Goodness-of-fit in economic models
¿How much are we losing?

Felipe Del Canto M.

September 17, 2019

> (...) all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.
>
> *Empirical Model-Building and Response Surfaces (1987)*
> George Box and Norman Draper

Abstract

Aggregation is a tool used to reduce the complexity of economic models in order to draw more clear and succinct conclusions or simplify analyses. As any approximation, its use may be accompanied with errors researches may not be willing to tolerate if they were aware of them. In this work I present how these errors appear in simple models using aggregation across goods and across consumers. I also show some of their determinants in order to find ways to bound them. Finally, I briefly discuss a methodology to study the goodness-of-fit of aggregate models in more general settings.

1. Introduction

Every model in science is by definition a simplified reality. On the bright side, abstracting from the complexity of the real world has allowed society to understand the sometimes subtle mechanisms that rule nature and human behavior. This does not mean that a model is useful for every purpose. Evidently, whilst some of them can illustrate certain dynamics of the real world very clearly, the approximation may carry errors that harm future predictions. The previous comment points directly to the question of which model is the most useful for some given problem. In particular, when the answer is *many* the modeler needs to make a choice based on the results she expects to highlight and the channels to study. The dilemma is by no means alien to the field of economics. When describing an economy, the researcher is faced with several possible assumptions that shape the complexity of the model. Although some of them may be made by feasibility reasons (for example, because a highly detailed model cannot be solved or simulated or because data will not be available to calibrate it) there may be others that serve a transparency purpose, that is, they intend to make clear the results without dwelling on the unnecessary details. Consequently, in the process of constructing a model, the investigator may choose to follow the Occam's Razor principle: among the models that are consistent with the evidence, choose the one that makes the fewest possible assumptions. This criterion implies that the measure of a (correct) model is its complexity. However, as Milton Friedman said, "The ultime goal of a positive science is the development of a 'theory' or, 'hypothesis' that yields valid and meaningful (...) predictions about phenomena not yet observed" and thus "Its performance is to be judged by the precision, scope, and conformity with experience of the predictions it yields".[1] Hence, in evaluating the validity of a theory, the robustness of its conclusions should be of critical importance.

Different strands in the economic literature have studied when predictions of some models are robust to different specifications. In Sutton (2007), the author discusses which mechanisms in the context of industrial organization still hold in conditions outside the classical models of, for example, Cournot and Bertrand. A similar motivation can be found in Kajii and Morris (1997), where they study how sensitive game theory conclusions are to the assumption of common knowledge of payoffs in a game. The interest in robustness in the context of mechanism design can be also found in ter Vehn and Morris (2011). An interesting approach is the one in Basu and Fernald (1997) where the authors try to estimate discrepancies due to "aggregation effects" when considering a model with a representative firm and one where

[1] Friedman (1953).

heterogeneous effects are considered. Similarly, in Hanushek et al. (1996) the authors try to reconcile contradictory results in the school literature arguing an important role of aggregation in the magnitude of omitted variable bias, which can in principle invalidate previous estimations. Related to these aggregation literature there is a concern with models that make use of aggregated data and different authors have studied what is called "aggregation bias" arguing that these models hide important mechanisms that could explain these differences.[2] All in all, different authors have tried to untangle the differences between predictions and realizations by appealing to the goodness-of-fit of the models used to produce such estimations. In the cases where the deviations are substantial, a revision to the model must be made.

As an example, consider the case of the representative agent model. The assumption that there is only one consumer in the economy is useful and has been key to understand important qualitative results, especially in macroeconomics. Nevertheless, employing this model to predict future realizations of certain key variables such as aggregate demand or marginal propensity to consume (MPC) may be inaccurate if heterogeneity effects are in place. In other words, there is a shadow price in the approximation (which the investigator could be willing to pay or not) if she wishes to use the model for another, more quantitative-driven purpose. This point is made clearly in Carroll (2000) in the context of the buffer-stock model: "Representative-agent models are typically calibrated to match an aggregate wealth-to-income ratio" but "the typical household's wealth is much smaller than the wealth of such a representative agent (...), this would lead one to expect that the behavior of the median household may not resemble the behavior of a representative agent with a wealth-to-income ratio similar to the aggregate ratio". The evidence quickly backs up this view: while the annual MPC predicted by the representative agent model is about 0.04, many empirical analysis estimate this parameter to lie between 0.2 and 0.5.[3]

The aforementioned model is a particular case of a common practice in economics: aggregation. The other canonical example of its use is aggregation across goods, where instead of describing the myriad of goods available in an economy they are grouped into one or several categories. Regarding these two implementations, previous theoretical literature focused in one side of the problem: When is it possible to carry out this practice and describe precisely the same economy?. In the case of the representative agent, the necessary and sufficient condition is that the indirect utility function of every consumer has the Gorman form.[4] When aggregation is applied to goods, the answer has been more elusive but two results arise. First,

[2] See for example Lee et al. (1990); Ravallion (1998); Feenstra and Hanson (2000); Imbs et al. (2005).
[3] Carroll (2000).
[4] Gorman (1953).

the Hicks-Leontief (composite commodity) theorem allows aggregation if relative prices are constant in the group of goods that are to be bundled.[5] Although a somewhat weaker requirement is proposed by Lewbel (1996): bundling is possible if all group relative prices are independent of price indexes and income. The second answer states that grouping some goods is possible if preferences between them are "independent" of the remaining goods present in the economy.[6]

