Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 3: Diagnostics and Remedial Measures

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. The features of the model, such as _____ of the regression function or _____ of the error terms, may not be appropriate for the particular data.
2. It is important to examine the aptness of the model for the data before _____ based on that model are undertaken.
3. Use some simple _____ methods to study the appropriateness of a model, as well as some _____.
4. Consider some _____ techniques that can be helpful when the data are not in accordance with the conditions of regression model (2.1).

3.1 Diagnostics for Predictor Variable

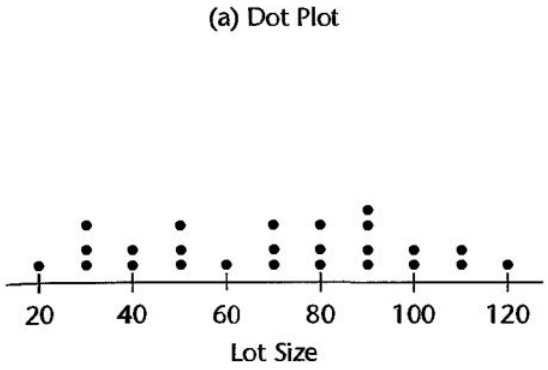
1. Diagnostic for the predictor variable to see if there are any _____ that could influence the appropriateness of the fitted regression function.
2. **Example**: Toluca Company Example
   (a) (Figure 3.1a) _____: the minimum and maximum lot sizes are 20 and 120, respectively, that the lot size levels are spread throughout this interval, and that there are no lot sizes that are far _____.

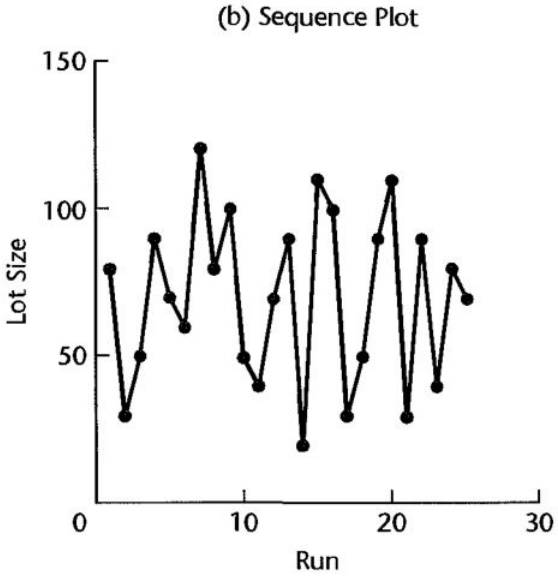
(b) (Figure 3.1b) _____: lot size is plotted against production run (i.e., against time sequence). The plot had shown that smaller lot sizes had been utilized early on and larger lot sizes later on.

(c) (Figure 3.1c) _____: provides information similar to a frequency _____. The letter M denotes the median, and the letter H denotes the first and third quartiles.

(d) (Figure 3.1d) _____: the middle half of the lot sizes range from 50 to 90, and that they are fairly _____ distributed because the median is located in the middle of the central box.

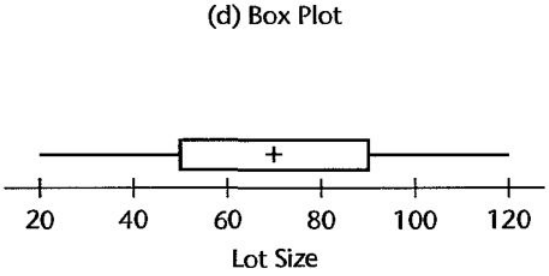
FIGURE 3.1 MINITAB and SYGRAPH Diagnostic Plots for Predictor Variable—Toluca Company Example.





(c) Stem-and-Leaf Plot

2	0
3	000
4	00
5H	000
6	0
7M	000
8	000
9H	0000
10	00
11	00
12	0



3.2 Residuals

1. Diagnostics for the response variable are usually carried out indirectly through an examination of the _____.
2. The residual e_i is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$: _____.
3. The residual may be regarded as the _____, in distinction to the unknown true error ϵ_i in the regression model: _____.
4. For regression model (2.1), the error terms ϵ_i are assumed to be _____ _____ random variables, with mean _____ and constant variance _____·If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ϵ_i.

Properties of Residuals

1. **Mean**
   (a) The mean of the n residuals e_i for the simple linear regression model (2.1) is always 0: _____.
   (b) It provides _____ as to whether the true errors ϵ_i have expected value _____.
2. **Variance**
   (a) The variance of the n residuals e_i for regression model is
   $$s^2 = \_\_\_\_\_ .$$
   (b) If the model is appropriate, MSE is an _____ estimator of the variance of the error terms σ^2.
3. **Nonindependence**
   (a) The residuals e_i are _____ random variables because they involve the fitted values $\hat{Y}_i$ which are based on the _____ fitted regression function.

(b) The residuals for regression model (2.1) are subject to two constraints. These are constraint (1. 17) - _____ - and constraint (1.l9) - _____.

(c) When the _____ is large in comparison to the number of _____ in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be _____ for most purposes.

Semistudentized Residuals

1. Since the standard deviation of the error terms ϵ_i is σ, which is estimated by _____, it is natural to consider the _____ residuals:

$$e_i^* = \frac{\qquad\qquad}{\qquad\qquad}$$

2. Both semistudentized residuals and studentized residuals can be very helpful in identifying _____ observations. (details in Chapter 10)

Departures from Model to Be Studied by Residuals

1. 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:

   (a) The regression function is not _____.

   (b) The error terms do not have _____.

   (c) The error terms are not _____.

   (d) The model fits all but one or a few _____ observations.

   (e) The error terms are not _____ distributed.

   (f) One or several _____ variables have been omitted from the model.

3.3 Diagnostics for Residuals

1. Some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model (2.1)

(a) Plot of residuals against _____ variable.

(b) Plot of _____ or _____ residuals against predictor variable.

(c) Plot of residuals against _____. (the most important)

(d) Plot of residuals against _____ or other sequence.

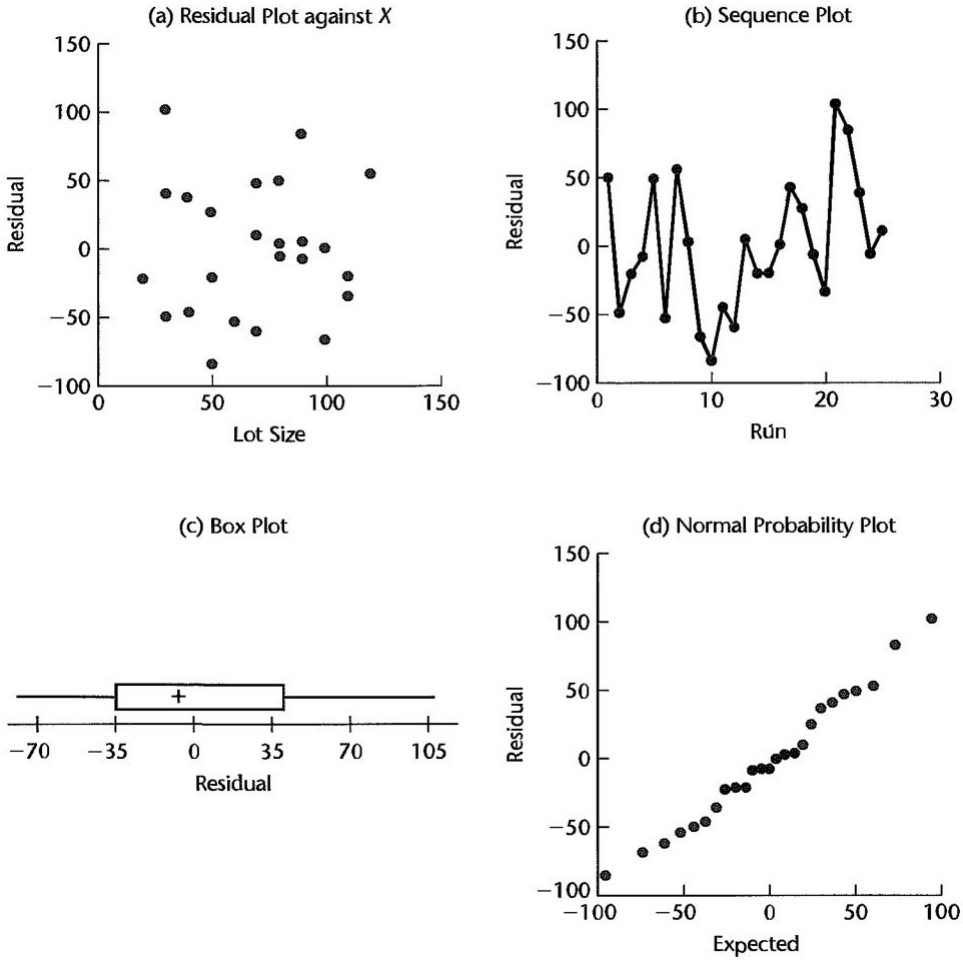
(e) Plots of residuals against _____ variables.

