

Causal Inference Crash Course

Defining Some Causal Models

Julian Hsu

Overview

- This presentation will define some propensity-matching based models:
 1. Ordinary Least Squares (OLS)
 2. Propensity Binning with Regression adjustment
 3. Inverse propensity weighting
 4. Double machine learning - Partial Linear Model (PLM)
 5. Double machine learning - Interactive Regression Model (IRM)

- For each model, we will define the estimator and its properties.
 - So yes, there will be a lot of math.

Ordinary Least Squares (OLS)

- We have an outcome Y_i , pre-treatment features X_i , and a treatment indicator W_i . We want to know the causal relationship between Y_i and W_i .
- We can estimate this relationship by estimating an OLS model:

$$Y_i = \beta X_i + \tau W_i + \epsilon_i$$

- Where $(\hat{\beta}, \hat{\tau})$ are estimated to minimize the mean squared error:

$$\operatorname{argmin}_{\hat{\beta}, \hat{\tau}} \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\beta} X_i - \hat{\tau} W_i)^2 \right\}$$

- We know OLS is simple but why is it causal?

Why is OLS Causal?

- The OLS model estimate $(\hat{\beta}, \hat{\tau})$ based on the assumption that the mean squared error is zero, conditional on (X_i, W_i) .

- This is defined in the moment conditions:

$$E[(Y_i - \hat{\beta}X_i - \hat{\tau}W_i) \times X_i] = 0 \text{ and}$$

$$E[(Y_i - \hat{\beta}X_i - \hat{\tau}W_i) \times W_i] = 0.$$

- After conditioning on X_i , the unexplained variation in Y_i is mean independent of treatment.

- Therefore, they corresponds to the unconfoundedness assumption
- Under mean independence, $\hat{\tau}$ is an unbiased ATE/ATET estimate.

OLS as a propensity-based matching model

- OLS implicitly estimates a propensity score
- Recall that the OLS estimator is:

$$\hat{\beta} = (X_i' X_i)^{-1} X_i' Y_i = \frac{\text{cov}(X_i, Y_i)}{\text{var}(X_i)}$$

- From the Frisch-Waugh-Lovell theorem, we can be more specific about $\hat{\tau}$. (Appendix has more details)

$$\hat{\tau} = \frac{\text{cov}(\tilde{W}_i, Y_i)}{\text{var}(\tilde{W}_i)}$$

- Where $\tilde{W}_i = E[W_i|X_i] - W_i$. Well, $E[W_i|X_i]$ is the propensity score!

Propensity Binning with Regression Adjustment

- What is Regression Adjustment?
- Recall the potential outcome (Neyman-Rubin) framework:
 - Average Treatment Effect on the Treated
 $(ATET) = Y_i(1, 1) - Y_i(1, 0)$
 - Average Treatment Effect (ATE)
 $= E_1[Y_i(1, 1), Y_i(0, 1)] - E_0[Y_i(0, 0), Y_i(1, 0)]$
 - Where E_x is the weighted average of observed and counterfactuals for $W = x$

- The problem is that we do not observe $y_i(0, 1)$ and $Y_i(1, 0)$
- Regression Adjustment model asks: "What if we treat estimating $Y_i(0, 1)$ and $Y_i(1, 0)$ as a pure prediction problem and predict out of sample?"

Regression Adjustment Algorithm

1. Start with the control ($W_i = 0$) sample. Train your favorite ML model to predict $Y_{i,W_i=0}$ using $X_{i,W_i=0}$. Call this trained model $g_0(X_i)$. Do the same with the treatment ($W_i = 1$) sample, call the trained model $g_1(X_i)$.
2. Estimate the outcomes under treatment and control with $g_0(X_i)$ and $g_1(X_i)$. This gives you $\hat{Y}_i(1, 1)$, $\hat{Y}_i(0, 1)$, $\hat{Y}_i(1, 0)$ and $\hat{Y}_i(0, 0)$.
3. Estimate ATE or ATET:

