

# Inference for Simple Regression Models

Andrew Zieffler

---

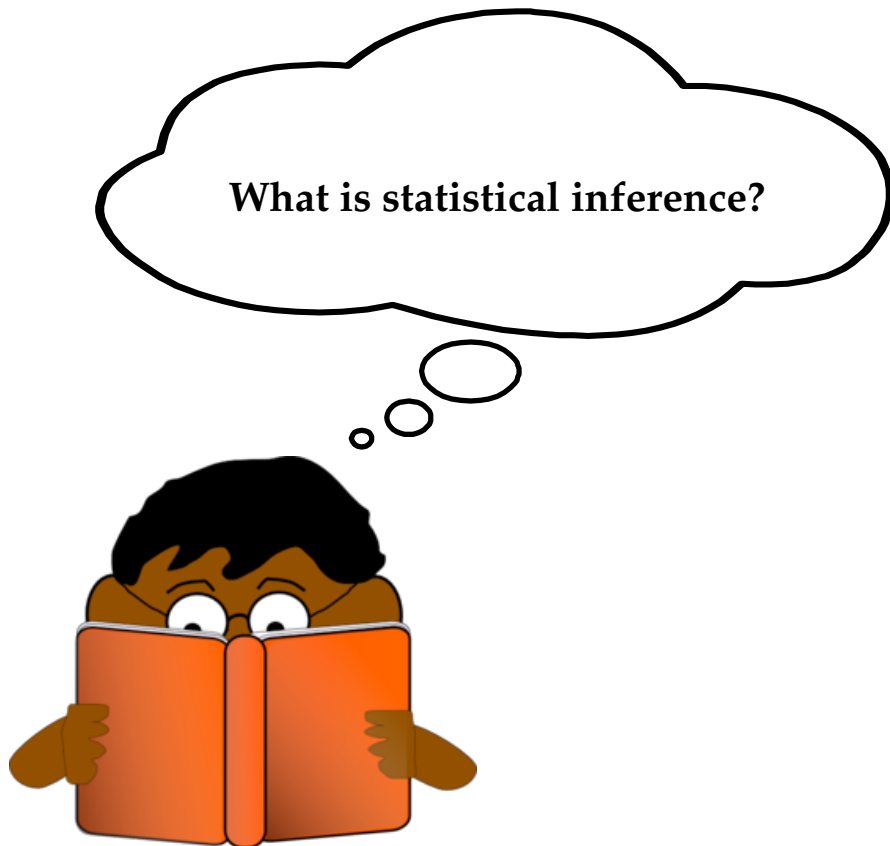


This work is licensed under a  
[Creative Commons Attribution  
4.0 International License](https://creativecommons.org/licenses/by/4.0/).

# Prepare

```
# Load the data (homework-achievement.csv)
> city = read.csv("~/epsy-8251/riverside_final.csv")

# Load libraries; Note: you may need to install them first
> library(ggplot2)
```



"Statistical inference is used to learn from incomplete or imperfect data." – Gelman & Hill (2006, p. 16)

- **Sampling model:** The primary interest is to learn about one or more characteristics about a population. These characteristics must be estimated from sample data.
- **Measurement error model:** The primary interest is in learning about some underlying pattern or law (maybe to test a theory), but the data are measured with error.

Despite these being very different paradigms, in practice they are often combined (e.g., we measure imperfectly *and* we want to make generalizations)

There are many goals of statistical inference, however for conventional regression analysis the goals are: (1) to **estimate the parameters** of the proposed model; and (2) to summarize the amount of **uncertainty** in those estimates.

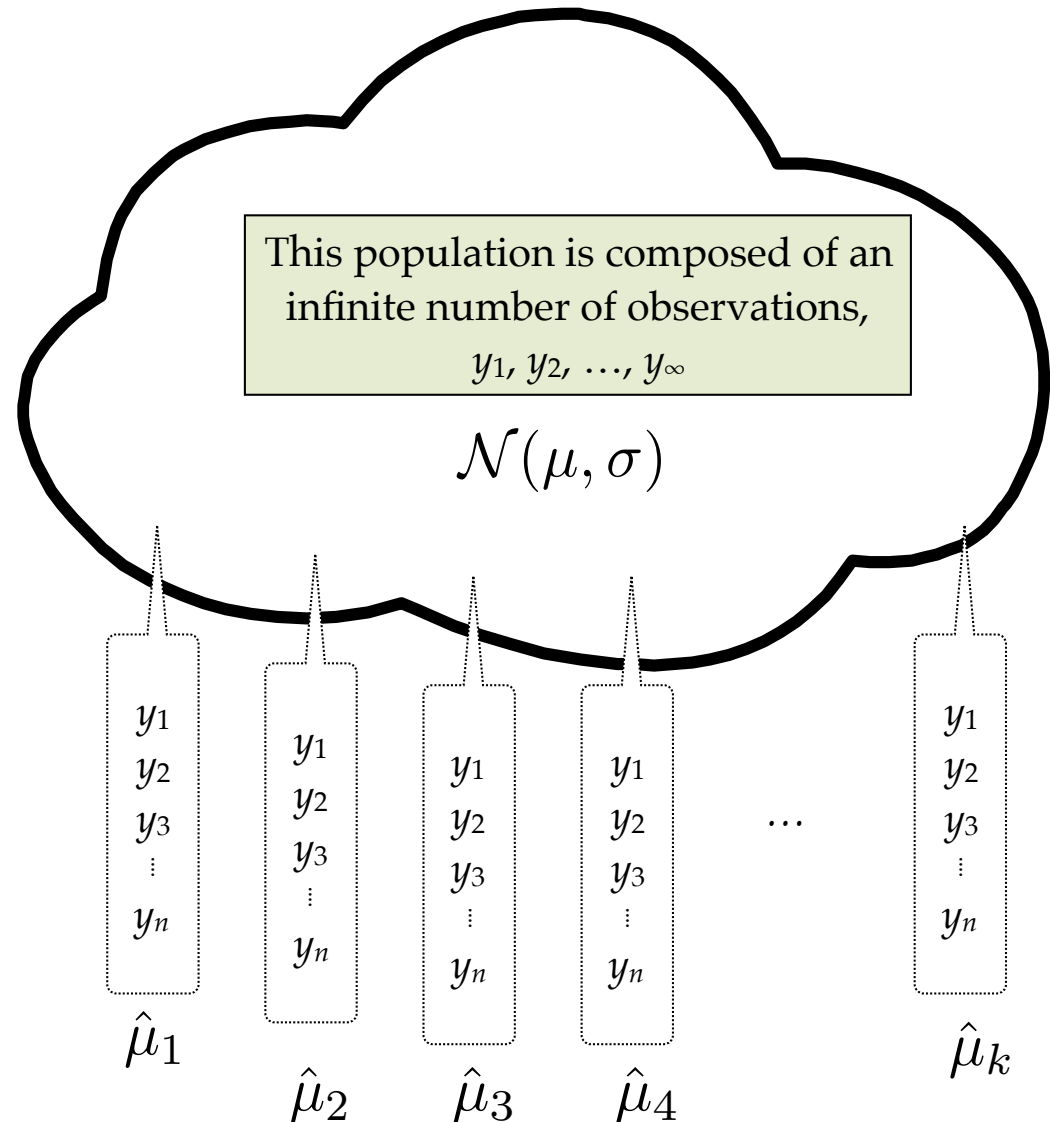
# QUANTIFICATION OF UNCERTAINTY

# Example from Stat I

**Problem:** Different random samples produce different estimates for the parameter!

**Resolution:** Quantify how much the sample estimates vary across all the different samples you could possibly draw.

This variation in estimates that is due to random sampling is referred to as **sampling error**.



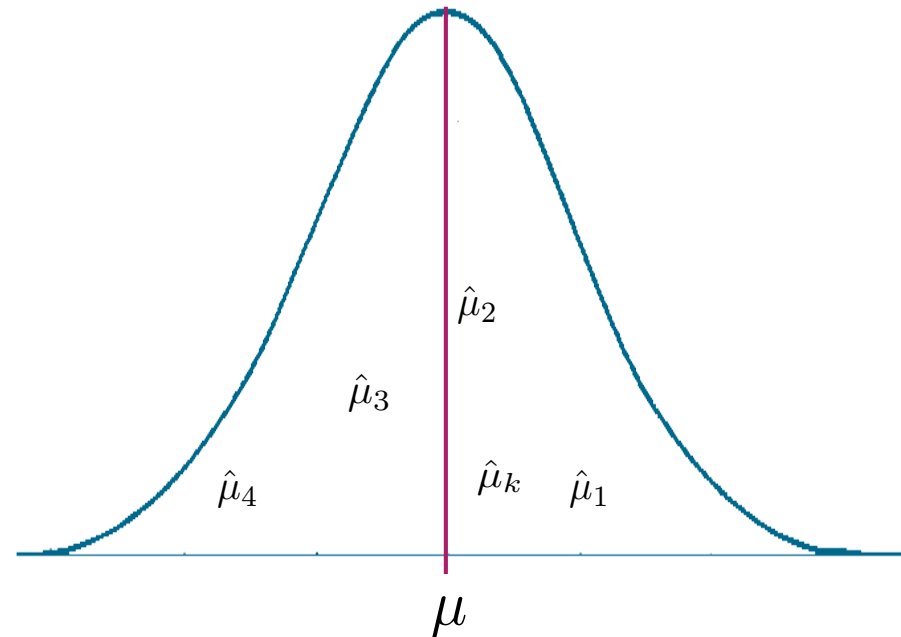
Goal is to quantify the amount of sampling error.

Statistical theory tells us that:

$$E(\hat{\mu}) = \mu$$

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

Distribution of the sample estimates

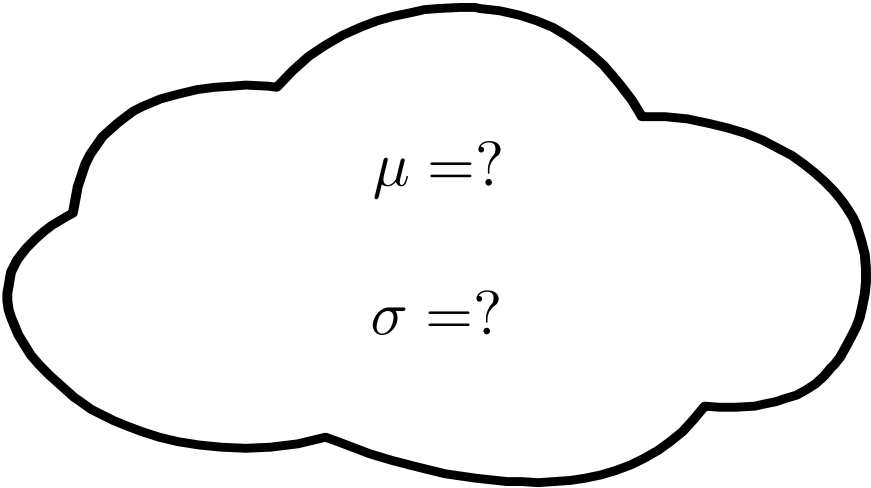


Key is that the amount of variation due to random sampling can be quantified.

