EXAMINATION AND ANALYSIS OF RESIDUALS: DIAGNOSTIC CHECKING OF RESIDUALS FOR DETECTING A SPECIAL TYPE OF HETEROSCEDASTICITY IN LINEAR REGRESSION THROUGH THE ORIGIN

A. Hedayat
Cornell University, Ithaca, New York

Abstract

The main purpose of this paper is to use the Theil residuals for detecting a monotonic relation between mean and variance by means of the peak test applied to linear regression through the origin. A numerical illustration is provided.

Paper No. BU-213 in the Biometrics Unit Series, and No. 597 in the Plant Breeding Department Series, Cornell University, Ithaca, New York.

EXAMINATION AND ANALYSIS OF RESIDUALS: DIAGNOSTIC CHECKING OF RESIDUALS FOR DETECTING A SPECIAL TYPE OF HETEROSCEDASTICITY IN LINEAR REGRESSION THROUGH THE ORIGIN

A. Hedayat
Cornell University, Ithaca, New York

SUMMARY

The main purpose of this paper is to use the Theil residuals for detecting a monotonic relation between mean and variance by means of the peak test applied to linear regression through the origin. A numerical illustration is provided.

1. INTRODUCTION

Consider the simple linear model $Y = X\beta + \epsilon$ where Y represents an n-dimensional random vector, X is an n-dimensional column vector with known coefficients consisting of nonstochastic elements, or elements which are distributed independently of the error terms, β is an unknown scalar, and ϵ is an n-dimensional random vector having multivariate normal distributions with

$$E\epsilon = 0, \quad E\epsilon\epsilon' = \sigma^2 I_n \tag{1}$$

where $\sigma^2 > 0$ is an unknown parameter and I_n is used to denote the $n \times n$ identity matrix.

Research was supported by NIH GM-05900

The least square (LS) estimates $\hat{\beta}$ of β and $\hat{\epsilon}$ of ϵ are

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2$$

and

$$\hat{\epsilon} = Y - X\hat{\beta} = Y - X\left(\sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2\right) = PY$$

where

$$P = I_n - \left(1/\sum_{i=1}^{n} x_i^2\right) XX' \quad .$$

Under the above assumptions

$$E\hat{\epsilon} = PX\beta = 0\beta = 0$$

$$E\hat{\epsilon}\hat{\epsilon}' = PEYY'P = PE\epsilon\epsilon'P' = \sigma^2 P \quad .$$

We see that even if assumption (1) is true, the LS estimates of residuals are neither independent nor do they have constant variance, since $P \pm I_n$.

Theil [1965] has presented an estimator of ϵ which has all the ordinary properties of $\hat{\epsilon}$ except that the covariance of the Theil estimator is $\sigma^2 I_{n-1}$ under assumption (1). The dimension of the Theil estimator of ϵ is n-1 due to the fact that the residual has n-1 degrees of freedom. While Theil [1965] has given the general procedure for deriving uncorrelated residuals with constant variance under the homoscedasticity assumption in multiple linear regression, Koerts [1967] has derived the explicit form of the Theil estimator for the simple linear model through the origin. The

latter will be used here in our diagnostic checking of residuals to detect a monotonic relation between mean and variance by means of the peak test introduced by Goldfeld and Quandt [1965].

2. THEIL ESTIMATOR OF RESIDUALS

We denote this estimator by ϵ^* and for a simple linear model through the origin it can be represented simply as

$$\epsilon_i^* = y_i - b^* x_i \qquad i=1,2,\ldots,k-1,\ k+1,\ldots,n$$

where

$$b^* = (1-a)\,\hat{\beta}_{n-1} + a y_k / x_k$$

and

$$a = |x_k| \Big/ \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} \text{ and } \hat{\beta}_{n-1} = \sum_{\substack{i=1 \\ i\neq k}}^{n} x_i y_i \Big/ \sum_{\substack{i=1 \\ i\neq k}}^{n} x_i^2 \ .$$

k can take any value from 1 to n and the choice is largely a matter of power with respect to a specific alternative hypothesis. Properties of Theil residuals are:

(1) ϵ_i^* is a linear function of y ,

(2) $E\epsilon_i^* = 0, \qquad i=1,2,\ldots,k-1,\ k+1,\ldots,n$,

(3) $\mathrm{Cov}(\epsilon_i^*, \epsilon_j^*) = 0$ if $i \neq j$

$= \sigma^2$ if $i = j$ $\qquad i,j = 1,2,\ldots,k-1,\ k+1,\ldots,n$,

(4) have the minimum expected sum of squares in the class of estimators with properties 1, 2, and 3,

(5) $$\sum_{\substack{i=1 \\ i \neq k}}^{n} \epsilon_i^{*2} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 \quad .$$

Properties (3) and (5) make Theil residuals very interesting indeed. Theil residuals have been derived based on the first four properties. Koerts [1967] has shown that Theil residuals also have the fifth property.

