Universal Residuals: A Multivariate Transformation *

A.E. Brockwell (abrock@stat.cmu.edu)
Dept. of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA

August 14, 2006

Abstract

Rosenblatt's transformation has been used extensively for evaluation of model goodness-of-fit, but it only applies to models whose joint distribution is continuous. In this paper we generalize the transformation so that it applies to arbitrary probability models. The transformation is simple, but has a wide range of possible applications, providing a tool for exploratory data analysis and formal goodness-of-fit testing for a very general class of probability models. The method is demonstrated with specific examples.

Keywords: residuals, hypercube, uniform, generalized linear models, time series, survival analysis

Rosenblatt (1952) described a transformation[1] mapping a k-variate random vector with a continuous distribution to one with a uniform distribution on the k-dimensional hypercube. The transformation is particularly important for generating residuals in nonlinear and/or non-Gaussian time series analysis (Smith, 1985; Shephard, 1994; Diebold et al., 1998; Kim et al., 1998), but can also be used to obtain residuals for more general probability models. This allows for formal goodness-of-fit testing of these models (see, among others, Darling,

*This work was supported in part by NIH Grants R21 EB005967-01A1 and R01 EB005847-01, as well as NSF Grant CCR-0326453.

[1]Rosenblatt pointed out in his 1952 paper that J.H. Curtiss and I.H. Savage had also considered the same transformation.

1957; Justel et al., 1997; Liang et al., 2001), and also provides a means of generating diagnostic plots useful for exploratory data analysis. In this paper, Rosenblatt's transformation is generalized so that it can be applied to arbitrary probability distributions, instead of just continuous distributions. This extends the scope of the aforementioned procedures to cover models with non-continuous distributions, such as generalized linear models, time series models with discrete observations, survival analysis models, and many others.

Following the notation of Rosenblatt (1952), let $X = (X_1, \ldots, X_k)$ be a random vector defined on a probability space $(\Omega, \mathcal{F}, P)$. Define the conditional cumulative distribution functions

$$\begin{aligned} F_1(x_1) &= P(X_1 \leq x_1), \\ F_2(x_2|x_1) &= P(X_2 \leq x_2|X_1 = x_1), \\ &\ldots \\ F_k(x_k|x_1, \ldots, x_{k-1}) &= P(X_k \leq x_k|X_1 = x_1, \ldots, X_{k-1} = x_{k-1}). \end{aligned}$$

We will also generically define $f(x-)$ to be the left limit $\lim_{u \uparrow x} f(u)$, so that $F_1(x_1-) = P(X_1 < x_1)$, $F_2(x_2 - |x_1) = P(X_2 < x_2|X_1 = x_1)$, and so on. Rosenblatt's transformation is given by $z = (z_1, \ldots, z_k) = T_1 x = T_1(x_1, \ldots, x_k)$, where

$$\begin{aligned} z_1 &= F_1(x_1), \\ z_2 &= F_2(x_2|x_1), \\ &\ldots \\ z_k &= F_k(x_k|x_1, \ldots, x_{k-1}). \end{aligned}$$

When the distribution of X is continuous, it is straightforward to show that $Z = T_1 X$ has a uniform distribution on the k-dimensional unit hypercube. However, when the distribution of X is discrete, or mixed discrete and continuous, this is not generally the case. This is easily verified by considering the trivial counterexample where $k = 1$ and X has a Bernoulli distribution with $P(X = 1) = p$.

We introduce a new transformation, T_2, that can be used when X has any k-variate distribution. The transformation is random, in the sense that it depends on auxiliary random variables $(U_1, \ldots, U_k)$, which are independent identically distributed (iid) uniform random variables on the interval $[0, 1]$, assumed to be independent not only of each other, but also of $(X_1, \ldots, X_k)$. The new transformation is specified by $z = (z_1, \ldots, z_k) = T_2(x_1, \ldots, x_k)$, where

$$z_1 = (1 - U_1)F_1(x_1-) + U_1 F_1(x_1),$$

$$
\begin{aligned}
z_2 &= (1-U_2)F_2(x_2 - |x_1) + U_2 F_2(x_2|x_1), \\
&\dots \\
z_k &= (1-U_k)F_k(x_k - |x_1, \dots, x_{k-1}) + U_k F_k(X_k|x_1, \dots, x_{k-1}). \qquad (1)
\end{aligned}
$$

The following result is proved at the end of this document.

