CHAPTER 1

INTRODUCTION

Multiple linear regression is a widely used statistical technique that allows us to estimate models that describe the distribution of a response variable with the help of a number of other variables usually called explanatory variables, or independent variables. The use of multiple regression mainly regards the interpretation of the regression coefficients. In case of independent coefficients the least-squares solution gives stable estimates and useful results.

However, data are not always "well behaved". We often come across cases where the regressors (explanatory variables) are nearly collinear. This condition is called multicollinearity and is one of the most oftenly encountered in econometrics. The major problem with multicollinearity is that it leads to estimates with inflated variances in the estimation of regression coefficients and thus unacceptably large prediction intervals. High estimated variances (and therefore high estimated standard errors) also mean small observed test statistics. That is the analyst will accept too many null hypotheses. Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model. In addition, the least-squares estimates are usually inflated with wrong signs- though they remain the best linear unbiased estimates (BLUE).

Note that, if the aim of the analyst is to generate forecasts, and if it is assumed that the multicollinearity problem will not be different for the forecast period, then multicollinearity may be considered not to be a problem at all. This is because multicollinearity will not affect the forecasts of a model but only the estimation of the coefficients (Koutsoyiannis, 1977).

In order to detect the presence of collinear variables many diagnostics have been proposed in the literature, for instance the condition number, variance inflation factors, variance decomposition proportions etc. Approaches to remedy the problem of

multicollinearity have also been proposed. Model respecification, variable selection, and biased estimation are some of them.

In 1970 Arthur Hoerl and Robert Kennard published a paper on ridge regression, also known as the biased estimation method, that became the most commonly used method for the remedy of multicollinearity. Hoerl and Kennard's method was in fact a crude form of regularization, a technique developed by Andre Tikhonov (Tikhonov and Arsenin, 1977). Ridge regression involves the introduction of some bias into the regression equation in order to reduce the variance of the estimators of the parameters. The bias is introduced by the use of a ridge constant which controls the extent to which ridge estimates differ from the least squares estimates. Depending on the method used to calculate the optimum ridge constant, i.e. the constant that provides the greatest amount of explained variance in the parameter estimators, different ridge estimators are defined. Ridge estimators have been proposed by many authors. McDonald and Galarneau (1975) proposed an estimator whose squared length equals an estimated squared length of $\boldsymbol{\beta}$. Based on the mean square error property of the ridge estimator Hoerl, Kennard and Baldwin (1975), Guilkey and Murphy (1975), Goldstein and Smith (1974) and others have also proposed ridge estimators. Furthermore, considering the Bayesian approach, Lindley and Smith (1972), Lawless and Wang (1976) and others have also introduced ridge estimators.

The purpose of this thesis is to present the properties of ridge regression as a way to tackle the multicollinearity problem. More specifically, chapter 2 is devoted to the description of multicollinearity. First, we recall the fundamentals of linear regression and provide a description of multicollinearity and its effects. Furthermore, this chapter deals with the diagnostics proposed to detect multicollinearity as well as the remedial measures available, while for better illustration an example is provided.

In chapter 3 we concentrate on ridge regression. Beginning with the reasoning given by Hoerl and Kennard, we proceed to the presentation of some properties of the ridge estimator as well as existence theorems that ensure that the ridge constant always exists. In addition, part of this chapter presents the available methods for calculating the ridge constant as well as the results of ridge regression applied to real data.

Chapter 4 deals with different interpretations of ridge regression as well as its use in cases different from the multiple linear regression model. For example, the use of ridge regression in the logistic model or in simple linear regression with a small number of observations or in the context of generalized linear models. Moreover, we discuss the effects of collinearity when, in addition, influential cases are present in the data set. Finally, chapter 5 provides a simulation experiment for the comparison of certain types of ridge estimators.