3. APPLICATION OF THEIL RESIDUALS

Consider the case where x's have been ordered such that $x_i < x_j$ if $i < j$. And suppose we are interested in testing the following hypothesis:

$$H_o: E\epsilon_i^2 = \sigma^2 \quad \text{for all } i$$

versus

$$H_1: E\epsilon_i^2 = \sigma_i^2 < E\epsilon_j^2 = \sigma_j^2 \quad \text{for } i < j \quad .$$

Note that alternative hypothesis says that as x increases the variance of ϵ or y increases also. We are considering the case where we have only a single observation for each level of x, as is the case in most experiments.

Two alternative tests for testing H_o against H_1 are suggested by Goldfeld and Quandt [1965].

(1) F test

The obvious choice for k is then the middle observation, so that one can compute the ratio of the sum of squares of the first (n-1)/2 estimated residuals to that of the last (n-1)/2, which is F distributed. When n-1 is not even, one can use either (n-2)/2 first and n/2 last observations or n/2 first and the (n-2)/2 last observations, and for this choice see Theil [1965].

(2) Peak Test

While the F test is a general test, the peak test has been constructed specifically for testing H_o against the particular H_1 under consideration and therefore one expects to obtain greater sensitivity from this test than the F test. Especially when the number of observations is small the greater sensitivity of the peak test becomes important.

The idea of this test was originally given by Goldfeld and Quandt [1965]. For residuals ordered by the ordering of x_i, $x_i < x_{i+1}$, they define a peak at x_i to be an instance where $|\hat{\epsilon}_i| > |\hat{\epsilon}_j|$ for $j = 1,2,\ldots,i-1$. Hedayat and Robson [1966] have demonstrated the failure of the Goldfeld and Quandt peak test applied to the LS residuals $\hat{\epsilon}$, but application to ϵ^* is valid and appropriate because

(a) under H_o, ϵ_i^*'s are uncorrelated and therefore under the normality assumption will be independent,

(b) under H_1, $\text{var}\, \epsilon_i^* (= E\epsilon_i^{*2}) < \text{var}_{i+1} \epsilon_{i+1}^* (= E\epsilon_{i+1}^{*2})$ and hence one expects

$$\epsilon_i^{*2} < \epsilon_{i+1}^{*2} \text{ or equivalently } |\epsilon_i^*| < |\epsilon_{i+1}^*| .$$

Proof of part (b) stems from the following theorem, a proof of which is given in the appendix.

Theorem: If $E\epsilon_i\epsilon_j = 0$, $i \neq j$ and $E\epsilon_i^2 = \sigma_i^2 < E\epsilon^2 = \sigma^2$, then var ϵ_i^* < var ϵ_j^* for $i < j$.

4. A NUMERICAL ILLUSTRATION

We apply the peak test using Theil residuals to the example (see table 1) given on page 180 of Steel and Torrie [1960]. As Steel and Torrie [1960] have pointed out, in this instance regression of y on x should pass through the origin. Therefore,

$$\hat{\beta} = \sum_{i=1}^{13} x_i y_i \Big/ \sum_{i=1}^{13} x_i^2 = 3.67$$

and hence the regression line is given by $y = 3.67x$. The reduction in sum of squares attributable to regression is

$$\left(\sum_{i=1}^{13} x_i y_i\right)^2 \Big/ \sum_{i=1}^{13} x_i^2 = 351{,}819 \quad .$$

And the residual sum of squares is 356,259 - 351,819 = 4,440 . The individual least square residuals are:

$\hat{\epsilon}_1 = +\ 2.088$	$\hat{\epsilon}_8 = +\ 11.887$
$\hat{\epsilon}_2 = -\ 3.013$	$\hat{\epsilon}_9 = +\ 17.527$
$\hat{\epsilon}_3 = -\ 5.437$	$\hat{\epsilon}_{10} = +\ 29.756$
$\hat{\epsilon}_4 = -\ 4.952$	$\hat{\epsilon}_{11} = +\ 19.716$
$\hat{\epsilon}_5 = -\ 16.888$	$\hat{\epsilon}_{12} = -\ 44.688$
$\hat{\epsilon}_6 = -\ 16.066$	$\hat{\epsilon}_{13} = +\ 6.541$
$\hat{\epsilon}_7 = -\ 8.167$	

TABLE 1

Induced Reversions to Independence per 10^7 Surviving Cells y per Dose (ergs/Bacterium) 10^{-5}x of Streptomycin Dependent Escherichia Coli Subjected to Monochromatic Ultraviolet Radiation of 2,967 Angstroms Wavelength.

x	y
13.6	32
13.9	48
21.1	72
25.6	89
26.4	30
39.8	130
40.1	139
43.9	173
51.9	208
53.2	225
65.2	259
66.4	199
67.7	255
$\sum_{i=1}^{13} x_i$ = 528.8	$\sum_{i=1}^{13} y_i$ = 1,929
$\sum_{i=1}^{13} x_i^2$ = 26,062.10	$\sum_{i=1}^{13} y_i^2$ = 356,259

First of all, examination of these residuals suggests, that there is a pattern for the distribution of plus and minus signs among the $\hat{\epsilon}_i$'s . Second, one gets the impression that the absolute value of $\hat{\epsilon}_i$ increases as i increases. Now suppose we suspect the assumption $E\epsilon_i^2 = \sigma^2$ for all i and in particular we suspect that the variance may increase with the mean; i.e., that the variance of y increases as x increases. To test against this alternative hypothesis we first find the Theil residuals

$$\epsilon_i^* = y_i - b^* x_i \quad , \qquad i=1,2,\ldots,12 \qquad (k=13)$$

where

$$b^* = (1-a)\hat{\beta}_{12} + ay_{13} \,/ x_{13} = 3.7014 \quad .$$

Thus,

$$\epsilon_i^* = y_i - 3.7014 x_i \quad , \qquad i=1,2,\ldots,12 \quad .$$

and

ϵ_1^*	= + 1.66096	ϵ_7^*	= - 9.42614
ϵ_2^*	= - 3.44946	ϵ_8^*	= + 10.50854
ϵ_3^*	= - 6.09954	ϵ_9^*	= + 15.89734
ϵ_4^*	= - 5.75584	ϵ_{10}^*	= + 28.08552
ϵ_5^*	= - 17.71696	ϵ_{11}^*	= + 17.66872
ϵ_6^*	= - 17.31572	ϵ_{12}^*	= - 46.77296

No. of peaks = 3

$\sum_{i=1}^{12} \epsilon_i^{*2} = 4,439.4133$ where the slight departure from 4,440 is due to rounding error in computing b^* .

The ϵ_i^*'s are independent and identically distributed under the homoscedasticity and normality assumption of the ϵ_i's . Now we can compute the probability of obtaining five or more peaks in a sequence of 12 independent and identically distributed random variables using table 2 from Goldfeld and Quandt [1965].

By interpolation from this table we see that the probability is about .036 that a sequence of 12 independent and identically distributed random variables produces five or more peaks. If we can accept a risk of 3.6 percent and if our suspicion about homoscedasticity has biological support, then we should fit a weighted regression rather than the unweighted one for obtaining an efficient estimate of β and hence the regression line.

5. DISCUSSION

If we neglect checking the textbook assumptions related to the linear model, this means either that we are willing to accept these assumptions or simply that we are not aware of their importance. In practice, however, there is rarely good reason for supposing that conventional assumptions are satisfied, and methods are therefore needed for detecting and measuring any sort of departure from these ideal conditions.

As Anscombe and Tukey [1960] have put it, "If we are to improve our analysis of data to which the conventional technique can be applied, it is not likely that we shall do this by improving the techniques themselves. Rather we must learn either to go further, beyond the place where the conventional techniques stop, or we must learn to use the technique better. Either path demands the analysis of residuals, where

(residual) = (observed value) - (fitted value) .

