Achieving Reliable Causal Inference with Data-Mined Variables

Edward McFowland III
March 6, 2023
Texas A&M Institute of Data Science Seminar
Agenda

• Motivation: Value of ML models for Variable Creation
• Challenges in Estimation: Measurement Error
• ForestIV
  • Conceptually
  • Procedurally
• Simulation Results
• Conclusions
Predictive machine learning facilitates more efficient and cost-effective information extraction from (often unstructured) data. E.g.,

- Predict sentiment from large amount of texts
- Identify gender/age from profile pictures
- Measure demographics using Google Street View images (previously ~$250 million done by survey)
- Predict poverty levels using satellite images (previously barely available for some developing countries)

Gebru T, et al. (2017) Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. Proceedings of the National Academy of Sciences; Jean N, et al. (2016) Combining satellite imagery and machine learning to predict poverty. Science 353(6301):790–794.
Combining ML with Econometrics

Opportunity to study empirical questions with ML-generated measures

- **Stage 1**: Use predictive ML to create new variables of interest
- **Stage 2**: Add these “mined” variables into econometric frameworks, typically as independent variables, to make inference

A increasingly popular strategy across the Social and Statistical Sciences

- Goh et al. (2013) [*Information Systems*]
  - **Mined Variable**: Textual sentiment
  - **Inference**: Effect of user- and marketer-generated content on consumer spending

- Cefalu and Dominici (2014) [*Epidemiology and Biostatistics*]
  - **Mined Variable**: Pollution exposure (given fixed sensors at monitoring locations)
  - **Inference**: Effect of air pollution exposure on health outcomes

- Dube and Lindner (2022) [*Labor Economics*]
  - **Mined Variable**: Exposure to minimum wage changes
  - **Inference**: Effect of the minimum wage on labor market outcomes
Combing ML with Econometrics

Tepper School of Business

Digital Marketing and Social Media Strategy

Course Number: 45882

This course posits to explore issues related to digital and social media marketing. This is a hands-on class where you will use real world data. We will cover the following topics in class.

**Search Engine Optimization:** how search engines, keyword auctions, and search engine marketing work, and how to optimize your pay per click advertisement efforts.

**Social Media Marketing:** How to design a social media campaign? How is social media marketing different from traditional marketing? What are the key ingredients that make social media campaigns successful? How can you make your product and your campaign viral? How to design a real Viral campaign?

**Forecasting using Online Search Trends Data:** How and when to build better forecasting models for demand using Google search data (Google Trends and Insights).

**Econo-mining User Generated Content:** How firms are and can use get useful information from user generated content using text mining and opinion mining capabilities to drive their product development, placement and advertisement decisions. Using real world data you will analyze whether the traditional approaches for driving product development strategy are in alignment with what you learn from user generated content.
Agenda

• Motivation: Value of ML models for Variable Creation
• Challenges in Estimation: Measurement Error
• ForestIV
  • Conceptually
  • Procedurally
• Simulation Results
• Conclusions
Common Practice and Pitfall: An Example

\[ Y = f(\beta \text{Sentiment} + \Pi \text{Controls}) \]

Data (e.g., UGC) → Predictive model → Predicted Sentiment

\[ \hat{\beta}, \hat{\Pi} \]

Estimation & Inference

Measurement Error

Imperfect

Biased
Common Practice and Pitfall: An Example

Data (e.g., UGC) → Predictive model → Predicted Sentiment

\[ Y = f(\beta X + \Pi Z) \]

Econometric model
Measurement Error

\[ \hat{\beta}, \hat{\Pi} \]

Biased
Estimation & Inference

Imperfect
The Prediction/Measurement Error

Linear Regression Example:

\[ Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon \]

- We do not observe \( X \), only observe ML-generated predictions \( \hat{X} \) (continuous)
- \( \hat{X} \) contains prediction/measurement error, e.g., \( \hat{X} = X + e \) (\( e \) is random)
- The regression actually being estimated:

\[ Y = \beta_0 + \beta_1 \hat{X} + \beta_2 Z + (\varepsilon - e \beta_1) \]

Endogeneity!

\[ \hat{\beta}_1^{OLS} - \beta_1 \propto \text{Cov}(\hat{X}, (\varepsilon - e \beta_1)) \]

\[ = \text{Cov}(X + e, (\varepsilon - e \beta_1)) \]

\[ \neq 0 \]

- In general, the bias can be over-/under-estimation
Agenda

- Motivation: Value of ML models for Variable Creation
- Challenges in Estimation: Measurement Error
- ForestIV
  - Conceptually
  - Procedurally
- Simulation Results
- Conclusions
Instrumental Variables

Estimable Linear Regression: \( Y = \beta_0 + \beta_1 \hat{X} + \beta_2 Z + (\varepsilon - e\beta_1) \)

If we have an instrumental variable, \( W \)

- We can mitigate the bias via IV regression
- \( Cov(W, \hat{X}) \neq 0 \) and \( Cov(W, e) = 0 \).
  - \( Cov(W, X) \neq 0 \)

Problem: good IVs are hard to find

- Our claim: build a Random Forest to predict \( X \), it will also produce potential IVs to correct for estimation biases
How does Random Forest Work?

Training Data
(Rows: instances
Columns: features)

Sample

→

Build Decision Trees

→ Predictions

→ Predictions

→ Predictions

Aggregate Predictions
Why does Random Forest Work?
• Each tree is somewhat accurate (strength)
• Different trees make somewhat different mistakes (diversity)
• Greater strength and diversity lead to more accurate forest (Breiman 2001)
• Suppose \( \{p_i, p_j\} \) are predictions from two trees in a forest, \( \{e_i, e_j\} \) are errors, then on average:
  • \( \text{Cov}(p_i, p_j) \) is relatively strong
  • \( \text{Cov}(e_i, e_j) \) is relatively weak

What’s a Good Instrumental Variable?
• \( Y = \beta_0 + \beta_1 \hat{X} + \beta_2 Z + (\epsilon - e\beta_1) \)
• A good IV, \( W \), needs to satisfy two conditions:
  • \( \text{Cov}(W, \hat{X}) \neq 0 \) (relevance)
  • \( \text{Cov}(W, \epsilon) \approx 0 \) (exclusion)
• Key idea: use \( p_i \) as endogenous regressor and \( p_j \) as candidate IVs
  • I.e., \( p_i \equiv \hat{X}, e_i \equiv e, p_j \equiv W \)
  • \( \text{Cov}(p_j, p_i) \neq 0 \) is satisfied
  • \( \text{Cov}(p_j, e_i) = \text{Cov}(e_j + a, e_i) \) is likely small
A random forest of 100 trees:

- $\text{Corr}(p_j, p_i)$ is relatively strong
- $\text{Corr}(p_j, e_i)$ is relatively weak on average, but occasionally strong

Good IVs are available, but need to be picked out

$\text{Corr}(p_j, p_i)$

$\text{Corr}(p_j, e_i)$
Our goal: select valid and strong IVs from a pool of candidates

- Belloni et al. (2012) proposed to use Lasso regression for IV selection
- Suppose a pool of valid IVs \( W_1, \ldots, W_K \) for a endogenous variable \( \hat{X} \).
- Run Lasso \( \hat{X} \sim \{W_1, \ldots, W_K\} \) and keep the ones with non-zero coefficients
  - Run 2SLS with only the selected ones
- We re-purpose this approach for our goal
Agenda

- Motivation: Value of ML models for Variable Creation
- Challenges in Estimation: Measurement Error
- ForestIV
  - Conceptually
  - Procedurally
- Simulation Results
- Conclusions
Suppose we build a random forest of 100 trees, their predictions denoted as \( \{p_1, \ldots, p_{100}\} \). We use \( p_1 \) as the endogenous regressor and \( \{p_2, \ldots, p_{100}\} \) as candidate IVs:

1. On Testing data, estimate Lasso \( e_1 \sim \{p_2, \ldots, p_{100}\} \), drop the candidates with non-zero coefficients, they are likely invalid IVs. Denote the remaining candidates as \( V_1 \).

2. On Testing+Unlabeled data, estimate Lasso \( p_1 \sim V_1 \), drop the candidates with zero coefficients, they are likely weak IVs. Denote the remaining candidates as \( S_1 \).

3. Repeat steps 1 and 2 until the selection does not change anymore. Return the selection as the IVs for \( p_1 \).

4. Repeat the selection procedure for each tree used as the endogenous regressor. Ultimately, we will have 100 pairs of \( \{p_i, S_i\} \).
ForestIV: IV Estimation Procedure

| Training | Testing | Unlabeled |
|----------|---------|-----------|
| Estimate econometric model without bias (small data/power) | | Estimate econometric model with IVs |

The unbiased coefficients, \( \beta_{unbias} \), can serve as a benchmark

1. On Unlabeled data, use \( \{p_i, S_i\} \) to obtain IV estimations \( \beta_{IV}^i \)
2. Use Hotelling \( T^2 \) test to check if \( \beta_{IV}^i \) is significantly different from \( \beta_{unbias} \). If not, keep it
3. Among “surviving” \( \beta_{IV}^i \), pick the one that has the smallest empirical risk, measured as

\[
(\beta_{IV}^i - \beta_{unbias})^2 + Variance(\beta_{IV}^i) = bias^2 + variance
\]

We call our IV selection and estimation approach ForestIV

Mochen Yang, Edward McFowland, III, Gordon Burtch, Gediminas Adomavicius (2022) Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest Approach to the Measurement Error Problem. INFORMS Journal on Data Science 1(2):138-155.
Motivation: Value of ML models for Variable Creation

Challenges in Estimation: Measurement Error

ForestIV
  • Conceptually
  • Procedurally

Simulation Results

Conclusions
Simulation Experiments

Setup

- “Bike Sharing” Dataset (UCI repository): 17,379 instances of bike rental activities and weather/seasonal information
- 1,000 training obs, 200 testing obs, 16,179 unlabeled obs
- Build a random forest of 100 trees to predict number of hourly rentals (lnCnt)
- Simulate an econometric model:
  \[ Y = 1.0 + 0.5lnCnt + 2.0Z_1 + 1.0Z_2 + \varepsilon \]
- lnCnt is the ML-predicted variable, \( Z_1 \sim Unif, Z_2 \sim Binary, \varepsilon \sim N(0,2^2) \)
- Regression coefficients are fixed in order to measure the degree of biases and efficacy of correction
- Simulation repeated for 100 times, we report averaged estimation results
Simulation Results

| True | Biased       | Unbiased    | ForestIV    |
|------|--------------|-------------|-------------|
| $\beta_0 = 1.0$ | 0.702 (0.063) | 1.018 (0.204) | 0.957 (0.059) |
| $\beta_{ML} = 0.5$ | 0.566 (0.013) | 0.498 (0.043) | 0.512 (0.012) |
| $\beta_{z_1} = 2.0$ | 2.000 (0.003) | 1.999 (0.011) | 2.000 (0.003) |
| $\beta_{z_2} = 1.0$ | 1.000 (0.002) | 0.999 (0.006) | 1.000 (0.002) |

Observations

- Directly using ML predictions in regressions produces biased estimates
- Our approach is effective in mitigating biases
- ForestIV estimates are not significantly different from “True” coefficients in 87% of simulation rounds (F-test, $\alpha = 0.001$)
- ForestIV estimates have $3x-4x$ smaller standard errors than unbiased estimates
Performance Sensitivity: Size of Unlabeled Data

Observations

- Bias does not go away with larger samples
- IV estimates have converged with sufficient unlabeled data
- ForestIV is advantageous when unlabeled data is much larger than labeled data

95% CI on the ML-generated variable for different sizes of unlabeled data, averaged over 100 simulation runs

Coefficient estimate on the ML-generated variable for different sizes of unlabeled data, from a single simulation run
Performance Sensitivity: Number of Trees

|             | True | 25 Trees | 50 Trees | 100 Trees | 150 Trees | 200 Trees |
|-------------|------|----------|----------|-----------|-----------|-----------|
| $\beta_0 = 1.0$ | 0.884 (0.060) | 0.919 (0.060) | 0.957 (0.058) | 0.941 (0.059) | 0.889 (0.061) |
| $\beta_{ML} = 0.5$ | 0.526 (0.013) | 0.519 (0.013) | 0.512 (0.012) | 0.514 (0.013) | 0.524 (0.013) |
| $\beta_{z_1} = 2.0$ | 2.000 (0.003) | 2.000 (0.003) | 2.000 (0.003) | 2.000 (0.003) | 2.000 (0.003) |
| $\beta_{z_2} = 1.0$ | 1.000 (0.002) | 1.000 (0.002) | 1.000 (0.002) | 1.000 (0.002) | 1.000 (0.002) |
| Hotelling $T^2$ | 3.0817 | 3.0501 | 2.7094 | 2.7816 | 2.9357 |

Observations

- ForestIV requires a reasonably large forest to work well
- However, having too many trees can hurt correction performance
1. Use data-mined variables directly in econometric models results in biased estimations and misleading inferences

2. Random Forest, as a data mining technique, can produce individual trees with correlated predictions but weakly correlated errors

3. Such property can be leveraged to discover useful instrumental variables

4. We design data-driven procedures to select proper IVs. Our approach can effectively mitigate estimation biases.

Key Takeaways
THANK YOU