Estimating prediction error for complex samples

Andrew Holbrook[1], Thomas Lumley[2] and Daniel Gillen[1]

[1]*Department of Statistics, University of California, Irvine, CA, U.S.A.*
[2]*Department of Statistics, University of Auckland, Auckland, New Zealand*

Key words and phrases: AIC; Generalization error; Generalized linear models; Horvitz-Thompson; NHANES III.

MSC 2010: Primary 62D05; secondary 62F40

Abstract: With a growing interest in using non-representative samples to train prediction models for numerous outcomes it is necessary to account for the sampling design that gives rise to the data in order to assess the generalized predictive utility of a proposed prediction rule. After learning a prediction rule based on a non-uniform sample, it is of interest to estimate the rule's error rate when applied to unobserved members of the population. Efron (1986) proposed a general class of covariance penalty inflated prediction error estimators that assume the available training data are representative of the target population for which the prediction rule is to be applied. We extend Efron's estimator to the complex sample context by incorporating Horvitz-Thompson sampling weights and show that it is consistent for the true generalization error rate when applied to the underlying superpopulation. The resulting Horvitz-Thompson-Efron (HTE) estimator is equivalent to dAIC, a recent extension of AIC to survey sampling data, but is more widely applicable. The proposed methodology is assessed with simulations and is applied to models predicting renal function obtained from the large-scale NHANES survey.

1. INTRODUCTION

The goal of building prediction models using empirical samples has become ubiquitous throughout all areas of business and science. With the exponential rise in statistical and machine learning methods for training flexible prediction models, increasing interest has been devoted to assessing the *extra-sample* (Efron and Tibshirani, 1997) performance of candidate models when they are applied to unobserved members of the population of interest. Analytic assessments of the performance of a prediction rule commonly focus on the expected loss associated with the rule, where the expectation is taken with respect to the underlying distribution of a new, independently sampled response conditional upon the observed support of the sampled predictors giving rise to the rule (Hastie et al., 2001). It is widely recognized that computation of the loss function solely based on the training sample is optimistically biased for this expectation.

* *Author to whom correspondence may be addressed.*
E-mail: aholbroo@uci.edu

To appear in the Canadian Journal of Statistics.

Statisticians have developed a number of techniques to adjust for this bias and thus ascertain extra-sample prediction performance. In this paper, we are interested in a class of unbiased estimators obtained by inflating the observed prediction error by a 'covariance penalty'. In linear regression under squared-error loss, Mallows's Cp (Efron, 1986) inflates the observed mean squared error by the covariance between each observed response and its fitted value. Relying on a similar covariance penalty, Stein's unbiased risk estimator is unbiased for mean squared error for differentiable prediction rules under Gaussian errors (Stein, 1981). Akaike's Information Criteria (AIC) (Akaike, 1998) penalizes the observed negative log-likelihood (or deviance) by a term that is asymptotically equivalent to a covariance penalty and thus achieves an asymptotically unbiased prediction error estimate. A more general treatment of covariance penalty inflated estimators for arbitrary prediction rules and loss functions has been considered by Efron (Efron, 1986, 2004).

But how do we ascertain extra-sample prediction performance when the *data* themselves are biased? Ignoring complex, non-uniform, unbalanced or otherwise biased sampling will cause these 'unbiased' estimators to be biased for future data arising from the true population of interest. Policy makers, economists, statisticians, and health care professionals will choose sub-optimal prediction rules and make sub-optimal predictions.

Biased samples are common. Non-uniform random sampling designs are commonly employed throughout multiple empirical sciences because they afford researchers greater efficiency in estimating parameters specific to less prevalent sub-populations. Classic examples of complex sampling designs include those implemented by the United States Census Bureau (United States Bureau of the Census, 2000) and the National Health and Nutrition Examination Study (NHANES) (National Center For Health Statistics, 1996). In each case, specific sub-populations are over-sampled by design and sampling weights are used to correct population prevalence estimates and draw inference for estimands at the population level.

Despite the growing interest in prediction modeling, there is a paucity of prediction error estimators for data arising from such complex sampling designs. *This dearth is not for a lack of need.* Prediction models trained on health survey data are common and of interest within the medical and health-care communities. For example, Wang et al. (2015) proposes statistical modeling methodologies for forecasting based on non-representative samples but does not propose any sort of model comparison criterion that would take sample complexity into account. Yu et al. (2010) trains support vector machine models using the NHANES survey to predict diabetes (and boasts 140 citations). Zhang et al. (2016) uses NHANES to train a predictive model for peripheral arterial disease, and Zhang et al. (2017) does the same for predicting atherosclerotic cardiovascular disease mortality. There are many, many more such examples we do not list here.

We address this deficiency by extending Efron's covariance penalty inflated prediction error estimator to account for complex data. The resulting estimator provides a unified framework for prediction model assessment that can be used for arbitrary loss functions and can be applied to regression based predictive models as well as algorithmically deduced prediction rules such as random forests and k-nearest neighbour approaches. We establish consistency for the true error rate relative to the super-population and further show equivalence to dAIC (Lumley and Scott, 2015), an extension of AIC for survey samples, as a special case in the context of generalized linear regression models (GLM) (McCullagh and Nelder, 1989).

2. PREDICTION ERROR ESTIMATION FOR SIMPLE RANDOM SAMPLES

We begin with the classic problem of estimating the prediction error rate where data, (y, X), is obtained via a simple random sample with y denoting the vector of outcomes of interest obtained on multiple sampling units and X denoting a matrix of explanatory variables on sampling units. Assume that an unknown data generating mechanism defined by g has produced y, from which we estimate the expectation $\mu = E_{\mathbf{g}}(y)$ with $\hat{\mu} = m(y)$, where $m(\cdot)$ is an arbitrary function potentially obtained from the data. The in-sample error is given by

$$err \equiv \frac{1}{n}\sum_{i=1}^{n} err_i = \frac{1}{n}\sum_{i=1}^{n} Q\big(y_i, \hat{\mu}_i\big)\,, \tag{1}$$

where n is the length of y, the number of observations, and $Q(\cdot,\cdot)$ denotes a specified loss function. Given the in-sample error, a common goal is to estimate

$$Err \equiv \frac{1}{n}\sum_{i=1}^{n} Err_i = \frac{1}{n}\sum_{i=1}^{n} E_0\, Q\big(y_i^0, \hat{\mu}_i\big) \tag{2}$$

for fixed $\hat{\mu}_i$. Here E_0 denotes the expectation over an unobserved random variable y_i^0 drawn independently from mechanism g but conditioning on observed support x_i. Note that although y_i^0 shares the same covariates x_i as observation y_i, the true data generating mechanism g may or may not be a function of observed covariates. Typically, err is a biased estimate for Err, but a correction is available by way of the covariance penalty.

