Chapter 5

Transformation and Weighting to Correct Model Inadequacies

The graphical methods help in detecting the violation of basic assumptions in regression analysis. Now we consider the methods and procedures for building the models through data transformation when some of the assumptions are violated.

Variance stabilizing transformations

In regression analysis, it is assumed that the variance of disturbances is constant, i.e., $Var(\varepsilon_i) = \sigma^2,\ i = 1, 2, ..., n.$ Suppose this assumption is violated. A common reason for such isolation is that the study variable follows a probability distribution in which the variance is functionally related to mean.

For example, if the study variable (y) in the model is Poisson random variable in a simple linear regression model, then its variance is the same as the mean. Since mean of y is related to the explanatory variable x, so the variance of y will be proportional to x. In such cases, variance stabilizing transformations are useful.

In another example, if y is proportion, i.e., $0 \le y_i \le 1$ then in such cases the variance of y is proportional to $E(y)[1 - E(y)]$. In such case, the variance – stabilizing transformation is useful.

Some commonly used variance-stabilizing transformations in the order of their strength are as follows:

Relation of σ^2 to $E(y)$	**Transformation**
$\sigma^2 \propto$ constant	$y^* = y$ **(**no transformation**)**
$\sigma^2 \propto E(y)$	$y^* = \sqrt{y}$ (Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y^* = \sin^{-1}(\sqrt{y})$ **(**Binomial proportion $0 \le y_i \le 1$**)**
$\sigma^2 \propto [E(y)]^2$	$y^* = \ln(y)$
$\sigma^2 \propto [E(y)]^3$	$y^* = 1/\sqrt{y}$
$\sigma^2 \propto [E(y)]^4$	$y^* = \frac{1}{y}$

After making a suitable transformation, use y^* as a study variable in the respective case.

The strength of a transformation depends on the amount of curvature present in the curve between the study and the explanatory variable. The transformation mentioned here ranges from relatively mild to relatively strong. The square root transformation is relatively mild and reciprocal transformation is relatively strong. The square root transformation is relatively mild and reciprocal transformation is relatively strong.

In general, a mild transformation applied when the minimum and maximum values do not range much (e.g. $y_{\max} / y_{\min} < 2,3)$ and such transformation has little effect on the curvature. On the other hand, when the minimum and maximum vary much, then a strong transformation is needed that will have a substantial impact on the analysis.

In the presence of non-constant variance, the OLSE will remain unbiased but will looses the minimum variance property.

When the study variable has been transformed as y^*, then the predicted values are in the transformed scale. It is often necessary to convert the predicted values back to the original units (y).

When the inverse transformation is applied directly to the original values, then it gives an estimate of the median of the distribution of the study variable instead of the mean. So one needs to be careful while doing so.

Confidence interval and prediction interval may be directly converted from one metric to another. The reason being that the interval estimates are percentile of distribution and percentiles are unaffected by the transformation. One may note that the resulting intervals may or may not be or remain the shortest possible intervals.

Transformations to linearize the model

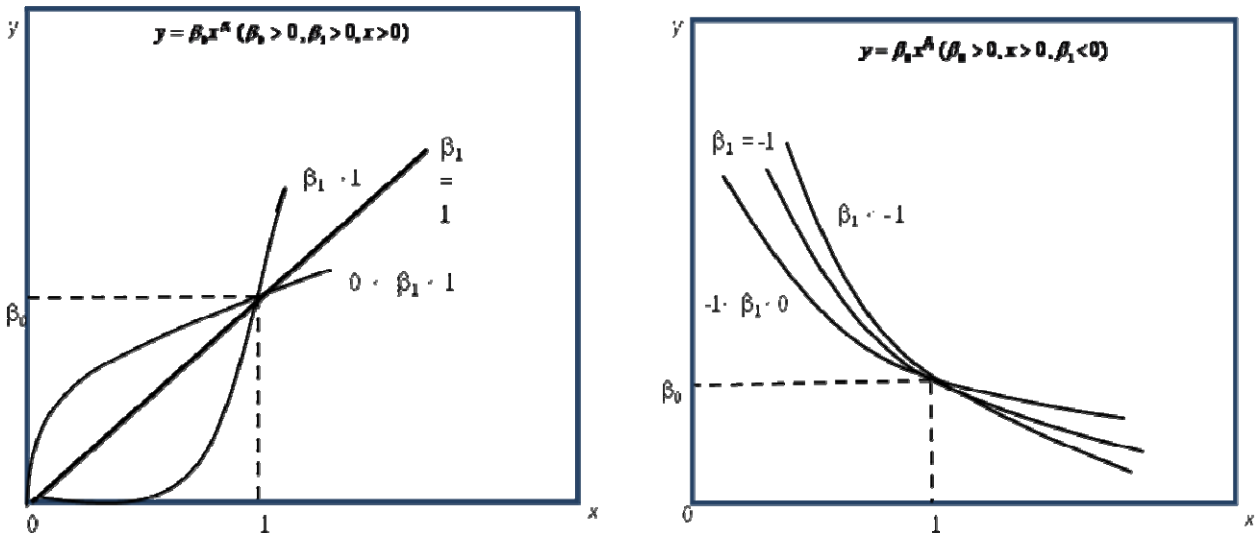
The basic assumption in linear regression analysis is that the relationship between the study variable and explanatory variables is linear. Suppose this assumption is violated. Such violation can be checked by scatter plot matrix, scatter diagrams, partial regression plots, lack of fit test etc.

