11
The normal linear regression model

Inference from probabilistic assumptions

One of the reasons we undertook a study of probability theory was for its own sake: to provide a language for understanding randomness and making smart decisions in the face of that randomness. We now turn to the second reason for studying probability: to sharpen our understanding of statistical modeling.

We recall that a statistical model decomposes variation into two parts: a predictable part and an unpredictable part. A *parametric statistical model* goes a bit further: it says that the unpredictable part can be described as a random variable arising from some probability distribution. This allows us to use probability theory to address the three important questions about uncertainty we studied several chapters ago:

(1) How confident are we in our estimate of a model parameter?

(2) How confident are we in our prediction for an observable quantity?

(3) Could an observed effect plausibly be explained by chance?

To help you understand this approach, let's see how it works in simple linear regression. You'll recall our original description of the regression relationship between a response y_i and a predictor x_i 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

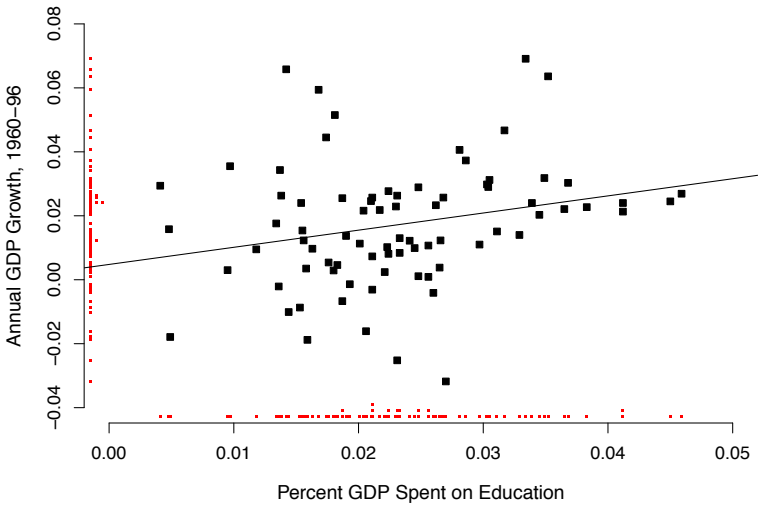


Figure 11.1: GDP growth (1960–1996) versus percentage of GDP spent on education for 79 countries.

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.

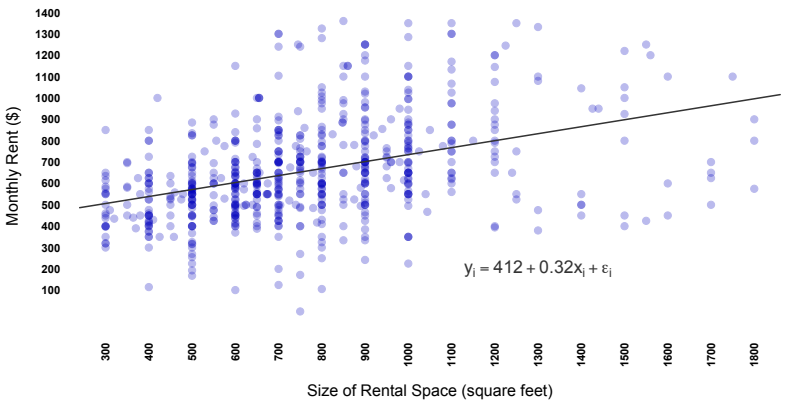
In principle, we could make any assumptions we want about the residuals. These assumptions will be embodied by the choice of the error distribution E. Of course, some assumptions are more reasonable than others And just as importantly, some assumptions are easier to leverage than others, if we wish to use probability theory to understand their consequences.

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. Many people believe that it strikes a nice balance between the goals of interpretability and calculational tractability. To justify it, we'll follow the same chain of reasoning that Gauss did. This chain has three links.

