HERCMA '98
E. A. Lipitakis (Editor)
1998, LEA
Printed in Hellas

DIAGNOSTIC TECHNIQUES FOR REGRESSION MODELS

Michael Perakis and Evdokia Xekalaki
Department of Statistics
Athens University of Economics and Business
76 Patision St., 10434 Athens, GREECE

Abstract

Diagnostic techniques have been developed for the detection of problems arising in the application of regression models. These problems may be associated with violations of one or more assumptions of the model, with the presence of outliers in the data, with the inappropriate choice of the functional form of the model e.t.c. Most of the proposed diagnostic techniques are suitable for the linear model. However, there are several techniques that may be applied to more complex types of models.Our aim is to present the major developments in the area of diagnostic methods for linear regression models, as well as diagnostic methods which are applicable to some more general types of models. The latter include generalized linear models, nonlinear models and errors in variables models.

Keywords and Phrases: diagnostics; residuals; linear model; generalized linear model; nonlinear model; errors in variables models; heteroscedasticity; autocorrelation; multicollinearity; influence measures

1. Introduction

Diagnostic methods intend to diagnose the existence of some potential problems in a regression model. These problems may be associated with some violations of the assumptions of the model, with the presence of outliers that may change substantially the estimates of the parameters of the model, with the choice of an inappropriate set of explanatory variables or with the choice of the wrong type of model. Most of the proposed diagnostic techniques have been developed for the linear model. However, various other types of models have also been used in the literature and several diagnostic techniques for these models have been developed. A detailed discussion of various techniques that one can apply to various types of regression models is given by Perakis (1997). In this article, we provide a succinct presentation of as many as possible from the numerous techniques that have been developed for four broadly used types of models.

The *linear model*, is defined as $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the $n\times 1$ vector of the dependent variable, $\boldsymbol{\beta}$ is a $p\times 1$ vector of parameters and $\mathbf{X}$ is an $n\times p$ matrix whose columns are the p independent variables. The elements of the $n\times 1$ vector of errors $\boldsymbol{\varepsilon}$ are assumed to be i.i.d. $N(0,\sigma^2)$ random variables (rv's). An extension of the linear model, is the *generalized linear*

model (GLM) (McCullagh and Nelder 1989) in which the distribution of the error term can be any member of the exponential family of distributions. The *normal nonlinear model* is defined as: $y_i=f(\mathbf{x}_i,\boldsymbol{\beta})+\varepsilon_i$, $i=1,...,n$, where y_i denotes the i-th response, $\mathbf{x}_i$ is its corresponding $p\times 1$ vector of explanatory variables and $\boldsymbol{\beta}$ is a $q\times 1$ vector of parameters. The errors ε_i are assumed to be i.i.d. $N(0,\sigma^2)$ rv's.

In an *errors in variables* (EIV) model (Fuller, 1987) it is assumed that all the variables are stochastic. A *linear* EIV model is defined as follows: Let $\boldsymbol{\psi}$ be an error-free response variable and $\boldsymbol{\Pi}=(\boldsymbol{\pi}_1,...,\boldsymbol{\pi}_p)$ be a matrix of error free covariates. The variables of $\boldsymbol{\psi}$ and $\boldsymbol{\Pi}$ are assumed to be related through $\boldsymbol{\psi}=\boldsymbol{\Pi}\boldsymbol{\beta}+\mathbf{q}$, where $\boldsymbol{\beta}$ is a $p\times 1$ vector of parameters and $\mathbf{q}$ is the equation error vector whose elements are i.i.d. rv's with zero mean and variance σ_{qq}. Instead of the unobservable $\boldsymbol{\psi}$ and $\boldsymbol{\Pi}$, we observe $\mathbf{y}$ and $\mathbf{X}$ respectively, so that $\mathbf{y}=\boldsymbol{\psi}+\mathbf{v}$ and $\mathbf{X}=\boldsymbol{\Pi}+\mathbf{U}$. We assume that the elements of $\mathbf{v}$ are i.i.d. rv's with zero mean and variance σ_{vv} and that the rows of $\mathbf{U}$ are independent random vectors with mean the zero vector and covariance matrix $\boldsymbol{\Sigma}_{UU}$.

