7.5 Statistical Significance: Uses and Abuses

7.5.1 *p*-Values

The test statistic for the Wald test,

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]}$$

has the nice, intuitive property that it ought to be close to zero when the null hypothesis $\beta_1 = \beta_1^*$ is true, and take large values (either positive or negative) when the null hypothesis is false. When a test statistic works like this, it makes sense to summarize just how bad the data looks for the null hypothesis in a ***p*-value**: when our observed value of the test statistic is T_{obs}, the p-value is

$$P = \mathbb{P}(|T| \geq |T_{obs}|)$$

calculating the probability under the null hypothesis. (I write a capital P here as a reminder that this is a random quantity, though it's conventional to write the phrase "p-value" with a lower-case p.) This is the probability, under the null, of getting results which are at least as extreme as what we saw. It should be easy to convince yourself that rejecting the null in a level-α test is the same as getting a p-value $< \alpha$.

It is not too hard (Exercise 4) to show that P has a uniform distribution over $[0,1]$ under the null hypothesis.

7.5.2 *p*-Values and Confidence Sets

When our test lets us calculate a p-value, we can form a $1-\alpha$ confidence set by taking all the β's where the p-value is $\geq \alpha$. Conversely, if we have some way of making confidence sets already, we can get a p-value for the hypothesis $\beta = \beta^*$; it's the largest α such that β^* is in the $1-\alpha$ confidence set.

7.5.3 Statistical Significance

If we test the hypothesis that $\beta_1 = \beta_1^*$ and reject it, we say that the difference between β_1 and β_1^* is **statistically significant**. Since, as I mentioned, many professions have an overwhelming urge to test the hypothesis $\beta_1 = 0$, it's common to hear people say that "β_1 is statistically significant" when they mean "β_1 is difference from 0 is statistically significant".

This is harmless enough, as long as we keep firmly in mind that "significant" is used here as a technical term, with a special meaning, and is *not* the same as "important", "relevant", etc. When we reject the hypothesis that $\beta_1 = 0$, what we're saying is "It's really implausibly hard to fit this data with a flat line, as opposed to one with a slope". This is informative, if we had serious reasons to think that a flat line was a live option.

It is incredibly common for researchers from other fields, and even some statisticians, to reason as follows: "I tested whether $\beta_1 = 0$ or not, and I retained the null; *therefore* β_1 is *in*significant, and I can ignore it." This is, of course, a complete fallacy.

To see why, it is enough to realize that there are (at least) two reasons why our hypothesis test might retain the null $\beta_1 = 0$:

1. β_1 is, in fact, zero,
2. $\beta_1 \neq 0$, but $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ is so large that we can't tell anything about β_1 with any confidence.

There is a very big difference between data which lets us say "we can be quite confident that the true β_1 is, if not perhaps exactly 0, then very small", and data which only lets us say "we have no earthly idea what β_1 is, and it may as well be zero for all we can tell"[11]. It is good practice to always compute a confidence interval, but it is *especially* important to do so when you retain the null, so you know whether you can say "this parameter is zero to within such-and-such a (small) precision", or whether you have to admit "I couldn't begin to tell you what this parameter is".

Substantive vs. statistical significance Even a huge β_1, which it would be crazy to ignore in any circumstance, can be statistically insignificant, so long as $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ is large enough. Conversely, any β_1 which isn't exactly zero, no matter how close it might be to 0, will become statistically significant at any threshold once $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ is small enough. Since, as $n \rightarrow \infty$,

$$\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] \rightarrow \frac{\sigma}{s_X \sqrt{n}}$$

we can show that $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] \rightarrow 0$, and $\frac{\hat{\beta}_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} \rightarrow \pm\infty$, unless β_1 is exactly 0 (see below).

Statistical significance is a weird mixture of how big the coefficient is, how big a sample we've got, how much noise there is around the regression line, and how spread out the data is along the x axis. This has so little to do with "significance" in ordinary language that it's pretty unfortunate we're stuck with the word; if the Ancestors had decided to say "statistically detectable" or "statistically distinguishable from 0", we might have avoided a lot of confusion.

If *you* confuse substantive and statistical significance in this class, it will go badly for you.

