Evaluating Significance of Predictors Hypothesis Testing

Pavlos Protopapas, Ignacio Becker

How reliable are the model interpretation

Suppose our model for advertising is:

$$y = 1.01x + 120$$

Where y is the sales in $1000, x is the TV budget.

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

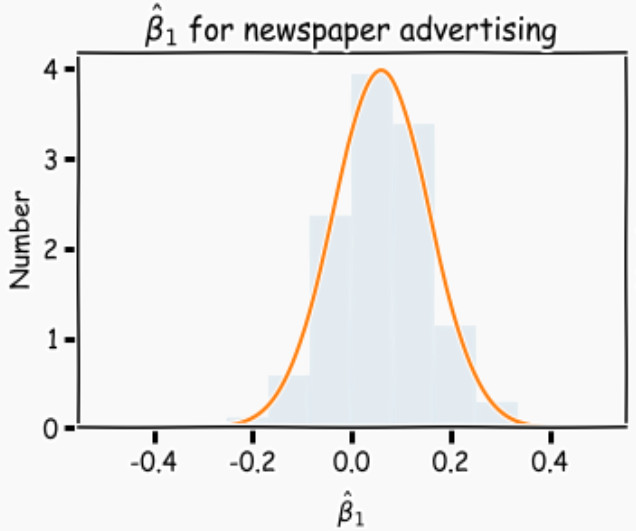
But how certain are we in our estimation of the coefficient 1.01?

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

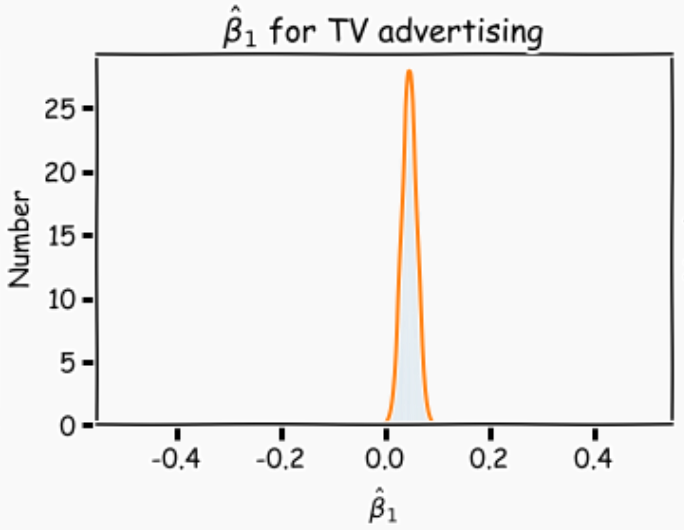
Feature importance

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

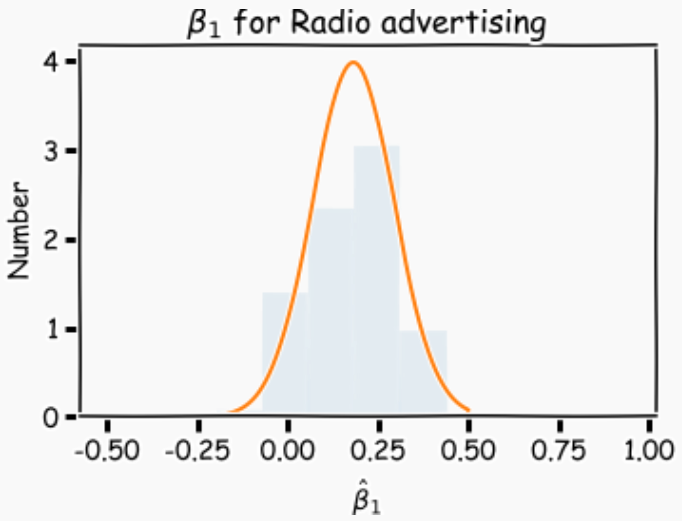
A. Which predictors are most important?
B. And which of them really affects the outcome?



