Uncertainty and Mortality Estimates in the Latin American and Caribbean (LAC) region: 1850-2010

Alberto Palloni[*] Hiram Beltrán-Sánchez[†]

Abstract

We propose simple techniques to account for uncertainty of estimates of adult mortality indicators into statistical analyses. We consider two dimensions of uncertainty: errors due to violation of assumptions on which estimates are based and probability of departures from assumptions in a particular empirical case. We use simulations to estimate the entire distribution of errors of selected indirect estimates of adult mortality due to violations of some or a combination of assumptions on which the methods rely. We then describe strategies to assess the validity of some or a combination of assumptions in empirical cases and, finally, propose methods to combine knowledge about estimates' errors and exogenous information about validity of assumptions. As an illustration we use the classic problem of identification of effects of improvements of living standards and medical technology on secular mortality decline. In addition, we introduce model uncertainty and assess its effects on inferences when occurring jointly with adult mortality estimates uncertainty. We then compare our results with those from standard practice that ignore imprecision of estimates and alternative choices of model. Although the techniques introduced here are designed to facilitate analysis of mortality estimates in countries with deficient statistics, they are quite general and could be generalized to other applications.

1 Introduction

It is not uncommon for analysis of mortality in population with defective data to handle multiple estimates of the same parameters, usually summary statistics of survival patterns such as life expectancy at fixed ages. Some of these estimates are the outcome of methods to estimate adjustments for completeness of intercensal death registration, others are the product of a mixture of vital statistics and adjustment factors derived from exogenous sources, some emerge from the combination of independent estimates of adult and child mortality plus reliance on model mortality patterns and, yet others are the product of third party computations that rest on (partially or fully) disclosed computational rules. As a consequence, there is a heterogeneous space containing k_i estimates for a single country-year i or observation point and, superficially at least, they may all seem equally plausible. Choosing one among them is conventionally done more or less explicitly using arbitrary rules and, in most cases, without ever considering the uncertainty that the choice implies. The task of the researcher usually comes to a close when a single value out of k_i possible is exhibited as a suitable choice. This need not be the case. We aim to show that, under some conditions, it is possible to implement procedures that consider all available estimates while

[*] Center for Demography of Health & Aging, University of Wisconsin-Madison. Email: palloni@ssc.wisc.edu.

[†] Department of Community Health Sciences & California Center for Population Research, UCLA. Email: beltrans@ucla.edu

simultaneously associating with each a level of uncertainty that can be explicitly incorporated into analyses of the data.

The approach rests on the idea that an estimate θ_j of a population parameter Θ obtained with method j can be associated with a measure of error that depends on (a) known distribution of errors given departures from assumptions and (b) prior knowledge of probabilities than the observed data violates assumptions. The target quantity is $P(\theta_j \epsilon I|D)$, the probability that $|\theta_j - \mathit{\Theta}| < I$ given the data set D at hand. The quantity is defined as

$$P(\theta_j \epsilon I|D) = P(\theta_j \epsilon I|A)P(A|D) + P(\theta_j \epsilon I|\overline{A})(1 - P(A|D)) \tag{1.1}$$

where A stands for "assumption is met by the data" and $\overline{A}$ for "assumption is violated by the data". Once we assign values to the elements of this expression, we compute an estimate of $P(\theta_j \epsilon I|D))$ and then proceed to use it jointly with the value of θ_j in hypotheses testing and model estimation.

Although (1.1) appears simple enough, it leads to massive complications. In most applications of the sort we deal with here $P(\theta_j \epsilon I|\overline{A})(1-P(A|D)$ is small and can be ignored. However, defining and estimating the remaining terms of the expression is far from trivial. First, A is likely to be a set of assumptions, not just a single assumption. Even if these assumptions were summarized as an array of binary variables (satisfied/not satisfied in the observed data) the number of possible combinations may mount quickly. Second, the idea that one can reduce the concordance between assumptions and data to a binary categorization may be misleading. In some cases what matters is not only whether or not an assumption is met but what are the nature and magnitude of departures from conditions that generate the data. Third, when the above issues are all settled, there still remains the problem of assigning numerical values to both $P(\theta_j \epsilon I|A)$ and $P(A|D)$. In the sections that follow we outline simple procedures to generalize (1.1) to make it applicable to situations encountered when using standard indirect estimates of adult mortality levels and patterns. We illustrate with applications to model estimation. Although the procedures are designed to facilitate analysis of mortality estimates, they are quite general and could, in principle at least, be exported to different applications.

2 Uncertainty of estimates of relative completeness of death registration

There are close to 10-15 different classes of estimates of relative completeness of adult death registration suitable for populations with defective vital statistics and census counts[1]. Although some of the techniques to compute these estimates are variants of each other and somewhat interdependent, most are characterized by non-overlapping features that, under some conditions, may make them more (less) desirable relative to candidate estimates. To the extent that variability of empirical conditions under which these estimates are applied may improve (worsen) their performance, it might be to the researcher's advantage to establish a formal, unambiguous link between empirical conditions producing the observed data and assumptions supporting a class of estimates, on the one hand, and their actual performance, on the other. The linkage could be operationalized as levels of uncertainty and explicitly accounted for in protocols to choose estimates that are subsequently employed in more complex analyses.

[1] Relative completeness refers to the net error of an observed mortality rate, e.g. the product of errors in both the numerator, or the observed number of events, (deaths), and the denominator, or the observed population years of exposure. It is defined as the ratio of M_x^o/M_x^T where M_x^o is the observed mortality rate in the age group $(x, x+1)$ and M_x^T is the true, unknown, mortality rate.

To fix ideas let θ_j be the value of an estimate produced by method j and $\mathit{\Lambda}_j = \{A_1, A_{2,}...A_J\}$ be the set of assumptions on which the estimate relies. If we let each assumption be represented by a binary variable that attains a value 1 when the assumption is violated and 0 otherwise, $\mathit{\Lambda}_j$ can be rendered as a vector of 0's and 1's. We assume throughout that if all assumptions in set $\mathit{\Lambda}_j$ are satisfied by the data, then θ_j is an optimal estimate of the population parameter. Assume also we have known measures of errors when the data departs from the assumptions contained in $\mathit{\Lambda}_j$. To simplify description we define a vector Φ_j=$(\phi_1, \phi_2, ...\phi_{J_j})$ where each entry is a simple error measure, e.g mean squared error, incurred when the corresponding assumption is violated by the available data (but all others are satisfied). In addition, faced with a given data produced by well-defined empirical conditions, the researcher is able to associate with each assumption in $\mathit{\Lambda}_j$ a probability P_j that the assumption is satisfied in the current data, D. Let the corresponding vector of probabilities be $\Omega_j = \{P(A_1|D), \ldots, P(A_J|D)\}$.

2.1 The nature of Φ_j and Ω_j

A simulation study is the best tool to associate each assumption with a measure of the error that obtains when the assumption is violated (but all others are satisfied). This defines the vector Φ_j. In turn, either prior knowledge about conditions that generate the data or results from diagnostic tests can be used to identify whether or not an assumption is violated or the probability that it is violated. This defines the vector Ω_j. Once the two vectors are specified it is in principle possible to compute the expected value of error associated with method j as the product of a row vector Ω_j and column vector Φ_j^T,e.g.

$$\varepsilon_j|D = \sum_{k=1}^{J} P(A_k|D)\phi_k. \tag{2.1}$$

This approach is simplistic for two reasons. First, in theory at least, *ex-ante* knowledge produced, for example, by a simulation study will yield complete information about the distribution of errors for each method, not just a point estimate of it. Thus, one could compute any number of parameters from the error distributions, including the mean but also centiles as well as measures of spread and skewness. To make this explicit one can redefine the vector Φ so that its entries are themselves vectors containing information on as many parameters of the error distribution as deemed necessary.

A second simplification is that in any empirical situation it is likely that not just one but a combination of assumptions will be violated and that each of these combinations is (in theory at least) associated with an error distribution of estimates. Thus, for example, violation of assumption A_j, $A_j = 1$, when A_k is met, $A_k = 0$, may yield a different error distribution than when it is not or $A_k = 1$. This implies that the space of departures from assumptions (and, therefore, the space of potential errors associated with each estimate) is the much larger set of all possible combinations of realizations of assumptions. For each estimate this space is composed of 2^J possible configurations of assumptions. We now have a new vector Ω_j defined as $\Omega_j = \{P(C_1|D), ...P(C_K|D)\}$, where $P(C_r|D)$ stands for the probability that assumption configuration r prevails given data D and K is the total number of relevant assumption configurations for the estimator θ_j. There is also a vector of errors $\Delta_j = (\delta_1, \delta_2, ...\delta_K)$ where the elements are measures of errors associated with a configuration of assumptions rather than single assumptions. Expression (2.1) becomes a weighted sum of errors associated with configurations of assumptions, e.g.

$$\varepsilon_j | D = \sum_{r=1}^{K} P(C_r | D) \delta_r. \tag{2.2}$$

The generalization above is desirable but impractical. If the number of assumptions exceeds say 3 or 4, the number of possible configurations becomes very large[2]. Also a complex set of configurations complicates the assignment of values to the vector Ω_j even if we use a $1/0$ treatment for assumptions. In the application below we choose a small subset of possible (with non- zero probability of occurring) configurations of assumptions per estimate, namely, those deemed to be the more 'realistic' or 'plausible' for each empirical case[3].

2.2 Simulation of errors

Consider the situation encountered with a selected set of methods to estimate relative completeness of adult (older than age 5) death registration in the Latin American and Caribbean region (LAC) during the period 1950-2010. The complete set of assumptions invoked by at least one of the methods contains the following:

1. population stability
2. age-invariance of relative completeness of census and death registration
3. equal completeness of censuses
4. closure to migration
5. no age misreporting

Altogether we consider 13 different methods to estimate relative completeness of death registration (Palloni et al. 2015). Only two of these rely on the assumption of stability and all rely on assumptions 2 to 5. For methods that do not require stability the error produced by lack of stability is always set to 0.

We design a simulation study to evaluate errors under two sets of general conditions: (a) when assumptions are violated one at a time and (b) when assumptions are violated in combinations defined a priori. Case (a) is the simplest and the expected error of a method is given by expression (2.1). Case (b) considers configuration of assumptions and is more realistic but is also less tractable unless one limits the combination of assumptions over which errors must be computed. To do so we ignore configurations where assumptions (2), (3) and (5) are not violated, e.g. we only choose configurations associated with vectors where the elements for these assumptions are set to 1. This limits the total number of relevant configurations to 4. Finally, to account for types and magnitude of departures from assumptions we include a limited number of simulated populations that depart from assumptions in different ways and then average errors over them. For example, lack of stability

[2] Some configurations are implausible and are assigned probability 0. For example a population cannot simultaneously be stable (not violate stability assumption) but be open to migration (violate the assumption of closure to migration).

[3] The proposed treatment assumes that a precise assessment of errors is possible knowing only whether or not a particular assumption is met by the data and ignoring (potentially available) information about the type (or magnitude) of departures from assumptions. As described below we handle this additional complication in an admittedly crude way. However, because in most cases we study in this paper it is sufficient to use a binary approach, the simplification is harmless.

(with no migration) is represented by two non stable regimes, one where only mortality declines over time and another where both mortality and fertility decline following a regime believed to apply to the average LAC country. Similarly, departures from assumption (5) are represented by populations where age misreporting follows a fixed age pattern but the levels of misreporting can attain one of five possible values[4].

2.2.1 Empirical definition of Δ

Vector Δ_j summarizes information from the simulation. It contains knowledge of the distribution of errors given known departures from assumptions. Although there are alternative ways of defining errors, in the application that follows we consider only one measure, namely, the probability that the estimated parameters is within 5 percent of the true parameter, that is, $\Delta_j = (\delta_1, \delta_2, ...\delta_K) = (P(|\theta_j - \mathit{\Theta}| < .05|C_1), P(|\theta_j - \mathit{\Theta}| < .05|C_2), \ldots, P(|\theta_j - \mathit{\Theta}| < .05|C_K))$where $P(|\theta_j - \mathit{\Theta}| < .05|C_K)$ is the probability that $|\theta_j - \mathit{\Theta}| < .05$ given the assumption configuration C_k. In this case the quantity of interest to us is

$$p_j = \sum_{c=1}^{K_j} P(C_r|D)\delta_r \tag{2.3}$$

and the value of p_j contains all the information about error associated with estimate θ_j and can be used as a measure of the uncertainty of the estimate.

2.2.2 Empirical definition of Ω

Vector $\mathit{\Omega}$ embodies prior knowledge of the case being studied and should be evaluated accordingly. For example, it is well known that mortality decline in most LAC countries began almost surely after 1950 and, in some of them at least, shortly after the turn of the XIX century. In any of these cases the assumption of stability will be met only approximately right after the onset of mortality decline and not at all in decades close to the end of the XXth century. Thus, prior knowledge dictates that for some countries any configuration of assumptions including $(A_1 = 0)$ (assumption of stability holds) must have a probability close to 0 for any data produced after 1970 or so. By the same token, we know that censuses in LAC countries are distorted by systematic age overstatement. It is also suspected that these errors decline over time. Thus, any configuration of assumptions that includes the assumption of reliable adult age reporting $A_5 = 0$ should be assigned a probability 0 for all years before 1980 and larger than zero after 1980.

One of the inputs to define vector $\mathit{\Omega}$ is expert knowledge about the demographic history of the population and/or knowledge of the quality of censuses and vital statistics. The use of expert knowledge to assign probability values to configurations of assumptions requires to handle variability of judgments across experts and this is an additional source of estimates' variances (see below).

In addition to *ex ante* knowledge about empirical conditions generating the data, analysts have the benefit of information derived from diagnostic tests. These tests are designed to ascertain whether or not the data that generates estimates produces indications that point toward violations of one or more assumptions. For example, adult age overreporting leads to telltale signs as it forces adult mortality patterns to taper off too rapidly at older ages or induces intercensal survival ratios that are grossly anomalous. When such diagnostic tests are applied, the researcher is provided

[4] The complete set of simulated populations is described in (Palloni et al. 2015)

with additional sources of prior information that she can use to assign probabilities $P(C_r|D)$.[5] Diagnostic tests are not always unambiguous and, as is the case with expert judgment, they can generate heterogeneous signals. Thus, whether the main source of prior information is expert judgment or diagnostic tests, the vector Ω may not be unique. This suggests that the values of $P(C_r|D)$'s themselves are subject to variability that must be explicitly accounted for in the analyses.

3 Uncertainty about uncertainty

There are situations when we cannot compute precise values of the probabilities $p_{j.}$ associated with alternative estimates of adult mortality. For example, most mortality estimates in LAC countries for the period 1850-1950 employ indirect techniques whose performance has not yet been assessed via simulations. Thus, estimates from the generalized ogive, a useful tool when there is no access to vital events, rely on assumptions about the accuracy of intercensal rates of growth and model mortality patterns. However, we lack quantitative information about errors when assumptions are not met. Similarly, one may have access to a stock of estimates from third parties but have only coarse information about data used, assumptions invoked, or methods applied.

To handle these cases we need different and, in most cases, more ambiguous rules. We first assemble the set of all estimates not associated with precise measures of uncertainty, including third parties candidates. For example, we can pool estimates from adjusted generalized procedures and those computed by demographers, officials from statistical offices, international agencies. For each point i, defined as a country/year of observation, we assemble a set of estimates $\{\theta_{ij}, j = 1, ...k_i\}$ that rests on a set of fully identifiable assumptions (those associated with a well-defined method such as the ogive) or on partial information about data and method employed by third parties. Each pair composed of method and assumptions can be scored on two dimensions: 'realism' and 'sensitivity'. Realism refers to the likely degree of concordance between an assumption l used by method j at point i and the objective conditions at i. Sensitivity is defined as the likelihood that departures of assumption l from objective conditions at i lead to non trivial errors in the estimate j. For simplicity we assign the following values or scores, σ_{ijl}: 3 when assumption is realistic and method is insensitive to assumption, 2 when assumption is realistic but the method is highly sensitive to assumption, 1 when assumption is unrealistic and method is insensitive to violations and, finally, 0 when assumption is unrealistic and method is sensitive to violations. The score or weight for estimate θ_{ij} is proportional to $S_{ij} = (1/L_j * 3) \sum_{\forall l} \sigma_{ijl}$ where L_j is the maximum number of identifiable assumption of method j . These scores can then be used in a form analogous to the uncertainty weights computed before.

4 Handling uncertainty of estimates

The values of p_j associated with each estimate θ_j are measures of uncertainty we wish to account for in the analysis. There are a number of ways of doing this.

4.1 A naive approach

The simplest solution is to choose for each point an estimate that satisfies $\arg max_{\{\theta_j\}}(p(\theta_j))$, that is, the estimate with the highest probability of being within 5 percent of the true value. This is similar to the standard approach used in most research in this field, though in practice researchers do not actually compute $p(\theta_j)$ but instead assign subjective values and then choose the estimate

[5] Note that $P(C_r|D)$ perform a role similar to that of an "informed" prior in Bayesian analysis in the sense that it encapsulates whatever is known about the mechanisms producing the data.

that is believed to have the highest probability of being the "right one" ("gold standard"). The problem with this is that all available information contained in alternative estimates is discarded.

4.2 Weighted estimates

A relatively simple solution is to use weights proportionally to some suitably standardized function of p_j, $g(p_j)$, to define a point estimate of the unknown parameter as the weighted average $\tilde{\theta} = \sum g(p_j)\theta_j$. This strategy can be deployed both in simple analyses, e.g. estimating a mortality trend over time, or in more complex cases, e.g. estimating models to identify determinants of mortality trends. While the strategy is appealing for its simplicity, it actually masks the uncertainty underlying estimates unless one computes and considers explicitly $var(\tilde{\theta})$. This quantity depends not only on the observed variance of θ_j but also on the variability of p_j created by heterogeneous results from diagnostic tests and/or experts judgments.

A second method to incorporate uncertainty embedded in the θ_j's is to use weights directly in the estimation of the model. Suppose, for example, we wish to estimate a country's slope of the time trend of life expectancy over 100 years and that there are multiple estimates θ_j per year each associated with values p_j. One can then fit a ML or GLS model that assigns to each estimate a weight p_j.[6] Similarly, if one is interested in the effects of a country's per capita income and literacy rates on life expectancy over a period of 50 years, we could use weighted structural equation models.

4.3 Bootstrap

A more natural solution is to use a form of bootstrap that considers the set of estimates θ_j for a particular point (country-year) as the result of a draw from a meta population of estimates. From our simulation results we know the details of the distribution of errors associated with each estimate under a number of conditions regarding violation of assumptions. If the population under study belongs to the space of simulated populations, then the estimate of the target parameter is associated with a known (from simulation) distribution of errors. This information is embedded in the quantities p_j that express the probability that the corresponding estimate is a "correct" measure of the underlying parameter given the data at hand[7]. Alternatively, one could think of p_j as proportional to the number of times that θ_j would be the "correct choice" if we computed all estimates for a given point a large number of times.

The next step is to implement a bootstrap. For each point we sample an estimate θ_j from the set of available estimates with probability proportional to p_j. We then repeat for all observation points and when all are populated with one estimate each, we fit the desired model. The procedure is repeated N times yielding N estimates of the target parameters (and standard errors) of the model. We then compute mean or median of estimates and retrieve its bootstrapped standard error.

The key difference (other than computational) between weighted estimation and the bootstrap-based procedure is that the former generates measures of uncertainty associated with multiple estimates of parameters *conditional on a model* whereas the uncertainty of estimates from the bootstrap is non-parametric, e.g. the estimates from the model and their standard errors rest on the bootstrap and are model independent (see below).

[6] As multiple estimates for a given time period are not independent, one ought to use cluster-adjusted estimated variances.

[7] According to our definition, "correct" means that the estimate if within 5 percent of the true value.

5 Uncertainty of estimates, uncertainty of models

The sections above dealt with uncertainty of adult mortality estimates. The strategy suggested there is useful but also limited and incomplete since it ignores sources of uncertainty associated with model choice. Model uncertainty creates problems of its own but its consequences may be aggravated if mixed with estimates' uncertainty. Past research on mortality trends and determinants ignored both. It is only recently that some researchers began explicitly introducing model and demographic estimates uncertainty but seldom treating them jointly (Wheldon et al. 2013; Gerland et al. 2014; Raftery et al. 2012; Alkema et al. 2012, 2008).

5.1 Model uncertainty

Model uncertainty can be handled using Bayesian or frequentist approaches. The Bayesian approach can be implemented using Bayes factors (Raftery 1996; Gelman and Rubin 1995) and Bayesian Model Averaging (BMA) (Raftery et al. 1999; Madigan and Raftery 1994; Raftery et al. 1997) and has found its way in a number of applications in social sciences (Kaplan and Chen 2014; Sala-i Martin et al. 2004). There are also several frequentists approaches (Hjort and Claeskens 2003; Breiman 1996; Buja and Stuetzle 2006) which have become popular in the prediction literature. In this paper we opt for a non-parametric bootstrap approach to model selection along the lines proposed by Buckland (Buckland et al. 1997)(see also section 2 in Efron (2014)). This approach smooths predicted values from alternative models thought it does not remove "jumpiness" of estimates in the boundaries between alternative models.[8] We use (unadjusted) bootstrap smoothing because it is computationally simpler than BMA and because its execution is very much in line with the method proposed to handle estimates' uncertainty. In fact, our proposed strategy consist of chaining together the bootstrap for estimates and a bootstrap for models to arrive at a more realistic assessment of errors and more conservative hypotheses testing.

5.2 A three-stage estimator

We combine the strategy suggested in section 4 to handle uncertainty of estimates and bootstrap smoothing to treat model uncertainty. We do so in three stages. First, for each point i (country-year) of the M available we have a set of k_i estimates of life expectancy, $\{\theta_{ij}, j = 1, ...k_i\}$. For each observation (point) i in our sample we draw a single value θ_{ij} with probability p_{ij} associated and generate a bootstrap sample containing as many observations as the original sample. We repeat this N times to generate a space of N samples with M observations each (of estimates and independent variables). In a second stage we apply bootstrap smoothing to each of the N replicas. To do so we draw K bootstrap samples for each of the N replicas and in each case choose an optimal model using the mean square error[9]. We then compute statistics of interest and their standard errors from the optimal model. This yields K estimates of the desired statistic for each of the N samples from which we compute bootstrapped mean and variances. Finally, in the third stage, we compute averages and standard deviations for the desired statistics across all N replicas. These estimates and their

[8] Our application requires computation of predicted values for each point or observation (country-year). When applying bootstrap smoothing to compute predicted values we will use conventional bootstrapped errors rather than those recently proposed by Efron (2014) that adjust for model discontinuities. Our aim here is not to produce optimal standard errors for predicted values of individual observations (as is the case in bagging in general and Efron's approach in particular) but rather to illustrate the consequences of facing jointly model and parameter uncertainty when the target statistic is sample-based.

[9] The competing models we choose for the application involve 3 parameters so there is no need to use penalized measures of fit.

standard errors reflect and contain information on uncertainty about both mortality parameters estimates and models.

Section 6 describes an application. It shows that conventional, naive analysis that ignores uncertainty of adult mortality estimates and/or models underestimates standard errors and could lead to excessively liberal hypotheses testing. The application deals with the classic problem in population health about the magnitude of the contribution of improvements in economic conditions, on one hand, and public health and medical technology, on the other, to secular mortality declines.

6 The role of living standards and medical technology

In this section we draw from the long-running discussion about whether the secular mortality decline was driven by changes in standards of living and well-being, public health interventions or diffusion of medical knowledge and technology. The first formal rendition of the controversy was developed by Preston (Preston 1980, 1976) and occupies a central place in the history of theories of health and mortality and is a key piece in any theory of the evolution of human mortality (special issue of International Journal of Epidemiology, 2005, vol 34 and 2007,vol. 36) Our goal here is much less ambitious that to add to it, much less to resolve it. The aim is to show that inferences from empirical data rest on less solid grounds than suspected, that consideration of uncertainty may alter inferences and should play an more central role in the discussion than they have so far[10].

6.1 Competing explanations

Sometime after 1930 mortality decline accelerated in high income countries. The timing of the acceleration occurs later in low income countries, in general, and LAC countries in particular. The most accepted accounting of mortality decline in England and Wales since 1870 (McKeown 1976) singles out improvements in nutrition and, more generally, economic well-being as the main determinants of changes. The diffusion of medical technologies (vaccination, antibiotics, sulfa drugs) could not possible have played a role. But they did so after 1930. Preston's classic piece and extensions (Preston 1976, 1980) suggest that this is indeed the case and that the bulk of post-1930 life expectancy gains is associated with advances in medical technology, that is, that levels of life expectancy post 1930 could not have been attained if driven by shifts in economic well-being alone. Following the seminal paper by Preston there have been discussions and new empirical contributions sometimes supporting the main claim and sometimes disputing it. In particular, work by historians (Szreter 1988) argues that hidden or ignored in the discussion is the important role of public health and of the political and institutional shifts needed for public health to exert a powerful influence. Although historians addressed the situation in pre and post industrial England, they apply equally well to low income countries.