2.1. The covariance penalty inflated estimator

Covariance penalty inflated prediction error estimators are unbiased for extra-sample prediction error. We assume that the loss function $Q(\cdot,\cdot)$ belongs to the q-class of loss functions (Efron, 1986) . A member of the q-class of loss functions is constructed from some concave function $q(\cdot)$. Given this function, the error for outcome y_i and prediction $\hat{\mu}_i$ is given by

$$Q(y_i, \hat{\mu}_i) = q(\hat{\mu}_i) + \dot{q}(\hat{\mu}_i)\,(y_i - \hat{\mu}_i) - q(y_i)\,,$$

where $\dot{q}(\cdot)$ is the derivative of $q(\cdot)$. This is not a limiting assumption. For example, the deviance functions of exponential family distributions belong to this class. Suppose y_i follows an exponential family distribution

$$g_{\mu_i}(y_i) = \exp\big(\lambda_i\, y_i - \psi(\lambda_i)\big)\,.$$

Here, λ_i is the natural parameter, ψ enforces density function integration constraints, and μ_i and λ_i are related by canonical link functions. Then the choice

$$q(y_i) = 2\left(\psi(\hat{\lambda}_i) - y_i\,\hat{\lambda}_i\right)\,, \tag{3}$$

with $\hat{\lambda}_i$ estimated from the observed data, renders the deviance (McCullagh and Nelder, 1989) as in-sample error:

$$err = \frac{1}{n}\sum_{i=1}^{n} Q\big(y_i, \hat{\mu}_i\big) = \frac{2}{n}\big(\log g_y(y) - \log g_{\hat{\mu}}(y)\big)\,, \tag{4}$$

where $g_y(\cdot)$ and $g_{\hat{\mu}}(\cdot)$ are the exponential family likelihoods with mean parameters y and $\hat{\mu}$, respectively. Note that since only the second term in the deviance depends on estimate $\hat{\mu}$, one may also consider concave function (3) as inducing the negative log-likelihood loss. Since the in-sample error is often an underestimate of Err, we define the

$$\textit{Optimism} \quad O_i = O_i(\mathbf{g}, y) = Err_i - err_i\,,$$

and the

$$\textit{Expected optimism} \quad \Omega_i(\mathbf{g}) = E_{\mathbf{g}}\, O_i(\mathbf{g}, y)\,.$$

If one is able to to obtain a consistent estimate of the optimism pertaining to a prediction rule, then the true generalization error may be estimated by adding the estimated optimism to the in-sample error err. Within the q-class of loss functions, the optimism can be analytically estimated using the term

$$\hat{\lambda}_i = -\dot{q}(\hat{\mu}_i)/2\,.$$

When $Q(y, \hat{\mu})$ is the deviance for an exponential family distribution, $\hat{\lambda}_i$ is the estimated natural parameter for the ith observation (Efron, 1986, 2004). Indeed, the following result identifies the expected optimism with the covariance between y and $\hat{\lambda}$.

Theorem 2.1 (The optimism theorem (Efron, 1986, 2004)) *For error measure $Q(y, \hat{\mu})$, we have*

$$E_{\mathbf{g}}(Err_i) = E_{\mathbf{g}}(err_i + \Omega_i)\,,$$

where

$$\Omega_i = 2\,cov_{\mathbf{g}}(y_i, \hat{\lambda}_i)\,. \tag{5}$$

It is a corollary (Efron, 1986) that when $\hat{\mu}$ is the MLE of μ for a correctly specified GLM, and when prediction error is given by the deviance, then

$$\frac{1}{n}\sum_{i=1}^{n}\Omega_i = E_{\mathbf{g}}\big(Err - err\big) \approx 2\,p/n \tag{6}$$

for p the number of model parameters. This approximation is obtained through the Taylor expansion of the link function and is exact for the Gaussian case. Following Equation (6) the reader might not be surprised that the unbiased estimator arising from Formula (5) is at least asymptotically equivalent to AIC (Efron, 1986). The same paper discusses the scenarios in which analytic estimators of the covariance penalty are available and how the parametric bootstrap may be used when they are not.

3. PREDICTION ERROR ESTIMATION FOR COMPLEX SAMPLES

While the covariance penalty inflated prediction error estimation procedure is useful for prediction rules derived from simple random samples, it is no longer accurate in a complex sample context. Hence, there is a need for a modified prediction error estimator that is applicable to models based on, say, large-scale health surveys or political polling. We now consider how to incorporate knowledge about the complex sampling design giving rise to data for accurate estimation of Err.

In the following, we make use of the *superpopulation* framework (Horvitz and Thompson, 1952) for finite population analysis. That is, we assume that the finite population $y_1\ \ldots, y_N$ is generated independently (not necessarily identically) by the same mechanism g, and that the data are then obtained via a (not necessarily uniform) sampling distribution, denoted π, where $\pi_i = Pr(y_i \in s)$, for sample s. Note that in prior sections we used g to denote the distribution producing the data, but we now use the same symbol to denote the distribution producing the finite population.

If we know $\text{cov}_{\mathbf{g}}(\hat{\lambda}_i, y_i)$—or if we can obtain a consistent estimate of it—then Theorem 2.1 provides an analytic, consistent estimator of Err_i in the case of uniform sampling. However, if the individual y_is are collected according to a non-uniform sampling scheme with known or estimable sampling probabilities, it is still possible to obtain consistent estimates of generalization error by incorporating Horvitz-Thompson (HT) sampling weights into the error estimate. In order to address this issue, we must distinguish between different kinds of generalization error for the finite population framework. We now use Err to denote the finite population prediction error rate:

$$Err = \frac{1}{N} \sum_{i=1}^{N} Q(y_i, \hat{\mu}_i) \, .$$

The superpopulation prediction error rate is the expected value of the finite population error rate $E_{\mathbf{g}}(Err)$. Next, define the Horvitz–Thompson–Efron (HTE) estimator of the predictive error rate

$$\widehat{Err} = \frac{1}{N} \sum_{i=1}^{n} \frac{1}{\pi_i} \big(err_i + 2 \, \mathrm{cov}_{\mathbf{g}}(\hat{\lambda}_i, y_i) \big) \ ,$$

where $\sum_{i=1}^{n} 1/\pi_i = N$, and err_i is the same as in Equation (1). The following corollary follows easily from Theorem 2.1.

