Review on A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data.

Jingrui Mu
Student ID: 261019374

Abstract

This project aims to study the paper from Joseph D.Y.Kang and Joseph L. Schafer (2007) (KS) [2] and discuss these comment papers. These methods in the literature of Missing data problem can be parallelly implemented in the casual inference. Therefore, I linked the knowledge and ideas in the missing data prolem with what we learned in the lecture. The same simulation study is also implemented and we can get the same conclusions as the paper stated from the simulation study in Section 3. R code is attached in the appendix and MyCourses. In the end, these discussion papers' main ideas are also studied in the Section 4.

1 Introduction

This project aims to review a variety of incomplete-data estimation strategies, describe various ways on how to estimate when we get missing values as Joseph D.Y.Kang and Joseph L. Schafer (2007) [2] discussed. The paper is couched in the language of missing data rather then casual inference. However, all of the methods described and implemented have precise parallels in the casual inference literature as we discussed in the lecture. For the sake of alliance of mathematical language between lectures and the main paper, I will translate the language in the paper to keep them same.

When outcomes are missing for reasons beyond an investigator's control, or when we can just observe the outcomes of patients received treatments, there are two different ways to adjust a parameter estimate for covariates that may be related to the outcome and missingness. One method is to model the relationships between covariates and outcome, which is called $y-$ model or outcome regression in the lecture. And then those relationships can be used to predict the missing values. Another one is to model the probabilities of missingess given the covariates and include them into a weighted estimate. This method is motivated by importance sampling, or change of measure.

Doubly Robust (DR) methods require specification of two models described above. The important feature of DR estimates is that they remain asymptotically unbiased if either of the two models has been correctly specified. The main paper discussed a variety of incomplete-data estimation strategies, such as, Inverse-Propensity Weighting (IPW), Stratification, Propensity Regression Estimation, and Augmented Inverse-Propensity Weighting (APIW). They also investigated the practical behavior of these estimators not only when either of the underlying models is correct, but also in a scenario where both models are moderately mis-specified.

The project reviewed these methods above and connected them to same techniques found in the literature on casual inference. Section 2 describes the problem setup, model assumption, and methodologies. Simulation study will be implemented in Section 3 in use of same data generating procedure. Section 4 discussed the main points raised by the authors in the Discussion and the Rejoinder papers.

2 Methodology

2.1 Problem Set up

In the main paper, they are focusing on the problem of estimating a population mean from an incomplete dataset. They suppose that they have a random sample of units i = 1,2,...,n from an infinite population. The variable of primary interest is y_i, and t_i is the response indicator for y_i.

$$t_i = \begin{cases} 1 \text{if } y_i \text{ is observed} \\ 0 \text{if } y_i \text{ is not observed} \end{cases}$$

For each unit, there is an observed p-dimensional vector of covariates x_i that may be related both to y_i and t_i. In general, it is hard to estimate the mean for nonrespondents, $\mu^{(0)} = E(y_i|t_i = 0)$, or the mean of the entire population based on the observed alone. Naive estimates, such as, the mean The sample mean of observed y_i's, $\bar{y}^{(1)} = \frac{1}{n^{(1)}} \sum_{i=1}^{n} t_i y_i$, may work well enough if the portion of non-responding is small and the relationships among x_i, t_i, and y_i are weak.

This problem is closely related to estimating an average casual effect from observational study. Suppose that there is a pair of potential outcomes associated with unit i: the response y_{i1} that is realized if $t_i = 1$, and another response y_{i0} that is realized if $t_i = 0$. As we discussed in the lecture, we let $Y(z)$ denote the random variable recording the (potential or counterfactual) outcome Y that would be observed if there is an intervention to set $Z = z$. If the treatment is a binary variable, the casual contrast of interest is $E[Y(1) - Y(0)]$, which is known as the average treatment effect (ATE) or it can be the average casual effect (ACE) in the population. We can translate t_i, response indicator, as the binary treatment variable in the lecture.

2.2 Assumptions

Strong Ignorability: Assume that the response mechanism is unconfounded in the sense that y_i and t_i are conditionally independent given x_i. Recall the assumption of strong ignorability in the lecture: $\{Y(1), Y(0)\} \perp\!\!\!\perp Z|X$. Strong ignorability implies that the missing y_i's are missing at random (MAR). Under strong ignorability, the joint distribution of the complete data can be written as

$$P(X, T, Y) = \prod_i P(x_i)P(t_i|x_i)P(y_i|x_i)$$

or

$$P(X, Z, Y) = \prod_i P(x_i)P(z_i|x_i)P(y_i|x_i)$$

MAR is not realistic in many applications. However, this assumption provides an important foundation upon which DR procedures rest.

Assumptions about $P(t_i|x_i)$ and $P(y_i|x_i)$

$$P(t_i|x_i) = \pi_i(x_i) = \pi_i$$

This probability is called the propensity score. A proposed functional form for $P(t_i|x_i)$ will be called $\pi-$ model, or propensity score regression.

$$\hat{\pi}_i = \text{expit}(x_i^T \hat{\alpha}) = \frac{exp(x_i^T \hat{\alpha})}{1 + exp(x_i^T \hat{\alpha})}$$

where $\hat{\alpha}$ is the maximum-likelihood (ML) estimate of the coefficients from the logistic regression of $t_1, ..., t_n$ on $x_1, ..., x_n$. In real applications, sometimes the assumed form of the $\pi-$ model or the propensity score regression is not correct.

Also, there is another issue in casual inference literature: it is not necessary to predict propensity scores very precisely. For example, as the DAG shows:

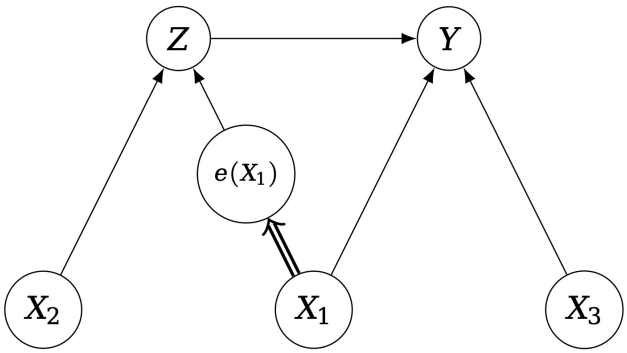


Figure 1

The propensity score does not need to be a function X_2 or X_3; making it depend only on X_1 is sufficient to block the back-door path.

