Statistical Modelling 2003; **3**: 193–213

Regression analysis of variates observed on (0, 1): percentages, proportions and fractions

Robert Kieschnick[1] **and BD McCullough**[2]
[1]Department of Finance and Managerial Economics, University of Texas at Dallas, Richardson, Texas, USA
[2]Department of Decision Sciences, Drexel University, Philadelphia, PA, USA

Abstract: Many types of studies examine the influence of selected variables on the conditional expectation of a proportion or vector of proportions, for example, market shares, rock composition, and so on. We identify four distributional categories into which such data can be put, and focus on regression models for the first category, for proportions observed on the open interval (0, 1). For these data, we identify different specifications used in prior research and compare these specifications using two common samples and specifications of the regressors. Based upon our analysis, we recommend that researchers use either a parametric regression model based upon the beta distribution or a quasi-likelihood regression model developed by Papke and Wooldridge (1997) for these data. Concerning the choice between these two regression models, we recommend that researchers use the parametric regression model unless their sample size is large enough to justify the asymptotic arguments underlying the quasi-likelihood approach.

Key words: fractions; percentages; proportions; regression

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received July 2002; **revised** October 2002; **accepted** March 2003

1 Introduction

Many studies in different disciplines examine how different variables influence some percentage, or proportion, or fraction or vector of such variates. For example, Webb (1983) studies the determinants of cable television subscribership as measured by the proportion of houses that are passed by a cable system that subscribe to that system. As another example, DeSarbo *et al.* (1993) study the proportions of household television viewing across different streams of programming.

We surveyed many such studies and arrived at two broad conclusions. Our first broad conclusion is that the data being analysed can be put into one of four distributional categories. The first distributional category comprises proportions on the open interval (0, 1). Figure 1 illustrates such data using cable penetration data (the proportion of homes that have cable in each of 278 different market areas). The second distributional category comprises proportions observed on the closed interval [0, 1].

Address for correspondence: R Kieschnick, University of Texas at Dallas, PO Box 830688, JO5.1 Richardson, Texas 75083-0688, USA. E-mail: rkiesch@utdallas.edu

10.1191/1471082X03st053oa

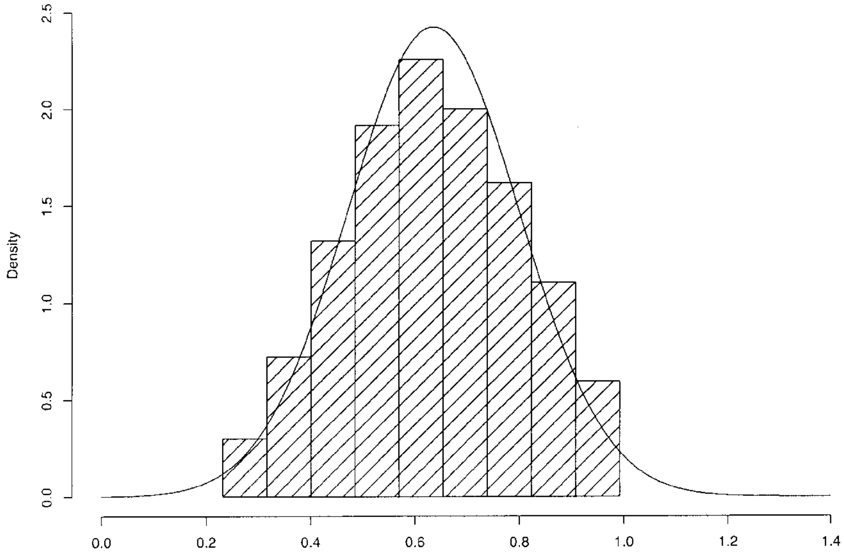

Figure 1 Distribution of cable penetration with superimposed normal distribution

Figure 2 illustrates such data using US corporate capital structure data. A comparison of Figures 1 and 2 reveals a critical difference between these two types of data: observations at 0 or 1 are typically mass points. (Most software procedures for producting histograms do not automatically capture point masses. The leftmost bin

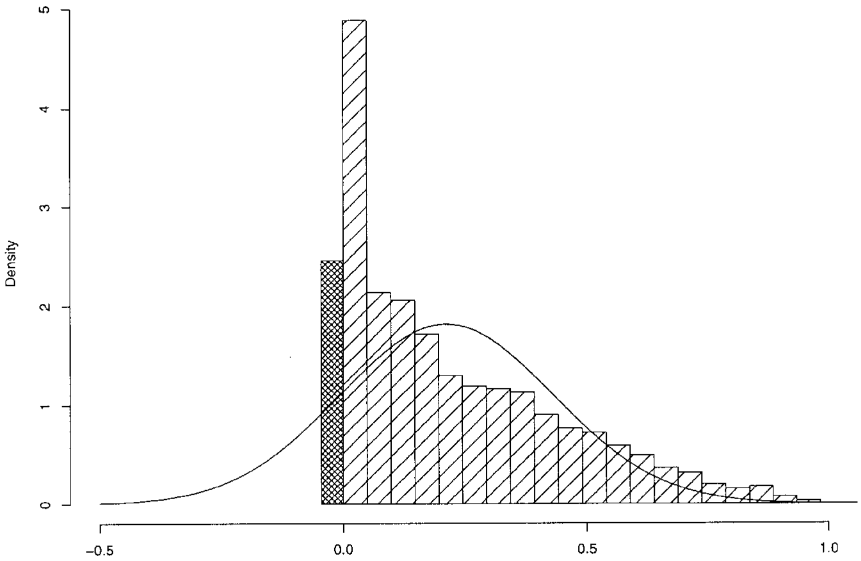

Figure 2 Distribution of corporate debt loads with superimposed normal distribution (crosshatched bin represents the mass point at zero)

in this histogram represents 454 observations with the value zero. The number of bins for all histograms was determined using the method of Scott (1979).) Consequently, while the data in the first distributional category can be modeled using a continuous distribution, the data in the second distributional category cannot, because they follow a mixed discrete–continuous distribution.

