Residual Analysis and Data Transformations: Important Tools in Statistical Analysis

George C.J. Fernandez
Department of Agricultural Economics, University of Nevada-Reno, Reno, NV 89557-0107

Analysis of variance (ANOVA) is a commonly used statistical analysis in agricultural experiments. Additivity, variance homogeneity, and normality are often considered prerequisites for ANOVA (Cochran, 1943; Eisenhart, 1947). The interpretation of ANOVA is valid when the random effects are independently distributed according to a normal distribution with zero mean and an unknown but fixed variance (Kempthorne, 1952; Scheffe, 1959; Steel and Torrie, 1980). Failure to meet one or more of these assumptions affects the significance levels and the sensitivity of the F test (Gomez and Gomez, 1984; Kempthorne, 1952; Little and Hills, 1978) Thus, strong deviations from one or more of the assumptions must be checked and corrected before the statistical analysis and interpretation of the results.

Discrepancies of many kinds between an assumed model and the data can be detected by studying the error component or residuals (Anscombe and Tukey, 1963; Emerson and Stoto, 1983). The residuals are the deviation from the observed and the predicted values according to the assumed model. If the assumptions about the validity of the model are valid, a residual plot (scatter plot between the residuals and the predicted values) will have a random distribution. If the residual plot has an unexplained systematic pattern, then the ANOVA model is not appropriate. Residual plots can be used to detect the violation of assumptions in ANOVA, such as variance heterogeneity (unequal variance), nonnormality, and auto-correlated error (nonindependence), and the presence of outliers. Thus, it is crucial to examine the residuals before interpreting the data.

Violation of assumptions in ANOVA

Nonadditivity, variance heterogeneity, and nonnormality. The additivity requirement implies that the block and treatment effects should be additive. For example, in a randomized complete-block design, the differences observed for any two treatments should be the same in every block except for the experimental error component (Finney, 1989). Equality of variance refers to the variance of the error component, which should be the same for all treatments and all blocks. Nonadditivity implies that every individual error component should be derived from a normal frequency distribution. The existence of a relationship between the size of the residuals and the predicted value indicates that the variance of the residuals is functionally related to the mean. This type of variance heterogeneity is usually associated with nonadditivity and/or nonnormally associated data (Box et al., 1978; Gomez and Gomez, 1984), and a wedge or fan shaped pattern is seen in the residual plots (Emerson and Stoto, 1983). Ott (1988) proposed an alternate test, the Hartley’s test for homogeneity of variance, to verify the assumption on equality of variance. The residuals also can be examined for normality and homogeneity by drawing normal probability and box plots by treatments, respectively, using the PROC UNIVARIATE in SAS (SAS Institute, Inc., 1988).

Auto-correlations. One essential requirement of ANOVA is that the “error” components of the observed responses should be independent of each other. The lack of auto-correlation assumption is secured by a proper randomization. If the “error” components are not independent, the validity of the F test of significance can be seriously impaired (Finney, 1989; Sokal and Rohlf, 1987). There is no simple adjustment or transformation to overcome the lack of independence of error. The basic design of the experiment or the way in which the analysis is performed must be changed to deal with this problem. A cyclic pattern in the residual plot is an indication for auto-correlation, nonindependent error (Fernandez, 1990a; Gomez and Gomez, 1984). If auto-correlated errors are observed in residual plots in special experimental layouts, a repeated measure of ANOVA (Fernandez, 1991) or moving mean covariance analysis (Fernandez, 1990a) may be appropriate to make adjustments for auto-correlation.

Outliers. Outliers or the influential observation can be detected by plotting the standardized residuals against predicted values. If the absolute value of the standardized residual is >2.5, that observation can be treated as an outlier (Freund and Littell, 1986; SAS, 1988). If the residual analysis revealed the presence of influential or extreme observations (outliers), check first whether the outlier is due to a recording error. Do not seek an excuse for the possible rejection of the outlier, but investigate the possibility that the outlier may have unexplained implication worthy of further investigation.

Data transformations

Data are transformed to make them conform more closely to the assumptions underlying the ANOVA (Bartlett, 1947). It is undertaken with three objectives: i) to make the error variances more nearly homogenous; ii) to improve additivity; iii) to produce a more nearly normal error distribution (Finney, 1989). The transformation of data implies the replacement of each observation by some simple function of its magnitude, followed by a standard ANOVA. Thus, the original data are transformed to a new scale, resulting in data that are expected to satisfy the assumptions of additivity, normality, and homogeneity of variance. Because a common transformation scale is used for all observations in the data, treatments cannot be altered, and the mean comparisons remain valid.