$\mu_{\beta_1} = 0.03$
σ_{β_1}=0.13

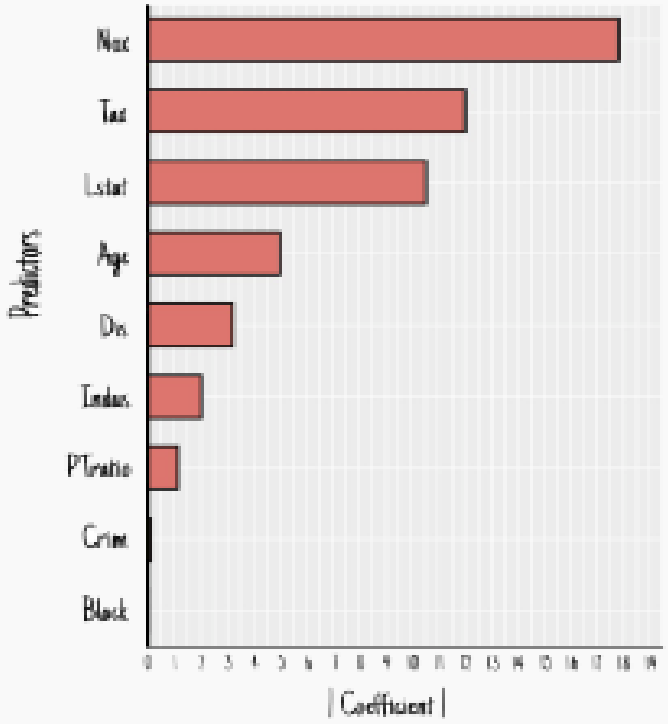


$\mu_{\beta_1} = 0.033$
σ_{β_1}=0.01

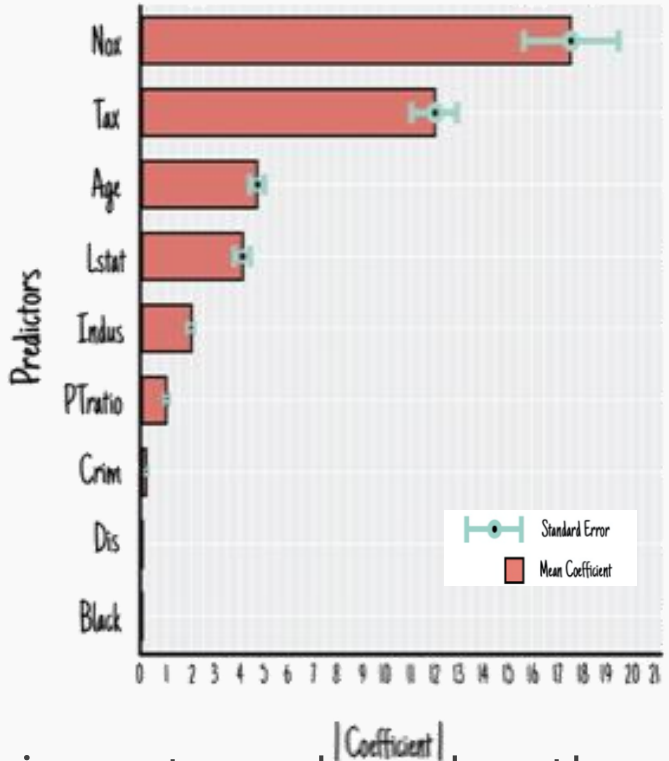


$\mu_{\beta_1} = 0.23$
σ_{β_1}=0.25

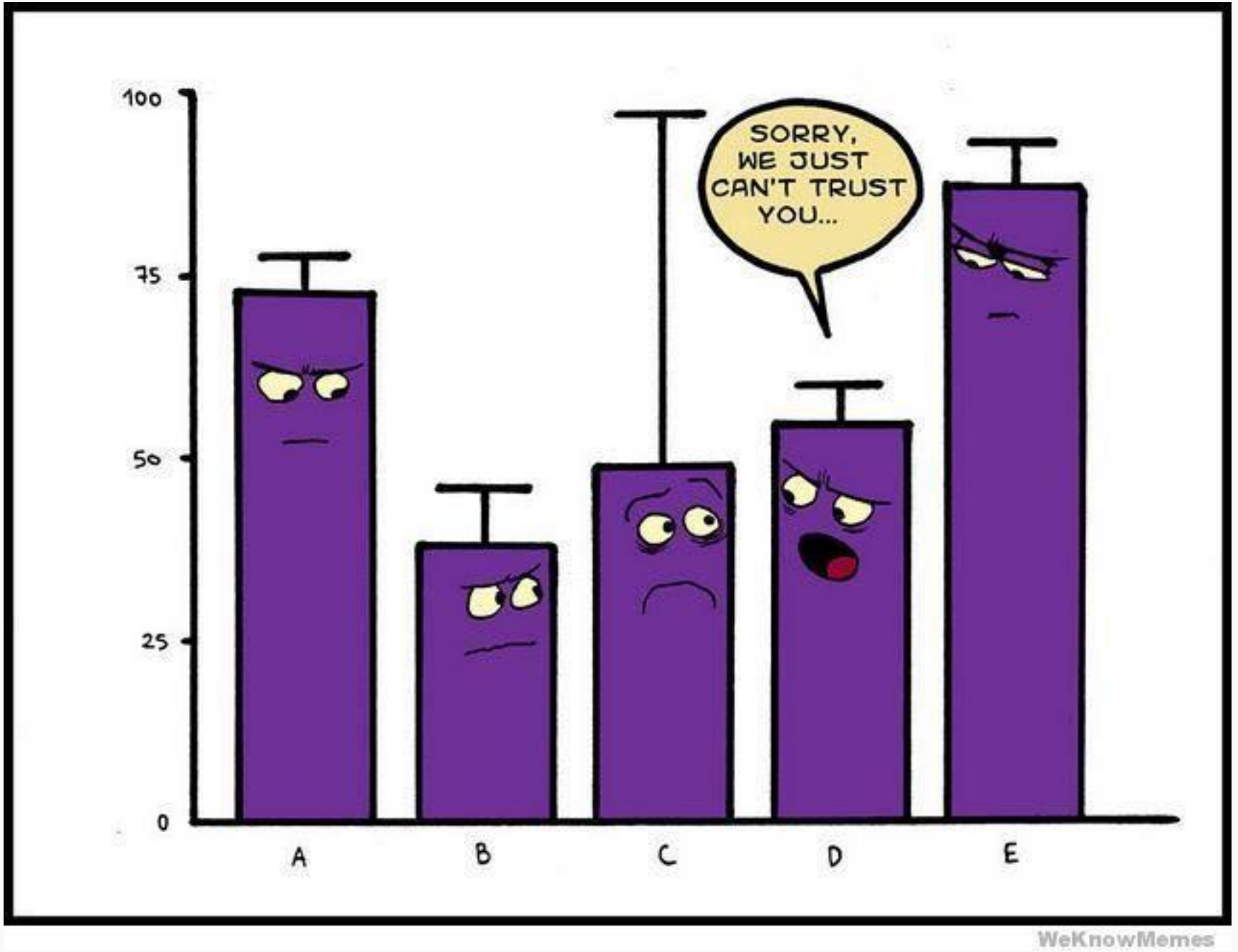
The example below is from Boston housing data. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston. The coefficients below are from a model that predicts prices given house size, age, crime, pupil-teacher ratio, etc.