Define $E(y_i|x_i) = m(x_i) = m_i$, so that

$$y_i = m(x_i) + \epsilon_i$$

with $E(\epsilon_i) = 0$. A functional form for $m(x_i)$ will be called a y-model, which is identical to the outcome regression model in casual inference. m_i can be estimated from $x_i\hat{\beta}$, where $\hat{\beta}$ is the vector of coefficients from the linear regression of y_i on x_i from the respondents. In casual inference, we usually write the assumption as $E_X^{\mathcal{O}}[\mu(X,z)] = \mu(z)$. But when we misspecify $y-$ model, the implication can be serious if $P(x_i|t_i = 1)$ and $P(x_i|t_i = 0)$ are very different, because the $\hat{m}_i$'s for the non-respondents will then be based on extrapolation. Therefore, there is another standard assumption behind the $y-$ model or outcome regression: $P(x_i|t_i = 1)$ and $P(x_i|t_i = 0)$ are not very different, or $E_X^{\epsilon}[\mu(X,z)] \equiv E_X^{\mathcal{O}}[\mu(X,z)]$ in the casual inference literature.

2.3 Weighting and Regression

2.3.1 Inverse-Probability Weighting

Strong ignorability implies that

$$E(t_i\pi_i^{-1}y_i) = E(E(t_i\pi_i^{-1}y_i)) = E(\pi_i\pi_i^{-1}m_i) = \mu$$

It is easy to see that $n^{-1}\sum_i t_i\pi_i^{-1}y_i$ unbiasedly estimates the mean of entire population because of the strong ignorability. Precision can be enhanced if a denominator of $\sum_i t_i\pi_i^{-1}$ is used rather than n. So the estimates becomes a weighted average of y_i's for the respondents. By replacing the unknown propensities by estiamates from $\pi-$ model or the propensity regression model, the inverse-propensity weighted (IPW) estimate becomes

$$\hat{\mu}_{IPW-POP} = \frac{\sum_i t_i\hat{\pi}_i{}^{-1}y_i}{\sum_i t_i\hat{\pi}_i{}^{-1}}$$

Alternatively, an estimate of $\mu^{(0)}$, the population of nonrespondents, based on the idea is:

$$\hat{\mu}_{IPW-NR} = \frac{\sum_i t_i\hat{\pi}_i{}^{-1}(1-\hat{\pi})y_i}{\sum_i t_i\hat{\pi}_i^{-1}(1-\hat{\pi})}$$

In casual inference literature, the same idea can be used. The techiniques used are importance sampling, or change of measure.

$$E_{Y|Z}^{\epsilon}[Y|Z=z] = \frac{\iiint I_z(z)yf_{Y|X,Z}^{\epsilon}(y|x,z)f_Z^{\epsilon}(z)f_Z^{\epsilon}(x)dydxdz}{\iiint I_z(z)f_{Y|X,Z}^{\epsilon}(y|x,z)f_Z^{\epsilon}(z)f_Z^{\epsilon}(x)dydxdz}$$

$$\frac{f_{X,Y,Z}^{\epsilon}(x,y,z)}{f_{X,Y,Z}^{\mathcal{O}}(x,y,z)} = \frac{f_X^{\epsilon}(x)}{f_X^{\mathcal{O}}(x)}\frac{f_Z^{\epsilon}(z)}{f_{Z|X}^{\mathcal{O}}(z|x)}\frac{f_{Y|X,Z}^{\epsilon}(y|x,z)}{f_{Y|X,Z}^{\mathcal{O}}(y|x,z)}$$

$$= \frac{f_Z^{\epsilon}(z)}{f_{Z|X}^{\mathcal{O}}(z|x)}$$

We therefore have that

$$E_{Y|Z}^{\epsilon}[Y|Z=z] = \frac{\iiint I_z(z) y \frac{f_Z^{\epsilon}(z)}{f_{Z|X}^{\mathcal{O}}(z|x)} f_{X,Y,Z}^{\mathcal{O}}(x,y,z) dy dx dz}{\iiint I_z(z) \frac{f_Z^{z}(z)}{f_{Z|X}^{\mathcal{O}}(z|x)} f_{X,Y,Z}^{\mathcal{O}}(x,y,z) dy dx dz}$$

Then the formulation is the same as the paper shows since $\hat{\pi}_i$ can be $e(X_i)$ or $1 - e(X_i)$ in the casual inference literature. The IPW estimator can also regarded as the solution to a simple weighted estimating equation $\sum_i w_i U_i = 0$ and $U_i = \frac{y_i - \mu}{\sigma^2}$.

There are two issues in IPW: IPW estimates are very sensitive to misspecification of $\pi-$ model or propensity score regression and IPW methods assigning large weights to respondents who closely resemble nonrespondents will cause estimates to have very high variance. Also, in real applications of IPW methodology, weights are obtained by logistic regression. But logistic models can be a poor way to estimate response propensities because ML estimates from the logistic regression are not resistant to outliers (Pregibon, 1982)[3].

2.4 Doubly Robust (DR) Estimates

2.4.1 Regression Estimation with Residual Bias Correction

Let's consider the outcome regression model holds in the sense that $E(y_i|x_i) = x_i^T \beta$ for some $\beta \in R^p$, then the mean of estimated residuals $\hat{\epsilon_i} = y_i - \hat{m}_i$ in the population will be zero. Residuals are only be seen for sampled respondents. However, the mean residual for the full population can be consistently estimated in use of $\pi-$ model. Cassel, Sanrndal and Wretman proposed the bias-corrected estimate

$$\hat{\mu} = \hat{\mu}_{OLS} + \frac{1}{n} \sum_i t_i \hat{\pi_i}^{-1} \hat{\epsilon_i}$$

A more general version was independently proposed by Robins, Rotnitzky and Zhao (1994)[5]. Suppose U_i is the contribution of sample unit i to a vector-valued quasi-score function for the regression of y_i on a set of covariates.

$$\frac{1}{n} \sum_i \hat{U}_i + \frac{1}{n} t_i \hat{\pi_i}^{-1} (U_i - \hat{U}_i) = 0$$

Which can be compared parallelly with we discussed in casual inference lecture:

$$E_{X,Z}^{\mathcal{O}}[E_{Y|X,Z}^{\mathcal{O}}[\frac{I_z(Z)}{f_{Z|X}^{\mathcal{O}}(z|x)}(Y - \mu(X,Z))|X, Z=z]] + E_X^{\mathcal{O}}[\mu(X,z)] = 0$$

In the procedure, we used IPW technique and the assumption $E_X^{\mathcal{O}}[\mu(X,z)] = E_X^{\epsilon}[\mu(X,z)]$

From the simulation study in the main paper, we can see that the DR procedure is substantially worse than OLS when $\pi-$ model and $y-$ model are misspecified.