When the observations in a distribution are summaries (i.e., statistics) the standard deviation is referred to as a **standard error**.

## In Practice...

**Complication:** The parameters in the population, are unknown to us.


$$\mu = ?$$

$$\sigma = ?$$

Using the theoretical formula, the numerator is thus unknown.

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

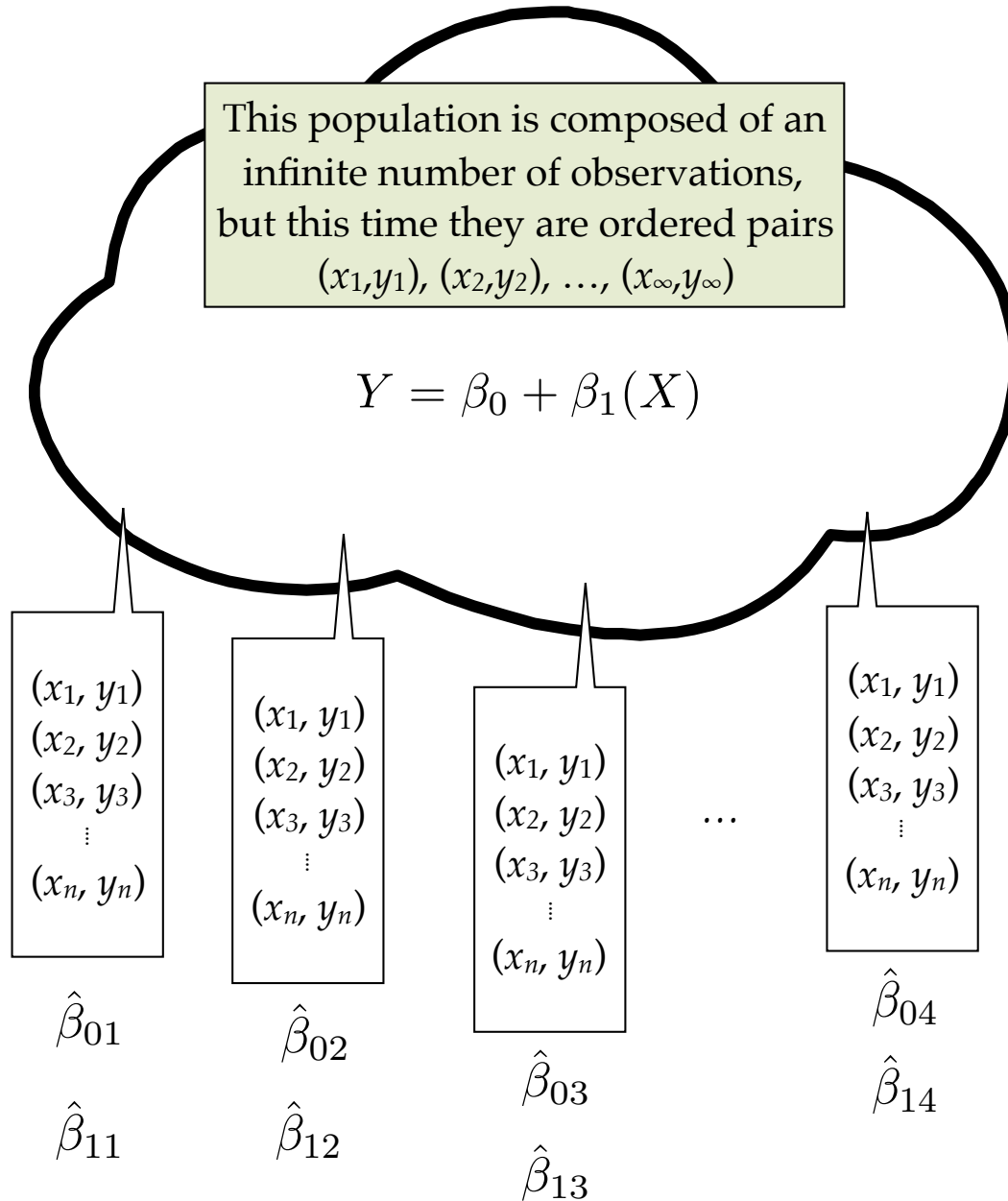
To deal with this, we use our sample estimate for the standard deviation in the numerator. But... sample estimates have error. This introduces additional error in the numerator. Because of this, the distribution of sample estimates is no longer normally distributed, it is *t*-distributed. ... and so on.

In practice, things are complicated, but we can still estimate the amount of sampling error.

# **QUANTIFICATION OF UNCERTAINTY IN REGRESSION**



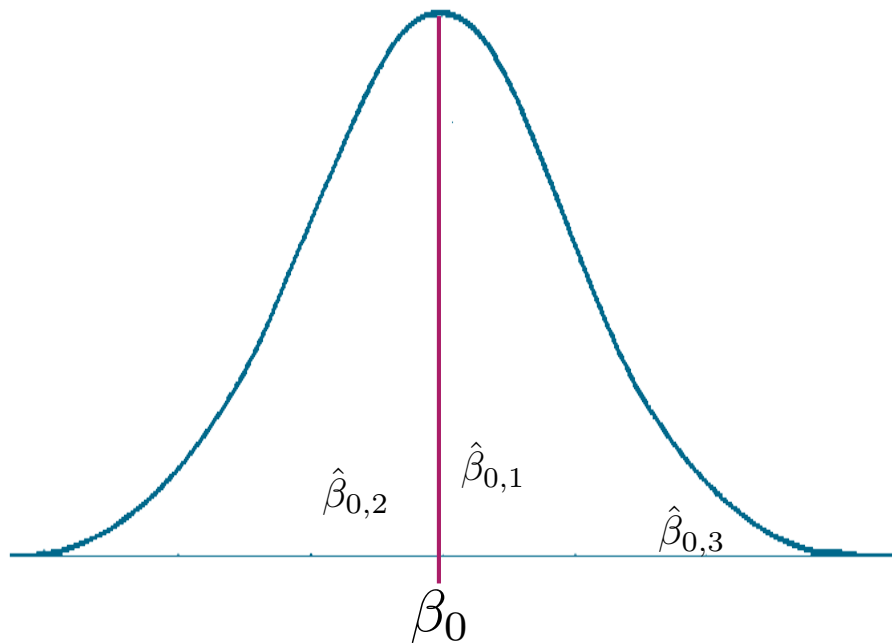
# Same Idea, Now with Regression



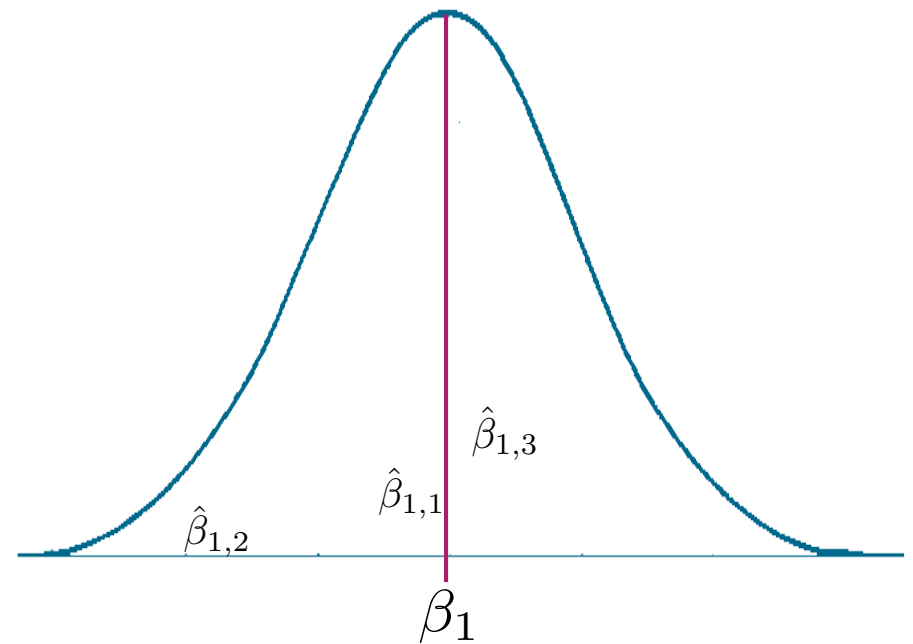
**Same Problem:** Different random samples produce different estimates for the parameter!

**Same Resolution:** Quantify how much the sample estimates vary across all the different samples you could possibly draw.

### Distribution of the **sample intercepts**



### Distribution of the **sample slopes**



Statistical theory tells us that:

$$E(\hat{\beta}_0) = \beta_0$$

$$\sigma_{\hat{\beta}_0} = \sigma_{\epsilon} \sqrt{\frac{1}{n} + \frac{\mu_X^2}{\sum (X_i - \mu_X)^2}}$$

Statistical theory tells us that:

$$E(\hat{\beta}_1) = \beta_1$$

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_{\epsilon}}{\sigma_x \sqrt{n-1}}$$

# Fit the Regression Model and Examine the Output

Use the `summary()` function to display the fitted regression coefficients and their standard errors.

```
# Fit the regression model
> lm.1 = lm(income ~ 1 ~ edu, data = city)

> summary(lm.1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11321.4      6123.2    1.849   0.0743 .
edu           2651.3       369.6    7.173 5.56e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317, Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF, p-value: 5.562e-08
```

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11321.4	6123.2	1.849	0.0743	.
edu	2651.3	369.6	7.173	5.56e-08	***

$$\hat{\beta}_0 = 11321$$

$$\hat{\sigma}_{\hat{\beta}_0} = 6123$$

**Intercept:** Based on the observed data, our estimate for the population intercept is 11,321. We recognize this is based on incomplete information (a sample), so we also attempt to quantify how uncertain about this estimate that we are. Again, based on the observed data, the estimate of the standard error for the intercept is 6,123.

$$\hat{\beta}_1 = 2651$$

$$\hat{\sigma}_{\hat{\beta}_1} = 370$$

**Slope:** Based on the observed data, our estimate for the population slope is 2,651. We recognize this is based on incomplete information (a sample), so we also attempt to quantify how uncertain about this estimate that we are. Based on these data, the estimate of the standard error for the slope is 370.

# INTERVAL ESTIMATES

## Using the Estimate and the SE Together

$$\hat{\beta}_1 = 2651$$

$$\hat{\sigma}_{\hat{\beta}_1} = 370$$

Rather than present the estimate and the estimate of the standard error as two separate measures, sometimes we combine them.

