Chapter 7

Transformation

7.1. Transformation

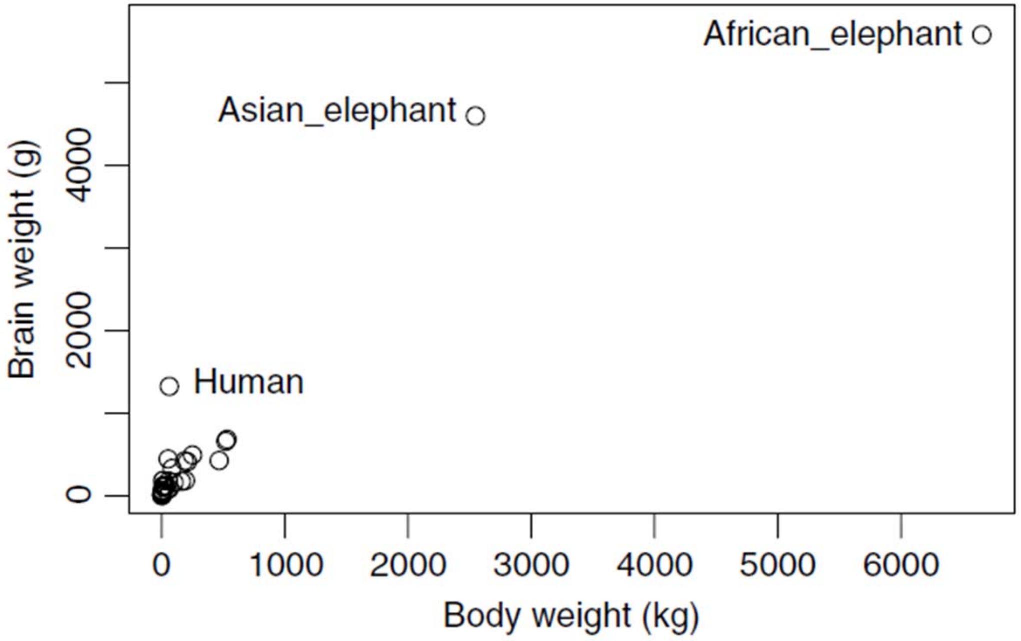
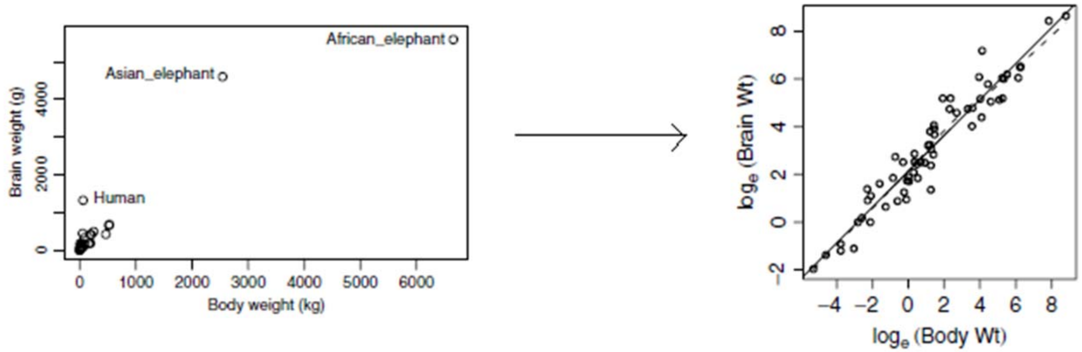


FIG. 7.1 Plot of *BrainWt* versus *BodyWt* for 62 mammal species.

- Is linear regression appropriate?

