Asymptotically Unbiased Off-Policy Policy Evaluation when Reusing Old Data in Nonstationary Environments

Vincent Liu University of Alberta
Yash Chandak Stanford University
Philip Thomas University of Massachusetts Amherst
Martha White University of Alberta

Abstract

In this work, we consider the off-policy policy evaluation problem for contextual bandits and finite horizon reinforcement learning in the nonstationary setting. Reusing old data is critical for policy evaluation, but existing estimators that reuse old data introduce large bias such that we can not obtain a valid confidence interval. Inspired from a related field called survey sampling, we introduce a variant of the doubly robust (DR) estimator, called the regression-assisted DR estimator, that can incorporate the past data without introducing a large bias. The estimator unifies several existing off-policy policy evaluation methods and improves on them with the use of auxiliary information and a regression approach. We prove that the new estimator is asymptotically unbiased, and provide a consistent variance estimator to a construct a large sample confidence interval. Finally, we empirically show that the new estimator improves estimation for the current and future policy values, and provides a tight and valid interval estimation in several nonstationary recommendation environments.

1 INTRODUCTION

Off-policy policy evaluation (OPE) is the problem of estimating the expected return of a target policy from a dataset collected by a different behavior policy. OPE has been used successfully for many real world systems, such as recommendation systems (Li et al., 2011) and digital marketing (Thomas et al., 2017), to select a good policy to be deployed in the real world. A variety of estimators have been proposed, particularly based on importance sampling (IS) (Hammersley and Handscomb, 1964) and modifications to reduce variance, such as self-normalization (Swaminathan and Joachims, 2015b), direct methods that use reward models and variance reduction techniques like the doubly robust (DR) estimator (Dudík et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2016). Often high-confidence estimation is key, with the goal to estimate confidence intervals around these value estimates that maintain coverage without being too loose (Thomas et al., 2015a,b; Swaminathan and Joachims, 2015a; Kuzborskij et al., 2021).

Much less work has been done, however, for the nonstationary setting where the reward and transition dynamics change over time. Extending these approaches to the nonstationary setting is key as most real world systems change with time, or appear to due to partial observability. In this setting, we face a critical bias-variance tradeoff: using past data introduces bias, but not using past data introduces variance. Jagerman et al. (2019) introduced the sliding-window IS and exponential-decay IS estimator, that gradually reduces the impact of older data to control the bias-variance tradeoff. There is some other work predicting future OPE values for a target policy in a nonstationary environment, by using time-series forecasting (Thomas et al., 2017; Chandak et al., 2020); the goal there, however, is to forecast future policy values using past value estimates, rather than to estimate the current value.

Much of the other work tackling nonstationary problems has been for policy optimization. There is a relatively large body of work on nonstationary bandits in the on-policy setting (e.g., see Yu and Mannor (2009)). More pertinent to this work is a recent approach in the off-policy setting (Hong et al., 2021). Their focus, however, is on the use of change point detection and hidden Markov models for policy optimization in the online phase. As a result, these ideas do not directly apply to nonstationary OPE.

In this work, we propose a new approach for nonstationary OPE by exploiting ideas from a related field called *survey sampling* (Cochran, 1977), where handling nonstationary data has been a bigger focus. We propose a variant of the DR estimator, called the regression-assisted DR estimator, for nonstationary environments. We exploit two ideas: (1) using auxiliary variables from the past data to build a proxy value and incorporate the proxy value in the

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

estimator without introducing bias, and (2) a regression approach on top of the proxy value to reduce variance further. Using the regression approach introduces some bias, however, we prove that the estimator is asymptotically unbiased and provide a consistent variance estimator to construct a large sample confidence interval (CI). Moreover, we show that this regression-assisted estimator unifies several existing OPE methods, including the weighted IS estimator. We empirically show that in several recommendation problems, formalized as contextual bandits, that the new estimator improves the estimation and provides a tighter and valid CI empirically compared to the sliding-window estimators. We then extend the idea to finite horizon reinforcement learning, and highlight similar improvements.

2 PROBLEM SETUP

In this section, we describe our main problem setting: off-policy policy evaluation (OPE) in the nonstationary setting. To convey the core insights of our paper precisely, we first focus on contextual bandits.

Notation. We start by describing the standard stationary setting for OPE in the contextual bandit setting. Let $\mathcal{S}$ be a set of contexts, $\mathcal{A}$ be a set of actions, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be the reward function. The goal is to evaluate the value of a target policy π, that is, estimate $J(\pi) = \mathrm{E}_{S\sim P, A\sim\pi(\cdot|S)}[r(S,A)]$, using an offline (off-policy) dataset. The dataset is created through the interaction of a behavior policy with the environment: (1) the environment draws a context s_i from $P \in \Delta(\mathcal{S})$ and (2) the behavior policy draws an action a_i from $\pi_b(\cdot|s_i)$ and observes $r_i = r(s_i, a_i)$. This process repeats n times, giving dataset $D = \{(s_i, \mathbf{x}_i, a_i, r_i)\}_{i=1}^n$. We assume that we also observe the context feature $\mathbf{x}_s \in \mathbb{R}^d$ for each context s in the dataset.

Nonstationary OPE. Dealing with arbitrary nonstationarity may not be possible. Fortunately, many real world environments have structures than can be exploited. We consider a piecewise stationary setting with known change points, where the reward function changes across intervals but remains stationary within each interval. For example, an environment can be stationary within each day or each week or for a number of interactions. We assume the set of contexts and the set of actions do not change over time.

Let r_k denote the reward function for the k-th interval and $D_k = \{(s_i, a_i, r_k(s_i, a_i))\}_{i=1}^{n_k}$ denote the data of size n_k collected over the k-th interval. The goal is to estimate

$$J_k(\pi) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} P(s)\pi(a|s)r_k(s,a)$$

given previous datasets $D_1, \ldots, D_{k-1}$ and a newly sampled D_k. The problem mirrors the real world where we have plenty of past data $D_1, \ldots, D_{k-1}$ but only a small amount of new data D_k to estimate the current value $J_k(\pi)$. We consider a stationary context distribution to present the paper succinctly, however, our methods described in the paper are applicable to the settings where the context distribution is also changing.

It is often necessary in high-stakes applications to provide confidence intervals. Let $\mathcal{D} = (D_t)_{t=1}^k$ denote the set of all data collected across different intervals. Given $\mathcal{D}$ and a desired level of failure probability $\alpha \in (0,1)$, it would be ideal to estimate a high confidence lower bound CI^- and a high confidence upper bound CI^+ such that

$$\Pr\left(\mathrm{CI}^-(\mathcal{D},\alpha) \le J_k(\pi) \le \mathrm{CI}^+(\mathcal{D},\alpha)\right) = 1-\alpha$$

where the probability is under the randomness of D_k and conditional on all old data $D_1, \ldots, D_{k-1}$.

3 BACKGROUND

In this section, we review existing estimators for stationary OPE and describe how OPE can be written using the survey sampling formulation. We use this survey sampling formulation to introduce the proposed estimators in the next section.

3.1 Estimators for Stationary OPE

A foundational strategy to estimate $J(\pi)$ in stationary OPE is to use importance sampling. The IS estimator is given by

$$\hat{J}_{\mathrm{IS}}(\pi) = \frac{1}{n}\sum_{i=1}^n \frac{\pi(a_i|s_i)}{\pi_b(a_i|s_i)} r(s_i,a_i).$$

This IS estimator can have high variance since the importance ratio can be very large. The weighted IS (WIS) estimator (Sutton and Barto, 1998), also known as the self-normalized estimator (Swaminathan and Joachims, 2015b), normalizes the importance weights and is more commonly used. The WIS estimator is given by

$$\hat{J}_{\mathrm{WIS}}(\pi) = \sum_{i=1}^n \frac{\pi(a_i|s_i)/\pi_b(a_i|s_i)}{\sum_{j=1}^n \pi(a_j|s_j)/\pi_b(a_j|s_j)} r(s_i,a_i).$$

Besides these IS-based estimators, another common estimator is the direct method (DM). We learn a reward prediction model $\hat{r}$ and use

$$\hat{J}_{\mathrm{DM}}(\pi) = \frac{1}{n}\sum_{i=1}^n \sum_{a\in\mathcal{A}} \pi(a|s_i)\hat{r}(s_i,a).$$

The doubly robust (DR) estimator (Dudík et al., 2011) combines the DR and the IS estimator,

$$\hat{J}_{\mathrm{DR}}(\pi) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\pi(a_i|s_i)}{\pi_b(a_i|s_i)}(r(s_i,a_i)-\hat{r}(s_i,a_i)) + \sum_{a\in\mathcal{A}}\pi(a|s_i)\hat{r}(s_i,a)\right].$$

There are other OPE estimators such as estimators with clipping (Bottou et al., 2013) or shrinkage (Su et al., 2020). Dudík et al. (2012) studied the setting where the policies are non-stationary (history-dependent) but the environment is stationary, which is different from our setting. Chandak et al. (2021) focus on estimating the reward distribution and do not discuss how to efficiently leverage past data under non-stationarity.

3.2 OPE as Survey Sampling

Survey sampling can be dated back to Hansen and Hurwitz (1943); Horvitz and Thompson (1952), where they consider the problem of selecting a sample of units from a finite population to estimate unknown population parameters. Formally, let $\mathcal{U} = \{1, \dots, N\}$ be the population of interest, y_i be the study variable and $\mathbf{x}_i$ be the auxiliary variable for the unit $i \in \mathcal{U}$. A subset of the population, called a sample, is selected according to a sampling design. We observe the study variable for units in the sample, and the goal is to estimate the population total of the study variables $t_y = \sum_{i \in \mathcal{U}} y_i$.

To formalize OPE under survey sampling, let the population be $\mathcal{U} = \mathcal{S} \times \mathcal{A}$ and the study variable be $y_{s,a} = P(s)\pi(a|s)r(s,a)$. The population total of y is the value of the policy: $t_y = \sum_{(s,a) \in \mathcal{U}} y_{s,a} = J(\pi)$. The weighting $P(s)\pi(a|s)$ goes into the study variable since the goal is to estimate the total of study variable without weighting. Even though we have $P(s)$ in the study variable, the term often cancels out in the estimator.

This formulation has some subtle differences from the standard OPE formulation. First, it assumes that $\mathcal{S} \times \mathcal{A}$ is finite, since $\mathcal{U}$ is finite in survey sampling. Second, the study variable is fixed, that is, the reward function is deterministic. These limitations can be overcome by assuming that the finite population is generated as a random sample from an infinite superpopulation; this superpopulation model is discussed in the appendix. For the main body, we assume a finite population with fixed study variables.

Of particular interest for nonstationarity is the *model-assisted* approach for survey sampling (Särndal et al., 1992). The key idea is to use the auxiliary variable $x_{s,a}$ to form a proxy value $\hat{y}_{s,a}$ such that $\hat{y}_{s,a}$ is close to the study variable $y_{s,a}$. A simple example is that the auxiliary variable $x_{s,a}$ might be the value of $y_{s,a}$ at a past time and we can use proxy value $\hat{y}_{s,a} = x_{s,a}$. A general form for a model-assisted estimator, assuming the population total of the proxy value is known, is the difference estimator (Cassel et al., 1976): $\sum_{(s,a) \in \mathcal{U}} \hat{y}_{s,a} + \sum_{(s,a) \in D} \frac{y_{s,a} - \hat{y}_{s,a}}{n p_{s,a}}$ where $p_{s,a}$ is the probability of selecting the pair (s,a). This estimator is unbiased and can be much lower variance, if the proxy value is close to the study variable. This strategy is like adding a control variate, but specific to survey sampling since the source of stochasticity is different than the typical Monte Carlo setting.

4 OPE ESTIMATORS UNDER NONSTATIONARITY

In the section, we propose an estimator for nonstationary environments. There are two popular strategies that consider the bias-variance tradeoff when reusing the past data in non-stationary environments: sliding window IS and exponential decay IS (Jagerman et al., 2019). The sliding window IS estimator directly uses the IS estimator for the data in the most recent B intervals. Though not proposed in the original work, it is natural to extend this idea to other estimators. For example, for the direct method, we can build a reward model from the data in the most recent B intervals and evaluate the policy with the reward prediction.

The window size B controls the bias-variance tradeoff. If $B = 0$ then we only use the most recent data D_k: the estimator does not introduce bias by using past data but suffers high variance due to having a small sample size. If we use a large B, the estimator might introduce large bias but might have lower variance due to a larger sample size. Sliding window estimators require carefully choosing B to balance the bias from using past data and the variance from not using past data. The balance usually depend on how fast the environment is changing, which is usually unknown. Moreover, even with a small value of B, the bias of the sliding window estimator can be so large that the confidence interval is invalid, as we will show in the experiment section.

Therefore, the main question that we aim to address is:

How can we reuse the past data for nonstationary OPE without introducing large bias?

One natural way to leverage the past data would be to use the DR estimator with a reward prediction learned from the past data, as described in the following section. However, naively using the past data to construct a reward prediction may not be the best approach in the nonstationary setting. This raises a followup question: *How can we obtain a good reward prediction to both reduce the error of estimation and also obtain tight CIs?* To address this challenge we draw inspiration from the survey sampling literature, and propose the regression-assisted DR estimator, that helps reduce variance further and provides tighter CIs.

4.1 The Difference and DR estimator

We can leverage the idea of the difference estimator in survey sampling, for our nonstationary OPE setting. We can use the past data $D_{k-B}, \dots, D_{k-1}$ to build a reward prediction $\hat{r}_k$ as a function of the context feature and the action: $\hat{r}_k(s,a) = m(\mathbf{x}_s, a; \theta)$ for some function m parameterized by θ. The reward prediction can be used as the proxy value in the estimator. The resulting difference esti-

mator, for interval k, is

$$\hat{t}_{\text{Diff},k} = \sum_{(s,a)\in\mathcal{U}} P(s)\pi(a|s)\hat{r}_k(s,a) + \frac{1}{n}\sum_{(s,a)\in D_k} \frac{\pi(a|s)}{\pi_b(a|s)}(r(s,a) - \hat{r}_k(s,a)). \quad (1)$$

The elegance of this approach is that we can leverage past data by incorporating it into the proxy value in the difference estimator, without introducing any bias.

While the estimator is unbiased, the variance depends on the quality of the reward prediction $\hat{r}_k$ (Thomas and Brunskill, 2016). The environment is nonstationary, so past data has to be used carefully to get a good estimate, and in some cases, the estimate may be poor. In the next section, we discuss how to obtain a better prediction by fitting a regression on top of the reward prediction.

A careful reader would have noticed one other nuance with the above difference estimator: it requires the population total of the proxy value $\hat{y}_{s,a} = P(s)\pi(a|s)\hat{r}_k(s,a)$, which is the first term in Eq (1). In some cases, this information may be known and it should be leveraged to get a better estimator for OPE. In other cases, we will need to estimate it. In the standard contextual bandit setting, given a sample D_k, we often assume that we know the auxiliary variable $\mathbf{x}_{s,a}$ for all units in the set $\{(s,a) : s \in D_k, a \in \mathcal{A}\}$. If we estimate the population total from D_k with the information about the auxiliary variables, the estimator becomes

$$\hat{t}_{\text{DR},k} = \frac{1}{n}\sum_{s\in D_k}\sum_{a\in\mathcal{A}} \pi(a|s)\hat{r}_k(s,a) + \frac{1}{n}\sum_{(s,a)\in D_k} \frac{\pi(a|s)}{\pi_b(a|s)}(r_k(s,a) - \hat{r}_k(s,a)). \quad (2)$$

This estimator reduces to the DR estimator. Therefore, the DR estimator is the difference estimator when the population total of the proxy value is estimated by sample D_k.

However, there are other options to estimate the population total, that do not result in the standard DR estimator. Of particular relevance here is that we can use past data D' to estimate this population total: $\frac{1}{|D'|}\sum_{s\in D'}\sum_{a\in\mathcal{A}} \pi(a|s)\hat{r}_k(s,a)$. This term does not rely on rewards in the past data—which might not be correct due to nonstationarity—and only requires access to the auxiliary variables $\mathbf{x}_s$ in these datasets. If we assume only the rewards are nonstationary, rather than the context distribution, making these old datasets a perfectly viable option to estimate this population total. In survey sampling, this is usually motivated by assuming that there might be another survey that contains the auxiliary variables (Yang and Kim, 2020).

4.2 The Regression-Assisted DR Estimator

We consider a model on top of the reward prediction from the past data to mitigate variance further. Let $\phi_k(s,a)^\top = (1, \hat{r}_k(s,a))$ be the augmented feature vector with the reward prediction and define $\mathbf{z}_{s,a} = P(s)\pi(a|s)\phi_k(s,a)$. Note that $\mathbf{z}_{s,a}$ is a function of the auxiliary variable $\mathbf{x}_s$. We consider a (heteroscedastic) linear regression model such that the study variables $y_{s,a} := P(s)\pi(a|s)r_k(s,a)$ are realized values of the random variables $Y_{s,a}$ with $\mathrm{E}_\xi[Y_{s,a}] = \mathbf{z}_{s,a}^\top\beta$ and $\mathrm{V}_\xi(Y_{s,a}) = \sigma_{s,a}^2 = P(s)\pi(a|s)\sigma^2$ where the expectation and variance are with respect to the model ξ, and β, σ are the model coefficients. These are the assumptions underlying the regression estimator, rather than assumptions about the real world. Further, even though we consider a linear regression on the feature vector ϕ for the regression-assisted DR estimator, the reward prediction itself can be non-linear (e.g., a neural network).

The weighted least squares estimate of β is $\beta_k = \arg\min_\beta \sum_{(s,a)\in\mathcal{U}} \frac{1}{\sigma_{s,a}^2}\left(\mathbf{z}_{s,a}^\top\beta - y_{s,a}\right)^2$. Suppose the relevant matrix is invertible, β_k can be estimated using sample data D_k:

$$\hat{\beta}_k = \left(\sum_{(s,a)\in D_k} \frac{\pi(a|s)}{\pi_b(a|s)}\phi_k(s,a)\phi_k(s,a)^\top\right)^{-1} \left(\sum_{(s,a)\in D_k} \frac{\pi(a|s)}{\pi_b(a|s)}\phi_k(s,a)r_k(s,a)\right). \quad (3)$$

If we know the population total of $\mathbf{z}_{s,a}^\top\hat{\beta}_k$, then the regression-assisted DR (Reg) estimator is

$$\hat{t}_{\text{Reg},k} = \sum_{(s,a)\in\mathcal{U}} P(s)\pi(a|s)\phi_k(s,a)^\top\hat{\beta}_k + \frac{1}{n}\sum_{(s,a)\in D_k} \frac{\pi(a|s)}{\pi_b(a|s)}\left(r_k(s,a) - \phi_k(s,a)^\top\hat{\beta}_k\right). \quad (4)$$

More generally, we can use the same data or the past data to estimate the population total, as described above.

This $\hat{\beta}_k$ consists only of the weight on the past reward model and the bias unit. This may not seem like a particularly useful addition, but because it is estimated using D_k, it allows us to correct the past reward prediction.

Further, the regression-assisted DR estimator actually provides a natural way to combine existing estimators in the OPE literature, depending on the choice of the feature vector ϕ and the coefficients β. To see this, we first show how WIS can be seen as an instance of this estimator.[1]

[1]Mahmood et al. (2014) have a similar observation that the solution $\hat{\beta}$ is the WIS estimator if $\phi(s,a) = 1$ for all (s,a). They also extend the estimate with linear features ϕ and use $\phi(s,a)\hat{\beta}$ directly, which is more related to the model-based approach. In our work, the model prediction is used as the proxy value so the resulting estimators are different.

We provide the theoretical result in the stationary setting where ϕ is fixed, so we can drop the subscript k for simplicity. For the nonstationary setting, the inference for $\hat{t}_{Reg,k}$ is conducted *conditional on* the past data $D_1, \ldots, D_{k-1}$, so ϕ is again fixed and all results extends to the nonstationary setting. The proofs can be found in Appendix B.

Theorem 1 (WIS as a special case of the regression-assisted estimator)**.** Suppose we use a linear regression model with univariate feature $\phi(s,a) = 1$. Then the regression-assisted DR estimator with estimated coefficient $\hat{\beta}$ from Eq (3) has the same form as the WIS estimator:

$$\hat{t}_{\text{Reg}} = \sum_{(s,a)\in D} \frac{\pi(a|s)/\pi_b(a|s)}{\sum_{(s',a')\in S} \pi(a'|s')/\pi_b(a'|s')} r(s,a). \quad (5)$$

The result provides a novel perspective for the WIS estimator: it can be viewed as fitting a regression to predict the reward with a constant feature. As a result, the only difference between the regression-assisted DR estimator and the WIS estimator is the choice of feature vector for reward prediction. If there are other features that might be useful for predicting the reward, we can include it with the regression approach and potentially improve the WIS estimator.

In Table 1, we show that we can recover other estimators based on different choices for the coefficients $\beta = (\beta_1, \beta_2)^\top$ with the feature vector $\phi(s,a)^\top = (1, \hat{r}(s,a))$. If $\beta_1 = 0, \beta_2 = 0$, we recover the IS estimator. If $\beta_1 = 0, \beta_2 = 1$, we recover the difference estimator or the DR estimator. If $\beta_2 = 0$ and β_1 is learned from data, we recover the WIS estimator.

Table 1: A Unifying View of Existing Estimators.

	IS	DR	WIS	Reg
β_1	0	0	$\hat{\beta}_1$	$\hat{\beta}_1$
β_2	0	1	0	$\hat{\beta}_2$

There are other approaches to estimate the coefficients from data. The more robust DR estimator (Farajtabar et al., 2018) minimizes the estimated variance $\hat{\text{V}}(\hat{t}_{\text{Reg}})$ with respect to the coefficient to achieve the lowest asymptotic variance among all coefficient β under some mild conditions. Kallus and Uehara (2019) further consider an expanded model class on top of the reward prediction and minimize the estimated variance among both the expanded model class and the reward prediction model class. However, it often unclear how large the sample size needs to be such that the estimator with the lowest asymptotic variance indeed has a lower variance against other estimators in practice. On the other hand, there is a considerable literature in survey sampling on improving estimation for the total and variance estimator when the sample size is small or the feature vector is high dimensional. For example, Breidt and Opsomer (2000); McConville et al. (2017) propose different models as an alternative to the linear regression model. These regression models can be potentially more useful for feature selection or to find a model that fits the population well.

