session12 Assign2

ANUJ

a. What are the assumptions of ANOVA test it out?

Assumptions and Effects of Violating Assumptions

- Deviation from Normal Distribution
- Homogeneity of Variances
- Homogeneity of Variances and Covariances
- Sphericity and Compound Symmetry

DEVIATION FROM NORMAL DISTRIBUTION

Assumptions. It is assumed that the dependent variable is measured on at least an interval scale level . Moreover, the dependent variable should be normally distributed within groups.

Effects of violations. Overall, the *F* test is remarkably robust to deviations from normality . If the kurtosis is greater than 0, then the *F* tends to be too small and we cannot reject the null hypothesis even though it is incorrect. The opposite is the case when the kurtosis is less than 0. The skewness of the distribution usually does not have a sizable effect on the *F* statistic. If the *n*per cell is fairly large, then deviations from normality do not matter much at all because of the *central limit theorem*, according to which the sampling distribution of the mean approximates the normal distribution, regardless of the distribution of the variable in the population. A detailed discussion of the robustness of the *F* statistic can be found in Box and Anderson (1955), or Lindman (1974).

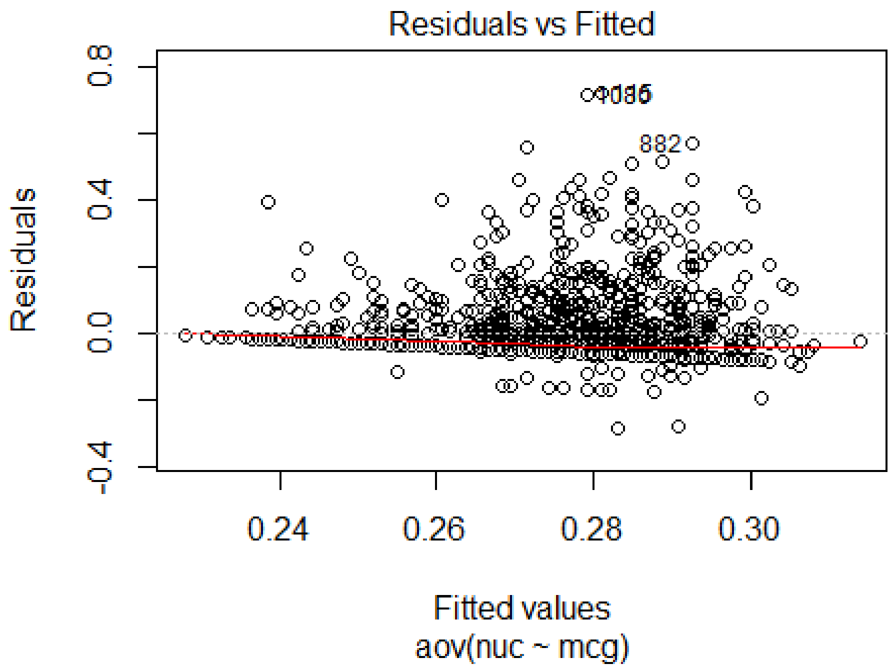
HOMOGENEITY OF VARIANCES

Assumptions. It is assumed that the variances in the different groups of the design are identical; this assumption is called the *homogeneity of variances* assumption. Remember that at the beginning of this section we computed the error variance (*SS* error) by adding up the sums of squares within each group. If the variances in the two groups are different from each other, then adding the two together is not appropriate, and will not yield an estimate of the common within-group variance (since no common variance exists).