7.1. Transformation



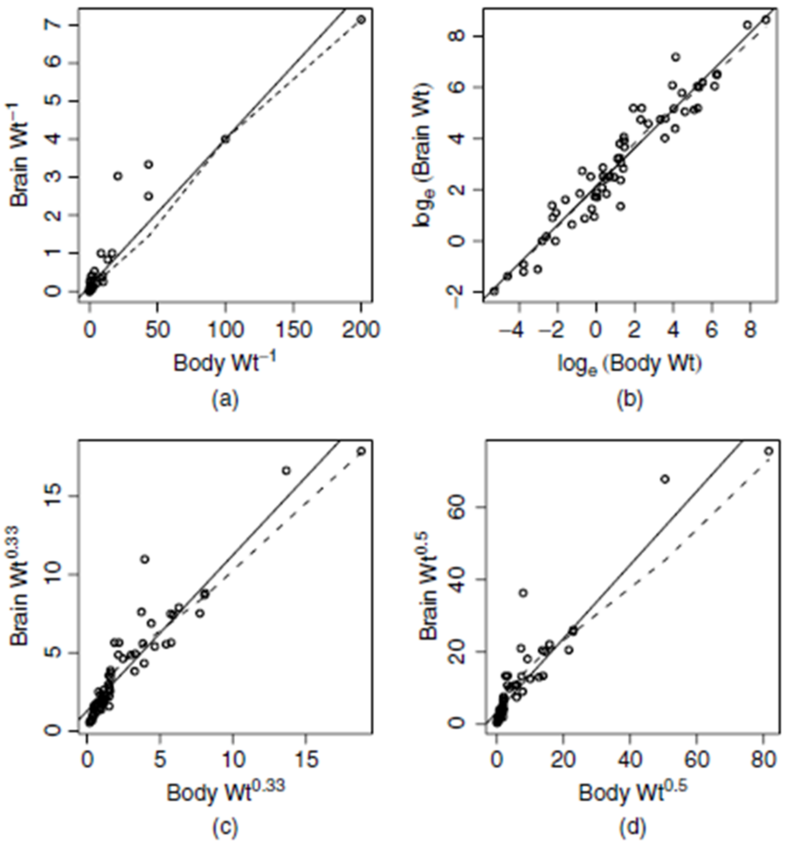
- The assumption of linear relationship does not always hold
- We can transform
  - The predictor
  - The response
  - Both

  to achieve the linear relationship

Power transformation

- Power transformation

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



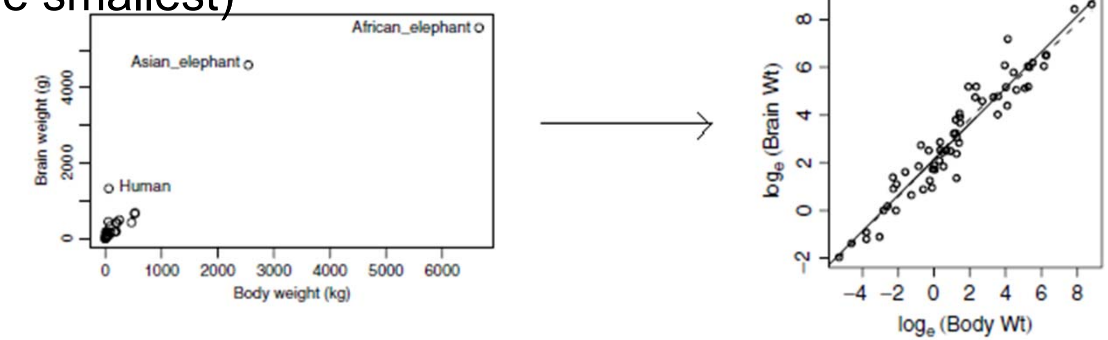
- Want a linear relationship

$$\psi(BrainWt,\lambda) = \alpha + \beta\psi(BodyWt,\lambda) + e$$

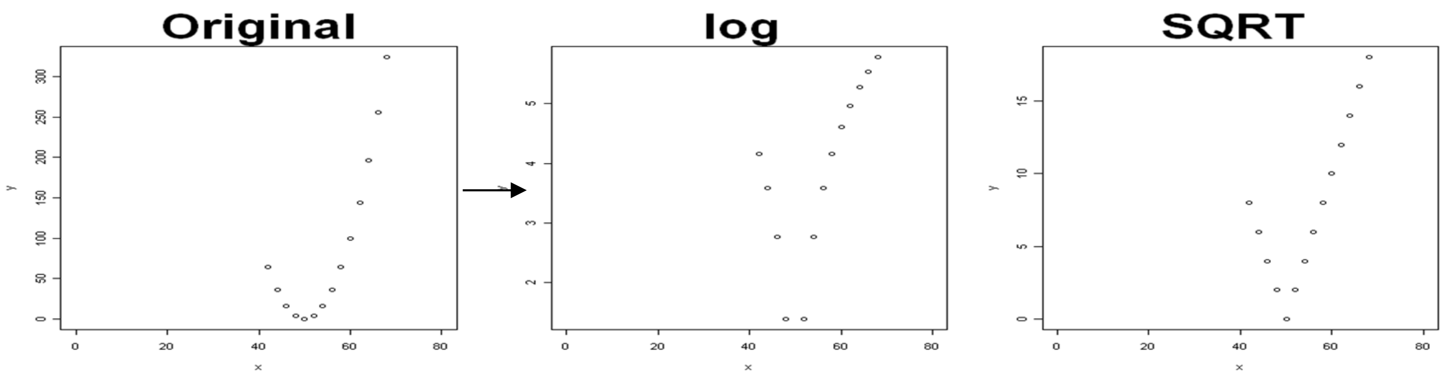
- λ=
  - a) -1
  - b) 0 (i.e. log U)
  - c) 0.33
  - d) 0.5
- Which λ will you choose?

