Chapter 3

Diagnostics and Remedial Measures

We take our first look at how check (diagnose) the assumptions necessary to carry out a linear regression and how to make corrections (take remedial measures) if these assumptions are not valid for the data.

3.1 Diagnostics for Predictor Variable

In an estimated linear regression,

$$\hat{Y} = b_0 + b_1 X,$$

it is important to check the predictor variable, X, for outlying observations and to check the range and concentration of observations as these things may have an impact of the validity of the regression analysis. Dot plots, percentiles and box–plots are useful tools when looking at the predictor variable. Although not discussed, histograms and stem and leaf plots are also useful in this regard.

Exercise 3.1 (Percentiles and Box–Plots)

1. *Dot Plots: Algae* Consider the algae weights (in ounces) given in the following three shipments of six containers each:

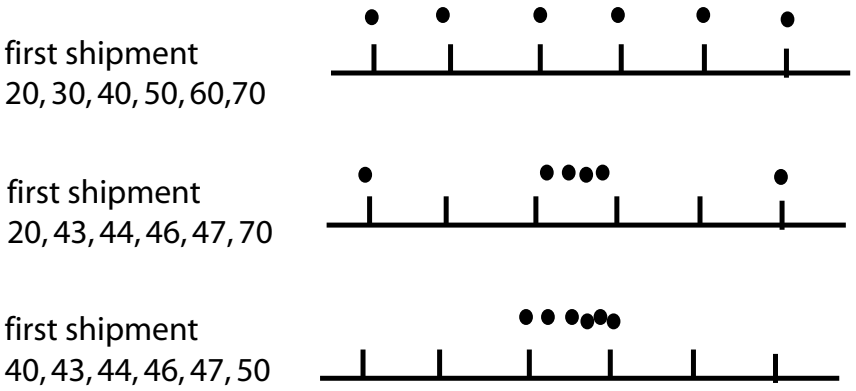


Figure 3.1 (Three Shipments of Six Containers of Algae Data)

It is clear from the set of deviations that the variability in the algae weights for the first shipment is
(circle one) **more than** / **the same as** / **less than**
the variability in the algae weights for the third shipment.

2. *Percentiles: Temperatures.* Reconsider the following sample of 10 temperatures,

$$0, 0, 0, 1, 1, 2, 2, 3, 3, 4.$$

   (a) The 50th percentile is that temperature such that 50% of the data is *below* this temperature. If 50% of the data is below this temperature, then 50% must be above this temperature. In other words, the 50th percentile must be the
   (circle one) **median** / **average**.

   (b) Assume the 50th percentile is the median. The *location* of the median can be calculated by adding one to the total number of data points, n, and dividing by 2, the $\frac{n+1}{2}$ formula. The *location* of the median for the $n = 10$ temperatures is
   (circle one) **5** / **5.5** / **6**.
   And so the median is equal to
   $\tilde{y}$ = (circle one) **0** / **1.5** / **2**.

   (c) The 25th percentile is that temperature such that 25% of the data is below this temperature and 75% is above this temperature. The 75th percentile must be
   (circle one) **above** / **equal to** / **below** the 50th percentile.

   (d) The 65th percentile is that temperature such that what percentage of all the temperatures is below this temperature?
   (circle one) **55%** / **65%** / **70%**.

   (e) The location formula for the 50th percentile, $\frac{n+1}{2}$, could be rewritten as $\frac{1}{2}(n + 1)$ or interpreted as "50% of $(n + 1)$". Not surprisingly, then, the

location formula for the 25th percentile could be interpreted as "25% of $(n+1)$" or $\frac{1}{4}(n+1)$. Consequently, the *location* of the 25th percentile for the $n = 10$ temperatures is
(circle one) **1.25** / **2.75** / **3.50**.
And so the 25th percentile is equal to
(circle one) **0** / **1** / **2**.

(f) The location formula for the 65th percentile could be interpreted as "65% of $(n+1)$" or $0.65\times(n+1)$. Consequently, the *location* of the 65th percentile for the $n = 10$ temperatures is
(circle one) **5.25** / **5.55** / **7.15**.
And so the 65th percentile is equal to
(circle one) **2.5** / **2.75** / **3.25**.

(g) The 25th percentile has a special name called the lower (or first) quartile and denoted $y^{(1q)}$. Similarly, the 50th percentile is often called
(circle one or more) **middle** / **second** / **upper** / **third** quartile.
Finally, the 75th percentile is often called the
(circle one or more) **middle** / **second** / **upper** / **third** quartile and denoted $y^{(3q)}$.

(h) The *five–number summary* is given by,

minimum, lower quartile, median, upper quartile, maximum,

and so, for the 10 temperatures, the five number summary
(circle one) **is** / **is not** $\{0, 0, 1.5, 3, 4\}$.
(Use your calculator: STAT ENTER; type 0, 0, ..., 4, into L_1; then STAT CALC ENTER 2nd L_1 ENTER; then arrow down to read minX = 0, $Q_1 = 0$, Med = 1.5, $Q_3 = 3$, maxX = 4.)

(i) The *interquartile range* is equal to the upper quartile minus the lower quartile. Consequently, the interquartile range for the set of 10 temperatures is given by,
$3 - 0 =$ (circle one) **1** / **2** / **3**.

(j) The interquartile range is robust to outliers whereas the variance and standard deviation are both sensitive to outliers. This means, for example, if one of the 10 temperatures above was mistyped as 40, instead of 4, the interquartile range would
(circle one) **not change much** / **would change a lot**,
but both the variance and standard deviation would
(circle one) **not change much** / **change a lot**.

3. *Box and Whisker Plot: Ph Levels Of Soil.* Consider the ordered set of the Ph levels of soil data given below.

