

# Linear Regression: Inference

**EC 339**

---

Marcio Santetti

Spring 2023

Motivation

# A critique

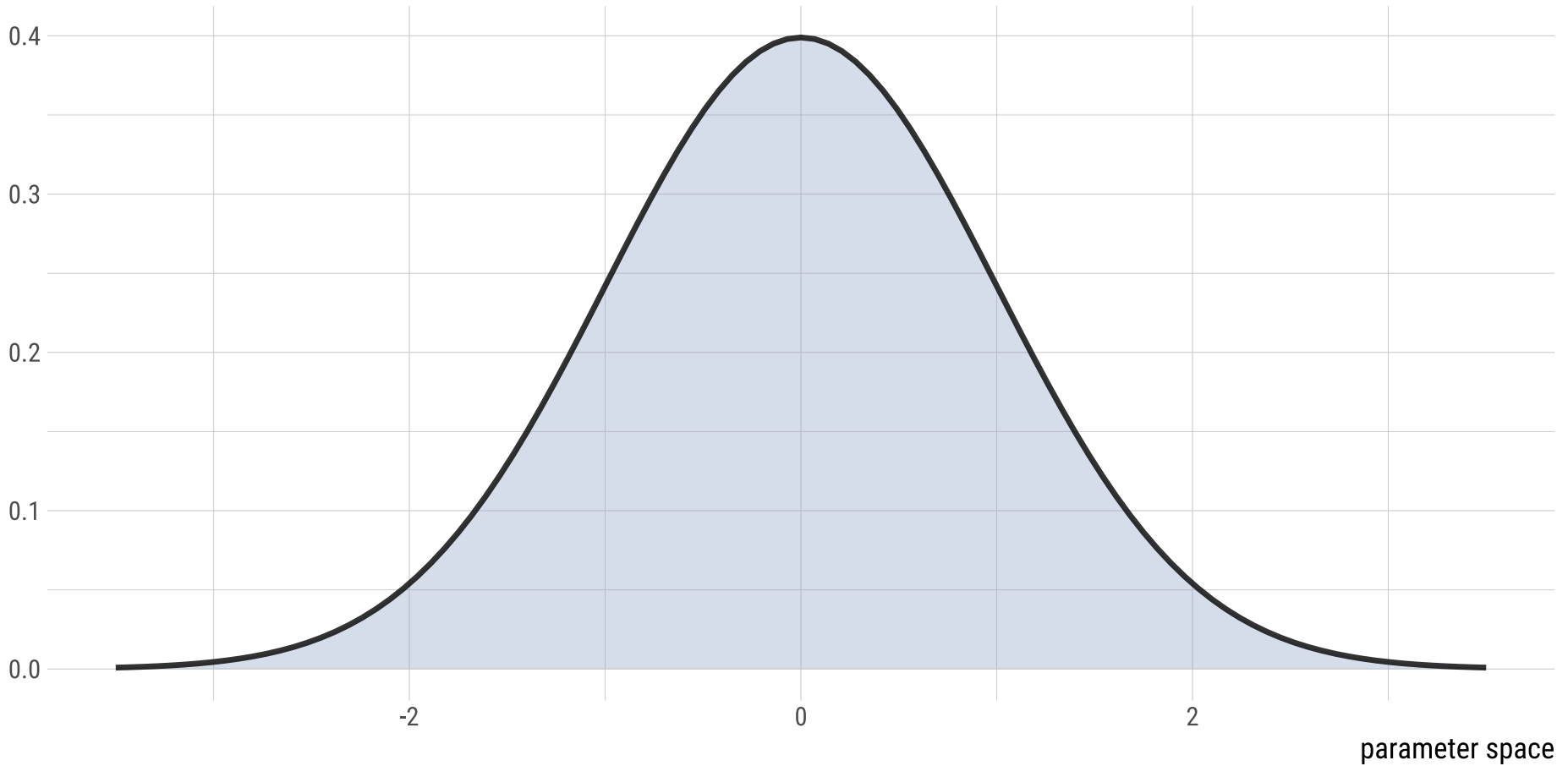
Here, we are dealing with the so-called **frequentist** approach to Statistics/Econometrics.

It assumes that there exists an underlying **true population parameter** in nature.

Therefore, while this **population parameter** value is fixed in nature, **samples** are variable.

And **using samples** is the best we can do.

# Where does this come from?



# Confidence Intervals

# Confidence intervals

In practical terms, a regression returns a **point estimate** of our desired parameter(s).

Supposedly, it **represents**, to the best of our efforts, the "true" population parameter.

But wouldn't it be better if we could have a **range** of values for  $\beta_i$ ?

Given a **confidence level**  $(1 - \alpha)$ , we can easily construct a **confidence interval** for  $\beta_i$ .

