ON THE FORMATION OF WEIGHTING ADJUSTMENT CELLS FOR UNIT NONRESPONSE

Sonya Vartivarian, Department of Statistics, University of Michigan. E-mail: slvartiv@umich.edu
Roderick J. Little, Department of Biostatistics, University of Michigan. E-mail: rlittle@umich.edu

KEY WORDS: Sampling weights; Survey Inference; Unit Nonresponse Adjustment

1. Introduction

Unit nonresponse occurs when entire interviews are missing due to noncontact of a sampled individual or refusal to answer the questionnaire. Weighting is a standard method of unit nonresponse adjustment and is a natural extension of weighting for unequal probabilities of selection. However, unlike the sample weight, the nonresponse rate is usually unknown and must be estimated.

In forming nonresponse weights, respondents and nonrespondents are often classified into adjustment cells based on covariate information recorded for both groups. Respondents in cell *c* are then weighted by the inverse of the response rate in cell *c*. For example, a cell is defined as "married women living in the south" with 80 respondents and 20 nonrespondents. Then, the response rate is 80/100 = 0.8 and the response weight is 1/0.8 = 1.25.

Let $D = (X,Z)$ be all fully-observed survey variables X and design variables Z, Y be the outcome variable and R be a response indicator. In principle, adjustment cells might be based on a joint classification of the variables D. We consider here situations where this leads to too many cells, so that some cells have no respondents or a small counts of respondents that lead to excessively variable nonresponse weights. This situation is common in the case of attrition in panel surveys, where extensive survey information from earlier waves is available for creating adjustment cells for later waves. In surveys involving differential probabilities of selection, adjustment cells are often formed within which the probability of selection is not constant. The usual weighting adjustment is then proportional to the inverse of the weighted response rate, defined as the sum of the weights for respondents divided by the sum of the weights for respondents and nonrespondents. Simulations in Little and Vartivarian (2002) show that improved inferences are obtained by forming adjustment cells that crossclassify on the survey design variables, rather than incorporating these variables by weighting the rates. However, this strategy may lead to too many adjustment cells to be practical. For example in the Health Interview Survey (Botman et al, 2000), weighted response weights are calculated within second stage sampling unit (SSU), a variable that has many levels. Joint classification by Z and X would correspond to stratifying households within SSU according to race, which would yield many small adjustment cells, including perhaps some with no respondents.

2. Two key dimensions for forming adjustment cells

We consider coarsening strategies based on classification by grouped values of linear combinations of Z. Two linear combinations of Z are particularly useful in this regard: (a) *response propensity* stratification, as defined in Section 2.1, which aims to form cells that are homogeneous with respect to the probability of response, and (b) *predictive mean* stratification, as defined in Section 2.2, which aims to form cells that are homogeneous with respect to the predicted mean of a particular outcome variable Y. Both of these approaches have the important property that if nonresponse is missing at random (MAR; Rubin, 1976; Little and Rubin, 2002), that is

$$R \perp\!\!\!\perp Y \mid D, \qquad (1)$$

where $\perp\!\!\!\perp$ denotes independence, then the same property applies approximately to the coarsened classification. That is, if A is a coarsening of D based on response propensity or predictive mean stratification, then approximately,

$$R \perp\!\!\!\perp Y \mid A, \qquad (2)$$

so adjustment based on A controls nonresponse bias.

2.1 Response Propensity Stratification

The first strategy for reducing the potential number of adjustment cells is response propensity stratification, where Little (1986) defines the *response propensity* as

$$p(D) = pr(R = 1 \mid D), \qquad (3)$$

and supposes that $p(D) > 0$ for all observed values of D. Following Rosenbaum and Rubin's (1983) theory for matching in observational studies, define a *balancing score* as a function b of the observed covariates D such that D is conditionally independent of response given the balancing score $b(D)$,

$$D \perp\!\!\!\perp R \mid b(D). \qquad (4)$$

Rosenbaum and Rubin's (1983) theory shows that (a) the finest balancing score is the full set of covariates, $b(D) = D$; (b) the coarsest balancing score is the propensity score $p(D)$; and (c) if the data are MAR, as in (1), then

$$Y \perp\!\!\!\perp R \mid p(D). \qquad (5)$$

Therefore, adjustment cells based on grouping units according to the propensity to respond adjusts for nonresponse bias if the data are MAR (Little, 1986).

