Published in final edited form as:

Am Stat. 2014 October ; 68(4): 271–276. doi:10.1080/00031305.2014.940467.

A Paradoxical Result in Estimating Regression Coefficients

Thaddeus Tarpey,

Professor in the Department of Mathematics and Statistics, Wright State University, Day-ton, Ohio 45435

R. Todd Ogden,

Professor in the Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Eva Petkova, and

Associate Professor in the Department of Child and Adolescent Psychiatry, New York University, New York, NY 10016-6023 and a Senior Researcher at the Nathan S. Kline Institute for Psychiatric Research, Orangeburg, NY 10962

Ronald Christensen

Professor in the Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 8713

Abstract

This paper presents a counterintuitive result regarding the estimation of a regression slope coefficient. Paradoxically, the precision of the slope estimator can deteriorate when additional information is used to estimate its value. In a randomized experiment, the distribution of baseline variables should be identical across treatments due to randomization. The motivation for this paper came from noting that the precision of slope estimators deteriorated when pooling baseline predictors across treatment groups.

Keywords

Jensen's inequality; moderator analysis; pooled variance estimator; ratio estimator

1 Introduction

Conventional statistical wisdom holds that if a model parameter is known, then it should be used in place of an estimate when computing statistics. For example, in introductory statistics courses, the classic t -test is often introduced by replacing the (typically unknown) population standard deviation σ in the denominator of a z -statistic by the sample standard deviation s . De Veaux *et al.* (2014) write "...in the rare situation that we *know* σ , it would be foolish not to use that information" (page 480). Similarly, in a one-way ANOVA setting with k groups, to compare two of the means, students are taught that the test will have higher power if the data across all levels of the factor are pooled to estimate the variance (assuming the variances are equal across groups). The message is: you can increase precision and power when estimating the variance by pooling the data and you can further increase

precision and power if an oracle were available to tell you the value of σ . Paradoxically this conventional wisdom does not hold in the regression setting, as this paper will show.

In a randomized experiment (e.g., a clinical trial) where subjects are randomly assigned to treatment at baseline, the distribution of baseline predictors should be identical in each treatment group. In this setting one might want to estimate the effect of a baseline characteristic on the outcome, and this effect could be different in each treatment group. Thus, following the conventional statistical wisdom, in order to estimate the regression coefficients for baseline characteristics in the different treatment groups, it seems reasonable to pool the baseline predictors across the treatment groups to obtain more accurate coefficient estimates.

Consider two random variables x and y that are related via a simple linear regression, i.e.,

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad E[\varepsilon] = 0, \quad \text{var}(\varepsilon) = \sigma^2, \quad (1)$$

with x independent of ε . In particular,

$$E[y|x] = \beta_0 + \beta_1 x, \quad \text{var}(y|x) = \sigma^2.$$

Let

$$E \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \text{cov} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

It is well known that $E[y|x]$ is the best predictor of y and in this case the best predictor is a linear function so it must coincide with the best linear predictor of y . When the conditional expectation is linear, it follows that

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} \text{ and } \beta_0 = \mu_y - \beta_1 \mu_x. \quad (2)$$

If the regression of y on x , namely $E[y|x]$, is not linear, then the slope β_1 in (2) is the slope of the best linear approximation to $E[y|x]$ (Tarpey, 2011). In what follows, it is important to highlight the distinction between σ_y^2 , the variance of the outcome y , and the error variance σ^2 which is the conditional variance of y given x . In particular, the relation of the conditional variance σ^2 to σ_y^2 is given by

$$\sigma_y^2 = \text{var}(y) = \text{var}(E[y|x]) + E[\text{var}(y|x)] = \text{var}(\beta_0 + \beta_1 x) + E[\sigma^2] = \beta_1^2 \sigma_x^2 + \sigma^2, \quad (3)$$

from which it follows that

$$\sigma^2 = \sigma_y^2 - \beta_1^2 \sigma_x^2 = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2},$$

a relation we will use below.

Consider a random sample $(x_i, y_i)', i = 1, \dots, n$, and let

$$\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

denote the usual unbiased estimates of the mean and covariance parameters where

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{and } s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

The standard approach for estimating β_0 and β_1 is least-squares and the least-squares estimator of the slope is given by

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (4)$$

We will show that using a pooled estimate of the variance σ_x^2 across treatment groups to estimate the slope in (2) actually increases the variance of the slope estimator. In fact, using the actual (true population) variance σ_x^2 in place of the sample variance further inflates the variance of the slope estimator. We relate this phenomenon with the idea of the ratio estimator from sampling design.

We begin by considering an idealized case when σ_x^2 is known in Section 2 and then consider the case of estimating the slope from independent samples in Section 3. A simulation illustration is provided in Section 4. The multiple regression setting is discussed in Section 5 and a discussion is given in Section 6.

2 Estimating the Slope When the Variance is Known

To highlight the paradoxical result mentioned in the Introduction, consider an ideal scenario when the variance σ_x^2 of the predictor x is known. Although the coverage of regression in statistics textbooks often treats predictor variables as fixed, in many settings, including randomized experiments, the predictors may actually be random instead of fixed by design. It will be helpful to recall that if $W \sim \chi^2(r)$, then

$$\begin{aligned} E[W] &= r, \quad \text{var}(W) = 2r, \\ E[W^{-1}] &= \frac{1}{r-2}. \end{aligned} \quad (5)$$

Let $\mathbf{x} = (x_1, \dots, x_n)'$. We now derive the variance of the least-squares slope estimator (4) assuming the x_i are normal. First note that

$$E[\hat{\beta}_1 | \mathbf{x}] = \beta_1 \quad \text{and} \quad \text{var}(\hat{\beta}_1 | \mathbf{x}) = \frac{\sigma^2}{s_x^2(n-1)}.$$

Then

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \text{var}(E[\hat{\beta}_1 | \mathbf{x}]) + E[\text{var}(\hat{\beta}_1 | \mathbf{x})] \\ &= \text{var}(\beta_1) + E\left[\frac{\sigma^2}{s_x^2(n-1)}\right] \\ &= 0 + \frac{\sigma^2}{\sigma_x^2} E\left[\frac{(n-1)s_x^2}{\sigma_x^2}\right]^{-1} \\ &= \frac{\sigma_y^2 - \beta_1^2 \sigma_x^2}{(n-3)\sigma_x^2}, \\ &= \frac{1}{n-3} \left(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2 \right) \end{aligned} \quad (6)$$

which follows from (3) and noting that $(n-1)s_x^2/\sigma_x^2$ has a $\chi^2(n-1)$ distribution and using (5).

However, if σ_x^2 is known and used in place of the sample variance in the denominator of (4), we get an alternative estimator

$$\tilde{\beta}_1^* = \frac{s_{xy}}{\sigma_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x^2}, \quad (7)$$

which is an unbiased estimator of β_1 since the sample covariance s_{xy} is unbiased for σ_{xy} . The variance of $\tilde{\beta}_1^*$ can be computed as

$$\begin{aligned} \text{var}(\tilde{\beta}_1^*) &= E\left[\text{var}(\tilde{\beta}_1^* | \mathbf{x})\right] + \text{var}\left(E[\tilde{\beta}_1^* | \mathbf{x}]\right) \\ &= E\left[\frac{\sigma^2 s_x^2}{\sigma_x^4(n-1)}\right] + \text{var}\left(\frac{s_x^2}{\sigma_x^2} \beta_1\right) \\ &= \frac{\sigma^2}{(n-1)\sigma_x^2} + \frac{2\beta_1^2}{n-1} \\ &= \frac{1}{n-1} \left[\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2 \right] \end{aligned} \quad (8)$$

Except for a minor difference in leading fractions, from equations (6) and (8), it is clear that the usual least squares estimate of the slope has a smaller variance than the estimator that uses a known value of σ_x^2 which highlights the counterintuitive result. In fact, this is easy to see in an extreme special case when there is a perfect linear relationship $y = \beta_0 + \beta_1 x$. In this case $y_i - \bar{y} = \beta_1(x_i - \bar{x})$ and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1,$$

so we get the perfect estimation that we would expect from a perfect linear relationship.

However, for $\tilde{\beta}_1^*$, we have

$$\begin{aligned}\tilde{\beta}_1^* &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x^2} \\ &= \beta_1 \frac{s_x^2}{\sigma_x^2},\end{aligned}$$

which is not a perfect estimator of β_1 whenever $s_x^2 \neq \sigma_x^2$.

Another insight comes from considering the Cauchy-Schwartz inequality $\sigma_{xy}^2 \leq \sigma_y^2 \sigma_x^2$, which can be written equivalently as

$$\beta_1^2 = \frac{\sigma_{xy}^2}{\sigma_x^4} \leq \frac{\sigma_y^2}{\sigma_x^2}.$$

As β_1^2 reaches its maximum value, from (6) $\text{var}(\hat{\beta}_1)$ approaches 0, whereas from (8), $\text{var}(\tilde{\beta}_1^*)$ approaches $2\sigma_y^2/(\sigma_x^2(n-1))$. Only when the slope is near 0 are the variances comparable.

What is driving the superiority of the least-squares estimate over the estimate that uses more information, is that the least-squares estimate is exploiting the dependence between s_{xy} and s_x^2 to obtain a more efficient estimate of β_1 than the one obtained by using only s_{xy} and knowing σ_x^2 . The idea of the ratio estimator from sampling theory can also shed light on this result (e.g. Cochran, 1977, Chapter 6). It is easy to show that the sample covariance s_{xy} between x and y will be positively correlated with s_x^2 when $\beta_1 > 0$ (and will be negatively correlated when $\beta_1 < 0$). We can write

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{\sigma_x^2} \left(\frac{\sigma_x^2}{s_x^2} \right) = \tilde{\beta}_1^* w,$$

where

$$w = \frac{\sigma_x^2}{s_x^2}$$

is the ratio of the true variance to the estimated variance. Therefore the usual least-squares estimator of the slope is equal to the $\tilde{\beta}_1^*$ times a factor w . Suppose $\beta_1 > 0$ and a sample produces an estimate $\tilde{\beta}_1^* > \beta_1$. Then $s_{xy} > \sigma_{xy}$ and $w = \sigma_x^2 / s_x^2$ will tend to be less than one due to the positive correlation between s_{xy} and s_x^2 . Thus, when $\tilde{\beta}_1^*$ over-estimates β_1 , the ratio w will tend to pull $\hat{\beta}_1 = w\tilde{\beta}_1^*$ down towards the true slope β_1 .

3 Independent Samples

In the previous section the least-squares slope was shown to be a ratio-type estimator when σ_x^2 is known. In this section we consider situations, such as randomized clinical trials (RCTs), where there are two or more independent samples for which the predictor is recorded and the distribution of the predictor is identical across samples. Think of the predictor as a baseline characteristic of the subjects in an RCT. The slope estimator considered in this section then parallels the double-sampling estimator from sampling design (e.g., Thompson, 2012, Chapter 14).

Consider the case of a single predictor variable x and two treatment groups. The extension of the results to the case of more than two groups is straightforward. Let $\sigma_{y_j}^2$ denote the variance of the outcome y in the j th group, and let σ_{xj} denote the covariance between x and y in the j th group, $j = 1, 2$. Then the covariance matrix for the predictor and the outcome in the j th group can be expressed as

$$\Psi_j = \begin{pmatrix} \sigma_x^2 & \sigma_{xj} \\ \sigma_{xj} & \sigma_{y_j}^2 \end{pmatrix} \quad (9)$$

It will be convenient to use vector notation. Let $\mathbf{x}_{cj}$ denote the $n_j \times 1$ vector of the predictor variable for the j th group centered at the mean for the j th group. Here n_j is the sample size from the j th group. One could use the overall mean of the predictor across all treatments to center the x 's but this makes little difference for large sample sizes. The pooled estimate of the slope coefficient in the first group ($j = 1$ say) is

$$\tilde{\beta}_1 = \frac{n_1 + n_2 - 2}{n_1 - 1} \left[\sum_{j=1}^2 \mathbf{x}_{cj}' \mathbf{x}_{cj} \right]^{-1} \mathbf{x}_{c1}' \mathbf{y}_1. \quad (10)$$

Employing the variance-covariance decomposition formula again, and letting σ_1^2 denote the error variance in group $j = 1$, we find

$$\begin{aligned}
\text{var}(\tilde{\beta}_1) &= E \left[\text{var}(\tilde{\beta}_1 | \mathbf{x}_{c1}, \mathbf{x}_{c2}) \right] + \\
\text{var} \left(E[\tilde{\beta}_1 | \mathbf{x}_{c1}, \mathbf{x}_{c2}] \right) &= \left(\frac{n_1 + n_2 - 2}{n_1 - 1} \right)^2 \left\{ \sigma_1^2 E \left[\left(\sum_{j=1}^2 \mathbf{x}'_{cj} \mathbf{x}_{cj} \right)^{-2} (\mathbf{x}'_{c1} \mathbf{x}_{c1}) \right] + \right. \\
&\quad \left. \text{var} \left[\left(\sum_{j=1}^2 \mathbf{x}'_{cj} \mathbf{x}_{cj} \right)^{-1} (\mathbf{x}'_{c1} \mathbf{x}_{c1}) \beta_1 \right] \right\}.
\end{aligned} \tag{11}$$

