Journal of the Korean
Data & Information Science Society
2001, Vol. 12, No.2, pp. 35 ~ 45

Percentile Envelope and Its Characteristic of Error Distribution for Supernormality

Jea-Young Lee [1] · **Seong-Won Rhee** [2]

Abstract

We introduce a new percentile envelope for diagnosing supernormality in regression analysis. Furthermore, we compare this percentile envelope, which is much simpler and easier, with Atkinson's and Flack and Flores' envelopes.

Using percentile envelope, we investigate characteristics of normal probability plots with envelope for error distributions when supernormality is occurred. We give cautions that test result for normality assumption of errors can be reached the wrong conclusion by supernormality.

Key Words and Phrases: supernormality, Atkinson's envelope, Flack and Flores' envelope

1. Introduction

In regression analysis, the error terms $\epsilon_i (i = 1, \cdots, n)$ are assumed to have normal distribution with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = \sigma^2$. To detect non-Gaussian error distribution, normal probability plot of the standardized residuals are often used. This is a plot of the standardized residuals against their expected values. Ryan(1997) described the reason using standardized residuals. It is to give the random variables that correspond to the plotted points to have the same distribution, but it will not occur if we use the raw residuals, which have different standard deviations, and hence have different distributions.

Although the normality assumption of errors is not rejected by Shapiro-Wilk test or normal probability plot, it can be reached the wrong conclusion by supernormality that was studied by Gnanadesikan(1977), Ryan(1997) and Weisberg(1980, 1985). Atkinson(1981, 1985) and Flack and Flores(1989) described the method that

[1] Associate Professor, Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea.
[2] Department of Statistics, Yeungnam University, Kyongsan, 712-749, Korea.

supernormality can be diagnosed by simulated envelope. We introduce and discuss about a new percentile envelope for detecting supernormality in regression analysis.

In section 2, we describe supernormality and its property. In section 3, we describe simulated envelopes for Atkinson's method and Flack and Flores' method and suggest a new percentile envelope. Section 4 gives characteristics of normal probability plot with a new percentile envelope for error distributions when supernormality is occurred. The conclusion is given in section 5.

2. Supernormality

We have linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, with $\mathbf{Y}(n \times 1)$, $\mathbf{X}(n \times p)$, and $\epsilon(n \times 1)$. The error terms ϵ_i are assumed to be independent and identically distributed $N(0, \sigma^2)$. The externally standardized ith residual is

$$r_i^* = \frac{(y_i - \hat{y}_i)}{s_{(i)}\sqrt{1 - v_{ii}}}, \tag{2.1}$$

where y_i is the ith observed y value, $\hat{y}_i$ is its ordinary least squares estimator, $s_{(i)}$ is the estimator of σ when the ith observation has been omitted in the estimation, and v_{ii} is the ith diagonal element of the hat matrix, $\mathbf{V} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

The normality assumption of errors are very important in regression analysis, because it is used to justify F- and t-tests and confidence procedures. Since ϵ is never observed, the test of normality assumption of errors must be based on the residuals, $\hat{\epsilon}$. In matrix notation, $\hat{\epsilon}$ can be written as

$$\hat{\epsilon} = (\mathbf{I} - \mathbf{V})\mathbf{Y} = (\mathbf{I} - \mathbf{V})(\mathbf{X}\beta + \epsilon) = (\mathbf{I} - \mathbf{V})\epsilon,$$

since $(\mathbf{I} - \mathbf{V})\mathbf{X} = \mathbf{0}$.

Let v_{ij} be the (i, j)th element of $\mathbf{V}$, the scalar form of last equation is

$$\hat{\epsilon}_i = \epsilon_i - \sum_{j=1}^{n} v_{ij}\epsilon_j, \tag{2.2}$$

where

$$v_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

The term ϵ_i is equal to $\hat{\epsilon}_i$ that is added to a weighted sum of all the ϵ_j's. If the number of degrees of freedom for error, n-p, is small and some of the v_{ij} have large values, $\sum_{j=1}^{n} v_{ij}\epsilon_j$ in (2.2) may be more important than ϵ_i to determine the distribution of $\hat{\epsilon}_i$. Although the ϵ_i's are not normally distributed, the sum will behave normally because of the central limit theorem. Therefore any test for the normality of errors based on the residuals cannot be expected to be quite good,

at least in small samples. This property of residuals, that is, residuals will have a distribution which is closer to normal than will be the distribution of the errors when the errors are not normally distributed, has been referred to as **supernormality** by many authors(Ryan, 1997).

