Chapter 8 Transformations

Applied Linear Regression

- Often no theory tells us the correct form for the mean function, and any parametric form we use is little more than an approximation that we hope is adequate for the problem at hand.
- Replacing either the predictors, the response, or both by nonlinear transformations of them is an important tool that the analyst can use to extend the number of problems for which linear regression methodology is appropriate.
- The most frequent purpose of transformations is to achieve a mean function that is linear in the transformed scale.

Chapter 8.1 Transformation Basics

* The most frequent purpose of transformations is to achieve a mean function that is linear in the transformed scale.

* In 8.1, we focus on simple regression with one predictor.
- 8.1.1 Transforming both predictor and response, Power Transformations
- 8.1.2 Transforming Only the Predictor Variable,
- 8.1.3 Transforming the Response Only, The Box and Cox Method

Chapter 8.1 Transformations for simple regression case

- We seek a transformation so if X is the transformed predictor and Y is the transformed response, then the mean function in the transformed scale is

$$E[Y|X = x] \sim \beta_0 + \beta_1 x$$

- With only one predictor and one response, the mean function can be visualized in a scatterplot.

Example: BrainWt v.s. BodyWt

- BodyWt (in Kg) and BrainWt (in g) for 62 species of mammals.
- Both variables range over several orders of magnitude from species with BodyWt of just a few grams to over 6600 kg.
- No clear linear relationship. Transformations can help in this problem.

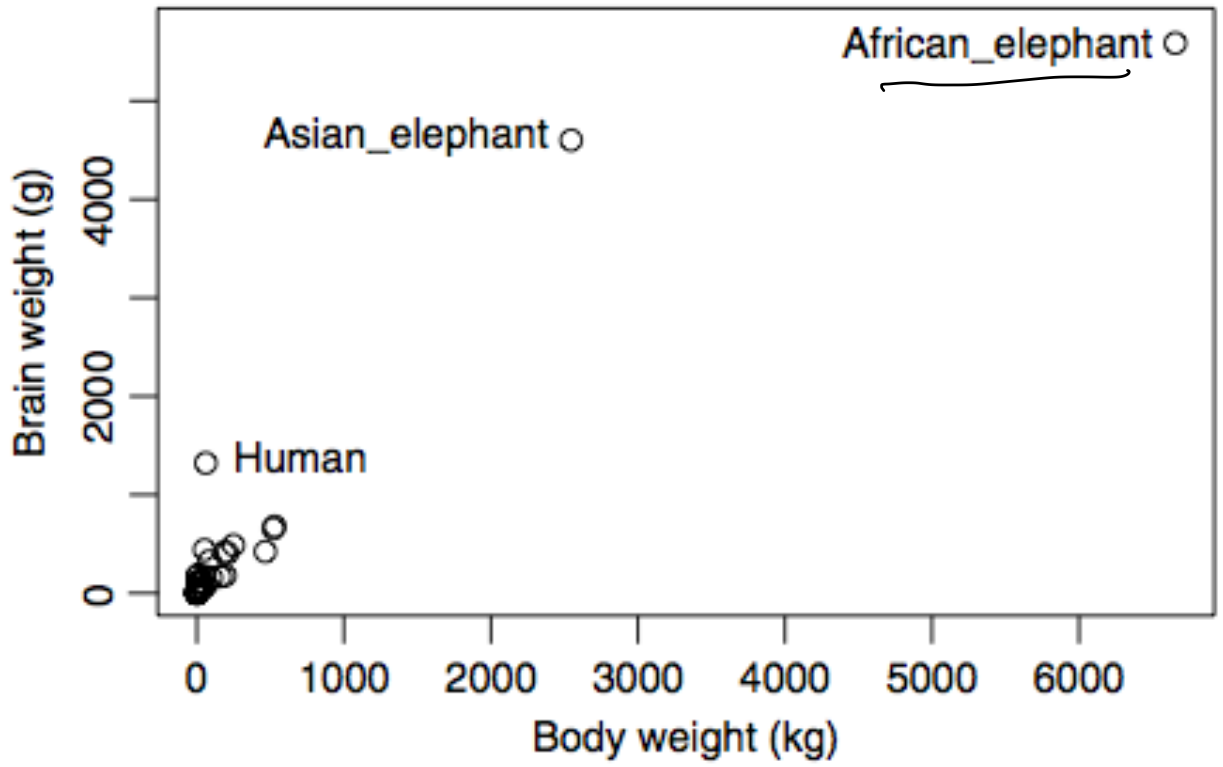


FIG. 7.1 Plot of *BrainWt* versus *BodyWt* for 62 mammal species

8.1.1 Power Transformation Method

- A transformation family is a collection of transformations that are indexed by one or a few parameters that the analyst can select. The family that is used most often is called the power family, defined for a strictly positive variable U ($U > 0$) by

power index parameter to be chosen/estimated from data.

$$\psi(U, \lambda) = U^{\lambda}.$$

- As the power parameter λ is varied, we get different transformation functions.
- E.g., $\lambda = 1/2$, $\psi(U, 1/2) = U^{1/2}$.
- When $\lambda = 0$, we let

$$\psi(U, 0) = \log_e U.$$

$\ln$

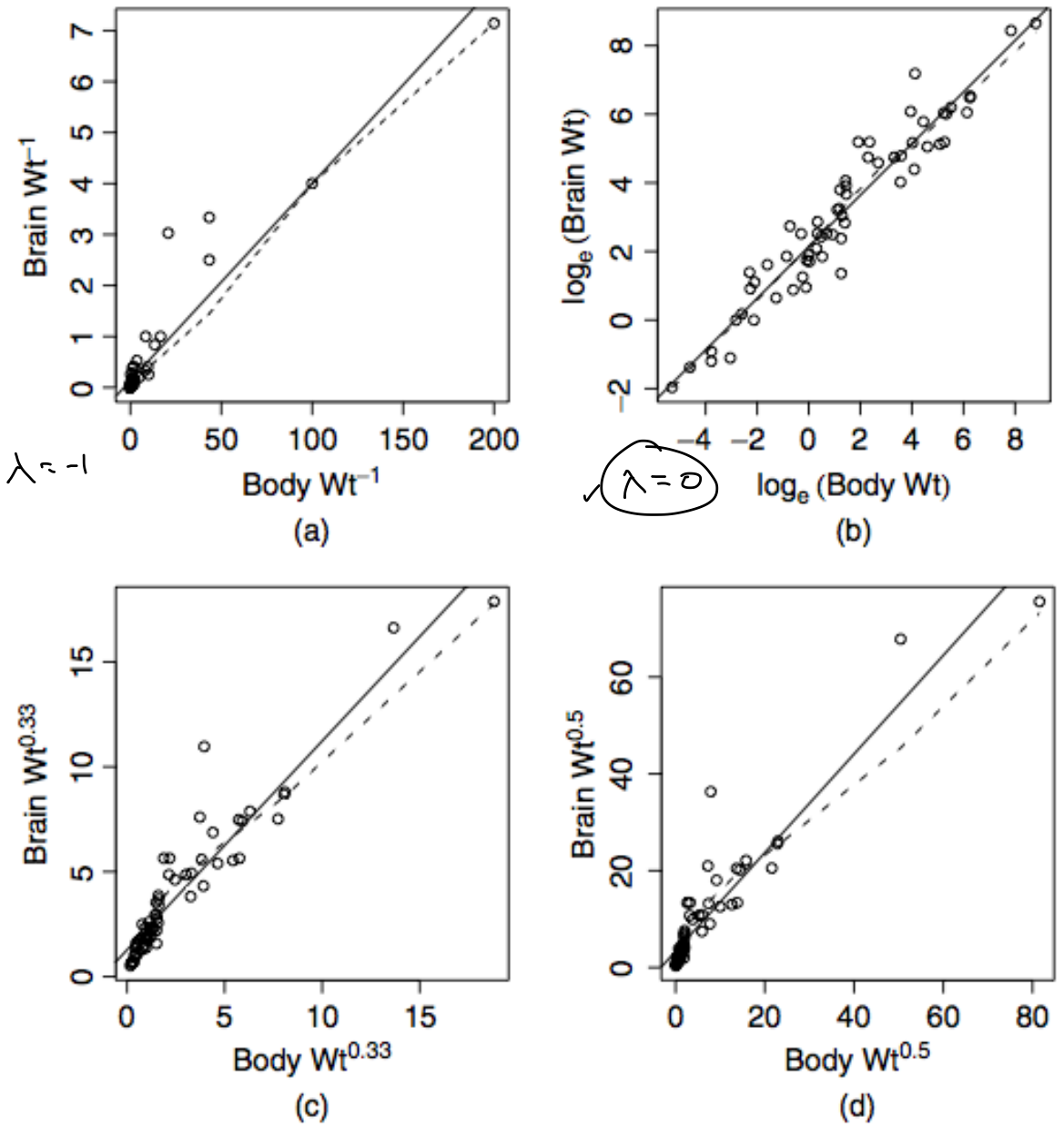
Power Transformation Method

More about power transformation

$$\psi(U, \lambda) = U^{\lambda}.$$

- The usual values of λ that are considered are in the range from -2 to 2, but values in the range from -1 to +1 are ordinarily selected.
- The value of $\lambda = 1$ corresponds to no transformation.
- The variable U must be strictly positive for these transformations to be used. (later we will discuss about transforming variables that may be zero or negative.)

Example: BrainWt v.s. BodyWt . Transformations with $\lambda = -1, 0, 1/3, 1/2$.



Power Transformation Method

We have the following two empirical rules that are often helpful in linear regression modeling:

- *The log rule: If the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful.*
- *The range rule: If the range of a variable is considerably less than one order of magnitude, then any transformation of that variable is unlikely to be helpful.*
- The log rule is satisfied for both BodyWt, with range 0.005 kg to 6654 kg, and for BrainWt, with range 0.14 g to 5712 g, so log transformations is a starting point for transformations.

Example: BrainWt v.s. BodyWt .

Simple linear regression seems to be appropriate with both variables in log scale. This corresponds to the *physical model*

$$\texttt{BrainWt} = \alpha \times \texttt{BodyWt}^{\beta_1} \times \delta \tag{8.2}$$

where δ is a *multiplicative error*. For example, if $\delta = 1.1$ for a particular species, then the BrainWt for that species is 1.1 times the expected BrainWt for all species with the same BodyWt. On taking logarithms and setting $\beta_0 = \log(\alpha)$ and $e = \log(\delta)$,

$$\log(\texttt{BrainWt}) = \beta_0 + \beta_1 \log(\texttt{BodyWt}) + e$$

$\beta_0 = \log(\alpha)$ $\quad e = \log(\delta)$

Area $= W \cdot L \Rightarrow \log \text{Area} \approx \log W + \log L$

Transforming Only the Predictor Variable

- If we want to use a family of power transformations, it is convenient to introduce the family of scaled power transformations, defined for strictly positive X by

$$\psi_S(X, \lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(X) & \text{if } \lambda = 0 \end{cases}$$

Example:
if $Cov(X, Y) > 0$
$\lambda = -1$
$Cov(X^{-1}, Y) < 0$
$Cov\left(\frac{X^{-1}-1}{(-1)}, Y\right) > 0$

- Scaled power transformations $\psi_S()$ preserve the direction of association. However, for basic power transformations $\psi()$, the direction of association changes when $\lambda < 0$.
- When $\lambda \to 0$, we let

$$\psi_S(U, 0) \to \log U.$$

$\lim_{\lambda \to 0} \frac{X^\lambda - 1}{\lambda} \to \log X$.

