Chapter 3:

Statistical and Econometric Foundations of Causal Economics

3.1 Introduction to Statistical Foundations of Causality

Statistics is a very important part of causal reasoning since it gives formal mechanisms to measure the relationship, estimate uncertainty, and examine evidence in economic analysis. Whereas descriptive statistics summarize observed statistics, predictive statistics are concerned with predicting future developments and causal statistics with cause-effect relationships explaining how and why economic developments change with interventions of different kinds. This means that correlation does not suffice economic decisions, which are statistically associated variables due to the possibility of confounding variables, reverse causation, or spurious relationship, which can make wrong policy inferences [41]. In recognition of these shortcomings, the discipline has developed out of classical statistical tools based on association and hypothesis testing to contemporary causal inference models that explicitly reflect assumptions, counterfactual logic and identification procedures to be able to make plausible causal assertions in economics.

3.2 Probability Theory and Causal Interpretation

Causal inference in economics is based on probability theory to provide a strict mathematical framework used to measure uncertainty and model the fluctuation of economic results amongst either individuals, firms, or markets. Fundamentally, probability theory brings on board the concept of random variables, which are economic quantities whose value is not deterministic but may have various outcomes with respect to some probability distribution [42]. The discrete and continuous forms of this distributions enable the researcher to specify the probability of different events to occur as well as to calculate the expectations, variances, and other measures of central tendencies, and dispersion of economic events [43]. In making uncertainty formal in such a manner, probability theory offers the language and the tools to be able to get off the simple observation of correlation that can be made to a more accurate view of causation relations. What is especially important to causal reasoning is conditional probability, a measure of the probability of an event given that another event has occurred, and conditional expectation, a measure of the expected value of a variable, given that a sequence of explanatory events has occurred. An example of this is when testing the impact of a policy intervention on household consumption or labor supply, conditional probability will allow economists to control the impact of the treatment by other factors that may also have the same effect, thus giving an approximation to the concept of a counterfactual situation in which the intervention does not take place [44]. It is also vital that there is a difference between independence and conditional independence. Though seeming to be correlated in isolation, by considering confounding factors, two variables that seem to be correlated may turn out to be actually not correlated after conditioning confounding factors is done, hence the need to condition in causal inferences. This is the core difference between the contemporary econometric techniques, such as matching, regression adjustment, and instrumental variables, all of which are based on the assumptions of conditional independence to determine the unbiased effects of causality. The Bayesian theory also advances the analysis of causality with a methodical approach to updating beliefs on the basis of the novel information. In economics, it enables researchers to improve estimates of causal effects as more data is received and combine previous knowledge in conjunction with the observed data in order to come up with probabilistic statements regarding the cause-effect relationships. As an example, Bayesian techniques can be applied by policymakers to revise the forecasts of the effects of a tax reform on investment behavior as additional quarters of economic data become known, and better information is available to aid decision-making under uncertainty. These probabilistic ideas are the basis of the concept of probabilistic causation, a concept that is inherently different than deterministic causation by admitting that causes do not always result in a given outcome with certainty but instead tend to increase or decrease the probability of that outcome or its distribution. Applied to the economy, this is a manifestation of the complexity and heterogeneity of human behavior, market interactions and institutional environments, where the same intervention can produce diverse outcomes in different people or different places at different times. More complex causal models, including structural equation modeling, potential outcomes analysis and graphical models, are also based on probability theory, in which individuals can express their causal assumptions and causal expectations in terms of random variables, conditional probabilities and expectations [45]. With these mathematical instruments combined with empirical data, economists can no longer achieve the tasks of description and prediction, but they can build models that give some approximation to what they believe are the causal processes that underlie the observed phenomena. Also, the sensitivity analysis, robustness, and quantification of uncertainty surrounding the causal estimates, which are essential to academic research and policy evaluation, are easier to perform through probabilistic reasoning. In general, probability theory allows entrepreneurs to formulate, test, and develop causal hypotheses in a rigorous and systematic way, closing the gap between abstract economic theory and the results of research, and having become the mathematical foundation of causal inference in economics today. As Figure 3.1 shows, probability theory can be used to cause a causal interpretation, which points the assumptions and counterfactual reasoning as well as the identification strategies used to establish credible cause-effect relationships in economics.

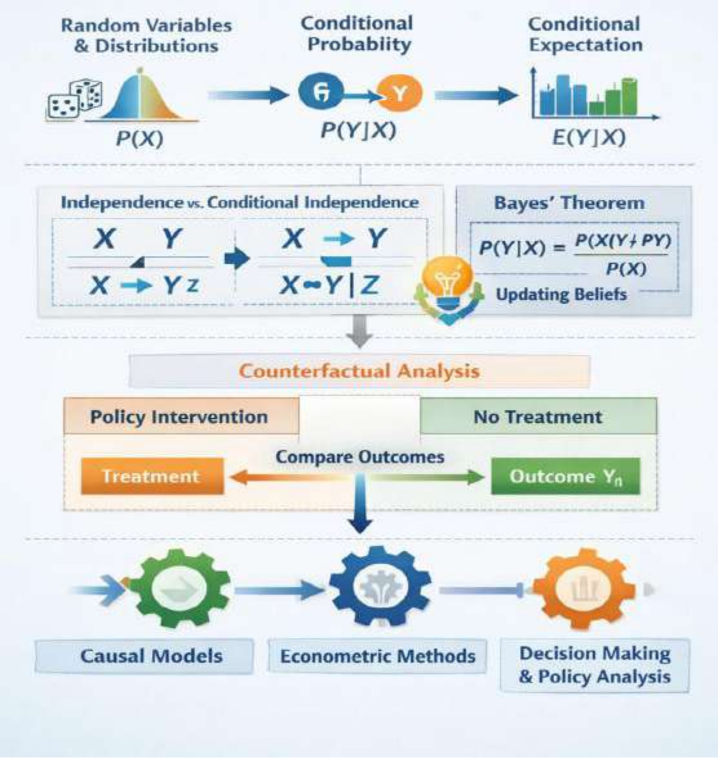

Figure 3.1: Probability Theory and Causal Interpretation

3.3 Classical Statistical Inference and Its Limitations

The foundation of applying empirical economics uses classical statistical inference based on estimation, hypothesis testing, and confidence interval to make conclusions with respect to sample data. In this context, Null Hypothesis Significance Testing (NHST) is often referred to as a method of testing the significance of observed relationships against zero, whereas Type I and Type II errors are the measures of risks of false positives and false negatives, respectively [46]. These tools are, however, useful in evaluating the statistical association, although they are not sufficient to make causal assertions since they fail to deal with problems of endogeneity, confounding or reverse causality. P-values in the economics literature are usually misunderstood to have a causal strength or policy relevance, although the probability of data to occur under a null hypothesis is all that they reflect [47]. Consequently, without the explicit inclusion of causal assumptions and identification strategies, classical inference may be used to come to overconfident or misleading inferences. After giving the key concepts and limitations of Classical Statistical Inference, table 3.1 will demonstrate it.

Table 3.1: Key Concepts and Limitations of Classical Statistical Inference

Concept	Description	Limitation for Causal Analysis
Estimation	Quantifies relationships between variables using sample data	Estimates associations, not necessarily causal effects
Hypothesis Testing	Evaluates whether effects differ statistically from zero	Does not establish cause–effect relationships
Confidence Intervals	Provide a range of plausible parameter values	Reflect sampling uncertainty, not causal validity
NHST	Tests null hypotheses using p-values	Encourages binary thinking without causal context
Type I Error	Rejecting a true null hypothesis	May falsely suggest a causal effect
Type II Error	Failing to reject a false null hypothesis	May overlook meaningful causal relationships
p-value Interpretation	Measures statistical significance	Often misused as evidence of economic or causal importance