(f) Box plot of residuals.

(g) _____ of residuals.

2. (Figure 3.2) Toluca Company example: plots of the residuals against the predictor variable and against time, a box plot, and a normal probability plot.

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots—Toluca Company Example.



Nonlinearity of Regression Function

1. **Residual plot**: whether a linear regression function is appropriate for the data being analyzed can be studied from a _____ against the _____.

2. Nonlinearity of the regression function can also be studied from a _____, but this plot is not always as effective as a residual plot.

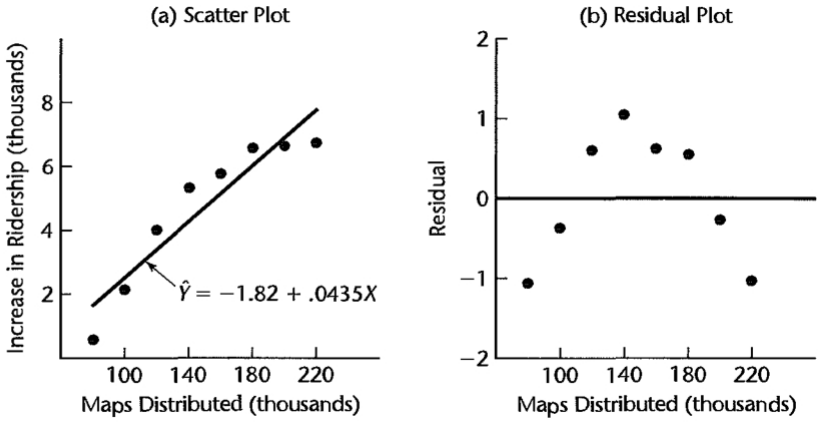
3. Example Ridership - Transit Example (Figure 3.3)(TABLE 3.1)

   (a) One would like to study the relation between maps distributed and bus ridership in eight test cities. Let X be the number of bus transit maps distributed free to residents of the city at the beginning of the test period and Y be the increase during the test period in average daily bus ridership during nonpeak hours.

   (b) (Figures 3.3) the lack of linearity of the regression function.

   (c) In general, the residual plot is to be preferred. It can clearly show any _____ in the deviations around the fitted regression line.

FIGURE 3.3 Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.



4. (Figure 3.4a) the residual plot against X when a linear regression model is _____. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative.

5. (Figure 3.4b) a departure from the linear regression model that indicates the need for a _____ regression function. Here the residuals tend to vary in a systematic fashion between being _____.

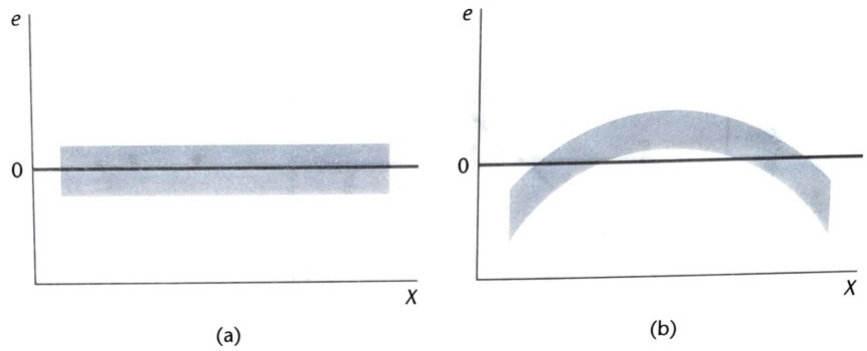


FIGURE 3.4 Prototype Residual Plots.

Nonconstancy of Error Variance

1. The residuals plot is also helpful to examine whether the variance of the error terms is _____.

2. Plots of the _____ values of the residuals or of the _____ residuals against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing _____ of the error variance since the _____ of the residuals are not meaningful for examining the constancy of the error variance.

3. Example Blood Pressure - Age Example

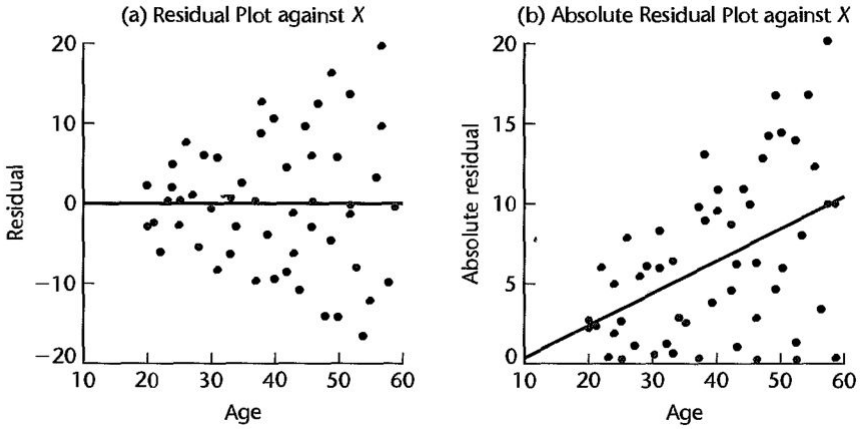
   (a) A study of the relation between diastolic blood pressure of healthy, adult women (Y) and their age (X).

   (b) (Figure 3.5) The residual plot suggests that the older the woman is, the more _____ the residuals are.

   (c) Since the relation between blood pressure and age is positive, this suggests that the error variance is _____ women than for younger ones.

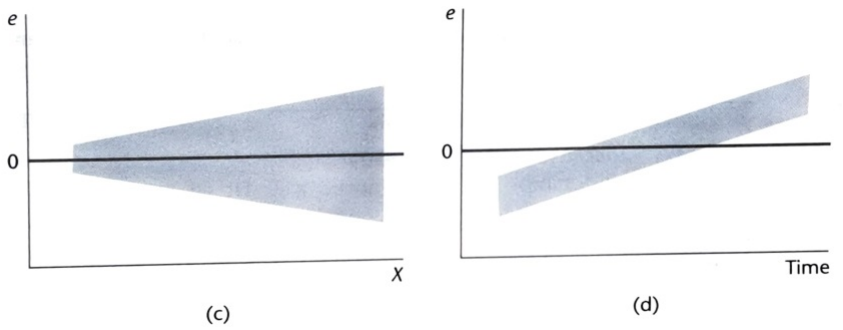
   (d) (Figure 3.5b) a plot of the absolute residuals against age for the blood pressure shows more clearly that the residuals tend to be larger in absolute magnitude for older-aged women.

FIGURE 3.5
Residual Plots Illustrating Nonconstant Error Variance.



4. (Figure 3.4c) a residual plots when the error variance increases with X. One can also encounter error variances _____ with increasing levels of the predictor variable and occasionally varying in some more complex fashion.

FIGURE 3.4
Prototype Residual Plots.



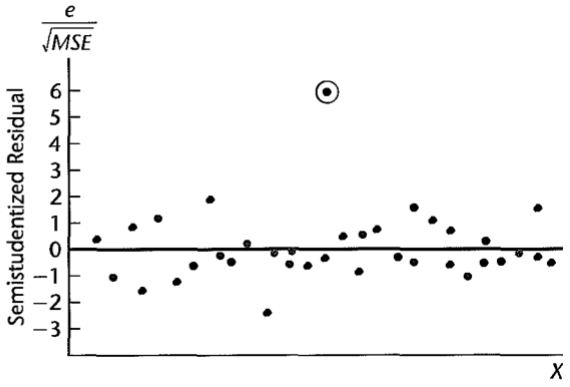
Presence of Outliers

1. Residual _____ (extreme observations) can be identified from residual plots against X or Y, as well as from box plots, stem-and-leaf plots, and dot plots of the residuals.

2. A rough rule of thumb when the number of cases is large is to consider _____ with absolute value of _____ to be outliers. (details in Chapter 10).

3. (Figure 3.6) The residual plot in presents semistudentized residuals and contains one outlier, which is circled.

FIGURE 3.6
Residual Plot with Outlier.



