Evaluating Significance of Predictors
Hypothesis Testing

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Photo: Greg Mccutcheon
Mount Fuji, Japan

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 [not]

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 a 1% net increase.

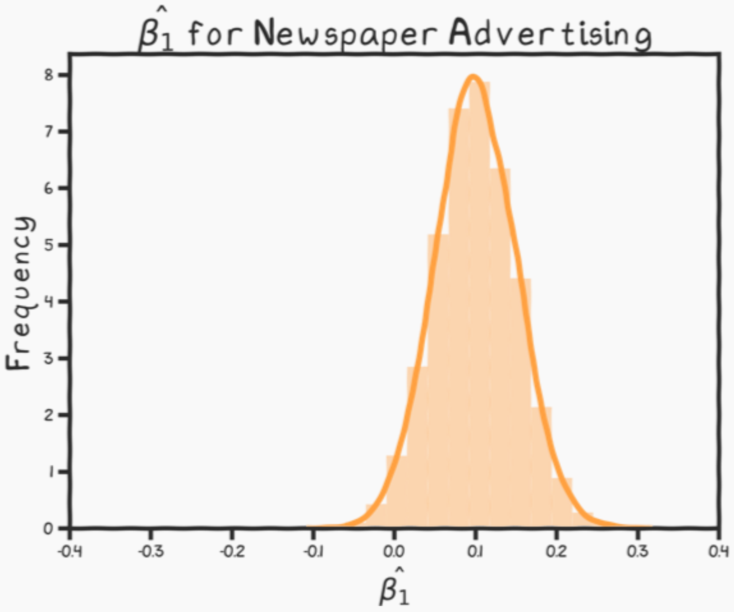
But how certain are we in our estimation of the coefficient 1.01?

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

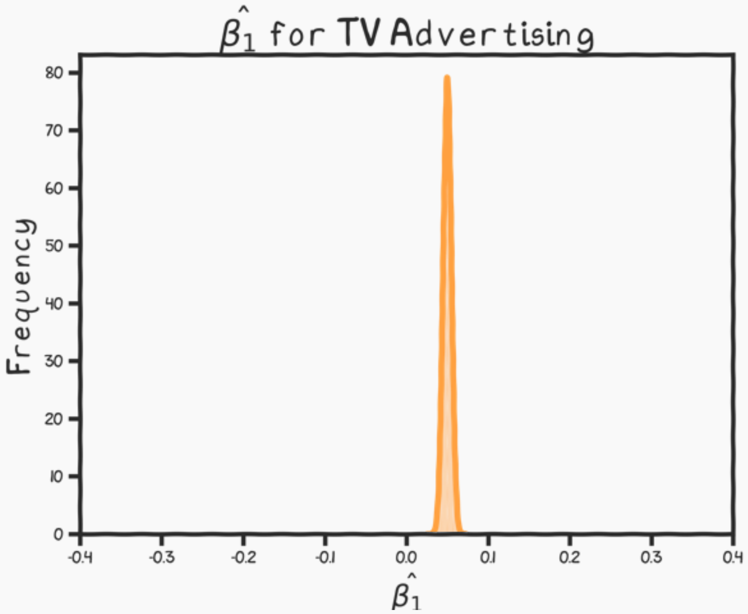
Feature Importance

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

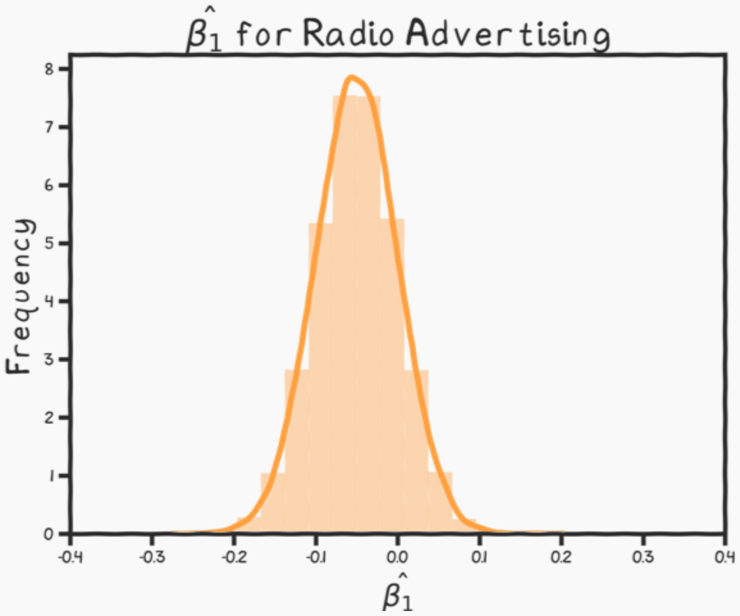
A. Which predictors are most important?
B. And which of them really affects the outcome?



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



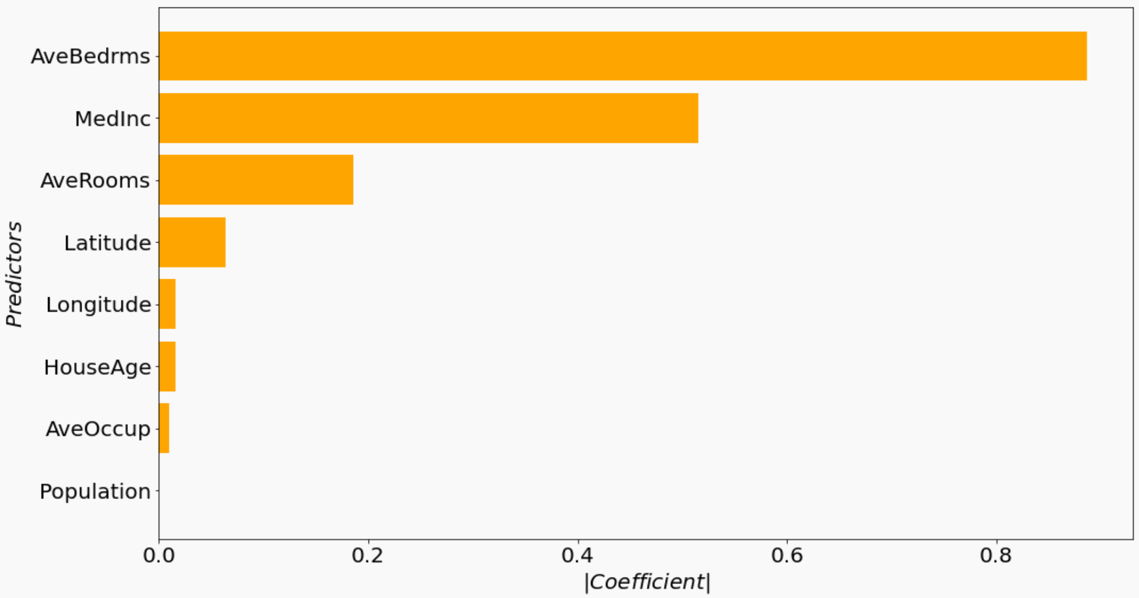
$\mu_{\beta_1} = 0.05$
$\sigma_{\beta_1} = 0.005$



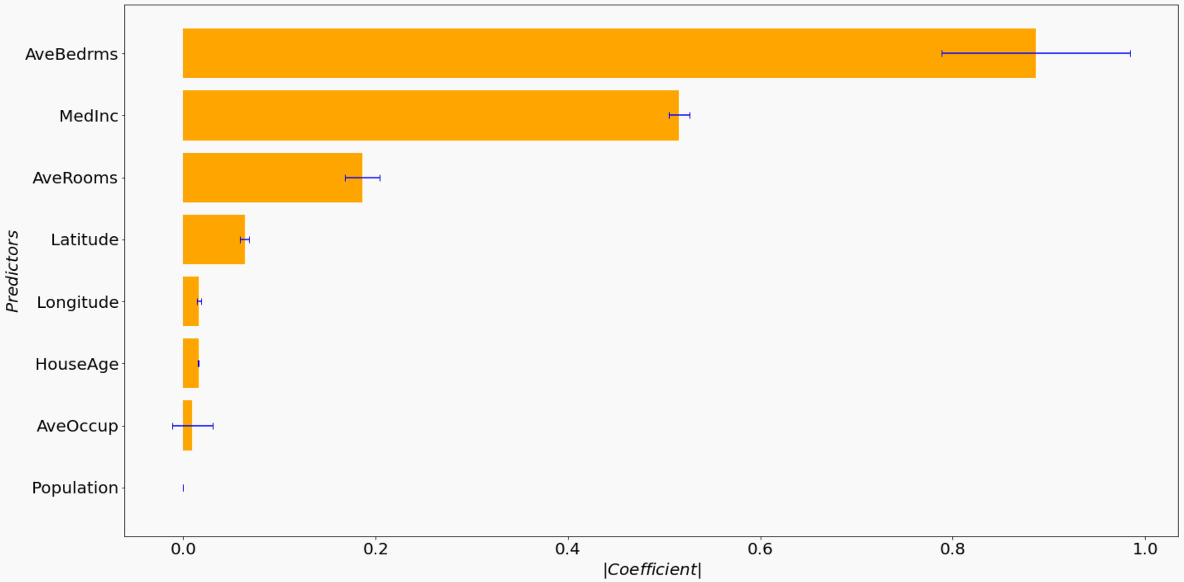
$\mu_{\beta_1} = -0.05$
$\sigma_{\beta_1} = 0.1$

