Chapter 7

Inference on Parameters

Having gone over the Gaussian-noise simple linear regression model, over ways of estimating its parameters and some of the properties of the model, and over how to check the model's assumptions, we are now ready to begin doing some serious statistical inference within the model[1]. In previous chapters, we came up with **point estimators** of the parameters and the conditional mean (prediction) function, but we weren't able to say much about the margin of uncertainty around these estimates. In this chapter we will focus on supplementing point estimates with *reliable* measures of uncertainty. This will naturally lead us to testing hypotheses about the true parameters — again, we will want hypothesis tests which are unlikely to get the answer wrong, whatever the truth might be.

To accomplish all this, we first need to understand the sampling distribution of our point estimators. We can find them, mathematically, but they involve the unknown true parameters in inconvenient ways. We will therefore work to find combinations of our estimators and the true parameters with fixed, parameter-free distributions; we'll get our confidence sets and our hypothesis tests from them.

Throughout this chapter, I am assuming, unless otherwise noted, that all of the assumptions of the Gaussian-noise simple linear regression model hold. After all, we checked those assumptions last time....

7.1 Sampling Distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$

The Gaussian-noise simple linear regression model has three parameters: the intercept β_0, the slope β_1, and the noise variance σ^2. We've seen, previously, how to estimate all of these by maximum likelihood; the MLE for the βs is the same as their least-

[1] Presuming, of course, that the model's assumptions, when thoroughly checked, do in fact hold good.

squares estimates. These are

$$\hat{\beta}_1 = \frac{c_{XY}}{s^2_X} = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s^2_X} y_i \tag{7.1}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{7.2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{7.3}$$

We have also seen how to re-write the first two of these as a deterministic part plus a weighted sum of the noise terms ϵ:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s^2_X} \epsilon_i \tag{7.4}$$

$$\hat{\beta}_0 = \beta_0 + \frac{1}{n} \sum_{i=1}^{n} \left(1 - \overline{x} \frac{x_i - \overline{x}}{s^2_X}\right) \epsilon_i \tag{7.5}$$

Finally, we have our modeling assumption that the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$.

7.1.1 Reminders of Basic Properties of Gaussian Distributions

Suppose $U \sim N(\mu, \sigma^2)$. By the basic algebra of expectations and variances, $\mathbb{E}[a + bU] = a + b\mu$, while $\mathrm{Var}[a + bU] = b^2\sigma^2$. This would be true of any random variable; a special property of Gaussians[2] is that $a + bU \sim N(a + b\mu, b^2\sigma^2)$.

Suppose $U_1, U_2, \ldots U_n$ are *independent* Gaussians, with means μ_i and variances σ^2_i. Then

$$\sum_{i=1}^{n} U_i \sim N(\sum_i \mu_i, \sum_i \sigma^2_i)$$

That the expected values add up for a sum is true of all random variables; that the variances add up is true for all uncorrelated random variables. That the sum follows the same type of distribution as the summands is a special property of Gaussians[3].

7.1.2 Sampling Distribution of $\hat{\beta}_1$

Since we're assuming Gaussian noise, the ϵ_i are independent Gaussians, $\epsilon_i \sim N(0, \sigma^2)$. Hence (using the first basic property of Gaussians)

$$\frac{x_i - \overline{x}}{n s^2_X} \epsilon_i \sim N(0, \left(\frac{x_i - \overline{x}}{n s^2_X}\right)^2 \sigma^2)$$

[2] There some other families of distributions which have this property; they're called "location-scale" families.

[3] There are some other families of distributions which have this property; they're called "stable" families.

```
# Simulate a Gaussian-noise simple linear regression model Inputs: x
# sequence; intercept; slope; noise variance; switch for whether to return
# the simulated values, or run a regression and return the coefficients
# Output: data frame or coefficient vector
sim.gnslrm <- function(x, intercept, slope, sigma.sq, coefficients = TRUE) {
    n <- length(x)
    y <- intercept + slope * x + rnorm(n, mean = 0, sd = sqrt(sigma.sq))
    if (coefficients) {
        return(coefficients(lm(y ~ x)))
    } else {
        return(data.frame(x = x, y = y))
    }
}

# Fix an arbitrary vector of x's
x <- seq(from = -5, to = 5, length.out = 42)
```

FIGURE 7.1: *Code setting up a simulation of a Gaussian-noise simple linear regression model, along a fixed vector of x_i values.*

Thus, using the second basic property of Gaussians,

$$\sum_{i=1}^{n} \frac{x_i - \overline{x}}{n s_X^2} \epsilon_i \sim N(0, \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{n s_X^2}\right)^2) \tag{7.6}$$