[11] Imagine hearing what sounds like the noise of an animal in the next room. If the room is small, brightly lit, free of obstructions, and you make a thorough search of it with unimpaired vision and concentration, not finding an animal in it is, in fact, good evidence that there was no animal there to be found. If on the other hand the room is dark, large, full of hiding places, and you make a hurried search while distracted, without your contact lenses and after a few too many drinks, you could easily have missed all sorts of things, and your negative report has little weight as evidence. (In this parable, the difference between a large $|\beta_1|$ and a small $|\beta_1|$ is the difference between looking for a Siberian tiger and looking for a little black cat.)

7.5.4 Appropriate Uses of p-Values and Significance Testing

I do not want this section to give the impression that p-values, hypothesis testing, and statistical significance are unimportant or necessarily misguided. They're often used badly, but that's true of every statistical tool from the sample mean on down the line. There are certainly situations where we really do want to know whether we have good evidence against some *exact* statistical hypothesis, and that's just the job these tools do. What are some of these situations?

Model checking Our statistical models often make very strong, claims about the probability distribution of the data, with little wiggle room. The simple linear regression model, for instance, claims that the regression function is *exactly* linear, and that the noise around this line has *exactly* constant variance. If we test these claims and find very small p-values, then we have evidence that there's a detectable, systematic departure from the model assumptions, and we should re-formulate the model.

Actual scientific interest Some scientific theories make very precise predictions about coefficients. According to Newton, the gravitational force between two masses is inversely proportional to the *square* of the distance between them, $\propto r^{-2}$. The prediction is exactly $\propto r^{-2}$, not $\propto r^{-1.99}$ nor $\propto r^{-2.05}$. Measuring that exponent and finding even tiny departures from 2 would be big news, if we had reason to think they were real and not just noise[12]. One of the most successful theories in physics, quantum electrodynamics, makes predictions about some properties of hydrogen atoms with a theoretical precision of one part in a trillion; finding even tiny discrepancies between what the theory predicts and what we estimate would force us to rethink lots of physics[13]. Experiments to detect new particles, like the Higgs boson, essentially boil down to hypothesis testing, looking for deviations from theoretical predictions which should be exactly zero if the particle doesn't exist.

Outside of the natural sciences, however, it is harder to find examples of interesting, exact null hypothesis which are, so to speak, "live options". The best I can come up with are theories of economic growth and business cycles which predict that the share of national income going to labor (as opposed to capital) should be constant over time. Otherwise, in the social sciences, there's usually little theoretical reason to think that certain regression coefficients should be *exactly* zero, or *exactly* one, or anything else.

Neutral models A partial exception is the use of **neutral models**, which comes out of genetics and ecology. The idea here is to check whether some mechanism is at work in a particular situation — say, whether some gene is subject to natural selection. One constructs two models; one incorporates all the mechanisms (which we think are) at work, including the one under investigation, and the other incorporate all the *other* mechanisms, but "neutralizes" the one of interest. (In a genetic example, the neutral

[12] In fact, it *was* big news: Einstein's theory of general relativity.

[13] Feynman (1985) gives a great conceptual overview of quantum electrodynamics. Currently, theory agrees with experiment to the limits of experimental precision, which is only about one part in a billion (https://en.wikipedia.org/wiki/Precision_tests_of_QED).

model would probably incorporate the effects of mutation, sexual repdouction, the random sampling of which organisms become the ancestors of the next generation, perhaps migration, etc. The non-neutral model would include all this *plus* the effects of natural selection.) Rejecting the neutral model in favor of the non-neutral one then becomes evidence that the disputed mechanism is needed to explain the data.

In the cases where this strategy has been done well, the neutral model is usually a pretty sophisticated stochastic model, and the "neutralization" is not as simple as just setting some slope to zero. Nonetheless, this is a situation where we do actually learn something about the world by testing a null hypothesis.

7.6 Any Non-Zero Parameter Becomes Significant with Enough Information

(This section is optional, but strongly recommended.)

Let's look more close at what happens to the test statistic when $n \rightarrow \infty$, and so at what happens to the p-value. Throughout, we'll be testing the null hypothesis that $\beta_1 = 0$, since this is what people most often do, but the same reasoning applies to departures from any fixed value of the slope. (Everything carries over with straightforward changes to testing hypotheses about the intercept β_0, too.)

