Fan Plot

Introduction

The fan plot is a **graphical** procedure for determining the effect of one or more observations on the **transformation** parameter λ in the Box and Cox family of power transformations of the **response** in **regression**. Such transformations, for example, from $\boldsymbol{y}$ to $\log \boldsymbol{y}$, are often important for ensuring that the assumptions behind **least squares** are satisfied and that therefore, efficient use is made of data (*see* **Power Transformations**). The fan plot is based on a **forward search** through the data to fit subsets of increasing numbers of observations, with any **outliers** being included toward the end of the search. The plot monitors the behavior of the approximate score test for five different transformations and reveals whether the evidence for a transformation depends on a few observations or is, preferably, spread throughout the data.

Interest is in transformation of the response $\boldsymbol{y}$ in the **multiple regression** model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

$\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters and it is assumed that the additive errors of observation ε are independently distributed with constant variance σ^2. Also in (1) $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of **explanatory variables** and perhaps functions of them, such as quadratics and interaction terms. To obtain the approximate score test we add a "constructed variable" (*see* **Residuals**) to the regression model and obtain the augmented model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{w}$ is $n \times 1$ and γ is a scalar parameter. The approximate score test is the **Student *t*-test** t_γ for testing that γ in (2) equals zero. The constructed variable for the transformation is derived in the next section. Testing that $\gamma = 0$ is testing that there is no evidence for any transformation of the response.

A Score Test for Transformations

The analysis of the data on mandible length in the article on **residuals** shows appreciable evidence not only of the normality of the residuals (*see* **Normality, Tests of**, Figure 3) but also of increasing variance with fitted value, Figure 1. Often, normality and constant variance can be achieved by fitting the regression model not to $\boldsymbol{y}$ but to a function of $\boldsymbol{y}$; Figure 1 of the article on **diagnostics** shows the beneficial effect of the transformation to $\log(\boldsymbol{y})$ combined with quadratic regression (*see* **Polynomial Regression**) on the residuals from the mandible length data. The appropriate transformation frequently, but, as will be seen later, not always, also leads to a simple linear model, without quadratic or **interaction** terms.

The logarithmic transformation is one special case of the normalized power transformation [4]

$$z(\lambda) = \begin{cases} \dfrac{\boldsymbol{y}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y} \log \boldsymbol{y} & \lambda = 0, \end{cases} \tag{3}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i / n)$. For inference about the transformation parameter λ, Box and Cox suggest **likelihood ratio tests**. A computationally simpler alternative test is the approximate score statistic (*see* **Likelihood**) derived by Taylor series expansion of (3) as

$$\begin{aligned} z(\lambda) &\doteq z(\lambda_0) + (\lambda - \lambda_0) \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0} \\ &= z(\lambda_0) + (\lambda - \lambda_0)\boldsymbol{w}(\lambda_0). \end{aligned} \tag{4}$$

In (4), $\boldsymbol{w}(\lambda_0)$ is the "constructed variable" for the transformation and can be treated as is the extra-explanatory variable in (2). To test the transformation $\lambda = \lambda_0$ the response $\boldsymbol{y}$ is transformed to $z(\lambda_0)$ in (3). The approximate score statistic, $T_p(\lambda_0)$, is then the t statistic t_γ for regression of the transformed response on $\boldsymbol{w}(\lambda_0)$ in (2). Details of the constructed variables are in the article on **residuals**.

The Fan Plot

In the **forward search**, the p parameters of the regression model (1) are estimated by least squares applied to a carefully chosen subset of m observations. We start the search with m small, usually p or perhaps $p + 1$, and randomly select 1000 subsamples. The initial subset provides the least median of squares estimator, that is it minimizes the median squared

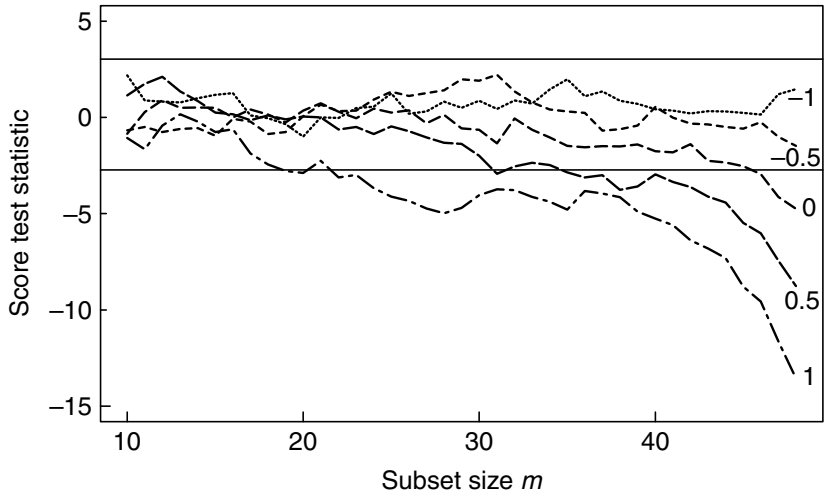

Figure 1 Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: both $\lambda = -1$ and $\lambda = -0.5$ are acceptable. There is no evidence of any outliers or influential observations

residual [5]. We then order the residuals and augment the subset.

When m observations are used in fitting, the optimum subset yields n residuals $\boldsymbol{e}(m^*)$. We order the squared residuals $\boldsymbol{e}^2(m^*)$ and take the observations corresponding to the $m + 1$ smallest as the new subset. Usually, this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. Owing to the form of the search, outliers, if any, tend to enter as m approaches n.

We combine calculation of the test statistic $T_p(\lambda_0)$ with the forward search. Since observations that are outlying on one scale may not be outlying for a different transformation, we conduct several searches for different values of λ_0. In most applications, including the examples here, we use five searches for the values $\lambda = -1, -0.5, 0, 0.5$, and 1. If there are outliers for a particular λ, they will enter the search last and influence the value of the test statistic.

As a first example, we use the Poisson Data from Box and Cox [4], partly analyzed in the article on **residuals**. These data are well behaved: there are no outliers or influential observations that cannot be reconciled with the greater part of the data by a suitable transformation. Our fan plot clearly indicates the reciprocal transformation. We then consider a series of modifications of the data in which an increasing number of outliers is introduced. The fan plot reveals the structure in all instances.

The data are the times to death of animals in a 3×4 **factorial experiment** with four observations at each factor combination. All our analyses use an additive model, that is, without interactions, so that $p = 6$, the model used by Box and Cox when finding the reciprocal transformation. The implication is that the model should be additive in death rate, not in time to death.

The fan plot of the values of the approximate score statistic $T_p(\lambda)$ for the five searches as the subset size m increases is given in Figure 1 and shows that the reciprocal transformation is acceptable as is the inverse square root transformation ($\lambda = -0.5$). The horizontal lines are at ± 2.58, corresponding to 1% significance, assuming the statistics have a **standard normal** distribution. The results of Atkinson and Riani [3] show that this is a good working approximation.

Initially, for small subset sizes, there is no evidence against any transformation. During the whole forward search, there is never any evidence against either $\lambda = -1$ or $\lambda = -0.5$ (for all the data $\hat{\lambda} = -0.75$). The log transformation is also acceptable until the last four observations are included by the forward search. These are some of the largest observations, which will be informative about the need to transform. Evidence that some transformation is needed is spread throughout the data, less than half of the observations being sufficient to reject the hypothesis that $\lambda = 1$. There are no jumps in this curve, just an increase in evidence against $\lambda = 1$ as each observation is introduced into the subset. The relative smoothness of the curves reflects the lack of outliers and exceptionally influential cases and

the general shape of the plot justifies the name of "fan plot".

For the introduction of a single outlier into the Poisson data, we follow Andrews [1] and change observation 8, one of the readings for Poison II, group A, from 0.23 to 0.13. This is not one of the larger observations, so the change does not create an outlier in the scale of the original data. The effect on the estimated transformation of all the data is, however, to replace the reciprocal with the logarithmic transformation: $\hat{\lambda} = -0.15$. And, indeed, the fan plot of the score statistics from the forward searches in Figure 2 shows that, at the end of the forward search, the final acceptable value of λ is 0, with -0.5 on the boundary of the acceptance region.

Figure 2 clearly reveals the altered observation and the differing effect it has on the five searches. Initially, the curves are the same as those of Figure 1. But for $\lambda = 1$, there is a jump due to the introduction of the outlier when $m = 41$, which provides evidence for higher values of λ. For other values of λ, the outlier is included further on in the search. When $\lambda = 0.5$, the outlier comes in at $m = 46$, giving a jump to the score statistic in favor of this value of λ. For the other values of λ, the outlier is the last value to be included. Inclusion of the outlier has the largest effect on the inverse transformation. It is clear from the figure how this one observation is causing an appreciable change in the evidence for a transformation.

We now further modify the Poisson data; in addition to the previous modification, we also change observation 38 (Poison I, group D) from 0.71 to 0.14. This creates an example of masking, in which one outlier hides the effect of another, so that neither is evident when using the methods for the deletion of single observations described in the article on **diagnostics**.

The effect of the two outliers is clearly seen in the fan plot, Figure 3. Here, only $\lambda = 0$ is acceptable at the end of the search. The plot also reveals the differing effect the two altered observations have on the five searches. Initially, the curves are again similar to those of the original data shown in Figure 1. The difference is greatest for $\lambda = -1$ where addition of the two outliers at the end of the search causes the statistic to jump from an acceptable 1.08 to 10.11. The effect is similar, although smaller, for $\lambda = -0.5$. It is most interesting, however, for the log transformation. Toward the end of the search this statistic is trending downwards, below the acceptable region. But addition of the last two observations causes a jump in the value of the statistic to a nonsignificant value. The incorrect log transformation is now acceptable.

For these three values of λ, the outliers are the last two observations to be included in the search. They were created by introducing values that are too near zero when compared with the model fitted to the rest of the data. For the log transformation, and more so for the reciprocal, such values become extreme and so have an appreciable effect on the fitted model. For the other two values of λ, the outliers are included earlier in the search. The effect is most clearly seen when $\lambda = 1$; the outliers come in at $m = 40$ and 46, giving upward jumps to the score statistic in favor of

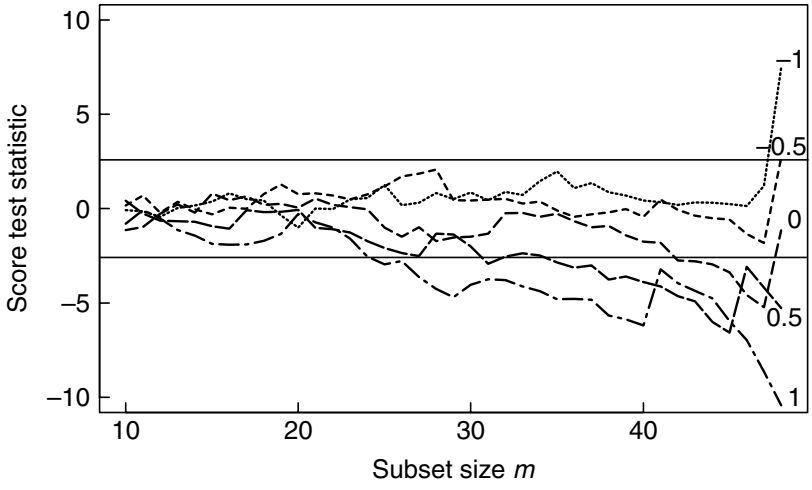

Figure 2 Modified Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: the effect of the outlier is evident in making $\lambda = 0$ appear acceptable at the end of the search

