

Evaluating Significance of Predictors

Hypothesis Testing



CS109A Introduction to Data Science
Pavlos Protopapas, Natesh Pillai and Chris Gumb

Outline

Part A and B: Assessing the Accuracy of the Coefficient Estimates

Bootstrapping and confidence intervals

Part C: Evaluating Significance of Predictors

Does the outcome depend on the predictors?

Hypothesis testing

Part D: How well do we know \hat{f}

The confidence intervals of \hat{f}

How Reliable are the Model Interpretations

Suppose our model for advertising is:

$$y = 1.01x + 0.005$$

where y is the sales in units (each unit is \$1000) and x is the TV budget.

Interpretation: For every dollar invested in advertising gets you 1.01 back in sales, which is 1% net increase.

But how **certain** are we in our estimation of the coefficient 1.01?

Now you know how **certain** you are in your estimates, will you want to change your answer?

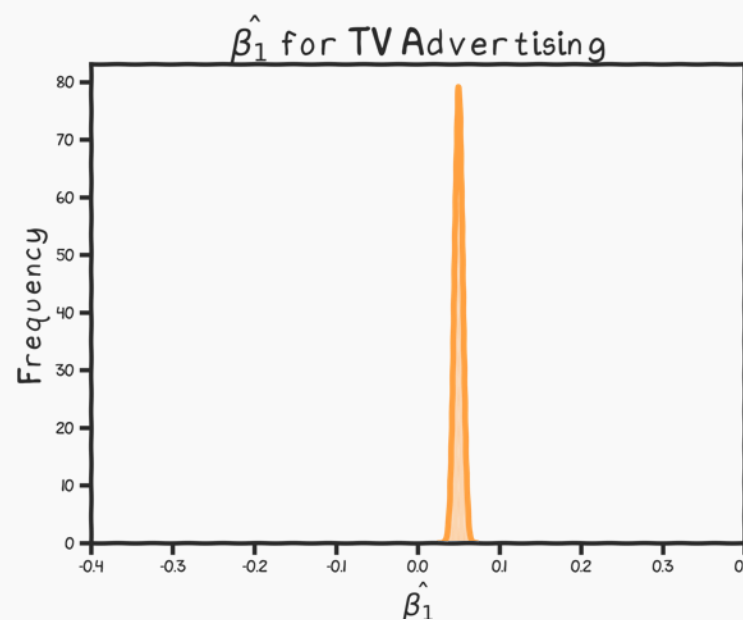
Feature Importance

Now we know how to generate these distributions we are ready to answer *two important questions*:

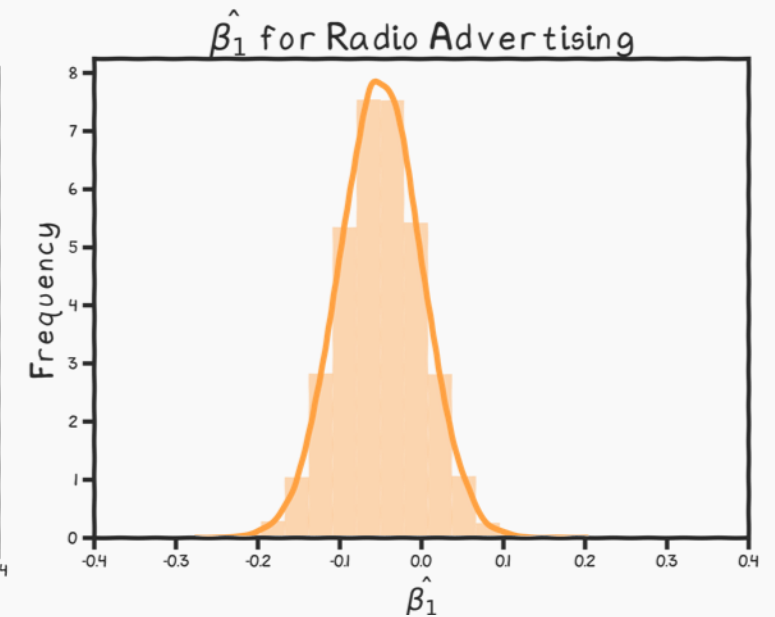
- A. Which predictors are most important?
- B. And which of them really affects the outcome?



$$\mu_{\beta_1} = 0.1$$
$$\sigma_{\beta_1} = 0.05$$



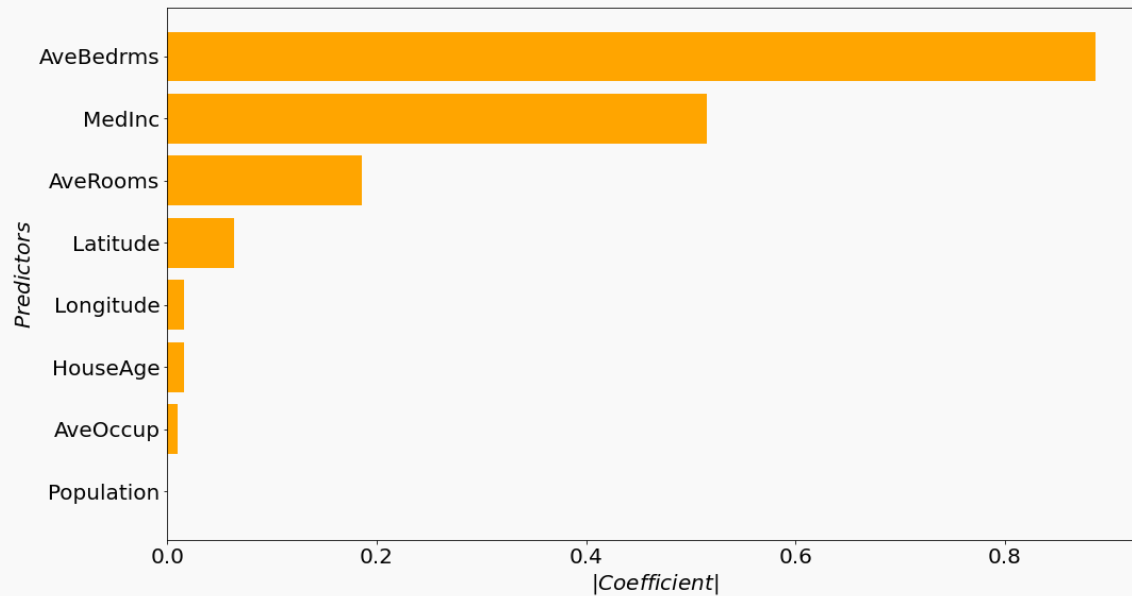
$$\mu_{\beta_1} = 0.05$$
$$\sigma_{\beta_1} = 0.005$$



