RESEARCH REPORT SERIES
(Statistics #2014-01)

Moderate-Sample Behavior of Adaptively Pooled Stratified Regression Estimators

Eric V. Slud

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: June 19, 2014

Moderate-Sample Behavior of Adaptively Pooled Stratified Regression Estimators

by Eric V. Slud, Census Bureau, CSRM, & Math. Dept., UMCP

Abstract. This report concerns the use of a preliminary test of equality of regression slopes in two-sample simple linear regression datasets for the purpose of deciding whether to pool the two datasets and perform a single analysis. This work was performed as a supplement to the paper of Shao, Slud et al. (2014), The latter paper cites this one as a Census Bureau preprint of 2012, and applies its results in the context of model-assisted design-based survey eastimation. For a broader overview of estimation following preliminary testing, see the book of Saleh (2006).

1 Problem Setting

We consider two-sample data of the form $\{(X_i, W_i)\}_{i=1}^m$ and $\{(Z_j, V_j)\}_{j=1}^n$ where each sample is *iid* with

$$E(W|X) = f(X), \;\; E(V|Z) = g(Z), \;\; \mathrm{Var}(W|X) = \sigma_1^2, \;\; \mathrm{Var}(V|Z) = \sigma_2^2$$

and both σ_1^2, σ_2^2 are constants. The restriction to constant conditional variances is the only serious assumption here, and can equivalently be termed an assumption of additive independent prediction errors (respectively with respect to X and to Z), with homoscedastic within-sample errors. In addition, we assume that these samples represent two portions ('strata') of a combined theoretical population with the (X, W) stratum a known proportion $\lambda \in (0, 1)$ of the whole, and that as the sample sizes m, n become large,

$$m/(m + n) \;\to\; \pi \;\in (0, 1)$$

where π is also known and may be different from λ. Finally, we assume that $\mu_X \equiv E(X)$ and $\mu_Z \equiv E(Z)$ are *known.*

Regarding W_i, V_j as stratumwise observations on a response variable and each of X_i, Z_j as the corresponding stratumwise observations on a scalar predictor variable, our objective can now be stated as estimation of the unknown mean $\mu_Y \equiv \lambda E(W) + (1 - \lambda)E(V)$.

Two regression-type unbiased estimators of μ_Y are given by

$$T \;=\; \lambda\,(\bar{W} + \hat{b}(\mu_X - \bar{X})) \;+\; (1-\lambda)\,(\bar{V} + \hat{\beta}(\mu_Z - \bar{Z})) \tag{1}$$

where

$$\hat{b} \;=\; \frac{\sum_{i=1}^m (X_i - \bar{X})W_i}{\sum_{i=1}^m (X_i - \bar{X})^2} \quad , \qquad \hat{\beta} \;=\; \frac{\sum_{j=1}^n (Z_j - \bar{Z})V_j}{\sum_{j=1}^n (Z_i - \bar{Z})^2}$$

and

$$S \;=\; \lambda\,(\bar{W} + \hat{\gamma}(\mu_X - \bar{X})) \;+\; (1-\lambda)\,(\bar{V} + \hat{\gamma}(\mu_Z - \bar{Z})) \tag{2}$$

where

$$\hat{\gamma} \;=\; \frac{\sum_{i=1}^m (X_i - \frac{m\bar{X}+n\bar{Z}}{m+n})W_i \;+\; \sum_{j=1}^n (Z_j - \frac{m\bar{X}+n\bar{Z}}{m+n})V_j}{\sum_{i=1}^m (X_i - \frac{m\bar{X}+n\bar{Z}}{m+n})^2 \;+\; \sum_{j=1}^n (Z_j - \frac{m\bar{X}+n\bar{Z}}{m+n})^2}$$

Here we use the standard notations $\bar{X} = m^{-1}\sum_{i=1}^m X_i$ and $\bar{Z} = n^{-1}\sum_{j=1}^m Z_j$, and S_X^2, S_Z^2 for the corresponding sample variances. In terms of these notations, the denominator of $\hat{\gamma}$ is readily checked to be equal to

$$D \;\equiv\; D(\mathbf{X},\mathbf{Z}) \;=\; (m-1)\,S_X^2 \;+\; (n-1)\,S_Z^2 \;+\; \frac{mn}{m+n}\,(\bar{X}-\bar{Z})^2$$