Feature Importance

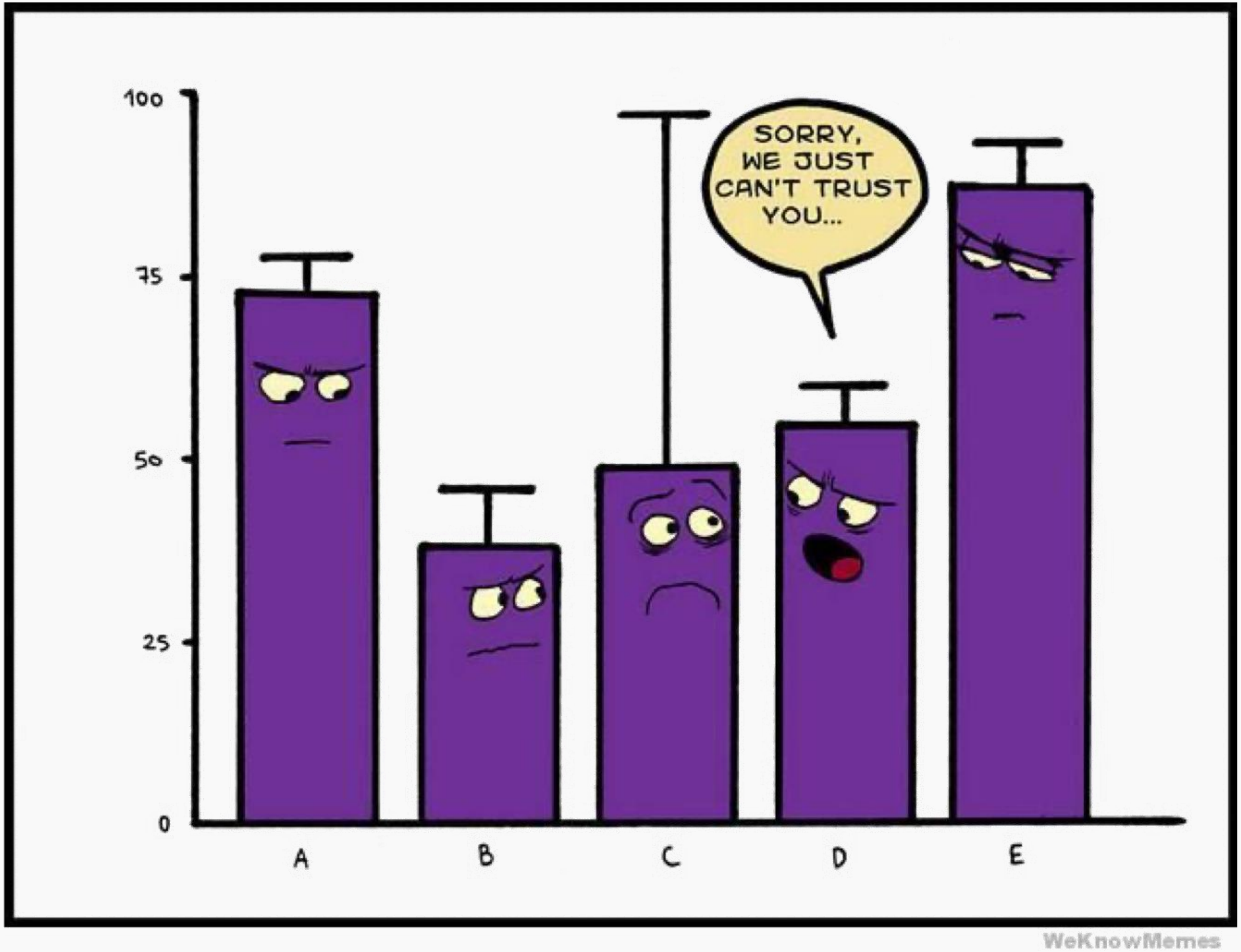
The example below is from California housing data. The coefficients below are from a model that predicts prices given house size, age, crime, etc.



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.

