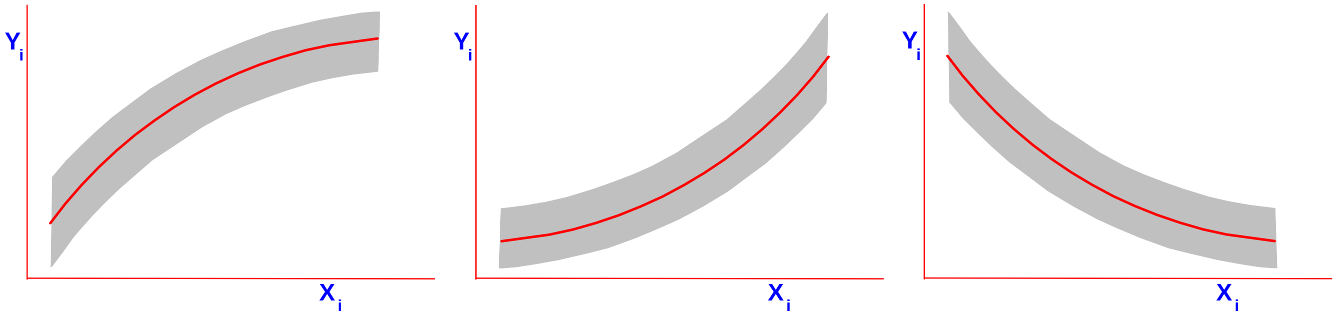
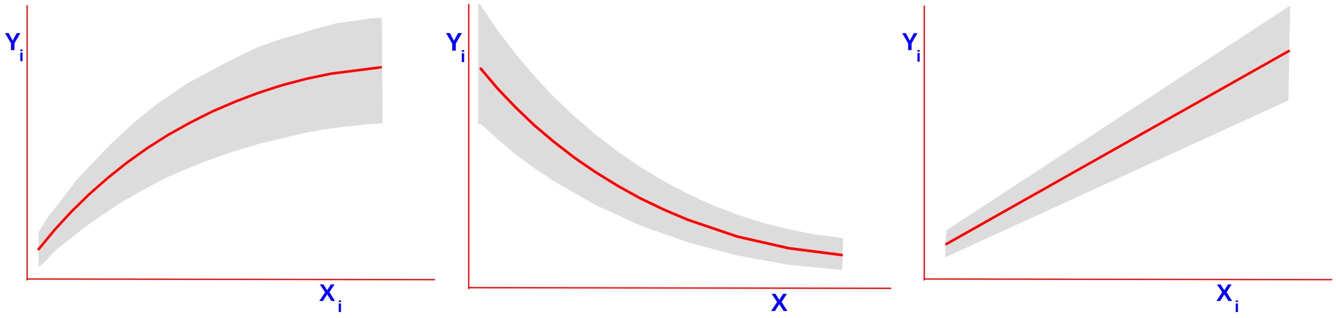
Chapter 8 (A little more on Assumptions for the Simple Linear Regression)

Linearity assumption – the best way to determine an appropriate model is to examine the literature for a theoretical model, or to see what model other investigators have found appropriate for a particular type of relationship.

There are some guidelines for selecting a transformation. Transformations in X do not affect the homogeneity of variance. The first model below would be fitted by a transformation of X like log(X) or ln(X) or √X. The second would be fitted by X^2 or e^X, and the third by X^2, 1/X or e^{-X}.



If the original data is homogeneous, the transformed model will still be homogeneous if only the X variable is transformed. Transformations of Y, on the other hand, will influence the homogeneity of variance. The model below would be transformed by log(Y) or ln(Y) or √Y or 1/Y.

