Lab Tutorial 3: General Linear Model and ANOVA

The General Linear Model

Most statistical tests (including t- and F- tests) can be understood as variations and extensions of the General Linear Model. It is important to understand the concept of *model fitting* in which an observed set of data can be conceptualized as resulting from the *linear* combination of a set of predictor variables. You can think of modelling as building a template of what you would expect the data to look like, and then assessing the degree to which the observed data fit this template. The general form of the General Linear Model is:

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + ... + \beta_k X_{k_i} + \epsilon_i$$

where Y_i is the observed score for individual i, X_{k_i} is a predictor variable for individual i in the k_{th} condition, β_0 is the model intercept, β_{k_i} is the model coefficient that indicates how much change in Y is expected from being in condition k, and ϵ_i is the deviation of the observed score from the predicted score. This model is considered *linear* because the outcome variable Y_i is predicted from the sum of independent predictor variables and the "error" associated with the predicted score. The β components of the model are known as the model *parameters* and are estimated from the observed data.

Significance Testing: Assessing The Relative Fit of Models

You can think of NHST significance testing as assessing the degree to which the addition of a model parameter(s) yields a significantly greater fit to the data than a model without this parameter(s). In other words, we are interested in comparing the relative fit of two different models - a *restricted model* which does not contain the parameter(s) of interest, and a *full* model which is identical to the restricted model, except for the addition of the parameter(s) of interest. The primary question is then: **does the full model provide a significant improvement in fit relative to the restricted model?** This can be assessed by examining the reduction in prediction error associated with the full model relative to the restricted model. The following example will illustrate this concept.

In a between-subject design with three levels ($k = 3$) of an independent variable, we can construct the following restricted and full models:

Restricted Model:

$$Y_{ij} = \mu + \epsilon_{ij}$$

Full Model:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

where μ is the model intercept (grand mean of all observations), and α_j is the difference between the mean of condition $\bar{Y}_j$ and the grand mean (μ). This method of constructing the models is known as *effects coding* where the α_j parameters being estimated from the data represent how far each condition mean ($\bar{Y}_j$) deviates from the grand mean (μ).

$$\alpha_j = \bar{Y}_j - \mu$$

The sum of α_j's are constrained to zero:

$$\sum_{j=1}^{k} \alpha_j = 0$$

Once we have specified our restricted and full models, we can test to see whether the full model significantly reduces prediction error relative to the restricted model. This is tantamount to testing the null hypothesis

(H_0) that $\alpha_1 = \alpha_2 = \alpha_3 = 0$, or equivalently that $\mu_1 = \mu_2 = \mu_3$. The alternative hypothesis (H_a) is that at least one $\alpha_j \neq 0$, or equivalently that at least one population mean (μ) is different than the others.

Now we need to quantify the relative goodness-of-fit of the models. To do this we can calculate a test statistic, F that captures the difference in fit between the restricted and full models:

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F}$$

where E_F and E_R are the *sum of squared residuals* (or prediction error) associated with the full and restricted models respectively, and df_F and df_R are the degrees of freedom associated with the full and restricted models, respectively. It can be shown that:

$$(df_R - df_F) = k - 1$$

$$df_F = N - k$$

Under the null hypothesis (H_0), the F-statistic follows a characteristic distribution defined by the degrees of freedom in the numerator and the denominator. When all $\alpha_j = 0$, large values of F should be relatively rare. If our F-statistic is sufficiently large such that it would only occur when H_0 is true <5% of the time, then we reject H_0.

