Twentieth vs. twenty first century statistics: evaluating the degree of improvement in model fit between general and generalized linear modelling approaches.

Rachel Buxton, Jens Currie, Sheena Fry, Megan Rector, Laura Siegwart, Linda Takahashi and Nyssa Trip

Biology 7932 Final Examination Report, Department of Biology, Memorial University of Newfoundland

Table of Contents

List of Tables

List of Figures

Abstract

Twentieth century data analysis in the context of normal errors and identity link is limited to datasets that assume normality and homogeneous error. In reality, many datasets describing natural processes and patterns do not meet these assumptions. General linear model (GLM) violations have traditionally been corrected through variable transformations and blurring interpretation of true cause and effect relationships. In the late 20th century, the generalized linear model (GzLM) was developed to accommodate a variety of error distributions and response variables under one unifying framework. Our objective was to compare 20th and 21st century statistics to evaluate the benefits and limitations of the GzLM. This was achieved by applying the GzLM to data previously analyzed using GLMs from 20th century statistics textbooks and our own research. We predicted that residual homogeneity would increase and dispersion parameters would decrease for GzLMs. We also predicted that GzLMs would meet the dispersion parameter criteria of the AIC approach more often than GLMs. Binomial tests on improvement of residual vs. fits plots indicated that GzLM did not improve residual error more often than the GLM. We attribute this to use of predominantly continuous data, which is more likely to meet the assumptions of homogeneity. However, GzLM significantly improved dispersion factors due to shrinking standard error and better fit of the data to the new model structure. When change and direction of change in significance of Type I errors were analyzed using binomial and multinomial logistic regression, neither revised error distribution nor data type significantly affected changes in significance or direction of Type I error change. Unlike the GzLM, original models (GLM) with normal error were rarely suitable for AIC analyses, since dispersion factors were almost always greater than 4. Our comparison shows that GzLMs provide a better model fit to the data with a lower dispersion factor that allows for AIC analysis. The GzLM is useful for obtaining models that better fit the data, and subsequently allow for the assembly of a parsimonious model with an optimal error structure.

Introduction

Why do we need Generalized Linear Models?

A brief examination of the literature in the field of biology will show that dependent variables are rarely normally distributed and continuous. This suggests that the general linear model, with assumptions of normally distributed errors and continuous response variables, is an inadequate approach to model selection and inference for many types of biological datasets. The generalized linear model (GzLM) provides an elegant solution to this problem by accommodating many types of dependent variables and error structures into one linear based model (Guisan et al. 2002; Hoffman 2004). In addition, GzLMs can address issues related to overdispersion and residual deviance (Guisan et al. 2002). Here, we compare linear and generalized linear models based on a series of datasets in order to quantify changes in model fit and Type I error as well as provide an overview of the benefits and limitations of GzLMs in the context of biological data analysis.

Linear Regression

The GzLM is an extension of classic linear models, which is based on least squares regression and can be summarized by the following equation (1):

(1) $Y = \alpha + X^T\beta + \varepsilon$

Where:

Y = response variable
α = *a* constant known as the intercept
$X = (X_1, \ldots X_p)$ a vector of *p* predictors
$\beta = \{\beta 1, \ldots, \beta p\}$ vector of *p* regression coefficients for each predictor
ε = normal error

Traditionally, violation of the assumptions that errors are normally distributed and homogeneous has been addressed using transformations of the response variable and/or predictors (Guisan et al. 2002).

One of the problems inherent in this approach is that interpretations of results are based on the transformed structure rather than the actual data. In addition, transformations don't always result in homogeneous and normal errors. The most significant advance in regression analysis in the last 30 years has been the development of the GzLM, which overcomes many inherent limitations of ordinary least squares regression.

The Generalized Linear Model

Nelder and Wedderburn first introduced the GzLM in 1972. Although various techniques for analyzing non-normal data had been developed by this time, including probit and logit regression (analyses that remain special cases of the GzLM (Gill 2000)), the GzLM provided a unified approach for fitting these models. The systematic component of all of these techniques had a linear structure, which allowed them to create a streamlined approach to non-normal data analysis based on maximum likelihood estimates that can be obtained using iteratively weighted least squares (Nelder and Wedderburn 1972). The GzLM has 3 basic components:

1. Random variables that share the same error distribution from the exponential dispersion family and a constant scale parameter, including: normal, binomial, multinomial (poisson with constraints), and gamma error.
2. A linear component (2) that describes how the response distribution responds to changes in the explanatory variables.

(2) $Y = \mathrm{X}\beta$

3. A link function connecting the linear component and the expected value of the dependent variable.

(3) $Y_{\mathrm{i}} = g_i(\mu_i)$

The main advantage of the GzLM is its capacity to accommodate a variety of error distributions, allowing analysis of data with many types of response variables to be unified under one framework.

This approach is especially beneficial in biology since counts, binomial, multinomial responses, and zero inflated data are common in this field. The GzLM also eliminates the need for transformations by using a desired error distribution and a link function to tie the error structure to the linear part of the model rather than using a default Gaussian distribution. There are several error and link combinations that are commonly used (Table 1).

Table 1. Summary of error distributions and link functions used in the GzLM approach.

Error structure	***Canonical link***	***Formal equation***
Gaussian	inverse	$g(\mu) = \mu^{-1}$
Negative binomial	Log	$g(\mu) = \log\mu$
Binomial	logit	$g(\mu) = \log(\pi/1-\pi)$
Poisson	Log	$g(\mu) = \log\mu$
Gamma	inverse	$g(\mu) = \mu^{-1}$
Gamma	identity	$g(\mu) = \mu$
Inverse Gaussian	inverse square	$g(\mu) = \mu^{-2}$
Quasi	Log	$g(\mu) = \log\mu$

*Modified from Pregibon (1980)

Although the GzLM has many advantages, it is not without its limitations. Use of the GzLM is restricted to the exponential dispersion family because iteratively weighted least squares (the machinery behind the GzLM) works only within the exponential family. As well, the linear component of the model is retained and responses must be independent. Strategies to overcome these limitations exist, but accessible software is not widely available (Lindsey 1997).

Information Theoretic Approach

Information Theoretic Approaches such as Akaike's Information Criterion (AIC) is another recent development in data analysis that is recommended for analysis of observational data, and is therefore of interest to research biologists. Burnham and Anderson (1998) argue that traditional statistical inference is a limited approach because the model structure is assumed and only the parameters in that model are estimated. In contrast, AIC allows data-based selection of a “best” or most parsimonious model that can then be used on new data for traditional hypothesis testing (Lindsey 1997).

In order to perform AIC analysis, a global model (incorporating all potentially relevant predictors) and set of candidate models (incorporating all biologically relevant predictor sets) are created. The global model must have an appropriate structure; meaning that the data should not be overdispersed (should have a dispersion factor of less than 4). Once models are identified AIC values are assigned to all candidate models using the following equation (4):

(4) $\text{AIC} = -2 \log(L(\ddot{\Theta}|y) + 2K$

Where:

$\log(L(\ddot{\Theta}|y)$ = numerical value of the log likelihood at its maximum point
likelihood = probability model with parameters $\ddot{\Theta}$
$y = x,g$
x = empirical data, g = approximate model
K = number of estimable parameters in the model

The AIC value represents an estimate of the relative distance between the fitted model and the unknown true model that underlies the observed data. The AIC value of each model is compared to the lowest AIC value in order to generate a ΔAIC. Values less than two indicate that there is substantial support for the model while values greater than ten indicate that there is no support for the model.

One of the advantages of AIC is the necessity of thoughtful consideration of the system under analysis in order to select an appropriate global model including all potentially relevant predictors and candidate models that make biological sense. However AIC is limited to the global and candidate models that are included in analysis; other possible models are not considered. One limitation of AIC is its relative nature. AIC values are compared among candidate models; therefore models are compared to each other. Although one model may be the “best” out of the set, it still may not be a good model in an absolute sense.

Objectives and Hypotheses

Our objective was to compare 20th and 21st century statistics to evaluate the benefits and limitations of the GzLM. This was achieved by applying the GzLM to data previously analyzed using a linear model. Data from 20th century statistics textbooks, as well as from our own research, were used to assess changes in residual plots, Type I error, and dispersion factors. We also examined the use of GzLMs as global models in AIC analyses as a way to meet the dispersion criteria of AIC modelling, and to compare the final model selection of AIC analyses based on traditional regression models versus GzLMs. We predicted that residual homogeneity would increase with application of the GzLM due to improved model fit and that dispersion parameters would decrease due to reduced discrepancy between observed and expected values resulting in a lower chi-square statistic. We also predicted that GzLMs would meet the dispersion parameter criteria of the AIC approach more often than linear models.

Methods

We compared the general linear model (GLM) with the GzLM by examining various assumptions about the residuals. We initially analysed data sets using a GLM (normal error structure) from textbooks, refereed literature, and our own data. The suitability of GLM results were assessed for appropriateness by analyzing homogeneity and normality of the residuals (McCullagh and Nelder 1989). All datasets, especially those that did not meet the assumptions of normal and homogenous residuals, were subjected to further analysis using the GzLM. Analysis with the GzLM allowed us to incorporate different error structures to try and improve the homogeneity and normality of the residuals. We considered gamma, poisson, binomial, quasi and negative binomial distributions when re-analysing the data sets using a GzLM. Homogeneity and normality of the residuals was again used as a criterion to determine whether the new error structure was appropriate. Of all error structures (families and links) used in the GzLM, the one that gave the most homogenous deviance residuals and lowest dispersion parameter was considered the most appropriate (McCullagh and Nelder 1989; Figure 1). The data sets were analysed using a variety of statistical programs including R, S-plus, and SPSS. Changes in the estimates of Type I error, from the original model (GLM) to the revised model (GzLM), were recorded for each parameter estimate, if the two models were comparable. Additional information including changes in the dispersion factor and the AIC values were also recorded (Figure 1).

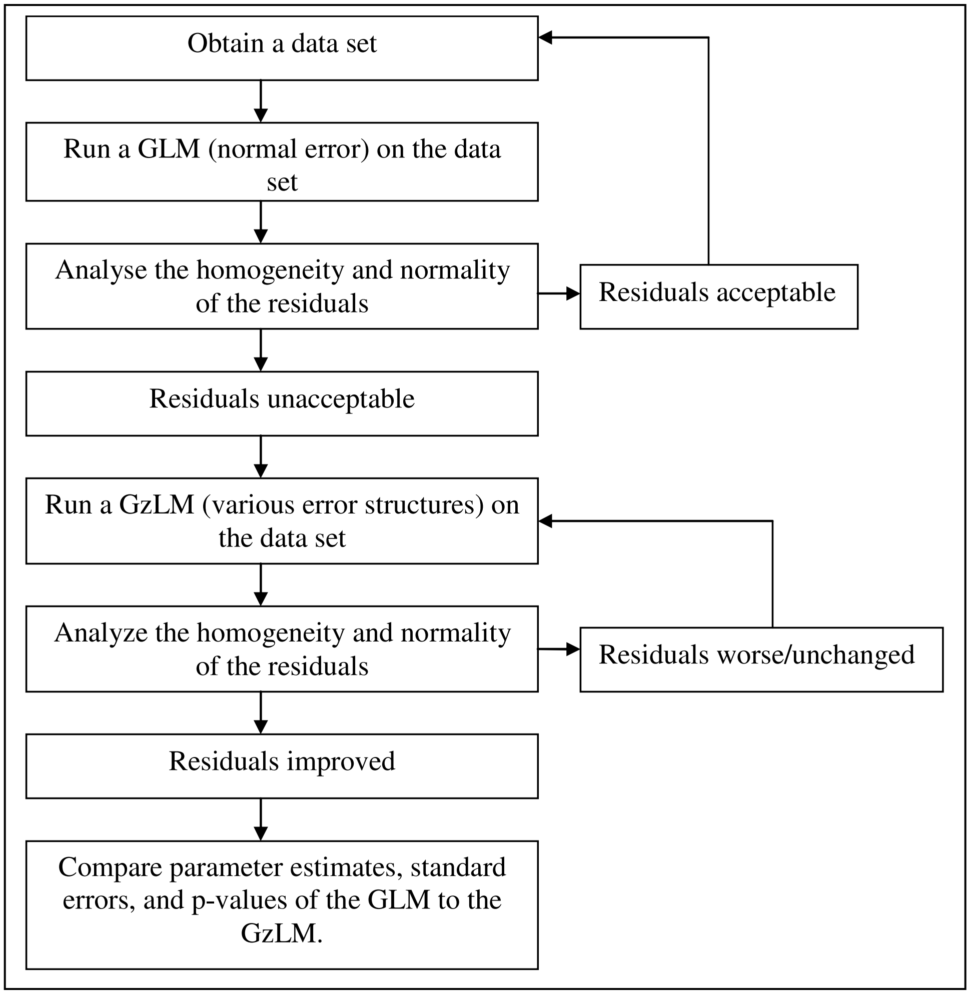

Obtain a data set
Run a GLM (normal error) on the data set
Analyse the homogeneity and normality of the residuals
Residuals acceptable
Residuals unacceptable
Run a GzLM (various error structures) on the data set
Analyze the homogeneity and normality of the residuals
Residuals worse/unchanged
Residuals improved
Compare parameter estimates, standard errors, and p-values of the GLM to the GzLM.

Figure 1. General approach for comparing 20th century statistic to 21st century statistics.

Data analysis

Differences among Inspection Techniques

A binomial test was done in S-plus to determine if the use of the GzLM to revise the GLM was warranted (significant p-value). This was done using two criteria: the graphical inspection of a residual vs. fit plot and a statistical inspection of dispersion parameters. The binomial was run twice, once to determine if the improvement in graphical inspection was significance and a second time to determine if the improvement in statistical inspection was significant.

Change in Predictor Variable Significance

Binomial logistic regression was used to determine the effects of the revised error structure (gamma, poisson, binomial, quasi and negative binomial) and response variable data type (count, continuous, percent) on the change in significance of Type I error (significant to non-significant or vice-versa).

(5) $\text{Change.in.Significance} = \mu + \text{Binomial error}$

(6) $\mu = ß_0 + ß_{\text{Revised.error.structure}}\text{Revised.error.structure} + ß_{\text{Data.type}}\text{Data.type}$

Change in Type I Error

The overall change in Type I error of all the parameters tested was analysed by looking at the change in p-value, whether it increased, decreased, or remained unchanged when going from the original model to the revised model. This change was then expressed as a percentage (% increase, % decrease, and % unchanged).

(7) $\text{Change.of.Type.I.error} = \mu + \text{gamma error}$

(8) $\mu = ß_0 + ß_{\text{Revised.error.structure}}\text{Revised.error.structure} + ß_{\text{Data.type}}\text{Data.type}$

Influence of Revised Error Structure and Data Type on Type I Error

A multinomial regression was then used to determine the effects of the revised error structure (gamma, poisson, binomial, quasi and negative binomial) and data type (count, continuous, percent) on change in direction of the p-value as an ordinal variable (whether it remained unchanged = 0, increased

= 1, or decreased = 2 when going from the GLM to the GzLM). P-values were obtained from the Deviance residuals, using a chi-squared distribution in Excel.

(9) $\text{Direction.of.change.in.TypeI} = \mu + \text{gamma error}$

(10) $\mu = ß_0 + ß_{\text{Revised.error.structure}}\text{Revised.error.structure} + ß_{\text{Data.type}}\text{Data.type}$

AIC Model Selection

AIC analysis was done in SPSS to compare model selection using a GLM and a GzLM. A set of candidate models were compared first using the GLM (normal error structure, with over dispersion). The most parsimonious model selected in this analysis was then compared to the most parsimonious model selected using a GzLM (revised error structure, with an improved dispersion).

Results

Differences among Inspection Techniques

Binomial logistic analysis of graphical inspection (i.e. fitted values vs. residuals) and statistical inspection (i.e. dispersion parameters) revealed significant differences in only some aspects of model improvement. Graphical inspection was found to be statistically insignificant when data were re-analyzed under a generalized linear model (GzLM) framework (p = 1.000, Table 2). The GzLM framework was found not to improve the appearance of fitted values vs. residuals plots. In contrast statistical inspection was found to be statistically significant when data were re-analyzed under the GzLM framework (p = 0.000, Table 2). The GzLM framework greatly reduced the value of the dispersion parameters.

Table 2. Differences among inspection techniques for model improvement.

Inspection Technique	Total Number of Observations	Number of Observations showing no improvement	Number of Observations showing improvement	Probability of improvement	Probability of no improvement
Graphical	56	28	28	1	1
Statistical	56	11	45	0*	0*

*Indicates statistical significance (p<0.05).

Change in Predictor Variable Significance

Using binomial logistic regression change in predictor variable significance was regressed onto revised error structure, the error structure under the GzLM framework and type of response variable. The original error structure under the GLM framework is always Gaussian. Predictors, revised error structure and type of response variable were found to not have a significant effect on change in predictor variable significance (Table 3). Figure 2 is a graphical inspection of the residuals, which indicates a violation of homogeneity in residuals. However this violation should not affect the p-values since they were found to be so far from significance (Table 3).

Table 3. Results of binomial regression on predictor variable significance.

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.656e+01	2.400e+03	-0.007	0.994
Revised : Binomial	3.313e+01	2.683e+03	0.012	0.990
Revised: gamma	1.532e+01	2.400e+03	0.006	0.995
Revised: gamma	1.367e+01	2.400e+03	0.006	0.995
Revised: gaussian	1.462e+01	2.400e+03	0.006	0.995
Revised : ngbinomial	1.656e+01	2.400e+03	0.007	0.994
Revised: poisson	1.766e+01	2.400e+03	0.007	0.994
Revised: quasi	1.628e+01	2.400e+03	0.007	0.995
Type: continuous	-2.144e-03	5.283e-01	-0.004	0.997
Type: count	NA	NA	NA	NA

NA= test appears not to recognize multiple response variables coded as count. NA error did not occur when test was performed with count variables alone. Revised=Revised Error Structure. Revised error structure is the error structure under the GzLM framework. Type=Type of Response Variable.

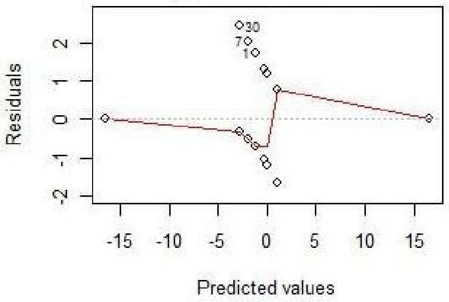


Figure 2. Plot of the relationship between predicted values and standardized deviance residuals for binomial regression on predictor variable significance.

Change in Type I Error

Figure 3 shows the percentage change in direction of p-values, a measure of Type I Error inflation. Switching from the GLM framework to the GzLM framework increased the p-value in 79 out of 156 observations or 50.6% of the time. In 65 observations or 41.7% the p-value remained unchanged and in only 12 observations or 7.6% of the time did p-values decrease. This indicates the GzLM framework deflates Type I Error.

Influence of Revised Error Structure and Data Type on Type I Error

A multinomial regression of the change in direction of the p-value revealed no significant influence of revised error structure or data type (Table 4). Table 5 is an aggregate the sub-categories of each predictor. When examined as an aggregate, type of response approaches significance (p=0.051). Figure 4 are graphical inspections of the assumptions under the GzLM framework in which the test was conducted. No serious violations of the assumptions appear to have occurred.

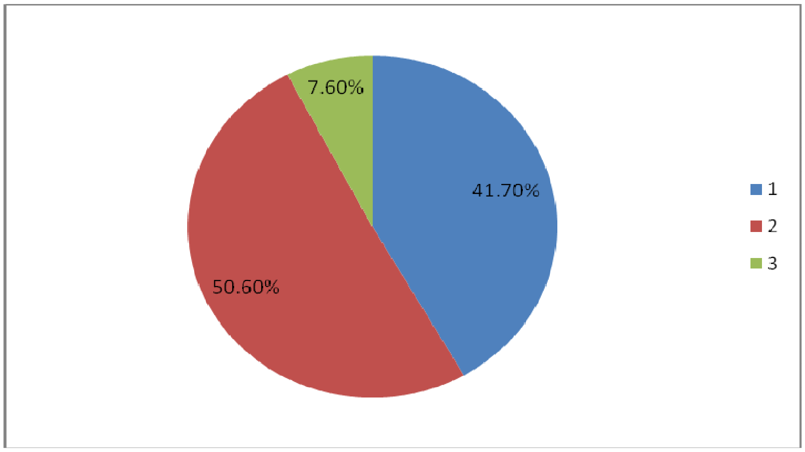


Figure 3.Pie chart of percentage of change in direction of the p-value (1=Unchanged, 2=increase, 3=decrease).

Table 4. Results of multinomial regression of the effect of revised error structure and data type on Type I Error.

Coefficients	Value	Std. Error	t value	p-value
(Intercept)	2.5	0.318905	7.839323	8.21E-13
Revised: gamma	0.213359	0.513679	0.415355	0.678484
Revised: gaussian	0.166667	0.557924	0.298727	0.765568
Revised: ngbinomial	-0.60491	0.554146	-1.09161	0.276777
Revised: poisson	-0.47991	0.604367	-0.79408	0.428423
Revised: quasi	0.480094	0.534109	0.898869	0.370183
Type: continuous	-0.5	0.601709	-0.83097	0.407331
Type: count	-0.02009	0.616584	-0.03258	0.974055

Revised=Revised Error Structure. Revised error structure is the error structure under the GzLM framework.
Type=Type of Response Variable.

Table 5. Aggregated results of multinomial regression of the effect of revised error structure and data type on Type I Error.

Predictor variable	Deviance residuals	p-value
revised	5.268569	0.38399
type	5.924363	0.051706

Revised=Revised Error Structure. Revised error structure is the error structure under the GzLM framework.
Type=Type of Response Variable.

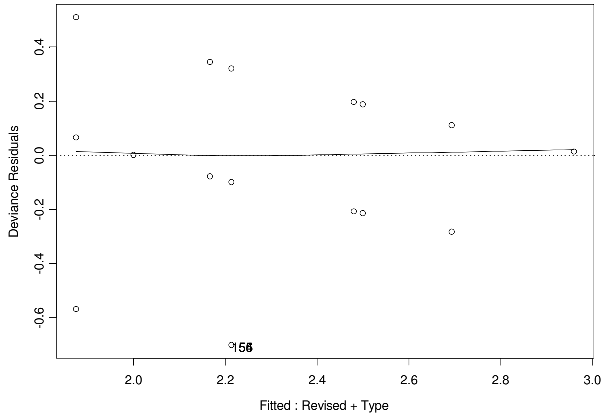


Figure 4. Plot of the relationship between predicted values and standardized deviance residuals for multinomial regression of the effect of revised error structure and data type on Type I Error.

AIC Model Selection

Table 6 displays results from an AIC analysis conducted on data from the CORT dataset under the GLM framework (see Appendix). Under the GLM framework the global model (norm_GM) was selected as the most parsimonious, w_i=1. Table 7 contains the same set of candidate models as Table 6 however the analysis was run under the GzLM framework, specifically gamma and log link. In Table 7 the most parsimonious model was gam_1, w_{i}=0.57. Clearly AIC model selection is heavily influenced by the GzLM framework since the most parsimonious model differed under the two frameworks.

Table 6. AIC Analysis of Gull Island (CORT) Data under GLM framework.

Model Codes	Model	LogLikelihood	K	AIC	AICc	DAICc	exp	wi
norm_GM	Cort = Ld+Y+S+T+M+Ld*Y+S*M +S*M*T+T*M	-112927	20	225894	225900.32	0	1	1
norm_1	Cort = M	-140352	1	280706	280706.03	54805.71	0	0
norm_2	Cort = Ld+M	-133109	2	266222	266222.08	40321.76	0	0
norm_3	Cort = Ld+M+S	-125620	4	251248	251248.27	25347.95	0	0
norm_4	Cort = Ld	-134479	1	268960	268960.03	43059.71	0	0
norm_5	Cort = S	-137552	1	275106	275106.03	49205.71	0	0
norm_6	Cort = Ld+Y+S+T+M	-119994	10	240008	240009.54	14109.22	0	0
norm_7	Cort = Ld+Y+S+T+M+S*M*T	-117244	13	234514	234516.6	8616.28	0	0

GLM framework assumed to have Gaussian error structure. LogLikelihood=Likelihood function of the model parameters, K= the number estimable parameters in an approximating model, AIC=uncorrected AIC value, AIC_c=corrected AIC value, exp=exponent of delta, where delta is AIC differences relative to the smallest AIC value in the set, w_i: Akaike Weights= estimates of the probability of model *i* being the most parsimonious model given the data and the model set. Variable Codes for models are: Ld = ordinal date, Y = year, S = sex, T = time, M = month.

Table 7. AIC Analysis of Gull Island (CORT) Data under GzLM framework.

Model Codes	Model	LogLikelihood	K	AIC	AICc	DAICc	exp	wi
gam_1	Cort = M	-877.764	1	1757.53	1757.55	0	1	0.57
gam_2	Cort = Ld+M	-877.239	2	1758.48	1758.56	1	0.61	0.35
gam_3	Cort = Ld+M+S	-876.609	4	1761.22	1761.49	3.93	0.14	0.08
gam_4	Cort = Ld	-883.289	1	1768.58	1768.6	11.05	0	0
gam_5	Cort = S	-883.558	1	1769.12	1769.14	11.59	0	0
gam_6	Cort = Ld+Y+S+T+M	-876.121	10	1772.24	1773.78	16.23	0	0
gam_7	Cort = Ld+Y+S+T+M+S*M*T	-875.943	13	1777.89	1780.49	22.93	0	0
gam_GM	Cort = Ld+Y+S+T+M+Ld*Y+S*M+S*M*T+T*M	-875.405	20	1794.81	1802.54	44.98	0	0

GzLM framework is gamma with log link. LogLikelihood=Likelihood function of the model parameters, K= the number estimable parameters in an approximating model, AIC=uncorrected AIC value, AIC_c=corrected AIC value, exp=exponent of delta, where delta is AIC differences relative to the smallest AIC value in the set, w_i: Akaike Weights: estimates of the probability of model *i* being the most parsimonious model given the data and the model set. Variable Codes for models are: Ld = ordinal date, Y = year, S = sex, T = time, M = month.

Discussion

Quantitative methods in biology have changed at a particularly rapid rate in the past few decades. With the advent of powerful computer software allowing the possibility of running iterative algorithms, statistical tests have become more unified and sophisticated. In particular, these new statistics include the innovation of the Generalized Linear Model (GzLM) by Nelder and Wedderburn (1972) and further by McCullagh and Nelder (1989), and a paradigm shift to an information theoretic approach by Burnham and Anderson (1998) utilizing a method called Aikaike's Information Criterion (AIC). GzLM is a technique that allows for analysis of data with different types of response variable distributions and errors structures. The GzLM uses maximum likelihood, as opposed to least squares estimation, which makes it more applicable then the GLM. Maximum Likelihood estimates use an algorithm for successively improving estimates for the likelihood function near the maximum. On the other hand, Information Theoretic Approach throws away the idea of hypothesis testing and instead attempts to select the most parsimonious model structure out of a set of candidate models. In this study, we set out to determine whether these 21st century statistical methods are more successful in generating an appropriate model structure than 20th century statistics.

The GLM assumes that the variance of the error is constant for all combinations of the independent variable, termed homogeneous errors. When this is true, a plot of the residual versus fits will display an even band. A common situation, especially when dealing with behavioural data, is that the variability of the error increases with an increase in the values of the independent variable, which results in a conical shape in the residual versus fits plot (Hoffman 2004). When dealing with heterogeneous errors one typically finds that standard errors of the coefficients are biased, thus significance tests are incorrect and one's ability to make inferences from the model is compromised. In the current study, we observed no significant improvement in shape from conical to band in the residual versus fit plot when a GzLM approach was taken compared to the original GLM approach. This could be due to the fact that we used mostly continuous data (75%) which are more likely to meet the assumptions of homogeneity. Furthermore, much of this data was from older textbooks (pre-2000) which most likely used this data because it fit theses assumptions.

One common problem in biological data is overdispersion. This arises naturally when experimental units are subject to random variations due to biological or environmental factors. Typically, these are factors that are affecting the response but which have not been measured. In this case, standard errors of estimated regression co-efficients will be smaller than they should and tests of hypotheses will have inflated probabilities of Type I error (smaller p-values) (Quinn and Keough 2002). Standard errors are often inflated by a factor of $\surd(\chi^2/df)$, or rather, are adjusted based on a dispersion parameter. We observed a significant improvement in the dispersion factors when the error structure and link were revised in a GzLM (p=0.000). This is most likely due to a shrinking standard error and better fit of the data to the new model structure. If the model is revised and the data fit the revised model more closely, the standard deviation of the sample mean (distance of each data point to the fitted model) will decrease. The dispersion factor is calculated by dividing the Pearsons Chi square (χ^2) by the degrees of freedom and is a measure similar to the mean squared error (D. Schneider class notes). The Chi square statistic is the squared difference of the observed (or data) and expected (or model) value divided by the expected value and summed across classes. A perfect fit would equal 0 and the Chi squared statistic increases as the difference between observed and expected values increase. The degrees of freedom must be taken into account to evaluate this statistic because the chi squared increases as the number of categories (degrees of freedom) becomes more numerous. The drastic reduction in dispersion factors when revising model structure from GLM to GzLM was evident when we compared the average values for the GLM (average = 59622.43) to that of the GzLM (average = 23.18). This is likely due to better fit of the data with the GzLM structure. If the data fit these revised models better the difference between the observed and expected values would decrease, thereby decreasing the Chi squared statistic.

The ability of the GzLM to accommodate various types of error distributions (normal, binomial, multinomial, and gamma) allows for analysis of data with many types of response variables. In the present study we looked at the effect of error distribution using the GzLM and data type (continuous, count, etc.) on the direction of change in Type I error. In the majority of cases Type I error increased from the GLM to the GzLM (51%) with relatively few comparisons resulting in a decrease (7%). There were also many cases in which there was no change in Type I error between the original and revised model (42%). When the effect of revised error distribution and original data type on the change in direction of Type I error was tested, both explanatory variables were found to be insignificant. Also, we wanted to determine if a change in error distribution caused a change in Type I error such that the decision to accept or reject the null hypothesis changed. That is, does a change in error distribution cause a variable that was significant in the original analysis to become insignificant with a new error structure or vice versa. We found that error distribution and data type did not significantly influence Type I error. Although no significant effect of the explanatory variables (revised error distribution and data type) was found, there is still an obvious trend (increase or no change) in the direction of change in Type I error. To better understand the reasoning behind this it is first important to understand the mechanics of the GzLM, particularly maximum likelihood.

The GzLM is based on likelihood-ratio statistic to test for goodness of fit. A goodness of fit test compares the model fit with the data. This approach regards the data as representing the fit of the most complex model possible (the saturated model, which has a separate parameter for each observation). The measure of goodness of fit, or G-statistic, is based on the theoretical underpinning of the likelihood theory, which considers how likely the data are given the model. The G-statistic (11) uses the ratio of the observed to fitted value taken as a likelihood (L) ratio (12):

(11) G = 2 ∑ lnL

= 2 ∑ observed(log(observed/expected))

(12) L = (observed/expected)$^{\text{observed}}$, where observed = data, expected = model

When the fit of the data to the model is perfect the likelihood ratio is L = 1 and thus the deviation of the data to the model will be zero, resulting in a G- statistic of zero. As previously mentioned, the GzLM allows for the use of different error distributions. The deviance of these error distributions is as follows:

(13) Normal $\sum(z-\hat{\mu})^2/\sigma^2,$

(14) Poisson $2\{\sum z \ln(z/\hat{\mu}) - \sum(z-\hat{\mu})\},$

(15) Binomial $2[\sum z \ln(z/\hat{\mu}) + \sum(n-z)\ln\{(n-z)/(n-\hat{\mu})\}],$

(16) Gamma $2p\{-\sum \ln(z/\hat{\mu}) + \sum(z-\hat{\mu})/\hat{\mu}\}.$

The correct use of these error distributions will create a model that better fits the data, which mean the likelihood ratio will be closer to 1. This in turn will provide a more accurate G- statistic, close to zero, which in turn is used to calculate the change in deviance (ΔG). This deviance is used to calculate the level of Type I error. Thus a change in model structure, using a better fitting error, will result in a lower change in deviance and a model that better fits the data. In summary, a GzLM with the correct error distribution that more closely fits the data will result in a smaller G-statistic and a p-value that is more accurate.

