HC3

dea

1/3/2022

Contents

Chapter 1

Acknowledgement

I have been compiling notes and tips on R programming from almost everywhere. I have taken most of these notes in Data Science Specialization Course from Coursera.

Chapter 2

HC3

Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model

by J Scott Long and Laurie H. Ervin

In the presence of heteroscedasticity, ordinary least squares (OLS) estimates are unbiased, but the usual tests of significance are generally inappropriate and their use can lead to incorrect inferences. Tests based on a heteroscedasticity consistent covariance matrix (HCCM), however, are consistent even in the presence of heteroscedasticity of an unknown form. Most applications that use a HCCM appear to rely on the asymptotic version known as HC0. Our Monte Carlo simulations show that HC0 often results in incorrect inferences when N 250, while three relatively unknown, small sample versions of the HCCM, and especially a version known as HC3, work well even for N's as small as 25. We recommend that: (1) data analysts should correct for heteroscedasticity using a HCCM whenever there is reason to suspect heteroscedasticity; (2) the decision to use HCCM-based tests should not be determined by a screening test for heteroscedasticity; and (3) when N 250, the HCCM known as HC3 should be used. Since HC3 is simple to compute, we encourage authors of statistical software to add this estimator to their programs.

Chapter 3

Cross Validation

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data.

It is a popular method because it is simple to understand and because it generally results in a less biased than other methods, such as a simple train/test split.

3.1 k-Fold Cross-Validation

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:

a. Take the group as a hold out or test data set
b. Take the remaining groups as a training data set
c. Fit a model on the training set and evaluate it on the test set
d. Retain the evaluation score and discard the model
e. Summarize the skill of the model using the sample of model evaluation scores

ISL: This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds

The results of a k-fold cross-validation run are often summarized with the mean of the model skill scores. It is also good practice to include a measure of the variance of the skill scores, such as the standard deviation or standard error.

See more at tutorial