In some cases, a nonlinear model can be linearized by using a suitable transformation. Such nonlinear models are called **intrinsically or transformable linear.** The advantage of transforming the nonlinear

function into the linear function is that the statistical tools are developed for the case of a linear regression model. For example, exact tests for the test of hypothesis, confidence interval estimation etc. are developed for the case of a linear regression model. Once the nonlinear function is transformed to a linear function, all such tools can be readily applied, and there is no need to develop them separately.

Some linearizable functions are as follows:

1. If the curve between y and x is like as follows:



then the possible linearizable function is of the form

$$y = \beta_0 x^{\beta_1}.$$

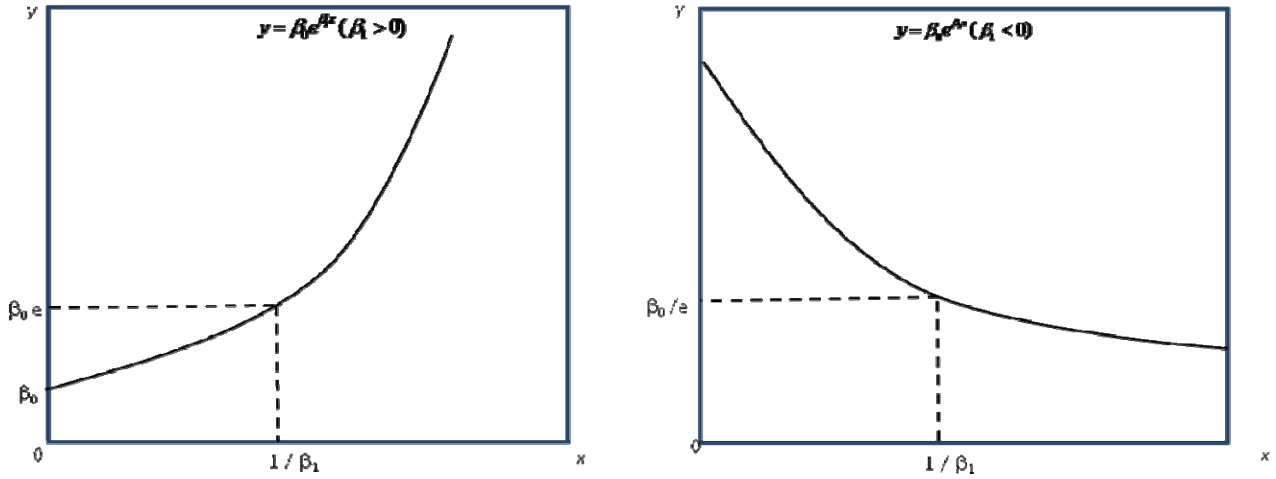
Using the transformation $y^* = \ln y,\ x^* = \ln x,$ i.e., by taking log on both sides, the model becomes

$$\log y = \log \beta_0 + \beta_1 \log x$$

or $y^* = \beta_0^* + \beta_1 x^*$

where $\beta_0^* = \log \beta_0$ and the model becomes a linear model. Note that the parameter β_0 changes to log β_0 in the transformed model.

