What is the PRESS Statistic?

Authored by

stats writer

April 19, 2024

The PRESS Statistic is a tool used in statistical analysis to evaluate the predictive power of a model. It stands for Predicted Residual Error Sum of Squares and is calculated by comparing the predicted values of a model to the actual values. The lower the PRESS Statistic, the more accurate the model is at predicting outcomes. This statistic is often used in regression analysis and can help researchers determine the best model to use for making predictions. It is a valuable tool for assessing the reliability and effectiveness of statistical models.

What is the PRESS Statistic?

In statistics, we fit regression models for two reasons:

(1) To *explain* the relationship between one or more explanatory variables and a response variable.

(2) To *predict* values of a response variable based on the values of one or more explanatory variables.

When our goal is to (2) *predict* the values of a response variable, we want to make sure that we're using the best possible regression model to do so.

One metric that we can use to find the regression model that will make the best predictions on new data is the PRESS Statistic, which stands for the "Predicted REsidual Sum of Squares."

It is calculated as:

PRESS = Σ(ei / (1-hii))2

where:

ei: The ith residual.hii: A measure of the influence (also called "leverage") of the ith observation on the model fit.

Given several regression models, the one with the lowest PRESS should be selected as the one that will perform best on a new dataset.

The following example shows how to calculate the PRESS statistic for three different linear regression models in R.

Example: Calculating the PRESS Statistic

Suppose we have a dataset with three explanatory variables, x1, x2, and x3, and one response variable y:

```
data <- data.frame(x1 = c(2, 3, 3, 4, 4, 6, 8, 9, 9, 9),
x2 = c(2, 2, 3, 3, 2, 3, 5, 6, 6, 7),
x3 = c(12, 14, 14, 13, 8, 8, 9, 14, 11, 7),
y = c(23, 24, 15, 9, 14, 17, 22, 26, 34, 35))
```

The following code shows how to fit three different regression models to this dataset using the lm() function:

```
model1 <- lm(y~x1, data=data)

model2 <- lm(y~x1+x2, data=data)

model3 <- lm(y~x2+x3, data=data)
```

The following code shows how to calculate the PRESS statistic for each model.

```
#create custom function to calculate the PRESS statistic
PRESS <- function(model) {
i <- residuals(model)/(1 - lm.influence(model)$hat)
sum(i^2)
}

#calculate PRESS for model 1
PRESS(model1)

590.2197

#calculate PRESS for model 2
```

PRESS(model2)

519.6435

#calculate PRESS for model 3
PRESS(model3)

537.7503

It turns out that the model with the lowest PRESS statistic is model 2 with a PRESS statistic of 519.6435. Thus, we would choose this model as the one that is best suited to make predictions on a new dataset.

Introduction to Simple Linear Regression
What is a Parsimonious Model?
What is a Good R-squared Value?