# Confidence intervals

From **Stats**, we know:

$$\text{CI} = \bar{x} \pm t_c \cdot \sigma$$

$$\text{CI} = \bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}}$$

And now:

$$\text{CI} = \hat{\beta}_k \pm t_c \cdot SE(\hat{\beta}_k)$$

where  $t_c = t_{1-\alpha/2, n-k-1}$ .

It denotes the  $1 - \alpha/2$  **quantile** of a  $t$  distribution, with  $n-k-1$  **degrees-of-freedom**.

# Confidence intervals

- The **standard error (SE)** of an estimate:

$$\text{SE}(\hat{\beta}_2) = \sqrt{\frac{s_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

where  $s_u^2 = \frac{\sum_i \hat{u}_i^2}{n - k - 1}$  is the variance of  $u_i$ .

The standard error of an estimate is nothing but its **standard deviation**.



# Confidence intervals

- **Informal interpretation:**

- The confidence interval is a region in which we are able to place some **trust** for containing the parameter of interest.

- **Formal interpretation:**

- With **repeated sampling** from the population, we can construct confidence intervals for each of these samples. Then  $(1 - \alpha) \cdot 100$  percent of our intervals (e.g., 95%) will contain the population parameter ***somewhere in this interval***.

# Confidence intervals - An example

```
#>
#> =====
#>                               Dependent variable:
#>                               -----
#>                               lsalary
#> -----
#> age                               -0.001
#>                               (0.005)
#> lsales                           0.225***
#>                               (0.028)
#> Constant                         5.005***
#>                               (0.303)
#> -----
#> Observations                      177
#> R2                               0.281
#> Adjusted R2                      0.273
#> Residual Std. Error      0.517 (df = 174)
#> F Statistic              34.004*** (df = 2; 174)
#> =====
#> Note:                *p<0.1; **p<0.05; ***p<0.01
```

# Confidence intervals - An example

From the previous regression output, we have:

- $\hat{\beta}_{lsales_i}$ : 0.225
- $SE(\hat{\beta}_{lsales_i})$ : 0.0277

In addition, the sample size ( $n$ ) is 177.

# Confidence intervals - An example

- Then, we can calculate a 95% confidence interval for  $\beta_{lsales_i}$ :

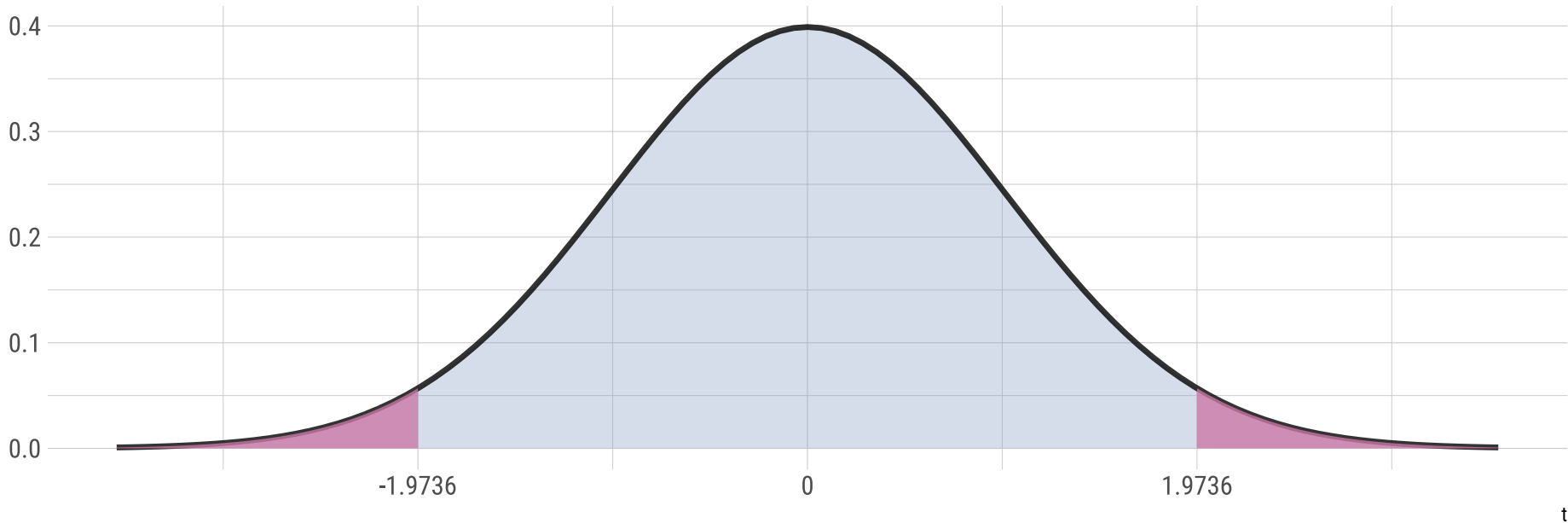
$$CI = \hat{\beta}_{lsales_i} \pm t_c \cdot SE(\hat{\beta}_{lsales_i})$$

