Displays for Statistics 5401/8401

Lecture 21

October 24, 2005

Christopher Bingham, Instructor

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

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

MANACOVA

MANOVA Model

In MANOVA a linear model has the form

$\mathbf{y}$ = (μ+$Term_1$+$Term_2$ + ...) + {Error terms}

where a term consists of main effects, interactions or nested effects due to *factors*, that is, *categorical variables*.

One-way MANOVA:

$$\mathbf{y}_{ij} = (\boldsymbol{\mu} + \boldsymbol{\alpha}_j) + \{\boldsymbol{\varepsilon}_{ij}\},\ j = 1,\ldots,g$$

or

$$\mathbf{y}_{ij} = (\boldsymbol{\mu}_j) + \{\boldsymbol{\varepsilon}_{ij}\},\ j = 1,\ldots,g$$

You can always write a MANOVA model in regression form as E[**Y**] = **ZB** where

- **B** is a k+1 by p matrix of means and main effects and interaction effects
- **Z** is a N by k+1 matrix whose columns are "dummy" variables coding for main effects and possibly interactions.

Each hypothesis matrix **H** in the MANOVA corresponds to a null hypothesis of the form H_0: **LB** = **0**, where each row $\boldsymbol{\ell}_i'$ of **L** = $[\ell_{ij}]$ defines a linear combination $\boldsymbol{\ell}_i'$B = $\sum_{0\leq j\leq k}\ell_{ij}\boldsymbol{\beta}_j'$ of the *rows* $\boldsymbol{\beta}_j'$ of **B**.

MANACOVA

In MANACOVA ,in addition you have m ≥ 1 *numerical* variables or covariates u_1, ..., u_m which are correlated with $\mathbf{y}$.

You can arrange these data in a N×m matrix **U** = [$\mathbf{U}_1$,..., $\mathbf{U}_m$] = [$\mathbf{u}_1$,..., $\mathbf{u}_N$]'.

Each variable u_j is to be viewed as a predictor (independent) variable rather than as a response (dependent) variable.

Caution on my notation: These **Z**'s and **U**'s have nothing to do with canonical variables or eigenvectors.

MANACOVA assumes that the dependence of $\mathbf{y}$ on $\mathbf{u}$ is linear.

You can combine the $\mathbf{U}_j$'s with the design matrix **Z** to get an larger linear model.

In pre-computer days, there were special analysis of covariance computations. These were based on MANOVA computations, which were easier than regression computations, at least for balanced designs.

It's now easier just to fit a combined model involving both the MANOVA dummy variables **Z** and the covariates **U**. In the context of this model you test a null hypothesis in the usual linear model way, using the principle of reduction of SSCP matrix of residuals. MacAnova uses the same command `manova()` for this.

The **analysis of covariance** assumes the following:

- Expectation $E[\mathbf{Y} \mid \mathbf{Z}]$ of $\mathbf{Y}$ given $\mathbf{Z}$ but *ignoring* $\mathbf{U}$ is linear in $\mathbf{Z}$:

  $E[\mathbf{Y} \mid \mathbf{Z}] = \mathbf{Z}\,\mathbf{B} = \sum_{0\leq j\leq k} \mathbf{Z}_j \boldsymbol{\beta}_j'$, $\mathbf{B}$ k+1 by p matrix with rows $\boldsymbol{\beta}_j'$

- Expectation $E[\mathbf{U} \mid \mathbf{Z}]$ of $\mathbf{U}$ given $\mathbf{Z}$ but *ignoring* $\mathbf{Y}$ is linear in $\mathbf{Z}$:

  $$E[\mathbf{U} \mid \mathbf{Z}] = \mathbf{Z}\mathbf{D} = \sum_{0\leq j\leq k} \mathbf{Z}_j \boldsymbol{\delta}_j',$$

  $\mathbf{D} = [\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k]'$, k+1 by m, $\boldsymbol{\delta}_j$ m by 1

  $\mathbf{D}$ contains means an effect coefficients in a MANOVA of $\mathbf{U}$.

  If the rows $\boldsymbol{\beta}_j'$ of $\mathbf{B}$ are group means $\boldsymbol{\mu}_j'$ for $\mathbf{Y}$, the rows $\boldsymbol{\delta}_j'$ of $\mathbf{D}$ group means for the covariates in $\mathbf{U}$.