A key component of the discussion is the empirical estimation that aims at disentangling the contributions of improvements in levels of well-being and standards of living, on one hand, and those of public health and medical technology, on the other. Estimates of these contributions are sensitive to the composition (by country and time periods) of the sample of countries and time periods included in the analyses as well as to the nature of the models estimated. To our knowledge, the role played by the accuracy of estimates of the main variables has not been part of the controversy. We show below that considering both model and estimate's uncertainty should be part and parcel of the discussion, even in the simple case we choose to illustrate here.

[10] This section provide a very brief, highly stylized summary of the problem. We stripped nuances and details, all very important, and presented only a skeleton that is useful to support the application.

6.2 Competing models

We propose two assess the performance of competing models, the original one estimated by Preston and a generalization formulated by (Palloni and Wyrick 1981). The key dependent variable is life expectancy at birth, $E(i,t)$, and the models include only one independent variable, adjusted GDP per capita. The two models we employ, logistic and Box-Cox, are as follows[11]

$$\begin{aligned} E(i,t) &= \alpha/(1+\exp(\beta(GDP(i,t)-\gamma)) \qquad (6.1) \\ \frac{E(i,t)^{\lambda}-1)}{\lambda} &= \mu+\varphi\ln(GDP(i,t)). \qquad (6.2) \end{aligned}$$

Given a sample of observed values $(E(i,t), GDP(i,t))$ $(t = 1900, \ldots, 2010)$ we estimate both models using NLS on two subsamples: all observations before 1950 (subsample 1 for period T_1) and all observations after 1950 (subsample 2 for period T_2). To estimate the contribution of medical technology and public health to changes in life expectancy experienced between the two periods we follow the classic shift analysis.[12] We first estimate the models' parameters separately in the two periods[13] and proceed as follows:

- Compute quartiles of GDP per capita in subsample 1, $(\{Q_{11}, \ldots, Q_{14}\}$ and in subsample 2, $\{Q_{21}, \ldots, Q_{24}\}$;
- Compute predicted values for each quartile in subsample 1, $\{P_{11}, \ldots, P_{14}\}$ and in subsample 2 $\{P_{21}, \ldots, P_{24}\}$;
- Compute predicted values for each quartile in subsample 1 using parameters estimated in subsample 2, $\{\pi_{11}, \ldots, \pi_{14}\}$;
- Compute predicted values for each quartile in subsample 2 using parameters estimated in subsample 1, $\{\pi_{21}, \ldots, \pi_{24}\}$;
- Compute the ratios $\delta_{1j} = (\pi_{1j} - P_{1j})/(P_{2j} - P_{1j})$ and $\delta_{2j} = (P_{2j} - \pi_{2j})/(P_{2j} - P_{1j})$ [14]
- Finally, compute the target statistic in the sample as the mean

$$\Delta = (1/8) * \sum_{j=1}^{4} (\delta_1(i) + \delta_2(i)) \qquad (6.3)$$

or, alternatively, as the median value of $(\delta_1(i) + \delta_2(i))/2$ across the four quartiles.

[11] Neither the Box-Cox nor the logistic functions are the most general models. The Box-Cox model defined here is restricted in that we fix ex-ante the functional form of the independent variable (log form) rather than searching for a (best fitting) transformation. The logistic function is a variant of the originally estimated by Preston and we also consider it below when dealing with model uncertainty.

[12] The legitimacy of the shift analysis and of the inferences drawn from it are another point of contention in this controversy. To limit the number of model that come into play, we choose to ignore the issue in our application.

[13] There are countries that will appear multiple times in both subsamples. To simplify estimation we do not adjust estimates for clustering effects that result from this.

[14] If the relation between mortality and GDP shifts due to medical technological improvements, both these quantities should be positive.

Inevitably, this choice of models plays against the role of model uncertainty since they represent a very narrow class of models consistent with the theories that could be brought into play. Neither includes covariates other then GDP and the functional forms fall within a rather restrictive range. It follows that our application may underplay the role of model uncertainty.

6.3 Data and Methods

6.4 Data sources

We use the Latin American Mortality Database LAMBdA (http://www.ssc.wisc.edu/cdha/latinmortality/) created to support the empirical study of the history of mortality trends in Latin American countries after independence. The database documents the period between 1848 and 2014 and contains about 500 life tables. The data to compute life tables for the period 1930-2010 are age-specific death rates adjusted for relative completeness of census enumeration and death registration as well as for age overstatement at adult ages. These life tables are available yearly. Life tables for most countries before 1930 are based on the application of generalized stable population to partial vital statistics and census information. We choose a maximum of 25 different estimates of life expectancy at birth for every five-year period between 1900 to 2015 for a total of 19 countries. Not all country-years contain all estimates; in addition, periods before 1950 include estimates computed with methods that are not needed nor used in the period after 1950.

To generate unconditional probabilities of estimates being within the 5 percent range we followed two rules. The first rule applies to estimates computed using two methods that apply only to ages over 5. For these (and for other methods we do not use in this paper) we have precise estimates retrieved from extensive simulations (Palloni et al. 2015). The only difficulty is that to calculate life expectancy at birth we use estimates of mortality below age five retrieved from other sources (DHS, WFS) for which we do not have guidance from simulations. Instead, we assigned to each country-year the probability corresponding to estimates of adult mortality, a rule that does not downplay errors only if the accuracy of mortality estimates below 5 is at worst as good as the estimates of adult mortality.[15]

The second rule applies to estimates form third parties or from methods that have not been scrutinized and verified with numerical simulations. Most of these apply to the period before 1950. In these cases we assigned probabilities according to our own judgment about the quality of data used and the sensitivity of methods to violations of two key assumptions on which all of them rely, accuracy of (adjusted for migration) intercensal rates of growth and degree of certainty regarding model mortality patterns.[16]

To replicate the results that would obtain from a naive approach ignoring uncertainty of estimates we construct three benchmark data sets. The first and second are composed of the medians and (unweighted) means of each country-year set of estimates. The third data set contains one randomly drawn estimate from each country-year set. This choice of benchmarks for comparison is an admittedly easy way to solve a complex issue. Indeed, life expectancy time trends built from medians (or means or randomly selected) estimates may not be the optimal choice for any researcher

[15] This is very likely to be the case since estimates of mortality below age 5 depend on multiple survey data and a non-parametric procedure that return yearly estimates. The only errors that could come into place are associated with sampling and the fitting algorithm.

[16] We are keenly aware that our assignment of probabilities of accuracy of these methods is not the best. In an ideal situation we ought to use simulations and expert judgment to assess the concordance of methods' assumptions and observed conditions. For our objective in this paper, however, it suffices to use an assignment rule that roughly approximates what an average expert judgment would conclude.

who ignores uncertainty. Before choosing a credible observed time trend an industrious researcher explores its nature, performs consistency checks and discards those that appear to be implausible. Our benchmark data could contain some implausible time trends that would be discarded in any analysis, regardless of whether or not the researcher admits uncertainty. To the extent that this is so the evaluation of the impact of uncertainty discussed here is weakened. There is no straightforward way to handle this problem without experimentally replicating (many times) what an analyst uninterested in uncertainty would chose to do.

6.5 Features of the data

We start out with a large data base including 19 countries with yearly estimates of life expectancy from 1850 to 2015. We draw a total of 382 observations or country-year points choosing country-specific estimates centered in the middle of each five-year period during the interval 1900-2015. Each observation (country-year) contains a variable number of estimates that can be as low as 1 and as high as 25 and each one of these is associated with a probability of being within 5 percent of the target parameter.

Figures 1a-1b display the distribution of available estimates in the data set. About 25 percent of observations for the period before 1950 (n=155) contain a single estimate and the median number is 6. In contrast, nearly 85 percent of the post 1950 observations include at least 2 estimates with a median of three per case. This is consistent with the fact that vital statistics before 1950 are weak almost everywhere in the region and that the number of feasible methods to generate estimates during that period is more heterogeneous than for the period after 1950.

Figure 2a-2b display distribution of the ratio of the interquartile range of estimates of life expectancy to the median estimate. This statistic is analogous to the coefficient of variation but uses the median instead of the mean and the interquartile range instead of the standard deviation. The contrast between Figure 2a and 2b is stark: whereas the statistic's distribution in the most recent period is left-skewed (low variance of alternative estimates), the distribution for the earlier period is right-skewed and has a considerably thicker tail. In fact, the median value of the statistic before 1950 is about 0.20 whereas for the more recent period is as low as 0.05.

Figures 3a and 3b show the distribution of values for the probabilities associated with estimates. For the period before 1950 the values are spread out over a wide range whereas values for the period after 1950 are highly concentrated in the very small, extreme valued range 0.80-0.90. This contrast reflects well the much larger uncertainty associated with estimates for the earlier period when the data are defective, there are multiple estimates, and all of them rely on assumptions that are difficult to ascertain. Instead, estimates for the post 1950 period are associated with a small set of methods for which we were able to assess errors associated with violation of assumptions (via simulation) and the assumptions on which these methods are based are fewer and are more easily adjudicated.

Figures 4a and 4b display the median estimate and interquartile range in two countries, Uruguay and Honduras. The former is known to have high quality vital statistics reflected in small interquartile ranges, particularly during the most recent period. Instead Honduras represents well conditions found in a subset of about 10 countries that have poor vital statistics and where the uncertainty is much larger, particularly during the earlier period.

This brief summary of descriptive statistics confirms that the variability of alternative estimates and associated probabilities is larger for the earlier than for the later period and that there are sharp inter-country contrasts in the magnitude of uncertainty of estimates in both periods. This is as expected for the quality of vital statistics has improved over time but continues to differ sharply across countries and the methods to estimate mortality for the earlier period are based on more rigid assumptions whose validity is more difficult to ascertain.

6.6 Three stage estimator

To compute the three-stage estimator of our target statistic, Δ, we first implement the first stage and create N bootstrapped replicas of the set of estimates for country-years. The sampling is done with replacement with weights proportional to the probabilities of errors associated with each estimate. In each bootstrap sample we then generate K replicas (with replacement) and use each of these to estimate both models, choose the optimal one, and compute in each case the statistic of interest Δ. We then compute the average of Δ and its standard error across K replicas. The estimates account for model but not for mortality estimate uncertainty. Finally we repeat the procedure in each of the N bootstrapped replicas. The result will be a distribution of Δ values that account for all the uncertainty associated with models and mortality estimates. In a final step we compare these results and associated inferences with the estimate and inferences one would obtain had we ignored both model and estimates uncertainty.

6.7 Results

6.7.1 Demographic parameter uncertainty

The first two panels of Table 1 display parameters of best fitting models using the median of estimates for observations before 1950 (Box-Cox with shape parameter (lambda) fixed at 0 or log form) and after 1950 (logistic model). These estimates are obtained ignoring uncertainty of demographic parameters and will be benchmarks for comparisons. To assess the impact of uncertainty on model estimate we implement the first of the three stage procedure suggested above. For the period before 1950 the relation GDP and life expectancy is best captured by models within the family of Box-Cox transforms. Figure 5a displays the empirical bootstrap distribution of the Box-Cox parameter (lambda) from the best fitting model in each replica. It is a highly skewed distribution with a median of about 0.20 and an interquartile range close to 0.25. This reflects a rather large margin of uncertainty due to variability of estimates. Had one estimated the model using the median or mean of estimates or a randomly chosen value, the estimates of lambda would be 0, 0.03 and 0.31 respectively. However, the central parameter of the model is not lambda but the slope of life expectancy relative to log of GDP. Figure 5b display its empirical distribution. Not surprisingly it too is highly skewed and with a rather large interquartile range (0.40) relative to the median (0.42). The third panel of Table 1 shows key parameters of the bootstrapped replicas.