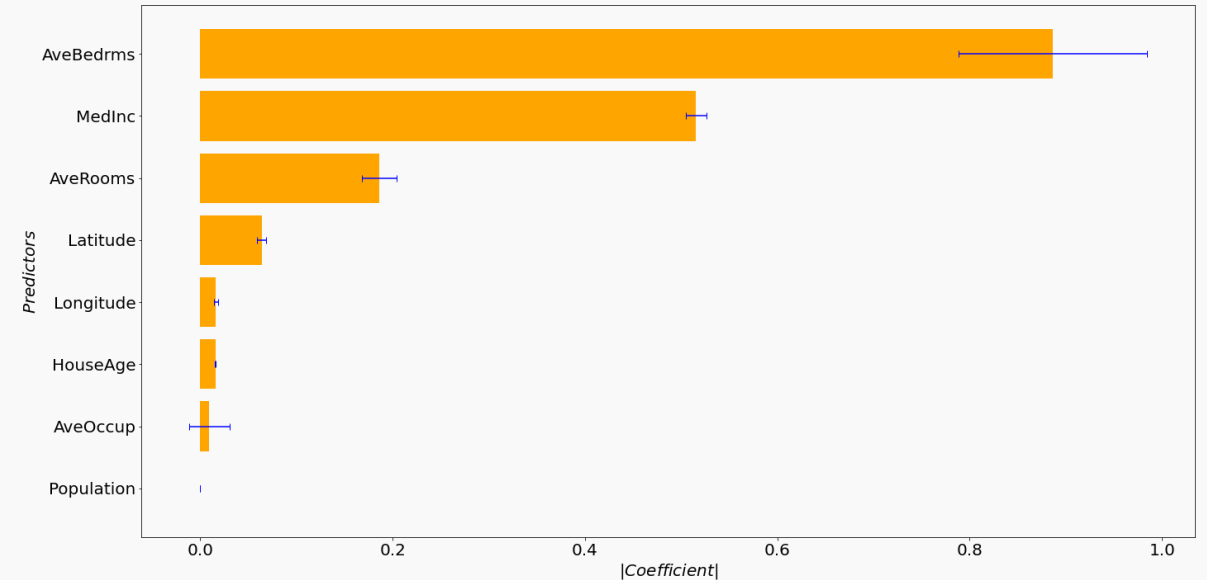
$$\mu_{\beta_1} = -0.05$$
$$\sigma_{\beta_1} = 0.1$$