2 Large-sample behavior

We begin by considering the large-sample behavior of these statistics. First, define population slope parameters b_0, β_0 by

$$\begin{aligned} b_0 \;&=\; E((X_1-\mu_X)\,W_1)/E(X_1-\mu_X)^2 \;=\; E((X_1-\mu_X)\,f(X_1))/\sigma_X^2 \\ \beta_0 \;&=\; E((Z_1-\mu_Z)\,V_1)/E(Z_1-\mu_Z)^2 \;=\; E((Z_1-\mu_Z)\,g(Z_1))/\sigma_Z^2 \end{aligned} \tag{3}$$

The corresponding population intercept parameters are

$$a_0 \;=\; \mu_W - b_0\mu_X \quad , \qquad \alpha_0 \;=\; \mu_V - \beta_0\mu_Z \tag{4}$$

Then it is easy to see that, up to terms converging to 0 in probability, for large samples of sizes m, n satisfying $m/(m+n) \to \pi \in (0,1)$, the estimators $\hat{b}$ and $\hat{\beta}$ are respectively consistent for b_0, β_0, and $\sqrt{m+n}\,(T-\mu_Y)$

$$\approx\; \frac{\lambda}{\sqrt{\pi}}\,\sqrt{m}(\bar{W} - E(W) - b_0(\bar{X}-\mu_X)) \;+\; \frac{1-\lambda}{\sqrt{1-\pi}}\,\sqrt{n}(\bar{V} - E(V) - \beta_0(\bar{Z}-\mu_Z))$$

2.1 Null hypothesis of equal slopes

Under the further restriction of the same-regression null hypothesis $H_{00}:\; b_0 = \beta_0,\; a_0 = \alpha_0$, which equalizes both slopes and intercepts for the two-sample regression lines, one similarly checks that $\gamma_0 = b_0$ and $\sqrt{m+n}\,(S-\mu_Y) \;\approx$

$$\approx\; \frac{\lambda}{\sqrt{\pi}}\,\sqrt{m}(\bar{W} - E(W) - b_0(\bar{X}-\mu_X)) \;+\; \frac{1-\lambda}{\sqrt{1-\pi}}\,\sqrt{n}(\bar{V} - E(V) - b_0(\bar{Z}-\mu_Z))$$

from which it follows immediately that

$$\sqrt{m+n}\,(T-\mu_Y) \;-\; \sqrt{m+n}\,(S-\mu_Y) \;\overset{P}{\to}\; 0 \tag{5}$$

The conclusion (5) which holds under H_{00} persists also under contiguous alternatives (van der Vaart 2000, Ch. 6), e.g., assuming suitable regularity conditions on the two-sample joint densities, under alternatives in which $\beta_0 - b_0 = O(1/\sqrt{m+n})$. This result has been proved in a superpopulation survey-sampling framework by Shao et al. (2011).

2.2 Fixed alternatives: distinct slopes

More generally, when b_0 and β_0 are unequal, the asymptotic form of S has a different centering: the population parameter γ_0 is given as

$$\frac{\lambda\, b_0\, E(X_1-\mu_X)^2 + (1-\lambda)\,\beta_0\, E(Z_1-\mu_Z)^2 + \lambda(1-\lambda)(\mu_X-\mu_Z)(a_0+b_0\mu_X-\alpha_0-\beta_0\mu_Z)}{\lambda E(X_1-\mu_X)^2 + (1-\lambda)E(Z_1-\mu_Z)^2 + \lambda(1-\lambda)(\mu_X-\mu_Z)^2}$$

and

$$\sqrt{m+n}\,(S-\mu_Y) \;\approx\; \frac{\lambda\sqrt{m}}{\sqrt{\pi}}(\bar{W}-\mu_W) \;+\; \frac{(1-\lambda)\sqrt{n}}{\sqrt{1-\pi}}(\bar{V}-\mu_V)$$
$$-\,\gamma_0\sqrt{m+n}(\lambda(\bar{X}-\mu_X)+(1-\lambda)(\bar{Z}-\mu_Z))$$

