Analysis of Variance

Analysis of variance is another commonly used technique for data analysis. This method refers to the partition of the total sum of squares into sum of squares due to effects (or treatments).

One Way ANOVA

The simplest kind of experiments are those in which a single continuous response variable is measured a number of times for each of several levels of some experimental factor. For example, consider the data in Table which consists of numerical values of blood coagulation times for each of four diets. Coagulation time is the continuous response variable, and diet is a qualitative variable, or factor, having four levels: A, B, C, and D. The diets corresponding to the levels A, B, C, and D were determined by the experimenter. Your main interest is to see whether or not

A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
65	68	63	
66	68	64	
63			
59			

Table 1: Blood coagulation times for four diets.

the factor diet has any effect on the mean value of blood coagulation time. The experimental factor, diet in this case, is often called the treatment.

In order to analyze the data, you need to get it into a form that `S-Plus` can use for the analysis of variance. You do this by setting up a data frame. First create a numeric vector `coag`:

```
> coag <- scan()     #scan the data
```

```
1: 62 60 63 59
5: 63 67 71 64 65 66
11: 68 66 71 67 68 68
17: 56 62 60 61 63 64 63 59
25:
> coag
 [1] 62 60 63 59 63 67 71 64 65 66 68 66 71 67 68 68 56 62 60 61 63 64 63 59
> diet <- factor(rep(LETTERS[1:4],c(4,6,6,8)))  #create a factor
> diet
 [1] A A A A B B B B B B C C C C C C D D D D D D D D
> coag.df <- data.frame(diet,coag)              #create a data frame
> coag.df
   diet coag
 1    A   62
 2    A   60
 3    A   63
 4    A   59
 5    B   63
 6    B   67
 7    B   71
 8    B   64
 9    B   65
10    B   66
11    C   68
12    C   66
13    C   71
14    C   67
15    C   68
16    C   68
17    D   56
18    D   62
19    D   60
20    D   61
21    D   63
22    D   64
23    D   63
24    D   59
```

The first step in the data analysis is to graphically explore whether or not there are differences among the factor levels. Figures 1 and 2 display the means and medians of each treatment group and the corresponding boxplots. The vertical line of Figure 1 is the overall mean (median) of the data. It should be clear that levels A and D form one group of levels while the other is formed from B and C.

```
> par(mfrow=c(1,2))
> plot.design(coag.df)
> plot.design(coag.df, fun= median)
```

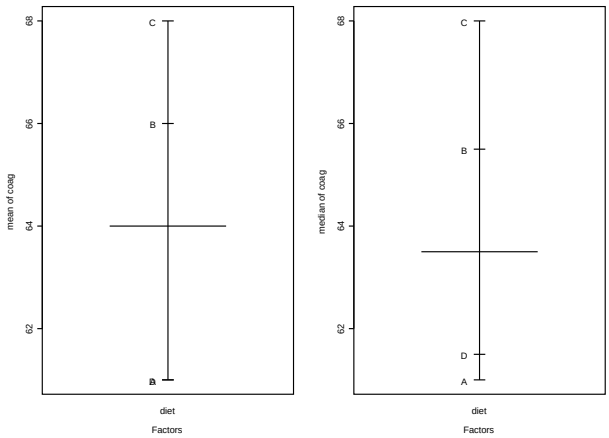

Figure 1: Treatment means and medians

```
> par(mfrow=c(1,1))
> plot.factor(coag.df)
```

You carry out the analysis of variance with the function `aov`

```
> aov.coag <- aov(coag ~ diet, coag.df)  #the same formula as in regression
> aov.coag                               # display the results
Call:
   aov(formula = coag ~ diet, data = coag.df)

Terms:
                diet Residuals
 Sum of Squares  228       112
Deg. of Freedom    3        20

Residual standard error: 2.366432
Estimated effects may be unbalanced
> summary(aov.coag)                     #ANOVA table
          Df Sum of Sq Mean Sq  F Value         Pr(F)
     diet  3       228    76.0 13.57143 0.00004658471
Residuals 20       112     5.6
```

Notice that the first argument to `aov` above is the formula `coag ~ diet` which is a symbolic representation of the one-way layout model equation. The second argument is the data frame you created, `coag.df`. To display the ANOVA table, we used the function `summary`. The result is significant leading to the conclusion that there are differences between the diets.

Some other useful commands are given by the following:

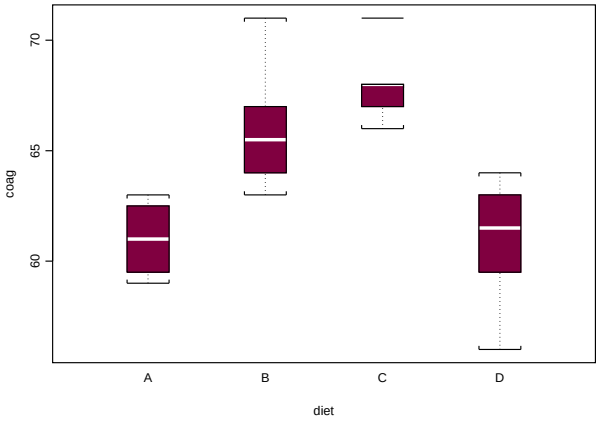

Figure 2: Boxplots for each treatment

```
> fitted.values(aov.coag)     #predicted values
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 61 61 61 61 66 66 66 66 66 66 68 68 68 68 68 68 61 61 61 61 61 61 61 61
> hist(resid(aov.coag))      #histogram of residuals
> qqnorm(resid(aov.coag))    # QQ-plot of residuals
> qqline(resid(aov.coag))
> plot(fitted(aov.coag), resid(aov.coag))  #plot predicted vs residuals
```

Multiple Comparisons

The previous analysis indicates that there are differences between the levels of the factor diet. Hence it is of practical interest to identify these differences. The function `multicomp` has been designed especially for this task.

```
> mca.coag <- multicomp(aov.coag, focus ="diet")
> plot(mca.coag)
> mca.coag

95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.7987
response variable: coag

intervals excluding 0 are flagged by '****'
```

```
      Estimate Std.Error Lower Bound Upper Bound
A-B -5.00e+000      1.53       -9.28      -0.725 ****
A-C -7.00e+000      1.53      -11.30      -2.720 ****
A-D -1.64e-014      1.45       -4.06       4.060
B-C -2.00e+000      1.37       -5.82       1.820
B-D  5.00e+000      1.28        1.42       8.580 ****
C-D  7.00e+000      1.28        3.42      10.600 ****
```

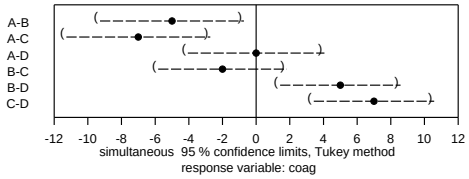

Figure 3: 95 % simultaneous confidence intervals for all mean differences

As the output and plot in Figure 3 indicate, this default call to multicomp has resulted in the calculation of simultaneous 95% confidence intervals for all pairwise differences between diet means, based on the Tukey's method. Hence treatments A and D form one group while B and C another, as alluded earlier.

In general the `multicomp` function can calculate critical points for simultaneous intervals or bounds by the following methods:

- Tukey (`method = tukey`),
- Dunnett (`method = dunnett`),
- Sidak (`method = sidak`),
- Bonferroni (`method = bon`),
- Scheffe (`method = scheffe`)