Feature Importance

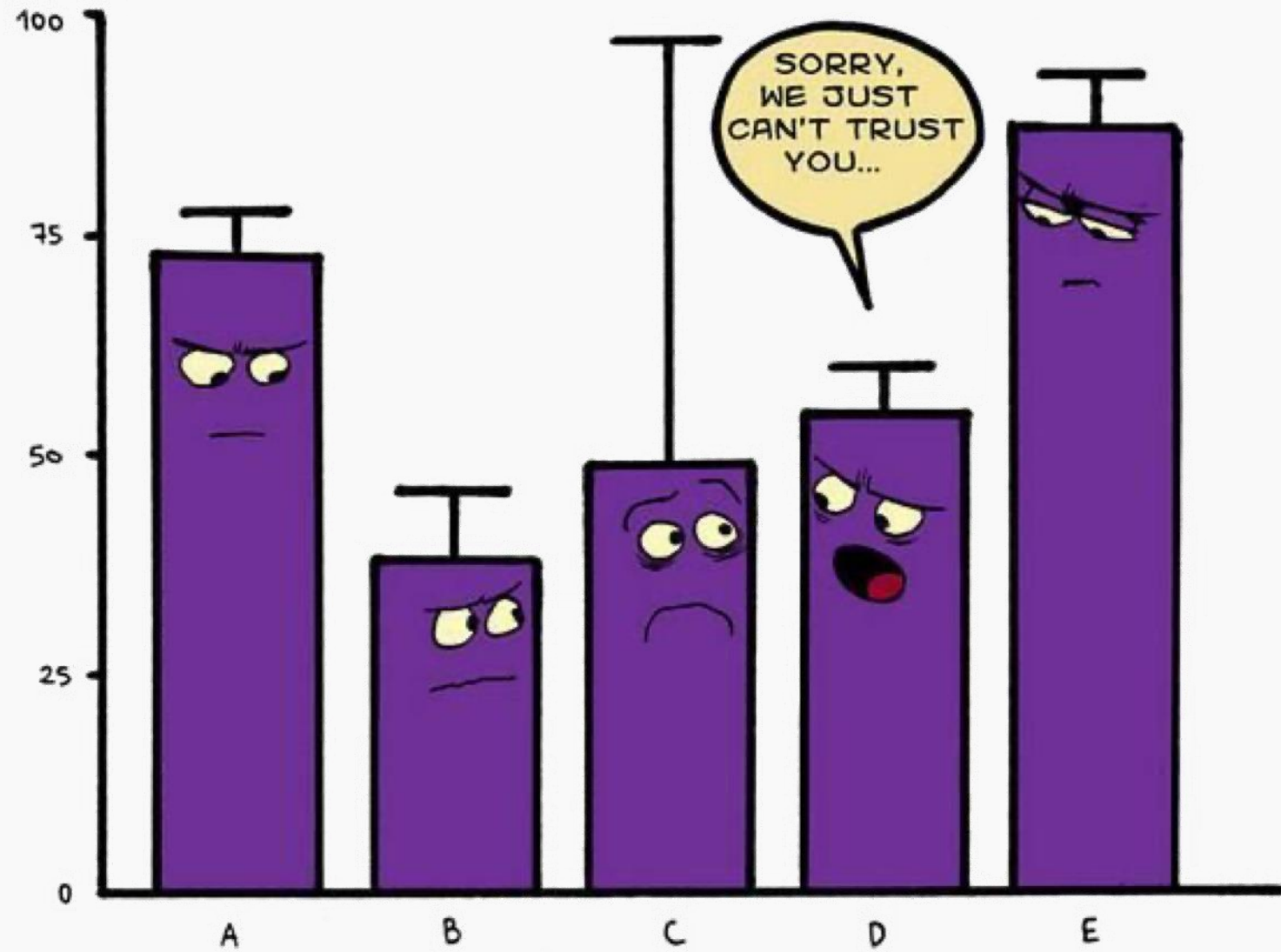
The example below is from [California housing data](#). The coefficients below are from a model that predicts prices given house size, age, crime, etc.



Feature importance based on the [absolute value](#) of the coefficients.



Feature importance based on the absolute mean value of the coefficients over multiple bootstraps including uncertainties.

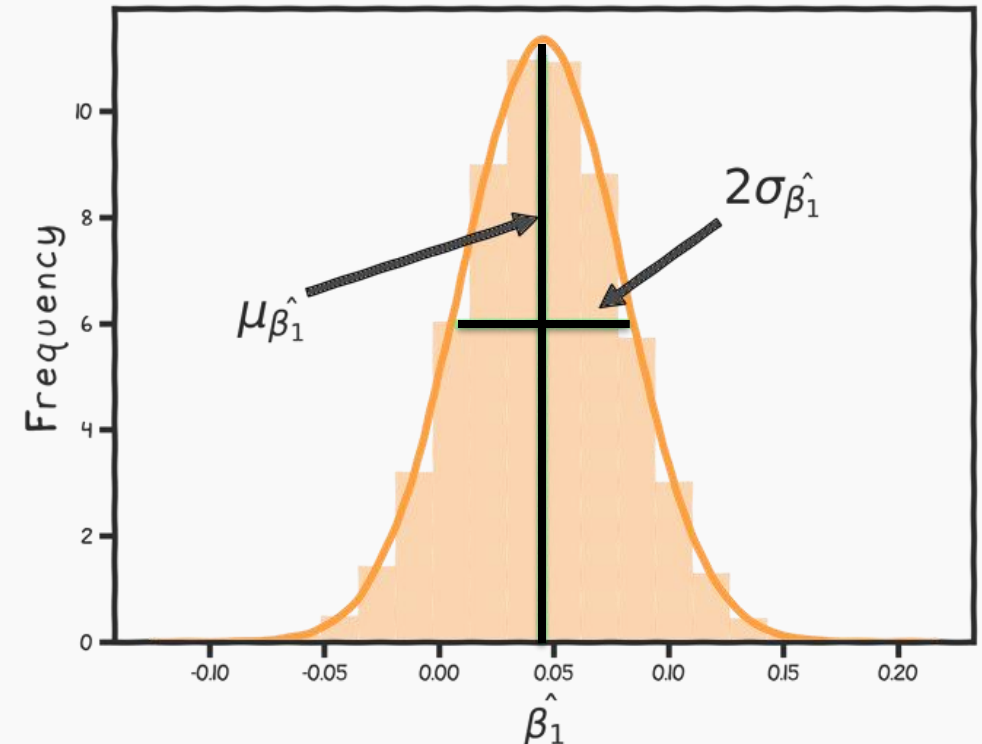


Feature Importance

To incorporate the coefficients' uncertainty, we need to determine whether the estimates of β 's are sufficiently **far from zero**.

