CHAPTER 6

TRANSFORMATION OF VARIABLES

6.1 INTRODUCTION

Data do not always come in a form that is immediately suitable for analysis. We often have to transform the variables before carrying out the analysis. Transformations are applied to accomplish certain objectives such as to ensure linearity, to achieve normality, or to stabilize the variance. It often becomes necessary to fit a linear regression model to the transformed rather than the original variables. This is common practice. In this chapter, we discuss the situations where it is necessary to transform the data, the possible choices of transformation, and the analysis of transformed data.

We illustrate transformation mainly using simple regression. In multiple regression where there are several predictors, some may require transformation and others may not. Although the same technique can be applied to multiple regression, transformation in multiple regression requires more effort and care.

The necessity for transforming the data arises because the original variables, or the model in terms of the original variables, violates one or more of the standard regression assumptions. The most commonly violated assumptions are those concerning the linearity of the model and the constancy of the error variance. As mentioned in Chapters 2 and 3, a regression model is linear when the parameters

present in the model occur linearly even if the predictor variables occur nonlinearly. For example, each of the four following models is linear:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon, \\ Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon, \\ Y &= \beta_0 + \beta_1 \log X + \varepsilon, \\ Y &= \beta_0 + \beta_1 \sqrt{X} + \varepsilon, \end{aligned}$$

because the model parameters $\beta_0, \beta_1, \beta_2$ enter linearly. On the other hand,

$$Y = \beta_0 + e^{\beta_1 X} + \varepsilon$$

is a nonlinear model because the parameter β_1 does not enter the model linearly. To satisfy the assumptions of the standard regression model, instead of working with the original variables, we sometimes work with transformed variables. Transformations may be necessary for several reasons.

1. Theoretical considerations may specify that the relationship between two variables is nonlinear. An appropriate transformation of the variables can make the relationship between the transformed variables linear. Consider an example from learning theory (experimental psychology). A learning model that is widely used states that the time taken to perform a task on the ith occasion (T_i) is

$$T_i = \alpha\beta^i, \quad \alpha > 0, \quad 0 < \beta < 1. \tag{6.1}$$

The relationship between (T_i) and i as given in (6.1) is nonlinear, and we cannot directly apply techniques of linear regression. On the other hand, if we take logarithms of both sides, we get

$$\log T_i = \log \alpha + i \log \beta, \tag{6.2}$$

showing that $\log T_i$ and i are linearly related. The transformation enables us to use standard regression methods. Although the relationship between the original variables was nonlinear, the relationship between transformed variables is linear. A transformation is used to achieve the linearity of the fitted model.

2. The response variable Y, which is analyzed, may have a probability distribution whose variance is related to the mean. If the mean is related to the value of the predictor variable X, then the variance of Y will change with X, and will not be constant. The distribution of Y will usually also be non-normal under these conditions. Non-normality invalidates the standard tests of significance (although not in a major way with large samples) since they are based on the normality assumption. The unequal variance of the error terms will produce estimates that are unbiased, but are no longer best in the sense of having the smallest variance. In these situations we often transform the data so as to ensure normality and constancy of error variance. In

Table 6.1 Linearizable Simple Regression Functions with Corresponding Transformations

Function	Transformation	Linear Form	Graph
$Y = \alpha X^{\beta}$	$Y' = \log Y, X' = \log X$	$Y' = \log \alpha + \beta X'$	Figure 6.1
$Y = \alpha e^{\beta X}$	$Y' = \ln Y$	$Y' = \ln \alpha + \beta X$	Figure 6.2
$Y = \alpha + \beta \log X$	$X' = \log X$	$Y = \alpha + \beta X'$	Figure 6.3
$Y = \frac{X}{\alpha X - \beta}$	$Y' = \frac{1}{Y}, X' = \frac{1}{X}$	$Y' = \alpha - \beta X'$	Figure 6.4(a)
$Y = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$	$Y' = \ln \frac{Y}{1-Y}$	$Y' = \alpha + \beta X$	Figure 6.4(b)

In Chapter 6 we describe an application using the transformation in the last line of the table.

practice, the transformations are chosen to ensure the constancy of variance (*variance-stabilizing transformations*). It is a fortunate coincidence that the variance-stabilizing transformations are also good normalizing transforms.

3. There are neither prior theoretical nor probabilistic reasons to suspect that a transformation is required. The evidence comes from examining the residuals from the fit of a linear regression model in which the original variables are used.

Each of these cases where transformation is needed is illustrated in the following sections.

6.2 TRANSFORMATIONS TO ACHIEVE LINEARITY

One of the standard assumptions made in regression analysis is that the model which describes the data is linear. From theoretical considerations, or from an examination of scatter plot of Y against each predictor X_j, the relationship between Y and X_j may appear to be nonlinear. There are, however, several simple nonlinear regression models which by appropriate transformations can be made linear. We list some of these linearizable curves in Table 6.1. The corresponding graphs are given in Figures 6.1 to 6.4.

When curvature is observed in the scatter plot of Y against X, a linearizable curve from one of those given in Figures 6.1 to 6.4 may be chosen to represent the data. There are, however, many simple nonlinear models that cannot be linearized. Consider for example, $Y = \alpha + \beta\delta^X$, a modified exponential curve, or

$$Y = \alpha_1 e^{\theta_1 X} + \alpha_2 e^{\theta_2 X},$$

which is the sum of two exponential functions. The strictly nonlinear models (i.e., those not linearizable by variable transformation) require very different methods for fitting. We do not describe them in this book but refer the interested reader to Bates and Watts (1988) and Seber and Wild (1989), and Ratkowsky (1990).

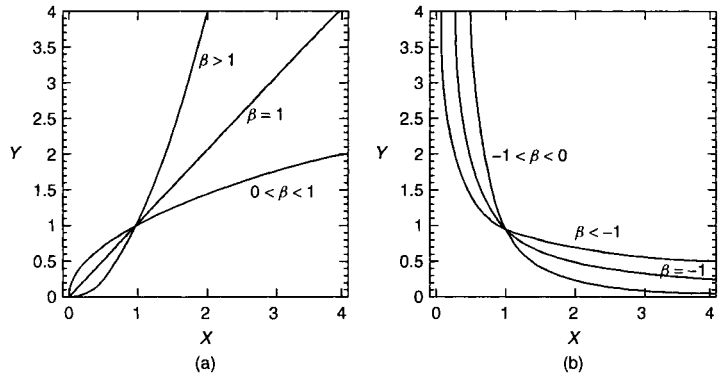

Figure 6.1 Graphs of the linearizable function $Y = \alpha X^{\beta}$.

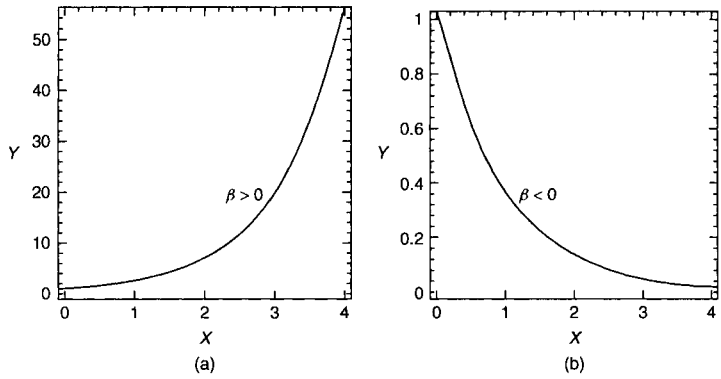

Figure 6.2 Graphs of the linearizable function $Y = \alpha e^{\beta X}$.

