POLSCI 630 Introduction to Regression Methods Spring 2026

Lecture 1 — Roadmap and Population Model

Jiawei Fu, Duke University *Scribe: Jiawei Fu*

1 Overview

In social science, we seek to study the causal effects of one or more variables on an outcome of interest. For example: What is the causal effect of campaign spending on voting behavior? or What is the causal effect of economic development on democracy? Empirical research uses data to answer such questions in a systematic and scientific manner. Traditionally, regression has been the primary tool for this purpose. However, we are currently in a methodological transition in which traditional statistical methods are increasingly integrated with modern causal inference frameworks, particularly in political science. Accordingly, in this course, we will study foundational regression techniques from both statistical and econometric perspectives and combine them with a causal-inference approach.

Of course, data can also be used to answer other types of questions, such as descriptive and predictive ones. In this course, however, our primary focus is on causal questions. Prediction problems are better studied using nonparametric methods and machine-learning techniques, which will be covered in advanced class like computational social science.

In this lecture, we introduce the overall roadmap of empirical research. We highlight four key methodological components that researchers must consider: the estimand, identification, estimation, and inference. We then introduce the linear model and clarify the assumptions underlying four different modeling approaches. This material lays the foundation for the rest of the course.

2 Empirical Studies are Challenging!

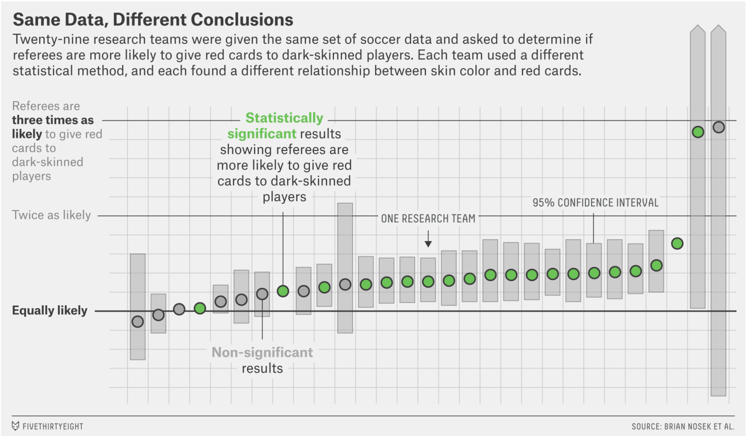

Figure 1: Cite from https://fivethirtyeight.com/features/science-isnt-broken/

3 Roadmap of Empirical Study

Step 1: Estimand. What is your target parameter of interest? This choice should be driven by the substantive purpose of the study. The estimand may be the average treatment effect of a treatment variable on an outcome of interest, or it may be a structural parameter in a regression model—for example, β_1 in the model $Y = \beta_0 + \beta_1 X + \epsilon$.

In traditional econometrics, the choice of estimand is often guided by the structure of the data. For instance, panel data may lead researchers to adopt random- or fixed-effects models, while binary outcomes often motivate the use of logistic regression. We emphasize, however, that the estimand should be determined by the underlying scientific question and the research design, rather than by the data structure alone.

Step 2: Identification. Suppose you have access to infinite data. Under what assumptions can your estimand be uniquely expressed as a function of the observed data? If this is not possible, the estimand is not identified, meaning that no amount of data can be used to learn the target parameter.

Identification necessarily relies on assumptions. Data alone are insufficient to deliver informative conclusions. For example, suppose we are interested in the causal effect of Z on Y. In an observational study, the data reveal only the joint distribution of Z and Y. Additional assumptions are required to use this joint distribution to identify the causal effect of Z on Y. This naturally raises further questions: Are the identification assumptions substantively meaningful? Do we have evidence to support them?

As we will see, especially in observational studies, identification assumptions are often difficult or impossible to test directly. This is why randomized experiments play a central role in causal inference: they provide a research design in which causal quantities can be identified under relatively weak and transparent assumptions.

Here, we focus on point identification. There are multiple notions of identification in the literature; see, for example, the discussion in Lewbel (2019).

