CHAPTER 6

Adjustments in Analysis

6.1 INTRODUCTION

When the principal confounding x variables have been noted, the main alternative to matching, as discussed in Section 5.1, is to make adjustments in the course of the statistical analysis. The objectives remain the same—to protect against bias and to increase the precision of the comparision between treatment means or proportions.

In some situations an adjustment method is the only possibility because matching is not feasible or is obviously unattractive. The economics of the study may require that the x's and y be measured simultaneously, after the samples have been chosen, so that advance creation of matched samples is ruled out. As mentioned, matching is confined mainly to smaller-sample and simpler studies, often two-group comparisons. Matching becomes troublesome with large samples, when subjects enter the study only over an extended time period, and also as the number of treatments to be compared or the number of variables to be matched increases.

This chapter describes the principal methods of adjustment in the simpler situations. Where possible, comparisons of the performance of matching and adjustment methods will be noted, since many studies could use either method. Once again, the details of the adjustment method depend on the scales in which y and the x's are measured.

6.2 y CONTINUOUS: x's CLASSIFIED

With two populations we assume that the adjustment method starts with independent random samples. Having selected the classified x variables for which adjustment is to be made, we first arrange the data from the two

samples into the cells created by this classification. Let the sample numbers in the ith cell of the classification be n_{1i} and n_{2i}. The response means are $\bar{y}_{1i}$ and $\bar{y}_{2i}$ and the proportions (if y is 0 and 1) are p_{1i} and p_{2i}. The only difference between this situation and within-class matching is that in the latter, $n_{1i} = n_{2i} = n_i$ in every cell.

If $d_i = \bar{y}_{1i} - \bar{y}_{2i}$, the estimates of the overall treatment difference for the two methods are

$$\bar{d} = \bar{y}_1 - \bar{y}_2 = \sum \frac{n_i}{n} d_i \qquad \text{(matched samples)}$$

and

$$\bar{d}_a = \bar{y}_{1a} - \bar{y}_{2a} = \sum W_i d_i \qquad \text{(random samples)}$$

The matched-sample weights, n_i/n, weight each class mean by its class size, resulting in simply the difference between two means; the weights W_i, with $\Sigma W_i = 1$, are chosen by the investigator; and the subscript a denotes adjustment. In both methods any remaining bias arises from the fact that $E(d_i) \neq 0$ when the underlying confounding x variables have different distributions in the two populations. If $E(d_i)$ were constant from cell to cell, matching and adjustment would be equally effective in reducing bias for any choice of weights W_i. In fact, in the presence of bias, $E(d_i)$ varies from cell to cell and is usually greater at the high and low extremes of the x distributions than near the medians. However, for the weights likely to be used in practice, within-class matching and adjustment may be regarded as roughly equally effective in reducing bias. Matching, though, has the advantage of a simpler estimate $\bar{y}_1 - \bar{y}_2$ that avoids weighting.

We now consider the choice of weights. Suppose first that the mean difference $\delta = \tau_1 - \tau_2$ between the effects of the two treatments is the same in every cell. It follows that in random samples *any* weighted mean $\Sigma W_i d_i$ is an estimate of δ, apart from within-cell bias in d_i. If this situation holds, the choice of weights may be determined by considering convenience or statistical precision. If, however, $\tau_{1i} - \tau_{2i} = \delta_i$ varies from cell to cell, $\Sigma W_i d_i$ becomes an estimate of $\Sigma W_i \delta_i$, a quantity whose value now depends on the choice of weights. It may be clear on inspection of the data, particularly with large samples, that there are real differences between the effects of the treatments and that these differences vary from cell to cell. Section 6.4 discusses this further. For now, we assume δ constant and discuss the estimation and testing of δ.

If σ_{1i}^2 and σ_{2i}^2 are the two population variances within the ith cell,

$$V(\bar{d}_a) = V\left(\sum W_i d_i\right) = \sum W_i^2 \left(\frac{\sigma_{1i}^2}{n_{1i}} + \frac{\sigma_{2i}^2}{n_{2i}} \right) \qquad (6.2.1)$$

This variance is minimized by taking W_i proportional to

$$w_i = \frac{1}{\sigma_{1i}^2/n_{1i} + \sigma_{2i}^2/n_{2i}}$$

Of course, $W_i = w_i/\Sigma w_i$. The resultant minimum variance is $1/\Sigma w_i$.

These general formulas are needed when assumptions of equality of within-class variances cannot be made. If all n_{1i} and n_{2i} exceed 30, choice of w_i inversely proportional to the estimated variance

$$\hat{w}_i = \frac{1}{s_{1i}^2/n_{1i} + s_{2i}^2/n_{2i}} \tag{6.2.2}$$

should do almost as well, where $V(\bar{y}_{1a} - \bar{y}_{2a}) \doteq 1/\Sigma\hat{w}_i$ [Meier (1953)].

Various particular cases arise when assumptions can be made about the within-cell variances. In the formulas below, s_i^2, s_1^2, s_2^2, and s^2 are the appropriate pooled estimates of σ_i^2, σ_1^2, σ_2^2, and σ^2 respectively

$$\sigma_{1i}^2 = \sigma_{2i}^2 = \sigma_i^2; \qquad \hat{w}_i = \frac{n_{1i}n_{2i}}{(n_{1i} + n_{2i})s_i^2} \tag{6.2.3}$$

$$\sigma_{1i}^2 = \sigma_1^2;\ \sigma_{2i}^2 = \sigma_2^2; \qquad \hat{w}_i = \frac{n_{1i}n_{2i}}{n_{2i}s_1^2 + n_{1i}s_2^2} \tag{6.2.4}$$

and

$$\sigma_{1i}^2 = \sigma_{2i}^2 = \sigma^2; \qquad \hat{w}_i = \frac{n_{1i}n_{2i}}{(n_{1i} + n_{2i})s^2} \tag{6.2.5}$$

In this case [(6.2.5)] it is customary to take $w_i' = \hat{w}_i s^2 = n_{1i}n_{2i}/(n_{1i} + n_{2i})$ as relative weights. The variance of $\Sigma w_i'\bar{d}_i/\Sigma w_i'$ is

$$\frac{s^2}{\Sigma w_i'} = \frac{s^2}{\Sigma[n_{1i}n_{2i}/(n_{1i} + n_{2i})]}$$

Formula (6.2.5) is often used when the σ_{ti}^2 are thought not to vary much.

When the σ_{ti}^2 vary little and the two total sample sizes are equal, the simpler weights, w_i proportional to $n_{1i} + n_{2i} = n_i$, the combined sample size in cell i, seldom do much worse than the optimum weights $n_{1i}n_{2i}/n_i$ in (6.2.5) for this case. With weights proportional to n_i and with σ_{ti}^2 constant, for instance,

$$\hat{V}(\bar{y}_{1a} - \bar{y}_{2a}) = \frac{s^2\Sigma(n_i^3/n_{1i}n_{2i})}{(\Sigma n_i)^2} \tag{6.2.6}$$

In general, no simple statement can be made about the relative precision of the comparison $\bar{y}_1 - \bar{y}_2$ for matched samples and $\bar{y}_{1a} - \bar{y}_{2a}$ for adjusted random samples both of total size n, because this depends on the way in which the σ_{ti}^2's vary from cell to cell and on the choice of weights. However, this comparison is of interest when any underlying x has the same distribution in the two populations. In this situation the purpose of matching or adjustment is to increase precision, because there is no danger of bias. Any difference in variances of $\bar{y}_1 - \bar{y}_2$ and $\bar{y}_{1a} - \bar{y}_{2a}$ is likely to be minor because differences between n_{1i} and n_{2i} in any cell will arise only from random-sampling variation. If $n_{1i} = n_{2i} = n_i/2$, then the variance of (6.2.6) becomes $4s^2/\Sigma n_i$ as does the variance derived from the conditions of (6.2.5).

6.3 *y* BINOMIAL: *x*'s CLASSIFIED

The adjusted mean difference with two random samples is of the form $\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})$, where the $\hat{p}_{ti}$ are the observed proportions of "ones" (successes) in cell i. As with y continuous, the difference in effectiveness of matching and adjustment depends on the within-cell biases and the choice of the w_i. The difference should be small for most choices of the w_i in practice.

In choosing the w_i for an initial test of significance or estimation of an overall difference (ignoring within-cell bias), some theoretical issues have to be considered. As indicated in Section 5.10, the assumption of an additive model of the form

$$p_{1i} = \mu + \tau_1 + \gamma_i; \qquad p_{2i} = \mu + \tau_2 + \gamma_i$$

is unreasonable on logical grounds, since the p_{ti} must lie between 0 and 1. Most recent work on the analysis of proportions in multiple classifications has assumed an additive model in the scale of $\log(p_{ti}/q_{ti})$, where $q_{ti} = 1 - p_{ti}$. In this situation the investigator may still wish to estimate an overall difference $p_1 - p_2$, particularly if it is not clear that there is a real difference in treatment effects, so that a test of significance is desired.

