June 2003: Correlations need to be substantial to gain advantage in ANCOVA. (New Rule, 6.14).

Introduction

Analysis of covariance has been discussed in the last two months. This month’s discussion also involves ANCOVA. The question is, when does it pay to use a covariate. This question will be addressed only in the context of potential gain in precision. The point can be nicely illustrated through sample size calculations.

Rule of Thumb

ANCOVA requires a substantial correlation of the covariate with the outcome variable in order to gain in efficiency.

Illustration

This example uses the last lines of data for each package from the May 2003 discussion—the data are given only for illustrative purposes. The data are reproduced here.

Table 1. Aspirin content by package and length of storage in months.

Package	Time (X)	Content(mg)
Bottle	0	345
	4	325
	8	342
	12	334
	16	325
	20	317
	24	319
Blister	0	328
	4	334
	8	335
	12	325
	16	331
	20	330
	24	328

These data are analyzed by analysis of variance and covariance. For comparison purposes they are shown together in Table 2.

The average (over the two packet types) correlation is r = –0.5816. This is a substantial value. However, the residual variation is reduced by r^2 = 0.3383 or about 34%. This is precisely the ratio 270.161/798.571=0.3383 of the SS(Regression)/SS(Residual of the ANOVA part from Table 2). That is, the residual sum of squares in Table 2 has been reduced by 270.161, the sum of squares attributable to regression.

Table 2. Analysis of variance and analysis of covariance of the data in Table 1.

	ANOVA		ANCOVA	
Source of Variation	D.F.	S.S.	D.F.	S.S.
Packets	1	1.143	1	1.143
Time			1	270.161
Residual	12	798.571	11	528.411
Total	13	799.714	13	799.714

Basis of the Rule

Suppose that the variability of an endpoint *Y*, without taking a covariate into account, is σ_y^2. If the correlation of *Y* with a covariate *X* is ρ then the variance of *Y* at a specified value of *X* is

$$\sigma_{y|x}^2 = \sigma^2(1-\rho^2).$$

Let *n* be the sample size per group in the usual two-sample comparison, as discussed in Chapter 2, and n_ρ the sample size when a covariate is used, then the two sample sizes are related by.

$$n_\rho = n(1-\rho^2).$$

The variance is reduced by the quantity $(1-\rho^2)$. Since the correlation is bounded by (-1, +1) the square will always be closer to 0 than the original value and $(1-\rho^2)$ closer to 1.

Discussion and Extensions

Table 3 relates n_ρ and *n* for specified values of ρ.

Table 3. Reduction in sample size when covariate is used to reduce variation.

Correlation ρ	$n_\rho = n(1-\rho^2)$
0.0	1*n*
0.1	0.99n
0.2	0.96n
0.3	0.91*n*
0.4	0.84*n*
0.5	0.75*n*
0.6	0.64*n*
0.7	0.51*n*
0.8	0.34*n*
0.9	0.19*n*

This table shows that correlations less than, say, 0.30 are not going to improve the precision of the analysis very much by the use of a covariate. With a correlation of 0.30 the savings in sample size is only 9%—not a huge amount. There is also the loss of one degree of freedom due to the estimation of the correlation. This is not important when the sample sizes are greater than 20, say.

The partitioning of sum of squares attributable to regression and the residual sum of squares is nicely additive in this example because of the balance in the covariate values in the two groups. This may not turn out to be the case when this balance does not exist and/or when the group sizes are not equal.

Of course, if the covariates are there for the taking, there is no reason not to take advantage of them. If there is more than one covariate the multiple correlation coefficient R^2 can be used instead of ρ^2 in calculating the reduction in the residual variance.

An analysis of covariance is usually carried out for two reasons. The first is the reduction of variance. This has formed the basis for the argument above. But there may be a second reason: adjustment for covariate value. For example, adjusting cognitive scores of Alzheimer patients for level of education or age. In randomized experiments there should be no need for this kind of covariate adjustment if the covariate is unaffected by the treatment—as would be the case in the example.

If the treatment affects the covariate as well, then adjustment for the covariate value can be carried out by the ANCOVA procedure. For example, suppose a treatment affects body weight but organ weight is of primary interest. Then organ weights can be compared adjusting for body weight. This is not an innocuous statistical procedure. It should only be carried out when there is a clear understanding of the biological mechanism. Adjustment by body weight could be "overadjustment," that is, removing the effect that is of interest. The same provisos apply to observational studies that typically do not involve randomization and thus there may be differences in the covariates. Again, adjusting for the covariate value may not be appropriate—it depends on the hypothesized causal chain leading from the covariates to the outcome variable. In epidemiology this problem crops up all the time and epidemiologists have developed an extensive vocabulary and methodology for dealing with this. A new book that illustrates these concerns nicely is the text by Koepsell and Weiss; see especially Chapter 11. The methodology of causal models in the social sciences formalizes these issues as well. See for example, the work of Greenland and Brumback (2002).

References

Greenland, S. and Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology,* **31**: 1030-1037.

Koepsell, T.D. and Weis, N.S. (2003). *Epidemiologic Methods—Studying the Occurrence of Illness*. Oxford University Press, Oxford.