Variable Selection, Doubly Robust Estimation, and Closing Comments

Beth Ann Griffin

Variable Selection

Modern Methods to Estimate Propensity Score Weights

Session 2: Other Methods for Variable Selection

Beth Ann Griffin
Dan McCaffrey

Review of Variable Selection for Causal inference

- ***The goal:*** **select a subset of observed covariates to include in the propensity score model**
- ***The issue:*** **traditional variable selection methods are designed for prediction, not effect estimation**
- ***The solution(s):*** **???**

Review of Variable Selection for Causal inference

- While propensity score methods have gained widespread use, there still remains confusion and a lack of guidance on how best to carry out variable selection
- The tools we have discussed today seek to optimize pretreatment covariate balance using GBM
  - Variable selection is imbedded within the algorithm
- What are the alternatives?

Controlling for the Widest Set Possible

- Rubin (2009) argues that controlling for the widest set of pretreatment characteristics protects against unobserved confounding
- What happens if the number of pretreatment covariates is large?
- VanderWheele and Shpitser (2011) show that controlling for all pretreatment covariates may lead to biased treatment effect estimates

The Use of Expert Knowledge

- Robins (2001) and Hernan et al. (2002) argue that full or partial knowledge of the underlying causal structure is required to select confounders
- In other words, we must use substantive knowledge to select the covariates to include in the analysis
- What happens if the number of covariates is large?

VanderWheele and Shpitser (2011)

- Argue that bias is removed by adjusting for all covariates that cause treatment or outcome
- A quote: ***An investigator simply need ask, ”Is the covariate a cause of the treatment?” and ”Is the covariate a cause of the outcome?”***
  - If the answer is yes to either, include in the analysis
- What happens if this set of covariates is large?
- Suggest iteratively discarding variables unassociated with the outcome

Data Driven Approaches for Variable Selection

- Previous methods may be sufficient for removing bias but are likely to be inefficient for estimating treatment effects
- Recent data driven approaches for variable selection improve efficiency without sacrificing unbiasedness
- *General theme:* confounders are related to both treatment and outcome

Data Driven Approaches for Variable Selection

- **Three possibilities:**
  - **Variable selection based only on the propensity score model**
  - **Variable selection based only on the outcome model**
  - **Combination of the two**

Selection based on propensity score model fit

Procedures that exclusively optimize the fit of the propensity score are inefficient (Schnitzer, 2015)

- **Variable selection on the propensity score model may remove true confounders if the confounders are relatively weak predictors of treatment**
- **They will be more likely to select instruments into the propensity score model, which are known to increase the variance**
- **They will not select covariates that only predict the outcome – including strong predictors of the outcome is thought to reduce the variance**

Selection based on outcome model fit

Selection only based on the outcome model suffers from similar issues (Schnitzer, 2015)

- **Variable selection on the outcome model may remove true confounders if the confounders are relatively weak predictors of outcome**

Collaborative targeted maximum likelihood

- Consider using a doubly robust estimator
  - The propensity score and outcome models need to be specified in such a way that the combined models collaboratively adjust for a sufficient confounder set
- van der Laan and Gruber (2010) argue that when the models jointly contain a sufficient confounder set, the doubly robust estimator might be consistent
  - Even if neither model contains a sufficient confounder set on its own
- C-TMLE exploits this property to perform variable selection
- Suite of R functions available at: http://www.stat.berkeley.edu/~laan/Software/

de Luna, Waernbaum, and Richardson (2011)

- **Described the necessary assumptions for the existence and identification of minimal sufficient adjustment sets of confounders in the nonparametric setting**
- **Proposed generic variable selection algorithms to obtain such a minimal confounder set using conditional independences:**
  - **First, partition the covariates X into** (X_1, X_2) **such that** $T \perp X_2 | X_1$
  - **Then, for** $j = 0, 1$ **select** $Q_j \subset X_1$ **such that** $(Y_j \perp X_1 \backslash Q_j \mid Q_j, T = j)$
- **Software available in the R package CovSel**

Bayesian model averaging

- Series of papers have looked at modifying Bayesian model averaging to select confounders:
  - Crainiceanu, Dominici, and Parmigiani (2008)
  - Wang, Parmigiani, and Dominici (2012)
  - Zigler and Dominici (2014)
  - Wang, Parmigiani, Dominici, and Zigler (2015)
  - Cefalu, Dominici, Arvold, and Parmigiani (2015)

Also see

- Vansteelandt, Stijn, Maarten Bekaert, and Gerda Claeskens. ”On model selection and model misspecification in causal inference.” Statistical methods in medical research 21, no. 1 (2012): 7-30.
- Ertefaie, Ashkan, Masoud Asgharian, and David A. Stephens. ”Variable Selection in Causal Inference Using Penalization.” arXiv preprint arXiv:1311.1283 (2013).
- Wilson, Ander, and Brian J. Reich. ”Confounder selection via penalized credible regions.” Biometrics 70, no. 4 (2014): 852-861.

Doubly Robust Estimation

Approaches Discussed Do Not Directly Use Relationship Between Covariates and Outcomes

- Often cited as a benefit
  - Adjustments determined prior to studying outputs avoiding potential post hoc tuning to achieve desired results
- Traditional methods often directly model the outcome
- We might be throwing away useful information

Combining Models for Treatment Assignment and Models for the Outcomes

- If we had the correct model for the outcomes, modeling approaches would be most efficient
  - Provide smallest standard errors or smallest mean-squared error (MSE)
- But there is a risk for large bias if the model is incorrect
- Weighting can yield unbiased (consistent) estimates even if we don’t know the model for the outcomes
- Can be inefficient relative to modeling with the correct model even if we know the correct weights
- Can be biased if the weighting function is wrong

Combine Models for Treatment Assignment and Models for the Outcomes (cont.)

- Combining the methods can potentially
  - Improve efficiency over weighting alone
  - Remove bias if the outcomes model is incorrect but the model for treatment assignment is correct
  - Remove bias if the outcomes model is correct but the model for treatment assignment is incorrect
- Methods that achieve these goals are called ***Doubly Robust* (DR)**

Doubly Robust Methods

- Early examples derived in survey sampling and estimating finite population quantities
  - Model-assisted estimation
  - Generalized regression estimation
- Robins and colleagues developed theory for generating an entire class of doubly robust estimators
  - Consistent and asymptotically normal

DR Estimation

- $E(Y_Z|X) = h(Z, X; \beta)$ **for** $Z = 0, 1$ **– mean model**
- $Pr(Z = 1|X) = p(X; \delta)$ **– selection model**
- **We will estimate both of these models and combine them to obtain estimates that are DR**
- **Focus on estimating ATE and note modifications for ATT**

Multiple Ways to Combine Treatment and Mean Modeling

- Multiple ways to combine models for treatment and models for the mean that are DR
- Weighted linear regression
- Include function of the propensity scores in mean model
- Combine predictions and weighted means

Weighted Linear Regression

- Suppose $h(Z, X; \beta)$ is linear
- A weighted regression analysis of the outcome Y on the treatment indicator Z and the covariates X is DR
  - Weights equals standard propensity score weights or the inverse probability of treatment weights
  - $w_i = Z_i/p(X_i; \widehat{\delta}) + (1 - Z_i)/(1 - p(X_i; \widehat{\delta}))$ for ATE
  - $w_i = Z_i + (1 - Z_i)p(X_i; \widehat{\delta})/(1 - p(X_i; \widehat{\delta}))$ for ATT
- Standard result that if the model is correct weighted least squares is unbiased or consistent
- Let $\bar{Y}_{w1}$, $\bar{Y}_{w0)}$, $\bar{X}_{w1}$, $\bar{X}_{w)}$ equal the weighted means of the outcomes and covariates for treatment and control groups
- $\widehat{TE} = \bar{Y}_{w1} - \bar{Y}_{w0} - \widehat{\beta}'(\bar{X}_{w1} - \bar{X}_{w0})$

Weighted Regression When Mean Model Is Not Linear

1. **Estimate model coefficients ($\widehat{\beta}_w$), weighting each case by propensity score weight, e.g.,**
   $w_i = Z_i/p(X_i;\widehat{\delta}) + (1-Z_i)/(1-p(X_i;\widehat{\delta}))$

2. **Estimate the expected outcome assuming treatment for all cases:** $m_{w,i}(1) = h(1, x_i; \widehat{\beta}_w)$

3. **Estimate the expected outcome assuming control for all cases:**
   $m_{w,i}(0) = h(0, x_i; \widehat{\beta}_w)$

4. **The treatment effect on the population equals**
   $\sum_i m_{wi}(1)/n - \sum_i m_{wi}(0)/n$ **– the difference in the averages of the predicted values**

- **Sometimes called "recycling"**

Augmented Regression Can Be DR

- **Augmented regression includes a function of propensity score in the model**
  - **Fit** $h(Z, X, \phi(p(X;\delta)))$ **to the outcomes**
  - **Involves no weighting**
- **Motivation:** $E(Y_z|p(x), Z = z) = E(Y_z|p(X))$ **so that models which include** $p(X)$ **or appropriate functions of it fit to the sample where** $Z = z$ **can yield estimates of the mean for the counterfactual, even if the rest of the model is incorrect**
- **There are many variants of this approach including:**
  - **Smooth functions or** $log(p(x)/(1 - p(x)))$ **or** $log((1 - p(x))/p(x))$ **(Little and An, 2004)**
  - **Bin probabilities into small number of categories (Kang and Schafer, 2008)**
  - $Z/p(X) + (1 - Z)/(1 - p(X))$ **(Bang and Robins, 2005)**

Bias Corrected DR

- **Fit mean and propensity score models (no weighting)**
- **Calculate** $\widehat{\mu}_1 = \frac{1}{n}\sum_i \frac{(Y_i - h(1, X_i; \widehat{\beta}))Z_i}{p(X_i; \widehat{\delta})} + \frac{1}{n}\sum_i h(1, X_i; \widehat{\beta})$
- **Calculate** $\widehat{\mu}_0 = \frac{1}{n}\sum_i \frac{(Y_i - h(0, X_i; \widehat{\beta}))(1 - Z_i)}{1 - p(X_i; \widehat{\delta})} + \frac{1}{n}\sum_i h(0, X_i; \widehat{\beta})$
- **Estimate ATE by** $\widehat{\mu}_1 - \widehat{\mu}_0$
- **Modification: Replace** $\frac{1}{n}\sum_i \frac{(Y_i - h(1, X_i; \widehat{\beta}))Z_i}{p(X_i; \widehat{\delta})}$ **by** $\frac{\sum_i (Y_i - h(1, X_i; \widehat{\beta}))Z_i / p(X_i; \widehat{\delta})}{\sum_i Z_i / p(X_i; \widehat{\delta})}$ **and** $\frac{1}{n}\sum_i \frac{(Y_i - h(0, X_i; \widehat{\beta}))(1 - Z_i)}{1 - p(X_i; \widehat{\delta})}$ **by** $\frac{\sum_i (Y_i - h(0, X_i; \widehat{\beta}))Z_i / (1 - p(X_i; \widehat{\delta}))}{\sum_i Z_i / (1 - p(X_i; \widehat{\delta}))}$
- **For ATT, use the sample mean treatment and**

$$\widehat{\mu}_{0,\mathrm{ATT}} = \frac{1}{n}\sum_i \frac{(Y_i - h(0, X_i; \widehat{\beta}))(1 - Z_i)p(X_i; \widehat{\delta})}{1 - p(X_i; \widehat{\delta})} + \frac{1}{n}\sum_i h(0, X_i; \widehat{\beta})Z_i$$

Checking that Bias Corrected DR is Doubly Robust

- $\widehat{\mu}_1 = \frac{1}{n}\sum_i \frac{(Y_i - h(1,X_i;\widehat{\beta}))Z_i}{p(X_i;\widehat{\delta})} + \frac{1}{n}\sum_i h(1,X_i;\widehat{\beta})$
- **If h is correct, then**
  - $(Y_i - h(1,X_i;\widehat{\beta})$ **estimates the residual errors which, by strong ignorability, have mean zero for $Z = 1$ and are independent of X, so that $\frac{1}{n}\sum_i \frac{(Y_i - h(1,X_i;\widehat{\beta}))Z_i}{p(X_i;\widehat{\delta})}$ converges to zero**
  - $\frac{1}{n}\sum_i h(1,X_i;\widehat{\beta})$ **converges to $E[h(1,X;\beta)] = E(Y_1)$**
- **If $p(X_i)$ is correct, then**
  - $\frac{1}{n}\sum_i Y_i Z_i / p(X_i;\widehat{\delta})$ **converges to $E(Y_1)$**
  - **Both $\frac{1}{n}\sum_i h(1,X_i;\widehat{\beta}) Z_i / p(X_i;\widehat{\delta})$ and $\frac{1}{n}\sum_i h(1,X_i;\widehat{\beta})$ are estimating the expected value of predictions from misfit mean model on the whole population**
    - **Their difference converges to zero**

Combination Common in Practice

- Fit linear model with covariates with poor balance
- Combine GBM for mean and weighting
  - Weighted GBM for mean and use recycling
  - Unweighted GBM for mean and bias corrected DR
- Specify process prior to observing mean or or use automated mean fitting model

Closing Remarks

Conclusions

- Hopefully today has increased your understanding of causal effects and the role that propensity score weighting can play in estimation of those effects
- This is an active and rich field
- Forthcoming work from my team:
  - Causal Mediation
  - Causal Moderation
  - Continuous Treatments
  - TWANG for Big Data

TWANG Home
- Frequently Asked Questions (FAQ)
- Downloads
- Tutorials
- Publications
- Workshops
- Project Members
- Contact Us

RAND > Statistics Group >

The Toolkit for Weighting and Analysis of Nonequivalent Groups, or **TWANG**, contains a set of functions to support causal modeling of observational data through the estimation and evaluation of propensity score weights. The TWANG package was first developed in 2004 by RAND researchers for the R statistical computing language and environment. The R version of the package contains functions for creating high-quality propensity score weights which can be used to estimate treatment effects with two or more treatment groups and time-varying treatments.

In 2014, TWANG macros were developed for SAS and Stata to support the use of these tools

Stay Informed

Sign up to get notified about updates and new features in TWANG.

Your participation helps TWANG support future development efforts.

Email Address: *(required)*

SUBSCRIBE

Privacy Statement »

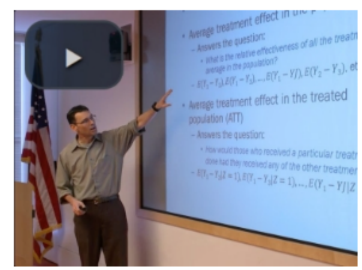
Featured Researchers

Beth Ann Griffin

Senior Statistician

Beth Ann Griffin is a senior statistician at the RAND Corporation. Her research has largely focused on causal effects estimation when using observational data. Her substantive research has primarily fallen into three areas: (1) substance abuse treatment for adolescents, (2) the impact of...

Instructional Videos



TWANG Short Course/Educational Videos: Three Videos — Introduction, Propensity Score Weighted Analyses with 2 Groups, and Propensity Score Weighted Analyses with More Than 2 Groups