Similarities between Nonnested Regression Estimation and Ridge Regression

Caleb Leedy

May 8, 2024

1 Nonnested Regression

In the case of non-nested two phase sampling, we have $A_1 = (X_i)_{i=1}^{n_1}$ and $A_2 = (X_1, Y_i)_{i=1}^{n_2}$ with A_1 and A_2 being selected independently from the same sampling frame. The regression estimator is then

$$\begin{aligned} \hat{\bar{Y}}_{reg} &= \hat{\bar{Y}}_{HT} + (\hat{\bar{X}}_c - \hat{\bar{X}}_2)'\hat{\beta}_2 && \text{for } \hat{\beta}_2 = \left(\sum_{i \in A_2} x_i x_i'\right)^{-1} \sum_{i \in A2} x_i y_i \text{ with } x_{1i} = 1 \\ &= \hat{\bar{Y}}_{HT} + (\hat{\bar{X}}_1 - \hat{\bar{X}}_2)'W'\hat{\beta}_2 && \text{if } \hat{\bar{X}}_c = W\hat{\bar{X}}_1 + (I - W)\hat{\bar{X}}_2 \\ &= \sum_{i \in A_1} x_i' W' \hat{\beta}_2 + \sum_{i \in A_2} (y_i - x_i' W' \hat{\beta}_2). \end{aligned}$$

While the matrix W controls the interaction between A_1 and A_2 it also plays the role of shrinkage on $\hat{\beta}_2$.

If the population total of X is known, then the regression estimator

$$\hat{\bar{Y}}_{reg} = \hat{\bar{Y}}_{HT} + (\bar{X} - \hat{\bar{X}}_2)'\hat{\beta}_2$$

can be used. In this case, $\hat{\beta}_2$ does not have the shrinkage interpretation. Now, if $\bar{X}$ is unknown, we need to use an estimator. If we use the best estimator $\hat{\bar{X}}_c$, then the residual is orthogonal. Otherwise, if we use an inefficient estimator $\hat{\bar{X}}_1$, we pay its price. It is more related to attenuation effect in the measurement error model problem.

2 Ridge Regression

Given a sample A, ridge regression solves the optimization problem,

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in A} (y_i - x_i' \beta)^2 + \beta^T (\lambda I_p) \beta.$$

Differentiating with respect to β and setting this equal to zero yields a solution,

$$\hat{\beta} = \left(\sum_{i \in A} x_i x_i' + \lambda I_p \right)^{-1} \sum_{i \in A} x_i' y_i = (X'X + \lambda I_p)^{-1} X'Y.$$

Let $\hat{\beta}_{OLS} = \hat{\beta}_2 = (X'X)^{-1} X'Y$, then using the Sherman-Morrison-Woodbury inverse formula,

$$\hat{\beta} = (I_p - (X'X)^{-1} (\lambda^{-1} I_p + (X'X)^{-1}))^{-1} \hat{\beta}_{OLS}.$$

3 Discussion

The previous two sections suggest that if we let

$$\begin{aligned} W &= (I_p - (X'X)^{-1} (\lambda^{-1} I_p + (X'X)^{-1}))^{-1} \\ \iff \lambda W &= (X'X)(I_p - W) \\ \iff \lambda &= (X'X)(W^{-1} - I_p) \end{aligned}$$

then we would have equivalent results. In this way, non-nested two phase sampling is like ridge regression.

- We could connect this to LASSO?
- This could be a connection to Bayesian statistics and choosing priors on our estimates?
- Maybe this can help us choose priors for ridge regression better?

4 Alternative idea for non-parametric regression estimation

One possible idea is to develop a debiased estimation under two-phase sampling. If the covariates are high dimensional or the regression employs nonparametric regression (such as spline or random forest, etc), then the resulting two-phase regression estimator can be biased. To correct for the bias, we can use the following approach.

1. Split the sample A_2 into two parts: $A_2 = A_2^{(a)} \cup A_2^{(b)}$. We can use SRS to split the sample, but other sampling designs can be used.

2. Use the observations in $A_2^{(a)}$ only to obtain a predictor of y_i, $\hat{f}^{(a)}(\mathbf{x}_i)$. Also, use the observations in $A_2^{(b)}$ only to obtain a predictor of y_i, $\hat{f}^{(b)}(\mathbf{x}_i)$.

3. Let
$$\hat{f}(\mathbf{x}_i) = \left(\hat{f}^{(a)}(\mathbf{x}_i) + \hat{f}^{(b)}(\mathbf{x}_i)\right)/2$$
be the predictor combining two samples.