5 THEORETICAL ANALYSIS

In the regression approach, if the coefficients are estimated from the same data D_k, the estimator becomes biased. For example, the DR estimator is unbiased since the coefficients are fixed, and the WIS estimator is biased since one of the coefficients is estimated. In this section, we show that even if we run the regression on the same data we use to build the estimator, the regression-assisted DR estimator still enjoys asymptotic properties.

To prove these theoretical properties, there are a number of results from the survey sampling literature that we build on. For completeness, we provide a brief overview of survey sampling in Appendix A, and the proof of these properties under survey sampling notation in Appendix B.

Theorem 2 (Properties of the estimator)**.** Let $\text{AV}(\cdot)$ denote the asymptotic variance in term of the first order, that is $\text{V}(\cdot) = \text{AV}(\cdot) + o(n^{-1})$, we have (1) $\hat{t}_{\text{Reg}}$ is asymptotically unbiased with a bias of order $O(n^{-1})$, and (2)

$$\text{AV}(\hat{t}_{\text{Reg}}) = \frac{1}{n} \left(\sum_{(s,a)\in U} P(s) \frac{\pi(a|s)^2}{\pi_b(a|s)} (r(s,a) - \phi(s,a)^\top \beta)^2 - t_e^2 \right)$$

where $t_e = \sum_{\mathcal{U}} P(s)\pi(a|s)(r(s,a) - \phi(s,a)^\top \beta)$.

Variance estimation for the regression-assisted DR estimator. The exact form of the variance of the regression-assisted DR estimator is often difficult to obtain, so we use the approximate variance from Theorem 2. Replacing the unknown β by the sample-based estimate $\hat{\beta}$, we have a variance estimator

$$\hat{\text{V}}(\hat{t}_{\text{Reg}}) = \frac{1}{n(n-1)} \left[\sum_{(s,a)\in D} \left(\frac{\pi(a|s)}{\pi_b(a|s)} (r(s,a) - \phi(s,a)^\top \hat{\beta}) \right)^2 - n\hat{t}_e^2 \right]$$

where $\hat{t}_e = \sum_D \frac{\pi(a|s)}{n\pi_b(a|s)} (r(s,a) - \phi(s,a)^\top \hat{\beta})$. Särndal et al. (1989) propose the weighted residual technique which can potentially result in better interval estimation. See Appendix C for a derivation.

Finally, we show the variance estimator is consistent and the regression-assisted estimator is asymptotically normal.

Theorem 3. The variance estimator $\hat{\text{V}}(\hat{t}_{\text{Reg}})$ is consistent, and $\frac{\hat{t}_{\text{Reg}} - t_y}{\sqrt{\hat{\text{V}}(\hat{t}_{\text{Reg}})}} \xrightarrow{D} \mathcal{N}(0,1)$.

Based on Theorem 3, we can construct a large sample CI.

Corollary 1. Let $\hat{\sigma} = \sqrt{\hat{\mathrm{V}}(\hat{t}_{\mathrm{Reg}})}$ and z_α denote the $100(1-\alpha)$ percentile of the standard normal distribution, then

$$\Pr\left(\hat{t}_{\mathrm{Reg}} - z_{\alpha/2}\hat{\sigma} \leq J(\pi) \leq \hat{t}_{\mathrm{Reg}} + z_{\alpha/2}\hat{\sigma}\right) \to 1-\alpha.$$

6 EXPERIMENTS

In this section, we demonstrate the effectiveness of the regression-assisted DR estimators in a semi-synthetic and a real world recommendation environment. We compare the proposed estimator to existing estimators, including IS, WIS, DM and Diff (which is DR without estimating the population total). We also include the IS, WIS, DM with the sliding window (SW) approach of window size B. When $B = 0$, SW-IS and SW-WIS is the standard IS and WIS. For Diff, Reg, we use the past data $D_{k-B}, \ldots, D_{k-1}$ to learn a reward prediction.

For the semi-synthetic dataset, we follow the experimental design from Dudík et al. (2011). We use the supervised-to-bandit conversion to construct a partially labeled dataset from the YouTube dataset in the LibSVM repository. We construct a nonstationary environment by generating a sequence of reward functions based on the environment design in Chandak et al. (2020). For each true positive context-action pair in the original classification dataset, the reward follows a sine wave with some noise over time. We use PCA to reduce the dimension of the context features to 32. The target policy is obtained by training a classifier on a small subset of the original classification dataset.

We adapt the Movielens25m dataset (Harper and Konstan, 2015) for the real world experiment. To construct a nonstationary environment, we divide the rating data chronologically. Each interval contains the rating data for 60 days and we use the rating data for $K = 24$ intervals ending November 21, 2019. We consider only active users who gave at least one rating for at least half of the K intervals, resulting in a total number of around 2000 users. During each interval, we compute the rating matrix $r(u, g)$ for each user and genre by averaging the user u's rating for all rated movies in the genre g. As a result, we have a sequence of rating functions which represent users' average rating for each genre over time. The user features are built by matrix factorization on the average rating data with hidden size 32, and the target policy is obtained by training a classifier on a small subset of the average rating data.

For the OPE objective, we consider an uniform weighting $P(s) = 1/|\mathcal{S}|$ for all users s. We also let $n_k = \alpha|\mathcal{S}|$ for all k and $\alpha \in \{0.1, 1.0\}$. For each interval $k = 0, \ldots, K$, we sample data D_k using a random policy. For estimators that require a reward prediction, we build the reward prediction by linear regression on historical data for each action separately, which is the same approach used in Dudík et al. (2011). More experiment details can be found in Appendix E, and Algorithm 1 describes the experimental procedure.

In nonstationary OPE, the aim is to estimate $J_k(\pi)$ with data from D_1 to D_k. All of the estimators discussed in this paper, however, can be extended to predict the future values using the ideas from Thomas et al. (2017); Chandak et al. (2020). Suppose we have the OPE estimators for each interval up to interval k, that is, $\hat{J}_1(\pi), \ldots, \hat{J}_k(\pi)$, we can fit these data to a forecasting model to predict the future value $J_{k+1}(\pi), \ldots, J_{k+\delta}(\pi)$. We therefore test both settings: estimating $J_k(\pi)$ and predicting $J_{k+1}(\pi)$. For the experiments predicting $J_{k+1}(\pi)$, we adapt the method proposed in Chandak et al. (2020) and predict the future values by fitting a regression. That is, $\hat{J}_{k+\delta}(\pi) = \psi(k+\delta)^\top \hat{\mathbf{w}}_k$ where $\hat{\mathbf{w}}_k$ is the OLS estimator for the regression problem with feature map $\psi(t) = (cos(2\pi t n))_{n=0}^{d-1}$ and target $\hat{J}_t(\pi)$ for $t = 1, \ldots, k$, where we set $d = 5$ in the experiment.

Sensitivity to window size and sample size. We vary the window size B and sample size n, and report the sensitivity plot in Figure 1. The error is averaged over K intervals, that is, $\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{J}_k(\pi) - J_k(\pi))^2}$. We can see that the sliding window (SW) estimators, including SW-IS, SW-WIS and SW-DM, are sensitive to the window size, while Diff and Reg are robust to the window size. Reg outperforms IS and WIS and simply using $B = 1$ reduces the error by a large margin. Reg also has a lower error compared to Diff, especially with small window size and sample size. This suggests that Reg is more robust to a bad reward prediction from the past data, which implies it is more robust to the speed of the nonstationarity.

We report the error for predicting the future value $J_{k+1}(\pi)$ in Figure 1. Reg has the lowest error for predicting the future values except in MovieLens with small sample size. We also find that even SW-DM and SW-IS have low error for estimating the current value J_k for some hyperparameters, they still have high error for predicting the future value J_{k+1}. We hypothesize that approximately unbiased estimators generally have better future prediction even though they might have high variance. It is possible that the forecasting model cancels out the noise in approximately unbiased estimators and results in better future value prediction.

Empirical validation of the interval estimation. We use $\hat{J}_k(\pi) \pm 1.96\sqrt{\hat{\mathrm{V}}(\hat{J}_k(\pi))}$ as the approximate 95% CI. We report the empirical coverage of the CI using the estimated variance in Figure 2. The empirical coverage is defined as the number of rounds such that the CI contains the true value divided by the total number of rounds K. The results shown here are with $n = 1.0|\mathcal{S}|$. IS ($B = 0$), WIS ($B = 0$), Diff and Reg all have the desired coverage and Reg has the smallest width. All sliding window estimators have large bias when $B > 0$, so the coverage is poor and it is unclear how to compare to estimators with the desired coverage. Note that even with a small value of B, for example, $B = 1$

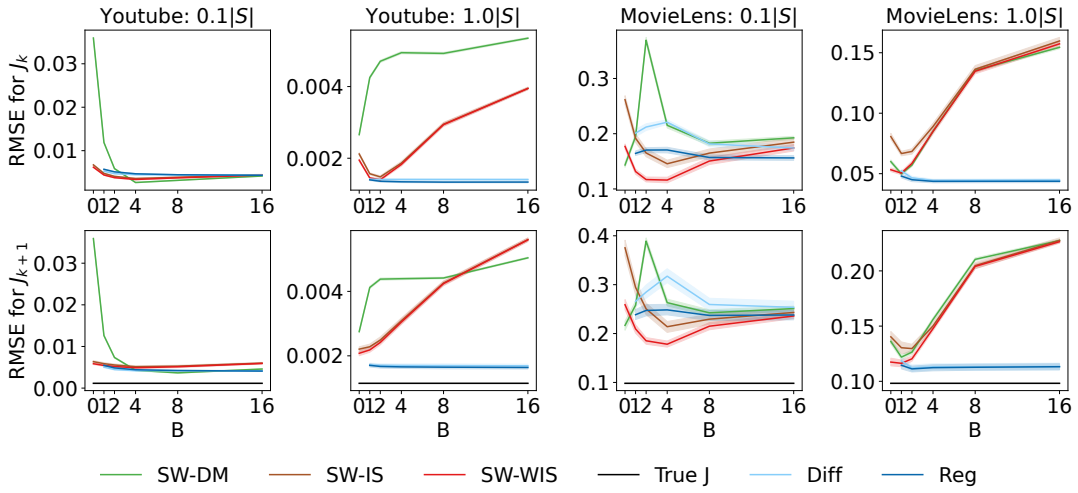

Figure 1: Sensitivity curves. **Top row: estimating $J_k(\pi)$. Bottom row: predicting $J_{k+1}(\pi)$.** "True J" is the baseline if we use the true values to predict the future values. The number are averaged over 30 runs with one standard error. Across runs, the target policy and the sequence of reward functions are fixed, but the sampled data is random.