3.4 Regression Analysis as a Statistical Tool

One of the most basic and most commonly used statistical instruments in economics is regression analysis which provides a conceptual framework of analyzing relationships among variables and quantifying the network of changes in a dependent variable in relation to changes in one or more of the explanatory variables. Simple linear regression is used to model the relationship between one independent variable and a dependent variable in its simplest case, which can be simply interpreted as the slope coefficient indicating the amount of change in the dependent variable that is expected in case of a one-unit change in the independent variable [48]. The multiple linear regression is an extension of this framework that allows incorporating multiple explanatory variables at the same time, which allows researchers to explain the impact of other factors that are important and isolate the partial effect of each variable when other variables are held constant. This characteristic is especially critical in economic analysis where results like consumption, supply of labor or investment behavior are usually determined by several different related factors. The validity and consistency of regression estimate however, is highly subject to assumptions of the Classical Linear Regression Model (CLRM). These are linearity that gives the relationship between an independent and a dependent variable a linear form of a combination of coefficients, exogeneity that requires that the explanatory variables and the error term are not correlated, homoscedasticity that requires the errors to have equal variances across observations and the non-existence of multicollinearity which assumes that the explanatory variables are not correlated amongst themselves. In order to assess the sufficiency of those assumptions, as well as to measure the overall goodness of the model, economists use the goodness-of-fit indicators like the coefficient of determination (R^2) and probing of residuals based on a sequence of behavior that could reflect the misspecification of a model, heteroscedasticity, and other breaches. Although regression analysis can be widely used and easy to conceptualize, it does not imply causality. When the assumptions like exogeneity have been disobeyed because of omitted variable bias, measurement error, simultaneous or unobserved confounding, the estimated coefficients can be biased and inconsistent causing misleading interpretations [49]. As an illustration, the spurious result of the effect of income on consumption could be obtained because of the failure to consider an unobservable factor that influences both income and consumption. Consequently, although regression can be a great method of relation description and making predictions, believable causal inference demands other methods, such as instrumental variables, fixed effects, difference-in-differences, or other identification methods, to attain the effects as true causal actions, but not a simple statistical correlation. In this regard, the regression analysis is a fundamental tool, which has to be supplemented with the intensive use of econometric analysis in order to be able to draw credible causes and effects in economic studies. It can be represented as follows

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon}$$

Where:

- **Y** = dependent variable (outcome of interest)
- $\boldsymbol{\beta_0}$ = intercept (constant term)
- $\boldsymbol{\beta_1, \beta_2, ..., \beta_k}$ = regression coefficients (partial effects of each explanatory variable)
- $\mathbf{X_1, X_2, ..., X_k}$ = independent variables (explanatory variables)
- $\boldsymbol{\varepsilon}$ = error term (captures unobserved factors)
- **k** = number of explanatory variables

Key interpretation: Each coefficient $\boldsymbol{\beta_j}$ represents the expected change in Y for a one-unit change in X_j, holding all other variables constant (ceteris paribus).

3.5 Econometric Foundations of Causal Analysis

The approach, which gives cause-and-effect analysis to the study of economics, is econometrics, which seeks to estimate and evaluate economic theories using empirical data. One of the major distinctions of both econometric models is the idea that structural models are based on economic theory and represent explicitly the causal mechanisms, whereas reduced-form models simply describe the empirical relationship without explicitly describing the underlying causal structure [50]. One of the most significant problems in the causal econometric analysis is endogeneity, caused by the correlation of explanatory variables with the error term because of omitted variable bias, measurement error, or because cause and effect happen simultaneously. Such endogeneity sources cause biased and inconsistent parameter estimates, which negatively affect causal interpretation [51]. As a result, the question of identification to establish the possibility of unique causal inference of observed data is at the center of econometrics and the reason to use specialized tools and assumptions in order to obtain plausible causal inference. Table 3.2 depicts Econometric Foundations of Causal Analysis

Table 3.2: Econometric Foundations of Causal Analysis

Concept	Description	Implication for Causal Inference
Econometrics	Application of statistical methods to economic data to test theories	Enables empirical evaluation of causal relationships
Structural Models	Theory-driven models that specify causal mechanisms among variables	Allow direct interpretation of causal parameters
Reduced-Form Models	Empirical models capturing relationships without explicit causal structure	Useful for prediction but limited for causal explanation
Endogeneity	Correlation between explanatory variables and the error term	Leads to biased and inconsistent estimates
Omitted Variable Bias	Excluding relevant variables correlated with regressors	Produces spurious causal effects
Measurement Error	Inaccurate measurement of explanatory variables	Attenuates or biases estimated coefficients
Simultaneity	Mutual causation between dependent and independent variables	Obscures direction of causality

