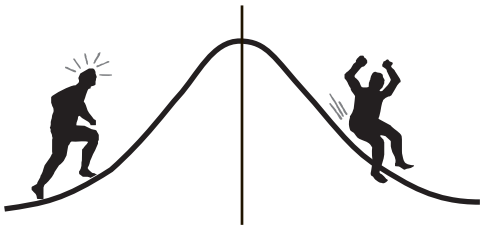

BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. 62 **DATE:** June 15, 2005

SUBJECT: The Box-Cox Transformation

Classical parametric statistical procedures such as ANOVA and regression assume that the data are independent and normally distributed with constant variance. These assumptions can be checked by examining the residuals[1] from an analysis − often called diagnostic checking. When the data, or more specifically, the residuals seriously violate these assumptions, the researcher has five options:

1. Use a non-parametric (rank based) analog to the test, if available. For example, see Hollander and Wolfe (1973) or Conover (1980).
2. Use a randomization procedure based on the observed data, to estimate the distribution of the unknown test statistic when the null hypothesis is true. For example, see Manly (1997).
3. Use a resampling technique such as the bootstrap or jackknife to assess the variability of the estimate(s). For example, see Efron (1982) or Potvin (1993).
4. Fit a different model, one that requires different (and probably more elaborate) distributional assumptions.
5. Apply a transformation to the response variable that will make it conform more closely to the assumed distribution.

This pamphlet will discuss the last option, which is a common approach[2]. There are many traditional transformations (e.g. log, reciprocal, square root, etc.) that researchers are familiar with – see for example, Chapter 13 of Krebs (1989) for an excellent discussion. But if no particular transformation seems obvious, then a solution is to use the Box-Cox transformation (Box and Cox, 1964) – a general procedure for finding a suitable transformation to make the data achieve a normal distribution.

The technique requires that the raw data be continuous and strictly positive. This means that none of the data can be negative or include zeros. If a few observations are less than or equal to zero, simply add a positive number to all observations and proceed. Note that adding a constant to data only shifts the distribution, without affecting its shape. Similarly, multiplying by a constant only affects the scale of the data, without affecting the shape.

[1] A residual is the observed response variable minus the value predicted from the fitted model.
[2] The first four options will not be discussed further in this pamphlet, because each would require substantial exposition to cover adequately.

BRITISH COLUMBIA
Ministry of Forests
Research Program

The Box-Cox transformation is actually a family of power transformations of the form:

$$y' = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{when } \lambda \neq 0 \\ \log(y) & \text{when } \lambda = 0 \end{cases}$$

where λ is an unknown to be estimated from the data.

This family includes several familiar transformations. For example, when λ = 1, there is essentially no transformation since $y' = y - 1$ (i.e. a simple shift to the left by one unit). A square root transformation is produced when λ = 0.5, and λ = -1 is equivalent to a reciprocal transformation[3].

An estimate of λ is obtained by finding the value of λ that maximizes the log-likelihood function (below), which is proportional to the probability of observing the raw data, when a normal independent model properly describes the transformed observations[4]:

$$\log(L(\lambda \mid y_1, y_2, \ldots, y_n)) = -\frac{n}{2}\log(s^2) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i)$$

where

- $\log(L(\lambda \mid y_1, y_2, \ldots, y_n))$ is the log-likelihood function
- n is the number of observations
- s^2 is the estimated variance (using n as the divisor) of the transformed observations y_i'
- y_i denotes the original observations
- λ is the interim estimate of the unknown transformation parameter

The estimated variance of the transformed observations will depend on what model is being fit to the data. For example, in regression and ANOVA, the variance of the residuals is estimated by the residual sums of squared error (SSE) divided by n. If no model is being fit then the variance can be estimated by the squared standard deviation with n as the divisor:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i' - \overline{y}')^2$$

In SAS, Proc Transreg[5] can find the estimate of λ that 'best' transforms the data. This is the maximum likelihood estimate (mle) of λ, or the value of λ that maximizes the above log-likelihood.

---

[3] The main reason for subtracting y^{λ} by 1 and dividing by λ (when $\lambda \neq 0$) is to make the full transformation a continuous function, which assists in the determination of the likelihood function. In application, a simple power transformation (i.e. y^{λ}) is usually employed when $\lambda \neq 0$.

[4] This log-likelihood function is derived using the 'change-of-variable' technique. The probability model for the transformed observations is assumed to be a product of normal distributions with equal variance and known mean (possibly specified by a model). However, to obtain the probability of seeing the untransformed observations (i.e. the raw data), the change-of-variable technique is necessary. See § 3 of Box and Cox (1964) for further details.

[5] Proc Transreg is a multipurpose procedure that fits linear models using a variety of nonlinear transformations. For further reference see the SAS online help (SAS Institute Inc., 2004).

Example

The following example is taken from Box and Cox (1964) where a 3×4 factorial experiment was conducted to test the effect of three poisons and four levels of a treatment on the survival times of animals (measured in units of 10 hours, coded here as "s_time"). The two treatments are allocated to the animals in a completely randomized manner and each treatment combination is replicated four times.