Set

$$\begin{aligned}
V_1^2 &= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 / \sigma_x^2 \text{ and} \\
V_2^2 &= \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 / \sigma_x^2,
\end{aligned} \tag{12}$$

and note that V_1^2 and V_2^2 are independent chi-square random variables with $n_1 - 1$ and $n_2 - 1$ degrees of freedom respectively, assuming a normally distributed predictor. The variance of the pooled slope estimator (10) becomes

$$\begin{aligned}
\text{var}(\tilde{\beta}_1) &= \left(\frac{n_1 + n_2 - 2}{n_1 - 1} \right)^2 \left\{ \frac{\sigma_1^2}{\sigma_x^2} E \left[\frac{V_1^2}{(V_1^2 + V_2^2)^2} \right] + \right. \\
&\quad \left. \beta_1^2 \text{var} \left[\frac{V_1^2}{V_1^2 + V_2^2} \right] \right\}.
\end{aligned} \tag{13}$$

The ratio in the variance in the second term in (13) has a beta type I distribution (e.g. Mardia *et al.*, 1979, page 487), that can be written

$$\frac{V_1^2}{V_1^2 + V_2^2} = \frac{(\frac{n_1-1}{n_2-1})F_1}{1 + (\frac{n_1-1}{n_2-1})F_1}, \tag{14}$$

where

$$F_1 = \frac{V_1^2 / (n_1 - 1)}{V_2^2 / (n_2 - 1)},$$

with $F_1 \sim F(n_1 - 1, n_2 - 1)$. From well-known results relating the F -distribution to the beta-distribution it follows that the beta distribution of the ratio (14) has parameters $a = (n_1 - 1)/2$ and $b = (n_2 - 1)/2$ and consequently, the variance in the second term of (13) equals

$$\text{var} \left[\frac{V_1^2}{V_1^2 + V_2^2} \right] = \frac{2(n_1-1)(n_2-1)}{(n_1+n_2)(n_1+n_2-2)^2}. \quad (15)$$

Computing the expectation term in (13) is not straightforward, but the following lemma (whose proof is provided in the appendix) gives a lower bound.

Lemma 1

Let V_1 and V_2 denote independent chi-square random variables on ν_1 and ν_2 degrees of freedom respectively. Then

$$E \left[\frac{V_1^2}{(V_1^2 + V_2^2)^2} \right] \geq \frac{\nu_1}{(\nu_1 + \nu_2 + 2)^2}. \quad (16)$$

From (13), (15) and (16) from the lemma we have

$$\text{var}(\tilde{\beta}_1) \geq \frac{\sigma_x^2}{\sigma_x^2(n_1-1)} \left[\frac{n_1+n_2-2}{n_1+n_2} \right]^2 + \frac{2\beta_1^2}{n_1-1} \left[\frac{n_2-1}{n_1+n_2} \right]. \quad (17)$$

It is interesting to compare this lower-bound on the variance to the variance of the regular least-squares estimator (6) and the variance of the slope estimator based on a known predictor variance (8). Table 1 highlights these comparisons. The expression for the pooled estimator in the table is an approximation to (17).

Table 1 shows that the variance of the pooled estimator tends to lie in-between these two other variance expressions. In particular, pooling the data to estimate the slope as in (10) tends to make the slope estimator less precise compared to not pooling the data – the simulation results in Section 4 highlight this point. If n_1 is fixed and $n_2 \rightarrow \infty$, then the expression for the variance of the pooled estimator converges to the variance obtained when σ_x^2 is known (column 2 of the table). However, for finite n_1 and n_2 , the term involving β_1^2 in the variance of the pooled estimator is smaller than the corresponding term when σ_x^2 is known indicating that the pooled estimator performs better than the estimator that uses a known variance of x . If n_1 is much larger than n_2 , then the expression for the pooled estimator approximates the variance of the usual least-squares estimator. These comparisons for the pooled estimator may seem counterintuitive initially. However, in light of the previous results, the pooled estimator can exploit the dependence between the numerator and denominator in the slope estimator leading to a smaller variance compared to using a known variance when estimating the slope.

4 Simulation Illustration

This section presents a simulation experiment to highlight the results in the two previous sections. In this simulation, data are generated from two independent bivariate normal distributions with identical distributions for the predictor x . Specifically, sample sizes of $n_1 = n_2 = 50$ were used for each group. Using the notation introduced in Section 1 for the

parameters of the bivariate distribution of (x, y) , the mean μ_x of x was set to 1 and μ_y was set to 2 and 2.5 in groups 1 and 2 respectively. The variance of x was set to $\sigma_x^2 = 1.3^2$ and the variance σ_y^2 in the two groups was set to 4 and 6.25 respectively; the correlations between x and y in groups 1 and 2 were set to 0.8 and 0.7 respectively. 1000 data sets were simulated using the `rnorm` function in R (R Core Team, 2013).

In the simulation results, normal theory maximum likelihood estimators (mle) of the parameters were also obtained under the constraint that the distribution of the predictors was identical in both groups (i.e., the means and variances of x in the two groups are equal). Closed form expressions for the mle's of all of the parameters in the constrained model do not exist, so numerical solutions were found using the `optim` function in R. The maximum likelihood approach is not the primary focus of this paper, but for asymptotic results on mle's under constraints in the multi-sample setting, the interested reader is referred to Section 7 of Silvey (1959).

Figure 1 illustrates the results. The true slope in group 1 equals $\beta_1 = 1.23$ and is indicated by the thick vertical line. The solid (nonparametric) density curve is for the usual (non-pooled) least-squares slope estimate (4) (this density curve is essentially indistinguishable from the density curve for the mle's indicated by the short-dashed curve); the long-dashed density curve is from the pooled estimates of the slope (10) (this is based on estimating the mean of x separately in each group; using a single overall pooled estimate for the mean of x produces a density curve indistinguishable from the dashed curve.) Finally, the dotted curve is the density for the slope estimate (7) in group 1 assuming the variance of x is known.

Figure 1 gives a graphic illustration of how the different methods of estimating the slope compare with one another. All methods produce unbiased (or nearly unbiased) estimates of the slope, but the pooled method is much less precise than simply computing the usual least-square slope estimate using data from group 1 only. Also, using the known variance of x to compute the slope produces the least efficient estimator with the most spread in its density. The standard deviation from the 1000 simulated estimates of the slope in group 1 was 0.1395 for the mle which is virtually the same value (0.1372) as that obtained with the usual least-squares slope estimator using group 1 data only. The standard deviation for the pooled method was 0.2179 when all the data were pooled to estimate the common mean of x and was essentially the same (0.2199) when the mean of x was estimated separately in each group. The standard deviation for the estimates using the known variance of x was 0.2851 which was much greater than all other methods.

The results from estimating the slope for group 2 were similar and not shown here.

5 Multiple Regression Setting

Generalizations of the previous results for a single predictor carry over to the multiple regression setting with $p \geq 2$ predictors $\mathbf{x} = (x_1, \dots, x_p)'$. For a sample of n independent observations, let $\mathbf{X}$ denote the $n \times p$ matrix of predictors and let $\mathbf{X}_c$ denote the matrix of centered predictor variables (without a column of one's for the intercept). A vector of responses $\mathbf{y}$ is related to the vector of regression coefficients $\boldsymbol{\beta}$ by

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \mathbf{X}_c \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (18)$$

where it is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ is independent of $\mathbf{X}_c$. We will assume the models are full rank. Let $\boldsymbol{\Psi}_x$ denote the covariance matrix for $\mathbf{x}$ and let $\boldsymbol{\Psi}_{xy}$ denote the vector of covariances between $\mathbf{x}$ and the response y . Then the multiple regression generalization of (2) is

$$\boldsymbol{\beta} = \boldsymbol{\Psi}_x^{-1} \boldsymbol{\Psi}_{xy}. \quad (19)$$

The covariance matrix of the vector of least-square estimated coefficients

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_c' \mathbf{X}_c)^{-1} \mathbf{X}_c' \mathbf{y} \quad (20)$$

is easily shown to be $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 E[(\mathbf{X}_c' \mathbf{X}_c)^{-1}]$. From the law of large numbers, for large n ,

$$\text{cov}(\hat{\boldsymbol{\beta}}) \approx \frac{\sigma^2}{n-1} \boldsymbol{\Psi}_x^{-1} = \frac{\sigma_y^2 - \boldsymbol{\Psi}_{yx} \boldsymbol{\Psi}_x^{-1} \boldsymbol{\Psi}_{xy}}{n-1} \boldsymbol{\Psi}_x^{-1}, \quad (21)$$

and if $\mathbf{x}_i$ are normal, we get an equality with

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \frac{\sigma^2}{n-p-2} \boldsymbol{\Psi}_x^{-1} = \frac{\sigma_y^2 - \boldsymbol{\Psi}_{yx} \boldsymbol{\Psi}_x^{-1} \boldsymbol{\Psi}_{xy}}{n-p-2} \boldsymbol{\Psi}_x^{-1},$$

(see e.g. Anderson (1984), Lemma 7.7.1).

The multivariate generalization of (7) when $\boldsymbol{\Psi}_x$ is known is

$$\tilde{\boldsymbol{\beta}}^* = \boldsymbol{\Psi}_x^{-1} \hat{\boldsymbol{\Psi}}_{xy}, \quad (22)$$

where $\hat{\boldsymbol{\Psi}}_{xy}$ is the usual vector of sample covariances. Applying the (co)variance decomposition once again, it follows that when n is sufficiently larger than p ,

$$\text{cov}(\tilde{\boldsymbol{\beta}}^*) > \text{cov}(\hat{\boldsymbol{\beta}}),$$

(by this we mean that the matrix $[\text{cov}(\tilde{\boldsymbol{\beta}}^*) - \text{cov}(\hat{\boldsymbol{\beta}})]$ is positive definite) because, as shown in the appendix,

$$\begin{aligned} \text{cov}(\tilde{\boldsymbol{\beta}}_1^*) &= \frac{\sigma_y^2 + \boldsymbol{\Psi}_{yx} \boldsymbol{\Psi}_x^{-1} \boldsymbol{\Psi}_{xy}}{n-1} \boldsymbol{\Psi}_x^{-1} \\ &\approx \text{cov}(\hat{\boldsymbol{\beta}}) + \frac{2\boldsymbol{\beta}' \boldsymbol{\Psi}_x \boldsymbol{\beta}}{n-1} \boldsymbol{\Psi}_x^{-1} \end{aligned}$$

If data are available from two independent (e.g., treatment) groups, let $\mathbf{X}_{cj}$ denote the $n_j \times p$ matrix of centered predictor variables for the j th group ($j = 1, 2$). The pooled estimate of the vector of slope coefficients in the first group say is

$$\tilde{\beta}_1 = \frac{n_1 + n_2 - 2}{n_1 - 1} (\mathbf{X}'_{c1} \mathbf{X}_{c1} + \mathbf{X}'_{c2} \mathbf{X}_{c2})^{-1} \mathbf{X}'_{c1} \mathbf{y}_1. \quad (23)$$

From the variance-covariance decomposition formula again, we find

$$\text{cov}(\tilde{\beta}_1) = \left(\frac{n_1 + n_2 - 2}{n_1 - 1} \right)^2 \left\{ \sigma_1^2 E [(\mathbf{X}'_{c1} \mathbf{X}_{c1} + \mathbf{X}'_{c2} \mathbf{X}_{c2})^{-1} (\mathbf{X}'_{c1} \mathbf{X}_{c1}) (\mathbf{X}'_{c1} \mathbf{X}_{c1} + \mathbf{X}'_{c2} \mathbf{X}_{c2})^{-1}] + \text{cov} [(\mathbf{X}'_{c1} \mathbf{X}_{c1} + \mathbf{X}'_{c2} \mathbf{X}_{c2})^{-1} (\mathbf{X}'_{c1} \mathbf{X}_{c1}) \beta_1] \right\},$$

where σ_1^2 is the error variance for group 1. Evaluating this expression is not straightforward, but for n_1 and n_2 sufficiently large, the first term is approximately equal to (21) (with $n = n_1$), the approximate value of $\text{cov}(\hat{\beta})$ showing again that pooling the predictor variables over the two groups leads to a less precise estimator of the vector of regression coefficients.

6 Discussion

The motivation for this work originated from the problem of finding linear combinations of baseline covariates in clinical trials that produce strong modifying effects between treatments (i.e. linear combinations of baseline variables that maximize interaction effects between treatments when the outcomes are regressed on these linear combinations). From this work, we were surprised to find that pooling the baseline predictors to estimate a slope degraded statistical performance. This paper has illustrated the counterintuitive result that when extra information is used to estimate the slope, the performance of the estimator deteriorates. Usually, when extra information is used to estimate a quantity, the performance of the estimator improves - a couple well-known examples from sampling design are the ratio estimator when the population mean of an auxiliary variable is known (e.g. Cochran, 1977, Chapter 6) or double-sampling when data on an auxiliary variable is available (e.g. Thompson, 2012, Chapter 14). Interestingly, even if μ_x is known, the unbiased estimator of β_1

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \mu_x) y_i}{n \sigma_x^2},$$

performs worse than when μ_x is estimated.

