CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Linear or real line data is a common data found in the literature and in our daily life. However, some data sets are not of linear type such as direction of the wind data set where the observations are measured by angle. Such data is said to belong to circular data category due to the bounded property of circular variable.

A circular observation can be regarded as a point on a circle of unit radius, or a unit vector (i.e. a direction) in a plane. Once an initial direction and an orientation of the circle have been chosen, each circular observation can be specified by the angle from the initial direction to the point on the circle corresponding to the observation. Circular data are usually measured in degree but sometimes it is useful to measure in radians.

Circular data arise in various ways including those corresponding to two circular measuring instruments: the compass and the clock. The observations measured by the compass include the directions of wind and migrating birds. Typical observations measured by clock include the arrival times (on a 24-hour clock) of patient at a casualty unit in a hospital. Data of similar type may correspond to times of year (or times of month) of appropriate events.

Closely related to circular data are the axial data. They are usually given as observations on the circle for which each direction is considered as equivalent to the opposite direction, so that the angles θ and $\theta + 180^0$ are equivalent. The standard way of handling axial data is to convert them to circular data by 'doubling the angles', i.e. transforming θ to 2θ and so removing the ambiguity in direction.

The difference between linear data and circular data can be clearly observed by plotting the following set of data on a linear and a circular plot given in Figures 1.1 and 1.2 respectively:

0 90 180 270 360 720

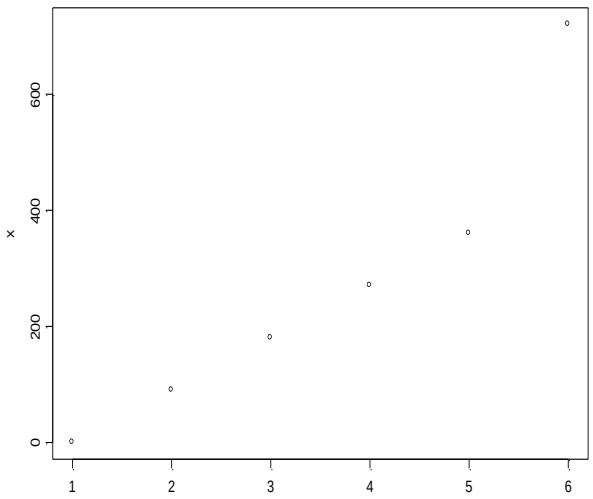

Figure 1.1: Linear plot

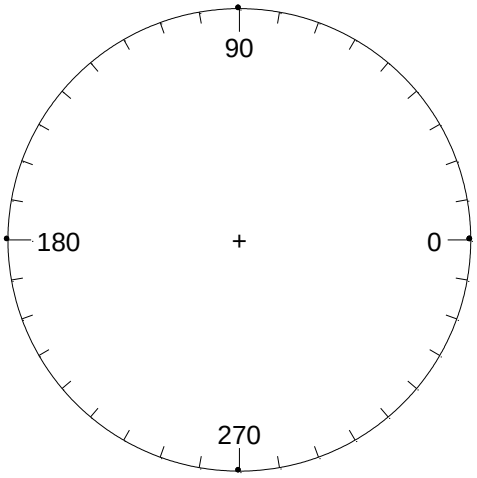

Figure 1.2: Circular plot

Figure 1.1 suggests that all observations are equally separated between each others. The trend for this data is more likely to be increasing linearly. However, if we treat the data as a sample of circular data where observations are in the range 0°-360°, we note that only four points are in the circular plot as shown in Figure 1.2. Remember that, circular data take the value between $[0°,360°)$ or $[0,2\pi)$ in unit radian. Hence, 360° and 720° are located at the same position as observation value of 0°. In general, we have

$$x = x + k(2\pi) \qquad k \in N$$

where $x = [0,2\pi)$. Hence, we need to have different statistical methods when dealing with circular data due to this bounded property of circular variable. Therefore, statisticians have developed different statistical measures and methods to describe and analyse circular data. For examples, we have mean direction and median direction as

measures of central tendency as well as concentration parameter as a measure of dispersion. Details on circular descriptive statistics are outlined in Chapter Two.

In the linear case, different types of distribution are available such as uniform, exponential, gamma, Cauchy, normal and many others. Normal distribution is commonly found in the literature due to its useful properties. On the other hand, von Mises, wrapped stable family including wrapped Cauchy and wrapped normal distributions are the examples of circular distributions. The von Mises distribution is the common distribution being used in circular data and can be considered to be as important as the normal distribution for linear case. The distribution is symmetrical and its shape depends on the mean direction and concentration parameter. Meanwhile, the wrapped Cauchy distribution has a heavy tail with a number of outlying observations in the data. In this study, we focus on two types of distributions only; the von Mises distribution and the wrapped normal distribution.

