Chapter Five
Multivariate Statistics

Jorge Luis Romeu
IIT Research Institute
Rome, NY 13440

April 23, 1999

Introduction

In this chapter we treat the multivariate analysis problem, which occurs when there is more than one piece of information from each subject, and present and discuss several materials analysis real data sets. We first discuss several statistical procedures for the bivariate case: contingency tables, covariance, correlation and linear regression. They occur when both variables are either qualitative or quantitative: Then, we discuss the case when one variable is qualitative and the other quantitative, via the one way ANOVA. We then overview the general multivariate regression problem. Finally, the non-parametric case for comparison of several groups is discussed. We emphasize the assessment of all model assumptions, prior to model acceptance and use and we present some methods of detection and correction of several types of assumption violation problems.

The Case of Bivariate Data

Up to now, we have dealt with data sets where each observation consists of a single measurement (e.g. each observation consists of a tensile strength measurement). These are called univariate observations and the related statistical problem is known as univariate analysis. In many cases, however, each observation yields more than one piece of information (e.g. tensile strength, material thickness, surface damage). These are called multivariate observations and the statistical problem is now called multivariate analysis.

Multivariate analysis is of great importance and can help us enhance our data analysis in several ways. For, coming from the same subject, multivariate measurements are often associated with each other. If we are able to model this association then we can take advantage of the situation to obtain one from the other. This is especially useful when one measurement or variable is easier, cheaper, faster or more accurately obtained than another one with which it is associated in some way.

For example, if the tensile strength of some material is associated with its thickness or its surface damage, we may be able to model this association accurately. If this is the case, then we can obtain an estimation of the tensile strength of a material having, say, a specific thickness, without the need to actually measure it. This advantage may save us the cost and the time of having to actually carry out this experimentation.

In the general case of multivariate analysis, each observation consists of n pieces of information represented by a vector x = (x_{1} , … , x_{n}). These information elements can be qualitative, quantitative or a combination of both. For example, we may have a vector of n=3 elements where the first component is tensile strength, the second is the batch (Id number) and the third is the material thickness. In this case, the first and third vector components are quantitative and the second (e.g. batch number) is really qualitative.

There are different statistical procedures according to whether the information is qualitative, quantitative or a combination of both. We start with the simpler, more straight forward multivariate case, with only two components: the bivariate analysis.

In the rest of this chapter, we first discuss and develop bivariate analysis examples for qualitative and quantitative data. Then, we develop a mixed qualitative/quantitative case. The case where both variables or information vector components are qualitative is developed via contingency tables and the analysis of categorical data. The case where both variables are quantitative is analyzed via correlation methods. The case where there is a combination of both, quantitative and qualitative variables, is approached using analysis of variance methods (ANOVA). Finally, we develop an example of the AD non-parametric test for several independent samples, that is an alternative to ANOVA.

Contingency Tables

Assume we have a bivariate information vector with two qualitative components. For example, we may want to analyze different sources that collect materials data information with the objective of classifying these sources into “reliable” or “deficient” data providers according to their data pedigree and their data handling track record.

We may then look at different characteristics of these data sources, in order to see if there are specific activities that these data sources do (or don’t do) that are somehow associated with they being classified as “reliable” or “deficient”. If such activity exists, we can look at this qualitative information and use it to pre-classify an incoming data set, or at least to have an initial idea of the quality of the data we will be handling.

Let’s assume that, in general, data sets may or may not have been “validated” (defined according to the Munro and Chen [14] methodology). Such “validation” includes the confirmation of the values obtained via the application of correlation and other statistical models to the data obtained. Let’s assume that there is reason to believe that those organizations submitting their data to such additional investigative procedures are able to check and correct any errors and hence display research characteristics that enhance and help insure the quality of the information they provide. Let’s assume that we want to use statistics to assess such a belief (i.e. our working hypothesis that we must test).

Assume that we have a set of n= 26 (fictitious) bivariate observations consisting of the qualitative pair (assessed data quality, validation practices) obtained from 26 data organizations. We will use these data to test the null hypothesis H_0 that there is NO association between the qualitative variables or factors “assessed data quality” and “validation practices”. The alternative hypothesis H_1 states that there is some kind of association (positive or negative) between the two factors above mentioned.

We can build a contingency table containing, as row entries, the values (reliable and unreliable data) and, by column, the values (validated and unvalidated data). Each organization will be classified in one of four possible classes (cells): reliable and validates data; unreliable and validates data, reliable and does not validate and unreliable and validates. Under the null hypothesis of no association, it would be equivalent to assign each organization, by chance, to any one of the four classification groups or cells.

At first glance, it may seem that all four classes would have the same expected frequency. But this is not the case, except when there is the same number of observations from each of the classes, which is rarely the case. For example this would occur if we had exactly the same number of organizations that had validated their data, as there were that hadn't.

But in general, the groups vary because they are selected at random and there is no reason that they should come out in the same proportion. In such case, the probability that they are classified, by chance, in any of the four classes depends on the expected (theoretical) class (cell) "size" in the sample chosen. This cell "size" is obtained by multiplying the total of the corresponding row by the total of the column, of the class or cell in question, and dividing the result by the sample size (general total). The resulting value (which in general is not an integer) yields the number of "expected" elements in the cell, i.e. those that would have fallen in that cell, were there no association between the two variables or factors. In our example, this is the "expected" number of organizations within each class, if there were no association between validation practices and data quality.

The reason for this variation in cell "size" is that row and column totals, which define the opportunities to fall in the cell, also differ. One can think of a square, painted on the floor and subdivided into four cells of different sizes. If one drops, at random, a bunch of beans on top of this square, the number of (expected) beans that end up in each of the four cell subdivisions is proportional to the (length times width) cell area or size.

