Chapter 9: Transformations

Some common transformations

Polynomial regression

ex: second degree polynomial in two variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \varepsilon_i$$

fourth degree in one variable

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i1}^3 + \beta_4 x_{i1}^4 + \varepsilon_i$$

However, polynomial regression can present 2 types of problems.

(1) The number of parameters increases rapidly with both the degree of the polynomial and the number of the (original) variables.

(2) Even with one independent variable, while high degree polynomials can be made to fit the data very well, such relationships usually mean

little and are often worthless for predictive purposes.

Multiplicative models.

By a multiplicative form of $x_1, \ldots x_K$, we mean a function of the form

$$y = A \prod_{j=1}^{K} x_j^{\beta_j}$$

A and β_j's are parameters.

Taking logs of both sides, we get

$$\log(y) = \beta_0 + \sum_{j=1}^{K} \beta_j \log(x_j) \quad (*).$$

$$\beta_0 = \log(A)$$

With an appropriate error term added, (*) can frequently be estimated by linear least squares.

This is a (linearized) Cobb-Douglas model commonly used in econometrics.

- Consistent yield if $\sum_{j=1}^{K} \beta_j = 1$.
- decreasing yield if $\sum_{j=1}^{K} \beta_j < 1$
- increasing yield if $\sum_{j=1}^{K} \beta_j > 1$.

eg: $y = A x_1^{\beta_1} x_2^{\beta_2}$

y: production. x_1: capital x_2: work.

A, β_1, β_2: technology.

$$\ln y = \beta_0 + \beta_1 \cdot \ln(x_1) + \beta_2 \ln(x_2) + \cdots \beta_K \ln(x_K)$$

$$\frac{\partial \ln(y)}{\partial x_j} = \beta_j \frac{\partial \ln(x_j)}{\partial x_j}$$

$$\Rightarrow \frac{\partial \ln(y)}{\partial y} \cdot \frac{\partial y}{\partial x_j} = \beta_j \cdot \frac{\partial \ln(x_j)}{\partial x_j}$$

$$\Rightarrow \frac{1}{y} \frac{\partial y}{\partial x_j} = \beta_j \cdot \frac{1}{x_j}$$

$$\Rightarrow \boxed{\beta_j = \frac{x_j}{y} \frac{\partial y}{\partial x_j}}$$

In economics, β_j, the exponent of the factor of production x_j is called the elasticity of output with respect to the variable x_j. It measures the response sensitivity of output to initial changes in these factors.

Multiplicative Errors.

In the econometrics literature, the usual procedure is to use the model

$$y_i = A\, x_{i1}^{\beta_1}\, x_{i2}^{\beta_2} \cdots x_{ik}^{\beta_k}\, \varepsilon_i$$

Such models can arise in many ways.

Suppose the real underlying model is $y_i = A \prod_{j=1}^{k} z_{ij}^{\beta_j}$, but for some reason, we do not know or cannot measure the z_{ij}'s.

Instead, we rely on the surrogates $x_{ij} = z_{ij}\eta_{ij}$ where η_{ij} is an unobservable random variable.

$$x_{ij} = z_{ij}\,\eta_{ij}$$

$$\Rightarrow x_{ij}^{\beta_j} = z_{ij}^{\beta_j}\,\eta_{ij}^{\beta_j}$$

$$\Rightarrow z_{ij}^{\beta_j} = x_{ij}^{\beta_j} \cdot \eta_{ij}^{-\beta_j}$$

$$\Rightarrow \quad y_i = A \prod_{j=1}^{k} x_{ij}^{\beta_j} \, \eta_{ij}^{-\beta_j}$$

$$= A \prod_{j=1}^{k} x_{ij}^{\beta_j} \prod_{j=1}^{k} \eta_{ij}^{-\beta_j}$$

$$= A \prod_{j=1}^{k} x_{ij}^{\beta_j} \cdot \varepsilon_i$$

$$\Rightarrow \quad \varepsilon_i = \prod_{j=1}^{k} \eta_{ij}^{-\beta_j}$$

Linearization of Model and parameter estimation.

