On regression adjustments to experimental data

David A Freedman

Statistics Department, University of California, Berkeley CA 94720-3860

Abstract

Regression adjustments are often made to experimental data. Since randomization does not justify the models, almost anything can happen. Here, we evaluate results using Neyman's non-parametric model, where each subject has two potential responses, one if treated and the other if untreated. Only one of the two responses is observed. Regression estimates are generally biased, but the bias is small with large samples. Adjustment may improve precision, or make precision worse; standard errors computed according to usual procedures may overstate the precision, or understate, by quite large factors. Asymptotic expansions make these ideas more precise.

Keywords and phrases: Models, randomization, multiple regression, balance, intention-to-treat

AMS classification numbers: 62A01, 62J99

Advances in Applied Mathematics vol. 40 (2008) pp. 180–93.

1. Introduction

Experimental data are often analyzed using regression models. In this paper, we examine the behavior of regression estimates in Neyman's model [5, 18], where each subject has two potential responses. One is observed if the subject is assigned to treatment, the other is observed if the subject is assigned to control. The "intention-to-treat" parameter, b_{ITT}, is the average response if all subjects are assigned to treatment, minus the average response if all subjects are assigned to control. In the design we consider, m out n subjects are chosen at random for treatment, and the remaining $n-m$ are assigned to control. (This excludes stratified designs, blocking, and so forth.)

In brief, let Y be the observed response. Let X be the assignment variable, taking the value 1 if the subject is assigned to treatment and 0 otherwise. We compare three estimators of b_{ITT}. The intention-to-treat estimator, $\hat{b}_{\text{ITT}}$, is the difference between the average response in the treatment group and the control group. The simple regression estimator, $\hat{b}_{\text{SR}}$, is obtained by running a regression of the observed response Y on the assignment variable X; there is an intercept in

the equation. For the third estimator, let Z be a covariate which is not affected by treatment: for example, Z_i might be a characteristic of subject i measured before assignment to treatment or control. The multiple regression estimator, $\hat{b}_{\mathrm{MR}}$, is obtained by running a regression of Y on X and Z; again, there is an intercept.

As is well known, the intention-to-treat estimator is exactly unbiased; furthermore, the simple regression estimator coincides with the ITT estimator. The following results, however, may be somewhat surprising.

(i) The multiple regression estimator is biased; the bias tends to 0 as the number of subjects increases.
(ii) Asymptotically, the multiple regression estimator may perform worse than the simple regression estimator.
(iii) "Nominal" standard errors (computed from the usual formulas) can be severely biased.
(iv) The nominal standard error for the simple regression estimator may differ from the nominal standard error for the intention-to-treat estimator—even though the two estimators coincide.

The reason for the breakdown is not hard to find: randomization does not not justify the assumptions behind the OLS model. Indeed, the assignment variable (to treatment or control) and the error term in the model will generally be strongly related. This will be detailed below, along with some asymptotic expansions that provide analytical proofs for the results listed above.

2. Simple regression

Index the subjects by $i = 1, \ldots, n$. Let T_i be the response of subject i if i is assigned to treatment, and let C_i be the response of subject i if i is assigned to control. For now, these are fixed numbers. (The extension to random responses is easy, and will not be considered here.) The investigator can choose to observe either T_i or C_i, but the two responses cannot be observed simultaneously. Let X_i be the assignment variable: $X_i = 1$ if subject i is assigned to treatment, and $X_i = 0$ if subject i is assigned to control. The observed response is

$$Y_i = X_i T_i + (1 - X_i) C_i. \tag{1}$$

If i is assigned to treatment, then $X_i = 1$ and $Y_i = T_i$: it is the response to treatment that is observed. If i is assigned to control then $X_i = 0$ and $Y_i = C_i$: the response to the control condition is observed. The ITT estimator is

$$\hat{b}_{\mathrm{ITT}} = \left(\frac{1}{m} \sum_i \{Y_i : X_i = 1\} \right) - \left(\frac{1}{n-m} \sum_i \{Y_i : X_i = 0\} \right), \tag{2}$$

with n being the number of subjects and $m = \sum X_i$ the size of the treatment group. The simple regression estimator $\hat{b}_{\mathrm{SR}}$ is the coefficient of X in a regression of Y on

1 and X. The following well known theorem is arithmetic in nature. There are no conditions on the joint distribution of the X_i: what matters is that $X_i = 0$ or 1.

Theorem 1. *If* $0 < m < n$, *then* $\hat{b}_{\text{SR}} = \hat{b}_{\text{ITT}}$.

