SCAFFOLDING ASSIGNMENT: RESEARCH DESIGN

Jason Tran

Data Analytics Department, Denison Univeristy

DA 401

Professor Matt Lavin

10/13/2024

1. Purpose

The primary objective of this research is to examine how California's carbon pricing policies from 2015 to 2021 have affected the transportation energy burden among low-income households. Specifically, the study focuses on the differential impacts on female-headed versus male-headed households in urban and rural areas using publicly available data.

To achieve this, the research aims to carefully outline the selected methods, measurements, and statistical analyses to address the research question. The methods section plays a pivotal role in clarifying these choices. It begins by specifying the techniques used in the study, providing an evaluation of both their advantages and disadvantages. This helps in justifying the chosen methods and ensures transparency in the decision-making process. Additionally, the methods section defends the methodological choices by referencing relevant literature, particularly concerning parameter assumptions, common validation techniques, best practices, and robustness checks, as discussed by Angrist and Pischke (2009). This solidifies the validity of the selected approach, ensuring that the methods are grounded in established research standards.

Furthermore, the section establishes a clear connection between the datasets and the methods, demonstrating that the data is well-suited to the chosen analytical techniques. By doing so, it highlights how these methods will advance the understanding of the research question and provide meaningful insights. Finally, the methods section prioritizes clarity and reproducibility by offering detailed descriptions that allow other researchers to replicate the study, reinforcing the rigor and credibility of the research.

1. Research Question and Causality

This study explores the impact that carbon pricing policies of California had on the use of gasoline among the vulnerable groups. Extra attention is paid to gender and geographical differentiation (subgroups include female-headed and male-headed households and urban and rural populations) since relevant research limitations are mentioned to be filled in Marron et al. (2015) and Ohlendorf et al. (2021). Therefore, the finalized research question is as follows:

- **Research Question**: How have California's carbon pricing policies from 2015 to 2021 affected the transportation energy burden among low-income households, and what are the differential impacts on female-headed versus male-headed households in urban and rural areas?

This research aims to fill the gaps identified in previous studies (Marron, Toder, & Austin, 2015; Ohlendorf et al., 2021) by providing a nuanced analysis of gender and geographical subgroups, expanding on the initial plan by identifying the datasets, statistical techniques and measurement approach to be employed in the analysis of the research question.

Key objectives of the assignment are as follows:

Clarify Selected Methods

Given the observational nature of the data and the policy context, a quasi-experimental design is appropriate. Because of this, the study will employ the Difference-in-Differences (DiD) technique, which is a quasi-experimental method that enables the researcher to compare outcomes (amount of gasoline use) before and after the treatment (California’s carbon pricing

policies) between the treatment group, which is California, and the control group, which consists of other states that did not implement carbon pricing policies. The key assumption is that, in the absence of the policy, both groups would have followed parallel trends in gasoline consumption. While DiD controls for unobserved time-invariant differences, it cannot account for time-varying unobserved factors. Thus, the results suggest associations that imply causality but do not definitively establish it. If the parallel trends assumption is violated, **Propensity Score Matching (PSM)** will be employed prior to DiD to balance covariates between treatment and control groups (Caliendo & Kopeinig, 2008). This enhances the validity of causal inferences by reducing selection bias.

Highlight Strengths and Limitations

All the chosen methods will be assessing on its advantage and disadvantage. DiD models have the ability to make causal inferences about the policy effects on treated groups and obscures time-invariant characteristics. Some limitations are that trends are assumed to be parallel and there might be an omitted variable bias and will be tested with both external validators and sensitivity analyses.

Justify the Methods Using Relevant Literature

The above choices of method will be justified based on empirically informed integrations from the applied econometric and policy analysis literature including but not limited to Angrist and Pischke (2009). This approach guarantees that the chosen statistical methods provide academic soundness regarding both parametric assumptions and best practice recommendations and robustness checks (e.g., Blundell & Dias, 2009).

Demonstrate Dataset Suitability

The research aims to use all of the following datasets (subjected to further changes):

- Consumer Expenditure Survey (CES): Provides data on household gasoline expenditures, income, and demographics (U.S. Bureau of Labor Statistics, n.d.).
- American Community Survey (ACS) PUMS: Offers individual-level data on gender, income, and urban/rural status (U.S. Census Bureau, n.d.).
- California Air Resources Board (CARB): Contains details on carbon pricing policies and implementation timelines (California Air Resources Board, n.d.).
- U.S. Energy Information Administration (EIA): Supplies state-level gasoline prices (U.S. Energy Information Administration, n.d.).

CES, ACS PUMS, and CARB data are ideal for asking the research question because CES captures consumers' energy expenditure, ACS PUMS reflects consumers' characteristics which could influence their spending patterns, and CARB is exclusive to California, where focus group is located. These datasets are basic and contain factors such as household expense on gasoline, income per households, and demographic details (gender and urban/rural split) which facilitates a complete analytical review of policy measures linked to carbon pricing.