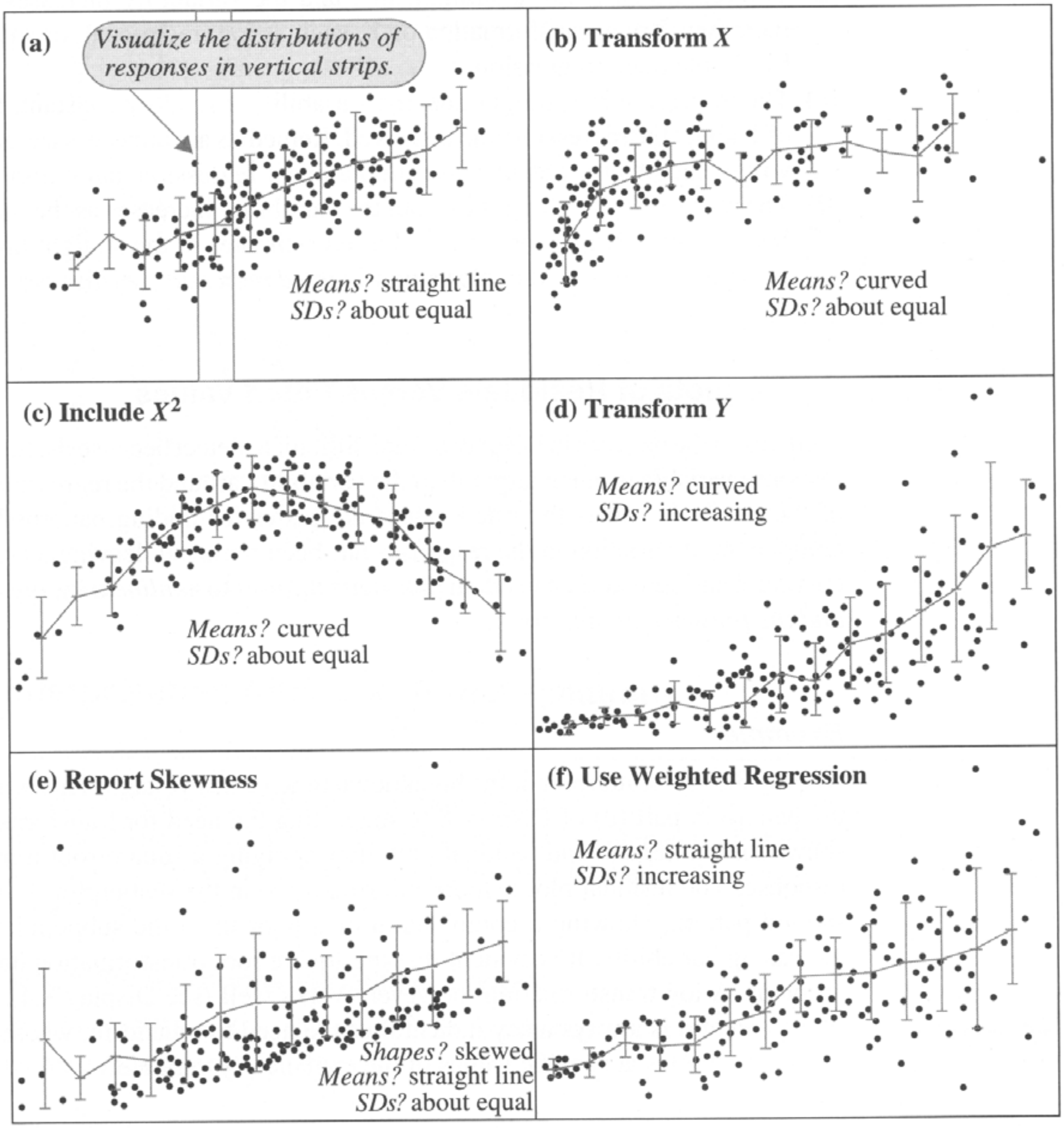


Transformations of Y, used to fit curvature when variance is NOT homogeneous and NOT normal

We saw several examples of these types of models, $Y_i = \beta_0 e^{\beta_1 X_i} \varepsilon_i$ and $Y_i = \beta_0 X_i^{\beta_1} \varepsilon_i$. Transformations of X_i are not as common except for polynomials (e.g. X, X^2, X^3, X^4, etc, covered in Chapter 10). Logarithmic transformations of X_i are used for some toxicology studies where the effect of a toxin is evaluated at a logarithmic progression of levels (e.g. 0.01mg/l, 0.1mg/l, 1mg/l, 10mg/l, 100mg/l, 1000mg/l, etc.) A logarithmic transformation (base 10) will then produce equally spaced values on the X axis and frequently linearize the data.

