University of Illinois
Fall 1998

Economics 472

Department of Economics
Roger Koenker

Lecture 2
Transformations and the Specification of Econometric Models

A fundamental aspect of interpreting any parametric statistical model is choice of functional form. Let's begin a consideration of this topic with the following simple example. Suppose

$$\log y_i = \alpha + \beta \log x_i + u_i$$

but unaware of this convenient formulation we instead estimate

$$y_i = a + bx_i + v_i.$$

What relationship does $(\hat{a}, \hat{b})$ bear to (α, β) in the original model and can we hope to say anything reasonable having made this initial specification error?

In Figure 1, we can examine a specific version of this situation in which $(\alpha, \beta) = (1, .5)$ and the variance of u_i is quite small. Clearly we don't do a very good job of estimating the curve represented by the observed points by the line indicating the least squares fit, but it is useful to look at this more carefully.* On a more optimistic note it might appear that the slope of the linear fit might provide a decent approximation to the tangent of the curve at a point roughly corresponding to $\bar{x}$. Figure 2 illustrates this phenomenon on the elasticity scale. Were we to estimate the log-linear model we would have an easily interpreted constant elasticity estimate. However, since we have estimated the model in the linear form, the implied elasticity of y with respect to x varies as we move along the fitted line. More explicitly, the elasticity is defined as

$$\eta = \frac{dy}{dx}\ \frac{x}{y}$$

and according to the linear specification the derivative, $dy/dx = b$ is constant, so the natural estimate of the elasticity of y with respect to x, at any point x, is given by

$$\hat{\eta}(x) = \hat{b} \cdot \frac{x}{\hat{y}(x)}$$

where $\hat{y}(x) = \hat{a} + \hat{b}x$. If we were going to offer only one such elasticity estimate for expository purposes, we would typically choose $x = \bar{x}$, but sometimes it is useful to choose several such points of evaluation for purposes of comparison. Recall $\hat{y}(\bar{x}) = \bar{y}$ as long as the estimated model has an intercept. This is done for each of the observed values of x in Figure 2. The horizontal line at $\beta = .5$ is the "true" elasticity according to which the data was generated, while the dots represent $\hat{\eta}(x)$ at the various deserved x's. Obviously these estimates are rather poor in the extremes, but reasonably good in the center of the x's. The two vertical lines represent the arithmetic and geometric means of x and we note that one yields a small overestimate while the other yields a small underestimate of β. This is the first of many lessons which can be roughly formulated by the

Maxim: It is dangerous to draw inferences too far away from the center of your data.

*One way to do this is to ask: suppose the x_i's are generated randomly from some distribution, F, and that $E(y|x) = g(x)$, then $(\hat{a}, \hat{b})$ solves $\min E_x(g(x) - a - bx)^2$, i.e., $\hat{a} + \hat{b}x$ is the best linear approximation (a, b) to $g(x)$ in quadratic mean.