A convenient rule of thumb for deciding whether transformation would be effective is to find the ratio between the largest and the smallest data values. Transformation could be helpful when the ratio is large, >20 (Emerson and Stoto, 1983). Tests of homogeneity of variance for two or more samples can be performed using Bartlett’s test of homogeneity (Snedecor and Cochran, 1956). SAS codes for performing Bartlett’s homogeneity are available in SAS/STAT sample library examples (SAS Institute, Inc., 1988). If variance stabilization is the primary objective of transformation, then efforts should be made to find the transformation that best achieves it. Logarithmic, square-root, and arcsin conversions are the most commonly used transformations for ANOVA of problem data (Gomez and Gomez, 1984; Sokal and Rohlf, 1987; Steel and Torrie, 1980).
Log transformation: When the treatment standard deviation (S) is small relative to the treatment mean and the treatment effects are multiplicative, a log transformation is recommended (Steel and Torrie, 1980; Gomez and Gomez, 1984). Data consisting of “whole” numbers that cover a wide range of values (number of diseased plants per plot, number of pods per square meter) often need a log transformation.

Square-root transformation: It is appropriate for data consisting of small whole numbers from rare events, e.g., number of insects captured in a trap. For such data, the variance is proportional to the mean. If the data contain zeros, 0.5 or 1 is added to the original data before performing square-root or log transformations, respectively.

Arcsin transformation: A typical characteristic of percentages based on counts is that the variances of means near 0% and 100% tend to be smaller than the variances of means near the middle range (30% to 70%) (Finney, 1989). Thus, percentage data based on counts are discrete and have a binomial distribution. The arcsin or the angular transformation is appropriate for these types of data obtained from a count and is expressed as a decimal fraction or percentage. If the percentages range near the middle range (30% to 70%), the arcsin transformation is not needed. Arcsin transformations convert the percentages to angles whose sines are square-roots of percentages expressed as decimals. Arcsin (Y) = (1/sin) (Y^0.5)

where the Ys are the decimal fractions. If the data include values of 0% and 100%, these values are replaced by (1/4n) and [100 - (1/4n)], respectively, where n is the total number of units upon which the percentage data were based. Tables of arcsin values can be obtained from statistical text books (Gomez and Gomez, 1984; Steel and Torrie, 1980) or by using the ARSIN option in SAS (1988). Arcsin transformation is inappropriate for unconstrained percentages involving rate of growth increases that might have values > 100% or even negative values (Finney, 1989).

Power transformation. When the functional relationship between the treatment means and variances is unknown, it is possible to use the data to estimate the suitable transformation. Box and Cox (1964) proposed the power transformation where:

Ym = \left \{ \begin{array}{ll} Y^\lambda \lambda > 0 \\ \log Y \lambda < 0 \end{array} \right.

Y is the transformed response, A vary over the range of – 2 to 2, and residual sums of squares SSE (I) should be minimal. Little et al. (1978) described a relatively simple method to determine the suitable power transformation using the data in question. The following steps describe this method: 1) Estimate the treatment means (\overline{Y}_m), single factor or treatment combination means (two or more factors) and treatment standard deviations (S). 2) Calculate the logs of the Ys and the logs of the Ss. 3) Plot log(S) on log(\overline{Y}^\lambda) and examine for a linear relationship. A strong nonlinear relationship indicates that a simple power transformation is not appropriate for such data, and distribution-free, nonparametric methods such as the Kruskal-Wallis test, Wilcoxon rank-sum test, and Mann-Whitney U test (SAS, 1988) should be considered as alternative methods (Kempthorne, 1952; Sokal and Rohlf, 1987). 4) Regress log(S) on log(\overline{Y}) and test for a significant linear relationship. If the regression is not significant (P > 0.05), log transformation usually is not necessary. A significant regression (P < 0.05) indicates the data should be transformed and the regression coefficient estimated. 5) Estimate the power (\lambda) of the transformation by subtracting the regression coefficient estimated. 5) Estimate the power (\lambda) of the transformation by subtracting the regression coefficient (\beta) from 1. The value of the power (\lambda) indicates the appropriate transformation. For example, if \beta approximately equals 2, then \lambda = 1 - \beta = -1. Thus, the appropriate transformation form would be reciprocals. Some commonly used transformations and their power (\lambda) values are:

\begin{align*}
\text{Transformations} & \\
\lambda & \\
2.0 & \text{Reciprocal} \\
1.0 & \text{Log} \\
0.66 & \text{Cubic root} \\
0.5 & \text{Square root} \\
\end{align*}

The value of the power (\lambda) indicates the appropriate transformation. For example, if \beta approximately equals 2, then \lambda = 1 - \beta = -1. Thus, the appropriate transformation would be reciprocals. Some commonly used transformations and their power (\lambda) values are:

\begin{align*}
\text{Transformations} & \\
\lambda & \\
2.0 & \text{Reciprocal} \\
1.0 & \text{Log} \\
0.66 & \text{Cubic root} \\
0.5 & \text{Square root} \\
\end{align*}

If the appropriate transformation is estimated by Box’s method using the data in question, one df is usually taken away from the error df in the ANOVA, since the same computations and graphical display of residuals, further, choosing the appropriate transformation is not straightforward, and without examining the residuals, it is difficult to confirm the appropriateness of the transformation.

The use of SAS software in statistical analysis is rapidly increasing with the availability of command-driven SAS for personal computers (PC-SAS). In a recent study, PC-SAS was identified as one of the more versatile and easy-to-use software programs available on the market (Milliken and Remenga, 1989). In addition, PC-SAS provides powerful data management and is flexible in formatting output (Fernandez, 1990b). With SAS available to perform the residual analysis and to estimate the appropriate transformation due to extreme observations in the data, ranks of the observations can be used in ANOVA (Conover and Iman, 1981). Many nonparametric statistical methods, Wilcoxon rank sum test, Kruskal-Wallis K-sample test, and Friedman’s two-way analysis using ranks are often better than original observations (Quade, 1969).

Tests of significance and mean separation should be carried out on the transformed data. Care should be taken in interpreting means of transformed data. To keep the metric interpretation to the original scale, the transformed means and associated confidence intervals can be back-transformed (Gomez and Gomez, 1984) and reported within parenthesis along with the transformed means.

Analysis of rating scale data. Rating scales can be defined as a series of numbers representing degree of intensity of some characteristic based on visual or sensory estimate. (Little, 1985). A comprehensive account of how to analyze rating scale data can be found in Little (1985). Checking for violations of assumptions in ANOVA by residual analysis is very important but is practiced less commonly in agricultural research since it involves additional computations and graphical display of residuals. Further, choosing the appropriate transformation is not straightforward, and without examining the residuals, it is difficult to confirm the appropriateness of the transformation.
priate power transformation, the horticulturist can easily perform residual analyses. Therefore, the purpose of this paper is to emphasize the importance of residual analysis and to present PC-SAS program statements to perform residual analysis and to estimate the appropriate power transformation.

Example

Data on mungbean (*Vigna radiata* L. Wilko) plant height at 50% flowering for eight genotypes grown in two separate experiments in the spring and the fall season at the Asian Vegetable Research and Development Center in Taiwan were used here as a worked example to emphasize the importance of the residual analysis. The design was a randomized complete block with three replications. The experiment was conducted in two distinct growing seasons, summer and fall, and a combined ANOVA over season was carried out. Statistical analysis was carried out using the PC-SAS (SAS, 1988), and the SAS program statements for this and the subsequent analysis (log, square-root, inverse transformed) are presented in Figs. 1 and 2.

The “wedge-shaped” residual plot (Fig. 3a) for the untransformed data clearly indicated the presence of the unequal variance or heterogeneity. The residual plots of the commonly used transformations (log and square-root) (Fig. 3 b-c) also indicated the presence of heterogeneity even after the transformation. Therefore, the statistical significance levels and the sensitivity of the F test for the untransformed log and square root-transformed mungbean plant height data are biased.

The method of Box et al. (1978) for power transformation was used to estimate the appropriate power transformation for this dataset. The SAS statements are given in Fig. 2. The means (Y) and the standard deviations (S) for genotype x season combinations were estimated. The log(S) was regressed on the log(Y) The regression coefficient \( \beta = 2.142 \) is significant (P > 0.0067). From the regression coefficient, the power (\( \lambda \)) was estimated:

\[ \lambda = 1 - \beta_1 = 1 - 2.142 = -1.142 \]

Thus, Box’s method on the plant height data suggested that an inverse transformation would be appropriate.

The results of the ANOVA on the untransformed log and square root, and inverse transformed plant height data were compared (Table 1). One df was taken away from the error df for the inverse transformed data since the same data had been used to choose the appropriate power transformation (Box et al., 1978). The appropriate SAS statements for making adjustments in the error df are given in Fig. 2. The random distribution between the residuals and the predicted values of the inverse transformed data in the residual plot (Fig. 3d) clearly show that the inverse transformation removed the heterogeneity of variance.

The P values for the original and the transformed plant height data indicated the significance levels for the main effects of season and genotype were in agreement for the original and the transformed data. However, the interaction between season x genotype was significant (\( P = 0.0267 \)) for the inverse transformed data, whereas the P values on the original, log, and square root transformation indicated that the interaction was not significant (\( P > 0.39 \)). Because of the heterogeneity of variance, the differential responses of genotypes across the two seasons were not detected in the original, log, or the square-root-transformed data. This example clearly indicates that if the assumption of homogeneity of variance is not met in ANOVA, both the significance levels and the sensitivity of the F tests are biased.