*Estimate  $\pm$  Uncertainty*

$$2651 \pm 370$$

$$[2281, 3021]$$

Remember our goal is to estimate the value of the population slope. This interval combines the information in our sample estimate together with the amount of uncertainty in the estimate. It gives us a range of candidates for the population slope. We believe the population slope is between 2,281 and 3,021.

In practice, we typically **double the amount of uncertainty**

$$2651 \pm 2(370)$$

$$[1911, 3391]$$

The range of candidates for the population slope is now larger (interval is wider). This expresses **more uncertainty**. But, it also makes us feel **more confident** that the actual population slope (which we don't know) is one of the candidates.

# Confidence Interval

```
> confint(lm.1)

                2.5 %      97.5 %
(Intercept) -1183.935 23826.693
edu          1896.425  3406.168
```

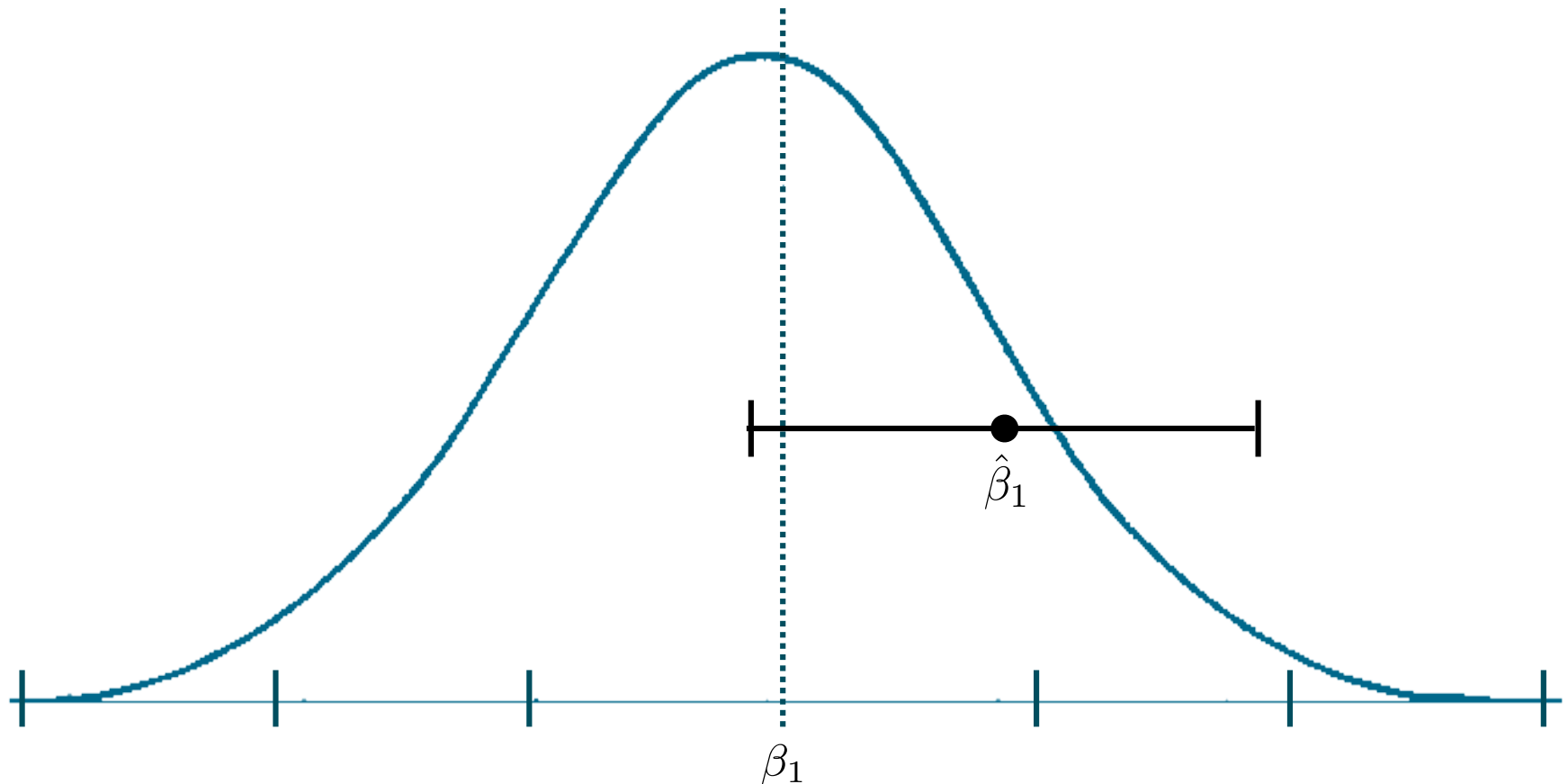
The `confint()` function gives us the endpoints for the interval straightaway.

**Practical Interpretation:** Based on our sample data, we believe the value of the population slope is probably between 1,896 and 3,406.

**Formal Interpretation:** With 95% confidence, the value of the population slope is between 1,896 and 3,406.

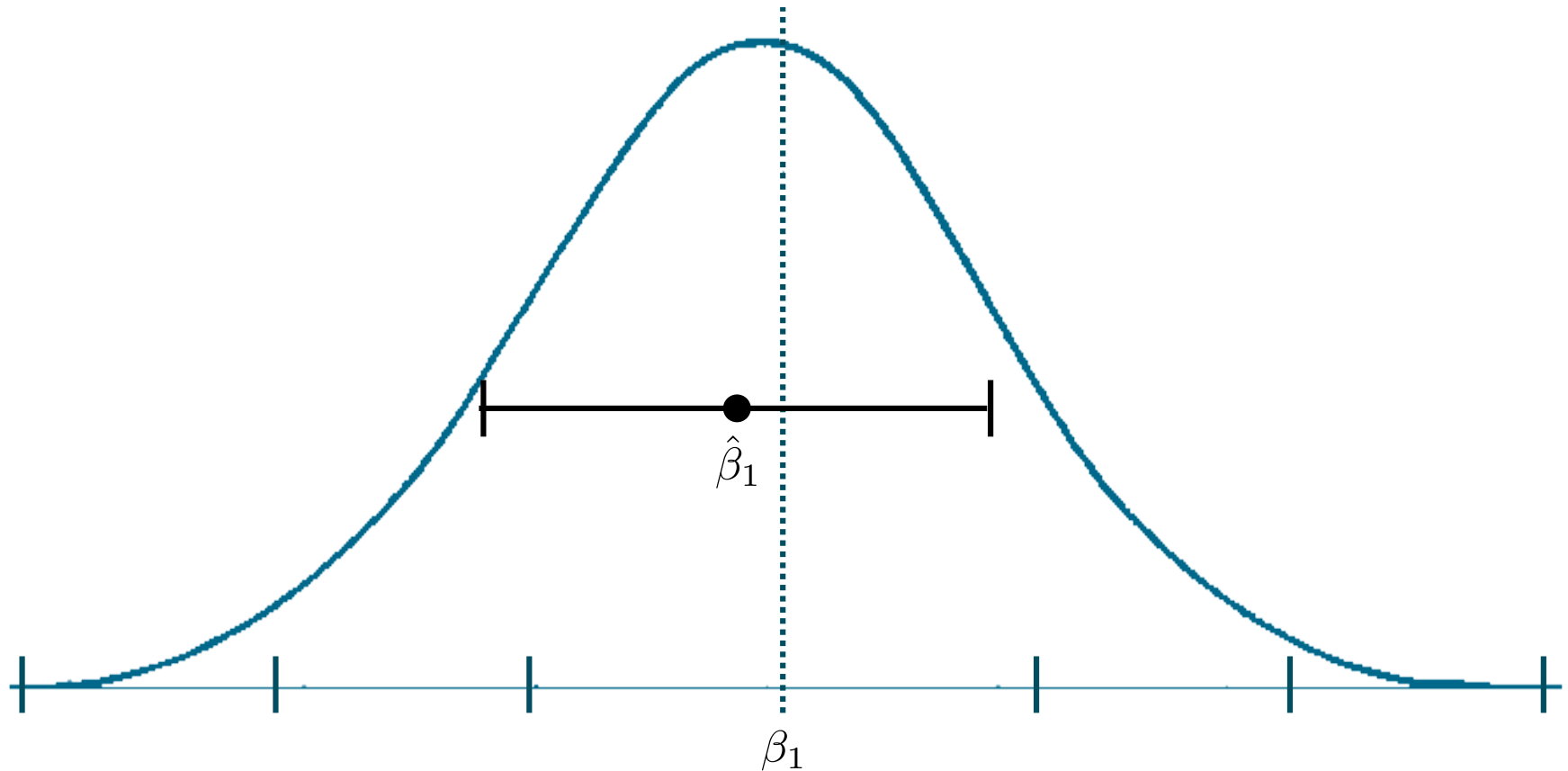


Consider the interval estimate for one of these possible slope estimates.