2.4.2 Regression Estimation with Inverse-Propensity Weighted Coefficients

In previous subsection, the correction term in $\hat{\mu}$ repairs the bias in $\hat{\mu_{OLS}} = \frac{1}{n} \sum_i x_i^T \hat{\beta}$ by estimating the mean residual in the full population. If based on the full sample, the OLS coefficients will be

$$\hat{\beta} = (\sum_i x_i x_i^T)^{-1} (\sum_i x_i y_i)$$

However, we can not compute it from the observed data. But with propensities estimated from $\pi-$ model, we can compute a weighted least-squares (WLS) estimate

$$\hat{\beta} = (\sum_i t_i \hat{\pi_i}^{-1} x_i x_i^T)^{-1} (\sum_i t_i \hat{\pi_i}^{-1]} x_i y_i)$$

And then the regression estimate for μ based on the WLS coefficients will be

$$\hat{\mu}_{WLS} = \frac{1}{n}\sum_i x_i^T \hat{\beta}_{WLS}$$

$\hat{\mu}_{WLS}$ can be a consistent estimator for μ when either π-model or $y-$ model is true. In the simulated exmpale from the paper, the WLS regression estimate is sometimes inferior to the bias corrected OLS estimate when one of the models is true, but much better when both models are misspecified.

2.4.3 Regression Estimation with Propensity-Based Covariates

A third general strategy for constructing a DR estimate is to incorporate functions of estimated propensities into the y-model or outcome regression as covariates (Joseph D. Y. Kang and Joseph L. Schafer, 2007) [2]. The idea is the same as we saw in the lecture, Scharfstein, Rotnizky and Robbins (1999) [6]proposed using

$$E[Y|X,Z] = \mu(X,Z) + \phi_1(\frac{Z}{e(X)}) + \phi_0(\frac{1-Z}{1-e(X)})$$

The idea behind the model is they constructed orthogonal inverse probability-weighted (AIPW) estimators and it can be seen from estimation equations in kint 11. These two estimating equations imply the orthogonality between $\frac{Z_i}{e(X_i)}$ and $Y_i - \mu(X_i, Z_i) - \phi_1 \frac{Z_i}{e(X_i)}$, $\frac{1-Z_i}{1-e(X_i)}$ and $Y_i - \mu(X_i, Z_i) - \phi_0 \frac{1-Z_i}{1-e(X_i)}$

Also, Robins and Bang (2005)[1] pointed out that using the inverse-propensity as a single additional covariate is sufficient to achieve double robustness. And they proposed

$$E[Y|X,Z] = \mu(X,Z) + \phi(\frac{Z}{e(X)} - \frac{1-Z}{1-e(X)})$$

to produce the ATE estimator:

$$\frac{1}{n}\sum_i^n(\mu(X_i,1) - \mu(X_i,0)) + \hat{\phi}\sum_i^n(\frac{1}{e(X_i)} + \frac{1}{1-e(X_i)})$$

where $\hat{\phi} = \frac{\sum_{i=1}^n(\frac{Z_i}{e(X_i)} - \frac{1-Z_i}{1-e(X_i)})(Y_i - \mu(X_i,Z_i))}{\sum_{i=1}^n(\frac{Z_i}{e(X_i)} - \frac{1-Z_i}{1-e(X_i)})^2}$

The regression estimate of Scharfstein, Rotnizky and Robbins (1999) that uses the inverse propensity as a covariate behaves poorly under a misspecified $\pi-$ model from Table 8 in the main paper. The simulation result keeps the same conclusion as we saw in knit 11.

3 Simulation Study

Consider the following simulated dataset: for each unit $i = 1, 2, ..., n$, suppose that $(z_{i1}, z_{i2}, z_{i3}, z_{i4})^T$ is independently distributed as $N(0, I)$ where I is the 4 x 4 identity matrix. The y_i's are generated as

$$y_i = 210 + 27.4z_{i1} + 13.7z_{i2} + 13.7z_{i3} + 13.7z_{i4} + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$, and the true propensity scores are

$$\pi_i = \text{expit}(-z_{i1} = 0.5z_{i2} - 0.25z_{i3} - 0.1z_{i4})$$

The data generating procedure produces an average response rate of $r^{(1)} = 0.5$, and the means are $\mu = 210.0$, $\mu^{(1)} = 200.0$, and $\mu^{(0)} = 220.0$. A logistic regression of t_i on the z_{ij}'s would be a correct $\pi-$ model, and a linear regression of y_i on the z_{ij}'s would be a correct y-model. However, the covariates actually seen by the data analyst are

$$x_{i1} = exp(z_{i1}/2),$$

$$x_{i2} = z_{i2}/(1 + exp(z_{i1})) = 10,$$
$$x_{i3} = (z_{i1}z_{i3}/25 + 0.6)^3,$$
$$x_{i4} = (z_2 + z_4 + 20)^2$$

This implies logit(π_i) and m_i are linear functions of $log(x_{i1})$, x_2, $x_1^2x_2$, $1/log(x_1)$, $x_3/log(x_1)$ and $x_4^{1/2}$. Scatterplots of y_i versus x_{ij}, j = 1,2,3,4, for the 100 respondents are shown in Figure 1.