- The expectation $E[\mathbf{Y} \mid \mathbf{U}, \mathbf{Z}]$ of $\mathbf{Y}$ given *both* $\mathbf{U}$ and $\mathbf{Z}$ is linear in $\mathbf{Z}$ *and* $\mathbf{U}$:

  $$E[\mathbf{Y} \mid \mathbf{U}, \mathbf{Z}] = \mathbf{Z}\mathbf{B}^* + \mathbf{U}\boldsymbol{\Gamma} = \sum_{0\leq j\leq k} \mathbf{Z}_j \boldsymbol{\beta}_j^{*\prime} + \sum_{1\leq j\leq m} \mathbf{U}_j \boldsymbol{\gamma}_j'$$

  $\boldsymbol{\Gamma}$ with rows $\boldsymbol{\gamma}_j'$ is a m by p matrix of regression coefficients of $\mathbf{Y}$ on $\mathbf{U}$ in a linear model with both $\mathbf{Z}$ and $\mathbf{U}$ and

  $\mathbf{B}^* = [\boldsymbol{\beta}_0^*, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_k^*]' \equiv \mathbf{B} - \mathbf{D}\boldsymbol{\Gamma}$, k+1 by p is the matrix of means and effects in this larger model

There are at least two different situations where you might use MANACOVA.

Situation 1: $E[\mathbf{u}_i \mid \mathbf{Z}] = \boldsymbol{\delta}_0$ is constant

That is, the means of covariates don't differ among factor levels. The "treatments" don't affect the covariates.

This would be the case, for example, when the covariates are measured *before* treatments were randomly assigned.

Since $\mathbf{Z}_0 = \mathbf{1}_N$, this means

$$\mathbf{D}' = [\boldsymbol{\delta}_0, \mathbf{0}, \ldots, \mathbf{0}],\ \boldsymbol{\delta}_0 = E[\mathbf{u}]$$

$$\mathbf{B}^* = \mathbf{B} - \mathbf{D}\boldsymbol{\Gamma} = [\boldsymbol{\beta}_0 - \boldsymbol{\Gamma}'\boldsymbol{\delta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k]'$$

In this case $\mathbf{B}$ and $\mathbf{B}^*$ are the same except for the intercepts (coefficients of $\mathbf{1}_N$) which are usually of no interest.

Whether you include $\mathbf{U}$ (MANACOVA) or ignore it (MANOVA) in your analysis, $E[\mathbf{Y}]$ has the same dependence on the *non-constant* columns of $\mathbf{Z}$.

You may be able to see what's going on by looking at both models complete with errors for the one-way MANOVA situation with $\boldsymbol{\beta}_0 = \boldsymbol{\mu}$, $\boldsymbol{\beta}_j = \boldsymbol{\alpha}_j$, j=1,...,g-1

Model ignoring $\mathbf{U}$:

$$\mathbf{Y} = \mathbf{1}_N\boldsymbol{\mu}' + \sum_{1\leq j\leq g-1} \mathbf{Z}_j\boldsymbol{\alpha}_j' + \boldsymbol{\varepsilon}$$

Model including $\mathbf{U}$:

$$\mathbf{Y} = \mathbf{1}_N\boldsymbol{\mu}' + \sum_{1\leq j\leq g-1} \mathbf{Z}_j\boldsymbol{\alpha}_j' + (\mathbf{U} - \mathbf{1}_N\boldsymbol{\mu}_u')\boldsymbol{\Gamma} + \boldsymbol{\varepsilon}^*$$

where $\boldsymbol{\varepsilon}^* = \boldsymbol{\varepsilon} - (\mathbf{U} - \mathbf{1}\boldsymbol{\mu}_u')\boldsymbol{\Gamma}$ is the part of $\mathbf{Y}$ that doesn't depends on the factors encoded in $\mathbf{Z}$ or on the covariates $\mathbf{U}$.

$\boldsymbol{\varepsilon}^*$ has a "smaller" variance matrix than $\boldsymbol{\varepsilon}$ in the sense that $V[\boldsymbol{\varepsilon}] - V[\boldsymbol{\varepsilon}^*]$ is positive definite. Other things being equal, the MANACOVA (errors $\boldsymbol{\varepsilon}^*$) is *more sensitive* and *precise* than MANOVA (errors $\boldsymbol{\varepsilon}$).

