Journal of Empirical Finance 37 (2016) 186–195

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Private information and limitations of Heckman's estimator in banking and corporate finance research

Randall C. Campbell [a,*], Gregory L. Nagel [b]

[a] College of Business, Mississippi State University, Mississippi State, MS 39762, USA
[b] Jones College of Business, Middle Tennessee State University, Murfreesboro, TN 37132, USA

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 22 March 2016
Accepted 25 March 2016
Available online 4 April 2016

JEL classification:
C15
C24
C53
G30

Keywords:
Private information
Self-selection
Sample selection
Omitted variables

ABSTRACT

Private information is a common problem in banking and corporate finance research. Heckman's (1979) two-step estimator is commonly used to test for sample selection using a simple *t*-test on the inverse Mills ratio (IMR) coefficient. Following Puri (1996), this test is often interpreted as a test for private information. We conduct a series of Monte Carlo simulations to show that researchers can reliably use the Heckman estimator to test for private information when this private information is random. However, private information often takes the form of an omitted variable with a deterministic relationship to selection and outcomes. In this case, we show that the IMR coefficient is biased and inconsistent and that *t*-tests lead to incorrect conclusions regarding the significance of private information as well as its impact on selection and outcomes. We illustrate our results using a unique case in prior literature in which a bank's prior information was revealed. In conclusion, the Heckman model cannot be interpreted as a test for private information (or sample selection) when private information takes the form of an omitted variable in the first-stage regression.

1. Introduction

Private information affects selection for bank loans and therefore outcomes such as the offered rate, as well as other areas of finance including yield spreads for reputable underwriters, diversification decisions, public debt offerings, analyst coverage, and bankruptcy.[1] Succinctly put, Agarwal and Hauswald (2010) state, "Private information and its distribution are among the fundamental forces shaping economic exchange." Puri (1996) proposes a method for estimating private information and its effect on outcomes in sample selection models, which are common in banking and corporate finance. We use Monte Carlo experiments to examine the performance of sample selection models when there is private information. Our focus is on private information that takes the form of an omitted variable and its impact on estimation and interpretation of the selection term (inverse Mills ratio) in the Heckman model.

Sample selection models, whereby the outcome of interest is observed only for a proportion of the sample, are common in banking and corporate finance. For example, we only observe the interest rate paid on a loan for individuals who are granted a loan, but not for those whose loan application is denied. Sample selection models can be consistently estimated using either

* Corresponding author at: Department of Finance and Economics, Mississippi State University, Mississippi State, MS, 39762, USA. Tel.: +1 662 325 1516.
E-mail addresses: rcampbell@business.msstate.edu (R.C. Campbell), Greg.Nagel@mtsu.edu (G.L. Nagel).

[1] See Puri (1996), Campbell and Loumioti (2013), Chen et al. (2013), Agarwal and Hauswald (2010), Goss and Roberts (2011), Andres et al. (2014), Campa and Kedia (2002), and Li and Prabhala (2007).

http://dx.doi.org/10.1016/j.jempfin.2016.03.007
0927-5398/

maximum likelihood or Heckman's (1979) two-step estimator. In our experiments, we focus on the Heckman estimator for two reasons. First, although the maximum likelihood estimator is efficient, as Greene (2012, p. 878) notes "the MLE is by far less common than the two-step estimator in the received applications." While maximum likelihood estimation of sample selection models has become more common in recent years, Puri (1996) and other papers that test for private information tend to use the Heckman model.[2] Moreover, we note that our results would apply to tests for private information in either case; the pattern of bias does not depend on the choice of estimator. A second reason for focusing on the Heckman estimator is ease of exposition. The inclusion of the inverse Mills ratio (hereafter IMR) directly into the outcome equation provides a simple and intuitive measure of the observed biases when there is an omitted variable in the selection equation.

