Chapter 20

Outlier Detection in Linear Regression

A. A. M. Nurunnabi
University of Rajshahi, Bangladesh

A. H. M. Rahmatullah Imon
Ball State University, USA

A. B. M. Shawkat Ali
Central Queensland University, Australia

Mohammed Nasser
University of Rajshahi, Bangladesh

ABSTRACT

Regression analysis is one of the most important branches of multivariate statistical techniques. It is widely used in almost every field of research and application in multifactor data, which helps to investigate and to fit an unknown model for quantifying relations among observed variables. Nowadays, it has drawn a large attention to perform the tasks with neural networks, support vector machines, evolutionary algorithms, et cetera. Till today, least squares (LS) is the most popular parameter estimation technique to the practitioners, mainly because of its computational simplicity and underlying optimal properties. It is well-known by now that the method of least squares is a non-resistant fitting process; even a single outlier can spoil the whole estimation procedure. Data contamination by outlier is a practical problem which certainly cannot be avoided. It is very important to be able to detect these outliers. The authors are concerned about the effect outliers have on parameter estimates and on inferences about models and their suitability. In this chapter the authors have made a short discussion of the most well known and efficient outlier detection techniques with numerical demonstrations in linear regression. The chapter will help the people who are interested in exploring and investigating an effective mathematical model. The goal is to make the monograph self-contained maintaining its general accessibility.

DOI: 10.4018/978-1-60960-551-3.ch020
INTRODUCTION

With the remarkable development of computer and information technologies, the number of databases, as well as their volume, dimensionality and complexity grow rapidly, resulting in the analysis of outliers in heterogeneous structured data. When analyzing data, some observations are often occurred that are different from the majority. Generally such observations are called outliers. Sometimes, the outlying observations are not incorrect rather they are made under exceptional circumstances, or they belong to other population(s). So identification of outliers is often by itself the primary goal, without any intention of a regression model. Outlier detection has been suggested for numerous applications, such as credit and fraud detection, clinical trials, medical imaging, voting irregularity analysis, network intrusion, severe weather prediction, geographic information system, and other data mining tasks.

At the beginning, different methods were originally developed arbitrarily in individual fields, but now the systematic approaches are used for outlier detection from the full gamut of computer science and statistics. Machine learning community has shown growing interest in outlier detection. Kernel based methods (Breiman et al., 1977; Terrell & Scott, 1992), distance-based methods (Knorr et al., 2000; Angiulli et al., 2006), density-based methods (Breunig et al., 2000), support vector machines (Eskin et al., 2002) and neural networks (Barron, 1993, 1994; Hawkins et al., 2002) are used as outlier detection techniques. Interested readers are suggested to see the excellent survey by Hodge and Austin (2004). Igelnik (2009) shows neural networks can be efficiently utilized for dynamic modeling of time-variant data. Very recently Chiang et al (2010) observe the limitations of neural networks in real time water level predictions of sewerage systems. Although a non-linear model can sometimes provide accurate prediction result, linear methods are easy to interpret and have got well acceptance to the applied researchers. Linear regression is one of the most popular outlier detection techniques to the computer science and statistical community as well.

A good number of statistical measures have been proposed to study outliers and influence of individual observations in regression analysis. Two different but complementary remedies: robust regression and regression diagnostics with same objectives are well recognized to the statistics community. We make discussion of the most popular outlier detection techniques in linear regression. This is far from exhaustive, but we try to deal with the most popular techniques. Rest of the paper is arranged as in following order.

We briefly describe basics of linear regression, least squares estimation and the idea of unusual (outliers) observations in regression analysis. Classification of outliers is introduced, and consequences of outliers are shown by using a simulated data. We discuss a number of efficient graphical and numerical diagnostic measures. The measures are described and arranged sequentially according to the classification of unusual observations and their structural construction. We use the measures to serve their purposes and to show the efficiency through several artificial and well-referred data sets. Findings and future research issues are attached hereafter.

REGRESSION ANALYSIS AND LEAST SQUARES ESTIMATION

Regression analysis is a statistical technique, which helps us to investigate and to fit an unknown model, quantifies relations among observed variables in multifactor data. More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly,