CHAPTER THREE NOTES: (Residual Plots and Transformations)

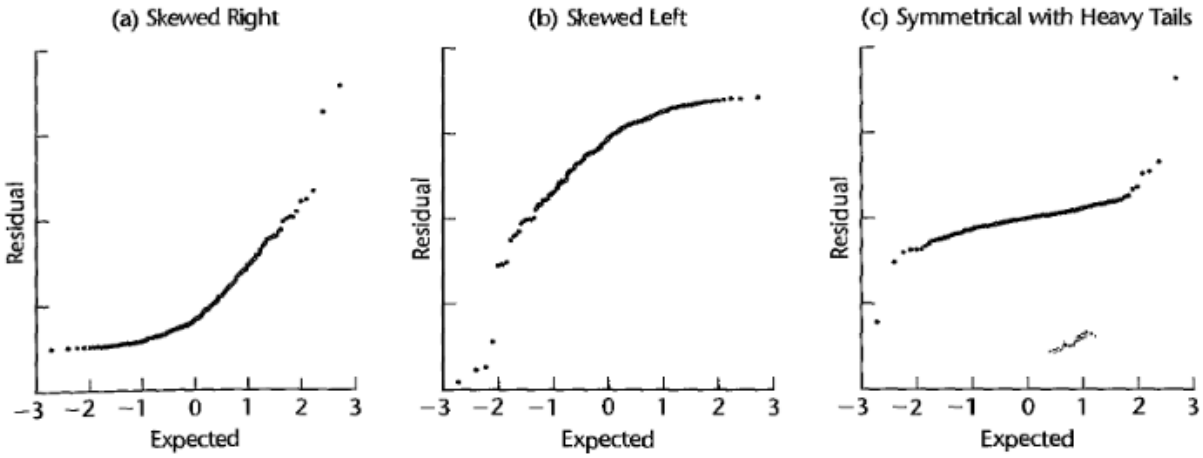
Departures from Model to Be Studied by Residuals

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model (2.1) with normal errors:

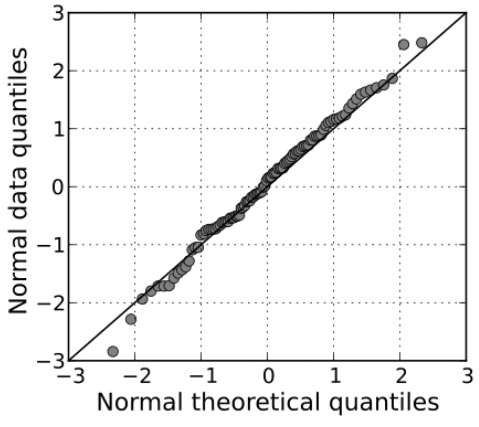
1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outlier observations.
5. The error terms are not normally distributed.
6. One or several important predictor variables have been omitted from the model.

NORMALPROBABILITY PLOT (QQPLOT) ANALYSIS

FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.

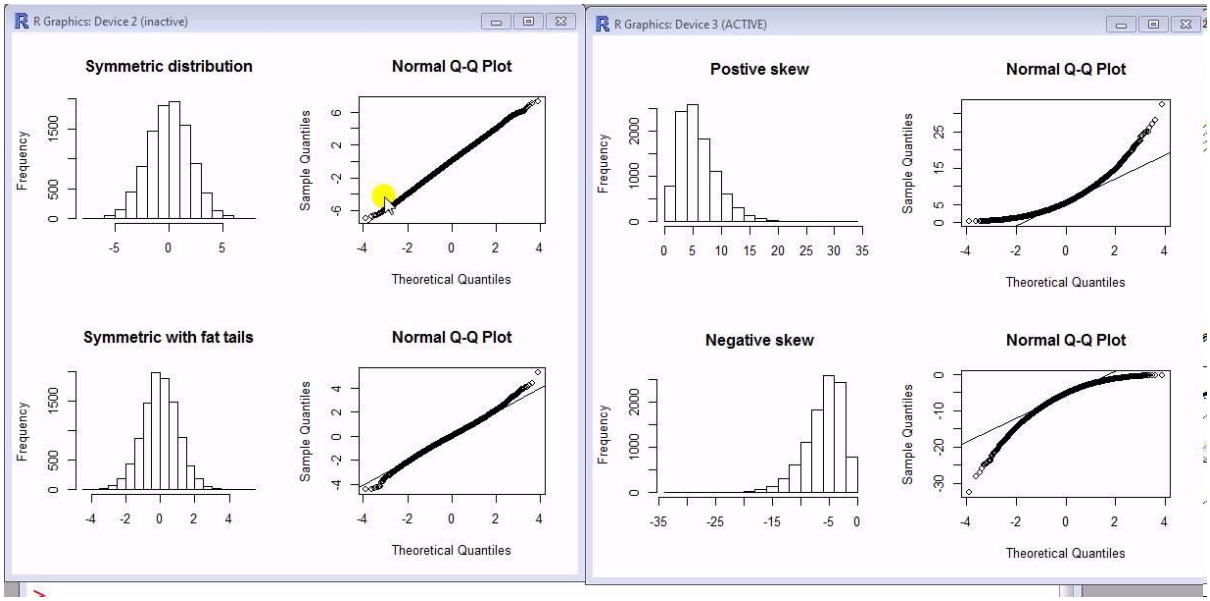


NORMAL PROBABILITY PLOT (Error Terms are normal)



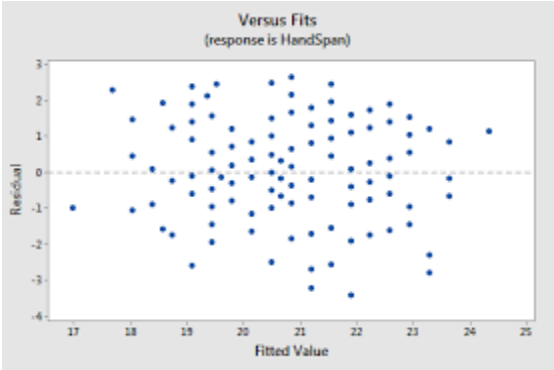
s

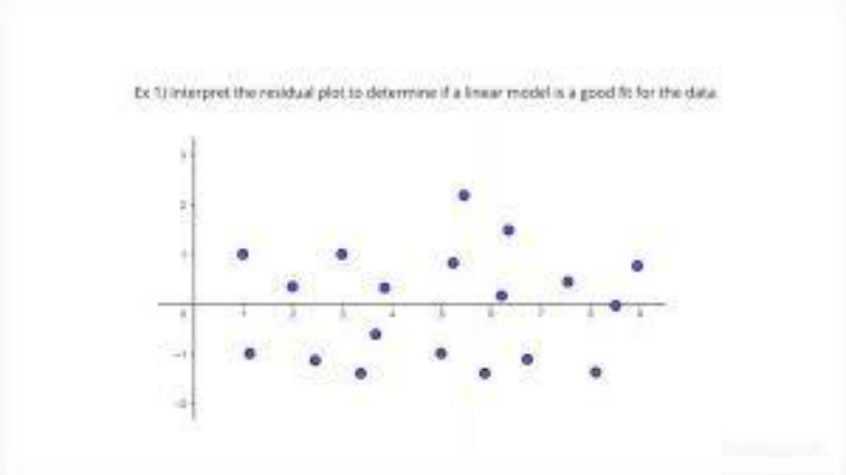
NORMAL PROBABILITY PLOTS (QQ PLOT) HISTOGRAMS

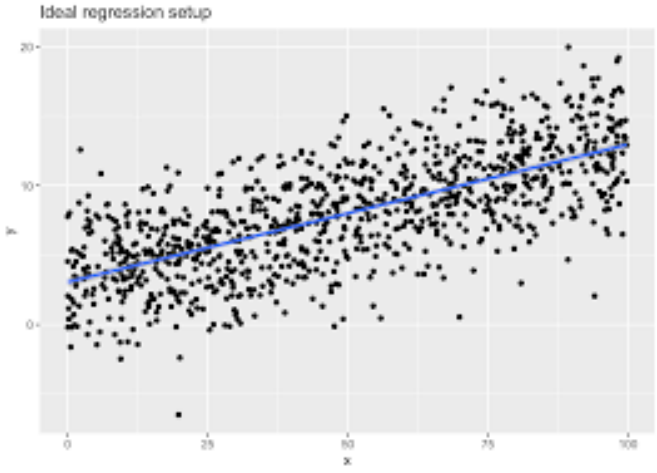


RESIDUAL PLOT ANALYSIS (CHECKING FOR CONSTANT VARIANCE)

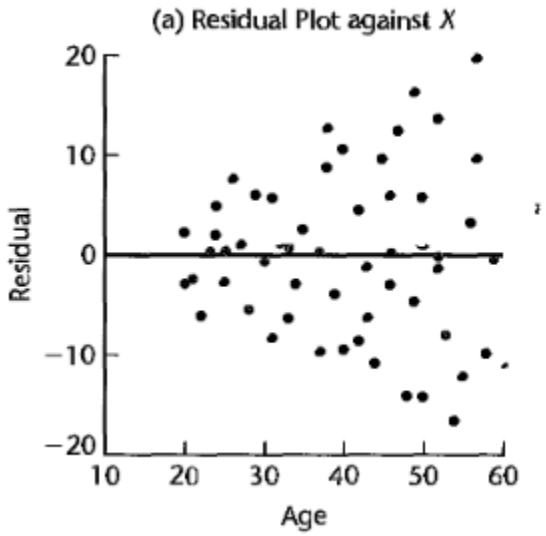
Residual Plots that support linearity (Constant Variance)

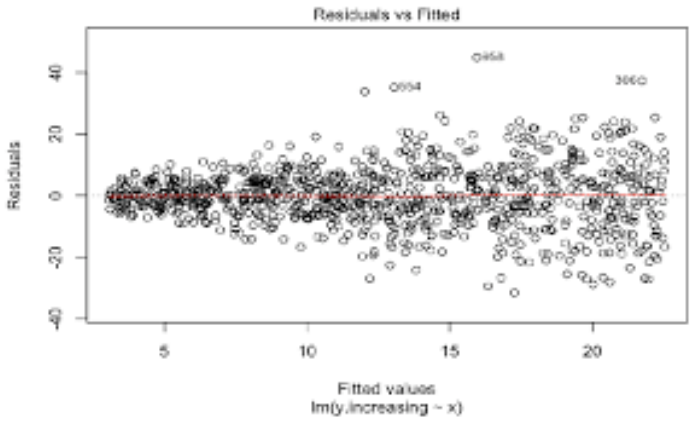


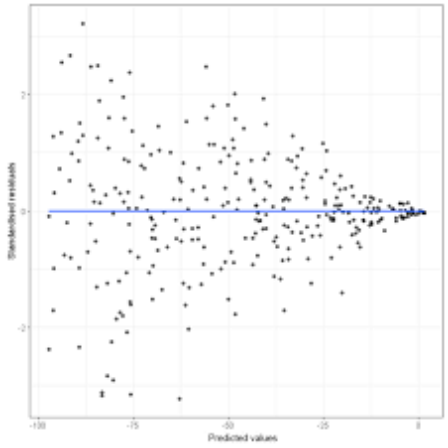




Residual Plots that do not support linearity (Non-Constant Variance)







3.9 Transformations

We now consider in more detail the use of transformations of one or both of the original variables before carrying out the regression analysis. Simple transformations of either the response variable Y or the predictor variable X, or of both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

Figure 3.13 contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y. Several alternative transformations may be tried. Scatter plots and residual plots based on each transformation should then be prepared and analyzed to decide which transformation is most effective.

Example

Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study. A scatter plot of these data is shown in Figure 3.14a. Clearly the regression relation appears to be curvilinear, so the simple linear regression model (2.1) does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on the prototype plot in Figure 3.13a, we shall consider initially the square root transformation $X' = \sqrt{X}$. The transformed values are shown in column 3 of Table 3.7.

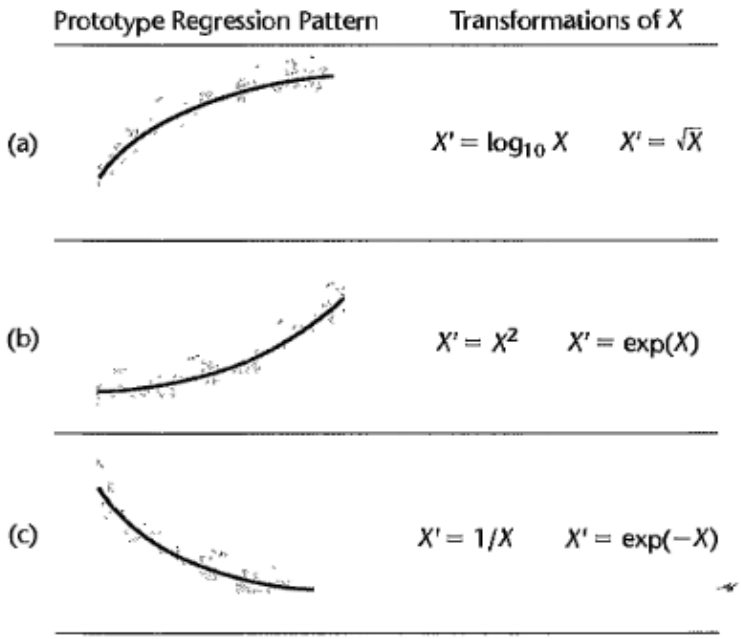


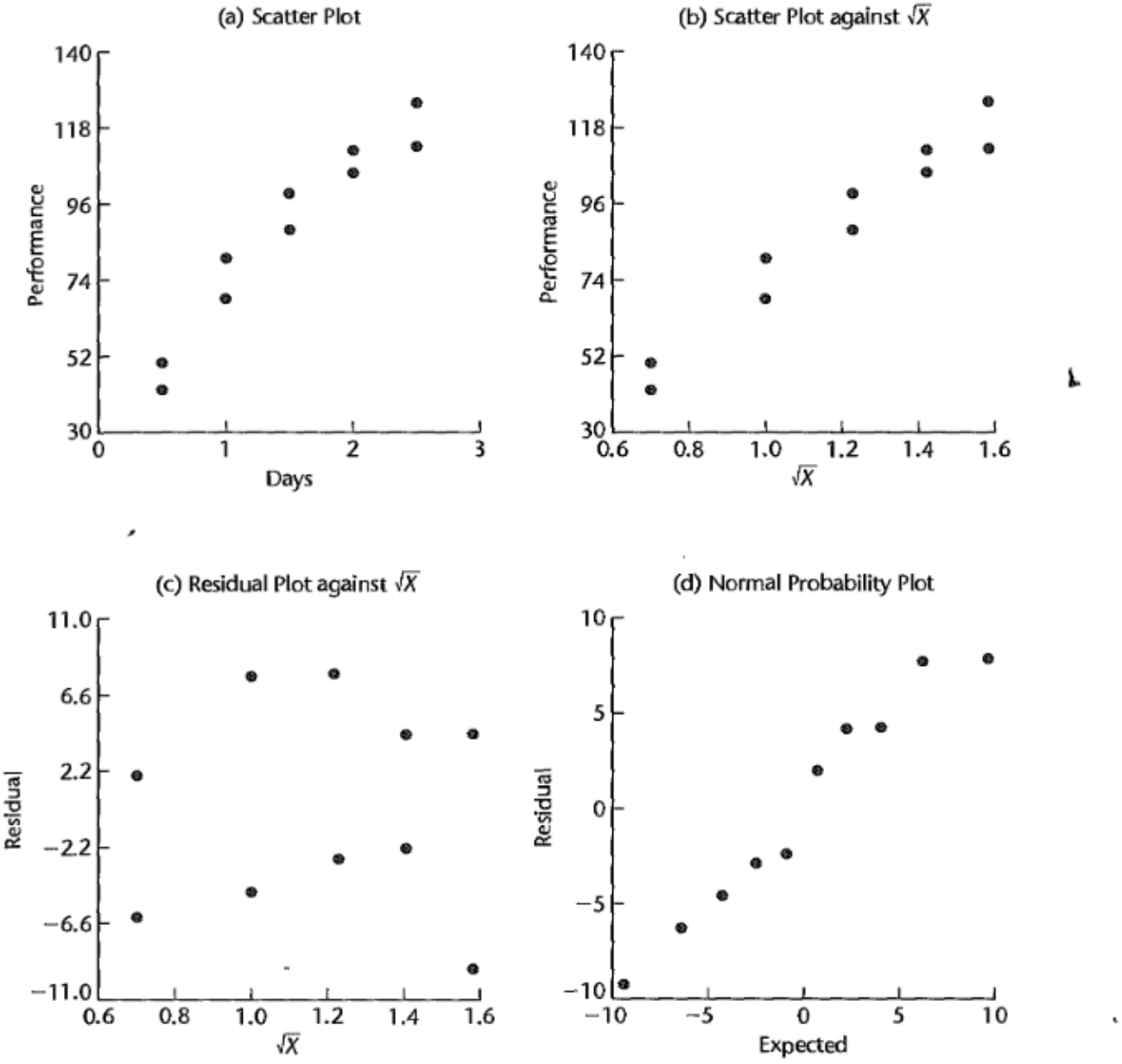
FIGURE 3.13 **Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.**

TABLE 3.7 **Use of Square Root Transformation of X to Linearize Regression Relation—Sales Training Example.**

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

FIGURE 3.14 Scatter Plots and Residual Plots—Sales Training Example.



Transformations for Nonnormality and Unequal Error Variances

Unequal error variances and nonnormality of the error terms frequently appear together. To remedy these departures from the simple linear regression model (2.1), we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response $E\{Y\}$ increases. For example, in a regression of yearly household expenditures for vacations (Y) on household income (X), there will tend to be more variation and greater positive skewness (i.e., some very high yearly vacation expenditures) for high-income households than for low-income households, who tend to consistently spend much less for vacations. Figure 3.15 contains some prototype regression relations where the skewness and the error variance increase with the mean response $E\{Y\}$. This figure also presents some simple transformations on Y that may be helpful for these cases. Several alternative transformations on Y may be tried, as well as some simultaneous transformations on X. Scatter plots and residual plots should be prepared to determine the most effective transformation(s).

most effective transformation(s).

Example

Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study are presented in columns 1 and 2 of Table 3.8. These data are plotted in Figure 3.16a as a scatter plot. Note the distinct curvilinear regression relationship, as well as the greater variability for younger children than for older ones.

TABLE 3.8 Use of Logarithmic Transformation of Y to Linearize Regression Relation and Stabilize Error Variance—Plasma Levels Example.

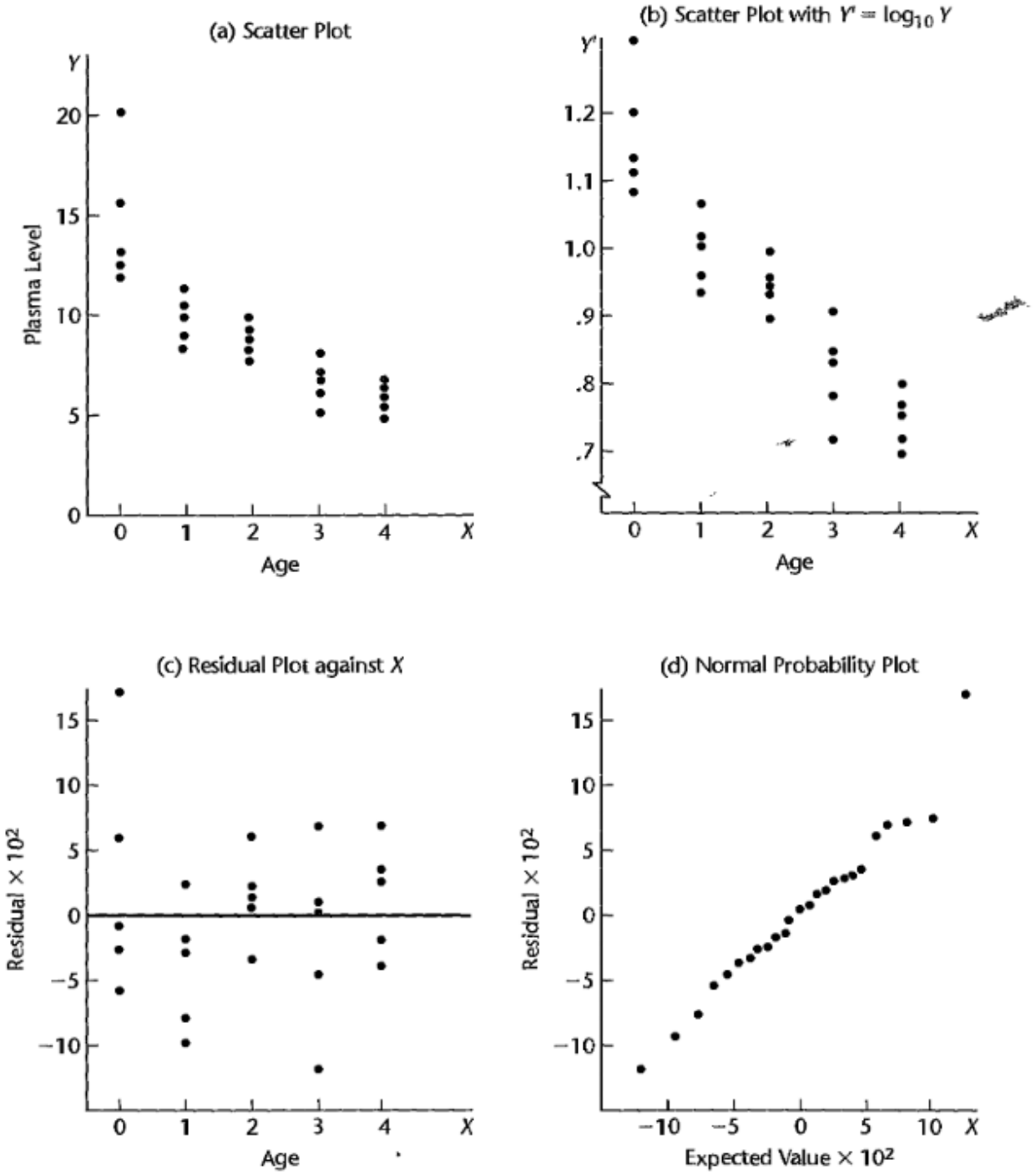
Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

On the basis of the prototype regression pattern in Figure 3.15b, we shall first try the logarithmic transformation $Y' = \log_{10} Y$. The transformed Y values are shown in column 3 of Table 3.8. Figure 3.16b contains the scatter plot with this transformation. Note that the transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant.

To further examine the reasonableness of the transformation $Y' = \log_{10} Y$, we fitted the simple linear regression model (2.1) to the transformed Y data and obtained:

$$\hat{Y}' = 1.135 - .1023X$$

FIGURE 3.16 Scatter Plots and Residual Plots—Plasma Levels Example.



Box-Cox Transformations

It is often difficult to determine from diagnostic plots, such as the one in Figure 3.16a for the plasma levels example, which transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variances, and nonlinearity of the regression function. The Box-Cox procedure (Ref. 3.9) automatically identifies a transformation from the family of power transformations on Y. The family of power transformations

is of the form:

$$Y' = Y^{\lambda} \tag{3.33}$$

where λ is a parameter to be determined from the data. Note that this family encompasses the following simple transformations:

$$\begin{array}{lll} \lambda = 2 & Y' = Y^2 & \\ \lambda = .5 & Y' = \sqrt{Y} & \\ \lambda = 0 & Y' = \log_e Y \quad \text{(by definition)} & \\ \lambda = -.5 & Y' = \dfrac{1}{\sqrt{Y}} & \\ \lambda = -1.0 & Y' = \dfrac{1}{Y} & \end{array} \tag{3.34}$$

The normal error regression model with the response variable a member of the family of power transformations in (3.33) becomes:

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.35}$$

EXAMPLE:

Install required package

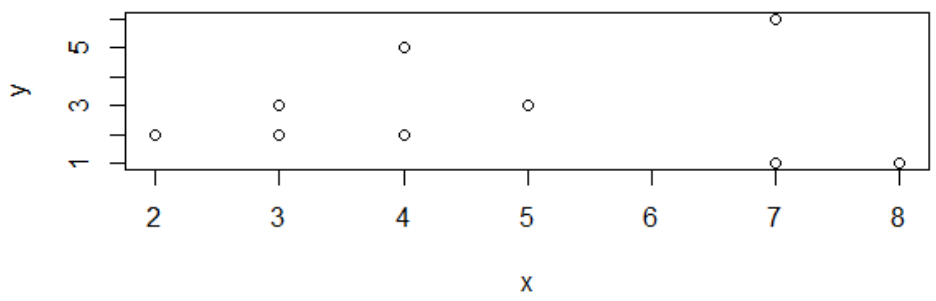
```
library(MASS)
```

Generate Data

```
y <- c(1, 1, 2, 2, 2, 2, 3, 3, 5, 6)  # dependent variable
y
x <- c(8, 7, 3, 2, 3, 4, 5, 3, 4, 7)  # independent variable
x
```

Examine the Scatter Plot

```
plot(y ~ x)
```



Create a linear regression mode

```
model <- lm(y~x)
model
```

Coefficients:

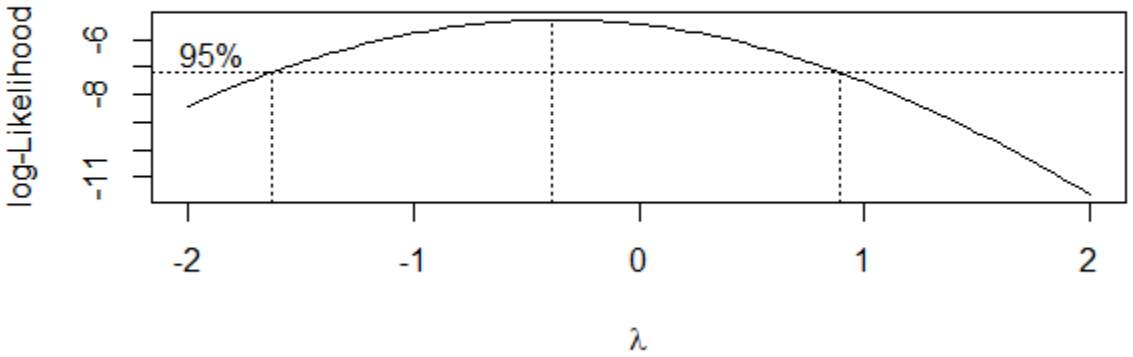
(Intercept) x

2.60417 0.02083

Use the boxcox()

Find optimal lambda for Box-Cox transformation

```
box_cox  <- boxcox(y ~ x)
box_cox
```




```
lambda <- box_cox$x[which.max(box_cox$y)]
lambda
```

-0.3838384

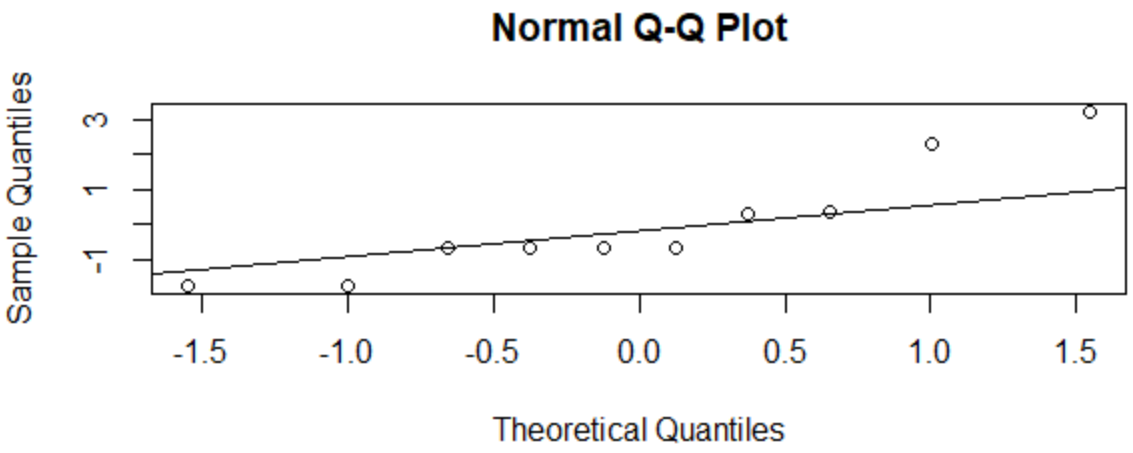
Now, fit new linear regression model using the Box-Cox transformation

```
new_model <- lm(((y^lambda-1)/lambda) ~ x)
new_model
```

Plot the old and new model Normal Probability Plots (QQ Plots)

#Q-Q plot for original model

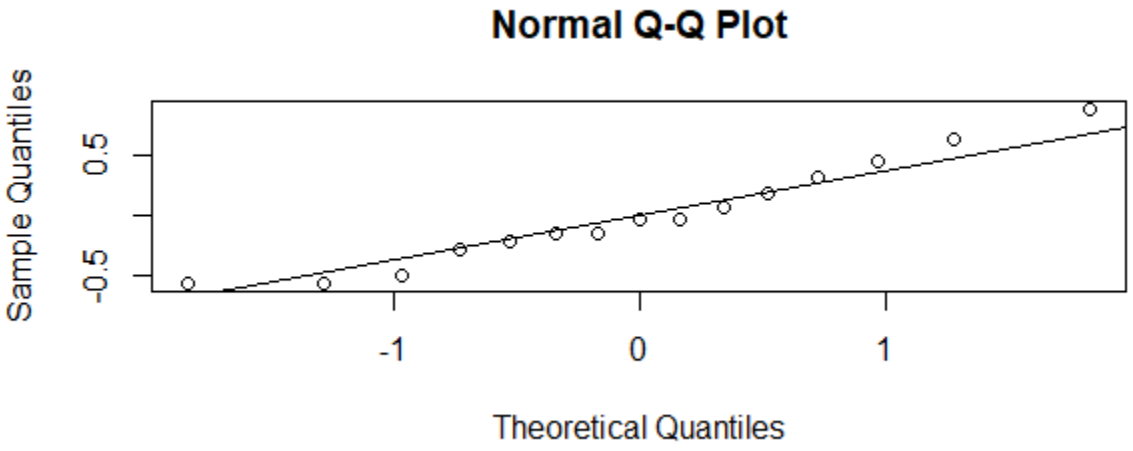
```
qqnorm(model$residuals)
qqline(model$residuals)
```



#Q-Q plot for Box-Cox transformed model

qqnorm(new_model$residuals)

qqline(new_model$residuals)



If in the Q-Q plot the data points fall in a straight line, the data points are said to follow normality. The new model produces a Q-Q plot which has a straighter line compared to the original plot.