4. How to deal with outliers:

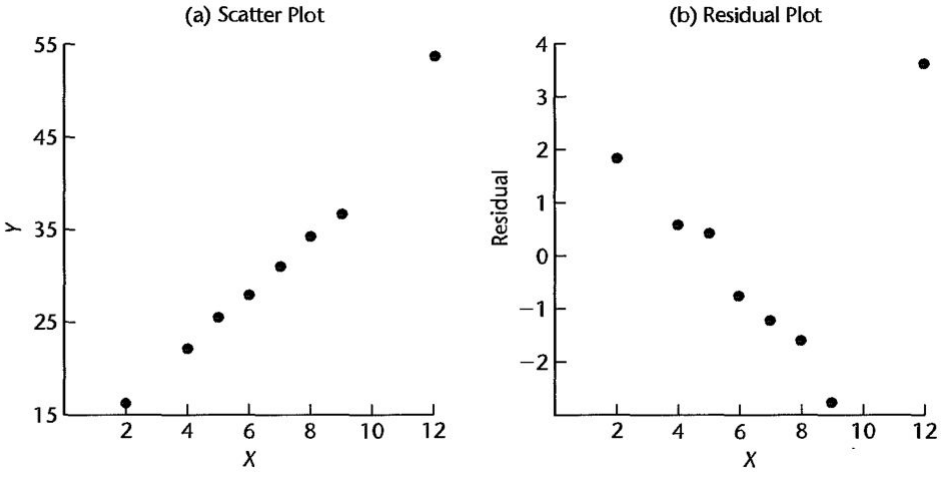
   (a) A safe rule frequently suggested is to _____ only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

   (b) Under the least squares method, a fitted line may be pulled disproportionately _____ an outlying observation because the sum of the squared deviations is minimized.

   (c) This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause.

5. (Figure 3.7) The fitted regression is _____ by the outlier that the residual plot suggest a lack of fit of the linear regression model.

FIGURE 3.7
Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.



Nonindependence of Error Terms

1. **A sequence plot of the residuals**: the purpose of plotting the residuals against time or in some other type of sequence is to see if there is any _____ between error terms that are near each other in the sequence.

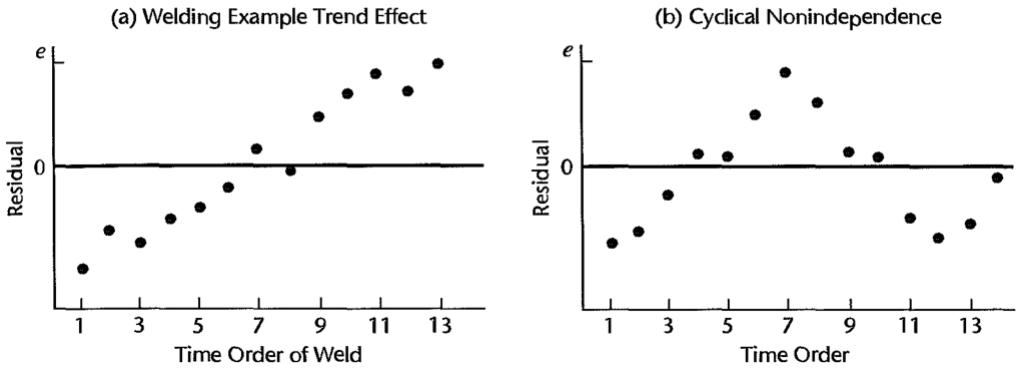
2. Example *Linear Time-related Trend Effect*

   (a) (Figure 3.8a) contains a time sequence plot of the residuals in an experiment to study the relation between the diameter of a weld (X) and the shear strength of the weld (Y).

   (b) An evident correlation between the error terms stands out. _____ residuals are associated mainly with the early trials, and _____ residuals with the later trials.

   (c) It is sometimes useful to view the problem of nonindependence of the error terms as one in which an important variable (in this case, _____) has been omitted from the model.

FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.



3. Example **Cyclical Nonindependent**

   (a) (Figure 3.8b) the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present.

   (b) When the error terms are _____, we expect the residuals in a sequence plot to _____ in a more or less random pattern around the base line 0.

Nonnormality of Error Terms

1. Small departures from normality do not create any serious problems.

2. The normality of the error terms can be studied informally by examining the residuals in a variety of _____ ways.

3. **Distribution Plots** A box plot, histogram, dot plot, or stem-and-leaf plot of the residuals can be helpful for detecting gross departures from normality. Note that the number of cases in the regression study must be _____ for any of these plots to convey reliable information about the _____ of the distribution of the error terms.

4. **Comparison of Frequencies** Another possibility when the number of cases is reasonably large is to compare _____ frequencies of the residuals against _____ frequencies under _____. For example, one can determine whether, say, about 68 percent of the residuals e_i fall between _____ or about 90 percent fall between _____.

5. **Normal Probability Plot of the residuals** Each residual is plotted against its _____ under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

✎ Question ..... (p111)

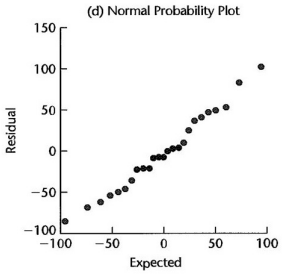
In Toluca Company example, find the expected values of the ordered residuals under normality.

sol:

TABLE 3.2 Residuals and Expected Values under Normality—Toluca Company Example.

Run i	(1) Residual e_i	(2) Rank k	(3) Expected Value under Normality
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots —Toluca Company Example.



5. Three normal probability plots when the distribution of the error terms departs substantially from normality.

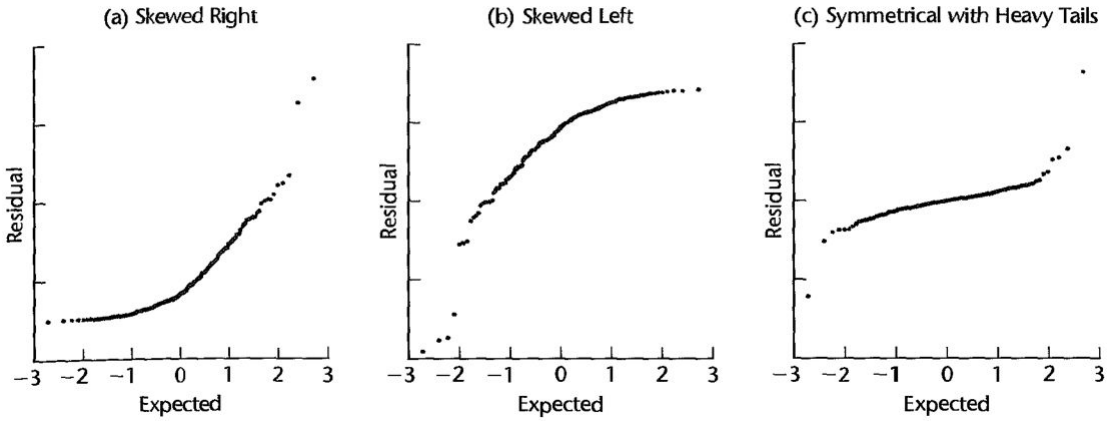
   (a) (Figure 3.9a) shows a normal probability plot when the error term distribution is highly _____. Note the _____ shape of the plot.

   (b) (Figure 3.9b) shows a normal probability plot when the error term distribution is highly _____. Here, the pattern is _____.

   (c) (Figure 3.9c) shows a normal probability plot when the distribution of the error tenus is _____ but has _____; in other words, the distribution has higher probabilities in the tails than a normal distribution.

`https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html`

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



6. **Difficulties in Assessing Normality**

   (a) The analysis for model departures with respect to normality is, in many respects, _____ than that for other types of departures.

   (b) It is usually a good strategy to investigate these other types of departures first, before concerning oneself with the normality of the error terms.

Omission of Important Predictor Variables

1. Residuals should also be plotted against variables omitted from the model that might have important effects on the response.

2. Example One would like to study the relation between output (Y) and age (X) of worker in an assembling operation for a sample of employees. In this study, the machines produced by two companies (A and B) are used in the assembling operation.

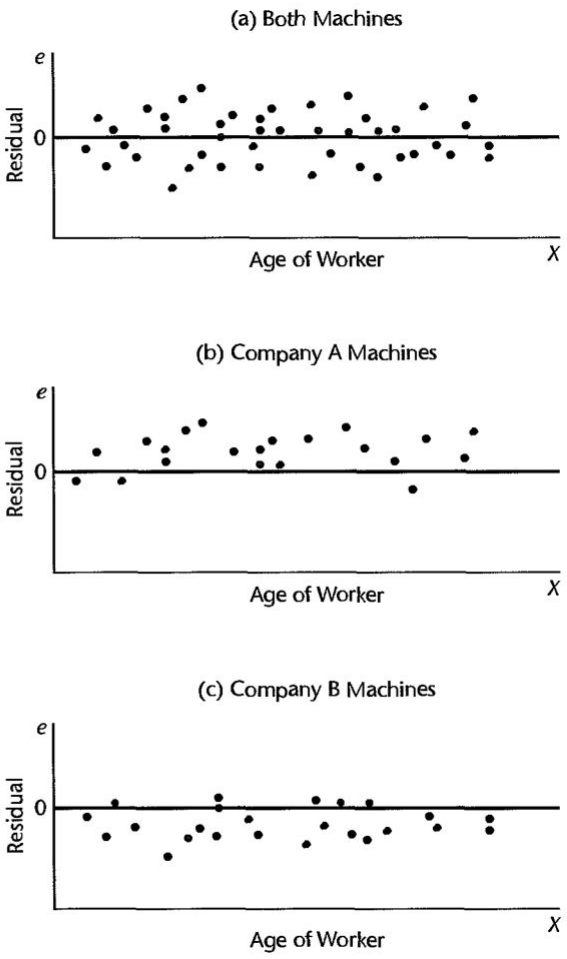


FIGURE 3.10 Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.

(a) (Figure 3.l0a) no ground for suspecting the appropriateness of the linearity of the regression function or the constancy of the error variance.

(b) (Figure 3.l0b, 3.l0c)The residuals for Company A machines tend to be positive: while those for Company B machines tend to be negative.

(c) Type of machine appears to have a definite effect on productivity, and output predictions may turn out to be far superior when this variable is added to the model.

Some Final Comments[1]

1. Several types of departures may occur _____.
2. Although graphic analysis of residuals is only an informal method of analysis, in many cases it _____ for examining the aptness of a model.
3. The basic approach to residual analysis explained here applies not only to simple linear regression but also to more _____ and other types of _____.
4. Model misspecification due to either _____ or the _____ of important predictor variables tends to be serious, leading to _____ estimates of the regression parameters and error variance.
5. _____ of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates.
6. The presence of _____ can be serious for smaller data sets when their influence is large.
7. The _____ of error terms results in estimators that are unbiased but whose variances are seriously _____.

3.4 Overview of Tests Involving Residuals

1. Graphic analysis of residuals is inherently _____.
2. Most statistical tests require independent observations. The residuals are _____. The dependencies become quite small for _____, so that one can usually then ignore them.

Tests for Randomness

1. A _____ is frequently used to test for lack of randomness in the residuals arranged in time order.

[1] Some will be discussed in other Chapters.

2. _____: designed for lack of randomness in least squares residuals. (Chapter 12).

Tests for Constancy of Variance

1. When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to X or $E(Y)$, a simple test is based on the _____ between the absolute values of the residuals and the corresponding values of the predictor variable.

2. Tests for constancy of the error variance: the _____ test and the _____ test (Section 3.6.)

Tests for Outliers

1. A simple test for identifying an outlier observation: detail in (Chapter 10).

2. Many other tests to aid in evaluating outliers have been developed (Reference 3.1.)

Tests for Normality

1. _____ (the chi-square test, the Kolmogorov-Smirnov test and its modification, the Lilliefors test) can be employed for testing the normality of the error terms by analyzing the residuals.

2. A simple test based on the _____ of the residuals (Section 3.5.)

3.5 Correlation Test for Normality

1. A formal test for normality of the error terms can be conducted by calculating the coefficient of _____ between the residuals e_i and their _____ under normality.

2. A high value of the correlation coefficient is indicative of normality.

3. (Table B.6) (Looney and Gulledge) (Ref. 3.2) contains _____ (percentiles) for various sample sizes for the distribution of the coefficient of correlation between the ordered residuals and their expected values under normality when the error terms are normally distributed.

4. If the observed coefficient of correlation is _____ as the tabled value, for a given a level, one can conclude that the error terms are reasonably normally distributed.

5. Example Toluca Company Example: the coefficient of correlation between the ordered residuals and their expected values under normality is _____. Controlling the a risk at _____, we find from Table B.6 that the critical value for $n = 25$ is _____. Since the observed coefficient exceeds this level, we have support for our earlier conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

☺ Normality test: `https://en.wikipedia.org/wiki/Normality_test`.

3.6 Tests for Constancy of Error Variance

Brown-Forsythe Test

1. *Assumption*: the sample size needs to be large enough so that the dependencies among the residuals can be ignored.

2. The Brown-Forsythe test is based on the _____ of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be.

3. The Brown-Forsythe test then consists simply of the _____ based on test statistic (A.67)

   to determine whether the _____ for one group differs significantly from the mean absolute deviation for the second group. Steps:

   (a) Divide the data set into two groups, according to the _____, so that one group consists of cases where the X level is comparatively _____ and the other group consists of cases where the X level is comparatively _____.

(b) If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be _____ than those in the other group.

(c) Equivalently, the _____ of the residuals around their group mean will tend to be larger for one group than for the other group.

(d) In order to make the test more _____, we utilize the absolute deviations of the residuals around the _____ for the group (Ref. 3.5).

4. Although the distribution of the absolute deviations of the residuals is usually _____, it has been shown that the t^* test statistic still follows approximately the _____ when the variance of the error terms is _____ and the sample sizes of the two groups are not extremely small.

5. Notations: the ith residual for group 1 (2) by e_{i1} (e_{i2}), the sample sizes of the two groups by n_1 and n_2, the medians of the residuals in the two groups by $\bar{e}_1$ and $\bar{e}_2$.

6. The Brown-Forsythe test uses the absolute deviations of the residuals around their group _____, to be denoted by d_{i1} and d_{i2}:

_____ and _____

7. The The two-samplet test statistic (called the Brown-Forsythe test statistics t^*_{BF}) becomes:

$$t^*_{BF} = \frac{}{\text{\_\_\_\_\_}}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2} respectively, and the pooled variance $s2$ becomes:

$$s^2 = \frac{}{\text{\_\_\_\_\_}}$$

8. If the error terms have constant variance and n_1 and n_2 are not extremely small, t^*_{BF} follows approximately the _____ distribution with _____ degrees of freedom.

9. Large absolute values of t^*_{BF} indicate that the error terms do not have constant variance.

✎ Question ..... (p117)

Use the Brown-Forsythe test for the Toluca Company example to determine whether or not the error term variance varies with the level of X. (Note that since the X levels are spread fairly uniformly, you can divide the 25 cases into two groups with approximately equal X ranges. The first group consists of the 13 runs with lot sizes from 20 to 70. The second group consists of the 12 runs with lot sizes from 80 to 120. ($\alpha = 0.05, t_{0.975,23} = 2.069$)

sol:

TABLE 3.3 Calculations for Brown-Forsythe Test for Constancy of Error Variance—Toluca Company Example.

Group 1

i	Run	(1) Lot Size	(2) Residual e_{i1}	(3) d_{i1}	(4) $(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	Total			582.60	12,566.6

$\tilde{e}_1 = -19.88 \quad \bar{d}_1 = 44.815$

Group 2

i	Run	(1) Lot Size	(2) Residual e_{i2}	(3) d_{i2}	(4) $(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	Total			341.40	9,610.2

$\tilde{e}_2 = -2.68 \quad \bar{d}_2 = 28.450$

Breusch-Pagan Test*

3.7 F Test for Lack of Fit

Assumptions

1. F test for _____ is a formal test for determining whether a specific type of regression function adequately fits the data.

2. The lack of fit test assumes that the observations Y for given X are (1) _____ and (2) _____ distributed, and that (3) the distributions of Y have the _____.

3. **Replications, Replicates**: the lack of fit test requires _____ observations at one or more X levels. Repeat trials for the same level of the predictor variable, of the type described, are called _____. The resulting observations are called _____.

4. Example Bank Example

   (a) In an experiment involving 12 similar but scattered suburban branch offices of a commercial bank, holders of checking accounts at the offices were offered gifts for setting up money market accounts. Minimum initial deposits in the new money market account were specified to qualify for the gift. The value of the gift was directly proportional to the specified minimum deposit. Various levels of minimum deposit and related gift values were used in the experiment in order to ascertain the relation between the specified minimum deposit and gift value, on the one hand, and number of accounts opened at the office, on the other. Altogether, six levels of minimum deposit and proportional gift value were used, with two of the branch offices assigned at random to each level. One branch office had a fire during the period and was dropped from the study. Table 3.4a contains the results, where X is the amount of minimum deposit and Y is the number of new money market accounts that were opened and qualified for the gift during the test period.

TABLE 3.4 Data and Analysis of Variance Table—Bank Example.

(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

(b) A linear regression function was fitted:

$$\hat{Y} = 50.72251 + 0.48670X$$

(Table 3.4b): The analysis of variance table.

(c) (Figure 3.11) A scatter plot, together with the fitted regression line, indicates that a linear regression function is _____. We use the general linear test approach to do a formal test.

FIGURE 3.11 Scatter Plot and Fitted Regression Line—Bank Example.

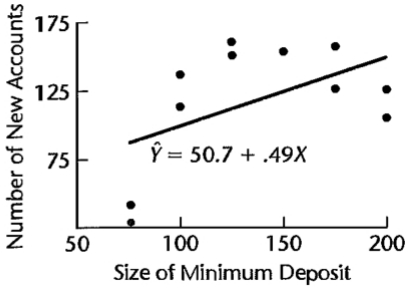


TABLE 3.5 Data Arranged by Replicate Number and Minimum Deposit—Bank Example.

Replicate	Size of Minimum Deposit (dollars) $j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$j=1$	28	112	160	152	156	124
$j=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

Notation

1. (Table 3.5) presents the same data but in an arrangement that recognizes the replicates. We shall denote the different X levels in the study, whether or not replicated observations are present, as $X_1, \cdots, X_c$.

2. There are six minimum deposit size levels in the study ($c = 6$), for five of which there are two observations and for one there is a single observation. We shall let $X_1 = 75$ (the smallest minimum deposit level), $X_2 = 100$, $\cdots$, $X_6 = 200$.