Graphical tools for Model Assessment
Scatterplot of Y on X (See graphics from the text).

Display 8.6 Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (a) is ideal



a) the text describes this as ideal, a straight line with homogenous variance

b) curved, but monotonic and homogeneous variance. Text recommends transforming X.

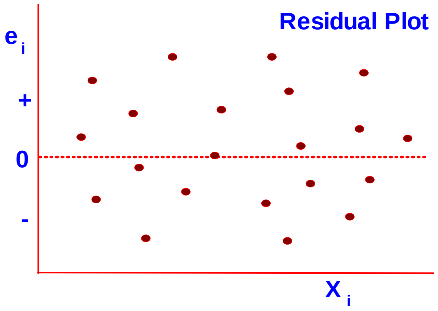
c) curved and homogeneous variance, not monotonic. Recommends quadratic regression (Chapter 10).

d) curved, monotonically increasing with nonhomogeneous variance, use a transformation in Y (log, reciprocal or square root).

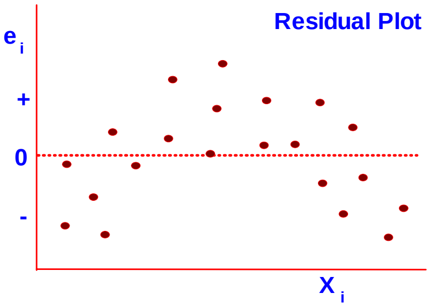
e) linear and monotonically increasing, but skewed. Recommends SLR but reporting skewness.

f) linear and monotonically increasing, but nonhomogeneous. Recommends weighted regression (Chapter 11).

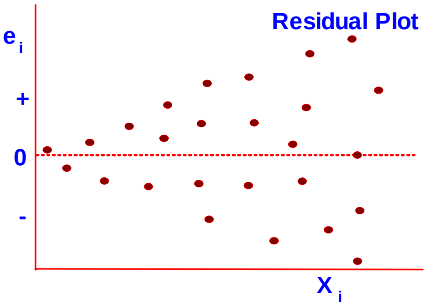
Scatterplot of residuals (See graphics from the text).



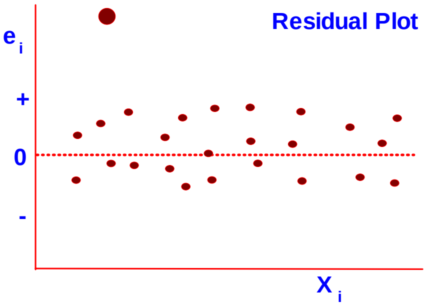
Normal residual plot, no problem



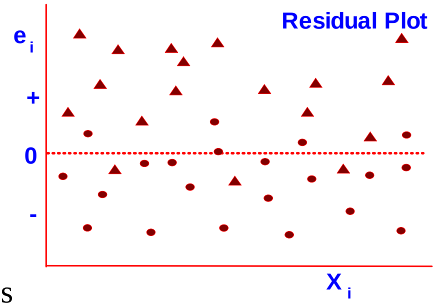
Residual plot shows curvature



Nonhomogeneous variance



Presence of outliers



Presence of several separate lines or levels

Assessment of Fit

Your book recognizes three models at this point.

$\mu_{Y|X_i} = \mu$ equal means model (single mean), one degree of freedom is correction factor adjustment

