Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach

Anastasios A. Tsiatis1, Marie Davidian1,*,†, Min Zhang1, and Xiaomin Lu2, Statisitics in Medicine

Improving Efficiency of Inferences in Randomized Clinical Trials Using Auxiliary Covariates

- Author(s): Min Zhang, Anastasios A. Tsiatis and Marie Davidian
  Source: Biometrics, Vol. 64, No. 3 (Sep., 2008), pp. 707-715

Basic Idea

- Estimand is unconditional average causal treatment effect for a randomized clinical trial, with a univariate outcome variable
- There may be important covariates that predict outcome
- The covariates will be balanced by randomization
- The method will be semiparametric and asymptotic, the estimator will be asymptotically unbiased for the average causal treatment effect and will be asymptotically normal.
- The method will be optimal over all similar estimators

Notation

Unconditional Average Causal Effect Estimand:

$$\beta = E(Y \mid Z=1) - E(Y \mid Z=0),$$

conditional Causal Effect Estimand:

$$\beta_x = E(Y \mid Z=1, X=x) - E(Y \mid Z=0, X=x).$$

Form of estimator

$$\overline{Y}^{(1)} - \overline{Y}^{(0)} - \sum_{i=1}^{n}(Z_i - \overline{Z})\left\{n_0^{-1}h^{(0)}(X_i)+n_1^{-1}h^{(1)}(X_i)\right\},$$

1. Second term has mean zero from randomization
2. Unbiased estimate of treatment difference
3. Second term might reduce variance

Estimator with lowest variance

$$h^{(k)}(X_i)=E(Y_i|Z_i=k,X_i),\ \ k=0,1;$$

Heuristic rationale

- E[Y|z,x]=$h_z(X)+\beta z$
- $E[h_1(X)] = E[h_0(X)]$
- Note: $^{-z}$ implies the average over the respective value of z, $^{-}$ over the all the data
- Raw Difference: $\bar{Y}^1 - \bar{Y}^0 = \beta + \overline{h_1(X)}^1 - \overline{h_0(X)}^0$
- Second Term: $\sum(Z - \bar{z})(\frac{E[Y|z=0,X]}{n0} + \frac{E[Y|z=1,X]}{n1})$
- $n0/(n0+n1)(\left(\frac{n1}{no}\right)\ \overline{h_0(X)}^1 + \beta + \overline{h_1\ (X)}^1)$

$$-n1/(n0+n1)(\overline{h_0(X)}^0 + \left(\frac{n0}{n1}\right)(\beta + \overline{h_1\ (X)}^0))$$

- Estimate = difference between raw and second term= $\beta + \overline{h_1(X)}\ - \overline{h_0(X)}$
- Has lower variance because $^{-}$ has a lower variance then $^{-z}$ because it as a larger sample size. Adjusts the comparison to the population X rather than X in each group. Conjecture: Yields an unbiased estimator, if X is not independent of Z

Estimation Strategy

- Divide the data into the two treatment groups
- In each treatment group find the best estimate of E(Y|X) using any method you like. It does not have to be preplanned
- Note that the method used on treatment group 1 will be used on treatment group 0, so overfitting could be a problem, this suggests methods that avoid overfitting, regression with prespecified covariates, LASSO, CART, Random Forests
- Predict using both methods on all the data, to create $h_1(X)$ and $h_0(X)$

Simple Example

- E(Y|z=1,x)=$\mu + \beta x$
- E(Y|z=0,x)=μ
- Note(E(Y|z=1)-E(Y|z=0)=$\beta E(X)$
- $\overline{Y^1}$-$\overline{Y^0}$ - $\{\sum_{z=1} 1/2(\frac{\mu}{n} + \frac{(\mu+\beta x)}{n})$-$\sum_{z=0} \frac{1}{2}(\frac{\mu}{n} + \frac{\mu+\beta x}{n})\}$
- $\overline{Y^1}$-$\overline{Y^0}$ -$\{\beta/2(\bar{x}^1 - \bar{x}^0)$

Simple Example-that turns out to be complicated

- Simple regression E(Y|z,x)=u+ $E(Y|z,x) = \mu + \beta z + \gamma x$
- E(y|z=1,x)=$\mu + \beta + \gamma x$
- E(y|z=0,x)=$\mu + \gamma x$

$$h^{(0)}(X_i)=h^{(1)}(X_i)=\sum_{XY}^{T}\sum_{XX}^{-1}X_i,$$

$$\sum_{XY}=E[\{X-E(X)\}\{Y-E(Y)\}],\ \sum_{XX}=E[\{X-E(X)\}\{X-E(X)\}^T],$$

Extension to non-linear models

- Logistic model E(Y|z)=$Exp(\mu + \beta z)/(1 + Exp(\mu + \beta z))$
- Note unconditional model(above) and conditional model(below) have different estimands for β
- $E(Y|z,x) = Exp(\mu + \beta z + \gamma x)/(1 + Exp(\mu + \beta z + \gamma x))$
- This is an example of Simpsons paradox, but in a clinical trial both estimands are of interest. The first is what is the average effect of treatment expressed as an odds ratio, the second is the effect of treatment for a person of with covariate x, when we aggregate the effect of treatment gets smaller

Borrowed slide from the master himself

Estimating functions using auxiliary covariates

Main result: For a given *semiparametric model* members of the *class of all unbiased estimating functions for* θ using *all of* (Y, Z, X) may be written

$$m^*(Y, Z, X; \theta) = m(Y, Z; \theta) - \{Z - \pi\}a(X)$$

- $m(Y, Z; \theta)$ is a *fixed* unbiased estimating function for θ without auxiliary covariates
- $a(X)$ is an arbitrary function of X
- $a(X) \equiv 0 \Rightarrow$ "*unadjusted estimator*" $\widehat{\theta} = (\widehat{\beta}, \widehat{\alpha})$
- "*Augmentation term*" effects the "*adjustment*"

Optimal estimating function in the class: Elements of the estimator have *smallest asymptotic variance*

- Take $a(X) = E\{m(Y, Z; \theta) \,|\, X, Z = 1\} - E\{m(Y, Z; \theta) \,|\, X, Z = 0\}$
- *Optimal estimating equation*

$$\sum_{i=1}^{n} \Big(m(Y_i, Z_i; \theta) - (Z_i - \pi)\big[E\{m(Y, Z; \theta) \,|\, X_i, Z = 1\} - E\{m(Y, Z; \theta) \,|\, X_i, Z = 0\}\big] \Big) = 0$$

- $E\{m(Y, Z; \theta) \,|\, X, Z = g\}, g = 0, 1$ are *unknown functions of* X $\Rightarrow$ *model them*...

Implementation

Approach: *Adaptive algorithm*

(1) Solve $\sum_{i=1}^{n} m(Y_i, Z_i; \theta) = 0 \Rightarrow \widehat{\theta}$

(2) For *each group* $g = 0, 1$ *separately*, using the "*data*" $m(Y_i, Z_i; \widehat{\theta})$ for $Z_i = g$, develop a *regression model*

$$E\{m(Y, g; \widehat{\theta}) \,|\, X, Z = g\} = q_g(X, \zeta_g),$$

$$q_g(X, \zeta_g) = \{1, c_g^T(X)\}^T \zeta_g,$$

and obtain $\widehat{\zeta}_g$ by *OLS separately*

(3) For each $i = 1 \ldots, n$, form *predicted values* $q_g(X_i, \widehat{\zeta}_g)$ for each $g = 0, 1$ and solve in θ with $\widehat{\pi} = n^{-1} \sum_{i=1}^{n} Z_i$

$$\sum_{i=1}^{n} \Big[m(Y_i, Z_i; \theta) - (Z_i - \widehat{\pi})\{q_1(X_i, \widehat{\zeta}_1) - q_0(X_i, \widehat{\zeta}_0)\} \Big] = 0 \;\Rightarrow\; \text{"}\textit{adjusted}\text{"}\; \widetilde{\theta}$$

Differences for linear and general case

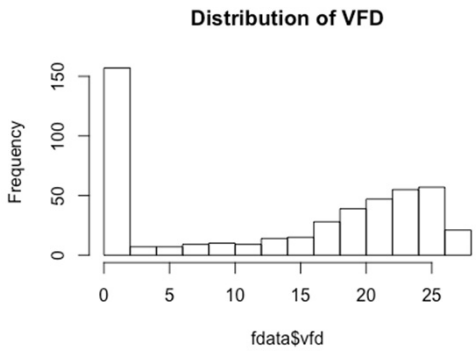
- In general case you first estimate treatment effect(and other parameters) without covariates
- Second step is to use this estimate to model the estimating equations separately for each treatment group. In the linear case this would be equivalent to modeling the residuals.(For proportional hazards model it would be the **Schoenfeld** residuals)
- Use the predictions to calculate the correction term and with this re-estimate the parameters

Software isn't rocket science

```
correctedEstimate=function(y,z,e0,e1,dat){
 ##y is outcome, z is treatment e0,e1 are predictions
 dat=dat[is.finite(dat[,y])&is.finite(dat[,z])&is.finite(dat[,e0])&is.finite(dat[,e1]),] #remove missing values
 ys=lapply(c(1,0),function(x) dat[dat[,z]==x,y])  #extract data for each treatment group
 convEstimate=mean(ys[[1]])-mean(ys[[2]])  #conventional estimate and SEE
 stderrConvEstimate=sqrt(var(ys[[1]])/length(ys[[1]])+var(ys[[2]])/length(ys[[2]]))
 rx=lapply(c(1,0),FUN=function(x) dat[,z]==x)  #treatment predicate
 n=sapply(rx,function(x) sum(x,na.rm=TRUE))  #sample sizes
 nn=sum(n)
 correction=(dat[,z]-mean(dat[,z]))*(dat[,e0]/n[1]+dat[,e1]/n[2])
 newEstimate=convEstimate-(sum(correction))
 C=(nn-8)/(nn-4)
 yy=sapply(rx,FUN=function(rx) mean(dat[rx,y],na.rm=TRUE))
 zz=rx[[1]]-mean(rx[[1]])
 f=list(dat[,e1],dat[,e0])
 vbeta=C*sum(((rx[[1]]/n[1]-rx[[2]]/n[2])*dat[,y]-newEstimate/nn-correction
          -zz*((yy[[1]]-mean(f[[1]][rx[[1]]]))/n[1]+(yy[[2]]-mean(f[[2]][rx[[2]]]))/n[2]))^2)
 return(c(convEstimate=convEstimate,stderr=stderrConvEstimate,
       newEstimate=newEstimate,stderr=sqrt(vbeta),
       pConv=pchisq( (convEstimate/stderrConvEstimate)^2,1,lower.tail=FALSE),
       pNew=pchisq(newEstimate^2/vbeta,1,lower.tail=FALSE)))
```

Example Vent Free Days in ARDS

- Vent Free Days: 28-Number of days on ventilator; 0 if Died



Covariates

- Pafi0: P/F ratio, measure of lung function
- Age;
- Apache3; Measure of Burden of Ilness
- Map; Mean arterial pressure-blood pressure

Run Software

- modr=lm(vfd~liberal+pafi0+age+apache3+map,data=fdata)
- mods=lapply(c(1,0),function(x) predict(lm(vfd~pafi0+age+apache3+map,data=fdata[fdata$liberal==x,]),fdata))
- vs=data.frame(fdata,e1=mods[[1]],e0=mods[[2]])
- correctedEstimate("vfd","liberal",'e1','e0',dat=vs)
- ###ARDS randomForests
- mods=lapply(c(1,0),function(x) predict(randomForest(vfd~pafi0+age+apache3+map,data=fdata[fdata$liberal==x,]),fdata))
- vs=data.frame(fdata,e1=mods[[1]],e0=mods[[2]])
- correctedEstimate("vfd","liberal",'e1','e0',dat=vs)

Conventional Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	23.795418	3.527437	6.746	4.5e-11	***
liberal	-1.122659	0.856627	-1.311	0.19065	
pafi0	0.020075	0.007691	2.610	0.00934	**
age	-0.064458	0.029051	-2.219	0.02698	*
apache3	-0.127791	0.014307	-8.932	< 2e-16	***
map	0.038073	0.032023	1.189	0.23507	

Linear Model

convEstimate	stderr	newEstimate	stderr	pConv	pNew
-1.48	0.962	-1.12	0.847	0.122	0.184

Using a Linear Model there is no benefit of this method

Random Forests

- mods=lapply(c(1,0),function(x) predict(randomForest(vfd~pafi0+age+apache3+map,data=fdata[fdata $liberal==x,]),fdata))

Results using Random Forests

convEstimate	stderr	newEstimate	stderr	pConv	pNew
-1.488	0.962	-1.215	0.657	0.122	0.0645

Uptake

- About 200 citations
- Seems to be used by only a few disease areas
- I think we need to develop a disease specific modeling strategy then develop software to use this as a standard analysis

Issues

- The ideal is to have independent groups model each treatment
- One group could do the analysis as long as the p-value for the method isn't a criteria in the modeling. They don't do what I just did☺
- I prefer prespecifying a general algorithm that guards against overfitting such as LASSO or Random Forests.
- Playing around with different algorithms guarantees overfitting