Simple Regression Model

Juergen Meinecke

Roadmap

Statistical Inference

- The Four OLS Assumptions

OLS estimators are statistics that vary with the underlying random sample

Three natural questions to ask are:

- What is the expected value of the OLS estimator?
- What is the variance of the OLS estimator?
- What is the sampling distribution of the OLS estimator?

To answer these questions, we need to impose four assumptions

They are known as the OLS Assumptions

Assumption (OLS Assumption 1)

The error term u_i is conditionally mean independent (CMI) of X_i, meaning:

$$E[u_i|X_i] = E[u_i] = \mu_u.$$

Note: many textbooks simply set $\mu_u = 0$, which is without loss of generality

CMI restricts the expected value of the error terms

Although we do not observe the error terms u_i, and therefore we do not know their distribution, we are willing to impose a restriction on their expected values

The essential restriction in Assumption 1 is that the expected value of u_i *is not a function of* X_i

When we write that $E[u_i|X_i] = E[u_i]$, we are saying that $E[u_i|X_i]$

- is not dependent on X_i, and instead
- is constant with value μ_u

Assumption 1 says that X_i is not informative for the mean of u_i

Benchmark for thinking about Assumption 1: a *randomized controlled experiment* in which X_i is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments)

If X_i is assigned randomly then the error term u_i must be statistically independent of X_i, and thus $E[u_i|X_i] = E[u_i]$ seems plausible

With observational data, however, we will need to think hard about whether $E[u_i|X_i] = E[u_i]$ is persuasive

Assumption (OLS Assumption 2)

The sample data (X_i, Y_i), $i = 1, \dots, n$, are independent and identically distributed (i.i.d.) draws from the joint population distribution.

Assumption 2 arises automatically if the entity i is sampled by simple random sampling:

- The entities are selected from the same population, so (X_i, Y_i) are identically distributed for all $i = 1, \ldots n$
- The entities are selected at random, so the values of (X_i, Y_i) for different entities are independently distributed.

The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity
(panel data and time series data)

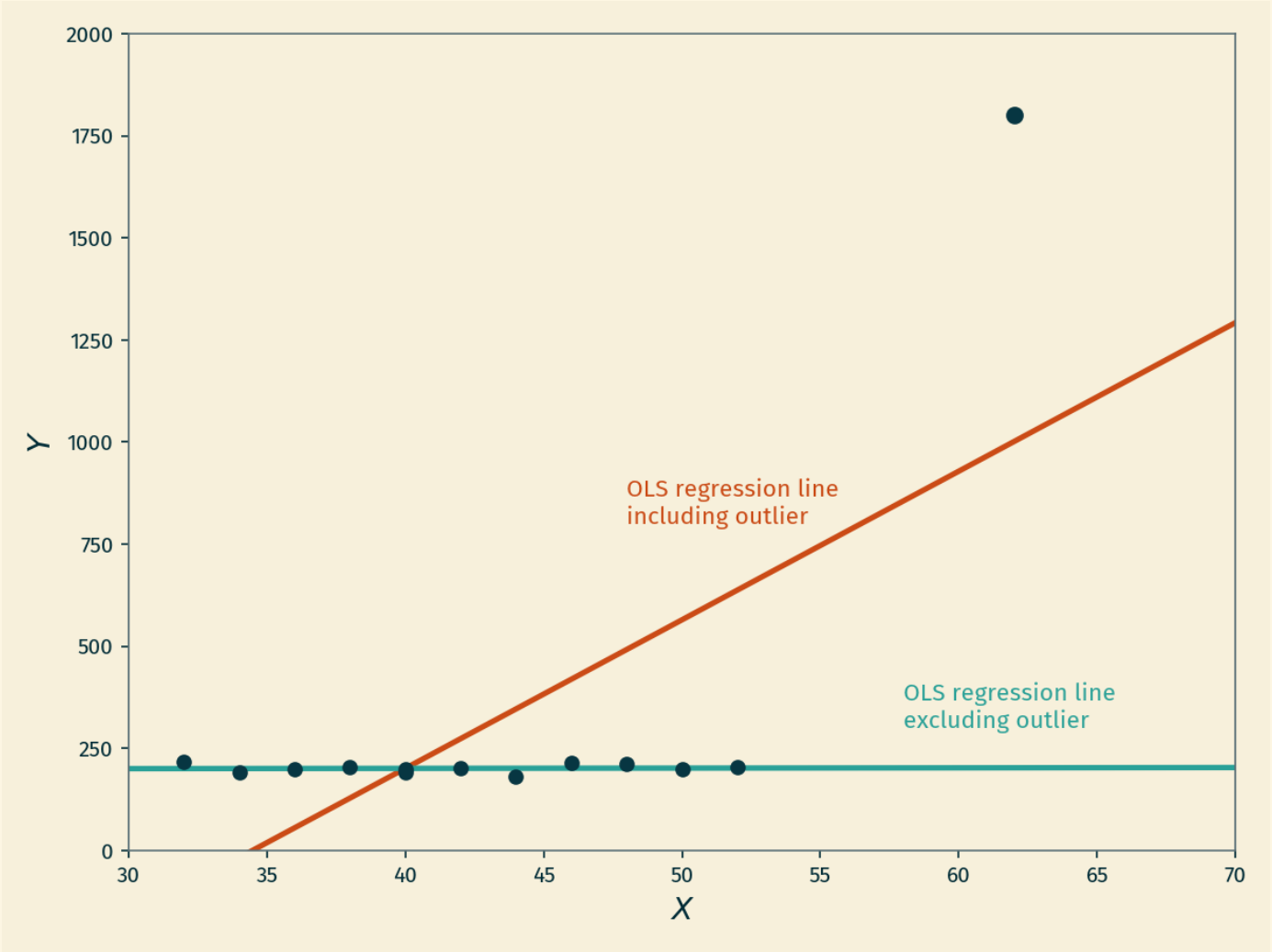
Assumption (OLS Assumption 3)

