States of Affordability: Methodology

Cost thresholds

Food

This analysis estimates food costs by household type using data from the United States Department of Agriculture (USDA) and Feeding America's Map the Meal Gap, an annual assessment of food insecurity for every U.S. county.

To create a baseline estimate of food costs by household type, we download and harmonize state-level cost data for USDA's Thrifty and Low-Cost food plans, available with disaggregation by age bracket. Because USDA provides separate cost thresholds for the Thrifty plan in Alaska and Hawaii but not separate thresholds for the Low-Cost plan in these states, we estimate distinct Low-Cost plan estimates for Alaska and Hawaii using the ratio between the Thrifty plan in these states compared to the rest of the continental U.S. These cost estimates are therefore defined as $L_{ak} = L_{us} \cdot \left(\frac{T_{ak}}{T_{us}}\right)$ and $L_{hi} = L_{us} \cdot \left(\frac{T_{hi}}{T_{us}}\right)$ for Alaska and Hawaii, respectively, where L represents the Low-Cost plan and T represents the Thrifty plan.

To convert these thresholds into family-specific cost estimates, we convert USDA's age brackets to their closest possible alignment with the age categories used throughout this analysis:

- 1 year and under ~ infants
- 2 to 5 years ~ preschoolers
- 6 to 11 years ~ school-aged children
- 13 to 18 years ~ teenagers
- 19-plus ~ adults

These age-adjusted cost thresholds are then aggregated into annual averages, crossed with our family archetype matrix, and adjusted based on USDA's recommended cost factors by family size to account for diminishing marginal costs. The total state-level cost threshold for each food plan and family archetype therefore becomes:

$$c_{u,y} = \lambda \cdot \Big(\big(A_{u,y} \cdot a\big) + \big(I_{u,y} \cdot i\big) + \big(P_{u,y} \cdot p\big) + \big(S_{u,y} \cdot s\big) + \big(T_{u,y} \cdot t\big) \Big),$$

where c represents total cost, u represents the USDA plan, y represents the period, and λ represents the diminishing marginal cost factor. The number of adults, infants, preschoolers, school-aged children, and teenagers are represented by a, i, p, s, and t, respectively, and their uppercase counterparts represent the cost threshold for that age group within a given period and USDA plan.

Finally, these cost thresholds are normalized into county-specific estimates using Feeding America data, where each county's cost-per-meal total is annualized, crossed with our family archetype matrix, and adjusted using the same USDA cost factors utilized above.

Housing

This analysis utilizes fair-market (40th percentile) rental costs provided on an annual basis by the U.S. Department of Housing and Urban Development (HUD). To assign housing units to each family in our family archetype matrix, we assume that each family requires a number of bedrooms equal to $bedrooms = \left\lceil \frac{adults}{2} \right\rceil + \left\lceil \frac{children}{2} \right\rceil$. This assumes a maximum occupancy of two family members per bedroom, and assumes that adults and children do not share bedrooms under any family archetype. Values are rounded up to the ceiling integer (e.g., a single-adult household with no children is still assumed to occupy a one-bedroom rental unit).

By default, HUD provides fair-market rental rates for units up to four bedrooms. For unusually large mixed-age households requiring larger units, we predict forward from smaller units, assuming that $\widehat{fmr} = exp\left(\hat{b}_0 + \hat{b}_1 \cdot bedrooms\right)$. Note that while this model is applied to ensure that all families in the archetype matrix are assigned a valid housing cost value, these larger families are uncommon in population-representative datasets such as the American Community Survey.

Child care

Data for child care costs used in this analysis are derived from the U.S. Department of Labor's (DOL) National Database of Childcare Prices (NDCP), which provides annual cost-of-care estimates for infants, preschoolers, and school-aged children in nearly every county in the United States. For each county and age group, we assume thresholds equal to the average cost of center-based and family-based care. Data for missing counties are imputed using chained random forests utilizing dozens of other predictors provided directly by DOL on county-level earnings data and other labor market outcomes.

After populating the dataset with all necessary imputations, we cross each cost threshold with our family archetype matrix and calculate total costs based on each family's number of infants, preschoolers, and school-aged children. This analysis does not assume that families receive any multi-child discounts for child care, nor do we assume that families with multiple adults cease incurring child care costs when one adult is not working. This ensures that the child care thresholds set in this analysis account for the costs that would be incurred if all working-age adults in a given household were willing and able to participate in the labor force.

Transportation

Data for transportation costs are derived from the Consumer Expenditure Survey (CEX) and the Center for Neighborhood Technology's (CNT) Housing and Transportation Affordability Index.

The CEX, a nationally representative panel survey on consumer spending, groups members within each respondent consumer unit into the following age groups:

Variable	*Description*
AS_COMP5	Number of members under age 2 in consumer unit
$AS\_COMP3 + AS\_COMP4$	Number of members aged 2 through 15 in consumer unit
$AS\_COMP1 + AS\_COMP2$	Number of members aged 16 and over in consumer unit
$PERSLT18$	Number of members aged 0 through 17 in consumer unit
$PERSOT64$	Number of members aged 65 and over in consumer unit

Separately, CEX provides a categorical variable representing summary age characteristics for all children of consumer unit's reference person, including adults:

Variable	*Description*
$CHILDAGE_0$	No children in consumer unit
$CHILDAGE_1$	Oldest child aged 0 to 6
$CHILDAGE_2$	Oldest child aged 6-11 and at least one child aged 0 to 6
$CHILDAGE_3$	All children aged 6 to 11
$CHILDAGE_4$	Oldest child aged 12 to 17 and at least one child aged 0 to 12
$CHILDAGE_5$	All children aged 12 to 17
$CHILDAGE_6$	Oldest child aged 18-plus and at least one child aged 0 to 17
$CHILDAGE_7$	All children aged 18-plus

For the purposes of this analysis, we do not differentiate between $CHILDAGE_0$ and $CHILDAGE_7$, as our model is contingent upon only the age composition of each consumer unit, rather than parent-child relationships between members.

The categories provided in $CHILDAGE$ are themselves mutually exclusive and collectively exhaustive, but cannot be used to definitively disaggregate each consumer unit into the number of children within more granular age categories. To synthetically estimate these age categories, we combine the sample with annual age-level population data from Lightcast, and weight each age group based on relevant number of infants (0 to 2), preschoolers (3 to 5), school-aged children (6 to 12), and teenagers (13 to 17), so that where a_k represents the total number of children in each age group k, where $(k = 2, \dots 15)$:

$$x = \begin{bmatrix} I \\ C_{2,15} \\ T_{16,17} \end{bmatrix} = \begin{bmatrix} infants \\ children\ 2-15 \\ children\ 16-17 \end{bmatrix}, \text{ and } y = \begin{bmatrix} I' \\ P \\ S \\ t \end{bmatrix} = \begin{bmatrix} infants\ (adj.) \\ preschoolers \\ schoolagers \\ teenagers \end{bmatrix},$$

where I' represents an adjusted number of infants so that it is inclusive of two-year-old children. These estimates are then populated into the log-linear prediction model detailed below:

$$log(v_h) = \beta_0 + \gamma Y_h + \delta u_h + \beta_1 log(max(n_h, 0) + 1) + \beta_2 log(\max(a_h, 0)) + \sum_{k \in \{i,p,s,t\}} \beta_k log\big(\max(C_{hk,0}) + 1\big) + \sum_k \theta_k log\big(\max(a_{h,0}\big) \cdot log(\max(c_{hk}, 0) + 1) + \varepsilon_h$$

where for each household h, where v represents annual spending, γ represents fixed effects for each year Y, δ represents fixed effects for binary rural/urban status u, n represents the number of vehicles in the household, a represents number of adults (18-plus), and c represents number of children. Additionally, for each child age group k, infants (0 to 2), preschoolers (3 to 5), school-aged children (6 to 12) and teenagers (13 to 17) are represented by i, p, s, and t, respectively, and θ represents the interaction effects between the number of adults and the number of children in each age group. Note that any term $log(\max(a_h, 0))$ cannot be used in any prediction model for households or consumer units with zero adults. Using coefficients from this prediction model allows us to estimate costs for a broader range of family types than may be natively available in the CEX data.

Finally, to regionalize the estimates obtained through the national panel data, we modify the model to utilize CNT's county-level data on average vehicles per household and costs for vehicle purchases and maintenance, gasoline, and public transportation, and re-estimate spending based on these revised coefficients. As prescribed by the above prediction model, county-level estimates are partially dependent on the urban/rural categorization of each county.

Health care

This analysis estimates health care costs by county using data from the Institute for Health Metrics and Evaluation's (IHME) U.S. Health Spending Estimates database, a comprehensive record of county-level spending on insurance and out-of-pocket medical costs by age, health condition, and payer. Individual per-capita spending estimates for each age are then aggregated into the topline age categories used in this analysis and crossed with our family archetype matrix. We assume that health insurance costs reflect spending on private health plans. For more information on IHME's data and methodology, see https://vizhub.healthdata.org.

Other

This analysis incorporates two primary components into the "other" category: 1) phone and internet costs; and 2) all other necessities. Phone and internet costs are derived using a similar methodology to that described in Transportation (above), by regionalizing nationally representative CEX data with local estimates provided by BroadbandNow Research in their

United States County Broadband Statistics report. These costs are totaled with all others described above to provide a preliminary cost-of-living threshold, absent other necessities. In alignment with the methodology utilized by the University of Washington in their Self-Sufficiency Standard, we assume that all other necessities account for 10% of each household's budget, so that:

$$other = (food + childcare + healthcare + housing + transportation + tech) \cdot \frac{1}{9}.$$

Time series imputation

While all data sources described above provide data in some time-series format, these sources are updated at different frequencies, and therefore cover disparate periods of analysis. To fill in gaps in the time series, we perform a multi-stage imputation on each cost element by covering it into log form and extrapolating across missing values over time through both: 1) random walk; and 2) Kalman smoothing to account for more complex temporal dynamics within each county and family archetype. Grouping by county, missing values for each cost factor were then imputed sequentially using trends in area median income (AMI), cost-specific inflation trends, and the random walk/Kalman-predicted trends for that cost factor. This model can be described as follows:

$$\hat{y}_{c,f,t}^{(k_j)} = \mathcal{F}^{(k_j)}\left(x_{c,f,t}, \iota_{c,t}^{(k_j)}, \hat{y}_{c,f,t}^{(k_j),rw}, \hat{y}_{c,f,t}^{(k_j),ks}\right), t \in \mathcal{M}_{c,f}^{(k_j)}, \text{ where:}$$

- $\tilde{y}$ represents the log-transformed value $log(1+y)$
- $\mathcal{F}^{(k_j)}$ represents a random forest with predictive mean matching (PMM)
- $x_{c,f,t}$ represents any other demographic, labor market, or structural coefficients where applicable (e.g., population)
- $\iota_{c,t}^{(k_j)}$ represents inflation indices for each cost factor k_j
- $\hat{y}_{c,f,t}^{(k_j),rw}$ and $\hat{y}_{c,f,t}^{(k_j),ks}$ represent random-walk and Kalman predictors, where:
    - $\hat{y}_{c,f,t}^{(k_j),rw} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 t & t \leq t^* \\ \hat{\beta}_0 + \hat{\beta}_1 t^* + \hat{\beta}_1 (t - t^*) & t > t^* \end{cases}$, where t^* represents the maximum time period
    - $\hat{y}_{c,t}^{(k_j),ks} = E\left[\tilde{y}_{c,t}^{(k)} \mid \tilde{Y}_{c,1:T}^{obs}\right]$, where $\tilde{Y}_{c,1:T}^{obs}$ represents partially observed sequences across all T periods.