Step 3: Estimation. Once the parameter is identified, how should it be estimated? In most settings, multiple estimators are available. Which one should we use? Not all estimators are equally appropriate. Ideally, we would like an estimator to possess desirable statistical properties, such as unbiasedness, consistency, and efficiency. In particular, obtaining an optimal estimator is not straightforward. Some modern estimation techniques, such as double machine learning, will be covered in the advanced CSS course.

Step 4: Inference. Statistical analysis inevitably involves uncertainty arising from randomness in the data or the research design. As a result, we must quantify this uncertainty. Common inferential tasks include constructing confidence intervals and conducting hypothesis tests.

Different perspectives exist regarding how uncertainty should be understood and quantified. For example, there is a long-standing debate between frequentist and Bayesian approaches to inference. Researchers also differ in their views on which sources of randomness should be accounted for. These differences give rise to distinct inferential paradigms, including sampling-based inference, design-based inference, and model-based inference.

In this class, we will learn all of these techniques. Have fun!

4 Different Linear Models

In empirical work, researchers often write the same regression equation $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + u$, but this equation can represent different underlying models. We will introduce four such models: the structural model, the (linear) conditional expectation function (CEF) model, the best linear projection (BLP) model, and the causal inference reduced-form model. These models differ in their parameters (estimands), interpretations, assumptions, and in the implications for the properties of estimators. The material is mainly drawn from Hansen (2022); Wooldridge (2010).

(Y, X) are random variables with a joint distribution F, which we refer to as the population. This population is assumed to be infinitely large. From F, we randomly draw n observations, forming a dataset (or sample) $\{(Y_i, X_i) : i = 1, ..., n\}$. In this lecture, we focus exclusively on the population. How to use data to learn about the population will be discussed in the next lecture.

Throughout, we assume $\mathbb{E}\left[Y^2\right] < \infty$ and $\mathbb{E}\,||X||^2 < \infty$, which imply that Y and X have finite means, variances, and covariances.

4.1 Conditional Expectation Function

Because the variables are random, we are mainly interested in the conditional expectation, which captures the systematic part of the relationship between Y and $X = (X_1, X_2, ..., X_k)^T$. The conditional expectation $\mathbb{E}\left[Y|X = x\right] = m(x)$ is a function of $x \in \mathbb{R}^k$, usually called regression function, or the regression of Y on X. It states that: “When X takes the value x then the average value of Y is $m(x)$.” Sometimes it is useful to view the conditional expectation function (CEF) as a function of the random variable X, and write it as $m(X)$ or $\mathbb{E}\left[Y|X\right]$.

The difference between Y and the CEF at X is the CEF error $r = Y - m(X)$. It has conditional expectation zero,

$$\mathbb{E}\left[e|X\right] = \mathbb{E}\left[Y - m(X)|X\right] = \mathbb{E}\left[Y|X\right] - \mathbb{E}\left[m(X)|X\right] = m(X) - m(X) = 0. \tag{1}$$

This is sometimes called a **conditional mean restriction** or **mean independence**, in the sense that the conditional mean of e is zero and thus independent of X. However, it does not imply that the distribution of e is independent of X.

Equation (1) also implies that $\mathbb{E}\left[e\right] = \mathbb{E}\left[\mathbb{E}\left[e|X\right]\right] = 0$. This result follows directly from the definition of $m(X)$ as the conditional expectation; it is not an additional assumption.

One important feature of the conditional expectation is that it is the best predictor of Y given X in the sense that it has the lowest mean squared error among all predictors. Suppose that, given a random vector X, we want to predict or forecast Y. Any predictor can be written as a function $g(X)$ of X.

Theorem 1. *If $\mathbb{E}\left[Y^2\right] < \infty$, then for any predictor $g(X)$,*

$$\mathbb{E}\left[(Y - g(X))^2\right] \geq \mathbb{E}\left[(Y - m(X))^2\right]$$

where $m(X) = \mathbb{E}\left[Y|X\right]$.