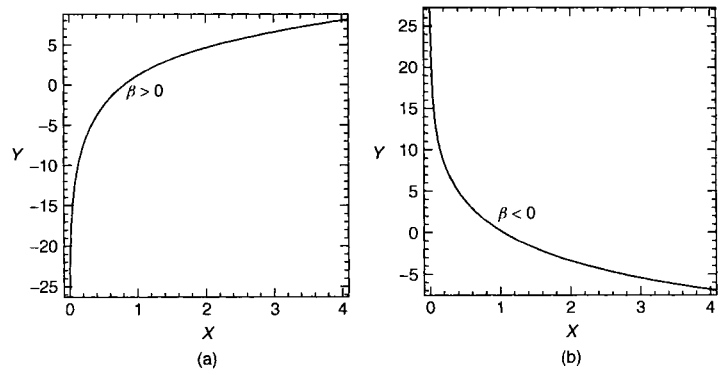

Figure 6.3 Graphs of the linearizable function $Y = \alpha + \beta \log X$.

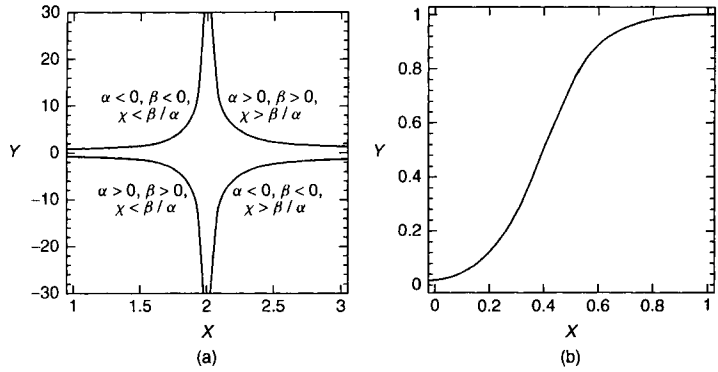

Figure 6.4 Graphs of the linearizable functions: (a) $Y = X/(\alpha\ X - \beta)$, and (b) $Y = (e^{\alpha+\beta\ X})/(1 + e^{\alpha+\beta\ X})$.

Table 6.2 Number of Surviving Bacteria (Units of 100)

t	n_t	t	n_t	t	n_t
1	355	6	106	11	36
2	211	7	104	12	32
3	197	8	60	13	21
4	166	9	56	14	19
5	142	10	38	15	15

In the following example, theoretical considerations lead to a model that is nonlinear. The model is, however, linearizable and we indicate the appropriate analysis.

6.3 BACTERIA DEATHS DUE TO X-RAY RADIATION

The data given in Table 6.2 represent the number of surviving bacteria (in hundreds) as estimated by plate counts in an experiment with marine bacterium following exposure to 200-kilovolt X-rays for periods ranging from $t = 1$ to 15 intervals of 6 minutes. The data can also be found in the book's Web Site.[1] The response variable n_t represents the number surviving after exposure time t. The experiment was carried out to test the single-hit hypothesis of X-ray action under constant field of radiation. According to this theory, there is a single vital center in each bacterium, and this must be hit by a ray before the bacteria is inactivated or killed. The particular bacterium studied does not form clumps or chains, so the number of bacterium can be estimated directly from plate counts.

[1] http://www.ilr.cornell.edu/~hadi/RABE4

If the theory is applicable, then n_t and t should be related by

$$n_t = n_0 e^{\beta_1 t}, \quad t \geq 0, \tag{6.3}$$

where n_0 and β_1 are parameters. These parameters have simple physical interpretations; n_0 is the number of bacteria at the start of the experiment, and β_1 is the destruction (decay) rate. Taking logarithms of both sides of (6.3), we get

$$\ln n_t = \ln n_0 + \beta_1 t = \beta_0 + \beta_1 t, \tag{6.4}$$

where $\beta_0 = \ln n_0$ and we have $\ln n_t$ as a linear function of t. If we introduce ε_t as the random error, our model becomes

$$\ln n_t = \beta_0 + \beta_1 t + \varepsilon_t \tag{6.5}$$

and we can now apply standard least squares methods.

To get the error ε_t in the transformed model (6.5) to be additive, the error must occur in the multiplicative form in the original model (6.3). The correct representation of the model should be

$$n_t = n_0 e^{\beta_1 t} \varepsilon_t', \tag{6.6}$$

where ε_t' is the multiplicative random error. By comparing (6.5) and (6.6), it is seen that $\varepsilon_t = \ln \varepsilon_t'$. For standard least squares analysis ε_t should be normally distributed, which in turn implies that ε_t', has a log-normal distribution.[2] In practice, after fitting the transformed model we look at the residuals from the fitted model to see if the model assumptions hold. No attempt is usually made to investigate the random component, ε_t', of the original model.

6.3.1 Inadequacy of a Linear Model

The first step in the analysis is to plot the raw data n_t versus t. The plot, shown in Figure 6.5, suggests a nonlinear relationship between n_t and t. However, we proceed by fitting the simple linear model and investigate the consequences of misspecification. The model is

$$n_t = \beta_0 + \beta_1 t + \varepsilon_t, \tag{6.7}$$

where β_0 and β_1 are constants; ε_t's are the random errors, with zero means and equal variances, and are uncorrelated with each other. Estimates of β_0, β_1, their standard errors, and the square of the correlation coefficient are given in Table 6.3. Despite the fact that the regression coefficient for the time variable is significant and we have a high value of R^2, the linear model is not appropriate. The plot of n_t against t shows departure from linearity for high values of t (Figure 6.5). We see this even more clearly if we look at a plot of the standardized residuals against time

[2]The random variable Y is said to have a log-normal distribution if $\ln Y$ has a normal distribution.

Table 6.3 Estimated Regression Coefficients From Model (6.7)

Variable	Coefficient	s.e.	t-test	p-value
Constant	259.58	22.73	11.42	< 0.0001
TIME (t)	-19.46	2.50	-7.79	< 0.0001
	$n = 15$	$R^2 = 0.823$	$\hat{\sigma} = 41.83$	$d.f. = 13$

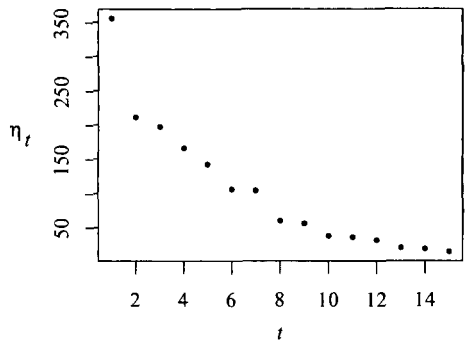

Figure 6.5 Plot of n_t against time t.

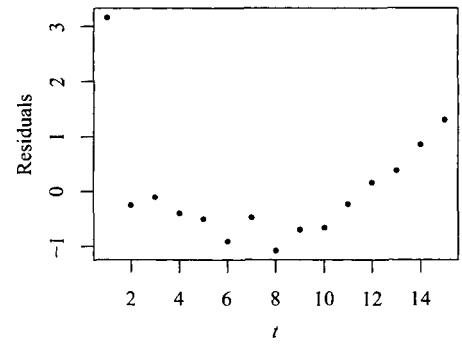

Figure 6.6 Plot of the standardized residuals from (6.7) against time t.