Proof. Write "ave" for the average across all subjects, and let Σ run over all subjects too. Let $p = m/n$. Now

$$\begin{aligned}
\hat{b}_{\text{ITT}} &= \frac{\Sigma X_i Y_i}{\Sigma X_i} - \frac{\Sigma(1 - X_i)Y_i}{\Sigma(1 - X_i)} \\
&= \frac{\text{ave}(XY)}{\text{ave}(X)} - \frac{\text{ave}(Y) - \text{ave}(XY)}{1 - \text{ave}(X)} \\
&= \frac{\text{ave}(XY) - \text{ave}(X)\text{ave}(Y)}{p(1 - p)} \\
&= \frac{\text{cov}(X, Y)}{\text{var}(X)} \\
&= \hat{b}_{\text{SR}}.
\end{aligned}$$

Here,

$$\begin{aligned}
\text{cov}(X, Y) &= \text{ave}(XY) - \text{ave}(X)\text{ave}(Y). \\
\text{var}(X) &= \text{ave}(X^2) - [\text{ave}(X)]^2.
\end{aligned}$$

Finally

$$\text{ave}(X) = p, \quad \text{var}(X) = p(1 - p),$$

because m of the X_i are equal to 1, and $m/n = p$. QED

Discussion

(i) The "nominal variance" for the simple regression estimator is obtained by the usual computation, as the (2,2) element of $\hat{\sigma}^2(M'M)^{-1}$ where $\hat{\sigma}^2$ is the mean square of the residuals and M is the design matrix, which will be defined more carefully below. The nominal variance for the ITT estimator is $\hat{v}_T/m + \hat{v}_C/(n - m)$, where $\hat{v}_T$ is the sample variance in the treatment group and $\hat{v}_C$ is the sample variance in the control group. Although $\hat{b}_{\text{SR}} = \hat{b}_{\text{ITT}}$, the two variances may be quite different: the regression formulas assume homoscedasticity, whereas the ITT formulas adjust for heteroscedasticity.

(ii) Even if $Y_i = 0$ or 1, Theorem 1—like the other theorems below—covers OLS; logits and probits would require a separate discussion.

3. The statistical model

As before, each subject has two potential responses T_i and C_i, and X_i is the assignment variable. The observed response is $Y_i = X_i T_i + (1 - X_i)C_i$. The T_i and C_i are fixed, subject-level parameters. Population-level parameters are defined

as follows:

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n} T_i, \quad \overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i, \quad b = \overline{T} - \overline{C}. \tag{3}$$

The parameter b is the *intention-to-treat* parameter, also called the *average treatment effect*, or the *average causal effect*. See, for instance, Holland [14]. What b represents is a differential: the effect of assigning everybody to treatment, minus the effect of assigning them to control. This parameter is the one of interest here.

We assume that m out n subjects are assigned at random to treatment, the remaining $n-m$ subjects being assigned to control. Under this assumption, Theorem 2 is a well-known result: the intention-to-treat estimator is unbiased.

Theorem 2. $E\big(\hat{b}_{\text{ITT}}\big) = b$.

The proof is omitted, as it boils down to an even better known fact: with simple random samples, the sample average is an unbiased estimator for the population average. To investigate the regression estimator, it will convenient to rewrite (1) as follows:

$$Y_i = a + b(X_i - p) + \delta_i, \tag{4}$$

where

$$a = p\overline{T} + (1-p)\overline{C}, \quad b = \overline{T} - \overline{C}, \tag{5a}$$

$$\alpha_i = p(T_i - \overline{T}) + (1-p)(C_i - \overline{C}), \quad \beta_i = (T_i - \overline{T}) - (C_i - \overline{C}), \tag{5b}$$

$$\delta_i = \alpha_i + \beta_i(X_i - p). \tag{5c}$$

Centering X_i at p in (4) does not affect the estimators, and simplifies the asymptotics below. Equation (4) is nothing like a standard regression model. The randomness in δ_i is due entirely to randomness in X_i, so the error term is strongly dependent on the explanatory variable. The δ's are not IID, nor do they have mean 0. On the other hand, by (5b),

$$\sum_{i=1}^{n} \alpha_i = \sum_{i=1}^{n} \beta_i = 0. \tag{6}$$

So $E[(X_i - p)\delta_i] = p(1-p)\beta_i$ sums to 0 over all subjects. Finally,

$$E(\delta_i) = \alpha_i \tag{7}$$

sums to 0 over all subjects. These are weak forms of orthogonality.

We turn now to the assignment variables, which are a little dependent because their sum is fixed. However, they are exchangeable, and behave in other ways very much like coin-tossing, at least when n is large. For example,

$$P\{X_i = 1\} = p, \quad P\{X_i = 0\} = 1 - p, \tag{8a}$$

$$E(X_i) = p, \quad \text{var}(X_i) = p(1-p), \tag{8b}$$

$$\text{cov}(X_i, X_j) = -\frac{p(1-p)}{n-1} \text{ if } i \neq j. \tag{8c}$$

In the display, $p = m/n$, while $\text{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$ and $\text{var}(X_i) = \text{cov}(X_i, X_i)$.

The setup here applies when there is one treatment group, one control group, and subjects are chosen at random without replacement for the treatment group. More complex designs with blocking and stratification are not covered by the present theorems.

The distinction between "observables" and "unobservables" is important. Formally, our estimators are defined in terms of observable random variables like X_i, Y_i, Z_i; unobservable parameters like T_i and C_i do not enter into the formulas.

In the simple regression model (4), and the multiple regression model below, the random element is the assignment to treatment or control. Conditional on the X_i, the Y_i are fixed (not random)—and so are the "error terms" δ_i in (4): see (5c).

4. Asymptotics: simple regression

We turn now to the asymptotics of the simple regression estimator, using the notation of the previous section. In principle, our inference problem is embedded in an infinite sequence of such problems, with the number of subjects n increasing to infinity. Parameters like p, the fraction of subjects assigned to treatment, should be subscripted by n, with the assumption $p_n \to p$ and $0 < p < 1$. Instead, we say that np subjects are assigned to treatment. Similarly, parameters like α_i in (5) should be doubly subscripted, and we should assume that

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_{i,n}^2 \to \overline{\alpha^2},$$

rather than the simpler formula in (9a) below. The additional rigor is not worth the notational price. In the same spirit, our moment conditions are fairly restrictive, the object being to minimize technicalities rather than maximize generality. The

symbol $\overline{\alpha^2}$ in the display merely denotes the value of a limit; likewise for $\overline{\alpha\beta}$ and $\overline{\beta^2}$, introduced below.

With these understandings, we write the conditions as follows:

$$(9a) \qquad \frac{1}{n}\sum_{i=1}^{n}\alpha_i^2 \to \overline{\alpha^2}, \quad \frac{1}{n}\sum_{i=1}^{n}\alpha_i\beta_i \to \overline{\alpha\beta}, \quad \frac{1}{n}\sum_{i=1}^{n}\beta_i^2 \to \overline{\beta^2},$$

where $\overline{\alpha^2}$, $\overline{\alpha\beta}$, and $\overline{\beta^2}$ are fixed real numbers. Plainly, $\overline{\alpha^2} \geq 0$, $\overline{\beta^2} \geq 0$. We also require bounded fourth moments:

$$(9b) \qquad \frac{1}{n}\sum_{i=1}^{n}\alpha_i^4 < K < \infty, \quad \frac{1}{n}\sum_{i=1}^{n}\beta_i^4 < K < \infty,$$

and

$$(9c) \qquad 0 < p < 1.$$

Condition (9) may seem unfamiliar, but similar conditions are used to derive consistency and asymptotic normality for OLS estimators. See Drygas [7], Anderson and Taylor [1], Freedman [8], or pp. 66ff in Greene [12]. Let

$$(9d) \quad \gamma = \lim \frac{1}{n}\sum_{i=1}^{n}[\alpha_i + (1-2p)\beta_i]^2 = \overline{\alpha^2} + 2(1-2p)\overline{\alpha\beta} + (1-2p)^2\overline{\beta^2} \geq 0.$$

Theorem 3. *Under condition* (9)*, the simple regression estimator is asymptotically normal with mean b and variance* $\gamma/[np(1-p)]$*, i.e., the distribution of* $\sqrt{n}(\hat{b}_{\mathrm{SR}} - b)$ *converges to* $N\big(0, \gamma/[p(1-p)]\big)$.

Proof. The design matrix is

$$M = \begin{pmatrix} 1 & X_1 - p \\ 1 & X_2 - p \\ \vdots & \vdots \\ 1 & X_n - p \end{pmatrix}.$$

(Centering X_i at p does not change $\hat{b}_{\mathrm{SR}}$ and does simplify the calculation.) Since $X_1 + \cdots + X_n$ is fixed at np by construction,

$$M'M/n = \begin{pmatrix} 1 & 0 \\ 0 & p(1-p) \end{pmatrix}, \quad (M'M/n)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/[p(1-p)] \end{pmatrix}.$$

If Y is the column vector of responses Y_i, then

$$M'Y/n = \begin{pmatrix} \Sigma_i Y_i/n \\ \Sigma_i (X_i - p)Y_i/n \end{pmatrix}$$

and

$$p(1-p)\hat{b}_{\mathrm{SR}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)Y_i. \tag{10}$$

Recall from (4) that $Y_i = a + b(X_i - p) + \delta_i$. Furthermore, $\Sigma_i(X_i - p) = 0$ while $\Sigma_i(X_i - p)^2 = np(1-p)$, because np subjects have $X_i = 1$ and the rest have $X_i = 0$. Substitution into (10) gives

$$p(1-p)(\hat{b}_{\mathrm{SR}} - b) = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)\delta_i. \tag{11}$$

