

Response Transformations for Models with Additive Errors

David J. Olive*

Southern Illinois University

October 6, 2010

Abstract

The applicability of a regression model of the form $Y = m(\mathbf{x}) + e$ can be expanded by allowing a response transformation of the form $Y = t_{\lambda_0}(Z) = m(\mathbf{x}) + e$. A graphical method for selecting the transformation is given. Note that models with additive errors include linear models such as multiple linear regression and many experimental design models, nonlinear and nonparametric regression, generalized additive models and single index models.

KEY WORDS: Generalized Additive Models, Linear Models, Nonlinear Regression, Nonparametric Regression, Single Index Models.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu.

1 INTRODUCTION

Regression is the study of the conditional distribution $Y|\mathbf{x}$ of the scalar response Y given the $p \times 1$ vector of predictors \mathbf{x} . An important regression model is

$$Y_i = m(\mathbf{x}_i) + e_i \tag{1}$$

for $i = 1, \dots, n$ where m is a function of \mathbf{x}_i and the errors e_i are iid. Many of the most important regression models have this form, including the multiple linear regression model, experimental design models, nonlinear regression, nonparametric regression and many time series and semiparametric models. If \hat{m} is an estimator of m , then the i th residual is $r_i = Y_i - \hat{m}(\mathbf{x}_i) = Y_i - \hat{Y}_i$.

The single index model with additive error has the form $Y = g(\alpha + \mathbf{x}^T \boldsymbol{\beta}) + e = g(SP) + e$ where the *sufficient predictor* $SP = \alpha + \mathbf{x}^T \boldsymbol{\beta}$. The linear model is a special case with $Y = \alpha + \mathbf{x}^T \boldsymbol{\beta} + e = SP + e$. The generalized additive model (GAM) analog of the linear model is $Y = AP + e$ where the *additive predictor* $AP = \alpha + \sum_{j=1}^p S_j(x_j)$ for some functions S_j . See Hastie and Tibshirani (1990), Wood (2006) and Zuur, Ieno, Walker, Saveliev and Smith (2009). Note that the linear model is a special case of the GAM where $S_j(x_j) = x_j \beta_j$. A nonlinear regression has $m(\mathbf{x}) = g_{\boldsymbol{\theta}}(\mathbf{x})$, a known function except for k unknown parameters $(\theta_1, \dots, \theta_k)^T = \boldsymbol{\theta}$.

Response plots are like residual plots but replace the residuals r_i by the response variable Y_i on the vertical axis. For single index models, the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}$ while the estimated additive predictor $EAP = \hat{\alpha} + \sum_{j=1}^p \hat{S}_j(x_j)$ is used for the GAM. For single index models and the GAM, the response plot is the plot of ESP or EAP versus Y , and is used to visualize the model in the background of the data

since regression is the study of $Y|SP$ or $Y|AP$. This type of response plot is also called an estimated sufficient summary plot, and often the estimated conditional model mean function and a scatterplot smoother are added as visual aids. See Brillinger (1983), Cook (1998, p. 10). Cook and Weisberg (1997, 1999: ch. 18), and Olive and Hawkins (2005). A second type of response plot is a plot of $\hat{Y}_i = \hat{m}(\mathbf{x}_i)$ versus Y_i . For the linear model and the GAM, the two types of response plot coincide, while a residual plot of the ESP or EAP versus the residuals is used to visualize $e|SP$ or $e|AP$.

Suppose the zero mean constant variance errors e_1, \dots, e_n are iid from a unimodal distribution that is not highly skewed. For the linear model and the GAM analog of a linear model, the estimated mean function is the identity line with unit slope and zero intercept. Suppose the ESP or EAP take on many values. For both models and large sample size n , the plotted points should scatter about the identity line and the residual $= 0$ line in an evenly populated band for the response and residual plots, with no other pattern. For linear models, the two plots often look good if $n > 5p$.

For alternative models, often need much larger n . Also even if the model is a good approximation to the data, for moderate n there are often cases with rather large or rather small \hat{Y} that tend to be influential with small absolute residuals. Then the plotted points do not scatter about the reference line in an evenly populated band, and for moderate n the assumption of constant error variance for the model can be difficult to check with response or residual plots.

