Chapter 3:
Improved! Now with Probabilities

Overview:

This chapter extends the mathematics from last chapter by adding a probability distribution to the residuals. This results in the independent variable having a probability distribution.

Please keep in mind that the independent variables are not random variables. The researcher specifically selects their values. Adhering to this paradigm allows us to more easily determine the resulting distributions. As such, this chapter continues this requirement.

Should we not adhere to this requirement, the results of this chapter will technically be wrong, but will be close if the independent variable is statistically independent of the dependent variable.

Forsberg, Ole J. (November 18, 2023). "Improved! Now with Probabilities." In *Linear Models and Řurità Kràlovstvì*. Version 0.704442ζ.

Chapter Contents

❧❧❧

In the previous chapter, we explored the mathematical consequences of our choice of definition of "best." In this chapter, we will acknowledge that the residuals are observations from a random variable, specify its distribution, and see where that takes us.

And so, let us return to our scalar model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.1}$$

and see what we can learn if the ε_i are generated from a Normal distribution. Specifically, let us assume/require

$$\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}\left(0;\ \sigma^2\right) \tag{3.2}$$

The expected value of ε_i is a constant 0. No matter the values of the other variables, the expected value of the residual is 0 at that point.

The variance of the ε_i is a constant σ^2. No matter the values of the other variables, the variance of the residual is σ^2 at that point.

iid

The abbreviation "iid" on top of the distribution sign means "independent and identically distributed." It indicates that the ε_i are independent of each other, and that the distribution of each ε_i is the same, $\mathcal{N}\left(0;\ \sigma^2\right)$.

On the right side of equation (3.1), the ε_i is the only random variable. The β_0 and β_1 are population parameters we are trying to estimate. The x_i are values selected by the experimenter, so they are also not random variables. This last sentence is rather important for a lot of the calculations we make. The values of the independent variable are selected by the researcher, they are *not* realizations of a random variable.

non-stochastic

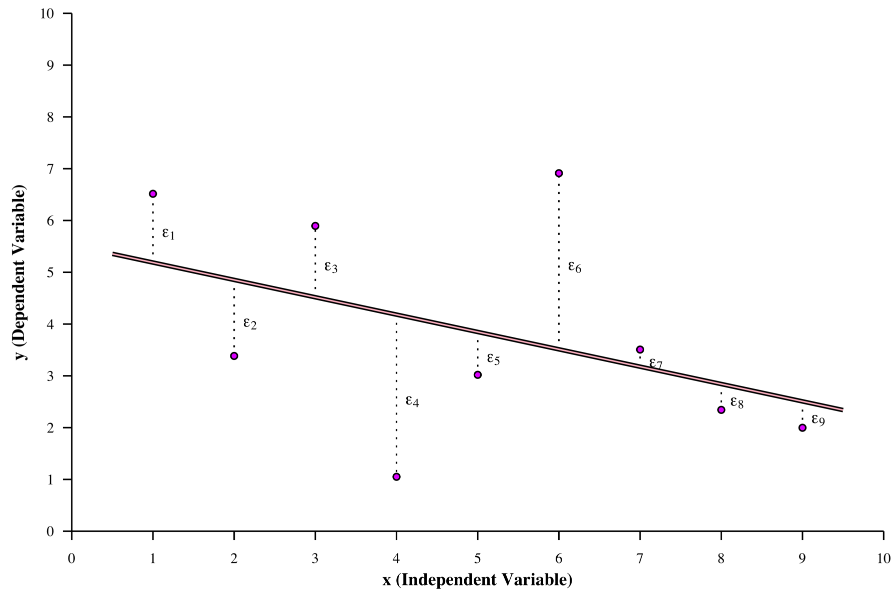

Figure 3.1: *The basic scatter plot. This provides the observed values of the data as well as the line of best fit according to the Ordinary Least Squares method. The residuals are also indicated, with the values represented by dotted segments.*

Since the only thing on the right hand side that is a random variable is the ε_i, then it is rather easy to determine the distribution of Y. And, with that, we are able to determine the distribution of almost all parameters we find important.

3.1: Probability Distributions

From this one assumption/requirement, and the math from the previous chapter, we have many consequences. This section provides the results regarding the distribution of our estimators. The next sections build on this.

Theorem 3.1. *The distribution of Y, conditional on the value of x, is*

$$Y \mid x \overset{ind}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x;\ \sigma^2\right) \tag{3.3}$$

Proof. We are given $Y = \beta_0 + \beta_1 x + \varepsilon$, with the only random variable on the right being ε. Since ε follows a Normal distribution, so too does Y (see Corollary S.39).

