How to Develop an Economics Research Paper

Key Steps, Workflow, Data Sources, and Examples

CLABE 2025/2026

Marco Rosso

4 December 2025

Learning Goals

- Understand the research pipeline: from idea to publication
- Learn where to find and evaluate data
- Master empirical identification strategies
- Apply marginal effects concepts to real data (Mroz 1987)
- Develop skills for independent research

What Makes a Good Research Idea?

Three essential ingredients:

- **Economically relevant:** Addresses a real-world question that matters
- **Novel contribution:** Something new relative to existing literature
- **Feasible:** Data available and identification strategy credible

Red flags:

- Too broad ("How does education affect income?")
- No clear data source
- Same as 50 other papers (no new angle)

Sources of Inspiration

Where to look for ideas:

- **Top journals:** AER, QJE, JPE, ReStud, Econometrica
- **Policy reports:** OECD, World Bank, IMF
- **Seminars and conferences:** Discussions, feedback loops
- **Replication with new data:** Take a classic paper, apply to new context
- **Advisor discussions:** Brainstorming sessions

Pro tip: Read extensively in your field. Ideas often emerge from gaps between papers.

Formulating Research Questions

A good research question is:

- **Answerable with data:** Don't ask "Should government do X?" (normative)
- **Narrow and focused:** Can be addressed in one paper (or one chapter)
- **Theoretically motivated:** Grounded in economic mechanisms
- **Empirically testable:** Clear predictions from theory

Example (good): *"How does access to childcare subsidies affect female labor force participation?"*
Example (bad): *"What policies improve the economy?"* (too vague)

Building Your Theoretical Framework

Why theory matters?

- Guides empirical design
- Generates testable predictions
- Helps interpret results

Your checklist:

1. Outline the economic mechanisms (how does A lead to B?)
2. Decide: Do I need a formal model?
    - YES if predictions are non-obvious or competing theories exist
    - NO if mechanisms are straightforward
3. List potential confounders and channels
4. Identify testable hypotheses

Types of Data

- **Micro data:** Individuals, households, firms
- **Administrative data:** Tax records, education, health, voting
- **Survey data:** Cross-sectional, panel (repeated over time)
- **Geospatial data:** Maps, satellite imagery, GPS coordinates
- **Historical archives:** Old documents, newspapers
- **Experimental data:** RCTs, field experiments

Key trade-off: Precision vs. coverage $\rightarrow$ Administrative data is rich but restricted.

Public Data Sources

Where to find free, high-quality datasets

- World Bank Microdata — `https://microdata.worldbank.org/`
- OECD Data — `https://data.oecd.org/`
- Eurostat — `https://ec.europa.eu/eurostat/`
- IPUMS (Census data) — `https://ipums.org/`
- DHS (health, demographics) — `https://dhsprogram.com/`
- Harvard Dataverse — `https://dataverse.harvard.edu/`
- Gapminder — `https://www.gapminder.org/data/`
- *...and many others*

Pro tip: Always check documentation, sample size, and coverage before committing.

Evaluating Datasets

Checklist before analysis:

- Sample size and representativeness
- Geographic and time coverage
- Variable definitions and coding
- Missing data patterns
- Potential measurement error
- Data quality reports

Red flag: If documentation is unclear or minimal, move on.

Core Identification Strategies

1. **Randomized Controlled Trials (RCTs):** Gold standard (exogenous treatment)
2. **Difference-in-Differences (DiD):** Exploit policy timing
3. **Instrumental Variables (IV):** Use exogenous variation in instrument
4. **Regression Discontinuity (RD):** Exploit cutoff rules
5. **Panel Fixed Effects:** Control for time-invariant confounders
6. **Synthetic Control:** Construct comparison group for policy evaluation

The choice depends on your research question and data available.

Requirements for Credible Identification

Your **empirical design** must satisfy:

- **Exogeneity:** Treatment is not correlated with unobservables
- **Transparency:** Clearly state assumptions (parallel trends, exclusion restriction, etc.)
- **Balance:** Treatment and control groups are similar pre-treatment
- **Robustness:** Results hold with alternative specs

Always include:

- Pre-treatment comparisons (balance tests, pre-trends)
- Falsification tests
- Robustness checks (alternative specs, placebo tests)

Paper Structure

Standard organization:

1. **Introduction:** Hook + motivation + contribution
2. **Literature Review:** Position your paper
3. **Institutional Background:** Context and institutions
4. **Data:** Sources, definitions, summary statistics
5. **Empirical Strategy:** Identification approach
6. **Results:** Main findings + robustness
7. **Mechanisms:** Heterogeneity and channels
8. **Conclusion:** Implications

Golden rule: Big-picture intuition BEFORE technical details.

