Confidence Intervals in Linear Regression

Renato Assunção
ESRI and DCC/UFMG

Back to the compressive strength of concrete example

- The first two highlighted columns have the OLS estimates $\hat{\beta}_j$ and their estimated standard deviations $\sqrt{\mathbb{V}(\hat{\beta}_j)}$.
- In this set of slides, we will learn about the last two highlighted columns, the extremes of 95% confidence intervals.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: Statsmodel output for linear regression analysis

The stochastic view of the coefficients

- Assume the linear regression model is the data-generating mechanism.
- There exists an unknown coefficient vector $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)'$.
- Our goal is to learn it from the observed data, and we use a linear regression model to do so.
- Our LS estimate $\hat{\boldsymbol{\beta}}$ is a random vector, and it is different from the true and unknown $\boldsymbol{\beta}^*$. But we hope they are close, meaning the estimation error is small.
- The estimation error is $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}$.

Distribution of the estimator

- It is somewhat surprising that, although we do not know what is the exact value of the estimation error we have in this example, we know how much we can typically expect for this estimation error.
- This typical value is obtained by considering conceptually all other samples that could be generated by the model (same fixed $\boldsymbol{\beta}^*$) and the different estimates $\hat{\boldsymbol{\beta}}$ we would get for each of those samples.
- Why?
- Because we obtained the distribution of the random vector:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta}^*, \sigma^2(X^T X)^{-1}\right).$$

One coordinate at a time

- For one coordinate, $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, \sigma^2[(X^T X)^{-1}]_{jj}\right)$.
- For the cement coefficient, the output table gives that its standard deviation is 0.008.
- Each coordinate oscillates as a Gaussian around its true unknown value.

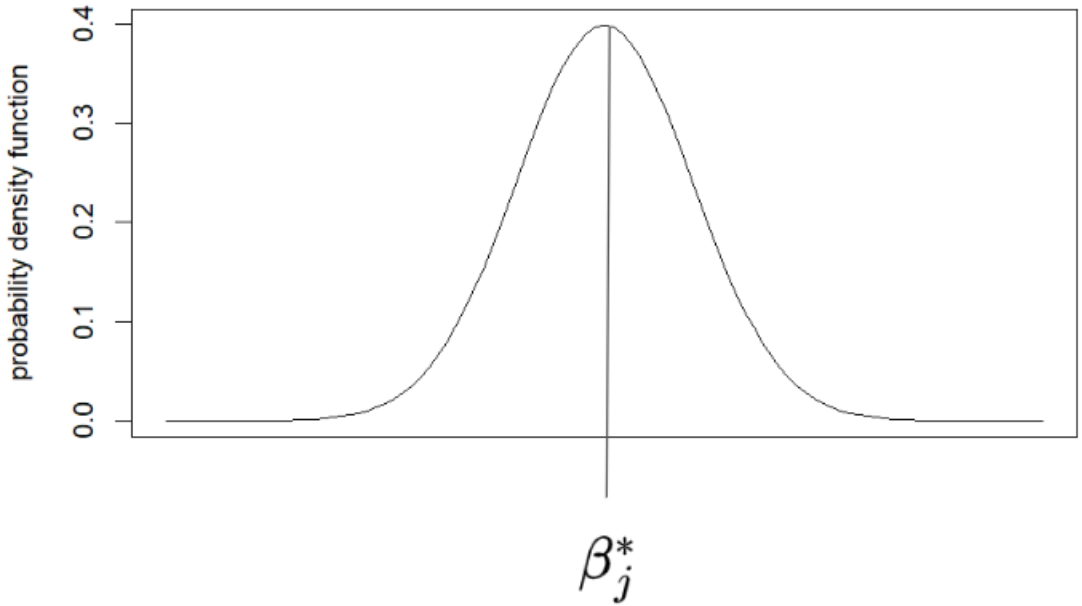

Figure: Distribution of $\hat{\beta}_j$. The center β_j^* is the true and unknown value.

The observed (instantiated) $\hat{\beta}_j$

- For the specific sample we used to fit the model, we get a specific value.
- For the cement coefficient, the output table gives $\hat{\beta}_j = 0.1198$.
- This value can be assigned to a specific position on the real axis.