However, $\lim_{\lambda \to 0} \frac{X^\lambda \cdot \log X}{\lambda} \to \log X$

How to get λ?

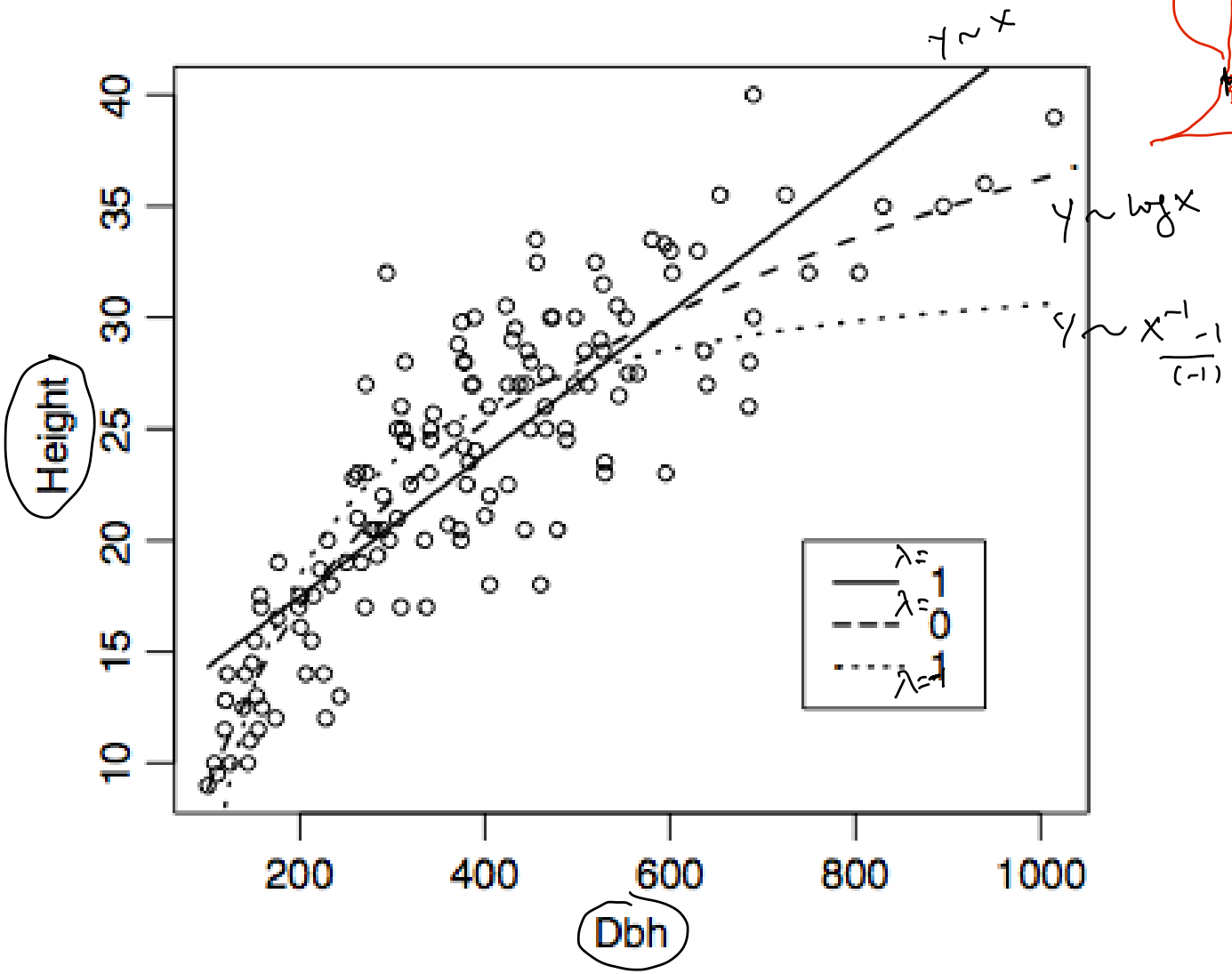
If transforming only the predictor and using a choice from the power family, we begin with the mean function

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 \psi_S(X, \lambda) \tag{8.4}$$

If we know λ, we can fit (8.4) via OLS and get the residual sum of squares, $\mathtt{RSS}(\lambda)$. An estimate $\hat{\lambda}$ of λ is the value of λ that minimizes $\mathtt{RSS}(\lambda)$. We do not need to know λ very precisely, and selecting λ to minimize $\mathtt{RSS}(\lambda)$ from $\lambda \in \{-1, -1/2, 0, 1/3, 1/2, 1\}$ is usually adequate.