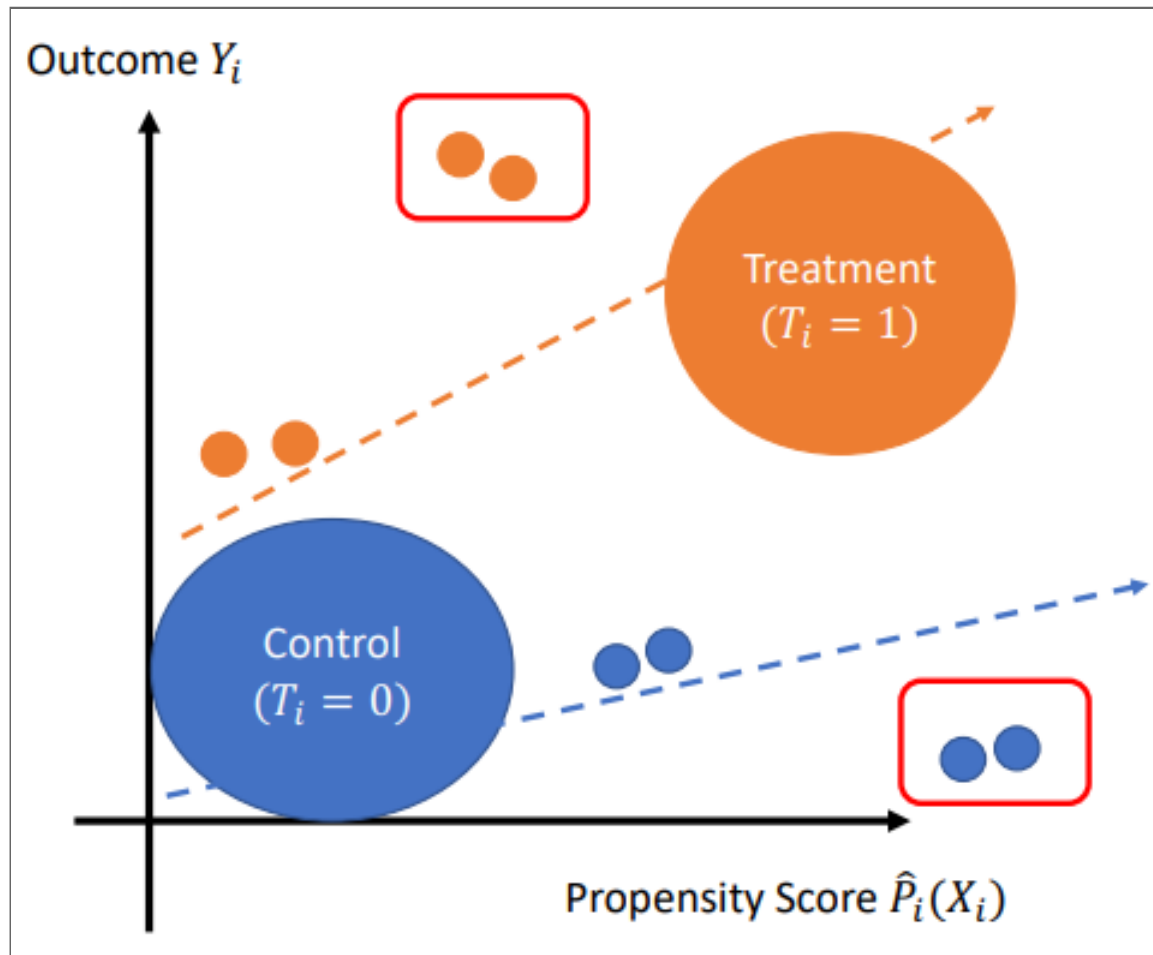
$$\text{A. } \hat{ATE} = \hat{Y}_i(1, 1) - \hat{Y}_i(1, 0)$$

$$\text{B. } ATE = E_1[\hat{Y}_i(1, 1), \hat{Y}_i(0, 1)] - E_0[\hat{Y}_i(0, 0), \hat{Y}_i(1, 0)]$$

Regression Adjustment is OLS

- Regression Adjustment is OLS in disguise.
- Intuitively, this is because the OLS is also extrapolating the potential outcomes with a single model, instead of multiple. (Appendix has the technical explanation.)
- Therefore, Regression Adjustment implicitly estimates a propensity score because OLS does too.

Propensity Binning as Insurance Against Outliers



- We show in the dotted lines the predicted outcomes for control and treatment.
- But we have outliers, based on propensity score, which are influencing our predicted outcomes.
- How can we flexibly accommodate the outliers?