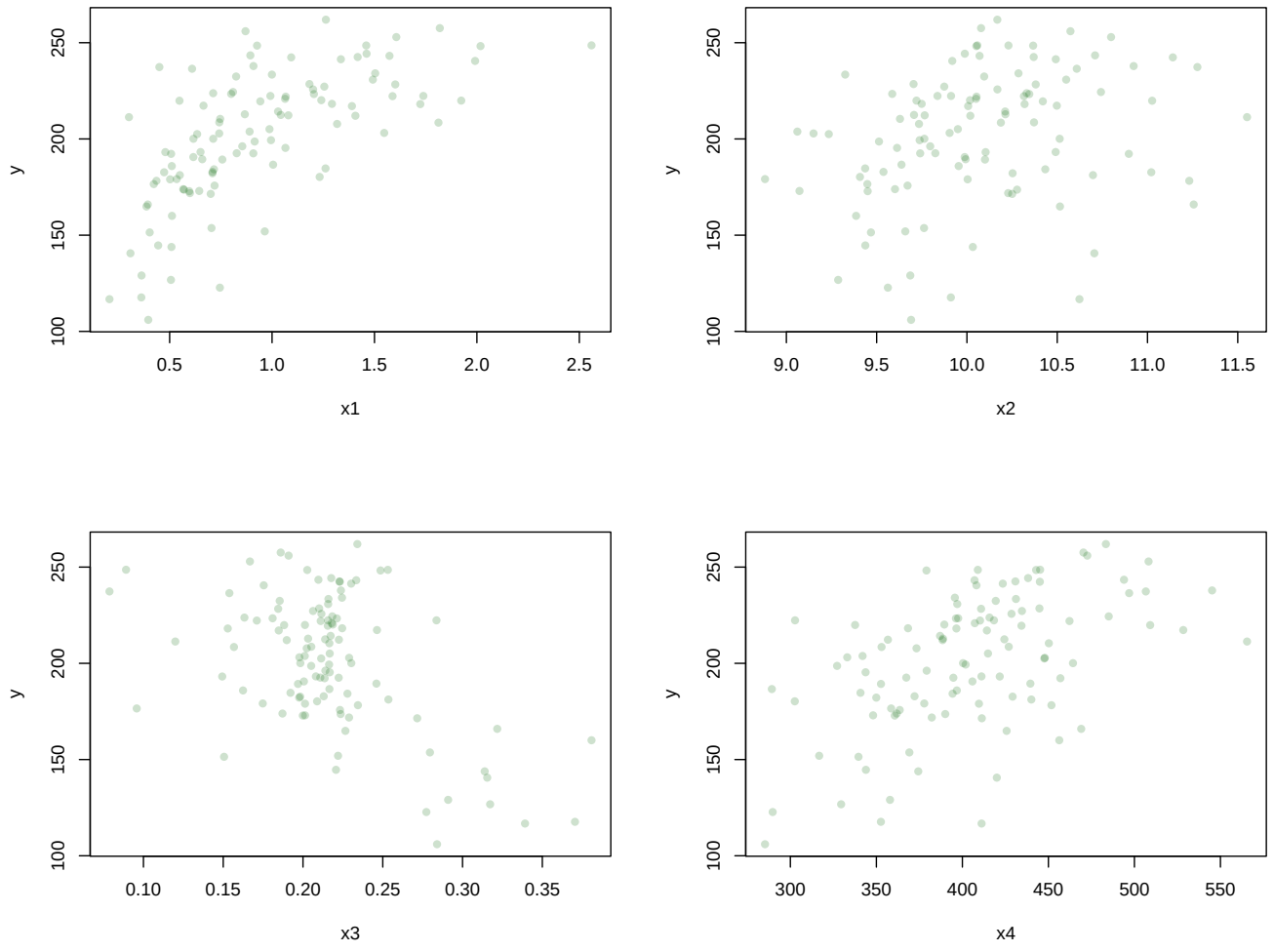
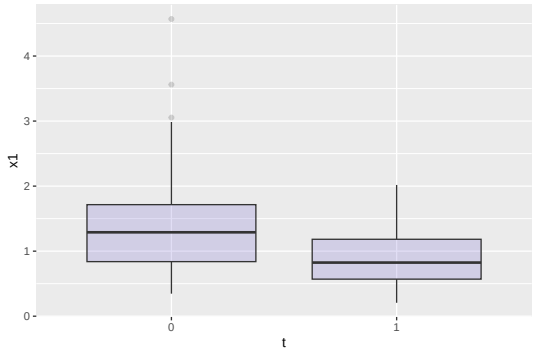
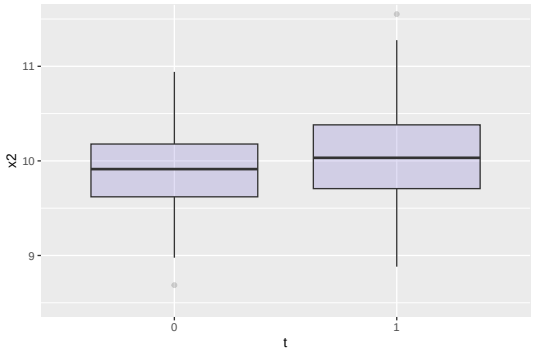


Figure 2: Scatterplots of response versus observed covariates for respondents in a sample of 200 units

The covariates seen by the analyst are also related to t_i's. Side-by-side boxplots of the x_{ij}'s for the $t_i = 0$ and $t_i = 1$ groups are shown in Figure 2.

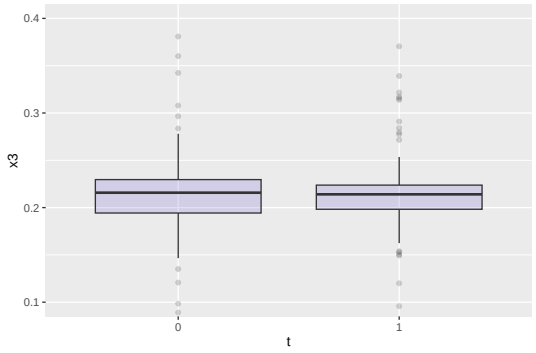


(a)

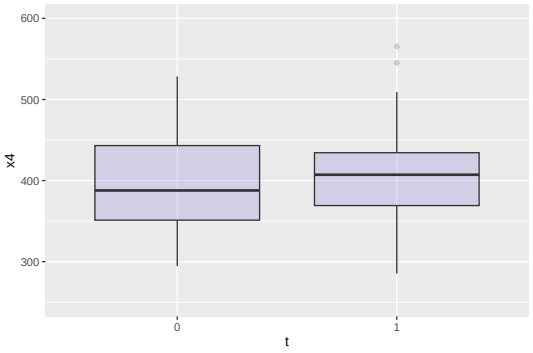


(b)

Figure 3: Distributions of (a-b) observed covariates



(a)



(b)