Our third and fourth distributional categories are multivariate extensions of the first two distributional categories. For example, one might be interested in the proportion of a household's wealth invested in different classes of assets. If the investment classes are broad enough, then typically each household has some of its wealth invested in each of the categories and there are no boundary observations. Thus, our third distributional category comprises vectors of proportions for which no component proportion is a boundary observation (e.g., 0s or 1s). However, if the investment classes are more narrowly defined, then a number of households will not have some of their wealth in some of these classes. Consequently, our fourth distributional category comprises vectors of proportions in which some component proportions are boundary observations. Each of these distributional categories presents additional problems that do not occur in their univariate analogs, such as the structure of the correlations between elements of these vectors.

The second broad conclusion that we derive from our survey of the published literature is that there are no commonly accepted distributional models for these data, nor any commonly accepted regression models for these data. Further, some of the regression models for these data appear dubious at best. For example, we find that researchers most frequently estimate the parameters of a linear regression model for the first two distributional categories using ordinary least squares (OLSs). However, such an approach contravenes two conditions: the conditional expectation function must be nonlinear since it maps onto a bounded interval; and its variance must be heteroskedastic since the variance will approach zero as the mean approaches either boundary point.

The diversity of practices and the questionable nature of some of these practices motivate this study. We, however, only focus on our first distributional category, proportions observed on (0, 1), because the issues presented by our other three categories are quite different, and yet might build in some degree upon how one addresses this first category. For this category, we surveyed the literature to identify the various specifications employed in prior research and synthesized the various practices into a number of groups. We present the essentials of these groups in section 2.

In section 3, we use a common data set, cable penetration data, and a common set of regressors to compare the different regression models. We do this, rather than perform a Monte Carlo based comparison, because a Monte Carlo study would have required us to assume some data generating process. However, there is no clear agreement on the data generating process for such data. Thus, we fit the different regression models to a common data set using a common specification of the regressors to determine which regression model best describes the data.

In section 4, we repeat the analysis reported in section 3, but this time using a different data set, voting for President Bush in the 2000 Presidential election. We do this to check the robustness of our conclusions about which regression model best describes this distributional category of proportional data.

Section 5 concludes our paper by providing a summary of our findings and directions for further research. Our basic recommendation is that researchers use either the beta regression model developed herein or a quasi-likelihood model developed in Papke and Wooldridge (1996). Our case studies suggest that the larger the sample, the closer the results of these models will be, but that in small samples the beta regression model is preferred.

2 Alternative regression models

We searched the literature to identify how different researchers conducted regression analyses of proportions observed on $(0, 1)$, and what their regression models implied about their maintained assumptions. (Specifically, we searched various electronic databases for the keywords percentage, proportion and fraction, to identify potentially relevant research.)

While there are a variety of ways that we could organize the different approaches we observed in published papers, we use the likelihood principle to organize our discussion. From the likelihood principle viewpoint, such research assumes that $h(\mathbf{x}, y)$ implies $f(y|\mathbf{x})g(\mathbf{x})$ and $E(y|\mathbf{x}) = k(\mathbf{x})$, where $k(\cdot)$ is a function of a vector of exogenous variables and $h(\mathbf{x}, y)$ is the joint distribution of $\mathbf{x}$ and y. (Sometimes $k(\mathbf{x})$ is called the response function, and sometimes it is called the conditional expectation function.) Thus we will distinguish approaches according to what the researcher assumed about $k(\mathbf{x})$ and $f(y|\mathbf{x})$. Further, we will distinguish between parametric and quasi-parametric specifications of $f(y|\mathbf{x})$.

2.1 Parametric regression models

We organize our discussion of the identified parametric regression models in order of their frequency of use. In this regard, we should note that the first three regression models account for most of what we observe being used in published papers. The last three regression models are rarely used. Because the normal distribution figures prominently in the commonly used models, all our histograms have superimposed a normal distribution parameterized by the sample mean and sample variance.

2.1.1 Normal distribution: linear response function

By far the most common practice of researchers is to apply OLS to their data (Mehran, 1995). The use of OLS presents a problem in characterizing these studies, as sometimes the sample sizes were large enough to invoke asymptotic arguments to rationalize less stringent characterizations of their regression models. Nevertheless, we focus on their most stringent characterization, for as Godfrey (1988) points out, when these researchers examine t tests or F tests, they are implicitly assuming that the conditional distribution is a normal distribution unless the sample size is large. Further, some commonly reported tests (e.g., Breusch–Pagan's test for heteroskedasticity) assume that the conditional distribution is a normal distribution regardless of sample size.

Consequently, we categorize all such studies as implicitly assuming a conditional normal distribution for their regression model (i.e., $f(y|\mathbf{x})$ is $\mathrm{N}(k(\mathbf{x}), \sigma^2)$). In addition to assuming a conditional normal distribution, these researchers also assume that $k(\mathbf{x})$ is equal to $\mathbf{x}'\beta$, that is, that the conditional expectation function is linear.

Conceptually this approach is subject to a number of flaws. First, it is obvious that proportions are not normally distributed because they are not defined over $\Re$, which is the domain over which the normal distribution is defined. A quick examination of any of our figures confirms this point. Further, as mentioned earlier, the fact that these variables are only observed over a closed interval implies that the conditional expectation function must be nonlinear, and that the conditional variance must be a function of the mean. Clearly, both of these conditions are violated by the assumptions of this regression model.

2.1.2 Additive logistic normal distribution

The next most frequent practice we observed is that the researcher would transform the dependent variable and then fit a linear response function to the transformed dependent variable using the least squares principle (Demsetz and Lehn, 1985). While it is not always clear what the researcher is assuming about the conditional distribution of the untransformed variable, it was possible to infer their assumptions in all the studies we examined, because they all used the logit transformation. (Atkinson (1985) devotes a chapter to discussing various transformations for percentages and proportions. While we cannot report finding any studies that follow Atkinson's proposed approach for implementing these transformations, we do note that most of the transformations he studied have the logit transformation as a limiting case.)

