Joining Panel Data with Cross-Sections for Efficiency Gains: an Application to a Consumption Equation for Nicaragua*

Randolph Bruno† and Marco Stampini ‡

June 5, 2008

Abstract

This paper explores how cross-sectional data can be exploited jointly with longitudinal data, in order to increase estimation efficiency while properly tackling the potential bias due to unobserved individual characteristics. We propose an innovative procedure and we show its implementation by analysing the determinants of consumption in Nicaragua, based on data from three Living Standard Measurement Study surveys from 1993, 1998 and 2001. The last two rounds constitute an unbalanced longitudinal data set, while the first is a cross-section of different households. Under the assumption that the relationship between observed and unobserved characteristics is homogeneous across time, information from longitudinal data is used to clean the bias in the unpaired sample. In a second step, corrected unpaired observations are used jointly with panel data. This reduces the standard errors of the estimation coefficients and might increase their significance as well, otherwise compromised by the limited variation provided by the short longitudinal data.

*We would like to thank Stefano Della Vigna, Giovanni Dosi, Christopher Flinn, Roberto Golinelli, Chang-Tai Hsieh, Andrea Ichino, Monica Merito, Edward Miguel and the participants to the seminars at Sant'Anna School of Advanced Studies and UC Berkeley for fruitful comments. We would also like to thank the World Bank for making the LSMS data available. Eventually, a very special thank goes to Giorgio Calzolari, Enrico Moretti and Michael P. Murray for fundamental suggestions on the econometric approach. Shortcomings are our own.

†Corresponding Author: Randolph Bruno, DARRT, Dipartimento di Scienze Economiche, Università di Bologna, Strada Maggiore 45 - 40125 Bologna, Italy and IZA (Bonn), E-mail: randolph.bruno@unibo.it.

‡Marco Stampini, Development Research Department, African Development Bank, B.P. 323 - 1002 Tunis Belvedere, Tunisia, E-mail: m.stampini@afdb.org.

JEL Classification Numbers: C33, C42, I38.

Keywords: Panel Data, Estimation Efficiency, Pseudo-Panel, Consumption Model, Nicaragua.

1 Introduction

The availability of high quality longitudinal data sets is rare in developing countries. In most cases, researchers rely on short panel data or unpaired cross sections. Nonetheless, properly accounting for individual effects is fundamental in order to avoid bias due to the omission of unobservable characteristics.

Estimation efficiency increases with the size of the cross-sectional sample, and with the number of survey rounds across time. With short panels, especially if the time variation in the variables of interest is not high, estimation relies on a small number of observations and is affected by high standard errors. As a result, coefficients often look statistically not significant.

When proper panel data are not available, the literature has attempted various approaches, for example by matching locations across different cross-sections (Pitt, Rosenzweig and Gibbons (1993)), or by creating pseudo-panels of birth cohorts of individuals (Deaton (1985)).

In this paper, we develop a new approach, which can be applied when short panel data and other unpaired cross-sectional data are available. We show that non-longitudinal sources of the data can be exploited to increase panel models estimation efficiency. The underlying idea is that the proper panel model - based on the short longitudinal data - provides information that the researcher can use, in order to address the problem of bias from unobservable characteristics in the non-longitudinal data. In fact, cross-sectional data are likely to lead to the estimation of biased coefficients, and consequently to poor policy decisions. Our methodology suggests a fashion to exploit the panel nature of part of the data to correct the bias for all periods, i.e. also for those years in which only cross-section data are available.

The validity of the results depends on the key assumption of time-invariant relationship between observed and unobservable characteristics. We use pseudo-panels *à la* Deaton to validate this hypothesis, i.e. that the relationship between observable and unobservable characteristics is comparable in longitudinal and non-longitudinal data sets. We argue that the benefit of the correction of the bias is higher than the cost of the underlying assumption.

The idea is applied to the analysis of the determinants of consumption in Nicaragua, by using data from the Living Standard Measurement Study (LSMS) surveys of 1993, 1998 and 2001. The last two rounds constitute an unbalanced longitudinal data set, while the first is a cross-section of different households. We show that exploiting non-longitudinal data allows reducing the standard errors in the panel model, increasing efficiency while controlling for unobserved heterogeneity.

The paper is organized as follows. Section 2 describes the structure of the data

set, and introduces the consumption model on which the econometric exercise will be performed. Section 3 outlines the empirical strategy, by explaining how information from longitudinal data can be used to clean the unpaired observations from the bias due to the omission of household fixed effects. We define a new estimator and derive the expression for its variance-covariance matrix. Eventually, we explain how pseudo-panels can be used to test the underlying hypothesis that the relationship between observed and unobserved characteristics is homogenous across time. Section 4 applies the procedure to a consumption model for Nicaragua, and leads the reader along the different results provided by alternative techniques. Section 5 concludes.

2 Data and Consumption Model

We use data from three Living Standard Measurement Study (LSMS) surveys carried out in Nicaragua in 1993, 1998 and 2001 by the World Bank, which collected information on demographic characteristics, assets, economic activities, income and consumption. The LSMSs from 1998 and 2001 provide an unbalanced panel of more than 4000 households [1] per-period. Due to attrition of about 25%, only 3015 households are surveyed in both periods. However, previous work (Davis and Stampini (2002)) shows that attrition is quite random in nature and is not expected to produce a bias in the analysis of household consumption. After cleaning outliers and missing values, we are left with a balanced panel of two periods and 2791 households, a cross section of 1240 households in 1998 and 1399 households in 2001.

Data from 1993 are *not* part of the *longitudinal set*. In 1993, the LSMS surveyed 4454 households. After cleaning for outliers and missing values, we can exploit information on 4201 households. The structure of our data set is summarized in Table (1).

The main variable of interest in our application is per-capita household consumption. Values from the three years are normalized to 1995 prices[2]. We estimate a standard consumption equation:

$$C_{it} = \alpha + \beta X_{it} + [\phi_i + u_{it}] \tag{1}$$

where C is the logarithm of per capita household consumption, X is the vector of household characteristics, ϕ_i represents the household idiosyncratic effect, which could

[1] A household is defined as physical address where a family lives within a "segmento censal".

[2] According to the Consumer Price Index FP.CPI.TOTL of World Development Indicators for Nicaragua.

be either fixed or random, and u_{it} is a normally distributed error term. α and β contain the parameters to be estimated.

Consumption is a function of the following variables: demographic characteristics (household size and composition, age and gender of the head of the household); human capital (share of adult members over 15 years old with primary, secondary and higher education -excluded category illiterate); labour market participation (share of adult working members in non-agricultural self-employment, agricultural self-employment in large farms, and non-agricultural wage-employment -excluded category agricultural wage-employment and small farming); infrastructure and assets (availability of electricity and water in the house, quality of the dwelling as proxied by dirt floor, property of the house (registered or not), land size and number of heads of cattle).

Mean values and standard deviations of the above variables are reported in columns 1 and 2, Table (2), where 1993-1998-2001 sample and panel data statistics are compared. There is no systematic difference between the two samples.

3 The Empirical Strategy

If panel data were not available, equation (1) would be estimated omitting the term ϕ_i. This may create two kinds of problems, depending on the fact that the household idiosyncratic effect might be fixed or random in nature. In the former case, β may be affected by omitted variable bias, as observed household characteristics would pick up the effect of omitted unobserved and unobservable variables (upward bias). In the latter, the problem would be an improper treatment of heteroscedasticity: β would be consistent, but estimates of its standard errors would be biased.

When T cross sections are available -in our case three-, the estimation of equation (1) for each time period produces T estimates of β, potentially either biased, or with biased standard errors. Longitudinal data allow estimating unbiased coefficients and standard errors, although producing a single estimate for each element of β. The Hausman test can be used to determine if a random or fixed effect model is more appropriate. The former is preferable (being both consistent and efficient) if there is no correlation between household characteristics and idiosyncratic effect, in other words if this correlation does not alter the regression coefficients. Otherwise, fixed effects are preferable.

Unfortunately, the panel model exploits only a small part of available information (5,582 observations out of 12,422). Information provided by the 1993 survey, as well as the cross-sectional parts of 1998 and 2001, remains unexploited. We will show that this information has the potential to increase estimation efficiency. Depending on the

relationship between household observed and unobserved characteristics - i.e. the fact that the covariance $cov(X_{it}, \phi_i)$ is different from zero or not- we need to proceed in two different ways, examined in the next two sections.

3.1 Uncorrelated case

If the observed household characteristics X_i are uncorrelated with the unobserved household effects ϕ_i, both fixed and random effect techniques provide consistent estimates of the coefficients in equation (1). However, the random effect estimation is more efficient, i.e. the standard errors of the estimators are lower. In fact the within-fixed effects 'throw away' variability, either by differencing (in the case of two observations across time) or by computing the distance from the mean (in the case of more then two). The random effect estimation, on the other hand, simply accounts for the fact that the errors for paired observations in the balanced panel are correlated with one another. For this reason, proper weights are applied. When adding unpaired observations, we need to correct for the fact that (a) these are not correlated with paired observations and (b) are characterized by a different variance. Hence, unpaired observations need to be weighted differently, in order to account for heteroskedasticity[3].

3.2 Correlated Case

If household unobserved and observed characteristics *are* correlated, a fixed effect model is required. As we will show in section (4), the Hausman test reported in Table (5) rejects the null hypothesis of non-systematic difference in coefficients between RE and FE. It follows that omission of household fixed effects in equation (1) produces biased estimates of β. A bias-correction procedure (such as Least Squares Dummy Variables (LSDV) for longitudinal data) is required. This is the focus of the next section.

3.2.1 Correcting the Bias

Estimating model (1) including household dummy variables using only the 2,791 households of the balanced panel (LSDV) provides consistent estimates of the parameters. This is equivalent to estimating the model in differences with OLS. Only *time variant* explanatory variables can be included in the model, as every characteristic constant over time ends up in the household specific fixed effect.

We estimate the following model, equivalent to (1):

[3]Something similar happens when first order autoregressive disturbances are tackled through the Prais-Winsten procedure - the first observation and the partial differenced observations (for periods 2 through T) need being weighted differently.

$$\Delta_{01-98}C_i = \beta\Delta_{01-98}X_i + \Delta_{01-98}u_i \tag{2}$$

where $\Delta_{01-98}C = C_{01} - C_{98}$ and $\Delta_{01-98}X = X_{01} - X_{98}$, $\Delta_{01-98}u_i = u_{i01} - u_{i98}$ and recover unbiased estimates of β through the standard OLS formula:

$$\widehat{\beta}_{OLS\Delta_{01-98}} = (\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C \tag{3}$$

Applying OLS to pooled longitudinal data -using only the households of the balanced panel (9801BAL)- provides instead biased estimates of the coefficients. The expression for the estimators is as follows:

$$\widehat{\beta}_{OLS9801BAL} = (X'_{9801BAL}X_{9801BAL})^{-1}X'_{9801BAL}C_{9801BAL} \tag{4}$$

The difference between these two sets of estimates is a consistent estimate of the *OLS bias* for each coefficient, and defines the following vector:

$$\widehat{bias} = \widehat{\beta}_{OLS9801BAL} - \widehat{\beta}_{OLS\Delta_{01-98}} = (X'_{9801BAL}X_{9801BAL})^{-1}X'_{9801BAL}C_{9801BAL} - (\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C \tag{5}$$

Subtracting $X \times \widehat{bias}$ from consumption C for each household belonging to the balanced panel and applying expression (5) allows reproducing the unbiased coefficients $\widehat{\beta}_{OLS\Delta_{01-98}}$. In fact, this procedure purges the dependent variable from the correlation between explanatory and omitted variables. Under the assumption that the relationship between observed and unobserved characteristics is the same for panel and unpaired households (an assumption we will discuss in Section 3.2.2), the same procedure can be applied to all data from 1993, 1998 and 2001. The new dependent variable is defined as follows (where the superscript C stands for "cleaned"):

$$C^C_{93,98,01} = C_{93,98,01} - X_{93,98,01} \times \widehat{bias} \tag{6}$$

From this point onward, we will use the subscript Δ_{01-98} to indicate differences in the balanced panel, *98,01* for pooled observations of the balanced panel, and *93,98,01* for all pooled observations. For our procedure, the fact that some unpaired observations are recorded in 1993, and others in 1998 and 2001 (the unbalanced component of the panel) is not relevant. The reader may think of all unpaired observations as recorded in 1993. Applying the standard OLS formula to the new 'cleaned' dependent variable, we derive the new estimator as follows (again C means "cleaned"):

$$
\begin{aligned}
\widehat{\beta}^{C}_{OLS93,98,01} &= (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}C^{C}_{93,98,01} \\
&= (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}\{C_{93,98,01} - X_{93,98,01}[(X'_{98,01}X_{98,01})^{-1}X'_{98,01}C_{98,01} \\
&\quad -(\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C]\} = \\
&\quad (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}C_{93,98,01} - \\
&\quad (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}X_{93,98,01}[(X'_{98,01}X_{98,01})^{-1}X'_{98,01}C_{98,01} - \\
&\quad (\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C]\} = \\
&\quad (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}C_{93,98,01} - I_k\{(X'_{98,01}X_{98,01})^{-1}X'_{98,01}C_{98,01} - \\
&\quad (\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C]\} = \\
&\quad (X'_{93,98,01}X_{93,98,01})^{-1}X'_{93,98,01}C_{93,98,01} - (X'_{98,01}X_{98,01})^{-1}X'_{98,01}C_{98,01} + \\
&\quad (\Delta_{01-98}X'\Delta_{01-98}X)^{-1}\Delta_{01-98}X'\Delta_{01-98}C \Rightarrow \\
\widehat{\beta}^{C}_{OLS93,98,01} &= \widehat{\beta}_{OLS93,98,01} - \widehat{\beta}_{OLS98,01} + \widehat{\beta}_{OLS\Delta_{01-98}} = \\
&\quad \widehat{\beta}_{OLS93,98,01} - \widehat{\beta}_{OLS98,01} + \widehat{\beta}_{FE98,01} = \\
&\quad \widehat{\beta}_{OLS93,98,01} - \widehat{bias} \qquad (7)
\end{aligned}
$$

The new estimator is a linear combination of three separate OLS estimates.

If $\widehat{\beta}_{OLS93,98,01} = \widehat{\beta}_{OLS98,01}$ - i.e. if the estimates from pooled 93-98-01 and pooled 98-01 are affected by the same bias - the cleaned 93-98-01 estimates will simply collapse to the fixed effect estimates for 1998-2001. Exploiting the greater number of observations will allow reducing the standard errors of the estimation. However, this is only a special case.

In general, OLS applied to the corrected data for 1993, 1998 and 2001 yield new consistent estimates of β - even if $\widehat{\beta}_{OLS93,98,01} \neq \widehat{\beta}_{OLS98,01}$. The new estimator is indeed unbiased:

$$
\begin{aligned}
E[\widehat{\beta}^{C}_{OLS93,98,01}] &= E[\widehat{\beta}_{OLS93,98,01}] - E[\widehat{\beta}_{OLS98,01}] + E[\widehat{\beta}_{FE98,01}] \\
&= \beta_{OLS93,98,01} - \beta_{OLS98,01} + \beta_{FE98,01} \\
&= \beta_{OLS93,98,01} - bias = \beta^{C}_{OLS93,98,01} \qquad (8)
\end{aligned}
$$

We define

$$
W = X_{93,98,01}; \quad Z = X_{98,01}; \quad \Delta = \Delta_{01-98}X;
$$

$$
\varepsilon_{1(3n\times 1)} = \varepsilon_{93,98,01}; \quad \varepsilon_{2(2n\times 1)} = \varepsilon_{98,01}; \quad \varepsilon_{3(n\times 1)} = \Delta_{01-98}\varepsilon
$$

$$I_{3n}\sigma_1^2 = V(\varepsilon_1) = E[\varepsilon_1\varepsilon_1']; \quad I_{2n}\sigma_2^2 = V(\varepsilon_2) = E[\varepsilon_2\varepsilon_2']; \quad I_n\sigma_3^2 = V(\varepsilon_3) = E[\varepsilon_3\varepsilon_3']$$

and we know that the variance-covariance matrix of the cleaned estimator -under the assumption of homoskedasticity and no-autocorrelation- is given by the following expression:

$$VAR(\widehat{\beta}_{OLS}^{C}) = (W'W)^{-1}[\sigma_1^2 - 2\sigma_2^2 + 2\sigma_3^2] + (Z'Z)^{-1}[\sigma_2^2 - 2\sigma_3^2] + (\Delta'\Delta)^{-1}\sigma_3^2 \quad (9)$$

The formal derivation of expression (9), as well as the formula for the most general case characterized by heteroskedasticity and autocorrelation, is presented in Appendix.

3.2.2 Relationship between observable and unobservable characteristics

Our bias-correction procedure relies on the fundamental assumption that the relationship between observable and unobservable characteristics is homogenous in the sub-sample of unpaired observations and in the balanced panel.

In order to attribute to unpaired observations (referred to as 1993) the same bias computed on the balanced panel 1998-2001, we need to assume that the covariance between the observable characteristics (X_{it}) and the omitted unobservable variables (ϕ_i) is the same in the two samples. This hypothesis can be formalised as follows:

$$\begin{gathered} COV(X_{93}, \phi_{93}) = COV(X_{98,01}, \phi_{98,01}) \Rightarrow \\ E[X_{93}\phi_{93}] - E[X_{93}]E[\phi_{93}] = E[X_{98,01}\phi_{98\&01}] - E[X_{98,01}]E[\phi_{98,01}] \end{gathered} \quad (10)$$

Unfortunately, there is no way to test this hypothesis directly on household data, because ϕ_{93} cannot be estimated.

To validate our proposal, we replicate the analysis and test the key hypothesis within a pseudo-panel *à la* Deaton (1985). Pseudo-panels are cohort panels, in which the structure of the fixed effect is homogenous across time by definition.

We create a cohort database (pseudo panel or synthetic panel), subdividing the sample in groups defined by the same value of a few key observed time-invariant characteristics. The cohort panel is a real panel (with unobservable characteristics stable across time), in which $\phi_{93} = \phi_{98,01}$. For the cohort panel we can write:

$$H_0 : COV(X_{93}, \phi) = COV(X_{98,01}, \phi) \quad (11)$$

Being the fixed effect in 1998 and 2001 the same by construction [4], we can split the test in two parts:

$$
\begin{aligned}
H_0 : COV(X_{93}, \phi) &= COV(X_{98}, \phi) \\
E[X_{93}\phi] - E[X_{93}]E[\phi] &= E[X_{98}\phi] - E[X_{98}]E[\phi] \\
E[X_{93}\phi - X_{98}\phi] - \{E[X_{93}] - E[X_{98}]\}E[\phi] &= 0 \\
E[X_{93}\phi - X_{98}\phi] - E[X_{93} - X_{98}]E[\phi] &= 0 \\
E[(X_{93} - X_{98})\phi] - E[X_{93} - X_{98}]E[\phi] &= 0 \\
H_0 : COV[(X_{93} - X_{98}), \phi] &= 0
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
H_0 : COV(X_{93}, \phi) &= COV(X_{01}, \phi) \\
E[X_{93}\phi] - E[X_{93}]E[\phi] &= E[X_{01}\phi] - E[X_{01}]E[\phi] \\
E[X_{93}\phi - X_{01}\phi] - \{E[X_{93}] - E[X_{01}]\}E[\phi] &= 0 \\
E[X_{93}\phi - X_{01}\phi] - E[X_{93} - X_{01}]E[\phi] &= 0 \\
E[(X_{93} - X_{01})\phi] - E[X_{93} - X_{01}]E[\phi] &= 0 \\
H_0 : COV[(X_{93} - X_{01}), \phi] &= 0
\end{aligned} \tag{13}
$$

We test the double hypotheses through Hausman tests, comparing fixed versus random-effect estimates separately for the balanced pseudo-panel data for 1993-1998 and 1993-2001. This double test verifies a sufficient -stronger than necessary- condition for (11) to hold.