Proof. The LHS is

$$\begin{aligned}
\mathbb{E}\left[(Y-g(X))^2\right] &= \mathbb{E}\left[(e+m(X)-g(X))^2\right] \\
&= \mathbb{E}\left[e^2\right] + 2\mathbb{E}\left[e(m(X)-g(X))\right] + \mathbb{E}\left[(m(X)-g(X))^2\right] \\
&= \mathbb{E}\left[e^2\right] + \mathbb{E}\left[(m(X)-g(X))^2\right] \\
&\geq \mathbb{E}\left[e^2\right] \\
&= \mathbb{E}\left[(Y-m(X))^2\right]
\end{aligned}$$

□

Recall, the prediction error is $e = Y - m(X)$. By construction, this yields the formula

$$Y = m(X) + e \tag{2}$$

An important special case is when CEF is linear in x. We often add a constant 1 in the vector $X = (1, X_1, X_2, ..., X_k)^T$, so that $m(X) = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_k\beta_k = X'\beta$, where $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$.

In sum, the Linear CEF model is

$$Y = X'\beta + e, \quad \mathbb{E}\left[e|X\right] = 0$$

4.1.1 Marginal Effects

We are particularly interested in how changes in X_1 affect Y in expectation, holding other variables $X_2, ..., X_k$ fixed. This marginal effect is captured by the partial derivative of the CEF

$$\frac{\partial}{\partial x_1} m(x_1, ..., x_k)$$

For a linear CEF, the marginal effect of X_1 is simply β_1. What does this mean? Are marginal effects causal? Note that a CEF can be defined for any set of variables.

In general, marginal effects need not be causal. A marginal effect from the CEF only describes how Y changes with one variable while holding the other variables in the regression constant. This is the notion of **ceteris paribus**. Whether such an effect is causal depends on additional conditions. Intuitively, if we could truly hold all other relevant variables constant—including those not included in the regression—then the ceteris paribus comparison would have a causal interpretation. At this stage, however, the concept of causality remains vague because we have not yet formally defined a causal effect. We return to this issue later.

For example, suppose Y is income, X_1 is education level. In the regression model, we control for X_2 parents' education level. In the Linear CEF model $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + e$, β_1 is the marginal effect of education level holding parents' education level fixed. Is this effect causal? Generally, no. One concern is omitted variables, such as ability X_3: individuals with higher ability are more likely to obtain higher education and earn higher income. If ability is not controlled for, an observed association between education and income may reflect ability rather than the causal effect of education. Suppose ability were observable and included in the regression. Consider the expanded model $Y = \gamma_0 + X_1\gamma_1 + X_2\gamma_2 + X_3\gamma_3 + e$. In general, $\gamma_1 \neq \beta_1$ (see section 4.2.1). That is, once we hold both X_2 and X_3 constant, the marginal effect of education changes. But does γ_1 now represent a causal effect? Not necessarily, because there may still exist other unobserved confounding variables beyond ability.

In short, in the linear CEF model, we can interpret the parameter β_k only as the marginal effect of X_k holding the other variables included in the regression fixed. This marginal effect may not have a causal interpretation.

4.2 Best Linear Projection

The function form of CEF $m(X) = \mathbb{E}[Y|X]$ is typically unknown. However, we can always use a linear model $X'\beta$ to approximate it. How shall we determine β? A natural objective is to select β so that the prediction error is minimized.

We focus on the mean-squared prediction error

$$S(\beta) = \mathbb{E}[(Y - X'\beta)^2].$$

Then, β minimizes mean-squared prediction error $S(\beta)$ called **linear projection coefficient**, and such $X'\beta$ is the **best linear predictor** of Y given X.

Proposition 2. *Suppose $\mathbb{E}[XX']$ is positive definite. The unique Linear projection coefficient is $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$.*

Proof. $S(\beta) = \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta$. The FOC is

$$\nabla_\beta S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta = 0$$

$$\Rightarrow \mathbb{E}[XY] = \mathbb{E}[XX']\beta$$

Because $\mathbb{E}[XX']$ is invertible implied by positive definite (see remark 1), we obtain $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$.

We leave SOC for minimization in the remark 2. □

Remark 1. *$\mathbb{E}[XX']$ is positive definite means that for any non-zero $\alpha \in \mathbb{R}^k$, $\alpha'\mathbb{E}[XX']\alpha = \mathbb{E}[(\alpha'X)^2] > 0$. This implies that there is no non-zero vector α such that $\alpha'X = 0$. Therefore, X are linearly independent, and thus invertible.*

Remark 2. *In the optimization problem $\min S(\beta)$, we also need to check SOC. Suppose β is one-dimensional. Then for minimization problem, we need $\frac{\partial^2}{\partial\beta^2}S(\beta) > 0$. When β is a vector, what we need is positive definite. To see why, consider the Taylor expansion around optimal β^*, roughly speaking, $S(\beta^* + h) \approx S(\beta^*) + h'\nabla_\beta S(\beta^*) + \frac{1}{2}h'\nabla^2_\beta S(\beta^*)h$, where $\nabla_\beta S(\beta^*) = 0$ by FOC, and $h'\nabla^2_\beta S(\beta^*)h > 0$ by positive definite. Therefore, any deviation from optimal β^* will increase the value. We conclude that positive definite implies that β^* is the minimizer.*

Similar to the prediction error in the previous section, we can also define the projection error as

$$e = Y - X'\beta$$

Note that, projection error is equivalent to the error from CEF (2) when and only when conditional expectation $m(X)$ is linear.

What are the properties of the projection error? One important property is $\mathbb{E}[Xe] = 0$. You will prove it in the homework. This is true for all X_k, $\mathbb{E}[X_k e] = 0$. When the regressor vector contains a constant, say $X_1 = 1$, then $\mathbb{E}[e] = 0$. Therefore, the projection error has mean zero when the regressors include a constant. This further implies that $Cov(X, e) = \mathbb{E}[Xe] - \mathbb{E}[X]\mathbb{E}[e] = 0$. In other words, the projection error has zero mean and is uncorrelated with every regressor.

Remark 3. *$E[e|X] = 0$ is stronger than $\mathbb{E}[Xe] = 0$.*

In sum, we define the Linear Projection Model:

$$Y = X'\beta + e, \quad \mathbb{E}[Xe] = 0, \quad \beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$$

It is useful to understand $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$ in the simple linear regression model with one regressor and one intercept.

Example 1. *Consider the simple linear BLP model $Y = \beta_0 + \beta_1 X + e$. In the homework, you will solve for $\beta_1 = \frac{Cov(X_1,Y)}{Var(X_1)}$. Therefore, β_1 is larger when X_1 and Y have a larger covariance, and when X_1 has a smaller variance.*

The BLP is often used as a working model, as we will see later. Clearly, there is no causal interpretation here; it is simply a linear approximation.

4.2.1 Omitted Variable Bias

We show that, given a BLP model, omitting a variable generally changes the projection coefficient. This difference is known as population omitted variable bias (OVB). The difference is known as population omitted variable bias (OVB). It is the consequence of omission of a relevant correlated variable. The term OVB has been used loosely in many areas. Here we emphasize that it refers specifically to a change in the underlying parameter induced by variable omission.

Consider the long BLP model

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

and the short model in which we omit X_2

$$Y = X_1'\gamma_1 + u$$

Here, X_1 and X_2 are vectors. In general, $\beta_1 \neq \gamma_1$, except in special cases. To see this, we calculate

$$\begin{aligned}
\gamma_1 &= (\mathbb{E}[X_1X_1'])^{-1}\mathbb{E}[X_1Y] \\
&= (\mathbb{E}[X_1X_1'])^{-1}\mathbb{E}[X_1(X_1'\beta_1 + X_2'\beta_2 + e)] \\
&= \beta_1 + (\mathbb{E}[X_1X_1'])^{-1}\mathbb{E}[X_1X_2]\beta_2 \\
&= \beta_1 + \Gamma_{12}\beta_2
\end{aligned}$$

What is Γ_{12}? It is the coefficient from the projection of X_2 on X_1 ($X_2 = X_1'\Gamma_{12} + \xi$).

Observe that $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2$ is different from β_1 unless $\Gamma_{12} = 0$ (X_1 and X_2 are uncorrelated) or $\beta_2 = 0$ (X_2 is not correlates with Y).