Ultimately, what is of interest is the elasticity of life expectancy relative to GDP. When lambda is close to 0 the elasticity is the estimated slope of the regression of the log of the variables. When lambda is larger than 0, the elasticity changes with values of the dependent variable. To provide a sense of magnitude behind the uncertainty levels we calculated numerically estimated elasticities using slopes within the interquartile range, the corresponding estimate of lambda and the median life expectancy during the period. We obtain elasticities as small as 0.05 and as large as 0.30, values reflecting an extreme case where life expectancy at birth is insensitive to GDP and one where a 10 percent increase in GDP can increase life expectancy by 3 percent.[17]

A different way of making the same point is this: a naive analyst interested in testing the null hypotheses that GDP does not influence life expectancy (slope=0) would proceed with a *t-test* using

[17] The estimated elasticity is given by $\varepsilon = \beta/Y^{\lambda}$ where β is the estimated slope and λ, the Box-Cox parameter. We assume a value of Y equal to the mean of life expectancy before 1950.

the estimated slope from the log form of the model and the sample of median estimates. From the first panel of Table 1 the t-statistic is computed as $t = 0.23/0.033 = 6.96$ and the researcher would reject the null hypothesis in favor of the alternative that GDP does indeed influence mortality levels. However, if one accounts for the estimates' variance associated with uncertainty and uses the variance of estimates from the bootstrap (third panel of Table 1), the t statistic turns out to be $t = 0.53/0.42 = 1.43$, not large enough to reject the null hypothesis.[18]

For the most recent period the best fitting model is the logistic. To assess the role of uncertainty of estimates we evaluate the empirical distribution of the threshold and slope parameters. These are shown in Figures 6a and 6b respectively. Key parameters of the bootstrap are in the fourth panel of Table 1. The range of the threshold parameter is quite small, 75-77, and the empirical bootstrap is centered close to 76. The naive estimate using medians comes close to 75, somewhat to the left of the empirical distribution. Figure 6b displays the empirical distribution of the logistic slope parameter. Its median value is 0.053 and the interquartile range is 0.0048, or about 7 percent of the median value. This does not seem to be large enough to derail inferences. And it does not: if we ignore variability due to uncertainty of estimates we would use the parameter estimate from a model fitted to the median of estimates (0.0004), the estimated standard error (0.0001) (second panel of Table 1) and conclude that it is legitimate to reject the null. If we use instead the adjusted standard error the t-statistic is about 4.5 (from fourth panel of Table 1) and the null hypothesis can be rejected. In this case, uncertainty of estimates does not lead to misleading inferences.

We also compute numerical values of the elasticity of life expectancy relative to GDP at the first, second and third quartiles of GDP's empirical distribution (results not shown). The mean value is 0.20 and its (bootstrapped) standard deviation is 0.030. A simple t-statistic would conclude that the elasticity GDP is significantly different from 0 also in the post 1950 period.

In summary, and as should be expected, uncertainty of demographic parameters plays an influential role in model parameter selection during the period when vital statistics and censuses where most defective and estimates of mortality are too heterogeneous and depend on assumptions that are difficult to verify. It plays a secondary role in the most recent period.[19]

6.7.2 Model uncertainty only

Model uncertainty alone, independently of demographic parameter uncertainty, may also play an important role. To illustrate this we choose the median estimate from the set associated with each country-year. We then apply the method proposed for the second stage (see above) and compute the empirical distributions of estimates and their standard errors using K sample replicas of the median estimates. As an illustration we use separately the two subsamples and in each case focus on three competing models, two logistic functional forms and Box-Cox.[20] For the period before 1950 all 500 replicas lead to selection of a Box-Cox model with scale parameter equal to 0. In this case, then, there is no variance due to model (at least within the class considered here). In the period after 1950 things are quite different: all three models perform best in some of the bootstrapped

[18] According to the third panel of Table 1 the mean value of the estimate from the bootstrapped distribution is 0.53, the standard error of the bootstrapped sample is $s(bs) = 0.40$, and the mean of the estimates standard errors across bootstrapped samples is $\widehat{s(bs)} = 0.12$. We then compute an adjusted standard error as $s(adj) = 0.42[(s(bs)^2 + \widehat{s(bs)^2}]^{1/2}$.

[19] We hasten to emphasize that this conclusion holds given the models and target demographic parameters used in the application. Other parameters of the life table, for example life expectancy at age 60 or adult Gompertz slope, are subject to higher levels of uncertainty both before and after 1950. An application using these parameters as dependent variables may well yield very different conclusions than those reached here.

[20] As stated before (see footnote 11, the second logistic model we focus on was originally estimated by Preston.

samples. To visualize the impact of model uncertainty we computed the numerical value of the GDP elasticity of life expectancy relative at the median of GDP[21]. Figure 7 displays the frequency distribution of the statistic. Plainly model uncertainty influences choices of estimates as the range of values is quite large. Even when considering the means of the distributions corresponding to each optimal model the boundaries of the range [0.02-0.20] differ by a factor of ten. Averaging our estimates across bootstraps and using the estimated variance leads to acceptance of the null hypothesis of no GDP effects during the period, an inference that is at odds with consensus and the naive analysis based on the median of estimates.

6.7.3 Model and parameter uncertainty

We now evaluate the patterns of results of models' statistics under conditions characterized by uncertainty of demographic parameters and models. First, we focus on model-specific parameters (not strictly comparable across models), that is, the slopes in the Box-Cox and logistic models. We compute the median value of the parameter of interest across all K=100 replicas when the model was optimal. The empirical distribution of these median values (one for each of the N=1000 bootstraps) is then estimated across all 1000 bootstraps.

We then study patterns of estimates of GDP elasticity of life expectancy, a quantity that, unlike model-specific parameters, is comparable across the models we study. In the case of a Box-Cox with shape parameter (lambda) equal to 0, the elasticity is simply the slope of the regression of the log form of the variables. In the logistic or when the Box-Cox model has a shape parameter different from 0 we calculate the elasticity numerically at various points of the observed distribution of GDP. Finally, we assess the behavior of the target statistic, Δ, the fraction of total changes in life expectancy between the two periods that is attributable to (unmeasured) improvements in medical technology.

Figure 8a displays frequency distributions of Box-Cox slopes for years before and after 1950. These distributions are computed over N=1000 bootstrap samples using the median parameter values of the Box-Cox models when this was the optimal choice over K=100 replicas. The distribution for the period after 1950 reflects little heterogeneity, a result of both small levels of uncertainty of demographic parameters and little or no uncertainty regarding the Box-Cox shape parameter (concentrated close to 0). The distribution of estimates in the subsample for the years before 1950 has a much larger variance attributable to higher levels of uncertainty in both estimates and optimal Box-Cox shape parameter. This is confirmed in Figure 8b displaying the distribution of the median value of the shape parameter for replicas in which Box Cox was the optimal model. The spread of the shape parameters is solely due to uncertainty in the estimates of mortality. Note that the results we obtain here are just the opposite of those we produce when working with the median of estimates (see 6.7.2).

The next figure (Figure 8c) shows the distributions of median values of the logistic slope when the logistic optimal model in a particular replica among the K possible replicas for years before and after 1950. The variance in each case is a reflection of demographic parameter uncertainty and is, as before, much larger in the subsample for earlier years.

In sum, as stated before, when one fixes the model (in this case to the optimal) the parameters of interest have distributions whose variance is influenced by parameter uncertainty. In the particular case of the Box-Cox model, where different shape parameters represent different models, parameter uncertainty leads to model uncertainty.

[21] We also computed elasticities at the first and third quartiles of GDP and obtained very similar results.

Figures 9a and 9b show the distribution of GDP elasticity of life expectancy obtained from both subsamples in the optimal models chosen among three candidates, Box-Cox and two logistic models (classic and standard). The spread of all three distributions in each figure can only reflect uncertainty due to mortality estimates whereas the range of values across all three distributions reflects model uncertainty. The elasticities computed from the new and classic logistic models range from about 0.2 to about 0.7 in the earlier years and from 0.1 to 0.2 in the more recent period. The sharp contrast is again due to more significant uncertainty in the years before 1950. However, model uncertainty is equally significant in both periods: the statistics differ by a factor of at least ten across Box-Cox and logistic models. Thus, if one is interested in inferences about changes in elasticities over time, we should be aware that model and parameter uncertainty have very large influences and should be accounted for.

Figure 9c displays the distribution of delta, the target parameter interpreted as the fraction of changes in life expectancy between two periods attributable to improvements in medical technology. Because the statistic is computed simultaneously using two optimal models, one for the period before 1950 and the other for years after 1950, the spread of the distribution reflects joint uncertainty of models and parameter estimates. The distribution if concentrated around a median of 0.71 with a rather small standard error (0.05). However, if we follow the strategies implemented in the past and compute delta using the best fitting models in the two subsamples we obtain an estimate of delta equal to 0.93 with a tiny standard deviation (0.01). Thus, whether we follow the naive route or we consider uncertainty, the inference would have been the same, namely, shifts in mortality regimes due to forces exogenous to changes in living standards is massive. The details of the shift, however, are different as the estimates of delta differ on average by about 30 percent and the standard errors of the estimates differ by a factor of five.

7 Discussion

Our aim in this paper is a modest one, namely, to illustrate the consequences of accounting for uncertainty of demographic parameters computed with multiple techniques, all of which rely on assumptions of variable rigidity. In addition, we superimposed model uncertainty, even though our treatment stayed within the somewhat limited confines of a frequentist approach.

We describe a framework to deal with parameter uncertainty in cases where the researcher can rely on extensive simulations that generate numerical estimates of biases when combinations of assumptions are violated. Measures of errors distributions are, however, insufficient. To use them properly one requires probabilities that one, several, or multiple combinations of assumptions are being violated in any particular historical situations. These empirical probabilities could be retrieved from two sources. The first are specially designed diagnostics tests returning statistics whose values can be easily translated into approximations to probabilities that individual or combination of assumptions are met. The second source is expert judgment about the empirical conditions to which the demographic parameter applies. Our application rests on probabilities assigned by us rather than on probabilistic judgments issued by a sample of suitably chosen experts.

In this paper we also handled a second class of demographic estimates, namely, those for which there is no extant numerical evaluation of the relation between estimates' biases and inconsistency, on one hand, and assumptions, on the other. These include, but are not limited to, third party estimates that are fully documented, that is, cases where the producer of estimates discloses faithfully the data and the techniques employed to compute the target parameter. The set of estimates pertaining to this class we use in the paper were all properly defined and could be replicated by any knowledgeable demographer. However, although we were able to ascertain and evaluate the realism of the suite of assumptions on which the estimates were based, we could not assign precise probabilities of errors when single or combination of assumptions are violated. As a result, ours remains a somewhat defective evaluation of the sensitivity of inferences to, in this case, mortality parameter uncertainty.

We handle model uncertainty in a rather cursory way, without relying on BMA but, instead, tailoring a frequentist approach to model selection. The bootstrap-method employed here is not originally designed to select between models but has a more pedestrian goal, namely, to generate more precise standard errors of individual predictions under model uncertainty. Instead, our application deviates from this goal and turns the method into a tool to generate estimates of based-sample statistics (model parameters and quantities computed from predicted values) and their standard errors rather than individual predictions. This may be somewhat problematic for a couple of reason. First, the statistics we use as target quantities (slopes, elasticities, predicted fractions) are averages over a large number of observations and, as a consequence, may conceal more than what they reveal about implications of model selection. For example, it could well be the case that some mortality parameters, such as adult mortality slopes, are more sensitive both to parameter and model uncertainty and that the distributions resulting from model and parameter heterogeneity have much larger variances than we were able to ascertain for a coarser statistic. The second reason our approach is not fully satisfactory is that there are no systematic studies showing the performance of model selection with the bootstrap and no way to distinguish between situations where the competing models can be clearly discriminated and situations where model boundaries are too fuzzy to proceed with model selection[22]. Thus, our desire to simplify the task of posing alternative models from which to choose, for example, may have derailed out efforts since the "bagging" tool we employ may be an effective one to discriminate between models that are discordant in some profound away but not at all when the models belong to the same class of exponential families. Or, it could be that the bootstrap is an efficient tool when models are nested and differ in terms of number of predictors, as is the case in the original applications by its proponents, but not in other situations. Instead in this paper we considered one independent variable and two or three models with the same number of parameters.

Despite the above limitations, the exercise performed here leads to four conclusions that are unlikely to vary much even if one pursues a different approach for evaluation. First, uncertainty of demographic parameters matters quite a bit when deciding between alternative models. In our case, the rather large amount of uncertainty plaguing pre-1950 mortality estimates resulted in conflicting statistical tests and contradictory inferences about the role of GDP. The main source of confounding is the much larger variance of the target estimate than one would use in a naive analysis that ignores uncertainty. The excess variance is all due to uncertainty of demographic parameters. Second, model uncertainty also matters and significantly influences interpretation of underlying processes. In our example, estimates of the post-1950 GDP elasticity of life expectancy varied wildly depending on the optimal model. If estimates are averaged over models, the key inference leads to negate effects of GDP that are surely there. Third and unsurprisingly, combining parameter and model uncertainty complicates inferences. Finally, neither model nor parameter uncertainty mattered very much for assessing the magnitude of the shift in the relation between GDP and life expectancy. It turns out than in countries we examine here the shift of the relation is so massive that even large variance of estimates in one period will not mislead the investigator into misidentifying the role of medical technology. Yet, the differences in the magnitude of estimates is

[22] This includes but goes beyond the choice of criterion to discriminate between competing models.

quite large and their standard errors much larger than when using a naive approach. This result is both cause for consolation but also for caution: the target statistic that adjudicates between theories that advocate the importance of medical technology and those that emphasize the role of living standards is coarse, computed over the entire sample, and will not reveal, as could examination on individual predictions by country year, subtle differences implied by alternative models and parameter uncertainty.