Figure 4: Distributions of (c-d) observed covariates

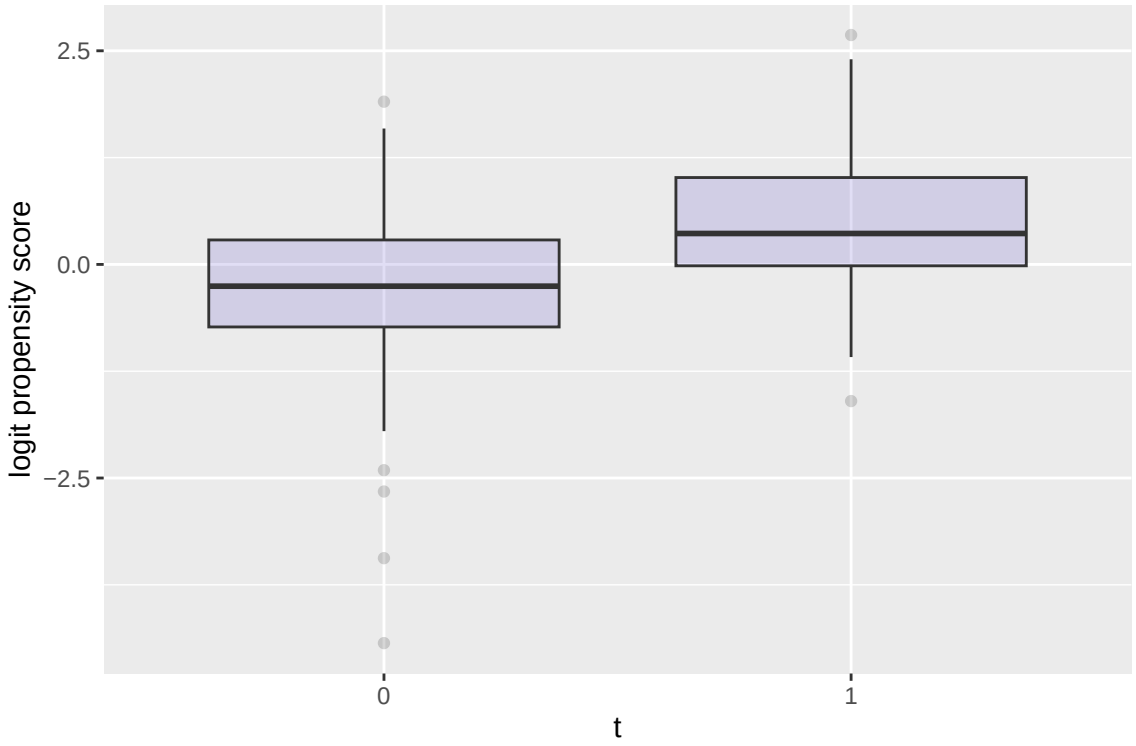


Figure 5: Estimated propensity scores for nonrespondents and respondents in a sample of 200 units

Table 1: Performance of IPW estimators of μ over 1000 samples from the artificial population

Sample Size	**π-model**	**Method**	**Bias**	**RMSE**	**MAE**
(1) n =200	Correct	IPW-POP	-0.20	3.92	2.43
		IPW-NR	-0.06	3.64	2.24
	Incorrect	IPW-POP	1.87	8.52	3.38
		IPW-NR	2.67	6.16	2.62
(2) n = 1000	Correct	IPW-POP	-0.12	1.68	1.08
		IPW-NR	-0.07	1.51	1.01
	Incorrect	IPW-POP	4.70	11.52	2.57
		IPW-NR	4.14	6.80	2.98

Examining Table 1, we can see that the IPW estimates are biased when $\pi-$ model is misspecified. The bias and RMSE actually get worse as the sample size grows.

Table 2: Performance of ordinary least-squares regression estimators over 1000 samples from the artifical population

Sample Size	**π-model**	**Method**	**Bias**	**RMSE**	**MAE**
(1) n =200	Correct	OLS	0.05	2.59	1.77
	Incorrect	OLS	-0.59	3.38	2.27
(2) n = 1000	Correct	OLS	0.01	1.12	0.74
	Incorrect	OLS	-0.83	1.71	1.21

From Table 2, we can see that the bias is indeed removed when the $y-$ model is correct. Comparing RMSE values from Table 1 and Table 3, we can see that the estimates based on the incorrect y-model are more stable and efficient than those based on the incorrect $\pi-$ model.

Table 3: Performance of bias-corrected regression estimators over 1000 samples from the artificial population

Sample Size	π**-model**	y**-model**	**Method**	**Bias**	**RMSE**	**MAE**
(1) n =200	Correct	Correct	BC-OLS	-0.13	2.54	1.67
		Incorrect	BC-OLS	0.23	3.46	2.24
	Incorrect	Correct	BC-OLS	-0.13	2.54	1.67
		Incorrect	BC-OLS	-4.19	7.65	3.62
(2) n = 1000	Correct	Correct	BC-OLS	-0.005	1.13	0.78
		Incorrect	BC-OLS	0.04	1.61	1.07
	Incorrect	Correct	BC-OLS	-0.005	1.14	0.81
		Incorrect	BC-OLS	-8.30	15.79	4.84

The bias of this estimate is indeed reduced when either of the two models is correct. Moever, comparing these results to those from the IPW-POP estimate in Table 1, we can see that augmenting the IPW procedure by information from a correct $y-$ model increases the efficiency.

4 Discussion

4.1 Is the Standard Logistic Regression Good? and Stronger Interaction Effects on the Outcome Regression Model

Joseph D. Y. Kang and Joseph L. Schafer (2007)[2] states the strong performance of OLS and they found no method outperformed OLS. However, Greg Ridgeway and Daniel F. McCaffrey (2007) [4] investigated if the high variance reported by Joseph D. Y. Kang and Joseph L. Schafer (2007)[2] when using propensity score weights results from their use of standard Logistic regression. Also, they consider interaction terms if they will favor the DR approach.

The original main paper stated that none of the various IPW methods could overcome the problems with estimated propensity scores near 0 and 1. Probably it indicated that the problem results from propensity score estimator rather than IPW methods. Therefore, Greg Ridgeway and Daniel F. McCaffrey (2007) [4] investigated the performance of using a generalized boosted model (GBM) estimating weights and tested the performance of IPW and DR estimators based on GBM in the presence of omitted interactions.

They did two simulation studies, one is for the IPW methods, the other one is for DR estimators. The simulation results for IPW methods show that IPW estimators with logistic regression using "mis-transformed" X variables have the largest RMSEs and IPW outperforms OLS when the OLS models exclude an important interaction.

4.2 Understanding OR, PS and DR

Tan (2007) [7] systematically studied the PS and the OR and DR estimation. They pointed out OR and PS are two approaches with different characteristics, and one does not necessarily dominate the other. The problem for OR approach is implicitly making extrapolation. The PS-weighting method tends to yield large weights, explicitly indicating uncertainty in the estimate. Viewing DR estimation in the PS approach by incorporating an OR model is more constructive than incorporating a PS model in the OR approach.

Tan (2007) [7] also answered the two questions intuitively:

- Which task is more likely to be accomplished, to correctly specify an OR model or a PS model?
- Which mistake (even a mild one) can lead to worse estimations, mis-specification of an OR model or a PS model?

For the first question, these two models both involve the same explanatory variables X. However, they deal with different difficulties. The OR-model works on the "truncated" data within treated subjects. Therefore, any OR models rely on extrapolation to predict $m_1(X)$ at values of X that are different from those for most values of X that are different from those for most treated subjects. It is not possible to detect OR-model mis-specification in use of the usual model checking. In contrast, PS-modelling works on the "full" data and does not need to deal with the presence of data truncation. For the second question, KS stated that (A)IPW estimator is sensitive to mis-specification of the PS model when $\pi(X) \approx 0$ for some values of X. In this case, the

estimator has inflated standard error, which can be much greater than its bias. For a misspecified OR model, the bias of OR estimator can be of similar magnitude to its standard deviation since the bias of OR estimator is the average of those $\hat{m_1}(X)$ across individual subjects in the original scale.