TABLE 2

CUMULATIVE PROBABILITIES FOR THE DISTRIBUTION OF PEAKS

	P (number of peaks ≤ x)										
n	x = 0	x = 1	x = 2	x = 3	x = 4	x = 5	x = 6	x = 7	x = 8	x = 9	x = 10
5	.2000	.6167	.9083	.9917	1.0000						
10	.1000	.3829	.7061	.9055	.9797	.9971	.9997	1.0000			
15	.0667	.2834	.5833	.8211	.9433	.9866	.9976	.9997	1.0000		
20	.0500	.2274	.5022	.7530	.9056	.9720	.9935	.9988	.9998	1.0000	
25	.0400	.1910	.4441	.6979	.8705	.9559	.9879	.9973	.9995	.9999	1.0000
30	.0333	.1654	.4001	.6525	.8386	.9395	.9815	.9953	.9990	.9998	1.0000
35	.0286	.1462	.3654	.6144	.8098	.9234	.9745	.9929	.9984	.9997	.9999
40	.0250	.1313	.3373	.5818	.7837	.9078	.9674	.9903	.9975	.9995	.9999
45	.0222	.1194	.3138	.5536	.7600	.8930	.9601	.9874	.9966	.9992	.9998
50	.0200	.1096	.2940	.5288	.7383	.8788	.9530	.9844	.9956	.9989	.9998
55	.0182	.1014	.2769	.5068	.7184	.8653	.9456	.9813	.9944	.9986	.9997
60	.0167	.0944	.2620	.4871	.7001	.8524	.9384	.9780	.9932	.9982	.9996

In the first path we analyze residuals to learn what they can tell us of direct interest. In the second path we must analyze the residuals from a first application of conventional methods to learn how a second application might be better made."

Least square residuals, even under ideal conditions, are in general correlated and have different variances. Perhaps for graphical examination of residuals in certain cases we can neglect both the covariance and heterogeneity of variances which exists among the least squares residuals, but certainly for constructing tests or for any rigorous examination of residuals, we prefer to work with a new type of residuals which are free from the above criticism.

As data analysis becomes more sophisticated, computation will surely get more extensive rather than simpler, and if sophisticated data analysts are to gain in depth and power, they must have both the time and stimulation to try out new procedures of analysis. We take advantage here to quote Tukey's [1962] words of wisdom, "The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge?"

ACKNOWLEDGEMENT

The author wishes to thank D. S. Robson for the suggestion that orthogonal residuals might preserve monotonicity of variance.

REFERENCES

Anscombe, F. J. [1960]. Examination of residuals. Proc. Fourth Berkeley Symp. Math. Statist. I, 1-36.

Anscombe, F. J. and Tukey, J. W. [1963]. The examination and analysis of residuals. Technometrics 5, 141-159.

Goldfeld, S. M. and Quandt, R. E. [1965]. Some tests for homoscedasticity. J. Amer. Statist. Assoc. 60, 539-547.

Hedayat, A. [1966]. Homoscedasticity in linear regression analysis with equally spaced x's. M.S. Thesis, Cornell University, Ithaca, N.Y.

Hedayat, A. and Robson, D. S. [1966]. Independent transformed residuals for testing homoscedasticity. Paper No. BU-135, Cornell University, Ithaca, N.Y.

Koerts, J. [1967]. Some further notes on disturbance estimates in regression analysis. J. Amer. Statist. Assoc. 62, 169-183.

Steel, R. G. D. and Torrie, J. H. [1960]. Principle and procedures of statistics. McGraw-Hill Book Company, Inc., New York.

Theil, H. [1965]. The analysis of disturbances in regression analysis. J. Amer. Statist. Assoc., 60, 1067-1079.

Tukey, J. W. [1962]. The future of data analysis. Ann. Math. Statist., 33, 1-67.

APPENDIX

We now prove that if $E\epsilon_i\epsilon_j = 0$, $i \neq j$ and $E\epsilon_i^2 = \sigma_i^2 < E\,\epsilon_{i+1}^2 = \sigma_{i+1}^2$ then $\operatorname{var} \epsilon_i^* < \operatorname{var} \epsilon_{i+1}^*$. Under these assumptions and by the definition of ϵ_i^* we have

$$\operatorname{Var} \epsilon_i^* = E\epsilon_i^{*2} - (E\epsilon_i^*)^2$$

$$= E\epsilon_i^{*2}$$

$$= E\left[y_i - x_i\hat{\beta}_{n-1} - ax_i\left(\hat{\beta}_{n-1} - \frac{y_n}{x_n}\right)\right]^2$$

