Review of An Analysis of Transformations

Tuhin Sarkar 09007030, Sagar Chordia 09005013, Amal Agarwal 09D11001

March 11, 2013

Abstract

In this report we attempt to summarize the paper - "An Analysis of Transformations, G.E.P Box and D.R. Cox". Additionally, to test the transformations discussed in this paper, we conduct simulations of hypothesis testing and linear regressions as part of this report.

1 Introduction

The usual techniques for linear regression are justified by the following underlying assumptions:

- simplicity of structure of E(y)
- constancy of error variance
- normality of distributions
- independence of observations

However, sometimes these assumptions may not hold. The motivation behind this paper is to derive transformations such that these not only justify these assumptions but rather find, wherever possible, a metric in terms of which the finding may be succinctly expressed.

Tukey (1949, 1950) used transformations to achieve additive in the analysis of variance. To stabilize the variance, the usual method of transformation by Bartlett (1947) is used where the relation between mean and variance is determined. Anscombe (1961) and Tukey and Anscombe (1983) have employed analysis of residuals to detect departures from standard assumptions and have indicated how transformations might be devised from functions of the residuals. There are many problems where both dependent and independent variables are transformed. Box and Tidwell (1962) transformation can be employed without affecting the constancy of variance and normality of error distribution. It is useless to try to linearize a relation which is not monotonic, but a transformation is sometimes useful in such cases.

2 General Idea

The general idea is to study transformed variables y^{λ} indexed by an unknown parameter λ, estimate λ and other parameter of our model. There are two

primary lines of analyses. First is where the particular λ is of interest. Second is where we study the factor effects of a choice of λ.
The important family of transformations discussed here are:

$$y^{\lambda} = \frac{y^{\lambda} - 1}{\lambda}$$

when $\lambda \neq 0$

$$y^{\lambda} = \log(y)$$

when $\lambda = 0$
Now this transformation is valid only when $y \geq 0$. There is also a second set of transformations which hold for $y \geq \lambda_2$

$$y^{\lambda} = \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1}$$

when $\lambda_1 \neq 0$

$$y^{\lambda} = \log(y + \lambda_2)$$

when $\lambda_1 = 0$
But the second transformation is nothing but a linear transformation of the first one. Therefore we can analyze the first one itself.
The basic idea is that once we have the transformation we can use our general linear regression model:

$$E[y^{\lambda}] = a\theta$$

where a is the independent variable matrix and θ are our parameters.
There are two approaches to find λ:

- Use the maximum likelihood method
- Use a Bayesian approach

In the first method we use the likelihood function:

$$\frac{1}{\sqrt{(2\pi)}^{n/2}} \exp\left\{-\frac{(y^{\lambda} - a\theta)'(y^{\lambda} - a\theta)}{2\sigma^2}\right\} J(\lambda; y)$$

$$J(\lambda; y) = \Pi_{i=1}^{n} \left(\frac{dy_i^{\lambda}}{dy_i}\right)$$

Note that $J(\lambda; y)$ is the Jacobian of transformation from y to y^{λ}. In this first, we apply "orthodox" large-sample maximum likelihood theory. This approach leads to point estimates of the parameters and to approximate tests and confidence intervals based on chi-squared distribution.
Now, we can use the normal linear regression techniques on the transformed dependent variables, where we get:

$$\hat{\sigma^2} = y^{\lambda'} a_r y^{\lambda}/n$$

$$a_r = I - a(a'a)^{-1}a'$$

The log-likelihood function looks like:

$$L_{max}(\lambda) = -\frac{1}{2} n \log \hat{\sigma^2}(\lambda) + log J(\lambda; y)$$

To get λ from this, plot the maximized log-likelihood function for a series of trial λ values. From the plot the maximizing value of λ can be taken as:

$$L_{max}(\hat{\lambda}) - L_{max}(\lambda) < \frac{1}{2}\chi^2_{\nu}(\alpha)$$

Here ν is the number of degrees of freedom of λ
The above results can be simplified by using the further transformation given by:

$$z^{\lambda} = y^{\lambda}/J^{1/n}$$

We now discuss how to approach the problem of finding λ using the Bayesian analysis. Consider

$$p(y|\theta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\{-\frac{\nu s^2(\lambda) + (\theta - \hat{\theta_\lambda})' a' a (\theta - \hat{\theta_\lambda})}{2\sigma^2}\} J(\lambda; y)$$

Where $\hat{\theta_\lambda}$ is the least square estimate of θ for given λ.
In this method we use the Bayes theorem to get the posterior distribution of λ by using the following:

$$I(\lambda|y) = \int_{-\infty}^{\infty} d(\log\sigma) \int_{-\infty}^{\infty} d\theta p(y|\theta, \sigma^2)$$

$$p(\lambda|y) = K'_y \frac{I(\lambda|y) p_0(\lambda)}{\{J(\lambda; y)\}^{(n-\nu)/n}}$$

Note K'_y is a normalizing constant. Here is also assumed that θ and $\log\sigma$ are effectively uniform over the range where likelihood is appreciable. By taking $p_0(\lambda)$ as uniform (prior distribution), we maximize the posterior distribution. That is required λ.

3 Further Analysis of Transformations

This analysis indicates the following things:

- how simple a model we are justified in using
- what weight is given to considerations (simple expectation structure, constant variance and normal distributions) in choosing λ
- whether different transformations are really needed

The basic ideas behind this analysis is to use the two approaches as before. We first discuss the log-likelihood method. Consider the constraint C. Now:

$$L_{max}(\lambda|C) = L_{max}(\lambda) + \{L_{max}(\lambda|C) - L_{max}(\lambda)\}$$

The two terms are examined separately. Similarly when more number of constraints are imposed.
Now in the Bayesian case:

$$p(\lambda|C) = p(\lambda)\frac{p(C|\lambda)}{p(C)}$$

4 Experiments

4.1 Aim

For non-trivial application of Box-cox transformation, data should not satisfy all properties of Linear normal regression model.

4.2 Data Generation

X is generated using uniform random distribution. We generate $X = U(1, 21)$. Let $Y = X^2 + \varepsilon$. Squaring X ensures that model is not linear. In normal model $\varepsilon = N(0, \sigma^2)$. Hence we use t-distribution to generate ε guaranteeing non-normality.

Code to generate data

```
n=100;
r = 10*ones(n,1);
X_1=ones(n,1);
X_2=1+20*rand(n,1);
X=[X_1,X_2];
m=0.1*trnd(r);
Y =(X_2.^2)+m;
```

4.3 Box-Cox transformation

We apply box-cox transformation on given dataset. To estimate λ we use maximum likelihood method. Note X is generated from uniform distribution and hence rank of X is most of times equal to N number of data points. Our choice of dataset ensures that we can use maximum likelihood method.

On evaluation of concepts in section 2 it can be shown that $\widehat{\lambda}$ can be found by minimizing following function

$$RSS(V(\lambda)) = [V(\lambda) - \widehat{V}(\lambda)]^t[V(\lambda) - \widehat{V}(\lambda)]$$

where

$$\widehat{V}(\lambda) = X(X^tX)^{-1}X^tV(\lambda)$$

$$V(\lambda) = \begin{bmatrix} V_1(\lambda) \\ V_2(\lambda) \\ \vdots \\ V_n(\lambda) \end{bmatrix}$$

$$V_i(\lambda) = (Y_i^{\lambda} - 1)/\lambda \overbrace{Y}^{\lambda-1}, \lambda \neq 0$$

$$V_i(\lambda) = \overbrace{Y} log(Y_i), \lambda = 0$$

$$\overbrace{Y} = (Y_1 Y_2 \cdots Y_n)^{1/N}$$

Matlab code for λ estimation

```
GM=1;
for i=1:n
    GM=GM*Y(i);
end
GM=GM^(1/n);
V = @(lambda) (((Y.^lambda)-1)./(lambda*(GM^(lambda-1))));
V_cap = @(lambda) (X*(inv(X'*X))*(X')*V(lambda));
RSS=@(lambda) ((V(lambda)-V_cap(lambda))'*(V(lambda)-V_cap(lambda)));
lambda_max = fminbnd(RSS, -100, 100);
```

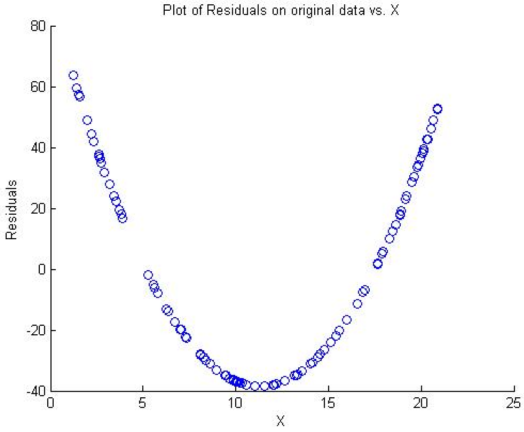
The estimated value of λ is **0.5002**

4.4 Analysis

We plot residual graph for original data and transformed data. Following is matlab code to plot graphs

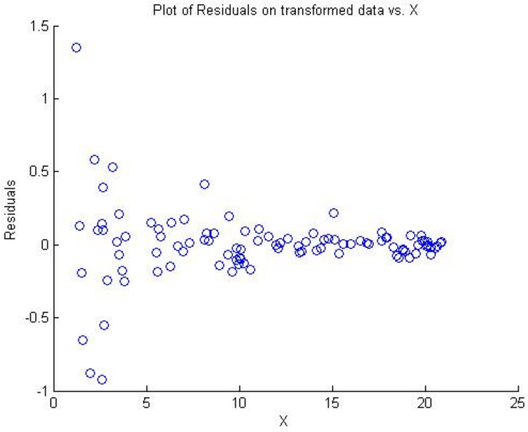
```
%residual plots on the original data
k = (X_2-mean(X_2))./(sum((X_2-mean(X_2)).^2));
b1 = sum(k.*Y);
b0=mean(Y)-b1*mean(X_2);
Y_cap = b0+b1*X_2;
Res_org = (Y-Y_cap);
figure, scatter(X_2,Res_org);
% axis([0 1800 0 2.5]);
title ('Plot of Residuals on original data vs. X');
xlabel('X');
ylabel('Residuals');

%residual plots on the transformed data
Res_new = V(lambda_max)-V_cap(lambda_max);
```

```
figure, scatter(X_2,Res_new);
title ('Plot of Residuals on transformed data vs. X');
xlabel('X');
ylabel('Residuals');
Err = RSS(lambda_max);
```



