LINEAR REGRESSION ANALYSIS

MODULE – V

Lecture - 20

Correcting Model Inadequacies Through Transformation and Weighting

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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 study variable *(y)* in the model is Poisson random variable in a simple linear regression model, then its variance is same as mean. Since mean of *y* is related to 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 the suitable transformation, use *y** as a study variable in respective case.

The strength of a transformation depends on the amount of curvature present in the curve between study and explanatory variable. The transformation mentioned here range from relatively mild to 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 strong effect on the analysis.

In the presence of nonconstant variance, the OLSE will remain unbiased but will loose 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 inverse transformation is directly applied to the original values, then it gives an estimate of the median of the distribution of study variable instead of mean. So one needs to be careful while doing so.

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

Transformations to linearize the model

The basic assumption in linear regression analysis is that the relationship between study variable and explanatory variables is linear.

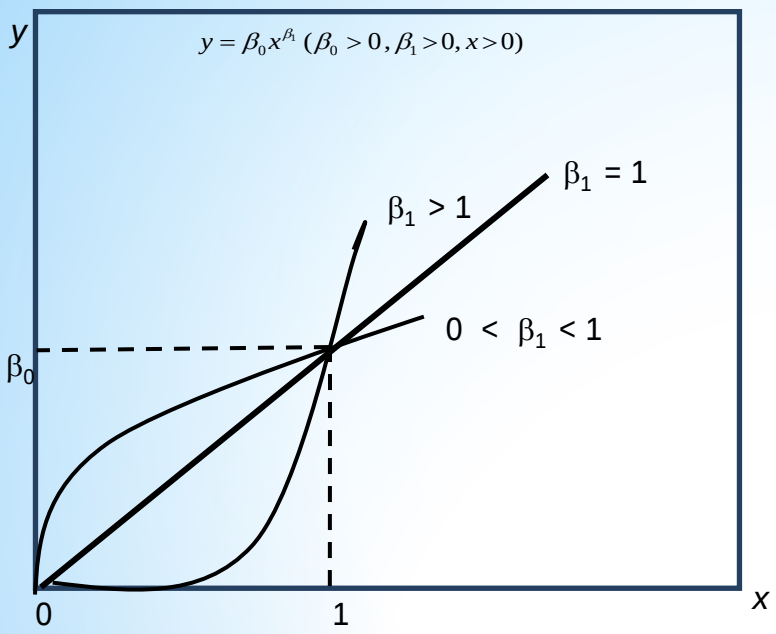
Suppose this assumption is violated. Such violation can be checked by scatter plot matrix, scatter diagrams, partial regression plots, lack of fit test etc.

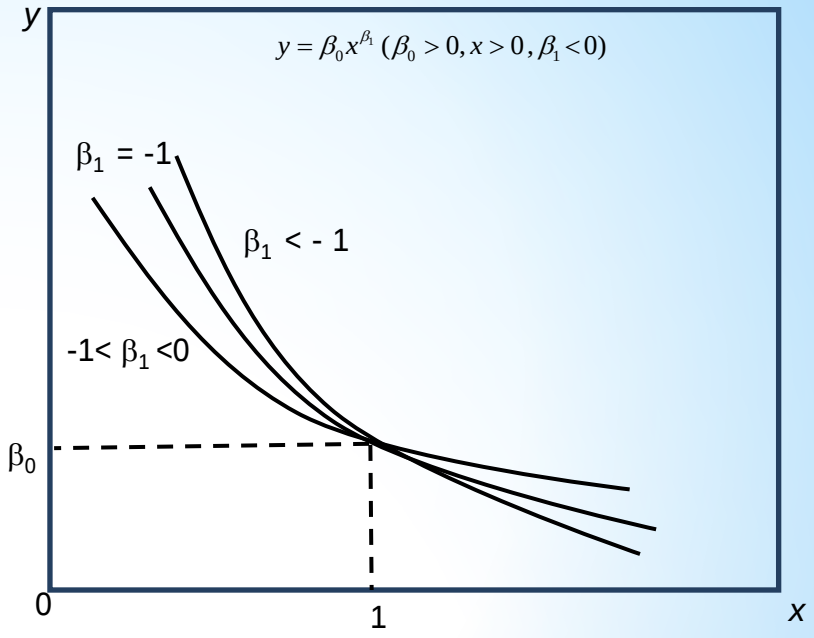
In some cases, a nonlinear model can be linearized by using a suitable transformation. Such nonlinear models are called **intrinsically or transformably linear.**

The advantage of transforming the nonlinear function into linear function is that the statistical tools are developed for the case of linear regression model. For example, exact tests for test of hypothesis, confidence interval estimation etc. are developed for the case of 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 $y = \beta_0 x^{\beta_1}$.