Propensity Binning with Regression Adjustment

- We estimate a propensity score, $\hat{P}_i(X_i)$ and then divide the data into segments of $\hat{P}_i(X_i)$.
- For each segment, implement the Regression Adjustment model.

Inverse Propensity Weighting

- This approach is inspired by sampling methods.
- Suppose you have :
 - (A) treatment observation with a propensity score of 0.99
 - (B) treatment observation with a propensity score of 0.01
- You most likely have a lot of (A), but not a lot of (B). So you want to give (B) more weight in your analysis because it happens very rarely.

Inverse Propensity Weighting Model Definition

- The Inverse Propensity Weighting (IPW) estimator for the ATE is:

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i Y_i}{\hat{P}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{P}(X_i)} \right]$$

- For ATET:

$$\frac{1}{N} \sum_{i=1}^N \hat{P}(X_i) \left[\frac{W_i Y_i}{\hat{P}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{P}(X_i)} \right]$$

- We will unpack the ATE to intuitively understand it.

Inverse Propensity Weighting Intuition

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i Y_i}{\hat{P}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{P}(X_i)} \right]$$

- When we divide a treatment observation by the propensity score, we are increasing its importance when it is less likely to be treated.
- Note that the denominator of $\frac{1}{N} \frac{1}{\hat{P}(X_i)}$ approximates taking the average of just the treated observations.

Advantages and Disadvantages

- **Advantage:** Weighting provides flexible form and only requires diagnosing with propensity score
- **Disadvantage:** since we divide by the propensity score, you risk imprecise estimates if you have a lot of propensity scores near zero or one.
 - This means the variance can explode and be very large.
- Solutions are:
 - Drop these observations;

- Replace these observations' propensity scores with a pre-determined value (like 0.001 or 0.999) or unconditional probability of treatment.

Double Machine Learning (DML)

- Yes, we are finally here.
- At a high-level, DML is a more flexible version of the previous models
- Note that DML relies on the same assumptions as the other models presented here
- We will cover the propensity-matching based models from the Chernozhukov et al. (2016) paper.
 - Partial Linear Model
 - Interactive Regression Model

What can ML do for causal inference?

- Off-the-shelf ML models also cannot be used for inference.
 - Exceptions: generalized random forests (Athey et al. 2018); neural nets (Farrell et al. 2020)
- We can use ML in two ways:
 1. Estimate a better propensity score. Recall that for OLS: $\hat{\tau} = \frac{\text{cov}(\tilde{W}_i, Y_i)}{\text{var}(\tilde{W}_i)}$

2. Estimate better counterfactuals as we do for
Regression Adjustment models

High-Level Strategy for DML

- We use two ML strategies so we can do inference:
 1. Regularization based on residualization
 - Based on Frisch-Waugh-Lovell theorem (recall this from the OLS slides?)
 - Compare residuals that are constructed to be independent except due to the variation of interest (Neyman orthogonality)
 2. Sample-splitting to prevent overfitting
 - Cross-validation is important to make sure prediction is not biased

DML – Partial Linear Model Motivation

- The partial linear model is a form of OLS. We allow for a more flexible prediction of Y_i and W_i :
- OLS:

$$\begin{aligned}Y_i &= \hat{\beta} X_i + \hat{\tau} W_i + \epsilon_i \\Y_i &= g_0(X_i) + \hat{\tau} W_i + \epsilon_i \\Y_i - g_0(X_i) &= g_0(X_i) - g_0(X_i) + \hat{\tau} W_i + \epsilon_i \\Y_i - g_0(X_i) &= \hat{\tau} W_i + \epsilon_i\end{aligned}$$

- We replace $\hat{\beta} X_i$ with $g_0(X_i)$. From this we can show that estimating a regression of the residualized Y_i based on $g_0(X_i)$ on W_i estimates the treatment effect. • OLS also requires a residualized W_i , from the Frisch-Waugh-Lovell theorem.

