ON THE DISTRIBUTION OF RESIDUALS IN FITTED PARAMETRIC MODELS

C. P. Quesenberry and Charles Quesenberry, Jr.

Results of a simulation study of the fit of data to an estimated parametric model are reported. Three particular models including the two-parameter normal and exponential distributions, and the simple linear regression model are considered. A number of scaled versions of the least squares residuals from the regression model, and quantities that may be called residuals from the other two models are seen to follow the parent distribution form too well, i.e., to be supernormal and superexponential. A point of particular interest is that this tendency does not decrease with increasing sample size.

KEY WORDS: Fitted parametric models; Residuals; Supernormality; Superexponentiality.

1. INTRODUCTION AND SUMMARY

Let an observed set of data be given, along with a parametric model which is alledged to have given rise to the data. Then a common practice in statistical analysis may be discribed as follows. First, the data is used to estimate the unknown parameters in the model. Secondly, especially if sample sizes are large, the estimated model thus obtained is considered to be approximately the true underlying model, and used as such in further analyses of the original data.

It has long been recognized that this procedure is inexact, at least for small samples. The "bias" in such a procedure may be viewed as arising from the following considerations. For any good method of

estimating the parameters of the model at hand (ML, MVU, least squares, etc.), the resulting estimated model will fit the data used in its own estimation better than the true underlying model of the same parametric form. To study the matter, we consider three particular parametric models in this paper, including two-parameter normal and exponential parents, and the simple normal linear regression model. In each case it is clearly seen that the fitted models fit too well as measured by our criteria.

2. DESCRIPTION OF DISTRIBUTION THEORY, NORMAL SAMPLES

We describe our method first for the normal sample problem. Let $x_1, \cdots, x_n$ be a random sample from a $N(\mu, \sigma^2)$ parent distribution, and consider the residuals $y_i = (x_i - \bar{x})/s$ for $\bar{x}$ and s the usual sample mean and variance. (Throughout this paper i takes values $1, \cdots, n$.) According to our remarks in section 1, if these y-values are plotted on the real line they will tend to be better fitted by a $N(0, 1)$ density than would observed values on iid $N(0, 1)$ rv's. However, if the y's are approximately iid $N(0, 1)$ rv's, and if Φ is the distribution function of a $N(0, 1)$ distribution, then $u_i = \Phi(y_i)$ defines a set of rv's that are approximately iid $U(0, 1)$, (uniformly distributed on the $(0, 1)$ interval) rv's. If these y residuals are indeed too normal, or supernormal, then the u-values are too uniform, or superuniform. In order to study this phenomenon, we can generate sets of u-"samples," and study their distribution on the unit interval.

To study the distribution of the u's on the unit interval, we will compute a statistic which is a measure of the uniformity of the u's on the unit interval. Our choice of a statistic to measure uniformity of

the u's is the Neyman smooth statistic, p_4^2. (See Neyman (1937), Kendall and Stuart (1961), p. 444, or Miller and Quesenberry (1979).) This statistic has several properties which are important here. It takes small values when the u_i's are uniformly distributed on the unit interval, and increases in value as the points become less uniformly distributed on the unit interval. Also, for the sample sizes which we consider here, $n \geq 10$, the p_4^2 statistic is well approximated by a $\chi^2(4)$ rv. Therefore, when the values of $y_1, \cdots, y_n$ are approximately iid $N(0,1)$, then the value of $G_4(p_4^2)$ is approximately a $U(0,1)$ rv, where $G_4(\cdot)$ is the df of a $\chi^2(4)$ rv. If the y_i's are supernormal, then the rv $v = G_4(p_4^2)$ will be stochastically less than a $U(0,1)$ rv. On the other hand, if the y_i's have a nonuniform pattern, then p_4^2 will be stochastically greater than a $U(0,1)$ rv. Thus, if we generate m samples, the values $v_1, \cdots, v_m$ contain important information about the normality of the y's. If the y's are iid $N(0,1)$, then $v_1, \cdots, v_m$ are iid $U(0,1)$, if the y's are supernormal then $v_1, \cdots, v_m$ should tend to be less than a set of m iid $U(0,1)$ rv's. Thus we will generate v-"samples" for a number of sample sizes n. We will then need to study the distributions described by these sets of values, and we have used two slightly different approaches in studying the v-samples.

