



Regression analysis of proportions in finance with self selection

Douglas O. Cook^{a,1}, Robert Kieschnick^{b,*}, B.D. McCullough^{c,2}

^a Department of Economics, Finance, and Legal Studies, Culverhouse College of Business, University of Alabama, Tuscaloosa, AL 35487-0224, United States

^b University of Texas at Dallas, 2601 N. Floyd Rd, SM 31, Richardson, TX 75080, United States

^c Drexel University, LeBow College of Business, Philadelphia, PA 19104-2875, United States

ARTICLE INFO

Article history:

Received 30 March 2006

Received in revised form 12 December 2007

Accepted 11 February 2008

Available online 4 March 2008

JEL classification:

G3

C2

Keywords:

Proportions

Zero-inflated beta

Capital structure

ABSTRACT

Numerous papers in finance study the conditional mean of some proportion or fraction with a mass point at zero. We argue that most, if not all, of these studies use mis-specified statistical models, especially when firms or individuals choose to not do something for different reasons. To address these issues, we develop a new statistical model, the zero-inflated beta model, and apply it to the analysis of corporate capital structure decisions to demonstrate its applicability.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Numerous papers in finance study the conditional mean of some proportion or fraction with a mass point at zero. Table 1 provides a range of such examples. All of these, and similar studies, share a number of specification errors that motivate our paper. First, all of these studies model the proportion under study as a linear function of the explanatory variables. Second, they all ignore the fact that the conditional variance must be a function of the conditional mean. Third, most, but not all, of these studies fail to recognize that firms or individuals might make the choice to do or not do something for different reasons. As a result, the estimates of the regression coefficients and their variances in these studies are biased and inconsistent; which raises questions about their conclusions. The purpose of this study is to propose a statistical model that addresses these concerns and demonstrate its relevance to these types of data using a case study.

Cox (1996), Papke and Wooldridge (1996), Paolina (2001), Kieschnick and McCullough (2003), and Ferrari and Cribari-Neto (2004) discuss and provide evidence on the first two specification errors. First, the conditional expectation of a continuous proportion or fractional variate is only defined on the bounded interval, [0,1]. Therefore, the conditional expectation must be a nonlinear function of the explanatory variables. Second, the conditional variance must be a function of the conditional mean since the conditional variance must change as the conditional mean approaches either boundary. Consequently, studies of the conditional mean of a proportion in finance using a linear conditional expectation function estimated by least squares or

* Corresponding author. Tel.: +1 972 883 2799.

E-mail addresses: dcook@cba.ua.edu (D.O. Cook), rkiesch@utdallas.edu (R. Kieschnick), bdmccullough@drexel.edu (B.D. McCullough).

¹ Tel.: +1 205 750 8887.

² Tel.: +1 215 895 2134.

Table 1

Examples of studies in finance that study the conditional mean of a proportion

	Dependent variable	Dependent variable range	Regression equation	Estimation procedure
Baker and Wurgler (2002)	Debt/Assets (where debt is both market and book)	[0,1]	Linear	OLS and Fama-MacBeth
Barclay and Smith (1995)	Percentage of total debt maturing in more than 3 years	[0%, 100%]	Linear	Pooled, Cross-Sectional, and Fixed Effects Regressions
Johnson (1997)	Type of Debt/Total Debt where type is debt held publicly, privately, or by banks	[0, 1]	Linear	
Houston and James (1996)	Bank Loans/Total Debt	[0,1]	Linear	OLS
Agrawal and Jayaraman (1994)	Dividend Payout (Dividends/Earnings)	[0,1]	Linear	OLS
Demsetz and Villalonga (2001)	Management's share of the firm's stock	[0,1]	Linear	Instrumental Variable
Arthur (2001)	The proportion of outside directors on the board of directors	[0,1]	Linear	OLS
Aussenegg, Pichler, and Stomper (2006)	IPO price revisions	[0,1]	Linear	OLS
Agnew (2005)	Company stock allocation	[0,1]	Linear	Tobit
Guiso, Sapienza, and Zingales (2007)	Share of portfolio invested in stock	[0,1]	Linear	Tobit

instrumental variables typically mis-specify both the mean and the variance structure of the conditional distribution under study and so use biased and inconsistent estimators of the coefficients and their standard errors.

To address these two specification errors, Cox (1996), Papke and Wooldridge (1996), Paolina (2001) and Kieschnick and McCullough (2003), Ferrari and Cribari-Neto (2004) examine the specification of regression models for proportional or fractional data observed on (0,1). Consistent with the above points, they find evidence confirming that for such data the conditional expectation function is nonlinear, and the conditional variance is a function of the mean. Of the various econometric specifications that Kieschnick and McCullough (2003) test, they fail to reject the applicability of either a regression model based upon the beta distribution that they propose or the quasi-likelihood model proposed by Papke and Wooldridge (1996). Either of these regression models uses a link function that is consistent with the evidence report in Cox (1996) as the preferred specification.

