UW Biostatistics Working Paper Series

9-28-2009

Robustness of Semiparametric Efficiency in Nearly-Correct Models for Two-Phase Samples

Thomas Lumley
tlumley@u.washington.edu

Robustness of semiparametric efficiency in nearly-correct models for two-phase samples

Thomas Lumley

September 28, 2009

Abstract

Augmented inverse-probability weighted (AIPW) estimators for incomplete-data models typically do not have full semiparametric efficiency, but do have model-robustness properties not shared by the efficient estimator. We examine the performance of efficient and AIPW estimators when the complete-data model is nearly correctly specified, in the sense that the misspecification is not reliably detectable from the data by any possible diagnostic or test. Asymptotic results for these nearly true models are obtained by representing them as sequences of misspecified models that are mutually contiguous with a correctly specified model. For some least favorable direction of model misspecification the bias in the efficient estimator induced by even this amount of model misspecification is comparable to the extra variability in the AIPW estimator, so that the mean squared error of the efficient estimator is no longer lower, at least in a local asymptotic minimax sense.
Keywords: semiparametric efficiency, model misspecification, contiguity, two-phase sampling, regression models

Robins, Rotnitzky & Zhao (1994) characterized all the regular asymptotically linear estimators for a marginal mean model in an incomplete data or two-phase sampling problem. In this problem, variables W are measured on a sample of N individuals and further variables Z on an unequal-probability subsample of size n, and we are interested in estimating a finite dimensional parameter θ in the presence of typically infinite-dimensional nuisance parameters. They characterized the semiparametric efficient estimator, which is often difficult to construct, and also described a more convenient class of Augmented Inverse-Probability Weighted (AIPW) estimators. Robins, Hsieh, & Newey (1995) extended this result to more general semiparametric problems. In addition to computational convenience the AIPW estimators are design-consistent. That is, if the model for the complete data is misspecified, the AIPW estimators based on incomplete data converge to the same limiting value as if complete data were available. On the other hand, even the most efficient AIPW estimator may be far from the semiparametric efficiency bound.

For example, consider a logistic regression model in a nested case–control study. A binary outcome Y is measured for a whole cohort of N people, and predictors Z are measured

for all the cases ($Y = 1$) and a fraction π of the controls ($Y = 0$). The outcome model is

$$\text{logit}\, P[Y = 1|Z = z] = z'\theta. \tag{1}$$

If Z were available for the entire cohort, the efficient estimator would be the maximum likelihood estimator, $\hat{\theta}_{CH}$, an unweighted logistic regression.

Standard practice in epidemiology is to estimate θ by unweighted logistic regression on the case–control sample. The estimator $\hat{\theta}_{CC}$ is a maximum likelihood estimator (Prentice & Pyke, 1979) and is semiparametric-efficient (Breslow et al, 2000) under the semiparametric model that assumes equation 1 is exactly true. In survey sampling, the natural estimator would be the weighted likelihood estimator with weights of 1 for cases and $1/\pi$ for controls. This estimator $\hat{\theta}_{SS}$ is often substantially less efficient than $\hat{\theta}_{CC}$ when equation 1 is exactly true.

On the other hand, if equation 1 is not exactly true, $\hat{\theta}_{SS}$ converges to the same limit that $\hat{\theta}_{CH}$ does, and $\hat{\theta}_{CC}$ converges to a different limit. If we regard the intended complete-data anlaysis (rather than the model) as defining the parameter of practical interest, $\hat{\theta}_{SS}$ is consistent and $\hat{\theta}_{CC}$ is inconsistent. This point of view was taken by survey statisticians, eg, Korn & Graubard (1999), Xie & Manski (1989), Scott & Wild (1986).

This example shows that a genuine controversy can exist as to the best approach. Recently, Scott & Wild(2002) re-examined the case–control design in detail and concluded that the bias of $\hat{\theta}_{CC}$ under various types of model misspecification was relatively harmless and that $\hat{\theta}_{CC}$ should often be preferred. This sort of detailed examination has not been done for many models, as the efficient estimator is often hard to characterize and even harder to compute.

Related issues arise in genetic epidemiology. Both independence of alleles at a locus (Hardy–Weinberg Equilibrium) and independence of genetic and environmental risk factors allow more efficient estimation of genetic associations, with the risk of bias if the assumptions are not met. In this example it is still not clear which estimators are to be preferred, and the recent dramatic increases in sample size for genetic association studies have reinforced concerns that the bias from misspecified assumptions may be non-negligible.

A common response to concerns about model misspecification is to claim that the assumptions are empirically verifiable, and so misspecification can be detected and removed. This response implicitly assumes that an appropriate estimator under a correctly specified model will also be appropriate when the model is close to being correctly specified. While gross misspecification will be detectable, the situation is less clear for nearly correct models. Comparisons of efficiency for $\sqrt{n}$–consistent estimators of a parameter θ involve differences in $\hat{\theta}$ of size $O_p(n^{-1/2})$, and misspecification resulting in biases of this order is thus not negligible. A closely-related class of parametric problems was studied by Claeskens & Hjort (2008, section 5.2). They considered parametric models and submodels such as the Weibull and exponential models for failure times. Under the assumption that the larger parametric model was correct, they computed the mean squared error of the MLE in the larger model and the submodel and gave conditions for

predictions based on the submodel to be more accurate even if the submodel was not correctly specified. We derive similar bounds in the semiparametric two-phase sampling problem.

This paper shows that departures from the outcome model too small to be reliably detected can still introduce sufficient bias in the efficient estimator to outweigh the precision advantage it has over the best AIPW estimator. Section 1 presents a simpler motivating example from genetic epidemiology, where the distributions are discrete and estimators and their asymptotic distributions are available in closed form. The remainder of the paper shows that essentially the same phenomena exist in semiparametric incomplete data models. Section 2 specifies the notation and the classes of models and estimators developed by Robins, Rotnitzky, & Zhao (hereafter RRZ). Section 3 gives the the mathematical idealization of 'nearly true' models in terms of contiguous sequences, and describes how the target of estimation is defined in these misspecified models. In Section 4 the asymptotic behaviour of the efficient estimator under the 'nearly true' model is derived, and compared to that of the AIPW estimators. Section 5 compares the asymptotic results to finite-sample simulations for a two-phase design where the efficient influence function is known explicitly, the classical case–control design. Finally, section 6 discusses some of the implications of these results for data analysis in two-phase studies.

1 Motivating example

A simple example for looking at nearly-true models is estimating a drug–gene interaction in case–control data. Our data (Table 1) come from a study by Psaty *et al.*, 2002. The *Gly460Trp* mutation in the α-adducin gene has been linked to salt-sensitive hypertension in both animal and human studies. Theory, and experiments in animals, suggest that this form of hypertension might be more responsive to thiazide diuretics than to other blood pressure drugs. Psaty et al. collected data on the α-adducin genotype and medication use for treated hypertensives who had heart attack or stroke and for controls, and fitted a logistic regression model

$$\text{logit } P[Y=1] = \alpha + \beta_G G + \beta_D D + \gamma G \times D$$

where $Y=1$ is an indicator for case status, G is an indicator for a carrier of the variant form of α-adducin and D is an indicator for treatment with diuretics.

With binary drug and gene data the case–control interaction estimate is the ratio of the drug–gene odds ratio $\hat{\psi}_{\text{case}} = e^{\hat{\beta}_{\text{case}}}$ in cases to the drug–gene odds ratio $\hat{\psi}_{\text{ctrl}} = e^{\hat{\beta}_{\text{ctrl}}}$in controls.

It is often plausible that drug and genetic variant are at least approximately independent in the population. For $2 \times 2 \times 2$ table and rare events, case-only estimation exploits the independence [Piegorsch et al, 1994].

Table 1: Interaction between thiazide diuretics and the α-adducin Gly460Trp polymorphism (Psaty et al, 2000)

		G	
	D	0	1
Case	0	103	85
	1	94	41
Control	0	248	131
	1	208	128

If drug use and genetic variant are independent in the population and the disease is rare, the population odds ratio in controls ψ_{ctrl} is unity and the interaction term can be estimated simply by $\hat{\psi}_{\text{case}}$. This case-only estimator is more efficient [Piegorsch et al], often by a substantial margin. Psaty et al found $\hat{\psi}_{\text{case}} = 0.45$ with 95% confidence interval 0.33–0.84, and $\hat{\psi}_{\text{case}}/\hat{\psi}_{\text{ctrl}} = 0.53$ (0.26–0.79). The case–only estimator has a narrower confidence interval, with the ratio of upper to lower endpoint being 2.5, vs 3.0 for the case–control estimator. On the other hand, the point estimate is different, raising the question of bias. It is plausible that the choice of antihypertensive drug is independent of the genotype, since the genotype is not known and does not have obvious phenotypic effects, but this is a weaker argument than would be provided by randomization.

To examine the large-sample theoretical behavior we consider the two estimators of the logarithm of the interaction odds ratio γ: the case–control estimator $\hat{\gamma}_{\text{cc}} = \hat{\beta}_{\text{case}} - \hat{\beta}_{\text{ctrl}}$ and the case-only estimator $\hat{\gamma}_{\text{case}} = \hat{\beta}_{\text{case}}$. As cases and controls are independent

$$\text{var}[\hat{\gamma}_{\text{cc}}] = \text{var}[\hat{\beta}_{\text{case}}] + \text{var}[\hat{\beta}_{\text{ctrl}}].$$

The case–control estimator is asymptotically unbiased, but the case–only estimator has asymptotic bias β_{ctrl}, which is non-zero when taking the drug and carrying the genetic variance are not independent in the population.

When comparing the mean squared error of these estimators the reduction in variance from using $\hat{\gamma}_{\text{case}}$ is exactly the variance of the efficient estimator $\hat{\beta}_{\text{ctrl}}$ of the bias β_{ctrl}. This means that a squared bias of the same size as the reduction in variance is not reliably detectable: the test for association in the controls, based on the null distribution

$$\hat{\beta}^2_{\text{ctrl}}/\text{var}[\hat{\beta}_{\text{ctrl}}] \sim \chi^2_1$$

is the most powerful test and it has a non-centrality parameter of 1.0 and thus has power of only 18% at 5% level. Figure 1 shows the relative asymptotic mean squared error for the case-only estimator and estimators with 1, 2, or 3 controls per case, when the independence assumption is true and when it is untrue but there is limited power to detect the violation. We see that the efficiency advantage of the case-only estimator can be lost with an effectively undetectable level of misspecification, and so is reliable only