The expected value of $Y \mid x$ is

$$\begin{aligned} \mathbb{E}[Y \mid x] &= \mathbb{E}[\beta_0 + \beta_1 x + \varepsilon] && (3.4)\\ &= \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 x] + \mathbb{E}[\varepsilon] && (3.5)\\ &= \beta_0 + \beta_1 x + 0 && (3.6)\\ &= \beta_0 + \beta_1 x && (3.7) \end{aligned}$$

The variance of Y, conditional on x, is

$$\begin{aligned} \mathbb{V}[Y \mid x] &= \mathbb{V}[\beta_0 + \beta_1 x + \varepsilon] && (3.8)\\ &= \mathbb{V}[\beta_0] + \mathbb{V}[\beta_1 x] + \mathbb{V}[\varepsilon] && (3.9)\\ &= 0 + 0 + \mathbb{V}[\varepsilon] && (3.10)\\ &= \sigma^2 && (3.11) \end{aligned}$$

Thus, putting this together, we have

$$Y \mid x \overset{\text{ind}}{\sim} \mathcal{N}\left(\beta_0 + \beta_1 x; \sigma^2\right) \tag{3.12}$$

□

Note that the Y are only 'independently distributed' and not 'independent and identically distributed.' This is because the expected value of Y depends on the value of x. Since the Y do not all have the same (identical) distribution, they are only 'independently distributed.'

As for the results of the theorem above, they may not be too interesting. However, as our estimators depend on the Y_i, so too do their distributions. And *that* is where the interest arises.

We see this in the next theorem.

Theorem 3.2. *The distribution of b_1 is $\mathcal{N}\left(\beta_1; \sigma^2 \frac{1}{S_{xx}}\right)$.*

Proof. Before we start, we need to note that b_1 can be written as a linear combination of the Y_i:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\, Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.13}$$

exercise I leave the proof of this as an exercise.

Since our b_1 is a linear combination of Y_i, and since the Y_i come from independent Normal distributions, we have that b_1 also follows a Normal distribution (see Corollary S.39).

The expected value of b_1 is

$$\mathbb{E}[b_1] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})\ Y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right] \tag{3.14}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})\ \mathbb{E}[Y_i]}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.15}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})\ (\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.16}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})\ \beta_0}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\ \beta_1 x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.17}$$

$$= \beta_0 \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})\ x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.18}$$

$$= \beta_0 \frac{0}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.19}$$

$$= 0 + \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.20}$$

$$= \beta_1 \tag{3.21}$$

In this sequence, note that

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0 \tag{3.22}$$

and that

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})\ x_i \tag{3.23}$$

Thus, at this point, we know $\mathbb{E}[b_1] = \beta_1$; our estimator is unbiased.

The final step is to determine the variance of b_1:

$$\mathbb{V}[b_1] = \mathbb{V}\left[\frac{\sum_{i=1}^n (x_i - \overline{x})\ Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \tag{3.24}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x})^2\ \mathbb{V}[Y_i]}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} \tag{3.25}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\left(\sum_{i=1}^n (x_i - \overline{x})^2\right)^2}\ \sigma^2 \tag{3.26}$$

$$= \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}\ \sigma^2 \tag{3.27}$$

$$= \sigma^2 \frac{1}{S_{xx}} \tag{3.28}$$

Recall that since we will be coming across $\sum_{i=1}^n (x_i - \overline{x})^2$ many, many, many times in the future, we denote it by S_{xx}. So, putting all of these parts together gives us

$$b_1 \sim \mathcal{N}\left(\beta_1; \sigma^2 \frac{1}{S_{xx}}\right) \tag{3.29}$$

□

Theorem 3.3. *The covariance between our b_1 estimator and $\overline{Y}$ is 0.*

Proof. I leave this as an exercise. □

Theorem 3.4. *The distribution of b_0 is $\mathcal{N}\left(\beta_0; \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$.*

Proof. Our estimator is

$$b_0 = \overline{Y} - b_1 \overline{x} \tag{3.30}$$

Since we have previously shown $\mathbb{C}ov\left[\overline{Y}, b_1\right] = 0$ (Theorem 3.3), the proof is straight forward. First, we note that b_0 is a linear combination of the Y_i. Thus, it follows a Normal distribution. (Again, see Corollary S.39 for a proof of this.)