Identification Problem	Difficulty in isolating true causal effects from observed data	Necessitates assumptions and specialized econometric techniques

3.6 Causal Assumptions in Econometric Models

Econometric models are highly dependent on a series of assumptions, which support causal inference that researchers can identify estimated relationships as causal effects and not as associations. Strict exogeneity and Exogeneity are critical in order to ascertain that the explanatory variables are not correlated with the error term, and that the feedback effect and omitted variable bias are excluded [52]. Conditional Independence Assumption (CIA) is an assumption that treatment assignment is independent of potential outcomes, conditional on observed covariates, which is the basis of matching and selection-on-observables methods. The Stable Unit Treatment Value Assumption (SUTVA) states that the treatment administered to a unit has no impact on other units and treatments are always determined [53]. Furthermore, the common support and overlap conditions also presuppose that there is enough similarity between treated and control groups, which can be compared. Collectively, these assumptions are of central importance in causal identification since their breaches can null causal interpretations despite the use of advanced econometric methods. Table 3.3 shows causal Assumptions in Econometric Models.

Table 3.3: Causal Assumptions in Econometric Models

Assumption	**Description**	**Role in Causal Identification**	**Consequence if** violated
Exogeneity	Explanatory variables are uncorrelated with the error term	Ensures unbiased and consistent estimates	Leads to endogeneity and biased estimates
Strict Exogeneity	No feedback from past, present, or future errors to regressors	Required for panel data consistency	Dynamic bias in estimates
Conditional Independence Assumption (CIA)	Treatment is independent of potential outcomes given covariates	Enables matching and propensity score methods	Selection bias persists

SUTVA	No interference between units and well-defined treatments	Ensures valid comparison of potential outcomes	Spillover and contamination effects
Common Support	Overlap in covariate distributions across groups	Allows credible counterfactual comparison	Extrapolation beyond data
Overlap Condition	Positive probability of treatment for all covariate values	Prevents perfect prediction of treatment	Unidentifiable treatment effects

3.7 Instrumental Variables and Identification

The application of Instrumental Variables (IV) methods in econometric analysis measures causal effects, but when endogeneity cannot be estimated consistently in ordinary least squares. An instrument is a variable that is correlated with the endogenous explanatory variable (condition of relevance) and only influences the outcome indirectly, via that explanatory variable and not directly (exclusion restriction). Under these assumptions, IV methods isolate exogenous variation which can be utilized in the recovering of causal effects [35]. One major problem of IV estimation is the weak instruments problem where the instrument has a weak relationship with the endogenous variable thus giving biased and imprecise estimates. The most often IV estimation is Two-Stage Least Squares (2SLS), in which the first stage is used to predict the endogenous variable on the basis of the instrument, and the second stage is used to estimate the causal effect based on the predicted values. IV estimates are usually construed as local average treatment effects (LATE) which captures the causal effect on the subpopulation on which the behavior is impacted by the instrument [55]. Instrumental Variables and identification is shown in table 3.4.

Table 3.4: Instrumental Variables and Identification

Concept	Description	Role in Causal Identification	Limitation
Instrument	Variable correlated with endogenous regressor	Provides exogenous variation for identification	Must satisfy strict conditions
Relevance Condition	Instrument significantly affects the endogenous variable	Ensures meaningful first-stage estimation	Weak correlation reduces reliability
Exclusion Restriction	Instrument affects outcome only through the endogenous variable	Guarantees causal interpretation	Difficult to test empirically
Weak Instruments	Instruments with low explanatory power in first stage	Causes biased and unstable IV estimates	Leads to large standard errors
Two-Stage Least Squares (2SLS)	Two-step estimation procedure using instruments	Produces consistent causal estimates	Sensitive to instrument validity
IV Estimates (LATE)	Effect for compliers influenced by the instrument	Clarifies scope of causal inference	Limited external validity

