

Regression

Statistic Modeling & Causal Inference – Oswald | Ramirez-Ruiz

Agenda

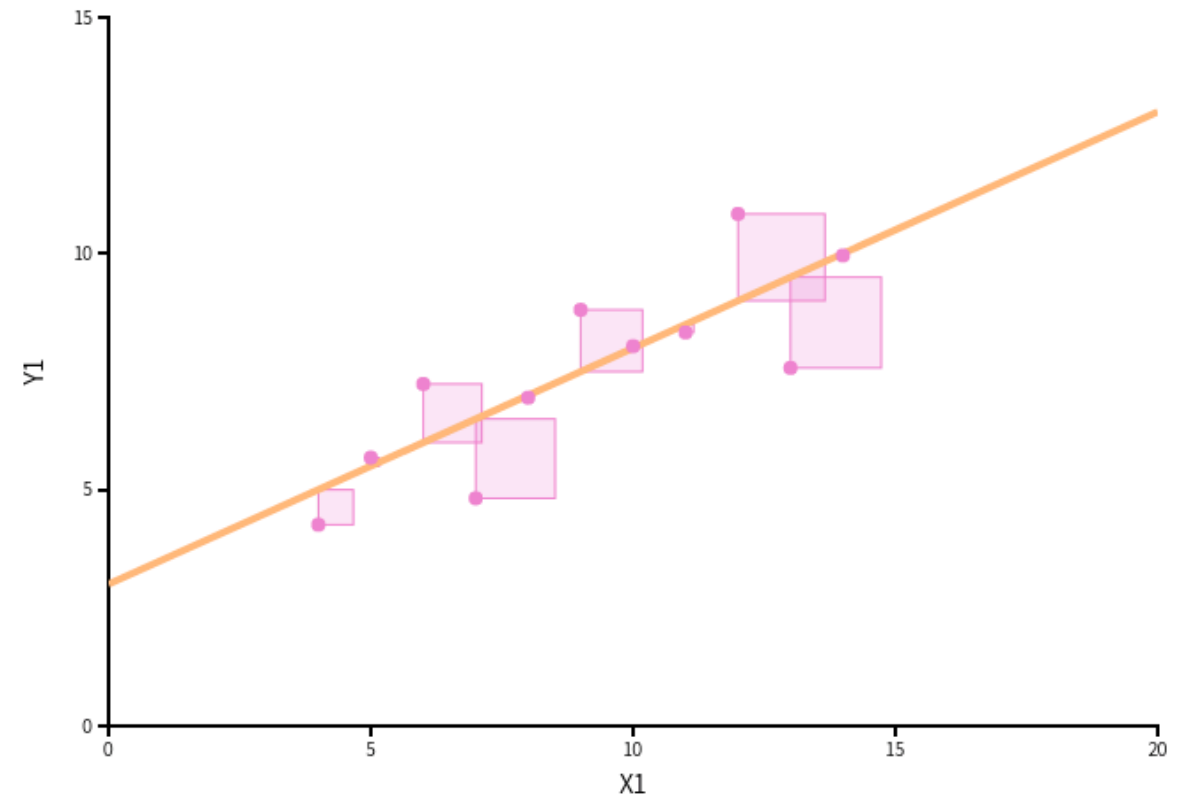
- Regression from a causal perspective
- Back-door criterion

- Regression in R
- DAGs in R

Ordinary Least Squares (OLS)

- Minimizing sum of squares of residuals (differences between observed and predicted values)
- Finding the best (linear) guess for y given a particular x value

Minimize $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$



<https://seeing-theory.brown.edu/regression-analysis/index.html>

Regression Coefficients

Calculation of intercept

Y value given $X = 0$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

Calculation of slope

How much does Y change when X increases by 1 unit

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

(in bivariate regression)

Omitted Variable Bias (OVB)

