MODEL-BASED ANALYSIS OF RECORDS LINKED USING MIXTURE MODELS

Partha Lahiri, University of Nebraska-Lincoln, and Michael D. Larsen, The University of Chicago
Michael D. Larsen, Department of Statistics, 5734 University Avenue, Chicago, Illinois 60637,
larsen@galton.uchicago.edu

Key Words: Fellegi-Sunter, Latent Class Model, Measurement Error, Mixture Models, Record Linkage

1 Introduction

There has been an increasing use of computerized record linkage (CRL) methods in various federal statistical systems (Alvey and Jamerson 1997). The methods quickly link records available from different data sources to create new, enhanced files. Since CRL utilizes already existing databases, it saves the substantial time and resources needed to collect new data. Various government agencies have developed sophisticated software to implement CRL, usually attaching weights reflecting the likelihood of a match to pairs of records. Statistics Canada uses a software called CANLINK for this purpose. The U. S. Bureau of the Census uses software by Winkler (1994, 1995) and Jaro (1989, 1995).

If mismatch errors are introduced by CRL, statistical analyses based on linked data can be adversely affected. Relatively little work has done to address this important issue. Neter et al. (1965) studied the effect of mismatch errors in finite population sampling. They observed that relatively small mismatch error could lead to a substantial bias in estimating the relationship between response errors and true values. Theoretical and computational advances in estimating matching probabilities (see, e.g., Belin and Rubin 1995; Winkler and Thibaudeau 1991) motivated Scheuren and Winkler (1993), henceforth referred to as SW, to revisit the work of Neter et al. (1965). Specifically, they investigated the effect of mismatch errors on the bias of ordinary least squares estimators of regression coefficients in a standard regression model and proposed a method of adjusting for the bias. Scheuren and Winkler (1997) advanced the work further with an iterative procedure that adjusted the regression and matching results for apparent outliers.

The purpose of this article is to consider an alternative to the bias correction method considered in SW (1993). In Section 2, we review the SW method and then propose a new method of estimating regression coefficients in the presence of mismatch errors. Our proposed estimator involves matching probabilities and is unbiased when the matching probabilities are all known. In Section 3, we propose a variance estimator of our proposed estimator. Simulation results are presented in Section 4. Our method improves on a naive, a robust, and the SW (1993) method. Technical proofs are deferred to a technical report (Lahiri and Larsen 2000).

2 Estimation of Regression Coefficients

Consider the following regression model:

$$y_i = x_i'\beta + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where $x_i' = (x_{i1}, \cdots, x_{ip})'$ is a vector of p known covariates and $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j, \quad i, j = 1, \ldots, n$. We are interested in a situation where the response variable (y) is in file A and all the covariates (x) are in file B and these two files are linked using a CRL technique. Thus, in our situation, the true responses $y_i's$ are not observable. Instead, we observe $z_i's$ which may or may not correspond to x_i. SW (1993) considered the following model:

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } i \neq j, \end{cases} \tag{2}$$

where $\sum_{j=1}^{n} q_{ij} = 1, \quad i, j = 1, \ldots, n (i \neq j)$. The naive estimator of β which ignores mismatch errors is given by $\hat{\beta}_N = (X'X)^{-1}X'Z$, where $X = (x_1', \ldots, x_n')'$ and $Z = (z_1, \ldots, z_n)'$. An alternative to this naive estimator would be to use a robust estimator such as an estimator that minimizes the sum of absolute deviations.