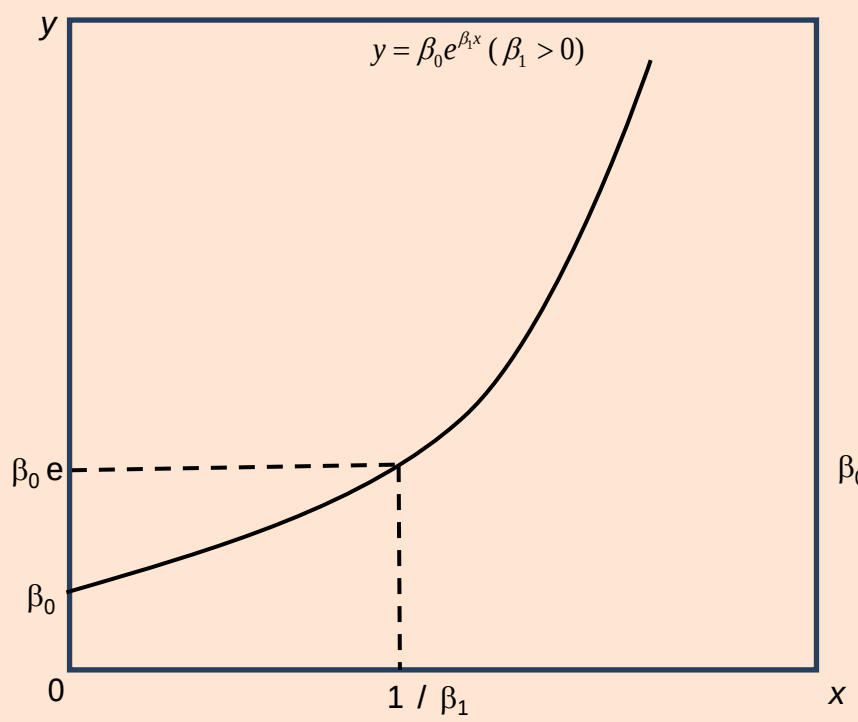
Using the transformation $y^* = \ln y,\ x^* = \ln x,$ i.e., by taking log on both sides, the model becomes

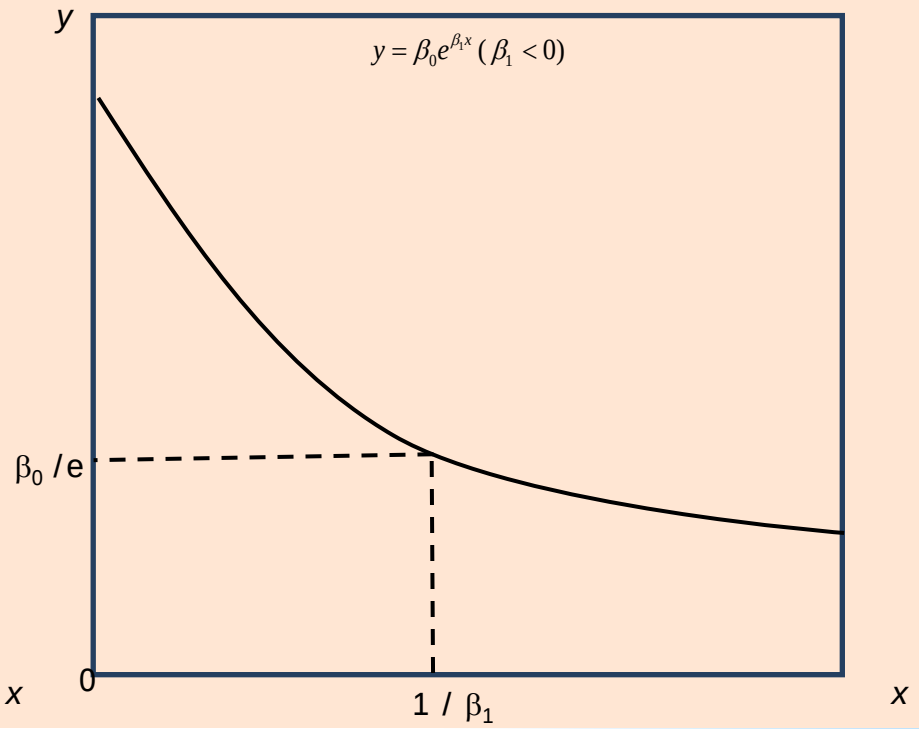
$$\log y = \log \beta_0 + \beta_1 \log x$$