Feature importance based on the absolute value of the coefficients.



Feature importance based on the absolute mean value of the coefficients over multiple bootstraps and includes the uncertainty of the coefficients.


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.

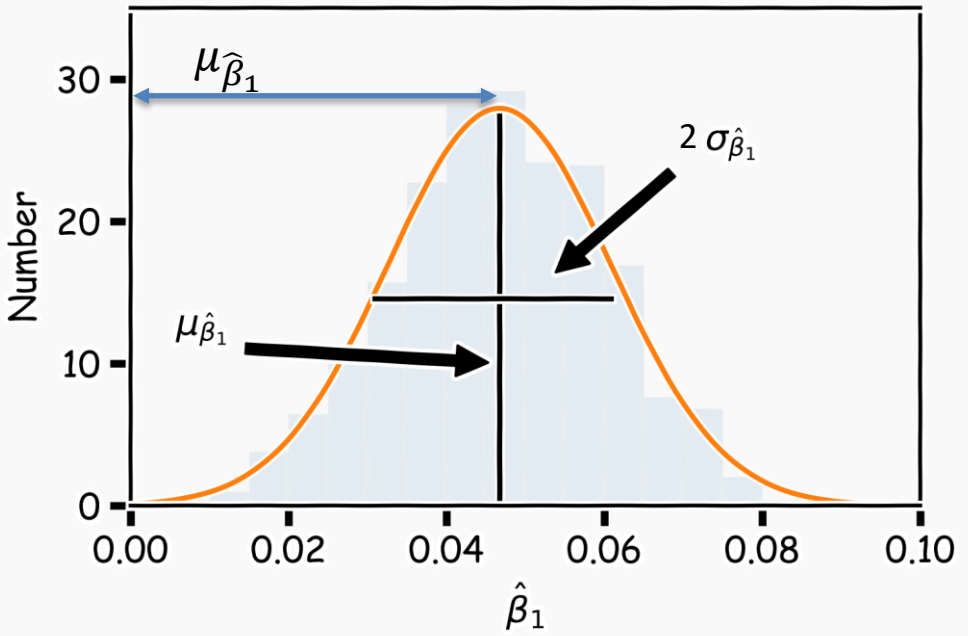
To do so, we define a new metric, which we call $\hat{t}$-test statistic:

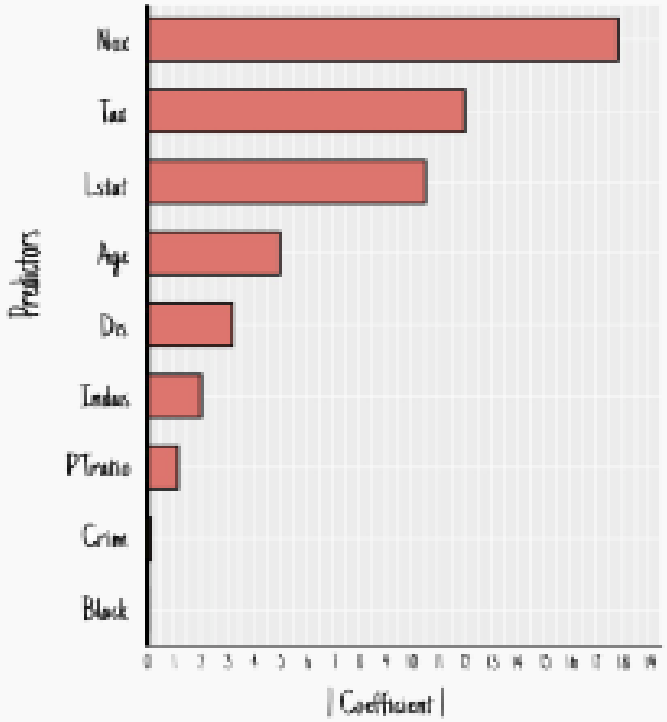
$$\hat{t}\text{–test} = \frac{\mu_{\hat{\beta}_1}}{\sigma_{\hat{\beta}_1}}$$

which measures the distance from zero in units of standard deviation.

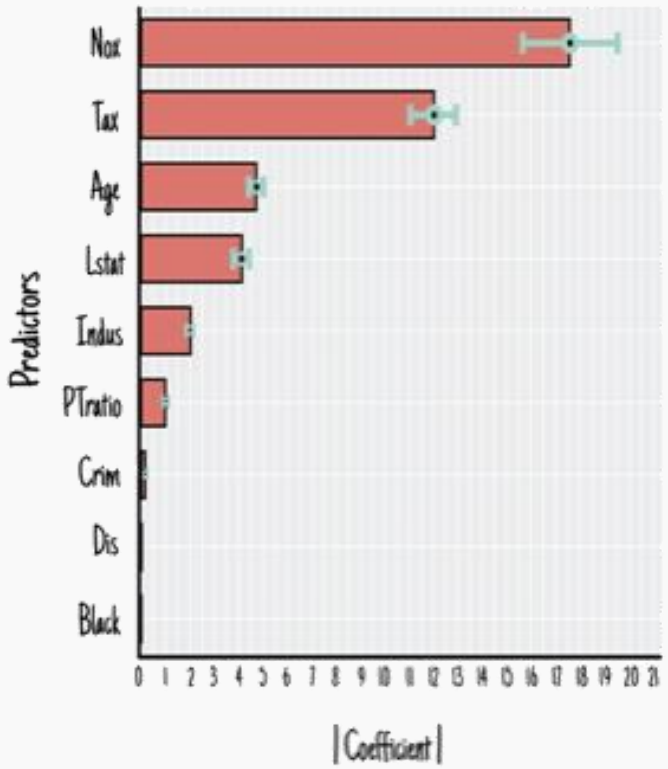
$\hat{t}$–test is a scaled version of the usual t-test