Practical suggestions

- Log rule: log transform is useful when
  - Observations are positive
  - Range of variable is huge (i.e. the biggest observations is a much bigger than the smallest)



- Range rule: No transformation is useful if
  - Range of variable is too small



Interpretation

$$\psi(BrainWt, \lambda) = \alpha + \beta\psi(BodyWt, \lambda) + e$$

- λ > 0

$$(BrainWt)^{\lambda} = \alpha + \beta(BodyWt)^{\lambda} + e$$

  - Artificial , usually has no physical meaning

- λ = 0 : log transformation
  - Corresponding to a physical model – allometric model

$$\log(BrainWt) = \alpha + \beta \log(BodyWt) + e$$

$$\Rightarrow BrainWt = \alpha(BodyWt)^{\beta}\delta$$

Multiplicative error

Improving Power transformation

- Power transformation

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

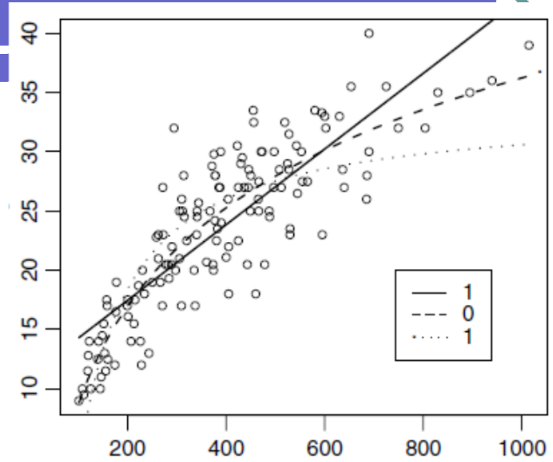
- Scaled power transformation

$$\psi_s(X,\lambda)=\begin{cases}\dfrac{X^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \log(X) & \lambda=0\end{cases}$$

- Advantage
  - Continuous function of λ : $\lim_{\lambda\to 0}\dfrac{X^{\lambda}-1}{\lambda}=\log(X)$
  - Preserve the direction of association
    - True model : $E(Y\mid X)=\beta X^{-\frac{1}{2}}$ (negative assocation b/w Y and X)
    - Power transform: $E(Y\mid X)=\beta\psi(X,-1/2)$ (positive assocation b/w Y and ψ)
    - Scaled power transform:

$$E(Y\mid X)=\beta-\frac{1}{2}\beta\psi_s(X,-\frac{1}{2})$$ (negative assocation b/w y and $\psi_s(X,-\frac{1}{2})$)

Procedures to look for transformation

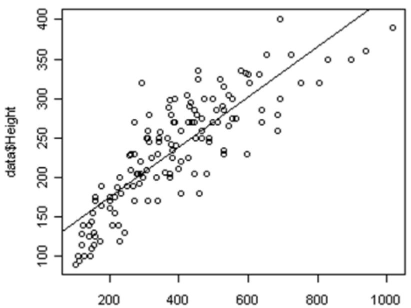


- Method 1: Draw many fitted curves

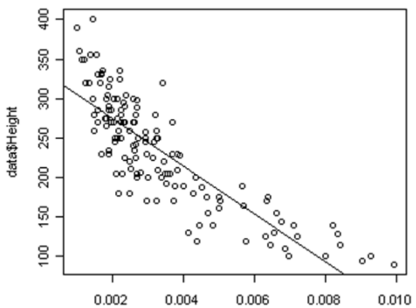
  i.e. plot $(x, \hat{y})$ for various x, where