The third specification error arises from the fact that prior research has presumed that the influences on a firm's decision to issue some type of financing, for example, are the same, in magnitude and composition, as those that influence its decision on how much of this type of financing to use. Violation of this assumption is the essence of sample selection bias, which Heckman (1979) demonstrates is another type of specification error. Li and Prabhala (2005) provide an extensive discussion of why violation of this assumption is to be expected in many corporate finance topics.

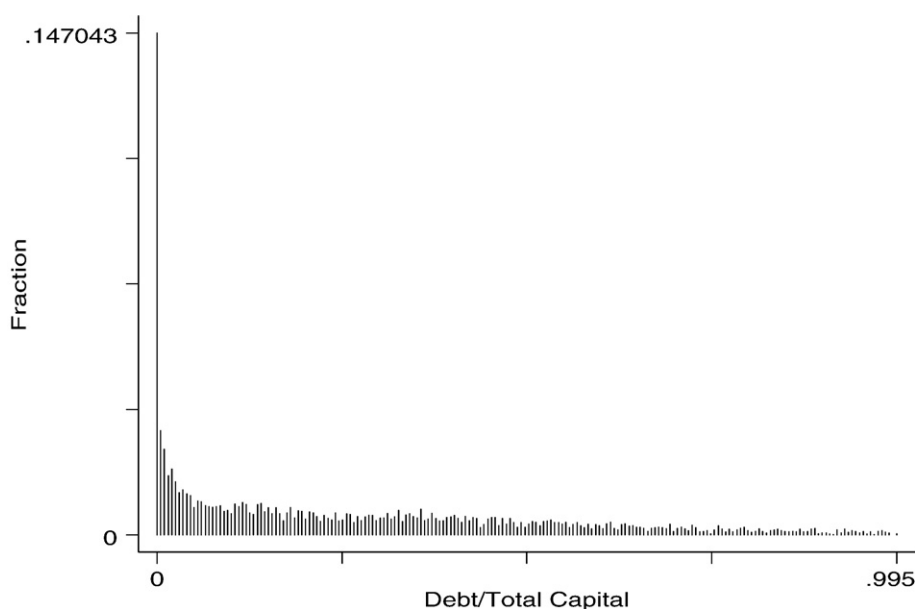


Fig. 1. Debt to total capital. This figure represents the proportion of a firm's total capital accounted for by its debt capital, as defined in Rajan and Zingales (1995), using a sample of public corporations with positive sales in 1991.

To see the prevalence of such boundary observations in studies of the conditional mean of some proportion in corporate finance, we display in Fig. 1 a graph of the proportion of capital accounted for by debt.³ Clearly this figure illustrates that boundary observations can be an important feature of these data. As explained in most textbook treatments (e.g., Peracchi, 2001) of limited dependent variables, studies that ignore this fact will produce coefficient and standard error estimates that are biased and inconsistent.

The censored normal regression model (i.e., the Tobit model) is sometimes used to address such observations (e.g., Rajan and Zingales, 1995). While seemingly appropriate for modeling the conditional expectation of a continuously measured proportion, the censored normal regression model is a conceptually flawed model for proportional data. As Maddala (1991) observes, these data are not observationally censored but rather are *defined* only over the interval [0,1]. Therefore, as noted earlier, the conditional mean must be a nonlinear function of the regressors, and heteroskedasticity becomes a concern.

In addition to these concerns, the Tobit model ignores the potential heterogeneity of the sample by assuming that each observation is drawn from a common distribution. This restrictive assumption, as Lin and Schmidt (1984) point out, implies that “in the Tobit model any variable that increases the probability of a non-zero value must also increase the mean of the positive values,” and by the same amount.⁴ Or put differently, the Tobit model imposes the restriction that the factors that influence whether or not a firm uses long-term debt, for example, exercise exactly the same influence on how much long-term debt the firm uses. Sample selection and mixture models avoid imposing this assumption, and so, we argue, should researchers examining the conditional mean of some proportion that shows evidence of clustering at 0.

To address the above issues, we develop a new statistical model, the zero inflated beta regression model, and demonstrate its applicability to these types of data using a case study. To present this material we organize our paper as follows. Section 2 provides an alternative statistical model for studies that examine the conditional mean of a proportion that allows for zeros to be generated by a different process. Section 3 uses a case study to provide empirical evidence that the conjectured specification errors do arise, and to illustrate the applicability of our proposed regression model for these data. Section 4 concludes with a summary of our results.

Using data on corporate capital structures, we find provide evidence that for the proportion of capital accounted for by debt: (1) the conditional expectation function is nonlinear, (2) the conditional variance is heteroskedastic, and (3) the boundary observations are drawn from a different regime. All of these results motivate the need to use a statistical model that addresses these issues, which we also develop and demonstrate to fit these data better. Further, we show through our case study that the use of this model leads to different estimates and inferences than one would derive using the standard regression models applied to these types of data in financial research. Regardless of whether one uses the model we propose, or some alternative, it is important that the regression model used address the specification issues that we raise for these types of data.