Large outliers in X_i or Y_i are rare. Technically, both X_i and Y_i have finite fourth moments:

$$E[X_i^4] < \infty, \qquad E[Y_i^4] < \infty.$$

A large outlier is an extreme value (positive or negative)

Large outliers for X_i and Y_i can strongly influence OLS estimates



In practice, outliers are often data glitches
(coding or recording problems)

Simple suggestions:

- literally look at your data spreadsheet
- are there any suspicious numbers?
  (for example, somebody's age was accidentally recorded as a negative number)
- do a scatterplot to spot outliers

Technically, if X_i and Y_i are bounded, then they will have finite fourth moments
(Standardized test scores automatically satisfy this;
student-teacher ratio, family income, and many other real world variables satisfy this too.)

Assumption (OLS Assumption 4a)

The error term u_i is ***homoskedastic****, meaning:*

$$Var\,(u_i|X_i) = \sigma_u^2.$$

Homoskedasticity restricts the variance of the error terms

Although we do not observe the error terms u_i, and therefore we do not know their distribution, we are willing to impose a restriction on their variances

The essential restriction in Assumption 4a is that the variance of u_i *is not a function of* X_i

When we write that Var $(u_i|X_i) = \sigma_u^2$, we are saying that Var $(u_i|X_i)$

- is not dependent on X_i, and instead
- is constant with value σ_u^2

Homoskedasticity is not absolutely essential as an assumption, but it makes the derivation of the variance of the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ easier

Simple Regression Model

Juergen Meinecke

Roadmap

Statistical Inference

- Sampling Distribution of the OLS Estimator

OLS estimators are statistics that vary with the underlying random sample

Three natural questions to ask are:

- What is the expected value of the OLS estimator?
- What is the variance of the OLS estimator?
- What is the sampling distribution of the OLS estimator?

Using the four OLS assumptions, we will answer these questions now

In what follows, we will only focus on the coefficient β_1 and its OLS estimator $\hat{\beta}_1$

We will ignore the coefficient β_0 and its OLS estimator $\hat{\beta}_0$

Main reason: the math is tedious and focusing our attention on only one coefficient is sufficient

Doing the math for the other coefficient would not be substantially more difficult

Some preliminary algebra before we derive the sampling distribution

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u},$$

So therefore,

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$$

Thus,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^{n} (X_i - \bar{X})[\beta_1 (X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \beta_1 \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}
\end{aligned}$$

Copying and pasting the last equation and simplifying...

