On some concepts of residuals

Georgi N. Boshnakov

Mathematics Department
University of Manchester Institute of Science and Technology
P O Box 88
Manchester M60 1QD, UK

E-mail: georgi.boshnakov@umist.ac.uk

Tel: (+44) (0)161 200 3684
Fax: (+44) (0)161 200 3669

June 16, 2003

Abstract

We introduce confidence residuals and standardised confidence residuals. These residuals may be especially useful for asymmetric and multimodal distributions.

Keywords: concentration function, confidence density, confidence residual, highest density region

1 Introduction and preliminaries

Residuals measure discrepancy between a statistical model and observed data and are used to evaluate the quality of the model either through summary statistics or directly, e.g., via plots. In prediction, measures of uncertainty are often based on residuals as well.

Let a quantity of interest be modelled as a random variable, Y, with distribution μ. When an observation y of Y becomes available the ordinary definition of residual is

$$y - \hat{y}, \qquad \text{((location) residual)}, \tag{1}$$

where $\hat{y}$ is a measure of location such as the mean, median, or the tallest mode of μ. This residual can be standardised with a scale parameter, σ,

$$\frac{y - \hat{y}}{\sigma}, \qquad \text{(standardised residual)}, \tag{2}$$

in order to remove heteroscedasticity and make residuals for different distributions from a given class comparable. In the case when σ is the standard deviation this residual is called the Pearson residual.

The term residual is often reserved for the case when μ contains estimated parameters. Numerous variations of the residuals mentioned here arise when some of the quantities are replaced by their sample analogues or other approximations (see, for example, McCullagh and Nelder (1989), Pierce and Schafer (1986)). For our purposes it is sufficient to work with the "theoretical" characteristics of μ since we are concerned with the concepts here.

The above residuals work well in normal linear models but may be of limited value in non-linear models. This has been long recognised in generalized linear models, where residuals have received considerable attention (see McCullagh and Nelder (1989, pp. 37–40), Pierce and Schafer (1986)). One approach is to replace y in equation (2) by a function $t(y)$ chosen so that residuals from a non-Gaussian distribution behave like those from Gaussian distributions:

$$\frac{t(y) - t(\hat{y})}{\mathrm{Var}(t(Y))}.$$

An example of this is the Anscombe residual (McCullagh and Nelder, 1989, p. 37). Even more sound theoretical basis has the deviance residual which for our purposes may be defined by

$$d = \mathrm{sgn}(y - a)\sqrt{2(\log f(M) - \log f(y))},$$

where f is the density of μ, M is its maximal mode, and a is the mean. For a normal density d is $(y - a)/\sigma$.

A recent discussion of residuals for survival data is given by Therneau and Grambsch (2000, Chapter 4).

Many parametric models express a new or future observation as a function of, among others, error (or, in time series context, innovation) term. A general definition of residuals can be based on inverting this function with respect to the error term (Cox and Snell, 1968), it that is possible. In some models there is no explicit error term or there are more of them. For example, there is no explicit error term in a mixture model specified by a mixture distribution but there are two such terms if the mixture is specified by two random variables, one to choose the mixture component, the other being the outcome of a draw from the distribution of that component. In such cases the very concept of residual becomes vague.

We briefly discuss residuals from a general viewpoint in Section 3, emphasise their relation to transformations (such as power transformations) and indicate how new types of residuals may be introduced. Seemingly remote quantities qualify for this extended notion of residual. The treatment seems novel even though it is based on the elementary fact that if X is a random variable with continuous distribution function F, then $F(X)$ is a uniform random variable.

In Section 4 we define a confidence residual as (roughly) the length of the region where the density is larger or equal to the density at the observed value. In Section 2 we provide some background material about the confidence transform (Boshnakov, 2003) which is a very powerful tool for studying "concentration" properties of probability distributions. Examples of such properties are highest density regions (Hyndman, 1996) and highest posterior density regions (Box and Tiao, 1973), for more details see Boshnakov (2003). The confidence residuals complement the tools for presentation of predictors in the multimodal case developed by Hyndman (1996).

The confidence residual measures quite naturally how "far" a particular value is from the "most probable" value. Confidence residuals for variables having different distributions are comparable as long as the variables are measured in the same units, in contrast to deviance residuals which, in general, are not. We also define standardised confidence residuals in the spirit of Section 3 which may be useful in assessing goodness of fit.

2 The confidence transform

The confidence transform (Boshnakov, 2003) maps the probability distributions to a smaller class of distributions with excellent analytical properties. It preserves many "concentration" and "spread" properties of the originals.

For absolutely continuous distributions it "rearranges", in the sense of Hardy et al. (1952), the (possibly weird) original densities into nice monotonically decreasing densities. The sup operation below can be interpreted as taking the "densest" region of the distribution μ for any given size l.

Definition 1. *Let μ be a probability measure and λ be the Lebesgue measure on R. The confidence characteristic of μ is the distribution ν whose distribution function is defined by*

$$G(l) = \begin{cases} 0, & \text{if } l < 0, \\ \sup\{\mu(A) : \lambda(A) \leq l, A \in \mathcal{B}\}, & \text{if } l \geq 0. \end{cases} \tag{3}$$

$G(\cdot)$, its derivative $g(\cdot)$, and ν are called the confidence distribution function, confidence density, and confidence characteristic of μ, respectively. The map from μ to ν is called confidence transform. More generally, any property of the confidence characteristic is prefixed with the qualifier "confidence" when referred to as a property of μ.

The function G is known as the decomposition concentration function of μ (Hengartner and Theodorescu, 1973, p. 110) but there are many concepts of "concentration function". The above definition provides consistent terminology for any property of the confidence characteristic.

On $(0, \infty)$, the confidence distribution function is continuous, concave and has everywhere left- and right- derivatives, the confidence density exists and is monotonic and continuous except possibly on a countable set.

In what follows we will assume that μ is absolutely continuous with density f. In that case the confidence density is obtained, effectively, by reordering the values of the original density in decreasing order so that the probability content and length of regions of the form $\{x : f(x) > c\}$ and $\{l : g(l) > c\}$ are the same, the precise meaning of this is discussed in Section 4. The confidence transform preserves the entropy. More generally,

$$\int U(f)dF = \int U(g)dG$$

for any function U for which the integrals exist (Boshnakov, 2003, Theorem 3).

3 Residuals and transformations

Let Y be a random variable with continuous distribution function F and H be a strictly monotonic distribution function. The following definition

probably stretches the notion of residual to the limit but it is consistent with the usual practice of centering and standardising (see below) and will be used to define standardised confidence residuals in Section 4.

Definition 2. *The H-residual of F at x is a value $z = z(x)$ such that*

$$H(z) = F(x). \tag{4}$$

In principle, the H-residual may be obtained by taking the inverse of H:

$$z = H^{-1}(F(x)).$$

Assuming that F and H are absolutely continuous and z differentiable at x, we get from the above

$$h(z)\frac{dz}{dx} = f(x). \tag{5}$$

The solution of this differential equation, for given f and h, will normally be given by the implicit equation (4). For given $z(x)$ and h equation (5) is a change of variables formula and may be explored to see which distributions can be converted to a given distribution, h, by a given transformation z. Care must be taken in applying formula (5) in this way to ensure that $z(x)$ indeed covers the support of H.

For example, let $z = (x - a)/b$, $b > 0$. Then $dz/dx = 1/b$ and from (5) we get $f(x) = \frac{1}{b}h((x-a)/b)$, i.e., this transformation works for distributions from the shift-scale family generated by h. More exotic distributions result for other standard transformations, such as the power or Box-Cox family of transformations (indexed by λ),

$$z_{\lambda,\mathrm{BC}} = \frac{x^\lambda - 1}{\lambda}, \qquad x \geq 0,$$

where $z_{0,\mathrm{BC}} := \log x$ by definition.

The close relationship between residuals and transformations (to normality) can be seen particularly well when H is the distribution function of a normal distribution. The main difference is in the purpose of these tools. Residuals are often used to provide diagnostics for a model. In this case the distribution (i.e., F) suggested by the model may be taken as given and an exact transformation applied with the desired H. On the other hand, transformations like $z_{\lambda,\mathrm{BC}}$ are applied to raw data in order to make them more conformable to (usually) a normal distribution. In this case a particular form of F can rarely be justified and one looks for transformations that "generally" work well and thus are approximate. For example, the power transformations $z_{\lambda,\mathrm{BC}}$ are used to make the data "more normal" even though $z_{\lambda,\mathrm{BC}}$ cannot

be exactly normal for $\lambda \neq 0$. If for some reason an exact transformation is preferred, then a "target" may be chosen, such as a χ^2-distribution. For example, x^λ converts densities of the form $cx^{\lambda\alpha-1}e^{-\beta x^\lambda}$ to Gamma densities which may be close to normal densities, such as the χ^2-density obtained for $\beta = \frac{1}{2}$ and large α.

Although the normal distribution would appear to be the most natural choice of H, for non-negative variables the exponential distribution may be considered. Also, the choice of the uniform distribution on $[0, 1]$ ($H(x) = x$) gives $z = F(x)$, which is not called a residual but is used in diagnostic tools, such as probability plots (see also at the end of Section 4).

4 Confidence residuals

We wish to define a residual at a point x_0 by measuring the length of the region were the density, f, of μ is larger than the density at x_0 but we do not wish our definition to depend on values of the density at individual points. We therefore use a possibly modified value of the density arranging for it to correspond to the intuitive idea of having positive mass around the chosen level.

Definition 3. *We say that the height of f at x_0 is at least c if the intersection of the set $\{x : f(x) \geq c\}$ with each open interval containing x_0 has positive λ-measure. We say that the height of f at x_0 is $f^*(x_0)$ if*

$$f^*(x_0) = \sup\{c \geq 0 : \textit{the height of } f \textit{ at } x_0 \textit{ is at least } c\}.$$

If f is continuous at x_0 then $f^*(x_0) = f(x_0)$. Also, if f has left and right limits at x_0 then $f^*(x_0)$ is equal to the larger of them.

The argument of the confidence density and confidence distribution function has the meaning of a residual. We formalise this as follows.

Definition 4. *The confidence residual of μ at x is defined by*

$$r(x) = \begin{cases} \sup\{l : g(l) > f^*(x)\}, & \textit{if } f^*(x) < \infty, \\ 0, & \textit{if } f^*(x) = \infty. \end{cases}$$

A unimodal density decreases monotonically on each side of the mode. So, the confidence residual is the length of a stretch connecting two points on either side of the mode having the same density. If, in addition, the density is symmetric the confidence residual is twice the absolute value of the location residual. If the density is asymmetric, the difference between the

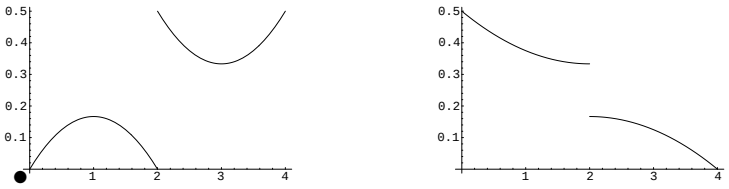

Figure 1: A density (left), and the corresponding confidence density (right). The height of f at the local maximum $x_1 = 1$ and at the local minimum $x_2 = 3$ is $f^*(x_1) = f(x_1)$ and $f^*(x_2) = f(x_2)$, respectively. Both points have the same confidence residual, $l_0 = l(x_1) = l(x_2) = 2$. The confidence density "plunges" from $f(x_2)$ to $f(x_1)$ at l_0 since f does not take values in the interval $(f(x_1), f(x_2))$.

two types of residual becomes essential, e.g., equal confidence residuals do not necessarilly correspond to equal location residuals. The two types of residual become even more different when the mean is used in the computation of the location residual, rather than the mode. They are incomparable for multimodal distributions where the confidence residual may be the sum of the lengths of several pieces.

If g is continuous at $r(x)$, the "usual" case, then $g(r(x)) = f^*(x)$. Note that the confidence density may be discontinuous at a point l_0 only if the set where f takes values between the left and right limits of g at l_0 is of zero λ-measure (Boshnakov, 2003). Figure 1 depicts this for the density

$$f(x) = \begin{cases} \frac{1}{6}(1 - (x-1)^2), & \text{for } 0 \le x < 2, \\ \frac{1}{6}(2 + (x-3)^2), & \text{for } 2 < x \le 4, \end{cases}$$

whose confidence density is

$$g(l) = \begin{cases} \frac{1}{24}(12 - 4l + l^2), & \text{for } 0 \le l < 2, \\ \frac{1}{24}(4l - l^2), & \text{for } 2 < l \le 4. \end{cases}$$

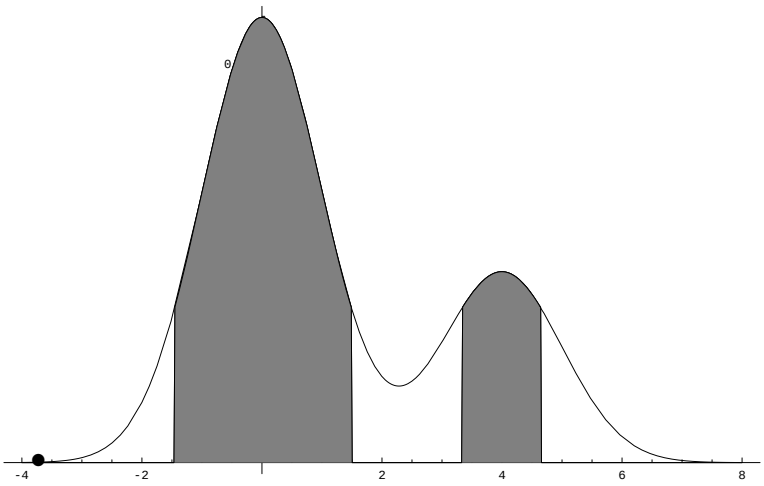

Figure 2: The pdf of a mixture of two normal distributions with the 75% highest density region.

Figure 2 shows the probability density of a mixture of two normal densities used by Hyndman (1996). The shaded area is the 75% highest density region. The corresponding confidence density and confidence distribution function are given in Figures 3 and 4. The confidence density shows the height of the density at a point with given confidence residual, e.g., $f(4.1) = g(4.1) \approx 0.1$, while the confidence distribution function shows the corresponding cumulative probability, e.g., $F(4.1) = G(4.1) \approx 0.75$. These values correspond to the height of the density at the end points of the confidence region in Figure 1.

The following theorem shows that the confidence residual does not depend on the particular choice of the density. Above we used the height $f^*(x_0)$ as a means to define the confidence residual. Now it is convenient to look at $f^*(x)$ as a function defined on the domain of μ.

Theorem 1. *The height function $f^*(x)$ does not depend on the particular choice of the density of the absolutely continuous measure μ.*

Proof. The result may be obtained from the definition of f^* and the fact that any two densities of μ may differ on a set of λ-measure 0 only. We use another argument which may be useful in other circumstances as well. Assume that the claim is not true. Then there exist two densities f_1 and f_2 and a point x_0 such that $f_1^*(x_0) < f_2^*(x_0)$. Let c_1 and c_2 be such that $f_1(x_0) \leq c_1 < c_2 \leq f_2(x_0)$. From the definition of height it follows that