$$= N(0, \frac{\sigma^2}{n s_X^2}) \tag{7.7}$$

Using the first property of Gaussians again,

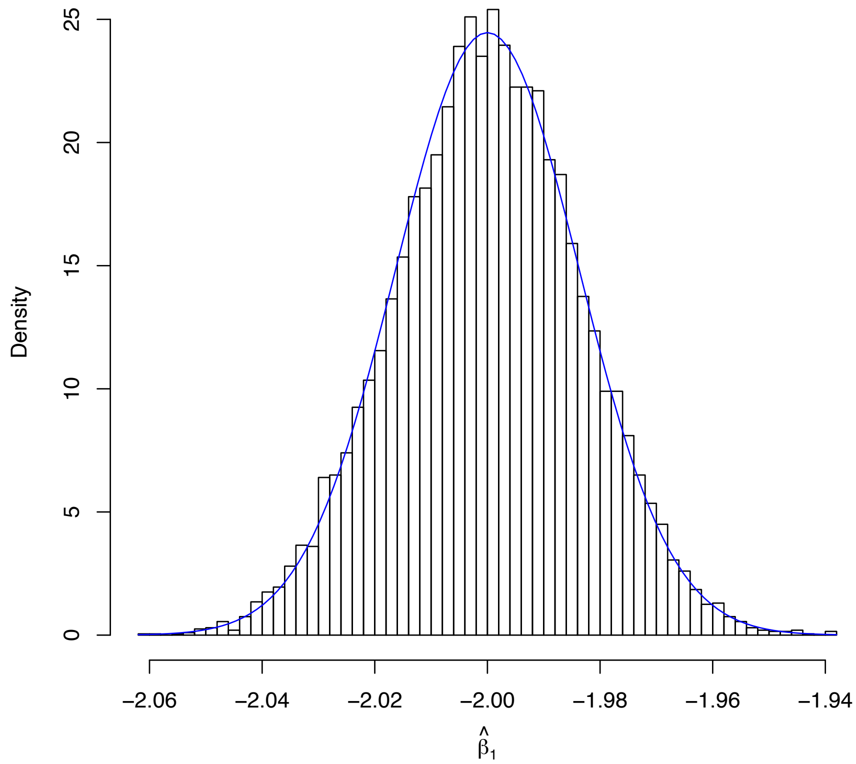
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n s_X^2}) \tag{7.8}$$

This is the distribution of estimates we'd see if we repeated the experiment (survey, observation, etc.) many times, and collected the results. Every particular run of the experiment would give a slightly different $\hat{\beta}_1$, but they'd average out to β_1, the average squared difference from β_1 would be $\sigma^2 / n s_X^2$, and a histogram of them would follow the Gaussian probability density function (Figure 7.2).

It is a bit hard to use Eq. 7.8, because it involves two of the unknown parameters. We can manipulate it a bit to remove one of the parameters from the probability distribution,

$$\hat{\beta}_1 - \beta_1 \sim N(0, \frac{\sigma^2}{n s_X^2})$$

but that still has σ^2 on the right hand side, so we can't actually calculate anything. We



```
# Run the simulation 10,000 times and collect all the coefficients What
# intercept, slope and noise variance does this impose?
many.coefs <- replicate(10000, sim.gnslrm(x = x, 5, -2, 0.1, coefficients = TRUE))
# Histogram of the slope estimates
hist(many.coefs[2, ], breaks = 50, freq = FALSE, xlab = expression(hat(beta)[1]),
    main = "")
# Theoretical Gaussian sampling distribution
theoretical.se <- sqrt(0.1/(length(x) * var(x)))
curve(dnorm(x, mean = -2, sd = theoretical.se), add = TRUE, col = "blue")
```

FIGURE 7.2: *Simulating 10,000 runs of a Gaussian-noise simple linear regression model, calculating $\hat{\beta}_1$ each time, and comparing the histogram of estimates to the theoretical Gaussian distribution (Eq. 7.8, in blue).*

could write

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma^2/\sqrt{ns_X^2}} \sim N(0,1)$$

but now we've got two unknown parameters on the left-hand side, which is also awkward.

7.1.3 Sampling Distribution of $\hat{\beta}_0$

Starting from Eq. 7.5 rather than Eq. 7.4, an argument exactly parallel to the one we just went through gives

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right))$$

It follows, again by parallel reasoning, that

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_X^2}\right)}} \sim N(0,1)$$

The right-hand side of this equation is admirably simple and easy for us to calculate, but the left-hand side unfortunately involves two unknown parameters, and that complicates any attempt to use it.

7.1.4 Sampling Distribution of $\hat{\sigma}^2$

It is mildly challenging, but certainly not too hard, to show that

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-2}{n}\sigma^2$$

As I have said before, this will be a problem on a future assignment, so I will not give a proof, but I will note that the way to proceed is to write

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 ;$$

then to write each residual e_i as a weighted sum of the noise terms ϵ; to use $\mathbb{E}\left[e_i^2\right] = \mathrm{Var}\left[e_i\right] + (\mathbb{E}\left[e_i\right])^2$; and finally to sum up over i.

Notice that this implies that $\mathbb{E}\left[\hat{\sigma}^2\right] = 0$ when $n = 2$. This is because any two points in the plane define a (unique) line, so if we have only two data points, least squares will just run a line through them exactly, and have an in-sample MSE of 0. In general, we get the factor of $n-2$ from the fact that we are estimating two parameters.

We can however be much more specific. When $\epsilon_i \sim N(0, \sigma^2)$, it can be shown that

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Notice, by the way, that this equation involves no unknown parameters on the right-hand side, and only one on the left-hand side. It lets us calculate the probability that $\hat{\sigma}^2$ is within any given *factor* of σ^2. If, for instance, we wanted to know the probability that $\hat{\sigma}^2 \geq 7\sigma^2$, this will let us find it.

I will offer only a hand-waving explanation; I am afraid I am not aware of any truly elementary mathematical explanation — every one I know of either uses probability facts which are about as obscure as the result to be shown, or linear-algebraic facts about the properties of idempotent matrices[4], and we've not seen, *yet*, how to write linear regression in matrix form. I do however want to re-assure you that there are actual proofs, and I promise to include one in these notes once we've seen how to connect what we're doing to matrices and linear algebra.

