Displays for Statistics 5401/8401

Lecture 22

October 28, 2005

Christopher Bingham, Instructor

612-625-1024, kb@umn.edu
372 Ford Hall

Class Web Page

http://www.stat.umn.edu/~kb/classes/5401

Analysis of covariance computations are useful even when there are no covariates.

They provide a way to test H_0: $\mathbf{LB} = \mathbf{0}$ that is different from either of

- tests based on the eigenvalues of $\mathbf{H}$ relative to $\mathbf{E}$
- tests based on Bonferronized univariate F-statistics for each response.
- tests based on Bonferronized univariate t-statistics for each coefficient of each response

Reminder: H_0: $\mathbf{LB} = \mathbf{0}$ is often stated more understandably in terms of means or effects. For example

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$, that is, no differences among group means
- H_0: $(\alpha\beta)_{jk} = 0$, all j and k, that is, no AB interactions

Suppose $\mathbf{Y} = [\mathbf{Y}_1 \quad \mathbf{Y}_2]$ consists of two groups of variables (columns of $\mathbf{Y}_1$ and columns of $\mathbf{Y}_2$).

The analysis of covariance approach can answer the following question:

- Does $\mathbf{Y}_2$ add information about violation of H_0 beyond information in $\mathbf{Y}_1$?

Example: H_0: $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$ in one-way MANOVA of $\mathbf{Y}$. Do the variables in $\mathbf{Y}_2$ provide information about differences of means that is not provided by $\mathbf{Y}_1$.

For the Fisher iris data, $\mathbf{Y}_1$ might contain sepal lengths and widths and $\mathbf{Y}_2$ petal lengths and widths. The question would be, do petal sizes help distinguish varieties once you know sepal sizes.

More specifically, suppose $\mathbf{Y}_1$ and $\mathbf{Y}_2$ have p_1 and p_2 columns respectively. For the iris data example, $\mathbf{Y}_1$ $p_1 = 2$ and $p_2 = 2$.

Then you can partition the coefficient and variance matrices as

$$\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2] \text{ and } \boldsymbol{\Sigma} = \begin{matrix} \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ & \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{array}\right] \end{matrix} \begin{matrix} p_1 \text{ rows} \\ \\ p_2 \text{ rows} \end{matrix}$$

Then $\mathbf{Y} = \mathbf{ZB} + \boldsymbol{\varepsilon}$, $V[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma}$ becomes

$$\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2] = [\mathbf{ZB}_1 \quad \mathbf{ZB}_2] + [\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2],$$
$$V[\boldsymbol{\varepsilon}_1] = \boldsymbol{\Sigma}_{11},\ V[\boldsymbol{\varepsilon}_2] = \boldsymbol{\Sigma}_{22},\ \mathrm{Cov}[\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2] = \boldsymbol{\Sigma}_{12}$$

The matrix of regression coefficients of the residuals $\mathbf{Y}_2 - \mathbf{ZB}_2$ on the residuals $\mathbf{Y}_1 - \mathbf{ZB}_1$ is $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

Note that $\boldsymbol{\Gamma} = \mathbf{0}$ if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

When the errors $[\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2]$ are multivariate normal, the conditional distribution of $\mathbf{Y}_2$ given both $\mathbf{Z}$ *and* $\mathbf{Y}_1$ is

$$N_{p_2}(\mathbf{Z}\,\mathbf{B}_2^* + \mathbf{Y}_1\boldsymbol{\Gamma}, \tilde{\boldsymbol{\Sigma}}_{22}) \text{ with } \mathbf{B}_2^* \equiv \mathbf{B}_2 - \mathbf{B}_1\boldsymbol{\Gamma}$$

This is essentially the same model as the MANACOVA model, with $\mathbf{Y}_1$ as covariates. Here's how the components match up:

Notation correspondence

MANACOVA	$\mathbf{U}$	$\mathbf{Y}$	$\mathbf{D}$	$\mathbf{B}$	$\mathbf{B}^*$	$\boldsymbol{\Gamma}$
Above	$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_2^*$	$\boldsymbol{\Gamma}$

$\mathbf{B}_2^*$ measures the effect of $\mathbf{Z}$ on $\mathbf{Y}_2$ that is not mediated through $\mathbf{Y}_1$. If $\mathbf{LB}_2^* = \mathbf{0}$, $\mathbf{Y}_2$ provides no additional information about violation of H_0.