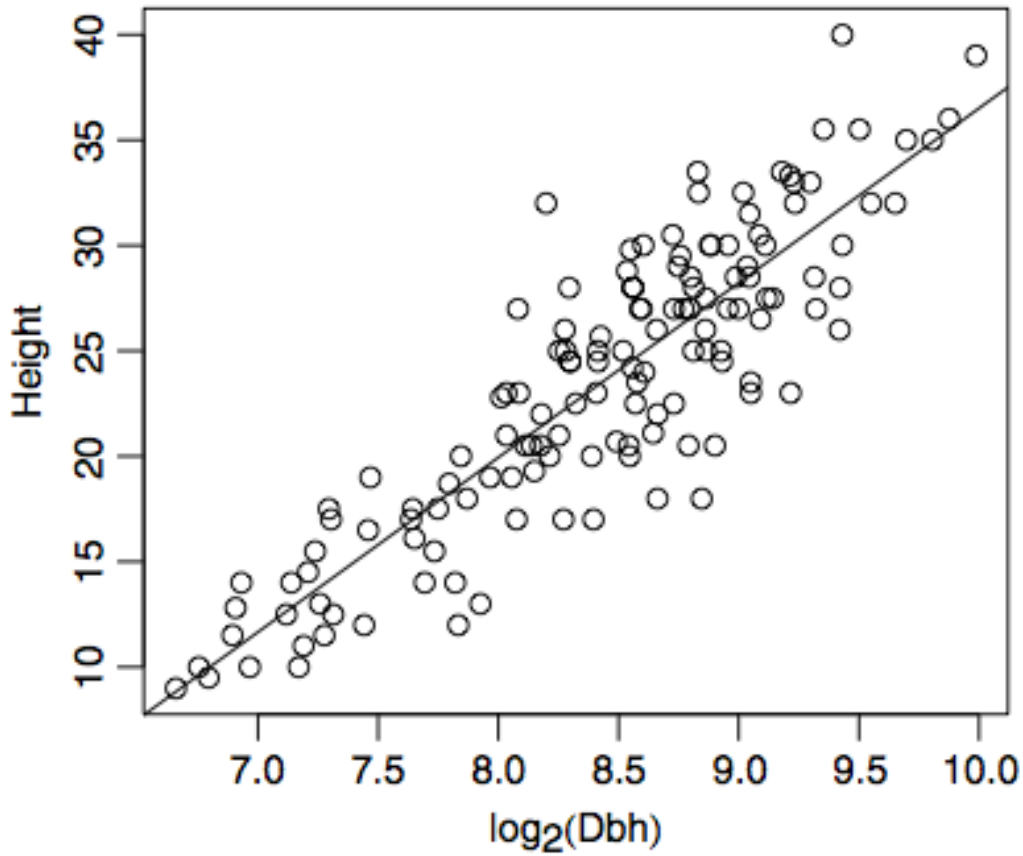
The invTranPlot function provides a graphical method to transform a single predictor for linearity.

An example of transforming only the predictor: tree Height in decimeters v.s. Dbh, the diameter of the tree in mm at 137 cm above the ground. $\lambda = 1, 0, -1$.



The choice of $\lambda = 0$ seems the best. The choice of $\lambda = 1$ does not match the data for large and small trees, while $\lambda = -1$ is too curved to match the data for larger trees.
For $\lambda =$ -1, 0, and 1, $RSS =$ 197352, 152232, and 193740, respectively.

Scatterplot of *Height* and $log_2(Dbh)$



8.1.3 The Box-Cox Method

- Box and Cox (1964) provided another general method for selecting transformations of the response that is applicable both in simple and multiple regression.
- For strictly positive Y

$$\psi_M(Y, \lambda_y) = \psi_S(Y, \lambda_y) \times \text{gm}(Y)^{1-\lambda_y}$$
$$= \begin{cases} \text{gm}(Y)^{1-\lambda_y} \times (Y^{\lambda_y} - 1)/\lambda_y & \text{if } \lambda_y \neq 0 \\ \text{gm}(Y) \times \log(Y) & \text{if } \lambda_y = 0 \end{cases} \quad (8.5)$$

where $\text{gm}(Y)$ is the *geometric mean* of the untransformed variable: if the values of Y are $y_1, \ldots, y_n$, the geometric mean of Y is $\text{gm}(Y) = \exp(\sum \log(y_i)/n)$.

Suppose that the mean function

$$\text{E}(\psi_M(Y, \lambda_y)|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \quad (8.6)$$

holds for some λ_y. How to estimate λ_y?

8.1.3 The Box-Cox Method

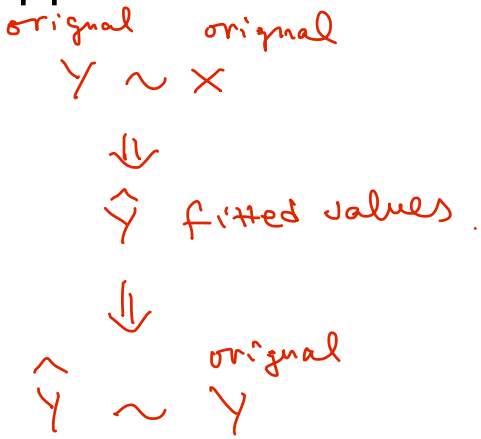
- If λ_y were known, we could fit the mean function (8.6) using OLS. Write the residual sum of squares from this regression as $RSS(\lambda_y)$.
- We estimate λ_y to be the value of the transformation parameter that minimizes $RSS(\lambda_y)$. From a practical point of view, we can again select λ_y from $\lambda_y \in \{-1, -1/2, 0, 1/3, 1/2, 1\}$.

Transforming Only the Response Variable

- A transformation of the response only can be selected using an inverse fitted value plot, in which we put the fitted values from the regression of Y on X on the vertical axis and the response on the horizontal axis.
- Scaled power transformations $\psi_S(Y, \lambda)$ then can be applied to this inverse problem.
- We use mean function

$$E[\hat{y}|Y] = \alpha_0 + \alpha_1 \psi_S(Y, \lambda)$$

and estimate λ by minimizing RSS.



