# A short note on conditional models, robustness, and missingness

Yen-Chi Chen
University of Washington
August 10, 2020

In this note, I will describe an interesting (and perhaps a well-known result to some researchers) property of a conditional model under missing data. This phenomenon can be viewed as a special case of the doubly-robust property from the semi-parametric inference but I will avoid using any techniques from semi-parametric inference[1].

Consider a simple regression problem where $Y \in \mathbb{R}$ is the response variable and $X \in \mathbb{R}^{d+1}$ is the covariate including 1's as the first covariate (to account for the intercept). Our data will be

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

that are IID from some distributions. Let $m(x) = \mathbb{E}(Y|X = x)$ be the regression function.

Now suppose we want to use a linear model to analyze the data and use the conventional least square method. Note that here we do not assume the linear model to be correct. The least square approach finds the estimator

$$\widehat{\beta}_n = \mathsf{argmin}_\beta \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2,$$

which is known to converge to the population limit

$$\beta^* = \mathbb{E}(XX^T)^{-1}\mathbb{E}(XY). \tag{1}$$

Now we make the problem a bit more complicated. Suppose that some response variable are missing. We use $R$ as the binary variable for observing $Y$ or not. Namely, when $R = 1$, we observe $Y$ (and $X$) and when $R = 0$, we only observe $X$.

Consider the conventional missing at random (MAR) assumption, which in this context, is equivalent to

$$(MAR) \qquad R \perp Y|X.$$

Our goal is to recover the idealized population limit $\beta^*$ in equation (1) under the missing data scenario. Note that even if the linear model is incorrect, such parameter is still useful–it is the coefficients corresponding to the best linear predictor.

**Summary of the results.**

- The inverse probability weighting estimator is always consistent (under suitable conditions) regardless of the assumed conditional model being correct or not.
- The complete-case analysis is generally biased but it will be consistent when the conditional model is correct.
- If we mis-specified the weights, the resulting estimator is biased. However, if the parametric conditional model is correct, a mis-specified weight will still lead to a consistent estimator.

[1] Personally, I will use this note as an introduction/motivation of the semi-parametric inference.

# 1 The inverse probability weighting approach

In the missing data literature, a simple approach to account for the missingness is to apply an inverse probability weighting (IPW) approach to the analysis. The idea is very simple. Suppose we want to estimate $\mathbb{E}(Y)$, the IPW estimator uses the following fact (under MAR assumption):

$$\mathbb{E}(Y) = \mathbb{E}\left(\frac{RY}{P(R=1|X)}\right), \tag{2}$$

which suggests to use the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(X_i)},$$

where $\pi(x) = P(R=1|X=x)$ is known as the propensity score.

As a result, we should estimate $\beta^*$ using the idea of inverse probability weighting:

$$\widehat{\beta}_{IPW} = \mathsf{argmin}_{\beta}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}(Y_i - X_i^T\beta)^2.$$

You can easily show that $\widehat{\beta}_{IPW} \xrightarrow{P} \beta^*$. An intuitive explanation is that the population version of $\widehat{\beta}_{IPW}$ is

$$\beta_{IPW} = \mathbb{E}\left(\frac{R}{\pi(X)}XX^T\right)^{-1}\mathbb{E}\left(\frac{R}{\pi(X)}XY\right).$$

Under MAR, we can apply equation (2) to show that $\beta_{IPW} = \beta^*$.

# 2 Robustness when the model is correct

Now we turn to a naive estimator called the *complete case analysis*, which uses only those observations that we have complete data and proceed without applying the inverse probability weighing. Namely, the complete case analysis uses the estimator

$$\widehat{\beta}_{CC} = \mathsf{argmin}_{\beta}\sum_{i=1}^{n}R_i(Y_i - X_i^T\beta)^2,$$

and its population version is

$$\beta_{CC} = \mathbb{E}\left(RXX^T\right)^{-1}\mathbb{E}\left(RXY\right) = \mathbb{E}\left(XX^T|R=1\right)^{-1}\mathbb{E}\left(XY|R=1\right).$$

In general, $\beta_{CC} \neq \beta^*$.

