Econometrics

In this lecture we will cover some basic characteristics of econometrics, what it is, how it differs from “data science” or pure statistics, and I will make an argument for why econometricians are the original data scientists.

What better way to define a field than by its criticisms. Here are a few critiques of econometrics[1]:

- I invite the reader to try to identify a meaningful hypothesis about economic behavior that has fallen into disrepute because of a formal statistical test ~ Larry summers, 1991
- Very little of what economists will tell you they know, and almost none of the content of elementary texts, has been discovered by running regressions. The support econometrics might give to a theory is weak, inconclusive, and easily countered by someone else’s regression ~ Bergmann, 1986
- [econometrics is] The art of drawing a crooked line from an unproved assumption to a forgone conclusion ~ General critics paraphrasing

You read these, and soon you question why this field even exists. There doesn’t seem to be much going for it. But these criticisms are naive, and short-sighted. Their observations rest on a unique limitation in economics: bad data. We cannot run experiments, so we cannot curate clean datasets, and we cannot isolate all of the possible effects for any given event. Very true, but quickly changing since the advent of cheap compute power and big data.

You all have had, I hope, first-hand experience with lack of good data in this course. Datasets are curated by organizations, like FRED or the government itself, but the exact methodology used to create the data series is often obscure. The frequency is not as granular as we would like, how can we hope to predict what will happen next year with confidence if we only have 30 annual observations or 360 monthly observations? For reference, your average machine learning model ingests no less than a few thousand observations. Finally, what do we do when data is simply not available or unreliable like in developing economies? The answer to all these questions is that we assume things about the underlying **data generating process.**

Here lies the first distinguishing characteristic of econometrics. **Econometric models are theoretically grounded on economic theories and are, hence, primarily used for testing hypotheses implied by such assumptions**. In the final project demo notebook, we used econometrics (a simple linear regression) to test the assumptions and predictions of the solow-swan model. But a myriad of things could have gone wrong: what if the data series chosen are not appropriate? What about the time period? What about the Cobb-Douglas functional form? …

[1] Taken from “A guide to econometrics” 4th edition by Peter Kennedy

There is an important fact that follows from this characteristic. **Traditional econometrics is about explaining** (trends, theories, assumptions), rather than predicting. We want to understand, in detail, each of the variables in our model and exactly how they relate to each other to tease out causal relationships from messy data. In contrast, data science is all about prediction; often at the cost of interpretability.

This backward looking feature of econometrics is, in my opinion, its greatest strength and its greatest weakness. Who cares about explaining an economic crisis once it already happened? Nature guarantees that the exact sequence of events that led to the crisis will never occur again, so what lessons can we possibly learn that will generalize into concrete policy prescriptions? On the other hand, being able to identify cause and effect from observational (i.e., non-experimental) data can be extremely powerful in constructing counterfactual analysis and guiding simulations. That is, we should care about explaining past events to the degree that we can pinpoint exactly what caused it. **Causal inference is what has differentiated econometrics from data science**.

Causality vs Correlation

The world is a complex system, which means that everything is interconnected. Therefore, any two variables we might be interested in will show some degree of correlation. But, the fact that two variables move together does not imply that the movement of one follows (or is caused by) the movement of the other. Correlation is a bidirectional relationship, but causation is all about finding direct and unidirectional relationships. Consider a canonical example: Ice cream sales and shark attacks both increase during the summer in Australia, does this mean that ice cream sales cause shark attacks? Maybe people eat more ice cream, cramp in the sea, and get eaten by sharks. Plausible, but very unlikely. Although note how easy it is to just come up with a *plausible rationalization* of the relationship observed.

Econometrics provides a framework for evaluating what are *credible* and *plausible* explanations by grounding the interpretation in economic theory. Consider the example of the well-known Phillips curve. The original observation by A.W. Phillips in 1958 was striking: for nearly a century (1861-1957), there appeared to be a ***stable negative relationship between unemployment rates and wage growth in the United Kingdom***. Economists later adapted this to show a relationship between unemployment and price inflation.

Let's pause here and think about what we're actually observing. We see two variables moving together in a systematic way across time. This is correlation. But does high unemployment *cause* low inflation? Does low unemployment *cause* high inflation? Or is there something else driving both variables? This is where econometrics becomes essential.