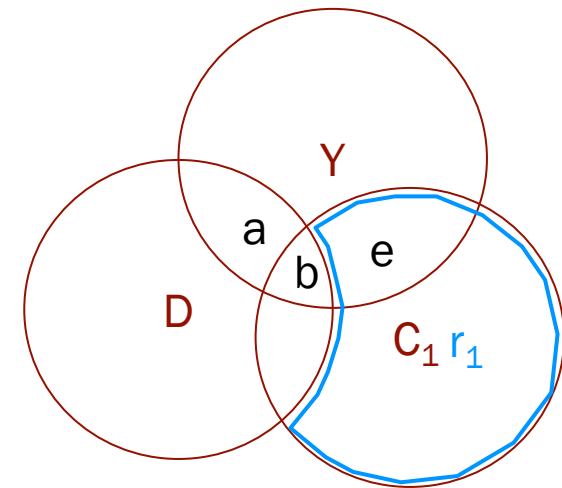
- Bivariate relationships can be confounded by other variables
 - Occurs when Z is a common cause of both X and Y
- Include Z in regression to (partially) deal with the issue

Multiple Regression

Multiple Regression

- Multiple regression only uses the unique variation in each regressor (X_i) not explained by other regressors
- Thus, β_1 can be estimated by

$$\hat{\beta}_1 = \frac{\sum \hat{r}_{i1} y_i}{\sum \hat{r}_{i1}^2}$$



- where r is the residual from a regression of C_1 on the other explanatory variables
- r_{i1} is c_{i1} after the effects of c_{ij} (and d_i) have been “partialled out”

Regression from a POF perspective

- Regression can be utilized without thinking about causes as a predictive or summarizing tool.
- It would not be appropriate to give causal interpretations to any, unless we establish the fulfilment of certain assumptions.

$$Y_i = \beta_0 + \beta_1 D + e_i \quad ?$$

$$E(Y^0|D = 0) = \beta_0$$

$$E(Y^1|D = 1) = \beta_0 + \beta_1$$

$$\beta_1 = E(Y^1|D = 1) - E(Y^0|D = 0) \\ = \text{NATE}$$

Regression Error Terms

$$Y_i = \beta_0 + \beta_1 D + e_i \neq Y_i = \beta_0 + \beta_1 X + r_i$$

- Error term in causal perspective: “summary” random variable representing all causes other than D (and other modeled regressors)
- regression residual r , which is uncorrelated with the regressors by construction

→ Only if D and e were independent (e.g., due to random assignment of D), the regression estimate of β_1 could be given a causal interpretation $\beta_1 = ATE$

Back to OVB

- Violations of the assumption that D and e are independent:
→ OVB problem

“true” causal model: $Y_i = \beta_0 + \beta_1 D_i + \beta_2 Z_i + e_i$,

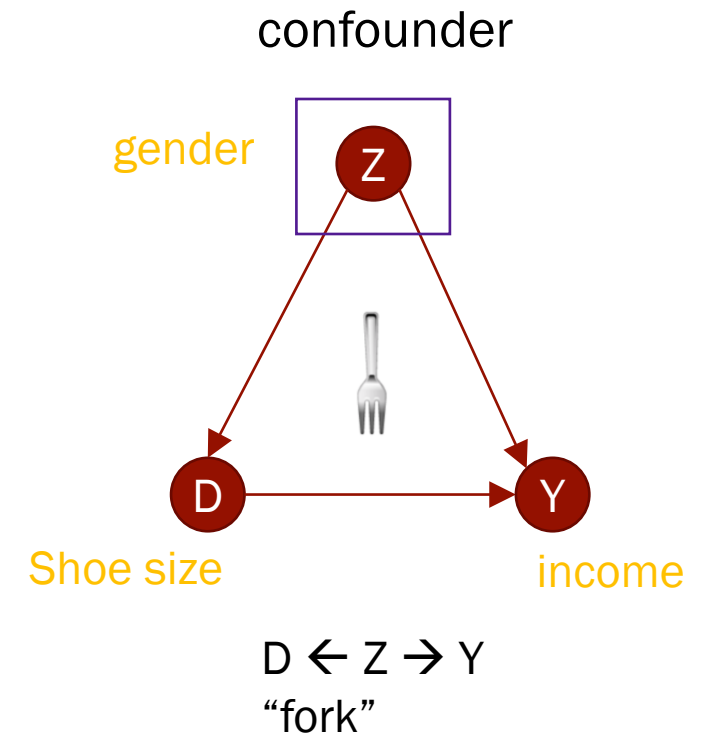
Fitted empirical model: $Y_i = \gamma_0 + \gamma_1 D_i + r_i$