Let U be the quantity on the right hand side of (11). By (5c),

$$U = \frac{1}{n}\sum_{i=1}^{n}\left[\alpha_i(X_i - p) + \beta_i(X_i - p)^2\right]. \tag{12}$$

We have now arrived at the crux of the proof, and must show that U is asymptotically normal, with mean 0 and asymptotic variance $\gamma p(1-p)/n$. To begin with, $X_i^2 = X_i$, so

$$\alpha_i(X_i - p) + \beta_i(X_i - p)^2 = \left[\alpha_i + (1-2p)\beta_i\right]X_i - \alpha_i p + \beta_i p^2. \tag{13}$$

Next, $\Sigma_i\,\alpha_i = \Sigma_i\,\beta_i = 0$, by (6). The last two terms on the right in (13) can therefore be dropped, i.e.,

$$U = \frac{1}{n}\sum_{i=1}^{n}\left[\alpha_i + (1-2p)\beta_i\right]X_i. \tag{14}$$

Visualize U in (14) as $1/n$ times the sum of np draws made at random without replacement from a box of n tickets, the ith ticket being marked with the number $\alpha_i + (1-2p)\beta_i$. The average of the tickets is 0, again by (6). Now U has mean 0 and variance

$$\frac{1}{n^2} \times np \times \frac{n(1-p)}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}[\alpha_i + (1-2p)\beta_i]^2 \approx \frac{\gamma p(1-p)}{n}, \tag{15}$$

where $c_n \approx d_n$ means that $c_n/d_n \to 1$. The third factor on the left side of (15) is the "finite sample correction factor." Asymptotic normality follows, e.g., from Högland [13], who gives a Berry-Esseen type of bound. Look back at (11): divide the right side of (15) by $[p(1-p)]^2$ to complete the proof of Theorem 3. QED

Discussion.

(i) Since $\hat{b}_{\mathrm{ITT}} = \hat{b}_{\mathrm{SR}}$, the theorem also gives the asymptotic distribution of the ITT estimator.

(ii) As (9d) shows, $\gamma \geq 0$. If $\gamma = 0$, the theorem asserts that $\sqrt{n}(\hat{b}_{\mathrm{SR}} - b) \to 0$ in probability. A little more is true. If $\alpha_i = \beta_i = 0$ for all i then $T_i = \overline{T}$ and $C_i = \overline{C}$ for all i; there is no variance in $\hat{b}_{\mathrm{SR}}$: see (5). If $\alpha_i = 0$ for all i and $p = 1/2$, then $T_i - \overline{T} = -(C_i - \overline{C})$ for all i; there is again no variance in $\hat{b}_{\mathrm{SR}}$. Either way, $\hat{b}_{\mathrm{SR}} = b$ for all assignments.

(iii) The condition $\beta \equiv 0$ will recur. The meaning is simple: $T_i = C_i + b$ for all i. In other words, for any subject, treatment adds the constant b. There is still variance in $\hat{b}_{\mathrm{ITT}}$, because C_i can vary from one subject to another, so the average of the C_i across the treatment and control groups will depend on the X_i. The deviation of C_i from the population average $\overline{C}$ is captured by α_i.

5. Asymptotics: multiple regression

Let Z_i be a covariate defined for each subject i. This Z_i is observable. Implicit in the notation is the idea that Z_i remains the same, whether $X_i = 1$ or $X_i = 0$. Without loss of generality, we may standardize:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} Z_i^2 = 1. \tag{16a}$$

In addition, we assume

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i Z_i \to \overline{\alpha Z}, \quad \frac{1}{n}\sum_{i=1}^{n} \beta_i Z_i \to \overline{\beta Z}, \quad \frac{1}{n}\sum_{i=1}^{n} Z_i^4 < K < \infty, \tag{16b}$$

where α, β were defined in (5). As before, $\overline{\alpha Z}$ and $\overline{\beta Z}$ are fixed real numbers—the limiting values in (16b). Let

$$\gamma' = \gamma - (\overline{\alpha Z})^2 - 2(1-2p)(\overline{\alpha Z})(\overline{\beta Z}), \tag{16c}$$

where γ was defined in (9d). The multiple regression estimator $\hat{b}_{\mathrm{MR}}$ is the coefficient of X in a regression of Y on 1, X, and Z; equivalently, the coefficient of $X - p$ in a regression of Y on 1, $X - p$, and Z. The latter formulation will be more convenient.

Theorem 4. *Under conditions* (9) *and* (16)*, the multiple regression estimator is asymptotically normal with mean b and variance* $\gamma'/[np(1-p)]$*, i.e., the distribution of* $\sqrt{n}(\hat{b}_{\mathrm{MR}} - b)$ *converges to* $N\big(0, \gamma'/[p(1-p)]\big)$.