The residuals play an important role in the development of diagnostic methods as they carry information on the appropriateness of the assumptions of a model. They can be used either in plots depicting their general features or in various types of tests. So, before introducing various types of diagnostic techniques, it is essential to provide a brief description of the five most widely used types of residuals. (A detailed discussion of residuals and their applications is given in Cook and Weisberg (1982) and Atkinson (1985)).

The five most widely used types of residuals for the linear model are: The *ordinary least squares* (OLS) residuals $e_i=y_i-\hat{y}_i$, the *standardized residuals* $e_i'=e_i/[S(1-h_i)]^{1/2}$, where h_i denotes the i-th diagonal element of the hat matrix $\mathbf{H}$ defined as $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and its i-th diagonal element can be used as a leverage measure for the i-th row of the matrix $\mathbf{X}$, $S^2=\mathbf{e}^T\mathbf{e}/(n-p)$, and $\mathbf{e}$ is the vector of e_i's, the *predicted residuals* $e_i^*=y_i-\hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is the predicted value for the i-th observation which is derived after its deletion, the *studentized* residuals $e_i^{*'}=e_i/[S_{(i)}^2(1-h_i)]^{1/2}$, where $S_{(i)}^2$ is given in Atkinson (1985) and the *recursive residuals* $v_i=\left(y_i-\mathbf{x}_i^T\mathbf{b}_{i-1}\right)\Big/\sqrt{1+\mathbf{x}_i^T\left(\mathbf{X}_{i-1}^T\mathbf{X}_{i-1}\right)^{-1}\mathbf{x}_i}$, $i=p+1,...,n$, where $\mathbf{b}_i$ is an estimate of $\boldsymbol{\beta}$ based on the first i observations, $\mathbf{X}_i$ is a matrix that consists of the first i rows of $\mathbf{X}$ and $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$.

In a GLM one can assess three types of residuals (Williams 1987). The i-th *standardized Pearson residual* is defined as $r_{Pi}=(y_i-\hat{\mu}_i)/[v_i(1-h_i)]^{1/2}$, where $\hat{\mu}_i$, v_i are the expected value and the variance of Y_i respectively, both evaluated at $\hat{\boldsymbol{\beta}}$ and h_i is the i-th diagonal element of the weighted hat matrix. The *standardized deviance residuals* are based on the *deviance* defined as $D=\sum d_i^2$, where $d_i^2=2\varphi[\log f(y_i;y_i,\varphi)-\log f(y_i,\hat{\mu}_i,\varphi)]$, f is the distribution of Y_i and φ is its dispersion parameter. The i-th *standardized deviance residual* is then given by $r_{Di}=d_i/[\varphi(1-$

$h_i)]^{1/2}$ and takes the sign of $y_i-\hat{\mu}_i$. Finally, the i-th *likelihood residual* is defined as $r_{Gi}=[(1-h_i)\, r_{Di}^2+h_i\, r_{Pi}^2\,]^{1/2}$ and takes the sign of $y_i-\hat{\mu}_i$. In a nonlinear model the *ordinary residuals* are defined as $e_i=y_i-f(\mathbf{x}_i,\hat{\beta})$. Cook and Tsai (1985) suggested a new type of residuals, termed *projected residuals* and they considered the use of *predicted residuals* in nonlinear models. Finally, the residuals that are used in an EIV model are $\hat{e}_i=y_i-\hat{\beta}^T\mathbf{x}_i$ while Wellman and Gunst (1991) proposed a method for obtaining predicted residuals for such models.

Section 2 is devoted to tests for two of the basic assumptions of the linear model: homoscedasticity and independence of the errors. Section 3 describes the problem of multicollineariry and, finally, Section 4 examines the assessment of the influence of an observation on the estimated vector of parameters.

2. Tests for Homoscedasticity and Autocorrelation

One of the basic assumptions of the linear model, is that the variance of the errors is constant. The term that is used to describe such cases, is *homoscedasticity*. On the contrary, if this assumption is violated then the situation is termed as *heteroscedasticity*. The null hypothesis in all of the tests to be presented in this section is homoscedasticity. The alternative hypothesis is heteroscedasticity, general or of a particular form. As will be seen, the first step in most of the test procedures is the calculation of e_i and $\hat{y}_i$. Several require, an auxiliary regression of these quantities.