I am afraid I do not have even a hand-waving explanation of a second important property of $\hat{\sigma}^2$: it is statistically independent of $\hat{\beta}_0$ and $\hat{\beta}_1$. This is *not* obvious — after all, all three of these estimators are functions of the same noise variables ϵ — but it *is* true, and, again, I promise to provide a genuine proof in these notes once we've gone over the necessary math.

7.1.4.1 The Hand-Waving Explanation for $n-2$

Let's think for a moment about a related (but strictly different!) quantity from $\hat{\sigma}^2$, namely

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$

This is a weighted sum of independent, mean-zero squared Gaussians, which is where the connection to χ^2 distributions comes in.

Some reminders about χ^2 If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$ *by definition* (of the χ^2_1 distribution). From this, it follows that $\mathbb{E}\left[\chi^2_1\right] = 1$, $\mathrm{Var}\left[\chi^2_1\right] = \mathbb{E}\left[Z^4\right] - (\mathbb{E}\left[Z^2\right])^2 = 2$. If $Z_1, Z_2, \ldots Z_d \sim N(0,1)$ and are independent, then the χ^2_d distribution is *defined* to be the distribution of $\sum_{i=1}^{d} Z_i^2$. By simple algebra, it follows that $\mathbb{E}\left[\chi^2_d\right] = d$ while $\mathrm{Var}\left[\chi^2_d\right] = 2d$.

Back to the sum of squared noise terms ϵ_i isn't a standard Gaussian, but ϵ_i/σ is, so

$$\frac{\sum_{i=1}^{n}\epsilon_i^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{\epsilon_i}{\sigma})^2 \sim \chi^2_n$$

The numerator here is *like* $n\hat{\sigma}^2 = \sum_i e_i^2$, but of course the residuals e_i are not the same as the noise terms ϵ_i.

The reason we end up with a χ^2_{n-2} distribution, rather than a χ^2_n distribution, is that we're estimating two parameters from the data removes two degrees of freedom, so two of the ϵ_i end up making no real contribution to the sum of squared errors. (Again, if $n = 2$, we'd be able to fit the two data points *exactly* with the least squares

[4] Where $M^2 = M$.

line.) If we had estimated more or fewer parameters, we would have had to adjust the number of degrees of freedom accordingly.

(There is also a geometric interpretation: the sum of squared errors, $\sum_{i=1}^{n} e_i^2$, is the squared length of the n-dimensional vector of residuals, $(e_1, e_2, \ldots e_n)$. But the residuals must obey the two equations $\sum_i e_i = 0$, $\sum_i x_i e_i = 0$, so the residual vector actually is confined to an $(n-2)$-dimensional linear subspace. Thus we only end up adding up $(n-2)$ *independent* contributions to its length. If we estimated more parameters, we'd have more estimating equations, and so more constraints on the vector of residuals.)

7.1.5 Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

The **standard error** of an estimator is its standard deviation[5]. We've just seen that the true standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are, respectively,

$$\mathrm{se}\left[\hat{\beta}_1\right] = \frac{\sigma}{s_x \sqrt{n}} \tag{7.9}$$

$$\mathrm{se}\left[\hat{\beta}_0\right] = \frac{\sigma}{\sqrt{n} s_X} \sqrt{s_X^2 + \overline{x}^2} \tag{7.10}$$

Unfortunately, these standard errors involve the unknown parameter σ^2 (or its square root σ, equally unknown to us).

We can, however, *estimate* the standard errors. The maximum-likelihood estimates just substitute $\hat{\sigma}$ for σ:

$$\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{s_x \sqrt{n}} \tag{7.11}$$

$$\widehat{\mathrm{se}}\left[\hat{\beta}_0\right] = \frac{\hat{\sigma}}{s_X \sqrt{n}} \sqrt{s_X^2 + \overline{x}^2} \tag{7.12}$$

For later theoretical purposes, however, things will work out slightly nicer if we use the de-biased version, $\frac{n}{n-2}\hat{\sigma}^2$:

$$\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] = \frac{\hat{\sigma}}{s_x \sqrt{n-2}} \tag{7.13}$$

$$\widehat{\mathrm{se}}\left[\hat{\beta}_0\right] = \frac{\hat{\sigma}}{s_x \sqrt{n-2}} \sqrt{s_X^2 + \overline{x}^2} \tag{7.14}$$

These standard errors — approximate or estimated though they be — are one important way of quantifying how much uncertainty there is around our point estimates. However, we can't use them, *alone* to say anything terribly precise[6] about, say, the probability that β_1 is in the interval $[\hat{\beta}_1 - \widehat{\mathrm{se}}\left[\hat{\beta}_1\right], \hat{\beta}_1 - \widehat{\mathrm{se}}\left[\hat{\beta}_1\right]]$, which is

[5] We don't just call it the standard deviation because we want to emphasize that it is, in fact, telling us about the random errors our estimator makes.

[6] Exercise to think through: Could you use Chebyshev's inequality (the extra credit problem from Homework 1) here?

the sort of thing we'd want to be able to give guarantees about the reliability of our estimates.