Second, $\mathbb{E}[b_0] = \beta_0$:

$$\begin{aligned}
\mathbb{E}[b_0] &= \mathbb{E}\left[\overline{Y} - b_1\overline{x}\right] && (3.31)\\
&= \mathbb{E}\left[\overline{Y}\right] - \mathbb{E}[b_1\overline{x}] && (3.32)\\
&= (\beta_0 + \beta_1\overline{x}) - \beta_1\overline{x} && (3.33)\\
&= \beta_0 && (3.34)
\end{aligned}$$

Next, $\mathbb{V}[b_0] = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$:

$$\begin{aligned}
\mathbb{V}[b_0] &= \mathbb{V}\left[\overline{Y} + b_1\overline{x}\right] && (3.35)\\
&= \mathbb{V}\left[\overline{Y}\right] + \mathbb{V}[b_1\overline{x}] + 2\mathbb{C}ov\left[\overline{Y}, b_1\overline{x}\right] && (3.36)\\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\overline{x}^2 + 0 && (3.37)\\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) && (3.38)
\end{aligned}$$

Thus, putting these three parts together gives us what we want:

$$b_0 \sim \mathcal{N}\left(\beta_0; \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right) \tag{3.39}$$

□

There is another parameter in our model that we may like to estimate. That is the variance of ε. The ordinary least squares estimator of σ^2 is called the mean square error. It is defined as **MSE**

$$MSE = \frac{1}{n-p}\sum_{i=1}^{n}\varepsilon_i \tag{3.40}$$

Here, p is the number of parameters in the regression. So far, we have dealt with estimating β_0 and β_1. Thus, $p = 2$ in simple linear regression.

Theorem 3.5. *The distribution of the mean square error, MSE, is $\left(\frac{\sigma^2}{n-p}\right)\chi^2_{n-p}$.*

Proof. The first thing to do is remind ourselves of the definition of a χ^2 random variable. From Definition S.20, we have that if $Z_i \sim \mathcal{N}(0;1)$, then

$\sum Z_i^2 \sim \chi^2_\nu$, where ν is the number of those Z_i that are independent (the degrees of freedom).

With this definition, we just need to find a random variable with a Normal distribution and transform it into the proper form. To that end:

$$\varepsilon_i \sim \mathcal{N}\left(0; \sigma^2\right) \tag{3.41}$$

$$\frac{\varepsilon_i}{\sigma} \sim \mathcal{N}(0; 1) \tag{3.42}$$

$$\frac{\varepsilon_i^2}{\sigma^2} \sim \chi^2_{\nu=1} \tag{3.43}$$

$$\frac{\sum \varepsilon_i^2}{\sigma^2} \sim \chi^2_{\nu=n-p} \tag{3.44}$$

$$\frac{(n-p)\frac{1}{n-p}\sum \varepsilon_i^2}{\sigma^2} \sim \chi^2_{n-p} \tag{3.45}$$

$$\frac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p} \tag{3.46}$$

and finally,

$$MSE \sim \frac{\sigma^2}{n-p}\chi^2_{n-p} \tag{3.47}$$

As usual, knowing the distribution of a sample statistic like the MSE allows us to create confidence intervals and perform hypothesis testing. □

With that said, however, the importance of the previous theorem lies more in how we can use it to obtain confidence intervals and test hypotheses about the OLS estimators of the intercept and slope parameters.

Theorem 3.6. *The distribution of Y for an observed value of x_i, which we will term $\hat{Y}_i$, is*

$$\hat{Y}_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i;\ \sigma^2\left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\right)$$

Note: What this means is that if we were to collect an infinite number of dependent variable values (Y_i) for the specified independent variable values (all given x_i), and calculate $\hat{Y}_i$ for each of those experiments using our formulas, then those many $\hat{Y}_i$ would follow the given distribution.

Proof. Remember that $\hat{Y}_i = b_0 + b_1 x_i$ and that x is non-stochastic (it is not a random variable). With this, we have that $\hat{Y}_i$ is a linear combination of Normally distributed random variables (b_0 and b_1). As such, the name of the distribution of $\hat{Y}_i$ is "Normal." What remains is to calculate the expected value and variance.

stochastic

$$\begin{aligned}
\mathbb{E}\left[\hat{Y}_i\right] &= \mathbb{E}\left[b_0 + b_1 x_i\right] && (3.48)\\
&= \mathbb{E}\left[b_0\right] + \mathbb{E}\left[b_1 x_i\right] && (3.49)\\
&= \beta_0 + \beta_1 x_i && (3.50)
\end{aligned}$$

As expected, the estimator is unbiased.

What about the variance? That is a bit more difficult, because we must deal with the covariance between b_0 and b_1.

$$\begin{aligned}
\mathbb{V}\left[\hat{Y}_i\right] &= \mathbb{V}\left[b_0 + b_1 x_i\right] && (3.51)\\
&= \mathbb{V}\left[b_0\right] + \mathbb{V}\left[b_1 x_i\right] + 2\ \mathbb{C}ov\left[b_0, b_1 x_i\right] && (3.52)\\
&= \mathbb{V}\left[b_0\right] + \mathbb{V}\left[b_1\right] x_i^2 + 2\ \mathbb{C}ov\left[b_0, b_1\right] x_i && (3.53)\\
&= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) + \sigma^2\left(\frac{1}{S_{xx}}\right) x_i^2 + 2\frac{-\overline{x}\sigma^2}{S_{xx}} x_i && (3.54)
\end{aligned}$$

