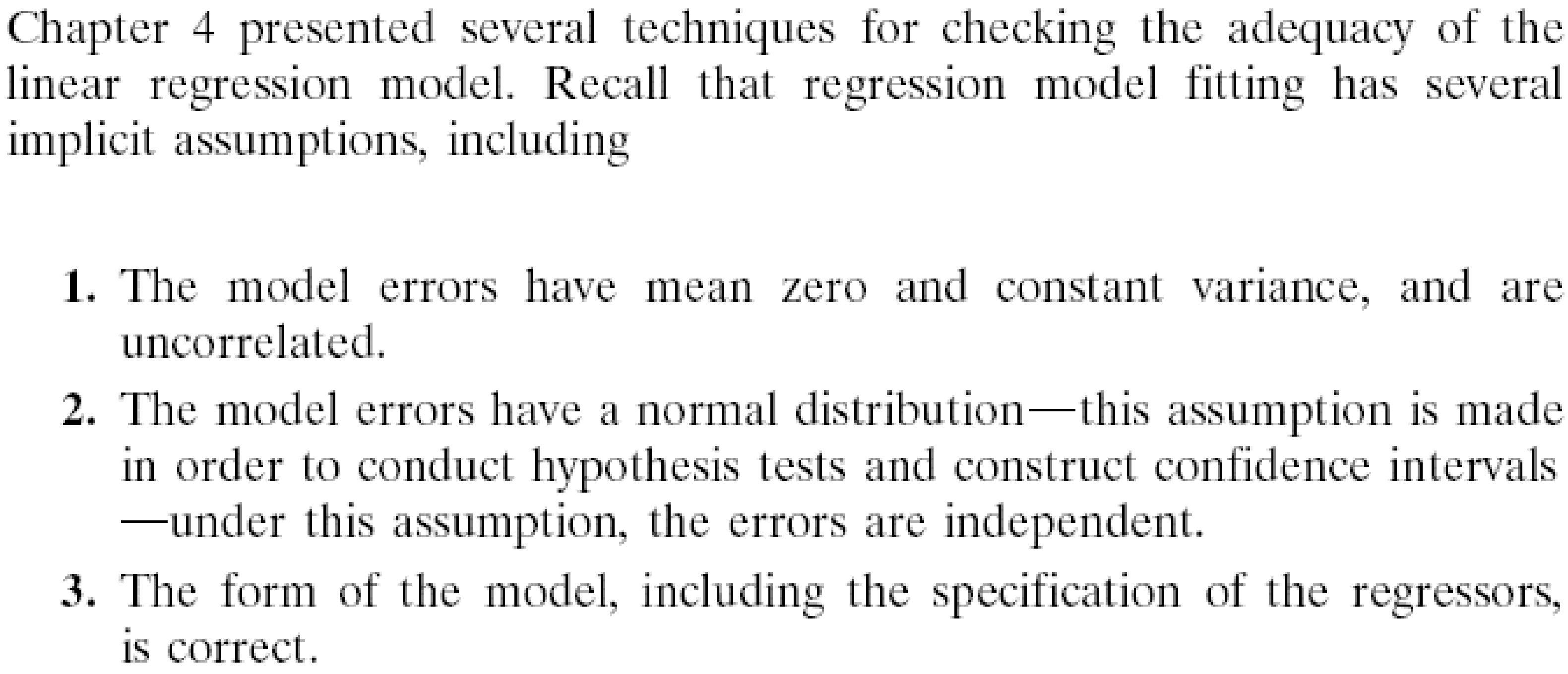
CHAPTER 5

Transformations and Weighting to Correct Model Inadequacies

Introduction



Chapter 4 presented several techniques for checking the adequacy of the linear regression model. Recall that regression model fitting has several implicit assumptions, including

1. The model errors have mean zero and constant variance, and are uncorrelated.
2. The model errors have a normal distribution—this assumption is made in order to conduct hypothesis tests and construct confidence intervals —under this assumption, the errors are independent.
3. The form of the model, including the specification of the regressors, is correct.

Introduction

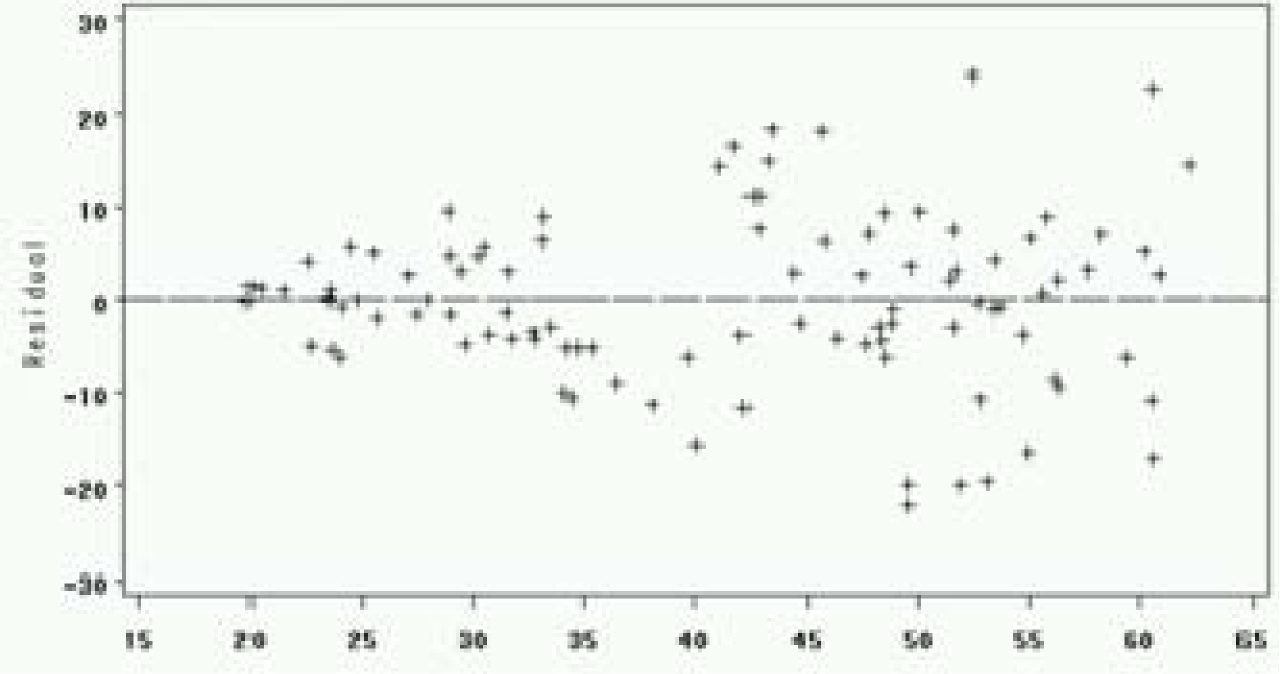
- Data Transformation
- Weighted Least Squares

Variance Stabilizing Transformations

- Constant variance assumption
  - Often violated when the variance is functionally related to the mean
  - Transformation on the response may eliminate the problem
  - The strength of the transformation depends on the amount of curvature that is induced
  - If not satisfied, the regression coefficients will have larger standard errors (less precision)

Residual Plots

- **Nonconstancy of Error Variance**
  - **Changing dispersion as the predictor variable changes in the plot of the residual against the predictor variable**



Variance Stabilizing Transformation

Useful Variance-Stabilizing Transformations

Relationship of σ^2 to E(y)	Transformation
$\sigma^2 \propto$ constant	$y' = y$ (no transformation)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$ (square root; Poisson data)
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$ (arcsin; binomial proportions $0 \leq y_i \leq 1$)
$\sigma^2 \propto [E(y)]^2$	$y' = \ln(y)$ (log)
$\sigma^2 \propto [E(y)]^3$	$y' = y^{-1/2}$ (reciprocal square root)
$\sigma^2 \propto [E(y)]^4$	$y' = y^{-1}$ (reciprocal)

Example 5.1

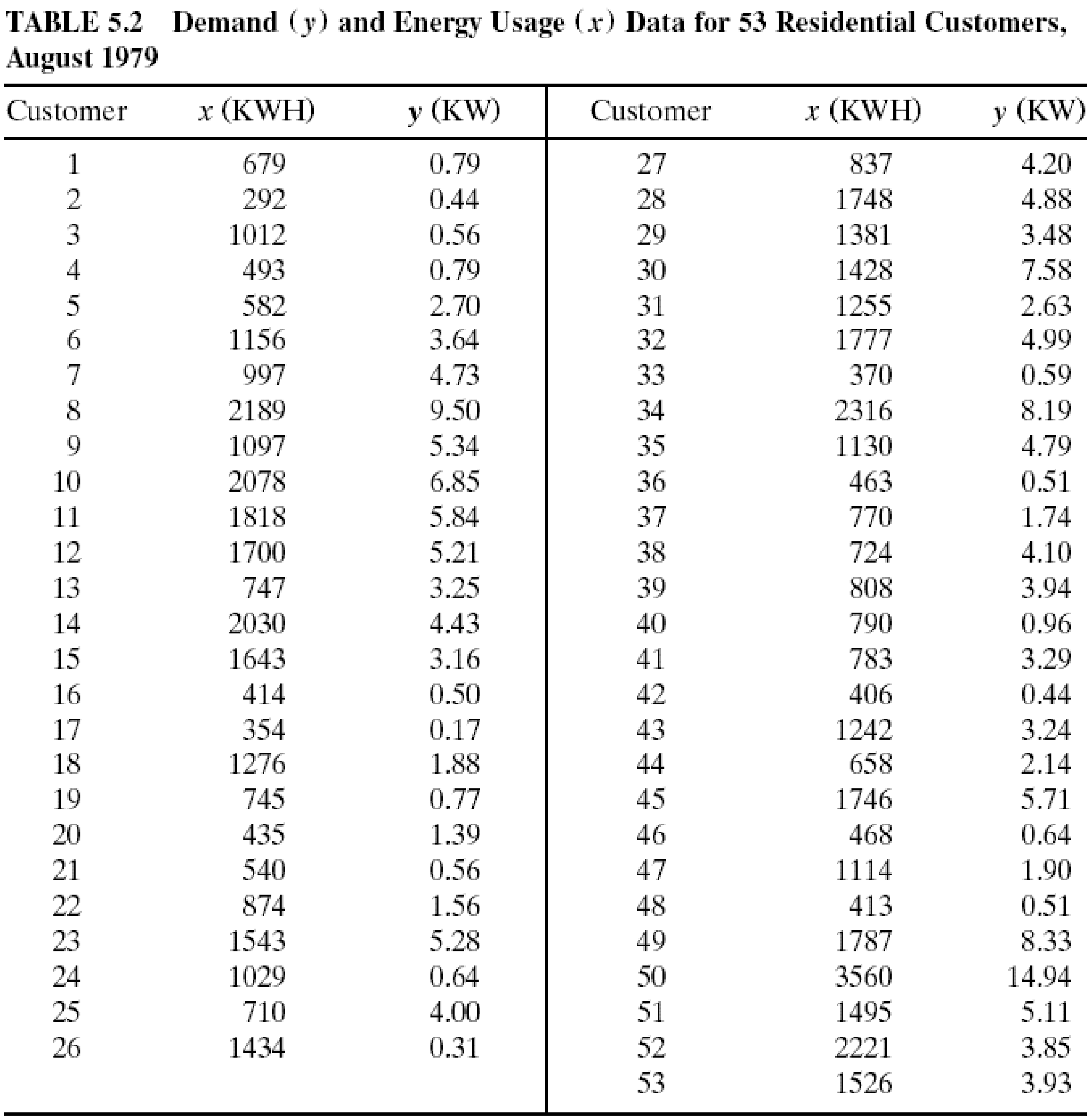


TABLE 5.2 Demand (y) and Energy Usage (x) Data for 53 Residential Customers, August 1979

Customer	x (KWH)	y (KW)
1	679	0.79
2	292	0.44
3	1012	0.56
4	493	0.79
5	582	2.70
6	1156	3.64
7	997	4.73
8	2189	9.50
9	1097	5.34
10	2078	6.85
11	1818	5.84
12	1700	5.21
13	747	3.25
14	2030	4.43
15	1643	3.16
16	414	0.50
17	354	0.17
18	1276	1.88
19	745	0.77
20	435	1.39
21	540	0.56
22	874	1.56
23	1543	5.28
24	1029	0.64
25	710	4.00
26	1434	0.31
27	837	4.20
28	1748	4.88
29	1381	3.48
30	1428	7.58
31	1255	2.63
32	1777	4.99
33	370	0.59
34	2316	8.19
35	1130	4.79
36	463	0.51
37	770	1.74
38	724	4.10
39	808	3.94
40	790	0.96
41	783	3.29
42	406	0.44
43	1242	3.24
44	658	2.14
45	1746	5.71
46	468	0.64
47	1114	1.90
48	413	0.51
49	1787	8.33
50	3560	14.94
51	1495	5.11
52	2221	3.85
53	1526	3.93

R code

```
# example 5.1
rm(list=ls())
# read data
Energy=read.csv("data-ex-5-1-(Electric-Utility).csv")
# visualize data
plot(Energy$Usage,Energy$Demand,pch=20)
# fit regression
model1=lm(Energy$Demand~Energy$Usage)
# residual plots based on x
plot(Energy$Usage,model1$residuals)
# residual plots based on fitted values
plot(model1$fitted.values,model1$residuals)
# residual plots based using R student
plot(model1$fitted.values,rstudent(model1))
```

Example 5.1

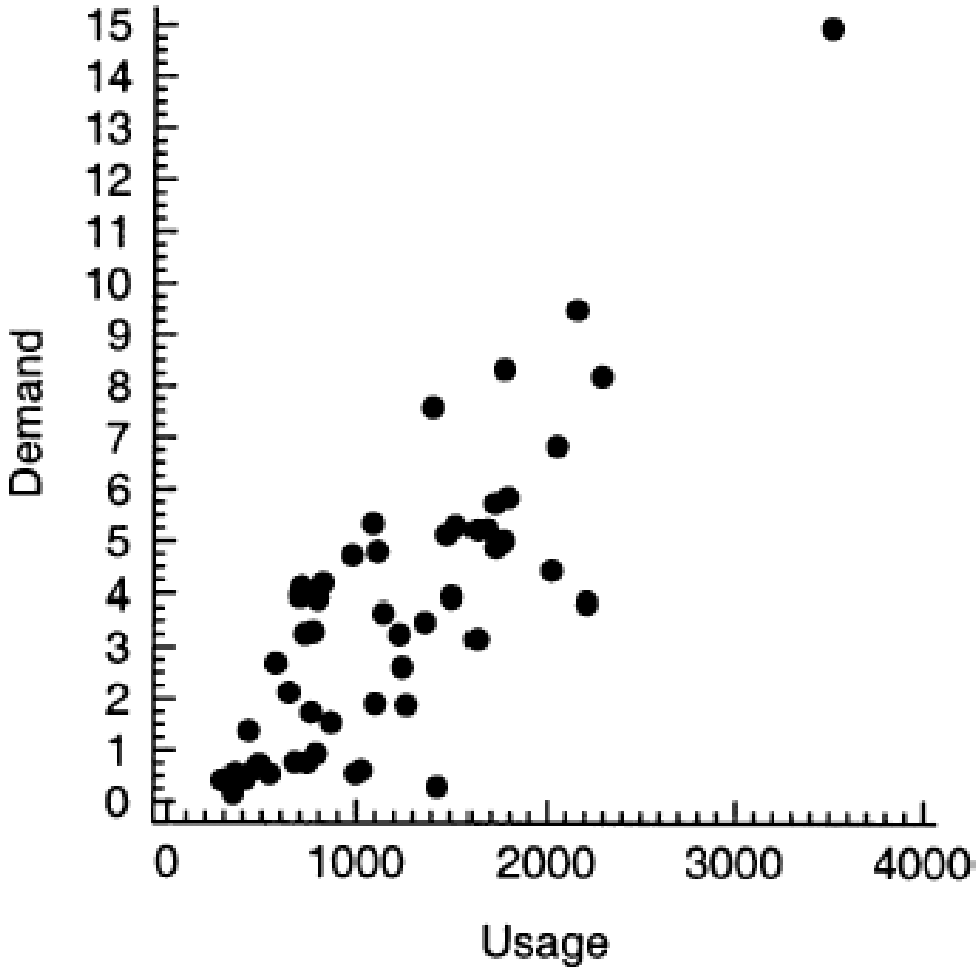


Figure 5.1 Scatter diagram of the energy demand (kW) versus energy usage (kWh), Example 5.1.

Example 5.1

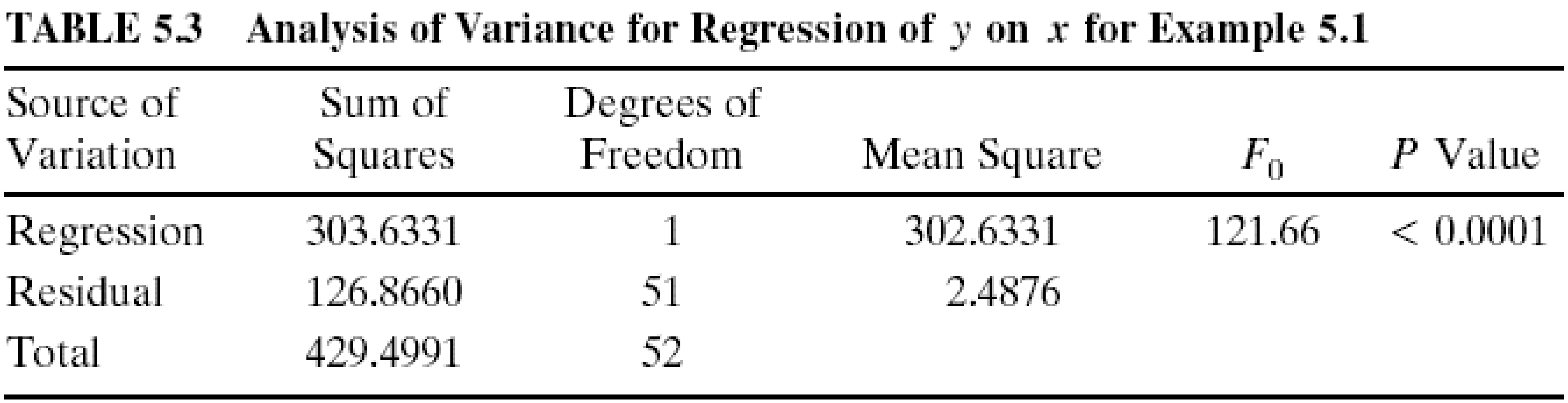


TABLE 5.3 Analysis of Variance for Regression of y on x for Example 5.1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P Value
Regression	303.6331	1	302.6331	121.66	< 0.0001
Residual	126.8660	51	2.4876		
Total	429.4991	52			

Example 5.1

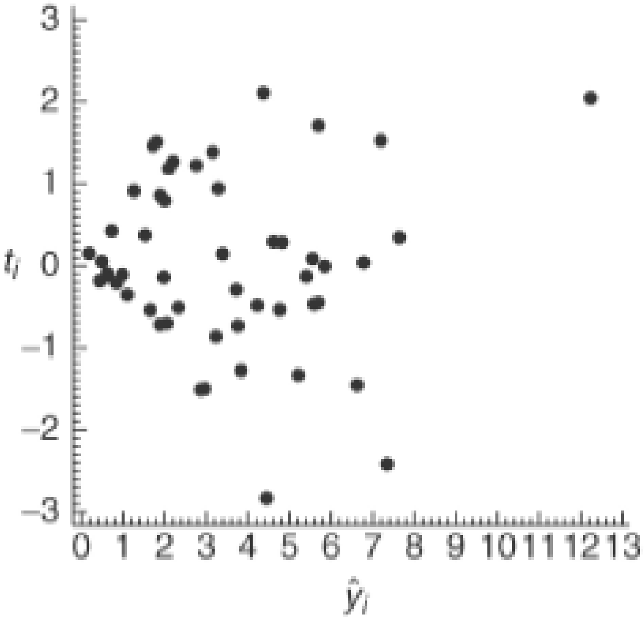


Figure 5.2 Plot of *R*-student values t_i versus fitted values $\hat{y}_i$, Example 5.1.

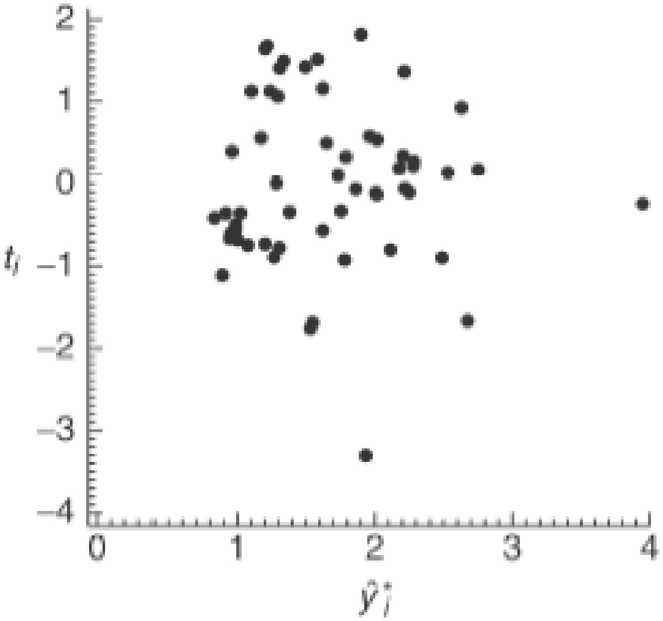


Figure 5.3 Plot of *R*-student values t_i versus fitted values y_i^* for the transformed data, Example 5.1.

Transformations to Linearize the Model

- Nonlinearity may be detected via the lack-of-fit test
- If a transformation of a nonlinear function can result in a linear function – we say it is *intrinsically* or *transformably* linear
- Example:

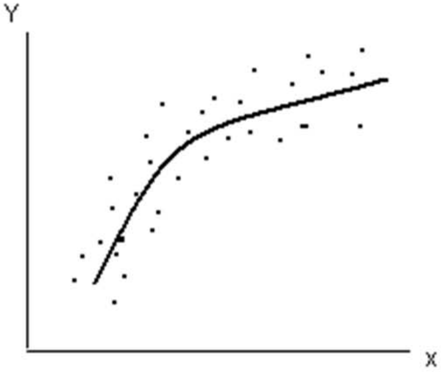
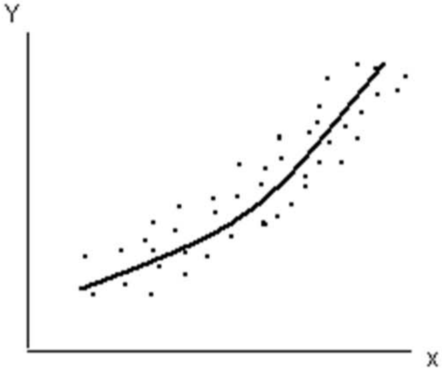
$$y = \beta_0 e^{\beta_1 x} \varepsilon$$

$$\ln y = \ln \beta_0 + \beta_1 x + \ln \varepsilon$$

$$y' = \beta_0' + \beta_1 x + \varepsilon'$$

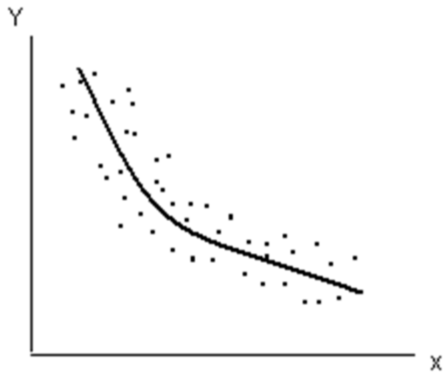
Transformations

- **Transformations for Nonlinear Relation Only**

Scatter Plot	Transformation of x to try
	$x` = log_{10}\, x, \;\; x` = \sqrt{x}$
	$x` = x^2, \;\; x` = e^x$

Transformations

- **Transformations for Nonlinear Relation Only**

	$x` = 1/x$, $x` = e^{-x}$

Transformations to Linearize the Model

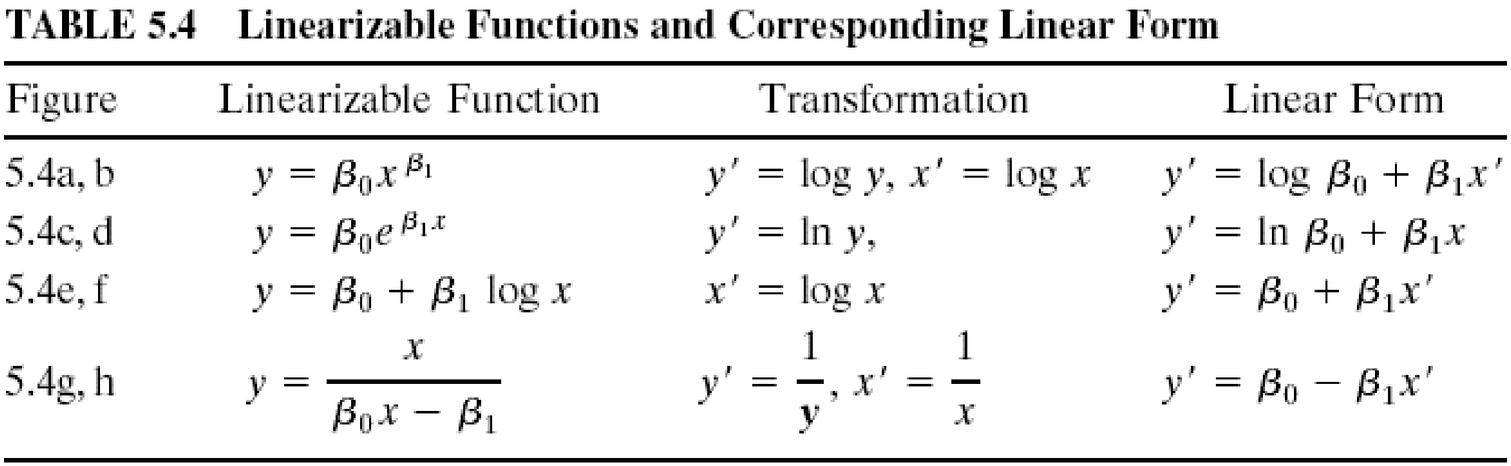


TABLE 5.4 Linearizable Functions and Corresponding Linear Form

Figure	Linearizable Function	Transformation	Linear Form
5.4a, b	$y = \beta_0 x^{\beta_1}$	$y' = \log y,\ x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4c, d	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y,$	$y' = \ln \beta_0 + \beta_1 x$
5.4e, f	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4g, h	$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y},\ x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Figure 5.4 (a)-(d)
Linearizable Functions [Daniel and Wood (1980)]

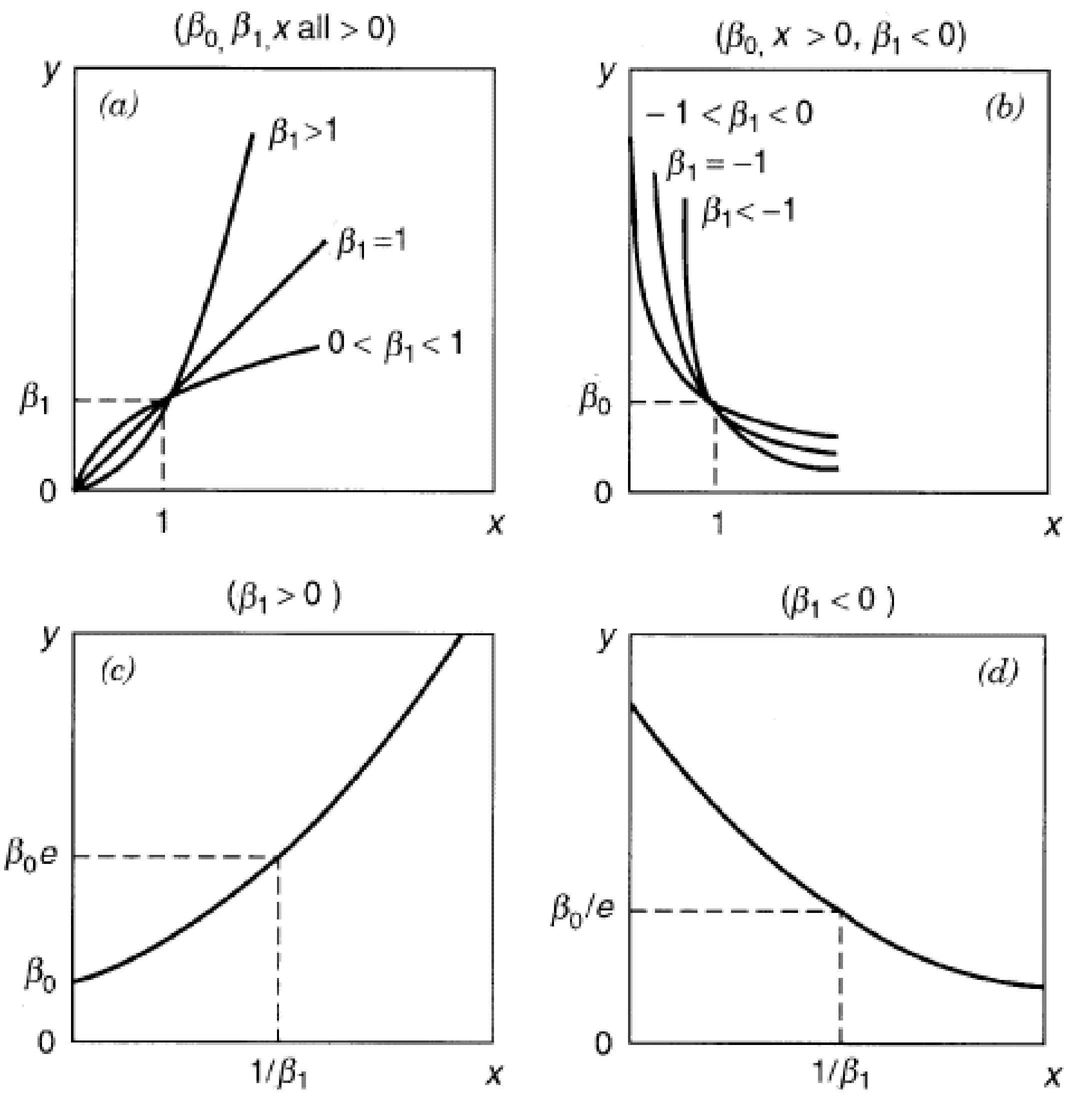
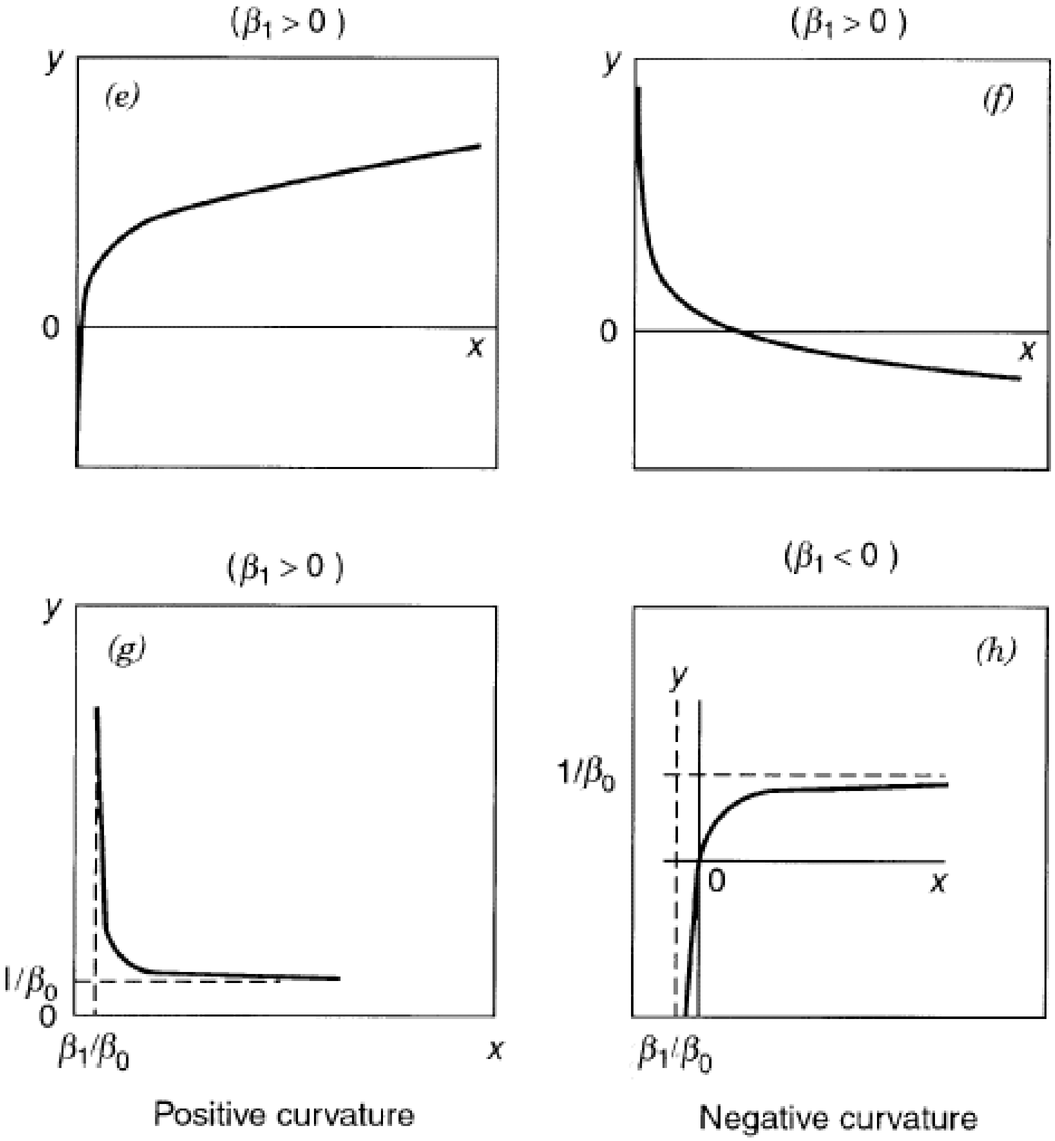


Figure 5.4 (e)-(h) Linearizable Functions [Daniel and Wood (1980)]



Transformations to Linearize the Model

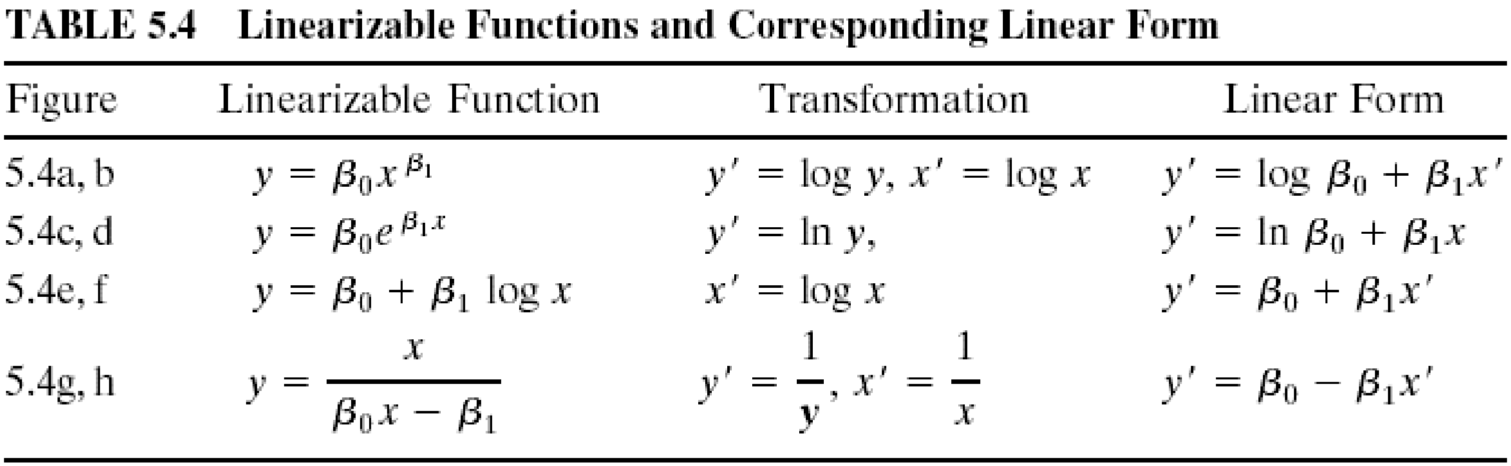


TABLE 5.4 Linearizable Functions and Corresponding Linear Form

Figure	Linearizable Function	Transformation	Linear Form
5.4a, b	$y = \beta_0 x^{\beta_1}$	$y' = \log y,\ x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4c, d	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y,$	$y' = \ln \beta_0 + \beta_1 x$
5.4e, f	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4g, h	$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{\mathbf{y}},\ x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Example 5.2 The Windmill Data

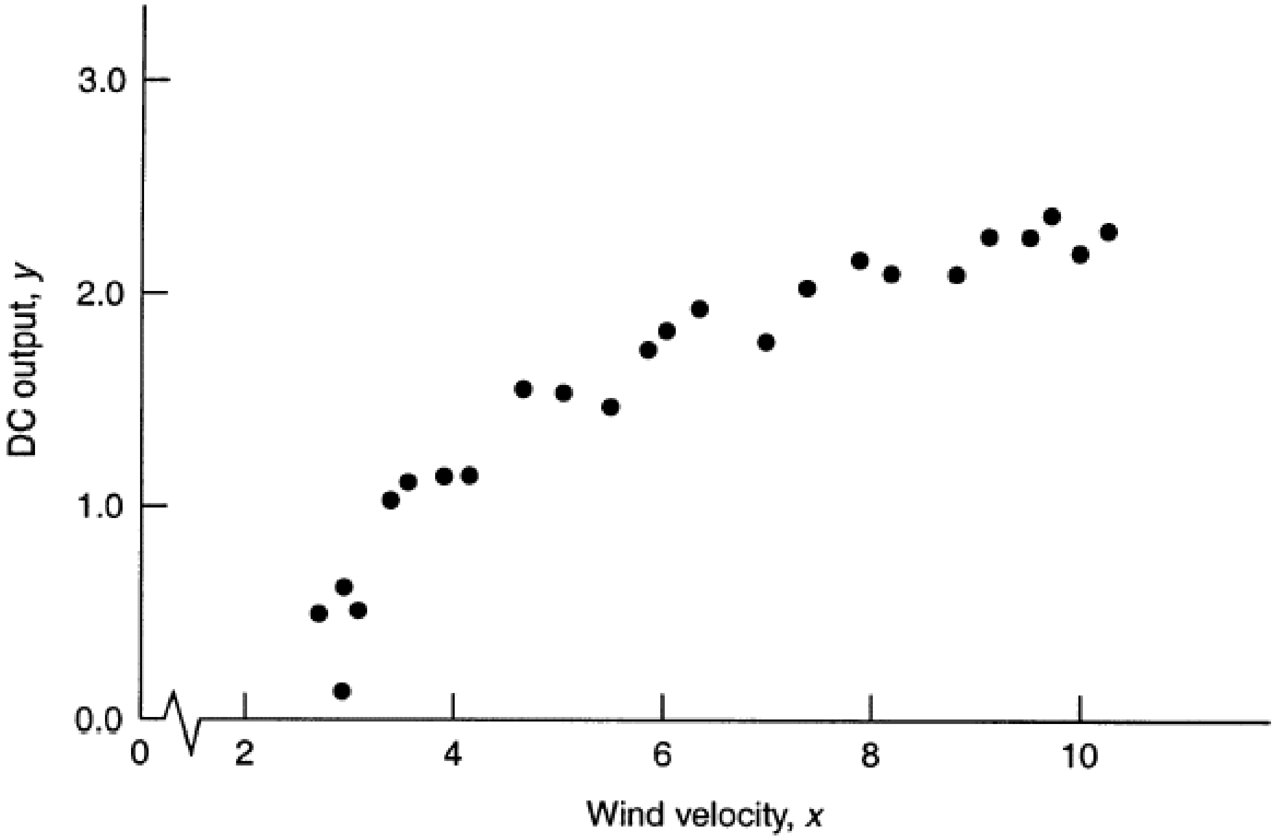


Figure 5.5 Plot of DC output y versus wind velocity x for the windmill data.

R code

- # example 5.2
- Wind=read.csv("data-ex-5-2-(Windmill).csv")
- # visualize data
- plot(Wind$Velocity,Wind$Output,pch=20)
- model2=lm(Wind$Output~Wind$Velocity)
- summary(model2)
- # residual plots based on fitted values
- plot(model2$fitted.values,model2$residuals)
- abline(h=0,col="grey",lwd=3)
- # residual plots based on R student
- plot(model2$fitted.values,rstudent(model2))

- # transform velocity
- Wind$InvVelo=1/Wind$Velocity
- model3=lm(Wind$Output~Wind$InvVelo)
- summary(model3)
- # residual plots based on fitted values
- plot(model3$fitted.values,model3$residuals)
- abline(h=0,col="grey",lwd=3)
- # residual plots based on R student
- plot(model3$fitted.values,rstudent(model3))

Example 5.2 The Windmill Data

- A straight line fit to the data resulted in:

$$\hat{y} = 0.1309 + 0.2411x$$

- The summary statistics are

$R^2 = 0.8745$,

$MS_{Res} = 0.0557$, and

$F_0 = 160.26$ (with a P-value < 0.0001)

Example 5.2 The Windmill Data

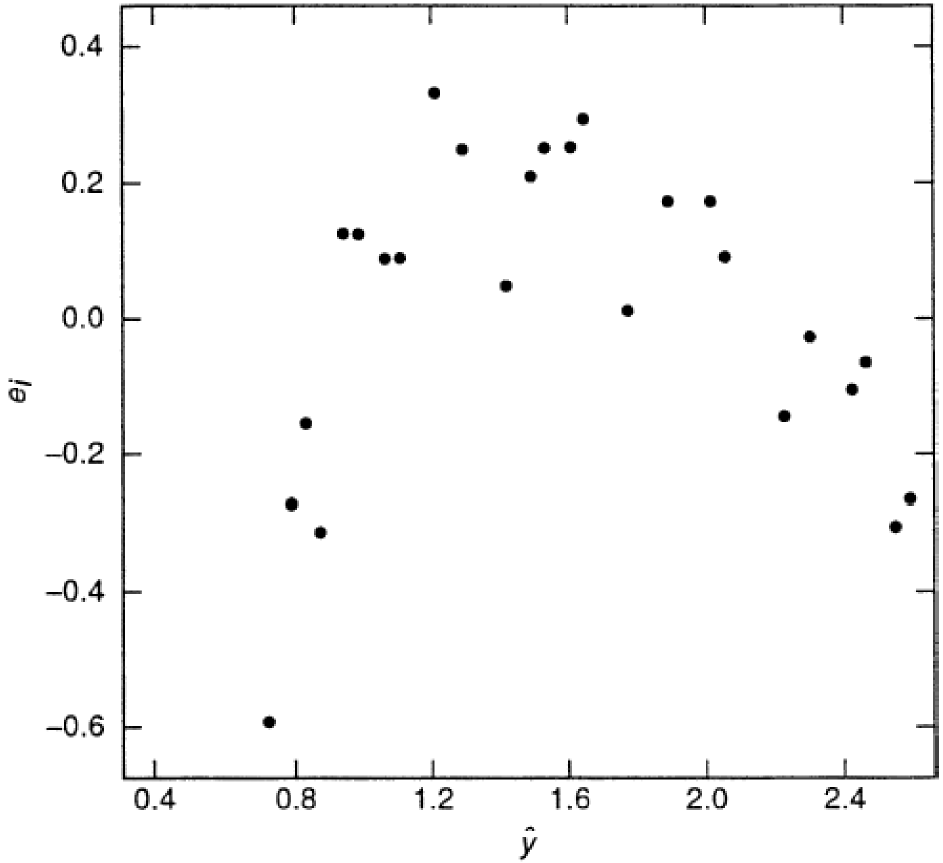


Figure 5.6 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

Example 5.2 The Windmill Data

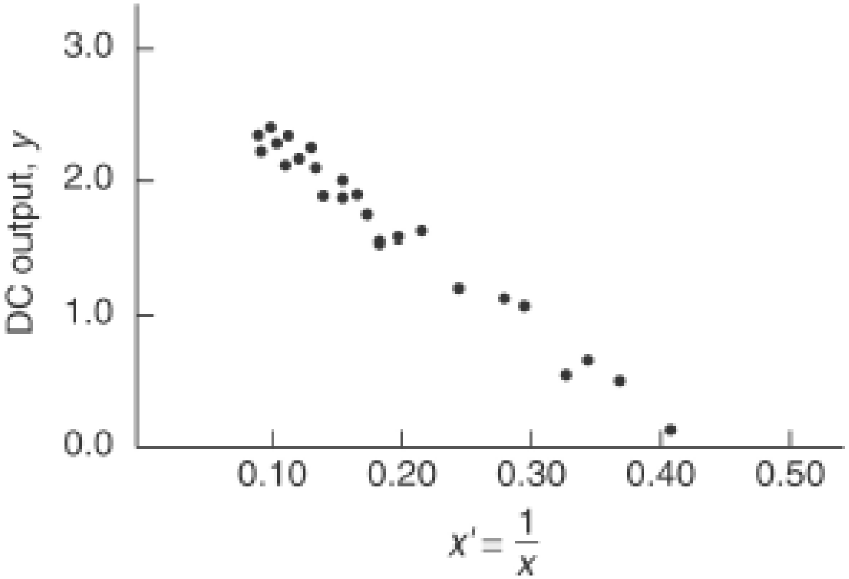


Figure 5.7 Plot of DC output versus $x' = 1/x$ for the windmill data.

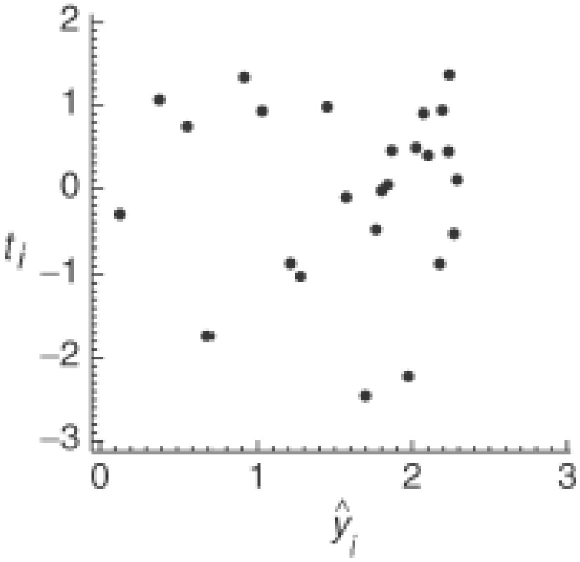


Figure 5.8 Plot of *R*-student values t_i versus fitted values $\hat{y}_i$ for the transformed model for the windmill data.

Example 5.2 The Windmill Data

Using a reciprocal transformation, the resulting fitted model is:

$$\hat{y} = 2.9789 - 6.9345x'$$

The summary statistics are

$R^2 = 0.9800$,

$MS_{Res} = 0.0089$, and

$F_0 = 1128.43$ (with a P-value < 0.0001)

Transformations to Linearize the Model

Note:

- Least-squares estimator has least-squares properties with respect to the transformed data, not original data.

Analytical Methods for Selecting a Transformation

- Transformations on *y*: The Box-Cox Method
- Transformations on the Regressor Variables

Transformations on y: The Box-Cox Method

- Suppose that we wish to transform *y* to correct nonnormality and/or nonconstant variance.
- A useful class of transformations is the **power transformation**, y^{λ} where λ is a parameter to be determined.
- The parameters of the regression model and λ can be estimated simultaneously using the method of maximum likelihood.

