BIOSTATISTICS
TOPIC 10: ANALYSIS OF COVARIANCE

I. INTRODUCTION

We have been concerned with general linear model in the analysis of data. In this class of models, we have learned about t-test, analysis of variance and analysis of covariance. Generally speaking, when these variables are all numerical, the linear model is called a *regression model*. When the variables are all categorical, we refer to the *analysis of variance* (ANOVA). While both regression and analysis of variance can be formally subsumed under the GLM, the two techniques have traditionally been treated as distinct. This historical separation occurred for two reasons. First, before high-speed computers were in general use, computational aspect of statistical techniques were of much interest. The most efficient computational procedures for regression and ANOVA were quite different. Second, the two methods tended to be applied to different sorts of problems.

The analysis of variance is usually thought of as a technique for comparing the means of two or more populations on the basis of samples from each. In practice, these populations often correspond to different treatment groups, so that differences in population means may be evidence for corresponding differences in treatment effects.

The ANOVA calculations involves q division of the total sample variance into within-group and between-group components. The within-group component provides an estimate of error variance, while the between-group component estimates error variance plus a function of the differences among treatment means. The ratio of between- to within-group variances provides a test of the null hypothesis that all means are equal. Moreover, the differences among group means provides unbiased estimates of the corresponding population mean differences, and standard errors based on the within-group variance provide confidence intervals for these differences and tests of their significance.

Regression analysis, on the other hand, is primarily used to model relationships between variables. With it, we can estimate the form of a relationship between a response variable and a number of independent variables. We can try to find that combination of variables which is most strongly related to the variation in the response.

The analysis of covariance represents marriage of these two techniques. Its first use in the literature was by R A Fisher (1932), who viewed the technique as one that "combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance".

Combining regression and ANOVA provides the powerful advantage of making possible comparisons among treatment groups differing prior to treatment. Suppose that we can identify a variable X that is related to the outcome Y, and on which treatment groups have different means. We shall assume for simplicity that X is the only variable on which the group differ. Then, if we knew the relationship between Y and X, we could appropriately adjust the observed differences on Y to take account of the differences on X.

II. EXAMPLE

Consider the following data obtained from a nutrition study designed to compare growth of children in an urban environment with that of rural children(Greenberg 1983). Data were height of children in the two samples: one from urban private school and one from rural public school. Differences in growth between these groups might be the result of the different environmental influences operating on the children. In particular, the rural children might be experiencing some nutritional deprivation relative to their urban counterparts. In the terminology of this note, height would be the response or outcome factor and nutrition the risk factor of interest.

The data are shown in the following table.

Table 1: Age and height among urban and rural students.

Urban School			**Rural School**		
Student	Age (months)	Height (cm)	Student	Age (months)	Height (cm)
1	109	137.6	1	121	139.0
2	113	147.8	2	121	140.9
3	115	136.8	3	128	134.9
4	116	140.7	4	129	149.5
5	119	132.7	5	131	148.7
6	120	145.4	6	132	131.0
7	121	135.0	7	133	142.3
8	124	133.0	8	134	139.9
9	126	148.5	9	138	142.9
10	129	148.3	10	138	147.7
11	130	147.5	11	138	147.7
12	133	148.8	12	140	134.6
13	134	133.2	13	140	135.8
14	135	148.7	14	140	148.5
15	137	152.0			
16	139	150.6			
17	141	165.3			
18	142	149.9			

Statistical summary of the data are as follows:

	Mean ± SD	
	Urban	Rural
Height (cm)	144.5 ± 8.6	141.7 ± 6.1
Age (months)	126.8 ± 10.2	133.1 ± 6.5

As a first start, we calculate the simple difference between the two groups:

Difference in height = 144.5 - 141.7 = 2.87

SD of difference = $\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

$= \sqrt{\frac{(18-1)(8.6)^2 + (14-1)(6.1)^2}{18+14-2}}$

= 7.62

The standardised difference is $\frac{2.87}{7.62}$ = 0.38 with p = 0.28.