  $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \psi(x, \lambda), \quad \lambda = -1, 0, 1 \ldots$$

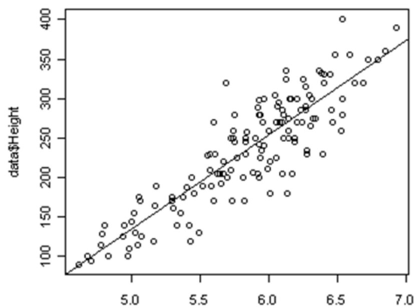
- Method 2: Draw many scatter plots



Y vs X



Y vs 1/X



Y vs log(X)

- Method 3: plot λ against RSS of fitting y against ψ(X, λ)

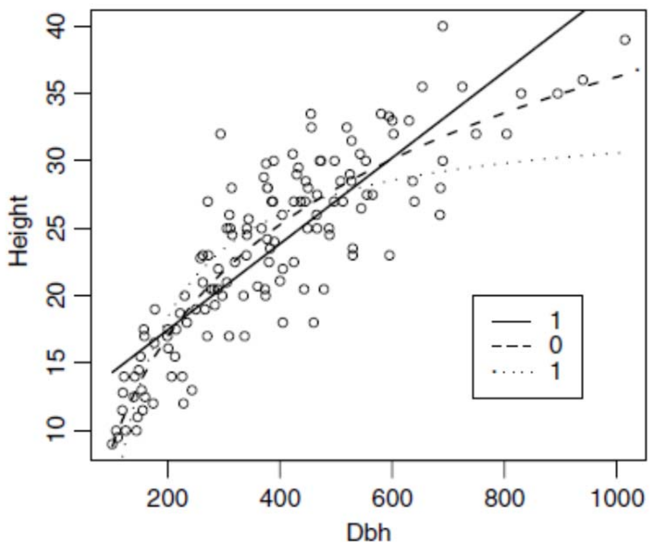
  then find the λ that minimizes RSS.

  Or choose λ in the set (-1,-1/2,0,1,2)

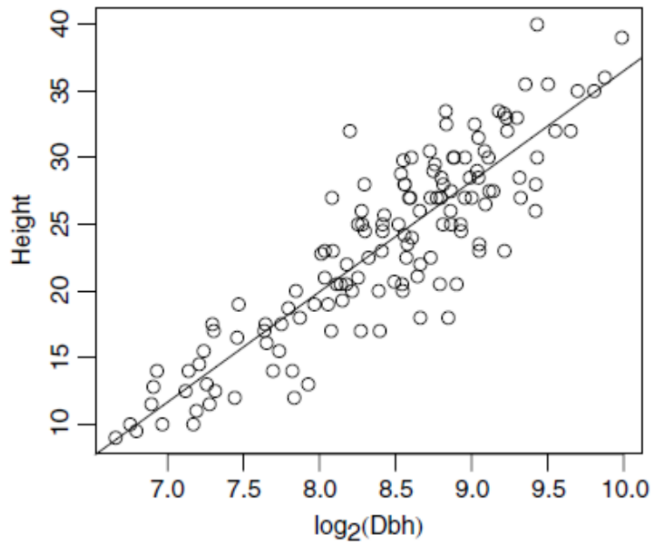
Example

- Y=Height of tree
- X=diameter of tree

M1: Draw many curves



M2: The best scatterplots



M3: Minimize RSS:

- **RSS(λ=0)=132.2, RSS(λ=1)=144.5, RSS(λ=-1)=254.8….**

Conclusion: Height = $\beta_o + \beta_1$ log(Diameter) + e

Methods for multiple regression

- Three approaches
  - Inverse fitted value plot
    - Plot $\hat{Y}$ against Y
    - Find transformation for Y that matches the above pattern
  - Box Cox transformation
    - A modification of scaled power transformation, but applied to Y.
  - Modified power transform for each predictor

