WEEK 2

Causality, Theory and Data

The Fundamental Question in Economics and ML

ECO447 — Machine Learning for Economists

Assoc. Prof. Dr. Onur Polat
Hacettepe University, Institute of Informatics

Today's Outline

01

Correlation vs. Causation

KEY CONCEPT

The Fundamental Problem

Correlation ≠ Causation. Spurious correlations arise from confounders, reverse causality, and selection bias.

Economics requires causal answers: Does a policy work? What is the effect of education on earnings? ML traditionally focuses on prediction, but modern econometrics leverages ML for causal inference.

Sources of Spurious Correlation

- Confounding variables: Z causes both X and Y (e.g., ability → education & earnings)
- Reverse causality: Y causes X instead of X → Y (e.g., police ↔ crime)
- Selection bias: non-random sample (e.g., survivors, volunteers)
- Measurement error: systematic mismeasurement of X or Y
- Omitted variable bias: relevant variable excluded from the model
- Simpson's paradox: relationship reverses when controlling for a variable
- Classic example: ice cream sales and drowning deaths (both driven by temperature)

Famous Spurious Correlations in Economics

- Per capita cheese consumption correlates with death by bedsheet tangling
- Super Bowl indicator: NFC wins predict stock market returns
- Hemline index: skirt lengths predict economic conditions
- Lesson: with enough data, you can find correlation between ANYTHING
- Data mining (p-hacking) amplifies this in research
- ML models can memorize spurious patterns if not carefully designed
- Economic theory provides guardrails against absurd predictions

02

The Potential Outcomes Framework

Rubin Causal Model (Potential Outcomes)

$Y_i(1)$ = potential outcome if unit i receives treatment

What WOULD have happened to treated unit if not treated

$Y_i(0)$ = potential outcome if unit i receives control

What WOULD have happened to control unit if treated

$\tau_i = Y_i(1) - Y_i(0)$ (individual treatment effect)

Fundamental problem: we never observe both for the same unit

$ATE = E[Y_i(1) - Y_i(0)]$ (average treatment effect)

KEY CONCEPT

The Fundamental Problem of Causal Inference

We can never observe both $Y_i(1)$ and $Y_i(0)$ for the same individual at the same time. The counterfactual is inherently unobservable.

Solution approaches: randomization (makes groups comparable on average), matching (find similar units), instrumental variables (exploit exogenous variation), regression discontinuity, difference-in-differences.

03

Experiments and Identification

Randomized Controlled Trials (RCTs)

- Gold standard for causal inference: randomly assign treatment
- Randomization ensures treatment is independent of potential outcomes
- $E[Y_i(0)|D=1] = E[Y_i(0)|D=0]$ — groups are comparable in expectation
- Simple difference in means estimates ATE without bias
- Examples: job training programs (LaLonde), microfinance, UBI experiments
- Limitations: ethical constraints, compliance issues, external validity
- Increasingly common in development economics (Duflo, Banerjee — Nobel 2019)

Natural Experiments & Quasi-Experiments

- Instrumental Variables (IV): find exogenous variation in X
- Regression Discontinuity (RD): sharp cutoff in treatment assignment
- Difference-in-Differences (DiD): compare treatment/control before/after
- Synthetic Control: construct counterfactual from weighted donors
- Event Studies: examine outcomes around a specific event
- Examples: draft lottery & earnings, minimum wage (Card & Krueger), EU enlargement
- Key insight: identification strategy must be credible for causal claims

04

Directed Acyclic Graphs (DAGs)

Causal Graphs: DAG Fundamentals

- DAG: visual representation of causal assumptions
- Nodes = variables, directed edges = causal relationships
- Three basic structures: chain ($X \to Z \to Y$), fork ($X \leftarrow Z \to Y$), collider ($X \to Z \leftarrow Y$)
- Fork (confounder): creates spurious association — must control for Z
- Collider: no association between X and Y — controlling for Z creates bias!
- d-separation: formal criterion for conditional independence
- DAGs make causal assumptions explicit and testable