Writing Tips

- Use simple, direct language (avoid jargon)
- Place results in tables and figures (easier to parse)
- Include graphical abstracts (event studies, maps, before-after plots)
- Replicate all results with do-files / scripts
- Get feedback from co-authors, advisors

Useful tools:

- **LaTeX:** Professional typesetting (Overleaf for cloud editing)
- **Git/GitHub:** Version control
- **Stata/R/Python:** Statistical analysis

Getting Your Work Out

- **Present:** Seminars, brown-bags, conferences
- **Upload:** SSRN, NBER (working paper versions)
- **Provide replication materials:**
  - Data documentation
  - Code with comments (master do-file / main script)
  - README file explaining everything
- **Submit:** Target journals based on your topic and field

Pro tip: Early feedback on working papers saves time later.

From Template to Real Paper

So far:

- We built a general roadmap for an economics paper.
- We discussed where to find data and how to think about identification.

Next: See how a real paper (Mroz 1987) fits this template and use it as a lab for binary choice models and marginal effects.

$\longrightarrow$ Mroz, T. A. (1987). *The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions*. *Econometrica*, 55(4), 765–799.

Empirical Application: Mroz (1987)

Big picture:

- Topic: Female labor supply (married women in the U.S.).
- Data: PSID 1975 (interview year 1976), 753 married women.
- Outcome: Labor supply (hours / participation), wages, non-wife income, children.
- Goal: Show how different economic/statistical assumptions (Tobit, exogeneity, selection) change estimated wage and income effects.

Why we use it here: canonical dataset, clean example of our pipeline:

$$\text{question} \rightarrow \text{data} \rightarrow \text{model} \rightarrow \text{robustness.}$$

From Template to Real Paper: Mroz (1987) (1)

Our generic structure vs. **Mroz's paper sections**

Our template	**Mroz (1987)**
Introduction: motivation, contribution, literature	Intro pages: motivates wide range of labor-supply estimates, shows Table I with previous studies, states contribution as a systematic sensitivity analysis.
Institutional background / context	Short discussion of female labor supply literature and PSID data context (Panel Study of Income Dynamics, 1975 wave).
Data: sources, definitions, summary stats	PSID sample description, definition of hours, wages, nonwife income; Table III with means and standard deviations.

From Template to Real Paper: Mroz (1987) (2)

Our template	**Mroz (1987)**
Empirical strategy / model	Section *"The Basic Labor Supply Model":* linear labor supply equation, instruments, selection issues, assumptions (Tobit, exogeneity).
Results + robustness	Tables IV–VIII: alternative specifications, exogeneity tests, selection corrections, tax controls; discussion of sensitivity.
Conclusion	Final section: summarizes main conclusions about wage and income elasticities and implications for female labor supply.

Paper Structure in Practice: Mroz (1987)

When you read Mroz (1987), try to **locate our checklist:**

1. **Research question & contribution**
   $\longrightarrow$ *Opening paragraphs and discussion around Table I.*
2. **Data and variables**
   $\longrightarrow$ *Description of PSID sample, construction of hours, wages, nonwife income, and Table III.*
3. **Model and identification**

   $\longrightarrow$ *Section "The Basic Labor Supply Model": choice of functional form, instruments, assumptions (Tobit, exogeneity, selection).*
4. **Robustness / sensitivity**
   $\longrightarrow$ *Comparisons across Tables IV–VIII: how wage and income effects change with different assumptions.*
5. **Conclusion**
   $\longrightarrow$ *Final pages: main message that wage and income effects are smaller and more stable than many previous studies suggest.*

Key Variables in the Mroz Dataset

Outcome variables:

- **inlf**: labor force participation ($1 =$ worked, $0 =$ did not work).
- **hours**: annual hours of work.

Main regressors:

- **educ**: years of schooling.
- **exper**, **expersq**: labor market experience.
- **nwifeinc**: non-wife income (other household income).
- **kidslt6**, **kidsge6**: number of young and older children.

Link to our empirical exercise: same variables used in the logit/probit models and marginal effects do-file.

Economic Hypotheses

Predictions from theory:

1. **Income effect:** Higher household income $\Rightarrow$ less likely to work
   - Intuition: Can afford to substitute away from market work
2. **Education effect:** More education $\Rightarrow$ more likely to work
   - Intuition: Higher wage, stronger incentive
3. **Childcare constraint:** Young children $\Rightarrow$ less likely to work
   - Intuition: Childcare is costly; must work elsewhere

$\longrightarrow$ **Model:** Binary choice (Logit or Probit)

Binary Choice Model

Latent variable framework:

$$\text{inlf}_i^* = \beta_0 + \beta_1 \text{nwifeinc}_i + \beta_2 \text{educ}_i + \cdots + \varepsilon_i$$

$$\text{inlf}_i = \begin{cases} 1 & \text{if inlf}_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

Key challenge: β_k is NOT the marginal effect!