(1) Each residual is the aggregate result of forces not explicitly modeled. On the data set on GDP growth in Figure 11.1, 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—other forms of spending, life expectancy, 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 these random up-or-down nudges can be modeled with a normal distribution. 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. We can now

invoke Abraham de Moivre's celebrated result: independent coin flips can be described by a binomial distribution, and the binomial distribution can be approximated with a normal distribution.

The principle of maximum likelihood

This is precisely the line of argument that Gauss pursued. In his classic work of 1809 on fitting equations to the orbits of comets,[1] 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:

$$\begin{aligned} L(\beta_0, \beta_1) &= p(\epsilon_1) \cdot p(\epsilon_2) \cdots p(\epsilon_n) \\ &= \prod_{i=1}^{n} p(y_i - \beta_0 - \beta_1 x_i)\,, \end{aligned}$$

where the second equation follows from the fact that $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$. The closer the residuals are to the line, the higher the product of these probabilities.[2]

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*.

As a matter of fact, Gauss's argument to this effect was incomplete, for reasons not worth going into here. 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. 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.

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

[2] 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 our Russian friend 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.

(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.[3]

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

The normal linear regression model

Invoking Laplace's synthesis of de Moivre's and Gauss's earlier results, 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"? That is, 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; it's just a model. Like all models, it will be good at some things and bad at others. The important question is not whether the model is 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. Happily, it is straightforward to check whether the assumptions of the simple regression model look reasonable for a given data set. We'll see how to do this later by looking at the residuals from the fitted model.

Point estimation

Going forwards: from model to data

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?

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 four 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).

We then 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.

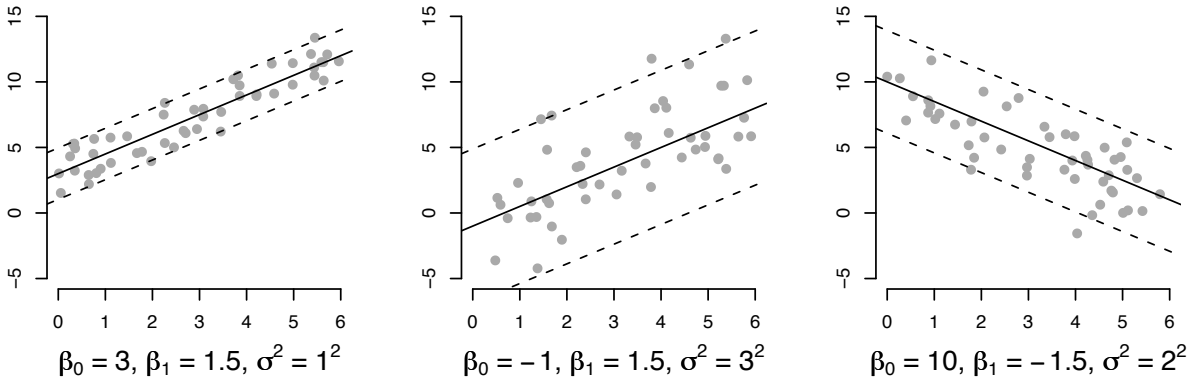


Figure 11.2: Three examples of data simulated according to the simple regression model.

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 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." This kind of inductive reasoning from known outcomes to plausible origins involves uncertainty. That's why we need probability theory to quantify just how far wrong we are likely to be.

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 11.4), we see five different data sets of size $n = 10$. All five have been simulated from the same model:

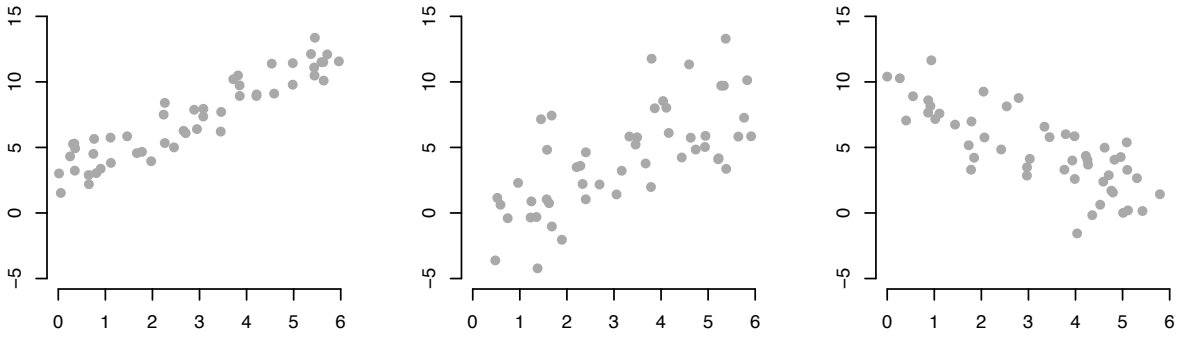


Figure 11.3: The same three data sets, without the true lines and true error bars.

$\beta_0 = -1$, $\beta_1 = 1.5$, 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 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.

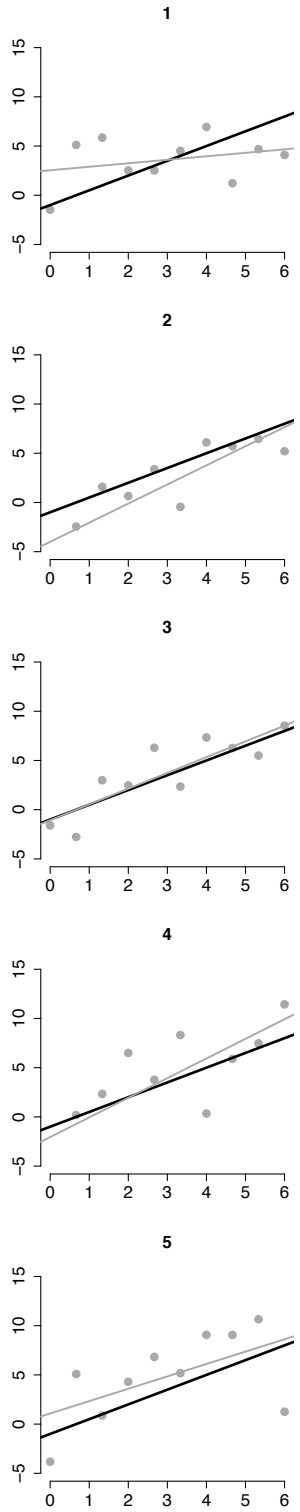


Figure 11.4: Five data sets of size $n = 10$ simulated from the same regression model: $\beta_0 = -1$, $\beta_1 = 1.5$, and $\sigma^2 = 3^2$.

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:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n \left\{y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right\}^2}{n-2}. \qquad (11.1)$$

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.

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.