The Information Theoretic Approach discards the idea of hypothesis testing, estimating model parameters, and model precision; instead it ranks various candidate models in order to choose the most parsimonious model that best accounts for patterns in the data. This alternative analytic approach rejects the assumption of hypothesis testing - that the model structure is known and correct and that only parameters in that model are to be estimated. Although these two methods (GzLM versus AIC) are very different, they do have some similarities. Both attempt to find the model and error structure that best fit the data, eliminate over-dispersion, and require a low dispersion factor. In fact, to perform AIC analysis the dispersion factor must be under 4 (Burnham and Anderson 1998). Based on our analyses, a dispersion factor under 4 using a normal error structure is unlikely, meaning that we can rarely perform AIC analysis on data analyzed using normal error structure. Furthermore, AIC analysis requires a log likelihood value for each candidate model which can only be obtained using the maximum likelihood framework of a GzLM. As we observed, using an over-dispersed global model (GLM) versus a non-dispersed global model (GzLM) greatly affects the final choice of a most parsimonious candidate model. In one respect, hypothesis testing using GzLM has facilitated the Information Theoretic Approach by innovating ways to decrease over-dispersion and utilize iterative maximum likelihood functions.

Our data suggest that GzLMs provide a better model fit to the data, in turn providing a model with a lower dispersion factor that allows for AIC analysis. AIC analysis then allows us to find the most parsimonious model to explain complex observational data. In conclusion, the GzLM is useful for obtaining models that better fit the data (lower dispersion factor), allow for AIC analysis, and subsequently allow for the assembly of a parsimonious model with an optimal error structure (not necessarily normal).

Literature Cited

Burnham, K. P. and D. R. Anderson. 1998. Model selection and multimodel inference: a practical Information-Theoretic Approach. Springer Series: New York.

Gill, J. 2000. Generalized Linear Models: A Unified Approach. (Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-134). Thousand Oaks, CA: Sage.

Guisan, A., Thomas, C., Edwards, Jr. and H. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89-100.

Hoffman, J.P. 2004. Generalized Linear Models: An Applied Approach. Pearson: USA.

Lindsey, J.K. 1997. Applying Generalized Linear Models. Springer: New York.

McCullagh, P. and J.A. Nelder. 1989. Generalized Linear Models 2nd edition. Chapman and Hall: London. (mathematical statistics of generalized linear model) 511 Pp.

Nelder, J.A. and R. W. M. Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society (Series A) 135: 370-384.

Pregibon, D. 1980. Goodness of Link Tests for Generalized Linear Models. Applied Statistics 29: 15-24.

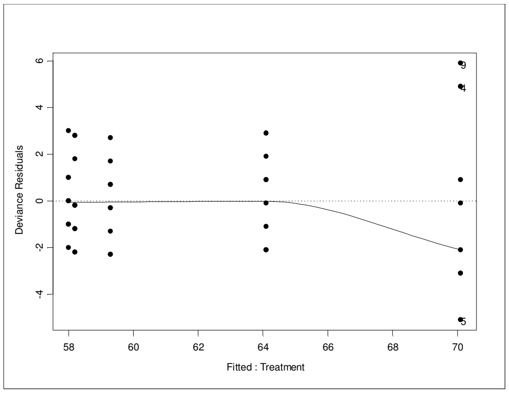
Quinn, G. P., and Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge, UK: Cambridge University Press.

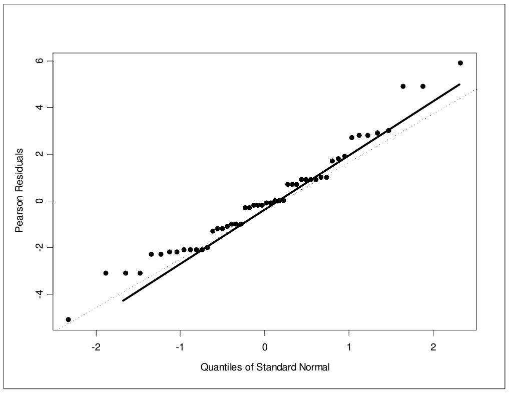
Appendices

<u>Peas:</u> The effect of different sugars (*treatments*) on length in ocular units (*length*) of pea sections grown in tissue culture with auxin present (n=10 replications/group).

Data from pg 218 in Sokal & Rohlf, 1995. Biometry. Freeman and Company, 887 Pp.

Model: Length = μ + normal error $\mu = ß_0 + ß_{treatment}$



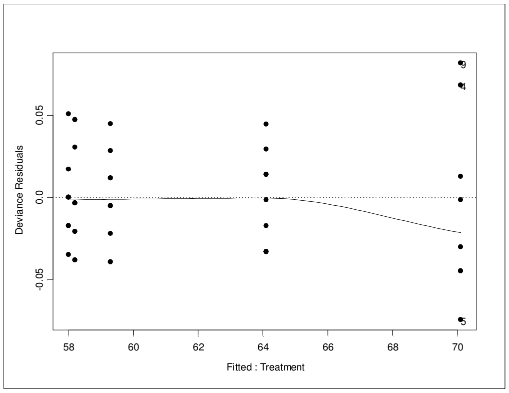


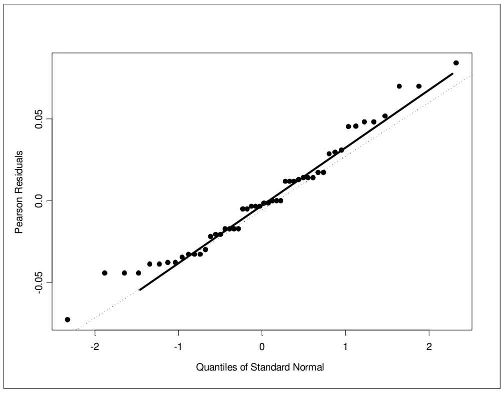
	Df	Value	Std. Error	t value	p-value
(Intercept)		70.1	0.738617	94.907061	
TreatmentGlucose	45	-10.8	1.044563	-10.339255	0.0000
TreatmentFructose	45	-11.9	1.044563	-11.392328	0.0000
TreatmentGluFru	45	-12.1	1.044563	-11.583795	0.0000
TreatmentSucrose	45	-6	1.044563	-5.744031	0.0000

Dispersion Parameter for Gaussian family taken to be 5.455556.

Null Deviance: 1322.82 on 49 df, Residual Deviance: 245.5 on 45 df.

Revised model: e^{μ} + gamma error (log link) $\mu = ß_0 + ß_{treatment}$





	Df	Value	Std. Error	t value	p-value
(Intercept)		70.1	0.793415	88.352221	
TreatmentGlucose	45	-10.8	1.039224	-10.392369	0.0000
TreatmentFructose	45	-11.9	1.031227	-11.539651	0.0000
TreatmentGluFru	45	-12.1	1.029783	-11.750054	0.0000
TreatmentSucrose	45	-6	1.075112	-5.580814	0.0000

Dispersion Parameter for Gamma family taken to be 0.001281.

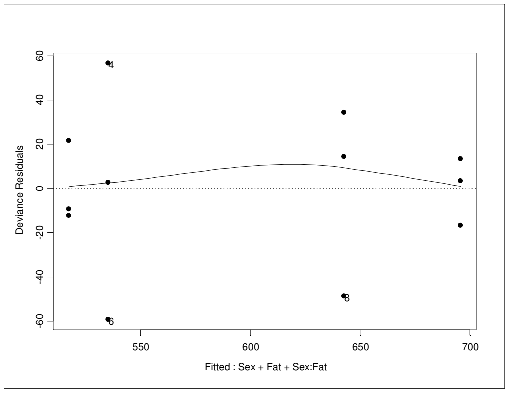
Null Deviance: 0.3276964 on 49 df, Residual Deviance: 0.0571755 45 df

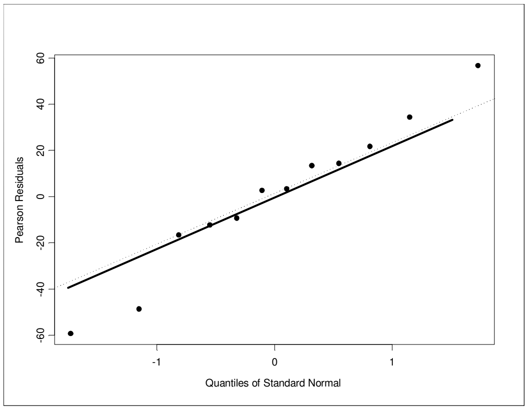
Summary: GLM assumptions not met since residuals are slightly heterogeneous and non-normal. Model revision resulted in no improvement in the residuals, however, it did make the null deviance acceptable.

<u>Rat diet:</u> Differences in food consumption when rancid lard was subsitituted for fresh lard in the diet of rats. Data classified by fat (fresh vs. rancid) and by sex (male vs. female).

Data from pg 324 in Sokal & Rohlf, 1995. Biometry. Freeman and Company, 887 Pp.

Model: Consumption = μ + normal error $\mu = ß_{0} + ß_{sex} + ß_{fat} + ß_{sex*fat}$



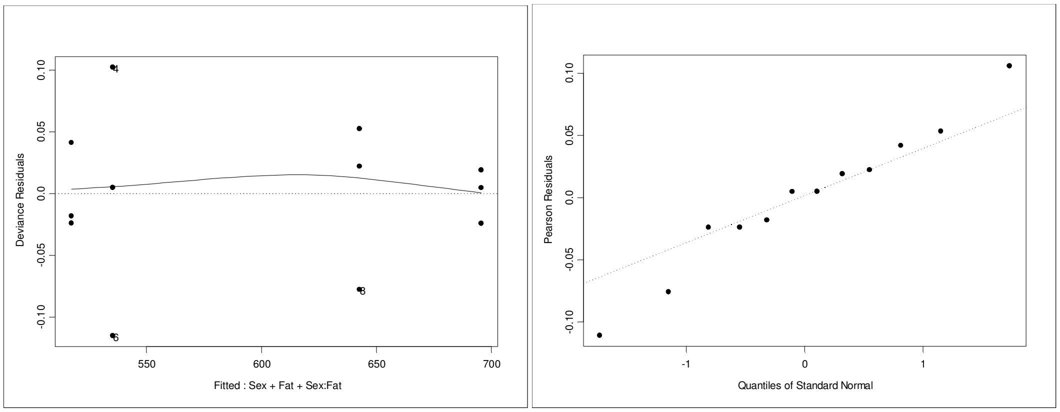


		Value	Std. Error	t value	p-value
(Intercept)		695.6667	22.04793	31.5524742	
Sex	8	-53	31.18048	-1.6997815	0.127595055
Fat	8	-160.3333	31.18048	-5.1421063	0.000882843
Sex:Fat	8	35	44.09586	0.7937254	0.450254579

Dispersion Parameter for Quasi-likelihood family taken to be 1458.333.

Null Deviance: 77570.25 on 11 df, Residual Deviance: 11666.67 on 8 df

Revised model: e^{μ} + gamma error (identity link) $\mu = ß_0 + ß_{sex} + ß_{fat} + ß_{sex*fat}$



	Df	Value	Std. Error	t value	p-value
(Intercept)		6.54487062	0.038844	168.493287	
Sex	8	-0.07924443	0.054933	-1.442565	0.187126202
Fat	8	-0.26198101	0.054933	-4.769099	0.00141024
Sex:Fat	8	0.04504224	0.077687	0.579791	0.578014315

Dispersion Parameter for Gamma family taken to be 0.0045265

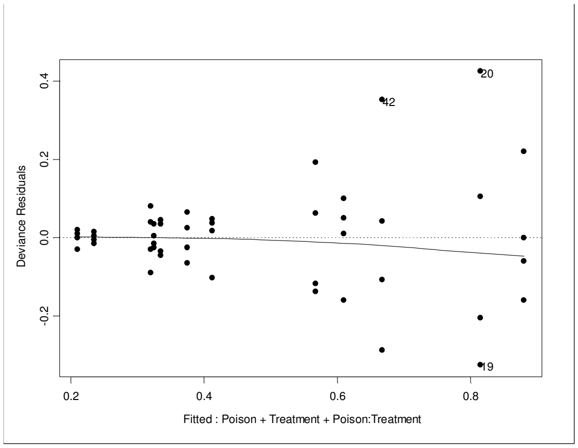
Null Deviance: 0.2203415 on 11 df, Residual Deviance: 0.0366419 on 8 df

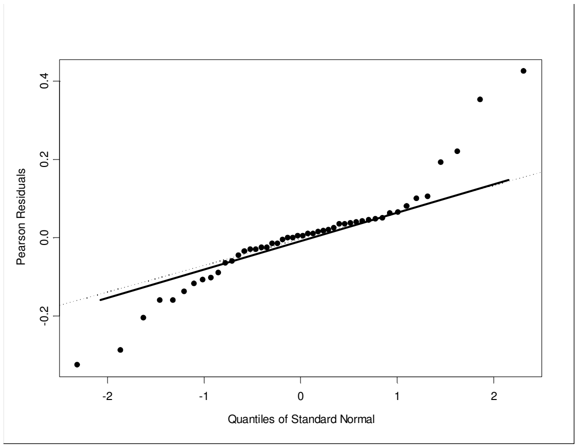
Summary: GLM assumptions not met since residuals are slightly heterogeneous and non-normal. Model revision resulted in no improvement in the residuals, however, it did make the null deviance acceptable.

<u>Poison:</u> The effect of poison (3 types) and Treatment (4 types) on survival (units of 10 hours)

Data from Box, Hunter, and Hunter, 1978 *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley and Sons.

Model: Survival = μ + normal error $\mu = ß_0 + ß_{poison} + ß_{treatment} + ß_{poison*treatment}$



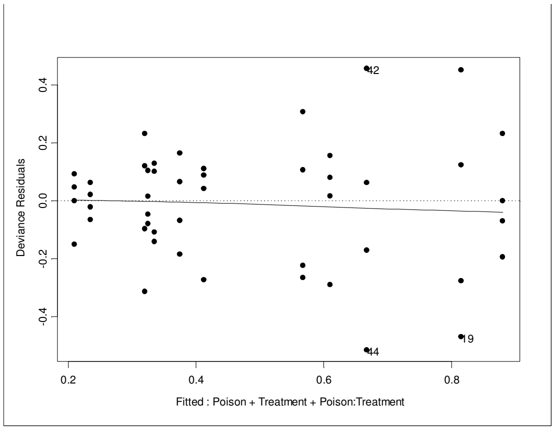


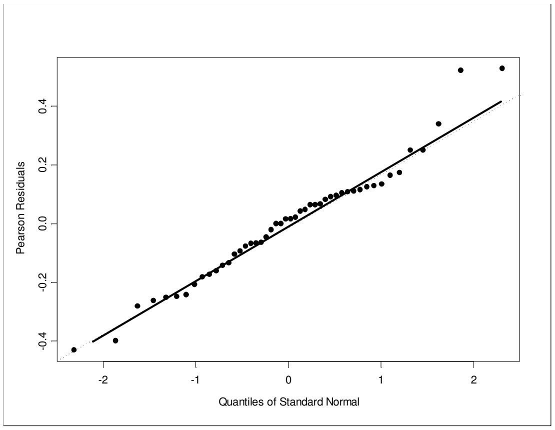
Value		Std.	Error	t-value	p-value
(Intercept)		0.4125	0.074569	5.531762	
Poison2	36	-0.0925	0.105457	-0.87713	0.3862
Poison3	36	-0.2025	0.105457	-1.92021	0.0628
Treatmentb	36	0.4675	0.105457	4.433086	0.0001
Treatmentc	36	0.155	0.105457	1.469793	0.1503
Treatmentd	36	0.1975	0.105457	1.872801	0.0692
Poison2Treatmentb	36	0.0275	0.149139	0.184392	0.8547
Poison3Treatmentb	36	-0.3425	0.149139	-2.29652	0.0276
Poison2Treatmentc	36	-0.1	0.149139	-0.67052	0.5068
Poison3Treatmentc	36	-0.13	0.149139	-0.87167	0.3892
Poison2Treatmentd	36	0.15	0.149139	1.005775	0.3212
Poison3Treatmentd	36	-0.0825	0.149139	-0.55318	0.5836

Dispersion Parameter for Gaussian family taken to be 0.0222424

Null Deviance: 3.005081 on 47 df, Residual Deviance: 0.800725 on 36 df.

Revised model: e^{μ} + gamma error (identity link)$\mu = ß_0 + ß_{poison} + ß_{treatment} + ß_{poison*treatment}$





Value		Std.	Error	t-value	p-value
(Intercept)		0.4125	0.047076	8.762398	
Poison2	36	-0.0925	0.059581	-1.55252	0.1293
Poison3	36	-0.2025	0.052826	-3.83337	0.0005
Treatmentb	36	0.4675	0.110915	4.214933	0.0002
Treatmentc	36	0.155	0.080067	1.93588	0.0608
Treatmentd	36	0.1975	0.084039	2.350108	0.0244
Poison2Treatmentb	36	0.0275	0.149288	0.184208	0.8549
Poison3Treatmentb	36	-0.3425	0.119742	-2.86031	0.0070
Poison2Treatmentc	36	-0.1	0.097857	-1.0219	0.3136
Poison3Treatmentc	36	-0.13	0.087774	-1.48107	0.1473
Poison2Treatmentd	36	0.15	0.119161	1.258806	0.2162
Poison3Treatmentd	36	-0.0825	0.094935	-0.86902	0.3906

Dispersion Parameter for Gamma family taken to be 0.0520972.

Null Deviance: 11.571 on 47 df, Residual Deviance: 1.920462 on 36 df.

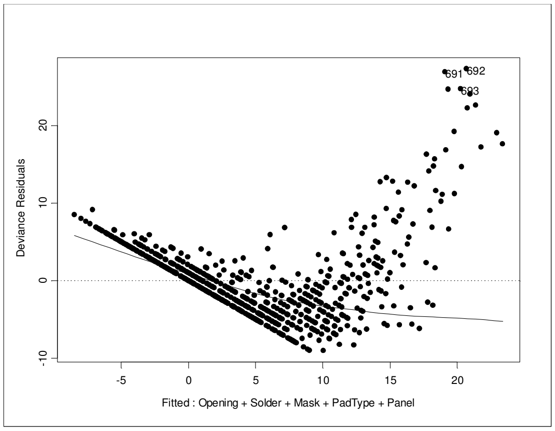
Summary: GLM assumptions not met since residuals are extremely heterogeneous and non-normal.

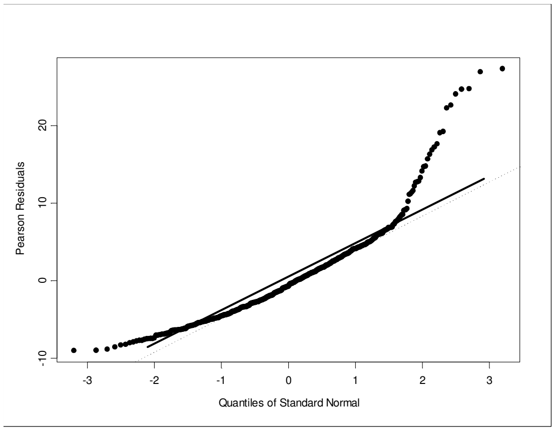
Best revised model (most homogenous errors) was Gamma error with identity link.

The conclusions based on the revised model differed for Posion3 and Treatmentd, both of which are significant in the GzLM and were not significant in the GLM.

<u>Solder Skips:</u> Does the number of solder skips on a circuit board depend on opening (3 categories), amount of solder (2 categories), Mask (4 categories), pad types (10 categories), and /or panel (3 categories). Data taken from Comizzoli R.B., Landwehr J.M., and Sinclair J.D. (1990). Robust materials and processes: Key to reliability. *AT&T Technical Journal,* 69(6): 113--128.

Model: Skips = μ + normal error $\mu = ß_0 + ß_{opening} + ß_{solder} + ß_{mask} + ß_{padtype} + ß_{panel}$





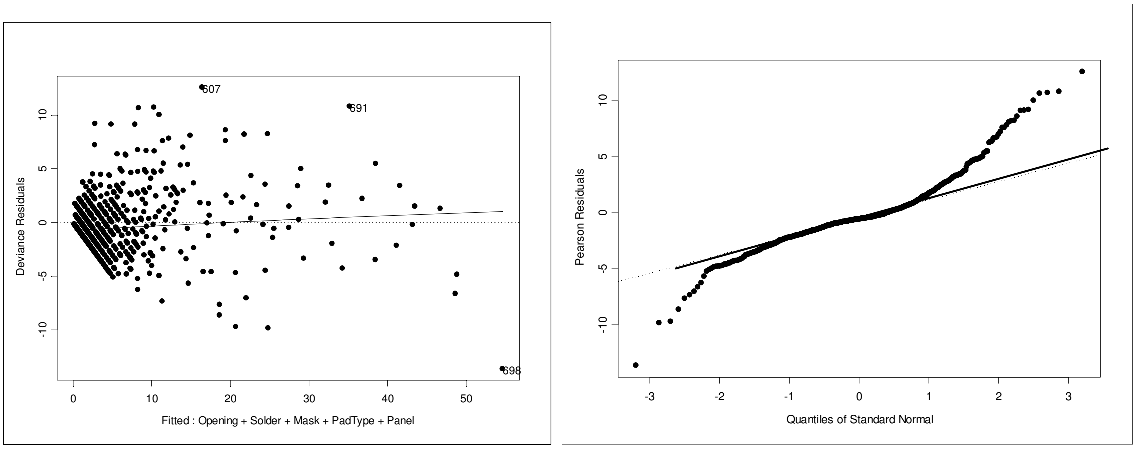
Value	Df	Std.	Error	t-value	p-value
(Intercept)		1.694444	0.747131	2.267935	
Opening.L	702	-6.64975	0.334127	-19.9019	0.0000
Opening.Q	702	3.437791	0.334127	10.28887	0.0000
Solder	702	-3.51393	0.272814	-12.8803	0.0000
MaskA3	702	0.861111	0.545627	1.578204	0.1150
MaskB3	702	3.75	0.545627	6.872822	0.0000
MaskB6	702	8.805556	0.545627	16.1384	0.0000
PadTypeD4	702	0.694444	0.862713	0.804955	0.4211
PadTypeL4	702	2.694444	0.862713	3.123224	0.0019
PadTypeD6	702	-1.36111	0.862713	-1.57771	0.1151
PadTypeL6	702	-2.55556	0.862713	-2.96223	0.0032
PadTypeD7	702	0.069444	0.862713	0.080495	0.9359
PadTypeL7	702	-1.88889	0.862713	-2.18948	0.0289

PadTypeL8	702	-0.88889	0.862713	-1.03034	0.3032
PadTypeW9	702	-4.38889	0.862713	-5.08731	0.0000
PadTypeL9	702	-2.44444	0.862713	-2.83344	0.0047
Panel2	702	1.6	0.472527	3.386049	0.0007
Panel3	702	1.170833	0.472527	2.477812	0.0135

Dispersion Parameter for Gaussian family taken to be 26.79383

Null Deviance: 48116.13 on 719 df, Residual Deviance: 18809.27 on 702 df.

Revised model: e^{μ} + quasi error (log link) $\mu = ß_{0} + ß_{opening} + ß_{solder} + ß_{mask} + ß_{padtype} + ß_{panel}$



Value	Df	Std.	Error	t-value	p-value
(Intercept)		0.375602	0.081303	4.619774	
Opening.L	702	-1.17627	0.043393	-27.1075	0.0000
Opening.Q	702	0.593093	0.050691	11.70016	0.0000
Solder	702	-0.64979	0.022623	-28.7223	0.0000
MaskA3	702	0.352193	0.084686	4.158797	0.0000
MaskB3	702	1.12062	0.072921	15.36768	0.0000
MaskB6	702	1.651356	0.070579	23.39735	0.0000
PadTypeD4	702	0.044798	0.038758	1.155851	0.2481
PadTypeL4	702	0.156692	0.03683	4.25449	0.0000

PadTypeD6	702	-0.19302	0.044128	-4.37416	0.0000
PadTypeL6	702	-0.69251	0.06258	-11.066	0.0000
PadTypeD7	702	-0.07787	0.041252	-1.8877	0.0595
PadTypeL7	702	-0.52984	0.055196	-9.59926	0.0000
PadTypeL8	702	-0.34838	0.048539	-7.17719	0.0000
PadTypeW9	702	-1.20488	0.097634	-12.3408	0.0000
PadTypeL9	702	-0.63478	0.059758	-10.6224	0.0000
Panel2	702	0.283596	0.028144	10.07647	0.0000
Panel3	702	0.167168	0.02944	5.678173	0.0000

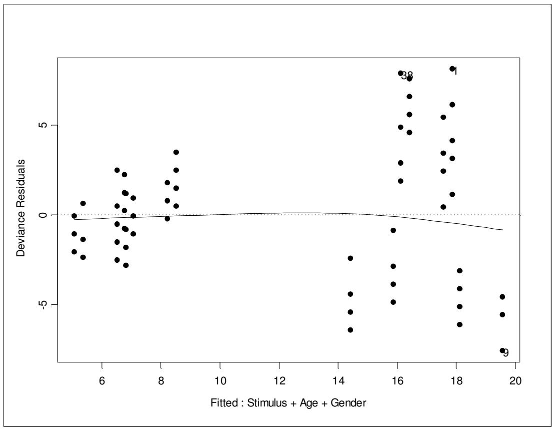
Dispersion Parameter for Quasi-likelihood family taken to be 7.36591

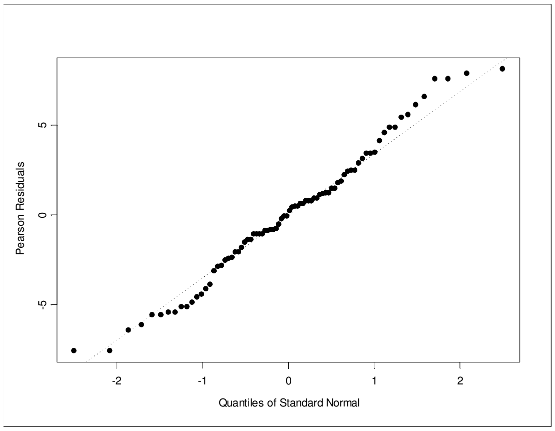
Null Deviance: 48116.13 on 719 df, Residual Deviance: 5170.836 on 702 df.

Summary: GLM assumptions not met since residuals are extremely heterogeneous (obvious bowl) and non-normal. Best revised model (most homogenous errors) was quasi error with log link. The conclusions based on the revised model differed for MaskA3 and PadTypeL8, both of which are significant in the GzLM and were not significant in the GLM.

Attention span: The affect of Stimulus (4 categories), Age (2 categories), and Gender (2categories) on looking time. Data from Pg 147 in Bogartz, 1994. An Introduction to the Analysis of Variance. Praeger Publisher, 565 Pp.

Model: Look = µ + normal error $\mu = ß_0 + ß_{\text{Stimulus}} + ß_{\text{Age}} + ß_{\text{Gender}}$



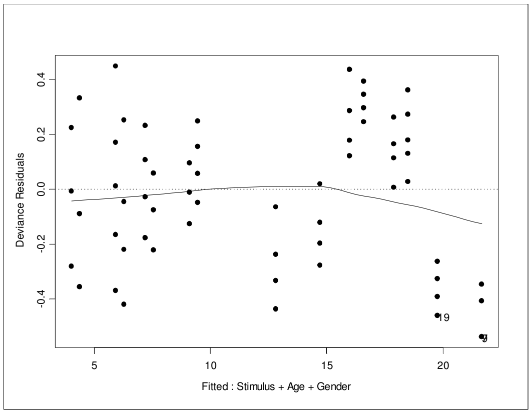


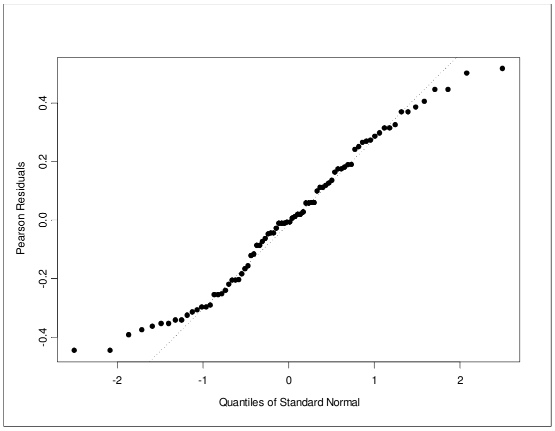
	Df	Value	Std. Error	t value	p-value
(Intercept)		16.175	1.61408	10.02119	
StimulusUnscrDia	74	-2	1.22013	-1.63917	0.1054
StimulusScraPic	74	-11.35	1.22013	-9.30229	0.0000
StimulusScraDia	74	-11.05	1.22013	-9.05642	0.0000
Age	74	0.566667	0.287587	1.970416	0.0525
Gender	74	-1.45	0.862762	-1.68065	0.0970

Dispersion Parameter for Gaussian family taken to be 14.88716

Null Deviance: 3323.2 on 79 df, Residual Deviance: 1101.65 on 74 df.

Revised model: e^{μ} + gamma error (identity link) $\mu = ß_{0} + ß_{Stimulus} + ß_{Age} + ß_{Gender}$





	Df	Value	Std. Error	t value	p-value
(Intercept)		15.32131	1.358386	11.27906	
StimulusUnscrDia	74	-3.77904	1.408532	-2.68296	0.0090
StimulusScraPic	74	-12.5633	1.156602	-10.8622	0.0000
StimulusScraDia	74	-12.2111	1.162961	-10.5	0.0000
Age	74	1.057728	0.172236	6.141166	0.0000
Gender	74	-1.90249	0.485329	-3.91999	0.0002

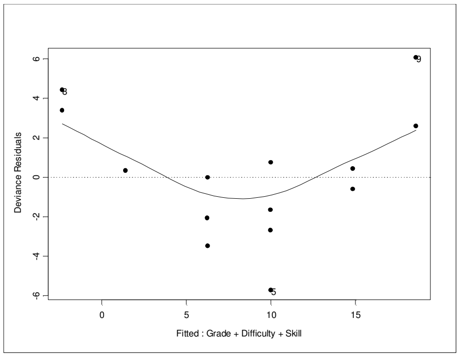
Dispersion Parameter for Gamma family taken to be 0.0681966

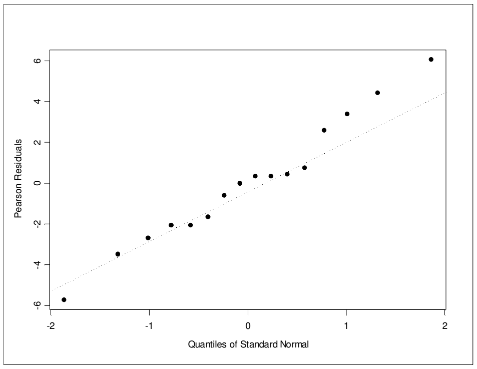
Null Deviance: 25.31959 on 79 df, Residual Deviance: 5.29482 on 74 df.

Summary: GLM assumptions not met since residuals are extremely heterogeneous (cone) and non-normal. Best revised model (most homogenous errors) was gamma error with identity link. The conclusions based on the revised model differed for Stimulus unscrambled diagram, Age and Gender, all of which are significant in the GzLM and were not significant in the GLM.

<u>Reading errors:</u> The affect of Grade (2 categories), Difficulty (2 categories), and Skill Level (2 categories) on mean percentage of reading errors. Data from Bowey, 1985. Contextual facilitation in children's oral reading in relation to grade and decoding skill. *J. of Exp. Chil Psych, 40*, pp 23-48.

Model: Look = μ + normal error $\mu = ß_0 + ß_{Grade} + ß_{Difficulty} + ß_{Skill}$



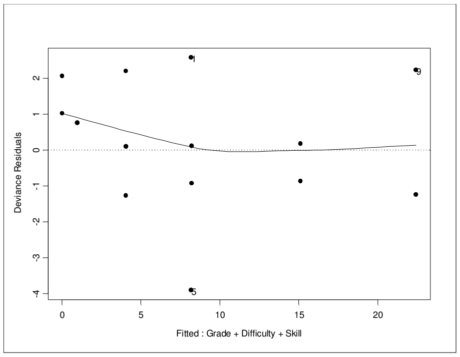


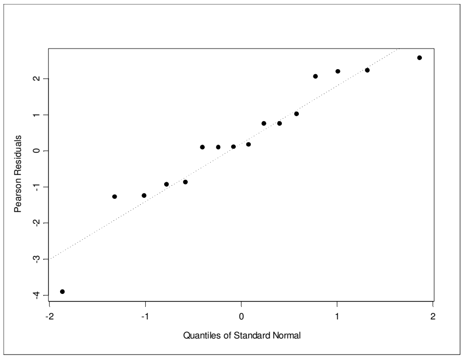
	Df	Value	Std. Error	t value	p-value
(Intercept)		24.99	7.806223	3.201292	
Grade	12	-3.745	1.703458	-2.19847	0.048269129
Difficulty	12	8.58	1.703458	5.036815	0.000290847
Skill	12	-8.6075	1.703458	-5.05296	0.000283119

Dispersion Parameter for Gaussian family taken to be 11.60707

Null Deviance: 786.2068 on 15 df, Residual Deviance: 139.2849 on 12 df.

