Using the Gamma Generalized Linear Model for Modeling Continuous, Skewed and Heteroscedastic Outcomes in Psychology

Victoria Ng & Robert A. Cribbie

Quantitative Methods Program

Department of Psychology

York University

Please send correspondence regarding this article to: Rob Cribbie, cribbie@yorku.ca.

Abstract

Some researchers in psychology have ordinarily relied on traditional linear models when assessing the relationship between predictor(s) and a continuous outcome, even when the assumptions of the traditional model (e.g., normality, homoscedasticity) are not satisfied. Of those who abandon the traditional linear model, some opt for robust versions of the ANOVA and regression statistics that usually focus on relationships for the typical or average case instead of trying to model relationships for the full range of relevant cases. Generalized linear models, on the other hand, model the relationships among variables using all available and relevant data and can be appropriate under certain conditions of non-normality and heteroscedasticity. In this paper, we summarize the advantages and limitations of using generalized linear models with continuous outcomes and provide two simplified examples that highlight the methodology involved in selecting, comparing, and interpreting models for positively skewed outcomes and certain heteroscedastic relationships.

Keywords: generalized linear model, robust statistics, heteroscedasticity, skewness, continuous outcomes

Modeling Continuous Nonnormal Clinical Outcomes:

The Generalized Linear Model Approach

Analysis of variance (ANOVA) and regression are commonly adopted methods for assessing the relationships between a continuous outcome and a categorical or continuous predictor, respectively. Both ANOVA and regression fall under the umbrella of the general linear model (GLM). An important limitation of the GLM is that statistical inference is only valid when errors are normally distributed and when variability is constant across the levels of the predictors (Fox, 2008). However, it is well known that skewed and heteroscedastic data are common in clinical psychology (Keselman et al., 1998). For example, non-negative data in psychology may be distributed with positive skewness and heteroscedasticity (e.g., Blanca et al., 2013; Grissim, 2000; Lindeberg et al., 2008), and variability that is proportional to the level of the outcome has been reported in clinical research, such as with depression scores (Mall et al., 2014; Sayer et al., 1993). Analyzing such data using methods that require the classical assumptions of normality and homoscedasticity would yield the consequence that estimates are less valid for statistical inference.

For analyzing such skewed outcomes that are potentially heteroscedastic, two popular approaches include use of robust statistics or of the GLM following transformations of the outcome variable. However, robust methods are usually only practical for simple designs (e.g., few covariates) and often ‘discard’ data that are potentially informative of the construct. For example, application of trimmed mean estimators to general population depression data (which are often positively skewed because of the presence of a few extreme cases) remove the extreme scores and thus focus on prediction based on the bulk of the data. However, one may consider whether the goals of the study are consistent with an approach that removes the less common, though potentially relevant, cases. Transformations of the outcome can sometimes restore normality and homoscedasticity, though care is required in interpreting the coefficients. For example, in using log transformations, a naive back-transformation of estimated coefficients would lead to interpretive issues for the response in original scale (Manning, 1998). If one prioritized generalizability, then the trimmed mean inference is unsuitable. If one desired inference in original scale because the inference in transformed scale lacks interpretability, then transformations are also unsuitable.

An alternative method that allows for inference to the population mean is the generalized linear model (GzLM, Nelder & Wedderburn, 1972), an extension of the GLM. All hypotheses that can be tested under the GLM can be tested under the GzLM. The GLM requires that errors approximate normality and homoscedasticity, but the GzLM does not. The GzLM also requires statistical assumptions, but the variety of statistical assumptions that data may conform to allows for considerable flexibility.

Therefore, we seek to recommend that researchers more frequently consider the GzLM as an alternative for analyzing continuous outcomes when the assumptions of GLMs (ANOVAs, regressions) are violated. Since the popularization of the GzLM in 1972, psychology has lagged somewhat in the frequency of its use. Although it is not uncommon for psychologists to correctly use the GzLM for categorical responses (i.e., logistic and Poisson regressions), it is uncommon to see the GzLM considered for continuous outcomes. In this paper, we briefly review the alternative approaches of robust statistics and transformations, present a non-technical introduction to the GzLM approach, and provide two simple, applied examples that utilize the GzLM on continuous outcomes. Because we consider the GzLM as an alternative to even the most basic designs like the ANOVA, the applied examples highlight considerations for model comparison, selection, and diagnostics in modeling skewed continuous outcomes against single predictors, when the single predictor is either categorical or continuous.

Popular Approaches for Dealing with Assumption Violations

Robust Statistics. Although there are numerous procedures that fall under the category of robust alternatives for regression and ANOVA, we introduce only a few that have been shown to have adequate statistical properties while preserving and/or improving power. Cribbie, Fiksenbaum, Wilcox and Keselman (2012), Grissim (2000), Rosopa, Schaffer and Schroeder (2013), Wilcox and Keselman (2011), among others, discuss regression and ANOVA alternatives for modeling heteroscedastic relationships, including weighted least squares, the Theil-Sen estimator, and the Welch (1938) adjusted degrees of freedom procedure. Although weighted least squares and the original Welch procedure are not robust to violations of normality (Cribbie et al., 2012, Sohn & Kim, 1997), the Theil-Sen estimator, i.e., the median of the slopes for all possible sample (x,y) points, and the Welch test with trimmed means and Winsorized variances or ranked data, provide good alternatives for regression and ANOVA, respectively, when both homoscedasticity and normality assumptions are violated (Cribbie et al., 2012). In substantive applications, such methods have since gained favour; skewed distributions for Beck Depression Inventory scores have been examined using the modified Theil-Sen estimator (Clewett, Bachman, & Mather, 2014) for reducing sensitivity to outliers, and other analyses in psychobiology, cannabis use, and depression have used the trimmed mean estimator to recover statistical power when classical assumptions were violated (Barnwell, Earleywine, & Wilcox, 2006; Denson & Earleywine, 2006; Wilcox, Granger, Szanton, & Clark, 2014).

Although the Welch test with trimmed means and Winsorized variances and ranked data procedures retain necessary statistical properties and has been extended beyond ANOVA-type designs (e.g., Welsh, 1987), there are certain limitations. Parameters from the Theil-Sen procedure, for example, are difficult to interpret in multiple predictor models (e.g., Wang, Dang, Peng, & Zhang, 2009); as a median-based procedure, it incorporates only the middle points into the analysis. Similarly, the oft recommended 20% symmetrical trimmed mean (Erceg-Hurn & Mirosevich, 2008; Keselman, Wilcox, Othman, & Fradette, 2002), though representative of the majority of data values, yields confidence intervals that generalize to only 60% of the sampled population (Bonett & Price, 2002). This loss of validity can be considered critical when the variables of interest are indeed skewed at the population level, such as depression, risk behaviours, and reaction times. It seems sensible to suggest that the persistent presence of extreme scores in such samples are theoretically and meaningful important for the response of interest; for analysis then, it would be ideal to include all potentially relevant data by making inferences to the population of interest.