7.2 Sampling distribution of $(\hat{\beta}-\beta)/\widehat{\mathrm{se}}\left[\hat{\beta}\right]$

It should take only a little work with the properties of the Gaussian distribution to convince yourself that

$$\frac{\hat{\beta}_1-\beta_1}{\mathrm{se}\left[\hat{\beta}_1\right]} \sim N(0,1)$$

the standard Gaussian distribution. If the Oracle told us σ^2, we'd know $\mathrm{se}\left[\hat{\beta}_1\right]$, and so we could assert that (for example)

$$\begin{aligned} & \mathbb{P}\left(\beta_1-1.96\mathrm{se}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 \leq \beta_1+1.96\mathrm{se}\left[\hat{\beta}_1\right]\right) && (7.15)\\ = \; & \mathbb{P}\left(-1.96\mathrm{se}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1-\beta_1 \leq 1.96\mathrm{se}\left[\hat{\beta}_1\right]\right) && (7.16)\\ = \; & \mathbb{P}\left(-1.96 \leq \frac{\hat{\beta}_1-\beta_1}{\mathrm{se}\left[\hat{\beta}_1\right]} \leq 1.96\right) && (7.17)\\ = \; & \Phi(1.96)-\Phi(-1.96)=0.95 && (7.18) \end{aligned}$$

where Φ is the cumulative distribution function of the $N(0,1)$ distribution.

Since the oracles have fallen silent, we can't use this approach. What we *can* do is use the following fact[7]:

Proposition 1 *If $Z \sim N(0,1)$, $S^2 \sim \chi^2_d$, and Z and S^2 are independent, then*

$$\frac{Z}{\sqrt{S^2/d}} \sim t_d$$

(I call this a proposition, but it's almost a definition of what we mean by a t distribution with d degrees of freedom. Of course, if we take this as the definition, the proposition that this distribution has a probability density $\propto (1+x^2/d)^{-(d+1)/2}$ would become yet another proposition to be demonstrated.)

Let's try to manipulate $(\hat{\beta}_1-\beta_1)/\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ into this form.

[7] When I messed up the derivation in class today, I left out dividing by d in the denominator. As I mentioned at the end of that debacle, this was stupid. As $d \to \infty$, t_d converges on the standard Gaussian distribution $N(0,1)$. (Notice that $\mathbb{E}\left[d^{-1}\chi^2_d\right]=1$, while $\mathrm{Var}\left[d^{-1}\chi^2_d\right]=2/d$, so $d^{-1}\chi^2_d \to 1$.) Without the normalizing factor of d inside the square root, however, looking just at Z/S, we've got a random variable whose distribution doesn't change with d being divided by something whose magnitude *grows* with d, so $Z/S \to 0$ as $d \to \infty$, not $\to N(0,1)$. I apologize again for my error.

$$\begin{aligned}
\frac{\hat{\beta}_1-\beta_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} &= \frac{\hat{\beta}_1-\beta_1}{\sigma}\frac{\sigma}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} && (7.19)\\
&= \frac{\frac{\hat{\beta}_1-\beta_1}{\sigma}}{\frac{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]}{\sigma}} && (7.20)\\
&= \frac{N(0,1/ns_X^2)}{\frac{\hat{\sigma}}{s_x\sigma\sqrt{n-2}}} && (7.21)\\
&= \frac{s_X N(0,1/ns_X^2)}{\frac{\hat{\sigma}}{\sigma\sqrt{n-2}}} && (7.22)\\
&= \frac{N(0,1/n)}{\frac{\hat{\sigma}}{\sigma\sqrt{n-2}}} && (7.23)\\
&= \frac{\sqrt{n}N(0,1/n)}{\frac{\sqrt{n}\hat{\sigma}}{\sigma\sqrt{n-2}}} && (7.24)\\
&= \frac{N(0,1)}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}\frac{1}{n-2}}} && (7.25)\\
&= \frac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} && (7.26)\\
&= t_{n-2} && (7.27)
\end{aligned}$$

where in the last step I've used the proposition I stated (without proof) above.

To sum up:

Proposition 2 *Using the* $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ *of Eq. 7.13,*

$$\frac{\hat{\beta}_1-\beta_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} \sim t_{n-2} \quad (7.28)$$

Notice that we can compute $\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ without knowing any of the true parameters — it's a pure statistic, just a function of the data. This is a key to actually using the proposition for anything useful.

By exactly parallel reasoning, we may also demonstrate that

$$\frac{\hat{\beta}_0-\beta_0}{\widehat{\mathrm{se}}\left[\hat{\beta}_0\right]} \sim t_{n-2}$$

7.3 Sampling Intervals for $\hat{\beta}$; hypothesis tests for $\hat{\beta}$

Let's trace through one of the consequences of Eq. 7.28. For any $k > 0$,

$$\begin{aligned} & \mathbb{P}\left(\beta_1 - k\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 \leq \beta_1 + k\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]\right) & (7.29) \\ = \; & \mathbb{P}\left(k\widehat{\mathrm{se}}\left[\hat{\beta}_1\right] \leq \hat{\beta}_1 - \beta_1 \leq k\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]\right) & (7.30) \\ = \; & \mathbb{P}\left(k \leq \frac{\hat{\beta}_1 - \beta_1}{\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]} \leq k\right) & (7.31) \\ = \; & \int_{-k}^{k} t_{n-2}(u)du & (7.32) \end{aligned}$$

where by a slight abuse of notation I am writing $t_{n-2}(u)$ for the probability density of the t distribution with $n-2$ degrees of freedom, evaluated at the point u.