In practice the response propensity is unknown and needs to be modeled, for example via a logistic regression of R on D, yielding estimated propensities $\hat{p}(D)$. One can then form a grouped version of

$\hat{p}(D)$, say $\hat{p}_G(D)$, and use this grouped version of the response propensity as a basis for forming nonresponse adjustment cells; for example one might base the groups on the quintiles of the distribution $\hat{p}(D)$. Since these adjustment cells do not depend on *Y* they are the same for each outcome variable, and only one regression of *R* on *Y* needs to be modeled.

Little (1986) considers properties of weighting class adjustments for *domain* means that do not cut across adjustment cells, and for *cross-class* means that do cut across adjustment cells. For example, the overall mean or the mean outcome for $C = 1$ in Table 1 are examples of domain means, but the mean outcome for homeowners is a cross-class mean since it cuts across the three levels of adjustment cells.

Table 1: Cross-classes of Home Ownership Status

Home Ownership Status	Adjustment Cell Variable *C*		
	$C = 1$	$C = 2$	$C = 3$
Owner			
Renter			

Means based on response propensity stratification have the following properties:

- Weighting using the propensity score stratification yields approximately unbiased estimates of domain and cross-class means, where the approximation arises from estimating and grouping $\hat{p}(D)$.
- Response propensity stratification does not control variance, and may be very inefficient, especially when the covariate set *D* includes variables that have a strong association with *Y*, but $\hat{p}(D)$ has a weak association with *Y*.

2.2 Predictive Mean Stratification

The second strategy for coarsening the set of adjustment cells is *predictive mean* stratification. First, a regression of the outcome *Y* on *D* is fitted to respondents, yielding a predicted mean $\hat{y}(D)$ for each respondent and nonrespondent. Then adjustment cells are based on a grouped version $\hat{y}_G(D)$ of $\hat{y}(D)$; for example, one possible choice is to base the groups on the quintiles of the distribution of $\hat{y}(D)$.

To motivate this form of coarsening, note that the MAR assumption (1) implies that the distribution of *Y* for respondents and nonrespondents are the same given *D*. Pooling over values of *D* with the same distribution of *Y* results in subpopulations where the outcome *Y* and response indicator *R* are still independent. Suppose the distribution of *Y* given *D* differs only in the mean $y^*(D)$ for different values of *D*. Then pooling over adjustment cells with the same value of $y^*(D)$ yields coarsened cells within which *Y* and *R* are independent. That is,

$$Y \perp\!\!\!\perp R \mid y^*(D).$$

Hence if $\hat{y}(D)$ is an estimate of $y^*(D)$ based on a well-specified regression model, then

$$Y \perp\!\!\!\perp R \mid \hat{y}(D), \qquad (6)$$

approximately. The variance of *Y* within adjustment cells is also minimized by classifying on the predicted mean. Little (1986) summarizes the properties of weighted means from predictive mean stratification as follows:

- The bias and variance of the overall mean of the outcome *Y* is approximately controlled, again the approximation arising due to the estimating and grouping of $\hat{y}(D)$.
- The variance is smaller than that obtained with response propensity stratification, since predictive mean stratification minimizes within cell variation.
- Weighted means for cross-classes have potentially nonzero large sample bias (LSB).

Thus predictive mean stratification gives better estimates of domains means than response propensity stratification, since it controls both bias and variance, but unlike response propensity stratification, it does not yield unbiased estimates of cross-class means. Another drawback with predictive mean stratification is that it leads to different choices of adjustment cells for each survey outcome *Y*. A single set of adjustment cells that is relatively efficient for a set of key outcomes would be desirable (Little, 1986; Goksel, Judkins and Mosher, 1992). One possible approach is to use a principal component analysis or some other form of factor analysis to reduce the number of outcome variables used in forming adjustment cells, thus reducing the number of regressions and sets of weights necessary.

2.3 Joint Classification by Response Propensity and Predictive Mean Stratification

Since response propensity and predictive mean stratification have attractive features, we propose to cross-classify on both the response propensity scores $\hat{p}(D)$ and the best linear predictor $\hat{y}(D)$ to form adjustment cells. The motivation is to capture the bias-reduction property of response propensity stratification and the gains in efficiency of predictive mean stratification. The joint classification also has potential gains in robustness if the model for the response propensity or predictive mean is misspecified, as discussed in section 2.3.2.