Future work should proceed along two routes. First, a contrast between results based on frequentist and Bayesian approaches is needed. The obvious choice would be to use in jointly BMA and the bagging tool and proceed to carry out extensive and detailed comparisons. Second, we should explore better the conditions that magnify (attenuate) the role of uncertainty of demographic parameters. Do they matter as much, more, or less when we are estimating country-specific trends? When projecting or forecasting population totals or age distributions? How much more relevant can it be when one estimates demographic parameters that depend on fine-tuned quantities, such as age specific mortality rates, where errors may be considerably more influential?

References

Alkema, L., Raftery, A., Gerland, P., Clark, S., and Pelletier, F. (2008), "Estimating the total fertility rate from multiple imperfect data sources and assessing its uncertainty," .

— (2012), "Estimating trends in the total fertility rate with uncertainty using imperfect data: examples from West Africa," *Demographic Research*, 26, 331–362.

Breiman, L. (1996), "Bagging predictors," *Machine Learning*, 53, 123–140.

Buckland, S., Burnham, K., and Augusin, N. (1997), "Model selection: anintegral part of inference," *Biometrics*, 53, 603–618.

Buja, A. and Stuetzle, W. (2006), "Observations on bagging," *Statistics Sinica*, 16, 323–351.

Efron, B. (2014), "Estimation and accuracy after model selection," *Journal of the American Stattistical Association*, 109, 991–1007.

Gelman, A. and Rubin, D. B. (1995), "Avoinding model selection in Bayesian social research," *Sociological Methodology*, 25, 165–174.

Gerland, P., Raftery, A., Sevcikova, H., and Li, N. e. a. (2014), "World population stabilzation unlikely this century," *Science*, 346, 2343–237.

Hjort, N. and Claeskens, G. (2003), "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

Kaplan, D. and Chen, J. (2014), "Bayesian model averaging for propensity score analysis," *Multivariate Behavioral Research*, 49, 505–517.

Madigan, D. and Raftery, A. (1994), "Model selection and accounting for model uncertainty in graphical model using Occam's window," *Journal of the American Statistical Association*, 89, 1535–1546.

McKeown, T. (1976), *The Modern Rise of Population*, New York: Academic Press.

Palloni, A., Pinto, G., and Beltrán-Sánchez, H. (2015), "Estimation of Life Tables 1850-2010: Adjustments for Relative Completeness and Age Misreporting," .

Palloni, A. and Wyrick, R. (1981), "Mortality decline in Latin America: changes in the structure of causes of deaths, 1950-1975," *Social Biology*, 28, 187–216.

Preston, S. (1976), *Mortality patterns in national populations: with special reference to recorded causes of death*, New York: Academic Press.

— (1980), "Causes and Consequences of Mortality Declines in Less Developed Countries During the Twentieth Centrury," in *Population and Economic Change in Developing Countrie.*, ed. Easterlin, R., Chicago, IL: University of Chicago Press, book section 5, pp. 289–360.

Raftery, A. (1996), "Approximate Bayes factors and accounting for model uncertainty in generalized linear models," *Biometrika*, 83, 251–266.

Raftery, A., Li, N., Sevcikova, H., Gerland, P., and Heiling, G. (2012), "Bayesian probabilistic population projections for all countries," *Proceedingf the National Academy of Sciences*, 109, 13915–13921.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.

Raftery, A., Madigan, D., and Volinsky, C. (1999), "Bayesian model averaging: A tutorial," *Statistical Science*, 14, 382–417.

Sala-i Martin, S., Doppelhoper, G., and Miller, R. (2004), "Determinants of long-term growth: a Bayesian model averaging of classic estimates (BACE) approach," *The American Economic Review*, 94, 813–835.

Szreter, S. (1988), "The importance of social intervention in Britain's mortality decline:c.1850-1914: A reinterpretation of the role of public health," *Social History Medicine*, 1, 1–38.

Wheldon, M., Raftery, A., Clark, S., and Gerland, P. (2013), "Reconstructing past populations with uncertainty from fragmentary data," *Journal of the American Statistical Association*, 108, 96–110.

Figure 1a: Frequency distribution of available estimates

(years before 1950)

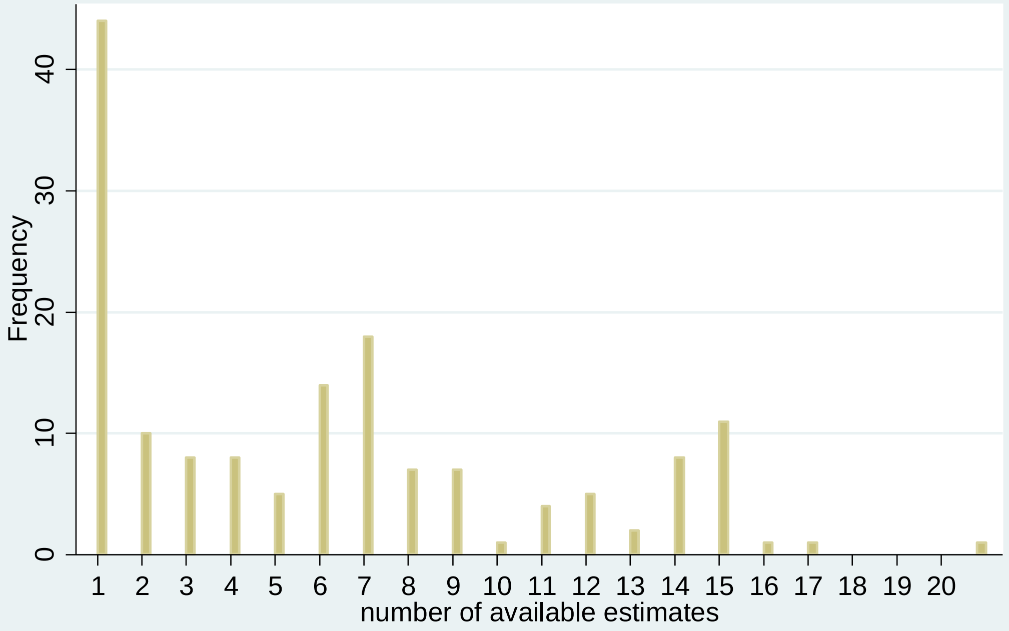
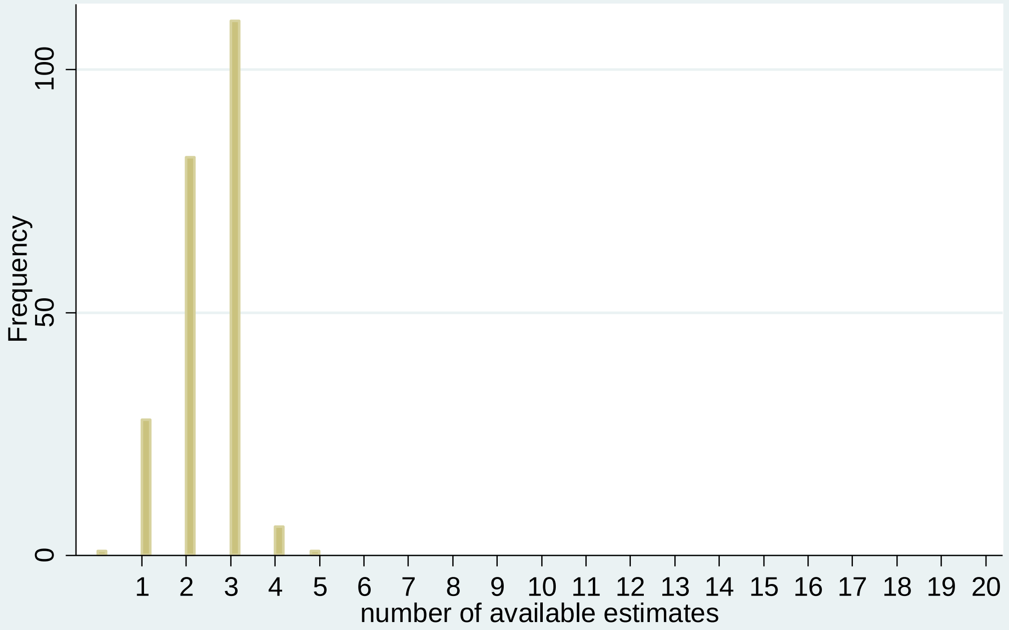


Figure 1b: Frequency distribution of available estimates

(Years after 1950)



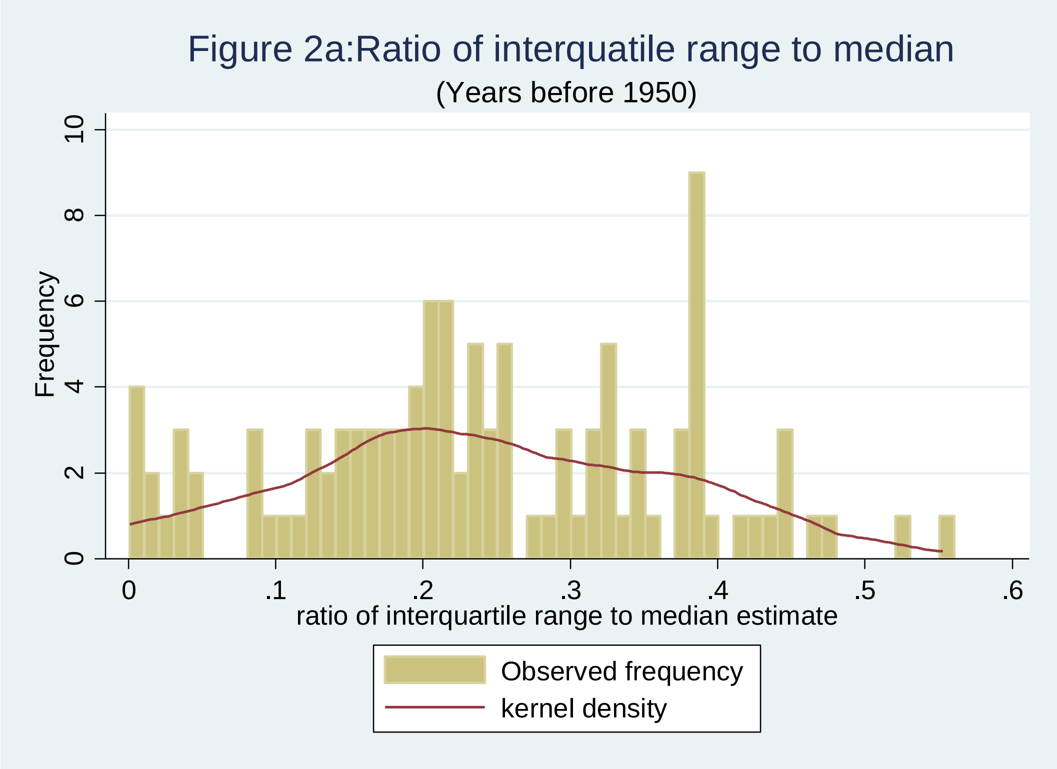


Figure 2a:Ratio of interquatile range to median

(Years before 1950)

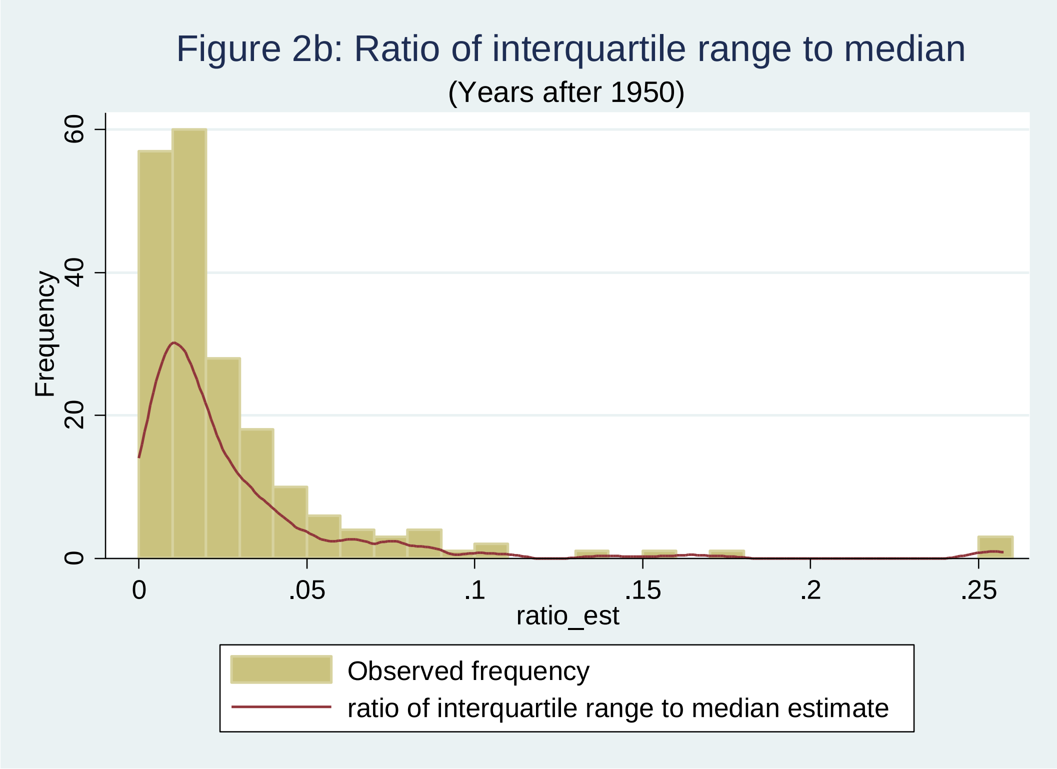


Figure 2b: Ratio of interquartile range to median

(Years after 1950)

