Simple Linear Models

## Model and Model Assumptions

Given the data

|  |  |
| --- | --- |
|  |  |
|  |  |
| … | … |
|  |  |

We would like to approximate with a function of , . Assume that this function is linear

Let the error/residual of each approximation be

To do inferential statistics, estimating the error of or finding confidence interval for , we need to bring in a probability distribution. Assume that is a known constant (non-random) and that

where

. From this equation, we see that . and are random, is non-random.

## Coefficient Estimators

There are multiple ways to define the quality of the approximation. For example, in the least square method, one wants to minimize the total sum of the errors or residual sum square (RSS) or sum square errors (SSE)

Notice that we use the summation notation () to present a sum of all for from 1 to . A full version of is . We will use instead of for simplicity.

In the least absolute error or least absolute deviations, one wants to minimize

The values of and that minimizes RSS is denoted by and , respectively.

We have some properties for the and as follows.

$S^2=\frac{SSE}{n-2} = \frac{S\_{yy}}{\hat\beta\_1\S\_{xy}}$ is an unbiased estimator for

## Goodness of Fit

We have the following formula

The quantity can be though as the total amount of information (variance) in the response variable . This quantity is decomposed into which can be thought as the information explained by the resgression models and , which is the amount of information that can not be explained by the model (the sum of squared erros of the models).

This formula leads to the definition of

Another way to view this formula is that TSS is actually the sum squared errors of the naive model $ y = {y}$ (predicting value by its average, disregarding ). Thus,

Thus, also measure how good the linear model is when comparing with the naive model. This definition of can be applied to measure the goodness of fit of other models, not just linear model.

To measure the goodness of fit of this linear approximation, we compare the approximation with a naive approximation, which is to approximate all with their average, . The naive model is a special case of the linear model where . In the other words, the naive models is

The sum square error of the naive models is . This quantity is also called the total sum square (SST or TSS). We use the following to measure the goodness of fit of the linear model

It’s noticed that if the linear approximation fit the data perfectly (), .

## Inferences on Coeffient Estimators

### F- Test

We are interested in testing the following hypotheses for

### t-test

$$
H\_0: \beta\_1 = d \\
H\_{\alpha}: \beta\_1 \neq d
$$

We have that under

It’s noticed that the F-test and the t-test have the same rejection region when