It should be evident that if you pick any α between 0 and 1, I can find a $k(n,\alpha)$ such that

$$\int_{-k(n,\alpha)}^{k(n,\alpha)} t_{n-2}(u)du = 1-\alpha$$

I therefore define the (symmetric) $1-\alpha$ **sampling interval** for $\hat{\beta}_1$, when the true slope is β_1, as

$$[\beta_1 - k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right], \beta_1 + k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]] \qquad (7.33)$$

If the true slope is β_1, then $\hat{\beta}_1$ will be within this sampling interval with probability $1-\alpha$. This leads directly to a test of the null hypothesis that the slope $\beta_1 = \beta_1^*$: reject the null if $\hat{\beta}_1$ is outside the sampling interval for β_1^*, and retain the null when $\hat{\beta}_1$ is inside that sampling interval. This test is called the **Wald test**, after the great statistician Abraham Wald[8].

By construction, the Wald test's probability of rejection under the null hypothesis — the **size**, or **type I error rate**, or **false alarm rate** of the test — is exactly α. Of course, the other important property of a hypothesis test is its **power** — the probability of rejecting the null when it is false. From Eqn. 7.28, it should be clear that if the true $\beta_1 \neq \beta_1^*$, the probability that $\hat{\beta}_1$ is inside the sampling interval for β_1^* is $< 1-\alpha$, with the difference growing as $|\beta_1 - \beta_1^*|$ grows. An exact calculation could be done (it'd involve what's called the "non-central t distribution"), but is not especially informative. The point is that the power is always $> \alpha$, and grows with the departure from the null hypothesis.

If you were an economist, psychologist, or something of their ilk, you have a powerful drive — almost a spinal reflex not involving the higher brain regions — to

[8] As is common with eponyms in the sciences, Wald was not, in fact, the first person to use the test, but he made one of the most important early studies of its properties, and he was already famous for other reasons.

test whether $\beta_1 = 0$. Under the Wald test, you would reject that point null hypothesis when $|\hat{\beta}_1|$ exceeds a certain number of standard deviations. As an intelligent statistician in control of your own actions, you would read the section on "statistical significance" below, before doing any such thing.

All of the above applies, *mutatis mutandis*, to $\frac{\hat{\beta}_0-\beta_0}{\widehat{\mathrm{se}}\left[\hat{\beta}_0\right]}$.

7.4 Building Confidence Intervals from Sampling Intervals

Once we know how to calculate sampling intervals, we can plot the sampling interval for every possible value of β_1 (Figure 7.3). They're the region marked off by two parallel lines, one $k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$ above the main diagonal and one equally far below the main diagonal.

The sampling intervals (as in Figure 7.3) are theoretical constructs — mathematical consequences of the assumptions in the the probability model that (we hope) describes the world. After we gather data, we can actually calculate $\hat{\beta}_1$. This is a random quantity, but it will have some particular value on any data set. We can mark this realized value, and draw a horizontal line across the graph at that height (Figure 7.4).

The $\hat{\beta}_1$ we observed is within the sampling interval for some (possible or hypothetical) values of β_1, and outside the sampling interval for others. We define the **confidence set**, with **confidence level** $1-\alpha$, as

$$\left\{\beta_1 \,:\, \hat{\beta}_1 \in [\beta_1 - k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right], \beta_1 + k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]]\right\} \tag{7.34}$$

This is precisely the set of β_1 which we retain when we run the Wald test with size α. In other words: we test every possible β_1; if we'd reject that null hypothesis, that value of β_1 gets removed from the hypothesis test; if we'd retain that null, β_1 stays in the confidence set[9]. Figure 7.5 illustrate a confidence set, and shows (unsurprisingly) that in this case the confidence set is indeed a confidence *interval*. Indeed, a little manipulation of Eq. 7.34 gives us an explicit formula for the confidence set, which is an interval:

Confidence set = Test all the hypotheses!

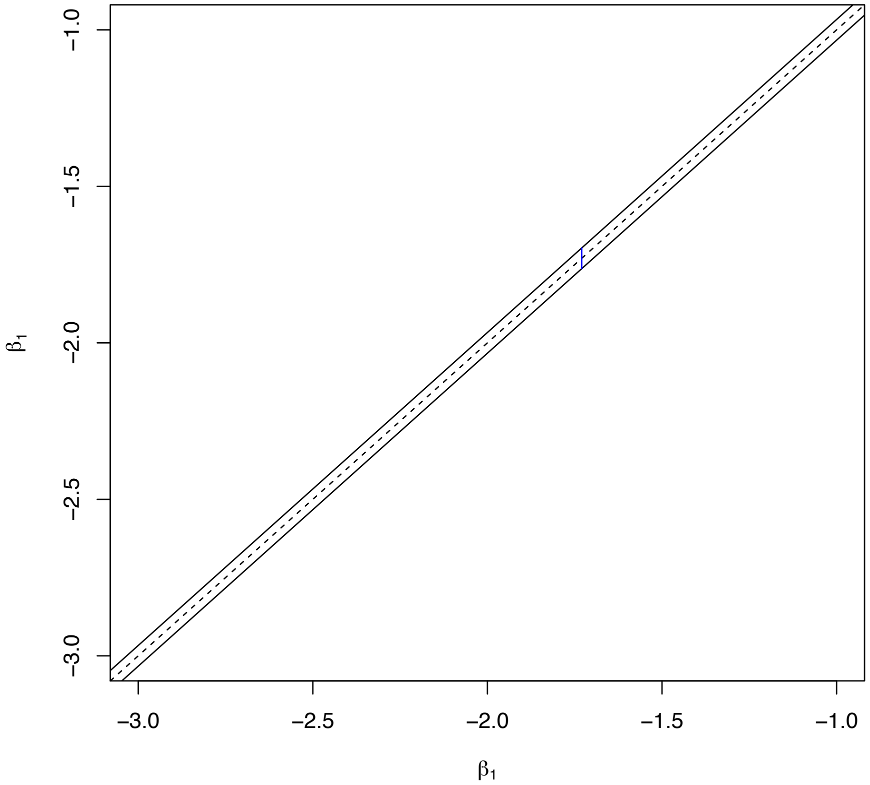
$$[\hat{\beta}_1 - k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right], \hat{\beta}_1 + k(n,\alpha)\widehat{\mathrm{se}}\left[\hat{\beta}_1\right]$$

The correct interpretation of a confidence set is that it offers us a dilemma. One of two[10] things must be true:

1. The true β_1 is inside the confidence set.

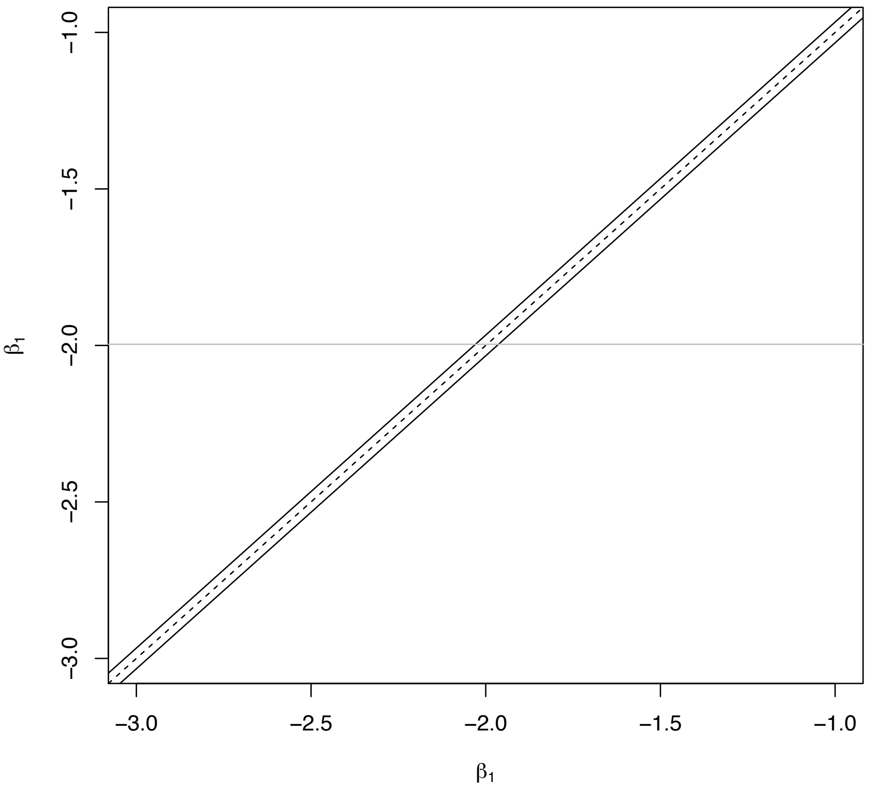
[9] Cf. the famous Sherlock Holmes line "When you have eliminated the impossible, whatever remains, however improbable, must be the truth." In forming the confidence set, we are eliminating the merely *unlikely*, rather than the absolutely impossible. This is because, not living in a detective story, we get only noisy and imperfect evidence.

[10] Strictly speaking, there is a third option: our model could be wrong. Hence the importance of model checking *before* doing within-model inference.



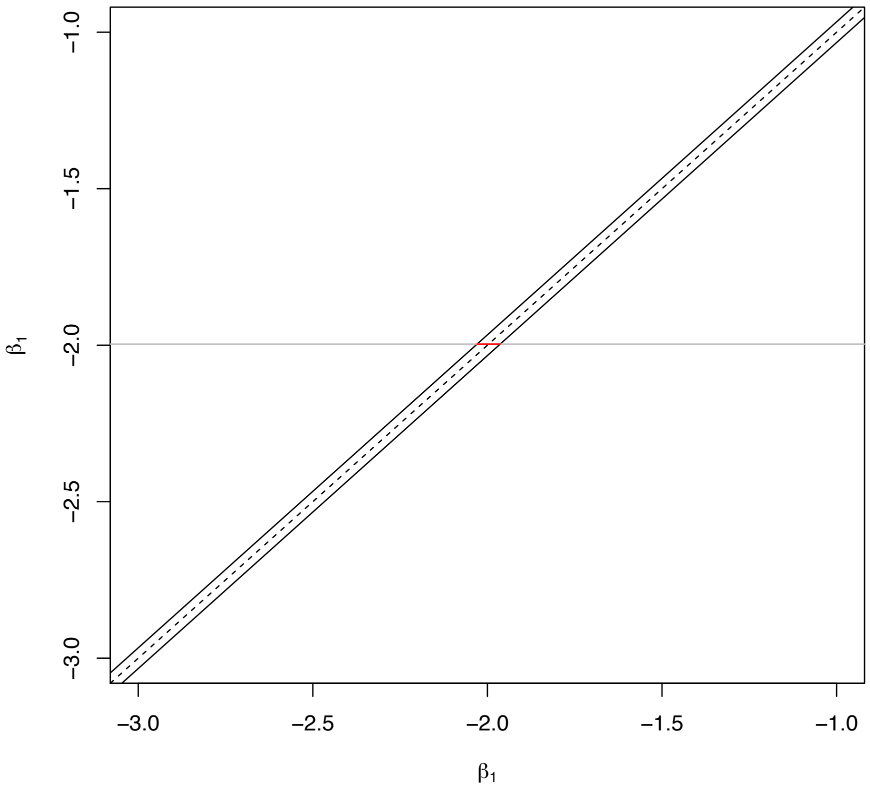
```
lm.sim <- lm(y ~ x, data = sim.gnslrm(x = x, 5, -2, 0.1, coefficients = FALSE))
hat.sigma.sq <- mean(residuals(lm.sim)^2)
se.hat.beta.1 <- sqrt(hat.sigma.sq/(var(x) * (length(x) - 2)))
alpha <- 0.02
k <- qt(1 - alpha/2, df = length(x) - 2)
plot(0, xlim = c(-3, -1), ylim = c(-3, -1), type = "n", xlab = expression(beta[1]),
    ylab = expression(hat(beta)[1]), main = "")
abline(a = k * se.hat.beta.1, b = 1)
abline(a = -k * se.hat.beta.1, b = 1)
abline(a = 0, b = 1, lty = "dashed")
beta.1.star <- -1.73
segments(x0 = beta.1.star, y0 = k * se.hat.beta.1 + beta.1.star, x1 = beta.1.star,
    y1 = -k * se.hat.beta.1 + beta.1.star, col = "blue")
```