Thus, the simple t-test analysis reveals that the observed difference between the groups is not statistically significant. It might be concluded that there is no evidence for a difference in nourishment between the urban and rural school children.

Before reaching this conclusion, however, we should consider whether there are likely to be confounding factors. One variable that comes immediately to mind is age. The data on age are also presented in Table 1. The mean age for the rural children is 6.3 months greater than that of the urban children. In a sense, then, the rural children have an "unfair advantage" conferred by their greater average age. Thus, we might expect that if the age distributions were the same, the difference in average height between the groups would be even larger than the observed 2.8 cm. The analysis of covariance allows us to adjust the 2.8 cm difference to obtain a better (less biased) estimate of the difference between groups that would have been observed had the mean ages in the two groups been equal.

In addition to the bias reduction described as above, another benefit results from the combination of regression analysis and ANOVA. Suppose that within treatment groups, a substantial proportion of the variance in Y can be explained by variation in X. In carrying out an ANOVA, we would like the within-group variance to reflect only random error. Regression analysis can be used to remove that part of the error attributable to X and thereby to increase the precision of group comparisons.

It is clear from the data that a substantial proportion of the variation in height is attributed to variation in age. Put differently, if all children in a group were of the same age, the variation in heights within that group would be substantially reduced. Since the relationship between height and age over this range is quite linear, we can estimate the pure error variation by taking the residuals around the regression line relating the two variables. In effect, this is what ANCOVA does, and when a high proportion of within-group variance is explained by the covariate, a large increase in precision results.

III. The ANCOVA Model

To understand the rationale underlying the use of ANCOVA in nonrandomised studies, it is helpful to begin with a somewhat idealised situation. Suppose that on the basis of extensive prior research, the relationship between an outcome and confounding factor can be specified. For example, it might be known that for rural children, the relationship between height and age over the age range being studied can be expressed as:

$$\text{Average height} = 75 + 0.5\ (\text{Age})$$

Suppose that a particular group of rural children have been exposed to some treatment, such as dietary supplement. At the time they are measured this group has a mean age of 132 months and a mean height of 147 cm. Suppose further that another group has been exposed to a different treatment and is measured when the children are 120 months old and on the average the height of this group was 133 cm.

Since the groups differ on mean age, it is not obvious which treatment has been more effective. To make a fair comparison, we must remove the effect of the confounding variable age. However, using the relationship specified above we know that the expected height for the two groups without any special treatment is given by:

Group 1: $\text{Average height} = 75 + 0.5(132) = 141\text{ cm}$

Group 2: Average height = 75 + 0.5(120) = 135 cm

Therefore, the effects of the treatments are:

Group 1: Effect = observed - expected = 147 - 141 = 6 cm.
Group 2: Effect = observed - expected = 133 - 135 = -2 cm.

And the difference between them is (6 - (-2)) = 8 cm.

Alternatively, we can say that because the groups differ by 12 months in age, the relationship predicts that they will differ by 6 cm. So, we could effectively "adjust" the comparison between the two groups by subtracting 6 cm from the difference between them. Since the observed difference is 14 cm, this would leave 8 cm attributable to the difference in treatments received.

Because we are assuming in this example that a known baseline relationship against which to measure performance under the treatments, we can obtain an absolute measure of effect for each treatment (6 cm and -2 cm). In most practical situations, we do not have available such an absolute standard, and we must use only data obtained during the study. Thus, an absolute measure of effect for each group is impossible. On the other hand, it may still be possible to obtain from the data an estimate of the coefficient (0.5 cm/month in our example) relating outcome level to confounding variable. So it may be possible to adjust the observed difference to remove the effect of age from the comparison. In effect, this is how ANOVA is used to estimate treatment effects in non-randomised studies.