The simplest test procedure is to regress e_i^2 on $\hat{y}_i$ (or $\hat{y}_i^2$ or $\log(\hat{y}_i^2)$) and a constant term. Its test statistic is nR^2 and is χ_1^2 distributed. An alternative test is based on the computation of the Spearman's rank correlation coefficients (r_s) of the residuals and each of the covariates. If $n>8$, the test statistic used is $r_s(n-2)^{1/2}/(1-r_S^2)^{1/2}$ and is, under H_0, t_{n-2} distributed. Another test is White's test (see, e.g. Greene, 1993). This is based on fitting an auxiliary regression model of e_i^2 on the variables $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_k$. The test statistic is nR^2 and is asymptotically χ_{k-1}^2 distributed, where k is the number of covariates in the auxiliary model, without the inclusion of the constant. The variables $\mathbf{u}_j$ are the covariates of the initial model their squares and their cross-products.

Often, an auxiliary model is used whose dependent variable is $z_i=\log(e_i^2)$, while the covariates are the p explanatory variables of the initial model. The utilized test statistic is SSR/4.9348, where SSR denotes the regression sum of squares of the auxiliary model and is χ_{p-1}^2 distributed. An alternative test procedure uses an auxiliary model whose the dependent variable is $z_i=|e_i|$ and its covariates coincide with those of the original model. The test statistic is $SSR/[(1-2/\pi)S^2]$, where S refers to the initial and SSR to the auxiliary model. Its distribution is the χ_{p-1}^2. A further test proposed is based on the recursive residuals whose test

statistic is: $\sum_{i=n-m+1}^{n} v_i^2 \Big/ \sum_{i=p+1}^{p+m} v_i^2$ that under the null hypothesis is $F_{m,m}$ distributed. The choice of m is arbitrary, but the value (n-p)/3 is considered as a "good" choice.

As cited above, a null hypothesis is often tested against an alternative of a specific form of heteroscedasticity. Park's test (see, e.g. Gujarati, 1978) deals with $H_1: \ln(\sigma_i^2)=\ln(\sigma^2)+\beta\ln(x_i)+v_i$ and involves the estimation of the parameters of the regression equation $\ln(e_i^2)=\alpha+\beta\ln(x_i)+v_i$. If β is significantly different from zero, the null hypothesis is rejected. An alternative test with a specific alternative hypothesis was proposed by Breusch and Pagan (see, e.g. Greene, 1993). The null hypothesis is tested versus the hypothesis that $\sigma_i^2=g(\gamma_0+\gamma_1 z_{i1}+...+\gamma_p z_{ik})$, where g is an arbitrary function. The variables $\mathbf{z_j}$ are usually the covariates of the initial model. The test procedure of Breusch and Pagan involves the utilization of the regression of $e_i^2/\hat{\sigma}^2$ on $\mathbf{z_j}$, j=1,...,k, where $\hat{\sigma}^2=n^{-1}\sum e_i^2$. The test statistic is SSR/2 and, under the null hypothesis, is χ_k^2 distributed. A drawback of this test is its sensitivity to violations of the assumption of normality of the errors. Because of this sensitivity Koenker and Bassett (see, e.g. Greene, 1993) suggested a modification of the test statistic.

Finally, other useful homoscedasticity tests, are the Goldfeld-Quandt test (see, e.g. Greene, 1993), used when one is interested in detecting whether the variances of the errors can be written as $\sigma_i^2=\sigma^2 x_i^2$ for a particular dependent variable **x** and the Bartlett homoscedasticity test (see, e.g Kmenta, 1986) used in cases where there are several available observations on **y**, for each different value of **x**.

A further assumption of the linear model is that the successive values of the random variable ε, which refers to the error term, are independent. If this assumption is not satisfied, then we say that there exists *autocorrelation* in the errors. The term autocorrelation refers to a special case of correlation, since it refers to the relationship between the successive values of the same variable. In the tests that are discussed in the sequel, the null hypothesis is that there is no autocorrelation and the alternative is that the errors are autocorrelated. In some of the tests the form of autocorrelation is specified and it may be that the errors constitute an autoregressive (AR) or a moving average (MA) process of order h.