As n increase for fixed p, the v_{ii} will tend to be zero and the ϵ_i in (2.2) will tend to dominate, because the sum will have relatively small variance. If sample size is large, methods based on residuals can be expected to give as much information as the same methods based on the errors.

Ryan(1997) illustrated an example of supernormality. He generated the Y values as $Y = 3 + X_1 + X_2 + \epsilon$, where ϵ has a chi-square distribution with 2 degrees of freedom, and the values for X_1 and X_2 were generated using the same distribution. Thus Y has a chi-square distribution with 6 degrees of freedom. The p-values of Shapiro-Wilk tests for residuals and errors are 0.3676 and 0.0011, respectively. Here, we conclude that the residuals are normally distributed, whereas the hypothesis of normality is rejected for the errors. Thus, we would reach the wrong conclusion if we relied on the residuals to determine the distribution of the error term.

We use simulation as an aid to diagnose supernormality. Because it gives statisticians a basis for comparison of the observed plot with an expected plot which is derived from Gaussian-distributed residuals. Atkinson(1981) and Flack and Flores(1989) suggested two diagnostic envelopes and in next section, we will discuss them and introduce a new percentile envelope based on Atkinson's and Flack and Flores' methods.

3. A New Percentile Envelope

Atkinson(1981) suggested a method that puts simulated envelope on the normal probability plot to give the statisticians a yardstick for comparing an observed plot with an expected plot.

We use terminology **diagnostic envelope** rather than confidence envelope that has been applied to this type of envelope. For given linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, the procedure for Atkinson's envelope is as follows:

Step 1. Generate M(in here, $M = 19$) $n \times 1$ vectors $\mathbf{Z_j}$ $(j = 1, ..., M)$ from the $N(0, \mathbf{I})$ distribution.

Step 2. For each of these vectors, fit the model $\mathbf{Z} = \mathbf{X}\beta + \epsilon$ to obtain M simulated residual vectors $\mathbf{r}^*_{\mathbf{z_j}}$.

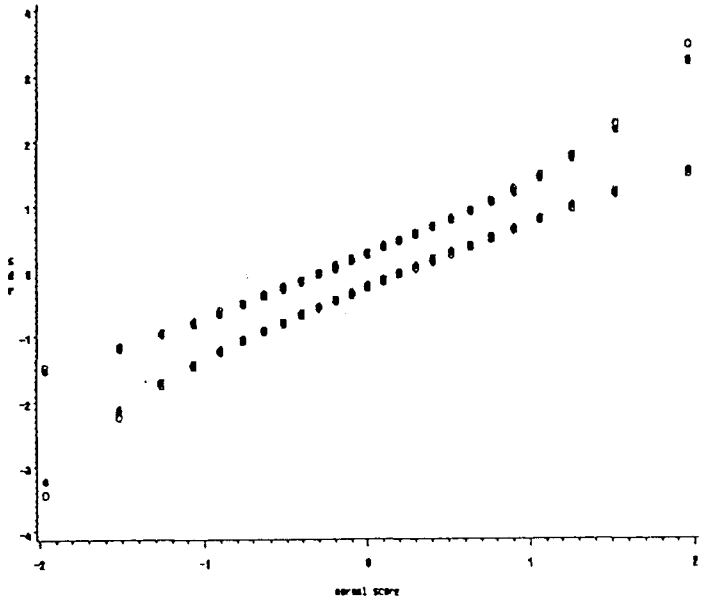
Step 3. Order the elements of each simulated residual vector.

Step 4. For each $i = 1, \cdots, n$, select $l_i = \min\limits_j r^*_{z_j}(i)$ and $u_i = \max\limits_j r^*_{z_j}(i)$, where $r^*_{z_j}(i)$ is the ith-order statistic of the jth simulated residual vector $\mathbf{r}^*_{\mathbf{z_j}}$.

Flack and Flores added 2 steps to stabilize the bounds even further.

Step 5. Repeat **step 1** to **step 4** 2000 times.