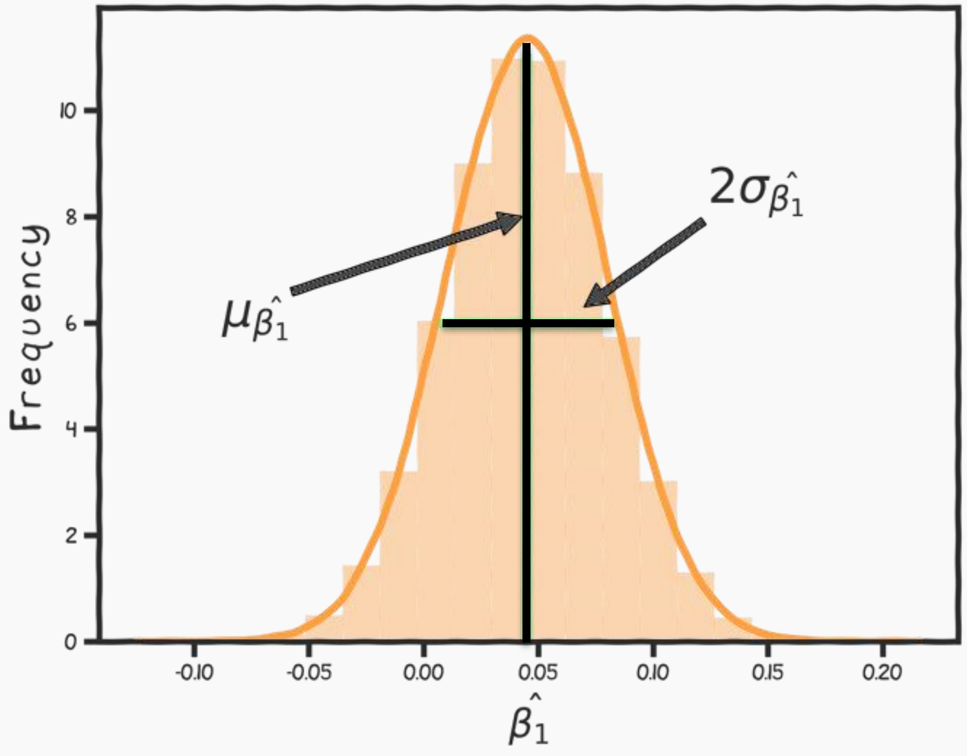

100
75
50
25
0
A
B
C
D
E
SORRY, WE JUST CAN'T TRUST YOU...
WeKnowMemes

Feature Importance

To incorporate the coefficients' uncertainty, we need to determine whether the estimates of $\beta's$ 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:

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



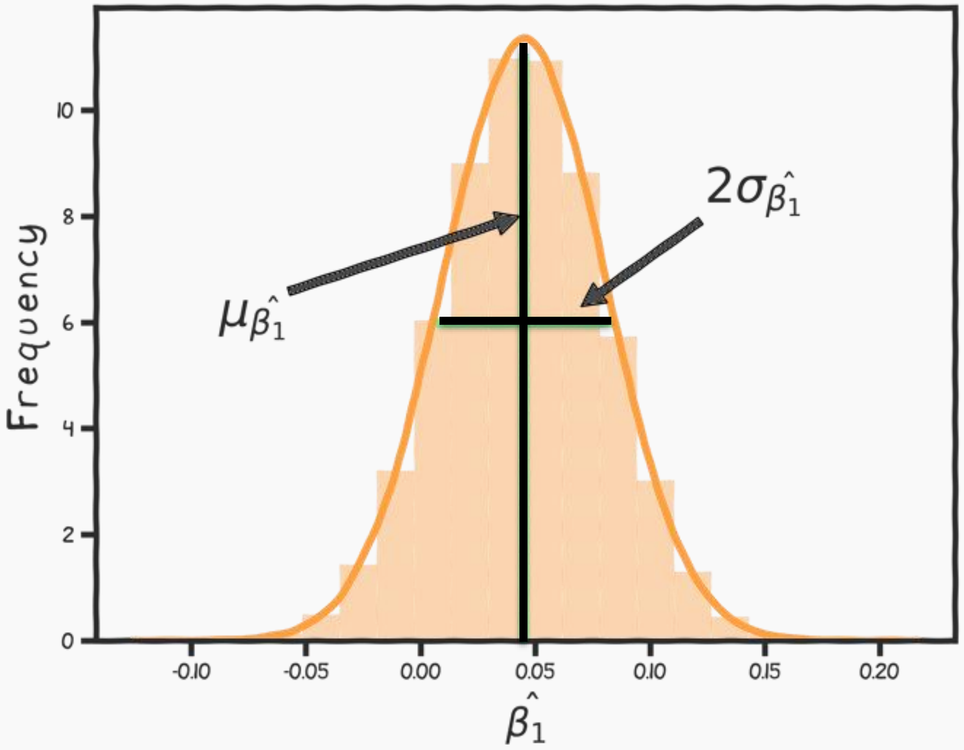
Feature Importance

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

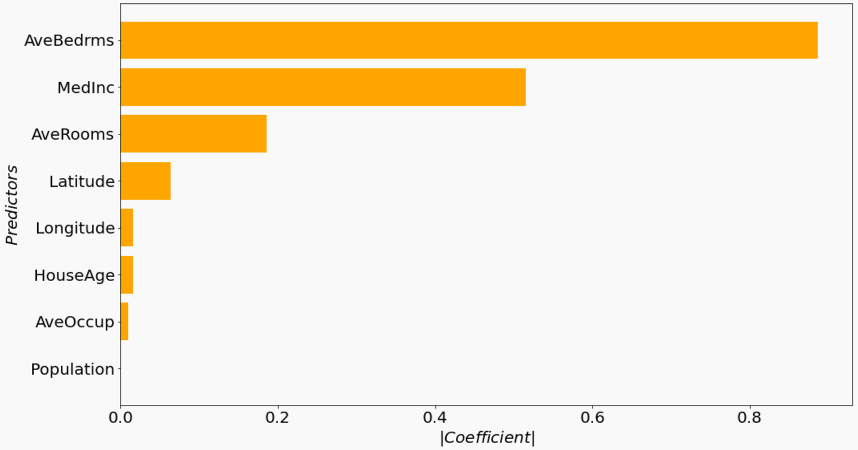
$\hat{t}\text{–}\boldsymbol{test}$ is a scaled version of the usual t-test:

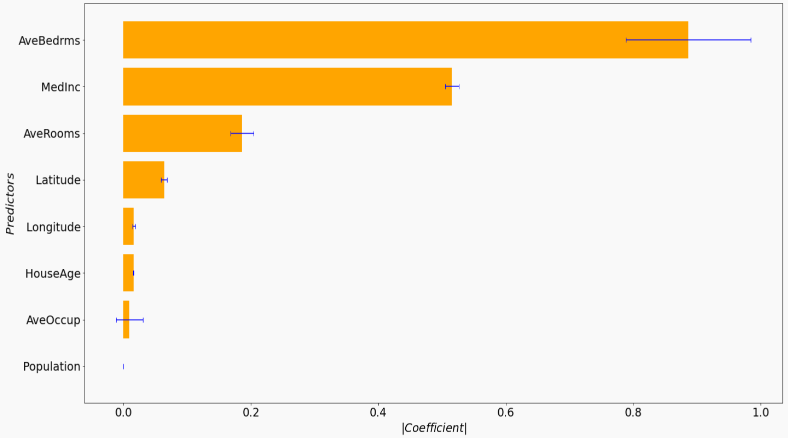
$$t\text{–test} = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}/\sqrt{n}} = \sqrt{n}\,\hat{t}\text{–test}$$

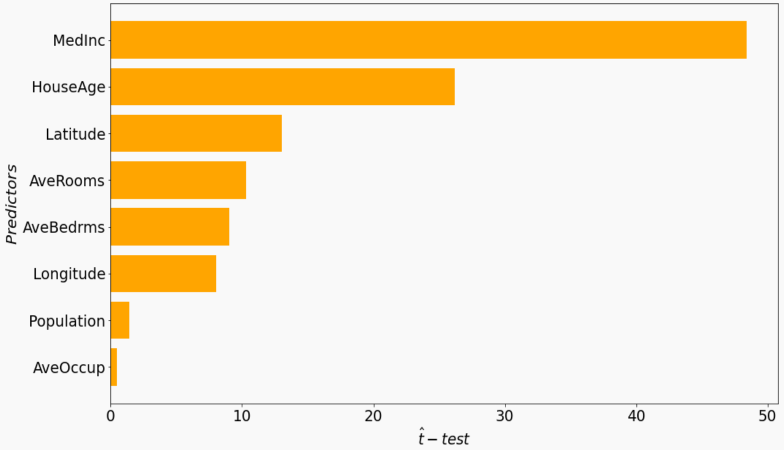
n is the number of bootstraps.



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:

Feature Importance

What do we mean by 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.