4.3	5	5.9	6.5	7.6	7.7	7.7	8.2	8.3	9.5
10.4	10.4	10.5	10.8	11.5	12	12	12.3	12.6	12.6
13	13.1	13.2	13.5	13.6	14.1	14.1	15.1		

(a) Six pieces of information are required to draw a box and whisker plot: median, upper and lower quartiles, the interquartile range, upper and lower fences. The lower quartile, median and upper quartile are 7.95, 11.15 and 13.05, respectively and, so, the interquartile range is
$13.05 - 7.95 =$(circle one) **4.95** / **5.10** / **6.15**.

(b) The "upper fence" is determined by adding 1.5×(interquartile range) to the upper quartile,
$13.05 + 1.5(5.1) =$(circle one) **20.95** / **20.7** / **6.15**.

(c) The "lower fence" is determined by subtracting 1.5×(interquartile range) from the lower quartile,
$7.95 - 1.5(5.1) =$(circle one) **0.1** / **0.2** / **0.3**.

(d) Use your calculator to show the box and whiskers plot is given below.

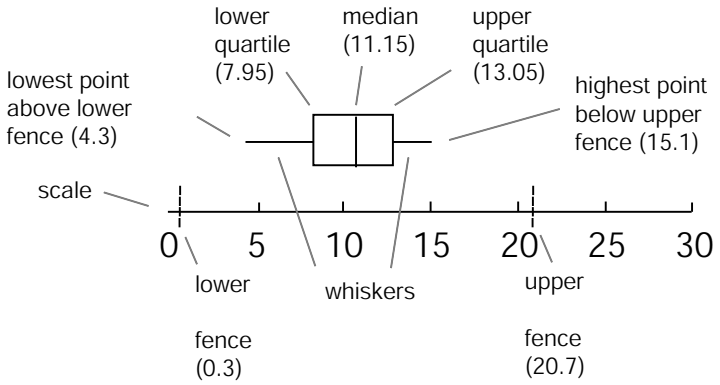


Figure 3.2 (Box and Whisker Plot For Ph Levels Of Soil Data)

The purpose of the "fences" is to determine if any data points are outliers or not. Data points outside the fences, are considered outliers. Any outliers are indicated as "dots" that are in line with, but out and away from, the whiskers.
(After STAT ENTER, type Ph levels into L_1; then 2nd STAT PLOT, choose second box plot; then ZOOM 9:ZoomStat ENTER; TRACE to see five number summary[1].)

[1]The calculator *can* give a slightly different box and whisker plot which uses hinges, rather than quartiles; in this case, the calculator gives the same box and whiskers as given above. Either box and whiskers plot is acceptable.

3.2 Residuals

The *residual*,

$$e_i = Y_i - \hat{Y}_i$$

could be considered an observed estimate of the *error*, where the error is given by

$$\varepsilon = Y_i - E\{Y_i\}$$

We will find out in later sections that residual plots are very important diagnostic tool to determine if the assumed model fits the observed data.

Exercise 3.2 (Residuals: Reading Ability Versus Level of Illumination)

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	75	88	91	94	100	92	90	85
predicted $\hat{Y}$	74.6	77.0	79.5	81.9	84.3	86.7	89.1	91.5	94.0	96.4
residual, $e = Y - \hat{Y}$	-4.6	-7.0	-4.5	6.1	6.7	7.3	10.9	0.5	-4.0	-11.4

1. **True** / **False** the residual plot of residuals versus the predictor variable, is really just the scatter plot with its regression line (and all the accompanying data points) "tilted" down to the horizontal *zero* line. The scatter plot is given on the left in the figure below, and the *residual plot* is given on the right in the figure below.

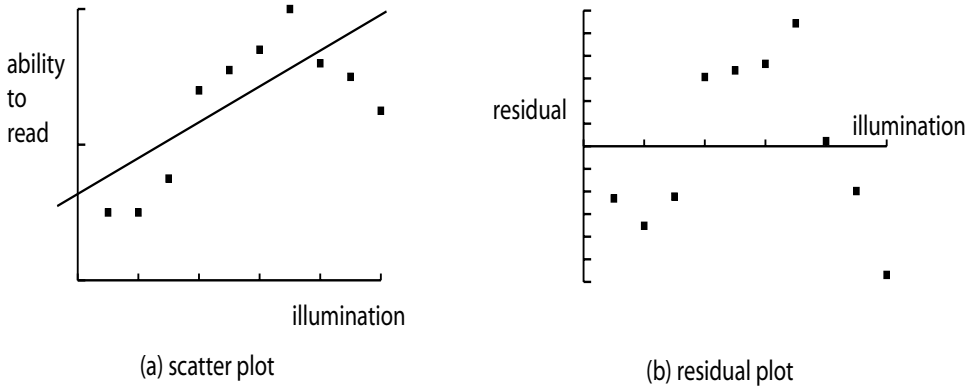


Figure 3.3 (Scatter Plot and Residual Plot for Reading Versus Illumination Data)

2. At $X = 3$, the *observed* value is
   $Y =$ (circle one) **70** / **75** / **88**,
   Since the least-squares line is $\hat{Y} = 72.2 + 2.418X$, at $X = 3$, the *predicted* value is $\hat{Y} =$ (circle one) **70** / **79.5** / **88**,
   and so the residual is
   $e = Y - \hat{Y} =$ (circle one) $\mathbf{-2.0}$ / $\mathbf{-4.5}$ / $\mathbf{-8.6}$.

3. **True / False**

   (a) The mean of the residuals is zero, $\bar{e} = 0$.