Under an additive model in the logit scale, Cochran (1954) has shown that an effective choice is to take w_i proportional to $n_{1i}n_{2i}/n_i$, where $n_i = n_{1i} + n_{2i}$. An approximate test of significance of the null hypothesis,

$p_{1i} = p_{2i}$ (for all i), is made as follows. [This test does not fully address the null hypothesis stated, but uses its assumptions. The test is directed to the more-general null hypothesis $\Sigma w_i(p_{1i} - p_{2i}) = 0$, which includes the stated null hypothesis.] With $w_i = n_{1i}n_{2i}/n_i$, the weighted difference $\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})$ has approximate estimated variance

$$\sum w_i \hat{p}_i \hat{q}_i$$

where $\hat{p}_i$ is the overall proportion of successes in cell i. The test is made by treating

$$\frac{\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})}{(\Sigma w_i \hat{p}_i \hat{q}_i)^{1/2}}$$

as a normal deviate.

Two refinements by Mantel and Haenszel (1959), who also developed this test from a different viewpoint, are worth using when some n_{ti} are small, as often occurs. One technique involves inserting a correction for continuity; the other uses a slightly different variance formula. In this form the normal deviate for a test of significance is taken as

$$\frac{|\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})| - \frac{1}{2}}{[\Sigma n_{1i}n_{2i}\hat{p}_i\hat{q}_i/(n_{1i} + n_{2i} - 1)]^{1/2}}$$

If we conclude that there is a real overall difference, the null hypothesis being rejected, and if we wish to attach a standard error to the weighted mean difference $\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})/\Sigma w_i$, then we can no longer regard $p_{1i} = p_{2i}$. The estimated standard error of the weighted mean difference is

$$\left[\sum w_i^2\left(\frac{\hat{p}_{1i}\hat{q}_{1i}}{n_{1i} - 1} + \frac{\hat{p}_{2i}\hat{q}_{2i}}{n_{2i} - 1}\right)\right]^{1/2} \Big/ \sum w_i$$

6.4 TREATMENT DIFFERENCE VARYING FROM CELL TO CELL

In this situation the choice of weights should not be dictated by considerations of precision, particularly in large-sample studies in which any reasonable weighting gives adequate precision. Several situations in choosing weights may arise. The investigator may find that on trying several likely sets of weights, the estimates $\Sigma w_i(\bar{y}_{1i} - \bar{y}_{2i})$ or $\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})$, while differing from set to set, agree sufficiently well that any conclusions to be drawn, or any action to be taken, would be the same. If the estimates disagree more widely, the choice of weights is more critical. This problem is old and familiar in vital statistics, for example, in the international comparision of overall death rates. One device used there is to take the weights from some standard population that is regarded as a target population. The process of adjustment is called *standardization*. To illustrate, Keyfitz (1966) reports that the mortality rate for French females in 1962 exceeded the rate for American females in 1963 by 14, 16, or 51%, according to the different standard populations used for weighting.

In the face of substantial differences between estimates based on different sets of weights, any overall estimate $\hat{\delta}$ may be, to some extent, arbitrary and liable to misinterpretation unless there is a specific target population with known weights for which an estimate of δ is clearly relevant. Otherwise, the most useful report on the data may be to summarize and try to interpret how d_i varies from cell to cell.

In both matched and independent samples, rough tests of the null hypothesis that the δ_i are the same in each cell are possible, and sometimes helpful. The simplest case is one in which y is continuous and the within-cell variance can be assumed constant. This is unlikely to be strictly true in observational studies, but might not be seriously wrong. Let s^2 denote the pooled within-cell variance and

$$w_i' = \frac{n_{1i}n_{2i}}{n_i}; \qquad d_i = \bar{y}_{1i} - \bar{y}_{2i}; \qquad \bar{d}_a = \frac{\Sigma w_i' d_i}{\Sigma w_i'}$$

Calculate the weighted sum of squares

$$Q = \Sigma w_i'(d_i - \bar{d}_a)^2 = \Sigma w_i' d_i^2 - \frac{(\Sigma w_i' d_i)^2}{\Sigma w_i'} \qquad (6.4.1)$$

With c classes, assuming normality, the quantity $Q/(c - 1)s^2$ is distributed on the null hypothesis as F with $(c - 1)$ and $(n_1 + n_2 - 2c)$ d.f. (degrees of freedom). Large values of F cause rejection of the null hypothesis.

With y binomial, and with y continuous, when within-cell variances vary, a method is available which leads to a large-sample χ^2 test. Let $d_i = \bar{y}_{1i} - \bar{y}_{2i}$, or $\hat{p}_{1i} - \hat{p}_{2i}$. Compute an unbiased estimate of the variance of d_i according to the assumptions that seem reasonable. In the general case,

$$\hat{V}(d_i) = \frac{s_{1i}^2}{n_{1i}} + \frac{s_{2i}^2}{n_{2i}} \qquad (y \text{ continuous})$$

and

$$\hat{V}(d_i) = \frac{\hat{p}_{1i}\hat{q}_{1i}}{n_{1i}-1} + \frac{\hat{p}_{2i}\hat{q}_{2i}}{n_{2i}-1} \qquad (y \text{ binomial})$$

Assume $\hat{w}_i = 1/\hat{V}(d_i)$. Then on the null hypothesis, $\Sigma\hat{w}_i(d_i - d_w)^2$ is approximately χ^2 with $(c - 1)$ d.f. where $d_w = \Sigma\hat{w}_i d_i$. [Sometimes assumptions such as those given in Eqs. (6.2.3), (6.2.4), and (6.2.5) seem natural and lead to slightly different formulas.]

The large-sample χ^2 test may be directed at more-specific alternatives by breaking χ^2, or the numerator Q of F in (6.4.1), into components. For instance, the c cells might subdivide into three sets, with reason to expect d_i to be constant within each set, but to vary from set to set. Then χ^2 is broken down into four components—one for "between sets" and one for "within each set." Similarly, if the cells represent an ordered classification, with scores z_i assigned to the cells, the test of the linear regression of d_i on z_i may be of interest. (Remember to take the different weights into account.) Thus in the linear-regression test for χ^2 with 1 d.f., we calculate

$$N = \sum \hat{w}_i d_i z_i - \frac{(\Sigma\hat{w}_i d_i)(\Sigma\hat{w}_i z_i)}{\Sigma\hat{w}_i}$$

$$D = \sum \hat{w}_i z_i^2 - \frac{(\Sigma\hat{w}_i z_i)^2}{\Sigma\hat{w}_i}$$

and

$$\chi_1^2 = \frac{N^2}{D}$$

When the cells or classes represent a single x variable, interpretation of the finding of significant variation in δ_i by the preceding methods is straightforward. With two x's, rejection of the null hypothesis does not reveal whether the variation in δ_i is associated primarily with x_1, with x_2, or partly with both. Further, if δ_i varies moderately with one of the x's but not with the other, the F test may lack the power to reject the null hypothesis; however, a test directed at x_1 and x_2 separately may reveal the correct state of affairs.

6.5 *y* AND *x*'s QUANTITATIVE: ADJUSTMENTS BY REGRESSION (COVARIANCE)

When y and the x's are quantitative, an approach that avoids matching first constructs a mathematical model for the regression of y on the x's, usually assumed of the same form in each population. This regression is then estimated from the sample data and used to adjust $\bar{y}_1 - \bar{y}_2$ in the unmatched samples for differences between the x distributions in the two populations.

In practice, linear regression is the most-frequent form. With k x variables, where x_{j1u} and x_{j2v} denote sample members of the jth x variate in populations 1 and 2, the linear model is

$$y_{1u} = \tau_1 + \sum_{j=1}^{k} \beta_j x_{j1u} + e_{1u}; \qquad y_{2v} = \tau_2 + \sum_{j=1}^{k} \beta_j x_{j2v} + e_{2v} \qquad (6.5.1)$$

This assumes a constant effect $\delta = \tau_1 - \tau_2$ of the difference between the two treatments. Assuming that e_{1u} and e_{2v} have the same variance, the sums of squares and products $\Sigma(yx_j)$, $\Sigma(x_j x_j)$ and $\Sigma(x_j x_m)$ in the normal equations are the pooled within-treatment values. The notation $\Sigma(yx_j)$ is shorthand; in a notation used earlier it means $(yx_j)_1 + (yx_j)_2$ and each of these terms is a sum of products of deviation scores for the treatment group indicated by the trailing subscript. The adjusted mean difference is

$$\bar{y}_{1a} - \bar{y}_{2a} = \bar{y}_1 - \bar{y}_2 - \sum_{j=1}^{k} b_j(\bar{x}_{j1} - \bar{x}_{j2}) \qquad (6.5.2)$$

where the b_j are the estimated regression coefficients. With random samples from each population and a correct mathematical model, the adjusted mean difference is an unbiased estimate of δ under this model.