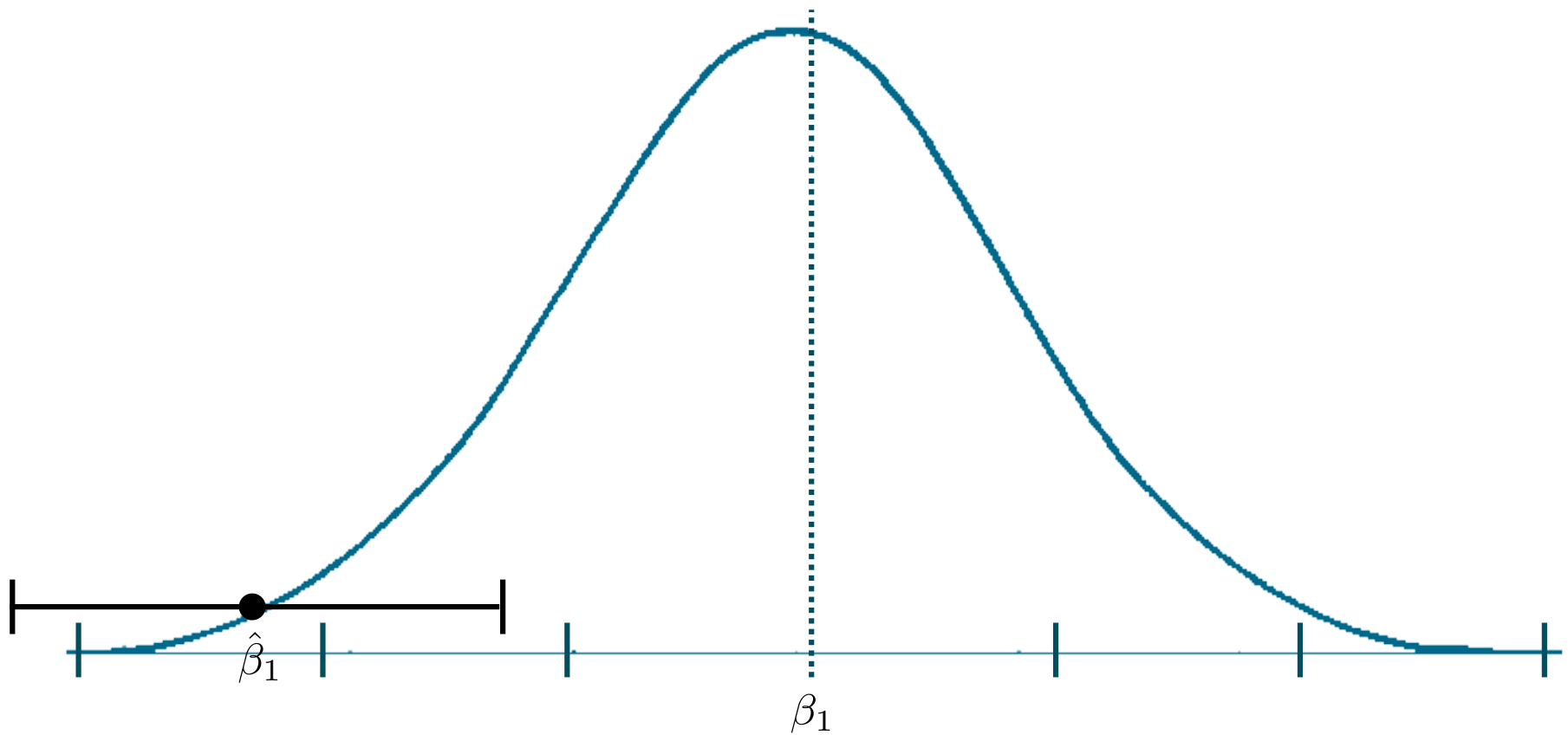


The interval visually shows the range of candidate values for the population slope. In this interval it turns out that the actual population slope was indeed one of the possible candidates.

Consider the interval estimate for another one of these possible slope estimates.

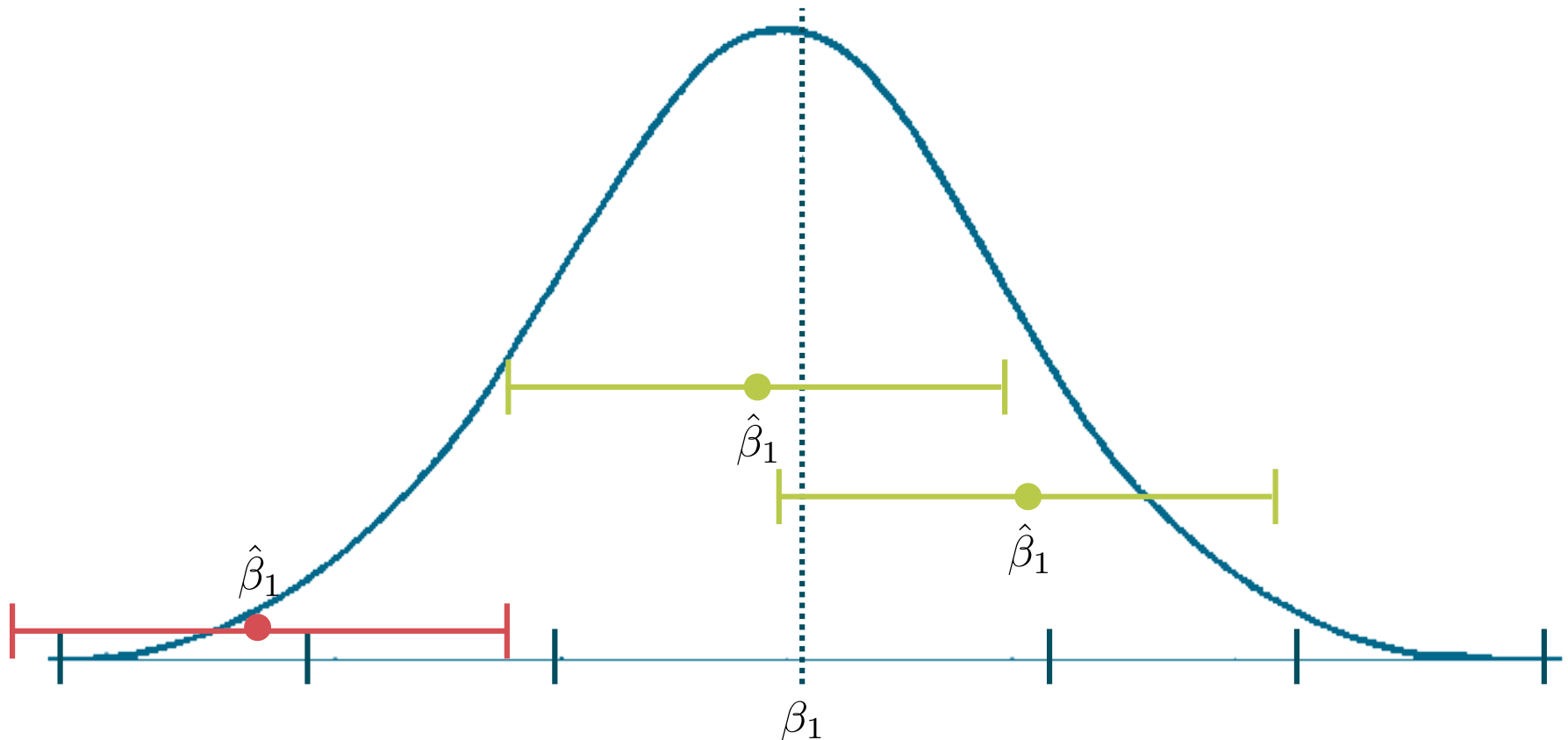


In this interval it turns out that the actual population slope was indeed one of the possible candidates.



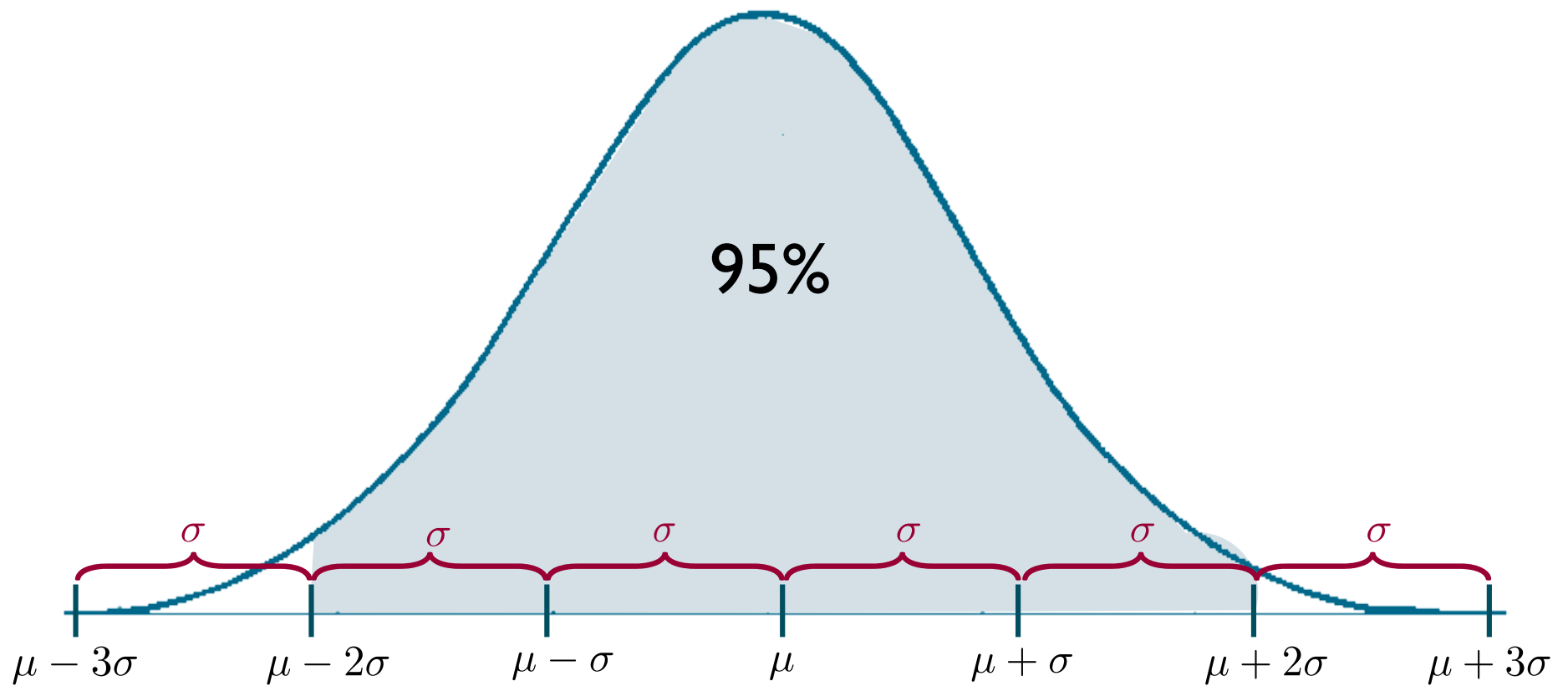
In this interval it turns out that the actual population slope was **NOT** one of the possible candidates.

Some of the slope estimates produce intervals that include the population slope in their range of candidate values. Some do not.