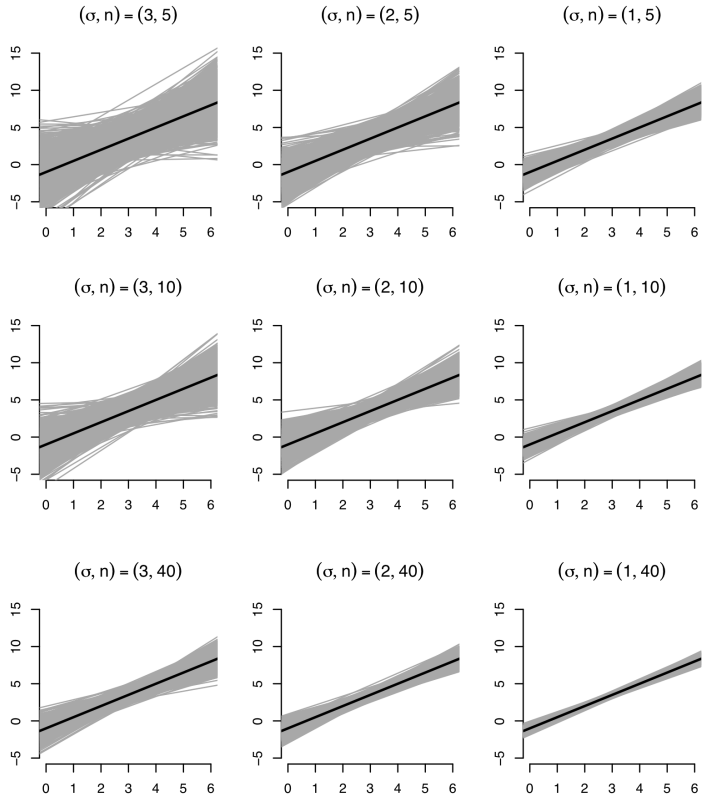


Figure 11.5: In each frame we see the true line, the least-squares fits for 1000 different simulated data sets, and the sample size (n) and residual standard deviation (σ) that were used to generate each data set.

Figure 11.5 can give you some idea of how these two crucial factors, σ and n, control the variation of the fitted line about the 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 11.4.

The direct parallel here is with Figure 5.3, back from our discussion of bootstrapping. 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,

$$\widehat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma_0^2\right) \tag{11.2}$$

$$\widehat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \sigma_1^2\right) . \tag{11.3}$$

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:

$$\sigma_0^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \tag{11.4}$$

$$\sigma_1^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} . \tag{11.5}$$

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:

$$\left(\frac{\widehat{\beta}_0 - \beta_0}{\sigma_0} \right) \sim \mathrm{N}(0,1) \tag{11.6}$$

$$\left(\frac{\widehat{\beta}_1 - \beta_1}{\sigma_1} \right) \sim \mathrm{N}(0,1) . \tag{11.7}$$

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 11.1. What if we made use of this fact by plugging in $\hat{\sigma}^2$ in lieu of the true (unknown) σ^2 in Equations 11.4 and 11.5?

$$\hat{\sigma}_0^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \tag{11.8}$$

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{11.9}$$

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. These are the estimated standard errors of the regression coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively. They approximate 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 chapter: 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 11.8 and 11.9). Under what circumstances will the standard errors be large? 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 11.6: 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.

Student's t distribution

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 11.6 and 11.7) 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:

$$t_0 = \left(\frac{\widehat{\beta}_0 - \beta_0}{\hat{\sigma}_0}\right) \sim t_{n-2} \tag{11.10}$$

$$t_1 = \left(\frac{\widehat{\beta}_1 - \beta_1}{\hat{\sigma}_1}\right) \sim t_{n-2}\,. \tag{11.11}$$

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

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^\star \pm z^\star \sigma \left\{ 1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}, \tag{11.12}$$

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 5.9.

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

$$\widehat{\beta}_0 + \widehat{\beta}_1 x^\star \pm t^\star \hat{\sigma} \left\{ 1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}^{1/2}. \tag{11.13}$$

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:

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

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.

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

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,

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

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

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

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

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

Therefore,

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

Now we put all these pieces together, giving us

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

Take the square root of this variance, and you end up with the predictive interval in Equation 11.12. Use $\hat{\sigma}^2$ instead, and you must use a t distribution rather than a normal, giving you Equation 11.13.

Hypothesis testing for regression coefficients

Recall that the t statistics are the standardized least-squares estimates of β_0 and β_1. To standardize, we subtract the mean and divide by the standard error:

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

These depend upon the unknown values β_0 and β_1—which means that the t statistics themselves must be unknown. When reporting t-statistics, the typical software package assumes you are interested in the null hypothesis that the corresponding coefficient is zero. In other words, it is computing

$$t_0 = \frac{\widehat{\beta}_0}{\hat{\sigma}_0} \quad \text{and} \quad t_1 = \frac{\widehat{\beta}_1}{\hat{\sigma}_1}\,.$$