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and therefore

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Intuitively, what are the numerator and the denominator measuring?

- numerator: sample covariance between X_i and u_i
- denominator: sample variance of X_i

The difference between $\hat{\beta}_1$ and β_1 is therefore equal to the ratio of the sample covariance between X_i and u_i and the sample variance of X_i

But we can simplify even further!

Rewrite the numerator

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) &= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left(\sum_{i=1}^{n}(X_i - \bar{X})\right)\bar{u} \\ &= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u} \\ &= \sum_{i=1}^{n}(X_i - \bar{X})u_i\end{aligned}$$

Substituting back the numerator results in

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

We can now turn to the derivations of $\mathrm{E}[\hat{\beta}_1|X_i]$ and $\mathrm{Var}\,(\hat{\beta}_1|X_i)$

$$
\begin{aligned}
\mathsf{E}[\hat{\beta}_1 | X_i] &= \mathsf{E}\left[\beta_1 + \frac{\sum_{i=1}^{n} (X_i - \bar{X}) u_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \middle| X_i \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \mathsf{E}\left[\sum_{i=1}^{n} (X_i - \bar{X}) \cdot u_i \middle| X_i \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} \mathsf{E}\left[(X_i - \bar{X}) \cdot u_i | X_i \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X}) \cdot \mathsf{E}\left[u_i | X_i \right] \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X}) \cdot \mu_u \\
&= \beta_1 + \frac{\mu_u}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X}) \\
&= \beta_1
\end{aligned}
$$

The first equality holds by definition

The second equality holds by property (2) of the expected value and because the denominator can be treated as a constant (because the expected value is conditional on X_i)

The third equality holds by property (3) of the expected value

The fourth equality holds because the factor $X_i - \bar{X}$ can be treated as a constant

The fifth equality holds by Assumption 1

The final equality follows because $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$

Overall, this means that $\hat{\beta}_1$ is an unbiased estimator of β_1

Deriving the variance of $\hat{\beta}_1$ is equally awkward, but we need to do it

Recall our earlier expression for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Now instead of $\mathrm{E}[\hat{\beta}_1|X_i]$ we are interested in $\mathrm{Var}(\hat{\beta}_1|X_i)$

Plugging in and solving, step by step, we get

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1|X_i) &= \mathrm{Var}\left(\beta_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big|X_i\right) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big|X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\mathrm{Var}\left(\sum_{i=1}^{n}(X_i - \bar{X})u_i\Big|X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\sum_{i=1}^{n}\mathrm{Var}\left((X_i - \bar{X})u_i\Big|X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\mathrm{Var}(u_i|X_i) \\
&= \frac{1}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\sigma_u^2 \\
&= \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}
$$

The first equality holds by definition

The second equality holds by property (2) of the variance

The third equality holds by property (2) of the variance

The fourth equality holds by property (3) of the variance
(all covariances are zero because of i.i.d. sampling)

The fifth equality holds because the factor $X_i - \bar{X}$ can be treated as a constant

The sixth equality holds by Assumption 4a

The final equality results from algebraic simplification

Recall that the sample variance of X_i is $s_X^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$

Then

$$\begin{aligned}\text{Var}(\hat{\beta}_1 | X_i) &= \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\frac{1}{n}\sigma_u^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{1}{n}\frac{\sigma_u^2}{s_X^2}\end{aligned}$$

This means that the conditional variance of $\hat{\beta}_1$ is:

- proportional to the population variance of the error term u_i
- inversely proportional to the variance of the regressor X_i and the sample size n

Now there is only one additional step to turn this into a result that will be useful for applying the central limit theorem

Instead of the sample variance of X_i, plug in its *asymptotic approximation*

$$s_X^2 \simeq \sigma_X^2$$

This means, that for large samples the sample variance of X_i is almost equal to the population variance of X_i