Relationship between Z and D : $Z_i = \delta_0 + \delta_1 D_i + u_i$

OVB: $\gamma_1 - \beta_1 = \beta_2 \delta_1$ → does Z influence Y ? β_2
→ does Z influence D ? δ_1

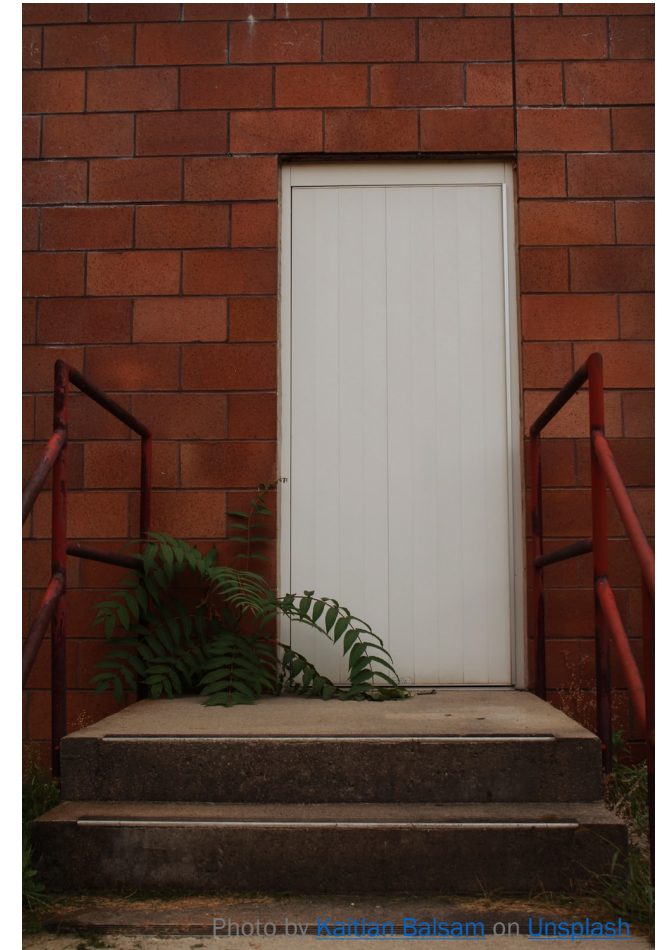
Tackling OVB with Regression

- “Control for” / “Condition on” confounders by including them in the model
- find the factors responsible for different baseline values (or differential treatment effects), and to include these variables in the equation in the hope that an unbiased estimate of β_1 is obtained
- **But:** we need to **include all relevant covariates** and there has to be a large enough overlap in covariate values across different values of D (“common support”)



Selecting Covariates

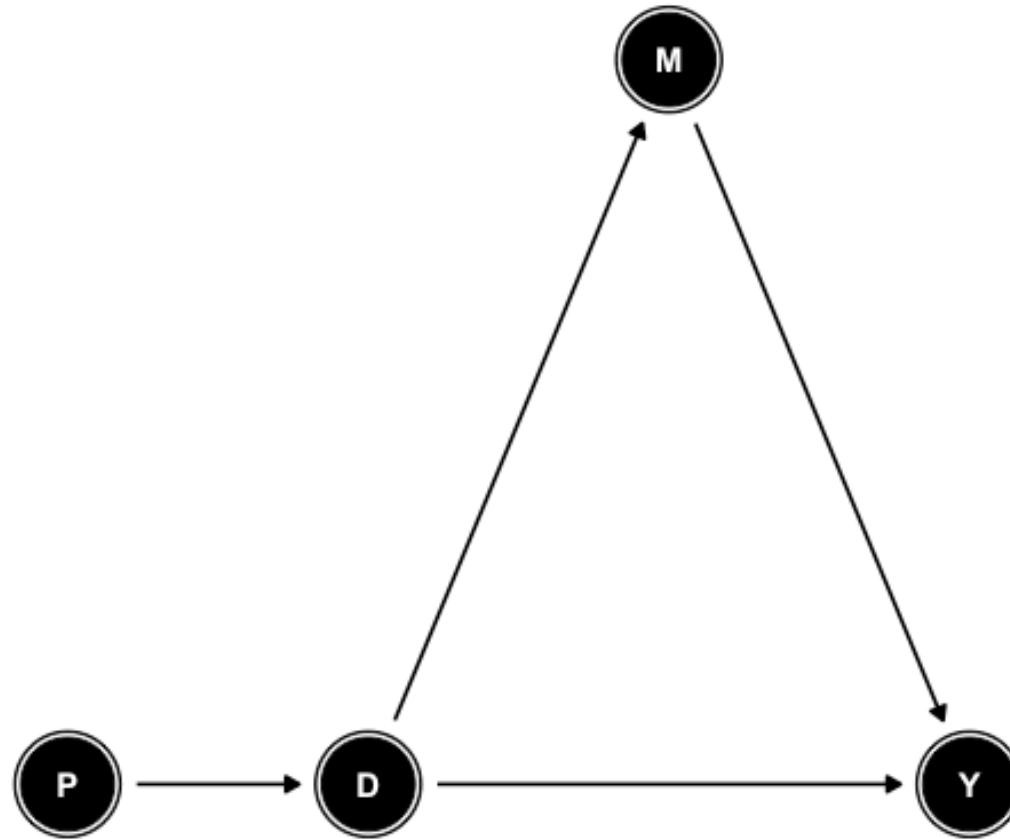
- Draw DAG
- Write down all paths between D and Y
- Identify conditions that satisfy **back-door-criterion**
- Control for the identified variables in model
- Only interpret D causally! The status of covariates is path-specific