3.8 Panel Data Methods for Causal Inference

The techniques of causal inference in economics cannot do without the panel data methods as they merge the advantages of cross-sectional and time-series data, enabling the researcher to observe the same economic unit, be it individuals, firms, or regions, across the periods. This design is less biased than gross cross-sectional data and allows analysts to accommodate unobserved heterogeneity, detect dynamic relationships, and estimate causal effect with a higher degree of precision. The fixed effects model is one of the key methodologies of panel data analysis and it attempts to manage unobserved and time-invariant attributes, which might be correlated with explanatory variables e.g. innate product differences between firms or inherent preferences between individuals [56]. Fixed effects models allow differencing out these invariant factors thereby minimizing the omitted variable bias and increasing the validity of the causal estimates. Contrary to this, random effects models assume that the unobserved heterogeneity is independent of the explanatory variables which are more efficient to estimate, but highly demands a strong exogeneity assumption. The decision of whether to use a fixed or a random effect is usually informed by the statistical tests like the Hausman test, and theoretically through the consideration of the probable correlation between unobserved factors and regressors. In addition to these classical panel models, another quasi-experimental design, the Difference-in-Differences (DiD) framework has emerged as a popular method of establishing causation of policy interventions, reforms or shocks. DiD is a method used to compare the outcome change in time between the treated and control groups, thus, eliminating confounding factors that remain the same between the groups or through time. Assumption on parallel trend is a critical factor in estimating DiD because in non-treated conditions, it is expected that both groups would have taken similar paths [57]. Breaks in this assumption like the underlying differences in trends can bias the estimated treatment effects and so pre-testing and robustness checks should be carefully considered. Fixed effects, random effects, and DiD are commonly used in the labor economics, health economics, and annual policy, and development research to evaluate the causal effects of programs, regulation, or macroeconomic shock. These techniques have the benefit of enabling economists to make more plausible inferences regarding the impact of interventions and produce policy recommendations based on empirical data as opposed to mere associations by relying on repeated observations and addressing unobserved heterogeneity [58]. Overall, panel data methods contribute to the strength and credibility of causal analysis through a combination of the 1-2 effects of time and cross-section along with confounding variables combined with a structure of systematic analysis to apply relations of dynamism across time and spatial economic relationships. Figure 3.2 shows how the methods used in panel data allow one to make a cause-effect inference by following the same units (e.g. people, firms, or regions) across time.

