Evaluating Significance of Predictors Hypothesis Testing

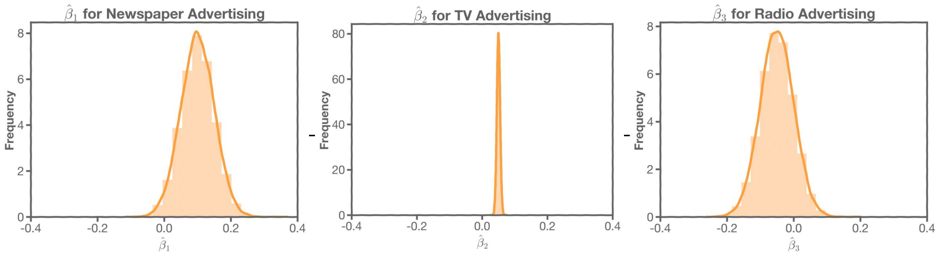
Bookmark this page

Feature importance

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

1. Which predictors are most important?
2. which of them really affects the outcome?

The three charts below show three different histograms for beta, for newspaper, TV, and radio advertising. Each one has a different mean and standard deviation. How can we tell which of these three predictors is the most important?



To incorporate the uncertainty of the coefficients, 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**:

> 📘 THE $\hat{t}$ -TEST
>
> The $\hat{t}$ -test statistic measures the distance from zero in units of standard deviation.
>
> $$\hat{t} - test = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

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

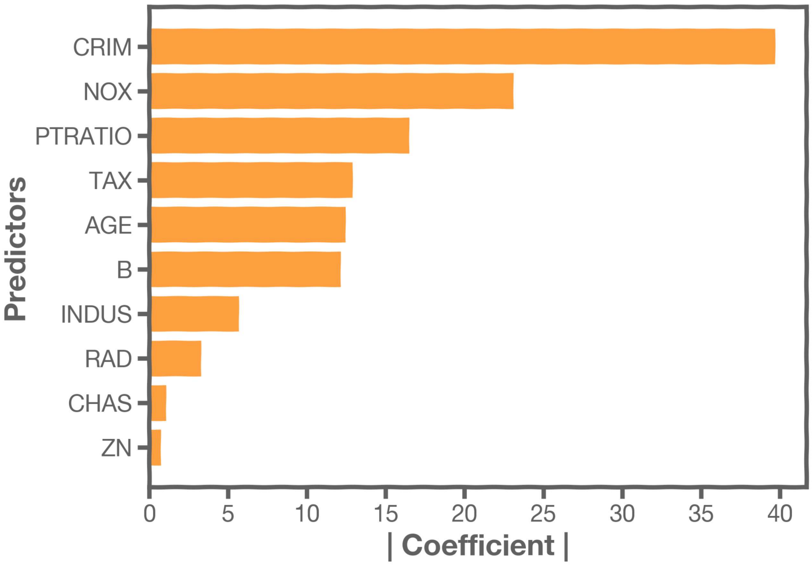
$$\hat{t} - test = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}/\sqrt{n}} = \sqrt{n} - test$$

where n is the number of bootstraps.

Feature Importance

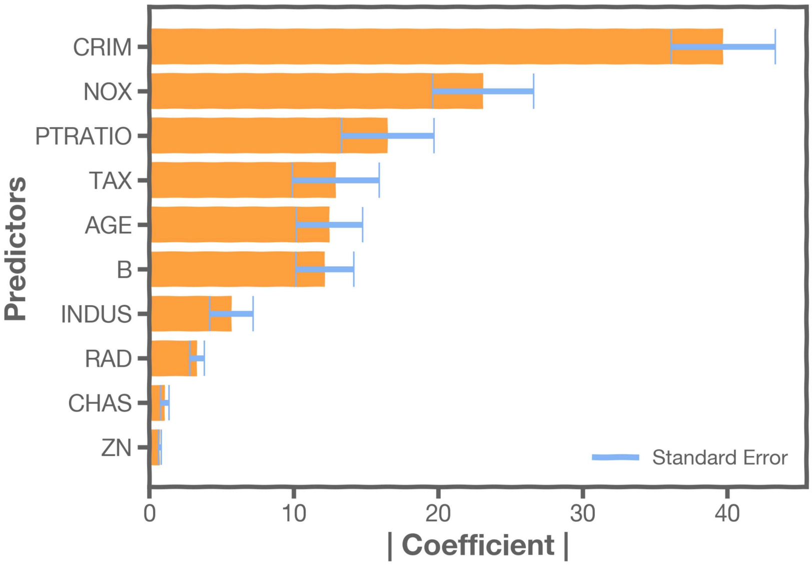
Consider the following example using the Boston Housing data. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston. The coefficients below are from a model that predicts prices given house size, age, crime, pupil-teacher ratio, etc.

