Randomized Quantile Residuals

Peter K. Dunn and Gordon K. Smyth
Department of Mathematics, University of Queensland,
Brisbane, Q 4072, Australia.

24 April 1996

Abstract

In this paper we give a general definition of residuals for regression models with independent responses. Our definition produces residuals which are exactly normal, apart from sampling variability in the estimated parameters, by inverting the fitted distribution function for each response value and finding the equivalent standard normal quantile. Our definition includes some randomization to achieve continuous residuals when the response variable is discrete. Quantile residuals are easily computed in computer packages such as SAS, S-Plus, GLIM or LispStat, and allow residual analyses to be carried out in many commonly occurring situations in which the customary definitions of residuals fail. Quantile residuals are applied in this paper to three example data sets.

Keywords: deviance residual; exponential regression; generalized linear model; logistic regression; normal probability plot; Pearson residual.

1 Introduction

Residuals, and especially plots of residuals, play a central role in the checking of statistical models. In normal linear regression the residuals are normally distributed and can be standardized to have equal variances. In non-normal regression situations, such as logistic regression or log-linear analysis, the residuals, as usually defined, may be so far from normality and from having equal variances as to be of no practical use. A particular problem occurs when the response variable is discrete and takes on a small number of distinct values, as for Poisson data with mean not far from zero or binomial data with mean close to either zero or the number of trials. In such situations the residuals lie on nearly parallel curves corresponding to distinct response values, and these spurious curves distract the eye seriously from any meaningful message that might be contained in a residual plot.

In this paper we give a general definition of residuals for regression models with independent responses. Our definition produces residuals which are exactly normal, apart from sampling variability in the estimated parameters, by inverting the fitted distribution function at each response value and finding the equivalent standard normal quantile. This approach is closely related to that of Cox and Snell (1968), but whereas Cox and Snell concentrate on mean and variance corrections we concentrate on the transformation to normality. Our definition includes some randomization to achieve continuous residuals when the response variable is discrete. Quantile residuals are easily computed in computer packages such as SAS, S-Plus, GLIM or LispStat, and allow residual analyses to be carried out in many commonly occurring situations in which the customary definitions of residuals fail.

Special cases of quantile residuals have been used by Brillinger and Preisler (1983) and Brillinger (1996). For other work on residuals for non-normal regression models see Pierce and Schafer (1986) or McCullagh and Nelder (1989) and the references therein. In the discussion at the end of the paper we briefly indicate how quantile residuals may be extended to models with dependent responses.

2 Pearson and Deviance Residuals

Let $y_1, \ldots, y_n$ be responses and for each i let $\mathbf{x}_i$ be a vector of covariates. The y_i are assumed to be independent and to follow a distribution $\mathcal{P}(\mu_i, \boldsymbol{\phi})$ where $\mu_i = E(y_i)$ and $\boldsymbol{\phi}$ is a parameter vector common to all the y_i. The μ_i are assumed to depend on the $\mathbf{x}_i$ and a vector of regression parameters $\boldsymbol{\beta}$. We have particularly in mind generalized linear models (McCullagh and Nelder, 1989) in which the probability density or mass function of y_i has the form

$$f(y; \theta_i, \phi) = a(y, \phi) \exp[\{y\theta_i - \kappa(\theta_i)\}/\phi]$$

where $a()$ and $\kappa()$ are known functions and $\mu_i = \kappa'(\theta_i)$. In this model we have $\mathrm{var}(y_i) = \phi V(\mu_i)$ where $V(\mu_i) = \kappa''(\theta_i)$. It is customary to assume that $g(\mu_i) = \mathbf{x}^T\boldsymbol{\beta}$ where $g()$ is a known link function. The parameter ϕ is the proportionality constant in the mean-variance relationship and is known as the dispersion parameter.

In the context of generalized linear models, two definitions of residuals have been commonly used in practice. The Pearson residual is defined by

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{V(\hat{\mu}_i)^{1/2}}$$

where $\hat{\mu}_i$ is the fitted value for μ_i. The Pearson residual has the advantage that its mean and variance are exactly zero and ϕ respectively, if sampling variability in $\hat{\mu}_i$ is small. The deviance residuals are defined in terms of the unit deviances. For the above model, let $t(y, \mu) = y\theta - \kappa(\theta)$. Assuming that y is in the domain of μ, the unit deviance is

$$d(y, \mu) = 2\{t(y, y) - t(y, \mu)\}$$

The deviance residual is

$$r_{d,i} = d(y_i, \hat{\mu}_i)^{1/2} \mathrm{sign}(y_i - \hat{\mu}_i)$$

Pierce and Schafer (1986) have argued on theoretical grounds that the deviance residuals should be more nearly normal than the Pearson. Indeed both converge to normality as $\phi \to 0$ relative to the μ_i, the Pearson residuals at rate $O(\phi^{1/2})$ by the Central Limit Theorem and the deviance residuals at $O(\phi)$ by the saddle-point approximation to $f(y;\theta_i,\phi)$. The Pearson and deviance residuals coincide and are exactly normal, ignoring variability in $\hat{\mu}_i$, for the normal linear model. The deviance residual is also exactly normal when the response is inverse-Gaussian. In other cases and for ϕ/μ large however, neither type of residual can be guaranteed to be closely normal, and the deviance residuals do not generally have zero means or equal variances even at the true values μ_i.

3 Randomized Quantile Residuals

Let $F(y;\mu,\phi)$ be the cumulative distribution function of $\mathcal{P}(\mu,\phi)$. If F is continuous, then the $F(y_i;\mu_i,\phi)$ are uniformly distributed on the unit interval. In this case, the quantile residuals are defined by

$$r_{q,i} = \Phi^{-1}\{F(y_i;\hat{\mu}_i,\hat{\phi})\}$$

where $\Phi()$ is the cumulative distribution function of the standard normal. Apart from sampling variability in $\hat{\mu}_i$ and $\hat{\phi}$, the $r_{q,i}$ are exactly standard normal. This implies that the distribution of $r_{q,i}$ converges to standard normal if $\boldsymbol{\beta}$ and ϕ are consistently estimated. The above definition is a special case of Cox and Snell's (1968) "crude" residuals.

Example 1: Leukemia data. Feigl and Zelen (1965) discuss some data relating the survival times y_i of leukemia patients to their initial white blood cell counts x_i and to existence of AG-factor. Following Feigl and Zelen, we treat the survival times as exponential, $y_i \sim \mathrm{Exp}(\mu_i)$. We work with a log-linear model for the means, including separate intercepts for the two AG-factor groups,

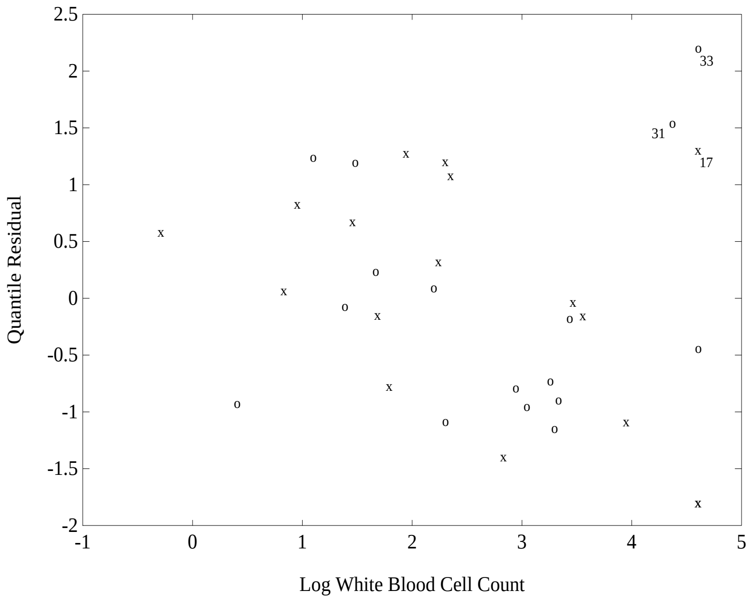
$$\log\mu_i = \begin{cases} \alpha_1 + \beta\log x_i & \text{AG positive} \\ \alpha_2 + \beta\log x_i & \text{AG negative} \end{cases}$$