2. If the curve between y and x is like as follows



then the possible linearizable function is of the form

$$y = \beta_0 \exp(\beta_1 x)$$

Taking $\log_e(\ln)$ on both sides,

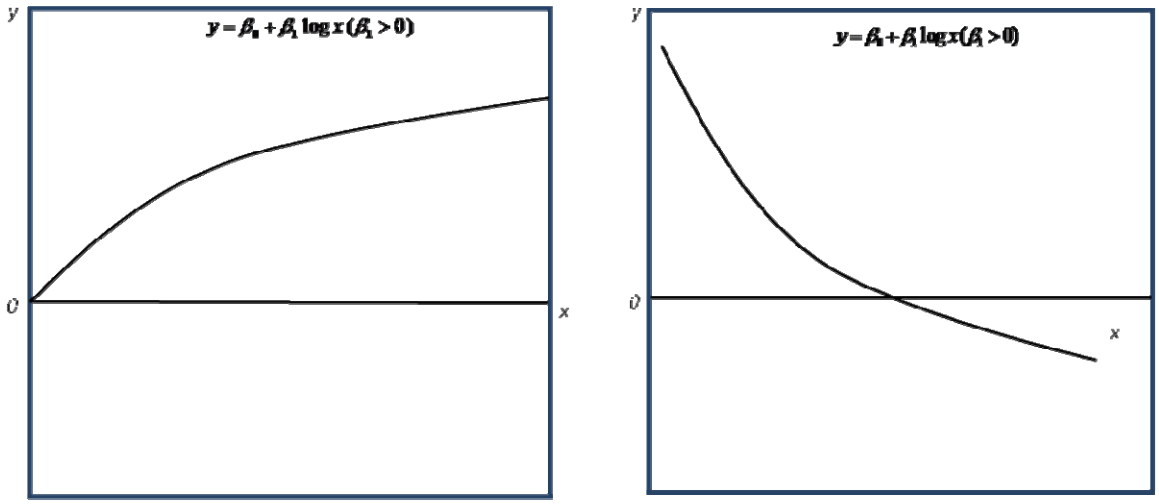
$$\ln y = \ln \beta_0 + \beta_1 x$$

or $y^* = \beta_0^* + \beta_1 x$

where $y^* = \ln y$ and $\beta_0^* = \ln \beta_0$.

So $y^* = \ln y$ is the transformation needed in this case. The intercept term β_0 becomes $\ln \beta_0$ in the transformed model.

3. If the curve between y and x is like as follows



then the possible linearizable function is of the form

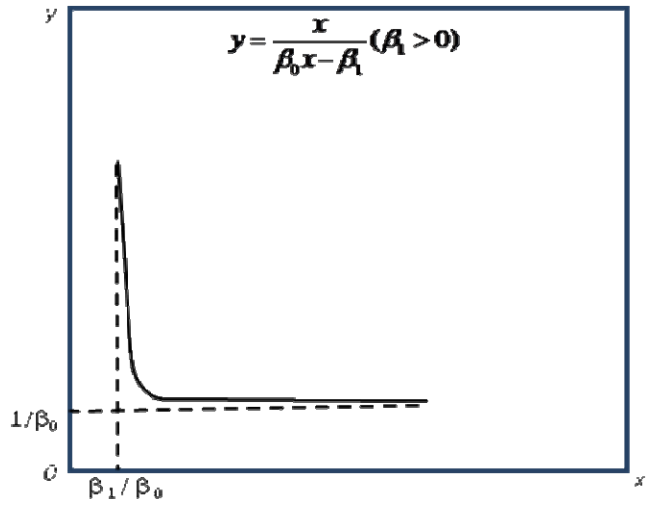
$$y = \beta_0 + \beta_1 \log x$$

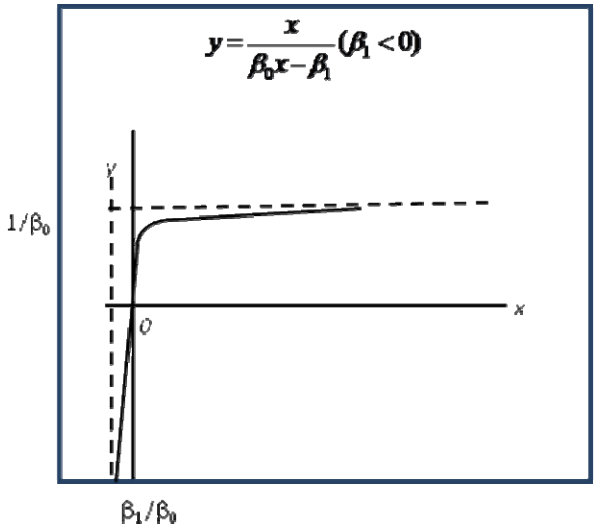
which can be written as

$$y = \beta_0 + \beta_1 x^*$$

using the transformation $x^* = \log x$.

