Bayesian Approaches to Capturing Heterogeneous Effects in Difference-in-Differences and Triple Differences: Analyzing Pre-Post COVID Household Income Changes

Introduction

In natural experiments, key challenges in difference-in-differences (DiD) analysis include addressing heterogeneous treatment effects across subgroups (Card & Krueger, 1994; Cintron et al., 2022) and over multiple time points (Callaway & Sant'Anna, 2021; Meer & West, 2024). Methods like two-way fixed effects (de Chaisemartin & d'Haultfoeuille, 2020) and staggered DiD (Callaway & Sant'Anna, 2021; Baker et al., 2022) help us understand these heterogeneous effects. Triple differences (Gruber, 1994; Muralidharan & Prakash, 2017) further complicate the analysis, as they typically require large samples and multiple models due to three-way interaction terms. While many studies have evaluated policy and economic shock effects, recent papers (e.g., Callaway & Sant'Anna, 2021; Griffin, 2021) have started incorporating simulation studies to rigorously assess method robustness. Additionally, the use of Bayesian hierarchical methods in the context of difference-in-differences (DiD) analysis has gained attention (Normington, 2019, Gelman 2021, Normington, 2021).

In this paper, we focus on COVID-19's impact on household income by investigating job disruptions (e.g., layoffs, wage cuts, hour cuts) across income-level subgroups (low, middle, high) using the Future of Families and Child Wellbeing Study (FFCWS) data. We develop a Bayesian approach to estimate heterogeneous effects in difference-in-differences and compute triple differences without relying on three-way interaction terms. This approach enables flexible modeling of group-specific effects without requiring large sample sizes typically needed for traditional interaction-based models. We evaluate the performance of our method using simulated data that mimics the FFCWS dataset under varying sample sizes. We report both the absolute and percentage changes in household income, with the latter measured using the log-transformed outcome, to assess the robustness of the method.

Data Explanation and Simulation

The FFCWS Study that is based on a stratified, multistage sample of 4,898 children born in large U.S. cities (population over 200,000) between 1998 and 2000, where births to unmarried mothers were oversampled by a ratio of 3 to 1 with the inclusion of a large number of Black, Hispanic, and low-income families. Follow-up interviews were conducted across seven waves when children were approximately ages 1, 3, 5, 9, 15, and 22. For our analysis, we use data from Wave 6 (2018) and Wave 7 (conducted between October 2020 and January 2024). Motivated by the FFCWS study, the simulated data consists of pre- and post-COVID household income variable, y, a job disruption variable, z, a time variable, t, and an income-level group variable, g, that is based on pre-COVID household income.

To reflect the size of the available FFCWS observations (2277), we generate a data set of 2500, which makes 5000 for pre and post time points. The generation of the simulated data as follows:

The income before COVID for each individual is drawn from a normal distribution with a log-mean of \$58k and a log-standard deviation of 1.5. To obtain the actual income values that are between \$0 to \$250k, we then take the exponential of the drawn values from this distribution.

$$y_0 \sim N\big(\log(58k),\log(1.5)\big)$$

The variable t indicates whether the observation corresponds to the pre- or post-COVID period.

$$t = \{0,1\}$$

The job disruption variable z is a binary treatment indicator of whether an individual experienced a job disruption due to COVID (e.g., layoffs, wage cuts, or hour cuts). The choice of a Bernoulli distribution with a 50% probability is based on the fact that about the half of the FFCWS observations experience a job disruption across all three income-level subgroups.

$$z \sim \text{Bernoulli}(0.5)$$

For income level grouping, the individuals are divided into three income-level groups: Low income (income below the 50th percentile), middle income (income between the 50th and 75th percentiles), high income (income above the 75th percentile). The choice to use the 50th percentile for defining the low-income group is based on observed differences in the DiD results between the 0–25th and 25–50th percentiles, which were found to be minimal.