   (b) The estimator used for the variance, σ^2, of the error is given by the *mean square error* [2] (or mean square *residual*)

   $$MSE = \frac{\sum e_i^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

   where the sum of squares of the error (residuals) are $SSE = \sum (Y_i - \hat{Y}_i)^2$.

   (c) The residuals, e_i, are *dependent* because they involve the fitted values, $\hat{Y}_i$, which are all based on the same regression model.

4. *Semistudentized Residuals.*
   True / False Since $\bar{e} = 0$, the semistudentized[3] residuals are

   $$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

   which, notice, measure the size the residuals in standard deviation units, $\sqrt{MSE}$.

5. *Types of Residual Plots.* There are a number of different types of residual plots, including (circle none, one or more)

   (a) residual versus predictor variable

   (b) absolute or squared residual versus predictor variable

   (c) residual versus fitted

   (d) residual versus time (or other sequences)

   (e) residual versus omitted predictor

   (f) box plot of residuals

   (g) normal probability plot of residuals

6. *Uses of Residual Plots.* Residual plots are used to check that the simple linear regression model with normal errors satisfies a number of assumptions. These plots are used to check (circle none, one or more)

[2] Remember STAT TESTS **E:LinRegTTest...** ENTER gives the square root of the *MSE*

[3] Studentized residuals,

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

where h_{ii} are used to detect outliers, are more slightly more involved than *semi*studentized residuals. Studentized residuals are discussed later on.

(a) for linearity of the regression function.

(b) that the error terms have constant variance.

(c) that the error terms are normal.

(d) for outliers.

(e) that the error terms are independent.

(f) if one or more predictor variables are missing.

3.3 Diagnostics For Residuals

SAS program: att3-3-3-residual-plots

We use residual plots to check if the simple linear regression model fits the data or not. We find these various plots are sometimes ambiguous to read. In later sections, we use numerical analysis to pin down these ambiguous results. In general, graphical methods, such as residual plots, are useful to get a broad picture of any problems that arise when fitting a model to the data; numerical methods are useful when analyzing any particular problems that arise when fitting a model to the data.

Exercise 3.3 (Diagnostics For Residuals)

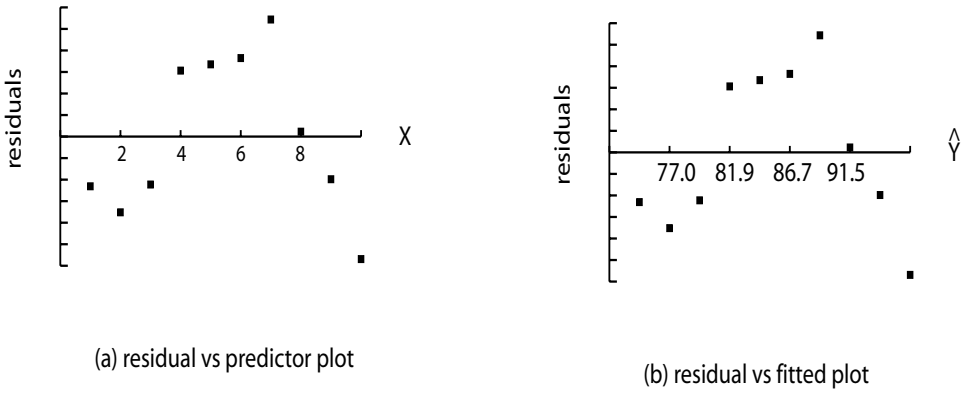
1. *Checking For Linearity of Regression Function: Reading Ability Versus Level of Illumination.* Is the data linear? Does it make sense to *assume* this data will be adequately represented by the *linear* regression,

$$\hat{Y} = b_0 + b_1 X.$$

We will use the residuals from the following reading ability data to help us perform this diagnosis.

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	75	88	91	94	100	92	90	85
predicted $\hat{Y}$	74.6	77.0	79.5	81.9	84.3	86.7	89.1	91.5	94.0	96.4
residual, $e = Y - \hat{Y}$	-4.6	-7.0	-4.5	6.1	6.7	7.3	10.9	0.5	-4.0	-11.4

The two residual plots given below are used to check for linearity of a simple linear regression function.



(a) residual vs predictor plot

(b) residual vs fitted plot

Figure 3.4 (Scatter Plot and Residual Plot for Reading Versus Illumination Data)

Both indicate that the data is (choose one) **linear / not linear**
because the residuals to *not* appear distributed at random both above and below the x–axis, but, in fact, appear in a curved concave patter.
Furthermore, both give (circle one) **the same / different**
information, since $\hat{Y}$ is a linear function of X in this case[4].

2. *Checking For Consistency of Variance.*
We *assume* the following simple linear regression model will fit the observed data,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

where the ε_i are independent $N(0, \sigma^2)$, $i = 1, \ldots, n$, and, in particular, σ^2 is *constant*; that is, we assume the variability in the error, measured by σ^2, does *not* change with i[5] A residual plot where the residuals appear distributed at random both above and below the x–axis, in a *horizontal band*, provides some evidence that σ^2 is, in fact, *constant*. Consider the following five residual plots.

[4]If $\hat{Y}$ was a curvilinear function of X, then the two plots would have given different information about the residuals.

[5]If the error variance did vary with i, this would be indicated by σ_i^2. Notice, by the way, that we are also assuming that the average error remains constant for different i and, in particular, this error mean, μ, is assumed to be zero (0)!