Theorem 1. *Let X be a k-dimensional random vector X defined on $(\Omega, \mathcal{F}, P)$, with an arbitrary distribution. Then T_2X has a uniform distribution on the k-dimensional unit hypercube.*

In the special case where X has a continuous distribution, then T_2X is equal to T_1X.

Applications. The transformation T_2 can be used to test goodness-of-fit for any probability model. For many models, it is easily applied to the observations $X = x$. Under the null hypothesis H_0 that the data were generated by the specified model, the residual vector $Z = T_2X$ will be uniformly distributed on the unit (If desired, one can further apply the inverse cumulative distribution function of a standard normal distribution, and then under H_0, the resulting values will be realizations of independent and identically distributed standard normal random variables.) Kolmogorov-Smirnov or other tests (Darling, 1957; Justel et al., 1997; Liang et al., 2001) can then be used.

Example 1: Generalized Linear Models. Consider a typical generalized linear model (see, e.g. McCullagh and Nelder, 1989; Dobson, 2001) of the form $Y \sim \text{Poisson}(\exp(V\beta))$ for some $Y = (Y_1, \ldots, Y_k)$, a $k \times p$ design matrix $V = [V_{ij}]_{i=1,\ldots,k,\ j=1,\ldots,p}$ and a p-dimensional vector of coefficients β. Suppose that for some estimate $\hat{\beta}$, we wish to test the null hypothesis $H_0 : Y \sim \text{Poisson}[\exp(V\hat{\beta})]$. In this case, the conditional structure in T_2 vanishes since observed values are assumed to be conditionally independent given $X_{ij} = x_{ij}$, and the residuals for the fitted model are simply given by

$$Z_i = \sum_{n=0}^{y_i - 1} f(n, (v_{i1}, \ldots, v_{ik}) \cdot \hat{\beta}) + U_j f(y_i, (v_{i1}, \ldots, v_{ik}) \cdot \hat{\beta}), \tag{2}$$

where $f(n, \lambda) = \exp(-\lambda)\lambda^n / n!$, with the convention that the sum vanishes when $y_i = 0$. One could then define $W_i = \Phi^{-1}(Z_i)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution, and under H_0, W_i should be iid standard normal random variables. Interestingly, in this context, Anscombe residuals (Anscombe, 1961) were originally intended to provide residuals whose distributions were "as normal as possible" under H_0. The transformation T_2 goes a step further toward this goal, in the sense that residuals are exactly normally distributed under H_0.

To illustrate the behaviour of these residuals, consider a special case with $k = 5000$,

$$Y_i \sim \text{Poisson}(\exp(v_i)), \tag{3}$$

with $v_i = \cos(2\pi i/k) - 1$, for $i = 1, \ldots, k$. A simulated realization $\{y_1, \ldots, y_{5000}\}$ is shown in Figure 1(a). Residuals $\Phi^{-1}(z_i)$ obtained using (2), under the correct model (that is,

$\hat{\beta} = 1$) are shown in Figure 1(b). These clearly behave as expected, that is, they appear to be iid realizations of standard normal random variables. Anscombe residuals are shown in Figure 1(c), and residuals $\Phi^{-1}(z_i)$ obtained using (2) under the (incorrect) assumption that $v_i \sim \text{Poisson}(\exp(-1))$ are shown in Figure 1(d). The Anscombe residuals are somewhat difficult to interpret. They are based on use of the correct model, and although their serial correlation is close to zero, they exhibit apparent structure induced by $\{v_i,\ i = 1, \ldots, 10000\}$. Furthermore, their time-varying discrete nature makes it difficult to see how one might use them to construct a formal goodness-of-fit test. It is not even easy to determine visually whether or not the model is appropriate. The T_2-residuals based on the incorrect model, on the other hand, are highly interpretable. Low values indicate the the quantiles of the observations are lower than expected, and hence are an indication of over-fitting. Conversely, high values indicate under-fitting. From Figure 1(d), it is easily seen that the model is under-fitting the data for small and large i, and over-fitting the data in between.