econometric model:

$$y_i = A \, x_{i1}^{\beta_1} x_{i2}^{\beta_2} \cdots x_{iK}^{\beta_K} \, \varepsilon_i$$

$$\Rightarrow \quad \ln(y_i) = \ln A + \beta_1 \ln(x_{i1}) + \beta_2 \ln(x_{i2}) + \cdots + \beta_K \ln(x_{iK}) + \ln(\varepsilon_i) \quad (*)$$

Suppose that $\eta_i = \ln(\varepsilon_i) \sim N(\mu, \sigma^2)$

Then it can be shown that $1 = E(\varepsilon_i) = E(e^{\eta_i})$

$= e^{\mu + \frac{1}{2}\sigma^2}$

$\Rightarrow e^{\mu + \frac{1}{2}\sigma^2} = 0 \qquad E[\ln(\varepsilon_i)] = \mu = -\frac{\sigma^2}{2}.$

we need to rewrite ⊛ as

$$\ln(y_i) = \left(\ln(A) - \frac{\sigma^2}{2}\right) + \beta_1 \ln(x_{i1}) + \cdots + \beta_K \ln(x_{iK}) + \left[\ln(\varepsilon_i) + \frac{\sigma^2}{2}\right]$$

$E\left[\ln(\varepsilon_i) + \frac{\sigma^2}{2}\right] = 0.$

$\mathrm{var}\left[\ln(\varepsilon_i) + \frac{\sigma^2}{2}\right] = \mathrm{var}[\ln(\varepsilon_i)] = \sigma^2$

OLS estimate of β: $\hat{\beta}_0 = \ln(\hat{A}) - \frac{s^2}{2}.$

$$\hat{A} = e^{\hat{\beta}_0 + \frac{s^2}{2}}$$

Choosing Transformations.

Graphical method : One independent variable

When there is only one independent variable, we can examine a plot of the dependent variable against it. Sometimes this plot will immediately suggest a course of action : the use of the logit function or broken line regression.

If no obvious course is apparent, the following method has been found useful.

Ladder of transformation

Divide the range of the independent variable into three portions, making a good compromise between getting equal numbers of data points in each portion. and making the three portions roughly equal.

For each of the three sets of data points thus created, find a point (which may or may not be one of the data points) which is good representative of the set. For each set, a good choice is the point whose coordinates are the medians of the x and y values for the points in the set. Find the slope of line joining the first two points (going from left to right) and slope for the line joining the last two. If the two scopes are equal, then the data points should be describing a straight line. If not, the middle of three points will be below (the convex case) or above (the concave case) the line joining the other two.

$-\frac{1}{y^2}$	$\Uparrow$	$\vdots$
$-\frac{1}{y}$	If convex	x^4
$-\frac{1}{\sqrt{y}}$	up the Ladder	
$\log(y)$		x^3
$\sqrt{y}$		x^2
y	you are here	x
y^2	If concave	$\sqrt{x}$
y^3	Down the	$\log x$
y^4	Ladder.	$-\frac{1}{\sqrt{x}}$
$\vdots$		$-\frac{1}{x}$
		$-\frac{1}{x^2}$
		$-\frac{1}{x^3}$

We may now transform either the dependent variable y or the independent variable x using the ladder of power transformations. If the three points are in a convex configuration we move up the ladder; if they are in a concave configuration we move down. In either case, we apply the transformations to the chosen coordinates (x or y) of the three points.

If the two slopes mentioned above become roughly equal we stop; if the configuration changes from convex to concave or vice versa, we have gone too far.

Graphical method: many independent variables.

Component plus residual plots

Consider the estimated model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK} + e_i.$$

$$\Rightarrow \hat{y}_i - \hat{\beta}_0 - \sum_{\substack{j=1 \\ j \neq m}}^{K} \hat{\beta}_j x_{ij} = \hat{\beta}_m x_{im} + e_i$$

Thus $\hat{\beta}_m x_{im} + e_i$ is essentially $\hat{y}_i$ with the linear effects of the other variables removed.

