5
Quantifying Uncertainty: Part I

Key terms and concepts: estimator; sampling distribution; standard error; resampling/bootstrapping; confidence intervals and the frequentist coverage property; critical values and tail areas; normal linear regression model; prediction intervals

Three archetypal questions

In coming this far through the book, you've already learned many valuable skills: how to summarize evidence both graphically and numerically; how to fit basic group-wise and linear statistical models to data; how to use these models to explore trends and predict new outcomes; and how to summarize the information content of a model, by quantifying the model's predictive power in the context of the variation that remains unexplained.

But we're missing a crucial piece of the puzzle. From the introduction you'll recall our working definition of statistical modeling as: the structured quantification of uncertainty. We've focused a lot so far on the "structure" part. Now we'll focus on the uncertainty part. Here are three archetypal questions about uncertainty that we will now take up in earnest.

(1) How confident are we in our estimate of an effect size? Take the following study of a new therapeutic regime for esophogeal cancer, from the New England Journal of Medicine in July of 2006:

> We randomly assigned patients with resectable adenocarcinoma of the stomach, esophagogastric junction, or lower esophagus to either perioperative chemotherapy and surgery (250 patients) or surgery alone (253 patients). . . . With a median follow-up of four years, 149 patients in the perioperative-chemotherapy group and 170 in the surgery group had died. As compared with the surgery group, the perioperative-

chemotherapy group had a higher likelihood of overall survival (five-year survival rate, 36 percent vs. 23 percent).[1]

[1] Cunningham, et. al. "Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer." *New England Journal of Medicine*, 2006 July 6; 355(1):11-20.

Thus the chemotherapy regime appears to save 1 additional person in 8, compared to surgery alone. But what if the physicians running the trial had enrolled a different sample of patients? Might the effect size have looked more like 1 patient in 6, or 1 in 16? Chemotherapy has nasty side effects and is very expensive. If you're a cancer patient or a Medicare administrator, uncertainty about the effect size matters.

(2) Could this association plausibly be due to chance? We noticed that a country's spending on education seemed to predict its GDP growth rate (Figure 2.13). We quantified this by observing the positive slope of the least-squares line when modeling growth versus education. But if we had compiled a different random sample of 79 countries, would we have seen the same effect? Or could the positive slope of the line for these particular countries be caused merely by luck in the selection of the sample?

Another example: did a Titanic passenger's cabin class really correlate with his or her likelihood of survival? Or is this impression conveyed by Table 2.2 (reprinted at right) just an illusion attributable to chance?

	Cabin Class	1st	2nd	3rd
Female	Survived	139	94	106
	Died	5	12	110
Male	Survived	61	25	75
	Died	118	146	418

(3) How confident are we in our forecast? We've touched on this one a little bit, in the context of naïve prediction intervals. But even then, we ignored a crucial fact: namely, that uncertainty about parameters in the model translates into additional uncertainty about predictions, above and beyond the contribution of the residual. This is a subtle point, and we'll consider it at greater length soon.

Sampling distributions, estimators, and alternate universes

DIFFERENT versions of these three questions come up again and again, in just about every data set you'll encounter. Luckily, once you know how to answer one of them, you will know how to answer all of them.

That's because all three questions boil down to the same counterfactual: "if our data set had been different merely due to chance," each asks, "would our answer have been different, too?" In fitting

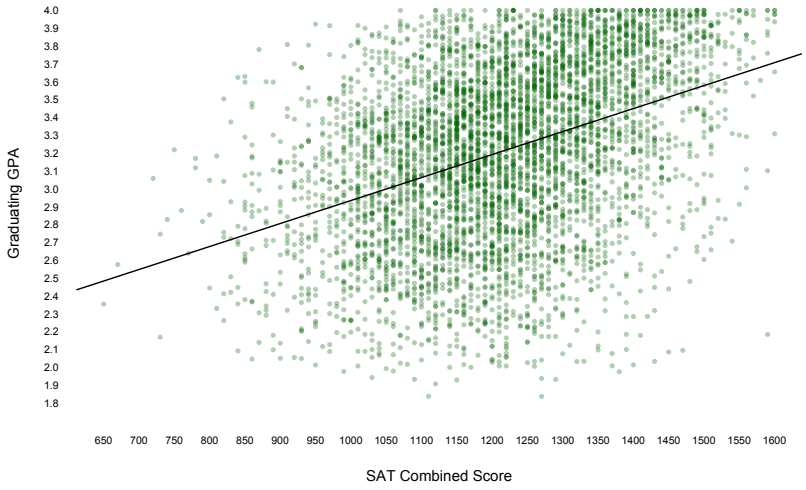


Figure 5.1: Graduating GPA versus high-school SAT score for all students who entered UT–Austin in the fall of 2000 and went on to earn a bachelor's degree within 6 years. The black line shows the least-squares fit.

statistical models, we typically equate the trustworthiness of a procedure with its stability under the influence of luck—that is, the degree to which that estimate might have changed if the forces of randomness had made the world look a bit different.

$$\text{Confidence in your estimates} \iff \text{Stability of those estimates under the influence of chance}$$

One obvious source of instability is when the individual observations themselves are subject to the forces of randomness. For example, suppose we wish to characterize the relationship between SAT score and graduating GPA for the entering class of 2000 at the University of Texas. Figure 5.1) shows the entire population, yet there is still randomness to worry about—for, as the teacher in Ecclesiastes puts it, "time and chance happeneth to them all." If any of these 5,191 students had taken the SAT on a different day, or eaten a healthier breakfast on the day of their chemistry finals, we would be looking at a slightly different data set, and thus a slightly different least-squares line—even if the underlying SAT–GPA relationship had stayed the same.

Another source of instability is the effect of sampling variability, which arises when we're unable to study the entire population of interest. The key insight here is that a different sample would have led to different estimates of the model parameters. Consider the example above about the study of a new chemotherapy regime for esophogeal cancer. If doctors had taken a different sample of 503 cancer patients and gotten a drastically different estimate of

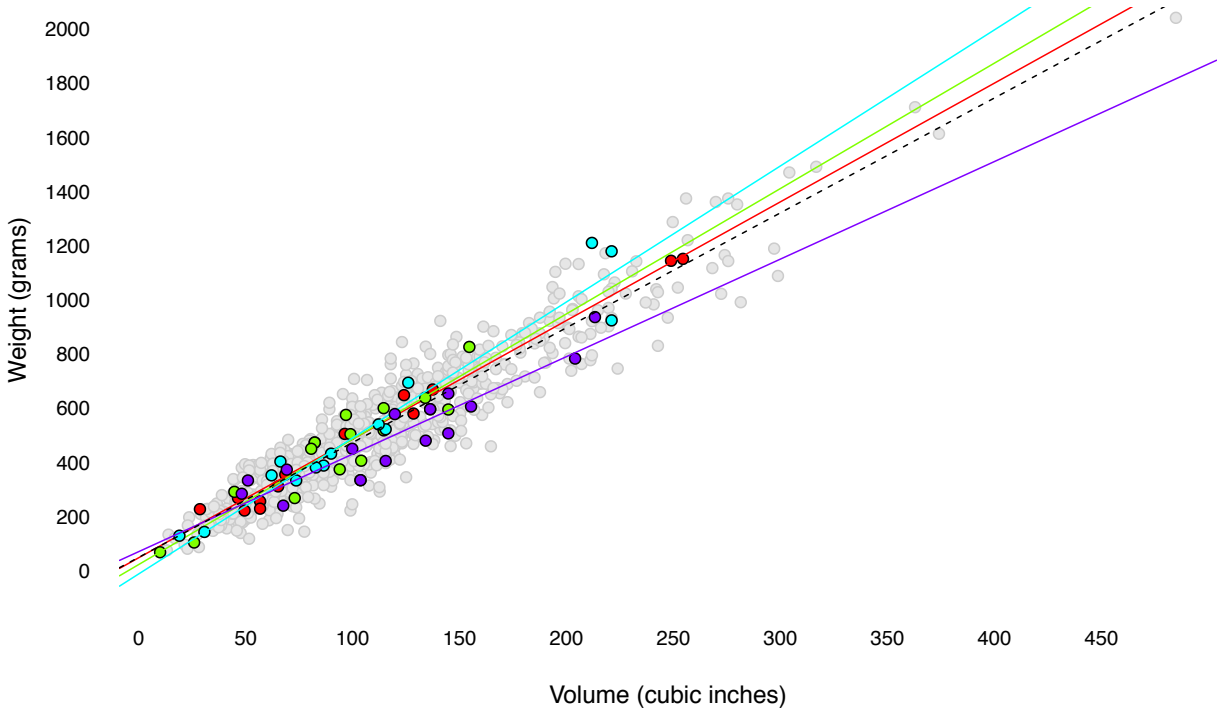


Figure 5.2: Four different days of fishing, coded by color, on an imaginary lake home to a population of 800 fish. On each day's fishing trip, you catch 15 fish, and end up estimating a slightly different weight–volume relationship. The dashed black line is the true relationship for the entire population.

the new treatment's effect, then the original estimate isn't very trustworthy! If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for *this particular* subset of 503 is likely to be accurate.

To get a little better intuition for this way of thinking, imagine that you go on a four-day fishing trip to a lovely small lake out the woods. The lake is home to a population of 800 fish of varying size and weight, depicted in Figure 5.2. On each day, you take a random sample from this population. That is, you catch-and-release 15 fish, recording the weight of each one, along with its length, height, and width (which multiply together to give a rough estimate of volume). You then use the day's catch to compute a different estimate of the linear volume–weight relationship for the entire population of fish in the lake. These four different days—and the four different least-squares fits—show up in different colors in Figure 5.2.

Four days of fishing give us some idea of how the estimates for β_0 and β_1 vary from sample to sample. But 2500 days of fishing, simulated by computer, give us a much better idea! Figure 5.3

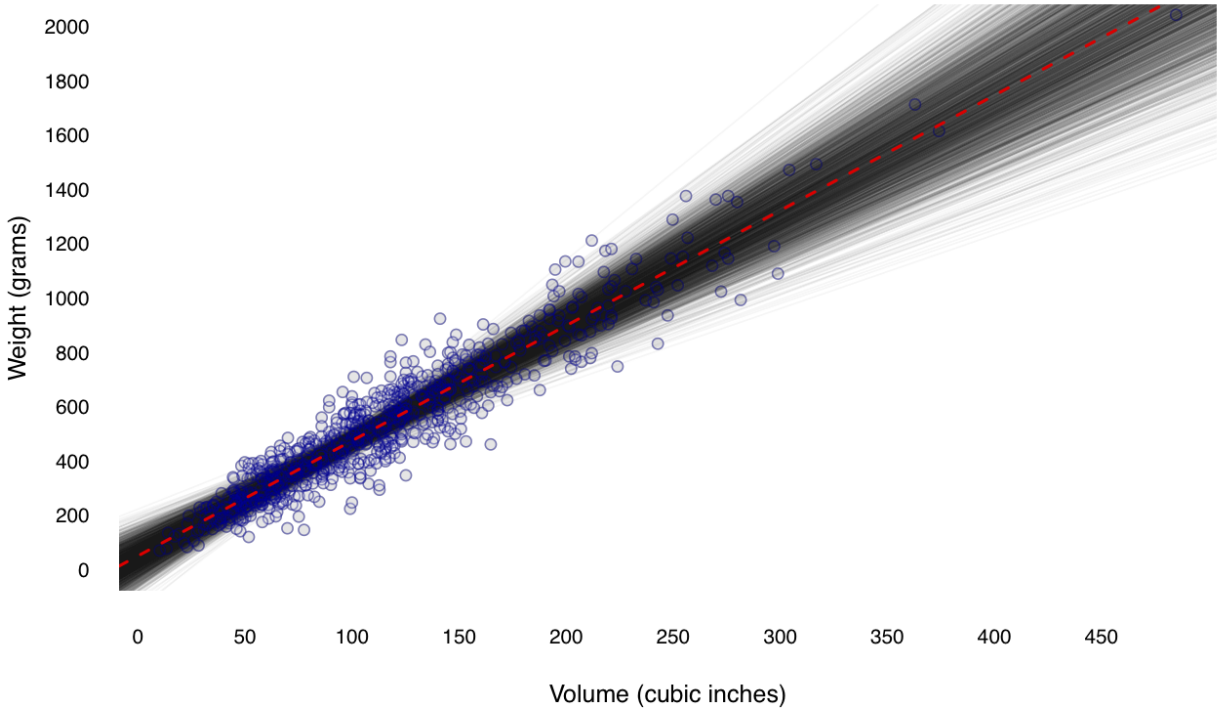
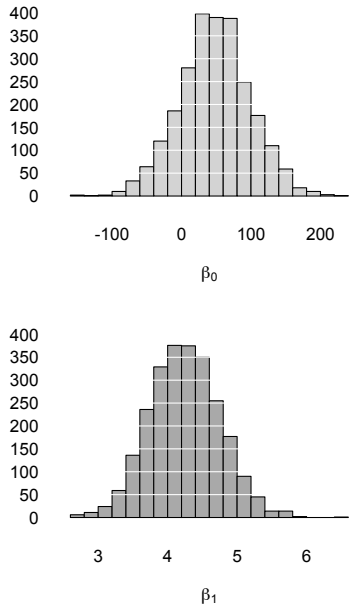


Figure 5.3: 2500 simulated days of fishing, together with the 2500 different estimates of β_0 and β_1 (below).

shows just this: 2500 different samples of size 15 from the population, together with 2500 different least-squares estimates of the weight–volume relationship.