Inverse fitted value plot

1. Fit a linear regression between Y and X, get the fitted value $\hat{Y} = X\hat{\beta}$
2. Plot $\hat{Y}$ (y-axis) against Y (x-axis)
3. Fix a λ, fit $\hat{Y}$ against $\psi_s(Y,\lambda)$ and obtain

$$\hat{Y}_\lambda = \hat{\beta}_0 + \hat{\beta}_1 \psi_s(Y,\lambda)$$

4. Draw the fitted curve $(Y, \hat{Y}_\lambda)$ on the graph, see if it matches the pattern in 2).
   - Match → $\hat{\beta}_0 + \hat{\beta}_1 \psi_s(Y,\lambda) = \hat{Y}_\lambda \approx \hat{Y} = X\hat{\beta}$
5. Repeat 3)-4) to search for the best λ, say λ*

$\psi_s(Y,\lambda^*)$ and X are linearly related $\Rightarrow$ Regress $\psi_s(Y,\lambda^*)$ against X

Example of Inverse fitted value

- **Read data**
  - highway.data=read.table("C:/highway.txt",header=T) #Or library(alr3); highway.data=highway
- **Step 1: Multiple regression**

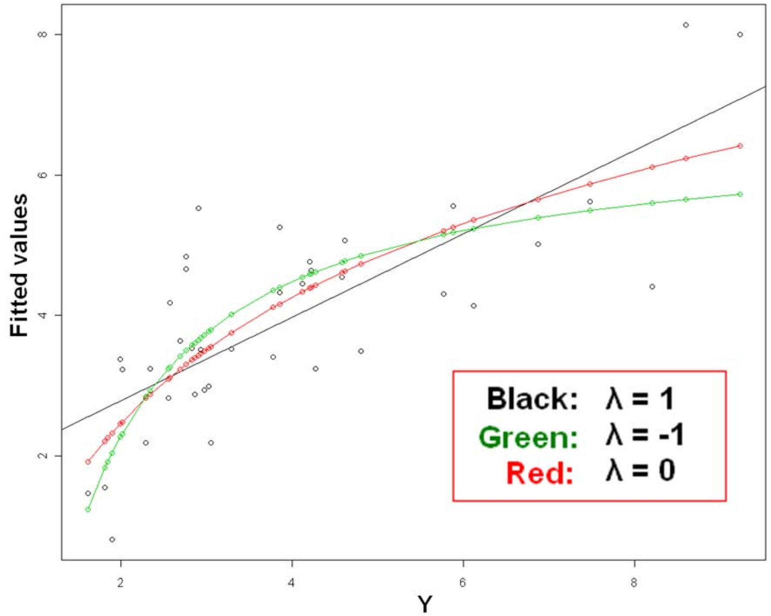
  fit=lm(Rate~log(ADT)+log(Trks)+Shld+log(Len),data=highway.data)
- **Step 2: Plot fitted values against Y**

  ```
  y.hat=fit$fitted.values
  y=highway.data$Rate
  plot(y,y.hat)
  abline(lm(y.hat~y))
  ```
- **Step 3+4: Regression: Fitted value against transformed Y, and plot the Newly fitted values**

  ```
  Psi.0=log(y)
  fit1=lm(y.hat~Psi.0)
  points(y,fit1$fitted.values,col=2)
  ```
- **Trial 2: Step 3+4:**

  ```
  Psi.minus1=-(1/y-1)
  fit2=lm(y.hat~Psi.minus1)
  points(y,fit2$fitted.values,col=3)
  ```
- **More R techniques: Sort y to draw the line.**

  ```
  order.y=order(y)
  ordered.y=y[order.y]
  ordered.fit1=fit1$fitted.values[order.y]
  ordered.fit2=fit2$fitted.values[order.y]
  lines(ordered.y,ordered.fit1,type="l",col=2)
  lines(ordered.y,ordered.fit2,type="l",col=3)
  ```



- In this case λ=0 seems to be the best.

Box-Cox transformation

1. Modified power family

$$\psi_M(Y,\lambda)=\psi_S(Y,\lambda)(\sqrt[n]{y_1y_2\ldots y_n})^{1-\lambda}$$

$$=\begin{cases}(\sqrt[n]{y_1y_2\ldots y_n})^{1-\lambda}\dfrac{Y^\lambda-1}{\lambda} & \text{if } \lambda\neq 0\\ \sqrt[n]{y_1y_2\ldots y_n}\log(Y) & \text{if } \lambda=0\end{cases}$$