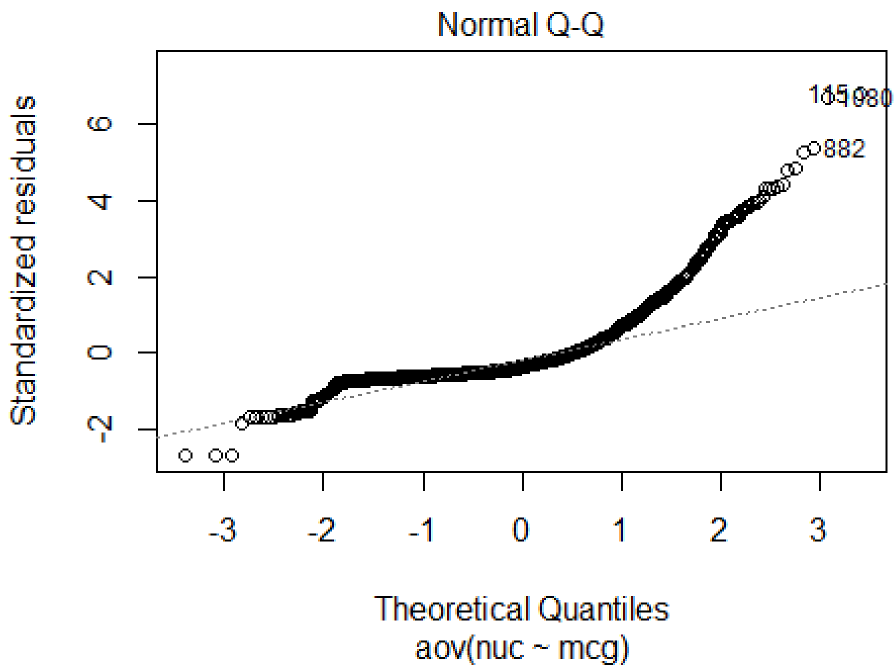
Special case: correlated means and variances. However, one instance when the *F* statistic is *very misleading* is when the means are correlated with variances across cells of the design. A scatterplot of variances or standard deviations against the means will detect such correlations. The reason why this is a "dangerous" violation is the following: Imagine that we have 8 cells in the design, 7 with about equal means but one with a much higher mean. The *F* statistic may suggest a statistically significant effect. However, suppose that there also is a much larger variance in the cell with the highest mean, that is, the means and the variances are correlated across cells (the higher the mean the larger the variance). In that case, the high mean in the one cell is actually quite unreliable, as is indicated by the large variance. However, because the overall *F* statistic is based on a *pooled* within-cell variance estimate, the high mean is identified as significantly different from the others, when in fact it is not at all significantly different if we based the test on the within-cell variance in that cell alone.

This pattern - a high mean and a large variance in one cell - frequently occurs when there are *outliers* present in the data. One or two extreme cases in a cell with only 10 cases can greatly bias the mean, and will dramatically increase the variance.

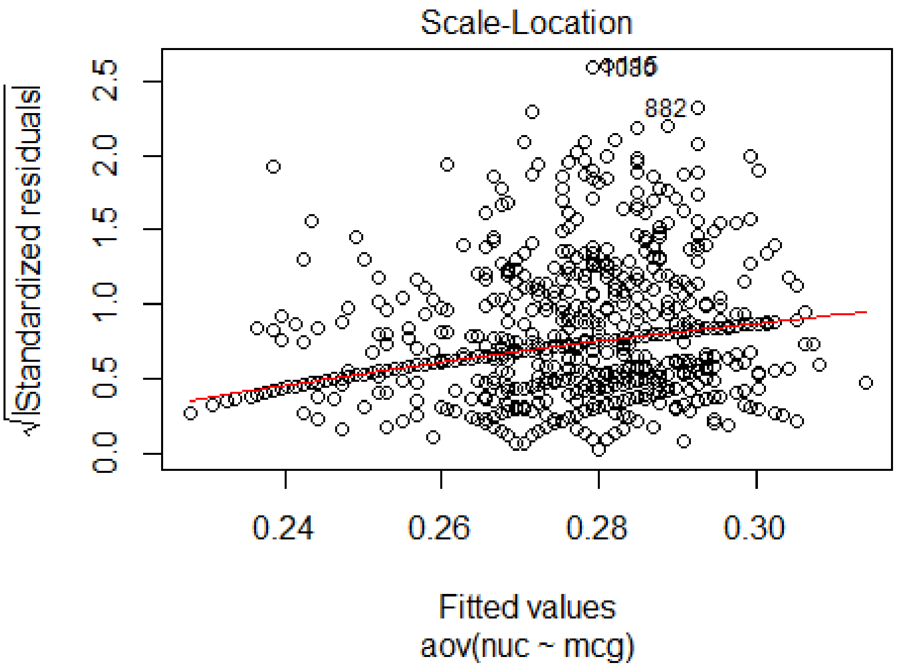
```r

# 1. Homogeneity of variances

# 2. Normality
#Normality plot of residuals. In the plot below, the quantiles of the residua
ls are plotted against the quantiles of the normal distribution. A 45-degree
reference line is also plotted.

#The normal probability plot of residuals is used to check the assumption tha
t the residuals are normally distributed. It should approximately follow a st
raight line.
plot(res.aov, 1)
```