If we fail to reject the null hypothesis of homogenous relationship between observable and unobservable characteristics in the pseudo-panel, we may feel comfortable assuming that the same holds for household data.

3.2.3 Pseudo Panel and the Homogeneity Hypothesis

Deaton (1985), Nijman and Verbeek (1990), Verbeek and Nijman (1992) have investigated the pros and cons of the creation of cohorts from repeated cross sections. Deaton (1985) defines a cohort as a "group with fixed membership" across time, and proposes the construction of an artificial panel of cohorts, called pseudo-panel. For example, all families with a 30-year old female head, living in the rural area of Rio San Juan, in 1993, are 'grouped' in a new unit -with each variable assuming the mean value within the cohort. The number of observations drops.

[4] We will formally test the hypothesis that household effects for 1998-2001 are fixed rather than random through a Hausman test.

When the database is 'collapsed' in cohorts, the model becomes:

$$C_{Ct} = \alpha + \beta X_{Ct} + [\phi_C + u_{Ct}] \tag{14}$$

where the subscript Ct stands for cohort at time t [5]. The specification is the same as in model (1), but now ϕ_C captures the cohort rather than household fixed effect.

As most literature on pseudo panels, we define cohorts on the basis of the year of birth of the household head. All households whose head was born in the same year form a cohort and, as such, can be tracked over time. However, not all cohorts exist at all times: for example, cohorts with very high initial values of age are likely to be lost across time. Such attrition makes the panel unbalanced.

Widening the window of the date of birth to intervals of more than one year reduces attrition. This happens because of an increase in the number of households per cohort, which also improves estimation's precision of cohort means. On the other hand, for a natural trade-off , this decreases the overall size of the balanced panel.

In addition to the date of birth of the head, for the definition of cohorts we exploit two time-invariant characteristics: the gender of the household head and the region of residence.

In theory, considering the observed values, 2,464 combinations of the three characteristics are possible (2 genders by 88 years of birth by 14 regions). In practice, only about 1,300 of these combinations are observed in each period. Of these, only 844 (about 63 percent) are observed in all three data sets. Our balanced pseudo-panel is hence made of three observations across time of 844 cohorts.

Two potential problems need being addressed: selection and time invariance.

Selection. Cohorts that do not have three time observations are not part of the balanced pseudo panel. Only if this selection is random, representativeness is preserved. Summary statistics for the pseudo panel sample are presented in Table (2), together with those of the full cohort sample: close correspondence between the mean value of basic characteristics suggests random attrition, so that selection does not seem to be an issue. The standard deviations are often smaller than for household data, because cohorts wash out part of the heterogeneity across household (the intra-cohort variability).

Time invariance. Two dynamic events deserve particular attention: change of the head of the family (e.g. marriage, death) and migration of households across regions. In our sample few household change the head of the family between 1998 and 2001 and

[5] Deaton (1985) points out that C_{Ct}, X_{Ct} are error-ridden measures of the true cohort means. Under the assumption of normally distributed and independent errors, he suggests a procedure of correction of the variance-covariance matrix.

according to Ambler (2005) migration across regions was very limited in those years, even after the Hurricane Mitch.

Overall, the criteria chosen for the definition of the cohorts seem to ensure good representativeness and size of the balanced panel.

We use the balanced pseudo panel to test the key hypothesis that the relationship between observed characteristics and fixed effects is homogenous across time (we call homogeneity hypothesis). We perform three tests. The Hausman test for 1998-2001 confirms that a fixed effect specification for the model is the most appropriate for the pseudo panel, consistently with our findings for the household panel. The Hausman test for 1993-1998 shows that the difference between the observable characteristics in the two years is uncorrelated with the fixed effect, which implies that the correlation between observable characteristics and fixed effect is the same in both years. The same holds for 1993-2001. The results reported in Table (3) support the null hypothesis that $COV(X_{93}, \phi) = COV(X_{98,01}, \phi)$. Tests performed on the unbalanced pseudo panel lead to the same conclusion. This is an indirect assessment of the validity of the key assumption for household data, on the basis of which we now proceed applying our bias-correction technique.

4 Results

The first way to estimate the consumption model described in equation (1) is to split the data by year and obtain three separate estimates of the coefficients β for 1993, 1998 and 2001. We focus our comments on the variables measuring education, both for practical reasons -which advise against discussing all coefficients-, and because of the relevance of the question asking wether a poor household can be lifted out of poverty by increasing its level of human capital [6].

Estimates of selected coefficients are presented in Table 4. In 1993, a household in which all adults hold a primary school degree consumes 26 percent more than if all adults had no degree (omitted category) [7]. Secondary education (always relative to

[6] Deaton (1985) warns that: "Expenditure differences between poor and rich consumers are not likely to be replicated by making a poor man rich unless the poor and the rich are otherwise identical". Berhman (1990) states: "In many cases schooling appears in substantial part to be a *proxy for other characteristics*, such as ability and motivation and family background, rather than representing purely the effects of schooling *per se*. Also the economic impact on human resource investments of the poor appears to be more through price effects and less through income effects than often is claimed". Nonetheless, Stampini and Davis (2006) show that education represents one of the fundamental assets that allow households exiting poverty in Nicaragua.

[7] The dependent variable is in log and we can interpret the coefficients as semi-elasticity.

no degree) is associated with an increase in consumption by 60 percent, and college and higher education with an increase by 116 percent -suggesting increasing returns to schooling. Evidence from 1998 and 2001 broadly confirms these results.

Cross-sectional estimates are likely to be biased because of the omission of time-invariant household unobserved and unobservable characteristics. Such bias is null only if unobserved characteristics are orthogonal to all regressors included in the model. Our data allow removing the bias by exploiting the balanced panel component of 1998 and 2001 data for the estimation of a panel fixed effect model.

A panel model will provide a single estimate of the β vector. It is worth noting that this may be problematic in case year-by-year estimation led to very different measures -for example because the structure of returns to education is changing. Panel model estimates provide a time-average of unbiased marginal effects.

Table (5) presents fixed-effect and random-effect estimates on the balanced panel sample. The latter do not differ substantially from those obtained with yearly OLS (Table 4). Fixed-effect estimates are radically different. They show that primary education has no significant impact on consumption. Secondary and higher education have positive significant effects, though of much smaller size than suggested by OLS estimates. A household in which all adults have a college degree consumes 15 percent more than if all adults had no school degree at all. Most of the effect estimated by OLS seem to be due to household time-invariant unobserved characteristics. The Hausman test confirms that fixed-effect and random-effect estimates are systematically different. The fixed-effect model is the most appropriate in our case.