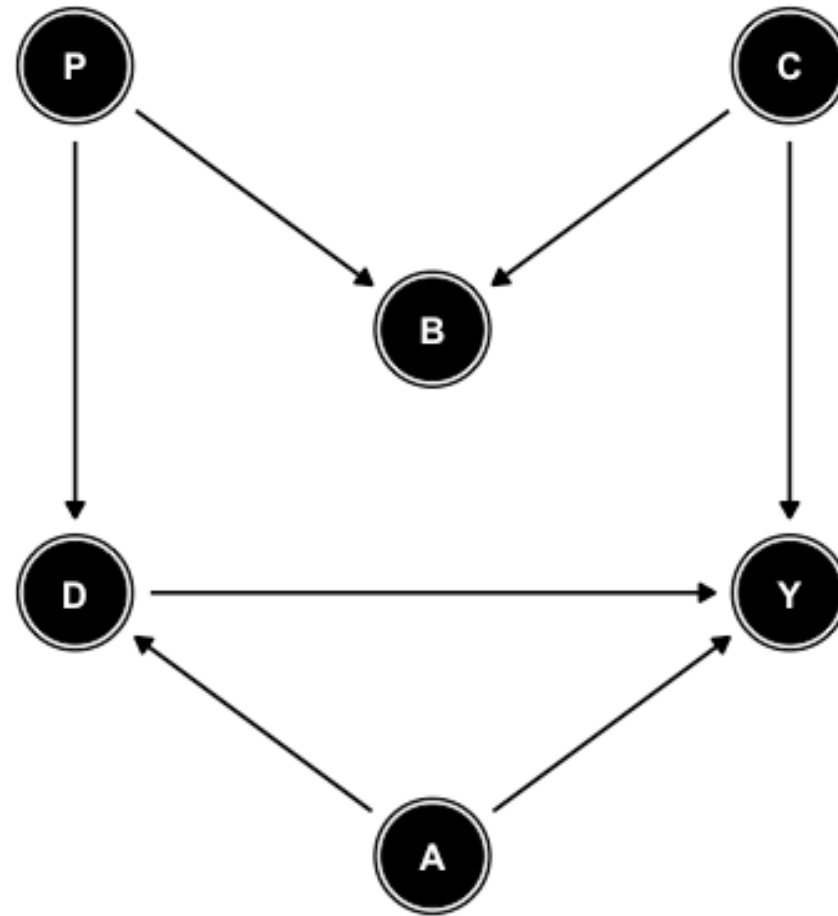
Back-door criterion

- Potential shortcut to formal adjustment criterion
- Focus on **non-causal paths that start with an arrow into D** (back-door-paths)
- To identify the total effect of D on Y , you need to condition on observed variables Z so that
 - no element of Z is a descendant of D , and
 - Z blocks all back-door paths from D to Y
- Remember: **Confounders, Mediators, Colliders...**