Revised model: e^{μ} + quasi error (sqrt link) $\mu = ß_0 + ß_{Grade} + ß_{Difficulty} + ß_{Skill}$





	Df	Value	Std. Error	t value	p-value
(Intercept)		6.254751	0.778935	8.029874	
Grade	12	-0.84835	0.175402	-4.83661	0.0004
Difficulty	12	1.873926	0.210332	8.909361	0.0000
Skill	12	-1.86849	0.210024	-8.89653	0.0000

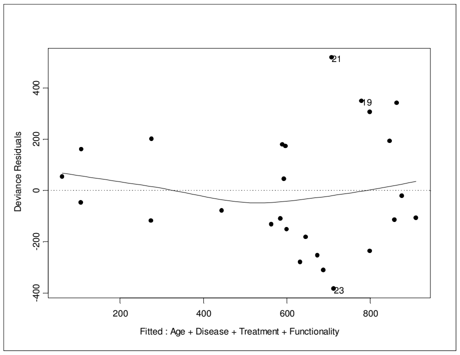
Dispersion Parameter for Quasi-likelihood family taken to be 3.577677

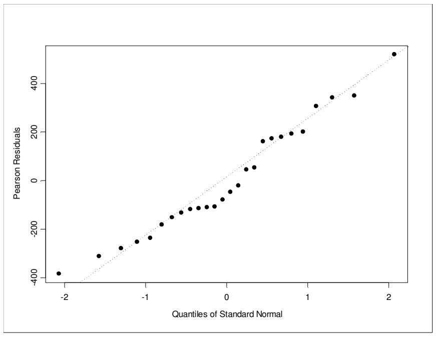
Null Deviance: 786.2068 on 15 df, Residual Deviance: 42.93206 on 12 df

Summary: GLM assumptions not met since residuals are heterogeneous (bowl) and non-normal. Best revised model (most homogenous errors) was quasi error with sqrt link. The conclusions for the revised model don't differ from those of the original model, however, the Std. Error is much more reasonable with the quasi than the normal error.

Ovarian cancer: The affect of Age (Years, ratio), Extent of Disease (2 categories), Treatment (2 categories), and Functional Status (2 categories) on Survival in days in patients with ovarian cancer. Data from pg 275, S-PLUS® 8 Guide to Statistics, Volume 2, Insightful Corporation, Seattle, WA.

Model: Survival = μ + normal error $\mu = ß_0 + ß_{Age} + ß_{Disease} + ß_{Treatment} + ß_{Functionality}$



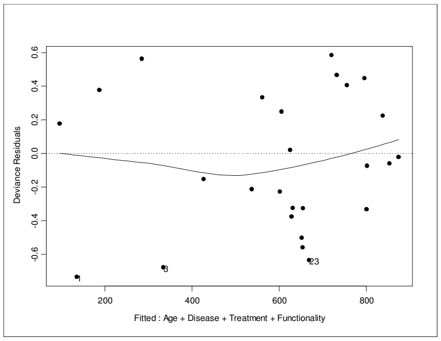


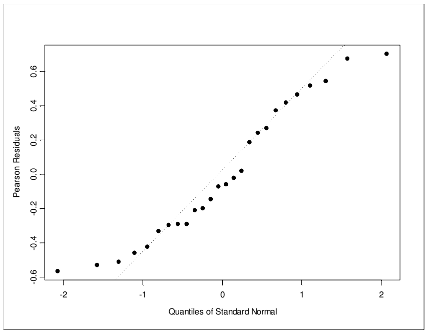
	Df	Value	Std. Error	t value	p-vale
(Intercept)		1623.42	355.6574	4.564561	
Age	21	-20.7742	5.350801	-3.88244	0.0009
Disease	21	-117.13	107.7199	-1.08736	0.2892
Treatment	21	173.4456	99.9558	1.735223	0.0974
Functionality	21	46.14655	102.4559	0.450404	0.6570

Dispersion Parameter for Gaussian family taken to be 64195.49

Null Deviance: 2884706 on 25 df, Residual Deviance: 1348105 on 21 df.

Revised model: e^{μ} + gamma error (identity link)$\mu = ß_0 + ß_{Age} + ß_{Disease} + ß_{Treatment} + ß_{Functionality}$





	Df	Value	Std. Error	t value	p-value
(Intercept)		1445.644	374.9647	3.855413	
Age	21	-18.2493	4.320428	-4.22395	0.0004
Disease	21	-113.626	121.3887	-0.93605	0.3599
Treatment	21	146.4784	97.16505	1.507521	0.1466
Functionality	21	91.41124	67.32944	1.357671	0.1890

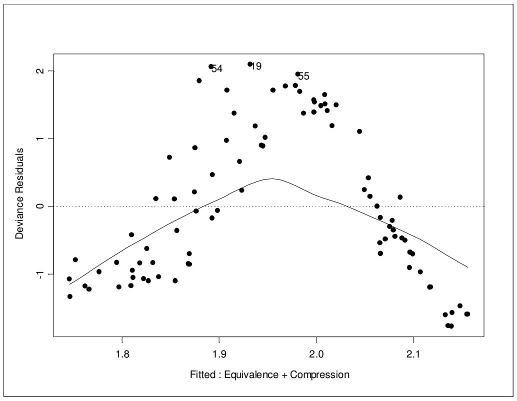
Dispersion Parameter for Gamma family taken to be 0.1876016

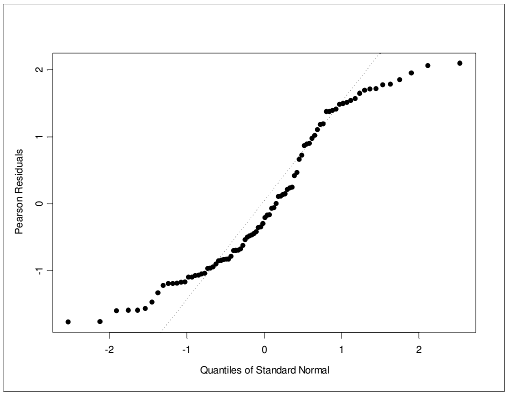
Null Deviance: 10.66275 on 25 df, Residual Deviance: 4.198008 on 21 df

Summary: GLM assumptions not met since residuals are heterogeneous (cone) and non-normal. Best revised model (most homogenous errors) was gamma error with identity link. The conclusions for the revised model don't differ from those of the original model. The Std. Error is more reasonable in some cases with the normal error structure and and in other cases with the gamma error structure.

NOx emmisions: The affects of the compression ratio, and equivalence ratio on nitric oxide emissions. Data from S-PLUS® 8 Guide to Statistics, Volume 2, Insightful Corporation, Seattle, WA.

Model: Survival = μ + normal error μ = ß$_0$ + ß$_{\text{Equivalence}}$ + ß$_{\text{Compression}}$



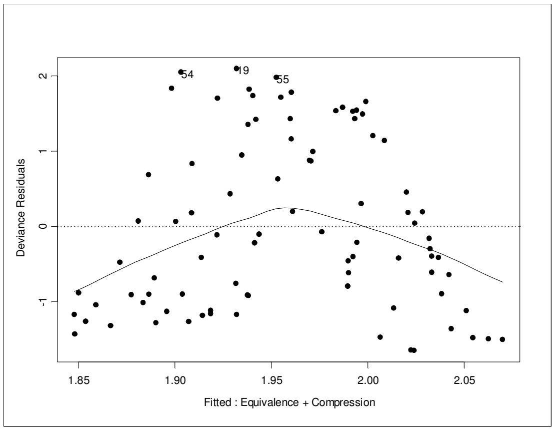


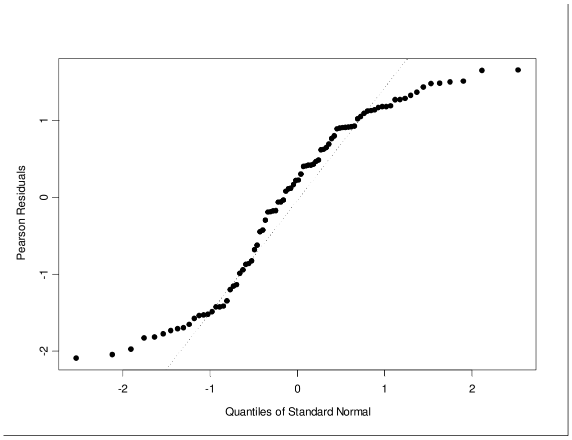
	Df	Value	Std. Error	t value	p-value
(Intercept)		2.559101	0.662396	3.863399	
Equivalence	85	-0.55714	0.601464	-0.9263	0.3569
Compression	85	-0.00711	0.031135	-0.22833	0.8199

Dispersion Parameter for Gaussian family taken to be 1.298838

Null Deviance: 111.6238 on 87 df, Residual Deviance: 110.4012 on 85 df.

Revised model: e$^{\mu}$ + quasi error (1/mu^2 link) μ = ß$_0$ + ß$_{\text{Equivalence}}$ + ß$_{\text{Compression}}$





	Df	Value	Std. Error	t value	p-value
(Intercept)		0.178158	0.17419	1.022782	
Equivalence	85	0.063975	0.160382	0.398892	0.6910

Compression	85	0.001999	0.008379	0.238579	0.8120

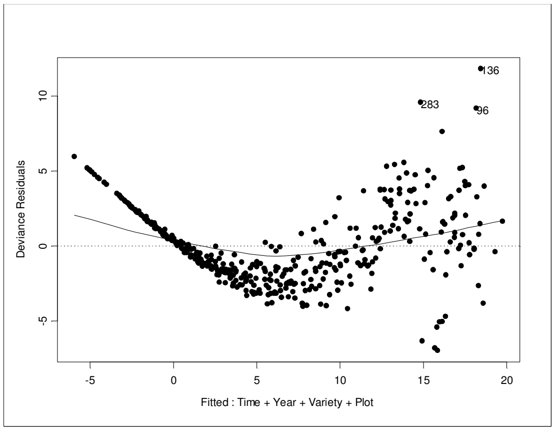
Dispersion Parameter for Quasi-likelihood family taken to be 1.306547

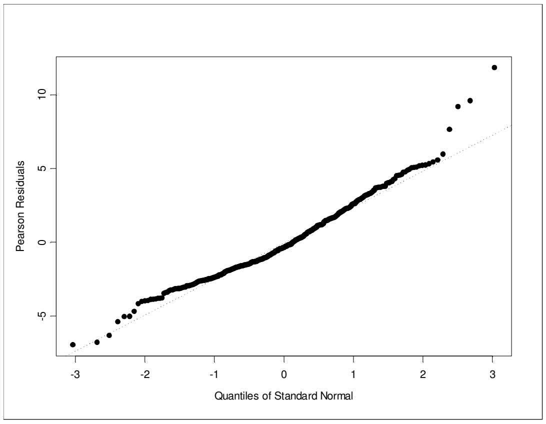
Null Deviance: 111.6238 on 87 df, Residual Deviance: 111.0569 on 85 df.

Summary: GLM assumptions not met since residuals are heterogeneous (obvious arch) and non-normal. Best revised model (most homogenous errors) was quasi error with 1/mu^2 link. The conclusions for the revised model don't differ from those of the original model, however, the Std. Error is much more reasonable with the quasi than the normal error.

<u>Soy bean:</u> The affect of Plot (16 plots), Year (3 years), Variety (2 categories) and Time on leaf Weight in soybean plants. Data from Davidian & Giltinan, 1995. *Nonlinear Models for Repeated Measurement Data.* London: Chapman & Hall.

Model: Weight = μ + normal error $\mu = ß_{0} + ß_{Plot} + ß_{Year} + ß_{Variety} + ß_{Time}$





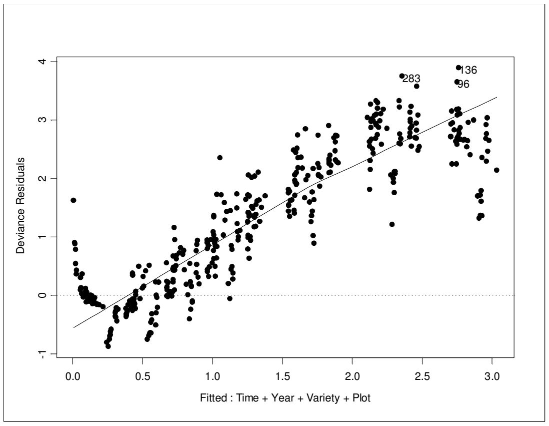
		Value	Std.Error	t-value	p-value
(Intercept)		-6.37981	11.20634	-0.5693	
Time	363	0.298113	0.006354	46.91587	0.0000
Year1989	363	0.005789	14.21442	0.000407	0.9997
Year1990	363	-6.49345	17.50497	-0.37095	0.7109
Variety	363	3.651714	4.284626	0.852283	0.3946
Plot	363	0.173277	0.056091	3.089187	0.0022

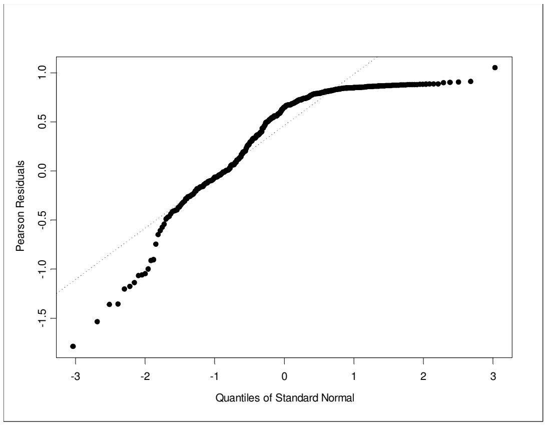
Note: Only one Plot estimate shown to save space.

Dispersion Parameter for Gaussian family taken to be 7.300964

Null Deviance: 19650.14 on 411 df, Residual Deviance: 2650.25 on 363 df.

Revised model: e^{μ} + gamma error (identity link) $\mu = ß_0 + ß_{Plot} + ß_{Year} + ß_{Variety} + ß_{Time}$





		Value	Std.Error	t-value	p-value
(Intercept)		-0.4755	0.078199	-6.08064	
Time	363	0.041427	0.002512	16.49286	0.0000
Year1989	363	-0.12251	0.166198	-0.73711	0.4615
Year1990	363	-0.02402	0.100973	-0.23792	0.8121
Variety	363	0.09687	0.14177	0.683294	0.4949
Plot	363	0.000175	0.00268	0.065464	0.9478

Note: Only one Plot estimate shown to save space.

Dispersion Parameter for Gamma family taken to be 0.4862491

Null Deviance: 860.934 on 411 df, Residual Deviance: NA on 363 df

Summary: GLM assumptions not met since residuals are heterogeneous (obvious bowl) and non-normal. Best revised model (most homogenous errors) was gamma error with identity link. The conclusions based on the revised model differed for some of the Plots which are not significant in the GzLM and were significant in the GLM. Additionally an error occurred when interpreting the Residual Deviance in the GzLM.

The Following Datasets were obtained from: Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series.

Website: www.ilr.cornell.edu/~hadi/RABE4/#

Milk Dataset

The current month's milk production in pounds (currentm) in cows is analyzed in relation to the previous month's milk production (previous), percent fat in the milk (fat), percent protein in the milk (protein), number of days since present lactation (days), frequency of lactation (lactatio) and an indicator variable (i79) recorded as zero if Days<79, recorded as 1 if Days>79.

Model: currentm=μ + normal error

$\mu=\beta_o + \beta_{previous}$*Previous + β_{fat}*fat +$\beta_{protein}$*Protein +β_{days}*Days+ $\beta_{lactatio}$*lactatio +β_{i179}*i79

currentm	previous	fat	protein	days	lactatio	i79
45	45	5.5	8.9	21	5	0
86	86	4.4	4.1	25	4	0
50	50	6.5	4	25	7	0
42	42	7.4	4.1	25	2	0
61	61	3.8	3.8	33	2	0
93	93	4.2	3	45	3	0

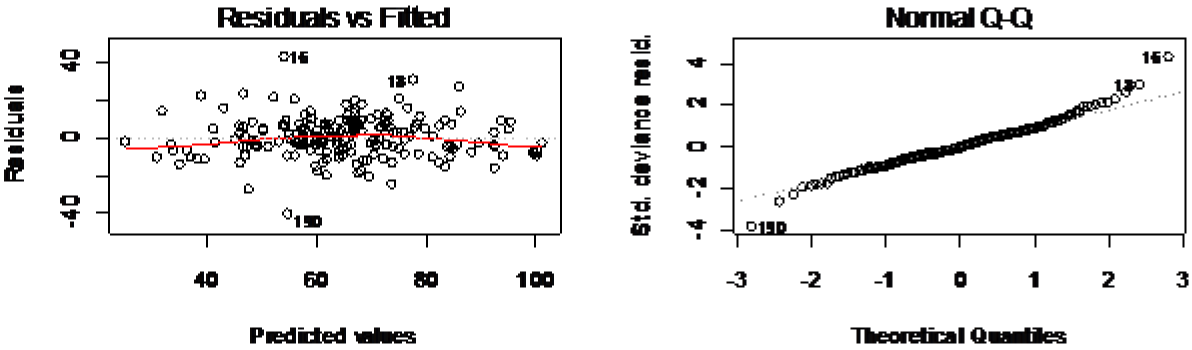
Note: Dataset quite large see website for full dataset

Coefficients:		Estimate Std.	Error t value	Pr(>|t|)
(Intercept)	51.0304	7.30322	6.987	4.48e-11 ***
DAYS	-0.0317	0.01548	-2.045	0.042231 *
FAT	0.79789	0.95244	0.838	0.403224
I79	-10.298	2.84711	-3.617	0.000381 ***
LACTATIO	0.52677	0.54881	0.96	0.338347
PREVIOUS	0.69846	0.05203	13.423	< 2e-16 ***
PROTEIN	-6.4156	1.6092	-3.987	9.51e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 113.9273)

Residual deviance: 21874 on 192 degrees of freedom



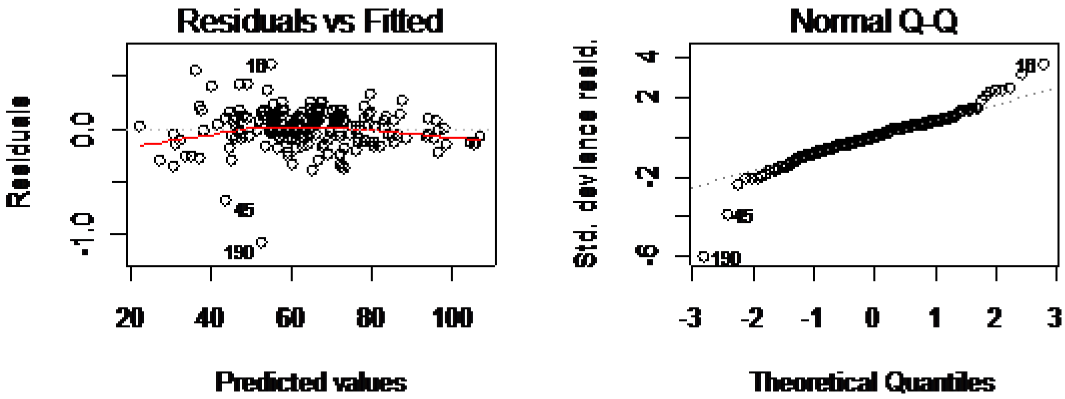
Revised Model: currentm= e^{μ} + gamma error (identity link used)

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.36587	6.89669	6.723	1.98e-10 ***
DAYS	-0.0289	0.01499	-1.929	0.0552
FAT	1.49947	0.96322	1.557	0.1212
I79	-15.1899	3.42533	-4.435	1.55e-05 ***
LACTATIO	0.29157	0.56072	0.52	0.6037
PREVIOUS	0.77441	0.04752	16.296	< 2e-16 ***
PROTEIN	-5.86321	1.26802	-4.624	6.90e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for Gamma family taken to be 0.03302502)

Residual deviance: 6.9574 on 192 degrees of freedom



Summary: GLM (general linear model) assumptions do not appear to be violated when inspected graphically. However the dispersion parameter was quite large. The revised model greatly decreased the dispersion parameter but had little impact on the appearance of residual plots. Parameter estimates decreased overall in the revised model (can be compared because identity link was used). Gamma error decreased the probability of Type I Error since p-values decreased.

Work Law Dataset

Analysis of cost of living for a four person family (COL) in the US in relation to Population density (person per square mile, PD), State unionization rate in 1978 (URate), Population in 1975 (Pop), Property taxes in 1972 (Taxes), Per capita income in 1974 (Income) and Right to Work Law (RTWL) an indicator variable (1 if there is right-to-work laws in state and 0 otherwise).

Model: COL= μ + normal error

$\mu=\beta_o+\beta_{PD}*PD+\beta_{URate}*URate+\beta_{Pop}*Pop+\beta_{taxes}*Taxes+_{Incoome}*Income*\beta_{RTWL}*RTWL$ + normal error

COL	PD	Urate	Pop	Taxes	Income	RTWL
169	414	13.6	1790128	5128	2961	1
143	239	11	396891	4303	1711	1
339	43	23.7	349874	4166	2122	0
173	951	21	2147850	5001	4654	0
99	255	16	411725	3965	1620	1
363	1257	24.4	3914071	4928	5634	0

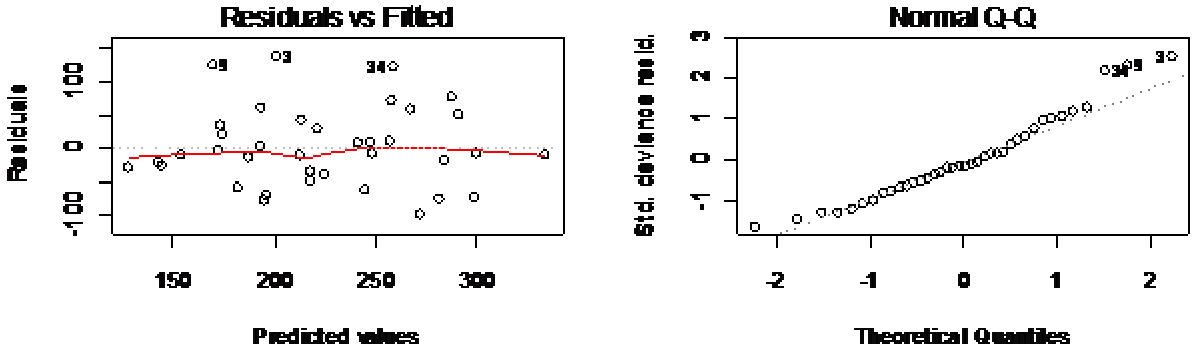
Note: Dataset quite large see website for full dataset

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.89E+02	1.62E+02	1.782	0.0846
INCOME	9.74E-03	6.39E-03	1.524	0.1376
PD	1.11E-02	1.77E-02	0.628	0.5344
POP	1.16E-05	1.03E-05	1.12	0.2713
RTWL	-1.10E+02	4.27E+01	-2.573	0.0151 *
TAXES	1.52E-03	2.76E-02	0.055	0.9566
URATE	-5.05E+00	2.38E+00	-2.12	0.0421 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 3976.749)

Residual deviance: 123279 on 31 degrees of freedom



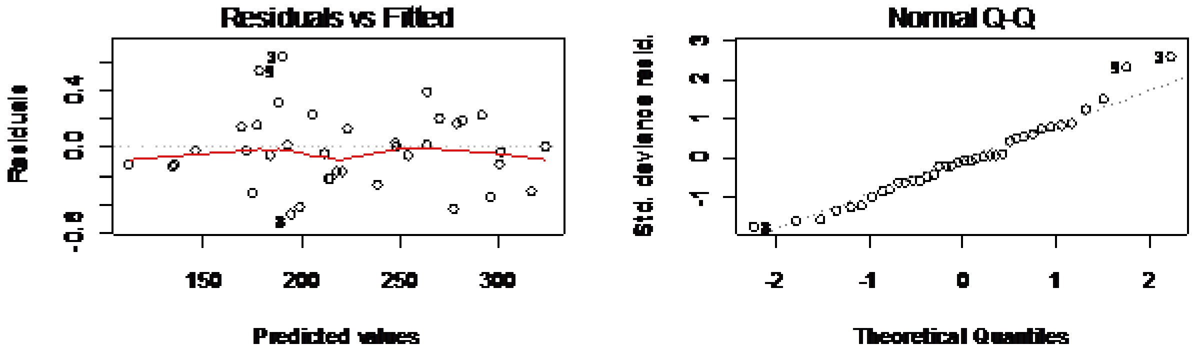
Revised Model: COL=e^{μ} + gamma error (identity link used)

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.55E+02	1.63E+02	1.569	0.12682
INCOME	1.30E-02	6.35E-03	2.053	0.04863 *
PD	1.15E-02	2.29E-02	0.501	0.62006
POP	1.07E-05	1.21E-05	0.888	0.38157
RTWL	-1.17E+02	4.19E+01	-2.799	0.00873 **
TAXES	1.01E-02	2.87E-02	0.352	0.72693
URATE	-5.84E+00	2.51E+00	-2.327	0.02666 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for Gamma family taken to be 0.08387292)

Residual deviance: 2.4404 on 31 degrees of freedom



Summary: GLM (general linear model) assumptions do not appear to be violated when inspected graphically. However the dispersion parameter was quite large. The revised model greatly decreased the dispersion parameter (from 3976 to 0.083) but had little impact on the appearance of residual plots. Parameter estimates were increased in the revised model (comparable since identity was used). In this case Type I Error was increased since p-values in the revised model decreased, giving stronger evidence to reject the null hypothesis.

Skulls Dataset

How to estimate the age of historical objects based on some age-related characteristic of the objects?

Approximate Year of Skull Formation (Year) in relation to maximum breath of skull (MB), basibregmatic Height of skull (BH), Nasal Height of Skull (NH).

Year coded as positive if A.D., negative if B.C.

Model: Year=μ + normal error

$\mu=\beta_o+\beta_{MB}*MB+\beta_{BH}*BH+\beta_{NH}*NH$

Year	MB	BH	BL	NH
-200	135	130	100	51
150	137	123	91	50
150	136	131	95	49

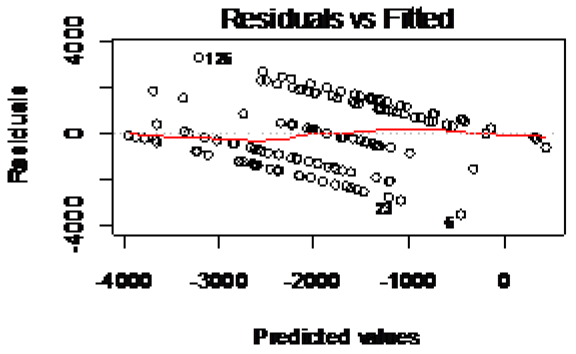
Note: Dataset quite large see website for full dataset

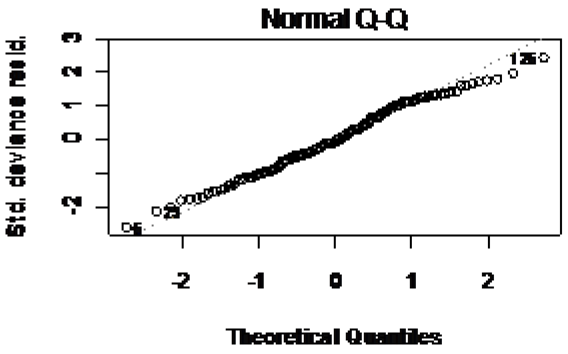
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3687.4	4946.86	-0.745	0.457235
BH	-29.38	24.4	-1.204	0.230384
BL	-109.03	22.35	-4.877	2.8e-06 ***
MB	96.4	24.19	3.986	0.000106 ***
NH	65.64	36.85	1.782	0.076918

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 1959493)

Residual deviance: 284126532 on 145 degrees of freedom





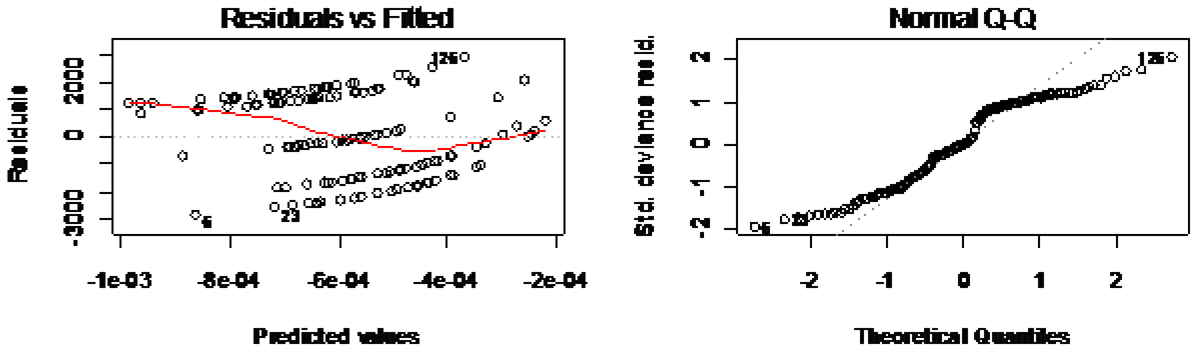
Revised Model: Year=µ+ normal error (with an inverse link)

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.02E-04	9.58E-04	-0.106	0.9156
BH	3.34E-06	5.14E-06	0.649	0.5171
BL	2.03E-05	4.27E-06	4.74	5.06e-06 ***
MB	-1.53E-05	5.26E-06	-2.917	0.0041 **
NH	-1.61E-05	9.24E-06	-1.743	0.0834

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 2129065)

Residual deviance: 308752383 on 145 degrees of freedom



Summary: The Original model did not violate the assumptions when inspected graphically. However the dispersion parameter was extremely high. The revised model made the residual plot appear worse, less homogeneous and more skewed. As well the dispersion parameter was not improved. Due to the type of response variable (discrete with negative and positive values) error structure choices were limited since several error structures require variables to be bounded above zero. Parameter estimates cannot be compared since the model structure was changed, different link was used. Changing the link increased the probability of Type I Error since p-values decreased.

New York Rivers Dataset

Analysis of mean nitrogen concentration (mg/litre) Nitrogen, in relation to the percentage of land currently in agricultural use (agr), the percentage of forested land (forest), the percentage of land in residential use (rsdntial) and the percentage of land in commercial/industrial use (comindl).

Nitrogen= µ + normal error

$\mu=\beta_o+\beta_{agr}*agr+\beta_{forest}*forest+\beta_{rsdntial}*rsdntial+\beta_{comindl}*comindl$

Nitrogen	agr	forest	rsdntial	comindl
1.1	26	63	1.2	0.29
1.01	29	57	0.7	0.09
1.9	54	26	1.8	0.58
1	2	84	1.9	1.98
1.99	3	27	29.4	3.11

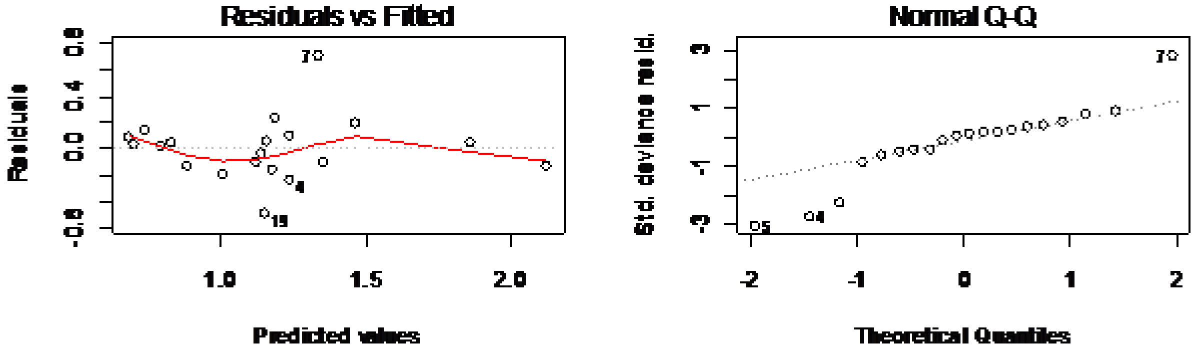
Note: For full dataset see website

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.722214	1.234082	1.396	0.1832
agr	0.005809	0.015034	0.386	0.7046
comindl	0.305028	0.163817	1.862	0.0823
forest	-0.012968	0.013931	-0.931	0.3667
rsdntial	-0.007227	0.03383	-0.214	0.8337

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 0.07018191)

Residual deviance: 1.0527 on 15 degrees of freedom



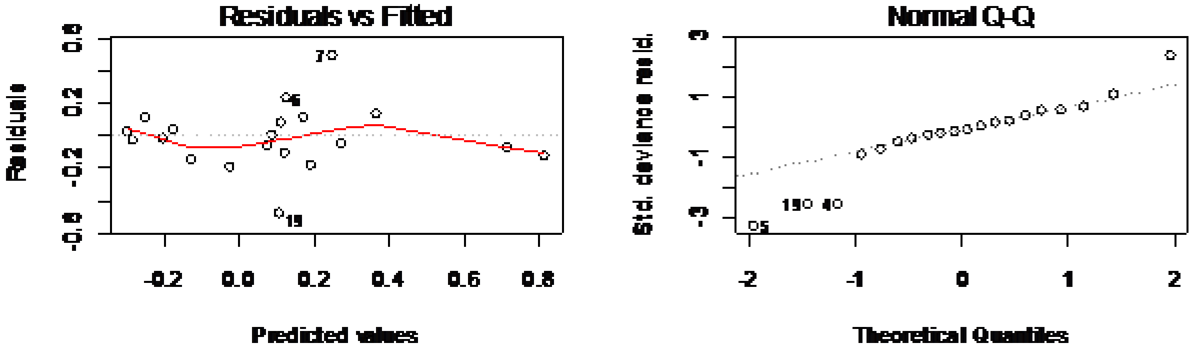
Revised Model: Nitrogen= e^{μ} + gamma error (with log link)

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.601857	1.050364	0.573	0.5751
agr	0.004913	0.012796	0.384	0.7064
comindl	0.274425	0.139429	1.968	0.0678
forest	-0.011187	0.011857	-0.943	0.3604
rsdntial	-0.012023	0.028794	-0.418	0.6822

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for Gamma family taken to be 0.05084133)

Residual deviance: 0.73158 on 15 degrees of freedom



Summary: The original model (GLM) did violate the assumption of normality of residuals. The Q-Q plot showed a heavy skew. The revised model reduced the dispersion parameter and corrected some of the skew. Parameter estimates cannot be compared because of the change in model structure. P-values were decreased and thus the probability of making a Type Error I (rejecting the null hypothesis when it is true) was increased.