Fixed-effect estimates are reported in Column 1 of Tables (6) & (7). Such estimates are unbiased. To measure the bias due to the omission of household time-invariant unobserved characteristics, we calculate OLS estimates on the pooled sample of the balanced panel. These estimates -reported in Column 2- resemble those of Table (4). The bias -given by the difference between Column 2 and Column 1- is reported in Column 3. Column 4 shows the results of the application of the bias-correction procedure to the households of the balanced panel. It perfectly replicates Column 1[8].

As the tests outlined in Section 3.2.3 with reference to pseudo-panel data suggest that the relationship between observed characteristics and fixed effects is homogenous across time, we proceed by applying the same bias-correction technique to the pooled sample of all observations from the three years.

Column 5 in Tables (6) & (7) reports biased estimates from OLS on the pooled sample of all households (1993, 1998 and 2001). These are a weighted average of the

[8]If standard errors were not corrected according to formula 9, they would on the contrary reproduce those from OLS estimation reported in Column 2.

three columns in Table (4). Eventually, Column 6 presents the results of the application of our bias-correction procedure to the same pooled sample. Corrected standard errors are calculated according to formula (9).

As expected, the new coefficients are not identical to those of the fixed-effect model. However, they always have the same sign and confirm all significant effects. Remarkably, they are estimated more efficiently than fixed-effect coefficients on the balanced panel: all standard errors are lower than in Column 1. In some cases, the combination of a slightly larger estimate of the coefficient and a smaller standard error make a previously non-significant variable significant. For example, access to running water in the house is associated with a significant increase in consumption by 4.5 percent -while the coefficient from the fixed-effect estimation on the balanced panel was not statistically different from zero. Living in a house with unpaved floor is associated with a significantly lower (by 7.5 percent) level of consumption -while the effect estimated with the fixed-effect model was not significant. In addition, the significance level of some already significant coefficients grows. For example, the effect of higher than secondary education is now significant at the 1 percent level of confidence -while the significance level for the fixed effect estimate was 5 percent.

Although the differences are not large, the above results confirm our intuition that the combination of unpaired observations and longitudinal data can increase estimation efficiency, relative to traditional panel data models.

5 Conclusions

We proposed and explored an innovative econometric procedure, which exploits non-longitudinal data to increase efficiency in the estimation of panel data models.

We focused mainly on fixed effect models. We are aware that random effect models are a valuable alternative, especially in contexts where the structure of the family is subject to high time variation. For this reason, we briefly discussed the case of random (time variant) unobservable characteristics as a theoretical alternative in section (3.1). The reason of our focus on fixed effect models is twofold. First, they were the most appropriate in our case, according to the tests we performed. Second, they represent the most interesting case from the theoretical point of view, as in the case of time variant household effects the procedure of integration of unpaired data and balanced panels boils down to a properly designed Feasible Generalized Least Square estimation.

We showed that, under the key assumption that the covariance between observed characteristics and fixed effects in the unpaired observations is the same as in the

balanced panel (homogeneity hypothesis), information from the latter can be used to clean the bias due to the omission of time invariant unobservable characteristics. In a second step, corrected unpaired data can be exploited to improve the precision of panel model estimators. We are aware that the underlying assumption is not testable in our household data. Nonetheless, we propose the use of a pseudo panel to test the hypothesis for cohorts from the same data.

We applied the new methodology to the estimation of a consumption model using data from three Living Standard Measurement Study surveys carried out in Nicaragua in 1993, 1998 and 2001. The last two rounds constitute an unbalanced longitudinal data set, while the first is a cross-section of different households. We showed that the application of the 'fixed-effect correction' to unpaired data and the integration with the balanced panel lead to efficiency gains, relative to the case in which the estimation relies only on the balanced panel.

The magnitude of the efficiency gain is likely to be a function of the size and variability of non-longitudinal data relative to the balanced panel, although additional research is needed to better understand the determinants of the result.

References

Ambler, C. (2005). "The Distribution of Emercency Relief in Post Hurricane Mitch Nicaragua". Thesis, Williams College, Williamstown, Massachusetts.

Berhman, J. R. (1990). "The Action of Human Resources and Poverty on one an Other: What We have yet to Learn" Living Standard Meausurement Study n. 74.

Davis, B. and Marco Stampini, (2002). "Pathways towards prosperity in rural Nicaragua; or why households drop in and out of poverty, and some policy suggestions on how to keep them out", ESA-FAO Working Paper Series, #12, http://www.fao.org/es/ESA/pdf/wp/ESAWP02_12.pdf.

Deaton, Angus (1985). "Panel Data from Time Series of Cross-Sections", *Journal of Econometrics*, 30 (1985), 109-126.

Nijman, Theo and Marno Verbeek (1990). "Estimation of Time-Dependent Parameters in Linear Models Using Cross-Sections, Panel, or Both", *Journal of Econometrics*, 46 (1990), 333-346.

Pitt, Mark M., Mark R. Rosenzweig and Donna M. Gibbons (1993), "The Determinants and consequences of the Placement of Goverment Programs in Indonesia" *The World Bank Economic Review*, Volume 7, pp. 319-348.

Stampini, M. and B. Davis (2006). "Discerning Transient From Chronic Poverty In Nicaragua: Measurement With A Two Period Panel Data Set". *European Journal of Development Research*, 18(1): 105-130.

Verbeek, Marco and Theo Nijman (1992). "Can Cohort Data be Treated as Genuine Panel Data", *Empirical Economics*, 17 , 9-23.

Appendix: The variance of the OLS estimator Corrected for unobserved heterogeneity

Defining

$$W = X_{93,98,01}; \quad Z = X_{98,01}; \quad \Delta = \Delta_{01-98}X;$$

$$\varepsilon_{1(3n\times 1)} = \varepsilon_{93,98,01}; \quad \varepsilon_{2(2n\times 1)} = \varepsilon_{98,01}; \quad \varepsilon_{3(n\times 1)} = \Delta_{01-98}\varepsilon$$