Cox and Snell (1968) considered a subset of this data, and defined approximately exponential crude residuals $R_i = y_i/\hat{\mu}_i$, where the $\hat{\mu}_i$ are the estimated means. In this case the quantile residuals

$$r_{q,i} = \Phi^{-1}\{1 - \exp(y_i/\hat{\mu}_i)\}$$

are a simple transformation of the R_i. A normal probability plot of the quantile residuals confirms the assumption of an exponential distribution. Figure 1 plots the quantile residuals versus the covariate. The three residuals (cases 17, 31 and 33) in the upper right-hand corner of the plot are relatively separate from the body of the other residuals, and without them there appears to be a marked negative trend. While the pattern is not sufficient to contradict the model assumptions, it raises the possibility that cases 17, 31 and 33 may be outliers, or that the dispersion of the residuals increases at the largest white blood cell counts. In any case, the three cases identified appear from the residual plot to be jointly influential. Assigning the identified cases zero weight increases $\hat{\beta}$ nearly three-fold, from -0.30 to -0.84 compared with a standard error of 0.14.

Figure 1: Plot of quantile residuals versus the covariate for the leukemia data. Circles represent patients which are AG-positive, crosses AG-negative.



If F is not continuous, a more general definition of quantile residuals is required. Let $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i, \hat{\phi})$ and $b_i = F(y_i; \hat{\mu}_i, \hat{\phi})$. We define the randomized quantile residual for y_i by

$$r_{q,i} = \Phi^{-1}(u_i)$$

where u_i is a uniform random variable on the interval $(a_i, b_i]$. Again, the $r_{q,i}$ are exactly standard normal, apart from sampling variability in $\hat{\mu}_i$ and $\hat{\phi}$. The randomization strategy employed here is similar to the strategy of jittering (Chambers et al, 1983) to prevent masses of overlapping points in plots. Whereas jittering applies a uniform random component to the response, our uniform random component is on the cumulative probability scale and is tailored to the actual probability mass at the point in question. Our randomization is the minimum necessary so that no granularity remains in the resulting residual distribution.

Example 2: Simulated binomial data. A logistic linear regression was used to model 60 binomial observations with binomial denominator $n = 3$, i.e., the responses were assumed to be independently distributed as $y_i \sim \text{bin}(n, p_i)$, with $n = 3$ and $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$ were x_i is a covariate. The first plot of Figure 2 displays the deviance residuals versus the covariate. The points in this plot lie on four parallel curves corresponding to the four possible values for the response. The curves make it difficult to see any other pattern in the data. The second plot displays the quantile residuals versus the covariate. In this plot is clear that the residuals follow a quadratic pattern. The data for this example was in fact computer generated with $\text{logit}(p_i)$ depending quadratically on the x_i. Figure 3 shows the residual plots once the quadratic term has been included in the regression. The

Figure 2: Deviance and quantile residuals versus the covariate from a logistic regression. The response is simulated $\mathrm{bin}(3, p)$ with logit p depending quadratically on the covariate.