The logit regression model has a long history in economics and related disciplines (see Dyke and Patterson (1952) for its earliest development). Using this approach, researchers (e.g., Webb, 1983) will estimate:

$$\ln\left(\frac{y}{1-y}\right) = \mathbf{x}'\beta + \epsilon \tag{2.1}$$

where $\ln(y/(1-y))$ is the logit transform of the dependent variable using the least squares principle. Thus, as noted earlier, these researchers are assuming that ϵ is distributed $N(0, \sigma)$.

Aitchison (1986) calls the above transformation the additive logratio transformation, and shows that $z = \ln(y/(1-y))$ will follow a normal distribution, $N(\mu, \sigma^2)$, if y follows an additive logistic normal distribution. Thus, if y is distributed according to an additive logistic normal distribution, then ϵ is distributed according to the standard normal distribution. (Aitchison (1986) proposes that one should test if y is distributed as an additive logistic normal distribution by testing if z is normally distributed. We follow his proposed testing strategy in our subsequent analysis by testing if ϵ is distributed as a standard normal random variate.)

The obvious concern with the application of this regression model to the analysis of proportions is that it assumes, first, that the link function (using generalized linear model terminology) is the logit function, and second, that this transformation stabilizes the conditional variance. The first concern is lessened by evidence reported in Cox (1996) on different link functions for these data. The second concern, however, remains important since alternative distributional models for these data (e.g., the beta and simplex distributions) imply that such a transformation will not stabilize the variance.

2.1.3 Censored normal distribution

Recently, researchers have begun to apply the censored normal model, or Tobit model, to proportional data (e.g., Barclay and Smith, 1995). Specifically, such researchers typically assume that:

$$y_i^* = x_i'\beta + u_i, \qquad i = 1, 2, \ldots, n \tag{2.2}$$

and

$$y_i = \begin{cases} 0, & y_i^* \leq 0 \\ y_i^*, & 0 < y_i^* < 1 \\ 1, & y_i^* \geq 1 \end{cases} \tag{2.3}$$

where $\{u_i\}$ are assumed to be i.i.d. draws from a $N(0, \sigma^2)$ distribution.

There are problems with the use of this approach to examining the conditional expectation of a proportion observed over the interval (0, 1). First, one is making the assumption that y_i^* is normally distributed, but only observes values within a specified range. We fail to observe values outside the [0, 1] range for proportional data not because they are censored, but because they are not defined outside this interval. Thus, there is no censoring, and the censored normal model is inappropriate for these data (see Maddala (1991) for further discussion of this point). Second, for the data observed on the interval (0, 1), the Tobit regression is observationally equivalent to the normal regression model. Thus the Tobit regression model is subject to the same criticisms as the linear normal regression model.

2.1.4 Normal distribution: nonlinear response function

Another alternative is to fit a nonlinear regression model to these data using least squares. For example, Hermalin and Wallace (1994) use the cumulative normal function as their conditional expectation function and estimate its parameters using least squares.

Rather than follow their practice, we will model the conditional expectation function in our study using the cumulative logistic function because of the evidence reported in Cox (1996) and because it will allow us to focus more on the effect of distributional assumptions. Further, this specification is consistent with the specification recommended in Kmenta (1986) for these data. Specifically, we assume:

$$y_i = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} + \epsilon_i \tag{2.4}$$

where ϵ_i follows $N(0, \sigma^2)$ distribution. We will estimate this equation using the least squares principle, as that comports with prior practice.

2.1.5 Beta distribution

Johnson *et al.* (1995) provide over a dozen examples from different physical sciences in which the beta distribution was found to be a better fitting distribution for the proportional data under study than considered alternatives. Hviid and Villadsen (1995) provide a recent example from economics. Even textbooks have suggested that the beta distribution is a good distributional model for proportional data. For example, Mittelhammer (1996, pp. 195–97) uses an example of the proportion of a tank containing heating oil measured at different times to illustrate the beta distribution.

All of these examples assume that y is distributed as follows:

$$f(y) = \frac{1}{B(p, q)} y^{p-1}(1 - y)^{q-1} \tag{2.5}$$

where $0 \leq y \leq 1$ and $B(p, q)$ is the beta function. We will focus on this two-parameter beta distribution, rather than the generalized beta distribution developed in McDonald and Xu (1995), for two reasons. First, it is the distribution most often fitted to proportional data in prior literature and so it is the distributional assumption that has the most empirical support. Second, this family of distributions is part of the class of exponential distributions, which have served as the basis for the generalized linear model paradigm. (see Mittelhammer (1996, pp. 213–15) for a discussion of the exponential class of distributions, and see McCullagh and Nelder (1989) for a discussion of the generalized linear model paradigm.)

One approach to specifying a beta regression model is to follow the linear regression model and assume that the mean is a linear function of the exogeneous variables. This is the approach taken in the econometrics program SHAZAM version 7 (Vancouver, Canada) and is also the approach taken in McDonald and Xu (1995). Specifically, SHAZAM's user manual sets out the following model. They assume that y is distributed as a beta random variate and that

$$E(y|\mathbf{x}) = \frac{p}{p + q} = \mathbf{x}'\beta \tag{2.6}$$

Further they assume that q is the parameter that is conditional on $\mathbf{x}$ and so they derive:

$$q(\mathbf{x}) = \frac{p}{\mathbf{x}'\beta} - p \tag{2.7}$$

They substitute (2.7) into (2.5) above to derive the conditional density function (i.e., $f(y|\mathbf{x})$). They then derive the log-likelihood function for the beta regression model using this specification.

This approach does not, however, restrict the range of the conditional mean. Thus it implicitly requires restrictions on the values of the exogeneous variables to give sensible results. Such restrictions are not recognized in SHAZAM's estimation of this model, and so the beta regression model specified in SHAZAM is not a very good approach to specifying a beta regression model.