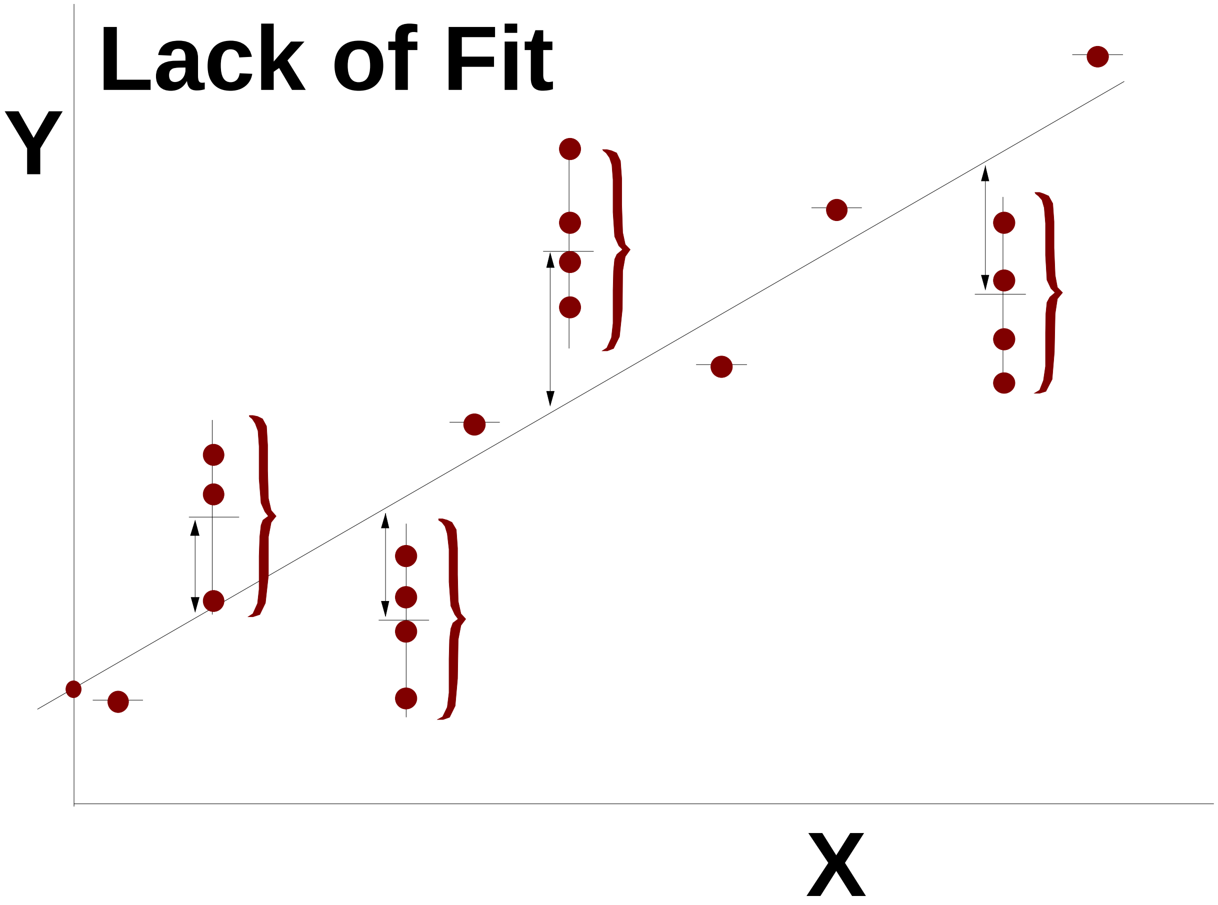
$\mu_{Y|X_i} = \mu_i$ separate means model (ANOVA), one d.f. for each mean

$\mu_{Y|X_i} = \beta_o + \beta_1 X_i$ simple linear regression, 2 d.f.

Extra SS for the difference between SLR and single mean model is the **SSRegression**

Extra SS for the difference between ANOVA and single mean is the **SSModel** for analysis of variance

Extra SS for the difference between SLR and ANOVA is lack of fit (SSLOF)



Calculating Lack of Fit for the fluid breakdown experiment – first for untransformed data

Regression analysis

```
    Analysis of Variance                Sum of            Mean
    Source                    DF       Squares          Square    F Value    Pr > F
    Model                      1       2150408         2150408      24.27    <.0001
    Error                     74       6557345           88613
    Corrected Total           75       8707754
```

ANOVA

```
    Dependent Variable: TIME            Sum of
    Source                    DF       Squares     Mean Square    F Value    Pr > F
    Model                      6   5082509.892      847084.982      16.12    <.0001
    Error                     69   3625243.641       52539.763
    Corrected Total           75   8707753.533
```

Composite (breakdown of the 6 d.f. of ANOVA into SSRegression (1 d.f.) and SSLOF (5 d.f.)

```
    Dependent Variable: TIME            Sum of
    Source                    DF       Squares     Mean Square    F Value    Pr > F
    Model                      6   5082509.892      847084.982      16.12    <.0001
       Regression              1   2150408.256     2150408.256      24.27     <.0001
       Lack of Fit             5   2932101.636      586420.327      11.16     <.0001
    Error                     69   3625243.641       52539.763
    Corrected Total           75   8707753.533
```

This test shows a significant LOF (F=11.16, P>F<0.0001), indicating that the regression does not provide an adequate description of the means.