DML – Partial Linear Model Setup

- The partial linear model is a form of OLS.
- OLS:

$$Y_i = \hat{\beta} X_i + \hat{\tau} W_i + \epsilon_i$$

- DML - Partial Linear Model:

$$Y_i = g_0(X_i) + \hat{\tau} W_i + \epsilon_i$$

$$W_i = m_0(X_i) + \nu_i$$

- Assumptions are still the same as OLS (especially unconfoundedness)
- We still assume the treatment effect is linearly additive

Partial Linear Model Procedure

$$Y_i = g_0(X_i) + \hat{\tau}W_i + \epsilon_i$$

$$W_i = m_0(X_i) + \nu_i$$

- First Stage:
 1. Predict Y_i using X_i with sample-splitting, get \hat{Y}_i .
 2. Predict w_i using X_i with sample-splitting, get \hat{W}_i .
 3. Calculate the residuals for Y_i and w_i .
Specifically, $\tilde{Y}_i = Y_i - \hat{Y}_i$ and $\tilde{W}_i = W_i - \hat{W}_i$.
- Second Stage:
 - Estimate the OLS model:

$$\tilde{Y}_i = \hat{\tau} \tilde{W}_i + \zeta_i$$

What ML models can be used?

- You can use essentially any ML model for the first stage to generate \hat{Y}_i and \hat{W}_i .
- These are called the “nuisance parameters” because we care about the quality of the prediction, not the theoretical properties of the ML model

DML – Interactive Regression Model

- The partial linear model is simple and intuitive because it is a more flexible version of OLS.
- Despite this, we are restricted by the assumption that treatment linearly interacts with the outcome.
- It also assumes that the treatment effect is the same for all observations. Specifically, that the average treatment effect (ATE) is the same as the average treatment effect on the treated (ATET).

Interactive Regression Model (IRM) Specification

$$\hat{\tau}_{ATE} = E\left[(\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(\hat{Y}_i - \hat{Y}_{1,i})}{\hat{P}(X_i)} - \frac{(1 - W_i)(\hat{Y}_i - \hat{Y}_{0,i})}{1 - \hat{P}(X_i)}\right]$$

$$\hat{\tau}_{ATET} = E\left[\frac{W_i(\hat{Y}_i - \hat{Y}_{1,i})}{P \times \hat{P}(X_i)} - \frac{\hat{P}(X_i)(1 - W_i)(\hat{Y}_i - \hat{Y}_{0,i})}{P \times 1 - \hat{P}(X_i)}\right]$$

- When we estimate ATET, we no longer need $\hat{Y}_{1,i}$
- Estimating the IRM model has the same first stage as PLM. The only difference is the second stage.
- Now we will explain the components of ATE, which generalize to ATET.

Explaining the IRM Model, Part 1

$$\hat{\tau}_{ATE} = E \left[(\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(\hat{Y}_i - \hat{Y}_{1,i})}{\hat{P}(X_i)} - \frac{(1 - W_i)(\hat{Y}_i - \hat{Y}_{0,i})}{1 - \hat{P}(X_i)} \right]$$

- This first part is just regression adjustment, which is biased if there is large estimation error
- The second parts are the estimation error of the outcome for treatment and control units, which are weighted by propensity scores
- We combine regression adjustment and propensity weighting for a “doubly robust” approach (more

details in the appendix) where we can correct for our regression adjustment estimates.

Explaining the IRM Model, Part 2

- Re-writing the previous equation:

$$\begin{aligned}\hat{\tau}_{ATE} &= E\left[(\hat{Y}_{1,i} - \hat{Y}_{0,i}) + \frac{W_i(\hat{Y}_i - \hat{Y}_{1,i})}{\hat{P}(X_i)} - \frac{(1 - W_i)(\hat{Y}_i - \hat{Y}_{0,i})}{1 - \hat{P}(X_i)}\right] \\ &= E\left[(\hat{Y}_{1,i} + \frac{W_i(\hat{Y}_i - \hat{Y}_{1,i})}{\hat{P}(X_i)}) - (\hat{Y}_{0,i} - \frac{(1 - W_i)(\hat{Y}_i - \hat{Y}_{0,i})}{1 - \hat{P}(X_i)})\right]\end{aligned}$$

- Therefore, we are correcting our estimates of $\hat{Y}_{1,i}$ and $\hat{Y}_{0,i}$

- Applying this observation level correction means that there is variation at treatment estimates at the observation level.