(a) Atkinson's envelope



(b) Flack and Flores' envelope

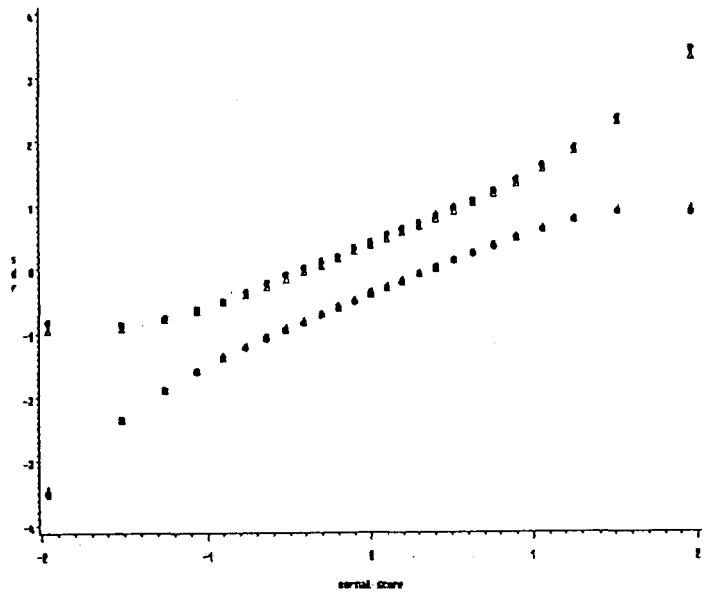


Figure 1. (a), (b) compare Percentile envelope with Atkinson's envelope and Flack and Flores' envelope, respectively. (sdr means studentized deleted residual, • : Percentile envelope, o : Atkinson's envelope, △ : Flack and Flores' envelope)

Step 6. Get mean envelop which is the pointwise average over the 2000 replications of envelopes.

These upper and lower values of mean envelope consist of the upper and lower of Atkinson's envelope, respectively. Atkinson suggested the number of M be 19 to give envelope boundaries estimates of the 5th and 95th percentiles of the distribution of the ith-order statistic of the externally studentized residual vector, given $\mathbf{X}$ and *iid* $N(0, \sigma^2)$ errors. However, Atkinson's envelope has problems. First, envelope boundaries for ith-ordered residual are estimates of a $100\frac{j}{(M+1)}$ percentile, where $j = 1, ..., M$. If estimates need to be further into the tail of the distribution M must have larger value, which means that larger or many number of simulations are needed. Furthermore, envelope is quite sensitive to extreme-order statistics among simulated residuals since such statistics would also become part of the envelope. Flack and Flores(1989) applied outlier-resistant rule proposed by Tukey(1977) and Hoaglin, Iglewicz and Tukey(1986) to Atkinson's method to get less sensitive to extreme values among the simulated residuals than the Atkinson's envelope. The procedure is as follows: All step is same as Atkinson's method except of selection of l_i and u_i $(i = 1, 2, \cdots, n)$. The lower and upper envelope boundaries for the ith-ordered externally studentized residuals are

$$l_i = F_i^L - k(F_i^U - F_i^L)$$

and

$$u_i = F_i^U + k(F_i^U - F_i^L)$$

where F_i^L and F_i^U are the 25th and 75th percentiles of the M ith-order statistics and k is an appropriately chosen constant. Hoaglin *et al.*(1986) chose $k = 1.5$ in the univariate case as a good rule for flagging outliers in moderately sized samples. For each i, F_i^L and F_i^U are the f-order and $(M + 1 - f)$-order statistics of the M simulated values, $r_{z_1}^*(i), \ldots, r_{z_M}^*(i)$. Hoaglin and Iglewicz(1987) suggested the ideal $f = M/4 + 5/12$ for estimating the fourths.

Here, we suggest a new percentile envelope that applies parametric percentile which is studied by Efron and Tibshirani(1993). Using Percentile Envelope, we can reduce significant number of simulations, comparing to two methods. The procedure is as follows:

Step 1 to Step 3 is same as Atkinson's method, where $M = 2000$.

Step 4. For each $i = 1, \cdots, n$, select, for Atkinson's envelope,

$$l_i = P_i^{LA}$$

and

$$u_i = P_i^{UA},$$

or for Flack and Flores' envelope,

$$l_i = P_i^{LF} - k\{P_i^{UF} - P_i^{LF}\}$$

Table 1. Values of 5th, 95th percentiles of Atkinson's envelope and Percentile envelope for Atkinson's method. Values of 25th, 75th percentiles of Flack and Flores' envelope and Percentile envelope for Flack and Flores' method. (P^{LA}, P^{LF}, P^{UF}, P^{UA} : 5th, 25th, 75th, 95th percentiles of Percentile envelope, A^L, A^U : 5th, 95th percentiles of Atkinson's envelope, F^L, F^U : 25th, 75th percentiles of Flack and Flores' envelope)