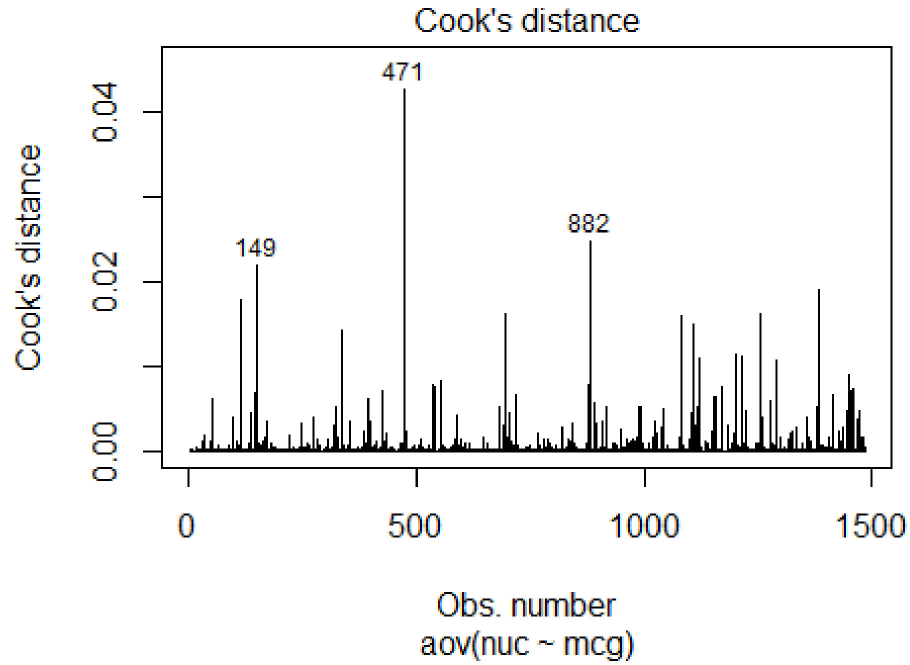
```r
plot(res.aov, 2)
```



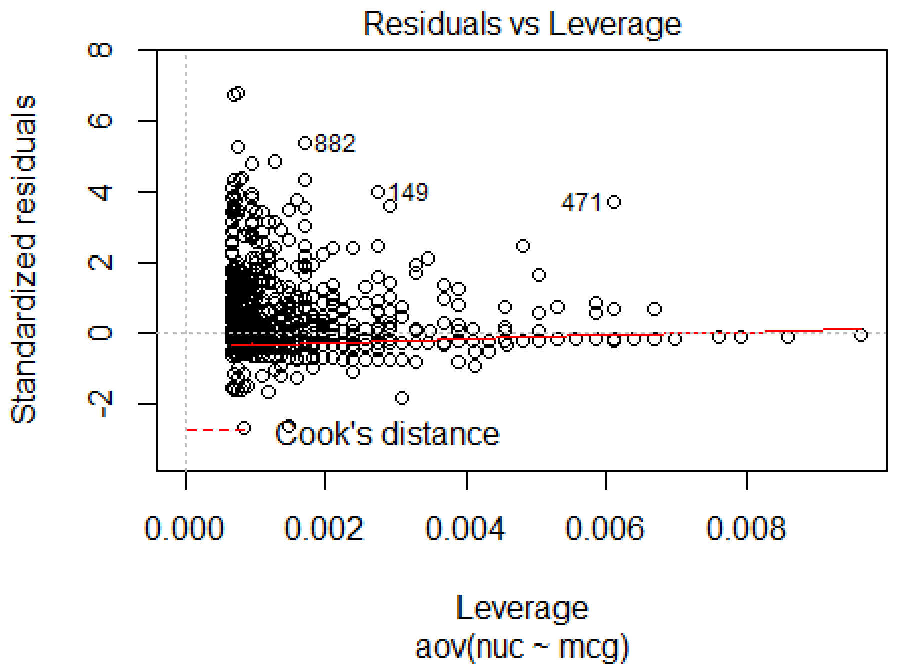
```r
plot(res.aov, 3)
```



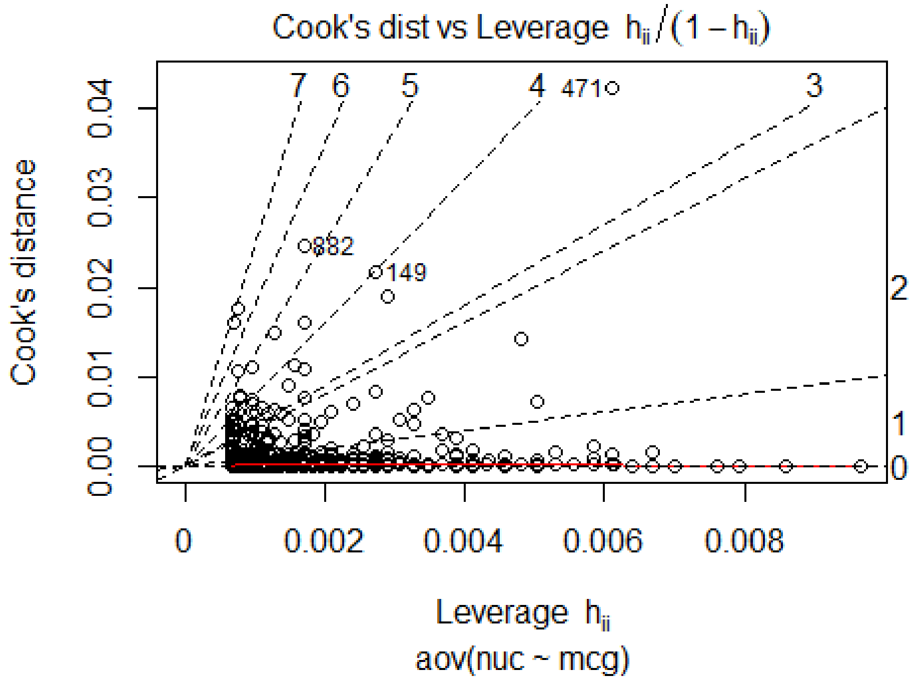
```
plot(res.aov, 4)
```



```
plot(res.aov, 5)
```



```r
plot(res.aov, 6)
```



b. Why ANOVA test? Is there any other way to answer the above question?

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group mean in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher.

The Analysis of Variance table is just like any other ANOVA table. The Total Sum of Squares is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the overall sample mean. The ANOVA table partitions this variability into two parts. One portion is accounted for (some say "explained by") the model. It's the reduction in uncertainty that occurs when the ANOVA model,

Yij = μ + αi + εij

is fitted to the data. The remaining portion is the uncertainty that remains even after the model is used. The model is considered to be statistically significant if it can account for a large amount of variability in the response.

Model, Error, Corrected Total, Sum of Squares, Degrees of Freedom, F Value, and Pr F have the same meanings as for multiple regression. This is to be expected since analysis of variance is nothing more than the regression of the response on a set of indicators definded by the categorical predictor variable.

The degrees of freedom for the model is equal to one less than the number of categories. The F ratio is nothing more than the extra sum of squares principle applied to the full set of indicator variables defined by the categorical predictor variable. The F ratio and its P value are the same regardless of the particular set of indicators (the constraint placed on the α-s) that is used.

Sums of Squares: The total amount of variability in the response can be written $\sum_{i,j}(y_{ij}-\bar{y}_{..})^2$, the sum of the squared differences between each observation and the overall mean. If we were asked to make a prediction without any other information, the best we can do, in a certain sense, is the overall mean. The amount of variation in the data that can't be accounted for by this simple method of prediction is the Total Sum of Squares.