The statistic used for testing this hypothesis is the Chi-Square, with degrees of freedom (d.f.) equal to the number of rows minus one (r-1) times the number of columns (c-1) minus one, that form the analysis table. The Chi-Square statistic has the form:

$$\chi = \sum_{i=1,k} \frac{(E_i-O_i)^2}{(E_i)}$$

Where E_i are the "expected" and O_i are the "observed" values (i.e. the actual values in the corresponding cell) for each of the k contingency table cells in the analysis.

For illustration, the example on data source classification is developed below. In each of the four contingency table cells, the expected counts are printed under the observed counts. The Chi-Square statistic has one degree of freedom (df = (2-1)*(2-1) = 1).

```
          good      bad    Total
Validate     9        2       11
Exp       5.50     5.50

NoValidate   4       11       15
Exp       7.50     7.50

Total       13       13       26

ChiSq =  2.227 +  2.227 + 1.633 +  1.633 = 7.721
```

From the Chi-Square table, the critical value (c.v.) for a significance level of $\alpha=0.05$ is: $\chi(0.05,1) = 3.84$. Since the test statistic (7.72) is larger than 3.84, it falls in the critical or rejection region. We reject the null hypothesis of no association between factors “data classification” and “validation practices”. Hence, there is an association and it becomes apparent from the contingency table: good data tends to be from validating organizations and bad data from non validating organizations. The assumed (ficticious) claim that validating the data has some positive bearing in data having a better quality, seems to be supported by the data. Further readings on categorical data analysis and the use of contingency tables can be found in the references [8, 9, 10] in this order of difficulty.

Regression

In the previous section, the bivariate data used was qualitative (categorical). That is, each observation or data point $P_{ij} = (i , j)$ provided two qualitative pieces of information (e.g. the data quality is good or bad and the organization validates or doesn’t validate its data). When, instead, both measurements are quantitative, i.e. when $P_i = (X_i , Y_i)$, where X_i , Y_i are quantitative variables, then the association between the two variables can be established more efficiently. For, now we have more information in the form of a stronger measurement scale for the variables under analysis.

Assume now that two quantitative measurements X_i and Y_i, are obtained from each data point P_i ; $1 \le i \le n$. Assume they correspond to the surface damage of a ceramics material (indexed 1 through 6, per the number of blemishes per unit area) and its corresponding tensile strength. Also assume that variable X_i (damage) is easier, faster, cheaper or more accurately obtained than Y_i (tensile strength). If X and Y were actually associated then we could find and use such relation to obtain an alternative or improved estimation of Y (tensile strength) given X (material surface damage). This is the philosophy behind regression analysis and what makes it a really useful working tool. Assume that the data for this problem consists of the 31 bivariate data points below, obtained by modifying the Example #2 of Program RECIPE (http://www.itl.nist.gov/div898/software/recipe/ex2.dat/) so as to serve as illustration of this statistical procedure.

```
 ROW  damage   matstr

   1       1   325.117
   2       1   331.767
   3       1   344.783
   4       1   343.266
   5       1   335.731
   6       2   291.039
   7       2   287.460
   8       2   302.042
   9       2   320.486
  10       2   312.130
  11       2   303.049
  12       3   328.093
  13       3   308.732
  14       3   312.429
  15       3   308.265
  16       3   282.888
  17       4   285.694
  18       4   315.397
  19       4   291.863
  20       4   301.098
  21       4   311.277
  22       5   297.974
  23       5   309.519
  24       5   319.596
  25       5   299.946
  26       5   307.719
  27       6   273.121
  28       6   291.785
  29       6   286.850
  30       6   270.018
  31       6   276.199
```

The descriptive statistics for the raw data are given below:

```
                 N     MEAN    MEDIAN STDEV  MIN      MAX       Q1       Q3
damage          31    3.452     3.000 1.729  1.000  6.000    2.000    5.000
matstr          31   305.66    307.72 19.67 270.02 344.78   291.04   319.60
```

The statistical analysis philosophy in this problem is similar to the one followed with qualitative variables. However, we can now implement more and better procedures because we have better and stronger information. We thus start by plotting Y_i (strength) vs. X_i (damage) for each $1\leq i \leq n$ (observation). If there is no association between variables X and Y (the same null hypothesis H_0 as before) then the resulting set of points $P_i = (X_i ,Y_i)$, will be uniformly and randomly scattered all over the X-Y plane.

We can also draw two lines, one vertical through the average of the projections over the X-axis (3.45) and one horizontal, through the average (305.6) of the projections over the Y-axis. They will divide the plane into four quadrants. Under H_0 (i.e. no association between damage/strength) the set of points P_i should be equally and randomly distributed among these four quadrants. As can be seen below, this is not the case. Actually, we have double the number of points in quadrants II and IV than in I and III. This indicates a possible negative association between X and Y (i.e. as one increases the other decreases).

```
        -     2
matstr  -
        -     x
        -     x                   x
     325+     x
        -               x                             x
        -                                   x
        -               x         3         x         2
        -               x
     300+               x                   x         2
        -
        -               x                   x                   x
        -               x         x         x                   x
        -
     275+                                                       2
        -                                                       x
        -
          ----+---------+---------+---------+---------+--------damage
            1.0       2.0       3.0       4.0       5.0       6.0
```

In general, if there is an association between X and Y (i.e. if we reject H_0) the number of points in each quadrant will differ. If there is a positive association (i.e. when X increases/decreases, so does Y) then the points will tend to cluster in the upper right and lower left quadrants. If there is a negative association between X and Y, the points will cluster in the upper left and lower right quadrants. Such is the case above.