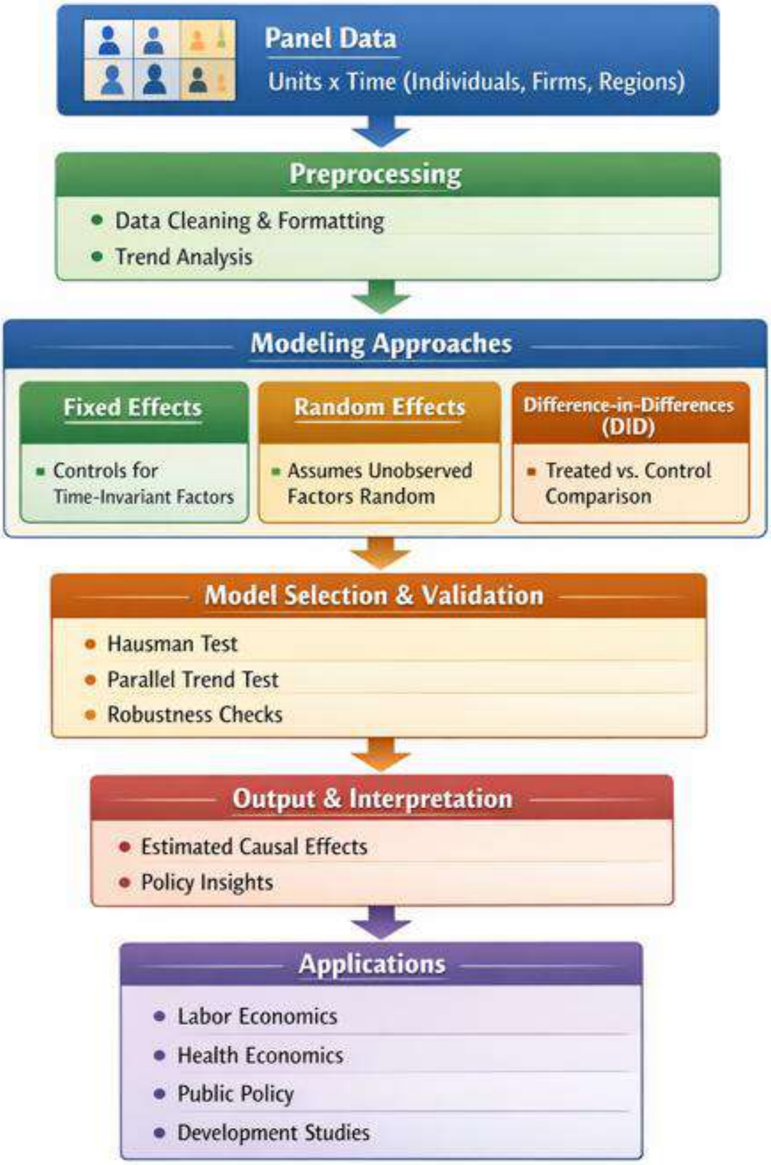

Figure 3.2: Panel Data Methods for Causal Inference

3.9 Causal Inference Using Quasi-Experimental Designs

The quasi-experimental designs have established a set of valuable instruments that are invaluable in the field of economics to estimate a causal influence in cases in which random controlled experiments are not practical or unethical or impossible. The designs capitalize on existing differences or policy manipulations which are close to the conditions of a randomized motivation enabling a researcher to make causal inferences based on observational data. One of the main methods within this framework is that of natural experiments which are exogenous shocks, institutional changes, or rules of regulation that create variation in treatment assignment that is effectively random to the viewpoint of the population that is impacted. As an illustration, unexpected policy changes, natural catastrophes or administrative limits can place individuals or areas in a position where certain people or regions receive a specific treatment, and similar units do not receive one, which is an adequate foundation of causal study. Another common quasi-experimental design called Regression Discontinuity Design (RDD) estimates causal effects based on the comparison of units just above and just below an established cutoff or threshold and assumes that these units are similar enough, except regarding their status with regard to the treatment [59]. The strong point of RDD is that it generates very plausible local causal inferences, but lacks much external validity to observations beyond the cutoff. Alternative methods, such as propensity score matching, offer another method, which creates a counterfactual group and approximates the treated group with regard to observed covariates. Such methods minimize selection bias by matching treated and untreated groups with comparable features, and are able to approximate the conditions of a randomized study, but cannot control unobserved confounders. Although they have their benefits, quasi-experimental designs are demanding to carry out and assume validity. The quality and granularity of data, whether in the proper specification of models, and comprehensive testing of identification assumptions, e.g. the continuity of covariates in RDD or the sufficiency of covariate balance in matching, determine the credibility of findings. Moreover, although these methods give practical information that can be directly applicable to policy assessment, their estimates may be limited in the external validity due to the context or the population to be studied [60]. In general, quasi-experimental designs can be seen as a compromise between a simple observational investigation and a randomized experiment that provides the researcher and policy-makers with the means to point to the estimate of the effects of causation in a real economic environment without disregarding the significance of rigorous design, testing of assumptions, and consideration of implications that can derive credible and actionable conclusions.

