Multivariate Analysis - MANOVA

Joseph J. Luczkovich

February 25, 2014

Introduction

Sometimes, when we have multiple variables measured on a coastal system, we wish to know how they inter-relate. For example, we may have predictors like temperature, salinity, nitrogen concentration, turbidity, runoff, precipitation, winds, currents, wave energy, sediment type, dissolved oxygen measured in multiple locations. We may also have measurements of plankton, seagrass, fishes collected at each of multiple life stages (larvae, juvenile, adult), and humans fishing success or catch rates in the same locations. We may want to know if a management plan to reduce nutrient inputs (N) has an effect on plankton, seagrass, fishes, and catch rates. How do the physical factors relate to the biological measurements and human use of the coastal locations? Is there a correlation between these variables, and can one or more physical factors predict the biological and human responses? The individual physical variables may be correlated with one another, exhibiting multi-collinearity. In addition, the response variables (adult fish abundance and fisher catch rates) may be correlated as well. Fortunately, we can deal with this issue of inter-variable correlation by creating new variables that are **linear combinations** of the original variables. These new combined variables are created using the familiar least squares methods, and the variation they explain in a response variable or variables is greater, without a multi-collinearity problem. This area of data analysis is known as **multivariate data analysis**. We have already been introduced to multi-factorial ANOVA (two-factor or more), and multiple regression analysis. These are simple kinds of multivariate analyses, but let's look at a situation like the example given here where there are multiple continuous response variables.

MANOVA

When multiple variables are used as response and predictor variables, we can use a new set of procedures that fall under the banner of multivariate analysis. In this section, I will introduce the concept of **multivariate analysis of variance or MANOVA**. MANOVA is an extension of the concepts developed in ANOVA, with multiple means compared across factors, but for multiple response variables, rather than a single response variable. So ANOVA looks at the differences among group means (sum of squares between groups) relative to the differences among observations (sum of squares within groups), MANOVA looks at the differences among **vectors** of multiple response variables between and within groups. The math involves using matrices of group means for each variable and their coefficients that are weighted in such a way as to reduce multicollinearity in the response variables. We will get a multivariate F-test, called the Wilks' Lambda λ as a result of a MANOVA. It tests the significance of the hypothesis that all multivariate vectors are equal. We can look at the univariate F-tests as well, to see which ones are most responsible for the multivariate effects.

Why not use multiple univariate ANOVAs alone? First, there may be significant effects not detected in the univariate tests that reveal themselves when multiple responses are measured. Second, there are type I errors (rejecting the null hypothesis when you should not reject, when the means do not differ) that are inflated when doing the multiple univariate F-tests - if you do enough of them, you will be wrong 5 percent of the time. Finally, it makes sense to measure multiple response variables when you are doing such a large-scale experiment and you are trying to maximize the chance of detecting a difference among groups. What if you choose the wrong response variable, and the difference would have been observed in another variable you did not measure? If you can measure all possible responses of the experimental units, you are better off. In the *Juncus* dredge spoil study, we also measured other plants (*Spartina*) and animal responses (invertebrates and fishes). A MANOVA is the appropriate way to measure the joint responses of the various species involved.

0.1 The Assumptions and Math of MANOVA

The assumptions of MANOVA are similar to ANOVA:

- The variables are multivariate normally distributed (transform if not, look for outliers)
- The variables are linearly related to one another - this allows for the construction of the linear combinations
- The variance is homogeneous across groups (if there is heteroscedacitity of variances across groups, means that you cannot add the sums of squares across groups).
- There is homogeneity of the responses (covariance matrix of the response variables)

For computations and matrix algebra in MANOVA see:
http://userwww.sfsu.edu/efc/classes/biol710/manova/MANOVAnewest.pdf