The most commonly used test for autocorrelation is the *Durbin-Watson test* (see, e.g. Greene, 1993) whose test statistic is $d=\sum_{i=2}^{n}(e_i-e_{i-1})^2 \Big/ \sum_{i=1}^{n} e_i^2$. Its alternative hypothesis is that the errors are AR(1). The values of d are between 0 and 4. A value of d close to 0 (4), is indicative of positive (negative) autocorrelation while if d is close to 2, no autocorrelation can be deduced. The exact distribution of d cannot be tabulated, but it can be shown that the true distribution of d lies between the distributions of two other rv's that can be tabulated. The next test that is presented was designed by Berenblut and Webb (see, e.g. Judge et al., 1985) for

H_1:AR(1). The test statistic is g= RSS^*/RSS, where RSS is the residual sum of squares of the original model and RSS^* is the residual sum of squares of the regression of the first differences of y on the first differences of **X** without a constant term. If the initial model contains a constant term, then the bounds of the Durbin-Watson test can be used. Note that this test is more powerful than the Durbin-Watson test when the autocorrelation is high.

Breusch and Godfrey proposed a *Lagrange Multiplier test* (see, e.g. Greene, 1993), for testing the null hypothesis, versus the alternative that ε_i=AR(h) or ε_i=MA(h). According to this test we regress e_i on $\mathbf{x_j}$'s and $e_{i-1},\ldots,e_{i-h}$. The values of the residuals that are missing are replaced by zeros. The test statistic is nR^2 and is distributed according to the χ^2_h distribution. An alternative test for autocorrelation was proposed by Box and Pierce (*Q-test*) (see, e.g. Kmenta, 1986). The alternative hypothesis of this test is that ε_i=AR(h) or ε_i=MA(h) and its test statistic is $Q=n\sum_{j=1}^{L} r_j^2$, where $r_j=\sum_{t=j+1}^{n} e_t e_{t-j} \Big/ \sum_{t=1}^{n} e_t^2$. The distribution of Q under the null hypothesis is the χ^2_L distribution. A drawback of this test, is that the choice of L is arbitrary. A modification that makes the Q-test more powerful was proposed by Box, Pierce and Ljung (see, e.g. Kmenta, 1986) who proposed the use of $Q'=n(n+2)\sum_{j=1}^{L} r_j^2/(n-j)$ instead of Q.

Durbin developed a test for models of the form $y_i=\alpha+\boldsymbol{\beta}^T\mathbf{x}_i+\gamma y_{i-1}+\varepsilon_i$. The test (see, e.g. Kmenta 1986) is known as *Durbin's h test*, its H_1: ε_i=AR(1) and its test statistic is $h=(1-d/2)[n/(1-ns_c^2)]^{1/2}$, where d is the Durbin-Watson test statistic, and s_c^2 is the variance of the estimated coefficient of the lagged variable. Its asymptotic distribution under H_o is N(0,1). An alternative test procedure, which is also used when there is a lagged dependent variable and performs better than Durbin's h test (see, e.g. Kmenta 1986), is the *m test*. Its alternative hypothesis is that ε_i=AR(h). According to this test we regress e_i on $\mathbf{x_i}$, y_{i-1}, $e_{i-1},\ldots,e_{i-h}$ and a constant and we do not reject the null hypothesis if the coefficients of $e_{i-1},\ldots,e_{i-h}$ are zero.

We conclude this section with three tests of randomness of the residuals. According to the *turning point test* we count the number of turning points where e_i is a turning point if $e_{i-1}<e_i$ and $e_i>e_{i+1}$ or if $e_{i-1}>e_i$ and $e_i<e_{i+1}$. If T is the number of turning points, its asymptotic distribution under H_0 is $N(\mu_T,\sigma_T^2)$, where $\mu_T=2(n-2)/3$ and $\sigma_T^2=(16n-29)/90$. In the *difference sign test* a count of the number S of instances where $e_i>e_{i-1}$ is made. Under H_0, the asymptotic distribution of S is normal with mean $\mu_S=(n-1)/2$ and variance $\sigma_S^2=(n+1)/12$. Finally, the test statistic of the *recursive t-test* is $[\hat{\sigma}^2(n-p)]^{-1/2}\sum_{i=p+1}^{n} v_i$, where v_i is the i-th recursive residual and $\hat{\sigma}^2=\frac{1}{n-p-1}\sum_{i=p+1}^{n}(v_i-\bar{v})^2$. Its distribution is t_{n-p-1}.