3. Denote the number of replicates for the jth level of X as n_j; for our example, $n_1 = n_2 = n_3 = n_5 = n_6 = 2$ and $n_4 = 1$. Thus, the total number of observations n is given by: $n = \sum_{j=1}^{c} n_j$.

4. Denote the observed value of the response variable for the ith replicate for the jth level of X by Y_{ij}, where $i = 1, \cdots, n_j$, $j = 1, \cdots, c$.

5. (Table 3.5), $Y_{11} = 28, Y_{21} = 42, Y_{12} = 112$, and so on. Denote the mean of the Y observations at the level $X = X_j$ by $\bar{Y}_j$. Thus, $\bar{Y}_1 = (28 + 42)/2 = 35$ and $\bar{Y}_4 = 152/1 = 152$.

Full model

1. The full model used for the lack of fit test makes the _____ as the simple linear regression model (2.1) except for assuming a linear regression relation, the subject of the test.

   _____,

   where μ_j are parameters $j = 1, \cdots, c$, ϵ_{ij} are independent _____.

2. Since the error terms have expectation zero, it follows that:

   $$E(Y_{ij}) = \_\_\_\_\_.$$

   Thus, the parameter μ_j ($j = 1, \cdots, c$) is the mean response when $X = X_j$.

3. The full model states that each response Y is made up of two components: the _____ when $X = X_j$ and a _____ term.

4. The difference between the two models is that in the full model (3.13) there are no restrictions on the _____, whereas in the regression model (2.1) the mean responses are linearly related to X (i.e., _____).

5. The least squares or maximum likelihood estimators for the parameters μ_j: _____.

6. The estimated expected value for observation Y_{ij} is _____ , and the error sum of squares (also called the pure error sum of squares, SSPE) for the full model:

$$SSE(F) = \_\_\_\_\_ = SSPE$$

7. Note that $SSPE$ is made up of the sums of squared deviations _____. At level $X = X_j$, this sum of squared deviations is:

$$\sum_i (Y_{ij} - \bar{Y}_j)^2$$

These sums of squares are then added over all of the X levels ($j = 1, \cdots, c$).

8. Example For the bank example, we have:

$$SSPE = (28 - 35)2 + (42 - 35)^2 + (112 - 124)^2 + \cdots + (104 - 114)^2 = 1,148$$

Note that any X level with no replications makes _____ to $SSPE$ because $\bar{Y}_j = Y_{1j}$ for $j = 4$.

9. The degrees of freedom associated with $SSPE$ can be obtained by recognizing that the sum of squared deviations (3.17) at a given level of X is like an ordinary total sum of squares based on n observations, which has _____ degrees of freedom associated with it. Here, there are n_j observations when $X = X_j$; hence the degrees of freedom are _____.

10. Just as $SSPE$ is the sum of the sums of squares (3.17), so the number of degrees of freedom associated with $SSPE$ is the sum of the component degrees of freedom:

$$df_F = \_\_\_\_\_$$

Reduced Model

1. For testing the appropriateness of a linear regression relation, the alternatives are:

$$H_0 \ : \ \_\_\_\_\_$$
$$H_a \ : \ \_\_\_\_\_$$

Thus, H_0 postulates that μ_j in the full model (3.13) is linearly related to X_j

_____

The reduced model under H_0 therefore is:

_____

2. Note that the reduced model is the ordinary simple linear regression model (2.1), with the subscripts modified to recognize the existence of _____.

3. We know that the estimated expected value for observation Y_{ij} with regression model (2.1) is the fitted value $\hat{Y}_{ij}$

_____

Hence, the error sum of squares for the reduced model is the usual error sum of squares SSE:

$$SSE(R) = \_\_\_\_\_$$

We also know that the degrees of freedom associated with $SSE(R)$ are: _____.

4. Example For the bank example, we have from Table 3.4b: $SSE(R) = SSE = 14741.6,\ df_R = 9$

Test Statistic

1. The general linear test statistic (2.70):

$$F^* = \_\_\_\_\_$$

here becomes:

$$F^* = \_\_\_\_\_$$

2. The difference between the two error sums of squares is called the _____ _____ ($SSLF$):

$$SSLF = \_\_\_\_\_$$

3. We can then express the test statistic as follows:

$$F^* = \_\_\_\_\_$$

where $MSLF$ denotes the lack of fit mean square and $MSPE$ denotes the pure error mean square.

4. We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\text{If } F^* \leq F_{(1-\alpha;c-2,n-c)}, \text{ conclude } H_0$$

$$\text{If } \_\_\_\_\_$$

5. Example For the bank example, the test statistic:

$$\begin{aligned} SSPE &= 1148.0, \quad n-c = 11-6 = 5 \\ SSE &= 14741.6, \\ SSLF &= 14741.6 - 1,148.0 = 13,593.6, \quad c-2 = 6-2 = 4 \\ F^* &= \frac{13,593.6}{4} \div \frac{1148.0}{5} = \frac{3,398.4}{229.6} = 14.80 \end{aligned}$$

If the level of significance is to be $\alpha = 0.01$, we require $F_{(0.99;4,5)} = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. The P-value for the test is 0.006.

ANOVA Table

1. The error deviations in SSE are made up of a pure error component and a lack of fit component: _____.

$$\begin{aligned} Y_{ij} - \hat{Y}_{ij} &= \_\_\_\_\_ \\ \text{Error deviation} &= \_\_\_\_\_ \end{aligned}$$

2. Example (Figure 3.12) illustrates this partitioning for the case $Y_{13} = 160, X_3 = 125$ in the bank example.

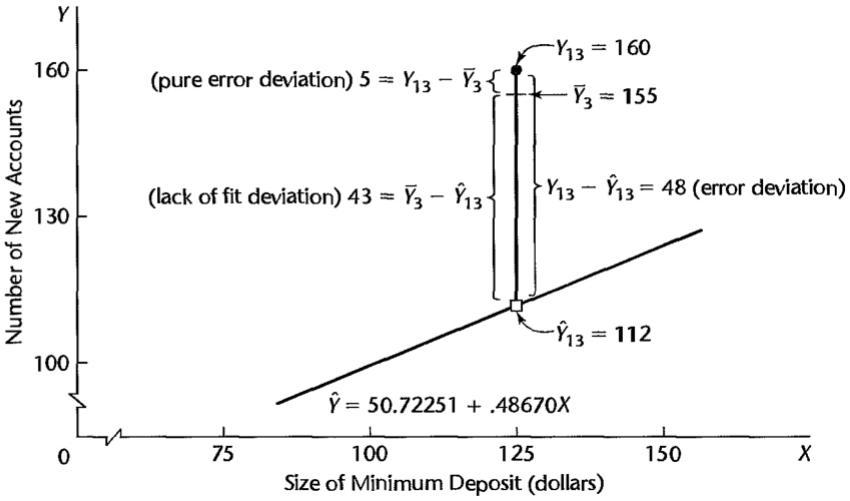


FIGURE 3.12 Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$—Bank Example.

3. When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\begin{aligned} \sum\sum(Y_{ij} - \hat{Y}_{ij})^2 &= \underline{\qquad\qquad\qquad\qquad\qquad\qquad} \\ SSE &= SSPE + SSLF \end{aligned}$$

4. *Why $SSLF$ measures lack of fit?* If the linear regression function is appropriate, then the _____ will be near the _____ calculated from the estimated linear regression function and $SSLF$ will be _____.

5. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function and $SSLF$ will be large.

6. SSLF has $c-2$ degrees of freedom: there are _____ means $\bar{Y}_j$ in the sum of squares, and _____ degrees of freedom are lost in estimating the parameters β_0 and β_1, of the linear regression function to obtain the fitted values $\hat{Y}_j$.

7. (Table 3.6) contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.

(a) General

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. Not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are _____.

2. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$. For the bank example (Table 3Ab), test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5141.3}{1638.0} = 3.14$$

For $\alpha = .10$, $F_{(0.90;1,9)} = 3.36$, and we would _____, that $\beta_1 = 0$ or that there is _____ between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no relation between these variables would be improper, however. Such an inference requires that regression model (2.1) be _____. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of *always examining the appropriateness of a model before any inferences are drawn.*

3. The alternative H_a in (3.19) includes all regression functions other than a _____ one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of _____ can be helpful in identifying an appropriate function.

4. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as _____, and the test for lack of fit is then carried out using these groupings of adjacent cases. (Reference 3.8.)

3.8 Overview of Remedial Measures

1. If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:

   (a) Abandon regression model (2.1) and develop and use a _____.

   (b) Employ some _____ on the data so that regression model (2.1) is appropriate for the transformed data.

Nonlinearity of Regression Function

Section 3.9, Section 3.10. Chapter 7.

Nonconstancy of Error Variance

Section 3.9, Chapter 11.

Nonindependence of Error Terms

Chapter 12.

Nonnormality of Error Terms

Section 3.9.