Corollary 3.3 (Optimism theorem for biased samples) *The HTE estimator $\widehat{Err}$ is unbiased for the* superpopulation *generalization error, i.e.*

$$E(\widehat{Err}) = E_{\mathbf{g}}(Err) \, .$$

Proof. Let E_{π} denote expectation with respect to sampling mechanism. Then,

$$\begin{aligned} E(\widehat{Err}) &= E_{\pi,\mathbf{g}} \Big(\frac{1}{N} \sum_{i=1}^{n} \frac{1}{\pi_i} \big(err_i + 2 \, \mathrm{cov}_{\mathbf{g}}(\hat{\lambda}_i, y_i) \big) \Big) \\ &= E_{\mathbf{g}} \Big(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} \, E_{\pi|\mathbf{g}}(I_{i \in s}) \, \big(err_i + 2 \, \mathrm{cov}_{\mathbf{g}}(\hat{\lambda}_i, y_i) \big) \Big) \\ &= E_{\mathbf{g}} \Big(\frac{1}{N} \sum_{i=1}^{N} \big(err_i + 2 \, \mathrm{cov}_{\mathbf{g}}(\hat{\lambda}_i, y_i) \big) \Big) = E_{\mathbf{g}}(Err) \, , \end{aligned}$$

where the last equality results from Theorem 2.1 and the linearity of the expectation operator. ■

Thus, the HT extension (HTE) of the covariance penalty inflated estimator gives an unbiased estimator for the superpopulation prediction error irrespective of sample design. As a simple application of the law of large numbers, we know that $Err \overset{a.s.}{\to} E_{\mathbf{g}}(Err)$, and hence in the limit as $n, N \to \infty$ we have that $\hat{Err}$

is consistent for finite population generalization error Err. For more details on asymptotics of the superpopulation framework see Fuller (2011, Section 1.3).

3.1. dAIC and the HTE

Efron (1986) shows that AIC and the covariance penalty inflated estimator of prediction error are asymptotically equivalent. Here, we show the same is true for our HTE estimator and dAIC (Lumley and Scott, 2015), a reweighted extension of AIC that accounts for complex sampling. We introduce dAIC in detail in the Appendix, and only provide a brief example here after proving the main result.

The design based Akaike's information criterion is

$$\text{dAIC} = -2\hat{\ell}(\hat{\theta}) + 2\,\text{tr}\big\{\hat{\mathcal{J}}(\hat{\theta})\hat{V}(\hat{\theta})\big\}\,,$$

where the first term is proportional to the HT reweighted likelihood

$$\hat{\ell}(\theta) = \frac{1}{N}\sum_{i=1}^{n} w_i\,\ell_i(\theta)\,,$$

and the second term is the trace of the matrix product of

$$\hat{\mathcal{J}}(\theta) = -\frac{1}{N}\sum_{i=1}^{n} w_i\frac{\partial^2\ell_i(\theta)}{\partial\theta\partial\theta^T}\,,$$

the HT reweighted log-likelihood Hessian and $\hat{V}(\hat{\theta})$, the regular 'sandwich' estimator of the asymptotic covariance of the MLE. The Appendix contains a thorough summary of dAIC and its terms, and Section 3.1.1 provides a worked example.

In the following theorem we establish the canonical result that, under non-uniform sampling, dAIC is a special case of the HTE estimator for standard generalized linear models (GLMs).

Theorem 3.1 (Equivalence of dAIC and HTE) *The dAIC and HTE penalty terms correspond exactly, provided that: (i) a generalized linear model with canonical link is specified; (ii) the weighted deviance loss is used and the model is fit by minimizing this loss function (which corresponds to maximizing the weighted log-likelihood).*

Proof. Let λ and μ denote the natural and mean parameters of the exponential family model

$$g_\mu(y) = \exp\big(\lambda\,y - \psi(\lambda)\big)\,.$$

Given observations $(y_i, x_i), i = 1, \ldots, n$, we adopt the GLM framework and assume that, for each observation, the canonical parameter is given by a linear combination of covariates: $\lambda_i = x_i^T\theta$. We show that the dAIC penalty is equal to the HTE penalty, i.e., that

$$\text{tr}\big(\hat{V}(\hat{\theta})\,\hat{\mathcal{J}}(\hat{\theta})\big) = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}\widehat{\text{cov}}(\hat{\lambda}_i, y_i)\,,$$

when the estimate $\widehat{\text{cov}}$ is obtained using the analytic estimate (below) and *not* obtained from the parametric bootstrap (although similar estimates are obtained in practice). We first use the following two facts about exponential family distributions to obtain the forms of $\hat{\mathcal{J}}(\hat{\theta})$ and $\hat{V}_U(\hat{\theta})$:

$$\frac{\partial\psi}{\partial\lambda} = \mu \quad \text{and} \quad \frac{\partial\mu}{\partial\lambda} = \frac{\partial^2\psi}{\partial\lambda^2} = \text{cov}_{\mu}(y)\,. \tag{7}$$

Let Π be the n by n diagonal matrix with $\Pi_{ii} = \pi_i$. $\hat{\mathcal{J}}(\hat{\theta})$ is defined as the HT weighted, observed Fisher information:

$$\begin{aligned}\hat{\mathcal{J}}(\hat{\theta}) &= -\frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}\frac{\partial^2\ell_i(\theta)}{\partial\theta\partial\theta^T}\\ &= \frac{1}{N}X^T(\Pi^{-1}\frac{\partial^2\psi}{\partial\lambda^2}\big|_{\hat{\lambda}})\,X\\ &= \frac{1}{N}X^T(\Pi^{-1}\hat{\Sigma}_M)\,X\,.\end{aligned}$$

Here $\hat{\Sigma}_M = \Sigma(\hat{\mu})$ is a diagonal matrix with elements given by the model-based covariances $\Sigma(\hat{\mu})_{ii} = \text{cov}_{\hat{\mu}_i}(y_i)$. Now, $\hat{V}_U(\hat{\theta})$ takes the form

$$\frac{1}{N^2}X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X\,,$$

for $\hat{\Sigma}_O$ a matrix of observed residuals with specific form depending on sample characteristics (see Appendix).

Next we need a formula for $\hat{\theta}$. Suppose that $\hat{\theta}$ is obtained by maximizing the weighted log-likelihood. Then $\hat{\theta}$ takes the form of the WLS solution

$$\hat{\theta} = (X^T(\hat{\Sigma}_M\Pi^{-1})X)^{-1}X^T(\hat{\Sigma}_M\Pi^{-1})z\,, \tag{8}$$

where z is the linearization of the canonical link applied to y called the 'adjusted dependent variable' or the 'working residual':

$$z = \hat{\lambda} + (y-\hat{\mu})\frac{\partial\lambda}{\partial\mu}\Big|_{\hat{\mu}}\,. \tag{9}$$

Combining the formulas gives

$$\begin{aligned}
\mathrm{tr}\big(\hat{V}(\hat{\theta})\,\hat{\mathcal{J}}(\hat{\theta})\big) &= \mathrm{tr}\big(\hat{V}_U(\hat{\theta})\,\hat{\mathcal{J}}(\hat{\theta})^{-1}\big) \\
&= \frac{1}{N}\mathrm{tr}\Big(X^T(\Pi^{-1}\hat{\Sigma}_O\Pi^{-1})X\,\big(X^T(\hat{\Sigma}_M\Pi^{-1})X\big)^{-1}\Big) \\
&= \frac{1}{N}\mathrm{tr}\Big(\Pi^{-1}\,X\big(X^T(\hat{\Sigma}_M\Pi^{-1})X\big)^{-1}X^T\Pi^{-1}\hat{\Sigma}_M\hat{\Sigma}_M^{-1}\hat{\Sigma}_O\Big) \\
&= \frac{1}{N}\mathrm{tr}\Big(\Pi^{-1}\,X\big(X^T(\hat{\Sigma}_M\Pi^{-1})X\big)^{-1}X^T\Pi^{-1}\hat{\Sigma}_M\,\frac{\partial\lambda}{\partial\mu}|_{\hat{\mu}}\,\hat{\Sigma}_O\Big) \\
&= \frac{1}{N}\mathrm{tr}\Big(\Pi^{-1}\,X\big(X^T(\hat{\Sigma}_M\Pi^{-1})X\big)^{-1}X^T\Pi^{-1}\hat{\Sigma}_M\,\widehat{\mathrm{cov}}(z,y)\Big) \\
&= \frac{1}{N}\mathrm{tr}\Big(\Pi^{-1}\widehat{\mathrm{cov}}(\hat{\lambda},y)\Big) = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}\widehat{\mathrm{cov}}(\hat{\lambda}_i,y_i)\,,
\end{aligned}$$

completing the proof. ■

3.1.1. Example: linear regression model based on weighted independent sample

Let $f_\theta(y|x)$ be the homoskedastic linear regression model with Gaussian errors and regression coefficients θ. Then

$$\hat{\theta} = (X^T\Pi^{-1}X)^{-1}X^T\Pi^{-1}Y\,,$$

and

$$\hat{\mathcal{J}}(\hat{\theta}) = (X^T\Pi^{-1}X)/(N\hat{\sigma}^2)\,. \tag{10}$$

where X is the $n \times p$ observed design matrix. Next, the estimated covariance matrix of $\hat{\theta}$ is given by

$$\begin{aligned}
\hat{V}(\hat{\theta}) &= (N\hat{\sigma}^2)(X^T\Pi^{-1}X)^{-1}\times \\
&\quad \frac{X^T\Pi^{-1}\mathrm{Diag}\big((Y-X\hat{\theta})(Y-X\hat{\theta})^T\big)\,\Pi^{-1}X}{(N\hat{\sigma}^2)^2}(X^T\Pi^{-1}X)^{-1}(N\hat{\sigma}^2) \\
&= (X^T\Pi^{-1}X)^{-1}X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}\,.
\end{aligned}$$

for $\hat{\Sigma}_O = \mathrm{Diag}\big((Y-X\hat{\theta})(Y-X\hat{\theta})^T\big)$. The design-effect corrected penalty term is then given by

$$\begin{aligned}
\mathrm{tr}\big\{\hat{\mathcal{J}}(\hat{\theta})\hat{V}(\hat{\theta})\big\} &= \mathrm{tr}\big\{(X^T\Pi^{-1}X)\times \\
&\quad (X^T\Pi^{-1}X)^{-1}X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}\big\}/(N\hat{\sigma}^2) \\
&= \mathrm{tr}\big\{X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}\big\}/(N\hat{\sigma}^2)\,.
\end{aligned}$$

It follows that the dAIC for the classical linear regression case is given by

$$\frac{1}{N}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\pi_i\,\hat{\sigma}^2}+\frac{2}{N\hat{\sigma}^2}\,\mathrm{tr}\big\{X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}\big\}\,. \tag{11}$$

We now derive the HTE estimator and show it to be the same. Let

$$err_i = Q(y_i,\hat{\mu}_i) = (y_i-\hat{\mu}_i)^2$$

be the loss function, then the inflation term for the HTE estimator is given by

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{n}\frac{2}{\pi_i}\widehat{\mathrm{cov}}_{\mathbf{g}}(\hat{\mu}_i,y_i) &= \frac{2}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}\left(X(X^T\Pi^{-1}X)^{-1}X^T\Pi^{-1}\hat{\Sigma}_O\right)_{ii}\\ &= \frac{2}{N}\mathrm{tr}\big\{\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}X^T\Pi^{-1}\hat{\Sigma}_O\big\}\,.\end{aligned}$$

By the cyclic property of the trace, one has

$$\begin{aligned}\hat{Err} &= \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}(y_i-\hat{\mu}_i)^2+\frac{2}{N}\,\mathrm{tr}\big\{X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X(X^T\Pi^{-1}X)^{-1}\big\} \qquad (12)\\ &= \mathrm{dAIC}\cdot\hat{\sigma}^2\,.\end{aligned}$$

Thus equivalence between the covariance penalty inflated prediction error estimator and dAIC clearly holds in this context. This is expected as a special case of Theorem 3.1.

4. SIMULATION STUDIES

4.1. Consistency of the HTE estimator

In this section we illustrate the properties of the HTE estimator via Monte Carlo simulation under four potential non-uniform sampling designs and consider performance under a linear regression and logistic regression model fit. The simulated experiments encompass four simplified scenarios in which the relationship between model covariates, sampling probabilities, and model noise are allowed to differ. In most scenarios, the HTE estimator is shown to provide a useful estimate for the generalization error. We also show that the HTE estimator fails when sampling probabilities and model errors are strongly correlated.

In all four scenarios, the finite population is first generated and then subsampled. The exact distributions of the data are given in Table 1. In Scenario 1, the non-uniform sampling mechanism is independent of X_N^i and y_N^i. For Scenario 2, both the mean and variance are functions of the predictor, but data generation is independent of the sampling mechanism. In Scenario 3, the model errors and the sampling probabilities are a function of the same random variable z^i; in sim-

TABLE 1: Data generation schemes. For Scenario 1, the distribution of response y_N^i is independent of the sampling probability π^i. In Scenario 2, the predictor influences the mean and variance of y_N^i, but is independent from the sampling probability. For Scenario 3, the distributions of y_N^i and π^i both depend on the index i. For Scenario 4, the distributions of y_N^i and π^i both depend on X_N^i. Since the variance of y_N^i and the sampling probability both grow with X_N^i, one might expect optimism to be negative. This turns out to be the case (see Table 2).