Again, this situation is analogous to setting free some beans, at random, from a location on the intersection of the mentioned two lines drawn on the plane shown above. Since the four quadrants are of the same size and there is no other external force guiding the beans, the expected number of beans in each of the four quadrants is the same. On the other hand, if there is a driving external force (such as an association between the two factors analyzed) then the number of beans in each of the quadrants will differ.

In the scatter plot above, the second and fourth quadrant concentrate most of the points at the expense of the first and third quadrants. This indicates that, in this example, there is a negative association between factors X and Y (i.e. that as the material surface damage increases, its tensile strength decreases). We proceed to investigate this analytically.

The indicator "covariance between X and Y" characterizes such relationship. It is defined as: Cov(X, Y) = S_{xy} = $\sum(x_i\text{-}x^*)(y_i\text{-}y^*)$ / (n-1) ; where x^* and y^* are the corresponding sample averages. The covariance indicator is positive when a positive association between X and Y exists; negative when a negative association exists and zero if no association exists. In our surface damage and tensile strength example we have:

```
Covar (damage,  matstr) =  -22.92
```

As a measure of association between two variables, the covariance is difficult to interpret. For, it depends heavily on the units in which variables X and Y are being measured. The above covariance of 386.8 is no exception. We would like to have a more interpretable measure of the degree of association between variables X and Y than this number.

The correlation coefficient, defined as $r_{xy} = S_{xy} / S_x S_y$ (where S_x and S_y are the sample standard deviations of variables X and Y) is sort of a “normalized” covariance. It also measures the association between X and Y, like the covariance. However, r_{xy} is “dimensionless” therefore, easier to interpret, because $-1 \le r_{xy} \le 1$. In our example:

```
Correlation between damage and matstr = -0.674
```

This again means that, as X (surface damage) increases, Y (tensile strength) decreases. In addition, r_{xy} is a measure of “linear” association between X and Y. If $r_{xy} > 0$ (and close to unit) then there is a “linear” trend that models the association between X and Y, with positive slope. If $r_{xy} < 0$ (and close to -1) this linear trend has a negative slope. In our example, the correlation value (–0.674) indicates that there is a linear trend that models the relationship between material surface damage and its associated tensile strength, with a negative slope. The Y vs. X plot above has been analytically corroborated.

It is therefore very useful to obtain the analytical form of such a linear trend (when it exists) called the linear regression. Then, we use it to obtain a better estimate of Y (the dependent variable) given X (the predictor variable). For, lacking any other information, the best estimate of Y (tensile strength) given X (material surface damage) is always the (point estimator) mean of Y (or at best, a confidence interval for its mean). With the additional information that a linear relationship exists between X and Y, and having a specific value of X, we can (by using the regression) improve in the estimation of Y.

From the tensile strength boxplot we obtain a sample of its distribution and its median:

```
                  -------------------
    --------------I          +      I------------------
                  -------------------
----+---------+---------+---------+---------+---------+matstr
  270       285       300       315       330       345
```

The above distribution looks symmetric; its sample average and standard deviation are known to be: (305.66, 19.67). With no further information, the mean and variance would provide the best estimation for tensile strength. However, if an association between surface damage and tensile strength exist, estimations can be improved with regression.

In mathematical terms, the (theoretical) linear regression model has the expression:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad ; \quad 1 \le i \le n$$

Constants β_0 and β_1 are called regression coefficients. Variable ε_i represents the random error term, which is distributed Normally with mean zero (which means it is symmetric and averages out). The estimated regression line is obtained by minimizing the sums of squares ($\sum\varepsilon^2{}_i$) of the distances (called residuals) from every data point to the said line. This one is not a statistical procedure. It is just a mathematical procedure for driving a line through and closest to a cloud of points on a plane. The point estimations for the linear regression coefficients are also random variables and are denoted by b_0 and b_1.

In addition to the above mathematical work, one imposes a specific statistical distribution on the residuals (the distances from each point to the line). This other is a statistical procedure. Mathematically, linear regression is just the line of best fit to the data points. Hence, all its estimations (e.g. of coefficients β_0 and β_1 or of the mean Y values given specific values X) are always acceptable point estimators.

However, the tests of hypotheses, confidence intervals and other regression outputs are all statistical concepts. They are only valid when the statistical assumptions of Normality, equal variance and independence of the regression residuals are met. This is an important differentiation, to be taken into account whenever implementing a regression model.

From the above data, we have obtained the linear regression for tensile strength, given the surface damage of the material in question. The results are given below:

```
                       matstr = 332.13 - 7.67 damage

Predictor       Coef       Stdev    t-ratio        p
Constant     332.135       6.002      55.34    0.000
damage        -7.671       1.560      -4.92    0.000

s = 14.77       R-sq = 45.5%     R-sq(adj) = 43.6%
```

The material tensile strength (matstr) is estimated by the line with intercept 332.13 and slope –7.67. If surface damage were not a consideration (not significant) then the best estimate for tensile strength would always be the mean of tensile strength values. The two regression coefficient estimators: b_0 (constant term or intercept, 332.13) and b_1 (slope term –7.67) are R.V. with standard deviations 6.002 and 1.56, respectively. If residuals are independent and Normally distributed with mean zero and same variance, then b_0 and b_1 are Normally distributed, with means β_0 and $\beta_{1.}$ We can then use the (small sample) t statistic, described in the previous chapter, to test H_0 that β_0 and β_1 are zero or not. The values of the t-test statistics (55.34 and –4.92) are given under t-ratio and the probability of erroneously rejecting H_0 is given under p-value (0.000).