Scots Race

Data are from a Scottish hills race, time (in seconds) is analyzed in related to distance in miles and climb in feet.

Model: Time=μ + normal error

$\mu=\beta_o+\beta_{distance}$*distance + β_{climb}*climb

Time	Distance	Climb
965	2.5	650
2901	6	2500
2019	6	900
2736	7.5	800
3736	8	3070
4393	8	2866

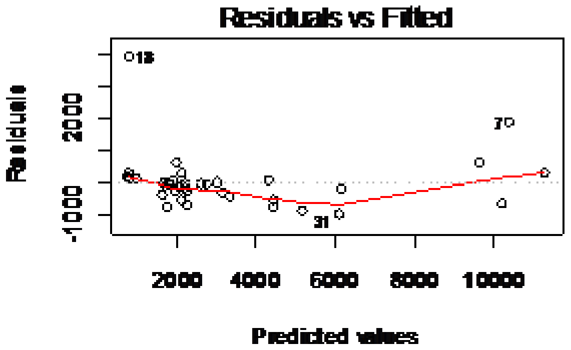
Note: See website for full dataset

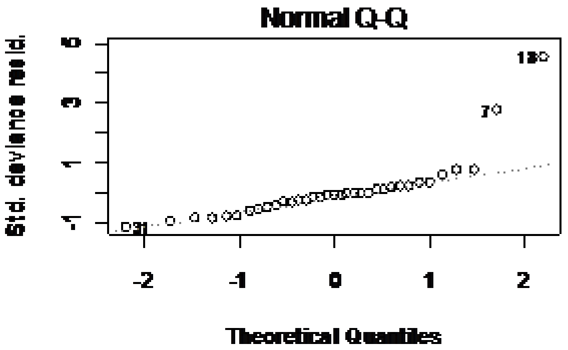
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-539.4829	258.1607	-2.09	0.0447 *
distance	373.0727	36.0684	10.343	9.86e-12 ***
climb	0.6629	0.1231	5.387	6.44e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 775315)

Residual deviance: 24810082 on 32 degrees of freedom





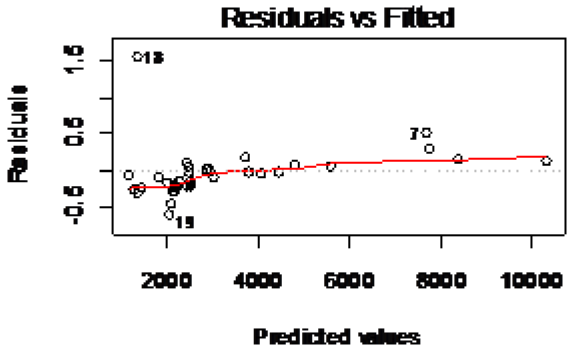
Revised Model: Time=e^{μ} + gamma error (with log link)

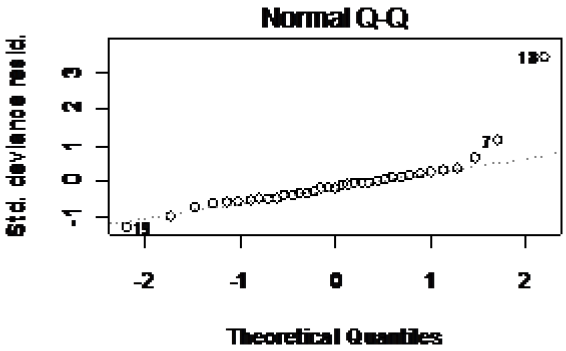
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.10E+00	1.19E-01	59.497	< 2e-16 ***
distance	7.57E-02	1.67E-02	4.54	7.53e-05 ***
climb	1.47E-04	5.69E-05	2.581	0.0147 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for Gamma family taken to be 0.1655957)

Residual deviance: 3.5553 on 32 degrees of freedom





Summary: The original model violated the assumptions since the residuals are heterogeneous and skewed. The revised model did help with the skew and made the residuals more homogeneous. Also the revised model greatly reduced the dispersion factor. Parameter estimates cannot be compared since the model structure was changed. P-values decreased in the revised model which lowered the probability of Type I Error.

Computer Repair Dataset

Analysis of number of computer repaired (units) in relation to time (minutes).

Units=μ + normal error

$\mu=\beta_o+\beta_{minutes}$*Minutes

Minutes	Time
1	23
2	29
3	49
4	64
4	74

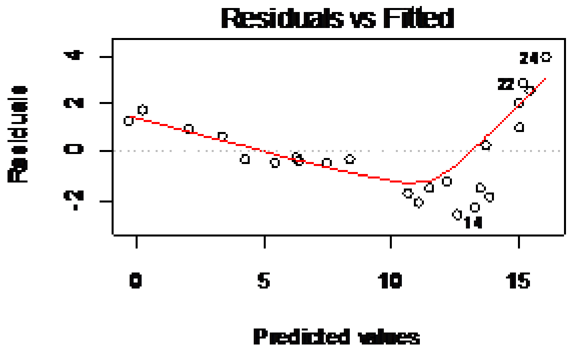
Note: Full dataset can be obtained from website

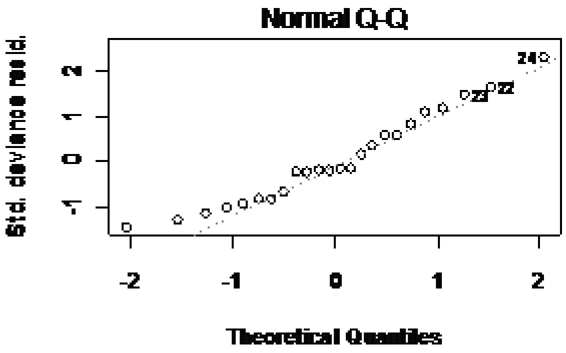
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.342748	0.945277	-2.478	0.0214 *
MINUTES	0.089933	0.006512	13.811	2.56e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 88.63)

Residual deviance: 69.796 on 22 degrees of freedom





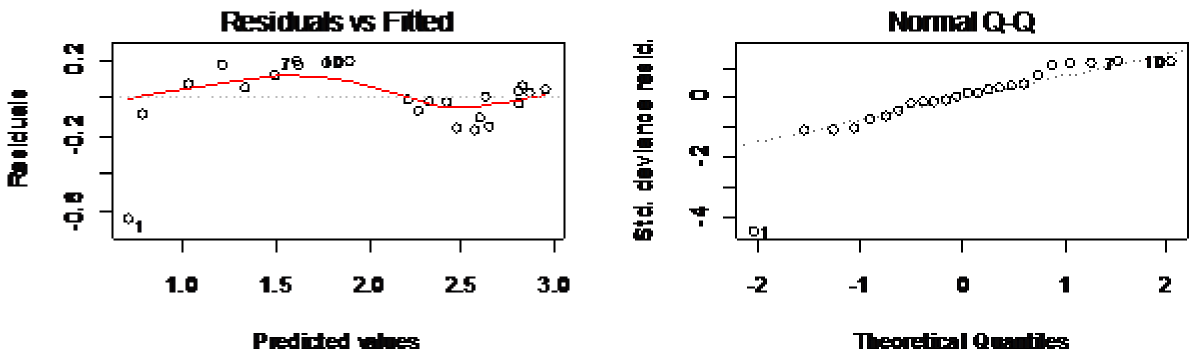
Revised Model: Units=e^{μ} + gamma error (log link)

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4296492	0.0852805	5.038	4.80e-05 ***
MINUTES	0.0123163	0.0005875	20.965	4.96e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for Gamma family taken to be 0.02582205)

Residual deviance: 0.7029 on 22 degrees of freedom



Summary: In the original mode the assumptions were violated since the residuals are strongly heterogeneous. The revised model was able to largely correct this. Also the revised reduced the dispersion factor. Parameter estimates cannot be compared since the model structure was changed. In the revised model the p-value for minutes decreased several orders of magnitude. Thus the revised model increases the chance of Type I Error.

Stock Dataset

Consumer expenditure, (measured in billions of US dollars) is regressed onto stock of money (also measured in billions of US dollars).

Model: Expendit=μ + normal error

$M=\beta_o+\beta_{stock}$*stock

expendit	stock
214.6	159.3
217.7	161.2
219.6	162.8
227.2	164.6
230.9	165.9

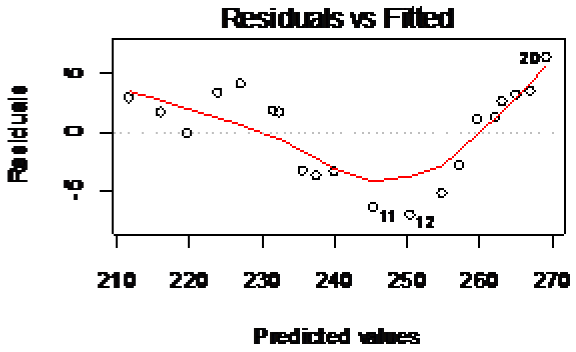
Note: See website for full dataset

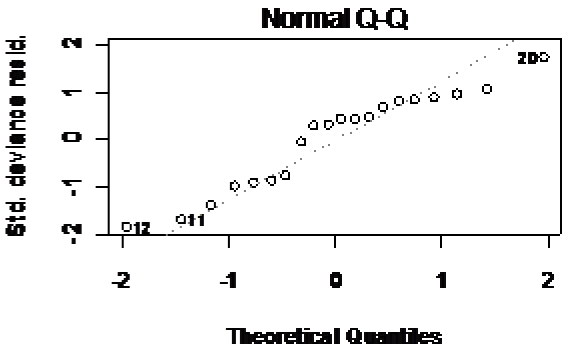
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-154.7191	19.85	-7.794	3.54e-07 ***
STOCK	2.3004	0.1146	20.08	8.99e-14 ***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 15.86173)

Residual deviance: 285.51 on 18 degrees of freedom





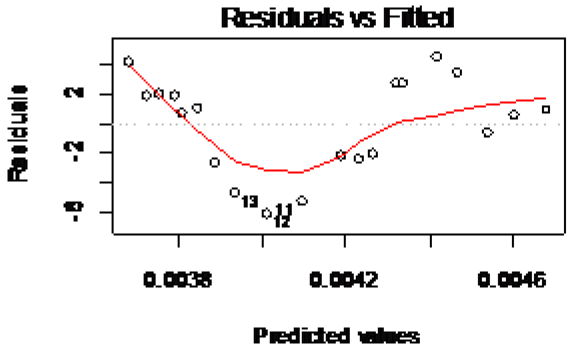
Revised Model: Expendit= μ + normal error (inverse link)

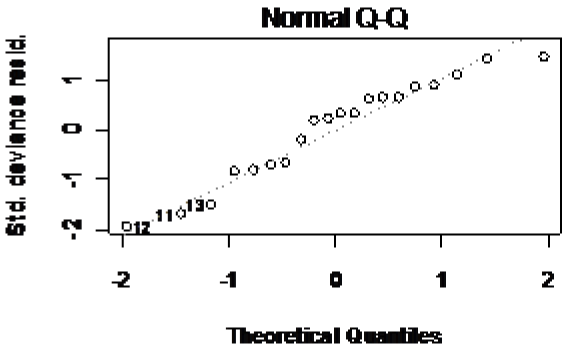
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.10E-02	2.85E-04	38.68	< 2e-16 ***
STOCK	-3.98E-05	1.62E-06	-24.52	2.78e-15 ***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

(Dispersion parameter for gaussian family taken to be 10.47296)

Residual deviance: 188.51 on 18 degrees of freedom





Summary: In the original model the assumptions are violated since the residuals are strongly heterogeneous and skewed. The revised model still has strongly heterogeneous residuals but residuals appear less skewed. Also the dispersion parameter has decreased from 15.86 to 10.47 Parameter estimates cannot be compared since the link was changed. The p-value of stock in the revised model decreased thus increasing Type I Error. Variables may be non-orthogonal.

arcsine

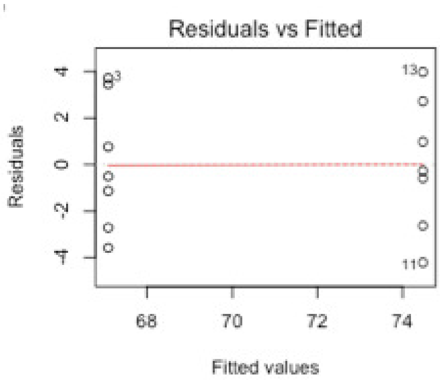
Analysis of percent (Y) in relation to group (X) – no biological source

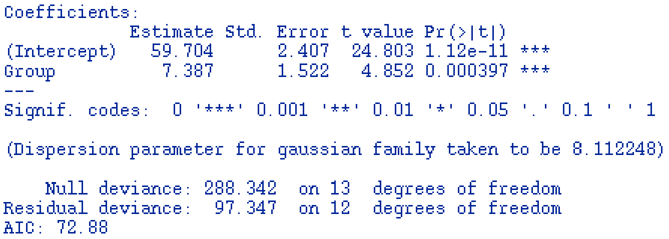
Data from example 13.4 in Zar, J.H. (1996). *Biostatistical Analysis*, Pretice Hall: New Jersey, p. 281

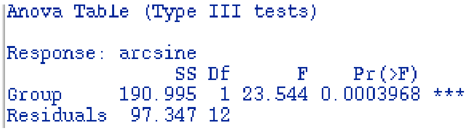
Model: Y = μ + normal error $\mu = ßo + ß_X \cdot X$

Using square arcsine transformation for percentage data: p' = arcsin√p

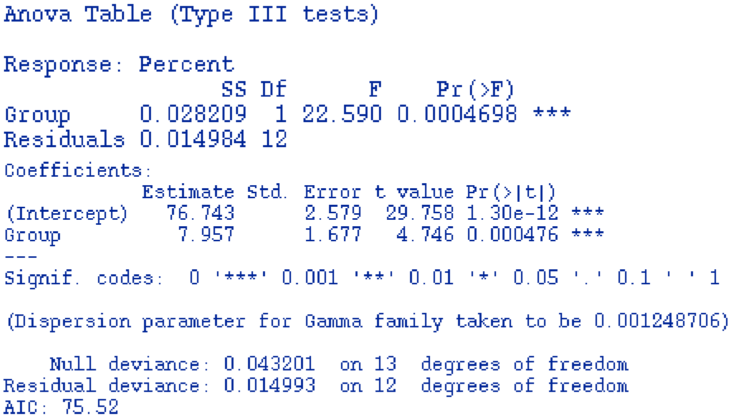
Group	Percent	arcsine
1	84.2	66.58
1	88.9	70.54
1	89.2	70.81
1	83.4	65.96
1	80.1	63.51
1	81.3	64.38
1	85.8	67.86
2	92.3	73.89

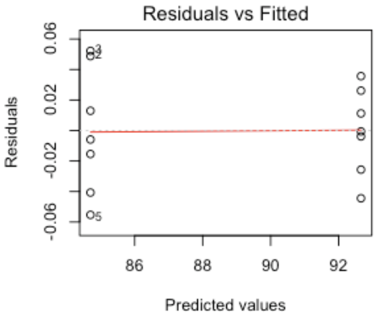



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.704      2.407  24.803 1.12e-11 ***
Group          7.387      1.522   4.852 0.000397 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 8.112248)

    Null deviance: 288.342  on 13  degrees of freedom
Residual deviance:  97.347  on 12  degrees of freedom
AIC: 72.88
```


```
Anova Table (Type III tests)

Response: arcsine
              SS Df      F    Pr(>F)
Group     190.995  1 23.544 0.0003968 ***
Residuals  97.347 12
```

Revised Model: Y = μ + gamma error (identity link) $\mu = ßo + ß_X \cdot X$


```
Anova Table (Type III tests)

Response: Percent
                SS Df      F    Pr(>F)
Group     0.028209  1 22.590 0.0004698 ***
Residuals 0.014984 12
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   76.743      2.579  29.758 1.30e-12 ***
Group          7.957      1.677   4.746 0.000476 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.001248706)

    Null deviance: 0.043201  on 13  degrees of freedom
Residual deviance: 0.014993  on 12  degrees of freedom
AIC: 75.52
```



Summary: GLM assumptions met, but dispersion factor and residual deviance are high. Best revised model was gamma error with identity link. Type I error increased from .0397% to .0466%

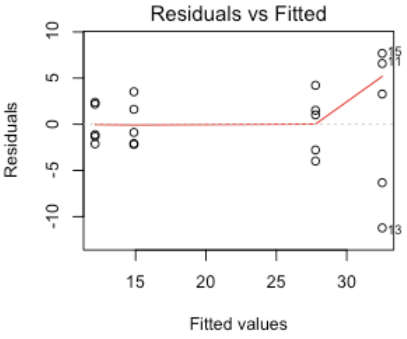
calcium

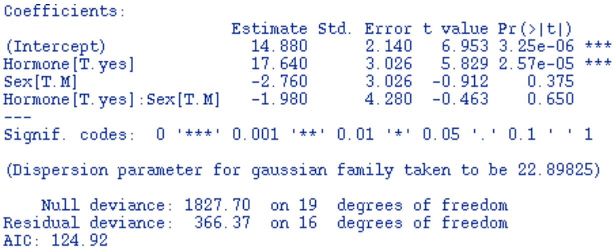
Analysis of calcium plasma concentration in birds (C) in relation to hormone treatment (T) and sex (S)

Data from example 12.1 in Zar, J.H. (1996). *Biostatistical Analysis*, Pretice Hall: New Jersey, p. 237

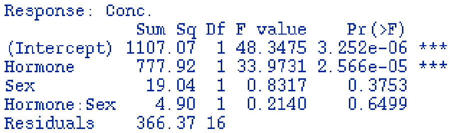
Model: C = μ + normal error $\mu = ßo + ß_T \cdot T + ß_S \cdot S + ß_{T \cdot S} \cdot T \cdot S$

Sex	Hormone	Conc.
2	2	16.5
2	2	18.4
2	2	12.7
2	2	.
.	.	.
.	.	.
.	.	.



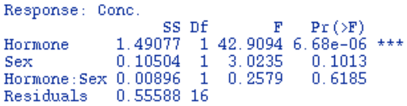


```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                14.880      2.140   6.953 3.25e-06 ***
Hormone[T.yes]             17.640      3.026   5.829 2.57e-05 ***
Sex[T.M]                   -2.760      3.026  -0.912    0.375
Hormone[T.yes]:Sex[T.M]    -1.980      4.280  -0.463    0.650
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 22.89825)

    Null deviance: 1827.70  on 19  degrees of freedom
Residual deviance:  366.37  on 16  degrees of freedom
AIC: 124.92
```

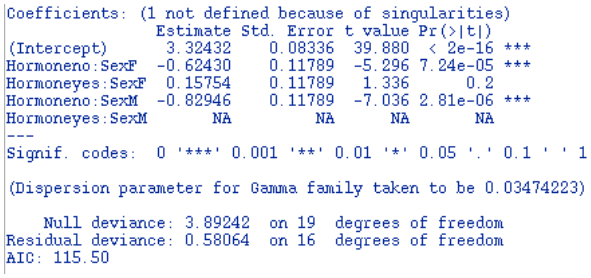


```
Response: Conc.
            Sum Sq Df F value    Pr(>F)
(Intercept) 1107.07  1 48.3475 3.252e-06 ***
Hormone      777.92  1 33.9731 2.566e-05 ***
Sex           19.04  1  0.8317    0.3753
Hormone:Sex    4.90  1  0.2140    0.6499
Residuals    366.37 16
```

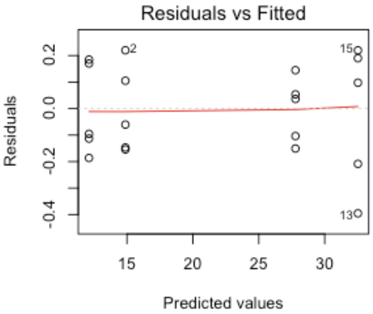
Revised Model: C = μ + gamma error $\mu = ßo + ß_T \cdot T + ß_S \cdot S + ß_{T \cdot S} \cdot T \cdot S$



```
Response: Conc.
                 SS Df       F   Pr(>F)
Hormone     1.49077  1 42.9094 6.68e-06 ***
Sex         0.10504  1  3.0235   0.1013
Hormone:Sex 0.00896  1  0.2579   0.6185
Residuals   0.55588 16
```



```
Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.32432    0.08336  39.880  < 2e-16 ***
Hormoneno:SexF -0.62430    0.11789  -5.296 7.24e-05 ***
Hormoneyes:SexF  0.15754    0.11789   1.336      0.2
Hormoneno:SexM -0.82946    0.11789  -7.036 2.81e-06 ***
Hormoneyes:SexM       NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.03474223)

    Null deviance: 3.89242  on 19  degrees of freedom
Residual deviance: 0.58064  on 16  degrees of freedom
AIC: 115.50
```



Summary:

GLM assumptions not met because of residual heterogeneity. Best revised model was Gamma error with log link. Parameter estimates cannot be compared because of change in model structure. Estimate of Type I error for main effect of hormone reduced from 0.0026% to 0.0007%. Estimate of Type I error for main effect of sex reduced from 38% to 62%. Estimate of Type I error for interaction between sex and hormone reduced from 65% to 62%

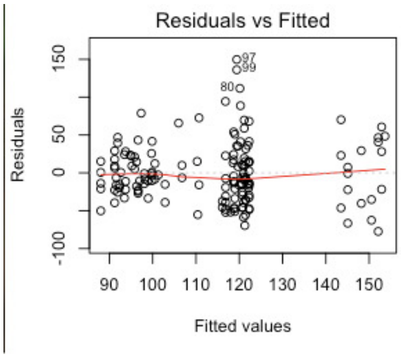
cort

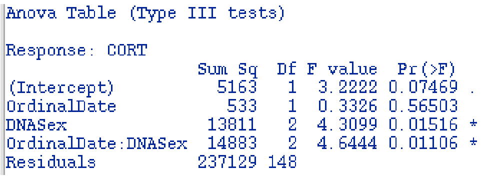
Analysis of corticosterone level of Atlantic Puffins (CORT) in relation to ordinal date (T) and sex (S)

Data from Storey/Walsh Lab, 1998-2003

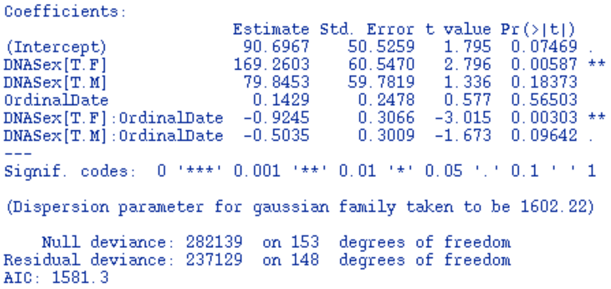
Model: CORT = μ + normal error $\mu = ßo + ß_T \cdot T + ß_S \cdot S + ß_{T \cdot S} \cdot T \cdot S$

OrdinalDate	CORT	DNASex
147	152.474	F
147	122.787	F
196	89.6	M
196	96.1	M
196	143.1	.
196	100.3	F
196	141.5	M
.	.	.
.	.	.
.	.	.



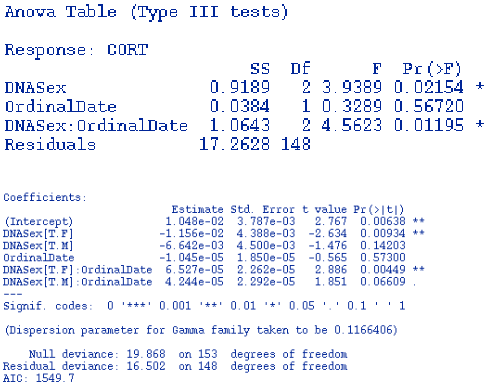


```
Anova Table (Type III tests)

Response: CORT
                  Sum Sq  Df F value  Pr(>F)
(Intercept)         5163   1  3.2222 0.07469 .
OrdinalDate          533   1  0.3326 0.56503
DNASex             13811   2  4.3099 0.01516 *
OrdinalDate:DNASex 14883   2  4.6444 0.01106 *
Residuals         237129 148
```

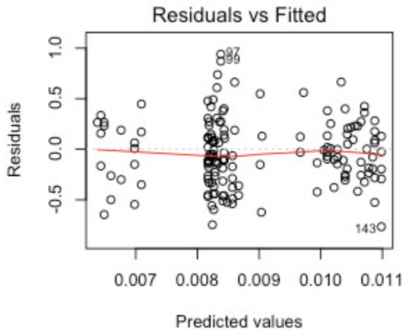


```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              90.6967    50.5259   1.795  0.07469 .
DNASex[T.F]             169.2603    60.5470   2.796  0.00587 **
DNASex[T.M]              79.8453    59.7819   1.336  0.18373
OrdinalDate               0.1429     0.2478   0.577  0.56503
DNASex[T.F]:OrdinalDate  -0.9245     0.3066  -3.015  0.00303 **
DNASex[T.M]:OrdinalDate  -0.5035     0.3009  -1.673  0.09642 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1602.22)

    Null deviance: 282139  on 153  degrees of freedom
Residual deviance: 237129  on 148  degrees of freedom
AIC: 1581.3
```

Revised Model: CORT = µ + gamma error (identity) $\mu = ßo + ß_T \cdot T + ß_S \cdot S + ß_{T \cdot S} \cdot T \cdot S$



```
Anova Table (Type III tests)

Response: CORT
                        SS  Df      F   Pr(>F)
DNASex              0.9189   2 3.9389 0.02154 *
OrdinalDate         0.0384   1 0.3289 0.56720
DNASex:OrdinalDate  1.0643   2 4.5623 0.01195 *
Residuals          17.2628 148


Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.048e-02  3.787e-03   2.767  0.00638 **
DNASex[T.F]             -1.156e-02  4.388e-03  -2.634  0.00934 **
DNASex[T.M]             -6.642e-03  4.500e-03  -1.476  0.14203
OrdinalDate             -1.045e-05  1.850e-05  -0.565  0.57300
DNASex[T.F]:OrdinalDate  6.527e-05  2.262e-05   2.886  0.00449 **
DNASex[T.M]:OrdinalDate  4.244e-05  2.292e-05   1.851  0.06609 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1166406)

    Null deviance: 19.868  on 153  degrees of freedom
Residual deviance: 16.502  on 148  degrees of freedom
AIC: 1549.7
```



Summary:

GLM assumptions met, but dispersion factor and residual deviance are high. Best revised model was gamma error with identity link. Type I error from increased from 1.106% 1.195%.

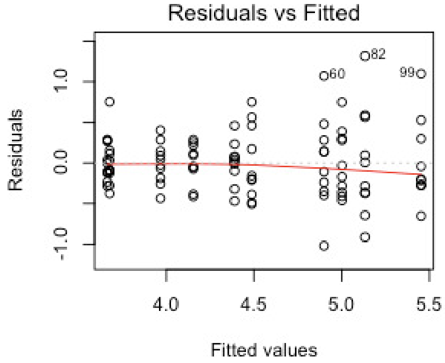
dairy

Analysis of butterfat (B) in relation to age (A) and breed (T)

Data from Sokal, R. R. and Rohlf F. J. (1981). *Biometry*, 2nd edition, San Fransisco: WH Freeman.

Model: B = µ + normal error $5\mu = ßo + ß_A \cdot A + ß_T \cdot T + ß_{A \cdot T} \cdot A \cdot T$

Butterfat	Breed	Age
3.74	Ayrshire	Mature
4.01	Ayrshire	2year
3.77	Ayrshire	Mature
3.78	Ayrshire	2year
.	Ayrshire	Mature
.	Ayrshire	2year
.	Ayrshire	Mature



Response: Butterfat

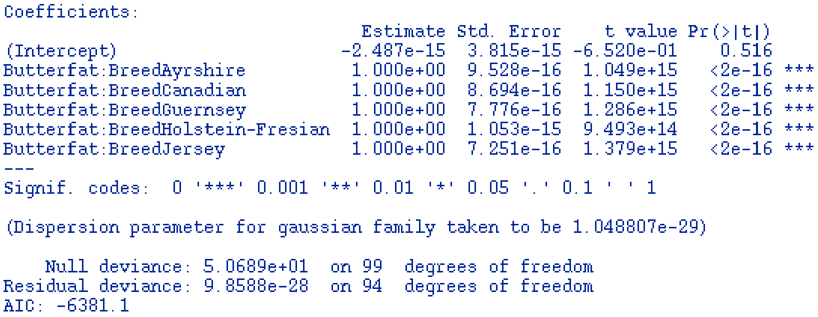
Sum Sq Df F value Pr(>F)

(Intercept) 157.292 1 908.6086 < 2.2e-16 ***

Age 0.177 1 1 .0208 0.3150

Breed 15.211 4 2 1.9672 1.128e-12 ***

Age:Breed 0.514 4 0 .7421 0.5658



```
Coefficients:
                                Estimate Std. Error    t value Pr(>|t|)    
(Intercept)                   -2.487e-15  3.815e-15 -6.520e-01    0.516    
Butterfat:BreedAyrshire        1.000e+00  9.528e-16  1.049e+15   <2e-16 ***
Butterfat:BreedCanadian        1.000e+00  8.694e-16  1.150e+15   <2e-16 ***
Butterfat:BreedGuernsey        1.000e+00  7.776e-16  1.286e+15   <2e-16 ***
Butterfat:BreedHolstein-Fresian 1.000e+00  1.053e-15  9.493e+14   <2e-16 ***
Butterfat:BreedJersey          1.000e+00  7.251e-16  1.379e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.048807e-29)

    Null deviance: 5.0689e+01  on 99  degrees of freedom
Residual deviance: 9.8588e-28  on 94  degrees of freedom
AIC: -6381.1
```

Summary: GLM assumption not met because residuals are heterogeneous. Revised models did not increase residual homogeneity.

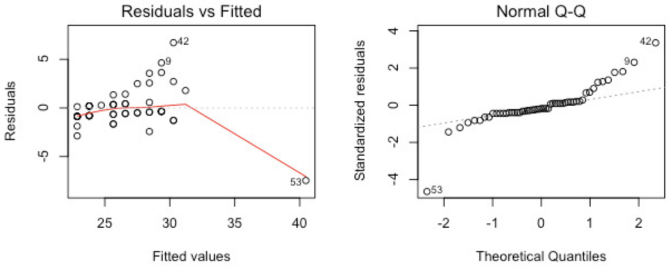
gannets

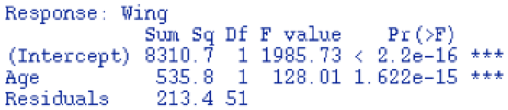
Analysis of chick wing length on the Gannet Islands in 1996 (WL) in relation to chick age (A)

Data from Mark Hipfner, Canadian Wildlife Service

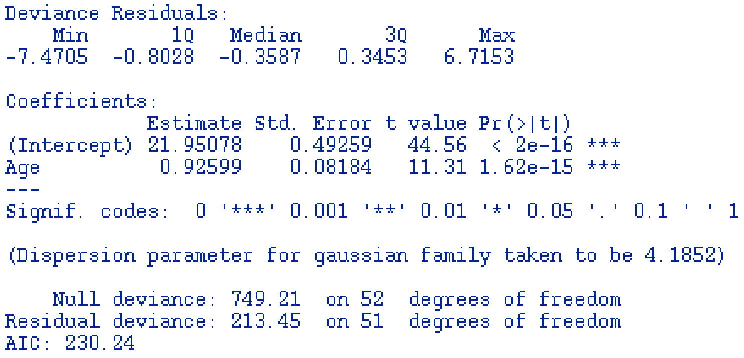
Model: WL = μ + normal error $\mu = ß_o + ß_A \cdot A$

Chick	Age	Wing
1	2	23
1	5	26
1	8	29
.	.	.
.	.	.
.	.	.



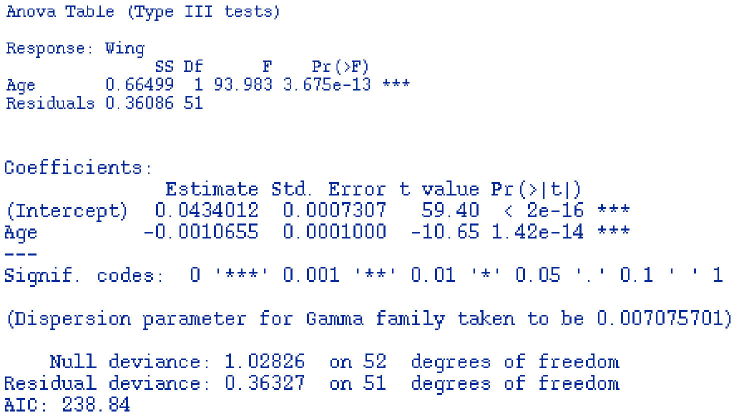


```
Response: Wing
            Sum Sq Df F value    Pr(>F)    
(Intercept) 8310.7  1 1985.73 < 2.2e-16 ***
Age          535.8  1  128.01 1.622e-15 ***
Residuals    213.4 51                      
```

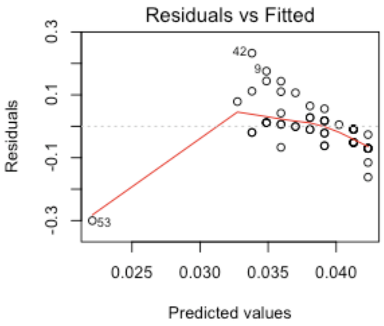


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4705  -0.8028  -0.3587   0.3453   6.7153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.95078    0.49259   44.56  < 2e-16 ***
Age          0.92599    0.08184   11.31 1.62e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.1852)

    Null deviance: 749.21  on 52  degrees of freedom
Residual deviance: 213.45  on 51  degrees of freedom
AIC: 230.24
```

Revised model: WL = μ + gamma error (inverse link) $\mu = ßo + ß_{A} \cdot A$



```
Anova Table (Type III tests)

Response: Wing
          SS Df      F    Pr(>F)
Age      0.66499  1 93.983 3.675e-13 ***
Residuals 0.36086 51


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0434012  0.0007307   59.40  < 2e-16 ***
Age         -0.0010655  0.0001000  -10.65 1.42e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.007075701)

    Null deviance: 1.02826  on 52  degrees of freedom
Residual deviance: 0.36327  on 51  degrees of freedom
AIC: 238.84
```



Summary: GLM assumptions not met because residuals are not homogeneous. Dispersion factor and residual deviance are also high in LM. Best revised model was gamma error with inverse link. Although residual heterogeneity was not resolved, dispersion factor and residual deviance were improved. Type I error increased from .000 000 000 000 162 2% to .000 000 000 001 42%

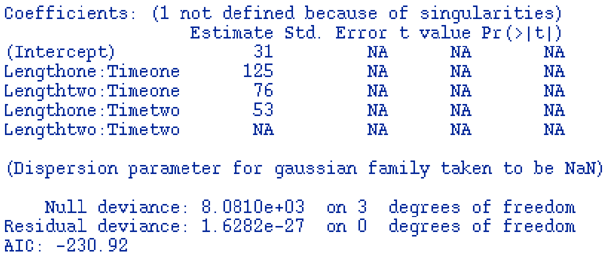
root

Analysis of number of plum root-stocks (B) in relation interaction between length (L)of cutting and time of planting (T).