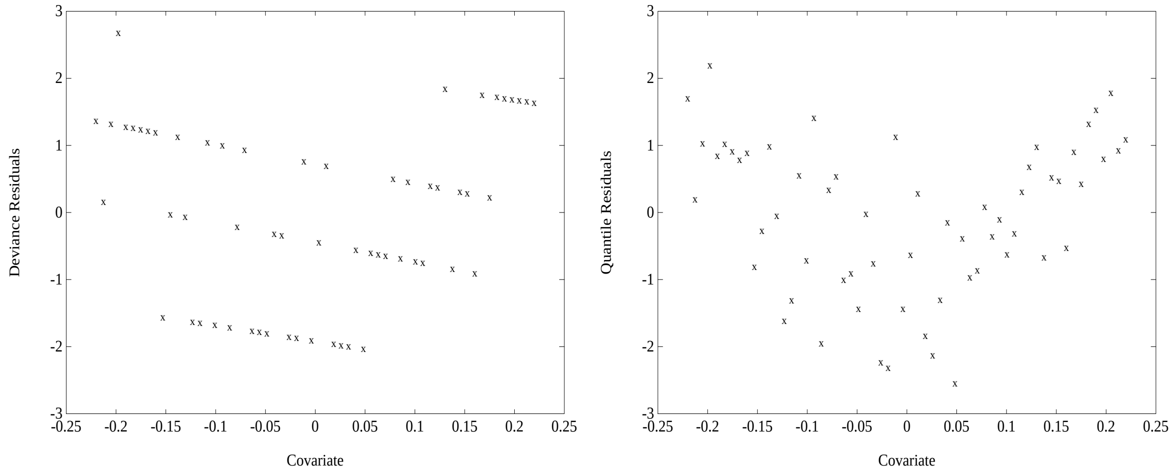


Figure 3: Deviance and quantile residuals versus the covariate for a well fitting logistic regression.

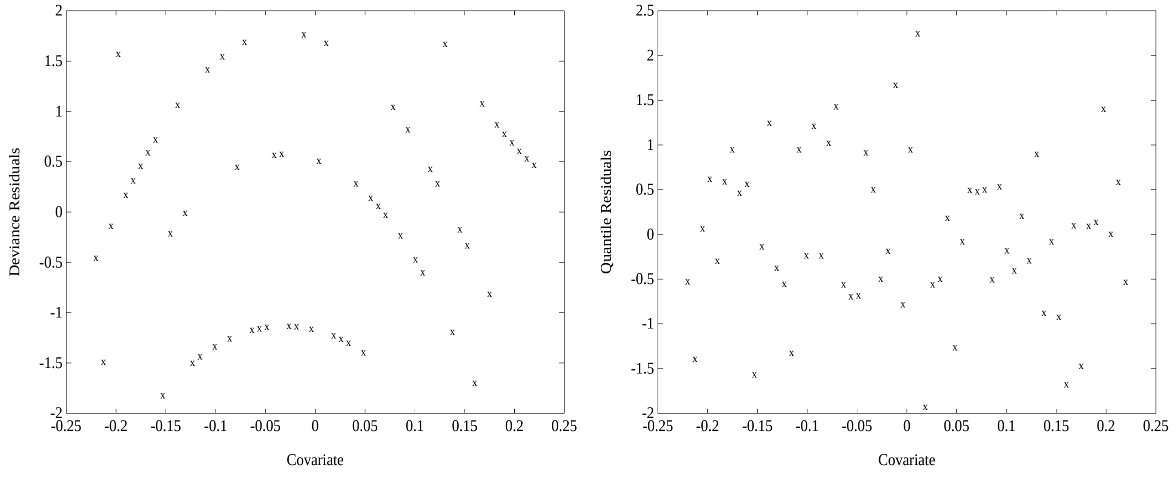
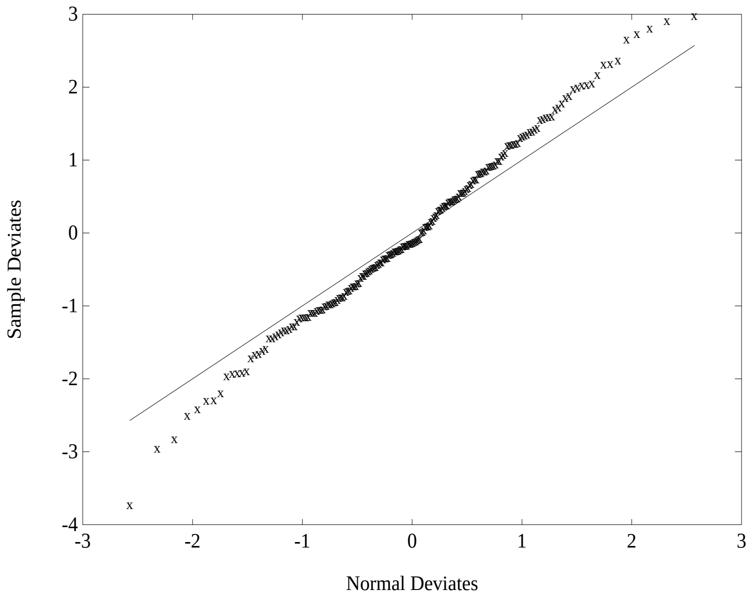


