UCLAAnderson
SCHOOL of MANAGEMENT

PMDA

Lecture 10: Regression Trees & Recap

Agenda

- Final Exam logistics
- Regression Trees
- Recap questions
- Class evaluation
- Workshop: Recap coding/Final exam like questions

Final Exam Logistics

- Logistics:
  - 40% of final grade
  - **Thursday, March 21st, 8.30-11:30am**
  - **Section 1: Room C301**
  - **Section 2: Room B313**
  - Open book, open notes, open everything
  - No Internet
  - I'll stop by the class every hour or 45min, the TAs will be proctoring
- Structure:
  - Similar to final practice problem set (available on BruinLearn by the end of the day today)
  - No coding beyond "cheat sheet" list of commands
  - Not designed for "time pressure"

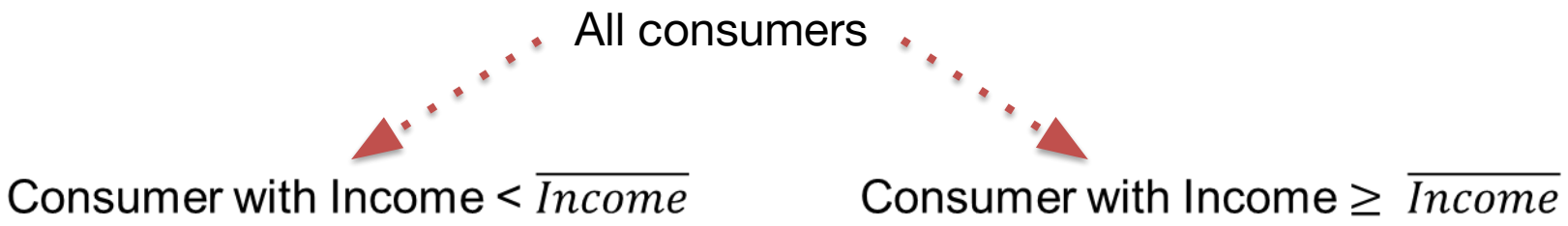
Agenda

- Final Exam logistics
- Regression Trees
- Recap questions
- Class evaluation
- Workshop: Recap coding/Final exam like questions

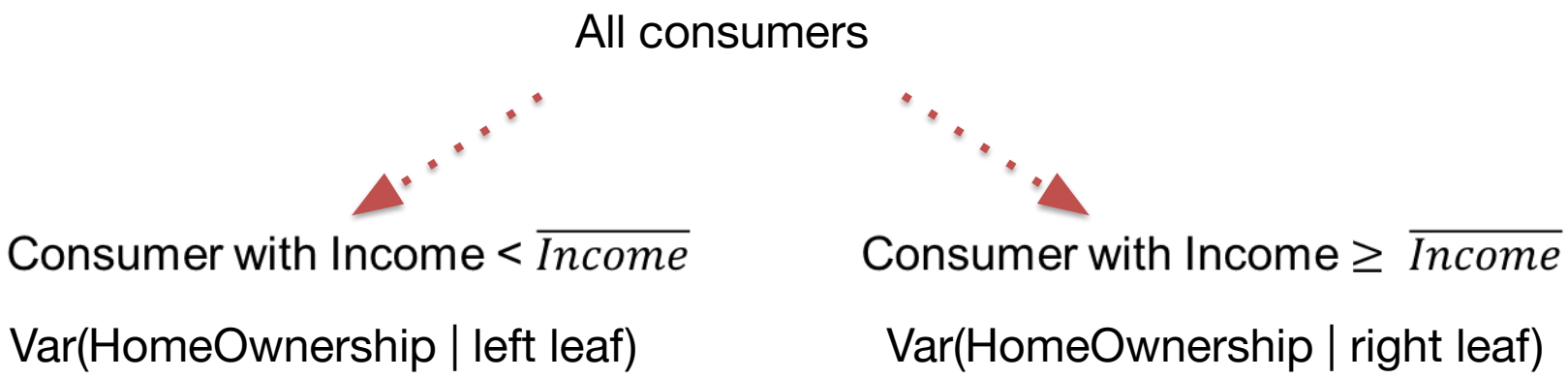
Regression Trees

- Non-parametric method for prediction (Lasso is parametric)
- In plain English: there are no regression coefficients, i.e. parameters, here