We argue that a better approach is derived by considering the work of Cox (1996), who tested various link functions for regression models of continous proportions using the quasi-likelihood framework. Based upon Cox's evidence, we use the logit link specification. (Separate from Cox's evidence, we are also motivated to use the logit link specification so that there is a consistency in our first moment specification across different parametric and quasi-parametric models.) Specifically, we assume that

$$E(y_i|\mathbf{x}_\mathrm{i}) = \mu_i = h(\eta_i) = \frac{1}{1+\exp(-\eta_i)} = \frac{1}{1+\exp(-\mathbf{x}_\mathrm{i}'\beta)} \tag{2.8}$$

This equation then implies the following re-expression:

$$\eta_i = g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_\mathrm{i}'\beta \tag{2.9}$$

Note that this specification restricts the conditional mean of a beta distributed regressand to the interval (0, 1), which is appropriate for this distributional model.

In order to derive an estimatable regression model, we must now relate the above relationships to the parameters of the beta distribution. Specifically, for the beta distribution defined in (2.5), we have

$$E(y_i) = \frac{p}{p+q} \tag{2.10}$$

We map $\mathbf{x}_\mathrm{i}'\beta$ into q because q is the shape parameter for the beta distribution. Given this approach, we develop the following expression for q that is consistent with equation (2.10) above:

$$q(\mathbf{x}_i) = p\,\exp(-\mathbf{x}_i'\beta) \tag{2.11}$$

We then substitute this expression for q into (2.5) to derive the conditional distribution of the beta distributed random variate:

$$\begin{aligned} f(y_i|\mathbf{x}_\mathrm{i}) &= \left[\frac{\Gamma(p)\Gamma(q(\mathbf{x}_i))}{\Gamma(p+q(\mathbf{x}_i))}\right]^{-1} y_i^{p-1}(1-y_i)^{q(\mathbf{x}_\mathrm{i})-1} \\ &= \left[\frac{\Gamma(p+q(\mathbf{x}_i))}{\Gamma(p)\Gamma(q(\mathbf{x}_i))}\right] y_i^{p-1}(1-y_i)^{q(\mathbf{x}_\mathrm{i})-1} \end{aligned} \tag{2.12}$$

To estimate the effect of the different conditioning variables $(x_1, x_2, \ldots, x_r)$ we can use the maximum likelihood estimation principle to derive estimates of the vector β by maximizing the implied log-likelihood function with respect to the parameters β and p.

Because the beta distribution is a member of the exponential class of distributions, these maximum likelihood estimators have all the statistical properties established for maximum likelihood estimators for this class of distributions (see Jorgensen (1983) or Fahrmeir and Tutz (1994) for further discussion of these issues and the literature addressing them).

2.1.6 Simplex distribution

Our final parametric regression model for proportions measured on (0, 1) is based upon the simplex distribution developed by Barndorff-Nielsen and Jorgensen (1991) for such data. Jorgensen (1997) argues that this distribution has the virtue of being a dispersion model, while the beta distribution is not. Consequently the analysis of deviance approach to generalized linear models can be applied to regression models based upon the simplex distribution, and not to regression models based upon the beta distribution.

We follow Jorgensen (1997) and define the univariate simplex distribution as:

$$f(y;\,\mu,\,\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2}\exp\left\{-\tfrac{1}{2}d(y;\,\mu)\right\} \tag{2.13}$$

for $0 < y < 1$, where

$$d(y;\,\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2} \tag{2.14}$$

is the unit deviance, and $0 < \mu < 1$. We can either maximize the associated log-likelihood function or minimize the associated unit deviance function to estimate a regression model based upon this distribution. Again, following the evidence of Cox (1996) and our specification of the beta regression model, we assume that the link function, $g(\mu)$, is the logit link function.

2.2 Quasi-parametric regression models

The prior approaches have presumed some specific family of distributions for the conditional distribution for the proportions under study. Cox (1996) and Papke and Wooldridge (1996) take a different tack and use the quasi-likelihood approach. The quasi-likelihood approach specifies the first and second moments of the conditional distribution, but does not specify the full distribution.

Cox (1996) examines the fit of four specifications of the first two moments of the conditional distribution to two samples of proportional data observed over the interval (0, 1). Specifically, he examines the use of the logit and complementary log-log link functions, with canonical and orthogonal specifications for the variance functions. He concludes that the logit link function with the orthogonal variance function is the preferred combination for his data sets. (Cox describes his use of the terms canonical and orthogonal specifications and specifies on p. 455 his examined combinations.

His preferred combination, which he calls the orthogonal pair, is: $\mu(\theta) = 1/(1 + e^{-\theta})$; $\upsilon(\mu) = \mu^2(1 - \mu)^2$.)

Papke and Wooldridge (1996) use a similar approach, but for a slightly different problem. They are interested in specifying a quasi-likelihood regression model for continuously measured proportions with a finite number of boundary observations (i.e., 0s and 1s). They use the following log-likelihood specification,

$$\ell_i(\beta) = y_i \ln[G(\mathbf{x_i})] + (1 - y_i)\ln[1 - G(\mathbf{x_i})] \tag{2.15}$$

which they point out is well defined for $0 < G(\cdot) < 1$. While Papke and Wooldridge discuss the use of different specifications for $G(\cdot)$, they use the logistic function in their analysis, which is equivalent to Cox's logit link specification. Further, they use a slightly more robust approach to the estimation of the standard errors. Consequently we will focus on their approach in the subsequent analysis.

2.3 Comparing regression models

We have presented the essentials of various regression approaches that have been applied to the analysis of proportions observed on the interval (0, 1). While we have given reasons to question the application of one or another of these approaches to this kind of data, we now turn to a comparison of their application to real data to judge their relative merits. We do this because, as stated earlier, there is no agreement in the literature on the proper probability models for these data.

3 Case study 1: cable penetration

3.1 Description of the data

We first examine influences on cable penetration. To do this we will use a sample that was collected by the Federal Communications Commission (FCC) in conjunction with the implementation of the Cable Television Consumer Protection and Competition Act of 1992. On 23 December 1992 the FCC mailed out 748 questionaires to a stratified sample of 'cable community units,' which are essentially individual franchise areas. Of the 748 questionaires sent out, 687 respondents supplied usable responses on prices, costs and other cable operator data for each franchise. Because the sampling methodology and data elements requested in the questionaire are described in detail in Appendix E of FCC 93-177, we will refer the reader to this source for further discussion of these topics (Federal Communications Commission, 1993).