4.3 Semi-parametric Theory Perspective

Anastasios A. Tsiatis and Marie Davidian (2007)[8] viewed inference from the point of semi-parametric theory, which focus on the assumptions in the statistical models leading to different types of estimators, rather than the forms of estimators themselves. As the MAR assumption or strong ignorability assumption, y and t are conditionally independent given x. Under the assumption, all joint densities for the observed data have the form

$$p(z) = p(y|z)^{I(t=1)}p(t|x)p(x)$$

Different statistical models may be proposed by making different assumptions on the components of p(z). Anastasios A. Tsiatis and Marie Davidian (2007) [8]focused on three such models:

- a) Make no assumptions on the forms of p(x) or p(t|x), but make a specific assumption on p(y|x), which is $E(y|x) = m(x, \beta)$.
- b) Make no assumptions on the forms of p(x) or p(y|x). Make a specific assumption on $p(t|x)$ that $p(t = 1|x) = E(t|x) = \pi(x, \alpha)$ for some given function $\pi(x, \alpha)$ depending on parameters α.
- c) Make no assumptions on the form of p(x), but make spscific assumptions on p(y|x) and p(t|x).

They consider estimators discussed by KS paper in a point view of influence functions. In the view of influence function, they found these estimators $\hat{\mu}_{BC-OLS}$, $\hat{\mu}_{BC-POP}$, $\hat{\mu}_{WLS}$ can be expressed in AIPW form. Influence functions' idea can provide us useful insights to construct and study these DR estimators. As we discussed in class, G-estimation is actually a idea of building orthogonality to find the most efficient influence function given the nuisance space.

References

[1] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–972, 2005. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3695907.

[2] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523 – 539, 2007. doi: 10.1214/07-STS227. URL https://doi.org/10.1214/07-STS227.

[3] Daryl Pregibon. Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, 38(2): 485–498, 1982. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530463.

[4] Greg Ridgeway and Daniel F. McCaffrey. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):540–543, 2007. ISSN 08834237. URL http://www.jstor.org/stable/27645859.

[5] James Robins, Andrea Rotnitzky, and Lue Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of The American Statistical Association - J AMER STATIST ASSN*, 89:846–866, 09 1994. doi: 10.1080/01621459.1994.10476818.

[6] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999. ISSN 01621459. URL http://www.jstor.org/stable/2669923.

[7] Zhiqiang Tan. Comment: Understanding or, ps and dr. *Statistical Science*, 22(4):560–568, 2007. ISSN 08834237. URL http://www.jstor.org/stable/27645861.

