Quant II Recitation

Ye Wang yw1576@nyu.edu

Feb 7, 2018

Today's Plan

- Causal inference from a machine learning perspective
- Regression
- Simulation (Regression in R)

Today's Plan

- Causal inference from a machine learning perspective
- Regression
- Simulation (Regression in R)

Causal inference from a machine learning perpective

- Now we have been familiar with the Rubin model

$$Y_i = \begin{cases} Y_i(1) \text{ if } D_i = 1 \\ Y_i(0) \text{ if } D_i = 0 \end{cases}$$

Causal inference from a machine learning perpective

- Now we have been familiar with the Rubin model

$$Y_i = \begin{cases} Y_i(1) \text{ if } D_i = 1 \\ Y_i(0) \text{ if } D_i = 0 \end{cases}$$

- For each i, we observe either $Y_i(0)$ or $Y_i(1)$ ("Fundamental problem of causal inference")

Causal inference from a machine learning perpective

- Now we have been familiar with the Rubin model

$$Y_i = \begin{cases} Y_i(1) \text{ if } D_i = 1 \\ Y_i(0) \text{ if } D_i = 0 \end{cases}$$

- For each i, we observe either $Y_i(0)$ or $Y_i(1)$ ("Fundamental problem of causal inference")
- Suppose we are interested in ATT, then we just need to know $Y_i(0)$ for each treated unit

Causal inference from a machine learning perpective

- Now we have been familiar with the Rubin model

$$Y_i = \begin{cases} Y_i(1) \text{ if } D_i = 1 \\ Y_i(0) \text{ if } D_i = 0 \end{cases}$$

- For each i, we observe either $Y_i(0)$ or $Y_i(1)$ ("Fundamental problem of causal inference")
- Suppose we are interested in ATT, then we just need to know $Y_i(0)$ for each treated unit
- It is a prediction problem: $\hat{Y}_i(0) = f(\mathbf{X}, \mathbf{Y}_{(-\mathbf{i})})$

Causal inference from a machine learning perpective

- That's where machine learning enters!

Causal inference from a machine learning perpective

- That's where machine learning enters!
- The target of machine learning algorithms is to find a prediction function $\hat{f}$ that minimizes the expected squared prediction error (ESPE), $E[(f - \hat{f})^2]$

Causal inference from a machine learning perpective

- That's where machine learning enters!
- The target of machine learning algorithms is to find a prediction function $\hat{f}$ that minimizes the expected squared prediction error (ESPE), $E[(f - \hat{f})^2]$
- It is easy to see that

$$\begin{aligned} E[(f - \hat{f})^2] &= E[f^2 - 2 * f * \hat{f} + \hat{f}^2] \\ &= f^2 - 2 * f * E[\hat{f}] + E[\hat{f}^2] \\ &= f^2 - 2 * f * E[\hat{f}] + E[\hat{f}]^2 - E[\hat{f}]^2 + E[\hat{f}^2] \\ &= (E[\hat{f}] - f)^2 + E[\hat{f}^2] - E[\hat{f}]^2 \\ &= (Bias(\hat{f}))^2 + Var(\hat{f}) \end{aligned}$$

Causal inference from a machine learning perpective

- That's where machine learning enters!
- The target of machine learning algorithms is to find a prediction function $\hat{f}$ that minimizes the expected squared prediction error (ESPE), $E[(f-\hat{f})^2]$
- It is easy to see that

$$
\begin{aligned}
E[(f-\hat{f})^2] &= E[f^2 - 2*f*\hat{f} + \hat{f}^2] \\
&= f^2 - 2*f*E[\hat{f}] + E[\hat{f}^2] \\
&= f^2 - 2*f*E[\hat{f}] + E[\hat{f}]^2 - E[\hat{f}]^2 + E[\hat{f}^2] \\
&= (E[\hat{f}] - f)^2 + E[\hat{f}^2] - E[\hat{f}]^2 \\
&= (Bias(\hat{f}))^2 + Var(\hat{f})
\end{aligned}
$$

- This is called bias-variance trade-off
- A method with smaller bias usually has larger variance

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment
- If $\hat{f} = \bar{Y}_{D_i=0,\mathbf{X}=\mathbf{x}}$, what do we have?

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment
- If $\hat{f} = \bar{Y}_{D_i=0,\mathbf{X}=\mathbf{x}}$, what do we have?
  Blocking experiment or matching

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment
- If $\hat{f} = \bar{Y}_{D_i=0,\mathbf{X}=\mathbf{x}}$, what do we have?
  Blocking experiment or matching
- Now, what is the assumption behind regression?