If primary interest is on estimating a covariance σ_{xy} between x and y and data are available from two groups as in Section 3, then pooling the data across groups can yield a better estimator of the covariance by using the ideas presented in this paper. Specifically, if s_{xy} is

the usual sample covariance of σ_{xy} from, say group 1 (with sample size n_1), and if s_p^2 is the usual pooled estimator of σ_x^2 , then s_{xy} can be scaled to form a new estimator

$$\tilde{\sigma}_{xy} = [s_p^2 s_x^{-2}] s_{xy} = s_p^2 \hat{\beta}_1, \quad (24)$$

where s_x^2 and $\hat{\beta}_1$ are respectively the sample variance of x and the usual least-squares estimator of the slope from group 1. It is straightforward to show $\tilde{\sigma}_{xy}$ is unbiased for σ_{xy} . Under the normality assumption for x , it follows that

$$\text{var}(s_{xy}) = \frac{2\sigma_{xy}^2}{n_1 - 1} + \frac{\sigma_1^2 \sigma_x^2}{n_1 - 1}.$$

However, with n_2 denoting the sample size in group 2, a somewhat tedious derivation shows that

$$\text{var}(\tilde{\sigma}_{xy}) = \frac{2\sigma_{xy}^2}{n_1 + n_2 - 2} + \frac{\sigma_1^2 \sigma_x^2 (n_1^2 - 6n_1 - 5n_2 + 2n_1 n_2 + 10)}{(n_1 - 3)(n_1 + n_2 - 2)^2},$$

which can be quite a bit smaller than $\text{var}(s_{xy})$.

Acknowledgments

We would like to thank the Associate Editor and two reviewers for their helpful comments. This work was supported by NIMH grant R01 MH099003.

References

- Anderson, TW. An Introduction to Multivariate Statistical Analysis. 2. Wiley; New York: 1984.
- Cochran, WG. Sampling Techniques. 3. Wiley; 1977.
- De Veaux, RD.; Velleman, PF.; Bock, DE. Intro Stats. Pearson; Boston: 2014.
- Eaton, ML. Multivariate Statistics: A Vector Space Approach. Wiley; New York: 1983.
- Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate Analysis. Academic Press; London: 1979.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013.
- Silvey SD. The lagrangian multiplier test. Annals of Mathematical Statistics. 1959; 30:389–407.
- Tarpey T. On the meaning of parameters in approximation models. Journal of Probability and Statistical Science. 2011; 9:139–151.
- Thompson, SK. Sampling. 3. Wiley; New York: 2012.

7 Appendix

This appendix contains a proof of Lemma 1 using Jensen's inequality and also contains the derivation of $\text{cov}(\tilde{\beta}_1^*)$.

Proof of Lemma 1

The function $f(x_1, x_2) = x_1/(x_1 + x_2)^2$ for $x_1, x_2 > 0$ is not convex, so Jensen's inequality cannot be used directly. However, for a fixed value of x_1 , $f(x_2) = x_1/(x_1 + x_2)^2$ is a convex function of x_2 ; applying Jensen's inequality for conditional expectations we get

$$\begin{aligned} E \left[\frac{V_1^2}{(V_1^2 + V_2^2)^2} \right] &= E \left\{ E \left[\frac{V_1^2}{(V_1^2 + V_2^2)^2} \middle| V_1^2 \right] \right\} \\ &\geq E \left\{ \frac{V_1^2}{(V_1^2 + E[V_2^2])^2} \right\} \quad (\text{since } V_1^2 \text{ is independent of } V_2^2) \\ &= E \left\{ \frac{V_1^2}{(V_1^2 + \nu_2)^2} \right\} \end{aligned}$$

Let $g_\nu(x)$ denote a chi-square density function on ν -degrees of freedom. This last expectation equals

$$\begin{aligned} \int_0^\infty \frac{x}{(x + \nu_2)^2} g_{\nu_1}(x) dx &= \frac{\Gamma(\frac{\nu_1+2}{2}) 2^{(\nu_1+2)/2}}{\Gamma(\frac{\nu_1}{2}) 2^{\nu_1/2}} \int_0^\infty \frac{1}{(x + \nu_2)^2} g_{\nu_1+2}(x) dx \\ &= \nu_1 \int_0^\infty \frac{1}{(x + \nu_2)^2} g_{\nu_1+2}(x) dx \\ &= \nu_1 E_{\nu_1+2}[\{X + \nu_2\}^{-2}] \\ &\geq \nu_1 \frac{1}{(E_{\nu_1+2}[X] + \nu_2)^2} \\ &= \frac{\nu_1}{((\nu_1+2) + \nu_2)^2} \\ &= \frac{\nu_1}{(\nu_1 + \nu_2 + 2)^2}, \end{aligned}$$

where the probabilistic form of Jensen's inequality has been applied in the last integral and $E_{\nu_1+2}[X]$ denotes the expectation of a chi-square random variable on $\nu_1 + 2$ degrees of freedom.