3.10 Structural Econometric Models and Causality

Structural econometric models present a more rigorous structure of the explanation of causal relationship in economics as they directly combine economic theory and empirical analysis. Structural models, unlike purely statistical or reduced-form models, are developed to represent the underlying behavioral processes which produce observed data, and therefore to enable the researcher to interpret estimated parameters as structural effects (and not simply correlations). Structural equations, a system of which is used to formalize these models, are used to describe how endogenous variables are mutually determined in the presence of other endogenous and exogenous variables, which represent interdependence and feedback mechanisms of the economic systems. One of the inherent issues with the application of structural models is the identification of structural parameters, which necessitates placing enough theoretical restriction, exclusion or exploiting instrumental variables to isolate true causal effects to those caused by confounding or simultaneity. Identification is a property that is important to ensure that the estimated coefficients are unique results of the postulated causal process of an economic theory, rather than arbitrary results of the data-generating process. In particular, important structural models are Simultaneous equation models, which are useful when several endogenous variables are interdependent, e.g. supply and demand in a market, investment and income in a macroeconomic model [61]. Simultaneous equation frameworks explain feedback loops by jointly modelling these relationships and eliminates the biases that would otherwise be experienced when a single equation is attempted to be estimated independently using ordinary least squares. The other important aspect of structural econometric modeling is that structural econometric modeling is based on theory-driven assumptions that direct the specification, estimation and interpretation. Founding empirical analysis on a properly developed economic theory contributes to the believability of causal inferences by giving a logical explanation of the connections which are estimated in terms of causation and also by making sure that the estimates do not contradict the well-established economic principles. The structural models also enable counterfactual and policy analysis because they enable the economist to imitate the impact of hypothetical interventions or policy alterations by manipulating the explanatory variables in the theoretical structure. Structural models fill in the divide between theory and data, combining strict methods of econometric analysis with theoretical advice to present solid and readable estimates on the effects of causal relationships [62]. Finally, such a strategy enhances the validity and policy applicability of empirical research, and structural econometric modeling is one of the foundations of causal analysis in applied and theoretical economics.

3.11 Linking Statistical Foundations to Causal Economics

The connection of statistical grounds to causal economics is core to the realization of the application of empirical analysis in informing the decision-making and policy testing based on theories. Statistical and econometric methods give the necessary framework in terms of the quantification of relations between variables, uncertainty assessment, and controlling confounding variables that may mask real causal effects. Classical statistical techniques, such as descriptive statistics, correlations, simple regression analysis provide convenient measures of association but cannot be considered to establish causality because of the possibility of endogeneity, omitted variable bias and unobserved heterogeneity. Econometric methods fill this gap by introducing formal identification methods that solve these problems and enable the researcher to go beyond correlation to believable causal inference. Exogenous variation around leverage leverages Instrumental variables are used to identify causal effects in correlated regressors, whereas panel data models leverage repeated time series variation, which is used to account for the unobserved heterogeneity of individuals, firms, or regions [62]. Quasi-experimental designs like natural experiments, regression discontinuity, and difference-in-differences techniques go further to enhance causal inference because they replicate the conditions of randomized experiments using observational data. The combination of these approaches with counterfactual models gives a researcher an opportunity to explicitly model what-if scenarios, aiding in the estimation of possible outcomes conditioned on alternative interventions or policy regimes. Counterfactual reasoning enables economists to answer the question of how the outcome would have differed given some change of variables or policies and this is a systematic way of measuring the causal effects. It is a combination of statistical rigor, econometric identification, and counterfactual modeling that guarantees that causal estimates are internally valid, as well as informative to real-world decision-making. In addition, connecting statistical grounds with causal models helps to increase the readability and believability of empirical findings, since it connects numerical approximations with theoretical predictions and economic processes. This combination is essential to policymakers and decision scientists: they can have solid evidence on which to draw their interventions, measure the effectiveness of their programs, and develop the strategies that are both evidence-based and cause-oriented [63]. With mathematical rigor of the probability theory, the empirical strength of the econometrics, and the intuitive clarity of the counterfactual reasoning, researchers can develop models that are able to reflect the complexity of economic behavior and yet provide actionable insights. In this regard, the connection between causal economics and statistical origins is not only methodological but also foundational as the way between observed facts, economic theory, and informed policy choices which can have an impact on the economic outcomes in practice.

3.12 Summary and Chapter Implications

Chapter 3 has also focused on the statistical and econometric underpinnings of a causal economic analysis with particular emphasis placed on the shift of descriptive and predictive approaches to frameworks that can determine relationships between causes and effects. Among these are the significance of assumptions (exogeneity, conditional independence, and SUTVA); the difficulties of endogeneity and identification, and the application of sophisticated methods (instrumental variables, panel data models, and quasi-experimental designs) to estimate plausible causal effects. These pillars bear great implications to empirical economic studies, as they offer the needed methodological rigor to design a study, interpret findings, and make a policy decision.

This chapter sets the stage of Chapter 4 where causal modeling using observational data will be discussed in detail.