To do so, we define a new **metric**, which we call \hat{t} – *test* statistic which measures the distance from zero in units of standard deviation:

$$\hat{t}\text{-test} = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$



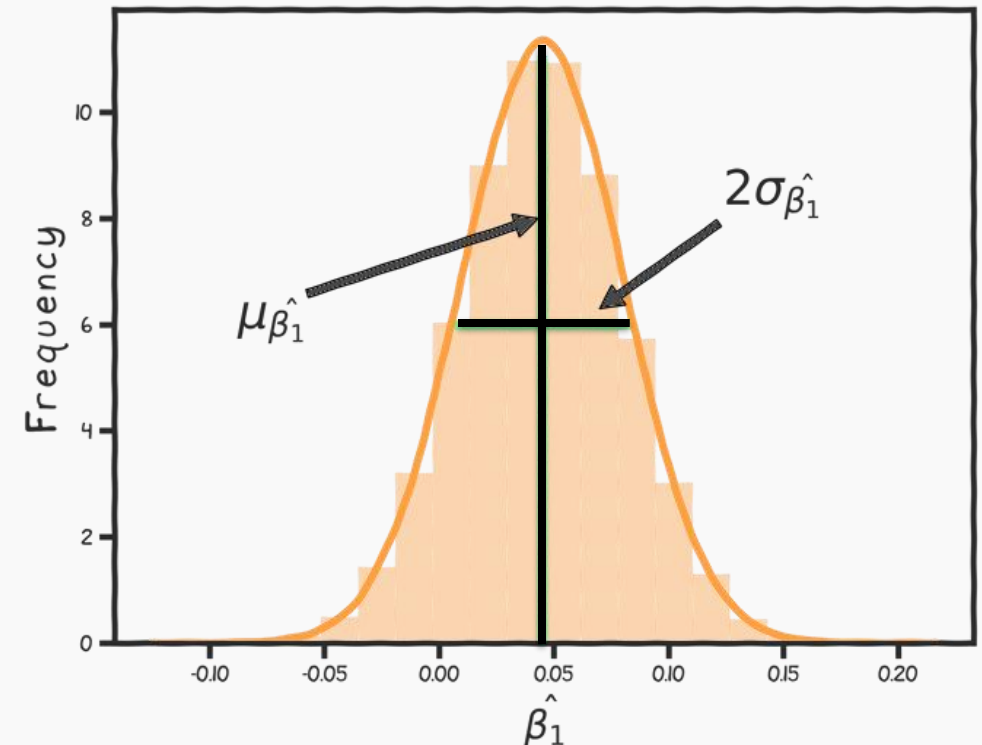
Feature Importance

$$\hat{t}\text{-test} = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

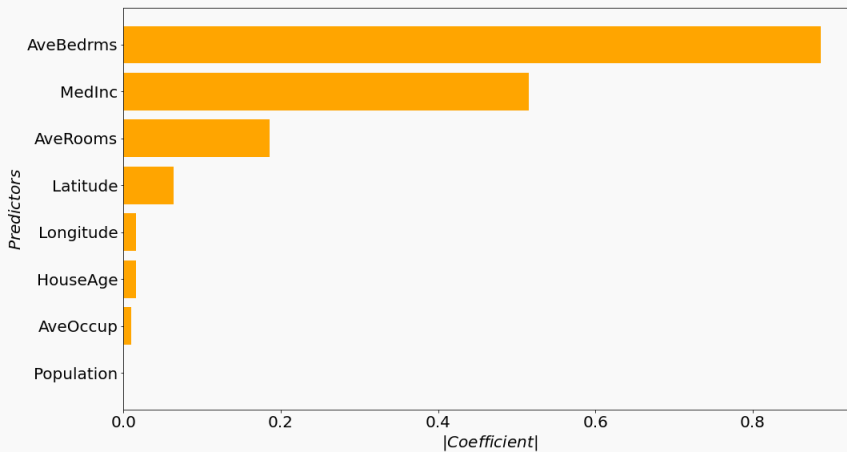
$\hat{t}\text{-test}$ is a scaled version of the usual t-test:

$$t\text{-test} = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}/\sqrt{n}} = \sqrt{n} \hat{t}\text{-test}$$

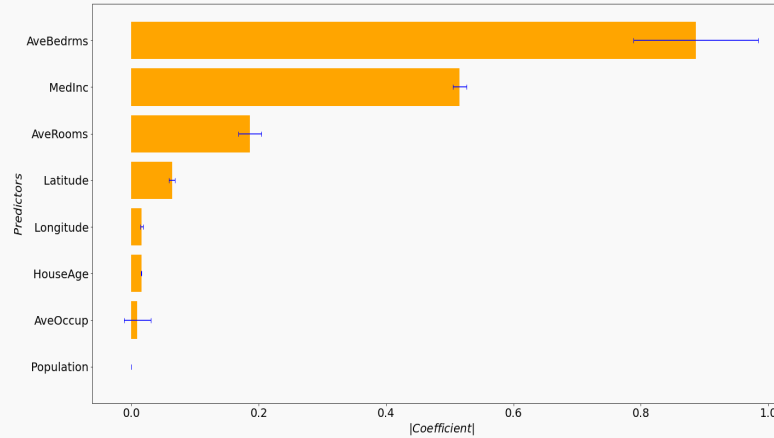
n is the number of bootstraps.



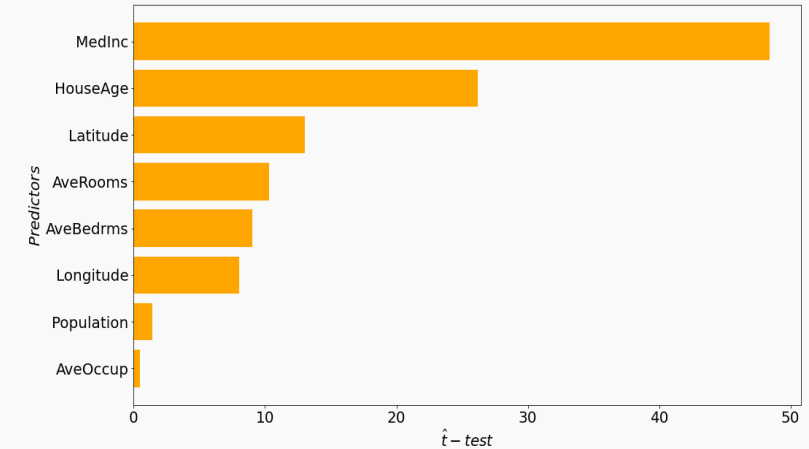
Feature Importance



The **absolute value** of the coefficients.