Fact With $\mathbf{B} = [\mathbf{B}_1 \;\; \mathbf{B}_2]$, $\mathbf{B}_2^* \equiv \mathbf{B}_2 - \mathbf{B}_1\boldsymbol{\Gamma}$

H_0: $\mathbf{LB} = \mathbf{0}$ is true if and only if *both*

$H_0^{(1)}$: $\mathbf{LB}_1 = \mathbf{0}$ and $H_0^{(2)}$: $\mathbf{LB}_2^* = \mathbf{0}$

are true.

- Test $H_0^{(1)}$: $\mathbf{LB}_1 = \mathbf{0}$ by MANOVA of $\mathbf{Y}_1$ with design matrix $\mathbf{Z}$, ignoring $\mathbf{Y}_2$. $\mathbf{H} = \mathbf{H}^{(1)}$, $\mathbf{E} = \mathbf{E}^{(1)}$
- Test $H_0^{(2)*}$: $\mathbf{LB}_2^* = \mathbf{0}$ by MANACOVA of $\mathbf{Y}_2$ with design matrix $\mathbf{Z}$ and $\mathbf{Y}_1$ as covariates. $\mathbf{H} = \mathbf{H}^{(2)*}$, $\mathbf{E} = \mathbf{E}^{(2)*}$

For both tests, you can use any available test - Bonferronized F or tests based on relative eigenvalues.

Note: This is different from testing

$$\mathbf{LB}_1 = \mathbf{0} \text{ and } \mathbf{LB}_2 = \mathbf{0}$$

by Bonferronizing multivariate tests based on MANOVA of $\mathbf{Y}_1$ and MANOVA of $\mathbf{Y}_2$.

Fact: Under multivariate normality, $\mathbf{H}^{(1)}$ and $\mathbf{E}^{(1)}$ are *independent* of $\mathbf{H}^{(2)*}$ and $\mathbf{E}^{(2)*}$.

This means you can combine P-values from each test more advantageously than by Bonferronizing.

If you use $\alpha' = 1 - (1 - \alpha)^{1/2} \tilde{=} \alpha/2 + \alpha^2/8 > \alpha/2$, the Bonferrronized α, the <u>overall</u> significance level of your test is exactly α.

An overall P-value is

$$P = 1 - (1 - \min(P_1, P_2))^2$$

where P_1 and P_2 are the P-values for the individual tests of $H_0^{(1)}$ and $H_0^{(2)}$. This is smaller than the Bonferronized P-values $2 \times \min(P_1, P_2)$

Suppose you reject $H_0^{(1)}$: $\mathbf{LB}_1 = \mathbf{0}$ using only $\mathbf{Y}_1$.

Then the test of $H_0^{(2)*}$: $\mathbf{LB}_2^* = \mathbf{0}$ based on $\mathbf{Y}_2$ with covariates $\mathbf{Y}_1$, attempts to answer our question

> Does $\mathbf{Y}_2$ add evidence against the overall H_0 beyond the evidence already provided by $\mathbf{Y}_1$?

When you <u>reject</u> $H_0^{(1)}$ using $\mathbf{Y}_1$ but can't reject $H_0^{(2)*}$, you have rejected the overall H_0 but find no evidence that $\mathbf{Y}_2$ provides additional information about violation of the overall H_0.

When you <u>reject</u> $H_0^{(2)*}$ you can conclude that $\mathbf{Y}_2$ *does* have information about violation of H_0 that $\mathbf{Y}_1$ does not provide.

Example with Fisher iris data,

- Y_1 = sepal data (y[,run(2)])
- Y_2 = petal data (y[,-run(2)])

```
Cmd> irisdata <- read("","t11_05",quiet:T)
Read from file "TP1:Stat5401:Data:JWData5.txt"

Cmd> varieties <- factor(irisdata[,1]);y <- irisdata[,-1]

Cmd> manova("{y[,run(2)]} = varieties",silent:T)#Sepal MANOVA

Cmd> h <- SS[2,,]; e <- SS[3,,]

Cmd> fh <- DF[2]; fe <- DF[3]; p1 <- 2 # p1 = ncols(Y1)

Cmd> vals1 <- releigenvals(h,e); vals1
(1)      4.1718       0.161 Relative eigen values

Cmd> cumtrace(sum(vals1),fh,fe,p1,upper:T)
(1) 5.7321e-131          Highly significant
```

Conclusion 1:

Variety means differ very significantly with respect to sepal dimensions.


```
Cmd> # Do MANACOVA of petal vars with sepal vars as covariates

Cmd> manova("{y[,-run(2)]} = {y[,1]} + {y[,2]} + varieties",\
          silent:T)

Cmd> TERMNAMES # helps to find numbers of terms
(1) "CONSTANT"
(2) "{y[,1]}"        Term for covariate y[,1]
(3) "{y[,2]}"        Term for covariate y[,2]
(4) "varieties"      Hypothesis term = term 4
(5) "ERROR1"         Error term = term 5

Cmd> h2 <- SS[4,,]; e2 <- SS[5,,]

Cmd> vals2 <- releigenvals(h2,e2); vals2
(1)      5.8018    0.044657 Relative eigen values

Cmd> cumtrace(sum(vals2),DF[4], DF[5], 2, upper:T)
(1) 2.2139e-178
```

Conclusion 2:

- Petal dimensions differ among varieties, even after adjusting for sepal dimensions.
- They do add information about differences among varieties.

