The Korean Communications
in Statistics Vol. 8, No. 2, 2001
pp. 311-317

Regression Diagnostics Using Residual Plots[1)]

Kwang-Sik Oh[2)]

Abstract

It is necessary to check the linearity of selected covariates in regression diagnostics. There are various graphical methods using residual plots such as partial residual plots, augmented partial residual plots and combining conditional expectation and residual plots. In this paper, we propose the modified pseudolikelihood ratio test statistics based on these residual plots to test linearity of selected covariate. These test statistics which measure the distance between the nonparametric and parametric models are derived as a ratio of quadratic forms. The approximate distribution of these statistics is calculated numerically by using three moments. The power comparision of these statistics is given.

Keywords : Residual Plots, CERES plots, pseudolikelihood ratio test,

1. Introduction

The problem of checking the linearity of selected covariates is fundamental in regression diagnostics and there are various methods of formal and informal approachs. The common aim is to examine the relationship between the residuals from a linear model and some selected covariate for patterns indicating non-linearity. A popular formal method is the Durbin-Watson test whose original motivation was to detect first-order autocorrelation in errors. However if a linear model is fitted to a non-linear relationship then the patterns in the residuals are in practice indistinguishable from those of positive autocorrelation, and so the test is often used to detect non-linearity by replacing ordering in time with ordering in the covariate. Munson and Jernigan(1989) develop this technique into a curvature test by measuring the roughness of a spline which interpolates the residual plot. Azzalini and Bowman(1993) proposed a pseudolikelihood ratio test which measures the distance between the true regression function and a fitted parametric model. Here the true regression function is estimated by nonparametric smoothing.

Informal approachs are based on graphical methods and particular on various kinds of

1) This research is supported by Catholic University of Daegu in 1999.
2) Professor, Department of Informational Statistics, Catholic University of Daegu, 712-702.
E-mail : ohkwang@cuth.cataegu.ac.kr

plots such as added-variable plots, component-plus-residual (C+R) plots, which are known as partial residual (PRES) plots, augmented partial residual (APRES) plots, and combining conditional expectation and residuals (CRRES) plots. Nonparametric regression can be used as a visual means of assisting the detection of trends in residual plots, as illustrated by LOESS of Cleveland(1979) or by kernel.

In this paper we investigate residual plots and apply these plots to the pseudolikelihood ratio test which was proposed by Azzalini and Bowman(1993). Power comparision is givn by simulation. In Section 2 we explore various residual plots. The modified pseudolikelihood ratio test statistics are suggested in Section 3. Simulation study and conclusions are given in Section 4.

2. Residual Plots

Suppose that the "true" model is

$$y = \alpha_0 + \boldsymbol{\alpha}_1^T \boldsymbol{x}_1 + g(x_2) + \varepsilon \quad (2.1)$$

where y is response, $p\times 1$ covariates $\boldsymbol{x}$ partitioned into a $(p-1)\times 1$ covariates vector $\boldsymbol{x}_1$ and a single covariate x_2, g is an unknown function with $E(g)=0$, $E(\varepsilon)=0$, and $Var(\varepsilon)=\sigma^2$. Many authors considered the following linear regression model to detect curvature of $g(.)$:

$$y = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_1 + \beta_2 x_2 + error \quad (2.2)$$

Let e denote OLS residual, $\boldsymbol{b}_1$ and b_2 denote OLS estimator of $\boldsymbol{\beta}_1$ and β_2, $e_{y|1}$ denote a residual from the OLS regression of y on $\boldsymbol{x}_1$, and $e_{2|1}$ denote a residual from the OLS regression of x_2 on $\boldsymbol{x}_1$.

An added-variable plot for x_2 is a two-dimensional plot of $e_{y|1}$ versus $e_{2|1}$. The slope of the OLS regression of this plot is b_2 and the residuals from this regression are the same as the residuals e from model (2.2). Thus, the presence of curvature in this plot means that there is curvature in the plot e versus $e_{2|1}$, which has implications for the lack of fit of model (2.2). According to Cook(1996), added-variable plots seem particularly well suited for studying influence, and curvature in such plots is an indication of a model deficiency. Added-variable plots should not be used as basis for selecting a covariate transformation or for diagnosing curvature. The absence of curvature in an added-variable plot should not necessarily be taken as assurance that $g(.)$ is linear.

A component-plus-residual (C+R) plot of x_2, which are also known as a partial residual (PRES) plot, is a plot of $e + b_2 x_2$ versus x_2. This plot should perform well when the model (2.2) is a good description of the data or when the conditional expectations $E(x_j | x_2)$,

j=1,...,p-1, are all essentially linear. But nonlinear relationships among the covariates are also a problem for PRES plots.

