MA4605 Lab I

Part 1 Testing Assumptions for ANOVA.

The three assumptions that require testing are as follows:

- The samples have equal variance
- The residuals are normally distributed
- The residuals have mean of zero, and have a constant variance.

We will test the validity of these assumptions for the ANOVA model fitted to the Paracetamol example data in previous labs.

```
a = c(84.55, 84.61, 84.26, 84.36, 84.66, 84.31, 84.65, 84.41, 84.52, 84.44);
b = c(84.12, 84.04, 83.95, 84.51, 84.08, 84.07, 84.35, 83.99, 84.25, 84.14);
c = c(84.44, 84.48, 84.14, 84.17, 84.31, 84.60, 84.44, 84.24, 84.64, 84.47);
d = c(84.05, 84.14, 84.53, 84.07, 84.45, 83.95, 84.10, 84.29, 84.13, 83.98);
e = c(84.09, 84.53, 84.60, 84.48, 84.42, 84.57, 84.35, 84.30, 84.37, 84.63);

y = c(a, b, c, d,e);
group = rep(c("a","b","c","d","e"), each = 10);
group = factor(group,c("a","b","c","d","e"));

Model1 = aov(y~group);

summary(Model1);
plot(Model1);
```

Bartlett's test: Bartlett's test is used to test if multiple samples have equal variances.
Equal variances across samples is called ***homogeneity of variances***.
Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

The null hypothesis is that the samples have equal variance.
The alternative hypothesis is that at least one sample has a significantly different variance.

To implement the Bartlett test in R, we simply use the command bartlett.test(), with the model specification as an argument.

We can also use boxplots to implement a graphical complement to the procedure.

```
boxplot(y~group);

bartlett.test(y~group);
```

What is your conclusion for this procedure?

Testing the residuals

To test the other two assumptions, we can use the diagnostic plots provided by the plot() command, and also by performing the Shapiro-Wilk test for residuals.

```
Residuals=resid(Model1);
plot(Residuals);
shapiro.test(Residuals);

plot(Model1);
```

Shapiro-Wilk Test

Are the residuals normally distributed? Base your conclusion on the p-value?

Diagnostic plot 1

Interpretation: Does the red trend-line stay consistently around the "zero level"?
Are residuals uniformly distributed across the plot? Or is there an indication of heteroscedascity?

Which data points are mentioned for further attention by this plot?

Diagnostic plot 2

Interpretation: This plot is a QQ plot used to test the normality of the residuals?
Do the points follow the diagonal trend-line?
Which data points are mentioned for further attention by this plot?

(For diagnostic plots 3 and 4 - we will just consider whether or not any new data points are mentioned.)

A Counter Example

A sixth sample (***f***) is added to the analysis. We will run the analysis again to consider how this affects the analysis.

You can update the data using the following code.

```
f = c(83.79, 84.23, 83.69, 84.48, 83.88, 84.57, 84.85, 84.50, 84.77, 84.93);
y = c(a,b,c,d,e,f);
group = rep(c("a","b","c","d","e","f"), each = 10);
group = factor(group,c("a","b","c","d","e","f"));

boxplot(y~group);
bartlett.test(y~group);
```

Perform the Bartlett test again for the updated data.

Part 2: Factorial design and Interaction Plots

Question 6

An experiment is run on an operating chemical process in which the aim is to reduce the amount of impurity produced. Three continuous variables are thought to affect impurity, these are concentration of NaOH, agitation speed and temperature. As an initial investigation two settings are selected for each variable these are

Factor:	level -1	level +1
concentration of NaOH	40%	45%
agitation speed (rpm)	10	20
temperature (^oF)	150	180

Readings were recorded of the impurity produced from the chemical process for each combination of the levels of these factors, and each combination was tested in duplicate.

Conc NaOH	agitation	temperature	Impurity replicate 1	Impurity replicate 2
-1	-1	-1	38	30
1	-1	-1	40	62
-1	1	-1	23	45
1	1	-1	25	30
-1	-1	1	85	89
1	-1	1	56	75
-1	1	1	20	53
1	1	1	20	20

To implement this in R, we use the following code.

```
A = c("L","H","L","H","L","H","L","H","L","H","L","H","L","H","L","H");
B = c("L","L","H","H","L","L","H","H","L","L","H","H","L","L","H","H");
C = c("L","L","L","L","H","H","H","H","L","L","L","L","H","H","H","H");

A=factor(A);B=factor(B);C=factor(C);

y=c(38, 40, 23, 25, 85, 56, 20, 20, 30, 62, 45, 30, 89, 75, 53, 20);

cbind(y,A,B,C);
```

Let us fit the model. Of specific interest today is the interaction terms? Are they significant?

```
Model2 = aov(Y~A*B*C);

summary(Model2);
```

Interaction plots

Sketch the following interaction plots? Are they parallel? Do they intersect? Does the interpretation accord with the model summary output?

```
interaction.plot(A,B,y);
interaction.plot(A,C,y);
interaction.plot(B,c,y);
```