Sequential F-tests

You can extend this approach to testing H_0, sequentially, variable by variable.

1. Use univariate ANOVA to test H_0: $\mathbf{L}\boldsymbol{\beta}_1 = \mathbf{0}$ for scalar response variable $\mathbf{Y}_1$ (N by 1 vector) in $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p]$.
2. Use univariate ANACOVA to test H_0 for $\mathbf{Y}_2$, *adjusted* for $\mathbf{Y}_1$.
3. Use univariate ANACOVA to test H_0 for $\mathbf{Y}_3$, *adjusted* for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, etc.

At stage j you have a *univariate* problem with test statistic F_j, j = 1,...p.

When the errors are MVN, the F_j are independent and have central (H_0 true) or non-central (H_0 not true) F-distributions.

Fact: When H_0: $\mathbf{LB} = \mathbf{0}$ is true, each F_j is distributed as $F_{f_h,\, f_e - j + 1}$.

- Numerator d.f. = f_h are all the same
- Denominator d.f. = $f_e - j + 1$ drop by 1 for each additional covariate.

Because of independence, to get overall significance level α, for each F-test you use

$$\alpha' = 1 - (1 - \alpha)^{1/p}$$

instead of the Bonferronized $\alpha' = \alpha/p$.

This is better than Bonferronizing since

$$1 - (1 - \alpha)^{1/p} > \alpha/p.$$

The P-value for the overall test of H_0 is

$$P = 1 - (1 - \min(P_1, P_2, \ldots, P_p))^p$$

where P_j is the P-value computed from the observed F_j. $P < p \times \min(P_1, P_2, \ldots, P_p)$, the Bonferronized P-value.

Notes:

- Except for F_1, the sequential F-statistics are *different* from the F-statistics $(h_{jj}/f_h)/(e_{jj}/f_e)$ computed from each variable ignoring all the others.
- Each successive F_j tests whether $\mathbf{Y}_j$ provides information on the violation of H_0 additional to that provided by $\mathbf{Y}_1,\ldots,\mathbf{Y}_{j-1}$.
- The F's depend on the order of the variables so the result of sequential F-tests may depend on the specific ordering of the variables $\mathbf{Y}_j$.

  When you are primarily interested in testing the overall H_0: $\mathbf{LB} = \mathbf{0}$, you can stop once you have found F_j with P-value $< 1 - (1 - \alpha)^{1/p}$.

Here is how the sequential test works with the Fisher iris data.

```
Cmd> anova("{y[,1]}=varieties",fstat:T,pval:F)
Model used is {y[,1]}=varieties
             DF          SS          MS           F
CONSTANT      1      5121.7      5121.7 19326.50528
varieties     2      63.212      31.606   119.26450 = F1
ERROR1      147      38.956     0.26501

Cmd> anova("{y[,2]}={y[,1]}+varieties",fstat:T,pval:F)
Model used is {y[,2]}={y[,1]}+varieties
WARNING: summaries are sequential
             DF          SS          MS           F
CONSTANT      1      1402.1      1402.1 16788.59502
{y[,1]}       1     0.39128     0.39128     4.68513
varieties     2      15.723      7.8613    94.13036 = F2
ERROR1      146      12.193    0.083515

Cmd> anova("{y[,3]}={y[,1]}+{y[,2]}+varieties",fstat:T,pval:F)
Model used is {y[,3]}={y[,1]}+{y[,2]}+varieties
WARNING: summaries are sequential
             DF          SS          MS           F
CONSTANT      1      2118.4      2118.4 26395.47270
{y[,1]}       1      352.87      352.87  4396.78014
{y[,2]}       1      50.022      50.022   623.28867
varieties     2        49.8        24.9   310.25674 = F3
ERROR1      145      11.637    0.080256

Cmd> anova("{y[,4]}={y[,1]}+{y[,2]}+{y[,3]}+varieties",\
 fstat:T,pval:F)
Model used is {y[,4]}={y[,1]}+{y[,2]}+{y[,3]}+varieties
WARNING: summaries are sequential
             DF          SS          MS           F
CONSTANT      1      215.76      215.76  7772.09243
{y[,1]}       1      57.918      57.918  2086.30627
{y[,2]}       1      6.3975      6.3975   230.45189
{y[,3]}       1      16.874      16.874   607.84862
varieties     2      1.3827     0.69137    24.90433 = F4
ERROR1      144      3.9976    0.027761  144
```

Underlined values are sequential F-statistics. All are huge, very significant.