8.2 A General Transformation Procedure

* The most frequent purpose of transformations is to achieve a mean function that is linear in the transformed scale.

* Next we focus on regression with more than one predictors.

Example: Highway Accident Data

Table 8.1 The Highway Accident Data

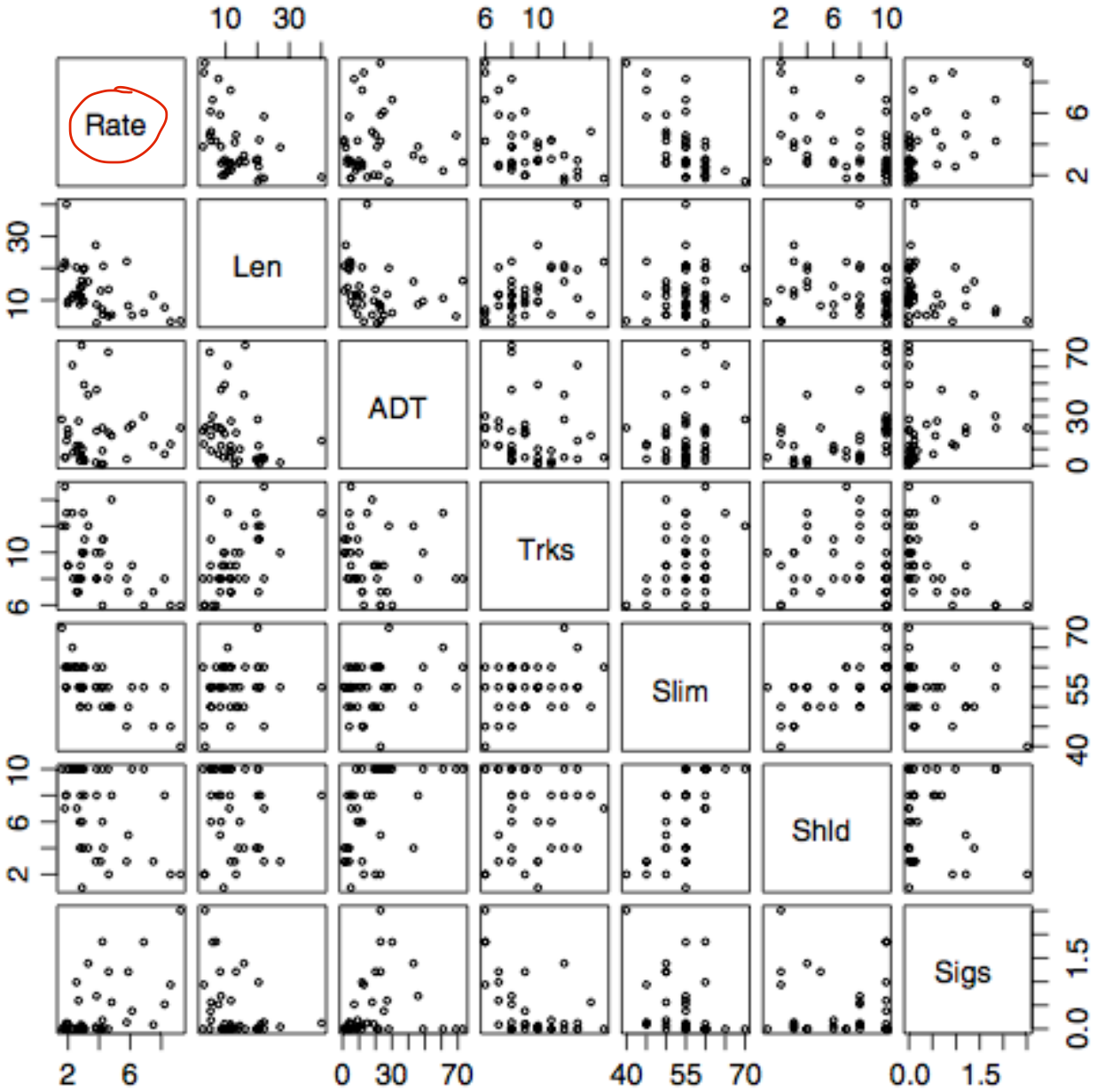
Variable	Description
`rate`	1973 accident rate per million vehicle miles
`len`	Length of the segment in miles
`adt`	Estimated average daily traffic count in thousands
`trucks`	Truck volume as a percentage of the total volume
`slim`	1973 speed limit
`shld`	Shoulder width in feet of outer shoulder on the roadway
`sigs`	Number of signalized interchanges per mile in the segment

response <- rate

predictors: len, adt, trucks, slim, shld, sigs

- We have no particular reason to believe that Rate will be a linear function of the predictors, or any theoretical reason to prefer any particular form for the mean function.
- An important first step in this analysis is to examine the scatterplot matrix of all the predictors and the response.

Example: Highway Accident Data



Example: Highway Accident Data

Some observations from Scattermatrix

- Each of the predictors seems to be at least modestly associated with Rate.
- Many of the predictors are also related to each other.
- The variable Sigs, the number of traffic lights per mile, is zero for freewaytype road segments but can be well over 2 for other segments. Transformations may help with this variable, but since it has non positive values, we use

$$Sigs1 = \frac{Sigs \times Len + 1}{Len}.$$

if sigs = 0

sigs1 = $\frac{1}{Len}$

- ADT and Len have a large range, and logarithms are likely to be appropriate for them.
- Slim varies only from 40 mph to 70 mph, with most values in the range 50 to 60. Transformations are unlikely to be much use here.

Step 1: transform X

The powerTransform function in the **car** package is the central tool for helping to choose predictor transformations.

```
> library(car)
> summary(ans <- powerTransform(cbind(Len, ADT, Trks, Shld, Sigs1) ~ 1, data=highway))
```

Table 8.2 Power Transformations to Normality for the Highway Data

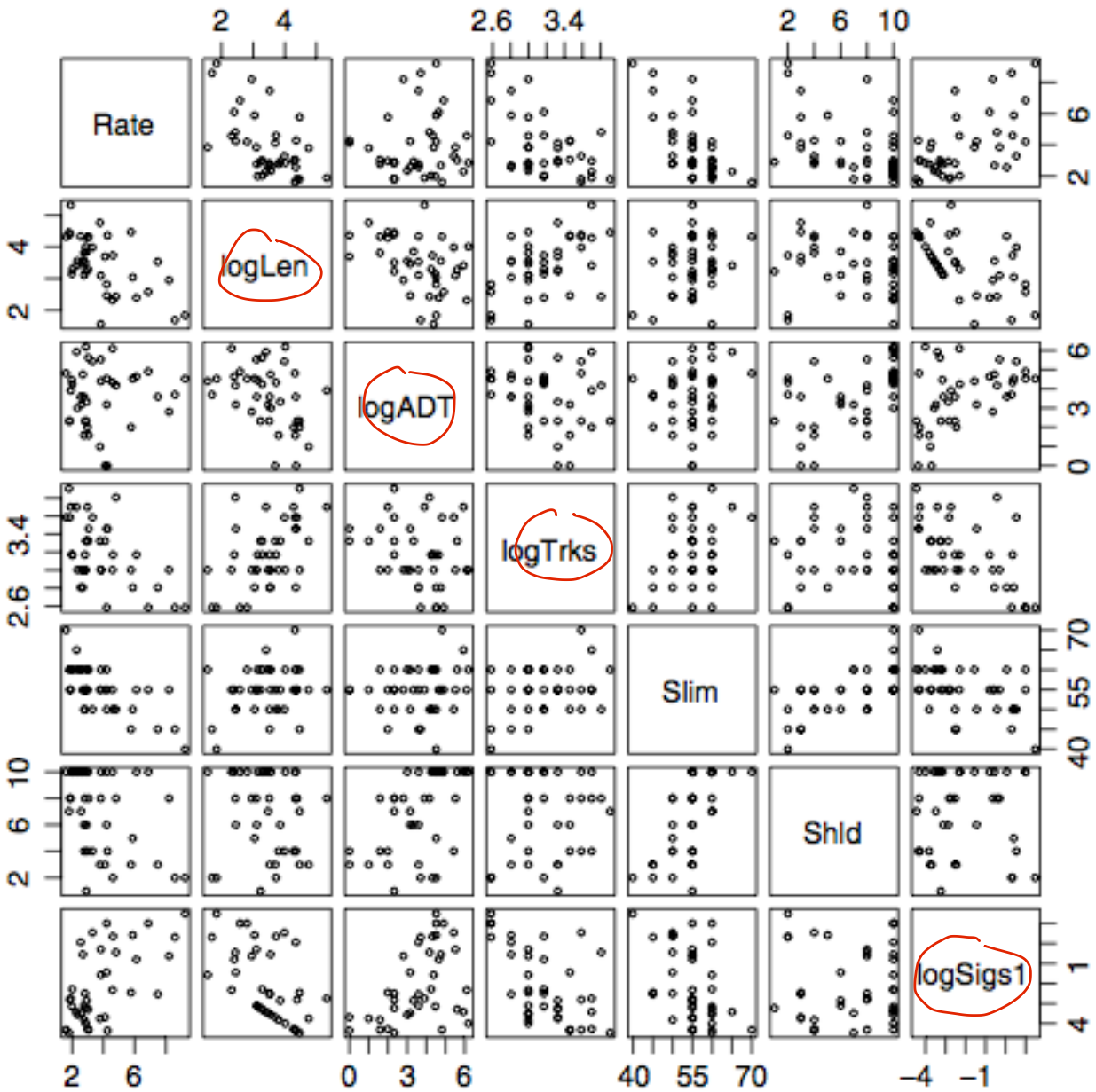
(a) Estimated Powers

	Est.Power	Std.Err.	Wald Conf. Int. Lower	Upper
len	0.144	0.213	–0.273	0.561
adt	0.051	0.121	–0.185	0.287
trks	–0.703	0.618	–1.913	0.508
shld	1.346	0.363	0.634	2.057
sigs1	–0.241	0.150	–0.534	0.052

(b) Test Statistics

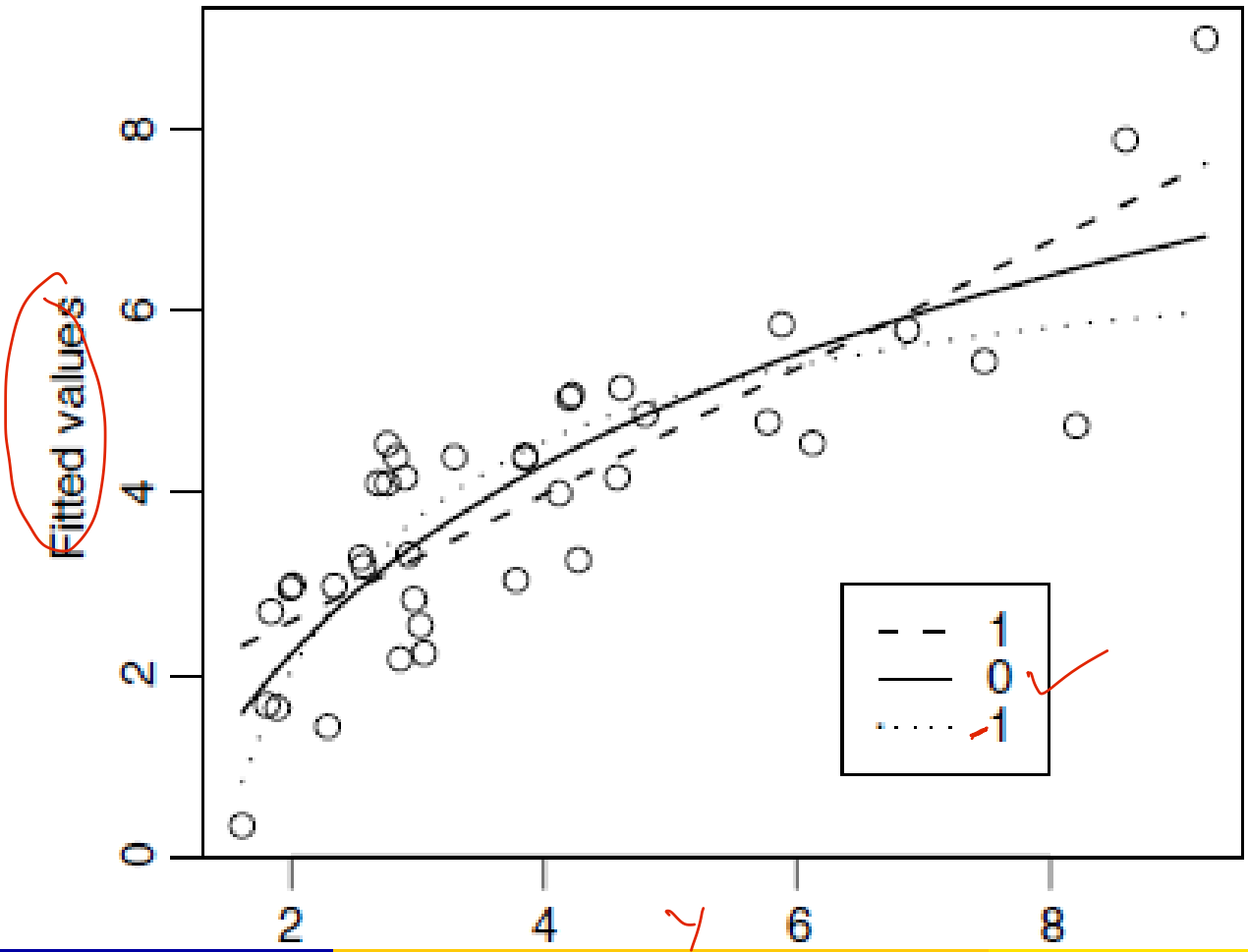
	LRT	df	p-Value
LR test, lambda = (0 0 0 0 0)	23.32	5	0.00
LR test, lambda = (1 1 1 1 1)	132.86	5	0.00
LR test, lambda = (0 0 0 1 0)	6.09	5	0.30