For the standard error of $\bar{y}_{1a} - \bar{y}_{2a}$ we need s^2, the pooled mean-square deviation from the multiple regression, with $(n_1 + n_2 - k - 2)$ d.f. and the inverse $C = \|c_{jm}\|$ of the matrix $\|\Sigma(x_j x_m)\|$ in the normal equations. The standard error of $\bar{y}_{1a} - \bar{y}_{2a}$ equals

$$s\left(\frac{1}{n_1} + \frac{1}{n_2} + \sum_{j=1}^{k} c_{jj}\bar{d}_j^2 + 2\sum_{j=1}^{k}\sum_{m>j}^{k} c_{jm}\bar{d}_j\bar{d}_m\right)^{1/2} \qquad (6.5.3)$$

where $\bar{d}_j = (\bar{x}_{j1} - \bar{x}_{j2})$.

Two precautions are worth noting. Although linear-regression adjustments are the most widely used and are often assumed to hold without checking, we can examine and test for the simpler types of curvature in the regression of y on any x_j by adding a variate $x_{k+1} = x_j^2$ to the model. With large samples and a good computer program, the practice of adding a term in x_j^2 is worthwhile when there are reasons to expect curvature or indications of it. Sometimes a linear regression on a simple transform of x such as $\log x$ or e^{-x} is a satisfactory alternative.

Another precaution in a two-sample regression is to estimate the regression separately in each sample and compare the regression coefficients. The method of adjustment in this section assumes that the regression is the same in the two populations. Possible alternatives when the regression is linear, but differs in the two populations, are outlined in Section 6.7.

Comparison of the regressions in the two samples might reveal that although the regressions appeared to have the same shape, the residual mean squares s_1^2 and s_2^2 are substantially different. In this event, more-precise estimates of the b_j would probably be obtained by weighting the contribution from each sample by $1/s_i^2$ when forming $\Sigma(yx_j)$ and $\Sigma(x_jx_m)$, instead of simply adding. The gain in precision as it affects the estimated treatment effect is, however, usually small.

6.6 REGRESSION ADJUSTMENTS WITH SOME x's CLASSIFIED

The regression method applies most naturally when y and all the x's are quantitative. If one or more of the x's are classified while the others are quantitative, there are two almost-equivalent methods of making the adjustments. To take the simplest case, suppose that one of the x's is a two-class variate. The subscripts $t = 1, 2$ denote the populations or treatments, $i = 1, 2$ the classes, and $j = 1, 2, \ldots, k$ the quantitative x variates. The linear model is assumed to be

$$y_{tiu} = \tau_t + \gamma_i + \sum_{j=1}^{k} \beta_j x_{jtiu} + e_{tiu} \qquad (u = 1, 2, \ldots, n_{ti}) \qquad (6.6.1)$$

1. The first method is the *analysis of covariance*. For the quantitative x's, calculate the quantities $\Sigma(yx_j)$ and $\Sigma(x_jx_m)$ from the pooled sums of squares or products within classes and treatments. These will have $(n_1 + n_2 - 4)$ d.f. with two classes and two treatments. Having computed the b_j, take the adjusted y difference, the estimate of $(\tau_1 - \tau_2)$ as

$$\bar{y}_{1a} - \bar{y}_{2a} - \sum_{j=1}^{k} b_j(\bar{x}_{j1a} - \bar{x}_{j2a})$$

Here $\bar{y}_{1a}$, $\bar{y}_{2a}$, $\bar{x}_{j1a}$, and $\bar{x}_{j2a}$ are adjusted means over the two classes, with weights proportional to $n_{1i}n_{2i}/n_i$ in class i.

2. Instead, the adjustments can be performed by an ordinary one-sample multiple regression. The primary advantage of this method is that, at present, computer programs for one-sample multiple regressions are more widely available than those for the combination of analysis of variance and multiple regression. Construct two dummy x variables: $x_{k+1,tiu}$ which has the value $+1$ for all observations from treatment 1 and the value 0 for all observations from treatment 2, and $x_{k+2,tiu}$ which has the value $+1$ for all observations in class 1 and the value 0 for all observations in class 2. Fit the multiple-regression model

$$y_{tiu} = \mu + \sum_{j=1}^{k+2} \beta_j x_{jtiu} + e_{tiu} \qquad (6.6.2)$$

Then b_{k+1} is the adjusted estimate of $\tau_1 - \tau_2$. In fact, many computer programs perform the calculations by constructing a third dummy variable, say x_{0tiu}, which takes the value $+1$ for all observations, so that μ in (6.6.2) is replaced by $\beta_0 x_{0tiu}$. It is easily verified that models (6.6.1) and (6.6.2) are identical. The computations in methods 1 and 2, as presented here, are not exactly identical. In method 1, the quantities $\Sigma(yx_j)$ and $\Sigma(x_jx_m)$ for the quantitative x's are based on $(n_1 + n_2 - 4)$ d.f., while in method 2 they are, in effect, based on $(n_1 + n_2 - 3)$ d.f.—the extra d.f. being that for the treatments-by-classes interaction. Any difference in results should be very minor in practice.

With three classes, two dummy x variables are needed for the classification: The first can take the value 1 in class 1 and the value 0 elsewhere; the second takes the value 1 in class 2 and the value 0 elsewhere. Cohen (1968), in describing this technique, has illustrated five equivalent sets of three dummy x variables when there are four classes. Any two sets that are linear transforms of one another are equivalent.

If x is an ordered classification with c classes, a possible alternative is to assign a score $x_{k+1,i}$ to the ith class, creating a single x instead of $(c - 1)$ dummy x's to describe class effects, as suggested by Billewicz (1965). The success of this method depends, of course, on how well the assigned scores are linearly related to y.

Suppose now that there are two classified x's—one with four classes and one with three classes—creating 12 individual cells. The possibilities are to have 11 dummy x's for the effects of the 12 cells, to assume that the effects of the two classifications on y are additive, creating $3 + 2 = 5$ dummy x's for the individual effects of each classification, or, with ordered classifications, to create two sets of scores defining two x variables. Billewicz (1965) reports that the score method did well in removing between-cell bias in a constructed example in which the effects of the two x's were not strictly additive.

6.7 EFFECT OF REGRESSION ADJUSTMENTS ON BIAS IN $\bar{y}_1 - \bar{y}_2$

With y and the x's quantitative, the conditions necessary for fully effective performance of the regression adjustments in removing bias in $\bar{y}_1 - \bar{y}_2$ are (1) the regression of y on the x's is the same in both populations (apart from any difference in the level of the means due to the difference in treatments), (2) the correct mathematical form of the regression has been fitted, and (3) the x's have been measured with negligible error (see Section 6.10).

If these conditions hold, the regression adjustment removes all initial bias. Its performance in this respect is superior to matching and to adjustment by subclassification.

We now consider the failure of condition (1) for linear regressions with different slopes in the two populations. This case was discussed briefly in Section 5.7 with respect to matching, where the conclusion was reached that matching is not appropriate. Regression adjustments are capable of treating this case, but require a judgment as to whether the difference between the regressions in the two populations actually represents confounding effects in treatment. To take the simplest illustration, suppose that the model is

$$y_{1u} = \tau_1 + \beta_1 x_{1u} + e_{1u}; \quad y_{2u} = \tau_2 + \beta_2 x_{2u} + e_{2u} \qquad (6.7.1)$$

It follows that

$$E(\bar{y}_1 - \bar{y}_2) = \tau_1 - \tau_2 + \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$$

Unbiased estimates of β_1 and β_2 can be obtained and substituted to give an unbiased estimate of $\tau_1 - \tau_2$. The extension to multivariate linear regression is straightforward.