Consider these three scenarios regarding the Phillips Curve:

1. Unemployment directly affects inflation through wage pressures
2. Inflation affects unemployment through monetary policy responses

3. Both variables are affected by broader economic conditions (like aggregate demand shocks or supply chain shocks or pandemics)

The fundamental problem is that we observe all these relationships simultaneously in our data. This is what econometricians call the **identification problem**: how do we isolate the specific causal channel we're interested in?

At this point in your major, you probably have used a regression before. One insight I had during my PhD studies is that ***this tool requires that we assume the relationship to be stable***. Remember that throughout this semester we studied the assumptions of several models, all of which were supposed to be stable (like MPC, Consumption increasing with income, or investment decreasing with interest rate). Regressions work with a fixed sample of data, implying that the time period of analysis is fixed a priori, and therefore whatever gets estimated will only strictly hold for that specific time period. Change the time period, and you might get different estimations. We already encountered this in the R advanced notebook, where we tested the assumptions of the quantity theory of money empirically.

Look at the data for unemployment rate and inflation rate since the 1970s. You will find that:

- The 1970s show High unemployment AND high inflation (stagflation)
- The 2010s show Low unemployment AND low inflation
- The Post-COVID era shows Initially high unemployment AND low inflation, followed by low unemployment AND high inflation

So the Phillips curve sometimes holds, and sometimes it doesn't. This breakdown of an allegedly "stable" relationship illustrates three very important lessons about econometric analysis:

1. **Time-varying relationships**: Economic relationships that appear stable in one period may break down in another. This is particularly challenging for econometric analysis because we often assume some degree of stability in our relationships.
    a. My contribution to you is that this is only because of the constraints our instruments impose on us. Look back to our first lecture of the semester. Modern instruments include dynamic equations, simulations, and behavioral models that can capture changes over time and changes in underlying micro behavior.
2. **Omitted variables**: The original Phillips Curve omitted crucial variables like inflation expectations. When these expectations changed dramatically in the 1970s, the simple correlation broke down.
    a. This you will never fully work around. There is always some important variable you are missing (either because the data is unreliable, biased, or unavailable).
3. **Lucas Critique**: Robert Lucas argued that the parameters we estimate in our models may change when policy changes because people's behavior adapts to policy.

The core challenge of econometrics is moving from correlation to causation in a setting where we:

1. Cannot run experiments - we can't randomly assign different inflation rates to different countries or create multiple versions of an economy to test different policies
2. Have complex, simultaneous relationships
3. Deal with changing behavioral responses
4. Must rely on often imperfect data

Econometrics has developed several tools to address causality:

1. **Natural Experiments**: Using unexpected events or policy changes
2. **Instrumental Variables**: Finding variables that affect our cause but not our effect through other channels
3. **Structural Models**: Using economic theory to impose restrictions that help identify causal effects

Yet again, these traditional econometric approaches have their own set of challenges. Modern methods in macroeconomic modeling are increasingly relying on a new philosophical framework and leveraging cheap computing power available today. Two modeling approaches are worth noting:

1. **Dynamic Stochastic General Equilibrium** (DSGE): They represent a sophisticated attempt to address the Lucas Critique head-on. Instead of estimating reduced-form relationships, these models:
    - Start with microfoundations (utility maximization, profit maximization)
    - Explicitly model expectations
    - Include random shocks (the "stochastic" part)
    - Allow for dynamic adjustment over time
    - Estimate structural parameters that should be policy-invariant

The key econometric innovation here is that we're not just estimating individual relationships (like the Phillips Curve) in isolation. Instead, we're estimating an entire system of equations that represents our theoretical understanding of how the macro-economy works.

2. **Agent-Based Modeling** (ABM): A simulation approach to causality that:
    - Models individual agents (households, firms, banks) directly
    - Allows for heterogeneity and complex interactions
    - Generates macro patterns from micro behavior
    - Enables us to observe how system-level causality emerges from individual decisions

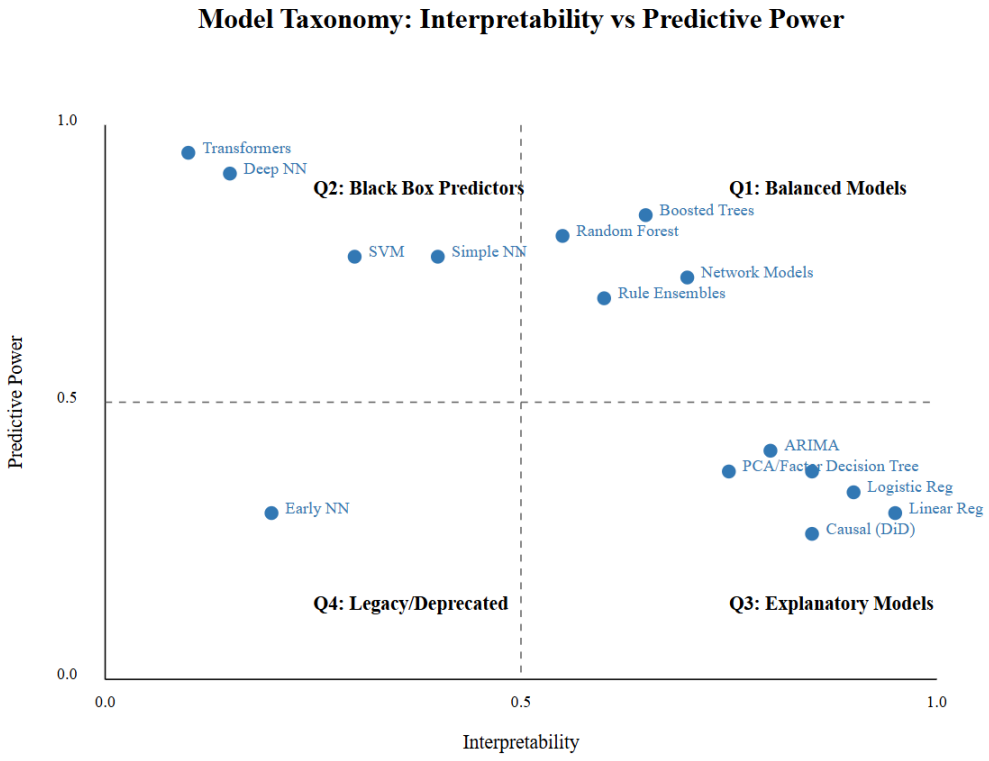
The econometric challenge with ABMs is different: instead of estimating parameters directly, we often need to: 1) Calibrate the model to match observed patterns, 2) Test whether the model generates realistic emergent behavior, and 3) Validate the model's predictions against real data.

Modern frameworks like DSGE and ABM represent different philosophical approaches to these challenges. DSGE models maintain the traditional econometric approach but with more sophisticated theory, while ABMs suggest a new way of thinking about causality in complex systems. The common denominator is that both rely on simulations and A LOT of compute power to make the estimations. Something that was impossible with state-of-the-art technology even 20 years ago.

This tension between different approaches to understanding causality will become clearer as we explore the distinction between inference and prediction in our next section.

Two Cultures: Inference vs Prediction

In 2001, statistician Leo Breiman published a seminal paper titled "Statistical Modeling: The Two Cultures"[2] where he identified a fundamental divide in how we approach data analysis. On one side stood the traditional statistical culture, focused on using data to infer the underlying process that generated it. On the other stood what he called the algorithmic modeling culture, focused on treating the data generating process as a black box and simply trying to predict outcomes accurately. I made up the following taxonomy, it helps me distinguish statistics/econometrics vs data science models and choose which model applies best to my problem and data, as Breiman prescribes. Notice there are no highly predictive and highly interpretable models, here lies modern efforts in causal machine learning. Check out the Appendix for more details.



[2] Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)

This distinction is particularly relevant for economics. Consider what happens when we observe a relationship between two variables, say, education levels and income. The "algorithmic" approach might employ sophisticated machine learning techniques, often including a bunch of non-linear relationships, to predict income based on education and numerous other variables. It might achieve impressive predictive accuracy, perhaps explaining 90% of the variation in incomes. But from an economist's perspective, this isn't enough.

Why? Because economics is fundamentally concerned with understanding mechanisms. When we observe that more educated people tend to earn higher incomes, **we want to know why**. Is it because education makes workers more productive? Is it because education serves as a signal to employers? Or is it because people with higher inherent ability tend to get more education? These questions matter tremendously for policy – if education is purely a signal, then universal education initiatives might not raise overall productivity.