However, the complete case (CC) estimator *will be consistent if the linear model is correct!* To see this, we assume that the linear model is correct, i.e.,

$$\mathbb{E}(Y|X) = m(X) = X^T\beta_0. \tag{3}$$

One can easily shown that under equation (3), $\beta_0 = \beta^*$ so the IPW estimator is consistent, regardless of the modeling being correct or not.

Now we consider the population version of the CC estimator. Note that the MAR assumption implies $m(X, R = 1) = \mathbb{E}(Y|X, R = 1) = \mathbb{E}(Y|X) = m(X)$. As a result,

$$\begin{aligned}\mathbb{E}\left(XY|R=1\right) &= \mathbb{E}\left(Xm(X,R=1)|R=1\right)\\ &= \mathbb{E}\left(Xm(X)|R=1\right)\\ &= \mathbb{E}\left(XX^T\beta_0|R=1\right)\\ &= \mathbb{E}\left(XX^T|R=1\right)\beta_0,\end{aligned}$$

which implies that

$$\begin{aligned}\beta_{CC} &= \mathbb{E}\left(XX^T|R=1\right)^{-1}\mathbb{E}\left(XY|R=1\right)\\ &= \mathbb{E}\left(XX^T|R=1\right)^{-1}\mathbb{E}\left(XX^T|R=1\right)\beta_0 = \beta_0 = \beta^*.\end{aligned}$$

So the CC estimator is still consistent even if we do not correctly specify the missing probability $\pi(x)$ (the CC estimator implicitly assume $\pi(x) = \pi_0$ for some constant $\pi_0$).

A more interesting fact is that you can show that if we use another weighting function $\omega(x) \neq \pi(x) = P(R = 1|X = x)$. Then the resulting estimator has a population limit

$$\begin{aligned}\beta_\omega &= \mathbb{E}\left(\frac{R}{\omega(X)}XX^T\right)^{-1}\mathbb{E}\left(\frac{R}{\omega(X)}XY\right)\\ &= \mathbb{E}\left(\frac{R}{\omega(X)}XX^T\right)^{-1}\mathbb{E}\left(\frac{R}{\omega(X)}XX^T\right)\beta_0\\ &= \beta_0 = \beta^*.\end{aligned} \tag{4}$$

So it will lead to a consistent estimator!

Thus, when the parametric model is correct, the *nuisance* part–the component that we are not interested in (also known as the nuisance parameter)–$\pi(x)$ can be mis-specified and we still obtain a consistent estimator. A similar phenomenon also occurs in the *doubly-robust estimator*, where as long as one of the model is correct, we are able to recover the parameter of interest.

## 3 A general conditional model

The linear model is not a special case. Other conditional parametric models also have a similar property (a more general case is the so-called semi-parametric model).

To see this, consider a conditional parametric model

$$p(y|x;\theta),$$

where $\theta \in \Theta$. Let

$$\ell(\theta|x,y) = \log p(y|x;\theta)$$

be the log-likelihood function and

$$S(\theta|x,y) = \nabla_\theta \ell(\theta|x,y)$$

be the score function. The MLE $\widehat{\theta}$ solves the score equation, i.e.,

$$\widehat{\theta} : 0 = \frac{1}{n}\sum_{i=1}^{n} S(\widehat{\theta}|X_i,Y_i).$$

So the population MLE solves the population score equation

$$\theta^* : 0 = \mathbb{E}(S(\theta^*|X,Y)).$$

We will focus on estimating $\theta^*$. It is still well-defined even if the parametric model is incorrect–the distribution corresponding to $\theta^*$ is the one that minimizes the Kullback-Leibler divergence to the distribution that generates the data.

As is argued in the linear model, when some $Y$'s are missing at random, we should use the IPW approach, which corresponds to an estimator that converges to the population limit

$$\theta_{IPW} : 0 = \mathbb{E}\left(\frac{R}{\pi(X)} S(\theta_{IPW}|X,Y)\right).$$

One can show that $\theta_{IPW} = \beta^*$ by the law of iterated expectation.

The CC estimator converges to the population limit

$$\theta_{CC} : 0 = \mathbb{E}\left(RS(\theta_{CC}|X,Y)\right) = \mathbb{E}\left(S(\theta_{CC}|X,Y)|R=1\right),$$

which in general, $\theta_{CC} \neq \beta^*$ so it is biased.

However, when the model is correct, we will show that the CC estimator, again, becomes consistent!

