Economic Development, Undernutrition and Diabetes

Nancy Luke, Kaivan Munshi, Swapnil Singh and Anu Mary Oommen*

December 8, 2022

Abstract

This research connects two seemingly unrelated facts that have recently been documented in developing countries, with important consequences for global health: (i) the weak association between nutritional status, which we measure by BMI, and income, and (ii) the elevated risk of diabetes among normal-weight individuals. Our model is based on a set point for BMI that is adapted to food supply in the pre-modern economy, but which subsequently fails to adjust to rapid economic change. During the process of development, some individuals thus remain at their low-BMI set point, despite the increase in their income (food consumption), while others who have escaped their set point (but are not necessarily overweight) are at increased risk of diabetes. The model and the underlying biological mechanism, which are validated with micro-data from many countries, can jointly explain inter-regional (Asia versus Africa) differences in nutritional status and the prevalence of diabetes.

*Luke: Pennsylvania State University, nkl10@psu.edu. Munshi: Yale University and Toulouse School of Economics, kaivan.munshi@yale.edu [corresponding author]. Singh: Bank of Lithuania and Kaunas University of Technology, ssingh@lb.lt. Oomen: Christian Medical College, anuoommen@cmcvellore.ac.in. Research support from the National Institutes of Health through grant R01-HD046940, Cambridge-INET, the Keynes Fund and the Newton Trust at the University of Cambridge, and the Agence Nationale de la Recherche (ANR) under the EUR Project ANR-17-EURE-0010 is gratefully acknowledged.

1 Introduction

Two recently documented facts run counter to the conventional wisdom that economic development leads to better health: first, the absence of a clear link between nutritional status and income in developing countries (Deaton, 2007; Swaminathan et al., 2019) and second, a surge in diabetes, with a surprisingly high prevalence of this condition among *normal* weight individuals, in these countries (Narayan and Kanaya, 2020). Our objective in this paper is to develop and test a model with three ingredients – adaptation, mismatch and a set point – that can explain these seemingly unrelated observations. We discuss these ingredients in sequence below.

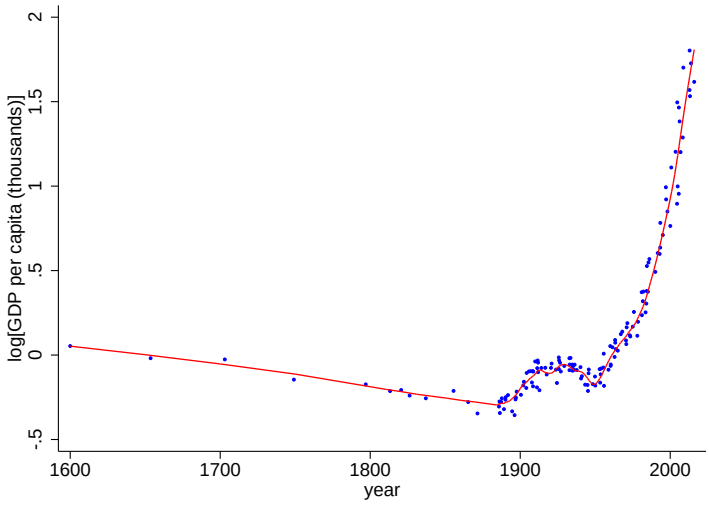
Two models of early-life adaptation or developmental plasticity have been proposed by evolutionary biologists: (a) The *developmental constraints* model in which developing organisms in severely resource limited environments make immediate tradeoffs to protect critical functions and improve survival in early life (Barker, 1995). (b) The *predictive* model in which maternal cues *in utero* predict the (normal) adult environment and the organism evolves accordingly in anticipation of future conditions (Gluckman and Hanson, 2006). While the developmental constraints model will apply to birth cohorts that face extreme or novel nutritional insults (Bateson et al., 2014), the predictive model will be relevant when conditions in past generations are an accurate predictor of (average) conditions in the current generation (Burgess and Marshall, 2014; Lind et al., 2020). The pre-modern (Neolithic) economy was characterized by wide short-term fluctuations in food supply, but had growth rates close to zero for centuries. The predictive model would have been especially relevant in such an environment, resulting in a population whose body size was adapted to long-term (low) food supply, with the adaptation varying across space with agroclimatic conditions (Pomeroy et al., 2019).

With economic development, there is a substantial increase in income. Figure 1, for example, plots GDP per capita (in logs) for India from 1600 to 2016. Income is stable (declining mildly) for the first 350 years, after which it starts to increase steeply. This increase in income would have translated into an increase in food consumption. The developmental origins of adult disease literature posits that the resulting mismatch between current and ancestral consumption (to which the population is adapted) has contributed to the high rates of diabetes in developing countries; e.g. Gluckman and Hanson (2004); Wells et al. (2016); Narayan and Kanaya (2020). Our model places additional restrictions on the relationship between diabetes and the mismatch, while simultaneously explaining the persistence of undernutrition in these countries by characterizing the initial adaptation by a set point.

Many individuals (but not all individuals, as we will see) have a relatively stable set point for their bodyweight throughout adult life (Müller et al., 2010, 2018). This set point is part of a homeostatic (stabilizing) system that maintains the body's energy balance against fluctuations in food intake by making metabolic and hormonal adjustments.[1] We posit that the set point for a given dynasty (family)

[1] Homeostasis is a fundamental concept in biology, which describes how physiological systems maintain an equilibrium set point by counteracting environmental stresses. As discussed in Müller et al. (2010) and Speakman et al. (2011), numerous studies indicate that when the energy balance is perturbed in either direction through a change in diet, the body returns to its original weight once the nutritional constraint is released. Furthermore, energy expenditures are modulated to resist the perturbation, indicating that the body is actively defending its set point.

Figure 1: Evolution of Income in India



Source: Maddison Project Database (2018)
GDP per capita is measured in 2011 US dollars.

is determined by food supply in the pre-modern economy. While the adapted set point would have allowed pre-modern populations to maintain their energy balance, and to survive and reproduce, in an environment characterized by low and fluctuating food supply, it becomes a liability if it persists for multiple generations after the onset of economic development, as made precise below.

A property of all – physical and biological – homeostatic systems is that they can only self-regulate within fixed bounds and will malfunction when the environmental stresses to which they are subjected exceed a threshold level (Kültz, 2020).[2] This implies that as long as current and pre-modern (ancestral) consumption or, equivalently, income remain sufficiently close to each other, the body will successfully defend its bodyweight set-point. Once the gap between current and pre-modern income crosses a threshold, however, the body will no longer be able to defend the set point. Escape from the set point is associated with imbalance in energy regulation, which will be accompanied by imbalance in related (inter-linked) homeostatic systems. Failure of glucose homeostasis, in particular, manifests directly as diabetes.

Although it may be appropriate to characterize the set point with respect to weight for a given individual, we account for possible variation in height across generations by specifying a common set point for members of a dynasty with respect to their BMI; i.e. weight conditional on height. This normalization is especially useful for our analysis because BMI is a standard measure of adult nutritional status and is also associated with the risk of diabetes.[3] Based on the discussion above, it

[2]Stebbing (2009) uses this description of homeostasis (with a threshold) to explain the typical 'dose-response' relationship observed in toxicology: there is no effect of a toxin on the functioning of an organism until the concentration reaches a threshold, after which the effect is increasing linearly with concentration. In our application, the toxin is replaced by food intake, but the principle is the same.

[3]Height is another common measure of nutritional status and archaeological evidence indicates that Neolithic populations adapted to low food supply by adjusting their stature (Pomeroy et al., 2019). Although this is not the focus

follows that there will be two types of individuals in a developing economy: (i) Those individuals who remain at their pre-modern set point, despite the increase in their consumption, are partly responsible for the weak association between nutritional status, which we measure by BMI, and income. (ii) Those individuals who have escaped their set point, but are not necessarily overweight, are at increased risk of diabetes and related metabolic disorders.

The partition of the population that we have described may not be permanent. The assumption in many models of developmental plasticity is that the initial adaptation is epigenetic; i.e. it involves changes in gene expression and, hence, will persist for a limited number of generations (Jablonka and Raz, 2009; Lind and Spagopoulou, 2018). This would explain why European populations, which were also under-nourished historically, no longer exhibit the traits we document in developing-country populations.[4] It would also explain the health experience of migrants from developing countries to substantially wealthier advanced economies. For example, Alacevich and Tarozzi (2017) document that the nutritional status of immigrants from South Asia (a historically poor region) to the U.K. converges to the level of the native population very swiftly, presumably because they have escaped their set points. Given the persistence in these set points, South Asian immigrants residing in the U.K. and the U.S. are nevertheless many times more likely to have diabetes, conditional on their BMI, than the native population (McKeigue et al., 1991; Oza-Frank and Narayan, 2010).

If data on income, BMI, and diabetes were available for each family (dynasty) over many generations, going back to the pre-modern era, then we could test the preceding argument directly. For a given dynasty, we would expect to observe a discrete increase in BMI in a particular generation in which the gap between current and ancestral income exceeded a threshold, with an accompanying increase in the risk of diabetes. In the absence of such multi-generational household-level data, we take a deductive, model-based, approach familiar to economists (but not biologists) that proceeds in four steps.

First, by characterizing the evolution of income in the population across generations during the process of development, the dynamic model laid out in Section 2 generates implications at any point in time that do not require knowledge of ancestral income: (i) Although BMI is increasing in current household income at all levels, there is a discontinuous increase in the slope of this relationship at a particular income threshold. (ii) The risk of diabetes is constant below the same threshold and increasing in current income above the threshold. Viewed through the lens of the model, these cross-sectional relationships across households are informative about underlying causal relationships within households (dynasties) over generations. However, such causal interpretations are only appropriate

of our analysis, we provide evidence supporting this complementary adaptation in Section 3.1. Moreover, while we specify historical adaptation and contemporary diabetes risk with respect to BMI, Pomeroy et al. focus on a correlated but distinct measure: lean mass. There is evidence that the risk of diabetes is negatively associated with lean mass, conditional on BMI (Haines et al., 2022) and we will account for this when interpreting our results below.

[4]Cutler et al. (2006) note that there is a negative association between the risk of cardiovascular disease and income in advanced economies and Deaton (2007) notes that the weak association between height and income that he documents in developing countries is in sharp contrast with the corresponding associations in European populations. In these populations, it has been postulated that the epigenetically determined pre-modern set point is replaced by (i) "settling-points," which the body does not defend (Müller et al., 2010), or (ii) by genetically determined upper and lower "intervention points," between which nutritional status responds flexibly to food intake (Speakman, 2007).

if the model is correctly specified and, thus, much of the analysis will be devoted to validating the model.

We verify the cross-sectional implications of the model in Section 3.1, using Hansen's (2017) slope-threshold test, with nationally representative household data from the India Human Development Survey (IHDS). The weak association between BMI and household income below the estimated threshold, which is located close to the median income level in the Indian population, explains (in part) the persistence of undernutrition in that population. The steep increase in the risk of diabetes with income above the same threshold, which corresponds to a BMI that is in the middle of the normal range, helps explain the second stylized fact.[5] Our interpretation of these twin findings is that BMI and the risk of diabetes increase simultaneously and independently when an underlying homeostatic system (maintaining a low BMI) breaks down. As discussed in Section 3.2, alternative determinants of nutritional status and diabetes in developing countries, such as childhood illness, diet, and lifestyle cannot plausibly explain these twin findings in the absence of a set point. Providing additional support for our interpretation, the test of the model's internal validity in Section 3.3 verifies not only that a set-point threshold is present, but also the specific structure that is imposed on the threshold function in the BMI-income relationship.

Next, we assess the external validity of the model. The core analysis focuses on the Indian population because it is simultaneously characterized by high levels of undernutrition and a high prevalence of diabetes. However, we expect the model to apply more generally, with the fraction of the population having escaped the set point in a given country depending on its stage in the process of development or, equivalently, the gap between current incomes and historical incomes. In line with the observation in Section 4.1 that the income-gap is greater in Indonesia than in India, the location of the precisely estimated income threshold with Indonesia Family Life Survey (IFLS) data indicates that three-quarters of the population has escaped its set point in that country. In contrast, per capita incomes have changed very little over time in Africa, which implies that African populations remain largely at their set points and explains why we are unable to detect a discontinuity with data from the Ghana Socioeconomic Panel Survey (GSPS).

Although information on income, BMI and diabetes is only available in a limited number of data sets, the Demographic Health Survey (DHS) and the WHO-STEPS surveys provide individual-level information on BMI and diabetes (with biomarkers) for many countries. We use these data to examine an additional implication of the model, which is that the positive association between the risk of diabetes and BMI is characterized by a slope discontinuity (precisely estimated at a BMI of 21.8 with Indian data) in Section 4.2. Moving from micro-data to cross-country comparisons in Section 4.3, we document that BMI conditional on current income is greater in African than in Asian countries, mirroring Deaton's (2007) findings with height as the measure of nutritional status, whereas the

[5]Diabetes is self-reported and, hence, under-reported in the IHDS. For all analyses that utilize self-reported health data in this paper, we thus construct a composite variable that indicates whether an individual has been diagnosed with diabetes or with either of two highly correlated comorbidities: hypertension and cardiovascular disease (Petrie et al., 2018). We validate this measure, which we refer to as "metabolic disease" to distinguish it from our measures of diabetes based directly on biomarkers, in Sections 3.1 and 4.2.

prevalence of diabetes conditional on BMI is greater in Asia (this is also observed with the micro-data). We show that these seemingly unrelated findings can be interpreted through the lens of our model, once we account for the fact that African populations remain largely at their set points and that historical per capita incomes, which determine the set points, were higher in Africa than in Asia.

Finally, we validate the biological relationships that serve as the starting point for our model: (a) BMI is determined by ancestral income, which is associated with the set point, below a threshold and by current income above the threshold. (b) The risk of diabetes is increasing in the difference between current and ancestral income, above but not below the same threshold. We do this by constructing exogenous measures of ancestral (pre-modern) per household income at (i) the district level with FAO-GAEZ crop suitability data in Section 5.1, using a method suggested by Galor and Özak (2016), and (ii) at the village level in Section 5.2, using data on the agricultural revenue tax that was collected by the British colonial government in 1871, based on its independent assessment of local agricultural productivity. The district-level measures of ancestral income are merged with the IHDS and IFLS datasets that we use to test the cross-sectional implications of the model for India and Indonesia, respectively. The village-level measures, which are available for villages in the modern Indian state of Tamil Nadu, are merged with data from the South India Community Health Study (SICHS) which provides information on income, BMI and diabetes for a representative sample of households in rural Vellore district. For these validation exercises, the location of the threshold is derived from the cross-sectional tests discussed above.

CDC statistics indicate that 9.5% of diabetics in the U.S. are normal weight (with a BMI below 25). Using a more stringent BMI cutoff of 23, recommended for Asian populations, we find that 55% of diabetics in the 2015-16 round of the India DHS are normal weight. Our model provides an explanation for this striking difference, based on a low-BMI set point that is specific to developing-country populations and which determines both the BMI distribution and the risk of diabetes. We will return to this observation in the concluding section where we discuss the policy implications of our analysis.

2 The Model

2.1 Population and Income

The population consists of a large number of infinitely lived dynasties (families). Each dynasty consists of a single individual in each generation, who is replaced by a single descendant in the generation that follows. There is a fixed return on wealth in each generation; i.e. an income flow, which is consumed, so that the stock is passed on (without depletion) to the next generation. We will thus use the terms (permanent) income and wealth interchangeably in the discussion that follows. Income is the same in each generation during the pre-modern era, in which epigenetic adaptation takes place, but subsequently evolves. Denote the logarithm of the dynasty's initial income by y_0. Permanent income in the modern economy is well approximated by the log-normal distribution (Battistin et al., 2009).

We thus assume that each dynasty receives a permanent, additive and independent income shock u_τ in each subsequent period or generation τ, where $u_\tau \sim N(\mu, \sigma^2)$. Solving recursively, log-income of a dynasty in period t is

$$y_t = y_0 + U_t, \tag{1}$$

where $U_t = \sum_{\tau=1}^{t} u_\tau \sim N(t\mu, t\sigma^2)$.[6] For ease of exposition, we will denote $t\mu$ by μ_t and $t\sigma^2$ by σ_t^2.

2.2 Biological Relationships

We now characterize the biological relationships between (i) BMI and income, and (ii) the risk of diabetes and income, during the process of economic development. This characterization is based on the verbal description from the preceding section.

There is a positive and continuous relationship between (food) consumption and income in all time periods.[7] Focussing first on the initial period in which the set point is determined, it follows that nutritional status, which we measure by BMI z_0, is increasing continuously in pre-modern income y_0, as specified below:

$$z_0 = a + by_0. \tag{2}$$

In subsequent periods, each descendant's body will defend her dynasty's set point z_0 in the face of fluctuations in consumption that arise due to the permanent income shocks. However, as noted, the body can only respond up to a point to deviations in income from the initial level, y_0, that determined the set point. There is thus a threshold α, such that BMI in period t,

$$z_t = \begin{cases} a + by_0 & \text{if} \quad U_t \leqslant \alpha \\ a + by_t & \text{if} \quad U_t > \alpha \end{cases} \tag{3}$$

Equation (3) imposes the restriction that the (linear) relationship between BMI and income is the same, below and above the threshold; what changes is the relevant measure of income, from y_0 to y_t. Later in the analysis, we will validate the structure we have imposed in equation (3) by separately estimating the b parameter, below and above the (estimated) threshold.[8]

Notice that we do not specify a lower threshold for the set point. Given low levels of food supply in the pre-modern era, the population would have been adapted to defend the set point especially

[6] We do not include a dynasty-specific identifier when deriving and characterizing the income equation to simplify notation.

[7] The implicit assumption is that individuals do not alter their behavior to account for the effect of the set point on their nutritional status and the risk of diabetes during the process of development. This seems reasonable, given that the effect of the set point on these outcomes is the subject of our inquiry and, thus, is not known to the general population. This assumption also allows us to specify the biological relationships with respect to income rather than (more proximate) consumption.

[8] While we focus on adaptation with respect to body size, in line with the modern evolutionary biology literature, the "thrifty genotype" hypothesis (Neel, 1962) posits that body weight in historically undernourished populations will be more responsive to the increase in food consumption that accompanies economic development. This implies that there should be an additional $y_o \cdot y_t$ term above the threshold in equation (3), with a negative coefficient. If that were the case, however, then we would fail the validation test that follows in Section 3.3.

vigorously against downward fluctuations in consumption.[9] Although mean income is increasing with economic development in our model, the distribution of income shocks is unbounded and, hence, a small number of dynasties could, nevertheless, face a sequence of very negative shocks that the body could not defend. However, all societies have consumption-smoothing mechanisms in place to insure against precisely such negative outcomes and these mechanisms improve with economic development. We thus assume that dynasties always successfully defend the set point z_0 in the face of negative income shocks, either biologically or by taking advantage of social safety nets to augment their consumption.[10]

As long as income remains within the threshold associated with the dynasty's set point, metabolic and hormonal adjustments ensure that the increases in consumption that accompany the increases in income due to economic development do not translate into increases in BMI. Once income crosses the threshold, however, the body can no longer defend the set point and BMI starts to track current income. As discussed in the preceding section, this simultaneously increases the risk of diabetes. As in the developmental origins of adult disease literature, this risk is specified to be increasing in the mismatch between current income, y_t, and initial income, y_0. The additional feature of our model is that the income-gap only determines the risk of diabetes when it exceeds a threshold (and the individual escapes the set point). The relationship between the probability of diabetes, $P(D_t)$, and income can thus be characterized as follows:[11]

$$P(D_t) = \begin{cases} \gamma_1 & \text{if} \quad U_t \leqslant \alpha \\ \gamma_1 + \gamma_2(y_t - y_0) & \text{if} \quad U_t > \alpha \end{cases} \tag{4}$$

2.3 Cross-Sectional BMI-Income Relationship

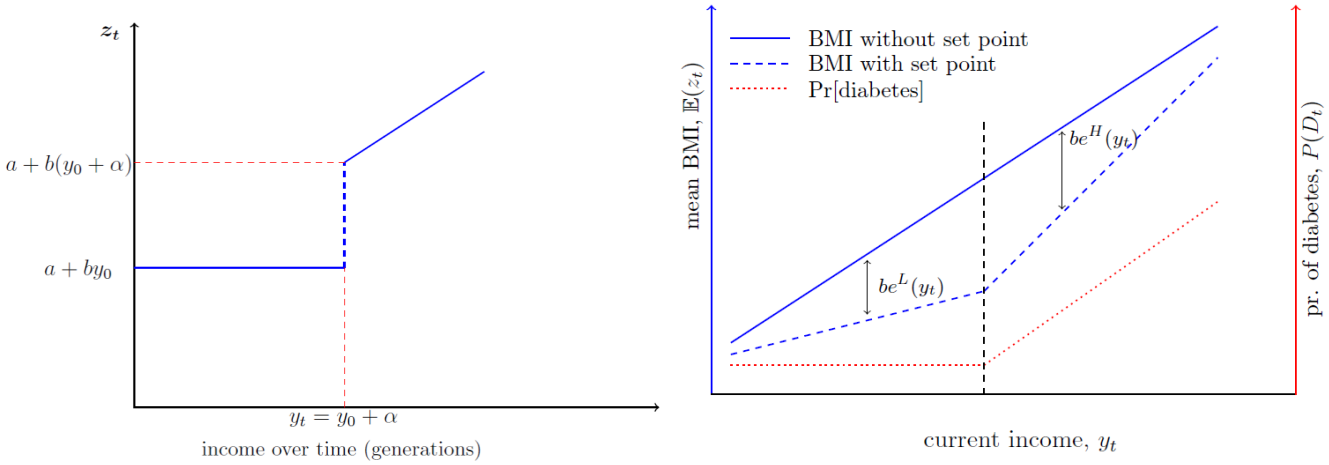
Figure 2a describes the evolution of BMI across multiple generations (time periods) for a single dynasty, based on the biological relationship specified above. For expositional convenience, we assume that the dynasty only receives positive income shocks. Starting from an initial income, y_0, the dynasty's income thus increases monotonically across generations. However, its members' BMI will remain at the dynasty's set point, $z_0 = a + by_0$, until y_t exceeds $y_0 + \alpha$. At that point in time, there will be a discrete increase in BMI, after which BMI will track y_t. If data over many generations, going back to the pre-modern period, were available for each dynasty, then these implications could be tested directly. In the absence of such multi-generational data, we proceed to derive the cross-sectional association between BMI and income, as implied by equation (3), when a dynasty-specific set point for BMI is present.

[9] This is consistent with the conventional view that the regulation of bodyweight is more responsive to weight loss than to weight gain (Müller et al., 2010). For example, despite repeated weight cycling in response to seasonal fluctuations in food supply, minimal bodyweight in a sample of rural Gambian women remained extremely stable (within 1.5 kg.) over a period of 10 years (Prentice et al., 1992).

[10] Given that income shocks are positive on average and their distribution is symmetric, such redistribution is feasible. We are effectively ignoring catastrophic common shocks, such as famines, that can shift set points in an entire birth cohort. Such events have always been rare and are less relevant in the modern economy.

[11] $\gamma_1 > 0$, $\gamma_2 > 0$ in equation (4). The implicit assumption, which is consistent with recent evidence on diabetes reversal is that the risk of diabetes can change in both directions over time as the individual's BMI shifts on either side of the threshold.