These pictures show the *sampling distribution* of the least-squares line—that is, how the estimates for β_0 and β_1 change from sample to sample. (In theory, to know the sampling distributions exactly, we'd need to take an infinite number of samples. But here, 2500 is probably good enough.) Of crucial importance here is the distinction between an *estimator* and an *estimate*. Just like a trial is a procedure for reaching a verdict about guilt or innocence, an estimator is a procedure for reaching an estimate of some population-level quantity on the basis of a sample. The least-squares procedure yields estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the slope and intercept of a population-wide linear trend; the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ you get for a specific data set are the estimates. A sampling distribution summarizes how an estimator behaves under repeated sampling from a particular population. Good estimators are those that usually yield estimates close to the truth!



We typically summarize a sampling distribution using two

quantities: its mean and its standard deviation. If the mean of an estimator's sampling distribution is equal to the true population value, we say that the estimator is *unbiased*.[2]

Meanwhile, the standard deviation of an estimator's sampling distribution is referred to as the *standard error*. In quoting the standard error of an estimator's sampling distribution, you are saying: "If I were to take repeated samples from the population and use this estimator for every sample, my estimate is typically off from the truth by about this much." Notice again that this is a claim about a procedure, not a particular estimate. The bigger the standard error, the less stable the estimator across different samples, and the less you can trust that estimator for any particular sample. (You can see why it makes sense to equate stability with trustworthiness if you imagine a suspect who gives the police 6 different answers to the question, "Where were you last Tuesday night?")

Of course, if you really could take repeated samples from the population, life would be easy. You could simply peer into all of those alternate universes, tap each version of yourself on the shoulder, and ask, "What slope and intercept did you get for *your* sample?" By tallying up these estimates and seeing how much they differed from one another, you could discover precisely how much confidence you should place in your own estimates of β_0 and β_1, and report appropriate error bars.[3]

Most of the time, however, we're stuck with one sample, and one version of reality. We cannot know the actual sampling distribution of our estimator, because we cannot peer into all those other lives we might have lived, but didn't:

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth. . . .[4]

Yet quantifying our uncertainty requires knowing the road not taken. Strictly speaking, this is impossible. But there are two different ways of trying, each of which we will discuss in turn:

(1) By pretending that the sample itself is the population, which allows one to assess the effect of sampling variability through the process of *resampling*.

(2) By assuming that the forces of randomness obey certain mathematical regularities, which allows one to assess their effect using the rules of *probability theory*.

[2] This term has a precise mathematical meaning, but also an unwarranted connotation of desireability that many statisticians find deeply problematic. Alas, for historical reasons, we're basically stuck with the term. It turns out that unbiasedness is not always a good property of an estimator—indeed, there can be very good reasons to use estimators that we know to be biased. But that's for another book.

For the 2500 samples in Figure 5.3, the standard error of $\hat{\beta}_0$ is about 50, while the standard error of $\hat{\beta}_1$ is about 0.5.

[3] Let's ignore the obvious fact that, if you had access to all those alternate universes, you'd also have more data. The presence of sample-to-sample variability is the important thing to focus on here.

[4] Robert Frost, *The Road Not Taken*, 1916.

Inference via resampling

Bootstrapped standard errors

At the core of the resampling approach to statistical inference lies a simple idea. In most cases we can't repeatedly take samples of size n from the population. But we can repeatedly take samples of size n *from the sample itself*, and compute our estimator afresh for each notional sample. The idea is that the variability of the estimates across all the bootstrapped samples can be used to approximate the sampling distribution of the corresponding estimator.

This process is often called *bootstrapping*, and each block of n resampled data points is called a bootstrapped sample.[5] Modern software makes a non-issue of the calculational tedium involved.

[5] The term is a metaphor. Imagine trying to climb over a tall fence. If you don't have a rope, just "pull yourself up by your own bootstraps"!

You might be puzzled by something here. If there are n data points in the original sample, and we resample n data points from this "pseudo-population," won't each bootstrapped sample be precisely equal to the original sample? It turns out that the answer is no—as long as the resampling is done *with replacement* from the original sample. Sampling with replacement means that virtually all the bootstrapped samples will have duplicates and omissions from the original sample. These duplicates and omissions induce variation from one bootstrapped sample to the next that mimics the variation across the real repeat samples you're unable to take.

Resampling won't yield the true sampling distribution of an estimator, but it is often good enough for approximating the standard error. The quality of the approximation depends almost entirely on one thing: how closely the original sample resembles the wider population. Alas, this often isn't under your control, and is almost always the limiting factor in the accuracy of the bootstrap. You can't magic your way to sensible error bars by bootstrapping a biased, woefully small, or otherwise poor sample.

The approximation also depends on how many bootstrapped samples you take from the original sample. More bootstrapped samples help—up to a point. But taking more bootstrapped samples is never a substitute for having more *actual* samples in the real data set.

A natural question is: how well does bootstrapping work? To see the procedure in action, let's reconsider the least-squares estimator of the slope (β_1) for the weight–volume line describing the fish in our hypothetical lake. The top row of Figure 5.4 shows three actual sampling distributions, for samples of size $n = 15$, $n = 50$, and $n = 100$ from the entire population. Below each true sampling distribution are four replications of 2500 bootstrapped samples, each corresponding to the sample size from the top row.

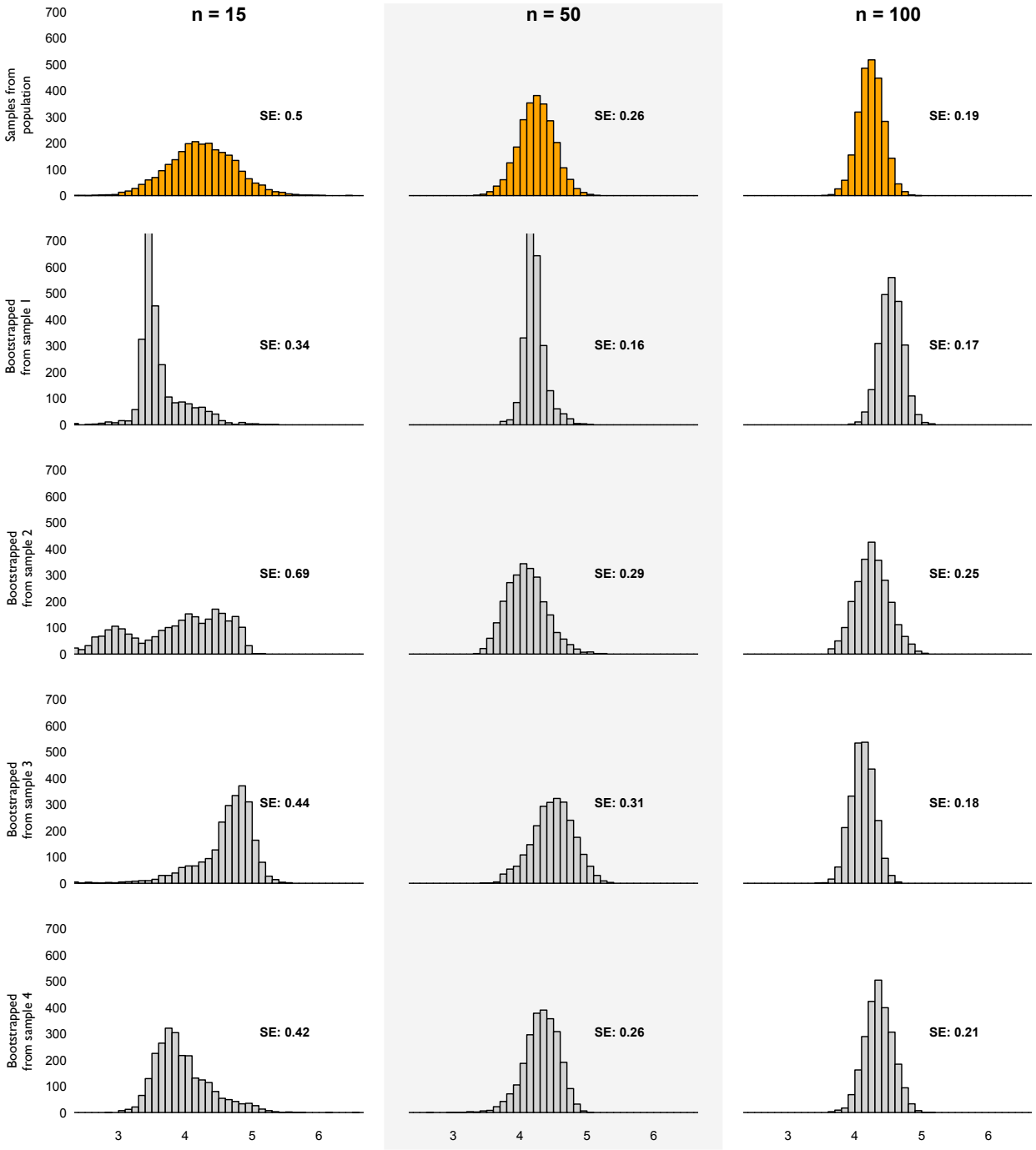


Figure 5.4: Actual (top, in orange) and bootstrapped sampling distributions (four replications) for the least-squares estimator of β_1 from Figure 5.2.

Recall how each of these bootstrapped sampling distributions is constructed. First, we take a real sample of size n from the population—for example, the 15 blue points in Figure 5.2. (We actually take 4 different real samples for each value of n, corresponding to the different columns in the figure.) Then we take 2500 bootstrapped samples, treating this original sample of size n as a pseudo-population. For each bootstrapped sample, we compute the least-squares line for weight versus volume. These 2500 estimates of β_1 are what you see in each grey-colored panel of Figure 5.4. If bootstrapping were perfect, each grey panel would look exactly like the corresponding orange panel above.

If you study these pictures closely, you'll notice a few things.

(1) The bootstrapped sampling distribution can differ substantially from one original sample to the next (top to bottom), especially when the original sample size is small (left column).

(2) The bootstrapped sampling distribution gets both closer to the truth, and less variable from one original sample to the next, as the original sample size gets larger (left to right).

(3) The bootstrapped standard errors (printed next to each histogram) are often closer to the true standard error than you would expect, based on the visual correspondence of the bootstrapped sampling distribution to the true one.

A rough rule of thumb is the following. If the bootstrapped sampling distribution has more than one peak or looks highly skewed—as in the left-hand column of Figure 5.4—then the resulting standard errors are highly suspect. There's a better chance, although no guarantee, that you're getting a decent estimate of the true standard error if the bootstrapped sampling distribution looks approximately bell-shaped. This will often, though not always, happen if you have at least 30 observations per parameter in the model. (In a group-wise model, for example, that means 30 observations per group.) Sometimes fewer are OK, and sometimes more are necessary, although it's hard as a general rule to say when.

The moral of the story is: don't take any such rule of thumb too seriously. In general, the only good reason to have confidence in the bootstrapped sampling distribution is because you believe your original sample is representative of the wider population. This isn't something that a statistical test—or a statistician—can verify. Rather, it's a question of judgment best answered by someone with subject-area expertise.