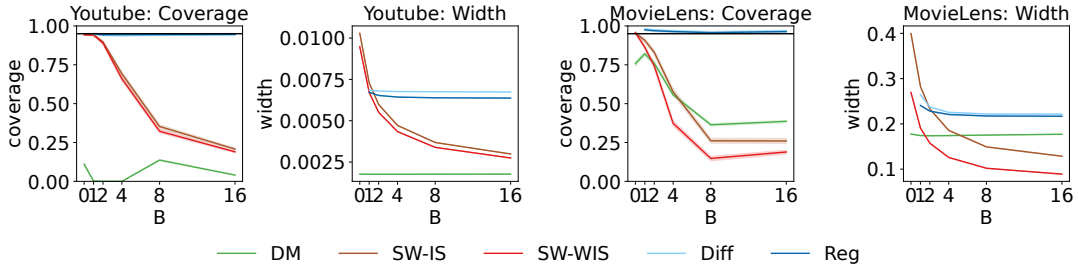

Figure 2: The empirical coverage and the width of CIs. Higher coverage and lower width is better.

in MovieLens, sliding window estimators fail to provide a valid CI. The result suggests that Reg provides an accurate and tight CI.

Empirical investigation of the feature vectors. Besides using one past reward prediction as the only feature, we also investigate the utility when we (1) include the context features; and (2) include separate past reward predictions, that is, $\phi_k(s,a) = (1, \hat{r}_{k-B}(s,a), \ldots, \hat{r}_{k-1}(s,a))$ where we learn a reward model $\hat{r}_t$ for interval t from data D_t separately. Since these additional features could be correlated, we use ridge regression when estimating the coefficients. The regularization parameter is chosen by cross-validation.

We aim to answer two questions: (1) whether including the context feature or the past reward predictions helps, and (2) how we should include the past reward information. To answer the questions, we test five different feature vectors: (a) Reg: we use one past reward prediction as described in Section 4.2, with and without the context features, (b) Reg-AR: we use separate past reward predictions with and without context features, and (c) Reg-Feature: we use context features only. We show the comparison in Figure 3. We find that including the context features helps in general, however, using only the context features is not sufficient. The past reward information helps deal with nonstationarity. Moreover, using separate predictions only improves the accuracy in MovieLens with $n = 0.1|\mathcal{S}|$. In these experiments, there was no one dominant way to include past reward information, and more experimentation is needed to understand when one might be preferred.

We provide an ablation study to investigate the impact when the population total of the proxy value is estimated in Appendix F. We found that using the past data to estimate the population total results in very similar performance as we know the population total.

7 EXTENSION TO REINFORCEMENT LEARNING

The estimators for contextual bandits can be extended to finite horizon reinforcement learning (RL). Let $M = (\mathcal{S}, \mathcal{A}, P, r, H, \nu)$ be a finite horizon finite MDP. Our goal is to estimate the value of a policy $J(\pi) = \sum_{\tau \in (\mathcal{S}\times\mathcal{A})^H} \mathbb{P}^\pi_M(\tau) R(\tau)$ where

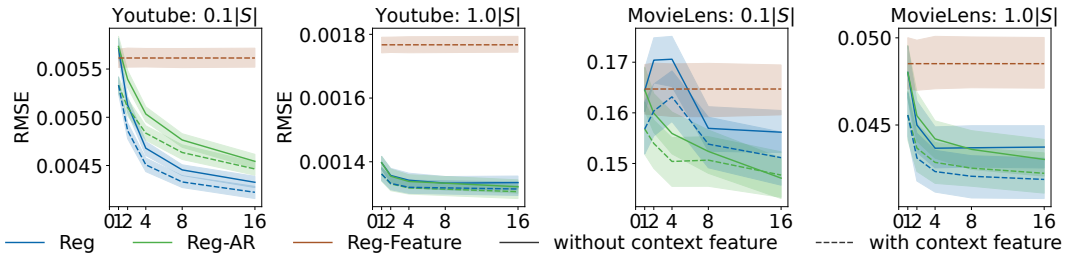

Figure 3: Comparison of different feature vectors for estimating $J_k(\pi)$.

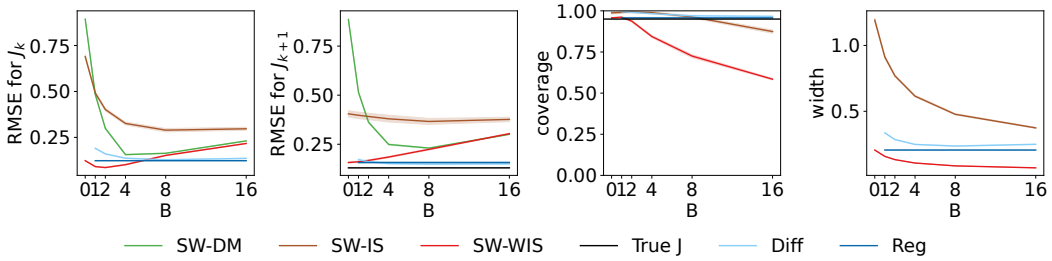

Figure 4: Results for the RL environment. **First column: estimating $J_k(\pi)$. Second column: predicting $J_{k+1}(\pi)$. Third and fourth column: coverage and width of CI**.

$\mathbb{P}_M^\pi(\tau) = \nu(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\dots\pi(a_{H-1}|s_{H-1})$ is the probability of seeing the trajectory $\tau = (s_0, a_0, \dots, s_{H-1}, a_{H-1})$ by running π in M, and $R(\tau) = \sum_{h=0}^{H-1} r(s_h, a_h)$.

To formalize OPE for RL under survey sampling, let $\mathcal{U} = (\mathcal{S}\times\mathcal{A})^H$ be the population containing all trajectories and $y_\tau = \mathbb{P}_M^\pi(\tau)R(\tau)$ be the study variable. Note that there are many ways to view OPE for RL in the survey sampling framework, which corresponds to different existing estimators for RL such as the trajectory-wise IS, per-decision IS (PDIS) estimator and marginalized IS estimator. We provide more discussion in Appendix D.

The regression-assisted estimator with fitted Q evaluation. We use fitted Q evaluation (FQE), which has been shown to be effective for several stationary OPE benchmarks empirically (Voloshin et al., 2021), to build a proxy value $\hat{R}(\tau)$ for each trajectories $\tau \in \mathcal{U}$.

In nonstationary environments, FQE outputs $\hat{Q}_{k-1}(s,a)$ from the past data $D_{k-B}, \dots, D_{k-1}$, and we use $\hat{R}(\tau) = \hat{V}_{k-1}(s_0) = \sum_{a\in\mathcal{A}}\pi(a|s_0)\hat{Q}_{k-1}(s_0,a)$ as the proxy value where s_0 is the initial state of the trajectory τ. Similar to the estimator for contextual bandits, we first estimate the coefficient with the feature vector $\phi(s_0)^\top = (1, \hat{V}_{k-1}(s_0))$ and use the regression-assisted DR estimator

$$\hat{t}_{Reg\text{-}FQE,k} = \sum_{s_0\in\mathcal{S}} \nu(s_0)\phi(s_0)^\top\hat{\beta}_k + \frac{1}{n}\sum_{\tau\in D_k}\frac{\pi(a_0|s_0)\dots\pi(a_{H-1}|s_{H-1})}{\pi_b(a_0|s_0)\dots\pi_b(a_{H-1}|s_{H-1})}(R(\tau) - \phi(s_0)^\top\hat{\beta}_k).$$

When ν is unknown, we can estimate the population total of the proxy value by $1/|D'|\sum_{s_0\in D'}\phi(s_0)^\top\hat{\beta}$ from the past data D' or the same data D_k. The regression-assisted DR estimator can be viewed as a biased-corrected FQE estimator for nonstationary environments.

Experimental results. We consider an RL environment with a binary tree structure, that is, a finite horizon MDP with $H = 10$, $|\mathcal{A}| = 2$, $|\mathcal{S}| = |\mathcal{A}|^H - 1$, and an initial state s_0. For each state, taking action 1 leads to the left child and taking action 2 leads to the right child. The reward for each state-action pair follows a sine wave with different frequency and amplitude. The environment mimics the session-aware recommendation problem where we take a sequence of actions for one customer during a short session. We use a random policy to collect 10 trajectories for every interval. The target policy is a trained policy using Q-learning on the underlying environment.

From Figure 4, Reg has the lowest error for estimating the current value and predicting the future value in general. We show the coverage of the one-sided CI since we mainly care about the lower bound on the policy value for safe policy improvement. The results show that Reg again provides a valid and tight interval estimation, and is promising for safe policy improvement in nonstationary RL environments.

8 CONCLUSION

We proposed the regression-assisted DR estimator for OPE in the nonstationary setting, inspired by estimators from the survey sampling literature. The estimator incorporates

past data into a proxy value without introducing large bias, and uses a regression approach to build a reward prediction well suited for nonstationary environments. As far as we know, these two ideas have not been applied to nonstationary OPE. We theoretically show that we can construct a large sample confidence interval and empirically demonstrate that the proposed estimator provides tight and valid high-confidence estimation in several recommendation environments in contextual bandits and finite horizon RL.

References

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 2013.

F Jay Breidt and Jean D Opsomer. Local polynomial regression estimators in survey sampling. *Annals of statistics*, 2000.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 1976.

Ray Chambers and Robert Clark. *An introduction to model-based survey sampling with applications*. Oxford University Press, 2012.

Yash Chandak, Georgios Theocharous, Shiv Shankar, Martha White, Sridhar Mahadevan, and Philip Thomas. Optimizing for the future in non-stationary mdps. In *International Conference on Machine Learning*, 2020.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. In *Neural Information Processing Systems*, 2021.

William G. Cochran. *Sampling Techniques, 3rd Edition.* John Wiley, 1977.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Sample-efficient nonstationary policy evaluation for contextual bandits. In *Uncertainty in Artificial Intelligence*, 2012.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, 2018.

Martín H Félix-Medina. Asymptotics in adaptive cluster sampling. *Environmental and Ecological Statistics*, 2003.

John M Hammersley and David C Handscomb. *Monte Carlo Methods*. Springer Dordrecht, 1964.

Morris H Hansen and William N Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 1943.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 2015.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, and Amr Ahmed. Non-stationary off-policy optimization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.

Rolf Jagerman, Ilya Markov, and Maarten de Rijke. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *International Conference on Web Search and Data Mining*, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Nathan Kallus and Masatoshi Uehara. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Neural Information Processing Systems*, 2019.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *International Conference on Web Search and Data Mining*, 2011.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Neural Information Processing Systems*, 2018.

A Rupam Mahmood, Hado P Van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Neural Information Processing Systems*, 2014.

Kelly McConville. *Improved estimation for complex surveys using modern regression techniques*. PhD thesis, Colorado State University, 2011.

Kelly S McConville, F Jay Breidt, Thomas Lee, and Gretchen G Moisen. Model-assisted survey regression

estimation with the lasso. *Journal of Survey Statistics and Methodology*, 2017.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 2000.

Carl-Erik Särndal, Bengt Swensson, and Jan H Wretman. The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 1989.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer New York, 1992.

Amode R Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 1953.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, 2020.

Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT press Cambridge, 1998.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, 2015a.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems*, 2015b.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015a.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, 2015b.

Philip S Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *IAAI Conference*, 2017.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Shu Yang and Jae Kwang Kim. Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 2020.

Frank Yates and P Michael Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1953.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *International Conference on Machine Learning*, 2009.

A OVERVIEW OF SURVEY SAMPLING

In this section, we introduce the survey sampling terminology and how to use it for OPE. Survey sampling can be dated back to Hansen and Hurwitz (1943); Horvitz and Thompson (1952), where they consider the problem of selecting a sample of units from a finite population to estimate unknown population parameters. For example, if the goal is to estimate the customer satisfaction rate for a product, survey sampling is concerned with selecting a subset of customers to conduct surveys. Since then, the field has investigated a variety of practical scenarios, including dealing with missing data, handling non-stationarity and understanding to how to leverage auxiliary information.