- Simple example:
  - Say we want to predict whether an individual owns a house based on income
  - A regression tree attempts to split the sample based on income such that outcomes within each "leaf" are as similar as possible



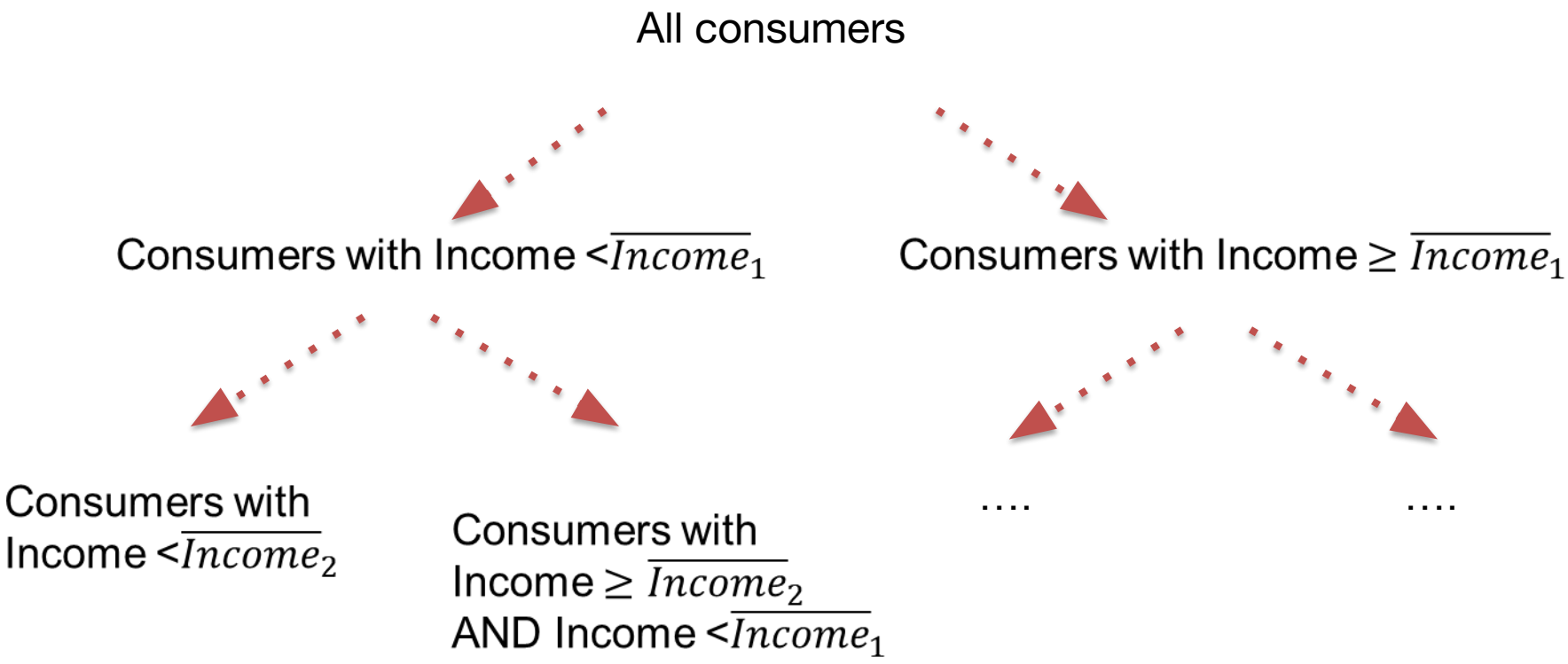
Regression Trees



- Regression tries to minimize sum of variances in both leafs
- This leads to leafs being predictive
- Take extreme case
  - We find cut-off such that everybody in the right leaf owns a house, nobody in the left leaf does
  - Therefore, knowing the leafs gives me a perfect prediction for home ownership