This focus on mechanisms rather than predictions is deeply embedded in the econometric approach. When we specify a regression model, we're not just trying to find the best-fitting line through our data. Instead, **we're trying to represent our theoretical understanding of how the world works in a form that we can confront with data**. This is why econometrics puts such emphasis on careful model specification and identification strategies.

Let's make this concrete with an example from growth economics. Consider two approaches to studying economic convergence:

In a pure prediction framework, we might gather data on GDP per capita, investment rates, education levels, and numerous other variables across countries and time periods. Using machine learning techniques, we could likely build a model that makes reasonable predictions about future growth rates. This might be useful for investors or international organizations.

In contrast, the Solow model gives us specific predictions about convergence based on diminishing returns to capital. When we test these predictions econometrically, we're not just asking whether poor countries grow faster than rich ones. We're asking whether they do so for the reason the theory suggests. This might seem like a subtle distinction, but it's crucial for policy. If convergence occurs because of diminishing returns to capital, that suggests one set of policies. If it occurs because of technology diffusion, that suggests another.

This brings us to a crucial point about econometric modeling: **our goal is to estimate parameters that have theoretical interpretations**. These parameters help us understand the mechanisms at work in the economy, which in turn helps us evaluate policies and make better decisions. **This is fundamentally different from the machine learning approach of finding patterns that predict well by treating the Data Generating Process (DGP) as a black box.**.

As we move into our discussion of mathematical versus statistical models, keep this distinction in mind. The mathematical model represents our theoretical understanding of how the economy

works. The statistical model is our strategy for confronting that theory with data. Both are essential to the econometric enterprise.

Mathematical vs Statistical models

Consider our earlier discussion of the Phillips Curve. The mathematical model posits a clean relationship: higher unemployment leads to lower inflation. We can write this as a simple equation:

$$\pi = \beta_0 - \beta_1 \cdot u + \epsilon$$

where π is inflation, u is unemployment, and ε is... well, this is where things get interesting. In a purely mathematical model, we might treat ε as a minor deviation or even ignore it entirely. The equation represents an idealized relationship, what we believe to be the true "Data Generating Process" (DGP) of the economy. This is similar to how the Solow model presents a **deterministic** relationship between capital, labor, and output.

But reality is messier. When we try to estimate this relationship using actual data, we're forced to confront uncertainty head-on. This is where we transition from a mathematical to a statistical model. The same equation now takes on a different interpretation: it becomes a **probabilistic** statement about the relationship between inflation and unemployment. It also forces us to decide which data to use to represent each variable. Our phillips curve model could become:

$$CPI = \beta_0 + \beta_1 \cdot UNRATE + \epsilon$$

The term ε, called the error or disturbance term, now becomes the point of focus. In a statistical model, it represents our acknowledgment that the relationship we're studying is not deterministic. Some of this uncertainty comes from measurement error - we don't measure inflation or unemployment perfectly. Some comes from omitted variables - factors affecting both inflation and unemployment that we haven't included in our model. And some comes from the inherent randomness in economic relationships.

This distinction becomes particularly important when we return to the models you've studied in macroeconomics[3]:

1. The Classical Model: As a mathematical model, it posits perfect wage and price flexibility leading to continuous market clearing. As a statistical model, we need to account for the fact that different markets adjust at different speeds and that we observe prices with error.

[3] As an exercise, try writing down the mathematical models we learnt in class and then writing the statistical model you would set up for testing it with data.

2. IS-LM: The mathematical version gives us clean equilibrium conditions where the goods and money markets clear simultaneously. The statistical version must grapple with how to estimate these relationships when we observe them adjusting dynamically over time.

3. Mundell-Fleming: The mathematical model assumes perfect capital mobility and immediate interest rate equalization. The statistical version needs to account for the fact that capital flows respond to many factors beyond interest differentials and that adjustment isn't instantaneous.

The transformation from mathematical to statistical model isn't just about adding an error term. It fundamentally changes how we think about our parameters. **In a mathematical model, β_0 and β_1 are fixed**, unknown constants waiting to be discovered. **In a statistical model, our estimates of these parameters are themselves random variables**, subject to uncertainty that we need to quantify.