Omission of Important Predictor Variables

Chapter 6.

Outlying Observations

Chapter 11.

3.9 Transformations

Simple transformations of either the response variable _____ or the predictor variable _____, or of _____, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

1. We first consider transformations for linearizing a nonlinear regression relation when the distribution of the _____ is reasonably close to a _____ distribution and the error terms have approximately _____.

2. In this situation, transformations on _____ should be attempted. Transformation on 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.

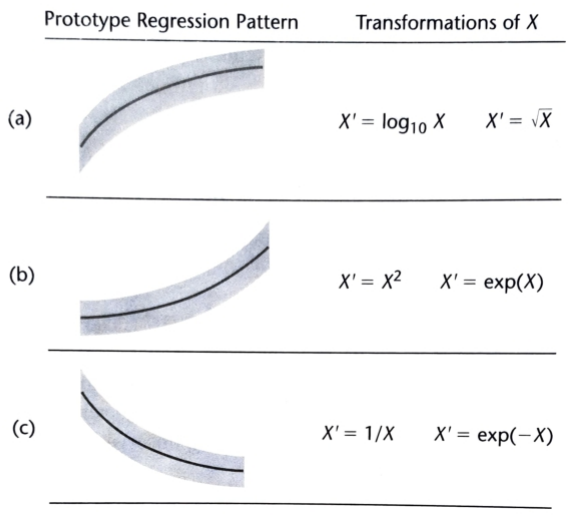


FIGURE 3.13 Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.

3. (Figure 3.13) some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to _____ the regression relationship without affecting the _____.

4. Example A battery of simulated sales

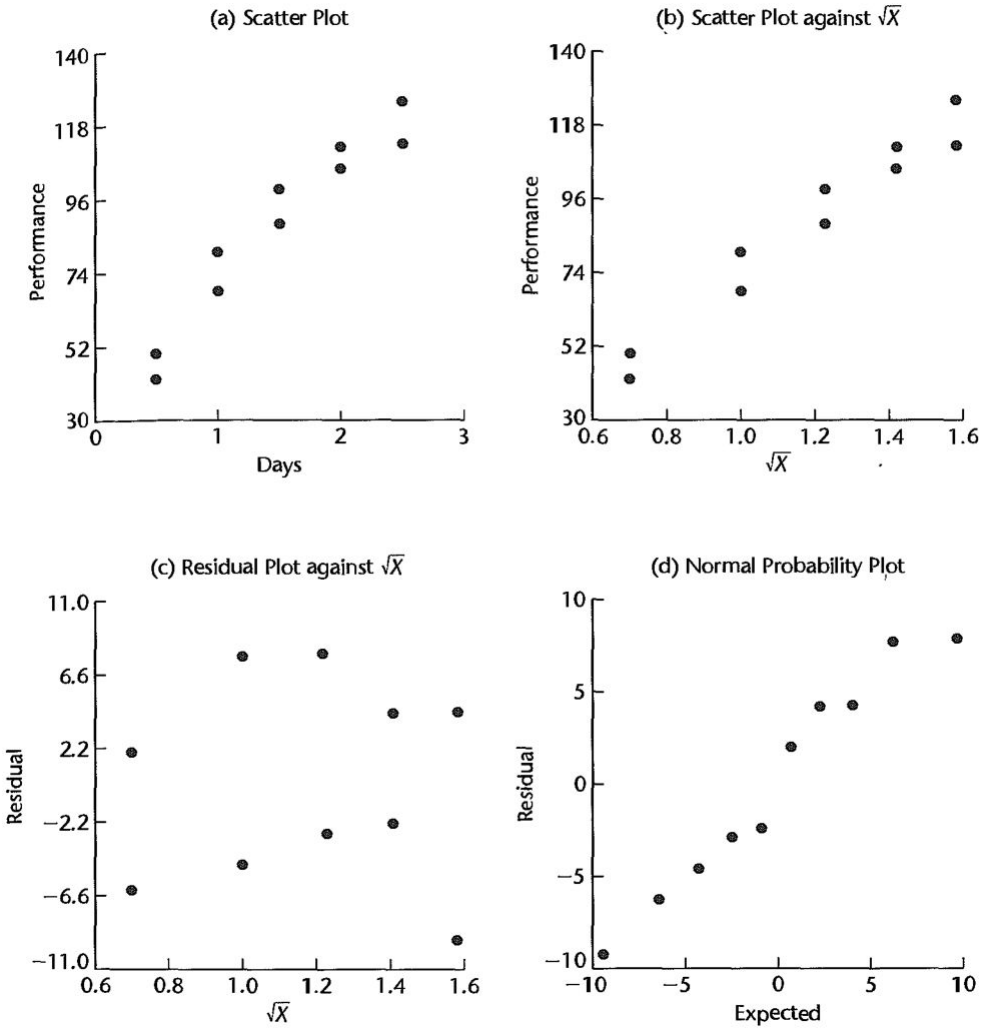
   (a) Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study.

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

   (b) (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 _____ at the different X levels appears to be fairly _____, we shall consider a transformation on X. Based on Figure 3.13a, consider initially the square root transformation _____.

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



(c) (Figure 3.14b), the same data are plotted with the predictor variable transformed to $X' = \sqrt{X}$. Note that the scatter plot now shows a reasonably _____ relation. The variability of the scatter at the different X levels is the same as before, since we did not make a transformation on _____.

(d) To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data:

_____.

(e) (Figure 3.14c) the plot of residuals against X' shows _____ of lack of fit or of strongly unequal error variances.

(f) (Figure 3.14d) a normal probability plot of the residuals. No strong indications of substantial departures from _____. This conclusion is supported by the _____ between the ordered residuals and their expected values under normality, 0.979.

(g) For $\alpha = 0.01$, Table B.6 shows that the critical value is 0.879, so the observed coefficient is substantially larger and supports the reasonableness of normal error terms. Thus, the simple linear regression model (2.1) appears to be appropriate here for the transformed data.

(h) The fitted regression function in the _____ can easily be obtained, if desired:

$$\hat{Y} = \_\_\_\_\_$$

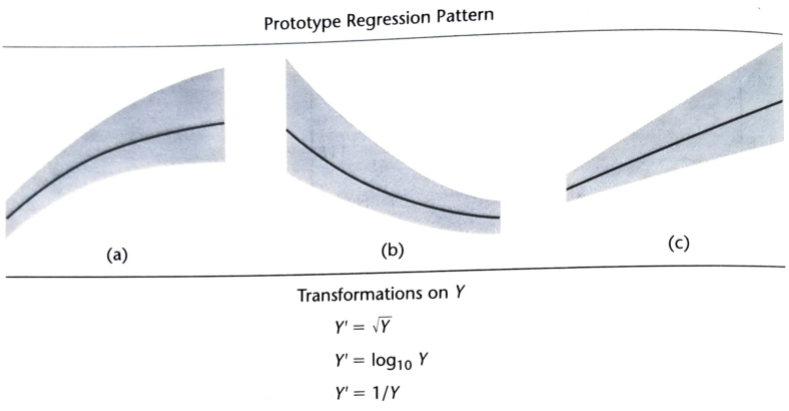
Transformations for Nonnormality and Unequal Error Variances

1. 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 _____, since the _____ and _____ of the distributions of Y need to be changed.

2. A simultaneous _____ may be needed to obtain or maintain a linear regression relation.

3. (Figure 3.15) Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of _____ and _____ of the distributions of the error terms as the mean response $E(Y)$ increases.

FIGURE 3.15 Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of Y.



Note: A simultaneous transformation on X may also be helpful or necessary.

4. _____ and _____ should be prepared to determine the most effective transformations.

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

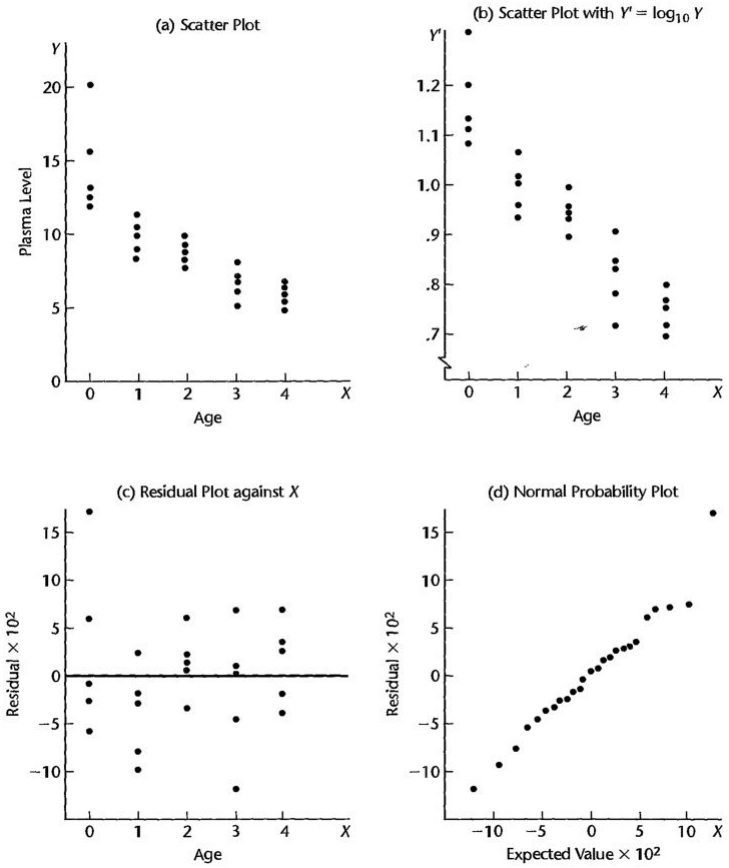
5. Example Plasma Level Example

   (a) (Table 3.8) Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study.