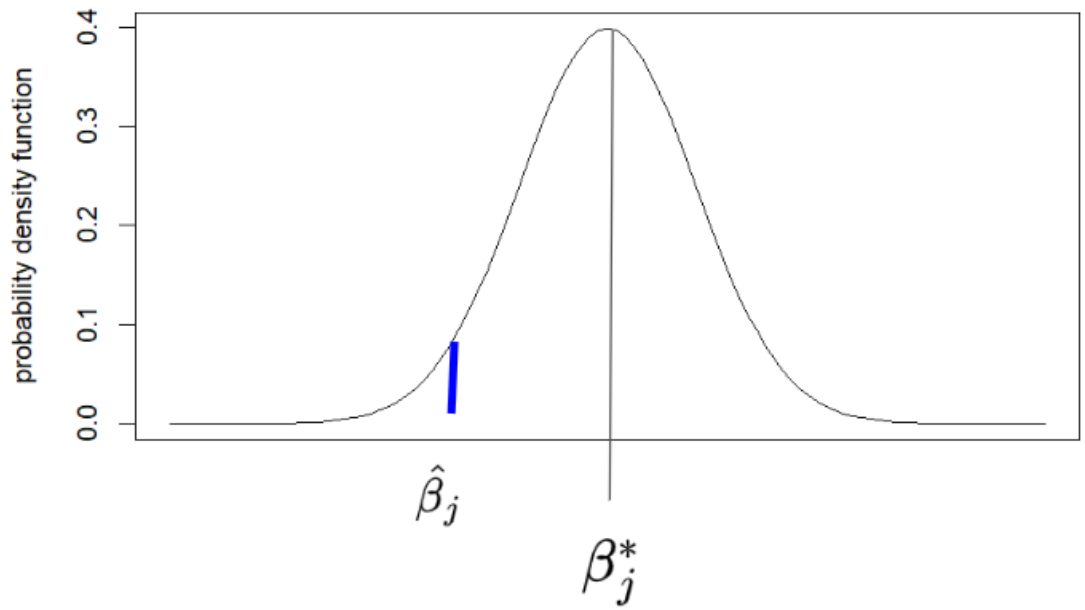

Figure: Distribution of $\hat{\beta}_j$ and its observed $\hat{\beta}_j$

What we REALLY have

- The previous figure is purely conceptual and could mislead by giving the impression that we know more than we really do.
- What we really know (for the cement coefficient):
  - The observed (realized, instantiated) value for the estimate: we had $\hat{\beta}_j = 0.1198$ with this sample.
  - We also know the probability distribution of this random variable that we have just known its realized value: $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, (0.008)^2\right)$.
- What we do not know: the true value β_j^* and therefor, the center of the Gaussian distribution

Figure: The observed value of $\hat{\beta}_j$. The distribution is known but not its center β_j^*

What we REALLY have

- In fact, writing that (for the cement coefficient): $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, (0.008)^2\right)$ is not completely correct.
- The variance 0.008^2 is an estimate of the true one that is unknown because we do not know σ^2 exactly.
- As we saw previously, σ^2 is estimated using the residuals. This estimate of σ^2 is used on the formula for the variance in $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, \sigma^2[(X^T X)^{-1}]_{jj}\right)$.
- So we know neither β_j^* nor the variance $\sigma^2[(X^T X)^{-1}]_{jj}$.
- We have only an approximation for this variance based on the residuals.
- However, we will assume for now that σ^2 and the variance of $\hat{\beta}_j$ can be calculated exactly and it is equal to 0.008^2.
- We return to this uncertainty about this variance and fix this problem after explaining the basic idea of a confidence interval.

What we REALLY know and don't know

- This next figure is a more realistic description of what we have.
- We know the observed value of $\hat{\beta}_j$ in the sample (for cement, it is 0.1198) but we do not know which of these Gaussians is the true one because we do not know the value of β_j^*, the center of the Gaussian distribution.