Conclusion

- This presentation defines some propensity-matching based models:
 1. Ordinary Least Squares (OLS)
 2. Propensity Binning with Regression adjustment
 3. Inverse propensity weighting
 4. Double machine learning - Partial Linear Model (PLM)
 5. Double machine learning - Interactive Regression Model (IRM)

- Remember they all rely on the same assumptions!

Appendix Slides

Frisch-Waugh-Lovell Theorem

- The OLS estimator is based on the Frisch-Waugh-Lovell theorem, or “partialing out”.
- Pay attention. This is the same trick behind double machine learning. •* Start with:

$$\text{cov}(\tilde{W}_i, Y_i) = \text{cov}(\tilde{W}_i, \beta X_i + \tau W_i + \epsilon_i)$$

- Know that $\text{cov}(\tilde{W}_i, \beta X_i) = 0$ because $\tilde{W}_i = E[W_i|X_i] - W_i$, so it already conditions on X_i .
- Also know that $\text{cov}(\tilde{W}_i, \epsilon_i) = 0$ for the same reason.

- Then $cov(\tilde{W}_i, Y_i) = cov(\tilde{W}_i, \tau W_i) = \tau var(\tilde{W}_i)$

Is LASSO an improvement on OLS?

model

Ordinary Least Squares

Objective
Function

$$\operatorname{argmin}_{\hat{\beta}, \hat{\tau}} \left\{ \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\beta} X_i - \hat{\tau})^2 \right\}$$

- LASSO regression coefficients are chosen to maximize prediction, subject to a constraint in the parameters.
- Intuitively, it assumes that coefficients are zero and there are penalties non-zero coefficients.

- Certainly, LASSO has better out-of-sample prediction. But can we use it for causal inference?

Can we use LASSO for causal inference?

- No, we can't. Here is a technical and intuitive explanation.
- Technically, OLS identifies the causal estimate because of this moment condition you can get from solving the optimization problem:

$$E[(Y_i - \hat{\beta}X_i - \hat{\tau}W_i) \times X_i] = 0$$

But you can't get this from a LASSO.

- Intuitively, a LASSO coefficient has two interpretations: the causal estimate of $\hat{\tau}$ and a

feature selection of whether w_i is important to the prediction problem.

- Then the unconfoundedness assumption may no longer hold.

Appendix Slides – Doubly Robust Models

Doubly Robust

- What if the propensity score is wrong?
- It can be wrong because we do not have enough features, or we have the incorrect model specification.
- This causes a problem because recall that we can estimate the treatment effect using the difference between treatment status and the propensity score.

Recall from the OLS slides:

$$\hat{\tau} = \frac{cov(\tilde{W}_i, Y_i)}{var(\tilde{W}_i)}$$

where $\tilde{W}_i = E[W_i|X_i] - W_i$.

- Ideally \tilde{W}_i represents the conditionally random variation in treatment. But if the propensity model is wrong, then it also incorporates model error.
- Then, we have the wrong value for $\hat{\tau}$!

Model mis-specification

- The same logic applies to predicting the counterfactual outcome.
- Model mis-specification isn't completely solved with a good ML model.
- For example, a counterfactual outcome prediction is an extrapolation exercise.

Doubly Robust Mental Model

- A “doubly robust” model attempts to address model mis-specification.
- It incorporates:
 1. propensity score; and
 2. counterfactual models such that it will give you the correct value for $\hat{\tau}$, as long as one of the models is correct.
- There are a lot of functional forms, but we’ll show a popular one next.

Augmented Inverse Propensity Weight (AIPW) Model

Definition

- from Robins, Rotnitzky, and Zhao (1994)
- ATE:

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i Y_i}{\hat{P}} - \frac{(1 - W_i) Y_i}{1 - \hat{P}} \right] - \frac{W_i - \hat{P}}{\hat{P}(1 - \hat{P})} [(1 - \hat{P})(\hat{Y}_{i,1}) + \hat{P}(\hat{Y}_{i,0})]$$

- We will build intuition why this works under misspecification of either \hat{P} or $\hat{Y}_{i,1}/\hat{Y}_{i,0}$.
- This is similar to a double machine learning implementation