With this approximation

$$\text{Var}(\hat{\beta}_1 | X_i) \simeq \frac{1}{n} \frac{\sigma_u^2}{\sigma_X^2}$$

Summarizing the sampling distribution thus far

We have restricted attention to $\hat{\beta}_1$
(dealing with $\hat{\beta}_0$ is not more difficult)

We have learned that

$$\mathsf{E}\left[\hat{\beta}_1 | X_i\right] = \beta_1$$

$$\text{Var}\left(\hat{\beta}_1 | X_i\right) \simeq \frac{1}{n}\frac{\sigma_u^2}{\sigma_X^2}$$

Putting things together, the OLS estimator has a distribution

$$\hat{\beta}_1 \overset{\text{approx.}}{\sim} P\left(\beta_1, \frac{1}{n}\frac{\sigma_u^2}{\sigma_X^2}\right),$$

where P is a placeholder for some unknown distribution

The central limit theorem tells us what P is

Theorem

*The **asymptotic distribution of the OLS estimator** $\hat{\beta}_1$ under OLS Assumptions 1,2,3, and 4a is*

$$\hat{\beta}_1 \overset{approx.}{\sim} N\left(\beta_1, \frac{1}{n}\frac{\sigma_u^2}{\sigma_X^2}\right).$$

A similar result holds for $\hat{\beta}_0$

A quick corollary is

Corollary

$$\sqrt{n}\frac{\hat{\beta}_1 - \beta_1}{\sigma_u/\sigma_X} \overset{approx.}{\sim} N(0,1)$$

This is entirely analogous to the sample average in univariate statistics:

$$\sqrt{n}\frac{\bar{Y} - \mu_Y}{\sigma_Y} \overset{\text{approx.}}{\sim} N(0,1)$$

Simple Regression Model

Juergen Meinecke

Roadmap

Statistical Inference

Confidence Intervals and Hypothesis Testing

The central limit theorem led to the following sampling distribution for the OLS estimator of the slope coefficient β_1:

$$\hat{\beta}_1 \overset{approx.}{\sim} \mathsf{N}\left(\beta_1, \frac{1}{n}\frac{\sigma_u^2}{\sigma_X^2}\right).$$

With this result we can

- derive confidence intervals for β_1,
- propose hypothesis tests for β_1

Notice that the approximate variance of $\hat{\beta}_1$ is

$$\text{Var}\,(\hat{\beta}_1) = \frac{1}{n}\frac{\sigma_u^2}{\sigma_X^2}$$

Therefore, $\text{std}(\hat{\beta}_1) = \frac{1}{\sqrt{n}}\frac{\sigma_u}{\sigma_X}$

By the virtue of the normal distribution, we should therefore consider the range $\hat{\beta}_1 \pm 1.96 \cdot \text{std}(\hat{\beta}_1)$

This, of course, results in the confidence interval

$$CI(\beta_1) := \left[\hat{\beta}_1 - 1.96 \cdot \frac{\sigma_u}{\sqrt{n}\sigma_X}, \hat{\beta}_1 + 1.96 \cdot \frac{\sigma_u}{\sqrt{n}\sigma_X}\right]$$

Only problem: we do not know σ_u and σ_X (why?)

But can estimate them easily instead:

- σ_u is estimated by s_u
- σ_X is estimated by s_X

Do you remember the definition of s_u and s_X?

It should be obvious to you that

$$s_u := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2} \qquad s_X := \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

An operational version of the confidence interval therefore is given by

$$CI(\beta_1) := \left[\hat{\beta}_1 - 1.96 \cdot \frac{s_u}{\sqrt{n}s_X}, \hat{\beta}_1 + 1.96 \cdot \frac{s_u}{\sqrt{n}s_X}\right]$$

The ratio $s_u/(\sqrt{n}s_X)$ has a special name

Definition

The **standard error of** $\hat{\beta}_1$ is defined as $SE(\hat{\beta}_1) := s_u/(\sqrt{n}s_X)$.
It is the estimated standard deviation of the OLS estimator $\hat{\beta}_1$.