SW investigated the bias of $\hat{\beta}_N$ conditional on the $y_i's$. It can be seen from the calculations of SW that

$$\mathrm{Bias}(\hat{\beta}_N) = E[\hat{\beta}_N - \beta | y] = (X'X)^{-1}X'B, \tag{3}$$

where $B = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij} y_j$. If an estimator of B, say $\hat{B}$, is available, then the SW estimator is given by

$$\mathrm{Bias}(\hat{\beta}_{SW}) = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}. \tag{4}$$

Let q_{ij_1} and q_{ij_2} denote the highest and the second highest elements of the vector $q_i = (q_{i1}, \cdots, q_{in})'$ and let z_{j_1} and z_{j_2} denote the corresponding elements in the vector Z. Then an estimator of D_i is given by $\hat{D}_i =$

$(q_{ij_1}-1)z_{j_1}+q_{ij_2}z_{j_2}$ (see SW 1993). The SW estimator does not produce an exact unbiased estimator of β even when the matching probabilities are available. In order to obtain an exact unbiased estimator of β, we first note that under the model described by (1) and (2),

$$\mathrm{E}(z_i) = w_i'\beta, \quad i = 1, \ldots, n,$$

where $w_i = \sum_{j=1}^{n} q_{ij}x_j, \quad i = 1, \ldots, n$ (see Lahiri and Larsen 2000). Thus, an unbiased estimator of β is given by

$$\hat{\beta}_U = (W'W)^{-1}W'Z,$$

where $W' = (w_1, \ldots, w_n)$.

3 Estimation of Var($\hat{\beta}_U$)

Define $c_{ij} = \sum_{l=1}^{n} q_{il}q_{jl}$, $d_{ij} = x_j - w_i$, and $A_i = \sum_{j=1}^{n} q_{ij}d_{ij}d_{ij}'$ for $(i, j = 1, \ldots, n, \quad i \neq j)$. Then we have (see Lahiri and Larsen 2000),

$$\mathrm{Var}(Z) = \Sigma = ((\sigma_{ij})),$$

where $\sigma_{ii} = \mathrm{var}(z_i) = \sigma^2 + \beta' A_i \beta$, and $\sigma_{ij} = \mathrm{cov}(z_i, z_j) = \sigma^2 c_{ij}$ for $i, j = 1, \ldots, n; i \neq j$. Hence $\mathrm{Var}(\hat{\beta}_U) = (W'W)^{-1}W'\Sigma W(W'W)^{-1}$.

Let us now turn our attention to the estimation of $Var(\hat{\beta}_U)$. This requires estimation of σ^2. A naive estimator of mean squared error (MSE) based on Z is given by

$$\mathrm{MSE} = \frac{1}{n-p} Z'(I - X'(X'X)^{-1}X)Z.$$

Note that MSE is not unbiased for σ^2 under the model described by (1) and (2).

In order to obtain an estimator of σ^2, consider

$$S^2 = Z'(I - W(W'W)^{-1}W')Z.$$

Note that (see Lahiri and Larsen 2000)

$$\mathrm{E}(S^2) = (n - p - a)\sigma^2 + \beta' D \beta, \tag{5}$$

where $a = \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_{ij}c_{ji}$, $D = \sum_{i=1}^{n}(1 - h_{ii})A_i$ and $h_{ij} = w_i'(W'W)^{-1}w_j$. The equation (5) motivates us to consider the following estimator of σ^2 :

$$\hat{\sigma}^2 = \max\{0, \frac{S^2 - \hat{\beta}_U' D \hat{\beta}_U}{n - p + a}\}.$$

It can be shown that $\hat{\sigma}^2$ is consistent for σ^2 under the model described by (1) and (2) and under certain mild regularity conditions (see Lahiri and Larsen 2000).

4 Simulation

The performance of the regression methods will be studied through a simulation. First, the simulation conditions are described. Then, results focusing on bias are presented. Variance esimation will be demonstrated in later work.

4.1 Simulation conditions

One hundred replications are performed under each of four conditions. In cases 1 and 2, files A and B each have 1200 records, which are grouped into 80 blocks of 15. It is assumed that every record has a match, so 1/15 of record pairs are matches. In cases 3 and 4, the two files each have 1000 records grouped in to blocks of 20. In the latter cases, 1/20 of record pairs are matches. The regression slope is .3 in cases 1 and 3 and .6 in cases 2 and 4. The standard deviation on the normal error term is .5. R-squared values are close to the regression slope value for the four cases. In cases 1 and 2 there are 10 matching fields, whereas in cases 3 and 4 there are only 8. The probability that a matched pair agrees on a specific comparison is always .75, but the probability that a nonmatch pair agrees on a comparison increases from .20 for cases 1 and 2 to .25 for cases 3 and 4. Table 6 displays the simulation conditions.