Growing a regression tree

- After splitting the sample on income once, we repeat the same procedure within each leaf
- Repeat until improvements are small / few observations in each leaf



Regression Trees with multiple variables

- Example so far had only one X variable, usually we have many X variables

- Trees with many X variables
  - Choose one X variable and its cutoff so that leaf variance is minimized
  - Repeat for each leaf
  - Sub-sequent splits could be based on the same variable or a different one
  - Repeat procedure until improvement is small or number of observations at node reaches limit

Regression Trees

- Regression tries to minimize sum of variances in both leafs
- Given a parent node $\{(\mathbf{x}_k, y_k)\}_{k=1}^{n}$ the optimal split is an x_ij that splits the data

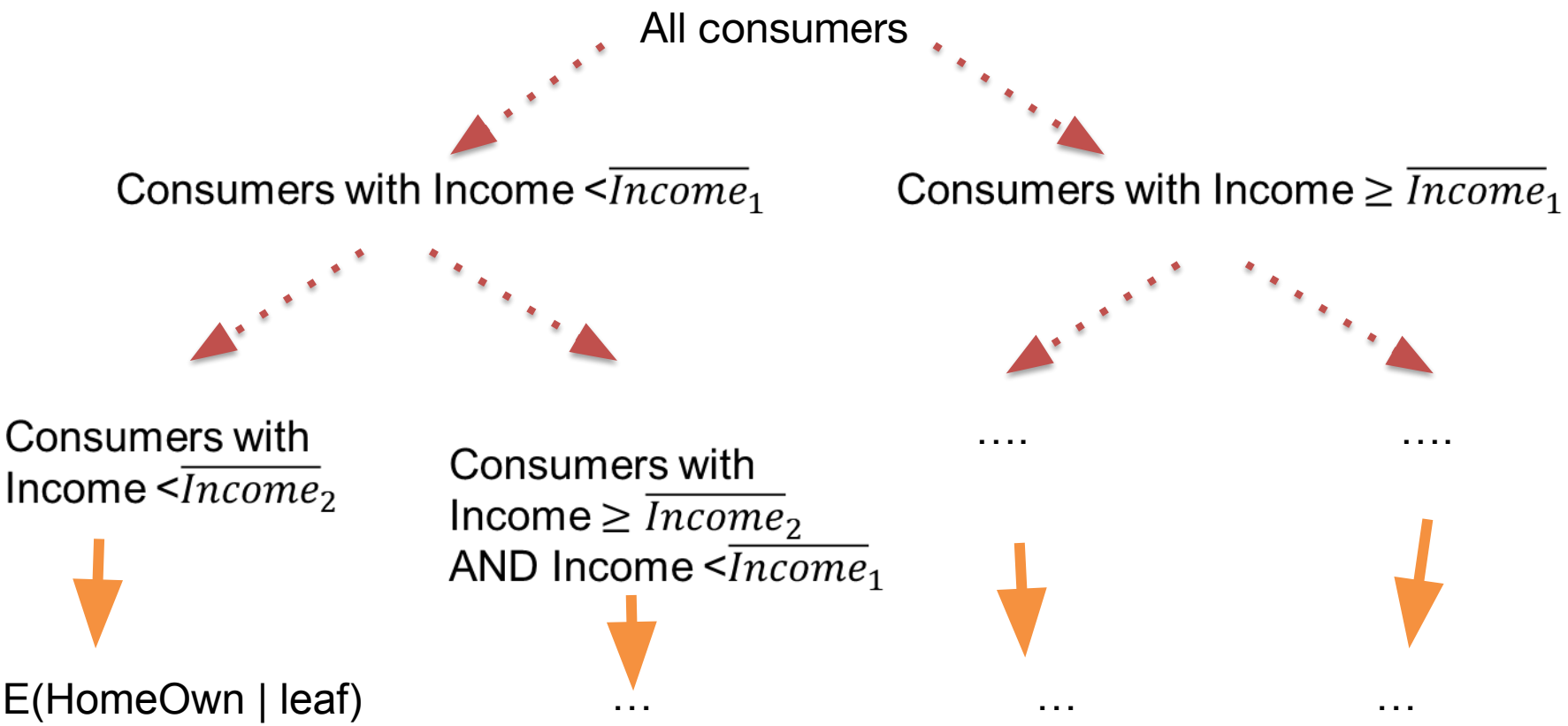
$$left : \{(\mathbf{x}_k, y_k) : x_{kj} \leq x_{ij}\} \qquad right : \{(\mathbf{x}_k, y_k) : x_{kj} > x_{ij}\}$$