   (b) (Figure 3.16a) a scatter plot shows the distinct _____ regression relationship, as well as the greater variability for younger children than for older ones.

   (c) (Figure 3.16b) the scatter plot of the logarithmic transformation _____. 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 _____.

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



(d) 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:

_____

(e) (Figure 3.16c, d) the evidence supports the appropriateness of regression model (2.1) for the transformed Y data: (i) A plot of the residuals against X, and a normal probability plot of the residuals. (ii) The coefficient of correlation between the ordered residuals and their expected values under normality is _____. (iii) For $\alpha = 0.05$, Table B.6 indicates that the critical value is _____ so that the observed coefficient supports the assumption of normality of the error terms.

(f) NOTE: When Y is negative, the logarithmic transformation to shift the origin in Y and make all Y observations positive would be _____, where k is an appropriately chosen constant.

(g) NOTE: When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient while such a transformation may _____ the error variance, it will also change the linear relationship to a _____ one. A transformation on X may therefore also be required.

Box-Cox Transformations

1. The Box-Cox procedure (Ref. 3.9) automatically identifies a transformation from the family of power transformations on Y. The family of _____ is of the form:

   _____

   where λ is a parameter to be determined from the data.

2. Note that this family encompasses the following simple transformations:

$$\begin{array}{rll} \lambda = 2 & Y' = Y^2 & \\ \lambda = 0.5 & Y' = \sqrt{Y} & \\ \lambda = 0 & \underline{\qquad\qquad} & \text{(by definition)} \\ \lambda = -0.5 & Y' = \dfrac{1}{\sqrt{Y}} & \\ \lambda = -1.0 & Y = \dfrac{1}{Y} & \end{array}$$

☺ Power transform (Box-Cox transformation) - Wikipedia:
`https://en.wikipedia.org/wiki/Power_transform`.

3. The normal error regression model with the response variable a member of the family of power transformations becomes:

$$Y^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i$$

Note that above regression model includes an additional parameter, λ, which needs to be estimated.

4. The Box-Cox procedure uses the method of _____ to estimate λ, as well as the other parameters β_0, β_1, and σ^2.

5. A simple procedure for obtaining $\hat{\lambda}$:

   (a) search in a range of potential λ values; for example, $\lambda = -2, \lambda = -1.75, \cdots, \lambda = 1.75, \lambda = 2$. For each λ value, the Y_i^λ observations are first _____ so that the magnitude of the error sum of squares does not depend on the value of λ.

   (b) Once the standardized observations have been obtained for a given λ value, they are regressed on the predictor variable X - and _____ is obtained.

   (c) It can be shown that the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE is a minimum.

6. After a transformation has been tentatively selected, residual plots and other analyses described earlier need to be employed to ascertain that the simple linear regression model (2.1) is appropriate for the transformed data.

3.10 Exploration of Shape of Regression Function*

lowess Method*

Use of Smoothed Curves to Confirm Fitted Regression Function*

3.11 Case Example – Plutonium Measurement

1. *Background Description*: Some environmental cleanup work requires that nuclear materials, such as plutonium 238 (鈽-238), be located and completely removed from a restoration site. When plutonium has become mixed with other materials in very small amounts, detecting its presence can be a difficult task. Even very small amounts can be traced, however, because plutonium emits subatomic particles – alpha particles – that can be detected. Devices that are used to detect plutonium record the intensity of alpha particle strikes in counts per second (#/sec). The regression relationship between alpha counts per second (the response variable) and

plutonium activity (the explanatory variable) is then used to estimate the activity of plutonium in the material under study.

2. *Data Description*: (Table 3.10) In a study to establish the regression relationship for a particular measurement device, four plutonium standards were used. These standards are aluminum/plutonium rods containing a fixed, known level of plutonium activity. The levels of plutonium activity in the four standards were 0.0, 5.0, 10.0, and 20.0 picocuries per gram (pCi/g). Each standard was exposed to the detection device from 4 to 10 times, and the rate of alpha strikes, measured as counts per second, was observed for each replication.

TABLE 3.10 Basic Data—Plutonium Measurement Example.

Case	Plutonium Activity (pCi/g)	Alpha Count Rate (#/sec)
1	20	.150
2	0	.004
3	10	.069
...	...	...
22	0	.002
23	5	.049
24	0	.106

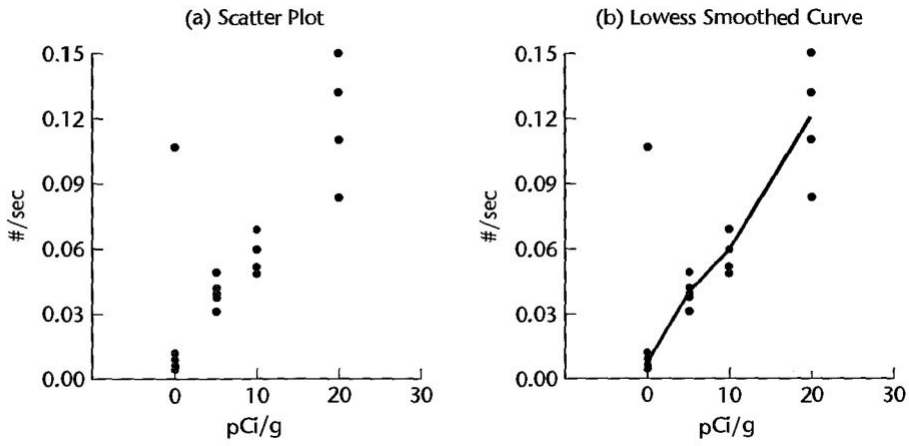
3. *Goal*: The task here is to estimate the regression relationship between alpha counts per second (Y) and plutonium activity (X).

4. *Assumption Before Doing Analysis*: the level of alpha counts increases with plutonium activity, but the exact nature of the relationship is generally unknown.

5. *Exploratory Data Analysis, EDA*:

   (a) *Scatter plot*: (Figure 3.20a) The strike rate tends to increase with the activity level of plutonium. Notice also that nonzero strike rates are recorded for the standard containing no plutonium. This results from background radiation and indicates that a regression model with an intercept term is required here.

FIGURE 3.20 SAS-JMP Scatter Plot and Lowess Smoothed Curve—Plutonium Measurement Example.



(b) *Investigate Relationship*: The regression relationship may be linear or slightly curvilinear in the range of the plutonium activity levels included in the study.

(c) *Outlier Detection*: An examination of laboratory records revealed that the experimental conditions were not properly maintained for the last case, and it was therefore decided that _____. A linear regression function was fitted next, based on the remaining 23 cases.

6. *Parameters Estimation and ANOVA*: (Figure 3.21a) the slope of the regression line is not zero ($F^* = 228.9984$, P-value$= 0.0000$) so that a regression _____.

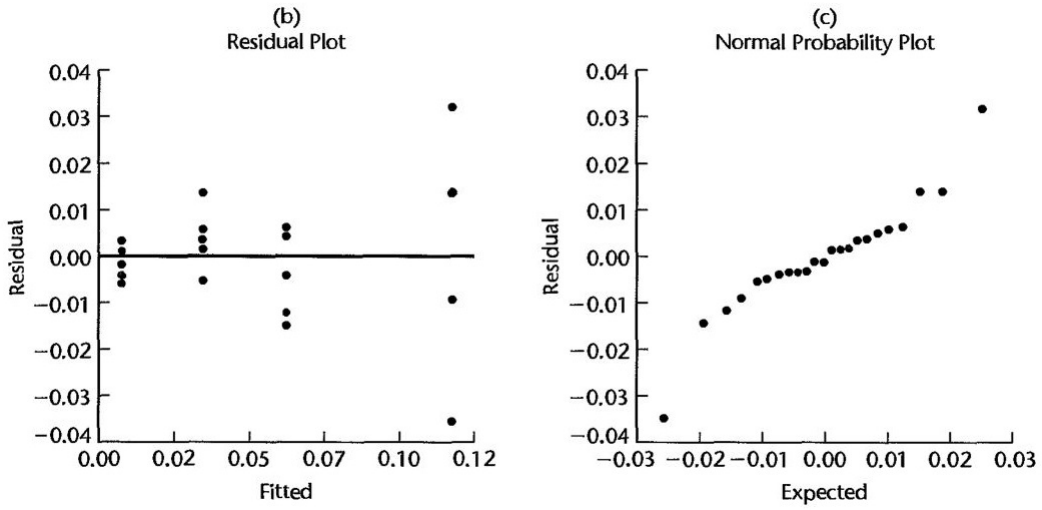
FIGURE 3.21 SAS-JMP Regression Output and Diagnostic Plots for Untransformed Data—Plutonium Measurement Example.