Enhance Understanding through Statistical Methods:

The research will use econometric modeling to sort out the impact of carbon pricing policies on gasoline usage in households of different categories. With interaction terms in the DiD model, the study will determine how the outcomes differ by gender and regions, policy

impact. This will contribute to the development of knowledge on effects of environmental policies on subpopulations of different economic status.

Ensure Clarity and Reproducibility:

To achieve the highest level of transparency, any method used, dataset analyzed, or statistical procedure applied will be described with sufficient detail to enable replication. The methods used in the study will be conductive to replicability where other scholars can conduct similar analysis using the same datasets and procedure.

Causality Considerations

This research utilizes a quasi-experimental research design based on the Difference-in-Differences (DiD). This technique is applicable to observational data, policy exposed (California) versus policy non-exposed (other states without carbon pricing). DiD achieves the elimination of time-fixed effects between the treatment and control groups, if indeed it is assumed that selection bias is time-fixed (Angrist & Pischke, 2009). The method supposes that without the policy the changes in the gasoline consumption should be similar for both groups.

Another technique can be used before DiD, which is Propensity Score Matching (PSM) that can reduce baseline differences as well (Caliendo & Kopeinig, 2008). Less strict than randomized experiments, DiD does work well for studying policy changes, but therefore the results highlight relationships more than causality (Blundell & Dias, 2009).

2. Data Description

Using publicly available databases, the study estimates household gasoline expenditure. These include:

Data Sources and Collection

- **Consumer Expenditure Survey (CES):** Contains information on expenditure on gasoline by a household, household income, and other demographic characteristics. Information is obtained from the U.S. Bureau of Labor Statistics (U.S. Bureau of Labor Statistics, n.d.).
- **American Community Survey (ACS) PUMS:** Issues gender, income and urban/rural splits which are readily provided by the United States Census Bureau (U.S. Census Bureau, n.d.).
- **California Air Resources Board (CARB):** Contains information on carbon pricing policies and the years when such policies are to take effect.
- **U.S. Energy Information Administration (EIA):** Offers state level gasoline prices to analyse policies (U.S. Energy Information Administration, n.d.).

Acquisition and Organization of Data

Both data sources contain state identifiers, and the datasets will be matched to allow for the comparison of different time periods. Potential variables will be selected with caution; costs will be presented in constant dollars. Key variables include:

Variable	Type	Description
Gasoline Expenditure	Continuous	Annual household spending on gasoline (in constant dollars)
Post-Tax Indicator	Binary	0 = Pre-policy period, 1 = Post-policy period
Household Gender	Binary	1 = Female-headed households, 0 = Male-headed
Urban/Rural Status	Binary	1 = Rural areas, 0 = Urban areas
Income Level	Binary	1 = ≤200% of Federal Poverty Level, 0 = Otherwise
Household Size	Continuous	Number of individuals in the household
Employment Status	Categorical	Employment status of the household head
Education Level	Categorical	Highest educational attainment of the household head
Vehicle Ownership	Continuous	Number of vehicles owned by the household
Gasoline Price	Continuous	State-level average gasoline price

Data Structure

The data encompass spatial (state-level), temporal (2015–2021), and household-level dimensions, allowing for panel data analysis.

Data Wrangling Steps

The pre-processing of data will entail converting the format of the data, filling in missing values either through imputation or by excluding records with long gaps, and optionally excluding outliers. All variables will be deflated using the Bureau of Labor Statistics Consumer Price Index (CPI). Geographic identifiers will be used to match the collected data from ACS, CES, CARB, and EIA in order to combine household and policy data. Some of the steps planned

for this phase include: handling missing values through appropriate imputation or exclusion, adjusting monetary variables to constant dollars and creating binary indicators, integrating CES, ACS PUMS, CARB, and EIA data using state and year identifiers, and identifying and removing extreme values that may skew results.

Ethical Considerations

This paper uses only secondary data and thus no communication takes place with human individuals. In analyzing CES data, identifiably information such as names and social security numbers of the respondents will not be used. The above shown potential bias will be avoided through the following; employment status, and the accessibility to the public transport (U.S Bureau of Labor Statistics, n.d.).

Study Selection Criteria

Sampling is targeted at low income families (with incomes below 200 percent of the federal poverty level). Information in the database ranges from 2015 (considered time ‘t-2’) to 2021 (time ‘t+1’) as required to undertake DiD analysis.

3. Statistical Analysis Plan

The primary analysis will apply Difference-in-Differences (DiD) regression to compare change in gasoline consumption between treatment that is low income female headed households in California and the comparison group of households in states without carbon pricing such as Texas or Florida. DiD successfully captures the effect of carbon pricing on gasoline consumption by including other factors connecting carbon pricing to gasoline expenditures.