Formally, let $\mathcal{U} = \{1, \dots, N\}$ be the population of interest, y_i be the study variable and $\mathbf{x}_i$ be the auxiliary variable for the unit $i \in \mathcal{U}$. Continuing the above example, the population could be all customers, the study variable could be the satisfaction level, and the auxiliary variable could be the information about the customer. A subset of the population, called a sample, is selected according to a sampling design. We observe the study variable for units in the sample, and the goal is to estimate the population total of the study variables $t_y = \sum_{i \in \mathcal{U}} y_i$.

A sampling design $\mathbf{I} = (I_1, \dots, I_N)$ is a random vector describing how the sample is drawn from the population: $I_i > 0$ means that the unit i is selected in the sample and $I_i = 0$ means the unit is not selected. For example, a multinomial design is a with-replacement and fixed-size design where we draw n units independently and identically according to probability p_i with $\sum_{i \in \mathcal{U}} p_i = 1$. In this case, the design vector $\mathbf{I}$ follows the multinomial distribution with parameters n and $(p_i)_{i=1}^N$, that is, $P(\mathbf{I}_1 = i_i, \dots, \mathbf{I}_N = i_N) = \frac{n!}{\Pi_{i=1}^N i_i!} p_i^{i_i} \dots p_N^{i_N}$ if $\sum_i i_i = n$ and 0 otherwise. In survey sampling, the study variable is fixed and the randomness comes from the sampling design $\mathbf{I}$.

Given a sample D of fixed size n, the Hansen-Hurwitz (HH) estimator (Hansen and Hurwitz, 1943) for multinomial design is $\hat{t}_{\text{HH}} = \sum_{i \in D} \frac{y_i}{\mathrm{E}[I_i]} = \sum_{i \in D} \frac{y_i}{np_i}$. The estimator $\hat{t}_{\text{HH}}$ is an unbiased estimator for t_y if $p_i > 0$ for all $i \in \mathcal{U}$.

This formalize OPE under survey sampling, let the population be $\mathcal{U} = \mathcal{S} \times \mathcal{A}$ and the study variable be $y_{s,a} = P(s)\pi(a|s)r(s,a)$. The population total of y is the value of the policy: $t_y = \sum_{(s,a) \in \mathcal{U}} y_{s,a} = J(\pi)$. The weighting $P(s)\pi(a|s)$ goes into the study variable since the goal is to estimate the total of study variable without weighting. Even though we have $P(s)$ in the study variable, the term often cancels out as we will see for the HH estimator.

For OPE, the sampling design is the multinomial design with sampling probability $p_{s,a} = P(s)\pi(a|s)$. Given a sample $D = \{(s_i, a_i, r(s_i, a_i))\}_{i=1}^n$, the HH estimator is

$$\hat{t}_{\text{HH}} = \sum_{(s,a) \in D} \frac{y_{s,a}}{np_{s,a}} = \sum_{(s,a) \in D} \frac{P(s)\pi(a|s)r(s,a)}{nP(s)\pi_b(a|s)} = \frac{1}{n} \sum_{(s,a) \in D} \frac{\pi(a|s)}{\pi_b(a|s)} r(s,a).$$

It has the same form as the IS estimator. In the case where the sampling design p is not known and needs to be estimated by a propensity model, it is called the inverse propensity score (IPS) estimator.

The HH estimator is called the *design-based* estimator in the survey sampling literature. This is because the primary source of randomness is from the sampling design. Another approach is called the *model-based* approach which assumes the study variables are generated by a superpopulation model. The goal is to model the relationship between the study variable and the auxiliary variable. The resulting estimator is similar to the direct method.

The model-based approach. The model-based approach is a popular approach in the survey sampling literature. Chambers and Clark (2012) provide an introduction for the model-based approach. Different from design-based and model-assisted approaches, the study variables are assumed to be generated by a superpopulation model and typically depends on the auxiliary variable. More previously, we assume the values $y_i, \dots, y_n$ are realization of random variables $Y_1, \dots, Y_n$. The joint distribution of $Y_1, \dots, Y_n$ is denoted by ξ, which is called the superpopulation distribution. For example, we assume $\mathrm{E}_\xi[Y_i|\mathbf{x}_i] = \mathbf{x}_i^\top \beta$ and $\mathrm{V}_\xi(Y_i|\mathbf{x}_i) = \sigma_i^2$ for some unknown model parameter β and σ_i. The selected sample D is treated as a constant and the sample values of y_i are random. Estimation and inference are deduced conditional on the selected sample and the model.

For OPE, we have $y_{s,a} = P(s)\pi(a|s)r(s,a)$ and auxiliary vector $\mathbf{x}_{s,a} = P(s)\pi(a|s)\phi(s,a)$. We assume a linear model: $\mathrm{E}_\xi[Y_{s,a}|\mathbf{x}_{s,a}] = \mathbf{x}_{s,a}^\top \beta$, $\mathrm{V}_\xi(Y_{s,a}|\mathbf{x}_{s,a}) = \sigma_{s,a}^2 = (P(s)\pi(a|s)\sigma)^2$, and $Y_{s,a}$'s are independent. Using the WLS estimator to estimate β

$$\hat{\beta} = \left(\sum_{(s,a) \in D} \frac{\mathbf{x}_{s,a}\mathbf{x}_{s,a}^\top}{\sigma_{s,a}^2} \right)^\dagger \left(\sum_{(s,a) \in D} \frac{\mathbf{x}_{s,a} y_{s,a}}{\sigma_{s,a}^2} \right) = \left(\sum_{(s,a) \in D} \phi(s,a)\phi(s,a)^\top \right)^\dagger \left(\sum_{(s,a) \in D} \phi(s,a) r(s,a) \right),$$

we have the model-based estimator

$$\hat{t}_{\text{MB}} = \sum_{(s,a)\in D} y_{s,a} + \sum_{(s,a)\notin D} \mathbf{x}_{s,a}^\top \hat{\beta}.$$

That is, the population total is estimated by the total of study variables in the sample and the total of the study variables of units not in the sample.

The model-based estimator is similar to the direct method (DM) in OPE. The key difference is that DM does not use the sample value of $y_{s,a}$ but uses the prediction for all units, that is,

$$\hat{t}_{\text{DM}} = \sum_{(s,a)\in \mathcal{U}} \mathbf{x}_{s,a}^\top \hat{\beta}.$$

The model-based survey sampling framework provide a way to do inference for the DM estimator, which is conditional on the selected sample and the model ξ. Let $t_{\mathbf{x}} = \sum_{(s,a)\in\mathcal{U}} \mathbf{x}_{s,a}$, then

$$\mathrm{V}_\xi(\hat{t}_{\text{DM}}) = \mathrm{V}_\xi\left(\sum_{(s,a)\in\mathcal{U}} \mathbf{x}_{s,a}^\top \hat{\beta}\right) = t_{\mathbf{x}}^\top \mathrm{V}_\xi(\hat{\beta}) t_{\mathbf{x}} = \sigma^2 t_{\mathbf{x}}^\top \left(\sum_{(s,a)\in D} \phi(s,a)\phi(s,a)^\top\right)^\dagger t_{\mathbf{x}}.$$

Plugging in the estimator $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{(s,a)\in D}(r(s,a) - \phi(s,a)^\top\hat{\beta})^2$ for σ, we have an estimated variance

$$\hat{\mathrm{V}}(\hat{t}_{\text{DM}}) = \hat{\sigma}^2 t_{\mathbf{x}}^\top \left(\sum_{(s,a)\in D} \phi(s,a)\phi(s,a)^\top\right)^\dagger t_{\mathbf{x}}.$$

B TECHNICAL DETAILS

For the theoretical analysis, we make the following assumptions:

1. $\forall(s,a)\in\mathcal{U}$, $L_p \le p_{s,a}$ for some real number $L_p > 0$.
2. $\forall(s,a)\in\mathcal{U}$, $L_y \le y_{s.a} \le U_y$ for some real number L_y, U_y.
3. $\forall(s,a)\in\mathcal{U}$, $L_x \le \phi(s,a) \le U_x$ for some real vector L_x, U_x. The inequality holds element-wise.
4. The estimated matrix of the covariates $\sum_{(s,a)\in D} \frac{\pi(a|s)}{\pi_b(a|s)}\phi(s,a)\phi(s,a)^\top$ and the finite population matrix $\sum_{(s,a)\in\mathcal{U}} \phi(s,a)\phi(s,a)^\top$ are invertible.

In short, we need to make sure the data collection policy chooses each action with a non-zero probability, and the reward and feature vector are bounded.

B.1 Proof of Theorem 1

Definition 1 (The ratio estimator)**.** Let $z_{s,a} \in \mathbb{R}$ be the auxiliary variable, t_z be the populating total of the auxiliary variable, which is assumed to be known, and $\hat{t}_{\text{HH}}$ and $\hat{t}_z$ be the HH estimator for t_y and t_z respectively. The ratio estimator is given by

$$\hat{t}_{\text{Ratio}} = t_z \frac{\hat{t}_{\text{HH}}}{\hat{t}_z}.$$

Now we show that the ratio estimator is a special case of the regression estimator.

Lemma 1. Suppose we have univariate auxiliary information $z_{s,a}$. Under the linear model $\mathrm{E}_\xi[Y_{s,a}] = \beta z_{s,a}$ and $\mathrm{V}_\xi(Y_{s,a}) = \sigma_{s,a}^2 = z_{s,a}\sigma^2$ for some $\beta \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$, the regression estimator is equivalent to the ratio estimator.

Proof. First note that the regression estimator has an alternative expression as

$$\begin{aligned}
\hat{t}_{\text{Reg}} &= \sum_{(s,a)\in\mathcal{U}} z_{s,a}\hat{\beta} + \sum_{(s,a)\in D} \frac{y_{s,a} - z_{s,a}\hat{\beta}}{np_{s,a}} \\
&= \sum_{(s,a)\in D} \frac{y_{s,a}}{np_{s,a}} + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a} - \sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)\hat{\beta} \\
&= \sum_{(s,a)\in D} \frac{y_{s,a}}{np_{s,a}} + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a} - \sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}z_{s,a}}{np_{s,a}\sigma^2_{s,a}}\right)^{-1}\left(\sum_{(s,a)\in D} \frac{z_{s,a}y_{s,a}}{np_{s,a}\sigma^2_{s,a}}\right) \\
&= \sum_{(s,a)\in D} \frac{y_{s,a}}{np_{s,a}} \underbrace{\left[1 + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a} - \sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}z_{s,a}}{np_{s,a}\sigma^2_{s,a}}\right)^{-1} \frac{z_{s,a}}{\sigma^2_{s,a}}\right]}_{g_{s.a}} \\
&= \sum_{(s,a)\in D} \frac{g_{s,a}y_{s,a}}{np_{s,a}}
\end{aligned}$$

where $g_{s,a}$ can be viewed as the weight for each unit in the sample.