Typically, under alternatives to H_{00}, in particular when $b_0 \neq \beta_0$, a hypothesis test of equality of slopes based on $\hat{b}-\hat{\beta}$ will reject with probability approaching 1 for large sample size. Now regardless of the validity of regression model assumptions, the estimators S,T are both asymptotically $\sqrt{m+n}$ unbiased estimators for μ_Y. One might intuitively expect the estimator S to be better in the sense of smaller variance, under the assumption H_{00}, and T to be better under alternatives. It is the purpose of this Note to examine whether that intuition is correct.

3 Moderate Samples: Conditional Variance and Unconditional MSE

Using the large-sample equivalent forms for S and T developed in the previous Section, there are no large-sample settings in which the top-order variance or Mean-Squared Error (MSE) for T as an estimator of μ_Y will be worse than that of S. However, numerical experience shows that in small or moderate sized samples, a unified regression analysis can confer a benefit in providing an estimator less sensitive to outliers, and we explore this formally by studying MSE's for T versus S, taking lower-order terms into account.

We assume the additive constant-variance error structure of the two-sample problem in Section 1, condition on $\{X_i\}_i$, $\{Z_j\}_j$, and treat T and S as linear estimators respectively in the variables W_i and V_j. To economize on lengthy expressions, we define the notation

$$\Delta \;=\; \lambda\,(\mu_X - \bar{X}) \;+\; (1-\lambda)\,(\mu_Z - \bar{Z})$$

and recall the notation $D = D(\mathbf{X}, \mathbf{Z})$ defined above. Then

$$T \;=\; \lambda \sum_{i=1}^{m} \Big(\frac{1}{m} + \frac{(\mu_X - \bar{X})(X_i - \bar{X})}{(m-1)S_X^2}\Big) W_i + (1-\lambda) \sum_{j=1}^{n} \Big(\frac{1}{n} + \frac{(\mu_Z - \bar{Z})(Z_j - \bar{Z})}{(n-1)S_Z^2}\Big) V_j \tag{6}$$

and

$$\begin{aligned} S \;&=\; \sum_{i=1}^{m} \Big(\frac{\lambda}{m} \;+\; \frac{\Delta}{D}\,(X_i - \frac{m\bar{X} + n\bar{Z}}{m+n})\Big) W_i \qquad (7) \\ &+ \sum_{j=1}^{n} \Big(\frac{1-\lambda}{n} \;+\; \frac{\Delta}{D}\,(Z_j - \frac{m\bar{X} + n\bar{Z}}{m+n})\Big)\, V_j \end{aligned}$$

Then we obtain, by direct calculation,

$$\mathrm{Var}(T|\mathbf{X},\mathbf{Z}) \;=\; \frac{\lambda^2\,\sigma_1^2}{m} + \frac{(1-\lambda)^2\,\sigma_2^2}{n} + \frac{\lambda^2\sigma_1^2(\mu_X - \bar{X})^2}{(m-1)S_X^2} + \frac{(1-\lambda)^2\sigma_2^2(\mu_Z - \bar{Z})^2}{(n-1)S_Z^2} \tag{8}$$

and

$$\begin{aligned} \mathrm{Var}(S|\mathbf{X},\mathbf{Z}) \;\;&=\;\; \frac{\sigma_1^2}{m}\Big(\lambda \;+\; \frac{mn(\bar{X} - \bar{Z})\Delta}{(m+n)D}\Big)^2 \;+\; \frac{\sigma_2^2}{n}\Big(1 - \lambda + \frac{mn(\bar{Z} - \bar{X})\Delta}{(m+n)D}\Big)^2 \\ &+\;\; \Big((m-1)S_X^2\,\sigma_1^2 \;+\; (n-1)S_Z^2\,\sigma_2^2\Big)\,(\Delta/D)^2 \qquad (9) \end{aligned}$$