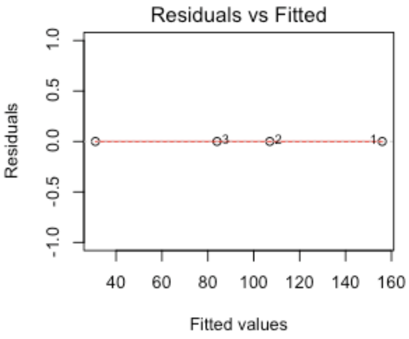
Data from Bartlett MS (1935). Contingency table interactions, in *Journal of the Royal Statistical Society Supplement*, **2**, p 248–252.

Model: B = μ + normal error $\mu = ßo + ß_L \cdot L + ß_T \cdot T + ß_{T \cdot L} \cdot T \cdot L$

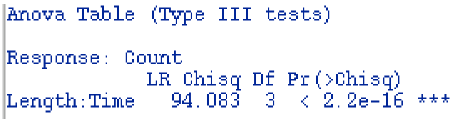
Count	Time	Length
156	one	one
107	one	two
84	two	one
31	two	two



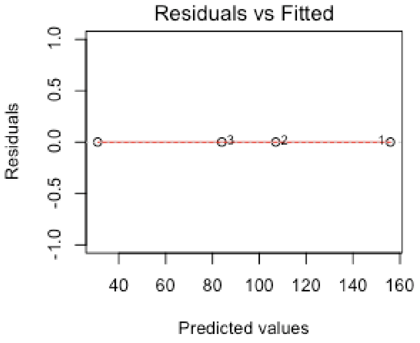
```
Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)               31         NA      NA       NA
Lengthone:Timeone        125         NA      NA       NA
Lengthtwo:Timeone         76         NA      NA       NA
Lengthone:Timetwo         53         NA      NA       NA
Lengthtwo:Timetwo         NA         NA      NA       NA

(Dispersion parameter for gaussian family taken to be NaN)

    Null deviance: 8.0810e+03  on 3  degrees of freedom
Residual deviance: 1.6282e-27  on 0  degrees of freedom
AIC: -230.92
```

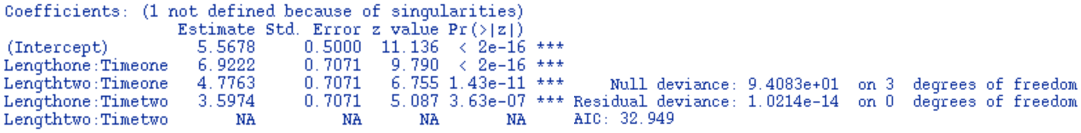


Revised Model: B = μ + poisson error (log link) $\mu = e^{(ßo)} + e^{(ßL \cdot L)} + e^{(ßT \cdot T)} + e^{(ßT \cdot L \cdot T \cdot L)}$



```
Anova Table (Type III tests)

Response: Count
            LR Chisq Df Pr(>Chisq)
Length:Time   94.083  3  < 2.2e-16 ***
```





```
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.5678     0.5000  11.136  < 2e-16 ***
Lengthone:Timeone   6.9222     0.7071   9.790  < 2e-16 ***
Lengthtwo:Timeone   4.7763     0.7071   6.755 1.43e-11 ***     Null deviance: 9.4083e+01  on 3  degrees of freedom
Lengthone:Timetwo   3.5974     0.7071   5.087 3.63e-07 *** Residual deviance: 1.0214e-14  on 0  degrees of freedom
Lengthtwo:Timetwo       NA         NA      NA       NA     AIC: 32.949
```

Dispersion parameter = 1

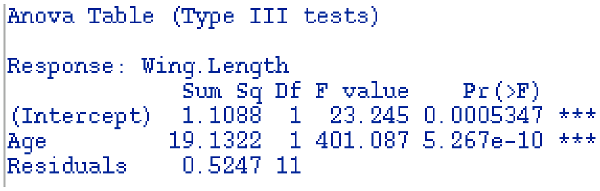
Summary: GLM assumptions met, but no type I errors were estimated. Best revised model was poisson error with log link.

sparrows

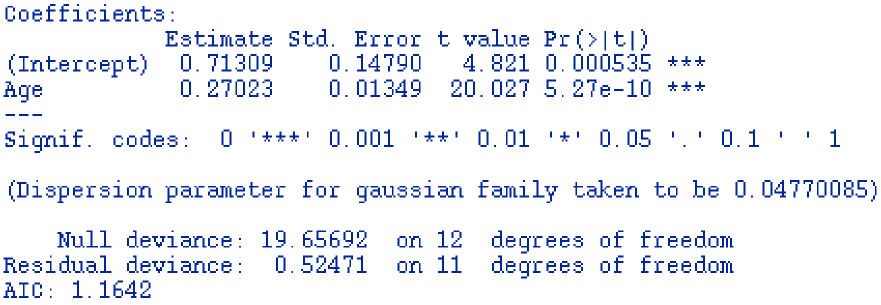
Analysis of wing length of sparrows (L) in relation to age (A).

Data from example 16.1 in Zar, J.H. (1996). *Biostatistical Analysis*, Pretice Hall: New Jersey, p. 319

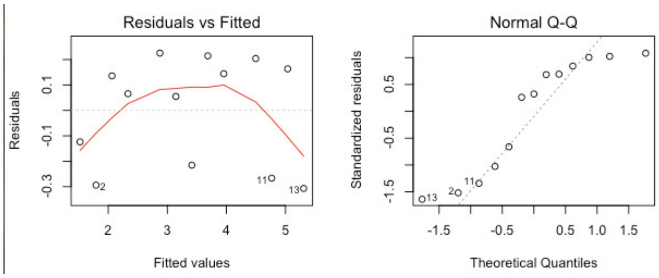
Model: C = μ + normal error $\mu = ßo + ß_L \cdot L$



```
Anova Table (Type III tests)

Response: Wing.Length
             Sum Sq Df F value    Pr(>F)
(Intercept)  1.1088  1  23.245 0.0005347 ***
Age         19.1322  1 401.087 5.267e-10 ***
Residuals    0.5247 11
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.71309    0.14790   4.821 0.000535 ***
Age          0.27023    0.01349  20.027 5.27e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.04770085)

    Null deviance: 19.65692  on 12  degrees of freedom
Residual deviance:  0.52471  on 11  degrees of freedom
AIC: 1.1642
```



Summary: GLM assumptions not met because residuals are heterogeneous and non-normal. Revised models did not improve residual homogeneity. Gamma error with inverse link improved normality, but increased heterogeneity of residuals.

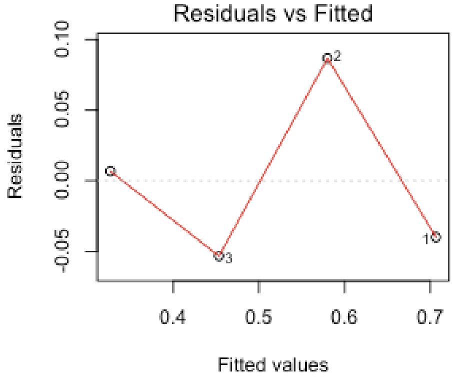
texts

Analysis of binomial responses of biology professors (R) in relation to different textbooks (B)

Data from exercise 12.4 in Zar, J.H. (1996). *Biostatistical Analysis*, Pretice Hall: New Jersey, p. 276

Model: R = μ + normal error $\mu = ßo + ß_B \cdot B$

Textbook	Response
1	1
1	1
1	1
1	1
1	1
1	0
1	0



```
Anova Table (Type III tests)

Response: Count
             Sum Sq Df F value  Pr(>F)
(Intercept) 0.46296  1  77.160 0.01271 *
Textbook    0.08022  1  13.370 0.06733 .
Residuals   0.01200  2
---

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.83333    0.09487   8.784   0.0127 *
Textbook    -0.12667    0.03464  -3.657   0.0673 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.006)

    Null deviance: 0.092222  on 3  degrees of freedom
Residual deviance: 0.012000  on 2  degrees of freedom
AIC: -5.8851
```

Revised Model: Odds = $e^{(\beta o)}$ + $e^{(\beta B)}$ + binomial error (logit)

```
Anova Table (Type III tests)

Response: Response
         LR Chisq Df Pr(>Chisq)
Textbook   5.6344  3     0.1308
```

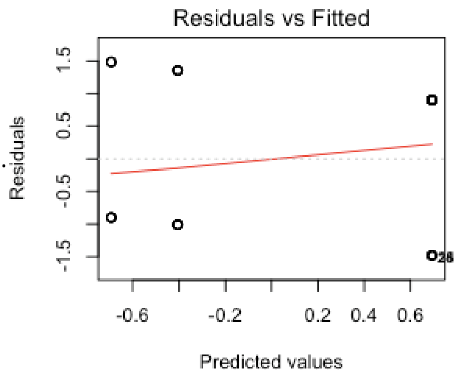
```
Coefficients:
               Estimate Std. Error  z value Pr(>|z|)
(Intercept)   6.931e-01  5.477e-01    1.266   0.2057
Textbook[T.B]  8.399e-16  7.746e-01 1.08e-15   1.0000
Textbook[T.C] -1.099e+00  7.601e-01   -1.445   0.1484
Textbook[T.D] -1.386e+00  7.746e-01   -1.790   0.0735 .
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.111  on 59  degrees of freedom
Residual deviance: 77.477  on 56  degrees of freedom
AIC: 85.477
```

Summary: GLM assumptions not met because of residual heterogeneity.

Best revised model was binomial error with logit link.

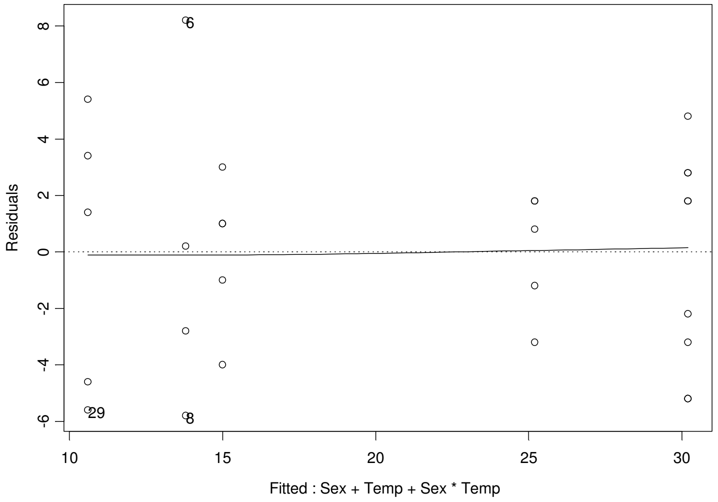
Type I error increased from 6.733% to 13.08%



1. Analysis of fly survival depending on temperature where flies were raised and sex
250 males and 250 females were selected from 3 population cages
1 = kept at 24 degrees
2 = kept at 17 degrees
3 = kept at natural outdoor temperature
Flies were captured and placed in bottles kept at 4 degrees under natural light conditions. The number of survivors was counted at the end of 45 days.
Data from table 13.1 Rao 1998

Number of Survivors	Sex	Temp of cage	Bottle
26	F	1	1
22	F	1	2
27	F	1	3
24	F	1	4
27	F	1	5
22	M	1	1
11	M	1	2
.	.	.	.

Count of flies that survived = F_s
Sex = S
Temperature of cage that flies were raised in = T
$F_s = \mu$ + normal error $\mu = \alpha + \beta_t * T + \beta_s * S + \beta_{T*S} * T*S$

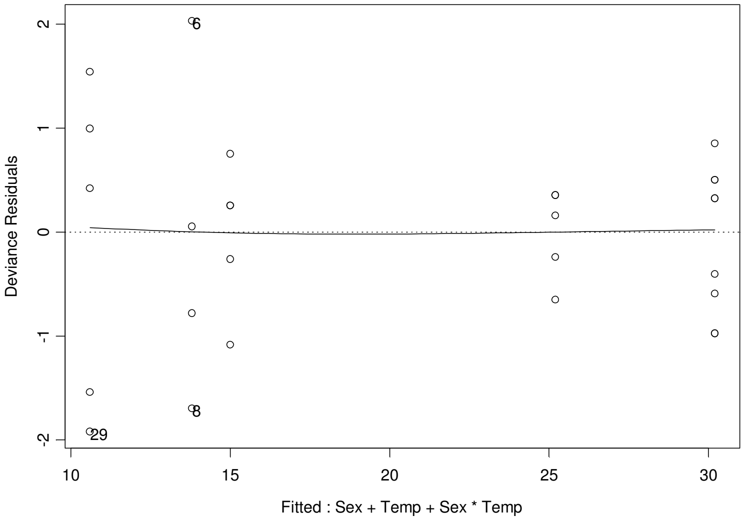


(Dispersion Parameter for Gaussian family taken to be 15.35)
Residual Deviance: 368.4 on 24 degrees of freedom

	Value	Std. Error	t value
(Intercept)	25.2	1.752142	14.3824
Sex	-11.4	2.477902	-4.60067
Temp2	5	2.477902	2.017836
Temp3	5	2.477902	2.017836
SexTemp2	-3.8	3.504283	-1.08439
SexTemp3	-8.2	3.504283	-2.33999

	Df	Deviance Res	Df	Resid. Dev	Type I error
NULL			29	2282.167	
Sex	1	1778.7	28	503.467	0
Temp	2	50.867	26	452.6	9.00267E-12
Sex:Temp	2	84.2	24	368.4	5.20238E-19

Revised Model: $F_s = e^{\mu}$ + poisson error $\mu = \alpha + \beta_t * T + \beta_s * S + \beta_{T*S} * T*S$



(Dispersion Parameter for Poisson family taken to be 1)
Residual Deviance: 23.8751 on 24 degrees of freedom

	Value	Std. Error	t value
(Intercept)	3.22684399	0.08908688	36.22132
Sex	-0.60217429	0.14969163	-4.02277
Temp2	0.18099793	0.12065988	1.500067
Temp3	0.18099793	0.12066019	1.500063
SexTemp2	-0.09761743	0.20582066	-0.47428

SexTemp3	-0.44480573	0.21868467	-2.03401

	Df	Deviance Resid.	Df	Resid. Dev	Type I error
NULL			29	118.3458	
Sex	1	87.43614	28	30.9097	8.70454E-21
Temp	2	2.4204	26	28.4893	0.298137646
Sex:Temp	2	4.61415	24	23.8751	0.099552016

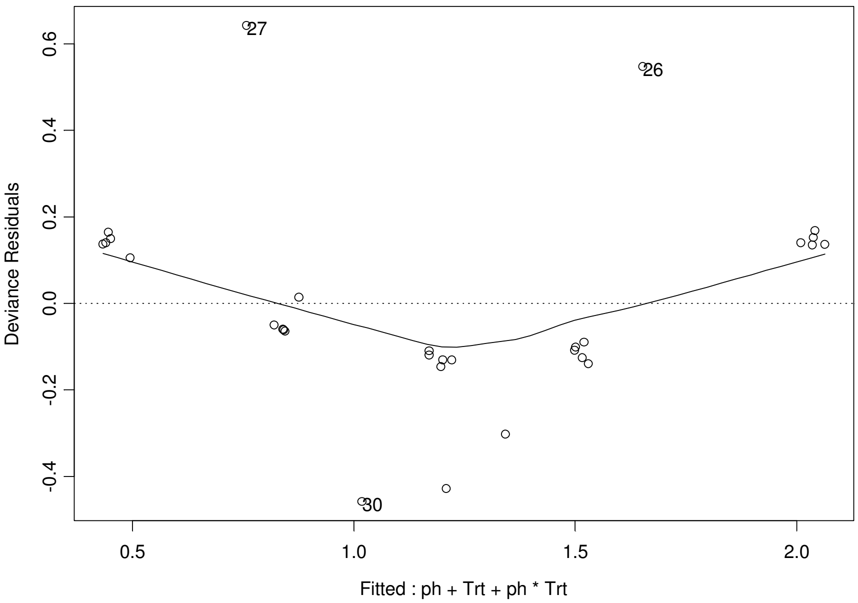
Summary:
Original test: ANOVA or GLM with gaussian error structure and identity link
GLM assumptions met according to Res vd Fits plot, however the dispersion parameter is very large and the res dev is >> than 2 times the degrees of freedom
Best revised model was poisson error with log link (survival is a count)
Res vs Fit plot still looks ok, however, the dispersion parameter and res dev both improved to within acceptable numbers
Type I error increased from p << 0.001 for both Temperature and the interaction term, to p = 0.298 and p = 0.996 for Temperature and the interaction term respectively.
This changes the original decision from rejecting null hypothesis to accepting it.

2. Analysis of calcium ion activities based on the ph value of the milk and the preheat treatment applied during manufacture of milk powder at 6 levels:
1 = none, 2 = low heat, 3 = medium heat, 4 = high heat, 5 = indirect UHT, and 6 = direct UHT

Data from table 12.1 in Rao 1998

Ca activity	pH	trt
2.21	6.07	1
1.39	6.35	1
1.05	6.52	1
0.78	6.71	1
0.61	6.92	1
2.19	6.08	2
1.39	6.36	2
1.09	6.53	2
.	.	.

Ca ion activity = Ca
Preheat Treatment = T
Ph = P
Ca = μ + normal error $\mu = \alpha + \beta_t * T + \beta_p * P + \beta_{T*P} * T*P$

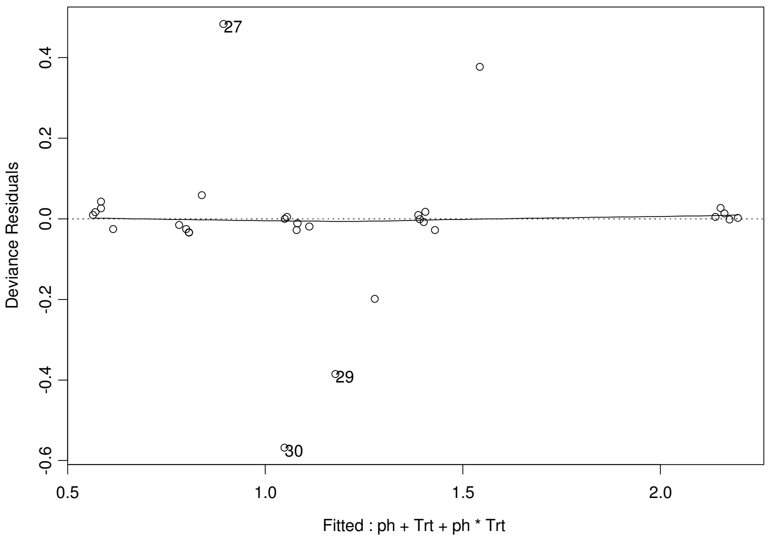


(Dispersion Parameter for Gaussian family taken to be 0.0870107)
Residual Deviance: 1.566193 on 18 degrees of freedom

	Value	Std. Error	t value
(Intercept)	13.43905	2.9423508	4.567454
ph	-1.87766	0.4512423	-4.16108
Trt2	-0.36364	4.1643012	-0.08732
Trt3	0.027967	4.1453887	0.006746
Trt4	-0.24654	4.1126203	-0.05995
Trt5	-0.3675	4.1392589	-0.08878
Trt6	-7.50162	3.6034171	-2.08181
phTrt2	0.062295	0.6381529	0.097618
phTrt3	0.002117	0.634868	0.003335
phTrt4	0.036522	0.6305122	0.057924
phTrt5	0.05228	0.6350843	0.082319
phTrt6	1.172925	0.5466533	2.145646

				Resid.	
	Df	Deviance Res	DF	Dev	Type I error
NULL			29	9.44852	
ph	1	6.664595	28	2.783925	0.009834701
Trt	5	0.376689	23	2.407236	0.995947412
ph:Trt	5	0.841043	18	1.566193	0.974341966

Revised Model: Ca = e^{μ} + gamma error $\mu = \alpha + \beta_t * T + \beta_p * P + \beta_{T*P} * T*P$



(Dispersion Parameter for Gamma family taken to be 0.0487433)
Residual Deviance: 0.9005756 on 18 degrees of freedom

	Value	Std. Error	t value
(Intercept)	10.06887	2.202244	4.572097
ph	-1.53251	0.337739	-4.53756
Trt2	-0.31412	3.11683	-0.10078
Trt3	0.070981	3.102674	0.022877
Trt4	0.09655	3.078148	0.031366
Trt5	0.004996	3.098086	0.001613
Trt6	-7.02596	2.697028	-2.60508
phTrt2	0.054917	0.477635	0.114978
phTrt3	-0.00588	0.475176	-0.01236
phTrt4	-0.01675	0.471916	-0.0355
phTrt5	-0.00434	0.475338	-0.00913
phTrt6	1.103401	0.40915	2.69681

	Df	Deviance Resid.	Df	Resid. Dev	Type I error
NULL			29	6.264378	
ph	1	3.907554	28	2.356824	0.04806952
Trt	5	0.48883	23	1.867994	0.992526585
ph:Trt	5	0.967418	18	0.900576	0.965150694

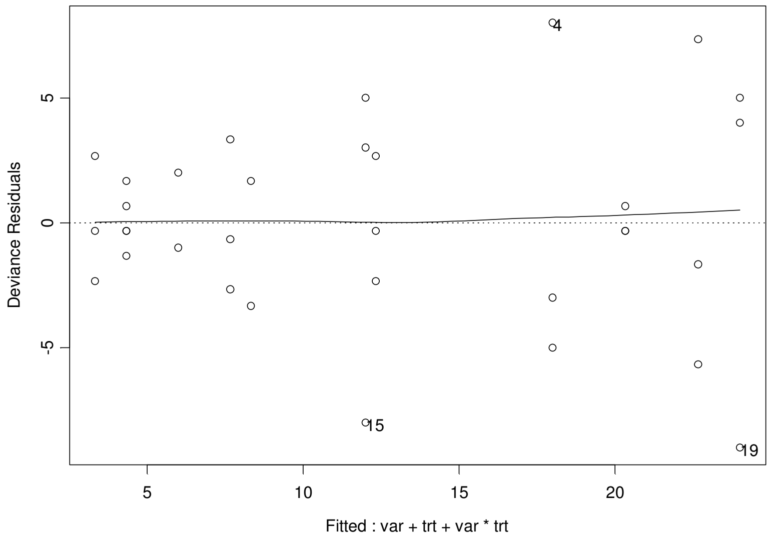
Summary:
GLM assumptions met according to Res vs Fits plot, dispersion parameter and the res dev

Revised model was gamma error with log link
This model lowered the dispersion parameter, but did not benefit the model in any other way.
The type 1 error increased from $p << 0$ to $p = 0.048$

3. Drought resistance of four crop varieties was compared.
First, plants were given 3 different types of pre-treatments to stimulate root growth. Cuttings of each of the 4 different plant varieties were then exposed to drought conditions. Average root lengths after 4 months of growth were measured.
Data from table 13.2 Rao 1998

length	variety	Pretreatment
11	1	1
5	1	1
7	1	1
26	2	1
13	2	1
15	2	1
.	.	.

Length of root = L
Pretreatment type = P
Variety = V
$L = \mu$ + normal error $\mu = \alpha + \beta_P * P + \beta_V * V + \beta_{L*V} * L*V$

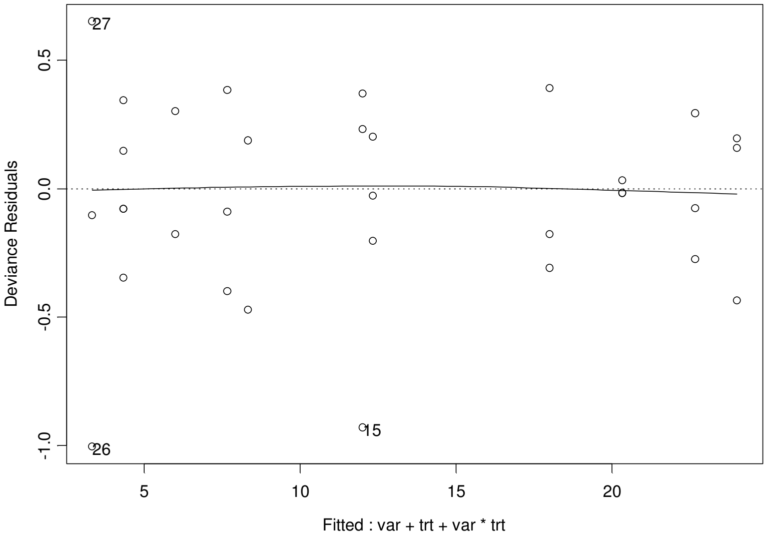


(Dispersion Parameter for Gaussian family taken to be 19.97222)
Residual Deviance: 479.3333 on 24 degrees of freedom

	Value	Std. Error	t value
(Intercept)	7.666667	2.580195	2.971351
var2	10.33333	3.648947	2.831867
var3	15	3.648947	4.110775

var4	-1.66667	3.648947	-0.45675
trt2	4.333333	3.648947	1.187557
trt3	-4.33333	3.648947	-1.18756
var2trt2	-2	5.16039	-0.38757
var3trt2	-3	5.16039	-0.58135
var4trt2	2	5.16039	0.387568
var2trt3	-9.33333	5.16039	-1.80865
var3trt3	-14	5.16039	-2.71297
var4trt3	6.666667	5.16039	1.291892

		Deviance			
	Df	Resid.	Df	Resid. Dev	Type 1 error
NULL			35	2351.889	
var	3	525.4444	32	1826.444	1.4596E-113
trt	2	924.3889	30	902.056	1.8685E-201
var:trt	6	422.7222	24	479.333	3.6322E-88



(Dispersion Parameter for Gamma family taken to be 0.1499696)
Residual Deviance: 4.32228 on 24 degrees of freedom

		Std.	
	Value	Error	t value
(Intercept)	2.036882	0.223584	9.110138
var2	0.85349	0.316196	2.699246
var3	1.084014	0.316196	3.4283
var4	-0.24512	0.316196	-0.77522
trt2	0.448025	0.316196	1.416922
trt3	-0.83291	0.316196	-2.63416
var2trt2	-0.32613	0.447168	-0.72933
var3trt2	-0.39087	0.447168	-0.87409

var4trt2	0.272521	0.447168	0.609438
var2trt3	-0.59113	0.447168	-1.32193
var3trt3	-0.82165	0.447168	-1.83745
var4trt3	1.161413	0.447168	2.597262

	Df	Deviance Resid.	Df	Resid. Dev	Type 1 error
NULL			35	19.19537	
var	3	3.804517	32	15.39085	0.283361128
trt	2	7.391258	30	7.9996	0.02483183
var:trt	6	3.677315	24	4.32228	0.720248544

Summary:

GLM assumptions are not met due to cone in res vs fits plot, dispersion parameter > 10, and the residual deviance >> than 2 times the degrees of freedom.

The best revised model had a gamma error with log link

This model lowered the dispersion parameter to less than 1, lowered the residual deviance below the degrees of freedom, and eliminated the cone shape in the res vs fits plot.

The type 1 error increased from p << 0 for all explanatory variables to p = 0.283, p = 0.025, and p = 0.720 for variety, treatment, and the interaction term respectively

This changes the decision reject all null hypotheses to instead accept only the alternative hypothesis for treatment.

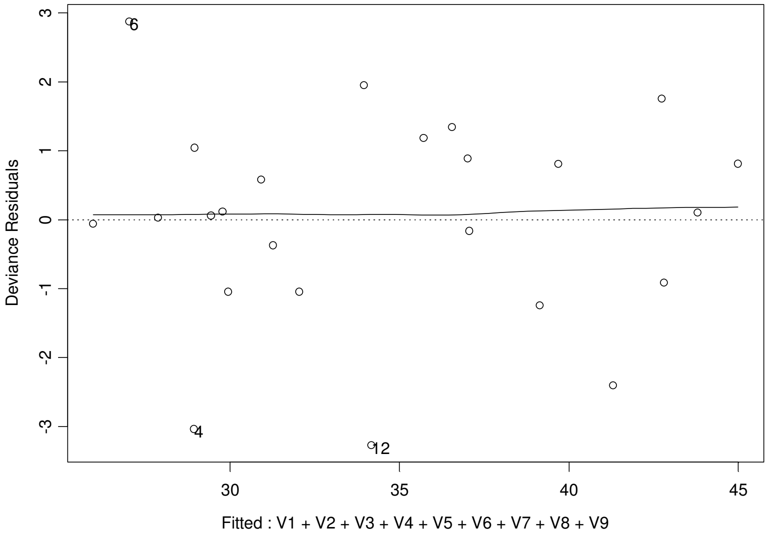
4. Analysis of selling price of houses based on characteristics of the house and taxes.