4. If the curve between y and x is like as follows





then the possible linearizable function is of the form

$$y = \frac{x}{\beta_0 x - \beta_1}$$

which can be written as

$$\frac{1}{y} = \beta_0 - \frac{\beta_1}{x}$$

or $y^* = \beta_0 + \beta_1 x^*$.

which becomes a linear model by using the transformation $y^* = \frac{1}{y}$, $x^* = -\frac{1}{x}$.

- With the observed behaviour of the plots, one can choose any such curve and use the linearized form of the function.
- When such transformations are used, many times the form of ε also gets changed. For example, in the case of

$$y = \beta_0 \exp(\beta_1 x)\varepsilon$$

$$\ln y = \ln \beta_0 + \beta_1 x + \ln \varepsilon$$

or $y^* = \beta_0^* + \beta_1 x + \varepsilon^*$.

This implies that the multiplicative error in the original model is log normally distributed in the transformed model. Many times, we ignore this aspect and continue to assume that the random errors are still normally distributed. In such cases, the residuals from the transformed model should be checked for the validity of the assumptions.

- When such transformations are used, the OLSE has the desired properties with respect to the transformed data and not the original data.

Analytical methods for selecting a transformation on study variable

The Box-Cox method

Suppose the normality and/or constant variance of the study variable y can be corrected through a power transformation on y. This means y is to be transformed as y^{λ} where λ is the parameter to be determined. For example, if $\lambda = 0.5,$ then the transformation is the square root and $\sqrt{y}$ is used as a study variable in place of y.

Now the linear regression model has parameters β, σ^2 and λ. Box and Cox method tells how to estimate simultaneously the λ and parameters of the model using the method of maximum likelihood.

Note that as λ approaches zero, y^{λ} approaches to 1. So there is a problem at $\lambda = 0$ because this makes all the observation y to be unity. It is meaningless that all the observation on the study variable are constant. So there is a discontinuity at $\lambda = 0$. One approach to solve this difficulty is to use $\frac{y^{\lambda} - 1}{\lambda}$ as a study variable. Note that as $\lambda \to 0, \frac{y^{\lambda} - 1}{\lambda} \to \ln y$. So a possible solution is to use the transformed study variable as

$$W = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln y & \text{for } \lambda = 0. \end{cases}$$

So the family W is continuous. Still, it has a drawback. As λ changes, the value of W change dramatically. So it is difficult to obtain the best value of λ. If different analysts obtain different values of $\lambda,$ then it will fit different models. It may then not be appropriate to compare the models with different values of λ. So it is preferable to use an alternative form

$$y^{(\lambda)} = V = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda y_*^{\lambda-1}} & \text{for } \lambda \neq 0 \\ y_* \ln y & \text{for } \lambda = 0 \end{cases}$$

where y_* is the geometric mean of y_i's as $y_* = (y_1 y_2 ... y_n)^{1/n}$ which is constant.

For calculation purpose, we can use

$$\ln y_* = \frac{1}{n}\sum_{i=1}^{n} \ln y_i .$$

When V is applied to each y_i, we get $V = (V_1, V_2, ..., V_n)'$ as a vector of observation on transformed study variable, and we use it to fit a linear model

$$V = X\beta + \varepsilon$$

using the least squares or maximum likelihood method.

The quantity $\lambda y_*^{\lambda-1}$ in the denominator is related to the n^{th} power of Jacobian of transformation. See how:

We want to convert y_i into $y_i^{(\lambda)}$ as

$$y_i^{(\lambda)} = W_i = \frac{y_i^{\lambda}-1}{\lambda}; \quad \lambda \neq 0 .$$

Let $y = (y_1, y_2, ..., y_n)'$, $W = (W_1, W_2, ..., W_n)'$.