2.3.1 Gains in Efficiency from Crossclassifying by $\hat{y}(D)$ as well as $\hat{p}(D)$

Consider two aspects that contribute to the inefficiency of forming weighting classes based on

$\hat{p}(D)$: (i) the R^2 from the regression of Y on D; (ii) the correlation between $\hat{p}(D)$ and $\hat{y}(D)$, say ρ. The asterisked cell in Table 2 corresponds to the situation where adding $\hat{y}(D)$ to $\hat{p}(D)$ improves the efficiency of $\hat{p}(D)$.

Table 2. Efficiency of the Response Propensity; **The case where $\hat{p}(D)$ is inefficient.*

		$R^2(Y,D)$ Low	$R^2(Y,D)$ High
$\rho(\hat{p}(D), \hat{y}(D))$	Low		*
	High		

Adding $\hat{y}(D)$ to $\hat{p}(D)$ improves the efficiency when $\hat{y}(D)$ is a good predictor of Y and $\hat{p}(D)$ and $\hat{y}(D)$ are weakly correlated. (If $\hat{p}(D)$ and $\hat{y}(D)$ are highly correlated, then the joint crossclassifying will have sparse off-diagonal cells leading to increased variance.)

2.3.2 Reduction in Bias from Crossclassifying by $\hat{p}(D)$ as well as $\hat{y}(D)$

If $\hat{p}(D)$ is misspecified, then further crossclassification on $\hat{y}(D)$ will control bias for the overall mean. On the other hand, if $\hat{y}(D)$ is misspecified, then stratifying on $\hat{y}(D)$ alone may lead to bias, which can be reduced by further classification on $\hat{p}(D)$. The response propensity may be misspecified because important predictors are omitted from the logistic regression of R on D, or the form of this regression is incorrect. The predicted mean $\hat{y}(D)$ may be misspecified for a number of reasons:

(1) Mismodeling of $\hat{y}(D)$ itself (e.g., omitting interaction terms);
(2) With multiple outcomes, it is not practical to create separate weights for each outcome. Choosing a single compromise predictive mean stratification for all outcomes entails a misspecification error for each individual outcome.
(3) Even if $\hat{y}(D)$ is correctly specified, the bias of the cross-class mean is not controlled by predictive mean stratification, so adding $\hat{p}(D)$ reduces the bias of weighted estimates of cross-class means.

Crossclassifying by $\hat{p}(D)$ and $\hat{y}(D)$ has the following "double robustness" property:

(a) If $\hat{p}(D)$ is correctly specified and $\hat{y}(D)$ is incorrectly specified, joint classification controls bias of estimates of the mean for the whole sample and for cross-classes;
(b) If $\hat{p}(D)$ is incorrectly specified and $\hat{y}(D)$ is correctly specified, then joint classification controls the bias of the overall mean and leads to gains in efficiency.

Similar "double robustness" properties were discussed by Robins et al (2000) in the context of estimating equations, and by Zeng (2001) in the context of survival analysis. In conclusion, a joint classification on $\hat{p}(D)$ and $\hat{y}(D)$ has the potential of yielding greater robustness to model misspecification, improved bias reduction, and improved efficiency. In the next section we assess the empirical validity of these theoretical properties by a simulation study.

3. Simulation I

A finite population of size N = 10,000 was generated. Four covariates were generated such that $[D1, D2, D3, D4] \sim N_4(0, I)$, where N_4 is a multivariate normal and I is the identity matrix. A stratifier Z and a cross-class variable C were generated, with Z and C each based on dichotomized independent standard normal variates. The outcome is as follows:

$$Y1 = N + C + Z + 0.5 * \varepsilon_Y,$$

where N and ε_Y are independent standard normal variates, the correlation r(Y1,D1)=0.7 and $Y1 \amalg [D2, D3, D4]$. The probit response probability depends on covariates $[D1, D2, D3, D4]$ and a standard normal variate ε_R such that

$$P(R = 1 | D1, D2, D3, D4) =$$
$$\Phi\{0.2 + 0.1 * D1 + 0.6 * (D2 + D3 + D4) + 0.5 * \varepsilon_R\}.$$

The resulting response rate is approximately 55%. One hundred replicate stratified random samples of size n = 2200 were taken from this population.