Data from Chapter 11 Chatterjee and Hadi 2006

X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
4.918	1	3.472	0.998	1	7	4	42	0	25.9
5.021	1	3.531	1.5	2	7	4	62	0	29.5
4.543	1	2.275	1.175	1	6	3	40	0	27.9
.	.	.	.	.	.	.	.	.	.

Building Prices = Y
Taxes (in the thousands) = X1
Number of Bathrooms = X2
Lot size (in thousands of square feet) = X3
Living Space (in thousands of square feet) = X4
Number of garage stalls = X5
Number of rooms = X6
Number of bedrooms = X7
Age of home = X8
Number of fireplaces = X9

$Y = \mu + \text{normal error} \quad \mu = \alpha + \beta_{X1} * X1 + \beta_{X2} * X2 + \beta_{X3} * X3 + \beta_{X4} * X4 + \beta_{X5} * X5 + \beta_{X6} * X6 + \beta_{X6} * X6 + \beta_{X7} * X7 + \beta_{X8} * X8 + \beta_{X9} * X9$



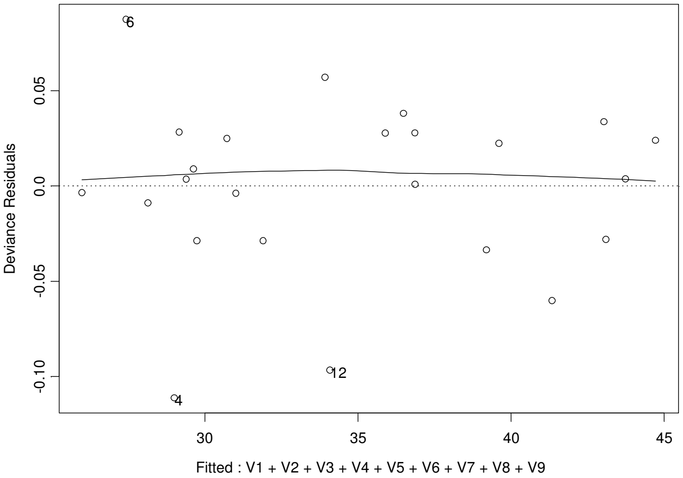
(Dispersion Parameter for Gaussian family taken to be 5.232225)
Residual Deviance: 52.32225 on 10 degrees of freedom

		Std.	
	Value	Error	t value
(Intercept)	9.07156	7.168366	1.265499
V1	2.515235	0.972022	2.587632
V2	2.577285	1.948415	1.32276
V3	0.294573	0.447692	0.65798
V4	1.690261	3.699984	0.456829
V51	3.047836	2.551729	1.19442
V51.5	10.12117	3.256685	3.107813
V52	4.817388	2.587746	1.861615
V66	1.665148	4.160063	0.40027
V67	-2.46749	3.169324	-0.77855
V68	-5.93554	3.8296	-1.54991
V73	-1.16553	2.977188	-0.39149
V74	NA	NA	NA
V8	0.029255	0.064932	0.45055
V9	2.639697	1.61624	1.633233

		Deviance		Resid.	
	Df	Resid.	Df	Dev	Type I error
NULL			23	831.5096	
V1	1	635.0419	22	196.4677	4.0019E-140
V2	1	28.5559	21	167.9118	9.1032E-08
V3	1	4.6546	20	163.2572	0.030970451
V4	1	0.0294	19	163.2278	0.863858713
V5	3	62.25	16	100.9779	1.94272E-13
V6	3	28.5887	13	72.3892	2.7324E-06

V7	1	0.8904	12	71.4987	0.345368941
V8	1	5.2198	11	66.2789	0.022331122
V9	1	13.9567	10	52.3223	0.00018707

Revised Model:
$Y = e^{\mu}$ + gamma error $\mu = \alpha + \beta_{X1} * X1 + \beta_{X2} * X2 + \beta_{X3} * X3 + \beta_{X4} * X4 + \beta_{X5} * X5 + \beta_{X6} * X6 + \beta_{X6} * X6 + \beta_{X7} * X7 + \beta_{X8} * X8 + \beta_{X9} * X9$



(Dispersion Parameter for Gamma family taken to be 0.0045661)
Residual Deviance: 0.0466435 on 10 degrees of freedom

	Value	Std. Error	t value
(Intercept)	2.813598	0.211764	13.2865095
V1	0.069482	0.028715	2.41973106
V2	0.086264	0.057559	1.49871088
V3	0.007795	0.013225	0.58940618
V4	0.059191	0.109303	0.54153012
V51	0.08748	0.075382	1.1604898
V51.5	0.276327	0.096207	2.87221262
V52	0.156293	0.076446	2.04450442
V66	0.001742	0.122894	0.01417426
V67	-0.10669	0.093626	-1.139554
V68	-0.20269	0.113132	-1.7916276
V73	-0.00207	0.08795	-0.0235478
V74	NA	NA	NA
V8	0.000846	0.001918	0.44126094
V9	0.070295	0.047746	1.47227759

	Df	Deviance Resid.	Df	Resid. Dev	Type I error
NULL			23	0.6873635	
V1	1	0.517343	22	0.1700201	0.471977119

V2	1	0.024866	21	0.1451543	0.874701921
V3	1	0.004837	20	0.1403169	0.944550705
V4	1	6.12E-05	19	0.1402557	0.993758179
V5	3	0.043913	16	0.0963426	0.997584563
V6	3	0.032966	13	0.0633765	0.998423735
V7	1	0.002552	12	0.0608247	0.959711722
V8	1	0.004109	11	0.0567156	0.948888801
V9	1	0.010072	10	0.0466435	0.920058642

Summary:
GLM assumptions met according to the res vs fits plot, however the dispersion parameter is > 5 and the deviance residual is >> than 2 times the degrees of freedom
The best revised model had a gamma error with log link
This model lowered the dispersion parameter to less than 1 and lowered the residual deviance below the degrees of freedom.
The type 1 error increased for all explanatory variables.
V1, V2, V5, V6, V9 went from having an extremely low p-value to a very high p-value/
This changes the decision reject the null hypotheses for V1,V2, V3, V5, V6, V8, V9 to instead accept only the alternative hypothesis for all variables.
The lower dispersion value now allows AICc to be run instead of QAICc

AIC ANALYSIS

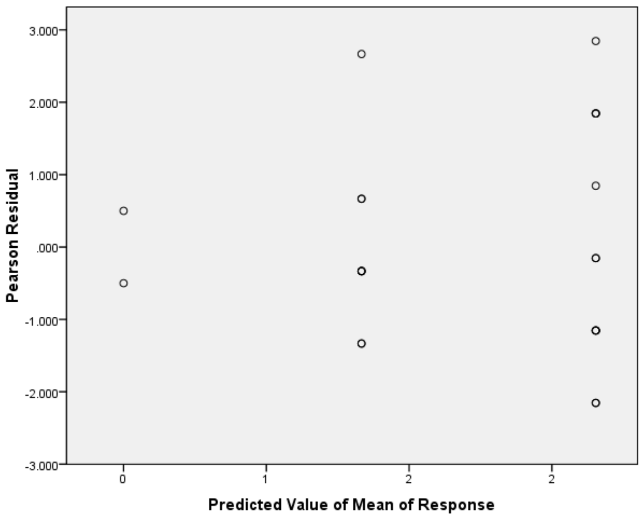
Global model:
Y = X1+X2+X3+X4+X5+X6+X7+X8+X9

5. Cabbage aphid distribution was analyzed depending on the density of a predatory beetle.
Data from Krebs (collected by N Gilbert) 1999

Cabbage aphid	Predatory beetle
5	0
4	0
5	0
1	0
2	1
1	0
0	2
.	.

Count of Cabbage aphids per leaf = C_A
Count of Beetles per leaf = C_B

$C_A = \mu$ + normal error $\mu = \alpha + \beta_{CB} * C_B$



Pearson Chi Square (Dispersion parameter for gaussian family) = 2.2

Omnibus Test[a]		
Likelihood Ratio Chi-Square	df	Sig.
6.766	2	0.034

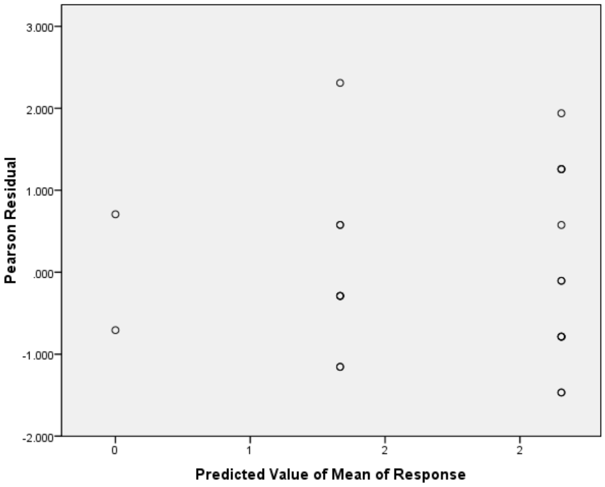
Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	23.106	1	.000
pred	6.766	2	.034

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.500	.7071	-.886	1.886	.500	1	.480
[pred=0]	1.654	.7596	.165	3.143	4.741	1	.029
[pred=1]	.833	.7817	-.699	2.366	1.136	1	.286
[pred=2]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Revised Model:
Y = μ + negative binomial error $\mu = \alpha + \beta_{CB} * C_B$



Pearson Chi Square (Dispersion parameter for Negative binomial family) = 0.452

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
1.889	2	.389

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.072	1	.788
pred	1.799	2	.407

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.693	1.2247	-3.094	1.707	.320	1	.571
[pred=0]	1.460	1.2699	-1.029	3.949	1.323	1	.250
[pred=1]	.981	1.3017	-1.570	3.532	.568	1	.451
[pred=2]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

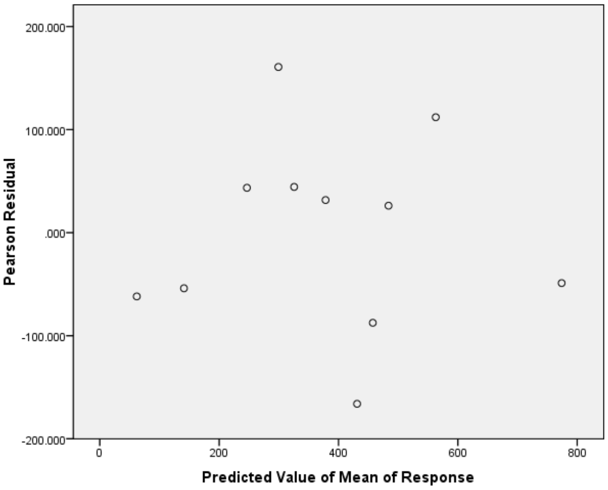
(Negative binomial)	1						

Summary:
Original GLM with gaussian error structure did not meet the assumptions due to cone in res vs fits plot
Revised error structure was negative binomial (count data) with a log link
GzLM improved the Res vs Fits plot
“Omnibus Test” is now very large, meaning the model is not significant??

6. The impacts of wolf numbers on moose populations in British Columbia were measured.
Data from Krebs 1999

Wolves	Moose
8	190
15	370
9	460
27	725
14	265
3	87
12	410
19	675
7	290
10	370
16	510

Number of Wolves = W
Number of Moose = M
M = μ + normal error $\mu = \alpha + \beta_W * W$



Pearson Chi Square (Dispersion parameter for gaussian family) = 9814.210

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
393666.659	1	.000

Goodness of Fit[b]

	Value	df	Value/df
Deviance	88327.886	9	9814.210
Scaled Deviance	88327.886	9	
Pearson Chi-Square	88327.886	9	9814.210
Scaled Pearson Chi-Square	88327.886	9	
Log Likelihood[a]	-44174.052		
Akaike's Information Criterion (AIC)	88352.103		
Finite Sample Corrected AIC (AICC)	88353.603		
Bayesian Information Criterion (BIC)	88352.899		
Consistent AIC (CAIC)	88354.899		

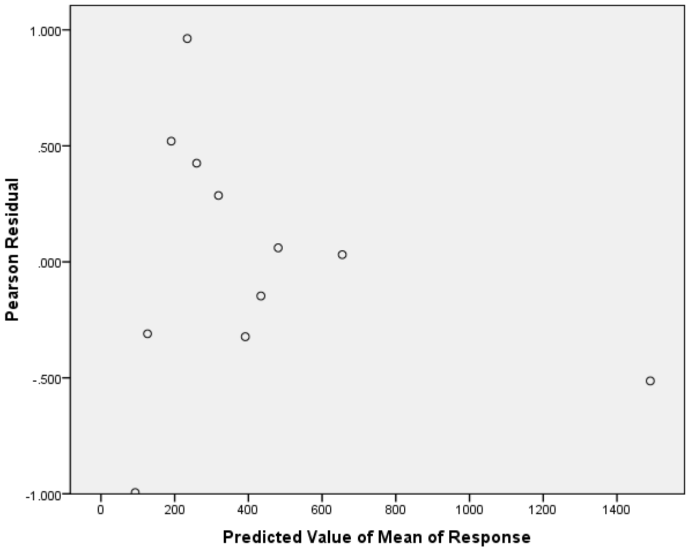
Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	11092.094	1	.000
wolf	393666.659	1	.000

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	61.890	.5876	60.738	63.042	11092.094	1	.000
wolf	26.373	.0420	26.290	26.455	393666.659	1	.000
(Scale)	1[a]						

Revised model:
M = μ + negative binomial error $\mu = \alpha + \beta_W * W$



Pearson Chi Square (Dispersion parameter for negative binomial family) = 0.327

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
3.763	1	.052

Goodness of Fit[b]

	Value	df	Value/df
Deviance	10.771	9	1.197
Scaled Deviance	10.771	9	
Pearson Chi-Square	2.940	9	.327
Scaled Pearson Chi-Square	2.940	9	
Log Likelihood[a]	-74.427		
Akaike's Information Criterion (AIC)	152.855		
Finite Sample Corrected AIC (AICC)	154.355		
Bayesian Information Criterion (BIC)	153.650		
Consistent AIC (CAIC)	155.650		

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	36.414	1	.000
wolf	3.228	1	.072

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	4.530	.7507	3.059	6.002	36.414	1	.000
wolf	.103	.0572	-.009	.215	3.228	1	.072
(Scale)	1[a]						
(Negative binomial)	1						

Summary:
GLM assumptions met according to the res vs fits plot, however the dispersion parameter is very high (>9000) suggesting the data is extremely overdispersed
The best revised model had a negative binomial error with log link
This model lowered the dispersion parameter to less than 1
The type 1 error increased from significant (p = 0.000) to non-significant (p = 0.072).
This changes the decision reject the null hypotheses

7. Patterns in Fork-tailed Storm-petrel flight call behaviour was analyzed based on different weather and light variables.
Number of flight calls were tallied every 15 minutes from 1230 to 6 every night at 4 different sites on an island.

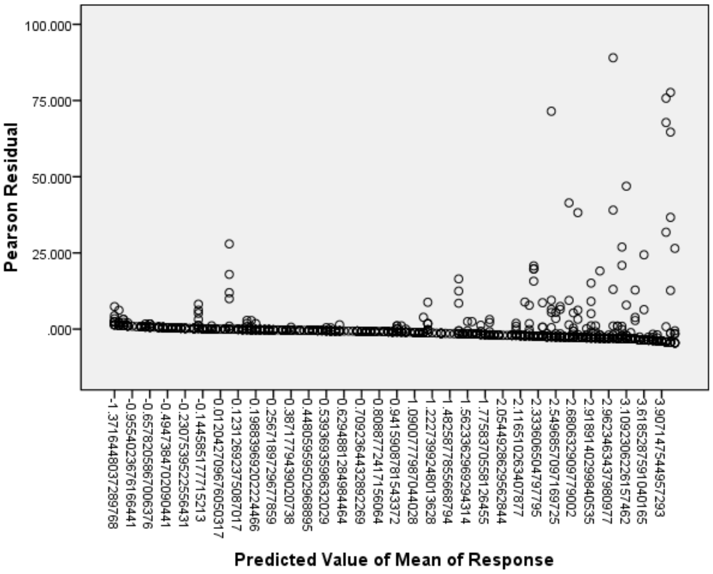
Data from RBuxton thesis

# Calls	**Time**	**Site**	Date	Moon Phase	Wind Speed	Wave Height	Cloud Cover	Precipitation
0	1:00:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
2	1:30:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
2	2:00:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
8	2:30:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
0	3:00:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no

1	3:30:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
3	4:00:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
0	4:30:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
0	5:00:00 AM	East	**6/18/2008**	1	6.31	1.15	light 100%	no
0	1:00:00 AM	East	**6/19/2008**	1	3.51	1.67	light 100%	showers
0	1:30:00 AM	East	**6/19/2008**	1	3.51	1.67	light 100%	showers
24	2:00:00 AM	East	**6/19/2008**	1	3.51	1.67	light 100%	showers
1	2:30:00 AM	East	**6/19/2008**	1	3.51	1.67	light 100%	showers

Number of flight calls per 15 minutes = F_c
Site = S
Moon Phase = M
Wind Speed = W_s
Wave Height = W_H
Cloud Cover = C
Precipitation = Ppt

$F_c = \mu$ + normal error $\mu = \alpha + \beta_S * S + \beta_M * M + \beta_{Ws} * W_S + \beta_{Wh} * W_h + \beta_C * C + \beta_{Ppt} * Ppt$



Pearson Chi Square (Dispersion parameter for gaussian family) = 46.58

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
2421.580	14	.000

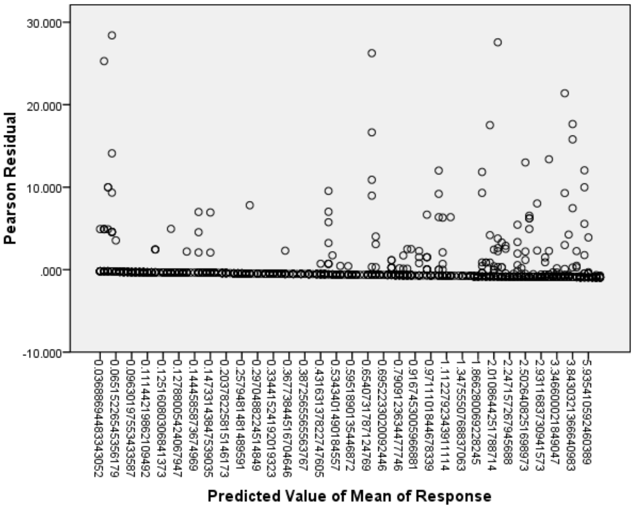
Goodness of Fit[b]

	Value	df	Value/df
Deviance	54637.782	1173	46.580
Scaled Deviance	54637.782	1173	
Pearson Chi-Square	54637.782	1173	46.580
Scaled Pearson Chi-Square	54637.782	1173	
Log Likelihood[a]	-28410.590		
Akaike's Information Criterion (AIC)	56851.180		
Finite Sample Corrected AIC (AICC)	56851.590		
Bayesian Information Criterion (BIC)	56927.381		
Consistent AIC (CAIC)	56942.381		

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	53.410	1	.000
Site	1756.705	2	.000
MoonPhase	233.394	3	.000
CloudCover	183.758	4	.000
Ppt	182.251	3	.000
WindSpeed	7.971	1	.005
Waveheight	.546	1	.460

Revised Model:
$Y = e^{\mu}$ + negative binomial error $\mu = \alpha + \beta_S * S + \beta_M * M + \beta_{Ws} * W_S + \beta_{Wh} * W_h + \beta_C * C + \beta_{Ppt} * Ppt$



Pearson Chi Square (Dispersion parameter for negative binomial family) = 6.89

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
814.723	14	.000

Goodness of Fit[b]

	Value	df	Value/df
Deviance	2101.355	1173	1.791
Scaled Deviance	2101.355	1173	
Pearson Chi-Square	8084.812	1173	6.892
Scaled Pearson Chi-Square	8084.812	1173	
Log Likelihood[a]	-1474.978		
Akaike's Information Criterion (AIC)	2979.956		
Finite Sample Corrected AIC (AICC)	2980.366		
Bayesian Information Criterion (BIC)	3056.157		
Consistent AIC (CAIC)	3071.157		

Tests of Model Effects

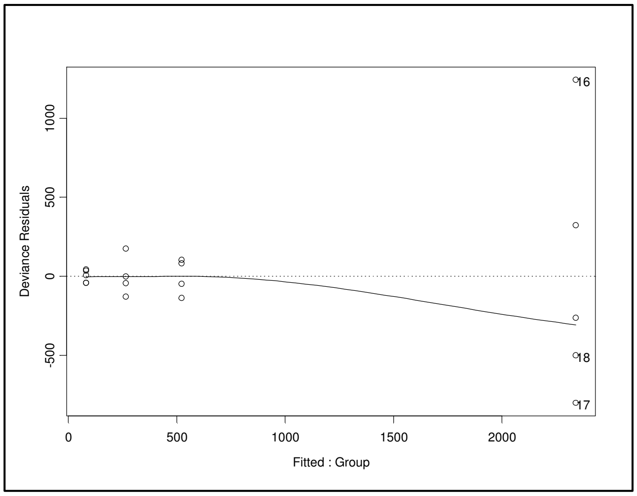
Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	15.623	1	.000
Site	440.626	2	.000
MoonPhase	54.114	3	.000
CloudCover	68.128	4	.000
Ppt	25.229	3	.000
WindSpeed	2.322	1	.128
Waveheight	.889	1	.346

Summary:
GLM assumptions not met according to the res vs fits plot (strong cone) and the dispersion parameter is very high (>40)
The residuals are homogenous and the data is extremely overdispersed
The best revised model had a negative binomial error with log link
This model lowered the dispersion parameter to 6 (still not low enough for AIC… must still revise the model)
The type 1 errors mostly stayed the same, but the type 1 error for wind speed increased from significant (p = 0.005) to non-significant (p = 0.128) and the type 1 error for wave height decreased from 0.460 to 0.346.

Analysis of the concentration of serum progesterone (*C*) in four groups of dogs (*G*), treated with estrogon, progesterone, estrogen plus progesterone, or an untreated control.
Data from Table 5 (Chapter 8.2) in Daniels (1995).
Model: *C* = μ + normal error μ = α + β*G*

Group (code)	Means
Untreated (0)	81.8
Estrogen (1)	265.25
Progesterone (2)	522.75
estrogen +progesterone(3)	2341.2

Residuals vs fits, normal error, identity link

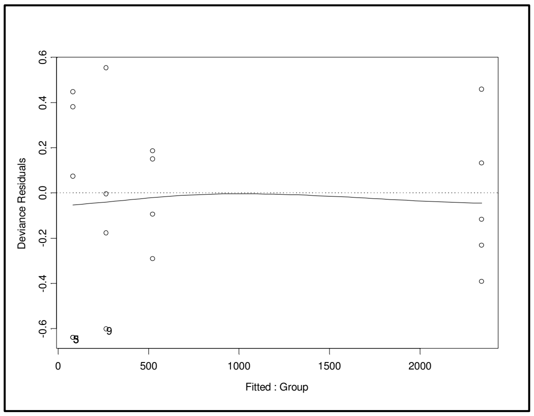


```
Coefficients reported by SPlus:
Source       Value   Std. Error  t value      Type I Error
(Intercept)   81.80   196.5991   0.4160751
    Group1  183.45   294.8987   0.6220781     0.2719
    Group2  440.95   294.8987   1.4952593     0.07852
    Group3  2259.40 278.0331   8.1263692     5.70E-05
```

(Dispersion Parameter for Gaussian family taken to be 193256.1)
Residual Deviance: 2705585 on 14 degrees of freedom Change in deviance of 1586083 on 3 df, p = 0

Residuals vs fits, gamma error, identity link

Revised Model: *C* = μ + gamma error μ = α + β*G*



```
Coefficients reported by SPlus:
Source       Value    Std. Error   t value      Type I error
(Intercept)  81.80      14.81643   5.520897
    Group1  183.45    55.72160   3.292260    0.002672
    Group2  440.95    106.89365 4.125128    0.0005151
    Group3  2259.40  424.32025 5.324752    5.36E-05
```

(Dispersion Parameter for Gamma family taken to be 0.1640403)
Residual Deviance: 2.464634 on 14 degrees of freedom Change in deviance of 26.41 on 3 df, p = 7.84E-06

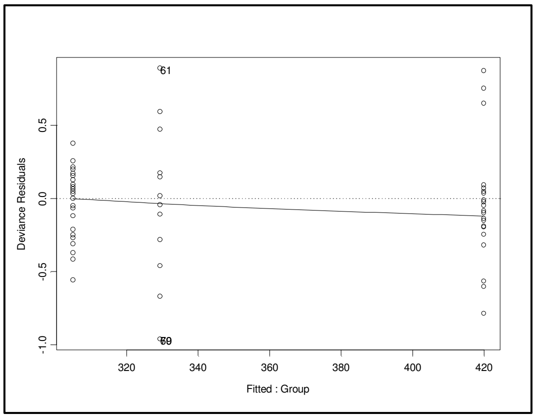
Summary: GLM assumptions not met because residuals strongly heterogeneous. Best revised model (most homogeneous errors) was Gamma error with identity link. Parameter estimates do not change but standard error decreases with a change in error structure.
Estimate of Type I error increased from 0.000409% to 0.000784%.

Analysis of serum concentration of an antigen (*C*) in three groups of children (*G*), autistic, normal, and mentally retarded.
Data from exercise 8.2.1 (Chapter 8.2) in Daniels (1995).

Model: *C* = μ + normal error μ = α + β*G*

Group (code)	Mean
Autistic (1)	419.913
Normal (2)	305
Mentally retarded (3)	329.3333

Residuals vs fits, normal error, identity link



Coefficients reported by SPlus:

```
Source       Value        Std. Error   t value      p-value
(Intercept)  419.91304   28.11987    14.932965
   Group2 -114.91304   36.63113    -3.137032  0.001261
   Group3 -90.57971     44.75685    -2.023818  0.02346
```

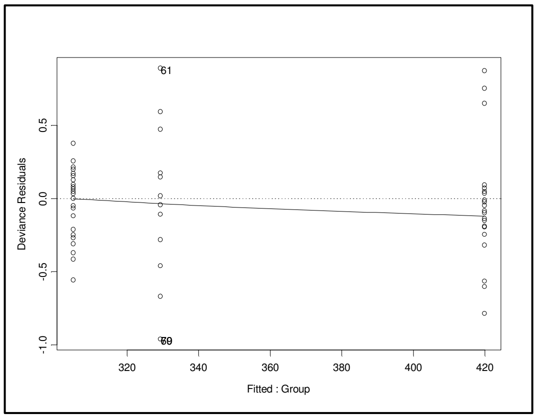
(Dispersion Parameter for Gaussian family taken to be 18186.72)
Residual Deviance: 1236697 on 68 degrees of freedom   Change in deviance of 185159.3 on 2 df, p = 0

Revised Model: *C* = μ + gamma error μ = α + β*G*

Residuals vs fits, gamma error, identity link

Coefficients reported by SPlus:

```
 Source       Value        Std. Error   t value      p- value
(Intercept)   419.91304   32.07837    13.090225
   Group2  -114.91304   37.51525    -3.063102   0.001569
   Group3  -90.57971     44.71646    -2.025646   0.02336
```



(Dispersion Parameter for Gamma family taken to be 0.1342251)
Residual Deviance: 9.065589 on 68 degrees of freedom Change in deviance of 1.459014 on 2 df, p = 0.4821

Summary: GLM assumptions not met because residuals strongly heterogeneous. Best revised model (not overdispersed but still has heterogenous errors) was Gamma error with identity link. Parameter estimates do not change but standard error increases with a change in error structure. Estimate of Type I error increased from 0.087% to 48.2%.

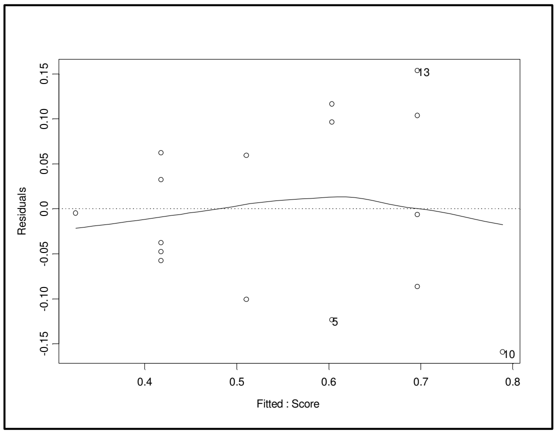
Analysis of rheologic measurements (*M*) in relation to end organ failure score (*S*).
Data from exercise 9.3.3 (Chapter 9.3) in Daniels (1995).
Model: *M* = μ + normal error μ = α + β*S*

Coefficients reported by SPlus:

```
Source           Value    Std. Error   t value    p-value
(Intercept)      -1.6636   0.7646       -2.1759
Measurement  7.4397   1.3302        5.5928    0.0001
```

Dispersion Parameter for Gaussian family taken to be 0.7495215
Residual Deviance: 10.4933 on 14 degrees of freedom

Residuals vs fits, normal error, identity link



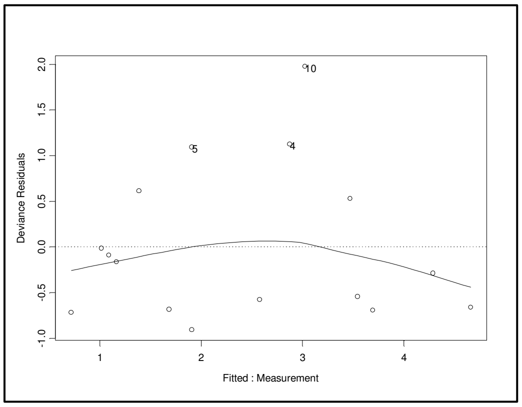
Model: *M* = μ + gamma error μ = α + β*S*

Coefficients reported by SPlus:

```
 Source          Value         Std. Error      t value     p-value
(Intercept)     -0.5456782  0.4308115   -1.266629
Measurement  2.4845291  0.6193907    4.011247 0.000643
```

Dispersion Parameter for Quasi-likelihood family taken to be 0.9712929. Residual Deviance: 13.59813 on 14 degrees of freedom

Residuals vs fits, gamma error, identity link



Summary: GLM assumptions not met because residuals strongly heterogeneous. Best revised model (most homogeneous errors) was Gamma error with identity link. Parameter and standard error decrease with the change in error structures. Estimate of Type I error increased from 0.01% to 0.06%.

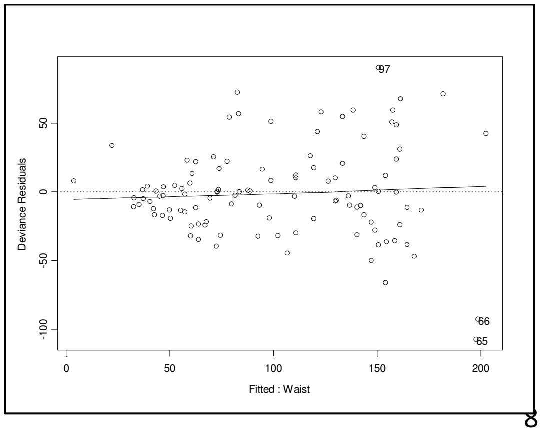
Analysis of deep abdominal adipose tissue (*AT*) in relation to Waist Circumference (*WC*).
Data from Table 9.3.1 (Chapter 9.3) in Daniels (1995).
Model: *AT* = μ + normal error μ = α + β*WC*

Coefficients reported by SPlus:

```
Source        Value         Std. Error     t value   p-value
(Intercept)  -215.9814  21.7962708  -9.9091
     Waist     3.458859  0.2346521    14.7403   0.00
```

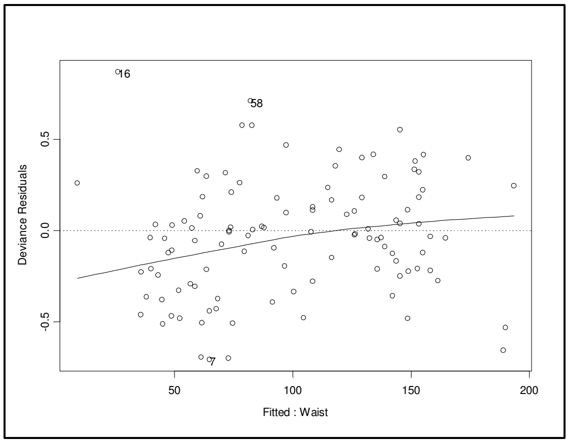
(Dispersion Parameter for Gaussian family taken to be 1093.29. Residual Deviance: 116982 on 107 degrees of freedom

Residuals vs fits, normal error, identity link



Model: *AT* = μ + gamma error μ = α + β*WC*

Residuals vs fits, gamma error, identity link



Coefficients reported by SPlus:

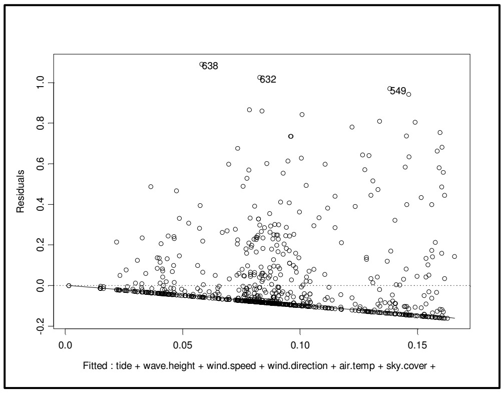
Source	Value	Std. Error	t value	p-value
(Intercept)	-194.8858	11.575748	-16.83570	
Waist	3.209408	0.153534	20.90356	7.1E-40

(Dispersion Parameter for Gamma family taken to be 0.104635. Residual Deviance: 11.13423 on 107 degrees of freedom.

