Estimating Food Waste at the Individual Household Level *

Yang Yu and Edward C. Jaenicke §

Abstract

We estimate food waste at the individual household level indirectly using a stochastic production frontier approach. The estimated average percentage of waste is about 30%-32%. In addition to a baseline model, we also develop two models that tackle the issue of missing information on physical activities–one model uses proxy and instrument variables, and the other applies data imputation technique. Based on the results, we are able to explore the relationship between food waste and important demographic variables, and we find that household food insecurity, SNAP participation, and larger household sizes are associated with less food waste, whereas healthy diet practices and higher income lead to more waste.

Keywords: Food waste, Stochastic frontier, Household production.

1 Introduction

As a global economic and environmental problem, unnecessary food waste deserves attention from researchers in academic and government institutions, as well as nonprofit organizations. Institutions such as the USDA's Economic Research Service (USDA-ERS) and the National

*This research is funded, in part, by the USDA-NIFA-AFRI Exploratory Program grant 2017-67030-26611.

§Department of Agricultural Economics, Sociology, and Education, Penn State University, Unversity Park, PA 16802. Yang Yu: yuy138@psu.edu, Edward Jeanicke: ecj3@psu.edu

Resources Defense Council (NRDC) have reported aggregate food waste on the national level (Buzby et al., 2014; Leib et al., 2013; Muth et al., 2011). Their annual estimates of food waste range from 30% to 40% of the total food supply in U.S., which is about $120-160 billion in value.

The estimates in these reports attempt to document the importance of food waste. More generally, scholarly research papers on food waste can arguably be classified into the following four types: (i) measurement of aggregate level estimates of percentage waste or critiques of these estimates (Bellemare et al., 2017; Buzby and Guthrie, 2002; Buzby et al., 2009,1; Garrone et al., 2014; Leib et al., 2013; Muth et al., 2011; Quested and Parry, 2011); (ii) attempts to identify reasons for household wasting behavior based on survey data or behavioral experiments (Neff et al., 2015; Porpino et al., 2015; Qi and Roe, 2016; Reynolds et al., 2014; Secondi et al., 2015; Stefan et al., 2013; Wilson et al., 2017); (iii) impact on environment and sustainable growth, including greenhouse gas emissions from decomposition of wasted food (Beretta et al., 2013; Chapagain and James, 2011; Quested and Parry, 2011; Venkat, 2011); and (iv) theoretical supply-chain analysis of perishable goods by operational research methods (Akçay et al., 2010; Van Donselaar and Broekmeulen, 2012; Wang and Li, 2012).

Recently, Bellemare et al. (2017) take aim at categories (i) and (iv) by proposing that definitions of food waste and efforts to measure it need to better account for what is truly wasted versus what is merely diverted in the supply chain. Their definition implies that "the cost of food waste is equal to total value of the food that goes to the landfill at each stage of the supply chain" (Bellemare et al., 2017). While efforts like these can create better precision when discussing or estimating food waste, and better linkages across categories of food-waste research, they do not address one fairly glaring gap in the research. Despite the importance of this topic and the emerging body of literature, we have not seen anyone successfully estimates food waste at the individual household level. Consequently, little is known about the role that heterogeneous demographics across households play in determining food waste. One reason for this gap is data: In general, there is little or no observable micro-level data

on food waste at the individual household level. Therefore estimating food waste at such a scale has been a difficult task.

In this paper, we propose a novel, indirect way to overcome the data obstacle. Instead of attempting to directly measure food waste, we start by employing a household production function in which food waste is considered as an inefficiency component and estimate the waste indirectly. Specifically, we consider the household food consumption as a process that converts food inputs into chemical energy that meets the requirement of human body's metabolic process and additional energy demand from physical activities. This production function tells us, from a nutrition science perspective, how various food-group contents are transformed into energy expenditure. The specific econometric technique is based on mature research methods of stochastic frontier analysis that typically investigate production efficiency analysis. Within the stochastic production framework, we also develop one of the first empirical applications of instrumental variables method in this literature using Limited Information Maximum Likelihood (LIML).

Aside from the scientific foundations that inspire our model, another novelty that sets our paper apart from the existing research is our choice of directly measurable quantities, e.g., energy requirement and food purchases. These measurements are made feasible by utilizing the USDA's National Household Food Acquisition and Purchase Survey (FoodAPS). For a sample of 4,826 households, FoodAPS provide reasonably complete information on (i) household demographic variables, including income, education, and health outcomes, (ii) biological measures of each household member, and (iii) detailed data, including food categories, food quantities, and nutrition information, on food purchased for at-home and away-from-home consumption, for a period of seven days.

Our first model considers a stochastic production of nine groups of food inputs. The distributional assumption is normal—half-normal, with the variance of the inefficiency conditioned on three household-level demographic variables–income, self-reported quality of diet, and self-report, survey-based measurement of household food insecurity. Our second model

adds a proxy variable and an instrumental variable for the household's physical activities, while in the third and final model, we impute physical activity levels using the National Health and Nutrition Examination Survey (NHANES).

Our estimates show that the average amounts of food wasted at the household level are 31.9%, 30.4%, and 30.1% in the three models. By adding physical activities into the model, the waste estimates are slightly decreased by about 1.5%-2%. In addition, we examine how household-specific attributes explain the variation of our food-waste estimates. We find that food insecurity and SNAP participation are associated with less food waste, while healthy diet practices and higher income lead to more household-level food waste. An examination of the data confirms our hypothesis– healthier diets include significantly larger shares of perishable fresh produce, which adds to food waste. Furthermore, larger households achieve better food management and therefore less waste. These results allow us to further investigate the feasibility and effectiveness of possible food-waste prevention policies that are aimed at particular food types, retail environment, and, more importantly, at particular household types.

The rest of the paper is organized as follows: Section 2 presents model specification and econometric methods. Section 3 provides detailed discussions on the data and main results. Section 4 conducts several robustness checks on choice of variables, and Section 5 concludes the paper.

2 Model and Estimation

2.1 Baseline Model (Model 1)

In most cases, directly measuring food waste is not feasible due to the difficulty of tracking and recording. More practical approaches would consider indirect implications of food waste and trace back to its source. Our model takes this direction. We model the household food consumption as a production process that converts food inputs into chemical energy

that meets the requirement of human body's metabolic process and additional energy demand from physical activities. We then treat the lower-than-predicted output or production inefficiency as a consequence of uneaten food, taking heterogeneous demographics into consideration. Thus, uneaten food, indirectly measured, becomes our operational definition of household-level food wasted[1].

Household h's production process is assumed to take the form in equation (1). The output, $Y(b_h, PA_h)$, is a function that takes into account of household members' biological measures and physical activities. The vector b_h contains every individual's weight, height, age, and gender that serve as a means to capture basic metabolism rate. And PA_h represents the physical activities. The production technology $F(x_h, d_h)$ is a function of food input vector x_h, measured either in weight or calorie contents, and a set of household demographic variables d_h that determine inefficiency (food waste).

$$Y(b_h, PA_h) = F(x_h, d_h) \tag{1}$$

The ideal choice of the functional form of $Y(b_h, PA_h)$ would be one that exhibits both scientific basis and intuitive interpretation. If we assume, as in Hall et al. (2009), that each individual maintained a state of energy balance during the survey period, then a natural starting point of $Y(\cdot)$ would be the sum of the total energy expenditures across all household members. The most commonly used method in medical research to calculate total expenditure is based on the Basal Metabolic Rate (BMR) and Physical Activity Level (FAO/WHO/UNU, 1985; Institute of Medicine, 2005; Scrimshaw et al., 1996). The BMR is the amount of energy required to maintain basic body functioning, calculated through the revised Harris-Benedict Equation using weight, height, age, and gender (Roza and Shizgal, 1984)[2]. Typically, BMR accounts for 65 to 75% of an individual's total energy expenditures (Institute of Medicine, 2005). The Physical Activity Level is a multiplier, ranging from 1

[1]Note that this operational definition ignores the main points raised by Bellemare et al. (2017), namely that some portion of this uneaten food may be diverted to a productive use instead of ending up in a landfill.

[2]The equations are provided in the appendix.

to 2.5, that represents the ratio of total energy expenditure to BMR. It includes thermal effect of food and additional energy needed to perform daily activities and exercises such as household task, walking, and cycling. For example, suppose a person's BMR is 70% of his/her total energy expenditure, the physical activity level is then $1/0.7 \approx 1.43$.

Let us denote $y_h(b_h)$ as the total BMR of all members in household h, and PA_h as the household average physical activity level, and propose the following specification for $Y(b_h, PA_h)$:

$$Y(b_h, PA_h) = y_h(b_h) \cdot PA_h \tag{2}$$

Note that both y_h and PA_h are at the household level. Ideally, we would calculate total energy expenditure for each individual before aggregating. Since FoodAPS does not provide information on physical activities, this specification allows us to rely on y_h as the operational measure of output after taking logarithm. In the baseline model, we impose distributional assumptions on PA_h to complete the specification. In the second model, a proxy variable for PA_h will be provided, as well as an instrumental variable to cope with potential endogeneity of the proxy. However, this aggregation concern is resolved in our third model when we impute values of physical activity for each household member using information in NHANES data.

The full specification of the baseline model (Model 1) is an extension of the Stochastic Frontier model (Aigner et al., 1977; Fried et al., 2008; Jondrow et al., 1982). Denote $x_h = (x_{1,h}, x_{2,h}, ...x_{I,h})'$ as a vector of amount of group i food, in weight or calorie content, purchased by this household[3]. We formulate the production technology $F(x_h, d_h)$ in the translog form where v_h is the white noise and u_h is production inefficiency due to food waste:

[3]Our main results in Section 3 are based on weight (grams). The calorie-content-based estimation is presented in the robustness check in Section 4.

$$\log y_h = \alpha_0 + \sum_{i=1}^{I} \alpha_i \log x_{i,h} + \sum_{i=1}^{I} \sum_{j \leq i} \beta_{i,j} \log x_{i,h} \log x_{j,h} + v_h - u_h \tag{3}$$

Note that the output was originally $\log Y(b_h, PA_h) = \log y_h + \log PA_h$. Hence there was a $-\log PA_h$ term on the right-hand side of equation (3). In Model 1, we tackle the issue of missing information on physical activities by assuming that $-\log PA_h$ is independent of all the explanatory variables and its population distribution is completely captured by the distribution of α_0 and v_h[4]. However, in case this assumption fails, there lacks mechanism to strictly prevent the impact of the missing variable on the predicted value of food waste. The issue will be addressed in Model 2 and Model 3.

As usually assumed in normal–half-normal stochastic frontier models, the white noise v_h is drawn from a normal distribution $N(0, \sigma_v^2)$. In addition, the inefficiency term, u_h, is drawn from a half-normal distribution $N^+(0, \sigma_{u_h}^2)$ and is heteroskedastic:

$$\sigma_{u_h}^2 = \exp(\gamma_0 + \gamma' d_h)$$

d_h is a set of demographic variables that may affect food wasting behavior. It is noteworthy that $\sigma_{u_h}^2$ does not only determine the variance of u_h but also its mean. When a demographic variable in d_h results larger $\sigma_{u_h}^2$, it induces more food waste, on average. In addition, we do not impose restrictions on the parameters–α, β, and γ's since we need not assume this production function to be homogeneous or concave. Hence typical issues involved with translog, e.g., monotonicity and global concavity, do not undermine the validity of the results.

Commonly used estimation approaches for stochastic frontier models are the corrected OLS and maximum likelihood. We choose the later as it better fits our purpose of extending the model to accommodate proxy and instrumental variables in Model 2 which uses Limited Information Maximum Likelihood (LIML). Since the white noise v_h and the inefficiency term

[4]As discussed in Section 2.4, the logarithm of imputed physical activity levels approximate a normal distribution, yet negatively skewed.

u_h are not disentangled before the estimation, the likelihood function is based on $\varepsilon_h = v_h - u_h$. Its density can be derived straightforwardly from the independence assumption between v_h and u_h , and a change of variable integration:

$$f_{\epsilon_h}(\varepsilon_h) = \frac{2}{\sigma_h}\phi(\frac{\varepsilon_h}{\sigma_h})\Phi(-\frac{\lambda_h \varepsilon_h}{\sigma_h})$$

where $\sigma_h^2 = \sigma_v^2 + \sigma_{u_h}^2$ and $\lambda_h = \sigma_{u_h}/\sigma_v$. $\phi(\cdot)$ and $\Phi(\cdot)$ are density and cumulative distribution functions of the standard normal distribution, respectively. The maximum likelihood is then performed on $\sum_h \log f_{\varepsilon_h}(\varepsilon_h)$ to obtain parameter estimates $(\hat{\alpha}_0, \hat{\alpha}, \hat{\gamma}_0, \hat{\gamma}, \hat{\sigma}_v^2)$. Intermediate household-specific parameters $\hat{\sigma}_{u_h}^2$, $\hat{\sigma}_h^2$, and $\hat{\lambda}_h$ are then calculated for each observation.

The translog specification in equation (3) is a flexible functional form that is adequate in most cases. Nonetheless, as we use the household total energy expenditure as the output, one would wonder if we can simply treat the total calorie content from all food groups as the input, that is $\log y_h = \alpha_0 + \alpha_1 \log(\text{total calories}) + v_h - u_h$[5]. This single-input production suffers from several weaknesses due to its excessive simplification. First, aggregating calorie values of food products based on their nutrition labels have been criticized for ignoring other substantial factors such as food composition (Trivedi, 2009). Food digestion itself requires energy (the thermal effect of food) which typically accounts for about 10% of total energy expenditure (McArdle et al., 1986). Different types of food take different amounts of energy to digest, even when they contain the same calorie content on the nutrition labels. For instance, protein-intense food generates more heat in postprandial thermogenesis than carbohydrate and lipids-intense food, thereby provides less "effective" chemical energy that is used by the body (Johnston et al., 2002). Consequently, calorie contents from different types of food are not perfect substitutes, hence not linearly additive. Finally, dividing food into categories allows a closer examination on the impact of food composition on food waste. For instance, our results show that more consumption of fruit and vegetables is associated with significantly more food waste. For the reasons provided above, we do not suggest the

[5]This specification yields a slightly higher estimate of average food waste at about 40%.

use of the overly simplified single-input model.

2.2 Percentage Food Waste

Our primary goal is to provide an estimate of percentage food waste at the individual household level. The idea is to transform the output distance function into an input distance function. This task can be efficiently accomplished once we have an estimate of the output inefficiency term $\hat{u_h}$ for each household. Note that since the dependent variable is the logarithm of total household BMR, $\hat{u_h}$ also represents the approximate percentage waste in household BMR. The closed form prediction of u_h post-estimation is well established in the stochastic frontier literature (for example, Jondrow et al. (1982)). The solution is given as the following, where $\hat{b}_h = \hat{\varepsilon}_h\hat{\lambda}_h/\hat{\sigma}_h$:

$$\begin{aligned} \hat{u_h} &= E(u_h|\hat{\varepsilon}_h) \\ &= \frac{\hat{\sigma}_{u_h}\hat{\sigma_v}}{\hat{\sigma_h}}[\frac{\phi(\hat{b}_h)}{1-\Phi(\hat{b}_h)} - \hat{b}_h] \end{aligned} \tag{4}$$

Recall that the household production function is given in equation (3). For exploratory purpose, let us assume that, for household h, food from all I groups are wasted in the same proportion, δ_h. Then we have the following relationship:

$$\begin{aligned} &\sum_{i=1}^{I}\alpha_i \log x_{i,h} + \sum_{i=1}^{I}\sum_{j\leq i}\beta_{i,j}\log x_{i,h}\log x_{j,h} - u_h \\ =&\sum_{i=1}^{I}\alpha_i \log(1-\delta_h)x_{i,h} + \sum_{i=1}^{I}\sum_{j\leq i}\beta_{i,j}\log(1-\delta_h)x_{i,h}\log(1-\delta_h)x_{j,h} \end{aligned} \tag{5}$$

On the second line, the $-u_h$ term is transformed into a multiplication factor $(1-\delta_h)$ on each $x_{i,h}$ [6]. Rearrangements of the equation results the quadratic solutions for $\log(1-\delta_h)$:

[6] Similar transformation is used in Reinhard et al. (1999) for a single-input case, whereas Kurkalova and Carriquiry (2003) considers transformation for multiple inputs in a Cobb-Douglas model. Alternatively, Kumbhakar and Tsionas (2006) provide an approach that directly formulates input-inefficiency therm δ_h as a random variable and uses a simulated ML estimation.

$$\log(1-\delta_h) = \frac{-B_h \pm \sqrt{B_h^2 - 4AC_h}}{2A}$$

where $A = \sum_{i=1}^{I}\sum_{j\leq i}\beta_{i,j}$, $B_h = \sum_{i=1}^{I}\alpha_i + \sum_{i=1}^{I}\sum_{j\leq i}\beta_{i,j}(\log x_{i,h} + \log x_{j,h})$, and $C_h = u_h$. This gives us two sets of predicted food waste as a result of quadratic solutions. Even though we did not impose any theoretical restrictions on parameters ex ante, only one of them makes economic sense: $\log(1-\delta_h) = (-B_h + \sqrt{B_h^2 - 4AC_h})/2A$. The reason is that we expect a positive correlation between $\hat{u}_h$ and $\hat{\delta}_h$–more output inefficiency implies more input waste. A simple verification through partial derivatives proves that the other solution gives a negative relationship between $\hat{u}_h$ and $\hat{\delta}_h$. [7]

The predicted percentage food waste for each household, $\hat{\delta}_h$, is then calculated based on parameter values of $\hat{\alpha}$ and $\hat{\beta}$, the predicted output inefficiency $\hat{u_h}$, and the observation-level food purchase $x_{i,h}$'s:

$$\%\text{ food waste} = \hat{\delta}_h = 1 - \exp\left(\frac{-\hat{B}_h + \sqrt{\hat{B}_h^2 - 4\hat{A}\hat{C}_h}}{2\hat{A}}\right) \tag{6}$$

This estimate sets our study apart from the existing research on food waste for several reasons. First and foremost, by our knowledge, this is the first study that provides individual household-level estimates of food waste. Moreover, it opens a channel to conducting post-estimation analysis on sub-group comparisons based on various demographic measures, as well as implications for waste prevention policies that are aimed at particular food types and retail environment.

2.3 Physical Activities – Proxy and Instrumental Variables (Model 2)

In this section, we turn to the issue of missing information on physical activities (PA_h). From an energy expenditure perspective, Archer et al. (2016) suggest to consider physical

[7] In fact, our estimation shows that the other set of the solutions yields values of $\hat{\delta}_h$ infinitely close to 1.

activities in estimating food waste to reach more accurate results. From a technical point of view, if the omitted variable in Model 1 is normally distributed and independent of other explanatory variables including those determining ${\sigma^2}_{u_h}$, then our parameter estimates are consistent. On the other hand, if PA_h fails to meet the conditions, it poses inconsistent parameter estimates that would likely generate biased food waste estimates. Whether the percentage food waste is overestimated or underestimated is a rather complex matter that involves many factors including the signs of correlation and the distributional properties of $-\log PA_h$. We do not explicitly explore the econometric mechanism that determine the bias in this paper. Nonetheless, the results of Model 2 and 3 suggest that Model 1 overestimates waste by about 1.5%.

In the second model (Model 2), we propose a proxy variable for the missing variable PA_h. Though FoodAPS does not contain direct measures of physical activities, it provide some highly indicative variables. One example is the employment status of all working-age individuals. It is a discrete variable of four levels, with 1 meaning unemployed while not searching for a job, and 4 representing employed and working regularly. For each household, we take the average across all working-age members and normalize it to a value between 0 and 1. The rationale of using employment as a proxy is that employed people generally have a higher level of mandatory physical activities. Moreover, among the unemployed individuals in FoodAPS, about 44% is due to retirement, health issues, or disability, who are likely to have less physical activity than those employed[8].

The validity of employment as a proxy is further supported by the NHANES 2011-2012 data. This dataset follow the same coding rule of employment as in FoodAPS. In addition, NHANES contains valuable records on physical activities. As Table A.6 shows, higher value of employment implies higher level of physical activity. In fact, the average Physical Activity Level is 1.64 for the employed and 1.52 for the unemployed.

[8]The implication of health issues and disability are straightforward. As for retirement, it can be regarded as an indicator of age. Our results in the next section on NHANES show that age is negatively correlated with physical activity level (Table A.6).

Despite these features of the employment status variable, the proxy itself is not completely free of endogeneity concerns. Indeed, employment does not represent all types of physical activities. Recreational activities, for instance, may not be fully explained by employment status. To minimize the endogeneity issue of the proxy variable, we adopt an instrumental variable approach and apply a version of the Limited Information Maximum Likelihood (LIML) that is derived specifically for the stochastic frontier analysis.

Our choice of the instrument is the frequency of weekend shopping. It is measured as the percentage of a household's shopping trips that occurred during weekends. On the one hand, whether a household shops on weekends or weekdays is highly correlated with its employment status. In FoodAPS data, households of the highest 25% employment status spend 34% of their trips on weekends, while the percentage of weekend trips is 26% for those of the lowest 25% employment status. On the other hand, the instrument is exogenous in a sense that it merely represents a choice of shopping schedule, not purchase decisions. For instance, it is unlikely to affect the total food purchases over a whole week. In addition, it is reasonable to assume that such shopping schedule is uncorrelated with physical activities not represented by employment status such as recreational activities. Hence the instrument is connected with the output only through the proxy variable.

There are several recent papers that tackle the issue of endogeneity in stochastic frontier models. Maximum likelihood methods are studied in Kutlu (2010) and Amsler et al. (2016), while Tran and Tsionas (2015) develop a copula approach without requiring external instruments. Our notation for the LIML estimation is adopted from Amsler et al. (2016). In the context of our framework, let us specify Model 2 by first adding a proxy variable for physical activities, $\widetilde{PA}_h$, which is the household employment status:

$$\log y_h = \alpha_0 + \alpha_{PA} \log \widetilde{PA}_h + \sum_{i=1}^{I} \alpha_i \log x_{i,h} + \sum_{i=1}^{I} \sum_{j \leq i} \beta_{i,j} \log x_{i,h} \log x_{j,h} + v_h - u_h \qquad (7)$$

We assume that after adding the proxy, the potential endogeneity resides in $\widetilde{PA}_h$ whereas the food purchases $x_{i,h}$'s are exogenous. In addition, we do not include the interactions of the proxy variable and food purchases as it leads to significant difficulty in finding enough instruments to compensate. Similar to linear regressions, the idea of LIML in stochastic frontier models is to add a set of reduced-form equations for the endogenous variables, and to estimate them jointly with the original equation. Specifically, we add the following reduced-form equation:

$$\log \widetilde{PA}_h = \pi_0 + \pi_{IV} \log z_h + \sum_{i=1}^{I} \pi_i \log x_{i,h} + \sum_{i=1}^{I} \sum_{j \leq i} \pi_{i,j} \log x_{i,h} \log x_{j,h} + \eta_h \tag{8}$$

where the instrument z_h is the percent of household weekend shopping trips. Since we have one endogenous variable and one instrument, the identification is exact here. For cases of multiple endogenous variables, the formulation and derivation of likelihood can be found in Amsler et al. (2016).

As in Kutlu (2010) and the LIML case in Amsler et al. (2016), we assume that η_h is correlated with v_h, but not with u_h. Let us denote $\psi_h = (v_h, \eta_h)$ and assume its distribution as following:

$$\psi_h \sim N(0, \Omega), \ \Omega = \begin{bmatrix} \sigma_v^2 & \sigma_{v\eta} \\ \sigma_{\eta v} & \sigma_\eta^2 \end{bmatrix}$$

Then the presence of endogeneity corresponds to the case when $\sigma_{\eta v} = \sigma_{v\eta} \neq 0$. The likelihood function is the joint density of ε_h and η_h, which can be derived analytically by change of variables integration. The key assumptions needed to derive this density is the independence between η_h and u_h, the normal distribution, as well as the independence between v_h and u_h:

$$
\begin{aligned}
f_{\varepsilon_h,\eta_h}(\varepsilon_h,\eta_h) &= \text{constant}\cdot\sigma_\eta\cdot\exp\left(-\frac{\eta_h^2}{2\sigma_\eta^2}\right)\cdot\sigma_h^{-1} \\
&\quad\cdot\phi\left(\frac{\varepsilon_h-\mu_{ch}}{\sigma_h}\right)\cdot\Phi\left(-\frac{\lambda_h(\varepsilon_h-\mu_{ch})}{\sigma_h}\right)
\end{aligned}
$$

where $\mu_{c,h}=(\sigma_{v\eta}/\sigma_\eta^2)\eta_h$, $\sigma_h^2=\sigma_{u_h}^2+\sigma_{c,h}^2$, $\sigma_{c,h}^2=\sigma_v^2-\sigma_{v\eta}^2/\sigma_\eta^2$, and $\lambda_h=\sigma_{u_h}/\sigma_{c,h}$. Finally, we can predict the inefficiency term u_h by its mean conditional on ε_h and η_h.

$$
\begin{aligned}
\hat{u_h}^{LIML} &= E(u_h|\hat{\varepsilon}_h,\hat{\eta}_h) \\
&= \hat{\sigma}_h^*[\Lambda(\hat{h}_h)-\hat{h}_h]
\end{aligned}
$$

where $\hat{\varepsilon_h}$ and $\hat{\eta_h}$ are residuals from the LIML estimation, $\hat{\sigma}_h^*=\frac{\hat{\sigma}_{u_h}\hat{\sigma}_{c,h}}{\hat{\sigma}_h}$, $\hat{h}_h=\frac{\hat{\lambda}_h}{\hat{\sigma}_h}(\hat{\varepsilon}_h-\hat{\mu}_{c,h})$, and $\Lambda(\hat{h}_h)=\phi(\hat{h}_h)/[1-\Phi(\hat{h}_h)]$. The percentage food waste is carried out the same way as the baseline model (equation (6)).

2.4 Physical Activities–Data Imputation (Model 3)

In this section, we present an alternative approach to cope with the missing physical activities in FoodAPS data. The National Health and Nutrition Examination Survey (NHANES) 2011-2012 provides valuable records of different types of physical activities of each survey participant that can be transformed to the standard Physical Activity Levels that range from 1 to 2.5. Though different datasets, FoodAPS and NHANES are both nationally representative. Additionally, they follow the same coding rules for many demographic variables, for instance, employment status. This enables us to obtain estimates of correlations between physical activities and various individual-specific characteristics, and apply these estimates to impute the missing values in FoodAPS for each household member. The output in equation (2), $Y(b_h,PA_h)$, now takes the form of the sum of individual member's total energy expenditures, where m is the index for household members and $Size_h$ is household size:

$$
Y_h=\sum_{m=1}^{Size_h} BMR_{m,h}\cdot\widehat{PA}_{m,h} \tag{9}
$$

Here, $\widehat{PA}_{m,h}$ is the second-stage imputed physical activity level for member m in household h, and $BMR_{m,h}$ is this member's basal metabolic rate. Note that the dependent variable in the stochastic frontier estimation (equation 3) is now $\log Y_h$.

In the first stage, we run the regression of physical activity level PA_t^{NHANES} on a set of individual characteristics g_t^{NHANES}. There are two age groups: For ages of 12 to 19, g_t^{NHANES} contains weight, height, age, and gender; and for ages 20 and above, employment status and education level are added[9]. Employment status has the same four values and in FoodAPS. And education represents the highest degree received by the survey participant with 1 corresponds to 9th grade and 5 for college degree or higher.

$$PA_t^{NHANES} = \theta_0 + \theta' g_t^{NHANES} \qquad \text{(First Stage)} \tag{10}$$

$$\widehat{PA}_{m,h} = \widehat{\theta}_0 + \widehat{\theta}' g_{m,h}^{FoodAPS} \qquad \text{(Second Stage)} \tag{11}$$

The physical activity levels in NHANES PA_t^{NHANES} are calculated using the Metabolic Equivalents (METs) and its implied increase in physical activity levels. The METs represents the multiples of an individual's resting oxygen uptake. Each value of METs corresponds to a certain amount of increase in physical activity level, depending on how much time spent daily on such activities, as shown in Table 1.

Physical activities are categorized into three types in NHANES: sedentary, moderate, and vigorous. Each survey participant reports how much time he or she spends on each type of activities on a typical day. Moreover, NHANES contains suggested METs values of moderate and vigorous activities, whereas the METs of sedentary activities is taken from Table 12-3 of Institute of Medicine (2005). Then PA^{NHANES} is calculated by equation (12) for each individual. The number 1.1 reflects the base energy requirement plus 10% thermal effect of food. As the readers may notice, the per 1-hour values in the last column of Table

[9]NHANES does not report physical activities for ages under 12.

Table 1: METs and Increase in Physical Activity Level

Activity Type	METs	$\Delta PAL/10min$	$\Delta PAL/1h$
Sedentary Activities ($Type = 1$)	1.5	0.005	0.03
Moderate Activities ($Type = 2$)	4.0	0.029	0.17
Vigorous Activities ($Type = 3$)	8.0	0.067	0.4

Source: Table 12-1, Table 12-2 and Table 12-3 in citation Institute of Medicine (2005), and Appendix 1 in NHANES 2011-2012 Codebook of Physical Activity (PAQ_G). Note: time spent on activities is calculated based on a typical day.

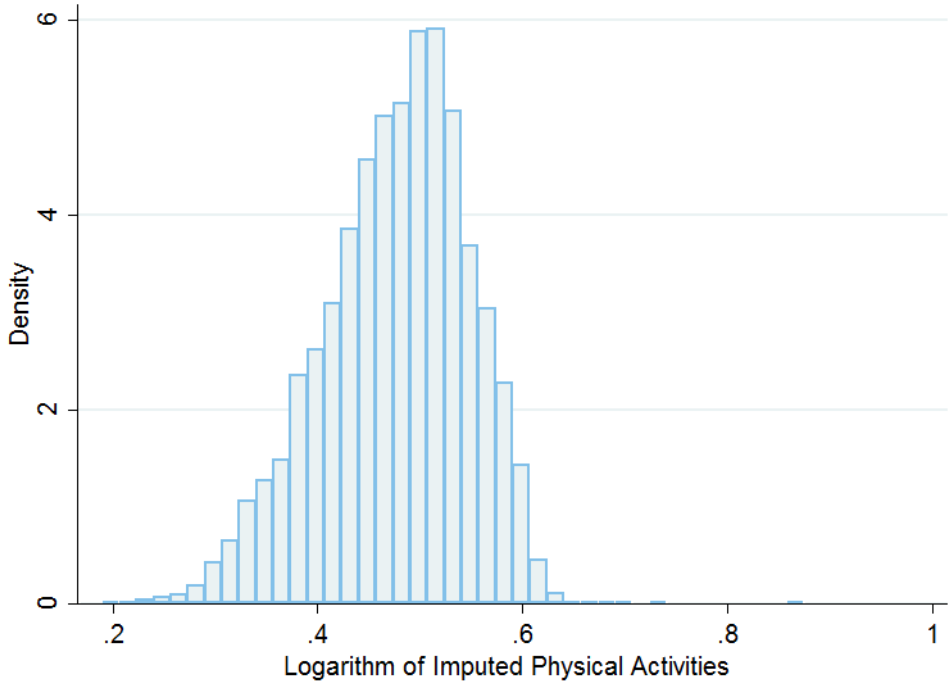
1 are not exactly six times the the per 10-min values due to the nonlinear relationship. For the $Time_{Type}$ variable, our paper uses the per 10-min values in the calculation, while the results change little if using the other.

$$PA^{NHANES} = 1.1 + \sum_{Type=1}^{3} \Delta PAL_{Type} \cdot Time_{Type} \tag{12}$$

The estimation results of first-stage regression are contained in Table A.6 in the appendix. Male consistently have higher physical activity levels than female, in both age groups. Weight is negatively correlated with physical activities while height has a positive correlation. For persons of age 20 and above, employment status and higher education are associated with higher activity levels, while age has a negatively impact.

Once we obtained estimates of coefficients, $\widehat{\theta}_0$ and $\widehat{\theta}'$, we substitute them into second-stage imputation. A summary of the imputed physical activity levels is contained in the appendix. As Figure 1 shows, the distribution of imputed values $\log \widehat{PA}_{m,h}$ has a negative skewness. If, for each household, we take the average value of $\log \widehat{PA}_{m,h}$ across all members, then this average has a similar skewed shape. Roughly speaking, the negative skewness would probably generate overestimated food waste in Model 1. This is because the inefficiency u_h has a half-normal distribution and the missing term $-\log PA_h$ would increase the estimated

Figure 1: **Distribution of** $\log \widehat{PA}_{m,h}$



$\sigma^2_{u_h}$. It is, however, noteworthy that this reasoning may not strictly hold for every case of negatively skewed $\log \widehat{PA}_{m,h}$. As we discussed in Section 2.1, the exact bias of food waste depends on many other factors.

3 Data and Results

3.1 Variables

The dependent variable, $\log y_h$, is the logarithm of the sum of household members' BMR. FoodAPS reports most individuals' body measures, age, and gender. Households with missing member BMR are dropped. In Model 3, we obtain the imputed physical activities and use they to calculate the household total energy expenditure $Y_h = \sum_{m=1}^{Size_h} BMR_{m,h} \cdot \widehat{PA}_{m,h}$ as a refined output measure. A summary of y_h and Y_h is provided in Table 2. The independent variables, $\log x_{1,h}, \log x_{2,h}, ... \log x_{I,h}$, are logarithms of total amount of food acquisition, for

nine groups of food[10]. The food groups are categorized based on the USDA's What We Eat in America (WWEIA) 9-group code. Note that FoodAPS has a tenth group that represents food not coded by the preceding criteria. We combine this last group with the ninth group in USDA code–the later being "infant formula and all other food without a category code". The summary statistics of the total amounts $x_{1,h}, x_{2,h}, ...x_{I,h}$ are presented in Table 3.

There are three demographic variables used to determine the inefficiency term's distribution: household monthly income per adult equivalent , overall self-evaluated diet healthiness, and household food security measure. The income variable is continuous. We use total family income in thousand dollars divided by adult equivalent household size. In calculating adult equivalence, we assign children under age of 6 years a weight of 0.2, between 7 to 12 years a weight of 0.3, and 13 to 17 years a weight of 0.5 (World Bank, 2005). The other two variables take discrete values and are normalized between 0 and 1. Diet healthiness has values 1 to 4, with 1 representing least healthy diet and 4 as the healthiest. Food security measure has three levels with highest value as the most secure[11]. These demographic variables enter the model as explanatory variables for $\sigma^2_{u_h}$.

Table 2: **Summary Statistics-Dependent Variable**

	Mean	Standard Deviation	5% percentile	95% percentile
y_h, Total Household BMR	4293.2	2401.8	1312.9	8699.6
Y_h, Total Household Energy Expenditure	6974.1	4019.82	1933.1	14300.0

[10]Since for some households, there are food groups with zero values, we used $\log(x_{i,h}+1)$ in estimation. The mean amounts of food in the data are typically in thousands. Hence we believe the bias, if any, is negligible. In fact, using $\log(x_{i,h}+0.001)$ would produce the same amount of food waste.

[11]The orders of number values in diet healthiness, food security and employment status are reversed in the original FoodAPS data. For instance, 1 represents the healthiest or most secure in FoodAPS. We reverse the orders to avoid confusion in relating the values and their meanings. Moreover, we reduced the number of levels of diet healthiness and food security to combine marginal small groups.

Table 3: **Summary Statistics–Food Acquisition**

	Mean	Standard Deviation	5% percentile	95% percentile
$x_{1,h}$, Milk and Dairy	3388.0	4276.6	0.0	11712.0
$x_{2,h}$, Protein Foods	1746.4	2264.6	0.0	5692.6
$x_{3,h}$, Mixed Dishes	2661.4	2629.0	0.0	7777.6
$x_{4,h}$, Grains	1548.0	2136.8	0.0	5264.3
$x_{5,h}$, Snacks	1428.4	1861.7	0.0	4951.2
$x_{6,h}$, Fruit and Vegetables	2594.7	2950.9	0.0	8361.0
$x_{7,h}$, Beverages	11099.4	11602.0	0.0	34852.2
$x_{8,h}$, Condiments	1509.3	2286.6	0.0	5972.0
$x_{9,h}$, Infant formula & Uncoded	90.1	610.7	0.0	340.2

Note: The amounts of food acquisition are in total grams.

Our second model contains an additional independent variable, household employment status, which serves as a proxy for physical activities. It has four values, with 1 meaning unemployed and not searching for a job and 4 representing employed. For each household, we take the average value of all working-age members and normalize it to a range of 0 to 1. In addition, the instrumental variable for the proxy is the frequency of household weekend shopping trips, as a percentage share of all shopping trips during the week. In Model 3, an additional demographic variable, education level, is used, which has five levels–from 9th grade graduate to college graduate. The original education variable in FoodAPS has more levels and was re-categorized to five-level as in NHANES. The detailed summary statistics for all demographic variables are contained in the appendix.

3.2 Main Results

Elasticities on Food Groups

Because the number of parameters in our model is more than 60 and the estimated coefficients do not have direct interpretations, we show the elasticities of each food group on the output,

as a means to display the direction and magnitude of the marginal effects[12]. For each household, we calculate the elasticity of group k food as follows:

$$e_{k,h} = \frac{\partial \log y_h}{\partial \log x_{k,h}} = \alpha_k + \sum_{j<k} \beta_{k,j} \log x_{j,h} + 2\beta_{k,k} \log x_{k,h} + \sum_{i>k} \beta_{i,k} \log x_{i,h} \tag{13}$$

Note that this elasticity is observation dependent. Hence we take the sample average elasticity for each food group. The results are shown in Table 4. The first model considers the baseline heteroskedastic stochastic frontier specification. The second model adds household employment status as a proxy variable for physical activities and frequency of weekend shopping as an instrument. Model 2 has more observations than the other two because it yields more households that have solutions for $\hat{\delta}_h$ (equation 6).

Most of the elasticities in these models are positive while the group 8 (Condiments) has negative values in all three models. The negativity on condiments does not undermine our results' validity. First, condiments are not significant in producing output in the sense that its first-order and second-order coefficients, α_8 and $\beta_{8,8}$ are not statistically significant (see Appendix). Moreover, condiments, by their nature, do not contribute to energy intake at a degree comparable to other groups, due to their relatively small share in food composition.

Among the groups with positive elasticities, group 3 (Mixed Dishes) persistently has the highest values, followed by group 7 (Beverages). This result is consistent with our common sense as they are major sources of gaining energy: mixed dishes are typical meals such as pizza and sandwiches, and the majority of beverage items consist of sweetened products such as soda and tea.

Percentage Food Waste

The estimates of average food waste across our sample and their sample standard deviations are presented in Table 5. The percentage food waste estimates in three models are 31.9%, 30.4%, and 30.1%, respectively.

[12]The full estimation results of three models are left in the Appendix.

Table 4: Mean Elasticities

Food Groups	**Model 1**	**Model 2**	**Model 3**
1. Milk and Dairy	0.0634	0.0491	0.0637
2. Protein Foods	0.0243	0.0096	0.0237
3. Mixed Dishes	0.1434	0.1957	0.1574
4. Grains	0.0395	0.0282	0.0408
5. Snacks	0.0045	0.0240	0.0026
6. Fruit and Vegetables	0.0341	0.0511	0.0306
7. Beverages	0.0818	0.1352	0.0881
8. Condiments	-0.0109	-0.0026	-0.0122
9. Infant formula & Uncoded	0.0103	-0.0046	0.0108
Number of Observations	3304	3579	3323

Note: Model 1: Baseline model. Model 2: Proxy-instrumental variable LIML estimation. Model 3: Imputed physical activity levels.

Table 5: Percentage Food Waste

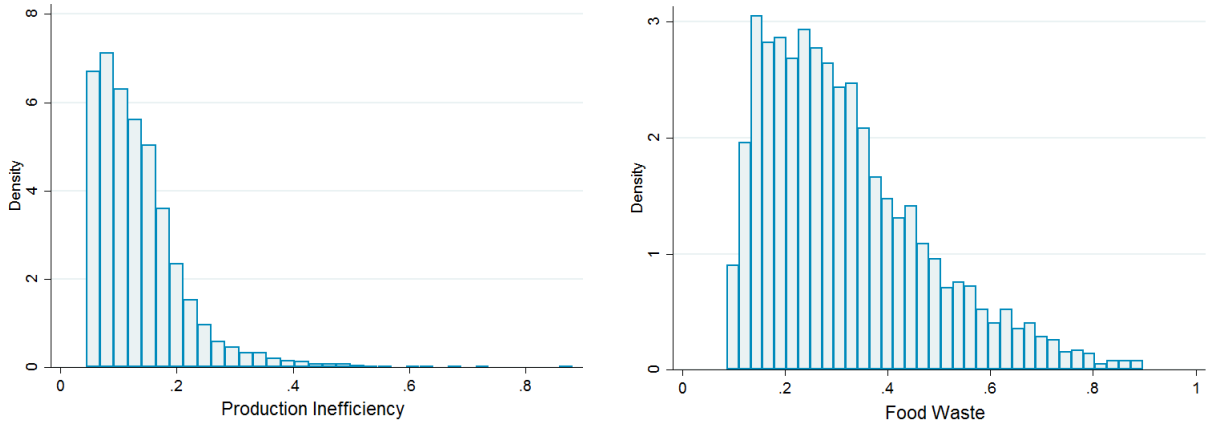
	Model 1	Model 2	Model 3
Average Waste	31.9%	30.4%	30.1%
Standard Deviation	15.8%	16.6%	15.4%

Note: Model 1: Baseline model. Model 2: Proxy-instrumental variable LIML estimation. Model 3: Imputed physical activity levels.

By taking physical activities into consideration, the average food waste decreases by about 2 percent in Model 2 and Model 3. This suggests that Model 1 overestimates food waste with a small bias. As we discussed in Section 2.4, one of the potential reasons is the negative skewness of the missing physical activity levels. Moreover, the coefficient on employment status, α_{PA} in Model 2, is negative, which is consistent with our specification.

Histograms of the waste in terms of inefficiency $\hat{u_h}$ and percentage food $\hat{\delta}_h$ from Model 1 are depicted in Figure 2. As the graphs suggest, the inefficiency $\hat{u_h}$ approximates the half-normal distribution as we assumed, and the percentage food waste assembles a truncated normal or a beta distribution.

Figure 2: Distribution of $\hat{u}_h$ and $\hat{\delta}_h$



Note: $\hat{u}_h$ is the inefficiency in terms of output, $\hat{\delta}_h$ is the food waste.

Food Waste Determinants

We have included three demographic variables as determinants of food waste. Specifically, the variance of the inefficiency term, σ_u^2 , is conditioned on household monthly income per adult equivalent, overall self-evaluated diet healthiness, and household food security measure: $\sigma_{u_h}^2 = \exp(\gamma_0 + \gamma' d_h)$. The estimated $\hat{\gamma}$'s are presented in Table 6. In all three models, they are all statistically significant except food security in Model 3. In fact, food security has a p-value of 0.137 in Model 3, marginally close to significance. Since higher inefficiency $\hat{u_h}$ yields higher food waste $\hat{\delta}_h$, the parameters also indirectly indicate the effects of demographic variables on percentage food waste.

The signs of these demographic variables make good economic sense. Income has a positive impact on food waste (Figure 3). Households facing less constrained budgets are more likely to spend less time managing food purchases and allocations among members. It is also reasonable to believe that they appreciate their food less because they can "afford" wasting.

The diet healthiness measure is positively correlated with food waste. Since a higher diet score represents a healthier diet, this means consuming healthier food leads to more waste (Figure 4). It seems to be inconsistent with common sense that people's awareness of nutritional facts may also suggest they take better care of food. However, one important component of healthy eating is perishable produce such as fruit and vegetables (group 6), which is a major source of food waste. Our data show a persistent relationship between diet healthiness and amount of group 6 food–the households with the highest self-stated diet quality consume 50% more fruit and vegetables than those with the lowest diet quality.

The last determinant, the food security measure, is also positively correlated with food waste in all three models, and significant in Models 1 and 2. We leave the discussion on food security to the next section as it pertains to policy issues.

Table 6: Food Waste Determinants

Description	Model 1	Model 2	Model 3
$\log \sigma^2_{u_h} = \gamma_0 + \gamma' d_h$			
Income	0.3458***	0.4089***	0.3010***
	(0.0601)	(0.0657)	(0.0635)
Healthy Diet	1.5147***	0.9888*	1.7914**
	(0.5871)	(0.5411)	(0.7222)
Food Security	1.9550*	2.3139**	2.1217
	(1.1395)	(1.0984)	(1.4286)

Figure 3: **Higher Income Leads to More Waste**

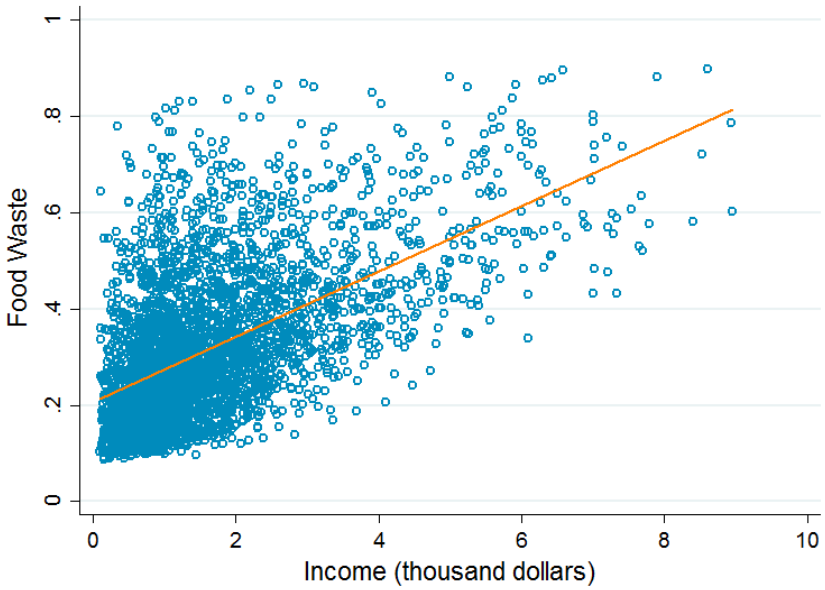
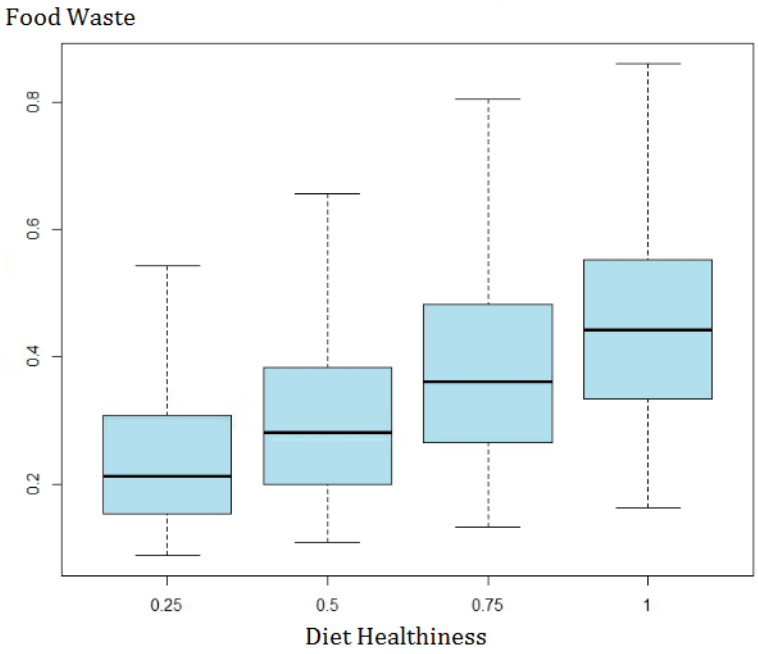


Figure 4: **Healthy Diet Leads to More Waste**



Note: higher values mean healthier diet.

3.3 Food Waste and Household Groups

In this section, we conduct several post-estimation analyses on food waste among various household groups. The discussion focuses on issues that may be of interest to policy makers.

Our results for the food security in the previous section have a sound and intuitive interpretation. Table 7 shows the average percentage waste along four levels of food security and Figure 5 displays a typical box plot of food waste at each level. In all three models, the less secure households waste significantly less than the more secure ones. In fact, the least secure households waste only about half the amount of the most secure, e.g. 20.5% vs. 39.9%, 18.0% vs. 38.9% and 18.6% vs. 38.1%. This persistent pattern is a robust evidence supporting the fact that food insecure households save and appreciate their food more.

Next we turn to two national food assistance programs and their relationship with food waste: the Supplemental Nutrition Assistance Program (SNAP) and the Women, Infants,

Table 7: Food Waste and Food Security

Food Security	Model 1	Model 2	Model 3
Low	20.5%	18.0%	18.6%
Medium	26.9%	25.3%	25.4%
High	39.9%	38.9%	38.1%

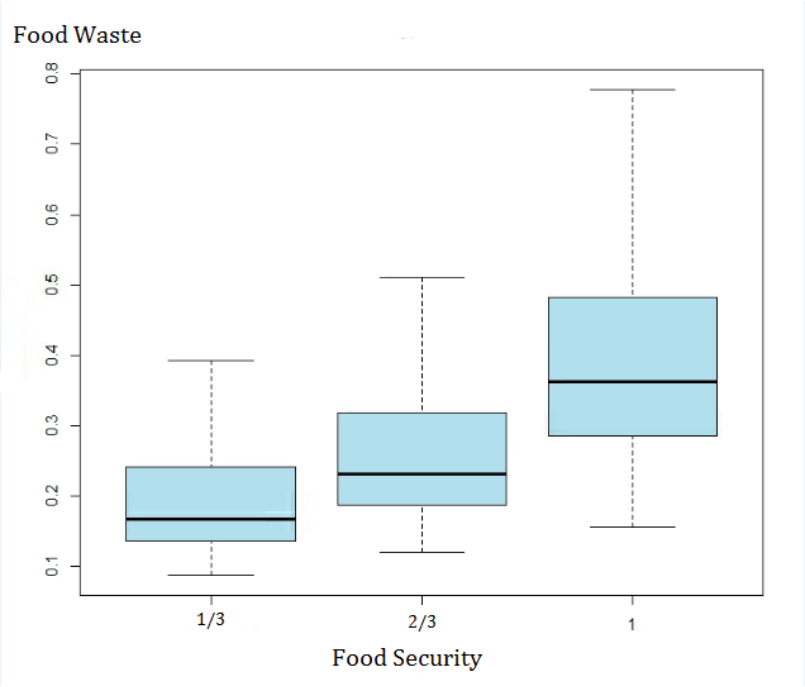
Note: These figures represent the average of the estimated percentage of food wasted for the three levels of stated household food security.

and Children Program (WIC). As Table 8 shows, in all three models, households receiving SNAP benefits waste significantly less (up to 30% less) than non-SNAP households. In the lower part of Table 8, only WIC categorically eligible households are considered. The average food waste for households receiving WIC is persistently less than those not receiving the benefit. However, careful interpretation is needed as to whether the SNAP and WIC households waste less because their income is low or because they take good management of the subsidized food. For instance, for those non-SNAP households, we do not distinguish if they are eligible for SNAP. As for WIC, FoodAPS only reports the categorical eligibility (i.e., female aged 14-49 years old and pregnant or children up to 5 years old) but not the income requirement. Therefore it is not clear how much of the difference can be attributed to each reason, i.e., income or management.

Finally, it is also interesting to observe that the average percentage food waste tends to decrease as the size of households gets larger. In Table 9, the single-member households are associated with the highest rate of food waste–more than 40%, and the rate is reduced to 20% for six-member households[13]. It suggests that larger households may spend more time managing food purchases and more efficiently allocate among the members. A single-member household, on the other hand, is less flexible to remedy over-purchased or near-expiring food.

[13]We only show households up to 6 members, as they account for 99% of the sample.

Figure 5: **Food Insecurity Leads to Less Waste**



Note: higher values mean higher security

Table 8: Food Waste, SNAP and WIC

	Model 1	Model 2	Model 3
SNAP			
Non-SNAP	35.2%	33.6%	33.2%
SNAP	24.7%	23.5%	23.6%
WIC			
Eligible, Not Receiving	28.0%	24.2%	26.2%
Eligible, Receiving	23.5%	20.7%	22.4%

Note: These figures represent the average of the estimated food waste for different SNAP and WIC categories.

Table 9: Food Waste and Household Size

Household Size	Model 1	Model 2	Model 3
1	44.7%	45.5%	41.8%
2	36.3%	34.8%	34.4%
3	29.7%	26.7%	27.9%
4	25.2%	22.2%	23.9%
5	23.0%	19.8%	21.6%
6	20.3%	18.1%	19.3%

Note: First column refers to the number of household members. These figures represent the average of the estimated food waste for various household sizes.

4 Robustness to Specification

In general, a translog model is a flexible form that provides the possibility for a good fit in most cases. In this study, we have obtained reasonable estimates for food waste and impacts of household-specific variables. However, besides the functional form, there are other perspectives of the model that deserve careful examination within this new approach to estimating food waste. In this section, we provide discussions on the robustness of two important specifications of our model–the choice of input units, and choice of demographic variables that determine productivity inefficiency.

4.1 Choice of Input Units

In our specification, the food inputs $x_h = (x_{1,h}, x_{2,h}, ...x_{I,h})'$ are measured by their weights in grams. Here we present results that are based on their calorie contents. Testing our method on a different input measure provides valuable insights on the robustness of the model specification.

Table 10 and Table 11 contain the new estimates of the percentage waste and the demographic determinants. The food waste estimates in all three models are very close to previous estimates. Moreover, the food waste determinants are also in line with our previous numbers, only except that the significance of food security measure are weakened. Finally, applying LIML estimation and imputation on physical activities produce lower food waste estimates than Model 1 as before. Overall, a different choice of input units does not produce a sensitive change in major estimates, which supports the robustness of our model specification.

Table 10: Percentage Food Waste–Calorie Contents

	Model 1	Model 2	Model 3
Average Waste	34.4%	30.7%	33.3%
Standard Deviation	15.8%	17.0%	15.5%
Number of Observations	3681	3678	3695

Table 11: Food Waste Determinants–Calorie Contents

Description	Model 1	Model 2	Model 3
$\log \sigma^2_{u_h} = \gamma_0 + \gamma' d_h$			
Income	0.3534***	0.4084***	0.3101***
	(0.0682)	(0.0692)	(0.0766)
Healthy Diet	1.0068*	0.8690*	1.1849*
	(0.5507)	(0.5233)	(0.6824)
Food Security	1.4681	1.8546*	1.4883
	(0.9540)	(0.9533)	(1.1676)

4.2 Choice of Demographic Variables

We included three demographic variables that determine the variance of the half-normal distributed inefficiency u_h: household monthly income per adult equivalent, overall self-evaluated diet healthiness, and household food security measure. We chose them as they directly influence food purchase and management. In addition, they have high impact on people's attitude towards wasting food.

As discussed in the preceding section, there are other demographics that are potentially correlated with food waste, such as SNAP benefits. We did not use them as variables in estimating $\sigma^2_{u_h}$ mainly because their effects on food waste are considered indirect. On the other hand, changing the demographic variables may shift our estimates on food waste while patterns across different household groups maintain. For instance, when adding SNAP into the model, the average food waste estimates are around 27% in the three models while other waste determinants' coefficients remain similar. Another interesting observation is that, for many combinations of demographic variables, when income is dropped from the model, the waste estimates will often decrease.

These modest robustness checks lead to two main conclusions: (i) Average food loss estimates may vary by several percentage points when using different combinations of demographics, and ii) Nonetheless, the post-estimation analysis shows the patterns across household groups persist, e.g., less food secure households waste less. As there lack theoretical restrictions of choosing demographics, systematic bias may not be completely avoided. However, as most heteroskedastic stochastic frontier studies show, what is of more importance is the relative difference among observations or groups.

5 Conclusion

Our estimates on average consumer-level food waste are in line with existing estimates on average food waste. Moreover, we are able to assign each individual household a waste

estimate. Model 2 and Model 3 also point out the importance of taking physical activities into consideration whenever feasible. Household-level estimates enable us to conduct a series of interesting analyses on the relationship between food waste and household characteristics. For example, we see a clear link between food waste and levels of dietary healthiness.

The model presented in this paper can serve as a foundation for further extensions and for conducting specific hypothesis testing. For example, it would be of considerable value if one can formulate and test different waste rates for each food group, possibly through a latent class Bayesian estimation.

Results discussed in the previous sections, taken together, help illustrate our contribution in the context of previous research on food waste. While the precise measurement of food waste is important, it may be equally important to investigate how household factors influence food waste. Our indirect method allows us to accomplish this section task. Thus, we hope that our approach provides other researchers working on the topic a new lens through which estimation on individual household level food waste is feasible; and that it encourages them to extend the idea of indirect measurement to applications on other datasets and interesting cases.

References

Aigner, D., Lovell, C., and Schmidt, P. (1977). Formulation And estimation Of Stochastic Frontier Production Function Models. *Journal of Econometrics*, 6(1):21–37.

Akçay, Y., Natarajan, H. P., and Xu, S. H. (2010). Joint Dynamic Pricing of Multiple Perishable Products Under Consumer Choice. *Management Science*, 56(8):1345–1361.

Amsler, C., Prokhorov, A., and Schmidt, P. (2016). Endogeneity in stochastic frontier models. *Journal of Econometrics*, 190(2):280–288.

Archer, E., Thomas, D. M., McDonald, S. M., Pavela, G., Lavie, C. J., Hill, J. O., and Blair, S. N. (2016). The Validity of US Nutritional Surveillance: USDA's Loss-Adjusted Food

Availability Data Series 1971-2010. *Current Problems in Cardiology*, 41(11-12):268–292.

Bellemare, M. F., Çakir, M., Peterson, H. H., Novak, L., and Rudi, J. (2017). On the Measurement of Food Waste. *American Journal of Agricultural Economics*, 99(5):1148–1158.

Beretta, C., Stoessel, F., Baier, U., and Hellweg, S. (2013). Quantifying food losses and the potential for reduction in Switzerland. *Waste Management*, 33(3):764–773.

Buzby, J., Muth, M. K., Kosa, K. M., Nielsen, S. J., and Karns, S. A. (2007). Exploratory Research on Estimation of Consumer-Level Food Loss Conversion Factors Final Report. *USDA-ERS Report*, (58).

Buzby, J. C. and Guthrie, J. F. (2002). Plate Waste in School Nutrition Programs: Report to Congress. *USDA-ERS Report*.

Buzby, J. C., Wells, H. F., Axtman, B., and Mickey, J. (2009). Supermarket Loss Estimates for Fresh Fruit , Vegetables , Meat , Poultry , and Seafood and Their Use in the ERS Loss-Adjusted Food Availability Data Data. *USDA-ERS Report*, (44).

Buzby, J. C., Wells, H. F., and Hyman, J. (2014). The Estimated Amount, Value, and Calories of Postharvest Food Losses at the Retail and Consumer Levels in the United States. *USDA-ERS Report*, (121).

Chapagain, A. and James, K. (2011). The water and carbon footprint of household food and drink waste in the UK A report containing quantification and analysis of the water and carbon. *WRAP Report*, (March).

FAO/WHO/UNU (1985). *Energy and protein requirements : report of a Joint FAO/WHO/UNU Expert Consultation [held in Rome from 5 to 17 October 1981]*. World Health Organization, Geneva.

Fried, H. O., Lovell, C. A. K., and Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*. Oxford University Press, New York.

Garrone, P., Melacini, M., and Perego, A. (2014). Opening the black box of food waste reduction. *Food Policy*, 46:129–139.

Hall, K. D., Guo, J., Dore, M., and Chow, C. C. (2009). The progressive increase of food waste in America and its environmental impact. *PLoS ONE*, 4(11):9–14.

Institute of Medicine (2005). *Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein, and amino acids*. National Academies Press, Washington, D.C.

Johnston, C. S., Day, C. S., and Swan, P. D. (2002). Postprandial Thermogenesis Is Increased 100Low-Fat Diet versus a High-Carbohydrate, Low-Fat Diet in Healthy, Young Women. *Journal of the American College of Nutrition*, 21(1):55–61.

Jondrow, J., Knox Lovell, C. A., Materov, I. S., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19(2-3):233–238.

Kumbhakar, S. C. and Tsionas, E. G. (2006). Estimation of stochastic frontier production functions with input-oriented technical efficiency. *Journal of Econometrics*, 133(1):71–96.

Kurkalova, L. A. and Carriquiry, A. (2003). Input- and Output-Oriented Technical Efficiency of Ukrainian Collective Farms, 1989-1992: Bayesian Analysis of a Stochastic Production Frontier Model. *Journal of Productivity Analysis*, 20(2):191–211.

Kutlu, L. (2010). Battese-coelli estimator with endogenous regressors. *Economics Letters*, 109(2):79–81.

Leib, E. B., Gunders, D., Ferro, J., Nielsen, A., Nosek, G., and Qu, J. (2013). The Dating Game : How Confusing Food Date Labels. *Nrdc Report*, (September).

McArdle, W. D., Katch, F. I., and Katch, V. L. (1986). *Exercise physiology: energy, nutrition, and human performance*. Lea & Febiger, Philadelphia, 2nd edition.

Muth, M. K., Karns, S. A., Nielsen, S. J., Buzby, J. C., and Wells, H. F. (2011). Consumer-Level Food Loss Estimates and Their Use in the Economic Research Service (ERS) Loss-Adjusted Food Availability Data (FAD). *USDA-ERS Report*.

Neff, R. A., Spiker, M. L., and Truant, P. L. (2015). Wasted food: U.S. consumers' reported

awareness, attitudes, and behaviors. *PLoS ONE*, 10(6):1–16.

Porpino, G., Parente, J., and Wansink, B. (2015). Food waste paradox: antecedents of food disposal in low income households. *International Journal of Consumer Studies*, 39:619–629.

Qi, D. and Roe, B. E. (2016). Household food waste: Multivariate regression and principal components analyses of awareness and attitudes among u.s. consumers. *PLoS ONE*, 11(7):1–19.

Quested, T. and Parry, A. (2011). New estimates for household food and drink waste in the UK. *WRAP Report*, (November).

Reinhard, S., Lovell, K., and Thijssen, G. (1999). Econometric Estimation of Technical and Environmental Efficiency An Application to Dutch Dairy Farms. *American Journal of Agricultural Economics*, 81(1):44.

Reynolds, C. J., Mavrakis, V., Davison, S., Høj, S. B., Vlaholias, E., Sharp, A., Thompson, K., Ward, P., Coveney, J., Piantadosi, J., Boland, J., and Dawson, D. (2014). Estimating informal household food waste in developed countries: the case of Australia. *Waste management and research*, 32(12):1254–8.

Roza, A. M. and Shizgal, H. M. (1984). The Harris Benedict equation reevaluated: Resting energy requirements and the body cell mass. *American Journal of Clinical Nutrition*, 40(1):168–182.

Scrimshaw, N. S., Waterlow, J. C., and Schürch, B. (1996). Energy and Protein Requirements: Proceedings of an IDECG Workshop Held in London, UK, October 31-November 4, 1994. *European Journal of Clinical Nutrition*, 50(Supplement 1).

Secondi, L., Principato, L., and Laureti, T. (2015). Household food waste behaviour in EU-27 countries: A multilevel analysis. *Food Policy*, 56:25–40.

Stefan, V., van Herpen, E., Tudoran, A. A., and Lähteenmäki, L. (2013). Avoiding food waste by Romanian consumers: The importance of planning and shopping routines. *Food Quality and Preference*, 28(1):375–381.

Tran, K. C. and Tsionas, E. G. (2015). Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters*, 133:85–88.

Trivedi, B. (2009). Why Food Labels are Wrong. *New Sciencist*, 15.

Van Donselaar, K. H. and Broekmeulen, R. a. C. M. (2012). Approximations for the relative outdating of perishable products by combining stochastic modeling, simulation and regression modeling. *International Journal of Production Economics*, 140(2):660–669.

Venkat, K. (2011). The Climate Change and Economic Impacts of Food Waste in the United States. *International Journal on Food System Dynamics*, 2(4):431–446.

Wang, X. and Li, D. (2012). A dynamic product quality evaluation based pricing model for perishable food supply chains. *Omega*, 40(6):906–917.

Wendel-Vos, W., Droomers, M., Kremers, S., Brug, J., and Van Lenthe, F. (2007). Potential environmental determinants of physical activity in adults: A systematic review. *Obesity Reviews*, 8(5):425–440.

Wilson, N. L., Rickard, B. J., Saputo, R., and Ho, S. T. (2017). Food waste: The role of date labels, package size, and product category. *Food Quality and Preference*, 55:35–44.

World Bank (2005). Introduction to Poverty Analyses. *Poverty Manual, All, JH Revision*, (August):1–218.

Appendix.

The Revised Harris-Benedict Equation (Roza and Shizgal, 1984):

For Male: BMR=88.362 + 13.397*weight(kg) + 4.799*height(cm) - 5.677*age(year)

For Female: BMR = 447.593 + 9.247*weight(kg) + 3.098*height(cm) - 4.33*age(year)

Table A.1: Summary Statistics–Continuous Variables

	Mean	Standard Deviation	5% percentile	95% percentile
d_1, Income	1.716	1.393	0.345	4.611
$\widetilde{PA}_h$, Employment Status	0.649	0.291	0.250	1.000
z_h, Weekend Shopping Frequency	0.294	0.343	0.000	1.000
$\widehat{PA}_{m,h}$, Imputed Physical Activity	1.614	0.1134	1.412	1.792

- Income is household monthly total income divided by adult equivalent household size, in thousand dollars.
- Employment Status: for each working-age household member, 1= not working; 2= looking for work; 3= with a job but not at work; 4= working. PA_h is calculated by taking the average value of working-age household members and normalized to 0-1.
- Weekend Shopping Frequency is the percentage share of household shopping trips that occur during the weekends.

Table A.2: Summary Statistics–Diet Healthiness

Values	1	2	3	4	Total
Frequency	1093	1730	986	263	4072
Percentage	26.84%	42.49%	24.21%	6.46%	100%

Note: First row: higher values represent healthier diet.

Table A.3: Summary Statistics–Food Security

Values	Low	Medium	High	Total
Frequency	1105	798	2171	4074
Percentage	27.12%	19.59%	53.29%	100%

Note: First row: higher values represent higher food security.

Table A.4: Summary Statistics–Household Size

Size	1	2	3	4	5	6	7	Total
Frequency	904	1199	726	629	347	151	69	4074
Percentage	22.19%	29.43%	17.82%	15.44%	8.52%	3.71%	1.69%	98.38%

Note: First row: number of household members.

Table A.5: Summary Statistics–Education

Values	1	2	3	4	5	Total
Frequency	361	1243	2707	2513	1671	8495
Percentage	4.25%	14.63%	31.87%	29.58%	19.67%	100%

Note: for each individual of age 20 and above, value ranges from 1 to 5, representing different levels of highest degrees, with 1= up to 9th degree, 2= 9-11th grade, 3=high school, 4=associate degree, 5=college graduate or above. It is normalized to 0-1.

Table A.6: First-Stage Regression on NHANES

	Age 12-19	Age>=20
Weight	-0.0010	-0.0007***
	(0.0006)	(0.0002)
Height	0.0043***	0.0037***
	(0.0016)	(0.0007)
Age	-0.0058	-0.0033***
	(0.0052)	(0.0003)
Male	0.0813***	0.0596***
	(0.0254)	(0.0125)
Employment Status		0.0731***
		(0.0136)
Education		0.1837***
		(0.0187)
Constant	1.1369***	0.9706***
	(0.2223)	(0.1057)
Number of Observations	1028	4475

Table A.7: Full Estimation Results

Description	Model 1	Model 2	Model 3
Production Equation			
α_{PA}, (Employment Status)		-0.7138	
		(0.4884)	
α_1, (Milk and Dairy)	-0.0719***	-0.021***	-0.0738***
	(0.0136)	(0.0289)	(0.0143)
α_2, (Protein Foods)	-0.0250*	0.0151	-0.0235
	(0.0144)	(0.0344)	(0.0152)
α_3, (Mixed Dishes)	-0.0937***	-0.1339***	-0.1025***
	(0.0142)	(0.0419)	(0.0149)
α_4, (Grains)	-0.0706***	-0.0185	-0.0762***
	(0.0147)	(0.0333)	(0.0154)
α_5, (Snacks)	-0.0228	0.0137	-0.0225
	(0.0145)	(0.0401)	(0.0153)
α_6, (Fruit and Vegetables)	-0.0276**	-0.0957***	-0.0262*
	(0.0141)	(0.0349)	(0.0148)
α_7, (Beverages)	-0.0671***	-0.1535***	-0.0694***
	(0.0128)	(0.0401)	(0.0135)
α_8, (Condiments)	0.0114	0.0685*	0.0141
	(0.0138)	(0.0374)	(0.0145)
α_9, (Infant formula & Uncoded)	-0.0021	0.0215	-0.0062
	(0.0388)	(0.0768)	(0.0407)
α_0, (Constant)	8.1370***	7.4556***	8.5869***
	(0.0826)	(0.4495)	(0.0890)
$\beta_{1,1}$	0.0100***	0.0104***	0.0103***
	(0.0012)	(0.0023)	(0.0013)
$\beta_{2,1}$	-0.0015	-0.0017	-0.0016
	(0.0011)	(0.0024)	(0.0012)
$\beta_{2,2}$	0.0030*	-0.0045	0.0029*
	(0.0016)	(0.0040)	(0.0016)
$\beta_{3,1}$	-0.0006	-0.0023	-0.0009
	(0.0011)	(0.0026)	(0.0012)
$\beta_{3,2}$	0.0015	-0.0013	0.0017
	(0.0014)	(0.0034)	(0.0015)
$\beta_{3,3}$	0.0153***	0.0199***	0.0167***
	(0.0015)	(0.0040)	(0.0016)
$\beta_{4,1}$	0.0006	0.0006	0.0006
	(0.0010)	(0.0020)	(0.0010)
$\beta_{4,2}$	0.0012	0.0065	0.0012
	(0.0013)	(0.0031)	(0.0014)

$\beta_{4,3}$	0.0018	0.0017	0.0017
	(0.0014)	(0.0035)	(0.0014)
$\beta_{4,4}$	0.0065***	0.0049	0.0069***
	(0.0016)	(0.0031)	(0.0016)
$\beta_{5,1}$	0.0017*	0.0100**	0.0019*
	(0.0010)	(0.0040)	(0.0011)
$\beta_{5,2}$	-0.0016	-0.0017	-0.0018
	(0.0014)	(0.0033)	(0.0015)
$\beta_{5,3}$	0.0002	0.0030	0.0002
	(0.0014)	(0.0027)	(0.0014)
$\beta_{5,4}$	0.0004	-0.0023	0.0005
	(0.0016)	(0.0028)	(0.0013)
$\beta_{5,5}$	0.0008	0.0042	0.0006
	(0.0015)	(0.0030)	(0.0016)
$\beta_{6,1}$	0.00004	-0.0005	0.0001
	(0.0012)	(0.0034)	(0.0013)
$\beta_{6,2}$	0.0019	0.0049	0.0020
	(0.0014)	(0.0036)	(0.0015)
$\beta_{6,3}$	0.0013	0.0056	0.0017
	(0.0014)	(0.0043)	(0.0015)
$\beta_{6,4}$	-0.0015	-0.0085*	-0.0015
	(0.0015)	(0.0050)	(0.0015)
$\beta_{6,5}$	-0.0005	-0.0067	-0.0005
	(0.0015)	(0.0042)	(0.0016)
$\beta_{6,6}$	0.0031**	0.0084***	0.00236*
	(0.0015)	(0.0030)	(0.00185)
$\beta_{7,1}$	0.0005	-0.0042	0.0006
	(0.0011)	(0.0030)	(0.0012)
$\beta_{7,2}$	0.0010	0.0043	0.0010
	(0.0014)	(0.0036)	(0.0015)
$\beta_{7,3}$	-0.0014	0.0010	-0.0012
	(0.0013)	(0.0031)	(0.0014)
$\beta_{7,4}$	0.0013	0.0002	0.0015
	(0.0014)	(0.0034)	(0.0015)
$\beta_{7,5}$	0.0013	-0.0062**	0.0014
	(0.0014)	(0.0032)	(0.0014)
$\beta_{7,6}$	0.0001	0.0079**	-0.0001
	(0.0015)	(0.0037)	(0.0015)
$\beta_{7,7}$	0.0070***	0.0145***	0.0074***
	(0.0012)	(0.0042)	(0.0013)
$\beta_{8,1}$	-0.0008	-0.0009	0.0008
	(0.0009)	(0.0019)	(0.0009)

$\beta_{8,2}$	-0.0017 (0.0013)	-0.0051 (0.0032)	-0.0020 (0.0013)
$\beta_{8,3}$	-0.0015 (0.0013)	-0.0031 (0.0026)	-0.0015** (0.0014)
$\beta_{8,4}$	-0.0004 (0.0011)	0.0009 (0.026)	-0.0003 (0.0012)
$\beta_{8,5}$	0.0009 (0.0011)	0.0004 (0.0024)	0.0009 (0.0012)
$\beta_{8,6}$	0.0007 (0.0013)	-0.0043 (0.0030)	0.0006 (0.0014)
$\beta_{8,7}$	0.0005 (0.0014)	-0.0002 (0.0030)	0.0006 (0.0014)
$\beta_{8,8}$	-0.0007 (0.0013)	0.0014 (0.0026)	-0.0012 (0.0014)
$\beta_{9,1}$	-0.0001 (0.0016)	0.0008 (0.0026)	-0.0001 (0.00187)
$\beta_{9,2}$	-0.0047* (0.0029)	-0.0040 (0.0054)	-0.0048 (0.0031)
$\beta_{9,3}$	-0.0038* (0.0023)	-0.0085** (0.0041)	-0.0039 (0.00274)
$\beta_{9,4}$	-0.0017 (0.0023)	-0.0056 (0.0040)	-0.0021 (0.0024)
$\beta_{9,5}$	-0.0021 (0.0023)	0.0004 (0.0041)	-0.0026 (0.0025)
$\beta_{9,6}$	0.0019 (0.0029)	0.0028 (0.0060)	0.0027 (0.0030)
$\beta_{9,7}$	0.0079*** (0.0030)	0.0068* (0.0065)	0.0081*** (0.0032)
$\beta_{9,8}$	0.0028 (0.0019)	0.0014 (0.0036)	0.0033* (0.0020)
$\beta_{9,9}$	-0.0021 (0.0039)	0.0013 (0.0069)	-0.0023 (0.0041)
White Noise σ_v^2	0.4845*** (0.0080)	0.3788*** (0.2117)	0.3262*** (0.0014)
Inefficiency $\log \sigma_{u_h}^2 = \gamma_0 + \gamma' d_h$			
Income	0.3458*** (0.0601)	0.4089*** (0.0657)	0.3010*** (0.0635)
Healthy Diet	1.5147*** (0.5871)	0.9888* (0.5411)	1.7914** (0.7222)
Food Security	1.9550* (1.1395)	2.3139** (1.0984)	2.1217 (1.4286)

Constant	-6.5729*** (1.7434)	-8.1155*** (1.8586)	-7.7571*** (2.0310)

Note: Model 1: Baseline model. Model 2: Proxy-instrumental variable LIML estimation. Model 3: Imputed physical activity levels.