The quantities of interest are the root mean square error (RMSE), the average bias (AB) over replicates relative to the mean before deletion of cases due to nonresponse and the RMSE relative to the correct model for the response propensity (RMSErpF), defined as RMSErpF = 100*((RMSE/RMSEpF)-1)).

3.1 Modeling the Response Propensity and Predictive Mean

The response propensity probit model includes the cross-class variable C and stratifier Z. In addition, if the model includes D1, then the model is denoted pD1. If the model includes D1 and D2, then the model is referred to as pD12, and so on. If the model includes $D1$,$D2$,$D3$ and $D4$, the model is referred to as the full model, pF. A summary of all five models examined is listed in Table 3.

Table 3. Models for the Response Propensity

Model	Mean Classification	Covariates Included in Model
1.	pD1	$D1, C, Z$
2.	pD12	$D1, D2, C, Z$
3.	pD34	$D3, D4, C, Z$
4.	pD234	$D2, D3, D4, C, Z$
5.	pF	$D1, D2, D3, D4, C, Z$

For example, the model pF is a probit regression of R on *D*1, *D*2, *D*3, *D*4, *C* and *Z*.

Similarly, the five predictive mean models for Y1 are as in Table 3, but with p replaced by y1. For example, the model y1F is a multiple regression of Y1 on *D*1, *D*2, *D*3, *D*4, *C* and *Z* fit to the respondent data. A grouped version of the predicted values for all models of the response propensity and predictive mean form the five adjustment cells, where groups are based on the quintiles of the distribution. The response rate in an adjustment cell is then the number of respondents in a cell divided by the number of sampled individuals in that cell.

The only correct model for the response propensity is pF, the model that includes all covariates. The correct models for the predictive mean are y1D1, y1D12 and y1F.

3.2 Simulation I Results

Since gains in efficiency by further classifying on the predictive mean after classifying on the response propensity are of interest, we examine the RMSErpF, the root mean square error of models relative to the root mean square error of classifying on the correct model for the response propensity pF. Figure 1 contains the RMSErpF and AB for all jointly classified adjustment cells, broken down into five cases: (1) the case where both the response propensity and the predictive mean are correctly modeled; (2) the case where only the predictive mean is correctly modeled; (3) the case where only the response propensity is correctly modeled; (4) the case where both the predictive mean and response propensity are incorrectly modeled; (5) the last case that includes both the mean before deletion of cases due to nonrespone (meanbd) and the respondent mean (meanr) .

The mean before deletion of cases due to nonresponse (meanbd) performs 35% better with respect to the RMSErpF than the mean based on the model pF. When the predictive mean is correctly modeled (cases (1) and (2)), we see that the all estimates are more efficient than the estimate based on pF (i.e., RMSErpF < 0). In fact, when the predictive mean is correctly modeled, using the joint classification of the predictive mean in addition to the response propensity brings us between one-third to two-thirds of the efficiency achieved by using meanbd, the mean based on data before nonresponse. Another interesting feature apparent in Figure 1 is the "double robustness" property. The robustness to model misspecification that is enjoyed by the joint classification of the predictive mean and the response propensity is apparent. Specifically, correctly modeling at least one of the two models allows for a sensible estimate, unlike in case (4), where the RMSErpF and AB are unacceptably high. Correctly modeling both models allows for gains in efficiency in addition to unbiasdness.

4. Simulation II

The simulation results presented thus far are based on one population, so general conclusions are not warranted. In order to explore the issues of unbiasedness, efficiency and "double robustness" more systematically, the population characteristics were varied. The outcome is of similar form to that in Simulation I,

$$Y1 = N + C + Z + 0.5 * \varepsilon_Y ,$$

where $Y1 \amalg [D2, D3, D4]$, but the correlation between *Y1* and *D1* is either moderate or low (i.e., r(*Y*1,*D*1)=0.68 or r(*Y*1,*D*1)=0.38, respectively).

The response propensity model is changed to

$$P(R = 1 \mid D1, D2, D3, D4) =$$
$$\Phi\{0.2 + \beta_1 * D1 + \beta_2 * (D2 + D3 + D4) + 0.3 * C + 0.5 * \varepsilon_R\},$$

where the coefficients β_1 and β_2 are as in Table 4.