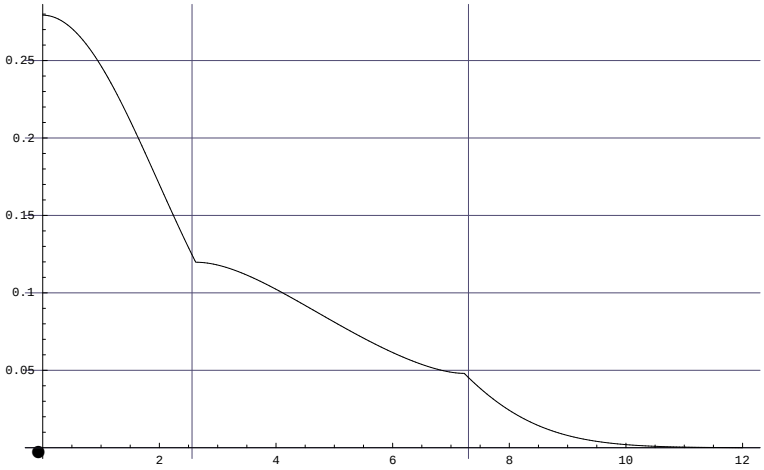

Figure 3: The confidence density of the mixture from Figure 2 gives information about the speed of decrease of the original density. For example, the length of the region where the density is more than half its maximal height is about 2.5. The slope may change abruptly at heights corresponding to peaks and valleys of the original density.

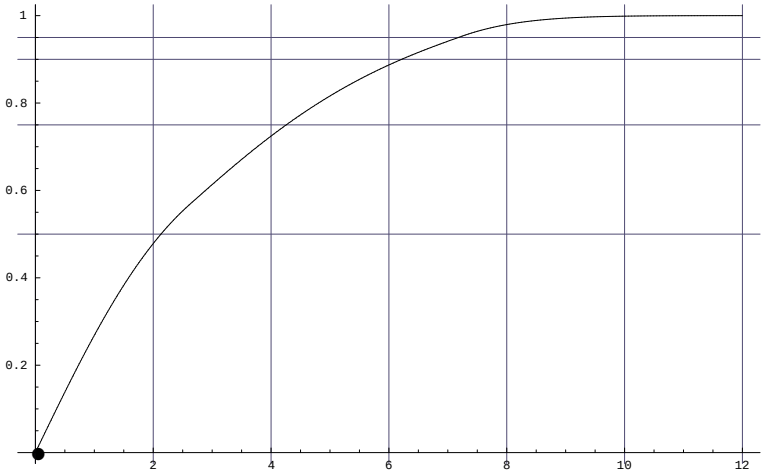

Figure 4: The confidence distribution function, G, of the mixture from Figure 2. The length of the 75% highest density region is approximately 4.2 since $G(4.2) \approx 0.75$.

there exists a set S with positive λ-measure γ, say, such that $f_2(x) \geq c_2$ for $x \in S$ and at the same time $f_1(x) \leq c_1$ on S except possibly for a subset of λ-measure 0. But this means that $\int_S f_1 \leq c_1\gamma < c_2\gamma \leq \int_S f_2$, i.e., the μ-measure of S is different for f_1 and f_2 which contradicts the assumption. □

Without additional assumptions, the height function of f cannot be guaranteed to be a density and if it is, it may be that of a different distribution. This is not essential for the definition of the confidence residuals since it is based on the confidence density. Nevetheless, the property is desirable and it makes sense to formulate the following result.

Theorem 2. *The height function, f^*, of μ is a density of μ if and only if $f(x) \neq f^*(x)$ on a set of λ-measure 0.*

The condition of this theorem holds in the most important for applications case when the density f is continuous almost everywhere.

It may be difficult to compare densities having different shapes, for example when comparing predictors from different non-linear models. The confidence density neatly represents the concentration of a distribution. By superimposing confidence densities meaningful comparisons are made possible, for example by their peakedness (Boshnakov, 2003).

One of the virtues of the confidence residuals is that they are on the same scale as the original variables. Thus confidence residuals from non-identically distributed data may be pooled together. If desired, the confidence residuals may be standardised in the spirit of Section 3 as follows.