When $\mathbf{y}$ depends only weakly on $\mathbf{u}$ ($\boldsymbol{\Gamma} \approx \mathbf{0}$), the gain from using covariates may be offset by lost degrees of freedom in $\mathbf{E}$.

R.A. Fisher pioneered correct analysis when there are covariates. He used a univariate example of this type.

- y was the yield of rice subjected to treatments which were randomly assigned to plots which had been used in the previous year in a uniformity trial when all plots were treated the same.
- The covariate u was the yield of rice on the same plot the previous year.

Because of the randomization, there is no way that *last* year's yield u could be affected by *this* year's treatment so E[u] would not differ among treatments.

The purpose of using the previous year's yield was to decrease the MSE in the analysis. This allowed more powerful tests and shorter confidence intervals.

Situation 2: $E[\mathbf{u}_i \mid \mathbf{Z}]$ is not constant

That is, $E[\mathbf{u}]$ differs among the levels of factors in the experiment.

The "treatments" *do* affect $\mathbf{u}$ and consequently there is no simple relationship between $\mathbf{B}$ and $\mathbf{B}^* = \mathbf{B} - \mathbf{D}\boldsymbol{\Gamma}$.

In this case there are *two different* matrices of coefficients, $\mathbf{B}$ and $\mathbf{B}^*$ that describe the dependence of $\mathbf{y}$ on $\mathbf{Z}$ (the effect of the treatments).

Vocabulary

$\mathbf{B}^*$ is the matrix of means and factor effects *adjusted* for $\mathbf{U}$.

In a one-way MANACOVA parametrized by treatment means $\boldsymbol{\mu}_j$ the rows of $\mathbf{B}^*$ would be the treatment means $\boldsymbol{\mu}_j^*$ *adjusted for* $\mathbf{U}$.

Depending on the analyst's goal, you might use *either* $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_k$ or $\boldsymbol{\beta}_1^*,\ldots,\boldsymbol{\beta}_k^*$ to describe dependence of $\mathbf{Y}$ on the factors encoded in $\mathbf{Z}$.

That means that $\mathbf{L}$ specifies two different null hypothesis:

$$H_0: \mathbf{LB} = \mathbf{0}$$
$$(\text{e.g. } \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_g = \mathbf{0})$$

or

$$H_0^*: \mathbf{LB}^* = \mathbf{LB} - \mathbf{LD}\boldsymbol{\Gamma} = \mathbf{0}$$
$$(\text{e.g. } \boldsymbol{\alpha}_1^* = \boldsymbol{\alpha}_2^* = \ldots = \boldsymbol{\alpha}_g^* = \mathbf{0})$$

If one of these is true, the other probably is not unless $\mathbf{LD} = \mathbf{0}$.

You need to decide, on non-statistical grounds, which is the appropriate null hypothesis.

The means and effects in $\mathbf{B}$ describe the overall dependence of $\mathbf{y}$ on the experimental factors, including any *indirect* effects mediated by $\mathbf{u}$ from factors which affect $\mathbf{u}$.

$\mathbf{B}^*$ describes the direct effect of the factors on $\mathbf{y}$ in addition to indirect effects mediated by $\mathbf{u}$.

Interpretation of two situations:

- H_0 is false and H_0^* is true
  The effects being tested are not zero but they are entirely mediated through $\mathbf{u}$.

- Both H_0 and H_0^* are false: The effects being tested are not zero, and are *not* entirely mediated through $\mathbf{u}$.

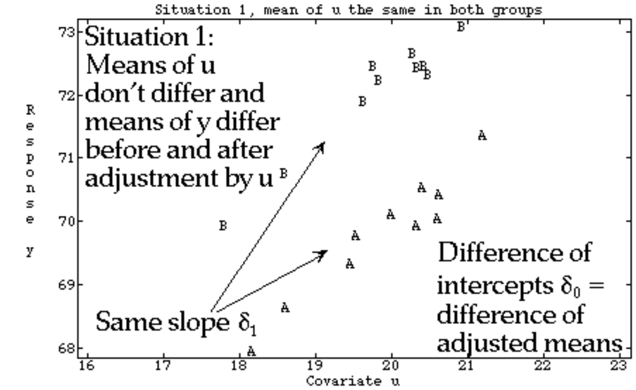
If you are not aware of this difference and routinely do MANACOVA whenever you have covariates, you may "throw out the baby with the bath water" by failing to conclude a treatment has an effect because you can't reject H_0* even though you can reject H_0.