Scenario	X_N	y_N	π
1	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} N(X_N^i, 1)$	$\pi^i \propto \log(i)$
2	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} N(X_N^i, \lvert X_N^i \rvert)$	$\pi^i \propto \log(i)$
	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} Bern\, \Phi(X_N^i)$	$\pi^i \propto \log(i)$
3	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} N(X_N^i, \log(i))$	$\pi^i \propto \log(i)$
4a	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} N(X_N^i, \lvert X_N^i \rvert)$	$\pi^i \propto \lvert X_N^i \rvert$
	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} Bern\, \Phi(X_N^i)$	$\pi^i \propto \lvert X_N^i \rvert$
4b	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} N(X_N^i, \lvert X_N^i \rvert)$	$\pi^i \propto \lvert X_N^i \rvert^{-1}$
	$X_N^i \overset{iid}{\sim} N(0,1)$	$y_N^i \overset{ind}{\sim} Bern\, \Phi(X_N^i)$	$\pi^i \propto \lvert X_N^i \rvert^{-1}$

TABLE 2: Population optimism versus HTE estimates. For both the finite population based optimism 'Err-err' and the HTE estimated optimism $\hat{\Omega}$, means, medians, and empirical intervals based on 10,000 independent simulations are shown. Based on their mutual consistency for the superpopulation optimism, one expects the empirical means of the finite population optimisms to be close to the HTE estimates. Similar to other prediction error estimation methods, the HTE conditions on the observed support of X, and is inaccurate for Scenario 4b (see text).

		Err-err		$\hat{\Omega}$	
	Scenario	mean {median}	95% Interval	mean {median}	95% Interval
Gaussian	1	0.004 {0.004}	$(-0.087,\ 0.091)$	0.004 {0.004}	$(\ 0.003, 0.005)$
	2	0.002 {0.007}	$(-0.192,\ 0.168)$	0.004 {0.004}	$(-0.019, 0.026)$
	3	0.392 {0.524}	$(-9.792, 10.114)$	0.451 {0.449}	$(\ 0.360, 0.555)$
	4a	0.023 {0.068}	$(-0.674,\ 0.446)$	0.009 {0.008}	$(-0.004, 0.024)$
	4b	-0.007 {-0.008}	$(-0.164,\ 0.156)$	0.052 {0.036}	$(\ 0.014, 0.189)$
Bernoulli	2	0.001 {0.001}	$(-0.027,\ 0.028)$	0.001 {0.001}	$(\ 0.000, 0.001)$
	4a	0.015 {0.015}	$(-0.041,\ 0.065)$	0.010 {0.007}	$(\ 0.002, 0.038)$
	4b	-0.001 {-0.001}	$(-0.050,\ 0.044)$	0.000 {0.000}	$(-0.000, 0.001)$

ulations z_i was degenerate at the logarithm of index i. In Scenario 4, the model errors and the sampling probabilities are both dependent on the absolute value of X_N^i.

For each scenario, 10,000 simulations are run. Within each simulation, a finite population of size $N = 100,000$ is generated from which a sample of size $n = 1,000$ is obtained. Then, weighted least squares regression or weighted logistic regression (using inverse probabilities as weights) are performed. In-sample error, extra-sample error, and the HTE estimator are recorded. Thus, the simulations provide a 10,000 large empirical sample of the difference between extra-sample and in-sample errors and the HTE estimator. Results are shown in Table 2.

In the first column of Table 2, the mean, median, and 0.025 and 0.975 quantiles of the simulated true optimism are presented. Note that in each scenario the optimism varies greatly between each of the 10,000 simulation iterations. The variance of the HTE estimates is much smaller in comparison, and, in most cases, their means and medians are extremely close to the means and medians of the empirical optimisms. The HTE estimator is not directly consistent for the population optimism but *is* consistent for the *superpopulation* optimism. We therefore expect that the means and medians of Table 2 should become arbitrarily close as the number of simulation iterations gets large.

Here we address the performance of the HTE estimator for Scenario 4. It is important to note that, in covariance penalty inflated prediction error estimation, the estimate is not just *based* on the observed data, but *explicitly for* the error rate of future observations over the exact same support as the observed data. For the covariance penalty inflated prediction error methodology, it is easy to forget that the accuracy of the error estimate deteriorates as the distance (given by some metric on the data space) between observations and future observations grows. This is another way of saying that error is a function of—among other things—the amount by which we use our error estimate inappropriately, i.e. as an estimate of something it is not an explicit estimator of. This fact explains why sampling probabilities that are strongly correlated with model errors would cause ostensibly inaccurate results (as in Table 2). Strong correlations between the sampling mechanism and model errors are problematic for the proposed methodology (and for dAIC) insofar as we choose to use it to generalize to future observations with drastically different support from that of the data observed. Despite these facts, the estimator performs adequately in Scenario 4a. However, Scenario 4b is simulated with unusually perverse dependencies between sampling probabilities and model errors, enforcing a negative optimism that 'breaks' our methodology.

4.2. Performance of the HTE estimator over AIC

Here we demonstrate the proposed HTE estimator's properties as a function of sample size and design efficiency in comparison to naive approaches that assume uniform sampling. We simulate a situation in which high risk individuals are oversampled. For binary outcome, this corresponds to a case-control study. We compare the *estimation* error of the *prediction* error estimator to that of AIC for: 1) different sample sizes (50 through $10,000$) keeping population prevalence of high risk individuals (1 in 200) and ratio (20%) of high risk to normal individuals fixed; and 2) different population prevalence of high risk individuals (1 in $1,000$ to 1 in 10) keeping sample size ($1,000$) and ratio (again, 20%) of high risk to

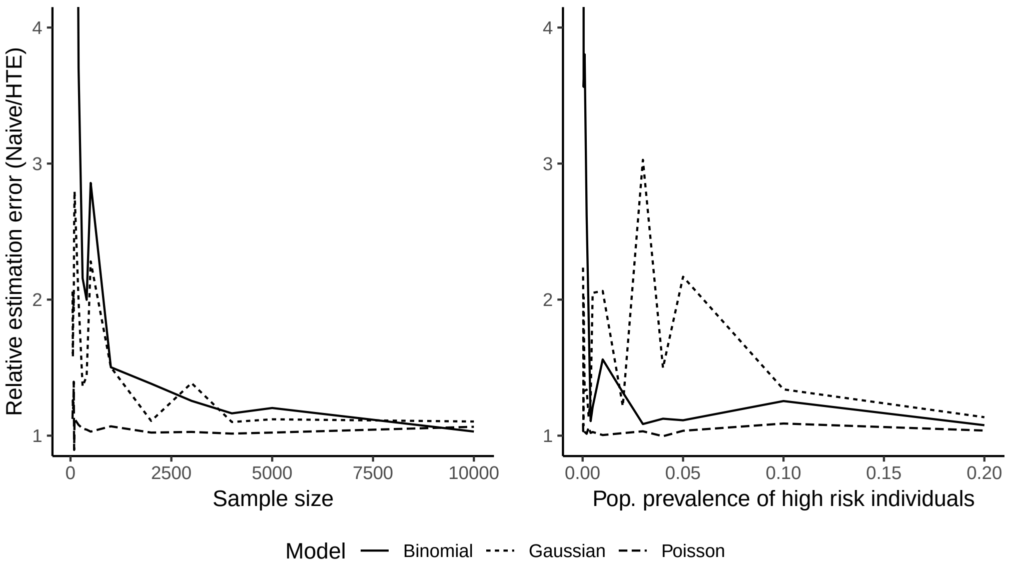

FIGURE 1: Relative performance of HTE to 'naive' AIC. For both plots, vertical axis is ratio between prediction error estimators *estimation* error for true prediction error. Each setting is simulated independently 100 times and means are communicated. Right plot varies *population* prevalence of high risk individuals keeping sampling proportions constant. HTE is particularly beneficial for small samples and efficient designs.

normal individuals fixed.