Example 2: Time Series Analysis. T_2 also has natural applications in time series analysis. In this context, one typically builds a probability model specifying the joint distribution of random variables $(X_1, \ldots, X_{k+h})$. The random variables $X_1, \ldots, X_k$ are observed, and random variables $X_{k+1}, \ldots, X_{k+h}$ represent some future horizon of interest. Forecasting is carried out by determining the conditional distribution of these future values, given available observations, but clearly relies on the quality of the specified probability model. For the family of models where $(X_1, \ldots, X_k)$ has a continuous distribution, use of Rosenblatt's transformation T_1 to compute residuals is a simple matter of determining each one-step predictive distribution function $P(X_j \leq u | X_{j-1} = x_{j-1}, \ldots, X_1 = x_1)$ and evaluating it at the observed value $u = x_j$. Indeed this approach has been suggested, used and discussed in this context by a number of authors, including Smith (1985); Shephard (1994); Kim et al. (1998); Diebold et al. (1998). For many of these models, calculation of the one-step predictive distributions is a well-studied problem, partly because the likelihood is often computed using the factorization of the joint density

$$p(x_1, \ldots, x_j) = p(x_j | x_{j-1}, \ldots, x_1) p(x_{j-1} | x_{j-2}, \ldots, x_1) \ldots p(x_2 | x_1) p(x_1),$$

and also because recently-developed sequential Monte Carlo methods (see e.g. Gordon et al., 1993; Kitagawa, 1996; Doucet et al., 2001) yield good numerical approximations to these distributions. The transformation T_2 allows the use of the same goodness-of-fit diagnostics used by the aforementioned authors, but for any time series, including those whose marginal distributions are not continuous, for example, the saturated Gaussian models described in Brockwell and Chan (2006).

Example 3: Survival Analysis. Suppose that lifetimes T_i of objects are independent, with distribution function $G(t) = P(T_i \leq t)$. Observations in this case may be censored. Let t_i denote the age of object i at the time of observation, if the object is "alive", and the lifetime of the subject, if the subject is no longer alive. If the object is still alive, we can only infer that $T_i > t_i$, otherwise we know that $T_i = t_i$. Regardless of whether or not the object is alive at the time of observation, let a_i denote its age at that time. We also introduce indicator variables $\{W_i\}$, with W_i equal to one if object i had died by the time of measurement, and zero if the object was still alive.

Assume that we have n observations. To apply T_2 in this context, we can define $k = 2n$, and

$$X_1 = W_1, \ldots, X_n = W_n, X_{n+1} = T_1, \ldots, X_{2n} = T_n.$$

Then to evaluate the conditional distribution functions required for T_2, we have, for $j = 1, \ldots, n$,

$$\begin{aligned} & P(X_j \leq x | X_{j-1} = x_{j-1}, \ldots, X_1 = x_1) \\ = \;\; & P(W_j \leq x) = \begin{cases} 0, & x < 0 \\ 1 - G(a_j), & 0 \leq x < 1, \\ 1, & x \geq 1, \end{cases} \end{aligned} \tag{4}$$

and

$$\begin{aligned} & P(X_{j+n} \leq t | X_{j+n-1} = x_{j+n-1}, \ldots, X_1 = x_1) \\ = \;\; & P(T_j \leq t | W_j = x_j) = \begin{cases} I_{[a_j, \infty)}(t), & x_j = 0 \\ G(t)/G(a_j), & x_j = 1. \end{cases} \end{aligned} \tag{5}$$

The quantities $P(X_j < x | X_{j-1}, \ldots, X_1)$ are easily obtained as the left limits of the functions on the right-hand sides of (4) and (5) above. Note that we do not necessarily require that $G(\cdot)$ itself be the distribution function of a continuous random variable.

Discussion. The transformation T_2 provides a means of generating residuals for any probablity model. Several examples have been given, and in each of these, the computational issues are trivial. We have shown (in the generalized linear model example) that the approach, beyond providing methods of performing formal goodness-of-fit tests, can yield informative plots for purposes of exploratory data analysis. Of course, in other cases, the computations may not be so trivial, and in certain cases, the sequencing of the data may make these computations more or less tractable.

