

Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption

EEA-ESEM Congress, August 24th, 2022

Mateus Souza

Universidad Carlos III de Madrid, Department of Economics

Motivation

- Increased data availability for research in economics
 - May lead to more nuanced insights about the problem at hand
 - But how should we control for confounders? How do we estimate heterogeneity?
- “Traditional” approaches were not designed for heterogeneity:
 - Recent critiques of Two-Way Fixed Effects (TWFE) for staggered difference-in-differences (Borusyak and Jaravel, 2017; Goodman-Bacon, 2021)

$$Y_{i,t} = \beta D_{i,t} + \alpha_i + \alpha_t + u_{i,t}$$

- Some solutions:
 - Sun and Abraham (2021): no covariates
 - Callaway and Sant’Anna (2021); Wooldridge (2021): only allow *pre-determined* covariates

Contribution

- I propose to **predict untreated counterfactuals** with machine learning techniques
- **Allows the inclusion of time-varying covariates**
- **Fixes the weighting** issue from TWFE

My proposal is similar to parallel work from Borusyak, Jaravel, and Spiess (2021)

- They propose OLS rather than ML for counterfactual predictions
- ML algorithms are **“agnostic” about the functional forms** of control variables
- ML results in **more efficient estimation** of treatment effects

Proposed approach

The treatment effect for unit i at time t is $b_{i,t} = Y_{i,t}(1) - Y_{i,t}(0)$
But we only observe $Y_{i,t} = Y_{i,t}(1)$.

My proposal is based on the estimation of a counterfactual function $g(\mathbf{X}_{i,t})$, so that we can impute $Y_{i,t}(0) = g(\mathbf{X}_{i,t}(0)) + \varepsilon_{i,t}$

Step 1: build and select a model to estimate $g()$

- using pre-treatment data
- machine learning techniques for better prediction accuracy

Step 2: Estimate the full distribution of treatment effects:

$$\hat{b}_{i,t} = Y_{i,t} - \hat{g}(\mathbf{X}_{i,t})$$

Step 3: Estimate causal parameters of interest by taking conditional averages of the treatment effects

Causal parameters of interest

- Average Treatment Effect on the Treated (ATT):

$$ATT = \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0) | D_{i,t} = 1]$$

- ATT for r periods of exposure relative to the treatment time ($t = q_i$):

$$ATT(r) = \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0) | D_{i,t} = 1, t - (q_i - 1) = r]$$

- Conditional ATT for groups defined based on a set of covariates $\mathbf{X}_{i,t}$:

$$CATT(\mathbf{c}) = \mathbb{E}[Y_{i,t}(1) - Y_{i,t}(0) | \mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1]$$

Key assumptions

Assumption 1: No anticipatory effects.

$$Y_{i,t} = Y_{i,t}(0), \text{ for all } i, t < q_i$$

Assumption 2: Covariates are not affected by the treatment.

$$\mathbf{X}_{i,t} = \mathbf{X}_{i,t}(0) = \mathbf{X}_{i,t}(1), \text{ for all } t$$

Assumption 3: Stability of the counterfactual function.

$$\mathbb{E}[Y_{i,t}(0) | \mathbf{X}_{i,t}, D_{i,t} = 1] = g(\mathbf{X}_{i,t}(1))$$

For CATT, we need Assumption 3 to hold for each group defined by $\mathbf{X}_{i,t} = \mathbf{c}$.

Illustration of the Approach

- Outcome: daily air pollution (PM_{10}) concentration measured at ~ 400 monitors across Spain, from 2014 to 2019
- Rich set of covariates:
 - Daily weather realizations (temp, precip, wind speed, etc.)
 - Daily national-level power generation by fuel type
 - Annual national-level fire activity
 - Characteristics about pollution monitors
 - Date fixed effects

Semi-Synthetic Simulations:

- Assign random treatment dates for each air pollution monitor
 - Staggered treatment
- Force a PM_{10} reduction for post-treatment observations
 - Can be thought of as staggered implementation of low-emissions zones across Spain