Why? The answer to this is the answer to the following hypothetical question: what kind of data would we expect to see if our x and y variable had no linear relationship whatsoever? Mathematically, this would mean that β_1, the true underlying slope parameter of the simple regression model, is exactly zero. Let's see what happens when we simulate from a model like this:

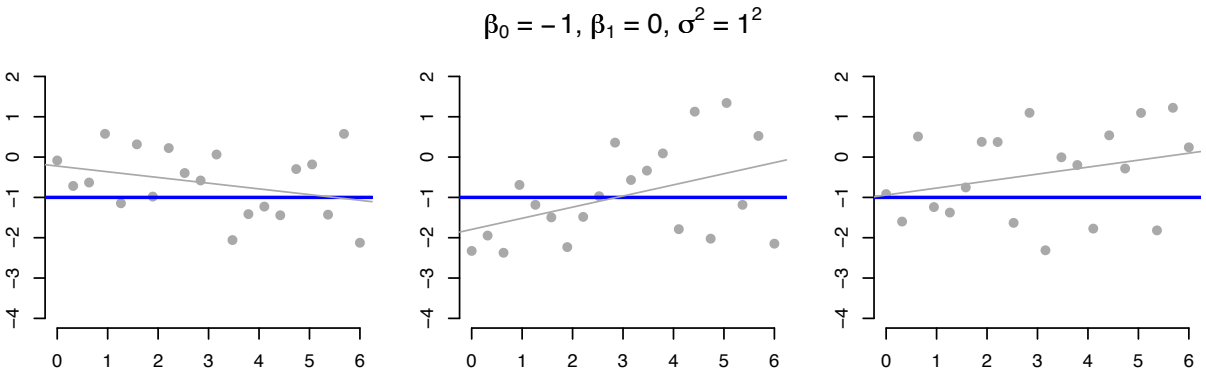


Figure 11.7: Three data sets of size $n = 20$ where the true slope is exactly zero.

If $\beta_0 = -1$ and $\sigma^2 = 1^2$, we get data that looks something like the above. The estimated slope will never be exactly zero, even if

the true slope is. And the same will be true of the intercept: $\widehat{\beta}_0$ will never be zero, even if β_0 is.

By how much will the estimated coefficients vary when the true coefficients are zero? We now have a whole theoretical apparatus for answering this question. If β_1 is zero, then the value of t_1 from your data set is just a random draw from a t distribution with $n-2$ degrees of freedom. To put this in terms that hark back to an earlier chapter: we know the *sampling distribution* of the t statistic under the null hypothesis that $\beta_1 = 0$.[4]

[4] Confusingly, t statistic is not short for test statistic, but refers to the t distribution. "Test statistic" is a more general term, applicable to any hypothesis-testing problem.

Using the t-statistic in a Neyman–Pearson test

This is all the information we need in order to conduct a Neyman–Pearson test of the null hypothesis that $\beta_1 = 0$. We could also test whether $\beta_0 = 0$, but let's focus on the slope; everything here also applies to testing the interception, for with $\widehat{\beta}_0$ and $\hat{\sigma}_0$ in place of $\widehat{\beta}_1$ and $\hat{\sigma}_1$.

To avoid confusion, let's use T_1 (with a capital letter) to denote a hypothetical value of the t statistic—that is, the value we might see in some parallel universe generated from the same underlying model. And let's use t_1 to denote the value of the t statistic for $\widehat{\beta}_1$ that you observed for your particular data set.

As in all Neyman–Pearson tests, there are three sub-steps to follow in deciding whether to reject the null hypothesis:

1. Choose an α that encodes your tolerance for false positives. Recall that α is the probability that you will reject the null hypothesis, given that the null hypothesis is true. The "industry standard" tends to be $\alpha = 0.05$, but as always, you should pick an α that you personally can live with.