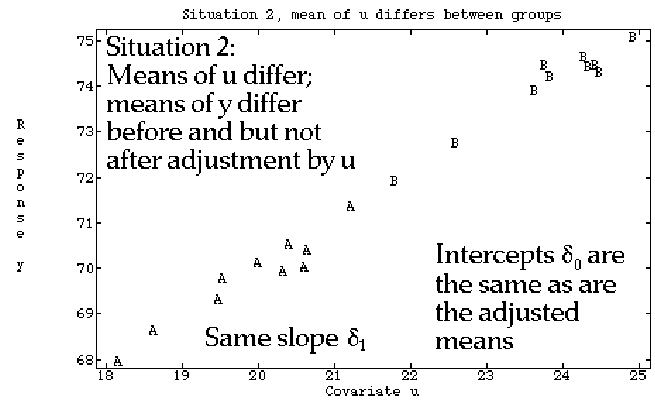
An <u>example</u> is an experiment comparing crop varieties where the response y was the yields on a plot, and the covariate u was a count of the number of "shoots" on the plot, which differed greatly between varieties.

An ANOVA indicated a big difference between varieties in mean yields. But after adjusting for u in ANACOVA, the variety effects lost significance.

The correct conclusion was that yield differed greatly among varieties, but the yield differences were caused by variety differences in the number of shoots.

Here are plots illustrating the two situations when p = m = 1.





Parallelism assumption

The assumption that linear combinations of $\mathbf{Z}\boldsymbol{\beta}_j$ and $\mathbf{U}\boldsymbol{\delta}_j$ enter additively into the model is called the *parallelism assumption*.

With only one covariate (m=1), it states that the slopes of the regressions of **Y** on **U** are the same for *all* treatments groups, that is, the regression lines are parallel.

When parallelism doesn't hold, you may be able to "enlarge" the model to include **Z** by **U** "interaction" terms. The null hypothesis that these additional terms are zero is the parallelism assumption and can be tested.

Without parallelism, the hypothesis of no treatment effect depends on the levels of the covariates. You need to pick a level of the **u** at which to test for or estimate treatment effects.

Parallelism assumption

The assumption that linear combinations of $\mathbf{Z}\boldsymbol{\beta}_j$ and $\mathbf{U}\boldsymbol{\delta}_j$ enter additively into the model is called the *parallelism assumption*.

With only one covariate (m=1), it states that the slopes of the regressions of **Y** on **U** are the same for *all* treatments groups, that is, the regression lines are parallel.

When parallelism doesn't hold, you may be able to "enlarge" the model to include **Z** by **U** "interaction" terms. The null hypothesis that these additional terms are zero is the parallelism assumption and can be tested.

Without parallelism, the hypothesis of no treatment effect depends on the levels of the covariates. You need to pick a level of the **u** at which to test for or estimate treatment effects.

Using manova() for MANACOVA

Suppose the response variables are columns of matrix **y** and there are three <u>covariates</u> in *vectors* **u1**, **u2**, **u3** (*not* factors or columns of a matrix).

And suppose you have a single <u>factor</u> **groups**, that is, you are in a one-way MANOVA/MANACOVA situation

- You compute ordinary MANOVA by
  Cmd> *manova("y=groups")*
  SS[2,,] and **SS[3,,]** are the <u>unadjusted</u> $\mathbf{H}_{\text{groups}}$ and **E**, <u>ignoring covariates</u>
- You compute MANACOVA by
  Cmd> *manova("y=u1+u2+u3+groups")*
  with **groups** the last term (term 5, counting **CONSTANT** as term 1).
  SS[5,,] and **SS[6,,]** are the <u>adjusted</u> $\mathbf{H}_{\text{groups}}$ and **E** matrices since **manova()** fits **groups** after **u1**, **u2**, and **u3** (terms 2, 3 and 4).

After **manova("y=groups")**, **secoefs()** computes *unadjusted* effects and their standard errors, <u>ignoring covariates</u>.

After **manova("y=u1+u2+u3+groups")**, **secoefs()** computes *adjusted* effects and their standard errors.

When covariates are columns of a matrix, you can use **makecols()** to create vectors.

For example, if covariates are in columns 1 through 3 of matrix **data**,

```
Cmd> makecols(data[,run(1,3)], u1, u2, u3)
Column 1 saved as vector u1
Column 2 saved as vector u2
Column 3 saved as vector u3
```

creates vectors **u1**, ..., **u3** containing covariates.