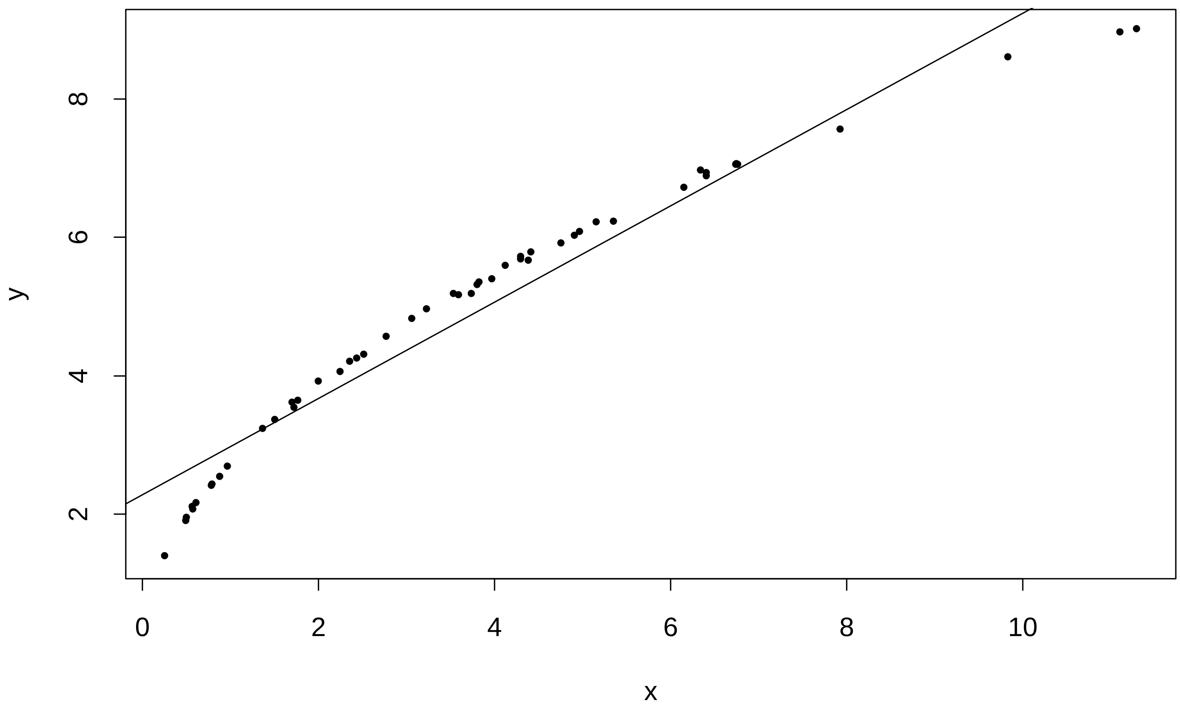

FIGURE 1. A linear fit to a log-linear model: The figure illustrates 50 observations from a log-linear model and a superimposed least-squares linear fit of the observations. Note that the fit provides a rough estimate of the tangent of the curve near the "center" of the x's, but cannot be considered very reliable unless the range of the x's is quite restricted.

A corollary, which is often offered as advice to young novelists is "Write what you know," another pithy corollary is "Extrapolate at your peril." A nice introduction to a more general formulation of these issues is White (1980).

Having seen this example it is natural to ask whether there is a systematic strategy for deciding on appropriate functional forms. This is obviously a big topic and I will try only to briefly survey the basic idea in the simplest bivariate regression setting.

The classical approach to dealing with this ‘ ̀transformation problem" involves the family of power transformations

$$h(x,\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x & \lambda = 0 \end{cases}$$