The importance of testing H_0 that the regression coefficient β_1 is zero lies in the fact that, if it is zero, there is no slope. Hence, there is no regression and variable Y is independent of X. In our case, the t-test statistic for testing that β_1 is zero is –4.92 and very highly significant. In addition, the regression index of fit or coefficient of multiple determination $R^2 = 0.455$ (denoted above as R-sq= 45.5%=100xR^2), indicates that this regression model "explains" 45.5% of the problem variation (the remaining 55% is "unexplained" by the model and hence explained by the remaining random variation in the problem).

Therefore, we reject H_0 and our probability of error is practically zero. We then derive a c.i. for the true (unknown) value of the slope, using the regression slope b_1 = -7.67 point estimator and a function (denoted S(b$_1$)) of the regression model standard deviation s. We then follow the small sample c.i. procedure for (b$_1$ - t$_{\alpha/2}$ S(b$_{1)}$, b$_1$ + t$_{\alpha/2}$ S(b$_1$)), where t (α/2 , n-2) = t (0.025, 31-2) = 2.045 and S(b$_1$)=1.56 from the regression table above.

None of these statistical results, however, hold unless the regression model hypotheses of Normality, independence and equal variance of the residuals also holds. Therefore, no regression work is entirely complete unless a thorough analysis of residuals is also performed. In our case, the residuals (obtained by subtracting each of the regression model estimations from their corresponding original values) are the following (RESI1):

```
   0.6539    7.3039   20.3198   18.8026   11.2679  -25.7534  -29.3326
 -14.7502    3.6943   -4.6624  -13.7433   18.9724   -0.3887    3.3086
  -0.8554  -26.2325  -15.7548   13.9482   -9.5862   -0.3507    9.8278
   4.1967   15.7415   25.8188    6.1682   13.9417  -12.9847    5.6790
   0.7437  -16.0878   -9.9067

                N      MEAN    MEDIAN    STDEV   MIN     MAX      Q1       Q3
RESI1          31     -0.00      0.74    14.52 -29.33   25.82   12.98   11.27
```

Notice how the mean is zero but the standard deviation is 14.52. For convenience, we standardize them (i.e. divide by the standard deviation) and they become Normal (0,1):

```
SRES1
  0.04665   0.52104   1.44955   1.34132   0.80382  -1.79432  -2.04370
 -1.02769   0.25740  -0.32484  -0.95755   1.30726  -0.02678   0.22798
 -0.05894  -1.80751  -1.08617   0.96161  -0.66089  -0.02418   0.67754
  0.29290   1.09865   1.80198   0.43049   0.97304  -0.92907   0.40634
  0.05322  -1.15110  -0.70884

                N      MEAN    MEDIAN    STDEV    MIN     MAX     Q1      Q3
SRES1          31     0.002     0.053    1.015 -2.044 1.802 -0.929   0.804
```

Now the residuals have mean zero and variance unit. We will use them to assess the three statistical assumptions of Normality, independence and equality of variance of regression. We first obtain the boxplot, stem-and-leaf and Anderson Darling statistic for them:

```
                            ------------------------
           -----------------I            +         I---------------
                            ------------------------
          +---------+---------+---------+---------+---------+-----SRES1
      -2.10     -1.40     -0.70      0.00      0.70      1.40

    1    -2 0
    3    -1 87
    6    -1 100
   10    -0 9976
   14    -0 3000
   (7)    0 0022244
   10     0 56899
    5     1 0334
    1     1 8
```

Both the boxplot and stem-and-leaf suggest that residuals are unimodal and symmetric about zero, as expected from a Normal Standard distribution. In addition, the Anderson Darling test for Normality yields a statistic AD = 0.313, with p-value=0.53. Hence, residuals can be assumed Normal Standard. Regarding the independence of the residuals, we plot each residual with respect to its predecessor, below, to assess this assumption.

```
        -
        -                                               x
      20+                 x                    x            x
        -                                  x
i+1     -               x                     x            x
        -                              xx                        x
        -                x x           x         x
       0+                     x            x  x            x
        -                                  x
        -               x                             x
        -                         x                   x
        -  x  x                          x
     -20+
        -                                         x
        -     x                        x
        -
        -
          +---------+---------+---------+---------+---------+------ i
        -30       -20       -10         0        10        20
```

The resulting random pattern is characteristic of an independent sample (the pattern of a serially correlated sample would instead appear as a defined form, e.g. linear, sinusoidal, etc. according to the existing correlation). We can also obtain their serial correlation (x_i , x_{i+1}) =0.335 and time series plot. All these procedures help assess residual randomness (or lack thereof) which are related to their independence. Since in our example there is small evidence against it, we are willing to assume the independence of the residuals. The last assumption to assess is that of equality of variance. Some residual plots may aid in assessing this assumption. One is the plot of residuals versus the estimations:

```
        -
        -             x
     1.5+                                                         x
        -             x                     x                     x
SRES    -             x          x                                x
        -                        x                                x
        -  x          2                     x          x
     0.0+  x                     x          2                     x
        -                                              x
        -  x                     x
        -  x                                           2
        -  x                     x
    -1.5+
        -                                   x          x
        -                                              x
        -
        -
          --+---------+---------+---------+---------+---------+---FITS
        287.0     294.0     301.0     308.0     315.0     322.0
```

The pattern to observe is one of randomness of the residuals about mean zero. The pattern above is, for the purposes of this example, reasonable enough to support the equality of variance assumption. If the pattern were funnel-like (i.e. the range of residual values increases as the values of the Fits increase) then the residual variance probably depends on the mean. When this occurs, the assumption of equal variance no longer holds, in which case a variance stabilizing data transformation is one problem solving possibility.