Figure 2: BMI, Diabetes and Income



(a) Within a dynasty over time

(b) Across households in the cross-section

We normalize so that the initial income distribution is bounded below at zero. We also do not specify a lower threshold for the set point. It follows that all individuals with $y_t \leq \alpha$ must be at their set point; some of these individuals will belong to dynasties that had initial incomes below α and which subsequently increased their income by relatively little, whereas others will belong to dynasties whose income has drifted down over time. Mean BMI at any given level of income $y_t \leq \alpha$ can then be characterized by the following expression:

$$\mathbb{E}(z_t|y_t) = \int_{-\infty}^{y_t} [a + b(y_t - U_t)] P(U_t \mid y_t) \ \mathrm{d}\, U_t$$

where $P(U_t|y_t)$ is the conditional density function of U_t given y_t. As shown in Appendix A, our distributional assumptions together with a simplifying (empirically validated) analytical approximation allow us to express the preceding equation as follows:

$$\mathbb{E}(z_t|y_t) = \int_{-\infty}^{y_t} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \ \mathrm{d}\, U_t = a + b\left(y_t - e^L(y_t)\right) \tag{5}$$

where $e^L(y_t) = \frac{1}{\Phi(y_t;\mu_t,\sigma_t^2)} \int_{-\infty}^{y_t} U_t \phi(U_t; \mu_t, \sigma_t^2) \ \mathrm{d}\, U_t = \mu_t - \sigma_t \Lambda\left(\frac{y_t - \mu_t}{\sigma_t}\right)$ and $\Lambda(\cdot)$ is the inverse Mills ratio.

For individuals with $y_t > \alpha$, some will have crossed their set point threshold, while others (who started with a higher initial income) will remain at their set point. The expression for mean BMI at a given level of income $y_t > \alpha$ thus includes both types of individuals. Incorporating the same analytical

approximation and distributional assumptions as above:

$$\begin{aligned}\mathbb{E}(z_t|y_t) &= \int_{-\infty}^{\alpha} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &\quad + \int_{\alpha}^{y_t} [a + by_t] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + b\left(y_t - e^H(y_t)\right) \end{aligned} \tag{6}$$

where $e^H(y_t) = \frac{1}{\Phi(y_t;\mu_t,\sigma_t^2)} \int_{-\infty}^{\alpha} U_t \phi(U_t; \mu_t, \sigma_t^2) \, \mathrm{d}\, U_t = \frac{\mu_t \Phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right) - \sigma_t \phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)}{\Phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)}$

Given the specifications of the $e^L(y_t)$, $e^H(y_t)$ functions, we can derive the following result (the proof is in Appendix A):

Proposition 1 *(i) The slope of the BMI-income relationship is positive but less than b for $y_t \leqslant \alpha$ and greater than b for $y_t > \alpha$. (ii) There is a discontinuous change in the slope of the BMI-income relationship at $y_t = \alpha$. (iii) There is no level discontinuity in the BMI-income relationship at $y_t = \alpha$.*

The association between BMI and income implied by Proposition 1 is described graphically in Figure 2b. Each dynasty transitions discretely to a higher BMI level, at a particular point in time, in Figure 2a. This level-shift is smoothed out, and translates into a slope change at a particular income level, when we derive the corresponding cross-sectional BMI-income relationship across dynasties, at any point in time.

The preceding implication is robust to alternative specifications of the set point. Although an epigenetically determined set point may be heritable, it will ultimately cease to be relevant once a changed economic environment has been in place for a sufficient number of generations. Our model thus describes the relationship between nutritional status and income over a finite number of generations during the initial phase of economic development. During this phase, we assume that the set point, z_0, determined in period 0, is fixed. However, an alternative specification would allow the set point to adjust gradually across generations until it is no longer relevant. For example, the set point could be specified as a weighted average of y_0 and y_t, with the weight on y_t increasing over time. Alternatively, the set point could be determined by initial conditions (income) in each generation. Since income does not vary within periods in our setup, the set point in period t with this alternative specification will then be parental income, y_{t-1}. As shown in Appendix A, the alternative specifications generate the same qualitative predictions as Proposition 1. What distinguishes the benchmark specification in equation (3) from the alternatives, as verified empirically in Section 5, is that BMI below the estimated current-income threshold is determined exclusively by y_0.

The test of internal validity, reported in Section 3.3, will provide additional and independent support for our preferred specification of the set point and, more generally, the structure we have imposed on the model. This stringent test is based on the observation, from equations (5) and (6), that once $e^L(y_t)$ and $e^H(y_t)$, respectively, are subtracted from y_t, the slope of the BMI-adjusted income association will be the same (equal to b) below and above the estimated threshold. If the specification

of the BMI-income relationship in equation (3), the distributional assumptions, or the analytical approximation that we use to derive closed-form expressions for $e^L(y_t)$, $e^H(y_t)$ were invalid, then we would fail the test of internal validity.

2.4 Cross-Sectional Diabetes-Income Relationship

Taking as given the biological relationship between the probability of diabetes, $P(D_t)$, and income, as specified in equation (4) for a single dynasty, the corresponding association in the cross-section across dynasties can be derived as follows:

Proposition 2 *(i) There is no relationship between $P(D_t)$ and y_t for $y_t \leqslant \alpha$, and a positive relationship for $y_t > \alpha$. (ii) There is a discontinuous change in the slope of the $P(D_t) - y_t$ relationship at $y_t = \alpha$. (iii) There is no level discontinuity in the $P(D_t) - y_t$ relationship at $y_t = \alpha$.*

The proof in Appendix A follows the same steps as the proof of Proposition 1. The $P(D_t) - y_t$ relationship specified by Proposition 2 is described graphically in Figure 2b. This relationship is qualitatively the same as the $\mathbb{E}(z_t) - y_t$ association, except that the slope is zero below the threshold. This is because the risk of diabetes is constant (and the same) for all individuals who remain at their set point and because all individuals below the income threshold are at their set point. Above the threshold, in contrast, the risk of diabetes is increasing in income. This is due to (i) the greater fraction of individuals who have escaped their set point, and (ii) the increased risk for those who have escaped. Note that the model predicts that the $\mathbb{E}(z_t) - y_t$ and $P(D_t) - y_t$ associations will exhibit a slope discontinuity at the same income level: $y_t = \alpha$.[12]

Proposition 1 indicates that BMI is increasing with income at all levels, more steeply above a threshold, while Proposition 2 indicates that the risk of diabetes is only increasing in income above the same threshold. Bringing the two implications together, it follows that there will be no association between the risk of diabetes and BMI up to a BMI threshold (which corresponds to the underlying income threshold) and a positive association thereafter. Although our analysis focuses on the BMI-income and diabetes-income relationships, we will examine this additional implication of the model, which is especially relevant for policy in Section 4.2.

3 Testing the Model

3.1 Cross-Sectional Analysis

The core data set that we use to test the model is the India Human Development Survey (IHDS). This nationally representative household survey, which was conducted in 2004-2005 and 2011-2012, includes detailed information on household income, nutritional status for adults residing in the household at

[12] Although we normalize so that the initial income distribution is bounded below at zero, it can more generally be bounded below at some income level $\underline{y}_0$, in which case the threshold would be located at $y_t = \underline{y}_0 + \alpha$. This would change the interpretation of the threshold location, but otherwise leave the analysis unchanged.

the time of the survey, and the self-reported prevalence of different diseases among adult members of the household. The survey includes, in addition, information on household composition, food consumption expenditure in the last month, morbidity among the children in the last month, and district-level geographic locators, which will be used to supplement the analysis.[13]

The key variables in our analysis – income, BMI and diabetes – are measured as follows:(i) Although a dynasty consists of a single individual in each generation in our model, multiple individuals will reside in a household in practice. Income is thus measured at the household level, as the average over the 2004 and 2012 rounds.[14] This smooths out noise in the round-specific income measures and given that the rounds were conducted nearly a decade apart, provides a more accurate estimate of the household's permanent income. We also report instrumental variable estimates below that account for reverse causality and for measurement error in the income variable. (ii) Nutritional status is measured by the BMI of the household head and his spouse in each survey round. BMI is defined as weight conditional on height (kg./m^2) and both height and weight are directly measured. (iii) Given that diabetes is self-reported and, hence, under-reported in the IHDS, we construct a composite variable, "metabolic disease," which indicates whether a given individual has been diagnosed with diabetes or with either of two highly correlated comorbidities: hypertension and cardiovascular disease. This indicator is constructed for the household head and his spouse in each survey round, consistent with the implicit assumption in the model that diabetes is reversible, and with recent experimental evidence (Taylor, 2013). Later in Section 4.2 we will validate our composite measure of diabetes by comparing it with a direct indicator, based on biomarkers, obtained from the India DHS. We also verify below that the tests of the model go through with self-reported diabetes alone.

We test the implications of the model by nonparametrically estimating the BMI-income and metabolic disease-income relationships using the measures described above. Although our analysis focuses on the association with income, other individual and household characteristics, which are omitted from the model for expositional convenience, could independently determine BMI and the risk of diabetes. All of the estimating equations in our analysis thus include the following standard set of covariates: age in years (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district and survey-round. These covariates are partialled out using the Robinson (1988) procedure prior to the nonparametric estimation reported in Figure 3a.[15]

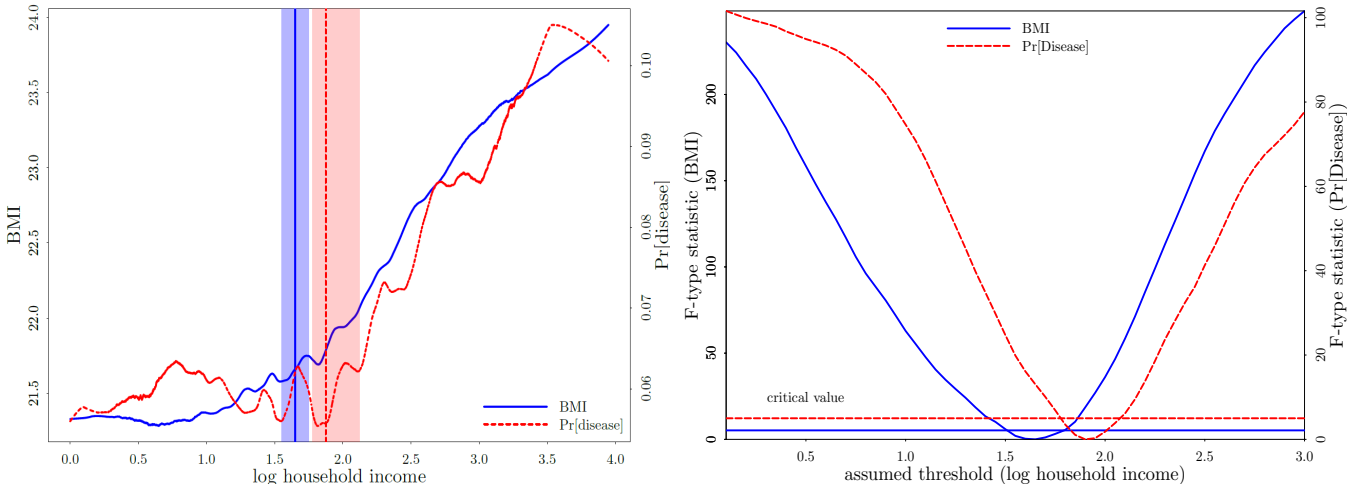
The vertical lines in Figure 3a mark the point where we locate an income threshold, based on the

[13]The Demographic Health Survey (DHS), which is used by Deaton (2007) also contains many of these variables. However, the DHS is not suitable for our purposes because it only collects indicators of asset ownership, which must then be converted into a crude wealth statistic using principal component analysis. The tests of the model, particularly the statistical tests to locate a slope-change at an income threshold, cannot be implemented without fine-grained income data. We will, however, use DHS data in Section 4.2 to examine the diabetes–BMI association that is implied by the model.

[14]Household income, measured in thousands of Rupees per month, includes farm income, non-farm business income, wage income, remittances, and government transfers. To make incomes in the two rounds comparable, we adjust 2004-2005 incomes to 2011-2012 prices. For rural areas, the correction is based on the Consumer Price Index (CPI) for agricultural wage labor and for urban areas it is based on the CPI for industrial workers.

[15]Observations in the top and bottom 1% of the outcome distribution are excluded from the estimation sample in all of our analyses. This ensures that the estimation results are not driven by extreme outliers.

Figure 3: Nutritional Status and Metabolic Disease with respect to Household Income



(a) Nonparametric relationship

(b) Threshold test

Source: India Human Development Survey (IHDS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are partialled out prior to nonparametric estimation. The same set of covariates are included in the estimating equation at each assumed threshold for the threshold test.
The vertical lines mark the estimated threshold location and the shaded areas demarcate the corresponding confidence intervals. Cluster bootstrapped 5% critical values are used to bound the threshold location.

statistical test described below. The shaded area around each line marks the 95% confidence interval for the threshold location, based on the same test. It is evident with each outcome that the association with income is relatively weak below the estimated threshold, and much stronger above the threshold. A slope discontinuity is not readily apparent with BMI as the outcome in Figure 3a. However, we can detect its presence with a high degree of statistical confidence and sharper discontinuities will be observed with other datasets (IFLS, SICHS) below. Notice also that the estimated threshold location is slightly lower with BMI as the outcome. Such minor differences are to be expected, given that BMI is directly measured, whereas metabolic disease (although diagnosed) is self reported. Nevertheless, this discrepancy is not observed in the robustness tests that follow and in the subsequent analyses with South Indian (IHDS) and Indonesian (IFLS) data.

The threshold locations and confidence intervals in Figure 3a are estimated using a procedure developed by Hansen (2017). This procedure involves sequential estimation of the following piecewise linear equation:

$$z_i = \beta_0 + \beta_1 y_i + \beta_2 (y_i - \tau) \times \mathbb{I}(y_i - \tau > 0) + x_i \lambda + \epsilon_i, \tag{7}$$

where z_i is an outcome of interest; e.g. BMI, y_i is household $i's$ income, τ is the location of the income threshold (which must be estimated), $\mathbb{I}(\cdot)$ is an indicator function, β_1, β_2 are slope parameters, and x_i is a vector of additional covariates (the same covariates that are partialled out prior to nonparametric estimation). This equation is estimated at different assumed income thresholds (values of τ), starting at a very low income level and then covering the entire income range in small increments. An F-type

Table 1: Piecewise Linear Equation Estimates - nutritional status and metabolic disease

Dependent variable:	BMI	metabolic disease
	(1)	(2)
Baseline slope (β_1)	0.239**	0.002
	(0.057)	(0.002)
Slope change (β_2)	0.940**	0.028**
	(0.066)	(0.003)
Threshold location (τ)	1.65	1.90
	[1.55, 1.75]	[1.80, 2.05]
Threshold test $p-$value	0.000	0.000
Mean of dependent variable	22.002	0.074
N	76,949	148,928

Source: India Human Development Survey (IHDS)
Metabolic disease indicates whether the individual has been diagnosed with diabetes, hypertension, or cardiovascular disease. BMI is measured for adults present in the household at the time of the survey.
Logarithm of household income is the independent variable.
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in the estimating equation.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
Cluster bootstrapped 95% confidence bands for the threshold location are in brackets.
** significant at 5%, based on cluster bootstrapped confidence intervals.

statistic is computed at each assumed threshold, based on a comparison of the sum of squared residuals at that assumed threshold and the minimized value across all assumed thresholds. This statistic will have a minimum value of zero by construction, and the assumed income threshold corresponding to that value is thus our best estimate of the true threshold. If there is indeed a slope-change, then the F-type statistic will increase steeply as the assumed threshold moves away (on either side) from the income level at which it is minimized.

Figure 3b plots the F-type statistic across the range of assumed thresholds for each outcome. The assumed threshold (income level) at which the statistic is minimized corresponds to the location of the threshold in Figure 3a. The confidence interval for each threshold location in that figure is determined by the points of intersection between the F-type statistic and the 5% critical value line for the corresponding outcome in Figure 3b. The F-type statistic increases steeply as the assumed threshold moves away from the income level at which it is minimized for both outcomes, allowing us to locate the thresholds with a high degree of statistical confidence.

The same (wild) bootstrap procedure, clustered at the level of the primary sampling unit, that is used to compute the critical values and, hence, the 95% confidence interval for the threshold location in Figure 3b can also be used to compute standard errors for the slope coefficients, β_1 and β_2, in a piecewise linear equation estimated at the threshold we have located.[16] Moreover, a similar bootstrap

[16]Following Hansen (2017) and Roodman et al. (2019), a coefficient's significance at the 5% level is determined by cluster bootstrapped 95% confidence intervals. For ease of exposition we report cluster bootstrapped standard errors for each coefficient.

procedure can be used to test our statistical model with a slope change at an income threshold, as described in equation (7), against the null hypothesis that there is a linear relationship between household income and each of the outcomes. These results are reported in Table 1. We can easily reject the null that the relationship is linear, without a discontinuity at a threshold, with each outcome. This does not rule out the possibility that the true relationship is actually (highly) nonlinear, without a discontinuity. However, the test of internal validity that follows in Section 3.3 will provide statistical support for the specific structure we have imposed on the model.

The reported point estimates of the baseline slope coefficient (β_1) and the slope-change coefficient (β_2) are obtained at our best estimate of the true threshold, τ, for each outcome. As implied by our model with a set point, the slope increases to the right of the threshold with each outcome (the slope-change coefficient is positive and significant). Moreover, the slope to the left of the threshold is positive and significant with BMI, but not with the risk of diabetes (measured by metabolic disease) as the outcome.[17] The estimated threshold location ranges from 1.65 to 1.9 for the two outcomes and the median income in our nationally representative sample of households is 1.8. This implies that the lower half of the income distribution in India remains at its pre-modern BMI set point, whereas the upper half is at risk of diabetes.

We complete this section by verifying the robustness of this evidence in a number of ways. (i) Appendix B1: We include measures of household composition, which could independently determine decisions and behaviors that are relevant for nutritional status and health outcomes as additional covariates in the estimating equation.[18] We also construct a nonparametric shift-share instrument for household income, based on national-level growth in agricultural crop values over the 1966-2015 period, weighted by crop acreage shares at the district level in 1966, and then interacted with the rural dummy and household land ownership.[19] The instrumental variable estimates, which are based on exogenous changes in the U_t component of current income, account for measurement error in the permanent income variable, as well as for possible reverse causality; i.e. the effect of BMI or metabolic disease on household income. (ii) Appendix B2: We separate men and women. (iii) Appendix B3: We separately examine the components of BMI (height, weight) and metabolic disease (diabetes, hypertension, cardiovascular disease), as described below.

Although height is not the focus of our analysis, archaeological evidence indicates that stature was also adapted to pre-modern food supply (Pomeroy et al., 2019). Replacing BMI by height in our model, we would then expect a discontinuous association between height and income and this is indeed what we see in Appendix B3. Height varies relatively little across the range of incomes in our data. This implies that the weight-income relationship should match the BMI-income relationship reported above, which, once again, is what we find (with a slope discontinuity at almost the same

[17] The number of observations in Column 2 is substantially greater than in Column 1 for two reasons: (i) BMI, based on height and weight, can only be measured for adult individuals who were physically present at the time of the survey interview. (ii) BMI data were only collected for a small number of adult men in the 2004-2005 round.

[18] Household income and household composition are closely related, which is why we exclude these variables from the estimating equation in the benchmark specification.

[19] The nonparametric instrumental variable estimation follows Newey et al. (1999) and the estimates are validated with tests of the shift-share instrument developed by Goldsmith-Pinkham et al. (2020). See Appendix B1 for details.

income level). Unpacking the components of our metabolic disease measure, we observe in Appendix B3 that the risks of diabetes and hypertension track very closely with income and that the precisely estimated threshold location is the same for both disorders. Although a slope discontinuity cannot be detected statistically with cardiovascular disease, it exhibits the same qualitative association with income.

3.2 Alternative Explanations

The additional covariates in the estimating equations are included to account for independent determinants of nutritional status and metabolic disease in India. For example, spatial variation in food tastes, as emphasized by Atkin (2013, 2016), or in the disease environment, as documented by Dandona et al. (2017), are captured by the district dummies and the rural dummy. However, such controls may not be complete and the discussion that follows thus considers alternative explanations for our results. Any alternative explanation must first generate the discontinuous association between BMI and income that we have uncovered (and will further validate in Section 3.3).

We begin by examining the possibility that there is a discontinuous relationship between income and two important proximate determinants of nutritional status (BMI) in developing countries: nutrient intake and childhood illness, particularly diarrhoeal disease (Scrimshaw et al., 1968). Nonparametric estimates of the nutrient intake-household income relationship are reported in Figure 4a and corresponding estimates of the children's illness-household income relationship are reported in Figure 4b, using IHDS data. Nutrient intake is measured by the consumption of calories and fat (in grams) at the household level. Childhood illness is measured by whether the child is reported to have had diarrhea and cough in the past month. The standard set of covariates, plus household composition and the number of adults engaged in physical labor are partialled out prior to estimation using Robinson's procedure. The additional covariates are included to condition for energy expenditures, since energy (nutrient) intake net of these expenditures determines nutritional status.[20] We see that there is a positive and continuous relationship between the intake of calories and fat and household income in Figure 4a, as assumed in our model. In addition, there is a negative and continuous relationship between the incidence of both diarrhea and cough with household income in Figure 4b. Indeed, Hansen's test fails to locate a slope-change at any assumed threshold in Figures 4c and 4d.[21] The same result (not reported) is obtained with other measures of nutrient intake – sugar consumption – and children's illness – the incidence of fever.[22]

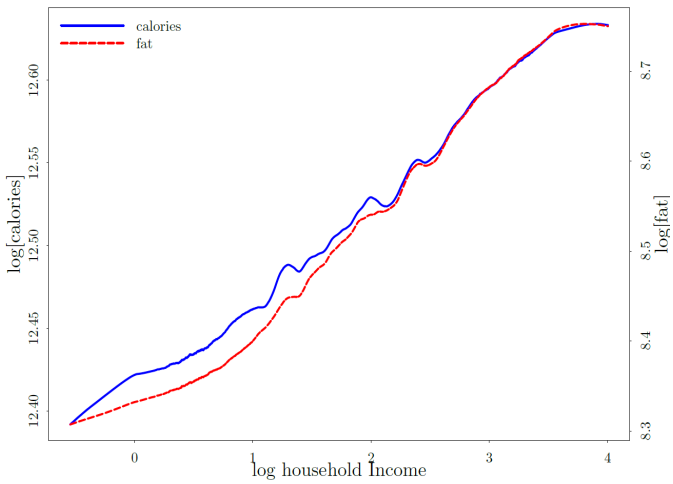
Selective child mortality, which Deaton (2007) considers as an explanation for his findings, could

[20] Ng and Popkin (2012) decompose total energy expenditures into types of activity: work, active leisure, travel, and domestic tasks. The work category accounted for over 80% of the total energy expenditure in 2000 and 2005 in India.

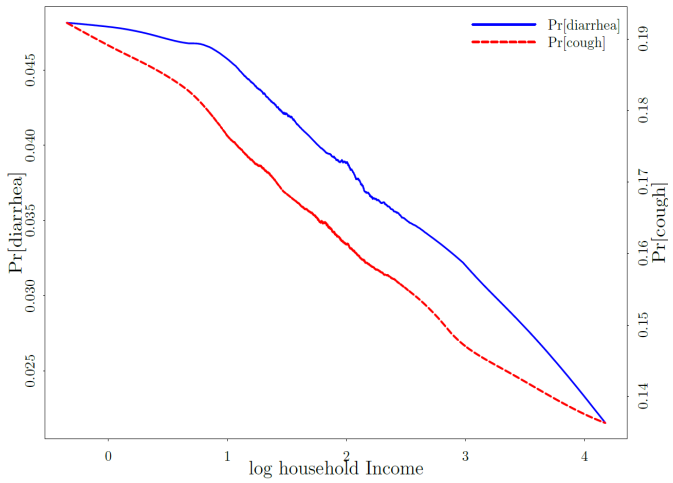
[21] We increase the sample size and, hence, the likelihood of detecting a threshold when estimating the children's illness-income relationship by including children aged 0-19. Tests that separate the children into 0-5 year olds and 5-19 year olds also fail to detect a threshold.

[22] Appendix Figure B5 examines the relationship between household income and expenditures on nine food categories: wheat, rice, cereals and derivative products, meat and eggs, milk and derivative products, pulses, vegetables, sugar and derivative products, and oil. Although a positive association is observed with each category, a slope discontinuity cannot be detected with any category.