Regression models are often used to discover the relationship between two variables when both independent and dependent variables are circular. We need to consider special regression model for circular variable. A number of such regression models have been proposed. In 1993, Jammalamadaka and Sarma proposed a circular regression model by utilizing the definition of characteristic function of a complex number which is expressed in Fourier series. Hussin *et al.* (2004) considered a simple circular regression model which attempts to describe a linear relationship between the two circular variables. Recently, Kato *et al.* (2006) had proposed another circular regression model which is expressed as a form of the Mobius transformation. In this study, we consider the circular regression model proposed by Down and Mardia (2002). The proposed regression model is relatively tractable and possesses a one-to-one mapping between the independent angle and the mean of dependent angle.

The existence of outlying values or unexpected observations is one of the important problems in statistical analysis. These values can occur due to error while collecting the data set, or due to natural odd phenomena such as earthquake, war or environmental change. Fisher (1993) summarized ways in which outliers can occur; due to mis-recording, sampling from a second population, or vagaries of sampling resulting in the occasional isolated values. Beckman and Cook (1983) highlighted that such observations, in the opinion of the investigator, stand apart from the bulk of the data. These observations have been called "outliers," "discordant observations," "rogue values," "contaminants," "surprising values," "mavericks," and "dirty data," to mention only a few of the terms that have been coined over the years. Observation which affects the modelling is usually referred as influential observation. They also reviewed the literature on outliers and the available approaches to deal with outliers in different areas of statistics.

Here, the challenge is to detect the existence of outlier in univariate data. The first attempt to develop an objective statistical method to deal with outliers in univariate data was proposed by Peirce (1852). Wright (1884) suggested that any observation whose residual exceeds 3.37 times the standard deviation is to be rejected. There were numerous ways to define the outliers at that time, but the only general agreement was that outliers in a set of data refer to observations which appear to be inconsistent with the remaining observations.

Collett (1979) described the existence of outlier in circular data and pointed out that an extreme numerical value may not imply an extreme angular observation. He also stated that we can only expect to encounter outliers in samples of angular data when the main mass of the data is sufficiently concentrated about a particular direction. In the case of a sample having a low concentration, it would be very difficult to find a single observation which is sufficiently separated from the rest to provide evidence of being an

outlier. To date, there are several numerical statistics such as *C*, *M*, *D* and *A* statistics that are able to detect the existence of outlier in univariate circular data. We will discuss these in detail in Chapter Three.

Another problem of interest is the detection of outliers in circular regression models. This is usually decided by studying their influence on forecasting and modelling. For circular regression models, we use the circular residuals as a tool of detecting outliers. Recently, Abuzaid *et al*. (2010) discussed the issues and proposed several statistics for the above purpose. We will discuss them in Chapter Five.

1.2 Problem Statement

A number of studies have been done on the occurrence of outliers in circular statistics. Several discordancy tests have been developed for circular data. However, no comprehensive study on the performance of the tests when applied on different circular distributions has been carried out. Similarly, no attempt has been made in identifying outliers in circular regression problem beyond the simple circular regression model. Thus, in this study, we investigate the performance of the discordancy tests on a family from the wrapped stable circular distributions via simulation. We also extend the method developed for simple regression model to other circular regression models such as a row deletion approach which is a procedure to investigate the existence of the influential observation.

1.3 Objectives

The objectives of this study include:

(1) To investigate the performance of four discordancy tests when applied to data from the wrapped normal distribution.

(2) To develop procedures for identifying outliers in circular regression models based on the covariance matrix and circular residuals.

(3) To extend the different procedures for identifying outliers in linear regression models to circular regression model such as a row deletion approach.

(4) To apply the tests on real data set.

1.4 Thesis outline

This research attempts to develop statistical methodologies for outlier detection in circular data and circular regression models proposed by Down and Mardia. The research is outlined as follows:

Chapter two provides a literature review on the descriptive statistics of univariate circular data. It introduces several distributions of circular data and their properties.

Chapter three presents the development of four statistical techniques, namely *C*, *M*, *D* and *A* statistics to detect possible outliers in univariate circular data. The cut-off points and the power of performance are discussed.

Chapter four reviews the circular regression model proposed by Down and Mardia (DM regression model) and its covariance matrix. Then, we use the maximum likelihood estimation (MLE) to estimate the parameters of that model.

Chapter five looks at the robustness of the MLE method and the problem of outliers in DM regression models. It applies the *COVRATIO* statistics to detect the existence of

influential observation in DM regression models. Via simulation, the cut off points are obtained and the power of performance is investigated.

Chapter six applies the Difference Mean Circular Error Statistics (*DMCEs*) procedure to detect the presence of outliers in DM regression models. Via simulation, the cut off points are obtained and the power of performance is investigated. This chapter provides a literature review on the existence of outliers in circular regression.

Chapter seven presents the summary of this research work and suggestions for extending the research work.