(Figure 6.6). The distribution of residuals has a distinct pattern. The residuals for $t = 2$ through 11 are all negative, for $t = 12$ through 15 are all positive, whereas the residual for $t = 1$ appears to be an outlier. This systematic pattern of deviation confirms that the linear model in (6.7) does not fit the data.

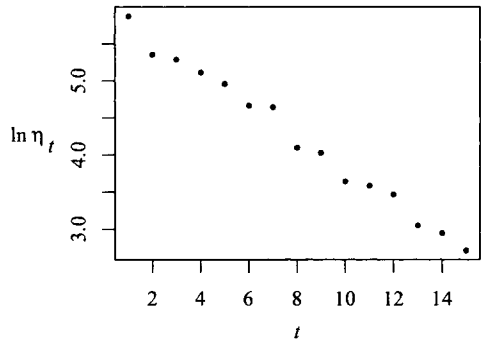

Figure 6.7 Plot of $\ln n_t$ against time t.

Table 6.4 Estimated Regression Coefficients When $\ln n_t$ Is Regressed on Time t

Variable	Coefficient	s.e.	t-test	p-value
Constant	5.973	0.0598	99.9	< 0.0001
TIME (t)	-0.218	0.0066	-33.2	< 0.0001
	$n = 15$	$R^2 = 0.988$	$\hat{\sigma} = 0.11$	$d.f. = 13$

6.3.2 Logarithmic Transformation for Achieving Linearity

The relation between n_t and t appears distinctly nonlinear and we will work with the transformed variable $\ln n_t$, which is suggested from theoretical considerations as well as by Figure 6.7. The plot of $\ln n_t$ against t appears linear, indicating that the logarithmic transformation is appropriate. The results of fitting (6.5) appear in Table 6.4. The coefficients are highly significant, the standard errors are reasonable, and nearly 99% of the variation in the data is explained by the model. The standardized residuals are plotted against t in Figure 6.8. There are no systematic patterns to the distribution of the residuals and the plot is satisfactory. The single-hit hypothesis of X-ray action, which postulates that $\ln n_t$ should be linearly related to t, is confirmed by the data.

While working with transformed variables, careful attention must be paid to the estimates of the parameters of the model. In our example the point estimate of β_1 is -0.218 and the 95% confidence interval for the same parameter is $(-0.232, -0.204)$. The estimate of the constant term in the equation is the best linear unbiased estimate of $\ln n_0$. If $\hat{\beta}_0$ denotes the estimate, $e^{\hat{\beta}_0}$ may be used as an estimate of n_0. With $\hat{\beta}_0 = 5.973$, the estimate of n_0 is $e^{\hat{\beta}_0} = 392.68$. This estimate is not an unbiased estimate of n_0; that is, the true size of the bacteria population at the start of the experiment was probably somewhat smaller than 392.68. A correction can be made to reduce the bias in the estimate of n_0. The estimate $exp[\hat{\beta}_0 - \frac{1}{2}\mathrm{Var}(\hat{\beta}_0)]$

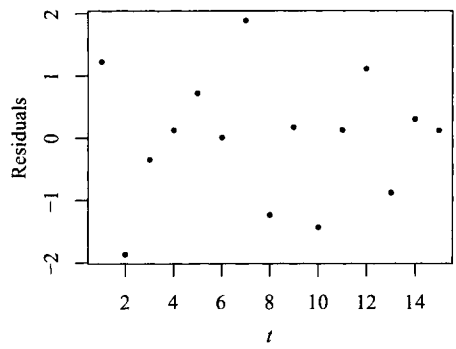

Figure 6.8 Plot of the standardized residuals against time t after transformation.

is nearly unbiased of n_0. In our present example, the modified estimate of n_0 is 381.11. Note that the bias in estimating n_0 has no effect on the test of the theory or the estimation of the decay rate.

In general, if nonlinearity is present, it will show up in a plot of the data. If the plot corresponds approximately to one of the graphs given in Figures 6.1 to 6.4, one of those curves can be fitted after transforming the data. The adequacy of the transformed model can then be investigated by methods outlined in Chapter 4.

6.4 TRANSFORMATIONS TO STABILIZE VARIANCE

We have discussed in the preceding section the use of transformations to achieve linearity of the regression function. Transformations are also used to stabilize the error variance, that is, to make the error variance constant for all the observations. The constancy of error variance is one of the standard assumptions of least squares theory. It is often referred to as the assumption of *homoscedasticity*. When the error variance is not constant over all the observations, the error is said to be *heteroscedastic*. *Heteroscedasticity* is usually detected by suitable graphs of the residuals such as the scatter plot of the standardized residuals against the fitted values or against each of the predictor variables. A plot with the characteristics of Figure 6.9 typifies the situation. The residuals tend to have a funnel-shaped distribution, either fanning out or closing in with the values of X.

If heteroscedasticity is present, and no corrective action is taken application of OLS to the raw data will result in estimated coefficients which lack precision in a theoretical sense. The estimated standard errors of the regression coefficients are often understated, giving a false sense of accuracy.

Heteroscedasticity can be removed by means of a suitable transformation. We describe an approach for (a) detecting heteroscedasticity and its effects on the

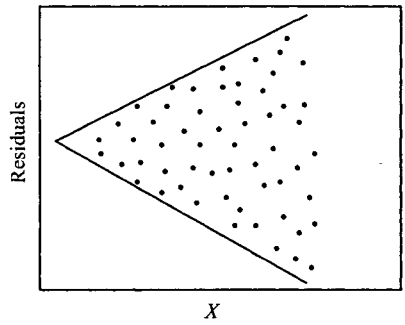

Figure 6.9 An example of heteroscedastic residuals.

analysis, and (b) removing heteroscedasticity from the data analyzed using transformations.

The response variable Y, in a regression problem, may follow a probability distribution whose variance is a function of the mean of that distribution. One property of the normal distribution, that many other probability distributions do not have, is that its mean and variance are independent in the sense that one is not a function of the other. The binomial and Poisson are but two examples of common probability distributions that have this characteristic. We know, for example, that a variable that is distributed binomially with parameters n and π has mean $n\pi$ and variance $n\pi(1 - \pi)$. It is also known that the mean and variance of a Poisson random variable are equal. When the relationship between the mean and variance of a random variable is known, it is possible to find a simple transformation of the variable, which makes the variance approximately constant (stabilizes the variance). We list in Table 6.5, for convenience and easy reference, transformations that stabilize the variance for some random variables with commonly occurring probability distributions whose variances are functions of their means. The transformations listed in Table 6.5 not only stabilize the variance, but also have the effect of making the distribution of the transformed variable closer to the normal distribution. Consequently, these transformations serve the dual purpose of normalizing the variable as well as making the variance functionally independent of the mean.

As an illustration, consider the following situation: Let Y be the number of accidents and X the speed of operating a lathe in a machine shop. We want to study the relationship between the number of accidents Y and the speed of lathe operation X. Suppose that a linear relationship is postulated between Y and X and is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is the random error. The mean of Y is seen to increase with X. It is known from empirical observation that rare events (events with small probabilities of occurrence) often have a Poisson distribution. Let us assume that Y has a Poisson

Table 6.5 Transformations to Stabilize Variance

Probability Distribution of Y	$Var(Y)$ in Terms of Its Mean μ	Transformation	Resulting Variance
Poisson[a]	μ	$\sqrt{Y}$ or $(\sqrt{Y} + \sqrt{Y+1})$	0.25
Binomial[b]	$\mu(1-\mu)/n$	$\sin^{-1}\sqrt{Y}$ (degrees)	$821/n$
		$\sin^{-1}\sqrt{Y}$ (radians)	$0.25/n$
Negative Binomial[c]	$\mu + \lambda^2\mu^2$	$\lambda^{-1}\sinh^{-1}(\lambda\sqrt{Y})$ or	
		$\lambda^{-1}\sinh^{-1}(\lambda\sqrt{Y} + 0.5)$	0.25

[a] For small values of Y, $\sqrt{Y+0.5}$ is sometimes recommended.