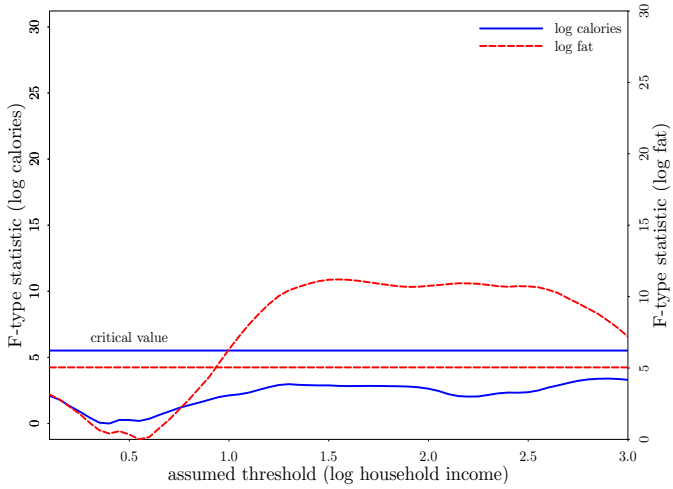
Figure 4: Nutrient Intake and Children's Illness with respect to Household Income



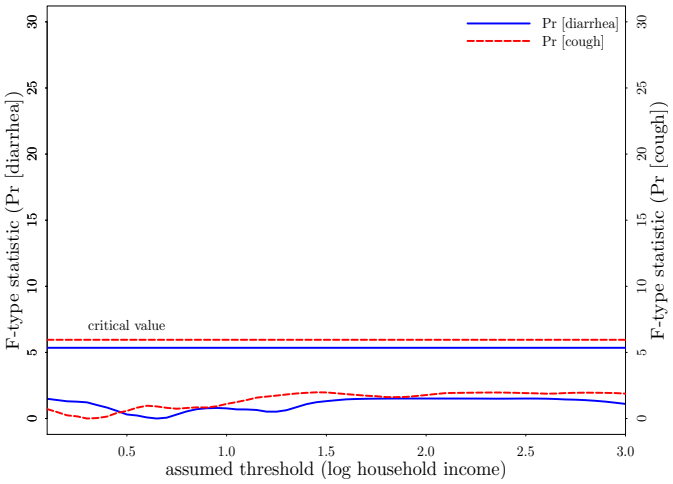
(a) Nutrient intake



(b) Children's illness



(c) Nutrient intake



(d) Children's illness

Source: India Human Development Survey (IHDS).
For the nutrient intake figures, the following covariates are partialled out prior to nonparametric estimation and included in the estimating equation at each assumed threshold: reported local price of rice, wheat, cereals and their derivative products, pulses, meat, sugar, oil, eggs, milk and its derivative products, vegetables and dummies for the number of children, adults, and teens in the household, dummies for the number of adults engaged in physical labor, caste group, rural area, district, and survey-round.
For the children's illness figures, age (linear, quadratic, and cubic terms) and dummies for gender, birth order, caste group, rural area, district and survey-round are partialled out prior to nonparametric estimation and included in the estimating equation at each assumed threshold.
Cluster bootstrapped 5% critical values are used to bound the threshold location.

generate a discontinuous BMI-income association, as shown in Appendix C. Poverty trap models generated by undernutrition; e.g. Dasgupta and Ray (1986) could also generate a discontinuity because of the feedback from BMI to income below a threshold.[23] However, neither of these explanations

[23]When poverty traps are generated by credit constraints and non-convexities, as in Galor and Zeira (1993) and Banerjee and Newman (1993), households with sufficiently low initial income will remain permanently at that level.

stands up to closer scrutiny. As discussed in Appendix C, the discontinuity generated by selective child mortality is driven by households at the lower end of the BMI distribution at each income level and quantile regressions do not detect such variation in the BMI-income association. The poverty trap model does not imply a role for ancestral income, conditional on current income, below the threshold, which is at odds with our model and the results in Section 5. Moreover, our instrumental variable estimates, reported in Appendix B1, account for possible reverse causality from BMI to income in any case.

Although it is difficult to come up with an alternative explanation for the discontinuous BMI-income association, it is possible (albeit unlikely) that an unobserved component of nutrient intake, or physical activity, is changing discontinuously at the income level at which we observe the discontinuous increase in BMI. However, any alternative explanation would also need to explain why (i) BMI and metabolic disease do not initially track together with respect to income, (ii) why both outcomes increase discontinuously at the same income level, and (iii) if BMI is the source of forcing variation, why a marginal increase in BMI, starting from a base level below 22 as we will observe, should generate an increase in the risk of metabolic disease. In our model, BMI and the risk of diabetes change simultaneously, at a BMI level that is well within the normal range, because they are independently impacted by the failure of an underlying homeostatic system, which is specific to developing country populations. The tests of external validity in Section 4 and the biological mechanism in Section 5 provide additional support for this interpretation of the results.

3.3 Internal Validity

Our assumption that the body defends its inherited (pre-modern) set point up to a threshold has not been previously verified in developing country populations. Moreover, the model places additional structure on the threshold function in equation (3) by specifying that there is a linear relationship, with slope b, between BMI, z_t, and income, both below and above the threshold, with the relevant income measure switching from y_0 to y_t. The analysis that follows empirically validates the threshold assumption, the specific structure we have imposed on the threshold function in the BMI-income relationship, the distributional assumptions underlying the income generating process, and the analytical approximation that is used to derive closed-form expressions for e^L, e^H.

Given our modeling assumptions, equation (3) implies the following cross-sectional $z_t - y_t$ relationships, below and above the threshold, respectively:

$$\mathbb{E}(z_t|y_t) = a + b(y_t - e^L(y_t))$$

$$\mathbb{E}(z_t|y_t) = a + b(y_t - e^H(y_t)).$$

Expressions for the adjustment terms, $e^L(y_t)$, $e^H(y_t)$, as functions of y_t and the parameters α, $\mu_t \equiv t\mu$, and $\sigma_t^2 \equiv t\sigma^2$ are derived in equations (5) and (6). If the parameter values can be independently

This will change the distribution of current income, but without a set point there will be no discontinuity in the cross-sectional BMI-income association.

obtained, then the appropriate adjustment term can be computed for each y_t. Once the adjustment term is included in the estimating equation, the structural slope parameter, b, can be independently estimated, below and above the income threshold. If the structure we have imposed on the model is empirically valid, the estimated b parameter will be statistically indistinguishable below and above the threshold.

The value of the α parameter can be obtained directly from the estimated location of the threshold in the cross-sectional tests. To determine the value of t, recall from Figure 1 that economic development in India commenced in the middle of the twentieth century. If each generation spans 30 years, then the grandparents of current working-age adults would have been the first generation to experience development; i.e. we are now in generation $t = 3$ of the model. To estimate the parameters of the distribution of income shocks, μ and σ^2, we require data on the income distribution over multiple time periods or generations. The distribution of pre-tax national income is available from the World Inequality Database from 1951 onwards for India (Chancel and Piketty, 2017). Assuming that each generation spans 30 years, as above, we use the (real) income distribution in 1951, 1981, and 2011 and, in particular, the change in these distributions, to estimate the μ and σ parameters.[24]

Table 2 reports coefficient estimates from a piecewise linear equation, using IHDS data, with adult BMI as the outcome. The standard covariates, in addition to household income, are included in each estimating equation. The slope-change in the estimating equation is imposed at the income level where the threshold was previously located. Column 1 reports benchmark estimates without including the $e^L(y_t)$, $e^H(y_t)$ adjustment terms. This specification is essentially the same as what we estimated earlier in Table 1, except that we now report the slopes below and above the threshold (rather than the slope-change). Column 2 reports estimates with the adjustment terms included in the estimating equation. The slope coefficients can now be interpreted as the structural, b, parameter in the model. Although we can easily reject the null hypothesis that the slopes below and above the threshold are equal in Column 1, without the adjustment, we cannot reject the null once the adjustment terms are included. Indeed, the point estimates of the slope coefficient are now remarkably similar, below and above the threshold. A comparison of the point estimates indicates, in addition, that the slope without the adjustment term is less than (greater than) b, below (above) the threshold, as implied by Proposition 1.

Figure 5a examines the sensitivity of the slope coefficient estimates in Table 2, Column 2 to different values of the threshold, α, parameter. We see that the slope coefficients below (above) the specified threshold are increasing (decreasing) in α and coincide just around the value that we assign to that parameter in Table 2 (marked by the vertical lines in Figure 5a). Appendix Figure B6 repeats this exercise for the three remaining parameters of the model: μ, σ, t. As with α, we see that the slope

[24]The World Inequality Database provides the 99 fractiles of the income distribution; p_0p_1, ..., $p_{98}p_{99}$, where p_xp_y refers to the average income between percentiles x and y, in each of the three years. We set the number of dynasties in the economy to be equal to 10,000. We draw 10,000 times from the 1951 income distribution, with each fractile being equally represented, to generate the initial income distribution. For a given value of μ and σ^2 this allows us to simulate the income distribution in 1981 and 2011. Our best estimate of the parameters of the income-shock distribution is the value of μ and σ^2 for which the simulated income distribution in 1981 and 2011 matches most closely with the actual distribution.

Table 2: Piecewise Linear Equation Estimates - with and without adjustment terms

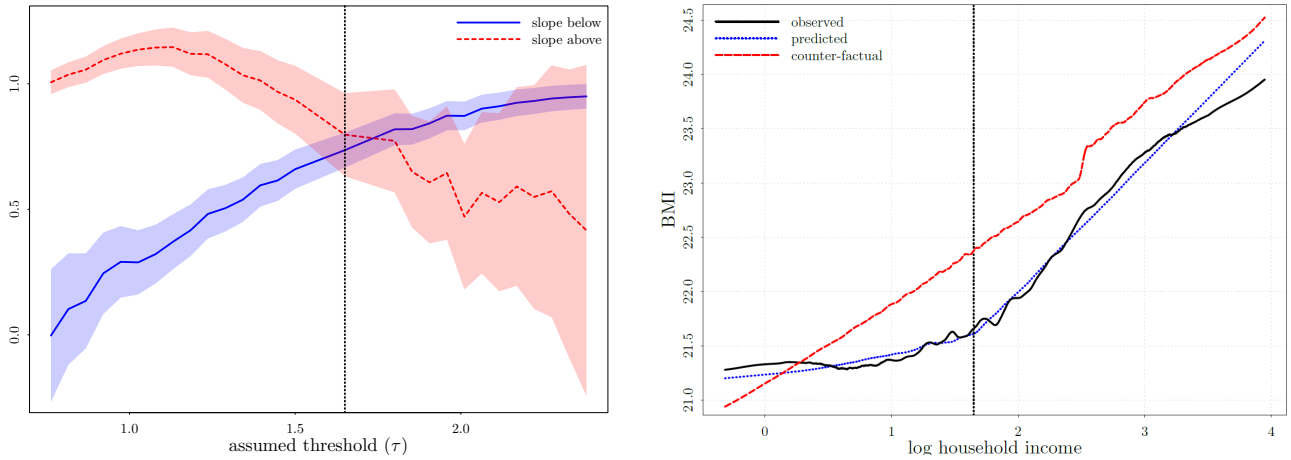
Dep. variable:	BMI	
Specification:	without adjustment (1)	with adjustment (2)
Slope below threshold (β_L)	0.223***	0.735***
	(0.048)	(0.035)
Slope above threshold (β_H)	1.140***	0.797***
	(0.035)	(0.084)
$F-$statistic ($\beta_L = \beta_H$)	234.45	0.45
	[0.000]	[0.502]
Imposed threshold	1.65	1.65
N	76,949	76,949

Source: India Human Development Survey (IHDS)
Logarithm of household income is the independent variable.
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in the estimating equation.
Least squares standard errors are reported in parentheses and $p-$values associated with F-statistic are in square brackets.
$*$ significant at 10%, $**$ at 5% and $***$ at 1%

coefficients coincide just around the values that we assign to the μ, t parameters in Table 2 (the slope coefficients are largely insensitive to the value of σ). These results indicate that all three parameter values need to line up precisely to equalize the slope coefficients in Table 2, which is especially striking given that these values are derived independently from different sources: the value of α is based on the income threshold location estimated with IHDS data, the value of μ is derived from the World Inequality Database, and t is based on the changes in per capita income over many centuries reported in Figure 1.

One benefit of the structural estimation is that it allows us to validate our modeling assumptions. An additional benefit is that it allows us to quantify the consequences of the set point for nutritional status. If the set point is irrelevant, there will be a linear relationship between BMI and household income: $\mathbb{E}(z_t) = a + by_t$. Figure 5b reports the relationship between income and (i) observed BMI, (ii) predicted BMI based on the estimated model, and (iii) counter-factual BMI in the absence of a set point. The standard set of covariates are partialled out, and the dotted vertical line in the figure marks the location of the estimated income threshold. Despite the model's parsimonious structure, and the simplifying assumptions we need to make to estimate its parameters, we see that the model fits the data very well. In our data, 20% of adults are underweight (with a BMI below 18.5). Based on the parameter estimates, the fraction of underweight adults would decline by 24% if the set point

Figure 5: Sensitivity of Slope Coefficients with respect to Parameter Values and Counter-factual Nutritional Status



(a) Threshold (b) BMI.

Source: India Human Development Survey (IHDS)
Panel (a) plots the estimated slope coefficients, below and above the threshold, with respect to the value of the threshold location. The vertical line marks the parameter value (threshold location) that we use for estimation in Table 2. Panel (b) plots the nonparametric relationship between actual, predicted and counter-factual BMI against the logarithm of household income. The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are partialled out prior to nonparametric estimation.

were absent. The observed dampening of the nutritional status-current income relationship below the threshold, which we attribute to a predetermined set point, has important consequences for adult nutritional status in India.

4 External Validity

4.1 Tests With Respect to Income

The next step in the analysis assesses the applicability of the model to other developing countries. To test the cross-sectional implications of the model, the following data are required: (i) Household income, preferably at multiple points in time. (ii) Nutritional status (BMI). (iii) Indicators of metabolic disease. (iv) Individual characteristics and detailed geographical indicators. The additional requirement is that a large sample is needed to locate a slope-change with precision. A search of publicly available data sets from other countries recovered two data sets that are suitable to test our model: the Indonesia Family Life Survey (IFLS) and the Ghana Socioeconomic Panel Survey (GSPS), although the GSPS does not contain information on metabolic disease.[25]

While a set point may be present in other developing countries, the fraction of the population

[25]Other well known data sets that we considered, but were determined to be unsuitable, include the Demographic Health Survey (DHS), the Living Standards Measurement Study (LSMS), Young Lives, and the China Health and Nutrition Survey (CHNS).

that has escaped its pre-modern set point in a given country will depend on the difference between current and historical (pre-modern) income. While roughly half the Indian population has escaped its set point, at what stage in the process of development are Indonesia and Ghana? To answer this question, we proceed to compare current and historical incomes across countries. It is standard practice to use adult height as a proxy for income, and the standard of living, in historical research. We thus use historical adult height to measure historical income.[26] Figure 6a plots the relationship between per capita GDP in 2010 and adult height for individuals born in 1900, which is available for a number of developing countries including India, Indonesia, and Ghana.[27] Figure 6b plots per capita GDP in 2010 and 1960 (the first available year) against adult height for the 1900 birth cohort. The first point to take away from the figures is that historical per capita incomes, measured by adult heights, were higher in Africa.[28] The second point to take away from the figures is that the gap between current income in 2010 and historical income, measured by height in 1900 or even income in 1960, is greater in Asian countries than in African countries. This is also true for the specific countries that we care about, with a larger income-gap in India and Indonesia than in Ghana (where per capita incomes were largely unchanged from 1960 to 2010).

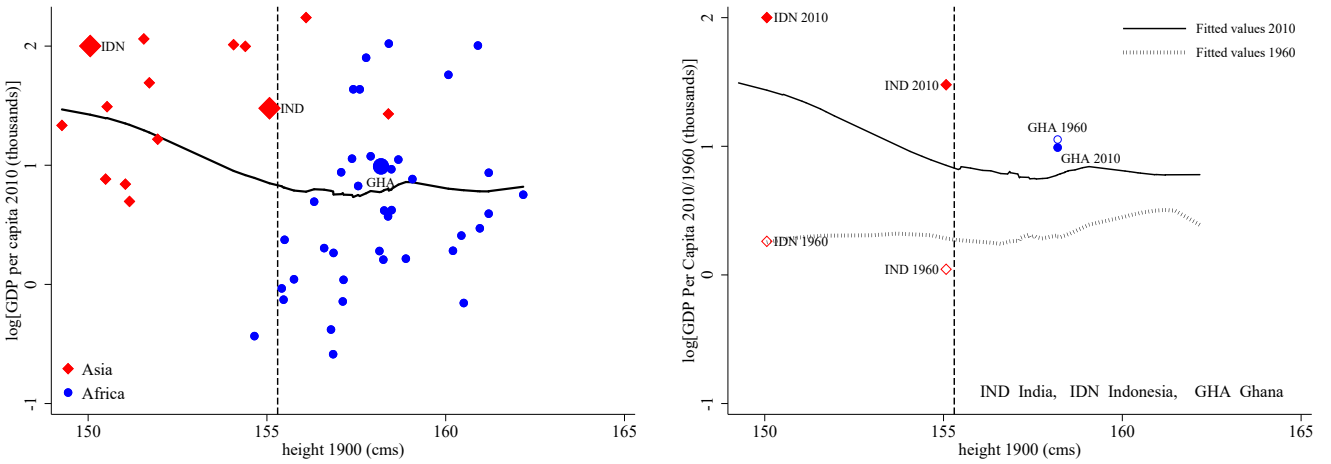
Figure 7a nonparametrically estimates the relationships between adult BMI, the risk of metabolic disease, and household income using Indonesia Family Life Survey (IFLS) data. The same set of covariates that were included in the estimating equation with Indian data are included here as well, except that the district is replaced by the regency and caste is replaced by ethnicity. These covariates are partialled out, using Robinson's procedure, prior to nonparametric estimation. The IFLS has been conducted in five waves. To be consistent with the analysis using IHDS data in 2005 and 2011, the outcomes with IFLS data are measured in the last two (2007 and 2014) waves. However, household income is averaged over all available waves to span as wide a time-window as possible and to smooth out transitory income shocks. The vertical lines in the figure mark the income levels at which Hansen's test locates thresholds for each outcome in Appendix Figure B7a and the shaded areas demarcate the corresponding confidence intervals. The estimated threshold locations are extremely close to each other, with an almost complete overlap in the confidence intervals. Moreover, as documented formally in Appendix Table B4, there is a weak association between household income and each outcome below the estimated threshold and a positive and significant slope-change above the threshold. The gap

[26] As noted by Deaton (2007), genes are important determinants of individual height (and nutritional status more generally) but cannot explain variation across populations. Recall from the model that nutritional status, which we measure by BMI but which also includes stature, is increasing continuously in contemporaneous income in the pre-modern economy. This relationship only weakens in subsequent periods (generations) with economic development on account of the persistent set point.

[27] We include all countries in South and South East Asia and Sub-Saharan Africa that satisfy the following requirement: their GDP per capita must be less than $10,000, which roughly corresponds to the upper bound for lower-middle income countries set by the World Bank. The same criterion is applied in the cross-country analysis below.

[28] This is consistent with archeological evidence that South Asian populations had relatively low nutritional status (Pomeroy et al., 2019), although we make a broader Asia-Africa comparison. Notice in Figure 6b that there is no apparent relationship between 1960 income and 1900 height across countries, in contrast with the negative relationship that is observed with 2010 income. Given the change in the slope over time, we expect that the sign would have reversed – turned positive – if cross-country income data were available a few decades prior to 1960, consistent with the assumption that historical heights and incomes are positively correlated.

Figure 6: Current and Historical Income Across Countries



(a) GDP 2010.

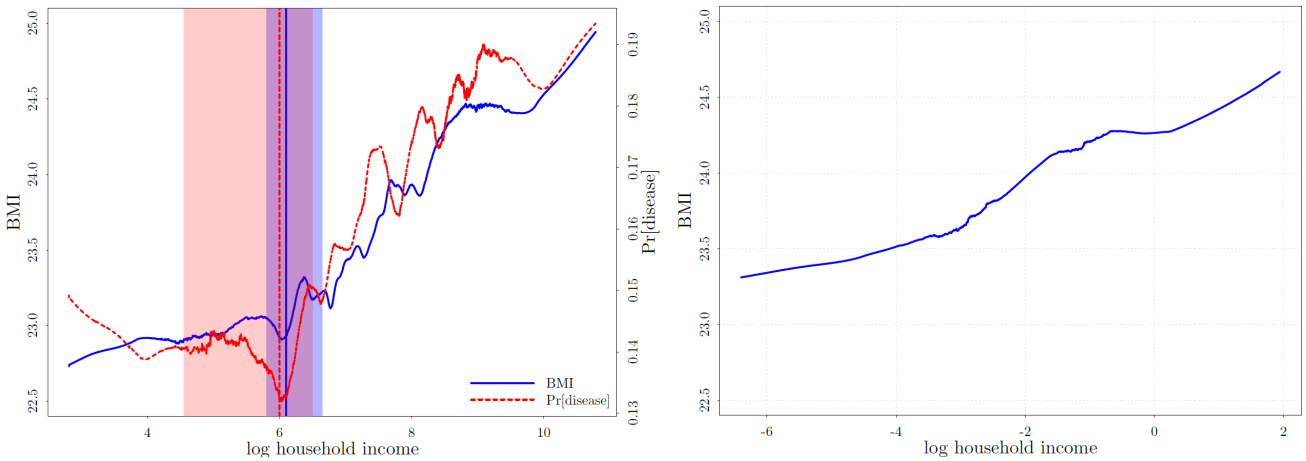
(b) GDP 1960 and 2010.

Source: NCD-RisC and Penn World Table 9.0
Historical income is measured by height in the 1900 birth cohort.

between current and historical income is even greater in Indonesia than in India in Figure 6. We would thus expect a larger fraction of the population to have escaped its set point in Indonesia and, based on our estimates of the threshold location with respect to the income distribution, it appears that three-quarters of the Indonesian population has indeed crossed the threshold.

Figure 7b reports the nonparametric relationship between adult BMI and household income, using data from the Ghana Socioeconomic Panel Survey (GSPS). As noted, the GSPS does not collect data on metabolic disease. However, the full set of covariates that were used in the Indian and Indonesian analyses are available, with tribal affiliation replacing caste category and ethnicity, respectively. These covariates are partialled out prior to nonparametric estimation, as usual. The GSPS was conducted in three waves; 2009-2010, 2013, and 2017. The outcomes are measured in the 2009-2010 and 2013 waves, which correspond most closely to the IHDS waves, while household income is averaged over all three waves. In contrast with the discontinuous relationships that we estimated with Indian and Indonesian data, nutritional status is increasing smoothly with income in Figure 7b. Formal statistical support for this observation is provided in Appendix Figure B7b, where the Hansen test is unable to detect an income threshold. As reported in Appendix Table B4, there is a positive and statistically significant association between adult BMI and household income in Ghana. Where the Ghana data differ from the Indian and Indonesian data is that there is no slope change. Our interpretation of this finding, which is in line with the observation that current and historical incomes are relatively close in Africa (and in Ghana) is that the bulk of the Ghanaian population remains at its pre-modern set point. Additional support for this interpretation is provided below.

Figure 7: Nutritional Status and Metabolic Disease with respect to Income (Indonesia and Ghana)



(a) Nonparametric relationships (Indonesia). (b) Nonparametric relationship (Ghana)

Source: Indonesia Family Life Survey (IFLS), Ghana Socioeconomic Panel Survey (GSPS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, ethnicity (Indonesia) or tribe (Ghana), rural area, regency (indonesia) or district (Ghana), and survey-round are partialled out prior to nonparametric estimation.
The vertical line marks the threshold location and the shaded region demarcates the cluster bootstrapped confidence interval.