Several analytical procedures exist that also assess the equality of variance assumption. There are formal tests such as the one proposed by Lehman, described in Section 8.6.3.2 of the old (Version 1D) of [7] and by Bartlett (in [9]). The equality of variance test proposed by Levine is described in Section 8.3.5.2 (Version 1E of [7]). It is a non-parametric test that transforms the original observations X_{ij} (in this case the model residuals) into:

$$W_{ij} = \left| X_{ij} - X_i \right| \quad ; \quad \text{where} \quad X_i \text{ is the median of the ith group}$$

Then an (ANOVA) F-test is performed for the transformed data. If the F statistic is larger than the tabulated F for the corresponding ANOVA test, equality of variance is rejected.

In our tensile strength example, the medians for the six standardized residual groups are 0.804, -0.993, -0.027, -0.024, 0.973 and -0.709. The transformed variable W is:

```
  0.75735   0.28296   0.64555   0.53732   0.00018   0.80132   1.05070
  0.03469   1.25040   0.66816   0.03545   1.33426   0.00022   0.25498
  0.03194   1.78051   1.06217   0.98561   0.63689   0.00018   0.70154
  0.68010   0.12565   0.82898   0.54251   0.00004   0.22007   1.11534
  0.76222   0.44210   0.00016
```

The F-test (analysis of variance) for the above data set yields the result:

```
ANALYSIS OF VARIANCE ON levine2
SOURCE     DF        SS        MS        F        p
damage      5     0.336     0.067     0.26    0.929
ERROR      25     6.367     0.255
TOTAL      30     6.702
                                   INDIVIDUAL 95% CI'S FOR MEAN
                                   BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  -+---------+---------+---------+----
-
     1      5    0.4447    0.3043  (-------------*------------)
     2      6    0.6401    0.5099         (-----------*-----------)
     3      5    0.6804    0.8219         (-------------*-------------)
     4      5    0.6773    0.4195         (-------------*-------------)
     5      5    0.4355    0.3577  (------------*-------------)
     6      5    0.5080    0.4412    (-------------*------------)
                                   -+---------+---------+---------+----
-
POOLED STDEV =   0.5046           0.00      0.35      0.70      1.05
```

The F-test statistic is F=0.29, lower than the table value and with a p-value of 0.91. The individual 95% c.i. for the (transformed) means overlap each other, indicating that they do not differ. All of which indicates that the assumption of equal variance is reasonable.

We have reviewed the three regression model assumptions and they are not strongly disproved by the data (in real life where perfect models are rarely found such is the case). Therefore we will assume they are valid and proceed to use the model statistical results.

However, since the explanation provided by the linear regression ($100R^2 = 45.5\%$) is not very high, we will also try a quadratic regression, to see whether we can improve in the model fit (or explanation). Such is the scheme for multiple linear regression which is an extension of the simple one when there are two or more "predictor" variables, X_1, X_2, etc.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots \beta_k X_{ik} + \varepsilon_i \quad ; \quad 1 \leq i \leq n$$

As before, ε_i is the error term, distributed Normally with mean 0 and variance σ^2. The β_j ; $0 \leq j \leq k$, are again regression coefficients. In general, regression requires three or more levels of measurements for the predictor variable X. Hence, if there are less than three levels of X, then we cannot proceed, until more data (levels) are gathered. If there are enough levels, then we can fit a multiple regression model. The dependent or response variable Y is now a function of k predictor independent variables $X_1, \dots , X_k$.

Such regression models are adequate if they are statistically significant. This occurs when the null hypothesis $H_0:\beta_1=\beta_2=\dots=\beta_k=0$ is rejected (i.e. one of the regression coefficients is not zero). Not all independent variables, however, need to be statistically significant (e.g. it is enough that $\beta_j \neq 0$ for some j). Some predictor variables (X_j) may be highly significant (i.e. have a coefficient $\beta_j \neq 0$) while others may be redundant (i.e. not significant or $\beta_j=0$). The multiple regression model as a whole has to remain statistically significant. To choose the adequate subset of regressors X, we use variable selection methods. They weed out redundant variables and the resulting regression model improves its efficienty.

If there exist four or more levels of measurements for the independent variable (X) then it is possible to fit a quadratic regression model to the data. Its equation is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \quad ; \quad 1 \leq i \leq n$$

We now develop this (multiple regression) model, with the tensile strength example. Here, independent variable X_2 corresponds to the square of the first independent variable, i.e. X_i^2. The regression results obtained are:

```
                 matstr = 338.03 - 12.1 damage + 0.63 dam-2

Predictor        Coef        Stdev    t-ratio         p
Constant       338.03        11.92      28.35     0.000
damage        -12.070        7.821      -1.54     0.134
dam-2           0.627        1.093       0.57     0.570
```

This second (quadratic) model does not significantly improve on the first one (linear regression). The model fit ($100R^2 = 46.1\%$) is barely larger than the previous fit (45.5%) and the regression coefficient t-tests (t_{b1}=-1.54 and t_{b2}=0.57) are now no longer statistically significant (0.13 and 0.57) but the whole quadratic regression remains significant. This indicates that it includes redundant variables. This modeling attempt has not been productive and could have been identified from the residual analysis. Had the residual plot been concave (up or down) a quadratic regression would have been justified. In the present case, the residual plot pattern did not encourage trying a quadratic model.

However, when both, linear and quadratic regression models are statistically significant, we want to compare them to select the best of the two. This is accomplished by comparing the sums of squares of the residuals of both models, divided by their respective degrees of freedom. We recall that the residuals are the distances from the actual data points to the regression function estimates. The sum of their squares provides a good regression performance measure, since it measures the overall distance from the cloud of data to the postulated regression model.