Since the transformation involves the use of additional random variables $U_1, \ldots, U_k$, a potentially interesting future line of work could address the problem of determining how

much information is contained in the residuals. Consider the extreme case, for instance, where X is a univariate constant. Then the residual obtained by T_2 is simply a uniformly distributed random variable on the interval $[0, 1]$. In this case, there is no information in the model itself, and there is also no useful information contained in the residual. For distributions with both a continuous and a discrete component, it's not immediately obvious whether or not one should consider all residuals to be equally important.

Acknowledgements. The author is also grateful to Peter Brockwell, Chris Genovese, and Larry Wasserman for helpful discussions related to the work, and to an anonymous reviewer for additional comments.

Proof of Theorem 1. Let $f(x-)$ denote the left limit $\lim_{u \uparrow x} f(u)$, and let $\nu(\cdot)$ denote the Lebesgue measure. We will write $Y \sim \text{Unif}\{A\}$ to indicate that the random variable Y has a uniform distribution over the set A, that is, $P(Y \in B) = \nu(B \cap A)/\nu(A)$. The core of the argument is encapsulated in the following result.

Lemma 2. *Let X be a random variable with cumulative distribution function F. Let U be a uniformly distributed random variable on the interval $[0, 1)$, independent of X. Then*

$$Z(X) = (1 - U)F(X-) + UF(X)$$

is also uniformly distributed on $[0, 1]$.

Proof. Lebesgue's decomposition allows us to write the distribution μ of X as $\mu = \alpha_1\mu_c + \alpha_2\mu_s + \alpha_3\mu_d$, where $\alpha_1 + \alpha_2 + \alpha_3 = 1$, and μ_c, μ_s and μ_d represent the continuous, singular, and discrete components of μ, respectively. Thus we can express X as a mixture random variable with two components,

$$X = \begin{cases} C, & \text{with probability } \alpha, \\ D, & \text{with probability } (1 - \alpha). \end{cases} \tag{6}$$

Here $\alpha = \alpha_1 + \alpha_2$, C is a random variable with distribution $(\mu_c + \mu_s)\alpha^{-1}$, and D is a random variable with distribution $\mu_d\alpha_3^{-1}$. Let $F_C(\cdot)$ and $F_D(\cdot)$ denote, respectively, the distribution functions of C and D, so that

$$F(x) = \alpha F_C(x) + (1 - \alpha)F_D(x). \tag{7}$$

Since C is continuous, we must have, for all x,

$$F_C(x) = F_C(x-). \tag{8}$$

The discrete random variable D can take only countably many different values with positive probability. Without loss of generality, let us denote the ordered sequence of these possible values by $\{d_j,\ j = 1, 2, \ldots\}$, with $P(D = d_j) = p_j > 0$, $j = 1, 2, \ldots$, and $\sum_j p_j = 1$.

Define the sets

$$\begin{aligned} H_D &= \bigcup_j [F(d_j-), F(d_j)), \\ H_C &= [1, 0] \setminus H_D. \end{aligned}$$

Then making use of (7) along with (8), we have, for each j,

$$\begin{aligned} Z(d_j) &= [F(d_j) - F(d_j-)]U + F(d_j-) \\ &= (1-\alpha)[(F_D(d_j) - F_D(d_j-))U + F_D(d_j-)] + \alpha F_C(d_j). \end{aligned}$$

In other words,

$$\begin{aligned} Z(d_j) &\sim (1-\alpha)\text{Unif}\{\ [F_D(d_j-), F_D(d_j))\ \} + \alpha F_C(d_j) \\ &\sim \text{Unif}\{\ [F(d_j-), F(d_j))\ \}. \end{aligned} \tag{9}$$