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 $\log \beta_0$ 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,

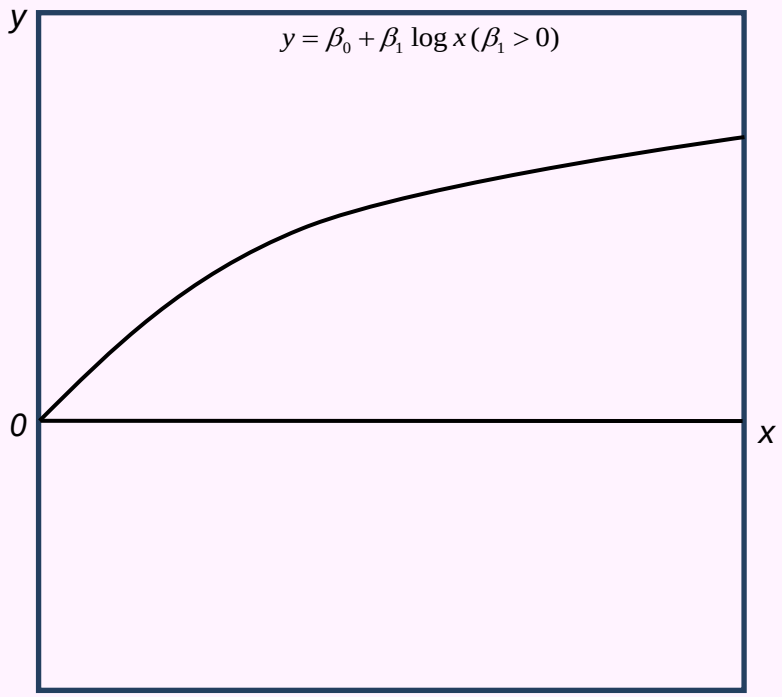
$$\ln y = \ln \beta_0 + \beta_1 x$$

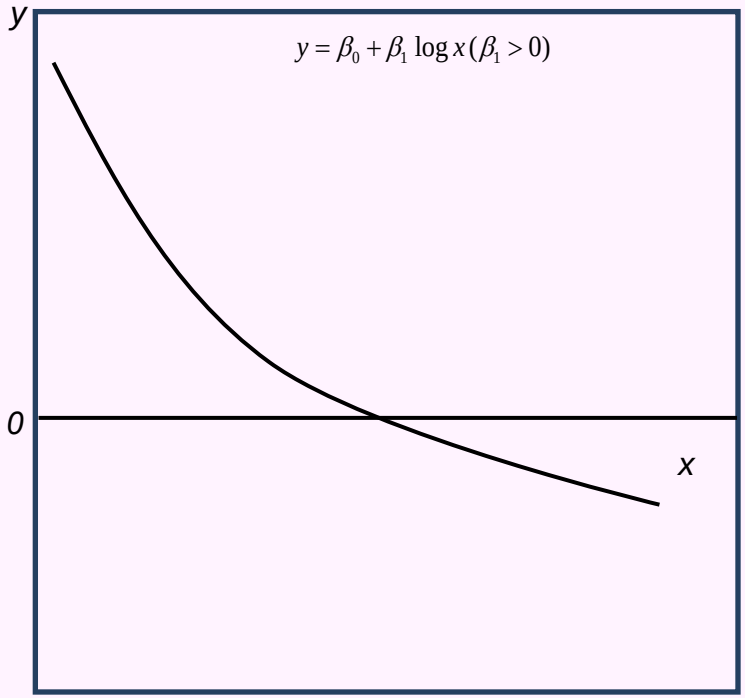
or $y^* = \beta_0^* + \beta_1 x$

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

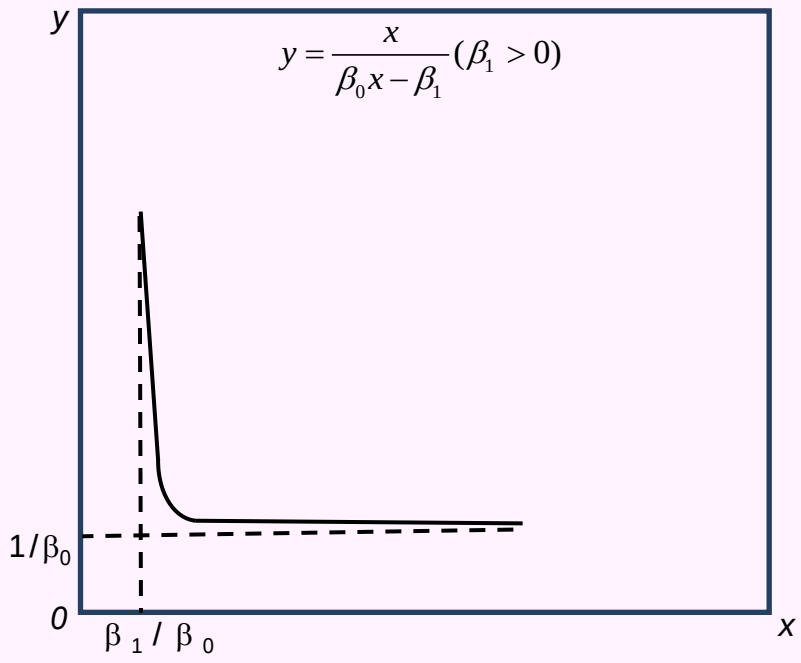
$$y = \beta_0 + \beta_1 \log x$$

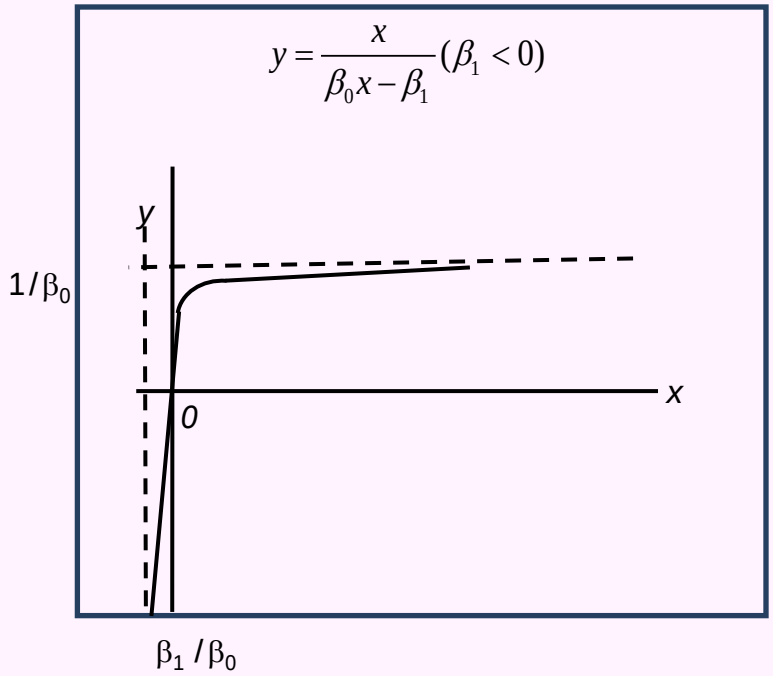
which can be written as

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

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

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 modelby using the transformation $y^* = \dfrac{1}{y},\ x^* = -\dfrac{1}{x}.$

- 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 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 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.