2011

Conditional and unconditional models in model-assisted estimation of finite population totals

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Conditional and unconditional models in model-assisted estimation of finite population totals

Abstract

The well known Godambe-Joshi lower bound for the anticipated variance of design unbiased estimators of population totals treats the auxiliary variables as constants. We extend the result to models where these variables are random and show that the generalized difference estimator using the expected values conditional on all auxiliary values is optimal. This has several implications including the fact that collecting multiple survey variables does not reduce the lower bound.

Disciplines

Physical Sciences and Mathematics

Publication Details

Steel, D. G. & Clark, R. Graham. (2011). Conditional and unconditional models in model-assisted estimation of finite population totals. Pakistan Journal of Statistics, 27 (4), 529-541.

Pak. J. Statist.
2011 Vol. 27(4), 529-541

CONDITIONAL AND UNCONDITIONAL MODELS IN MODEL-ASSISTED ESTIMATION OF FINITE POPULATION TOTALS

David G. Steel and Robert G. Clark
Centre for Statistical and Survey Methodology
University of Wollongong Australia

ABSTRACT

The well known Godambe-Joshi lower bound for the anticipated variance of design unbiased estimators of population totals treats the auxiliary variables as constants. We extend the result to models where these variables are random and show that the generalized difference estimator using the expected values conditional on all auxiliary values is optimal. This has several implications including the fact that collecting multiple survey variables does not reduce the lower bound.

KEY WORDS

Cluster sampling; Generalized regression estimator; Optimal estimation; Probability sampling

AMS Classification 62D05

1. INTRODUCTION

1.1 Inferential Frameworks for Survey Sampling

This article extends a well known result relating to the best possible estimator of a finite population total according to a particular criteria, the anticipated variance. Consider a population U consisting of N units, where units may be people, households, businesses or other entities. A sample $s \subseteq U$ is selected from U using some probability sampling method, with $\pi_i = P[i \in s]$ denoting the probability of selection for unit $i \in U$. A variable of interest Y_i is defined for each unit $i \in U$. The aim is to estimate $Y = \sum_{i \in U} Y_i$ using data on Y_i for sampled units $i \in s$ only. We denote the vector of Y_i for $i \in s$ by Y_s. A vector of auxiliary variables X_i may also be available and observed for all units $i \in U$. We write X_U and Y_U to denote the collection of all of these variables for all units in the population.

For example, the aim may be to measure the total employment for working age adults in a country. In this case, the units i would be people and the population U would be all working age adults. The variable Y_i would be defined to equal 1 if person i is employed and 0 otherwise. A survey would be administered to determine the employment status of all people in a sample s. There is often external information on age and sex for all people in the population, so the elements of X_i would be indicator variables indicating the age and sex of person i. The objective is to combine information on Y_i for $i \in s$ with X_U to give the best estimator of Y.

A number of frameworks have been used in sample survey theory, including the design-based, model-based and model-assisted frameworks. In the design-based framework, X_U and Y_U are regarded as constants, and the only source of randomness is the selection of the sample s. This framework avoids any explicit modelling of X_U and Y_U. The "design expectation" of an estimator is defined to be its expected value over all possible samples, with X_U and Y_U treated as constants. Estimators of Y are required to be exactly or approximately "design-unbiased", that is their design expectation (expectation over all possible samples) must be exactly or approximately equal to Y. For a discussion of the design-based framework, see for example Cochran (1977) and Sarndal, Swensson, and Wretman (1992, ch.1-2). Design-variances are similarly defined to be the variances over the possible values of s, with X_U and Y_U treated as constants.

In the model-based, or prediction framework, a model is postulated for Y_U conditional on X_U. Estimators are usually evaluated in terms of the expectation conditional on X_U and on the particular sample s selected, so that the only sources of randomness are the population values Y_U. Expectations conditional on s and X_U are called model-expectations; model-variances are defined similarly. Predictors of Y are generally required to be model-unbiased, in the sense that the model expectation of the estimator minus Y should equal zero, and to have low model variance. For a discussion of the model-based framework, see for example Brewer (1963), Royall (1970) and Valliant, Dorfman, and Royall (2000).

In the model-assisted approach to estimating a finite population total, both probability sampling methods and population models have a role (Sarndal et al., 1992, pp.227,238-239). Both s and Y_U are regarded as random, with a model of some kind being assumed for Y_U. X_U are regarded as constants. Estimators are required to be (at least approximately) design unbiased, that is unbiased over repeated sampling conditional on Y_U and X_U. Subject to this constraint, it is desirable to reduce the variance over both repeated sampling and repeated realisations of Y_U based on an assumed model. The most commonly used estimator in the model-assisted framework is the generalized regression estimator. Most estimators of population totals used in practice are special cases of the generalized regression estimator. Sarndal et al. (1992) contains an extensive discussion of this estimator and the model-assisted framework.

1.2 Optimal Estimation

The existence of optimal estimators has been investigated for all of these frameworks. In the design-based framework, an optimal estimator of Y should ideally minimize the design-variance in the class of all design-unbiased estimators. However, it has been shown that there is no such estimator (Godambe, 1955). One proof of this result (Basu, 1971) shows that for every population, there is in theory a design-unbiased estimator of a total which has zero variance for that particular population, but there is no estimator which has the lowest variance for all populations. (As noted in (Basu, 1971), this result is not usable in practice because it requires perfect knowledge of the population values, but it nevertheless provides a useful benchmark.) The design-based framework is therefore not suitable to give guidance on best estimators without restricting the range of populations being considered. This can be done informally, for example ratio estimators have lower design variance than the simpler number-raised estimator for populations satisfying a simple condition, for simple random sampling (Cochran, 1977, p.157).

The non-existence of an optimal design-based estimator of Y was one of the motivations for the development of the model-assisted framework. In this framework, estimators are still required to be exactly or approximately design-unbiased. Unlike the design-based framework, however, the aim is to minimize the model expectation of the design variance, under an assumed model for Y_U. This quantity, which has been called the anticipated variance (Isaki & Fuller, 1982), can be minimized for design-unbiased estimators (Godambe & Joshi, 1965). An estimator with this optimality property would be design-unbiased regardless of the correctness of the population model, and would be optimal in the sense of minimising the anticipated variance provided the model is true.

Consider the following model for Y_U:

$$\left.\begin{array}{l} E\left[Y_i\right] = \mu_i \\ var\left[Y_i\right] = \sigma_i^2 \\ Y_i, Y_j \text{ independent } \forall i \neq j \end{array}\right\} \tag{1}$$

Design-unbiased estimators $\hat{Y}$ satisfy the inequality

$$var\left[\hat{Y} - Y\right] \geq \sum_{i \in U} \left(\pi_i^{-1} - 1\right) \sigma_i^2 \tag{2}$$

where variance is over both repeated sampling and repeated realisations of Y_U (Godambe & Joshi, 1965). The estimator which meets the lower bound is

$$\hat{Y}_\mu = \sum_{i \in U} \mu_i + \sum_{i \in s} \pi_i^{-1} \left(Y_i - \mu_i\right). \tag{3}$$

The estimator $\hat{Y}_\mu$ generally cannot be calculated in practice, because the values of μ_i and their sum over the population would be unknown in almost all cases. The usual strategy in practice is to assume that μ_i are known functions of X_U and some unknown parameters. These parameters can then be estimated from sample data, and estimates $\hat{\mu}_i$ can then be used in place of μ_i in (3). Estimators formed in this way may have variance which tends to the lower bound (2) as the sample size and population size both tend to infinity, given a correct model for $\{\mu_i\}$. For example, the generalized regression estimator has this property (Sarndal, 1980). Estimators with this property can be called asymptotically optimal.

This paper extends result (2) to the case where the auxiliary variables X_U are random variables, for certain types of designs. Section 2 contains a theorem for this case. The theorem allows the possibility that multiple survey variables are collected and also allows a model with dependencies between different values of the survey variable, subject to a restriction. A difference estimator similar to (3) is shown to be optimal, but with μ_i replaced by $E\left[Y_i \,|X_U\,\right]$. Hence the apparently arbitrary treatment of X_U as constants by Godambe and Joshi (1965) is reasonable, as the same estimator is optimal under a more general model. Section 3 gives four implications of this result and Section 4 summarises the conclusions.

2. A THEOREM ON THE AV UNDER AN UNCONDITIONAL MODEL

This section contains a theorem giving a lower bound for the anticipated variance (AV) of design-unbiased estimators of Y, and an optimal estimator of Y. The theorem allows for the possibility that other survey variables, Z_i, are observed for $i \in s$.