$$\sum_{k \in left} (y_k - \bar{y}_{left})^2 + \sum_{k \in right} (y_k - \bar{y}_{right})^2$$

- We are making the child nodes as homogeneous as possible
- This is known as CART (classification and regression tree) which applies this logic recursively

Prediction

- For each final node / leaf of the tree, calculate average outcome
- For new consumer: find which leaf she belongs to, use average (leaf-specific) outcome as prediction



Regression Trees: Summary

- Different, nonparametric, method to make predictions
- Sample splitting instead of linear relationships

- Advantage / Disadvantages:
  - More flexible (non-linear relationships, interactions)
  - Requires more data (#variables << #observations)
  - Could test it against Lasso (on hold-out sample)
  - Could lead to overfitting

Regression Trees: Overfitting

- Can use cross-validation
  - We grow a deep tree
  - Prune the tree backwards by removing leaf splits that yield the lowest in sample error reduction
  - This process generates a sequence of models with different number of leaf nodes
  - We run cross validation on each model and pick the one with the smallest out of sample error (like in Lasso)

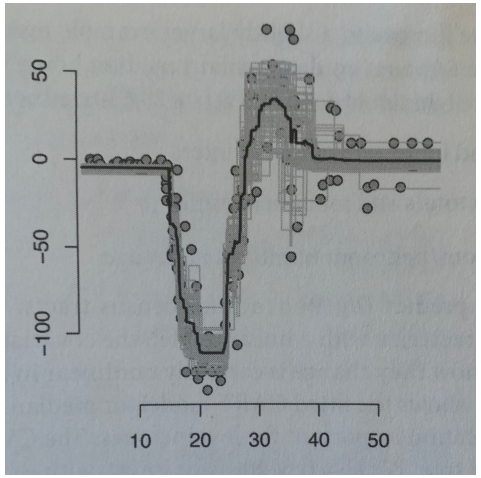
- Problem: does not work reliably in practice because of stability issues
  - Prediction rules tend to change (the tree changes) with different samples
  - Due to the nature of CV this will mean that the model choice and predictions will have high variance
  - In Lasso this is addressed with the regularizing penalty, but here we cannot add a penalty

Regression Trees: Random Forest

- Random Forests:
  - Trees can lead to overfitting when splitting “too often”
  - Solution: Grow many trees, i.e. a forest, then average across them
  - This is done by sampling (with replacement) from the raw data, then fitting a tree on each sample
  - This technique is called “bagging”

- Random Forests Algorithm: Let B be the bootstrap size
  - Run the next steps for b=1,...,B
  - Sample with-replacement n observations from the data
  - Fit a CART tree, T_b, to this sample
  - You get a a set of trees (a forest) T_1,...,T_B
  - Prediction for x:

$$\hat{y}(x) = \frac{1}{B} \sum_{b} \hat{y}_b(x)$$



Causal Trees (a Brief Intro)