The basic model underlying the use of the standard ANOVA assert that there is a linear relationship between the outcome Y and the covariate X with identical slopes in the two groups, but possibly different intercepts. With two treatment groups, we can write the basic model as:

$$\begin{aligned} &Y = \alpha_1 + \beta X + e && \text{in group 1} \\ &Y = \alpha_2 + \beta X + e && \text{in group 2.} \qquad [1] \end{aligned}$$

where

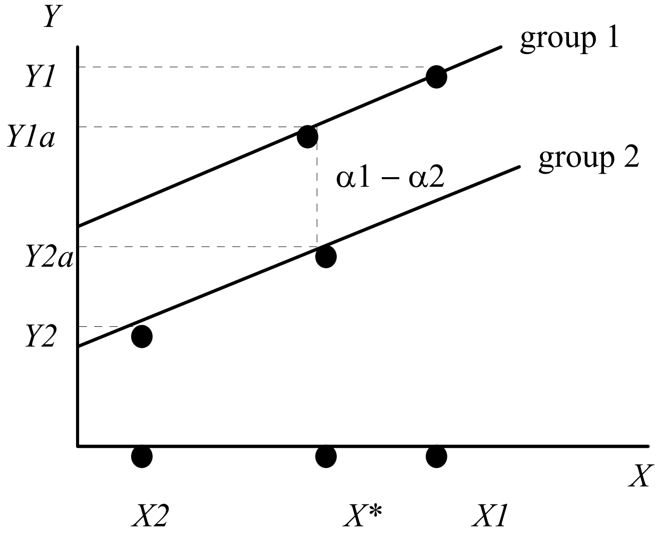
α_1: expected value of Y when $X = 0$ in group 1;
α_2: expected value of Y when $X = 0$ in group 2;
β: the common slope of relationship between Y and X;
e: random variable representing error (with mean 0 for any given X).

Let $\overline{X}$ represent the sample mean of all the X observations in both groups, $\overline{X}_1$, the mean for group 1 and $\overline{X}_2$, the mean for group 2, the following figure illustrates this situation. Note that the direct comparison of $\overline{Y}_1$ and $\overline{Y}_2$ will be biased since $\overline{X}_1 \neq \overline{X}_2$. In fact, taking means in [1] yields:

$$\overline{Y}_1 = \alpha_1 + \beta\overline{X}_2 + e_1$$
$$\overline{Y}_2 = \alpha_2 + \beta\overline{X}_2 + e_2$$

so that the difference between them is:

$$\overline{Y}_1 - \overline{Y}_2 = \alpha_1 - \alpha_2 + \beta\left(\overline{X}_1 - \overline{X}_2\right)$$



Note that in [1] we can interpret $\alpha_1 - \alpha_2$ as the expected difference between the outcomes of the individuals with the same value of X but in two different groups. This difference will represent the differential effect of the two treatments unless there is other variable related to Y which distinguishes the two subjects. To estimate $\alpha_1 - \alpha_2$, we can not simply subtract $\overline{Y}_1$ from $\overline{Y}_2$, but must adjust each of these to move them, in effect, to a common *X* value, say *X**. Let us define the "adjusted" mean of *Y* for group 1 as:

$$\overline{Y}_{1a} = \overline{Y}_1 - \beta\left(\overline{X}_1 - X^*\right)$$

$\overline{Y}_{1a}$ may be interpreted as an estimate of the mean outcome for members of group 1 whose *X* value is *X** . Similarly,

$$\overline{Y}_{2a} = \overline{Y}_2 - \beta\left(\overline{X}_2 - X^*\right)$$

estimate the mean outcome for members of group 2 whose *X* value is *X**. To estimate the difference between the means of the two groups at the same value of *X* (in this case *X**) we can simply take the difference of these two adjusted means:

$$\overline{Y}_{1a} - \overline{Y}_{2a} = \overline{Y}_1 - \overline{Y}_2 - \beta\left(\overline{X}_1 - \overline{X}_2\right)$$

For simplicity, we have not discussed how the value of β necessary for perform the adjustments is actually obtained. In practice, we rarely have any a priori theoretical basis for determining the value β and must therefore use the data to obtain an estimate $\hat{\beta}$. The ANCOVA calculations provide us with an unbiased estimator based on the relationship between *Y* and *X* within the two groups. Thus the adjusted difference is of the form:

$$\overline{Y}_{1a} - \overline{Y}_{2a} = \overline{Y}_1 - \overline{Y}_2 - \hat{\beta}\left(\overline{X}_1 - \overline{X}_2\right)$$

It can be shown that the substitution of an unbiased estimate $\hat{\beta}$ for the unknown true β still yields an unbiased estimate of $\alpha_1 - \alpha_2$ under the model specified by [1].