The absolute value of the coefficients over multiple **bootstraps** and includes the **uncertainty** of the coefficients.



The \hat{t} -test. Notice the rank of the importance has changed.

Feature Importance

Because a predictor is ranked as the most important, it does not necessarily mean that the **outcome depends on that predictor**.

How do we assess if there is a true relationship between outcome and predictors?

As with R^2 score, we should compare its significance ($\hat{t} - test$) to:

slido

Please download and install the Slido app on all computers you use



As with the R^2 score, we should compare its significance (using a \hat{t} -test) to:

① Start presenting to display the poll results on this slide.

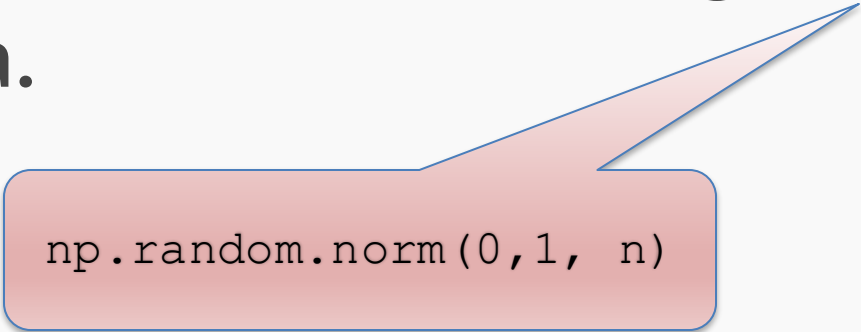
Feature Importance

What do we mean
random data?

We want to compare the $\hat{t} - test$ of the predictors from our model with $\hat{t} - test$ values calculated using random data.

Feature Importance

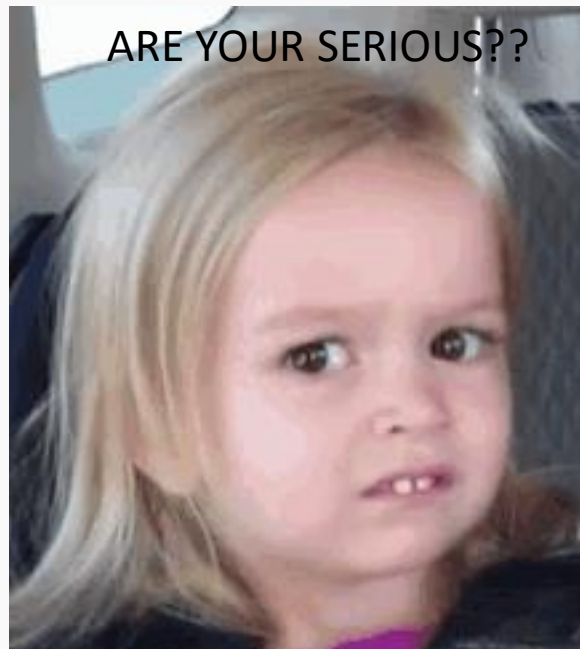
We want to compare the $\hat{t} - test$ of the predictors from our model with $\hat{t} - test$ values calculated using **random** data.



```
np.random.norm(0, 1, n)
```

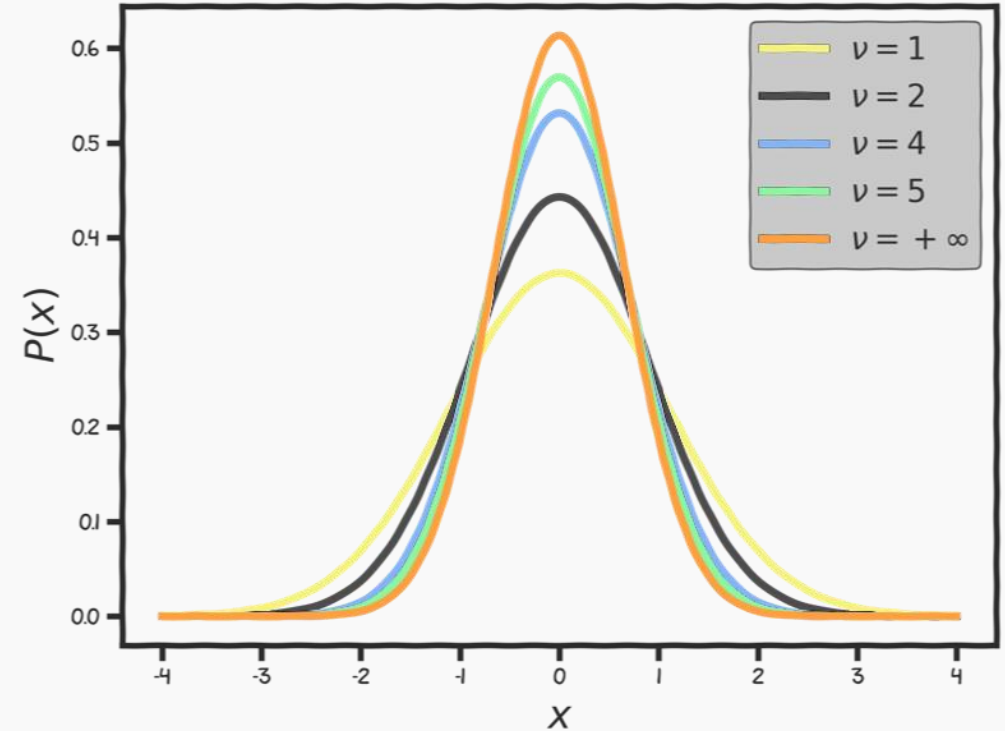
Feature Importance

1. For ***m*** random datasets fit ***m*** models.
2. Generate distributions for all predictors and calculate the means and standard errors ($\mu_{\hat{\beta}}, \widehat{SE}(\hat{\beta}_1)$).
3. Calculate the $\hat{t}_{test_random} = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$.
4. Repeat steps 1-3 and create a histogram for all the $\hat{t} - test - random$.