Factoring things out to make it look more simple gives

$$\begin{aligned}
\mathbb{V}\left[\hat{Y}_i\right] &= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left(\overline{x}^2 + x_i^2 - 2\overline{x}x_i\right) && (3.55)\\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left(\overline{x} - x_i\right)^2 && (3.56)\\
&= \sigma^2\left(\frac{1}{n} + \frac{\left(\overline{x} - x_i\right)^2}{S_{xx}}\right) && (3.57)
\end{aligned}$$

And so, putting these three things together gives us our hoped-for result... as we expected. □

Note: There are a couple of things interesting about this result. First, the uncertainty in $\hat{Y}_i$ is a function of n, S_{xx}, and $\overline{x} - x_i$. Larger sample sizes (larger n) produce a better (more precise) estimate.

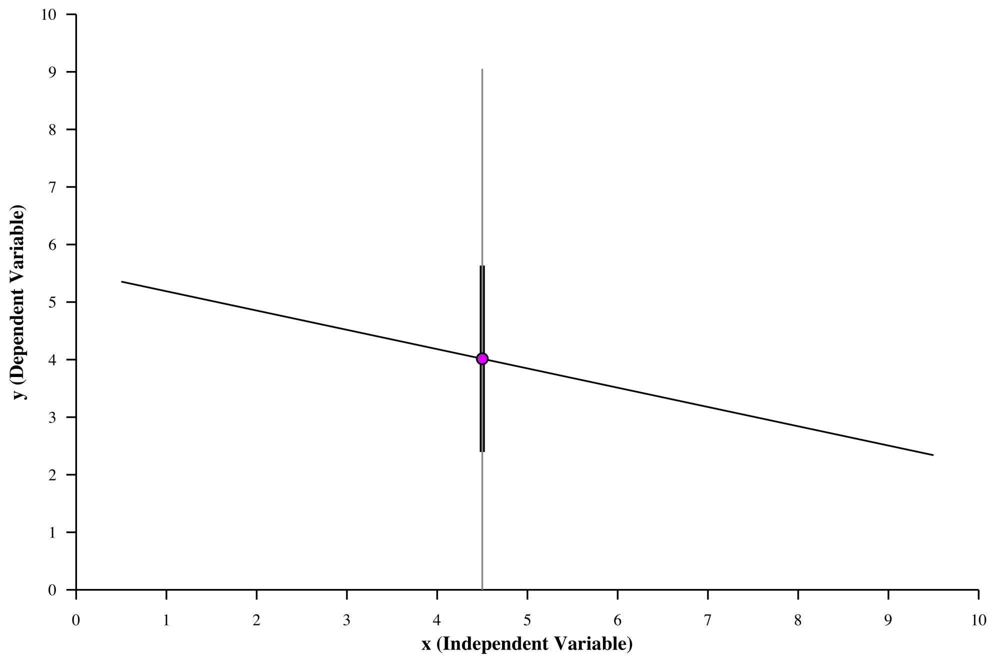

Figure 3.2: *The basic scatter plot with the confidence and prediction intervals for* $x = 4.5$ *provided. Note that the prediction interval (thin line) is much wider than the confidence interval (thick line). This is because the prediction interval uncertainty includes both the uncertainty in the mean value (confidence interval) and the inherent variation in the residuals* (σ^2).

Samples with larger values of S_{xx} also produce better estimates. To maximize S_{xx}, the researcher must have half of the x_i values at the minimum and half at the maximum.[1]

Finally, the precision of the estimate also depends on how far that x value is from the center of gravity, $(\overline{x},\ \overline{y})$. Note that the uncertainty in $\hat{Y}_i$ when $x = \overline{x}$ *only* comes from the uncertainty in the value of $\overline{Y}$. Convince yourself that this makes sense (non-mathematically).

[1] Unfortunately, the drawback to doing this is that one is not able to detect a curvature in the expected values of Y. Thus, we again see that there is a trade off in statistics. The important part is to understand what you are trying to understand… and use your statistical understanding to understand it.

