Tutorial 11. Multivariate Analysis of Variance (MANOVA)

Tutor: Daniel Yanan ZHOU

As you might expect, we use a multivariate analysis of variance (MANOVA) when we have one or more categorical independent variables with two or more treatment levels AND more than one continuous response variable (that's what makes it "multivariate"). MANOVA is commonly used species communities, where the frequency of each species represents a response variable.

MANOVA is based on the same principles as a discriminant analysis, which is a rotational technique designed to maximise variance between groups, rather than across an entire data set (which would be a principal component analysis). Essentially, MANOVA examined the variance explained between your groups (treatment levels) by comparing components (called "eigenvectors") which separate the data.

Generally, MANOVA is used for two things: 1) to implement a single inferential test on multiple dependent variables which may be correlated, and more frequently 2) determine the change in arrangement of dependent variables in response to the independent variable(s).

1. Prepare the Data

For a change let's take a break from environmental data and do something with more of a human aspect, let's consider data on triathlon performance. In case you're not familiar with triathlon, it is a multi-sport race where competitors complete a swim course, bike course, and run course, in that order. We want to know if gender or age category has an effect on the times for the individual sports. So we have a multi- factor MANOVA. This seems complicated, but it actually runs exactly the same as a regular multi-factor ANOVA.

- Download the "triathlon.csv" data from the class website. This is a multivariate data set containing triathlon race times for each sport (SWIM, BIKE, and RUN), which are our continuous responses. As well, we have as categorical information about the racers' genders (M or F) and age categories (CAT1, CAT2, or CAT3), which are our independent predictors.

```
dat <- read.csv("triathlon.csv")
head(dat)
```

- The data format for MANOVA is slightly different than we saw in ANOVA. R needs each independent variable in its own vector of factors. It also needs all the continuous response variables together in a separate matrix. Not to worry, we can make those files easily:

```
gender <- as.factor(dat[,1])
cat <- as.factor(dat[,2])
times <- as.matrix(dat[,3:5])
```

- Excellent. Now we're ready to proceed with the analysis. Well, almost ready. First we have to test our assumptions.

2. Testing the Assumptions of MANOVA

- Just as with a standard ANOVA, the MANOVA analysis assumes both normality and homoscedasticity (equality of variance) of your experimental errors (residuals). Again, just as before, we can test both of these assumptions statistically or graphically. Remember, however, that we must test the assumptions for each level of each factor. For MANOVA, we also must test each response variable within each level of each treatment. I know. This is getting ridiculous, but that's the price of inferential stats!

- First, we will create R objects for the residuals from each treatment level:

```
SWIMresCAT=lm(dat$SWIM~CATEGORY)$residuals
BIKEresCAT=lm(dat$BIKE~CATEGORY)$residuals
RUNresCAT=lm(dat$RUN~CATEGORY)$residuals

SWIMresGEN=lm(dat$SWIM~GENDER)$residuals
BIKEresGEN=lm(dat$BIKE~GENDER)$residuals
RUNresGEN=lm(dat$RUN~GENDER)$residuals
```

- To test normality statistically, we can use a Shapiro Test on each response for each level of each treatment:

```
shapiro.test(SWIMresCAT)

# Repeat for each sport for CATEGORY and GENDER
```

- We can also check for normality graphically by looking at the boxplots of residuals. In this code, we will make residuals for each treatment and plot them, separated by the treatment levels:

```
boxplot(lm(dat$SWIM~cat)$residuals~cat) # For CATEGORY
boxplot(lm(dat$SWIM~cat)$residuals~cat) # For GENDER

# Repeat for each sport
```

- You probably found that almost all the groups are normally distributed. We're comfortable that our data meets the assumption of normality. Now, we'll look at homoscedasticity. This is a bit easier. We can visually examine all the levels of a single treatment in one shot by plotting the residuals and looking for equality:

- If we want to look at the variances statistically, we can use Bartlett's test:

```
bartlett.test(SWIMresCAT~cat)
bartlett.test(BIKEresCAT~cat)
bartlett.test(RUNresCAT~cat)

# Repeat for gender
```

- We can even look at the residual plots for all combinations of CATEGORY and GENDER for each sport:

```
plot(lm(dat$SWIM~dat$CATEGORY*dat$GENDER))
plot(lm(dat$BIKE~dat$CATEGORY*dat$GENDER))
plot(lm(dat$RUN~dat$CATEGORY*dat$GENDER))
```

- Looks like we meet the assumptions fairly well. Even though a few samples tested as non-normal or with slight inequality of variance, we know that ANOVA and MANOVA are both fairly robust to deviations from normality and homoscedasticity, especially when sample sizes are equal (which they are, in this case). Now, we can run the actual test (which takes much less time and less code!).

3. Run the MANOVA

- We'll run the MANOVA by generating a new R object, which we can subsequently query for various statistics and outputs. We're interested in the effect of both gender and age category on all three times, so we'll include them both in the model (just like ANOVA). But we're also interested in the interaction between them, so we'll separate them with an asterisk (*):

```
output <- manova(times~gender*cat)
summary.aov(output)
```

- In this summary, we can see how each response variable relates to each treatment. From this output, we can see that the swim and bike times for at least one level of age category is significantly different. Gender doesn't appear to have much effect on anything (maybe bike time). However, we have a significant interaction between gender and category for the swim and bike times. Unfortunately, as far as I know, there's no way to produce an interaction plot with a MANOVA. Nor is there a good way to run multiple comparisons, aside from looking at each effect individually (in separate MANOVAs). This isn't a bad thing, necessarily. If we're interested in a particular effect or interaction, we should be running a univariate ANOVA instead.

- Typically when we run ANOVA, we want to know if the arrangement of responses (if ALL responses together) are significant as a whole. To obtain this, we need to ask for specific statistics. MANOVA can be interpreted with one of three multivariate F-tests: 1) Pillai's trace (R default), 2) Wilk's lambda (Λ), and 3) Hotelling-Lawley's trace.

- In reality, any of these tests are fine. The most common MANOVA test is Wilk's lambda, which can be useful as 1-Λ is often interpreted as the proportion of the variance explained by the model. Before you choose a test, however, you should read up on them to determine which is the best metric for your purposes:

```
summary(output, test="Wilks")
summary(output, test="Pillai")
summary(output, test="Hotelling")
```

- You can also use the `summary()` command to obtain other statistics, such as the sums of squares. There is more info on this in the `?summary.manova` help file:

```
summary(output)$SS
```