Model Specification

The following Difference-in-Differences (DiD) model will be used: $GasolineExpenditure_{it} = \beta_0 + \beta_1 * PostTax_t + \beta_2 * TreatmentGroup_i + \beta_3 * (PostTax_t \times TreatmentGroup_i) + \beta_4 * HouseholdGender_i + \beta_5 * (PostTax_t \times HouseholdGender_i) + \beta_6 * UrbanRural_i + \beta_7 * Controls_{it} + \epsilon_{it}$, where the dependent variable is $GasolineExpenditure_{it}$, representing the annual gasoline expenditure for household *i* at time *t*. The independent variables include $PostTax_t$, which indicates the post-policy period; $TreatmentGroup_i$, an indicator for California households; $HouseholdGender_i$, an indicator for female-headed households; $UrbanRural_i$, which identifies rural areas; and $Controls_{it}$, a vector of control variables such as income level, household size, employment status, education level, and vehicle ownership. Additionally, interaction terms include $(PostTax_t \times TreatmentGroup_i)$, capturing the overall effect of the policy, and $(PostTax_t \times HouseholdGender_i)$, assessing the differential impact by gender.

Key Input/Output Variables

The key input variables for the study include policy implementation, household gender, urban/rural status, and control variables such as income level, household size, employment status, education level, and vehicle ownership. The output variable is the resulting changes **in gasoline expenditure.**

Justification of Statistical Choices

The statistical choices in this study are justified as follows: the Difference-in-Differences (DiD) method is appropriate for evaluating the impact of policy changes, particularly when randomization is not feasible. Control variables are included to adjust for potential confounding

factors that could influence gasoline consumption, ensuring a more accurate assessment of the policy's effect.

Validation and Assumption Checks

Validation and assumption checks will be conducted to ensure the reliability of the results. The parallel trends assumption will be tested by comparing pre-policy trends between the treatment and control groups. Sufficient data points will be ensured through a power analysis to detect significant effects. Additionally, robustness checks will be performed, including sensitivity analysis by exploring different income thresholds (e.g., 150%, 250% of the Federal Poverty Level) to test the stability of the results, as suggested by Caliendo and Kopeinig (2008). Placebo tests will also be applied by using the model on pre-policy periods, where no policy change occurred, to check for any spurious associations.

Sensitivity Analysis

To address the concern of sensitivity testing, other definitions for low-income families will be explored including families closely below 200%, 300% of FPL. This preserves the comparability of the results regardless of the income differentials (Caliendo & Kopeinig, 2008).

Handling Statistical Issues

To handle potential statistical issues, robust standard errors clustered at the state level will be used to correct for heteroskedasticity and autocorrelation, following the approach of Blundell and Dias (2009). Multicollinearity will be addressed by checking variance inflation factors (VIF) to ensure that the independent variables are not highly correlated, thereby maintaining the integrity of the regression model.

Post Hoc Tests

Post Hoc tests will include subgroup analyses, where separate models will be run for urban vs. rural households and female-headed vs. male-headed households to explore potential heterogeneous effects. Additionally, interaction effects will be examined by introducing further interaction terms to investigate the combined influence of gender and geographical location on gasoline expenditure.

Conclusion

This research design outlines a methodological framework to investigate the socioeconomic impacts of California's carbon pricing policies on gasoline expenditures among low-income households. By employing robust statistical methods and utilizing comprehensive datasets, the study aims to provide insights that can inform equitable policy development.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. Retrieved from https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion

Blundell, R., & Dias, M. C. (2009). *Alternative approaches to evaluation in empirical microeconomics*. Journal of Human Resources, 44(3), 565-640. Retrieved from https://www.jstor.org/stable/20648911

Caliendo, M., & Kopeinig, S. (2008). *Some Practical Guidance for the Implementation of Propensity Score Matching*. Journal of Economic Surveys, 22(1), 31-72. Retrived from

https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-6419.2007.00527.x?msockid=3066027e11916f70355411d410396e45

California Air Resources Board. (n.d.). *Cap-and-Trade Program*. Retrieved from https://ww2.arb.ca.gov/our-work/programs/cap-and-trade-program

Marron, D. B., Toder, E. J., & Austin, L. (2015). *Taxing Carbon: What, Why, and How*. Tax Policy Center. Retrieved from https://www.taxpolicycenter.org/publications/taxing-carbon-what-why-and-how

Ohlendorf, N., Jakob, M., Minx, J. C., Schröder, C., & Steckel, J. C. (2021). *Distributional Impacts of Carbon Pricing: A Meta-Analysis*. Environmental and Resource Economics, 78(1), 1–42. Retrieved from https://link.springer.com/article/10.1007/s10640-020-00521-1

U.S. Bureau of Labor Statistics. (n.d.). *Consumer Expenditure Survey (CES) – Restricted-Use Microdata*. Retrieved from https://www.bls.gov/rda/home.htm

U.S. Census Bureau. (n.d.). *American Community Survey (ACS) Public Use Microdata Sample (PUMS)*. Retrieved from https://www.census.gov/programs-surveys/acs/microdata.html

U.S. Energy Information Administration. (n.d.). *Gasoline and Diesel Fuel Update*. Retrieved from https://www.eia.gov/petroleum/gasdiesel/