Note that if $W_1 = \dfrac{y_1^{\lambda}-1}{\lambda}$, then

$$\frac{\partial W_1}{\partial y_1} = \frac{\lambda y_1^{\lambda-1}}{\lambda} = y_1^{\lambda-1}$$

$$\frac{\partial W_1}{\partial y_2} = 0.$$

In general,

$$\frac{\partial W_i}{\partial y_j} = \begin{cases} y_i^{\lambda-1} & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

The Jacobian of transformation is given by

$$J(y_i \to W_i) = \frac{\partial y_i}{\partial W_i} = \frac{1}{\left(\dfrac{\partial W_i}{\partial y_i}\right)} = \frac{1}{y_i^{\lambda-1}}.$$

$$J(W \to y) = \begin{vmatrix} \frac{\partial W_1}{\partial y_1} & \frac{\partial W_1}{\partial y_2} & \cdots & \frac{\partial W_1}{\partial y_n} \\ \frac{\partial W_2}{\partial y_1} & \frac{\partial W_2}{\partial y_2} & \cdots & \frac{\partial W_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial W_n}{\partial y_1} & \frac{\partial W_n}{\partial y_2} & \cdots & \frac{\partial W_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} y_1^{\lambda-1} & 0 & 0 & \cdots & 0 \\ 0 & y_2^{\lambda-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & y_n^{\lambda-1} \end{vmatrix}$$

$$= \prod_{i=1}^{n} y_i^{\lambda-1}$$

$$= \left(\prod_{i=1}^{n} y_i \right)^{\lambda-1}$$

$$J(y \to W) = \frac{1}{J(W \to Y)} = \left(\frac{1}{\prod_{i=1}^{n} y_i} \right)^{\lambda-1} .$$

Since this is a Jacobian when we want to transform the whole vector y to whole vector W. If an individual y_i is to be transform into W_i, then take its geometric mean as

$$J(y_i \to W_i) = \left(\frac{1}{\left(\prod_{i=1}^{n} y_i \right)^{\frac{1}{n}}} \right)^{\lambda-1} .$$

The quantity $J(Y \to W) = \frac{1}{\prod_{i=1}^{n} y_i^{\lambda-1}}$ ensures that unit volume is preserved moving from the set of y_i to the set of V_i. This is a factor which scales and ensures that the residual sum of squares obtained from different values of λ can be compared.

To find the appropriate family, consider

$$y^{(\lambda)} = V = X\beta + \varepsilon$$

where $y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda y_*^{\lambda-1}}$, $\varepsilon \sim N(0, \sigma^2 I)$.

Applying the method of maximum likelihood for likelihood function for $y^{(\lambda)}$,

$$L\left[y^{(\lambda)}\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}\varepsilon_i^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{\varepsilon'\varepsilon}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)}{2\sigma^2}\right]$$

$$\ln L\left[y^{(\lambda)}\right] = -\frac{n}{2}\ln\sigma^2 - \left[\frac{(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)}{2\sigma^2}\right] \qquad \text{(ignoring constant).}$$

Solving

$$\frac{\partial \ln L\left[y^{(\lambda)}\right]}{\partial\beta} = 0$$

$$\frac{\partial \ln L\left[y^{(\lambda)}\right]}{\partial\sigma^2} = 0$$

gives the maximum likelihood estimators

$$\hat{\beta}(\lambda) = (X'X)^{-1}X'y^{(\lambda)}$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}y^{(\lambda)}{}'\left[I - X(X'X)^{-1}X'\right]y^{(\lambda)} = \frac{y^{(\lambda)}{}'\bar{H}y^{(\lambda)}}{n}$$

for a given value of λ.

Substituting these estimates in the log-likelihood function $\ln L\left[y^{(\lambda)}\right]$ gives

$$L(\lambda) = -\frac{n}{2}\ln\hat{\sigma}^2 = -\frac{n}{2}\ln\left[SS_{res}(\lambda)\right]$$

where $SS_{res}(\lambda)$ is the sum of squares due to residuals which is a function of λ. Now maximize $L(\lambda)$ with respect to λ. It is difficult to obtain any closed form of the estimator of λ. So we maximize it numerically.