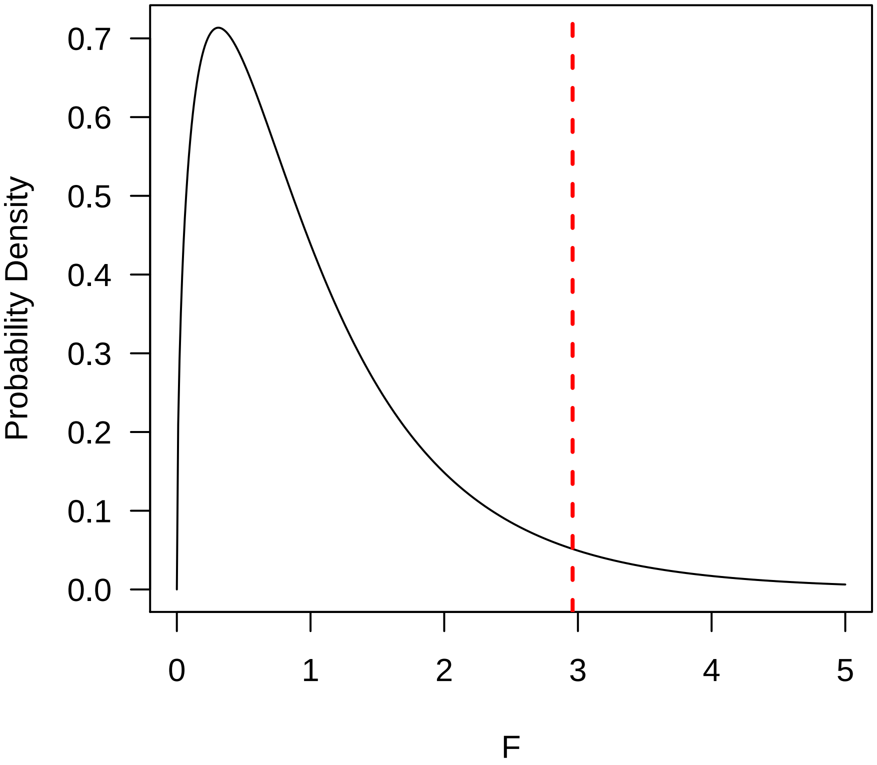


Figure 1: F-distribution with df1=3, df2=27. Dotted line represents 95th percentile, such that 95% of F-values lie to the left of this line.

Below is a graphical depiction of what E_F and E_R represent. First, let's consider the restricted model:

In this experiment with $k = 3$ levels, and $n_j = 5$, the restricted model aims to predict each individual's score using only the grand mean (μ; dotted red line). The residuals (E_R), are calculated by summing the squared difference between each individual's score, Y_{ij}, and the grand mean, μ, and represent the prediction error associated with the restricted model (see Figure 2).

Now consider the predictions from the full model (see Figure 3):