Unfortunately, the above simple characterization of omitted variable bias does not immediately carry over to more complicated setting.

4.3 Structural Model

When people say that $Y = X'\beta + e$ is a **structural model**, they are claiming that the model represents a causal relationship, as opposed to one that merely captures statistical associations. Such a model may be derived from formal theory—such as spatial voting models or principal–agent theory—or it may be motivated by informal reasoning.

An explanatory variable X_j is said to be **endogenous** if it is correlated with e (more precisely, $\mathbb{E}[Xe] \neq 0$). In traditional usage, a variable is endogenous if it is determined within the context of a model. The usage in regression, has evolved to describe any situation where an explanatory variable is correlated with the disturbance e.

If X_j is uncorrelated with e, then it is **exogenous**. Exogeneity is a strong assumption. It is a property of random variables relative to parameters of interest. Hence a variable may be validly treated as exogenous in one structural model but not in another. It may be justified as being a consequence of a natural experiment or a quasi-experiment in which the value of the variable is determined by an external intervention; For example, a government or regulatory authority may set a tax rate or a policy parameter.

We will see in later lectures that endogeneity implies that the (least squares) estimator $\hat{\beta}$ does not converge to the structure parameter as the sample size goes to infinity. In other words, it is inconsistent for the structural parameter: $\hat{\beta} \to \beta^* \neq \beta$. The inconsistency of least squares is typically referred to as endogeneity bias or estimation bias due to endogeneity.

Note, by definition, the CEF model ($\mathbb{E}[e|X] = 0$) and the BLP ($\mathbb{E}[eX] = 0$) does not suffer from an endogeneity problem. Therefore, endogeneity problem arises only in structural models.

4.3.1 Source of Endogeneity

The major source of endogeneity arises from omitted variables. Consider the canonical example in which Y denotes income, X_1 is a measure of education, and X_2 represents ability. We specify the structural model as $\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Because this is assumed to be a structural CEF, it implies that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e, \quad \mathbb{E}[e|X_1, X_2] = 0 \tag{3}$$

where e is the structural error. The parameter β_1 is the structural parameter of interest, as it has a causal interpretation by definition. Without loss of generality, assume $\mathbb{E}[X_2] = 0$, since an intercept is included in the model.

However, X_2 is unobserved. As a result, we can only work with the projection model

$$Y = \beta_0 + \beta_1 X_1 + u \tag{4}$$

where $u = \beta_2 X_2 + e$.

It is clear that $\mathbb{E}[X_1 u] = \mathbb{E}[X_1(\beta_2 X_2 + e)] = \beta_2 \mathbb{E}[X_1 X_2]$. If $\beta_2 \neq 0$ (ability X_2 has a non-zero effect on income Y) or $\mathbb{E}[X_1 X_2] = Cov[X_1, X_2] \neq 0$ (ability X_2 and education X_1 are correlated), then an endogeneity problem arises. We may expect $\beta_2 > 0$ and $\mathbb{E}[X_1 X_2] > 0$. Then, it implies that projection coefficient β_1 in (4) will be upward biased relative to the structural coefficient β_1 in (3). Thus least squares (which is estimating the projection coefficient) will tend to over-estimate the causal effect of education on wages. (I hope you have seen the connection with section 4.2.1.)

Other sources of endogeneity—such as measurement error and simultaneous equations bias—will be introduced in later lectures.

4.3.2 Identification

Whether β is identified is not an easy question. We will encounter many different identification assumptions for structural models. Today, we study the first case, in which there is no endogeneity. We will discuss identification with endogeneity in later lectures. Recall that X is a $k \times 1$ column vector.

The first assumption is orthogonality. You should recognize that this assumption is the same as the defining property of the projection error e in the BLP.

Assumption 3 (Population Orthogonality)**.** $\mathbb{E}[Xe] = 0$

Because X contains a constant, this condition is equivalent to saying that e has mean zero and is uncorrelated with each regressor. One sufficient condition is $\mathbb{E}[e \mid X] = 0$. For example, if Y is income and X is education level, this assumption implies that education is uncorrelated with ability, which is absorbed into e.