Theorem 3.7. *The distribution of Y_{new}, a new observation, for a new value of x, is*

$$Y_{new} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_{new};\ \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{new} - \overline{x})^2}{S_{xx}}\right)\right)$$

Note: Before we begin this proof, be aware that

$$Y_{new} = b_0 + b_1 x_{new} + \varepsilon = \hat{Y}_{new} + \varepsilon \tag{3.58}$$

Since we are estimating a new *observation* (as opposed to just an expected value), we need to include ε in our calculations. This is subtle and very important. It emphasizes the importance of ε. **subtle**

Proof. And now for the expected proof. See that Y_{new} is a linear combination of Normally distributed random variables (b_0, b_1, and ε). Thus, Y_{new} follows a Normal distribution. All that remains is to calculate its expected value and its variance. To do so, we rely on the previous theorem.

$$\mathbb{E}[Y_{new}] = \mathbb{E}[b_0 + b_1 x_{new} + \varepsilon] \tag{3.59}$$

$$= \mathbb{E}[b_0 + b_1 x_{new}] + \mathbb{E}[\varepsilon] \tag{3.60}$$

$$= \beta_0 + \beta_1 x_{new} \tag{3.61}$$

Next, for the variance:

$$\mathbb{V}[Y_{new}] = \mathbb{V}[b_0 + b_1 x_{new} + \varepsilon] \tag{3.62}$$

$$= \mathbb{V}\left[\hat{Y} + \varepsilon\right] \tag{3.63}$$

$$= \mathbb{V}\left[\hat{Y}\right] + \mathbb{V}[\varepsilon] + 2\ \mathbb{C}ov\left[\hat{Y}, \varepsilon\right] \tag{3.64}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(\overline{x} - x_{new})^2}{S_{xx}}\right) + \sigma^2 + 0 \tag{3.65}$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(\overline{x} - x_{new})^2}{S_{xx}}\right) \tag{3.66}$$

□

Note that the only difference in the uncertainties between Y_{new} and $\hat{Y}$ is an additional term of σ^2 due to the inclusion of the residual term. Thus, all of the things that affect the variance of $\hat{Y}$ also affect the variance of Y_{new}, and in the same way.

Also note that the uncertainty in an observation is higher than the uncertainty in the expected value (see Figure 3.2).

observation

The important difference between this theorem and the previous is that this theorem models a new observation, while the previous models the expected value of an observation. The difference is important.

3.2: Test Statistics and Hypothesis Testing

The previous section provided the distribution of several important estimators. With those distributions, and our knowledge of probability distributions, we can test individual hypotheses. For this section, we rely heavily on the definition of Student's t distribution given as Definition S.21. If we let $Z \sim \mathcal{N}(0; 1)$ and $V \sim \chi^2_\nu$, with Z and V independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \tag{3.67}$$

follows a Student's t distribution with ν degrees of freedom.

You have, most likely, come across this ratio in your elementary statistics course when you were investigating hypotheses about a single population mean, given that the data came from a Normal distribution.

Theorem 3.8. *The ratio* $T = \frac{b_1-\beta_1}{\sqrt{MSE/S_{xx}}}$ *follows a Student's t distribution with* $n-p$ *degrees of freedom.*

Proof. To prove this statement, one must show that it can be written in the form of Equation 3.67. First, let us look at the numerator.

$$b_1 \sim \mathcal{N}\left(\beta_1; \sigma^2/S_{xx}\right) \tag{3.68}$$

$$\implies \qquad \frac{b_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0; 1) \tag{3.69}$$

$H_0: \beta_1 = \beta_{10}$	$H_A: \beta_1 \neq \beta_{10}$	p-value = $\mathbb{P}[t \leq -\mid T \mid] \times 2$
$H_0: \beta_1 \leq \beta_{10}$	$H_A: \beta_1 > \beta_{10}$	p-value = $\mathbb{P}[t \geq T]$
$H_0: \beta_1 \geq \beta_{10}$	$H_A: \beta_1 < \beta_{10}$	p-value = $\mathbb{P}[t \leq T]$

Table 3.1: *Table of how to calculate p-values given the null and alternative hypotheses.*

Now for the denominator we use a previous theorem (Theorem 3.5):

$$\frac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p} \tag{3.70}$$

Next, we put these together

$$T = \frac{b_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \tag{3.71}$$

$$= \frac{\frac{b_1-\beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{MSE}{\sigma^2}}} \tag{3.72}$$

$$= \frac{\frac{b_1-\beta_1}{\sqrt{\sigma^2/S_{xx}}}}{\sqrt{\frac{(n-p)MSE}{\sigma^2}/(n-p)}} \tag{3.73}$$

Note that the numerator of this follows a standard Normal distribution. The denominator is the square-root of a chi-square distribution divided by its degrees of freedom. Thus, by Definition S.21, T follows a Student's t distribution with $n-p$ degrees of freedom. □

This result is important for two reasons. First, it allows us to test hypotheses regarding the β_1 parameter. Second, this result allows us to calculate confidence intervals for β_1 (see Section 3.3). This parameter is usually of most interest to researchers as it provides "the effect of the independent variable on the dependent variable."

Since we know the distribution of this ratio, we can calculate p-values for any hypothesis about β_1 using the same rules as from your elementary statistics course (see Table 3.1).