Using the same representations of T, S as linear estimators, we obtain exact formulas for conditional bias:

$$\begin{aligned} E(T|\mathbf{X},\mathbf{Z}) \;-\; \mu_Y \;&=\; \frac{\lambda}{m} \sum_{i=1}^{m} (f(X_i) - Ef(X_1)) \;+\; \frac{1-\lambda}{n} \sum_{j=1}^{m} (g(Z_j) - Eg(Z_1)) \qquad (10) \\ &+ \frac{\lambda(\mu_X - \bar{X})}{(m-1)S_X^2} \sum_{i=1}^{m} (X_i - \bar{X}) f(X_i) \;+\; \frac{(1-\lambda)(\mu_Z - \bar{Z})}{(n-1)S_Z^2} \sum_{j=1}^{n} (Z_j - \bar{Z}) g(Z_j) \end{aligned}$$

$$\begin{aligned} E(S|\mathbf{X},\mathbf{Z}) - \mu_Y \;&=\; \frac{\lambda}{m} \sum_{i=1}^{m} (f(X_i) - Ef(X_1)) + \frac{1-\lambda}{n} \sum_{j=1}^{m} (g(X_i) - Eg(X_1)) + \qquad (11) \\ &\frac{\Delta}{D}\Big[\sum_{i=1}^{m} (X_i - \frac{m\bar{X} + n\bar{Z}}{m+n}) f(X_i) \;+\; \sum_{j=1}^{n} (Z_j - \frac{m\bar{X} + n\bar{Z}}{m+n}) g(Z_j)\Big] \end{aligned}$$

We now continue with calculations based on these formulas in the most interesting cases of homoscedastic linear models within substrata.

3.1 *Linear Regression*

The main restriction allowing us to simplify and to compute and compare Mean Squared Errors of estimation is the restriction to stratumwise linear models. That is, we assume

$$W = a + bX + \epsilon_1, \quad f(X) = a + bX, \quad V = \alpha + \beta Z + \epsilon_2, \quad g(Z) = \alpha + \beta Z$$

In this case, substitution into formula (10) immediately shows $E(T|\mathbf{Z},\, \mathbf{X}) = 0$, and formula (8) is already in as simple a form as possible.

In addition, we assume for simplicity that the two strata are separated by a cut-point c, with $X < c < Z$ and with the linear regressions joining continuously at the known cut-point c. Then, if we define $\delta \equiv \beta - b$,

$$a + bc \;=\; \alpha + \beta c \quad \Rightarrow \quad a - \alpha = (\beta - b)c = \delta c$$

Formula (11) becomes

$$E(S|\mathbf{X},\, \mathbf{Z}) - \mu_Y = \lambda b(\bar{X} - \mu_X) + (1-\lambda)\beta(\bar{Z} - \mu_Z) + \frac{\Delta}{D}\Big\{(a-\alpha)\,\frac{mn}{m+n}\,(\bar{X} - \bar{Z}) +$$

$$+ \, b(m-1)S_X^2 + \beta(n-1)S_Z^2 + (b\bar{X} - \beta\bar{Z})\,\frac{mn}{m+n}\,(\bar{X} - \bar{Z})\Big\}$$

$$= \;\lambda b(\bar{X} - \mu_X) \,+ (1-\lambda)\beta(\bar{Z} - \mu_Z) +$$

$$+\,\frac{\Delta}{D}\Big\{bD \,+\, \delta\,((n-1)S_Z^2 - \bar{Z}\,\frac{mn}{m+n}\,(\bar{X} - \bar{Z})) \,+\, \delta c\,\frac{mn}{m+n}\,(\bar{X} - \bar{Z})\Big\}$$

$$= \;(1-\lambda)\delta(\bar{Z} - \mu_Z) \,+\, \frac{\Delta\,\delta}{D}\,((n-1)S_Z^2 \,+\, (c - \bar{Z})(\bar{X} - \bar{Z})\,\frac{mn}{m+n})$$

