

# Multivariate Analysis of Variance

---

Max Turgeon

STAT 4690—Applied Multivariate Analysis

# Quick Overview

*What do we mean by Analysis of Variance?*

- ANOVA is a collection of statistical models that aim to analyze and understand the differences in means between different subgroups of the data.
  - As such, it can be seen as a generalisation of the  $t$ -test (or of Hotelling's  $T^2$ ).
  - Note that there could be multiple, overlapping ways of defining the subgroups (e.g multiway ANOVA)
- It also provides a framework for hypothesis testing.
  - Which can be recovered from a suitable regression model.
- **Most importantly**, ANOVA provides a framework for understanding and comparing the various sources of variation in the data.

# Review of univariate ANOVA i

- Assume the data comes from  $g$  populations:

$$\begin{array}{ccc} X_{11}, & \dots, & X_{1n_1} \\ \vdots & \ddots & \vdots \\ X_{g1}, & \dots, & X_{gn_g} \end{array}$$

- Assume that  $X_{\ell 1}, \dots, X_{\ell n_\ell}$  is a random sample from  $N(\mu_\ell, \sigma^2)$ , for  $\ell = 1, \dots, g$ .
  - Homoscedasticity**
- We are interested in testing the hypothesis that  $\mu_1 = \dots = \mu_g$ .

# Review of univariate ANOVA ii

- *Reparametrisation*: We will write the mean  $\mu_\ell = \mu + \tau_\ell$  as a sum of an overall component  $\mu$  (i.e. shared by all populations) and a population-specific component  $\tau_\ell$ .
  - Our hypothesis can now be rewritten as  $\tau_\ell = 0$ , for all  $\ell$ .
  - We can write our observations as

$$X_{\ell i} = \mu + \tau_\ell + \varepsilon_{\ell i},$$

where  $\varepsilon_{\ell i} \sim N(0, \sigma^2)$ .

- **Identifiability**: We need to assume  $\sum_{\ell=1}^g \tau_\ell = 0$ , otherwise there are infinitely many models that lead to the same data-generating mechanism.

## Review of univariate ANOVA iii

- *Sample statistics:* Set  $n = \sum_{\ell=1}^g n_{\ell}$ .
  - Overall sample mean:  $\bar{X} = \frac{1}{n} \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} X_{\ell i}$ .
  - Population-specific sample mean:  $\bar{X}_{\ell} = \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} X_{\ell i}$ .
- We get the following decomposition:

$$(X_{\ell i} - \bar{X}) = (\bar{X}_{\ell} - \bar{X}) + (X_{\ell i} - \bar{X}_{\ell}).$$

- Squaring the left-hand side and summing over both  $\ell$  and  $i$ , we get

$$\sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (X_{\ell i} - \bar{X})^2 = \sum_{\ell=1}^g n_{\ell} (\bar{X}_{\ell} - \bar{X})^2 + \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (X_{\ell i} - \bar{X}_{\ell})^2.$$

# Review of univariate ANOVA iv

- This is typically summarised as  $SS_T = SS_M + SS_R$ :

- The **total sum of squares**:

$$SS_T = \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (X_{\ell i} - \bar{X})^2$$

- The **model** (or treatment) **sum of squares**:

$$SS_M = \sum_{\ell=1}^g n_{\ell} (\bar{X}_{\ell} - \bar{X})^2$$

- The **residual sum of squares**:

$$SS_R = \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (X_{\ell i} - \bar{X}_{\ell})^2$$

# Review of univariate ANOVA v

- Yet another representation is the *ANOVA table*:

Source of Variation	Sum of Squares	Degrees of freedom
Model	$SS_M$	$g - 1$
Residual	$SS_R$	$n - g$
Total	$SS_T$	$n - 1$

- The usual test statistic used for testing  $\tau_\ell = 0$  for all  $\ell$  is

$$F = \frac{SS_M/(g - 1)}{SS_R/(n - g)} \sim F(g - 1, n - g).$$

- We could also instead reject the null hypothesis for *small* values of

$$\frac{SS_R}{SS_R + SS_M} = \frac{SS_R}{SS_T}.$$

**This is the test statistic that we will generalize to the multivariate setting.**



# Multivariate ANOVA i

- The setting is similar: Assume the data comes from  $g$  populations:

$$\begin{array}{ccc} \mathbf{Y}_{11}, & \dots, & \mathbf{Y}_{1n_1} \\ \vdots & \ddots & \vdots \\ \mathbf{Y}_{g1}, & \dots, & \mathbf{Y}_{gn_g} \end{array}$$

- Assume that  $\mathbf{Y}_{\ell 1}, \dots, \mathbf{Y}_{\ell n_\ell}$  is a random sample from  $N_p(\mu_\ell, \Sigma)$ , for  $\ell = 1, \dots, g$ .
  - **Homoscedasticity** is key here again.
- We are again interested in testing the hypothesis that  $\mu_1 = \dots = \mu_g$ .

# Multivariate ANOVA ii

- *Reparametrisation*: We will write the mean as  $\mu_\ell = \mu + \tau_\ell$ 
  - $\mathbf{Y}_{\ell i} = \mu + \tau_\ell + \mathbf{E}_{\ell i}$ , where  $\mathbf{E}_{\ell i} \sim N_p(0, \Sigma)$ .
- **Identifiability**: We need to assume  $\sum_{\ell=1}^g \tau_\ell = 0$ .
- Instead of a decomposition of the sum of squares, we get a decomposition of the outer product:

$$(\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}})(\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}})^T.$$

# Multivariate ANOVA iii

- The decomposition is given as

$$\sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}})(\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}})^T = \sum_{\ell=1}^g n_{\ell} (\bar{\mathbf{Y}}_{\ell} - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_{\ell} - \bar{\mathbf{Y}})^T + \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}}_{\ell})(\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}}_{\ell})^T.$$

- **Between sum of squares and cross products matrix:**

$$\mathbf{B} = \sum_{\ell=1}^g n_{\ell} (\bar{\mathbf{Y}}_{\ell} - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_{\ell} - \bar{\mathbf{Y}})^T.$$

- **Within sum of squares and cross products matrix:**

$$\mathbf{W} = \sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}}_{\ell})(\mathbf{Y}_{\ell i} - \bar{\mathbf{Y}}_{\ell})^T.$$

## Multivariate ANOVA iv

- Note that  $W = \sum_{\ell=1}^g (n_{\ell} - 1)S_{\ell}$ .
- Similarly as above, we have a *MANOVA table*:

Source of Variation	Sum of Squares	Degrees of freedom
Model	$B$	$g - 1$
Residual	$W$	$n - g$
Total	$B + W$	$n - 1$

- To test the null hypothesis  $H_0 : \tau_{\ell} = 0$  for all  $\ell = 1, \dots, g$ , we will use *Wilk's lambda* as our test statistic:

$$\Lambda = \frac{|W|}{|B + W|}.$$

# Multivariate ANOVA v

- There is actually no closed-form for the null distribution of  $\Lambda$ , so we will use Bartlett's approximation:

$$-\left(n - 1 - \frac{1}{2}(p + g)\right) \log \Lambda \approx \chi^2((g - 1)p).$$

- In particular, if we let  $c = \chi^2_{\alpha}((n - 1)p)$  be the critical value, we reject the null hypothesis if

$$\Lambda \leq \exp\left(\frac{-c}{n - 1 - 0.5(p + g)}\right).$$

## Example i

```
## Example on producing plastic film
## from Krzanowski (1998, p. 381)
tear <- c(6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2,
          6.9, 6.1, 6.3, 6.7, 6.6, 7.2, 7.1,
          6.8, 7.1, 7.0, 7.2, 7.5, 7.6)
gloss <- c(9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0,
          9.9, 9.5, 9.4, 9.1, 9.3, 8.3, 8.4,
          8.5, 9.2, 8.8, 9.7, 10.1, 9.2)
opacity <- c(4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0,
            3.9, 1.9, 5.7, 2.8, 4.1, 3.8, 1.6,
            3.4, 8.4, 5.2, 6.9, 2.7, 1.9)
```