Then since $P(D = d_j) = p_j = F_D(d_j) - F_D(d_j-) = F(d_j) - F(d_j-)$, it follows directly that

$$Z(D) \sim \text{Unif}\{H_D\}. \tag{10}$$

Next consider an infinitesimal interval $dz \subset H_C$. Let $I_j = (d_{j-1}, d_j]$, with the convention that $d_0 = -\infty$. Thus dz lies in exactly one of the intervals I_j, say in interval I_{j^*}, and we have

$$\begin{aligned} P(Z(C) \in dz) &= \sum_m P(Z(C) \in dz | C \in I_m) P(C \in I_m) \\ &= \nu(dz) / \{\alpha[F_C(d_{j^*}) - F_C(d_{j^*-1})]\} \times [F_C(d_{j^*}) - F_C(d_{j^*-1})] \\ &= \nu(dz)/\alpha. \end{aligned}$$

It follows directly that

$$Z(C) \sim \text{Unif}\{H_C\}. \tag{11}$$

Finally, combining (6), (10), and (11), and using the property that $\nu(H_C) = \alpha$ and $\nu(H_D) = 1 - \alpha$, we see that

$$Z(X) \sim \text{Unif}\{[0, 1]\}. \tag{12}$$

This completes the proof of Lemma 2. □

Now we are in a position to prove the main result. Under the conditions of Theorem 1, defining $Z = (Z_1, \ldots, Z_k) = T_2(X_1, \ldots, X_k)$, it follows directly from Lemma 2 that $Z_1 \sim$ Uniform($[0, 1]$). Next observe that for $j = 2, \ldots, k$, for any Borel subset A of the unit interval $[0, 1]$, by conditioning and making use of Lemma 2 again,

$$\begin{aligned}
&P(Z_j \in A | Z_{j-1} = z_{j-1}, \ldots, Z_1 = z_1) \\
&= \int_{x_{j-1}} \ldots \int_{x_1} P(Z_j \in A | X_{j-1} = x_{j-1}, \ldots, X_1 = x_1) \\
&\quad \cdot P(X_1 \in dx_1, X_2 \in dx_2, \ldots, X_{j-1} \in dx_{j-1} | Z_{j-1} = z_{j-1}, \ldots, Z_1 = z_1) \\
&= \nu(A) \int_{x_{j-1}} \ldots \int_{x_1} P(X_1 \in dx_1, X_2 \in dx_2, \ldots, X_{j-1} \in dx_{j-1} | Z_{j-1} = z_{j-1}, \ldots, Z_1 = z_1) \\
&= \nu(A).
\end{aligned}$$

Thus the random variables $\{Z_1, \ldots, Z_k\}$ are independent of each other, and each has a uniform distribution on $[0, 1]$. That is, $(Z_1, \ldots, Z_k)$ has a uniform distribution on the unit hypercube.

References

F.J. Anscombe. Examination of residuals. In *Proc. Fourth Berkeley Symposium*, pages 1–36, 1961.

A.E. Brockwell and N.H. Chan. Long memory dynamic tobit models. *J. Forecasting*, 2006. To appear.

D. A. Darling. The Kolmogorov-Smirnov, Cramer-von Mises tests. *Annals of Mathematical Statistics*, 28(4):823–838, 1957.

F.X. Diebold, T.A. Gunther, and T. S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883, 1998.

A.J. Dobson. *Introduction to Generalized Linear Models*. Chapman and Hall/CRC, second edition edition, 2001.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F*, 140:107–113, 1993.

A. Justel, D. Pena, and R. Zamar. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters*, 35(3):251–259, 1997.

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65:361–393, 1998.

G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

Jia-Juan Liang, Kai-Tai Fang, Fred J. Hickernell, and Runze Li. Testing multivariate uniformity and its applications. *Math. Comp.*, 70:337–355, 2001.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, second edition, 1989.

M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23:470–472, 1952.

N. Shephard. Partial non-Gaussian state space. *Biometrika*, 81:115–131, 1994.

J.Q. Smith. Diagnostic checks of non-standard time series models. *Journal of Forecasting*, 4:283–291, 1985.