[8] Anastasios A. Tsiatis and Marie Davidian. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):569–573, 2007. ISSN 08834237. URL http://www.jstor.org/stable/27645862.

```
library(MASS)
N <- 200
dim_z <- 4
R_z <- matrix(c(1,0,0,0,
                0,1,0,0,
                0,0,1,0,
                0,0,0,1),
                nrow = 4, ncol = 4, byrow = TRUE)
Z <- mvrnorm(n = N, mu = rep(0,dim_z), Sigma = R_z)
beta <- c(210,27.4,13.7,13.7,13.7)
Y <- (cbind(rep(1,N),Z))%*%beta

expit <- function(x){
  return(1/(1+exp(-x)))
}

true_prs <- rep(0,N)
for(i in 1:N){
  true_prs[i] <- expit(Z[i,]%*%c(-1,0.5,-0.25,-0.1))
}

X <- matrix(data = NA, nrow = N, ncol = 4)
for(i in 1:N){
  X[i,1] <- exp(Z[i,1]/2)
  X[i,2] <- Z[i,2]/(1+exp(Z[i,1])) + 10
  X[i,3] <- ((Z[i,1]*Z[i,3])/25 + 0.6)^3
  X[i,4] <- (Z[i,2] + Z[i,4] + 20)^2
}

rs_index <- t <- rbinom(N,1,true_prs)
respondents_index <- which(rs_index == 1)
Y_respondents <- Y[respondents_index]
X_observed <- X[respondents_index,]

par(mfrow = c(2,2))
plot(X_observed[,1],Y_respondents,
     main="",
     xlab = "x1",
     ylab = "y",
     col=rgb(0,100,0,50,maxColorValue=255),
     pch=16)
plot(X_observed[,2],Y_respondents,
     main="",
     xlab = "x2",
     ylab = "y",
     col=rgb(0,100,0,50,maxColorValue=255),
     pch=16)
plot(X_observed[,3],Y_respondents,
     main="",
     xlab = "x3",
     ylab = "y",
     col=rgb(0,100,0,50,maxColorValue=255),
     pch=16)
plot(X_observed[,4],Y_respondents,
     main="",
     xlab = "x4",
     ylab = "y",
     col=rgb(0,100,0,50,maxColorValue=255),
     pch=16)
```

```
library(ggplot2)

x1 <- c(X_observed[,1],X[-respondents_index,1])
observed_x1 <- c(rep("1",97),rep("0",103))
dt_x1 <- data.frame(x1 = x1, observed = observed_x1)
ggplot(dt_x1, aes(x=observed_x1, y=x1)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
   # coord_cartesian(ylim=c(0.5,3.5))+
  xlab("t")+
  ylab("x1")

x2 <- c(X_observed[,2],X[-respondents_index,2])
observed_x2 <- c(rep("1",97),rep("0",103))
dt_x2 <- data.frame(x2 = x2, observed = observed_x2)
ggplot(dt_x2, aes(x=observed_x2, y=x2)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  coord_cartesian(ylim=c(8.5,11.5))+
  xlab("t")+
  ylab("x2")

x3 <- c(X_observed[,3],X[-respondents_index,3])
observed_x3 <- c(rep("1",97),rep("0",103))
dt_x3 <- data.frame(x3 = x3, observed = observed_x3)
ggplot(dt_x3, aes(x=observed_x3, y=x3)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  coord_cartesian(ylim=c(0.1,0.4))+
  xlab("t")+
  ylab("x3")

x4 <- c(X_observed[,4],X[-respondents_index,4])
observed_x4 <- c(rep("1",97),rep("0",103))
dt_x4 <- data.frame(x4 = x4, observed = observed_x4)
ggplot(dt_x4, aes(x=observed_x4, y=x4)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  coord_cartesian(ylim=c(250,600))+
  xlab("t")+
  ylab("x4")

data_XTY <- data.frame(X1 = X[,1], X2 = X[,2], X3 = X[,3], X4 = X[,4], Y = Y, t = t)

glm_tx <- glm(t~X1+X2+X3+X4, data = data_XTY, family = "binomial")

eta_fitted <- cbind(rep(1,N),X) %*% coefficients(glm_tx)

dt_eta <- data.frame(eta_fitted = eta_fitted, t = t)
ggplot(dt_eta, aes(x=as.factor(t), y=eta_fitted)) +
  geom_boxplot(fill="slateblue", alpha=0.2) +
  # coord_cartesian(ylim=c(250,600))+
  xlab("t")+
  ylab("logit␣propensity␣score")

## Simulation for Table 1 and Table 2
nreps<-1000
ests<-matrix(0,nrow=nreps,ncol=6)
for(rep in 1:nreps){
  N <- 200
  dim_z <- 4
  R_z <- matrix(c(1,0,0,0,
```

```
                    0,1,0,0,
                    0,0,1,0,
                    0,0,0,1),
                  nrow = 4, ncol = 4, byrow = TRUE)
  Z <- mvrnorm(n = N, mu = rep(0,dim_z), Sigma = R_z)
  beta <- c(210,27.4,13.7,13.7,13.7)
  Y <- (cbind(rep(1,N),Z))%*%beta
  true_prs <- rep(0,N)
  for(i in 1:N){
    true_prs[i] <- expit(Z[i,]%*%c(-1,0.5,-0.25,-0.1))
  }
  X <- matrix(data = NA, nrow = N, ncol = 4)
  for(i in 1:N){
    X[i,1] <- exp(Z[i,1]/2)
    X[i,2] <- Z[i,2]/(1+exp(Z[i,1])) + 10
    X[i,3] <- ((Z[i,1]*Z[i,3])/25 + 0.6)^3
    X[i,4] <- (Z[i,2] + Z[i,4] + 20)^2
  }
  rs_index <- t <- rbinom(N,1,true_prs)
  respondents_index <- which(rs_index == 1)
  Y_respondents <- Y[respondents_index]
  X_observed <- X[respondents_index,]

  dt_XTYZ <- data.frame(Z1 = Z[,1], Z2 = Z[,2], Z3 = Z[,3], Z4 = Z[,4], X1 = X[,1], X2 = X[,2
  dt_XTYZ_res <- dt_XTYZ[respondents_index,]
  eX_specified <- fitted(glm(t~Z1+Z2+Z3+Z4,data = dt_XTYZ, family=binomial))
  eX_misspecified <- fitted(glm(t~X1+X2+X3+X4,data = dt_XTYZ, family=binomial))

  w1<-t/eX_specified
  W1<-w1/sum(w1)
  w0<-(1-t)/(1-eX_specified)
  W0<-w0/sum(w0)
  r_1 <- length(respondents_index)/N
  ests[rep,1]<-sum(W1*Y)
  ests[rep,2]<-(1-r_1)*sum(W0*Y)+r_1*sum(W1*Y)

  w1_mis<-t/eX_misspecified
  W1_mis<-w1_mis/sum(w1_mis)
  w0_mis<-(1-t)/(1-eX_misspecified)
  W0_mis<-w0_mis/sum(w0_mis)

  ests[rep,3]<-sum(W1_mis*Y)
  ests[rep,4]<-(1-r_1)*sum(W0_mis*Y)+r_1*sum(W1_mis*Y)

  # Y_specified <- fitted(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))
  ests[rep,5] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),dt_

  # Y_misspecified <- fitted(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))
  ests[rep,6] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),dt_
}

nreps<-1000
ests_1000<-matrix(0,nrow=nreps,ncol=6)
for(rep in 1:nreps){
  N <- 1000
  dim_z <- 4
  R_z <- matrix(c(1,0,0,0,
                  0,1,0,0,
                  0,0,1,0,
```

```
                    0,0,0,1),
                  nrow = 4, ncol = 4, byrow = TRUE)
  Z <- mvrnorm(n = N, mu = rep(0,dim_z), Sigma = R_z)
  beta <- c(210,27.4,13.7,13.7,13.7)
  Y <- (cbind(rep(1,N),Z))%*%beta
  true_prs <- rep(0,N)
  for(i in 1:N){
    true_prs[i] <- expit(Z[i,]%*%c(-1,0.5,-0.25,-0.1))
  }
  X <- matrix(data = NA, nrow = N, ncol = 4)
  for(i in 1:N){
    X[i,1] <- exp(Z[i,1]/2)
    X[i,2] <- Z[i,2]/(1+exp(Z[i,1])) + 10
    X[i,3] <- ((Z[i,1]*Z[i,3])/25 + 0.6)^3
    X[i,4] <- (Z[i,2] + Z[i,4] + 20)^2
  }
  rs_index <- t <- rbinom(N,1,true_prs)
  respondents_index <- which(rs_index == 1)
  Y_respondents <- Y[respondents_index]
  X_observed <- X[respondents_index,]