## Example ii

```
Y <- cbind(tear, gloss, opacity)
Y_low <- Y[1:10,]
Y_high <- Y[11:20,]
n <- nrow(Y); p <- ncol(Y); g <- 2

W <- (nrow(Y_low) - 1)*cov(Y_low) +
      (nrow(Y_high) - 1)*cov(Y_high)
B <- (n-1)*cov(Y) - W
(Lambda <- det(W)/det(W+B))

## [1] 0.4136192
```

## Example iii

```
transf_lambda <- -(n - 1 - 0.5*(p + g))*log(Lambda)
transf_lambda > qchisq(0.95, p*(g-1))
```

```
## [1] TRUE
```

*# Or if you want a p-value*

```
pchisq(transf_lambda, p*(g-1), lower.tail = FALSE)
```

```
## [1] 0.002227356
```



## Example iv

```
# R has a function for MANOVA  
# But first, create factor variable  
rate <- gl(g, 10, labels = c("Low", "High"))  
  
fit <- manova(Y ~ rate)  
summary_tbl <- broom::tidy(fit, test = "Wilks")  
# Or you can use the summary function  
  
knitr::kable(summary_tbl, digits = 3)
```

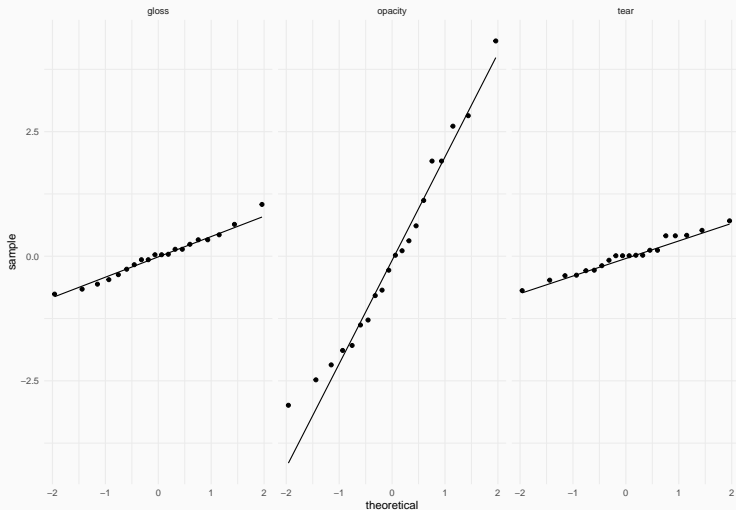
## Example v

term	df	wilks	statistic	num.df	den.df	p.value
rate	1	0.414	7.561	3	16	0.002
Residuals	18	-	-	-	-	-

## Example vi

```
# Check residuals for evidence of normality  
library(tidyverse)  
fit %>%  
  residuals %>%  
  as.data.frame() %>%  
  gather(variable, residual) %>%  
  ggplot(aes(sample = residual)) +  
  stat_qq() + stat_qq_line() +  
  facet_grid(. ~ variable) +  
  theme_minimal()
```

# Example vii



- The output from R shows a different approximation to the Wilk's lambda distribution, due to Rao.
- There are actually 4 tests available in R (we will discuss them in the next lecture):
  - Wilk's lambda;
  - Pillai-Bartlett;
  - Hotelling-Lawley;
  - Roy's Largest Root.