FIGURE 7.3: *Sampling intervals for $\hat{\beta}_1$ as a function of β_1. For compatibility with the earlier simulation, I have set $n = 42$, $s_X^2 = 9$, and (from one run of the model) $\hat{\sigma}^2 = 0.067$; and, just because $\alpha = 0.05$ is cliched, $\alpha = 0.02$. Equally arbitrarily, the blue vertical line illustrates the sampling interval when $\beta_1 = -1.73$.*



```
plot(0, xlim = c(-3, -1), ylim = c(-3, -1), type = "n", xlab = expression(beta[1]),
    ylab = expression(hat(beta)[1]), main = "")
abline(a = k * se.hat.beta.1, b = 1)
abline(a = -k * se.hat.beta.1, b = 1)
abline(a = 0, b = 1, lty = "dashed")
beta.1.hat <- coefficients(lm.sim)[2]
abline(h = beta.1.hat, col = "grey")
```

FIGURE 7.4: *As in Figure 7.3, but with the addition of a horizontal line marking the observed value of $\hat{\beta}_1$ on a particular realization of the simulation (in grey).*



```
plot(0, xlim = c(-3, -1), ylim = c(-3, -1), type = "n", xlab = expression(beta[1]),
    ylab = expression(hat(beta)[1]), main = "")
abline(a = k * se.hat.beta.1, b = 1)
abline(a = -k * se.hat.beta.1, b = 1)
abline(a = 0, b = 1, lty = "dashed")
beta.1.hat <- coefficients(lm.sim)[2]
abline(h = beta.1.hat, col = "grey")
segments(x0 = beta.1.hat - k * se.hat.beta.1, y0 = beta.1.hat, x1 = beta.1.hat +
    k * se.hat.beta.1, y1 = beta.1.hat, col = "red")
```

FIGURE 7.5: *As in Figure 7.4, but with the* **confidence set** *marked in red. This is the collection of all* β_1 *where* $\hat{\beta}_1$ *falls within the* $1-\alpha$ *sampling interval.*

2. $\hat{\beta}_1$ is outside the sampling interval of the true β_1.

We know that the second option has probability at most α, no matter what the true β_1 is, so we may rephrase the dilemma. Either

1. The true β_1 is inside the confidence set, or
2. We're very unlucky, because something whose probability is $\leq \alpha$ happened.

Since, most of the time, we're not very unlucky, the confidence set is, in fact, a reliable way of giving a margin of error for the true parameter β_1.

Width of the confidence interval Notice that the width of the confidence interval is $2k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_1\right]$. This tells us what controls the width of the confidence interval:

1. As α shrinks, the interval widens. (High confidence comes at the price of big margins of error.)
2. As n grows, the interval shrinks. (Large samples mean precise estimates.)
3. As σ^2 increases, the interval widens. (The more noise there is around the regression line, the less precisely we can measure the line.)
4. As s_X^2 grows, the interval shrinks. (Widely-spread measurements give us a precise estimate of the slope.)

What about β_0? By exactly parallel reasoning, a $1-\alpha$ confidence interval for β_0 is $[\hat{\beta}_0 - k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_0\right], \hat{\beta}_0 + k(n,\alpha)\widehat{\text{se}}\left[\hat{\beta}_0\right]]$.

What about σ^2? See Exercise 1.

What α should we use? It's become conventional to set $\alpha = 0.05$. To be honest, this owes more to the fact that the resulting k tends to 1.96 as $n \to \infty$, and $1.96 \approx 2$, and most psychologists and economists could multiply by 2, even in 1950, than to any genuine principle of statistics or scientific method. A 5% error rate corresponds to messing up about one working day in every month, which you might well find high. On the other hand, there is nothing which stops you from increasing α. It's often illuminating to plot a series of confidence sets, at different values of α.