Transformations on y: The Box-Cox Method

- The appropriate procedure to be used is

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda - 1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases}$$

where $\dot{y} = \ln^{-1}[(1/n\Sigma_{i=1}^{n} \ln y_i]$ is the **geometric mean** of the observations

- The model to be fit is $\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Example 5.3 The Electric Utility Data

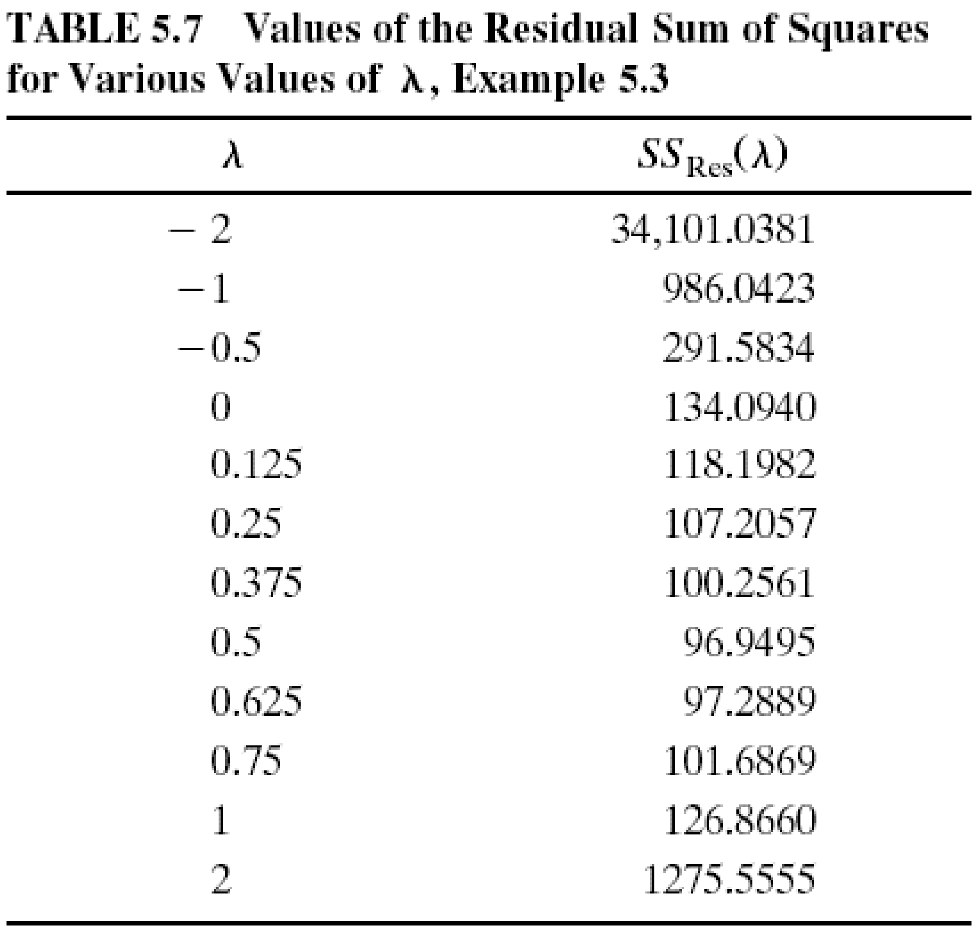


TABLE 5.7 Values of the Residual Sum of Squares for Various Values of λ, Example 5.3

λ	$SS_{\text{Res}}(\lambda)$
− 2	34,101.0381
−1	986.0423
−0.5	291.5834
0	134.0940
0.125	118.1982
0.25	107.2057
0.375	100.2561
0.5	96.9495
0.625	97.2889
0.75	101.6869
1	126.8660
2	1275.5555

R code

- # Box-Cox
- require(MASS)
- boxcox(model1)
- boxcox(model2)
- boxcox(model3)

Transformations on the Regressors

- Sometimes a transformation on one or more regressor variables is useful
- These transformations are often selected empirically

Weighting to Correct Model Inadequacies

- No suitable transformation found
- Use weighted least squares

Weighting to Correct Model Inadequacies

- **Weighted Least Squares**

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i$$

$$where:\ \beta_0, \beta_1, \dots, \beta_{p-1} are\ parameters$$

$$X_1, \dots, X_{p-1} are\ known\ constants$$

$$\varepsilon_i are\ independent\ N\left(0, \sigma_i^2\right) i = 1, \dots, n$$

Weighted Least Squares

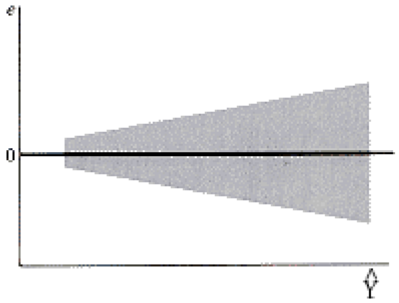
- **Error Variances Known**

$$w_i = {}^{1}\!/_{\sigma_i^2}$$

$$Q_w = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

Weighted Least Squares

- **Error Variances Unknown**
  - **Estimate σ_i^2**
    - **Megaphone residual plot**



$$\sigma_i^2 = E\left(\varepsilon_i^2\right) - \left(E(\varepsilon_i)\right)^2$$

$$\sigma_i^2 = E\left(\varepsilon_i^2\right)$$

- ε_i^2 is an estimator of σ_i^2
- $|\varepsilon_i|$ is an estimator of σ_i

Weighted Least Squares

- Error Variances Unknown
  - Summary
    - Fit the regression model by unweighted least squares and analyze the residuals
    - Estimate the variance or standard deviation function by regressing the squared or absolute residuals on the appropriate variables
    - Use the fitted values to obtain the weights
    - Estimate the regression coefficients using these weights
    - Use process above iteratively to stabilize the estimated regression coefficients

Weighted Least Squares

- **Comments**
  - **Nonconstant error variance --- heteroscedasticity**
  - **Constant error variance --- homoscedasticity**
  - R^2 **may be provided by software packages but does not have a clear-cut meaning for weighted least squares**

R code

```
# weight lm
Food=read.csv("table5.9.csv")
plot(Food$AdvertisingExpense,Food$Income)
# build linear regression without weights
model_noweight=lm(Food$Income~Food$AdvertisingExpense)
# check the coefficient
summary(model_noweight)
# generate residual plot
plot(Food$AdvertisingExpense,model_noweight$residuals)
# regress residual^2 on covariate
weight_coef=lm( model_noweight$residuals^2~Food$AdvertisingExpense )$coef
# build linear regression with weights
model_weight=lm(Food$Income~Food$AdvertisingExpense,weights=1/(weight_coef[1]+Food$AdvertisingExpense*weight_coef[2]))
# check coefficients
summary(model_weight)
# generate residual plot with weighted residuals
plot(Food$AdvertisingExpense,
     1/sqrt(weight_coef[1]+Food$AdvertisingExpense*weight_coef[2]) * model_weight$residuals)
abline(h=0,col="grey",lwd=3)
```