A Note on the Selection of Data Transformations

D. F. Andrews

Biometrika, Volume 58, Issue 2 (Aug., 1971), 249-254.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28197108%2958%3A2%3C249%3AANOTSO%3E2.0.CO%3B2-Q

Biometrika (1971), 58, 2, *p.* 249
Printed in Great Britain

A note on the selection of data transformations

By D. F. ANDREWS
Bell Telephone Laboratories, Murray Hill and Princeton University

Summary

Recently Box & Cox (1964) and Fraser (1967) have proposed likelihood functions as a basis for choosing a transformation of data to yield a simple linear model. Here a simple, exact, test of significance is proposed for the consistency of the data with a postulated transformation within a given family. Confidence sets can be derived from this test. The power of the test may be estimated and used to predict the sharpness of the inferences to be derived from such an analysis. The methods are illustrated with examples from the paper by Box & Cox (1964).

1. Introduction

Box & Cox (1964) considered the choice of a transformation among a parametric family of data transformations to yield a simple, normal, linear model. They investigated two approaches to this problem and derived a likelihood function and a posterior distribution for the parameters of the transformation. Draper & Cox (1969) have found approximations for the precision of the maximum likelihood estimate.

Fraser (1967) derived a different likelihood function which yields quite different inferences from those of Box & Cox (1964) in extreme cases where the number of parameters is close to the number of observations.

Likelihood methods require repeated computations using a number of transforms of the original data. This can be troublesome if there is a multiparameter family of transformations. A further defect of the likelihood methods is that confidence limits and tests based on them have only asymptotic validity; the number of parameters must be small compared with the number of observations. This will not be the case for small data sets, paired comparison experiments and extreme cases.

In the present paper, a method is proposed which has three possible advantages over direct calculation of likelihoods. Its main disadvantage is that it does not lead to such a clear graphical summary of conclusions as is given by a plot of a likelihood. The advantages are: (i) an 'exact' test of significance is obtained from which 'exact' confidence limits can be calculated; (ii) the amount of calculation is reduced if only one or a few transforms are to be tested; (iii) the precision with which the transformation can be estimated is capable of theoretical calculation.

2. Test of significance

Consider $\Lambda = \{\lambda\}$, a parametric class of transformations

$$\lambda: \mathbf{y} \rightarrow \mathbf{y}^{(\lambda)} \quad (\lambda \in \Lambda),$$

and suppose that for some λ the transformed response $\mathbf{y}^{(\lambda)}$ may be described by a linear model

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \sigma \mathbf{e}, \qquad (2\cdot1)$$

where $\mathbf{X}$ is a $n \times p$ matrix of independent variables with rows $\mathbf{x}_i'$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, σ is an unknown scale parameter and $\mathbf{e}$ is a vector whose elements are independent standard normal deviates.

The model for $\mathbf{y}$ is nonlinear; nevertheless a simple F test may be derived to test the hypothesis $\lambda = \lambda_0$, a given value of λ. The derivation of this test is analogous to those of Williams (1962) which are based on locally linear expansions of the nonlinear terms in the model. The present paper extends this method to cases where terms in the expansion depend on the response $\mathbf{y}$.

Assume that the transformation $\mathbf{y}$ involves q functionally independent parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$ and is sufficiently regular to be approximated by a linear expansion about the true value $\boldsymbol{\lambda}$:

$$y_i^{(\lambda_0)} = y_i^{(\lambda)} + \mathbf{v}_i(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda})$$

and so

$$\mathbf{y}^{(\lambda_0)} = \mathbf{X}\boldsymbol{\beta} + \mathbf{V}(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}) + \sigma\mathbf{e},$$

where

$$v_{ij} = \left[\frac{\partial\{y_i^{(\lambda)}\}}{\partial y_j}\right]_{\lambda=\lambda_0}.$$

The matrix $\mathbf{V}$ depends on $\mathbf{y}$ and must be modified to yield a simple test. The model (2·1) fitted to $\mathbf{y}^{(\lambda_0)}$ yields fitted values for $\mathbf{y}$, given by $\hat{y}_i^{(\lambda_0)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. Thus the matrix $\mathbf{V}$ may be approximated by $\hat{\mathbf{V}}$ calculated using these fitted values:

$$\hat{v}_{ij} = \left[\frac{\partial\{y_i^{(\lambda)}\}}{\partial \lambda_j}\right]_{\lambda=\lambda_0,\, y=\hat{y}}.$$

A test of the hypothesis $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ may now be constructed from the modified model

$$\mathbf{y}^{(\lambda_0)} = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{V}}(\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}) + \sigma\mathbf{e}.$$

The F statistic to test the hypothesis $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ is based on the regression of the least squares residuals $\mathbf{r} = \mathbf{y}^{(\lambda_0)} - \mathbf{X}\hat{\boldsymbol{\beta}}$ on $\hat{\mathbf{U}} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\hat{\mathbf{V}}$. The independence of $\mathbf{r}$ and $\hat{\boldsymbol{\beta}}$ implies the independence of $\mathbf{r}$ and $\hat{\mathbf{U}}$ since $\hat{\mathbf{U}}$ depends on $\mathbf{y}$ only through $\hat{\boldsymbol{\beta}}$. Thus the F statistic calculated in this way has a standard F distribution. It is important to note that the precision of the above approximations may affect the efficiency of the above test but it will not affect the exactness of this distribution; for a detailed development of this result see Milliken & Graybill (1970).

3. TWO EXAMPLES

In this section the foregoing theory is applied to two examples of Box & Cox (1964). They considered a family of transformations depending on a real parameter λ, equivalent to

$$y^{(\lambda)} = \begin{cases} y^{\lambda} & (\lambda \neq 0), \\ \log y & (\lambda = 0). \end{cases}$$

For this class of transformations the matrix of derivatives is a vector $\mathbf{v}$ which may be estimated by

$$\hat{v}_i = \lambda_0^{-1}\mathbf{x}_i'\hat{\boldsymbol{\beta}} \ln(\mathbf{x}_i'\hat{\boldsymbol{\beta}}).$$

Box & Cox (1964) consider in some detail a biological and a textile example.

The biological experiment was a 3 × 4 factorial experiment in which the factors were poisons and treatments and the data were 48 survival times of test animals.

There were some general reasons for analysing rates, which correspond to $\lambda_0 = -1$. The significance level associated with this hypothesis is 0·15. Had these authors chosen, say,

$\lambda_0 = 1$, the significance level would have been 0·01. Note that the calculations required for this test involve only one regression operation for each transformation tested.

The textile example was a single replicate of a 3^3 design in which the data consisted of 27 observations on the number of cycles to failure of yarn under repeated loadings. It was natural to try analyzing $\log y$, $\lambda_0 = 0$. The significance level associated with the hypothesis $\lambda_0 = 0$ is 0·6. The significance level associated with $\lambda_0 = -0{\cdot}25$, say, is 0·003.

4. Confidence sets

The significance test of § 2 may be used to generate a confidence set for λ defined by

$$C(\mathbf{y}) = \{\lambda : \alpha(\lambda) \geqslant \alpha_0\}.$$

Since the test has exact size α_0, the confidence intervals generated in this way are exact with confidence coefficient $1-\alpha_0$. The amount of calculation required to determine these confidence sets is much greater, although no maximization is required.

5. Examples

The confidence intervals for the two examples discussed by Box & Cox (1964) may be readily obtained graphically from Table 1. In Table 2 these are compared with the approximate intervals derived by Box and Cox based on the asymptotic properties of the likelihood function.

Table 1. *Significance levels for various values of* λ

A. Biological example

λ	−1·25	−1·05	−0·55	−0·05	0·45	0·95
α	0·03	0·11	0·89	0·25	0·04	0·01

B. Textile example

λ	−0·25	−0·15	−0·05	0·05	0·15	0·25
α	0·003	0·09	0·74	0·30	0·03	0·002

Table 2. *Ninety-five per cent confidence sets for the two examples: the significance and the likelihood methods compared*

	Confidence interval	
Method	Biological	Textile
Significance	$-1{\cdot}18 < \lambda < 0{\cdot}40$	$-0{\cdot}2 < \lambda < 0{\cdot}12$
Likelihood	$-1{\cdot}13 < \lambda < -0{\cdot}37$	$-0{\cdot}18 < \lambda < 0{\cdot}06$

6. General remarks

Tukey (1949) proposed a test for non-additivity with one degree of freedom. This one degree of freedom is chosen to be sensitive to second-order terms in $E(y)$, the expected response. The test is based on the F statistic associated with fitting one additional variable, a vector whose elements are $(\mathbf{x}'\hat{\boldsymbol{\beta}})^2 = \hat{y}_i^2$. This criterion could also be used to test a hypothesized transformation and to generate exact confidence intervals.

The 'exact' test is of a similar form and corresponds in general to a test of non-additivity with p degrees of freedom. These degrees of freedom are selected to be sensitive to small changes in the transformation, one degree of freedom for each independent transformation parameter. For the one-parameter family of power transformations this test is based on the F statistic associated with fitting one additional variable, a vector whose elements are $v_i = (\mathbf{x}_i'\hat{\boldsymbol{\beta}})\ln(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ or $\hat{y}_i \log \hat{y}_i$. By writing $\hat{y}_i = \bar{y}(1+d_i)$ and noting that the F statistic will be unchanged under scalar multiplication of $\mathbf{v}$ or addition of any linear combination of the columns of $\mathbf{X}$ to $\mathbf{v}$, we see that $\mathbf{V}$ is equivalent, for test purposes, to a vector with components $v_i = (1+d_i)\log(1+d_i)$ or

$$v_i = (1+d_i)(d_i - \tfrac{1}{2}d_i^2 + \tfrac{1}{3}d_i^3 - \ldots),$$

or

$$v_i = \tfrac{1}{2}d_i^2 - \tfrac{1}{6}d_i^3 + \ldots,$$

since $\mathbf{d}$ lies in the linear space generated by $\mathbf{X}$. If $d_i \ll 1$, then v_i is approximately equivalent to $v_i = d_i^2$ or $v_i = \hat{y}_i^2$, the components of the vector used in the test with one degree of freedom for non-additivity. It is not surprising then that the exact and the non-additivity tests yield very similar results in these examples. The two tests may lead to quite different results in some cases, particularly where some $d_i > 1$ or when the family of transformations involves $p > 1$ independent parameters.

This discussion indicates that much of the information in the exact tests derives from the assumption of additivity. Indeed, in the biological example with the completely non-additive model specifying 12 independent cell expectations, $\hat{\mathbf{U}} \equiv \mathbf{O}$ and the test does not exist. Often this should be the most important basis for the selection of a transformation.

However, an examination of Fig. 6 of Box & Cox (1964) reveals that, for the biological example, the likelihood inferences are almost the same, assuming an additive or a non-additive model. In this case the likelihood is determined by other features of the data, primarily homogeneity of variance. This may explain the narrower confidence interval given by the likelihood method in this example (Table 2).

To investigate how sensitive the likelihood and exact methods are to outliers, one observation in the biological example was changed. The response 0·23 for poison II, treatment A, corresponding to the largest residual when $\lambda = -1$, was changed to 0·13. The maximum likelihood estimate and the estimate obtained by minimizing the exact test criterion are given in Table 3 together with 75 % confidence limits calculated by both methods for the original and the perturbed data.

Table 3. *The effect of one outlier on estimates of* λ

	Original data	With one outlier modified
Likelihood method		
75 % confidence interval	$-0{\cdot}95 < \lambda < -0{\cdot}05$	$-0{\cdot}3 < \lambda < 0{\cdot}05$
Maximum likelihood estimate	$-0{\cdot}75$	$-0{\cdot}15$
Significance method		
75 % confidence interval	$-0{\cdot}9 < \lambda < 0{\cdot}05$	$-1{\cdot}2 < \lambda < 0{\cdot}0$
Minimum F estimate	$-0{\cdot}5$	$-0{\cdot}5$

The maximum likelihood estimate was affected much more than the minimum F estimate. The confidence limits of both methods were affected. The perturbed observation so inflated the residual sum of squares for small λ that the lower exact confidence limit for confidence coefficients $> 0{\cdot}9$ was very small.

This points to perhaps the most important difference between the two methods. The F statistic and hence the 'exact' method is relatively insensitive to some departures from normality which will remain in the transformed data where they can, perhaps, be detected and the model or data modified accordingly. The likelihood method is sensitive to such departures from normality. In data transformed by this method these departures will be harder to identify but their effect will be smaller.

The choice of the method used will depend in part on (i) the relative weights to be given normality and the simplicity of model; (ii) the interpretability of a particular transformation; (iii) the sample size, and through it the relevance of the asymptotic properties of the likelihood function; and (iv) the possible treatment of outliers and other departures from normality. There will be instances where one, the other, and both methods should be investigated.