An augmented partial residual (APRES) plot for x_2 suggested by Mallows(1986) is constructed from the model :

$$y = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_1 + \beta_2 x_2 + \beta_{22} x_2^2 + error \quad (2.3)$$

The APRES plot for x_2 is a two-dimensional plot of $e + b_2 x_2 + b_{22} x_2^2$ versus x_2. This plot might be expected to give a more accurate depiction of $g(.)$ than a partial residual plot, at least when $g(.)$ is strongly quadratic or when $E(x_j | x_2)$ is nonlinear.

A CERES plot for x_2 suggested by Cook(1993) is constructed from the model :

$$y = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_1 + \boldsymbol{\beta}_2^T \boldsymbol{m}(x_2) + error \quad (2.4)$$

where $\boldsymbol{m}(x_2) = E(\boldsymbol{x}_1 | x_2) - E(\boldsymbol{x}_1)$. The CERES plot for x_2 is a plot of $e + \boldsymbol{b}_2^T \boldsymbol{m}(x_2)$ versus x_2. This class of plots was called by CERES plots because of "combining Conditional Expectation and RESiduals". To obtain CERES plots, $\boldsymbol{m}(x_2)$ could be estimated by using nonparametric regression. An estimate $\widehat{E}(x_{1j} | x_2)$ of the j-th coordinate of $E(\boldsymbol{x}_1 | x_2)$ can be obtained by extracting the fitted values based on smoothing the plot of x_{1j} versus x_2. The nonparametric regression estimates $\widehat{E}(x_{1j} | x_2)$, j=1,...,p-1, are then used as a replacement for $E(\boldsymbol{x}_1 | x_2)$ in (2.4) and the CERES plot constructed in the usual way. We can obtain the estimate of $\boldsymbol{m}(x_2)$ by LOESS and nonparametric kernel methods in this paper. This plot might be useful for obtaining an impression of curvature of $g(.)$ when $E(\boldsymbol{x}_1 | x_2)$ are neither linear nor quadratic.

3. Testing linearity in residual plots

Since residual plots are useful for obtaining an impression of non-linearity, we can apply the test of Azzalini and Bowman(1993) to the problem of testing linearity of selected covariate x_2 using residual plots. We consider PRES(APRES or CERES, similary) vector as our new response vector $\boldsymbol{y}^*$ and assume the model as follows,

$$y_i^* = g(x_{2i}) + \varepsilon_i, \quad i=1,...,n \quad (3.1)$$

where y_i^* is the i-the element of PRES(APRES or CERES) vector $\boldsymbol{y}^*$ and ε_i are iid from unknown distribution with mean 0 and finite variance σ^2. The function $g(.)$ can be estimated without making parametric assumption on shape by employing nonparametric smoothing. The kernel approach provides a simple estimator through the formular

$$\hat{g}_n(x:h) = \sum_{i=1}^{n} [\, y_i^* K(\frac{x - x_i}{h}) / \sum_{j=1}^{n} K(\frac{x - x_j}{h})] \qquad (3.2)$$

where $K(.)$ is a kernel function and the smoothing parameter h controls the degree of smoothing which depends on n and trends to 0 as $n \to \infty$ but $nh \to \infty$. For convenience, the numerical work in this paper assumes the weight function to be the standard normal density. Because the estimator $\hat{g}$ is linear in the response variable y^*, the fitted values

$$\hat{g} = (\hat{g}(x_1), \ldots, \hat{g}(x_n)) = Wy^* \qquad (3.3)$$

where W is an $n \times n$ matrix of constants depending on x, h and $K(.)$.

The aim of this section is to assess whether model (3.1) can be reduced to the simple linear form. The pseudolikelihood ratio approach arises from the formal expression of the likelihood ratio for the hypotheses

$$H_0 : g(x) = a + bx \quad \text{for some } a \text{ and } b$$

$$H_1 : g(x) \text{ is a smooth fuction.}$$

The likelihood under H_0 is evaluated at $g(x) = \hat{a} + \hat{b}x$ where $\hat{a}, \hat{b}$ are LSE, and the likelihood under H_1 is evaluated at $g(x) = \hat{g}(x)$ in (3.3). After a standard transformation, we led to a test statistics of the form

$$F = \frac{(RSS_0 - RSS_1)}{RSS_1}$$

where RSS_0, RSS_1 denote the residual sums of squares after fitting the linear and smooth models respectively. An expression for F is given by

$$F = \frac{(\, y^{*\prime} M_0 \, y^* - y^{*\prime} M_1 \, y^*)}{y^{*\prime} M_1 \, y^*}$$