[b] n is an index describing the sample size; for $Y = r/n$ a slightly better transformation is $\sin^{-1}\sqrt{(r+3/8)/(n+3/4)}$.

[c] Note that the parameter $\lambda = 1/\sqrt{r}$.

distribution. Since the mean and variance of Y are the same,[3] it follows that the variance of Y is a function of X, and consequently the assumption of homoscedasticity will not hold. From Table 6.5 we see that the square root of a Poisson variable $(\sqrt{Y})$ has a variance independent of the mean and is approximately equal to 0.25. To ensure homoscedasticity we, therefore, regress $\sqrt{Y}$ on X. Here the transformation is chosen to stabilize the variance, the specific form being suggested by the assumed probability distribution of the response variable. An analysis of data employing transformations suggested by probabilistic considerations is demonstrated in the following example.

Injury Incidents in Airlines

The number of injury incidents and the proportion of total flights from New York for nine ($n = 9$) major United States, airlines for a single year is given in Table 6.6 and plotted in Figure 6.10. Let f_i and y_i denote the total flights and the number of injury incidents for the ith airline that year. Then the proportion of total flights n_i made by the ith airline is

$$n_i = \frac{f_i}{\sum f_i}.$$

If all the airlines are equally safe, the injury incidents can be explained by the model

$$y_i = \beta_0 + \beta_1 n_i + \varepsilon_i,$$

where β_0 and β_1 are constants and ε_i is the random error.

[3] The probability mass function of a Poisson random variable Y is $Pr(Y = y) = e^{-\lambda}\lambda^y/y!$; $y = 0, 1, \ldots$, where λ is a parameter. The mean and variance of a Poisson random variable are equal to λ.

Table 6.6 Number of Injury Incidents Y and Proportion of Total Flights N

Row	Y	N	Row	Y	N	Row	Y	N
1	11	0.0950	4	19	0.2078	7	3	0.1292
2	7	0.1920	5	9	0.1382	8	1	0.0503
3	7	0.0750	6	4	0.0540	9	3	0.0629

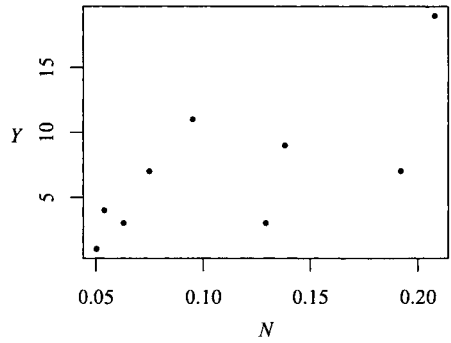

Figure 6.10 Plot of Y against N.

Table 6.7 Estimated Regression Coefficients (When Y is Regressed on N)

Variable	Coefficient	s.e.	t-test	p-value
Constant	−0.14	3.14	−0.045	0.9657
N	64.98	25.20	2.580	0.0365
	$n = 9$	$R^2 = 0.487$	$\hat{\sigma} = 4.201$	$d.f. = 7$

The results of fitting the model are given in Table 6.7. The plot of residuals against n_i is given in Figure 6.11. The residuals are seen to increase with n_i in Figure 6.11 and, consequently, the assumption of homoscedasticity seems to be violated. This is not surprising, since the injury incidents may behave as a Poisson variable which has a variance proportional to its mean. To ensure the assumption of homoscedasticity, we make the square root transformation. Instead of working with Y we work with $\sqrt{Y}$, a variate which has an approximate variance of 0.25, and is more normally distributed than the original variable.

Consequently, the model we fit is

$$\sqrt{y_i} = \beta_0' + \beta_1' n_i + \varepsilon_i. \tag{6.8}$$

The result of fitting (6.8) is given in Table 6.8. The residuals from (6.8) when plotted against n_i are shown in Figure 6.12. The residuals for the transformed model do not seem to increase with n_i. This suggests that for the transformed

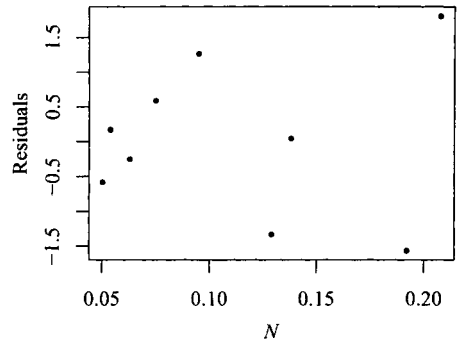

Figure 6.11 Plot of the standardized residuals versus N.

Table 6.8 Estimated Regression Coefficients When $\sqrt{y_i}$ Is Regressed on n_i

Variable	Coefficient	s.e.	t-test	p-value
Constant	1.169	0.578	2.02	0.0829
N	11.856	4.638	2.56	0.0378
	$n = 9$	$R^2 = 0.483$	$\hat{\sigma} = 0.773$	$d.f. = 7$

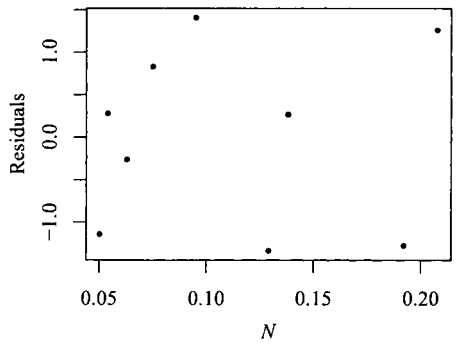

Figure 6.12 Plot of the standardized residuals from the regression of $\sqrt{y_i}$ on n_i.

model the homoscedastic assumption is not violated. The analysis of the model in terms of $\sqrt{y_i}$ and n_i can now proceed using standard techniques. The regression is significant here (as judged by the t statistic) but is not very strong. Only 48% of the total variability of the injury incidents of the airlines is explained by the variation in their number of flights. It appears that for a better explanation of injury incidents other factors have to be considered.

Table 6.9 Number of Supervised Workers and Supervisors in 27 Industrial Establishments

Row	X	Y	Row	X	Y	Row	X	Y
1	294	30	10	697	78	19	700	106
2	247	32	11	688	80	20	850	128
3	267	37	12	630	84	21	980	130
4	358	44	13	709	88	22	1025	160
5	423	47	14	627	97	23	1021	97
6	311	49	15	615	100	24	1200	180
7	450	56	16	999	109	25	1250	112
8	534	62	17	1022	114	26	1500	210
9	438	68	18	1015	117	27	1650	135

In the preceding example the nature of the response variable (injury incidents) suggested that the error variance was not constant about the fitted line. The square root transformation was considered based on the well-established empirical fact that the occurrence of accidents tend to follow the Poisson probability distribution. For Poisson variables, the square root is the appropriate transformation (Table 6.5). There are situations, however, when the error variance is not constant and there is no a priori reason to suspect that this would be the case. Empirical analysis will reveal the problem, and by making an appropriate transformation this effect can be eliminated. If the unequal error variance is not detected and eliminated, the resulting estimates will have large standard errors, but will be unbiased. This will have the effect of producing wide confidence intervals for the parameters, and tests with low sensitivity. We illustrate the method of analysis for a model with this type of heteroscedasticity in the next example.