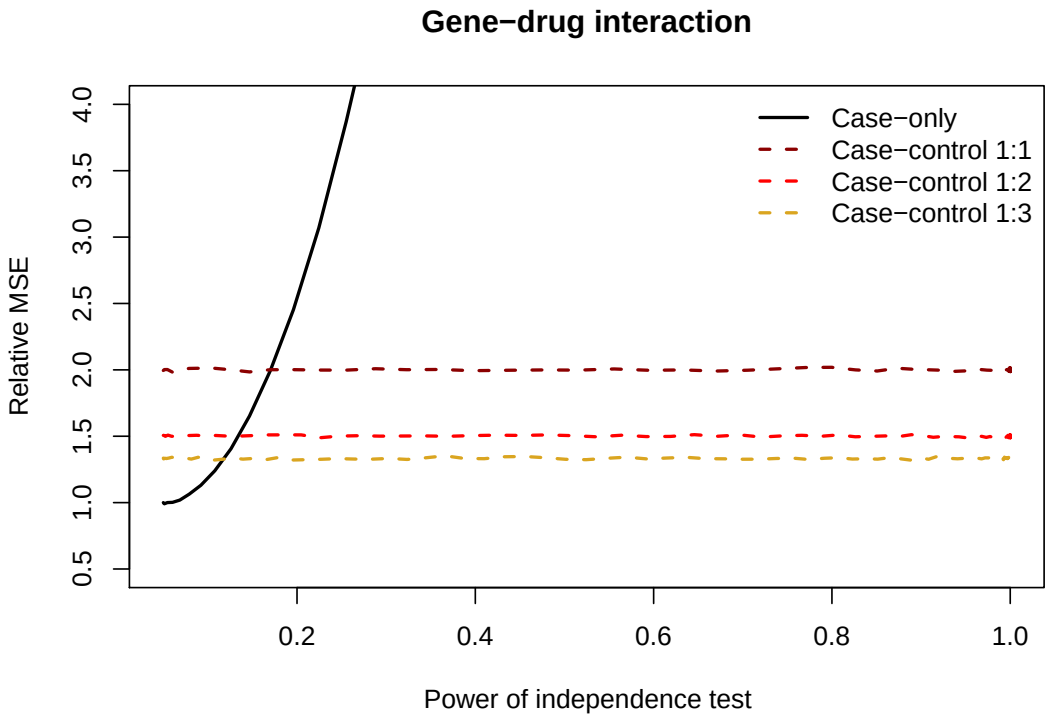

Figure 1: Asymptotic relative MSE for case-only estimator and case-control estimator when independence assumption is true, and when it is untrue but there is limited power to detect misspecification

when there are good *a priori* grounds to believe in independence, such as randomization. It is interesting to note that in our example data the test statistic is

$$\hat{\beta}^2_{\text{ctrl}}/\text{var}[\hat{\beta}_{\text{ctrl}}] = 0.95.$$

This is almost exactly the mean of the test statistic under the null hypothesis and the median when the non-centrality parameter is 1.0, so the test provides no useful guidance on which estimator to prefer.

In the remainder of this paper we extend these calculations semiparametric incomplete data models. We show that there is a way to misspecify a model so that the variance of the efficient estimator of the bias is the same as the variance reduction from using the efficient estimator. The efficient and inefficient estimators will have the same mean squared error when the bias is still too small to be reliably detected.

The case–only estimator is still of interest when genotype data are not available on controls, which can lead to substantial cost savings. The behavior that we have shown when the model is nearly true does not seriously affect the validity of the estimator, only

its relative efficiency. The case–only estimator is also useful when there are genuine *a priori* reasons, such as randomization, to expect β_{ctrl} to be zero or small. There are also other reasons that testing independence can fail, for example, strong marginal effects of the gene and drug with an outcome that is only moderately rare can lead to the distribution in controls being different from the population distribution (Gatto et al, 2004).

2 Incomplete data models

An outcome model $\mathcal{P}$ for $[Y|X,Z]$ in the complete data is indexed by the parameter of interest θ and a typically infinite-dimensional nuisance parameter η. We also observe other variables L that are not part of the efficient influence function for θ with complete data.

With complete data on (X,Y,Z,L), estimation of θ would be performed by solving

$$U(\theta) = U(\theta; X, Y, Z; \hat{\eta}) = 0,$$

where $U(\theta)$ is an estimate of the efficient influence function for θ giving at least locally efficient estimation with complete data at the true η_0.

In fact we obtain Z only for a sample of the observations, with R_i the indicator that Z is observed for observation i, and with known $\pi_i = E[R_i|\text{phase 1 data}]$. A simple consistent estimator is the Horvitz–Thompson or Inverse-Probability Weighted (IPW) estimator, which solves

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} U(\theta; X_i, Y_i, Z_i, \hat{\eta}) = 0. \tag{2}$$

The Horvitz–Thompson estimator does not use any information from (X,Y,L) on observations with $R_i = 0$. Robins, Rotnitzky & Zhao (1994) defined Augmented IPW estimators (AIPW) that solve

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} U(\theta; X, Y, Z, \hat{\eta}) + \sum_{i=1}^{N} \left(1 - \frac{R_i}{\pi_i}\right) A(\theta; X_i, Y_i, \hat{\phi}) = 0 \tag{3}$$

where A is an arbitrary function of phase-one data. The most efficient choice of $A_i(\theta)$ is $E[U_i(\theta)|\text{observed data}]$, though this will often not be feasible. In practice, reasonably efficient choices of A_i are conveniently available via connections with calibration of weights in survey sampling [Breslow *et al* 2009a,2009b, Deville & Särndal, 1992; Robins & Rotnitzky 1998].

RRZ also characterized the efficient estimator, which uses a complete-data influence function V_i obtained by projecting U orthogonal to the tangent space for nuisance parameters

$$\sum_{i=1}^{N} \frac{R_i}{\pi_i} V(\theta; X_i, Y_i, Z_i) + \sum_{i=1}^{N} \left(1 - \frac{R_i}{\pi_i}\right) A_i^{\dagger}(\theta) = 0 \tag{4}$$

with

$$A_i^\dagger(\theta) = E[V_i(\theta)|\text{observed data}].$$

If the distribution is in $\mathcal{P}$ and V can be estimated sufficiently accurately then $\hat{\theta}_{\text{eff}}$ is (locally) semiparametric efficient, and typically is strictly more efficient than the best AIPW estimator. We will write $\check{U}$ and $\check{V}$ for the influence functions of the two estimator

$$\check{U}(\theta) = \frac{R_i}{\pi_i} U(\theta; X_i, Y_i, Z_i) + \left(1 - \frac{R_i}{\pi_i}\right) A(\theta; X_i, Y_i, \hat{\phi})$$

and

$$\check{V}(\theta) = \frac{R_i}{\pi_i} V(\theta; X_i, Y_i, Z_i) + \left(1 - \frac{R_i}{\pi_i}\right) A_i^\dagger(\theta) = 0.$$