$$CI = 0.225 \pm t_{1-0.05/2, 177-2-1} \cdot 0.0277$$

$$CI = 0.225 \pm t_{1-0.05/2, 174} \cdot 0.0277$$

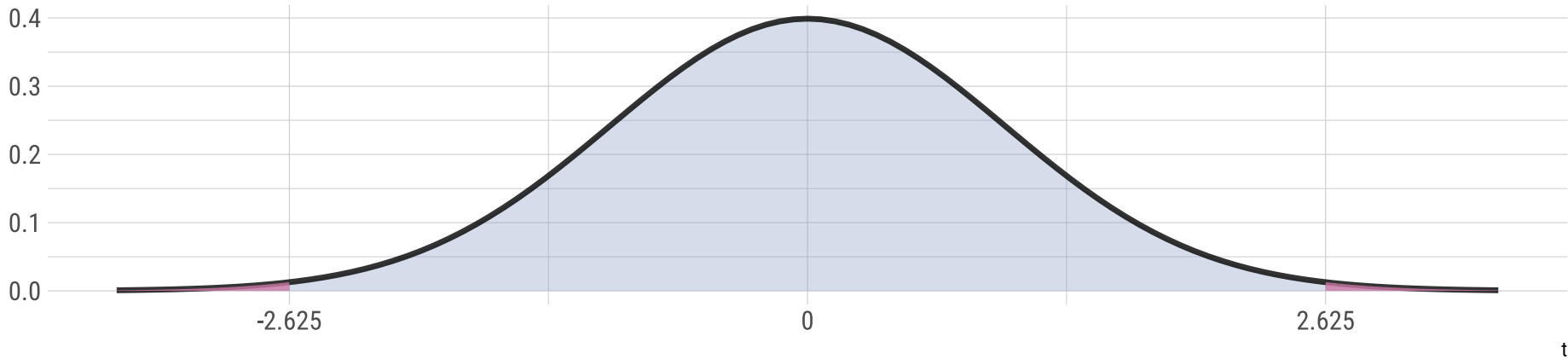
- $t_{1-0.05/2, 174} = -1.973691$
- The interval is  $[0.17, 0.28]$ .

# Confidence intervals - An example



With **repeated sampling** from the population, 95% of our intervals will contain the population parameter ***somewhere in this [0.17, 0.28] interval.***

# Confidence intervals - An example



- If we estimate a 99% confidence interval, we have:

$$CI = 0.225 \pm t_{1-0.01/2, 174} \cdot 0.0277$$

- $t_{1-0.01/2, 174} = 2.604379$
- The interval is  $[0.15, 0.29]$ .

# Hypothesis Testing

# Hypothesis testing

- When doing *hypothesis testing*, our aim is to determine whether there is enough **statistical evidence** to reject a hypothesized value or range of values.
- In Econometrics, we usually run **two-sided (tailed)** tests about *regression parameters*.
  - $H_0 : \beta_i = 0$
  - $H_a : \beta_i \neq 0$
- The above testing procedure is a test of **statistical significance**.
  - If we **do not reject**  $H_0$ , the coefficient is not statistically significant.
  - If we **reject**  $H_0$ , we have enough evidence to support the coefficient's statistical significance.



# Hypothesis testing

In Stata...

```
. reg wage educ exper tenure
```

Source	SS	df	MS	Number of obs	=	935
Model	22278193.8	3	7426064.59	F(3, 931)	=	53.00
Residual	130437974	931	140105.236	Prob > F	=	0.0000
				R-squared	=	0.1459
				Adj R-squared	=	0.1431
Total	152716168	934	163507.675	Root MSE	=	374.31

wage	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
educ	74.41486	6.286993	11.84	0.000	62.07654	86.75318
exper	14.89164	3.25292	4.58	0.000	8.507732	21.27554
tenure	8.256811	2.497628	3.31	0.001	3.355178	13.15844
_cons	-276.2405	106.7018	-2.59	0.010	-485.6444	-66.83653

# Hypothesis testing

- Where does the 11.8  $t$  value come from?

$$t = \frac{\hat{\beta}_k - \beta_{H_0}}{SE(\hat{\beta}_k)} = \frac{74.4 - 0}{6.29} = 11.8283$$

- Where does the 4.58  $t$  value come from?

