What is Mallows’ Cp and what is its definition? Can you provide an example?

Authored by

stats writer

April 30, 2024

Mallows' Cp is a statistical measure used in regression analysis that evaluates the accuracy and predictive power of a model. It is defined as the difference between the mean squared error of the model and the mean squared error of a full model that includes all possible independent variables. In other words, it measures the trade-off between model simplicity and accuracy. A lower Cp value indicates a more accurate and parsimonious model. For example, if we have a regression model with three independent variables and we calculate a Cp value of 4, it means that the model is 4 units away from the ideal model that includes all possible variables.

What is Mallows' Cp? (Defintion & Example)

Mallows' Cp is a metric that is used to pick the best among several different models.

It is calculated as:

Cp = RSSp/S2 - N + 2(P+1)

where:

RSSp: The residual sum of squares for a model with *p* predictor variablesS2: The residual mean square for the model (estimated by MSE)N: The sample sizeP: The number of predictor variables

Mallows' Cp is used when we have several potential predictor variables that we'd like to use in a regression model and we'd like to identify the best model that uses a subset of these predictor variables.

We can identify the "best" regression model by identifying the model with the lowest Cp value that is less than P+1, where P is the number of predictor variables in the model.

The following example shows how to use Mallows' Cp to pick the best regression model among several potential models.

Example: Using Mallows' Cp to Pick the Best Model

Suppose a professor would like to use hours studied, prep exams taken, and current GPA as predictor variables in a regression model to predict the score that a student will receive on the final exam.

He fits seven different regression models and calculates the value for Mallows' Cp for each model:

Predictor Variables	P+1	Mallows' Cp
Hours	2	45.5
Prep exams	2	31.4
GPA	2	29.3
Hours, Prep exams	3	3.4
Hours, GPA	3	2.9
Prep exams, GPA	3	2.7
Hours, Prep exams, GPA	4	4

If the value of Mallows' Cp is less than the number of coefficients in the model (P+1) then the model is said to be unbiased.

We can see that there are two models that are unbiased:

The model with Hours and GPA as the predictor variables (Mallows' Cp = 2.9, P+1 = 3)The model with Prep Exams and GPA as the predictor variables (Mallows' Cp = 2.7, P+1 = 3)

Among these two models, the model that uses Prep Exams and GPA as the predictor variables has the lowest value for Mallows' Cp, which tells us that it's the best model that leads to the least amount of bias.

Notes on Mallows' Cp

Models that have a Mallows' Cp value near P+1 are said to have low bias.If every potential model has a high value for Mallows' Cp, this is an indication that some important predictor variables are likely missing from each model.If several potential models have low values for Mallow's Cp, choose the model with the lowest value as the best model to use.

Also keep in mind that Mallows' Cp is only one way to measure the quality of fit of a regression model.

Another commonly used metric is adjusted R-squared, which tells us the proportion of variance in the that can be explained by the predictor variables in the model, adjusted for the number of predictor variables used.

When deciding which regression model is best among a list of several different models, it's a good idea to look at both Mallows' Cp and adjusted R-squared.