In our first method of studying v-samples, we generate $m = 500$ samples of size n from the original distribution and then make a 40 cell histogram of these 500 v-values. If the y-values from the original samples are approximately $N(0,1)$ rv's, then these 500 values are independent observations on an approximately $U(0,1)$ rv. Thus the 40 cells have expected cell frequencies of $(500/40) = 12.5$. A classic chi-squared goodness-of-fit statistic computed from these 40 cells will have a

$\chi^2(39)$ distribution, and the observed significance level (p-value) can be easily evaluated.

These histograms are shown in Figures 1 and 2 for n = 10 and 100, respectively. We have computed and plotted such histograms for a range of values of n up to 5,000 and the graphs in Figures 1 and 2 are typical of all these histograms. Observe particularly that these histograms show that the v-values tend to be stochastically less than they would be if the u-values were iid U(0, 1) rv's, or equivalently, if the y-values were iid N(0, 1) rv's. The observed chi-squared goodness of fit statistics were 590.72 (n = 10) and 743.84 (n = 100), which give observed p-values of 1.00000000 in both cases.

Figure 1 and 2 near here

Thus we see that in both cases the distribution of p-values of the Neyman smooth statistics computed from the u-samples tend to be much too small for the y-samples to be normal. Thus u-values tend to be too uniform, and the y-values are too normal (supernormal!).

Now, in order to present results in a compact form, we will report results from the rest of this study without histogram graphs. For selected values of n we have computed 100 u-samples and a v-sample as above. In Table 1 we give the means of these v-samples and the p-value of the Neyman smooth statistic computed from the v-sample. These two statistics summarize the information of prime interest from the histograms like those of Figures 1 and 2.

Two important points should be carefully observed in Table 1. First, recall that these are sample means of 100 v-values, and therefore if the