- We want to estimate the heterogeneous treatment effect (HTE)

$$\gamma(x) = E[y|d = 1, \mathbf{x}] - E[y|d = 0, \mathbf{x}]$$

- We assume that we are controlling for all covariates that may cause OVB
- Similar to CART but we choose tree splits to maximize the squared difference between estimated treatment effects in each child node
- Causal Tree Algorithm:
  - Given observation {*d_i,x_i,y_i*} at node *m* of a tree the estimated treatment effect is

$$\hat{\gamma}_m = \bar{y}_{m1} - \bar{y}_{m0}$$

  - We split nodes into left and right children on the variable observation x_ij that maximizes

$$\hat{\gamma}^2_{left} + \hat{\gamma}^2_{right}$$

  - where each value above is the squared of the estimated treatment effect in each new node respectively
  - Continue until minimum number of observation in a node is reached

Causal Trees (a Brief Intro)

- “Honest” version, which has good asymptotic properties, uses two samples:
  - One for determining the tree splits
  - Another to re-estimate the treatment effect within each leaf conditional upon this tree structure
- For pruning you can use cross validation: only valid if using shallow tree due to stability issues (good enough for some application with e.g., few customer segments)
- There is also a Generalized Random Forest Framework for more complex HTE modeling
  - In R: grf package
  - In python: conml.dml.CausalForestDML
- **For more details:**
  - *Recursive partitioning for heterogeneous causal effects.* Athey and Imbens 2017
  - *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*. Athey and Wager 2018

Agenda

- Final Exam logistics
- Regression Trees
- Recap questions
- Class evaluation
- Workshop: Recap coding/Final exam like questions

Overview …

- Brief overview of methods we covered together and where to apply them
- Two approaches address different marketing questions:
- Causal inference methods
  - A/B tests
  - Control variables
  - Panel data
  - Diffs-in-diffs
  - IV
  - Regression Discontinuity
  - Causal Lasso
  - Causal Trees
- Predictive methods
  - Lasso
  - Regression Trees & Random Forest

Causal inference

- You have a business decision where you change a specific variable (price, ad)
- Want to understand how manipulating this variable will affect a specific outcome variables (profits, sales, …)