The characterization of the efficient estimator given by RRZ is not necessarily the most convenient way to compute the efficient estimator. For example, computations using profile likelihoods are described by Scott &Wild (2006) for the estimators proposed by those authors and co-workers. For our purposes it is sufficient that equation 4 characterizes the semiparametric-efficient estimator up to asymptotic equivalence, we do not need to assume that equation 4 is used in implementation, and we do not need to know the efficient influence function V explicitly.

There are many important technical issues in constructing an efficient estimator that we will not cover in this paper since we are assuming that such an estimator has in fact been constructed. Modern discussions of many of these issues can be found in Tsiatis (2006) and Kosorok (2008).

3 Nearly true models

3.1 Definition

The practical question for data analysis underlying the concept of a nearly true model is whether it is sufficient to conduct tests or examine diagnostics for model misspecification in order to justify relying on the efficient estimator. This leads to the heuristic concept of a nearly true model as a model that cannot reliably be rejected by the available diagnostics. Since essentially all our tools for proving statements about efficiency are asymptotic, we need a formal characterization of 'nearly true' that captures this heuristic concept but allows relevant asymptotic arguments to be constructed.

The available tools for model criticism will vary by the model and the data collected, but a bound on the effectiveness of these tools is given by the Neyman–Pearson lemma. If we knew that the data came either from a specific distribution $P_{\theta,\nu}$ inside the model or a specific distribution Q outside the model, the most powerful test is based on the likelihood ratio $L = dQ/dP_{\theta,\nu}$. We can thus measure the distance from Q to the model based on $\inf_{\theta,\nu} dQ/dP_{\theta,\nu}$, the Kullback–Leibler divergence.

For any fixed Q and $P_{\theta,\nu}$ the test based on L will eventually reject with certainty as N and n increase. To construct an asymptotic setting that is relevant to the practical question we need a sequence Q_n of misspecified distributions where $dQ_n/dP_{\theta,\nu,n}$ is bounded. That is, the data at hand are considered as an element of a sequence of experiments in which Q_n is not reliably distinguishable from the model.

A formal characterisation of this condition is that the sequence of data distributions Q_n and some sequence of model distributions P_n are mutually contiguous [eg Chapter 6, van der Vaart, 1998]. The definition of mutual contiguity is that for any sequence of events A_n,

$$Q_n[A_n] \to 0 \iff P_n[A_n] \to 0.$$

In particular, this holds if A_n is the event that we find a satisfactory level of model fit after using some set of diagnostics.

When Q_n and P_n are mutually contiguous the sequence of likelihood ratios $L_n = dQ_n/dP_n$ is uniformly tight. If this sequence converges in distribution under P_n to a variable L_∞ then $E[L_\infty] = 1$. By taking a subsequence if necessary it is no loss of generality to assume that this convergence in distribution holds.

3.2 Target of estimation

When the model is correctly specified we assume that the target of estimation is the parameter value θ_0 at the data-generating distribution P_{θ_0}. When the model is misspecified this default choice is not available. The choice of the complete-data limiting value of $\hat{\theta}$ as the target of estimation is important in defining efficiency and different choices could lead to qualitatively different conclusions. In this paper we choose the limiting value of $\hat{\theta}$ that would be obtained with complete data as the target of estimation and write it θ^*.

In support of this choice we argue that two-phase sampling has historically been motivated by the idea of estimating the same associations that would be estimated in complete data, but at lower cost. The choice of outcome variable, adjustment variables, and model exploration strategy in a case–cohort design, for example, will be made in the same ways that these choices would be made for a simple cohort analysis. The estimators used in two-phase samples have been constructed as extensions of those used for cohort data, rather than the cohort-data estimators being constructed as specializations of two-phase estimators.

4 Estimation in nearly true models

LeCam's Third Lemma [LeCam 1960; van der Vaart 1998, p90] describes how to convert distributions of statistics under P_n to those under Q_n. If some sequence of random variables X_n satisfies $X_n \xrightarrow{d} X$ under P_n we may define a probability measure M by

$M[A] = E[L \times \{X \in A\}]$ and then $X_n \xrightarrow{d} M$ under Q_n.

An important special case is when X_n and $\log L_n$ are asymptotically multivariate Normal. If

$$\left(X_n, \log \frac{dQ_n}{dP_n}\right) \xrightarrow{P_n} N\left(\begin{pmatrix} \mu \\ -\kappa^2/2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \kappa^2 \end{pmatrix}\right)$$

then

$$X_n \xrightarrow{Q_n} N(\mu + \tau, \Sigma).$$

The change from P_n to Q_n shifts the limiting distribution but does not change the scale. The condition that $\log L_\infty$ is Normal is natural when the data $(W, R, Z \times R)$ are independent, as L_n is then a normalized sum of independent random variables.

In particular, if $\mu = 0$ and X is scalar we can write σ^2 for Σ and reparametrize τ in terms of a correlation $\tau = \rho\kappa\sigma$. We then have

$$X_n \xrightarrow{P_n} N(0, \sigma^2)$$

and

$$X_n \xrightarrow{Q_n} N(\kappa\rho\sigma, \sigma^2)$$

Here ρ is the correlation between $\log L_\infty$ and X under P_n. It describes whether the model is misspecified in a direction that affects θ. The size of the model misspecification, in terms of the power of the most powerful test for misspecification, is measured by κ.