First, a standard ANOVA table on non-transformed survival is constructed using proc GLM. The edited SAS output is:

Dependent Variable: s_time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2.20435625	0.20039602	9.01	<.0001
Error	36	0.80072500	0.02224236		
Corrected Total	47	3.00508125			

R-Square	Coeff Var	Root MSE	Mean
0.733543	31.11108	0.149139	0.479375

Source	DF	Type III SS	Mean Square	F Value	Pr > F
pois	2	1.03301250	0.51650625	23.22	<.0001
treat	3	0.92120625	0.30706875	13.81	<.0001
pois*treat	6	0.25013750	0.04168958	1.87	0.1123

A diagnostic assessment using Proc Univariate showed that the skewness (i.e. asymmetry) of the residuals was 0.592 and the kurtosis (i.e. thinness of the tails) was 2.968. Note that both of these values should be close to zero to be considered normal[6]. Futhermore, the Anderson-Darling statistic of 1.592 ($p<0.005$) indicates that the residuals are not normally distributed[7]. Figure 1 provides graphical evidence to support these statistics. For instance, there are more points around zero than expected for a normal distribution.

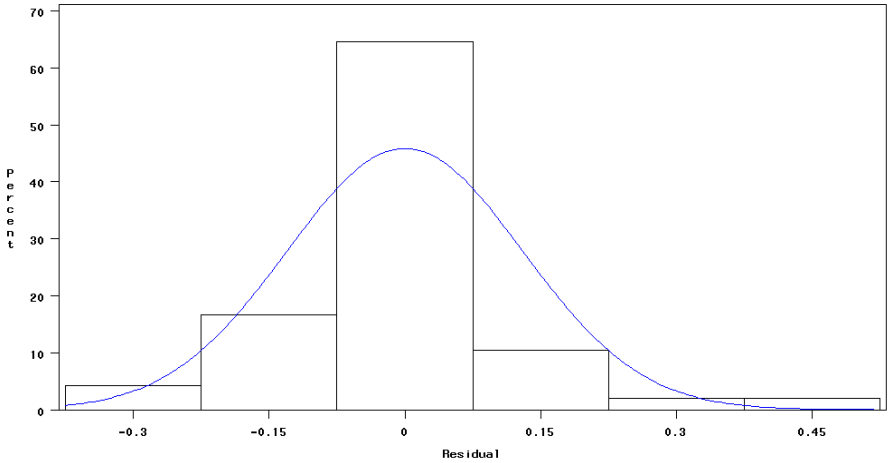

FIGURE 1. Histogram of model residuals based on raw data, with best-fit normal curve superimposed.

[6] To my knowledge, there is no hard-and-fast rule on an acceptable range for skewness and kurtosis in normal populations, although anything beyond -2 and 2 would seem suspicious.

[7] A proper and complete examination of the residuals entails a suite of graphical and statistical diagnostics. For the sake of comparison (and brevity), I have chosen to concentrate on the shape of the distribution and the Anderson-Darling test of normality.

Next, we will use Proc Transreg (see Appendix for complete SAS code) to determine the 'best' transformation:

```
proc transreg data=box;
 model boxcox(s_time) = class(treat pois treat*pois);
run;
```

Within Proc Transreg, the 'boxcox' function (on the left hand side of the equal sign) tells SAS to apply the Box-Cox transformation to the dependent variable s_time. Also, the 'class' function (on the right hand side of the equal sign) tells SAS to use dummy variables to handle the two factors poison and treatment, since they are categorical. Nested factors that involve parentheses are not acceptable arguments, so it is important to specify the full combination of factors that make-up your model. This way, the residual error can be the only possible factor leftover.

The mle of λ is estimated to be -0.82, with estimated 95% confidence interval (-1.29, -0.34); the SAS output is not shown to save space. Rather than applying the full Box-Cox transformation, a simple power version will be used as the new transformed variable. A power transformation is performed since the resulting variable has a more natural interpretation[8]. An exponent of -0.82 is close to -1, which corresponds to the reciprocal of survival time, and this can be interpreted as the rate of dying. The ANOVA table using the transformed variable then becomes:

Dependent Variable: d_rate=(1/s_time)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	56.86218143	5.16928922	21.53	<.0001
Error	36	8.64308307	0.24008564		
Corrected Total	47	65.50526450			

R-Square	Coeff Var	Root MSE	Mean
0.868055	18.68478	0.489985	2.622376

Source	DF	Type III SS	Mean Square	F Value	Pr > F
pois	2	34.87711982	17.43855991	72.63	<.0001
treat	3	20.41428935	6.80476312	28.34	<.0001
pois*treat	6	1.57077226	0.26179538	1.09	0.3867

The skewness and kurtosis of the residuals for this model are 0.544 and -0.169 respectively, and the Anderson-Darling statistic is 0.553 (p=0.1490) so the null hypothesis of normal residuals is not rejected. The residuals in Figure 2 also appear much closer to normal than Figure 1.