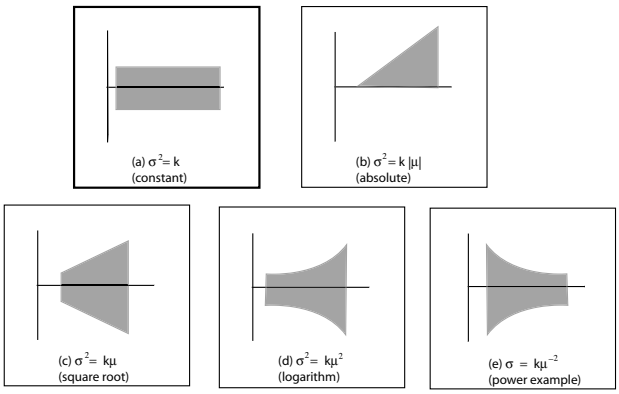


Figure 3.5 (Checking For Consistency of Variance)

Match residual (or absolute residual) versus predictor plots with type of residual plot.

residual plot	description
(a) $\sigma^2 = k$	**(a)** variance increases by square of mean
(b) $\sigma^2 = k\|\mu\|$	**(b)** variance increase linearly in mean
(c) $\sigma^2 = k\mu$	**(c)** absolute, variance increase linearly in mean
(d) $\sigma^2 = k\mu^2$	**(d)** constant variance
(e) $\sigma^2 = k\mu^{-2}$	**(e)** variance decreases by square of mean

residual plot	(a)	(b)	(c)	(d)	(e)
description					

Absolute residual plots accentuate any patterns that might be in the residual plot.

3. *Checking For Normality of Error Terms.*
We *assume* the following simple linear regression model will fit the observed data,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

where the ε_i are independent $N(0, \sigma^2)$, $i = 1, \ldots, n$, and, in particular, they are normally distributed; that is, we assume the chance of the size of the error follows a bell–shaped normal distribution and no other probability distribution such as a χ^2 or F distribution, say. A *normal probability plot* of residuals versus expected residuals,

$$e \quad \text{versus} \quad \sqrt{MSE}\left[z\left(\frac{k - 0.375}{n + 0.25}\right)\right],$$

which appears to be *linear* provides some evidence that the error is, in fact, normally distributed. Consider the following five normal probability plots.

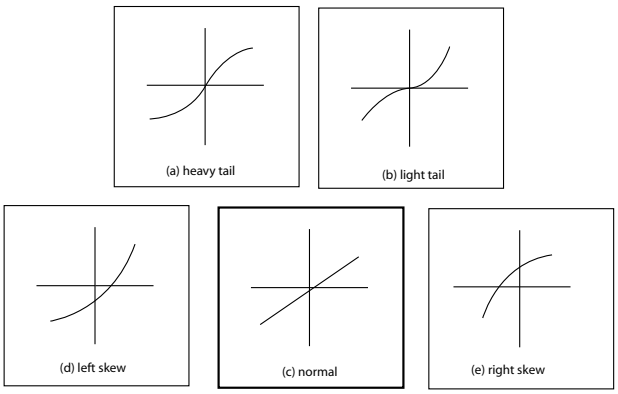


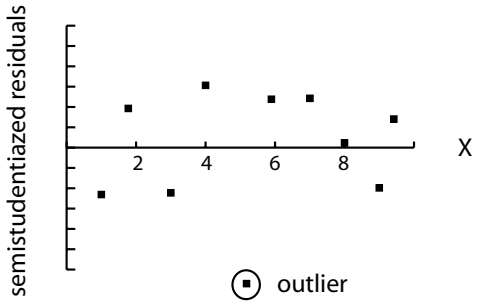
Figure 3.6 (Checking For Normality of Error Terms)

The SAS plot[6] indicates
heavy tail / light tail / normality
left skew / right skew / none of these Normality is checked last, after checking for constant variance and a linear regression model, since these first two problems can influence normality (such as in this case where nonlinearity is clearly influencing the normal probability plot). Normality of the error can also be detected using box plots, stem and leaf plots and histograms.

4. *Checking For Outliers.*
Although the assumed simple linear regression model,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n,$$

does *allow* for outliers, it is very *sensitive* to them in the sense that the slope of the linear regression can be tilted up or down very easily by an outlier[7]. Residual plots (or, even better, semistudentized residual plots) are useful in identifying outliers.



semistudentized residual vs predictor plot

Figure 3.7 (Checking For Outliers)

[6]PRGM NRMPLT could also be used in your calculator.
[7]There are different kinds of outliers, as we will discovery in later sections.

Most of the residuals in the *semistudentized* residual versus predictor plot are within three standard deviations of the regression function. The one circled outlier, though, is (circle one) **4** / **5** / **6** standard deviations from the regression function. Outliers can also be detected using box plots, stem and leaf plots and histograms.

5. *Checking For Independence of Error Terms.*
We *assume* the following simple linear regression model will fit the observed data,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

where the ε_i are *independent* $N(0, \sigma^2)$, $i = 1, \ldots, n$; where the value of the value of the ith error, ε_i, does *not* depend[8] on the value of the jth error, ε_j. A *pattern* on a residual plot indicates *dependent* errors; random scatter on a residual plot indicates *in*dependent errors.

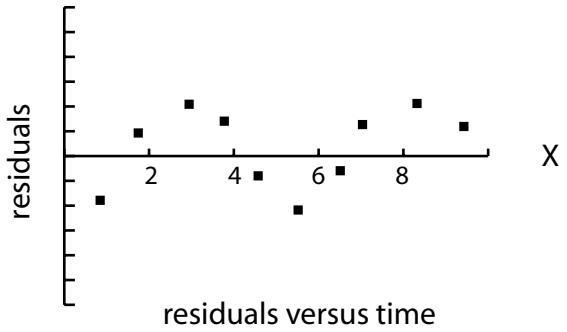


Figure 3.8 (Checking For Independence of Error Terms)

The residual versus time plot in indicates the error terms are
(circle one) **dependent** / **independent**
because of the wave–like pattern in the plot. Dependence of error terms often shows up with either too much alteration or too little alteration in residuals from positive to negative.