We should mention in passing that this pooled coefficient is not found by calculating a regression coefficient from the data on both groups taken together as a single group, as is sometimes proposed. This latter approach may be viewed as comparing the mean residuals for the two groups around the overall regression line fitted to the entire sample. It is incorrect, however, in the sense that it does not yield an unbiased estimate of β or of the effect $\alpha_1 - \alpha_2$ under the model given by [1].

IV. ESTIMATION

To estimate the parameters of model 1, we use the multiple regression technique. In this technique, let *Y* be height, *X* be age and a dummy variable *Z* for grouping. Variable *Z* will be coded as 0 for urban and 1 for rural group, then the regression model is:

$$Y = \alpha + \beta X + \gamma Z + e \qquad [2]$$

where α, β and γ are intercept, gradient associated with age, gradient associated with group and *e* is the random error term.

Note that if *Z* = 0 (urban group) then the equation is:

$$Y = \alpha + \beta X + e$$

if *Z* = 1 (rural group) then the equation is

$$Y = \alpha + \beta X + Z + e$$

Using SAS, the parameters of equation 2 are estimated as follows:

Parameter	Coefficient ± SE
Intercept	91.82 ± 17.92
Age (months) i.e. $\hat{\beta}$	0.42 ± 0.14
Group (0, 1)	-5.47 ± 2.57

Thus, the adjusted difference is:

$$\overline{Y}_{1a} - \overline{Y}_{2a} = \overline{Y}_1 - \overline{Y}_2 - \hat{\beta}\left(\overline{X}_1 - \overline{X}_2\right)$$

$$= (144.5 - 141.7) - 0.42(126.8 - 133.1)$$

$$= 5.5 \text{ cm}$$

You may notice that the initial difference was 2.8 cm in favour of the urban children has, after adjustment, been nearly doubled.

We may ask at this point whether this adjusted difference is statistically significant. To answer this question, we can look at the standard error provided as part of the ANCOVA calculations. This standard error can be used to perform a *t* test of

$$H_o : \alpha_1 = \alpha_2$$

More generally, when there are more than two treatment groups (say *k* groups), ANCOVA provides a test of $H_o : \alpha_1 = \alpha_2 = ... = \alpha_k$.

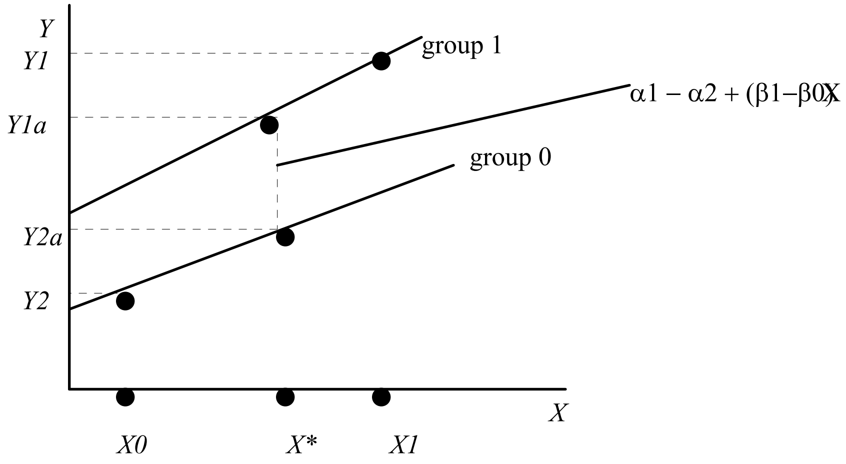
In the above example, the test statistic reveals that the difference was statistically significant at the p = 0.04 level.