2. Find the critical value $t^\star_\alpha$ corresponding to your chosen α. If your alternative hypothesis is that $\beta_1 \neq 0$, then you are performing a two-tailed test, and must find a $t^\star_\alpha$ such that

$$P(T_1 \geq t^\star_\alpha \mid \beta_1 = 0) = \alpha/2\,.$$

   Since the t distribution is symmetric, this will ensure that

$$P(T_1 \leq -t^\star_\alpha \mid \beta_1 = 0) = \alpha/2$$

   as well. In other words, you need to form a rejection region with $\alpha/2$ in the lower tail and $\alpha/2$ in the upper tail, thereby ensuring that the total probability of the rejection region is exactly α.

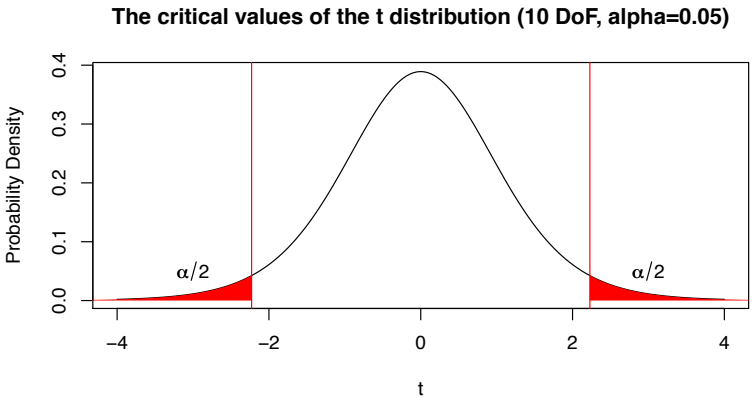


Figure 11.8: The critical values of the t_{10} distribution for a two-sided test at $\alpha = 0.05$.

If $n \geq 30$ and $\alpha = 0.05$, you can always just use the rule of thumb from before: $t^\star_\alpha \approx 2$. Otherwise, you can find a critical value using the R command `qt(1-alpha/2, n-2)`.

If, on the other hand, your alternative hypothesis is that $\beta_1 > 0$, then you are performing a one-tailed test, and must find a $t^\star_\alpha$ such that

$$P(T_1 \geq t^\star_\alpha \mid \beta_1 = 0) = \alpha\,.$$

That's because your rejection region must only account for values of T_1 greater than zero, since you're only interested in values of the slope that are greater than zero. In this case, use the R command `qt(1-alpha, n-2)`.

3. Finally, answer the question: does t_1 from your data set fall in the rejection region? For a two-sided test, the rejection region is all values of t_1 such that $|t_1| \geq t^\star_\alpha$. For a one-sided test, the rejection region is either $t_1 \geq t^\star_\alpha$ or $t_1 \leq t^\star_\alpha$, depending on whether your alternative hypothesis is $\beta_1 > 0$ or $\beta_1 < 0$. Either way, if t_1 falls in the rejection region, reject at the specified α level. If not, don't.

Finally, remember our rule of thumb: if n is 30 or more, then the t distribution is pretty close to the normal. Therefore, if the null hypothesis is true, then t_1 will be within $2\hat{\sigma}_1$ of zero 95% of the time.

This gives rise to a very simple guideline. If you're conducting a two-tailed test of $\beta_1 = 0$ at the $\alpha = 0.05$ level, then you should reject the null hypothesis if $\widehat{\beta}_1$ is more than twice its standard error. (Ditto $\widehat{\beta}_0$ and $\hat{\sigma}_0$.) Another way of phrasing this is: you should reject the null hypothesis at the $\alpha = 0.05$ level if the 95% confidence interval for β_1 fails to contain zero.

Testing for values other than zero

Regression software almost always assumes that you are testing the null hypothesis that $\beta_1 = 0$, and gives you the t statistics corresponding to this assumption. But with a simple calculation, you can easily test whether β_1 is equal to some value other than zero, instead.

Let's say your null hypothesis is that $\beta_1 = \theta$, where θ is a number other than zero. For example, in some situations, you might want to test whether a regression coefficient is exactly equal to one. How should you proceed?