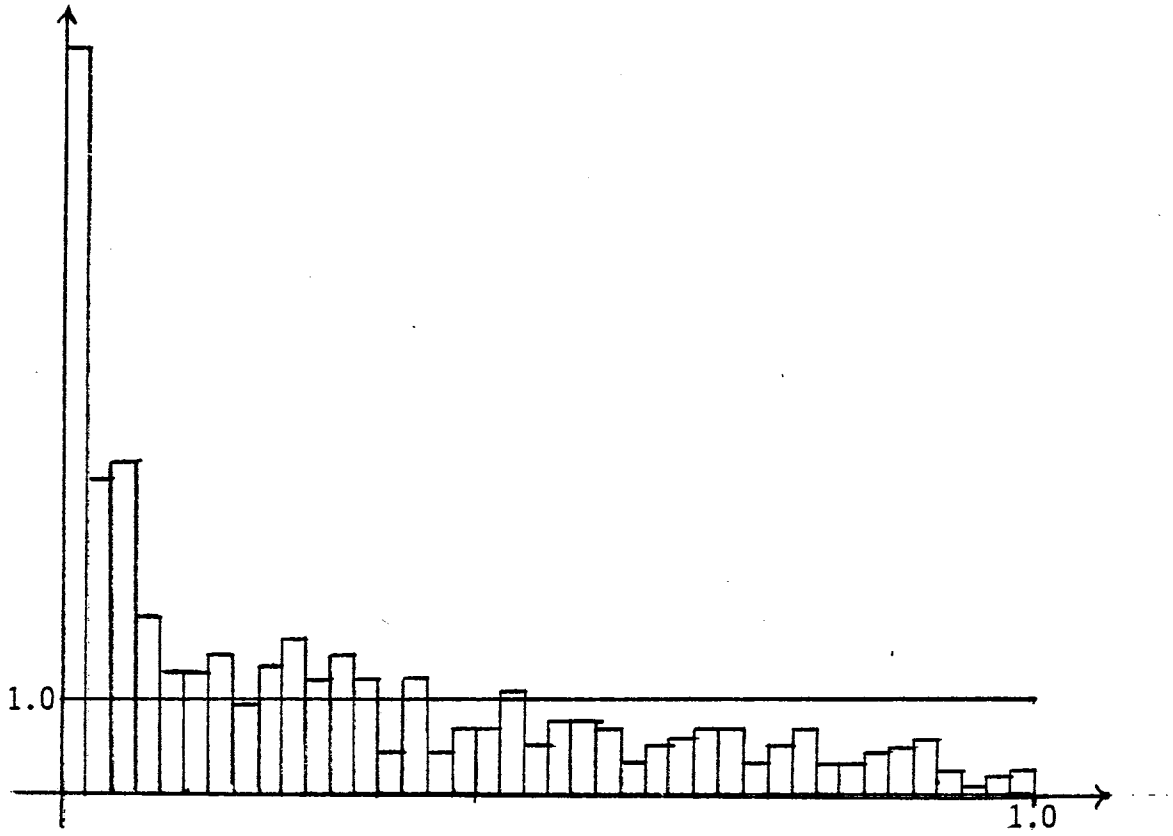

Figure 1. Histogram of v-values for n = 10

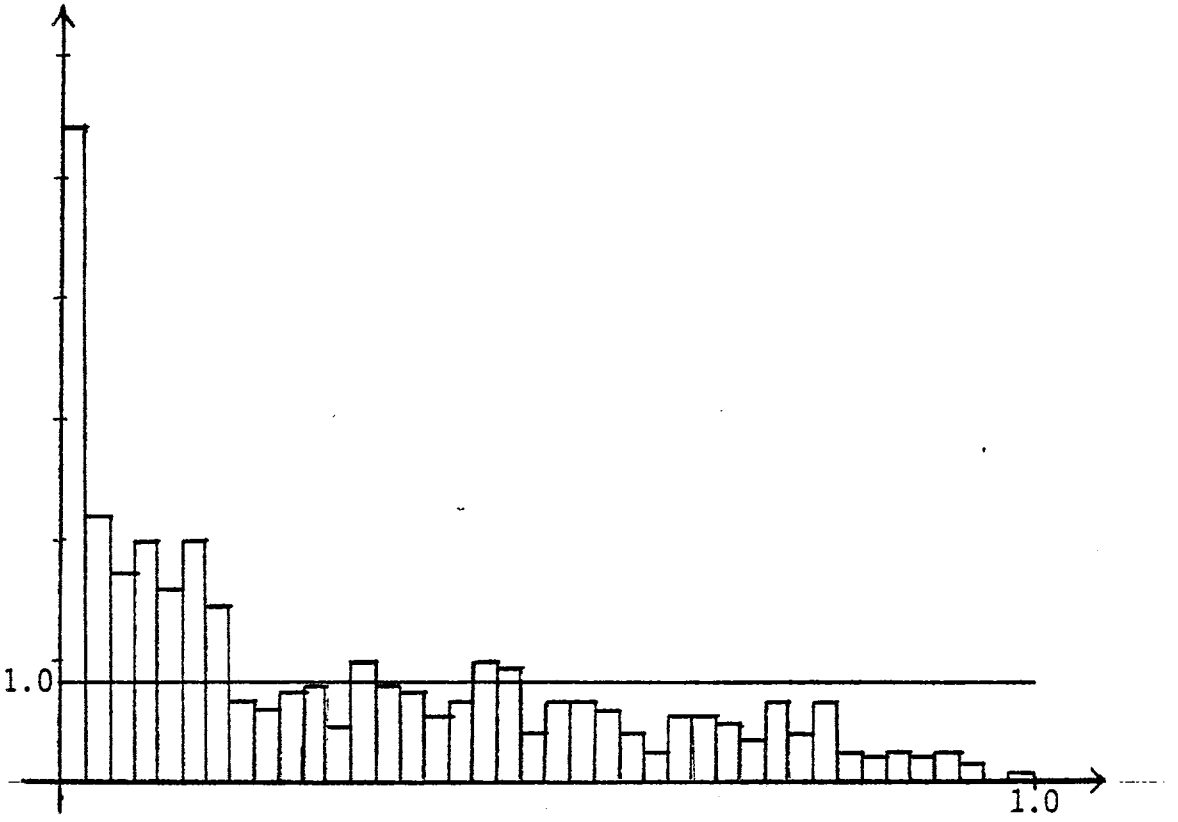

Figure 2. Histogram of v-values for n = 100

1. Normal Distribution Results

n	Mean	Neyman Smooth p-value
10	0.2576	1.00000000
30	0.2642	1.00000000
60	0.2496	1.00000000
100	0.2919	1.00000000
500	0.2702	1.00000000
2,000	0.2893	1.00000000
5,000	0.2987	1.00000000

v-values were iid U(0, 1) rv's, then these would be observations on (essentially) $N(0.5, 1/1200) \doteq N(0.5, (.03)^2)$ rv's. These means are much too small and the Neyman smooth p-values too large to claim that $y_1, \cdots, y_n$ are iid N(0, 1). Clearly, $y_1, \cdots, y_n$ are supernormal for all of the values of n in Table 1.

