Lecture 11 – Program

1. Data structure and model

2. Multivariate ANOVA

3. Multivariate analysis of covariance

Data structure

The new feature is that there is more than one response. On each unit there are p responses and q covariates. The data matrix can therefore be expressed as

unit	responses			covariates
1	y_{11}	$\cdots$	y_{1p}	$x_{11} \cdots x_{1q}$
2	y_{21}	$\cdots$	y_{2p}	$x_{21} \cdots x_{2q}$
.	.	$\cdots$	.	$\cdots$
.	.	$\cdots$	.	$\cdots$
.	.	$\cdots$	.	$\cdots$
n	y_{1n}	$\cdots$	y_{1p}	$x_{n1} \cdots x_{nq}$

Example: growth of tumors

18 mice in experiment

3 responses measured on each mice

y_1 initial weight

y_2 final weight minus tumor weight

y_3 final weight of tumor

Two factors

- Sex, two levels: male or female
- Temperature, three levels: 4, 20 and 34 centigrades

unit	init. wht.	finmintu. wht.	tum. wht	temp.	sex
1	18.15	16.51	0.21	4	1
2	18.68	19.50	0.32	4	1
3	19.54	19.84	0.20	4	1
4	19.15	19.49	0.16	4	2
⋮	⋮				⋮
18	20.85	19.90	0.17	34	2

- Objective: Explain variation in responses $y_1, \ldots, y_p$ by variation in covariates $x_1, \cdots, x_q$

- Covariates as earlier:
  - quantitative
  - qualitative
  - mixture of quantitative and qualitative

- p regression equations:

$$y_k = \beta_{0k} + \beta_{1k} x_1 + \cdots + \beta_{qk} x_q + \varepsilon_k$$

$$k = 1, \ldots, p$$

New feature: Have to take into account possible correlation in the error terms $\varepsilon_1, \ldots, \varepsilon_p$.

Parameters:

Var. no.	Regr. coeff.			Errors: SDs		Errors: Correlations			
1	β_{01}	$\cdots$	β_{q1}	σ_1	$\bullet$	τ_{12}	τ_{13}	$\cdots$	τ_{1p}
2	β_{02}	$\cdots$	β_{q2}	σ_2	τ_{21}	$\bullet$	τ_{23}	$\cdots$	τ_{2p}
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$				$\vdots$
p	β_{0p}	$\cdots$	β_{qp}	σ_p	τ_{p1}	τ_{p2}	τ_{p3}	$\cdots$	$\bullet$

Example:

$p = 3$ and $q = 2$

Number of parameters:

$3 \times (1 + 2) + 3 + 3 \times (3 - 1)/2 = 15$

Estimation method

First estimate $\beta_{0k}, \beta_{1k}, \cdots, \beta_{qk}$ from p separate linear regressions.

Residuals:

$$\widehat{\varepsilon}_k = y_k - (\widehat{\beta}_{0k} + \widehat{\beta}_{1k} x_1 + \cdots + \widehat{\beta}_{qk} x_q)$$

Estimate for variances:

$$\widehat{\sigma}_k^2 = (\widehat{\varepsilon}_{1k}^2 + \cdots + \widehat{\varepsilon}_{nk}^2)/(n - q - 1)$$

Estimate for covariances:

$$\widehat{\gamma}_{kk'} = (\widehat{\varepsilon}_{1k}\widehat{\varepsilon}_{1k'} + \cdots + \widehat{\varepsilon}_{nk}\widehat{\varepsilon}_{nk'})/(n - q - 1)$$

Estimate for correlations:

$$\widehat{\tau}_{kk'} = \frac{\widehat{\gamma}_{kk'}}{\widehat{\sigma}_k \widehat{\sigma}_{k'}}$$

Estimation contd.

As in multiple regression

$$\mathrm{Var}(\widehat{\beta}_{jk}) = t_{jj}\,\sigma_k^2$$

For covariances of estimators in the same regression

$$\mathrm{Cov}(\widehat{\beta}_{jk}, \widehat{\beta}_{j'k}) = t_{jj'}\,\sigma_k^2.$$

and for covariances of estimators in different regressions

$$\mathrm{Cov}(\widehat{\beta}_{jk}, \widehat{\beta}_{j'k'}) = t_{jj'}\,\gamma_{kk'}.$$

In the case that $(\varepsilon_1, \ldots, \varepsilon_p)$ is multivariate normally distributed

$$\frac{\widehat{\beta}_{jk} - \beta_{jk}}{\widehat{\sigma}_k \sqrt{t_{jj}}}$$

is t_{n-k-1}-distributed. This can be used as earlier to construct tests and confidence intervals

Multivariate analysis of variance (MANOVA)

Will decompose the total variation of each response.

Consider for ease of presentation the situation with only one factor, with levels indexed by i, and let u denote the replication number

The total sum of squares for response k is

$$\sum_i \sum_u (y_{iuk} - \bar{y}_{..k})^2.$$

These can be decomposed as in the univariate case.

A new aspect for MANOVA is that one also can consider sums of cross products:

$$\sum_i \sum_u (y_{iuk} - \bar{y}_{..k})(y_{iuk'} - \bar{y}_{..k'})$$

Decomposing the sum of squares is important for interpretation. When the focus is on testing the cross products must also be taken into account.

MANOVA in R

Illustrate by growth of tumors in mice.