The theorem requires that the sampling scheme is non-informative. A sampling scheme is said to be non-informative if

$$p(s) = P\left[s \text{ is selected} \,|\, Y_U, Z_U, X_U\right] = P\left[s \text{ is selected} \,|\, X_U\right],$$

i.e. the sampling process depends only on X_U, which is known at the time of design, and not on the survey variables Y_U and Z_U which are collected during the survey.

The population X_U, Y_U, Z_U will be assumed to be generated by a model such that

$$\left.\begin{array}{l} E\left[Y_i \,|\, X_U\right] = \mu_i = \mu_i\left(X_U\right) \\ var\left[Y_i \,|\, X_U\right] = \sigma_i^2 = \sigma_i^2\left(X_U\right) \end{array}\right\} \tag{4}$$

Note that the expectations in (4) are model-expectations. This model is extremely general because it allows the means and variances of Y_i conditional on X_U to be any functions of X_U. In practice, $\{\mu_i\}$ would be assumed to be known functions of X_U and some unknown parameters. The unknown parameters would need to be estimated from sample data, so that the lower bound in the Theorem could not be perfectly achieved in practice.

Theorem 1

Suppose that a variable Y_i and a number of other variables Z_i are observed for $i \in s$. It is assumed that sampling is non-informative. The population X_U, Y_U, Z_U is assumed to be generated by model (4). It is further assumed that (Y_i, Z_i) is conditionally independent of (Y_j, Z_j) given X_U for all $i \neq j$ such that there is at least one sample $s \in S$ with $p(s) > 0$ where $i \in s$ and $j \notin s$.

Let $\hat{Y} = \hat{Y}\left(s, X_U, D_s\right)$ be a design unbiased estimator of Y so that

$$E\left[\hat{Y} \,|\, X_U, Y_U, Z_U\right] = Y$$

where D_s denotes the sample data (Y_s, Z_s). Define $\hat{Y}_\mu$ by

$$\hat{Y}_\mu = \sum_{i \in s} \pi_i^{-1}\left[Y_i - \mu_i\left(X_U\right)\right] + \sum_{i \in U} \mu_i\left(X_U\right) \tag{5}$$

Then

$$var\left[\hat{Y} - Y\right] \geq var\left[\hat{Y}_\mu - Y\right]. \tag{6}$$

If Y_i and Y_j are conditionally independent given X_U for all $i \neq j$, then

$$var\left[\hat{Y}_\mu - Y \,\middle|\, X_U\right] = \sum_{i \in U}\left(\pi_i^{-1} - 1\right)\sigma_i^2\left(X_U\right) \tag{7}$$

$$var\left[\hat{Y}_\mu - Y\right] = E\left[\sum_{i \in U}\left(\pi_i^{-1} - 1\right)\sigma_i^2\left(X_U\right)\right]. \tag{8}$$

See Appendix for proof. The proof is similar to Godambe and Joshi (1965) except that the conditioning on X_U has been explicitly stated. With X_U a random variable, we have obtained essentially the same result as Godambe and Joshi (1965), provided that $\hat{Y} = \hat{Y}(s, X_U, D_s)$.

The condition regarding independence of (Y_i, Z_i) and (Y_j, Z_j) allows for these to be dependent for some $i \neq j$ provided that there is never a dependency between values in the sample and values not in the sample. This is is a generalization of the condition in Godambe and Joshi (1965), where Y_i was assumed to be independent of Y_j for every $i \neq j$. The generalization allows the result to be applied to one-stage cluster sampling where there are dependencies between (Y_i, Z_i) and (Y_j, Z_j) for i and j in the same cluster, provided that observations from different clusters are independent. The result does not cover multistage sampling, because in this case there may be correlations between the values of the sampled and unsampled units from each primary sampling unit. However, it would cover most other designs, including stratified single-stage sampling.

There is no requirement in Theorem 1 or in Godambe and Joshi (1965) for the estimator $\hat{Y}$ to be linear.

3. IMPLICATIONS OF THE THEOREM

No Value in Modelling X_U

The values X_U may follow some complex model, for example there may be interesting dependencies between the elements of X_i, or between the elements of X_i and X_j. The theorem shows that $\hat{Y}_\mu$ is the optimal estimator regardless of the distribution of X_U. The values of $E[Y_i|X_U]$ are required to calculate $\hat{Y}_\mu$, and generally modelling of Y_i conditional on X_U is needed to estimate $E[Y_i|X_U]$. There is no improvement in the lower bound on the AV of estimators of Y from considering the marginal distribution. Hence no benefit from knowing this marginal distribution would be expected, at least for sufficiently large samples.