When the Analysis of Variance model is used for prediction, the best that can be done is to predict each observation to be equal to its group's mean. The amount of uncertainty that remains is sum of the squared differences between each observation and its group's mean, $\sum_{i,j}(y_{ij}-\bar{y}_{i.})^2$. This is the Error sum of squares. In this outpur it also appears as the GROUP sum of squares. The difference between the Total sum of squares and the Error sum of squares is the Model Sum of Squares, which happens to be equal to $\sum_{i} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$.

Each sum of squares has corresponding degrees of freedom (DF) associated with it. Total df is one less than the number of observations, N-1. The Model df is the one less than the number of levels The Error df is the difference between the Total df (N-1) and the Model df (g-1), that is, N-g. Another way to calculate the error degrees of freedom is by summing up the error degrees of freedom from each group, n_i-1, over all *g* groups.

The Mean Squares are the Sums of Squares divided by the corresponding degrees of freedom.

The F Value or F ratio is the test statistic used to decide whether the sample means are within sampling variability of each other. That is, it tests the hypothesis H_0: $\mu_1 ... \mu_g$. This is the same thing as asking whether the model as a whole has statistically significant predictive capability in the regression framework. F is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability--that is, that all of the population means are equal--the F statistic follows an F distribution with *p* numerator degrees of freedom and *n-p-1* denominator degrees of freedom. The null hypothesis is rejected if the F ratio is large. This statistics and P value might be ignored depending on the primary research question and whether a multiple comparisons procedure is used