1 MANOVA Example and R Code

Multivariate comparison of universities and colleges admission standards. Say you have the data:

```
> schools <- read.csv("~/CRM7008/Multivariate ANOVA/MANOVA/schools.csv")
> View(schools)
> schools

           School School_Type  SAT Acceptance X..Student Top10. X.PhD Grad.
1         Amherst     LibArts 1315         22      26636     85    81    93
2      Swarthmore     LibArts 1310         24      27487     78    93    88
3        Williams     LibArts 1336         28      23772     86    90    93
4         Bowdoin     LibArts 1300         24      25703     78    95    90
5       Wellesley     LibArts 1250         49      27879     76    91    86
6          Pomona     LibArts 1320         33      26668     79    98    80
7        Wesleyan     LibArts 1290         35      19948     73    87    91
8      Middlebury     LibArts 1255         25      24718     65    89    92
9           Smith     LibArts 1195         57      25271     65    90    87
10       Davidson     LibArts 1230         36      17721     77    94    89
11         Vassar     LibArts 1287         43      20179     53    90    84
12       Carleton     LibArts 1300         40      19504     75    82    80
13      Claremont     LibArts 1260         36      20377     68    94    74
14        Oberlin     LibArts 1247         54      23591     64    98    77
15 Washington&Lee     LibArts 1234         29      17998     61    89    78
16      Grinnell      LibArts 1244         67      22301     65    79    73
17  Mount Holyoke     LibArts 1200         61      23358     47    83    83
18          Colby     LibArts 1200         46      18872     52    75    84
19       Hamilton     LibArts 1215         38      20722     51    86    85
20          Bates     LibArts 1240         36      17554     58    81    88
21      Haverford     LibArts 1285         35      19418     71    91    87
22        Colgate     LibArts 1258         38      17520     61    78    85
23      Bryn Mawr     LibArts 1255         56      18847     70    81    84
24     Occidental     LibArts 1170         49      20192     54    93    72
25        Barnard     LibArts 1220         53      17653     69    98    80
26        Harvard        Univ 1370         18      46918     90    99    90
27       Stanford        Univ 1370         18      61921     92    96    88
28           Yale        Univ 1350         19      52468     90    97    93
29      Princeton        Univ 1340         17      48123     89    99    93
30       Cal Tech        Univ 1400         31     102262     98    98    75
31            MIT        Univ 1357         30      56766     95    98    86
32           Duke        Univ 1310         25      39504     91    95    91
33      Dartmouth        Univ 1306         25      35804     86   100    95
34        Cornell        Univ 1280         30      37137     85    90    83
35       Columbia        Univ 1268         29      45879     78    93    90
36       Uchicago        Univ 1300         45      38937     74   100    73
37          Brown        Univ 1281         24      24201     80    98    90
38          Upenn        Univ 1280         41      30882     87    99    86
39       Berkeley        Univ 1176         37      23665     95    93    68
40  Johns Hopkins        Univ 1290         48      45460     69    58    86
41           Rice        Univ 1327         24      26730     85    95    88
```

```
42            UCLA        Univ 1142        43      26859     96  100    61
43             UVA        Univ 1218        37      19365     77   91    88
44      Georgetown        Univ 1278        24      23115     79   89    89
45             UNC        Univ 1109        32      19684     82   84    73
46       Umichigan        Univ 1195        60      21853     71   93    77
47  CarnegieMellon        Univ 1225        64      33607     52   84    77
48    Northwestern        Univ 1230        47      28851     77   79    82
49 Washington Univ        Univ 1225        54      39883     71   98    76
50     U Rochester        Univ 1155        56      38597     52   96    73
```

You wish to test the hypothesis that liberal arts colleges are different thn research universities in terms of the response variables: SAT scores, Acceptance Rate, Dollars per Student, Top 10 percent of high school class, Percent of faculty with PhDs, Graduation rate (percent):

```
> Y<-cbind(schools[,3],schools[,4],schools[,5],schools[,6],schools[,7],schools[,8])
> #This binds the school variables by columns into a new data frame Y
> Y2<-cbind(Y[,1:2],1/Y[,3],asin(sqrt(Y[,4:6]/100)))
> #This binds transformed school variables by columns into a new data frame Y2
> fit.Y<-manova(Y~schools[,2])
> fit.Y2<-manova(Y2~schools[,2])
> summary.manova(fit.Y,test="Wilks")
             Df   Wilks approx F num Df den Df    Pr(>F)
schools[, 2]  1 0.45919   8.4405      6     43 4.507e-06 ***
Residuals    48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.manova(fit.Y2,test="Wilks")
             Df  Wilks approx F num Df den Df    Pr(>F)
schools[, 2]  1 0.3623   12.614      6     43 3.743e-08 ***
Residuals    48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #summary specific for manova, test specifies the Wilks' lambda, like F-test
```

Note that the makers of the R **manova**() package like to use Pillai- Bartlett trace statistic, that is the default. Here I specified Wilks' λ, which is way more commonly used, but perhaps not as good. See: help(manova). The interpretation is that the two types of schools (colleges and universities) are significantly different in their student acceptance, expenditures per student, graduatation rates, and faculty doctoral metrics, taken as a whole. The two types are differnt in a multivariate sense. Univariate plots and ANOVA can be done to compare the individual metrics to see which one matters. In SYSTAT, these univarite F-tests are reported in teh MANOVA output, but not in R. R forces you to make teh decision to to them yourself.