2. Advantage: Unit of $\psi_M(Y,\lambda)$ is the same as Y for all λ
3. Model Assumption:

$$E(\psi_M(Y,\lambda)\,|\,X=x)=\beta' x \qquad (*)$$

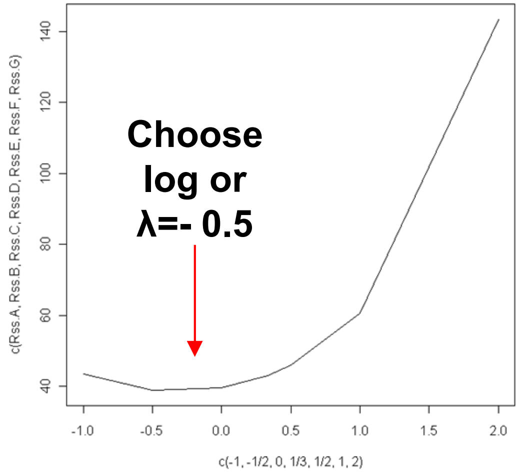
4. How to choose λ?
   - Fix a λ, fit model (*) for and obtain RSS(λ)
   - Try various λ and find the one which minimizes RSS(λ)

Example of Box-Cox transformation

$$E(\psi_M(Y,\lambda)\,|\,X = x) = \beta' x \qquad (*)$$

- Modified power family

$$\psi_M(Y,\lambda) = \begin{cases} (\sqrt[n]{y_1 y_2 \ldots y_n})^{1-\lambda} \dfrac{Y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \sqrt[n]{y_1 y_2 \ldots y_n} \log(Y) & \text{if } \lambda = 0 \end{cases}$$




```
highway.data=read.table("C:/highway.txt",header=T)
y=highway.data$Rate
n=length(y)
gm=prod(y)^{1/n}

#A) lambda=-1
Transform.A=-gm^2*(1/y-1)
fit.A=lm(Transform.A~log(ADT)+log(Trks)+Shld+log(Len),data=highway.data)
Rss.A=sum(fit.A$residuals^2)
… … …
… … …
#G) lambda=2
Transform.G=1/2/gm*(y^2-1)
fit.G=lm(Transform.G~log(ADT)+log(Trks)+Shld+log(Len),data=highway.data)
Rss.G=sum(fit.G$residuals^2)
```

- plot(c(-1,-1/2,0,1/3,1/2,1,2),c(Rss.A,Rss.B,Rss.C,Rss.D,Rss.E,Rss.F,Rss.G),type="l")

Example of Box-Cox transformation

```
# Read data
highway.data=read.table("C:/highway.txt",header=T)
y=highway.data$Rate
n=length(y)
gm=prod(y)^{1/n}

#A) lambda=-1
Transform.A=-gm^2*(1/y-1)
fit.A=lm(Transform.A~log(ADT)+log(Trks)+Shld+log(Le
        n),data=highway.data)
Rss.A=sum(fit.A$residuals^2)

#B) lambda=-1/2
Transform.B=-2*(gm^(3/2))*(y^(-1/2)-1)
fit.B=lm(Transform.B~log(ADT)+log(Trks)+Shld+log(Le
        n),data=highway.data)
Rss.B=sum(fit.B$residuals^2)

#C) lambda=0
Transform.C=gm*log(y)
fit.C=lm(Transform.C~log(ADT)+log(Trks)+Shld+log(L
        en),data=highway.data)
Rss.C=sum(fit.C$residuals^2)

#D) lambda=1/3
Transform.D=3*(gm^(2/3))*(y^(1/3)-1)
fit.D=lm(Transform.D~log(ADT)+log(Trks)+Shld+log(Len),data=
        highway.data)
Rss.D=sum(fit.D$residuals^2)

#E) lambda=1/2
Transform.E=2*(gm^(1/2))*(sqrt(y)-1)
fit.E=lm(Transform.E~log(ADT)+log(Trks)+Shld+log(Len),data=
        highway.data)
Rss.E=sum(fit.E$residuals^2)

#F) lambda=1
Transform.F=y
fit.F=lm(Transform.F~log(ADT)+log(Trks)+Shld+log(Len),data=h
        ighway.data)
Rss.F=sum(fit.F$residuals^2)

#G) lambda=2
Transform.G=1/2/gm*(y^2-1)
fit.G=lm(Transform.G~log(ADT)+log(Trks)+Shld+log(Len),data=
        highway.data)
Rss.G=sum(fit.G$residuals^2)

plot(c(-1,-
        1/2,0,1/3,1/2,1,2),c(Rss.A,Rss.B,Rss.C,Rss.D,Rss.E,Rs
        s.F,Rss.G),type="l")
```