Other Methodology to test

```
# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )

##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.7479, p-value < 2.2e-16

kruskal.test(nuc ~ mcg, data = yeast)

##
##  Kruskal-Wallis rank sum test
##
## data:  nuc by mcg
## Kruskal-Wallis chi-squared = 126.7, df = 80, p-value = 0.0006882

#The classical one-way ANOVA test requires an assumption of equal variances f
or all groups.An alternative procedure (i.e.: Welch one-way test), that does
not require that assumption have been implemented in the function oneway.test
().
#The assumptions on which f-test relies are:

#The population is normally distributed.
#Samples have been drawn randomly.
#Observations are independent.
#H0 may be one sided or two sided.

oneway.test(nuc ~ LocalizationSite , data = yeast)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  nuc and LocalizationSite
## F = 23.555, num df = 9.000, denom df = 83.991, p-value < 2.2e-16

#Pairewise t-test
#The function pairewise.t.test() can be also used to calculate pairwise comparisons between group levels with corrections for multiple testing.

pairwise.t.test(yeast$nuc, yeast$LocalizationSite,
                 p.adjust.method = "BH")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  yeast$nuc and yeast$LocalizationSite
##
##     CYT     ERL     EXC     ME1     ME2     ME3     MIT     NUC
## ERL 0.90278 -       -       -       -       -       -       -
## EXC 0.26144 0.90278 -       -       -       -       -       -
## ME1 0.78608 0.89913 0.23251 -       -       -       -       -
## ME2 0.73980 0.99135 0.73170 0.61121 -       -       -       -
## ME3 0.40617 0.89683 0.09106 0.94425 0.34402 -       -       -
## MIT 0.16646 0.94425 0.73980 0.29390 0.90278 0.02147 -       -
## NUC < 2e-16 0.21645 3.5e-08 0.00037 7.2e-08 4.9e-10 < 2e-16 -
## POX 0.65421 0.90278 0.90278 0.48903 0.90278 0.34402 0.90278 0.00014
## VAC 0.90278 0.94425 0.65421 0.76810 0.90278 0.65421 0.89683 0.00014
##     POX
## ERL -
## EXC -
## ME1 -
## ME2 -
## ME3 -
## MIT -
## NUC -
## POX -
## VAC 0.87658
##
## P value adjustment method: BH

#The result is a table of p-values for the pairwise comparisons. Here, the p-values have been adjusted by the Benjamini-Hochberg method.
#One-Way ANOVA with Pairwise Comparisons. ... Whereas a one-way omnibus ANOVA assesses whether a significant difference exists at all amongst the groups, pairwise comparisons can be used to determine which group differences are statistically significant.

#Roughly, paired t-test is a t-test in which each subject is compared with itself or, in other words, determines whether they differ from each other in a
```

```
significant way under the assumptions that the paired differences are indepen
dent and identically normally distributed.
```

R Markdown

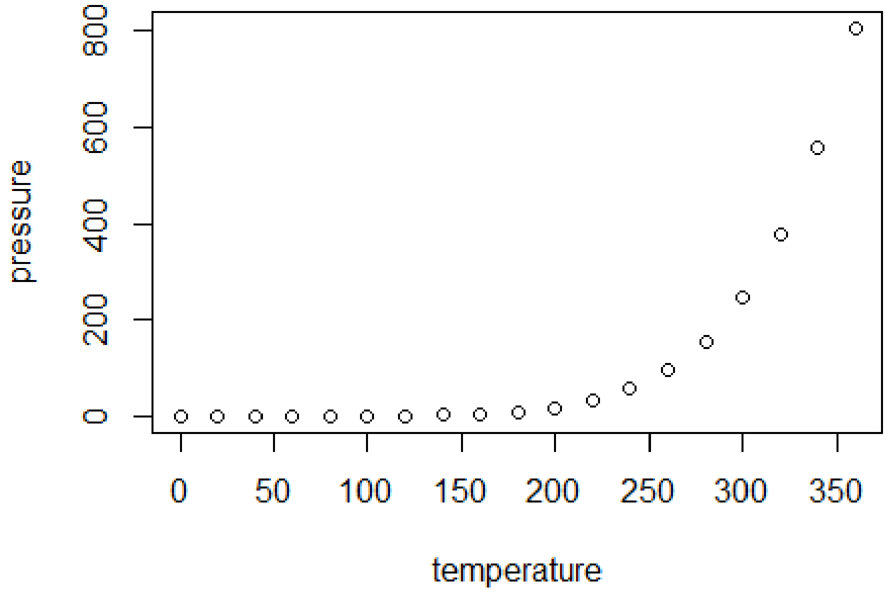
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)

##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.