Illustration of the Approach

- **Step 1:** build and select a model to estimate $g()$
 - Focus on **tree-based** methods (gradient boosted trees);
 - Other algorithms can be considered (LASSO, neural networks, etc.)
 - use *cross-validation* to assess out-of-sample accuracy
 - assess identifying assumptions
- **Step 2:** Estimate the full distribution of treatment effects:
$$\hat{b}_{i,t} = Y_{i,t} - \hat{g}(\mathbf{X}_{i,t})$$
- **Step 3:** Estimate the causal parameters of interest by taking conditional averages of the treatment effects

Step 1: Model selection

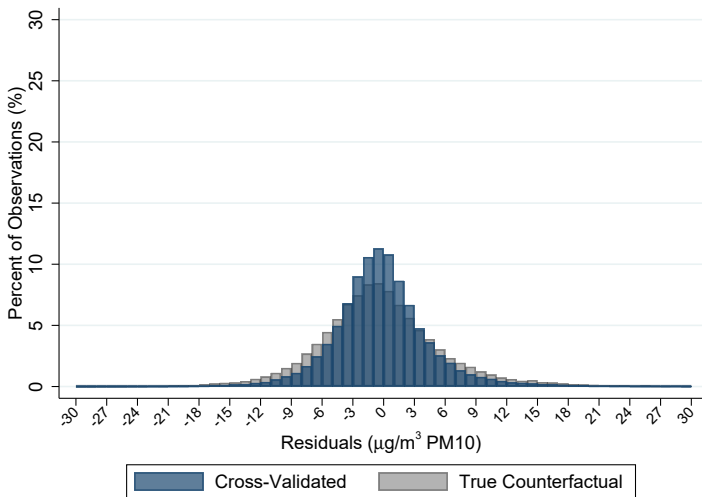
Tried 8 configurations of XGBoost (Chen and Guestrin, 2016)

Model ID	Number of Trees	Max Tree Depth	Min Obs per Node	Shrinkage	In Sample RMSE	Cross-Validated RMSE
1	2000	10	20	0.05	2.793	6.574
2	3000	10	20	0.05	2.149	6.524
3	2000	30	20	0.05	0.203	6.852
4	3000	30	20	0.05	0.068	6.853
5	2000	10	60	0.05	3.818	6.686
6	3000	10	60	0.05	3.248	6.601
7	2000	30	60	0.05	1.029	6.618
8	3000	30	60	0.05	0.598	6.627

Selected model ID 2, with **lowest cross-validated RMSE**.

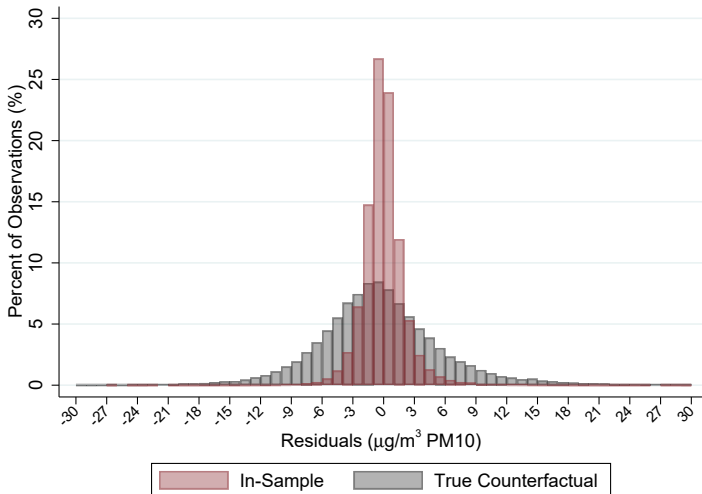
Step 1: Out-of-sample accuracy

Check the distribution of cross-validated residuals



Step 1: Out-of-sample accuracy

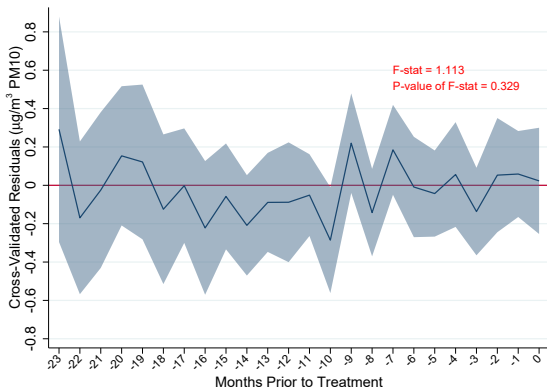
How in-sample residuals underestimate the variance of errors