Feature Importance

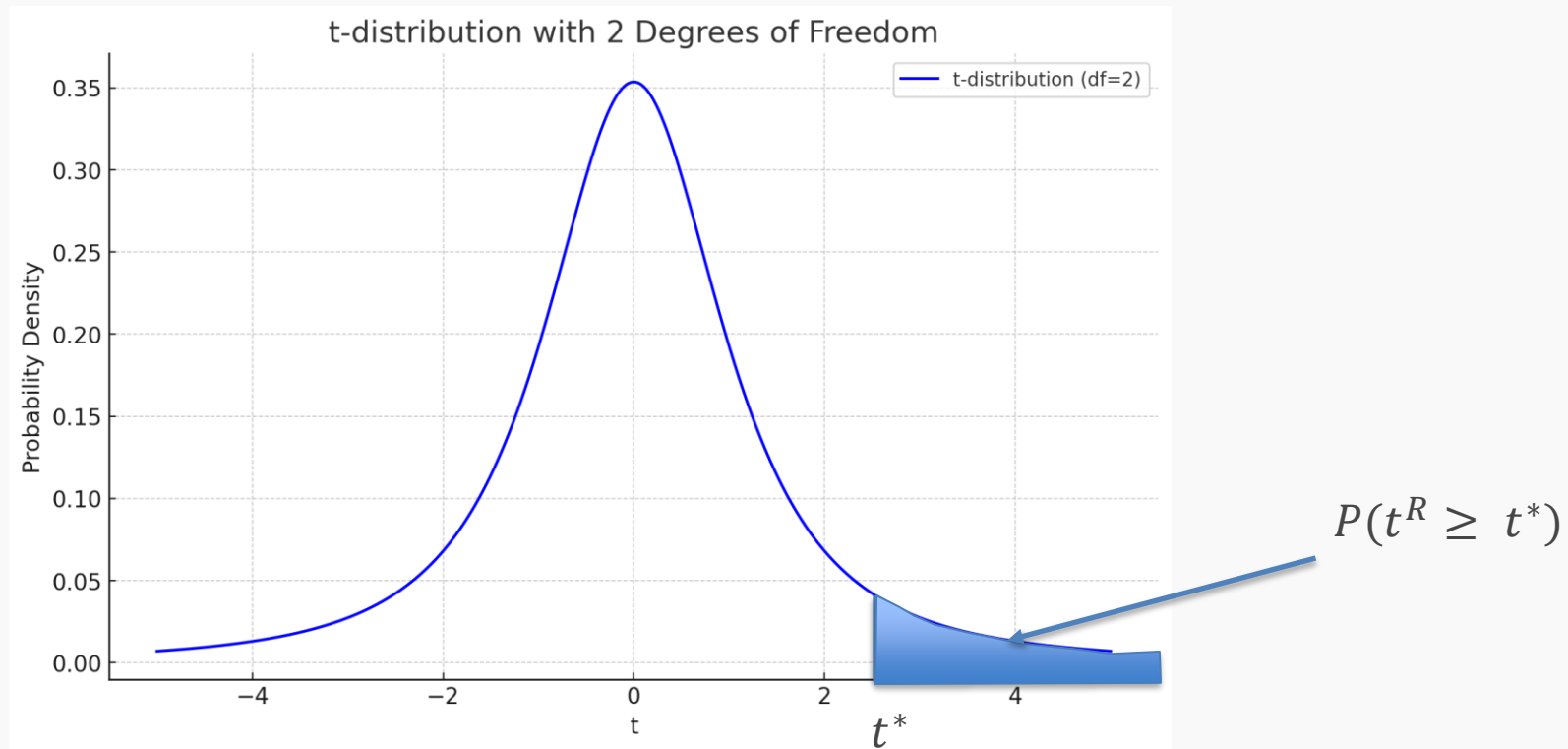
It turns out we do not have to do this, because this is a known distribution called student-t distribution.



Student-t distribution, where ν is the degrees of freedom (number of data points minus number of predictors) = $n - (p + 1)$.