4. Tests for Multicollinearity

A problem with which one may often be faced is that of *multicollinearity* which refers to high correlation among the covariates that does not allow one to examine the individual effect of each of them. In the presence of multicollinearity the estimates of the parameters are not stable, they have very large standard errors and their coefficients may have the wrong sign. If there exists multicollinearity, the matrix of regressors $\mathbf{X}$ is not of full rank. There are several methods that one can use in order to determine whether there is multicollinearity.

Judge et al. (1985) propose three such methods. The first is based on the calculation of the correlation coefficient (r) between any pair of covariates. A value of r far from zero is an indication of the presence of multicollinearity. Furthermore, one can compare these coefficients with the R^2 of the model. If any of the r's is greater than R^2, there may be multicollinearity. A disadvantage of this method is that it cannot detect multicollinearity that involves more than two variables. The second method is based on the comparison of the t-test, which tests if a parameter equals zero, to the F-test, which tests if all the parameters equal zero. The case where none of the null hypotheses of the t-test is rejected, while at the same time the null hypothesis of the F-test is rejected, may be indicative of the presence of multicollinearity. Nonetheless, this method does not always detect multicollinearity and does not provide any information on the involved regressors. According to the third method we regress $\mathbf{y}$ on all the sets of p-1 regressors that can be selected from the p regressors. If there is multicollinearity, the R^2 of the full model, must be very close to the highest of the R^2 coefficients of the partial models. A drawback of this method is that it does not show the relationship among the multicollinear variables. Furthermore, a small difference between the R^2 of the initial and the partial models may be the result of the wrong choice of regressors.

Another method proposed by Klein (see, e.g. Greene 1993), leads to the calculation of R_j^2, j=1,...,p, where R_j^2 is the R^2 of the regression of the j-th covariate on all the remaining ones. If any of these coefficients is greater than the R^2 of the examined model, there may exist multicollinearity. If the multicollinearity involves only a few variables, the estimated coefficients may reveal its form. However, if many regressors are involved then the same problem is faced in the fitting of the auxiliary model.

The next approach for the detection of multicollinearity is based on the *determinant* of the matrix $\mathbf{X}^T\mathbf{X}$ (after the normalization of $\mathbf{X}$). The value of $|\mathbf{X}^T\mathbf{X}|$ is between 0 and 1 and if the columns of $\mathbf{X}$ are orthogonal, it is close to 1, while if the columns of $\mathbf{X}$ are multicollinear it is close to 0. A disadvantage of this measure is that it does not provide information on the form of the present linear dependencies. Another method that involves the use of the matrix $\mathbf{X}^T\mathbf{X}$ is the calculation of its *condition number*. The condition number of a matrix is the square root of the ratio of its largest to its smallest eigenvalue, i.e. $(\lambda_{max}/\lambda_{max})^{1/2}$. Evidently, it is always greater than or equal to 1 and it is equal to 1 if all the eigenvalues are equal. If there is high correlation

among the regressors, the condition number becomes large. As Belsley et al. (1980) points out, the condition number is greater than 30, some problems are likely to arise due to the presence of multicollinearity.

Finally, a three-step procedure for testing multicollinearity hypotheses was proposed by Farrar and Glauber (see, e.g. Koutsoyiannis 1977). The first step amounts to checkning if there is mutlicollinearity among the p variables, the second to determining which of the regressors are collinear and the last step to determining the variables that cause multicollinearity.

A remedy for the problem of multicollinearity is achieved by the use of the ridge regression estimator. Although this estimator is biased, it is preferable to the maximum likelihood estimator because its errors are much smaller. The ridge regression estimator is defined as $\mathbf{b}_r=(\mathbf{X}^T\mathbf{X}+r\mathbf{D})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{D}$ is a p×p diagonal matrix whose diagonal elements are the diagonal elements of the matrix $\mathbf{X}^T\mathbf{X}$. The choice of r is arbitrary. In practice, one starts from a small value and then increases r slightly until the estimates of the parameters become stable.

Weissfeld and Sereika (1991) suggested a method for testing for multicollinearity in a GLM which is based on the observed information matrix. Finally, Gamboa and Gunst (1992) proposed a method that enables us to examine the existence of collinearity in EIV models and Rasekh and Fieller (1995) proposed two ridge estimators which can be used in such cases.

