Misclassification of the dependent variable in a debt–repayment behavior context

Carlos Aller [a,*], Jorge González Chapela [b,1]

[a] Department of Economics, University of Alicante, Campus de San Vicente, E-03080 Alicante, Spain
[b] Department of Economics, University of Girona, Campus de Montilivi, E-17071 Girona, Spain

A R T I C L E   I N F O

A B S T R A C T

This paper estimates a model of the household debt–repayment decision that accounts for the possibility of misclassification of self-reported debt–repayment status. It likewise estimates the extent of misclassification in a sample of data from different European countries. The evidence suggests that many households that are in arrears do not report this condition, so that the true level of arrears is, on average, 24% higher than that observed in our data. Furthermore, the effects on the incidence of arrears of adverse income and expense shocks are substantially greater than those predicted by estimators that ignore the possibility of misclassification.

1. Introduction

For good reasons, many micro data studies of households' debt–repayment behavior rely on survey data, as, for example, Bridges and Disney (2004), Duygan-Bump and Grant (2009), Fay et al. (2002), Lyons and Fisher (2006), or May and Tudela (2005). Like administrative (or lenders') data, survey data that incorporate a panel or retrospective dimension allow tracking households' debt–repayment decisions, but these data may better represent the population in studies of personal bankruptcy filings or broadly defined arrears. Moreover, surveys tend to gather a richer set of potentially relevant household characteristics for analyses of households' debt repayment behavior.

Debt–repayment survey data are not free of drawbacks, though. A concern is raised by their self-reported nature, as even the household's financially responsible person may not provide perfectly reliable answers.[2] For one thing, default seems a socially undesirable behavior, so that questions about debt–repayment might elicit underreporting of financial problems (i.e. false negatives). On the other hand, debt–repayment problems might leave a strong trace in memory, and there is some evidence that events' importance leads to over-reporting (i.e. false positives) on the part of the respondent (a phenomenon referred to in the literature as *telescoping*; see for example Chase and Harada, 1984). The significant underreporting of bankruptcy filings in the Panel Study of Income Dynamics (Fay et al., 2002) and of mortgage arrears in the British Household Panel Survey (May and

---

* Corresponding author. Tel.: +34 965903400x3307.
  E-mail addresses: caller@ua.es (C. Aller), jorge.gonzalez@udg.edu (J. González Chapela).
[1] Tel.: +34 972418781.
[2] This is not the only concern with survey data. Typically, information on the temporal extension and/or magnitude of the repayment problem is not asked, so that the empirical proxy for "repayment problems" covers a wide range of default behaviors, ranging from bankruptcy to being a few days behind on repayment.

Tudela, 2005) suggest that households are indeed misclassified as to debt–repayment status, as well as that the former type of misclassification is more prevalent than the latter.

Empirical studies of households' debt–repayment behavior typically run probit or logit regressions explaining whether households have experienced debt–repayment problems during certain periods of time. Hausman et al. (1998), however, have shown that these traditional estimation techniques generally lead to inconsistent coefficient estimates when the dependent variable (in our case, the household debt–repayment status) is misclassified. While it is difficult to calculate the degree of inconsistency analytically, the simulation results in Hausman et al. (1998) and Hug (2010) indicate that relatively small amounts of misclassification (as little as 2%) can lead to significant amount of bias even in large samples. For this reason, traditional estimates are to be reexamined in order to obtain consistent, superior estimates.

In this paper, the modified maximum likelihood estimator proposed in Hausman et al. (1998) (referred hereafter as HAS 98) is applied to a standard model of debt–repayment behavior with a commonly used data set. This estimator, which is a natural extension of the parametric methods employed in the debt–repayment literature, allows correcting the bias caused by misclassification and estimating its extent, under the assumption that the extent of misclassification in the data is not too high. We are not the first applying the HAS 98 estimator or some generalization of it, which has been used in studies concerning job change (Falaris, 2011; Hausman et al., 1998), smoking behavior (Brachet, 2008), student cheating behavior (Caudill and Mixon, 2005), insurance fraud (Artís et al., 2002), language proficiency (Coulon and Wolff, 2007; Dustmann and van Soest, 2001), or civil wars (Hug, 2010).