4. The final debiased two-phase regression estimator is given by
$$\hat{Y}_{\text{d,reg}} = \sum_{i \in A_1} w_{1i}\hat{f}(\mathbf{x}_i) + \sum_{i \in A_2^{(a)}} w_{1i}\pi_{2i|1}^{-1}\left\{y_i - \hat{f}^{(b)}(\mathbf{x}_i)\right\} + \sum_{i \in A_2^{(b)}} w_{1i}\pi_{2i|1}^{-1}\left\{y_i - \hat{f}^{(a)}(\mathbf{x}_i)\right\} \quad (1)$$

If I am not mistaken, there are several advantages of the debiased two-phase regression estimator in (1).

1. Unlike the classical two-phase regression estimator using nonparametric regression, we can establish asymptotic unbiasedness and $\sqrt{n}$-consistency.

2. Even if we use the sample split, there is no efficiency loss. That is, the asymptotic variance is equal to

$$V\left(\hat{Y}_{\text{d,reg}}\right) = V\left(\hat{Y}_1\right) + E\left[V\left\{\sum_{i \in A_2} w_{1i}\pi_{2i|1}^{-1}\left(y_i - f(\mathbf{x}_i)\right) \mid A_1\right\}\right]$$

where $f(\mathbf{x}_i)$ is the probability limit of $\hat{f}(\mathbf{x}_i)$.

3. Variance estimation is also straightforward. We can compute

$$\hat{\eta}_i = \hat{f}(\mathbf{x}_i) + \delta_i \pi_{2i|1}^{-1} I_i^{(a)} \left\{ y_i - \hat{f}^{(b)}(\mathbf{x}_i) \right\} + \delta_i \pi_{2i|1}^{-1} I_i^{(b)} \left\{ y_i - \hat{f}^{(a)}(\mathbf{x}_i) \right\}$$

and apply to the variance estimation formula for the first-phase sample, where $I_i^{(a)}$ is the indicator function for $A_2^{(a)}$ such that $I_i^{(a)} = 1$ if $i \in A_2^{(a)}$ and $I_i^{(a)} = 0$ otherwise. Also, $I_i^{(b)} = 1 - I_i^{(a)}$.

I also have an idea on how to implement the above debiased regression estimator using calibration. I will give more details once we are confident in the proposed idea.

4.1 Simulation Study

We generate a finite population of size $N = 10,000$ from a superpopulation model of:

$$X_i \overset{ind}{\sim} N(0,1)$$

$$\varepsilon_i \overset{ind}{\sim} N(0,1)$$

$$Y_i = m(x) + 0.3\varepsilon_i$$

for $m(x) = 0.2x + \sin(x)$. The Phase 1 sample is a simple random sample (SRS) of size $n_1 = 2000$ from the finite population and the Phase 2 sample is a Poisson sample with the probability of selection into the Phase 2 sample from the Phase 1 sample being $\max(\min(0.7, \Phi(X_i - 0.5)), 0.02)$ where Φ is the CDF function of a normal distribution. These sampling probabilities yield $E[n_2] \approx 700$.

The goal of this simulation is to estimate $\theta = E[Y]$, which is challenging because Y is a nonlinear function of X as seen in Figure 1 and Figure 2.

We estimate θ using the estimator:

$$\hat{Y} = \sum_{i \in A_1} \frac{\hat{\eta}_i}{\pi_{1i}}$$

where $\hat{\eta}_i$ differs between three methods:

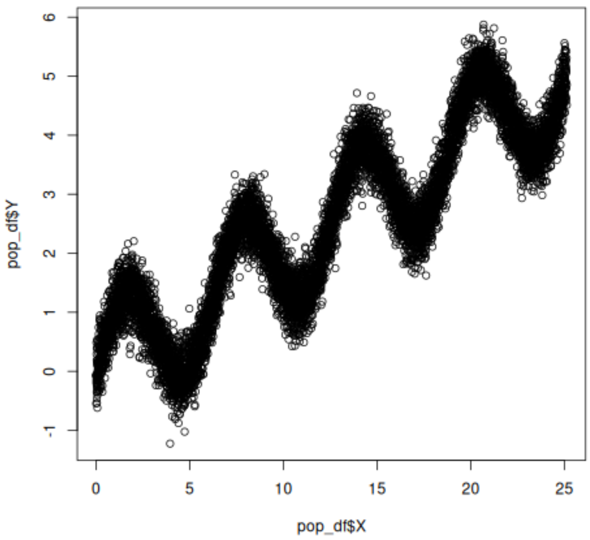

Figure 1: Population plot of relationship between X and Y.