What about power? The **coverage** of a confidence set is the probability that it includes the true parameter value. This is not, however, the only virtue we want in a confidence set; if it was, we could just say "Every possible parameter is in the set", and have 100% coverage no matter what. We would also like the *wrong* values of the parameter to have a high probability of *not* being in the set. Just as the coverage is controlled by the size / false-alarm probability / type-I error rate α of the hypothesis test, the probability of excluding the wrong parameters is controlled by the power / miss probability / type-II error rate. Test with higher power exclude (correctly) more parameter values, and give smaller confidence sets.

7.4.1 Confidence Sets and Hypothesis Tests

I have derived confidence sets for β by inverting a specific hypothesis test, the Wald test. There is a more general relationship between confidence sets and hypothesis tests.

1. Inverting any hypothesis test gives us a confidence set.
2. If we have a way of constructing a $1-\alpha$ confidence set, we can use it to test the hypothesis that $\beta = \beta^*$: reject when β^* is outside the confidence set, retain the null when β^* is inside the set.

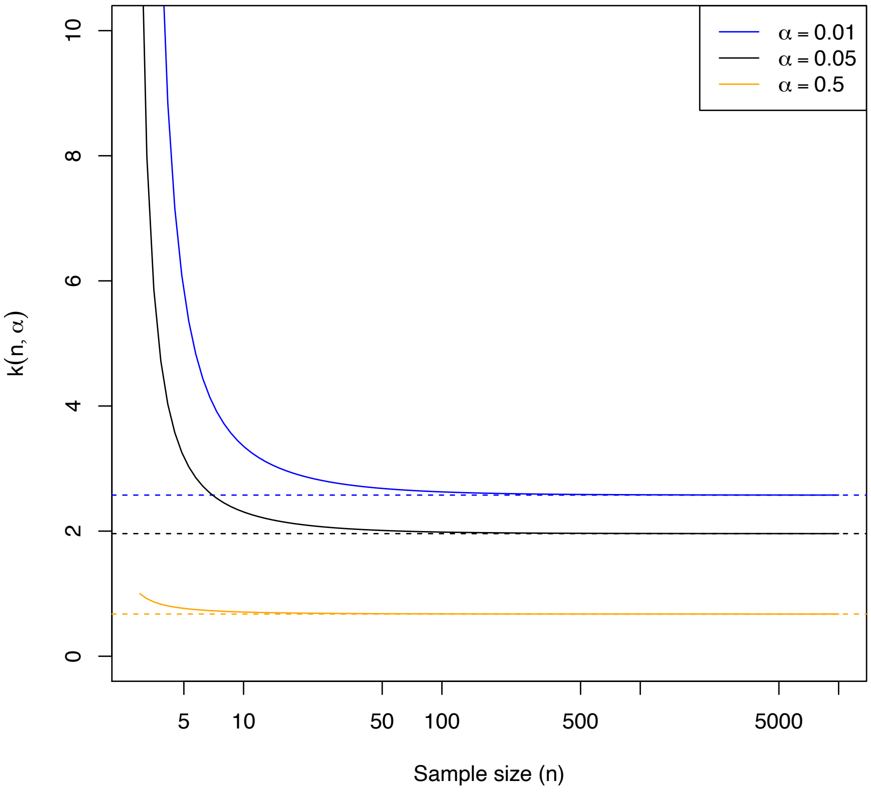
I will leave it as a pair of exercises (2 and 3) to that inverting a test of size α gives a $1-\alpha$ confidence set, and that inverting a $1-\alpha$ confidence set gives a test of size α.

7.4.2 Large-n Asymptotics

As $n \rightarrow \infty$, $\hat{\sigma}^2 \rightarrow \sigma^2$. It follows (by continuity) that $\widehat{\mathrm{se}}\left[\hat{\beta}\right] \rightarrow \mathrm{se}\left[\hat{\beta}\right]$. Hence,

$$\frac{\hat{\beta}-\beta}{\widehat{\mathrm{se}}\left[\hat{\beta}\right]} \rightarrow N(0,1)$$

which considerably simplifies the sampling intervals and confidence sets; as n grows, we can forget about the t distribution and just use the standard Gaussian distribution. Figure 7.6 plots the convergence of $k(n,\alpha)$ towards the $k(\infty,\alpha)$ we'd get from the Gaussian approximation. As you can see from the figure, by the time $n = 100$ —a quite small data set by modern standards — the difference between the t distribution and the standard-Gaussian is pretty trivial.



```
curve(qt(0.995, df = x - 2), from = 3, to = 10000, log = "x", ylim = c(0, 10),
    xlab = "Sample size (n)", ylab = expression(k(n, alpha)), col = "blue")
abline(h = qnorm(0.995), lty = "dashed", col = "blue")
curve(qt(0.975, df = x - 2), add = TRUE)
abline(h = qnorm(0.975), lty = "dashed")
curve(qt(0.75, df = x - 2), add = TRUE, col = "orange")
abline(h = qnorm(0.75), lty = "dashed", col = "orange")
legend("topright", legend = c(expression(alpha == 0.01), expression(alpha ==
    0.05), expression(alpha == 0.5)), col = c("blue", "black", "orange"), lty = "solid")
```

FIGURE 7.6: *Convergence of $k(n, \alpha)$ as $n \to \infty$, illustrated for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.5$. (Why do I plot the 97.5^{th} percentile when I'm interested in $\alpha = 0.05$?)*