Scientific Software International

Analysis of variance and covariance

Suppose there are G groups and we want to compare their means on a response variable y. We may also wish to know if the sensitivity of the analysis can be increased by the use of the covariates, $x_1, x_2, ..., x_q$. These analyses are usually done by one-way analysis of variance (ANOVA) or analysis of covariance (ANCOVA), respectively.

The essential results of analysis of variance can be obtained from LISREL by forming dummy variables $d_1, d_2, ..., d_{G-1}$ and regression y on these dummy variables. Those of analysis of covariance are obtained by regressing y on the covariates and the dummy variables.

The dummy variables represent group memberships such that d_{ig} =1 if case i belongs to group g and d_{ig} = 0 otherwise. The raw data is of the form:

Case	y	x	d_1	d_2	$\ldots$	d_G
1	y_1	x_{11}	d_{11}	d_{21}	$\ldots$	d_{1G}
2	y_2	x_{21}	d_{21}	d_{22}	$\ldots$	d_{2G}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	y_N	x_N	d_{N1}	d_{N2}	$\ldots$	d_{NG}

It is not necessary that the number of cases per group is the same.

The hypothesis that all group means are equal is the same as the hypothesis that $\gamma_1 = \gamma_2 = \cdots = \gamma_{G-1} = 0$. For the analysis of variance a formal F statistic can be computed as

$$F = \frac{R^2 / (G-1)}{\left(1-R^2\right) / \left(N-G\right)},$$

where R^2 is the squared multiple correlation in the regression of y on $d_1, d_2, ..., d_{G-1}$. Each γ_i measures the mean difference $\mu_i - \mu_G$, the significance of which can be tested with the corresponding t-value used as a t-statistic.

The analysis of covariance (ANCOVA) can be done in a similar way. First, regress the response variable *y* on one or more covariates, $x_1, x_2, ..., x_q$, to obtain the squared multiple correlation R^2_{yx}.

Second, regress *y* on $x_1, x_2, ..., x_q$ and $d_1, d_2, ..., d_{G-1}$ to obtain the squared multiple correlation R^2_{yxd}. Then

$$F = \frac{R^2_{yxd} - R^2_{yx} / (G-1)}{\left(1 - R^2_{yxd}\right) / \left(N - G - 1\right)},$$

can be used as a *F* statistic with (*G* – 1) and (N – G – 1) degrees of freedom for testing the hypothesis that the group means, adjusted for mean differences in the covariates, are zero.

Both the ANOVA and the ANCOVA considered above assume that the within-group variances of *y* are equal. In Addition, ANCOVA assumes that the regressions of *y* on $x_1, x_2, ..., x_q$ are the same for each group. These assumptions can be tested using multi-group analysis in LISREL.

The following example, adapted from Huitema (1980), illustrates both ANOVA and ANCOVA.

An experiment is performed to investigate the effects of three different types of study objectives on student achievement in freshman biology. The students are randomly assigned to three different groups and instructed as follows.

1. General: Students are told to know and understand everything in the text.
2. Specific: Students are provided with a clear specification of the terms and concepts they are expected to master and of the testing format.
3. Specific with study time allocations: The amount of time that should be spent on each topic is provided in addition to specific objectives that describe the type of behavior expected on examinations.

The dependent variable is a biology achievement test administered at the end of the course. In addition, an academic aptitude test score, to serve as a covariate, was obtained before the students were assigned to the treatment groups.

Simulated data for this study are shown in the table below. The dummy variables indicating each student's group assignment have been inserted.

Table: Fictitious data for ANOVA and ANCOVA (N = 30)

Biology score	Aptitude score	Dummy variables	Biology score	Aptitude score	Dummy variables
15	29	1 0 0	44	43	0 1 0
19	49	1 0 0	46	64	0 1 0
21	48	1 0 0	47	61	0 1 0
27	35	1 0 0	40	55	0 1 0
35	53	1 0 0	54	54	0 1 0
39	47	1 0 0	14	33	0 0 1

23	46	1 0 0	20	45	0 0 1
38	74	1 0 0	30	35	0 0 1
33	72	1 0 0	32	39	0 0 1
50	67	1 0 0	34	36	0 0 1
20	22	0 1 0	42	48	0 0 1
34	24	0 1 0	40	63	0 0 1
28	49	0 1 0	38	57	0 0 1
35	46	0 1 0	54	56	0 0 1
42	52	0 1 0	56	78	0 0 1

For the ANOVA part of the problem, regress the Biology score (*y*) on d_1 and d_2. The command file for this analysis is (**EX43A.LIS**):

```
Anova
DA NI=5
LA
Y X D1 D2 D3
RA FI=EX43.RAW
SE
1 3 4 /
MO NY=1 NX=2
OU SE TV
```

The squared multiple correlation, R^2, is 0.106. The hypothesis that the three-group means are equal (equivalent to the hypothesis that γ_1 and γ_2 are both zero) is tested with the *F*-statistic with 2 and 27 degrees of freedom. *F* becomes:

$$F = \frac{0.106/2}{(1-0.106)/27} = 1.60.$$

As this value is not significant, no differences in group means of the biology scores are detected.

For the ANCOVA, first regress the Biology scores on the Aptitude scores. The command file (**EX43B.LIS**) is:

```
Ancova Part 1
DA NI=5
LA
Y X D1 D2 D3
RA FI=EX43.TXT
SE
1 2 /
MO NY=1 NX=1
PD
OU ND=3
```

The result is: R^2 = 0.396.

Next, run the regression of y on x, d_1, and d_2 using the command file **EX43C.LIS**:

```
Ancova Part 2
DA NI=5
LA
Y X D1 D2 D3
RA FI=EX43.TXT
SE
1 2 3 4 /
MO NY=1 NX=3
PD
OU ND=3
```

Now the strength of the regression has increased to R^2 = 0.575.

The F-statistic with 2 and 26 degrees of freedom for testing the equality of means, adjusted by the covariate, is

$$F = \frac{(0.575 - 0.396)/2}{\left(1 - 0.575\right)/26} = 5.48.$$

Thus, the result of the more sensitive analysis of covariance is significant at the 5 percent level. There is some evidence that the group means of the Biology Achievement Score are different when they are adjusted for difference in the Aptitude Test Score. By controlling for the covariate, we get a more powerful test of the group differences than when using the response variable alone in the analysis of variance.