The function $-\frac{n}{2}\ln\left[SS_{res}(\lambda)\right]$ is called as the **Box-Cox objective function.**

Let $\lambda_{\max}$ be the value of λ which minimizes the Box-Cox objective function. Then under fairly general conditions, for any other λ

$$n\ln\left[SS_{res}(\lambda)\right]-n\ln\left[SS_{res}(\lambda_{\max})\right]$$

has approximately $\chi^2(1)$ distribution. This result is based on the large sample behaviour of the likelihood ratio statistic. This is explained as follows:

The likelihood ratio test statistic in our case is

$$\eta_n \equiv \eta = \frac{\underset{\Omega_o}{Max}\, L}{\underset{\Omega}{Max}\, L}$$

$$= \frac{\underset{\Omega_o}{Max}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}}{\underset{\Omega}{Max}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}}$$

$$= \frac{\left(\frac{1}{\hat{\sigma}^2(\lambda)}\right)^{\frac{n}{2}}}{\left(\frac{1}{\hat{\sigma}^2(\lambda_{\max})}\right)^{\frac{n}{2}}}$$

$$= \left(\frac{1/SS_{res}(\lambda)}{1/SS_{res}(\lambda_{\max})}\right)^{\frac{n}{2}}$$

$$\ln\eta = \frac{n}{2}\ln\left(\frac{SS_{res}(\lambda_{\max})}{SS_{res}(\lambda)}\right)$$

$$-\ln\eta = \frac{n}{2}\ln\left(\frac{SS_{res}(\lambda)}{SS_{res}(\lambda_{\max})}\right)$$

$$= \frac{n}{2}\ln\left[SS_{res}(\lambda)\right]-\frac{n}{2}\ln\left[SS_{res}(\lambda_{\max})\right]$$

$$= -L(\lambda)+L(\lambda_{\max})$$

where

$$L(\lambda) = -\frac{n}{2}\ln\left[SS_{res}(\lambda)\right]$$

$$L(\lambda_{\max}) = -\frac{n}{2}\ln\left[SS_{res}(\lambda_{\max})\right].$$

Since under certain regularity conditions, $-2\ln\eta_n$ converges in distribution to $\chi^2(1)$ when the null hypothesis is true, so

$$-2\ln\eta \sim \chi^2(1)$$

$$\text{or } -\ln\eta \sim \frac{\chi^2(1)}{2}$$

$$\text{or } L(\lambda_{\max}) - L(\lambda) \sim \frac{\chi^2(1)}{2}.$$

Computational procedure

The maximum- likelihood estimate of λ corresponds to the value of λ for which residual sum of squares from the fitted model $SS_{res}(\lambda)$ is a minimum. To determine such λ, we proceed computationally as follows:

- Fit $y^{(\lambda)}$ for various values of λ. For example, start with values in (-1, 1) then take the values in (-2, 2) and so on. Take about 15 to 20 values of λ which are expected to be sufficient for the estimation of optimum value.
- Plot $SS_{res}(\lambda)$ versus λ.
- Find the value of λ which minimizes $SS_{res}(\lambda)$ from the graph.
- A second iteration can be performed using a finer mesh of values of desired.

Note that the value of λ can not be selected by directly comparing the residual sum of squares from the regression of y^{λ} on x because for each λ, the residual sum of squares is measured on a different scale.

It is better to use simple values of λ. For example, the practical difference between $\lambda = 0.5$ and $\lambda = 0.58$ is likely to be small but $\lambda = 0.5$ is much easier to interpret.

Once λ is selected, then use

- y^{λ} as a study variable if $\lambda \neq 0$
- $\ln y$ as a study variable if $\lambda = 0$.

It is entirely acceptable to use $y^{(\lambda)}$ as a response to the final model. This model will have a scale difference and an origin shift in comparison to model using y^{λ} (or $\ln y$) as the response.

An approximate confidence interval for λ

We can find an approximate confidence interval for the transformation parameter λ. This interval helps in selecting the final value of λ. For example, if $\hat{\lambda} = 0.58$ is the value of λ which is minimizing the sum of squares due to residual. But if $\lambda = 0.5$ is in the confidence interval, then one may use the square root transformation because it is easier to explain. Furthermore, if $\lambda = 1$ is in the confidence interval, then it may be concluded that no transformation is necessary.

In applying the method of maximum likelihood to the regression model, we are essentially maximizing

$$L(\lambda) = -\frac{n}{2}\ln\left[SS_{res}(\lambda)\right]$$

or equivalently, we are minimizing $SS_{res}(\lambda)$.