Technically, we do need to show that b_1 and MSE are independent. If they are not, then Theorem 3.8 is not valid. For the proof, you will want to investigate Cochrane's Theorem and its uses (Bapat 2000, Cochrane 1934).

Theorem 3.9. *The ratio* $T = \frac{b_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$ *follows a Student's t distribution with* $n - p$ *degrees of freedom.*

Proof. I leave this as an exercise. □

This theorem allows us to easily prove the next.

Theorem 3.10. *The ratio* $T = \frac{\hat{y} - \hat{y}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right)}}$ *follows a Student's t distribution with* $n - p$ *degrees of freedom.*

Proof. I leave this as an exercise. □

That sure is a lot of exercise. How are those abs doing?

3.3: Confidence Intervals

dual

In the previous section, we examined hypothesis testing. This required that we created a test statistic and determined its distribution. One can think of confidence intervals as the dual of test statistics. Test statistics are functions of an unknown population parameter and have a distribution. Confidence intervals are for that unknown population parameter, where a probability is known (assumed). Once a person has the test statistic and its definition, the confidence interval can be determined by inverting the test statistic function (solve for the parameter).

From your elementary statistic course, you knew that the distribution of $T = \frac{\overline{x} - \mu}{s/\sqrt{n}}$ followed a Student's t distribution with $n - 1$ degrees of freedom. Solving the formula for the parameter of interest, μ, gives

$$\mu = \overline{x} - T\,\frac{s}{\sqrt{n}} \tag{3.74}$$

The interpretation of T here is that it contains the values (quantiles) that correspond to the confidence level claimed (Figure 3.3). For instance, if you desire a 95% confidence interval for a sample of size 10, the central T values are ±2.262 because the probability $\mathbb{P}[-2.262 < t < 2.262] = 0.95$.

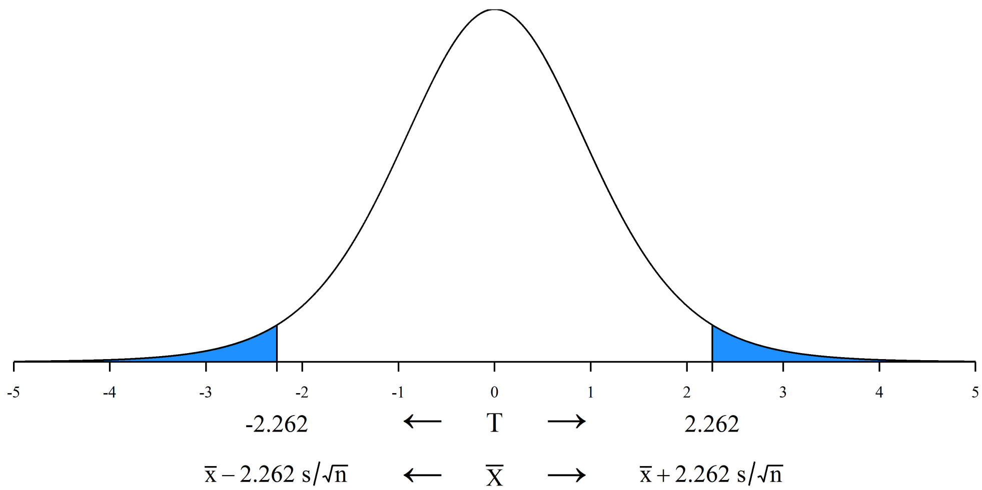

Figure 3.3: *An illustration of a confidence interval seen from the standpoint of T or from $\overline{X}$. The unshaded area constitutes 95% of the area under the curve. Thus, the vertical segments delimit the endpoints of a central 95% confidence interval.*

Thus, the interpretation of μ in Equation 3.74 is that it contains the values that correspond to the endpoints of the confidence level claimed for the distribution of the right-hand side of the formula.

This interpretation holds for *all* confidence intervals.

With this discussion, it is rather straight-forward to calculate the endpoints of confidence intervals for all of the population parameters we have explored thus far. When the distribution of the test statistic is unimodal and symmetric, the central confidence interval is *also* the narrowest. This may be important if the researcher desires the most precise estimate of the population parameter.

Theorem 3.11. *The endpoints of a central* $(1-\alpha)100\%$ *confidence for* β_1 *are defined by* $b_1 \pm t_{\alpha/2,n-p}\sqrt{MSE/S_{xx}}$.

Proof. From a previous theorem, we know $T = \frac{b_1-\beta_1}{\sqrt{MSE/S_{xx}}}$. Solving this for β_1 gives

$$\beta_1 = b_1 - T\sqrt{MSE/S_{xx}} \tag{3.75}$$

Because the distribution of T is symmetric unimodal, the endpoints of the minimum-width interval for T correspond to the quantiles $t_{\alpha/2,n-p}$ and $t_{1-\alpha/2,n-p}$. These two endpoints are equivalent to $\pm t_{\alpha/2,n-p}$.