Calculating Lack of Fit for the fluid breakdown experiment – transformed data

Regression analysis

```
    Analysis of Variance               Sum of           Mean
    Source                    DF      Squares         Square    F Value    Pr > F
    Model                      1    190.15149      190.15149      78.14    <.0001
    Error                     74    180.07484        2.43344
    Corrected Total           75    370.22633
```

ANOVA

```
                                          Sum of
    Source                    DF         Squares     Mean Square     F Value    Pr > F
    Model                      6     196.4774059      32.7462343       13.00    <.0001
    Error                     69     173.7489210       2.5181003
    Corrected Total           75     370.2263270
```

Composite (breakdown of the 6 d.f. of ANOVA into SSRegression (1 d.f.) and SSLOF (5 d.f.)

```
    Dependent Variable: LogTIME         Sum of
    Source                    DF         Squares     Mean Square     F Value    Pr > F
    Model                      6     196.4774059      32.7462343       13.00    <.0001
       Regression              1     190.1514905     190.1514905       78.14    <.0001
       Lack of Fit             5       6.3259154       1.2651831        0.50    0.7734
    Error                     69     173.7489210       2.5181003
    Corrected Total           75     370.2263270
```

This test shows the LOF to be not significant (F=0.50, P>F<0.7734), indicating that the regression does provide an adequate description of the means.

The calculation of Lack of Fit is an F test whose denominator is the Mean Square Difference between the full (ANOVA) and reduced (SLR) models. The denominator is the best available estimate of error, which is the full model error.

$$F = \frac{(SSError_{Regression} - SSError_{ANOVA})/(d.f.Error_{Regression} - d.f.Error_{ANOVA})}{MSError_{ANOVA}}$$

Choosing between SLR and ANOVA

In the example above the data was from a designed experiment, so the data was in groups. This analysis can either be done with a simple linear regression (1 d.f.) or as an ANOVA (6 d.f.). Generally, the model with fewer degrees of freedom is preferred. The regression analysis usually has a better interpretation. As a general rule, the simplest model is preferred.

Lack of fit can be used to examine the adequacy of the regression model even if the experiment is not a designed experiment. Any time there are some repeated observations at some values of X_i, lack of fit can be calculated. Repeated observations are not needed at all values of X_i.

R square, also called the Coefficient of Determination

The R^2 is the fraction of the corrected total SS that is accounted for by the model, or regression, SS.

$$R^2 = \frac{SSRegression}{SSTotal}$$

The proportion is usually expressed as a percent.