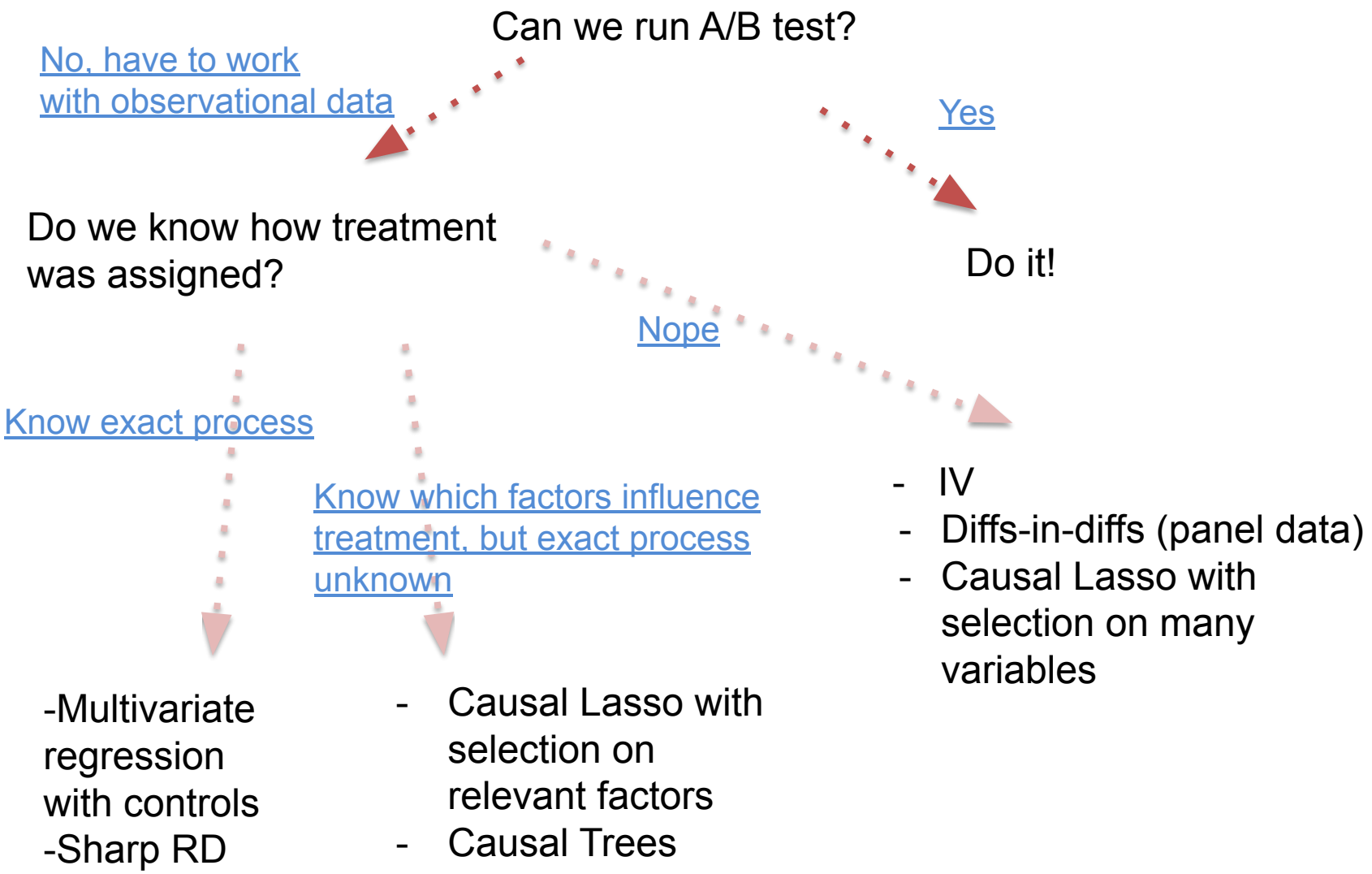
- Best solution:
  - Run A/B test
  - Always do this if feasible and not too costly
  - There is no conceptual downside

- When are A/B tests difficult?
  - Consumers do not like price experimentation
  - Experimentation requires collaboration with platform (Google, FB, …)
  - Often hard to implement in offline settings (TV advertising, shelf displays in brick-and-mortar retailers)

Measuring treatment effects in observational data

- Case 1: we know how the treatment was assigned (e.g. how ads were targeted)
  - Include variables that are correlated with treatment as controls
  - We can use Sharp RDD if we know the assignment "mechanism" (and we have the continuity assumption)
- Case 2: we know drivers but not the exact process by which they influence treatment (e.g. algorithmic targeting with known inputs)
  - Use causal Lasso (on inputs) to uncover/control for this process
  - Causal trees: deriving a partition of the population according to treatment effect heterogeneity
- Case 3: no idea what drives treatment OR know it but have no data
  - IV: find instrument that makes treatment "as if" randomly assigned
  - Diff-in-diff: if have panel data, find control group with the same assumed time evolution of the outcome variable (parallel trends)
  - Causal Lasso (on many variables): conservative approach to eliminating OVB

Decision-making flow chart for treatment effects



Predictive models

- Predictive models are useful for:

- Purely predictive exercise:
  - Which demographics predict credit default
  - Predictions can be used to deny credit / choose appropriate interest rate
  - We are not studying a variable we can manipulate (price, advertising,…), therefore just prediction is sufficient

- Targeting (in combination with A/B test)
  - We can interact demographics with treatment indicator to analyze which groups are most affected by the treatment
  - Again, we don’t aim to manipulate demographics, therefore prediction is fine

