

Multivariate Analysis of Variance

Multivariate analysis of variance

- Standard ANOVA has just one response variable.
- What if you have more than one response?
- Try an ANOVA on each response separately.
- But might miss some kinds of interesting dependence between the responses that distinguish the groups.

Packages

```
library(car) # may need to install first  
library(tidyverse)  
library(MVTests) # also may need to install
```

Small example

- Measure yield and seed weight of plants grown under 2 conditions: low and high amounts of fertilizer.
- Data (fertilizer, yield, seed weight):

```
url <- "http://ritsokiguess.site/datafiles/manova1.txt"  
hilo <- read_delim(url, " ")
```

- 2 responses, yield and seed weight.

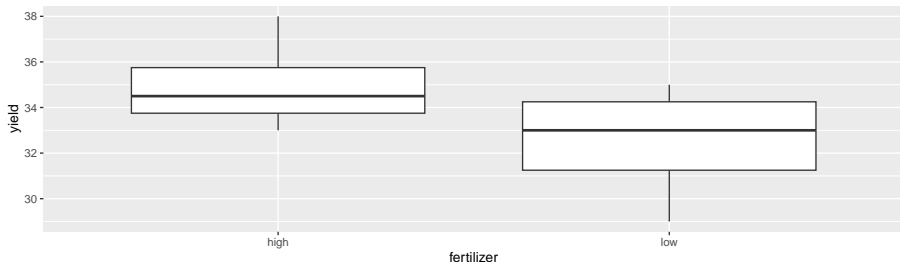
The data

```
hilo
```

```
# A tibble: 8 x 3
  fertilizer yield weight
  <chr>      <dbl>  <dbl>
1 low         34     10
2 low         29     14
3 low         35     11
4 low         32     13
5 high        33     14
6 high        38     12
7 high        34     13
8 high        35     14
```

Boxplot for yield for each fertilizer group

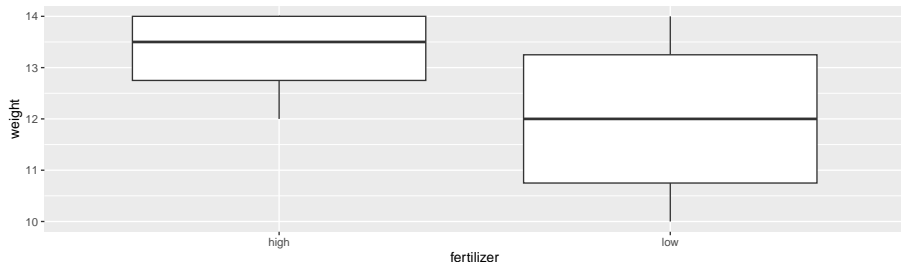
```
ggplot(hilo, aes(x = fertilizer, y = yield)) + geom_boxplot()
```



Yields overlap for fertilizer groups.

Boxplot for weight for each fertilizer group

```
ggplot(hilo, aes(x = fertilizer, y = weight)) + geom_boxplot()
```



Weights overlap for fertilizer groups.

ANOVAs for yield and weight

```
hilo.y <- aov(yield ~ fertilizer, data = hilo)
summary(hilo.y)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fertilizer	1	12.5	12.500	2.143	0.194
Residuals	6	35.0	5.833		

```
hilo.w <- aov(weight ~ fertilizer, data = hilo)
summary(hilo.w)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fertilizer	1	3.125	3.125	1.471	0.271
Residuals	6	12.750	2.125		

Neither response depends significantly on fertilizer. But...

Plotting both responses at once

- Have two response variables (not more), so can plot the response variables against *each other*, labelling points by which fertilizer group they're from.
- First, create data frame with points (31, 14) and (38, 10) (why? Later):

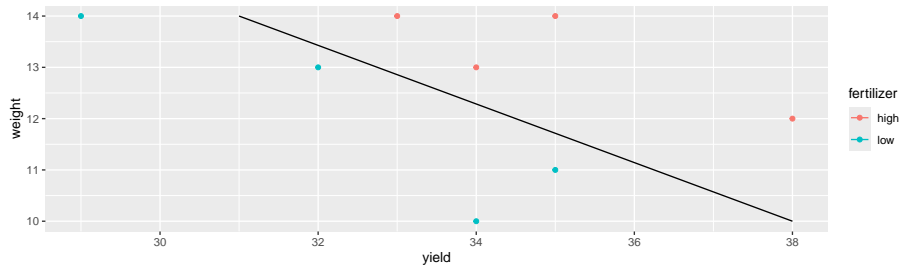
```
d <- tribble(  
  ~line_x, ~line_y,  
  31, 14,  
  38, 10  
)  
d
```

```
# A tibble: 2 x 2  
  line_x line_y  
  <dbl> <dbl>  
1     31     14  
2     38     10
```

- Then plot data as points, and add line through points in d:

```
ggplot(hilo, aes(x = yield, y = weight,
```

The plot



Comments

- Graph construction:
 - ▶ Joining points in `d` by line.
 - ▶ `geom_line` inherits colour from `aes` in `ggplot`.
 - ▶ Data frame `d` has no fertilizer (previous colour), so have to unset.
- Results:
 - ▶ High-fertilizer plants have both yield and weight high.
 - ▶ True even though no sig difference in yield or weight individually.
 - ▶ Drew line separating highs from lows on plot.

MANOVA finds multivariate differences

- Is difference found by diagonal line significant? MANOVA finds out.

```
response <- with(hilo, cbind(yield, weight))
response
```

	yield	weight
[1,]	34	10
[2,]	29	14
[3,]	35	11
[4,]	32	13
[5,]	33	14
[6,]	38	12
[7,]	34	13
[8,]	35	14

```
hilo.1 <- manova(response ~ fertilizer, data = hilo)
summary(hilo.1)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
fertilizer	1	0.80154	10.097	2	5	0.01755 *
Residuals	6					

Strategy

- Create new response variable by gluing together columns of responses, using `cbind`.
- Use `manova` with new response, looks like `lm` otherwise.
- With more than 2 responses, cannot draw graph. What then?
- If MANOVA test significant, cannot use Tukey. What then?
- Use *discriminant analysis* (of which more later).

Another way to do MANOVA

using Manova from package car:

```
hilo.2.lm <- lm(response ~ fertilizer, data = hilo)
hilo.2 <- Manova(hilo.2.lm)
summary(hilo.2)
```

Type II MANOVA Tests:

Sum of squares and products for error:

```
      yield weight
yield    35 -18.00
weight  -18  12.75
```

Term: fertilizer

Sum of squares and products for the hypothesis:

```
      yield weight
yield  12.50  6.250
weight  6.25  3.125
```

Multivariate Tests: fertilizer

	Df	test	stat	approx	F	num	Df	den	Df	Pr(>F)
Pillai	1	0.801542	10.09714			2	5	0.017546	*	
Wilks	1	0.198458	10.09714			2	5	0.017546	*	
Hotelling-Lawley	1	4.038855	10.09714			2	5	0.017546	*	
Roy	1	4.038855	10.09714			2	5	0.017546	*	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comments

- Same result as small-m manova.
- Manova will also do *repeated measures*, coming up later.

Assumptions

- normality of each response variable within each treatment group
 - ▶ this is actually *multivariate* normality, with correlations
- equal spreads: each response variable has same variances and correlations (with other response variables) within each treatment group. Here:
 - ▶ yield has same spread for low and high fertilizer
 - ▶ weight has same spread for low and high fertilizer
 - ▶ correlation between yield and weight is same for low and high fertilizer
- test equal spread using Box's M test
 - ▶ a certain amount of unequalness is OK, so only a concern if P-value from M -test is very small (eg. less than 0.001).

Assumptions for yield-weight data

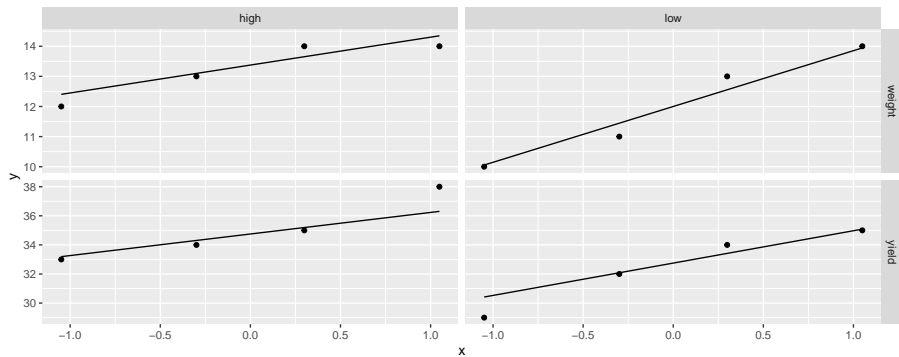
For normal quantile plots, need “extra-long” with all the data values in one column:

```
hilo %>%  
  pivot_longer(-fertilizer, names_to = "xname",  
               values_to = "xvalue") %>%  
  ggplot(aes(sample = xvalue)) + stat_qq() +  
    stat_qq_line() +  
    facet_grid(xname ~ fertilizer, scales = "free") -> g
```

There are only four observations per response variable - treatment group combination, so graphs are not very informative (over):

The plots

gg



Box M test

- Make sure package `MVTests` loaded first.
- inputs:
 - ▶ the response matrix (or, equivalently, the response-variable columns from your dataframe)
 - ▶ the column with the grouping variable in it (most easily gotten with `$`).

```
library(MVTests)
# hilo %>% select(yield, weight) -> numeric_values
summary(BoxM(response, hilo$fertilizer))
```

Box's M Test

Chi-Squared Value = 1.002964 , df = 3 and p-value: 0.801

- No problem at all with unequal spreads.

Another example: peanuts

- Three different varieties of peanuts (mysteriously, 5, 6 and 8) planted in two different locations.
- Three response variables: y , smk and w .

```
u <- "http://ritsokiguess.site/datafiles/peanuts.txt"  
peanuts.orig <- read_delim(u, " ")
```

The data

```
peanuts.orig
```

```
# A tibble: 12 x 6
```

	obs	location	variety	y	smk	w
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	1	5	195.	153.	51.4
2	2	1	5	194.	168.	53.7
3	3	2	5	190.	140.	55.5
4	4	2	5	180.	121.	44.4
5	5	1	6	203	157.	49.8
6	6	1	6	196.	166	45.8
7	7	2	6	203.	166.	60.4
8	8	2	6	198.	162.	54.1
9	9	1	8	194.	164.	57.8
10	10	1	8	187	165.	58.6
11	11	2	8	202.	167.	65
12	12	2	8	200	174.	67.2

Setup for analysis

```
peanuts.orig %>%  
  mutate(  
    location = factor(location),  
    variety = factor(variety)  
  ) -> peanuts  
peanuts
```

A tibble: 12 x 6

	obs	location	variety	y	smk	w
	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<dbl>
1	1	1	5	195.	153.	51.4
2	2	1	5	194.	168.	53.7
3	3	2	5	190.	140.	55.5
4	4	2	5	180.	121.	44.4
5	5	1	6	203	157.	49.8
6	6	1	6	196.	166	45.8
7	7	2	6	203.	166.	60.4

Analysis (using Manova)