6.5 DETECTION OF HETEROSCEDASTIC ERRORS

In a study of 27 industrial establishments of varying size, the number of supervised workers (X) and the number of supervisors (Y) were recorded (Table 6.9). The data can also be found in the book's Web site. It was decided to study the relationship between the two variables, and as a start a linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{6.9}$$

was postulated. A plot of Y versus X suggests a simple linear model as a starting point (Figure 6.13). The results of fitting the linear model are given in Table 6.10.

The plot of the standardized residuals versus X (Figure 6.14) shows that the residual variance tends to increase with X. The residuals tend to lie in a band that diverges as one moves along the X axis. In general, if the band within which the residuals lie diverges (i.e., becomes wider) as X increases, the error variance is also increasing with X. On the other hand, if the band converges (i.e., becomes

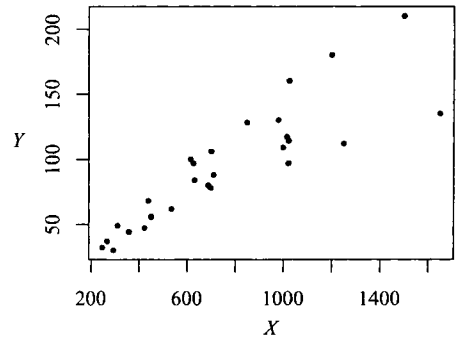

Figure 6.13 Number of supervisors (Y) versus number supervised (X).

Table 6.10 Estimated Regression Coefficients When Number of Supervisors (Y) Is Regressed on the Number Supervised (X)

Variable	Coefficient	s.e.	t-test	p-value
Constant	14.448	9.562	1.51	0.1350
X	0.105	0.011	9.30	< 0.0001
	$n = 27$	$R^2 = 0.776$	$\hat{\sigma} = 21.73$	$d.f. = 25$

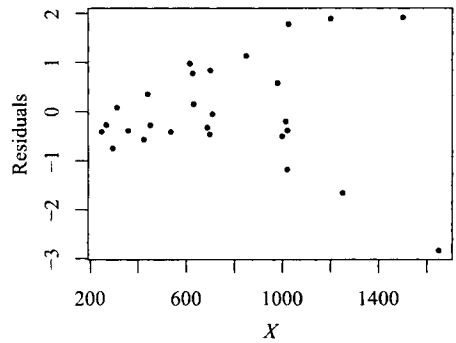

Figure 6.14 Plot of the standardized residuals against X when number of supervisors (Y) is regressed on the number supervised (X).

narrower), the error variance decreases with X. If the band that contains the residual plots consists of two lines parallel to the X axis, there is no evidence of heteroscedasticity. A plot of the standardized residuals against the predictor variable points up the presence of heteroscedastic errors. As can be seen in Figure 6.14, in our present example the residuals tend to increase with X.

6.6 REMOVAL OF HETEROSCEDASTICITY

In many industrial, economic, and biological applications, when unequal error variances are encountered, it is often found that the standard deviation of residuals tends to increase as the predictor variable increases. Based on this empirical observation, we will hypothesize in the present example that the standard deviation of the residuals is proportional to X (some indication of this is available from the plot of the residuals in Figure 6.14):

$$Var(\varepsilon_i) = k^2 x_i^2, \quad k > 0. \tag{6.10}$$

Dividing both sides of (6.9) by x_i, we obtain

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}\,. \tag{6.11}$$

Now, define a new set of variables and coefficients,

$$Y' = \frac{Y}{X},\ X' = \frac{1}{X},\ \beta_0' = \beta_1,\ \beta_1' = \beta_0,\ \varepsilon' = \frac{\varepsilon}{X}\,.$$

In terms of the new variables (6.11) reduces to

$$y_i' = \beta_0' + \beta_1' x_i' + \varepsilon_i'. \tag{6.12}$$

Note that for the transformed model, $Var(\varepsilon_i')$ is constant and equals k^2. If our assumption about the error term as given in (6.10) holds, to fit the model properly we must work with the transformed variables: Y/X and $1/X$ as response and predictor variables, respectively. If the fitted model for the transformed data is $\hat{\beta}_0' + \hat{\beta}_1'/X$, the fitted model in terms of the original variables is

$$\hat{Y} = \hat{\beta}_1' + \hat{\beta}_0' X. \tag{6.13}$$

The constant in the transformed model is the regression coefficient of X in the original model, and vice versa. This can be seen from comparing (6.11) and (6.12).

The residuals obtained after fitting the transformed model are plotted against the predictor variable in Figure 6.15. It is seen that the residuals are randomly distributed and lie roughly within a band parallel to the horizontal axis. There is no marked evidence of heteroscedasticity in the transformed model. The distribution of residuals shows no distinct pattern and we conclude that the transformed model is adequate. Our assumption about the error term appears to be correct; the transformed model has homoscedastic errors and the standard assumptions of least squares theory hold. The result of fitting Y/X and $1/X$ leads to estimates of β_0' and β_1' which can be used for the original model.

The equation for the transformed variables is $Y/X = 0.121 + 3.803/X$. In terms of the original variables, we have $\hat{Y} = 3.803 + 0.121X$. The results are summarized in Table 6.11. By comparing Tables 6.10 and 6.11 we see the reduction in standard errors that is accomplished by working with transformed variables. The variance of the estimate of the slope is reduced by 33%.

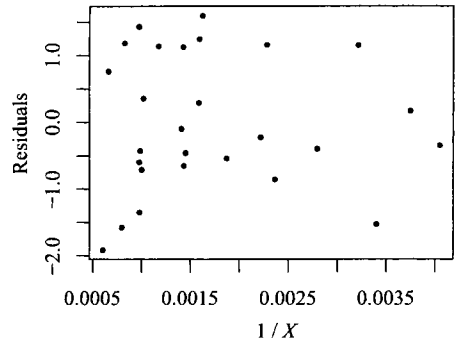

Figure 6.15 Plot of the standardized residuals against $1/X$ when Y/X is regressed on $1/X$.

Table 6.11 Estimated Regression Coefficients of the Original Equation When Fitted by the Transformed Variables Y/X and $1/X$

Variable	Coefficient	s.e.	t-test	p-value
Constant	0.121	0.009	13.44	< 0.0001
$1/X$	3.803	4.570	0.832	0.4131
	$n = 27$	$R^2 = 0.758$	$\hat{\sigma} = 22.577$	$d.f. = 25$

6.7 WEIGHTED LEAST SQUARES

Linear regression models with heteroscedastic errors can also be fitted by a method called the *weighted least squares* (WLS), where parameter estimates are obtained by minimizing a weighted sum of squares of residuals where the weights are inversely proportional to the variance of the errors. This is in contrast to ordinary least squares (OLS), where the parameter estimates are obtained by minimizing equally weighted sum of squares of residuals. In the preceding example, the WLS estimates are obtained by minimizing

$$\sum \frac{1}{x_i^2}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{6.14}$$

as opposed to minimizing

$$\sum (y_i - \beta_0 - \beta_1 x_i)^2. \tag{6.15}$$

It can be shown that WLS is equivalent to performing OLS on the transformed variables Y/X and $1/X$. We leave this as an exercise for the reader.

Weighted least squares as an estimation method is discussed in more detail in Chapter 7.