CHAPTER 4

when repeated samples are selected and the specified confidence intervals for the entire family are calculated for each sample. Thus, a family confidence coefficient corresponds to the probability, in advance of sampling, that the entire family of statements will be correct.

To illustrate the meaning of a family confidence coefficient further, consider again the joint estimation of β_0 and β_1. A family confidence coefficient of, say, .95 would indicate here that if repeated samples are selected and interval estimates for both β_0 and β_1 are calculated for each sample by specified procedures, 95 percent of the samples would lead to a family of estimates where *both* confidence intervals are correct. For 5 percent of the samples, either one or both of the interval estimates would be incorrect.

A procedure that provides a family confidence coefficient when estimating both β_0 and β_1 is often highly desirable since it permits the analyst to weave the two separate results together into an integrated set of conclusions, with an assurance that the entire set of estimates is correct. We now discuss one procedure for constructing simultaneous confidence intervals for β_0 and β_1 with a specified family confidence coefficient—the Bonferroni procedure.

Bonferroni Joint Confidence Intervals

The Bonferroni procedure for developing joint confidence intervals for β_0 and β_1 with a specified family confidence coefficient is very simple: each statement confidence coefficient is adjusted to be higher than $1 - \alpha$ so that the family confidence coefficient is at least $1 - \alpha$. The procedure is a general one that can be applied in many cases, as we shall see, not just for the joint estimation of β_0 and β_1.

We start with ordinary confidence limits for β_0 and β_1 with statement confidence coefficients $1 - \alpha$ each. These limits are:

$$b_0 \pm t(1 - \alpha/2; n - 2)s\{b_0\}$$
$$b_1 \pm t(1 - \alpha/2; n - 2)s\{b_1\}$$

$$P(\bar{A}_1 \cap \bar{A}_2) \geq 1 - \alpha - \alpha = 1 - 2\alpha \qquad \textbf{(4.2a)}$$

Thus, if β_0 and β_1 are separately estimated with, say, 95 percent confidence intervals, the Bonferroni inequality guarantees us a family confidence coefficient of at least 90 percent that both intervals based on the same sample are correct.

We can easily use the Bonferroni inequality (4.2a) to obtain a family confidence coefficient of at least $1-\alpha$ for estimating β_0 and β_1. We do this by estimating β_0 and β_1 separately with statement confidence coefficients of $1 - \alpha/2$ each. This yields the Bonferroni bound $1-\alpha/2-\alpha/2 = 1-\alpha$. Thus, the $1-\alpha$ family confidence limits for β_0 and β_1 for regression model (2.1) by the Bonferroni procedure are:

$$b_0 \pm Bs\{b_0\} \qquad b_1 \pm Bs\{b_1\} \qquad \textbf{(4.3)}$$

where:

$$B = t(1 - \alpha/4; n - 2) \qquad \textbf{(4.3a)}$$

and b_0, b_1, $s\{b_0\}$, and $s\{b_1\}$ are defined in (1.10), (2.9), and (2.23). Note that a statement confidence coefficient of $1 - \alpha/2$ requires the $(1 - \alpha/4)100$ percentile of the t distribution for a two-sided confidence interval.

mple

For the Toluca Company example, 90 percent family confidence intervals for β_0 and β_1 require $B = t(1 - .10/4; 23) = t(.975; 23) = 2.069$. We have from Chapter 2:

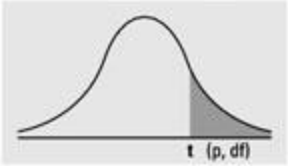
$$b_0 = 62.37 \qquad s\{b_0\} = 26.18$$
$$b_1 = 3.5702 \qquad s\{b_1\} = .3470$$

Hence, the respective confidence limits for β_0 and β_1 are $62.37 \pm 2.069(26.18)$ and $3.5702 \pm 2.069(.3470)$, and the joint confidence intervals are:

$$8.20 \leq \beta_0 \leq 116.5$$
$$2.85 \leq \beta_1 \leq 4.29$$

Thus, we conclude that β_0 is between 8.20 and 116.5 *and* β_1 is between 2.85 and 4.29. The family confidence coefficient is at least .90 that the procedure leads to correct pairs of interval estimates.

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%