Reject Inference, Augmentation and Sample Selection

John Banasik and Jonathan Crook

John.Banasik@ed.ac.uk

Abstract

Many researchers see the need for reject inference to come from a sample selection problem whereby a missing variable results in omitted variable bias. Specifically, the success in being accepted for a loan is related to subsequent repayment performance. Accordingly, the residuals of the previous scoring model by which the person is accepted may be correlated with those of a new model that predicts his repayment performance. Unless the correlation between the residuals of the new and old model are reflected in the new model its parameters will be biased. Alternatively, practitioners often see the problem as one of missing data where the relationship in the new model is biased because the behaviour of the omitted cases differs from that of those who make up the sample for a new model. To attempt to correct for this, differential weights are applied to the new cases. The aim of this paper is to see if the use of both a Heckman style sample selection model and the use of sampling weights, *together*, will improve predictive performance compared with either technique used alone. This paper will use a sample of applicants in which virtually every applicant was accepted. This allows us to compare the actual performance of each model with the performance of models which are based only on accepted cases.