$$g = \{1,2,3\}$$

The post-COVID income for each individual is defined as:

$$y_1 = \begin{cases} y_0 + \beta_t + \epsilon_g, & \text{if } z_i = 0 \\ y_0 + \beta_t + \beta_z + \delta_g + \epsilon_g, & \text{if } z_i = 1 \end{cases}$$

where $\beta_t = 12k$ represents the time effect (post-COVID), $\beta_z = -1k$ represents the treatment effect, and $\delta_g = \delta + \delta'_g = -2k + \{-1k, -4k, -10k\}_g$ represents the difference-in-differences (DiD) effect for each income-level subgroup. The term δ captures the baseline DiD effect, which is the common treatment effect across all groups, while δ'_g accounts for subgroup-specific variations in the treatment effect (i.e., low, middle, and high-income groups). The inclusion of both components allows for a more nuanced understanding of how job disruptions impact household income differently across income levels, reflecting heterogeneous treatment effects. Finally, $\epsilon_g \sim N\big(0, \sigma_g\big)$ where $\sigma_g = \{5k, 15k, 8k\}$ represents the error terms for each income group, with variability reflecting different levels of uncertainty in post-COVID household income. These values were chosen to approximate the observed variability in the FFCWS data. Importantly, the error variance does not depend on the treatment status z, but rather on the income group identities. This assumption captures the heterogeneity in income levels, reflecting the real-world complexity. This modeling decision is crucial because it aligns with our objective to estimate heterogeneous treatment effects across subgroups while mitigating overfitting or incorrect model specification. By incorporating these group-specific error terms,

our Bayesian model is better equipped to handle real-world complexities, providing more accurate and robust estimates.

Methods

To evaluate our simulated data, we initially considered a simple time-treatment interaction linear regression model as the smallest possible specification and a more complex time-treatment-group interaction model as the largest. After fitting these and several intermediate models, we observed consistent challenges in capturing the heterogeneity of treatment effects across subgroups. While larger interaction models incorporate more flexibility, they suffer from biased estimates of the difference-in-differences (DiD) effects due to overparameterization and potential overfitting, especially in smaller samples. In contrast, the time-treatment interaction model accurately captures the overall time and treatment effects but fails to address subgroup-specific variations.

These findings motivated the development of a hierarchical Bayesian model to incorporate the relational structure between subgroups and better estimate heterogeneity in DiD effects. This approach also leverages subgroup-specific time effects to reflect heterogeneous variability, alongside random intercepts to account for unobserved individual-level differences. By embedding these features in a hierarchical framework, the model balances flexibility and parsimony, effectively mitigating the limitations of traditional interaction-heavy models. This hierarchical approach, implemented in Stan, enables the incorporation of multiple group-customized likelihoods or models within a unified framework, eliminating the need to fit separate models for each group. This not only enhances computational efficiency but also more accurately captures relationships across groups by leveraging shared information. The model dynamically adapts to subgroup-specific variations, providing robust and unbiased estimates of heterogeneous treatment effects while preserving the relational structure between groups. The model is as follows:

$$y_i \sim N\left(\alpha + \beta_t t_i + \beta_z z_i + \delta_{g_i} t_i z_i + \eta_{g_i} t_i + r_i, \sigma\right)$$

The group-specific treatment effect is defined as $\delta_g = \delta + \delta'_g$, where the deviations δ'_g satisfy $\delta'_3 < \delta'_2 < \delta'_1$, based on insights from preliminary analysis. To enforce this ordering, we parameterize the deviations as follows:

$$\delta'_1 = \delta_{\text{raw}_1} \;\; \delta'_2 = \delta'_1 + \delta_{\text{raw}_2} \;\; \delta'_3 = \delta'_2 + \delta_{\text{raw}_3}$$

where δ_{raw_1}, δ_{raw_2}, and δ_{raw_3} are unconstrained negative parameters sampled from appropriate priors. For the group-specific time effects, we model η_g as:

$$\eta_g = \sigma_\eta \Sigma_\eta \eta_{\text{raw}}$$

where $\sigma_\eta = \left(\sigma_{\eta_1}, \sigma_{\eta_2}, \sigma_{\eta_3}\right)^T$ is a vector of standard deviations for group-specific effects. $\Sigma_\eta = L_\eta L_\eta^T$ is the $3x3$ covariance matrix constructed by the Cholesky factor L_η, which is drawn from an LKJ prior to model the correlation structure between groups. $\eta_{\text{raw}} = \left(\eta_{\text{raw}_1}, \eta_{\text{raw}_2}, \eta_{\text{raw}_3}\right)^T$