6. *Checking For Omission of Predictor Variables.*
We *assume* the following *simple* linear regression model will fit the observed data,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

Often, though, more complicated regression models, involving adding more predictor variables, say, are required to adequately represent the data,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, i = 1, \ldots, n$$

[8]The errors can be dependent on one another in different ways. For example, their variability may depend on one another and, in particular, the *co*variance of two errors would be indicated by σ^2_{ij}.

A missing *systematic* predictor component appears on a residual plot as a pattern in the residual plot. Consequently, the problem of missing predictor variables and the problem of dependence are (choose one) **easy** / **difficult** problems to distinguish between, when looking on a residual plot. Residual plots that show too much or too little alteration of positive and negative values could indicate the omission of the time predictor.

3.4 Overview of Tests Involving Residuals

We look at an overview of the various *significant tests* (as opposed to residual plots) are used to check that the various aspects of a simple linear regression model are satisfied.

Tests for Normality: goodness of fit tests such as chi–square, Kolmogorov–Smirnov test, Lilliefors test, simple test based on normal probability plot of residuals (discussed shortly).

Tests for constancy of variance: rank correlation between absolute values of residuals and predictor, modified Levene test (discussed shortly), Breusch–Pagan test.

Tests for Outliers: if chance of getting outlier, assuming other data are given, is small, reject outlier. Outliers are considered in greater detail, later, in Chapter 9.

Tests for randomness: runs test on residual versus time plot data, or Durbin–Watson test (considered in more detail, later, in Chapter 12).

3.5 Correlation Test for Normality

SAS program: att3-3-5-corr-test

A simple test based on normal probability plot of residuals is discussed which involves testing whether or not the correlation coefficient is 1 (indicating linearity and hence normality of error) or not.

Exercise 3.4 (Correlation Test for Normality: Reading Level Versus Level of Illumination)

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	75	88	91	94	100	92	90	85
predicted $\hat{Y}$	74.6	77.0	79.5	81.9	84.3	86.7	89.1	91.5	94.0	96.4
residual, $e = Y - \hat{Y}$	-4.6	-7.0	-4.5	6.1	6.7	7.3	10.9	0.5	-4.0	-11.4

Test if the coefficient of correlation between the ordered residuals and their expected values under normality is one (in other words, that the residuals are normally distributed) or not, at $\alpha = 0.05$.

1. *Statement.*
   The statement of the test, in this case, is (circle one)

   (a) $H_o : \rho = 1$ versus $H_1 : \rho > 1$

   (b) $H_o : \rho = 1$ versus $H_1 : \rho < 1$

   (c) $H_o : \rho = 1$ versus $H_1 : \rho \neq 1$

   In other words, the null is "residuals normal" and the alternative is "residuals not normal".

2. *Test.*
   From the SAS program, the test statistic is
   $r =$ (circle one) **0.909** / **0.937** / **0.972**
   The critical value at $\alpha = 0.05$ is (circle one) **0.934** / **0.918** / **0.901**
   (use Table B.6 of the Neter et al. text)

3. *Conclusion.*
   Since the test statistic, 0.972, is larger than the critical value, 0.901, (remember, this is a *left–sided* test!) we (circle one) **accept** / **reject** the null hypothesis that $\rho = 1$. In other words, the error terms *are* normally distributed.

3.6 Tests for Constancy of Error Variance

SAS program: att3-3-6-levene-breusch

We look at two tests for the constancy of error variance: the Levene test[9] and the Breusch–Pagan test.

Exercise 3.5 (Tests for Constancy of Error Variance: Reading Ability Versus Level of Illumination)

[9]The Neter et al. text does a *modified* Levene test which is suitable for the special case of splitting the data into only *two* groups and uses the t distribution. The test given here is the more general "actual" Levene test which is suitable for splitting the data into two *or more* groups and uses the F distribution. The Levene test is used later in the Neter et al. text. In any case, the results from the modified and unmodified Levene are the same. Students are only responsible for the SAS unmodified Levene version of this test.

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	75	88	91	94	100	92	90	85
predicted $\hat{Y}$	74.6	77.0	79.5	81.9	84.3	86.7	89.1	91.5	94.0	96.4
residual, $e = Y - \hat{Y}$	-4.6	-7.0	-4.5	6.1	6.7	7.3	10.9	0.5	-4.0	-11.4

1. *Levene Test, Two Groups.*
   Divide the data into two groups, $X < 6$ and $X \geq 6$, and test if the error variance is or is not constant between these two groups at $\alpha = 0.10$.

   (a) *Statement.*
   The statement of the test, in this case, is (circle one)

   i. H_0 : error variance constant *versus* H_1 : increases
   ii. H_0 : error variance constant *versus* H_1 : decreases
   iii. H_0 : error variance constant *versus* H_1 : not constant

   (b) *Test.*
   From SAS, the p–value is (circle one) **0.1344** / **0.3452** / **0.6348**
   The level of significance is (circle one) **0.05** / **0.10** / **0.15**

   (c) *Conclusion.*
   Since the p–value is (choose one) **smaller** / **larger** than the level of significance we (circle one) **accept** / **reject** the null hypothesis that the error variance is constant.

2. *Levene Test, Three Groups.*
   Divide the data into two groups, $X < 4$, $4 \leq X < 7$ and $X \geq 7$, and test if the error variance is or is not constant between these three groups at $\alpha = 0.10$.