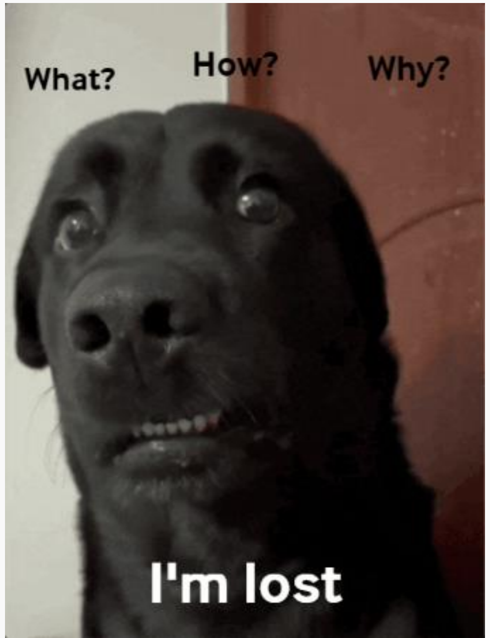
Feature Importance

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)`

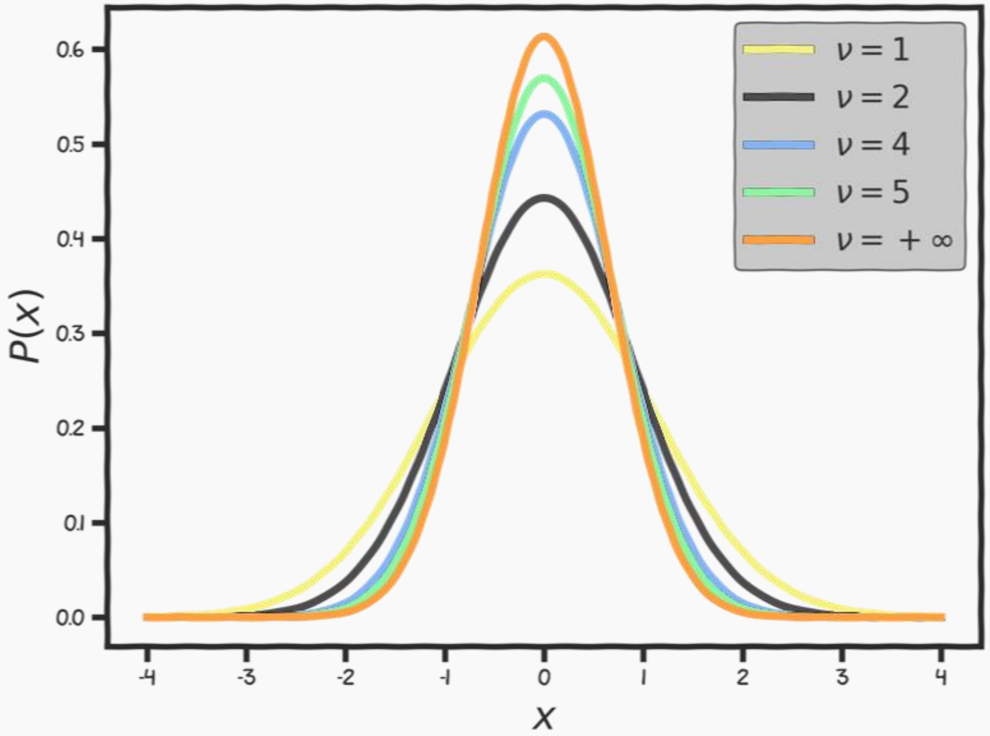
Feature Importance

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 $\hat{t}\_test\_random = \frac{\mu_{\hat{\beta}_1}}{\widehat{SE}_{\hat{\beta}_1}}$.
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.

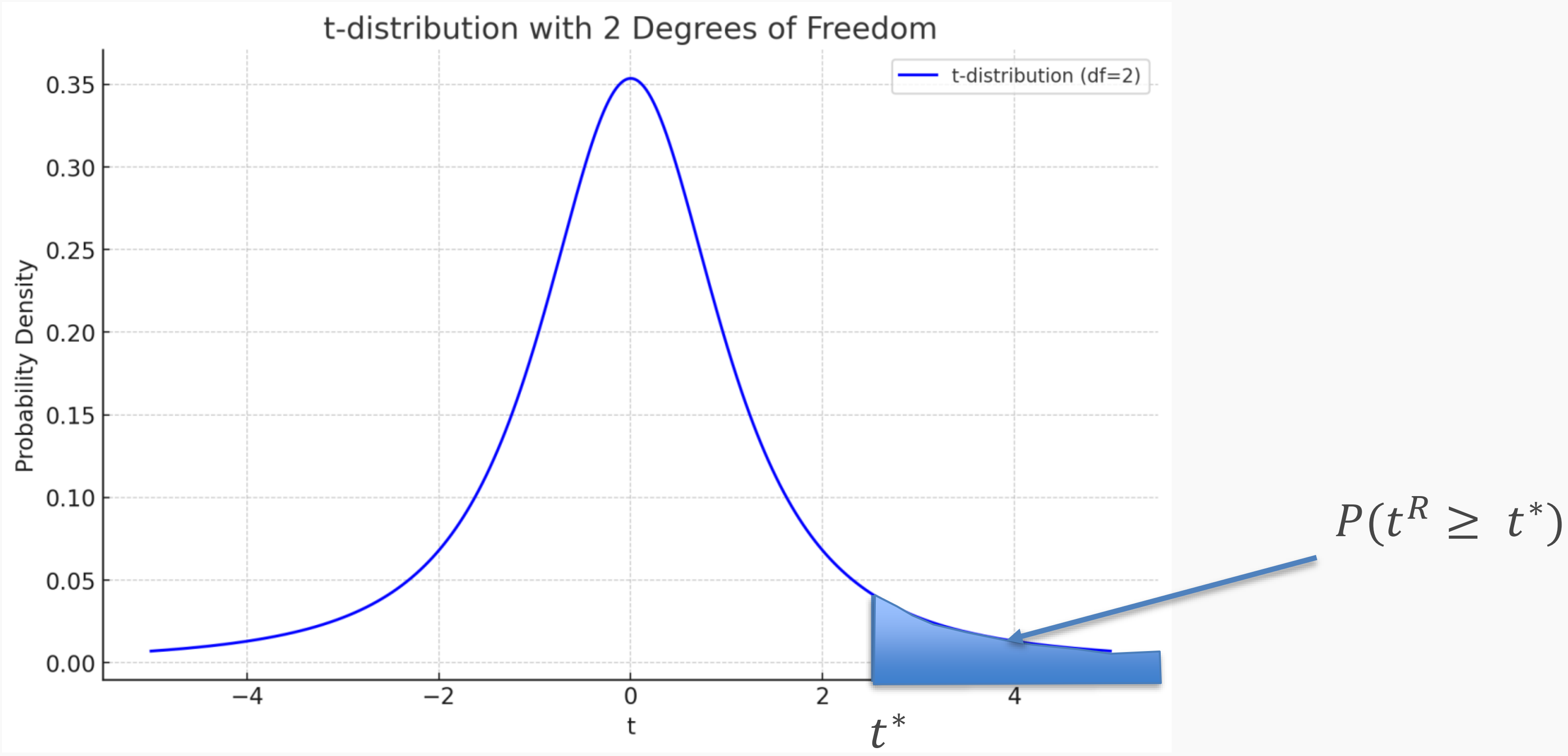


Student-t distribution, where ν is the degrees of freedom (number of data points minus number of predictors) = $\boldsymbol{n} - (\boldsymbol{p} + 1)$).