One consequence of these formulas is that the conditional variances are free of the quantity δ, while the conditional bias $E(S|\mathbf{X}, \mathbf{Z}) - \mu_Y$ is directly proportional to δ. (In particular, the conditional bias $E(S|\mathbf{X}, \mathbf{Z}) = 0$ under the null hypothesis $\delta = 0$.) Similarly, the conditional variances are linear in σ_1^2 and σ_2^2, while the conditional biases do not involve these variances at all. Since it turns out that the quantities

$$MSE(S) = E(Var(S|\mathbf{X}, \mathbf{Z})) \,+\, E(\big[E(S|\mathbf{X}, \mathbf{Z}) - \mu_Y\big]^2)$$

and $MSE(T) = E(Var(T|\mathbf{X}, \mathbf{Z}))$ are generally related by $MSE(S)\Big|_{\delta=0} < MSE(T)$, we can display the relationships for positive δ by telling

(i) the relative improvement $1 - MSE(S)/MSE(T)$ of S over T at $\delta = 0$, and

(ii) the value $\delta^2/(\sigma_1^2 + \sigma_2^2)$ at which $MSE(S) = MSE(T)$.

Since these quantities involve expectations which are difficult to find analytically, we provide accurate estimates through simulations of $R = 1000$ replications.

Table 1: Estimates based on simulations with $R = 1000$ replications, for various distributions of X_i and values $m, n,$ and $\gamma = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$, of relative MSE (delMSE $=$ 1-MSE(S)$|_{\delta=0}$/MSE(T)) and of the value $\delta_* = (\beta - b)/\sqrt{\sigma_1^2 + \sigma_2^2}$ for δ at which MSE(S)=MSE(T). In all cases, $q = \lambda = 0.8$.

Dist. of ξ	m	n	γ	delMSE	δ_*
$\mathcal{N}(4,1)$	100	50	.5	.010	.066
	50	30		.024	.147
	40	20		.027	.170
	100	50	.3	.013	.061
	50	30		.023	.114
	50	30		.032	.154
Expon(1)	100	50	.5	.013	.034
	50	30		.027	.071
	40	20		.035	.093
Gamma(2, 1)	100	50	.5	.011	.027
	50	30		.022	.053
	40	20		.032	.074
Weib(1.5, 1)	100	50	.5	.011	.072
	50	30		.021	.140
	40	20		.029	.186
Lognorm(0, 1)	100	50	.5	.016	.012
	50	30		.029	.022
	40	20		.043	.030

The results are tabulated below. In Table 1, the random variables X, Z are taken to be distributed with the conditional distribution of a random variable ξ respectively given $\xi < c$ and given $\xi > c$, where the distribution of ξ and the quantile $q = P(\xi \leq c)$ are specified. Note that the ratios MSE(S)/MSE(T) are invariant under location shifts in ξ or under scaling that multiplies ξ and each of σ_1^2, σ_2^2 by the same constant k.

Note that almost all reasonable parameter combinations result in $MSE(S) < MSE(T)$ under the null hypothesis $\delta = 0$, as the result proved in the next subsection indicates. Examples where $MSE(S) \geq MSE(T)$ are easily calculated to arise when λ is very small but $\lambda^2\sigma_1^2/\sigma_2^2$ is large, or when $1 - \lambda$ is small and $(1 - \lambda)^2\sigma_2^2/\sigma_1^2$ is large, but neither of these cases is very likely to occur in practice.

All of the numerical calculations described here were done in `R` (R Core Development Team, 2009).

3.2 *Further Restricted Cases*

It is worth remarking on the special outcomes of the previous conditional bias and variance formulas in a few special cases further restricting the linear regression setup of the previous subsection. First if $\mu_X = \bar{X}$ and $\mu_Z = \bar{Z}$, then $\Delta = 0$ and $E(S|X,Z) = \mu_Y$ for all values of δ, and $\mathrm{Var}(S|X,Z) = \mathrm{Var}(T|X,Z)$. The same result holds if $\Delta = \delta = 0$. In thes settings, the MSE's of S and T are identically equal. However, these cases rely on special data values. A more important case, where $Var(S|\mathbf{X},\mathbf{Z}) \le Var(T|\mathbf{X},\mathbf{Z})$ for all data values under an important general set of parameter values, is provided in the following result.