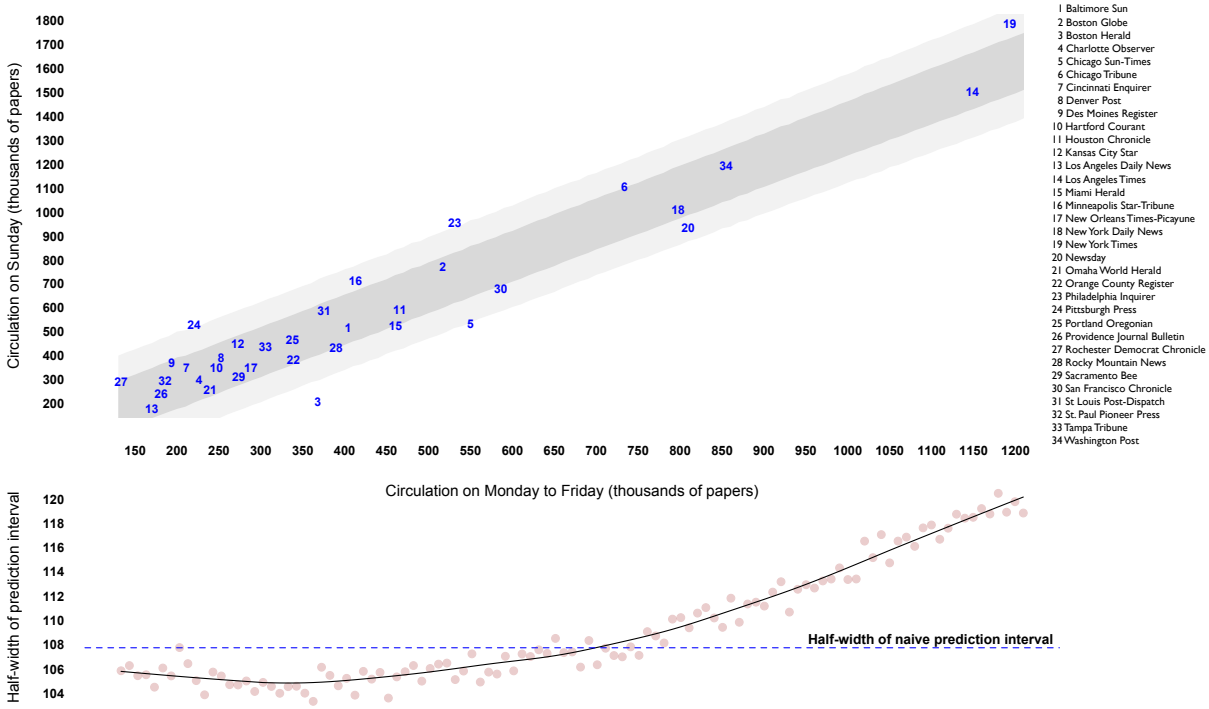


Figure 5.5: Sunday circulation versus daily circulation for 34 major metropolitan newspapers, together with one- and two-standard-deviation bootstrapped prediction intervals across the range of the X variable (top panel). Also shown is the half-width of the darker-grey prediction interval across the range of X (bottom panel), versus the half-width of the naïve prediction interval, shown by the dotted blue line.

You'll notice that the pink dots marking the half-width of each bootstrapped prediction interval wiggle up and down a bit from the black curve. This happens because we only took 2,500 bootstrap samples, which produces a bit of unwanted noise. Taking more bootstrapped samples would make the pink points fall closer to the black curve, but it wouldn't shift the black curve up or down.

Bootstrapped prediction intervals

Recall the problem of forecasting a future $y^\star$ corresponding to some predictor $x^\star$, using past data as a guide. (For example, how much should a used truck with 80,000 miles cost? How much can an Austin restaurant with a food rating of 7.5 charge for a meal?) Previously, we were content to quote a naïve prediction interval—for example,

$$\hat{y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x^\star \pm s\,,$$

or the best guess, plus-or-minus one residual standard deviation.

What made the prediction intervals naïve was the way we ignored uncertainty in our estimates for β_0 and β_1. For example, imagine that you work for a major metropolitan newspaper with a daily (Monday–Friday) circulation of 200,000 newspapers, and that your employer is contemplating a new weekend edition. You could certainly use the data in Figure 5.5, which correlates Sunday circulation with daily circulation for 34 major metropolitan newspapers, to inform your guess about the new Sunday edition's likely circulation. But the naïve prediction interval will mask real sources of

uncertainty. These may be large, and financially significant.

Luckily, now that we understand the logic of the bootstrap, we can try to account for this extra uncertainty. Just repeat the following steps a few thousand times:

(1) Take a single bootstrapped sample from the original sample, and compute the least-squares estimates $\hat{\beta}_0^{(r)}$ and $\hat{\beta}_1^{(r)}$. This gives you your best guess for the future y, given the information in the bootstrapped sample:

$$\hat{y}^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x^\star \,.$$

Here the superscript r denotes the r^{th} resample.

(2) Sample a residual $e^{(r)}$ at random from the bootstrapped least-squares fit, to mimic the unpredictable variation in the model.

(3) Set $y^{(r)} = \hat{y}^{(r)} + e^{(r)}$. This is your notional "future y" for the r^{th} bootstrapped sample.

If you take the standard deviation of all those $y^{(r)}$'s, you can directly quantify the uncertainty in your prediction—for example, by quoting the dark- and light-grey prediction intervals in Figure 5.5, which stretch to one and two standard deviations (respectively) on either side of the least-squares line.

One noticeable feature of the bootstrapped prediction intervals is the way they bend outwards as they get further away from the center of the sample. This is a bit hard to see in the top panel of Figure 5.5. To show this effect more clearly, the bottom panel explicitly plots the half-width of the dark grey bootstrapped prediction intervals at 109 different hypothetical X points: every increment of 10,000 newspapers across the entire range of daily circulation, from 130,000 to 1.2 million.

The black curve shows an unmistakeable trend. Prediction uncertainty increases when you move away from the mean of X. Figure 5.3, several pages earlier, will give you some intuition for why this is so: small differences in the slope get magnified when you move further away from the middle of the sample. The naïve prediction interval fails to capture this effect entirely! On this problem, for example, the naïve interval understates prediction uncertainty by 10,000 newspapers or more for large values of X.

A final point worth noting: all of the previous warnings about bootstrapped standard errors also apply to bootstrapped prediction intervals. If the observed data is unrepresentative of the population, bootstrapping will mislead rather than inform.

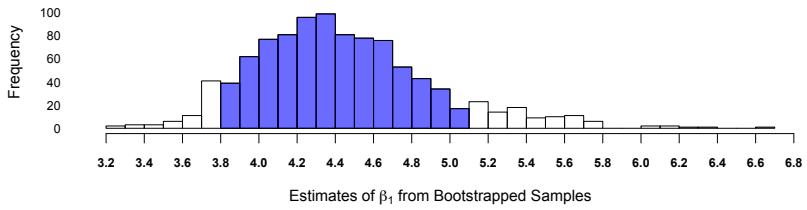


Figure 5.6: The estimated sampling distribution of $\hat{\beta}_1$ that arises from bootstrapping one sample of size 30 from the full fish population. The blue area reflects an 80% confidence interval, with symmetric tail areas of 10% above and 10% below the blue area.

Confidence intervals and coverage

We've seen how bootstrapping allows us to quantify the uncertainty in the least-squares estimates for the linear model parameters. A natural way to summarize this uncertainty is via a *confidence interval*, which will have an associated *confidence level* between 0% and 100%. A confidence interval is a summary of how precisely the data have allowed you to estimate the underlying population parameter. If your interval actually contains the true population value, we say that the interval *covers*. If it doesn't, the interval *fails to cover*. In real life, you won't know whether *your* interval covers. The confidence level quantifies how confident you are that it actually does!

You can express bootstrapped confidence intervals in two different ways. First, you could quote a symmetric error bar: the least-squares estimate, plus-or-minus some multiple k of the standard error. This number k is called the *critical value*, defined as the number of standard errors you must go out from the center to capture a certain percentage of the distribution. Typical values are $k = 1$ and $k = 2$. You could also aim for symmetric upper and lower tail areas, and report that 80% of the values from your bootstrapped sampling distribution yielded estimates of the Y-versus-X slope that fell on the interval $(3.8, 5.1)$, as above in Figure 5.6.

Be slightly careful here, for "confidence" has a notoriously tricky interpretation. To put it concisely but opaqely, confidence intervals are intervals generated by a method that satisfies the frequentist coverage principle. We state this roughly as follows.

The frequentist coverage principle: If you were to analyze one data set after another for the rest of your life, and were to quote 80% confidence intervals for every estimate you made, those intervals should cover their true values at least 80% of the time.

Let's imagine that your interval was generated with a procedure that, under repeated use on one sample after the next, tends to yield intervals that cover the true value with a relative frequency of at least 80%. Then, and only then, may you claim a bona fide 80% confidence level for your specific interval. (You may, of course, aim for whatever coverage level you wish in lieu of 80%. Many people seem to like 95%!) Thus confidence intervals involve something of a bait-and-switch: they purport to answer a question about an individual widget, but instead give you information about the assembly line that made the whole batch of widgets.

An obvious question is: do bootstrapped confidence intervals satisfy the frequentist coverage property? If your sample is fairly representative of the population, then the answer is a qualified yes. That is, the bootstrapping procedure yields nominal 80% intervals that cover the true value "approximately" 80% of the time (and so forth for other choices of the confidence level). Moreover, as the size of the original sample gets bigger, the quality of the approximation gets better. Alas, it is necessary to appeal to probability theory to place both of these claims on firmer footing. This is best deferred to another, more advanced book.

For our purposes, it is better to show the procedure in action. Figure 5.7, for example, depicts the results of running 100,000 regressions—1,000 bootstrapped samples for each of 100 different real samples from the population in Figure 5.2. The vertical black line shows the true population value of the weight–volume slope ($\beta_1 = 4.24$) for our population of fish. Each row is a different sample of size $n = 30$ from the population. Dots and crosses indicate the least-squares estimate of the slope arising from that sample, while the grey bars show the corresponding 80% bootstrapped confidence intervals (like the blue region in Figure 5.6).

The nominal confidence level of 80% for each individual interval must be construed as a claim about the *whole ensemble* of 100 intervals: 80% should cover, 20% shouldn't. The actual values were 83 and 17, so the claim is approximately correct!

If your bootstrapped sampling distribution looks bell-shaped, here's a nice rule of thumb, reflecting a pattern that statisticians have discovered to be true quite generally. An interval of the form "estimate $\pm$ 1 standard error" is an approximate 68% confidence interval; while one of the form "estimate $\pm$ 2 standard errors" is an approximate 95% interval. Another way of phrasing this is: $k = 1$ is the 68% critical value, while $k = 2$ is the 95% critical value.

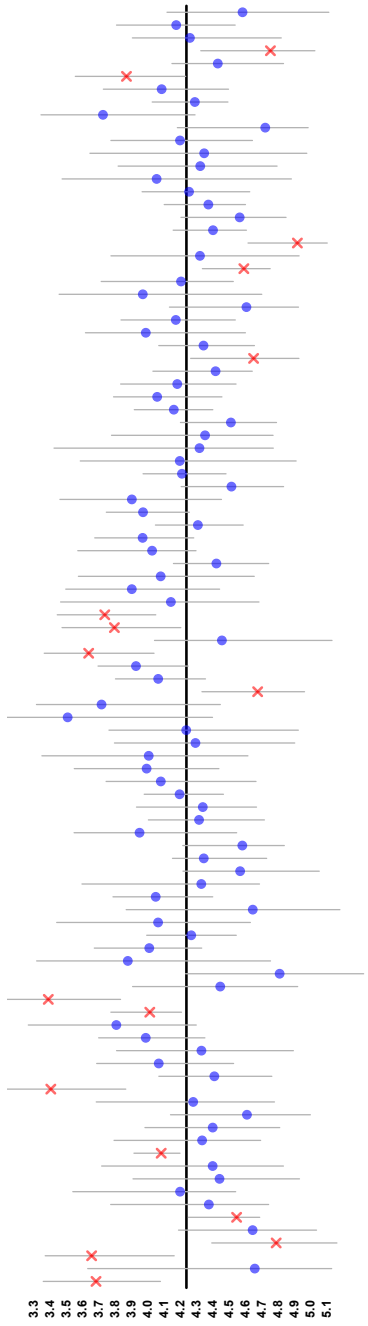


Figure 5.7: 100 different samples of size 30 from the population in Figure 5.2, along with each least-squares estimate of the weight–volume slope, and an 80% bootstrapped confidence interval, just like that at the top left. Blue dots show confidence intervals that cover; red crosses show those that don't.