2. An alternative approach to modeling the conditional mean of proportional data that have a probability mass at zero

To address the issues that we raise, it is important to recognize that discretely distributed or continuously distributed random variables belong to an even larger class of random variables called mixed, or mixed discrete-continuous, random variables.⁵ Going back to Aitchison (1955), such distributional models have had a long history in consumer expenditure studies, where clustering at zero is an issue.⁶ A generic representation of such probability density functions is:

$$g(X; \theta) = \begin{cases} 0, & \text{if } X < 0 \\ \delta, & \text{if } X = 0 \\ (1 - \delta)f(X; \theta) & \text{if } X > 0 \end{cases} \quad (1)$$

where $f(X; \theta)$ is any absolutely continuous probability density function defined over the positive real line. From this expression, we can represent the cumulative distribution function as:

$$G(X; \theta) = \delta + (1 - \delta)F(X; \theta), \quad \text{for } X \geq 0 \quad (2)$$

where $F(X; \theta)$ is the cumulative distribution function for the absolutely continuous part of the distribution.

A key point illustrated by the above example is that mixed discrete-continuous random variables indicate heterogeneity in the data; a point made more formally in Bock (1996). This point raises the issue of whether all the observations are generated by the same processes. In this regard, we note that Cragg (1971) or Blundell and Meghir (1987) point out the censored normal or Tobit model assumes that the sets of influences are the same and that they exercise the same influence. In contrast, mixed discrete-continuous models do not impose this restriction.

Effectively, mixed discrete-continuous models allow that the decision to use a particular form of financing, for example, to be influenced by either a different set of variables or influenced differently by the variables that determine how much of that

³ We can provide additional graphical illustrations of this pattern using managerial ownership, the proportion of debt that is private debt, etc.

⁴ Lin and Schmidt (1984), page 174.

⁵ See Theorem 4.1.6 of Dudewicz and Mishra (1988). This theorem shows that discrete or continuous distributions can be viewed as sub-sets of mixed distributions.

⁶ A reader familiar with corporate investment data will recognize that this raises questions about some of the statistical specifications employed in the study of the sensitivity of corporate capital expenditures to financial constraints. However, this is not an issue for this study.

particular form of financing to use. Conceptually, they imply that theoretical models which conflate the two decisions into one are misconceptions of the decision-making processes followed by individuals or firms, which is the essence of selection bias.

Regression models based upon such distributional assumptions have taken on a variety of names and forms. For example, the zero-inflated Poisson model is such a model and is used to account for excessive zeros in count data. Some, like Cragg (1971), would identify the zero-inflated Poisson model as a “single hurdle” model. In the context of consumer expenditure studies, such models have often been called type-2 Tobit models.⁷ Blundell and Meghir (1987) follow Wilks (1962) and identify these as bivariate distributions, whereas others, like Manning, Duan and Rogers (1987) identify these models as two-part models.⁸ Typically, the naming of such models depends upon the naming of the non-singular component distribution.

With respect to the continuous component of our hypothesized mixed discrete-continuous distribution, we use a beta distribution. There are two primary motivations for this specification. First, it is the standard statistical model for proportional or fractional data observed on the (0,1) interval (Spanos, 1999).⁹ Second, as Johnson, Kotz, and Balakrishnan (1995) document, many empirical studies have found it to be a good characterization of such data.¹⁰ Thus, we define the zero-inflated beta probability density as:

$$g(y; \theta) = \begin{cases} 0, & \text{if } y < 0 \\ \delta, & \text{if } y = 0 \\ (1 - \delta)f(X; \theta) & \text{if } 0 < y < 1 \end{cases} \quad (3)$$

where:

$$f(y : p, q) = \left[\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1} \right] \quad \text{for } 0 < y < 1. \quad (4)$$

This representation uses the two parameter beta distribution for the continuous portion of the distribution since both Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004) provide evidence that a regression model based upon the two parameter beta distribution is a reasonable specification for examining the conditional mean of data distributed over (0,1). We can write the mean and variance of this distribution as:

$$\begin{aligned} E(y) &= (1 - \delta)\mu \\ \text{Var}(y) &= (1 - \delta)V(\mu) + \delta(1 - \delta)\mu^2 \end{aligned} \quad (5)$$

where μ and $V(\mu)$ are the mean and variance of the two parameter beta distribution.

Studying this expression is useful because it illustrates the effect of a mass point on the expected value of a mixed discrete-continuous random variable. Since the mass point is 0 for most of the studies reported in Table 1, this means that researchers are likely to predict more corporate debt use, for example, than will be observed if they base their predictions on the standard linear regression model. This fact is relevant to studies that try to explain why corporations use less debt than predicted by the trade-off model (e.g., Berens and Cuny, 1995 or Ju et al., 2005).