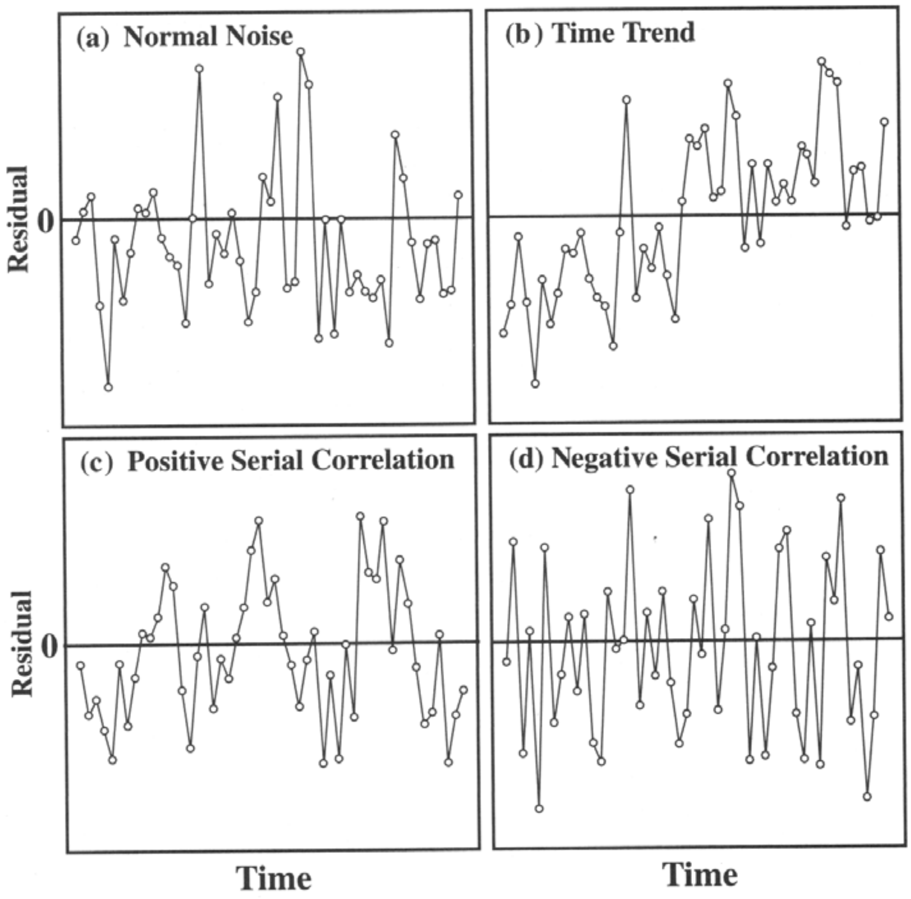
This statistic is not very meaningful unless the investigator is familiar with the context. Generally larger is better, but for some analyses a value of 30% might be very good while for other applications 90% is poor.

The square root (r) is the correlation coefficient between Y_i and X_i.

Other Residual plots for special situations (time order plot or ordered data plots)

In some situations we can detect problems by plotting the residuals on time or on the order that the data was taken. The problems detected have to do with a lack of independence, or in some cases the presence of variation due to an additional variable not included in the model (a time trend).

Display 8.12 Possible patterns in plots of residuals versus time order of data collection



a) **Random variation** (or "noise" as the book calls it): this is the expected pattern when the assumptions of independence is met.

b) **Time trend**: in this case there is some additional variable that is influencing the data and is related to time. It may be, for example, a function of "learning" as the experiment progresses. If the independent variable is randomized this will only add extra variation and will not bias the variable.

c) **Serial correlation (positive)**: if an observation has a positive residual there is a tendency for the next observation to also have a positive residual. Eventually the pattern switches, so negative residuals are followed by negative residuals. This indicates a lack of independence.

d) **Serial correlation (negative)**: this is the reverse of the previous pattern such that there is a tendency for positive residual is followed by negative residuals and vice versa. This pattern also indicates a lack of independence.

Normal probability plots

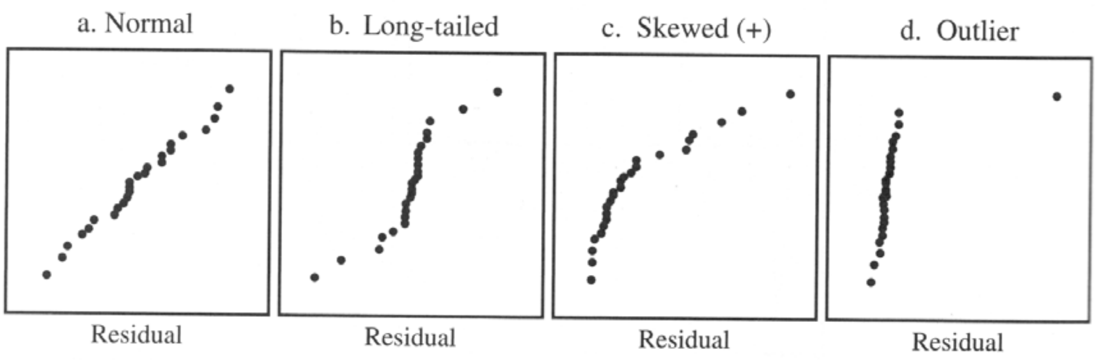
The best method of determining if you have met the assumption of normality is to use a statistical test. One of the better tests available is the Shapiro-Wilk test available in SAS PROC UNIVARIATE. There are also some graphical techniques (stem and leaf plot, box plot and normal probability plot) that will aid in determining if the assumption of normality is met and, if the assumption is not met, can aid in determining the nature of the departure from normality.