Inference from probabilistic assumptions

When we quantify uncertainty via bootstrapping, we are implicitly making an assumption about the sample: namely, that it resembles the population in the ways that matter.

A second way of quantifying uncertainty is by making assumptions about the underlying state of nature (e.g. the population). In doing so, we pay a price by losing flexibility. But we enjoy an enormous return on this investment by being able to use probability theory, a rich language for describing chance outcomes.

Actually, in fitting least-squares lines, we've already made one important assumption about the underlying state of nature: namely, that the X and Y variables follow a straight-line trend. But as we've seen, this assumption is insufficient to quantify uncertainty about our estimates of the model parameters. To do that, we will need some additional assumptions—not about the line itself, but rather about the manner in which the line misses. That is, we must invest in a probability model for the residuals.

Previously we were content to invoke, somewhat vaguely, the idea of "sampling variability" or "the forces of randomness" in describing the uncertainty that arises in estimating the parameters of a regression line. Our new assumptions will pinpoint the source of this uncertainty. The line itself, we will stipulate, is the underlying state of nature. Like some Platonic form, it remains stable and unchanging for all possible samples. It's the residuals that are random, and thus the source of all uncertainty. Re-run history again—eat a different breakfast on the morning of your chemistry final, take a different sample of trucks advertised on Craigslist—and you'll get different residuals, therefore different data, therefore different estimates for the model parameters.

To help you understand this approach, let's recast our original description of the regression relationship as a straight-line fit, plus some wiggle room:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\sim \mathrm{E}\,. \end{aligned}$$

The original assumption of linearity remains. But now we write each residual using a Greek letter epsilon (ϵ_i), to emphasize that each one is a random variable, modeled by some as-yet-unspecified probability distribution E (for "error distribution"). This involves only a tiny notational difference, but an enormous

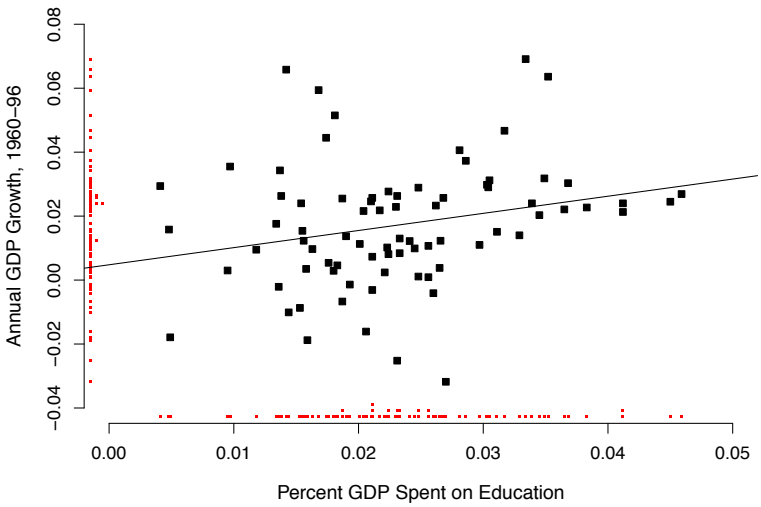
conceptual difference, from thinking of the residuals as the "misses" in the least-squares procedure. No longer will we interpret this equation merely as a claim about our particular sample. Now, we will interpret it as a much more ambitious claim about the underlying system we're studying—one that holds not just for our data set, but also for all the other data sets we might conceivably have collected for the same problem.

In principle, we could make any assumptions we want about the residuals. (These assumptions will be embodied in the choice of the error distribution E.) Of course, some assumptions are more reasonable than others! And just as importantly, some assumptions are easier to leverage than others, if we wish to use them to calculate the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Over the next several pages, we'll learn about a commonly used set of assumptions, often referred to collectively as the *normal (or Gaussian) linear regression model.* The model itself was first proposed by Gauss in 1809, and has more or less stuck! Many people believe that it strikes a nice balance between the goals of interpretability and calculational tractability.

Its justification requires three premises.

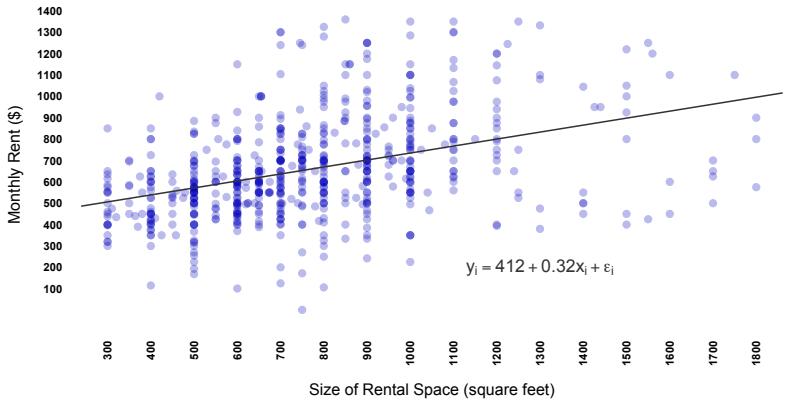
(1) Each residual is the aggregate result of forces not explicitly modeled. On the GDP growth data set, for example, we saw that education spending and GDP growth were related:



We can imagine, moreover, that the residuals from this regression line represent all those others factors—defense spending, life ex-

pectancy, population density, geography, natural resources, and so forth—that we've left out of the model.

(2) Each such force acts like a random up-or-down nudge to the residual, with no single nudge dominating. Take the following data on price versus size for 696 residential apartments for rent in Chicago:



Unsurprisingly, bigger units tend to be more expensive. But some apartments are shiny and new, nudging the price up from the line. Others have inefficient air conditioners, nudging the price down. Some buildings have doormen—upward nudge. Others have smelly laundry rooms—downward nudge. Granite countertops? Up. Ugly view? Down. And so on, for each of the hundreds of other little things that, taken together, determine whether an apartment will be cheap or expensive for its size.

An important caveat here is that no single nudge dominates. Is the apartment on top of an old nuclear waste dump? If so, the premise of roughly equal nudges probably doesn't hold! (More on this caveat in a bit.)

(3) The aggregation of random up-or-down nudges can be modeled with a normal (a.k.a. Gaussian) distribution. Of the three, this premise is the one most in need of explanation. Before we get there, however, take a moment to appreciate the conclusion. If we take this premise in conjunction with the first two, we are led to conclude that the residuals in a regression model can be described by a normal distribution. This is quite a substantial result. Therefore, let's take care to understand this third premise in more detail.

Aggregating up-or-down nudges

Nudges as coins

START with a single basic metaphor: *a nudge is like a coin flip*, where "heads" brings you up from the line, and "tails" brings you down. A residual, we'll assume, is the sum of many small nudges, where each nudge is equally likely to be up or down, and where successive nudges are independent of each other. Therefore—invoking premises 1 and 2—a residual is like a sequence of independent flips of a fair coin.

To begin with a simple case, suppose we flip the coin only twice. This would correspond, for example, to a situation where there are just three conceivable forces affecting the response variable Y, one of which we've explicitly modeled, and two of which we haven't. Then there are three possible outcomes—zero, one, or two—for the random variable K, the number of heads arising from 2 different flips of the coin. Since all four possible sequences for the two flips (HH, HT, TH, TT) are equally likely, the probability distribution for K is:

k	$P(K = k)$	Residual
0	0.25	Below line
1	0.50	Even with line
2	0.25	Above line

The logic of this simple case can be extended to the general case of m nudges: by accounting for every possible sequence of heads and tails that could arise from m flips of a fair coin. In general, the residual will end up $k - (m - k) = 2k - m$ nudges away from the line. This could be either a positive or negative number, depending on whether the heads or tails predominate.

Since successive flips are independent, every sequence of heads and tails has the same probability: $1/2^m$. Therefore,

$$P(k \text{ heads}) = \frac{\text{Number of sequences with } k \text{ heads}}{\text{Total number of sequences}}. \qquad (5.1)$$

There are 2^m possible sequences, which gives us the denominator. To compute the numerator, we must count the number of these sequences where we see exactly k heads.

How many such sequences are there? To count them, imagine distributing the k heads among the m flips, like putting k items in

m boxes, or handing out k cupcakes among m people who want one. Clearly there are m people to which we can assign the first cupcake. Once we've assigned the first, there are $n-1$ people to which we could assign the second cupcake. Then there are $m-2$ choices for the third, and so forth for each successive cupcake. Finally for the kth and final cupcake, there are $m-k+1$ choices. Hence we count

$$m \times (m-1) \times (m-2) \times \cdots \times (m-k+1) = \frac{m!}{(m-k)!}$$

possible sequences, where $m!$ is the factorial function. For example, if $m = 10$ and $k = 7$, this gives 604,800 sequences.

But this is far too many sequences. We have violated an important principle of counting here: don't count the same sequence more than once! The problem is that have actually counted all the ordered sequences, even though we were trying to count unordered sequences. For example, in the $m = 10, k = 7$ case, we have counted "Heads on flips $\{1,2,3,4,5,6,7\}$" and "Heads on flips $\{7,6,5,4,3,2,1\}$" as two different sequences. But they clearly both correspond to the same sequence: HHHHHHHTTT.

So how many times have we overcounted each unordered sequence in our tally of the ordered ones? The way to compute this is to count the number of ways we could order k objects. Given a group of k numbers which will be assigned to the "heads" category, we could have chosen from k of the objects to be first in line, from $k-1$ of them to be second in line, from $k-2$ of them to be third in line, and so forth. This means we have counted each unordered sequence $k!$ times, giving us

$$\frac{m!}{k!(m-k)!} = \binom{m}{k}$$

possible sequences.[6] For $m = 10$ and $k = 7$, this is 120 unordered sequences—the right answer, and a far cry from the 604,800 we counted above.

[6] These $\binom{m}{k}$ expressions are called *binomial coefficients*, and are read aloud as "m choose k."

Putting all these pieces together, we find that the probability of getting k heads in m flips of a fair coin is

$$P(k \text{ heads}) = \frac{m!}{k!(m-k)!}\,\frac{1}{2^m} = \binom{m}{k}\frac{1}{2^m}. \tag{5.2}$$

The binomial distribution

We can apply the same idea to construct the whole family of *binomial distributions*. A binomial distribution describes the number of

successes in m independent yes-or-no trials, where the probability of success in each trial is p (not necessarily 1/2). The only difference from above is that now, some sequences are more likely than others. For example, if p were small—say, 0.05—then sequences with more successes than failures would be comparatively unlikely.

Let's take a sequence of m trials where we observed k successes. Each success happens with probability p, and there are k of them. Each failure happens with probability $1 - p$, and there are $m - k$ of them. Because each trial is independent, we multiply all of these probabilities together to get the probability of the whole sequence: $p^k \ (1-p)^{m-k}$.

Now if we let K denote the (random) number of successes in m trials, then for any value of k from 1 to m,

$$P(K = k) = \binom{m}{k} \ p^k \ (1-p)^{m-k} .$$

The random variable K is said to have a binomial distribution, which we often write as $K \sim \text{Bin}(n, p)$. The binomial distribution has two parameters: m, the number of trials; and p, the probability of success on each trial. Figure 5.8 shows a histogram of 1000 samples from the binomial distribution with $m = 10$ and $p = 0.25$.

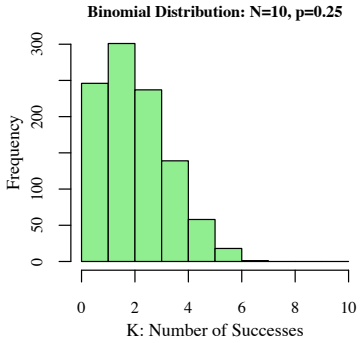


Figure 5.8: A histogram showing 1000 samples from a binomial distribution with $m = 10$ and $p = 0.25$.

de Moivre, Gauss, Laplace, and the first central limit theorem

Let's quickly review the thread of our argument here. Recall that we started by conceptualizing the residuals from a regression model as the aggregration of up-or-down nudges from the line. We have shown that these up or down nudges can be described using a binomial distribution. There is but one link left in the chain leading to the normal linear regression model, with residuals described using a normal distribution. That link belongs jointly to three famous mathematicians: de Moivre, Gauss, and Laplace.

In 1711, a Frenchman named Abraham de Moivre published a book called *The Doctrine of Chances*. The book was reportedly was prized by gamblers of the day for its many useful calculations that arose in dice and card games. In the course of writing about these games, de Moivre found it necessary to perform computations using the binomial distribution for very large numbers of independent trials. (Imagine flipping a large number of coins and making bets on the outcomes, and you too will see the necessity of this seemingly esoteric piece of mathematics.)