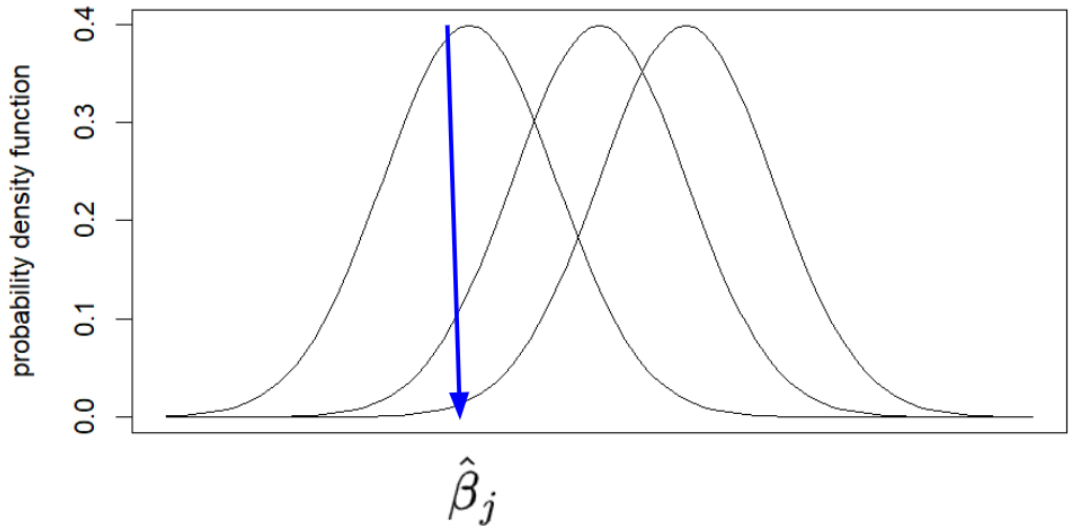

Figure: The observed $\hat{\beta}_j$ and some of the possible distributions it may follow.

We know quite a lot about Gaussians

- Consider the standard Gaussian $\mathcal{N}(0, 1)$.
- We know that random variables following this distribution hardly fall outside the interval $(-2, 2)$.
- Indeed, $\mathbb{P}(Z \in (-2, 2)) = 0.9544997$ (or very close to 95%).

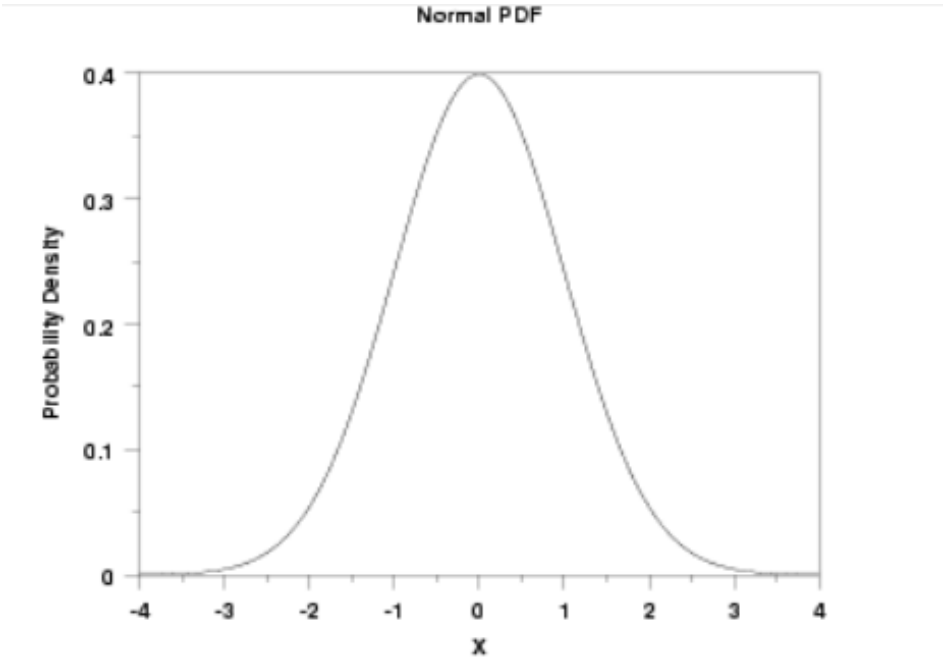

Figure: Density of the standard Gaussian $\mathcal{N}(0, 1)$.