However, an alternative interpretation of (6.7.1), as mentioned previously, is that the effect of the difference in treatments depends on the level of x. Suppose that the regression of y on x is β_2 in each population. Treatment 2 is a control treatment with effect τ_2, while the effect of treatment 1 is $\tau_1 + \delta x$ when applied to a subject whose level is x. The model is then

$$y_{1u} = \tau_1 + (\delta + \beta_2)x + e_{iu}; \qquad y_{2u} = \tau_2 + \beta_2 x + e_{2u} \qquad (6.7.2)$$

Belson (1956) has suggested that in making regression adjustments, the estimate b_2 of the regression coefficient from the control sample be used. That is, his adjusted mean difference is

$$\bar{y}_{1a} - \bar{y}_{2a} = (\bar{y}_1 - \bar{y}_2) - b_2(\bar{x}_1 - \bar{x}_2)$$

From (6.7.1) we find

$$E(\bar{y}_{1a} - \bar{y}_{2a}) = \tau_1 - \tau_2 + \delta\bar{x}_1 + \beta_2(\bar{x}_1 - \bar{x}_2) - \beta_2(\bar{x}_1 - \bar{x}_2)$$
$$= \tau_1 - \tau_2 + \delta\bar{x}_1$$

Thus Belson's method estimates the average effect of treatment 1 (as compared with the control) on the persons in sample 1. In some studies this is a quantity of interest to report [see Cochran (1969)]

6.8 EFFECT OF CURVATURE ON LINEAR-REGRESSION ADJUSTMENTS

The effect of linear-regression adjustment on the bias in $\bar{y}_1 - \bar{y}_2$ when the relationship between y and x is monotone and moderately curved has been investigated by Rubin (1973). He dealt with the functions $y = e^{\pm x/2}$ and $y = e^{\pm x}$, with y monotone and quadratic in x, assuming the same form of regression in both populations.

In such cases linear-regression adjustments are still highly effective in removing an initial bias in $\bar{y}_1 - \bar{y}_2$, provided that x has the same variance in the two populations and that the distribution of x is symmetric or nearly symmetric. However, as with matching, the condition $\sigma_{1x}^2 = \sigma_{2x}^2$ is important.

The situation when y is quadratic provides insight on these results. Assume the model

$$y_{tu} = \tau_t + c_1 x_{tu} + c_2 x_{tu}^2 + e_{tu} \qquad (6.8.1)$$

where $t = 1, 2$ and $u = 1, 2, \ldots, n$. Following Rubin we consider the bias in $\bar{y}_1 - \bar{y}_2$ conditional on the set of x's that arose in the two samples. For random samples the initial conditional bias is

$$E_c(\bar{y}_1 - \bar{y}_2) - (\tau_1 - \tau_2) = c_1(\bar{x}_1 - \bar{x}_2) + \frac{c_2\Sigma(x_{1u}^2 - x_{2u}^2)}{n}$$

It is convenient to use the notation $s_t^2 = \Sigma(x_{tu} - \bar{x}_t)^2/n$, and $k_{3t} = \Sigma(x_{tu} - \bar{x}_t)^3/n$. Then

$$E_c(\bar{y}_1 - \bar{y}_2) - (\tau_1 - \tau_2) = c_1(\bar{x}_1 - \bar{x}_2) + c_2(\bar{x}_1^2 - \bar{x}_2^2) + c_2(s_1^2 - s_2^2) \qquad (6.8.2)$$

For the regression adjustment we assume that the pooled within-samples regression coefficient is used, that is,

$$b_p = \frac{\Sigma y_{1u}(x_{1u} - \bar{x}_1) + \Sigma y_{2u}(x_{2u} - \bar{x}_2)}{\Sigma(x_{1u} - \bar{x}_1)^2 + \Sigma(x_{2u} - \bar{x}_2)^2}$$

By substitution for y from the model (6.8.1), b_p is found in large samples to be a consistent estimate of

$$c_1 + 2c_2\left(\frac{\bar{x}_1 s_1^2 + \bar{x}_2 s_2^2}{s_1^2 + s_2^2}\right) + c_2\left(\frac{k_{31} + k_{32}}{s_1^2 + s_2^2}\right)$$

Consequently, the remaining conditional bias in $\bar{y}_1 - \bar{y}_2$ after adjustment by $-b_p(\bar{x}_1 - \bar{x}_2)$ approximates

$$-c_2(\bar{x}_1 - \bar{x}_2)^2\left(\frac{s_1^2 - s_2^2}{s_1^2 + s_2^2}\right) + c_2(s_1^2 - s_2^2) - c_2(\bar{x}_1 - \bar{x}_2)\left(\frac{k_{31} + k_{32}}{s_1^2 + s_2^2}\right) \tag{6.8.3}$$

Suppose now that x has differing means but the same variance in the two populations. From (6.8.2) the initial bias approximates

$$c_1(\bar{x}_1 - \bar{x}_2) + c_2(\bar{x}_1^2 - \bar{x}_2^2)$$

From (6.8.3) the bias after adjustment approximates

$$-\frac{c_2(\bar{x}_1 - \bar{x}_2)(k_{31} + k_{32})}{(s_1^2 + s_2^2)}$$

which is small or negligible if the distribution of x is symmetric or nearly symmetric. Thus linear adjustments are highly effective in this case.

In more-general cases with a quadratic regression, comparison of (6.8.2) and (6.8.3) indicates that the average relative sizes of the initial and final biases depend on the sizes and signs of the linear- and quadratic-regression coefficients, c_1 and c_2, on the sizes and signs of $\mu_{1x} - \mu_{2x}$ and $\sigma_{1x}^2 - \sigma_{2x}^2$, and on the amounts of skewness. No simple overall summary statement is possible. With $e^{\pm x/2}$ Rubin found that linear-regression adjustment was successful if $\sigma_{1x}^2 = \sigma_{2x}^2$, but that the adjustment either overcorrected or undercorrected when these variances were unequal, as shown in the first line of Table 6.8.1.

Table 6.8.1. Percent Bias Removed by (1) Linear-Regression (LR) Adjustment on Random Samples, (2) "Nearest Available" Matching, and (3) Linear-Regression Adjustment on Matched Samples [Both Samples of Size 50; $B = \mu_{1x} - \mu_{2x} = \frac{1}{2}\sigma_x$, where $\sigma_x^2 = \frac{1}{2}(\sigma_{1x}^2 + \sigma_{2x}^2)$]

Percent Bias Removed by (1), (2), and (3)	$\sigma_{1x}^2 = \sigma_{2x}^2 = 1$		$\sigma_{1x}^2 = \frac{2}{3}, \sigma_{2x}^2 = \frac{4}{3}$		$\sigma_{1x}^2 = \frac{4}{3}, \sigma_{2x}^2 = \frac{2}{3}$	
	$e^{x/2}$	$e^{-x/2}$	$e^{x/2}$	$e^{-x/2}$	$e^{x/2}$	$e^{-x/2}$
(1) LR[a]	101	101	146	80	80	146
(2) Matching[b]:						
$N/n = 2$	74	94	96	99	45	81
$N/n = 3$	87	98	98	100	60	89
$N/n = 4$	92	99	99	100	65	94
(3) Both[b]:						
$N/n = 2$	102	100	101	100	100	111
$N/n = 3$	100	100	100	100	100	108
$N/n = 4$	100	100	100	100	100	107

[a] With LR on random samples, results are for the pooled within-sample regression; on matched samples, results are for the regression from differences between members of a pair.
[b] N is the size of the reservoir supplying matches to the n members of the target sample.

Table 6.8.1, taken from Rubin (1970), compares "nearest available matching," linear regression applied to random samples, and linear regression applied to the matched samples for $E(y) = e^{\pm x/2}$. The results shown are for a bias in x equal to half the average σ_x—a fairly substantial bias. When $\sigma_{1x}^2 = \sigma_{2x}^2$, linear regression is superior. When $\sigma_{1x}^2 \neq \sigma_{2x}^2$ both methods are erratic and neither method is consistently superior. However, linear regression applied to matched samples was superior to either method and was highly effective. The regression adjustments on matched samples usually performed best when the regression coefficients were estimated from the differences between members of each matched pair. This is the method that would normally be used in matched samples from the viewpoint of analysis of variance.