After **manova("y=groups+u1+u2+u3")** you can test $H_0{:}\mathbf{\Gamma} = \mathbf{0}$ (coefficients of **Y** on covariates in model $\mathbf{Y} = \mathbf{ZB} + \mathbf{U\Gamma} + \boldsymbol{\varepsilon}$) because the covariates are *last* in the model.

```
Cmd> hgamma <- SS[3,,] +´SS[4,,] + SS[5,,]#SSCP due to u1,u2,u3
Cmd> e <- SS[6,,]
```


Example with g = 4 groups, p = 3 response variables and 1 covariate.

```
Cmd> data <- read("","X5.9.1") # read from cbmorex.txt
X5.9.1        45     8 FORMAT
) Data on specific gravity and chemicals in urine specimens of
) young men classified into four groups according to their
degree
) of obesity or underweight.  The specific gravity is considered
) to be a concomitant variable (covariate).
) Data from morrison p. 224.  groups have been combined in one
) 45 by 8 matrix, with columns 1 - 4 dummy variables.
) Col. 1: constant column of 1's
) Col. 2: dummy variable for group 1 coded (1,0,0,-1)
) Col. 3: dummy variable for group 2 coded (0,1,0,-1)
) Col. 4: dummy variable for group 3 coded (0,0,1,-1)
) Col. 5: u = (specific gravity - 1) x 10~
) Col. 6: x1 = pigment creatinine
) Col. 7: x2 = chloride
) Col. 8: x3 = choline
Read from file "TP1:Stat5401:Data:cbmorex.txt"
```

I first had to <u>create a factor from the dummy variable columns</u>:

```
Cmd> group <- factor(1*(data[,2]==1) + 2*(data[,3]==1) +\
                     3*(data[,4]==1) + 4*(data[,2]==-1))

Cmd> print(paste(group)) # check that it's right
1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
3 3 3 3 3 4 4 4 4 4 4 4 4   12 1's, 15 2's, 11 3's, 8 4's

Cmd> n <- tabs(,group);n # sample sizes
(1)          12          14          11           8

Cmd> u <- data[,5]; y <- data[,-run(5)] # cols 6, 7, 8
```

MacAnova: When **a**, **b**, ... are factors with length N, **tabs(,a,b,...)** with no argument 1 computes the sizes of the "cells" defined by **a**, **b**, ...

MANACOVA to test groups adjusted for covariate **u**.

```
Cmd> manova("y=u + group") # MANACOVA
Model used is y=u + group
WARNING: summaries are sequential
              SS and SP Matrices
                DF
CONSTANT         1
(1,1)        11370       3752.9       5340.9
(2,1)       3752.9       1238.7       1762.9
(3,1)       5340.9       1762.9       2508.8
u                1
(1,1)       71.949      -65.991      -133.36
(2,1)      -65.991       60.527       122.31
(3,1)      -133.36       122.31       247.17
group            3
(1,1)       142.68       57.976      -25.818 SS[3,,] =
(2,1)       57.976       28.324      -31.295 Adjusted  H_groups
(3,1)      -25.818      -31.295       111.34
ERROR1          40
(1,1)       463.21       52.559      -132.57 SS[4,,] =
(2,1)       52.559       84.042      -46.163 Adjusted  E
(3,1)      -132.57      -46.163       1177.1

Cmd> vals <- releigenvals(SS[3,,], SS[4,,])

Cmd> vals # relative eigenvalues
(1)      0.48767      0.12049    0.0044973

Cmd> addmacrofile("") # make sure new Mulvar.mac is available

Cmd> cumwilks(1/prod(1+vals),DF[3],DF[4],ncols(y))
(1)     0.015923   < .05,  P-value for Wilks
```

$$\mathbf{1/prod(1+vals)} = \prod_\ell (1/(1+\hat{\lambda}_\ell)) = \det(\mathbf{E})/\det(\mathbf{H}+\mathbf{E}) = \Lambda^*.$$