[More about student-t and degrees of freedom]

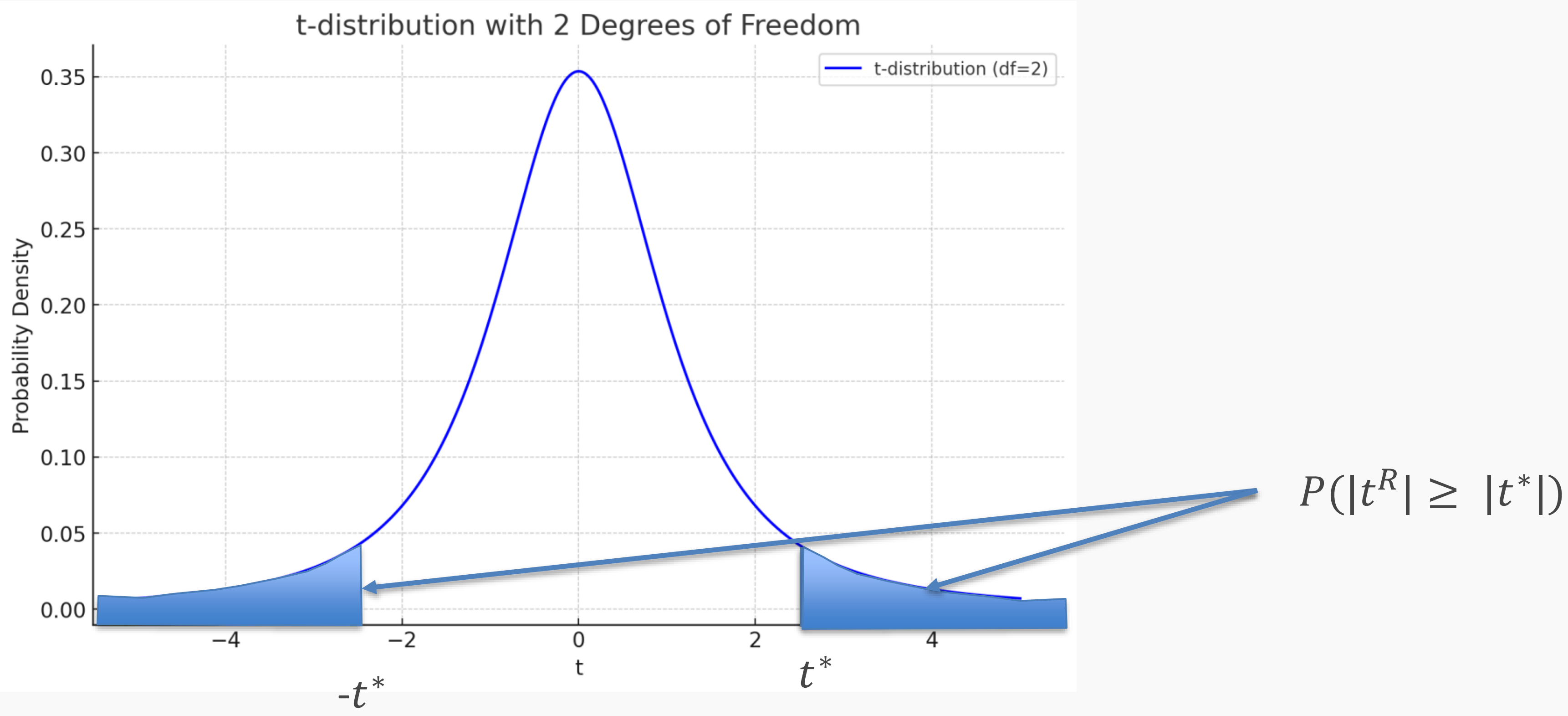
P-value

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 $|t^R| \geq |t^*|$.



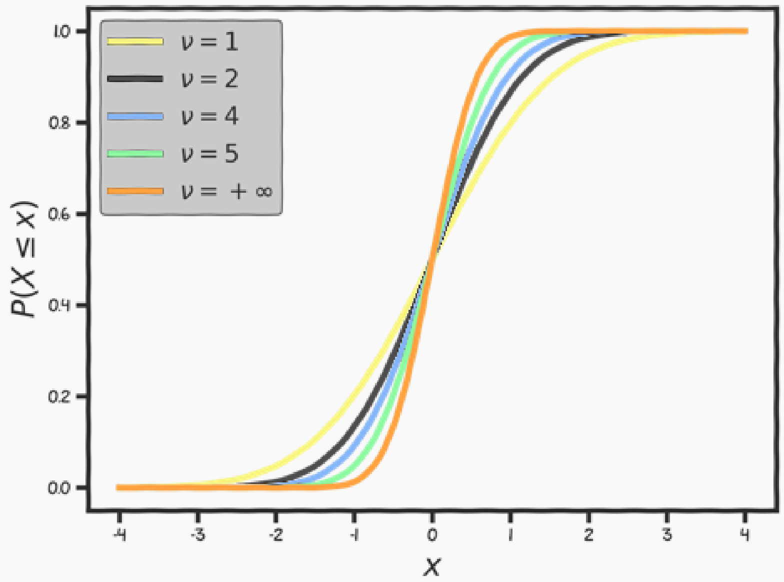
P-value

We call this probability the p-value:

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

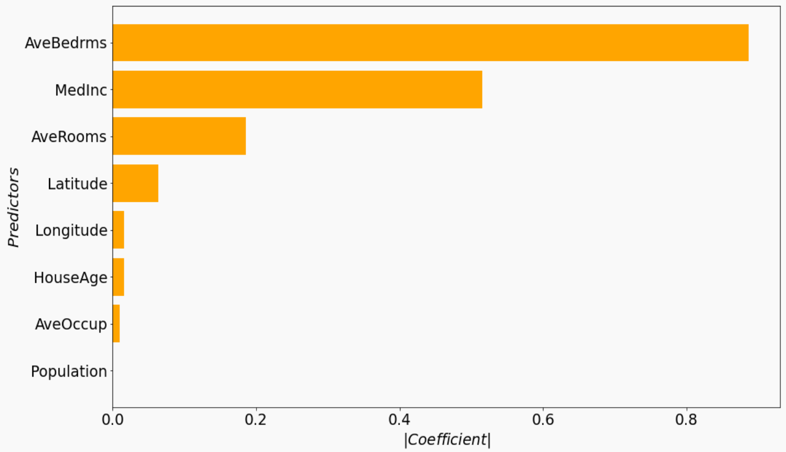
Small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.
It is common to use p-value<0.05 as the threshold for significance.

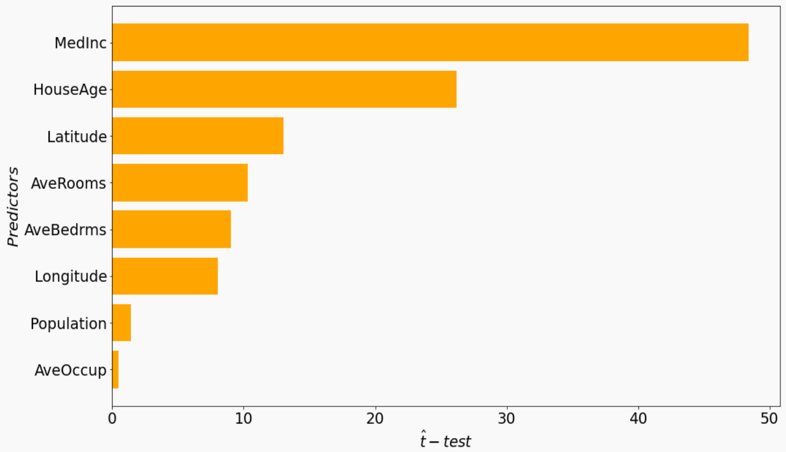
To calculate the p-value we use the cumulative distribution function (CDF) of the student-t. `stats model` a python library has a build-in function `stats.t.cdf()` which can be used to calculate this.

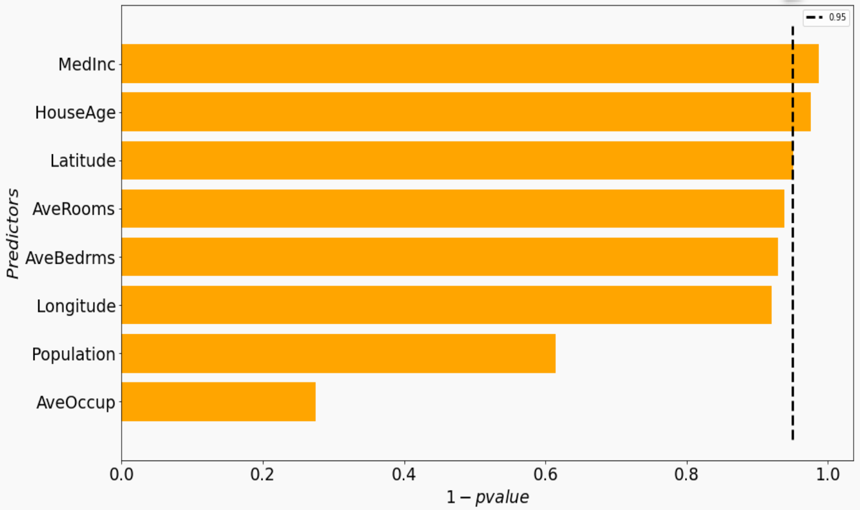


Feature Importance

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.

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 [not]

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

The confidence intervals of $\hat{f}$

Not Done Yet