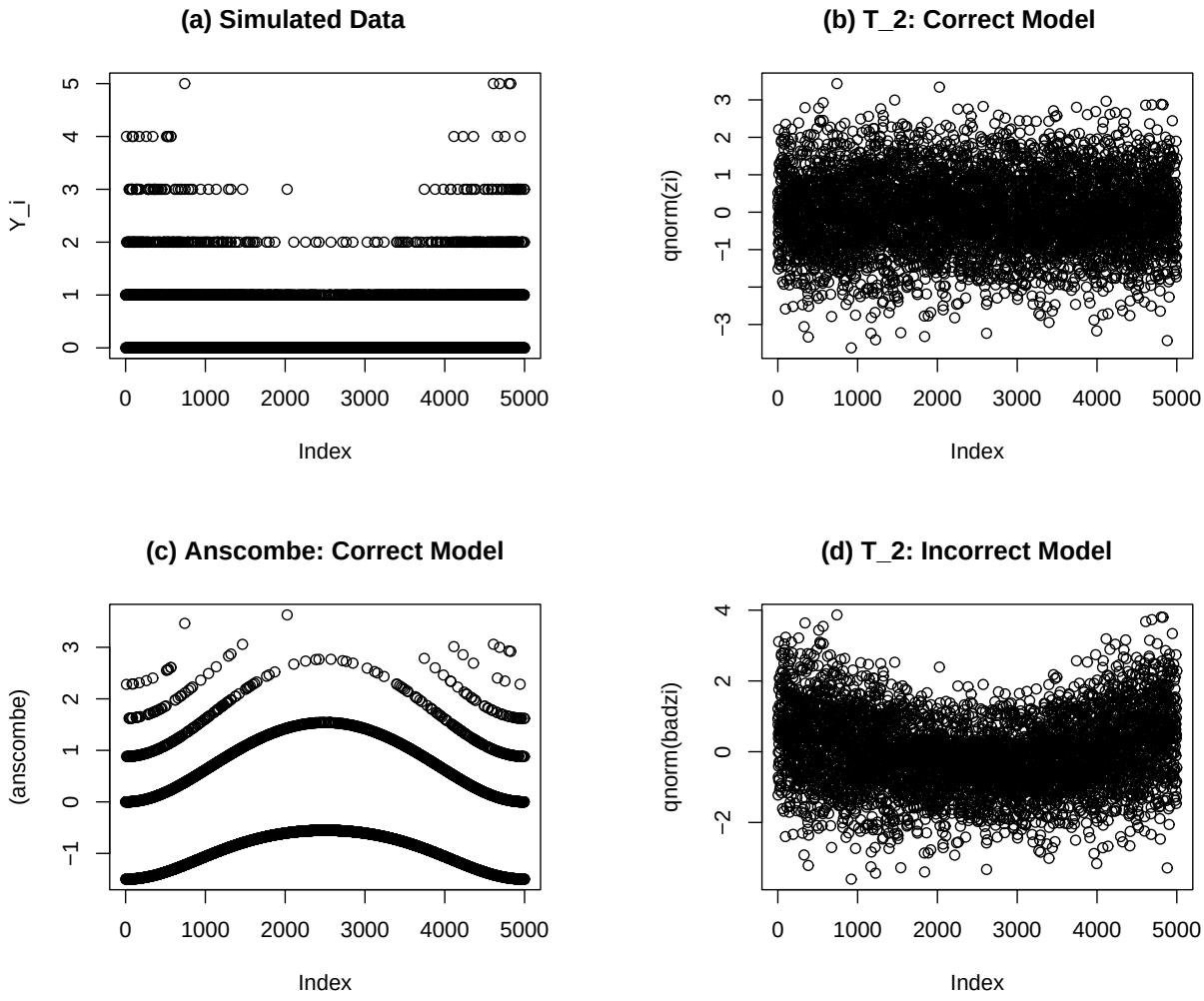

Figure 1: Test data generated from the generalized linear model (3), along with different types of residuals. **Top left:** simulated observations $\{y_i,\ i = 1, \ldots, 5000\}$. **Top right:** residuals, under correct model specification, obtained using T_2 and applying the inverse cumulative distribution Φ^{-1} of a standard normal. **Bottom left:** Anscombe residuals, under the correct model specification. **Bottom right:** residuals obtained using T_2 and applying Φ^{-1}, under a constant rate (incorrect) model specification.