Standardizing a generic Gaussian

- Any generic Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ can be transformed into a standard Gaussian $\mathcal{N}(0, 1)$.
- Define the random variable

$$Z = \frac{X - \mu}{\sigma}$$

- Then, as we know from FECD-A, the random variable Z has a $\mathcal{N}(0, 1)$ distribution.
- We can apply this to the random coefficient estimate.
- We know that $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j^*, \sigma^2[(X^T X)^{-1}]_{jj}\right) = \mathcal{N}\left(\beta_j^*, (0.008)^2\right)$.
- Then

$$Z = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\sigma^2(X'X)^{-1}[jj]}} = \frac{\hat{\beta}_j - \beta_j^*}{0.008} \sim N(0, 1)$$

Aha!

- Given that $Z = \frac{\hat{\beta}_j - \beta_j^*}{0.008} \sim N(0,1)$, we know that Z is hardly outside the interval $(-2, 2)$.
- With approximately 95% chance it is inside $(-2, 2)$.
- Hence, we say that with high probability (high confidence) Z is in $(-2, 2)$.
- So what?
- It seems a worth knowledge because we can not calculate Z: its numerator $\hat{\beta}_j - \beta_j^*$ can not be calculated because β_j^* is unknown.
- However, we can do a lot here.

Aha!

- Let us simplify the denominator denoting it simply by v:

$$Z = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\sigma^2(X'X)^{-1}[jj]}} = \frac{\hat{\beta}_j - \beta_j^*}{v} \sim N(0,1)$$

- With approximately 95% chance (or 0.95 probability) Z is between -2 and $+2$. That is, with high probability

$$-2 \leq \frac{\hat{\beta}_j - \beta_j^*}{v} \leq 2 \quad \text{or} \quad -2v \leq (\hat{\beta}_j - \beta_j^*) \leq 2v$$

- We can manipulate algebraically this expression moving any of its central terms to the extremes.

Aha!

- For example, add $-\beta_j^*$ in each side of the inequality:

$$-\hat{\beta}_j - 2v \leq -\hat{\beta}_j + (\hat{\beta}_j - \beta_j^*) \leq -\hat{\beta}_j + 2v$$

- Now, multiply by (-1) (remember that you need to change the inequality signs)

$$(-1) * (-\hat{\beta}_j - 2v) \geq \quad (-1) * (-\beta_j^*) \quad \geq (-1) * (-\hat{\beta}_j + 2v)$$

- That is,

$$\hat{\beta}_j + 2v \geq \beta_j^* geq \hat{\beta}_j - 2v$$

- I prefer the natural order to the numbers in the real axis:

$$\hat{\beta}_j - 2v \leq \beta_j^* \leq \hat{\beta}_j + 2v$$

- The two extremes in this expression represent a confidence interval:

$$IC = [\hat{\beta}_j - 2v, \hat{\beta}_j + 2v]$$

- Let us step back to understand how to interpret this confidence interval.

The ICs in the output

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.3312	26.586	-0.878	0.380	-75.500	28.837
cement	0.1198	0.008	14.113	0.000	0.103	0.136
slag	0.1039	0.010	10.247	0.000	0.084	0.124
flyash	0.0879	0.013	6.988	0.000	0.063	0.113
water	-0.1499	0.040	-3.731	0.000	-0.229	-0.071
superplasticizer	0.2922	0.093	3.128	0.002	0.109	0.476
coarseaggregate	0.0181	0.009	1.926	0.054	-0.000	0.037
fineaggregate	0.0202	0.011	1.887	0.059	-0.001	0.041
age	0.1142	0.005	21.046	0.000	0.104	0.125

Figure: The confidence interval $IC = [\hat{\beta}_j - 2v, \hat{\beta}_j + 2v]$ of each coefficient in the concrete dataset