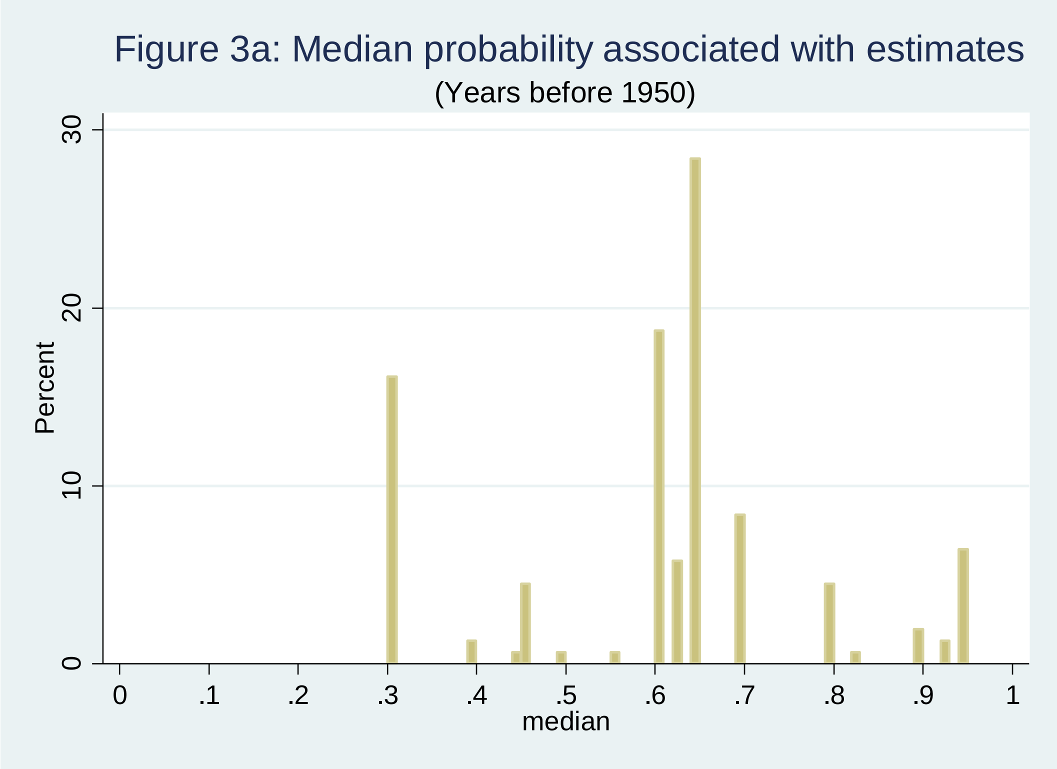


Figure 3a: Median probability associated with estimates

(Years before 1950)

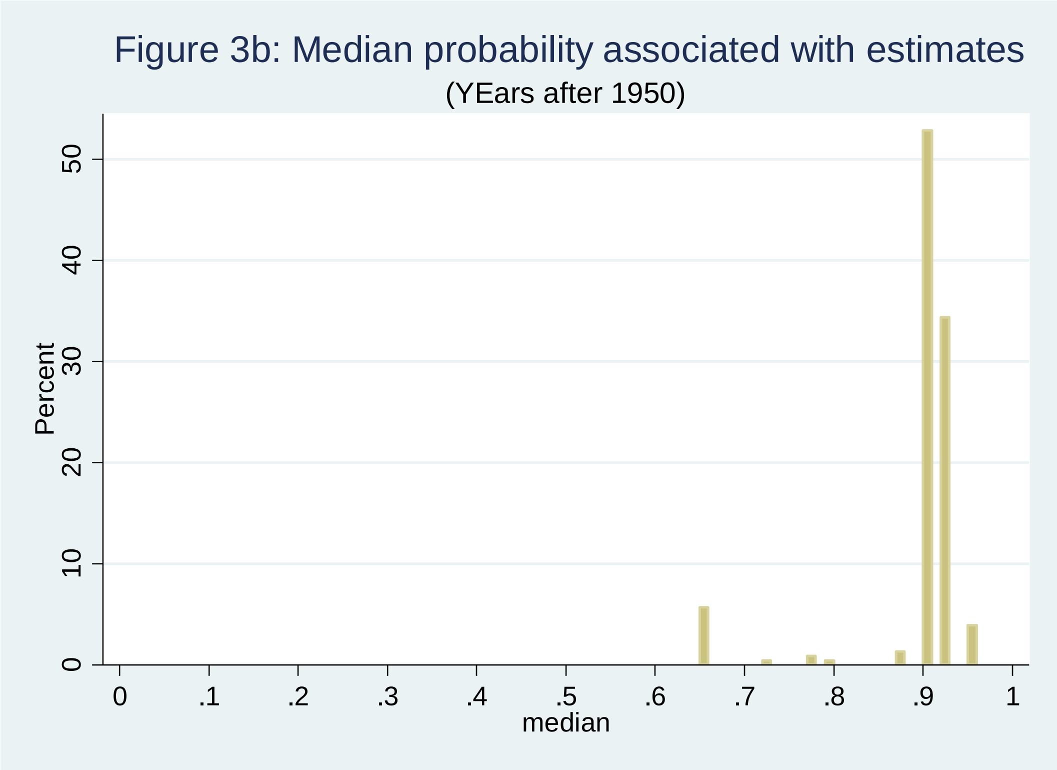


Figure 3b: Median probability associated with estimates

(YEars after 1950)

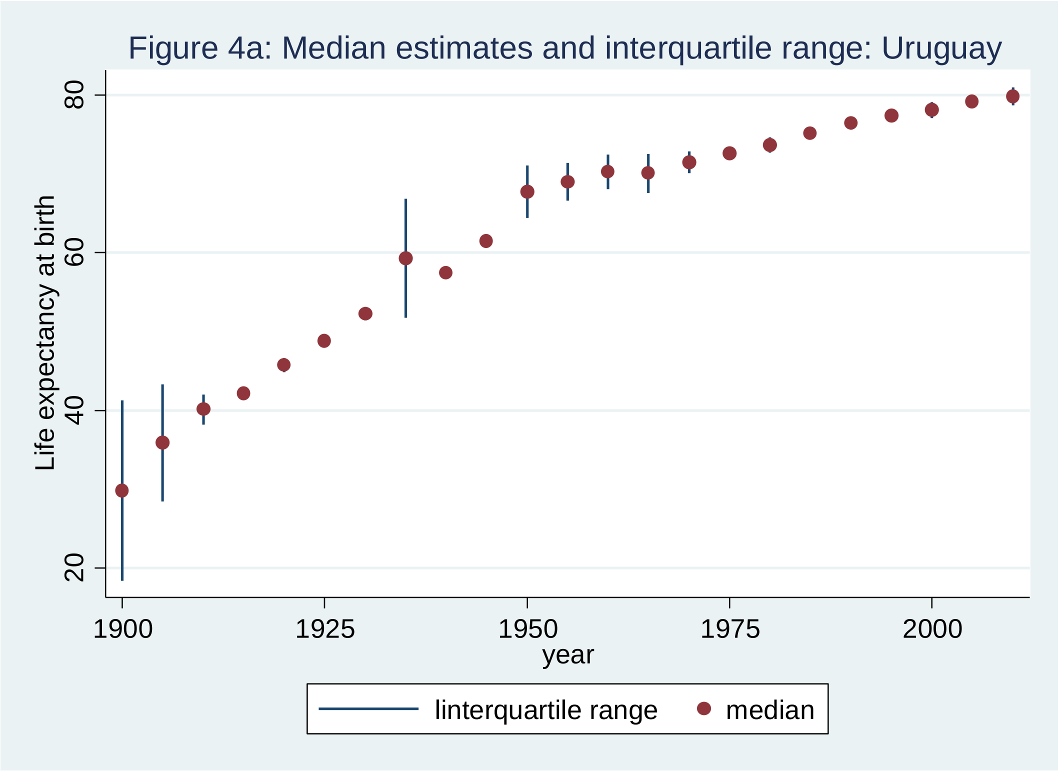


Figure 4a: Median estimates and interquartile range: Uruguay
Life expectancy at birth
20
40
60
80
1900
1925
1950
1975
2000
year
linterquartile range
median