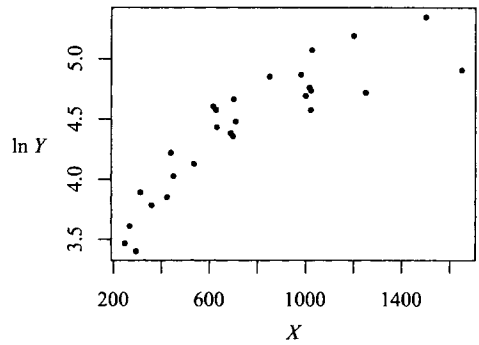

Figure 6.16 Scatter plot of $\ln Y$ versus X.

Table 6.12 Estimated Regression Coefficients When $\ln Y$ Is Regressed on X

Variable	Coefficient	s.e.	t-test	p-value
Constant	3.5150	0.1110	31.65	< 0.0001
X	0.0012	0.0001	9.15	< 0.0001
	$n = 27$	$R^2 = 0.77$	$\hat{\sigma} = 0.252$	$d.f. = 25$

6.8 LOGARITHMIC TRANSFORMATION OF DATA

The logarithmic transformation is one of the most widely used transformations in regression analysis. Instead of working directly with the data, the statistical analysis is carried out on the logarithms of the data. This transformation is particularly useful when the variable analyzed has a large standard deviation compared to its mean. Working with the data on a log scale often has the effect of dampening variability and reducing asymmetry. This transformation is also effective in removing heteroscedasticity. We illustrate this point by using the industrial data given in Table 6.9, where heteroscedasticity has already been detected. Besides illustrating the use of log (logarithmic) transformation to remove heteroscedasticity, we also show in this example that for a given body of data there may exist several adequate descriptions (models).

Instead of fitting the model given in (6.9), we now fit the model

$$\ln y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{6.16}$$

(i.e., instead of regressing Y on X, we regress $\ln Y$ on X). The corresponding scatter plot is given in Figure 6.16. The results of fitting (6.16) are given in Table 6.12. The coefficients are significant, and the value of R^2 (0.77) is comparable to that obtained from fitting the model given in (6.9).

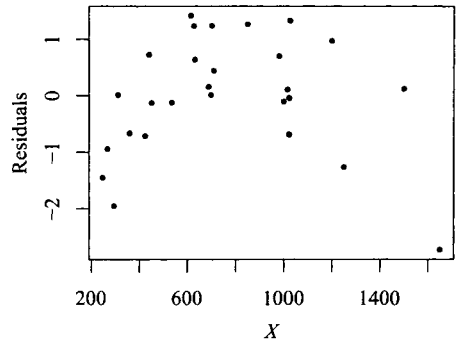

Figure 6.17 Plot of the standardized residuals against X when $\ln Y$ is regressed on X.

The plot of the residuals against X is shown in Figure 6.17. The plot is quite revealing. Heteroscedasticity has been removed, but the plot shows distinct nonlinearity. The residuals display a quadratic effect, suggesting that a more appropriate model for the data may be

$$\ln y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i. \tag{6.17}$$

Equation (6.17) is a multiple regression model because it has two predictor variables, X and X^2. As discussed in Chapter 4, residual plots can also be used in the detection of model deficiencies in multiple regression. To show the effectiveness of residual plots in detecting model deficiencies and their ability to suggest possible corrections, we present the results of fitting model (6.17) in Table 6.13. Plots of the standardized residuals against the fitted values and against each of the predictor variables X and X^2 are presented in Figures 6.18–6.20, respectively.[4]

Residuals from the model containing a quadratic term appear satisfactory. There is no appearance of heteroscedasticity or nonlinearity in the residuals. We now have two equally acceptable models for the same data. The model given in Table 6.13 may be slightly preferred because of the higher value of R^2. The model given in Table 6.11 is, however, easier to interpret since it is based on the original variables.

6.9 POWER TRANSFORMATION

In the previous section we used several types of transformations (such as the reciprocal transformation, $1/Y$, the square root transformation, $\sqrt{Y}$, and the logarithmic transformation, lnY). These transformation have been chosen based on theoretical

[4]Recall from our discussion in Chapter 4 that in simple regression the plots of residuals against fitted values and against the predictor variable X_1 are identical; hence one needs to examine only one of the two plots but not both. In multiple regression the plot of residuals against the fitted values is distinct from the plots of residuals against each of the predictors.

Table 6.13 Estimated Regression Coefficients When $\ln Y$ is Regressed on X and X^2

Variable	Coefficient	s.e.	t-test	p-value
Constant	2.8516	0.1566	18.2	< 0.0001
X	3.11267E−3	0.0004	7.80	< 0.0001
X^2	−1.10226E−6	0.220E−6	−4.93	< 0.0001
	$n = 27$	$R^2 = 0.886$	$\hat{\sigma} = 0.1817$	$d.f. = 24$

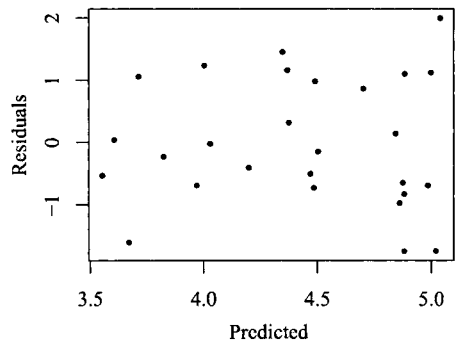

Figure 6.18 Plot of standardized residuals against the fitted values when $\ln Y$ is regressed on X and X^2.

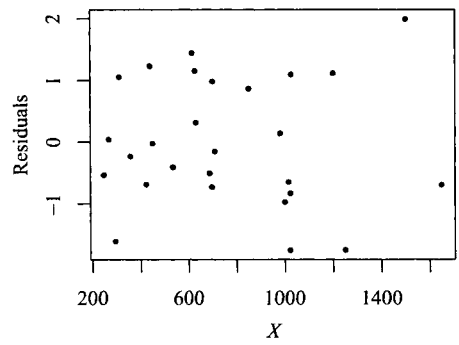

Figure 6.19 Plot of standardized residuals against X when $\ln Y$ is regressed on X and X^2.

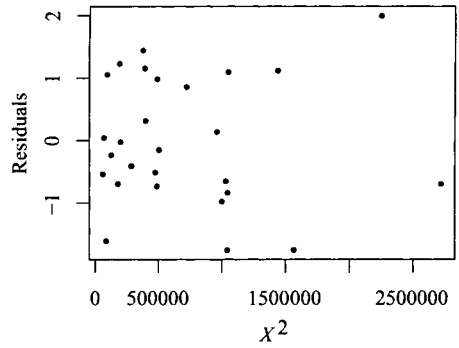

Figure 6.20 Plot of standardized residuals against X^2 when $\ln Y$ is regressed on X and X^2.

or empirical evidence to obtain linearity of the model, to achieve normality, and/or to stabilize the error variance. These transformation can be thought of as a general case of power transformation. In power transformation, we raise the response variable Y and/or some of the predictor variables to a power. For example, instead of using Y we use Y^{λ}, where λ is an exponent to be chosen by the data analyst based on either theoretical or empirical evidence. When $\lambda = -1$ we obtain the reciprocal transformation, $\lambda = 0.5$ gives the square root transformation, and when $\lambda = 0$ we obtain the logarithmic transformation.[5] Values of $\lambda = 1$ implies no transformation is needed.

If λ cannot be determined by theoretical considerations, the data can be used to determine the appropriate value of λ. This can be done using numerical methods. In practice, several values of λ are tried and the best value is chosen. Values of λ commonly tried are: 2, 1.5, 1.0, 0.5, 0, −0.5, −1, −1.5, −2. These values of lambda are chosen because they are easy to interpret. They are known as a *ladder of transformation.* This is illustrated in the following example.