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment
- If $\hat{f} = \bar{Y}_{D_i=0,\mathbf{X}=\mathbf{x}}$, what do we have?
  Blocking experiment or matching
- Now, what is the assumption behind regression?
  $\hat{f} = \mathbf{X}_{D_i=0}\beta$ (Linearity)
  $\gamma_i = \gamma$ for any i (Constant treatment effect)

Causal inference from a machine learning perpective

- In causal inference, we train a model based on the control group observations, then use the model to predict counterfactuals
- If $\hat{f} = \bar{Y}_{D_i=0}$, what do we have?
  Random experiment
- If $\hat{f} = \bar{Y}_{D_i=0,\mathbf{X}=\mathbf{x}}$, what do we have?
  Blocking experiment or matching
- Now, what is the assumption behind regression?
  $\hat{f} = \mathbf{X}_{D_i=0}\beta$ (Linearity)
  $\gamma_i = \gamma$ for any i (Constant treatment effect)
- Matching: low bias and high variance; regression: high bias and low variance

Causal inference from a machine learning perpective

- It is straightfoward to drop the constant treatment effect assumption
  $\hat{\gamma}_i = Y_i - \mathbf{X}_{D_i=0}\hat{\beta}$ (Regression with interaction)
- Replacing $\mathbf{X}_{D_i=0}\beta$ with $(\mathbf{X}_{D_i=0} - \bar{\mathbf{X}}_{D_i=0})\beta$, we get the more efficient Lin's regression

Causal inference from a machine learning perpective

- It is straightfoward to drop the constant treatment effect assumption
  $\hat{\gamma}_i = Y_i - \mathbf{X}_{D_i=0}\hat{\beta}$ (Regression with interaction)
- Replacing $\mathbf{X}_{D_i=0}\beta$ with $(\mathbf{X}_{D_i=0} - \bar{\mathbf{X}}_{D_i=0})\beta$, we get the more efficient Lin's regression
- Question: How to get rid of the linearity assumption?

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision
- Design-based perspective vs. Model-based perspective

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision
- Design-based perspective vs. Model-based perspective
- With no extra assumption: Agnostic, or non-parametric estimation

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision
- Design-based perspective vs. Model-based perspective
- With no extra assumption: Agnostic, or non-parametric estimation

  Group-mean difference, Matching
- When a complete model is specified: Parametric estimation

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision
- Design-based perspective vs. Model-based perspective
- With no extra assumption: Agnostic, or non-parametric estimation
  Group-mean difference, Matching
- When a complete model is specified: Parametric estimation
  Regression, Probit, Logit, All Bayesian approaches, etc.
- With some "structure" assumed for $\hat{f}$: Semi-parametric estimation

Some Jargons

- Causal inference is built upon the assumption of (strong) ignorability
- But assumptions imposed on $\hat{f}$ are up to the researcher's decision
- Design-based perspective vs. Model-based perspective
- With no extra assumption: Agnostic, or non-parametric estimation

  Group-mean difference, Matching
- When a complete model is specified: Parametric estimation
  Regression, Probit, Logit, All Bayesian approaches, etc.
- With some "structure" assumed for $\hat{f}$: Semi-parametric estimation

  Kernelized or serial estimation, factor models
- Two types of pre-treatment attributes: confounders and covariates

Problems with naive regression

- It is biased and inconsistent under treatment effect heterogeneity

Problems with naive regression

- It is biased and inconsistent under treatment effect heterogeneity
- What is its expectation then?
  Abadie et al. (2017): a weighted sum of the true individualistic effects under linearity, and a weighted sum of something without linearity

Problems with naive regression

- It is biased and inconsistent under treatment effect heterogeneity
- What is its expectation then?
  Abadie et al. (2017): a weighted sum of the true individualistic effects under linearity, and a weighted sum of something without linearity
- Should we add as many covariates as possible?
  No. Covariates may sometimes amplify the existing bias (Middleton et al., 2016)

1. X may absorb the variation of D and reduces its explanatory power of Y
2. If X is negatively correlated with Y and the unobservables are positively correlated with Y, leaving X outside the regression may offset the impact of the unobservables

Covariate Adjustment in sampling

- Imagine that we are biologists who are interested in leaf size.
- Finding the size of leaves is hard, but weighting leaves is easy.
- We can use auxilliary information to be smarter:
  - Sample from leaves on a tree.
  - Measure their size and weight.
  - Let $\bar{y}_s$ be the average size in the sample.
  - Let $\bar{x}_s$ be the average weight in the sample.
  - We know that $\bar{y}_s$ unbiased and consistent for $\bar{y}$
  - But we have extra information!
  - We also have $\bar{x}$ (all the weights)
  - This motivates the regression estimator:
    $\hat{\bar{y}} = \bar{y}_s + \beta(\bar{x} - \bar{x}_s)$
  - We get β by a regression of leaf area on weight in the sample.

A Social Science Example

- We are interested in the effect of a binary treatment on test scores.
- Let's set up a simulation.
- 200 students. Observed over two years.
- Half good tutors and half bad since the second year.
- We want to estimate the effect of the intervention in year 2.
- Treatment is assigned randomly
- Test score in the first year will be a covariate

Simulation

```
cat("Real ATE =", RealATE, "\n")
```

```
## Real ATE = 10.35974
```

```
round(summary(lm(Yr2Obs~Trt))$coefficients[2,], 4)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     8.9718     1.1365     7.8944     0.0000
```

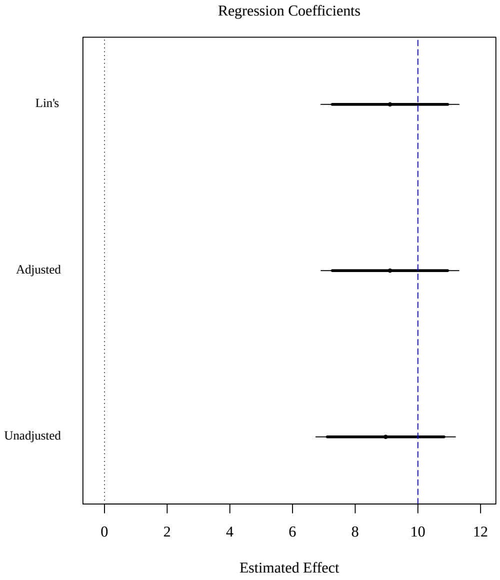
```
round(summary(lm(Yr2Obs~Trt+Yr1Score))$coefficients[2,], 4
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     9.1094     1.1231     8.1106     0.0000
```

```
round(summary(lm(Yr2Obs~Trt*demeaned_Yr1Score))$coefficient
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     9.1093     1.1258     8.0916     0.0000
```

Coefficient Plot Code



Regression Table

```
## Loading required package: stargazer

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regress

##  R package version 5.2.2. https://CRAN.R-project.org/pac

##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Ha
## % Date and time: Thu, Feb 07, 2019 - 13:56:02
## \begin{table}[!htbp] \centering
##   \caption{Regression Results}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
```

Regression Table

Table 1: Regression Results

	Dependent variable:		
	Yr2Obs		
	(1)	(2)	
Treatment	8.452^{***}	8.240^{***}	
	(1.247)	(1.236)	
Yr1 Score		0.271^{*}	
		(0.114)	
Yr1 Score (demeaned)			
Tr. * Yr1 Score			
Observations	200	200	
R^2	0.188	0.211	
Adjusted R^2	0.184	0.203	
Residual Std. Error	8.819 (df = 198)	8.717 (df = 197)	
F Statistic	45.928^{***} (df = 1; 198)	26.322^{***} (df = 2; 197)	18.

Unbiasedness

```
cat("Real ATE =", RealATE, "\n")
```

```
## Real ATE = 10.35974
```

```
mean(coefs[, 1]) - RealATE
```

```
## [1] -5.094933e-06
```

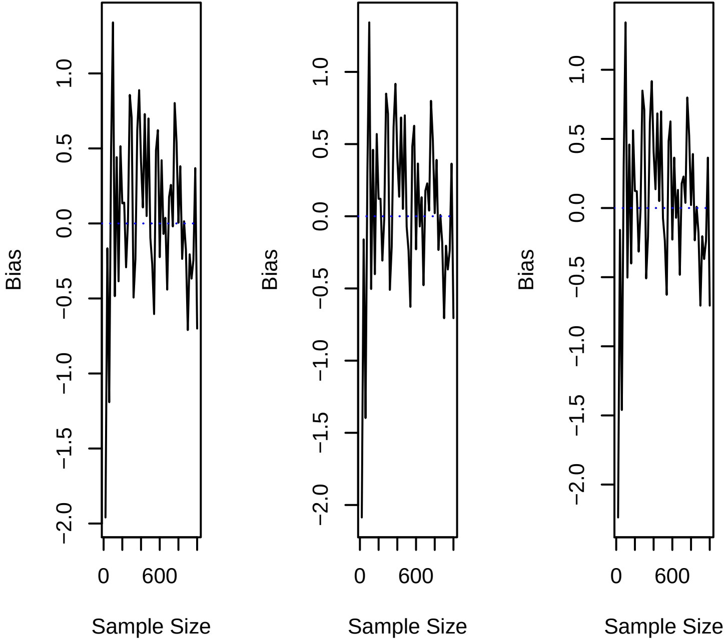
```
mean(coefs[, 2]) - RealATE
```

```
## [1] 0.005582937
```

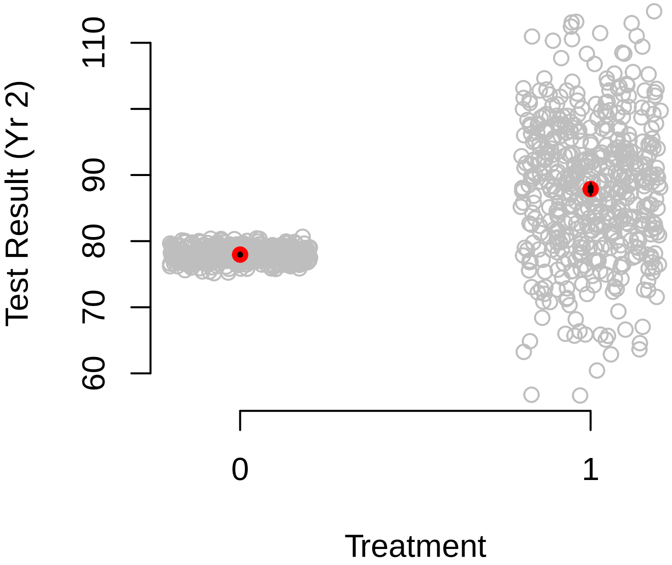
```
mean(coefs[, 3]) - RealATE
```

```
## [1] 0.005548307
```

Consistency



Plot Data



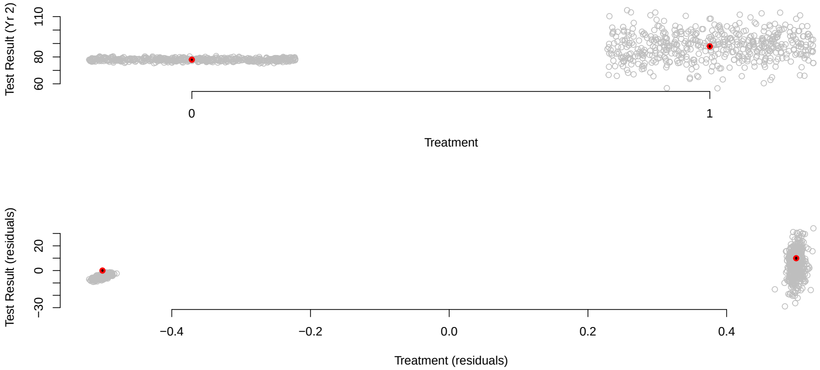
Partial Regression and Residualized Plot

- Can we make that plot a little more friendly?
- Let's residualize our outcome based on scores in the first period. This should remove a substantial amount of the variance in the outcome.

Partial Regression and Residualized Plot

- Can we make that plot a little more friendly?
- Let's residualize our outcome based on scores in the first period. This should remove a substantial amount of the variance in the outcome.

Partial Regression and Residualized Plot



Partial Regression for FEs

- We'll get to this later in the semester.
- The point is, partial regression is a fundamentally important tool that let's us do things that would otherwise be very hard.

Partial Regression for FEs

- We'll get to this later in the semester.
- The point is, partial regression is a fundamentally important tool that let's us do things that would otherwise be very hard.

Partial Regression for FEs

- We'll get to this later in the semester.
- The point is, partial regression is a fundamentally important tool that let's us do things that would otherwise be very hard.

- When the panel is unbalanced, this is not correct...

Testing linear Restrictions

- Hypothesis: $R\beta = r$
- $W = (R\hat{\beta} - r)'(R\hat{\boldsymbol{V}}R')^{-1}(R\hat{\beta} - r) \sim \chi^2_q$
- Or more conservatively: $W/q \sim F_{q,N-K}$
- In R:

Testing linear Restrictions

- Hypothesis: $R\beta = r$
- $W = (R\hat{\beta} - r)'(R\hat{\boldsymbol{V}}R')^{-1}(R\hat{\beta} - r) \sim \chi^2_q$
- Or more conservatively: $W/q \sim F_{q,N-K}$
- In R:

- Think about how these two might differ for different starting parameters (ex. sample size)