A selection of things you need to know

- You can be tested on everything in lectures/workshops/homeworks
- Here is a selection of concepts/skills we learned, in no particular order:
  - **Q:** How to interpret regression coefficients (univariate or multivariate regression; dummy or continuous variables; log transformations)
  - **A:** Pay attention to units of measurement; ”keeping everything else fixed” in multivariate regression; dummies with multiple categories; logs in either or both Y and X variables
  - **Q:** How to form predictions based on a regression model (expected value of Y for given values of X-variables)
  - **A:** Especially in complex regressions such as two-ways fixed effects
  - **Q:** Random assignment vs. partial randomization
  - **A:** Eliminate endogeneity by controlling for source of non-randomization

A selection of things you need to know

- **Q:** How to check for randomization
- **A:** Regress treatment variable on predictors (possibly using Lasso)
- **Q:** Endogeneity and possible causes for it
- **A:** Omitted variables (correlated with both X and Y); simultaneous causality
- **Q:** Omitted variable bias formula and its use to guess the direction of change of coefficients when controlling for an omitted variable
- **A:** Think of a specific omitted variable (observable or unobservable, e.g. unit- and time-fixed effects) and sign of correlation with X and Y. If bias positive, you are overestimating, so coefficient will go down once you include variable (same interpretation if coefficient is positive or negative)
- **Q:** Precision of estimators (in univariate and multivariate regressions) and what affects it
- **A:** Large sample size N, large variance of X, small variance of regression residuals. Can affect all three in experimental data. Can affect variance of residuals in observational data by including many control variables

A selection of things you need to know

- **Q:** Bias and precision in experimental vs. observational data
- **A:** Adding control variables improves precision in experimental data (no effect on bias because coeff already unbiased). It can correct bias and improve precision in observational data
- **Q:** When it is meaningful to compare the precision of the treatment coefficient in different regressions (e.g. before and after adding controls)
- **A:** In experimental data (because coefficient estimator unbiased). In observational data only if you are comparing two unbiased estimators. Not meaningful when going from one biased to one unbiased estimator
- **Q:** Frisch-Waugh theorem and how it explains how to isolate variation you want and eliminate variation you don't want
- **A:** Multivariate regression eliminates variation correlated with control variable and isolates variation not correlated with control variable

A selection of things you need to know

- **Q:** Observable vs. unobservable omitted variables and what to do about them
- **A:** Include observable omitted variable as a control. If you have panel data, can also eliminate unobservable omitted variables but only if they vary only in one dimension (across units or over time)
- **Q:** Demand estimation: challenges and solutions
- **A:** Classical example of simultaneous causality bias (prices are set based on potential demand, which can be seen as omitted variable). Solution: if you can measure potential demand (Uber: rides requested) control for it or run local regressions if surge pricing model (isolates demand changes driven by price jumps)
- **Q:** How to compute optimal sample size for A/B tests
- **A:** Need preliminary estimate of treatment effect (other studies or pilot study) and variance of regression residual (regression on controls only). Use estimate to decide SE you need to obtain significance of effect, set prob. of treatment=.5 and obtain optimal sample size N

A selection of things you need to know

- **Q:** Hypothesis testing, significance, confidence intervals, p-values
- **A:** Hypotheses on one or multiple coefficients; t- and F-tests; one-sided vs. two-sided hypotheses; how to do hypothesis testing with confidence intervals; interpretation in words of p-values, critical values, test statistics
- **Q:** False discoveries in multiple hypothesis testing
- **A:** Computing probabilities of rejecting or not rejecting one or multiple true or false null hypotheses (type I, type II error, power of a test)
- **Q:** Significance vs. causality vs. correlation vs. in-sample vs. out-of-sample fit
- **A:** Separate concepts. Correlation does not mean causation (except in experimental data or under assumptions in observational data: e.g. parallel paths in diffs-in-diffs estimator); causation or correlation may not be significant; causation or correlation may not improve out-of-sample fit (e.g., if coefficient is small Lasso will throw it out); in-sample fit always goes up when adding variables