Exercise: Verify using L'Hôpital's rule that

$$\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = \log x$$

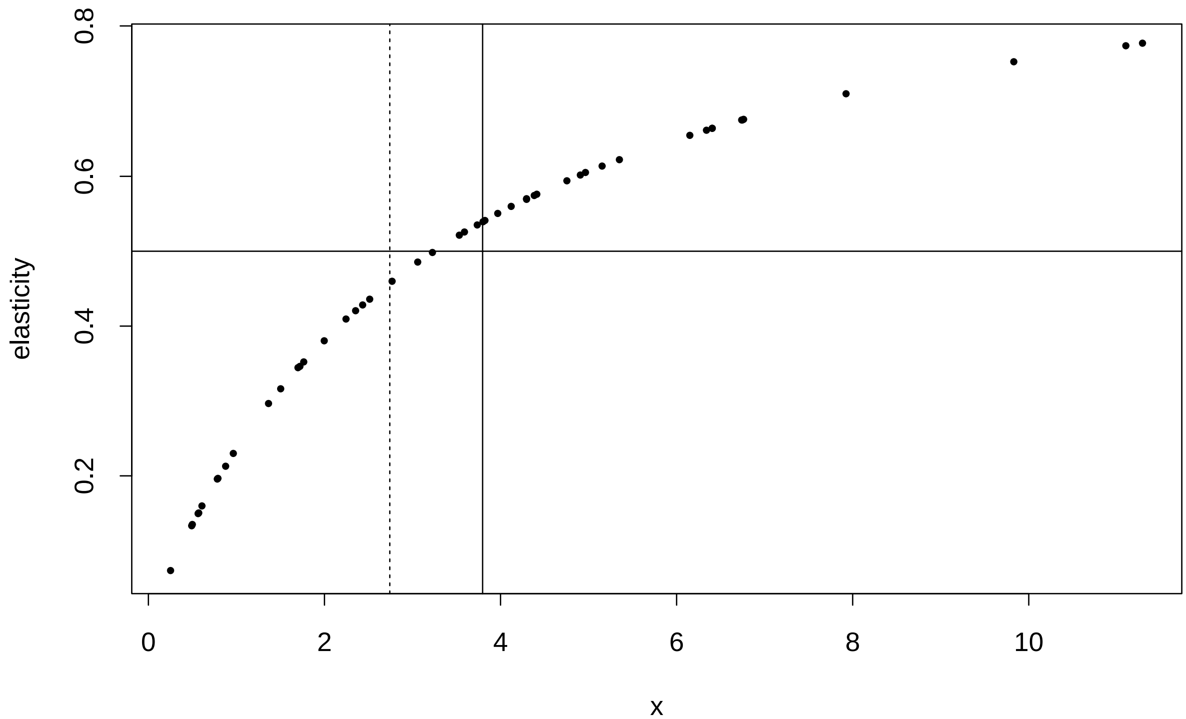

FIGURE 2. A linear fit to a log-linear model: This figure illustrates the bias introduced in estimating the elasticity parameter of the log-linear model by using the estimated linear model. The points in the figure represent elasticities implied by the fitted *linear* model at each of the observed x's. The horizontal line at $\beta = .5$ represents the true, constant elasticity for the model, and the two vertical lines indicate the mean (solid) and geometric mean (dotted) of the x's. Thus, at the mean of the x's the linear model slightly overestimates the elasticity, and at the geometric mean it slightly underestimates it.

Answer:

$$
\begin{aligned}
\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} &= \frac{\frac{d}{d\lambda}(e^{\lambda \log x} - 1)}{1}\Big|_{\lambda=0} \\
&= e^{\lambda \log x} \cdot \log x|_{\lambda=0} \\
&= \log x
\end{aligned}
$$

The family of Box-Cox transformations is illustrated in Figure 3 for 6 different values of λ. The family is quite flexible and useful, but it is somewhat limited because it is only fully applicable for $x \geq 0$. It has been suggested that one might extend the definition using

$$\lambda(x) = (|x|^\lambda \text{ sgn } (x) - 1)/\lambda$$

but this behaves rather strangely and is rarely used in applications.

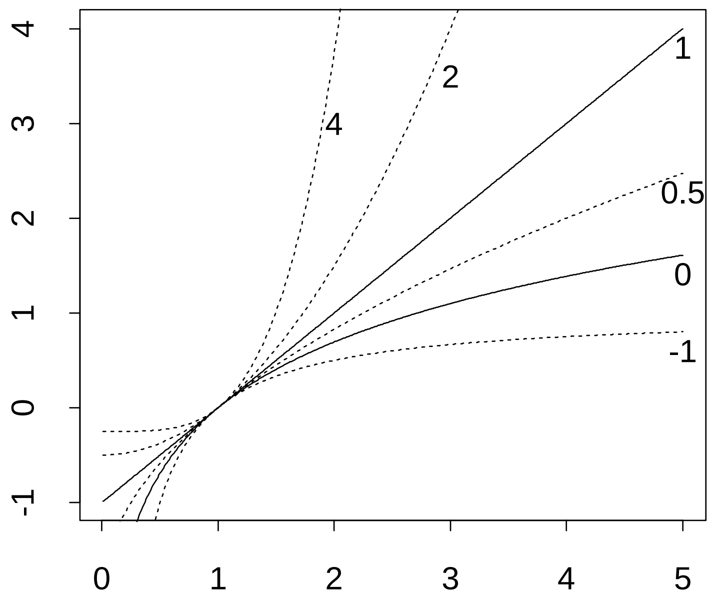

FIGURE 3. The Box-Cox Power Transformations: The Figure illustrates 6 versions of the Box-Cox Power family of transformations. Note that the log transformation fits nicely into the family with $\lambda = 0$.