5. Assessment of Influence

The influence that an observation has on the estimates of the parameters of a model is certainly of great interest and has motivated research efforts towards developing diagnostic tools for its detection. The basic influence measures are based on the deletion of the observation whose influence has to be examined. The statistic that Cook proposed for examining this influence is known as *Cook's distance* (see, e.g. Cook and Weisberg, 1982) and is defined as $D_i=e_i'^2 h_i/[p(1-h_i)]$. There is also a modified form of the Cook statistic (see, e.g. Atkinson 1985) given by $C_i=\{[(n-p)/p][h_i/(1-h_i)]\}^{1/2}\,|e_i^{*'}|$. An alternative influence measure can be obtained via a slight modification of C_i. The resulting measure is known as DFFITS (Belsley et al., 1980) and is defined as $DFFITS_i=[h_i/(1-h_i)]^{1/2}e_i^{*'}$. In order to determine the most influential observations, a list or a plot of the values of the preferred influence measure is constructed. If D_i (or equivalently C_i or DFFITS) is large, the i-th observation has high influence on the estimates of the vector $\boldsymbol{\beta}$. Andrews and Pregibon proposed an alternative influence diagnostic measure (see, e.g. Cook and Weisberg 1982). If $\mathbf{X}^*=(\mathbf{X},\mathbf{y})$, i.e. $\mathbf{X}^*$ is the matrix $\mathbf{X}$ with one additional column which is the vector $\mathbf{y}$, the proposed diagnostic measure is defined as $R_I=|\mathbf{X}_{(I)}^{*\,T}\mathbf{X}_{(I)}^{*}|/|\mathbf{X}^{*T}\mathbf{X}^{*}|$, where I denotes the cases whose influence is examined and $\mathbf{X}_{(I)}^{*}$ is the matrix $\mathbf{X}^*$ after the deletion of the I rows.

Atkinson (1985) and Cook and Weisberg (1982) considered the assessment of influence in the context of nonlinear models and they proposed two alternative influence measures which are based on a linearization of it. Finally, Wellman and Gunst (1991) proposed influence measures for EIV models.

References

Atkinson A.C. (1985). *Plots, Transformations and Regression*, Clarendon Press.

Belsley D.A., Kuh E. and Welsch R.E. (1980). *Regression Diagnostics*, Wiley.

Cook R.D. and Weisberg S. (1982). *Residuals and Influence in Regression*, Chapman and Hall.

Cook R.D. and Tsai C.L. (1985). Residuals in Nonlinear Regression, *Biometrika*, **72**, 23-29.

Fuller W.A. (1987). *Measurement Error Models*, Wiley.

Gamboa O.C. and Gunst R.F., (1992). Measurement Error Model Collinearities, *Technometrics*, **34**, 454-464.

Greene W.H. (1993). *Econometric Analysis*, Macmillan.

Gujarati D. (1978). *Basic Econometrics*, McGraw-Hill.

Judge G., Griffiths W., Hill R., Lutkepohl H. and Lee T. (1985). *The Theory and Practice of Econometrics*, Wiley.

Kmenta J. (1986). *Elements of Econometrics*, Macmillan.

Koutsoyiannis A. (1977). *Theory of Econometrics*, Macmillan.

McCullagh P. and Nelder J.A. (1989). *Generalized Linear Models, Second edition*, London: Chapman and Hall.

Perakis M. (1997). Diagnostic Techniques for Regression Models, M.Sc. Thesis, Athens University of Economics and Business, Department of Statistics.

Perakis M. and Xekalaki E. (1998). Diagnostic Techniques for Regression Models, *Technical Report*, **55**, Department of Statistics, Athens University of Economics and Business.

Rasekh A.R. and Fieller N.R.J., (1995). Ridge Estimation in Functional Errors in Variables Models, Research report No. 447/95, *University of Sheffield.*

Weissfeld L.A. and Sereika S.M., (1991). A Multicollinearity Diagnostic for Generalized Linear Models, *Communications in Statistics - Theory and Methods*, **20(4)**, 1183-1198.

Wellman J.M. and Gunst R.F.,(1991). Influence Diagnostics for Linear Measurement Error Models, *Biometrika*, **78**, 373-380.

Williams D.A. (1987). Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletion, *Applied Statistics*, **38**, 181-191.