Expressing the confidence interval for the population coefficient β_1 in terms of the standard error:

$$CI(\beta_1) := \left[\hat{\beta}_1 - 1.96 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \cdot \text{SE}(\hat{\beta}_1)\right]$$

This looks very similar to the confidence interval for the population mean from univariate statistics:

$$CI(\mu_Y) := [\bar{Y} - 1.96 \cdot \text{SE}(\bar{Y}), \bar{Y} + 1.96 \cdot \text{SE}(\bar{Y})]$$

Confidence intervals are always constructed that way:
Estimated population parameter plus/minus 1.96 times standard error

When you run a regression in `Python`, it will compute and display

- coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- standard errors for both $\hat{\beta}_0$ and $\hat{\beta}_1$
- confidence intervals for β_0 and β_1
- t-statistics
- p-values

Let's take a look...

Student-teacher ratio example

Python Code (output edited)

```
> import pandas as pd
> import statsmodels.formula.api as smf
> df = pd.read_csv('caschool.csv')
> formula = 'testscr ~ str'
> model1 = smf.ols(formula, data=df, missing='drop')
> reg1 = model1.fit(use_t=False)
> print(reg1.summary())

OLS Regression Results
==============================================================================
Dep. Variable:                testscr   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     22.58
No. Observations:                 420
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    698.9330      9.467     73.825      0.000     680.377     717.489
str           -2.2798      0.480     -4.751      0.000      -3.220      -1.339
==============================================================================
Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Python computes that $\hat{\beta}_0 = 698.93$ and $\hat{\beta}_1 = -2.28$

These are the estimates for the unknown population coefficients β_0 and β_1

We care in particular about β_1

If STR goes up by one then TestScore is expected to go down by 2.28

But is this number statistically different from zero?

The answer depends on how precise our estimate $\hat{\beta}_1$ is

Luckily, Python provides the standard error: 0.48

Going 1.96 times the standard error to the left and right of the coefficient estimate gives us the 95% confidence interval

We could easily calculate this ourselves, but again `Python` already does it for us

`Python` tells us that the confidence interval for β_1 is $[-3.22, -1.34]$

Because this confidence interval does not contain zero, we conclude that the the true population parameter β_1 is likely not equal to zero (at a 95% confidence level)

Alternatively, we could have used the p-value to arrive at the same conclusion

`Python` even tells us the *p-value*: it is `0.000` which means a value less than `0.050` which is enough information to reject the null hypothesis that $\beta_1 = 0$ at a 95% significance level

Remember: the p-value is the smallest significance level at which the null hypothesis can be rejected

There is another piece of information that `Python` calculates: the *t-statistic*

In the sample regression output above, that number equals -4.75 for STR

What does this mean?

Definition

The **t-statistic** is defined as

$$t := \frac{\hat{\beta}_1 - \beta_1^{H_0}}{\mathrm{SE}(\hat{\beta}_1)},$$

where $\beta_1^{H_0}$ is the value of the population parameter β_1 under the null hypothesis.

In the previous few slides we have considered the null hypothesis $H_0 : \beta_1 = 0$

Therefore, the t-statistic would translate to $\frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$

What is the approximate distribution of this expression?

Can you see that

$$\frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \overset{approx.}{\sim} N(0,1)?$$

In other words, the t-statistic has an approximate *standard normal distribution* under the null hypothesis $H_0 : \beta_1 = 0$

When we know the value of the t-statistic, all we need to do is check whether it is smaller than -1.96 or larger than 1.96

In the current example, the t-stat is -4.75

We therefore conclude that it is sufficiently far away from zero; the true population parameter is not equal to zero
(at a 95% confidence level)

In summary, there are three equivalent ways of testing the hypothesis $H_0 : \beta_1 = 0$, via the

- t-statistic
- p-value
- confidence interval

Each of these are constructed as functions of $\hat{\beta}_1$ and SE($\hat{\beta}_1$)

Each of these just presents that information in a different way

You can choose whichever method you prefer, they will all give you the same answer regarding the significance of β_1