Given this, you might expect that β is identified as $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$. However, as we recall from the proof of Proposition 2, we also require $\mathbb{E}[XX']$ to be invertible. This leads to the second assumption, full rank.

Assumption 4 (Rank)**.** $\mathbb{E}[XX'] = k$

This assumption fails if and only if at least one regressor can be written as a linear function of the other regressors (in the population).

Now recall that $Y = X'\beta + e$. Premultiplying both sides by X and taking expectations yields

$$\mathbb{E}[XY] = \mathbb{E}[XX']\beta + \mathbb{E}[Xe]$$

The last term satisfies $\mathbb{E}[Xe] = 0$, and thus we can invert $\mathbb{E}[XX']$ to obtain $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$.

4.4 Causal Inference (reduced-form) Model

Causal parameters based on counterfactuals provide statistically meaningful and operational definitions of causality that in some respects differ from the traditional structural model.

We introduce the potential outcomes framework to define causal effects. Let $Z_i \in \{0,1\}$ denote the binary treatment variable for individual i. Let $Z = (Z_1, \ldots, Z_n)$ denote the treatment assignment vector for the n individuals in the sample. The potential outcome for individual i is denoted by $Y_i(Z)$.

Assumption 5 (No Interference)**.** *Individual i's potential outcomes do not dependent on other individual's treatments.*

Therefore, we can simplify $Y_i(Z)$ to $Y_i(Z_i)$. In other words, for each individual i, there exists two potential outcomes: $(Y_i(1), Y_i(0))$.

Assumption 6 (Consistency)**.** *There are no different forms or versions of each treatment level, which lead to different potential outcomes.*

Therefore, the observed outcome is $Y_i = Z_iY_i(1) + (1 - Z_i)Y_i(0)$. If $Z_i = 1$, we observe $Y_i(1)$; if $Z_i = 0$, we observe $Y_i(0)$.

People often call Stable Unit Treatment Value Assumption (**STUVA**) assumption by combing no interference and consistency assumptions.

The individual treatment effect is $\tau_i = Y_i(1) - Y_i(0)$. In a single realized world, we can observe only one potential outcome for each individual. This is the fundamental problem of causal inference. Most of the time, we are interested in the average treatment (ATE): $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. This is a causal quantity under the potential outcome framework.

Under what conditions can we identify the ATE? We introduce some sufficient conditions.

Assumption 7 (Ignorability)**.** $Y(z) \perp Z|X$ *for* $z = 0, 1$

Remark 4. *This assumption has many names: unconfoundedness, selection on observables, conditional independence.*

This assumption emphasizes experimental, quasi-experimental, or natural-experimental randomization as a key source of identification.

Assumption 8 (Positivity)**.** $0 < \mathbb{P}[Z = 1|X] < 1$

Proposition 9. *Suppose STUVA holds. Under assumption 7 and 8, ATE is identified as* $\tau = \mathbb{E}\left[\mathbb{E}\left[Y_i|Z_i = 1, X_i\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i|Z_i = 0, X_i\right]\right]$.

Proof.

$$\begin{aligned}
\mathbb{E}[Y_i(1)] &= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i]] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1)|X_i, Z_i = 1]] \quad \text{ignorability and positivity} \\
&= \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 1]] \quad \text{STUVA}
\end{aligned}$$

By analogous, $\mathbb{E}[Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 0]]$.

Therefore, $\tau = \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 1]] - \mathbb{E}[\mathbb{E}[Y_i|X_i, Z_i = 0]]$.

□

Remark 5. *Note that the outer expectation is taken with respect to the marginal distribution of* X*, not the conditional distribution of* $X|Z = 1$.

It means that, to identify causal effects, we only need conditional expectations. We do not need a linear regression function at all. If X is discrete, we can calculate $\mathbb{E}[Y \mid Z = 1, X = x]$ for each x. However, if X is continuous, we need to rely on a regression function. For example, people often specify $\mathbb{E}[Y \mid Z = 1, X] = m(X) = X'\beta$. From here, we connect to regression techniques. We will return to this topic in later lectures.

References

Hansen, B. (2022). *Econometrics.* Princeton University Press.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4), 835–903.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.