- Since we only had two groups in the above example, we were only comparing two means.
  - Wilk's lambda was therefore equivalent to Hotelling's  $T^2$ .
  - But of course MANOVA is much more general.
- We can assess the normality assumption by looking at the residuals  $\mathbf{E}_{\ell i} = \mathbf{Y}_{\ell i} - \bar{\mathbf{Y}}_{\ell}$ .

# Testing for Equality of Covariance Matrices i

- Last lecture, when comparing two multivariate means, and again today, we talked about **homoscedasticity** as an important assumption.
- This is a *testable* assumption, i.e. we can devise a corresponding hypothesis test.
- Our null hypothesis:  $H_0 : \Sigma_1 = \dots = \Sigma_g$ , where  $\Sigma_\ell$  is the covariance matrix for population  $\ell$ .
- In this course, we will discuss *Box' M-test*
  - This test is based on a comparison of generalized variances.

## Testing for Equality of Covariance Matrices ii

- Under the normality assumption, the likelihood ratio statistic for the null hypothesis above is

$$\Lambda = \prod_{\ell=1}^g \left( \frac{|S_{\ell}|}{|S_{pool}|} \right)^{(n_{\ell}-1)/2}.$$

- Here,  $S_{\ell}$  is the sample covariance for population  $\ell$ , and  $S_{pool}$  is the pooled estimator:

$$S_{pool} = \frac{1}{n-1} \left( \sum_{\ell=1}^g (n_{\ell}-1) S_{\ell} \right) = \frac{1}{n-1} W.$$



## Testing for Equality of Covariance Matrices iii

- Box's M-statistic is defined as

$$M = -2 \log \Lambda.$$

- The general theory of Likelihood Ratio Tests tells us that  $M \approx \chi^2(\nu)$  for an appropriate value  $\nu > 0$ .

# Testing for Equality of Covariance Matrices iv

## Box's Test for Equality of Covariance Matrices

Set

$$u = \left( \sum_{\ell=1}^g \frac{1}{n_{\ell} - 1} - \frac{1}{n - g} \right) \left( \frac{2p^2 + 3p - 1}{6(p + 1)(g - 1)} \right).$$

Then  $C = (1 - u)M$  has approximate  $\chi^2(\nu)$  distribution, where

$$\nu = \frac{1}{2}p(p + 1)(g - 1).$$

## Comments about Box's M-test

- Good approximation if  $n_\ell > 20$  for all  $\ell$  and both  $g, p \leq 5$ .
  - Not very realistic for modern datasets...
- There is another approximation using the  $F$  distribution when the conditions above are not met.
  - See Rencher (1998), Section 4.3.
- However, Box's M-test is especially sensitive to departures from normality.
- In general, one can also use graphical tests.
- **Key result:** With large and approximately equal sample sizes, MANOVA is relatively robust to heteroscedasticity.

## Example (cont'd) i

```
S_low <- cov(Y_low)
S_high <- cov(Y_high)
S_pool <- W/(n - 1)

c("pool" = log(det(S_pool)),
  "low" = log(det(S_low)),
  "high" = log(det(S_high)))

##      pool      low      high
## -2.370911 -2.949096 -2.013061
```

## Example (cont'd) ii

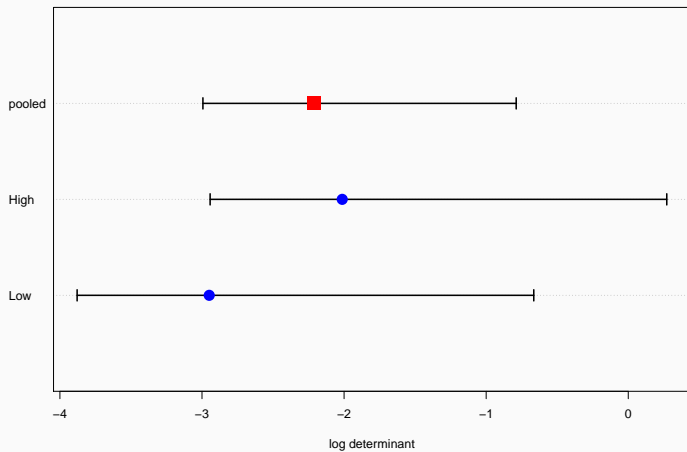
```
library(heplots)
(boxm_res <- boxM(Y, rate))

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data: Y
## Chi-Sq (approx.) = 4.0175, df = 6, p-value = 0.6743
```