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

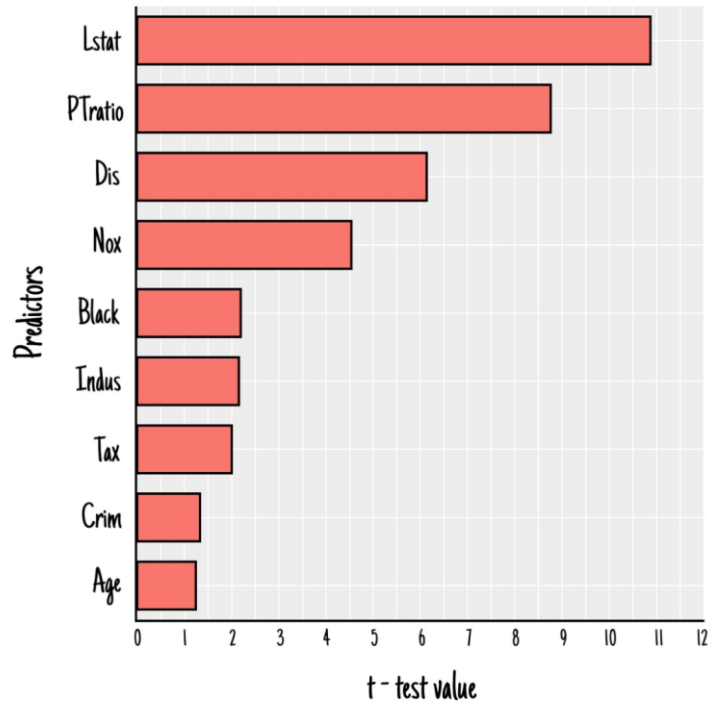




Feature importance base on the absolute value of the coefficients.



Feature importance base on the absolute value of the coefficients over multiple bootstraps and includes the uncertainty of the coefficients.



Feature importance base on $\hat{t}$-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?

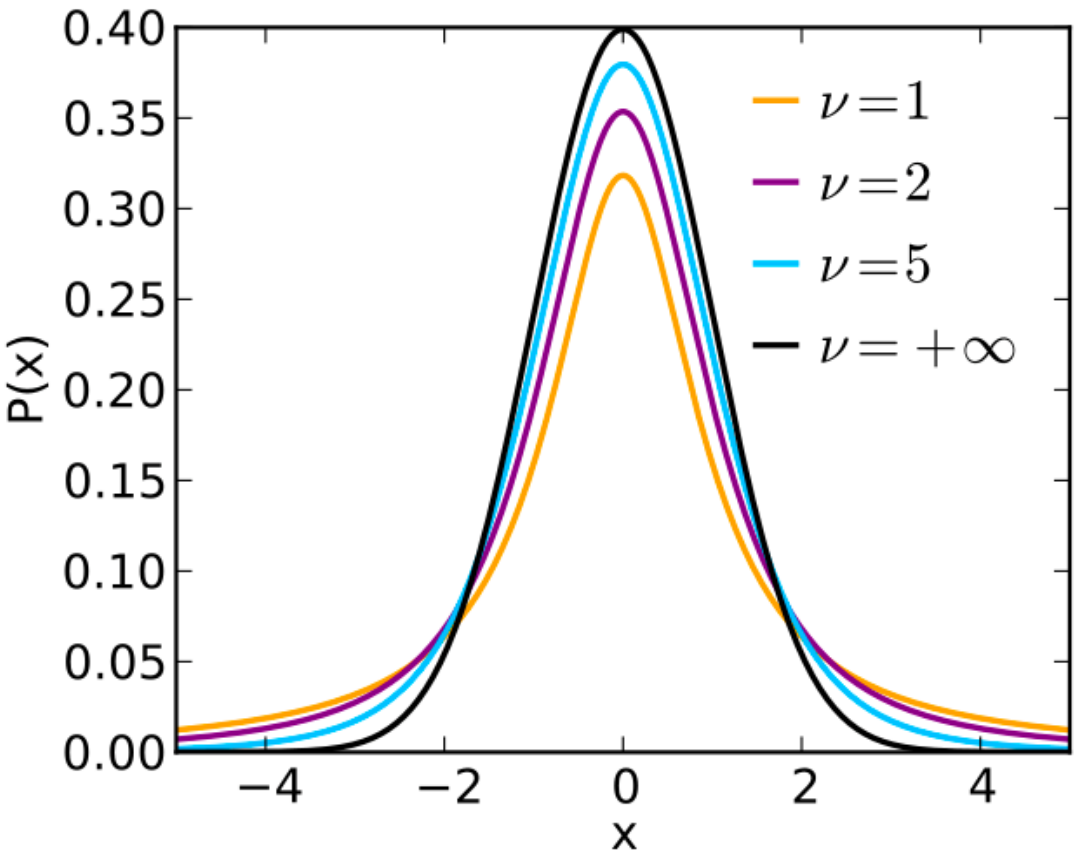
As with R-squared, we should compare its significance ($\hat{t}$-test) to the equivalent measure from a dataset where we know that there is no relationship between predictors and outcome.