AIPW - $\hat{Y}_{i,1}/\hat{Y}_{i,0}$ are wrong

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i Y_i}{\hat{P}} - \frac{(1 - W_i) Y_i}{1 - \hat{P}} \right] - \frac{W_i - \hat{P}}{\hat{P}(1 - \hat{P})} [(1 - \hat{P})(\hat{Y}_{i,1}) + \hat{P}(\hat{Y}_{i,0})]$$

- If $\hat{Y}_{i,1}/\hat{Y}_{i,0}$ are wrong, but \hat{P} is right, then then $\hat{P} = E[W_i|X_i]$ so term $W_i - \hat{P}$ is zero in expectation.
- Therefore, in expectation, we are left with this:

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i Y_i}{\hat{P}} - \frac{(1 - W_i) Y_i}{1 - \hat{P}} \right]$$

- Which is the inverse propensity weighting model

AIPW - \hat{P} is wrong

- Note we can re-arrange the AIPW estimator to look like this:

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{W_i(Y_i - \hat{Y}_{i,1})}{\hat{P}} + \hat{Y}_{i,1} - \frac{(1 - W_i)(Y_i - \hat{Y}_{i,0})}{W_i - \hat{P}} - \hat{Y}_{i,0} \right]$$

- If \hat{P} is wrong, but $\hat{Y}_{i,1}/\hat{Y}_{i,0}$ are right, then the difference in observed and predicted outcomes are zero in expectation.
- Therefore, in expectation, we are left with this:

$$\frac{1}{N} \sum_{i=1}^N [\hat{Y}_{i,1} - \hat{Y}_{i,0}]$$

- Which is the regression adjustment model

More Misspecification Problems

- Doubly robust methods allow us to have misspecification in \hat{P} or $\hat{Y}_{i,1}/\hat{Y}_{i,0}$.
- We can get even more flexible with ML models and relax functional forms.
- Flexibility becomes important if x_i becomes high dimensional.
- Note that x_i include generated features.
 - For example:
 - Interact statuses with previous spending

- Basis function transformations of previous spending (polynomials, segments, etc.)

Appendix Slides –Instrumental Variables

Instrumental Variables (IV) - Motivation

- The unconfoundedness assumption may not be satisfied. There is selection bias into W_i unexplained by X_i
- Are we blocked? Not necessarily.
- Suppose we have a feature Z_i that directly determines W_i but does not directly determine Y_i
- In other words, Z_i only affects Y_i through W_i

IV Two-Stage Model

$$Y_i = \beta X_i + \tau W_i + \epsilon_i$$

$$W_i = \alpha X_i + \gamma Z_i + \psi_i$$

- We can use this model to estimate τ in a two stage process.
 1. Estimate W_i , get \hat{W}_i
 2. Estimate $Y_i = \hat{\beta} X_i + \hat{\tau} W_i + \epsilon_i$
- Why does this work? We study the necessary assumptions and intuition

IV Assumptions

$$Y_i = \beta X_i + \tau W_i + \epsilon_i$$

$$W_i = \alpha X_i + \gamma Z_i + \psi_i$$

1. Exclusion Restriction: $cov(\epsilon, Z_i) = 0$, in other words Z_i is uncorrelated to Y_i conditional on X_i and W_i /

- Example of a violation: suppose that we want to know the impact of a missed delivery date (W_i) on future spending (Y_i). We propose using an extreme weather event (Z_i), like an earthquake, as an instrument for getting a missed promise. This would not work because an

earthquake would affect future spending for reasons unrelated to a missed delivery occurring.

2. Strong instrument: $cov(W_i, Z_i) \neq 0$, in other words Z_i is a strong predictor of W_i .

Intuitively, why does this work?

- In the second stage, we estimate $Y_i = \hat{\beta}X_i + \hat{\tau}W_i + \epsilon_i$
- Then, variation in \hat{W}_i depends on X_i and Z_i . Let's focus on the variation due to Z_i .
- If we think back to the Frisch-Waugh-Lovell theorem, the OLS coefficient is based on the difference between W_i and $\hat{W}(X_i)$. Here in the IV setting, we are looking at the difference between $\hat{W}(X_i, Z_i)$ and $\hat{W}(X_i)$.
- Therefore, we are relying in variation in W_i based on Z_i which we assume is exogenous to Y_i conditional

on X_i .

In []: