Inverse Probability Weighted Estimation and Augmentation

Inverse Probability Weighted Estimation

Motivation from survey sampling

- Recall that in suvey sampling, we collect data $(Y_1, ..., Y_n)$, each with sampling weights $(w_1, ..., w_n)$.
- The data can be viewed as missing data, where full data are the observations from the population $(Y_1, ..., Y_N)$, and the missing indicator R_i denotes whether subject i is sampled. Thus, $P(R_i = 1) = w_i$ for $i = 1, ..., N$.
- IPW estimator for the population sum is essentially Horwitz-Thompson estimator:

$$\sum_{i=1}^{N} \frac{R_i}{w_i} Y_i.$$

- The estimator is unbiased.

Further motivation from survey sampling

- Suppose that we also have auxiliary information X_i for each subject in the population.
- Furthermore, suppose that we have a model to predict Y_i based on X_i and the prediction is $\mu(X_i)$.
- One estimator (called model-assisted estimator or difference estimator) is

$$\sum_{i=1}^{N} \frac{R_i}{w_i}(Y_i - \mu(X_i)) + \sum_{i=1}^{N} \mu(X_i)$$

$$\equiv \sum_{i=1}^{N} \frac{R_i}{w_i} Y_i - \sum_{i=1}^{N} \frac{R_i - w_i}{w_i} \mu(X_i).$$

Further motivation from survey sampling

- Such an estimator is also unbiased even if $\mu(X_i)$ may be a bad prediction.
- What is the asymptotic variance? What $\mu(X_i)$ yields the smallest variance?
- The closer $\mu(X_i)$ is to $E[Y_i|X_i]$, the smllaer the variance is.
- In practice, we can fit a regression model using the sampled data and the regression is weighted by the sampling weights.

What can we learn?

- This is a robust approach to estimate the parameter of interest, as it doesn't require any additional model or assumption for the data distribution as required in the likelihood approach.
- IPW is a simple construction to produce unbiased (consistent) estimators; however, additional information/model can lead to a class of unbiased estimators through augmenting IPW.
- Some augmentation can yield the most efficient estimator in this class.

Regression model in i.i.d setting

- Consider the full data of $(Y_i, X_i), i = 1, ..., n$.
- The model of interest is $E[Y|X] = \beta X$.
- In the missing data, some Y_i are not observed. We use $R_i = 1$ to denote non-missing.
- IPW basically constructs an estimating equation for β using the complete data, while each subject is weighted by the probability of not-missing (like sampling survey, the probability of being selected into the analysis):

$$\sum_{i=1}^{n} \frac{R_i}{p_i} X_i (Y_i - \beta X_i) = 0.$$

Estimation of weights

- However, different from the sampling survey, p_i is usually not observed so it must be estimated using data.
- Under MAR that R_i is independent of Y_i given X_i and Z_i (Z_i contains auxiliary information but may be empty), $p_i = P(R_i = 1|X_i, Z_i)$.
- p_i can be estimated using the observations (R_i, X_i) for $i = 1, ..., n$.

Remarks on IPW estimator

- The IPW estimator is unbiased!
- The same idea can be applied to any missing data context, i.e., performing the analysis (regression, machine learning algorithms) using the complete data while assigning each subject the probability of not-missing.
- IPW is simple and easy to implement; however, it is usually not efficient because only uses partial data.

Augmented IPW estimator

- Following a similar construction, suppose that $\mu(X_i, Z_i)$ predicts $E[Y_i|X_i, Z_i]$. Then an augmented IPW estimating equation is

$$\sum_{i=1}^{n} \frac{R_i}{p_i} X_i(Y_i - \beta X_i) - \sum_{i=1}^{n} \frac{R_i - p_i}{p_i} X_i(\mu(X_i, Z_i) - \beta X_i) = 0.$$

- AIPW estimator is unbiased (asymptotically).
- The choice of $\mu(X_i, Z_i) = E[Y_i|X_i, Z_i]$ yields the most efficient estimator in this class.
- We can fit a model using the complete data weighted by $1/p_i$ to estimate $\mu(X_i, Z_i)$.

Special properties of AIPW

- First, AIPW is robust without specifying the full distribution of (Y_i, X_i, Z_i).
- If p_i is correct (the model for missingness), AIPW is unbiased even if $\mu(X_i, Z_i)$ is not correct for estimating $E[Y_i|X_i, Z_i]$.
- If $\mu(X_i, Z_i) = E[Y_i|X_i, Z_i]$, AIPW remains to be unbiased even if p_i is wrong.
- This is called the doubly robustness property.

Additional properties of AIPW

- If p_i is correctly estimated at $\sqrt{n}$-rate, then the AIPW estimator for β is $\sqrt{n}$-rate even if $\mu(X_i, Z_i)$ is estimated slower than $\sqrt{n}$-rate (but faster than $n^{-1/4}$).
- This is called the de-bias property.
- In other words, we can estimate $\mu(X_i, Z_i)$ using nonparametric or machine learning methods and allow the dimensionality of Z_i to be ultra-high.
- Such a property is particularly useful to make inference for β in high-dimensional settings or when applying machine learning methods to estimate $E[Y_i|X_i, Z_i]$.

Generalization

- The above construction only considers a simple missing pattern (missing vs not missing one particular variable).
- But what about more general and complex missing patterns? For example, in a longitudinal study, subjects may miss measurements at different time points.
- We need a general framework to obtain a similar class of estimating equations.
- This entails knowledge of semiparametric efficiency theory in the next module.