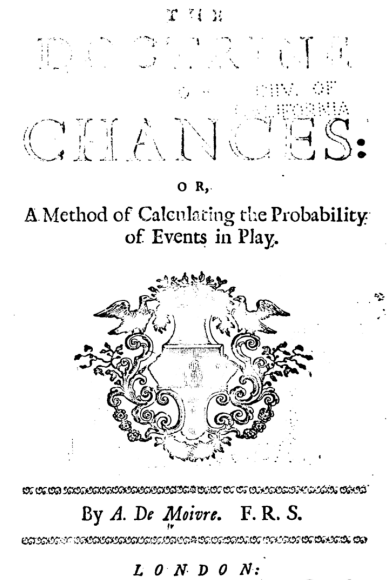

THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.
By *A. De Moivre.* F. R. S.
LONDON:

Figure 5.9: The title page of de Moivre's "The Doctrine of Chances" (1711), from an early edition owned by the University of California, Berkeley. One of the more interesting things about the history of statistics is the extent to which beautiful mathematical results came out of the study of gambling and parlor games!

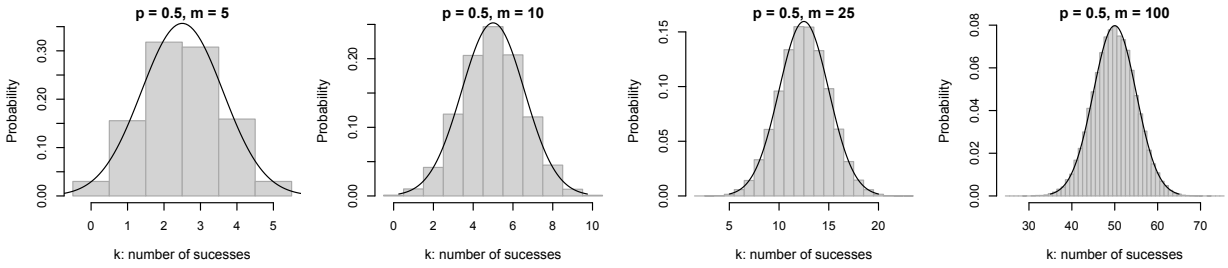


Figure 5.10: The binomial distribution for $p = 0.5$ and an increasingly large number of trials, together with de Moivre's normal approximation.

As you now know, this in turn requires computing binomial coefficients $\binom{m}{k}$ for very large values of m. But these computations were far too time-consuming without modern computers, which de Moivre obviously didn't have. So he derived an approximation based on the number $e \approx 2.7183$, the base of the natural logarithm.[7] He discovered that, if a random variable K has a binomial distribution, which we recall is written $K \sim \text{Bin}(m, p)$, then the approximate probability that $K = k$ is

$$P(K = k) \approx \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(k-\mu)^2}{2\sigma^2}}, \tag{5.3}$$

where $\mu = mp$ and $\sigma^2 = mp(1-p)$ are the expected value and variance, respectively, of the binomial distribution. When considered as a function k, this results in the familiar bell-shaped curve plotted in Figure 5.11—the famous *normal distribution*.

[7] Remember that the symbol $\approx$ means "approximately."

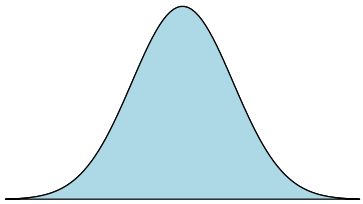


Figure 5.11: A normal distribution, often called a bell curve.

We can usually (though not always) avoid working with this expression directly, since every piece of statistical software out there can compute probabilities under the normal distribution. The important thing to notice is how the binomial samples in Figure 5.10 start to look more normal as the number of trials gets progressively larger: first 5, then 10, 25, and finally 100. The histograms show the binomial distribution itself, while the black curves show de Moivre's approximation. Clearly he was on to something!

This famous result of de Moivre's is usually thought of as the first *central limit theorem* in the history of statistics, where the word "central" should be understood to mean "fundamental." He essentially proved Premise 3 from the previous section, that the aggregation of random up-or-down nudges—that is, a binomial random variable—can be approximated with a normal distribution.

So if de Moivre invented the normal approximation to the binomial in 1711, and Gauss (1777–1855) did his work on statistics almost a century after de Moivre, why then is the normal distri-

bution often referred to as the Gaussian distribution? This quirk of eponymy arises because de Moivre only viewed his approximation as a narrow mathematical tool for performing binomial calculations. He gave no indication that he saw it as a more widely applicable "error curve," or a probability distribution for describing random deviations from an underlying trend.

But Gauss did see this, and far more. In his classic work of 1809 on fitting equations to the orbits of comets,[8] he reasoned as follows. Suppose that each residual from a regression model was an independent random variable arising from a probability distribution $p(\epsilon_i)$. If we then wanted to choose the "most probable" values of the regression parameters, we could do so by maximizing L, the product of the probabilities of the (independent) residuals:

$$L(\beta_0, \beta_1) = p(\epsilon_1) \cdot p(\epsilon_2) \cdots p(\epsilon_n)\,,$$

where of course $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. The closer the residuals are to the line, the higher the product of these probabilities.[9]

Considered as a function of β_0 and β_1, this quantity L is called a *likelihood*, and Gauss's proposal is known as the *principle of maximum likelihood*: maximize the probability of the data as a function of the model parameters. Gauss's marvelous result of 1809 was this: if $p(\epsilon_i)$ is de Moivre's normal distribution, *the maximum-likelihood solution for β_0 and β_1 is the same as the least-squares solution*.

In fact, Gauss's argument to this effect was unsatisfying. But Pierre-Simon Laplace, another famous mathematician, had been thinking about error curves and the notion of likelihood for almost 30 years, and Gauss's result would have electrified him head to toe. Laplace jumped in almost immediately to clean up Gauss's math, extending it in all sorts of beautiful ways. At a stroke, he unified three different lines of thinking about linear regression:

(1) that a residual, which captures the effect on the Y variable of what is left out of the model, can be viewed as the sum of many up or down nudges, where no single nudge dominates.

(2) that the aggregate of these up-or-down nudges can be described very well using a normal distribution.

(3) that the use of the normal distribution as an "error curve" validates the least-squares procedure for fitting lines, placing it on much more satisfying probabilistic foundations.

If anything, Laplace played a larger role than Gauss in this grand synthesis. But the details of his story are for another book.[10]

[8] Entitled *Theoria motus corporum coelestium in sectionibus conicis solem ambientum*.

[9] It is something of an anachronism to invoke the term "probability distribution" in the context of Gauss's work, for this notion wouldn't be defined precisely until a Russian mathematician named Kolmogorov did so in the 1930s. But it certainly conveys the essence of Gauss's argument, even if it is a slight abuse of modern terminology.

[10] For example, Stigler's *A History of Statistics* (ibid.), whence these historical details come.

The normal distribution

THE NORMAL distribution is now used in situations far more diverse than either de Moivre, Gauss, or Laplace ever would have envisioned. But it still bears the unmistakeable traces of its genesis as a large-sample approximation to the binomial distribution. That is, it tends to work best for describing situations where each normally distributed "error" can be thought of as an aggregation of many tiny, independent effects of about the same size, some positive and some negative. In cases where this description doesn't apply, the normal distribution may be a poor model of reality.

The main issue is the following: the normal distribution shares the property of a large-sample binomial distribution that huge deviations from the mean are very unlikely. It has, in statistical parlance, "thin tails." A normally distributed random variable has only a 5% chance of being more than two standard deviations away from the mean, and less than a 0.3% chance of being more than three standard deviations away from the mean. Large outliers are surpassingly rare. In the following histogram of daily returns for IBM stock, notice the huge outliers in the lower tail:

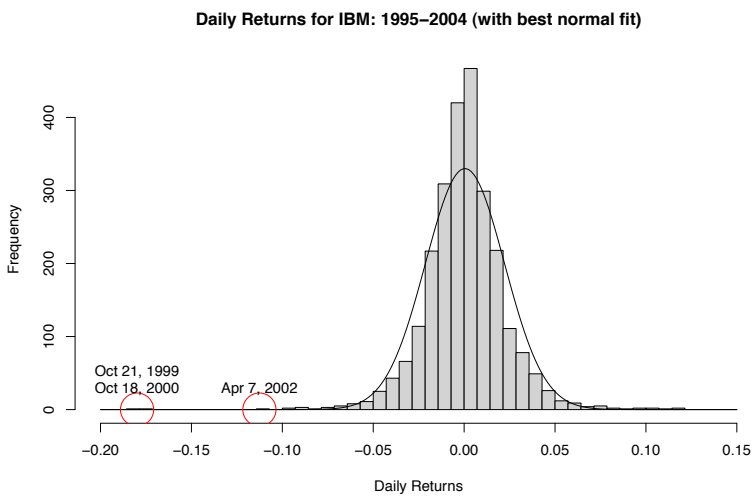


Figure 5.12: Daily stock returns for IBM from 1995 to 2004, together with the best-fitting normal approximation.

These returns would be wildly implausible if the returns really followed a normal distribution. A daily return tends to be dominated by one or two major pieces of information, and thus looks nothing like the aggregation of many independent up-or-down nudges.

The normal distribution works better for stock indices than it does for individual stocks. Take, for example, the best-fitting normal approximation for daily returns of the Dow Jones index, an aggregation of 30 individual stock returns:

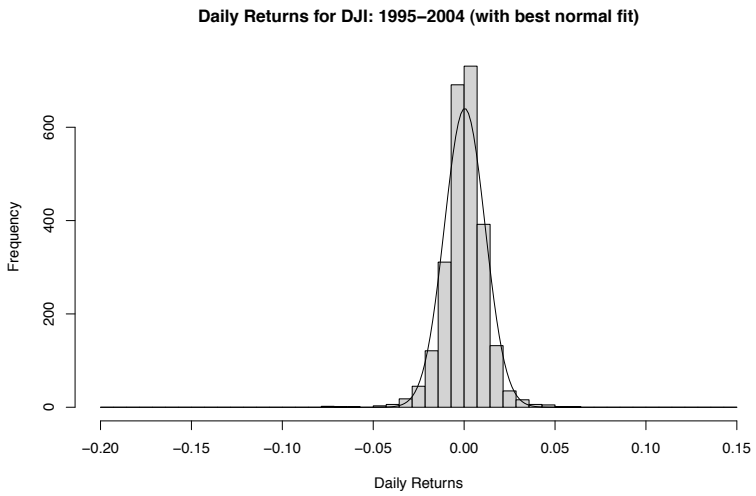


Figure 5.13: Daily stock returns for the Dow Jones stock index from 1995 to 2004, together with the best-fitting normal approximation.

Of course, the individual daily returns are not independent—stocks move together with the broader market—and so the fit is still not especially great.

Notation and important facts

If we want to use the normal distribution to describe our uncertainty about some random variable Y, we write $Y \sim \mathrm{N}(\mu, \sigma^2)$. The numbers μ and σ^2 are called the *parameters* of the distribution, and they describe which particular member of the family of normal distributions we're using to model Y. The first parameter, μ, describes where Y tends to be centered; it also happens to be the expected value (or mean) of the random variable. The second parameter, σ^2, describes how spread out X tends to be around its expected value; it also happens to be the variance of the random variable. Together, μ and σ^2 completely describe the distribution, and therefore completely characterize our uncertainty about Y.