Under the model $\sigma^2_{s,a} = z_{s,a}\sigma^2$, for each $(s',a') \in D$, we have

$$\begin{aligned}
g_{s',a'} &= 1 + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a} - \sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}z_{s,a}}{np_{s,a}\sigma^2 z_{s,a}}\right)^{-1}\left(\frac{z_{s',a'}}{\sigma^2 z_{s',a'}}\right) \\
&= 1 + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a} - \sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)^{-1} \\
&= 1 + \left(\sum_{(s,a)\in\mathcal{U}} z_{s,a}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)^{-1} - 1 \\
&= t_z \Big/ \left(\sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)
\end{aligned}$$

The second equality follows by cancelling out the right most term with the σ^2 in the inverse bracket. Note that the weight is the same for each unit in the sample. Plugging into the previous equation, we have

$$\hat{t}_{\text{Reg}} = t_z \frac{\sum_{(s,a)\in D} y_{s,a}/np_{s,a}}{\sum_{(s,a)\in D} z_{s,a}/np_{s,a}} = \hat{t}_{\text{Ratio}}.$$

□

Proof of Theorem 1. We first show that the WIS estimator belongs to a class of estimators called the ratio estimator in survey sampling in Definition 1. Suppose the auxiliary variable $z_{s,a} = P(s)\pi(a|s)$, and we know $t_z = \sum_{(s,a)\in\mathcal{U}} P(s)\pi(a|s) = 1$. Then, the ratio estimator is

$$\hat{t}_{\text{Ratio}} = t_z \frac{\hat{t}_{\text{HH}}}{\hat{t}_z} = \left(\sum_{(s,a)\in D} \frac{y_{s,a}}{np_{s,a}}\right)\left(\sum_{(s,a)\in D} \frac{z_{s,a}}{np_{s,a}}\right)^{-1} = \sum_{(s,a)\in D} \frac{\pi(a|s)/\pi_b(a|s)}{\sum_{(s',a')\in D} \pi(a'|s')/\pi_b(a'|s')} r(s,a)$$

which is the WIS estimator in the OPE literature.

Then, we prove a more general statement that a ratio estimator with univariate auxiliary information $z_{s,a}$ is a special case of the regression estimator under the linear model $\mathrm{E}_\xi[Y_{s,a}] = \beta z_{s,a}$ and $\mathrm{V}_\xi(Y_{s,a}) = \sigma^2_{s,a} = z_{s,a}\sigma^2$ for some $\beta \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ in Lemma 1. □

B.2 Proof of Theorem 2

Lemma 2 (Variance of the HH estimator for multinomial design)**.** Let z be a mapping from $\mathcal{S} \times \mathcal{A}$ to $[a, b]$ for two constants $a < b$, and $\hat{t}_{\text{HH}}$ be the HH estimator for the variable $y_{s,a} = P(s)\pi(a|s)z(s,a)$. With multinomial design n and $p_{s,a} = P(s)\pi_b(a|s)$, the variance is given by

$$\mathrm{V}(\hat{t}_{\text{HH}}) = \frac{1}{n}\left(\sum_{(s,a)\in U}\frac{y_{s,a}^2}{p_{s,a}} - t_y^2\right) = \frac{1}{n}\left(\sum_{(s,a)\in U} P(s)\pi(a|s)\frac{\pi(a|s)}{\pi_b(a|s)}z(s,a)^2 - t_y^2\right).$$

Proof. Recall the HH estimator is

$$\hat{t}_{\text{HH}} = \sum_D \frac{y_{s,a}}{np_{s,a}} = \sum_{\mathcal{U}} \frac{I_{s,a}y_{s,a}}{np_{s,a}}.$$

where $I_{s,a}$ is the (s,a)-th element of the design vector $\mathbf{I}$. The variance is

$$\mathrm{V}(\hat{t}_{\text{HH}}) = \mathrm{V}\left(\sum_{\mathcal{U}} I_{s,a}\frac{y_{s,a}}{np_{s,a}}\right) = \sum_{(s,a)\in\mathcal{U}} \mathrm{V}(I_{s,a})\left(\frac{y_{s,a}}{np_{s,a}}\right)^2 + \sum_{(s,a)\neq(s',a')} \mathrm{Cov}(I_{s,a}, I_{s',a'})\left(\frac{y_{s,a}}{np_{s,a}}\right)\left(\frac{y_{s',a'}}{np_{s',a'}}\right).$$

We know $\mathrm{V}(I_{s,a}) = np_{s,a}(1-p_{s.a})$ and $\mathrm{Cov}(I_{s,a}, I_{s',a'}) = -np_{s.a}p_{s',a'}$ from the properties of the multinomial distribution, hence, after some calculation, we have $\mathrm{V}(\hat{t}_{\text{HH}}) = \frac{1}{n}\left(\sum_{(s,a)\in\mathcal{U}}\frac{y_{s,a}^2}{p_{s,a}} - t_y^2\right)$. The proof is completed by plugging in the value of $y_{s,a}$ and $p_{s,a}$. □

Proof of Theorem 2. Note that we assume the first term in Eq (3) is invertible. We can write the estimator as $\hat{t}_{\text{Reg}} = \hat{t}_y + (t_z - \hat{t}_z)\hat{A}^{-1}\hat{C}$ where $\hat{t}_y$ is the HH estimator for $t_y = \sum_{\mathcal{U}} P(s)\pi(a|s)r(s,a)$ and $t_z = \sum_{\mathcal{U}} \mathbf{z}_{s,a}$ and $\hat{t}_z$ be the HH estimator for t_z. $\hat{A}$ and $\hat{C}$ denote the first and second matrix is Eq (3) respectively. Moreover, let t_{zj} be the j-th element of the vector t_z, and $\hat{t}_{zj}$ be the j-th element of the vector $\hat{t}_{zj}$.

Let $A = \sum_{\mathcal{U}} P(s)\pi(a|s)\phi(s,a)\phi(s,a)^\top$, $C = \sum_{\mathcal{U}} P(s)\pi(a|s)\phi(s,a)r(s,a)$ and $B = A^{-1}C$. Using the Taylor linearization technique (see Section 6.6 of Särndal et al. (1992)), and we can approximate $\hat{t}_{\text{Reg}}$ at $\hat{t}_y = t_y$, $\hat{t}_1 = t_1$, $\hat{t}_z = t_z$, $\hat{A} = A$ and $\hat{C} = C$:

$$\begin{aligned}
\hat{t}_{\text{Reg}} &= t_y + 1(\hat{t}_y - t_y) - \sum_j B_j(\hat{t}_{z,j} - t_{z,j}) + \sum_{i,j}(t_z - t_z)^\top[-A^{-1}E_{ij}A^{-1}]C(\hat{A}_{ij} - A_{ij}) + \\
&\quad \sum_j (t_z - \hat{t}_z)^\top e_j(\hat{C}_j - C_j) + \dots \\
&= \hat{t}_y + (t_z - \hat{t}_z)^\top B + ...
\end{aligned}$$

where E_{ij} is a matrix where the ij- and ji-th elements are one and all other elements are zero, and e_j is a vector where the j-th element is one and zero otherwise.

Since the random variable is bounded, the moments exist. Taking the expectation, we get

$$\mathrm{E}[\hat{t}_{\text{Reg}}] = \mathrm{E}[\hat{t}_y + (t_z - \hat{t}_z)^\top B] + O(n^{-1}) = t_y + O(n^{-1}). \tag{6}$$

The first equality follows from the remainder terms of the Taylor expansion are the expectations of $(\hat{t}_y - t_y)^p$ and $(\hat{t}_{z,j} - t_{z,j})^p$ for $p \geq 2$, which is of order $O(1/n)$. The second equality follows from $\hat{t}_z$ is an unbiased estimator for t_z. Therefore, $\hat{t}_{\text{Reg}}$ is asymptotically unbiased.

Furthermore,

$$\begin{aligned}
\mathrm{V}(\hat{t}_{\text{Reg}}) &= \mathrm{E}[(\hat{t}_{\text{Reg}} - \mathrm{E}[\hat{t}_{\text{Reg}}])^2] \\
&= \mathrm{E}[(\hat{t}_{\text{Reg}} - t_y + t_y - \mathrm{E}[\hat{t}_{\text{Reg}}])^2] \\
&= \mathrm{E}[(\hat{t}_{\text{Reg}} - t_y)^2 + (t_y - \mathrm{E}[\hat{t}_{\text{Reg}}])^2 + 2(\hat{t}_{\text{Reg}} - t_y)(t_y - \mathrm{E}[\hat{t}_{\text{Reg}}])] \\
&= \mathrm{E}[(\hat{t}_{\text{Reg}} - t_y)^2] + o(n^{-1}) \\
&= \mathrm{E}[(\underbrace{\hat{t}_y - \hat{t}_z B}_{(a)} - t_y + t_x B)^2] + o(n^{-1})
\end{aligned}$$

The last equality comes from Eq (6). Note that (a) is the HH estimator $\hat{t}_e = \sum_D \frac{\pi(a|s)}{n\pi_b(a|s)}(r(s,a) - \phi(s,a)^\top B)$ so the expectation (the first term in the last line) is the variance of (a) which is given by $\frac{1}{n}\left(\sum_{(s,a)\in U} P(s)\pi(a|s)\frac{\pi(a|s)}{\pi_b(a|s)}(r(s,a)-\phi(s,a)^\top B)^2 - t_e^2\right)$ with $t_e = \sum_{\mathcal{U}} P(s)\pi(a|s)(r(s,a) - \phi(s,a)^\top B)$ by Lemma 2.

Since the variance converges to zero and the estimator is asymptotically unbiased, we also know $\hat{t}_{\text{Reg}} \xrightarrow{p} t_y$. □

B.3 Proof of Theorem 3

Proof of consistency. Define

$$\hat{\mathrm{V}}_n(\beta) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{\pi(a|s)}{\pi_b(a|s)}(r(s,a)-\phi(s,a)^\top\beta)\right)^2 - n\hat{t}_e(\beta)^2\right], \text{ and}$$

$$\mathrm{V}_n(\beta) = \frac{1}{n}\left(\sum_{(s,a)\in U} P(s)\pi(a|s)\frac{\pi(a|s)}{\pi_b(a|s)}(r(s,a)-\phi(s,a)^\top\beta)^2 - t_e(\beta)^2\right)$$

where $\hat{t}_e(\beta) = \sum_D \frac{\pi(a|s)}{n\pi_b(a|s)}(r(s,a)-\phi(s,a)^\top\beta)$ and $t_e(\beta) = \sum_{\mathcal{U}} P(s)\pi(a|s)(r(s,a)-\phi(s,a)^\top\beta)$. Then it is sufficient to show that

$$n|\hat{\mathrm{V}}_n(\hat{\beta}_n) - \mathrm{V}_n(\beta_{WLS})| \xrightarrow{p} 0.$$

For $\epsilon > 0$, by the triangle inequality,

$$\Pr(n|\hat{\mathrm{V}}_n(\hat{\beta}_n) - \mathrm{V}_n(\beta_{WLS})| > \epsilon) \le \Pr(n|\hat{\mathrm{V}}_n(\hat{\beta}_n) - \hat{\mathrm{V}}_n(\beta_{WLS})| > \epsilon/2) + \Pr(n|\hat{\mathrm{V}}_n(\beta_{WLS}) - \mathrm{V}_n(\beta_{WLS})| > \epsilon/2).$$

For the first term, using the fact that $\hat{\beta}_n \xrightarrow{p} \beta_{WLS}$ and the continuous mapping theorem, we get $\hat{\mathrm{V}}_n(\hat{\beta}_n) \xrightarrow{p} \hat{\mathrm{V}}_n(\beta_{WLS})$, which implies $\lim_{n\to\infty}\Pr(n|\hat{\mathrm{V}}_n(\hat{\beta}_n) - \hat{\mathrm{V}}_n(\beta_{WLS})| > \epsilon/2) = 0$.