Figure 4b: Median estimate and interequartile range: Honduras

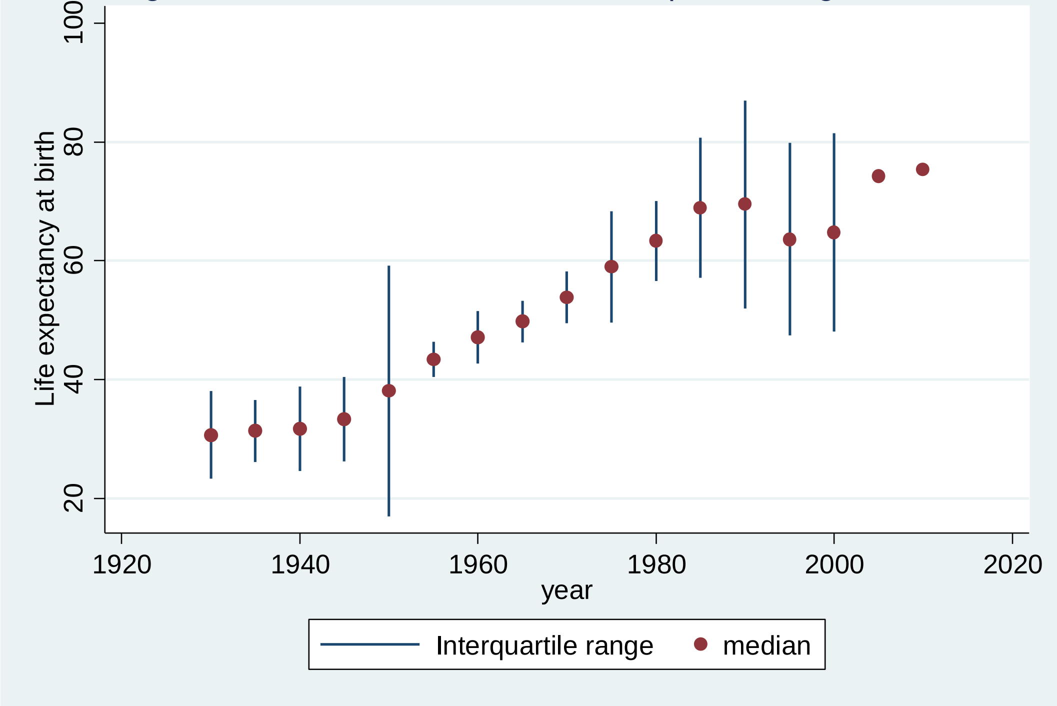


Figure 5a: Frequency distribution of Box-Cox parameter

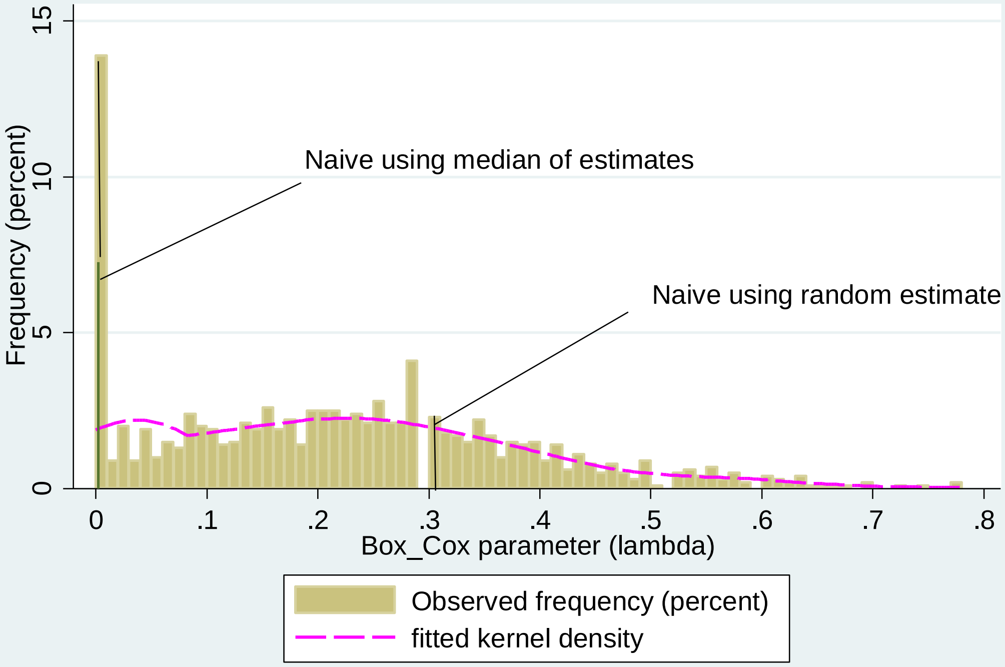


Figure 5b: Estimates of slopes in best fitting models

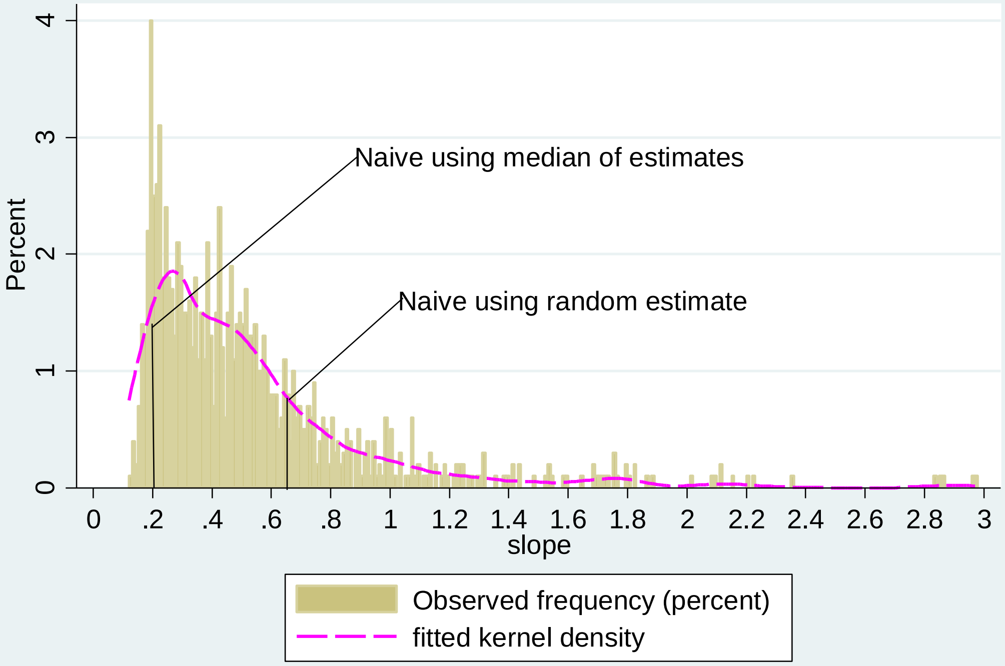


Figure6a: Frequency threshold logistic parameter

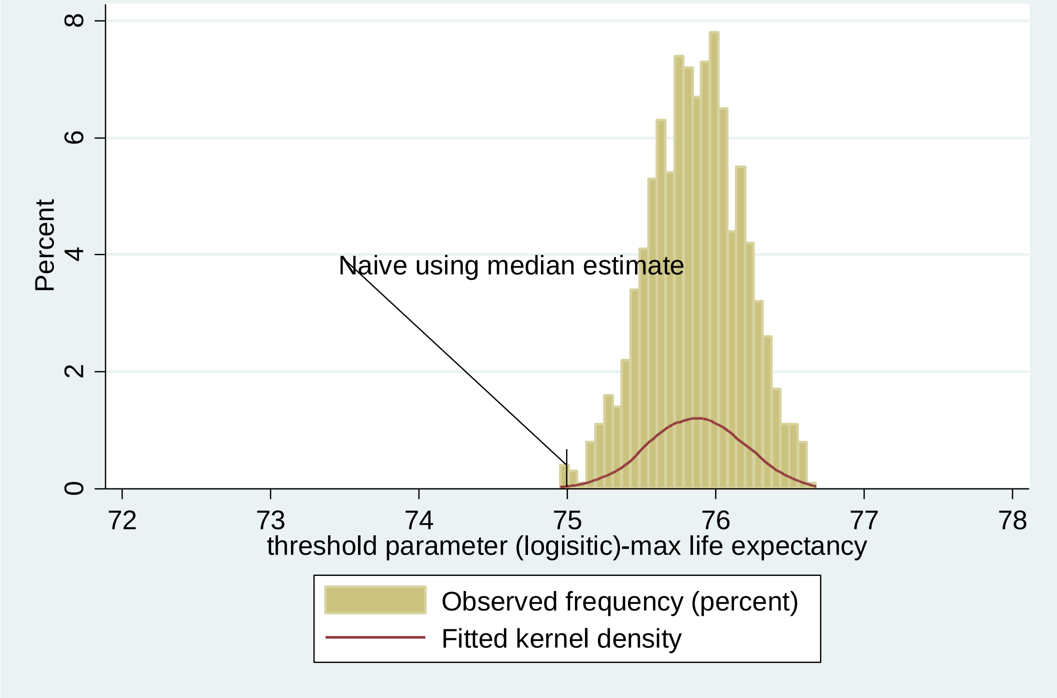


Figure 6b:Frequency distribution of logistic slope parameter

(Years after 1950)

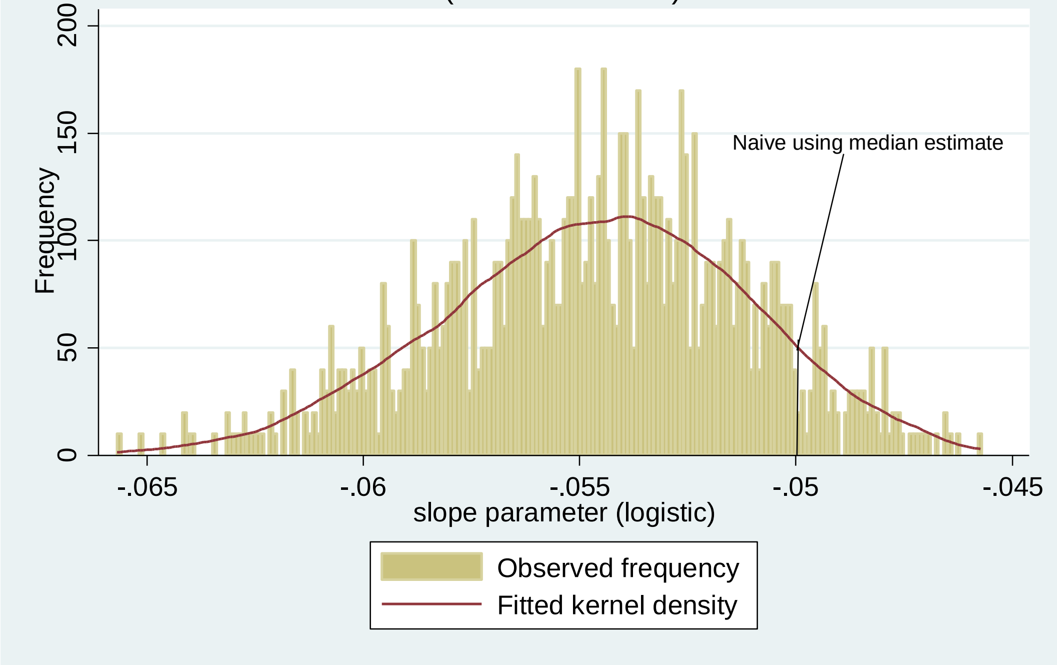


Figure 7: Distribution of elasticities from optimal models

(Years after 1950)

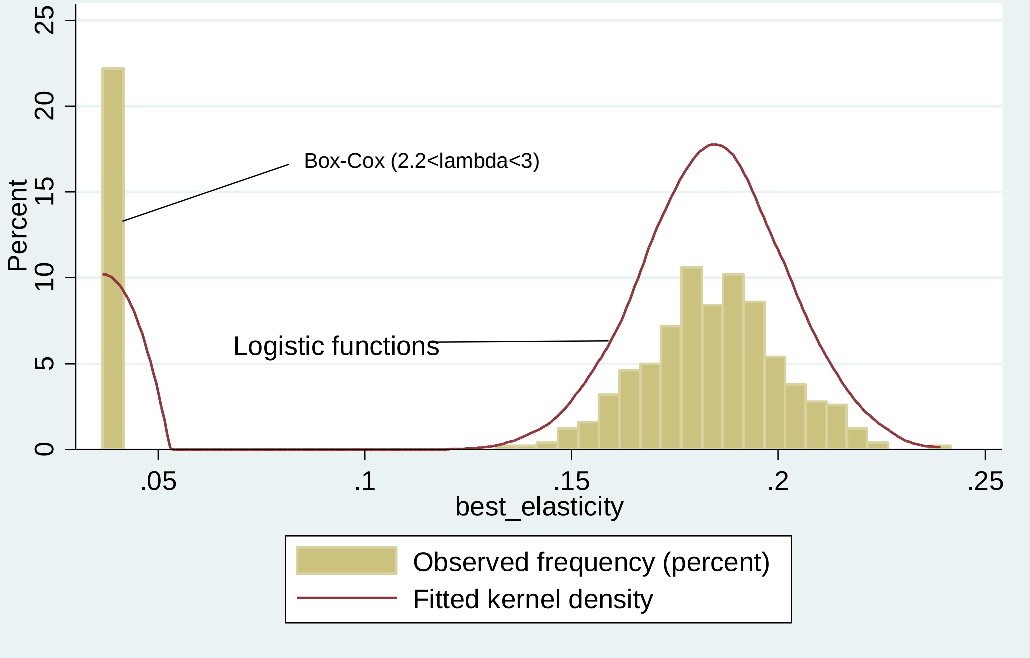


Figure 8a: Frequency distribution of slope parameter

(From Box Cox optimal model in each bootstrap)

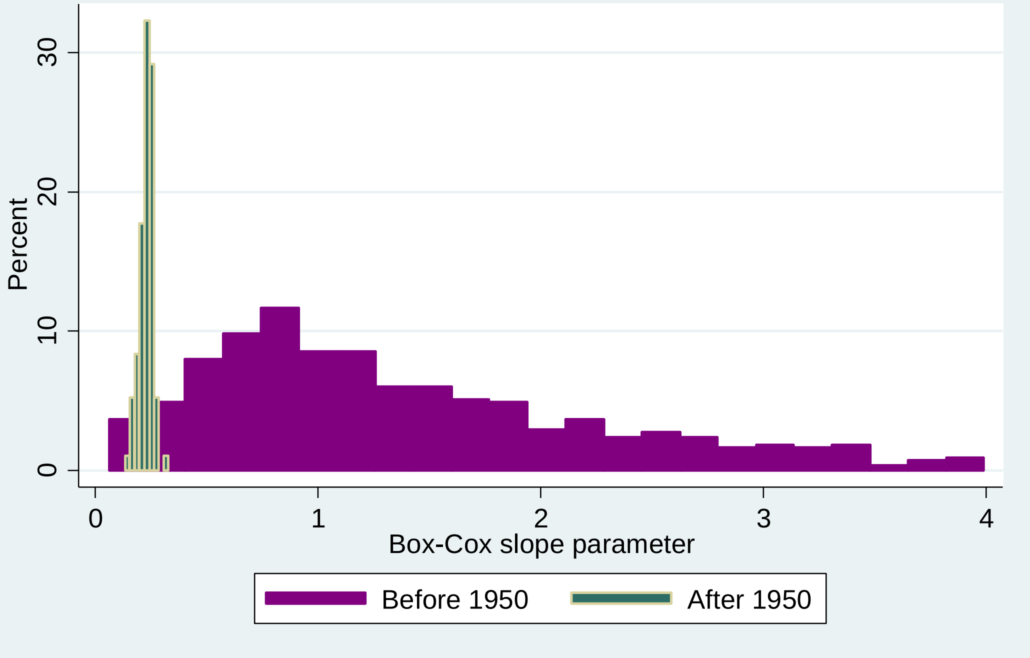


Figure 8b: Frequency distribution of Box-Cox shape parameter
(from replicas in which Box-Cox model is optimal; years before 1950)