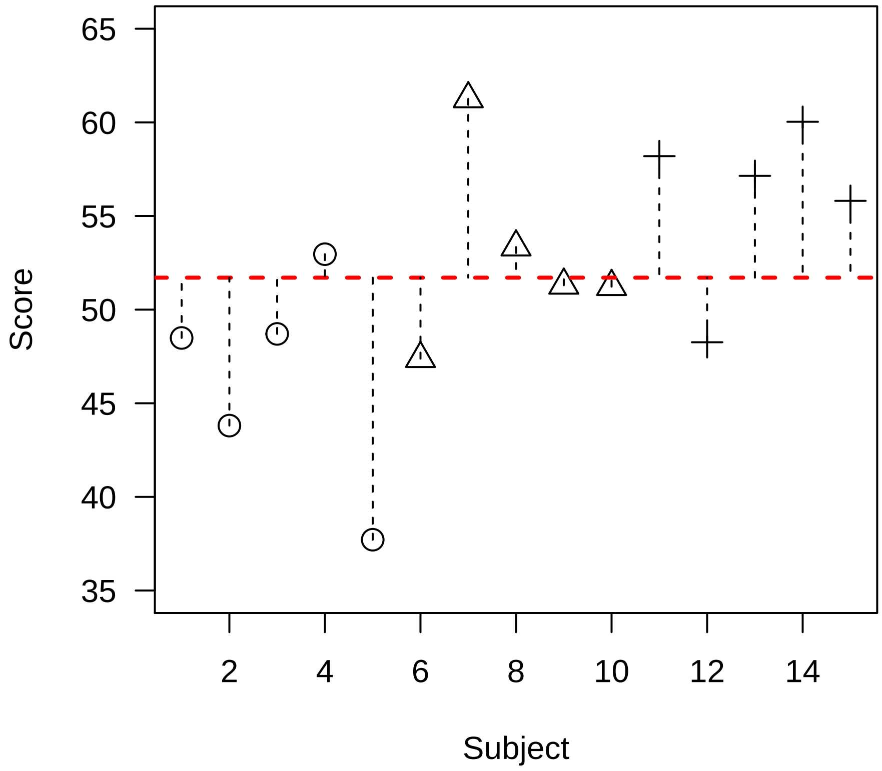


Figure 2: Predictions of individuals scores from the restricted model. The black dotted lines represent the residuals, which quantify the amount of prediction error associated with this model.

Inclusion of the α_j parameters in the full model allows our predictions to be more specific - that is, we can now use information about group membership to predict an individual's score. For individual i in group j, we predict that their score will be the sum of the grand mean (μ) plus the effect of being in group j (α_j). The remaining difference represents prediction error (residuals; ϵ_{ij}). The key is whether this prediction error is smaller in the full model (E_F) than in the restricted model (E_R), which we can formally assess using an F-test. Let's learn how to do this in R.

Analysis of Variance As Comparison of Nested Models

Example #1

In R we can build the restricted and full models using the **lm** command. Let's read in a .txt file containing data to illustrate how to use this command. In this example, there are $k = 3$ levels of one independent variable with 5 subjects in each of the 3 conditions.

```
options(contrasts=c("contr.sum","contr.poly")) # IMPORTANT!! Set sum-to-zero contrasts

my.data<-read.table(file="L3data1.txt",header=T)
my.data$Subj<-as.factor(my.data$Subj) # IMPORTANT!! Convert Subj variable to factor!!
str(my.data)
```

```
## 'data.frame':    15 obs. of  3 variables:
##  $ Subj     : Factor w/ 15 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Condition: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 2 2 2 2 2 ...
```

