Outlier detection is a long-lasting issue in regression analysis and the problem will become very complex when it come to multivariate regression (MVR) modeling where outliers, or extreme observations, are on one or a combination of variables. The presence of outliers can exert negative influences on the fit of the MVR regression model.

There are some traditional approaches to detecting outliers and they have to be done separately for outlying X and Y observations. To identify outlying Y observations, one often examines the studentized deleted residuals. Those observations whose studentized deleted residuals are large in absolute values are identified as outlying Y observations. The detection of outlying X observations depends on the leverage values in the Hat matrix. A leverage value more than twice as large as the mean leverage is usually considered to be a possible outlying X observation. However, in MVR modeling, the computations involved are complicated and expensive.

This study hopes to save the trouble of checking separately for two types of outliers by providing a new outlier detection technique using Akaike’s Information Criterion (AIC) and Bozdogan’s Information Complexity Criterion (ICOMP). These information criteria can be used as mathematical indicators of the quality of the fitted regression model. Since smaller values of the information criteria indicate a better model-data fit, the presence of one or more outliers will negatively impact the regression model and result in larger criterion values.

In view of this, the study will implement a jackknife technique in outlier detection. That is, the study will first settle with a tentative MVR regression model and fit it using all the data points with one excluded. The two information criteria AIC & ICOMP are scored after the model is built to evaluate the model-data fit. Then the procedure is repeated for each observation in the sample and a set of information criterion values are computed for each fitted model. If the point being excluded is an outlying observation, its absence from the MVR model will lead to smaller information criterion values, indicating a better model-data fit than when the observation is included.
The tentative MVR regression model used in the jackknife procedure is determined using a stochastic search technique termed Genetic Algorithm (GA) with all observations included. Once the tentative model is selected, the predictor subset will stay fixed and will be fitted repeatedly during the jackknife process. By using the same subset of predictors, put on an equal scale is the measurement of each observation in the sample.

The technique GA is used to select optimal subsets of variables and it represents the presence or absence of a predictor variable using 1 or 0, respectively, on a binary string representing the full MVR model. This algorithm makes possible the rapid selection of variable subsets and avoids the computational burden of exhausting all possible combinations of predictor variables. The decision on subset regression model selection is based on a fitness function which serves as the criterion for mapping the models of the candidate population into the domain of performance. This paper will implement Bozdogan’s Information Complexity Criterion (ICOMP) as the fitness function. This criterion is misspecification-resistant and penalizes interdependence between parameter estimates as well as over-parameterization.