If we take

$$X_n = \sqrt{n}(\hat{\theta}_{\text{eff}} - \hat{\theta}_{AIPW})$$

and

$$\sqrt{n}(\hat{\theta}_{\text{eff}} - \hat{\theta}_{AIPW}) \xrightarrow{P_n} N(0, \omega^2)$$

LeCam's third lemma gives

$$\sqrt{n}(\hat{\theta}_{\text{eff}} - \hat{\theta}_{AIPW}) \xrightarrow{Q_n} N(\kappa\rho\omega, \omega^2)$$

Under Q_n the outcome model is misspecified, so care is needed in defining the 'true' parameter value. We define θ^* as the value to which the outcome-model point estimator would converge with complete data as $N \to \infty$. The AIPW estimator is still asymptotically unbiased, so

$$\sqrt{n}(\hat{\theta}_{\text{AIPW}} - \theta^*) \xrightarrow{Q_n} N(0, \sigma^2 + \omega^2)$$

and

$$\sqrt{n}(\hat{\theta}_{\text{eff}} - \theta^*) \xrightarrow{Q_n} N(\kappa\rho\omega, \sigma^2)$$

Finally, we note that the likelihood ratio test for $H_0 : Q_n$ vs $H_1 : P_n$ prescribed by the Neyman–Pearson Lemma has null distribution $N(-\kappa^2/2, \kappa^2)$ and alternative distribution $N(\kappa^2/2, \kappa^2)$, so it is equivalent to detecting a location shift of κ in a $N(0, 1)$ distribution.

The Neyman–Pearson test is one-sided, and so is more powerful than the model misspecification test in section 1. Its power at level 0.05 is 13% for $\kappa = 0.5$, 26% for $\kappa = 1$, 64% for $\kappa = 2$, and 90% for $\kappa = 3$. For $\kappa \leq 3$ the Neyman–Pearson test could certainly not be described as reliable, and in most scenarios the available model diagnostics will be less powerful than the Neyman–Pearson test as the alternative will not be known precisely.

At this point we can distinguish two cases

1. $\rho = 0$, so that the efficient estimator is still consistent for θ^*
2. $\rho \neq 0$, so that the efficient estimator is inconsistent for θ^*

When $\rho = 0$, LeCam's Third Lemma shows that the asymptotic distribution of $\hat{\theta}_{\text{eff}}$ is still $N(0, \sigma^2)$. That is, when the estimator is consistent under a contiguous sequence of misspecified models , the asymptotic variance is also correct. It is not necessary to construct a new standard error estimator such as the bootstrap or a sandwich-type estimator for model misspecification that is close to the limit of detection.

The more interesting case is when $\rho \neq 0$, so that $\hat{\theta}_{\text{eff}}$ is not consistent for θ^*. If $\hat{\theta}_{\text{AIPW}}$ is locally efficient among AIPW estimators there will exist sequences Q_n with ρ arbitrarily close to 1 under only some moment assumptions, as shown in section 4.1. For AIPW estimators other than the most efficient one ρ will be bounded away from one. In particular, this will typically be true for the Horvitz–Thompson estimator. Figure 2 shows the asymptotic relative mean squared error for the efficient estimator and the best AIPW estimator when $\rho = 1$ and $\omega^2 = \sigma^2$

The asymptotic mean squared error of $\hat{\theta}_{\text{eff}}$ is

$$MSE_{\text{eff}} = \kappa^2 \rho^2 \omega^2 + \sigma^2$$

and of $\hat{\theta}_{\text{AIPW}}$ is

$$MSE_{\text{AIPW}} = \sigma^2 + \omega^2$$

If $\kappa^2 \rho^2 > 1$, $\hat{\theta}_{\text{AIPW}}$ has smaller mean squared error.

For the best AIPW estimator there are misspecified models Q_n with ρ arbitrarily close to1, so small amounts of model misspecification in an unfavorable direction are sufficient to remove the advantage of the efficient estimator. For the crude Horvitz–Thompson estimator, on the other hand, the maximum attainable ρ may be quite small and the efficient estimator may have substantially superior mean-squared error when the model is nearly correct.

4.1 Existence of $\rho \approx 1$

Sequences Q_n with $\rho > 1 - \epsilon$ will exist for any P and model where the best AIPW estimator is not fully efficient. They are constructed by taking one-dimensional parametric families through P_n with score function constructed using $\check{V} - \check{U}$.