```
peanuts.1 <- lm(response ~ location * variety, data = peanuts)
peanuts.2 <- Manova(peanuts.1)
summary(peanuts.2)
```

Type II MANOVA Tests:

Sum of squares and products for error:

	y	smk	w
y	104.205	49.365	76.480
smk	49.365	352.105	121.995
w	76.480	121.995	94.835

Term: location

Sum of squares and products for the hypothesis:

	y	smk	w
y	0.7008333	-10.6575	7.129167
smk	-10.6575000	162.0675	-108.412500
w	7.1291667	-108.4125	72.520833

Multivariate Tests: location

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.893484	11.18432	3	4	0.020502 *
Wilks	1	0.106516	11.18432	3	4	0.020502 *
Hotelling-Lawley	1	8.388243	11.18432	3	4	0.020502 *
Roy	1	8.388243	11.18432	3	4	0.020502 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comments

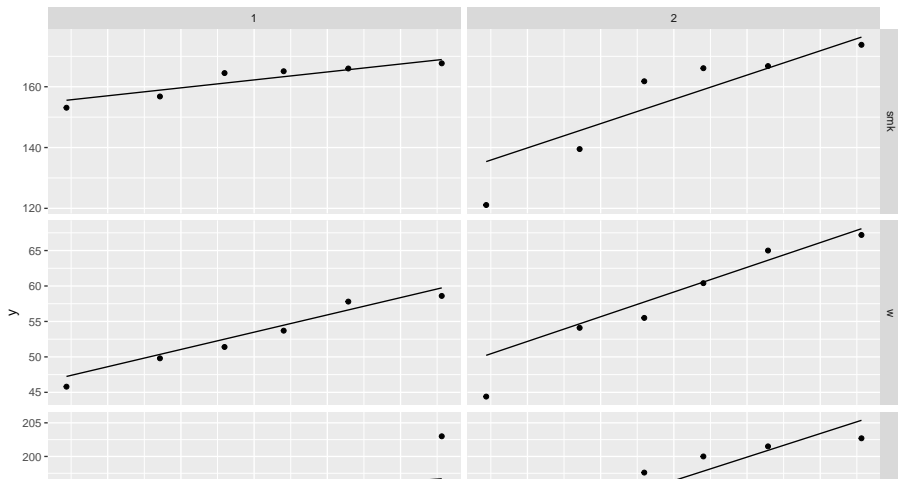
- Interaction not quite significant, but main effects are.
- Combined response variable (y, \mathbf{smk}, w) definitely depends on location and on variety
- Weak dependence of (y, \mathbf{smk}, w) on the location-variety *combination*.
- Understanding that dependence beyond our scope right now.

Comments

- this time there are only six observations per location and four per variety, so normality is still difficult to be confident about
- y at location 1 seems to be the worst for normality (long tails / outliers), and maybe y at location 2 is skewed left, but the others are not bad
- there is some evidence of unequal spread (slopes of lines), but is it bad enough to worry about? (Box M-test, over).

Assessing normality

```
peanuts %>% pivot_longer(y:w, names_to = "yname", values_to =  
  ggplot(aes(sample = y)) + stat_qq() + stat_qq_line() +  
  facet_grid(yname ~ location, scales = "free_y")
```



Box's M tests

- One for location, one for variety:

```
summary(BoxM(response, peanuts$location))
```

Box's M Test

Chi-Squared Value = 12.47797 , df = 6 and p-value: 0.0521

```
summary(BoxM(response, peanuts$variety))
```

Box's M Test

Chi-Squared Value = 10.56304 , df = 12 and p-value: 0.567

- Neither of these P-values is low enough to worry about. (Remember, the P-value here has to be *really* small to indicate a problem.)
- Box's M test does not work well (and can fail to work at all) if the sample sizes are too small.

Addendum: Box's M for interaction

- Create a combo column that is the combination of location and variety:

```
peanuts %>% mutate(combo = str_c(location, "-", variety)) -> c  
d
```

```
# A tibble: 12 x 7
```

	obs	location	variety	y	smk	w	combo
	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<chr>
1	1	1	5	195.	153.	51.4	1-5
2	2	1	5	194.	168.	53.7	1-5
3	3	2	5	190.	140.	55.5	2-5
4	4	2	5	180.	121.	44.4	2-5
5	5	1	6	203	157.	49.8	1-6
6	6	1	6	196.	166	45.8	1-6
7	7	2	6	203.	166.	60.4	2-6
8	8	2	6	198.	162.	54.1	2-6
9	9	1	8	194.	164.	57.8	1-8