Arguments 2, 3 and 4 are f_h, f_e, and p

Test H_0: Γ = 0

```
Cmd> manova("y=group+u") # test dependence on u (H_0: gamma=0)
Model used is y=group+u          Covariate u is now last
WARNING: summaries are sequential
                SS and SP Matrices
                  DF
CONSTANT           1
(1,1)        11370      3752.9      5340.9
(2,1)       3752.9      1238.7      1762.9
(3,1)       5340.9      1762.9      2508.8
group              3
(1,1)       181.07      40.037     -66.725
(2,1)       40.037      20.038     -41.372
(3,1)      -66.725     -41.372      103.81
u                  1                        fh for testing gamma
(1,1)       33.555     -48.052     -92.448
(2,1)      -48.052      68.812      132.39  H_gamma
(3,1)      -92.448      132.39      254.71
ERROR1            40
(1,1)       463.21      52.559     -132.57
(2,1)       52.559      84.042     -46.163  Same E as before
(3,1)      -132.57     -46.163      1177.1

Cmd> valsu <- releigenvals(SS[3,,],SS[4,,]); valsu
(1)     1.3824 -1.4236e-16 -4.4607e-16   s=min(fh,p) = 1

Cmd> cumwilks(1/prod(1+valsu),DF[3],DF[4],ncols(y))
(1)  2.6704e-07
```

Or since s = $\min(f_h,p) = 1$, you can treat $f_e\hat{\lambda}_1$ as T^2 so $(f_e-p+1)\hat{\lambda}_1/p = F_{p,f_e-p+1}$

```
Cmd> p <- ncols(y); fe <- DF[4]

Cmd> cumF(((fe-p+1)*valsu[1]/p),p,fe-p+1,upper:T)
(1)  2.6704e-07
```

This tests the hypothesis that the slopes of y_j on u are 0 in each group, under the assumption that they are the same in the four groups (parallelism assumption).

Test of parallelism

You can test departure from parallelism by including the term **groups.u** (interaction of **groups** by **u**) last in the model.

```
Cmd> manova("y = group + u + group.u")
Model used is y = group + u + group.u
WARNING: summaries are sequential
                SS and SP Matrices
                  DF
CONSTANT           1
(1,1)        11370      3752.9      5340.9
(2,1)       3752.9      1238.7      1762.9
(3,1)       5340.9      1762.9      2508.8
group              3
(1,1)       181.07      40.037     -66.725
(2,1)       40.037      20.038     -41.372
(3,1)      -66.725     -41.372      103.81
u                  1
(1,1)       33.555     -48.052     -92.448
(2,1)      -48.052      68.812      132.39
(3,1)      -92.448      132.39      254.71
group.u            3
(1,1)       4.9342     -10.888      8.8227
(2,1)      -10.888      25.846     -15.746    H for interaction
(3,1)       8.8227     -15.746      23.802
ERROR1            37        fh
(1,1)       458.28      63.447      -141.4
(2,1)       63.447      58.196     -30.417
(3,1)       -141.4     -30.417      1153.3

Cmd> H <- SS[4,,];E <- SS[5,,];fh <- DF[4]; fe <- DF[5]

Cmd> cumwilks(det(E)/det(E+H),fh,fe,p)
(1)    0.038727
```

It appears there is some evidence that the slope of at least one of the responses differs among groups.


```
Cmd> stats <- secoefs("group.u") # or secoefs(4)

Cmd> tstats <- matrix(stats$coefs/stats$se); tstats
(1,1)    -0.23489     0.62935    -0.58247    Group 1
(2,1)    -0.49734      3.7097    -0.29359    Group 2
(3,1)      0.1049    0.067863       0.532    Group 3
(4,1)     0.44219     -2.8658     0.35789    Group 4

Cmd> 12*twotailt(tstats,fe) # Bonferronized P-values
(1,1)      9.7871      6.3958      6.7654    Group 1
(2,1)      7.4627   0.0081415      9.2485    Group 2
(3,1)      11.004      11.355      7.1749    Group 3
(4,1)      7.9311    0.081864      8.6695    Group 4

Cmd> u0 <-  run(min(u),max(u),(max(u) - min(u))/5)#used in plot

Cmd> for(i,1,p){
        plot(u,y[,i],symbols:vector("\1","\2","\3","\4")[group],\
        title:paste("Variable",i,"by groups"),show:F)
        manova("y=group + group.u - 1",silent:T)
        b0 <- coefs(1), b1 <- coefs(2)
          for(j,1,4){
             addlines(u0,b0[j,i] + b1[j,,i]*u0,symbols:j,show:F)
          }
          showplot(window:i,ymin:?,ymax:?)
     }
```