Step 1: Tests for validity of assumptions

- Event-study regression, similar to analyzing “pre-trends”
- Perform F-test for joint significance of β_r

$$\hat{\varepsilon}_{i,t}^{CV} = \sum_{r \leq 0} \beta_r \mathbb{1}[r = t - (q_i - 1)] + u_{i,t}, \text{ for all } t < q_i$$



Approach

- **Step 1:** build and select a model to estimate $g()$
- **Step 2:** Estimate the full distribution of treatment effects:
$$\hat{b}_{i,t} = Y_{i,t} - \hat{g}(\mathbf{X}_{i,t})$$
- **Step 3:** Estimate the causal parameters of interest by taking conditional averages of the treatment effects

Approach

- **Step 1:** build and select a model to estimate $g()$
- **Step 2:** Estimate the full distribution of treatment effects:
$$\hat{b}_{i,t} = Y_{i,t} - \hat{g}(\mathbf{X}_{i,t})$$
- **Step 3:** Estimate the causal parameters of interest by taking conditional averages of the treatment effects; and estimate standard errors
 - Estimates from my approach are compared to standard TWFE and OLS imputation
 - Will assess bias and efficiency (standard errors)

Inference

Building on Borusyak, Jaravel, and Spiess (2021), I propose estimating the variance as:

$$\hat{\sigma}_{cv}^2 = \sum_i \left(\sum_{t; D_{i,t}=0} \gamma_{i,t} \hat{\varepsilon}_{i,t}^{cv} + \sum_{t; D_{i,t}=1} \gamma_{i,t} \tilde{\varepsilon}_{i,t} \right)^2$$

- $\hat{\varepsilon}_{i,t}^{cv}$ are cross-validated residuals from the first (prediction) step
- $\tilde{\varepsilon}_{i,t} = \hat{b}_{i,t} - \hat{b}$ are treatment effect deviations from an average
- $\gamma_{i,t}$ are the weights of each observation
- \hat{b} is up to the researcher
- A conservative choice is a single \hat{b} for the full post-treatment sample. Alternatively, we may take averages across subsamples for which heterogeneity is expected.
- Even more conservative: bootstrap the whole process.

Step 3: Estimating *ATT* under dynamic effects

Panel A: weaker effects over time (5% decrease per semester)

Panel B: stronger effects over time (5% increase per semester)

Panel A: Treatment Effects Decreasing Over Time						
	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
\widehat{ATT}	-3.1840	-3.1350	-4.9006	-4.5487	-3.0367	-2.7432
Standard Errors		(0.1281)	(0.2891)	(0.3457)	(0.2548)	(0.2118)
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Panel B: Treatment Effects Increasing Over Time						
\widehat{ATT}	-2.2199	-2.1709	-0.9287	-0.4429	-2.0691	-1.6989
Standard Errors		(0.1244)	(0.2869)	(0.3334)	(0.2563)	(0.2153)
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station \times Month FE		NA	No	Yes	No	Yes
Day of sample \times Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Average PM₁₀ concentration is about 20 $\mu\text{g}/\text{m}^3$