We know that $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/ns_X^2)$. This means[14]

$$\begin{aligned} \hat{\beta}_1 &\sim \beta_1 + N(0, \sigma^2/ns_X^2) && (7.35) \\ &= \beta_1 + \frac{\sigma}{s_X\sqrt{n}} N(0,1) && (7.36) \\ &= \beta_1 + O(1/\sqrt{n}) && (7.37) \end{aligned}$$

where $O(f(n))$ is read "order-of $f(n)$", meaning that it's a term whose size grows like $f(n)$ as $n \rightarrow \infty$, and we don't want (or need) to keep track of the details. Similarly, since $nsi\hat{g}ma^{2}/\sigma^2 \sim \chi^2_{n-2}$, we have[15]

$$\begin{aligned} n\hat{\sigma}^2 &\sim \sigma^2 \chi^2_{n-2} && (7.38) \\ \hat{\sigma}^2 &\sim \sigma^2 \frac{\chi^2_{n-2}}{n} && (7.39) \end{aligned}$$

Since $\mathbb{E}\left[\chi^2_{n-2}\right] = n-2$ and $\mathrm{Var}\left[\chi^2_{n-2}\right] = 2(n-2)$,

$$\begin{aligned} \mathbb{E}\left[\frac{\chi^2_{n-2}}{n}\right] &= \frac{n-2}{n} \rightarrow 1 && (7.40) \\ \mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] &= \frac{2(n-2)}{n^2} \rightarrow 0 && (7.41) \end{aligned}$$

[14] If seeing something like $\frac{\sigma}{s_X\sqrt{n}}N(0,1)$, feel free to introduce random variables $Z_n \sim N(0,1)$ (though not necessarily independent ones), and modify the equations accordingly.

[15] Again, feel free to introduce the random variable Ξ_n, which just so happens to have a χ^2_{n-2} distribution.

with both limits happening as $n \rightarrow \infty$. In fact $\mathrm{Var}\left[\frac{\chi^2_{n-2}}{n}\right] = O(1/n)$, so

$$\hat{\sigma}^2 = \sigma^2\left(1 + O(1/\sqrt{n})\right) \tag{7.42}$$

Taking the square root, and using the fact[16] that $(1+x)^a \approx 1 + ax$ when $|x| \ll 1$,

$$\hat{\sigma} = \sigma\left(1 + O(1/\sqrt{n})\right) \tag{7.43}$$

Put this together to look at our test statistic:

$$\begin{aligned} \frac{\hat{\beta}_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} &= \frac{\beta_1 + O(1/\sqrt{n})}{\frac{\sigma\left(1+O(1/\sqrt{n})\right)}{s_X\sqrt{n}}} && (7.44)\\ &= \sqrt{n}\frac{\beta_1 + O(1/\sqrt{n})}{(\sigma/s_X)\left(1+O(1/\sqrt{n})\right)} && (7.45)\\ &= \sqrt{n}\frac{\beta_1}{\sigma/s_X}\left(1+O(1/\sqrt{n})\right) && (7.46)\\ &= \sqrt{n}\frac{\beta_1}{\sigma/s_X} + O(1) && (7.47) \end{aligned}$$

In words: so long as the true $\beta_1 \neq 0$, the test statistic is going to go off to $\pm\infty$, and the rate at which it escapes towards infinity is going to be proportional to $\sqrt{n}$. When we compare this against the null distribution, which is $N(0,1)$, eventually we'll get arbitrarily small p-values.

We can actually compute what those p-values should be, by two bounds on the standard Gaussian distribution[17]:

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\frac{1}{\sqrt{2\pi}}e^{-x^2/2} < 1 - \Phi(x) < \frac{1}{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{7.48}$$

Thus

$$\begin{aligned} P_n &= \mathbb{P}\left(|Z| \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}a\right|\right) && (7.49)\\ &= 2\mathbb{P}\left(Z \geq \left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right|\right) && (7.50)\\ &\leq \frac{2}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}\frac{\hat{\beta}_1^2}{\hat{\sigma}^2/ns_X^2}}}{\left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right|} && (7.51) \end{aligned}$$

[16] From the binomial theorem, back in high school algebra.

[17] See Feller (1957), Chapter VII, §1, Lemma 2. For a brief proof online, see `http://www.johndcook.com/normalbounds.pdf`.