Macro `seqF()` in the new `Mulvar.mac` computes sequential F statistics.

```
Cmd> manova("y = varieties",sscp:F) # do before using seqF()
Model used is y = varieties
WARNING: summaries are sequential
                          SS and SP Matrices
               DF
CONSTANT        1
                    Type 'SS[1,,]' to see SS/SP matrix
varieties       2
                    Type 'SS[2,,]' to see SS/SP matrix
ERROR1        147
                    Type 'SS[3,,]' to see SS/SP matrix

Cmd> stats <- seqF(2); stats # or seqF("varieties")
component: f
      SepLen      SepWid      PetLen      PetWid
      119.26       94.13      310.26      24.904
component: fh
      SepLen      SepWid      PetLen      PetWid
           2           2           2           2
component: fe
      SepLen      SepWid      PetLen      PetWid
         147         146         145         144

Cmd> pvals <- cumF(stats$f,stats$fh, stats$fe, upper:T); pvals
(1)  1.6697e-31  5.4894e-27  4.0983e-53  5.1432e-10
```

These are the ordinary P-values of the 4 sequential F-statistics.

```
Cmd> p <- 4

Cmd> 1 - (1 - min(pvals))^p
(1)          0

Cmd> 4*min(pvals) # valid for small min(pvals)
(1)  1.6393e-52
```

This is the P-value for the test of the overall null hypothesis that the three varieties are identical.

`seqF()` can change the order.

Put petal variables ahead of sepal variables:

```
Cmd> seqF("varieties",order:vector(3,4,1,2))
component: f
      PetLen      PetWid      SepLen      SepWid
      1180.2      24.766      31.289      21.936
component: fh
      PetLen      PetWid      SepLen      SepWid
           2           2           2           2
component: fe
      PetLen      PetWid      SepLen      SepWid
         147         146         145         144
```

Conclusions are the same.

Multi-sample repeated measures profile analysis

Suppose you have g *independent random* samples of sizes n_1, n_2, ..., n_g of p-variable *repeated measures* data from populations with

- means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, ..., $\boldsymbol{\mu}_g$
- *common* variance matrix $\boldsymbol{\Sigma}$.

Example: Subjects randomly assigned to one of g = 3 treatments, with p = 6 measurements x_1, x_2, ..., x_6 of heart rate made on each subject at times 0000h, 0400h, 0800h, 1200h, 1600h, 2000h.

This situation may be viewed as a two-factor repeated measures design with

- a **within-subjects factor** (e.g., time of day) with p levels, and
- a **between-subjects factor** (e.g., treatment or variety) with g levels.

This is a type of g by p factorial experiment.

It is similar to but *not* the same as a split plot design with g whole plot treatments and p subplot treatments.

- Subjects or cases correspond to whole plots
- The between-subjects factor corresponds to a whole plot factor.
- Variables within a subject correspond to subplots
- The within-subjects factor corresponds to the subplot factor.

This differs from a split plot:

- There is no *randomization* of subplot treatments
- There is no assumption that the variance is the same for different subplot treatments ($\sigma_{11} = \ldots = \sigma_{pp}$).

As with any multi-factor design, you are usually interested in testing and estimating

- main effects of each factor
- interactions (differences in effect of one factor between different levels of the other)

Sometimes a univariate split plot ANOVA provides a correct analysis.

This is the case when

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \ldots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \ldots & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \ldots & \rho\sigma^2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \ldots & \sigma^2 \end{bmatrix}$$

- All variances are equal
- All correlations are equal.

When $\mathbf{\Sigma}$ is not of this form, univariate ANOVA may not "work as advertised."

Under somewhat broader conditions you *can* use ANOVA, but with adjustments in degrees of freedom.

The names associated with this are Geisser and Greenhouse (*Ann. Math. Stat* (1958) **29** 885-891, *Psychometrika* **24** (1959) 95-112)

There is an example of such an analysis, with two subplot factors, in Section 10.17 of the MacAnova Users' manual. and another in the profile analysis example handout posted on the web.