## Example (cont'd) iii

```
# You can plot the log generalized variances  
# The plot function adds 95% CI  
plot(boxm_res)
```

## Example (cont'd) iv

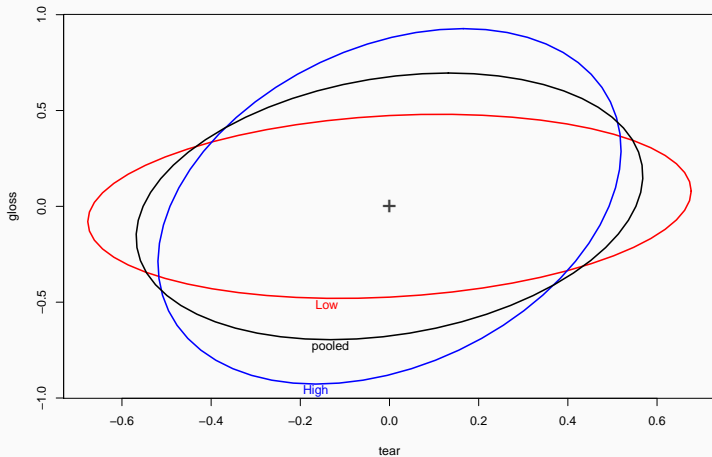


## Example (cont'd) v

```
# Finally you can also plot the ellipses  
# as a way to compare the covariances  
covEllipses(Y, rate, center = TRUE,  
             label.pos = 'bottom')
```



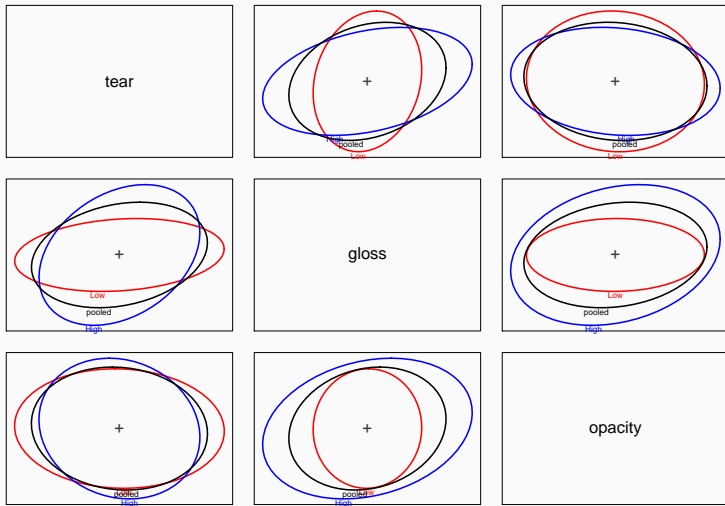
## Example (cont'd) vi



## Example (cont'd) vii

```
# Or all pairwise comparisons together  
covEllipses(Y, rate, center = TRUE,  
            label.pos = 'bottom',  
            variables = 1:3)
```

## Example (cont'd) viii



# Strategy for Multivariate Comparison of Treatments

1. Try to identify outliers.
  - This should be done graphically at first.
  - Once the model is fitted, you can also look at influence measures.
2. Perform a multivariate test of hypothesis.
3. If there is evidence of a multivariate difference, calculate Bonferroni confidence intervals and investigate component-wise differences.
  - The projection of the confidence region onto each variable generally leads to confidence intervals that are too large.