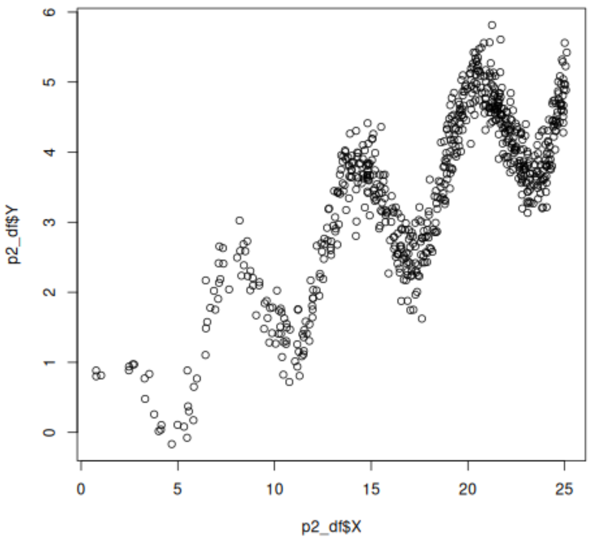

Figure 2: Phase 2 plot of relationship between X and Y.

1. Oracle: $\hat{\eta}_i = m(x_i) + \frac{\delta_{2i|1}}{\pi_{2i|1}}(y_i - m(x_i))$ and $m(x)$ is the true mean function of Y.

2. Naive: $\hat{\eta}_i = \hat{m}(x_i) + \frac{\delta_{2i|1}}{\pi_{2i|1}}(y_i - \hat{m}(x_i))$ where $\hat{m}(x)$ is estimated using the data from the Phase 2 sample using nonparametric regression.

3. Proposed: $\hat{\eta}_i = \hat{m}(x_i) + \frac{\delta_{2i|1}\delta_i^{(a)}}{\pi_{2i|1}}(y_i - \hat{m}^{(b)}(x_i)) + \frac{\delta_{2i|1}\delta_i^{(b)}}{\pi_{2i|1}}(y_i - \hat{m}^{(a)}(x_i))$ where $\hat{m}(x) = \frac{1}{2}(\hat{m}^{(a)}(x) + \hat{m}^{(b)}(x))$ and $\hat{m}^{(a)}(x)$ is the nonparametric model estimated using $A_2^{(a)}$ and $\hat{m}^{(b)}(x)$ is the nonparametric model estimated using $A_2^{(b)}$. We also have $\delta_i^{(a)}$ and $\delta_i^{(b)}$ which denote is $i \in A_2$ was part of $A_2^{(a)}$ or $A_2^{(b)}$ respectively.

For the Naive and Proposed methods, we use three nonparametric regression estimators: loess, random forest, and a smoothing spline. These functions come from the `stats`, `randomForest`, and `npreg` packages in `R`. To estimate the variance we use the following:

$$\hat{V} = \hat{S}_\eta^2 / n_1$$

where $\hat{S}_\eta^2$ is the estimated variane of $\hat{\eta}$ in A_1.

The results of this simulation for the mean are displayed in Table 1 and the variance results are in Table 2.

Est	Bias	RMSE	EmpCI	Ttest
Oracle	0.000	0.038	0.942	0.030
NaiveLoess	-0.006	0.061	0.923	2.985
PropLoess	0.000	0.062	0.944	0.023
NaiveRF	0.001	0.042	0.885	0.681
PropRF	-0.001	0.047	0.948	0.373
NaiveSpline	0.010	0.052	0.857	6.242
PropSpline	0.001	0.106	0.924	0.269

Table 1: This shows the results of our simulation. The Est column displays the name of the estimator with examples of NaiveLoess indicating the Naive method with the loess nonparametric estimator. The Bias column computes $E_{MC}(\hat{\theta}) - \theta_N$ where E_{MC} is the Monte Carlo mean of the 1000 simulations and θ_N is the true value from the population. The RMSE column computes $\sqrt{B^{-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \theta_N)^2}$ where $\hat{\theta}^{(b)}$ is the estimate of θ_N for the bth iteration of the Monte Carlo sample. The EmpCI column indicates the fraction of iterations for which $|\hat{\theta}^{(b)} - \theta_N| < 1.96 * \sqrt{\hat{V}^{(b)}}$ is true where $\hat{V}^{(b)}$ is the estimated variance of $\hat{\theta}^{(b)}$. The final column of Ttest gives the test statistic of a test for the bias of the estimator being zero.

