 coin tosses in which the Patriots have a 50% chance of winning each toss.
2. Count how many times out of 25 that the Patriots won the toss.

If you're counting, that's 250,000 coin tosses: 10,000 simulations of 25 tosses each.

Figure 7.1 shows a histogram of the number of coin tosses won by the Patriots across 10,000 simulations. Clearly 19 wins is an unusual, although not impossible, number under this distribution: in our simulation, the Patriots won at least 19 tosses only 62 of 10,000 times ($p = 0.0062$), shown as the red area in Figure 7.1.

So did the Patriots win 19 out of 25 coin tosses by chance? Well, nobody knows for sure—I report, you decide.[1] But unless you're a hard-core NFL conspiracy theorist, let me encourage you to forget the Patriots for a moment and focus instead on the process we've just gone through. This simple example has all the major elements of *hypothesis testing*, which is the subject of this chapter:

[1] Despite the small probability of such an extreme result, it's hard to believe that the Patriots cheated on the coin toss, for a few reasons. First, how could they? The coin toss would be extremely hard to manipulate, even if you were inclined to do so. Moreover, the Patriots are just one team, and this is just one 25-game stretch. There are 32 NFL teams, so the the probability that *one* of them would go on an unusual coin-toss winning streak over *some* 25-game stretch over a long time period is a lot larger than the number we've calculated. Finally, after this 25-game stretch, the Patriots reverted back to a more typical coin-toss winning percentage, closer to 50%. The 25-game stretch was probably just luck.

(1) We have a *null hypothesis*, that the pre-game coin toss in the Patriots' games was truly random.

(2) We use a *test statistic*, number of Patriots' coin-toss wins, to measure the evidence against the null hypothesis.

(3) There is a way of calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin.

(4) Finally, we used this probability distribution to assess whether the null hypothesis looked believable in light of the data.

All hypothesis testing problems have these same four elements. Usually the difficult part is Step 3: calculating the probability distribution of the test statistic, assuming that the null hypothesis is true. The essence of the problem is that, in most cases, we can't just run a simple simulation of coin flips. Luckily, there is a very general way of proceeding here, called the permutation test, which we will now learn about.

Permutation tests

Is gun violence correlated with gun policy?

Gun policy is an important and emotionally charged topic in 21st-century America, where gun violence occurs with far higher frequency than it does in other rich countries. Many people feel strongly that certain types of guns, like military-style assault weapons, should be banned, and that all gun purchases should be subject to stronger background checks. Others view gun ownership as both an important part of their cultural heritage and a basic right protected by the U.S. Constitution. Like with many issues, there seems to be little prospect of a national consensus.

Both gun laws, and the likelihood of dying violently as a result of gun crime, vary significantly from state to state. Figure 7.2 shows some of this variation in a *chloropleth map*, where discrete areas on the map are shaded according to the value of some numerical variable. Notice that the states are shown as a gridded tile of equal-sized hexagons, rather than as an actual map of the United States. This is common technique used to avoid the visual imbalances due to large differences in the states' total area.

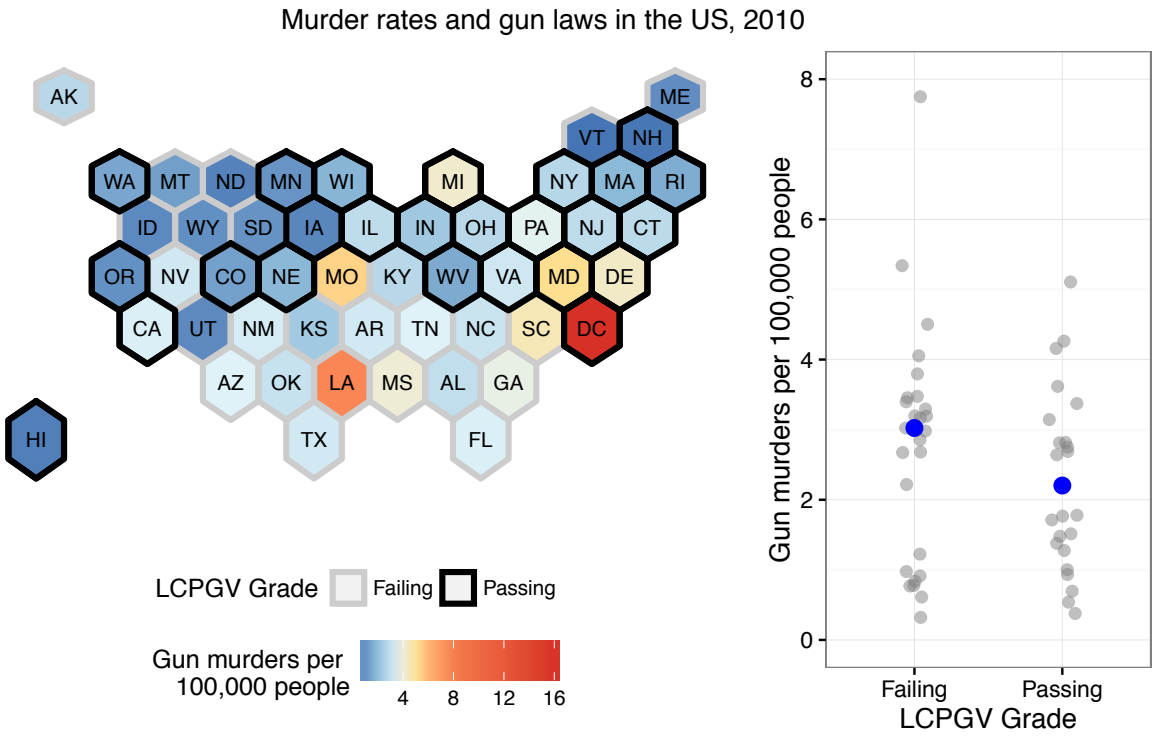


Figure 7.2: Left panel: a chloropleth map of murder rates versus gun laws across the U.S. states. The shaded color shows the state's gun-murder rate; blue is lower, and red is higher. The outline indicates whether a state's gun-control laws received a passing or a failing grade from the Law Center to Prevent Gun Violence (black for passing, grey for failing). The right panel shows a dot plot of the gun-murder rates across the two groups, together with the median for each group in blue. Washington (D.C.), at 16.2 gun murders per 100,000 people, is far off the top of the plot, but is still included in all calculations. According to its website, `http://smartgunlaws.org`, the LCPGV is "a national law center focused on providing comprehensive legal expertise in support of gun violence prevention and the promotion of smart gun laws that save lives." You can read a full description of the methodology used to grade states at this link.

In the chloropleth map in Figure 7.2, the fill color indicates each state's gun-murder rate in 2010: blue is lower, red is higher. The outline color indicates whether a state's gun-control laws received a passing or failing grade from the Law Center to Prevent Gun Violence (LCPGV). The center graded each state's gun laws on an A–F letter-grade scale; here "failing" means a grade of F. In the figure, a black outline means a passing grade, while a grey outline means a failing grade.

The right panel of Figure 7.2 summarizes the relationship between gun laws and gun violence via a dot plot, together with the median for each group in blue. We use the median rather than the mean to estimate the center of each group, because the median is more robust to outliers; a clear example of an outlier here is Washington (D.C.), which at 16.2 gun murders per 100,000 people has a drastically higher rate than everywhere else in the country.

This dotplot shows that the median murder rate of states with a failing gun-laws grade is 3 murders per 100,000 people, while the median murder rate of states with a passing grade is 2.2 per

100,000. On the face of it, it would seem as the states with stricter gun laws have lower murder rates.

Let's aside for a moment the fact that correlation does not establish causality. We will instead address the question: could this association have arisen due to chance? To make this idea more specific, imagine we took all 50 states and randomly divided them into two groups, arbitrarily labeled the "passing" states and the "failing" states. We would expect that the median murder rate would differ a little bit between the two groups, simply due to random variation (for the same reason that hands in a card game vary from deal to deal). But how big of a difference between these two groups could be explained by chance?

Null and alternative hypotheses

Thus there are two hypotheses that can explain Figure 7.2:

(1) There is no systematic relationship between murder rates and gun laws; the observed observed relationship between murder rates and gun laws is consistent with other unrelated sources of random variation.

(2) The observed relationship between murder rates and gun laws is too large to be consistent with random variation.

We call hypothesis 1 the *null hypothesis*, often denoted H_0. Loosely, it states that nothing special is going on in our data, and that any relationship we thought might have existed isn't really there at all.[2] Meanwhile, hypothesis 2 is *alternative hypothesis*. In some cases the alternative hypothesis may just be the logical negation of the null hypothesis, but it can also be more specific.

In the approach to hypothesis testing that we'll learn here, we don't focus a whole lot on the alternative hypothesis.[3] Instead, we set out to check whether the null hypothesis looks plausible in light of the data—just as we did when we tried to check whether randomness could explain the Patriots' impressive run of 19 out of 25 coin flips won.

A permutation test: shuffling the cards

In the Patriots' coin-flipping example, we could easily simulate data under the null hypothesis, by programming a computer to repeatedly flip a virtual coin and keep track of the winner. But of course, most real-life hypothesis-testing situations don't involve

[2] "Null hypothesis" is a term coined in the early twentieth century, back when "null" was a common synonym for "zero" or "lacking in distinctive qualities." So if the term sounds dated, that's because it is.

[3] Specifically, this approach is called the *Fisherian* approach, named after the English statistician Ronald Fisher. There are more nuanced approaches to hypothesis testing in which the alternative hypothesis plays a major role. These include the Neyman–Pearson framework and the Bayesian framework, both of which are widely used in the real world, but which are a lot more complicated to understand.

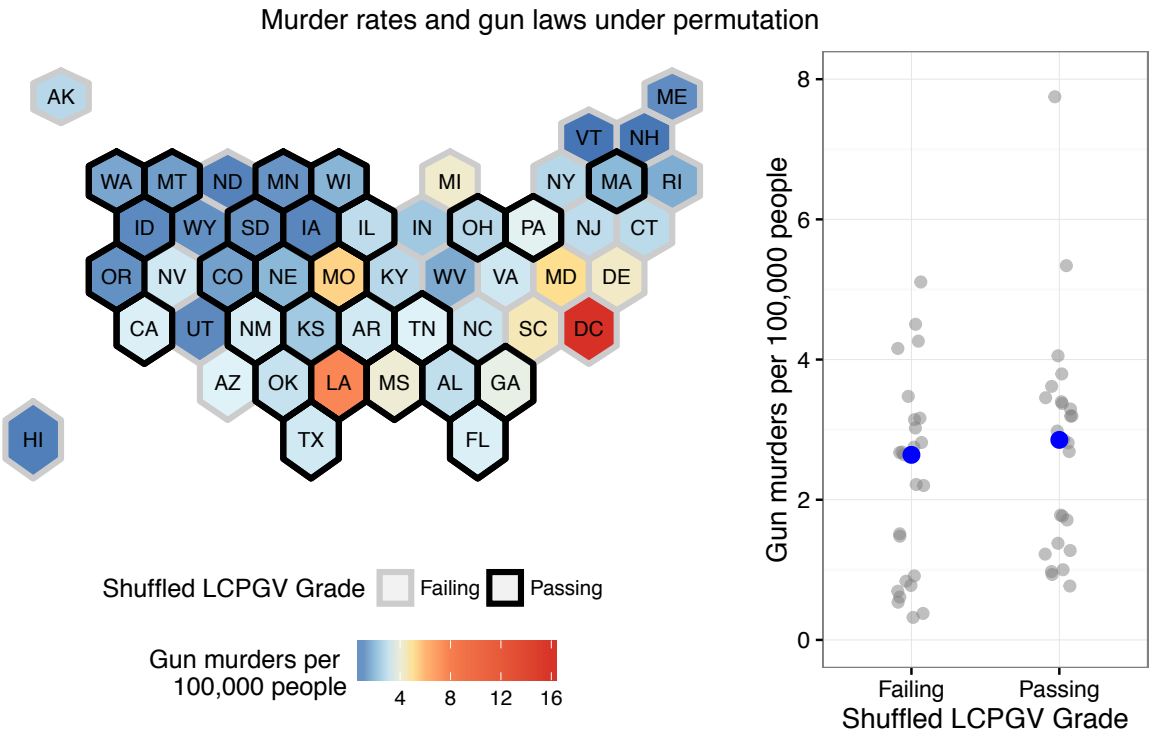


Figure 7.3: This map is almost identical to Figure 7.2, with one crucial difference: the identities of the states with passing and failing grades have been randomly permuted. There is still a small difference in the medians of the notionally passing and failing groups, due to random variation in the permutation process.

actual coin flips, which makes the virtual coin-flipping approach somewhat unhelpful as a general strategy.

It turns out, however, that in most situations, we can still harness the power of Monte Carlo simulation to understand what our data would look like if the null hypothesis were true. Rather than flipping virtual coins, we run something called a *permutation test*, which involves repeatedly permuting (or shuffling) the predictor variable and recalculating the statistic of interest.

To understand how this works, let's see an example. Figure 7.3 shows a map and dotplot very similar to those in Figure 7.2, with one crucial difference: in Figure 7.3, the identities of the states with notionally "passing" and "failing" gun laws have been randomly permuted. These grades bear no correspondence to reality. It's as though we took a deck of 51 cards, each card having some state's grade on it (treating D.C. as a state); shuffled the deck; and then dealt one card randomly to each state. The mathematical term for this is a *permutation* of the grades.

As expected, the median gun-murder rates of these two ran-

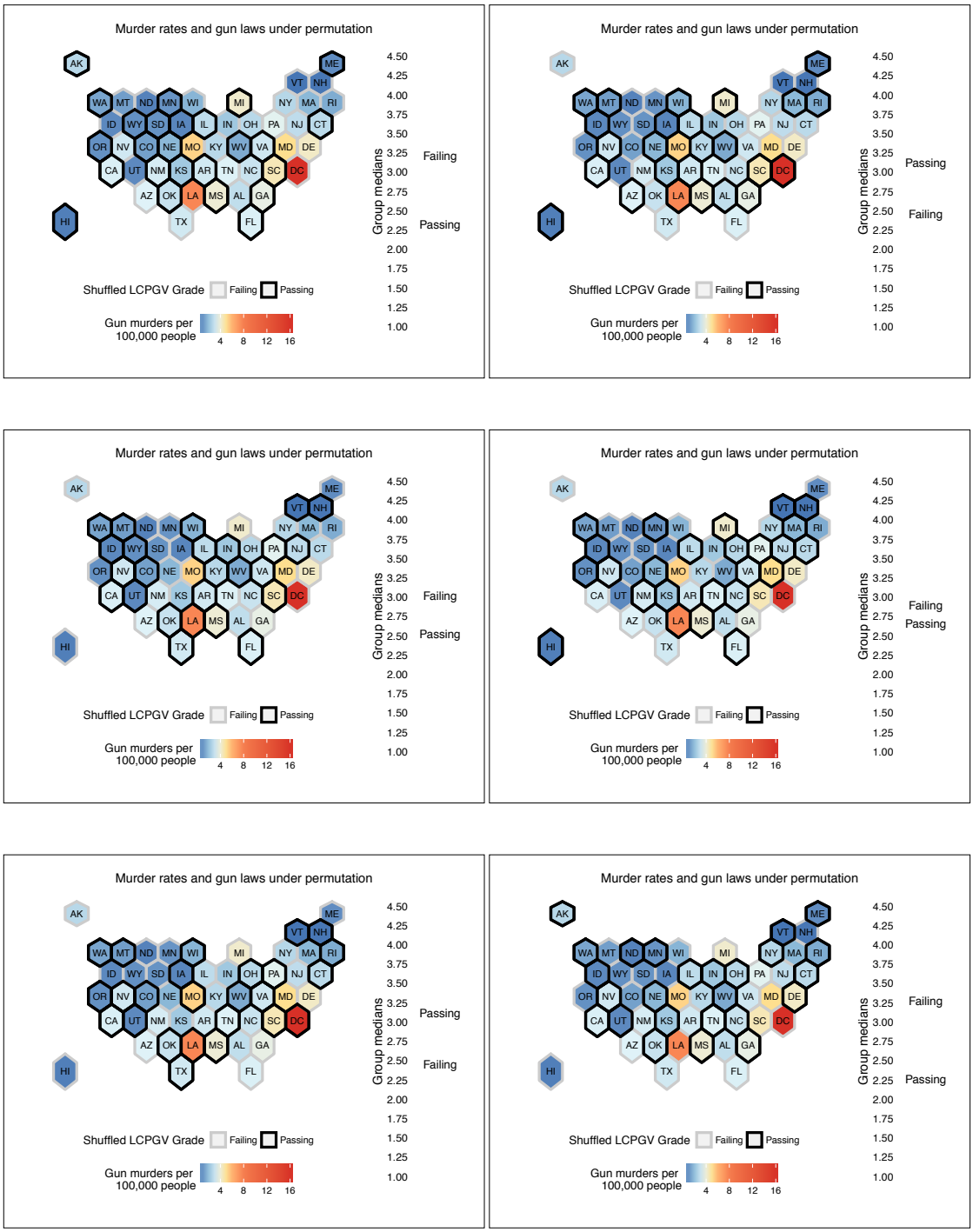


Figure 7.4: Six maps with permutated gun-law grades, with the medians for the passing and failing groups.

dom chosen "passing" and "failing" groups aren't identical (right panel). The randomly chosen "failing" states have a median of 2.6, while the randomly chosen "passing" states have a slightly larger median of 2.8. Clearly we can get a difference in medians of at least 0.2 quite easily, just by random chance—that is, when the null hypothesis is true by design.

But Figure 7.3 shows the difference in medians for only a single permutation of the states' gun-law grades. This permutation is random, and a different permutation would have given as a slightly different answer. Therefore, to assess whether could we get a difference in group medians as large as 0.8 just by random chance, we need to try several more permutations.

Figure 7.4 shows 6 more maps generated using the same permutation procedure. For each map, we shuffle the grade variables for all the states and recompute the median murder rates for the notionally "passing" and "failing" groups. Each map leads to its own difference in medians. In some maps, the difference is positive ("passing" states are higher), while in others it is negative ("failing" are states higher). In at least one of the 6 maps—the bottom right one—the median for the "failing" states exceeds the median for the "passing" states by more than 1 murder per 100,000 people, just by chance. This is a larger difference than we see for the real map, in Figure 7.2.

Six permutations give us some idea of how much a difference in the medians we could expect to see if the null hypothesis were true. But ideally we'd have many more than 6. Figure 7.5 addresses this need, showing the result of a much larger Monte Carlo simulation in which we generated 5,000 random maps, each one with its own random permutation of the states' gun-law grades. For each of these 5,000 maps, we computed the difference in medians between the notionally passing and failing groups. These 5,000 differences in group medians across the 5,000 maps are shown as a histogram in Figure 7.5.

Hypothesis testing: a four-step process

Let's review the vocabulary that describes what we've done here. First, we specified a null hypothesis: that the correlation between rates of gun violence and state-level gun policies could be explained by other unrelated sources of random variation. We decided to measure this correlation using a specific statistic: the difference in medians between the states with passing grades and

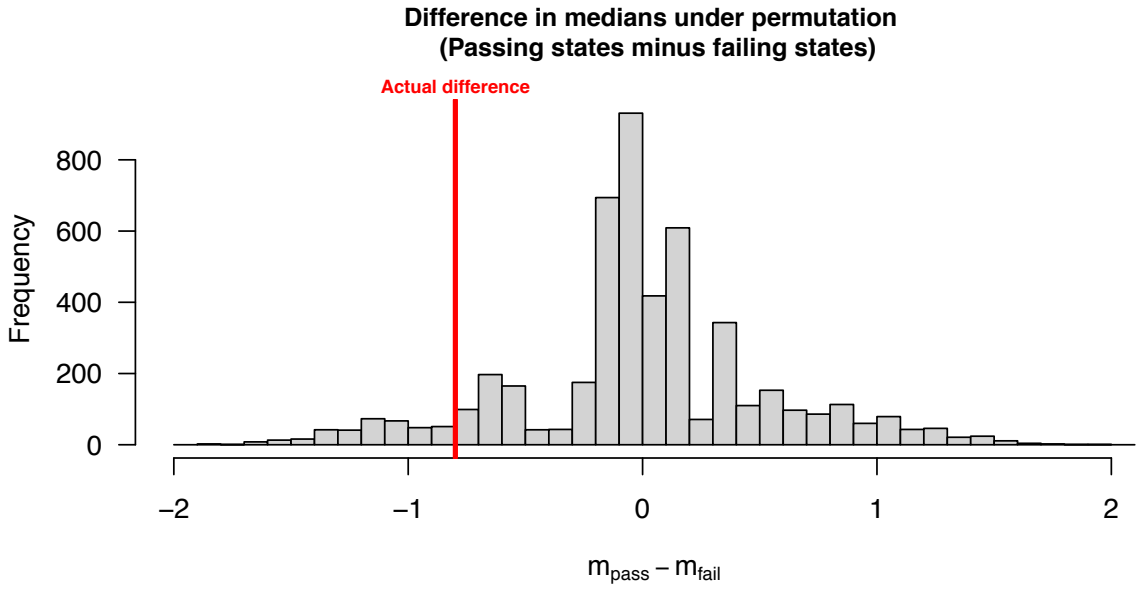


Figure 7.5: The histogram shows the difference in group medians for 5,000 simulated maps generated by the same permutation procedure as the 6 maps in Figure 7.4. Negative values indicate that the "failing" states had higher rates of gun violence than the "passing" states. The actual difference in medians for the real map in Figure 7.2 is shown as a vertical red line. This difference seems to be consistent with (although does not prove) the null hypothesis that other sources of random variation, and not necessarily state-level gun policy, explains the observed difference in murder rates.

those with failing grades. (Remember that a statistic is just some numerical summary of a data set.) To give this statistic a name, let's call it Δ (for difference in medians). It's intuitively clear that the larger Δ is, the less plausible the null hypothesis seems.

Figure 7.5 quantifies this intuition by giving us an idea of how much variation we can expect in the sampling distribution of our Δ statistic under the hypothesis that there is no systematic relationship between gun laws and rates of gun violence. As before, the sampling distribution is simply the probability distribution of the statistic under repeated sampling from the population—in this case, assuming that the null hypothesis is true.

There are two possibilities here, corresponding to the null and alternative hypotheses. First, suppose that we frequently get at least as extreme a value of Δ for a random map, like those in Figure 7.4, as we do in the real map from Figure 7.2. Then there's no reason to be especially impressed by the actual value of $\delta = -0.8$ we calculated from the real map.[4] It could have easily happened by chance. Hence we will be unable to reject the null hypothesis; it could have explained the data after all. (An important thing to remember is that *failing to reject* the null hypothesis is not the

[4] We use the lower-case δ to denote the value of the test statistic for your specific sample, to distinguish it from the Δ's simulated under permutation.

same thing as *accepting* the null hypothesis as truth. To use a relationship metaphor: failing to reject the null hypothesis is not like getting married. It's more like agreeing not to break up this time.)

On the other hand, suppose that we almost always get a smaller value of Δ in a random map than we do in the real map. Then we will probably find it difficult to believe that the correlation in the real map arose due to chance. We will instead be forced to reject the null hypothesis and conclude that it provides a poor description of the observable data.

Which of these two possibilities seems to apply in Figure 7.5? Here, the actual difference of -0.8 for the real map in Figure 7.2 is shown as a vertical red line. It's position on the histogram suggests possibility (1) here: $\delta = -0.8$ is consistent with (although does not prove) the null hypothesis that other sources of random variation unrelated to state-level gun policy can explain the observed difference in murder rates between the passing-grade and the failing-grade states.

To summarize, the four steps we followed above were:

(1) Choose a null hypothesis H_0, the hypothesis that there is no systematic relationship between the predictor and response variables.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Assess whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

For the gun-laws example, our test statistic in step (2) was the difference in medians between the "passing" states and the "failing" states. We then accomplished step (3) by randomly permuting the values of the predictor (gun laws) and recomputing the test statistic for the permuted data set. This shuffling procedure is called a permutation test when it's done in the context of this broader four-step process. There are other ways of accomplishing step (3)—for example, by appealing to probability theory and doing some math. But the permutation test is nice because it works for any test statistic (like the difference of medians in the previous example), and it doesn't require any strong assumptions.

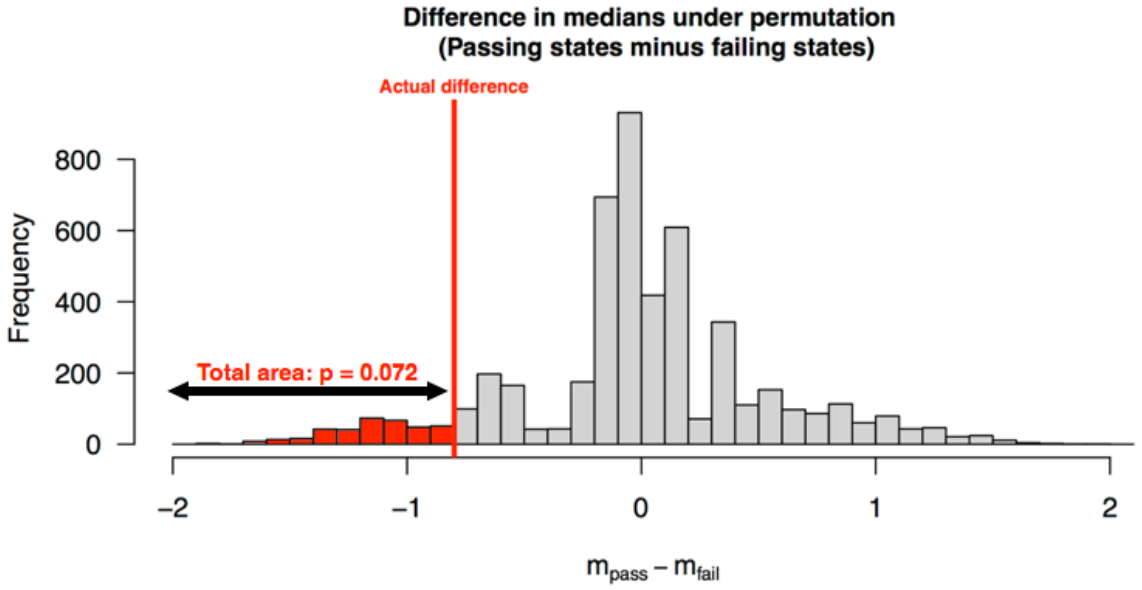


Figure 7.6: Assuming that the null hypothesis is true, the probability of observing a difference in medians at least as extreme as $\delta = -0.8$ is $p = 0.072$. This tail area to the left of $\delta = -0.8$ is the p-value of the test.

Using and interpreting p-values

There's one final question we haven't answered. How do we accomplish step (4) in the hypothesis test? That is, how can we measure whether the observed statistic for your data is consistent with the null hypothesis?

The typical approach here is to compute something called a *p-value*. Although we didn't call it by the name "p-value," this is exactly what we did for the Patriots' coin-flipping example at the beginning of the chapter.

Let's begin with a concise definition of a p-value, before we slowly unpack the definition (which is dense and non-intuitive). *A p-value is the probability of observing a test statistic as extreme as, or more extreme than, the test statistic actually observed, given that the null hypothesis is true.* The way to compute the p-value is to calculate a *tail area* indicating what proportion of the sampling distribution, $P(\Delta \mid H_0)$, lies beyond the observed test statistic δ.

This all sounds a bit abstract, but is much easier to understand by example. Let's go back to the gun-laws hypothesis test, where we observed a difference in the medians of $\delta = -0.8$. If the null hypothesis were true, the probability of getting $\delta = -0.8$ (or

something more extreme in the negative direction) would be $p = 0.072$. We calculate this by taking the tail area under the sampling distribution that to the left of our observed δ of -0.8. Figure 7.6 highlights this area in the left tail of the sampling distribution $P(\Delta \mid H_0)$. This is the p-value.

Using p-values has both advantages and disadvantages. The main advantage is that the p-value gives us a continuous measure of evidence against the null hypothesis. The smaller the p-value, the more unlikely it is that we would have seen our data under the null hypothesis, and therefore the greater the evidence the data provide that H_0 is false.

The main disadvantage is that the p-value is hard to interpret correctly. Just look at the definition—it's pretty counterintuitive! To avoid having to think too hard about what a p-value actually means, people often take $p \leq 0.05$ as a very important threshold that demarcates "significant" ($p \leq 0.05$) from "insignificant" ($p > 0.05$) results. While there are some legitimate reasons[5] for thinking in these terms, in practice, the $p \leq 0.05$ criterion can feel pretty silly. After all, there isn't some magical threshold at which a result becomes important: in all practical terms, $p = .049$ and $p = .051$ are nearly identical in terms of the amount of evidence they provide against a null hypothesis.

Because of how counterintuitive p-values are, people make mistakes with them all the time, even (perhaps especially) people with Ph.D's quoting p-values in original research papers. Here is some advice about a few common misinterpretations:

- The p-value is *not* the probability that the null hypothesis is true, given that we have observed our statistic.
- The p-value is *not* the probability of having observed our statistic, given that the null hypothesis is true. Rather, it is the probability of having observed our statistic, *or any more extreme statistic*, given that the null hypothesis is true.
- The p-value is *not* the probability that your procedure will falsely reject the null hypothesis, given that the null hypothesis is true.[6]

The moral of the story is: always be careful when quoting or interpreting p-values. In many circumstances, a better question to ask than "what is the p-value?" is "what is a plausible range for the size of the effect?" This question can be answered with a confidence interval.[7]

[5] If you are interested in these reasons, you should read up on the Neyman–Pearson school of hypothesis testing.

[6] To get a guarantee of this sort, you have to set up a pre-specified rejection region for your p-value (like 0.05), in which case the size of that rejection region—and not the observed p-value itself—can be interpreted as the probability that your procedure will reject the null hypothesis, given that the null hypothesis is true. As above: if you're interested, read about the Neyman–Pearson approach to testing.

[7] In this case, you could get a confidence interval by bootstrapping the difference in medians between the two groups of states.

Hypothesis testing in regression

To finish off this chapter, we will show how the permutation-testing framework can be used to answer questions about partial relationships in multiple regression modeling.

In a previous chapter, we asked the following question about houses in Saratoga, NY: what is the partial relationship between heating system type (gas, electric, or fuel oil) and sale price, once we adjust for the effect of living area, lot size, and the number of fireplaces? We fit a multiple regression model with these four predictors, which led to the following equation:

$$\begin{aligned}\text{Price} = \$29868 &+ 105.3 \cdot \text{SqFt} + 2705 \cdot \log(\text{Acres}) + 7546 \cdot \text{Fireplaces} \\ &- 14010 \cdot \mathbf{1}_{\{\text{fuel = electric}\}} - 15879 \cdot \mathbf{1}_{\{\text{fuel = oil}\}} + \text{Residual}\,.\end{aligned}$$

Remember that the baseline case here is gas heating, since it has no dummy variable. Our model estimated the premium associated with gas heating to be about $14,000 over electric heating, and about $16,000 over fuel-oil heating.

But are these differences due to heating-system type statistically significant, or could they be explained due to chance?

To answer this question, you could look at the confidence intervals for every coefficient associated with the heating-system variable, just as we learned to do in the chapter on multiple regression. The main difference is that before, we had one coefficient to look at, whereas now we have two: one dummy variable for fuel = electric, and one for fuel = oil. Two coefficients means two confidence intervals to look at.

Sometimes this strategy—that is, looking at the confidence intervals for all coefficients associated with a single variable—works just fine. For example, when the confidence intervals for all coefficients associated with a single variable are very far from zero, it's pretty obvious that the categorical variable in question is statistically significant.

But at other times, this strategy can lead to ambiguous results. In the context of the heating-system type variable, what if the 95% confidence interval for one dummy-variable coefficient contains zero, but the other doesn't? Or what if both confidence intervals contain zero, but just barely? Should we say that heating-system type is significant or not? This potential for ambiguous confidence intervals gets even worse when your categorical variable has more than just a few levels, because then there will be many more confi-

dence intervals to look at.

The core of the difficulty here is that we want to assess the significance of the heating-system variable itself, not the significance of any individual *level* of that variable. To assess the significance of the whole variable, with all of its levels, we'll use a permutation test. Specifically, we will compare two models:

- The *full model*, which contains variables for square footage, lot size, number of fireplaces, and heating system.
- The *reduced model*, which contains variables for square footage, lot size, and number of fireplaces, but not for heating system. We say that the reduced model is *nested* within the full model, since it contains a subset of the variables in the full model, but no additional variables.

As always, we must start by specifying H_0. Loosely speaking, our null hypothesis is that the reduced model provides an adequate description of house prices, and that the full model is needlessly complex. To be a bit more precise: the null hypothesis is that *there is no partial relationship* between heating system and house prices, once we adjust for square footage, lot size, and number of fireplaces. This implies that all of the *true* dummy variable coefficients for heating-system type are zero.

Next, we must pick a test statistic. A natural way to assess the evidence against the null hypothesis is to use improvement in R^2 under the full model, compared to the reduced model. This is the same quantity we look at when assessing the importance of a variable in an ANOVA table. The idea is simple: if we see a big jump in R^2 when moving from the reduced to the full model, then the variable we added (here, heating system) is important for predicting the outcome, and the null hypothesis of no partial relationship is probably wrong.

You might wonder here: why not use the coefficients on the dummy variables for heating-system type as test statistics? The reason is that there are two such coefficients (or in general, $K - 1$ coefficients for a categorical variable with K levels). But we need a single number to use as our test statistic in a permutation test. Therefore we use R^2: it is a single number that summarizes the predictive improvement of the full model over the reduced model.

Of course, even if we were to add a useless predictor to the reduced model, we would expect R^2 to go up, at least by a little bit, since the model would have more degrees of freedom (i.e. param-

Remember the four basic steps in a permutation test:

(1) Choose a null hypothesis H_0.

(2) Choose a test statistic Δ that is sensitive to departures from the null hypothesis.

(3) Repeatedly shuffle the predictor of interest and recalculate the test statistic after each shuffle, to approximate $P(\Delta \mid H_0)$, the sampling distribution of the test statistic T under the assumption that H_0 is true.

(4) Check whether the observed test statistic for your data, δ, is consistent with $P(\Delta \mid H_0)$.