1.jpg

Figure 1: Residual plot for original data



2.jpg

Figure 2: Residual plot for transformed data

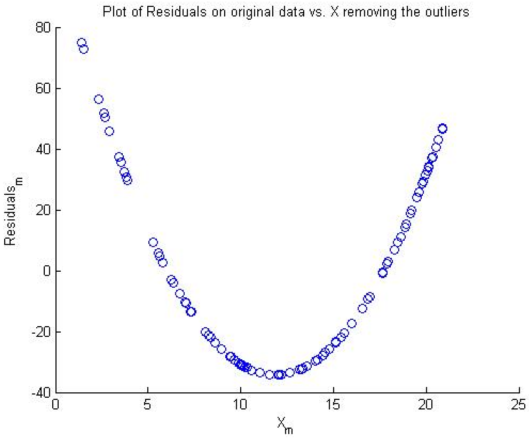
In residual plot for original data we see parabolic function of residual with respect to X. Hence it violates the linear regression model assumption. In residual plot for transformed data residuals is almost uniformly spread across 0 over range of X. But the residuals for data points close to zero is large compared to other datapoints. Therefore there is possibility of outliers being present in data.

4.5 Outlier detection and removal

We repeated the entire procedure for multiple dataset. In all experiments residual near zero were high indicating that data-generation method produced outliers near zero. Thus we remove all points from dataset whose residual is greater than some threshold in our case we took threshold as 0.3. Following is code for outlier removal.

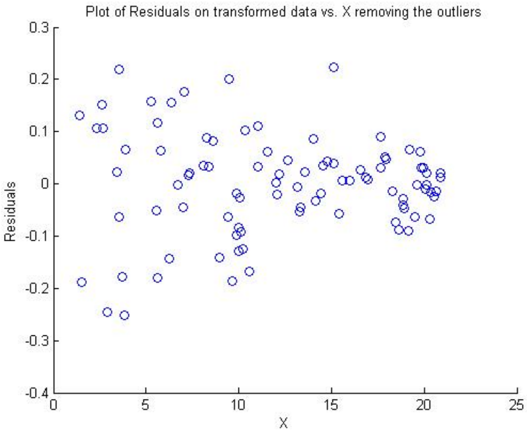
```
% Outlier Analysis
k=0;
for i=1:n
    if (abs(Res_new(i))>0.3)
        k=[i,k];
    end
end

s=length(k);
for i=1:s-1
    km(i) = k(i);
end

X_2_temp1=X_2;
tr_temp1=tr;
for i=1:n
    for j=1:s-1
        if (km(j)==i)
            X_2_temp1(i) = 0;
            tr_temp1(i) = 0;
        end
    end
end

X_2_temp2=[0];
tr_temp2=[0];
for p=1:n
    if (X_2_temp1(p)~=0)
        X_2_temp2 = [X_2_temp1(p);X_2_temp2];
    end
    if (tr_temp1(p)~=0)
        tr_temp2 = [tr_temp1(p);tr_temp2];
    end
end
```

```
y=length(X_2_temp2);
X_2_m=zeros((n-s+1),1);
tr_m=zeros((n-s+1),1);
for i=1:n-s+1
    X_2_m(i) = X_2_temp2(i);
    tr_m(i) = tr_temp2(i);
end
```



3.jpg

Figure 3: Residual plot for original data after removing outliers



4.jpg

Figure 4: Residual plot for transformed data after removing outliers

We repeated the entire analysis for this modified dataset. Residual plots were regenerated. As we can see the residuals are now evenly distributed with respect to range of X. And hence transformation was useful to remove non-linearity and heteroscedacity in the generated data.

4.6 Hypothesis testing

We did hypothesis test for $\beta_0 = 0$ and $\beta_1 = 0$ vs. corresponding alternatives for 95 percent confidence coefficient that is $\alpha = 0.05$ Calculated t-statistic in each case was mush greater leading to rejection of null hypothesis. Thus we conclude $\beta_0 \neq 0$ and $\beta_1 \neq 0$ after transformation.

5 Conclusions

This paper gives an idea of a transformation that can be employed to modify dependent variables so that they are workable (standard linear regression techniques can be applied). However, at the end of the paper there is also some discussion as to how we can also use a transformation where even the independent variables are modified.