Proof. The proof is like that for Theorem 3; some details are omitted. To begin with, the design matrix is

$$M = \begin{pmatrix} 1 & X_1 - p & Z_1 \\ 1 & X_2 - p & Z_2 \\ \vdots & \vdots & \\ 1 & X_n - p & Z_n \end{pmatrix}.$$

(Centering X_i at p doesn't affect $\hat{b}_{\mathrm{MR}}$ and does simplify the calculation.) Thus,

$$M'M/n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & p(1-p) & \xi \\ 0 & \xi & 1 \end{pmatrix},$$

where

$$\xi = \frac{1}{n}\sum_{i=1}^{n} Z_i(X_i - p)$$

is by previous arguments asymptotically normal with mean 0 and variance on the order of $1/n$. Now $\det(M'M/n) = p(1-p) + O(1/n)$ in probability, and

$$(M'M/n)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/[p(1-p)] & -\xi/[p(1-p)] \\ 0 & -\xi/[p(1-p)] & 1 \end{pmatrix} + O\Big(\frac{1}{n}\Big).$$

Next,

$$M'Y/n = \begin{pmatrix} \Sigma_i Y_i/n \\ \Sigma_i (X_i - p)Y_i/n \\ \Sigma_i Z_i Y_i/n \end{pmatrix}.$$

In consequence,

$$p(1-p)\hat{b}_{\mathrm{MR}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - p)Y_i - \xi\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i + O\Big(\frac{1}{n}\Big). \tag{17}$$

Substitute (4) into (17) and use the argument in (11–14):

$$p(1-p)(\hat{b}_{\mathrm{MR}} - b) = U - \xi V + O\Big(\frac{1}{n}\Big), \tag{18}$$

where

$$U = \frac{1}{n}\sum_{i=1}^{n}[\alpha_i + (1-2p)\beta_i]X_i$$

and

$$V = \frac{1}{n}\sum_{i=1}^{n} Z_i[a + b(X_i - p) + \delta_i].$$

The a-term in V vanishes because $\Sigma_i Z_i = 0$. The b-term in V is $b\xi$, which contributes $b\xi^2 = O(1/n)$ to ξV in (18). Thus, we may improve (18) as follows:

$$p(1-p)(\hat{b}_{\mathrm{MR}} - b) = U - \xi V' + O\Big(\frac{1}{n}\Big), \text{ where } V' = \frac{1}{n}\sum_{i=1}^{n} Z_i\delta_i. \tag{19}$$

Substitute $\delta_i = \alpha_i + \beta_i(X_i - p)$ from (5c) into the formula for V'. The α-term in δ contributes $\xi\theta$ to $\xi V'$, where

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\alpha_i Z_i. \tag{20}$$

The β-term contributes $\xi\zeta$, where

$$\zeta = \frac{1}{n}\sum_{i=1}^{n} \beta_i Z_i (X_i - p). \tag{21}$$

But $\xi\zeta = O(1/n)$. Indeed, ξ is asymptotically normal with mean 0 and variance on the order of $1/n$. The same is true of ζ. In more detail, the argument for asymptotic normality of U in the previous section can be adapted to cover ζ in (21): center $\beta_i Z_i$ and drop p; or, compute the mean and variance of ξ, ζ and use Chebychev's inequality. This completes our discussion of $\xi\zeta$. In sum, $\xi V' = \xi\theta + O(1/n)$.

Recall that $\xi = \frac{1}{n}\sum_i Z_i(X_i - p) = \frac{1}{n}\sum_i Z_i X_i$ because $\sum_i Z_i = 0$. On this basis, (19) shows that

$$\begin{aligned} p(1-p)(\hat{b}_{\mathrm{MR}} - b) &= \left(\frac{1}{n}\sum_{i=1}^{n}\big[\alpha_i + (1-2p)\beta_i\big]X_i\right) - \xi\theta + O\Big(\frac{1}{n}\Big) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\big[\alpha_i - \theta Z_i + (1-2p)\beta_i\big]X_i\right) + O\Big(\frac{1}{n}\Big). \end{aligned}$$

Conditions (9) and (16) entail $\theta \to \overline{\alpha Z}$, and then

$$\lim \frac{1}{n}\sum_{i=1}^{n}\big[\alpha_i - \theta Z_i + (1-2p)\beta_i\big]^2 = \gamma', \tag{22}$$

where γ' is defined by (16c) and (9d). The rest of the argument is the same as for Theorem 3. QED

Discussion

(i) By construction, $\frac{1}{n}\sum_1^n \alpha_i = \frac{1}{n}\sum_1^n Z_i = 0$ and $\frac{1}{n}\sum_1^n Z_i^2 = 1$, so θZ is the regression of α on Z, and $\alpha - \theta Z$ is the residual vector.