Define $e_{s,a}(\beta) = r(s,a) - \phi(s,a)^\top\beta$, $t_{we^2}(\beta) = \sum_{\mathcal{U}} P(s)\pi(a|s)\frac{\pi(a|s)}{\pi_b(a|s)}e_{s,a}(\beta)^2$ (the first term of $\mathrm{V}_n(\beta)$) and $\hat{t}_{we^2}(\beta) = \sum_D\left(\frac{\pi(a|s)}{n\pi_b(a|s)}e_{s,a}(\beta)\right)^2$ (the first term of $\hat{\mathrm{V}}(\beta)$). Then, for the second term, we have

$$\begin{aligned}
&\Pr(n\left|\hat{\mathrm{V}}_n(\beta_{WLS}) - \mathrm{V}_n(\beta)\right| > \epsilon/2)\\
&= \Pr(\left|\frac{n}{n-1}\hat{t}_{we^2}(\beta_{WLS}) - \frac{n}{n-1}\hat{t}_e(\beta_{WLS})^2 - t_{we^2}(\beta_{WLS}) + t_e(\beta_{WLS})^2\right| > \epsilon/2)\\
&\le \Pr(\left|\frac{n}{n-1}\hat{t}_{we^2}(\beta_{WLS}) - t_{we^2}(\beta_{WLS})\right| > \epsilon/4) + \Pr(\left|\frac{n}{n-1}\hat{t}_e(\beta_{WLS})^2 - t_e(\beta_{WLS})^2\right| < \epsilon/4).
\end{aligned}$$

Note that $\hat{t}_{we^2}(\beta_{WLS})$ and $\hat{t}(\beta_{WLS})^2$ are the HH estimators for $t_{we^2}(\beta_{WLS})$ and $t_e(\beta_{WLS})$ respectively, so they are consistent. As a result, we have

$$\lim_{n\to\infty}\Pr(n\left|\hat{\mathrm{V}}_n(\hat{\beta}_{WLS}) - \mathrm{V}_n(\beta_{WLS})\right| > \epsilon/2) = 0,$$

which completes the proof. □

Proof of asymptotic normality. It is known that the HH estimator for with-replacement sampling is asymptotically normal (for example, see Theorem 2 of Félix-Medina (2003) or McConville (2011)), that is,

$$\begin{bmatrix}\sqrt{n}(\hat{t}_y - t_y)\\ \sqrt{n}(\hat{t}_z - t_z)\end{bmatrix} \xrightarrow{D} \mathcal{N}\left(0, \begin{bmatrix}\Sigma^y & \mathbf{\Sigma}^{yz}\\ \mathbf{\Sigma}^{zy} & \mathbf{\Sigma}^z\end{bmatrix}\right)$$

where $\Sigma^y, \mathbf{\Sigma}^{yz}, \mathbf{\Sigma}^{zy}$ and $\mathbf{\Sigma}^z$ are the limiting covariance matrices. Then we follow the proof idea from Theorem 3.2 of McConville (2011). Using the Slutsky's Theorem and the fact that $\hat{\beta}_n \xrightarrow{p} \beta_{WLS}$, we have

$$\begin{bmatrix}\sqrt{n}(\hat{t}_y - t_y)\\ \sqrt{n}(\hat{t}_z - t_z)\hat{\beta}_n\end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \begin{bmatrix}\Sigma^y & \mathbf{\Sigma}^{yz}\beta_{WLS}\\ \beta_{WLS}^\top\mathbf{\Sigma}^{zy} & \beta_{WLS}^\top\mathbf{\Sigma}^z\beta_{WLS}\end{bmatrix}).$$

Note that $\sqrt{n}(\hat{t}_{\text{Reg}} - t_y) = \sqrt{n}(\hat{t}_y - t_y) - \sqrt{n}(\hat{t}_z - t_z)\hat{\beta}_n$. By the Delta method, we have $\sqrt{n}(\hat{t}_{\text{Reg}} - t_y) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = \Sigma^y - \Sigma^{yz}\beta_{WLS} - \beta_{WLS}^\top \Sigma^{zy} + \beta_{WLS}^\top \Sigma^z \beta_{WLS}$. Note that we can write the variance of $\hat{t}_y - \hat{t}_z\beta_{WLS}$ as $\text{V}(\hat{t}_y - \hat{t}_z\beta_{WLS}) = \frac{1}{n}(\Sigma^y - \Sigma^{yz}\beta_{WLS} - \beta_{WLS}^\top \Sigma^{zy} + \beta_{WLS}^\top \Sigma^z \beta_{WLS})$, and in the proof for Theorem 2, we show that the asymptotic variance of $\hat{t}_{\text{Reg}}$ is $\text{AV}(\hat{t}_{\text{Reg}}) = \text{V}(\hat{t}_y - \hat{t}_z\beta_{WLS})$. Therefore, $\text{AV}(\hat{t}_{\text{Reg}}) = \sigma^2/n$, and $(\hat{t}_{\text{Reg}} - t_y)/\sqrt{\text{AV}(\hat{t}_{\text{Reg}})} \xrightarrow{D} \mathcal{N}(0,1)$.

By the consistency of the variance estimator and Slutsky's theorem, we have

$$\frac{\hat{t}_{\text{Reg}} - t_y}{\sqrt{\hat{\text{V}}(\hat{t}_{\text{Reg}})}} = \frac{\hat{t}_{\text{Reg}} - t_y}{\sqrt{\text{AV}(\hat{t}_{\text{Reg}})}} \frac{\sqrt{\text{AV}(\hat{t}_{\text{Reg}})}}{\sqrt{\hat{\text{V}}(\hat{t}_{\text{Reg}})}} \xrightarrow{D} \mathcal{N}(0,1).$$

□

C VARIANCE ESTIMATION

In this section, we provide the variance estimation for all estimators used in our experiments.

Variance estimation for the IS estimators. For the IS estimator from the Monte Carlo literature, we first note that the estimator can be written as $\frac{1}{n}\sum_i W_i R_i$ where $W_i = \frac{\pi(A_i|S_i)}{\pi_b(A_i|S_i)}$ and $R_i = r(S_i, A_i)$. Then, the variance is given by $\text{V}(\hat{J}_{\text{IS}}(\pi)) = \frac{1}{n}\text{V}(WR)$ due to the i.i.d. property, and $\text{V}(WR)$ can be estimated by the sample variance. Therefore, we have an unbiased variance estimator

$$\hat{\text{V}}(\hat{J}_{\text{IS}}(\pi)) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(W_i R_i - \hat{J}_{\text{IS}}(\pi)\right)^2\right].$$

For the HH estimator from the survey sampling literature, we can use the Sen-Yates-Grundy variance estimator (Sen, 1953; Yates and Grundy, 1953) for the multinomial design, which is given by

$$\hat{\text{V}}(\hat{t}_{\text{HH}}) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{\pi(a|s)}{\pi_b(a|x)} r(s,a)\right)^2 - n\hat{t}_{\text{HH}}^2\right].$$

The variance estimator $\hat{\text{V}}(\hat{t}_{\text{HH}})$ is an unbiased estimator for the true variance $\text{V}(\hat{t}_{\text{HH}})$. Interestingly, it also has the same form as the variance estimator for the IS estimator.

Variance estimation for the WIS estimator. Using the Taylor linearization technique, the ratio estimator is approximately by $\hat{t}_{\text{Ratio}} = t_x \hat{R} t_x \approx R + (\hat{t}_y - R\hat{t}_x)$. Define $u_{s,a} = y_{s,a} - R x_{s,a}$, $t_u = \sum_{i\in\mathcal{U}} u_{s,a}$ and $\hat{t}_u = \frac{1}{n}\sum_D \frac{u_{s,a}}{p_{s,a}}$, then we have an approximate variance $\text{AV}(\hat{t}_{\text{Ratio}}) = \text{V}(\hat{t}_u)$. Based on the approximation, the estimated variance is given by

$$\hat{\text{V}}(\hat{t}_{\text{WIS}}) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{\pi(a|s)}{\pi_b(a|x)}(r(s,a) - \hat{t}_{\text{WIS}})\right)^2 - n\left(\frac{1}{n}\sum_{(s,a)\in D}\frac{\pi(a|s)}{\pi_b(a|s)}(r(s,a) - \hat{t}_{\text{WIS}})\right)^2\right].$$

Variance estimation for the difference and DR estimator. Since the first term of the difference estimator is fixed, the variance of the difference estimator equals to the variance of the HH estimator $\hat{t}_\Delta = \frac{1}{n}\sum_{(s,a)\in D}\frac{\pi(a|s)}{\pi_b(a|s)}\Delta(s,a)$ where $\Delta(s,a) = r(s,a) - \hat{r}(s,a)$. Then the variance estimator is

$$\hat{\text{V}}(\hat{t}_{\text{Diff}}) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{\pi(a|s)}{\pi_b(a|x)}\Delta(s,a)\right)^2 - n\hat{t}_\Delta^2\right].$$

For the DR estimator where the first term is also estimated from D, first note that the DR estimator can be written as

$$\hat{t}_{\text{DR}} = \frac{1}{n}\sum_{(s,a)\in D}\frac{(\pi(a|x)(r(s,a) - \hat{r}(s,a)) + \pi_b(a|s)\hat{r}_\pi(s))}{\pi_b(a|s)}$$

which is the HH estimator for $t = \sum_{\mathcal{U}} P(s)(\pi(a|s)(r(s,a) - \hat{r}(s,a)) + \pi_b(a|s)\hat{r}_\pi(s))$. Therefore, we have the variance estimator

$$\hat{\mathrm{V}}(\hat{t}_{\text{DR}}) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{(\pi(a|x)(r(s,a) - \hat{r}(s,a)) + \pi_b(a|s)\hat{r}_\pi(s))}{\pi_b(a|s)}\right)^2 - n\hat{t}_{\text{DR}}^2\right].$$

Variance estimation for the regression-assisted estimator. We briefly describe the weighted residual technique from Särndal et al. (1989). Using the definition of $g_{s,a} = 1 + (t_x - \hat{t}_x)(\sum_{i\in D} \frac{\pi(a|s)\phi(s,a)\phi(s,a)^\top}{n\pi_b(a|s)})^{-1}\phi(s,a)$, the regression estimator can be written as

$$\hat{t}_{\text{Reg}} = \sum_{(s,a)\in\mathcal{U}} P(s)\pi(a|s)\phi(s,a)^\top\beta_{WLS} + \sum_{(s,a)\in D}\frac{\pi(a|s)}{n\pi_b(a|s)}g_{s,a}(r(s,a) - \phi(s,a)^\top\beta_{WLS}).$$

