IDS 702: Module 6.4

Regression-based estimation and covariate balance

Dr. Olanrewaju Michael Akande

Estimation: regression-based

- With unconfoundedness and overlap, we can move on to estimation.
- Clearly, we need to adjust for any difference in the outcomes due to the differences in pre-treatment characteristics.
- Commonly via a regression model for the potential outcome on covariates
- However,

  1. validity of the analysis critically relies on the validity of the unconfoundedness assumption (which, remember is untestable); and
  2. usually, model parameters do not directly correspond to the causal estimand of interest.

Estimation: regression-based

- For example, consider two regressions, one for each potential outcome. Write the mean functions as

  $$\mathbb{E}[Y(1)|X=x]=\mu_1(x),\quad \mathbb{E}[Y(0)|X=x]=\mu_0(x).$$

  This need not be two separate regressions, but could be a regression with W included as a predictor.

- Let $\hat{\mu}_w(X_i)$ denote the fitted potential outcome for $Y_i(w)$ based on the regression models.

- For ATE, the covariate-adjusted estimator is then

  $$\hat{\tau}_{\text{adj}}=\sum_{i=1}^{N}\frac{W_i(Y_i^{\text{obs}}-\hat{\mu}_0(X_i))+(1-W_i)(\hat{\mu}_1(X_i)-Y_i^{\text{obs}})}{N}$$

- Unlike randomized experiments, the estimator is not consistent if the linear model is misspecified.

Estimation: regression-based

- Variance can be estimated using bootstrap.

- Note that regression itself does not take the lack of overlap into account.

- If the imbalance of the covariates between the two groups is large, the model-based results heavily relies on extrapolation in the non-overlap region, which is sensitive to the model specification assumption.

- Take away: Regression (or any model) here comes with a package. You need to know and acknowledge what assumptions—explicit or implicit—come with that model.

Strategies for mitigating model dependence

- To mitigate model dependence in the case of linear regression, there are two general strategies

  1. Attempt to fix the design - balance covariates
  2. Use more flexible model for analysis

- Best strategy is to actually use both jointly: first balance covariates in the design stage, then use flexible models in the analysis stage.

- However, in this class, we will not cover the kind of flexible models that would help, so we will focus on balancing the predictors/covariates instead.

Strategies for mitigating model dependence

- Covariate balance (our focus)
  - Stratification
  - Matching
  - Propensity score methods
- Flexible methods (we won't cover these)
  - Semiparametric models (e.g., power series)
  - Machine learning methods (e.g., CART, random forest, boosting, bagging, etc)
  - Bayesian non-parametric and semi-parametric models (e.g., Gaussian Processes, BART, Dirichlet Processes mixtures)

Covariate balance

- Under unconfoundedness and overlap, valid causal inference can be obtained by comparing the observed distributions of Y under treatment and control **if the covariates are "balanced"**.

- Thus, a good practice is always to first check balance. That is, how similar are the two groups?

- What metric to use? The most common one is the absolute standardized difference (ASD):

$$\text{ASD}_1 = \frac{\left| \frac{\sum_{i=1}^{N} X_i W_i}{N_1} - \frac{\sum_{i=1}^{N} X_i (1 - W_i)}{N_0} \right|}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}},$$

where s_w^2 is the sample variance of the covariate in group w for $w = 0, 1$, $N_1 = \sum_{i=1}^{N} W_i$, and $N_0 = \sum_{i=1}^{N} (1 - W_i)$.

IDS 702

Covariate balance

- For a continuous covariate, ASD_1 is the standard two-sample t-statistic, and the threshold is based on a t- or z- test (e.g. 1.96).
- There is some debate on whether N_1 and N_0 should be in the denominator.
- In some disciplines, the ASD is defined as

$$\mathrm{ASD}_2 = \frac{\left| \frac{\sum_{i=1}^{N} X_i W_i}{N_1} - \frac{\sum_{i=1}^{N} X_i (1 - W_i)}{N_0} \right|}{\sqrt{s_1^2 + s_0^2}}.$$

- The common threshold is 0.1.
- Limitation of ASD: only on the difference in means (1st moments), can not capture difference in higher order moments and interactions.

IDS 702

Covariate balance

- More general, multivariate, balance metrics are available.
- R package for balance assessment: `cobalt`.
- `cobalt` generates customizable balance tables, plots (marginal distribution and Love plots) for covariates, with balance metrics.
- Besides checking marginal balance, it is always good to also check higher order terms and interactions.
- However, most times ASD is still the only balance metric checked in practice...

The minimum wage analysis

- In 1992, New Jersey decided to raise it's minimum wage from \$4.25 an hour to \$5.05 an hour.
- What was the causal effect of this decision on employment in the fast food industry?
- To study this, economists from Princeton collected data from fast food restaurants along the New Jersey - Pennsylvania border, with the Pennsylvania restaurants acting as a control group for the New Jersey restaurants.
- They also collected data on several covariates for the restaurants.
- The outcome is the employment rate after the minimum wage was raised in New Jersey.
- For more information, see the NY Times article Supersize My Wage.

IDS 702

The minimum wage analysis

- The data is in the file `MinimumWageData.csv` on Sakai.

Variables	Description
NJ.PA	indicator for which state the restaurant is in (1 if NJ, 0 if PA)
EmploymentPre	measures employment for each restaurant before the minimum wage raise in NJ
EmploymentPost	measures employment for each restaurant after the minimum wage raise in NJ
WagePre	measures the hourly wage for each restaurant before the minimum wage raise
BurgerKing	indicator for Burger King
KFC	indicator for KFC
Roys	indicator for Roys
Wendys	indicator for Wendys

The minimum wage analysis

```
MinWage <- read.csv("data/MinimumWageData.csv",header=T,
                    colClasses=c("factor","numeric","numeric","numeric",
                                 "factor","factor","factor","factor"))
str(MinWage)
```

```
## 'data.frame':    372 obs. of  8 variables:
##  $ NJ.PA         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ EmploymentPost: num  18 29.5 24 30.5 9 6.5 13.5 25 26.5 23 ...
##  $ EmploymentPre : num  30 19 67.5 18.5 6 7 12.5 55 21.5 25.5 ...
##  $ WagePre       : num  5 5.5 5 5 5.25 5 5 5 5 5.5 ...
##  $ BurgerKing    : Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 2 2 2 ...
##  $ KFC           : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 1 1 ...
##  $ Roys          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ Wendys        : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
```

```
head(MinWage)
```

```
##   NJ.PA EmploymentPost EmploymentPre WagePre BurgerKing KFC Roys Wendys
## 1     0           18.0          30.0    5.00          0   0    0      1
## 2     0           29.5          19.0    5.50          0   0    0      1
## 3     0           24.0          67.5    5.00          1   0    0      0
## 4     0           30.5          18.5    5.00          1   0    0      0
## 5     0            9.0           6.0    5.25          0   1    0      0
## 6     0            6.5           7.0    5.00          0   1    0      0
```

The minimum wage analysis

```
summary(MinWage[,c(2:4)])
```

```
##  EmploymentPost  EmploymentPre      WagePre     
##  Min.   : 0.00   Min.   : 3.00   Min.   :4.250  
##  1st Qu.:11.25   1st Qu.:11.38   1st Qu.:4.250  
##  Median :17.00   Median :16.38   Median :4.500  
##  Mean   :17.33   Mean   :17.65   Mean   :4.611  
##  3rd Qu.:22.50   3rd Qu.:21.00   3rd Qu.:4.890  
##  Max.   :55.50   Max.   :80.00   Max.   :5.750  
```

```
summary(MinWage[,-c(2:4)])
```

```
##  NJ.PA   BurgerKing KFC     Roys    Wendys 
##  0: 73   0:218      0:295   0:280   0:323  
##  1:299   1:154      1: 77   1: 92   1: 49  
```

The minimum wage analysis

Let's examine covariate balance. First, summarize covariates by NJ and PA.

```
summary(MinWage[MinWage$NJ.PA == 0, 3:8]) #first PA
```

```
##  EmploymentPre     WagePre      BurgerKing KFC    Roys   Wendys
##  Min.   : 4.5   Min.   :4.250   0:40       0:63   0:56   0:60  
##  1st Qu.:12.5   1st Qu.:4.250   1:33       1:10   1:17   1:13  
##  Median :17.0   Median :4.500                                  
##  Mean   :20.1   Mean   :4.629                                  
##  3rd Qu.:25.0   3rd Qu.:5.000                                  
##  Max.   :67.5   Max.   :5.500                                  
```

```
summary(MinWage[MinWage$NJ.PA == 1, 3:8]) #now NJ
```

```
##  EmploymentPre      WagePre      BurgerKing KFC     Roys     Wendys 
##  Min.   : 3.00   Min.   :4.250   0:178      0:232   0:224   0:263  
##  1st Qu.:11.00   1st Qu.:4.250   1:121      1: 67   1: 75   1: 36  
##  Median :15.75   Median :4.500                                     
##  Mean   :17.05   Mean   :4.606                                     
##  3rd Qu.:20.38   3rd Qu.:4.870                                     
##  Max.   :80.00   Max.   :5.750                                     
```

The minimum wage analysis

Using the `bal.tab` function in the `cobalt` package, we have

```
bal.tab(list(treat=MinWage$NJ.PA,covs=MinWage[,3:8],estimand="ATE"))
```

```
## Balance Measures
##                   Type Diff.Un
## EmploymentPre Contin. -0.2937
## WagePre       Contin. -0.0645
## BurgerKing     Binary -0.0474
## KFC            Binary  0.0871
## Roys           Binary  0.0180
## Wendys         Binary -0.0577
##
## Sample sizes
##     Control Treated
## All      73     299
```

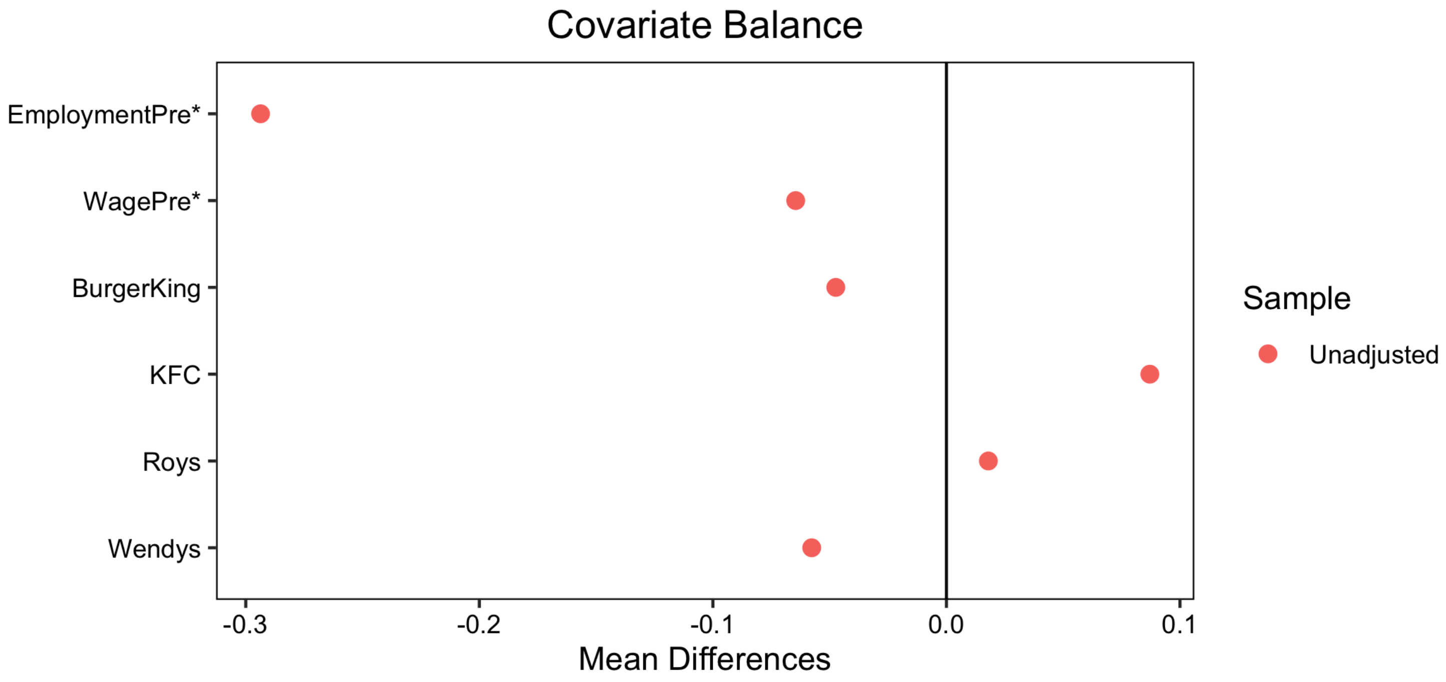
The default statistic for continuous variables is the standardized mean difference (without the absolute value). For binary variables, the default is the raw difference in proportion.

The distribution of prior employment is not well balanced across groups; other variables are pretty close, but we might be able to do better.

The minimum wage analysis

Can also use `love.plot` instead.

```
love.plot(list(treat=MinWage$NJ.PA,covs=MinWage[,3:8],estimand="ATE"),stars = "std")
```



Same conclusion. How can we improve the balance?

IDS 702

Acknowledgements

These slides contain materials adapted from courses taught by Dr. Fan Li.

What's next?

Move on to the readings for the next module!