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.

1. For n random datasets fit n models.
2. Generate distributions for all predictors and calculate the means and standard errors $(\mu_{\hat{\beta}}, \sigma_{\hat{\beta}})$.
3. Calculate the $\hat{t}$-tests.

Repeat and create a probability density function (pdf) for all the $\hat{t}$-test.

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).

To learn more about why student-t, what are degrees of freedom and more details see https://en.wikipedia.org/wiki/Student%27s_t-test

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

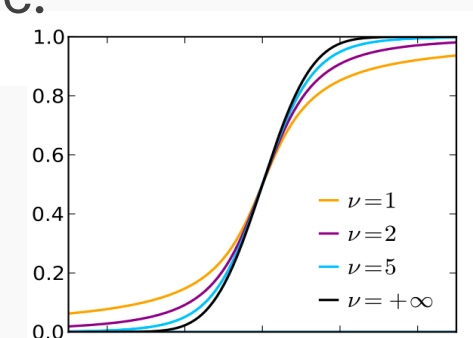
We call this probability the p-value.

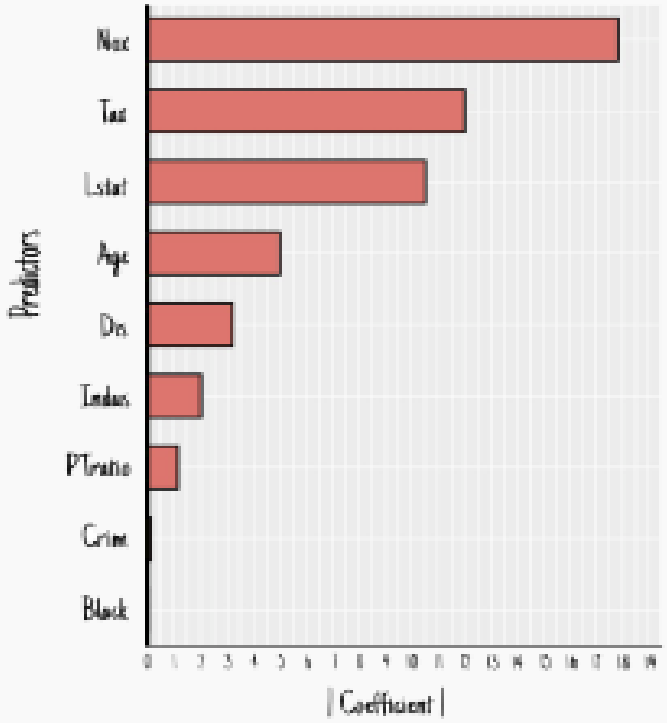
$$p - value = P(|t^R| \geq |t^*|)$$

small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.

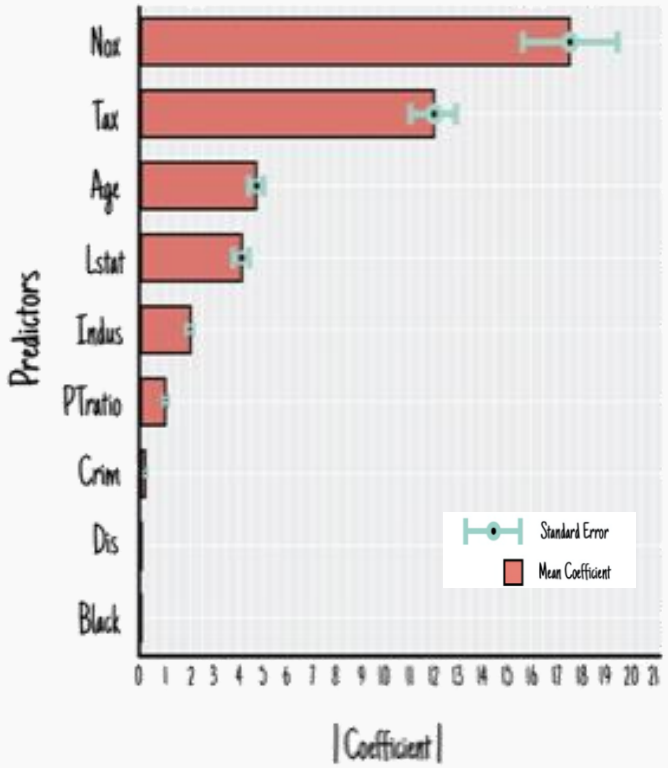
It is common to use p-value<0.05 as the threshold for significance.