6.9 EFFECTIVENESS OF REGRESSION ADJUSTMENTS ON PRECISION

As with matching, regression adjustments on random samples may be made in order to increase precision in studies in which the investigator is not concerned with the danger of bias. We assume first a linear regression of y

on a single x, the same in both populations. We suppose that there is an initial bias $B\sigma_x$ in x, since regression adjustments on random samples completely remove this bias. Therefore, there is interest in noting the precision of regression adjustments when $B \neq 0$ as well as when $B = 0$. Using the model, we find that

$$\bar{y}_1 - \bar{y}_2 = \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2) + (\bar{e}_1 - \bar{e}_2) \tag{6.9.1}$$

the adjusted estimate is

$$\begin{aligned}\bar{y}_{1a} - \bar{y}_{2a} = \bar{y}_1 - \bar{y}_2 - b(\bar{x}_1 - \bar{x}_2) &= (\tau_1 - \tau_2) + (\bar{e}_1 - \bar{e}_2)\\ &\quad - (b - \beta)(\bar{x}_1 - \bar{x}_2)\end{aligned} \tag{6.9.2}$$

Hence the conditional variance for two samples of size n is

$$V_c(\bar{y}_{1a} - \bar{y}_{2a}) = \frac{2}{n}\sigma_e^2 + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\Sigma_{xx}}\sigma_e^2 \tag{6.9.3}$$

where Σ_{xx} is the denominator of b, the pooled within-samples $\Sigma(x - \bar{x})^2$. The average value of (6.9.3) in random samples of size n is approximately

$$V(\bar{y}_{1a} - \bar{y}_{2a}) = \frac{2}{n}\sigma_e^2 + \left(B^2 + \frac{2}{n}\right)\frac{\sigma_e^2}{2(n-2)} \tag{6.9.4}$$

The expression is correct when x is normal.

In the "no bias" situation ($B = 0$) the leading term in (6.9.4) for large n is $V(\bar{y}_{1a} - \bar{y}_{2a}) = 2\sigma_e^2/n = 2\sigma_y^2(1 - \rho^2)/n$. For within-class matching with the model, $V(\bar{y}_1 - \bar{y}_2)$ is $2\sigma_y^2(1 - f\rho^2)/n$, as given in Section 5.8, where f is the fractional reduction in $V(\bar{x}_1 - \bar{x}_2)$ due to matching. Thus regression adjustments on large random samples give higher precision than within-class matching in the "no bias" case under a linear model. They should perform about as well as tight caliper matching and "nearest available" matching based on a large reservoir for which f is near 1.

When there is initial bias, $B \neq 0$, the leading term in (6.9.4) for large n is

$$V(\bar{y}_{1a} - \bar{y}_{2a}) \doteq 2\sigma_y^2(1 - \rho^2)(1 + \tfrac{1}{4}B^2)/n$$

In this case, regression applied to pair-matched samples would be expected to be more precise than regression on random samples, since pairing reduces $E(\bar{x}_1 - \bar{x}_2)^2$ in (6.9.3).

Under a linear model, the conclusions are little altered when regression adjustments are made on k x variables from random samples. If all x's have the same means in the two populations,

$$V(\bar{y}_{1a} - \bar{y}_{2a}) \doteq \frac{2\sigma_y^2}{n}(1 - R^2)\left(1 + \frac{k}{2n - k - 3}\right) \tag{6.9.5}$$

where R^2 is the squared multiple correlation coefficient between y and the x's. If $k/2n$ is negligible, this variance is practically $2\sigma_y^2(1 - R^2)/n$. Within-class matching with the same number of classes per variable gives $2\sigma_y^2(1 - fR^2)/n$, which is a larger value.

When there is initial bias, (6.9.5) also contains a quadratic expression in the biases B_j of the variables x_j; this term is of the same order as $2\sigma_y^2(1 - R^2)/n$. As before, regression applied to pair-matched samples should perform better than regression on random samples.

Using experimental sampling on a computer, Billewicz (1965) made comparisons of the precisions of within-class matching, regression on random samples, and regression applied to matched samples under a variety of situations. He uses two groups—treated and control. His results, reported here, concern relative precision in the "no bias" situation.

1. For a linear-regression model with y and x quantitative, regression was more precise than within-class matching with three or four groups, by amounts that agreed well with those given here.

2. Billewicz also made this comparison, with $n = 40$ in each sample, for three different nonlinear regression models, $y = 0.4x - 0.1x^2$, $y = 0.8x - 0.14x^2$, and $y = \tanh x$, with x following $N(0, 1)$ in both populations. These amounts of nonlinearity were detectable in 12.3, 20.3, and 19.8% of his samples. Despite the use of an incorrect model, *linear*-regression adjustments were superior in precision to matching with three or four classes.

3. When linear regressions have different slopes in the two populations, Billewicz indicates the importance of detecting this situation and the difficulties of interpretation to which we have referred. Matched pairs, regression analysis of random samples, and regression analysis applied to frequency matched samples were about equally effective in detecting the difference in slopes. He concludes that the average user of matched samples would be unlikely to examine his sample in this respect, since the concept of matching is directed toward finding a single overall effect of treatment.

6.10 EFFECT OF ERRORS IN THE MEASUREMENT OF X

Sometimes confounded x variables are difficult to measure and hence are measured with substantial errors. In large-sample studies a crude measuring device may be used for reasons of expense or because accurate measurement requires trained personnel who are in short supply. As noted by Lord (1960) and other investigators, regression adjustments fail to remove all the initial bias when the x's are measured with error. Their effectiveness in increasing precision is also reduced.

The symbol x denotes the fallible measurement actually made, while X denotes the correct value, and e, the error. The simplest model with two populations is

$$y_{1u} = \tau_1 + \beta X_{1u} + e_{1u}; \qquad y_{2u} = \tau_2 + \beta X_{2u} + e_{2u}$$

$$x_{1u} = X_{1u} + h_{1u}; \qquad x_{2u} = X_{2u} + h_{2u}$$

where h_{1u} and h_{2u} are the errors of measurement. The errors h are assumed independent of e, but h_{tu} and X_{tu} may be correlated.

Lindley (1947) has shown that even if h and X are independent, the regression of y on the fallible x is not linear unless the distributions of h and X belong, in a certain sense, to the same type (e.g., both χ^2 or normal). However, there is some evidence (Cochran, 1970a) that the linear component is dominating, and in this discussion, nonlinearity will be ignored. The slope β' of the linear component is

$$\beta' = \beta\left(\sigma_X^2 + \sigma_{Xh}\right)/\left(\sigma_X^2 + 2\sigma_{Xh} + \sigma_h^2\right) \qquad (6.10.1)$$

here σ_{Xh} is the population covariance of X and h. If b' is the estimated regression coefficient of y on x, then for given $\bar{x}_1 - \bar{x}_2$,

$$E_c(\bar{y}_{1a} - \bar{y}_{2a}) = E(\bar{y}_1 - \bar{y}_2) - \beta'(\bar{x}_1 - \bar{x}_2)$$

$$= \tau_1 - \tau_2 + (\beta - \beta')(\bar{x}_1 - \bar{x}_2)$$

Thus, conditionally, a fraction $(\beta - \beta')/\beta$ of the initial bias remains after adjustment.

If h and X are uncorrelated, $\beta' = \beta\sigma_X^2/\sigma_x^2 = G\beta$, where G is a quantity often called the "reliability" of the measurement. Thus $100G$ is the percentage of the initial bias that is removed and $100(1 - G)$ is the percentage remaining. One method attempts to remove all the initial bias by regression adjustments. It estimates σ_h^2 and σ_X^2 and hence G by an auxiliary study, and thus obtains a consistent estimate of β which is used instead of b' in making the regression adjustment. Lord (1960) addresses this problem.

Such errors of measurement also affect the performance of within-class matching and adjustment by weighted means when a fallible quantitative x is replaced by a classification in order to use these methods. These methods

still produce a fractional reduction of amount f in the initial bias of $\bar{x}_1 - \bar{x}_2$, as discussed in Sections 5.4 and 6.2, but because of the errors of measurement, this creates a fractional reduction of only fG in the initial bias of $\bar{X}_1 - \bar{X}_2$ and hence of $\bar{y}_1 - \bar{y}_2$ under the linear-regression model. The relative performance of regression on random samples and within-class matching is, therefore, unaffected by such errors of measurement in x.

The gain in precision due to regression adjustments in the "no bias" case is also affected by errors of measurement in x. With h and X uncorrelated, the population correlation ρ' between y and x is

$$\rho' = \frac{\sigma_{yx}}{\sigma_y\sigma_x} = \frac{\beta\left(\sigma_X^2 + \sigma_{Xh}\right)}{\sigma_y\sigma_x} = \frac{\rho\sigma_X}{\sigma_x} = \rho\sqrt{G}$$

Hence the residual variance from the regression is $\sigma_y^2(1 - G\rho^2)$, instead of $\sigma_y^2(1 - \rho^2)$. For a given reliability of measurement G, the relative loss of precision is greatest when ρ^2 is high, that is, when X is a very good predictor of y.