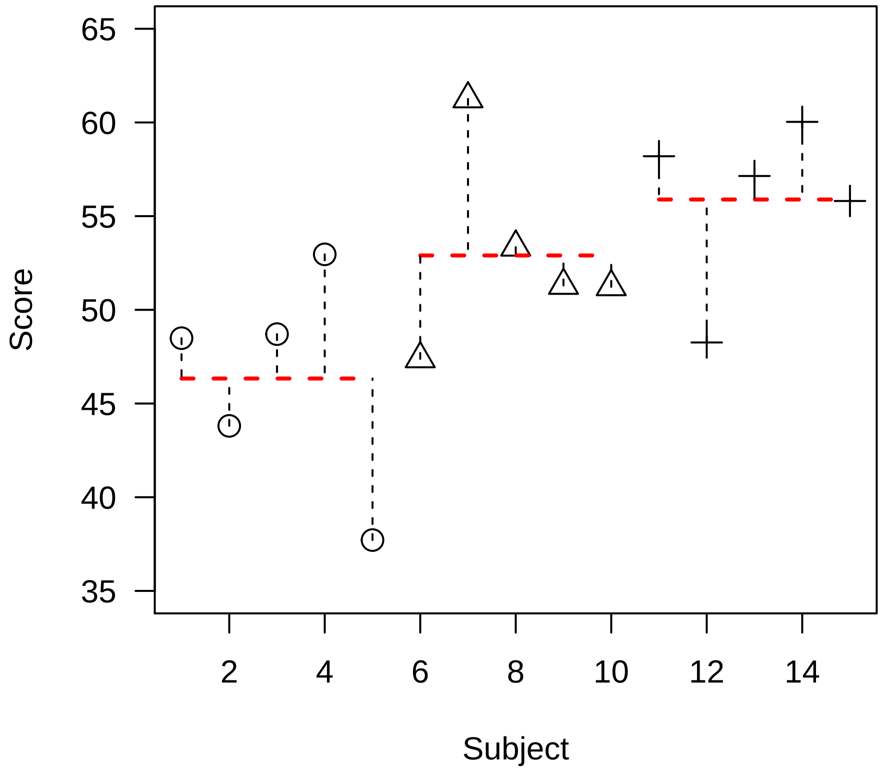


Figure 3: Predictions of individuals scores from the full model. The black dotted lines represent the residuals, which quantify the amount of prediction error associated with this model.

```
##  $ DV       : num  48.5 43.8 48.7 53 37.7 ...
```

```
model.restricted<-lm(DV~1,data=my.data) # Create restricted model
model.full<-lm(DV~1+Condition,data=my.data) # Create full model
```

Note that the ~ symbol can be taken to mean, 'as a function of'. In the case of the restricted model, we are modelling the dependent variable DV as a function of the grand mean (intercept) only, which is represented by a 1. In the full model, we are modelling DV as function of the grand mean (intercept) and Condition, where the effect of being in Condition j is represented by α_j.

We can formally compare the models with an F-test using the **anova** command:

```
my.aov<-anova(model.restricted,model.full)
print(my.aov)
```

```
## Analysis of Variance Table
##
## Model 1: DV ~ 1
## Model 2: DV ~ 1 + Condition
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     14 561.81
## 2     12 322.96  2    238.86 4.4376 0.03608 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table contains many components. First, note that $F(2,12) = 4.44$ and that the associated p-value is $<.05$. This means than we can reject the null hypothesis that $\alpha_1 = \alpha_2 = \alpha_3 = 0$, and that the addition of these α_j parameters in the full model significantly reduced the prediction error relative to the restricted model.

A quicker and easier way to do a one-way between-subjects ANOVA in R is to use the **aov** followed by the **summary** command as illustrated below:

```
my.aov2<-aov(DV~Condition,data=my.data)
print(summary(my.aov2))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## Condition    2  238.9  119.43   4.438 0.0361 *
## Residuals   12  323.0   26.91                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the values obtained are identical to those obtained using the **lm** and **anova** commands above. We can also take a look at our model coefficients α_j by using the **coef** command and inputting the *my.aov2* object:

```
print(coef(my.aov2))
```

```
## (Intercept)  Condition1  Condition2 
##   51.708116   -5.373607    1.193841
```

The intercept term represents the grand mean (μ) of all scores in our dataset, wheareas the labels Condition1 and Condition2 represent α_1 and α_2, respectively. α_3 is not shown, but can easily be calculated given that:

$$\sum_{j=1}^{k} \alpha_j = 0$$

Therefore the remaining α coefficient can be calculated as follows:

$$\alpha_3 = -\sum_{j=1}^{k-1} \alpha_j = -1(-5.37 + 1.19) = 4.18$$

Most students who have learned about analysis of variance (ANOVA) in the past might not recognize the approach to ANOVA described here. However, mathematically the two approaches are identical. Traditionally, ANOVA is taught by calculating sums-of-squares (SS) and mean-square (MS) terms that capture the relevant sources of variability in the data. This approach is directly related to the approach taken here. For instance it can be shown that:

$$SS_{Between} = E_R - E_F$$

$$SS_{Within} = E_F$$

This makes sense given that our equation for the F-statistic is:

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} = \frac{SS_{Between}/(df_R - df_F)}{SS_{Within}/df_F} = \frac{MS_{Between}}{MS_{Within}}$$

Post-Hoc Tests

A one-way ANOVA is known as an *omnibus* test because it tests whether there exists at least one difference between multiple condition means ($\bar{Y}_j$). Therefore, a statistically significant result with p <.05 only tells us that at least one of the condition means differs from one of the other condition means, or equivalently, that at least one $\alpha_j \neq 0$. Because there are three conditions, we need to perform additional tests to pinpoint which condition means differ from each other. These tests are known as *post-hoc* tests because they are conducted after the omnibus test.