Section 2 of the paper describes the data and the sample design. Our choice of data has been motivated by the contention that households that might eventually opt to default will take financial benefits and costs into account in making their debt–repayment decisions. Hence, and as in Duygan-Bump and Grant (2009), we have combined household-level information on being behind on repayment with country-level measures of default incentives. Although failure to do a monthly payment on time is generally not considered default, it may be viewed by lenders as a signal that credits could be at risk.

Section 3 presents the econometric specification and Section 4 discusses the results. The probabilities of false negatives that we have obtained range from essentially 0 in Ireland and The Netherlands to approximately 0.64 in Denmark, and tend to be higher for households with income above the country median income. Probabilities of false positives are much lower, ranging from almost 0 in The Netherlands to 0.04 in Greece, and differ little by income group. When the possibility of misclassification is taken into account, the overall level of arrears in the data (8.9%) is 24% higher than the observed amount (7.2%). Furthermore, the effects on the incidence of arrears of adverse income and expense shocks to households, and of countries' judicial and financial institutions, are generally larger. For example, a 10% cut in household income would increase the number of households being behind on repayment by about 13%, instead of the 9% increase predicted by traditional estimates. Likewise, increasing the cost of judicial proceedings from 0.5% of the claimed debt (as was the case of The Netherlands during the observation period) to 10.7% (as was the case of Spain) would increase the incidence of arrears after a negative health shock by about 48%, instead of the 18% increase predicted by traditional estimates. Section 5 concludes.

2. Data and sample design

The European Community Household Panel (ECHP) is an annual survey that interviewed a representative sample of households and individuals in each of 14 European countries between 1994 and 2001.[3] In addition to data on income, employment, health, and marital status of all household adults, which we will use to construct indicators for income and expense shocks at the household level, the ECHP asked the household respondent information about missed scheduled payments: "Has your household been in arrears at any time during the past 12 months, that is, unable to pay as scheduled any of the following?" The respondent then answered "Yes" or "No" to each of the following: Rent for accommodation, mortgage payments, utility bills, and hire purchase installments or other loan repayments. We consider any "Yes" answer to be indicative of arrears, whereby virtually any household may be in arrears. More restricted definitions could raise statistical concerns. Suppose, for example, that we wanted to analyze mortgage arrears only. If a sample of mortgagors were used, it would be probably selected in terms of repayment propensity. If, on the other hand, any household were considered a mortgage applicant and the whole sample of households were used, non-mortgagors would have to be arbitrarily assigned the status of "not in arrears", thus stirring up more misclassification.

The Doing Business (DB) project, launched by the World Bank in 2002, gathers quantitative indicators on the regulations that apply to local firms in 185 economies. The indicators cover a wide range of dimensions of the regulatory environment, including the complexity and cost of starting a business and the number of procedures, time, and cost needed to resolve a commercial dispute. More than 18,000 professionals assist in providing the data that inform the indicators, including legal professionals and credit bureaus officials. See www.doingbusiness.org for further details. To represent cross-country differences in the cost of default, and following Duygan-Bump and Grant (2009), we take advantage of three contract enforcement indicators and a measure of coverage by credit bureaus gathered by the DB project: The average number of calendar days it takes for dispute resolution; the cost, as a percentage of the debt, of judicial proceedings; the number of legal procedures mandated by law that must be followed to legally recover a debt; and the percentage of the adult population who has information on their repayment history, unpaid debts, or outstanding credit recorded in public or private registries.[4] The purpose of the first three indicators is to

---

[3] See http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/echp for detailed information on ECHP.

[4] The DB project data were accessed in May 2006. As to the coverage measure, Jappelli and Pagano (2002) have found that the impact on defaults of private arrangements to share credit information is similar to that of public credit registries.