An approximate $100(1-\alpha)\%$ confidence interval for λ consists of those values of λ that satisfy

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{\chi^2_\alpha(1)}{2}$$

where $\chi^2_\alpha(1)$ is the upper $\alpha\%$ point of the Chi-square distribution with one degree of freedom.

The approximate confidence interval is constructed using the following steps:

- Draw a plot of $L(\lambda)$ versus λ.
- Draw a horizontal line at the height

  $$L(\hat{\lambda}) - \frac{\chi^2_\alpha(1)}{2}$$

  on the vertical scale.
- This line would cut the $L(\lambda)$ at two points.
- The location of these two points on the λ-axis defines the two endpoints of the approximate confidence interval.
- If the sum of squares due to residuals is minimized and $SS_{res}(\lambda)$ versus λ is plotted, then the line must be plotted at the height

  $$SS^* = SS_{res}(\hat{\lambda})\exp\left[\frac{\chi^2_\alpha(1)}{n}\right]$$

  where $\hat{\lambda}$ is the value of λ which minimizes the sum of squares due to residuals. See how:

$$
\begin{aligned}
L(\hat{\lambda}) - \frac{\chi^2_\alpha(1)}{2} &= -\frac{n}{2}\ln\left[SS_{res}(\hat{\lambda})\right] - \frac{\chi^2_\alpha(1)}{2} \\
&= -\frac{n}{2}\left[\ln\left\{SS_{\text{Res}}(\hat{\lambda})\right\} + \frac{\chi^2_\alpha(1)}{n}\right] \\
&= -\frac{n}{2}\left[\ln\left\{SS_{res}(\hat{\lambda})\right\} + \ln\left\{\exp\left(\frac{\chi^2_\alpha(1)}{n}\right)\right\}\right] \\
&= -\frac{n}{2}\left[\ln\left\{SS_{res}(\hat{\lambda}).\exp\left(\frac{\chi^2_\alpha(1)}{n}\right)\right\}\right] \\
&= -\frac{n}{2}\ln SS^*.
\end{aligned}
$$

Using the expansion of exponential function as

$$
\begin{aligned}
\exp(t) &= 1 + t + \frac{t^2}{2!} + ... \\
&\approx 1 + t,
\end{aligned}
$$

we can approximate and replace $\exp\left[\frac{\chi^2_\alpha(1)}{n}\right]$ by $1+\frac{\chi^2_\alpha(1)}{n}$. So in place of $\exp\left[\frac{\chi^2_\alpha(1)}{n}\right]$ in applying the confidence interval procedure, we can use the following:

$$
\begin{aligned}
&1+\frac{Z^2_{\alpha/2}}{\gamma} && \left(\text{or } 1+\frac{Z^2_{\alpha/2}}{n}\right) \\
&\text{or } 1+\frac{t^2_{\alpha/2}}{\gamma} && \left(\text{or } 1+\frac{t^2_{\alpha/2}}{n}\right) \\
&\text{or } 1+\frac{\chi^2_{\alpha/2}}{\gamma} && \left(\text{or } 1+\frac{\chi^2_{\alpha/2}}{n}\right)
\end{aligned}
$$

where γ is the degrees of freedom associated with the sum of squares due to residuals.

These expressions are based on the fact that

$$\chi^2(1) = Z^2 \approx t^2_\gamma \text{ if } \gamma \text{ is small.}$$

It is debatable to use either γ or n but practically the difference is very little between the confidence interval results.

Box-Cox transformation was originally introduced to reduce the nonnormality in the data. It also helps in reducing the nonlinearity. The approach is to find out the transformations, which attempts to reduce the residuals associated with outliers and also reduce the problem of non-constant error variance if there was no acute nonlinearity, to begin with.

Transformation on explanatory variables: Box and Tidwell procedure

Suppose the relationship between y and one or more of the explanatory variables is nonlinear. Other usual assumptions normally and independently distributed study variable with constant variance are at least approximately satisfied.

We want to select an appropriate transformation on the explanatory variable so that the relationship between y and transformed explanatory variable is as simple as possible.