For some experimental design models, including the one way anova model, the ESP does not take on many values. Consider the one way fixed effects anova model. The *response plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$

versus r_{ij} .

For the one way anova model, the points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of Z_1, \dots, Z_m consists of an axis and m points each corresponding to the value of Z_i . The response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of Y_{i1}, \dots, Y_{i,n_i} . The p dot plots should have roughly the same amount of spread, and each $\hat{\mu}_i$ corresponds to level a_i . Similarly, the residual plot consists of p dot plots, and the dot plot corresponding to $\hat{\mu}_i$ is the dot plot of r_{i1}, \dots, r_{i,n_i} .

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i , each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot.

Section 1 extends the Olive (2004) graphical method for linear model response transformations to response transformations for regression models with additive errors, and section 2 gives examples.

2 A Graphical Method for Response Transformations

The applicability of the regression model (1) can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional

unknown transformation parameter λ_o , such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = m(\mathbf{x}_i) + e_i. \quad (2)$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow model (1). The function m depends on λ_o , the p predictors x_j are assumed to be measured with negligible error, and the zero mean constant variance errors e_i are assumed to be iid from a unimodal distribution that is not highly skewed.

Next, two important response transformation models are given. Assume that *all* of the values of the “response” Z_i are *positive*. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

The *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \quad (3)$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by Z_i for $\lambda = 1$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L . This family is a special case of the response transformations considered by Tukey (1957).

A graphical method for response transformations computes the “fitted values” \hat{W}_i using $W_i = t_\lambda(Z_i)$ as the “response.” Then a *transformation plot* of \hat{W}_i versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$. If the plotted points follow the identity line for λ^* , then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. After selecting the transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is a response plot, and a residual plot should also be

made. This technique is very simple and Olive (2004) suggested the method for linear models.

Each transformation plot is a “response plot” for the seven values of $W = t_\lambda(Z)$, and the method chooses the “best response plot” where the model (1) seems “most reasonable.” If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding “residual plots” of \hat{W} versus $W - \hat{W}$ look reasonable. According to Mosteller and Tukey (1977, p. 91), the values of λ in decreasing order of importance are 1, 0, 1/2, -1 and 1/3. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good. Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ_L . For linear models, Box and Cox (1964) is widely used.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L , then sometimes $\hat{\lambda}_n$ will converge (e.g. in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L . Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3 . Powers from numerical methods can also be added.

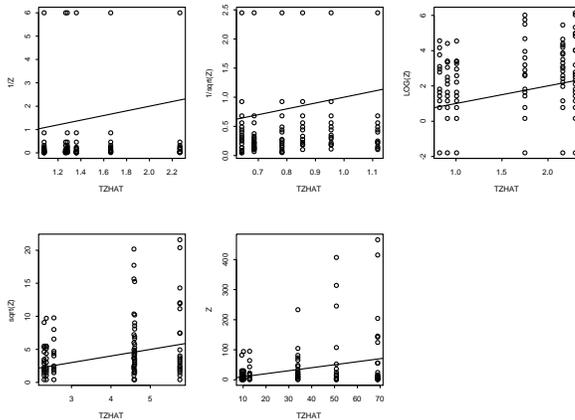


Figure 1: Transformation Plots for Crab Data

3 Examples

In the following examples, the plots show $t_\lambda(Z)$ on the vertical axis. The label “TZHAT” of the horizontal axis is for the fitted values that result from using $t_\lambda(Z)$ as the “response” in the software.

Example 1. For experimental design models, often use five transformations instead of seven: $\Lambda = \{-1, -1/2, 0, 1/2, 1\}$. Kuehl (1994, p. 128) gives data for counts of hermit crabs in six different coastline habitats, where C is the count of crabs and the “response” $Z = C + 1/6$. Each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. The one way anova model $W_{ij} = t_\lambda(Z_{ij}) = \mu_i + e_{ij} = \eta + \tau_i + e_{ij}$ was fit for $i = 1, \dots, 6$ with $n_i = 25$, and $j = 1, \dots, n_i$. Each of the six habitats was a level with 25 replicates. Figure 1 shows the five transformation plots. The transformation $Y = \log(Z)$ is used since the six dot plots have roughly the same shape and spread. The transformations $1/Z$ and $1/\sqrt{Z}$ do not handle the 0 counts well, while the transformations \sqrt{Z} and Z have variance that increases with the mean.