ith	P^{LA}	A^L	P^{UA}	A^U	P^{LF}	F^L	P^{UF}	F^U
1	-3.2152	-3.42035	-1.52619	-1.47785	-2.50229	-2.52023	-1.83252	-1.89378
2	-2.1469	-2.23784	-1.20036	-1.16781	-1.80007	-1.82352	-1.42924	-1.4654
3	-1.70799	-1.75844	-0.96888	-0.95155	-1.46715	-1.47474	-1.18526	-1.20221
4	-1.43392	-1.47007	-0.82472	-0.77503	-1.23552	-1.23681	-0.98904	-1.00398
5	-1.21428	-1.25061	-0.65461	-0.61794	-1.04944	-1.05495	-0.82893	-0.84006
6	-1.05264	-1.08003	-0.51708	-0.48337	-0.8981	-0.8997	-0.68338	-0.69546
7	-0.90473	-0.93446	-0.39022	-0.35931	-0.7577	-0.76313	-0.54743	-0.56814
8	-0.78233	-0.80556	-0.2765	-0.24402	-0.63871	-0.63876	-0.42397	-0.44786
9	-0.65467	-0.68416	-0.17169	-0.13502	-0.52573	-0.5261	-0.31548	-0.33557
10	-0.55541	-0.57545	-0.06553	-0.03328	-0.40733	-0.41441	-0.20313	-0.22895
11	-0.44047	-0.46984	0.02876	0.07034	-0.30301	-0.31031	-0.10577	-0.1235
12	-0.34214	-0.36829	0.14266	0.17207	-0.20334	-0.20858	0.00426	-0.02009
13	-0.23883	-0.27	0.2411	0.27063	-0.09087	-0.10676	0.1048	0.08122
14	-0.1339	-0.17303	0.35472	0.37111	0.00481	-0.00485	0.21299	0.18369
15	-0.03817	-0.07176	0.45233	0.46996	0.10046	0.09645	0.31096	0.28482
16	0.06352	0.03026	0.54918	0.57359	0.20954	0.2006	0.40887	0.39057
17	0.16703	0.13263	0.65873	0.68486	0.31637	0.30653	0.52139	0.50081
18	0.28254	0.2416	0.7797	0.80175	0.43133	0.42221	0.63926	0.61444
19	0.38342	0.35507	0.90545	0.93188	0.5481	0.54189	0.75407	0.73758
20	0.51361	0.4794	1.05012	1.07615	0.67556	0.67033	0.89077	0.87541
21	0.64919	0.61727	1.19903	1.24636	0.81964	0.81404	1.04646	1.02841
22	0.79863	0.77033	1.42261	1.46703	0.97999	0.97845	1.23107	1.20842
23	0.99797	0.95026	1.72258	1.7654	1.18602	1.17311	1.46519	1.44092
24	1.21147	1.17123	2.14831	2.24337	1.43705	1.42869	1.80139	1.77517
25	1.53262	1.48507	3.20058	3.44513	1.83636	1.83282	2.47741	2.42977

Table 2. General characteristics of normal probability plot

Shape of distribution compared to normal from the center	Slope of normal probability plot
shorter left-tail	increase
shorter right-tail	decrease
longer left-tail	decrease
longer right-tail	increase

and

$$u_i = P_i^{UF} + k\{P_i^{UF} - P_i^{LF}\},$$

where P_i^{LA} and P_i^{UA} are 5th and 95th percentiles of the M ith-order statistics and P_i^{LF} and P_i^{UF} are 25th and 75th percentiles of the M ith-order statistics.

Based on the above step 1 through step 4, we find much simpler and easier method than Atkinson's and Flack and Flores' methods, because the total number of simulations for Percentile envelope is 2000, but that of Atkinson's and Flack and Flores' methods are $19 \times 2000 = 38000$.

Figure 1-(a) shows that Percentile envelope for Atkinson's method is almost same as Atkinson's envelope, in exception of extreme points. Because Percentile envelope for Atkinson's method is based on estimates of 5th and 95th percentiles of the distribution of the 2000, ith-order statistic of the externally studentized residual vector, but Atkinson's envelope is based on extreme-order statistics.

Figure 1-(b) shows that Percentile envelope for Flack and Flores' method is almost same as Flack and Flores' envelope, in exception of extreme points. We see that widths of Percentile envelope for Flack and Flores' method are wider than Percentile envelope for Atkinson's.

Table 1 also shows that Atkinson's envelope and Flack and Flores' envelope are almost same as Percentile envelope.