Overall, these results seem very promising. From Table 1, we can see the there is evidence for bias in the naive algorithm in both the random forest and the spline method. Yet, this bias is no longer there under the proposed method and the empirical confidence intervals

Est	MCVar	EstVar	VarVar	Ttest
Oracle	0.0014	0.0013	0.0000000	16.8
NaiveLoess	0.0037	0.0033	0.0000006	15.4
PropLoess	0.0039	0.0042	0.0000021	6.1
NaiveRF	0.0018	0.0012	0.0000000	143.3
PropRF	0.0022	0.0023	0.0000005	6.0
NaiveSpline	0.0026	0.0013	0.0000000	246.6
PropSpline	0.0113	0.0068	0.0010567	4.4

Table 2: This displays the variance results from the simulation. The MCVar colum is the Monte Carlo variance of $\hat{\theta}^{(b)}$ while the EstVar column is the average of the estimated variance $\hat{V}^{(b)}$. The VarVar column is the Monte Carlo variance of the variance estimator and the Ttest column tests for equality of MCVar and EstVar.

are approximately correct. A further investigation of the variance in Table 2 indicates that the estimated variance for the proposed method is too large when the regression function is either the random forest or smoothing spline, but we might be approximately correct.

For this simulation, we are still doing well. The only problem is that the naive method is also working well (especially for the random forest). Maybe I can choose a more nonlinear function to estimate?

4.1.1 Simulation Update

I updated the simulation so that we use the same approach as Wang and Kim (2023). They use a multivariate approach with X_{1i}, X_{2i}, X_{3i}, and X_{4i} being simulated independently from a $U(1,3)$ distribution. They have three models for Y,

Model	Y
A	$Y_i = 3 + 2.5x_{1i} + 2.75x_{2i} + 2.5x_{3i} + 2.25x_{4i} + \sqrt{3}\varepsilon_i$
B	$Y_i = 3 + (1/35)x_{1i}^2 x_{2i}^3 x_{3i} + 0.1x_{4i} + \sqrt{3}\varepsilon_i$
C	$Y_i = 3 + (1/180)x_{1i}^2 x_{2i}^3 x_{3i} x_{4i}^2 + \sqrt{3}\varepsilon_i$

The first phase sample is a SRS with $n_1 = 1000$ and the second phase sample is a Poisson sample with $\pi_{2i|1} = \text{logistic}(\mathbf{x}_i^T \boldsymbol{\beta} + 2.5)$ for $\boldsymbol{\beta} = (-1.1, 0.5, -0.25, -0.1)^T$. We have modified the spline method to be a cubic spline with 15 knots for each coordinate and implemented by the mgcv package. Table 3 shows the results of the simulation under Model C. The results for Model B are similar.

Est	Bias	RMSE	EmpCI	Ttest
Oracle	-0.002	0.108	0.950	0.584
NaiveLoess	-0.001	0.110	0.942	0.379
PropLoess	-0.001	0.111	0.964	0.241
NaiveRF	-0.020	0.116	0.900	5.466
PropRF	0.004	0.119	0.972	0.930
NaiveSpline	0.000	0.121	0.945	0.014
PropSpline	0.000	0.121	0.961	0.058

Table 3: This shows the results of our simulation. The Est column displays the name of the estimator with examples of NaiveLoess indicating the Naive method with the loess nonparametric estimator. The Bias column computes $E_{MC}(\hat{\theta}) - \theta_N$ where E_{MC} is the Monte Carlo mean of the 1000 simulations and θ_N is the true value from the population. The RMSE column computes $\sqrt{B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \theta_N)^2}$ where $\hat{\theta}^{(b)}$ is the estimate of θ_N for the bth iteration of the Monte Carlo sample. The EmpCI column indicates the fraction of iterations for which $|\hat{\theta}^{(b)} - \theta_N| < 1.96 * \sqrt{\hat{V}^{(b)}}$ is true where $\hat{V}^{(b)}$ is the estimated variance of $\hat{\theta}^{(b)}$. The final column of Ttest gives the test statistic of a test for the bias of the estimator being zero.

Unfortunately, with this simulation (and also the simulation with Model B but not shown here), the naive models are not biased for the loess and spline methods. It does look like we can debias the result for the random forest though.

References

Wang, H. and J. K. Kim (2023). Statistical inference using regularized m-estimation in the reproducing kernel hilbert space for handling missing data. *Annals of the Institute of Statistical Mathematics 75*(6), 911–929.