Following de Moivre's original approximation, the probability

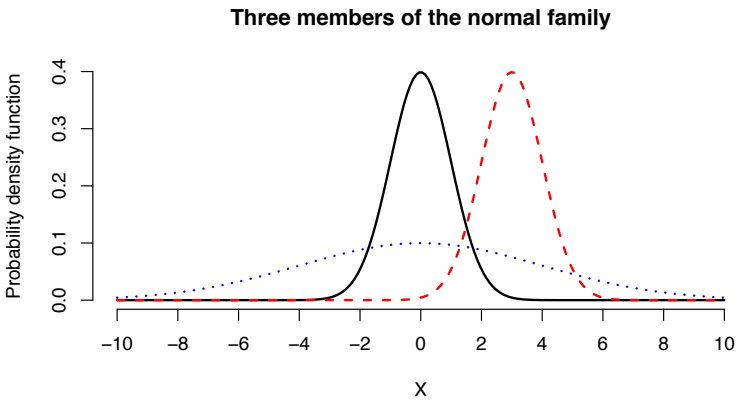


Figure 5.14: Three members of the normal family: $N(0, 1^2)$, $N(0, 4^2)$, and $N(3, 1^2)$. See if you can identify which is which using the guideline that 95% of the probability will be within two standard deviations σ of the mean. Remember, the second parameter is the variance σ^2, not the standard deviation. So $\sigma^2 = 4^2$ means a variance of 16 and a standard deviation of 4.

density function of the normal distribution is

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} . \tag{5.4}$$

If plot this as a function of y, you will see the familiar bell curve. How can you interpret a "density function" like this one? If you the take the area under this curve between two values z_1 and z_2, you will get the probability that the random variable Y will end up falling between z_1 and z_2. The height of the curve itself is a little more difficult to interpret. But intuitively, it corresponds to the probability of values that are within a very small region near the corresponding y value.

Recall from our discussion of critical values that a *tail area* refers to the probability of seeing an observation larger (or smaller) than some pre-specified value z. Here are two useful facts about normal tail areas—or more specifically, about the central areas under the curve, between the tails. If $Y \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} P(\mu - 1\sigma < Y < \mu + 1\sigma) &\approx 0.68 \\ P(\mu - 2\sigma < Y < \mu + 2\sigma) &\approx 0.95 . \end{aligned}$$

In words: the chance that Y will be within 1σ of its mean (i.e. a critical value of $z^\star = 1$) is about 68%, and the chance that it will be within 2σ of its mean (critical value $z^\star = 2$) is about 95%. Actually, it's more like $z^\star = 1.96$ rather than 2. So if your problem requires a level of precision to an order of 0.04σ or less, then don't use this rule of thumb, and instead go with the true critical value of 1.96.)

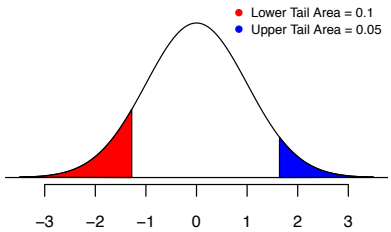


Figure 5.15: Examples of upper and lower tail areas. The lower tail area of 0.1 is at $z = -1.28$. The upper tail area of 0.05 is at $z = 1.64$

The normal linear regression model

FINALLY, we're there! Invoking Laplace's grand synthesis, we arrive at the normal linear regression model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \,. \end{aligned}$$

An equivalent way of writing this is:

$$(y_i \mid x_i, \beta_0, \beta_1, \sigma^2) \overset{iid}{\sim} \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \,.$$

This says that, conditional upon x_i and all three parameters of the model (β_0, β_1, and σ^2), each y_i value follows a normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

Let's be a bit more explicit about what else we are assuming here, in addition to the linearity of the regression relationship.

(1) *Independence of the residuals:* no residual provides any information about another residual.

(2) *Normality of the residuals:* the residuals ϵ_i come from a normal distribution with mean 0 and variance σ^2. (Invoking premises 1–3 from earlier: this is what you get if you sum many independent up-or-down nudges of comparable size.)

(3) *Homoskedasticity:* this one is a mouthful—an ornate Latin word that basically means "same variance." The assumption of homoskedasticity means that the variance of the residuals does not change as a function of x. Rather, σ^2 is the same for all observations, rather than being systematically larger for some observations and smaller for others.

These three assumptions about the residuals can be expressed in one succinct phrase: "i.i.d. normal," where "i.i.d." stands for *independent and identically distributed.* You see these letters above the twiddle ($\sim$) in the above equations. Sometimes for the sake of brevity we drop these three letters. But be aware that, in the regression output from most statistical software, the i.i.d. assumption is usually implicit unless stated otherwise.

Is the normal linear regression model ever "right"? Do we ever meet any real data sets where the x–y relationship is perfectly linear, and where the residuals all independently arise from the

same perfect normal distribution? Of course not! That's where this whole edifice of the normal linear regression model gets its name. Like all models, it will be good at some things and bad at others. The important question is not whether the model is wrong—because, taken literally, the model is always wrong—but whether, for a particular data set, it is so far wrong that it ceases to be useful for our purposes: summarizing trends, predicting future observations, and testing hypotheses about empirical relationships between quantities.

Of course, there's one obvious thing that the simple regression model isn't any good at: discovering causal relationships. Remember, the question of whether your model has a causal interpretation hinges on the design of your study and the substantive issue you are exploring—not in the statistical methods you bring to bear on your data. "Correlation $\neq$ causality" is just a rule of the game. The rules of the game don't change just because we can now play it with fancier mathematical equipment.

Luckily, it is straightforward to check whether the assumptions of the simple regression model look reasonable for a given data set. When we come to the the topic of model checking, we'll learn how to do just this. But to give you a preview: we can use the residuals to check for normality, independence, and homoskedasticity, just like we've already been using them to check for linearity.

Point estimation

FROM NOW on, we'll adopt the following thought process. Suppose that the normal linear regression model is true, but that we don't know the parameters. Now we see some data from the model. What do we think about β_0, β_1, and σ?

Going forwards: from model to data

In thinking about regression this way, it helps to see what kind of data sets that our model, if it were true, would produce. This is where computer simulation makes life easy. We can simulate from the simple regression model by following five steps:

(1) Choose particular values for the parameters β_0, β_1, and σ^2.

(2) Choose a particular value for the predictor variable x_i.

(3) Simulate a normally distributed residual $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$.

(4) Set $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, incorporating both the part predicted by line ($\beta_0 + \beta_1 x_i$) *and* the random deviation from the line (ϵ_i).

(5) Repeat Steps 2–4 until we have n different (x_i, y_i) pairs—as many as we want.

Let's follow this recipe. Below we see three different data sets of size $n = 50$ where the predictor variables (x) have been chosen randomly between 0 and 6. The true values of the parameters are given below each picture.

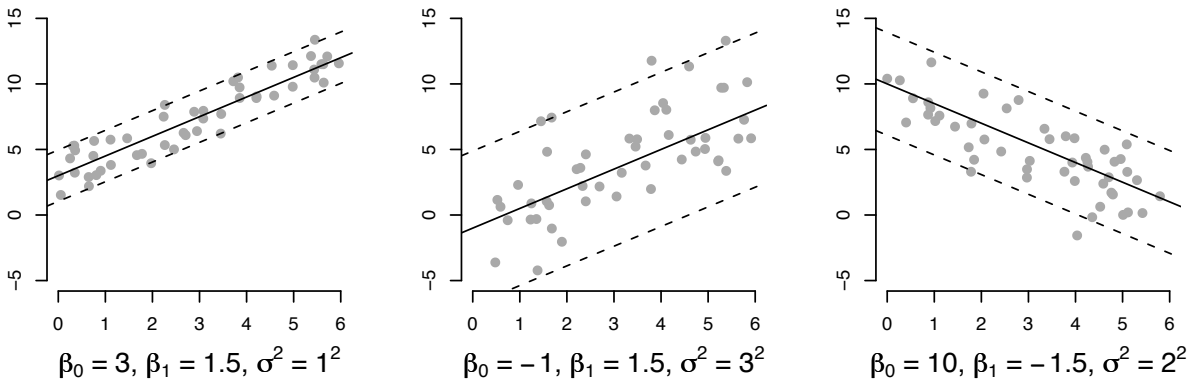


Figure 5.16: Three examples of data simulated according to the simple regression model.

The solid line is the true regression function, and the dotted lines mark a 2σ envelope to either side of the truth. On average, about 95% of the points should fall within this envelope. Notice the key role of σ in controlling how closely the values of y fall to the true line. (Try this yourself! Experiment with the parameters and see what different data sets you can generate.)

Going backwards: from data to model

Simulating from a known model to produce a specific (fake) data set is a deductive process. We start with some assumptions, and explore the consequences of these assumptions. Our reasoning goes something like: "If A, then B is likely."

But now imagine taking away the true line and the 2σ envelope from the plots above, so that all we see is the data. We can no longer peer "behind the curtain" to see the true values of the slope, intercept, and residual variance. Instead, we must reason

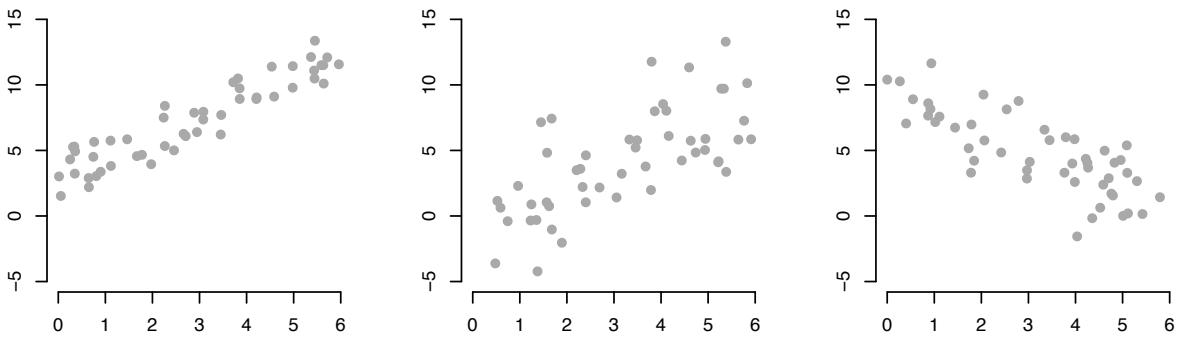


Figure 5.17: The same three data sets, without the true lines and true error bars.

about the likely values of these parameters, using only the traces they leave in a scatter plot, above.

This process is necessarily inductive: "If A, then B is likely. We see B in the data, therefore A is supported." Of course, this kind of inductive reasoning from known outcomes to plausible origins is never rock solid. That's why we need probability theory to quantify just how likely we are to be wrong.

And when we use the least-squares criterion to fit the line, we will always be wrong, at least by a little bit. In the margin on the next page (Figure 5.18), we see five different data sets of size $n = 10$. All five have been simulated from the same model: $\beta_0 = -1$, $\beta_1 = 1.5$, and $\sigma^2 = 3^2$. (This is the same model which gave rise to the middle data set in the previous two figures.) The true lines are in black, while the simulated data and least-squares fits are in grey.

In each case the data points are randomly scattered about the line in a now-familiar fashion. Each deviation from the line, we recall, is a single draw from a $N(0, \sigma^2)$ distribution. The least-squares fit—which "sees" only the 10 data points, and not the underlying model—usually comes close to the true line, but is never exact. Since we can see the true line, we can also see that the fitted line misses by a different amount, and in a slightly different way, for each simulated data set. The fitted line is always shifted slightly up or down, and always slightly tilted, with respect to the true line, suggesting that we're a bit off in our estimate of both the intercept and the slope.

Sampling variability, as we've come to learn, is just a fact of

life. This also has important implications for our understanding of the residuals. Under the simple regression model, both of the following relationships hold:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i \end{aligned}$$

The ϵ_i's are the deviations of the y_i's from the *true* line. The e_i's are the deviations of the y_i's from the *fitted* line. Since the fitted line is not quite equal to the true line, the fitted residuals e_i are not quite equal to the true residuals ϵ_i. From a conceptual standpoint, it's crucial to distinguish between these two quantities. (In general, our notational convention will be: if a quantity is written in Greek and doesn't wear a hat, then it's a "true" value of some probability model.)

Having said all of this, the fitted line does tend, *on average*, to be close to the true line. This fact suggests the following estimation strategy for tackling the inductive problem we've posed for ourselves, recalling that the least-squares estimate is also the maximum-likelihood estimate under the assumption of normal errors:

(1) Use the least-squares fitted intercept $\widehat{\beta}_0$ as an estimate of the true intercept β_0.

(2) Use the least-squares fitted slope $\widehat{\beta}_1$ as an estimate of the true slope β_1.

(3) Use the sample variance of the fitted residuals e_i as an estimate of the true variance σ^2.

Steps 1 and 2 are exactly what we've already been doing with the least-squares criterion. The only slight wrinkle here is with Step 3. Recall that σ^2, as the variance of the true residuals, is the expected squared deviation of the true ϵ_i's from the true line. Hence it would seem reasonable to estimate σ^2 using $s^2 = \sum_{i=1}^{n} e_i^2 / n$. (We've been doing just this when computing naïve prediction intervals.) Its appeal is obvious: if we need to estimate a theoretical average, let's use the corresponding sample average.