we can write

$$\begin{aligned}
VAR(\widehat{\beta}^{C}_{OLS}) &= E\{[(W'W)^{-1}W'\varepsilon_1 - (Z'Z)^{-1}Z'\varepsilon_2 + \\
&\quad (\Delta'\Delta)^{-1}\Delta'\varepsilon_3][\varepsilon_1'W(W'W)^{-1} - \varepsilon_2'Z(Z'Z)^{-1} + \varepsilon_3\Delta(\Delta'\Delta)^{-1}]'\} \\
&= E\{(W'W)^{-1}W'\varepsilon_1\varepsilon_1'W(W'W)^{-1} - (W'W)^{-1}W'\varepsilon_1\varepsilon_2'Z(Z'Z)^{-1} \\
&\quad +(W'W)^{-1}W'\varepsilon_1\varepsilon_3'\Delta(\Delta'\Delta)^{-1} - (Z'Z)^{-1}Z'\varepsilon_2\varepsilon_1'W(W'W)^{-1} \\
&\quad +(Z'Z)^{-1}Z'\varepsilon_2\varepsilon_2'Z(Z'Z)^{-1} - (Z'Z)^{-1}Z'\varepsilon_2\varepsilon_3'\Delta(\Delta'\Delta)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\varepsilon_3\varepsilon_1'W(W'W)^{-1} - (\Delta'\Delta)^{-1}\Delta'\varepsilon_3\varepsilon_2'Z(Z'Z)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\varepsilon_3\varepsilon_3'\Delta(\Delta'\Delta)^{-1}\} \\
&= (W'W)^{-1}W'E[\varepsilon_1\varepsilon_1']W(W'W)^{-1} - (W'W)^{-1}W'E[\varepsilon_1\varepsilon_2']Z(Z'Z)^{-1} \\
&\quad +(W'W)^{-1}W'E[\varepsilon_1\varepsilon_3']\Delta(\Delta'\Delta)^{-1} - (Z'Z)^{-1}Z'E[\varepsilon_2\varepsilon_1']W(W'W)^{-1} \\
&\quad +(Z'Z)^{-1}Z'E[\varepsilon_2\varepsilon_2']Z(Z'Z)^{-1} - (Z'Z)^{-1}Z'E[\varepsilon_2\varepsilon_3']\Delta(\Delta'\Delta)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'E[\varepsilon_3\varepsilon_1']W(W'W)^{-1} - (\Delta'\Delta)^{-1}\Delta'E[\varepsilon_3\varepsilon_2']Z(Z'Z)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'E[\varepsilon_3\varepsilon_3']\Delta(\Delta'\Delta)^{-1}
\end{aligned} \tag{15}$$

Formula (15) states that the variance of $\widehat{\beta}^{C}_{OLS}$ is lower than the variance of the FE model on the 98-01 balanced panel if the sum and subtraction of the first eight addenda is negative. As this general case is not easily treatable, we assume homoskedasticity and no autocorrelation.

$$I_{3n}\sigma_1^2 = V(\varepsilon_1) = E[\varepsilon_1\varepsilon_1']; \quad I_{2n}\sigma_2^2 = V(\varepsilon_2) = E[\varepsilon_2\varepsilon_2']; \quad I_n\sigma_3^2 = V(\varepsilon_3) = E[\varepsilon_3\varepsilon_3']$$

In this case, we can simplify as follows:

$$
\begin{aligned}
VAR(\widehat{\beta}^{C}_{OLS}) &= (W'W)^{-1}W'[I_{3n}\sigma_1^2]W(W'W)^{-1} - (W'W)^{-1}W'\left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & 0 \\ \hline I_n\sigma_2^2 & 0 \\ \hline 0 & I_n\sigma_2^2 \\ \hline 0 & 0 \end{array}\right] Z(Z'Z)^{-1} \\
&\quad +(W'W)^{-1}W'\left[\begin{array}{c} 0 \\ \hline 0 \\ \hline -I_n\sigma_3^2 \\ \hline I_n\sigma_3^2 \\ \hline 0 \end{array}\right] \Delta(\Delta'\Delta)^{-1} \\
&\quad -(Z'Z)^{-1}Z'\left[\begin{array}{c|c|c|c|c} 0 & 0 & I_n\sigma_2^2 & 0 & 0 \\ \hline 0 & 0 & 0 & I_n\sigma_2^2 & 0 \end{array}\right] W(W'W)^{-1} \\
&\quad +(Z'Z)^{-1}Z'\left[I_{2n}\sigma_2^2\right] Z(Z'Z)^{-1} - (Z'Z)^{-1}Z'\left[\begin{array}{c} -I_n\sigma_3^2 \\ \hline I_n\sigma_3^2 \end{array}\right] \Delta(\Delta'\Delta)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\left[\begin{array}{c|c|c|c|c} 0 & 0 & -I_n\sigma_3^2 & I_n\sigma_3^2 & 0 \end{array}\right] W(W'W)^{-1} \\
&\quad -(\Delta'\Delta)^{-1}\Delta'\left[\begin{array}{c|c} -I_n\sigma_3^2 & I_n\sigma_3^2 \end{array}\right] Z(Z'Z)^{-1} \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\left[I_n\sigma_3^2\right] \Delta(\Delta'\Delta)^{-1}
\end{aligned}
$$

With further simplification, we obtain formula (9):

$$
\begin{aligned}
VAR(\widehat{\beta}^{C}_{OLS}) &= (W'W)^{-1}W'W(W'W)^{-1}\sigma_1^2 - (W'W)^{-1}Z'Z(Z'Z)^{-1}\sigma_2^2 \\
&\quad +(W'W)^{-1}\Delta'\Delta(\Delta'\Delta)^{-1}\sigma_3^2 - (Z'Z)^{-1}Z'Z(W'W)^{-1}\sigma_2^2 \\
&\quad +(Z'Z)^{-1}Z'Z(Z'Z)^{-1}\sigma_2^2 - (Z'Z)^{-1}\Delta'\Delta(\Delta'\Delta)^{-1}\sigma_3^2 \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\Delta(W'W)^{-1}\sigma_3^2 - (\Delta'\Delta)^{-1}\Delta'\Delta(Z'Z)^{-1}\sigma_3^2 \\
&\quad +(\Delta'\Delta)^{-1}\Delta'\Delta(\Delta'\Delta)^{-1}\sigma_3^2 \\
&= (W'W)^{-1}\sigma_1^2 - (W'W)^{-1}\sigma_2^2 \\
&\quad +(W'W)^{-1}\sigma_3^2 - (W'W)^{-1}\sigma_2^2 \\
&\quad +(Z'Z)^{-1}\sigma_2^2 - (Z'Z)^{-1}\sigma_3^2 \\
&\quad +(W'W)^{-1}\sigma_3^2 - (Z'Z)^{-1}\sigma_3^2 \\
&\quad +(\Delta'\Delta)^{-1}\sigma_3^2 \\
&= (W'W)^{-1}[\sigma_1^2 - 2\sigma_2^2 + 2\sigma_3^2] + (Z'Z)^{-1}[\sigma_2^2 - 2\sigma_3^2] + (\Delta'\Delta)^{-1}\sigma_3^2
\end{aligned}
$$

Number of households, by year	1993	1998	2001
Unpaired	4201	-	-
Unpaired	-	1240	-
Balanced Panel	-	2791	2791
Unpaired	-	-	1399
Total	4201	4031	4190