$$= x_i^2\beta + \sigma_i^2 + \beta^2\left[x_i^2(1-a)^2 + x_i^2a^2 + 2a(1-a)x_i^2\right] + \frac{(1-a)^2x_i^2\sum_{i=1}^{n-1} x_i^2\sigma_i^2}{\sum_{i=1}^{n-1} x_i^2} + \frac{x_i^2a\sigma_n^2}{x_n^2}$$

$$- 2x_i(1-a)\left(x_i\beta^2 + \frac{x_i\sigma_i^2}{\sum_{i=1}^{n-1}}\right) - 2x_i^2\beta a$$

$$= \beta^2\left[x_i^2 + x_i^2(1-a)^2 + x_i^2a^2 + 2a(1-a)x_i^2 - 2x_i^2(1-a) - 2x_i^2a\right] + \sigma_i^2 + \frac{(1-a)^2x_i^2\sum_{i=1}^{n-1} x_i^2\sigma_i^2}{\sum_{i=1}^{n-1} x_i^2} + \frac{ax_i^2\sigma_n^2}{x_n^2} - 2x_i(1-a)\frac{x_i^2\sigma_n^2}{x_n^2}$$

$$= \sigma_i^2 + A^2x_i^2 \sum_{i=1}^{n-1} x_i^2\sigma_i^2 + Bx_i^2 - 2Ax_i^2\sigma_i^2$$

where

$$A = \frac{1-a}{\sum_{i=1}^{n-1} x_i^2}, \quad B = \frac{\sigma_n^2(1-a)^2}{\sum_{i=1}^{n-1} x_i^2}$$

$$= \sigma_i^2(Ax_i^2 + 1)^2 + x_i^2A^2 \sum_{\substack{j=1 \\ j\neq i,i+1}}^{n-1} x_j^2\sigma_j^2 + x_i^2A^2x_{i+1}^2\sigma_{i+1}^2 + Bx_i^2$$

$$= \sigma_i^2(Ax_i^2 + 1)^2 + CA^2x_i^2 + A^2x_i^2x_{i+1}^2\sigma_{i+1}^2 + Bx_i^2$$

where

$$C = \sum_{\substack{j=1 \\ j\neq i,i+1}}^{n-1} x_j^2\sigma_j^2 .$$

Now

$$\text{Var } \epsilon_{i+1}^* - \text{Var } \epsilon_i^* = \sigma_{i+1}^2 (Ax_{i+1}^2 + 1)^2 - \sigma_i^2(Ax_i^2 + 1)^2 + CA^2(x_{i+1}^2 - x_i^2)$$

$$+ A^2x_i^2x_{i+1}^2(\sigma_i^2 - \sigma_{i+1}^2) + B(x_{i+1}^2 - x_i^2) .$$

For $x_{i+1} > x_i$ we have

$$CA^2(x_{i+1}^2 - x_i^2) > 0 \quad \text{and} \quad B(x_{i+1}^2 - x_i^2) > 0 \quad .$$

Now in order to show that var ϵ_{i+1}^* > var ϵ_i^* we have to show that

$$\sigma_{i+1}^2(Ax_{i+1}^2 + 1)^2 - \sigma_i^2(Ax_i^2 + 1) + A^2x_i^2x_{i+1}^2(\sigma_i^2 - \sigma_{i+1}^2) > 0 \quad .$$

That is,

$$\sigma_{i+1}^2A^2x_{i+1}^2x_{i+1}^2 - \sigma_i^2A^2x_i^2x_i^2 + \sigma_{i+1}^2(1 + 2Ax_{i+1}^2)$$

$$- \sigma_i^2(1 + 2Ax_i^2) + A^2x_i^2x_{i+1}^2(\sigma_i^2 - \sigma_{i+1}^2) > 0 \quad .$$

Since $x_i < x_{i+1}$, then the left side of the above final expression will be greater than

$$\sigma_{i+1}^2A^2x_i^2x_{i+1}^2 - \sigma_i^2A^2x_i^2x_{i+1}^2 + \sigma_{i+1}^2(1 + 2Ax_{i+1}^2)$$

$$- \sigma_i^2(1 + 2Ax_i^2) + A^2x_i^2x_{i+1}^2(\sigma_i^2 - \sigma_{i+1}^2)$$

$$= \sigma_{i+1}^2(1 + 2Ax_{i+1}^2) - \sigma_i^2(1 + 2Ax_i^2) > 0 \quad .$$

Q.E.D.