The distinction between mathematical and statistical models also helps explain why economists put so much emphasis on model specification. When we write down a statistical model, we're making claims about:

- The functional form of relationships (linear vs non-linear)
- Which variables to include
- The properties of our error terms
- How our variables are measured
- What we assume about unobservable factors

These aren't just technical details - they're crucial assertions about how we believe the economy works. This is why econometricians spend so much time discussing assumptions about error terms and testing for various statistical properties. These tests aren't just mathematical exercises; they're ways of checking whether our statistical model is a reasonable representation of the economic relationships we're trying to understand.

This leads us naturally to our next topic: the linear regression model, which represents the workhorse tool for estimating statistical models in economics. ***But before we dive into the mechanics of regression, remember: the technique itself is just a statistical tool. What makes it useful for economics is how we embed it within economic theory.***

Linear Regression

We began this lecture by discussing the criticisms of econometrics, particularly the challenge of drawing reliable conclusions from non-experimental data. We then explored how econometricians approach causality, the distinction between inference and prediction, and the transformation of mathematical models into statistical ones. All of these threads come together in what is perhaps the most fundamental tool in traditional econometrics: the linear regression model.

It often happens that we have two sets of related values/variables/features, and we want to estimate or predict the value of one variable that would correspond with a given value on the other. That is, there are random variables X and Y that, we hypothesize, have a relationship between each other. For example, midterm grades and final grades for a class OR quality of sleep and positive emotions OR interest rates and inflation. Perfect estimations are only possible when all dots lie on the same straight line. With perfect correlation, we can say exactly what value of one variable will go with any given value of the other. This special case of correlation (for a long time believed to be analogous to causal inference) motivated the approach of "reducing" the data to one **line of best fit**. The question being asked is "what is the underlying straight line from which all these points deviate?"

The challenge is, of course, estimating this line of best fit. The underlying idea is simple: we have to apply some operation on X and Y that results in a number that encodes the main information of interest about the relationship we hypothesize. In other words, we reduce the dimensionality from 2D to 1D.

There are many methods, think of any of the statistics you have learnt that take in many inputs and return one single value. The average, and its variations, is the most commonly applied method because we are often interested in the average behavior of the relationship (but note how this means that we cannot, by design, obtain perfectly precise estimates. All of our results are "on average"). The collection of methods used to find the position (intercept + slopes) of the line of best fit fall under the **regression line** category. It was invented by Galton in the 19th century to study the relationship between the height of fathers and sons. His studies resulted in a finding called "regression to mediocrity", what we now know as regression to the mean, alluding to the tendency of trends to revert to their average behavior. Mathematically, the regression line is defined by a **regression equation** such as $y = mx + b$.

We now know that we can find correlations everywhere, the chicken and sunrise or sharks and ice cream sales are canonical examples. In order to claim that our predictions are sound and valid, we must impose a specific form on the structure of the regression equation. **A regression equation + constraints derived from economic theory + data = an econometric model**. The constraints imposed have two flavors:

1) on the functional form of the equation (linear or non-linear), and
2) on the variables chosen, which can be one or many.

While the form is often borrowed from existing validated forms (unless you are doing theory, in which case you might be interested in proposing a new form all together), the latter is more like a guideline rather than a constraint. We want to choose independent variables that are logically and empirically related to the dependent variable. This selection is based on economic theory, and your hypothesis which relies on experience + observations.

The model is encapsulated by the equation $Y=\beta_0 +\beta_1* X+U$, where:

- Y is the dependent variable we aim to predict or explain
- X is the independent variable that we use as a predictor.
- β_0 (the intercept) and β_1 (the slope) are parameters of the model that we seek to estimate.
- U represents the error term, capturing all other factors affecting Y that are not included in our model.

This equation represents a linear relationship between X and Y. The slope, β_1, indicates the average change in Y for a one-unit increase in X. The intercept, β_0 , signifies the expected value of Y when X equals zero.

This linear model describes the relationship we believe the variables have with each other. Now, we have a functional form that represents or describes the relationship we are interested in modeling. But we still need to find optimal values of the relevant parameters.

So, basically, we want to find an estimator (remember, a process or set of rules that allows us to approximate the population from sample data) that will takes us as close as possible to the true value of the coefficients in our model (β_0 and β_1). For this, we move from the mathematical formulation of our model to the statistical formulation.