Let's return to the definition of the t-statistic. If the true value of β_1 is really θ—in other words, if your null hypothesis is really true—then the following relationship holds:

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

This particular t_1 is not part of the standard regression output; the software is assuming that your θ is zero, and in this case it's not. But since you know θ, you can conduct a Neyman–Pearson test by just computing t_1 yourself according to the above formula: take the least-squares estimate, subtract θ, and divide by the standard error. With that one simple modification, you can then conduct a Neyman–Pearson test using the same steps above. Just remember to use your own "hand-calculated" t_1, and not the t_1 that the software calculates for you, in determining whether the data falls in the rejection region.

You can use a confidence interval to conduct a Neyman–Pearson test here, as well. If you want to test whether β_1 is significantly different from θ, then just check whether θ falls inside the $100(1 - \alpha)\%$ confidence interval for β_1.

Statistical versus practical significance

Remember not to confuse statistical significance with practical significance. Let's imagine the following example to illustrate the

difference. Suppose we take some data on an x variable and a y variable, and run a regression to understand the x–y relationship. We compute the least-squares estimate of the slope as $\widehat{\beta}_1 = 0.01$, and the standard error as $\hat{\sigma}_1 = 0.001$. Then if we test the hypothesis that $\beta_1 = 0$, our t statistic will be

$$t_1 = \frac{0.01}{0.001} = 10\,.$$

This is an enormous t statistic, and would lead us to reject the null hypothesis at pretty much any α level that wasn't absurdly small.

Should we reason, therefore, that x has a big effect on y—an effect that has been subjected to rigorous empirical scrutiny, and is substantiated by an enormous t statistic and a microscopic p value? Of course not! The 95% confidence interval for β_1 will be approximately $(0.008, 0.012)$. The data are telling us that, in all likelihood, x is useless for predicting y. After all, enormous changes in x are associated with only minuscule changes in y.

The moral of the story is: big t statistics and small p values can still happen even for tiny, insignificant effect sizes. This is especially likely to happen when you have a whole lot of data, since large samples allow you to estimate even tiny slopes with great precision.

Therefore, don't look merely at the t statistic or p value in summarizing an analysis. These quantities are not designed to distinguish a mountain from an anthill; they will only tell you that the ground isn't flat. Confidence intervals matter, too, and can save you the embarrassment of championing a result that is statistically significant, but practically useless.

F tests

Return to the regression model for a student's college GPA in terms of SAT scores and undergraduate college:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.691e+00  9.624e-02  17.566   <2e-16 ***
SAT.V                 1.486e-03  8.515e-05  17.455   <2e-16 ***
SAT.Q                 1.186e-03  9.098e-05  13.041   <2e-16 ***
SchoolBUSINESS        5.784e-03  7.827e-02   0.074   0.9411
SchoolCOMMUNICATIONS  8.565e-02  8.088e-02   1.059   0.2896
SchoolEDUCATION       4.492e-02  8.552e-02   0.525   0.5994
SchoolENGINEERING    -1.890e-01  7.851e-02  -2.408   0.0161 *
```