6.11 MATCHING AND ADJUSTMENT COMPARED: IN EXPERIMENTS

We start with $2n$ subjects, presumed drawn at random from the sampled population. In this population the linear-regression model is

$$y = \alpha + \beta(x - \mu) + e$$

where the residual e is assumed to have mean 0 and variance σ_e^2 for any fixed x. If n subjects are assigned at random to each of two treatments, T_1 and T_2,

$$\bar{y}_1 - \bar{y}_2 = \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2) + \bar{e}_1 - \bar{e}_2$$

Hence

$$E(\bar{y}_1 - \bar{y}_2) = \tau_1 - \tau_2 \qquad \text{(no bias)}$$

and

$$V(\bar{y}_1 - \bar{y}_2) = \frac{2}{n}\left(\beta^2\sigma_x^2 + \sigma_e^2\right) = \frac{2}{n}\left[\rho^2\sigma_y^2 + \left(1 - \rho^2\right)\sigma_y^2\right]$$

since $\beta\sigma_x = \rho\sigma_y$. The penalty for failure to control x is loss of precision, since $V(\bar{y}_1 - \bar{y}_2)$ is inflated by the term $2\beta^2\sigma_x^2/n$, or by a factor of $1/(1 - \rho^2)$.

Matching on x

The $2n$ values of x are ranked in decreasing order. Of the two highest x's, one x is assigned at random to T_1 and the other x to T_2, and so forth for succeeding pairs. This gives

$$(\bar{y}_1 - \bar{y}_2)_m = \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2)_m + \bar{e}_1 - \bar{e}_2.$$

The quantity $(\bar{x}_1 - \bar{x}_2)_m$ will not be exactly zero in this method of matching, but will have a variance that can be calculated from the variances and covariances of the order statistics. For n exceeding 50 it appears that this variance is negligible with x approximately normal, so

$$V(\bar{y}_1 - \bar{y}_2)_m = \frac{2}{n}\sigma_y^2(1 - \rho^2)$$

Regression Adjustment

When matching is not used, each treatment is assigned at random to n subjects. The pooled sample estimate b of β is

$$b = \frac{\Sigma_1 y(x - \bar{x}_1) + \Sigma_2 y(x - \bar{x}_2)}{\Sigma_1(x - \bar{x}_1)^2 + \Sigma_2(x - \bar{x}_2)^2} = \beta + \frac{\Sigma_1 e(x - \bar{x}_1) + \Sigma_2 e(x - \bar{x}_2)}{\Sigma_1(x - \bar{x}_1)^2 + \Sigma_2(x - \bar{x}_2)^2}$$

as is found when we substitute $y = \alpha + \beta(x - \mu_x) + e$ in the formula for b. For fixed x's the quantity $b - \beta$ is a random variable $\bar{e}_w$ in the e's with mean 0 and variance $\sigma_e^2/(\Sigma_1 + \Sigma_2)$, where $\Sigma_1 = \Sigma_1(x - \bar{x}_1)^2$, and so forth.

Hence the adjusted estimate

$$\begin{aligned}(\bar{y}_1 - \bar{y}_2) - b(\bar{x}_1 - \bar{x}_2) &= \tau_1 - \tau_2 + (\beta - b)(\bar{x}_1 - \bar{x}_2) + \bar{e}_1 - \bar{e}_2 \\ &= \tau_1 - \tau_2 + (\bar{e}_1 - \bar{e}_2) - (\bar{x}_1 - \bar{x}_2)\bar{e}_w\end{aligned}$$

The variance of the regression-adjusted estimate is, therefore, for fixed x's because $\bar{e}_w$ is uncorrelated with $\bar{e}_1$ and $\bar{e}_2$,

$$\frac{2}{n}\sigma_e^2 + \frac{(\bar{x}_1 - \bar{x}_2)^2\sigma_e^2}{\Sigma_1 + \Sigma_2} = \frac{2}{n}\sigma_y^2(1 - \rho^2)\left(1 + \frac{n(\bar{x}_1 - \bar{x}_2)^2}{2(\Sigma_1 + \Sigma_2)}\right)$$

For x normal the second term in the large parentheses may be shown to have mean $1/2(n - 2)$, which is only about 0.01 when n is 50.

Thus in experiments with $n = 50$ or more and a linear model, matching and regression adjustment are about equally effective. Their purpose is to increase the precision of the estimate of $\tau_1 - \tau_2$ and their effect is to reduce the term σ_y^2 in $V(\hat{\tau}_1 - \hat{\tau}_2)$ to $\sigma_y^2(1 - \rho^2)$.

6.12 MATCHING AND ADJUSTMENT COMPARED: IN OBSERVATIONAL STUDIES

In an observational comparison of two treatments, the investigator begins with *two* populations—one for each treatment. The investigator has chosen to study these two populations but did not create them. The investigator must suppose that, in general, the two populations will have different means (μ_{1y}, μ_{2y}) and (μ_{1x}, μ_{2x}). In their simplest form the regression models in the two populations become, for subject j in sample 1 and subject k in sample 2,

$$y_{1j} = \mu_{1y} + \tau_1 + \beta(x_{1j} - \mu_{1x}) + e_{1j} \qquad (6.12.1)$$

and

$$y_{2k} = \mu_{2y} + \tau_2 + \beta(x_{2k} - \mu_{2x}) + e_{2k} \qquad (6.12.2)$$

(For this illustration it is assumed that uncontrolled variables whose effects on y are summed in the terms e_{1j} and e_{2j} behave as random variables.)

Suppose first that random samples are drawn from the respective sampled populations making no attempt to control for x. Then

$$\bar{y}_1 = \mu_{1y} + \tau_1 + \beta(\bar{x}_1 - \mu_{1x}) + \bar{e}_1 \qquad (6.12.3)$$

and

$$\bar{y}_2 = \mu_{2y} + \tau_2 + \beta(\bar{x}_2 - \mu_{2x}) + \bar{e}_2 \qquad (6.12.4)$$

In repeated sampling, $E(\bar{x}_i) = \mu_{ix}$ and $E(\bar{e}_i) = 0$ $(i = 1, 2)$. Hence

$$E(\bar{y}_1 - \bar{y}_2) = \tau_1 - \tau_2 + (\mu_{1y} - \mu_{2y})$$

The estimate $\bar{y}_1 - \bar{y}_2$ is now biased by the amount $\mu_{1y} - \mu_{2y}$, with the bias favoring the treatment given to the population with the *higher* mean of y. As Campbell and Erlebacher (1970) and Campbell and Boruch (1975) have stressed, this bias produces an *underestimate* in the beneficial effect of a program given to a sample $\bar{y}_2$ from a disadvantaged population.

With random samples from the two populations the standard error (SE) of $\bar{y}_1 - \bar{y}_2$ is $\sqrt{\sigma_{1y}^2 + \sigma_{2y}^2}/\sqrt{n}$.

The ratio of the bias $\mu_{1y} - \mu_{2y}$ to this SE is

$$\frac{\sqrt{n}\,(\mu_{1y} - \mu_{2y})}{\sqrt{\sigma_{1y}^2 + \sigma_{2y}^2}}$$

This increases indefinitely as n grows. Tests of significance of the null hypothesis (NH) $\tau_1 - \tau_2$ are likely to reject the NH even when it is actually true, so that no clear interpretation can be given to rejection of the NH by the test. The interpretation of a nonsignificant result is also obscured by the possibility that $\tau_1 - \tau_2$ and $\mu_{1y} - \mu_{2y}$ have similar magnitudes and opposite signs. The relative sizes of the bias $\mu_{1y} - \mu_{2y}$ to the true treatment difference $\tau_1 - \tau_2$ obviously affects any conclusions drawn about the relative merits of the treatments.

Thus in observational studies, matching and regression have two objectives: to remove or reduce bias and to increase precision by reducing the SE of $\bar{y}_1 - \bar{y}_2$. Of these, it is reasonable to regard reduction of bias as the more-important objective. A highly precise estimate of the wrong quantity is of limited use.

Under this model of parallel linear regressions, the complete removal of bias by either matching or regression adjustment requires that the following condition hold:

$$\beta = (\mu_{1y} - \mu_{2y})/(\mu_{1x} - \mu_{2x}) \qquad (6.12.5)$$

This condition, in turn, is equivalent to each of the following:

1. Both populations (in the absence of treatment) have the same regression line.
2. The regression of y on x within the populations is equal to the regression between populations.

If (6.12.5) does not hold, the bias in estimating $\tau_1 - \tau_2$ is, after adjustment, equal to

$$(\mu_{1y} - \mu_{2y}) - \beta(\mu_{1x} - \mu_{2x})$$

(The appendix to this section gives the justification for these statements.)

Without evidence that the regression lines are the same in the two populations, the attitude of the investigator may have to be that matching and regression adjustment leave some residual bias. The investigator hopes that this bias is only a small fraction of the original bias $\mu_{1y} - \mu_{2y}$—sufficiently small in relation to $\tau_1 - \tau_2$ so that conclusions drawn about the treatments are little affected.