Plotting $\hat{\beta}_m x_{im} + e_i$ against x_{im} estimates some of the clutter mentioned above.

Wood (1973) called these plots component plus residual plots since $\hat{\beta}_m x_{im}$ may be seen as a component of $\hat{y}_i$.

The plot has also been called a partial residual plots.

If the full model is true, the graph shows a linear trend and the variable x_m is involved in the model in a linear fashion. If on the other hand the graph shows a nonlinear trend according to some function (f), it will be better to replace x_m with $f(x_m)$

Analytic methods: Transforming the response.

A good transformation should make residuals smaller. However, when we transform the dependent variable so that y_i becomes $f(y_i)$, we also change scales. Therefore, we cannot simply compare the s^2's after making various transformation. We need to make some adjustments to them. Such adjustments are used in the methods described below.

The Box and Cox method.

For the situations where all $y_i > 0$, Box and Cox considered following family of transformations:

$$y_i^{(\lambda)} = \begin{cases} (y_i^{\lambda} - 1) / \lambda & \text{when } \lambda \neq 0 \\ \log(y_i) & \text{when } \lambda = 0. \end{cases}$$

To simplify the notations, we note $y_i^{(\lambda)}$ by z_i.

Suppose that $z_i \sim N(x_i'\beta, \sigma^2)$

$$\prod_{i=1}^{n} f(z_i)\,dz_i = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(z - X\beta)'(z - X\beta)\right] J$$

$$z = (z_1, \dots z_n)'$$

$$J = \prod_{i=1}^{n} dz_i \qquad dz_i = \left(\frac{dz_i}{dy_i}\right) dy_i = \left(\frac{dy_i^{(\lambda)}}{dy_i}\right) dy_i$$

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

In the initial scale, the density function of

$y = (y_1, y_2, \dots y_n)'$ is

$$\prod_{i=1}^{n} g(y_i)\,dy_i = (2\pi\sigma)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)\right].$$

$$\prod_{i=1}^{n}\left(\frac{dy_i^{(\lambda)}}{dy_i}\right)\left(\prod_{i=1}^{n} dy_i\right)$$

$$\Rightarrow \prod_{i=1}^{n} g(y_i)\,dy_i = (2\pi\sigma)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y^{(\lambda)}-X\beta)'(y^{(\lambda)}-X\beta)\right] \cdot \left(\prod_{i=1}^{n} y_i\right)^{\lambda-1} \left(\prod_{i=1}^{n} dy_i\right)$$

To obtain $\beta, \sigma^2, \lambda,$

$$\max \quad l(\lambda) = -\frac{1}{2} n \ln[S^2(\lambda)]$$

$$= -\frac{1}{2} n \ln(\hat{\sigma}^2) + (\lambda-1) \cdot \sum_{i=1}^{n} \ln(y_i) \quad (*)$$

where $\hat{\sigma}^2 = \frac{1}{n} y^{(\lambda)'}[I-H]y^{(\lambda)}$, i.e. $\hat{\sigma}^2$ is the sum of squares of the residuals divided by n when $y_i^{(\lambda)}$'s are used as observations.

$S^2(\lambda)$ is essentially the S^2 for the transformed model adjusted for change of scale for the dependent variable $y^{(\lambda)}$.

The maximization is best carried out by computing ⊛ for several values of λ. We shall sometimes call $-\frac{1}{2}n\ln[s^2(\lambda)]$ the Box-Cox objective function.

Let λ_{max} be the value of λ which max ⊛

Then under fairly general conditions, for any other λ,

$$K = n\ln[s^2(\lambda)] - n\ln[s^2(\hat{\lambda}_{max})] \sim \chi_1^2.$$

The Box-Cox transformations were originally introduced to reduce non-normality in data.

⊛ when $y_i < 0$.

$$y_i^{(\lambda)} = \begin{cases} \dfrac{[(y_i+a)^{\lambda} - 1]}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y_i+a) & \text{if } \lambda = 0. \end{cases}$$