Step 3: Estimating $ATT(r)$

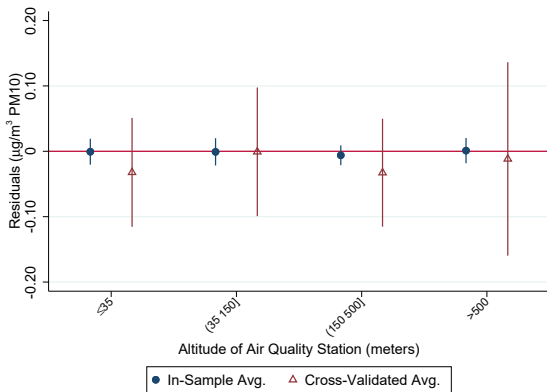
Simulating weaker effects over time (5% decrease per semester)

	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
$\widehat{ATT}(1)$: Semester 1	-4.2808	-4.2578	-4.5397	-4.0435	-4.2057	-3.5697
Standard Errors		(0.1568)	(0.2967)	(0.3733)	(0.2386)	(0.3176)
$\widehat{ATT}(2)$: Semester 2	-3.1807	-3.2693	-3.4408	-2.9358	-3.0351	-2.8437
		(0.1814)	(0.4054)	(0.4179)	(0.2969)	(0.3106)
$\widehat{ATT}(3)$: Semester 3	-2.1932	-2.0567	-2.3363	-1.7575	-1.9510	-1.6892
		(0.2619)	(0.5098)	(0.5406)	(0.4209)	(0.4430)
$\widehat{ATT}(4)$: Semester 4	-1.1405	-0.8112	-1.3540	-0.2796	-0.8908	-0.8406
		(0.3677)	(0.5721)	(0.5758)	(0.5169)	(0.4410)
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station \times Month FE		NA	No	Yes	No	Yes
Day of sample \times Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Average PM_{10} concentration is about $20 \mu\text{g}/\text{m}^3$

Setup for *CATT*

- More complex setting where treatment effects decrease over time and with altitude of pollution monitors
- First, we check the validity of the assumptions: are cross-validated errors correlated with altitude of stations?



Step 3: Estimating *CATT*

More complex setting where treatment effects decrease over time and with altitude of pollution monitors

	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
\widehat{CATT} : Altitude \leq 35m	-7.0897	-6.9100	-8.3251	-7.9242	-6.6679	-6.4911
Standard Errors		(0.2885)	(0.4644)	(0.4796)	(0.4471)	(0.4048)
\widehat{CATT} : 35m < Altitude \leq 150m	-4.1586	-4.0815	-5.5698	-5.4877	-4.6664	-4.2504
		(0.1982)	(0.3555)	(0.4692)	(0.5500)	(0.2942)
\widehat{CATT} : 150m < Altitude \leq 500m	-1.2035	-1.1198	-2.8135	-2.7120	-1.0951	-0.4605
		(0.2662)	(0.4679)	(0.6323)	(0.4719)	(0.5502)
\widehat{CATT} : Altitude > 500m	0.0000	-0.2662	-1.3353	-0.2348	0.8703	1.2154
		(0.3053)	(0.4479)	(0.8139)	(0.4615)	(0.4980)
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station \times Month FE		NA	No	Yes	No	Yes
Day of sample \times Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Average PM₁₀ concentration is about 20 $\mu\text{g}/\text{m}^3$

Conclusions

- I propose an approach for heterogeneous treatment effect estimation under staggered adoption
- I show how to build a model, based on pre-treatment data, to predict untreated counterfactuals
- I also show how machine learning can be leveraged in this setting for more efficient estimation

With the real data application (Weatherization Assistance Program):

- I find substantial heterogeneity in energy savings, depending on levels and types of investments
- In particular, wall insulation and furnace replacements are the measures associated with highest energy savings

Thank You!

Comments? Feedback? Questions?

mateus.nogueira@uc3m.es





<http://energyecolab.uc3m.es/>



European Research Council
Established by the European Commission

This Project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 772331).

References I

-  Borusyak, Kirill and Xavier Jaravel (2017). “Revisiting Event Study Designs”. *SSRN Working Paper*.
-  Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021). “Revisiting Event Study Designs: Robust and Efficient Estimation”. *arXiv Working Paper 2108.12419*.
-  Callaway, Brantly and Pedro H.C. Sant’Anna (2021). “Difference-in-Differences with Multiple Time Periods”. *Journal of Econometrics* 225(2), pp. 200–230.
-  Chen, Tianqi and Carlos Guestrin (2016). “XGBoost: A Scalable Tree Boosting System”. *arXiv:1603.02754*.
-  Goodman-Bacon, Andrew (2021). “Difference-in-differences with variation in treatment timing”. *Journal of Econometrics* 225(2), pp. 254–277.

References II



Sun, Liyang and Sarah Abraham (2021). “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects”. *Journal of Econometrics* 225(2). Themed Issue: Treatment Effect 1, pp. 175–199.

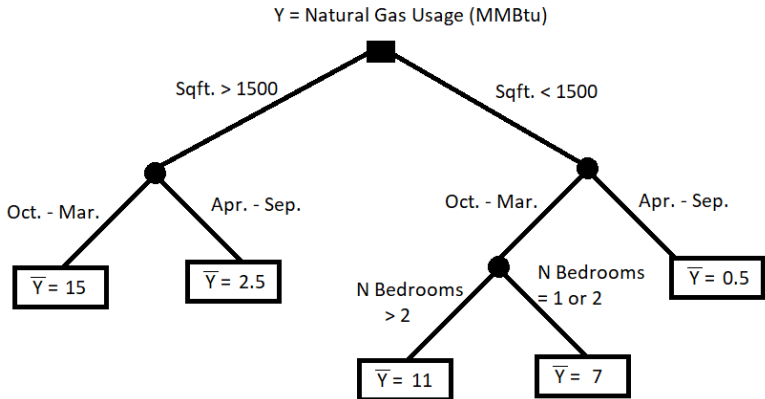


Wooldridge, Jeff (2021). “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators”. *SSRN Working Paper 3906345*.

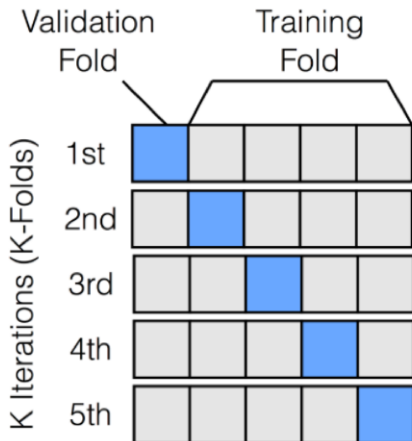
Why Machine Learning?

- Machine learning algorithms are designed for *accurate predictions* (measured with root-mean squared error, RMSE)
- Allow us to flexibly and systematically consider a wide range of variables, functional forms, and interaction terms
- Algorithms are “agnostic” about which variables to include
- Special concern about **out-of-sample** performance
- This paper focuses on **tree-based** methods
 - Specifically, gradient boosted trees
 - Other algorithms can be considered (LASSO, neural networks, etc.)

Illustration of a Single Regression Tree



Cross-Validation

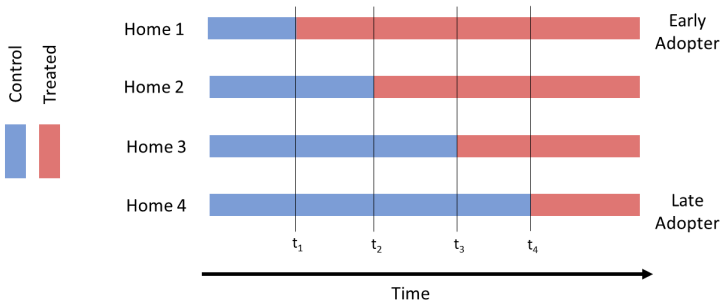


Assess cross-validated residuals ($\hat{\epsilon}_{i,t}^{CV}$) for observations excluded from the training fold

The problem of TWFE

$$Y_{i,t} = \beta D_{i,t} + \alpha_i + \alpha_t + u_{i,t}$$

- The problem:
 - β will capture a variance-weighted average of effects over time and across units
 - can cause near-term bias
 - and “mid-adopters” will receive greater weights



Estimators for causal parameters of interest

Step 3: Estimate the causal parameters of interest

$$\widehat{ATT} = \frac{\sum_{i=1}^I \sum_{t=1}^T \hat{b}_{i,t} \mathbb{1}\{D_{i,t} = 1\}}{\sum_{i=1}^I (T - (q_i - 1)) \mathbb{1}\{q_i \leq T\}}$$

$$\widehat{ATT}(r) = \frac{\sum_{i=1}^I \hat{b}_{i,t} \mathbb{1}\{t - (q_i - 1) = r\}}{\sum_{i=1}^I \mathbb{1}\{t - (q_i - 1) = r\}}, \quad r > 0$$

$$\widehat{CATT}(\mathbf{c}) = \frac{\sum_{i=1}^I \sum_{t=1}^T \hat{b}_{i,t} \mathbb{1}\{D_{i,t} = 1\} \mathbb{1}\{\mathbf{X}_{i,t} = \mathbf{c}\}}{\sum_{i=1}^I (T - (q_i - 1)) \mathbb{1}\{q_i \leq T\} \mathbb{1}\{\mathbf{X}_{i,t} = \mathbf{c}\}}$$

Real data application

Setting:

- Weatherization Assistance Program
- Large household energy efficiency program
- Fully subsidizes furnace repair/replacements, air sealing, wall insulation and others

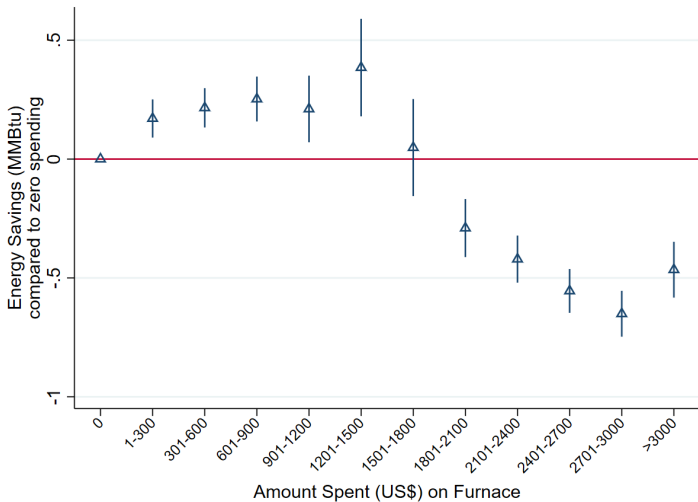
Data:

- Monthly energy billing data from thousand of homes treated in Illinois
- Program administrative data: household demographics, housing structure, costs of upgrades

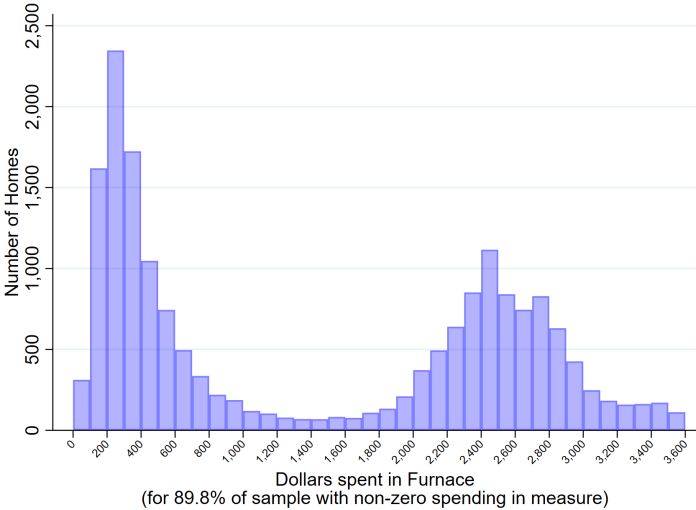
Objectives:

- Estimate upgrade-specific effects
- Provide insight on which upgrades are most cost-effective

Real Data Results – Furnace



Distribution of Furnace Spending



Real-Data Results – Wall Insulation