As Bartlett (1936) and Lord (1960) have stated, both matching and regression in this situation involve an element of unverifiable extrapolation. To take an extreme case, suppose $\mu_{1x} < \mu_{2x}$, the difference being so large that no member of sample 1 has a value as high as $\bar{x}_2$. We can still apply the regression adjustment. Formally, this adjusts $\bar{y}$ to its predicted value when the mean of the accompanying x's is $\bar{x}_2$, so that the adjusted $\bar{y}_1$ becomes comparable with $\bar{y}_2$. But this adjusted value is purely hypothetical when we have no y_{1j} value with an accompanying x as high as $\bar{x}_2$. In less-extreme cases the extrapolation is more moderate.

APPENDIX TO SECTION 6.12

Matching on x

We try to find matched pairs of subjects from the two populations such that $x_{1j} - x_{2j}$ is small in the jth pair ($j = 1, 2, \ldots, n$). Incidentally, if μ_{1x} and μ_{2x} differ substantially, matching is often a slow process, requiring large reservoirs of subjects. This may require, for instance, finding subjects with unusually high x's from population 1 to pair with low x's from population 2.

Successful matching will make $\bar{x}_1 - \bar{x}_2$ negligible. In this event, from (6.12.3) and (6.12.4),

$$E(\bar{y}_1 - \bar{y}_2)_m = \tau_1 - \tau_2 + \mu_{1y} - \mu_{2y} - \beta(\mu_{1x} - \mu_{2x})$$

Hence, all the bias is removed by matching if

$$\mu_{1y} - \mu_{2y} = \beta(\mu_{1x} - \mu_{2x}) \qquad (6.12.5)$$

This condition can be described in two equivalent ways:

1. From (6.12.1) the regression lines in the two populations may be written (in the absence of any treatment effect)

$$E(y_{1j}|x_{1j}) = \mu_{1y} - \beta\mu_{1x} + \beta x_{1j}$$

and

$$E(y_{2j}|x_{2j}) = \mu_{2y} - \beta\mu_{2x} + \beta x_{2j}$$

The condition $\mu_{1y} - \mu_{2y} = \beta(\mu_{1x} - \mu_{2x})$ then means that these two lines have the same intercepts and slopes, that is, they are identical.

From the results of the study we can test whether the slopes are the same. If they are $y_{1j} - y_{2j}$ should have no regression on x_{1j}. Since x_{1j} and x_{2j} often differ slightly in matched pairs, an approximation is to compute and test the regression of $y_{1j} - y_{2j}$ on $(x_{1j} + x_{2j})/2$. But given only y_{ij} (after treatment) and x_{ij}, we cannot check from the data whether the intercepts would be identical in the absence of treatment effects. If we fit separate parallel lines to the samples from the two populations, the intercepts on the fitted lines will be estimates of

$$\mu_{1y} - \beta\mu_{1x} + \tau_1; \qquad \mu_{2y} - \beta\mu_{2x} + \tau_2$$

They will thus differ by an estimate of the treatment difference $\tau_1 - \tau_2$, if condition (6.12.5) holds.

In the type of study called the *pretest–posttest study*, y is measured both before treatments are applied as well as after a period of application. With such data, coincidence of the regression lines in the absence of treatment can be tested from the pretest data.

2. The condition for the removal of bias

$$\beta = (\mu_{1y} - \mu_{2y})/(\mu_{1x} - \mu_{2x})$$

can also be described as meaning that the between-population regression of y on x must equal the within-population regression. If we were given the pairs of means μ_{iy} and μ_{ix} for a number of populations, the regression of μ_{iy} on μ_{ix} might appropriately be called the "between population" regression of y on x. With only two populations the slope of this regression is $(\mu_{1y} - \mu_{2y})/(\mu_{1x} - \mu_{2x})$.

Regression Adjustment

Here we assume random samples from the two populations, with no attempt at matching. The adjusted estimate of $\tau_1 - \tau_2$ is

$$(\bar{y}_1 - \bar{y}_2)_{\text{adj}} = \bar{y}_1 - \bar{y}_2 - b(\bar{x}_1 - \bar{x}_2)$$

With random samples and the linear model, $E(b) = \beta$ for any set of x. Further, $E(\bar{x}_1) = \mu_{1x}$ and $E(\bar{x}_2) = \mu_{2x}$. Hence

$$E(\bar{y}_1 - \bar{y}_2)_{\text{adj}} = \tau_1 - \tau_2 + \mu_{1y} - \mu_{2y} - \beta(\mu_{1x} - \mu_{2x})$$

The expression $\mu_{1y} - \mu_{2y} - \beta(\mu_{1x} - \mu_{2x})$ is the bias that remains after regression adjustment; it vanishes when condition (6.12.5) holds. Thus under the linear model both the residual bias and the condition for its complete removal are the same for adjustment by linear regression as for matching.

6.13 A PRELIMINARY TEST OF COMPARABILITY

In deciding whether to match or adjust for an x variable, it has been recommended that consideration be given first to x's in which it is suspected that there will be a bias arising from a difference $\mu_{1x} - \mu_{2x}$ in the means of x. If uncertain whether there is a danger of bias, we might first make a t test of significance of $\bar{x}_1 - \bar{x}_2$ from two random samples of size n. If t is nonsignificant, we judge that the risk of major bias is small and decide not to match or adjust for this x. Tests of significance are often employed as decision rules in this way.

This procedure has been examined [Cochran (1970b)] assuming a linear regression of y on x. If t is significant at some chosen level, a linear-regression adjustment on random samples is made. If t is not significant, the unadjusted estimate $\bar{y}_1 - \bar{y}_2$ is used.

Under the standard linear-regression model the conditional mean of the adjusted y difference given $\bar{x}_1$ and $\bar{x}_2$ is

$$E_c(\bar{y}_{1a} - \bar{y}_{2a}) = \tau_1 - \tau_2 - (\bar{x}_1 - \bar{x}_2)E_c(b - \beta)$$

Now $b - \beta = \Sigma e(x - \bar{x})/\Sigma(x - \bar{x})^2$, where the Σ's are the pooled within-sample sums of squares or products. Consider samples selected so that $t = \sqrt{n}\,|\bar{x}_1 - \bar{x}_2|/\sqrt{2}\,s_x$ is significant. Since this selection is based solely on the values of x and since e and x are independent, $E_c(b - \beta) = 0$ in samples selected in this way. (The conditional variance of b is affected, but not the conditional mean). It follows that the adjusted $\bar{y}_{1a} - \bar{y}_{2a}$ is free from bias when t is significant.

The remaining bias from this process is, therefore,

$$P(|t| < t_0)E_c(\bar{y}_1 - \bar{y}_2) = P(|t| < t_0)\beta E_c(\bar{x}_1 - \bar{x}_2)$$

where t_0 is the critical value of t and the conditional mean is for $|t| < t_0$. The intuitive idea behind the method is, of course, that if $\mu_{1x} - \mu_{2x}$ is large there should be little remaining bias because t is almost certain to be significant. If $\mu_{1x} - \mu_{2x}$ is small, t may be nonsignificant frequently, but the final bias should be small because the initial bias is small.

At some intermediate point we obtain the maximum final bias. Since $\beta(\bar{x}_1 - \bar{x}_2) = \beta\sqrt{2}\,s_x t/\sqrt{n}$ the final bias is of order $1/\sqrt{n}$.

The maximum final bias occurs when the probability of a nonsignificant t is around 0.70 for 5% tests, 0.65 for 10% tests, and 0.60 for 20% tests. Expressed for convenience as a fraction f of the quantity $\sqrt{2}\,\beta\sigma_x/\sqrt{n}$, the values of f vary between 0.72 (20 d.f. for t tests) and 0.66 (∞ d.f.) for 5% tests, 0.49 and 0.45 for 10% tests, and 0.26 and 0.25 for 20% tests. As expected, a larger, that is, less stringent, significance level of t gives a smaller final bias at the expense of more-frequent adjustments.

Is the procedure adequate? Suppose an investigator uses standard elementary formulas for tests of significance of $\bar{y}_1 - \bar{y}_2$ or confidence levels of $\mu_{1y} - \mu_{2y}$ after using this test. That is, the investigator assigns to $\bar{y}_1 - \bar{y}_2$ a standard error $\sqrt{2}\,s_y/\sqrt{n}$, if t is nonsignificant, and to $(\bar{y}_{1a} - \bar{y}_{2a})$ a standard error

$$\sqrt{2}\,s_{y\cdot x}\frac{\left[1 + (\bar{x}_1 - \bar{x}_2)^2/\Sigma\right]^{1/2}}{\sqrt{n}}$$

if t is significant. (In the preceding expression, $s_{y\cdot x}$ is the root of the residual variance about the regression line based on pooled within-group sums of squares and cross-products. Also Σ is the pooled within-group sum of squares of x.) Even with 5% tests, it is found that type-I errors and confidence probabilities are only slightly disturbed.