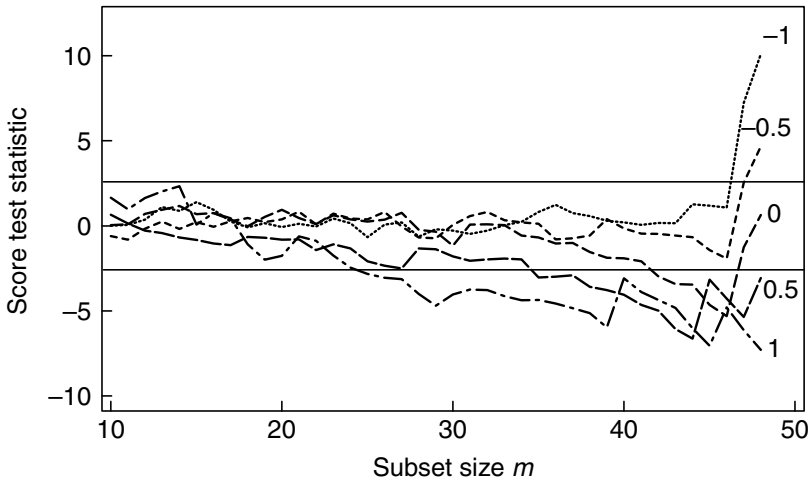

Figure 3 Doubly modified Poisson data: fan plot–forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost; the effect of the two outliers is clear

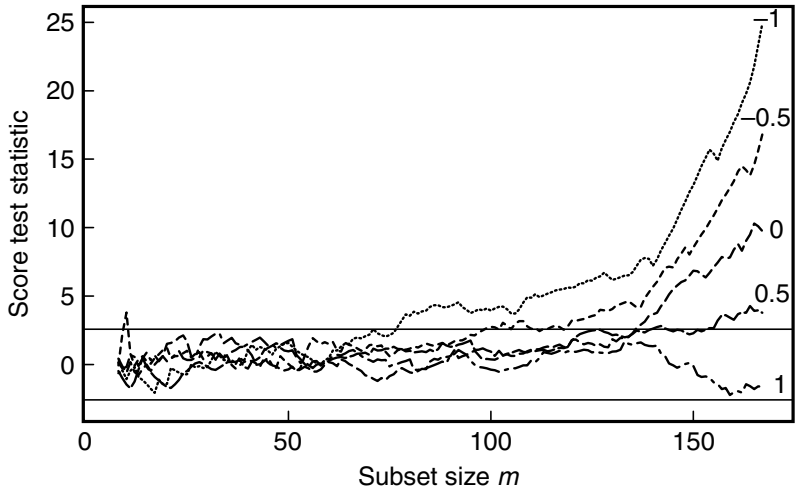

Figure 4 Mandible length data: fan plot–forward plot of $T_p(\lambda)$ for the five transformations of the data when the regression is on age

this value of λ. For the remaining value of 0.5, one of the outliers is the last value to be included.

These three plots exhibit the main features of the fan plot. Further analyses of the examples and comparison with other procedures are in Atkinson and Riani [2, Sections 4.4 and 4.7]. One conclusion is that alternative diagnostic procedures, such as the constructed variable plot in Figure 5 of **residuals**, can fail in the presence of masking and multiple outliers.

Mandible Length Data

The preceding examples calibrate the properties of the fan plot. We now use it to analyze transformations of the mandible length data.

The plot of the residuals of the untransformed data after regression on **gestational age**, for example, Figure 3 of **residuals**, showed three negative outliers as well as many smaller residuals lying outside the simulation envelope. In contrast, the residuals after regression of $\log \boldsymbol{y}$ on a quadratic in age, Figure 1 of **diagnostics**, are much more nearly normal. Is the evidence for this transformation largely dependent on the outlying observations? How is it affected by the linear model?