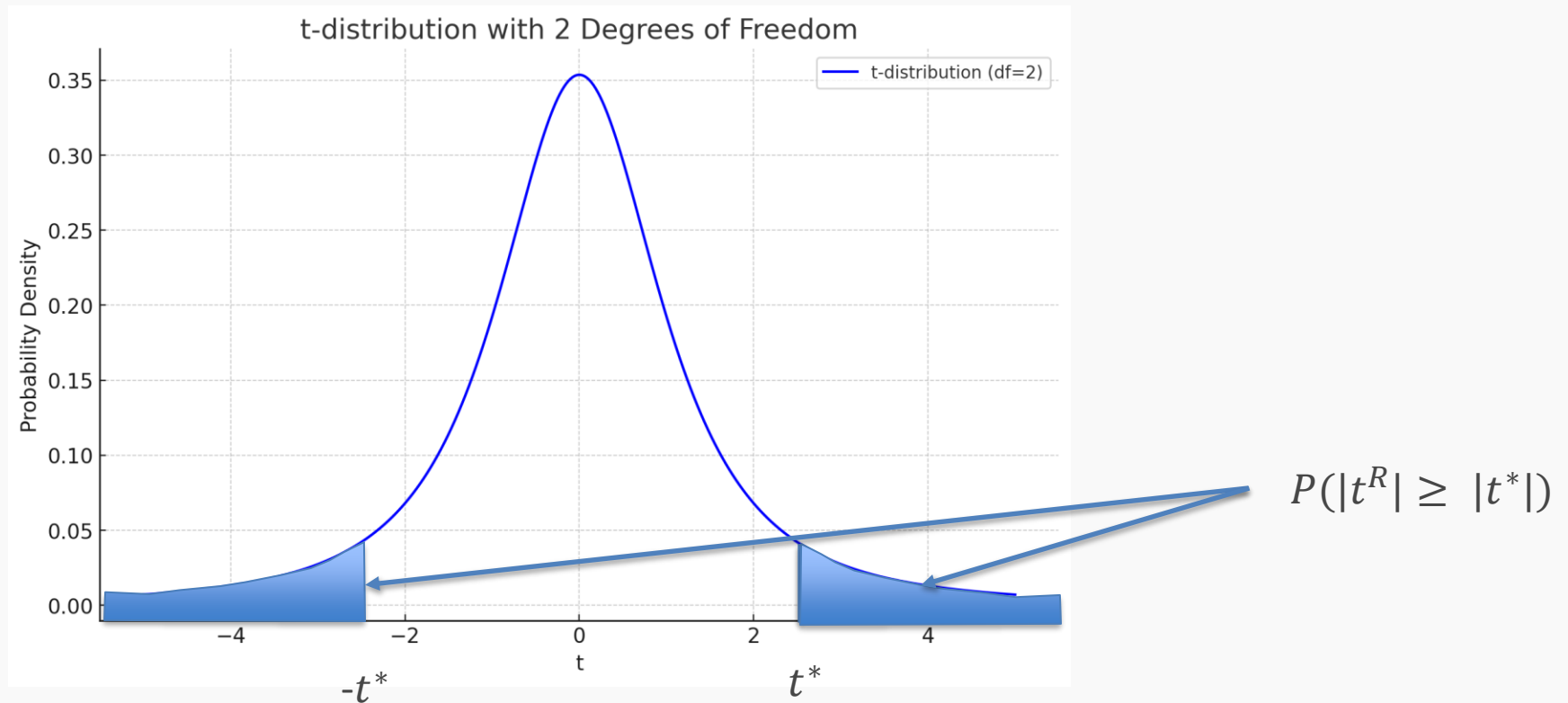
P-value

To compare the t -test values of the predictors from our model, t^* , with the t -tests calculated using random data, t^R , we estimate the probability of observing $t^R \geq t^*$.



P-value

Actually, we need compare $|t^*|$, with the t-tests calculated using random data, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t^*|$.



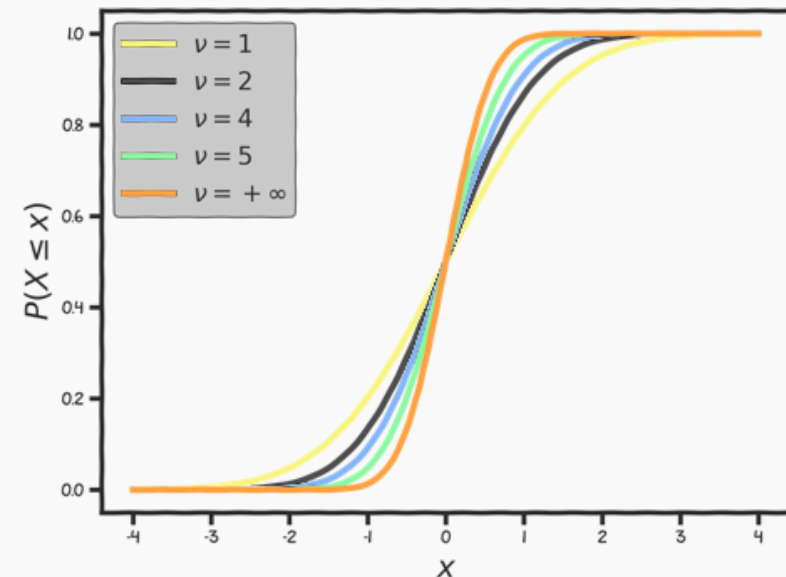
P-value

We call this probability the **p-value**:

$$p - value = P(|t^R| \geq |t^*|)$$

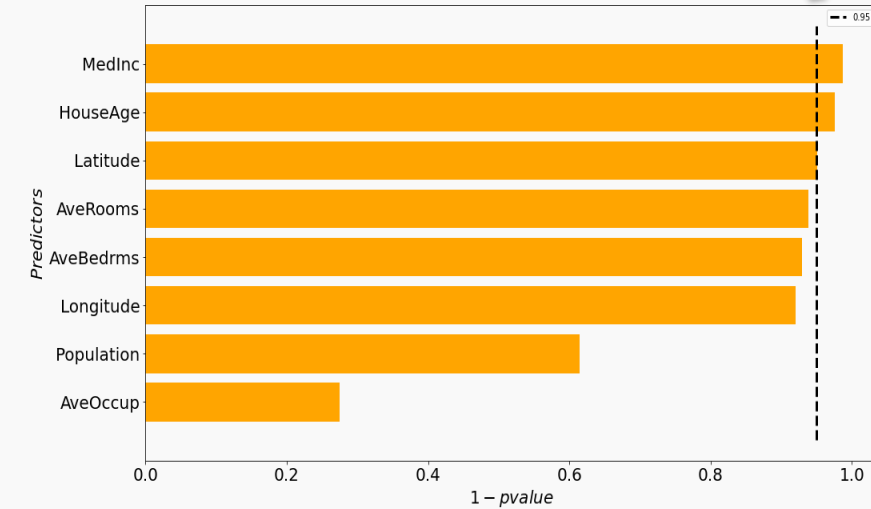
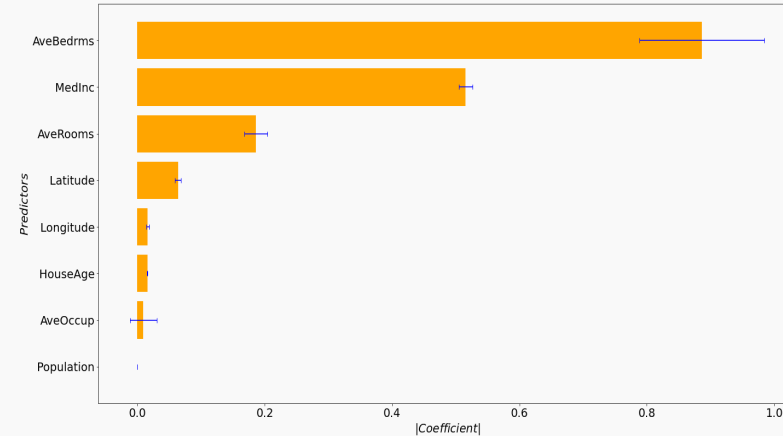
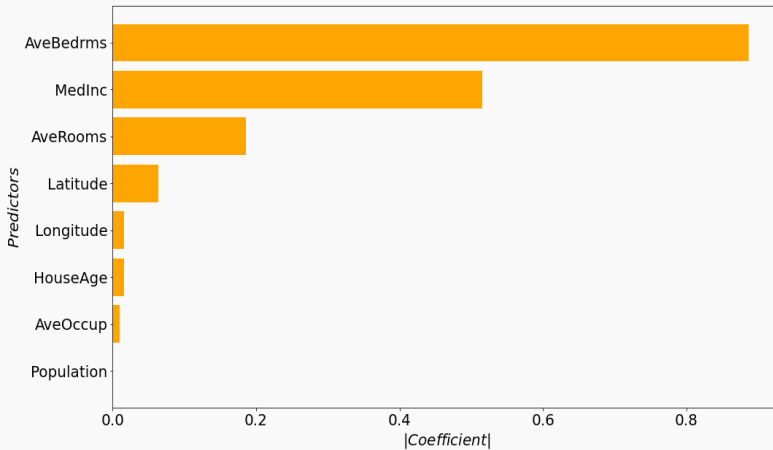
Small **p-value** indicates that it is **unlikely to observe such a substantial association** between the predictor and the response due to chance. It is common to use **p-value < 0.05** as the threshold for significance.

To calculate the p-value we use the cumulative distribution function (CDF) of the student-t. `stats` model a python library has a build-in function `stats.t.cdf()` which can be used to calculate this.



Feature Importance

Any predictor with a $1 - p$ value higher than this is considered important.



The absolute value of the coefficients over multiple **bootstraps** and includes the coefficients' **uncertainty**.

The \hat{t} -test. Notice the rank of the importance has changed.

Using the the **p-value** we also have which predictors are important. Note here we use $1 - p$.

Hypothesis Testing

Hypothesis testing is a formal process through which we evaluate the validity of a statistical hypothesis by considering evidence **for** or **against** the hypothesis gathered by **random sampling** of the data.

1. State the hypotheses, typically a **null hypothesis**, H_0 and an **alternative hypothesis**, H_1 , that is the negation of the former.
2. Choose a type of analysis, i.e. how to use sample data to evaluate the null hypothesis. Typically, this involves choosing a single test statistic.
3. **Sample** data and compute the test statistic.
4. Use the value of the test statistic to either **reject** or **not reject** the null hypothesis.

Hypothesis Testing

1. State Hypothesis:

Null hypothesis:

H_0 : There is no relation between X_j and Y , ($\beta_j = 0$).

The alternative:

H_a : There is some relation between X_j and Y , ($\beta_j \neq 0$).

2. Choose test statistics

$$\hat{t} - test = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$$

Hypothesis Testing

3. Sample:

Using bootstrap we can estimate $\hat{\beta}'_1$ s, and $\mu_{\hat{\beta}_1}$ and $\widehat{SE}(\hat{\beta}_1)$ and the \hat{t} – *test*.

4. Reject or not reject the hypothesis:

We compute ***p-value***, the probability of observing any value equal to $|t|$ or larger, from random data.

If **p-value < p-value-threshold** we **reject the null**.

Not Done Yet

Permutation Tests: a side note

Should you use a bootstrap approach to perform a hypothesis test?

While this is tempting, this is **not advisable**. Why?

It is a technical issue: the bootstrap approach is prone to inflating Type I error: you conclude there is an association when there really is not one.

In order to preserve the state Type I error (presumably at 5%), you should instead perform a permutation test: another resampling method.

In a permutation test, you resample the data assuming the null hypothesis is true. This can most easily be done by shuffling the response variable while keep the columns of the predictors as-is.