Table 4. Simulation II Response Probability Coefficients

	β_1	β_2
1.	0.6	0.6
2.	0.2	0.6
3.	0.6	0.2

Thus, there are a total of 2*3=6 populations representing the 2 outcome structures and the 3 nonresponse structures. One hundred replicate stratified random samples of size n=2200 were drawn from each of the six populations.

4.1 Simulation II Results

Figure 2 contains the RMSE for each joint classification of the response propensity and predictive mean, averaged over all six populations. The average RMSE for the classification based on pF alone is represented with a red line, RMSE=365. When the predictive mean is correctly modeled, the method of joint classification by the predictive mean and the response propensity in general improves the average RMSE relative to using only the correct model for the response propensity pF. Simulation II further supports conclusions based on Simulation I. Notice that the average RMSE is not reported for

cases where both models are incorrect or for the respondent mean as the average RMSE are unacceptably high, ranging from 563 to 1706, again demonstrating robustness of the joint classification to model misspecification when at least one of the two models is correct.

5. Cross-Class Means

The cross-class means for each of the six populations in Simulation II are calculated using the same weights obtained for the overall mean, but applied to the cross-classes of C (i.e., $C = 0$ or $C = 1$). The sample weighted cross-class mean before deletion of nonrespondents is used as a standard. Absolute total cross-class bias (ATOTCB) is defined to be the sum of the absolute bias in the cross-classes relative to the mean before deletion, and the standard deviation of the absolute total class bias (SDTOTCB) is the standard deviation of ATOTCB over replicates.

The "double robustness" is again evident when examining the average ATOTCB over the six populations, where a reasonable ATOTCB is obtained when at least one of the two models is correct as shown in Figure 3.

Figure 4 displays that using the crossclassification of the predictive mean and the response propensity, averaged over the six populations, results in lower ATOTCB for 10 of the 17 cases when one of the two models is correct; also, the model y1D234pF results in only two units larger ATOTCB than when using only pF.

6. Summary

This research proposes to construct unit nonresponse adjustments based on the joint classification of $\hat{p}(D)$ and $\hat{y}(D)$. Simulations suggest that an improvement in efficiency is gained in situations where the response propensity is inefficient, with negligible loss in efficiency when the response propensity is efficient. Further, our simulations demonstrate robustness of the joint classification to misspecification of the model for the response propensity or the predictive mean.

Our research focuses on the simple situation of a single outcome *Y*, where predictive mean stratification yields a one-dimensional classification variable. In real surveys with multiple key outcomes, the method proposed here would lead to a different set of weights for each outcome, which is practically cumbersome and leads to complications for multivariate analysis. Thus, in future work we plan to explore our method in conjunction with dimension reduction of a set of outcomes, using techniques such as principal component analysis.

ACKNOWLEDGEMENTS

This research was supported by grants SES-0106914 from the National Science Foundation and UR6/CCU517481 from the Centers for Disease Control. We appreciate helpful comments from Trivellore Raghunathan.

REFERENCES

1. Botman SL, Moore TF, Moriarty CL, Parsons VL. Design and Estimation of the National Health Interview Survey, 1995-2004. *National Center for Health Statistics, Vital Health Statistics* 2000; 2, 130.
2. Cochran WG. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 1968; 24: 295-313.
3. Goskel H, Judkins D, and Mosher, D. Nonresponse Adjustments for a Telephone Follow-up to a National In-Person Survey. *Journal of Official Statistics*, 1992; 4: 417-431.
4. Little RJ. Survey nonresponse adjustments. *International Statistical Review*, 1986; 54: 139-157.
5. Little RJ and Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Ann. Rev. Pub. Health*, 2000; 21:121-45.
6. Little RJ and Rubin DB. *Statistical Analysis with Missing Data*, 2nd Edition, John Wiley and Sons, Inc: New York, 2002.
7. Little RJ and Vartivarian S. To appear in the *Special Issue, Statistics in Medicine* on papers from the CDC/ATSDR Symposium on Statistical Methods, 2001.
8. Robins JM, Rotnitzky A and van der Laan M. Comment on "On Profile Likelihood" by Murphy S and van der Vaart A. W. *Journal of the American Statistical Association*; 2000; 95: 477-482.
9. Rosenbaum PR and Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*; 1983; 70: 41-55.
10. Rubin DB. Inference and missing data. *Biometrika*, 1976; 63, 3: 581-92.
11. Zeng D. Adjusting for Dependent Censoring using High-Dimensional Auxiliary Covariates. University of Michigan Press: Ann Arbor, MI, 200