- In linear models: $\beta_k = \frac{\partial y}{\partial x_k}$
- In logit/probit: $\beta_k \neq \frac{\partial P(y=1)}{\partial x_k}$
- Must compute marginal effects explicitly

Mroz Results: Illustration of Logit Output

Marginal effects (Logit, AME):

Variable	Coefficient	AME	Std. Error
nwifeinc	-0.021^{**}	-0.0038^{**}	0.0016
educ	$+0.221^{***}$	$+0.0395^{***}$	0.0075
exper	$+0.206^{***}$	$+0.0368^{***}$	0.0052
expersq	-0.003^{***}	-0.0006^{***}	0.0002
age	-0.088^{***}	-0.0157^{***}	0.0024
kidslt6	-1.443^{***}	-0.2578^{***}	0.0324
kidsge6	$+0.060$	$+0.0107$	0.0142

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

Interpretation (AME, in percentage points):

- Each additional \$1,000 of non-wife income $\Rightarrow$ about 0.4 pp $\downarrow$ in participation
- One more year of education $\Rightarrow$ about 4.0 pp $\uparrow$ in participation
- One more child < 6 years $\Rightarrow$ about 25.8 pp $\downarrow$ in participation

Why Marginal Effects Matter

Example: Suppose $\beta_{\text{nwifeinc}} = -0.10$ in logit

Wrong interpretation: "A $1,000 increase in non-wife income decreases participation by 10 percentage points."

Why wrong? The effect size depends on baseline probability:

- Near $P = 0.5$: effect is LARGE
- Near $P = 0$ or $P = 1$: effect is SMALL (flat CDF region)

Solution: Compute marginal effects at meaningful points

Average Marginal Effects (AME)

Definition:

$$\text{AME}_k = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial P_i}{\partial x_{ik}}$$

Interpretation: On average across the sample, a one-unit increase in x_k changes predicted participation by AME_k percentage points.

Advantage: Representative of typical effect

Disadvantage: Doesn't correspond to any single individual

Marginal Effects at the Means (MEM)

Definition:

$$\text{MEM}_k = \frac{\partial P}{\partial x_k}\bigg|_{x=\bar{\mathbf{x}}}$$

Interpretation: For an "average" woman (at sample means), a one-unit increase in x_k changes predicted participation by MEM_k percentage points.

Advantage: Interpretable as effect for typical person

Disadvantage: Mean individual may not exist

Logit Marginal Effects: Formulas

Logit CDF: $P_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{-x_i\beta}}$

Marginal effect on x_{ik}:

$$\frac{\partial P_i}{\partial x_{ik}} = P_i(1-P_i)\beta_k$$

AME:

$$\text{AME}_k = \frac{1}{N}\sum_{i=1}^{N} P_i(1-P_i)\hat{\beta}_k$$

MEM:

$$\text{MEM}_k = P_{\bar{x}}(1-P_{\bar{x}})\hat{\beta}_k$$

where $P_{\bar{x}}$ is predicted probability at sample means

Probit Marginal Effects: Formulas

Probit CDF: $P_i = \Phi(x_i\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i\beta} \exp\left(-\frac{t^2}{2}\right) \, dt \longrightarrow$ standard normal CDF

Marginal effect on x_{ik}:

$$\frac{\partial P_i}{\partial x_{ik}} = \phi(x_i\beta)\beta_k$$

where ϕ is the standard normal PDF.

AME:

$$\text{AME}_k = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i\hat{\beta})\hat{\beta}_k$$

MEM:

$$\text{MEM}_k = \phi(\bar{x}\hat{\beta})\hat{\beta}_k$$

Logit vs. Probit

Feature	**Logit**	**Probit**
Distribution	Logistic	Normal
CDF	$\frac{1}{1+e^{-z}}$	$\Phi(z)$
ME Formula	$P(1-P)\beta_k$	$\phi(z)\beta_k$
Tail Behavior	Heavier	Thinner

In practice: Results are usually very similar. Choice is often conventional.

Key Takeaways

- ✓ **Research process is iterative:** From idea → data → analysis → writing
- ✓ **Data quality matters:** Invest time in understanding your data
- ✓ **Identification is crucial:** Credible causal claims require careful design
- ✓ **Interpretation requires care:** In non-linear models, look at marginal effects, not coefficients
- ✓ **Transparency builds trust:** Share code, data, assumptions

⟹ **Next steps:** Start with a question that excites you. The rest follows.

Questions?

Let's see Mroz (1987) in Stata