Using DAGs for Identification

- Step 1: Draw the DAG based on domain knowledge
- Step 2: Identify all paths between treatment and outcome
- Step 3: Block backdoor paths (confounding) by conditioning
- Step 4: Keep front-door paths open (causal channels)
- Step 5: Check for collider bias — don't condition on descendants of both T and Y
- Backdoor criterion: set of variables that blocks all backdoor paths
- Tools: dagitty.net (web), DoWhy (Python), ggdag (R)

05

ML Meets Causality

Where ML Helps Causal Inference

- High-dimensional confounders: too many controls for traditional methods
- Double/Debiased ML (Chernozhukov et al., 2018): ML for nuisance, inference for causal
- Causal Forests (Athey & Imbens, 2019): heterogeneous treatment effects
- LASSO for IV: selecting instruments from many candidates
- Propensity score estimation: ML can model treatment assignment better
- Synthetic control with ML: data-driven donor weights
- Transfer learning for policy evaluation across contexts

Double/Debiased Machine Learning

$$Y = \theta_0 D + g_0(X) + U, \quad E[U|D,X] = 0$$

Partially linear model: θ_0 is the causal parameter of interest

$$D = m_0(X) + V, \quad E[V|X] = 0$$

Treatment equation: model selection into treatment

$$\text{Step 1: } \hat{Y} = \hat{g}(X) \text{ using ML}, \quad \text{residual } \tilde{Y} = Y - \hat{Y}$$

Partialling out the effect of controls on outcome

$$\text{Step 2: } \hat{D} = \hat{m}(X) \text{ using ML}, \quad \text{residual } \tilde{D} = D - \hat{D}$$

PYTHON

Double ML in Python with DoubleML

```python
from doubleml import DoubleMLPLR, DoubleMLData
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LassoCV

# Prepare data
dml_data = DoubleMLData(df, y_col="wage",
                        d_cols="training",
                        x_cols=controls)

# Set up learners for nuisance parameters
ml_l = RandomForestRegressor(n_estimators=200)  # E[Y|X]
ml_m = RandomForestRegressor(n_estimators=200)  # E[D|X]

# Estimate causal effect
dml_plr = DoubleMLPLR(dml_data, ml_l, ml_m,
                      n_folds=5, n_rep=10)
dml_plr.fit()
print(dml_plr.summary)
```

- DoubleML package implements Chernozhukov et al. (2018)
- Cross-fitting (n_folds) prevents overfitting bias
- Any scikit-learn compatible learner can be used

06

Applications in Economics

Case Study: Returns to Education

- Naive OLS: regress earnings on years of schooling — biased (ability confounds)
- IV approach: use compulsory schooling laws, distance to college as instruments
- ML enhancement: use LASSO to select relevant instruments from many candidates
- Double ML: partial out confounders with ML, estimate causal effect of education
- Heterogeneity: causal forests reveal who benefits most from additional education
- Policy implication: targeted education subsidies for high-return subpopulations
- Key reference: Angrist & Krueger (1991), Belloni et al. (2014)

Case Study: Policy Evaluation

- Question: Does a job training program increase employment and earnings?
- RCT data: LaLonde (1986) — experimental benchmark
- Observational challenge: non-random participation creates selection bias
- ML approach: propensity score estimation with gradient boosting
- CATE estimation: who benefits most from the program?
- Budget allocation: target program to highest-benefit individuals
- Practical framework: predict → identify → allocate

SUMMARY

Week 2 Key Takeaways

- Causation requires more than correlation — identification strategy is key
- Potential outcomes framework: counterfactuals define causal effects
- DAGs make causal assumptions explicit and guide variable selection
- RCTs are the gold standard but not always feasible
- ML enhances causal inference: Double ML, causal forests, IV selection
- Economists bring causal thinking; ML brings prediction power — synergy
- Next week: Data Analysis, Data Types, and Exploration