Table 1: Structure of the database

	Households				Cohorts			
	All 93-98-01		Panel 98-01		All 93-98-01		Panel 93-98-01	
Variable	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Real per-capita Consumption*	4563	6491	4409	4930	5020	5480	5140	5182
D. Female Head	0.27	0.44	0.27	0.45	0.41	0.49	0.32	0.47
Age Head	46.07	14.42	47.48	14.33	50.31	18.58	47.53	15.43
Family Size	6.81	3.06	6.86	3.08	5.40	2.31	5.60	2.06
D. Urban	0.57	0.50	0.56	0.50	0.58	0.49	0.59	0.49
D. Water	0.59	0.49	0.60	0.49	0.53	0.44	0.55	0.42
D. Electricity	0.68	0.46	0.70	0.46	0.64	0.41	0.65	0.39
D. Dirt Floor	0.46	0.50	0.46	0.50	0.49	0.40	0.47	0.37
D. Property Reg.	0.51	0.50	0.51	0.50	0.54	0.39	0.53	0.35
D. Property Not-Reg.	0.31	0.46	0.35	0.48	0.27	0.33	0.28	0.30
D. Not-Property	0.18	0.39	0.14	0.35	0.19	0.29	0.19	0.25
% Kids aged 0-4	0.13	0.14	0.12	0.13	0.13	0.12	0.13	0.11
% Kids 5-10	0.17	0.15	0.17	0.15	0.15	0.13	0.16	0.12
% Males 11-14	0.06	0.09	0.06	0.09	0.05	0.07	0.05	0.06
% Female 11-14	0.05	0.09	0.05	0.09	0.05	0.07	0.05	0.06
% Males 15-19	0.06	0.10	0.07	0.10	0.06	0.09	0.06	0.08
% Female 15-19	0.06	0.10	0.06	0.10	0.06	0.09	0.06	0.08
% Males 20-34	0.11	0.13	0.10	0.12	0.10	0.12	0.10	0.11
% Female 20-34	0.12	0.12	0.11	0.12	0.12	0.11	0.12	0.10
% Males 35-59	0.09	0.11	0.09	0.11	0.07	0.10	0.08	0.10
% Female 35-59	0.09	0.10	0.10	0.11	0.10	0.11	0.10	0.10
% Males > 60	0.03	0.08	0.03	0.08	0.05	0.13	0.04	0.11
% Female > 60	0.03	0.09	0.03	0.09	0.07	0.16	0.05	0.12
% Adults Prim. Ed.	0.23	0.28	0.23	0.27	0.21	0.22	0.22	0.19
% Adults Sec. Ed.	0.15	0.25	0.16	0.24	0.13	0.19	0.14	0.18
% Adults Higher Ed.	0.06	0.18	0.07	0.19	0.06	0.13	0.07	0.13
% Adults Self. Emp.	0.37	0.39	0.35	0.36	0.37	0.32	0.39	0.29
% Adults Big Farmers	0.04	0.16	0.05	0.18	0.04	0.12	0.04	0.11
% Adults Non-Agr. Wage Emp.	0.42	0.41	0.42	0.39	0.39	0.34	0.41	0.32
Land Use	5.82	35.81	6.85	41.72	6.32	32.84	6.17	21.62
# Cattle	1.64	10.78	1.57	9.66	2.09	10.64	1.89	6.59
# Observations	12422		5582 (2791x2)		4010		2532 (844x3)	

Notes: Time invariant variables for cohort construction: Female Headed household, Year of birth of the head, Region, Urban Rural. D. stands for Dummy. * Cordobas at 1995 prices.
Source: Our elaboration on LSMS data.

Table 2: Summary Statistics: Samples Comparison

	$chi^2(.) = (b-B)'[(V_b - V_B)^{(-1)}](b-B)$	$Prob > chi^2$
1998-2001	125.56(26)	0.0000
1993-2001	70.50(26)	0.0000
1993-1998	62.57(26)	0.0001
1993-1998-2001	78.14(26)	0.0000

Notes: Number of Coefficients in parenthesis.

Table 3: Hausman Test of the relationship between observable and unobservable characteristics: Balanced Pseudo Panel

	OLS, all households		
Variable	1993	1998	2001
Share with primary education	0.265***	0.292***	0.294***
	(0.037)	(0.032)	(0.030)
Share with secondary education	0.597***	0.566***	0.539***
	(0.045)	(0.037)	(0.034)
Share with higher education	1.159***	1.104***	0.969***
	(0.062)	(0.05)	(0.039)
Share in non-ag. self-employment	0.133***	0.120***	0.082***
	(0.031)	(0.028)	(0.027)
Share in large farming	0.499***	0.438***	0.307***
	(0.127)	(0.051)	(0.040)
Share in non-ag. wage employment	0.064*	0.052*	0.067**
	(0.033)	(0.029)	(0.028)
running water in or outside house	0.191***	0.140***	0.123***
	(0.026)	(0.021)	(0.019)
has electricity	0.340***	0.246***	0.239***
	(0.029)	(0.023)	(0.021)
dirtfloor	-0.298***	-0.234***	-0.228***
	(0.024)	(0.019)	(0.017)
yes/no own house, with escritura	-0.007	0.052**	0.046**
	(0.028)	(0.022)	(0.020)
yes/no own house, without escritura	-0.044	-0.02	-0.041*
	(0.031)	(0.023)	(0.021)
# manzanas land use	0.001	0.001**	0.001**
	(0.001)	(0.000)	(0.000)
# cattle vacuno	0.005***	0.003***	0.005***
	(0.001)	(0.001)	(0.001)
Constant	8.954***	8.731***	8.689***
	(0.114)	(0.097)	(0.086)
Observations	4201	4031	4190
R-squared	0.53	0.58	0.61

Notes: Standard errors in parenthesis.