   (a) *Statement.*
   The statement of the test, in this case, is (circle one)

   i. H_0 : error variance constant *versus* H_1 : increases
   ii. H_0 : error variance constant *versus* H_1 : decreases
   iii. H_0 : error variance constant *versus* H_1 : not constant

   (b) *Test.*
   From SAS, the p–value is (circle one) **0.0344** / **0.0452** / **0.0841**
   The level of significance is (circle one) **0.05** / **0.10** / **0.15**

   (c) *Conclusion.*
   Since the p–value is (choose one) **smaller** / **larger** than the level of significance we (circle one) **accept** / **reject** the null hypothesis that the error variance is constant.

3. *Breusch–Pagan Test.* Assume the error is related to X in the following way

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

The Breusch–Pagan test tests if $H_0 : \gamma_1 = 0$ (constant error variance) versus $H_1 : \gamma_1 \neq 0$.

(a) *Statement.*
The statement of the test, in this case, is (circle one)

i. H_0 : error variance constant *versus* H_1 : increases
ii. H_0 : error variance constant *versus* H_1 : decreases
iii. H_0 : error variance constant *versus* H_1 : not constant

(b) *Test.*
From SAS, the p–value is (circle one) **0.2288** / **0.2452** / **0.3841**
The level of significance is (circle one) **0.05** / **0.10** / **0.15**

(c) *Conclusion.*
Since the p–value is (choose one) **smaller** / **larger** than the level of significance we (circle one) **accept** / **reject** the null hypothesis that the error variance is constant.

3.7 F Test for Lack of Fit

SAS program: att3-3-7-drug-lackofit

The lack of fit test described in this section tests whether the *assumed linear* regression function,

$$\mu_j = \beta_0 + \beta_1 X_j$$

fits the observed data or not.

Exercise 3.6 (Test for Lack if Fit: Patient Responses Versus Drug Dosages) Twelve different patients are subjected to one drug at three dosages. Their responses are recorded below and a corresponding scatter plot is given.

10 mg	5.90	5.92	5.91	5.89	5.88
20 mg	5.51	5.50	5.50	5.49	5.50
30 mg	5.01	5.00	4.99	4.98	5.02

Test if there is a lack of fit between a linear (as opposed to quadratic, say) regression function and the data at $\alpha = 0.01$.

1. *Lack of Fit Test*

   (a) *Statement.*
   The statement of the test is (check none, one or more):

   i. $H_0 : \mu_j = \beta_0 + \beta_1 X_j$ versus $H_1 : \mu_j > \beta_0 + \beta_1 X_j$.
   ii. $H_0 : \mu_j = \beta_0 + \beta_1 X_j$ versus $H_1 : \mu_j < \beta_0 + \beta_1 X_j$.
   iii. $H_0 : \mu_j = \beta_0 + \beta_1 X_j$ versus $H_1 : \mu_j \neq \beta_0 + \beta_1 X_j$.

   (b) *Test.*
   The test statistic is

$$\begin{aligned} F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\ &= \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} \\ &= \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} \\ &= \frac{0.008670}{1} \div \frac{0.001880}{12} = \end{aligned}$$

   (circle one) **12.34** / **34.82** / **55.34**.
   The critical value at $\alpha = 0.01$, with 1 and 12 degrees of freedom, is
   (circle one) **5.32** / **9.33** / **11.26**
   (Use PRGM INVF ENTER 1 ENTER 8 ENTER 0.99 ENTER)

   (c) *Conclusion.*
   Since the test statistic is (choose one) **smaller** / **larger** than the critical value we (circle one) **accept** / **reject** the null hypothesis that the regression function is linear, $\mu_j = \beta_0 + \beta_1 X_j$; that is, it is *non*linear.

2. *ANOVA Table.*
   True / **False**. The ANOVA table is given by,

Source	Sum Of Squares	Degrees of Freedom	Mean Squares
Regression	2.02345	1	2.02345
Error	0.01055	13	0.000812
Lack of Fit	0.008670	1	0.008670
Pure Error	0.001880	12	0.000157
Total	2.034	14	

3.8 Overview of Remedial Measures

If the simple linear regression model is not appropriate for the data, we change it in a variety of ways to make it better fit the data. Some problems and possible solutions are given below.

Nonlinear Regression Function.
Instead of the simple linear regression function we have used in this chapter,

$$E\{Y\} = \beta_0 + \beta_1 X$$

we may add nonlinear variables to the model,

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2.$$

or we may transform the variables given in the model in a nonlinear way,

$$E\{Y\} = \beta_0 + \beta_1 \ln X.$$

in an attempt to get the model to better fit the data.

Error Variance Not Constant.
We look at changing the model to make the error variance constant using weighted least squares method and, later, and transforming the variables in model.

Error Terms Not Independent.
We look at changing the model to make the error terms independent.

Error Terms Not Normal.
We look at ways of changing the model to make non–normal error normal. We find out that nonconstant error variance and non normality often are interrelated: fixing the first often fixes the second.

Other Problems.
We look a ways of changing the model to fix other problems such as the possibility that important predictor variables are missing or that outlying observations are unduly influencing the regression model.

3.9 Transformations

SAS program: att3-3-9-read-transform

We look at transforming either the X variable or the Y variable or both variables to correct for nonlinearity, non–constant error variance and non–normality. We also look at identifying the "best" (simple regression with minimum[10] *SSE*) transformation in a class of transformations called Box–Cox transformations.

Exercise 3.7 (Transformations)

[10]The *SSE* is the sum of squared error and measures the total squared error between model and data.

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	76	88	91	94	100	92	90	85

Transform either the X and/or Y variables in the linear regression model[11] in an attempt to make the model not only fit the data more closely, but also to have more normal error and have more constant variance of the error.

1. *Transform X Variable to Linearize Model.*
   The X variable could be transformed in many different ways. Based on analysis and past experience, though, statisticians have found that if the scatter plot appears as one the patterned scatter plot given below, the X variable should be transformed as indicated in the equations given below each patterned scatter plot.