represents standardized latent variables, which are scaled and stretched by σ_η and Σ_η to produce the group-specific time effects. The priors as follows:

$$
\begin{aligned}
\alpha &\sim N(60k, 10k) \\
\beta_t &\sim N(12k, 1k) \\
\beta_z &\sim N(-1.2k, 0.5k) \\
\delta &\sim N(0, 4k) \\
\delta_{\text{raw}_1} &\sim N(0, \sigma_\delta) \\
\delta_{\text{raw}_2} &\sim N(0, \sigma_\delta) \\
\delta_{\text{raw}_3} &\sim N(0, \sigma_\delta) \\
\sigma_\delta &\sim N(8k, 4k) \\
\sigma_\eta &\sim N(0, 2k) \\
L_\eta &\sim \text{lkj\_corr\_cholesky}(2) \\
\eta_{raw} &\sim N(0, 1) \\
r &\sim N(0, \sigma_r) \\
\sigma_r &\sim N(0, 1k) \\
\sigma &\sim N(0, 10k)
\end{aligned}
$$

The prior for the baseline outcome α is set to the mean of pre-COVID household income, with a standard deviation chosen to reflect reasonable uncertainty around this estimate. The prior means for β_t and β_z are informed by the estimated coefficients from a time-treatment interaction model, which can be easily obtained from real-world data. For δ, we assign a non-informative prior to reflect minimal prior knowledge. The raw deviations δ_{raw_i} share a common standard deviation, which allows for flexibility in their relative magnitudes while maintaining consistency across groups. For σ_δ, we set a relatively large standard deviation to account for the uncertainty surrounding this value. For L_η, we use a weakly informative prior with $\eta = 2$, which shrinks correlations towards zero but still allows for non-zero correlations, providing some flexibility in modeling group-specific relationships. For the remaining parameters, we employ non-informative priors with reasonable standard deviations to minimize the influence of prior beliefs on the model, allowing the data to dominate the inference.

Results

Simulation Study. For the simulation study results, we present posterior mean estimates along with one-standard-deviation credible intervals. Additionally, we compare the performance of our Bayesian model with four time-treatment interaction regression models, using the full dataset for time and treatment coefficients and subsamples for estimating difference-in-differences effects.

The Bayesian model demonstrates strong performance in capturing both time and treatment effects and heterogeneous group-specific impacts. The difference-in-differences estimates, δ_g, are accurately recovered, with posterior means (-2.9k, -5.7k, -11.3k) closely approximating the true effects (-3.0k, -6.0k, -12.0k), and the credible intervals consistently encompassing the true parameters. Compared to the OLS model, the Bayesian model exhibits either less bias or

narrower one standard deviation intervals for all parameters. This highlights its superior performance in providing more precise and reliable estimates of the heterogeneous difference-in-differences effects.

	β_t	β_z	δ_1(DiD)	δ_2(DiD)	δ_3(DiD)
Population	12.0k	-1.0k	-3.0k	-6.0k	-12.0k
Bayesian	12.4k (0.9k)	-1.2k (0.4k)	-3.2k (1.1k)	-6.0k (1.4k)	-12.6k (1.8k)
M-Weighted Bayesian	**12.4k (0.9k)**	**-1.2k (0.4k)**	**-3.1k (1.0k)**	**-5.8k (1.4k)**	**-12.3k (1.8k)**
D-Weighted Bayesian	12.5k (1.0k)	-0.8k (0.1k)	-3.0k (0.2k)	-6.5k (0.3k)	-14.0k (0.3k)
OLS	11.8k (1.1k)	-1.2k (1.1k)	-4.0k (0.8k)	-6.9k (1.3k)	-13.0k (2.7k)
M-Weighted OLS	11.8k (1.1k)	-1.2k (1.1k)	-4.0k (0.8k)	-6.7k (1.3k)	-13.1k (2.7k)
D-Weighted OLS	12.1k (1.1k)	-0.8k (1.1k)	-4.2k (0.8k)	-7.7k (1.3k)	-13.3k (2.8k)

To evaluate the performance of triple differences, we compare the results of the Bayesian model presented above to those of three separate OLS models with three-way interaction terms.

	DDD (2,1)	**DDD (3,1)**	**DDD (3,2)**
Population	-3.0k	-9.0k	-6.0k
Bayesian	-2.8k	-9.4k	-6.6k
M-Weighted Bayesian	-2.7k	-9.2k	-6.5k
OLS	-2.8k	-9.0k	-6.1k
M-Weighted OLS	-2.6k	-9.0k	-6.4k

The OLS results for triple differences align with the differences observed in the two-way interaction models. While the individual estimates are highly biased, the differences between them provide reasonably accurate estimates. Similarly, the Bayesian model results yield reliable estimates for triple differences, with far less bias compared to the OLS models. Among the OLS approaches, the model-based weighted Bayesian model provides the least biased results, followed by the non-weighted model and the design-based model, consistent with the findings of Lee and Gelman (2025).

Among the Bayesian approaches, while the non-weighted Bayesian model performs well, the model-based weighted Bayesian model outperforms it in estimating both difference-in-differences and triple differences. Furthermore, when comparing the design-based weighted Bayesian model to the model-based weighted Bayesian model, the latter demonstrates superior performance, producing less biased estimates. However, the design-based weighted Bayesian model produces significantly smaller standard errors, which may seem incorrect compared to those from other models. Aside from this case, overall, the model-based weighted approaches yield the lowest standard errors, albeit with only a small difference, in both OLS and Bayesian frameworks compared to the other two approaches. Among them, the model-based weighted Bayesian method provides the most robust and least biased estimates, making it the most reliable approach.

FFCWS Study. The same model is used for the FFCWS study with slightly adjusted priors, which were determined in the same manner as those used in the simulation study. The only big

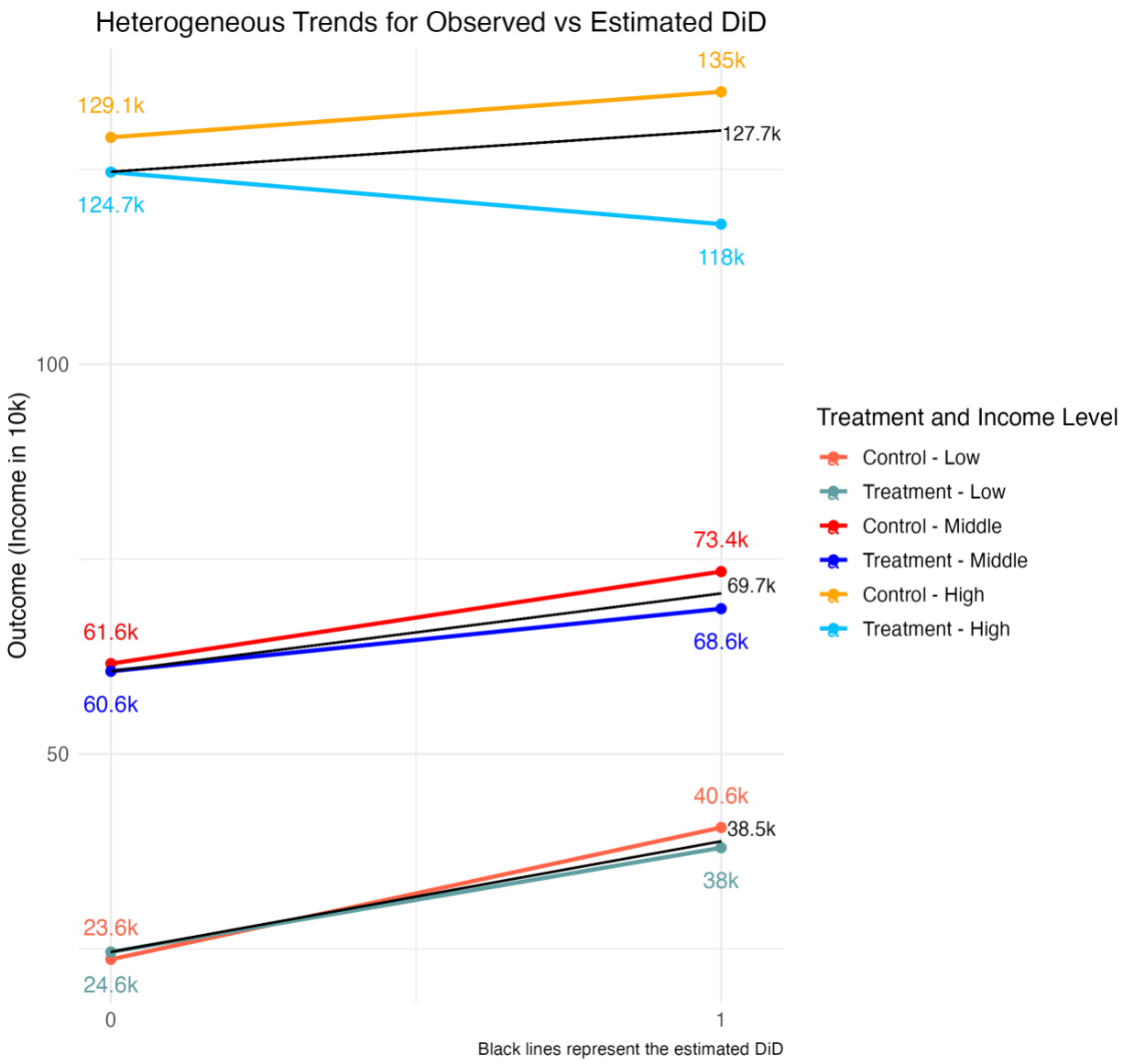
change in priors is $\sigma_\delta \sim N(9k, 4k)$. We set the prior with a mean representing the largest plausible value for δ_{raw_i} because the estimated DiD for the high-income group was much smaller than the expected value based on from our raw data and the estimated value from OLS. The following results present the change in post-COVID household income dollars.

	$\boldsymbol{\beta_t}$	$\boldsymbol{\beta_z}$	$\boldsymbol{\delta_1}$**(DiD)**	$\boldsymbol{\delta_2}$**(DiD)**	$\boldsymbol{\delta_3}$**(DiD)**
Bayesian	13.0k (11.2, 14.8)	-4.9k (-5.7, -4.1)	-2.1k (-3.2, -1.0)	-3.7k (-5.4, -2.0)	-7.3k (-10.3, -4.3)
OLS	12.7k (10.5, 14.9)	-5.3k (-7.4, -3.2)	-3.5k (-5.6, -1.4)	-3.7k (-7.7, 0.3)	-12.5k (-18.5, -6.5)

We now present the results using the log-transformed outcome, which reflects the percentage change in post-COVID household income.

	$\boldsymbol{\beta_t}$	$\boldsymbol{\beta_z}$	$\boldsymbol{\delta_1}$**(DiD)**	$\boldsymbol{\delta_2}$**(DiD)**	$\boldsymbol{\delta_3}$**(DiD)**
Bayesian	9.7 (-0.3, 19)	-3.1 (0.4, -6.7)	-7.7 (-2.8, -0.8)	-5.5 (-4.2, -1.4)	-12.2 (-19.2, -5.3)
OLS	7.0 (2.0, 12)	-6.7 (-1.7, -11.7)	-8.7 (-18, 0.5)	-3.4 (-12.9, 6.1)	-17.6 (-25.5, -9.7)

The time effects indicate an average increase of $13k in household income, corresponding to a 9.7% increase between Wave 6 and Wave 7, which aligns with the typical annual income growth of approximately 3% in the U.S. The treatment effect reveals a significant average income reduction of $4.9k due to job disruptions, emphasizing the financial strain caused by COVID-related income losses. The difference-in-differences (DiD) estimates show heterogeneous impacts across income groups, with income reductions becoming progressively larger: -$2.1k for the low-income group, -$3.7k for the middle-income group, and -$7.3k for the high-income group, all with credible intervals excluding zero. This highlights the varying degrees of financial hardship across income groups, with all experiencing substantial income losses during the COVID period. The DiD estimates for percentage changes further confirm these findings, with the high-income group facing the greatest loss at -12.2%, followed by the low-income group at -7.7%, and the middle-income group at -5.5%, all with credible intervals excluding zero. These results underscore the intensifying financial burden faced by households, particularly among higher-income groups.



The high-income group experienced the greatest loss at -12.2%, likely due to a larger share of income coming from volatile sources, such as bonuses or business profits, which were more heavily impacted during the pandemic. Additionally, high-income individuals were more likely to face job disruptions in sectors like finance and tech, where layoffs and salary cuts were more prominent. Conversely, Low-income households, although significantly affected by COVID-19, may have experienced a smaller percentage loss of -7.7% in income due to their greater reliance on government assistance programs, such as unemployment benefits and stimulus checks. These safety nets may have helped cushion the financial blow, reducing the overall percentage decline. Additionally, with many sectors in need of flexible labor, low-income individuals may have been able to take advantage of temporary or remote job opportunities that helped mitigate income losses. Additionally, the flexibility of starting small businesses, particularly those related to COVID-19 (e.g., delivery services, personal protective equipment), may have provided alternative income streams during periods of economic uncertainty. The middle-income group experienced the lowest percentage change of -5.5%, which may reflect a mix of both stable and

vulnerable employment sectors. While some middle-class individuals were able to maintain their jobs in essential sectors, others faced income reductions or job disruptions, especially in industries like retail, hospitality, and manufacturing. This group likely saw moderate losses compared to the high and low-income groups due to a combination of exposure to both economic stability and volatility during the pandemic.

Discussion

The Common Shocks assumption is met in our context because COVID-19 represents a global, external phenomenon that affected all groups uniformly, regardless of treatment (job disruption) status. This ensures that external factors influencing household income, such as economic policies and public health measures, impacted treatment and control groups similarly. While we acknowledge the Parallel Trends assumption, which posits that treatment and control groups would have followed similar trajectories in the absence of treatment, our Bayesian hierarchical approach does not rely on this assumption as strictly as the counterfactual framework. Instead, our method provides greater flexibility in capturing heterogeneous effects and addressing potential deviations from parallel trends.

The use of hierarchical Bayesian models to capture heterogeneous effects in difference-in-differences and triple differences without relying on traditional three-way interaction terms is a novel approach. This method not only addresses issues like overfitting in small sample sizes but also avoids the complexity of large interaction models that are often computationally intensive. Our approach allows for subgroup-specific effects and better flexibility in modeling, which is important for real-world data like the FFCWS dataset.

Reference:

Card, D., & Krueger, A. B. (1994). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania*. American Economic Review, 84(4), 772-793.

Cintron, E. G., He, J., & Zha, Z. (2022). *Heterogeneous effects of economic shocks over time*. Journal of Economic Analysis, 56(2), 123-145.

Callaway, B., & Sant'Anna, P. H. (2021). *Difference-in-differences with multiple time periods*. Journal of Econometrics, 225(1), 23-39.

Meer, J., & West, J. (2024). *Economic shocks and policy responses: Evidence from multiple time periods*. Journal of Economic Policy, 48(3), 271-296.

de Chaisemartin, C., & d'Haultfoeuille, X. (2020). *Two-way fixed effects estimators with heterogeneous treatment effects*. Journal of Econometrics, 217(2), 372-384.

Baker, M., Fernandez, J., & Lee, T. (2022). *Staggered difference-in-differences: A critical review of recent developments*. Econometric Reviews, 41(4), 505-531.

Gruber, J. (1994). *The incidence of mandated maternity benefits*. American Economic Review, 84(3), 622-641.

Muralidharan, K., & Prakash, N. (2017). *Influencing voter behavior with targeted political messages: Evidence from a field experiment*. American Economic Review, 107(10), 2740-2785.

Griffin, J. (2021). *Simulating the effects of policy shocks: New tools for rigorous DiD analysis*. Policy Analysis Journal, 35(2), 98-115.

Normington, R. (2019). *Bayesian hierarchical models in difference-in-differences studies: A practical guide*. Journal of Applied Economics, 43(5), 1673-1690.

Gelman, A. (2021). *Bayesian approaches to causal inference: A primer*. Annual Review of Political Science, 24(1), 389-404.

Normington, R. (2021). *Advancements in Bayesian modeling for policy analysis*. Economics and Statistics, 35(2), 217-237.