To supplement these data, the FCC collected additional economic and demographic information from the US Department of Commerce, Census Bureau. Appendix C of FCC 94-38 describes the data, matching procedures (i.e., county to cable community unit matching), and variable definitions used to extend the original sample data for additional economic and demographic information (Federal Communications

Commission, 1994). Again, we refer the reader to this source for detailed descriptions of the added data.

For the above sample, we will only use a portion of the data collected as we are more interested in comparing the results of different regression procedures than in exploring the best specification of regressors. Consequently we use the same set of regressors in each estimated regression model to focus on the effect of the statistical model on the results obtained. Specifically, we use:

$$E(y|\mathbf{x}) = h(\beta_0 + \beta_1 lin + \beta_2 child + \beta_3 ltv + \beta_4 dism + \beta_5 agehe) \tag{3.1}$$

In this equation, we are using the logarithm of franchise median income (*lin*), the percentage of franchise households with children (*child*), the number of local broadcast television signals (*ltv*), the age of the cable system headend (*agehe*), and a measure of consumer dissatisfaction with the cable operator (*dism*). The dependent variable, displayed in Figure 1, is the proportion of households within a market area that subscribe to cable television.

We include the logarithm of franchise median income because prior studies (e.g., Park, 1972; Reagan *et al.*, 1985, etc.) have shown household income to influence a household's decision to subscribe to cable television. We adjust for scale effects by using the logarithm of median franchise income. We include the percentage of franchise households with children because prior studies (e.g., Park, 1972; Reagan *et al.*, 1985) have shown that the presence of children in a household is a significant influence on a household's decision to subscribe to cable television. We include the number of local broadcast television signals because prior studies (e.g., Park, 1972) have found the number of local broadcast television signals available to a household to be a significantly negative influence on whether a household subscribes to cable television. We include the age of the cable system headend because this proxies for extent of diffusion of the cable system within its franchise (see Sparkes and Kang (1986) for further discussion of why this is important). Finally we create a consumer satisfaction variable, *dism*, by computing the number of net disconnectors (number of disconnecting households − number of reconnecting households) and substracting this number from the number of new subscribing households. We presume that if more households are leaving a cable system than entering it, then consumers are dissatisfied with the service provided by the cable system operator. LaRose and Atkin (1988), Atkin (1992), and Albarran and Umphrey (1994) provide evidence that such a variable significantly influences cable penetration.

3.2 Comparison of regression models

In Table 1, we report the results of fitting the various regression models described above to the data, also described above. Before discussing these results, we should address a couple of computational issues since a number of these regression models require nonlinear optimization to derive parameter estimates. First, starting values for the coefficients of the independent variables were obtained from the logit or logistic normal regression model as its conditional expectation is consistent with the functional form of their conditional expectation function. (We estimated the simplex regression model by

Table 1 Results for different regression models of cable penetration

$k(\mathbf{x})f(y\|\mathbf{x})$	Linear normal	Linear censored normal	Transformed logistic normal	Logistic normal	Logistic beta	Logistic simplex	Logistic unspecified
Estimation method[a]	LS	ML	LS	LS	ML	ML	QML
constant[b]	−0.1453	−0.1453	−3.0696*	−2.7391**	−1.9221*	−2.5985*	−2.8547**
	(0.2815)	(0.2784)	(1.5643)	(1.2177)	(1.1283)	(1.4054)	(1.2911)
lin	0.0763**	0.0763**	0.3679**	0.3215**	0.2424**	0.3497**	0.3340**
	(0.0288)	(0.0285)	(0.1600)	(0.1248)	(0.1113)	(0.1484)	(0.1318)
child	0.0007	0.0007	0.0027	0.0029	0.0031	−0.0017	0.0029
	(0.0010)	(0.0097)	(0.0054)	(0.0042)	(0.0046)	(0.0065)	(0.0041)
ltv	−0.0147**	−0.0147**	−0.077**	−0.0622**	−0.0592**	−0.0795**	−0.0642**
	(0.0033)	(0.0033)	(0.0184)	(0.0146)	(0.0145)	(0.0174)	(0.0136)
dism[c]	−0.4082	−0.4082	−1.7529	−1.6576	−1.9455**	−2.1407**	−1.6175**
	(0.2986)	(0.2986)	(1.6589)	(1.2450)	(0.4934)	(0.7250)	(0.5558)
agehe	0.0062**	0.0062**	0.0338**	0.0282**	0.0258**	0.0261**	0.0283**
	(0.0012)	(0.0012)	(0.0067)	(0.0048)	(0.0059)	(0.0071)	(0.0056)
AIC_c[d]	−1.4525	−1.4525	−2.4835	−2.6334	−2.7395	−2.6334	−2.7377

[a] LS, least squares; ML, maximum likelihood; QML, quasi-maximum likelihood.

[b] Standard errors are reported in parentheses below coefficient estimate. ** Significance at the 5% level; * significance at the 10% level.

[c] *dism* was scaled by 1/100 000 in order to produce coefficents of the same magnitude as the other coefficients.

[d] This statistic represents a version of the Akaike Information Criteria derived in McQuarrie and Tsai (1998).

both maximizing its associated log-likelihood function and by minimizing its unit deviance function. Each approach gave the same results, but the second approach converged to a solution in $\frac{1}{6}$ the number of iterations. Consequently we recommend the use of the second approach for estimating these types of regression models.) In addition, the beta regression model requires a starting value for p, which we derived by fitting the two-parameter beta distribution to our sample data using both maximum likelihood and method of moment estimators. (For this we use a STATA program developed by Nicholas Cox to implement procedures developed in Mielke (1975). Note, however, that the coefficient estimates of the beta regression model were very robust with respect to variations of our starting parameter estimates. Consequently it is clear that these starting values were not important to our final estimates.) Second, all functions were optimized and standard deviations estimated using first and second analytic derivatives where appropriate (derivatives were confirmed using Mathematica (version 4)). Third, we used both STATA (version 7) and TSP (version 4.5) on a Pentium 400 PC running under Windows 2000 to perform our computations in order to confirm that our results were robust across software packages. (In an earlier draft, we used S-PLUS v4.5 release 2 on a Pentium 400 PC running Windows 98 and obtained similar estimates to those reported.) Given these computational notes, we begin our discussion of our results.