   (a) *Picking an Appropriate X Transformation.*

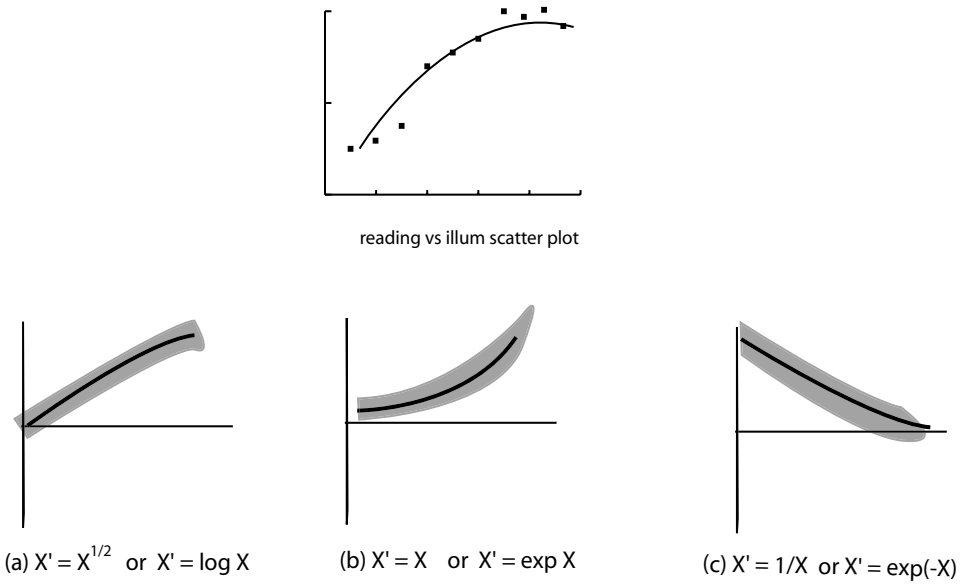


Figure 3.9 (Linearization Transformations on X)

   Which of the three transformations given in the figure above, appears to best represent the reading ability versus illumination scatter plot of data? Choose one **(a)** / **(b)** / **(c)**

   (b) **True** / **False**
   From SAS, two linearization transformations, $X' = \sqrt{X}$ and $X' = \ln X$, are attempted, with the resulting plots given below.

[11] The regression function is *linear* in the transformed variable X' (and/or transformed variable Y'), but *nonlinear* in the original variable X (and/or original variable Y).

illumination, X	1	2	3	4	5	6	7	8	9	10
$X' = \sqrt{X}$	1	1.41	1.73	2	2.24	2.45	2.65	2.83	3	3.16
$X' = \ln X$	0	0.69	1.10	1.39	1.61	1.79	1.95	2.08	2.20	2.30
ability to read, Y	70	70	76	88	91	94	100	98	99	97

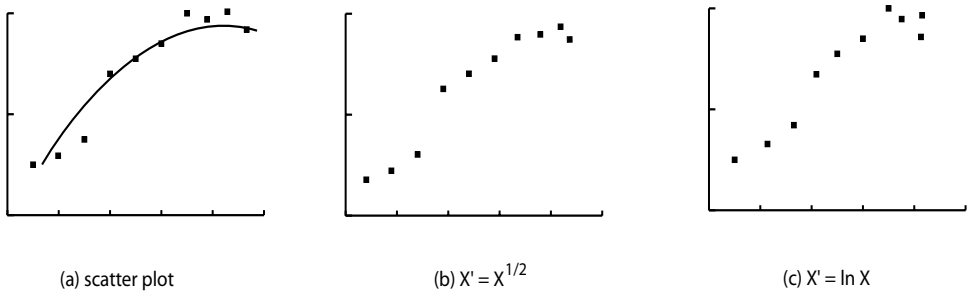


Figure 3.10 (Linearization Transformations)

(Type X and Y into L_1 and L_2 respectively. Define $L_3 = \sqrt{X}$ and $L_4 = \ln X$. Draw the three (X, Y) scatter plots of (L_1, L_2), (L_3, L_2) and (L_4, L_1).)

(c) **True** / **False**
The two linearization transformations, $X' = \sqrt{X}$ and $X' = \ln X$, are the *only* possible transformations.

(d) Assume we linearize the data using $X' = \sqrt{X}$. The new regression function is (circle *two!*)

i. $\hat{Y} = 60.05 + 11.32X'$
ii. $\hat{Y} = 60.05 + 11.32\sqrt{X}$
iii. $\hat{Y} = 60.05 + 11.32 \ln X$
iv. $\hat{Y} = 60.05 + 11.32X^2$

(STAT 2nd CALC LinReg(a+bx) L_4, L_2.)

(e) Assume we linearize the data using $X' = \ln X$. The new regression function is (circle *two*!)

i. $\hat{Y} = 68.09 + 11.52X'$
ii. $\hat{Y} = 68.09 + 11.52\sqrt{X}$
iii. $\hat{Y} = 68.09 + 11.52 \ln X$
iv. $\hat{Y} = 68.09 + 11.52X^2$

(f) From SAS, the correlation of both residual normal probability plots for the two transformations is
$r =$ (choose one) **0.98268** / **0.73453** / **0.34564**
This r value indicates the residuals (choose one) **are** / **are not** normal.

(g) **True / False**
To both linearize the model and, at the same time, *not* alter the error distribution, which depends on the Y, $e_i = Y_i - \hat{Y}_i$, it makes more sense to transform the X variable (as we have done here), than to transform the Y variable, when linearizing the data.