Puri (1996), in a bank underwriting model, is the first to measure the effect of private information using the Heckman (1979) two-step estimator. She interprets the IMR coefficient as expected private information; this interpretation has found acceptance in the literature. Li and Prabhala (2007) cite an emerging line of literature that interprets the IMR as expected private information in many areas of finance.[3]

Consider the standard selection equation $z_i^* = w_i{}'\gamma + u_i$, where the researcher only observes $z_i = 0$ or 1, depending on the sign of z_i^*, and the error term is assumed to be mean zero. As Puri (1996) shows, the conditional expectation $E(\mu_i|z_i^* \geq 0)$ is equal to the IMR, and so may be interpreted as expected private information. We refer to this as *random private information*. Note that the private information may differ, both in terms of data and how it is weighted, for different observations. For example, Puri (1996, p. 374) states that, when making an underwriting decision, commercial banks may obtain private information via means such as "scrutiny of internal budget statements and factory/inventory inspections."

An alternative type of private information is that of an omitted variable which is not available to the researcher (Roberts and Whited, 2012), such as a proprietary credit score (see Agarwal and Hauswald, 2010). This type of private information is uniformly interpreted by decision makers (banks, for example) and thus has a deterministic relationship to selection and outcomes; therefore we call it *deterministic private information*. In this case, the selection equation is $z_i^* = w_i{}'\gamma + h_i'\delta + u_i$, but the researcher does not observe h_i. Thus, the selection equation for the Heckman model (or the maximum likelihood estimator) suffers from omitted variable bias.

The distinction between these two types of private information is important in applied research. Although a *t*-test of the IMR term can be considered a test for private information under the assumptions of Puri (1996), applying this as a test of private information in all cases (i.e., when there is an omitted variable) leads to erroneous conclusions. We investigate the potential magnitude of such errors by conducting Monte Carlo experiments under both random and deterministic private information.

Our simulations confirm that under random private information, estimation of the IMR coefficient using the Heckman model is consistent and that *t*-tests tend to reject the null $H_0: \beta_\lambda = 0$ at reasonable rates. However, when there is an omitted variable in the selection equation, the IMR coefficient in the outcome equation is biased and *t*-tests involving the IMR coefficient are unreliable. We interpret the empirical results in Agarwal and Hauswald (2010) and show that they are consistent with our simulations. Their results are unique in that a bank's deterministic private information was obtained, which provides an illustration of the conclusions from our simulations. We show that the Heckman model performs well when private information is random in nature, but that applying this methodology when private information takes the form of an omitted variable leads to erroneous conclusions.

The rest of the paper is organized as follows. Section 2 gives a description of Heckman's selection model and describes the difference between random and deterministic information in the context of this model. Section 3 describes data generation for the Monte Carlo experiments and the experimental design. Experimental results regarding model estimation are discussed in Section 4. Section 5 concludes.

2. Modeling the effect of private information on selection and outcomes

We estimate the effect of private information on selection and outcomes using the standard Heckman (1979) model:

$$z_i^* = w_i{}'\gamma + u_i, \tag{1}$$

$$y_i = x_i{}'\beta + e_i, \tag{2}$$

where y_i is observed only if $z_i^* \geq 0$. In this model, a researcher only observes whether selection into the outcome population occurs; that is, they observe the sign of z_i^* not its magnitude. Eq. (1) is the latent selection equation, and Eq. (2) is the outcome equation. The random errors (u_i and e_i) are traditionally assumed to be bivariate normal with correlation ρ. If ρ is equal to zero, there is no selection bias. However, if $\rho \neq 0$ there is a selection bias, which causes the OLS estimator of the outcome equation to be biased and inconsistent. Typically, the variance of u_i is normalized to 1. Thus,

$$\begin{pmatrix} u_i \\ e_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_e \\ \rho\sigma_e & \sigma_e^2 \end{pmatrix}\right]. \tag{3}$$

[2] From 2010-2013, we count roughly 60 articles altogether in the *Journal of Finance*, *Review of Financial Studies*, *Journal of Financial Economics*, *Journal of Financial and Quantitative Analysis*, *Journal of Corporate Finance*, and *Journal of Banking and Finance* that test for selection bias and/or private information using Heckman's model.

[3] Papers that specifically interpret the IMR term as private information include Li and Prabhala (2007); Ayyagari et al. (2010); McCahery and Schwienbacher (2010); Bayar and Chemmanur (2012); Ongena et al. (2013), and Andres et al. (2014).