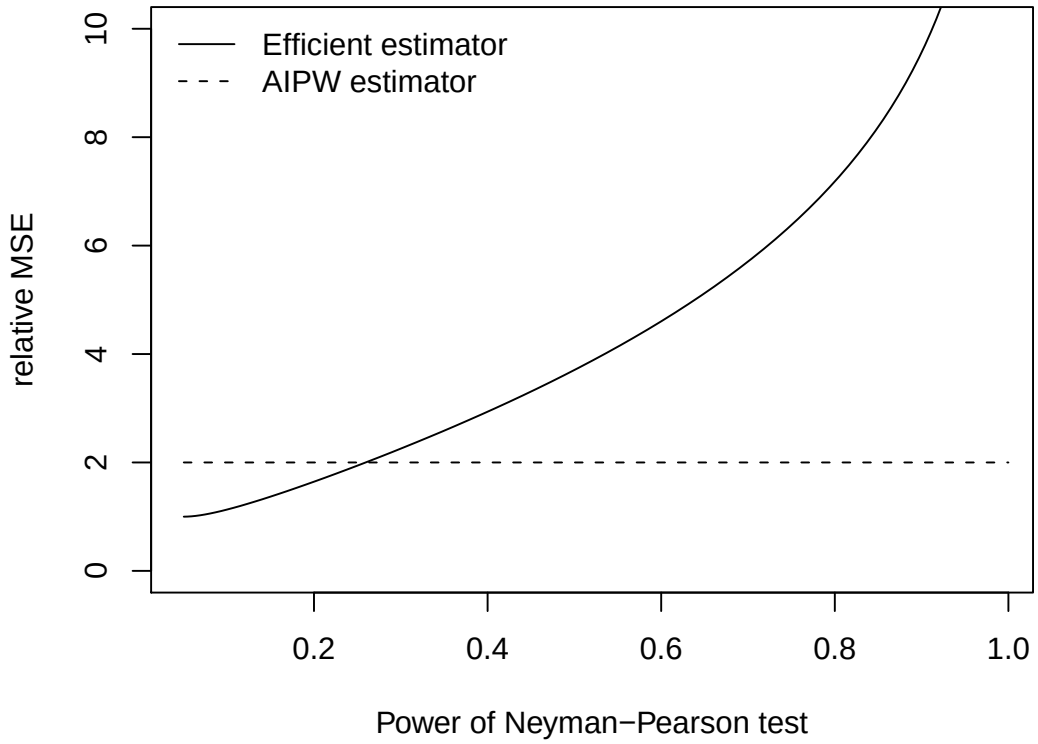

Figure 2: Asymptotic relative MSE for efficient estimator and best AIPW estimator shrinkage estimator, with $\rho = 1$, for $\omega^2 = \sigma^2$

The simplest case arises when $P[\delta \exp(\check{V}(\theta_0) - \check{U}(\theta_0))]$ is finite for δ in a neighbourhood of zero. Define a parametric family $\tilde{Q}_\delta$ by

$$\frac{d\tilde{Q}_\delta}{dP} = C_\delta e^{\delta(\check{V}-\check{U})}$$

where C_δ is an appropriate normalizing constant. Now $\tilde{Q}_{\delta/\sqrt{n}}$ is well-defined for any fixed δ and large enough n and

$$\frac{d\tilde{Q}_\delta}{dP} = C_\delta(1 + \delta(\check{V} - \check{U})) + O_p(n^{-1})$$

so the correlation between this and $\check{V} - \check{U}$ goes to 1 as n increases.

If $P[\delta \exp(V(\theta_0) - \check{U}(\theta_0))]$ is not finite, truncate it at M and define

$$\frac{d\tilde{Q}_{\delta,M}}{dP} = C_\delta e^{\delta((\check{V}-\check{U})\wedge M)}.$$

For any fixed h and M and all large enough n we can take $Q_n = \tilde{Q}_{h/\sqrt{n},M}$, and given any $\epsilon > 0$ we can choose M so that the $\rho > 1 - \epsilon$.

5 Simulation study

A simulation example to verify the theoretical results presented above requires the ability to compute the efficient estimator, an AIPW estimator that is close to optimal, and the efficient influence function. One incomplete-data design for which all the required quantities are known is the classical population-based case–control design discussed in the introduction to this paper. In phase one a binary outcome variable Y is measured on a large sample, and in phase two, predictors X are measured on all case subjects with $Y = 1$ and on a fraction π_0 of control subjects with $Y = 0$. We consider the case of a single predictor X. The model is

$$E[Y|X = x] = \mu(\theta, x) = \frac{e^{\alpha+x\beta}}{1 + e^{\alpha+x\beta}}$$

The complete data efficient influence function is

$$U(\theta) = E[XX^T\mu(\theta, X)(1 - \mu(\theta, X)]X\frac{1}{\pi}(Y - \mu(\theta, X))$$

so the IPW estimator solves

$$\sum_{i=1}^{n} \frac{1}{\pi_i} X_i(Y_i - \mu(\theta, X)) = 0$$

and

$$\check{U}(\theta) = E[XX^T \frac{\mu(\theta, X)(1 - \mu(\theta, X)}{\pi}|R = 1]X(Y - \mu(\theta, X))$$

and because there is no further phase-one information this is also the best AIPW estimator.

The efficient estimator for β is the unweighted case–control estimator, with influence function

$$\check{V}(\beta) = E[XX^T\tilde{\mu}(\theta, X)(1 - \tilde{\mu}(\theta, X)|R = 1]X(Y - \tilde{\mu}(\theta, X))$$

where

$$\tilde{\mu} = E[Y|X = x, R = 1] = \frac{e^{\alpha - \log \pi_0 + x\beta}}{1 + e^{\alpha - \log \pi_0 + x\beta}}$$

is the regression function conditional on being sampled in phase two.

The simulation results in Figure 3 are for Normally distributed x, with $(\alpha, \beta) = (-3.5,\ 1)$ giving $\omega^2 \approx 0.44\sigma^2$. The misspecified model was defined as described in section 4.1. A lowess smoother was used to estimate the difference $(U - V)(x)$ between the influence functions of the two estimators of the slope as a function of x, and this was used to generate $P(Y = 1|X = x)$ for the misspecified model. For each iteration of the simulations, a population of size 10000 was sampled from the superpopulation and a case–control sample of taken from the population. The expected number of cases was 450, and the number of controls was equal to the number of cases in each realization. The parameters β was estimated in the population by logistic regression and in the case–control sample by the MLE and by a design-based logistic regression. 1000 iterations were performed for each parameter setting.