One very interesting point to be observed in the above results is that the supernormality of the y-samples appears to be unrelated to sample size. In other words, this bias does not disappear with increasing sample size. The overfitting of the model appears to be as serious for n large (2,000 or 5,000) as for n small (10, 30, 60, 100). This same behavior will be observed in the two other examples to be reported below.

3. EXPONENTIAL SAMPLES

We consider next a sample $x_1, \cdots, x_n$ from a two-parameter exponential distribution $E(\mu, \theta)$ with density function

$$f(x) = (1/\theta) \exp\{-(x-\mu)/\theta\}, -\infty < \mu < \infty, \theta > 0, x > \mu.$$

2. Exponential Distribution Results

n	Mean	Neyman Smooth p-value
10	0.3504	0.99999881
30	0.3617	0.99999940
60	0,3694	0.99986041
100	0.4206	0.94826734
500	0.3958	0.99384993
2,000	0.3804	0.99999112
5,000	0.3697	0.99995726

We will use the ML estimators of μ and θ that have been adjusted to make them unbiased. These estimators are

$$\hat{\mu} = (n X_{(1)} - \overline{X})/(n - 1) ,$$

and

$$\hat{\theta} = n(\overline{X} - X_{(1)})/(n - 1) .$$

One hundred samples of size n were drawn from an $E(0, 1)$ distribution and transformed with a $E(\hat{\mu}, \hat{\theta})$ distribution function to obtain u-values as described in the last section for normal samples. Table 2 gives the results for exponential samples.

The trends observed in the last section in Table 1 for normal samples are observed here for exponential samples. The means of the v-values are less than 0.5, and the p-values of the Neyman smooth statistics computed on these samples are very large, implying that the values of the residuals $(x_i - \hat{\mu})/\hat{\theta}$ are "superexponential."

4. SIMPLE LINEAR REGRESSION

The third and final model we consider is the standard simple normal linear regression model. Let $(x_1, y_1), \cdots, (x_n, y_n)$ satisfy the usual

assumptions, i.e., that

$$y_i = \alpha + \beta x_i + \varepsilon_i, \; i = 1, \cdots, n \; , \tag{4.1}$$

the x_i's are nonrandom variables, and the ε_i's are iid $N(0, \sigma^2)$ rv's. Let a and b be the usual least squares estimates of α and β, and e_i the residual from the fitted regression line, i.e., $e_i = y_i - a - bx_i$.

Now, it is common practice to study the appropriateness of the above linear model, and also of its generalization to multiple linear regression, by analyzing the e_i residuals or some scaled version of them obtained by dividing each e_i by a scaling factor. These analyses are usually based upon an assumption that the residuals are approximately iid $N(0, 1)$ rv's or iid $N(0, \sigma^2)$ rv's.

When the model assumptions hold, then e_i has a marginal $N(0, \sigma^2 \delta_i^2)$ distribution, with

$$\delta_i^2 = 1 - (\Sigma x_j^2 - 2 x_i \Sigma x_j + n x_i^2)/(n \Sigma x_j^2 - (\Sigma x_j)^2) \; .$$

This has been the motivation for the choice of a number of scaling factors. We will consider here residuals using a number of scaling factors (see Seber (1977), section 6.6.2), as follows.

$$\begin{aligned} e_{1i} &= e_i/\sigma\delta_i \; , \\ e_{2i} &= e_i/s\,\delta_i \; , \\ e_{3i} &= e_i/s\sqrt{(n-2)/n} \; , \\ e_{4i} &= e_i/s \; , \\ e_{5i} &\text{ are independently generated iid } N(0, 1) \text{ rv's.} \end{aligned} \tag{4.2}$$

We have studied the normality of e_i and the other five residuals above (e_{5i} is included as a check and for comparison), by the methods

described above for n = 10 and 100. For n = 10, we generated 100 samples from the model

$$y = 5 + 2x + \varepsilon\,,\ \varepsilon \sim N(0,1)\,, \tag{4.3}$$

for x = 21, 22, ···, 30. The summary statistics for these samples are given in Table 3.

