

Looking into the Consumption Black Box: Evidence from Scanner Data

Krzysztof Pytka Ⓛ Daniel Runge
University of Mannheim

SED Annual Meeting, Copenhagen, 6/27/25



Generative prompt: Create a painting entitled
"Looking into the Consumption Black Box"
painted by Sandro Botticelli

Disclaimer

The researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project

“There is no accounting for taste”

Big Picture: Short Story of Non-Durable Consumption in Macroeconomics

Motto of the Project ☺

“There is **no** accounting for taste”

◊ Neoclassical model with representative agent

(Ramsey, EJ 1928; Cass, REStud 1965; Koopmans, 1965; Brock&Mirman, JET 1972; Kydland&Prescott, Ecta 1982)

- Early macroeconomic models considered a **single representative agent**.
- Consumption was aggregated into a **single product** for tractability.
- This approach ignored individual differences in consumption patterns.

◊ Models of Inequalities w/Heterogeneous Households

(Aiyagari, QJE 1994, JPE 1995; Huggett, JEDC 1993)

- Aimed to capture state-based consumption disparities.
- Generate **differences in quantities** consumed across households.
- Still focused on a single type of product.

◊ Models of Heterogeneous Goods (+Heterogeneity in Choices)

(Handbury, Ecta 2021; Michelacci, Paciello, and Pozzi, REStud 2022; Faber&Fally, REStud 2022; Neiman&Vavra, AEJ:Macro 2023; Mongey&Waugh, JPE R&R 2025)

- Recognize that households may consume different varieties of products due to:
either (i) latent heterogeneity or (ii) (intratemporal) non-homothetic preferences.
- **Differences in preferences** lead to diverse consumption baskets.
- Aim to provide more realistic representation of actual consumption behavior.

Motivation and Why is It Important?

Research Questions ②

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?  *Not discussed today*

Motivation and Why is It Important?

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ⚡ *Not discussed today*
3. What are implications for structural models of the answers to Questions 1 and 2?

Motivation and Why is It Important?

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ❌ *Not discussed today*
3. What are implications for structural models of the answers to Questions 1 and 2?
 - ◊ While there is a rich empirical literature documenting inequality in consumption expenditures, **very few studies systematically analyze the composition of consumption bundles across different groups of consumers**

Motivation and Why is It Important?

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ❌ *Not discussed today*
3. What are implications for structural models of the answers to Questions 1 and 2?
 - ◊ While there is a rich empirical literature documenting inequality in consumption expenditures, **very few studies systematically analyze the composition of consumption bundles across different groups of consumers**
 - ◊ This gap is particularly important given the direct implications for:
 - Price discrimination strategies by retailers (Mussa & Rosen, JET 1978; Maskin & Riley, RAND 1984)
 - Magnitude of consumption externalities (Pytka, REStud R&R 2024)
 - Redistribution through commodity taxation (Saez, JPubE 2003; Atkinson & Stiglitz, JPubE 1976)

Motivation and Why is It Important?

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ❌ *Not discussed today*
3. What are implications for structural models of the answers to Questions 1 and 2?
 - ◊ While there is a rich empirical literature documenting inequality in consumption expenditures, **very few studies systematically analyze the composition of consumption bundles across different groups of consumers**
 - ◊ This gap is particularly important given the direct implications for:
 - Price discrimination strategies by retailers (Mussa & Rosen, JET 1978; Maskin & Riley, RAND 1984)
 - Magnitude of consumption externalities (Pytko, REStud R&R 2024)
 - Redistribution through commodity taxation (Saez, JPubE 2003; Atkinson & Stiglitz, JPubE 1976)

👎 Key Challenges

1. *How to measure group differences in high-dimensional choices?*
2. *Potentially severe small-sample bias (yet to be discussed).*



◊ Heterogeneous Preferences

Neiman & Vavra (AEJ:Macro, 2023), Handbury (Ecta, 2021),
Michelacci, Paciello, & Pozzi (REStud, 2022), Faber & Fally (REStud, 2022),
Becker (JMP, 2024), Mongey & Waugh (JPE R&R, 2025)

◊ Polarization in Economics

Bertrand & Kamenica (AEJ: Applied, 2023), Alesina, Tabellini, & Trebbi (BP:EA, 2017),
Desmet & Wacziarg (EJ, 2021), Boar & Giannone (WP, 2023)

◊ Big Data with Small Sample Bias

Gentzkow, Shapiro, & Taddy (Ecta, 2019), Armenter & Koren (AER, 2014),
Conlon, Mortimer, & Sarkis (WP, 2023)

◊ Methodological Debates – “*Stirring up a hornet’s nest*”

- *Confronting Mainstream:* Armenter & Koren (AER, 2014), Menzio (JPE R&R, 2025),
Borovičková & Shimer (QJE R&R, 2024)
- *Defending Mainstream:* Becker (JPE, 1962)



⚖️ Indistinguishable Consumption Choices

- Consumption choices show **very limited polarization** between rich and poor households
- On average, observing how one dollar is spent provides only modest ability to identify whether the consumer is rich or poor
- This suggests that non-durable consumption polarization is **substantially lower** than commonly perceived, though not absent entirely

⌚ Unstable Individual Consumption Choices ⓘ Not discussed today

- Individual consumption choices are very unstable. On average, only 40% of products purchased in one year are bought again in the next year

_MOUSE_parsimonious Model of Consumption

- Perfect substitutes + random sampling
- Surprisingly good empirical fit, but fundamentally different policy implications than recently popular models with heterogeneous preferences
- **Identification problem:** if two very different models fit equally well, **how do we know which mechanisms drive consumption patterns?**

- ◊ **Source:** Kilts-NielsenIQ Consumer Panel
- ◊ **Sample Coverage:** 40,000–60,000 households per year (2004–2016) with detailed socio-economic characteristics
- ◊ **Product Universe:** 1.5 million unique products ▪ groceries ▪ drug products ▪ small appliances ▪ electronics
- ◊ **Data Quality:** Scanner-based collection ensures low measurement error
 - ▷ Data collected via handheld scanner or mobile app by scanning barcodes
- ◊ **Representativeness:** Projection weights ensure US economy representativeness

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ⓘ Not discussed today
3. What are implications for structural models of the answers to Questions 1 and 2?

Challenge 1

How to measure group differences in **high-dimensional** choices?

- ◊ Measuring polarization is challenging due to high-dimensional data and diverse product categories.
 - ▶ Some products are favored by specific income groups
 - ▶ Other products have universal appeal

Challenge 1

How to measure group differences in **high-dimensional** choices?

- ◊ Measuring polarization is challenging due to high-dimensional data and diverse product categories.
 - ▶ Some products are favored by specific income groups
 - ▶ Other products have universal appeal

💡 **Our Approach:** Gentzkow, Shapiro, & Taddy (Ecta, 2019) ▶ Methodology Details

Polarization in Consumption

Challenge 1

How to measure group differences in **high-dimensional** choices?