4. Supernormality Using Percentile Envelopes for Error Distributions

In this section, we investigate characteristics of normal probability plots with Percentile envelope for error distributions when supernormality is occurred. First, X_1 and X_2 fixed for all error distributions were generated for uniform distribution on the interval $(0, 10)$ for small number of observations by the property of supernormality, $n = 20$. Second, we get envelopes for Percentile via simulation as described in section 3. Third, the Y values are generated as $Y = 3 + X_1 + X_2 + \epsilon$, where

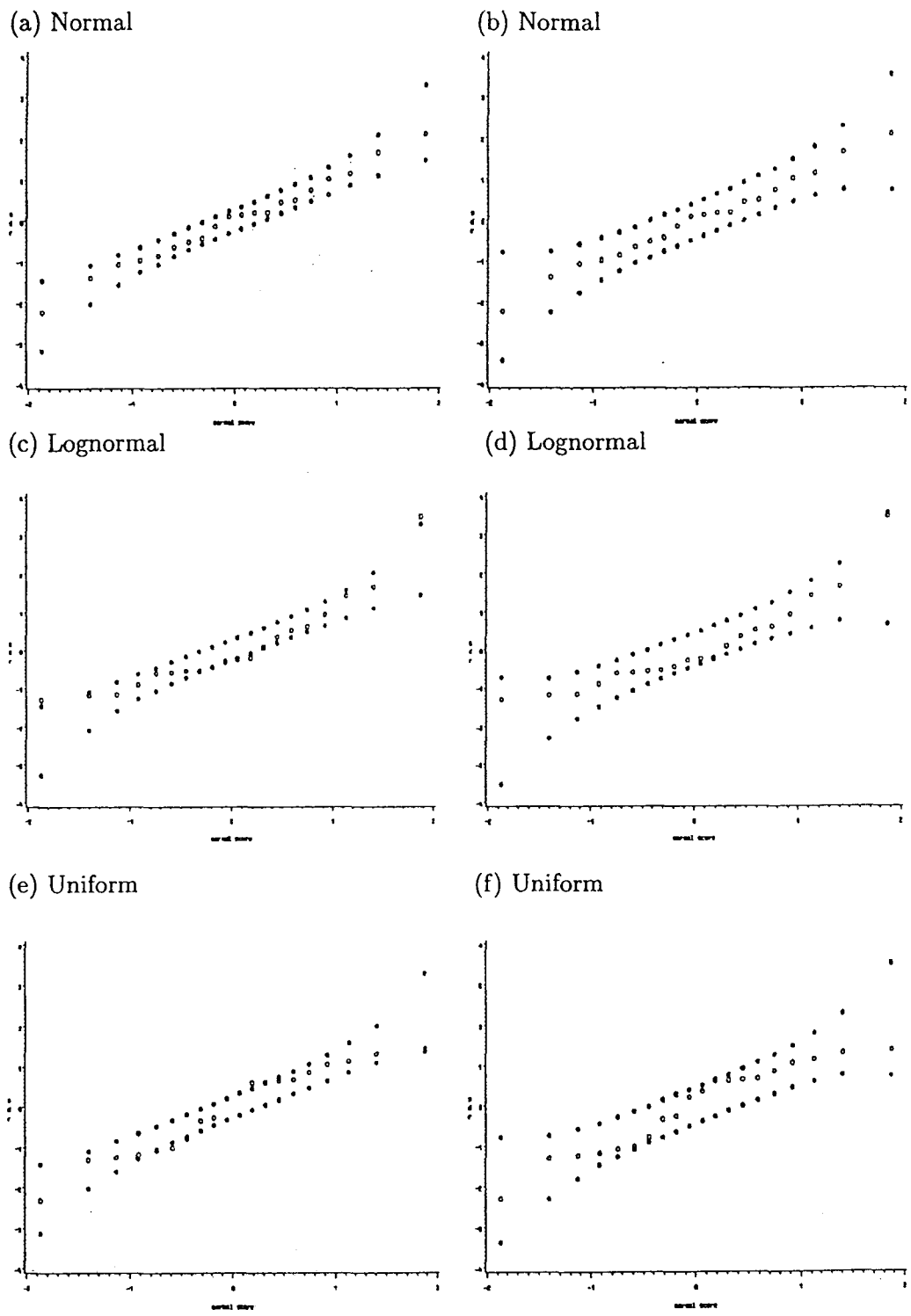


Figure 2. (a), (c), (e) are Percentile envelope for Atkinson's method for each error distribution. (b), (d), (f) are Percentile envelope for Flack and Flores' method for each error distribution.