### 3.1 When the model is correct

Let $\theta_0$ be the true parameter of the conditional model. Namely, the true conditional PDF $p(y|x) = p(y|x;\theta_0)$. One can easily show that in this context, $\theta^* = \theta_0$.

To demonstrate the consistency of the CC estimator, we consider its score equation $\mathbb{E}\left(S(\theta|X,Y)|R=1\right)$. All we need is to show that when we plug-in $\theta = \theta_0$, we will solve this equation. Using the fact that

$$S(\theta|X,Y) = \nabla_\theta \log p(Y|X;\theta) = \frac{1}{p(Y|X;\theta)}\nabla_\theta p(Y|X;\theta)$$

and the MAR assumption

$$(MAR) \Rightarrow p(y|x;\theta_0) = p(y|x,R=1),$$

we can rewrite the score equation as

$$\begin{aligned}
0 &= \mathbb{E}\left(S(\theta|X,Y)|R=1\right) \\
&= \int S(\theta|x,y)p(y|x,R=1)dyp(x|R=1)dx \\
&\overset{(MAR)}{=} \int S(\theta|x,y)p(y|x;\theta_0)dyp(x|R=1)dx \\
&= \int \frac{1}{p(y|x;\theta)}\nabla_\theta p(y|x;\theta)p(y|x;\theta_0)dyp(x|R=1)dx.
\end{aligned}$$

When we set $\theta = \theta_0$, we obtain

$$\begin{aligned}
\mathbb{E}\left(S(\theta_0|X,Y)|R=1\right) &= \int \frac{1}{p(y|x;\theta_0)}\nabla_\theta p(y|x;\theta_0)p(y|x;\theta_0)dyp(x|R=1)dx \\
&= \int \nabla_\theta \underbrace{\int p(y|x;\theta_0)dy}_{=1} p(x|R=1)dx \\
&= 0.
\end{aligned}$$

Therefore, the true parameter $\theta_0 = \theta^*$ also solves the score equation under the CC case. So the CC estimator is still consistent!

You can easily verify that a similar result occurs when we are using any other weighting function $\omega(x)$ as well! As a result, when the conditional model is correct, our analysis is pretty robust to the mis-specification of the nuisance parameter $\pi(x) = p(x|R=1)$.

# 4 Remarks

- **A statistical learning framework.** The above procedure can be generalized into a statistical learning framework with a nonparametric model. Suppose that we impose a loss function $L(y_0,y_1)$ that measures the loss between $y_0$ and $y_1$. And consider a class of predictor $f \in \mathcal{F}$ such that each predictor $f = f(x)$. The optimal predictor

  $$f^* = \mathsf{argmin}_{f\in\mathcal{F}}\mathbb{E}(L(f(X),Y)).$$

  Under the MAR condition, the IPW approach leads to

  $$f_{IPW} = \mathsf{argmin}_{f\in\mathcal{F}}\mathbb{E}\left(\frac{R}{\pi(X)}L(f(X),Y)\right) = \mathsf{argmin}_{f\in\mathcal{F}}\mathbb{E}(L(f(X),Y)),$$

  which is always consistent. And an incorrect weighting approach such as the CC analysis will lead to a biased estimator.

  Now we will show that when a nonparametric model is correct, an incorrectly specified weighting approach is still consistent. To simplify the problem, we assume that *X is categorical with a finite number of categories.* Then we can decompose the above expectation into

  $$\mathbb{E}(L(f(X),Y)) = \sum_x \mathbb{E}(L(f(X),Y)|X=x)p(x). \tag{5}$$

*Suppose that $\mathcal{F}$ contains all possible functions.* Then the optimal $f^*$ can be decomposed as

$$f^*(x) = \mathsf{argmin}_b \mathbb{E}(L(b,Y)|X=x) = \mathsf{argmin}_b \int L(b,y)p(y|x)dy$$

for each $x$. Consider a general weighting estimating

$$f_\omega = \mathsf{argmin}_{f\in\mathcal{F}} \mathbb{E}\left(\frac{R}{\omega(X)}L(f(X),Y)\right).$$

Equation (5) and the abundance of $\mathcal{F}$ implies that we can decompose it as

$$\begin{aligned} f_\omega(x) &= \mathsf{argmin}_b \mathbb{E}\left(\frac{R}{\omega(x)}L(b,Y)|X=x\right) \\ &= \mathsf{argmin}_b \frac{\pi(x)}{\omega(x)}\int L(b,y)p(y|x)dy \\ &= \mathsf{argmin}_b \int L(b,y)p(y|x)dy = f^*(x). \end{aligned}$$

Again, we recover the optimal predictor even if the weights are not correct.