It should be noted that the Theorem states a lower bound which is generally only attained asymptotically. In practice, the model for $E[Y_i|X_U]$ must be estimated from sample data, so that for small samples there could be some benefit from considering the marginal distribution of X_U.

Multiple Survey Variables

It is clear from Theorem 1 that considering several survey variables Y_i and Z_i together does not reduce the lower bound for estimates of Y. The same AV can be achieved by modelling one at a time each variable collected in the survey. The practical implication is that there is no large sample reduction in the AV if multivariate models for several survey variables conditional on X_U are used. Again, it should be noted that the lower bound is generally only attained asymptotically, so it is possible that multivariate models could give some benefit for small samples.

$\hat{Y}_\mu$ is Still Optimal under Cluster Sampling

In one-stage cluster sampling, a sample of clusters of units is selected, and all units in these clusters are selected. In this case, it would be usual to assume that there may be dependencies between the values of the survey variables for units in the same cluster,

but no dependencies across clusters. In this situation, Theorem 1 shows that $\hat{Y}_\mu$ is still the optimal estimator. Hence modelling the dependencies within the cluster cannot give any improvement in estimators of Y, at least for large samples. One practical example of one-stage cluster sampling might be all people in selected households, where households are selected by telephone sampling.

In multi-stage sampling, a sample of first stage units (e.g. suburbs) is selected, followed by a sample of second stage units (e.g. households) from selected first stage units, followed by a sample of third stage units (e.g. from selected second stage units, and so on. One-stage cluster sampling is a simple special case of multi-stage sampling, but in practice more complex multi-stage samples are often used. Theorem 1 would not generally apply in this case, because there could be correlations between sampled and non-sampled units. In some multi-stage household surveys, the final stage of selection consists of all people in selected households. In this case, the correlations across households may be negligible compared to the correlations between people in the same household, so that Theorem 1 may apply at least approximately.

Mukerjee and Sengupta (1989) derived the optimal *linear* model-assisted estimator for correlated data, which could potentially be applied to multi-stage samples in general.

Auxiliary Data for All Units Should Be Used

The lower bounds in Theorem 1 are achieved by the estimator $\hat{Y}_\mu$ in (5) with

$$\mu_i = E\left[Y_i \,|X_U\right]$$

which conditions on the values of the auxiliary variables for all units in the population. This may be different from $E\left[Y_i \,|X_i\right]$ which is widely used in motivating estimators, for example the generalized regression estimator (Sarndal, 1980) and the nonlinear estimators of Firth and Bennett (Firth & Bennett, 1998) and Lehtonen and Veijanen (Lehtonen & Veijanen, 1998). In other words, it is usually assumed that the expectation of Y_i conditional on X_U does not depend on X_j for $j \neq i$. This may not be reasonable for populations with natural clustering or a hierarchical structure (Steel & Welsh, 2006).

Cluster sampling (or stratified cluster sampling) of all people in selected households may be one case where

$$E\left[Y_i \,|X_U\right] \neq E\left[Y_i \,|X_i\right].$$

In some household surveys, X_U consists of demographic data. A person's survey variable could depend on their own demographic characteristics and those of other household members, because the latter could indicate family or living arrangements. Models of this kind are called contextual models; see for example Kreft and De Leeuw (1998).

Consider a regression estimator based on $E\left[Y_i \,|X_i\right]$,

$$\hat{Y}_A = \sum_{i \in s} \pi_i^{-1}\left(Y_i - \mu_i^*\right) + \sum_{i \in U} \mu_i^*$$

where $\mu_i^* = E\left[Y_i \,|X_i\right]$. This is not the same as $\hat{Y}_\mu$, because $\hat{Y}_\mu$ uses $\mu_i = E\left[Y_i \,|X_U\right]$. If it can be assumed that there are no correlations across different households, then Theorem 1 means that the AV of $\hat{Y}_A$ must be greater than or equal to the AV of $\hat{Y}_\mu$. To illustrate the possible loss from using $\hat{Y}_A$ instead of the optimal estimator, consider the special case

when: π_i are fixed constants not depending on X_U; X_i are independently and identically distributed for $i \in U$; $\{Y_i : i \in U\}$ are independent conditional on X_U; $\{(X_i, Y_i) : i \in U\}$ are identically distributed; the sample size, n, is non-random so that $n = E[n] = \sum_{i \in U} \pi_i$. It is further assumed that $\sigma_i^2 = var[Y_i|X_U]$ is equal to a constant, $V_{y|xU}$, for all $i \in U$, and that $var[Y_i|X_i]$ is equal to a constant $V_{y|x}$ for all $i \in U$. From Theorem 1,

$$var\left[\hat{Y}_\mu - Y\right] = E\left[\sum_{i \in U}\left(\pi_i^{-1} - 1\right)\sigma_i^2\left(X_U\right)\right] = V_{y|xU}\sum_{i \in U}\left(\pi_i^{-1} - 1\right).$$