(ii) $\gamma' \geq 0$. If $\gamma' = 0$, Theorem 4 asserts that $\sqrt{n}(\hat{b}_{\mathrm{MR}} - b) \to 0$ in probability.

(iii) Preliminary calculations suggest the bias in $\hat{b}_{\mathrm{MR}}$ is $B/n + O(1/n^{3/2})$, where $B = -\lim \frac{1}{n}\sum_{i=1}^n \beta_i Z_i^2$, assuming the limit exists and Z is standardized as above.

6. The gain from adjustment

Compare Theorems 3 and 4 to see that the asymptotic gain from adjustment—the reduction in asymptotic variance—is

$$\frac{\Delta}{np(1-p)}, \quad \text{where } \Delta = (\overline{\alpha Z})\Big[(\overline{\alpha Z}) + 2(1-2p)(\overline{\beta Z})\Big]. \tag{23}$$

If $p = 1/2$, adjustment is either neutral or helps, because $(\overline{\alpha Z})^2 \geq 0$. If $p \neq 1/2$, then adjustment may hurt. For example, take $\overline{\alpha Z} > 0$ and $p > 1/2$. Another option is to take $\overline{\alpha Z} < 0$ and $p < 1/2$. Either way, take $\overline{\beta Z}$ large and positive. If $T_i = C_i$ for all i (the "strict null hypothesis"), then $\beta \equiv 0$ and adjustment will help—unless $\overline{\alpha Z} = 0$, i.e., the remaining variation (in C_i) is orthogonal to the covariate. A more interesting case to consider is the analysis of covariance with unequal numbers of subjects in treatment and control, and limiting quantities in (9a) and (16b) nontrivial.

7. The nominal variance

We turn to the asymptotic behavior of the "nominal" variances, that is, variances computed using the conventional formulas. Details are omitted, being very similar to those in Sections 3 and 4. Only convergence in probability is claimed, although a.s. convergence seems within reach. We follow the notation of Section 3.

Theorem 5. *Assume* (9). *Let* $\hat{\sigma}^2$ *be the mean square residual from the regression of* Y *on* 1 *and* $X - p$. *Let* $\hat{v}$ *be the nominal variance for the coefficient of* $X - p$, *i.e.,* $\hat{\sigma}^2$ *times the* $(2, 2)$ *element of* $(M'M)^{-1}$, *where* M *is the design matrix. Then*

(i) $\hat{\sigma}^2 \to \sigma^2 = \overline{\alpha^2} + p(1 - p)\overline{\beta^2}$.

(ii) $np(1 - p)\hat{v} \to \sigma^2$.

Theorem 6. *Assume* (9) *and* (16). *Let* $\hat{\sigma}^2$ *be the mean square residual from the regression of* Y *on* 1, $X - p$, *and* Z. *Let* $\hat{v}$ *be the nominal variance for the coefficient of* $X - p$, *i.e.,* $\hat{\sigma}^2$ *times the* $(2, 2)$ *element of* $(M'M)^{-1}$, *where* M *is the design matrix. Then*

(i) The intercept tends to a, *the coefficient of* $X - p$ *tends to* b, *and the coefficient of* Z *tends to* $\overline{\alpha Z}$.

(ii) $\hat{\sigma}^2 \to \sigma^2 = \overline{\alpha^2} - (\overline{\alpha Z})^2 + p(1 - p)\overline{\beta^2}$.

(iii) $np(1 - p)\hat{v} \to \sigma^2$.

Discussion

(i) For the notation, the constants a, b were defined in (5); $\overline{\alpha^2}$, $\overline{\alpha\beta}$, and $\overline{\beta^2}$ were defined in (9); $\overline{\alpha Z}$ and $\overline{\beta Z}$ were defined in (16).

(ii) For the simple regression estimator, the asymptotic variance is

$$\gamma/[np(1 - p)] :$$

see (9) and Theorem 3. If $p = 1/2$, then $\gamma \leq \sigma^2$; the inequality is strict unless $\beta \equiv 0$. If $p \neq 1/2$, the inequality can go either way: the nominal variance can be too big, or too small.

(iii) For the multiple regression estimator, the asymptotic variance is

$$\gamma'/[np(1-p)]:$$

see (16) and Theorem 4. Again, if $p = 1/2$, then $\gamma' \leq \sigma^2$; the inequality is strict unless $\beta \equiv 0$. If $p \neq 1/2$, the inequality can go either way: the nominal variance can be too big, or too small.

(iv) Calculations like those above give the asymptotic nominal variance of the ITT estimator as $\gamma''/np(1-p)$, where

$$\gamma'' = \overline{\alpha^2} + 2(1-2p)\overline{\alpha\beta} + [p^3 + (1-p)^3]\,\overline{\beta^2}.$$

Compare this with (9d): asymptotically, the nominal variance for the ITT estimator is conservative, by a considerable margin when $\overline{\beta^2}$ is large.