There are many different types of post-hoc tests, but the one mostly commonly used when all pairwise comparisons of condition means are of interest is Tukey's Honestly Significant Difference (HSD). Importantly, this procedure controls for the inflation of the family-wise Type I error associated with multiple tests. We will discuss this concept in more depth in future lessons. Implementing Tukey's HSD in *R* is very straightforward using the **TukeyHSD** command and inputting the *my.aov2* object created with the **aov** command:

```
pairwise.posthoc<-TukeyHSD(my.aov2)
print(pairwise.posthoc)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = DV ~ Condition, data = my.data)
##
## $Condition
##         diff        lwr      upr     p adj
## B-A 6.567448 -2.1859157 15.32081 0.1541764
## C-A 9.553374  0.8000096 18.30674 0.0324591
## C-B 2.985925 -5.7674389 11.73929 0.6445035
```

Measures of Effect Size

Eta-Squared

Although a significant ANOVA suggests that there exists at least one difference among the condition means, this analysis does not address the question: *How large is the effect?* In the case of a one-way ANOVA, one appropriate measure of effect size is known as eta-squared, η^2 (your MDK textbook refers to η^2 as R^2), which represents the proportion of the total variability in the observed scores that can be attributed to our between-subjects manipulation. It is calculated in the following manner:

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

The value of $SS_{Between}$ can be found in the ANOVA table in the output from the **summary(aov)** command, and represents the variability in observed scores that is due to our experimental manipulation. The value of SS_{Total} is simply the sum of $SS_{Between}$ and SS_{Within}:

```
SS.between<-238.9 # Taken from ANOVA table above
SS.within<-323.0 # Taken from ANOVA table above
SS.total<-SS.between+SS.within

eta.sq<-SS.between/SS.total
print(round(eta.sq, digits=2))
```

```
## [1] 0.43
```

In other words, ~43% of the total variability in our dependent measure is attributable to the Condition manipulation.

Cohen's f

Another way to think of effect size is as a signal-to-noise ratio. In other words, we can express the standard deviation among group means (σ_m) as a proportion of the pooled within-group standard deviation (σ_e). σ_m

can be calculated directly from the values of α_j:

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^{k} \alpha_j^2}{k}}$$

Likewise, σ_e is simply the within-group standard deviation pooled across groups ($\sqrt{MS_W}$). These values are then used to calculate Cohen's *f*, which is our signal-to-noise ratio, and is useful for power calculations, which we will turn to later.

$$f = \frac{\sigma_m}{\sigma_e}$$

Alternatively, it is also possible to derive Cohen's f directly from η^2:

$$f = \sqrt{\frac{\eta^2}{(1-\eta^2)}}$$

According to Cohen, benchmarks for the f effect size measure define a "small" effect as $f = .10$, a "medium" effect as $f = .25$, and a "large" effect as $f = .40$.

Power

Just as we did in the case of a two independent groups design (see Lab Tutorial 2), we can also use power calculations to determine the required sample size (n) needed to detect an effect of size (f) with a pre-specified probability when our design includes three or more independent groups.