Example: The Brain Data

The data set shown in Table 6.14 represent a sample taken from a larger data set. The data can also be found in the book's Web site. The original sources of the data is Jerison (1973). It has also been analyzed by Rousseeuw and Leroy (1987). The average brain weight (in grams), Y, and the average body weight (in kilograms), X, are measured for 28 animals. One purpose of the data is to determine whether a larger brain is required to govern a heavier body. Another purpose is to see

[5]Note that when $\lambda = 0$, $Y^{\lambda} = 1$ for all values of Y. To avoid this problem the transformation $(Y^{\lambda} - 1)/\lambda$ is used. It can be shown that as λ approaches zero, $(Y^{\lambda} - 1)/\lambda$ approaches lnY. This transformation is known as the Box-Cox power transformation. For more details, see Carroll and Ruppert (1988).

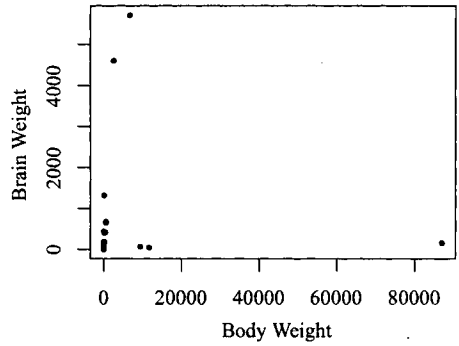

Figure 6.21 The Brain data: Scatter plots of Brain Weight versus Body Weight.

Table 6.14 The Brain Data: Brain Weight (Grams) and Body Weight (Kilograms)

Name	Brain Weight	Body Weight	Name	Brain Weight	Body Weight
Mountain beaver	8.1	1.35	African elephant	5712.0	6654.00
Cow	423.0	465.00	Triceratops	70.0	9400.00
Gray wolf	119.5	36.33	Rhesus monkey	179.0	6.80
Goat	115.0	27.66	Kangaroo	56.0	35.00
Guinea pig	5.5	1.04	Hamster	1.0	0.12
Diplodocus	50.0	11700.00	Mouse	0.4	0.02
Asian elephant	4603.0	2547.00	Rabbit	12.1	2.50
Donkey	419.0	187.10	Sheep	175.0	55.50
Horse	655.0	521.00	Jaguar	157.0	100.00
Potar monkey	115.0	10.00	Chimpanzee	440.0	52.16
Cat	25.6	3.30	Brachiosaurus	154.5	87000.00
Giraffe	680.0	529.00	Rat	1.9.0	0.28
Gorilla	406.0	207.00	Mole	3.0	0.12
Human	1320.0	62.00	Pig	180.0	192.00

whether the ratio of the brain weight to the body weight can be used as a measure of intelligence. The scatter plot of the data (Figure 6.21) does not show an obvious relationship. This is mainly due to the presence of very large animals (e.g., two elephants and three dinosaurs). Let us apply the power transformation to both Y and X. The scatter plots of Y^{λ} versus X^{λ} for several values of λ in the ladder of transformation are given in Figure 6.22. It can be seen that the values of $\lambda = 0$ (corresponding to the log transformation) is the most appropriate value. For $\lambda = 0$, the graphs looks linear but the three dinosaurs do not conform to the linear pattern suggested by the other points. The graph suggests that either the brain weight of the dinosaurs are underestimated and/or their body weight is overestimated.

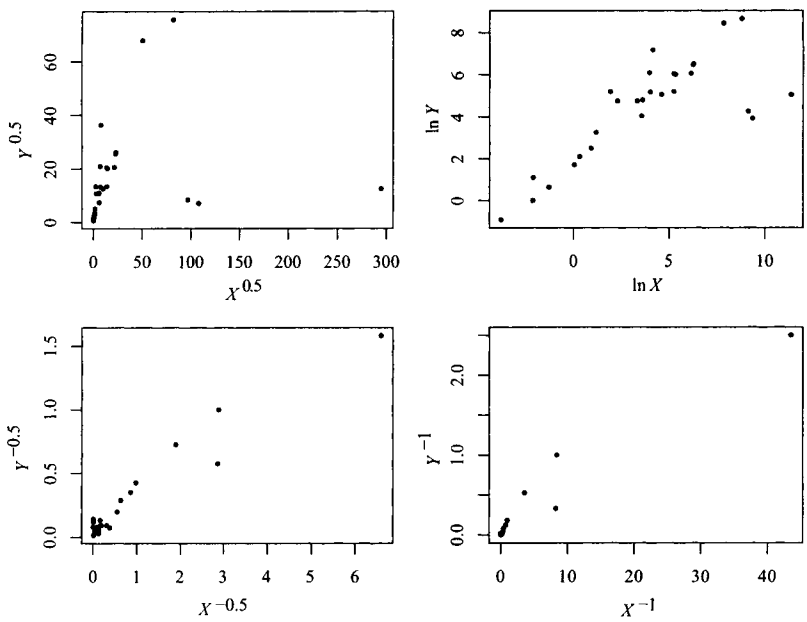

Figure 6.22 Scatter plots of Y^{λ} versus X^{λ} for various values of λ.

Note that in this example we transformed both the response and the predictor variables and that we used the same value of the power for both variables. In other applications, it may be more appropriate to raise each value to a different power and/or to transform only one variable. For further details on data transformation the reader is referred to Carroll and Ruppert (1988) and Atkinson (1985).

6.10 SUMMARY

After fitting a linear model one should examine the residuals for any evidence of heteroscedasticity. Heteroscedasticity is revealed if the residuals tend to increase or decrease with the values of the predictor variable, and is conveniently examined from a plot of the residuals. If heteroscedasticity is present, account should be taken of this in fitting the model. If no account is taken of the unequal error variance, the resulting least squares estimates will not have the maximum precision (smallest variances). Heteroscedasticity can be removed by working with transformed variables. Parameter estimates from the transformed model are then substituted for the appropriate parameters in the original model. The residuals from the appropriately transformed model should show no evidence of heteroscedasticity.

Table 6.15 Advertising Pages (P), in Hundreds, and Advertising Revenue (R), in Millions of Dollars) for 41 Magazines in 1986

Magazine	P	R	Magazine	P	R
Cosmopolitan	25	50.0	Town and Country	1	7.0
Redbook	15	49.7	True Story	77	6.6
Glamour	20	34.0	Brides	13	6.2
Southern Living	17	30.7	Book Digest Magazine	5	5.8
Vogue	23	27.0	W	7	5.1
Sunset	17	26.3	Yankee	13	4.1
House and Garden	14	24.6	Playgirl	4	3.9
New York Magazine	22	16.9	Saturday Review	6	3.9
House Beautiful	12	16.7	New Woman	3	3.5
Mademoiselle	15	14.6	Ms.	6	3.3
Psychology Today	8	13.8	Cuisine	4	3.0
Life Magazine	7	13.2	Mother Earth News	3	2.5
Smithsonian	9	13.1	1001 Decorating Ideas	3	2.3
Rolling Stone	12	10.6	Self	5	2.3
Modern Bride	1	8.8	Decorating & Craft Ideas	4	1.8
Parents	6	8.7	Saturday Evening Post	4	1.5
Architectural Digest	12	8.5	McCall's Needlework and Craft	3	1.3
Harper's Bazaar	9	8.3	Weight Watchers	3	1.3
Apartment Life	7	8.2	High Times	4	1.0
Bon Appetit	9	8.2	Soap Opera Digest	2	0.3
Gourmet	7	7.3			

EXERCISES

6.1 Magazine Advertising: In a study of revenue from advertising, data were collected for 41 magazines in 1986 (Table 6.15). The variables observed are number of pages of advertising and advertising revenue. The names of the magazines are listed.