For the two kinds of aggregation mentioned before, the conditions for them to hold are highly restrictive and not typically met in econometric or theoretical applications. As mentioned previously, the literature has assumed (disregarding these conditions) exact aggregation of both types in constructing models and making econometric estimations and this practice comes at a cost. As was stated in the preceding discussion, the validity of these models is directly attached to the size of such cost. If the question an investigator is seeking to answer allows the use of aggregate models without incurring in a severe deviation from predictions, then not having exact aggregation is of minor importance. In other words, compliance of the conditions for aggregation is not a problem as long as the parameters of the simplified model are calibrated in such a way that the estimation error is below some tolerance predefined by the modeler. Hence, understanding and quantifying possible these differences is crucial in determining and measuring the goodness-of-fit of these models. Consequently, and in contrast with some of the articles mentioned earlier, in this work I intend to give a theoretical look at how these deviations appear using simple models and understand how a researcher could limit them. The main insight employed in this work is that aggregation is trading off heterogeneity for simplicity and here is from where errors make their appearance. Next, drawing on the previous results, I will discuss a methodology about approaching this problem in more general settings.

The rest of the paper proceeds as follows. To motivate the following discussion, in Section 2 I present a summary of the problems and previous results about conditions under which aggregation is possible. Then, I present 3 settings to illustrate how approximation errors appear when using aggregate models to estimate future economic variables. First, in Section 3, I present a representative agent model in the context of aggregate demand estimation. Second, In Section 5 I study aggregation across goods when used in the context of demand forecasting. The third model is presented in Section 4 and mixes the previous two by presenting an economy of several individuals consuming various goods, where the good aggregation takes a different

[5] Leontief (1936); Hicks (1946).
[6] See for example Gorman (1959).

form. More sections?. Finally, in Section 6 I discuss some final thoughts about model fitness.

2. Previous results in aggregation

The economic literature has recognized two forms of aggregation that are usually taught in microeconomics courses around the world. First, the problem of consumer aggregation or the representative agent problem seeks to describe the aggregate demand of a multi-person economy by focusing only on the aggregate determinants of demand (i.e., the aggregate income), as opposed to the distribution of such variables. The second class of aggregation focuses on describing demands for categories of goods without distinguishing the individual consumption on each element in the category. I will refer to this last problem as the "Hicksian aggregation problem".

Both kinds of aggregation are widely used in economic models. The representative model agent was a salient feature of macroeconomic models in the decade of which one? I need literature here.[7] On the other hand, the two good setting (one particular example of aggregation of goods) has been widely used in buffer-stock models where their main conclusions arise from the interaction between consumption and savings.[examples] Empirical studies also make use of both forms of aggregation. Some of them assume all consumers are equal which is an example of consumer aggregation and others overcome problems in the data (e.g., availability or comprehensiveness) by assuming that agents choose consumption on the category and not on individual goods.[examples].

In what follows I will formally describe the problems about aggregation the early literature tried to answer. These questions aimed at finding conditions for which aggregation is possible. In order to describe both problems I will rely heavily on Varian (1992).

2.1. The representative agent problem

Consider an economy composed by n consumers indexed by $i = 1, \ldots, n$. Their demand functions of the k goods in the economy are summarized in the vector $\mathbf{x}_i(\mathbf{p}, y_i)$, where $\mathbf{p}$ is the price vector of the goods and y_i is the income of agent i. The aggregate demand vector is defined by

$$\mathbf{X}(\mathbf{p}, y_1, \ldots, y_n) = \sum_{i=1}^{n} \mathbf{x}_i(\mathbf{p}, y_i). \tag{1}$$

[7] Lucas 78', Kydland & Prescott 82' ciclos reales

The question that automatically arises is: Can this aggregate demand function be regarded as generated by a single (or "representative") consumer?. In terms of equation (1), the previous question is equivalent to looking for conditions under which $\mathbf{X}$ does not depend on the distribution but on the aggregate income

$$Y := \sum_{i=1}^{n} y_i.$$

The definitive answer to this problem came with Gorman (1953). According to his result, $\mathbf{X}$ is a function of Y if and only if for every $i \in \{1, \ldots, n\}$ the indirect utility function has the Gorman form

$$v_i(\mathbf{p}, y_i) = a_i(\mathbf{p}) + b(\mathbf{p})y_i,$$

where a_i, b are functions that must only depend on $\mathbf{p}$ and b has to be the same across consumers. This functional requirement is somewhat restrictive but at least two particular examples are worth mentioning: homothetic and quasilinear utility functions. For the first, the indirect utility functions is

$$v(\mathbf{p}, y) = b(\mathbf{p})y, \tag{2}$$

while for the second

$$v(\mathbf{p}, y) = a(\mathbf{p}) + y.$$

Both examples clearly have the Gorman form. However, some homothetic functions could differ in the function v between consumers and thus aggregation is not possible. For example, when $k = 2$, an economy of consumers with Cobb-Douglas utility functions

$$u(x_1, x_2) = x_1^{\alpha_i} x_2^{1-\alpha_i},$$

but where at least two α_i are different. In that case the function $v(\mathbf{p})$ in (2) is different between individuals and thus aggregation is not possible.