Table 3. Characteristics of normal probability plot and Supernormality by envelope

Distribution of ϵ	Characteristic of normal probability plot	Supernormality by envelope
Normal	Straight line	All points fall inside envelope.
Lognormal, Chisquare, Gamma, Exponential	Slope increases from left-tail	On boundary or fall outside upper envelope at left- and right- tail, lower envelop at right from center.
Extreme Value	Slope decreases from left-tail	On boundary or fall outside upper envelope at left from center, lower envelop at left- and right-tail.
Cauchy, Inverse Normal	Slope decreases from left-tail and increases from center	On boundary or fall outside upper envelope at left from center and right-tail, lower envelop at left-tail and right from center.
Uniform	Slope increases from left-tail and decreases from center	On boundary or fall outside upper envelope at left-tail and right from center, lower envelop at left from center and right-tail.

ϵ is generated for each error distribution. Here, we select data that p-values of Shapiro-Wilk tests for errors and residuals are less than 0.05 and between 0.1 and 0.15, respectively. It means that supernormality is occurred.

To diagnose supernormality from normal probability plot with envelope, we need to examine the shape of normal probability plot for distributions studied by Lee and Rhee(1997). Table 2 shows general characteristics of normal probability plot for shape of distribution which is compared with tails of normal distribution from center. For example, uniform distribution has shorter tails than normal from center, so the slope of normal probability plot increases from left-tail and decreases from center to right-tail.

Table 3 shows characteristics of normal probability plot and supernormality by envelope for error distributions. When the distribution of ϵ is normal, all points fall inside envelope, but if ϵ is not normally distributed some points are on boundary or fall outside envelope which means that supernormality is occurred.

Figure 2-(a), (b) are normal probability plots with Percentile envelope for Atkinson's method and Flack and Flores' method for normal distribution. There are nc points that fall outside of envelope as we expected. Figure 2-(c), (e) are normal probability plots with Percentile envelope for Atkinson's method. These have some points that are on boundary or fall outside envelope at the location described by Table 2 for each error distribution, so supernormality can be diagnosed. Figure 2-

(d), (f) are normal probability plots with Percentile envelope for Flack and Flores' method. These are wider than Percentile envelope for Atkinson's as described in section 3.

5. Conclusion

Atkinson and Flack and Flores suggested methods to diagnose supernormality by simulated envelope. We introduced a *new percentile envelope* which applies bootstrap percentile method. The procedure of percentile envelopes was much simpler and easier, comparing to two methods, because the total number of simulations for Percentile envelope, 2000, is much smaller than that of Atkinson's method and Flack and Flores' method, $19 \times 2000 = 38000$. Furthermore, Figure 1-(a), (b) show that Percentile envelopes for Atkinson's and Flack and Flores' methods are almost same as Atkinson's envelope and Flack and Flores' envelope, respectively.

We investigated supernormality using percentile envelopes for error distributions. We got some results of supernormality that some points are on boundary or fall outside envelope at the location described by Table 3 for distributions of ϵ. Therefore, we give cautions that test result for normality assumption of errors can be reached the wrong conclusion by supernormality.

References

1. Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression, *Biometrika*, 68, 13-20.

2. Atkinson, A. C. (1985). *Plots, Transformations, and Regression*, Oxford University Press, New York.

3. Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, Chapman & Hall, New York.

4. Flack, V. F. and Flores, R. A. (1989). Using simulated envelopes in the evaluation of normal probability plots of regression residuals, *Technometrics*, 31, 219-225.

5. Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*, John Wiley & Sons, New York.

6. Hoaglin, D. C. and Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling, *Journal of the American Statistical Association*, 82, 1147-1149.

7. Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling, *Journal of the American Statistical Association*, 81, 991-999.

8. Lee, J.-Y. and Rhee, S.-W. (1997). Plot Bimodality, *The Korean Communications in Statistics*, 4, 243-254.

9. Ryan, T. P. (1997). *Modern regression methods*, John Wiley & Sons, New York.

10. Tukey, J. W. (1976). *Exploratory Data Analysis*, Addison-Wesley, MA.

11. Weisberg, S. (1980). Comment on "Some large sample tests for nonnormality in the linear regression model" by White, H. and MacDonald, G. M., *Journal of the American Statistical Association*, 75, 28-31.

12. Weisberg, S. (1985). *Applied linear regression*, University of Minnesota, Minnesota.