- ◊ Measuring polarization is challenging due to high-dimensional data and diverse product categories.
 - ▶ Some products are favored by specific income groups
 - ▶ Other products have universal appeal

💡 **Our Approach:** Gentzkow, Shapiro, & Taddy (Ecta, 2019) ▶ Methodology Details

⚙️ Methodology:

Consumption-Based Polarization

1. Randomly select \$1 spent in Nielsen universe
2. Observe product allocation of this \$1
3. Predict household group membership from products
4. **Polarization** = expected predictive accuracy

Polarization in Consumption

Challenge 1

How to measure group differences in **high-dimensional** choices?

- ◆ Measuring polarization is challenging due to high-dimensional data and diverse product categories.
 - ▶ Some products are favored by specific income groups
 - ▶ Other products have universal appeal

💡 Our Approach: Gentzkow, Shapiro, & Taddy (Ecta, 2019)

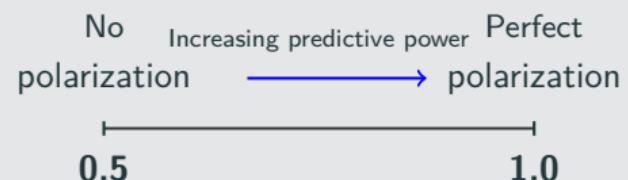
► Methodology Details

Methodology:

Consumption-Based Polarization

1. Randomly select \$1 spent in Nielsen universe
 2. Observe product allocation of this \$1
 3. Predict household group membership from products
 4. **Polarization** = expected predictive accuracy

Interpretation



Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. **50,000** households \Rightarrow **Sparse Data Problem**

Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. 50,000 households \Rightarrow Sparse Data Problem

- ◊ **Data Sparsity:** $\approx 30\%$ of consumption bought only once
- ◊ **Extreme Case:** Single purchases \Rightarrow 100% predictive power
- ◊ **Key Question:** Are rare purchases systematic or random?

 View Table

Challenge 2: Potentially Severe Small-Sample Bias

800,000 products vs. 50,000 households \Rightarrow Sparse Data Problem

- ◊ **Data Sparsity:** $\approx 30\%$ of consumption bought only once
- ◊ **Extreme Case:** Single purchases \Rightarrow 100% predictive power
- ◊ **Key Question:** Are rare purchases systematic or random?

 View Table

Solution Strategy

- ◊ **Cannot use raw data** \Rightarrow infrequent transactions cause **spurious polarization**
- ◊ Need to **filter systematic differences** and **reduce noise**
- ◊ **Solution:** Estimate theoretical distributions, then study polarization

Consumption Polarization Methodology

- ◊ **Approach:** Following Gentzkow, Shapiro, & Taddy (Ecta, 2019), we study polarization of estimated data generating processes rather than raw data to overcome small sample bias
 1. Estimate consumption choice model allowing for income group differences
 2. Compare predicted choice distributions between high and low income groups

Consumption Polarization Methodology

- ◊ **Approach:** Following Gentzkow, Shapiro, & Taddy (Ecta, 2019), we study polarization of estimated data generating processes rather than raw data to overcome small sample bias
 1. Estimate consumption choice model allowing for income group differences
 2. Compare predicted choice distributions between high and low income groups
- ◊ **Model:** Household i 's consumption vector $\mathbf{c}_{i,t}$ in year t :

$$\mathbf{c}_{i,t} \sim \text{Multinomial}(m_{i,t}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{i,t}))$$

where $m_{i,t}$ = total expenditure, $P(i) \in \{H, L\}$ = income group, $\mathbf{x}_{i,t}$ = household characteristics

Consumption Polarization Methodology

- ◊ **Approach:** Following Gentzkow, Shapiro, & Taddy (Ecta, 2019), we study polarization of estimated data generating processes rather than raw data to overcome small sample bias
 1. Estimate consumption choice model allowing for income group differences
 2. Compare predicted choice distributions between high and low income groups
- ◊ **Model:** Household i 's consumption vector $\mathbf{c}_{i,t}$ in year t :

$$\mathbf{c}_{i,t} \sim \text{Multinomial}(m_{i,t}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{i,t}))$$

where $m_{i,t}$ = total expenditure, $P(i) \in \{H, L\}$ = income group, $\mathbf{x}_{i,t}$ = household characteristics

- ◊ **Choice Probabilities:**

$$q_{jt}^{P(i)}(\mathbf{x}_{i,t}) = \frac{e^{u_{i,j,t}}}{\sum_{k \in J} e^{u_{i,k,t}}}, \quad u_{i,j,t} = \alpha_{j,t} + \mathbf{x}'_{i,t} \boldsymbol{\gamma}_{j,t} + \varphi_{j,t} \mathbb{1}_{i \in H_t}.$$

Consumption Polarization Methodology

- ◊ **Approach:** Following Gentzkow, Shapiro, & Taddy (Ecta, 2019), we study polarization of estimated data generating processes rather than raw data to overcome small sample bias
 1. Estimate consumption choice model allowing for income group differences
 2. Compare predicted choice distributions between high and low income groups
- ◊ **Model:** Household i 's consumption vector $\mathbf{c}_{i,t}$ in year t :

$$\mathbf{c}_{i,t} \sim \text{Multinomial}(m_{i,t}, \mathbf{q}_t^{P(i)}(\mathbf{x}_{i,t}))$$

where $m_{i,t}$ = total expenditure, $P(i) \in \{H, L\}$ = income group, $\mathbf{x}_{i,t}$ = household characteristics

- ◊ **Choice Probabilities:**

$$q_{jt}^{P(i)}(\mathbf{x}_{i,t}) = \frac{e^{u_{i,j,t}}}{\sum_{k \in J} e^{u_{i,k,t}}}, \quad u_{i,j,t} = \alpha_{j,t} + \mathbf{x}'_{i,t} \boldsymbol{\gamma}_{j,t} + \varphi_{j,t} \mathbb{1}_{i \in H_t}.$$

- ◊ **LASSO Regularization addresses sparsity:**
 - Parameter $\varphi_{j,t}$ penalized to **extract only systematic between-group differences**
 - **Noise-driven differences are shrunk to zero**
 - **Constraint:** $\sum_{j,t} |\varphi_{j,t}| \leq \Lambda$, where hyperparameter Λ is determined using AIC
- ◊ **Computational Solution:** Distributed Poisson approximation (Taddy, AoAS 2013) for 800,000+ product categories

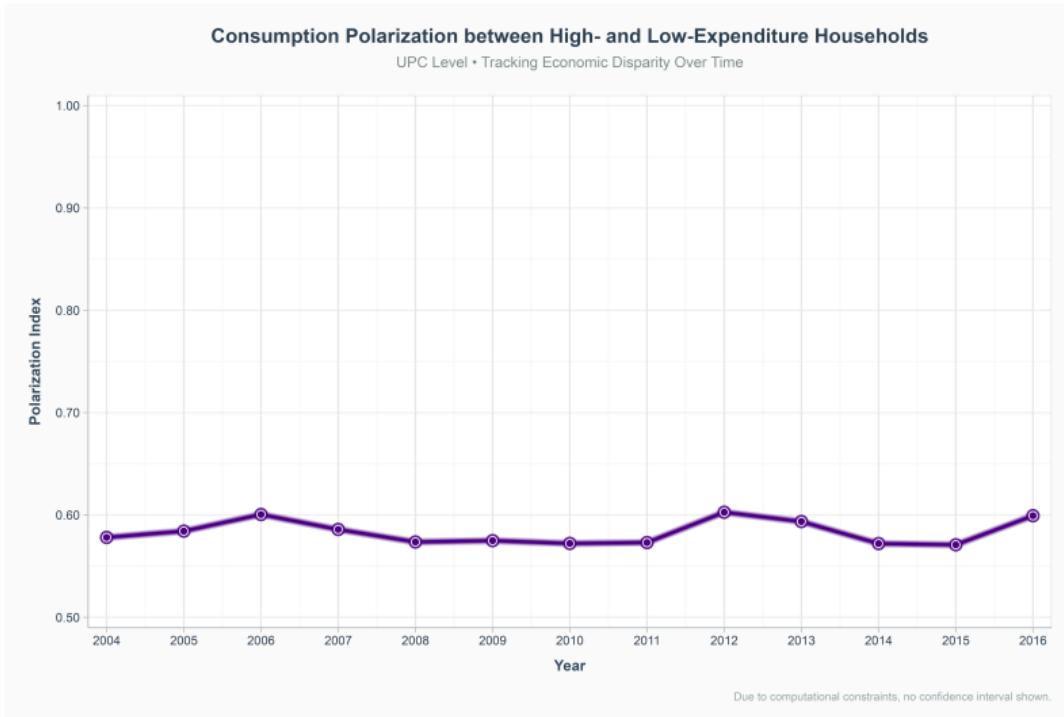
Consumption Baskets of the Rich and Poor: Not So Different After All

Fact (Consumption Polarization)

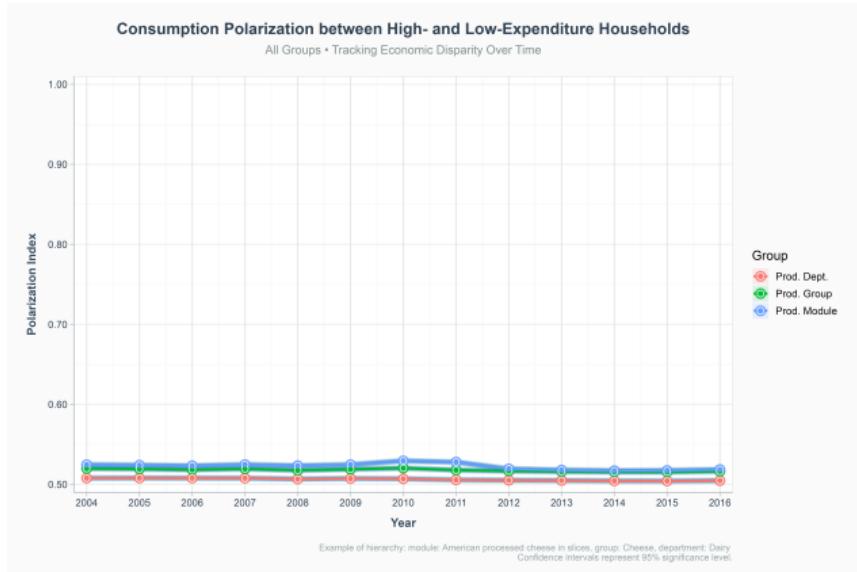
The average consumption polarization $\bar{\pi}$ is low and close to its lower bound of 50%. Specifically, if we randomly draw \$1 spent by a high- or low-income household in the NielsenIQ universe and observe how it was allocated, the probability of correctly classifying the spender is:

- ◊ 58.3% when the **barcode** of the purchased product is observed,
- ◊ 52.2% when the **product module** is observed,
- ◊ 51.7% when the **product group** is observed,
- ◊ 50.6% when the **product department** is observed.

Consumption Polarization: Predictive Power of UPC Codes



Consumption Polarization: Predictive Power of Broader Categories



- ◊ With our methodology, broader categories always result in **lower** polarization values
- ◊ UPC level provides the **upper bound** for consumption polarization

► Where are Confidence Intervals?

Comparing Predictability-Based Measures

- ◊ Bertrand & Kamenica (AEJ: Applied, 2023) find **high predictive power of consumption behavior** for income classification ($\approx 90\%$)
- ◊ Yet our results are **much more modest** ($< 60\%$, where 50% is the lower bound)

Comparing Predictability-Based Measures

- ◊ Bertrand & Kamenica (AEJ: Applied, 2023) find **high predictive power of consumption behavior** for income classification ($\approx 90\%$)
- ◊ Yet our results are **much more modest** ($< 60\%$, where 50% is the lower bound)

 **How come?**

Comparing Predictability-Based Measures

- ◊ Bertrand & Kamenica (AEJ: Applied, 2023) find **high predictive power of consumption behavior** for income classification ($\approx 90\%$)
- ◊ Yet our results are **much more modest** ($< 60\%$, where 50% is the lower bound)

 **How come?**

Cultural Distance

Bertrand & Kamenica (AEJ:Applied, 2023)

- ▷ **Focus:** Existence of distinguishing behaviors
- ▷ **Method:** Maximum classification accuracy across all possible product combinations

Consumption Polarization (Our Approach)

- ▷ **Focus:** Magnitude of distinguishing behaviors in overall spending
- ▷ **Method:** Expected information value of the “representative” dollar spent

Key Distinction

Bertrand & Kamenica: Identifies **whether** household groups have distinct consumption behaviors

Our approach: Quantifies **how much** goods that distinguish households contribute to overall grocery spending

🍴 Illustrative Example: Dinner at Beppa Fioraia

Consumption Breakdown

Item	Me	Pavel
Bistecca Fiorentina	30€	30€
Chianti Classico	10€	10€
Tiramisu	7€	7€
Spinacci	3€	—
French Fries	—	3€
Total	50€	50€

🍴 Illustrative Example: Dinner at Beppa Fioraia

Consumption Breakdown

Item	Me	Pavel
Bistecca Fiorentina	30€	30€
Chianti Classico	10€	10€
Tiramisu	7€	7€
Spinacci	3€	—
French Fries	—	3€
Total	50€	50€

Methodological Results

Bertrand & Kamenica (AEJ:Applied, 2023)

- Side dishes allow **100%** classification accuracy
- Consequently: Cultural Distance = **Maximum**

Our Consumption Polarization:

$$\text{Polarization} = \frac{2 \times 3 \times 100\% + 2 \times 47 \times 50\%}{100}$$
$$= 53\% \text{ (Close to lower bound, 50%)}$$

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time? ⓘ Not discussed today
3. What are implications for structural models of the answers to Questions 1 and 2?

▶ Skip to Shopping Spree

Persistence in Basket Composition

Defining the Persistence Measure

 **Core Concept:** We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

Defining the Persistence Measure

💡 **Core Concept:** We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

$$O_i^E = \frac{\sum_{j \in (\mathcal{U}_{i,t-1} \cap \mathcal{U}_{i,t})} e_{i,t}(j)}{\sum_{j \in \mathcal{U}_{i,t}} e_{i,t}(j)}$$

$\mathcal{U}_{i,t-1}$ = products consumed by household i in year $t - 1$

$\mathcal{U}_{i,t}$ = products consumed by household i in year t

$e_{i,t}(j)$ = expenditures on product j by household i in year t

Defining the Persistence Measure

💡 **Core Concept:** We measure persistence in household consumption by calculating the share of current-year expenditures spent on products already purchased in the previous year

$$O_i^E = \frac{\sum_{j \in (\mathcal{U}_{i,t-1} \cap \mathcal{U}_{i,t})} e_{i,t}(j)}{\sum_{j \in \mathcal{U}_{i,t}} e_{i,t}(j)}$$

$\mathcal{U}_{i,t-1}$ = products consumed by household i in year $t - 1$

$\mathcal{U}_{i,t}$ = products consumed by household i in year t

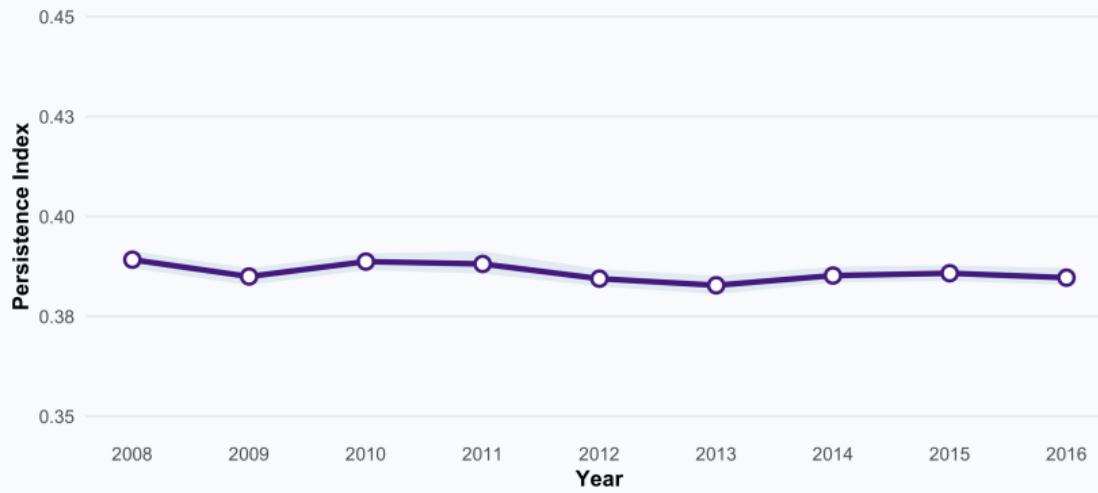
$e_{i,t}(j)$ = expenditures on product j by household i in year t

Robustness checks:

- ➡ Restrict analysis to products available in both years to control for market entry/exit effects
- ⌚ Use higher-order lags
- 🏷️ Use different product definitions: • UPC • modules • modules×brands

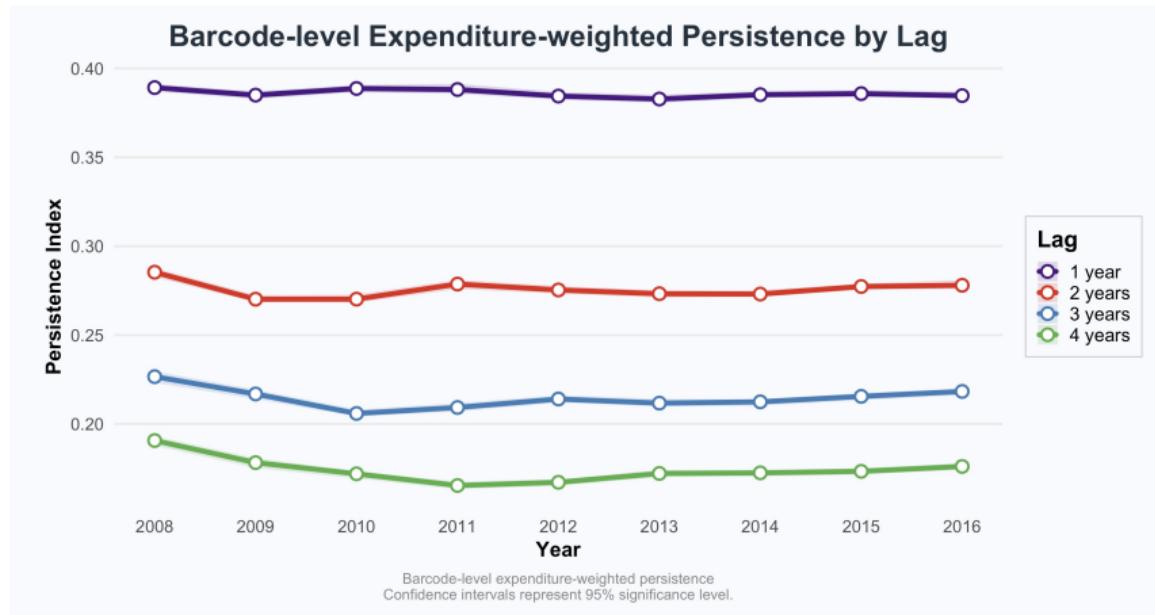
Individual Consumption Baskets: Constantly in Flux

Barcode-level Expenditure-weighted Persistence Over Time



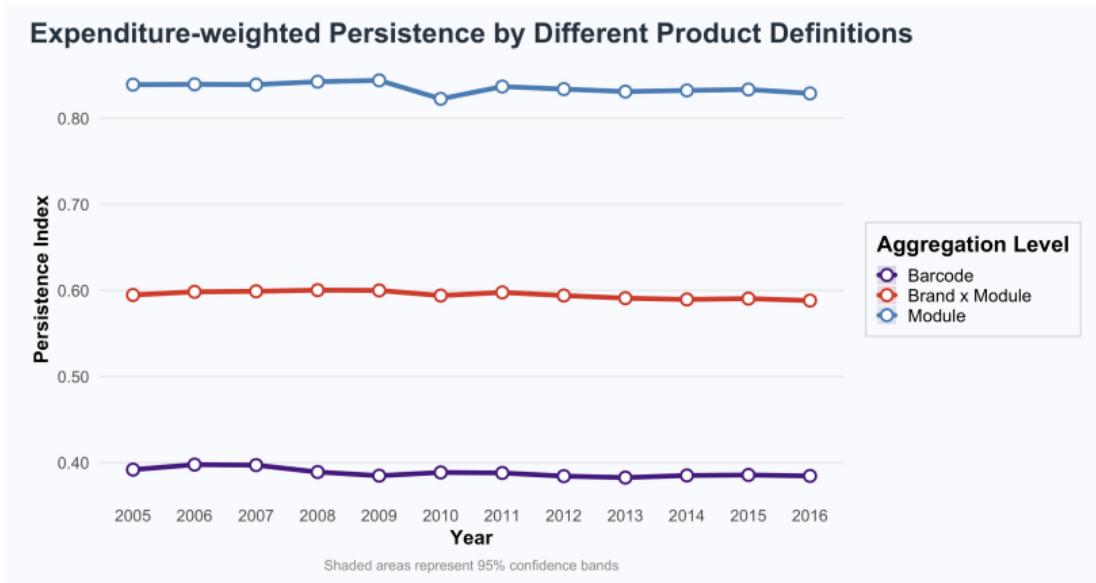
Key Insight: Household consumption patterns exhibit significant year-to-year variation, highlighting the dynamic nature of consumer choice behavior

Persistence Decreases with Longer Time Horizons



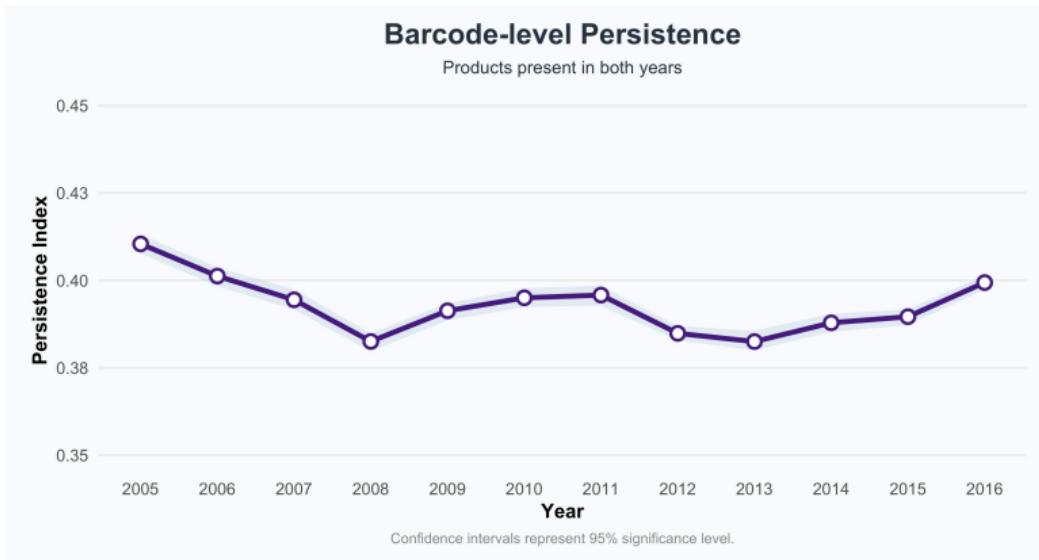
Pattern: Consumption persistence drops from 40% (1-year lag) to below 20% (4-year lag)

Product Definition Matters: Modules vs. Specific Products



Key Finding: Persistence jumps dramatically to over 80% when using product modules instead of specific UPC codes, revealing that households have stable *category* preferences but flexible *brand* choices

Robustness Check: Product Exit Does Not Drive Results



✓ **Robustness Check:** Restricting analysis to products present in both years yields similar persistence patterns, confirming that our baseline findings are not driven by product entry or exit dynamics

What This Tells Us About Consumer Behavior

Stable Category Preferences:

- ◊ Households know what general products they want (e.g., "American processed cheese in slices")
- ◊ These broader consumption patterns persist over time

Flexible Brand/Product Choices:

- ◊ Limited attachment to specific brands or UPC codes
- ◊ Willingness to substitute within product categories

 **Implication:** Consumers exhibit *categorical loyalty* rather than *brand loyalty*, maintaining consistent consumption needs while remaining flexible on specific product choices

Research Questions ⓘ

1. Do different groups of consumers buy different products?
2. How persistent are the individual choices over time?
3. What are implications for structural models of the answers to Questions 1 and 2?

Shopping Spree of Beckerian Consumers

Alternative View: Consumption as Random Sampling

- ◊ **Recent models:** Differences in consumption baskets result from **idiosyncratic substitution patterns** and heterogeneous preferences

Alternative View: Consumption as Random Sampling

- ◊ **Recent models:** Differences in consumption baskets result from **idiosyncratic substitution patterns** and heterogeneous preferences
- ◊ **Our extremely different thought experiment:** What if basket differences simply reflect **random sampling** because all products are (price-adjusted) perfect substitutes?
Data-driven version of Becker's (JPE, 1962) impulsive consumers
or "balls and bins" (Armenter & Koren, AER 2014) applied to consumption choices

Alternative View: Consumption as Random Sampling

- ◊ **Recent models:** Differences in consumption baskets result from **idiosyncratic substitution patterns** and heterogeneous preferences
- ◊ **Our extremely different thought experiment:** What if basket differences simply reflect **random sampling** because all products are (price-adjusted) perfect substitutes?
Data-driven version of Becker's (JPE, 1962) impulsive consumers
or "balls and bins" (Armenter & Koren, AER 2014) applied to consumption choices

Extreme example:



Shopping Spree Model

- ◊ Each household i has its own budget $m_{i,t}$ and nothing can be saved for the future.
- ◊ No information about products: to buy a product, they need to find it.
- ◊ There is no order in preferences; households randomly draw products during each purchase.
- ◊ Goal of each household: to spend as much as possible.

¹This reflects our empirical finding of no consumption polarization between rich and poor households.

² \mathbf{e}_j is a vector with 1 at the j -th position and 0 elsewhere.

Shopping Spree Model

- ◊ Each household i has its own budget $m_{i,t}$ and nothing can be saved for the future.
- ◊ No information about products: to buy a product, they need to find it.
- ◊ There is no order in preferences; households randomly draw products during each purchase.
- ◊ Goal of each household: to spend as much as possible.

Decision Problem of the Consumer

0. The consumer's consumption $\mathbf{c}_{i,t}$ is a zero vector at the beginning of the period.
1. The consumer randomly draws product j from the set of all products. The probability of drawing product j is product-specific and common across all consumers.¹ The number of purchased products, n_j , is drawn from a product-specific Poisson distribution.
2. Update $\mathbf{c}_{i,t} \leftarrow \mathbf{c}_{i,t} + \mathbf{e}_j n_j$.²
3. If the budget constraint for the new bundle is not violated ($\mathbf{p}'_{i,t} \mathbf{c}_{i,t} < m_{i,t}$), go to **Step 1**. Otherwise, **Stop**.

¹This reflects our empirical finding of no consumption polarization between rich and poor households.

² \mathbf{e}_j is a vector with 1 at the j -th position and 0 elsewhere.

Model Validation: Spending Concentration Plot

- ◊ Diagnostic introduced by Neiman & Vavra (AEJ:Macro, 2023):

Households concentrate spending on top-ranked products, but concentration varies with consumption diversity

- ◊ Plot structure:

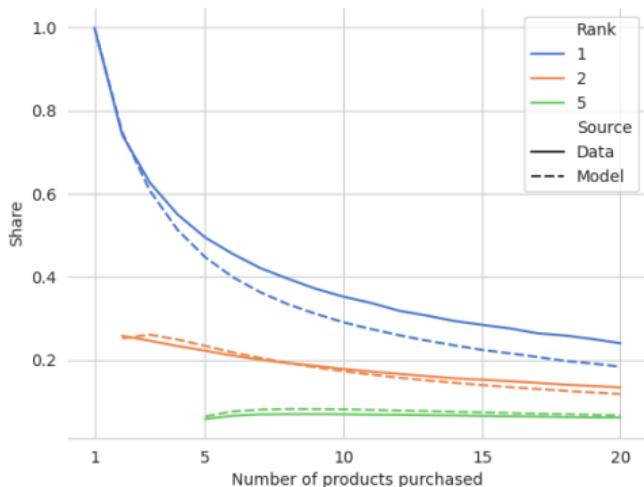
- ↳ X-axis: Total products consumed ($|\Omega_i|$)
 - ↳ Y-axis: Average spending share on ranked products
 - ↳ Lines: Different ranks (1st, 2nd, 5th)

- ◊ Validation approach:

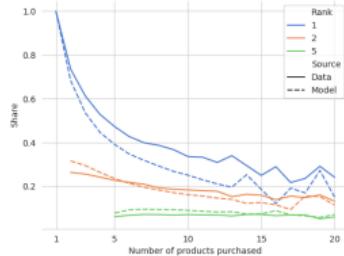
- **Solid** = data patterns
 - **Dashed** = model predictions

Model Validation: Spending Concentration Plot

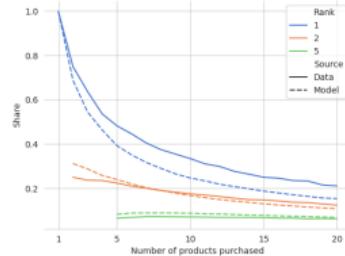
- ◊ Diagnostic introduced by Neiman & Vavra (AEJ:Macro, 2023):
Households concentrate spending on top-ranked products, but concentration varies with consumption diversity
- ◊ Plot structure:
 - ↳ X-axis: Total products consumed ($|\Omega_i|$)
 - ↳ Y-axis: Average spending share on ranked products
 - ↳ Lines: Different ranks (1st, 2nd, 5th)
- ◊ Validation approach:
 - Solid = data patterns
 - Dashed = model predictions



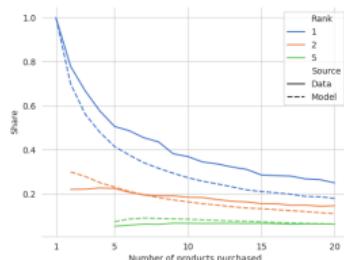
Spending Concentration within Different Product Categories



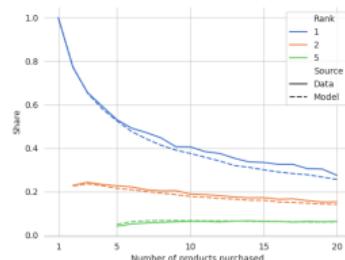
(a) Seafood Canned



(b) Cereal



(c) Yogurt



(d) Pet Food

☺ Ok... so what?

- ◊ **Observational equivalence:** Empirical patterns typically attributed to heterogeneous preferences can be generated by very different models.
- ◊ **Other examples of “stirring up a hornet’s nest”:** “Balls and bins” vs. gravity models (Armenter & Koren, AER 2014); price search vs. Dixit-Stiglitz (Menzio, JPE R&R 2025); irrational vs. rational consumers (Becker, JPE 1962); assortative wage matching vs. AKM (Borovičková & Shimer, QJE R&R 2024)
- ◊ **Our model:** Differences in consumption baskets result solely from different sampling histories— no heterogeneity in preferences, just random draws. An extremely different theoretical foundation.

☺ Ok... so what?

- ◊ **Observational equivalence:** Empirical patterns typically attributed to heterogeneous preferences can be generated by very different models.
- ◊ **Other examples of “stirring up a hornet’s nest”:** “Balls and bins” vs. gravity models (Armenter & Koren, AER 2014); price search vs. Dixit-Stiglitz (Menzio, JPE R&R 2025); irrational vs. rational consumers (Becker, JPE 1962); assortative wage matching vs. AKM (Borovičková & Shimer, QJE R&R 2024)
- ◊ **Our model:** Differences in consumption baskets result solely from different sampling histories— no heterogeneity in preferences, just random draws. An extremely different theoretical foundation.
- ◊ **Cautionary tale:** If two extremely different models generate similar empirical patterns, perhaps more investigation is warranted before attributing consumption differences to heterogeneous preferences
- ◊ **Policy implications diverge:** Increasing product variety is welfare decreasing in our setup vs. welfare increasing under heterogeneous preferences (Neiman & Vavra, AEJ:Macro 2023)

Concluding Thoughts

1. Consumption Similarity Across Income Groups

- Rich and poor households make surprisingly similar consumption choices
- Product purchase data alone cannot distinguish household income groups

👉 Challenge: Intratemporal non-homothetic preference models

2. Consumption Instability

- Individual choices are highly unstable over time
- Repeat purchase probability: only 40% year-over-year

👉 Challenge: Latent heterogeneous preference models

3. Beckerian Impulsive Consumers

- Randomness-driven model explains consumption basket differences
- (Surprisingly) strong empirical fit with scanner data

⚠ Key Takeaway

The paper strikes a cautious note on policy implications drawn from models with heterogeneous preferences.

Thank you for your attention!

Group Differences in High-dimensional Choices

- ◊ **Context:** Gentzkow, Shapiro, and Taddy (Ecta, 2019) identify bias in their study on the polarization of US politics using congressional speech data
- ◊ **Model:** They developed a formal model of the speech-generating process
- ◊ **Solution:** They analyzed the bias formally and introduced estimators to **overcome finite-sample bias**
- ◊ **Result:** Their methods help recover a more **accurate estimate of polarization**
- ◊ **Application:** Their methodology is well-suited for studying consumption polarization due to structural similarities between speech and scanner data

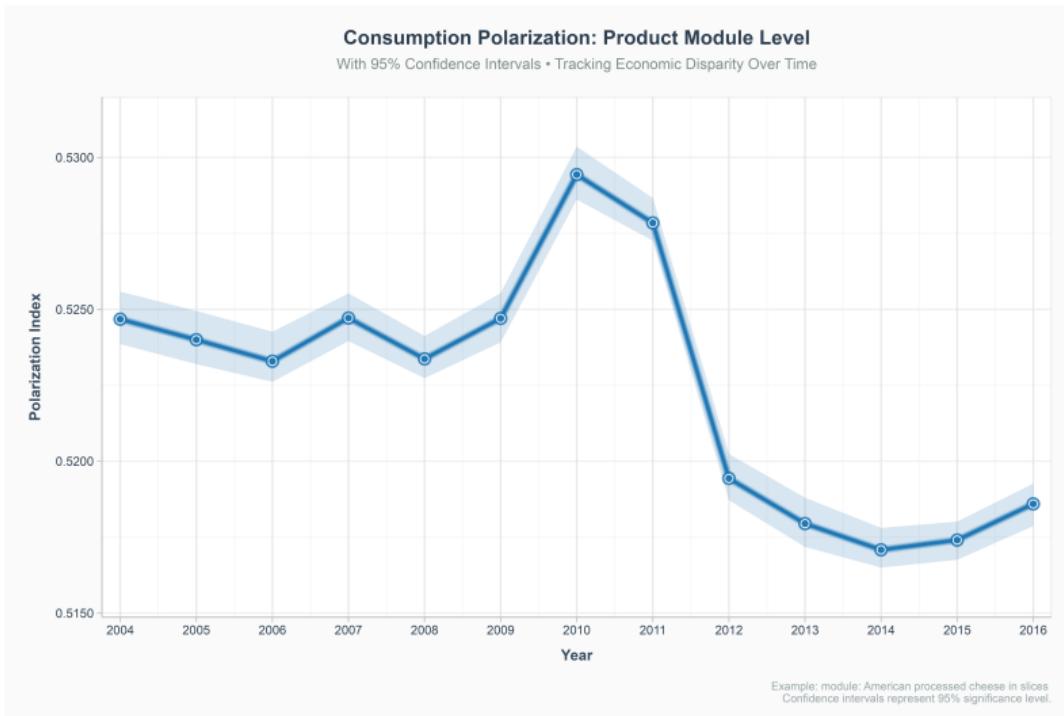
Gentzkow on Gentzkow et al. (Ecta, 2019)

Gentzkow on Gentzkow et al. (Ecta, 2019)



"Let's look at how different the choices of two different groups are in that high-dimensional space (...) It could be which UPCs did they buy in the supermarket."

Consumption Polarization: Predictive Power of Modules

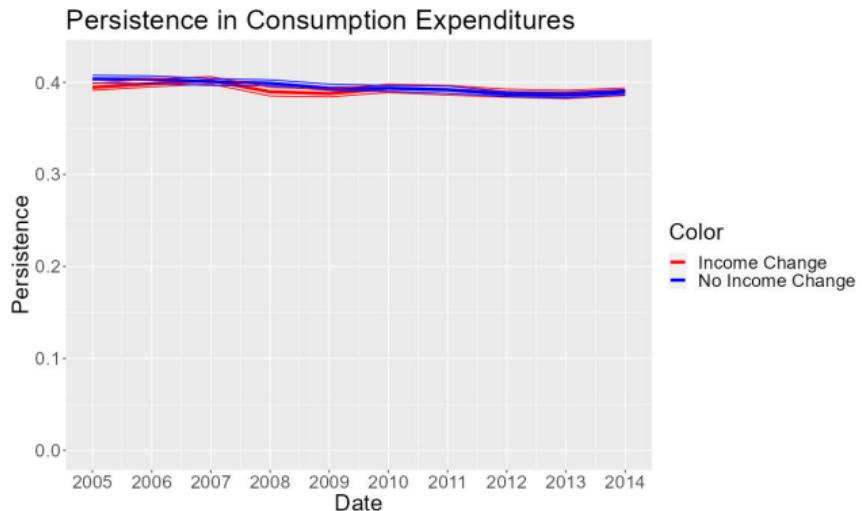


Group Averages

Quantile	Size	Male Age	Female Age	Race	Income
1st	2.16	4.38	6.24	1.38	20
2nd	2.26	5.22	6.54	1.29	20.2
3rd	2.32	5.65	6.63	1.25	20.1
4th	2.33	5.94	6.65	1.21	20
5th	2.19	6.16	6.37	1.17	19.3

▶ Back

Persistence and Income Change



▶ Back

Zooming into the Heterogeneity

Question

Can we attribute the heterogeneity to observable characteristics of the households?

Zooming into the Heterogeneity

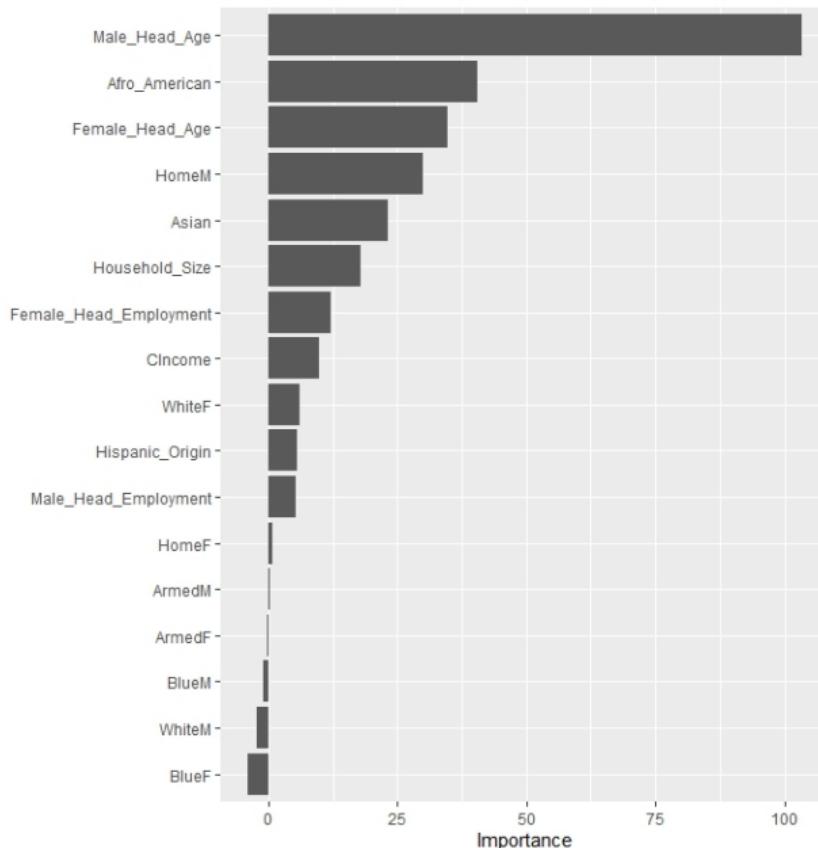
Question

Can we attribute the heterogeneity to observable characteristics of the households?

- ◊ We employ a random forest algorithm to find the household characteristics that are most important to explain differences within persistence
- ◊ We can then use accumulated local effects to see the size and direction of the influence of individual characteristics
 - Identifies the effect of singular variables while holding all else constant

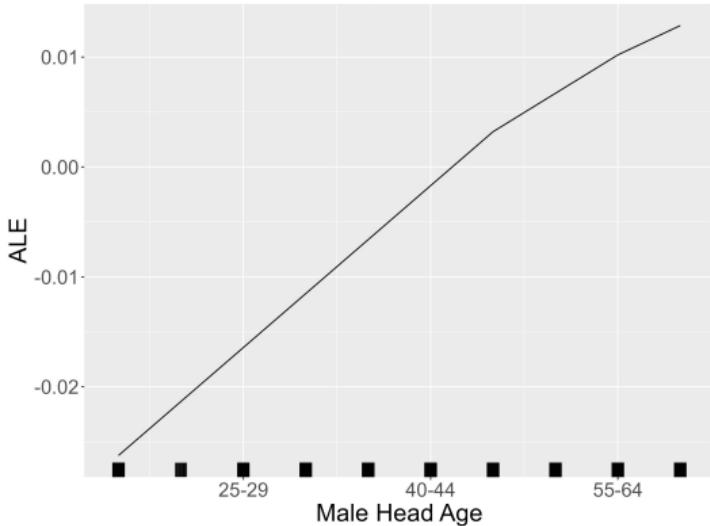
▶ Alternatively: Subgroup analysis

Variable Importance

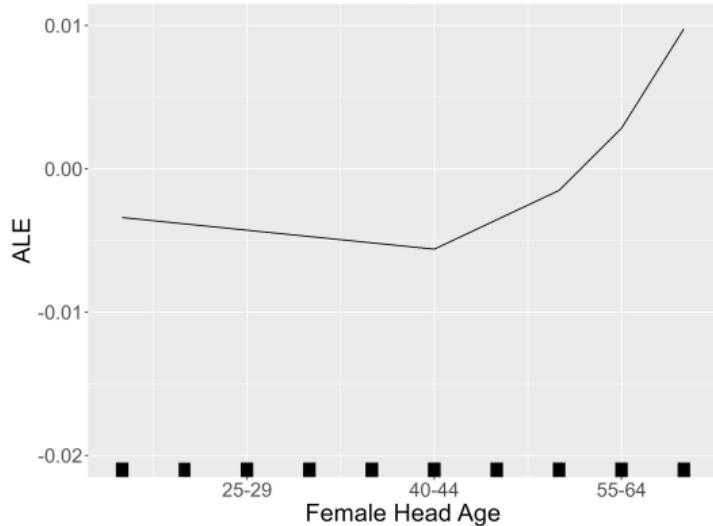


Effect Age

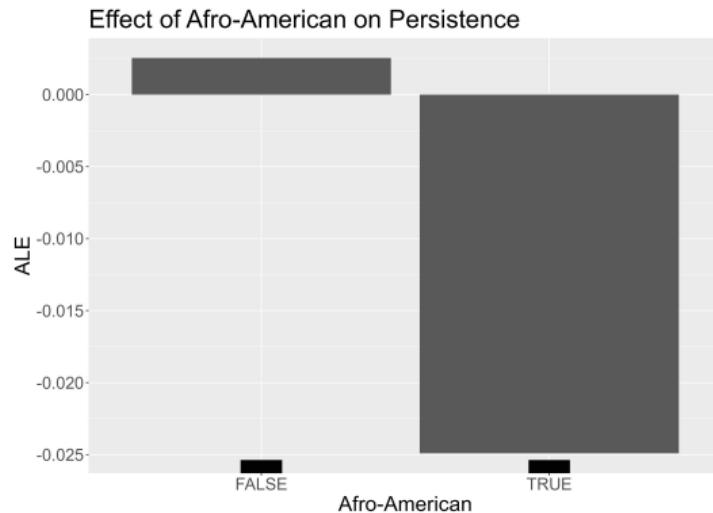
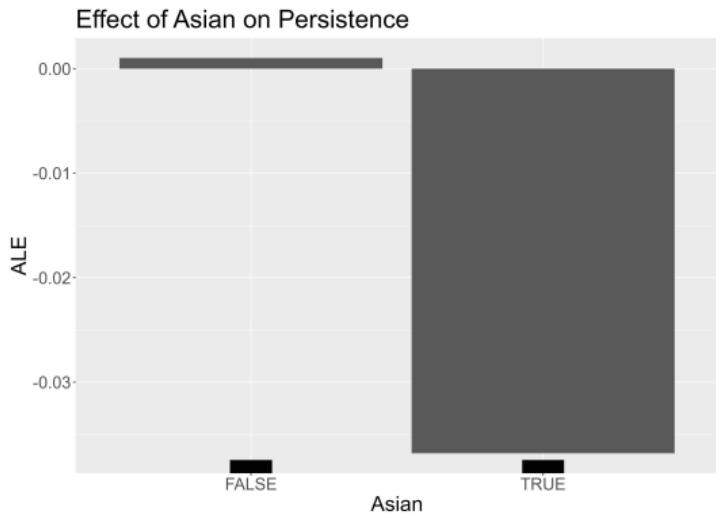
Effect of Male Head Age on Persistence



Effect of Female Head Age on Persistence



Effect Race



Results: Persistence Summary

- ◊ We have seen that households are **on average** highly impersistent with their consumption choices
- ◊ More than 50% of aggregate non-durable consumption is accounted for by households with a persistence of less than 50%
- ◊ **But** there is substantial heterogeneity within persistence
- ◊ Most of the heterogeneity is latent and only some can be explained by household characteristics.

Data Sparsity in Nielsen Dataset

Table 1: Fraction of Products by Minimum Transaction Threshold

Product Definition	Geographic Scope	Minimum No. of Transactions			
		≥ 1	≥ 2	≥ 10	≥ 20
UPC	Nationwide	1.000	0.963	0.814	0.723
UPC	Scantrack Market	1.000	0.713	0.287	0.209

- ◊ **Nationwide:** Even with national aggregation, 28% of products have fewer than 20 transactions
- ◊ **Local markets:** 79% of products have fewer than 20 transactions in local markets
- ◊ Sparsity increases dramatically when moving from national to local market analysis

Some Products Are More Popular in Certain Income Groups



Soft'n Gentle Tissue

Unscented, 2-Ply

Purchased **1.98× more frequently**
by **bottom expenditure quintile**
(vs. *top quintile, 2008*)



Tropicana Pure Premium

Orange Juice, 128 Oz

Purchased **7.56× more frequently**
by **top expenditure quintile**
(vs. *bottom quintile, 2008*)

[Go Back](#)

Universal Appeal Across Income Distribution



Kellogg's All-Bran

Original, 18.5 Oz

Purchased at **similar frequencies**
by both **high and low income**
consumers

(Less than one-percent difference between quintiles, 2008)

[Go Back](#)