Box and Tidwell procedure describes a general analytical procedure for determining the form of transformation on x

Suppose that the study variable y is related to the power of explanatory variables. Box and Tidwell procedures for explanatory variables choose the variables as

$$z_{ij} = \begin{cases} \dfrac{x_{ij}^{\alpha_j} - 1}{\alpha_j} & \text{when } \alpha_j \neq 0,\ i = 1,2,..,n;\ j = 1,2,...,k \\ \ln x_{ij} & \text{when } \alpha_j = 0. \end{cases}$$

We need to estimate α_j's. Since the dependent variable is not being transformed, we need not worry about the changes of scale and minimize

$$\sum_{i=1}^{n} \left[y_i - \beta_0 - \beta_1 z_{i1} - ... - \beta_k z_{ik} \right]^2$$

by using the nonlinear least-squares techniques.

We consider this for simple linear regression model instead of a nonlinear regression model.

Assume y is related to $\xi = x^{\alpha}$ as

$$E(y) = f(\xi, \beta_0, \beta_1) = \beta_0 + \beta_1 \xi$$

where $\xi = \begin{cases} x^{\alpha} & \text{if } \alpha \neq 0 \\ \ln x & \text{if } \alpha = 0 \end{cases}$

where β_0, β_1 and α are the unknown parameters.

Suppose α_0 is the initial guess of constant α.

Usually, the first guess is $\alpha_0 = 1$ so that $\xi = x$ or no transformation is applied in the first iteration.

Expand about the initial guess in a Taylor series and ignoring terms of order higher them one gives

$$E(y) = f(\xi_0, \beta_0, \beta_1) + (\alpha - \alpha_0)\left\{\frac{df(\xi, \beta_0, \beta_1)}{d\alpha}\right\}_{\substack{\xi=\xi_0 \\ \alpha=\alpha_0}}$$

$$= \beta_0 + \beta_1 x + (\alpha - 1)\left\{\frac{df(\xi, \beta_0, \beta_1)}{d\alpha}\right\}_{\substack{\xi=\xi_0 \\ \alpha=\alpha_0}}.$$

Suppose the term $\left\{\frac{df(\xi, \beta_0, \beta_1)}{d\alpha}\right\}_{\substack{\xi=\xi_0 \\ \alpha=\alpha_0}}$ is known, then it can be treated just like as an additional explanatory variable. Then the parameters β_0, β_1 and α can be estimated by least-squares method.

The estimate of α can be considered as an improved estimate of the transformation parameter.

This term can be written as

$$\left\{\frac{df(\xi, \beta_0, \beta_1)}{d\alpha}\right\}_{\substack{\xi=\xi_0 \\ \alpha=\alpha_0}} = \left\{\frac{df(\xi, \beta_0, \beta_1)}{d\xi}\right\}_{\xi=\xi_0} \left\{\frac{d\xi}{d\alpha}\right\}_{\alpha=\alpha_0}.$$

Since the form of transformation is known, i.e., $\xi = x^{\alpha}$, so $\frac{d\xi}{d\alpha} = x \ln x.$

Furthermore

$$\left\{\frac{df(\xi, \beta_0, \beta_1)}{d\xi}\right\}_{\xi=\xi_0} = \frac{d(\beta_0 + \beta_1 x)}{dx} = \beta_1.$$

So β_1 can be estimated by fitting the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

by least-squares method.

Then an "adjustment" to initial guess $\alpha_0 = 1$ is computed by defining a second regression variable as

$$\omega = x \ln x$$

estimating the parameter in

$$E(y) = \beta_0^* + \beta_1^* x + (\alpha - 1)\beta_1 \omega$$

$$= \beta_0^* + \beta_1^* x + \gamma\omega$$

by least-squares.

This gives the following:

$$\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^* x + \hat{\gamma}\omega$$

$$\hat{\gamma} = (\alpha - 1)\hat{\beta}_1$$

$$\text{or} \quad \alpha_1 = \frac{\hat{\gamma}}{\hat{\beta}_1} + 1$$

as the revised estimate of α.

Note that $\hat{\beta}_1$ is obtained from $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and $\hat{\gamma}$ is obtained from $\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^* x + \hat{\gamma}\omega$.

Generally, $\hat{\beta}_1$ and $\hat{\beta}_1^*$ will differ.

This procedure may be repeated using a new regression $x^* = x^{\alpha_1}$ in the calculation.

This procedure generally converges rapidly.

Usually, the first stage result α_1 is a satisfactory estimate of α. The round-off error is a potential problem. If enough decimal places are not taken care, then the successive values of α may oscillate badly. If the standard deviation of error (σ) is large or the range of the explanatory variable is very small relative to its mean, then the estimator may face convergence problems. This situation implies that the data do not support the need for any transformation.