IC interpretation

- Because $Z = (\hat{\beta}_j - \beta_j^*)/v \sim N(0,1)$,
- we known that

$$\mathbb{P}\left(-2 \leq \frac{\hat{\beta}_j - \beta_j^*}{v} \leq 2\right) \approx 0.95$$

- Manipulating algebrically the event inside the parenthesis we obtain a mathematically equivalent expression:

$$\mathbb{P}\left(\hat{\beta}_j - 2v \leq \beta_j^* \leq \hat{\beta}_j + 2v\right) \approx 0.95$$

or

$$\mathbb{P}\left(\beta_j^* \in (\hat{\beta}_j - 2v, \hat{\beta}_j + 2v)\right) \approx 0.95$$

Aha!

- Look closely to this last expression:

$$\mathbb{P}\left(\beta_j^* \in (\hat{\beta}_j - 2v, \hat{\beta}_j + 2v)\right) \approx 0.95$$

- What is random here?
- Not β_j^*: this is an unknown but a fixed value.
- The random random component is $\hat{\beta}_j$.
- What is random is the interval extremes.
- A simulation will make this more clear

Aha!

- I took the compressive strength of concrete dataset and fitted the linear regression model.
- Then, I took the coefficient estimates as if they were the true coefficients, used the $\hat{\sigma}$ estimate as the true value of σ and then simulated many datasets generating new $\boldsymbol{Y}$ vectors by simulating Gaussian errors $\mathcal{N}(0, \hat{sigma}^2)$.
- In these simulations the matrix $\boldsymbol{X}$ were kept fixed.
- After generating a new $\boldsymbol{Y}$ vector, I fitted the linear regression model, got the estimated coefficient and built the

$$IC = [\hat{\beta}_j - 2v, \hat{\beta}_j + 2v]$$

- The result is in the next slide.

Aha!