This plot gives the feature importance based on the absolute value of the coefficients.

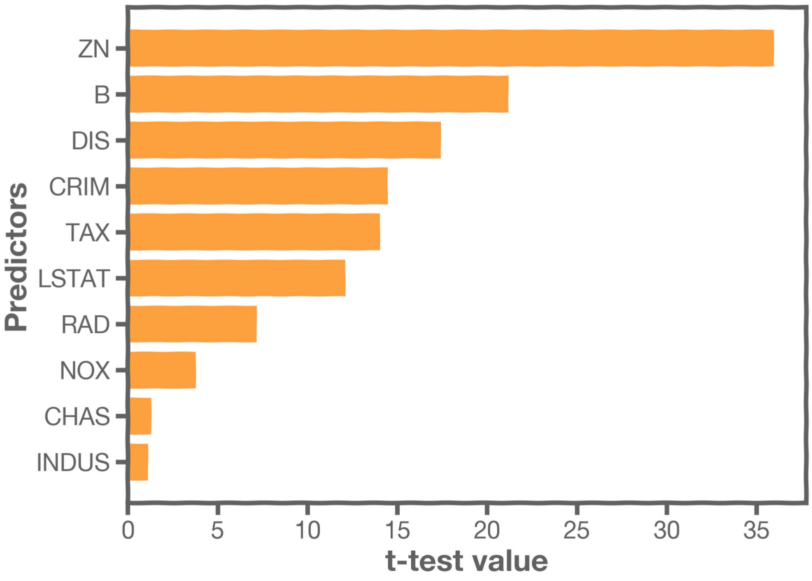


Calculator

The following plot gives the feature importance based not on the absolute value of the coefficients over multiple bootstraps and includes the uncertainty of the coefficients.



Finally, we have the feature importance based on the t-test. Notice that the rank of the importance has changed.



Just 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?

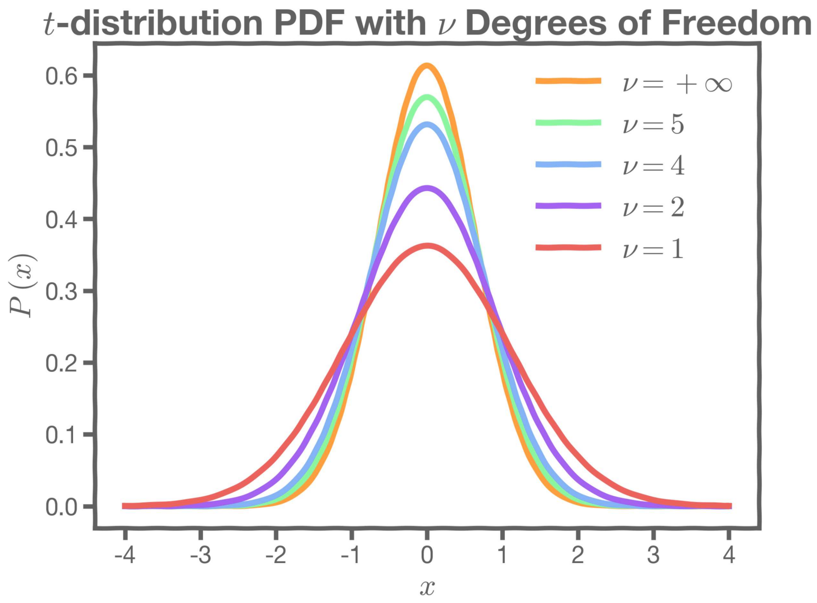
As with R-squared, we should compare its significance $\hat{t}$-test to the equivalent measure from a dataset where we know that there is no relationship between predictors and outcome.