For each setting, real, count, and binary data are simulated from Gaussian, Poisson, and Bernoulli distributions, respectively, using the same systematic component with respective canonical links. For the Gaussian and Poisson scenarios, "cases" are marked using a dummy variable indicating the top $.005$-quantile of the observed linear predictor values. Linear, logistic, and log-linear models are used for prediction. True prediction error is obtained by applying the prediction rules to the rest of the finite population. Each setting is simulated independently 100 times, and means are communicated.

Figure 1 exhibits the results in terms of the ratio of estimated prediction errors to the true prediction error. This ratio is labeled the 'relative estimation error'. In both plots, the ratio in accuracies is most pronounced for the binary data case and more pronounced for the Gaussian data case than the count data case. For small samples, HTE outperforms AIC. As sample sizes grow, the difference between in-sample prediction error and extra-sample prediction error grows small, so the estimators converge. When prevalence of high risk individuals in the population is small, the study design is more efficient and the HTE outperforms AIC.

TABLE 3: Comparing dAIC and HTE for prediction of GFR (< 60) with different covariates included. Results come from a logistic regression model using deviance loss. For each method, $\hat{p}$ is the estimated effective number of parameters. The HTE term $\hat{p} = n\,\hat{\Omega}/2$ is calculated using parametric bootstrap and presented as median from 100 simulations and 95% empirical interval. Penalties increase with the number of covariates but are too small to influence generalization error estimates because of the large sample size. Importantly, dAIC and HTE give similar results and have penalties that are larger that the usual AIC penalty.

Model	Weighted deviance	p	dAIC $\hat{p}$	HTE $\hat{p}$
Age	9017.07	2	3.652	3.442 (2.550, 4.470)
+ BMI	8929.49	3	5.472	5.180 (3.998, 6.463)
+ Gender	8748.63	4	7.163	6.385 (5.339, 7.870)
+ Race/ethnicity	8695.96	5	7.121	7.158 (6.036, 8.782)

TABLE 4: Estimated error for k-nearest neighbours (kNN) classification. Again, GFR (<60) is being predicted, but errors are based on 0-1 loss. As the number of voting neighbours increases, in-sample error increases, and the HTE optimism estimate $\hat{\Omega}$ (double the covariance penalty) decreases. Generalization error appears to hit its lowest point at around 30-NN.

Predictive model	err	$\hat{\Omega}/2$	$\hat{Err}$
10-NN	0.108	0.008	0.125
20-NN	0.113	0.004	0.121
30-NN	0.114	0.003	0.120
40-NN	0.116	0.002	0.121

5. PREDICTION OF RENAL FUNCTION USING DATA FROM THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

We consider data from the Third National Health and Nutrition Examination Survey (NHANES III) with the illustrative goal of constructing a model for predicting abnormal renal function as proxied by a estimated glomerular filtration rate (GFR) of less than 60 millilitres per minute per 1.73m^2. Briefly, NHANES III was one of several periodic surveys conducted by the National Center for Health Statistics (NCHS). The survey was conducted during 1988 to 1994, and was designed to provide national estimates of health and nutritional status in the civilian non-institutionalized United States population aged 2 months and older. Children ages two months to five years, persons 60 years and older, Mexican-American persons, and non-Hispanic black persons were sampled at rates substantially higher than their proportions in the general population National Center For Health Statistics (1996). To estimate renal function, we use the Modification of Diet in Renal Disease (MDRD) equation for GFR based on serum levels and demographic covariates Levy et al. (1999). Specifically, GFR was estimated as

$$\text{GFR}_{\text{MDRD}} = 170 \times \text{SCr}^{-0.999} \times \text{ageyrs}^{-0.176} \times \text{BUN}^{-0.170} \times$$

$$\text{Salb}^{0.318} \times 1.180^{\text{black}} \times 0.762^{\text{female}},$$

where SCr denotes serum creatinine, BUN denotes blood urea nitrogen, Salb denotes serum albumin, and black and female denote indicators of non-Hispanic black race and sex, respectively. Coresh et al. (2002) have previously reported that the assay used for measuring serum creatinine in the NHANES study resulted in creatinine levels systematically higher than those used to obtain the MDRD prediction model. As a consequence, they suggest creatinine values from NHANES III be recalibrated to account for an average overestimate of 0.23 mg/dL. All analyses presented here have performed the recommended recalibration.

Grade 3 chronic kidney disease (CKD) is defined as a GFR less than 60 millilitres per minute per 1.73m^2 and is associated with increased morbidity and risk of end stage renal disease. As such we consider building a model for predicting grade 3 CKD. To illustrate the generalizability of the proposed HTE estimator, we approach this binary prediction task from two separate vantage points and using two different loss functions. First, we use logistic regression to demonstrate the empirical equivalence between dAIC and the HTE approach; here, deviance loss is used. Second, we compare a number of prediction models with respect to in-sample and estimated extra-sample errors; here, 0-1 loss is used. Using the NHANES data, we first demonstrate that dAIC and HTE give similar results for a binary prediction task with deviance loss. Second, using 0-1 loss, we calculate the in-sample error and HTE estimator for the k-neareast neighbours method (for which dAIC is not available) and show how the HTE estimate might influence prediction model preference. Data and code for all empirical studies has been included in the supplement.

5.1. dAIC and the HTE

Deviance loss is used for the dAIC/HTE comparison, so $\hat{\lambda}_i$ is given by the log-odds. Table 3 shows the results from the dAIC/HTE comparison based on a logistic regression model with different covariate combinations. p/n, the number of covariates divided by sample size, is used as a reference that accords with a scaled traditional AIC, where the scaling is to meant for easy comparison to prediction error. dAIC is similarly scaled. In general, all optimism estimates grow with the model size, but all generalization error estimates get smaller with model size. Indeed, the optimism estimates are kept small by the large size of the data sample. We note that dAIC and HTE are close.