The AV of $\hat{Y}_A$ is given by

$$\begin{aligned}
var\left[\hat{Y}_A - Y\right] &= E\left[var\left[\hat{Y}_A - Y\,|s\right]\right] + var\left[E\left[\hat{Y}_A - Y\,|s\right]\right] \\
&= E\left[var\left[\hat{Y}_A - Y\,|s\right]\right] + 0 \\
&= E\left[var\left[\sum_{i \in s}\left(\pi_i^{-1} - 1\right)\left(Y_i - \mu_i^*\right) - \sum_{i \in U - s}\left(Y_i - \mu_i^*\right)|s\right]\right] \\
&= E\left[V_{y|x}\left\{\sum_{i \in s}\left(\pi_i^{-1} - 1\right)^2 + N - n\right\}\right] \\
&= V_{y|x}\left\{\sum_{i \in U}\pi_i\left(\pi_i^{-1} - 1\right)^2 + N - n\right\} \\
&= V_{y|x}\left\{\sum_{i \in U}\pi_i\left(\pi_i^{-1} - 1\right)^2 + \sum_{i \in U}\left(1 - \pi_i\right)\right\} \\
&= V_{y|x}\sum_{i \in U}\left(\pi_i^{-1} - 1\right)
\end{aligned}$$

Hence the inefficiency of $\hat{Y}_A$ is

$$var\left[\hat{Y}_A - Y\right]/var\left[\hat{Y}_\mu - Y\right] = V_{y|x}/V_{y|xU}.$$

So the ratio of the AVs is equal to the ratio of the variance of Y_i conditional on X_i to the variance of Y_i conditional on X_U, which depends on the predictive power of X_U for Y_i given X_i.

4. SUMMARY

The Godambe-Joshi lower bound, which defines optimal model-assisted estimators of total, treats the auxiliary variables as known constants. It is applicable when the values of the survey variable are independent for different units. Theorem 1 states a similar lower bound which allows the auxiliary variables to be random, allows for multiple survey variables, and allows for correlations between values for different units (provided that sample and non-sample values are always independent). This leads to four new conclusions:

- Modelling the marginal distribution of the auxiliary variables cannot give any improvement in the asymptotic AV of design-unbiased estimators of Y as the sample and population size tend to infinity.

- If several survey variables are collected for the sample, it does not reduce the lower bound if all are considered together. Therefore multivariate modelling of several survey variables has no benefit for the asymptotic AV.

- The optimal estimator is applicable to one-stage cluster sampling even if there are dependencies between values of the survey variables for different units in the same cluster. However the estimator is not optimal for more complex cases of multi-stage sampling which are often used in practice.

- Conditional models for the variable of interest for a unit should condition on the auxiliary data for all population units, not just the auxiliary data for that unit. This can make a difference for contextual models in cluster sampling.

Acknowledgements

This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics.

REFERENCES

1. Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In R. Holt & Winston (Eds.), *Foundations of Statistical Inference* (pp. 203–242). Toronto.
2. Brewer, K. R. W. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, *5*, 93–105.
3. Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: Wiley.
4. Firth, D., & Bennett, K. (1998). Robust models in probability sampling (with discussion). *Journal of the Royal Statistical Society Series B*, *60*(1), 3–56.
5. Godambe, V. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society Series B*, *17*, 269–278.
6. Godambe, V., & Joshi, V. (1965). Admissibility and Bayes estimation in sampling finite populations 1. *Annals of Mathematical Statistics*, *36*, 1707–1722.
7. Isaki, C., & Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*(377), 89–96.
8. Kreft, I., & De Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage.
9. Lehtonen, R., & Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, *24*(1), 51–55.
10. Mukerjee, R., & Sengupta, S. (1989). Optimal estimators of finite population total under a general correlated model. *Biometrika*, *76*(4), 789–794.
11. Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387.
12. Sarndal, C. (1980). On $\pi-$inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, *67*(3), 639–650.

13. Sarndal, C., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
14. Steel, D., & Welsh, A. (2006). *Joint models for clustered data.* (Preprint). University of Wollongong School of Mathematics and Applied Statistics.
15. Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

APPENDIX: PROOFS

Let $\hat{Y}_\pi = \sum_{i\in s}\pi_i^{-1}Y_i$. This is the Horvitz-Thompson or inverse probability estimator which is exactly design unbiased (e.g. Sarndal et al., 1992). Let S be the set of all possible samples s.

Lemma: Under the assumptions stated in Theorem 1,

$$\sum_{s\in S} p(s)cov\left[\hat{Y}_\pi - Y, \hat{Y}-\hat{Y}_\pi \middle| s, X_U\right] = 0.$$

Proof of Lemma:

$$\begin{aligned}
& \sum_{s\in S} p(s)cov\left[\hat{Y}_\pi - Y, \hat{Y}-\hat{Y}_\pi \middle| s, X_U\right] \\
= & \sum_{s\in S} p(s)E\left[\left(\hat{Y}_\pi - Y - E\left[\hat{Y}_\pi - Y \middle| s, X_U\right]\right)\left(\hat{Y}-\hat{Y}_\pi\right) \middle| s, X_U\right] \\
= & \sum_{s\in S} p(s)E\left[\left(\sum_{i\in s}\pi_i^{-1}(Y_i-\mu_i) - \sum_{i\in U}(Y_i-\mu_i)\right)\left(\hat{Y}-\hat{Y}_\pi\right) \middle| s, X_U\right] \\
= & \sum_{s\in S} p(s)E\left[\left(\sum_{i\in s}\left(\pi_i^{-1}-1\right)(Y_i-\mu_i) - \sum_{i\in U-s}(Y_i-\mu_i)\right)\left(\hat{Y}-\hat{Y}_\pi\right) \middle| s, X_U\right] \\
= & \sum_{s\in S} p(s)E\left[\left(\sum_{i\in s}\left(\pi_i^{-1}-1\right)(Y_i-\mu_i(X_U))\right)\right. \\
& \qquad \left.\left(\hat{Y}(s,D_s,X_U) - \hat{Y}_\pi(s,D_s,X_U)\right) \middle| s, X_U\right] \\
& - \sum_{s\in S} p(s)E\left[\left(\sum_{i\in U-s}(Y_i-\mu_i(X_U))\right)\right. \\
& \qquad \left.\left(\hat{Y}(s,D_s,X_U) - \hat{Y}_\pi(s,D_s,X_U)\right) \middle| s, X_U\right]
\end{aligned} \tag{9}$$

The second term of (9) is zero because (Z_i, Y_i) and (Z_j, Y_j) are independent for all $i \in s$ and $j \notin s$ conditional on X_U, so that Y_i is conditionally independent of

$$\left(\hat{Y}(s,D_s,X_U) - \hat{Y}_\pi(s,D_s,X_U)\right)$$

for $i \notin s$. So (9) becomes

$$\begin{aligned}
&\sum_{s\in S} p(s) cov\left[\hat{Y}_\pi - Y, \hat{Y} - \hat{Y}_\pi \middle| s, X_U\right] \\
&= \sum_{s\in S} p(s) E\left[\sum_{i\in s}\left(\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| s, X_U\right] \qquad (10)
\end{aligned}$$

Now, $p(s)$ is the conditional probability that sample s is selected given X_U. Hence

$$\begin{aligned}
&\sum_{s\in S} p(s) E\left[\sum_{i\in s}\left(\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| s, X_U\right] \\
&= E\left[\sum_{i\in s}\left(\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| X_U\right]
\end{aligned}$$

and so (10) becomes

$$\begin{aligned}
&\sum_{s\in S} p(s) cov\left[\hat{Y}_\pi - Y, \hat{Y} - \hat{Y}_\pi \middle| s, X_U\right] \\
&= E\left[\sum_{i\in s}\left(\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| X_U\right] \\
&= E\left[E\left[\sum_{i\in s}\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right.\right. \\
&\qquad \left.\left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| X_U, Y_U, Z_U\right] \middle| X_U\right] \\
&= E\left[\sum_{s\in S} p(s) \sum_{i\in s}\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| X_U\right] \\
&= E\left[\sum_{i\in U}\sum_{s\ni i} p(s)\left(\pi_i^{-1}-1\right)\left(Y_i - \mu_i\left(X_U\right)\right)\right. \\
&\qquad \left.\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) \middle| X_U\right] \qquad (11)
\end{aligned}$$

where $\sum_{s\ni i}$ denotes summation over all samples s which contain unit i. By assumption, $\hat{Y}$ is design unbiased so

$$\sum_{s\in S} p(s)\left(\hat{Y}\left(s, D_s, X_U\right) - \hat{Y}_\pi\left(s, D_s, X_U\right)\right) = 0$$

and therefore

$$\begin{aligned} 0 &= \sum_{s \ni i} p(s) \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \\ &\quad + \sum_{s \not\ni i} p(s) \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \end{aligned}$$

for any $i \in U$ so that

$$\begin{aligned} & \sum_{s \ni i} p(s) \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \\ = & -\sum_{s \not\ni i} p(s) \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \end{aligned}$$

for any $i \in U$. Substitution into (11) gives

$$\begin{aligned} & \sum_{s \in S} p(s) cov\left[\hat{Y}_\pi - Y, \hat{Y} - \hat{Y}_\pi \middle| s, X_U\right] \\ = & -E\left[\sum_{i \in U} \sum_{s \not\ni i} p(s) \left(\pi_i^{-1} - 1\right) (Y_i - \mu_i(X_U)) \right. \\ & \qquad \left. \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \middle| X_U\right] \\ = & -E\left[E\left[\sum_{i \in U} \sum_{s \not\ni i} p(s) \left(\pi_i^{-1} - 1\right) (Y_i - \mu_i(X_U)) \right.\right. \\ & \qquad \left.\left. \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \middle| s, X_U\right] \middle| X_U\right] \\ = & -E\left[\sum_{i \in U} \sum_{s \not\ni i} p(s) \left(\pi_i^{-1} - 1\right) E\left[(Y_i - \mu_i(X_U)) \right.\right. \\ & \qquad \left.\left. \left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right) \middle| s, X_U\right] \middle| X_U\right]. \end{aligned} \tag{12}$$

From the assumptions of the theorem, Y_i and Y_j are conditionally independent whenever $i \in s$ and $j \notin s$ for all samples s with $p(s) > 0$. Hence Y_i is conditionally uncorrelated with $\left(\hat{Y}(s, D_s, X_U) - \hat{Y}_\pi(s, D_s, X_U)\right)$ for all samples $s \not\ni i$ with $p(s) > 0$. Hence the right hand side of (12) is zero.

Proof of Theorem 1:

$$\begin{aligned} & var\left[\hat{Y} - Y \middle| X_U\right] \\ & \quad = E\left[var\left[\hat{Y} - Y \middle| s, X_U\right] \middle| X_U\right] + var\left[E\left[\hat{Y} - Y \middle| s, X_U\right] \middle| X_U\right] \\ & \quad \geq E\left[var\left[\hat{Y} - Y \middle| s, X_U\right] \middle| X_U\right] \\ & \quad = \sum_{s \in S} p(s) var\left[\hat{Y} - Y \middle| s, X_U\right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s\in S} p(s) var\left[\hat{Y}-\hat{Y}_\pi+\hat{Y}_\pi-Y\middle| s,X_U\right] \\
&= \sum_{s\in S} p(s)\left\{var\left[\hat{Y}-\hat{Y}_\pi\middle| s,X_U\right]+var\left[\hat{Y}_\pi-Y\middle| s,X_U\right]\right. \\
&\qquad\qquad \left.+2cov\left[\hat{Y}-\hat{Y}_\pi,\hat{Y}_\pi-Y\middle| s,X_U\right]\right\} \\
&\geq \sum_{s\in S} p(s) var\left[\hat{Y}_\pi-Y\middle| s,X_U\right] \\
&\quad +2\sum_{s\in S} p(s) cov\left[\hat{Y}-\hat{Y}_\pi,\hat{Y}_\pi-Y\middle| s,X_U\right] \qquad (13)
\end{aligned}
$$

The Lemma states that the second term of (13) is zero. Also, $\hat{Y}_\pi$ is equal to $\hat{Y}_\mu$ plus terms depending only on s and X_U and hence constant conditional on s and X_U. Hence $var\left[\hat{Y}_\pi-Y\,|s,X_U\right]=var\left[\hat{Y}_\mu-Y\middle| s,X_U\right]$. Making both of these substitutions into (13) gives:

$$
var\left[\hat{Y}-Y\middle| X_U\right] \geq \sum_{s\in S} p(s) var\left[\hat{Y}_\mu-Y\middle| s,X_U\right] \qquad (14)
$$

Inequality (14) implies that

$$
\begin{aligned}
var\left[\hat{Y}-Y\right] &= E\left[var\left[\hat{Y}-Y\middle| X_U\right]\right]+var\left[E\left[\hat{Y}-Y\middle| X_U\right]\right] \\
&\geq E\left[var\left[\hat{Y}-Y\middle| X_U\right]\right] \\
&\geq E\left[\sum_{s\in S} p(s) var\left[\hat{Y}_\mu-Y\middle| s,X_U\right]\right]. \qquad (15)
\end{aligned}
$$

Notice that $\hat{Y}_\mu$ is unbiased conditional on s and X_U:

$$
\begin{aligned}
E\left[\hat{Y}_\mu-Y\middle| s,X_U\right] &= E\left[\sum_{i\in s}\pi_i^{-1}(Y_i-\mu_i)+\sum_{i\in U}\mu_i-Y\middle| s,X_U\right] \\
&= \sum_{i\in s}\pi_i^{-1}(\mu_i-\mu_i)+\sum_{i\in U}\mu_i-\sum_{i\in U}\mu_i=0.
\end{aligned}
$$

Hence

$$
\begin{aligned}
var\left[\hat{Y}_\mu-Y\right] &= E\left[var\left[\hat{Y}_\mu-Y\middle| X_U,s\right]\right]+var\left[E\left[\hat{Y}_\mu-Y\middle| X_U,s\right]\right] \\
&= E\left[var\left[\hat{Y}_\mu-Y\middle| X_U,s\right]\right]+var[0] \\
&= E\left[var\left[\hat{Y}_\mu-Y\middle| X_U,s\right]\right] \\
&= E\left[\sum_{s\in S} p(s) var\left[\hat{Y}_\mu-Y\middle| s,X_U\right]\right] \qquad (16)
\end{aligned}
$$

Substituting (16) into (15) gives

$$
var\left[\hat{Y}-Y\right] \geq var\left[\hat{Y}_\mu-Y\right]
$$

which is result (6) of the Theorem.

If Y_i and Y_j are conditionally independent given X_U for all $i \neq j$, then the conditional variance of $\hat{Y}_\mu - Y$ is

$$\begin{aligned}
var\left[\hat{Y}_\mu - Y \middle| X_U\right] &= E\left[var\left[\sum_{i\in s}\pi_i^{-1}(Y_i-\mu_i)+\sum_{i\in U}\mu_i - Y \middle| s, X_U\right]\middle| X_U\right] \\
&\quad + var\left[0 \,|X_U\right] \\
&= E\left[var\left[\sum_{i\in s}\left(\pi_i^{-1}-1\right)Y_i - \sum_{i\in U-s} Y_i \middle| s, X_U\right]\middle| X_U\right] \\
&= E\left[\sum_{i\in s}\left(\pi_i^{-1}-1\right)^2\sigma_i^2 + \sum_{i\in U-s}\sigma_i^2 \middle| X_U\right] \\
&= \sum_{i\in U}\pi_i\left(\pi_i^{-1}-1\right)^2\sigma_i^2 + \sum_{i\in U}(1-\pi_i)\sigma_i^2 \\
&= \sum_{i\in U}\left\{\pi_i\left(\pi_i^{-1}-1\right)^2 + 1 - \pi_i\right\}\sigma_i^2 \\
&= \sum_{i\in U}\left(\pi_i^{-1}-1\right)\sigma_i^2
\end{aligned}$$

which is result (7) of the Theorem. The unconditional variance of $\hat{Y}_\mu - Y$ is

$$\begin{aligned}
var\left[\hat{Y}_\mu - Y\right] &= E\left[var\left[\hat{Y}_\mu - Y \middle| X_U\right]\right] + var\left[E\left[\hat{Y}_\mu - Y \middle| X_U\right]\right] \\
&= E\left[\sum_{i\in U}\left(\pi_i^{-1}-1\right)\sigma_i^2\right] + var\left[0\right] \\
&= E\left[\sum_{i\in U}\left(\pi_i^{-1}-1\right)\sigma_i^2\right]
\end{aligned}$$

which is result (8).