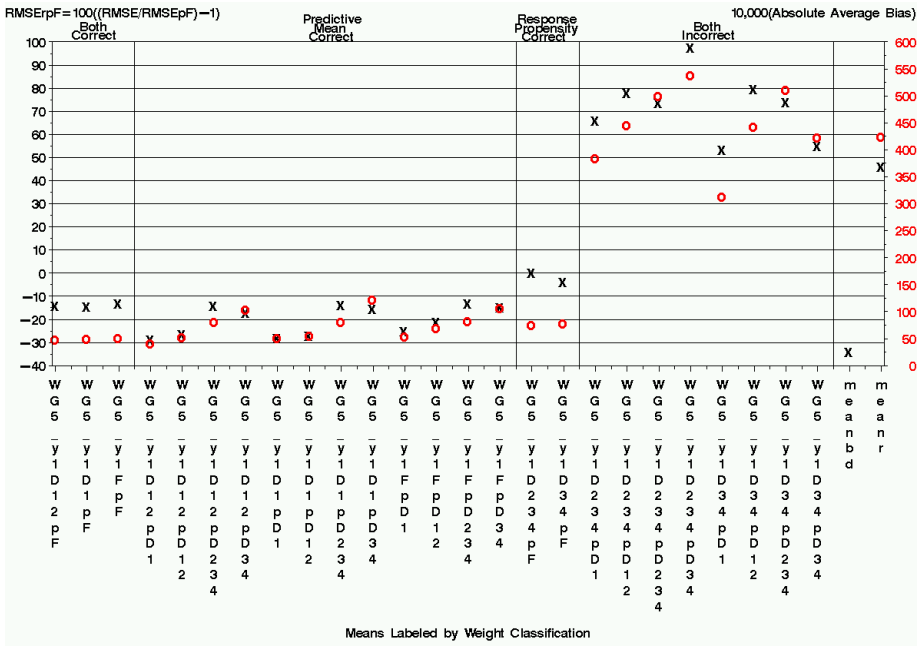

Figure 1: RMSErpF and AB by Model Correctness; Simulation I

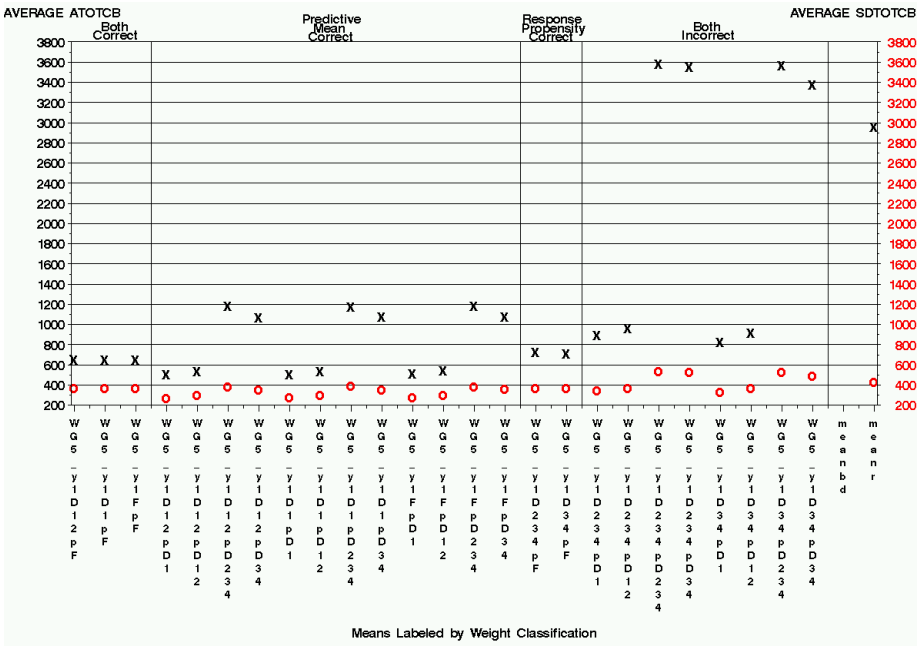

Figure 3. Average ATOTCB and SDTOTCB over 6 Populations for Cross-class Estimates; Simulation II

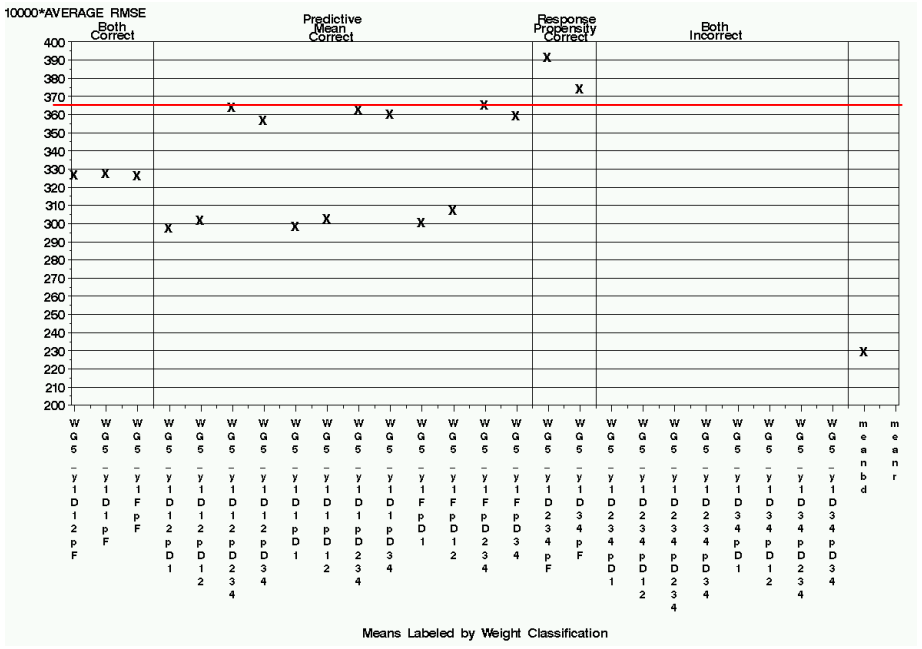

Figure 2. Average RMSE of Estimates over 6 Populations; Simulation II; —— Average RMSE = 365 for pF

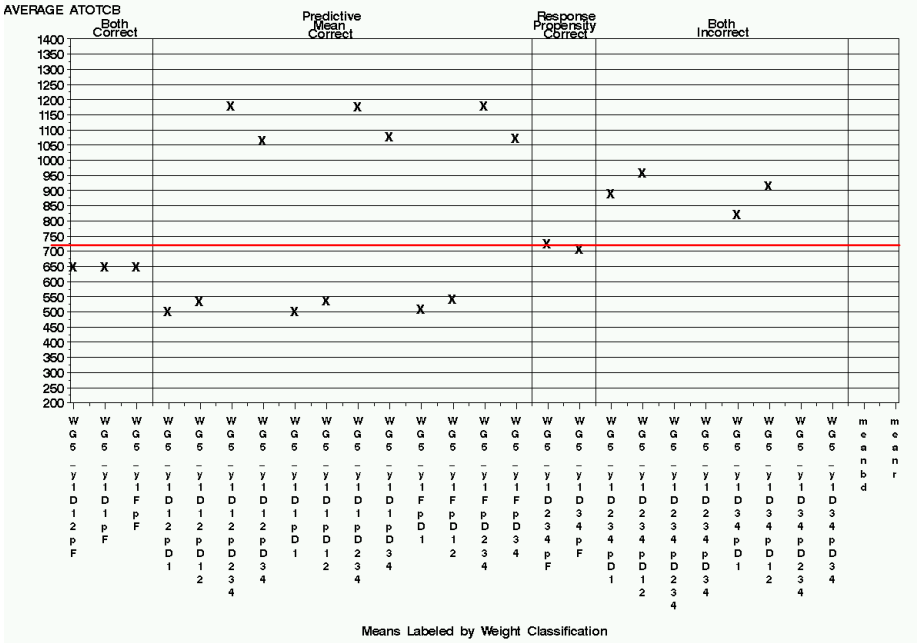

Figure 4. Average ATOTCB < 1400 over 6 Populations for Cross-class Estimates; Simulation II; —— Average ATOTCB = 724 for pF