To clarify the behavior, let's take the logarithm and divide by n:

$$\begin{aligned}\frac{1}{n}\log P_n &\leq \frac{1}{n}\log\frac{2}{\sqrt{2\pi}} \\ &\quad -\frac{1}{n}\log\left|\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{n}s_X}\right| \\ &\quad -\frac{1}{2n}\frac{\hat{\beta}_1^2}{\hat{\sigma}^2/ns_X^2}\end{aligned} \tag{7.52}$$

$$\begin{aligned}&= \frac{\log\sqrt{2\pi}}{n} \\ &\quad +\frac{\log\left|\frac{\hat{\beta}_1}{\hat{\sigma}/s_x}\right|}{n} \\ &\quad -\frac{\log n}{2n} \\ &\quad -\frac{\hat{\beta}_1^2}{2\hat{\sigma}^2/s_X^2}\end{aligned} \tag{7.53}$$

Take the limit as $n \to \infty$:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log P_n &\leq \lim_{n}\frac{\log\sqrt{2\pi}}{n} \\ &\quad +\lim_{n}\frac{\log\frac{\hat{\beta}_1}{\hat{\sigma}/s_x}}{n} \\ &\quad -\lim_{n}\frac{\log n}{2n} \\ &\quad -\lim_{n}\frac{\hat{\beta}_1^2}{2\hat{\sigma}^2/s_X^2}\end{aligned} \tag{7.54}$$

Since $\hat{\beta}_1/(\hat{\sigma}/s_X) \to \beta_1/(\sigma/s_X)$, and $n^{-1}\log n \to 0$,

$$\lim_{n\to\infty}\frac{1}{n}\log P_n \leq -\frac{\beta_1^2}{2\sigma^2/s_X^2} \tag{7.55}$$

I've only used the upper bound on $1-\Phi(x)$ from Eq. 7.48; if we use the lower bound from that equation, we get (Exercise 5)

$$\lim_{n\to\infty}\frac{1}{n}\log P_n \geq -\frac{\beta_1^2}{2\sigma^2/s_X^2} \tag{7.56}$$

Putting the upper and lower limits together,

$$\lim_{n\to\infty}\frac{1}{n}\log P_n = -\frac{\beta_1^2}{2\sigma^2/s_X^2}$$

Turn the limit around: at least for large n,

$$P_n \approx e^{-n \frac{\beta_1^2}{2\sigma^2/s_X^2}} \tag{7.57}$$

Thus, *any* $\beta_1 \neq 0$ will (eventually) give exponentially small p-values. This is why, as a saying among statisticians have it, "the p-value is a measure of sample size": any non-zero coefficient will become arbitrarily statistically significant with enough data. This is just another way of saying that with enough data, we can (and will) detect even arbitrarily small coefficients, which is what we *want*. The flip-side, however, is that it's just senseless to say that one coefficient is important because it has a really small p-value and another is unimportant because it's got a big p-value. As we can see from Eq. 7.57, the p-value runs together the magnitude of the coefficient ($|\beta_1|$), the sample size (n), the noise around the regression line (σ^2), and how spread out the data is along the x axis (s_X^2), the last of these because they control how precisely we can estimate β_1. Saying "this coefficient must be really important, because we can measure it really precisely" is not smart.

7.7 Confidence Sets and *p*-Values in R

When we estimate a model with `lm`, R makes it easy for us to extract the confidence intervals of the coefficients:

```
confint(object, level = 0.95)
```

Here `object` is the name of the fitted model object, and `level` is the confidence level; if you want 95% confidence, you can omit that argument. For instance:

```
library(gamair)
data(chicago)
death.temp.lm <- lm(death ~ tmpd, data = chicago)
confint(death.temp.lm)
##                   2.5 %     97.5 %
## (Intercept) 128.8783687 131.035734
## tmpd         -0.3096816  -0.269607
confint(death.temp.lm, level = 0.9)
##                     5 %        95 %
## (Intercept) 129.0518426 130.8622598
## tmpd         -0.3064592  -0.2728294
```

If you want p-values for the coefficients[18], those are conveniently computed as part of the `summary` function:

```
coefficients(summary(death.temp.lm))
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 129.9570512 0.55022802 236.18763  0.00000e+00
## tmpd         -0.2896443 0.01022089 -28.33845 3.23449e-164
```

[18] And, really, why do you?