As an exercise in reviewing some basic ideas about maximum likelihood estimation, let's consider, following Box and Cox (1964), the problem of estimating the model

$$h(y_i, \lambda) = x_i\beta + u_i$$

assuming that $\{u_i\}$ is iid $\mathcal{N}(0, \sigma^2)$. The log likelihood is

$$\ell(\beta, \lambda, \sigma) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(h(Y_i, \lambda) - x_i\beta)^2 + \log|J|$$

where $J = \prod_{i=1}^n |\partial\lambda(y_i)/\partial y_i|$ is the determinant of the transformation from u to $h(y, \lambda)$. Note that

$$\frac{\partial h(y_i, \lambda)}{\partial y_i} = y_i^{\lambda - 1}$$

so

$$\log|J| = (\lambda - 1)\sum_{i=1}^n \log(y_i).$$

Concentrating the likelihood we have

$$\ell(\lambda, \sigma) = -\frac{n}{2}\log\hat{\sigma}^2 + \log J + K$$

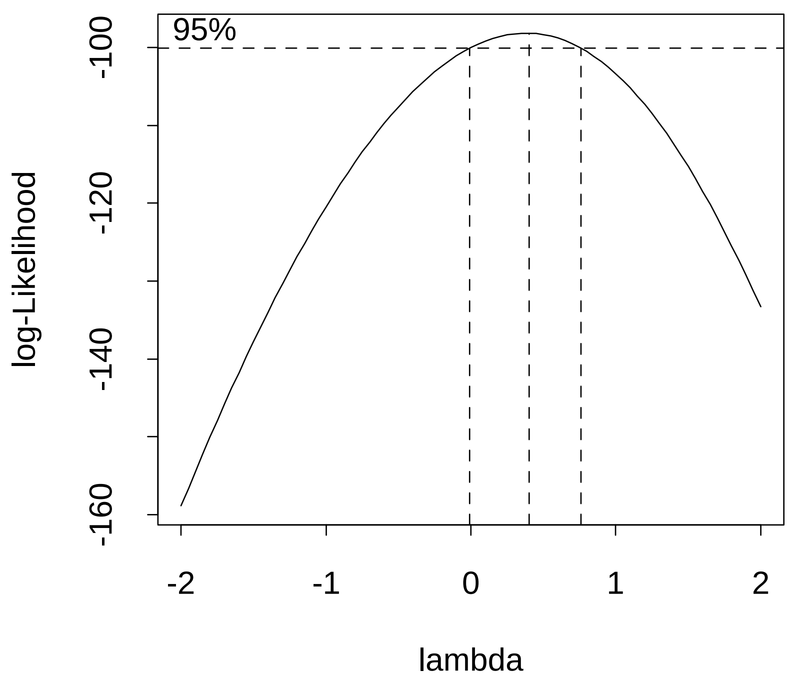

FIGURE 4. The Box-Cox Power Transformation: The Figure illustrates the profile log likelihood for a simple bivariate linear model, the confidence interval indicated for λ is based on the asymptotic theory of the likelihood ratio statistic.

where K doesn't depend on the parameters. Now let

$$\begin{aligned} z(y,\lambda) &= (y^\lambda - 1)/(\lambda J^{1/n}) \\ &= \begin{cases} (y^\lambda - 1)/(\lambda \tilde{y}^{\lambda-1}) & \lambda \neq 0 \\ \tilde{y}\log y & \lambda = 0 \end{cases} \end{aligned}$$

where $\tilde{y} = (\prod y_i)^{1/n}$ denotes the geometric mean of y_i's. Note $n^{-1}\log J = (\lambda - 1)n^{-1}\sum \log y_i = (\lambda - 1)\log \tilde{y}$.

Claim: $\ell(\lambda) = -\frac{n}{2}\log(R(\lambda)/n) + K$ where $R(\lambda) = z'(I - P_x)z$.

Pf.: We will show that $\hat{\sigma}^2(\lambda)/J^{2/n} = R(\lambda)/n$. Since $\hat{\sigma}^2(\lambda) = h'(I - P_x)h/n = S(\lambda)/n$ we need to show that $S(\lambda)/J^{2/n} = R(\lambda)$. But

$$\frac{S(\lambda)}{J^{2/n}} = \frac{S(\lambda)}{\tilde{y}^{2(\lambda-1)}} = z'(I - P_x)z \qquad \square$$

The function $\ell(\lambda)$ which we used to call the concentrated log likelihood we now call the profile (log) likelihood, terminology I believe introduced by Cox. The profile likelihood provides an extremely convenient and powerful means of doing inference in many problems. In the simple

Box-Cox problem under consideration we would often like to test the hypothesis $H_0 : \lambda = \lambda_0$. This is effectively done using the fact (whose proof is deferred to 476) that under H_0,

$$(*) \qquad \tau(\lambda_0) = 2(\ell(\hat{\lambda}) - \ell(\lambda_0)) \rightsquigarrow \chi_1^2$$

where $\hat{\lambda}$ denotes the maximum likelihood estimate of λ. The limiting behavior of this likelihood ratio statistic can also be used to construct confidence intervals for λ: we simply find the set of λ_0 such that $\tau(\lambda_0)$ fails to reject at a specified level of confidence. This is illustrate in Figure 4.

Sometimes we would rather not go to the bother of estimating the Box-Cox model, but instead we would like to estimate some "preferred" form and then test whether this choice of λ is reasonable. A simple test suggested by David Andrews (1971) handles this situation, and since it nicely illustrates an important principle of diagnostic test design we will develop it in some detail. Consider

$$h(y, \lambda) = x_i'\beta + u_i$$

with $\lambda = 1$ as our "preferred" value. Expanding in Taylor series we have

$$\begin{aligned} h(y, \lambda) &= y - 1 + (\lambda - 1)\frac{d\lambda(y)}{d\lambda}|_{\lambda=1} \\ &= (\lambda - 1)y \log y - (y - 1) \end{aligned}$$

Thus, for λ close to one,

$$y - 1 \simeq x'\beta + (\lambda - 1)y \log y$$

this seems rather strange since it suggests that we should regress y on $y \log y$ – this is clearly unsound. But if we instead proceed in two steps:

1. Estimate the linear model and compute $\hat{y}_i = x_i\hat{\beta}$ for $i = 1, \ldots, n$ and then
2. Reestimate the augmented model

   $$y_i = x_i'\beta + \gamma\hat{y}_i \log \hat{y}_i$$

   and test $H_0 : \gamma = 0$.

This procedure, in effect provides one-step approximation to the mle for λ i.e., $\hat{\lambda} = \hat{\gamma} + 1$.

Question: What about the 1?

Exercises (Review) For the OLSE $\hat{\beta}$ show (1.) $\hat{\beta}(\sigma y + X\gamma, X) = \sigma\hat{\beta}(y, X) + \gamma$, and (2.) $\hat{\beta}(y, XA) = A^{-1}\hat{\beta}(y, X)$.

On the other hand if $H_0 : \lambda = 0$ is the preferred version, then at $\lambda = 0$, we have,

$$\frac{d\lambda(y)}{d\lambda}|_{\lambda=0} = \frac{1}{2}(\log y)^2$$

so now we would fit

$$\log(y) = x'\beta + \delta \cdot (\widehat{\log y})^2$$

so here δ estimates $(1/2)\lambda$ under the alternative hypothesis.

Conflicting Objectives of Transformations

We have 3 possibly conflicting objectives in choosing a transformation. We would like the transformation to (simultaneously) yield a model

(i)	which is linear in parameters
(ii)	homoscedastic
(iii)	has approximately "normal" conditional density

Carroll and Ruppert have proposed a more general strategy which they call "transforming both sides". We begin with a model like

$$y_t = f(x_t, \beta).$$

One might think of this as the *systematic* part of the model before any noise is introduced. Now we might consider models of the form

$$h(y_t, \lambda) = h(f(x_t, \beta), \lambda) + u_t$$

This is quite different than the Box-Cox transformation we considered above. Here $f(x_t, \beta)$ is intended to deal with the non-linearity, while h is *hopefully* going to transform to homoscedastic and normal errors. How does $h(\cdot)$ work?

Suppose y_i has $E(y_i|x_i) = \mu_i, V(y_i|x_i) = \sigma_i^2$ and $\sigma_i = \sigma g(\mu_i)$, then

$$\begin{aligned} V(h(y_i)) &\simeq E(h(y_i) - h(\mu_i))^2 \\ &\simeq (h'(\mu_i))^2 E(y_i - \mu_i)^2 \\ &= (h'(\mu_i))^2 \sigma^2 (g(\mu_i))^2 \end{aligned}$$

[Note these approximations depend on σ being "small"]
Thus if we were to choose h so that

$$h'(\mu_i) = \frac{1}{g(\mu_i)}$$

then we would have (approximate) homoscedasticity. For example, in Poisson cases

$$g(\mu) = \mu^{1/2}$$

so

$$h(\mu) = 2\mu^{1/2} \Rightarrow h'(\mu) = \frac{1}{\mu^{1/2}}$$

and

$$g(\mu) = \mu \Rightarrow h(\mu) = \log(\mu) \Rightarrow h'(\mu) = \frac{1}{\mu}$$

and

$$g(\mu) = \mu^{(1-\lambda)} \Rightarrow h(\mu) = y^{(\lambda)} \Rightarrow h'(\mu) = \mu^{\lambda - 1}$$

Another way to look at this is to say that if σ^2 is small relative to the variability of μ_i's, then

$$h(y_i) = h(\mu_i) + h'(\mu_i)(y_i - \mu)$$

For this order of approximation we are back to a "simple" heteroscedastic model,

$$y_i = \mu_i + \sigma h'(\mu_i)\varepsilon_i$$

Note that the interpretation of the β's is quite different in this setup than in the classical Box-Cox setup. There the β's don't mean much independent of λ – recall $\partial y / \partial x$ expression, – but here they do.

Transformation *and* weighting: Consider the model

$$h(y_i, \lambda) = h(f(y_i, \beta), \lambda) + \sigma g(\mu_i(\beta), z_i\theta)\varepsilon_i$$

Now we can think of $g(\cdot)$ as modeling the heteroscedasticity and $h(\cdot)$ being exclusively for achieving normality, while $f(\cdot)$ fixes the non-linearity in the conditional mean relationship. This model is considerably more complicated to estimate, but may arise naturally in the process of diagnostic checking.

Interpreting Transformed Models

It is very important to be clear about what parameters "*mean*" in transformation models, In the normal linear model

$$y_i = x_i + \beta + \sigma\varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0,1)$$
$$P(y_i < y|x_i) = \Phi((y - x_i\beta)/\sigma)$$
$$Q_{y_i}(p|x_i) = x_i\beta + \sigma\Phi^{-1}(p)$$

In the Box-Cox framework we have,

$$h(y,\lambda) = x_i\beta + \sigma\varepsilon$$

so the pth quantile of $y_i|x_i$ is

$$y_i = h_\lambda^{-1}(x_i\beta + \sigma\varepsilon)$$
$$Q_{y_i}(p|x_i) = h_\lambda^{-1}(x_i\beta + \sigma\Phi^{-1}(p))$$

Thus if we wanted to estimate the effect of a change in x_{ij} on median y_i, we would write

$$\frac{\partial}{\partial x_{ij}} Q_{y_i}(1/2|x_i) = \frac{\partial}{\partial x_{ij}}[h_\lambda^{-1}(x_i\beta + \sigma\Phi^{-1}(p))]$$

For example, if

$$h_\lambda(h_i) = \frac{y_i^\lambda - 1}{\lambda}$$

then,

$$\begin{aligned} y_i^\lambda &= \lambda h_i + 1 \\ y_i &= (\lambda h_i + 1)^{1/\lambda} \\ y_i &= (\lambda(x_i\beta + \sigma\Phi^{-1}(p)) + 1)^{1/\lambda} \\ \frac{\partial y_i}{\partial x_{ij}} &= \frac{1}{\lambda}(\lambda(x_i\beta + \sigma\Phi^{-1}(p)) + 1)^{\frac{1}{\lambda}-1} \cdot \lambda\beta_j = (\lambda(x_i\beta + \sigma\Phi^{-1}(p)) + 1)^{\frac{1}{\lambda}-1}\beta_j \end{aligned}$$

This could then be used to generate a confidence interval. Note that models for expectations are less convenient here since $E(h(y)) \neq h(Ey)$.

Transformations for Proportions

Often we are interested in estimating models of proportions, for example, Engel Curves for proportions of expenditure, unemployment rates, etc. Two simple alternatives are logit: $h(y) = \log(y/(1-y))$ or more generally $h(y,\lambda) = y^\lambda - (1-y)^\lambda$ folded power transformation. Note $\lim_{\lambda\to 0} h(y,\lambda) = \log(y/(1-y))$

References

Andrews, D. A note on the selection of data transformations, *Biometrika*, 58, 249-54.

Box, G.E.P. and D.R. Cox (1964), Analysis of Transformations, (with discussion), *JRSS(B)*, 26, 211-52.

Carroll, R. and D. Ruppert (1988), *Transformation and Weighting in Regression*, Chapman-Hall.

White, H. (1980), Using least squares to approximate unknown regression functions, *IER*, 21, 149-170.