To derive the covariance matrix of $\tilde{\beta}^*$, recall that $\mathbf{X}'_c \mathbf{X}_c \sim W(n-1, \Psi_x, 0)$, so that

$$\text{cov}\{\text{vec}[(\mathbf{X}'_c \mathbf{X}_c)]\} = 2(n-1) (\Psi_x \otimes \Psi_x),$$

see Eaton (1983, Chapter 8).

$$\begin{aligned}
\text{COV}(\hat{\beta}_1^*) &= \text{COV}\left(\frac{\Psi_x^{-1} \mathbf{X}_c' \mathbf{y}}{n-1}\right) \\
&= \frac{\Psi_x^{-1} [E(\text{COV}(\mathbf{X}_c' \mathbf{y} | \mathbf{X})) + \text{COV}(E(\mathbf{X}_c' \mathbf{y} | \mathbf{X}))] \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2 \Psi_x^{-1} E\left[\frac{\mathbf{X}_c' \mathbf{X}_c}{n-1}\right] \Psi_x^{-1}}{n-1} + \frac{\Psi_x^{-1} \text{COV}(\mathbf{X}_c' \mathbf{X}_c \beta) \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + \frac{\Psi_x^{-1} \text{COV}[\text{vec}(\mathbf{X}_c' \mathbf{X}_c \beta)] \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + \frac{\Psi_x^{-1} \text{COV}\{(\beta' \otimes \mathbf{I}_p) \text{vec}[(\mathbf{X}_c' \mathbf{X}_c)]\} \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + \frac{\Psi_x^{-1} (\beta' \otimes \mathbf{I}_p) \text{COV}\{\text{vec}[(\mathbf{X}_c' \mathbf{X}_c)]\} (\beta \otimes \mathbf{I}_p) \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + \frac{\Psi_x^{-1} (\beta' \otimes \mathbf{I}_p) \{2(n-1) (\Psi_x \otimes \Psi_x)\} (\beta \otimes \mathbf{I}_p) \Psi_x^{-1}}{(n-1)^2} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + 2 \frac{\Psi_x^{-1} (\beta' \Psi_x \beta \otimes \Psi_x) \Psi_x^{-1}}{n-1} \\
&= \frac{\sigma^2}{n-1} \Psi_x^{-1} + \frac{2\beta' \Psi_x \beta}{n-1} \Psi_x^{-1} \\
&= \frac{\sigma_y^2 - \Psi_{yx} \Psi_x^{-1} \Psi_{xy}}{n-1} \Psi_x^{-1} + \frac{2\Psi_{yx} \Psi_x^{-1} \Psi_{xy}}{n-1} \Psi_x^{-1} \\
&= \frac{\sigma_y^2 + \Psi_{yx} \Psi_x^{-1} \Psi_{xy}}{n-1} \Psi_x^{-1} \\
&\approx \text{COV}(\hat{\beta}) + \frac{2\beta' \Psi_x \beta}{n-1} \Psi_x^{-1}
\end{aligned}$$

Figure 1.

Non-parametric densities of the least-square, the pooled, the mle and the estimate with known $\sigma_x^2=1.3^2$. The true slope $\beta_1 = 1.23$ is indicated with solid vertical line.

Table 1

Comparison of the variances of the different slope estimators

Least Squares Estimator	σ_x^2 known	2-Samples Pooled Estimator
$\text{var}(\hat{\beta}_1) = \frac{1}{(n-3)} \left(\frac{\sigma_y^2}{\sigma_x^2} - \beta_1^2 \right)$	$\text{var}(\tilde{\beta}_1^*) = \frac{1}{(n-1)} \left(\frac{\sigma_y^2}{\sigma_x^2} + \beta_1^2 \right)$	$\text{var}(\tilde{\beta}_1) \gtrsim \frac{1}{n_1-1} \left(\frac{\sigma_y^2}{\sigma_x^2} + \frac{n_2-n_1}{n_1+n_2} \beta_1^2 \right)$