In SAS PROC UNIVARIATE the null hypothesis for the tests of normality is that the distribution of the raw data is consistent with a normal distribution. In order to obtain the tests of normality the “NORMAL” option must be requested. To get the plots mentioned the option “PLOT” should be included. The residuals tested can be obtained from many SAS procedures including PROC REG and PROC MIXED.

```
PROC UNIVARIATE DATA=datasetname NORMAL PLOT; VAR residualname;
RUN;
```

The normal probability plot is useful in determining how a distribution departs from normality. Some examples are given below.

Display 8.13 Normal probability plots illustrating four distributional patterns



a) Typical pattern for normality

b) Symmetric distribution, but with long tails in both directions (– and +).

c) Asymmetric distribution, but with long tails in the + direction.

d) Presence of an outlier

Below are some PROC UNIVARIATE graphics for some of the analyses we have seen.

Note that the first plot is similar to “d” above (presence of outlier) but is in the opposite direction.

PROC UNIVARIATE plots for fluid breakdown experiment – untransformed data

```
Tests for Normality
Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.615009    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.248565    Pr > D     <0.0100
Cramer-von Mises      W-Sq  1.204042    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  6.998223    Pr > A-Sq  <0.0050
```

```
The UNIVARIATE Procedure
Variable:  resid

   Stem Leaf                        #  Boxplot                        Normal Probability Plot
     18 4                           1     *        1850+                                                 *
     17                                                |
     16                                                |
     15                                                |
     14                                                |
     13                                                |
     12                                                |
     11 0                           1     *            |
     10                                                |                                             *
      9                                                |
      8                                                |
      7                                                |                                                 ++
      6 9                           1     *            |                                           *  +++
      5                                                |                                          ++++
      4                                                |                                       +++
      3                                                |                                    +++
      2                                                |                                ++++
      1 66666777                    8     |            |                             +++    ***** *
      0 256666666666666678         18  +--+--+         |                         +++********
     -0 98776555555555544442221    23  *-----*         |                    *********
     -1 6666665543321              13  +-----+         |               ******+
     -2 776555422                   9     |            |       ** *****++++
     -3 1                           1     |            |     *      +++
     -4 8                           1     0        -450+ *       +++
        ----+----+----+----+---                         +----+----+----+----+----+----+----+----+----+----+
    Multiply Stem.Leaf by 10**+2                            -2        -1         0        +1        +2
```

PROC UNIVARIATE plots for fluid breakdown experiment – transformed data

```
Tests for Normality
Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.956759    Pr < W      0.0112
Kolmogorov-Smirnov    D      0.10799    Pr > D      0.0267
Cramer-von Mises      W-Sq   0.11647    Pr > W-Sq   0.0699
Anderson-Darling      A-Sq  0.834933    Pr > A-Sq   0.0311
```

```
The UNIVARIATE Procedure
Variable:  resid

   Stem Leaf                     #  Boxplot                        Normal Probability Plot
      2 567                      3     |        2.75+                                          +*+*   *
      2 02                       2     |            |                                      ++++*
      1 5678888999              10     |            |                                  *******
      1 0222334                  7  +-----+         |                               ***+
      0 66778                    5  |     |         |                             ***
      0 00000123334444          14  *--+--*         |                        ******
     -0 4432211100              10  |     |         |                     ****
     -0 9777776665              10  +-----+    -0.75+                 ****+
     -1 442                      3     |            |               **++
     -1 7777                     4     |            |            ****
     -2 10                       2     |            |         ++*+
     -2                                |            |      +++  *
     -3 410                      3     |            |   +++  * *
     -3 6                        1     0            |+++    *
     -4 00                       2     0       -4.25+ *   *
        ----+----+----+----+                         +----+----+----+----+----+----+----+----+----+----+
                                                         -2        -1         0        +1        +2
```

The graphics above show no clear indication of problems.