V. ASSUMPTIONS

Like any mathematical model attempting to represent reality, the ANCOVA model is never perfectly true. It is more or less accurate abstraction. So, although we may for simplicity discuss whether or not a particular condition holds in a particular situation, it should be remembered that such a statements are only approximate. The real question is whether the ANCOVA model is good enough not to result in misleading results. With this caveat in mind, we now proceed to list the ANCOVA assumptions:

1. Equality of regression slopes. ANCOVA assumes that the relationship between *Y* and *X* in each group differs only in terms of the intercept, not the slope. This assumption

is essential if we are to have the possibility of interpreting the difference between the lines $\left(\alpha_1 - \alpha_0\right)$ as a measure of treatment effect. The problem of non-parallel regressions in different treatment groups is illustrated in the following figure.



The expected difference between two individuals in different groups with identical X value depends on X. Thus, there is no unique summary value which can be interpreted as *the* treatment effect.

In such a situation, we say that there is an *interaction* between treatment groups and the covariate. If an interaction is suspected, it is worthwhile to examine carefully the graph of *Y* versus *X* in the two groups. Visual inspection will usually be adequate to detect serious departure from parallelism.

A formal statistical test for equality of slopes can also be conducted. If such a test is carried out, and the null hypothesis of slopes rejected, we can not apply ANCOVA. If, on the other hand, the null hypothesis is not rejected, we still can not be sure that the slopes are identical. This is a general property if statistical test. Our ability to assert that the null hypothesis in fact holds if it is not rejected is related to the "power" of the test, which is difficult to compute. Generally speaking, however, the power increases with the sample size. So, a statistical test can provide evidence on whether the slopes are equal, but no certainty unless the sample sizes are very large.

2. Linearity. The ANCOVA model assumes a linear relationship between *Y* and *X*. The simplest and usually adequate, test of linearity is to plot graph of *Y* versus *X* in each group. Formal statistical tests of linearity are available if there is any doubt. The simplest involves calculating the regression line in each group and examining the residuals (Topic 8).

3. Covariate measured without error. In some situations, the variable thought to be linearly related to *Y* can not be measured directly, and an imperfect substitute containing some measurement error must be used. When the observed *X*, consisting of in part error, is used in the ANCOVA model, both the estimates and tests may be affected. In both randomised and non-randomised studies, the power of statistical tests will generally decrease as the reliability decreases. As a general rule, it is desirable to use variables with high reliability.

4. No unmeasured confounding variables. The existence of unmeasured variables which are related to the outcome and have unequal distributions in the treatment groups is a general problem in the analysis of cross-sectional studies. To see the problem, let us consider what happens when an ANCOVA model is performed, which does not consider such a variable. Suppose that there exists a variable Z with mean $\overline{Z}_1$ and $\overline{Z}_0$ for the groups. Then, instead of [1], the true model might be described by:

$$Y = \alpha_i + \beta X + \beta Z + e \quad (i = 0,\ 1)$$

In this case, the appropriate adjustment becomes:

$$\overline{Y}_{ia} = \overline{Y}_i - \beta\left(\overline{X}_i - \overline{X}\right) - \gamma\left(\overline{Z}_i - \overline{Z}\right)$$

Thus, if we adjust using *X* only as a covariate, and if $\overline{Z}_1 \neq \overline{Z}_0$, we have adjusted for only part of the differences between groups which is related to *Y*.

5. Equality of error variance. Ordinarily, as in most applications of linear models, it is assumed that all error terms have the same variance. In ANCOVA situation, it is possible that the treatment groups have different error variances. The estimates of treatment effects will still be unbiased in this case, but the validity of tests may be

affected. If there is some reason to suspect this inequality of error variances, the residuals from the fitted lines in the two groups can be compared. If the variances of these residuals differ greatly, caution in the interpretation of test results is advised.