From the ANOVA table, notice that although the overall conclusions have not changed, the new model fits better (the R-Square has improved from 0.73 to 0.87) and the tests of the main effects on the transformed variables are more sensitive, as evidenced by observed F-values that have more than doubled.

[8] This seems justified since -1 falls within the estimated 95% confidence interval for λ. Also, Proc Transreg has identified -1 as a "convenient" transformation (output not shown).

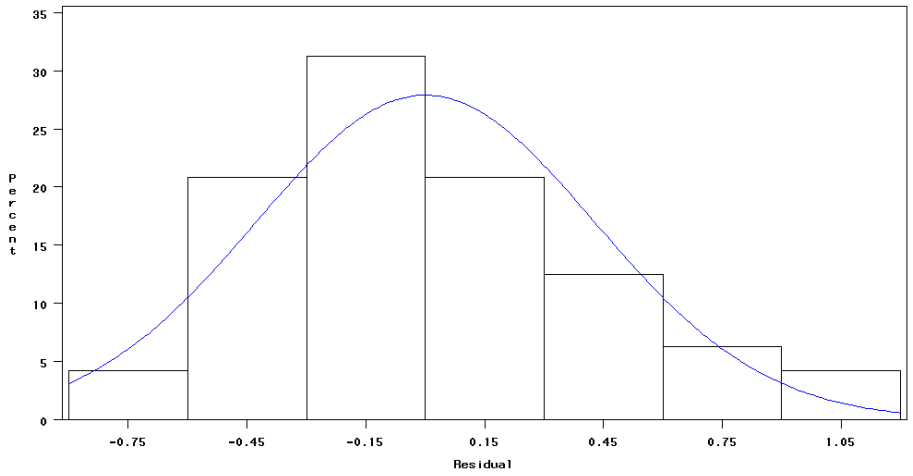

FIGURE 2. Histogram of model residuals based on transformed data, with best-fit normal curve superimposed.

We have seen that the Box-Cox transformation is an objective way to find the optimal (power) transformation to make your data normally distributed. Also, it's advantageous if you can find a transformation that has a meaningful interpretation. But when the transformation leads to little improvement in making your data normally distributed, a different approach (i.e. one of the first four options from the first page) is probably[9] needed.

Prepared by:

Peter Ott. Phone: 250-387-7982; e-mail: peter.ott@gov.bc.ca
B.C. Ministry of Forests, Research Branch.

References:

Box, G. E. P. and Cox, D. R. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B. 26:211-252.

Conover, W. J. 1980. Practical Nonparametric Statistics, 2nd Edition. John Wiley & Sons. NY.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics 38. SIAM. Philadelphia.

Hollander, M. and Wolfe, D. A. 1973. Nonparametric Statistical Methods. John Wiley & Sons. NY.

Krebs, C. J. 1989. Ecological Methodology. Harper Collins. NY.

Manly, B. F. J. 1997. Randomization and Monte Carlo Methods in Biology, Second edition. Chapman & Hall. NY.

Potvin, C. and Roff, D. A. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology 74(6):1617-1628.

SAS Institute Inc. 2004. SAS OnlineDoc® 9.1.3**.** Cary, NC. SAS Institute Inc.

[9] It is possible that a transformation outside of the Box-Cox family (e.g. arcsine square root, logit, etc) would work.

Appendix: SAS Program

```
data box;
 do pois = 1 to 3;
    do treat = 'a','b','c','d';
       do rep = 1 to 4;
          input s_time @;
          d_rate=1/s_time;
          output;
       end;
    end;
 end;
 cards;
 0.31  0.45  0.46  0.43
 0.82  1.10  0.88  0.72
 0.43  0.45  0.63  0.76
 0.45  0.71  0.66  0.62
 0.36  0.29  0.40  0.23
 0.92  0.61  0.49  1.24
 0.44  0.35  0.31  0.40
 0.56  1.02  0.71  0.38
 0.22  0.21  0.18  0.23
 0.30  0.37  0.38  0.29
 0.23  0.25  0.24  0.22
 0.30  0.36  0.31  0.33
 ;
run;

proc glm data=box;
 class pois treat;
 model s_time=pois|treat / ss3;
 *restrict output to only display type 3 (“last in”) sums of squares;
 output out=predicted r=resid;
run;

proc univariate data=predicted;
 var resid;
 histogram / normal;
run;

proc transreg data=box;
 model boxcox(s_time / lambda=-2 to 0 by 0.01) = class(treat|pois);
 *use the 'lambda' option to fine-tune estimation. The default is -3 to
  3 in increments of 0.25;
 *the pipe operator works like Proc Glm and Proc Mixed, where x1|x2
  implies x1 x2 and x1*x2.
run;

proc glm data=box;
 class pois treat;
 model d_rate=pois|treat / ss3;
 output out=predicted r=resid;
run;

proc univariate data=predicted;
 var resid;
 histogram / normal;
run;

quit;
```