Remember our final project demo notebook where we tested the Solow-Swan model's predictions? We used a simple linear regression to answer a fundamental question: to what extent does capital deepening explain output per worker in the United States? This empirical exercise exemplifies everything we've discussed today about econometrics.

Let's deconstruct what we did:

1. **Theory to Testing** We started with the Cobb-Douglas production function: Y = K^α * L^(1-α)

To test this, we:

- Converted to per-worker terms: y = k^α
- Took natural logs to linearize the equation: ln(y) = α*ln(k)
- Added an intercept and error term: ln(y) = β_0 + β_1*ln(k) + ε

This progression illustrates the journey from mathematical model (Cobb-Douglas) to statistical model (log-linear regression).

2. **Causality and Inference** Our hypotheses were carefully structured:
    - H1: Positive relationship between k and y (causality)
    - H2: α ≈ 1/3 (testing a specific theoretical prediction)
    - H3: Diminishing returns to capital (mechanism)

Notice how these align with our earlier discussion about econometrics being focused on testing theoretical mechanisms rather than just finding patterns.

3. **Data Challenges** Remember the data issues we encountered:
    - Had to merge different frequencies (annual capital stock with quarterly GDP)
    - Dealt with missing values
    - Needed to construct per-worker variables

These are exactly the kind of "messy data" challenges we discussed in critiquing econometrics.

Next class will focus on the method we used to estimate the values of this parameter, called Ordinary Least Squares (OLS), which underlies what the `lm()` function in R is doing in the background. We will work through the mathematical derivation of OLS, the Gauss Markov Assumptions (GMAs) which is a list of conditions that our data and model must meet to justify OLS, interpreting the summary results of a linear regression model, and using statistical tests to check if we meet each of the GMAs. *When you look at that regression output next class, remember: behind those coefficients and standard errors lies a powerful tool for testing economic theory. The challenge is knowing when we can trust what the regression tells us - and that's exactly what we'll learn about next time.*

Appendix

The following table contains the data used to create the plot shown in the lectures. Relative scores are subjective and estimated via a supervised LLM classification. That is, I came up with subjective classifications for each model and a rationale. The LLM is fed this rationale and prompted to calculate values that would visually illustrate these classifications. A next step in developing the taxonomy further would be training each model on a few benchmark datasets and comparing evaluation metrics for each dimension.

Model Scores Table (Normalized 0-1 Scale)

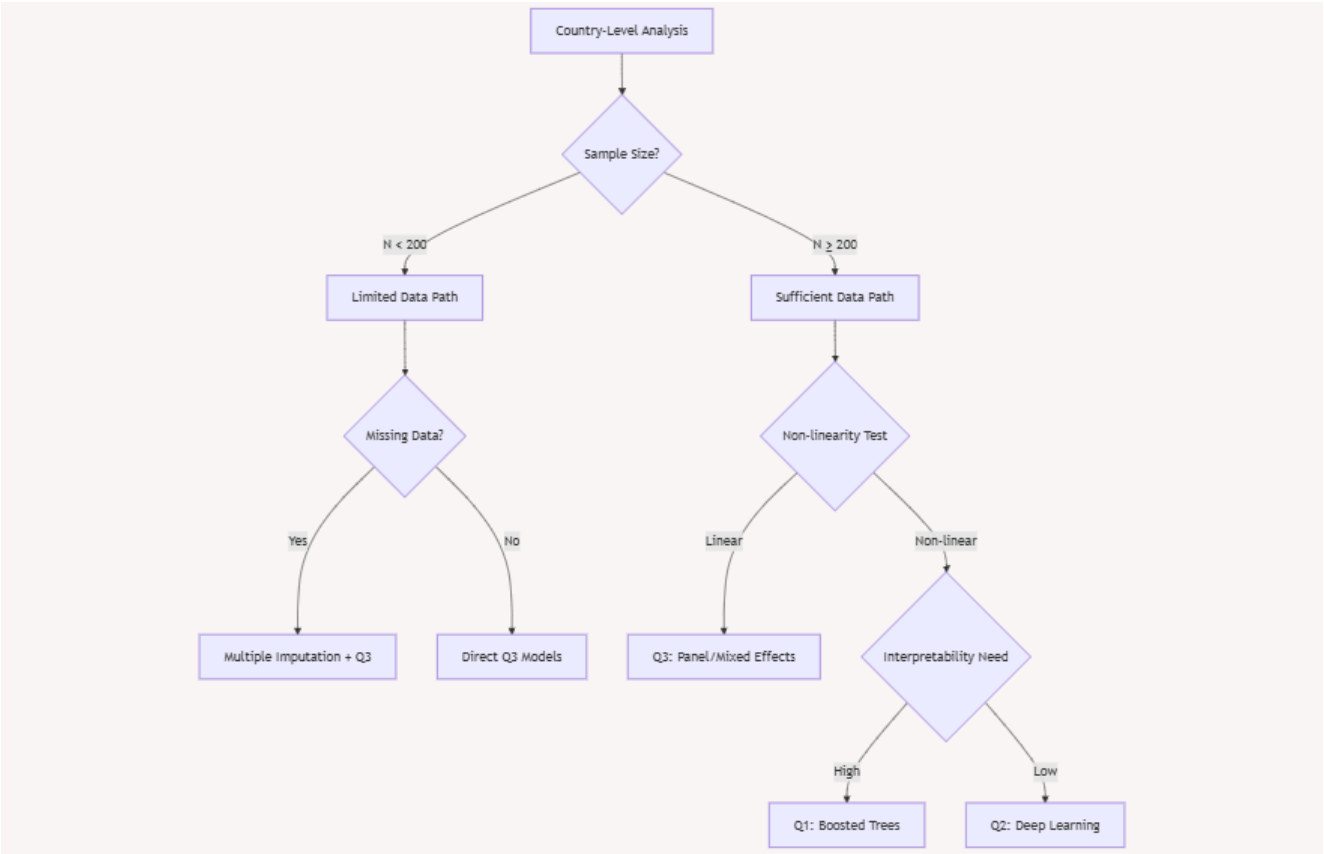
Model Type	Interpretability	Predictive Power	Reasoning
Linear Regression	0.95	0.30	Highest interpretability due to simple linear relationships, but limited predictive power
Logistic Regression	0.90	0.35	Very interpretable odds ratios, slightly better predictions than linear
Decision Trees (Single)	0.85	0.40	Highly interpretable rules, moderate predictions
Gradient Boosted Trees	0.65	0.85	Partially interpretable through feature importance, strong predictions
Random Forests	0.55	0.80	Less interpretable than boosted trees but still provides feature importance
Neural Networks (Simple)	0.40	0.75	Basic architectures maintain some interpretability
Deep Neural Networks	0.15	0.95	Very low interpretability but highest predictive power

Support Vector Machines	0.30	0.75	Complex decision boundaries reduce interpretability
Network Models	0.70	0.70	Balanced through visualization and community metrics
Causal Models (DiD)	0.85	0.25	High interpretability through causal effects, low prediction focus
Early Neural Nets	0.20	0.30	Poor on both dimensions due to architectural limitations
PCA/Factor Analysis	0.75	0.40	Interpretable components but limited predictive application
ARIMA Models	0.80	0.45	Clear temporal relationships but limited to specific patterns
Transformers	0.10	0.98	Lowest interpretability but state-of-art predictions
Rule-based Ensembles	0.60	0.65	Balance through explicit rules and ensemble power