We therefore compare the residual sum of squares obtained from the linear regression ($SSR_{L)}$ of the above example (6326.7) with that of the quadratic (SSR_Q) regression (6253.1). They have DF_L=29 and DF_Q=28 degrees of freedom. Their difference (divided by the reduction in d.f. and standardized by dividing by SSR_Q / DF_Q) provides a measure of how much explanation is gained by moving from one model to the other. In our case:

$$F = \frac{(SSR_L - SSR_Q)/(DF_L-DF_Q)}{(SSR_Q / DF_Q)} = \frac{(6326.7-6253.1)/(29-28)}{6253.1/28} = 0.33$$

Comparing 0.33 with the F-Table (critical) value F(α=0.05, dfnum=1, dfden=28) = 4.20 we see that the test result is not statistically significant. Therefore, the quadratic equation does not improve significantly our regression and is not selected as the best.

If selected, all regression assumptions (Normality, independence and equality of variance of the residuals) would have to be checked before using the quadratic regression results. If any of these procedures indicate that any model assumption has been violated (e.g. variances are not equal) the data should be transformed and the regression model should be recalculated. This procedure however, is beyond the scope of this SOAR.

Finally, and for comparison, let's assume that we are unaware of the relation between surface damage and tensile strength. Then, the best estimate of the mean tensile strength value for a ceramic material with surface damage value of two is still the tensile strength mean of 305.66, as indicated in the descriptive statistics above.

However, if the linear regression is obtained, this specific mean strength can be greatly improved by using the regression estimator instead, where:

$$\text{Strength} = 332 – 7.67 \text{ x damage} = 332 – 7.67 \text{ x } 2 = 332 – 15.34 = 316.66$$

Summarizing, regression models are based on two procedures. First, an optimization process selects a function such that the sums of the squares of the distances to each data point ($\sum\varepsilon^2_i$) is minimum. Then, a statistical model (i.e. distributional assumptions) is imposed on such distances (ε_i). If an invalid regression model is used (one where the assumption of independence, normality and equality of variance of the ε_i is not met) then the tests, the significance levels and confidence intervals derived (which are the statistical contributions provided by the regression model) are no longer valid or exact.

Restating, the regression point estimator for Y_i (given X_i) always applies (for it only depends on the optimization part of the regression procedure). However, if there are violations of the distributional assumptions of the residuals, the confidence (interval) estimation for Y_i and the probabilistic statements (e.g. tests of significance for the regression coefficients) are no longer exact nor valid. For these statistical results depend on the (now invalid) statistical assumptions of the regression model.

ANOVA (one way analysis of variance)

In materials data analysis we can work with a single or with several samples (batches). If we work with several samples, we want to pool them together, if possible. We can pool the samples and work with the combined data set as if it were one, if all the data come from the same population. We use the multiple sample A-D GoF test to assess the null hypothesis H_0, that the data (i.e. all samples) come from the same population. However, if A-D rejects H_0 then the observations become bivariate data. For, now each data point $P_{i,j}$ implicitly provides two pieces of information (i, Y_i): its batch or sample number and its materials property measurement (e.g. tensile strength).

ANOVA is the procedure used to assess whether, say, k independent batches of n elements each have the same mean, or whether the group means differ. The assessment is made via comparing two estimates of the variance. One estimate is obtained using the within groups variance estimator. The other one is obtained using the between groups estimator. If all k group means are equal, then these two variance estimators are close (for both estimate the same variance parameter) and their ratio is unit. If group means differ, the ratio of these two variance estimators (between and within groups) will be greater than unit. The one-way ANOVA or Analysis of Variance model is:

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \; ; \quad 1 \le i \le n \; ; \; 1 \le j \le k$$

where α_j is the contribution of the jth sample (group) to the general mean μ, and ε_{ij} is the error term, a R.V. distributed Normally, with mean 0 and variance σ^2. Under H_0, all group means (i.e. $\mu_j = \mu + \alpha_j$) are equal, hence all $\alpha_j = 0$; $1 \le j \le k$. If the null hypothesis H_0 is rejected, then at least one group effect (α_j) differs from zero.

As with regression, there are three key model assumptions: that errors are independent, Normal and with the same variance σ^2. A crucial ANOVA assumption is that all group variances are equal. This assumption must be carefully assessed before accepting the

ANOVA results. If this test fails (i.e. there is reason to believe that not all groups have the same variance σ^2) then data transformation or other procedures must be used.

Another important ANOVA model consideration is the number of data points (n_j ; $1 \leq j \leq k$) in each group or batch. ANOVA works better under “balanced” designs (i.e. $n_j = n$). This means that all (k) groups have equal size n. One way to visualize why this is so, is to think of the sample size n as the amount of information, of the k groups as informants and of the statistical test as an assessment procedure based on the information provided by k different informants. Optimally, we would like to give equal weight to all informants’ contribution. Therefore, it is better not to rely on more information from some (possibly biased) informants, over the others. This forces all batches to be of the same size.