(v) The multiple regression model can be compared explicitly to the simple regression model in (4): according to the multiple regression model,

$$Y_i = a + b(X_i - p) + \theta Z_i + \delta_i',$$

where

$$\delta_i' = \delta_i - \theta Z_i = (\alpha_i - \theta Z_i) + \beta_i(X_i - p).$$

The quantities a, b, δ were defined in (5), and θ was defined in (20). In essence, part of δ has been explained by Z. If the error term satisfied the usual assumptions—but it doesn't—explaining part of δ would reduce the variance in $\hat{b}$.

(vi) How can we get to the usual multiple regression model from here? The idea seems to be this. Let ϵ_i be independent and identically distributed across subjects i. Let $C_i = c + dZ_i + \epsilon_i$ while $T_i = b + c + dZ_i + \epsilon_i$, for suitable constants b, c, d. Randomness in T_i and C_i is easily accommodated. However, independence, common distributions, and linearity of response—these are strong assumptions, not justified by the randomization.

(vii) In a variety of examples, simulation results (not reported here) indicate the following. When the number of subjects n is 100 or 250, bias in the multiple regression estimator may be quite noticeable. If n is 500, bias is sometimes significant, but rarely of a size to matter. With $n = 1000$, bias is negligible, and the asymptotics seem to be quite accurate.

(viii) The simulations, like the analytic results, indicate a wide range of possible behavior. For instance, adjustment may help or hurt. Nominal variances for the regression estimators can be too big or too small, by factors that are quite large. The simple regression estimator and the ITT estimator are the same, but their nominal variances may differ. (The regression model assumes constant variance, but the

nominal variance for the ITT estimator allows the treatment and control groups to have different variances.)

(ix) The ultimate variables that need to be balanced are T_i and C_i. Other variables are merely proxies: see Robins [22]. On the other hand, in practice, T_i and C_i are unknown, and the available regressors may be only weakly related to T_i and C_i—in which case the gains or losses from adjustment are likely to be minor. With a real experiment, if adjustment made a substantial difference, we would suggest much caution when interpreting results. That is the principal take-home message from the theory developed here.

(x) Practitioners will doubtless be heard to object that they know all this perfectly well. Perhaps, but then why do they so often fit models without discussing assumptions?

(xi) It appears that the theorems in the present paper can be proved by probabilistic calculations, and extensions of Höglund's central limit theorem [13] to the multivariate case with multiple samples—rather than matrix asymptotics. Those proofs would cover several levels of treatment, and several covariates.

(xii) The multiple regression estimator will be exactly unbiased in finite samples under severe regularity conditions. For instance, with two levels of treatment, the design should be balanced ($p = 1/2$) and there should be no subject-by-treatment interactions ($\beta \equiv 0$).

8. Other literature

The Neyman model is reviewed in Freedman [9, 10], with pointers to current literature on statistical models for causal inference, and discussion of the extent to which randomization justifies the regression model. A useful text on the design and analysis of clinical trials is Friedman, Furberg, and DeMets [11]. On study design in the behavioral and social sciences, see Shadish, Cook, and Campbell [25], Brady and Collier [3].

Data from many clinical trials are now filtered through the prism of conventional models, even when study populations number in the tens of thousands, perhaps to improve the balance between treatment and control groups, perhaps due to habit. Some investigators explicitly recommend adjusting data from clinical trials, using regression models and the like. A particularly enthusiastic paper is Victora, Habicht, and Bryce [26]. However, the validity of assumptions behind the models is rarely considered.

Two large clinical trials that attracted much attention at the time of writing are Rossouw, Anderson, Prentice, et al. [23], Howard, Van Horn, Hsia, et al. [15]. These papers report data from the Women's Health Initiative on the effects of hormone replacement therapy and low-fat diets. The key tables only give estimates from proportional-hazards models. Intention-to-treat analyses are not reported. However, there is enough summary data so that intention-to-treat estimates can be

reconstructed, and there is almost no difference between the modeling results and the ITT estimates. Blocking cannot be accounted for without unpublished data, but the combined effect of blocking and modeling is minor.

Substantive results should be mentioned: the experiments found no good effects from any intervention tested, including hormone replacement therapy and low-fat diets. On the other hand, a a modified Mediterranean diet shows great promise: see de Lorgeril, Salen, Martin, et al. [6].

When there is a conflict between models and experiments, some investigators definitely prefer the models. See, for instance, Prentice, Langer, Stefanick, et al. [20]. In this example, the models seem to lack adequate substantive foundations, and were somewhat post hoc, as noted by Petitti and Freedman [19]. For additional discussion from various perspectives, see Prentice, Pettinger, and Anderson [21]. Many social scientists analyze experimental data using regression models; one recent example is Chattopadhyay and Duflo [4]. An interesting comparison of model-based and intention-to-treat analyses will be found in Arceneaux, Gerber, and Green [2].

For discussion from the modeling viewpoint, see Koch, Tangen, Jung, et al. [16]. (By "non-parametric" analysis, these authors seem to mean fitting less-restrictive parametric models.) Lesaffre and Senn [17] criticize [16], from a perspective similar to the one adopted here. These two papers focus on the analysis of covariance. Also see Schmoor, Ulm, and Schumaker [24], who compare proportional hazards to CART.

Acknowledgments. Persi Diaconis, Thad Dunning, Don Green, Winston Lin, Stephen Senn, and Terry Speed made many useful comments.

References

[1] Anderson, T.W., J.B. Taylor, Strong consistency of least squares estimates in dynamic models, Annals of Statistics 7 (1979) 484–89.

[2] Arceneaux, K., A.S. Gerber, D.P. Green, Comparing experimental and matching methods using a large-scale voter mobilization experiment, Political Analysis 14 (2006) 37–62.

[3] Brady, H.E., D. Collier, eds., Rethinking Social Inquiry: Diverse Tools, Shared Standards. Rowman & Littlefield Publishers, Inc., Lanham, Maryland, 2004.

[4] Chattopadhyay, R., E. Duflo, Women as policy makers: Evidence from a randomized policy experiment in India, Econometrica 72 (2004) 1409–43.

[5] Dabrowska, D., T.P. Speed, On the application of probability theory to agricultural experiments: Essay on principles, English translation of Neyman [18], Statistical Science 5 (1990) 463–80 (with discussion).

[6] de Lorgeril, M., P. Salen, J.L. Martin, et al., Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction: Final report of the Lyon Diet Heart Study, Circulation 99 (1999) 779–85.

[7] Drygas, H., Consistency of the least squares and Gauss-Markov estimators in regression models, Zeitschrift für Wahrscheinlichkeitstheorie 17 (1971) 309–26.

[8] Freedman, D.A., Bootstrapping regression models, Annals of Statistics 9 (1981) 1218–28.

[9] Freedman, D.A., Statistical Models: Theory and Practice, Cambridge University Press, New York, 2005.

[10] Freedman, D.A., Statistical models for causation: What inferential leverage do they provide? Evaluation Review 30 (2006) 691–713.

[11] Friedman, L.M., C.D. Furberg, D.L. DeMets, Fundamentals of Clinical Trials, 3rd ed. Springer, New York, 2006.

[12] Greene, W.H., Econometric Analysis, 5th ed. Prentice Hall, Upper Saddle River, N.J., 2003.

[13] Höglund, T., Sampling from a finite population: A remainder term estimate, Scandinavian Journal of Statistics 5 (1978) 69–71.

[14] Holland, P.W., Statistics and causal inference, Journal of the American Statistical Association 8 (1986) 945–70 (with discussion).

[15] Howard, B.V., L. Van Horn, J. Hsia, et al., Low-fat dietary pattern and risk of cardiovascular disease: The Women's Health Initiative randomized controlled dietary modification trial, Journal of the American Medical Association 295 (2006) 655–66.

[16] Koch, G., C. Tangen, J. Jung, I. Amara, Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them, Statistics in Medicine 17 (1998) 1863—92.

[17] Lesaffre, E., S. Senn, A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials, Statistics in Medicine 22 (2003) 3583—96.

[18] Neyman, J., Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes, Roczniki Nauk Rolniczych 10 (1923) 1–51, in Polish.

[19] Petitti, D.B., D.A. Freedman, Invited commentary: How far can epidemiologists get with statistical adjustment? American Journal of Epidemiology 162 (2005) 1–4.

[20] Prentice, R.L., R. Langer, M. Stefanick, et al., Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial, American Journal of Epidemiology 162 (2005) 404–414.

[21] Prentice, R. L., M. Pettinger, G.L. Anderson, Statistical issues arising in the Women's Health Initiative, Biometrics 61 (2005) 899–941 (with discussion).

[22] Robins, J.M., Association, causation, and marginal structural models, Synthese 121 (1999) 151–79.

[23] Rossouw, J.E., G.L. Anderson, R.L. Prentice, et al., Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial, Journal of the American Medical Association 288 (2002) 321–33.

[24] Schmoor, C., K. Ulm, M. Schumacher, Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial, Statistics in Medicine 12 (1993) 2351—68.

[25] Shadish, W.R., T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton Mifflin, Boston, 2002.

[26] Victora, C.G., J.P. Habicht, J. Bryce, Evidence-based public health: Moving beyond randomized trials, American Journal of Public Health 94 (2004) 400–405.