The superpopulation parameter values (α^*, β^*) were estimated by averaging the estimates from 1000 population realizations. Mean squared errors were computed with respect to this estimated superpopulation parameter. The Neyman–Pearson test compared the Bernoulli log likelihood for the true mean function with the Bernoulli log likelihood for a model

$$P[Y = 1|X = x] = \frac{\exp(\alpha^* + x\beta^*)}{1 + \exp(\alpha^* + x\beta^*)}.$$

The log likelihood ratios followed the expected $N(\kappa^2/2, \kappa^2)$ distribution quite accurately, but the power was estimated using the empirical variance of the log likelihood ratios rather than the simulation-specified κ^2. Figure 4 shows $P[Y = 1|X = x]$ as a function of x, for the correctly specified model ($\kappa = 0$) and for the misspecified model where the design-based estimator and the MLE have the same mean squared error. The curves separate for large x, which is where the MLE puts more weight, but the separation is small, reflecting the fact that the power for distinguishing the curves is only about 25%.

6 Discussion

The arguments based on contiguity and the Convolution Theorem may be applicable in more general settings than the incomplete data models that we have considered. The key difficulty in generalizing the argument is that it is necessary to identify a clear

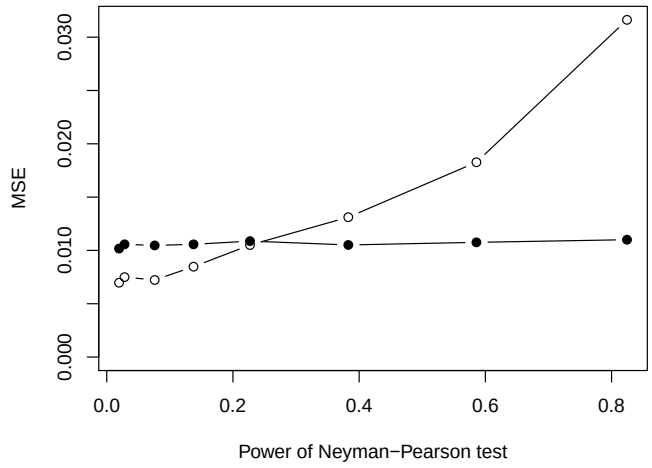

Figure 3: Efficiency of MLE (solid line) and design-based case–control estimator (dashed line) of β with $\rho \approx 1$, $\omega^2 \approx 0.44\sigma^2$, evaluated at a sample size of 450 cases and 450 controls.

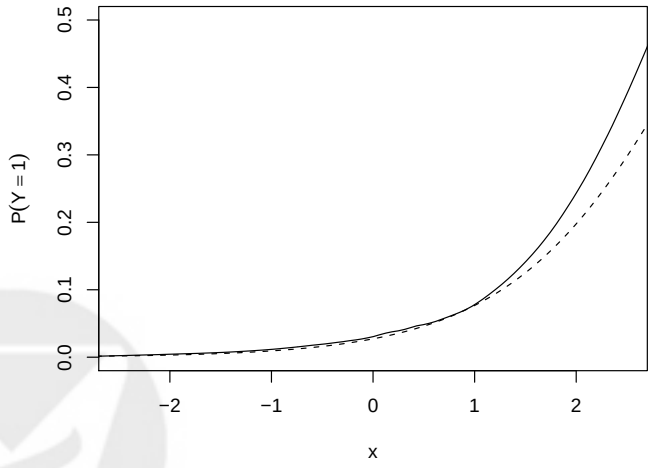

Figure 4: Population misspecified model (solid line) and best-fitting logistic model with $(\alpha^*, \beta^*) = (-3.574, 1.087)$, $\rho \approx 1$, $\kappa = 11$, sample size of 450 cases and 450 controls.

target of inference θ^* for the misspecified model. When estimating gene–environment interaction there is a clear definition of the interaction parameter ('synergy index') in the saturated model. In incomplete data models we argue that a natural definition of θ^* is the parameter that would be estimated with complete data. In other settings the target of inference would need to be defined in other ways, and the conclusions may depend on how θ^* is chosen. For 'nearly true' models as defined here the differences between plausible definitions of the target of estimation will be only $O(n^{-1/2})$, but caring about efficiency implies that differences of this size are meaningful. Alternatively, a data analyst might reasonably regard the subtle of non-linearity in Figure 4, and the resulting bias, as unimportant, but this suggests that the analyst should also not care about the difference in efficiency between the estimators, which is a difference of the same size.

There may be situations where it makes sense to use the efficient estimator even though it is biased. One plausible scenario is when the primary interest is in testing rather than estimation and the bias does not affect the null hypothesis. For example, in a case–control design, it is possible to test the null hypothesis that Y is independent of X using a likelihood ratio test, because if Y and X are independent the logistic regression model with $\beta = 0$ will be correctly specified. This example is not compelling because the design-based estimator is fully efficient when $\beta = 0$, so there is no increase in power, at least for large samples and contiguous alternatives. If there is a difference in power in small samples it would need to be demonstrated directly and would not follow automatically from the greater efficiency of the MLE. The efficiency bounds for the Cox model under case–cohort sampling (Nan *et al*, 2004, Figure 3) suggest that AIPW estimators have full or nearly full efficiency at the null hypothesis in this setting as well. Scott & Wild (2002) discuss this issue for case–control data more generally and argue that the efficient estimator may be useful even though it is biased. The case–control study is a special situation both because the bias is analytically relatively tractable and because there is so much practical experience with the design. It is clear that hard-to-detect levels of misspecification bias are virtually never one of the primary weaknesses of a case–control study: there are so many more important things that can easily go wrong.

A compromise estimator can be constructed along the lines proposed by Mukherjee & Chatterjee (2007) for the gene-environment interaction problem. They took a weighted average of the case-only and case–control estimators; this approach generalizes to taking a weighted average of the efficient estimator and an AIPW estimator. At least asymptotically the resulting compromise estimator recovers about half the extra efficiency of the efficient estimator when the outcome model is correctly specified, and reduces to the AIPW estimator under gross model misspecification. The compromise estimator does not dominate the best AIPW estimator, but it is never much worse. On the other hand, it will often require substantially more effort than computing an AIPW estimator, and its sampling distribution is not asymptotically Normal.