Table 4: OLS year by year

	Balanced Panel Households		
Variable	FE	RE	Difference
Share with primary education	0.032	0.278***	-0.246***
	(0.043)	(0.028)	(0.033)
Share with secondary education	0.118**	0.583***	-0.465***
	(0.058)	(0.032)	(0.048)
Share with higher education	0.150**	0.888***	-0.738***
	(0.073)	(0.040)	(0.061)
Share in non-ag. self-employment	0.029	0.069***	-0.040*
	(0.030)	(0.023)	(0.019)
Share in large farming	0.03	0.197***	-0.167***
	(0.046)	(0.036)	(0.028)
Share in non-ag. wage employment	-0.02	0.039	-0.059**
	(0.032)	(0.024)	(0.021)
running water in or outside house	0.029	0.137***	-0.108***
	(0.029)	(0.018)	(0.023)
has electricity	0.080**	0.237***	-0.157***
	(0.035)	(0.020)	(0.029)
dirtfloor	-0.032	-0.206***	0.174***
	(0.029)	(0.017)	(0.024)
yes/no own house, with escritura	-0.001	0.049**	-0.050**
	(0.027)	(0.020)	(0.019)
yes/no own house, without escritura	0.014	0.005	0.009
	(0.027)	(0.020)	(0.018)
# manzanas land use	0.000	0.000	0.000**
	(0.000)	(0.000)	(0.000)
# cattle vacuno	0.003***	0.006***	-0.002**
	(0.001)	(0.001)	(0.001)
Constant	-	8.734***	-
		(0.083)	
Observations	5582	5582	
R-squared	0.152	0.44	
Hausman Test			410.29
$Prob > chi^2$			0.0000

Notes: Standard errors in parenthesis. Hausman Test: chi2(27) = $(b-B)'[(V\_b - V\_B)^{(-1)}](b-B)$

Table 5: FE, RE and Hausman test

	Balanced Panel Households				All households	
Variable	FE	Pooled	Bias	Bias-C.	Pooled	Bias-C.
	98,01	98,01	Col. (2)-(1)	98,01	93,98,01	93,98,01
	Δl(pc Cons.)	l(pc Cons.)		l(pc Cons.)	l(pc Cons.)	l(pc Cons.)
D. Female Head	-0.047	-0.025	0.022	-0.047	0.011	-0.011
	(0.038)	(0.017)		(0.038)	(0.013)	(0.036)
Age Head	-0.004***	-0.004***	0.000	-0.004***	-0.005***	-0.005***
	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)
Family Size	-0.082***	-0.071***	0.011	-0.082***	-0.072***	-0.083***
	(0.005)	(0.003)		(0.005)	(0.002)	(0.005)
% Kids 0-4	-0.593***	-0.895***	-0.302	-0.593***	-1.041***	-0.739***
	(0.134)	(0.085)		(0.134)	(0.062)	(0.117)
% Kids 5-10	-0.525***	-0.865***	-0.34	-0.525***	-0.893***	-0.553***
	(0.128)	(0.077)		(0.128)	(0.058)	(0.115)
% Males 11-14	-0.506***	-0.819***	-0.313	-0.506***	-0.808***	-0.495***
	(0.146)	(0.096)		(0.146)	(0.072)	(0.126)
% Female 11-14	-0.467***	-0.593***	-0.126	-0.467***	-0.726***	-0.601***
	(0.138)	(0.096)		(0.138)	(0.074)	(0.119)
% Males 15-19	-0.538***	-0.661***	-0.123	-0.538***	-0.596***	-0.472***
	(0.143)	(0.088)		(0.143)	(0.067)	(0.128)
% Female 15-19	-0.406***	-0.449***	-0.043	-0.406***	-0.510***	-0.467***
	(0.135)	(0.088)		(0.135)	(0.066)	(0.117)
% Males 20-34	-0.208	-0.441***	-0.233	-0.208	-0.411***	-0.178
	(0.134)	(0.081)		(0.134)	(0.06)	(0.118)
% Female 20-34	-0.302**	-0.169**	0.133	-0.302**	-0.209***	-0.342***
	(0.13)	(0.08)		(0.13)	(0.06)	(0.115)
% Males 35-59	0.017	-0.09	-0.107	0.017	-0.078	0.029
	(0.131)	(0.077)		(0.131)	(0.057)	(0.117)
% Female 35-59	-0.186	-0.022	0.164	-0.186	-0.124**	-0.288***
	(0.117)	(0.069)		(0.117)	(0.052)	(0.105)
% Males > 60	-0.17	-0.034	0.136	-0.17	-0.043	-0.178
	(0.144)	(0.083)		(0.144)	(0.062)	(0.128)

Notes: Standard errors in parenthesis.

Table 6: Bias Corrected OLS: Panel, Pooled and bias-corrected OLS (part I)

	Balanced Panel Households				All households	
Variable	FE	Pooled	Bias	Bias-C.	Pooled	Bias-C.
	98,01	98,01	Col. (2)-(1)	98,01	93,98,01	93,98,01
	Δl(pc Cons.)	l(pc Cons.)		l(pc Cons.)	l(pc Cons.)	l(pc Cons.)
% Prim. Ed.	0.032	0.309***	0.277	0.032	0.281***	0.003
	(0.043)	(0.027)		(0.043)	(0.019)	(0.038)
% Sec. Ed.	0.118**	0.590***	0.472	0.118**	0.562***	0.091*
	(0.058)	(0.03)		(0.058)	(0.023)	(0.054)
% Higher Ed.	0.150**	0.981***	0.831	0.150**	1.025***	0.194***
	(0.073)	(0.037)		(0.073)	(0.029)	(0.070)
% Self. Emp.	0.029	0.075***	0.046	0.029	0.106***	0.061***
	(0.03)	(0.024)		(0.03)	(0.017)	(0.023)
% Big Farmers	0.03	0.298***	0.268	0.03	0.326***	0.058
	(0.046)	(0.038)		(0.046)	(0.032)	(0.040)
% No Agr. Emp.	-0.02	0.044*	0.064	-0.02	0.056***	-0.009
	(0.032)	(0.025)		(0.032)	(0.018)	(0.025)
D. Water	0.029	0.140***	0.111	0.029	0.157***	0.045*
	(0.029)	(0.017)		(0.029)	(0.013)	(0.026)
D. Electricity	0.080**	0.238***	0.158	0.080**	0.274***	0.117***
	(0.035)	(0.019)		(0.035)	(0.014)	(0.033)
D. Dirt Floor	-0.032	-0.218***	-0.186	-0.032	-0.260***	-0.075***
	(0.029)	(0.015)		(0.029)	(0.012)	(0.026)
D. Prop. Reg.	-0.001	0.060***	0.061	-0.001	0.042***	-0.019
	(0.027)	(0.019)		(0.027)	(0.014)	(0.023)
D. Prop. Not-Reg.	0.014	-0.01	-0.024	0.014	-0.028*	-0.004
	(0.027)	(0.02)		(0.027)	(0.015)	(0.022)
Land Use	0.000	0.000**	0.000	0.000	0.001***	0.000
	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)
# Cattle	0.003***	0.006***	0.003	0.003***	0.005***	0.002**
	(0.001)	(0.001)		(0.001)	(0.000)	(0.001)
Constant	-	8.728***	-	-0.000	8.813***	0.085
	-	(0.078)	-	(0.078)	(0.058)	(0.058)
Observations	2791	5582		5582	12422	12422
Number of households	2791	2791		2791	9631	9631
R-squared	0.152	0.596		0.152	0.56	0.283

Notes: Standard errors in parenthesis.

Table 7: Bias Corrected OLS: Panel, Pooled and bias-corrected OLS (continued)