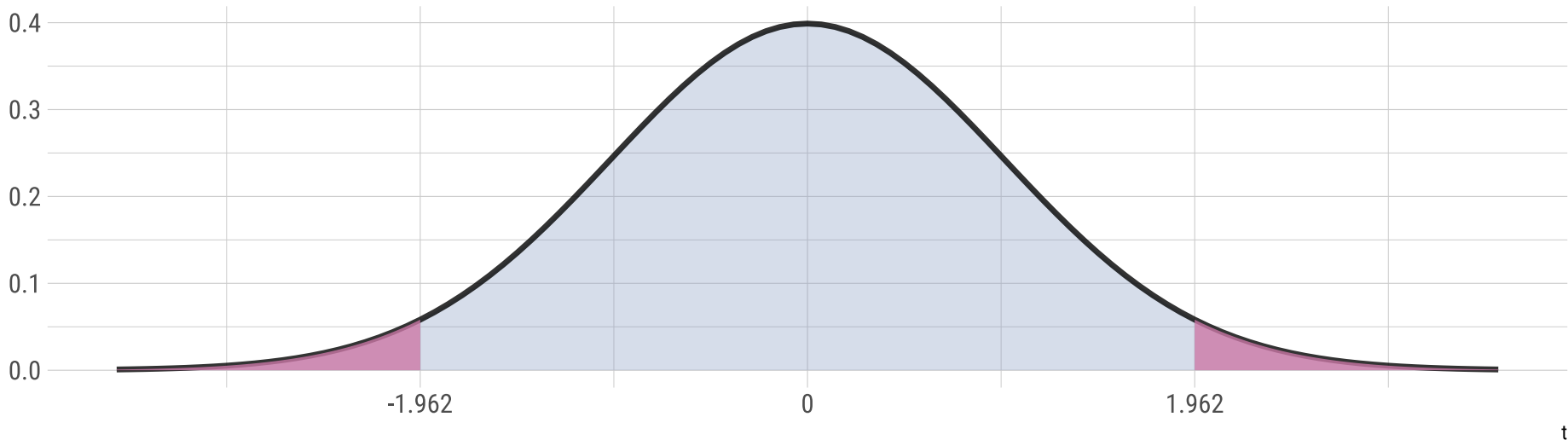
$$t = \frac{\hat{\beta}_k - \beta_{H_0}}{SE(\hat{\beta}_k)} = \frac{14.9 - 0}{3.25} = 4.584615$$

# Hypothesis testing

What are we supposed to do with these test statistics?

- $t_{\text{educ}} = 11.8$
- $t_{\text{exper}} = 4.58$
- $t_{\text{tenure}} = 3.31$

- $t_{\text{critical value}} = t_{.05/2, 931} = 1.962515$



# Hypothesis testing

## Interpretation

At 5% of significance, we have enough evidence to **reject the null hypothesis** that `educ` is not statistically significant.

At 5% of significance, we have enough evidence to **reject the null hypothesis** that `exper` is not statistically significant.

At 5% of significance, we have enough evidence to **reject the null hypothesis** that `tenure` is not statistically significant.

Therefore, all coefficients are (individually) **statistically significant**.

# Hypothesis testing

## The F-test

Sometimes, a coefficient on a **specific variable** may not be *statistically significant*.

However, it may be of use in the **model's context**.

Thus, a test of **joint** significance is appropriate to evaluate whether **all slope coefficients** are *jointly* significant within the model.

$$F = \frac{R_{\text{unr}}^2 - R_{\text{rest}}^2}{1 - R_{\text{unr}}^2} \cdot \frac{(n - k - 1)}{q}$$

# The F-test

Still with our **wage** model:

Suppose we want to test whether `educ` and `exper` are **jointly** significant.

For the purpose of this test, our previous model is the **unrestricted** (full) model.

Then, we estimate a **restricted** model, excluding `educ` and `exper`.

- Its R-squared is **0.0165**; while the unrestricted's is **0.146**.

We have imposed **2** restrictions to the full model. Thus,  $q=2$ .

And the **sample size** is  $n=935$ , which gives  $n-k-1 = 931$  for the full model.

# The F-test

$$F = \frac{R_{\text{unr}}^2 - R_{\text{rest}}^2}{1 - R_{\text{unr}}^2} \cdot \frac{(n - k - 1)}{q}$$
$$= \frac{0.146 - 0.0165}{1 - 0.146} \cdot \frac{935 - 3 - 1}{2} = 70.588$$

- 70.588 is the **test statistic** for the F-test
- Then, we compare the above value with the **critical values** given by the F-distribution table.
- Right-tail critical value:
  - $F_{1-.05/2, 2, 931} = 3.703535$
  - Thus, we **reject the null hypothesis**, meaning that we have enough evidence to infer that `educ` and `exper` are **jointly significant** in this model.

Next time: Inference in practice