It follows that

$$\mathrm{V}(\hat{t}_{\text{Reg}}) = \mathrm{V}\left(\sum_{(s,a)\in D}\frac{\pi(a|s)}{n\pi_b(a|s)}g_{s,a}(r(s,a) - \phi(s,a)^\top\beta_{WLS})\right)$$

which is the variance of the HH estimator $\hat{t}_e = \sum_D \frac{\pi(a|s)}{\pi_b(a|s)}g_{sa}(r(s,a) - \phi(s,a)^\top\beta)$ for $t_e = \sum_{\mathcal{U}} P(s)\pi(a|s)g_{s,a}(r(s,a) - \phi(s,a)^\top\beta)$. Ignoring the fact that the weight $g_{s,a}$ is sample dependent and replacing β_{WLS} with $\hat{\beta}$, we have the g-weighted variance estimator

$$\hat{\mathrm{V}}(\hat{t}_{\text{Reg}}) = \frac{1}{n(n-1)}\left[\sum_{(s,a)\in D}\left(\frac{\pi(a|s)}{\pi_b(a|x)}g_{sa}(r(s,a) - \phi_{s,a}\hat{\beta})\right)^2 - n\hat{t}_e^2\right].$$

D EXTENSION TO RL

In the main paper, we discuss OPE for RL by treating each trajectory as one unit in a population containing all possible trajectories. That is, let $\mathcal{U} = (\mathcal{S}\times\mathcal{A})^H$ be the population containing all trajectories and $y_\tau = \mathbb{P}_M^\pi(\tau)R(\tau)$ be the study variable. We use multinomial design with $p_\tau = \mathbb{P}_M^{\pi_b}(\tau)$ to obtain a sample of trajectories D. The resulting HH estimator has the same form as the trajectory-wise IS estimator (Sutton and Barto, 1998). , that is,

$$\hat{t}_{\text{IS}} = \frac{1}{n}\sum_{\tau\in D}\frac{\mathbb{P}_M^\pi(\tau)}{\mathbb{P}_M^{\pi_b}(\tau)}R(\tau) = \frac{1}{n}\sum_{\tau\in D}\frac{\pi(a_0|s_0)\dots\pi(a_{H-1}|s_{H-1})}{\pi_b(a_0|s_0)\dots\pi_b(a_{H-1}|s_{H-1})}R(\tau).$$

However, there are many other ways to view OPE for RL in the survey sampling framework. One way is to consider estimating the expected reward at each horizon separately. In this case, for $h = 0,\dots,H-1$, let $\mathcal{U}_h = (\mathcal{S}\times\mathcal{A})^{h+1}$ and our goal is to estimate

$$J_h(\pi) = \sum_{\tau\in\mathcal{U}_h}\nu(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\dots\pi(a_h|s_h)r(s_h,a_h)$$

which is the expected reward at horizon h under policy π. It is easy to see that $J(\pi) = \sum_{h=0}^{H-1} J_h(\pi)$. Therefore, this can be viewed as stratified sampling where stratum h contain all trajectories of horizon h. Using the IS estimator to estimate $J_h(\pi)$ for each horizon, we get the per-decision IS estimator (Precup et al., 2000). That is,

$$\hat{t}_{\text{PDIS}} = \sum_{h=0}^{H-1}\hat{J}_h(\pi) = \sum_{h=0}^{H-1}\frac{1}{n}\sum_{\tau\in D}\frac{\pi(a_0|s_0)\dots\pi(a_h|s_h)}{\pi_b(a_0|s_0)\dots\pi_b(a_h|s_h)}r(s_h,a_h). \tag{7}$$

The sampling at each horizon might depend on the sampling at the previous horizon. In that case, we can't easily get the variance or an variance estimator since $V(\hat{J}(\pi)) \neq \sum_{h=0}^{H-1} V(\hat{J}_h(\pi))$. However, we can still compute the variance and an variance estimator via a recursive form (Jiang and Li, 2016).

Another way is to consider sampling transitions instead of episode. Let $\mathcal{U} = \mathcal{S}\times\mathcal{A}$ and for $(s,a)\in\mathcal{U}$, $y_{s,a} = d^\pi(s,a)r(s,a)$ where $d^\pi(s,a) = (\sum_{h=0}^{H-1}\mathbb{P}_M^\pi(S_h = s, a_h = a))/H$. We use multinomial design with $p_{s,a} = \mu(s,a)$ where μ is a data distribution that we can use to sample data. However, for this formulation, $d^\pi(s,a)$ is usually unknown, and there are existing work on estimating the density ratio (Liu et al., 2018).

E EXPERIMENT DETAILS

The goal of the experiment design is to model the recommendation system where reward associated with each user-item pair changes over time. For the Youtube dataset, we generate a sequence of reward functions based on the non-stationary recommendation environment used in Chandak et al. (2020). For each positive context-action pair in the original classification dataset, the reward follows a sine wave with noises. More precisely, $r_k(s,a) = 0.5 + \text{amplitude}_{s,a} * \sin(k * \text{frequency}_{s,a}) + 0.01\varepsilon$ where $\varepsilon \sim \text{Unif}([0,1])$. For each interval, we also randomly sample some context-action pairs and set their rewards to positive random values to increase the noise.

To obtain a target policy for the Youtube and MovieLens dataset, we fist train a classifier on a small subset of the original multi-label classification dataset. Then we apply the softmax function on the outputs of the trained classifiers to obtain a probability distribution over actions for each context. The conditional distribution is used as the target policy.

Similarly for the RL environment, the reward follows $r_k(s,a) = \mu_{s,a} + 0.25 * \sin(k * \text{frequency}_{s,a}) + 0.01 * \varepsilon$ where $\varepsilon \sim \text{Unif}([0,1])$. To obtain a target policy, we fist train a Q-learning agent on the underlying environment for 1000 episodes and then apply the softmax function on the Q-value as the target policy.

The summary statistics of the dataset are:

	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{A}\rvert$
Youtube	31703	47
MovieLens	1923	19

We provide a pseudocode for our experiment procedure in Algorithm 1.

Algorithm 1 Non-stationary OPE experiment with regression-assisted DR estimator

Input: a non-stationary environment M, a target policy π, a behavior policy π_b, window size B, a prediction subroutine $Pred(X, Y, x_{test})$ using linear regression with basis function ψ
for $k = 0, \ldots, K$ **do**
  Collect a dataset $D_k = \{(s_i, a_i, r_k(s_i, a_i), \pi_b(a_i|s_i))\}_{i=1}^n$ from M
  if $k > 0$ **then**
    # Estimate the current value
    Build a reward prediction $\hat{r}_k$ from the past data $D_{k-B}, \ldots, D_{k-1}$
    Compute $\hat{\beta}_k$ from D_k using Eq (3) with $\phi(s,a) = (1, \hat{r}_k(s,a))$
    Compute $\hat{t}_{Reg,k}$ from D_k using Eq (4)
    # Predict the future value
    $\hat{t}_{Pred,k+1} = Pred(X = [\psi(1), \ldots, \psi(k)], Y = [\hat{t}_{Reg,1}, \ldots, \hat{t}_{Reg,k}], x_{test} = \psi(k+1))$
Compute the true value $J_1(\pi), \ldots, J_k(\pi)$
Output:
$\text{RMSE}_{Reg} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{t}_{Reg,k} - J_k(\pi))^2}$ *# Error for estimating the current value*
$\text{RMSE}_{Pred} = \sqrt{\frac{1}{K}\sum_{k=2}^{K}(\hat{t}_{Pred,k} - J_k(\pi))^2}$ *# Error for predict the future value*

F ADDITIONAL EXPERIMENTS

We provide more experiment results in this section.

Estimating the population total of the proxy values. For the experiments in the main paper, we use the population total of the proxy values for the DM, Diff and Reg estimator. In this experiment, we test the regression-assisted estimator when the population total of the proxy values are being estimated, that results in the estimator using Eq (2), which we call RegDR, and the estimator with an independent survey D', described in the last paragraph of Section 4.1, which we call RegDR2.

In Figure 5, we found that RegDR2 has a similar RMSE compared to Reg, and both are slightly better than RegDR. This suggests that using an independent survey has potential to improve the standard DR estimator in the non-stationary setting.

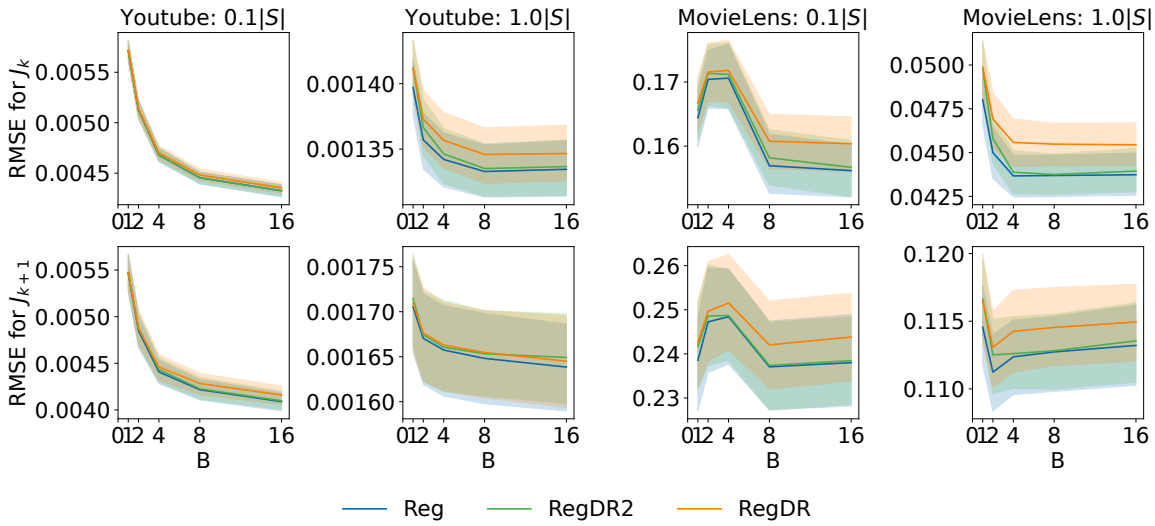

Figure 5: Comparison when the population total of the proxy value is estimated. **Top: estimating $J_k(\pi)$. Bottom: predicting $J_{k+1}(\pi)$.**

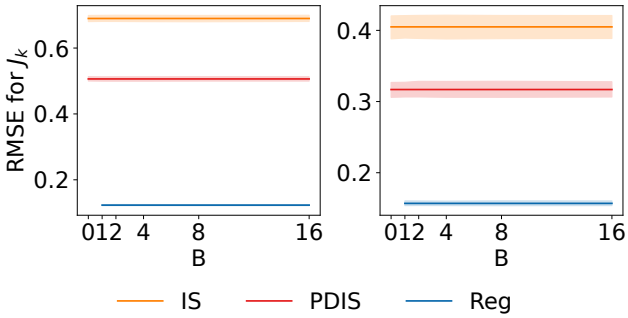

Figure 6: Comparison to PDIS in the simulated RL environment. **Left: estimating $J_k(\pi)$. Right: predicting $J_{k+1}(\pi)$.** Note that PDIS and IS are horizontal lines since they do not depend on B.

Moreover, even the population total needs to be estimated, the regression-assisted estimator can still perform well using RegDR2.

Comparison to PDIS. For the RL experiment, we compare Reg to PDIS in Eq (7), which is expected to improve over the standard trajectory-wise IS estimator. The result in Figure 6 shows that Reg still outperforms PDIS.