Following Cragg (1971), we formulate our zero-inflated beta regression model as:

$$f(y_i = 0 : \mathbf{X}_i) = 1 - C(\alpha' \mathbf{X}_i) \quad \text{for } y_i = 0, \quad (6)$$

and

$$f(y_i : \mathbf{X}_i) = C(\alpha' \mathbf{X}_i) \left[\frac{\Gamma(p + q(\mathbf{X}_i))}{\Gamma(p)\Gamma(q(\mathbf{X}_i))} y^{p-1} (1 - y)^{q(\mathbf{X}_i)-1} \right] \quad \text{for } 0 < y_i < 1, \quad (7)$$

where $q(\mathbf{X}_i) = p \exp(-\beta' \mathbf{X}_i)$ and p is a parameter of the beta distribution.¹¹ In this specification, $C(\alpha' \mathbf{X}_i)$ represents the probability of a firm choosing to use or do something. We use the cumulative logistic function for this probability in our study, as this is consistent with Papke and Wooldridge's (1996) treatment, and is consistent with the typical approach of researchers studying expenditure data (e.g., Yoo, 2004). We refer to this portion of our model as the selection equation following the nomenclature of the self-selection literature.

Note that we have allowed the coefficients of the exogenous variables to differ in their effect on the decision to use or do something, from their effect on how much to use or do (i.e., the vector α and β are not required to be the same). Further, we can allow the variables that influence a firm's choice of regime to differ from those that influence its choice of level. The second term of Eq. (7) (within brackets) uses the beta distribution to represent the distribution of proportions of something (e.g., the proportion of

⁷ See Peracchi (2001) for an excellent discussion of these and related bivariate models.

⁸ It is worth pointing out that such models directly address self-selection issues.

⁹ Further, the beta distribution nests a number of other distributions, such as the uniform distribution and so we let our data decide if one of the nested distributions is appropriate. For example, our data reject the uniform distribution as an appropriate distributional model for our debt use measure.

¹⁰ Johnson, Kotz, and Balakrishnan (1995) provide dozens of empirical studies for which the beta distribution has been found to be an appropriate distributional model for proportional data.

¹¹ We can show that the zero-inflated beta distribution is a member of the three parameter nonlinear exponential family. As a consequence, the zero-inflated regression model shares all the properties of regression models based on the nonlinear exponential family and so we do not explore its properties further in this paper (see Gay and Welsch, 1988 for instance).

Table 2
Sample summary statistics

	Mean	Median	Standard Deviation
Debt ratio	0.1780	0.0794	0.2906
Tang	0.2906	0.2336	0.2233
Mtbr	2.5256	1.5106	9.1026
Logsale	4.0577	4.1231	2.4175
Profit	0.0332	0.1087	0.6033

The sample of 4231 firms comprises firms with positive sales during 1992, complete data on Compustat, and excludes regulated (SICs 41–49) or financial (SICs 60–67) firms. *Debt ratio* equals the proportion of a firm's capital, long-term debt (book value) and equity (market value), accounted for by its long-term debt. *Tang* represents the ratio of fixed assets to the book value of total assets. *Mtbr* is the ratio of the book value of assets less the book value of equity plus the market value of equity all divided by the book value of assets. *Logsale* represents the logarithm of net sales. *Profit* equals EBITDA divided by the book value of assets.

capital accounted for by debt capital), given that the firm or individual uses that something (e.g., debt). We identify this regression model as the zero-inflated beta regression model.

3. Case study: the analysis of corporate capital structures

There are too many potential applications for which the above regression model might be appropriate for us to explore its applicability in all of its potential applications. Consequently we focus on one application; that of corporate capital structure analysis.¹²

With this purpose in view, we use [Rajan and Zingales \(1995\)](#) as the template for our study of corporate capital structures as it has served as the starting point for specification issues in a number of recent studies (e.g., [Baker and Wurgler, 2002](#) and [Fama and French, 2002](#)).¹³ This template allows us to defer issues concerning the specifications of the appropriate regressors and to focus only on the issues under study. Specifically, we collect Compustat data for 1992 including all firms with positive sales in 1992. Following Rajan and Zingales, we refine this sample further by dropping firms in either SICs 60–67 (financial service industries) or SICs 41–49 (regulated industries). Our resulting sample thus represents a cross-section of 4231 firms.¹⁴

With this sample, we use the same regressors as employed by Rajan and Zingales. We use *tang* (the ratio of fixed assets to the book value of total assets), *mtbr* (the ratio of the book value of assets less the book value of equity plus the market value of equity all divided by the book value of assets), *logsale* (the natural logarithm of net sales), and *profit* (EBITDA divided by the book value of assets). We report summary statistics on each of these variables in [Table 2](#).

In addition to these variables, we define our measure of corporate leverage as the ratio of long-term debt to the sum of long-term debt and the market value of equity, where the market value of equity is defined as the number of shares outstanding times the closing price at fiscal year-end. We focus on this specific leverage definition as it is consistent with that used in a number of studies and our results are robust to the typical market value definitions of financial leverage.¹⁵ For our sample, 813 firms or 19.22% of the firms use no long-term debt.

Using these variables, we then estimate the censored normal (Tobit), the Quasi-likelihood, and the zero-inflated beta (ZIB) regression models for these data. We estimate these models because all of these models allow for the clustering of observations at zero.¹⁶ However, the quasi-likelihood model relaxes the nonlinear mean and conditional variance assumptions of the Tobit model, and the ZIB model then relaxes sample selection assumption of the quasi-likelihood model. We report these results in [Table 3](#).

One way to determine which of these models best describes the data is to compare the correlation between the actual and predicted values of corporate leverage. Since it is obvious that Pearson correlation measure is inappropriate for these data, we compute the Spearman rank correlation between predicted and actual regressands. These correlations, reported in the last row of [Table 3](#), suggest that the zero-inflated beta model best describes the data. As in [Kieschnick and McCullough \(2003\)](#), the use of a nonlinear conditional expectation function (in the quasi-likelihood model) improves the fit between actual and predicted values over a linear conditional expectation function (the Tobit model in our study), and changes the inferences concerning the significance of different regressors. Consequently, comparing the Tobit and Quasi-likelihood results illustrates the importance of addressing the first two specification errors identified earlier while also accounting for boundary observations.

¹² It is worth noting that we also find evidence that proposed statistical model is applicable to both the analysis of debt structures and managerial ownership structures. However, we do not report these analyses as they raise issues about the specification of the appropriate set of explanatory variables that we avoid by focusing on the well known [Rajan and Zingales \(1995\)](#) study.

¹³ We recognize that [Welch \(2007\)](#) raises another set of concerns about these and similar capital structure studies, but his concerns do not negate ours and so we do not address them in this paper.

¹⁴ We focus on a single cross-section rather than a panel. Examining the cross-section allows us to test the basic specification issues that are of concern in our study. Examining a panel over time introduces additional statistical issues, which would require agreement on the basic specification issues that we identify. For example, in addition to issues concerning the correlation matrix, [Frank and Goyal \(2004\)](#) and [Welch \(2007\)](#) point out that panels introduce survival biases or missing-data induced errors.

¹⁵ Using book value definitions of financial leverage introduces an additional problem in that some firms have negative book values of equity which implies leverage use in excess of 100%, which is nonsensical.

¹⁶ We also estimated a linear model using least squares (OLS), but do not report the results because the results are not sufficiently different from the Tobit results to warrant separate attention.

Table 3

Regression results for different models

	Tobit regression model	Quasi-likelihood regression model	Zero-inflated beta regression model
Constant	−0.0660 (0.00)	−1.8309 (0.00)	−2.0360 (0.00)
Tang	0.3151 (0.00)	1.4505 (0.00)	0.9673 (0.00)
Mtbr	−0.0015 (0.00)	−0.4554 (0.04)	−0.0031 (0.01)
Logsale	0.0304 (0.00)	0.1555 (0.00)	0.1048 (0.00)
Profit	−0.0104 (0.15)	−0.7485 (0.00)	−0.1174 (0.00)
<i>Selection equation:</i>			
Constant			−0.2628 (0.00)
Tang			2.666 (0.00)
Mtbr			−0.0008 (0.81)
Logsale			0.2948 (0.00)
Profit			0.0484 (0.41)
Spearman Rank Correlation between predicted and actual	0.3677	0.4767	0.6174

The sample of 4231 firms requires positive sales during 1992, complete data on Compustat, and excludes regulated (SICs 41–49) or financial (SICs 60–67) firms. Following [Rajan and Zingales \(1995\)](#), the dependent variable in each of these regressions is the proportion of a firm's capital, long-term debt (book value) and equity (market value), accounted for by its long-term debt. *Tang* equals the ratio of fixed assets to the book value of total assets. *Mtbr* represents the ratio of the book value of assets less the book value of equity plus the market value of equity all divided by the book value of assets. *Logsale* is the logarithm of net sales. *Profit* equals EBITDA divided by the book value of assets. We report *p*-values associated with a two-sided *t*-test of the difference between the estimated coefficients and 0 in the parentheses.

Comparing the ZIB and Quasi-likelihood results illustrates the importance of addressing the third specification error as well: the ability to discern differences in the influence of variables on boundary observations from non-boundary observations. It is critically important to recognize that the coefficient estimates and their significance are different between level and selection equations; which is not the case for regression models that ignore their potential differences.¹⁷ These data demonstrate that our third specification issue is a relevant concern.

Altogether, these results also raise doubts about studies, like [Hovakimian, Opler, and Titman \(2001\)](#) or [Flannery and Rangan \(2006\)](#) that examine the effects of deviations from a firm's "target" capital structure on its issuance of debt or equity since they clearly mis-estimate the "target" capital structure for each firm. One of the critical points of nonlinear models is the estimates of the economic effects of a regressor depends on where that regressor is evaluated as it is varying over the range of the variable in a nonlinear manner.

Given these results, we next run a battery of specification tests. First, we examine whether the boundary observations are drawn from a different regime by following [Ruud \(1984\)](#) and [Lin and Schmidt \(1984\)](#) and applying a Hausman type comparison of the coefficient estimates from a probit and Tobit model. The Chi-square statistic of 303.71 with 4 degrees of freedom rejects the null hypothesis of their originating from a common regime. Given this result, we next focus on the specification of the conditional distribution of the non-boundary observations. Using a nonlinear link test (see [Fahrmeir and Tutz, 1996](#)), we find that the squared residuals are significant at the 1% level.¹⁸ Using the Breusch–Pagan test for heteroskedasticity, we obtain a Chi-Square statistic of 14.37 with 1 degree of freedom, which is significant at the 1% level. Altogether, our specification tests indicate that the data are consistent with the key features of a zero-inflated beta model.

Statistically, our conclusions imply that estimates of the coefficients and their standard deviations derived in prior research on corporate financing are invalid because they are biased and inconsistent. This fact is especially evident for those firms that do not use long-term debt, which illustrates the selection bias in these studies.

To reinforce this point, we expand our regression model to incorporate managerial stock ownership information as [Agrawal and Nagarajan \(1990\)](#) provide evidence that firms that do not use long-term debt are distinguished by managers who own more of their firm's stock. Such evidence is consistent with [Lewellen's \(2006\)](#) argument that managerial incentives influence their firm's capital structure decisions.

To do this, we match our sample firms against those available on Execucomp's database as it provides information on CEO and managerial shareholdings. For 1992, we are able to match 1,560 firms. Using these data, we create two new variables, *CEO's stock ownership* and *Managers' stock ownership*. The first variable represents the fraction of a firm's stock held by its CEO, and the second variable represents the fraction of a firm's stock held by its managers.

In addition to these ownership variables, we create a new variable to capture a firm's "free cash flows", or more specifically its operating cash flows after payments to shareholders. We add this variable because [Agrawal and Jayaraman \(1994\)](#) suggest that [Jensen's \(1986\)](#) agency costs of free cash flows are a concern in such firms. We measure this variable by the ratio of a firm's operating cash flows before interest, taxes, and depreciation less cash dividends and stock purchases to the firm's total assets, and label the result, *Freecf*.

¹⁷ We should note that it is possible to modify the quasi-likelihood model to model boundary observations separately. While not reported, we observe similar differences between the level and selection equation results, which reinforce the conclusion that models that ignore such potential differences are biased.

¹⁸ In addition, we test for the non-normality of the errors, assuming censoring, by following [Melenberg and van Soest \(1996\)](#) and using a Hausman type test to compare the coefficient estimates from the Tobit model with those from [Powell's \(1984\)](#) Censored Least Absolute Deviations estimators. A Chi-Square statistic of 865.19 with 4 degrees of freedom rejects the normality of the errors at the 1% level.

Table 4

Regression results for sample with ownership data

	OLS	ZIB	ZIB
Constant	0.178 (0.00)	−0.559 (0.01)	−0.936 (0.00)
Ceo's stk ownership	−0.097 (0.08)	0.044 (0.94)	
Managers' stk ownership			0.122 (0.73)
Logsale	0.010 (0.05)	0.061 (0.03)	0.053 (0.00)
Tang	0.179 (0.00)	0.649 (0.00)	0.871 (0.00)
Mtbr	−0.053 (0.00)	−0.901 (0.00)	−0.715 (0.00)
Freecf	0.00001 (0.25)	0.00004 (0.08)	0.0005 (0.06)
<i>Selection equation:</i>			
Constant		−0.833 (0.14)	−0.891 (0.00)
Ceo's stk ownership		−3.632 (0.00)	
Managers' stk ownership			−2.694 (0.00)
Logsale		0.301 (0.00)	0.339 (0.00)
Tang		3.417 (0.00)	3.722 (0.00)
Freecf		0.002 (0.13)	0.001 (0.15)
Adjusted R^2	0.32		
Wald's Chi-Square		833.21 (0.00)	1490.81 (0.00)

The sample of 1455 firms requires positive sales during 1992, complete data on Compustat, ownership data on Execucomp, and excludes regulated (SICs 41–49) or financial (SICs 60–67) firms. Following [Rajan and Zingales \(1995\)](#), the dependent variable (*DebtRatio*) in each of these regressions is the proportion of a firm's capital, long-term debt (book value) and equity (market value), accounted for by its long-term debt. *Tang* equals the ratio of fixed assets to the book value of total assets. *Mtbr* is the ratio of the book value of assets less the book value of equity plus the market value of equity all divided by the book value of assets. *Logsale* is the logarithm of net sales. *Freecf* equals the ratio of EBITDA less cash dividends and stock repurchases to the book value of assets. *CEO's stk ownership* represents the fraction of stock held by the company's CEO. *Managers' stk ownership* represents the fraction of stock held by managers of the company. We estimate the first regression as a linear regression using OLS and we estimate the last two regressions using the zero inflated beta regression model. We report *p*-values associated with a two-sided *t*-test of the difference between the estimated coefficients and 0 in the parentheses.

Using these variables in conjunction with *Logsale*, *Tang*, and *Mtbr* variables from our prior analysis, we fit an OLS linear regression model and the zero inflated beta regression model to these data and report the results in [Table 4](#). We exclude *Mtbr* from the hurdle equation as it was insignificant in the earlier regression.

The contrast between the results of applying the linear regression model and the ZIB model to these data is very interesting because it further illustrates the effects of not addressing our concerns. The two regression models disagree on the effect of CEO stock ownership on corporate capital structures and on the influence of a firm's free cash flows on its use of debt. Altogether this evidence reinforces our argument that it is necessary to recognize that the decision to issue long-term debt is not determined in the same fashion as the decision on how much long-term debt to use given that the firm has decided to use long-term debt.

Nevertheless, we check the robustness of this conclusion by estimating similar annual regressions for 1997 through 2004, and find similar results. In addition, we find, as did [Frank and Goyal \(2004\)](#), that using lagged explanatory variables leads to similar conclusions. In fact, using explanatory variables that are average over the prior two years lead to similar conclusions. Consequently, possible arguments that a firm's use of long-term debt and its managerial shareholding are jointly determined would not appear to undermine our results.¹⁹

4. Summary

Numerous papers in finance study the conditional mean of some proportion or fraction with a mass point at zero and commit one or more specification errors that raise doubts about their results. First, most, if not all of these studies model the proportion under study as a linear function of the explanatory variables. Second, most, if not all of these studies ignore the fact that the conditional variance must be a function of the conditional mean. Third, most, if not all of these studies fail to recognize that firms or individuals might make the choice to do or not do something for different reasons, i.e., they ignore selection bias.

We develop a mixed discrete-continuous distributional model to address these three specification concerns. Specifically, we develop a regression model using the beta and cumulative logistic distributions to model the continuous and discrete components of the mixture distribution. We call this model the zero-inflated beta model.

To illustrate the use and applicability of this new regression model, we fit the zero-inflated beta regression model along with the censored normal (Tobit) and quasi-likelihood model of [Papke and Wooldridge \(1996\)](#) to corporate capital structure data. We use these three statistical models because each addresses the bounded nature of the data. However, they differ in the degree to which they address the identified specification issues: the censored normal/linear regression model addresses none of the three, the quasi-likelihood model addresses the first two, and the zero-inflated beta model addresses all three.

Based upon our data, we conclude that the conditional expectation is a nonlinear function and the conditional variance is heteroskedastic: both conclusions are consistent with the evidence in [Cox \(1996\)](#), [Papke and Wooldridge \(1996\)](#), [Paolina \(2001\)](#), and [Kieschnick and McCullough \(2003\)](#). These specification errors alone cast suspicion upon prior regression results. While one

¹⁹ We should point out, however, that such arguments are rather dubious in light of the evidence reported in [Denis and Sarin \(1999\)](#).

might argue for different specifications of the conditional expectation function than we use, ours is consistent with the link tests in Cox (1996) and with the specification used in Papke and Wooldridge (1996), Paolina (2001), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004).

An important implication of these results is that prior studies that focus on incremental financing decisions in light of deviations from target capital structures (e.g., Hovakimian et al., 2001 or Flannery and Rangan, 2006) need to be re-examined as they likely mis-estimated the benchmark capital structure for a firm. Basically, such research has used a linear prediction equation when a nonlinear one is appropriate.

Further, and more central to our research focus, we find that the influence of different variables differ across boundary and non-boundary values, which is consistent with Cragg's (1971) hurdle model but inconsistent with a censoring model. Such evidence simply suggests that self selection is a potential issue when examining proportional data that possesses a cluster at zero. Reinforcing this point, our evidence appears to suggest that one observes different capital structure choices when managerial equity in their firms becomes substantial. Further, when their control appears contestable, different factors (e.g., free cash flows) influence how much debt is used.

In conclusion, we think that any study of the conditional expectation of a proportion or fraction in finance, particularly with a mass point at zero, needs to use either our proposed statistical model for these data, or any other alternative (e.g., a suitably modified version of Papke and Wooldridge's quasi-likelihood model) that addresses the three issues that we have raised.

References

- Agnew, J., 2005. Do behavioral biases vary across individuals? Evidence from individual level 401(k) data. *Journal of Financial and Quantitative Finance* 41, 939–962.
- Agrawal, A., Nagarajan, N., 1990. Corporate capital structure, agency costs, and ownership control: the case of all-equity firms. *Journal of Finance* 45, 1325–1331.
- Agrawal, A., Jayaraman, N., 1994. The dividend policies of all-equity firms: a direct test of the free cash flow theory. *Managerial and Decision Economics* 15, 139–148.
- Aitchison, J., 1955. On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* 50, 901–990.
- Arthur, N., 2001. Board composition as the outcome of an internal bargaining process: empirical evidence. *Journal of Corporate Finance* 7, 307–340.
- Aussenegg, W., Pichler, P., Stomper, A., 2006. IPO pricing with bookbuilding and a when-issued market. *Journal of Financial and Quantitative Finance* 41, 829–862.
- Baker, M., Wurgler, J., 2002. Market timing and capital structure. *Journal of Finance* 57, 1–32.
- Barclay, M., Smith Jr., C., 1995. The maturity structure of corporate debt. *Journal of Finance* 50, 609–632.
- Berens, J., Cuny, C., 1995. The capital structure puzzle revisited. *Review of Financial Studies* 8, 1185–1208.
- Blundell, R., Meghir, C., 1987. Bivariate alternatives to the Tobit model. *Journal of Econometrics* 34, 179–200.
- Bock, Hans-Hermann, 1996. Probability models and hypotheses testing in partitioning cluster analysis. In: Arabie, P., Hubert, L., DeSoete, G. (Eds.), *Clustering and Classification*. World Scientific Publishers, River Edge, NJ.
- Cox, C., 1996. Nonlinear quasi-likelihood models: applications to continuous proportions. *Computational Statistics & Data Analysis* 21, 449–461.
- Cragg, J., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39, 829–844.
- Demsetz, H., Villalonga, B., 2001. Ownership structure and corporate performance. *Journal of Corporate Finance* 7, 209–233.
- Denis, D.J., Sarin, A., 1999. Ownership and board structures in publicly traded corporations. *Journal of Financial Economics* 52, 187–223.
- Dudewicz, E., Mishra, S., 1988. *Modern Mathematical Statistics*. John Wiley & Sons, New York, N.Y.
- Fahrmeir, L., Tutz, G., 1996. *Multivariate Statistical Modelling Based on Generalized Linear Model*. Springer-Verlag, New York, N.Y.
- Fama, E., French, K., 2002. Testing tradeoff and pecking order predictions about dividends and debt. *Review of Financial Studies* 15, 1–33.
- Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modeling rates and proportions. *Journal of Applied Statistics* 31, 799–815.
- Flannery, M., Rangan, K., 2006. Partial adjustment toward target capital structures. *Journal of Financial Economics* 79, 469–506.
- Frank, M., Goyal, V., 2004. Capital structure decisions: which factors are reliably important? EFA 2004 Maastricht Meetings Paper No. 2464; Tuck Contemporary Corporate Finance Issues III Conference Paper.
- Gay, D., Welsch, R., 1988. Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. *Journal of the American Statistical Association* 83, 990–998.
- Guiso, L., Sapienza, P., Zingales, L., 2007. Trusting the stock market. NBER working paper.
- Heckman, J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–161.
- Houston, J., James, C., 1996. Bank information monopolies and the mix of private and public debt claims. *Journal of Finance* 51, 1863–1889.
- Hovakimian, A., Opler, T., Titman, S., 2001. The debt-equity choice. *Journal of Financial and Quantitative Analysis* 36, 1–24.
- Jensen, M., 1986. Agency costs of free cash flow, corporate finance and takeovers. *American Economic Review* 76, 323–329.
- Johnson, N., Kotz, S., Balakrishnan, N., 1995. *Continuous Univariate Distributions*, 2nd edition. John Wiley & Sons, New York.
- Johnson, S.A., 1997. An empirical analysis of the determinants of corporate debt ownership structure. *Journal of Financial and Quantitative Analysis* 32, 47–69.
- Ju, Parrino, Potesman, Weisbach, 2005. Horses and rabbits? Trade-off theory and optimal capital structures. *Journal of Financial and Quantitative Analysis* 40, 259–281.
- Kieschnick, R., McCullough, B.D., 2003. Regression Analysis of variates observed on (0,1): percentages, proportions, and fractions. *Statistical Modeling* 3, 193–213.
- Lewellen, K., 2006. Financing decisions when managers are risk averse. *Journal of Financial Economics* 82, 551–589.
- Li, K., Prabhala, J., 2005. Self-selection models in corporate finance. Chapter 2 In: Espen Eckbo, B. (Ed.), *Handbook of Corporate Finance: Empirical Corporate Finance* (Handbooks in Finance Series). Elsevier/North-Holland, New York.
- Lin, Tsai-Fen, Schmidt, P., 1984. A test of the Tobit specification against an alternative suggested by Cragg. *Review of Economics and Statistics* 1984, 174–177.
- Maddala, G.S., 1991. A perspective on the use of limited-dependent variables in accounting research. *Accounting Review* 66, 786–807.
- Manning, W., Duan, N., Rogers, W., 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35, 59–82.
- Melenberg, B., van Soest, A., 1996. Parametric and semi-parametric modelling of vacation expenditures. *Journal of Applied Econometrics* 11, 59–76.
- Paolina, P., 2001. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* 9, 325–346.
- Papke, L., Wooldridge, J., 1996. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* 11, 619–632.
- Peracchi, F., 2001. *Econometrics*. John Wiley & Sons, Inc., New York.
- Powell, J., 1984. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* 25, 303–325.
- Rajan, R., Zingales, L., 1995. What do we know about capital structure? Some evidence from international data. *Journal of Finance* 50, 1421–1460.
- Ruud, P., 1984. Tests of specification in econometrics. *Econometric Reviews* 3, 211–242.
- Spanos, A., 1999. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge University Press, NY.
- Welch, I., 2007. Common flaws in empirical capital structure research. NBER working paper.
- Wilks, S.S., 1962. *Mathematical Statistics*. John Wiley & Sons, Incorporated, NY.
- Yoo, S., 2004. A note on an approximation of the mobile communication expenditures distribution function using a mixture model. *Journal of Applied Statistics* 7, 747–752.