Now, let's use the **pwr.anova.test** function in the *pwr* package to calculate how many subjects we would need to detect an effect size of $f = .15$ (small) with a power = 0.80 in a study with three conditions manipulated between-subjects. This function requires us to input the number of groups in our study (k), the desired level of power, and the significance level we have chosen (0.05 by convention). The output will give us n, which is the number of subjects per group.

```r
library(pwr) # make sure the pwr package is installed first - install.packages("pwr")

pwr.anova.test(k=3,n=NULL,f=.15,sig.level=0.05,power=0.80)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 143.7394
##               f = 0.15
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

To detect an effect size of Cohen's $f = .15$ with a power of 0.80, we would need 144 subjects in each of our three groups!! **Take home message**: When you are trying to detect a small effect size, you probably need more subjects than you think you do if you want to have a good chance of detecting it!

Underlying Assumptions of ANOVA

ANOVAs make a number of assumptions about the structure of the data. If these assumptions are not met, it can threaten the validity of the ANOVA results. In this section I will describe some of the assumptions, and show you how to evaluate them in R.

1) Homogeneity of Variance

ANOVAs make the assumption that the variance of scores within each group are the same for all groups. This stems in part from the fact that the within-group variability estimates in an ANOVA are pooled across groups. Violations of this assumption can lead to erroneous results, particularly if the group *n's* are small or if there are unequal *n's* in each group.

To test the null hypothesis that the variances for each group are equal, we can use the **bartlett.test** function.

```
bartlett.test(DV~Condition,data=my.data)
```

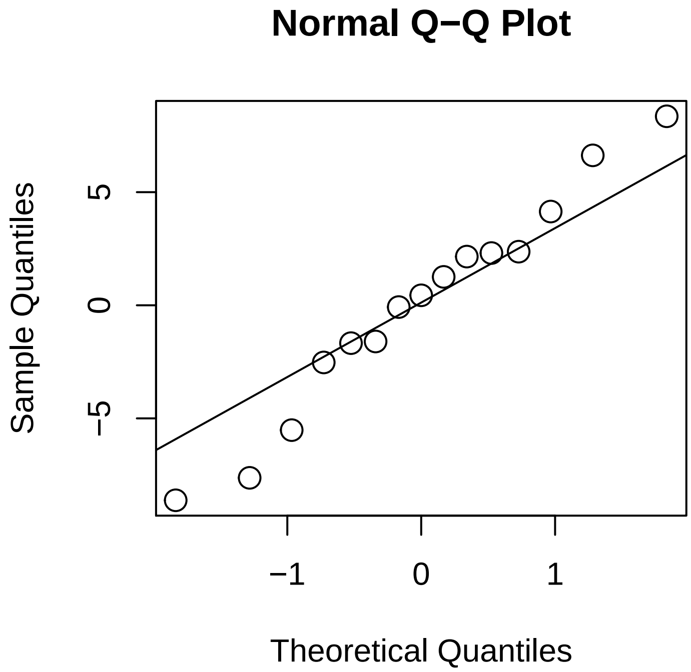
```
##
##  Bartlett test of homogeneity of variances
##
## data:  DV by Condition
## Bartlett's K-squared = 0.21834, df = 2, p-value = 0.8966
```

Our p-value = .89, so we do not reject the null hypothesis of equal variances. It appears that in our data, the homogeneity of variance assumption is satisfied.

2) The Residuals Are Distributed Normally

ANOVAs also assume that the residuals from the fitted ANOVA model are normally distributed (bell-shaped). To assess whether the residuals are indeed normally distributed, we can use the **qqnorm** function to generate a plot of the residuals. In this plot, the residuals should all fall along the diagonal line. Noticeable departures from the diagonal line are indicative of non-normality. To formally test the null hypothesis that the residuals are distributed normally, you can use the **shapiro.test** function.

```
resid<-residuals(my.aov2)
qqnorm(resid,cex=1.5) # make q-q (quantile-quantile) plot of residuals
qqline(resid) # plot diagonal line
```



```r
shapiro.test(resid) # test whether residuals differ significantly from normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97151, p-value = 0.8797
```

The p-value is > 0.50, so we're probably OK assuming that the residuals are distributed normally.

3) Scores are independent of one another

This assumption is typically satisfied by random assignment of subjects to condition.

Note: There are alternative tests that you can use if these assumptions are violated, but they are usually less powerful. We won't discuss them in this class, but you can find additional information on these tests in your textbook.