2. *Transform Y Variable to Normalize Error and Equalize Variance of Error.*

(a) *Picking an Appropriate Y Transformation.*

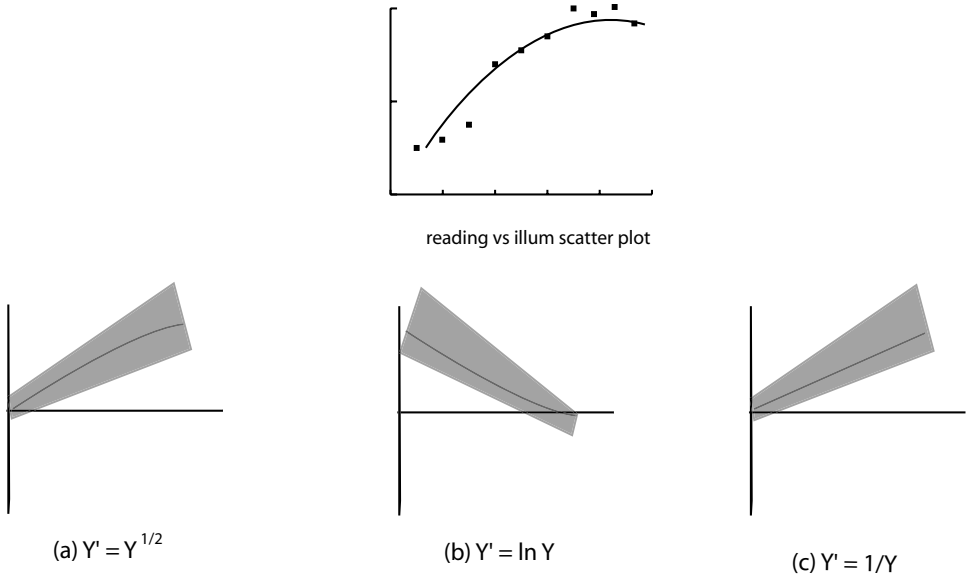


Figure 3.11 (Linearization Transformations on X)

Which, of the three transformations given in the figure above, appears to best represent the reading ability versus illumination scatter plot of data? Choose one **(a)** / **(b)** / **(c)**

(b) **True / False**
From SAS, the linearization transformation, $Y' = \sqrt{Y}$, is attempted, with the resulting results given below.

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	76	88	91	94	100	98	99	97
$Y' = \sqrt{Y}$	8.37	8.37	8.66	9.38	9.54	9.70	10.0	9.59	9.49	9.22

(c) **True / False**
The linearization transformations, $Y' = \sqrt{Y}$ is the *only* possible transformation.

(d) Assume we linearize the data using $Y' = \sqrt{Y}$. The new regression function is (circle *two*!)

i. $Y' = 8.49 + 0.13X$
ii. $\hat{\sqrt{Y}} = 8.49 + 0.13X$
iii. $\ln\hat{Y} = 8.49 + 0.13 \ln X$
iv. $\hat{Y} = 8.49 + 0.13\sqrt{X}$

(e) From SAS, the correlation of the normal residual plot is
$r =$ (choose one) **0.96825** / **0.73453** / **0.34564**
This r value indicates the residuals (choose one) **are** / **are not** normal.

(f) **True** / **False**
If the error distribution *needs* alteration, to make the error variance constant or normal, say, then it makes more sense to transform the Y variable, than to transform the X variable, when linearizing the data.

3. *Box–Cox Transformations.*
The Box–Cox transformations represents the power family of transformations on Y,

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

(a) From the SAS output, the various *SSE* values for given box–cox λ transformations, are given below.

λ	-0.2	-0.1	0	0.1	0.2
SSE	499.48	498.03	496.71	495.51	494.43

In this case, the best (minimum[12] *SSE*) box–cox transformation of the data is given by

(i) $\lambda = -0.2$, corresponding to the $Y' = Y^{-0.2}$ transformation
(ii) $\lambda = -0.1$, corresponding to the $Y' = Y^{-0.1}$ transformation
(iii) $\lambda = 0$, corresponding to the $Y' = \ln Y$ transformation
(iv) $\lambda = 0.1$, corresponding to the $Y' = Y^{0.1}$ transformation
(v) $\lambda = 0.2$, corresponding to the $Y' = Y^{0.2}$ transformation

(b) From SAS, the box–cox procedure suggests that the best estimated linear regression is given by (choose one)

(i) $\hat{Y}' = 236.17 + 2.51X$
(ii) $\hat{Y}' = 216.17 + 1.51X$
(iii) $\hat{Y}' = 206.17 + 0.51X$

(c) Since

$$\hat{Y}' = Y^{0.2} = 236.17 + 2.51X,$$

then (choose one)

[12]Remember, the *SSE* measures the error in the model relative to the data; that is, how far the data is away from the model. Small error means the model fits the data well.

(i) $Y = \left(\frac{1}{236.17+2.51X}\right)^5$
(ii) $Y = (236.17 + 2.51X)^5$
(iii) $Y = (136.17 + 0.51X)^5$

3.10 Exploration of Shape of Regression Function

SAS program: att3-3-10-loess

This section gives a description of one nonparametric regression curve method, the *loess* procedure, which fits a (nonlinear) smoothed curve to the data.

Exercise 3.8 (Loess Method: Nonparametric Regression Methods)

illumination, X	1	2	3	4	5	6	7	8	9	10
ability to read, Y	70	70	76	88	91	94	100	92	90	85

Use the loess method to fit a smoothed curve to the data. Although it is somewhat ambiguous, from SAS, the best fitting smoothed curve is (choose none, one or more)

(i) default (0.85)

(ii) 0.2

(iii) 0.3

(iv) 0.4

3.11 Case Example–Plutonium Measurements

This is an example which ties to together some of the remedial techniques discussed previously.