Alternatively, we might ask whether the maximum remaining bias is negligible with respect to the size of difference δ_y that we are trying to measure; or in other words, whether the ratio $\sqrt{2}\,f\rho\sigma_y/\sqrt{n}\,\delta_y$ is negligible. The answer here is less certain, since it depends on n, ρ, and the ratio δ_y/σ_y. For instance, in some applications an improvement of a new method of treatment over a standard method might be important in practice if $\delta_y/\sigma_y = 0.2$. Taking the maximum f as about 0.7 for 5% tests and $\rho = 0.4$, the ratio of the maximum bias to this δ_y is $\sqrt{2}\,(0.7)(0.4)/0.2\sqrt{n} = 1.98/\sqrt{n}$. If we want the ratio to be less than 10%, we need $n = (19.8)^2 = 392$ in each sample. Unless we have samples at least this large, the ratio will not be negligible (less than 10%).

6.14 SUMMARY

In comparing the means $\bar{y}_1 - \bar{y}_2$ or proportions $\hat{p}_1 - \hat{p}_2$ from two populations, an alternative to matching is to draw random samples from the two populations and make adjustments in the statistical analysis to $\bar{y}_1 - \bar{y}_2$ or $\hat{p}_1 - \hat{p}_2$ in order to reduce bias or increase precision. The method of adjustment depends on the scales in which the variables are measured.

If the y's are quantitative and the x's are classified (or have been made classified), let $\bar{y}_{1i}$ and $\bar{y}_{2i}$ be the sample means of y in the ith cell of this classification. If the effect $\tau_1 - \tau_2$ of the difference in treatments is the same in every cell, any weighted mean $\Sigma w_i(\bar{y}_{1i} - \bar{y}_{2i}) = \Sigma w_i d_i$, with $\Sigma w_i = 1$, controls bias to precisely the same extent as does within-class matching, so that there is little difference between the methods in this respect. The choices of weights that minimize the variance of $\Sigma w_i d_i$ are given in Section 6.2. In particular, optimum weights are proportional to $n_{1i}n_{2i}/(n_{1i} + n_{2i})$ if the within-cell and treatment variances are constant. Under this assumption, two matched samples of size n give a smaller variance than two weighted random samples of size n, but the difference is likely to be minor in the "no bias" case in which the comparision is of most interest.

If y is a (0, 1) variate, many workers have assumed a model in which the effect of the difference between treatments is constant from cell to cell on the scale of logit $p_{ti} = \log(p_{ti}/q_{ti})$. Under this model it may still be desirable to estimate and test a weighted mean difference of the form $\Sigma w_i(\hat{p}_{1i} - \hat{p}_{2i})/\Sigma w_i$. For this purpose, a good choice for testing significance is $w_i = n_{1i}n_{2i}/(n_{1i} + n_{2i})$.

If the treatment effect $\delta_i = \tau_{1i} - \tau_{2i}$ differs from cell to cell, the choice of weights determines the quantity $\Sigma w_i \delta_i$ that is being estimated. In the analysis, possibilities are (1) to use weights derived from a target population that is of interest; (2) to note that the values of $\Sigma w_i \delta_i$ agree well enough for different weighting systems so that the same conclusion or action is suggested; and (3) to decide against estimation of an overall mean and to summarize instead the way in which δ_i varies from cell to cell. A method of testing whether δ_i varies from cell to cell is given, but the interpretation of the test is simple only when a single x variable is involved.

When y and the x's are all quantitative, adjustments for bias may be made on random samples by means of the regression of y on the x's. This method can also include a classified x by the creation of dummy variables to represent class effects or (with ordered classifications) by assigning scores to the classes. In practice, a linear regression with the same slopes in both populations is most commonly assumed, but the method provides tests for differences in slopes and for nonlinearity which help to make the assumed model more nearly correct. Linear-regression adjustments can be used when there are differences in slopes, if these differences are due to the confounding x variables. However, another possible interpretation is that the differences may represent a relation between the effects of the treatments and the level of x.

With regard to the control of bias, the regression method removes all the initial bias, provided that the fitted model is correct in form and the x's are not subject to errors of measurement. In this situation, regression is superior to pair or within-class matching and to adjustment by weighted class means. If adjustment by linear regression is used when the true regression of y on x is monotone and moderately curved [e.g., a quadratic or $E(y) = e^{\pm x/2}$], the available evidence suggests that linear adjustment still removes almost all the bias, provided that $\sigma_{1x} = \sigma_{2x}$ and that the distribution of x is symmetrical. If $\sigma_{1x} \neq \sigma_{2x}$ the performance of linear-regression adjustments on $e^{\pm x/2}$ is erratic. However, linear-regression adjustments on matched samples were highly successful in this situation.

With regard to the precision of $\bar{y}_1 - \bar{y}_2$ in the "no-bias" situation, linear-regression adjustments on random samples were superior under a linear-regression model to within-class matching and almost as good as mean matching and tight caliper matching. With three monotone nonlinear population regressions (two quadratic and one $y = \tanh x$) Billewicz found linear-regression adjustments superior in precision to within-class matching with three or four classes.

By way of an overall comparison, the comparisons made indicate that, with y and the x's quantitative, regression adjustments based on random samples should be superior to within-class matching and probably also superior to a fairly tight caliper matching and "nearest available" pair matching based on a large reservoir, provided that care is taken to fit approximately the correct shape of regression. Even if linear adjustments are routinely applied, they appear to perform about as well as "nearest available" pair matching in the presence of monotone curved regressions. In such cases, however, linear-regression adjustments applied to pair-matched samples are consistently better in removing bias.

If the true regression of y on X is linear, but the measured x is subject to an independent error of measurement, the percentage of bias removed by regression adjustment is reduced, dropping to $100G$, where $G = \sigma_X^2/\sigma_x^2$. The performance of within-class matching is affected similarly.

Finally, one possibility is to adjust for the regression on x as a precaution against bias only if $\bar{x}_1 - \bar{x}_2$ is statistically significant. Under a linear model, this decision rule operates well enough so that type-I errors and confidence probabilities relating to $\bar{y}_1 - \bar{y}_2$ calculated by standard techniques are not much affected.

REFERENCES

Bartlett, M. S. (1936). A note on the analysis of covariance. *J. Agric. Sci.*, **26**, 488–491.

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Appl. Statist.*, **5**, 195–202.

Billewicz, W. Z. (1965). The efficiency of matched samples: an empirical investigation. *Biometrics*, **21**, 623–644.

Campbell, D. T. and R. F. Boruch (1975). Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six Ways in which Quasi-Experimental Evaluations in Compensatory Education Tend to Underestimate Effects, in C. A. Bennett and A. A. Lumsdaine, Eds. *Evaluation and Experience: Some Critical Issues in Assessing Social Programs*. Academic Press, New York.

Campbell, D. T. and A. E. Erlebacher (1970). How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful, in J. Hellmuth, Ed. *Compensatory Education: A National Debate, 3, Disadvantaged Child*. Brunner/Mazel, New York.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451 [Collected Works #59].

Cochran, W. G. (1969). The use of covariance in observational studies. *Appl. Statist.*, **18**, 270–275 [Collected Works #92].

Cochran, W. G. (1970a). Some effects of errors of measurement on linear regression. *Proceedings of the 6th Berkeley Symposium*, Vol. 1, University of California Press, pp. 527–539 [Collected Works #95].

Cochran, W. G. (1970b). *Performance of a Preliminary Test of Comparability in Observational Studies* ONR Technical Report No. 29, Department of Statistics, Harvard University, Cambridge, Mass. [Collected Works #96].

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychol. Bull.*, **70**, 426–443.

Keyfitz, N. (1966). Sampling variance of standardized mortality rates. *Human Biology*, **38**, 309–317.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *J. Roy. Statist. Soc. B*, **9**, 218–244.

Lord, F. (1960). Large-sample covariance analysis when the control variable is fallible. *J. Am. Statist. Assoc.*, **55**, 307–321.

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.

Meier, P. (1953). Variance of a weighted mean. *Biometrics*, **9**, 59–73.

Rubin, D. B. (1970). The Use of Matched Sampling and Regression Adjustment in Observational Studies. *Ph.D. Thesis, Harvard University, Cambridge, Mass.*

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.