```
SchoolFINE ARTS         8.423e-03  8.443e-02   0.100   0.9205
SchoolLIBERAL ARTS     -1.374e-01  7.763e-02  -1.770   0.0767 .
SchoolNATURAL SCIENCE  -1.495e-01  7.789e-02  -1.920   0.0549 .
SchoolNURSING           2.423e-02  1.022e-01   0.237   0.8126
SchoolSOCIAL WORK      -3.787e-02  1.391e-01  -0.272   0.7854
```

You might notice that only a few of the individual dummy variables looked close to significant at the $\alpha = 0.05$ level. We could throw out the others, and just keep in the ones for which we reject the null hypothesis that the regression coefficient is zero (at some pre-specified α level, of course).

But what if we'd like to test the significance of our dummy variables *as a block*, rather than one at a time? After all, we'd like to think of the category itself as a variable, and the bunches of dummy variables as just a trick for letting the regression model take category membership into account. We might decide that it doesn't make sense to allow some categories to enter the regression model, but not others.

One way of doing this is look at how much R^2 improved after adding the category. With just SAT Math and SAT Verbal, R^2 was 0.155. Now with the nine dummy variables added in, R^2 is 0.185. This looks like a modest change in absolute terms, but does represent a relative change of about 20%. Is this change significant? We know by now, of course, that adding *any predictor at all* to a regression model, even a random one, will make R^2 jump by at least a little bit.

Hence the test of our hypothesis that category membership provides additional predictive power can be phrased as follows: could the addition of random predictors have plausibly made R^2 jump by at least as much as it jumped when we added the real predictors?

This is where the F-test comes in handy. As it turns out, it is possible to quantify precisely how much R^2 is expected to increase when we add predictors to a model that are uncorrelated with the response. This requires defining something called the F statistic:

$$f = \frac{\Delta R^2}{1 - R_F^2} \cdot \frac{n - p_F - 1}{p_F - p_R}, \tag{11.14}$$

where:

- R_F^2 is the value of R^2 under the full model—that is, the one that includes the block of variables you're interested in testing.

- $\Delta R^2 = R_F^2 - R_R^2$ is the gain in R^2 in moving from the Reduced model (without the block of tested coefficients) to the Full model (with the block of tested coefficients).
- p_F and p_R are the number of parameters, not including the intercept, in the Full and Reduced models, respectively.
- n is the number of observations in the sample.

This formula has a lot of pieces, but you can recognize F as essentially a rescaled version of ΔR^2. The rescaling takes into account how many variables are at stake, how many observations are in the sample, and how big R^2 was to begin with (that is, before adding the block of questionable variables).

The F statistic is the analogue of the t statistic for testing a single regression coefficient. Under the null hypothesis that the entire block of coefficients is zero, this statistic has what is known as an F distribution with $(p_F - p_R, n - p_F - 1)$ degrees of freedom. That is, if H_0 is true, then

$$(f \mid H_0) \sim F(p_F - p_R, n - p_F - 1)\,.$$

(Yes, that is indeed two distinct degrees-of-freedom parameters, compared to one for the t distribution.) Larger values of the F statistic correspond to greater evidence against the null hypothesis; if f is large enough, you reject the null. To find the critical value of the appropriate F distribution at a specified α level, use the R function qf$(1-\alpha, d_1, d_2)$, where d_1 and d_2 are the two degrees of freedom parameters.

For example, in our data on UT students:

- $R_F^2 = 0.185$ from the full model with all nine dummy variables in it
- $\Delta R^2 = R_F^2 - R_R^2 = 0.03$ is the gain in R^2 over the reduced model without the nine dummy variables
- $p_F = 11$ (two SAT scores, plus nine dummy variables), and $p_R = 2$ (just the two SAT scores)
- $n = 5191$

Putting these pieces together, we compute that

$$f = \frac{\Delta R^2}{1 - R_F^2} \cdot \frac{n - p_F - 1}{p_F - p_R} = \frac{0.03}{1 - 0.185} \cdot \frac{5191 - 11 - 1}{11 - 2} = 21.2$$

Suppose we want to conduct a Neyman–Pearson test with $\alpha = 0.01$. We must compare $f = 21.2$, the observed value of our test statistic, to $f^\star$, the 1% critical value of an $F(9, 5179)$ distribution. Here $f^\star$ turns out to be roughly 2.41. Therefore $f > f^\star$; we reject

the null hypothesis at the $\alpha = 0.01$ level and conclude that the block of dummy variables should be in the model, after all.

You can also use the F test to determine whether the R^2 value for a multiple regression model is significant compared to a model with no predictors at all. Think of it as an omnibus test for all the coefficients at once. To conduct this test, just use the same formulas above, plugging in $p_R = 0$ and $R_F^2 = \Delta R^2 = R^2$ (since the "reduced model" has no parameters and an R^2 of zero by definition). Everything else, including the computation of the critical value, is the same.