Figure 4: Normal probability plot with identity line of the quantile residuals from the fathers' and sons' occupation data.



deviance residuals lie on prominent curves while the quantile residuals now show random scatter.

Example 3: Fathers' and sons' occupations. Brown (1974) and Kotze and Hawkins (1984) analyze a sparse 14×14 contingency table showing the cross-classification of occupations of fathers (rows) by occupations of sons (columns). The data was originally published by Pearson (1904) and appears also in Hand et al (1994). Brown, Kotze and Hawkins were interested in identifying those cells which are outliers relative to the independence model. We take a similar approach, with the difference that the quantile residual approach allows us to look for outliers relative to a more realistic model. Observing that there is an apriori expectation that sons will be influenced by their father's occupation, we fit a log-linear Poisson regression model to the counts with row and column effects and with an effect for equality of father's and son's occupation, i.e., $y_{ij} \sim \text{Pois}(\mu_{ij})$, with

$$\log \mu_{ij} = \mu_0 + \alpha_i + \beta_j + \delta x_{ij} \tag{1}$$

and $x_{ij} = 1$ if $i = j$ and 0 otherwise. Figure 4 is a normal probability plot of quantile residuals from this model. The largest positive residual corresponds to the (2,2) cell: sons almost always continue to work in the Arts if their father did. Figure 4 shows evidence of large negative residuals as well as large positive residuals. Although none of the negative residuals are individually significant, and the actual contingency table cells represented in the left tail of the probability plot varies with each realization of the quantile residuals, the overall pattern is preserved across realizations. The quantile residual plot shows in this way that there are too many small counts in the contingency table to be compatible with the above model. No other method which has been applied to this data in the literature is able to show this aspect of the data. Although Figure 4 shows clear evidence of lack of fit, the model (1) and the models which arise from it by deleting selected cells does give an appreciably better fit to this data than the independence models considered by earlier authors.

4 Discussion and Extensions

In this paper quantile residuals are computed by finding the equivalent standard normal deviate for each response observation. In principle, any reference distribution could have been chosen for the residuals. Cox and Snell (1968) for example computed exponential residuals for data of Example 1 and Brillinger and Preisler (1983) and Brillinger (1996) use uniform residuals. However asymmetry seems an unnecessary complication, and bounded distributions introduce a spurious pattern (the boundary itself) and make it difficult to distinguish between large residuals and outright outliers. The normal distribution is recommended in this paper on the basis that normal variation is that which most people have practice interpreting graphically.

Randomization is used to produce continuously distributed residuals when the response is discrete or has a discrete component. This means that the quantile residuals will vary from one realization to another for a given data set and fitted model. For the sake of brevity, we have given only one realization of the quantile residuals for each example in this paper. In practice though we have found it useful to routinely plot four realizations of the quantile residuals. Any pattern in the residuals which is not consistent across the realizations is then ignored. The idea of applying a continuous random component to discrete responses so that methods for continuous variables can be applied is in fact very old. See Pearson (1950) for a discussion. As used in this paper, randomization is a device through which the aggregate pattern of the residuals becomes apparent. Since decisions do not depend on individual realizations, the obvious objections to randomization which arise in the context of tests and confidence intervals do not seem to apply.