7. Power

In this section the power of the test is approximated and used to predict the size of the confidence intervals.

The test described in § 2 is an F test. If the linear expansions of § 2 were not approximations, the nonnull distribution of the statistic would be noncentral F. In the following, the approximation of the nonnull distribution by noncentral F will be made. The power of the test is then a function of the noncentrality parameter of the χ^2 distribution of the numerator of the F statistic. This noncentrality parameter may be obtained by replacing $\mathbf{y}$ with its expectation in the calculation of this term. When this is done the sum of squares entry associated with fitting $\hat{\mathbf{U}}$ after $\mathbf{X}$ is just $(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)'\,\hat{\mathbf{U}}'\hat{\mathbf{U}}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)$. The noncentrality parameter is approximated by

$$\hat{\gamma}^2 = \hat{\sigma}^{-2}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)'\,\hat{\mathbf{U}}'\hat{\mathbf{U}}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0), \qquad (7\cdot1)$$

where $\hat{\sigma}^2$ is an unbiased estimate of σ^2 obtained from the residual sum of squares. The size of the confidence interval may be predicted by

(i) finding the noncentrality parameter, γ^2, associated with power, say $\frac{1}{2}$;

(ii) replacing $\hat{\gamma}^2$ with γ^2 in (7·1) and solving for $(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0)$ to obtain an ellipsoid.

In the absence of any approximation, points lying on this surface would lie outside the confidence set with probability equal to the power used, here $\frac{1}{2}$.

The exact test is based on the regression of the least squares residuals on the matrix $\hat{\mathbf{U}} = \hat{\mathbf{U}}(\hat{\boldsymbol{\beta}})$. The power of this test is conditional, given an independent variable $\hat{\boldsymbol{\beta}}$ whose distribution is known. In principle the unconditional power may be calculated although this seems a very difficult task.

8. Examples

The noncentrality parameter associated with power $\frac{1}{2}$ is $\gamma^2 = 4$ for both examples. Thus $|\lambda-\lambda_0|$ may be found from

$$\gamma^2 \doteq \hat{\gamma}^2 = (\lambda-\lambda_0)^2\,\hat{\mathbf{U}}'\hat{\mathbf{U}}/\hat{\sigma}^2$$

using the original data ($\lambda = 1$), in the calculation of $\hat{\mathbf{U}}$. The corresponding values of $|\lambda-\lambda_0|$ are given in Table 4 for $\lambda_0 = 0$ and $\lambda_0 = 1$. These are compared with the half width of the observed 95 % confidence intervals given in Table 2.

Table 4. *The critical difference $|\lambda-\lambda_0|$ estimated with using two values of λ*

	Difference $\lvert\lambda-\lambda_0\rvert$	
Estimated using:	Biological	Textile
$\lambda = 1$	0·8	0·13
$\lambda = 0$	0·8	0·14
Actual from Table 2	0·7	0·12

The close agreement between the predicted critical difference for power $\frac{1}{2}$ and the observed width of the related confidence interval obtained by detailed calculation suggests that the above procedure may well be used to identify those instances for which a detailed calculation will yield only broad and hence not vital inferences about λ.

The estimated size is almost invariant under changes in λ (see Table 4). Thus its use does not require *a priori* knowledge of the true value.

The author gratefully acknowledges the comments of Professor D. R. Cox, Colin Mallows and the referees, which were most helpful in revising this note.

References

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *J.R. Statist. Soc.* B **26**, 211–52.

Draper, N. R. & Cox, D. R. (1969). On distributions and their transformation to normality. *J. R. Statist. Soc.* B **31**, 472–6.

Fraser, D. A. S. (1967). Data transformations and the linear model. *Ann. Math. Statist.* **38**, 1456–65.

Milliken, G. A. & Graybill, F. A. (1970). Extensions of the general linear hypothesis. *J. Am. Statist. Ass.* **65**, 797–807.

Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics* **5**, 232–42.

Williams, E. J. (1962). Exact fiducial limits in non-linear estimation. *J. R. Statist. Soc.* B **24**, 125–39.

[*Received June* 1970. *Revised March* 1971]

Some key words: Data transformations; *F* tests; Non-additivity; Normal distribution; Power transformations.