It turns out, however, that naïve variance estimator systematically underestimates the true value of σ^2. To use the proper statistical term: it is a *biased estimate* of the truth. And since this bias is in a direction that overstates the explanatory power of the

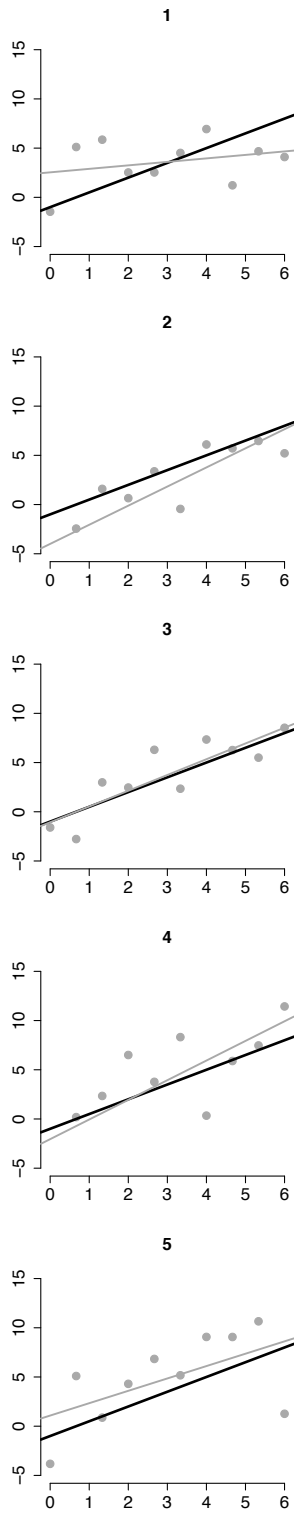


Figure 5.18: Five data sets of size $n = 10$ simulated from the same regression model: $\beta_0 = -1$, $\beta_1 = 1.5$, and $\sigma^2 = 3^2$.

least-squares line, it is particularly worrisome. (Remember the decomposition of variance: lower σ^2 means lower residual variance, and therefore higher R^2 for the fitted line.)

We'll not delve into the formal mathematical derivation of this fact. But the intuition is quite straightforward. Remember, in trying to estimate σ^2, we are trying to estimate the (theoretical) average squared deviation of the residuals from the true line. If we could observe an infinite number of deviations from the true line, this would be easy—with enough data, sample averages will converge to theoretical averages. But there are two problems: (1) we have only a finite number of observations; and (2) we can only observe the deviations from the fitted line, and the fitted line is not equal to the true line.

The naïve variance estimator runs afoul of this second crucial fact. If we were to use it, we'd be pretending as though the deviations of the y_i's from the fitted line are the same as the deviations from the true line. But they're not!

The standard fix here is to use a slightly more conservative estimator for σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} \left\{ y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right\}^2}{n-2}. \tag{5.5}$$

Notice what's different: instead of n in the denominator, we have $n-2$. This is smaller than n, meaning that the whole quantity $\hat{\sigma}^2$ is larger than the naïve estimator above. The square root of this quantity, denoted $\hat{\sigma}$, called the *residual standard error*, and is part of the routine output for every software package that does linear regression.

Why $n-2$ rather than n? This seemingly strange choice is hard to understand without delving into the mathematics. But it can be loosely interpreted using the following heuristic argument. What really matters is not the sample size n, but the *degrees of freedom* in our data set. We started with n data points, and therefore n degrees of freedom. But we "used up" two of these degrees of freedom in estimating β_0 and β_1. Hence we have $n-2$ degrees of freedom left, and should therefore divide by $n-2$ instead of n.

This is, admittedly, a somewhat murky justification. Don't think too hard about it; the only real justification is to be found in the geometry of high-dimensional Euclidean space, which shows that the residual standard error $\hat{\sigma}$ is an unbiased estimator of the true residual standard deviation σ. (It's easier to think in terms of σ rather than σ^2 because σ is on the scale of the original y data.)

The frequentist interpretation of the parameter estimates

Above, we said that the fitted line tends, on average, to be equal to the real line, even the equality is never exact for a specific data set. What do we mean by "on average" here? Let's assume that the following three conditions are met: (1) that we repeatedly take independent samples of size n from a simple linear regression model for the same fixed values of the predictor variables x_i; (2) that the true line and true residual variance of the model remain the same for all samples; and (3) that for each sample, we compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ according to the least-squares procedure, and $\hat{\sigma}$ according to Equation 5.5 above. (These assumptions describe how the five plots in Figure 5.18 were generated, except that we are imagining far more than five simulated data sets.)

Under these assumptions, the average value of $\widehat{\beta}_0$ will be β_0, the average value of $\widehat{\beta}_1$ will be β_1, and the average value of $\hat{\sigma}$ will be σ. The estimators, in other words, have a frequentist interpretation, just like bootstrapped confidence intervals. Here, the frequentist interpretation is the following: under repeated sampling from the same regression model, the average fitted values of the model parameters are equal to the true values. (This is precisely what is meant by the term *unbiased estimator*.) Just as before, the frequentist interpretation is not a property of an individual data set, but rather a property of our *procedure*, embodied here by the estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\hat{\sigma}$.

Uncertainty in estimation

We now return to the question that motivated us to posit the normal linear regression model in the first place: how much do our estimators vary from sample to sample under the assumption of normally distributed residuals? Just as with the bootstrapping procedure, the answer to this question provides a natural measure of the uncertainty associated with the estimators we actually have for the particular data set in hand.

From an intuitive standpoint, there are two obvious factors: (1) the true standard deviation σ, which measures the spread of the residuals around the true line; and (2) n, the size of each sample, since more data makes the true line easier to pin down.

Figure 5.19 can give you some idea of how these two crucial factors, σ and n, control the variation of the fitted line about the

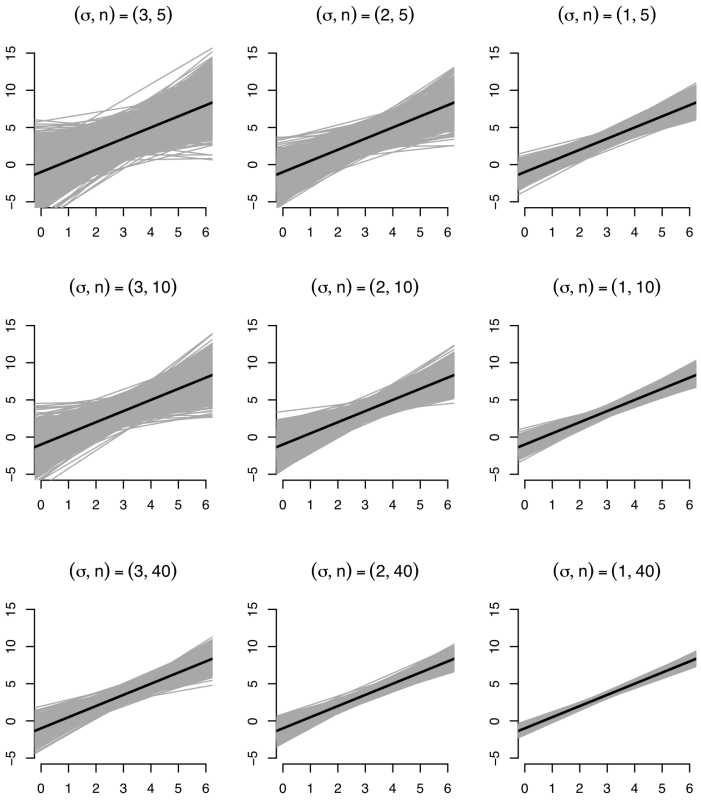


Figure 5.19: In each frame we see the true line, the least-squares fits for 1000 different simulated data sets, and the sample size (n) and residual standard deviation (σ) that were used to generate each data set.

true line. Each of these 12 frames shows the fitted lines for 1000 different simulated data sets. The black line is the same true line from before, and it always stays the same: $\beta_0 = -1$, $\beta_1 = 1.5$. Each thin grey line is a fitted line, based on data generated using a particular sample size and a particular true σ. As you move down the figure, the sample sizes (n) get bigger: first 5, then 10, then 40. As you move across the figure from left to right, the true residual standard deviations (σ) get smaller: first 3, then 2 and 1. In each frame, think of the 1000 grey lines as representing 1000 parallel universes—and while not quite infinite, 1000 universes is still many more than the 5 we looked at in Figure 5.18.

The direct parallel here is with Figure 5.3! In each case, the fans traced out by each set of 1000 grey lines are a visual reminder that sampling variability, while inescapable, can nonetheless be measured and quantified. In the upper left, where the sample size is small and the true σ large, the fitted lines vary quite a bit around the true line. Indeed, on rare occasions the estimated slope

is even negative, even though the slope of the true line is positive. (Chalk it up to bad luck, which sometimes happens.) But in the lower right, where the sample size is large and the true σ small, then the fitted line is always close to the true line.

It turns out that we can quantify, in a precise mathematical sense, the amount by which our estimators vary in all those parallel universes. We can do this, moreover, by directly invoking the assumptions of the normal linear regression model, without ever resorting to the bootstrapping procedure.

Let's first state the results outright, then try to understand them. It turns out that, under the assumption of i.i.d. normal residuals,

$$\widehat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma_0^2\right) \tag{5.6}$$

$$\widehat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \sigma_1^2\right). \tag{5.7}$$

These are the now-familiar *sampling distributions* of the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The variances of these two sampling distributions are:

$$\sigma_0^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \tag{5.8}$$

$$\sigma_1^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{5.9}$$

Remember that σ^2 is the true variance of the normal distribution governing the residuals ϵ_i, and that $\bar{x}$ is the sample mean of our predictor (x) variable.

An equivalent way of expressing these relationships is to use the following standardized formulas:

$$\left(\frac{\widehat{\beta}_0 - \beta_0}{\sigma_0}\right) \sim \mathrm{N}(0,1) \tag{5.10}$$

$$\left(\frac{\widehat{\beta}_1 - \beta_1}{\sigma_1}\right) \sim \mathrm{N}(0,1). \tag{5.11}$$

This is just like standardizing any normal random variable: subtract the mean and divide by the standard deviation, and the resulting quantity has a standard normal distribution with zero mean and unit variance. These quantities are often called z-statistics, and can be interpreted as a signal-to-noise ratio: the estimated size of the effect (signal), divided by how precisely you can estimate the effect (noise).

We'll not spend much time on the mathematical derivation of these results. But let's at least express the formulas in words.

In all of our parallel universes, the different values of $\widehat{\beta}_0$ follow a normal distribution with mean β_0 and variance σ_0^2. And the different values of $\widehat{\beta}_1$ follow a normal distribution with mean β_1 and variance σ_1^2. Larger values of σ_0^2 and σ_1^2 indicate greater uncertainty in the fitted regression line. Smaller values indicate less uncertainty, and therefore greater precision.

It's crucial that you keep straight the difference between the true values β_0 and β_1, and the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The true values are assumed to be fixed and unknown. The estimators are random variables, and are described by normal distributions. These two normal distributions have the same useful features as all other normal distributions. In particular, we know that in 95% of parallel universes, $\widehat{\beta}_0$ is within $2\sigma_0$ of β_0, and $\widehat{\beta}_1$ is within $2\sigma_1$ of β_1. (The corollary, of course, is that in 5% of parallel universes, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ will be outside of these bounds.)

From a mathematical standpoint, these formulas are a triumph. Using only the four assumptions of the normal linear regression model—linearity, independence, normality, and homoskedasticity—it is possible not only to derive estimators for the underlying slope and intercept of the simple regression model, but also to derive explicit uncertainty bands for these estimators. As long as the assumptions are met, these relationships hold regardless of the true values of β_0, β_1, and σ^2.

Of course, from a practical standpoint, the formulas have one flaw: they both depend upon σ^2, the true variance of the residuals, and we don't know what σ^2 is! But even though we don't know it, we can estimate it using $\hat{\sigma}^2$, the square of residual standard error from Equation 5.5. What if we made use of this fact by plugging in $\hat{\sigma}^2$ in lieu of the true (unknown) σ^2 in Equations 5.8 and 5.9?

$$\hat{\sigma}_0^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \tag{5.12}$$

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \,. \tag{5.13}$$

Notice how we've put hats on these two quantities to indicate that they are estimates, and not necessarily equal to the corresponding true values from above.