where $M_0 = I - X\,(X'X)^{-1}X'$, $M_1 = (I - W)'(I - W)$ and X is the design matrix which has first column $\mathbf{1}_n$ and second column $\boldsymbol{x}$. Since $(I - W)'\,\mathbf{1}_n = \mathbf{0}$, the distribution of F is free from a. But F is not free from b because of $(I - W)'\,\boldsymbol{x} \neq \mathbf{0}$. Therefore, we would like to construct a formal test of linearity based on F, large value being significant, the dependence of the distribution of F on the unknown parameter b makes F unsuitable for hypothesis testing. To overcome the above problem, Azzalini and Bowman(1993) consider the vector $\boldsymbol{e}^* = M_0\, \boldsymbol{y}^*$ and apply the pseudolikelihood ratio principle to $\boldsymbol{e}^*$ instead of $\boldsymbol{y}^*$.

The modified pseudolikelihood ratio now arises formally from the hypotheses

$$H_0^* : E(\, \boldsymbol{e}^*) = \mathbf{0} \text{ for some } x_i$$

$$H_1^* : E(\, \boldsymbol{e}^*) \text{ is a smooth function of the } x_i \text{ values,}$$

and leads to the test statistic

$$F^* = \frac{(\boldsymbol{e}^{*\prime}\boldsymbol{e}^* - \boldsymbol{e}^{*\prime}M_1\boldsymbol{e}^*)}{\boldsymbol{e}^{*\prime}M_1\boldsymbol{e}^*} \qquad (3.4)$$

Large values of this statistic are significant.

Since this statistic is similar to the statistic of Azzalini and Bowman (1993), we can obtain the probability by calculating numerically. Under a normal assumption on the distribution of $\boldsymbol{e}^*$ and

$$P(F^* > t) = P(\boldsymbol{e}^{*\prime}(I-(1-t)M_1)\boldsymbol{e}^* > 0) = P(\boldsymbol{e}^{*\prime}A\boldsymbol{e}^* > 0) \qquad (3.5)$$

the problem is reduced to the computation of the distribution of a quadratic form of normal variates which under the null hypotheses have mean $\mathbf{0}$ and $cov(\boldsymbol{e}^*) = M_0$ because the scale parameter σ may be set equal to 1 without loss of generality since statistic (3.4) is scale invariant. Johnson and Kotz (1972) provide some standard results on the distribution of quadratic forms like $\boldsymbol{e}^{*\prime}A\boldsymbol{e}^*$. Specifically, the evaluation of equation (3.5) can be shown to be equivalent to the computation of distribution of linear combination of independent χ^2 -variates with coefficients given by the eigenvalues of M_0A. However, for most practical purposes, the exact evaluation of (3.5) is at the same time unnecessary and computationally burdensome. It is widely recognized that one can obtain a reasonably accurate approximation to a distribution function replacing the original distribution by a member of a parametric class by matching the first three or four moments with the original ones. In our case, the s-th cumulant of $\boldsymbol{e}^{*\prime}A\boldsymbol{e}^*$ is

$$k_s = 2^{s-1}(s-1)!\,tr(M_0A)^s \qquad (3.6)$$

which does not require explicit computation of the eigenvalues of M_0A. And the s-th cumulant of $a+b\chi^2_c$ is $2^{s-1}cb^s(s-1)!$ for $s \geq 2$ and $cb+a$ for $s=1$. Let first three cumulants of $a+b\chi^2_c$ and those of $\boldsymbol{e}^{*\prime}A\boldsymbol{e}^*$ be equal, then we can calculater $P(a+b\chi^2_c > 0)$. Thus we can obtain the required distribution.

4. Simulation Study and Conclusion

A small power study was carried out to compare the performance of the modified pseudolikelihood ratio test. Data were simulated from the model

$$y = x_1 + x_2 + g(x_3), \quad g(x_3) = \frac{a}{1+e^{-x_3}} \qquad (4.1)$$

where the x_3 are generated from a uniform random variable on the interval (0, 30), $x_1 = x_3^{-1} + N(0, 0.1^2)$ and $x_2 = \log(x_3) + N(0, 0.25^2)$. No error was included so that y is a deterministic function of the three covariates. This allows the conclusions to

be illustrated more clearly than if an additive error were included but will not change the qualitative nature of the results. Sample sizes n=10, 15, 20, 25 were considered. In each case 1000 samples were generated and the proportion of times that the observed significance was below a= 0.05, 0.1 was counted. The bandwidth of kernel function is 0.5 and the number of smoothing data in LOESS is fixed at n/2 for obtaining conditional expectation in CERES. The bandwidth of kernel function for obtaining test statistic is 0.25. We choose a= 1.94, 9.68, 19.36, 96.82 in curve $g(.)$ using roughness measure $\int (g'')^2$.