There may also be situations where there are good reasons to believe the key assumptions of the outcome model are true or very close to true. Independence assumptions justified by Mendelian segregation in genetics or by randomization in clinical trials are

two examples. Such an argument must rely on substantive knowledge of a particular application, rather than on the observed data. Conversely, there are situations where the sampling probabilities that we have assumed to be known may be misspecified because of non-response. Misspecified sampling probabilities would lead to bias in the AIPW estimator and the relative impact of this on the mean squared error would need to be considered separately.

In many two-phase designs we do not have either analytic results or sufficient experience to trust intuition in handling misspecification bias. Most statisticians would agree that it is unwise to rely on gains in precision from model assumptions that are unverifiable and not strongly motivated by substantive arguments. The results for contiguous models show that correct model specification is effectively an unverifiable assumption at the level of precision at which discussions of relative efficiency take place. More study of this phenomenon is needed both to characterize the behavior under typical kinds of misspecification and in order to understand when it is appropriate to accept the changed target of estimation in order to increase precision.

There is sometimes a substantial difference in efficiency between the crude Horvitz–Thompson estimator and computationally simple AIPW estimators based on calibration of weights, although the gain will be small if the available auxiliary data are not very predictive [Deville & Särndal 1992; Breslow & Chatterjee, 1999; Breslow, Lumley et al 2009; Mark & Katki, 2006; Kulich & Lin, 2004]. In contrast to the use of the semiparametric efficient estimator, using an improved AIPW estimator, at least in large enough samples, is a 'free lunch': there is a gain in precision with no change in assumptions. This fact, combined with our results on misspecification, suggest that simulation studies of efficient estimators under two-phase sampling should consider contiguous model misspecification and should use a more efficient AIPW estimator rather than the 'straw man' Horvitz–Thompson estimator where possible.

References

Breslow NE (1996) Statistics in epidemiology: The case-control study. Journal of the American Statistical Association 91(433):14-28.

Breslow NE, Chatterjee N. (1999) "Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis," Applied Statistics 48:457-68

Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. (2009) Using the Whole Cohort in the Analysis of Case-Cohort Data. Am J Epidemiol. 169(11):1398-405.

Breslow NE, Robins JM, Wellner JA: On the semiparametric efficiency of logistic regression under case-control sampling. Bernoulli 6:447–455, 2000

Chatterjee N, Carroll RJ. Semiparametric maximum-likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 2005; 92:399-418.

Claeskens G, Hjort N. L. (2008) *Model selection and model averaging.* CUP: Cambridge.

Deville J-C, Särndal C-E (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87:376–382.

Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol. 2004 Oct;33(5):1014-24

Kosorok MR (2008). Introduction to Empirical Processes and Semiparametric Inference. Springer: New York.

Korn EL, Graubard BI. (1999) *Analysis of Health Surveys.* Wiley: New York.

Kulich M., Lin D. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. Journal of the American Statistical Association, 99(467):832–844.

Le Cam, L. (1960), Locally asymptotically normal families of distributions, University of California Publications in Statistics 3: 37–98

Mark SD, Katki, H. A. (2006). Specifying and implementing nonparametric and semiparametric survival estimators in two-stage (nested) cohort studies with missing case data. Journal of the American Statistical Association, 101(474):460–471

Mukherjee, B , Chatterjee, N (2007). Exploiting gene-environment independence for analysis of case-control studies: An empirical-Bayes type shrinkage estimator to trade off between bias and efficiency. Biometrics. 64(3):685694

Nan B, Emond MJ, Wellner JA (2004). Information Bounds for Cox Regression Models with Missing Data. Annals of Statistics. 32: 723-753

Piegorsch WW, Weinberg CR, Taylor JA. (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153-62

Prentice RL. and Pyke R. (1979) Logistic disease incidence models and case-control studies. Biometrika, 66, 403–411.

Psaty BM, Smith NL, Heckbert SR, Vos HL, Lemaitre RN, Reiner AP, Siscovick DS, Bis J, Lumley T, Longstreth WT, Rosendaal FR. (2002) "Diuretic therapy, the alpha-adducin variant, and the risk of myocardial infarction or stroke in subjects with treated hypertension" JAMA 287:1680-1689

Robins JM, Hsieh F, Newey W. (1995) Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates (1995) J. Roy. Statist. Soc. B 57(2):409-424.

Robins JM, Rotnitkzy A. (1998). Discussion of: Firth, D. Robust Models in Probability Sampling. Journal of the Royal Statistical Society, Series B. 60:51–52.

Robins JM, Rotnitzky A, Zhao LP. (1994) Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89, 846-866.

Scott AJ, Wild CJ (1986) Fitting logistic models under case-control or choice based sampling, Journal of the Royal Statistical Society, B 48, 170-182.

Scott AJ, Wild CJ. (2002) "On the robustness of weighted methods for fitting models to case-control data", Journal of the Royal Statistical Society, B, 64, 207-219.

Scott AJ, Wild CJ, (2006) Calculating efficient semiparametric estimators for a broad class of missing-data problems, In *Festschrift for Tarmo Pukkila on his 60th Birthday.* E. P. Liski, J. Isotalo, J. Niemel, S. Puntanen, and G. P. H. Styan (Eds), Dept. of Mathematics, Statistics and Philosophy, Univ. of Tampere, ISBN 978-951-44-6620-5, 301-314.

Tsiatis AA (2006) *Semiparametric Theory and Missing Data.* Springer: New York.

van der Vaart A (1998) *Asymptotic Statistics* CUP: Cambridge

Xie Y, Manski CF (1989). The logit model and response-based samples. Sociol. Meth. Res. 17: 283–302.