(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0070331	0.0036	1.95	0.0641
Plutonium	0.005537	0.00037	15.13	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.03619042	0.036190	228.9984
Error	21	0.00331880	0.000158	Prob>F
C Total	22	0.03950922		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00016811	0.000084	0.5069
Pure Error	19	0.00315069	0.000166	Prob>F
Total Error	21	0.00331880		0.6103



7. *Model Diagnostic*:

   (a) *Residuals Plot*: (Figure 3.21b) the flared, megaphone shape of the residual plot shows that the error variance appears to be increasing with the level of plutonium activity.

   (b) *The Normal Probability plot*: (Figure 3.21c) suggests non-normality _____, but the nonlinearity of the plot is likely to be related (at least in part) to the unequal error variances.

   (c) *Breusch-Pagan Test*: the existence of nonconstant variance is confirmed by the Breusch-Pagan Test statistic:

$$\chi^2_{BP} = 23.29 > \chi^2_{(0.95;1)} = 3.84$$

8. *Re-analysis After Data Transformation on Y*:

   (a) *Box-Cox transformation*: using the standardized variable, the maximum likelihood estimate of λ to be $\hat{\lambda} = 0.65$. The Box-Cox procedure supports the use of the _____ (i.e., use of $\lambda = 0.5$).

   (b) *Parameters Estimation and ANOVA*: (Figure 3.22a) The results of fitting a linear regression function when the response variable is $Y' = \sqrt{Y}$. The Lack of Fit Test statistic is $F^* = 10.1364$ with P-value $= 0.0010$.

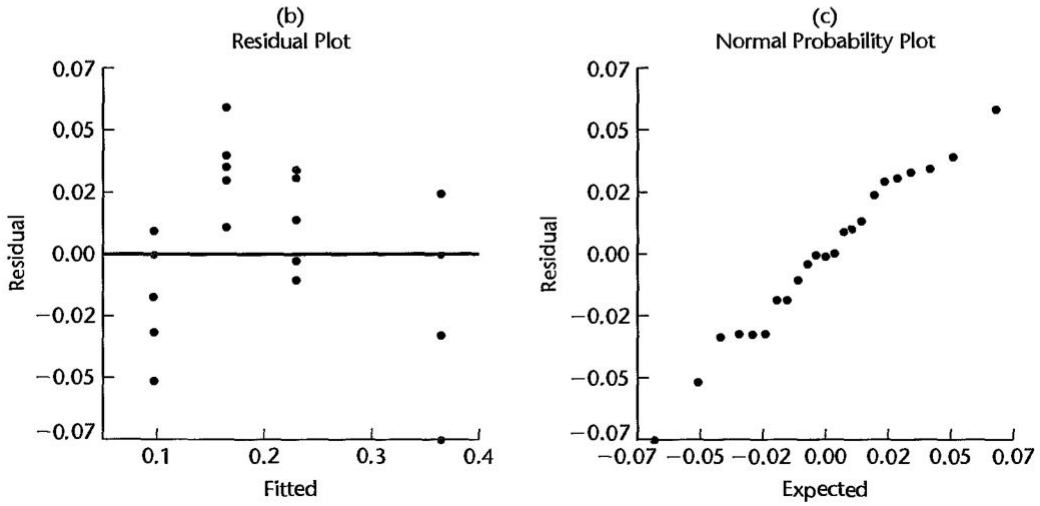
FIGURE 3.22 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response Variable—Plutonium Measurement Example.

(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0947596	0.00957	9.91	0.0000
Plutonium	0.0133648	0.00097	13.74	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.21084655	0.210847	188.7960
Error	21	0.02345271	0.001117	Prob>F
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.01210640	0.006053	10.1364
Pure Error	19	0.01134631	0.000597	Prob>F
Total Error	21	0.02345271		0.0010



(c) *Diagnostic Plots*: (Figure 3.22b, c) the residual plot shows that the error variance appears to be more _____, it also suggests the Y' is nonlinearly related to X. The points in the normal probability plot fall roughly on a _____ line.

9. *Re-analysis Again After Transformation on X*

(a) *Parameters Estimation and ANOVA*: (Figure 3.23a) The Lack of Fit Test ($F^* = 1.2868$ with P-value $= 0.2992$) supports the linearity of the regression relating _____ to _____.

FIGURE 3.23 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response and Predictor Variables—Plutonium Measurement Example.

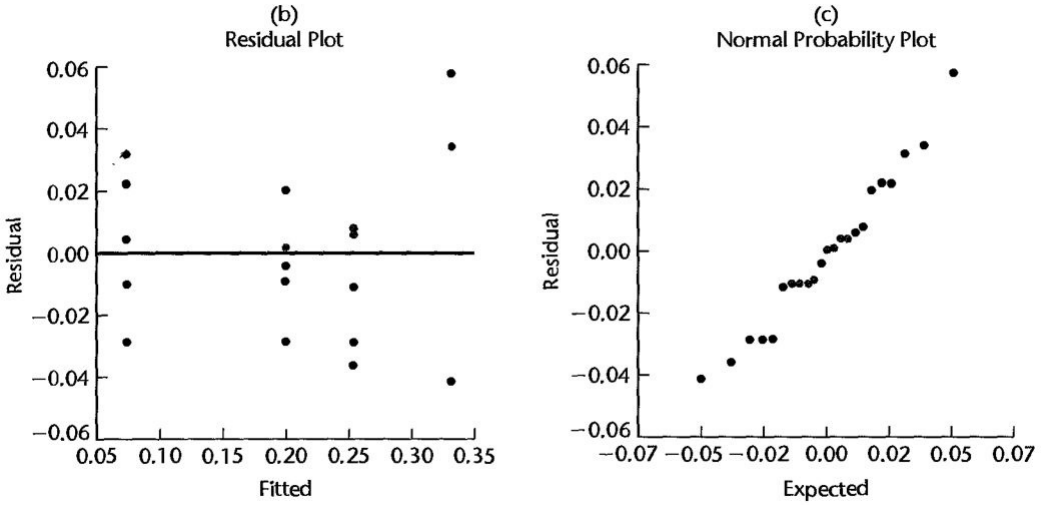
(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0730056	0.00783	9.32	0.0000
Sqrt Plutonium	0.0573055	0.00302	19.00	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.22141612	0.221416	360.9166
Error	21	0.01288314	0.000613	Prob>F
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00153683	0.000768	1.2868
Pure Error	19	0.01134631	0.000597	Prob>F
Total Error	21	0.01288314		0.2992

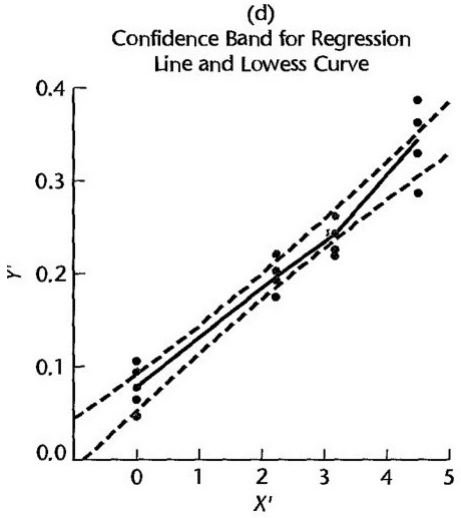
(b) *Diagnostic Plots* (Figure 3.23b, c) the residual plot shows that the square root transformation of the predictor variable has eliminated the lack of fit. It also suggests that some nonconstancy of the error variance may still remain; but if so, it does not appear to be _____. The normal probability plot of the residuals in Figure 3.23c appears to be satisfactory.



(c) *Diagnostic Tests*: the _____ ($r = 0.986$) supports the assumption of normally distributed error terms (the interpolated critical value in Table B.6 for $\alpha = 0.05$ and $n = 23$ is 0.9555). The _____ ($X^2_{BP} = 3.85$ with a P-value $= 0.05$) supports the conclusion from the residual plot that the nonconstancy of the error variance is not substantial.

(d) *Additional Results*: (Figure 3.23d) the scatter plot of X and Y with the con-

fidence band for the fitted regression line: _____. The regression line has been estimated fairly precisely. The lowess curve falls entirely within the confidence band, supporting the reasonableness of a linear regression relation between Y' and X'.



😍 TA Class

- **Problems**: 3.4, 3.9, 3.13, 3.15, 3.17
- **Exercises**: 3.20, 3.21
- **Projects**: 3.25