For n = 100, we generated sample values for the same linear model as above for n = 10, but with y replicated 10 times at each value of x. These results are also given in Table 3.

3. Normal Linear Regression Results, Normal Errors

Residual	n	Mean	P-value
e_i	10	0.3672	0.99997300
	100	0.3658	0.99999470
e_{1i}	10	0.3841	0.99947482
	100	0.3619	0.99999839
e_{2i}	10	0.2696	1.00000000
	100	0.2879	1.00000000
e_{3i}	10	0.2681	1.00000000
	100	0.2879	1.00000000
e_{4i}	10	0.2538	1.00000000
	100	0.2818	1.00000000
e_{5i}	10	0.4930	0.78271532
	100	0.4873	0.14018631

As we anticipated from our general remarks in section 1, the least squares residuals e_i show a rather strong supernormal tendency. We feel that it is interesting that all of the various scaled versions of the

least squares residuals also show this supernormal trend, and for some of them the trend is even more pronounced (e_2, e_3, e_4). Note that e_5, the control "residual," takes values for mean and p-value that are in the appropriate ranges.

Clearly, this supernormality of residuals has important implications that should be carefully considered when residuals of the type considered above are used in residuals analyses. If a correct model gives residuals that are supernormal, (superexponential, etc.) then it is reasonable to expect that some models that are not quite correct will give residuals that show normal (exponential, etc.) patterns. In order to study this tendency we have generated data from (4.3) where ε is a Laplace rv with density function

$$f(x) = (1/2)\, e^{-|x|}\,, -\infty < x < \infty\,. \tag{4.4}$$

These data were analyzed exactly the same way as the normal model data above. The results are given in Table 4. In this case the e_{5i} are iid rv's from the density (4.4). The other e's are the same as given in (4.2).

We call attention to certain points in Table 4. Note first that e_5, the iid Laplace rv's, are appropriately nonnormal with v-sample means much larger than 0.5 and p-values of one. Observe that e_i and e_{1i} behave very similarly, as they did also for normal errors. For $n = 10$, the v-sample means for e_i and e_{1i} are near enough to 0.5 for the y's to be normal, however, the p-values are too large for the v-sample to be uniform.

The most interesting results of Table 4 are those for e_{2i}, e_{3i}, and e_{4i}, which also are very similarly behaved. From this observation

4. Normal Linear Regression Results, Laplace Errors

Residual	n	Mean	P-value
e_i	10	0.5345	0.99952871
	100	0,9421	1.00000000
e_{1i}	10	0.5295	0.99999261
	100	0.9466	1.00000000
e_{2i}	10	0.3022	1.00000000
	100	0.8547	1.00000000
e_{3i}	10	0.3076	0.99999994
	100	0.8549	1.00000000
e_{4i}	10	0.3107	0.99999994
	100	0.8639	1.00000000
e_{5i}	10	0.6125	0.99999988
	100	0.9594	1.00000000

and the one above about e_i and e_{1i} it appears that an influential factor in the behavior of these residuals is the choice of σ or s as the divisor. Note that e_{2i}, e_{3i} and e_{4i} are all clearly supernormal for $n = 10$ and nonnormal for $n = 100$. Thus the general bias resulting from the estimation of parameters discussed in section 1 appears to be at work here, and tends to be more pronounced with the estimation of an additional parameter, as we should expect.

5. COMPUTATIONS

The computations for this study were performed at the Triangle Universities Computation Center (TUCC) using Fortran programs written for this project. The Monte Carlo subroutines of David A. Dickey were

used for generating random variates and plotting histograms, and some IMSL functions were also used. The random number generator in Dickey's subroutine is the "Super Duper" developed at McGill University.

REFERENCES

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, Vol. 2. Charles Griffin and Co., Ltd. London.

Miller, F. L., Jr. and Quesenberry, C. P. (1979). Power studies of tests for uniformity, II. *Commun. Statist. - Simula. Computa.* B8(3):271-290.

Neyman, Jerzy (1937). "Smooth" test for goodness of fit. *Skandinavisk Abtuarietidskrift*, 20, 149-199.

Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley & Sons, New York.

* C. P. Quesenberry is Professor of Statistics, North Carolina State University, Raleigh, NC, 27650. Charles Quesenberry, Jr. is presently a graduate student in Biostatistics at the University of California at Berkeley. The second author's work was performed at North Carolina State University as part of a Senior Honors Program project.