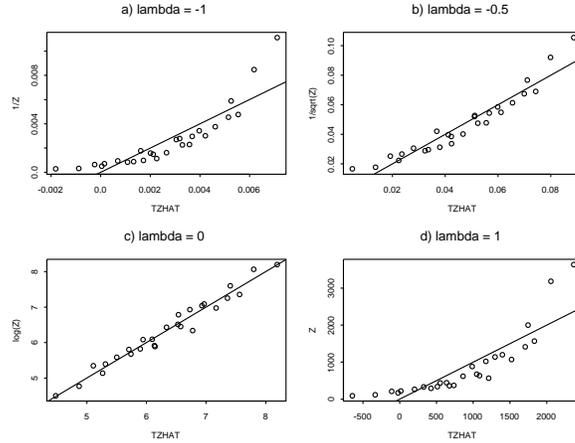


Figure 2: Transformation Plots for Textile Data

Example 2. Box and Cox (1964) analyze data from a 3^3 experiment on the behavior of yarn under cycles of repeated loadings. Here Z = number of cycles until failure while the three predictors are the length, amplitude and load. A constant and the three main effects were used. For this data set, there is one value of the response for each of the 27 treatment level combinations. Figure 2 shows four of the transformation plots. The plotted points curve away from the identity line in three of the four plots. The plotted points for the log transformation follow the identity line with roughly constant variance.

This transformation plot is the response plot where $Y = \log(Z)$. To visualize the conditional distribution of $Y|\mathbf{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose that $\log(\text{cycles to failure})$ given $\text{fit} = 6$ is of interest. Mentally examine the plot about a narrow vertical strip about $\hat{Y} = 6$, perhaps from 5.75 to 6.25. The cases in the narrow strip have a mean close to 6 since they fall close to the identity line. Similarly, when $\hat{Y} = \hat{y}$ for \hat{y} between 4.5 and 8.5, the cases have $\log(\text{cycles to failure})$ near \hat{y} , on average. Cases 19 and 20 had the largest Y values with long length, short

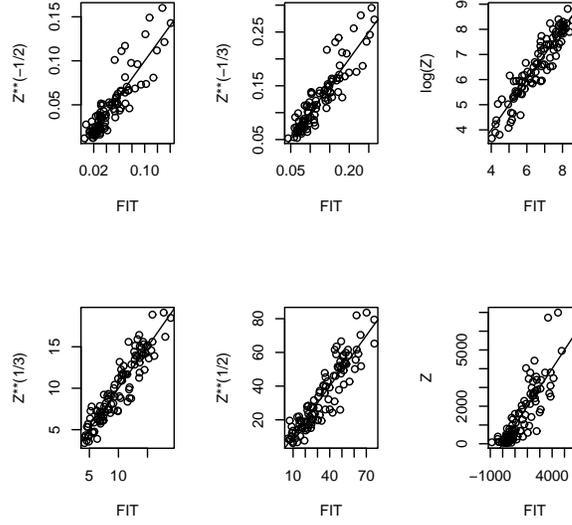


Figure 3: Transformation Plots for Lynx Data

amplitude of loading cycle and low load. Cases 8 and 9 had the smallest Y values with low length, high amplitude and high load.

For experimental design models, interest is often in finding the combination of predictors that result in the largest or smallest values of the response. This example illustrates that the response plot is useful for finding combinations of levels with desirable values of the response.

Example 3. The Moran (1953) lynx data is a well known time series of $n = 114$ cases concerning the number Z_t of lynx trapped in a section of Northwest Canada from 1821 to 1934. Autoregressive time series was used, and several of the transformation plots in Figure 3 look linear. The residual plots in Figure 4 suggest that the log, square root and cube root transformations are adequate.

Although the AR(2) model with $\log(\text{lynx})$ suggested by Moran (1953) has been heavily

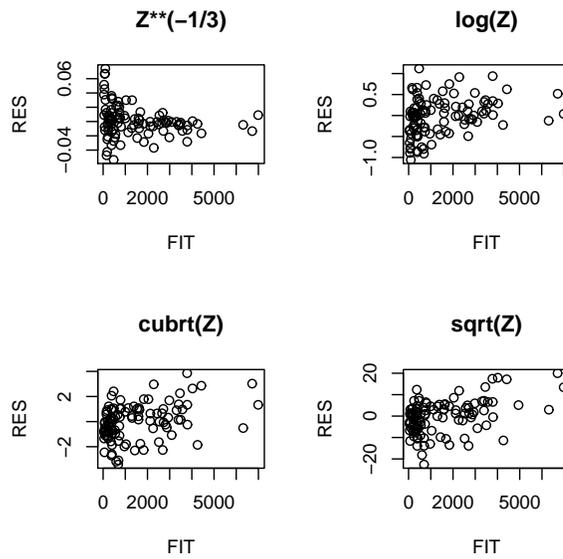


Figure 4: Residual Plots for Lynx Data

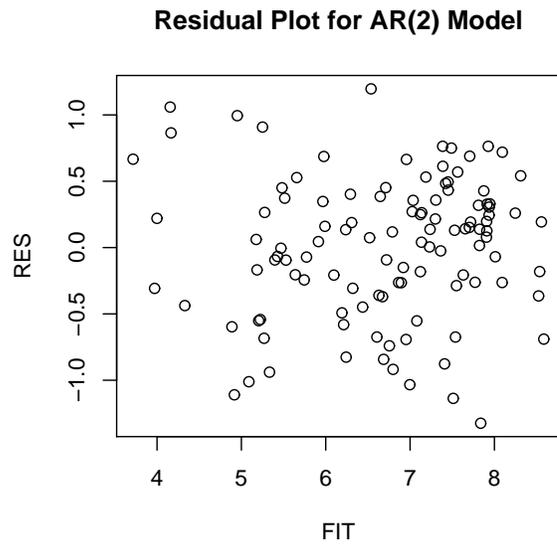


Figure 5: The AR(2) Model May Be Reasonable

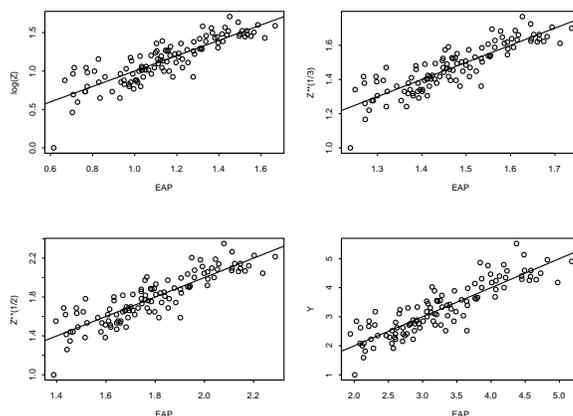


Figure 6: Transformation Plots for Ozone Data

criticized in the literature, the AR(2) model gives a good response plot and a better residual plot than those shown in Figure 4. See Figure 5.

Example 4. Chambers and Hastie (1993, pp. 251, 516) examine an environmental study that measured the four variables Z = ozone concentration, solar radiation, temperature, and wind speed for 111 consecutive days. Generalized additive models are fit using Z and $Z^{1/3}$ as the response. Figure 6 shows the four best transformation plots. The residual plots in Figure 7 suggest that no transformation, $Y = Z$ may be best since the other transformations do not fit the case in the lower left corner poorly.

4 Conclusions

To show that a transformation $Y = t(Z) = m(\mathbf{x}) + e$ is good, make a response plot of $\hat{Y} = \hat{m}(\mathbf{x})$ versus Y and the residual plot of $r = Y - \hat{Y}$ versus \hat{Y} . Also display at least three additional transformation plots. If more than one transformation plot is linear, display the corresponding “residual plots.”

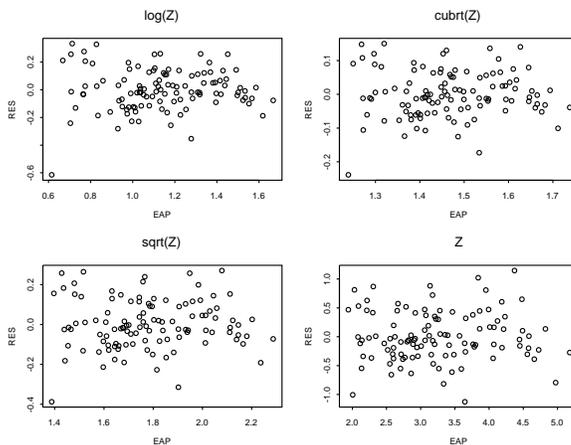


Figure 7: Residual Plots for Ozone Data

Cook and Olive (2001) used a similar graphical method for linear models where the “transformation plot” of \hat{Z}_i versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$. Cook and Weisberg (2004) give a graphical method for multiple linear regression, noting that an *inverse response plot* of Z versus \hat{Z} can often be used to visualize t_{λ_0} . Then the transformation plot of \hat{Z} versus Z can be used to visualize $t_{\lambda_0}^{-1}$. An advantage of this procedure is that the family of transformations need not be picked in advance, but the predictors need to be well behaved, and it may be difficult to generalize this method beyond the multiple linear regression model.

There is a massive literature on response transformations for the multiple linear regression model, but there is much less work for alternative models. The Box and Cox (1964) numerical procedure also works for many experimental design models. Hastie and Tibshirani (1990, ch. 7) give some numerical methods for fitting response transformations for the GAM, including the Breiman and Friedman (1985) ACE algorithm. Also see references in Carroll and Ruppert (1988), Castillo, Hadi, Lacruz and Pruneda (2008)

and Ruppert, Wand and Carroll (2003, section 2.9).

Once $Y = t(Z)$ and m are chosen by a numerical method, the response plot of $\hat{Y} = \hat{m}(\mathbf{x})$ versus Y is useful for checking whether the numerical method gave a reasonable model.

The graphical method is also very useful for outlier detection for linear models. Using the graphical method with the GAM is attractive for checking the linear model. Suppose the method gives $Y = t(Z) = \alpha + \sum_{j=1}^p S_j(x_j)$. If the response plot is linear and each plot of \hat{S}_j is linear, then use the simpler linear model $Y = \alpha + \mathbf{x}^T \boldsymbol{\beta}$. If the functional form of the \hat{S}_j is quadratic, then replace $S_j(x_j)$ by $\beta_{1j}x_j + \beta_{2j}x_j^2$.

5 References

- Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B*, 26, 211-246.
- Breiman, L., and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," (with discussion), *Journal of the American Statistical Association*, 80, 580-619.
- Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann*, eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.
- Carroll, R.J., and Ruppert, D. (1988), *Transformations and Weighting in Regression*, Wiley, New York, NY.
- Castillo, E., Hadi, A.S., Lacruz, B., and Pruneda, R.E. (2008), "Semi-parametric Non-

- linear Regression and Transformation Using Functional Networks,” *Computational Statistics & Data Analysis*, 52, 2129-2157.
- Chambers, J.M., and Hastie, T.J. (eds.) (1993), *Statistical Models in S*, Chapman & Hall, New York, NY.
- Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regression Through Graphics*, Wiley, New York, NY.
- Cook, R.D., and Olive, D.J. (2001), “A Note on Visualizing Response Transformations in Regression,” *Technometrics*, 43, 443-449.
- Cook, R.D., and Weisberg, S. (1994), “Transforming a Response Variable for Linearity,” *Biometrika*, 81, 731-737.
- Cook, R.D., Weisberg, S., 1997. Graphics for assessing the adequacy of regression models. *J. Amer. Statist. Assoc.* 92, 490-499.
- Cook, R.D., Weisberg, S., 1999. *Applied Regression Including Computing and Graphics*. Wiley, New York.
- Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall, London, UK.
- Kuehl, R. O. (1994), *Statistical Principles of Research Design and Analysis*, Belmont, CA: Duxbury Press.
- Moran, P.A.P (1953), “The Statistical Analysis of the Sunspot and Lynx Cycles,” *Journal of Animal Ecology*, 18, 115-116.
- Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, MA.
- Olive, D.J. (2004), “Visualizing 1D Regression,” in *Theory and Applications of Recent*

- Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst S., Series: Statistics for Industry and Technology, Birkhauser, Basel.
- Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.
- Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge, UK.
- Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics*, 28, 602-632.
- Wood, S.N. (2006), *Generalized Additive Models: an Introduction with R*, Chapman & Hall/CRC, Boca Rotan, FL.
- Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer-Science, New York, NY.