Highway Accident Data; Step 1



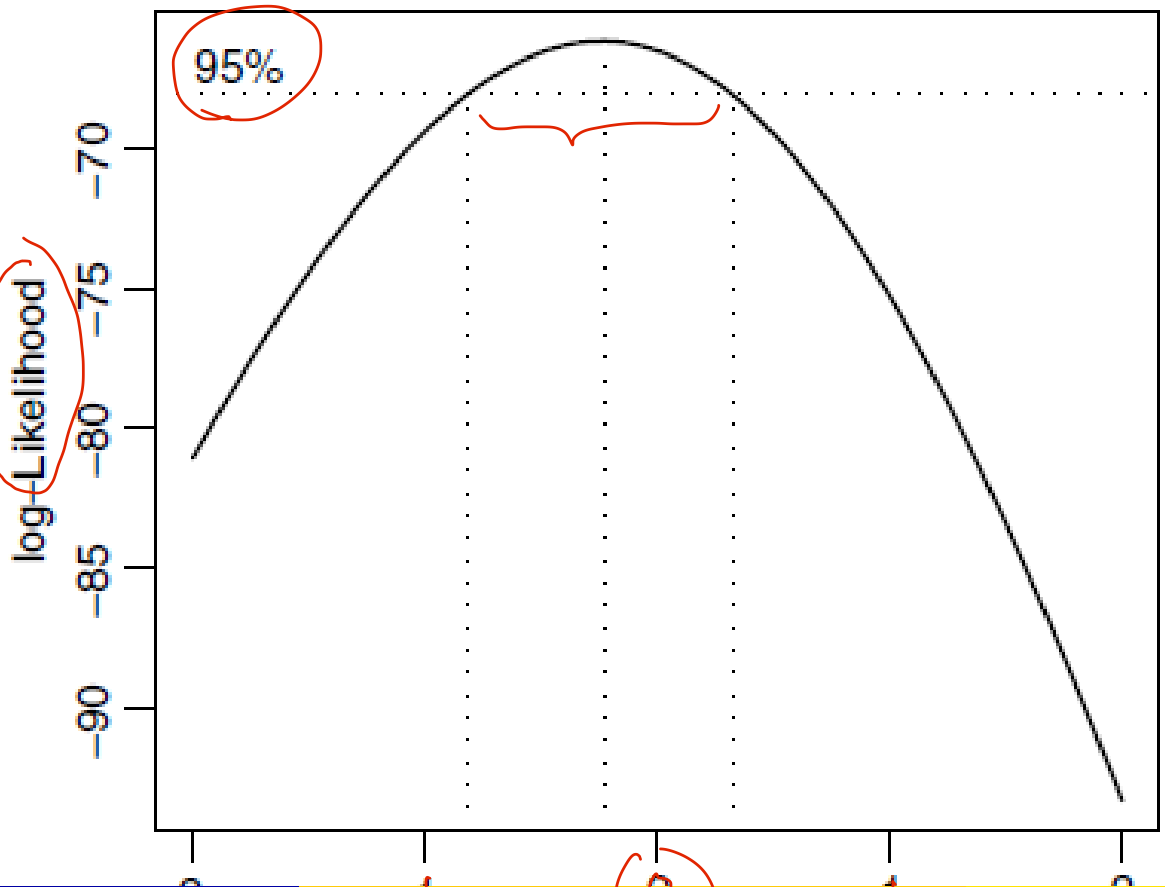
Step 2: transform Y

Draw the inverse response plot of fitted $\hat{Y}$ v.s. Y.
inverseResponsePlot in car package – invResPlot(model)



Highway Accident Data; Step 2: transform Y

Use Box-Cox Method. Choose λ_y minimizing $RSS(\lambda_y)$ or maximizing likelihood $-(n/2)\log(RSS(\lambda_y)/n)$.
(The boxcox function in the MASS package -boxCox(model))



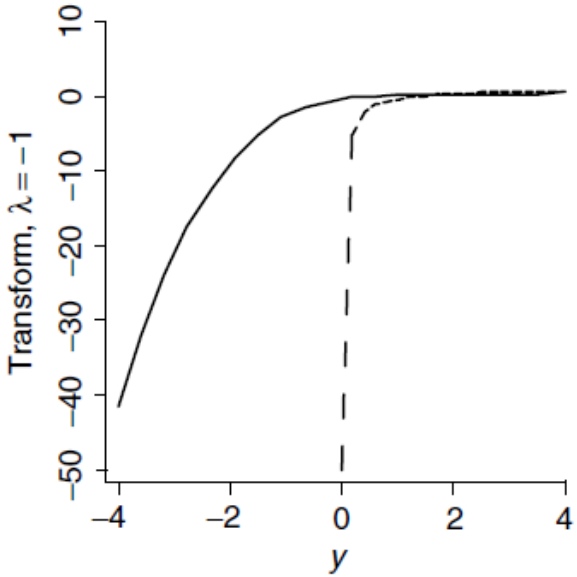
8.4 Transformations of Non-positive variables

- A natural choice is to use $(U+\gamma)^{\lambda}$ for some positive constant γ s.t. $U+\gamma>0$.
- Yeo and Johnson (2000) proposed another family of transformations

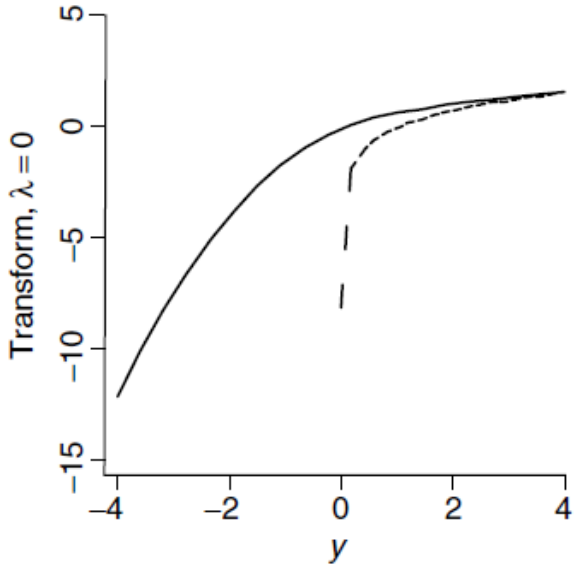
$$\psi_{YJ}(U,\lambda)=\begin{cases}\psi_M(U+1,\lambda) & \text{if } U\geq 0\\ \psi_M(-U+1,2-\lambda) & \text{if } U<0\end{cases}$$

If $U\geq 0$, this is same as the Box-Cox transform of $(U+1)$. If $U<0$ this is the Box-Cox transform of $(-U+1)$ with power $(2-\lambda)$.

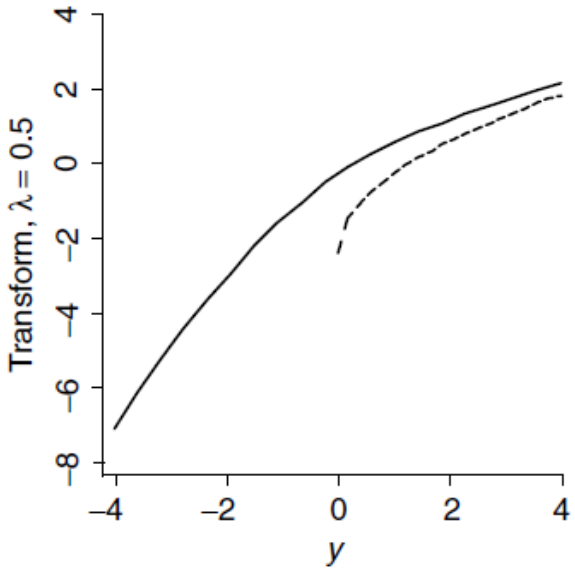
Transformations of Non-positive variables



(a)



(b)



(c)