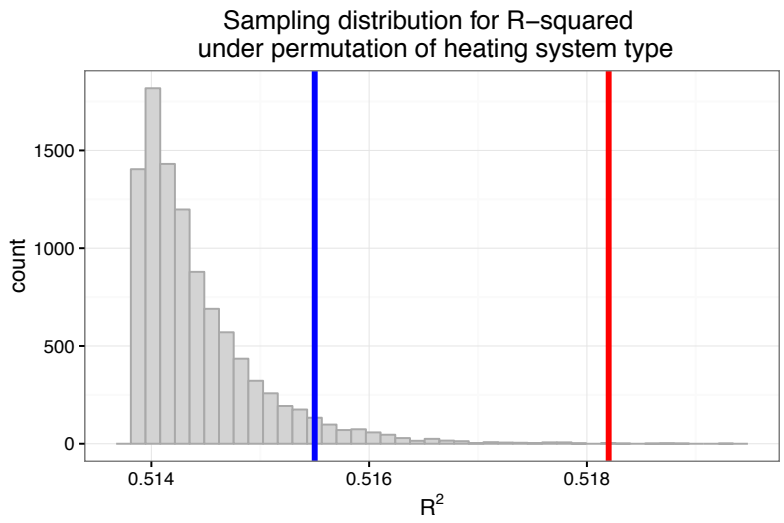


Figure 7.7: Sampling distribution of R^2 under the null hypothesis that there is no partial relationship between heating system and price after adjusting for effects due to square footage, lot size, and number of fireplaces. The blue vertical line marks the 95th percentile of the sampling distribution (and so corresponds to a rejection region at the 5% level). The red line marks the actual value of $R^2 = 0.518$ when we fit the full model by adding heating system to a model already containing the other three variables.

eters) that it can use to predict the observed outcome. Therefore, a more precise way of stating our null hypothesis is that, when we add heating system to a model already containing variables for square footage, lot size, and number of fireplaces, the improvement we see in R^2 could plausibly be explained by chance, even if this variable had no partial relationship with price.

To carry out a hypothesis test, we need to approximate the sampling distribution of R^2 under the null hypothesis. We will do so by repeatedly shuffling the heating system for every house (keeping all other variables the same), and re-fitting our model to each permuted data set. This breaks any partial relationship between heating system and price that may be present in our data. It tells us how big an improvement in R^2 we'd expect to see when fitting the full model, even if the null hypothesis were true.

This sampling distribution is shown in Figure 7.7, which was generating by fitting the model to 10,000 data sets in which the heating-system variable had been randomly shuffled, but where the response and the variables in the reduced model have been left alone. As expected, R^2 of the full model under permutation is always bigger than than the value of $R^2 = 0.513$ from the reduced model—but rarely by much. The blue line at $R^2 = 0.5155$ shows the 95th percentile of the sampling distribution (i.e. the critical value for a rejection region at the 5% level). The red line shows the actual value of $R^2 = 0.518$ from the full model fit the original

data set (i.e. with no shuffling). This test statistic falls far beyond the 5% rejection region. We therefore reject the null hypothesis and conclude that there is statistically significant evidence for an effect on price due to heating-system type.

One key point here is that we shuffled *only* heating-system type—or in general, whatever variable is being tested. We don't shuffle the response or any of the other variables. That's because we are interested in a partial relationship between heating-system type and price. Partial relationships are always defined with respect to a specific context of other control variables, and we have to leave these control variables as they are in order to provide the correct context for that partial relationship to be measured.

To summarize: we can compare any two nested models using a permutation test based on R^2, regardless of whether the variable in question is categorical or numerical. To do so, we repeatedly shuffle the extra variable in the full model—without shuffling either the response or the control variables (i.e. those that also appear in the reduced model). We fit the full model to each shuffled data set, and we track the sampling distribution of R^2. We then compare this distribution with the R^2 we get when fitting the full model to the *actual* data set. If the actual R^2 is a lot bigger than what we'd expect under the sampling distribution for R^2 that we get under the permutation test, then we conclude that the extra variable in the full model is statistically significant.

F tests and the normal linear regression model. Most statistical software will produce an ANOVA table with an associated p-value for all variables. These p-values are approximations to the p-values that you'd get if you ran sequential permutation tests, adding and testing one variable at a time as you construct the ANOVA table. To be a bit more specific, they correspond to something called an F test under the normal linear regression model that we met awhile back:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \,, \quad e_i \sim N(0, \sigma^2) \,.$$

You might want to revisit the discussion of the normal linear regression model starting on page 120. But the upshot is that an F test is conceptually similar to a permutation test based on R^2—and if you're happy with the assumption of normally distributed residuals, you can treat the p-values from these two tests as virtually interchangeable.[8]

[8] If you're not happy with this assumption, then you're better off with the permutation test.

8
Building predictive models

Building predictive models

Suppose you have a house in Saratoga, NY that you're about to put up for sale. It's a 1900 square-foot house on a 0.7-acre lot. It has 3 bedrooms, 2.5 bathrooms,[1] 1 fireplace, gas heating, and central air conditioning. The house was built 16 years ago. How much would you expect it to sell for?

[1] A half-bathroom has a toilet but no bath or shower.

Although we've been focusing on only a few variables of interest so far, our house-price data set actually has information on all these variables, and a few more besides. A great way to assess the value of the house is to use the available data to fit a multiple regression model for its price, given its features. We can then use this model to make a best guess for the price of a house with some particular combination of features—and, optionally, to form a prediction interval that quantifies the uncertainty of our guess.

We refer to this as the process of *building a predictive model.* Although we will still use multiple regression, the goal here is slightly different than in the previous examples. Here, we don't care so much about isolating and interpreting one particular partial relationship (like that between fireplaces and price). Instead, we just want the most accurate predictions possible.

The key principle in building predictive models is *Occam's razor*, which is the broader philosophical idea that models should be only as complex as they need to be in order to explain reality well. The principle is named after a medieval English theologian called Willam of Occam. Since he wrote in Latin, he put it like this: *Frustra fit per plura quod potest fieri per pauciora* ("It is futile to do with more things that which can be done with fewer.") A more modern formulation of Occam's razor might be the KISS rule: keep it simple, stupid.

In regression modeling, this principle is especially relevant for *variable selection*—that is, deciding which possible predictor variables to add to a model, and which to leave out. In this context,

Occam's razor is about finding the right set of variables to include so that we fit the data, without overfitting the data. Another way of saying this is that we want to find the patterns in the data, without memorizing the noise.

In this chapter, we'll consider two main questions:

(1) How can we measure the predictive power of a model?

(2) How can we find a model with good predictive power?

Measuring generalization error

To understand how we measure the predictive power of a regression model, we first need a bit of notation. Specifically, let's say that we have estimated a multiple regression model with p predictors $(x_1, x_2, \ldots, x_p)$ to some data, giving us coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$. Now we encounter a new case, not in our original data set. We'll let $x^\star = (x_1^\star, x_2^\star, \ldots, x_p^\star)$ be the predictor variables for this new case, and $y^\star$ denote the corresponding response. We will use the fitted regression model, together with $x^\star$, to make a prediction for $y^\star$:

$$\hat{y}^\star = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j^\star \,.$$

Our goal is to make the *generalization error*—that is, the difference between $y^\star$ and $\hat{y}^\star$—as small as possible, on average.

A natural way to measure the generalization error of a regression model is using a quantity called the *mean-squared predictive error*, or MSPE. The mean-squared predictive error is a property of a fitted model, not an individual data point. It summarizes the magnitude of the errors we typically make when we use the model to make predictions $\hat{y}^\star$ on new data:

$$\text{MSPE} = \text{Average value of } (y^\star - \hat{y}^\star)^2 \text{ when sampling new data points}\,.$$

Here a "new" data point means one that hasn't been used to fit the model. You'll notice that, in calculating MSPE, we square the prediction error $y^\star - \hat{y}^\star$ so that both positive and negative errors count equally.

Low mean-squared predictive error means that $y^\star - \hat{y}^\star$ tends to be close to zero when we sample new data points. This gives us a simple principle for building a predictive model: find the model (i.e. the set of variables to include) with the lowest mean-squared predictive error.

Estimating the the mean-squared predictive error

Conceptually, the simplest way to estimate the mean-squared predictive error of a regression model is to actually collect new data and calculate the average predictive error made by our model. Specifically, suppose that, after having fit our model in the first place, we go out there and collect $n^\star$ brand new data points, with responses $y_i^\star$ and predictors $(x_{i1}^\star, \ldots, x_{ip}^\star)$. We can then estimate the mean-squared predictive error of our model in two simple steps:

1. Form the prediction for each new data point:

$$\hat{y_i}^\star = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}^\star \,.$$

2. Calculate the average squared error of your predictions:

$$\widehat{\text{MSPE}}_{\text{out}} = \frac{1}{n^\star} \sum_{i=1}^{n^\star} (y_i^\star - \hat{y}_i^\star)^2 \,.$$

   Notice that we put a hat on MSPE, because the expression on the right-hand side is merely an *estimate* of the true mean-squared predictive error, calculated using a specific sample of new data points. (Calculating the *true* MSPE would require us, in principle, to average over all possible samples of new data points, which is obviously impractical.) We also use the subscript "out" to indicate that it is an *out-of-sample* measure—that is, calculated on new data, that falls outside of our original sample.

Conventionally, we report the square root of $\widehat{\text{MSPE}}_{\text{out}}$ (which is called *root mean-squared predictive error*, or RMSPE), because this has the same units as the original y variable. You can think of the RMSPE as the standard deviation of future forecasting errors made by your model.

Assuming your new sample size $n^\star$ isn't too small, these two steps are a nearly foolproof way to estimate the mean-squared predictive error of your model. The drawback, however, is obvious: you need a brand new data set, above and beyond the original data set that you used to fit the model in the first place. This new data set might be expensive or impractical to collect.

Thus we're usually left in the position of needing to estimate the mean-squared predictive error of a model, without having access to a "new" data set. For this reason, the usual practice is

make a *train/test split* of your data: that is, to randomly split your original data set into two subsets, called the *training* and *testing* sets.

- The training set is used only to fit ("train") the model—that is, to estimate the coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$.
- The testing set is used only to estimate the mean-squared predictive error of the model. It is not used at all to fit the model. For this reason, the testing set is sometimes referred to as the "hold-out set," since it is held out of the model-fitting process.

From this description, it should be clear that the training set plays the role of the "old" data, while the testing set plays the role of the "new" data.

This gives us a simple three-step procedure for choosing between several candidate models (i.e. different possible sets of variables to include).

(1) Split your data into training and testing sets.

(2) For each candidate model:

   A. Fit the model using the training set.
   B. Calculate $\widehat{\text{MSPE}}_{\text{out}}$ for that model using the testing set.

(3) Choose the model with the lowest value of $\widehat{\text{MSPE}}_{\text{out}}$.

Choosing the training and testing sets. A key principle here is that you must *randomly* split your data into a training set and testing set. Splitting your data nonrandomly—for example, taking the first 800 rows of your data as a training set, and the last 200 rows as a testing set—may mean that your training and testing sets are systematically different from one another. If this happens, your estimate of the mean-squared prediction error can be way off.

How much of the data should you reserve for the testing set? There are no hard-and-fast rules here. A common rule of thumb is to use about 75% of the data to train the model, and 25% to test it. Thus, for example, if you had 100 data points, you would randomly sample 75 of them to use for model training, and the remaining 25 to estimate the mean-squared predictive error. But other ratios (like 50% training, or 90% training) are common, too.

My general guideline is that the more data I have, the larger the fraction of that data I will use for training the predictive model.

Thus with only 100 data points, I might use a 75/25 split between training and testing; but with 10,000 data points, I might use more like a 90/10 split between training and testing. That's because estimating the model itself is generally harder than estimating the mean-squared predictive error.[2] Therefore, as more data accumulates, I like to preferentially allocate more of that data towards the intrinsically harder task of model estimation, rather than MSPE estimation.

[2] By "harder" here, I mean "subject to more sources of statistical error," as opposed to computationally more difficult.

Averaging over different test sets. It's a good idea to average your estimate of the mean-squared predictive error over several different train/test splits of the data set. This reduces the dependence of $\widehat{\text{MSPE}}_{\text{out}}$ on the particular random split into training and testing sets that you happened to choose. One simple way to do this is average your estimate of MSPE over many different random splits of the data set into training and testing sets. Somewhere between 5 and 100 splits is typical, depending on the computational resources available (more is better, to reduce Monte Carlo variability).

Another classic way to estimate MSPE it is to divide your data set into K non-overlapping chunks, called *folds*. You then average your estimate of MPSE over K different testing sets, one corresponding to each fold of the data. This technique is called *cross validation*. A typical choice of K is five, which gives us five-fold cross validation. So when testing on the first fold, you use folds 2-5 to train the model; when testing on fold 2, you use folds 1 and 3-5 to train the model; and so forth.

Can we use the original data to estimate the MSPE?

A reasonable question is: why do even we need a new data set to estimate the mean-squared prediction error? After all, our fitted model has residuals, $e_i = y_i - \hat{y}_i$, which tell us how much our model has "missed" each data point in our sample. Why can't we just use the residual variance, s_e^2, to estimate the MSPE? This approach sounds great on the surface, in that we'd expect the past errors to provide a good guide to the likely magnitude of future errors. Thus you might be tempted to use the *in-sample* estimate of MSPE, denoted

$$\widehat{\text{MSPE}}_{\text{in}} = s_e^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,,$$

where we recall that p is the number of parameters in the model.

Using $\widehat{\text{MSPE}}_{\text{in}}$ certainly removes the need to collect a new data set. This turns out, however, to be a false economy: $\widehat{\text{MSPE}}_{\text{in}}$ is usually too optimistic as an estimate of a model's generalization error. Practically speaking, this means the following. When we use $\widehat{\text{MSPE}}_{\text{in}}$ to quantify the *in-sample* error of a model, and then we actually go out and take new data to calculate the *out-of-sample* generalization error $\widehat{\text{MSPE}}_{\text{out}}$, we tend to discover that the out-of-sample error is larger—sometimes much larger! This is called overfitting, and it is especially likely to happen when the size of the data set is small, or when the model we're fitting is very complex (i.e. has lots of parameters).

An example

Let's see these ideas in practice, by comparing three predictive models for house prices in Saratoga, New York. Our models will draw from the following set of variables:

- lot size, in acres
- age of house, in years
- living area of house, in square feet
- percentage of residents in neighborhood with college degree
- number of bedrooms
- number of bathrooms
- number of total rooms
- number of fireplaces
- heating system type (hot air, hot water, electric)
- fuel system type (gas, fuel oil, electric)
- central air conditioning (yes or no)

We'll consider three possible models for price constructed from these 11 predictors.

Small model: price versus lot size, bedrooms, and bathrooms (4 total parameters, including the intercept).

Medium model: price versus all variables above, main effects only (14 total parameters, including the dummy variables).

Big model: price versus all variables listed above, together with all pairwise interactions between these variables (90 total parameters, include dummy variables and interactions).

Table 8.1 shows both $\widehat{\text{MSPE}}_{\text{in}}$ and $\widehat{\text{MSPE}}_{\text{out}}$ for these three models. To calculate $\widehat{\text{MSPE}}_{\text{out}}$, we used 80% of the data as a training

	In-sample RMSPE	Out-of-sample RMSPE	Difference
Small model: underfit	$76,144	$76,229	$85
Medium model: good fit	$65,315	$65,719	$403
Big model: overfit	$61,817	$71,426	$9,609

Table 8.1: In-sample versus out-of-sample estimates of the root mean-squared predictive error for three models of house prices in Saratoga, NY. The "difference" column shows the difference between the in-sample and out-of-sample estimates. The big model has a very large difference (over $9,000), indicating that the in-sample estimate is way too optimistic, and that the model is probably overfit to the data.

set, and the remaining 20% as a test set, and we averaged over 100 different random train/test splits of the data. The final column, labeled "difference," shows the difference between the in-sample and out-of-sample estimates of prediction error.

There are a few observations to take away from Table 8.1. The first is that that big model (with all the main effects and interactions) has the lowest in-sample error. With a residual standard deviation of $61,817, it seems nearly $3,500 more accurate than the medium model, which is next best. This is a special case of a very general phenomenon: a more complex model will always fit the data better, because it has more degrees of freedom to play with.

However, the *out-of-sample* measure of predictive error tells a different story. Here, the medium-sized model is clearly the winner. Its predictions on new data are off by about $65,719, on average, which is nearly $6,000 better than the big model

Finally, notice how severely degraded the predictions of the big model become when moving from old (in-sample) data to new (out-of-sample) data: about $9,600 worse, on average. This kind of degradation is a telltale sign of overfitting. The medium model suffers only a mild degradation in performance on new data, while the small model suffers hardly any degradation at all—although it's still not competitive on the out-of-sample measure, because it wasn't that good to begin with. This is also a special case of a more general phenomenon: *some* degradation in predictive performance on out-of-sample versus in-sample data is inevitable, but simpler models tend to degrade a lot less.