Additionally, since the conditional expectation function in the first two regression models differs from the conditional expectation function in the last five regression models, we cannot simply compare coefficient estimates across regressions to gain a sense of the estimates of the marginal effect of a change in a regressor on the regressand. The last five regression models assume: $k(\mathbf{x}) = 1/(1 + \exp(-(x'\hat{\beta})))$. Thus,

$$\frac{\partial k(\mathbf{x})}{\partial x_i} = \frac{\hat{\beta}_i \exp(-(x'\hat{\beta}))}{[1 + \exp(-(x'\hat{\beta}))]^2}$$

To facilitate comparison, we report in Table 2 the partial derivatives of the conditional expectation function of the nonlinear regression models with respect to the different regressors, evaluated at the sample means of the regressors.

Returning to Table 1, an inspection reveals that one does derive different inferences regarding the statistical significance of different regressors depending upon one's choice of regression model. Specifically, we see that the linear models (columns 2 and 3) reject the significance of the *dism* variable. The transformed/logistic normal (column 4) and the logistic/normal (column 5) regression models also reject the significance of the *dism*

Table 2 Estimates of the marginal effects of regressors

$k(\mathbf{x})$ $f(y\|\mathbf{x})$	Linear normal	Linear censored normal	Transformed logistic normal	Logistic normal	Logistic beta	Logistic simplex	Logistic unspecified
lin	0.0763	0.0763	0.0822	0.0740	0.0556	0.0775	0.0769
child	0.0007	0.0007	0.0006	0.0006	0.0007	−0.0003	0.0006
ltv	−0.0147	−0.0147	−0.0172	−0.0143	−0.0135	−0.0176	−0.0147
dism	-4.0×10^{-6}	-4.0×10^{-6}	-3.9×10^{-6}	-2.7×10^{-7}	-4.3×10^{-7}	-4.6×10^{-7}	-3.6×10^{-7}
agehe	0.0062	0.0062	0.0075	0.0064	0.0059	0.0057	0.0065

variable. In contrast, the beta, simplex and quasi-likelihood models (columns 5, 6 and 7) do not.

In addition to demonstrating that one's choice of regression model for these data influences one's inferences, these results illustrate two other important points. First, the fact that the beta, simplex and quasi-likelihood model pick up the significance of the *dism* variable while the other regression models do not, comports with our criticisms of the homoskedasticity assumption of these models. Consistent with this conjecture, we note that for the transformed/logistic normal (column 4) and the logistic/normal (column 5) regression models, the *dism* variable becomes statistically significant when we correct the standard errors for heteroskedasticity. Second, the results of the beta and quasi-likelihood models comport with discrete choice studies of the cable subscription decision. One of the reasons for examining cable penetration data is that we could relate our results using aggregate data to results using disaggregate data to discern which models for proportional data gave similar results to those of individual choice data. Several probit analyses have shown that consumer satisfaction with a cable service (our *dism* variable) is an important determinant of the cable subscription decision. Consequently we conclude that the beta, simplex and quasi-likelihood models provide similar conclusions to those derived in discrete choice studies.

Inspection of Table 2 suggests that there is not much difference across regression models in their estimates of the marginal effects of the various regressors when evaluated at the same means of the regressors. This result is not surprising in that the linear models should be first-order approximations to a nonlinear surface when evaluated at the sample means. Despite this point, we note that the nonlinear models tend to ascribe a much lower marginal effect to the *dism* variable than do the linear models. Further we must point out the linear and nonlinear models will give quite different estimates of the marginal effects of different regressors as those regressors are evaluated at points other than their means. This fact should not be overlooked when evaluating the economic significance or policy implications of one's regression results.

Turning to an analysis of the distributional assumptions of the different regression models, we find evidence to reject the normal, the censored normal and the logistic normal distributional assumptions. We find that either a Shapiro–Wilk or a Jarque–Bera test rejects the normality of the residuals of the first regression model at the 1% marginal significance level. Further, we tested for a nonlinear expectation function using by fitting a Box–Cox model to these data and testing whether the null hypothesis that $\delta = 1$ holds. (We use this testing procedure because it imposes fewer assumptions than alternatives (e.g., testing the significance of quadratic terms).) Consistent with our previous conclusion, we find that the likelihood ratio chi-square of 72.47 rejects the null hypothesis at the 1% marginal significance level. Hence, the data not only suggest that the residuals are not normally distributed, but also that the conditional expectation function is nonlinear. (We should note that this result is also consistent with Cox's (1996) evidence on the appropriateness of the logit link for the two data sets he examined.) Because the censored normal model provides estimates and residuals similar to those of the linear normal regression model, these criticisms also apply to it. (We should note that these criticisms also apply to the truncated normal distribution since it generates the same estimates and residuals as the normal regression models do for these data.)

The rejection of the conditional distribution as a normal distribution is not simply a consequence of the nonlinearity of the conditional expectation function. We tested the residuals of the nonlinear normal regression model. While a Shapiro–Wilk W test fails to reject the normality of the residuals at the 10% level, a separate test of skewness and kurtosis of the residuals of this model reject the hypothesis that the kurtosis is consistent with a normal distribution at the 5% level. These last results makes sense in that Godfrey (1988) points out that the least squares residuals will tend to fail to reject the normality of the residuals even when the generating model was not a normal distribution. Further, our earlier argument that the variance of bounded variables should tend to zero as the mean approaches the boundary points suggests that we might observe the distribution of the residuals of a nonlinear least squares regression model to show more kurtosis. Supplementing this evidence, we regress the squared residuals on the predicted values and find the coefficient (−0.0331) on $\hat{y}$ to be significant at the 1% marginal significance level. This result suggests that the variance is a function of the mean, which is consistent with our earlier discussion about bounded variables. (We should note that this conclusion is consistent with Cox's (1996) characterization of the variance of the two data sets he examined.)