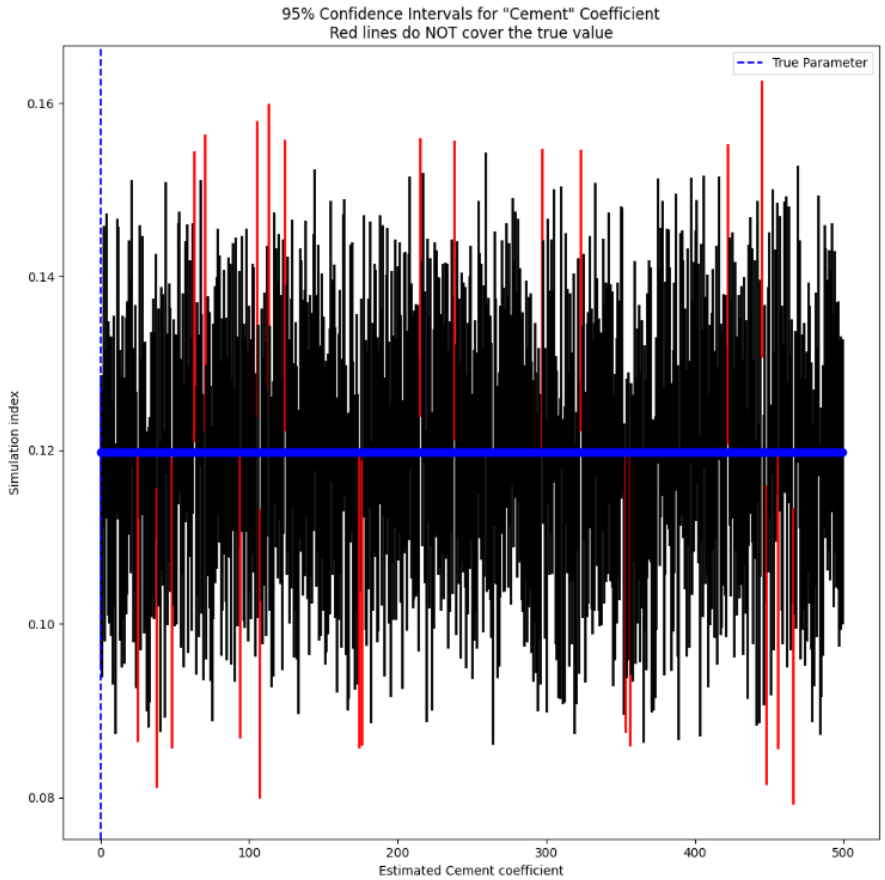

Figure: Confidence intervals for the Cement coefficient

Aha!

```python
# Step 3: Fit initial OLS model to get "true" parameters and residual std deviation
model_true = sm.OLS(Y, X).fit()
beta_hat = model_true.params           # True coefficients (Series with names)
sigma_hat = model_true.resid.std(ddof=X.shape[1])  # Estimated sigma (residual std dev

# Step 4: Monte Carlo simulation
n_simulations = 500
coeff_matrix = []

for _ in range(n_simulations):
    noise = np.random.normal(0, sigma_hat, size=len(Y))
    Y_sim = X @ beta_hat + noise
    model_sim = sm.OLS(Y_sim, X).fit()
    coeff_matrix.append(model_sim.params.values)

coeff_matrix = np.array(coeff_matrix)

# Step 5: Plot histograms for each coefficient
feature_names = X.columns
n_params = len(feature_names)

fig, axes = plt.subplots(3, 3, figsize=(18, 12))
axes = axes.flatten()

for i in range(n_params):
    ax = axes[i]
    ax.hist(coeff_matrix[:, i], bins=20, color='lightgray', edgecolor='black')
    ax.axvline(beta_hat[i], color='blue', linestyle='dashed', linewidth=2)
    ax.set_title(f"Sampling Distribution of: {feature_names[i]}")
    ax.set_xlabel("Estimated Value")
    ax.set_ylabel("Frequency")

plt.tight_layout()
plt.show()
```

Figure: Code for the previous figure

Aha!

- In practice, we have only one single IC that may contain the true unknown value of β_j^*.
- We never know if the interval shown in the output table really covers the true coeffcient value β_j^*.
- However, we know that the intervals were built with an algorithm that covers these true and unknown β_j^* values 95% of the times the algorithm is used.
- So, the confidence is attached to the *method* being used to build the ICs: it covers the true coefficient almost always (95% of the times we use it).

The constant "2"

- We defined the 95% confidence interval as

$$IC = [\hat{\beta}_j - 2v, \hat{\beta}_j + 2v]$$

- The constant "2" appears because, in the standard Gaussian distribution, approximately 95% of the density area is between -2 and $+2$.
- If we want to have exactly 95% we should use the constant 1.96 rather than "2".
- However, in most applications, the different ICs will be negligible and the choice of either constant will not have a noticeable impact.
- Also, we can build ICs with different confidence levels. For example, a 99% confidence interval requires the constant 2,32.
- This constant comes from $\mathbb{P}(-2.32 \leq Z \leq 2.32) = 0.99$ when $Z \sim \mathcal{N}(0, 1)$.

Another issue

- We defined the 95% confidence interval as

$$IC = [\hat{\beta}_j - 2v, \hat{\beta}_j + 2v]$$

- where $v^2 = \sigma^2(X'X)^{-1}[jj]$.
- However, σ^2 is also unknown.
- We solved this by substituting σ^2 in v^2 by its unbiased estimate

$$S^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- This has a consequence.
- The Gaussian distribution used to determine the constant "2" should be substituted by the t-Student distribution.

The t-Student distribtion

- The t-Student distribution is similar to the $\mathcal{N}(0,1)$ distribution.
- The main difference are the heavy tails of the t-Student
- As we gets farther away from zero, the t-Student density decreases to zero more slowly than a $\mathcal{N}(0,1)$ density.
- Rather than "2" for the 95% IC, the constant should be larger.
- The exact value depends of the degrees of freedom of the residual vector which is equal to $n-(p+1)$.

The 95% constant from the t-Student

- As the degrees of freedom increases, the constant converges to the standard normal 1.96 constant.
- In practice, with $df = 30$, the constant is very close to that of the standard Gaussian.

df	5	10	20	30	100	200	300	∞
constant	2.52	2.23	2.09	2.04	1.98	1.97	1.97	1.96

Table: Value for the constant in the 95% IC using the t-Student distribution