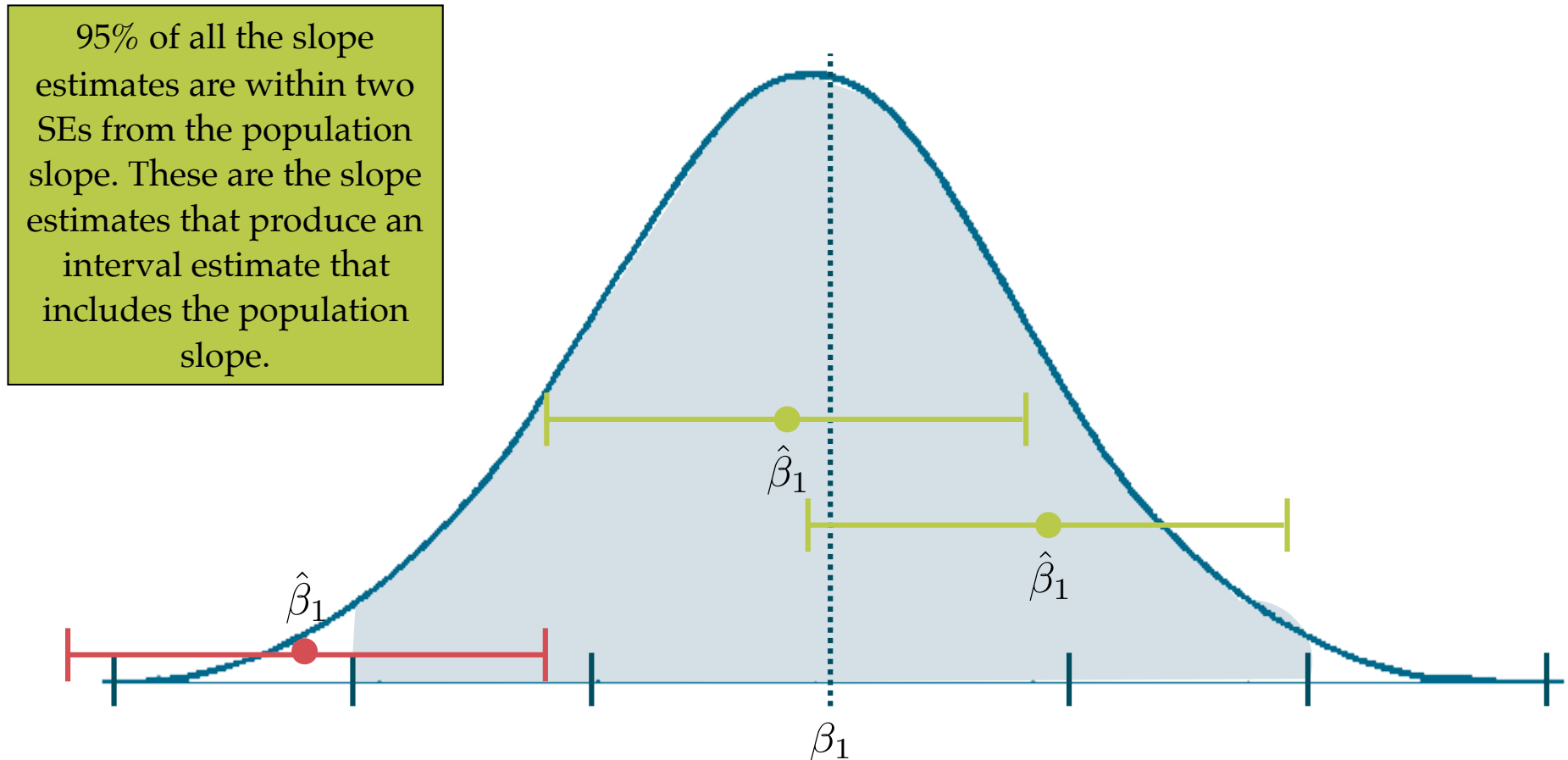
Consider the intervals produced for **all** of the slope estimates. What percentage would include the population slope value in their interval?

## Some Normal Theory (A Reminder)



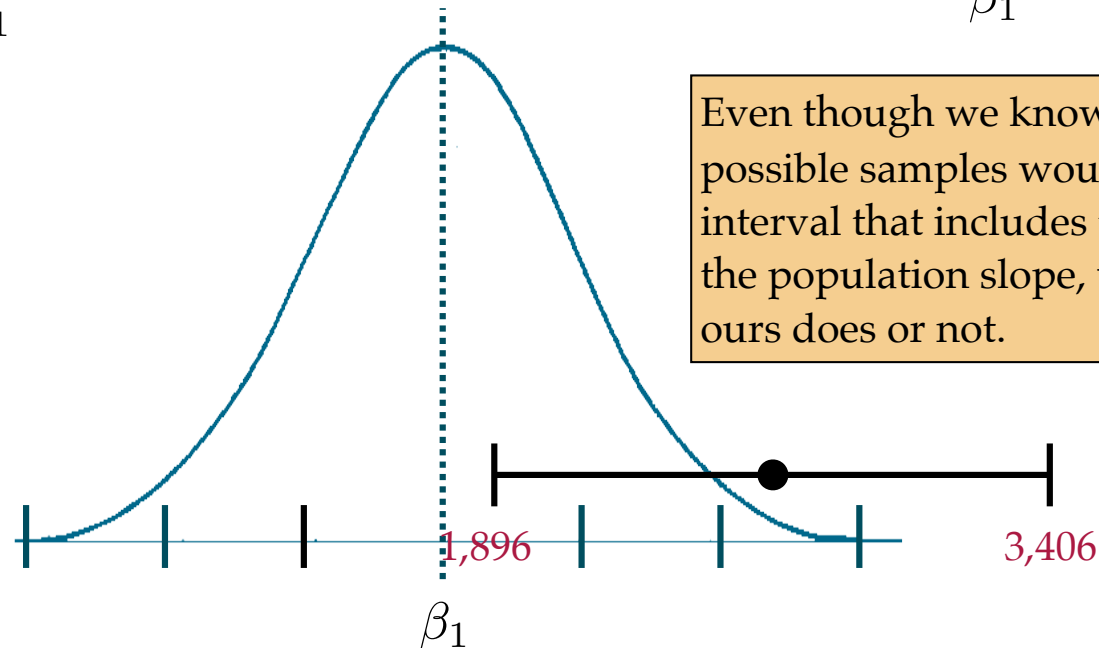
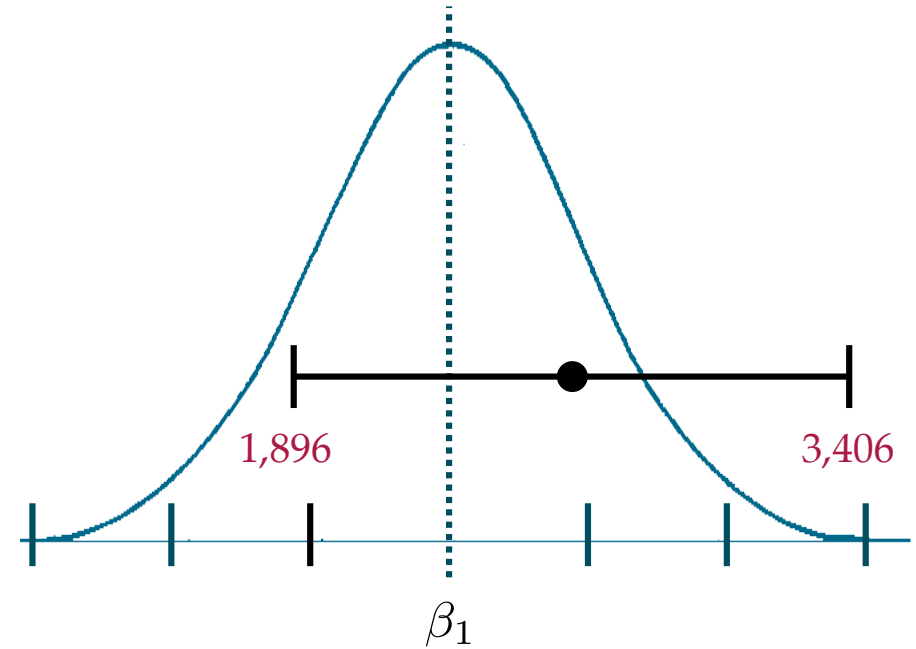
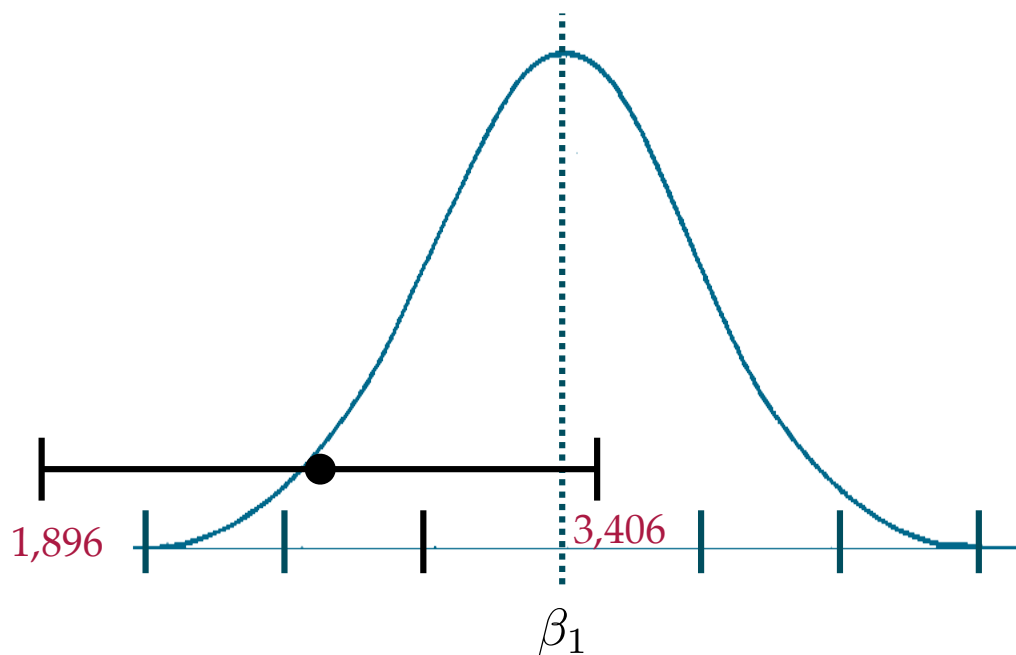
95% of the observations in a normal distribution are within two SDs from the mean.

Our Distribution of Slope Estimates was also Normal...



The 95% in the confidence interval really refers to the idea that across all possible random samples, 95% of the CIs produced will include the population slope within their limits.

In practice, we **do not know** is where that interval is relative to the distribution, since we do not know the value for the population slope.



Even though we know that 95% of all possible samples would produce an interval that includes the actual value of the population slope, **we have no idea** if ours does or not.

# **HYPOTHESIS TESTING**



# Testing Specific Values of the Parameter

Some research questions point to examining whether one of the regression parameters is a specific value. (e.g., Is  $\beta_1 = 0$ ?)

$$H_0 : \beta_1 = 0$$

We state the value we are testing in a statement called a *hypothesis*. When the value we are testing is zero, the statement is referred to as a *null hypothesis*.

It would seem logical that one could just examine the estimate of the parameter from the observed sample to answer this question...

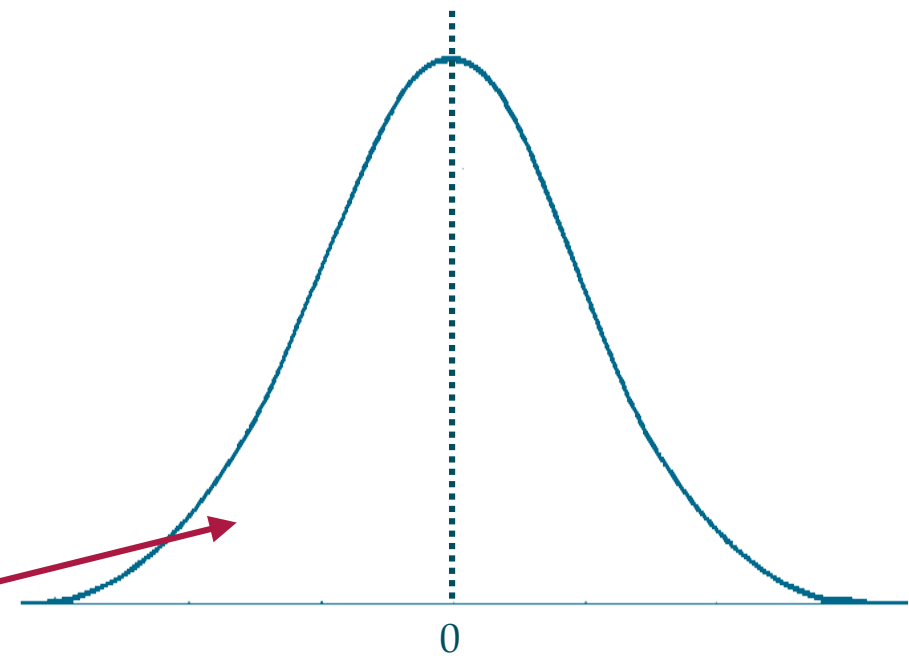
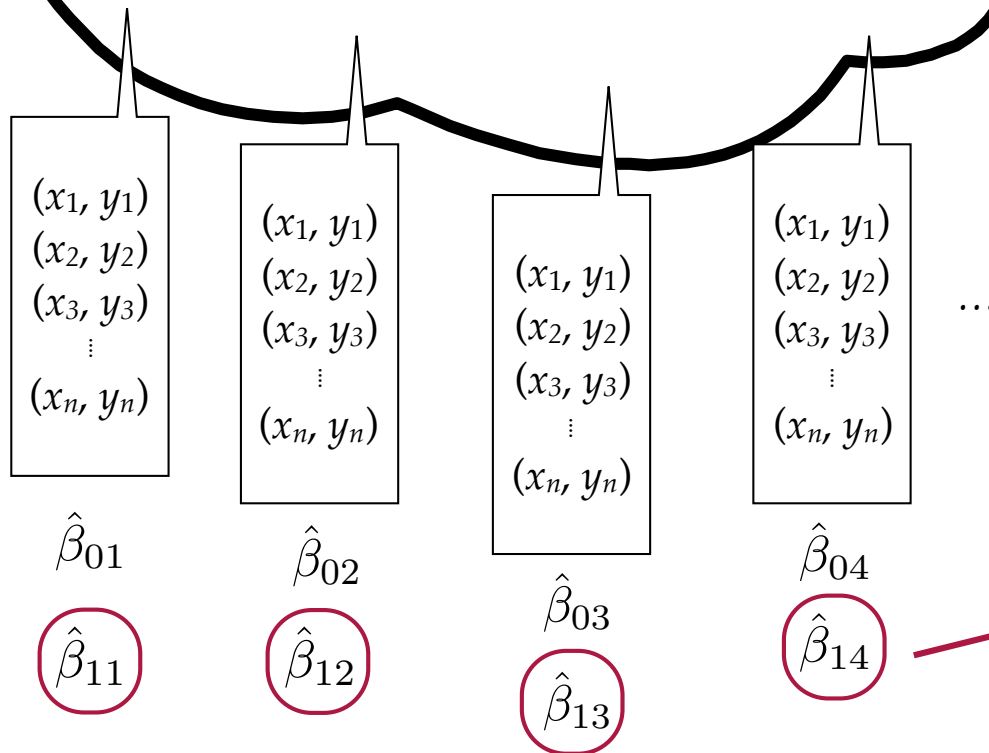
$$\hat{\beta}_1 = 2651$$

But, we also have to account for sampling uncertainty.

The hypothesis is a statement about the population. Here we hypothesize  $\beta_1 = 0$ .

$$Y = \beta_0 + (0)(X) = \beta_0$$

The distribution of the sample slopes is normally distributed with mean  $= \beta_1 = 0$  and some standard deviation.



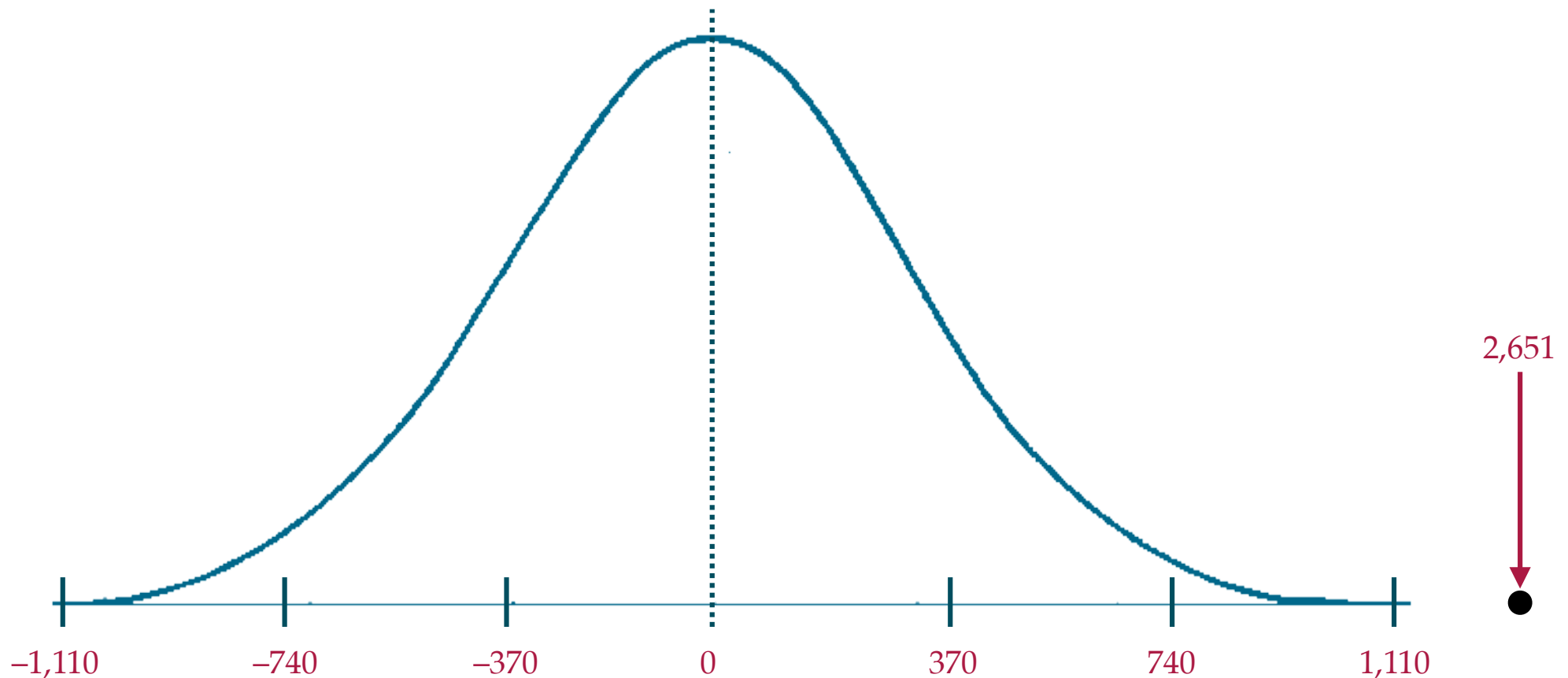
Distribution of the sample slopes

Note that even though the population has a slope of 0, we can still obtain a sample that produces an estimate for that slope that is **not** 0.

The **statistical question** is: If the population slope is 0, how likely is it to see an observed sample with an estimated slope of 1.99?

To answer this, essentially boils down to putting 2,651, the observed slope, in this distribution and then quantifying how likely that result is.

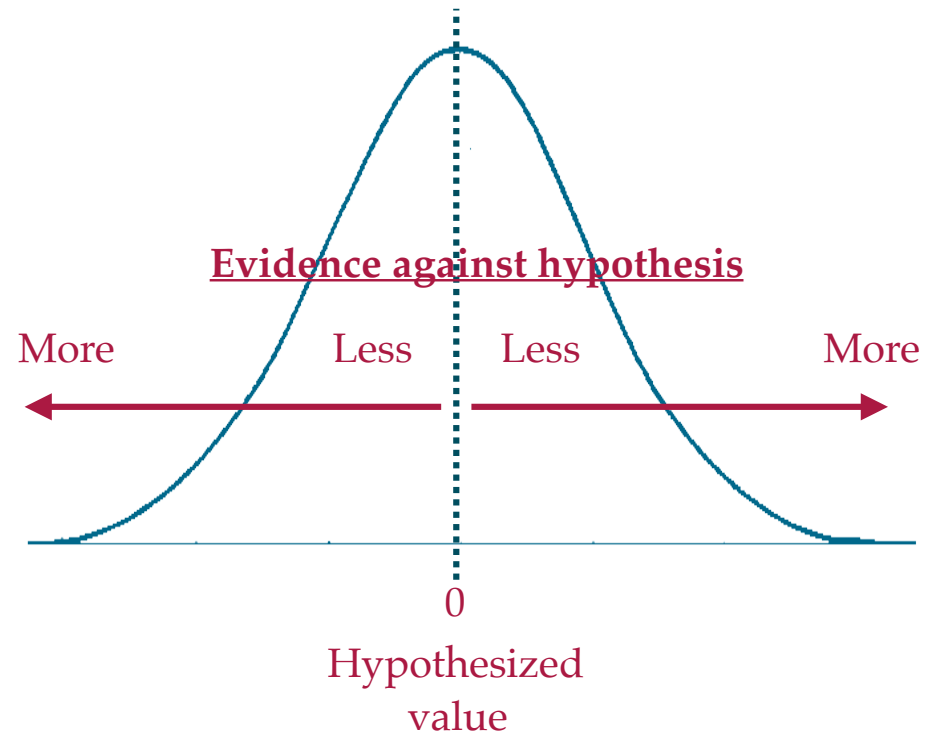
Where 2,651 falls in this distribution depends on the value of the standard error. Fortunately, the SE is primarily a function of sample size and doesn't change because of the slope value. Our estimate for the SE of 370 can be used here as well.



# Evidence

The observed data and estimates we get from that data are the evidence we use to judge a model/hypothesis.

In general, the further an estimate falls from the hypothesized value, the more evidence it provides against the hypothesis.



While evidence may lead us to reject a hypothesis, it can never lead us to confirm a hypothesis. A better scientific question may be, "Can I rule out this hypothesis?"

One way to say how far from the hypothesized value an observed result is, is to compute the distance between those values in the SE metric. In other words, answer the question: How many SEs from the hypothesized value is the observed result?

$$\frac{2651 - 0}{370} = 7.16$$

The observed slope of 2,651 is 7.16 SEs from the hypothesized slope value of 0.