A selection of things you need to know

- **Q:** In treatment effects estimation (in observational data) what are the differences between treated and control group that cause endogeneity
- **A:** Differences that are correlated with treatment. I.e., the groups can be different as long as treatment assignment does not depend on differences
- **Q:** For observational data, possible ways to eliminate differences between treated and control groups
- **A:** If you know the source of non-randomization (e.g. targeting mechanism) control for it (either as single variable or let Lasso uncover the mechanism) so you isolate the random part of treatment assignment.

A selection of things you need to know

- **Q:** How to deal with seasonality in treatment effect estimation
- **A:** Can control for it with diffs-in-diffs: find control group that is subject to same seasonal variation (can be a non-linear pattern and can have different level in the two groups: just needs to be parallel)
- **Q:** Know how to read regression output as typically reported in empirical studies (e.g., fixed-effects, diffs-in-diffs regression)
- **A:** Only reports coefficients for variables that vary both over time and across units. Coeffs for fixed effects not reported (only says if they have used one-way or two-ways fixed effects, which changes interpretation)
- **Q:** Think of what omitted variables panel data regression cannot control for
- **A:** Those that vary both across units and over time

A selection of things you need to know

- **Q:** How selecting the “wrong” control group can affect your estimates of treatment effects and what you can do about it (e.g. Philly soda tax example)
- **A:** Control group by definition must not be affected by treatment (if it is indirectly affected, it is not a good control group) + outcome must have same time variation as the treatment group would have had without the treatment

- **Q:** How to test for treatment effect heterogeneity
- **A:** Interaction terms between treatment and demographics. Test for significance of their coefficients. Causal trees.

A selection of things you need to know

- **Q:** What is a valid instrument and whether you can test for it
- **A:** Must be (strongly) correlated with X variable of interest (can test: strong if first stage F-stat >10). Must be exogenous (can only test if more than one instrument)
- **Q:** What standard errors to use for TSLS?
- **A:** SE from second stage regression are wrong and need to be adjusted for the fact that the regression is a generated variable (R and python do this)
- **Q:** Are more instruments better than fewer?
- **A:** If they are all valid yes: isolates more exogenous variation in X and increases precision (can also test for exogeneity if have more than one instrument)
- **Q:** What if you have weak instruments?
- **A:** Drop instruments (if have more than two) and repeat tests. Look for others

A selection of things you need to know

- **Q:** Regularization and what it does
- **A:** Selects predictors trading off fit and penalty for having too many (large) parameters. Sets to zero small coefficients that don't improve fit enough
- **Q:** When to standardize variables in Lasso
- **A:** When they have different scales. No need to standardize if you only have dummies
- **Q:** How to do inference in Lasso regressions
- **A:** Lasso is a predictive method so it doesn't give you standard errors, only coefficient estimates. Need to re-estimate model with only the predictors selected by Lasso

A selection of things you need to know

- **Q:** How to interpret coefficients when you have both slope- and intercept dummies for a demographic (interacted and non-interacted demographic) or only one of the two
- **A:** You need to have both interacted and non-interacted demographic to estimate the treatment effect for a specific demographic group.
- **Q:** Intuition for causal Lasso
- **A:** We leave out variables that are only mildly correlated with treatment and outcome. Variables with large effects are included. It is a conservative approach to removing OVB: OVB caused by variables that correlate with both X and Y, but here we control for correlation with either.

Agenda

- Final Exam logistics
- Regression Trees
- Recap questions
- Class evaluation
- Workshop: Recap coding/Final exam like questions

Class Evaluations

- Please check your e-mail or BruinLearn.
- I'll leave for 15min
- We will do the workshop after

Agenda

- Final Exam logistics
- Regression Trees
- Recap questions
- Class evaluation
- Workshop: Recap coding/Final exam like questions