To illustrate the implementation of one-way ANOVA models we will use a real data set taken from Example #2 of RECIPE. This is the NIST regression program developed for obtaining allowables for materials data, which is available free from the NIST Web Site (http://www.itl.nist.gov./div898/software/recipe/ex2.dat/). The data consists of 31 tensile strength observations from 6 independent batches. It is given below:

```
  328.117   334.767   347.783   346.266   338.731   297.039   293.460
  308.042   326.486   318.130   309.049   337.093   317.732   321.429
  317.265   291.888   297.694   327.397   303.863   313.098   323.277
  312.974   324.519   334.596   314.946   322.719   291.121   309.785
  304.850   288.018   294.199
```

The descriptive statistic and boxplot for the tensile strength data values are:

```
                   N     MEAN    MEDIAN  STDEV  MIN     MAX      Q1      Q3
strength          31   316.01    317.27  16.62 288.02  347.78  303.86  327.40

                                             -------------------
                              --------------I         +        I------------------
                                             -------------------

                          ----+---------+---------+---------+---------+-------strength
                            288       300       312       324       336       348
```

Notice how both mean and median are close and tails are similar, suggesting a possible symmetric, unimodal parent distribution, close to the Normal. The one way ANOVA model was implemented for factor “batches” with six levels (batches 1 through 6):

```
ANALYSIS OF VARIANCE ON strength
SOURCE      DF        SS        MS        F        p
batches      5      4915       983     7.30    0.000
ERROR       25      3369       135
TOTAL       30      8284
```

The above ANOVA table displays an F-test result of 7.3. The F-test is the ratio of the bewteen (batches MS=983) to the within (error MS=135) means variance estimators and is highly significant (i.e. p-value is practically zero). This result indicates that batch means are different and batches cannot be pooled together. Below, we present graphical 95% confidence intervals for the six batch means:

```
                                   INDIVIDUAL 95% CI'S FOR MEAN
                                   BASED ON POOLED STDEV
 LEVEL      N      MEAN     STDEV  -------+---------+---------+--------
     1      5    339.13      8.16                         (-----*----)
     2      6    308.70     12.44        (----*----)
     3      5    317.08     16.24            (-----*----)
     4      5    313.07     12.56          (-----*----)
     5      5    321.95      8.61               (----*----)
     6      5    297.59      9.31  (-----*----)
                                   -------+---------+---------+--------
POOLED STDEV =    11.61                 300       320       340
```

Notice how batches 2 through 5 are close but batches 1 and 6 are not. Standard deviations are also close. However, before accepting any of these ANOVA results, we first need to check the model assumptions. To do so we plot the standardized residuals vs. the fits:

```
 -1.03922  -0.41186   0.81605   0.67292  -0.03790  -1.10020  -1.43786
 -0.06216   1.67789   0.88952   0.03282   1.88788   0.06136   0.41016
  0.01733  -2.37673  -1.45015   1.35202  -0.86821   0.00306   0.96330
 -0.84686   0.24228   1.19297  -0.66087   0.07248  -0.61070   1.15003
  0.68443  -0.90344  -0.32032
```

```
                N     MEAN   MEDIAN  STDEV   MIN    MAX     Q1     Q3
Sres3          31    0.000    0.017  1.000 -2.377  1.888  -0.847 0.816
```

```
         -
         -                 x         x
      1.5+                      x
         -   x                             x
 Sres3   -                 x    x                                x
         -   x                                                   x
         -                           x     x
      0.0+                 2    x    2     x                     x
         -   x                                                   x
         -   x                             x
         -   x                  x          x                     x
         -                 x
     -1.5+                 x    x
         -
         -
         -                           x
         -
           +---------+---------+---------+---------+---------+---FITS3
       296.0     304.0     312.0     320.0     328.0     336.0
```

The above residual plot shows a stable (parallel) pattern within 2 standard deviations of the mean (zero) as expected from the Normal Standard residuals. In addition, the AD GoF Normality test statistic was 0.183, with a p-value of 0.9. Therefore, we will assume the Normality of residuals, as required by the model.

Had there been problems with the residuals (e.g. the above pattern was not stable) one can also plot the residuals versus the factors (e.g. batch) to assess potential problems:

```
            -
            -               x         x
         1.5+                                   x
            -                                             x         x
    Sres3   -     x         x                   x
            -     x                                                 x
            -                         x                   x
         0.0+     x         2         2         x         x
            -     x                                                 x
            -                                             x         x
            -     x                             x         x         x
            -               x
        -1.5+               x                   x
            -
            -
            -                         x
            -
              ----+---------+---------+---------+---------+---------batch
                1.0       2.0       3.0       4.0       5.0       6.0
```

In practice, groups (batches) are often (as in the above example) of different sizes. To correct for this problem we can use the "effective" sample sizes (n') obtained via the formula: $n' = (N-n^{*})/(k-1)$; where $n^{*} = \sum n^{2}_{j} /N$; $N=\sum n_j$ and $1 \le j \le k$. When $n_j = n$ (i.e. all groups have the same size) then $n^* = n' = n$. When $n_j \neq n$ (group sizes differ) then $n' < n$ (i.e. the test procedure is less efficient, for the samples are sub optimal). In statistical analysis we strive to obtain the most efficient and unbiased assessment (test) from the data (information). Therefore, we try to obtain samples as close in size as possible.

Summarizing, the one way ANOVA model is used to compare means of different groups (levels) for one variable or factor, as done above. One can also compare several factors at different levels, known as two, three, etc. way ANOVAs. It is important and useful to first carefully plan how the ANOVA is going to be carried out, i.e. design of (statistical) experiments or DoE. This exercise extracts the maximum results out of the information obtained and should be performed by (or with the assistance of) a professional statistician. If groups differ, they cannot be pooled together into a single analysis. Methods for c.i. estimation of the differences between group mean are available. Further readings on this topic are found in references [8, 9, 10] of the appendix.

Non Parametric Alternatives to ANOVA

As seen above, ANOVA requires that the data (or the residuals) are independent and Normally distributed, with the same variance, among other assumptions. When the Normality assumption is not appropriate, one alternative is to implement a K-sample Anderson Darling (AD) test. This distribution free procedure assesses whether k different samples (batches) come from the same distribution or not, without assuming that the parent distribution is Normal (nor any other specific distribution).