First we organize the data:

```
resp1<-c(18.15,18.68,19.54,21.27,19.57,20.15,20.74,20.02,
17.20,19.15,18.35,20.68,18.87,20.66,21.56,20.22,18.38,20.85)

resp2<-c(16.51,19.50,19.84,23.30,22.30,18.95,16.69,19.26,15.90,
19.49,19.81,19.44,22.00,21.08,20.34,19.00,17.92,19.90)

resp3<-c(0.24,0.32,0.20,0.33,0.45,0.35,0.31,0.41,0.28,0.16,0.17,
0.22,0.25,0.20,0.20,0.18,0.30,0.17)


resp<-cbind(resp1,resp2,resp3)

temp<-c(4,4,4,20,20,20,34,34,34,4,4,4,20,20,20,34,34,34)

sex<-c(rep(1,9),rep(2,9))

fsex<-factor(sex)

ftemp<-factor(temp)
```

The "aov" command does separate ANOVA analyses for each of the variables:

```
fit.aov<-aov(resp~ftemp+fsex+ftemp:fsex)
```

Brief summary with breakdowns of the total sums of squares for the three variables:

```
fit.aov
                    ftemp     fsex ftemp:fsex Residuals
resp1             4.81608  0.64222    0.27548  19.32640
resp2            32.58671  2.51627    3.20538  26.69880
resp3             0.01963  0.06009    0.00608   0.03920
Deg. of Freedom         2        1          2        12

Residual standard error: 1.269068 1.49161 0.05715476
```

Note that the residual sum of squares dominates for the first response, while this is not the case for the other two.

This should come as no surprise. (Why?)

More detailed results of the analysis:

```
summary(fit.aov)

 Response resp1 :
            Df  Sum Sq Mean Sq F value Pr(>F)
ftemp        2  4.8161  2.4080  1.4952 0.2632
fsex         1  0.6422  0.6422  0.3988 0.5396
ftemp:fsex   2  0.2755  0.1377  0.0855 0.9186
Residuals   12 19.3264  1.6105

 Response resp2 :
            Df Sum Sq Mean Sq F value   Pr(>F)
ftemp        2 32.587  16.293  7.3232 0.008342
fsex         1  2.516   2.516  1.1310 0.308503
ftemp:fsex   2  3.205   1.603  0.7203 0.506476
Residuals   12 26.699   2.225


 Response resp3 :
            Df   Sum Sq  Mean Sq F value   Pr(>F)
ftemp        2 0.019633 0.009817  3.0051 0.087493
fsex         1 0.060089 0.060089 18.3946 0.001052
ftemp:fsex   2 0.006078 0.003039  0.9303 0.421120
Residuals   12 0.039200 0.003267
```

Analysis using the "manova" command:

```
mod1<-manova(resp~ftemp+fsex+ftemp:fsex)
```

Brief summary (as for "aov"):

```
 mod1
                   ftemp     fsex ftemp:fsex Residuals
resp1            4.81608  0.64222    0.27548  19.32640
resp2           32.58671  2.51627    3.20538  26.69880
resp3            0.01963  0.06009    0.00608   0.03920
Deg. of Freedom        2        1          2        12

Residual standard error: 1.269068 1.49161 0.05715476
```

Tests for overall effects of the factors:

```
summary(mod1, test="Wilks")

           Df  Wilks approx F num Df den Df  Pr(>F)
ftemp       2 0.2617   3.1827      6     20 0.02332
fsex        1 0.3373   6.5503      3     10 0.01001
ftemp:fsex  2 0.7720   0.4605      6     20 0.82909
Residuals  12
```

(Wilks test is described on the next slides)

Wilks test

We may want to pool evidence in favour of factor effects from the different variables. One such test is Wilks test, which is equivalent to an ordinary F-test when there is only one response.

We first describe Wilks test statistic for the case of only one response, say the kth.

Then (e.g.) the hypothesis of no main effect of temperature is rejected by a F-test if

$$\frac{SS_{\text{temp},k}/2}{SS_{\text{res},k}/12} > c$$

or equivalently (after some algebra) if

$$\frac{SS_{\text{res},k}}{SS_{\text{res}} + SS_{\text{temp},k}} < \frac{1}{1 + 2c/12} = k$$

The statistic

$$\Lambda = \frac{SS_{\text{res},k}}{SS_{\text{res}} + SS_{\text{temp},k}}$$

is Wilks test statistic ("Wilks lambda") for the case of one response

Note that we reject for small values of Wilks lambda

For the case of more responses the sums of squares are replaced by determinants of matrices of sums of squares (also involving the cross products mentioned above). The details are beyond the scope of this course.

Multivariate analysis of covariance

The variables of primary interest in the example is

y_2 = final weight minus tumor weight

y_3 = final weight of tumor

The first response:

y_1 = initial weight

may also be considered as a covariate.

Example: growth of tumors in mice

```
resp23<-cbind(resp2,resp3)
mod2<-manova(resp23~resp1+ftemp+fsex+ftemp:fsex)
mod2
```

```
              resp1     ftemp      fsex ftemp:fsex Residuals
resp2    14.044531 22.629698  1.552768   2.623570 24.156594
resp3     0.000166  0.025480  0.055261   0.006772  0.037320
Df               1         2         1          2         11

Residual standard error: 1.481909 0.05824682
```

```
summary(mod2,test="Wilks")
```

```
           Df  Wilks approx F num Df den Df   Pr(>F)
resp1       1 0.5977   3.3657      2     10 0.076266
ftemp       2 0.3123   3.9472      4     20 0.016075
fsex        1 0.3464   9.4337      2     10 0.004988
ftemp:fsex  2 0.7830   0.6506      4     20 0.633068
Residuals  11
```

There some (but not significant) effect of introducing "resp1" as a covariate.