The square roots of these two quantities, denoted $\hat{\sigma}_0$ and $\hat{\sigma}_1$, have universally recognized names, and are part of the standard output from every software package that performs regression analysis. They are called the *estimated standard errors* of the regression

coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. They estimate the standard deviations of the distributions of errors that we make when, in each of our parallel universes, we use $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to estimate β_0 and β_1. (Recall the intuition behind this name from the bootstrapping section: an estimate of the standard deviation of the error is the standard error.)

Take a moment to inspect the formulas for the standard errors (Equations 5.12 and 5.13). Try to answer this question yourself: under what circumstances will the standard errors be large? Pause, and try it before reading on.

Now for the reveal. The two factors we mentioned before stick out immediately. When the true residual variance is large and the sample size is small, $\hat{\sigma}_0$ and $\hat{\sigma}_1$ will both be large, and our estimates for β_0 and β_1 will both have a lot of uncertainty.

But there's also a third factor that we might not have anticipated. Notice that the standard errors are small when the quantity $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is large compared to σ^2 (and vice versa). What's the intuition here? When our observed (x_i, y_i) pairs are spread out along the x axis—that is, whenever the x points have high variance—we will more easily be able to pin down the true line. If you imagine the difference in stability between balancing a stick with one finger, versus holding it on either end, you'll get the rough idea. See, for example, the two data sets at right. Both were generated from the same line, and both have the same number of points ($n = 20$) and the same residual variance ($\sigma^2 = 3^2$). Yet it will clearly be easier to estimate the true line using the top data set, because the x points are spread out compared to the residual variance.

Figure 5.20: The true line, sample size, and residual variance are the same. But the x points are more spread out in the top frame, making it easier to estimate the true line from the data.

We must now make one final, minor tweak to make the whole theory hang together. It turns out that the normality of the standardized regression estimators (Equations 5.10 and 5.11) no longer holds when we compute standard errors by plugging in $\hat{\sigma}^2$ and pretending as if it were exactly equal to the true σ^2. Instead of a normal distribution, the standardized regression estimators actually follow something called a t distribution:

$$t_0 = \left(\frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_0}\right) \sim t_{n-2} \tag{5.14}$$

$$t_1 = \left(\frac{\widehat{\beta}_1 - \beta_1}{\hat{\sigma}_1}\right) \sim t_{n-2}\,. \tag{5.15}$$

The t distribution is like a heavier-tailed version of the normal distribution. It has a single degrees-of-freedom parameter, which

is equal to $n-2$ in both of the above formulas. These quantities, appropriately enough, are called the *t statistics* associated with the regression coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The heavier tails of the t distribution are necessary because of the additional uncertainty we introduce by using $\hat{\sigma}$ in lieu of the real σ.

We won't worry too much about the difference between the t and the normal distribution. For moderate sample sizes—say, n about 30 or larger—there is almost no difference. Just keep in mind the same rule of thumb of "within $2\hat{\sigma}$ of the true value 95% of the time." (Notice the $\hat{\sigma}$ instead of the σ!) Even though the t distribution makes this rule not quite right, it's still close enough to be useful unless the sample size is very small.

Finally, we arrive at explicit expressions that give us confidence intervals associated with our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. To see this, let's review the interpretation of a standard error, starting with what we know about the t statistics:

$$\begin{aligned} t_0 &= \left(\frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_0}\right) \quad \sim \quad t_{n-2} \\ t_1 &= \left(\frac{\widehat{\beta}_1 - \beta_1}{\hat{\sigma}_1}\right) \quad \sim \quad t_{n-2}\,. \end{aligned}$$

We also know that if some number T follows a t distribution, then for any tail area α between 0 and 1, we can find an associated critical value, here denoted $t_{\alpha/2,n-2}$ such that

$$P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1-\alpha\,.$$

For example, for $\alpha = 0.05$ and $n = 100$, the critical value is $t_{\alpha/2,n-2} = 1.98$. This is basically the same as the $\alpha = 0.05$ critical value of 1.96 for the normal distribution.

Both of the t statistics corresponding to $\widehat{\beta}_0$ and $\widehat{\beta}_1$ follow a t distribution, so these results apply. Let's focus on $\widehat{\beta}_0$ for the moment. The above formula gives us

$$P\left(-t_{\alpha/2,n-2} < \frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_0} < t_{\alpha/2,n-2}\right) = 1-\alpha\,.$$

With a little bit of algebra inside the parentheses, we get

$$P\left(\widehat{\beta}_0 - \hat{\sigma}_0 \cdot t_{\alpha/2,n-2} < \beta_0 < \widehat{\beta}_0 + \hat{\sigma}_0 \cdot t_{\alpha/2,n-2}\right) = 1-\alpha\,.$$

This interval, $\widehat{\beta}_0 \pm \hat{\sigma}_0 \cdot t_{\alpha/2,n-2}$ can be interpreted as a confidence interval for the true value of β_0. The confidence level is

determined by α, and is typically quoted as the percentage corresponding to $1 - \alpha$ (for example, 95% for $\alpha = 0.05$). Just as with bootstrapping, higher confidence levels will produce wider intervals, since the corresponding critical value of the relevant t distribution will be larger in absolute value.

Of course, by a similar argument we can construct a confidence interval for β_1. This will simply be $\widehat{\beta}_1 \pm \hat{\sigma}_1 \cdot t_{\alpha/2,n-2}$. The fact that $t_{\alpha/2,n-2} \approx 2$ for moderate n and $\alpha = 0.05$ gives rise to the rule of thumb that, to construct a 95% confidence interval for a regression parameter, one simply takes the least-squares estimate, plus or minus twice the estimated standard error.

These confidence intervals also have a frequentist interpretation: they will contain the true values β_0 and β_1 in $100(1-\alpha)\%$ of all parallel universes in which the data have been generated from the same simple regression model. Just remember that it is the estimates and the interval endpoints which are the random variables here. The true values remain fixed, but unknown. We never know whether the confidence interval contains the true value for a given data set. We just know that it does so for $100(1-\alpha)\%$ of all possible data sets.

The standard errors in output from regression software

Both the standard errors and the t statistics are part of the usual output for all regression software. Here's an example of some regression output from R for one of the simulated data sets where $\beta_0 = -1$, $\beta_1 = 1.5$, and $\sigma^2 = 3^2$:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3570     1.5956  -0.850   0.4198
x             1.6665     0.4483   3.717   0.0059 **
---
Residual standard error: 2.715 on 8 degrees of freedom
Multiple R-squared: 0.6333,Adjusted R-squared: 0.5875
```

You can see the standard errors of the estimates in a column of their own, right next to the corresponding estimates of the slope and intercept parameters. And right next to them are the t statistics, here labeled as "t value." You can also see the residual standard error ($\hat{\sigma} = 2.715$) quoted at the bottom, right above R^2.

Uncertainty in prediction

We now come to the topic of prediction. The ability to take a known $x^\star$ value and use it to predict the corresponding $y^\star$ is one of the most powerful features of linear regression. We've done this at three ascending levels of sophistication:

(1) Point predictions, $\hat{y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x^\star$. This ignores all forms of prediction uncertainty.

(2) Naïve prediction intervals, $\hat{y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x \pm ks$, where s is the residual standard deviation, and k is some multiple of s that you get to choose. This interval takes into account uncertainty that arises from the residuals, but not parameter uncertainty.

(3) Bootstrapped prediction intervals, described several pages earlier. These account for all forms of uncertainty, assuming that the sample is representative of the population.

We will now explore a fourth way: by using the assumptions of the normal linear regression model to quantify both residual and parameter uncertainty, without appealing to the bootstrap.

First, let's assume we know the true β_0, β_1, and σ^2, in which case past data is irrelevant for future prediction. We know that there's a 95% chance that $y^\star$ will be within 2σ of its mean, which we'll recall is equal to $\beta_0 + \beta_1 x^\star$. Hence we would quote our 95% prediction interval as

$$\beta_0 + \beta_1 x^\star \pm 2\sigma$$

and call it a day. This symmetric interval is 2σ wide from center to endpoint, or equivalently 4σ wide from endpoint to endpoint. Of course, if we want a different confidence interval, such as 75% or 99%, there's only one additional step. Just compute the $z^\star$ corresponding to our confidence level—using, for example, R's `qnorm` function—and quote our prediction interval as

$$\beta_0 + \beta_1 x^\star \pm z^\star \sigma\,.$$

But we don't know β_0, β_1, or σ. We only have $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\hat{\sigma}$. We know that these estimates are off by a bit; as we've already seen, this adds a second layer of uncertainty to our prediction of $y^\star$, which must be taken into account.

Let's postpone the math and get straight to the answer. If we actually knew σ^2 but accounted for uncertainty in β_0 and β_1, it

turns out that our predictive confidence interval for the future $y^\star$ would be

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^\star \pm z^\star \sigma \left\{ 1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}, \tag{5.16}$$

where $z^\star$ is the normal critical value associated with our preferred confidence level (e.g. $z^\star = 1.96 \approx 2$ for a 95% interval). If you inspect this formula, you'll notice that the predictive interval gets wider as $x^\star$ gets further and further away from the average of the past observations ($\bar{x}$)—just like the bootstrapped prediction intervals in Figure 5.5.

But of course, since we don't know σ, we have to use $\hat{\sigma}$ instead. We therefore must also use $t^\star$ instead of $z^\star$, since in passing from a true variance to an estimated one, we pass from a normal distribution to a t distribution. This gives a prediction interval of

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^\star \pm t^\star \hat{\sigma} \left\{ 1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}. \tag{5.17}$$

This $t^\star$ is the critical value of a t_{n-2} distribution corresponding to our given confidence level. As always, if n is large and you want a 95% interval, feel free to use the normal approximation, $t^\star \approx 2$.

Advanced topic: derivation of the prediction interval

Where does this formula for the prediction interval come from? Let's write our prediction error as

$$e^\star = y^\star - \hat{y}^\star,$$

or the difference between the true value of $y^\star$ and what we will predict it to be. This quantity $e^\star$ is a random variable describing our predictive uncertainty. The old data is independent of the new data, so

$$\mathrm{Var}(e^\star) = \mathrm{Var}(y^\star) + \mathrm{Var}(\hat{y}^\star).$$

We can now explicitly see the two components of our uncertainty:

- the variance within the model due to the residuals, $\mathrm{Var}(y^\star)$.
- the variance due to parameter uncertainty, $\mathrm{Var}(\hat{y}^\star)$.

The first part is easy: the variance of the residuals is just σ^2. The second part, the variance due to our uncertainty in estimation, is a little harder. Let's write this part as

$$\mathrm{Var}(\hat{y}^\star) = \mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x^\star).$$

Now apply the following result from probability theory that describes the variance of a sum of random variables:

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\,\mathrm{Cov}(X_1, X_2)\,.$$

In other words, the variance of the sum is the sum of the variances, plus twice the covariance. Let's use this formula for $\widehat{\beta}_0$ and $\widehat{\beta}_1 x^\star$, which are both random variables.

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x^\star) &= \mathrm{Var}(\widehat{\beta}_0) + \mathrm{Var}(\widehat{\beta}_1 x^\star) + 2\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1 x^\star) \\ &= \mathrm{Var}(\widehat{\beta}_0) + \{x^\star\}^2\,\mathrm{Var}(\widehat{\beta}_1) + 2x^\star\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)\end{aligned}$$

But we know the variances of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$: these are just σ_0^2 and σ_1^2 from before. Therefore,

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x^\star) &= \mathrm{Var}(\widehat{\beta}_0) + \{x^\star\}^2\,\mathrm{Var}(\widehat{\beta}_1) + 2x^\star\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \left(\frac{\{x^\star\}^2\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + 2x^\star\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)\,.\end{aligned}$$

The covariance between the random variables $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = E\{(\widehat{\beta}_0 - \beta_0)(\widehat{\beta}_1 - \beta_1)\}\,.$$

A little algebra, which you are encouraged to try yourself, shows this to be

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\,.$$

Therefore,

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x^\star) &= \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\{x^\star\}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - 2\frac{x^\star\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\} \\ &= \sigma^2\left\{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\}\end{aligned}$$

Now we put all these pieces together, giving us

$$\mathrm{Var}(e^\star) = \mathrm{Var}(y^\star) + \mathrm{Var}(\hat{y}^\star) = \sigma^2\left\{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right\}\,,$$

Take the square root of this variance, and you end up with the predictive interval in Equation 5.16. Use $\hat{\sigma}^2$ instead, and you must use a t distribution rather than a normal, giving you Equation 5.17.