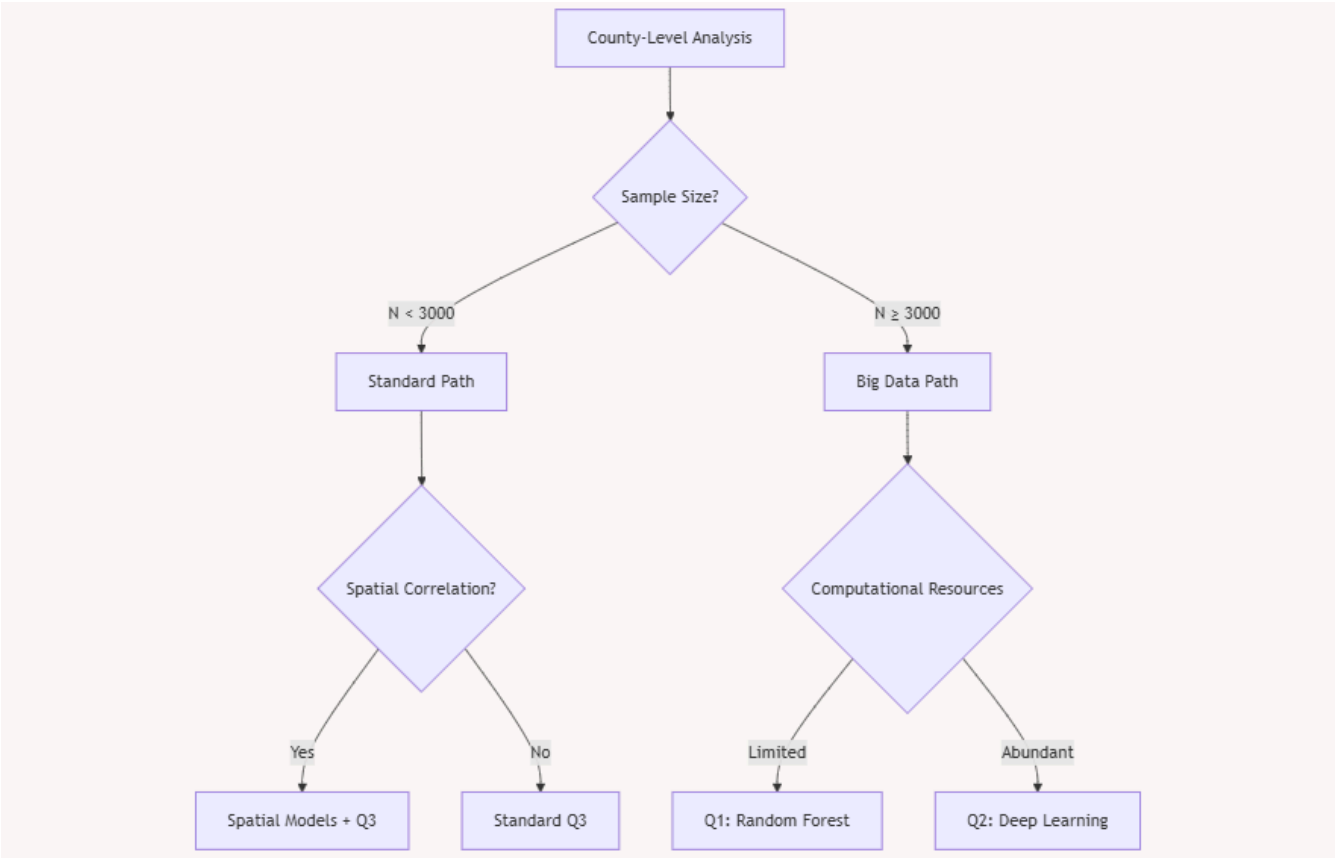
Based on this classification, I propose the following workflow to guide your model selection and approach:

https://shorturl.at/yxBtY

For example, suppose you are working with a country-level cross-sectional dataset (which would have at most 195 observations. Your sample size is small, so big models will be useless even if you have a lot of variables (although trees could be useful in that situation). Your best option in this case is a good old explanatory model. Estimate a regression model to predict or classify.



In contrast, suppose you are working with county-level data which is much more granular. If the units are highly correlated, then you are probably dealing with a highly non-linear scenario. If interpretability is very important, you should rely on regression based methods like linear regression, logistic regression, or decision trees. If the dataset is big enough, the cost in potential predictive power increases significantly relative to the benefits of interpretability. In these scenarios, you should aim for a highly predictive model first (maybe applying a random forest to balance out the trade-off initially) and then focus on explanation. Many policy scenarios require *knowing what will happen*, with some confidence level, and would rather predict first and explain later if the predictions are sufficiently accurate. But once you are dealing with big data, local computational resources become a decision point. This limitation is increasingly overcome by cheap compute power available through cloud providers or cloud based notebooks like google colaboratory.


County-Level Analysis
Sample Size?
N < 3000
N ≥ 3000
Standard Path
Big Data Path
Spatial Correlation?
Computational Resources
Yes
No
Limited
Abundant
Spatial Models + Q3
Standard Q3
Q1: Random Forest
Q2: Deep Learning