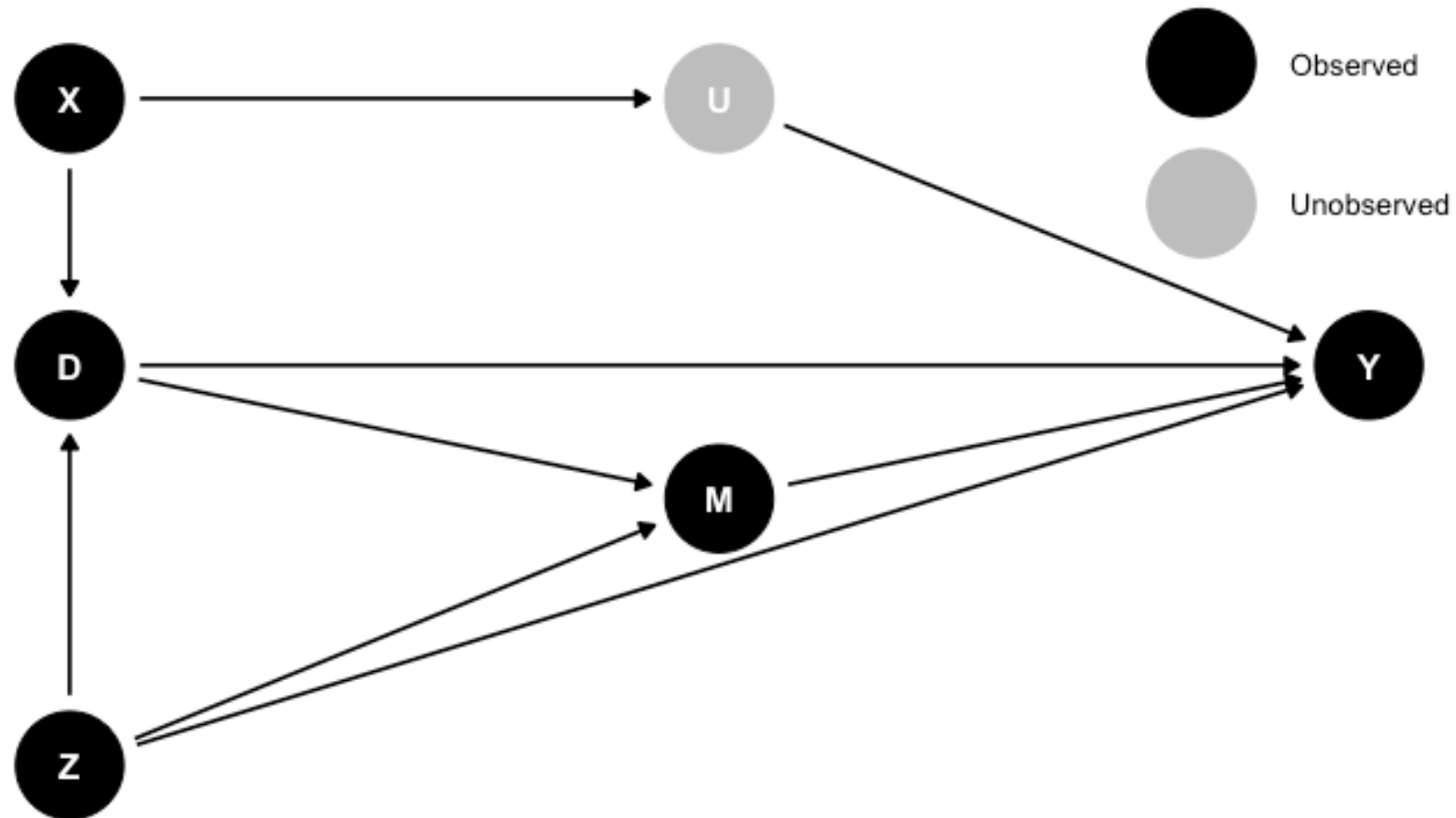
Example 1



Example 2



Example 3



Further Resources

For any coding issues – [Stackoverflow](#)

Hertie's Data Science Lab – [Research Consulting](#)