2.2. Hicksian aggregation problem

For this problem consider the following setting. Assume the consumption vector of some agent is divided in two bundles $(\mathbf{x}, \mathbf{z})$. Accordingly, the price vector is separated into $(\mathbf{p}, \mathbf{q})$.

Thus, if the utility function of the consumer is u, then the demand for the $\mathbf{x}$ goods is

$$\begin{aligned} \mathbf{x}(\mathbf{p}, \mathbf{q}, y) = \arg\max_{\mathbf{x},\mathbf{z}} \quad & u(\mathbf{x}, \mathbf{z}) \\ \text{s.a.} \quad & \mathbf{p}\mathbf{x} + \mathbf{q}\mathbf{z} = y. \end{aligned} \tag{3}$$

In numerous models, there is no interest in the consumption of each of the $\mathbf{x}$-goods but only in the demand for the group (i.e., focus on expenditure in savings against expenditure in different financial instruments). Hence, the Hicksian aggregation problem is finding conditions under which this approximation can be made without losing information. This implies finding a quantity index $X = g(\mathbf{x})$, a price index $P = f(\mathbf{p})$ and a new utility function $U(X, \mathbf{z})$ such that the solution

$$\begin{aligned} X(P, \mathbf{q}, y) = \arg\max_{X,\mathbf{z}} \quad & U(X, \mathbf{z}) \\ \text{s.a.} \quad & PX + \mathbf{q}\mathbf{z} = y. \end{aligned} \tag{4}$$

satisfies

$$X(f(\mathbf{p}), \mathbf{q}, y) = g(\mathbf{x}(\mathbf{p}, \mathbf{q}, y)).$$

In other words, the problem seeks to find an alternative model that is coherent with the original. Elaborate on this.

At least two situations exist under which these alternative models can be found: functional and Hicksian separability. In the first, assume that the preference relation represented by u has the following "independence" property

$$(\mathbf{x}, \mathbf{z}) \succ (\mathbf{x}', \mathbf{z}) \iff (\mathbf{x}, \mathbf{z}') \succ (\mathbf{x}', \mathbf{z}') \qquad \forall\ \mathbf{x}, \mathbf{x}', \mathbf{z}, \mathbf{z}'. \tag{5}$$

This independence property implies that there exists a function v such that

$$u(\mathbf{x}, \mathbf{z}) = U(v(\mathbf{x}), \mathbf{z}),$$

where $U(v, \mathbf{z})$ is increasing in v. In this case, the consumer values consumption of $\mathbf{x}$ only through v. For example, consider $\mathbf{x}$ to be different types of food. Independence in this context could manifest If the agent only values the amount of calories she gets from $\mathbf{x}$. In that situation, all kinds of food will be valued differently according to the amount of calories per unit each of them give but the consumer will only value the total caloric intake. The previous example suggest a particularity of this type of separability. Following the same nomenclature: The consumer will choose the amount of calories and the total amount to

spend on food first and then she will choose the individual consumption on each kind of food. Models that show this feature are often referred to as hierarchical consumption models.

By calling $m_{\mathbf{x}} := \mathbf{p} \cdot \mathbf{x}(\mathbf{p}, \mathbf{q}, y)$, then it can be shown that the following equality holds

$$\begin{aligned} \mathbf{x}(\mathbf{p}, \mathbf{q}, y) = \arg\max_{\mathbf{x}} \quad & v(\mathbf{x}) \\ \text{s.a.} \quad & \mathbf{p}\mathbf{x} = m_{\mathbf{x}}. \end{aligned}$$

Thus, if $e(\mathbf{p}, v)$ is the expenditure function of the previous problem, then

$$\begin{aligned} v(\mathbf{x}(\mathbf{p}, \mathbf{q}, y)) = \arg\max_{v, \mathbf{z}} \quad & U(v, \mathbf{z}) \\ \text{s.a.} \quad & e(\mathbf{p}, v) + \mathbf{q}\mathbf{z} = y. \end{aligned}$$

However, note that the latter problem does not have the exact same structure as (4). This only happens if v has a particular structure that

$$e(\mathbf{p}, v) = e(\mathbf{p})v.$$

This property holds if, for example, v is homothetic and thus the Cobb-Douglas utility function is an example in which this kind of aggregation is possible. The previous discussion shows that under functional separability this consumption model is hierarchical.

On the other hand, Hicksian separability is present in the following situation. Assume that $\mathbf{p} = t\mathbf{p}_0$, where t is a scalar and $\mathbf{p}_0$ is a fixed price vector. By defining $P := t$ and $X := \mathbf{p}_0\mathbf{x}$, we can define the following indirect utility function

$$\begin{aligned} V(P, \mathbf{q}, y) = \arg\max_{\mathbf{x}, \mathbf{z}} \quad & u(\mathbf{x}, \mathbf{z}) \\ \text{s.a.} \quad & P\mathbf{p}_0\mathbf{x} + \mathbf{q}\mathbf{z} = y. \end{aligned}$$

By using V we can find another function U by solving

$$\begin{aligned} U(X, \mathbf{z}) = \arg\min_{P, \mathbf{q}} \quad & V(P, \mathbf{q}, y) \\ \text{s.a.} \quad & PX + \mathbf{q}\mathbf{z} = y, \end{aligned}$$

that by definition satisfies

$$V(P, \mathbf{q}, y) = \arg\max_{X,\mathbf{z}} \quad U(X, \mathbf{z})$$
$$\text{s.a.} \quad PX + \mathbf{qz} = y.$$

and thus aggregation is possible with $g(\mathbf{x}) = \mathbf{p}_0\mathbf{x}$ and $f(\mathbf{p}) = t$. This result was both presented by Leontief (1936) and Hicks (1946) and thus has been called the Hicks-Leontief theorem. As was the case with functional separability, note that the theorem presents sufficient but not necessary conditions to find the functions g, f and U. In his work, Lewbel (1996) relaxes the assumptions of the Hicks-Leontief theorem by not asking for constant relative prices but instead that these are independent of the price index and income. In terms of data analysis, the generalized composite commodity theorem of Lewbel does not ask a correlation of one between the price of the $\mathbf{x}$-goods and their price index.

One of the most common applications of good aggregation under Hicksian separability assumptions are two-good models. Here, the interest is in the demand for one good while bundling the other goods in only one category. These kind of models are frequently used in the macroeconomic literature where the study of aggregate consumption dynamics is greatly simplified by assuming agents just consume and save (and hence there are only two goods available). Observe that for this to happen and in accordance with the conditions in Lewbel (1996) it must be true that relative prices of all goods in the economy have to be independent of income and the price index. I want to explain why this could be a strong assumption but I can't find a convincing argument yet (I was thinking about connecting the assumption to substitutability issues of price indices but I'm not sure if that works.)

3. Aggregation across consumers

3.1. Description of the economy

Consider a two period economy ($t = 0, 1$) composed by a continuum of agents that consume two goods $\mathbf{x}_t = (x_{1t}, x_{2t})$ valued at prices $\mathbf{p}_t = (p_{1t}, p_{2t}) \in \mathbb{R}^2_{++}$. For simplicity, all subindices t will be omitted unless needed. Each individual in this setting is identified with a pair (y, α), where $y \in (0, \infty)$ is her income and α determines the form of her Cobb-Douglas utility function, that is,

$$u_\alpha(\mathbf{x}) := u_{(y,\alpha)}(\mathbf{x}) = x_1^\alpha x_2^{1-\alpha}.$$

The random variables y and α follow a joint distribution, F, with support $S := (0,\infty)\times(0,1)$. Observe that in principle y and α could be correlated. The marginal distributions are F_y and F_α, respectively. As is usual in the literature, y_t will follow a *log*-Normal distribution with parameters μ_t and σ_t.

For every fixed vector $\mathbf{p}$, the agents choose the consumption bundle $\mathbf{x}(\mathbf{p}, y, \alpha)$ by maximizing u over $\mathbf{x}$ subject to their respective budget restriction. Specifically, the agent identified with the pair (y,α) solves

$$\begin{aligned} \max_{\mathbf{x}} \quad & u_\alpha(\mathbf{x}) \\ \text{s.a.} \quad & \mathbf{p}\cdot\mathbf{x} = y. \end{aligned}$$

Thus,

$$x_j(\mathbf{p}, y, \alpha) = \alpha\frac{y}{p_j}, \qquad j = 1, 2,$$

and thus the aggregate demand for good j is

$$X_j(\mathbf{p}, F_y, F_\alpha) = \frac{1}{p_j}\int_S y\alpha\, dF(\alpha, y) = \frac{1}{p_j}\mathbb{E}[y\alpha] = \frac{1}{p_j}\Big(\mathbb{E}[y]\mathbb{E}[\alpha] + \mathrm{Cov}(y,\alpha)\Big), \quad j = 1, 2.$$

3.2. The researcher's problem

Suppose now an investigator wishes to describe the aggregate demand of good 1 at time 1. She has access to a full description of the distribution income of the agents in the economy in $t = 0$ (that is, she knows F_y) but preferences at any time are not available to her. She also knows X_{10}, the aggregate demand of good 1 at time 0.[8] In this setting, if X_1 is the aggregate demand for cellphones then our researcher wishes to estimate consumption of mobiles at $t = 1$ by using data available at $t = 0$ on income and aggregate consumption of these devices.

Since individual preferences are unknown to her, she therefore assumes all individuals are equal, that is, there exists $\overline{\alpha}$ such that every agent in this economy has utility function $u_{\overline{\alpha}}$. This assumption is equivalent as saying there only exists one person in the economy given that aggregation is possible when the parameters of the Cobb-Douglas utility function are the same for all individuals (see Section 2.1). In more subtle interpretation of the mentioned assumption, what the researcher is doing is approximate F_α by $\delta_{\overline{\alpha}}$ where δ_x is the Dirac distribution centered at x. Analogously, she is approximating F with a joint distribution for (y,α) whose marginals are F_y and $\delta_{\overline{\alpha}}$.

[8] Observe that by assuming complete knowledge on the distribution of income and on the aggregate demand I am abstracting the discussion from the appearance of possible estimation errors due to sampling.

Under the previous assumption, the representative agent has utility function

$$U_{\overline{\alpha}}(X_1, X_2) = X_1^{\overline{\alpha}} X_2^{1-\overline{\alpha}},$$

and thus, since $E[y_t]$ is the aggregate income of this economy at time t, the best estimation for X_{1t} is

$$\widehat{X}_{1t}(p_{1t}, F_{y_t}, \overline{\alpha}) = \overline{\alpha}\, \frac{E[y_t]}{p_{1t}}.$$

3.3. The estimation error

As mentioned in the introduction, the aggregation assumption imposes a shadow price on the estimation. The error in this case will arise from the differences between the individual demands for x_1 and the ones estimated by the simplified model. Indeed, observe that

$$\begin{aligned}\left|X_1(p_1, F_y, F_\alpha) - \widehat{X}_1(p_1, F_y, \overline{\alpha})\right| &= \left| \int_S \frac{y\alpha}{p_1} - \frac{y\overline{\alpha}}{p_1}\, dF(y, \alpha) \right| \\ &= \left| \int_S x_1(p_1, y, \alpha) - x_1(p_1, y, \overline{\alpha})\, dF(y, \alpha) \right|,\end{aligned}$$

and thus the accuracy in the estimation, although observable in macroeconomic variables, has its roots in microeconomic differences. Moreover, conditional on having the correct distribution of income, this discrepancy will depend only on the choice of $\overline{\alpha}$.

Note that the monetized error in the estimation at time t, as a function of $\overline{\alpha}$ is

$$D_t(\overline{\alpha}) := p_{1t}\left|X_{1t}(p_{1t}, F_{y_t}, F_{\alpha_t}) - \widehat{X}_{1t}(p_{1t}, F_{y_t}, \overline{\alpha})\right| = \mathbb{E}[y_t] \left| \mathbb{E}[\alpha_t] + \frac{\mathrm{Cov}(y_t, \alpha_t)}{\mathbb{E}[y_t]} - \overline{\alpha} \right|, \quad (6)$$

and observe that this difference does not depend on prices. Consequently, the error could be reduced to zero if

$$\overline{\alpha} = \underbrace{\mathbb{E}[\alpha_t] + \frac{\mathrm{Cov}(y_t, \alpha_t)}{\mathbb{E}[y_t]}}_{:= M_t}, \quad (7)$$

in the case where the expression on the right hand side is between 0 and 1. Recalling the interpretation of the single consumer assumption as a way to approximate F, then (7) says that the only features of the distribution that are important to have an exact fit at time t are its first and second moments. This finding suggests that what matters when estimating aggregate variables using representative consumer models are a finite set of statistics of the underlying distribution of the population.

Since the researcher has access to X_{10}, then $\overline{\alpha}$ could be chosen optimally to reduce the

estimation error at $t = 0$. Assume that $M_0 \in (0, 1)$. Then, the optimal choice of $\overline{\alpha}$ given this information is M_0 and thus the estimation error in $t = 1$ is

$$E[y_t] \Big| M_1 - M_0 \Big|.$$

If preferences don't change, then

$$M_1 - M_0 = \frac{\text{Cov}(y_1, \alpha)}{\mathbb{E}[y_1]} - \frac{\text{Cov}(y_0, \alpha)}{\mathbb{E}[y_0]}.$$

Moreover, if $y_1 = Ry_0$ for some $R > 0$, then $\overline{\alpha} = M_0$ completely vanishes the estimation error at $t = 1$.

Dynamics? Simulation?

4. Consumer aggregation and category goods

4.1. Description

Consider a two period economy composed by a continuum of agents that consume $n + 1$ goods: an essential good z (i.e., water) and n different goods grouped in the vector $\mathbf{x}$. Good z is valued at price $q \in \mathbb{R}_{++}$ and the $\mathbf{x}$ goods are valued at prices $\mathbf{p} \in \mathbb{R}^n_{++}$. In what follows and for simplicity of notation, the subindices t for all variables will be omitted unless needed. Each individual in this setting is identified with a pair (y, α), where $y \in (0, \infty)$ is her income and $\alpha \in (0, 1)$ determines the form of her pseudo-Cobb-Douglas (PCD)[9] utility function,

$$u_\alpha(\mathbf{x}, z) := u_{(y,\alpha)}(\mathbf{x}, z) = \sum_{j=1}^{n} x_j^\alpha z^{1-\alpha} = \left(\sum_{j=1}^{n} x_j^\alpha \right) z^{1-\alpha}.$$

Consumers choose how much of the $\mathbf{x}$ goods to consume in each period according to

$$\begin{aligned} \mathbf{x}(\mathbf{p}, q, y, \alpha) := \arg\max_{\mathbf{x}, z} \quad & u_\alpha(\mathbf{x}, z), \\ \text{s.a.} \quad & \mathbf{p}\mathbf{x} + qz = y. \end{aligned} \tag{8}$$

The pairs (y, α) follow a distribution F (with marginal distributions F_y and D_α) over its support $S := (0, \infty) \times (0, 1)$. As was presented in Section 3, F_y will be a *log*-Normal distribution with parameters μ and σ. In this setting, it is possible to aggregate consumption

[9] The choice of this name will be clear shortly.

of the $\mathbf{x}$ goods into a single good X (see Section 2.2) given by:

$$g_\alpha(\mathbf{x}) = \left(\sum_{j=1}^{n} x_j^\alpha \right)^{1/\alpha}. \tag{9}$$

In that case, we have

$$U(g_\alpha(\mathbf{x}), z) := g_\alpha(\mathbf{x})^\alpha z^{1-\alpha}, \tag{10}$$

which is the usual Cobb-Douglas utility function. Hence, by defining $\varepsilon := (1-\alpha)^{-1}$ and

$$P_\alpha(\mathbf{p}) := \left(\sum_{j=1}^{n} p_j^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}}, \tag{11}$$

we have that the category demand

$$\begin{aligned} X(\mathbf{p}, q, y, \alpha) := \arg\max_{X,z} \quad & U(X, z) \\ \text{s.a.} \quad & P_\alpha(\mathbf{p})X + qz = y, \end{aligned} \tag{12}$$

satisfies

$$X(\mathbf{p}, q, y, \alpha) = g_\alpha\big(\mathbf{x}(\mathbf{p}, q, y, \alpha)\big). \tag{13}$$

This means that the disaggregate model is coherent with the aggregate one for each individual consumer. Indeed, observe that the price and quantity indices are specific for each agent as they depend on α. This means that any attempt to describe this economy using category demands instead of the disaggregate consumption needs knowledge about the heterogeneity of the population in order to avoid approximation errors.

4.2. The prediction problem

Consider now the following situation. An investigator at time $t = 0$ has data available on disaggregate consumption of the $\mathbf{x}$ and z goods, income y (that is, she knows F_y) and prices $(\mathbf{p}, q)$. Her objective is to estimate the aggregate category demand at $t = 1$ (i.e., the country demand for meat next year), $\mathbf{X}_1$. For simplicity she assumes that all agents have the same preferences, which means that all differences in consumption are accounted by differences in income. The last assumption also implies that she must choose a single parameter $\overline{\alpha}$ in order to obtain the category demand and the price index of each agent. In the distribution approximation sense proposed in Section 3.2, observe that the same-preference assumption has the same effect mentioned there: F_α is approximated by $\delta_{\overline{\alpha}}$. Along the same line, bundling

goods into a category also represents an approximation but of a different nature. Note that F and the function $\mathbf{x}(\mathbf{p}, q, y, \alpha)$ induce a distribution G over the vector $\mathbf{x}$. Then, g_α induces a distribution over X that is in some sense an approximation of G, due to (13). Elaborate more? I'm not sure I'm making a point

Under the previous assumption, the best estimation for the category demand of each agent at time t is

$$\widehat{X}_t(\mathbf{p}_t, y_t, \overline{\alpha}) = \overline{\alpha}\, \frac{y_t}{P_{\overline{\alpha}}(\mathbf{p}_t)}.$$

Thus, the best estimation for the aggregate category demand is

$$\widehat{\mathbf{X}}_t(\mathbf{p}_t, F_{y_t}, \overline{\alpha}) = \int_0^\infty \overline{\alpha}\, \frac{y_t}{P_{\overline{\alpha}}(\mathbf{p}_t)}\, dF_{y_t} = \overline{\alpha}\, \frac{\mathbb{E}[y_t]}{P_{\overline{\alpha}}(\mathbf{p}_t)} \tag{14}$$

4.3. A complex estimation error

To understand how the estimation errors appear in this setting I proceed in two steps. First, since category demands depend on α, then for agent (y, α) the true value will differ from the estimated one by

$$\mathcal{D}_t(y_t, \alpha_t, \overline{\alpha}) := X(\mathbf{p}_t, y_t, \alpha_t) - \widehat{X}_t(\mathbf{p}_t, y_t, \overline{\alpha}) = y_t \left(\frac{\alpha_t}{P_{\alpha_t}(\mathbf{p}_t)} - \frac{\overline{\alpha}}{P_{\overline{\alpha_t}}(\mathbf{p}_t)} \right).^{10}$$

Discuss the monotonicity (or absence of it).

The second step is to note that the estimation error (with sign) of $\mathbf{X}_{1t}$ is just the integral of these errors over S. In the simplest case where preferences do not change over time we have,

$$\begin{aligned}
D_t(\overline{\alpha}) :&= \int_S \mathcal{D}(y_t, \alpha, \overline{\alpha}) \\
&= \int_S \alpha\, \frac{y_t}{P_\alpha(\mathbf{p}_t)}\, dF(y_t, \alpha) - \overline{\alpha}\, \frac{\mathbb{E}[y_t]}{P_{\overline{\alpha}}(\mathbf{p}_t)} \\
&= \mathbb{E}\left[\alpha\, \frac{y_t}{P_\alpha(\mathbf{p}_t)} \right] - \overline{\alpha}\, \frac{\mathbb{E}[y_t]}{P_{\overline{\alpha}}(\mathbf{p}_t)} \\
&= \mathrm{Cov}\left(y_t, \frac{\alpha}{P_\alpha(\mathbf{p}_t)} \right) + \mathbb{E}[y_t]\mathbb{E}\left[\frac{\alpha}{P_\alpha(\mathbf{p}_t)} \right] - \overline{\alpha_t}\, \frac{\mathbb{E}[y_t]}{P_{\overline{\alpha}}(\mathbf{p}_t)} \\
&= \mathbb{E}[y_t] \left\{ \frac{\mathrm{Cov}\left(y_t, \frac{\alpha}{P_\alpha(\mathbf{p}_t)} \right)}{\mathbb{E}[y_t]} + \mathbb{E}\left[\frac{\alpha}{P_\alpha(\mathbf{p}_t)} \right] - \frac{\overline{\alpha}}{P_{\overline{\alpha}}(\mathbf{p}_t)} \right\}.
\end{aligned}$$

Let T_t be the random variable $\dfrac{\alpha}{P_\alpha(\mathbf{p}_t)}$ and let $\overline{T}_t := \dfrac{\overline{\alpha}}{P_{\overline{\alpha}}(\mathbf{p}_t)}$. Then we can (with some abuse

[10] Note that in this setting X does not depend on q and thus, to simplify notation, this variable was suppressed.

of notation on the function D_t) write

$$D_t\left(\overline{T}_t\right) = \mathbb{E}[y_t]\left\{\mathbb{E}\left[T_t\right] + \frac{\mathrm{Cov}\left(y_t, T_t\right)}{\mathbb{E}[y_t]} - \overline{T}_t\right\}, \tag{15}$$

which is in principle very similar to (6) in Section 3.3. However, note now that D_t does depend on prices and hence on their dynamics. In terms of distributions observe that in this case, the mixed setting implies that the distribution that matters is the one of (y_t, T_t) that depends also on prices. From (15) we know that the sufficient statistics needed to obtain an exact fit at time t are, just as in Section 3.3, the first and second moments of that distribution. The difference in this case is that now there is a third source of variation over time, which are prices.

From (15) an investigator interested in making the estimation error the lowest possible should choose $\overline{\alpha}$ such that

$$\overline{T}_t = \underbrace{\mathbb{E}\left[T_t\right] + \frac{\mathrm{Cov}\left(y_t, T_t\right)}{\mathbb{E}[y_t]}}_{:=\tilde{M}_t}. \tag{16}$$

Observe that changes in prices may affect the right hand side of (16). If the investigator only has access to $t = 0$ variables, knowing the dynamics of prices may not be enough to reduce the estimation error at $t = 1$ to zero. Indeed, note that the value of each term in $\tilde{M}_0$ is unknown at $t = 0$, thus any possible correction using dynamics of prices is not feasible. It is interesting to note however that if relative prices do not change between $t = 0$ and $t = 1$, that is, if $\mathbf{p}_1 = \lambda \mathbf{p}_0$, then using that $P_\alpha(\mathbf{p})$ is homogeneous of degree 1 we have

$$\lambda^{-1}\tilde{M}_0 = \mathbb{E}[T_1] + \frac{\mathrm{Cov}\left(y_0, T_1\right)}{\mathbb{E}[y_0]},$$

and

$$\overline{T}_1 = \lambda^{-1}\overline{T}_0. \tag{17}$$

Hence, in this scenario, setting $\overline{T}_0 = \tilde{M}_0$ as in (16) (assuming that reducing the estimation error to zero at $t = 0$ is possible) and then making the correction in (17) will further reduce the estimation error if

$$\begin{aligned} \left|\tilde{M}_1 - \lambda^{-1}\tilde{M}_0\right| &\leqslant \left|\tilde{M}_1 - \tilde{M}_0\right| \\ \left|\frac{\mathrm{Cov}\left(y_1, T_1\right)}{\mathbb{E}[y_1]} - \frac{\mathrm{Cov}\left(y_0, T_1\right)}{\mathbb{E}[y_0]}\right|, &\leqslant \left|\mathbb{E}\left[T_1\right] - \mathbb{E}\left[T_0\right] + \frac{\mathrm{Cov}\left(y_1, T_1\right)}{\mathbb{E}[y_1]} - \frac{\mathrm{Cov}\left(y_0, T_0\right)}{\mathbb{E}[y_0]}\right|. \end{aligned}$$

Moreover, if $y_1 = Ry_0$ for some $R > 0$, then implementing this correction will reduce the error

to zero, just as what happened in Section 3.3.

Compute the error as a function of $\overline{\alpha}$ for some parameters?

5. Aggregation across goods

5.1. Economy setting

Consider a two period economy with $n+1$ goods available, $(\mathbf{x}, z)$, valued at prices $(\mathbf{p}, q)$. In what follows and for simplicity of notation, the subindices t for all variables will be omitted unless needed. In this economy, consider one consumer whose income is $y \in (0, \infty)$.[11] The vector $\alpha \in (0,1)^n$ determines the form of our consumer's pseudo-Cobb-Douglas (PCD) utility function,

$$u_\alpha(\mathbf{x}, z) := u_{(y,\alpha)}(\mathbf{x}, z) = \sum_{j=1}^{n} x_j^{\alpha_j} z^{1-\alpha_j}.$$

Hence, the agent can be identified with the pair (y, α). This consumer chooses how much of the $\mathbf{x}$ goods to consume in each period according to

$$\begin{aligned} \mathbf{x}(\mathbf{p}, q, y) := \arg\max_{\mathbf{x}, z} \quad & u_\alpha(\mathbf{x}, z), \\ \text{s.a.} \quad & \mathbf{p}\mathbf{x} + qz = y. \end{aligned} \tag{18}$$

Observe that if all α_j are different, then it is not possible to bundle the $\mathbf{x}$ goods into a category. Indeed, as mentioned in Section 2.2, without assuming certain price dynamics, the only option is to be in presence of functional separability. A necessary condition for this to happen is that the $\mathbf{x}$ goods are independent of z in terms of the preference, that is, (5) holds. It is straightforward to see that for the preference relation represented by the function u_α this property is not satisfied unless all α_j are equal. Furthermore, when $\alpha_j = \alpha_0$ for every j, aggregation is possible as was seen in Section 4.

5.2. The estimation problem

An investigator at time $t = 0$ is interested in studying the evolution of the demand on the $\mathbf{x}$ goods. For that purpose, instead of focusing in the individual demands on the $\mathbf{x}$ goods she wishes to construct a quantity index for them, $g_{\overline{\alpha}}$, like the one defined in (9) in Section 3.1. In other words she wants to choose a value $\overline{\alpha}$ such that by defining U and $P_{\overline{\alpha}}$ as in (10) and

[11] I am not claiming there is only one consumer in the economy. Instead I am just focusing on one of them.

(11), respectively, the category demand

$$\begin{aligned} X(\mathbf{p}, q, y) := \underset{X,z}{\arg\max} \quad & U(X, z) \\ \text{s.a.} \quad & P_{\overline{\alpha}}(\mathbf{p})X + qz = y, \end{aligned} \tag{19}$$

is the closest possible to $g_{\overline{\alpha}}(\mathbf{x}(\mathbf{p}), q, y)$. In general, the equality

$$X(\mathbf{p}, q, y) = g_{\overline{\alpha}}\big(\mathbf{x}(\mathbf{p}, q, y)\big),$$

will not hold as was discussed in the previous section, but the choice of $\overline{\alpha}$ can be made optimally to reduce the difference as much as possible.

Following the discussion in the previous section, the question that arises in this setting is: Which distribution is being approximated by assuming there is only two, instead of $n+1$ goods in the economy?. The answer is straightforward but the interpretation is subtle. Since $\overline{\alpha}$ is replacing every α_j with a single parameter, then what is being approximated is the "distribution" of the α_j. To see this, we can regard the heterogeneous preferences between the $\mathbf{x}$ goods as an equiprobable distribution with mass at each α_j. This way of seeing this approximation is coherent with the previous examples where the distributions were representing the heterogeneity of the context, in particular the different preferences between individuals in the economy.

5.3. Measuring the error

6. Concluding remarks

References

S. Basu and J. G. Fernald. Returns to scale in u.s. production: Estimates and implications. *Journal of Political Economy*, 105(2):249–283, 1997. ISSN 00223808, 1537534X. 2

C. D. Carroll. Requiem for the representative consumer? aggregate implications of microeconomic consumption behavior. *American Economic Review*, 90(2):110–115, May 2000. doi: 10.1257/aer.90.2.110. 3

R. C. Feenstra and G. H. Hanson. Aggregation bias in the factor content of trade: Evidence

from u.s. manufacturing. *American Economic Review*, 90(2):155–160, May 2000. doi: 10.1257/aer.90.2.155. 3

M. Friedman. The methodology of positive economics. In *Essays in Positive Economics*, A Phoenix book. Business economics, pages 3–43. University of Chicago Press, 1953. ISBN 9780226264035. 2

W. M. Gorman. Community preference fields. *Econometrica*, 21(1):63–80, 1953. doi: 10.2307/1906943. 3, 6

W. M. Gorman. Separable utility and aggregation. 27(3):469–481, 1959. doi: 10.2307/1909472. 4

E. A. Hanushek, S. G. Rivkin, and L. L. Taylor. Aggregation and the estimated effects of school resources. *The Review of Economics and Statistics*, 78(4):611–627, 1996. ISSN 00346535, 15309142. 3

J. R. Hicks. *Value and Capital: An Inquiry Into Some Fundamental Principles of Economic Theory*. Clarendon paperbacks. Clarendon Press, 1946. ISBN 9780198282693. 4, 9

J. Imbs, H. Mumtaz, M. O. Ravn, and H. Rey. PPP Strikes Back: Aggregation And the Real Exchange Rate*. *The Quarterly Journal of Economics*, 120(1):1–43, 02 2005. ISSN 0033-5533. doi: 10.1162/0033553053327524. 3

A. Kajii and S. Morris. The robustness of equilibria to incomplete information. 65(6):1283–1309, 1997. doi: 10.2307/2171737. 2

K. C. Lee, M. H. Pesaran, and R. G. Pierse. Testing for aggregation bias in linear models. 100(400):137–150, 1990. doi: 10.2307/2234191. 3

W. Leontief. Composite commodities and the problem of index numbers. 4(1):39–59, 1936. doi: 10.2307/1907120. 4, 9

A. Lewbel. Aggregation without separability: A generalized composite commodity theorem. *The American Economic Review*, 86(3):524–543, 1996. doi: 10.2307/2118210. 4, 9

M. Ravallion. Does aggregation hide the harmful effects of inequality on growth? *Economics Letters*, 61(1):73 – 77, 1998. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(98)00139-6. 3

J. Sutton. Market structure: Theory and evidence. volume 3, chapter 35, pages 2301–2368. Elsevier, 1 edition, 2007. 2

M. M. ter Vehn and S. Morris. The robustness of robust implementation. *Journal of Economic Theory*, 146(5):2093 – 2104, 2011. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2011.03.011. 2

H. Varian. *Microeconomic Analysis*. Norton International edition. Norton, 1992. ISBN 9780393960266. 5