Modified power transformation for all predictors

- Modified power family

$$\psi_M(Y,\lambda) = \psi_S(Y,\lambda)(\sqrt[n]{y_1 y_2 \dots y_n})^{1-\lambda}$$

$$= \begin{cases} (\sqrt[n]{y_1 y_2 \dots y_n})^{1-\lambda} \dfrac{Y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \sqrt[n]{y_1 y_2 \dots y_n} \log(Y) & \text{if } \lambda = 0 \end{cases}$$

- Transform predictors so that each pair of variables in the scatterplot matrix has a linear relationship.

$$(X_1, X_2 \dots, X_p) \rightarrow \left(\psi_M(X_1,\lambda_1), \psi_M(X_2,\lambda_2), \dots, \psi_M(X_p,\lambda_p)\right)$$

Modified power transformation for all predictors

- Transformation with modified power family

$$(X_1, X_2 ..., X_p) \rightarrow \left(\psi_M(X_1, \lambda_1), \psi_M(X_2, \lambda_2), ..., \psi_M(X_p, \lambda_p)\right)$$

- Not an easy task.
- Only use it if other methods do not work well

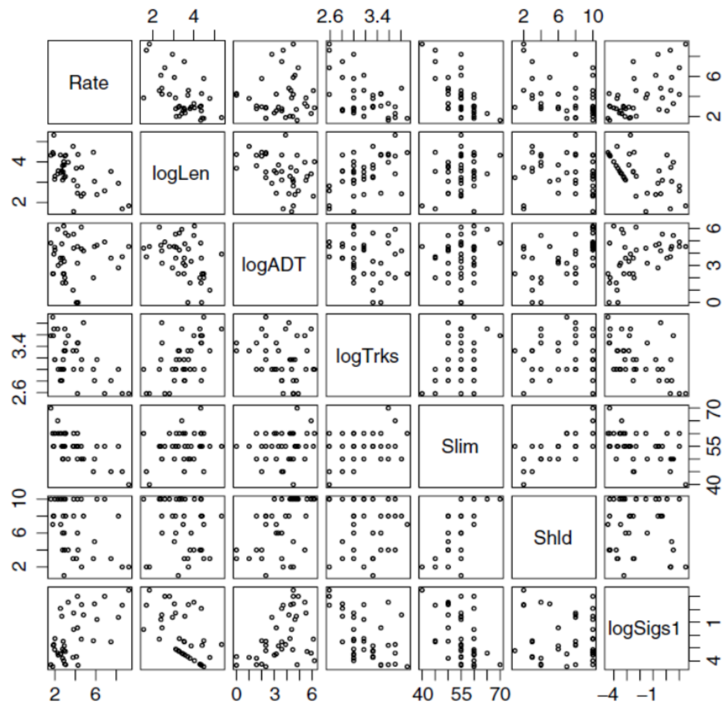


FIG. 7.6 Transformed predictors for the highway data.

Transformation of non-positive variables

- Problem of non-positive variables
  - e.g. λ=2, $\psi_S(x,2)=\psi_S(-x,2)=\frac{x^2-1}{2}$ we can't distinguish between x and –x.
  - log(x) is undefined if x<0.
- Solutions
  - Find a sufficiently large γ and transform U to

$$(U+\gamma)^{\lambda}$$

  - Yeo-Johnson transformation

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

Final Remarks

- No need to transform factors
  - e.g.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 F, \qquad F = \begin{cases} 1 & \text{group 1} \\ 0 & \text{group 2} \end{cases}$$

  we look at β_2 to see the mean different between the groups. Transforming the dummy doesn't help.

- There is no 'correct' way of transformation, once you come up with transformation

$$\left(\psi(X_1, \lambda_1), \ldots, \psi(X_p, \lambda_p), \psi(Y, \lambda_0)\right)$$

which looks roughly linear in the scatterplot matrix, then it is ok to fit.

$$\psi(Y, \lambda_0) = \beta_0 + \beta_1 \psi(X_1, \lambda_1) + \ldots + \beta_p \psi(X_p, \lambda_p)$$