Summary: GLM assumptions not met because residuals strongly heterogeneous. Best revised model (most homogeneous errors) was Gamma error with identity link. Parameter and standard error decrease with the change in error structure. Estimate of Type I error increased slightly from 0% to 7.1E

Analysis of haul-out % of harbor seals (*H*) against tide (*T*), wave intensity (*I*), wind speed (*WS*), wind direction (*D*), air temperature (*AT*), sky cover (*S*), and disturbance (*D*).
Data from Table 2 in Schneider & Payne (1983) J. Mamm. 64:518-520
Model: *H*= μ + normal error μ = α + β*T* + β*I* + β*WS* + β*D* + β*AT* + β*S* + β*D*

Residuals vs fits, normal error, identity link



Coefficients reported by SPlus:

Source	Value	Std. Error	t value	p-value
(Intercept)	0.1539	0.0232	6.6462	0.0000
tide	-0.0592	0.0774	-0.7646	0.4447
wave.height	-0.0154	0.0084	-1.8333	0.0671
wind.speed	-0.0014	0.0014	-1.0050	0.3152
wind.direction	-0.0523	0.0153	-3.4203	0.0007
air.temp	0.0013	0.0010	1.2180	0.2235
sky.cover	-0.0003	0.0018	-0.1450	0.8848
disturbance	-0.0468	0.0175	-2.6659	0.0078

(Dispersion Parameter for Gaussian family taken to be 0.03511. Residual Deviance: 30.30511 on 863 degrees of freedom

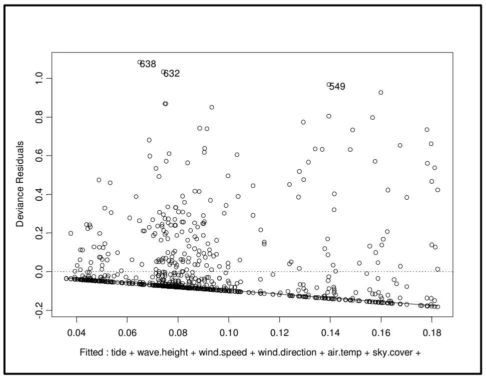
Model: *H*= μ + quasi error μ = α + β*T* + β*I* + β*WS* + β*D* + β*AT* + β*S* + β*D*

Coefficients reported by SPlus:

Source	Value	Std. Error	t value	p-value
(Intercept)	-1.9457	0.2277	-8.5463	
tide	-0.6544	0.7625	-0.8583	0.03433
wave.height	-0.2081	0.1321	-1.5751	0.07992
wind.speed	-0.0051	0.01519	-0.3295	0.3687
wind.direction	-0.5193	0.1442	-3.6018	0.001022
air.temp	0.01539	0.01114	1.3809	0.01755
sky.cover	-0.001351	0.01787	-0.07558	0.2611
disturbance	-0.5263	0.2649	-1.9867	0.2421

(Dispersion Parameter for Quasi-likelihood family taken to be 0.0351094)
Residual Deviance: 30.29926 on 863 degrees of freedom

Residuals vs fits, quasi error, log link



Summary: GLM assumptions not met because residuals strongly heterogeneous. Best revised model (most homogeneous errors) was Quasi error with log link. Parameter estimates cannot be compared because of change in model structure. Estimate of Type I error for tide decreased from 44.5% to 3.4 %, wave height increased from 6.7% to 7.9%, wind speed increased from 31.5% to 36.9%, wind direction increased from 0.07% to 0.10%, air temperature decreased from 22.4% to 1.8%, sky cover decreased from 88.5% to 26.1%, and disturbance increased from 0.78% to 24.2%.

Analysis of cytochrome P-450IA2 activities (*C*) in relation to urinary cotinine level (*U*) and cigarettes smoked per day (*S*).

Data from Table 10.3.1 (Chapter 10.3) in Daniels (1995).

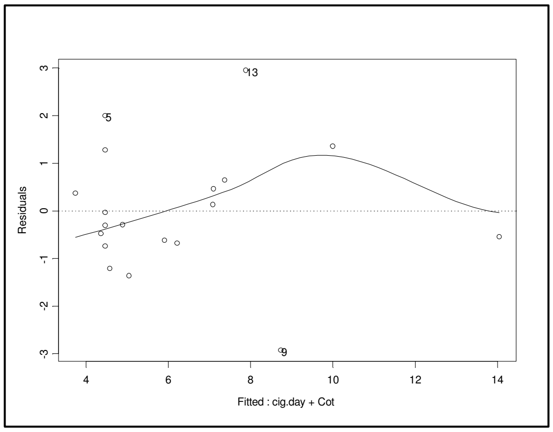
Model: $C = \mu$ + normal error $\mu = \alpha + \beta U + \beta S$

Coefficients reported by SPlus:

Source	Value	Std. Error	t value	p-value
(Intercept)	4.5234	0.5381	8.4066	0.0000
cig.day	-0.0517	0.0695	-0.7438	0.4678
Cot	0.1702	0.0301	5.6492	0.0000

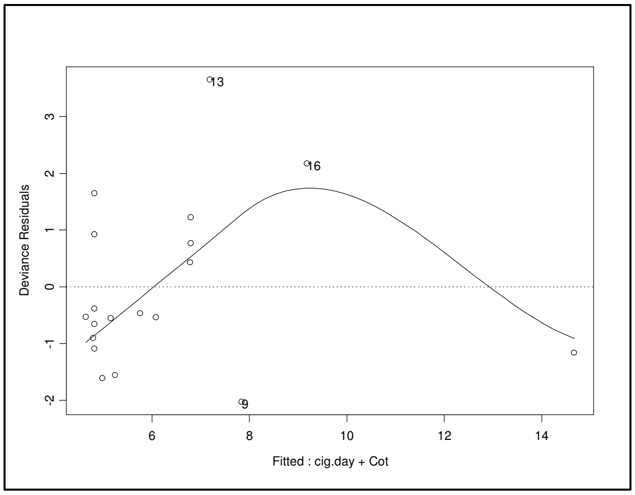
Dispersion Parameter for Gaussian family taken to be 1.932002. Residual Deviance: 30.91203 on 16 degrees of freedom

Residuals vs fits, normal error, identity link



Model: *C* = μ + quasi error μ = α + β*U* + β*S*

Residuals vs fits, quasi error, log link



Coefficients reported by SPlus:

Source	Value	Std. Error	t value	p-value
(Intercept)	1.5758	0.1078	14.62	
cig.day	-0.002666	0.01292	-0.2063	0.4195
Cot	0.01856	0.004429	4.19	0.000345

(Dispersion Parameter for Quasi-likelihood family

taken to be 2.376655. Residual Deviance: 38.02647

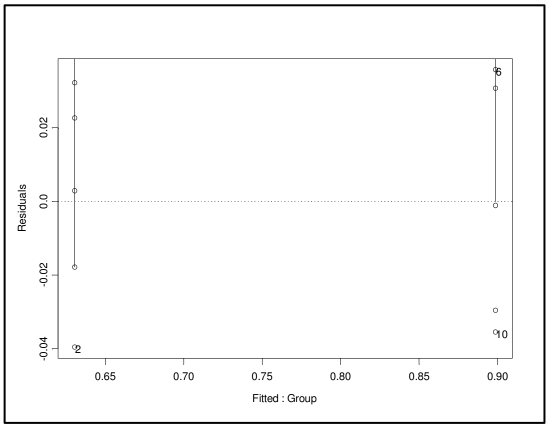
on 16 degrees of freedom.

Summary: GLM assumptions not met because residuals heterogeneous. Would chose the original model (guassian error with identity link) since a change in error structure caused an inceasre in the dispersion parameter but no change in heterogenaity. Parameter and standard error decrease with the change in error structure. Estimate of Type I error for cigarrettes smoked per day decreased from 46.8% to 42% and urine cotinine levels increased from 0% to 0.0345%.

Analysis of continuous variable Y (*Y*) in relation to group (G). No other description of data was provided Data from Exercise 13.2 (Chapter 13) in Zar (1996).

Model: log*Y* = μ + normal error μ = α + β*G*

Residuals vs fits, normal error, identity link



Source	Value	Std. Error	t value	p-value
(Intercept)	0.630656	0.01394751	45.21639	
Group	0.268164	0.01972476	13.59530	4.12E-07

(Dispersion Parameter for Gaussian family taken to be

0.00097. Residual Deviance: 0.0077813 on 8 degrees of

Freedom.

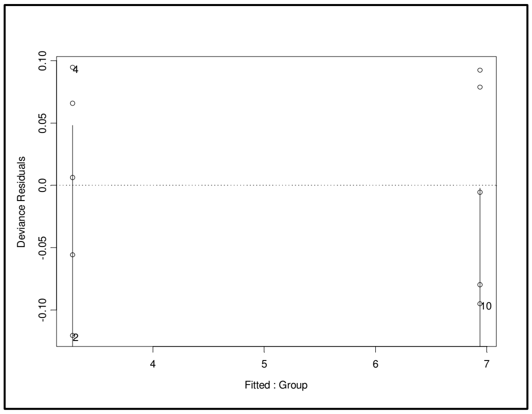
Model: $Y = \mu$ + gamma error $\mu = \alpha + \beta G$

Coefficients reported by SPlus:

```
Source      Value  Std. Error   t value      p-value
(Intercept)  3.28   0.1277044  25.68432
     Group   3.66   0.2988619  12.24646   9.18E-07
```

(Dispersion Parameter for Gamma family taken to be 0.0075794. Residual Deviance: 0.0611471 on 8 degrees of freedom.

Residuals vs fits, gamma error, identity link

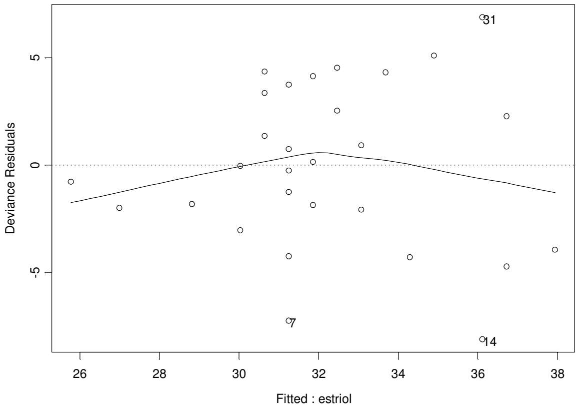


Summary: GLM assumptions were met but in the process the raw data was altered and thus became uninterpretable. The revised model is chosen because the assumptions are met without altering the raw data. Best revised model (most homogeneous errors) was Gamma error with identity link.
Parameter and standard error increased with the change in error structure.
Estimate of Type I error for group decreased from 0.09727% to 0.0000918%.

1. Analysis of estriol level (E) in pregnant women near term in relation to birthweight (BW).
Data from Table 11.1 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: BW = μ + normal error $\mu = \alpha + \beta E$

subject	Estriol (mg/24hr) (E)	Birthweight (g/100) (BW)
1	7	25
2	9	25
3	9	25
4	12	27
5	14	27
6	16	27
7	16	24
8	14	30
9	16	30
10	16	31
11	17	30
12	19	31
13	21	30
14	24	28
15	15	32
16	16	32
17	17	32
18	25	32
19	27	34
20	15	34
21	15	34
22	15	35
23	16	35
24	19	34
25	18	35
26	17	36
27	18	37
28	20	38
29	22	40
30	25	39
31	24	43



Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	21.52343	2.620417	8.213742	
estriol	0.608191	0.146812	4.142656	0.000135615

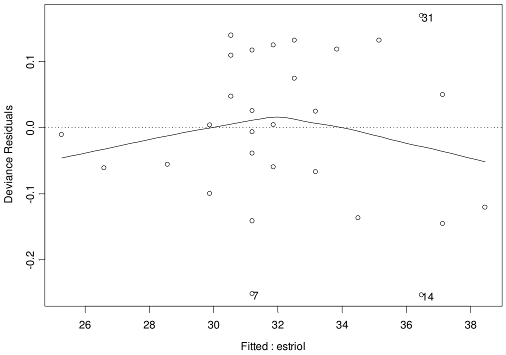
```
(Dispersion Parameter for Gaussian family taken to be
14.60088)
Residual Deviance: 423.4255 on 29 degrees of freedom
```

Revised Model: BW = e^{μ} + Gamma error $\mu = \alpha + \beta E$

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	20.65959	2.352858	8.780637	
estriol	0.658989	0.138593	4.754851	2.50609E-05

```
(Dispersion Parameter for Gamma family taken to be
0.0130887 )
Residual Deviance: 0.3932284 on 29 degrees of
freedom
```

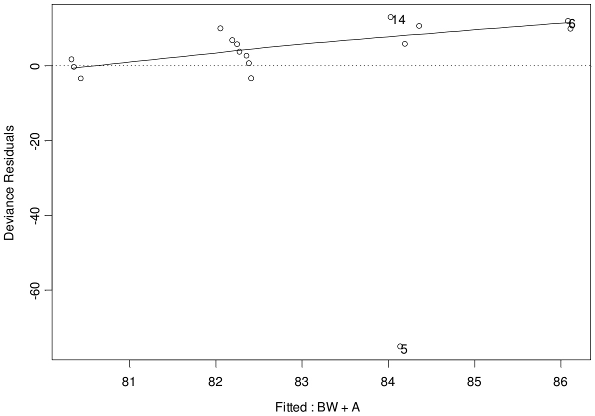


Summary:
GLM assumptions not met because residuals strongly heterogeneous.
Best revised model was Gamma error with an identity link.
Parameter estimates cannot be compared because of change in model structure.
Estimate of Type I error increased from 0.0135% to 2.5E^{-3}%

2. Analysis of birthweight (BW) and age (A) in infants in relation to systolic blood pressure (BP).
Data from Table 11.7 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: BP = μ + normal error μ = α + βBW + βA

subject	birthweight in oz (BW)	age in days (A)	systolic blood pressure (mm Hg) (BP)
1	135	3	89
2	120	4	90
3	100	3	83
4	105	2	77
5	130	4	9
6	125	5	98
7	125	2	82
8	105	3	85
9	120	5	96
10	90	4	95
11	120	2	80
12	95	3	79
13	120	3	86
14	150	4	97
15	160	3	92
16	125	3	88



Coefficients

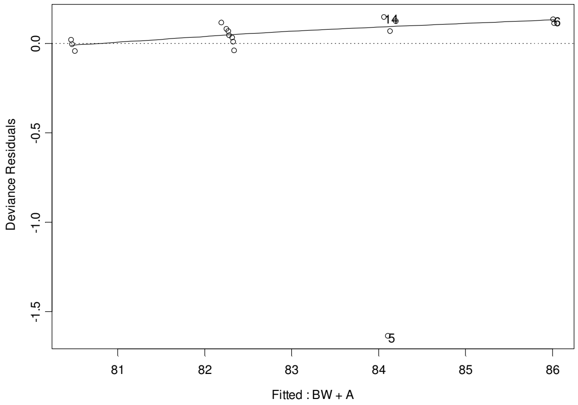
	Value	Std. Error	t value	Type I error
(Intercept)	77.17981889	40.6439303	1.8989261	
BW	-0.005486907	0.3079419	0.017818	0.493017745
A	1.918588571	6.100373	0.3145035	0.378888378

```
(Dispersion Parameter for Gaussian family taken to be 494.3636 )
Residual Deviance: 6426.727 on 13 degrees of freedom
```

Revised Model: BP = e^{μ} + Gamma error μ = α + βBW + βA

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	4.346812211	0.48316514	8.9965353	
BW	-2.7307E-05	0.00366074	0.0074594	0.497076773
A	0.022216599	0.07251975	0.3063524	0.381922499



```
(Dispersion Parameter for Gamma family taken to
be 0.0698628 )
Residual Deviance: 2.786742 on 13 degrees of
freedom
```

Summary:

GLM assumptions not met because residuals strongly heterogeneous.

Best revised model was Gamma error with a log link.

Parameter estimates cannot be compared because of change in model structure.

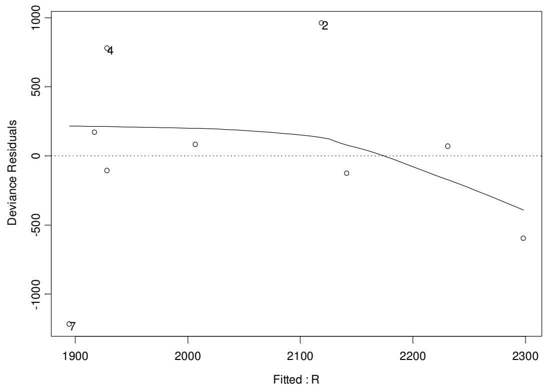
Estimate of Type I error increased from 49.3% to 49.7% for birthweight (BW)

Estimate of Type I error increased from 37.8% to 38.2% for age in days (A)

3. Analysis of reticulytes (R) in patients with aplastic anemia in relation to number of lymphocytes (L). Data from Table 11.28 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: L = μ + normal error $\mu = \alpha + \beta R$

subject	%reticulates (R)	lymphocytes (per mm2) (L)
1	3.6	1700
2	2	3078
3	0.3	1820
4	0.3	2706
5	0.2	2086
6	3	2299
7	0	676
8	1	2088
9	2.2	2013



Coefficients

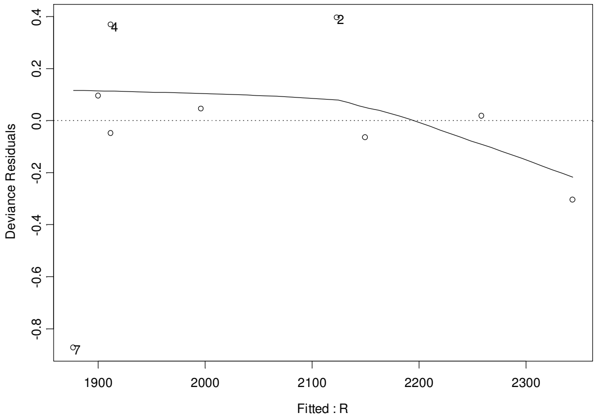
	Value	Std. Error	t value	Type I error
(Intercept)	1894.818	348.4739	5.437475	
R	112.114	184.7485	0.606847	0.281554693

```
(Dispersion Parameter for Gaussian family taken to be 490818.2 )
Residual Deviance: 3435727 on 7 degrees of
freedom
```

Revised Model: $L = e^{\mu}$ + Gamma error $\mu = \alpha + \beta R$

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	7.537414	0.176155	42.78848	
R	0.061692	0.093391	0.660581	0.265005924



```
(Dispersion Parameter for Gamma family taken to
be 0.1254212 )
Residual Deviance: 1.166087 on 7 degrees of
freedom
```

Summary:

GLM assumptions not met because residuals strongly heterogeneous.

Best revised model was Gamma error with a log link.

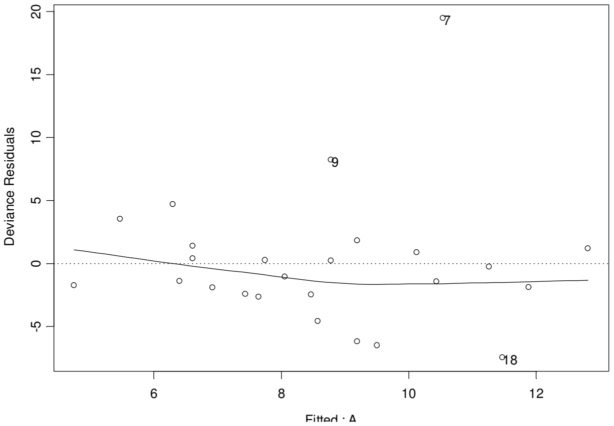
Parameter estimates cannot be compared because of change in model structure.

Estimate of Type I error reduced from 28.1% to 26.5% for % reticulytes

4. Analysis of age (A) of patients discharged from a Pennsylvania hospital to duration of hospital stay (D). Data from Table 2.11 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: $D = \mu$ + normal error $\mu = \alpha + \beta A$

subject	duration of hospital stay (D)	Age (A)
1	5	30
2	10	73
3	6	40
4	11	47
5	5	25
6	14	82
7	30	60
8	11	56
9	17	43
10	3	50
11	9	59
12	3	4
13	8	22
14	8	33
15	5	20
16	5	32
17	7	36
18	4	69
19	3	47
20	7	22
21	9	11
22	11	19
23	11	67
24	9	43
25	4	41



Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	4.337617	2.524058	1.718509	
A	0.103356	0.055229	1.871416	0.03703084

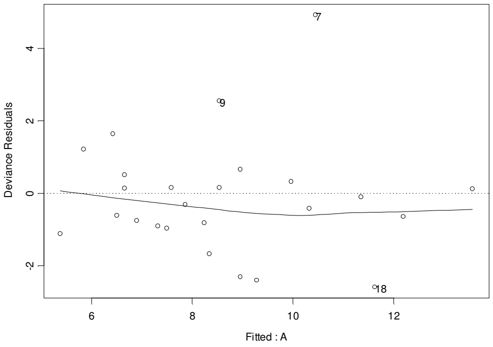
```
 (Dispersion Parameter for Gaussian family taken to be 29.58245 )
Residual Deviance: 680.3964 on 23 degrees of freedom
```

Revised Model: $D = e^{\mu}$ + Gamma error $\mu = \alpha + \beta A$

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	1.633372	0.267814	6.098903	
A	0.011899	0.00586	2.030503	0.02701171

```
(Dispersion Parameter for Gamma family taken to
be 0.3330445 )
Residual Deviance: 6.787932 on 23 degrees of
freedom
```



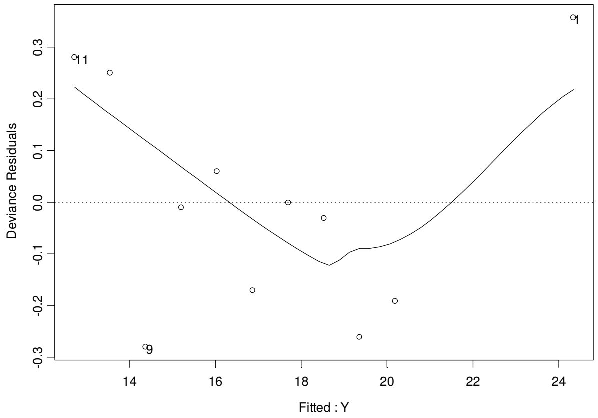
Summary:

GLM assumptions not met because residuals strongly heterogeneous.
Best revised model was Gamma error with a log link.
Parameter estimates cannot be compared because of change in model structure.
Estimate of Type I error reduced from 3.7% to 2.7% for for age

5. Analysis of year (Y) in the US from 1960 to 1979 to infant mortality rates per 1000 live births (IM). Data from Table 11.30 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: IM = μ + normal error $\mu = \alpha + \beta Y$

year	infant mortality
1960	26
1965	24.7
1970	20
1971	19.1
1972	18.5
1973	17.7
1974	16.7
1975	16.1
1976	15.2
1977	14.1
1978	13.8
1979	13



Coefficient

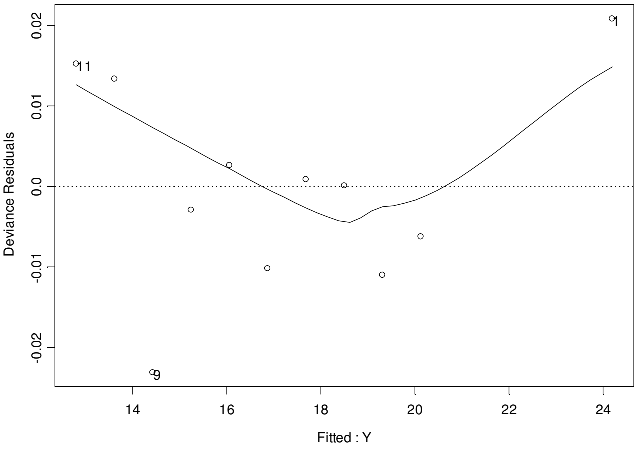
	Value	Std. Error	t value	Type I error
(Intercept)	1655.727	35.75682	46.30521	
Y	-0.83022	0.018117	45.82503	2.8075E-12

```
(Dispersion Parameter for Gaussian family taken to be 0.0540092 )
Residual Deviance: 0.4860829 on 9 degrees of freedom
```

Revised Model: IM = e^{μ} + Gamma error $\mu = \alpha + \beta Y$

Coefficient				
	Value	Std. Error	t value	Type I error
(Intercept)	1622.647	38.40046	42.25593	
Y	-0.81346	0.019444	41.83557	6.3506E-12

```
(Dispersion Parameter for Gamma family taken
to be 0.0001842 )
Residual Deviance: 0.001658 on 9 degrees of
freedom
```



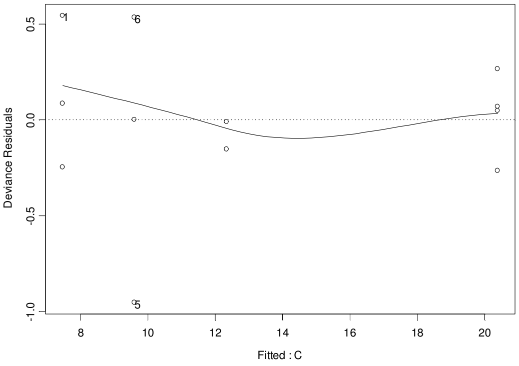
Summary:

GLM assumptions not met because residuals strongly heterogeneous.
Best revised model was Gamma error with an identity link.
Parameter estimates cannot be compared because of change in model structure.
Estimate of Type I error increased from $2.8E^{-10}$% to $6.3E^{-10}$% for for year (Y)

6. Analysis of cars per hour (C) at a particular street corner to CO concentrations (CO).
Data from Table 11.29 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: CO = μ + normal error $\mu = \alpha + \beta C$

cars/hr (x10^3) (C)	CO concentrations (CO)
1	9
1	6.8
1	7.7
1.5	9.6
1.5	6.8
1.5	11.3
2	12.3
2	11.8
3	20.7
3	19.2
3	21.6
3	20.6

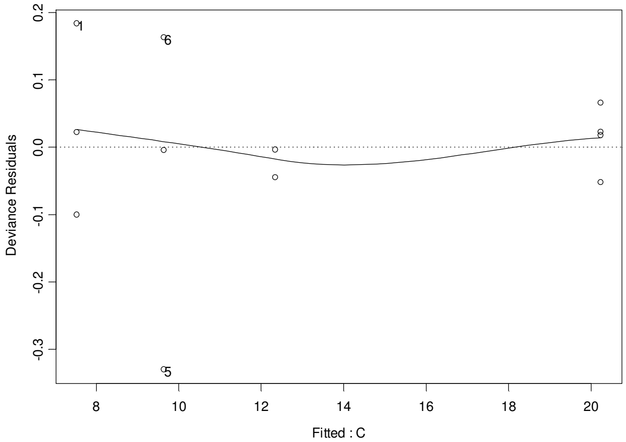


Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	0.116712	1.228386	0.095012	
C	6.638275	0.580412	11.43718	2.29183E-07

```
(Dispersion Parameter for Gaussian family taken to be 2.603784 )
Residual Deviance: 26.03784 on 10 degrees of freedom
```

Revised Model: CO = e^{μ} + Gamma error $\mu = \alpha + \beta C$



Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	1.524925	0.100505	15.17269	
C	0.49427	0.047488	10.40823	5.5012E-07

```
(Dispersion Parameter for Gamma family taken to be 0.0174304 )
Residual Deviance: 0.1897167 on 10 degrees of freedom
```

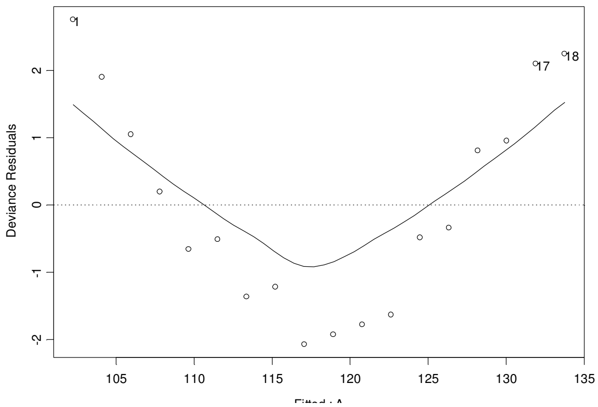
Summary:
GLM assumptions not met because residuals strongly heterogeneous.
Best revised model was Gamma error with a log link.
Parameter estimates cannot be compared because of change in model structure.
Estimate of Type I error increased from $2.3E^{-5}$% to $5.5E^{-5}$% for Cars/hr (X 10^3) (C).

7. Analysis of boys aged (A) 1 to 18 to the observed 90th percentile of systolic blood pressure (SPB) in a single year.

Data from Table 11.31 in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: SPB = μ + normal error $\mu = \alpha + \beta A$

Age (A)	SBP
1	105
2	106
3	107
4	108
5	109
6	111
7	112
8	114
9	115
10	117
11	119
12	121
13	124
14	126
15	129
16	131
17	134
18	136



Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	100.3922	0.79289945	126.614	
A	1.853457	0.07325143	25.30268	1.23983E-14

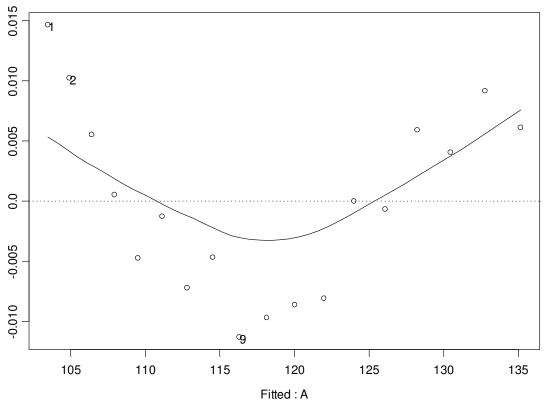
```
(Dispersion Parameter for Gaussian family taken to be 2.599716 )
Residual Deviance: 41.59546 on 16 degrees of freedom
```

Revised Model: SBP = e^{μ} + gamma error $\mu = \alpha + \beta A$

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	4.619251	0.00529768	871.939	
A	0.015593	0.00048942	31.86031	3.32436E-16

```
(Dispersion Parameter for Gamma family
taken to be 0.0001161 )
Residual Deviance: 0.0018533 on 16 degrees
of freedom
```



Summary:

GLM assumptions not met because residuals strongly heterogeneous.

Best revised model was Gamma error with a log link.

Parameter estimates cannot be compared because of change in model structure.

Estimate of Type I error decreased from 1.24^{-12} % to 3.32^{-14} % for Age (A).

8. Analysis of age (A) in years and height (H) in inches and personal smoking (PS) to the level of pulmonary function (SPB).