Two key assumptions are: 1. $X$ being categorical with finite number of categories and 2. $\mathcal{F}$ contains all possible functions. The first condition can be relaxed to continuous but when $X$ is continuous, we need to impose smoothness assumption on $p(y|x)$ as a function of $x$ since we will have 0 observations with exactly a value of $X=x$. The second assumption allows us to decompose the minimization to locally at each $x$ because it contains every possible functions. We can relaxed it into assuming $f^*\in\mathcal{F}$.

- **Relation to minimax causal inference.** The analysis in this note is also related to causal inference, in particular, the minimax framework of causal inference[2]. Note that the causal here refers to *the causal effect from $X$ on $Y$*. Under the MAR, the complete data distribution $p(y,x|R=1) = p(y|x)p(x|R=1)$ and the full-data distribution $p(y,x) = p(y|x)p(x)$ differ by the covariate distribution. Any weighting approach is implicitly assuming a hypothetical distribution of $p(x|R=1)$. When using the IPW approach, the resulting completely observed data behaves like from the distribution

$$\begin{aligned} \frac{p(y,x|R=1)}{P(R=1|x)} &= p(y|x)\frac{p(x|R=1)}{P(R=1|x)} \\ &= p(y|x)\frac{p(x|R=1)}{P(R=1|x)}\frac{P(R=1)}{P(R=1)}\frac{p(x)}{p(x)} \\ &= p(y|x)\frac{p(x,R=1)}{P(R=1,x)}\frac{p(x)}{P(R=1)} \\ &= p(y|x)\frac{p(x)}{P(R=1)} \\ &= p(y,x)/p(R=1). \end{aligned}$$

So any weighting $\omega(x)$ can be viewed as a change of the covariate distribution.

Also, from the above analysis, if the conditional model is correct, any weighting function will not change the target parameter. Under the minimax causal framework, this implies that we can learn more

[2] See http://faculty.washington.edu/yenchic/short_note/note_causalminimax.pdf.

causal effects by comparing several weighting functions. Take the linear regression as an example. For a given weight function $\omega \in \Omega$, the parameter in equation (4) can be expressed as

$$\beta_\omega = \mathsf{argmin}_\beta \mathbb{E}\left(\frac{R}{\omega(X)}(Y - X^T\beta)^2\right).$$

Then the parameter

$$\beta_\Omega = \mathsf{argmin}_\beta \sup_{\omega\in\Omega} \mathbb{E}\left(\frac{R}{\omega(X)}(Y - X^T\beta)^2\right)$$

is the minimax causal parameter under such a linear model.

- **Covariate shift in domain adaptation/transfer learning.** The above analysis is also related to the covariate shift problem in domain adaptation/transfer learning. A typical scenario in domain adaptation is that our training data is from one population while we want to make predictions on another population.

  In the covariate shift problem, the conditional relation $m(X) = \mathbb{E}(Y|X)$ will not change between the two populations but the marginal distribution of the covariates/features $X$ are often different.

  A common strategy in this case is to apply an re-weighting approach to account for the distributional differences. This idea is in spirit, the same as the IPW but it often takes the form of *importance* weighting. As a concrete example, suppose that in the training population, the covariates are from the PDF $p(x)$ while in the population we want to predict, the covariates are from the PDF $q(x)$. Then we will weight each observation by $W_i = \frac{q(X_i)}{p(X_i)}$.

  Similar to our analysis, if the conditional model is correct, i.e., $p(y|x;\theta)$ or $m(y|x;\beta)$ is correct, then even if our weights are incorrectly specified, we will still obtain a consistent predictor. So if the model is correct, not accounting for the population difference will not matter too much.

  However, in general, we will not assume the model to be correct but just a model that we use to train our predictor. In particular, in most of the domain adaptation problems, we will not assume the model to be correct. So accounting for the weights is still vital in the analysis.