GLM following transformations of the outcome. Typically introduced as a technique to normalize skewness, transformations can also have the effect of stabilizing variance, such that the relationship between the outcome and set of predictors is linearized. Once data are transformed, the researcher would proceed with the general linear model. When the variance changes with the mean by some power relationship (e.g., changes in the variance are proportional to changes in the mean squared) for a strictly positive variable, for example, the Box-Cox transformation class of power transformations may be implemented (Osborne, 2010). Of these transformations, the log and square root transformations are well known and commonly conducted, such as with alcohol use (Neal & Simons, 2007), depression and automatic thoughts (Cui, Shi, & Oei, 2013; Pirbaglou et al., 2013), and reaction times (Klein Entink, van der Linden, & Fox, 2009).

However, several issues arise with transformations, including interpretation of scale and the requirement for the simultaneous correction of both skew and heteroscedasticity. With transformations, inferences are also made in transformed scale (e.g., a unit increase in x predicts a 0.52 *log unit* change in y), but their interpretations might not be particularly meaningful. If one desires inference to the mean, then back-transformed estimates would require an adjustment. Though adjustments have been calculated in the health econometric literature (Duan, 1983; Manning, 1998), such adjustments have never gained prominence in psychology. More importantly, even though some transformation may indeed provide a full correction, the full and simultaneous correction for both skew and heteroscedasticity is not always guaranteed. Given the premise that one is interested in modeling, ideally, an outcome's distribution without exclusion of less typical responses, the concept of transforming each value is still appealing. We now discuss the GzLM structure and discuss how it adopts the concept of transformation.

Generalized Linear Modeling

Model Structure and Specification

As an extension of the GLM, the GzLM relaxes the assumption that the residuals be normally distributed and homoscedastic. Distributions that could be specified for continuous responses include the Gaussian, gamma, and inverse-Gaussian (Nelder & Wedderburn, 1972), all from the exponential family of probability distributions. Specification of any of these exponential distributions implies that assumptions are made about the error distribution and the relation between the mean and variance (Nelder & Wedderburn, 1972). The mean-variance relation is described with the variance function (see Table 1). For example, a Gaussian distribution, implies that the variance is unrelated to the mean (i.e., the assumption of homoscedasticity); with the gamma distribution, which models positively skewed, non-negative data, the variances are expected to be proportional to the square of the means. In this latter case, the gamma distribution implies a certain form of heteroscedasticity but not homoscedasticity.

The GzLM is further specified by the link function, given as

$$\eta = g(\mu),$$

where μ denotes expected values, η denotes values in transformed scale, and $g(\mu)$ represents that transformation of expected values. When the link function is correctly specified, the relation between the outcome and the set of predictors is linearized. Further, the GzLM allows for this transformation to be specified separately from the outcome's distribution (Fox, 2008), highlighting the flexibility of exploring various link functions for the same distribution. Table 1 shows the link functions most commonly used with each distribution, the ‘canonical links’. Although these are the most commonly used, one can easily alter the choice of the link function. For example, with a specification for the gamma family, one might explore the efficacy of either a log or an inverse transformation for obtaining linearity. Because η are values given some function on μ, the inverse function can also be obtained; that is, one may invert $g(\mu)$ to obtain $g(\mu)^{-1}$, the mean function. This is valuable for directly obtaining the expected value in original scale. Thus, the transformation of the expected value, and not of the outcome itself, provides the additional advantage of interpreting the model's estimates in the outcome's original scale (Blough, Madden, & Hornbrook, 1999).

Thus, the GzLM has three primary components: 1) a structural component, specifying the linear combination of predictors (e.g., $\beta_0 + \beta_1X$, where β_0 represents the population intercept [Y|X=0], β_1 represents the change in η, which is a function of the expected value of Y, for a unit change in X, and X represents the predictor of interest); 2) a random component, specified by some distribution family, which assumes some mean-variance relationship in the error distribution, conditional on the predictors in the model; and 3) a link function that transforms the expected value of the response, such that the mean value of Y is related to the structural component (i.e., the linear combination of predictors). As well, the link function has the corresponding inverse link function.

Taken together, model specification involves specification for the family distribution and for the link function. The choice of the family distribution is based on properties of the chosen probability distribution (e.g., the gamma distribution might be chosen for data that are positive and that exhibit the implied mean-variance relation). The choice of the link function may be more nuanced, involving both statistical and substantive considerations. Statistically, different link functions can affect the performance of models, as indicated by differences in model fit and residual deviance (detailed later). Substantively, fitted values should be meaningful --- they should remain within reasonable bounds for the response and ideally allows for interpretation. For example, the inverse link can be very useful for reaction time data because the inverse transformation on reaction time may be interpreted as reaction speed.

Evaluating the Parameters of the Model

Unlike the sums of squares calculations through ordinary least squares, GzLMs are fit using maximum likelihood (ML) estimation. ML estimation uses an iterative approach to determine estimates for population parameter values (i.e., estimates of the slopes, standard errors, etc.) that maximize the likelihood that the sample data came from a population with these parameter values (for more detail about estimation, see Coxe, West & Aiken, 2013).

The contribution of the predictors can be evaluated by comparing nested models with the log-likelihood test and/or by testing individual coefficients with Wald tests. A log-likelihood (or likelihood ratio) test compares the fit of a model with the parameter of interest to that of a reduced model without the parameter of interest, whereas a Wald test determines whether some individual coefficient is significantly different from zero. If the parameter contributes to the prediction of the outcome, then the likelihood ratio would be large (favoring the model with the parameter of interest), and the Wald test statistic would also be large. As expected, a *p*-value can be calculated for both of these tests. The p-value on the log-likelihood ratio test reflects whether the change in model fit is statistically significant; the p-value on the Wald test reflects whether the individual coefficient is significantly different from zero. Both indicate the value of adding some predictor. Generally, log-likelihood tests are recommended with non-normal and/or heteroscedastic distributions, with *F*-based tests more appropriate with estimated dispersion parameters (e.g., gamma) and chi-square based tests more appropriate with fixed dispersion parameters (e.g., Poisson) (Venables & Ripley, 2002).

Model Fit and Model Comparison

An important part of all GzLMs is to determine whether the specified model provides an acceptable fit to the data. Methods for assessing model fit can include residual analyses, diagnostic tests, and information criterion fit statistics. Although the GLM's ordinary fit statistic, R^2, is not used in the GzLM, there is some literature on pseudo- R^2 values for generalized linear models, [e.g., Mittlböck & Heinzl, 2002], but their use is controversial. Certainly, the estimates obtained from models are already effect sizes. Generally, model fit focuses on the deviance. The deviance can be thought of as roughly similar to the residual sums of squares in regression, and represents the degree of lack of fit relative to a saturated model where a separate parameter is estimated for each case (i.e., the saturated model can perfectly reproduce the sample data, see Long, 1997). The deviance is composed of deviance residuals, which represent the individual contribution of each case to the overall deviance and which play an important role in determining the fit of a GzLM.

Deviance residuals are useful for analysis of model residuals for continuous responses. A normal probability plot of the deviance residuals allows for an assessment of substantive departures from normality. For continuous responses, the deviance residuals should be roughly normal if the distribution is appropriately selected. A plot of the deviance residuals against predicted values allows for visual detection of potential non-linearity and unequal error variances. Such a plot show roughly equal variability in the deviance residuals across the levels of the predictor.

Other diagnostics are available for determining whether a model is misspecified in either the family or the link function, though there is no single test that identifies the ‘correct’ model. The Pregibon (1980) goodness-of-link test has been used in the econometric literature to test for nonlinearity. More specifically, the Pregibon tests uses the predicted values and squared predicted values in a second model predicting the outcome; if there is no significant non-linearity that is detected, then the coefficient for the squared predicted value should not be significant. The modified Hosmer-Lemeshow (MHL; Hosmer & Lemeshow, 1995) test is an alternative for assessing the goodness-of-link. The MHL test conducts an ANOVA on the mean of the residuals for each decile of the predictor; a nonsignificant test statistic provides no evidence that the residuals are unevenly dispersed across the levels of the fitted values. Pearson’s correlation assesses systematic bias in fit by correlating the raw residuals on the outcome with the predictor; again, a nonsignificant test statistic provides no evidence of a poorly fitting model because there is no evidence that the residuals are related to the value of predictor. Results in Manning, Basu & Mullahy (2005) provide some evidence that these tools are effective at identifying ill-fitting models. For models that include only a single categorical predictor, we explain later that these diagnostics are not as readily applied. Further, it should be noted that these aforementioned statistics are merely tools and that their results are only guidelines, not strict criteria.

In lieu of hypothesis tests for diagnostics, fit statistics are available. Information-theoretic approaches, such as Akaike's Information Criterion (AIC, Akaike, 1974), bias-corrected AIC (AIC-C; Hurvich & Tsai, 1989), and Bayesian Information Criterion (BIC, Schwartz, 1978), are relative fit statistics that are useful for comparing nested and non-nested models. The AIC has been suggested for discriminating among sets of candidate error distributions (Dick, 2004). Dick (2004) generated data from lognormal, gamma, Weibull, log-logistic, and inverse-Gaussian processes, and GzLMs for those distributions were fit to each data set; there, AIC, as a criterion for distributional selection (he did not investigate AIC-C or BIC), seemed to be effective at identifying the correct data-generating distribution, particularly with higher samples sizes. In application, the AIC may have limited utility for identifying the 'true' data generating mechanism, but the AIC is effective for selecting the best choice among models under consideration (Dick, 2004). The AIC-C is a modified statistic that is less likely to select an over-parameterized model than AIC, and BIC provides a stricter penalty for estimating more coefficients in a model (i.e., both AIC-C and BIC correct for the AIC's bias towards more complicated models). By themselves, the statistics are meaningless; these statistics are used only for the comparison of multiple models. Models with lower values of information criteria relative to others are models that support the data better.

Simple Empirical Examples using GzLMs

Continuous Predictor. Arpin-Cribbie et al. (2012) conducted a study exploring perfectionism and its correlates in a group of 87 university students with elevated maladaptive perfectionism. Suppose that one hypothesizes that socially-prescribed perfectionism (SPP) predicts variability in negative automatic thoughts (NAT). SPP is a type of perfectionism that develops due to expectations from significant others and was measured using the Multidimensional Perfectionism Scale (Hewitt & Flett, 1991). NAT are negative thoughts or images that arise spontaneously and occur more often in those with maladaptive perfectionism (Nylund, 2004) and were measured using the Automatic Thoughts Questionnaire (ATQ; Hollon & Kendall, 1980). Scores represent the perceived frequency of NAT and are computed as a sum of 30 Likert-type items with anchors of “1 – not at all” to “5 – all the time”. With sum scores that range from 30 to 150, NATs are regularly treated as continuous.

The distribution of NATs, as described in previous literature (e.g., Gibb, Benas, Crossett & Uhrlass, 2007; Pirbaglou et al., 2013), typically has a long right tail. Figure 1 displays a scatterplot of the relationship between NAT and SPP, as well as a histogram displaying the distribution of NAT scores. It appears that NAT and SPP are related, and the distribution of NAT scores is indeed positively skewed. The variability is also related to the level of the predictor; if we split SPP into tertiles, the variability of the first, second and third tertiles are 449.39, 598.75 and 624.63, respectively. Because the NAT variable may be treated as continuous and because there is evidence that variability systematically increases across levels of the predictors, the gamma family is a possibility.

In this example, we explore the following models: 1) gamma distribution with a log link; 2) gamma distribution with an inverse link; 3) Gaussian (normal) distribution with a log link (this should not be confused with the ordinary linear model with transformation of the raw outcome, as opposed to the expected values); and 4) a Gaussian distribution with an identity link (equivalent to a simple linear regression) to highlight a model that would be expected to fit very poorly. [Note: One could explore other models that allow for positive skewness and heteroscedasticity, such as the inverse-Gaussian distribution. However, upon testing, this yielded very poor fit because the observed nature of the heteroscedasticity did not match the heteroscedasticity implied by their structures; that is, the hypothesized mean-variance relationship [variance proportional to mean cubed] for the inverse-Gaussian distribution was too extreme for the observed data and was therefore not selected for presentation].

Figure 2 displays a normal probability plot of the deviance residuals. Residuals that lie close to the diagonal line on the plot indicate less departure from normality. Table 3 summarizes the Pregibon, MHL, and correlation results, as well as the AIC, AIC-C, and BIC values. As expected, the models with Gaussian distributions did not fit as well because the distribution of residuals is positively skewed. The gamma distribution with the inverse link function fit the data best because it yielded the lower information criteria values. However, this model also failed the Pregibon test.

Here, there is a disagreement between fit criteria and diagnostic tests. Certainly, hypothesis tests for diagnostic tests are subject to the limitations of significance testing. Thus, although residual diagnostics, tests, and information criteria all provide some information about model fit, the choice of model is not pure statistical. When disagreement in fit statistics does occur in practice, theoretical considerations should play a strong role in model selection. In this case, we selected the model with log link. Previous studies had used log transformations on the raw response for inferences about the geometric mean; the log link in this case provides some comparability while also providing the alternative inference to the arithmetic mean. Figure 3 displays the deviance residuals against the predicted values for the gamma (log link) model, which allows for an assessment for heteroscedasticity. There is no obvious evidence of heteroscedasticity.

The likelihood ratio test indicated that SPP significantly predicts NAT score, *b* = .011, *F* (1, 81) = 19.97, *p* < .001. For interpretation in original scale, recall that the model has the log link function. To predict NAT scores in the original scale, we use the inverse function, which inverts the model equation. The mean function corresponding to the log link involves exponentiation; therefore, we exponentiate the fitted model. With an intercept of 3.64, we write the prediction equation as NAT' = exp(3.64 + .011*SPP) = exp(3.64)*exp(.011*SPP) = 37.91*exp(.011*SPP). Higher levels of SPP are associated with higher levels of predicted NAT levels. For example, someone at the mean level of SPP (66.24) has a predicted NAT value of 37.91*exp(.011*66.24) = 78.94, whereas someone one standard deviation below the mean (52.35) has a predicted value of 37.91*exp(.011*52.35) = 67.75, and someone with a value one standard deviation above the mean (80.13) has a predicted value of 37.91*exp(.011*80.13) = 91.97. Figure 4 displays a scatterplot of the raw SPP and NAT scores, along with the fitted line from the significant Gamma model with a log link.

Categorical Predictor. Arpin-Cribbie et al. (2012) also compared the effects of three different interventions (cognitive behavioral therapy [CBT], stress reduction therapy [SRT], and a no treatment option [NT]) for students with elevated maladaptive perfectionism levels. Students were randomly assigned to one of the interventions. For this example, we focus on the post-test NAT (NATP). Figure 5 displays density plots of scores in each of the treatment conditions following the intervention. The positive skewness of NATP scores is evident for each of the treatment groups. Because we will model only a single categorical predictor, the distribution of group scores also correspond to the distribution of group residuals. For descriptive purposes, gamma density curves are overlaid on these histograms to show how the group distributions approximate gamma shapes. The means of the CBT, SRT, and NT groups are 57.17 (20% trimmed mean = 53.61, *SD* = 20.13), 69.62 (20% trimmed mean = 65.26, *SD* = 31.13), and 83.88 (20% trimmed mean = 81.88, *SD* = 25.66), respectively. Dummy variables were created with the CBT group as the reference group. Consistent with the earlier gamma models, the CBT group, which has the lowest mean, also has lower variability than the two groups.

We explore the same models as in the continuous predictor example, namely: 1) a Gamma distribution with a log link; and 2) a Gamma distribution with an inverse link; 3) a normal model with a log link; and 4) a normal (Gaussian) distribution with an identity link. A normal probability plot of the deviance residuals (Figure 6) indicates that the model with Gaussian distributions do not fit as well as the gamma distributions. The Gamma models (log/inverse link) had the lowest AIC, AIC-C and BIC value (765.67, 766.18 and 775.35, respectively). The Gaussian models (781.27, 781.78, and 790.94 for AIC, AIC-C and BIC, respectively) did not fit nearly as well. Note here that all Gamma models had the same fit statistics and deviance statistics, regardless of the link function; the same can be said for the Gaussian models. This implies that the use of deviance residuals, GOF tests and information criteria is less relevant for the comparison of models that have the same single categorical predictor. Whereas one would observe different fit statistics between two models of single continuous predictors with the same family but different link function (e.g., gamma with log link versus gamma with inverse link), one would observe identical fit statistics between two models of single categorical predictors, given the same family specification. Further, regardless of the link function and family specification, the predicted values for each group, in original scale, would be identical across models (by maximum likelihood, predicted values for each category are the groups' sample means). In other words, for a GzLM alternative to the one-way ANOVA case, one would only need to inspect residuals for assessing the family specification.

For these data, we select the gamma distribution with the inverse link for a different model interpretation, such that the inverse of negative automatic thoughts may be considered the lack of negative symptoms. Figure 7 displays the spread of the residuals by the predicted values for each group. We can see that there is slightly greater variability in the residuals for those in the STR group, likely a function of the variances not being proportional to the squared means (e.g., for the dummy variable comparing the CBT group to the STR group, the squared means for the STR group are about 1.5 times that for the CBT group, while the variances are more than twice as large in the STR group; thus, the larger residual variability is not surprising).

A likelihood ratio test for the gamma distribution with an inverse link provided evidence for a significant difference between the groups, F (2, 80) = 6.97, p=.002. (the two degrees of freedom reflect that the difference is two parameters, i.e., the two dummy variables). Further, it was also found that those receiving CBT (predicted value = 57.17) had significantly lower NATP than those receiving either the STR (predicted value = 69.62), t (80) = 2.009, p = .0492, b = -.003, $CI_{.95}$ = [-.0062, -.0001] or the NT options (predicted value = 83.88), t (80) = 3.717, p = .0004, b = -.006, $CI_{.95}$ = [-.0086, -.0026]. Because the intercept for this model was .017, we can write the prediction equation as NATP' = 1/(.017 - .003*SRT - .006*NT). Notice that although the sign of the slopes are negative, we are working with the inverse; so, being in either the STR or NT groups is predictive of higher NATP scores. As an example, we can reproduce the predicted value for the CBT group from above via NATP' = 1/(.017 - .003*SRT - .006*NT) = 1/(.017 - .003*0 - .006*0) = 57.17. It should be further noted that maximum likelihood estimates for each group (or levels of the predictor) are statistically unbiased. For the case of the single-categorical predictor, predictions in original scale for all models with the same family specification are equal to sample means by maximum likelihood estimation (i.e., predicted values are equal for gamma models with either the log or inverse link). Comparing across models with different families, inferential tests as obtained by the gamma models are more valid than those obtained in models with a poorer family specification (e.g., Gaussian models). Thus, for the one-way case, the appropriateness of the family specification is more important for obtaining better precision and better confidence in the results of inferential tests.

Summary and Conclusion

Researchers in psychology are often confronted with continuous outcomes that are distributed with positive skewness and variances that relate to the level of the predictor. Of those who prefer the use of alternative methods over mere reliance on the robustness of traditional models, most transform the outcome variable to try to achieve normality and/or homoscedasticity while others use some robust statistic. However, transformations do not always simultaneously normalize distributions and stabilize variances; further, when inference to means is of interest, the naïve back-transformation of coefficients is insufficient for recovering expected means. Trimmed mean estimators are also limited in that they often make use of a limited amount of information from the full distribution of relevant observations.

The GzLM offers certain advantages over these alternatives. Firstly, the separate specifications for the distribution and link function offers flexibility for achieving linearity, whilst raw data transformations require simultaneous corrections. Secondly, the link function and its inverse function allows for interpretation in both transformed scale and in original scale, while power transformations could have less clear interpretations in transformed scale. Thirdly, when distribution tails are of interest, the GzLM allows for the specification of distributions that explicitly model non-normality, while trimmed mean estimators typically trim away the more extreme cases; in other words, the GzLM offers estimator sufficiency while trimmed mean estimators do not. Fourthly, when strict statistical assumptions are met, parametric procedures are most powerful, and the modeling framework can provide richer interpretations. Finally, the GzLM framework allows for a variety of research designs with any number of predictors while trimmed mean estimators may currently be applied only to a limited number of designs (e.g., factorial designs). While the GzLM accommodates any number of potential predictors, the GzLM, as an overarching framework that encompasses the general linear model, is also suitable for simple designs. Even with a single predictors, it may be advantageous to implement the GzLM in place of regular linear regressions and one-way ANOVAs when it may be appropriate.

Although we have highlighted the advantages of interpretation in original scale and of GzLM estimator sufficiency, and though we advocate for researchers to consider the GzLM a potential avenue of analysis, even for designs with a single-grouping variable, we are not suggesting that the GzLM is a panacea that will be applicable for all possible scenarios in which skewness and heteroscedasticity occurs. Indeed, model misspecification, as with other modeling procedures, is a resounding problem when the observed residual structure does not well match the theoretical error structure; for example, if the observed variances are independent of the means, then it would not be logical to specify a gamma distribution just because one observes a positively skewed distribution. Indeed, the statistical assumptions pertain to error structure, not the distribution of the response variable itself. Further, the choices in model selection may sometimes be unclear, as formal diagnostic tests and fit statistics may disagree, particularly because diagnostic tests are also subject to potential issues with Type I error and power. It would be important for researchers to use all available statistical tools alongside theoretical consideration for model selection.

When the GzLM for continuous outcomes clearly does not fit well, then there are alternatives besides the less appropriate traditional methods. If one desires to follow a route that encompasses modeling, one could consider extensions of the GzLM. One could 'customize' one's own link function in the GzLM by either incorporating Box-Cox power transformations or other transformations that might induce linearity; or, one might go beyond the GzLM into other frameworks, such as the generalized gamma model (Manning, Basu, & Mullahy, 2005) or the generalized additive models (Hastie & Tibshirani, 1986). If, instead, there is little desire for model-based procedures and if one considers the cost of generalizability to be justified, then trimmed mean estimators could be used instead.

Overall, GzLMs can be a valuable alternative to traditional linear models and robust methods. We recommend that applied researchers add the GzLM to their statistical toolboxes and to be aware of situations where GzLMs might be a better tool than traditional approaches for modeling continuous outcomes.

Compliance with Ethical Standards

Funding: This research was supported by the Social Sciences and Humanities Research Council of Canada.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Arpin-Cribbie, C. A., Irvine, J., & Ritvo, P. (2012). Web-based cognitive-behavioral therapy for perfectionism: A randomized controlled trial. *Psychotherapy Research, 22*, 194–207. doi: 10.1080/10503307.2011.637242.

Basu, A., Manning, W. G., & Mullahy, J. (2004). Comparing alternative models: log vs Cox proportional hazard? *Health Economics*, *13*, 749–65. doi:10.1002/hec.852

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*, 78-84.http://dx.doi.org/10.1027/1614-2241/a000057

Blough, D. K., Madden, C. W., & Hornbrook, M. C. (1999). Modeling risk using generalized linear models. *Journal of Health Economics, 18*, 153–171. doi:10.1016/S0167-6296(98)00032-0

Coley, R. L., Votruba-Drzal, E., & Schindler, H. S. (2008). Trajectories of Parenting Processes and Adolescent Substance Use: Reciprocal Effects. *Journal of Abnormal Child Psychology, 36*, 613-625. doi: 10.1007/s10802-007-9205-5.

Coxe, S., Aiken, L. S., and West, S. G. (2013). Generalized linear models. In T. Little (Ed.), *Oxford Handbook of Quantitative Methods, Vol 2: Statistical Analysis*. New York: Oxford University Press.

Cribbie, R. A., Fiksenbaum, L., Keselman, H. J., & Wilcox, R. R. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal* *of Mathematical and Statistical Psychology, 65*, 56–73. doi:10.1111/j.2044-8317.2011.02014.x

Darlington, R. B. (1990). *Regression and linear models.* New York, NY: McGraw-Hill.

Dick, E. J. (2004). Beyond “lognormal versus Gamma”: discrimination among error distributions for generalized linear models. *Fisheries Research, 70,* 351–366. doi:10.1016/j.fishres.2004.08.013

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association, 78,* 605–610. doi:10.1080/01621459.1983.10478017

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models: Second Edition.* Thousand Oaks, CA: Sage Publications.

Gibb, B. E., Benas, J. S., Crossett, S. E., & Uhrlass, D. J. (2007). Emotional maltreatment and verbal victimization in childhood*. Journal of Emotional Abuse, 7*, 59-73. doi: 10.1300/J135v07n02_04

Gill, J. 2001. Generalized linear models: a unified approach. Sage University Paper: London.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165. doi: 10.1037/0022-006X.68.1.155

Hewitt, P. L., & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60,* 456-470. doi:10.1037/0022-3514.60.3.456

von Hippel, P. (2010). Skewness. In Lovric, M. (Ed.), *International Encyclopedia of*

Statistical Science. New York, NY: Springer.

Kappes, H. B., Sharma, E., & Oettingen, G. (2012). Positive fantasies dampen charitable giving when many resources are demanded. *Journal of Consumer Psychology, 23,* 128-135. doi:10.1016/j.jcps.2012.02.001

Keselman, H. J., Huberty, C., Lix, L., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., Levin, J. R. (1998). Statistical Practices of Educational Researchers: An Analysis of their ANOVA, MANOVA, and ANCOVA Analyses. *Review of Educational Research, 68,* 350–386. doi:10.3102/00346543068003350

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics and bootstrapping: Circumventing the biasing effects of heterescedasticity and nonnormality. *Journal of Modern Applied Statistical Methods, 1*, 288–309. Retrieved from http://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1670&context=jmasm

Lindeberg SI, Eek F, Lindbladh E, Ostergren PO, Hansen AM, & Karlson B (2008). Exhaustion measured by the SF-36 vitality scale is associated with a flattened diurnal cortisol profile. *Psychoneuroendocrinology, 33*, 471-477. doi: 10.1016/j.psyneuen.2008.01.005

Lindsey, J. K., & Jones, B. (1998). Choosing among generalized linear models applied to medical data. *Statistics in Medicine, 17*, 59–68. DOI: 10.1002/(SICI)1097-0258(19980115)17:1<59::AID-SIM733>3.0.CO;2-7

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.*

Thousand Oaks, CA: Sage.

Mall, S., Lund, C., Vilagut, G., Alonso, J., Williams, D. R., & Stein, D. J. (2015). Days out of role due to mental and physical illness in the South African stress and health study. *Social Psychiatry and Psychiatric Epidemiology, 50,* 461–468. doi:10.1007/s00127-014-0941-x

Manning, W. G. & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics, 20,* 461–94. doi:10.1016/S0167-6296(01)00086-8

Manning, W. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics, 17*, 283–295. doi: 10.1016/S0167-6296(98)00025-3

Manning, W., Basu, A., & Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics, 24,* 465–88. doi:10.1016/j.jhealeco.2004.09.011

Mittlböck, M, & Heinzl, H. (2002). Measures of explained variation in Gamma regression models. *Communications in Statistics: Simulation and Computation, 31,* 61-73. doi:10.1081/SAC-9687282

Neal, D., & Simons, J. (2007). Inference in regression models of heavily skewed alcohol u se data: A comparison of ordinary least squares, generalized linear models, and bootstrap resampling. *Psychology of Addictive Behaviors, 21,* 441–452. doi:10.1037/0893-164X.21.4.441

Nelder, J. A., & Wedderburn, R. W. M. (1972). General linearized models. *Journal of the*

Royal Statistical Society: Series A, 135, 370–384. doi:10.2307/2344614

Nevill, A., & Copas, J. (1991). Using generalized linear models (GLMs) to model errors in motor performance. *Journal of Motor Behavior, 23*, 241–50. doi:10.1080/00222895.1991.9942035

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15,* 1-9. Retrieved from http://pareonline.net/getvn.asp?v=15&n=12.

Pirbaglou, M., Cribbie, R. A., Irvine, J., Radhu, N., Vora, K., & Ritvo, P. (2013). Perfectionism, anxiety, and depressive distress: Evidence for the mediating role of negative automatic thoughts and anxiety sensitivity. *Journal of American College Health, 61,* 477-483. doi: 10.1080/07448481.2013.833932

Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology, 42*, 203–213. doi:10.1111/j.2044-8317.1989.tb00910.x

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*, 510–32. doi: 10.1037/0033-2909.114.3.510

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136, MR 468014.

Sayer, N.A., Sackheim, H.A., Moeller, J.R., Prudic, J., Devanand, D.P., Coleman, E.A., & Kiersky J.E. (1993). The relations between observer-rating and self-report of depressive symptomatology. *Psychological Assessment, 5,* 350–360. doi: 10.1037/1040-3590.5.3.350

Siegel, A. F. (1982). Robust regression using repeated medians. *Biometrika, 69*, 242–244. doi: 10.1093/biomet/69.1.242

Sohn, B. Y. & Kim, G. B. (1997). Detection of outliers in weighted least squares regression. *Journal of Applied Mathematics and Computing, 4*, 501-512. doi: 10.1007/BF03014491

Staudte, R. G. & Sheather, S. J. (1990). *Robust estimation and testing.* New York: Wiley.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S (Fourth edition).* New York: Springer.

Wang, X., Dang, X., Peng, H. & Zhang, H. (2009). The Theil-Sen estimators in multiple linear regression models. Manuscript available at: http://home.olemiss.edu/~xdang/papers/MTSE.pdf

Welsh, A. H. (1987). The trimmed mean in the linear model. *Annals of Statistics, 15*, 20--36. doi:10.1214/aos/1176350248.

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing (3rd ed.).* San Diego, CA: Academic Press.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254–74. doi:10.1037/1082-989X.8.3.254

Wilcox, R. R. & Keselman, H. J. (2012). Modern regression methods that can substantially increase power and provide a more accurate understanding of associations. *European Journal of Personality, 26*, 165–174. doi: 10.1002/per.860

Table 1

Common family specifications and corresponding variance and link functions for positively skewed distributions

Distribution	Variance Function	Canonical Link
Gaussian	Constant	Identity
Poisson	μ_i	Log
Gamma	μ_i^2	Inverse
Inverse-Gaussian	μ_i^3	Inverse-square

Table 2

Common link functions and their corresponding mean functions

Link Name	Link Function, η=g(μ_i)	Mean Function, μ = g^{-1}(η_i)
Identity	μ_i	η_i
Log	$\log_e\mu_i$	$e^{(\eta i)}$
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2

Table 3

Fit statistics for the four models assessing the relationship between Negative Automatic Thoughts (ANT) and Socially Prescribed Perfectionism (SPP)

Model	AIC	AIC-C	BIC	Pregibon	MHL	Corr
Gamma (log)	*758.16*	*758.46*	*765.41*	pass	pass	pass
Gamma (inverse)	**756.94**	**757.24**	**764.20**	fail	pass	pass
Gaussian (log)	761.84	762.15	769.10	fail	pass	pass
Gaussian (identity)	763.87	764.18	771.13	fail	pass	pass

Note: AIC = Akaike Information Criteria; AIC-C = bias-adjusted AIC; BIC = Bayesian Information Criteria; Pregibon = Pregibon link test; MHL = modified Hosmer-Lemeshow goodness of fit test; Corr = Pearson's correlation test between the raw residuals and the values of the predictor. Bolded texts indicate lowest values for the information criteria value; italicized text indicates values that are within two points of the lowest information criteria value.

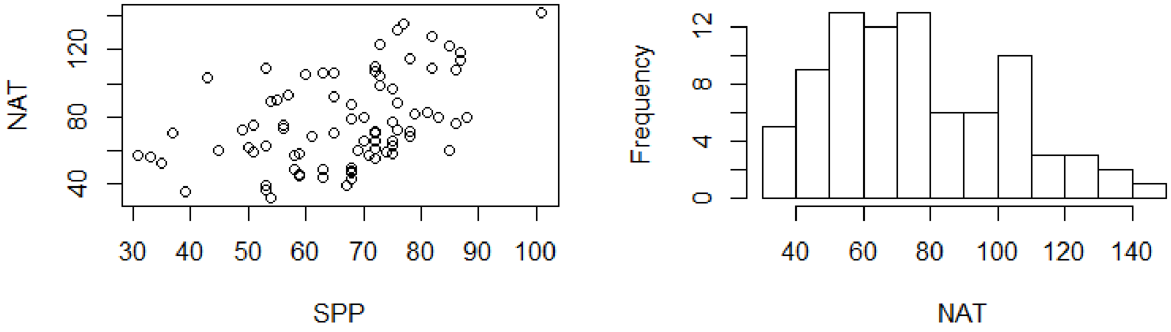

Figure 1. a) Scatterplot of the positive relationship between Negative Automatic Thoughts (NAT) and Socially-Prescribed Perfectionism (SPP). b) Histogram of NAT scores demonstrating nonnegative, positively skewed values.

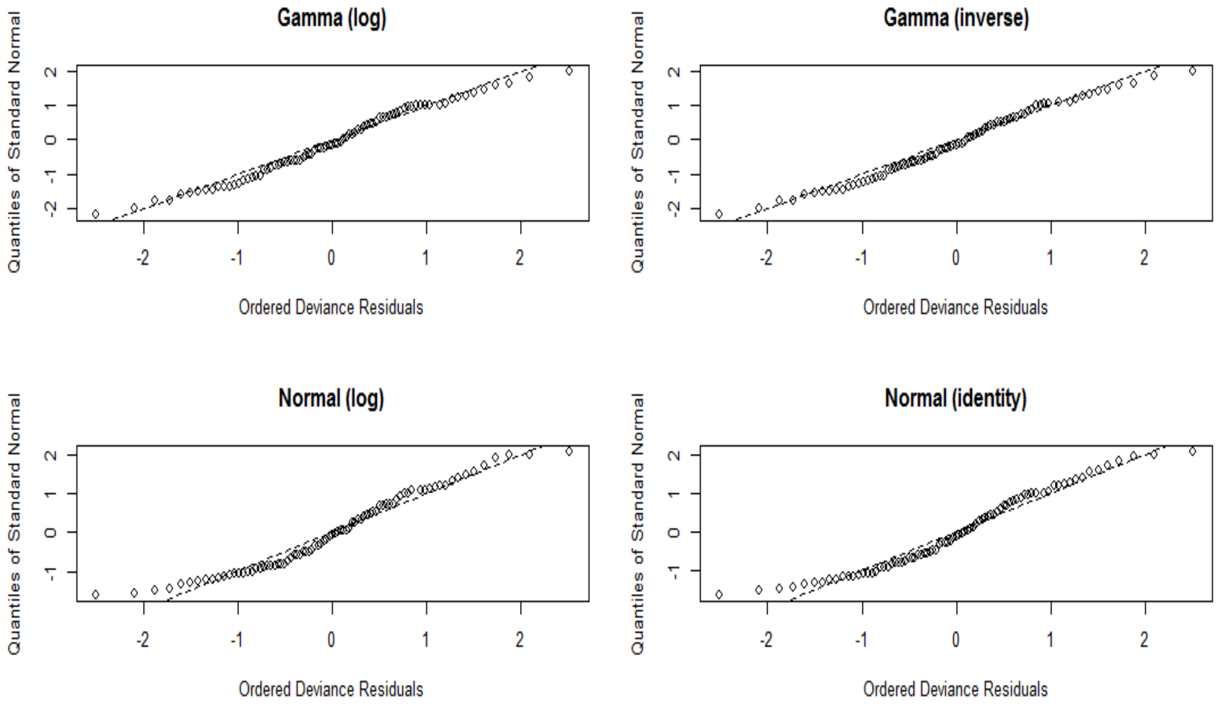

Figure 2. Normal probability plot of the standardized deviance residuals for the distributions (link functions) modeling the relationship between Negative Automatic Thoughts (NAT) and Socially Prescribed Perfectionism (SPP), using models

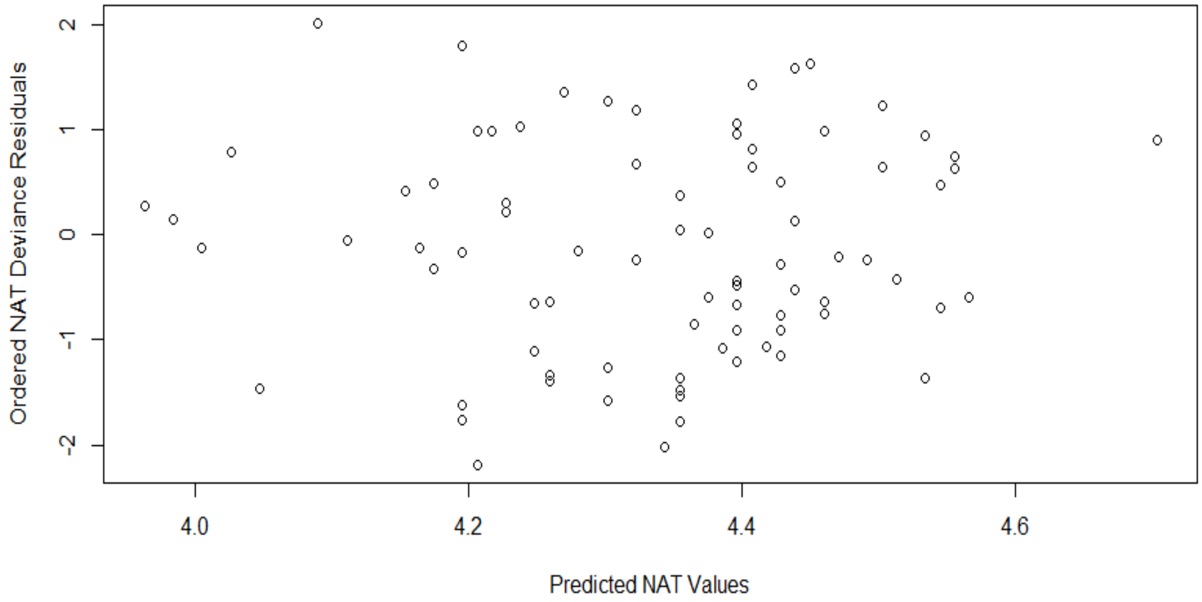

Figure 3. Plot of the standardized deviance residuals against the predicted values for Negative Automatic Thoughts (NAT)

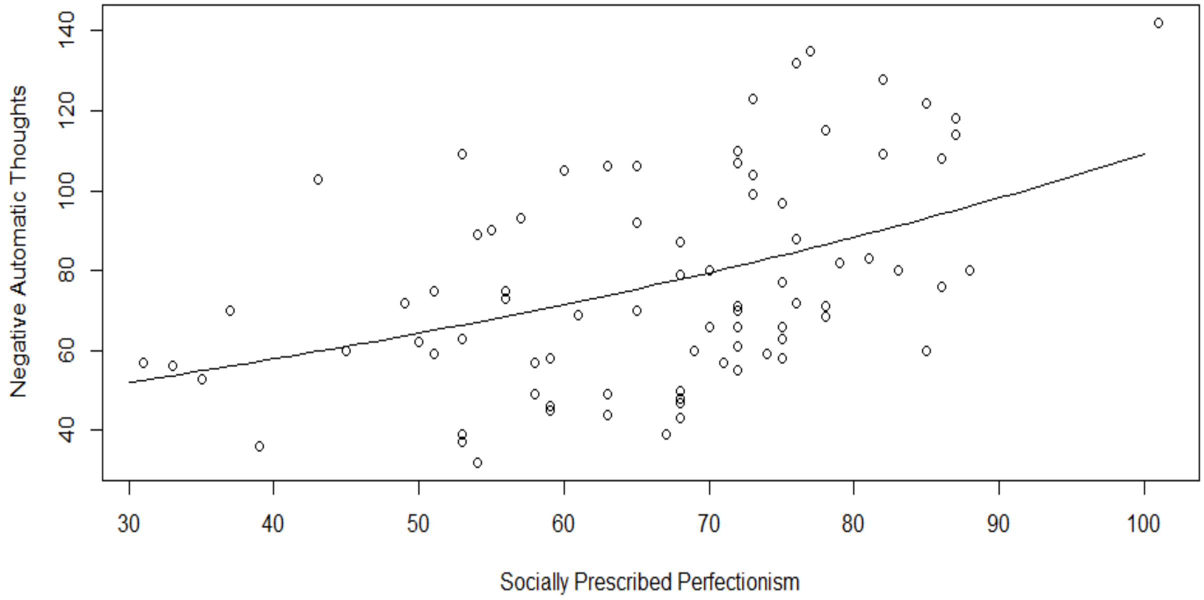

Figure 4. Scatterplot of the relationship between Negative Automatic Thoughts (NAT) and Socially Prescribed Perfectionism (SPP), along with the fit line for the gamma distribution with a log link

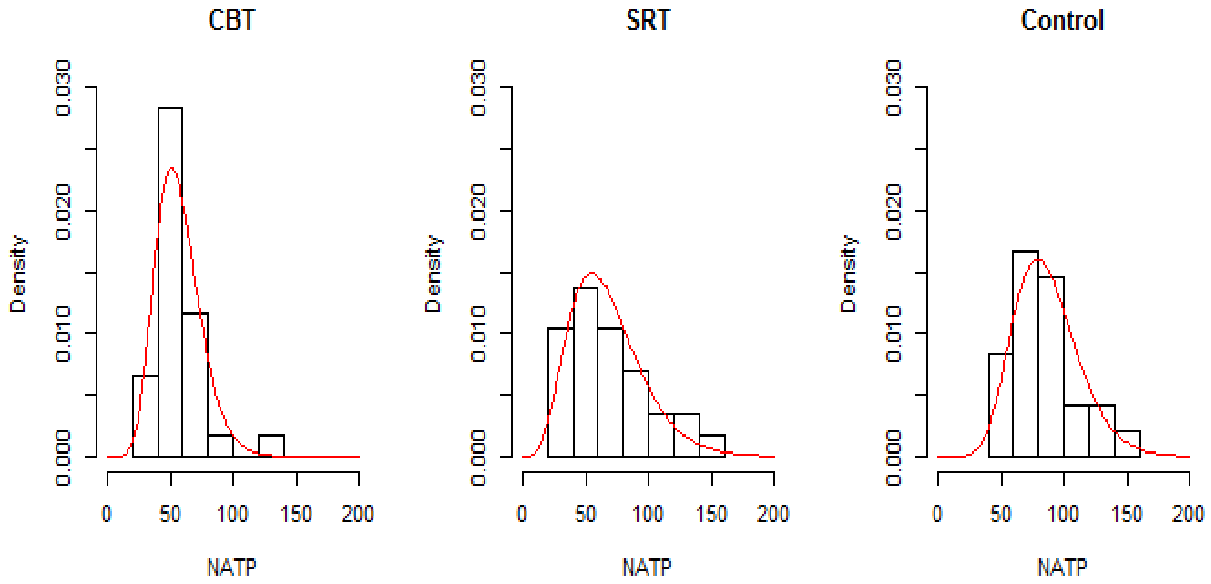

Figure 5. Histograms of the post-test Negative Automatic Thoughts (NATP) scores for each of the Cognitive Behavioral Therapy (CBT), Stress Reduction Therapy (SRT) and Control groups. The red line is the fit line for the gamma distribution.

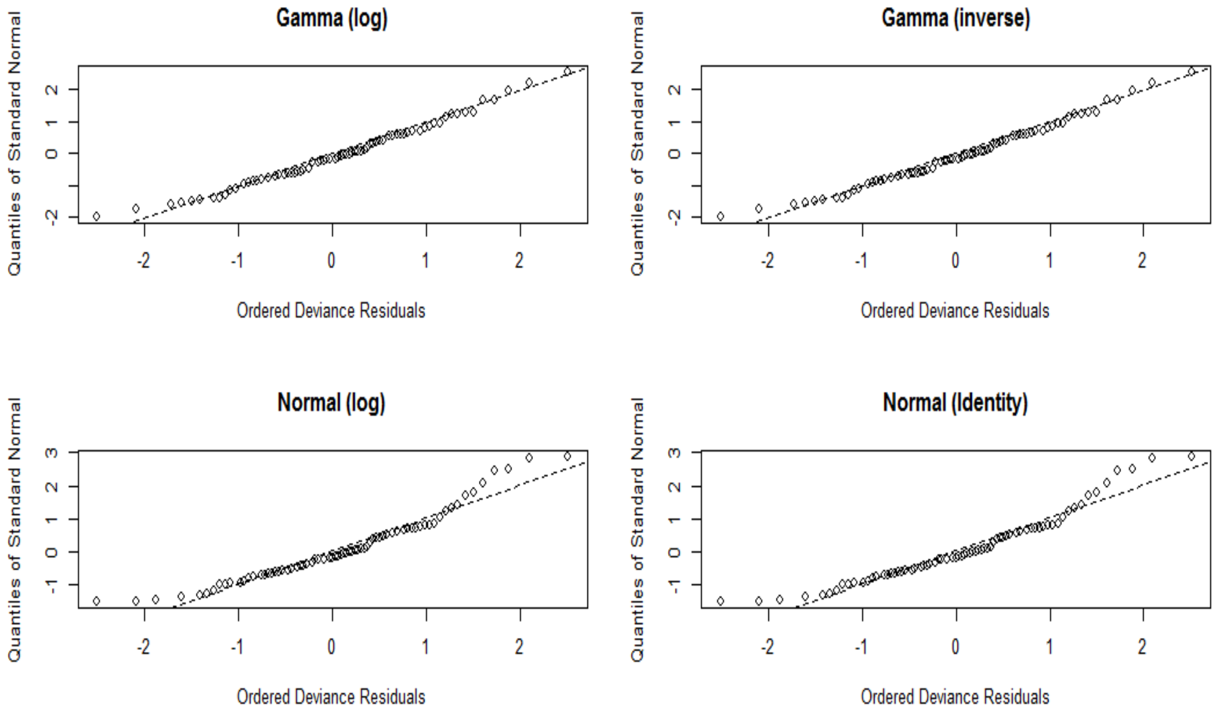

Figure 6. Normal probability plot of the standardized deviance residuals for the distributions (link functions) modeling the relationship between post-test Negative Automatic Thoughts (NATP) and the treatment groups, using competing models

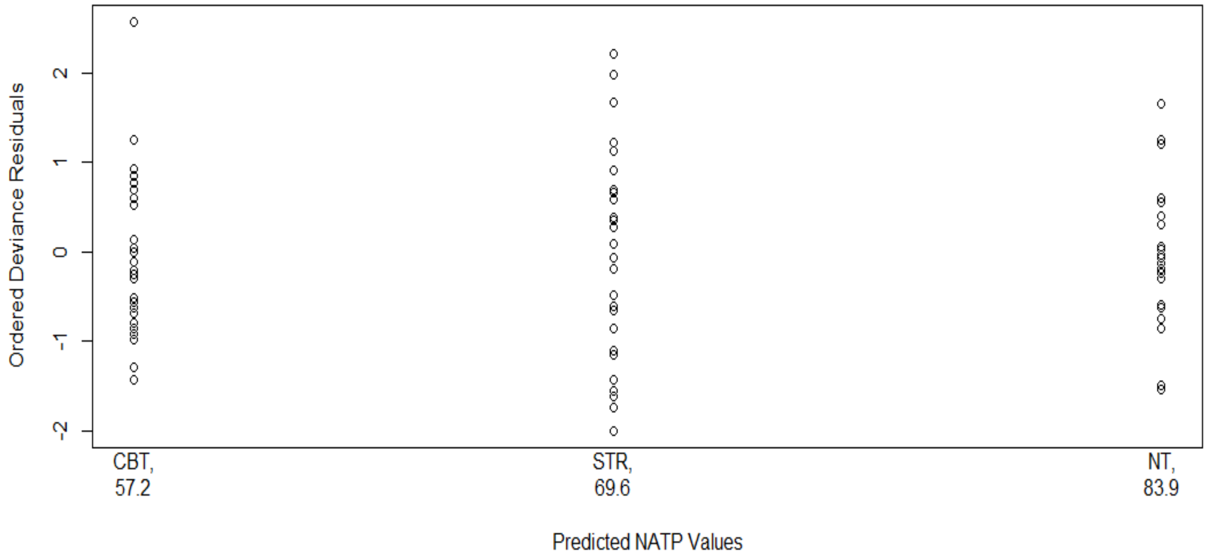

Figure 7. Ordered deviance residuals for each of treatment groups for the gamma model with an inverse link function