6. Normality of errors. For the ANCOVA tests to be strictly valid, it must be assumed that the errors follow a normal distribution. Departures from normality may affect statistical tests and the properties of estimators in a variety of ways, depending on the actual form of the error distribution of residuals. The normality assumption can be tested by examining the distribution of residuals. While severe departures from normality may affect the properties of tests, ANCOVA appears to be generally rather robust. Thus, most researchers assume that the normality assumption is not critical.

V. MATHEMATICAL DETAILS

We consider the general situation where *K* treatment groups are being compared. These will be indexed by $k = 1, 2, 3, \ldots, K$. Let X_{ik} and Y_{ik} represent the covariate and outcome values for individual *i* in group *k*. Let $\overline{X}_k$ and $\overline{Y}_k$ be the means for the n_k individuals in group *k*. The, we can define the between-group sums of squares and cross-products by:

$$T_{xx} = \sum_{k=1}^{K} n_k \left(\overline{X}_k - \overline{X}\right)^2$$

$$T_{yy} = \sum_{k=1}^{K} n_k \left(\overline{Y}_k - \overline{Y}\right)^2$$

$$T_{xy} = \sum_{k=1}^{K} n_k \left(\overline{X}_k - \overline{X}\right)\left(\overline{Y}_k - \overline{Y}\right)$$

where $\overline{X}$ and $\overline{Y}$ are overall (grand) means of *X* and *Y* across all groups. Similarly, we define within-group (error) sums of squares and cross-products by:

$$E_{xx} = \sum_{k=1}^{K} \sum_{i} \left(X_{ik} - \overline{X}_k\right)^2$$

$$E_{yy} = \sum_{k=1}^{K} \sum_{i} \left(Y_{ik} - \overline{Y}_k\right)^2$$

$$E_{xy} = \sum_{k=1}^{K} \sum_{i} \left(X_{ik} - \overline{X}_k\right)\left(Y_{ik} - \overline{Y}_k\right)$$

where $\sum_{i}$ denotes the sum over individuals within each group. We also define the quantity:

f = total number of subjects minus number of groups

= N - K

and, using the definitions above, we have:

$$S_{xx} = T_{xx} + E_{xx}$$

$$S_{xy} = T_{xy} + E_{xy}$$

$$S_{yy} = T_{yy} + E_{yy}$$

Then, we can calculate the residual mean squares for treatments and error:

$$s_e^2 = \frac{\left(E_{yy} - \dfrac{E_{xy}^2}{E_{xx}}\right)}{f - 1}$$

$$s_t^2 = \frac{\left(T_{yy} - \dfrac{S_{xy}^2}{S_{xx}} + \dfrac{E_{xy}^2}{E_{xx}}\right)}{K - 1}$$

These can be used to calculate an F statistic to test the null hypothesis that all treatment effects are equal:

$$F = \frac{s_t^2}{s_e^2}$$

Under the null hypothesis, this ratio has an F distribution with *K*-1 and *f* - 1 degrees of freedom. The estimated regression coefficient of *Y* on *X* is then:

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}}$$

From the definitions of E_{xx} and E_{xy} given above, it is clear why this is called pooled within-group estimator. The estimated standard error for the adjusted difference between group means (say group 0 and group 1) is given by:

$$s_d = s_e \sqrt{\frac{1}{n_0} + \frac{1}{n_1} + \frac{\left(\overline{X}_1 - \overline{X}_0\right)^2}{E_{xx}}}$$

where n_0 and n_1 are the sample size of the two groups. A test of the null hypothesis that the adjusted difference is zero is provided by the statistic:

$$t = \frac{\overline{Y}_1 - \overline{Y}_0 - \hat{\beta}\left(\overline{X}_1 - \overline{X}_0\right)}{s_d}$$

Under the null hypothesis, it has a *t* distribution with *f*-1 df.

VI. EXERCISES