4.2 Tests With Respect to BMI

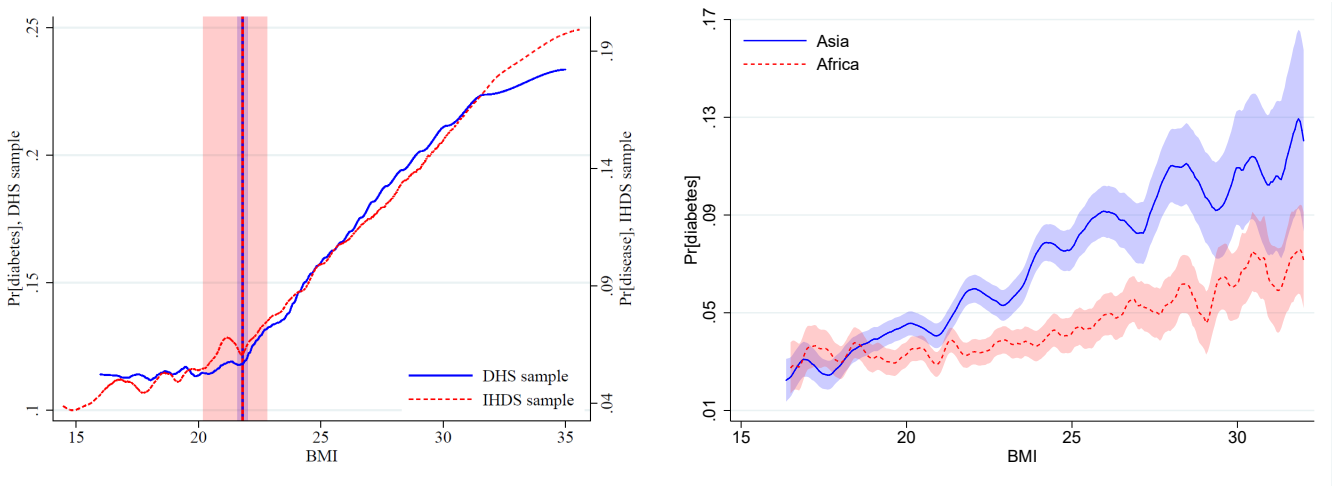
Although few surveys collect information on income, BMI and metabolic disease, many health-focussed surveys have collected information on BMI and specific diseases, including diabetes (with biomarkers). These include recent rounds of the DHS (India 2015-16, Bangladesh 2011 and 2017-18, Namibia 2013), WHO-STEPS surveys (16 African countries and 5 Asian countries) and the 2014 round of the IFLS. These data allow us to test an additional implication of the model derived in Section 2.4, which is that there is no association between the risk of diabetes and BMI up to a threshold BMI, and a positive association thereafter.[29]

Our tests of the diabetes-BMI association begin with data from India. The IHDS, which we use for the core analysis, includes 150,000 observations on metabolic disease (diabetes, hypertension, cardiovascular disease) over two rounds and the 2015-16 round of the DHS includes diabetes information (with biomarkers) for as many as 770,000 adults. Nonparametric estimates of the association between metabolic disease or diabetes and BMI are reported in Figure 8a, after partialling out the additional covariates in the estimating equation as usual. We locate a threshold at precisely the same BMI – 21.8 – with both datasets.[30] It has been recommended, without supporting statistical evidence, that

[29]Other obesity indicators; e.g. waist circumference, waist-hip ratio have also been associated with diabetes. However, these indicators are highly correlated and meta-analyses indicate that the three indicators have similar associations with diabetes (Vazquez et al., 2007).

[30]Matching the figure, we estimate a positive and significant slope change above the estimated threshold in Appendix Table B5. However, we also estimate a positive and significant (albeit small in magnitude) coefficient below the threshold, which is not implied by the model. This positive association is also observed in the inter-regional analysis

Figure 8: Reported Metabolic Disease, Measured Diabetes, and BMI



(a) India

(b) South/South-East Asia and Sub-Saharan Africa

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS), Demographic and Health Survey (DHS), WHO STEPS

For panel (a), the standard set of covariates: age (linear, quadratic and cubic terms), and dummies for gender, caste group, rural area, and survey-round for IHDS data are partialled out prior to the nonparametric estimation. For panel (b), age (linear, quadratic and cubic terms), and dummies for gender, country, and survey year are partialled out prior to the nonparametric estimation.

the lower bound for the overweight range in Asian populations be reduced from 25 to 23, to account for the fact that these populations are at elevated risk of diabetes at lower BMI (Deurenberg-Yap et al., 2002; Pan et al., 2004). Our estimates indicate that the risk of diabetes increases discontinuously at an even lower threshold and to put this difference in perspective, we note that 11% of adult Indians have BMI's between 21.8 and 23 (based on the DHS sample).

As discussed, diabetes with IHDS (and IFLS) data is measured by a composite variable, which indicates whether a given individual has been diagnosed with diabetes or with highly correlated comorbidities (hypertension, cardiovascular disease). In contrast, the analysis with DHS data is based on objective biomarkers (blood sugar levels exceeding 125 mg/dL) for diabetes. Although reported levels are higher with DHS data in Figure 8a, on account of the under-counting with self reported data, notice that the two measures track closely across the range of BMI's once the levels are adjusted, validating the composite measure of diabetes that we have used thus far in the analysis.

The discontinuity that we detect in Figure 8a arises because there is a discontinuous increase in both BMI and the risk of diabetes at a threshold income level. Lean mass, a component of BMI that is emphasized by Pomeroy et al. (2019), would also likely increase at that level, but this would have a dampening effect on the risk of diabetes (conditional on BMI). In contrast, the inter-regional comparisons of the diabetes-BMI association that follow could potentially be independently explained by variation in lean mass (conditional on BMI) across regions. We do not attempt to disentangle

that follows in Figure 8b. One explanation for the positive association, in line with the conventional view, is that an increase in BMI has a direct effect on the risk of diabetes.

these channels; our objective with the inter-regional comparisons is to verify that they are in accord with the implications of the model, while allowing for co-existing mechanisms.

Sample sizes for the diabetes-BMI analysis with Indian data, from the IHDS and DHS, are an order of magnitude larger than what are available for other countries from the DHS, WHO-STEPS, and IFLS (with biomarkers), ranging from 2,000 to 9,500 observations. Not surprisingly, we cannot locate a slope discontinuity with statistical confidence separately by country with these datasets and, hence, we proceed to pool individual-level data into two regions: Asia and Africa. Based on the inter-regional income dynamics reported in Figure 6 and the results for India in Figure 8a, we infer that African populations are largely at their set points, whereas Asians will start to escape when their BMI's cross a relatively low threshold. Once we pool Asian countries with different thresholds, we do not expect to observe a discontinuity, as in Figure 8a. The expectation, instead, is that diabetes prevalence will be increasing with BMI in both regions, with a divergence at higher BMI levels as an increasing fraction of Asian populations escape their set points. This is indeed what we observe in Figure 8b and the same broad cross-regional patterns are observed when the diabetes-BMI associations are reported country by country in Appendix Figure B8.[31]

4.3 Cross-Country Analysis

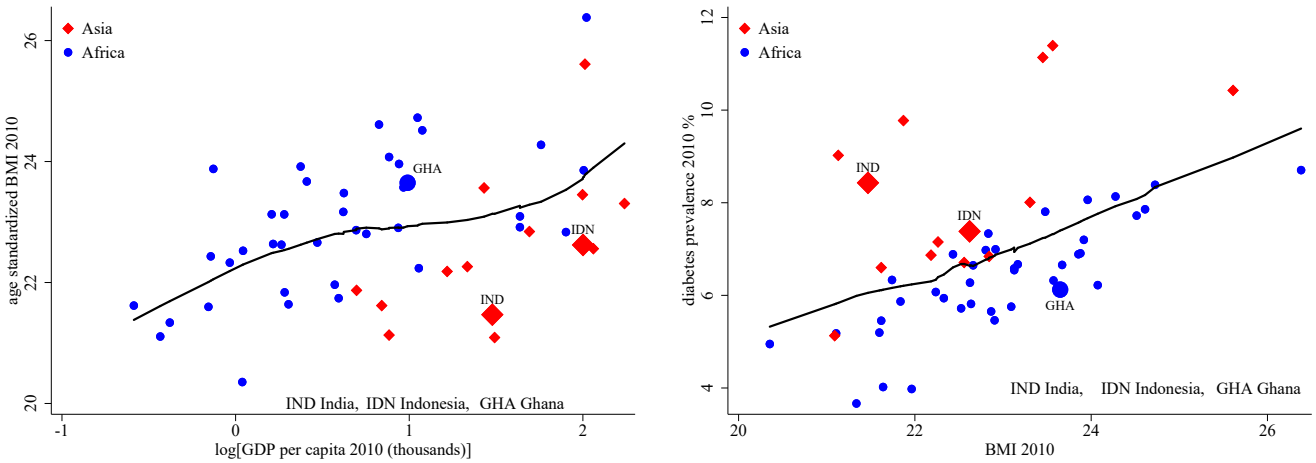
We complete the tests of external validity by shifting the analysis from the individual level to the country level. The nutrition-income puzzle that Deaton (2007) uncovered is that nutritional status, which he measures by height, is lower in South Asia than what would be predicted by GDP per capita, whereas the reverse is true for Africa. Our model, adapted to a cross-country setting with aggregate data (and a country-specific set point) can generate the same fact, but with BMI as the measure of nutritional status. Consider an Asian and an African country with the same current income (per capita GDP). Looking back at Figure 6, the African country will have higher historical income, which determines the set point. Assuming, as above, that the population in the African country is largely at its set point, its BMI will be determined by historical income. In contrast, BMI in the Asian population, which is on both sides of the set point, will be a weighted average of (lower) historical income and (higher) current income. If a sufficiently large fraction of the Asian population remains at its set point, then its BMI will be lower than the African population with the same current income. This is what we observe in Figure 9a using cross-country data from NCD-RisC. Focusing on the income region in which both Asian and African countries are represented, Asian BMI's lie disproportionately below the trend line.

Although other mechanisms have been proposed to explain the weak association between nutritional status and income in developing countries, an appealing feature of our mechanism, based on a

[31] Geographic indicators at the district level and ethnic affiliation are unavailable in the WHO-STEPS surveys. We thus partial out a reduced set of covariates – age (linear, quadratic and cubic terms) and dummies for gender, country and survey year – using the Robinson procedure, prior to nonparametric estimation in Figure 8b. Only STEPS countries that lie entirely south of the Sahel are included in Africa and two outlying Asian countries with implausibly low diabetes rates – Vietnam (2.3%) and Cambodia (1.8%) – are excluded from the sample. Depending on the survey, diabetes is defined by a blood sugar level exceeding 125 mg/dL or 7 mmol/L.

historically determined set point is that it also has implications for the emergence of diabetes. The micro evidence reported in Figure 8b indicates that diabetes rates are (weakly) higher in Asian populations than in African populations at each BMI level. Moving up to the country level, this implies that diabetes prevalence should be higher in Asia than in Africa, conditional on BMI, and this is what we observe in Figure 9b, with Asian countries clustered (almost without exception) above the trend line. Notice that while India is somewhat of an outlier in the figure, other Asian countries are even bigger outliers and not all of them are South Asian. Although the diabetes literature has tended to focus on South Asians as a particularly vulnerable group, our analysis, as with the analysis of the BMI-income association, indicates that inter-regional differences in diabetes prevalence extend to the Asian continent as a whole.

Figure 9: BMI - Current Income and Diabetes - BMI Relationship Across Countries



(a) BMI - current income

(b) Diabetes - BMI

Source: NCD-RisC and Penn World Table 9.0

5 The Mechanism

Two biological relationships serve as the starting point for our model: (a) BMI is determined by ancestral income below a threshold and by current income above the threshold. (b) The risk of diabetes is constant below the threshold and increasing in the difference between current and ancestral income above the threshold.[32] We next proceed to validate these relationships by constructing exogenous measures of ancestral income. The threshold location for this exercise is derived from the cross-sectional tests of the model; recall that households below the current income threshold remain at their

[32]Bringing the two relationships together, the risk of diabetes is increasing in the difference between current BMI and ancestral BMI. This is precisely the model proposed by Wells et al. (2016), with current BMI measuring the "metabolic load" and ancestral BMI measuring "metabolic capacity." We do not test this model because reliable measures of ancestral BMI are unavailable. Wells et al. use height as a proxy for metabolic capacity, but this is a measure of current nutritional status, together with BMI, in our analysis.

set point.[33] As with the cross-sectional tests, we focus on India in the analysis that follows, but verify that the results hold up with Indonesian data (with which a threshold can also be located).

An appealing feature of the cross-sectional tests of the model is that they do not require knowledge of the set point, y_0. This allowed us to include rural populations and urban populations (which include a large share of relatively recent migrants) in the analysis. When testing the biological relationships, however, we will need to link current income, y_t, to pre-modern ancestral income, y_0, and hence the tests of the mechanism are restricted to rural households who would have remained in their place of residence for many generations. Measures of ancestral income are unavailable at the family (dynasty) level. We thus construct measures of per household ancestral income at the district level and at the village level in the analysis that follows.

5.1 District-Level Evidence

Our first measure of y_0 is constructed at the district level and is based on historical food supply. Agriculture was the dominant activity in the pre-modern economy and aggregate wealth would thus have been determined by crop productivity. Galor and Özak (2016) convert potential crop yields, obtained from the Food and Agriculture Organization Global Agro-Ecological Zones (FAO-GAEZ) project, to caloric production and then average across crops to construct a Caloric Suitability Index (CSI) which they document is a good indicator of the historical level of economic development or, equivalently, aggregate wealth across countries. We use the same index to measure pre-modern wealth at the district level, except that the baseline specification restricts attention to two staple crops – wheat and rice – that dominated historical agricultural production (and continue to account for a large share of agricultural production) in India. If the CSI is a good measure of pre-modern aggregate wealth, then it should be closely related to historical population density (Diamond, 1997). Appendix Figure B9a verifies this hypothesis at the district level by estimating a positive association between population density in 1951, when the Indian economy was just starting to develop, and CSI.[34]

While Appendix Figure B9a provides empirical support for our measure of historical aggregate wealth, it also indicates that the positive relationship between population and CSI must be accounted for when constructing measures of ancestral per household income. We do this by specifying that ancestral per household income is a flexible function of the CSI, $f(CSI)$. We then estimate the following equation:

$$y_t = f(CSI) + \epsilon_t, \tag{8}$$

where y_t is current household income, which is obtained as in the cross-sectional tests from the India Human Development Survey (IHDS), and CSI is based on the household's district of residence (the

[33] Based on the model, some households above the current income threshold will also be at their set point. In a rapidly growing economy, however, most households above the threshold will have escaped their set point.

[34] State fixed effects are partialled out in Appendix Figures B9a and B9b and are included in the estimating equations that follow to account for independent state-level determinants of historical population density, current household income, and the outcomes of interest (BMI and metabolic disease). For the analysis with Indonesian data, state fixed effects are replaced by regency fixed effects.

IHDS does not provide location identifiers below the district level). Equation (8) can be compared with the income equation (1) in the model:

$$y_t = y_0 + U_t.$$

Predicted income in equation (8) corresponds to ancestral income, y_0, and the residual in that estimating equation corresponds to the income mismatch, $U_t \equiv y_t - y_0$.[35] The objective when specifying the $f(CSI)$ function is to capture that part of the variation in current income that is explained by historical conditions and, by extension, ancestral per household income. Our preferred measure of y_0 will thus be predicted household income based on the most flexible nonparametric specification of the $f(CSI)$ function. In our data we document a nonmonotonic relationship (reasonably approximated by a quadratic function) between predicted household income and CSI in Appendix Figure B9b.[36]

Table 3 reports the relationship between BMI and both ancestral income, y_0, and current income, y_t, below and above the estimated threshold. y_0 and y_t are normalized, by dividing by their respective standard deviations, to allow the magnitude of the income coefficients to be comparable. The standard set of covariates, with state fixed effects instead of district fixed effects since y_0 is measured at the district level, and with the exception of the rural dummy since this is now a rural sample, are included in the estimating equations. As observed in Columns 1-2 with IHDS data, ancestral income has a positive and significant effect on BMI below the threshold (where households are at their set point) but not above it. Although the current income coefficient is also significant below the threshold, it is substantially smaller than the ancestral income coefficient and, moreover, is four times larger above the threshold.

Table 3, Columns 3-4, reports the BMI-income relationship with Indonesian (IFLS) data. The analysis proceeds in exactly the same way as above, except that we restrict attention to a single staple crop – rice – which is by far the dominant crop in Indonesia. While a long history of internal migration in Indonesia could potentially weaken the relationship between our measure of ancestral income, which is based on the current place of residence, and nutritional status, the compensating advantage of the IFLS data is that they provide the sub-regency (sub-district) in which the household resides. The CSI can thus be constructed at a more disaggregate level than is possible with IHDS data. We see

[35] The residual, ϵ_t, is mean-zero by construction, whereas U_t has positive mean μ_t. Our estimates of y_0 and U_t are thus only identified up to a constant, but this has no bearing on the analysis that follows. Appendix Figure B10a uses binned scatter plots to (separately) describe the relationships between household income, y_t, and our measures of y_0 and U_t. These relationships are linear, matching the structure of the income equation (1) in the model. Note that failure of the separability assumption in equation (8), which allows us to construct measures of y_0 and U_t, would lead to false rejection of the model and not the converse.

[36] Spatial heterogeneity in y_0 is not inconsistent with the Malthusian model. Ashraf and Galor (2011) show theoretically that steady-state pre-modern per capita income would have varied with the predisposition towards having children and the cost of child rearing. Such heterogeneity in fertility could have varied with agricultural productivity. For example, Diamond (1997) argues that greater agricultural productivity in the pre-modern period was associated with higher population densities and with more complex (vertically stratified) societies. Fertility would have varied by social class in such societies, with the elites consuming above subsistence. Once social stratification and associated fertility regulation is incorporated in the Malthusian model, average per capita income (food consumption) will vary with agricultural productivity, but in a way that is theoretically ambiguous.

Table 3: Nutritional Status - Income Relationship (below and above the threshold)

Dependent variable:	BMI			
Country:	India		Indonesia	
Sample:	Below	Above	Below	Above
Ancestral income	0.899***	0.165	1.059***	0.464
	(0.243)	(0.283)	(0.254)	(0.337)
Current income	0.185***	0.852***	-0.048	0.591***
	(0.040)	(0.047)	(0.119)	(0.064)
Threshold location	1.65	1.65	6.1	6.1
Dep. var. mean	20.482	21.851	22.317	23.021
N	27,164	20,296	3,182	10,610

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia) and survey-round are included in the estimating equation.The rural-urban dummy is excluded, since the sample is restricted to rural households.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

that the results in Columns 3-4 match closely with the biological relationships specified in the model. Ancestral income has a positive and significant effect on adult BMI below but not above the estimated threshold, whereas the converse is true for current income.

Current income in equation (8) can be decomposed into two orthogonal components: ancestral income, y_0, which is measured by predicted income and the income mismatch, $U_t \equiv y_t - y_0$, which is measured by the residual in that equation. Table 4 reports the relationship between the risk of metabolic disease and (separately) each income component, below and above the threshold (τ). Results with Indian (IHDS) data are presented in Columns 1-2 and with Indonesian (IFLS) data in Columns 3-4. As specified in the model, the (uninteracted) income mismatch coefficient, which reflects the association with the risk of metabolic disease below the threshold, is economically and statistically insignificant in Columns 1 and 3. In contrast, the interaction coefficient, reflecting the change in the association above the threshold, is positive and significant in both columns. Moreover, the ancestral income coefficients in Columns 2 and 4 are insignificant, with one exception (the uninteracted coefficient with Indian data in Column 2). Summarizing the estimation results and in line with the specification of the metabolic disease -income relationship implied by the model, we observe that the uninteracted and interacted coefficients are jointly significant in Columns 1 and 3, which measure the association between metabolic disease and the income mismatch, but jointly insignificant in Columns 2 and 4, where we measure the corresponding association with ancestral income.[37]

We complete the district-level analysis in Appendix Tables B6-B9 by verifying the robustness of

[37] Although ancestral income does not determine the risk of diabetes in our model, initial conditions and the mismatch will jointly determine outcomes in other models of developmental plasticity (Malani et al., 2022). These factors can be examined independently in our analysis because they are orthogonal by construction.

Table 4: Metabolic Disease - Income Relationship

Dependent variable:	Pr(metabolic disease)			
Country:	India		Indonesia	
Income component:	income mismatch (1)	ancestral income (2)	income mismatch (3)	ancestral income (4)
Income component	0.001	0.012*	-0.004	-0.011
	(0.002)	(0.006)	(0.011)	(0.019)
Income component × $\mathbf{1\{}$current income $> \boldsymbol{\tau}\mathbf{\}}$	0.018***	-0.002	0.032**	0.001
	(0.004)	(0.002)	(0.011)	(0.008)
Joint signficance				
$F-$statistic $[p-$value$]$	14.983	1.889	13.811	0.170
	[0.000]	[0.153]	[0.000]	[0.844]
Threshold location (τ)	1.90	1.90	6.00	6.00
Dep. var. mean	0.054	0.054	0.162	0.162
N	90,879	90,879	11,001	11,001

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia) and survey-round are included in the estimating equation.The rural-urban dummy is excluded, since the sample is restricted to rural households.
$F-$statistic measures the joint significance of the uninteracted and interacted income component coefficients.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

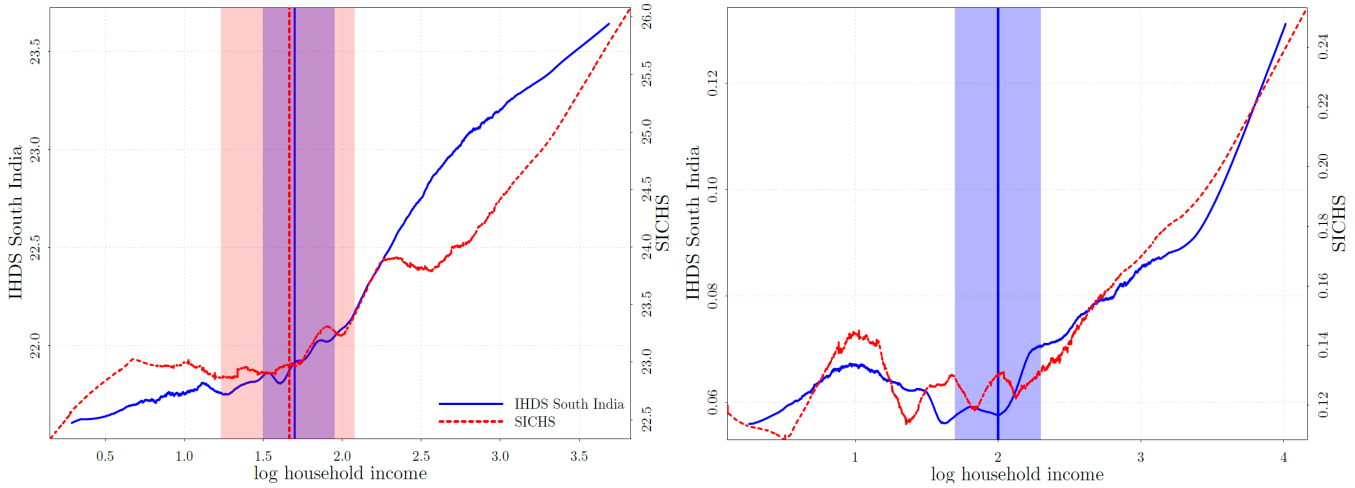
the results to (i) a less flexible quadratic specification of the $f(CSI)$ function, and (ii) to construction of the CSI with an expanded set of major crops; wheat, rice, barley, sorghum, rye and millet for India and rice, sorghum, cassava and maize for Indonesia.

5.2 Village-Level Evidence

The district-level measures of ancestral income, y_0, allow us to validate both biological relationships specified in the model. The advantage of these measures is that they can be constructed, in a consistent fashion, using nationally representative data from India and Indonesia. However, the district and the sub-regency are aggregate spatial units. Moreover, while we use low technology-rainfed agriculture to construct the CSI, as do Galor and Özak (2016), this measure is not based directly on pre-modern income. We improve on both of these dimensions by using data from the South India Community Health Study (SICHS), which we have collected, for the analysis that follows.

The SICHS covers a rural population of 1.1 million individuals residing in Vellore district in the South Indian state of Tamil Nadu. Two components of the SICHS are relevant for our analysis: a census of all 298,000 households residing in the study area, completed in 2014, and a detailed survey of 5,000 representative households, completed in 2016. The SICHS census collected each

Figure 10: Nutritional Status and Metabolic Disease with respect to Income (IHDS and SICHS)



(a) BMI (b) Metabolic disease

Source: India Human Development Survey (IHDS), South India Community Health Study (SICHS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, and (for IHDS) rural area, district and survey-round are partialled out prior to nonparametric estimation.
The vertical lines mark the estimated threshold location and the shaded areas demarcate the corresponding 95% confidence intervals.

household's income in the preceding year. The SICHS survey collected information on the marriage of the household head and his spouse, and their parents, and in addition covers all variables included in the analysis using IHDS and IFLS data above. More importantly, the SICHS data are supplemented with historical records, obtained from the British Library in London, on the agricultural revenue tax per acre of cultivated land that was collected from each village in the Northern Tamil Nadu region (encompassing the study area) in 1871.[38] The revenue tax was based on a detailed assessment, made by the colonial government, of crop suitability, soil quality, precipitation, and other growing conditions. Like the CSI, this is a measure of potential agricultural productivity, but it is (i) defined at the village level, (ii) based explicitly on pre-modern growing conditions, and (iii) provides a direct measure of pre-modern income; i.e. the monetary value of agricultural output.

We begin the analysis with SICHS data by establishing that the cross-sectional relationships estimated above with nationally representative IHDS data are obtained in the study area as well. Figure 10a reports the association between the BMI of the household head and his spouse and current household income. To smooth out transitory shocks, we take the average of the household income reported in the SICHS census and the SICHS survey as our measure of permanent household income. The standard set of covariates, excluding the district dummies and the rural dummy since the rural sample is drawn from a single district, are partialled out prior to nonparametric estimation. The SICHS

[38]There are 377 *panchayats* or village governments in the SICHS study area. These *panchayats* were historically single villages, which over time sometimes divided or added new habitations. The *panchayat* as a whole, which often consists of multiple modern villages, can thus be linked back to a single historical village. What we refer to as a "village" in the discussion that follows is thus a historical village or, equivalently, a modern *panchayat*.

study area was purposefully selected to be representative of rural South India, defined as in Munshi and Rosenzweig (2016) by the states of Tamil Nadu, Andhra Pradesh, Karnataka, and Maharashtra, with respect to socioeconomic and demographic characteristics.[39] As a basis for comparison, we thus report the corresponding nonparametric plot obtained with IHDS data, for the South Indian states, in Figure 10a. We go through the same steps as above to plot the relationship between the risk of metabolic disease and current income, with SICHS and IHDS South India data, in Figure 10b.

The estimated relationships, with SICHS and IHDS South India data, match very closely across the income distribution in both figures.[40] The vertical lines mark the spot where Hansen's test (shown in Appendix Figure B11) locates an income threshold, with the shaded area demarcating the associated 95% confidence interval. The threshold locations with adult BMI as the outcome are precisely estimated and almost identical with the two data sets. With the risk of metabolic disease as the outcome, in contrast, a threshold is precisely estimated with IHDS South India data but not SICHS data.[41] The tests that follow will thus be restricted to the BMI-income relationship.

One advantage of the SICHS analysis is that ancestral income can be measured at the village level. However, this creates a new complication because ancestors can be drawn from multiple villages. Epigenetic inheritance was traditionally assumed to occur along the female line; i.e. via the mother, although recent evidence indicates that paternal traits can also be transmitted epigenetically (Jablonka and Raz, 2009; Lind and Spagopoulou, 2018). We allow for both possibilities, in which case ancestral incomes along the male and female line are relevant. Marriage in India is patrilocal, with women often leaving their natal (birth) village when they marry. In a patrilocal society, men do not move when they marry and, hence, ancestral income along the male line is determined by historical income in the individual's natal village. Ancestral income along the female line, in contrast, will be determined by historical income in the (possibly) many different villages from which the female ancestors were drawn.

To construct a single measure of ancestral income, we take advantage of the fact that families in rural India match assortatively on wealth (permanent income) in the marriage market, as documented with SICHS data by Borker et al. (2021). Although ancestral income, y_0, will not match perfectly on the male and female side in any marriage on account of the $U_t \equiv y_t - y_0$ term in the income equation, it will still be highly correlated for husbands and wives. We verify that this is the case, with SICHS data,

[39] Borker et al. (2021) provide a detailed description of the study area, documenting that it is representative of rural Tamil Nadu and rural South India with respect to socioeconomic and demographic characteristics; e.g. age distribution, marriage patterns, literacy rates, labor force participation, child and adult sex ratios, and religious composition.

[40] BMI and the risk of metabolic disease are systematically higher with SICHS data relative to IHDS South India data (this can be observed by comparing the range of the Y-axes in Figure 10). In line with this finding, Alacevich and Tarozzi (2017) document that average heights for children under 5 are lower in the IHDS than in the Demographic Health Survey (DHS). They also document that heights and weight are measured with error in the IHDS, with heaping at particular focal points. Once we control for the level, however, the SICHS and the IHDS South India data track very closely with household income.

[41] This is because the sample size is much smaller with SICHS data and the threshold location is more difficult to estimate with the risk of metabolic disease as the outcome. For those outcomes for which thresholds can be located in Figure 10, the piecewise linear equation estimates at the estimated thresholds are reported in Appendix Table B10. In line with previous results, we cannot reject the hypothesis with South Indian (IHDS) data that the thresholds with BMI and metabolic disease as outcomes are located at the same income level.

for the household head and his spouse in Appendix Figure B12a and for their parents in Appendix Figure B12b, using the 1871 tax revenue in each individual's natal village to measure y_0.[42] The strong correlation in ancestral income that we document does not arise mechanically because couples are drawn from the same natal village. 80% of women in the SICHS study area leave their natal village when they marry, although almost all of them marry within the district, and we expect that similarly strong correlations in ancestral incomes would be observed if data from earlier generations were available. This implies that the 1871 tax revenue in any village from which ancestors were drawn could be used to construct y_0. To be consistent with our measure of current income, we use 1871 tax revenue in the current village of residence, both for the household head and his spouse, to construct their ancestral income.

The tax revenue per acre of cultivated land in 1871 measures historical wealth at the level of the village. As with the construction of the district level measure of ancestral per household income, we allow for an endogenous (village level) population response by specifying that per household ancestral income, y_0, is a flexible function, $g(R)$, of the 1871 tax revenue, R. The analysis thus proceeds in two steps: First, we estimate the relationship between current household income, y_t, and $g(R)$; the predicted income provides us with a measure of y_0, following the same argument as above. Second, we estimate the relationship between BMI and both y_0 and y_t, below and above the threshold located in Figure 10. As seen in Table 5, the ancestral income coefficient is positive and significant below, but not above, the threshold. In contrast, the current income coefficient is positive and significant above, but not below, the threshold. This result is robust to alternative (nonparametric and quadratic) specifications of the $g(R)$ function.

We close this section by considering alternative explanations for the results in Table 5. The statistical challenge when testing the mechanism underlying the model is that the set point is determined by fixed pre-modern conditions. Even if these conditions are exogenously determined, they could still be associated with factors that independently determine the outcomes of interest. For example, village-level tax revenue in 1871, which we use to construct ancestral income, is also associated with pre-modern aggregate wealth and contemporaneous levels of economic development. These historical economic conditions could potentially determine nutritional status today through a variety of channels. A second alternative argument posits that ancestral income proxies for poorly measured current income at low income levels. However, any alternative mechanism must explain the additional restrictions imposed by our model: ancestral income should only be relevant below the estimated threshold and current income should only be relevant above the threshold. These sharp discontinuities cannot be explained by historical development or by measurement error, and they continue to hold in Appendix Table B11 even when ancestral income and current income are included separately in piecewise linear equations. The ancestral income coefficient continues to be positive and statistically significant below (but not above) the threshold, whereas the converse is true with current income.[43]

[42] The household's ancestral income, y_0, is specified as a continuous function of the 1871 village-level tax revenue below. However, this has no bearing on the analysis of assortative matching.

[43] We include ancestral income and current income, above and below the threshold, in Tables 3 and 5 to avoid the possibility that one variable simply proxies for the other. This is because ancestral income and current income are

Table 5: SICHS Nutritional Status - Income Relationship (below and above the threshold)

Dependent variable:	BMI			
$g(R)$ specification:	nonparametric		quadratic	
Sample:	below	above	below	above
	(1)	(2)	(3)	(4)
Ancestral income	0.334***	0.170	0.375***	0.026
	(0.124)	(0.150)	(0.128)	(0.123)
Current income	0.012	0.834***	0.048	0.834***
	(0.190)	(0.119)	(0.191)	(0.120)
Threshold location	1.69	1.69	1.69	1.69
Dependent var. mean	23.033	23.755	23.033	23.755
N	1810	3844	1810	3844

Source: South India Community Health Study (SICHS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender and caste group are included in the estimating equation. The rural-urban dummy and district dummies are excluded, since the rural sample is drawn from a single district.
Bootstrapped standard errors, clustered at the level of the village, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

6 Conclusion

This research helps explain two seemingly unrelated facts that have been documented in developing countries: (i) the weak association between nutritional status, which we measure by BMI, and income, and (ii) the elevated risk of diabetes among normal weight individuals. Our explanation is based on a set point for BMI that is adapted to food supply in the pre-modern economy, but which fails to subsequently adjust to rapid economic change. The postulated mechanism is validated with micro data from many countries and can jointly explain inter-regional (Asia versus Africa) differences in nutritional status and the prevalence of diabetes.

Our structural estimates and accompanying counter-factual simulations indicate that the fraction of underweight adults in India, who comprise 20% of the population, would decline by 24% in the absence of a set point. At the same time, half the adult population who remain at their set point are protected from diabetes. While the health consequences of the set point are thus ambiguous, what is incontrovertible is the fact that in the coming decades an increasing fraction of the Indian population, and the adult population in other developing countries, will escape their pre-modern set points. Screening will be an important component of public health programs that attempt to address the resulting increase in diabetes. Our analysis, which documents a discontinuous increase in the risk of diabetes at a BMI below 22 in India, indicates that much of the adult population will need to be screened for this condition.

While the cost of screening may be greater than currently envisaged, the flip-side of this finding

correlated by construction.

is that many individuals detected with diabetes will have relatively low BMI's. A natural question to ask is how these lean diabetics should be treated. The recent medical literature has shifted focus away from BMI towards more proximate risk factors for diabetes that are common in developing-country populations, such as low insulin secretion and ectopic fat deposition (Narayan and Kanaya, 2020). But in our view, these factors are symptomatic of an underlying homeostatic-system failure, and efforts to reverse diabetes would be better served by correcting that fundamental failure. Evidence on diabetes reversal through a weight-loss program in the U.K. indicates that there is an individual-specific BMI threshold, which is independent of initial BMI, below which diabetes is reversed (Taylor and Holman, 2015). In a developing-country population, we expect that this threshold will be associated with the pre-modern set point, which for many (lean) diabetics will not be far from their existing BMI. This suggests a promising approach to diabetes control in such a population, involving relatively little weight loss, which we plan to explore in future research.

References

Alacevich, C. and A. Tarozzi (2017): "Child Height and Intergenerational Transmission of Health: Evidence from Ethnic Indians in England," *Economics & Human Biology*, 25, 65–84.

Ashraf, Q. and O. Galor (2011): "Dynamics and Stagnation in the Malthusian Epoch," *American Economic Review*, 101, 2003–41.

Atkin, D. (2013): "Trade, Tastes, and Nutrition in India," *American Economic Review*, 103, 1629–63.

——— (2016): "The Caloric Costs of Culture: Evidence From Indian Migrants," *American Economic Review*, 106, 1144–81.

Banerjee, A. V. and A. F. Newman (1993): "Occupational Choice and the Process of Development," *Journal of Political Economy*, 101, 274–298.

Barker, D. J. (1995): "Fetal Origins of Coronary Heart Disease," *BMJ: British Medical Journal*, 311, 171–174.

Bateson, P., P. Gluckman, and M. Hanson (2014): "The Biology of Developmental Plasticity and the Predictive Adaptive Response hypothesis," *The Journal of Physiology*, 592, 2357–2368.

Battistin, E., R. Blundell, and A. Lewbel (2009): "Why is Consumption More Log Normal Than Income? Gibrat's Law Revisited," *Journal of Political Economy*, 117, 1140–1154.

Borker, G., J. Eeckhout, N. Luke, S. Minz, K. Munshi, and S. Swaminathan (2021): "Wealth, Marriage, and Sex Selection," Tech. rep., Working Paper, Yale University.

Burgess, S. C. and D. J. Marshall (2014): "Adaptive Parental Effects: the iImportance of Estimating Environmental Predictability and Offspring Fitness Appropriately," *Oikos*, 123, 769–776.

Chancel, L. and T. Piketty (2017): "Indian Income Inequality, 1922-2014: From British Raj to Billionaire Raj?" *Working Paper*.

Cutler, D., A. Deaton, and A. Lleras-Muney (2006): "The Determinants of Mortality," *Journal of Economic Perspectives*, 20, 97–120.

Dandona, L., R. Dandona, G. A. Kumar, D. Shukla, V. K. Paul, K. Balakrishnan, D. Prabhakaran, N. Tandon, S. Salvi, A. Dash, and Others (2017): "Nations Within a Nation: Variations in Epidemiological Transition Across the States of India, 1990–2016 in the Global Burden of Disease Study," *The Lancet*, 390, 2437–2460.

Dasgupta, P. and D. Ray (1986): "Inequality as a Determinant of Malnutrition and Unemployment: Theory," *The Economic Journal*, 96, 1011–1034.

Deaton, A. (2007): "Height, Health, and Development," *Proceedings of the National Academy of Sciences*, 104, 13232–13237.

Diamond, J. (1997): *Guns, Germs and Steel: The Fates of Human Societies*, Vintage.

Galor, O. and Ö. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106, 3064–3103.

Galor, O. and J. Zeira (1993): "Income Distribution and Macroeconomics," *The Review of Economic Studies*, 60, 35–52.

Gluckman, P. D. and M. A. Hanson (2004): "Living With the Past: Evolution, Development, and Patterns of Disease," *Science*, 305, 1733–1736.

——— (2006): "The Conceptual Basis for the Developmental Origins of Health and Disease," in *Developmental Origins of Health and Disease*, ed. by P. Gluckman and M. Hanson, Cambridge University Press, 33 – 50.

Goldsmith-Pinkham, P., I. Sorkin, and H. Swift (2020): "Bartik instruments: What, when, why, and how," *American Economic Review*, 110, 2586–2624.

Haines, M. S., A. Leong, B. C. Porneala, J. B. Meigs, and K. K. Miller (2022): "Association between muscle mass and diabetes prevalence independent of body fat distribution in adults under 50 years old," *Nutrition & Diabetes*, 12, 1–6.

Hansen, B. E. (2017): "Regression Kink With an Unknown Threshold," *Journal of Business & Economic Statistics*, 35, 228–240.

Jablonka, E. and G. Raz (2009): "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution," *The Quarterly Review of Biology*, 84, 131–176.

Kültz, D. (2020): "Defining Biological Stress and Stress Responses based on Principles of Physics," *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology*, 333, 350–358.

Lind, M. I. and F. Spagopoulou (2018): "Evolutionary Consequences of Epigenetic Inheritance," *Heredity*, 121, 205–209.

Lind, M. I., M. K. Zwoinska, J. Andersson, H. Carlsson, T. Krieg, T. Larva, and A. A. Maklakov (2020): "Environmental Variation Mediates the Evolution of Anticipatory Parental Effects," *Evolution Letters*, 4, 371–381.

Malani, A., S. Rosenbaum, S. C. Alberts, and E. A. Archie (2022): "Seeing the Future: a Better Way to Model and Test for Aaptive Developmental Plasticity," Tech. rep., National Bureau of Economic Research.

McKeigue, P., B. Shah, and M. Marmot (1991): "Relation of Central Obesity and Insulin Resistance With High Diabetes Prevalence and Cardiovascular Risk in South Asians," *The Lancet*, 337, 382–386.

Müller, M. J., A. Bosy Westphal, and S. B. Heymsfield (2010): "Is There Evidence for a Set Point That Regulates Human Body Weight?" *F1000 Medicine Reports*, 2.

Müller, M. J., C. Geisler, S. B. Heymsfield, and A. Bosy-Westphal (2018): "Recent advances in understanding body weight homeostasis in humans," *F1000Research*, 7.

Munshi, K. and M. Rosenzweig (2016): "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap," *American Economic Review*, 106, 46–98.

Narayan, K. and A. M. Kanaya (2020): "Why are South Asians prone to type 2 diabetes? A hypothesis based on underexplored pathways," *Diabetologia*, 63, 1103–1109.

Neel, J. V. (1962): "Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"?" *American Journal of Human Genetics*, 14, 353.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Ng, S. W. and B. M. Popkin (2012): "Time Use and Physical Activity: A Shift Away From Movement Across The Globe," *Obesity Reviews*, 13, 659–680.

Oza-Frank, R. and K. V. Narayan (2010): "Overweight and Diabetes Prevalence Among US Immigrants," *American Journal of Public Health*, 100, 661–668.

Petrie, J. R., T. J. Guzik, and R. M. Touyz (2018): "Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms," *Canadian Journal of Cardiology*, 34, 575–584.

Pomeroy, E., V. Mushrif-Tripathy, T. J. Cole, J. C. Wells, and J. T. Stock (2019): "Ancient origins of low lean mass among South Asians and implications for modern type 2 diabetes susceptibility," *Scientific Reports*, 9, 1–12.

Prentice, A. M., S. A. Jebb, G. R. Goldberg, W. A. Coward, P. Murgatroyd, S. Poppitt, and T. Cole (1992): "Effects of Weight Cycling on Body Composition," *The American Journal of Clinical Nutrition*, 56, 209–216.

Robinson, P. M. (1988): "Root-N-consistent Semiparametric Regression," *Econometrica*, 931–954.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb (2019): "Fast and Wild: Bootstrap Inference in Stata Using Boottest," *The Stata Journal*, 19, 4–60.

Scrimshaw, N. S., C. E. Taylor, J. E. Gordon, W. H. Organization, et al. (1968): *Interactions of nutrition and infection*, World Health Organization.

Speakman, J. R. (2007): "A Nonadaptive Scenario Explaining the Genetic Predisposition to Obesity: the "Predation Release" Hypothesis," *Cell Metabolism*, 6, 5–12.

Speakman, J. R., D. A. Levitsky, D. B. Allison, M. S. Bray, J. M. De Castro, D. J. Clegg, J. C. Clapham, A. G. Dulloo, L. Gruer, S. Haw, et al. (2011): "Set points, settling points and some alternative models: theoretical options to understand how genes and environments combine to regulate body adiposity," *Disease models & mechanisms*, 4, 733–745.

Stebbing, A. (2009): "Interpreting 'Dose-Response' Curves using Homeodynamic Data: with an Improved Explanation for Hormesis," *Dose-Response*, 7, 221–233.

Swaminathan, S., R. Hemalatha, A. Pandey, N. J. Kassebaum, A. Laxmaiah, T. Longvah, R. Lodha, S. Ramji, G. A. Kumar, A. Afshin, et al. (2019): "The burden of child and maternal malnutrition and trends in its indicators in the states of India: the Global Burden of Disease Study 1990–2017," *The Lancet Child & Adolescent Health*, 3, 855–870.

Taylor, R. (2013): "Type 2 Diabetes: Etiology and Reversibility," *Diabetes Care*, 36, 1047–1055.

Taylor, R. and R. R. Holman (2015): "Normal Weight Individuals Who Develop Type 2 Diabetes: The Personal Fat Threshold," *Clinical Science*, 128, 405–410.

Vazquez, G., S. Duval, D. R. Jacobs Jr, and K. Silventoinen (2007): "Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis," *Epidemiologic Reviews*, 29, 115–128.

Wells, J. C., E. Pomeroy, S. R. Walimbe, B. M. Popkin, and C. S. Yajnik (2016): "The Elevated Susceptibility to Diabetes in India: An Evolutionary Perspective," *Frontiers in Public Health*, 4, 145.

Online Appendix

A Mathematical Appendix

A.1 Proofs of Propositions

Proof of Proposition 1: At any given level of income $y_t \leq \alpha$,

$$\mathbb{E}(z_t|y_t) = \int_{-\infty}^{y_t} [a + b(y_t - U_t)] P(U_t \mid y_t) \ \mathrm{d}\, U_t$$

Let $f(\cdot)$ denote the density of the y_0 distribution. Applying Bayes' rule:

$$P(U_t \mid y_t) = \frac{P(U_t)P(y_t \mid U_t)}{\int_{-\infty}^{y_t} P(\widetilde{U}_t)P(y_t \mid \widetilde{U}_t)} = \frac{\phi(U_t; \mu_t, \sigma_t^2) f(y_t - U_t)}{\int_{-\infty}^{y_t} \phi(\widetilde{U}_t; \mu_t, \sigma_t^2) f(y_t - \widetilde{U}_t) \ \mathrm{d}\, \widetilde{U}_t}$$

In the absence of any prior knowledge about the distribution of pre-modern income, we make the simplifying assumption that initial income is uniformly distributed; i.e. $f(\cdot)$ is constant. It follows that

$$\mathbb{E}(z_t|y_t) = \int_{-\infty}^{y_t} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \ \mathrm{d}\, U_t = a + b\left(y_t - e^L(y_t)\right) \tag{A.1}$$

where $e^L(y_t) = \frac{1}{\Phi(y_t; \mu_t, \sigma_t^2)} \int_{-\infty}^{y_t} U_t \phi(U_t; \mu_t, \sigma_t^2) \ \mathrm{d}\, U_t$.

Since the uniform distribution has bounded support, the lower range of integration should extend to $y_t - \overline{y}_0$, where $\overline{y}_0$ is the right support of the initial income distribution. The advantage of extending the range to $-\infty$ is that we can solve the model analytically and derive a closed-form expression for $e^L(y_t)$, with simulations reported below in Appendix A.2 indicating that this approximation has no discernable effect on predicted BMI (and the risk of metabolic disease) except in the right tail of the y_t distribution.

Making the same approximation as above, at any given level of income $y_t > \alpha$:

$$\begin{aligned} \mathbb{E}(z_t|y_t) &= \int_{-\infty}^{\alpha} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \ \mathrm{d}\, U_t + \int_{\alpha}^{y_t} [a + by_t] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} \ \mathrm{d}\, U_t \\ &= a + b\left(y_t - e^H(y_t)\right), \quad \text{where } \ e^H(y_t) = \frac{1}{\Phi(y_t; \mu_t, \sigma_t^2)} \int_{-\infty}^{\alpha} U_t \phi(U_t; \mu_t, \sigma_t^2) \, \mathrm{d}\, U_t \end{aligned} \tag{A.2}$$

We next derive closed-form expressions for $e^L(y_t)$, $e^H(y_t)$, which are given as

$$e^L(y_t) = \mu_t - \sigma_t \Lambda\left(\frac{y_t - \mu_t}{\sigma_t}\right) \tag{A.3}$$

$$e^H(y_t) = \frac{\mu_t \Phi\left(\frac{\alpha - \mu_t}{\sigma_t}; 0, 1\right) - \sigma_t \phi\left(\frac{\alpha - \mu_t}{\sigma_t}; 0, 1\right)}{\Phi\left(\frac{y_t - \mu_t}{\sigma_t}; 0, 1\right)} \tag{A.4}$$

where $\Lambda(\cdot)$ is the Inverse Mill's ratio. Focusing on the numerator of the $e^L(y_t)$ expression in (A.1) we can write

$$\begin{aligned}\int_{-\infty}^{y_t} U_t \phi(U_t; \mu_t, \sigma_t^2)\, \mathrm{d}\, U_t &= \int_{-\infty}^{y_t} U_t \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{1}{2}\left(\frac{U_t - \mu_t}{\sigma_t}\right)^2\right] \mathrm{d}\, U_t \\ &= \int_{-\infty}^{\frac{y_t - \mu_t}{\sigma_t}} (\sigma_t x_t + \mu_t) \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} x_t^2\right] \mathrm{d}\, x_t\end{aligned}$$

where the second equality comes from the substitution $x_t = \frac{U_t - \mu_t}{\sigma_t}$. The last equality can be written as

$$\mu_t \Phi\left(\frac{y_t - \mu_t}{\sigma_t}; 0, 1\right) - \sigma_t \phi\left(\frac{y_t - \mu_t}{\sigma_t}; 0, 1\right)$$

given that $\frac{\mathrm{d}\, \phi(x_t; 0,1)}{\mathrm{d}\, x_t} = -x_t \phi(x_t; 0, 1)$. A similar transformation of $\Phi(y_t; \mu_t, \sigma_t^2)$ in the denominator of the $e^L(y_t)$ expression in (A.1) gives us the closed-form expression for $e^L(y_t)$ in equation (A.3). The corresponding expression for $e^H(y_t)$ in equation (A.4) is derived by replacing y_t with α in the upper limit for integration.

To establish that the slope of the BMI-income relationship is positive but less than b below the threshold, substitute the expression for $e^L(y_t)$ from equation (A.3) in equation (A.1) and differentiate with respect to y_t. Given the properties of the inverse Mill's ratio, the slope at $y_t \leqslant \alpha$ is given as

$$\frac{\mathrm{d}\, \mathbb{E}(z_t | y_t)}{\mathrm{d}\, y_t} = b\left[1 + \Lambda'\left(\frac{y_t - \mu_t}{\sigma_t}\right)\right] \in (0, b)$$

Further, to demonstrate that the slope of the BMI-income relationship above the threshold is greater than b, observe from the expression for $e^H(y_t)$ in equation (A.4), that the numerator is independent of y_t and the denominator is increasing in y_t. Hence, $\frac{\mathrm{d}\, e^H(y_t)}{\mathrm{d}\, y_t} < 0$, which implies $\frac{\mathrm{d}\, \mathbb{E}(z_t | y_t)}{\mathrm{d}\, y_t} > b$ for $y_t > \alpha$.

Note, from equations (A.3) and (A.4), that $e^L(y_t) = e^H(y_t)$ at $y_t = \alpha$, and thus, from equations (A.1) and (A.2), there is no level discontinuity at the threshold. To prove that there is, nevertheless, a slope discontinuity at the threshold, $y_t = \alpha$, we need to show that

$$\lim_{y_t \uparrow \alpha} \frac{\mathrm{d}\, \mathbb{E}(z_t | y_t)}{\mathrm{d}\, y_t} \neq \lim_{y_t \downarrow \alpha} \frac{\mathrm{d}\, \mathbb{E}(z_t | y_t)}{\mathrm{d}\, y_t}$$

From equations (A.1) and (A.2), a necessary and sufficient condition for the preceding inequality to be satisfied is that $\frac{\mathrm{d}\, e^L(y_t)}{\mathrm{d}\, y_t} \neq \frac{\mathrm{d}\, e^H(y_t)}{\mathrm{d}\, y_t}$ at $y_t = \alpha$. Using equations (A.3) and (A.4), it can be established that this is indeed the case. For this result, first denote $v_t = \frac{y_t - \mu_t}{\sigma_t}$. From equation (A.3), $e^L(y_t) = \frac{\mathcal{L}(v_t)}{\Phi(v_t; 0, 1)}$, where $\mathcal{L}(v_t) = \mu_t \Phi(v_t; 0, 1) - \sigma_t \phi(v_t; 0, 1)$. From equation (A.4), $e^H(y_t) = \frac{\mathcal{L}(\overline{v})}{\Phi(v_t; 0, 1)}$ where $\overline{v} = \frac{\alpha - \mu_t}{\sigma_t}$. Given that the denominator and the numerator (evaluated at $y_t = \alpha$) of the $e^L(y_t), e^H(y_t)$ expressions are the same, a necessary condition for $\frac{\mathrm{d}\, e^L(y_t)}{\mathrm{d}\, y_t} \neq \frac{\mathrm{d}\, e^H(y_t)}{\mathrm{d}\, y_t}$ is that $\frac{\mathrm{d}\, \mathcal{L}(v_t)}{\mathrm{d}\, y_t} \neq \frac{\mathrm{d}\, \mathcal{L}(\overline{v})}{\mathrm{d}\, y_t}$ at $y_t = \alpha$. $\frac{\mathrm{d}\, \mathcal{L}(\overline{v})}{\mathrm{d}\, y_t} = 0$. From the property of the standard normal distribution, $\phi'(v_t; 0, 1) = -v_t \phi(v_t; 0, 1)$, and,

hence, $\frac{\mathrm{d}\,\mathcal{L}(v_t)}{\mathrm{d}\,y_t}\Big|_{y_t=\alpha} = \frac{\alpha}{\sigma_t}\phi(\overline{v};0,1) > 0$.

Proof of Proposition 2: The relationship between the probability of diabetes, $P(D_t)$, and income is given as

$$P(D_t) = \begin{cases} \gamma_1 & \text{if} \quad U_t \leqslant \alpha \\ \gamma_1 + \gamma_2(y_t - y_0) & \text{if} \quad U_t > \alpha \end{cases} \tag{A.5}$$

Hence, for any given $y_t \leqslant \alpha$, making the same analytical approximation and distributional assumptions as above:

$$P(D_t|y_t) = \int_{-\infty}^{y_t} \gamma_1 \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t = \gamma_1 \tag{A.6}$$

and for any given $y_t > \alpha$,

$$P(D_t|y_t) = \int_{-\infty}^{\alpha} \gamma_1 \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t + \int_{\alpha}^{y_t} (\gamma_1 + \gamma_2 U_t)\frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t$$

$$= \gamma_1 + \gamma_2 \int_{\alpha}^{y_t} U_t \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t$$

Following the same steps that we used to derive the expression for $e^L(y_t)$ in (A.3), we can write for any given $y_t > \alpha$,

$$P(D_t|y_t) = \gamma_1 + \gamma_2 \left[\mu_t - \sigma_t \Lambda\left(\frac{y_t-\mu_t}{\sigma_t}\right) - \frac{\mu_t \Phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right) - \sigma_t \phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)}{\Phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)}\right] \tag{A.7}$$

From equation (A.6), $\frac{\mathrm{d}\,P(D_t|y_t)}{\mathrm{d}\,y_t} = 0$ for $y_t \leqslant \alpha$, and from equation (A.7), $\frac{\mathrm{d}\,P(D_t|y_t)}{\mathrm{d}\,y_t} > 0$ for $y_t > \alpha$ because $\Lambda'(\cdot) < 0$ and $\Phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)$ is increasing in y_t. This also establishes that there is a slope discontinuity at $y_t = \alpha$. Further, substituting $y_t = \alpha$ in equation (A.7) eliminates the term inside square brackets, implying that there is no level discontinuity at $y_t = \alpha$.

A.2 Placing an upper bound on y_0

BMI-income relationship: Assume that the period 0 income has both lower and upper bounds i.e. $y_0 \in [0, \overline{y}_0]$. Hence the range of U_t for any given value of y_t is $[y_t - \overline{y}_0, y_t]$. The mean BMI at any given

$y_t \leqslant \alpha$ is given by

$$\begin{aligned}\mathbb{E}(z_t|y_t) &= \int_{y_t-\overline{y}_0}^{y_t} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2) - \Phi(y_t - \overline{y}_0; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + by_t - b \int_{y_t-\overline{y}_0}^{y_t} U_t \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2) - \Phi(y_t - \overline{y}_0; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + b(y_t - \overline{e}^L(y_t)) \end{aligned} \tag{A.8}$$

where $\overline{e}^L(y_t)$ corresponds to $e^L(y_t)$ in the model without an upper bound on y_0. Following the same steps as in the proof of Proposition 1 above:

$$\overline{e}^L(y_t) = \mu_t - \sigma_t \frac{\left[\phi\left(\frac{y_t-\mu_t}{\sigma_t}; 0, 1\right) - \phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t}; 0, 1\right)\right]}{\left[\Phi\left(\frac{y_t-\mu_t}{\sigma_t}; 0, 1\right) - \Phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t}; 0, 1\right)\right]} \tag{A.9}$$

For $y_t > \alpha$ there are two cases: (i) $y_t \in [\alpha, \overline{y}_0 + \alpha]$ and (ii) $y_t > \overline{y}_0 + \alpha$. In the first case, at each level of y_t, there are two types of individuals: those who remain at their set point and those who have crossed the threshold. The mean BMI at any given $y_t \in [\alpha, \overline{y}_0 + \alpha]$ is thus described by the following expression:

$$\begin{aligned}\mathbb{E}\left(z_t|y_t\right) &= \int_{y_t-\overline{y}_0}^{\alpha} [a + b(y_t - U_t)] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2) - \Phi(y_t - \overline{y}_0; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &\quad + \int_{\alpha}^{y_t} [a + by_t] \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2) - \Phi(y_t - \overline{y}_0; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + by_t - b \int_{y_t-\overline{y}_0}^{\alpha} U_t \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2) - \Phi(y_t - \overline{y}_0; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + b(y_t - \overline{e}^H(y_t)) \end{aligned} \tag{A.10}$$

where $\overline{e}^H(y_t)$ corresponds to $e^H(y_t)$ in the model without an upper bound. As above, this expression can be simplified as

$$\overline{e}^H(y_t) = \frac{\mu_t \left[\Phi\left(\frac{\alpha-\mu_t}{\sigma_t}; 0, 1\right) - \Phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t}; 0, 1\right)\right] - \sigma_t \left[\phi\left(\frac{\alpha-\mu_t}{\sigma_t}; 0, 1\right) - \phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t}; 0, 1\right)\right]}{\Phi\left(\frac{y_t-\mu_t}{\sigma_t}; 0, 1\right) - \Phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t}; 0, 1\right)} \tag{A.11}$$

For $y_t > \overline{y}_0 + \alpha$, everyone has escaped the set point. Hence, the mean BMI at any given $y_t > \overline{y}_0 + \alpha$ is

$$\begin{aligned}\mathbb{E}(z_t|y_t) &= \int_{\alpha}^{\infty} (a + by_t) \frac{\phi(U_t; \mu_t, \sigma_t^2)}{1 - \Phi(\alpha; \mu_t, \sigma_t^2)} \, \mathrm{d}\, U_t \\ &= a + by_t \end{aligned}$$

Diabetes-income relationship: For any given $y_t \leqslant \alpha$,

$$\begin{aligned} P(D_t|y_t) &= \int_{y_t-\overline{y}_0}^{y_t} \gamma_1 \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)-\Phi(y_t-\overline{y}_0;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t \\ &= \gamma_1 \end{aligned}$$

For any given $y_t \in [\alpha, \overline{y}_0+\alpha]$,

$$\begin{aligned} P(D_t|y_t) = &\int_{y_t-\overline{y}_0}^{\alpha} \gamma_1 \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)-\Phi(y_t-\overline{y}_0;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t + \\ &\int_{\alpha}^{y_t} (\gamma_1+\gamma_2 U_t) \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)-\Phi(y_t-\overline{y}_0;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t \\ = \gamma_1 + \gamma_2 &\int_{\alpha}^{y_t} U_t \frac{\phi(U_t;\mu_t,\sigma_t^2)}{\Phi(y_t;\mu_t,\sigma_t^2)-\Phi(y_t-\overline{y}_0;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t \end{aligned}$$

Solving the integral,

$$P(D_t|y_t) = \gamma_1 + \gamma_2 \frac{\mu_t\left[\Phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)-\Phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)\right]-\sigma_t\left[\phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)-\phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)\right]}{\Phi\left(\frac{y_t-\mu_t}{\sigma_t};0,1\right)-\Phi\left(\frac{y_t-\overline{y}_0-\mu_t}{\sigma_t};0,1\right)} \quad \text{(A.12)}$$

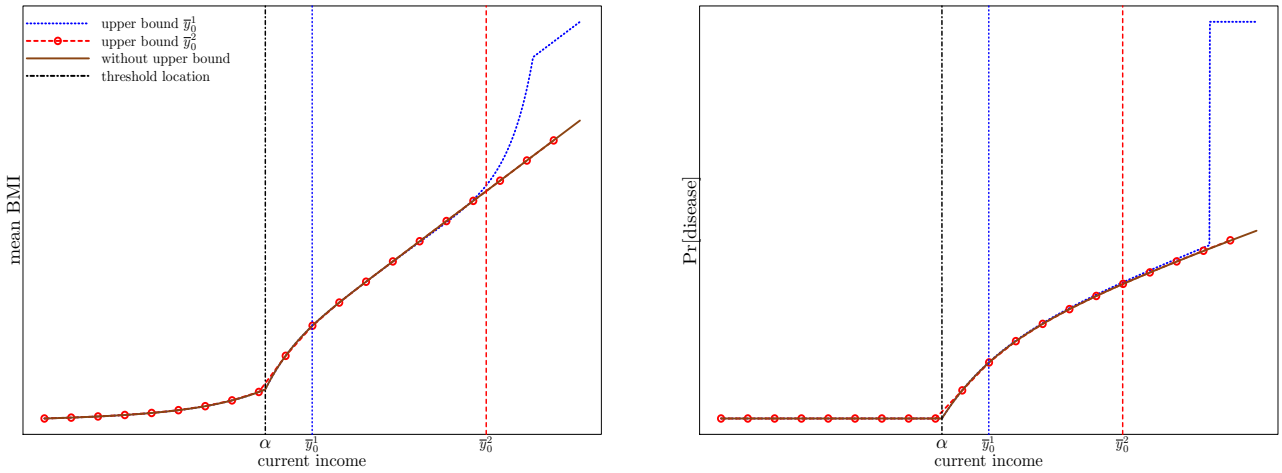
For any given $y_t > \overline{y}_0+\alpha$, as everyone has escaped their set point, we can write,

$$\begin{aligned} P(D_t|y_t) &= \int_{\alpha}^{\infty} (\gamma_1+\gamma_2 U_t) \frac{\phi(U_t;\mu_t,\sigma_t^2)}{1-\Phi(\alpha;\mu_t,\sigma_t^2)}\,\mathrm{d}\,U_t \\ &= \gamma_1 + \gamma_2 \left[\mu_t + \sigma_t \frac{\phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)}{1-\Phi\left(\frac{\alpha-\mu_t}{\sigma_t};0,1\right)}\right] \end{aligned} \quad \text{(A.13)}$$

which is independent of y_t.

Although analytical results can no longer be derived as in Propositions 1 and 2, expressions (A.8), (A.9), (A.11), (A.10), (A.12) and (A.13) can be used to simulate the relationship between current income and both BMI and the probability of metabolic disease. We use the actual income from the IHDS and the estimates of μ_t, σ_t from the structural estimation exercise for the simulation. The left panel in Figure A1 plots the relationship between BMI and current income, with and without the upper bound on y_0. The right panel plots the corresponding relationships between metabolic disease and income. For the upper bound we choose two values of $\overline{y}_0$. The first value $\overline{y}_0^1$, marked by the blue dotted vertical line, is close to the threshold α whereas the second value $\overline{y}_0^2$, marked by the red dashed line, is further to the right. The simulated BMI-income and metabolic disease-income relationships track together, almost exactly, with the three specifications, except in the right tail of the income distribution where we observe a second discontinuity with $\overline{y}_0^1$. In our data, we do not observe a second discontinuity, at a high income level, with either BMI or the risk of metabolic disease as outcomes.

Figure A1: Simulated Cross-Sectional Relationships with upper bound on y_0



(a) BMI

(b) Metabolic disease

A.3 Alternative Specifications for the Set Point

A.3.1 Set point determined by ancestral and current income

Assume that a dynasty's set point is determined, each period, by the weighted average of ancestral income and current income. The relationship between BMI and income can now be written as

$$z_t = \begin{cases} a + b[r_t y_0 + (1 - r_t)y_t] & \text{if } \; y_t - [r_t y_0 + (1 - r_t)y_t] \leqslant \widetilde{\alpha} \\ a + by_t & \text{if } \; y_t - [r_t y_0 + (1 - r_t)y_t] > \widetilde{\alpha} \end{cases} \tag{A.14}$$

where $r_1 = 1$ and $\lim_{t\to\infty} r_t = 0$. $y_t - [r_t y_0 + (1 - r_t)y_t] = r_t(y_t - y_0) = r_t U_t$. Hence, the threshold becomes time variant and is given by $\frac{\widetilde{\alpha}}{r_t}$. The mean BMI at any given $y_t \leqslant \frac{\widetilde{\alpha}}{r_t}$ can then be expressed as

$$\begin{aligned} \mathbb{E}\left[z_t | y_t\right] &= \int_{-\infty}^{y_t} (a + b[r_t y_0 + (1 - r_t)y_t]) \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} dU_t \\ &= \int_{-\infty}^{y_t} (a + b[y_t - r_t U_t]) \frac{\phi(y_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} dU_t \\ &= a + b(y_t - r_t e^L(y_t)) \end{aligned}$$

where $e^L(y_t)$ is defined in (A.3). Similarly, for any given $y_t > \frac{\widetilde{\alpha}}{r_t}$, we can write

$$\begin{aligned}
\mathbb{E}\left[z_t|y_t\right] &= \int_{-\infty}^{\frac{\widetilde{\alpha}}{r_t}} (a + b[y_t - r_t U_t]) \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} dU_t + \int_{\frac{\widetilde{\alpha}}{r_t}}^{y_t} (a + by_t) \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} dU_t \\
&= a + by_t - br_t \int_{-\infty}^{\frac{\widetilde{\alpha}}{r_t}} U_t \frac{\phi(U_t; \mu_t, \sigma_t^2)}{\Phi(y_t; \mu_t, \sigma_t^2)} dU_t \\
&= a + b(y_t - r_t \widetilde{e}^H(y_t))
\end{aligned}$$

where the expression for $\widetilde{e}^H(y_t)$ is the same as in equation (A.4) when α is replaced by $\frac{\widetilde{\alpha}}{r_t}$.

A.3.2 Set point determined by previous generation income

Assume that a dynasty's set point is determined, each period, by the previous generation's income. The relationship between nutritional status and income can be written as

$$z_t = \begin{cases} a + by_{t-1} & \text{if } y_t - y_{t-1} \leqslant \overline{\alpha} \\ a + by_t & \text{if } y_t - y_{t-1} > \overline{\alpha} \end{cases} \tag{A.15}$$

Assuming that $y_{t-1} \geqslant 0$, and using $u_t = y_t - y_{t-1}$ where $u_t \sim N(\mu, \sigma^2)$, we can write mean BMI for any given $y_t \leqslant \overline{\alpha}$ as

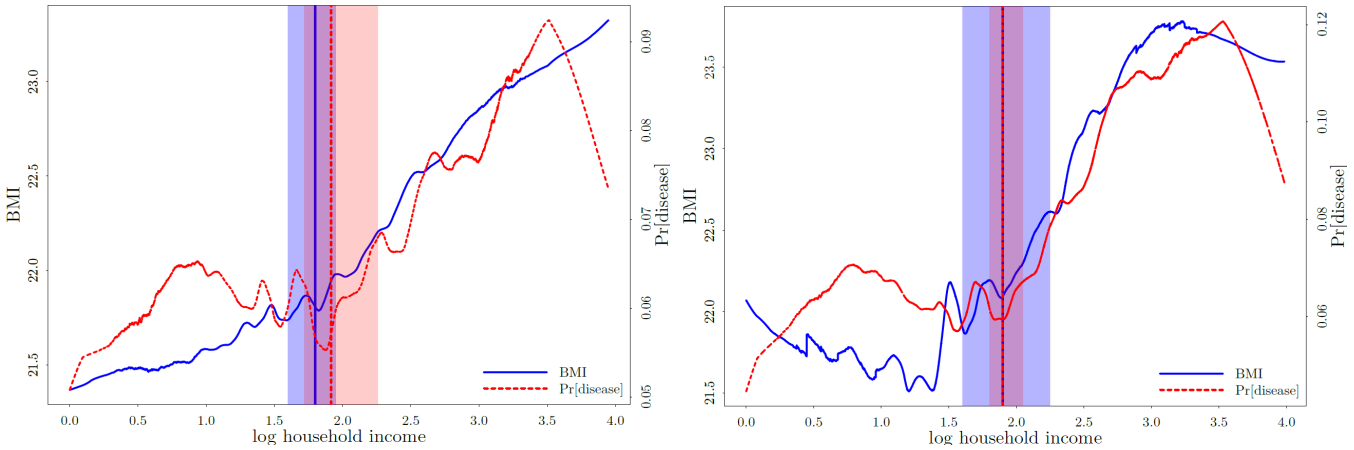
$$\begin{aligned}
\mathbb{E}[z_t|y_t] &= \int_{-\infty}^{y_t} [a + by_{t-1}] \frac{\phi(u_t; \mu, \sigma^2)}{\Phi(y_t; \mu, \sigma^2)} du_t \\
&= a + by_t - b \int_{-\infty}^{y_t} u_t \frac{\phi(u_t; \mu, \sigma^2)}{\Phi(y_t; \mu, \sigma^2)} du_t \\
&= a + b(y_t - e^L(y_t; \mu, \sigma^2))
\end{aligned}$$

Similarly, mean BMI at any given $y_t > \overline{\alpha}$ is given as

$$\begin{aligned}
\mathbb{E}[z_t|y_t] &= \int_{-\infty}^{\overline{\alpha}} [a + by_{t-1}] \frac{\phi(u_t; \mu, \sigma^2)}{\Phi(y_t; \mu, \sigma^2)} du_t + \int_{\overline{\alpha}}^{y_t} [a + by_t] \frac{\phi(u_t; \mu, \sigma^2)}{\Phi(y_t; \mu, \sigma^2)} du_t \\
&= a + by_t - b \int_{-\infty}^{\overline{\alpha}} u_t \frac{\phi(u_t; \mu, \sigma^2)}{\Phi(y_t; \mu, \sigma^2)} du_t \\
&= a + b(y_t - e^H(y_t; \mu, \sigma^2))
\end{aligned}$$

B Appendix Figures and Tables

Figure B1: Nutritional Status and Metabolic Disease with respect to Household Income (additional covariates and nonparametric shift-share instrument)



(a) Additional covariates

(b) Nonparametric shift-share instrument

Source: India Human Development Survey (IHDS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in panels (a) and (b).
For panel (a), additional covariates include dummies for the number of adults, teens, and children in the household, dummies for the number of individuals engaged in manual labor, and dummies for the highest education of adult females and males. For panel (b), additional covariates include land ownership, its interaction with the rural dummy, and the residual (linear, quadratic, and cubic terms) from the first-stage nonparametric regression, as described below. Covariates are partialled out prior to nonparametric estimation.
The vertical lines mark the estimated threshold location and the shaded areas demarcate the corresponding confidence intervals.

Table B1: Piecewise Linear Equation Estimates (additional covariates and nonparametric shift-share instrument)

Robustness exercise:	additional covariates		nonparametric shift-share instrument	
Dependent Variable:	BMI	metabolic disease	BMI	metabolic disease
	(1)	(2)	(3)	(4)
Baseline slope (β_1)	0.281**	0.003	0.192	0.005
	(0.052)	(0.002)	(0.411)	(0.006)
Slope change (β_2)	0.516**	0.014**	1.522*	0.044**
	(0.069)	(0.003)	(0.827)	(0.012)
Threshold location (τ)	1.80	1.95	1.95	1.90
	[1.60, 1.95]	[1.75, 2.30]	[1.60, 2.25]	[1.80, 2.05]
Threshold test $p-$value	0.000	0.000	0.022	0.000
Mean of dependent variable	22.002	0.074	22.275	0.073
N	76,949	148,928	73,708	138,782

Source: India Human Development Survey (IHDS)

Metabolic disease indicates whether the individual has been diagnosed with diabetes, hypertension, or cardiovascular disease.

The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in the estimating equation.

For columns 1-2, additional covariates include dummies for the number of adults, teens, and children in the household, dummies for the number of individuals engaged in manual labor, and dummies for the highest education of adult females and males.

For columns 3-4, additional covariates include land ownership, its interaction with the rural dummy and the residual (linear, quadratic, and cubic terms) from the first-stage nonparametric regression, as described below.

Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.

For all columns except for column (3), cluster bootstrapped 95% confidence bands for the threshold location are in brackets. For column (3), 90% confidence bands are provided.

**, * significant at 5%, 10%, based on cluster bootstrapped confidence intervals.

The instrumental variable estimation proceeds in the following steps:

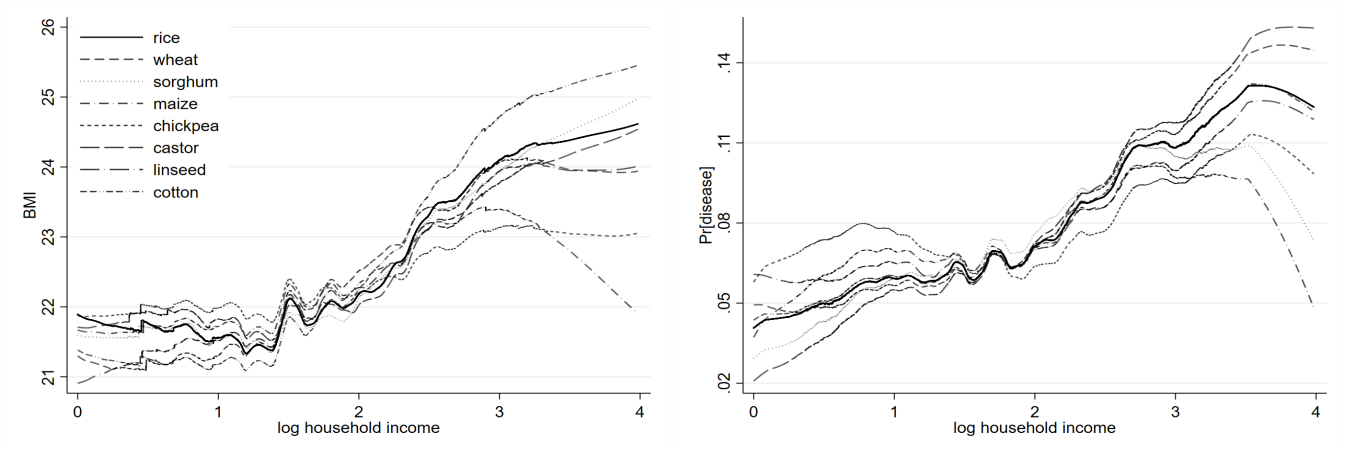
Step 1: We use ICRISAT District Level Data (DLD) for India to construct the growth in output value, at the national level over the 1966-2015 period, for each of the following crops: rice, wheat, sorghum, maize, chickpea, castor, linseed and cotton. We then construct a district-level measure of the growth in value by taking a weighted average of the growth of each crop, where the weight is the acreage allocated to that crop in 1965 divided by total cultivated acreage in that year. District-level growth is interacted with the rural dummy and land owned by the household (obtained from IHDS) to construct the shift-share instrument.

Step 2: We regress household income nonparametrically on the shift-share instrument, after partialling out district effects, the rural dummy, land ownership and the interaction of land ownership with the rural dummy, using the Robinson procedure. The coefficient on the shift-share instrument in a corresponding linear regression has a t-statistic of 3.75 (F=13.67), indicating that the instrument has sufficient statistical power.

Step 3: Following Newey et al. (1999), we include a polynomial (cubic) function of the residuals from the preceding step, land ownership, and its interaction with the rural dummy as additional covariates, which are partialled out together with the standard set of controls, when we nonparametrically estimate the BMI-income and metabolic disease-income relationships.

Step 4: Following Goldsmith-Pinkham et al. (2020), we validate the nonparametric instrumental variable estimates, reported in Figure B1b and Table B1, Columns 3-4 by using acreage shares of individual crops, rather than the growth in value, to construct crop-specific instruments. As Goldsmith-Pinkham et al. note, the estimates will be similar with each crop if the shift-share instrument is valid, and this is indeed what we observe below.

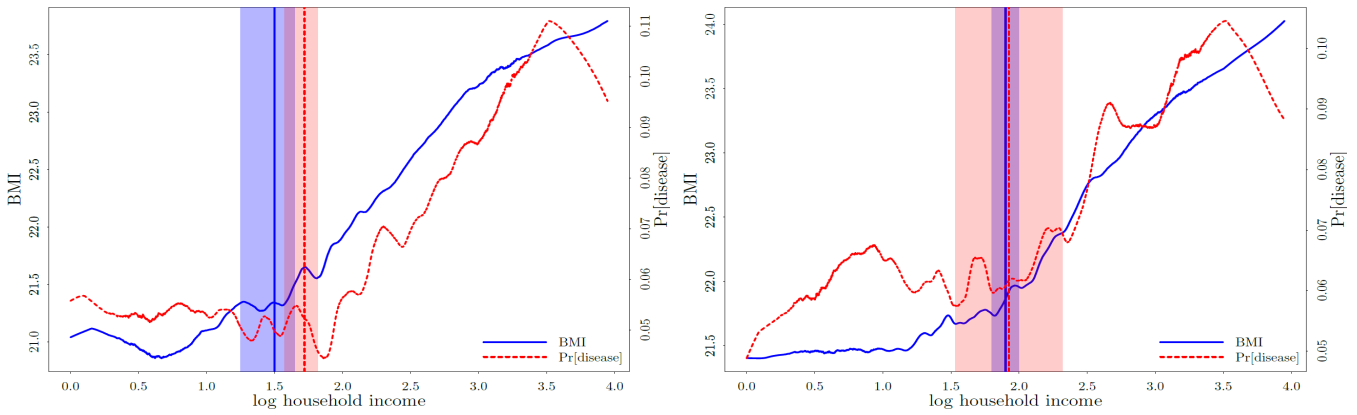
Figure B2: Nutritional Status and Metabolic Disease with respect to Household Income (instrument based on individual crop shares)



(a) BMI (b) Metabolic disease

Source: India Human Development Survey (IHDS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round, together with land ownership, its interaction with the rural dummy, and the residual from the first-stage nonparametric regression (linear, quadratic, and cubic terms) are partialled out prior to nonparametric estimation.

Figure B3: Nutritional Status and Metabolic Disease with respect to Household Income (separately for men and women)



(a) Men (b) Women

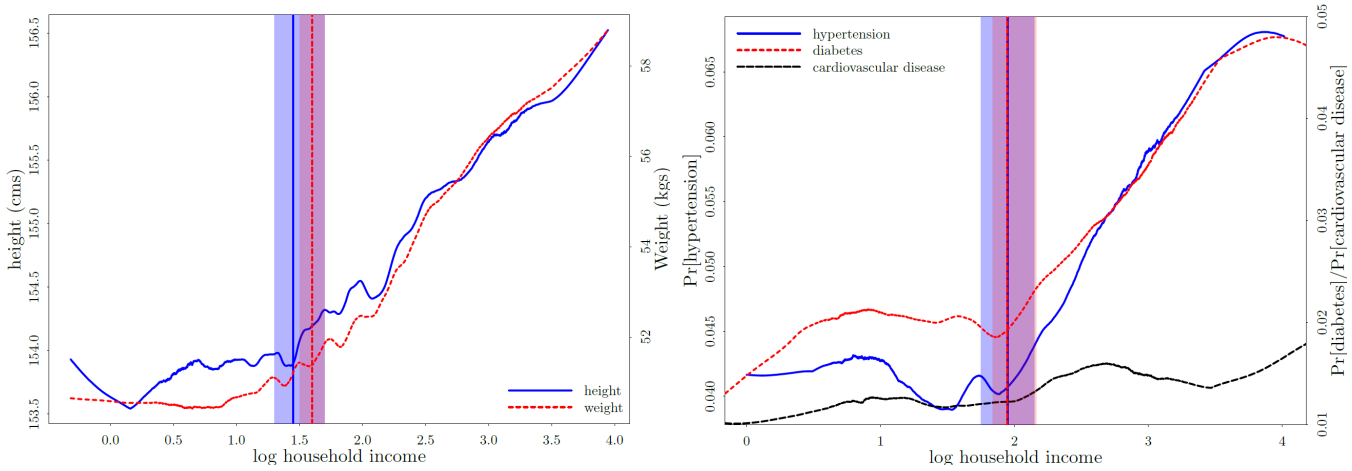
Source: India Human Development Survey (IHDS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for caste group, rural area, district, and survey-round are partialled out prior to nonparametric estimation.
The vertical lines mark the estimated threshold location and the shaded areas demarcate the corresponding confidence intervals.

Table B2: Piecewise Linear Equation Estimates (separately for men and women)

Dependent variable:	BMI		metabolic disease	
Sample:	men	women	men	women
	(1)	(2)	(3)	(4)
Baseline slope (β_1)	0.342**	0.225**	−0.001	0.005
	(0.104)	(0.062)	(0.003)	(0.003)
Slope change (β_2)	0.877**	0.980**	0.038**	0.018**
	(0.112)	(0.079)	(0.004)	(0.005)
Threshold location (τ)	1.50	1.75	1.90	1.95
	[1.25, 1.65]	[1.60, 1.85]	[1.80, 2.00]	[1.55, 2.35]
Threshold test $p-$value	0.000	0.000	0.000	0.002
Mean of dependent variable	21.854	22.060	0.071	0.077
N	20,596	56,044	71,768	77,160

Source: India Human Development Survey (IHDS)
Metabolic disease indicates whether the individual has been diagnosed with diabetes, hypertension, or cardiovascular disease.
Logarithm of household income is the independent variable.
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for caste group, rural area, district, and survey-round are included in the estimating equation.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
Cluster bootstrapped 95% confidence bands for the threshold location are in brackets.
** significant at 5%, based on cluster bootstrapped confidence intervals.

Figure B4: Alternative Nutritional Status Measures and Metabolic Diseases (separately) with respect to Household Income



(a) Height and weight

(b) Hypertension,diabetes, and cardiovascular disease

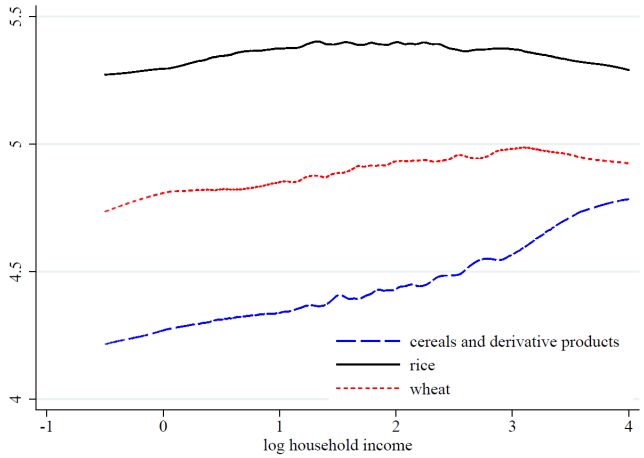
Source: India Human Development Survey (IHDS)
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in the estimating equation.
The vertical lines mark the estimated threshold locations and the shaded areas demarcate the corresponding cluster bootstrapped 95% confidence intervals.

Table B3: Piecewise Linear Equation Estimates (alternative nutritional status measures, hypertension, diabetes, and cardiovascular disease)

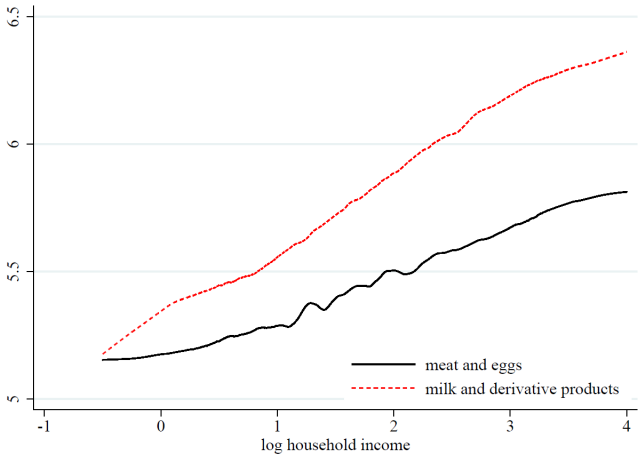
Measures:	alternative nutrition measure		metabolic disease		
Dependent variable:	height	weight	hypertension	diabetes	cardiovascular disease
	(1)	(2)	(3)	(4)	(5)
Baseline slope (β_1)	0.191	0.656**	0.001	0.001	0.001**
	(0.135)	(0.150)	(0.002)	(0.001)	(0.0005)
Slope change (β_2)	0.836**	2.863**	0.018**	0.017**	
	(0.144)	(0.174)	(0.003)	(0.002)	
Threshold location (τ)	1.45	1.60	1.95	1.95	
	[1.30,1.70]	[1.50,1.70]	[1.75, 2.15]	[1.85, 2.15]	
Threshold test $p-$value	0.000	0.000	0.000	0.000	
Mean of dependent variable	154.483	52.578	0.049	0.027	0.014
N	77,000	77,143	147,858	147,684	147,626

Source: India Human Development Survey (IHDS)
Logarithm of household income is the independent variable.
The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round are included in the estimating equation.
For columns (1)-(4), bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses. For column (5), standard errors are clustered at the primary sampling unit level.
Cluster bootstrapped 95% confidence bands for the threshold location are in brackets.
** significant at 5%, based on cluster bootstrapped confidence intervals.

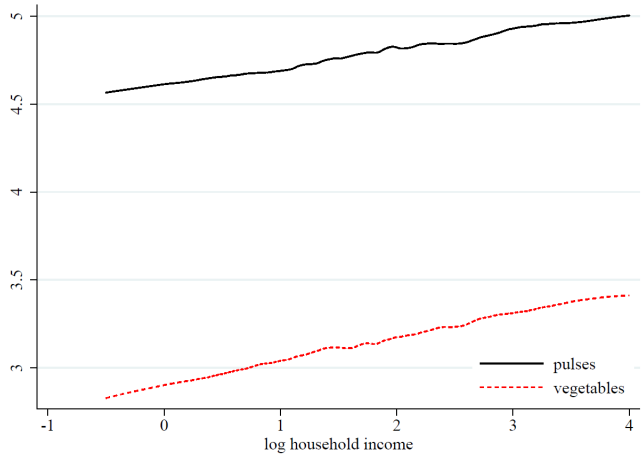
Figure B5: Expenditure on different food categories with respect to household income



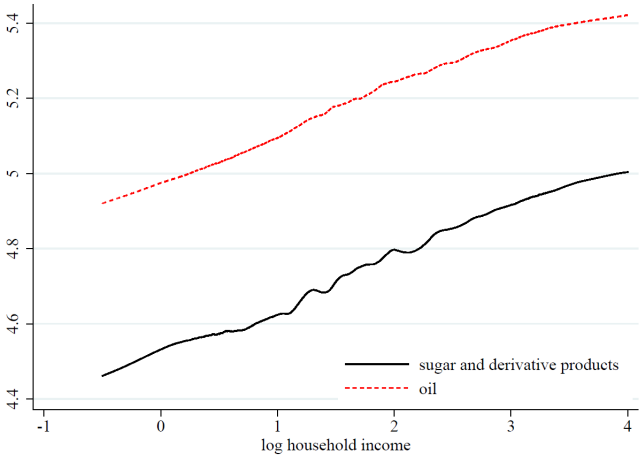
(a) Cereals, rice and wheat



(b) Meat, eggs and milk, including derivative products



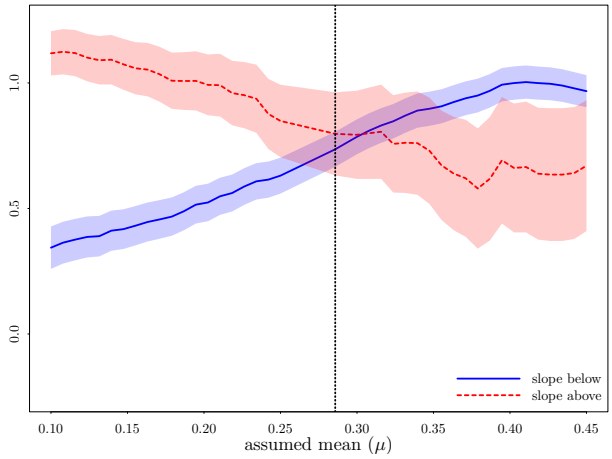
(c) Pulses and vegetables



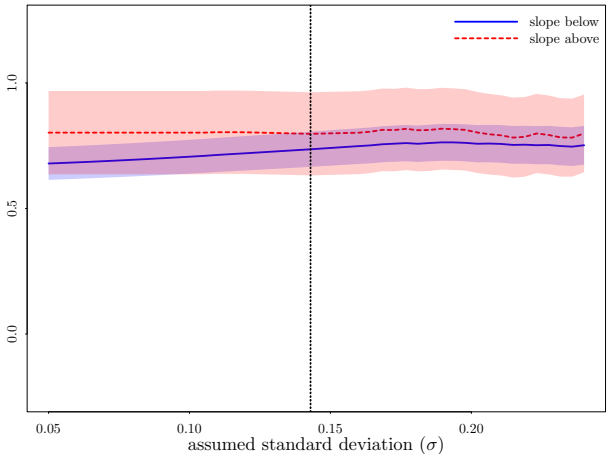
(d) Oil and sugar and derivative products

Source: India Human Development Survey (IHDS).
This figure plots the nonparametric relationship between expenditures on different food categories and household income. Food expenditures are measured as the log of monthly expenditures in Rupees. The following covariates are partialled out prior to the nonparametric estimation: reported local price of rice, wheat, cereals and their derivative products, pulses, meat, sugar, oil, eggs, milk and its derivative products, vegetables and dummies for the number of children, adults, and teens in the household, occupation, caste group, rural area, district, and survey-round.

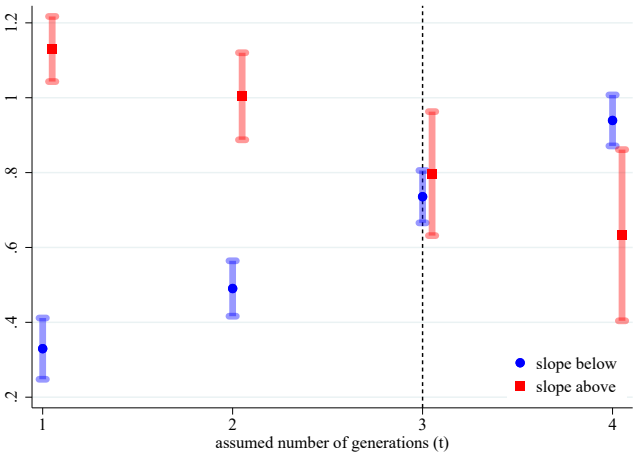
Figure B6: Sensitivity of Slope Coefficients with respect to Parameter Values



(a) Mean of the income shock



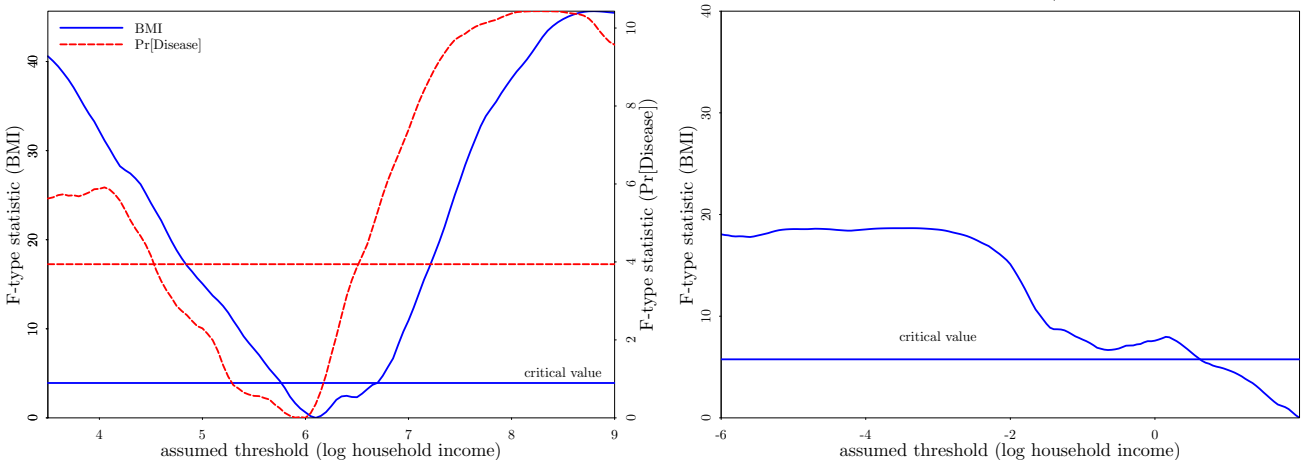
(b) Standard deviation of the income shock



(c) Number of generations

Notes: This figure plots the estimated slope coefficients, below and above the threshold, with respect to three parameters of the model: (i) mean of the income shock, (ii) standard deviation of the income shock, and (iii) the number of generations. The vertical line in each panel marks the parameter value that we use for estimation in Table 2.

Figure B7: Nutritional Status and Metabolic Disease with respect to Income (Indonesia and Ghana)



(a) Threshold tests (Indonesia)

(b) Threshold test (Ghana)

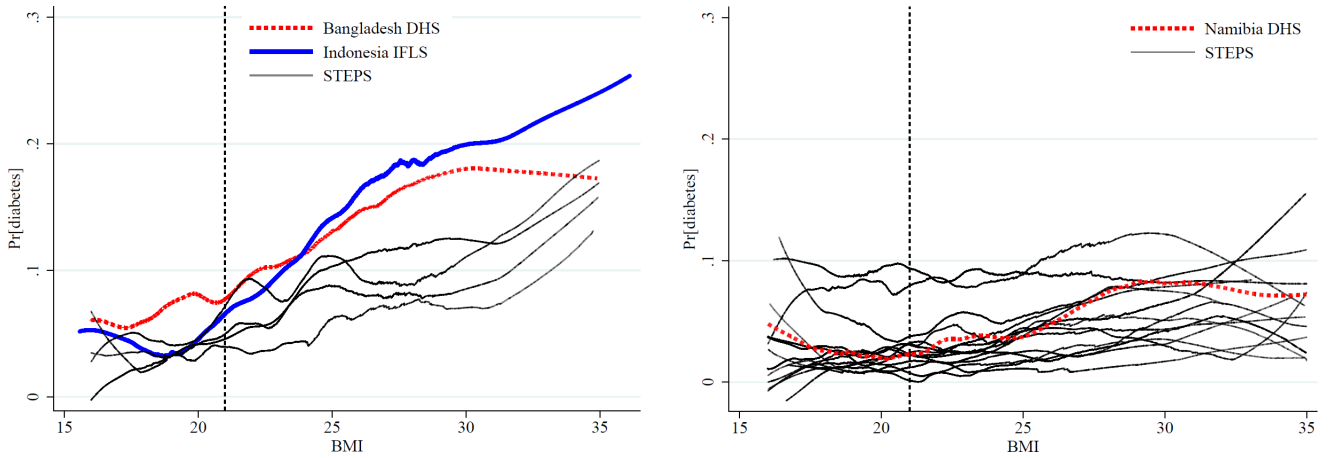
Source: Indonesia Family Life Survey (IFLS), Ghana Socioeconomic Panel Survey (GSPS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, ethnicity (Indonesia) or tribe (Ghana), rural area, regency (Indonesia) or district (Ghana), and survey-round are included in the estimating equation at each assumed threshold for the threshold test.
Indonesia: bootstrapped 5% critical values, clustered at the sub-regency level. Ghana: bootstrapped 5% critical values, clustered by enumeration area.

Table B4: Piecewise Linear Equation Estimates (Indonesia and Ghana)

Sample country:	Indonesia		Ghana
Dependent variable:	BMI	metabolic disease	BMI
	(1)	(2)	(3)
Slope below (β_L)	0.067	-0.001	0.165***
	(0.065)	(0.010)	(0.036)
Slope above (β_H)	0.398**	0.022**	–
	(0.069)	(0.011)	
Threshold location (τ)	6.10	6.00	–
	[5.80, 6.65]	[4.55, 6.50]	
Threshold test $p-value$	0.000	0.004	–
Dep. var. mean	23.532	0.181	23.934
N	30,812	24,788	11,372

Source: Indonesia Family Life Survey (IFLS), Ghana Socioeconomic Panel Survey (GSPS)
Metabolic disease indicates whether the individual has been diagnosed with diabetes, hypertension, or cardiovascular disease.
Logarithm of household income is the independent variable.
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, ethnicity (Indonesia) or tribe (Ghana), rural area, regency (Indonesia) or district (Ghana), and survey-round are included in the estimating equation.
Bootstrapped standard errors, clustered at the sub-regency level for Indonesia and by enumeration area for Ghana, are in parentheses.
Cluster bootstrapped 95% confidence bands for the threshold location are in brackets.
** significant at 5%, *** at 1%, based on cluster bootstrapped confidence intervals.

Figure B8: Measured Diabetes and BMI



(a) South and South-East Asia

(b) Sub-saharan Africa

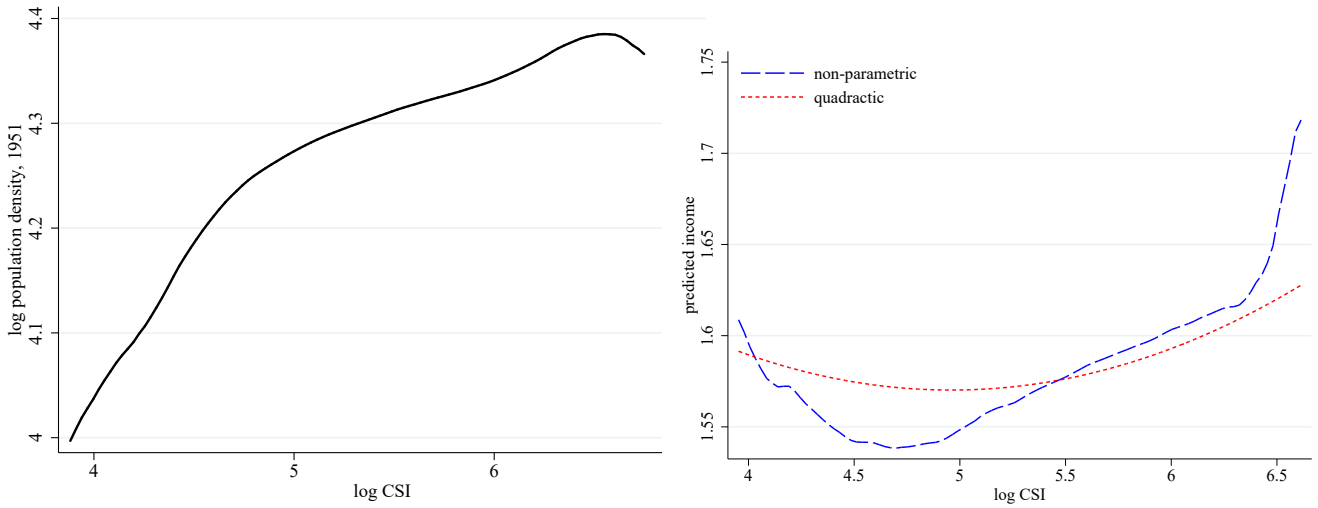
Source: Indonesia Family Life Survey (IFLS), Demographic and Health Survey (DHS), WHO-STEPS
The following covariates: age (linear, quadratic and cubic terms), and dummies for gender, country, and survey year are partialled out prior to nonparametric estimation.
WHO-STEPS African countries, with final sample size: Benin (9541), Botswana (2839), Central African Republic (5752), Cameroon (7026), Comoros (2359), Eritrea (5137), Ethiopia (7303), Ghana (2162), Guinea (1976), Mozambique (2447), Malawi (6859), Rwanda (5897), Tanzania (4613), Togo (3297), Uganda (3045), Zambia (4887).
WHO-STEPS Asian countries: Laos (2311), Sri Lanka (4028), Myanmar (7382), Nepal (4681).
DHS countries: Bangladesh (17207), Namibia (2955). IFLS with biomarkers (6240)

Table B5: Piecewise Linear Equation Estimates (reported metabolic disease and measured diabetes)

Dependent variable:	metabolic disease (IHDS)	diabetes (DHS)
	(1)	(2)
Baseline slope (β_1)	0.003**	0.001**
	(0.001)	(0.0004)
Slope change (β_2)	0.006**	0.010**
	(0.001)	(0.0005)
Threshold location (τ)	21.80	21.80
	[20.20, 22.80]	[21.60, 22.00]
Threshold test $p-$value	0.000	0.000
Mean of dependent variable	0.066	0.136
N	76,103	730,995

Source: India Human Development Survey (IHDS), Demographic Health Survey (DHS) 2015-16
BMI is the independent variable. The standard set of covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, rural area, district, and survey-round for IHDS are included in the estimating equation.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses. Cluster bootstrapped 95% confidence bands for the threshold location are in brackets.
** significant at 5%, based on cluster bootstrapped confidence intervals.

Figure B9: Relationship between Population Density, Predicted Household Income and Caloric Suitability Index (CSI)

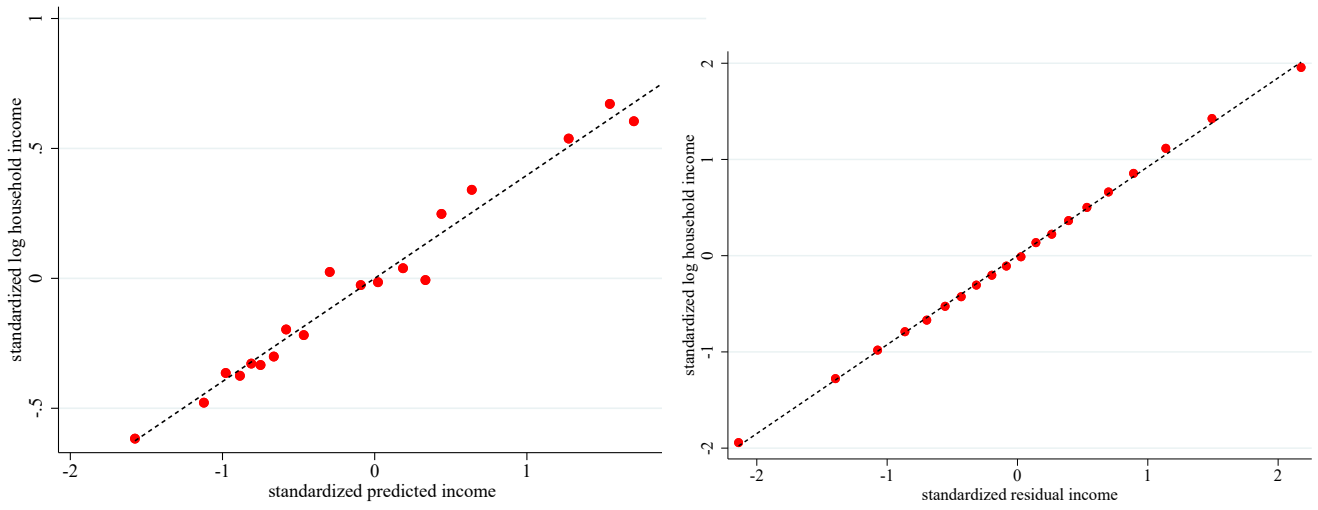


(a) Population density

(b) Predicted household income

Source: FAO-GAEZ dataset, 1951 population census, India Human Development Survey (IHDS)

Figure B10: Relationship between Household Income, Predicted Income, and Residual Income



(a) Predicted income

(b) Residual income

Source: FAO-GAEZ dataset, India Human Development Survey (IHDS)
This figure reports binned scatter plots describing the relationship between current household income, y_t, and (i) predicted income, which is our measure of y_0, and (ii) residual income, which is our measure of U_t. All variables are standardized.

Table B6: Nutritional Status - Income Relationship (below and above the threshold, quadratic $f(CSI)$ function)

Dependent variable:	BMI			
Country:	India		Indonesia	
Sample:	below	above	below	above
Ancestral income	0.530	-0.009	0.966***	0.339
	(0.243)	(0.202)	(0.354)	(0.475)
Current income	0.194***	0.854***	-0.048	0.601***
	(0.040)	(0.047)	(0.120)	(0.065)
Threshold location (τ)	1.65	1.65	6.1	6.1
Dep. var. mean	20.482	21.851	22.317	23.021
N	27,164	20,296	3,182	10,610

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia), and survey-round are included in the estimating equation.The rural-urban dummy is excluded, since the sample is restricted to rural households.
For India, staple crops are wheat and rice. For Indonesia, the staple crop is rice.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

Table B7: Nutritional Status - Income Relationship (below and above the threshold, additional crops)

Dependent variable:	BMI			
Country:	India		Indonesia	
Sample:	below	above	below	above
Ancestral income	0.538**	0.406	1.332***	0.552
	(0.193)	(0.210)	(0.281)	(0.381)
Current income	0.186***	0.848***	-0.051	0.589***
	(0.040)	(0.047)	(0.118)	(0.063)
Threshold location (τ)	1.65	1.65	6.1	6.1
Dep. var. mean	20.482	21.851	22.317	23.021
N	27,164	20,296	3,182	10,610

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia) and survey-round are included in the estimating equation.The rural-urban dummy is excluded, since the sample is restricted to rural households.
For India, staple crops are wheat, rice, sorghum, barley and millet. For Indonesia, staple crops are rice, sorghum, cassava and maize.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

Table B8: Metabolic Disease - Income Relationship (quadratic $f(CSI)$ function)

Dependent variable:	metabolic disease			
Country:	India		Indonesia	
Income component:	income mismatch	ancestral income	income mismatch	ancestral income
	(1)	(2)	(3)	(4)
Income component	0.001	0.006	-0.003	0.016
	(0.002)	(0.006)	(0.011)	(0.037)
Income component × $\mathbf{1\{}$income $> \boldsymbol{\tau}\mathbf{\}}$	0.018***	-0.002	0.030**	0.006
	(0.004)	(0.002)	(0.011)	(0.009)
Joint significance				
$F-$statistic [$p-$value]	16.070	0.734	12.699	0.341
	[0.000]	[0.481]	[0.000]	[0.711]
Threshold location (τ)	1.90	1.90	6.00	6.00
Dep. var. mean	0.054	0.054	0.162	0.162
N	90,879	90,879	11,001	11,001

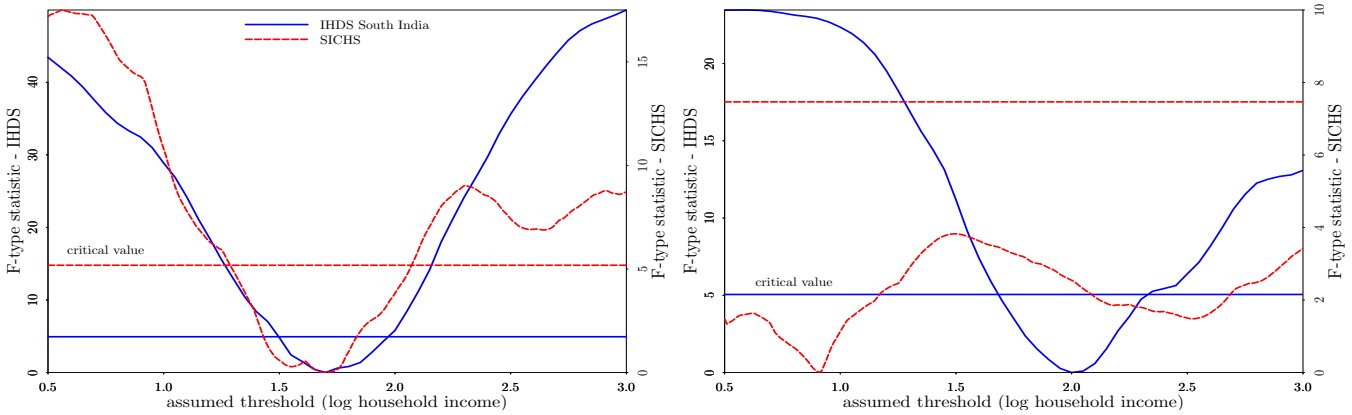
Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia) and survey-round are included in the estimating equation. The rural-urban dummy is excluded, since the sample is restricted to rural households.
For India, staple crops are wheat and rice. For Indonesia, staple crop is rice.
$F-$statistic measures the joint significance of the uninteracted and interacted income component coefficients.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

Table B9: Metabolic Disease - Income Relationship (additional crops)

Dependent variable:	metabolic disease			
Country:	India		Indonesia	
income component:	income mismatch	ancestral income	income mismatch	Ancestral income
	(1)	(2)	(3)	(4)
Income component	0.001	0.005	-0.004	0.006
	(0.002)	(0.007)	(0.011)	(0.020)
Income component ×				
$\mathbf{1\{}$income $> \boldsymbol{\tau}\mathbf{\}}$	0.018***	-0.002	0.031**	0.004
	(0.003)	(0.002)	(0.011)	(0.009)
Joint significance				
$F-$statistic $[p-$value$]$	15.646	0.435	13.121	0.236
	[0.000]	[0.648]	[0.000]	[0.790]
Threshold location (τ)	1.90	1.90	6.00	6.00
Dep. var. mean	0.054	0.054	0.162	0.162
N	90,879	90,879	11,001	11,001

Source: India Human Development Survey (IHDS), Indonesia Family Life Survey (IFLS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group (India) or ethnicity (Indonesia), state (India) or regency (Indonesia) and survey-round are included in the estimating equation.The rural-urban dummy is excluded, since the sample is restricted to rural households.
For India, staple crops are wheat, rice, sorghum, barley and millet. For Indonesia, staple crops are rice, sorghum, cassava and maize.
$F-$statistic measures the joint significance of the uninteracted and interacted income component coefficients.
. Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

Figure B11: Threshold Tests - Nutritional Status and Metabolic Disease (IHDS and SICHS)



(a) BMI (b) metabolic disease

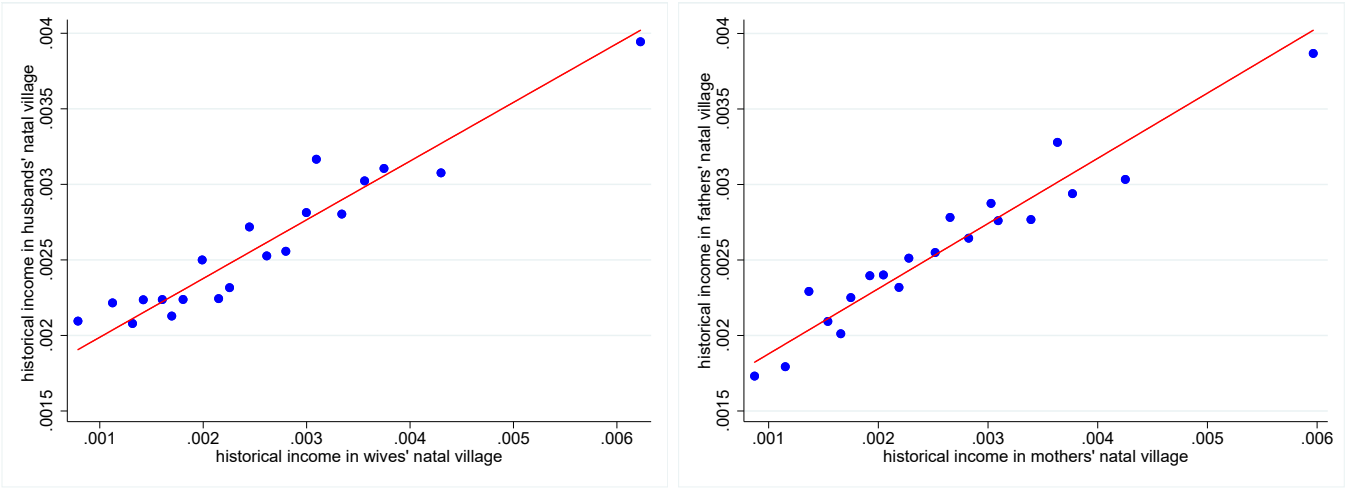
Source: India Human Development Survey (IHDS), South India Community Health Study (SICHS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, and (for IHDS) rural area, district and survey-round are included in the estimating equation at each assumed threshold for the threshold test.
Cluster bootstrapped 5% critical values are used to bound the threshold location.

Table B10: Piecewise Linear Equation Estimates – Nutritional Status and Metabolic Disease (South India)

Source:	IHDS		SICHS
Dependent variable:	BMI	metabolic disease	BMI
	(1)	(2)	(3)
Slope below (β_L)	0.200**	0.001	0.079
	(0.112)	(0.005)	(0.369)
Slope above (β_H)	0.803**	0.029**	1.148**
	(0.125)	(0.008)	(0.382)
Threshold location (τ)	1.70	2.00	1.69
	[1.50, 1.95]	[1.75, 2.25]	[1.29, 2.07]
Threshold test $p-$value	0.000	0.000	0.002
Dep. var. mean	22.186	0.074	23.449
N	22,316	41,198	7,634

Source: India Human Development Survey (IHDS), South India Community Health Study (SICHS)
Metabolic disease indicates whether the individual has been diagnosed with diabetes, hypertension, or cardiovascular disease.
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender, caste group, and (for IHDS) rural area, district and survey-round are included in the estimating equation.
Bootstrapped standard errors, clustered at the level of the primary sampling unit, are in parentheses.
** significant at 5%, based on cluster bootstrapped confidence intervals

Figure B12: Assortative Matching on Historical Income



(a) Current generation

(b) Parents' generation

Source: South India Community Health Study (SICHS)
Historical income is measured by tax revenue per acre of cultivated land in 1871 in the individual's natal village.
The number of bins in the binned scatter plot is set equal to 20.

Table B11: Nutritional Status - Income relationship below and above the threshold (SICHS)

Dep. var.:	BMI					
Income measure:	current income		ancestral income (non-parametric $g(R)$ function)		ancestral income (quadratic $g(R)$ function)	
Sample:	below	above	below	above	below	above
	(1)	(2)	(3)	(4)	(5)	(6)
Income coefficient	0.072	0.773***	0.334***	0.170	0.374***	0.014
	(0.161)	(0.102)	(0.123)	(0.159)	(0.129)	(0.127)
Dep. var. mean	23.034	23.755	23.034	23.755	23.034	23.755
N	1810	3844	1810	3844	1810	3844

Source: South India Community Health Study (SICHS)
The following covariates: age (linear, quadratic, and cubic terms) and dummies for gender and caste group are included in the estimating equation.
Bootstrapped standard errors, clustered at the level of the village, are in parentheses.
$*$ significant at 10%, $**$ at 5%, $***$ at 1%, based on cluster bootstrapped confidence intervals.

C Selective Child Mortality

Suppose that there is a positive and continuous relationship between mean nutritional status and household income, with a fixed dispersion in nutritional status at each level of income, as in Figure C1. If children can only survive above a subsistence nutrition level, and this constraint only binds at lower income levels, then as observed in the figure there will be a discontinuous relationship between mean nutritional status, which we measure by BMI, and income. If this relationship persists into adulthood, then the observed BMI-income relationship can be explained by selective mortality in childhood (without requiring a set point).

Notice, however, that the discontinuous relationship between mean BMI and income with this alternative model is driven entirely by households at the lower end of the nutritional status distribution, at each income level. Child mortality is concentrated in the first five years and, hence, if the nutritional status-income relationship is distorted by child mortality, this will show up most clearly among the 5-19 year olds. As with adult BMI, we precisely locate an income threshold at which the slope of the relationship between household income and child (aged 5-19) BMI-for-age changes discontinuously. Figures C2a and C2b report quantile regression estimates of the baseline slope coefficient (β_1) and the slope-change coefficient (β_2) in the piecewise linear equation estimated at that threshold. The conditional mean of the baseline slope coefficient and the slope change coefficient, evaluated at the mean of the dependent variable, BMI-for-age, match what we obtain with adult BMI as the outcome in Table 1, Column 1: the baseline slope is close to zero and the slope change is positive and significant. Moreover, it is evident from Figures C2a and C2b that these results are not driven by a small fraction of households at the bottom of the nutritional status distribution, as the alternative explanation based on selective child mortality would imply. We cannot statistically reject the hypothesis that the estimated coefficients at each quantile are equal to the corresponding conditional mean coefficient.

Figure C1: Child Nutritional Status with respect to Income (with selective child mortality)

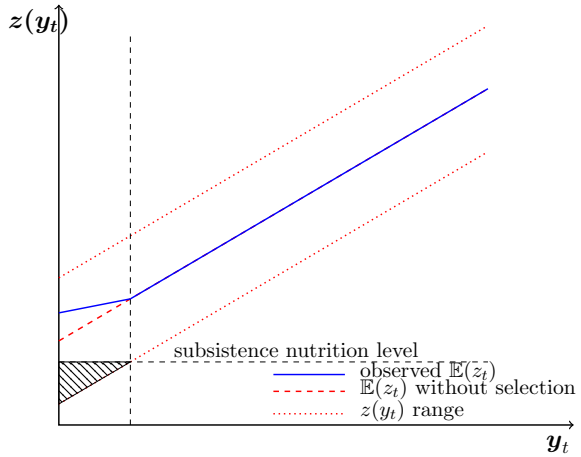
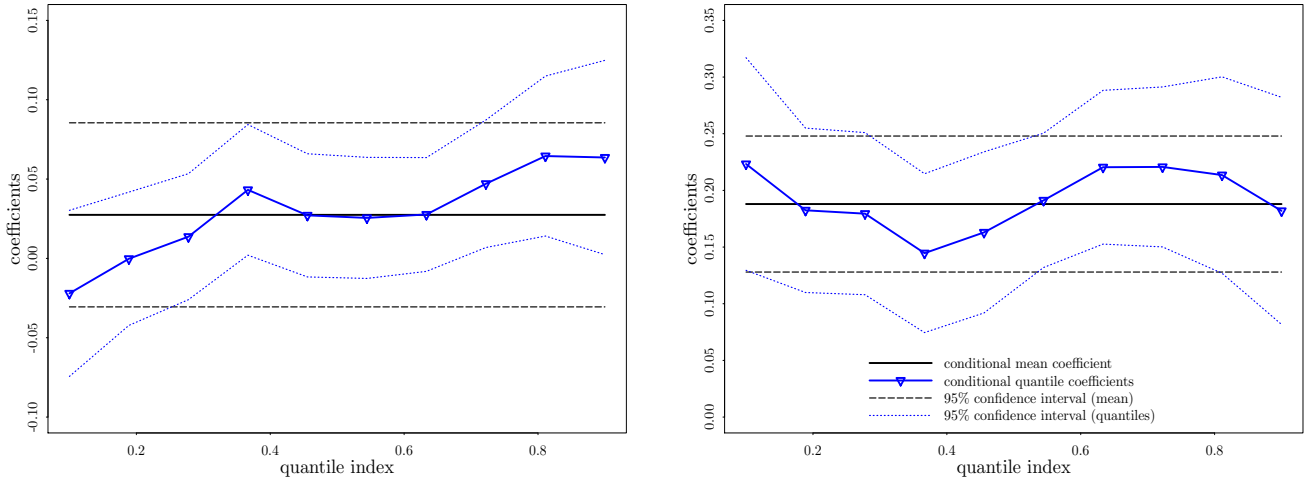


Figure C2: Conditional Mean and Conditional Quantile Coefficients (child nutritional status with respect to income)



(a) Baseline slope

(b) Slope change

Source: India Human Development Survey (IHDS)