Data from FEV.DAT on data disk in Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

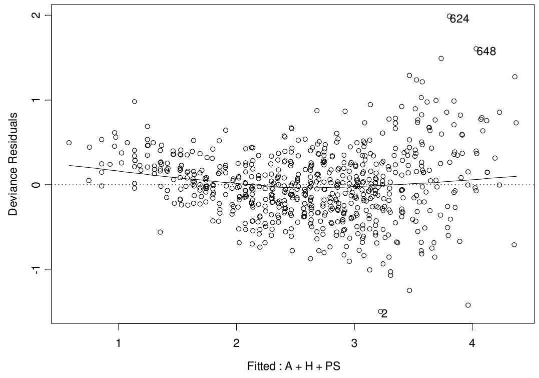
Model: SPB = μ + normal error $\mu = \alpha + \beta A + \beta H + \beta PS$

Data set contains 654 individuals (see data disk)

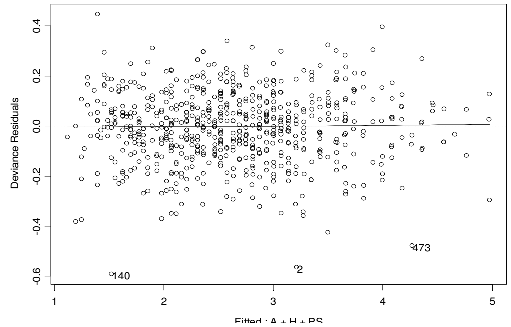
Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	-4.61600695	0.223883258	-20.617919	
A	0.05974105	0.009563412	6.246834	3.78595E-10
H	0.10909474	0.004719598	23.115259	4.85604E-87
PS	-0.11023193	0.060017457	1.836664	0.03335798

```
(Dispersion Parameter for Gaussian family taken
to be 0.175512 )
Residual Deviance: 114.0828 on 650 degrees of
freedom
```



Revised Model: SBP = e^{μ} + Gamma error



μ = α + βA + βH + βPS

```
(Dispersion Parameter for Gamma family taken to be 0.0201668 )
Residual Deviance: 13.54681 on 650 degrees of freedom
```

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	-1.97082496	0.07589034	-25.969378	
A	0.021707	0.003241737	6.6961	2.31722E-11
H	0.04392164	0.001599816	27.454187	4.5518E-111
PS	-0.04511184	0.020344287	2.217421	0.013469722

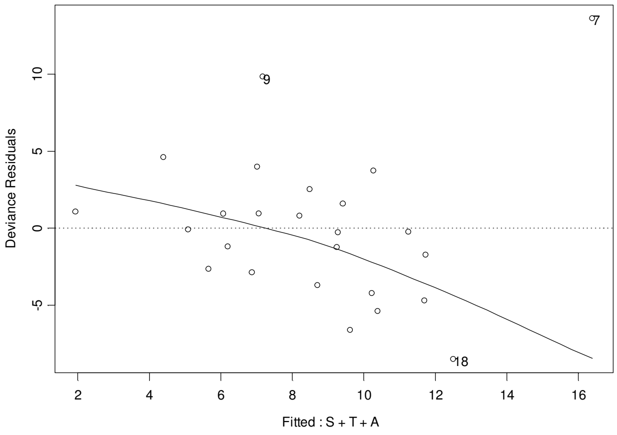
Summary:

GLM assumptions not met because residuals strongly heterogeneous.
Best revised model was Gamma error with a log link.
Parameter estimates cannot be compared because of change in model structure.
Estimate of Type I error reduced from $3.78E^{-8}$% to $2.31E^{-9}$ % for Age (A).
Estimate of Type I error reduced from $4.85E^{-85}$ % to $4.55E^{-109}$% for height (H).
Estimate of Type I error reduced from 3.3 % to 1.35% for personal smoking (PS).

9. Analysis of sex (S) and first temperature reading following admission (T) of patients discharged from a Pennsylvania hospital to duration of hospital stay (D).
Data from HOSPITAL.DAT Rosner, B.A. (1995). Fundamentals of Biostatistics 4th Edition. Duxbury Press.

Model: D = μ + normal error μ = α + βS + βT + βA

subject	duration of hospital stay (D)	age (A)	sex (S)	temperature (T)
1	5	30	2	99
2	10	73	2	98
3	6	40	2	99
4	11	47	2	98.2
5	5	25	2	98.5
6	14	82	1	96.8
7	30	60	1	99.5
8	11	56	2	98.6
9	17	43	2	98
10	3	50	1	98
11	9	59	2	97.6
12	3	4	1	97.8
13	8	22	2	99.5
14	8	33	2	98.4
15	5	20	2	98.4
16	5	32	1	99
17	7	36	1	99.2
18	4	69	1	98
19	3	47	1	97
20	7	22	1	98.2
21	9	11	1	98.2
22	11	19	1	98.6



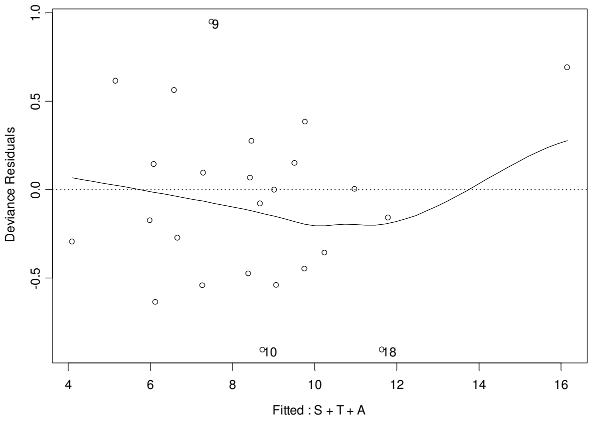
23	11	67	2	97.6
24	9	43	2	98.6
25	4	41	2	98

Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	-341.242	168.8159	-2.02138	
S	-1.37902	2.133161	0.646466	0.26249019
T	3.50283	1.709556	2.048971	0.02658314
A	0.151741	0.05766	2.631666	0.00779694

```
(Dispersion Parameter for Gaussian family taken to be 26.91583 )
Residual Deviance: 565.2323 on 21 degrees of freedom
```

Revised Model: $D = e^{\mu}$ + Gamma error $\mu = \alpha + \beta S + \beta T + \beta A$



Coefficients

	Value	Std. Error	t value	Type I error
(Intercept)	-28.8869	17.1271	-1.68662	
S	-0.04763	0.216418	0.220105	0.41395742
T	0.309175	0.173442	1.782586	0.04455912
A	0.015107	0.00585	2.582413	0.00868582

```
(Dispersion Parameter for Gamma family
taken to be 0.2770439 )
Residual Deviance: 5.771594 on 21 degrees
of freedom
```

Summary:

GLM assumptions not met because residuals strongly heterogeneous.

Best revised model was Gamma error with a log link.

Parameter estimates cannot be compared because of change in model structure.

Estimate of Type I error increased from 26.2% to 41.4% for sex (S).

Estimate of Type I error increased from 2.65 % to 4.44% for temperature (T).

Estimate of Type I error increased from 0.77% to 0.86% for age (A).

Analysis of final examination data in relation to first and second practice exam scores

Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P076.txt

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

μ = F = final examination scores

$B_1 = P_1$ = practice score #1

$B_2 = P_2$ = practice score #2

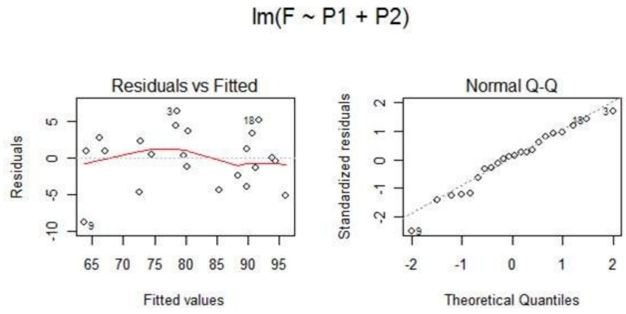
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14.5005	9.2356	-1.570	0.13290
P1	0.4883	0.2330	2.096	0.04971
P2	0.6720	0.1793	3.748	0.00136

Dispersion parameter for gaussian family taken to be 15.62263

Residual deviance: 296.83 on 19 degrees of freedom

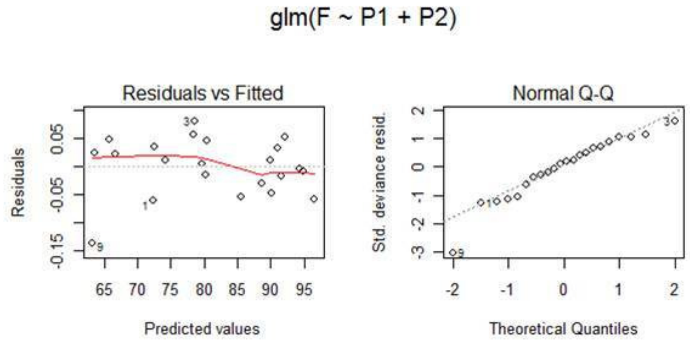
AIC: 127.68



Revised Model: $\beta = \mu$ + gamma error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.2612	9.0531	-1.907	0.07180
P1	0.4985	0.2480	2.010	0.05884
P2	0.6956	0.1965	3.541	0.00218

Dispersion parameter for Gamma family taken to be 0.002698299
Residual deviance: 0.052623 on 19 degrees of freedom
AIC: 130.78



Summary: GLM residual were somewhat homogeneous and normally distributed, and GzLM improved them slightly. GzLM with gamma error and identity link improved dispersion and deviance residuals. Parameters could not be interpreted due to change in error structure and link. Estimates of Type I error increased from 4.9 to 5.8 % for P1 and 0.1 to 0.2% for P2.

Analysis of salary data in relation to years of experience, level of education and the presence of management responsibility

Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P122.txt

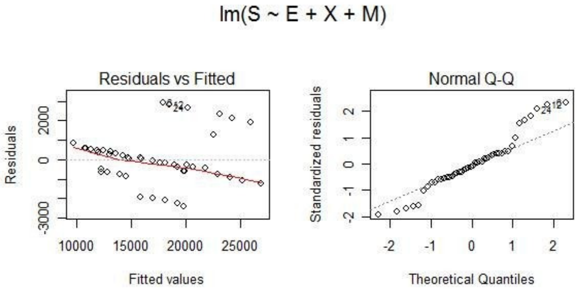
Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2 + B_3 X_3$

μ = S = salary
B_1 = X = experience (yrs)
B_2 = E = education (1 = high school diploma, 2=bachelor degree, 3=advanced degree)
B_3= M = management (1=management responsibility, 0 = no management responsibility)

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6963.48	665.69	10.460	2.88e-13 ***

X	570.09	38.56	14.785	< 2e-16 ***
E	1578.75	262.32	6.018	3.74e-07 ***
M	6688.13	398.28	16.793	< 2e-16 ***

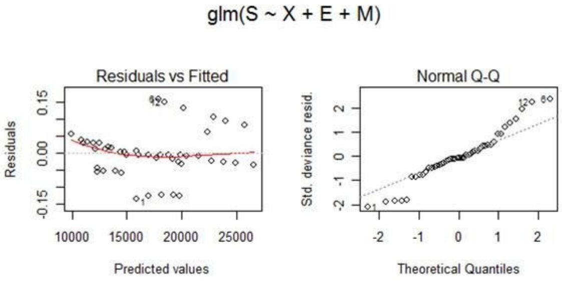
Dispersion parameter for gaussian family taken to be 1723415
Residual deviance: 72383410 on 42 degrees of freedom
AIC: 796.91



Revised Model: $\beta = \mu$ + gamma error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7405.22	537.62	13.774	< 2e-16 ***
X	555.55	35.56	15.621	< 2e-16 ***
E	1443.31	219.71	6.569	6.03e-08 ***
M	6495.50	372.10	17.456	< 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.004901762)
Residual deviance: 0.20413 on 42 degrees of freedom
AIC: 785.7



Summary: GLM residuals not homogeneous. GzLM with gamma error and identity link improved the residuals, although still not homogeneous. GzLM greatly improved dispersion and residual deviance. Parameters could not be interpreted due to change in error structure and link. Type I errors increased for all predictors, however the errors are all very small (<0.00001).

Analysis of temperature with wind chill factor in relation to the actual temperature of still air and wind speed

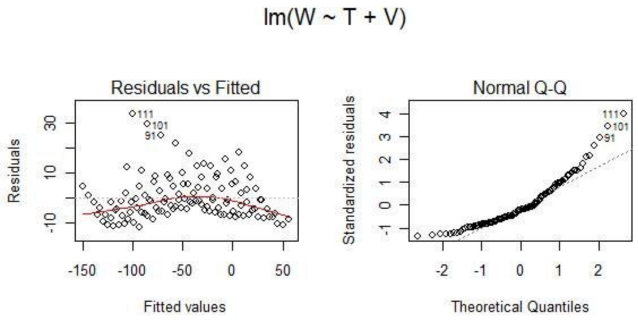
Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P175.txt

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

μ = W = Temperature with wind chill (°C)
B_1 = T = Temperature of still air (°C)
B_2 = V = Wind speed

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.05664	1.71960	-5.267	6.4e-07 ***
T	1.41867	0.02301	61.661	< 2e-16 ***
V	-1.10545	0.05530	-19.989	< 2e-16 ***

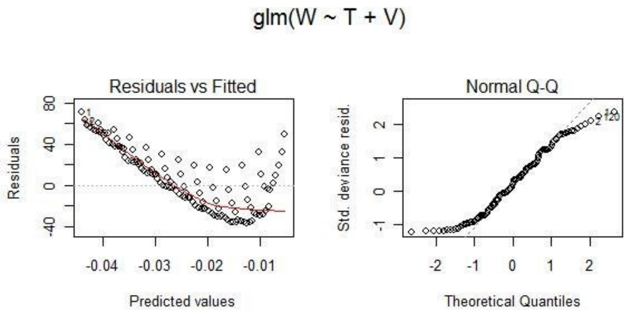
Dispersion parameter for gaussian family taken to be 75.69753
Residual deviance: 8856.6 on 117 degrees of freedom
AIC: 864.72



Revised Model: $\beta = 1/\mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.922e-02	2.148e-03	-13.599	< 2e-16 ***
T	-3.079e-04	3.496e-05	-8.805	1.39e-14 ***
V	1.120e-04	3.117e-05	3.592	0.000482 ***

Dispersion parameter for gaussian family taken to be 905.4633
Residual deviance: 105957 on 117 degrees of freedom
AIC=1162.5



Summary: GLM residuals heterogenous with a clear patter, however revised GzLM with Gaussian error and inverse link made residuals worse. The GzLM also greatly increased the dispersion factor and residual deviance. Type I errors increased, however still very close to zero. Accept the original model.

Analysis of an equal opportunity achievement index in relation to family status, peer influence and education opportunity indices

Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P224.txt

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

μ = ACHV = Achievement

B_1 = FAM = Family

B_2 = PEER= Peer

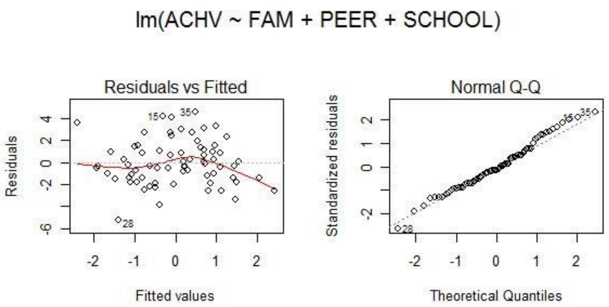
B_3 = SCHOOL= School

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.06996	0.25064	-0.279	0.781
FAM	1.10126	1.41056	0.781	0.438
PEER	2.32206	1.48129	1.568	0.122
SCHOOL	-2.28100	2.22045	-1.027	0.308

Dispersion parameter for gaussian family taken to be 4.285958

Residual deviance: 282.87 on 66 degrees of freedom

AIC: 306.41



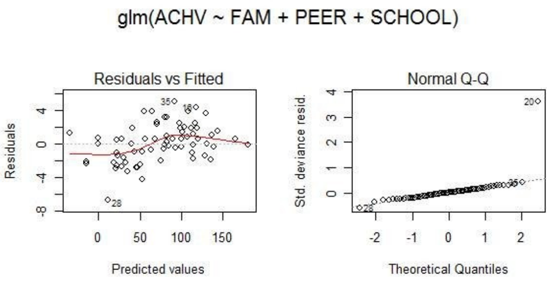
Revised Model: $\beta = 1/\mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.84	707.88	0.097	0.923
FAM	64.64	655.64	0.099	0.922
PEER	98.18	984.21	0.100	0.921
SCHOOL	-112.65	1130.17	-0.100	0.921

Dispersion parameter for gaussian family taken to be 154.6012

Residual deviance: 342.43 on 66 degrees of freedom

AIC: 319.78



Summary: Residuals of GLM somewhat homogeneous and normally distributed. GzLM with Gaussian error and inverse link did not affect the residuals greatly, but highlighted an influencial point, thus the data were no longer normally distributed. Standard errors increased substantially. Type I errors increased to almost 100% for all predictors from 43, 12 and 31 % for FAM, PEER and SCHOOL respectively. Accept original model.

Analysis of supervisor performance data from in relation to 6 survey responses categories regarding the supervisors attributes (see variables below).

Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P056.txt

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 + B_5 X_5 + B_6 X_6$

μ = Y= Overall rating of the job being done by the supervisor

$B_1 = X_1$ = Handles employee complaints

$B_2 = X_2$ = Does not allow special privileges

$B_3 = X_3$ = Opportunity to learn new things

$B_4 = X_4$ = Raises based on performance

$B_5 = X_5$ = Too critical of poor performance

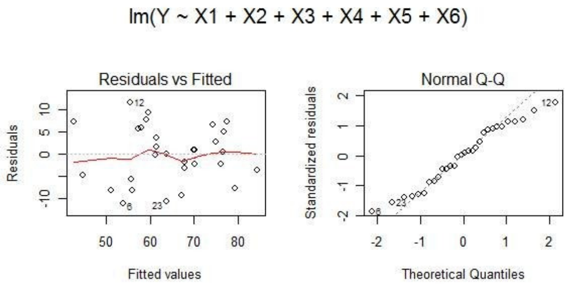
$B_6 = X_6$ = Rate of advancing to better jobs

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.78708	11.58926	0.931	0.361634
X1	0.61319	0.16098	3.809	0.000903 ***
X2	-0.07305	0.13572	-0.538	0.595594
X3	0.32033	0.16852	1.901	0.069925
X4	0.08173	0.22148	0.369	0.715480
X5	0.03838	0.14700	0.261	0.796334
X6	-0.21706	0.17821	-1.218	0.235577

Dispersion parameter for gaussian family taken to be 49.95654

Residual deviance: 1149 on 23 degrees of freedom

AIC: 210.5



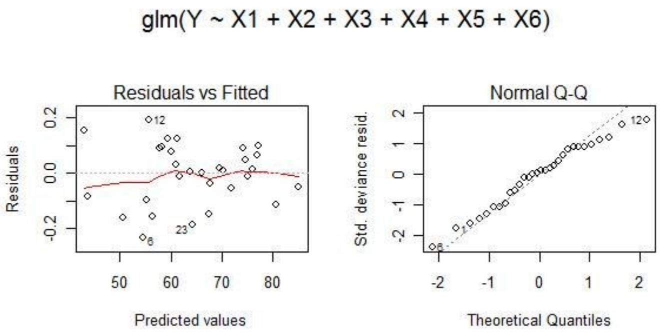
Revised Model: $\beta = \mu$ + gamma error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 + B_5 X_5 + B_6 X_6$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.17428	11.05906	1.191	0.24570
X1	0.62822	0.18426	3.409	0.00240 **
X2	-0.05152	0.15468	-0.333	0.74211
X3	0.29985	0.17843	1.680	0.10640
X4	0.07403	0.24174	0.306	0.76217
X5	0.03744	0.14905	0.251	0.80389
X6	-0.28253	0.19691	-1.435	0.16479

Dispersion parameter for Gamma family taken to be 0.01419289

Residual deviance: 0.33455 on 23 degrees of freedom

AIC: 215.30



Summary: GLM residuals appeared somewhat homogeneous but not normally distributed. GzLM with gamma error and identity link produced very similar plots, however greatly improved dispersion and residual deviance. The estimates, standard errors and p-values did not change greatly, but 4/6 Type I errors increased and 2/6 decreased.

Analysis of student weight data from estimates of their height, age and sex collected from personal information

Data from: Chatterjee, S. and A. Hadi. 2006. Regression Analysis by Example, Fourth Edition. Wiley Series. Link: http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P148.txt

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2 + B_3 X_3$

μ = W= weights

B_1 =H = heights

B_2 =A = age

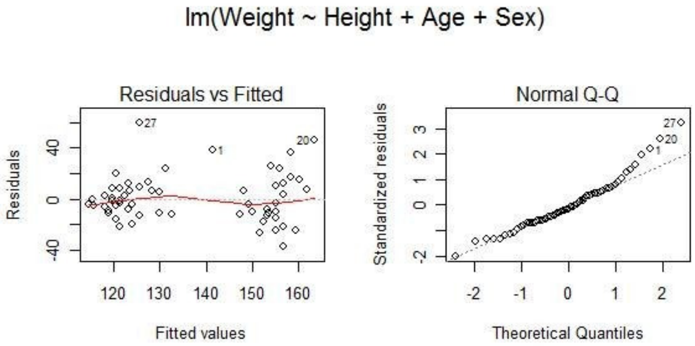
B_3 =S = Sex (male or female)

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	53.1514	56.2748	0.944	0.3490
Height	1.7025	0.7847	2.170	0.0343 *
Age	-0.8176	1.5738	-0.520	0.6054
Sex	-25.2002	5.7713	-4.366	5.5e-05 ***

Dispersion parameter for gaussian family taken to be 353.8688

Residual deviance: 19817 on 56 degrees of freedom

AIC: 528.27

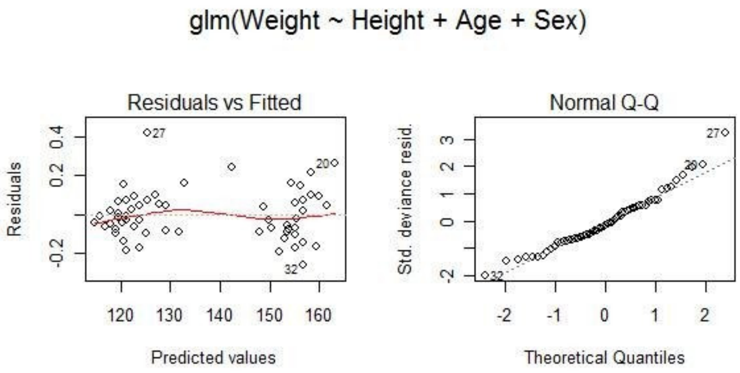


Revised Model: $\beta = \mu$ + gamma error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2 + B_3 X_3$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	57.5184	52.3090	1.100	0.2762

Height	1.5971	0.7490	2.132	0.0374 *
Age	-0.6606	1.3574	-0.487	0.6284
Sex	-26.0045	5.5428	-4.692	1.79e-05 ***

Dispersion parameter for Gamma family taken to be 0.01755271
Residual deviance: 0.9262 on 56 degrees of freedom
AIC: 519.22



Summary: GLM residuals are not homogeneous or normally distributed. GzLM with gamma error and identity link somewhat improved normality but residuals remained similar. The GzLM greatly improved dispersion and residual deviance. Standard error was not affected greatly, and all Type I errors increased slightly.

Analysis of crowberry productivity data in relation to birch height and site

Data from: Siegwart-Collier

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

μ = P= Productivity (counts/m2)

B_1 =Begl ht = Birch height

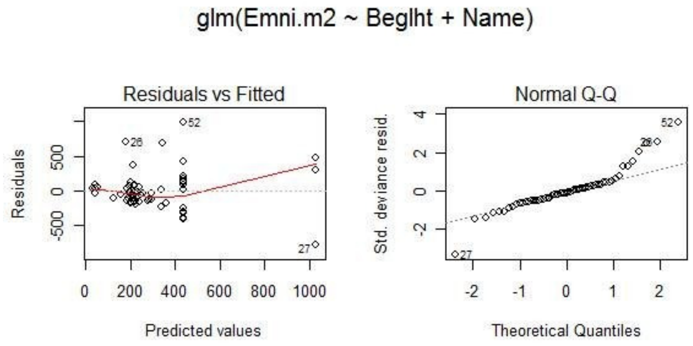
B_2 =Site = site location across eastern Arctic

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	161.15	129.81	1.241	0.219810
Beglht	-11.58	9.48	-1.222	0.227155
GeorgeRiver	277.11	127.31	2.177	0.033890 *
T.TorrBay	149.15	122.22	1.220	0.227633
T.Wakeham	868.18	210.12	4.132	0.000126 ***

Dispersion parameter for gaussian family taken to be 81893.27

Residual deviance: 4422237 on 54 degrees of freedom

AIC: 841.69



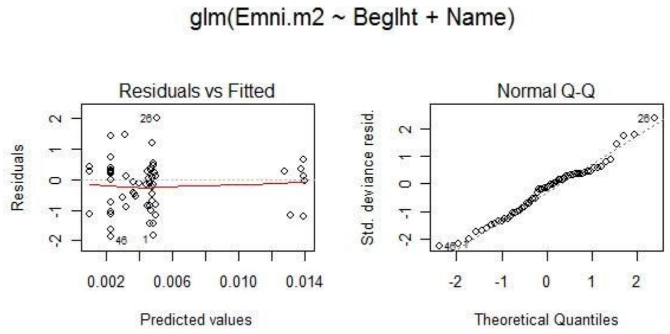
Revised Model: $\beta = 1/\mu$ + gamma error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0127894	0.0044916	2.847	0.00622 **
Beglht	0.0001105	0.0000982	1.125	0.26568
GeorgeRiver	-0.0105647	0.0044845	-2.356	0.02214 *
T.TorrBay	-0.0089924	0.0045071	-1.995	0.05108 .
T.Wakeham	-0.0118179	0.0045175	-2.616	0.01151 *

Dispersion parameter for Gamma family taken to be 0.7436435

Residual deviance: 41.049 on 54 degrees of freedom

AIC: 776.96



Summary: GLM residuals not homogenous or normally distributed. Gamma error and inverse link improved normality and expanded the spread of residuals, but still heterogeneous. GzLM greatly improved dispersion and residual deviance. Parameters could not be interpreted due to change in error structure and link. Type I error changed significantly as well: increased for birch height from 22% to 26%, decreased for George River from 2% to 3%, decreased from 22% to 5% for Torr Bay, and increased from ~0 to 1% for Wakeham Bay.

Analysis of bilberry fruit presence/absence in relation to birch height and site

Data from: Siegwart-Collier

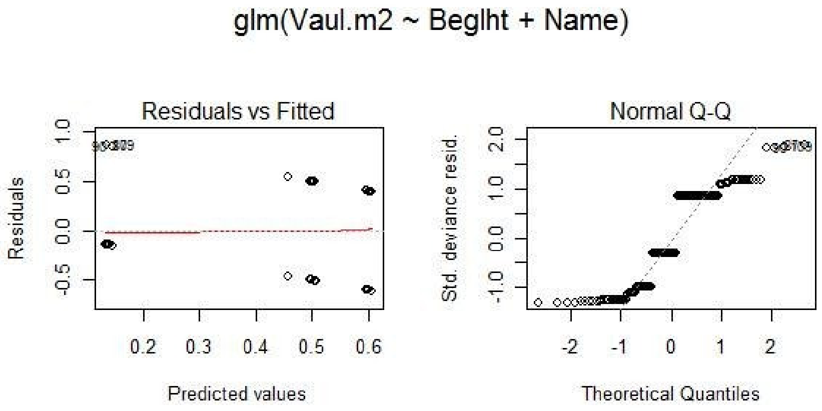
Original Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1 + B_2 X_2$

μ = Vaul/m^2= Presence/absence of bilberry (1 vs. 0)

B_1 =Begl ht = Birch height

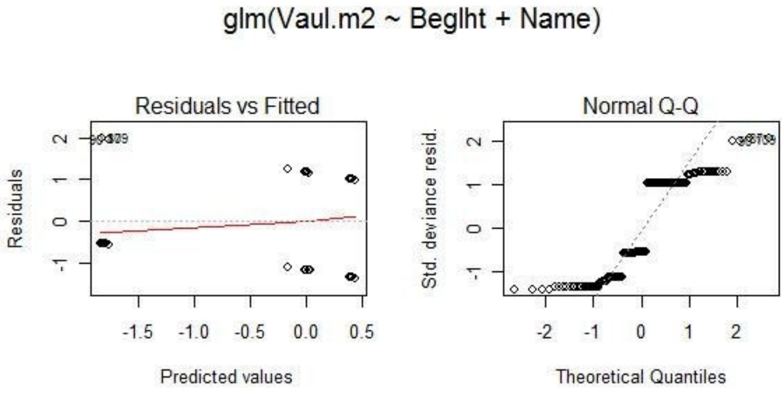
B_2 =Site = site location across eastern Arctic

```
Coefficients:     Estimate       Std. Error     t value        Pr(>|t|)
(Intercept)       0.5065390      0.1581145      3.204          0.00174 **
Beglht            -0.0008817     0.0106533      -0.083         0.93418
T.GeorgeRiver     -0.3588859     0.1650445      -2.174         0.03165 *
T.TorrBay         0.0995650      0.1501211      0.663           0.50846
T.Wakeham         -0.0482057     0.1854153      -0.260         0.79532
Dispersion parameter for gaussian family taken to be 0.2250871
Residual deviance: 26.785 on 119 degrees of freedom
AIC: 173.88
```



Revised Model: β = μ + binomial error; μ = $\beta_0 + B_1 X_1 + B_2 X_2$

```
Coefficients:     Estimate       Std. Error     z value        Pr(>|z|)
(Intercept)       0.036736       0.699717       0.053          0.9581
Beglht            -0.004953      0.053294       -0.093         0.9260
T.GeorgeRiver     -1.801881      0.798561       -2.256         0.0240 *
T.TorrBay         0.403054       0.635188       0.635          0.5257
T.Wakeham         -0.203790      0.810824       -0.251         0.8016
Dispersion parameter for binomial family taken to be 1
Residual deviance: 153.46 on 119 degrees of freedom
AIC: 163.46
```



Summary: GLM residuals neither homogenous nor normally distributed. GzLM with binomial error and logit link did not improve residuals. Parameters could not be compared because of change in error structure and link. Dispersion and residual deviance improved greatly by GzLM. Type I error increased slightly for all predictor variables. Thus, GzLM is much better fit for the data.

Analysis of plant species richness in relation to post-fire organic matter thickness

Data from: Siegwart-Collier

Model: $\beta = \mu$ + normal error; $\mu = \beta_0 + B_1 X_1$

μ = ROM=Residual organic matter depth (cm)

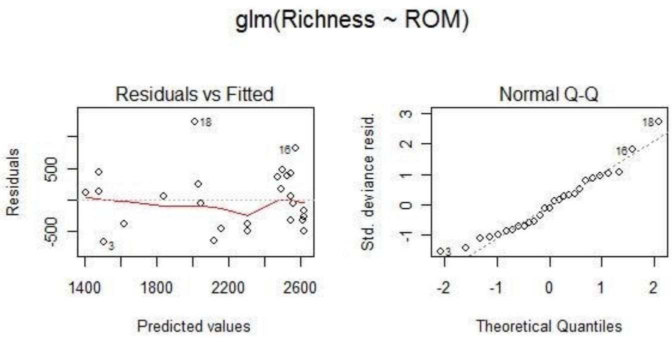
B_1 =Species Richness

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2641.7	126.9	20.819	< 2e-16 ***
ROM	-181.6	39.4	-4.609	0.000103 ***

Dispersion parameter for gaussian family taken to be 215892.0

Residual deviance: 5397299 on 25 degrees of freedom

AIC: 412.17



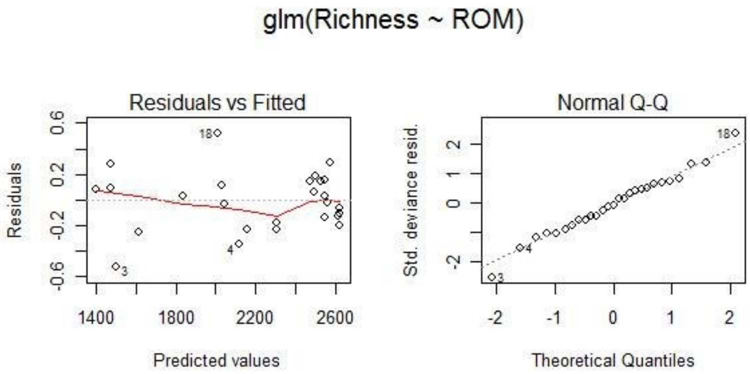
Revised Model: $\beta = \mu$ + gamma error; $\mu = \beta_0 + B_1 X_1$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2642.17	149.58	17.663	1.24e-15 ***
ROM	-181.82	35.53	-5.118	2.75e-05 ***

Dispersion parameter for Gamma family taken to be 0.05031124

Residual deviance: 1.2522 on 25 degrees of freedom

AIC: 413.84



Summary: GLM residuals neither homogeneous not normally distributed. GzLM with gamma error and identity link compressed some of the spread in the residuals and improved the normally plot. GzLM also improved dispersion and residual deviance. Estimates and standard error could not be compared due to change in error distribution and link, however Type I error decreased from 0.1% to 0%.