We are sure that there will be no such relationship in data that are randomly generated. Therefore, we want to compare the $\hat{t}$-test of the predictors from our model with $\hat{t}$-test values calculated using random data using the following steps:

- For n random datasets fit n models
- Generate distributions for all predictors and calculate the means and standard errors
- Calculate the t-tests
- Repeat and create a Probability Density Function (PDF) for all the t-tests

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

In this student-t distribution plot, v represents the degrees of freedom (number of data points minus number of predictors):



P-value

To compare the t-test values of the predictors form our model, $|t^R|$, we estimate the probability of observing $|t^R| \geq |t^*|$. This is called the probability of the p-value, defined as:

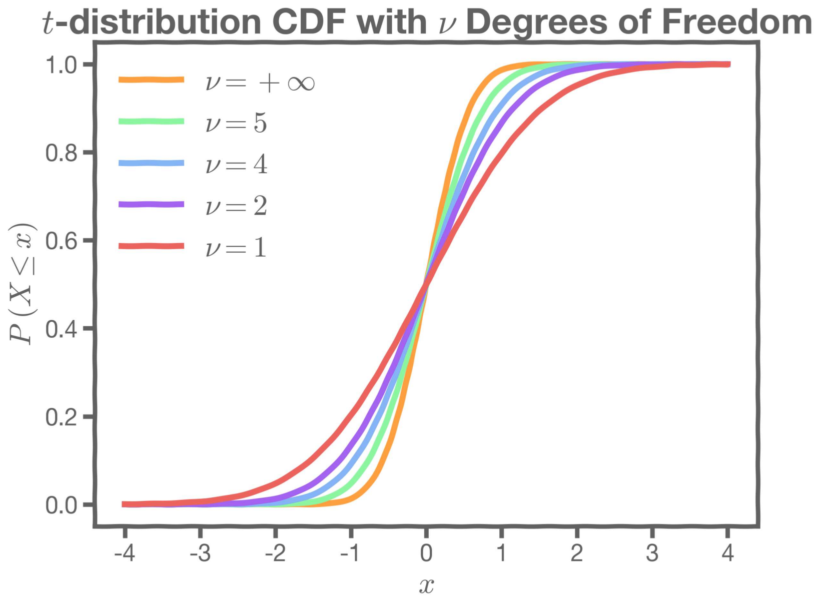
> **P-VALUE**
>
> $$p - value \ = \ P\left(\left|t^R\right| \geq \left|t^*\right|\right)$$

A 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 practice to use p-value < 0.05 as the threshold for significance.

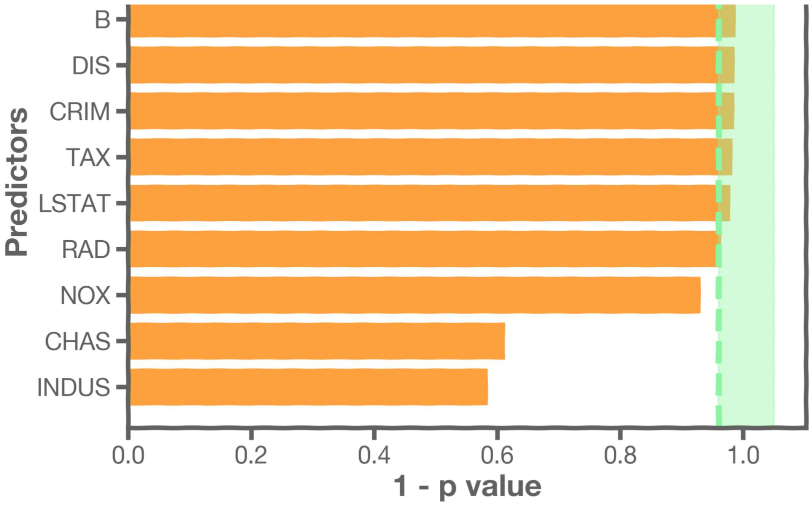
NOTE: To calculate the p-value we use the Cumulative Distribution Function (CDF) of the student-t.

`stats model` a Python library has a build-in function `stats.t.cdf()` which can be used to calculate this.



As a continuation of the Boston Housing data example used above, we now have the feature importance plotted using the p-value.

Calculator



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.

The involves the following steps:

1. State the hypothesis, 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.

Use the value of the test statistic to either reject or not reject the null hypothesis.

This is an example of the Hypothesis testing process:

1. **State Hypothesis**
    - **Null hypothesis:** H_0: There is no relation between X and Y
    - **Alternative hypothesis:** H_0: There is some relation between X and Y
2. **Choose Test Statistic:** the t-test
3. **Sample:** Using bootstrapping we can estimate $\hat{\beta}_1$s, $\mu_{\hat{\beta}_1}$, $\sigma_{\hat{\beta}_1}$ and the t-test
4. **Reject or accept the hypothesis**
    - We compute the p-value, the probability of observing any value equal to $|t|$ or larger from random data.
        - If p-value < (p-value - threshold), reject the null hypothesis
        - Else, accept the null hypothesis

Discussion Board (External resource)

Click OK to have your username and e-mail address sent to a 3rd party application.

OK

‹ Previous
Next ›

Calculator

Calculator