Definition 5. *Let μ be a distribution with distribution function F and confidence distribution function G. For any x from the domain of μ the standardised confidence residual of μ at x is defined as*

$$z(x) = H^{-1}(G(l)),$$

where $l = l(x)$ is the confidence residual of μ at x.

Choosing the uniform distribution with $H(x) = x$ gives $z = G(l)$, i.e., the standardised confidence residuals, $z = z(Y) = G(Y)$, are uniformly distributed. Of course, $F(Y)$ is also uniformly distributed for continuous F, a fact used for diagnostics such as probability plots. Note that while $F(y)$ is the probability to get a smaller observation, $G(y)$ is the probability to get a "more likely" observation. Thus $G(y)$ close to zero corresponds to an observation with high density and so corresponds to a "good" observation. For comparison, $F(y) < \frac{1}{2}$ shows that y is to the left of the median, while $G(y) < \frac{1}{2}$ coresponds to an observation that is "more likely" than more than

"half" of the observations. So, if we have data y_i from F_i with confidence distribution function G_i, $i = 1, \ldots, n$, then the standardised confidence residuals z_i are uniformly distributed on $[0, 1]$. For example, we may be worried if too many of them turn out to be larger than $\frac{1}{2}$. In theory, we should be equally worried if there are too many small z_is but one rarely worries when the data fit a model too well.

If the confidence density is constant, say c_0, over an interval, then points with $f^*(x) = c$ will be allocated a residual corresponding to the left-hand end of that interval. In the extreme case of a uniform distribution the confidence residuals are identically zero. This will happen also with deviance residuals and illusrates that it may be necessary to examine the confidence density along with the residuals. In fact, it does not make sense to compute residuals for uniformly distributed variables since these would introduce the false impression that some values are "better" than others. In a sense, uniformly distributed random variables constitute the ultimate residuals.

5 Conclusion

We introduced confidence residuals and standardised confidence residuals. Such residuals may be useful when accessing the quality of non-linear models particularly when the distributions involved have different shapes. The confidence residuals preserve the measurement scale of the variables. They also resolve the difficulty of defining residuals for some models. Applications of confidence residuals for diagnostics of some time series and other models are in preparation, e.g., Boshnakov (2003). Extention to the multivariate case is possible if in the definitions given here the Lebegue measure is taken to be in R^k even though in practice the numerical calculations may be plagued with the curse of dimensionality.

References

Boshnakov G. N. (2003) *Confidence characteristics of distributions*, Statistics & Probability Letters (to appear).

——— (2003) *Prediction with mixture autoregressive models*, (preprint, UMIST).

Box G. E. P., Tiao G. C. (1973) *Bayesian inference in statistical analysis*, Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, Mass. etc.: Addison-Wesley Publishing Company.

Cox D.R., Snell E.J. (1968) *A general definition of residuals*, J. R. Stat. Soc., Ser. B **30**, 248–265, Discussion 265–275.

Hardy G.H., Littlewood J.E., Pólya G. (1952) *Inequalities. 2nd ed.*, Cambridge: At the University Press.

Hengartner W., Theodorescu R. (1973) *Concentration functions.*, Probability and Mathematical Statistics. 20. New York - London: Academic Press, a subsidiary of Harcourt Brace Jovanovich, Publishers.

Hyndman R.J. (1996) *Computing and graphing with highest density regions*, The American Statist. **50**, 120–126.

McCullagh P., Nelder J.A. (1989) *Generalized linear models. 2nd ed.*, Monographs on Statistics and Applied Probability. 37. London etc.: Chapman and Hall, ISBN 0-412-31760-5.

Pierce Donald A., Schafer Daniel W. (1986) *Residuals in generalized linear models.*, J. Am. Stat. Assoc. **81**, 977–986.

Therneau Terry M., Grambsch Patricia M. (2000) *Modeling survival data: Extending the Cox model*, Statistics for Biology and Health. New York, NY: Springer.