Quantile residuals can generalize any of the usual diagnostic methods which use residuals. For example, an added variable plot (Cook and Weisberg, 1982) could be computed for a generalized linear model by plotting the quantile residuals, for the model excluding x, versus x_a, where x_a is x adjusted for the other covariates in the model. The vector x_a would be chosen to be orthogonal to the other covariates, relative to the covariance matrix of the y_i. It might be computed as the residuals from weighted least squares regression of x on the other covariates, using as weights the working weights from the generalized linear model.

Independence of the response observations was assumed in this paper. The method of quantile residuals can be extended to dependent data situations by expressing the multivariate likelihood as a sum of univariate conditional likelihoods. For example we might define the ith conditional quantile residual from the conditional distribution of y_i given $y_1, \ldots, y_{i-1}$ instead of from the marginal distribution of y_i as in the paper. This would provide independent, standard normal residuals.

Finally we consider the sampling variability of the $\hat{\mu}_i$, which has for simplicity been ignored throughout this paper. Treating the $\hat{\mu}_i$ as fixed is appropriate when good information is available on the model parameters, but may be unrealistic for example for designed experiments in which the number of parameters is not small compared to the number of observations. In normal linear models, REML estimation of the variance structure is obtained from the marginal distribution of any set of zero mean contrasts, $Z^T\mathbf{y}$ say. In a similar way, independent and identically distributed residuals could be obtained by transforming from the y_i to any orthonormal set of zero mean constrasts. Extending this idea to non-normal regression is more difficult, but could in principle be done using the conditional approach of Smyth and Verbyla (1995). In that paper, Smyth and Verbyla argue that REML estimation for generalized linear models should proceed by considering the conditional distribution of the y_i given $\hat{\boldsymbol{\beta}}$. Independent quantile residuals could therefore be defined by considering the conditional distribution of each y_i given $y_1, \ldots, y_{i-1}$ and $\hat{\boldsymbol{\beta}}$. For certain values of i this distribution would be degenerate; these values could be ignored without loss of information.

References

Brillinger, D. R. and Preisler, H. K. (1983). Maximum likelihood estimation in a latent variable problem. In S. Karlin, T. Amemiya and L. A. Goodman, eds., *Studies in Econometrics, Time Series and Multivariate Statistics*, Academic, New York, pp. 31–65.

Brillinger, D. R. (1996). An analysis of an ordinal-valued time series. In P. M. Robinson and M. Rosenblatt, eds., *Papers in Time Series Analysis: A Memorial Volume to Edward J Hannan. Athens Conference Volume 2.* Lecture Notes in Statistics, Volume 115, Springer, New York. To appear.

Brown, M. B. (1974). Identification of the sources of significance in two-way contingency tables. *Appl. Statist.*, **23**, 405–413.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont, California.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc.*, B, **30**, 248–275.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant observation. *Biometrics*, **21**, 826–838.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). *Handbook of Small Data Sets.* Chapman & Hall, London.

Kotze, T. J. v W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Appl. Statist.*, **33**, 215–223.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models, 2nd ed.* Chapman and Hall: London.

Pearson, E. S. (1950). On questions raised by the combination of tests based on discontinuous distributions. *Biometrika*, **37**, 383–398.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. Reprinted in 1948 in *Karl Pearson's Early Statistical Papers,* Cambridge University Press, Cambridge, 443–475.

Pierce, D. A., and Schafer, D. W. (1986). Residuals in generalized linear models. *J. Amer. Statist. Ass.*, **81**, 977–986.

Smyth, G. K. and Verbyla, A. P. (1996). A conditional approach to REML in generalized linear models. *J. Roy. Statist. Soc.* B, **58**, 565-572.