  dt_XTYZ <- data.frame(Z1 = Z[,1], Z2 = Z[,2], Z3 = Z[,3], Z4 = Z[,4], X1 = X[,1], X2 = X[,2
  dt_XTYZ_res <- dt_XTYZ[respondents_index,]
  eX_specified <- fitted(glm(t~Z1+Z2+Z3+Z4,data = dt_XTYZ, family=binomial))
  eX_misspecified <- fitted(glm(t~X1+X2+X3+X4,data = dt_XTYZ, family=binomial))

  w1<-t/eX_specified
  W1<-w1/sum(w1)
  w0<-(1-t)/(1-eX_specified)
  W0<-w0/sum(w0)
  r_1 <- length(respondents_index)/N
  ests_1000[rep,1]<-sum(W1*Y)
  ests_1000[rep,2]<-(1-r_1)*sum(W0*Y)+r_1*sum(W1*Y)

  w1_mis<-t/eX_misspecified
  W1_mis<-w1_mis/sum(w1_mis)
  w0_mis<-(1-t)/(1-eX_misspecified)
  W0_mis<-w0_mis/sum(w0_mis)

  ests_1000[rep,3]<-sum(W1_mis*Y)
  ests_1000[rep,4]<-(1-r_1)*sum(W0_mis*Y)+r_1*sum(W1_mis*Y)

  # Y_specified <- fitted(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))
  ests_1000[rep,5] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N

  # Y_misspecified <- fitted(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))
  ests_1000[rep,6] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N
}

## Simulation for Table 3
nreps_AIPW<-1000
ests_AIPW<-matrix(0,nrow=nreps,ncol=4)
for(rep in 1:nreps){
  N <- 200
  dim_z <- 4
  R_z <- matrix(c(1,0,0,0,
                  0,1,0,0,
                  0,0,1,0,
                  0,0,0,1),
```

```
                nrow = 4, ncol = 4, byrow = TRUE)
Z <- mvrnorm(n = N, mu = rep(0,dim_z), Sigma = R_z)
beta <- c(210,27.4,13.7,13.7,13.7)
Y <- (cbind(rep(1,N),Z))%*%beta
true_prs <- rep(0,N)
for(i in 1:N){
  true_prs[i] <- expit(Z[i,]%*%c(-1,0.5,-0.25,-0.1))
}
X <- matrix(data = NA, nrow = N, ncol = 4)
for(i in 1:N){
  X[i,1] <- exp(Z[i,1]/2)
  X[i,2] <- Z[i,2]/(1+exp(Z[i,1])) + 10
  X[i,3] <- ((Z[i,1]*Z[i,3])/25 + 0.6)^3
  X[i,4] <- (Z[i,2] + Z[i,4] + 20)^2
}
rs_index <- t <- rbinom(N,1,true_prs)
respondents_index <- which(rs_index == 1)
Y_respondents <- Y[respondents_index]
X_observed <- X[respondents_index,]

dt_XTYZ <- data.frame(Z1 = Z[,1], Z2 = Z[,2], Z3 = Z[,3], Z4 = Z[,4], X1 = X[,1], X2 = X[,2
dt_XTYZ_res <- dt_XTYZ[respondents_index,]
eX_specified <- fitted(glm(t~Z1+Z2+Z3+Z4,data = dt_XTYZ, family=binomial))
eX_misspecified <- fitted(glm(t~X1+X2+X3+X4,data = dt_XTYZ, family=binomial))

w1<-t/eX_specified
W1<-w1/sum(w1)
w0<-(1-t)/(1-eX_specified)
W0<-w0/sum(w0)
r_1 <- length(respondents_index)/N
w1_mis<-t/eX_misspecified
W1_mis<-w1_mis/sum(w1_mis)
w0_mis<-(1-t)/(1-eX_misspecified)
W0_mis<-w0_mis/sum(w0_mis)

# ests[rep,1]<-sum(W1*Y)
# ests[rep,2]<-(1-r_1)*sum(W0*Y)+r_1*sum(W1*Y)
#
residulas_cor <- Y-t(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
ests_AIPW[rep,1] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N
residulas_incor <- Y-t(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N
ests_AIPW[rep,2] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N

ests_AIPW[rep,3] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N
ests_AIPW[rep,4] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N

# w1_mis<-t/eX_misspecified
# W1_mis<-w1_mis/sum(w1_mis)
# w0_mis<-(1-t)/(1-eX_misspecified)
# W0_mis<-w0_mis/sum(w0_mis)
#
# ests[rep,3]<-sum(W1_mis*Y)
# ests[rep,4]<-(1-r_1)*sum(W0_mis*Y)+r_1*sum(W1_mis*Y)
#
# # Y_specified <- fitted(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))
# ests[rep,5] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
#
# # Y_misspecified <- fitted(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))
# ests[rep,6] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
```

```
}

nreps_AIPW<-1000
ests_AIPW_1000<-matrix(0,nrow=nreps,ncol=4)
for(rep in 1:nreps){
  N <- 1000
  dim_z <- 4
  R_z <- matrix(c(1,0,0,0,
                  0,1,0,0,
                  0,0,1,0,
                  0,0,0,1),
                nrow = 4, ncol = 4, byrow = TRUE)
  Z <- mvrnorm(n = N, mu = rep(0,dim_z), Sigma = R_z)
  beta <- c(210,27.4,13.7,13.7,13.7)
  Y <- (cbind(rep(1,N),Z))%*%beta
  true_prs <- rep(0,N)
  for(i in 1:N){
    true_prs[i] <- expit(Z[i,]%*%c(-1,0.5,-0.25,-0.1))
  }
  X <- matrix(data = NA, nrow = N, ncol = 4)
  for(i in 1:N){
    X[i,1] <- exp(Z[i,1]/2)
    X[i,2] <- Z[i,2]/(1+exp(Z[i,1])) + 10
    X[i,3] <- ((Z[i,1]*Z[i,3])/25 + 0.6)^3
    X[i,4] <- (Z[i,2] + Z[i,4] + 20)^2
  }
  rs_index <- t <- rbinom(N,1,true_prs)
  respondents_index <- which(rs_index == 1)
  Y_respondents <- Y[respondents_index]
  X_observed <- X[respondents_index,]

  dt_XTYZ <- data.frame(Z1 = Z[,1], Z2 = Z[,2], Z3 = Z[,3], Z4 = Z[,4], X1 = X[,1], X2 = X[,2
  dt_XTYZ_res <- dt_XTYZ[respondents_index,]
  eX_specified <- fitted(glm(t~Z1+Z2+Z3+Z4,data = dt_XTYZ, family=binomial))
  eX_misspecified <- fitted(glm(t~X1+X2+X3+X4,data = dt_XTYZ, family=binomial))

  w1<-t/eX_specified
  W1<-w1/sum(w1)
  w0<-(1-t)/(1-eX_specified)
  W0<-w0/sum(w0)
  r_1 <- length(respondents_index)/N
  w1_mis<-t/eX_misspecified
  W1_mis<-w1_mis/sum(w1_mis)
  w0_mis<-(1-t)/(1-eX_misspecified)
  W0_mis<-w0_mis/sum(w0_mis)

  # ests[rep,1]<-sum(W1*Y)
  # ests[rep,2]<-(1-r_1)*sum(W0*Y)+r_1*sum(W1*Y)
  #
  residulas_cor <- Y-t(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
  ests_AIPW_1000[rep,1] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(re
  residulas_incor <- Y-t(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N
  ests_AIPW_1000[rep,2] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(re

  ests_AIPW_1000[rep,3] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(re
  ests_AIPW_1000[rep,4] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(re

  # w1_mis<-t/eX_misspecified
  # W1_mis<-w1_mis/sum(w1_mis)
```

```
  # w0_mis<-(1-t)/(1-eX_misspecified)
  # W0_mis<-w0_mis/sum(w0_mis)
  #
  # ests[rep,3]<-sum(W1_mis*Y)
  # ests[rep,4]<-(1-r_1)*sum(W0_mis*Y)+r_1*sum(W1_mis*Y)
  #
  # # Y_specified <- fitted(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))
  # ests[rep,5] <- sum(coefficients(lm(Y~Z1+Z2+Z3+Z4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
  #
  # # Y_misspecified <- fitted(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))
  # ests[rep,6] <- sum(coefficients(lm(Y~X1+X2+X3+X4,data = dt_XTYZ_res))%*%t(cbind(rep(1,N),
}
```