We use the notations RES, PRES, APRES, CERES(L), and CERES(k) for residual, partial residual, augmented partial residual, conditional expectations residual based on LOESS, and conditional expactations residual based on kernel curve respectively. We show the empirical power obtained by the modified pseudolikelihood ratio test statistics in Table 4.1. In this table, RES means the test statistic of Azzalini and Bowman. In Table 4.1 the test statistic of Azzalini and Bowman is highly affected by roughness measure a , but the other test statistics are not affected by a . As the sample size is larger, the empirical power of all statistics is higher. In case $a=1.94$, the empirical power of the modified pseudolikelihood ratio test statistics is higher than that of Azzalini and Bowman. And the empirical power of modified pseudolikelihood ratio test statistics based on CERES(k) is highest.

In conclusion, the modified pseudolikelihood ratio test statistics based on CERES(k) is most sensitive to non-linearity. Thus we suggest the modified pseudolikelihood ratio test statistics based on CERES(k) to check the linearity of selected covariates in regression diagnostics.

Table 4.1 Comparing empirical power of the modified pseudolikehood ratio test statistics

alpha=0.1						
n	a	RES	PRES	APRES	CERES(L)	CERES(k)
10	1.94	0.247	0.346	0.593	0.809	0.951
	9.68	0.484	0.319	0.562	0.802	0.936
	19.36	0.646	0.327	0.578	0.799	0.929
	96.82	0.853	0.340	0.576	0.805	0.938
15	1.94	0.296	0.420	0.648	0.864	0.990
	9.68	0.653	0.412	0.636	0.855	0.980
	19.36	0.784	0.430	0.670	0.865	0.987
	96.82	0.956	0.428	0.659	0.876	0.981
20	1.94	0.357	0.513	0.740	0.931	0.998
	9.68	0.767	0.510	0.742	0.929	0.999
	19.36	0.877	0.473	0.746	0.941	0.997
	96.82	0.979	0.507	0.745	0.937	0.998
25	1.94	0.483	0.612	0.813	0.968	1.000
	9.68	0.866	0.612	0.816	0.958	1.000
	19.36	0.959	0.603	0.759	0.962	0.999
	96.82	0.997	0.609	0.799	0.967	1.000

alpha=0.05						
n	a	RES	PRES	APRES	CERES(L)	CERES(k)
10	1.94	0.183	0.263	0.507	0.763	0.924
	9.68	0.407	0.247	0.489	0.747	0.910
	19.36	0.588	0.241	0.496	0.732	0.898
	96.82	0.828	0.254	0.491	0.753	0.910
15	1.94	0.203	0.316	0.571	0.819	0.983
	9.68	0.561	0.316	0.560	0.798	0.970
	19.36	0.730	0.320	0.597	0.820	0.978
	96.82	0.942	0.328	0.581	0.823	0.970
20	1.94	0.248	0.402	0.643	0.907	0.997
	9.68	0.693	0.406	0.642	0.894	0.994
	19.36	0.843	0.371	0.645	0.905	0.995
	96.82	0.972	0.408	0.660	0.909	0.996
25	1.94	0.362	0.480	0.728	0.944	0.998
	9.68	0.792	0.499	0.726	0.930	1.000
	19.36	0.924	0.498	0.719	0.936	0.997
	96.82	0.955	0.480	0.729	0.948	1.000

References

[1] Azzalini, A. and Bowman, A.W.(1993). On the use of nonparametric regression for checking linear relationships. *Journal of The Royal Statistic Society* B.,55, No.2, 549-557.

[2] Cook, R.D.(1993). Exploring partial residual plots. *Technometrics*, 35, 351-362.

[3] Cook, R.D.(1996). Added-Variable Plots and Curvature in Linear Regression. *Technometrics*, 38, 275-278.

[4] Cook, R.D. and Weisberg, S.(1982). *Residuals and influence in regression.* London: Chapman and Hall.

[5] Johnson, N.L. and Kotz, S.(1972). *Distributions in statistics: Continuous multivariate distributions.* New York: Wiley.

[6] Larsen, W.A. and McCleary, S.J.,(1972). The use of partial residual plots in regression analysis. *Technometrics*, 14, 781-790.

[7] Mallows, C.L.(1986). Augmented partial residual plots. *Technometrics*, 28, 313-320.

[8] Munson, P.J. and Jernigan, R.W.(1989). A cubic spline extension of the Durbin-Watson test. *Biometrika*, 76, 39-47.