This example clearly shows that significance levels in ANOVA may be biased if the data violate the assumption of the ANOVA. Furthermore, this may lead one to draw incorrect conclusions from the analysis. Therefore, checking data for the violation of the ANOVA assumptions before discussing the results is very critical. Residual analysis is a powerful tool to detect the problems associated with the violation of the ANOVA assumptions. The commonly used transformations such as the log and the square-root conversions may not be appropriate for every data set. The method of power transformation according to Box et al. (1978) is recommended to estimate the appropriate transformation. Although the residual analysis and the estimation of the power transformation needs additional calculations, with the use of the PC-SAS program presented here, the proper NOVA can be performed without tedious calculations.

Table 1. Comparisons of ANOVA statistics (P values) for the original and the transformed data for mungbean plant heights obtained in two separate plantings.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>HT</th>
<th>SQRT(HT)</th>
<th>Log(HT)</th>
<th>1/HT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasons (S)</td>
<td>4</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Replicates(season)</td>
<td>7</td>
<td>0.924</td>
<td>9.917</td>
<td>0.906</td>
<td>0.866</td>
</tr>
<tr>
<td>Genotype (G)</td>
<td>7</td>
<td>0.0014</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>S x G</td>
<td>7</td>
<td>0.6007</td>
<td>0.5756</td>
<td>0.3993</td>
<td>0.0267</td>
</tr>
</tbody>
</table>

The df for the inverse-transformed data is one minus the error df since the same data have been used to choose the appropriate power transformation according to Box et al. (1978).

The P values for the original and the transformed plant height data indicated the significance levels for the main effects of season and genotype were in agreement for the original and the transformed data. However, the interaction between season x genotype was significant (\( P = 0.0267 \)) for the inverse transformed data, whereas the P values on the original, log, and square root transformation indicated that the interaction was not significant (\( P > 0.39 \)). Because of the heterogeneity of variance, the differential responses of genotypes across the two seasons were not detected in the original, log, or the square-root-transformed data. This example clearly indicates that if the assumption of homogeneity of variance is not met in ANOVA, both the significance levels and the sensitivity of the F tests are biased.

This example clearly shows that significance levels in ANOVA may be biased if the data violate the assumption of the ANOVA. Furthermore, this may lead one to draw incorrect conclusions from the analysis. Therefore, checking data for the violation of the ANOVA assumptions before discussing the results is very critical. Residual analysis is a powerful tool to detect the problems associated with the violation of the ANOVA assumptions. The commonly used transformations such as the log and the square-root conversions may not be appropriate for every data set. The method of power transformation according to Box et al. (1978) is recommended to estimate the appropriate transformation. Although the residual analysis and the estimation of the power transformation needs additional calculations, with the use of the PC-SAS program presented here, the proper NOVA can be performed without tedious calculations.
Fig. 3. Residual plots of the original, square-root and log, and inverse-transformed data for mungbean plant heights.

Literature Cited


Learn the five most important data analysis methods you need in order to interpret your data correctly (and what pitfalls to avoid in the process). But to sort through all this information, you need the right statistical data analysis tools. With the current obsession over “big data,” analysts have produced a lot of fancy tools and techniques available to large organizations. However, there are a handful of basic data analysis tools that most organizations aren’t using…to their detriment. We suggest starting your data analysis efforts with the following five fundamentals and learn to avoid their pitfalls before advancing to more sophisticated techniques.

1. Mean. Provides a comprehensive knowledge of the analytical tools for problem solving related to laboratory data analysis and quality control. The content of the book is motivated by the topics that a laboratory statistics course audience and others have requested over the years since 2003. This book presents a detailed discussion of important basic statistical concepts and methods of data presentation and analysis in aspects of biological experimentation requiring a fundamental knowledge of probability and the foundations of statistical inference, including basic statistical terminology such as simple statistics (e.g., means, standard deviations, medians) and transformations needed to effectively communicate and understand one’s data results.

Important Tools in Statistical Analysis. George C.J. Fernandez

1 Department of Agricultural Economics, University of Nevada-Reno, Reno, NV 89557-0107.

Analysis of variance (ANOVA) is a commonly used statistical analysis in agricultural experiments. Additivity, variance homogeneity, and normality are often considered prerequisites for ANOVA (Cochran, 1943; Eisenhart, 1947). Although the residual analysis and the estimation of the power transformation needs additional calculations, with the use of the PC-SAS program presented here, the proper NOVA can be performed without tedious calculations.

Fig. 3. Residual plots of the original, square-root and log, and inverse-transformed data for mungbean plant heights.