While not based upon the normal distribution, the regression model based upon the additive logistic normal distribution does suggest that its residuals should be normally distributed (as noted earlier, this point comes from Aitchison (1986)). This implication is not supported by the data as both the Shapiro–Wilk and Jarque–Bera tests reject the normality of its residuals at the 1% marginal significance level. Consequently, our earlier arguments for questioning the distributional assumptions underlying the logit regression model are borne out by our data.

While we reject the normal, censored normal, and logistic normal distribution models for our data, we cannot reject the distributional models underlying the beta, simplex or quasi-likelihood models. In the case of the quasi-likelihood model, this is because we have simply assumed the functional form of the conditional expectation function, which the data do not reject. For the beta and simplex distributions, *q–q* plots do not suggest that the data are inconsistent with either model.

Consequently, we compare the different regression models using Akaike's Information Criteria (AIC), as it is the most widely used and accepted model selection criteria. However, there are different variations of the AIC statistic depending upon the class of regression models being evaluated and the sample size. We use a variation of the AIC statistic named the AIC$_c$ statistic, as McQuarrie and Tsai (1998) show that it is the preferred Kullback–Leibler information-based model selection criterion for non-normal and quasi-likelihood regression models. The AIC$_c$ is defined as:

$$\log(\hat{\sigma}) + \frac{n+k}{n-k-2}$$

where $\log(\hat{\sigma})$ represents the natural logarithm of a regression's mean square error, n represents the number of observations, and k represents the number of parameters to be estimated. Note that this measure adjusts for sample size and so is appropriate for small as well as large samples.

An examination of Table 1 reveals that the beta regression model dominates the other regression in terms of the AIC$_c$ statistic. (The lower the value of the AIC$_c$ statistic the better the model fit to the data. Since the number of observations and regressors are fixed across regressions in our study, this statistic will choose the regression with the lowest mean square error.) However, the differences between the beta regression model and the quasi-likelihood model are so small as to suggest that these two models fit the data equivalently well. As we show in the next case study, this last inference is driven by sample size, and so consistent with quasi-likelihood methods being asymptotic approximations.

4 Case study 2: Presidential voting

4.1 Description of the data

As a second case study we examine factors that influenced the proportion of a state's votes for President George Bush in the 2000 Presidential Election. We obtained the voting data from the web site www.uselectionatlas.org, which provides data on past presidential elections. We then obtained demographic, employment and income data for each state for the year 2000 from the US Bureau of Census' web site, factfinder.census.gov. However, one can also obtain these data from the US Department of Agriculture's Economic Research Service web site, www.ers.usda.gov.

Using these data, we created the following variables. Our dependent variable is the fraction of a state's total counted vote that was for President George Bush. The histogram of these data, displayed in Figure 3, clearly suggests that the data do not follow a normal distribution. Our independent variables, or regressors, are the natural logarithm of a state's population (*lnpop*), a state's unemployment rate (*clfu*), the proportion of a state's population that a male (*male*), the proportion of a state's male population that as older than 18 years of age (*mgt18*), the proportion of a state's population that was older than 65 years (*pgt65*), the proportion of a state's population that lived in 'urban' areas (*urban*), the proportion of a state's households that had annual incomes greater than \$75 000 (*gt75k*), and finally the proportion of a state's households whose annual income was below the 'poverty' level (*bpovl*). Using these variables, we estimate the following common specification:

$$E(y|\mathbf{x}) = h(\beta_0 + \beta_1 lnpop + \beta_2 clfu + \beta_3 male + \beta_4 mgt18 + \beta_5 pgt65 \\ + \beta_6 urban + \beta_7 gt75k + \beta_8 bpovl)$$

We should note that we did not test to see whether this specification of variables was the best specification or whether there might have been a better selection of potential regressors. Rather, we simply chose a set of variables that seemed reasonable regressors and used them. This approach is adequate for our purposes since we are only interested in comparing the different regression models using a common data set and specification of regressors. Nevertheless, we should note that the predictions from our nonlinear

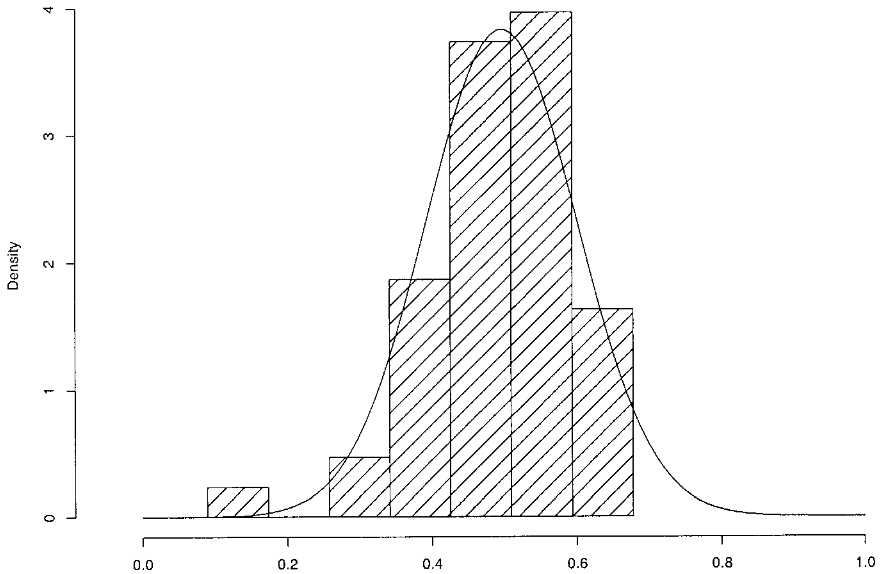

Figure 3 Distribution of proportion of votes going to George Bush with superimposed normal distribution

regression models are highly correlated with actual values (around 88%). Thus our specification has some merit.