(a) Fit a linear regression equation relating advertising revenue to advertising pages. Verify that the fit is poor.

(b) Choose an appropriate transformation of the data and fit the model to the transformed data. Evaluate the fit.

(c) You should not be surprised by the presence of a large number of outliers because the magazines are highly heterogeneous and it is unrealistic to expect a single relationship to connect all of them. Delete the outliers and obtain an acceptable regression equation that relates advertising revenue to advertising pages.

6.2 Wind Chill Factor: Table 6.16 gives the effective temperatures (W), which are due to the wind chill effect, for various values of the actual temperatures (T) in still air and windspeed (V). The zero-wind condition is taken as the rate of chilling when one is walking through still air (an apparent wind of four

Table 6.16 Wind Chill Factor (°F) for Various Values of Windspeed, V, in Miles/Hour, and Temperature (°F)

	Actual Air Temperature (T)											
V	50	40	30	20	10	0	−10	−20	−30	−40	−50	−60
5	48	36	27	17	5	−5	−15	−25	−35	−46	−56	−66
10	40	29	18	5	−8	−20	−30	−43	−55	−68	−80	−93
15	35	23	10	−5	−18	−29	−42	−55	−70	−83	−97	−112
20	32	18	4	−10	−23	−34	−50	−64	−79	−94	−108	−121
25	30	15	−1	−15	−28	−38	−55	−72	−88	−105	−118	−130
30	28	13	−5	−18	−33	−44	−60	−76	−92	−109	−124	−134
35	27	11	−6	−20	−35	−48	−65	−80	−96	−113	−130	−137
40	26	10	−7	−21	−37	−52	−68	−83	−100	−117	−135	−140
45	25	9	−8	−22	−39	−54	−70	−86	−103	−120	−139	−143
50	25	8	−9	−23	−40	−55	−72	−88	−105	−123	−142	−145

miles per hour (mph)). The National Weather Service originally published the data; we have compiled it from a publication of the Museum of Science of Boston. The temperatures are measured in degrees Fahrenheit (°F), and the wind-speed in mph.

(a) The data in Table 6.16 are not given in a format suitable for direct application of regression programs. You may need to construct another table containing three columns, one column for each of the variables W, T, and V. This table can be found in the book's Web Site.[6]

(b) Fit a linear relationship between W, T, and V. The pattern of residuals should indicate the inadequacy of the linear model.

(c) After adjusting W for the effect of T (e.g., keeping T fixed), examine the relationship between W and V. Does the relationship between W and V appear linear?

(d) After adjusting W for the effect of V, examine the relationship between W and T. Does the relationship appear linear?

(e) Fit the model

$$W = \beta_0 + \beta_1 T + \beta_2 V + \beta_3 \sqrt{V} + \varepsilon. \tag{6.18}$$

Does the fit of this model appear adequate? The W numbers were produced by the National Weather Service according to the formula (except for rounding errors)

$$W = 0.0817(3.71\sqrt{V} + 5.81 - 0.25V)(T - 91.4) + 91.4. \tag{6.19}$$

Does the formula above give an accurate numerical description of W?

[6] http://www.ilr.cornell.edu/~hadi/RABE4

Table 6.17 Annual World Crude Oil Production in Millions of Barrels (1880–1988)

Year	OIL	Year	OIL	Year	OIL
1880	30	1940	2,150	1972	18,584
1890	77	1945	2,595	1974	20,389
1900	149	1950	3,803	1976	20,188
1905	215	1955	5,626	1978	21,922
1910	328	1960	7,674	1980	21,722
1915	432	1962	8,882	1982	19,411
1920	689	1964	10.310	1984	19,837
1925	1,069	1966	12,016	1986	20,246
1930	1,412	1968	14,104	1988	21,338
1935	1,655	1970	16,690		

(f) Can you suggest a model better than those in (6.18) and (6.19)?

6.3 Refer to the Presidential Election Data in Table 5.17, where the response variable V is the proportion of votes obtained by a presidential candidate in United States. Since the response is a proportion, it has a value between 0 and 1. The transformation $Y = \log(V/(1-V))$ takes the variable V with values between 0 and 1 to a variable Y with values between $-\infty$ to $+\infty$. It is therefore more reasonable to expect that Y satisfies the normality assumption than does V.

(a) Consider fitting the model

$$\begin{aligned} Y &= \beta_0 + \beta_1 \cdot I + \beta_2 \cdot D + \beta_3 \cdot W + \beta_4 \cdot (G \cdot I) \\ &\quad + \beta_5 \cdot P + \beta_6 \cdot N + \varepsilon, \end{aligned} \tag{6.20}$$

which is the same model as in (5.11) but replacing V by Y.

(b) For each of the two models, examine the appropriate residual plots discussed in Chapter 4 to determine which model satisfies the standard assumptions more than the other, the original variable V or the transformed variable Y.

(c) What does the equation in (6.20) imply about the form of the model relating the original variables V in terms of the predictor variables? That is, find the form of the function

$$\begin{aligned} V &= f(\beta_0 + \beta_1 \cdot I + \beta_2 \cdot D + \beta_3 \cdot W + \beta_4 \cdot (G \cdot I) \\ &\quad + \beta_5 \cdot P + \beta_6 \cdot N + \varepsilon). \end{aligned} \tag{6.21}$$

[Hint: This is a nonlinear function referred to as the *logistic function*, which is discussed in Chapter 12.]

6.4 Oil Production Data: The data in Table 6.17 are the annual world crude oil production in millions of barrels for the period 1880–1988. The data are taken from Moore and McCabe (1993), p. 147.

Table 6.18 The Average Price Per Megabyte in Dollars From 1988–1998

Year	Price	Year	Price
1988	11.54	1994	0.705
1989	9.30	1995	0.333
1990	6.86	1996	0.179
1991	5.23	1997	0.101
1992	3.00	1998	0.068
1993	1.46		

Source: Kindly provided by Jim Porter, Disk/Trends in Wired April 1998.

(a) Construct a scatter plot of the oil production variable (OIL) versus Year and observe that the scatter of points on the graph is not linear. In order to fit a linear model to these data, OIL must be transformed.

(b) Construct a scatter plot of log(OIL) versus Year. The scatter of points now follows a straight line from 1880 to 1973. Political turmoil in the oil-producing regions of the Middle East affected patterns of oil production after 1973.

(c) Fit a linear regression of log(OIL) on Year. Assess the goodness of fit of the model.

(d) Construct the index plot of the standardized residuals. This graph shows clearly that one of the standard assumptions is violated. Which one?

6.5 One of the remarkable technological developments in computer industry has been the ability to store information densely on hard disk. The cost of storage has steadily declined. Table 6.18 shows the average price per megabyte in dollars from 1988–1998.

(a) Does a linear time trend describe the data? Define a new variable t by coding 1988 as 1, 1989 as 2, etc.

(b) Fit the model $P_t = P_0 e^{\beta t}$, where P_t is the price in period t. Does this model describe the data?

(c) Introduce an indicator variable which takes the value 0 for the years 1988–1991, and 1 for the remaining years. Fit a model to connecting $\log(P_t)$ with time t, the indicator variable, and the variable created by taking the product of time and the indicator variable. Interpret the coefficients of the fitted model.