5.1.1. Estimating the covariance penalty

The covariance between the linear predictor $\hat{\lambda}_i$ and the outcome y_i is rarely known outside of a few basic examples and approximations. As a result, Efron (1986) suggested the use of the parametric bootstrap to produce empirical covariances between simulated outcomes and their resulting fitted values. In this comparison, we use a stratified quasi-binomial GLM, where ‘quasi’ denotes a shared intra-primary sampling unit (PSU) dispersion parameter ϕ satisfying:

$$\mathrm{var}(y_i) = p_i(1-p_i)\phi \quad \text{and} \quad \phi = (1+(n_j-1)\rho)\,,$$

for n_j the size of PSU j and ρ the within-PSU correlation shared across all PSUs. If $\hat{\rho}$ differs from zero, then $\hat{\phi}$ should be multiplied by the naive parametric bootstrap estimated covariance

$$\sum_{b=1}^{B} \hat{\lambda}_i^{*b}(y_i^{*b} - y_i^{*\cdot})\,,$$

making the correct estimate

$$\widehat{\mathrm{cov}}_i = \hat{\phi} \sum_{b=1}^{B} \hat{\lambda}_i^{*b}(y_i^{*b} - y_i^{*\cdot})\,.$$

That said, this point is moot in a number of ways. First, when using a GLM (as is the case in this example), it is less computationally intensive to use dAIC, the fully analytic special case of the HTE. Second, in a study such as NHANES with PSUs on the order of 600, within-PSU correlations tend to be small (in this paper, $|\hat{\rho}| < 5 \times 10^{-4}$. Third, survey structure is often approximated for end-users. Fourth, if the purpose is model comparison, multiplying by a scalar effects each fit equally. Nonetheless, Table 3 presents dAIC (using sandwich estimator) and HTE (using parametric bootstrap) side-by-side for the to show that results are similar as is to be expected.

5.2. k-Nearest neighbours

We also compare the in-sample error and HTE estimated errors for k-nearest neighbours (kNN) models with different values of k. Here we adopt 0-1 loss, so $\hat{\lambda}_i$ is given by -1 for $p_i < .5$ and 1 for $p_i \geq .5$ (Efron, 2004). Results are shown in Table 4. In general, the larger k is, the smoother the decision rule. In terms of the bias-variance trade-off, this amounts to more bias and less variance. Indeed, as k increases from 10 to 40, in-sample prediction error increases, but the covariance-inflation decreases. The HTE estimator appears to achieve an optimum somewhere around $k = 30$.

6. DISCUSSION

Motivated by the increasing importance of algorithmic prediction methodologies and the need to effectively make predictions in the public health and medical sectors, we present a prediction error estimation methodology with the hope that it will provide for the rigorous comparison of competing prediction rules obtained from complex survey data. We show that our Horvitz–Thompson–Efron (HTE) estimator is accurate and somewhat robust for GLMs and algorithmic prediction methods. Moreover, we prove that the HTE generalizes dAIC (an AIC variant for complex samples) in the exact same way that Efron's covariance penalty inflated estimator generalizes AIC. We empirically demonstrate this fact via simulation and also by considering the prediction of chronic kidney disease using data from NHANES III, a large public health survey with prescribed sampling weights.

There is a trend in medicine toward increasingly personalized treatment. Such treatment is essentially a prediction task and, as such, is subject to the bias-variance trade-off. To help avoid over-fitting when training the necessary predictive models, it will be necessary to use large swaths of public health data, the majority of which arises from complex sampling procedures. We therefore expect that our proposed methodology and its extensions will become increasingly important for model scoring in the context of personalized medicine. Causal inference from observational data is in some ways the opposite challenge of personalized medicine, although the two are closely tied together. Moreover, methods in observational causal inference often make use of the Horvitz–Thompson reweighting procedure. We are particularly interested in the question of whether the proposed HTE estimator may be extended using reweighting procedures commonly used in causal inference and whether this methodology might be useful for effective personalized medicine.

We know of two immediate extensions to the methodology proposed here. Whereas the HTE estimator is based on the closed-form model optimism, algorithmic and non-analytic prediction error estimators are more common in the machine learning literature. We therefore anticipate the extension of both cross-validation and bootstrap prediction error estimation to the complex sample domain.

ACKNOWLEDGEMENTS

Research reported in this article was supported by the National Institutes of Health under award number R01AG053555.

BIBLIOGRAPHY

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Coresh, J., Astor, B.C., McQuillan, G., Kusek, J., Greene, T., Van Lente, F. & Levey, A.S. (2002). Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. *American Journal of Kidney Diseases, 39*(5), 920–929.

Deville, J.C. & Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(418), 376–382.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association, 81*(394), 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association, 99*(467), 619–632.

Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association, 92*(438), 548–560.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer Series in Statistics.

Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685.

Levy, A., Bosch, J., Lewis, J., Green, T., Rogers, N., Roth, D. & the Modification of Diet in Renal Disease Study Group (1999). A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. *Annals of Internal Medicine, 130*, 461–470.

Lumley, T. & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology, 3*(1), 1–18.

Lumley, T., Shaw, P.A. & Dai, J.Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review, 79*(2), 200–220.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Number 7 in Monograph on Statistics and Applied Probability. Chapman & Hall.

National Center For Health Statistics (1996). Third National Health and Nutrition Examination Survey (Documentation (catalogue number 76200)).

Robins, J.M., Rotnitzky, A. & Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846–866.

Seaman, S.R. & White, I.R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22*(3), 278–295.

Stein, C.M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1135–1151.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science, 153*, 12–18.

United States Bureau of the Census (2000). *Current Population Survey: Design and Methodology*, Volume 63. US Department of Commerce, Bureau of the Census.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association, 88*(421), 89–96.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980–991.

Yu, W., Liu, T., Valdez, R., Gwinn, M. & Khoury, M.J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making, 10*(1), 16.

Zhang, Y., Huang, J. & Wang, P. (2016). A prediction model for the peripheral arterial disease using NHANES data. *Medicine, 95*(16).

Zhang, Z., Gillespie, C., Bowman, B. & Yang, Q. (2017). Prediction of atherosclerotic cardiovascular disease mortality in a nationally representative cohort using a set of risk factors from pooled cohort risk equations. *PloS One, 12*(4), e0175822.

APPENDIX

Design-based AIC

Lumley and Scott (2015) proposed an extension to AIC under a non-uniform sampling regime by adopting the above superpopulation framework. We do not know the true distribution $\mathbf{g}$, and we seek to minimize the KL divergence between a plausible conditional distribution $f_\theta(y|\mathbf{x})$ and $g(y)$ for observed covariate vector $\mathbf{x}$. As in the uniform sampling case, this corresponds to maximizing the log-likelihood $\ell(\theta)$ of the assumed model. For complex samples, this may be estimated using Horvitz–Thompson (Horvitz and Thompson, 1952) weights:

$$\hat{\ell}(\theta) = \frac{1}{N}\sum_{i=1}^{n} w_i\, \ell_i(\theta)\,,$$

where n is the size of s, N is the size of the finite population, $w_i \propto 1/\pi_i$, and $\sum_{i=1}^{n} w_i = N$. In general, the weights need not be the true sampling weights and may be adjusted to account for nonresponse or calibrated toward population totals. Weight estimation methodologies are well understood. Important examples in the biostatistics literature include Robins et al. (1994) and Seaman and White (2013). Examples from the survey sampling literature include Valliant (1993) and Deville and Särndal (1992), and Lumley et al. (2011) consider connections between the areas.