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.

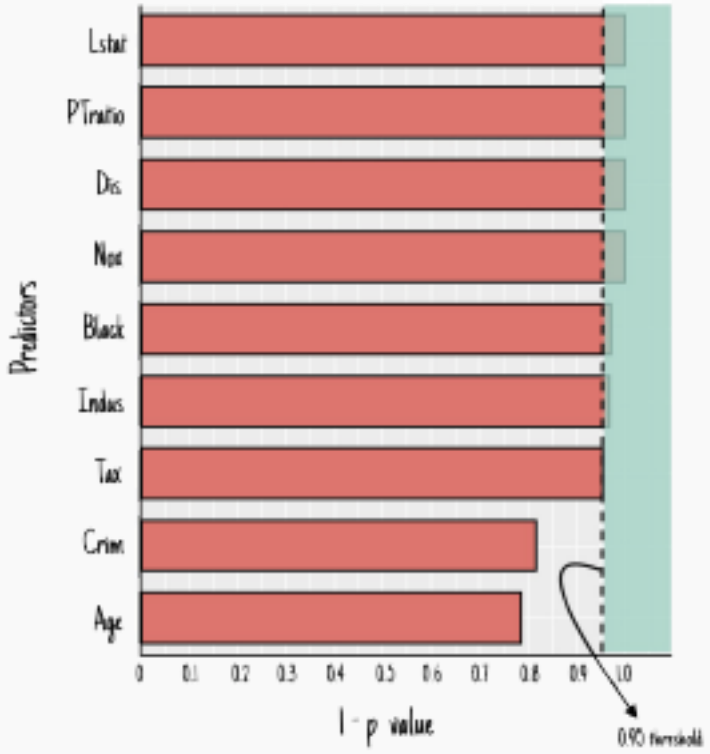




Feature importance based on the absolute value of the coefficients over multiple bootstraps and includes the coefficients' uncertainty.



Feature importance based on t-test. Notice the rank of the importance has changed.



Feature importance using p-value.

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 and Y

The alternative:

H_a: There is some relation between X and Y

2. **Choose test statistics**

t-test

3. Sample:

Using bootstrap we can estimate $\hat{\beta}_1'$s, and $\mu_{\hat{\beta}_1}$ and $\sigma_{\hat{\beta}_1}$ and the t-test.

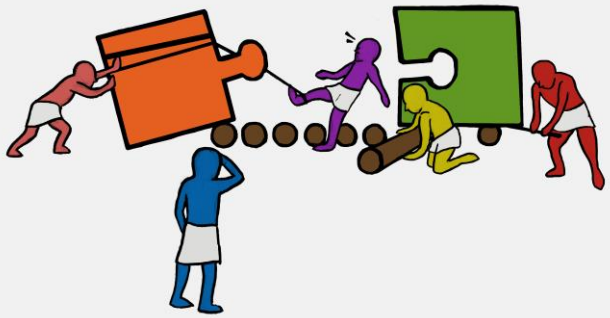
Hypothesis testing

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.

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

Thank you



Description

Exercise: D.1 - Computing the CI

You are the manager of the Advertising division of your company, and your boss asks you the question, ***"How much more sales will we have if we invest $1000 dollars in TV advertising?"***

The goal of this exercise is to estimate the **Sales** with a 95% confidence interval using the *Advertising.csv* dataset.

Instructions:

- Read the file `Advertising.csv` as a dataframe.
- Fix a budget amount of 1000 dollars for TV advertising as variable called *Budget*.
- Select the number of bootstraps.
- For each bootstrap:
  - Select a new dataframe with the predictor as *TV* and the response as *Sales*.
  - Fit a simple linear regression on the data.
  - Predict on the *budget* and compute the error estimate using the helper function `error_func()` .
  - Store the `sales` as a sum of the prediction and the error estimate and append to `sales_list` .
- Sort the `sales_list` which is a distribution of predicted sales over `numboot` bootstraps.
- Compute the 95% confidence interval of `sales_list` .
- Use the helper function `plot_simulation` to visualize the distribution and print the estimated sales.