1 Transforming the Predictors

The material in this lecture is covered in chapter 7 of your textbook, and appendix 11. Our main objective, here, is to transform the either the predictors or the response variables in a regression model, with two objectives:

i. We may wish to increase the linearity of the mean function.

ii. We may also want to improve the normality of the response.

All the methods presented here are restricted to positive random variables, although some related techniques can be used to transform non-positive variables.

1.1 Power Transformations

The family power transformations are defined for any \( \lambda \in \mathbb{R} \),

\[
\psi_P(X; \lambda) := \begin{cases} 
X^\lambda & \text{if } \lambda \neq 0, \\
\log(X) & \text{if } \lambda = 0.
\end{cases}
\]

This family of transformation, however, suffers from two limitations:

i. It is not continuous with respect to \( \lambda \), since as \( \lambda \to 0 \) we have \( \psi_P(X; \lambda) \to 1 \), which is distinct from \( \log(X) \).

ii. Moreover, the direction of the association between the covariate and the response will change, when \( \lambda < 0 \). That is, \( \psi_P(X; \lambda) \) is not monotonic increasing in \( X \), for every choice of \( \lambda \).

1.2 Scaled Power Transformations

We resolve this problem by defining a family of scaled power transformations, of the form,

\[
\psi(X; \lambda) := \begin{cases} 
(X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\
\log(X) & \text{if } \lambda = 0.
\end{cases}
\]

This particular transformation has the advantage of being continuous for every \( \lambda \). In particular, we have

\[
\lim_{\lambda \to 0} \psi(X; \lambda) = \log(X).
\]
In addition, the scaled power transformation preserves the direction of the association, such that a positive slope between $Y$ and $X$ will remain positive when considering $Y$ and $\psi(X; \lambda)$, for any choice of $\lambda \in \mathbb{R}$. That is, $\psi(X; \lambda)$ is monotonic increasing in $X$, for every choice of $\lambda$.

If we choose to transform the predictor variable, $X$, we will study the mean function

$$E[Y|X = x] = \beta_0 + \beta_1 \psi(X; \lambda).$$

In general, the value of $\lambda$ will be estimated by computing the $\text{RSS}(\lambda)$, given $\hat{\beta}_0$ and $\hat{\beta}_1$, for a small set of candidate values,

$$\lambda \in \{-2, -1, -1/2, 0, 1/2, 1, 2\}.$$

However, if we were to apply this transformation to the $y_i$'s, and then compute the $\text{RSS}(\lambda)$, we would not be preserving the units of measurements, thereby automatically favoring very small values of $\lambda$. The Box-Cox transformation addresses this problem by calibrating the transformations.

## 2 Transforming the Response Variable

### 2.1 Transforming for Normality

The Box-Cox transformation is a modified power transformation that can be applied to the response variable. When transforming the response in a regression model, we are faced with the problem that this particular variable is given a distribution, and therefore any transformation must respect the random generating process that had produced the data under scrutiny. Our main assumption, here, is that for some unknown $\lambda$, the transformed response variable,

$$\tilde{y}_i := \psi(y_i; \lambda)$$

is normally distributed. Here, $\psi(y_i; \lambda)$ is defined as the scaled power transformation, as in the previous section. Observe that this scaled power transformation is a continuous map from the positive real numbers to the entire real line, $\psi : \mathbb{R}^+ \mapsto \mathbb{R}$. Using the standard change of variable formula, this implies that the integral for any pdf $p(\tilde{y}_i)$ of the random variable, $\tilde{Y}_i$, can be written as

$$\int_{\psi(\mathbb{R}^+)} p(\tilde{y}_i) d\tilde{y}_i = \int_{\mathbb{R}^+} p(\psi(y_i; \lambda)) \frac{d\psi}{dy_i} dy_i.$$

The full likelihood function for the transformed set of $n$ observations, $\tilde{y}_i$, is therefore given by

$$L(\beta, \sigma^2, \lambda) := \prod_{i=1}^{n} \left(N(\tilde{y}_i; \mathbf{x}_i^T \beta, \sigma^2) J(y_i; \lambda)\right),$$

where $N(\tilde{y}_i; \mathbf{x}_i^T \beta, \sigma^2)$ denotes the Normal density estimated at the transformed observation $\psi(y_i, \lambda)$, for every $i = 1, \ldots, n$. Altogether, taking the product over all these terms, we obtain

$$L(\beta, \sigma^2, \lambda) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{y} - X\beta)^T(\tilde{y} - X\beta)\right\} J(y; \lambda),$$

where the Jacobian of the transformation is

$$J(y; \lambda) = \prod_{i=1}^{n} \frac{d\psi_i}{dy_i}.$$
By straightforward differentiation, this gives

\[ J(y; \lambda) = \prod_{i=1}^{n} \frac{d}{dy_i} \left( \left( y_i^\lambda - 1 \right) / \lambda \right) = \prod_{i=1}^{n} y_i^{\lambda-1}, \]

which is well-defined for every \( \lambda \in \mathbb{R} \), since all the \( y_i \)'s are here assumed to be positive.

### 2.2 Profile Likelihood

Now, we wish to maximize equation (1) for the triple \((\beta, \sigma^2, \lambda)\). Observe that for every choice of \( \lambda \), the likelihood is simply a standard normal likelihood function up to a constant factor depending on \( \lambda \). Therefore, we can simply use the OLS framework to find the estimates of \( \beta \) and \( \sigma^2 \), such that

\[ \hat{\beta}(\lambda) = (X^T X)^{-1} X^T \tilde{y}, \quad \text{and} \quad \hat{\sigma}^2(\lambda) = \frac{1}{n} (\tilde{y} - X\hat{\beta}(\lambda))^T (\tilde{y} - X\hat{\beta}(\lambda)), \]

where note that these estimates are a function of \( \lambda \), through the transformed values, \( \tilde{y}_i \)'s.

Secondly, we can plug in these estimates back into equation (1), in order to obtain the so-called profile likelihood of \( \lambda \). In a profile likelihood, the parameter of interest, which is here the one controlling the transformation, \( \lambda \), is written as a function of the other parameters up to a constant, such that

\[ \log L(\hat{\beta}, \hat{\sigma}^2, \lambda) := -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) + \log(J(y; \lambda)), \]

where we have here eliminated two constant terms that do not depend on \( \lambda \), composed of the following,

\[ C := -\frac{n}{2} \log(2\pi) - \frac{n\hat{\sigma}^2}{2\hat{\sigma}^2} = \frac{n}{2} (\log(2\pi) + 1). \]

Ignoring this constant term, and after some manipulations, we can express the profile likelihood of \( \lambda \) in a succinct manner,

\[ \log L(\hat{\beta}, \hat{\sigma}^2, \lambda) = -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) - \frac{n}{2} \log(J(y; \lambda)^{-2/n}) \]

\[ = -\frac{n}{2} \log \left( \hat{\sigma}^2(\lambda) J(y; \lambda)^{-2/n} \right), \]

\[ = \frac{n}{2} \log \left( \frac{1}{n} \tilde{y}^T (I - H) \tilde{y} J(y; \lambda)^{-2/n} \right), \]

where we have used the fact,

\[ \hat{\sigma}^2(\lambda) = \frac{1}{n} (\tilde{y} - X\hat{\beta})^T (\tilde{y} - X\hat{\beta}) = \frac{1}{n} \tilde{y}^T (I - H) \tilde{y}. \]

Finally, consider the definition of the geometric mean, \( \text{GM}(y) := (\prod_{i=1}^{n} y_i)^{1/n} \), and observe that

\[ J(y; \lambda)^{-1/n} = \left( \prod_{i=1}^{n} y_i^{\lambda-1} \right)^{-1/n} = \left( \prod_{i=1}^{n} y_i \right)^{(1-\lambda)/n} = \text{GM}(y)^{1-\lambda}. \]

Therefore, we obtain the simplified profile log-likelihood of \( \lambda \),

\[ \log L(\lambda) = -\frac{n}{2} \log \left( \frac{1}{n} z(\lambda)^T (I - H) z(\lambda) \right), \quad (2) \]

where

\[ z(\lambda) := \tilde{y} \times \text{GM}(y)^{1-\lambda}. \]
2.3 Box-Cox Transformations

Thus, altogether, we obtain the Box-Cox transformations, which are a normalized version of the scaled power transformations,

\[ \psi_{BC}(y; \lambda) := \psi(yi, \lambda) \times \text{GM}(y)^{1-\lambda}. \]

Thus, in full, this gives

\[ \psi_{BC}(y_i; \lambda) = \begin{cases} \frac{(y_i^\lambda - 1)}{\lambda} \times \text{GM}(y)^{1-\lambda} & \text{if } \lambda \neq 0, \\ \log(y_i) \times \text{GM}(y)^{1-\lambda} & \text{if } \lambda = 0. \end{cases} \]

The optimal value of \( \lambda \) can then be estimated using the profile likelihood in equation (2). Alternatively, the value of the transformation parameters could also be estimated using a non-linear least squares approach.

3 Multivariate Box-Cox for Predictors

Finally, we can also apply the Box-Cox framework to a set of predictor variables. Assume that we have \( p \) predictors, for instance. In this case, the transformation will be a vector-valued map of the form,

\[ \psi_M(x_i; \lambda) := \begin{bmatrix} \psi_M(x_{i1}, \lambda_1) \\ \vdots \\ \psi_M(x_{ip}, \lambda_p) \end{bmatrix}, \]

where we have used the subscript \( M \) to emphasize that this is the multivariate version of the Box-Cox transformation. Our normality assumption can be expressed as follows,

\[ \psi_M(x_i, \lambda) \sim \text{MVN}_p(\mu, V), \]

for some unknown positive definite matrix \( V \) and mean \( \mu \). Note that this is distinct from any assumptions made on the response variable. We are here solely attempting to transform our predictor variables, in order to approximate a multivariate normal distribution. The full log-likelihood for this model over \( n \) data points can be expressed as

\[ \log L(\mu, V, \lambda) := -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log(|V|) - \frac{1}{2} \sum_{i=1}^{n} (\tilde{x}_i - \mu)^T V^{-1} (\tilde{x}_i - \mu), \]

where

\[ \tilde{x}_i := \psi_M(x_i; \lambda). \]

Here, it suffices to use the cycling property of the trace, which states that for any matrix \( A \) and vector \( v \), we have \( \text{tr}(v^T Av) = \text{tr}(Avv^T) \). Hence, after defining the sample estimates of \( \mu \) and \( V \) as

\[ \hat{\mu}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i, \quad \text{and} \quad \hat{V}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - \mu)(\tilde{x}_i - \mu)^T; \]

we immediately obtain

\[ \frac{1}{2} \sum_{i=1}^{n} \text{tr}((\tilde{x}_i - \hat{\mu})^T \hat{V}^{-1} (\tilde{x}_i - \hat{\mu})) = \frac{1}{2} \sum_{i=1}^{n} \text{tr}(\hat{V}^{-1} (\tilde{x}_i - \hat{\mu})(\tilde{x}_i - \hat{\mu})^T) = \frac{1}{2} \text{tr}(\hat{V}^{-1} \sum_{i=1}^{n} (\tilde{x}_i - \hat{\mu})(\tilde{x}_i - \hat{\mu})^T) \]

\[ = \frac{n}{2} \text{tr}(\hat{V}^{-1} \hat{V}) = \frac{n}{2} \text{tr}(I_p) = \frac{np}{2}. \]

Therefore, the profile likelihood for the vector \( \lambda \) of order \( p \times 1 \), up to a constant, is then given by

\[ \log L(\lambda) = -\frac{n}{2} \log(|\hat{V}(\lambda)|). \]