The K-sample AD non parametric procedure, described in Section 8.3.2.2 of [7], tests the hypothesis H_0 that the populations from which the samples come from are identical. It only assumes the independence of the samples and (optimally) that the measurements are continuous (though this is not necessary) so there are no ties among their values

This freedom allows its application as a screening test (following handbooks [6 and 7]) for possibly pooling the batches. When the K-sample AD test rejects hypothesis H_0 then the different batches cannot be pooled together. The ANOVA procedure above developed is then applied to aid in obtaining the desired allowables.

To illustrate the use of the K-sample AD test, we use a subset of the same data set above, composed of the last three tensile strength batches. First, lets combine and sort these three tensile strength samples (batches) resulting in a single and sorted sample of N=31 data points. In the combined sample, N will be the total number of data points and L will be the total number of distinct or different data points. Only if there are no ties (as now occurs) is L=N. Denote by Z_i (for $1\leq i \leq L$) the L rearranged (sorted) distinct values in the combined sample. Notice the Z_i are all different and increasing.

Denote by h_j the number of data points in the combined, sorted sample that are equal to the value Z_j (if there are no ties, h_j is unit). Denote by H_j the number of values, in the combined sample, that are strictly less than Z_j (including ties in those smaller values) plus one half of the number of values that are equal to Z_j (i.e. tied with this value). Let F_{ij} denote the number of values, in the ith sample, that are strictly less than Z_j (including ties) plus one half of the number of values in the ith sample equal to Z_j (i.e. tied with it).

The K-Sample Anderson Darling statistic is defined:

$$\text{ADK} = \frac{N-1}{N^2(k-1)} \Sigma_i \left\{ \frac{1}{n_i} \Sigma_j h_j \frac{(NF_{ij} - n_i H_j)^2}{H_j (N-H_j) - Nh_j /4} \right\}$$

Where the i index runs from 1 to k and k is the number of groups or batches, where the j index runs from 1 to L, and where, as before, L is the number of distinct values of Z_j.

Under the null hypothesis H_0 (of no difference between the population distributions) the ADK statistic has a known mean and variance. Then, its distribution can be approximated by a well known one and percentiles or critical values for the test, can be obtained.

To illustrate its implementation, we will start the calculations for the last three batches of the RECIPE Example #2 above. Calculations are very tedious without a computer program, even for very small values. The three last batches have five observations each. They also have (see ANOVA) obviously different means and thus differ significantly.

The data in question (ksamp) from the last three batches of the ANOVA example, their sequence (N) and batch group (gord) numbers and their corresponding reordering after sorting in ascending order (ordsamp and newgord) are given below:

N	gord	ksamp	ordsamp	newgord
1	4	297.694	288.018	6
2	4	327.397	291.121	6
3	4	303.863	294.199	6
4	4	313.098	297.694	4
5	4	323.277	303.863	4
6	5	312.974	304.850	6
7	5	324.519	309.785	6
8	5	334.596	312.974	5
9	5	314.946	313.098	4
10	5	322.719	314.946	5
11	6	291.121	322.719	5
12	6	309.785	323.277	4
13	6	304.850	324.519	5
14	6	288.018	327.397	4
15	6	294.199	334.596	5

For this example, N=L=15 (for there are no ties); k=3 (there are three groups or batches); n_i = n = 5 (all batches have the same number of observations) and h_j=1 (there are no ties). In addition, H_j will always be the number of values that are strictly less than Z_j plus 0.5 (since h_j=1 and we add one half of unit) and F_{ij} is the number of values in the ith sample that are strictly less than Z_j plus one half of unit, by the same reason as before. Hence:

$$\text{ADK} = \frac{15-1}{15^2(3-1)} \Sigma_i \left\{ \frac{1}{5} \Sigma_j 1 \frac{(15F_{ij} - 5\,H_j\,)^2}{H_j\,(15\text{-}H_j)\text{-}15/4} \right\}$$

Once the above equation is set up, an iterative computer program will provide the results for the ADK statistic. Program SBMP17, a computer code for calculating Statistically Based Material Properties for MIL HDBK 17 that includes the AD statistic, is available from NIST (through their above mentioned Web Page).

If all batches have the same size then we compare the statistic ADK with the critical values in Table 8.5.6 of [7]. If the ADK statistic is smaller than the table value, then we reject hypothesis H_0 at significance level α=0.05, and conclude that the batches were drawn from different populations. If batch sizes differ, then we calculate the variance for statistic ADK, as well as the critical value to test it, both from the formulas in section 8.3.2.2 of [7]. Then, we compare the ADK statistic with this critical value (instead of with the Table 8.5.6 value) in the same way as explained above.

Summary.

In this chapter we have discussed some fundamental problems related to multivariate data analysis and multivariate statistics. When data consist of observations yielding more than one measurement (or pieces of information) they are called multivariate data. Their data analysis is more complex but also more informative than the univariate, especially if there are associations among the different variables that integrate the multivariate information vector. For, some of these variables may be easier, faster or cheaper to measure than the others. In such cases, we take advantage of the existing relationships to measure some variables and use these results to estimate the measurements for the others.

If the multivariate observations are qualitative then we implement categorical data analysis methods such as contingency tables, to assess possible associations. If data are quantitative, we can implement methods that quantify this associations, such as correlations and regressions. Finally, if data are both qualitative and quantitative we can implement other methods such as ANOVAs (analysis of variance). But, whichever the case we are dealing with (and whichever method we implement) it is extremely important to always assess whether the model assumptions have been met, before we are able to use any of the model derived statistical results.