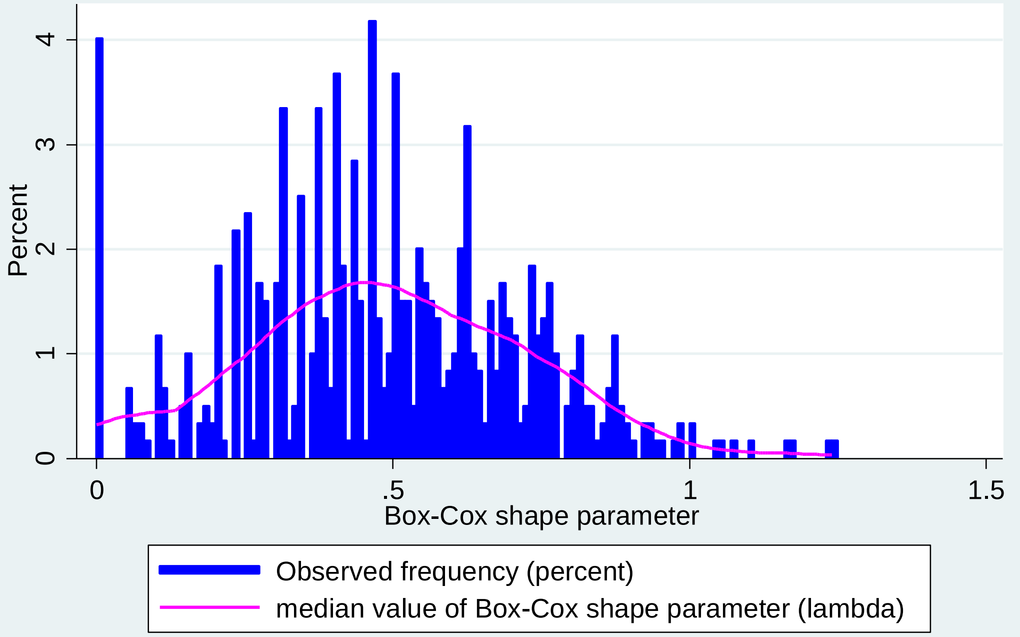


Figure 8c:Frequency distribution of estimated slopes

Logistic-new

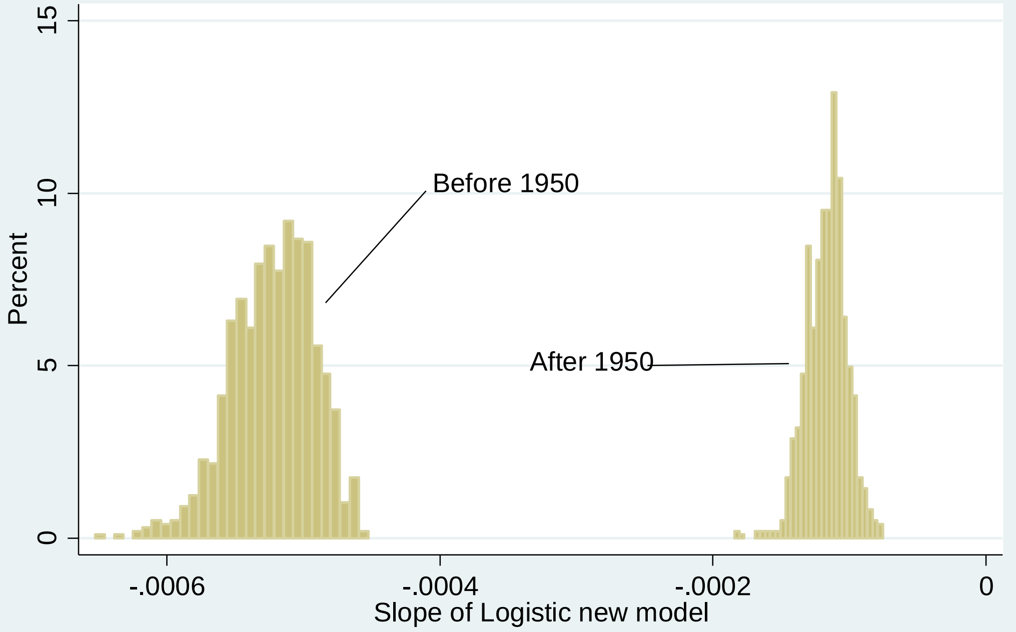


Figure 9a: Frequency distribution of estimated GDP elasticities

(From parameters of optimal model in each of 100 bootstraps; before 1950)

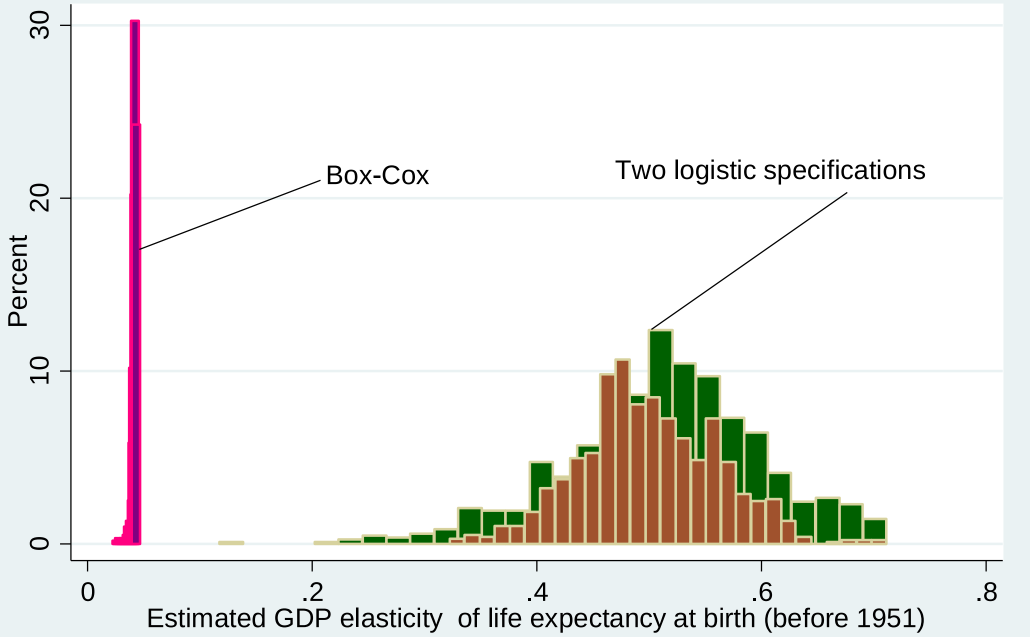


Figure 9b: Frequency distribution of estimated GDP elasticities

From parameters of optima model in each of 100 bootstraps; after 1950

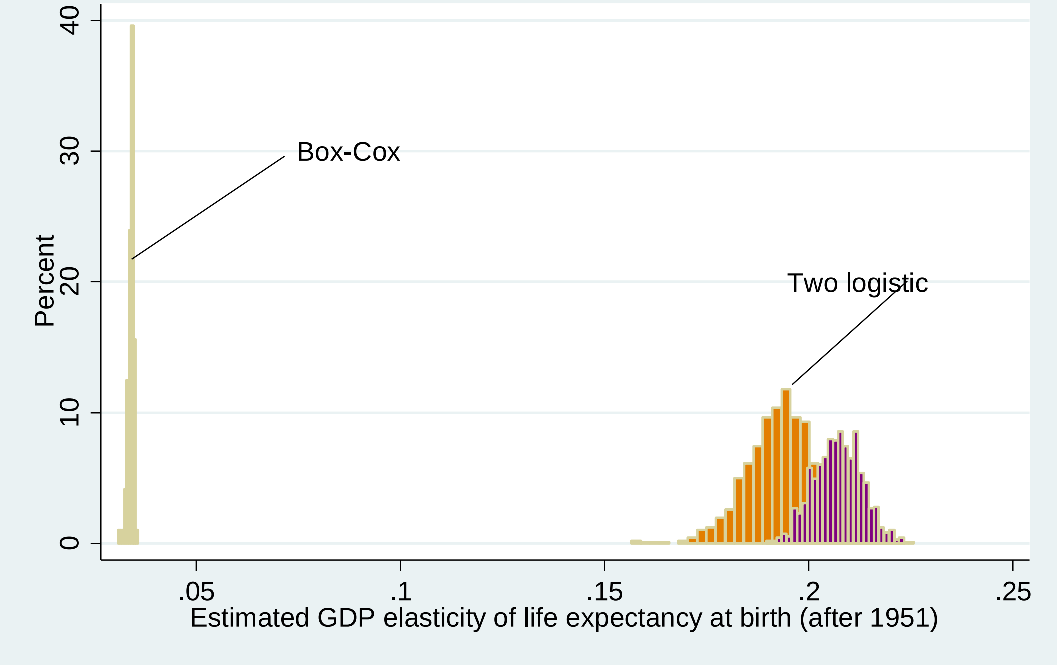


Figure 9c: Frequency distribution of estimated value of delta

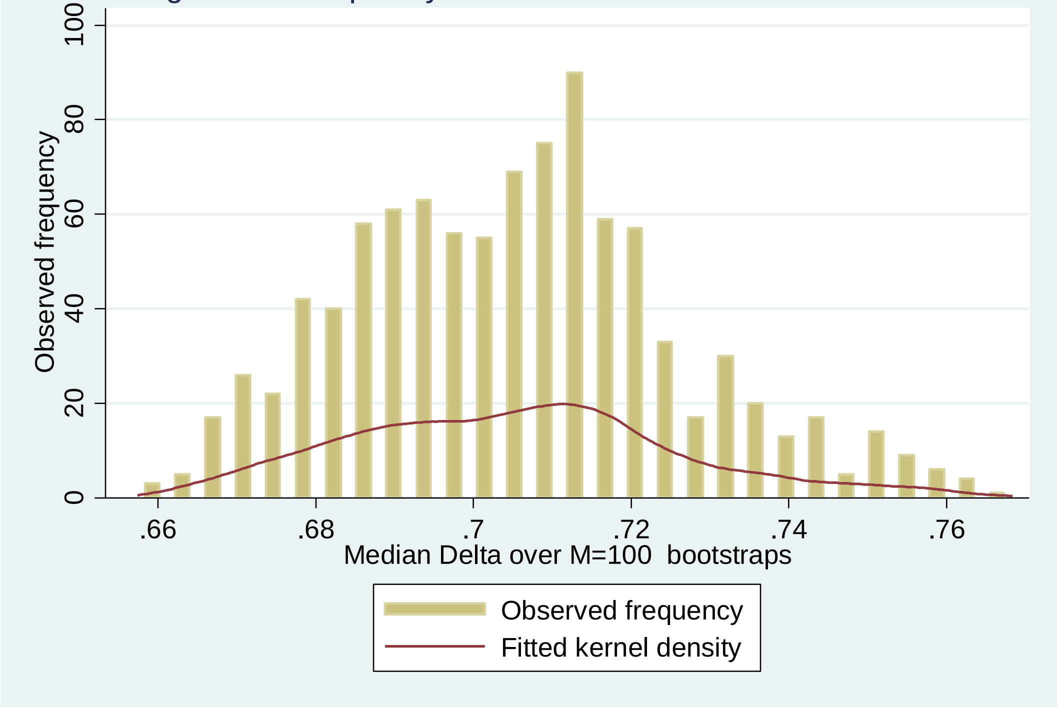


Table 1: Summary statistics of model estimation

Panel 1: Best model using median of estimates, years before 1950

Parameter	Box-Cox
Constant	1.91 (.25)
Slope	.23 (.033)
Lambda	0
MSE	56.88

Panel 2: Best model using median of estimates, years after 1950

Parameter	Logistic	
Threshold	76.95	(1.86)
Constant	-804	(502)
Slope	.000441	(.0001)

Panel 3: Summary statistics of N=1000 bootstraps for years before 1950 (Box-Cox)

Parameter	N	Mean	SE
Lambda	1,000	.216	.163
Slope	1,000	.531	.398
SE of slope	1,000	.117	.0866

Panel 4: Summary statistics of N=1000 bootstraps for years after 1950 (Logistic-new)

Parameter	N	Mean	SE
Threshold	1,000	75.87	.312
SE of threshold	1,000	1.639	.0837
Slope	1,000	-.00055	.000035
SE of slope	1,000	.00011	.00006