Figure 8.1 demonstrates this point visually. Starting from a very simple model of price (using only lot size as a predictor), we've added one variable or interaction at a time[3] from the list on page 168. For each new variable or interaction, we recalculated both the in-sample ($\widehat{\text{MSPE}}_{\text{in}}$) and out-of-sample ($\widehat{\text{MSPE}}_{\text{out}}$) estimates of the generalization error. As we add variables, the out-of-sample error initially gets smaller, reflecting a better fitting model that still generalizes well to new data. But after 15 or 20 variables,

[3] To be specific here, at each stage we added the single variable or interaction that most improved the fit of the model. See the next section on stepwise selection.

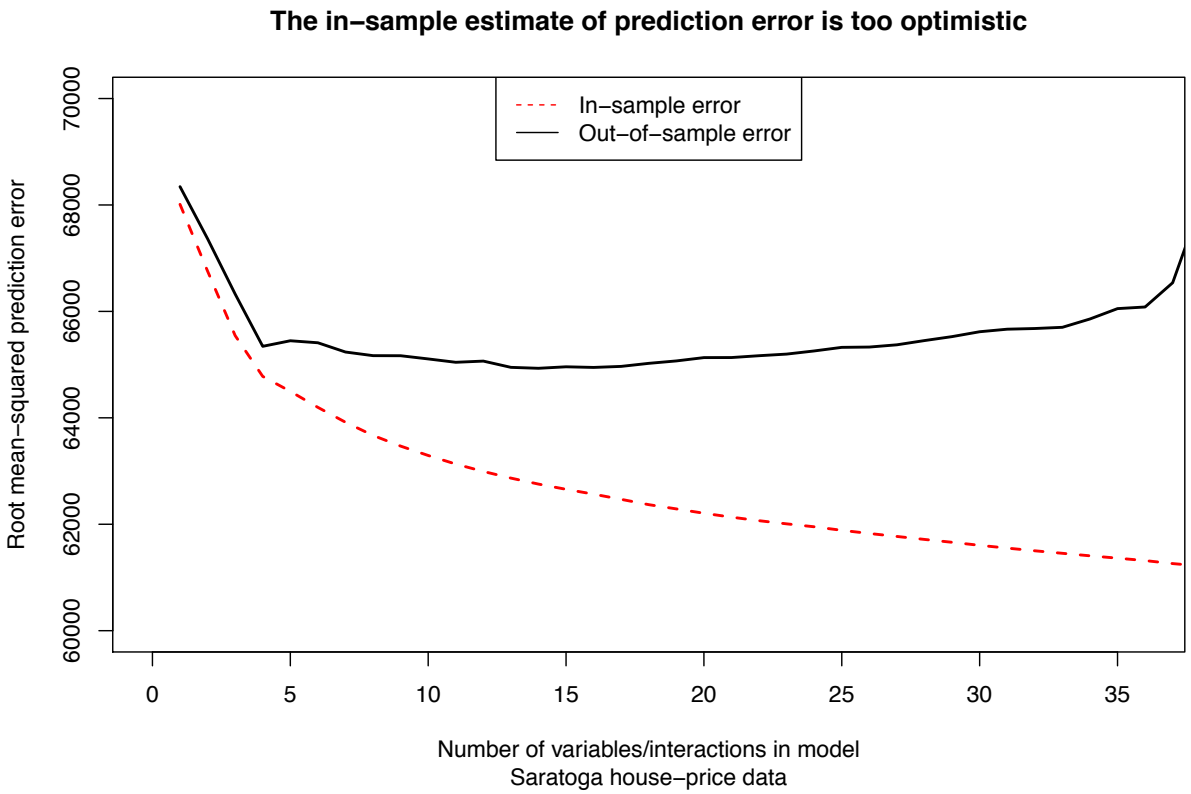


Figure 8.1: Starting from a small pricing model with just lot size as a predictor, we've added one variable or interaction at a time from the list on page 168. The red line shows the in-sample estimate of error, while the black line shows the out-of-sample estimate. After we add about 15 variables and interactions, the out-of-sample error starts to creep back up. Clearly the in-sample estimate is too optimistic, especially as the model gets more complex.

eventually the out-of-sample error starts creeping back up, due to overfitting. The in-sample estimate of error, however, keeps going down, falling even further out of line with the real out-of-sample error as we add more variables to the model.

In summary, you should remember the basic mantra of predictive model building: out-of-sample error is larger than in-sample error, especially for bigger models. If you care about minimizing out-of-sample error, you should always use an out-of-sample estimate of a model's MSPE, to make sure that you're not overfitting the original data. Our goal here should be obvious: to find the "turning point" in Figure 8.1, and to stop adding variables before we start overfitting.

Iterative model building via stepwise selection

Now that we know how to measure generalization error of a model, we're ready to introduce the overall steps in the process of building and using a predictive model from a set of candidate variables x_1, x_2, etc. We sometimes use the term *scope* to refer to this set of candidate variables.

The seemingly obvious approach is to fit all possible models under consideration to a training set, and to measure the generalization error of each one on a testing set. If you have only a few variables, this will work fine. For example, with only 2 variables, there are only $2^2 = 4$ possible models to consider: the first variable in, the second variable in, both variables in, or both variables out. You can fit and test those four models in no time. This is called *exhaustive enumeration*.

However, if there are lots of variables, exhaustive enumeration of all the models becomes a lot harder to do, for the simple fact that it's too exhausting—there are too many models to consider. For example, suppose we have 10 possible variables, each of which we could put in or leave out of the model. Then there are $2^{10} = 1024$ possible models to consider, since each variable could be in or out in any combination. That's painful enough. But if there are 100 possible variables, there are 2^{100} possible models to consider. That's *1 nonillion* models—about 10^{30}, or a thousand billion billion billion. This number is larger than the number of atoms in a human body.

You will quite obviously never be able to fit all these countless billions of models, much less compare their generalization errors on a testing set, even with the most powerful computer on earth. Moreover, that's for just 100 candidate variables *with main effects only*. Ideally, we'd like the capacity to build a model using many more candidate variables than that, or to include the possibility of interactions among the variables.

Thus a more practical approach to model-building is *iterative*: that is, to start somewhere reasonable, and to make small changes to the model, one variable at a time. Model-building in this iterative way is really a three-step process:

(1) Choose a baseline model, consisting of initial set of predictor variables to include in the model, including appropriate transformations, polynomial terms and interactions. Exploratory

data analysis (i.e. plotting your data) will generally help you get started here, in that it will reveal obvious relationships in the data. Then fit the model for y versus these initial predictors.

(2) Check the model. If necessary, change what variables are included, what transformations are used, etc.:

   (a) Are the assumptions of the model met? This is generally addressed using residual plots, of the kind shown in Figures 6.7 and 6.8. This allows you to assess whether the response varies linearly with the predictors, whether there are any drastic outliers, etc.

   (b) Are we missing any important variables or interactions? This is generally addressed by *adding* candidate variables or interactions to the model from step (1), to see how much each one improves the generalization error (MSPE).

   (c) Are there signs that the model might be overfitting the data? This is generally addressed by *deleting* variables or interactions that are already in the model, to see if doing so actually improves the model's generalization error.

   You may need to iterate these three questions a few times, going through many rounds of adding or deleting variables, before you're satisfied with your final model. Remember that the best way to measure generalization error is using an out-of-sample measure, like $\widehat{\text{MSPE}}_{\text{out}}$ derived from a train/test split of the data.

   Once you're happy with the model itself, then you can. . . .

(3) Use your fitted model to form predictions (and optionally, prediction intervals) for your new data points.

Can this process be automated?

In this three-step process, step 1 (start somewhere reasonable) and step 3 (use the final model) are usually pretty easy. The part where you'll spend the vast majority of your time and effort is step 2, when you consider many different possible variables to add or delete to the current model, and check how much they improve or degrade the generalization error of that model.

This is a lot easier than considering all possible combinations of variables in or out. But with lots of candidate variables, even this

iterative process can get super tedious. A natural question is, can it be automated?

The answer is: sort of. We can easily write a computer program that will automatically check for iterative improvements to some baseline ("working") model, using an algorithm called *stepwise selection*.

(1) from among a candidate set of variables (the scope), check all possible one-variable additions or deletions from the working model;

(2) choose the single addition or deletion that yields the best improvement to the model's generalization error. This becomes the new "working model."

(3) iteratively repeat steps (1) and (2) until no further improvement to the model is possible.

The algorithm terminates when it cannot find any one-variable additions or deletions that will improve the generalization error of the working model.

Some caveats. Stepwise selection tends to work tolerably well in practice. But it's far from perfect, and there are some important caveats. Here are three; the first one is minor, but the second two are pretty major.

First, if you run stepwise selection from two different baseline models, you will probably end up with two different final models. This tends not to be a huge deal in practice, however, because the two final models usually have similar mean-squared predictive errors. Remember, when we're using stepwise selection, we don't care too much about *which* combinations of variables we pick, as long as we get good generalization error. Especially if the predictors are correlated with each other, one set of variables might be just as good as another set of similar (correlated) variables.

Second, stepwise selection usually involves some approximation. Specifically, at each step of stepwise selection, we have to compare the generalization errors of many possible models. Most statistical software will perform this comparison *not* by actually calculating $\widehat{\text{MSPE}}_{\text{out}}$ on some test data, but rather using one of several possible heuristic approximations for MSPE. The most common one is called the AIC approximation:[4]

$$\widehat{\text{MSPE}}_{\text{AIC}} = \widehat{\text{MSPE}}_{\text{in}} \left(1 + \frac{p}{n}\right) = s_e^2 \left(1 + \frac{p}{n}\right) ,$$

[4] In case you're curious, AIC stands for "Akaike information criterion." If you find yourself reading about AIC on Wikipedia or somewhere similar, it will look absolutely nothing like the equation I've written here. The connection is via a related idea called "Mallows' C_p statistic," which you can read about here.

where n is the sample size and p is the number of parameters in the model.

The AIC estimate of mean-squared predictive error is not a true out-of-sample estimate, like $\widehat{\text{MSPE}}_{\text{out}}$. Rather, it is like an "inflated" or "penalized" version of the in-sample estimate, $\widehat{\text{MSPE}}_{\text{in}} = s_e^2$, which we know is too optimistic. The inflation factor of $(1 + p/n)$ is always larger than 1, and so $\widehat{\text{MSPE}}_{\text{AIC}}$ is always larger than $\widehat{\text{MSPE}}_{\text{in}}$. But the more parameters p you have relative to data points n, the larger the inflation factor gets. It's important to emphasize that $\widehat{\text{MSPE}}_{\text{AIC}}$ is just an approximation to $\widehat{\text{MSPE}}_{\text{out}}$. It's a better approximation than $\widehat{\text{MSPE}}_{\text{in}}$, but it still relies upon some pretty specific mathematical assumptions that can easily be wrong in practice.

The third and most important caveat is that, when using any kind of automatic variable-selection procedure like stepwise selection, we lose the ability to use our eyes and our brains each step of the way. We can't plot the residuals to check for outliers or violations of the model assumptions, and we can't ensure that the combination of variables visited by the algorithm make any sense, substantively speaking. It's worth keeping in mind that your eyes, your brain, and your computer are your three most powerful tools for statistical reasoning. In stepwise selection, you're taking two of these tools out of the process, for the sake of doing a lot of brute-force calculations very quickly.

None of these caveats are meant to imply that you *shouldn't* use stepwise selection—merely that you shouldn't view the algorithm as having God-like powers for discerning the single best model, or treat it as an excuse to be careless. You should instead proceed cautiously. Always verify that the stepwise-selected model makes sense and doesn't violate any crucial assumptions. It's also a good idea to perform quick a train/test split and compute $\widehat{\text{MSPE}}_{\text{out}}$ for your final model, just as a sanity check, to make sure that you're actually improving the generalization error versus your baseline model.

9
Understanding cause and effect

Statistical questions versus causal questions

WHY have some nations become rich while others have remained poor? Do small class sizes improve student achievement? Does following a Mediterranean diet rich in vegetables and olive oil reduce your risk of a heart attack? Does a "green" certification (like LEED, for Leadership in Energy and Environmental Design) improve the value of a commercial property?

Questions of cause and effect like these are, fundamentally, questions about *counterfactual statements.* A counterfactual is an if–then statement about something that has not actually occurred. For example: "If Colt McCoy had not been injured early in the 2010 National Championship football game, then the Texas Longhorns would have beaten Alabama." If you judge this counterfactual statement to be true—and who but the most hopelessly blinkered Crimson Tide fan doesn't?—then you might say that Colt McCoy's injury caused the Longhorns' defeat.

Statistical questions, on the other hand, are about correlations. This makes them fundamentally different from causal questions.

- Causal: "If we invested more money in our school system, how much faster would our economy grow?" Statistical: "In looking at data on a lot of countries, how are education spending and economic growth related?"

- Causal: "If I ate more vegetables than I do now, how much longer would I live?" Statistical: "Do people who eat a lot of vegetables live longer, on average, than people who don't?"

- Causal: "If we hire extra teachers at our school and reduce our class sizes, will our students' test scores improve?" Statistical: "Do students in smaller classes tend to have higher test scores?"

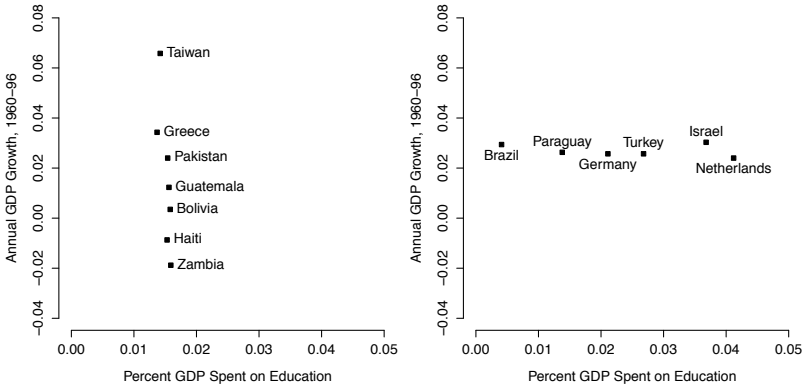


Figure 9.1: Two egregious examples of selective reporting.

Causal questions all invoke some kind of hypothetical intervention, where one thing is changed and everything else is held equal. In such a hypothetical intervention, there is no competing explanation for what might be causing the change we expect to see—in our economy, our lifespan, our students' test scores, a football game, or whatever outcome we're interested in.

Statistical questions, on the other hand, are about the patterns we observe in the real world. And the real world is rarely so simple as the hypothetical interventions we imagine. For example, people who eat more vegetables live longer—that's a clear pattern. But those same people also tend to exercise more, live in better housing, and have higher-status jobs. These other factors are *confounders*. A confounder is a competing explanation—some other factor correlated with both the "treatment" assignment (whether someone eats vegetables) and the response (lifespan). So in light of these confounders, how do we know it's the vegetables, rather than all that other stuff, that's making veggie-eaters live longer?

This is just a specific version of the general question we'll address in this chapter: under what circumstances can causal questions be answered using statistics?

Good evidence . . . and bad

Most of the cause-and-effect reasoning that you'll see out there in the real world is of depressingly poor quality. A common flaw is *cherry picking*: that is, pointing to data that seems to confirm some argument, while ignoring contradictory data.

Here's an example. In the left panel in Figure 9.1 we see a

group of seven countries that all spend around 1.5% of their GDP on education, but with very different rates of economic growth for the 37 years spanning 1960 to 1996. In the right panel, we see another group of six countries with very different levels of spending on education, but similar growth rates of 2–3%.

Both highly selective samples make it seem as though education and economic growth are barely related. If presented with the left panel alone, you'd be apt to conclude that the differences in growth rates must have been caused by something other than differences in education spending (of which there are none). Likewise, if presented with the right panel alone, you'd be apt to conclude that the large observed differences in education spending don't seem to have produced any difference in growth rates. The problem here isn't with the data—it's with the biased, highly selective *use* of that data.

This point seems almost obvious. Yet how tempting it is just to cherry pick and ignore the messy reality. Perhaps without even realizing it, we're all accustomed to seeing news stories that marshal highly selective evidence—usually even worse than that of Figure 9.1—on behalf of some plausible because-I-said-so story:

> [H]igher levels of education are critical to economic growth. . . . Boston, where there is a high proportion of college graduates, is the perfect example. Well-educated people can react more quickly to technological changes and learn new skills more readily. Even without the climate advantages of a city like San Jose, California, Boston evolved into what we now think of as an "information city." By comparison, Detroit, with lower levels of education, languished.[1]

[1] "Economic Scene." *New York Times* (Business section); August 5, 2004

And this from a reporter who presumably has no hidden agenda. Notice how the selective reporting of evidence—one causal hypothesis, two data points—lends an air of such graceful inevitability to what is a startlingly superficial analysis of the diverging economic fates of Boston and Detroit over the last half century.

Of course, most bad arguments are harder to detect than this howler from the New York Times. After all, using data to understand cause-and-effect relationships is hard. For example, consider the following summary of a recent neuroscience study:

> A study presented at the Society for Neuroscience meeting, in San Diego last week, shows people who start using marijuana at a young age have more cognitive shortfalls. Also, the more marijuana a person used in adolescence, the more trouble they had with focus and attention. "Early onset smokers

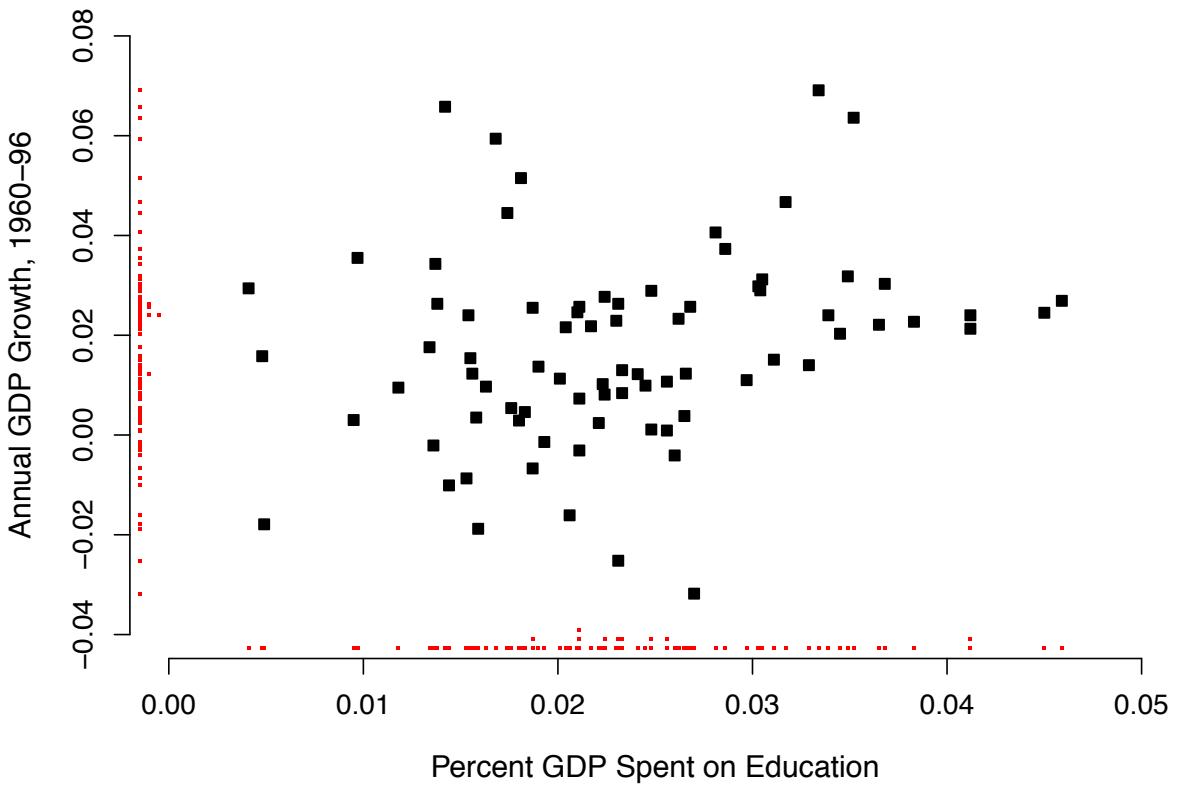


Figure 9.2: A scatter plot of GDP growth versus education spending for 79 countries. The tiny red dots clustered near the x and y axes are called *rug plots*. They are miniature histograms aligned with the axes of the predictor and the response.

have a different pattern of brain activity, plus got far fewer correct answers in a row and made way more errors on certain cognitive tests," says study author Staci Gruber.[2]

Did the marijuana smokers get less smart, or were the less-smart kids more likely to pick up a marijuana habit in the first place? It's an important question to consider in making drug policy, especially for states and countries where marijuana is legal. But can we know the answer on the basis of a study like this?

For another example, consider the bigger sample of countries in Figure 9.2, which provides a much more representative body of evidence on the GDP-versus-education story. This evidence takes the form of a scatter plot of GDP growth versus education spending for a sample of 79 countries worldwide. Notice the following two facts:

(1) Of the 29 countries that spent less than 2% of GDP on education, 18 fall below the median growth rate (1.58%).

[2] www.usatoday.com/yourlife/health/medical/pediatrics/2010-11-20-teendrugs22_ST_N.htm

(2) Of the 18 countries that spent more than 3% of GDP on education, 16 fall above the median growth rate.

These two facts, together with the upward trend in the scatter plot, suggest that economic growth and education spending are correlated. But this does not settle the causal question. For example, it might be that countries spend a lot on education because they are rich, rather the other way around.

The generic difficulty is that there are many different ways that two variables X and Y can appear correlated.

(1) One-way causality: the first domino falls, then the second; the rain falls, and the grass gets wet. (X causes Y directly.)

(2) Two-way causality: flowers and honey bees prosper together. (Both X and Y play a role in causing each other.)

(3) Common cause: People who go to college tend to get higher-paying jobs than those who don't. Does education directly lead to better economic outcomes? Or are a good education and a good job both just markers of a person's underlying qualities? (The role of X in causing Y is hard to distinguish from the role of C, which we may not have observed.)

(4) Common effect: either musical talent (X) or athletic talent (Y) will help you get into Harvard (Z). Among a population of Harvard freshmen, musical and athletic talent will thus appear negatively correlated, even if they are independent in the wider population. (X and Y both contribute to some common outcome C, inducing a correlation among a subset of the population defined by Z. This is often called Berkson's paradox; it is subtle, and we'll encounter it again.)

(5) Luck: the observed correlation is a coincidence.

This is the point where most books remind you that "correlation does not imply causation." Obviously. But if not to illuminate causes, what is the point of looking for correlations? Of course correlation does not imply causality, or else playing professional basketball would make you tall. But that hasn't stopped humans from learning that smoking causes cancer, or that lightning causes thunder, on the basis of observed correlations. The important question is: what distinguishes the good evidence-based arguments from the bad?

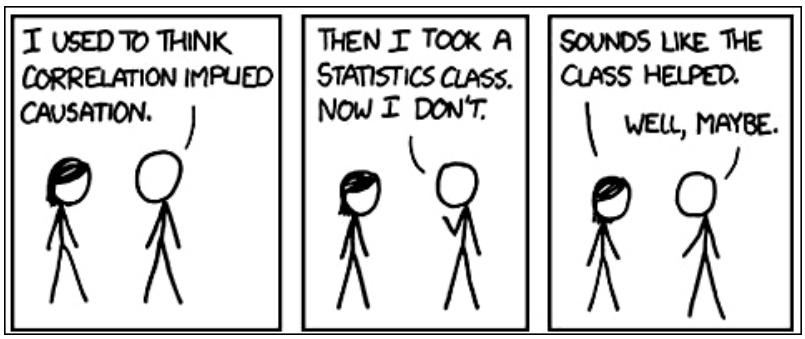


Figure 9.3: Originally published online at xkcd: http://xkcd.com/552/

Four common identification strategies

The key principle in using evidence to draw causal conclusions is that of a *balanced comparison*. To make things simple, we'll imagine that our predictor X is binary (i.e. has two groups), and we'll borrow the lingo of a clinical trial by referring to the two groups as the "treatment" and "control." To reach the conclusion that X causes Y, you must do two things: (1) *compare cases* in the treatment and control groups, to see how their Y values differ; and (2) *ensure balance*, by removing all other systematic differences between the cases in the treatment and control groups. Balance is crucial; it's what allows us to conclude that the differences in X (and not something else) cause the differences we observe in Y.

In general, there are four common ways to make a balanced comparison. These are often called *identification strategies*, in the sense that they are strategies for identifying a causal effect.

(1) Run a real experiment, randomizing subjects to the treatment and control groups. The randomization will ensure that, on average, there are no systematic differences between the two groups, other than the treatment.

(2) Find a natural experiment: that is, find a situation in which the the way that cases fall naturally into the treatment and control groups plausibly resembles a random assignment.

(3) Matching: artificially construct a balanced data set from an unbalanced data set, by explicitly matching treated cases with similar control cases, and discarding the cases without a good match. This will correct for lack of balance between control and treatment groups.

(4) Modeling: use multiple regression modeling to adjust for confounders and isolate a partial relationship between the response and the treatment of interest.

We'll take each of these four ideas in turn.

The power of experiment

THE idea of an experiment is simple. If you want to know what would happen if you intervened in some system, then you should intervene, and measure what happens. There is simply no better way to establish that one thing causes another.

Indeed, one kind of experiment—the randomized, controlled clinical trial—is one of the most important medical innovations in history. Suppose we want to establish whether a brand new cholesterol drug—we'll call it Zapaclot—works better than the old drug. Also suppose that we've successfully recruited a large cohort of patients with high cholesterol. We know that diet and genes play a role here, but that drugs can help, too. We express this as

$$\text{Cholesterol} \sim \text{Diet} + \text{Genes} + \text{Drugs}\,.$$

Interpret the plus sign as the word "and," not like formal addition: we're assuming that cholesterol depends upon diet, genes, and drugs, although we haven't said how. Of course, it's that third predictor in the model we care about; the first two, in addition to some others that we haven't listed, are potential confounders.

First, what not to do: don't proceed by giving Zapaclot to all the men and the old drug to all the women, or Zapaclot to all the marathon runners and the old drug to the couch potatoes. These highly non-random assignments would obviously bias any judgment about the relative effect of the new drug compared to the old one. We refer to this sort of thing as *selection bias*: that is, any bias in the selection of cases that receive the treatment.

To avoid selection bias, an experiment has two simple steps.

Randomize: randomly split the cohort into two groups, denoted the treatment group and the control group.

Intervene: allocate everyone in the treatment group to take the treatment (e.g. Zapaclot, the new drug), and everyone in the control group to take something else (e.g. the old drug or a placebo).[3]

[3] Everyone in the control group should be taking the *same* something else, whether it's the old drug or a placebo.

Randomize and intervene: a simple prescription, but the surest way to establish causality. The intervention allows you to pick up a difference between the new and old drug, if there's one to be found. The randomization ensures that other factors—even unknown factors, in addition to known ones like diet and lifestyle—do not lead us astray in our causal reasoning. The Latin phrase *ceteris paribus*, which translates roughly as "everything else being equal," is often used to describe such a situation. By randomizing and intervening, we have ensured that the only *systematic* difference between the groups is the treatment itself. The randomization gives us a balanced comparison.

This last point is crucial. It's not that diet, genes, and other lifestyle factors somehow stop affecting a patient's cholesterol level when we randomize and intervene. It's just that diet, genes, and lifestyle factors aren't correlated with the treatment assignment, and so they're balanced between the two groups, on average.

The need to avoid selection bias sounds obvious. But if selection bias in medical trials were not rigorously policed, then it would be easy for doctors to cherry pick healthy patients for newly proposed treatments. After all, a doctor who invents a new, seemingly effective form of treatment will almost surely become both rich and famous. As one physician reminisces:

> One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone successful operations for vascular reconstruction. At the end of the lecture, a young student at the back of the room timidly asked, "Do you have any controls?" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "Do you mean did I not operate on half of the patients?" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "Yes, that's what I had in mind." Then the visitor's fist really came down as he thundered, "Of course not. That would have doomed half of them to their death." God, it was quiet then, and one could scarcely hear the small voice ask, "Which half?"[4]

These last two words—"Which half?"—should echo in your mind whenever you are asked to judge the quality of evidence offered in support of a causal hypothesis. There is simply no substitute for a controlled experiment: not a booming authoritative voice, not even fancy statistics.

In fact, government regulators are so fastidious in their attention to possible selection biases that, in most real clinical trials, nei-

[4] Dr. E. Peacock, University of Arizona. Originally quoted in *Medical World News* (September 1, 1972). Reprinted pg. 144 of *Beautiful Evidence*, Edward Tufte (Graphics Press, 2006).

ther the doctors nor the patients are allowed to know which drug each person receives. Such a "double-blind" experiment avoids the possibility that patients might simply imagine that the the latest miracle drug has made them feel better, in a feat of unconscious self-deception called the placebo effect.

A placebo, from the Latin *placere* ("to please"), is a fake treatment designed to simulate the real one.

Some history

The notion of a controlled experiment was certainly around in pre-Christian times. The first chapter of the book of Daniel relates the tale of one such experiment. Daniel and his three friends Hananiah, Mishael, and Azariah arrive in the court of Nebuchadnezzar, the King of Babylon. They enroll in a Babylonian school, and are offered a traditional Babylonian diet. But Daniel wishes not to "defile himself with the portion of the king's meat, nor with the wine which he drank." He goes to Melzar, the prince of the eunuchs, who is in charge of the school. Daniel asks not to be made to eat the meat or drink the wine. But Melzar responds that he fears for Daniel's health if he were to let them follow some crank new-age diet. More to the point, Melzar observes, if the new students were to fall ill, "then shall ye make me endanger my head to the king."

So Daniel proposes a trial straight out of a statistics textbook:

> Prove thy servants, I beseech thee, ten days; and let them give us pulse to eat, and water to drink.
> Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants.[5]

[5] King James Bible, Daniel 1:12–13.

The King agreed. When Daniel and his friends were inspected ten days later, "their countenances appeared fairer and fatter in flesh" than all those who had eaten meat and drank wine. Suitably impressed, Nebuchadnezzar brings Daniel and his friends in for an audience, and he finds that "in all matters of wisdom and understanding," they were "ten times better than all the magicians and astrologers that were in all his realm."

As for a placebo-controlled trial, in which some of the patients are intentionally given a useless treatment (the "placebo"): that came much later.[6] The first such trial seems to have taken place in 1784. It was directed by none other than Benjamin Franklin, the American ambassador to the court of King Louis XVI of France. A German doctor by the name of Franz Mesmer had gained some degree of notoriety in Europe for his claim to have discovered a new force of nature that he called "magnétisme animal," and

[6] See "The Power of Nothing" in the December 12, 2011 edition of *The New Yorker* (pp. 30–6).

which was said to have magical healing powers. The demand for Dr. Mesmer's services soon took off among the ladies of Parisian high society, whom he would "Mesmerize" using a wild contraption involving ropes and magnetized iron rods.

Much to the king's dismay, his own wife, Marie Antoinette, was one of Mesmer's keenest followers. The king found the whole Mesmerizing thing frankly a bit dubious, and presumably wished for his wife to have nothing to do with the Herr Doctor's magnétisme animal. So he convened several members of the French Academy of Sciences to investigate whether Dr. Mesmer had indeed discovered a new force of nature. The panel included Antoine Lavoisier, the father of modern chemistry, along with Joseph Guillotin, whose own wild contraption was soon to put the King's difficulties with Mesmer into perspective. Under Ben Franklin's supervision, the scientists set up an experiment to replicate some of Dr. Mesmer's prescribed treatments, substituting non-magnetic materials—history's first placebo—for half of the patients. In many cases, even the patients in the control group would flail about and start talking in tongues anyway. The panel concluded that the doctor's method produced no effect other than in the patients' own minds. Mesmer was denounced as a charlatan, although he continues to exact his revenge via the dictionary.

A more recent and especially striking example of a placebo comes from Thomas Freeman, director of the neural reconstruction unit at Tampa General Hospital in Florida. Dr. Freeman performs placebo brain surgery. (You read that correctly.) According to the British Medical Journal,

> In the placebo surgery that he performs, Dr Freeman bores into a patient's skull, but does not implant any of the fetal nerve cells being studied as a treatment for Parkinson's disease. The theory is that such cells can regenerate brain cells in patients with the disease. Some colleagues decry the experimental method, however, saying that it is too risky and unethical, even though patients are told before the operation that they may or may not receive the actual treatment.[7]

[7] BMJ. 1999 October 9; 319(7215): 942

"There has been a virtual taboo of putting a patient through an imitation surgery," Dr. Freeman said. (Imagine that.) "This is the way to start the discussion." Freeman has performed 106 real and placebo cell transplant operations since 1992. Dr. Freeman argues that the medical history is littered with examples of unsafe and ineffective surgical procedures—think of that small voice at the back of the room, asking "which half?"—that were not tested

against a placebo and resulted in needless deaths, year after year, before doctors abandoned them.

Experimental evidence is the best kind of evidence

Let's practice here, by comparing two causal hypotheses arising from two different data sets. The first comes from a clinical trial in the 1980's on a then-new form of adjuvant chemotherapy for treating colorectal cancer, a dreadful disease that, as of 2015, has a five-year survival rate of only 60-70% in the developed world.

The trial followed a simple protocol. After surgical removal of their tumors, patients were randomly assigned to different treatment regimes. Some patients were treated with fluorouracil (the chemotherapy drug, also called 5-FU), while others received no follow-up therapy. The researchers followed the patients for many years afterwards and tracked which ones suffered from a recurrence of colorectal cancer.

The outcome of the trial are in Table 9.1, below. Among the patients who received chemotherapy, 39% (119/304) had relapsed by the end of the study period, compared with 57% of patients (177/315) in the group who received no therapy:

Chemotherapy?		Yes	No
Recurrence?	Yes	119	177
	No	185	138

Table 9.1: Data from: J. A. Laurie et. al. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. J. Clinical Oncology, 7:1447–56, 1989. There was also a third treatment arm of the study in which patient received a drug called levamisole, which isn't discussed here. Survival statistics on colorectal cancer from Cunningham et. al (2010). "Colorectal cancer." Lancet 375 (9719): 1030–47.

The evidence strongly suggests that the chemotherapy reduced the risk of recurrence by a substantial amount: the relative risk of a relapse under the treatment group is 0.7, with a 95% confidence interval of $(0.59, 0.83)$.

We can be confident that this evidence reflects causality, and not merely correlation, because patients were randomly assigned to the treatment and control groups. Randomization ensures *balance*: that is, it ensures that there are no systematic differences between the two groups with respect to any confounding factors that might be correlated with the patients' survival chances. This would obviously not be true if we had non-randomly assigned all the healthiest patients to the treatment group, and all the sickest patients to the control group.

It's worth emphasizing a key fact here. Randomization ensures balance both for the possible confounders that we can measure

(like a patient's age or baseline health status), as well as for the ones we might *not* be able to measure (like a patient's will to live). This is what makes randomization so powerful, and randomized experiments so compelling. We don't even have to know what the possible confounding variables are in order for the experiment to give us reliable information about the causal effect of the treatment. *Randomization balances everything*, at least on average.

Next, let's examine data from a study from the 1990's conducted in sub-Saharan Africa about HIV, another dreadful disease which, at the time, was spreading across the continent with alarming speed. Several studies in Kenya had found that men who were uncircumcised seemed to contract HIV in greater numbers. This set off a debate among medical experts about the extent to which this apparent association had a plausible biological explanation.

	Circumcised?	Yes	No
HIV positive?	Yes	105	85
	No	527	93

Table 9.2: Data from Tyndall et. al. Increased risk of infection with human immunodeficiency virus type 1 among uncircumcised men presenting with genital ulcer disease in Kenya. Clin. Infect. Dis. 1996 Sep; 23(3):449–53.

Table 9.2, above, shows some data from one of these studies, which found that among those recruited for the survey, 48% of uncircumcised men were HIV-positive, versus only 17% of circumcised men. The evidence seems to suggest that circumcision reduced a Kenyan man's chance of contracting HIV by a factor of 3.

Evaluating the evidence. If you suffer from colon cancer, should you get chemotherapy? Almost certainly: the researchers in the first study randomized and intervened, giving chemotherapy only to a random subset of patients. Unless you believe that the chemotherapy patients in this trial just happened to be much luckier than their peers, this result establishes that the reduction in recurrence must have been caused by the treatment.

But should all Kenyan men head straight to a surgeon? In this case we can't really be sure. The researchers in the second study neither randomized nor performed any snipping themselves. They merely ask whether each man was circumcised. It is therefore possible that they've been fooled by a confounder. To give one plausible example, a man's religious affiliation might affect both the likelihood that he is circumcised and the chances that he contracts HIV from unprotected sex. If that were true, the observed correlation between circumcisions and HIV rates might be simply

a byproduct of an imbalanced, unfair comparison, rather than a causal relationship.[8]

[8] The authors of the study were obviously aware of these possible confounders. They used a technique called logistic regression to attempt to account for some them and isolate the putative effect of circumcision on HIV infection. This is like our fourth method for making balanced comparisons: use a model to adjust for confounders statistically. See the original paper for details.

Natural experiments

A randomized, controlled experiment is the gold standard of evidence for a causal hypothesis. Yet many times an experiment is impossible, impractical, unethical, or too expensive in time or money. In these situations, it often pays to look for something called a *natural experiment*, also called a *quasi-experiment*. A natural experiment is not something that you, as the investigator, design. Rather, it is an "experiment" where nature seems to have done the randomization and intervention for you, thereby giving you the same type of balance between treatment and control groups that you'd expect to get out of a real experiment.

This idea is best understood by example. Suppose you want to study the effect of class size on student achievement. You reason that, in smaller classes, students can get more individual attention from the instructor, and that intructors will feel a greater sense of personal connection to their students. All else being equal, you believe that smaller class sizes will help students learn better.

A cheap, naïve way to study this question would be to compare the test scores of students in small classes to those of students in larger classes. Any of these confounders, however, might render such a comparison highly unbalanced, and therefore dubious: (1) students in need of remediation are sometimes put in very small classes; (2) highly gifted students are also sometimes put in very small classes; (3) richer school districts can afford both smaller classes and many other potential sources of instructional advantage; or (4) better teachers successfully convince their bosses to let them teach the smaller classes themselves.

An expensive, intelligent way to study this question would be to design an experiment, in conjunction with a scientifically inclined school district, that randomly assigned both teachers and students to classes of varying size. This would guarantee exogenous variation in class sizes. In fact, a few school systems have done exactly this. A notable experiment is Project STAR in Tennessee—an expensive, lengthy experiment that studied the effect of primary-school class sizes on high-school achievement, and showed that reduced class sizes have a long-term positive impact both on test scores and drop-out rates.[9]

[9] The original study is described in Finn and Achilles (1990). "Answers and Questions about Class Size: a Statewide Experiment." *American Educational Research Journal* 28, pp. 557–77

Question	Problem	Natural experiment	Lingering issues
Does being rich make people happy?	Even if richer people are happier on average, maybe happiness and success are the common effect of a third factor. Or maybe the rich grade on a different curve than the rest of us.	Compare a group of lottery winners with a similar group of people who played the lottery but didn't win.	Lottery winners may play the lottery far more often than people who played the lottery but didn't win, which might correlate with other important differences.
Does smoking increase a person's risk for Type-II diabetes?	People who smoke may also engage in other unhealthy behaviors at systematically different rates than non-smokers.	Compare before-and-after rates of diabetes in cities that recently enacted bans on smoking in public places.	Maybe the incidence of diabetes would have changed anyway.
Do bans on mobile phone use by drivers in school zones reduce the rate of traffic collisions?	Groups of citizens that enact such bans may differ systematically in their attitudes toward risk and behavior on the road.	Go to Texarkana, split by State Line Avenue. Observe what happens when Texas passes a ban and Arkansas doesn't.	There may still be systematic differences between the two halves of the city.

Table 9.3: Three hypothetical examples of natural experiments.

But suppose you are neither naïve nor rich, and yet still want to study the question of whether small class sizes improve test scores. If you're in search of a third way—one that's better than merely looking at correlations, yet cheaper than a full-fledged experiment—you might be interested to know the following fact about the Israeli school system.

> [I]n Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."[10]

[10] Angrist and Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press, p. 21

This is a lovely example of a natural experiment—something you didn't design yourself, but that is almost as good as if you had. The researchers in this study compared the students in a group of 40 ("control group," in one large class) versus the students in a group of 41 ("treatment group," split into two smaller classes). This is a plausibly random assignment: the "randomization mechanism" is whether a student fell into a peer group of 40 versus a peer group of 41, and we would not expect this difference to be confounded by anything else that might predict test scores. Therefore, if we see a big difference in performance between the

two groups, the most likely explanation is that class size caused the difference.

Some natural experiments, of course, are better than others. Consider the examples in Table 9.3, on page 188. For each one, ask yourself two questions. (1) What are the "treatment" and "control" groups? (2) How balanced are these two groups? (Said another way: how good is the quasi-randomization of cases to these groups?) Think carefully about each one, and you may begin to see "experiment" versus "non-experiment" as the black and white ends of a spectrum, with many shades of grey in between.

Matching

To estimate a causal effect by matching, we artificially construct a balanced data set out of an unbalanced one, by explicitly matching treated cases with similar control cases. We then compare the outcomes in treatment versus control groups, using only the balanced data set. This is most easily seen by example.

An example: the value of going green

For many years now, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious ("green") buildings. There are both ethical and economic forces at work here. To quote a recent report by Mercer, an investment-consulting firm, entitled "Energy efficiency and real estate: Opportunities for investors":

> Investing in energy efficiency has two intertwined virtues that make it particularly attractive in a world with a changing climate and a destabilized economy: It cuts global-warming greenhouse gas emissions and saves money by reducing energy consumption. Given that the built environment accounts for 39 percent of total energy use in the US and 38 percent of total indirect CO_2 emissions, real estate investment represents one of the most effective avenues for implementing energy efficiency.

This only scratches the surface. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. Every new project involves negotiating a trade-off between costs incurred and benefits realized over the lifetime of the building. In this context, the decision to invest in an eco-friendly building could pay off in at least four ways.

(1) Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.

(2) Green buildings are often associated with indoor environments that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify *ex ante*; you cannot simply ask an engineer in the same way that you could ask a question such as, "How much are these solar panels likely to save on the power bill?"

(3) Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.

(4) Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is "life-cycle analysis," which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market and comparing "green" versus "non-green" buildings. By "green," we mean that a commercial property has received some official certification, because its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks, as certified by outside engineers.[11]

[11] The two most common certifications are LEED and EnergyStar; you can easily find out more about these rating systems on the web, e.g. at www.usgbc.org.

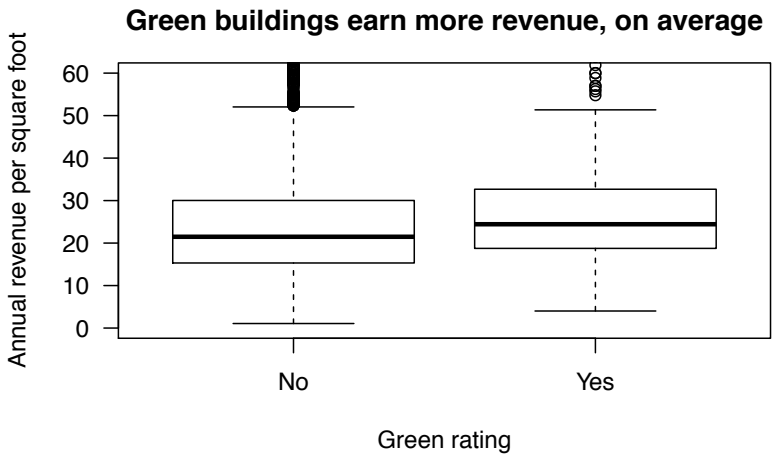


Figure 9.4: Green buildings seem to earn more revenue per square foot, on average, than non-green buildings.

Let's look at some data on 678 green-certified buildings in the United States, together with 6,298 non-green buildings in similar geographic areas. The boxplot above shows that, when we measure revenue by a building's rental rate per square foot per year, green buildings tend to earn noticeably higher revenue (mean = 26.97) than non-green buildings (mean = 24.51). That's a difference of $3.46 per square foot, or nearly a 15% market premium.

However, there's a problem with this comparison. As Table 9.4 shows, the green buildings tend to be newer than the non-green buildings, and are more likely to be "Class A" buildings.

So the important question is: do green buildings command a market premium *because* they are green, or simply because they are newer, better buildings in the first place? We can't tell by simply computing the average revenue in each group, because the

	Original data	
	Non-green buildings	Green buildings
Sample size	6928	678
Mean revenue/sq ft.	24.51	26.97
Mean age	49.2	23.9
Class A	37%	80%
Class B	48%	19%
Class C	15%	1%

Table 9.4: Covariate balance for the original data. Class A, B, and C are relative classifications within a specific real-estate market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

green ("treatment") and non-green ("control") groups are highly unbalanced with respect to some important confounders.

This is where matching comes in. Matching means constructing a balanced data set from an unbalanced one. It involves three steps:

(1) For each case in the treatment group, find the case in the control group that is the closest match in terms of confounding variables, and pair them up. Put these matched pairs into a new matched data set, and discard the cases in the original data set for which there are no close matches.

(2) Verify covariate balance for the matched data set, by checking that the confounders are well balanced between the treatment and control groups.

(3) Assuming that the confounders are approximately balanced, then compare the treatment outcomes with the control outcomes, using *only* the matched pairs.

Matching relies on a simple principle: compare like with like. In this example, that means if we have a 25-year-old, Class A building with a green rating, we try to find another 25-year old, Class A building without a green rating to compare it to.

In this particular example, once we've constructed the data set of matched pairs, the confounder variables are much more closely balanced between the treatment and control groups. A comparison of revenue rates for this matched data set makes the premium for green buildings look a lot smaller: $26.97 versus $25.94, or about a 4% premium. Compare that with the 15% premium we estimated from the original, unmatched data.

	Matched data	
	Non-green buildings	Green buildings
Sample size	678	678
Mean revenue/sq ft.	25.94	26.97
Mean age	23.9	23.9
Class A	80%	80%
Class B	19%	19%
Class C	1%	1%

Table 9.5: Covariate balance for the matched data.

How do we actually find matches? The nitty-gritty algorithmic details of actually finding good matched pairs of cases are best left to the experts who write the software for these things. The two most common types of matching are called *nearest-neighbor search* and *propensity-score matching*; follow the links if you'd like to know more. In R, the package `MatchIt` uses propensity-score matching as a default; this is a very commonly used algorithm in real-world data analysis. In addition, the paper linked here[12] has a much more detailed overview of different matching methods.

[12] "Matching Methods for Causal Inference: A Review and a Look Forward." Elizabeth A. Stuart, *Statistical Science*, 2010.

Matching isn't a silver bullet: a bigger example

If you've ever been admitted to the intensive-care unit at a hospital, you may have undergone a diagnostic procedure called right heart catheterization, or RHC. RHC is used to see how well a patient's heart is pumping, and to measure the pressures in that patient's heart and lungs. RHC is widely believed to be helpful, since it allows the doctor to directly measure what's going on inside a patient's heart. But it is an invasive procedure, since it involves inserting a small tube (the catheter) into the right side of your heart, and then passing that tube through into your pulmonary artery. It therefore poses some risks—for example, excessive bleeding, partial collapse of a lung, or infection.

A natural question is: do the diagnostic benefits of RHC outweigh the possible risks? But this turns out to be tricky to answer. The reason is that doctors would not consider it ethical to run a randomized, controlled trial to see if RHC improves patient outcomes. As the authors of one famous study from the 1990s pointed out:[13]

> Many cardiologistics and critical care physicians believe that the direct measurement of cardiac function provided by right heart catheterization (RHC) . . . is necessary to guide therapy for certain critically ill patients, and that such management leads to better patient outcomes. While the benefit of RHC has not been demonstrated in a randomized controlled trial (RCT), the popularity of this procedure, and the widespread belief that it is beneficial, make the performance of an RCT difficult. Physicians cannot ethically participate in such a trial or encourage a patient to participate if convinced the procedure is truly beneficial.

[13] "The effectiveness of right heart catheterization in the initial care of critically ill patients." Connors et. al. *Journal of the American Medical Association*. 1996 Sep 18; 276(11):889-97.

We're therefore left with only observational data on the effectiveness of RHC—which, on the surface, doesn't look good! Here's

	Original data		Matched data	
	No RHC	RHC	No RHC	RHC
Sample size	3551	2184	2184	2184
180-day survival rate	0.370	0.320	0.354	0.320
mean APACHE score	50.934	60.739	57.643	60.739
Trauma	0.005	0.016	0.008	0.016
Heart attack	0.030	0.043	0.036	0.043
Congestive heart failure	0.168	0.195	0.209	0.195
Sepsis	0.148	0.321	0.24	0.321

Table 9.6: A before-and-after table of summary statistics showing covariate balance for the observational study on right-heart catheterization. The entries for trauma, heart attack, etc. show rates of these complications in the two groups. The left half of the table shows the original data set, while the right half shows the matched data set.

the data from the study quoted above, showing that critically ill patients undergoing RHC actually have a *worse* 180-day survival rate (698/2184, or 32%) than patients not undergoing RHC (1315/3551, or 37%):

	No RHC	RHC
Survived 180 days	1315	698
Died within 180 days	2236	1486

What's going on here? Should we conclude that right heart catheterization is actually killing people, and that the doctors are all just plain wrong about its putative benefits?

Not so fast. The problem with this conclusion is that the treatment (RHC) and control (no RHC) groups are heavily unbalanced with respect to baseline measures of health. Put simply, the patients who received RHC were a lot sicker to begin with, so it's no surprise that they have a lower 6-month survival rate. To cite a few examples: the RHC patients were three times more likely to have suffered acute trauma, 50% more likely to have had a heart attack, and 16% more likely to be suffering from congestive heart failure. The RHC patients also had an average APACHE score that was 10 points higher than the non-RHC patients.[14] The left half of Table 9.6 shows these rates of various complications for the two groups in the original data set. They're quite different, implying that the survival rates of these two groups cannot be fairly compared.

And what about after matching? Unfortunately, Table 9.6 shows that, even after matching treatment cases with controls having similar complications, the RHC group still seems to have a lower

[14] The APACHE score is a composite severity-of-disease score used by hospital ICUs to estimate which patients have a higher risk of death. Patients with higher numbers have a higher risk of death.

survival rate. The gap looks smaller than it did before, on the unmatched data—a 32% survival rate for RHC patients, versus a 35.4% survival rate for non-RHC patients—but it's still there.

Again we find ourselves asking: what's going on? Is the RHC procedure actually killing patients? Well, it might be, at least indirectly! The authors of the study speculate that one possible explanation for this finding is "that RHC is a marker for an aggressive or invasive style of care that may be responsible for a higher mortality rate." Given the prevalance of overtreatment within the American health-care system, this is certainly plausible.

But we can't immediately jump to that conclusion on the basis of the matched data. In fact, this example points to a couple of basic difficulties with using matching to estimate a causal effect.

The first (and most important) difficulty is that *we can't match on what we haven't measured.* If there is some confounder that we don't know about, then we'll never be able to make sure that it's balanced between the treatment and control groups within the matched data. This is why experiments are so much more persuasive: because they also ensure balance for unmeasured confounders. The authors of the study acknowledge as much, writing:

> A possible explanation is that RHC is actually beneficial and that we missed this relationship because we did not adequately adjust for some confounding variable that increased both the likelihood of RHC and the likelihood of death. As we found in this study, RHC is more likely to be used in sicker patients who are also more likely to die.

Another possible explanation is that we simply haven't been able to match treatment cases with control cases very effectively. The right half of Table 9.6 shows that covariate balance for the matched data is noticeably better than for the unmatched data, but it's not perfect. We still see some small differences in complication rates and APACHE scores between the treatment and control group. There are two main reasons for this.

(1) First, and most importantly, although finding a match on one or two variables is relatively easy, finding a match on several variables is pretty hard. Think of this in terms of your own life experience—for example, in seeking a spouse or partner. It probably isn't too hard to find someone who's a good match for you in terms of your interests and your sense of humor. But if you require that this person *also* match you in terms of age, career, education, home town, height, weight, looks,

and favorite sport, then you're a lot less likely to find a match. *Picky people are less likely to find a satisfying match in life.* For this same reason, it's unlikely that we'll be able to find an exact match for each treatment case in a matching problem, especially with lots of possible confounders.

(2) Second, finding matches for cases with rare confounders is especially hard—by definition, since the confounder is rare!

These two points underline a basic difficulty with matching: perfect matches usually don't exist, and we have no choice but to accept approximate matches. In practice, therefore, we give up on the requirement that every single pair of matched observations is similar in terms of all possible confounders, and settle for having matched groups that are similar in their confounders, *on average*. That's why it's so important to check the covariate balance after finding matched pairs, to make sure that there's nothing radically different between the two groups.

Model-based statistical adjustment

A fourth identification strategy for estimating a causal effect is to build a regression model. If some important (and quite strong) assumptions are met, then such a model is capable of isolating a causal relationship between predictor and response, by adjusting for the effects of confounders *statistically*, rather than experimentally. You may have heard this process described as "statistical control" or "statistical correction," both in the popular media and in scientific publications:

- "Schatz's numbers are unique in that they evaluate each play against the league average for plays of its type, adjust for the strength of the opponents' defense, and even try to divide credit for a given play among teammates."[15]
- "The committee concluded that a statistical adjustment of the 1990 census leads to an improvement of the counts."[16]
- "Further adjustment for weight change and leukocyte count attenuated these risks substantially."[17]

Estimating a causal effect using a regression model is, in principle, no different than estimating a partial relationship, which we've already learned how to do:

[15] "Pigskin Pythagoras: A guy from Framingham tries to remake the muddy field of football statistics." *Boston Globe*, February 1, 2004

[16] "Judge must decide on census adjustment." *Chicago Tribune*, 6/8/1992

[17] "Smoking, Smoking Cessation, and Risk for Type 2 Diabetes Mellitus: A Cohort Study." *Annals of Internal Medicine*, January 4, 2010

(1) Build a multiple regression model for the outcome (y) versus the predictor of interest (x) and other possible confounders;

(2) Interpret the coefficient on the x variable of interest as the partial linear relationship between y and x, holding confounders constant.

The key question is: under what circumstances can we interpret the partial relationship in a multiple regression model as the *causal* effect of x on y? By *causal effect*, you should think in terms of the counterfactuals we entertained at the beginning of the chapter: *if* I were to intervene and change x by one unit, holding all other variables constant, *then* how much would y change on average?

There are three important assumptions that must be met in order to give a causal interpretation to a regression coefficient. First, you must have included all confounding variables (that is, variables that have a causal effect on both the treatment assignment and the outcome) in the model. Second, the model must be correct. In this context, "correct" means that you have included the right interactions among confounding variables, and that you have specified the right functional form of the model (linear, power law, etc.). Finally, you must *not* include any post-treatment effects as covariates in the model. A post-treatment effect is something causally "downstream" from the treatment variable, and that becomes known only as a result of receiving or not receiving the treatment. This is a subtle point, and we won't discuss it in detail. But the important thing is: include those confounders, and *only* those confounders, that affect the allocation of cases to the treatment and control groups.

If, and only if, these three assumptions about your model are true, then the regression coefficient of y on x has a causal interpretation. If, on the other hand, there are any unmeasured confounders affecting your x and y variable, then the coefficient of y on x measures association, not causation. This is called *omitted-variable bias*.[18]

[18] Or *lurking-variable bias.*

Another way of saying this is that *if* the possible confounders are all observed, then accurately estimating the causal effect of x and y really just boils down to modeling the data well, and not using that model to extrapolate beyond the range of available data. However, the assumption that we've observed all relevant confounders, and can therefore adjust for them appropriately, is very strong. It's also unverifiable using the data; as with matching,

you have to believe this assumption, and convince people of it, on extrinsic grounds.

Using regression analysis to estimate causal effects is a big, serious topic. Here are two full books about it:

- *Causality*, by Judea Pearl
- *Observational Studies*, by Paul Rosenbaum

For some additional, more easily digestible advice on choosing which covariates to include in a causal model, see Chapter 17 of Daniel Kaplan's book on statistical modeling.[19]

[19] Kaplan also has a good explanation for why it's not a good idea to include post-treatment effects (i.e. variables causally downstream of the treatment) as covariates in a regression model.

Matching versus regression, or matching and regression?

We've seen that it's easiest to infer causality if the cases in the treatment group are comparable to those in the control group. One way to do this is via matching: explicitly constructing a balanced data set from an unbalanced one. Another way to do this is via regression: adjust for confounders using a statistical model, so that we can evaluate the partial relationship between treatment and response, holding confounders constant.

This makes it sound as though regression and matching are competing identification strategies for causal inference. Sociologically speaking, there is certainly some truth to this, in that some people tend to use matching more often, and others tend to use regression more often. So which one should *you* use?

In the real world, if you're going to use only one strategy or the other, my advice is to use matching, mainly for three reasons:

(1) Matching is a lot easier for non-experts to understand, since you can point to the matched treatment and control groups and show that they are visibly balanced with respect to observed confounders. In other words, the nature of the "balanced comparison" being made via matching is much more transparent than the idea of a partial slope in a regression model. This will make it easier for you to convince others of your conclusions.

(2) Matching is a bit more robust than regression, at least in their "off the shelf" versions. The regression-based approach to causal inference relies on a whole bunch of hard-to-verify assumptions: linearity, all necessary interactions included, and so forth. By comparison, it's a lot easier to verify covariate balance using before-and-after tables of summary

statistics. (Of course, neither method is robust to unmeasured confounders—only an experiment can fix that problem.)

(3) Unwarranted extrapolations are more apparent when matching than when using regression. Suppose that the treatment and control groups have highly nonoverlapping distributions of confounders—for example, that most the men are in the treatment group and most of the women in the control group. In such cases, the data are inherently limited in what they can tell us about the treatment–response relationship in this region of nonoverlap (i.e. how the treatment will work for women). This lack of overlap will be obvious if you use matching, because you'll still have drastic post-match covariate imbalances that will stick out like a sore thumb. But the lack of overlap will be less obvious if you throw all the confounders into a multiple regression model without plotting your data.

In summary, it's easier to convince others with matching, and easier to fool yourself with regression. These aren't intrinsic *statistical* advantages to matching; they are merely *practical* advantages worth keeping in mind.

It turns out, however, that there's no need to choose between matching and regression. Better still is to use both matching *and* regression, to get better estimates of causal effects than either technique is capable of getting on its own. In other words: first run matching to get an approximately balanced data set. Then run a regression model for the response versus the treatment variable and the confounders, to correct for minor imbalances in the matched data set. Under this approach, the primary role of matching is to correct for major covariate imbalances between the groups, while the primary role of regression is to model the treatment–response relationship in a way that adjusts for any minor confounding that remains in the matched data set.

There's one other major advantage of using matching and regression together. By fitting a regression model to a matched data set, you are able to search for interactions *between* the treatment variable and possible confounders. For example, what if the treatment effect is different for men than for women? You can discover this kind of modulating effect much more easily using a regression model than you can with matching alone.

Matching and regression make for an excellent pair. There's rarely a good reason to use just one or other!

10
Generalized linear models

Binary responses

In many situations, we would like to predict the outcome of a binary event, given some relevant information:

- Given the pattern of word usage and punctuation in an e-mail, is it likely to be spam?
- Given the temperature, pressure, and cloud cover on Christmas Eve, is it likely to snow on Christmas Day?
- Given a person's credit history and income, is he or she likely to default on a mortgage loan?

In all of these cases, the y variable is the answer to a yes-or-no question. This is a bit different to the kinds of problems we've become used to seeing, where the response is a real number.

Nonetheless, we can still use regression for these problems. Let's suppose, for simplicity's sake, that we have only one predictor x, and that we let $y_i = 1$ for a "yes" and $y_i = 0$ for a "no." One naïve way of forecasting y is simply to plunge ahead with the basic, one-variable regression equation:

$$\hat{y}_i = \mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

Since y_i can only take the values 0 or 1, the expected value of y_i is simply a weighted average of these two cases:

$$\begin{aligned} \mathrm{E}(y_i \mid x_i) &= 1 \cdot P(y_i = 1 \mid x_i) + 0 \cdot P(y_i = 0 \mid x_i) \\ &= P(y_i = 1 \mid x_i) \end{aligned}$$

Therefore, the regression equation is just a linear model for the conditional probability that $y_i = 1$, given the predictor x_i:

$$P(y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i \,.$$

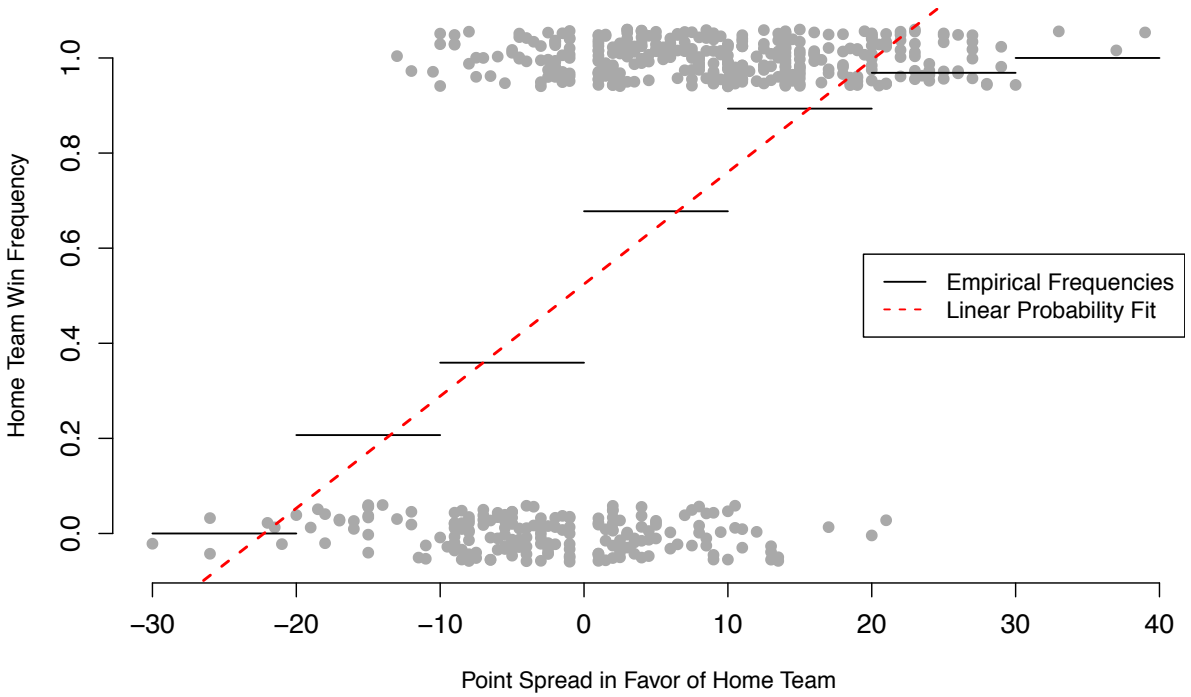


Figure 10.1: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

This model allows us to plug in some value of x_i and read off the forecasted probability of a "yes" answer to whatever yes-or-no question is being posed. It is often called the linear probability model, since the probability of a "yes" varies linearly with x.

Let's try fitting it to some example data to understand how this kind of model behaves. In Table 10.1 on page 203, we see an excerpt of a data set on 553 men's college-basketball games. Our y variable is whether the home team won ($y_i = 1$) or lost ($y_i = 0$). Our x variable is the Las Vegas "point spread" in favor of the home team. The spread indicates the betting market's collective opinion about the home team's expected margin of victory—or defeat, if the spread is negative. Large spreads indicate that one team is heavily favored to win. It is therefore natural to use the Vegas spread to predict the probability of a home-team victory in any particular game.

Figure 10.1 shows each of the 553 results in the data set. The

home-team point spread is plotted on the x-axis, while the result of the game is plotted on the y-axis. A home-team win is plotted as a 1, and a loss as a 0. A bit of artificial vertical jitter has been added to the 1's and 0's, just so you can distinguish the individual dots.

The horizontal black lines indicate empirical win frequencies for point spreads in the given range. For example, home teams won about 65% of the time when they were favored by more than 0 points, but less than 10. Similarly, when home teams were 10–20 point underdogs, they won only about 20% of the time.

Finally, the dotted red line is the linear probability fit:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.524435   0.019040   27.54   <2e-16 ***
spread      0.023566   0.001577   14.94   <2e-16 ***
---
Residual standard error: 0.4038 on 551 degrees of freedom
Multiple R-squared: 0.2884
```

This is the result of having regressed the binary y_i's on the point spreads, simply treating the 1's and 0's as if they were real numbers. Under this model, our estimated regression equation is

$$\mathrm{E}(y_i \mid x_i) = P(y_i = 1 \mid x_i) = 0.524 + 0.024 \cdot x_i \,.$$

Plug in an x, and read off the probability of a home-team victory. Here, we would expect the intercept to be 0.5, meaning that the home team should win exactly 50% of the time when the point spread is 0. Of course, because of sampling variability, the estimated intercept $\widehat{\beta}_0$ isn't exactly 0.5. But it's certainly close—about 1 standard error away.

The linear probability model, however, has a serious flaw. Try plugging in $x_i = 21$ and see what happens:

$$P(y_i = 1 \mid x_i = 21) = 0.524 + 0.024 \cdot 21 = 1.028 \,.$$

We get a probability larger than 1, which is clearly nonsensical. We could also get a probability less than zero by plugging in $x_1 = -23$:

$$P(y_i = 1 \mid x_i = -23) = 0.524 - 0.024 \cdot 23 = -.028 \,.$$

The problem is that the straight-line fit does not respect the rule that probabilities must be numbers between 0 and 1. For many values of x_i, it gives results that aren't even mathematically legal.

Game	Win	Spread
1	0	-7
2	1	7
3	1	17
4	0	9
5	1	-2.5
6	0	-9
7	1	10
8	1	18
9	1	-7.5
10	0	-8
⋮		
552	1	-4.5
553	1	-3

Table 10.1: An excerpt from a data set on 553 NCAA basketball games. "Win" is coded 1 if the home team won the game, and 0 otherwise. "Spread" is the Las Vegas point spread in favor of the home team (at tipoff). Negative point spreads indicate where the visiting team was favored.

Link functions and generalized linear models

THE PROBLEM can be summarized as follows. The right-hand side of the regression equation, $\beta_0 + \beta_1 x_i$, can be any real number between $-\infty$ and ∞. But the left-hand side, $P(y_i = 1 \mid x_i)$, must be between 0 and 1. Therefore, we need some transformation g that takes an unconstrained number from the right-hand side, and maps it to a constrained number on the left-hand side:

$$P(y_i \mid x_i) = g(\beta_0 + \beta_1 x_i) \,.$$

Such a function g is called a *link function*; a model that incorporates such a link function is called a *generalized linear model*, or GLM. The part inside the parentheses $(\beta_0 + \beta_1 x_i)$ is called the *linear predictor*.

We use link functions and generalized linear models in most situations where we are trying to predict a number that is, for whatever reason, constrained. Here, we're dealing with probabilities, which are constrained to be no smaller than 0 and no larger than 1. Therefore, the function g must map real numbers on $(-\infty, \infty)$ to numbers on $(0, 1)$. It must therefore be shaped a bit like a flattened letter "S," approaching zero for large negative values of the linear predictor, and approaching 1 for large positive values.

Figure 10.2 contains the most common example of such a link function. This is called the *logistic* link, which gives rise to the *logistic regression model*:

$$P(y_i = 1 \mid x_i) = g(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \,.$$

Think of this as just one more transformation, like the logarithm or powers of some predictor x. The only difference is that, in this case, the transformation gets applied to the whole linear predictor at once. The logistic regression model is often called the logit model for short.[1]

[1] The "g" in "logit" is pronounced softly, like in "gentle" or "magic."

With a little bit of algebra, it is also possible to isolate the linear predictor $\beta_0 + \beta_1 x_i$ on one side of the equation. If we let p_i denote

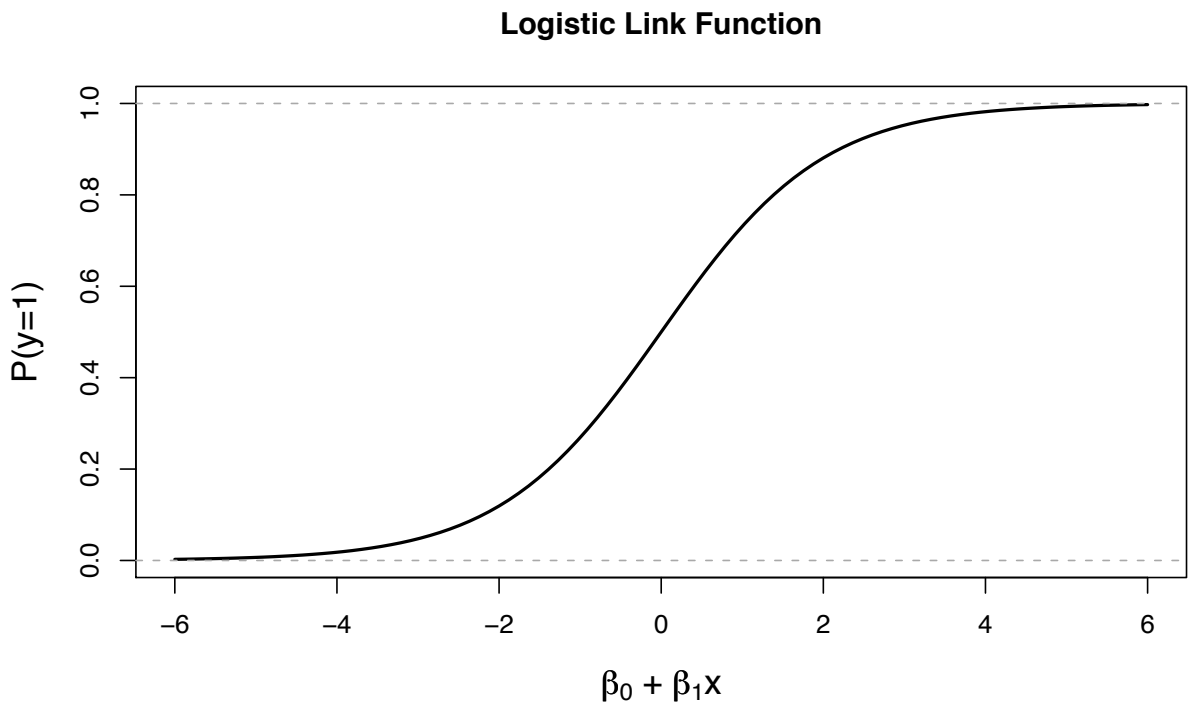


Figure 10.2: The logistic link function.

the probability that $y_i = 1$, given x_i, then

$$
\begin{aligned}
p_i &= \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \\
p_i + p_i e^{\beta_0+\beta_1 x_i} &= e^{\beta_0+\beta_1 x_i} \\
p_i &= (1-p_i)e^{\beta_0+\beta_1 x_i} \\
\log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 x_i
\end{aligned}
$$

Since $p_i = P(y_i = 1 \mid x_i$, we know that $1 - p_i = P(y_i = 0 \mid x_i)$. Therefore, the ratio $p_i/(1 - p_i)$ is the odds in favor of the event $y_i = 1$, given the predictor x_i. Thus the linear predictor $\beta_0 + \beta_1 x_i$ (on the right-hand side of the last equation) gives us the logarithm of the odds in favor of success ($y_i = 1$), on the left-hand side of the last equation.

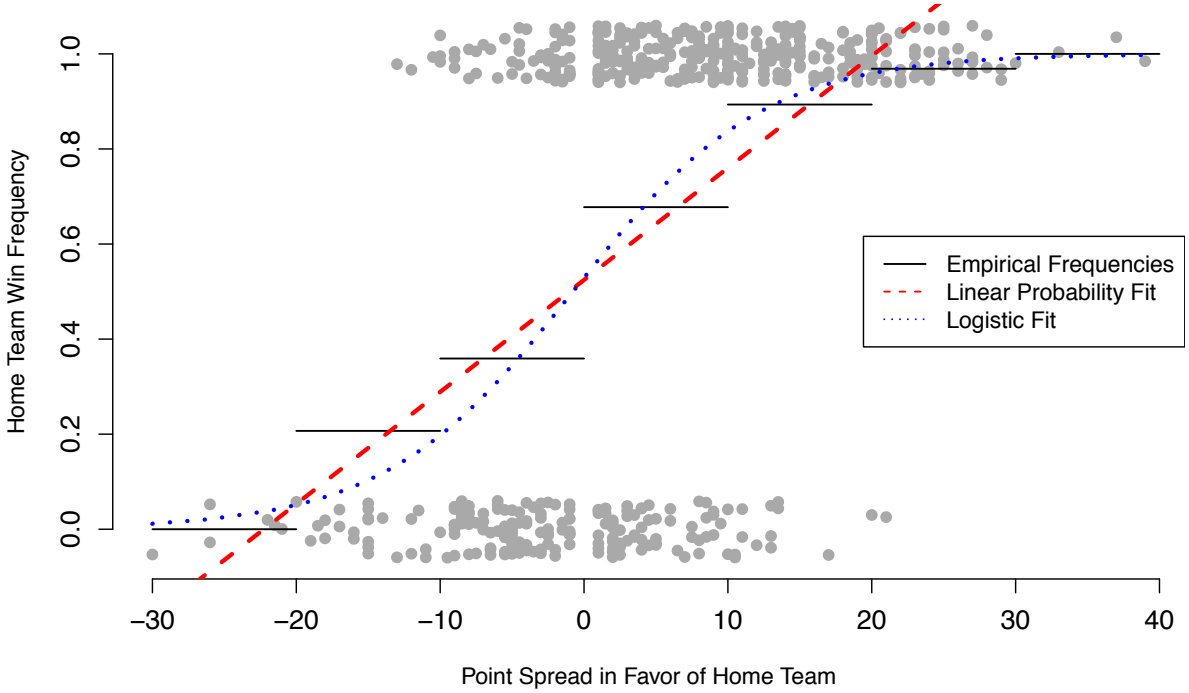


Figure 10.3: Win frequency versus point spread for 553 NCAA basketball games. Actual wins are plotted as 1's and actual losses as zeros. Some artificial vertical jitter has been added to the 1's and 0's to allow the dots to be distinguished from one another.

The logistic regression fit for the point-spread data

Let's return briefly to the data on point spreads in NCAA basketball games. The figure above compares the logistic model to the linear-probability model. The logistic regression fit ($\widehat{\beta}_0 = 0.117$, $\widehat{\beta}_1 = 0.152$) eliminates the undesirable behavior of the linear model, and ensures that all forecasted probabilities are between 0 and 1. Note the clearly non-linear behavior of the dotted blue curve. Instead of fitting a straight line to the empirical success frequencies, we have fit an S-shape.

Interpreting the coefficients

Interpreting the coefficients in a logistic regression requires a bit of algebra. For the sake of simplicity, imagine a data set with only a single regressor x_i that can take the values 0 or 1 (a dummy variable). Perhaps, for example, x_i denotes whether someone received

the new treatment (as opposed to the control) in a clinical trial.

For this hypothetical case, let's consider the ratio of two quantities: the odds of success for person i with $x_i = 1$, versus the odds of success for person j with $x_j = 0$. Denote this ratio by R_{ij}. We can write this as

$$\begin{aligned} R_{ij} &= \frac{O_i}{O_j} \\ &= \frac{\exp\{\log(O_i)\}}{\exp\{\log(O_j)\}} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot 1\}}{\exp\{\beta_0 + \beta_1 \cdot 0\}} \\ &= \exp\{\beta_0 + \beta_1 - \beta_0 - 0\} \\ &= \exp(\beta_1)\,. \end{aligned}$$

Therefore, we can interpret the quantity e^{β_1} as an *odds ratio*. Since $R_{ij} = O_i/O_j$, we can also write this as:

$$O_i = e^{\beta_1} \cdot O_j\,.$$

In words: if we start with $x = 0$ and move to $x = 1$, our odds of success ($y = 1$) will change by a multiplicative factor of e^{β_1}.

For this reason, we usually refer to the exponentiated coefficient e^{β_j} as the odds ratio associated with predictor j.

Advanced topic: estimating the parameters of the logistic regression model

In previous chapters we learned how to estimate the parameters of a linear regression model using the least-squares criterion. This involved choosing values of the regression parameters to minimize the quantity

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\,,$$

where $\hat{y}_i$ is the value for y_i predicted by the regression equation.

In logistic regression, the analogue of least-squares is Gauss's principle of maximum likelihood, which we introduced when discussing the normal linear regression model. The idea here is to choose values for β_0 and β_1 that make the observed patterns of 1's and 0's look as likely as possible.

To understand how this works, observe the following two facts:

- If $y_i = 1$, then we have observed an event that occurred with probability $P(y_i = 1 \mid x_i)$. Under the logistic-regression

model, we can write this probability as

$$P(y_i = 1 \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

- If $y_i = 0$, then we have observed an event that occurred with probability $P(y_i = 0 \mid x_i) = 1 - P(y_i = 1 \mid x_i)$. Under the logistic regression model, we can write this probability as

$$1 - P(y_i = 1 \mid x_i) = 1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

Since all of the individual 1's and 0's are independent, given the parameters β_0 and β_1, the joint probability of all the 1's and 0's is the product of their individual probabilities. We can write this as:

$$P(y_1, \ldots, y_n) = \prod_{i:y_i=1} \left(\frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right) \cdot \prod_{i:y_i=0} \left(1 - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right).$$

This expression is our *likelihood*: the joint probability of all our data points, given some particular choice of the model parameters.[2] The logic of maximum likelihood is to choose values for β_0 and β_1 such that $P(y_1, \ldots, y_n)$ is as large as possible. We denote these choices by $\widehat{\beta}_0$ and $\widehat{\beta}_1$. These are called the *maximum-likelihood estimates* (MLE's) for the logistic regression model.

[2] Remember that the big $\prod$ signs mean "product," just like $\sum$ means "sum." The first product is for the observations where y_i was a 1, and the second product is for the observations where y_i was a 0.

This likelihood is a difficult expression to maximize by hand (i.e. using calculus and algebra). Luckily, most major statistical software packages have built-in routines for fitting logistic-regression models, absolving you of the need to do any difficult analytical work.

The same is true when we move to multiple regression, when we have p predictors rather than just one. In this case, the logistic-regression model says that

$$P(y_i = 1 \mid x_{i1}, \ldots, x_{i,p} = g(\beta_0 + \beta_1 x_i) = \frac{e^{\psi_{ij}}}{1 + e^{e^{\psi_{ij}}}}, \quad , \psi_{ij} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

where ψ_{ij} is the linear predictor for observation i.

Extensions to the basic logit model

The ordinal logit model

We can modify the logistic regression model to handle ordinal responses. The hallmark of ordinal variables is that they are measured on a scale that can't easily be associated with a numerical

magnitude, but that does imply an ordering: employee evaluations, survey responses, bond ratings, and so forth.

There are several varieties of ordinal logit model. Here we consider the *proportional-odds* model, which is most easily understood as a family of related logistic regression models. Label the categories as $1, \ldots, K$, ordered in the obvious way. Consider the probability $c_{ik} = P(y_i \leq k)$: the probability that the outcome for the ith case falls in category k *or any lower category.* (We call it c_{ik} because it is a cumulative probability of events at least as "low" as k.) The proportional-odds logit model assumes that the logit transform of c_{ik} is a linear function of predictors:

$$\text{logit}(c_{ik}) = \log\left(\frac{c_{ik}}{1 - c_{ik}}\right) = \eta_k + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\,.$$

Crucially, this relationship is assumed to hold for all categories at once. Because $c_{iK} = 1$ for the highest category K, we have specified $K - 1$ separate binary logit models that all share the same predictors x_j and the same coefficients β_j. The only thing that differs among the models are the intercepts η_k; these are commonly referred to as the *cutpoints*. Since the log odds differ only by an additive constant for different categories, the odds differ by a multiplicative factor—thus the term "proportional odds."

To interpret the ordinal-logit model, I find it easiest to re-express individual fitted values in terms of covariate-specific category probabilities $w_{ik} = P(y_i = k)$:

$$w_{ik} = P(y_i \leq k) - P(y_i \leq k - 1) = c_{ik} - c_{i,k-1}\,,$$

with the convention that $c_{i0} = 0$. Good software makes it fairly painless to do this.

The multinomial logit model

Another generalization of the binary logit model is the multinomial logit model. This is intended for describing *unordered* categorical responses: PC/Mac/Linux, Ford/Toyota/Chevy, plane/train/automobile, and so forth. Without a natural ordering to the categories, the quantity $P(y_i \leq k)$ ceases to be meaningful, and we must take a different approach.

Suppose there are K possible outcomes ("choices"), again labeled as $1, \ldots, K$ (but without the implied ordering). As before, let $w_{ik} = P(y_i = k)$. For every observation, and for each of the K

choices, we imagine that there is a linear predictor ψ_{ik} that measures the preference of subject i for choice k. Intuitively, the higher ψ_{ik}, the more likely that $y_i = k$.

The specific mathematical relationship between the linear predictors and the probabilities w_{ik} is given the multinomial logit transform:[3]

$$\begin{aligned} w_{ik} &= \frac{\exp(\psi_{ik})}{\sum_{l=1}^{K} \exp(\psi_{il})} \\ \psi_{ik} &= \beta_0^{(k)} + \beta_1^{(k)} x_{i1} + \cdots \beta_p^{(k)} x_{ip} \,. \end{aligned}$$

[3] Some people, usually computer scientists, will refer to this as the *softmax* function.

Each category gets its own set of coefficients, but the same set of predictors x_1 through x_p.

There is one minor issue here. With a bit of algebra, you could convince yourself that adding a constant factor to each ψ_{ik} would not change the resulting probabilities w_{ik}, as this factor would cancel from both the numerator and denominator of the above expression. To fix this indeterminacy, we choose one of the categories (usually the first or last) to be the reference category, and set its coefficients equal to zero.

Models for count outcomes

The Poisson model. For modeling event-count data (photons, mortgage defaults in a ZIP code, heart attacks in a town), a useful place to start is the Poisson distribution. The key feature of counts is that they must be non-negative integers. Like the case of logistic regression, where probabilities had to live between 0 and 1, this restriction creates some challenges that take us beyond ordinary least squares.

The Poisson distribution is parametrized by a rate parameter, often written as λ. Let k denote an integer, and y_i denote the event count for subject i. In a Poisson model, we assume that

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i} \,,$$

and we wish to model λ_i in terms of covariates. Because the rate parameter of the Poisson cannot be negative, we must employ the same device of a link function to relate λ_i to covariates. By far the most common is the (natural) log link:

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} \,,$$

or equivalently,

$$\lambda_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip}\}\,.$$

As with the case of logistic regression, the model is fit via maximum-likelihood.

Interpreting the coefficients. Because we are fitting a model on the log-rate scale, additive changes to an x variable are associated with multiplicative changes in the y variable. As before, let's consider the ratio of two quantities: the rate of events for person i with $x_1 = x^\star + 1$, versus the rate of events for person j with $x_1 = x^\star$. Let's further imagine that all other covariates are held constant at values x_2 to x_p, respectively. This implies that the only difference between subjects i and j is a one-unit difference in the first predictor, x_1.

We can write their ratio of rates as

$$\begin{aligned} R_{ij} &= \frac{\lambda_i}{\lambda_j} \\ &= \frac{\exp\{\beta_0 + \beta_1 \cdot (x^\star + 1) + \beta_2 x_2 + \cdots \beta_p x_p\}}{\exp\{\beta_0 + \beta_1 \cdot x^\star + \beta_2 x_2 + \cdots \beta_p x_p\}} \\ &= \exp\{\beta_1 (x^\star + 1 - x^\star)\} \\ &= \exp(\beta_1)\,. \end{aligned}$$

Thus person i experiences events events e^{β_1} times as frequently as person j.

Overdispersion. For most data sets outside of particle physics, the Poisson assumption is usually one of convenience. Like the normal distribution, it is familiar and easy to work with. It also has teeth, and may bite if used improperly. One crucial feature of the Poisson is that its mean and variance are equal: that is, if $y_i \sim \text{Pois}(\lambda_i)$, then the expected value of y_i is λ_i, and the standard deviation of y_i is $\sqrt{\lambda_i}$. (Since λ_i depends on covariates, we should really be calling these the *conditional* expected value and standard deviation.)

As a practical matter, this means that if your data satisfy the Poisson assumption, then roughly 95% of observations should fall within $\pm 2\sqrt{\lambda_i}$ of their conditional mean λ_i. This is quite narrow, and many (if not most) data sets exhibit significantly more variability about their mean. If the conditional variance exceeds the

conditional mean, the data exhibits *overdispersion with respect to the Poisson*, or just *overdispersion* for short.

Overdispersion can really mess with your standard errors. In other words, if you use (i.e. let your software use) the Poisson assumption to calculate error bars, but your data are overdispersed, then you will end up overstating your confidence in the model coefficients. Sometimes the effect is dramatic, meaning that the blind use of the Poisson assumption is a recipe for trouble.

There are three common strategies for handling overdispersion:

(1) Use a quasi-likelihood approach ("family=quasipoisson" in R's glm function);

(2) Fit a different count-data model, such as the negative binomial or Poisson-lognormal, that can accommodate overdispersion;

(3) Fit a hierarchical model.

Alas, these topics are for a more advanced treatment of generalized linear models.