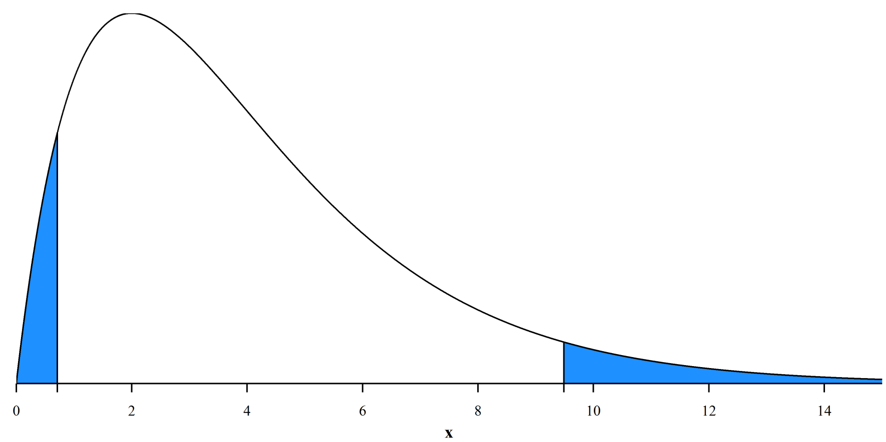

Figure 3.4: *A plot of the chi-square distribution with 4 degrees of freedom. The unshaded area constitutes 90% of the area under the curve. Thus, the vertical segments delimit the endpoints of a central 90% confidence interval.*

As such, the endpoints of a minimum-length $(1-\alpha)100\%$ confidence for β_1 are defined by $b_1 \pm t_{\alpha/2,n-p}\sqrt{MSE/S_{xx}}$. □

This is a typical result when dealing with the Student's t distribution.

efficiency

There is absolutely no reason we *need* a minimum-width confidence interval. It is, however, useful in maximizing the precision of the estimate.

When the distribution of the test statistic is unimodal symmetric, the central interval and the minimum-width interval are identical. When the distribution is *not* symmetric, they are not. The following illustrates this.

Theorem 3.12. *The endpoints of a central* $(1-\alpha)100\%$ *confidence for* σ^2 *are defined by* $\frac{(n-p)MSE}{\chi^2_{1-\alpha/2,n-p}}$ *and* $\frac{(n-p)MSE}{\chi^2_{\alpha/2,n-p}}$.

Proof. From Theorem 3.5, we know $\frac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p}$. Solving this for σ^2 gives $\sigma^2 = \frac{(n-p)MSE}{\chi^2_{n-p}}$.

Thus, a central $(1-\alpha)100\%$ confidence interval (see Figure 3.4) is defined by the endpoints

$$\frac{(n-p)MSE}{\chi^2_{n-p,1-\alpha/2}} \qquad \text{and} \qquad \frac{(n-p)MSE}{\chi^2_{n-p,\alpha/2}} \tag{3.76}$$

□

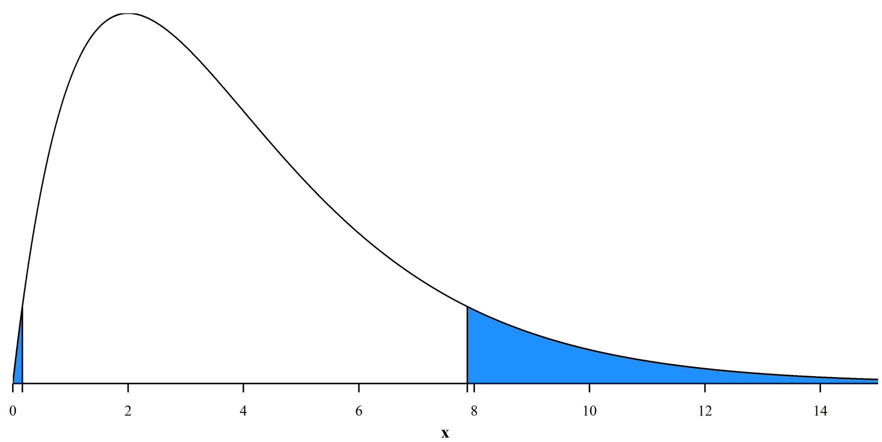

Figure 3.5: *A plot of the chi-square distribution with 4 degrees of freedom. The shaded area constitutes 10% of the area under the curve. Thus, the vertical segments delimits the endpoints of a 90% confidence interval.* This *confidence interval, however, is the minimum-width interval.*

> ***Note***: This is *not* the minimum-width interval. It is, however, the *usual* confidence interval provided. Calculating the minimum-width interval takes a little calculus that is beyond the scope of this section… and the typical coverage of this topic.