Proposition 1 *Assume as above that* $\{(W_i, X_i)\}_{i=1}^m$ *are iid with* $E(W_i \,|\, X_i) = a + bX_i$, $Var(W_i \,|\, X_i) = \sigma_1^2$, *and similarly that* $\{(V_j, Z_j)\}_{j=1}^n$ *are iid with* $E(V_j \,|\, Z_j) = \alpha + \beta Z_j$, $Var(V_j \,|\, Z_j) = \sigma_2^2$. *Further, assume that for some fixed* c, $a + bc = \alpha + \beta c$, *and define* $\delta = \beta - b$. *With* S, T *defined as above, in terms of* $\lambda \in (0,1)$, *assume further that*

$$\delta = 0 \quad , \qquad \sigma_1^2 = \sigma_2^2 = \sigma^2 \qquad \textit{and} \qquad \lambda = m/(m+n) \tag{12}$$

Then for all $\mathbf{X}$, $\mathbf{Z}$, $MSE(S) < MSE(T)$.

Proof. Under the assumptions (12), we check immediately from (9) that

$$\mathrm{Var}(T \,|\, \mathbf{X}, \mathbf{Z}) \;=\; \frac{\sigma^2}{m+n} \;+\; \sigma^2 \left\{ \lambda^2 \, \frac{(\mu_X - \bar{X})^2}{(m-1)S_X^2} \;+\; (1-\lambda)^2 \, \frac{(\mu_Z - \bar{Z})^2}{(n-1)S_Z^2} \right\}$$

and

$$\mathrm{Var}(S \,|\, \mathbf{X}, \mathbf{Z}) \;=\; \frac{\sigma^2}{m+n} \;+\; \sigma^2 \, \frac{\Delta^2}{D}$$

Moreover, by the Cauchy-Schwarz inequality,

$$\Delta^2 \;\le\; \left\{ \frac{\lambda^2 \, (\mu_X - \bar{X})^2}{(m-1)S_X^2} \;+\; \frac{(1-\lambda)^2 \, (\mu_Z - \bar{Z})^2}{(n-1)S_Z^2} \right\} ((m-1)S_X^2 \;+\; (n-1)S_Z^2)$$

The combination of the last three displayed expressions yields

$$\frac{\mathrm{Var}(S \,|\, \mathbf{X}, \mathbf{Z}) \;-\; \sigma^2/(m+n)}{\mathrm{Var}(T \,|\, \mathbf{X}, \mathbf{Z}) \;-\; \sigma^2/(m+n)} \;\le\; 1 \;-\; \frac{mn(\bar{X} - \bar{Z})^2}{(m+n)\, D}$$

Since the conditional Variances are the same as conditional MSE's at $\delta = 0$, the Proposition has been proved, and the inequality holds with probability 1 when X_i and Z_j are continuously distributed. □

4 Tentative Conclusions

Our provisional conclusion is that, at least in the case of substrata within which there are two similar linear regression models which join continuously at the cut-point, the MSE comparison between the one- and two- stratum estimators S and T is broadly similar: S is superior for alternatives $\delta = \beta - b$ very close to 0. But as soon as δ exceeds a proportion δ_* of $(\sigma_1^2 + \sigma_2^2)^{1/2}$ ranging from 3% to 20%, depending on the distribution of ξ, then T becomes superior. This break-even proportion δ_* does depend strongly on the distribution, and is much larger for highly skewed distributions (exponential, weibull, gamma) and if anything is smaller for less-skewed long-tailed distributions (log-normal). Note that this discussion takes no account of the special features of the survey-sampling origins of the problem studied here, especially the feature of biased sampling through unequal-probability weights, and those aspects of MSE comparisons will be studied by simulation elsewhere.

5 References

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Saleh, A.K. Md. (2006). **Theory of Preliminary Test and Stein-type Estimation, with Applications.** Wiley-Interscience, Hoboken.

Shao, J., Slud, E., Cheng, Y., Wang, S. and Hogue, C. (2014), Theoretical and empirical properties of model assisted decision-based regression estimators, to appear in *Survey Methodology*.

Van der Vaart, A. (2000) **Asymptotic Statistics**, Cambridge Univ. Press.