We start with just simple regression. Figure 4 is a fan plot for the five transformations of the data when the regression is on age. There is no evidence for a transformation – all values except $\lambda = 1$ are rejected by the end of the search. The

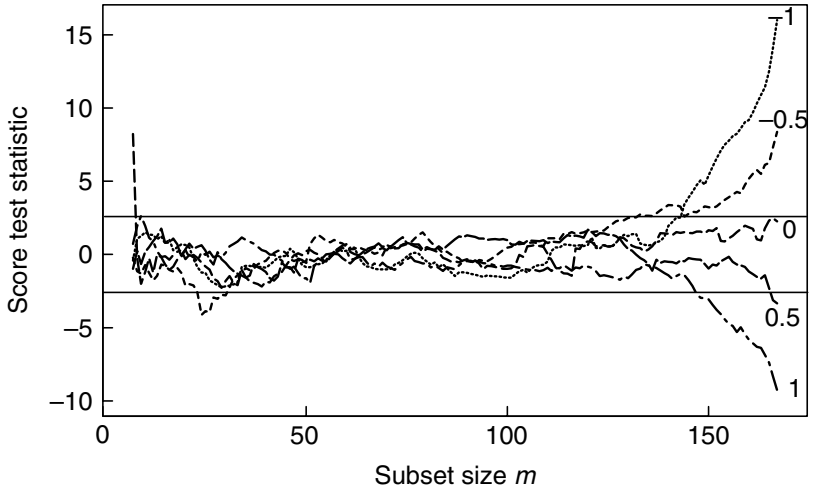

Figure 5 Mandible length data: fan plot–forward plot of $T_p(\lambda)$ for the regression of log $\boldsymbol{y}$ on a quadratic in age

statistic for this value remains within the bounds of ± 2.58 throughout the search. Although the values are toward the lower boundary at the end of the search, there is no obvious evidence of the effect of the three outlying observations, of the kind seen in Figure 3. Such jumps in the curve of the statistic are most in evidence for the reciprocal transformation $\lambda = -1$, where the observations giving negative residuals on the untransformed scale are even more extreme after transformation.

Although there is no evidence for transformation when regression is on age, we know from Table 1 of **diagnostics** that the quadratic term in this regression is significant. The final plot, Figure 5, is therefore the fan plot for the regression of log $\boldsymbol{y}$ on a quadratic in age. It shows that, for this more complicated model with an extra term, the log transformation is the only one that is acceptable. Although the last three observations to enter the search increase the value of the statistic, it does not change dramatically. There are no jumps in the other curves of the kind visible for $\lambda = -1$ in Figure 4.

The general conclusion is that the logarithmic transformation with a quadratic model is to be preferred to simple regression and no transformation. As the forward plots of t statistics for regression coefficients in Figure 4 of the article on the **forward search** show, this conclusion is supported by all the data and is in agreement with the $Q-Q$ plots of residuals mentioned above. An interesting feature of the analysis is that transformation has strengthened the evidence for a more complicated regression model. Often transformations result in a simpler model, but here there is a conflict between the linearity of the plot of $\boldsymbol{y}$ against $\boldsymbol{x}$ and the increasing variance with $\boldsymbol{y}$ evident in Figure 2 of the article on **Goodness of Fit**. This conflict was a reason for the fractional polynomial models used by Royston and Altman [6]. An alternative analysis is to keep the simple linear model, but to transform both sides of the model to obtain errors with constant variance (*see* **Power Transformations**). The forward search for this transformation is illustrated in [2, Section 4.12].

References

[1] Andrews, D.F. (1971). A note on the selection of data transformations, *Biometrika* **58**, 249–254.

[2] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[3] Atkinson, A.C. & Riani, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression, *Chemometrics and Intelligent Laboratory Systems* **60**, 87–100.

[4] Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B* **26**, 211–246.

[5] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[6] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON & MARCO RIANI