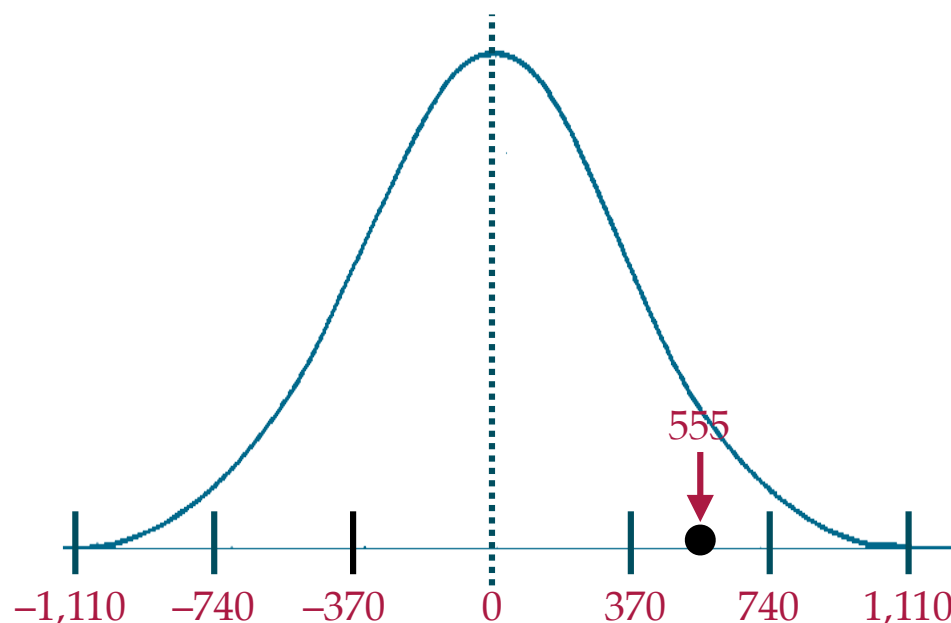
In general, absolute values of this distance over two imply **evidence against the hypothesized model**. Here, a distance of 7.16 is quite a bit of evidence against the hypothesized model that  $\beta_1 = 0$ .

We also quantify how likely the observed value (and more extreme values) are under the hypothesized model.

To understand how to do this, let's pretend we had a slope estimate of 555 (instead of 2,651)

$$\frac{555 - 0}{370} = 1.5$$

The slope estimate of 555 is 1.5 SEs from the hypothesized slope value of 0.



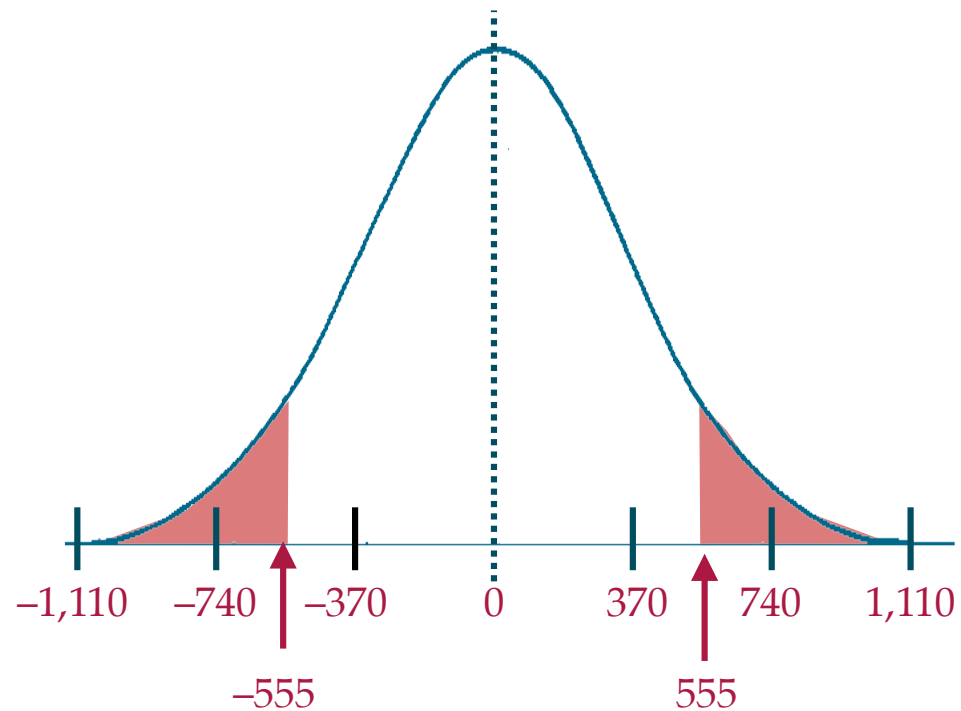
Remember, we are using 555 to measure the degree of evidence against the hypothesis of 0.

- If 555 indeed constitutes evidence against the hypothesis, values **more extreme** than 555 (e.g., 600, 700, 750, etc.) would also constitute evidence against the hypothesis.
- Similarly, since the distribution is symmetric, if we are considering evidence against the hypothesis, **regardless of which side of the distribution** the observed value is on, it constitutes that same degree of evidence against the hypothesis.

The red area constitutes all of the evidence in the distribution against the hypothesis that is at least as extreme as the observed value.

This area, called the  $p$ -value is 0.137 of the distribution.

**Interpretation:** If the hypothesis that  $\beta_1 = 0$  is true, then the probability of obtaining a sample slope *at least as extreme as 555* is 0.137.

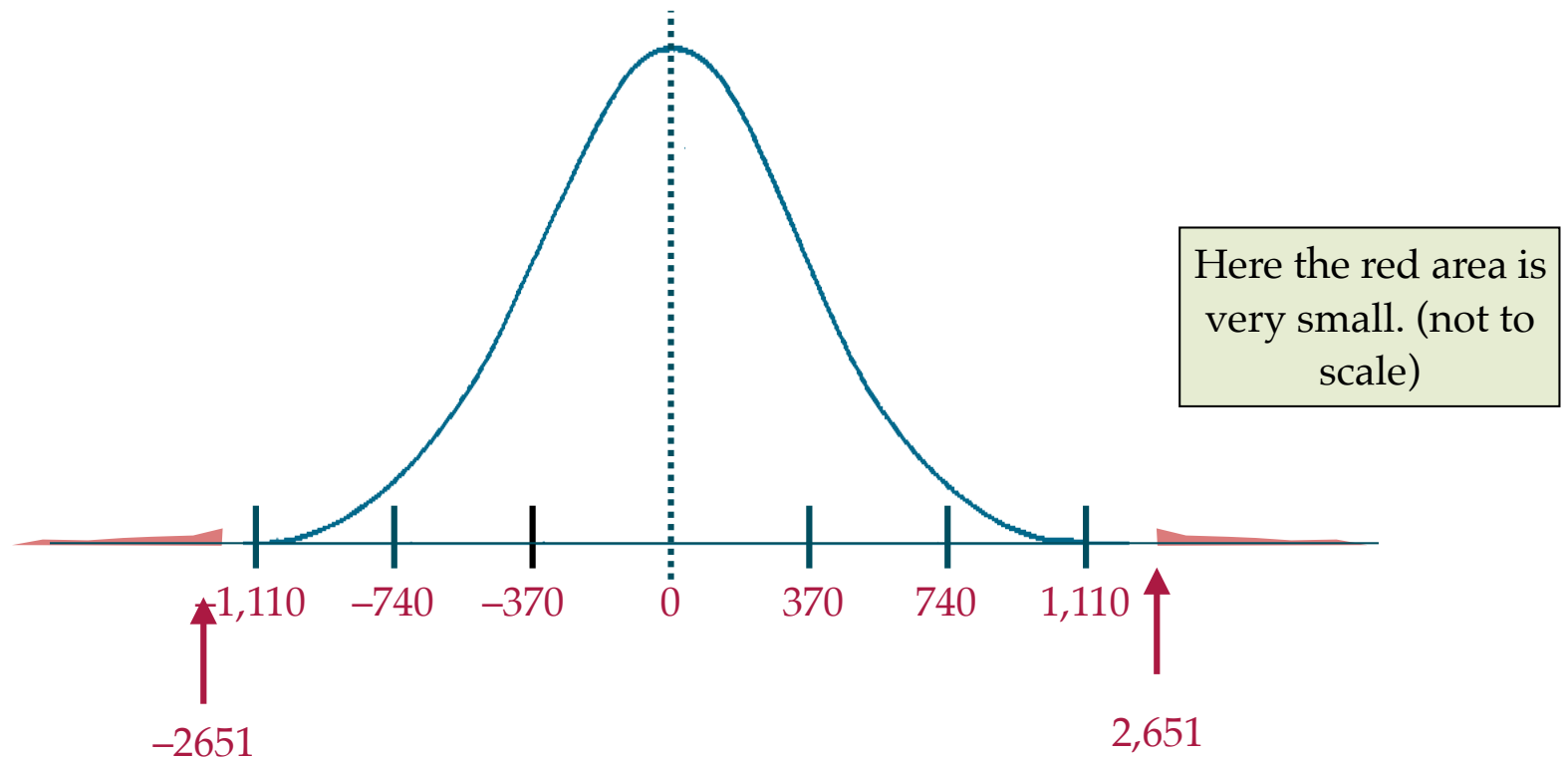


The quantification ( $p$ -value) we compute is based on the hypothesized model being TRUE.

Recall we are asking whether we can rule out the hypothesis in question. To do this we need evidence *against* the original hypothesis.

Thus, the  $p$ -value is small when there is **more evidence against the hypothesized model** and large when there is less evidence against the hypothesized model.

## Back to the Actual Data



### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11321.4	6123.2	1.849	0.0743	.
edu	2651.3	369.6	7.173	5.56e-08	***

The  $p$ -value is also displayed in the `summary()` output.



### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11321.4	6123.2	1.849	0.0743	.
edu	2651.3	369.6	7.173	5.56e-08	***

R uses **scientific notation** to output really small  $p$ -values. Here, 5.56e-08 means  $5.56 \times 10^{-8}$

If the hypothesized model that  $\beta_1 = 0$  is true, then the probability of obtaining a sample slope at least as extreme as 2,651 is 0.0000000556.

In practice, we would report this as  $p < .001$ .

This constitutes a great deal of evidence against the hypothesis, and **in practice**, we would reject the hypothesis that  $\beta_1 = 0$ .

How much evidence do we need to reject the hypothesis? In the social sciences,  $p$ -values that are less than 0.05 are often considered statistically significant.

The  $p$ -values given in software are typically tests of the hypothesized value of 0.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	11321.4	6123.2	1.849	0.0743	.
edu	2651.3	369.6	7.173	5.56e-08	***

### Intercept

$$H_0 : \beta_0 = 0$$

The sample intercept of 11,321 is 1.85 SEs from the hypothesized value of 0. If the hypothesized model that  $\beta_0 = 0$  is true, the probability of obtaining an intercept at least as extreme as 11,321 is 0.074.

Based on the  $p$ -value of  $<0.001$ , we reject the hypothesis that  $\beta_0 = 0$  and conclude that the population intercept is unlikely to be zero.

# Confidence Interval (Revisited)

Here, the confidence interval for the slope provided us a range of candidate values for the population slope.

$[1896, 3406]$

Consider the model

$$H_0: \beta_1 = 0$$

If you are only interested in whether you should reject (or fail to reject) this particular model, we can examine the confidence interval and ask whether the hypothesized value, in this case 0, is in the range of candidate values for the population slope.

⋮  
0



# **MODEL-LEVEL INFERENCE**

Sometimes you may want to carry out inference for the **model as a whole**, rather than for the individual parameters. The statistical question at the model level is: *Does the model explain variation in the outcome?*

$$H_0 : \rho^2 = 0$$

The model-level inferential information is shown at the bottom of the `summary()` output.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11321.4	6123.2	1.849	0.0743 .
edu	2651.3	369.6	7.173	5.56e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom

Multiple R-squared: 0.6317, Adjusted R-squared: 0.6194

F-statistic: 51.45 on 1 and 30 DF, p-value: 5.562e-08

Based on the evidence ( $p < .001$ ), we reject the hypothesis that the model does not explain variation in incomes in the population,  $F(1, 30) = 51.45$ . Our best guess for the amount of variation explained by the model is 63.2% (the Multiple R-squared value).

We can also get the model-level inferential information from the `anova()` output. This also gives us the ANOVA decomposition for the model.

```
> anova(lm.1)
```

### Analysis of Variance Table

Response: income

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
edu	1	4147330492	4147330492	51.452	5.562e-08 ***
Residuals	30	2418196934	80606564		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Note that the two  $df$  for the  $F$ -statistic correspond to the  $df$  in each row of the ANOVA table. The first  $df$  is the model degrees-of-freedom, and the second  $df$  is the residual degrees-of-freedom.

## Same Result for Test of Slope and Model???

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11321.4	6123.2	1.849	0.0743 .
edu	2651.3	369.6	7.173	5.56e-08 ***
---				

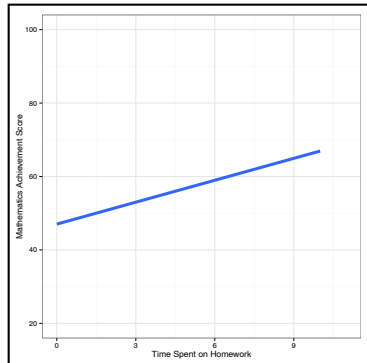
Multiple R-squared: 0.6317, Adjusted R-squared: 0.6194  
F-statistic: 51.45 on 1 and 30 DF, p-value: 5.562e-08

In simple regression models, the results of the model-level inference (i.e., the  $p$ -value) is exactly the same as that for the coefficient-level inference for the slope.

That is because the model is composed of a single predictor, so asking whether the *model accounts for variation* in achievement scores **is the same as** asking whether *differences in time spent on homework account for variation* in achievement scores.

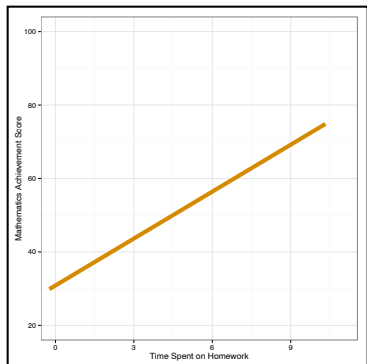
Once we have multiple predictors in the model, the model-level results and predictor-level results will not be the same.

# Confidence Envelope for the Model

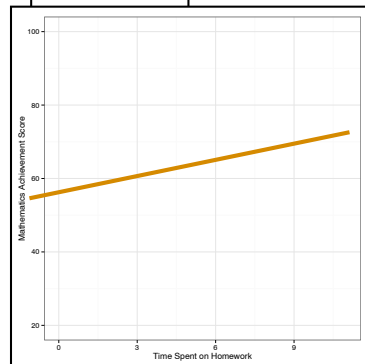


$$Y = \beta_0 + \beta_1(X)$$

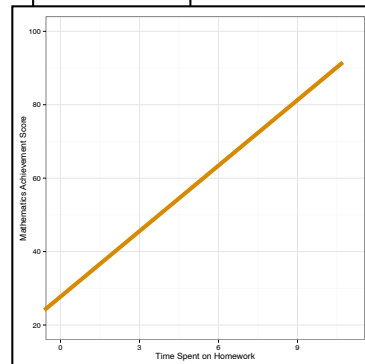
$(x_1, y_1)$   
 $(x_2, y_2)$   
 $(x_3, y_3)$   
 $\vdots$   
 $(x_n, y_n)$



$(x_1, y_1)$   
 $(x_2, y_2)$   
 $(x_3, y_3)$   
 $\vdots$   
 $(x_n, y_n)$

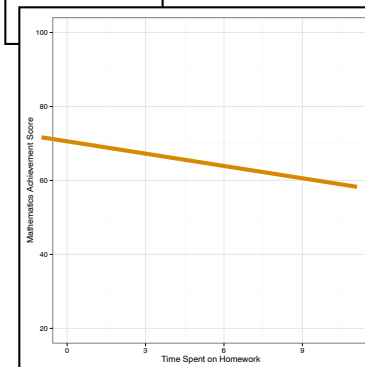


$(x_1, y_1)$   
 $(x_2, y_2)$   
 $(x_3, y_3)$   
 $\vdots$   
 $(x_n, y_n)$



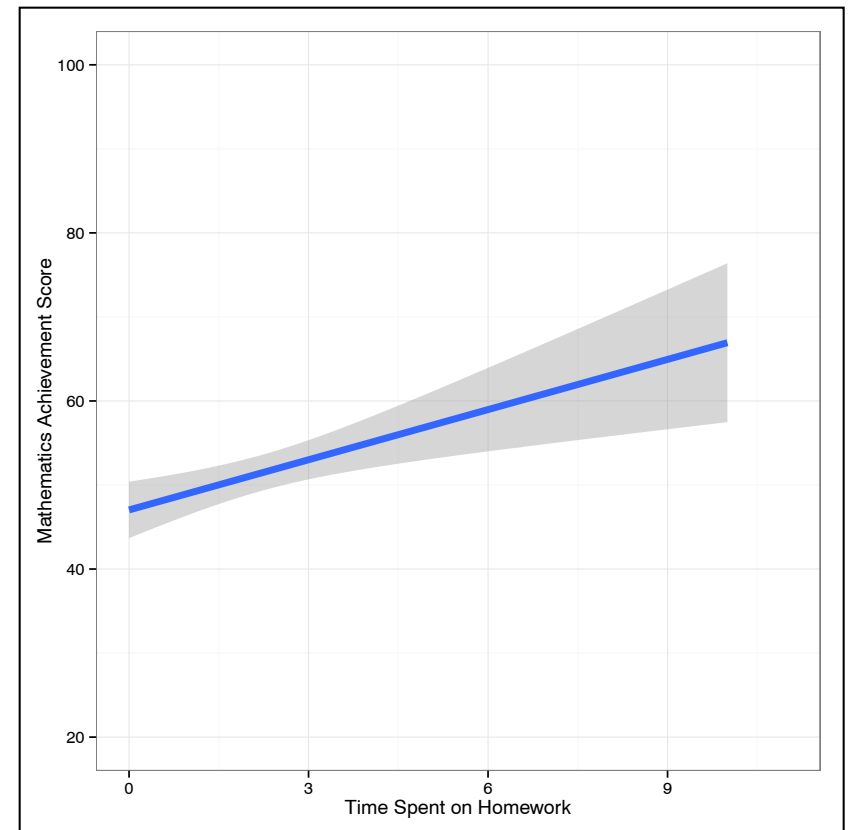
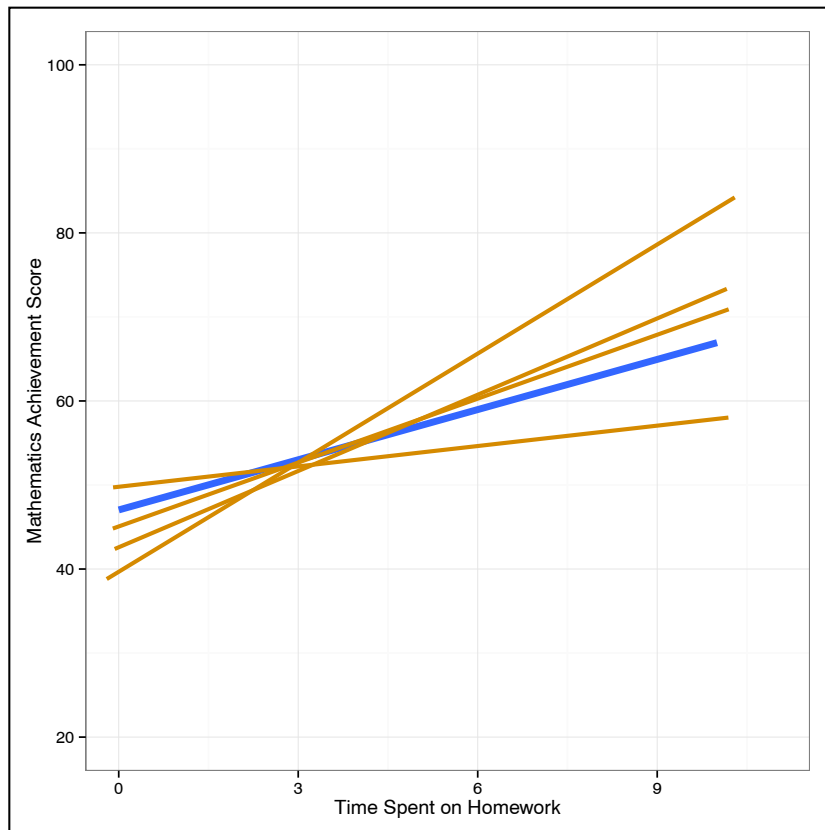
$(x_1, y_1)$   
 $(x_2, y_2)$   
 $(x_3, y_3)$   
 $\vdots$   
 $(x_n, y_n)$

...





As we superimpose these lines on the same plot we can visualize the uncertainty in the model (i.e., in the intercept and slope).



In practice, we estimate the uncertainty from the sample data.

To plot the simple regression model and the model uncertainty, we use the `geom_smooth()` function. The `method=` argument is "lm" to estimate the regression model, and `se=TRUE` adds the confidence envelope. If you only want the regression line, use `se=FALSE`.

```
> ggplot(data = city, aes(x = edu, y = income)) +  
  geom_smooth(method = "lm", se = TRUE) +  
  xlab("Education Level") +  
  ylab("Income") +  
  theme_bw()
```