The minimum-width interval is illustrated in Figure 3.5. Note that the area in the shaded area to the right is not the same as that to the left. However, the two areas still account for 10% of the area, leaving 90% unshaded in the middle.

The width of the central 90% confidence interval shown in Figure 3.4 is 8.777. This is wider than the width of the minimum-width confidence interval shown in Figure 3.5, which is 7.714. The minimum-width interval is 12% narrower than the central interval. That is an increase in estimator efficiency. It also requires some additional mathematics that we will skip.

However, as a teaser, notice that the value of the density function for each of the two endpoints is the same in the minimum-width interval. If the distribution is unimodal, then that observation will be true. That's enough of a hint. Feel free to explore this on your own. Calculus will serve you well here. **explore**

Theorem 3.13. *The endpoints of a central (and minimum width) confidence interval for β_0 are defined by*

$$b_0 \pm t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)} \tag{3.77}$$

Proof. I leave this as an exercise. □

Theorem 3.14. *The endpoints of a confidence interval for $\hat{y}$ are defined by*

$$b_0 + b_1 x \pm t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right)} \tag{3.78}$$

Proof. I leave this as an exercise. □

Theorem 3.15. *The endpoints of a prediction interval for y are defined by*

$$b_0 + b_1 x \pm t_{\alpha/2,n-p}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x-\overline{x})^2}{S_{xx}}\right)} \tag{3.79}$$

Proof. I leave this as an exercise. □

Note: This interval is termed a "prediction interval" because it is used to predict a *new observation* of y. It is not used to estimate the expected value of y — or trends in y. That would be the purpose of a confidence interval.

3.4: The Working-Hotelling Bands

One last confidence interval we may be interested in is a confidence interval for the entire regression line.

Note that all of the confidence intervals in this chapter (except for σ^2) have been of the form

$$\text{point estimate} \pm K \times se \tag{3.80}$$

That is because they were confidence intervals for a measure of center. The Working-Hotelling (1929) confidence band for the regression line follows this format. It is

$$(b_0 + b_1 x) \pm F(1-\alpha, 2, n-2) \times \sqrt{MSE\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)} \tag{3.81}$$

The proof is beyond the scope of this course.

3.5: Conclusion

This chapter started with the mathematics of the previous chapter, the mathematics based on our definition of "best." From that decision, we added a single assumption: $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0; \sigma^2)$.

That assumption/requirement about the residuals gave us the entire chapter. Probability distributions for each of the estimators arose from the mathematics and the assumption. From those probability distributions, we created test statistics.

Having test statistics allows us to calculate p-values and confidence intervals for the parameters of interest. That is the flow of statistics. Once we have a distribution for a test statistic, we know everything we want to know for inferential statistics.

The difficulty comes in finding a test statistic with a known distribution. The assumption of Normality (and of iid) were key in allowing us to find those test statistics.

3.6: End-of-Chapter Materials

Here are the expected materials to supplement the chapter.

3.6.1 Exercises

1. Prove that $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})\, x_i$.

2. Prove that $b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\, y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$.

3. Prove that the covariance between our b_1 estimator and $\overline{y}$ is 0.

4. Prove that the OLS estimators b_0 and MSE are independent.

5. Prove that the OLS estimators b_1 and MSE are independent.

6. Prove that the ratio $T = \frac{b_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}}$ follows a Student's t distribution with $n - p$ degrees of freedom.

7. Prove that the ratio $T = \frac{\hat{y} - \hat{y}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right)}}$ follows a Student's t distribution with $n - p$ degrees of freedom.

8. Prove that the endpoints of a central confidence interval for β_0 are defined by $b_0 \pm t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$.

9. Prove that the endpoints of a confidence interval for $\hat{y}$ are defined by $b_0 + b_1 x \pm t_{\alpha/2,n-p}\sqrt{MSE\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right)}$.

10. Prove that the endpoints of a prediction interval for y are defined by $b_0 + b_1 x \pm t_{\alpha/2,n-p}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}\right)}$.

3.6.2 Theory Readings

- R. B. Bapat (2000). *Linear Algebra and Linear Models* (Second ed.). New York: Springer.
- William G. Cochrane (1934). "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance." *Mathematical Proceedings of the Cambridge Philosophical Society*. **30**(2): 178–191.
- Franklin A. Graybill and David C. Bowden (1968). "Linear Segment Confidence Bands for Simple Linear Models." *Journal of the American Statistical Association*, 62(318): 403–408.
- Henry Scheffé (1959). *The Analysis of Variance*. New York: Wiley.
- Holbrook Working and Harold Hotelling (1929). "Applications of the Theory of Error to the Interpretation of Trends." *Journal of the American Statistical Association*, 24(165A): 73–85.