Now, suppose that θ^* and $\hat{\theta}$ are obtained by solving the score equation and pseudo-score equation, respectively:

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = 0\,, \quad \text{and} \quad \hat{U}(\theta) = \frac{\partial \hat{\ell}(\theta)}{\partial \theta} = 0\,.$$

In the context of AIC, we are interested in estimating $E_{\mathbf{g}}(\ell(\hat{\theta}))$, the expected value of the log-likelihood $\log f_\theta(y|\mathbf{x})$ evaluated at $\hat{\theta}$ with respect to the true superpopulation distribution $\mathbf{g}$. Then it is shown in Lumley and Scott (2015) that

$$E_{\mathbf{g}}\big(\hat{\ell}(\hat{\theta})\big) = E_{\mathbf{g}}\big(\ell(\hat{\theta})\big) + \frac{1}{n}\text{tr}(\Delta) + o_p(n^{-1}) \tag{1}$$

where $\Delta = I(\theta^*)V(\theta^*)$, $V(\theta^*)$ is the asymptotic covariance of $\sqrt{n}\hat{\theta}$, and

$$\mathcal{I}(\theta) = E_{\pi,\mathbf{g}}\big(\hat{\mathcal{J}}(\theta)\big) = -E_{\mathbf{g}}\Big(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\Big)\,,$$

for

$$\hat{\mathcal{J}}(\theta) = -\frac{1}{N}\sum_{i=1}^{n} w_i \frac{\partial^2 \ell_i(\theta)}{\partial\theta\partial\theta^T}\,.$$

Note that $\mathcal{I}$ is just the Fisher information corresponding to distribution g and that $\hat{\mathcal{J}}(\hat{\theta})$ reduces to the observed Fisher information when the sample is collected uniformly. Equation (1) results in a design-based formulation of AIC for complex data, dAIC:

$$\text{dAIC} = -2\hat{\ell}(\hat{\theta}) + 2\,\text{tr}\big\{\hat{\mathcal{J}}(\hat{\theta})\hat{V}(\hat{\theta})\big\}\,,$$

where $\hat{V}(\hat{\theta})$ is the sandwich estimator for $V(\theta^*)$:

$$\hat{V}(\hat{\theta}) = \hat{\mathcal{J}}(\hat{\theta})^{-1}\hat{V}_U(\hat{\theta})\hat{\mathcal{J}}(\hat{\theta})^{-1} \tag{2}$$

for $\hat{V}_U(\hat{\theta})$ a consistent estimator of $\text{cov}\big(\sqrt{n}\hat{U}(\theta)\big)$. Thus dAIC may be rewritten as

$$\text{dAIC} = -2\hat{\ell}(\hat{\theta}) + 2\,\text{tr}\big\{\hat{\mathcal{J}}(\hat{\theta})^{-1}\hat{V}_U(\hat{\theta})\big\}\,.$$

When the weights are uniform, dAIC reduces to a robust version of AIC called Takeuchi's information criterion Takeuchi (1976). If, in addition, the model is correctly specified, dAIC reduces to AIC Lumley and Scott (2015).

The meat of the sandwich The sandwich estimator for $V(\theta^*)$ is provided in Equation (2). The meat of this sandwich is the estimated asymptotic covariance $\hat{V}_U(\hat{\theta})$ of the score function $\hat{U}(\hat{\theta})$. In its most general form, we have

$$\hat{V}_U(\hat{\theta}) = \hat{U}(\hat{\theta})\hat{U}(\hat{\theta})^T\,.$$

Among other things, the example of Section 3.1.1 shows that for linear regression in the non-uniform, unstratified sampling case, $\hat{V}_U(\hat{\theta})$ takes the form

$$\frac{X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X}{(N\hat{\sigma}^2)^2}\,,$$

where $\hat{\Sigma}_O$ is the diagonal matrix of pointwise residuals. In general, for exponential family GLMs for non-uniform, unstratified samples, we have

$$\hat{V}_U(\hat{\theta}) = \frac{1}{N^2}X^T\Pi^{-1}\hat{\Sigma}_O\,\Pi^{-1}X\,,$$

for $\hat{\Sigma}_O$ the diagonal matrix of observed residuals. The NHANES data considered in Section 5 is obtained from a stratified sample with intra-stratum and inter-PSU correlations as well as intra-PSU correlations. In such a sample with strata $h = 1, \ldots, H$, $\hat{V}_U(\hat{\theta})$ takes the form

$$\hat{V}_U(\hat{\theta}) = \frac{1}{N^2}X^T\Pi^{-1}bdiag\big(\hat{\Sigma}^1_O, \ldots, \hat{\Sigma}^H_O\big)\Pi^{-1}X\,,$$

assuming observations are ordered according to stratum membership. Here, *bdiag* indicates a block-diagonal structure, and $\hat{\Sigma}_O^h$ itself has a block structure corresponding to individual PSUs:

$$\hat{\Sigma}_O^h = \begin{pmatrix} \hat{\Sigma}_1^h & \hat{\Sigma}_{12}^h \; \dots \\ \vdots & \ddots \\ \hat{\Sigma}_{1n_h}^h & \qquad \hat{\Sigma}_{n_h}^h \end{pmatrix} .$$

In this formula, blocks along the diagonal are given by

$$\hat{\Sigma}_j^h = (Y_j^h - \hat{\mu}_j^h)(Y_j^h - \hat{\mu}_j^h)^T = \sum_{i,i' \in j} (y_i^h - \hat{\mu}_i^h)(y_{i'}^h - \hat{\mu}_{i'}^h) ,$$

where i, i' denote individuals and j denotes the PSU. Off-diagonal blocks take the form

$$\hat{\Sigma}_{jj'}^h = (Y_j^h - \hat{\bar{\mu}}_j^h)(Y_{j'}^h - \hat{\bar{\mu}}_{j'}^h)^T = \sum_{i \in j,\, i' \in j'} (y_i^h - \hat{\bar{\mu}}_j^h)(y_{i'}^h - \hat{\bar{\mu}}_{j'}^h)$$

where $\hat{\bar{\mu}}_j^h$ is the average predicted value for PSU j, and $\hat{\bar{\mu}}_j^h$ is this average multiplied by an n_j-vector of ones. The upshot is that in the above (the proof of Theorem 3.1, in particular), we can ignore the particular sampling structure and simply write

$$\hat{V}_U(\hat{\theta}) = \frac{1}{N^2} X^T \Pi^{-1} \hat{\Sigma}_O \Pi^{-1} X ,$$

letting $\hat{\Sigma}_O$ take whichever form appropriate for whichever sampling structure.