The files A and B were generated and comparison vectors calculated. The EM algorithm (Dempster, Laird, Rubin 1977) was used to fit a two-class conditional independence mixture model to the comparison vectors to estimate probabilities for the Fellegi-Sunter (1969) algorithm. The EM algorithm for record linkage examples is described in many places, including Larsen and Rubin (2001), Armstrong and Mayda (1993), and Winkler (1988). Cases 1 and 2 and cases 3 and 4 have similar Fellegi-Sunter weights, respectively. Cases 1 and 2 can be described as the good matching scenario in which there is little overlap of weight distributions. Cases 3 and 4 are mediocre matching scenarios in comparison to cases 1 and 2.

4.2 Results

Results are presented for the four simulations separately in figures 1 through 4 and in tables 2 through 5. In figure 1, histograms of regression estimates in 100 simulation replications under case 1 conditions are presented. The naive regression estimates (regress Z on X) underestimate the regression slope of .3. The robust regression (least median regression) also underestimates. Scheuren and Winkler's (1993) method (method SW) overestimates the slope. Our unbiased method looks

better. Table 2 presents a numerical summary of bias, sum of squared errors, and mean absolute deviations in the 100 replications for case 1. Scheuren and Winkler's method using all pairs instead of just the best 2 (row SWext) seems to make additional adjustment in these simulation cases.

The regression estimates were computed with the true probabilities used to make the simulated data for comparison. Results using the true probabilities are very similar to the results reported here, because the probabilities estimated by EM are close to the true probabilities.

In fairness to the SW method (1993), it would perform better if not all record pairs, but record pairs with probability of matching above a threshold were used. Additionally, the next stage of research should be to compare Scheuren and Winkler's (1997) iterative method to the unbiased method. It could be considered to be an advantage of the unbiased method that it performs well using all record pairs.

The story in the other cases is basically the same as for case 1. In case 3, the SW method has the smallest bias, followed closely by our unbiased method. Further study is needed to learn why the methods are performing as they are and whether or not the results can be generalized to more realistic situations.

5 Conclusion

Computerized record linkage can introduce errors into the composite file when errors are made in matching records. The mismatch errors can cause problems for analyses of variables brought together from different source files. In the presence of matching errors, naive estimates of linear regression coefficients are biased toward zero because the errors attenuate the relationship between the predictors and response. In simulations, least median regression was not sufficient to guard against matching errors, whereas the method of Scheuren and Winkler (1993) as applied here made too much of an adjustment. Our unbiased method seemed to perform very well across a range of situations.

More work is needed to understand why our method seems to offer some improvement over the other methods. Please note that we have not implemented Scheuren and Winkler's (1997) iterative procedure. Our results are sensitive to estimates of probabilities used in the Fellegi-Sunter (1969) algorithm and the influence should be studied. Future work will involve comparing our method to Scheuren and Winkler's (1997) iterative method. We also plan to produce measures of uncertainty for our estimator and develop methods of regression adjustment that account for the uncertainty in the estimated probabilities. It might also be possible to incorporate the iterative clerical review method of Larsen and Rubin (2001) in the estimation process.

6 References

Alvey, W., and Jamerson, B. (1997), *Record Linkage Techniques – 1997*, Proceedings of an International Workshop and Exposition. Federal Committee on Statistical Methodology, Office of Management of the Budget.

Armstrong, J .B., and Mayda, J. E. (1993), "Model-Based Estimation of Record Linkage Error Rates, " *Survey Methodology*, 19, 137-147.

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False-Match Rates in Record Linkage, " *Journal of the American Statistical Association*, 90, 694-707.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414-420.

— (1995), "Probabilistic Linkage of Large Public Health Data Files, " *Statistics in Medicine*, 14, 491-498.

Lahiri, P., and Larsen, M. D. (2000), *Regression Analysis with Linked Data*, Technical Report.

Larsen, M. D., and Rubin, D. B. (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, Accepted for publication.

Neter, John, Maynes, E. Scott, and Ramanathan, R. (1965), "The effect of mismatching on the measurement of response errors," *JASA*, 60, 1005-1027.

Scheuren, Fritz, and Winkler, William E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58.

Scheuren, Fritz, and Winkler, William E. (1997), "Regression analysis of data files that are computer matched – Part II," *Survey Methodology*, 23, 157-165.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage, " in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 667- 671.

Table 1: Simulation conditions

	Conditions	case 1	case 2	case 3	case 4
m	number of replications	100	100	100	100
n	size of files A and B	1200	1200	1000	1000
β	regression slope	.3	.6	.3	.6
σ	regression SD	.5	.5	.5	.5
k	number of comparison fields	10	10	8	8
p_k	probability of agreement on a field for a match	.75	.75	.75	.75
q_k	probability of agreement on a field for a nonmatch	.20	.20	.25	.25
	size of blocks	15	15	20	20
	quality of matching situation	good	good	soso	soso

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," in *American Statistical Association Proceedings of Survey Research Methods Section*, pp. 1994.

— (1995), "Matching and Record Linkage," in *Business Survey Methods*, ed. Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., New York: Wiley Publications, pp. 355-384.

Table 2: Regression Results case 1

Method	Bias $\sum(\hat{\beta}-.3)/100$	sum of squared errors $\sum(\hat{\beta}-.3)^2$	mean absolute deviation $\sum\lvert\hat{\beta}-.3\rvert/100$
Naive	-.012	.039	.016
SW	.019	.077	.023
SWext	.020	.075	.023
Robust	-.012	.052	.018
Unbiased	-.004	.029	.014

Table 3: Regression Results case 2

Method	Bias $\sum(\hat{\beta}-.6)/100$	sum of squared errors $\sum(\hat{\beta}-.6)^2$	mean absolute deviation $\sum\lvert\hat{\beta}-.6\rvert/100$
Naive	-.023	.072	.024
SW	.030	.121	.030
SWext	.041	.200	.042
Robust	-.013	.045	.018
Unbiased	-.009	.034	.015

Table 4: Regression Results case 3

Method	Bias $\sum(\hat{\beta}-.3)/100$	sum of squared errors $\sum(\hat{\beta}-.3)^2$	mean absolute deviation $\sum\lvert\hat{\beta}-.3\rvert/100$
Naive	-.046	.239	.046
SW	.000	.043	.017
SWext	.019	.078	.023
Robust	-.038	.185	.038
Unbiased	.008	.044	.016

Table 5: Regression Results case 4

Method	Bias $\sum(\hat{\beta}-.6)/100$	sum of squared errors $\sum(\hat{\beta}-.6)^2$	mean absolute deviation $\sum\lvert\hat{\beta}-.6\rvert/100$
Naive	-.056	.335	.056
SW	.024	.102	.027
SWext	.038	.187	.039
Robust	-.034	.153	.035
Unbiased	-.010	.048	.017

Figure 1:

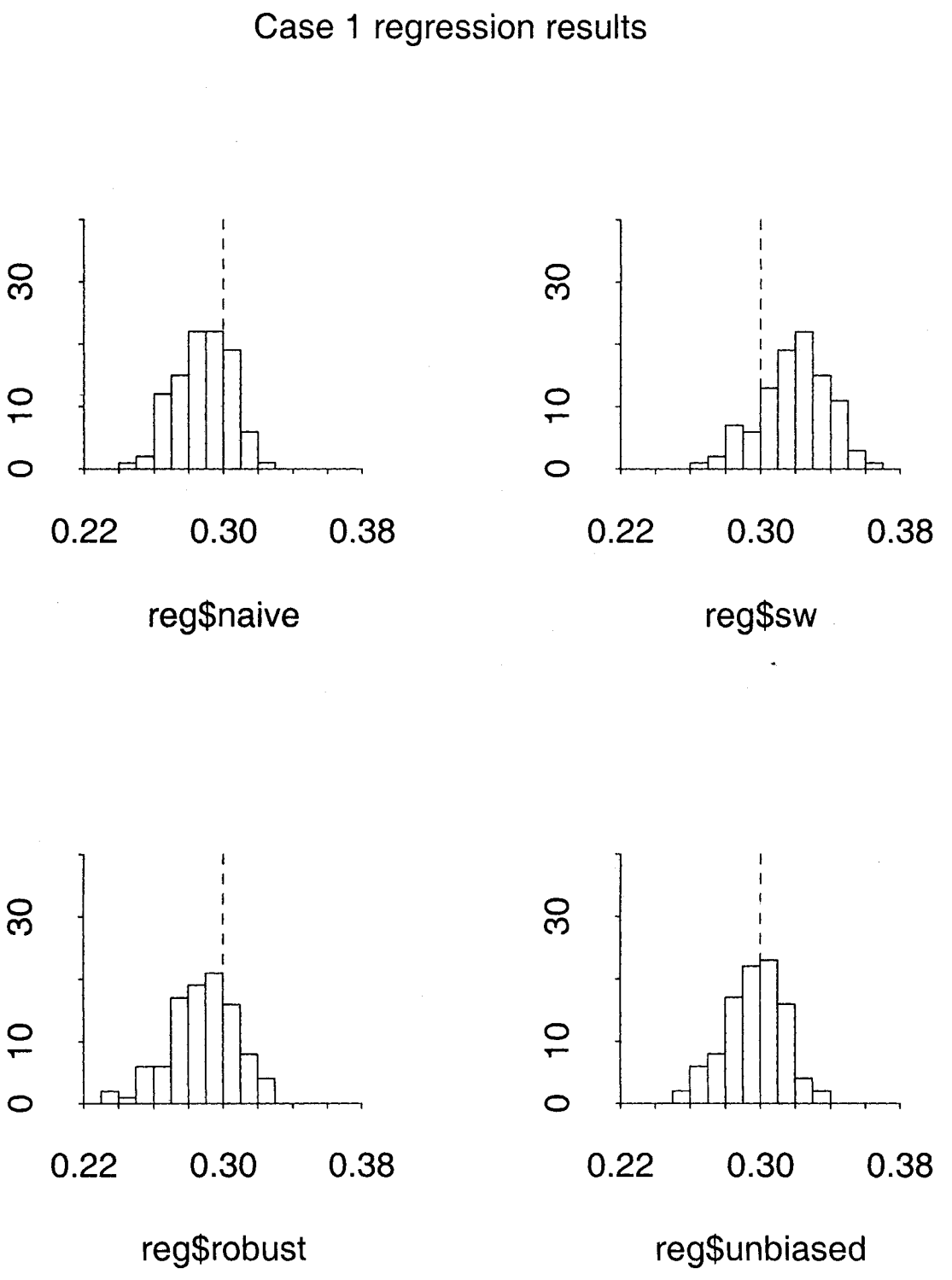

Case 1 regression results
0
10
30
0.22
0.30
0.38
reg$naive
reg$sw
reg$robust
reg$unbiased

Figure 2:

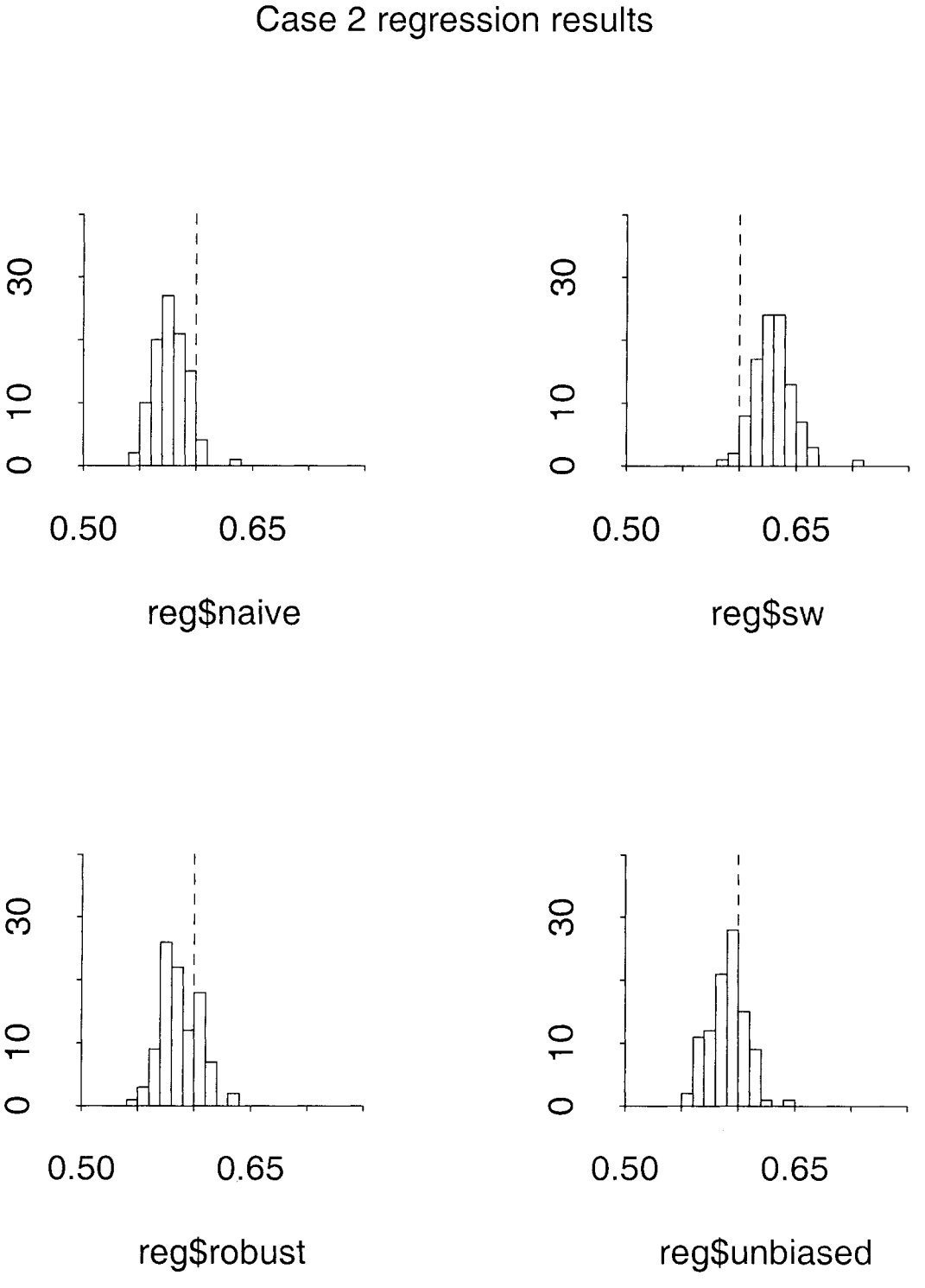

Case 2 regression results
0
10
30
0.50
0.65
reg$naive
reg$sw
reg$robust
reg$unbiased

Figure 3:

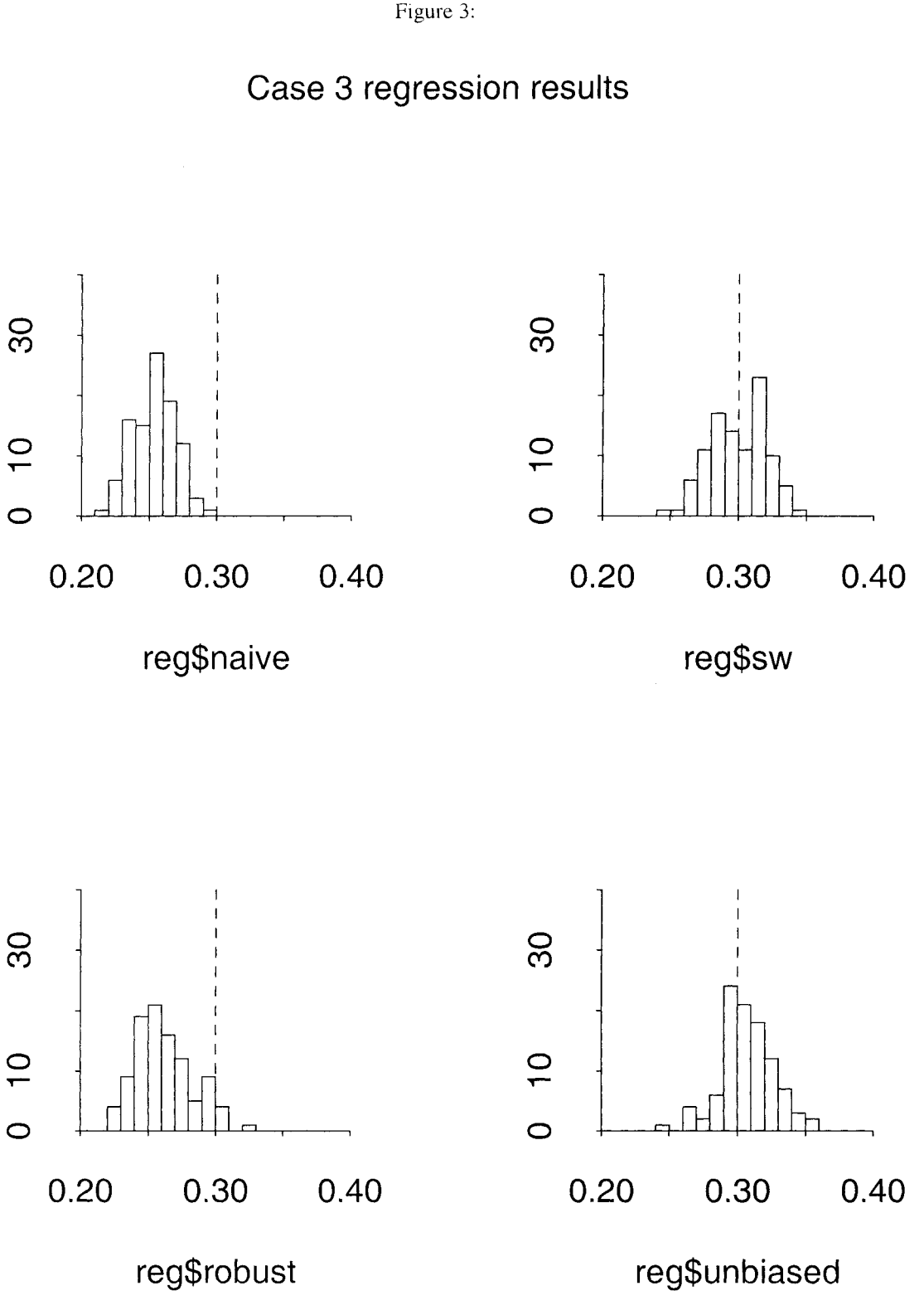
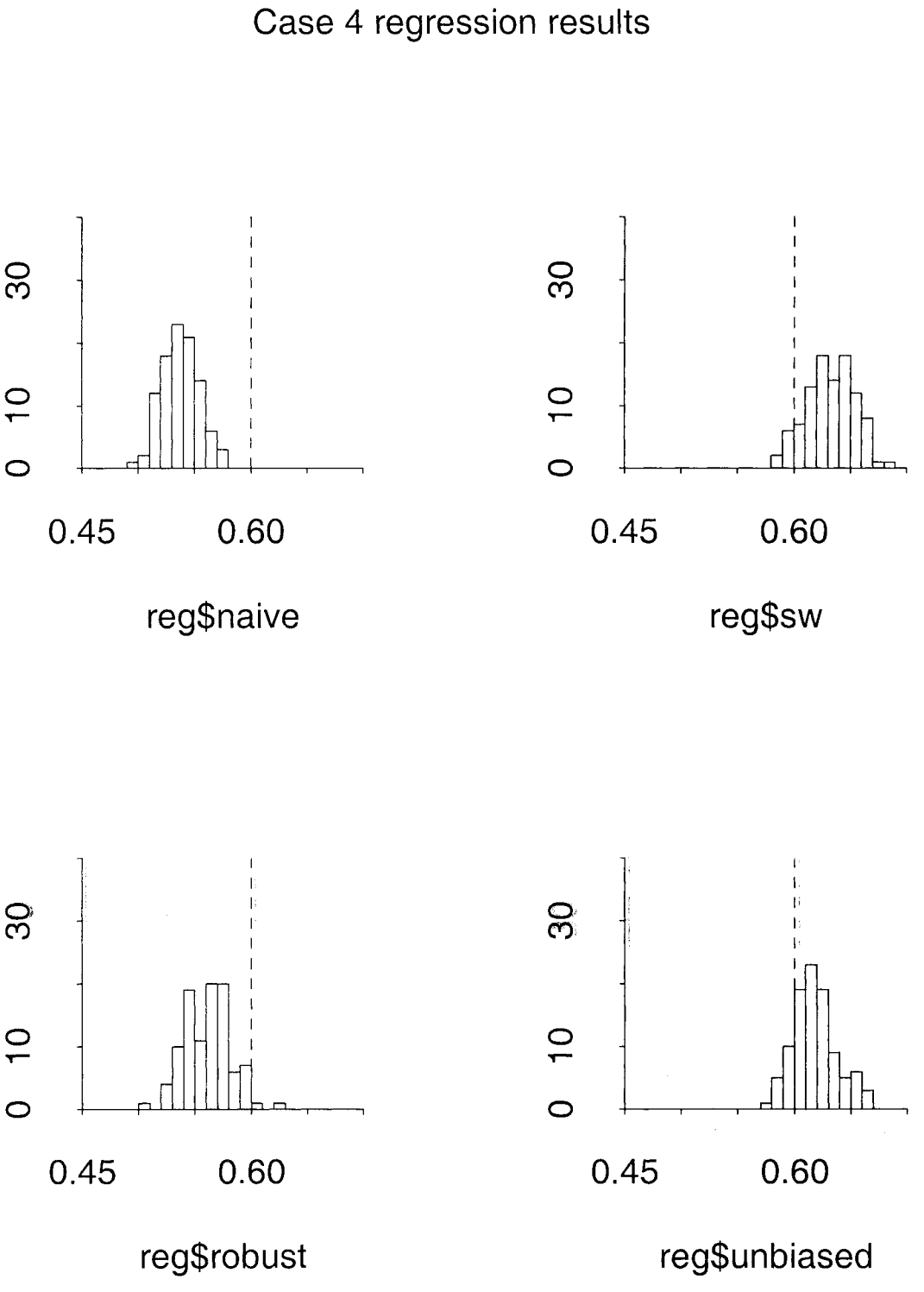

Case 3 regression results
0
10
30
0.20
0.30
0.40
reg$naive
reg$sw
reg$robust
reg$unbiased

Figure 4:


Case 4 regression results
0
10
30
0.45
0.60
reg$naive
reg$sw
reg$robust
reg$unbiased