4.2 Comparison of regression models

Our results from fitting the different regression models to the voting data are reported in Table 3. Because the patterns we observe in these data across models are so similar to those that we observe using our cable penetration data, we will not discuss them as fully as we did earlier. Thus, for example, we do not compute a table like Table 2, since it would produce no new insights.

As before, we can see in Table 3 that one would be led to draw different inferences about which regressors were statistically significant and which were not depending upon whether one used the beta, simplex or quasi-likelihood model or one of the other regression models. Again, as before, these differences make sense as the linear models fail nonlinearity tests (at the 5% level), and the residuals of regression models assuming a conditional normal distribution fail kurtosis tests of normality (at the 5% level). Consequently, we are once again able to reject the distributional assumptions of our first four regression models but not our last three regression models.

Given the similarity of evidence between these two case studies, we will focus our attention on the computed $\mathrm{AIC_c}$ statistic for each regression model. As before, the beta regression dominates the alternatives, but it dominates the quasi-likelihood model in a more pronounced way. The fact that a parametric model, such as the beta regression model, dominates the quasi-likelihood model in a smaller sample is consistent with the fact that quasi-likelihood models are expected to be better approximations to

Table 3 Results for different regression models of voting for George Bush in the 2000 Presidential election

$k(\mathbf{x})f(y\|\mathbf{x})$	Linear normal	Linear censored normal	Transformed logistic normal	Logistic normal	Logistic beta	Logistic simplex	Logistic unspecified
Estimation method[a]	LS	ML	LS	LS	ML	ML	QML
constant[b]	−5.5324**	−5.5324**	−37.5203**	−28.5359**	−38.4671**	−40.5041**	−32.8526**
	(1.9909)	(1.8067)	(8.4419)	(10.4428)	(7.4161)	(6.7868)	(9.542)
lnpop	0.0167	0.0167	0.0928*	0.0774	0.1083**	0.0983**	0.0852**
	(0.0114)	(0.0104)	(0.0486)	(0.0501)	(0.0410)	(0.0428)	(0.0451)
clfu	−2.2608*	−2.2608*	−10.1201*	−9.5962*	−9.8385**	−10.0596**	−9.8078**
	(1.3202)	(1.1980)	(5.5978)	(5.5269)	(4.1347)	(4.3956)	(4.5021)
male	40.2568**	40.2568**	258.553**	192.690**	260.816**	279.494**	224.048**
	(16.8208)	(15.2646)	(71.3241)	(84.1921)	(54.0491)	(53.3667)	(72.506)
mgt18	−28.2229**	−28.2229**	−185.143**	−136.255**	−186.137**	−200.541**	−159.487**
	(13.2065)	(11.9845)	(55.9986)	(65.0604)	(41.6349)	(41.8948)	(55.6544)
pgt65	−0.9682*	−0.9682*	−2.6945	−3.5968	−1.7759	−2.2237	−3.1051
	(0.5611)	(0.5092)	(2.3792)	(2.4339)	(2.4892)	(2.5229)	(2.8105)
urban	−0.0019**	−0.0019**	−0.0094**	−0.0086**	−0.0101**	−0.0095**	−0.0089**
	(0.0006)	(0.0006)	(0.0026)	(0.0027)	(0.0024)	(0.0023)	(0.0024)
gt75k	−0.0052	−0.0052	−0.0147	−0.0189	−0.0111	−0.01298	−0.0167
	(0.0034)	(0.0031)	(0.0145)	(0.0148)	(0.0127)	(0.0127)	(0.0134)
bpovl	0.6754	0.6754	2.8286	2.8330	2.6586	2.8266	2.8362
	(0.5497)	(0.4988)	(2.3308)	(2.2850)	(1.9224)	(1.8375)	(1.8635)
AIC_c[c]	−1.8038	−1.8038	−4.3387	−4.3538	−4.421	−4.3761	−4.3838

[a] LS, least squares; ML, maximum likelihood; QML, quasi-maximum likelihood.

[b] Standard errors are reported in parentheses below coefficient estimate. ** Significance at the 5% level; * significance at the 10% level.

[c] This statistic represents a version of the Akaike Information Criteria derived in McQuarrie and Tsai (1998).

parametric models in larger samples. Our evidence is consistent with this fact and so suggests that choice between the beta and quasi-likelihood regression models will turn in part on the size of the sample that one is studying.

5 Summary

Many types of studies examine the influence of selected variables on the conditional expectation of a proportion or vector of proportions. These studies can be sorted into four distributional categories. Of these categories, we focus on proportions observed on the open interval (0, 1).

Surveying what regression models prior researchers have used to analyse these types of data, we identify seven basic regression models, when categorized according to their characterization of the conditional distribution and the conditional expectation function. Several of these regression models ignore the fact that since these variates are bounded, their conditional expectation function must be nonlinear. Further, a number of these regression models also ignore the fact that since these variates are bounded, their error distributions must be heteroskedastic since their conditional variance must approach zero as their conditional mean approaches either of their boundary points.

We studied the application of these seven regression models to two common data sets with common specifications of the regressors. These case studies allow us to examine the conformance of the data to each model's distributional assumptions, how the parameter estimates and inferences differ across regression models, and which model best describes the data.

Based upon these comparisons, we recommend that researchers in the future use either a parametric regression model based upon the beta distribution or the quasi-likelihood model developed by Papke and Wooldridge (1997). Concerning the choice between the parametric and quasi-likelihood model, we recommend that researchers use the parametric regression model unless their sample is large enough to justify the asymptotic arguments underlying the quasi-likelihood approaches for the reasons discussed in Godfrey (1988).

Regardless of choice, we strongly recommend that future research recognize that the data are likely generated by a distribution for which the mean is a nonlinear function of the regressors and the variance is a function of the mean. Consequently, even a linear regression on a logit transformed dependent variable is preferable to a linear regression on a nontransformed variable. This recognition is critically important when one tries to derive policy implications from one's parameter estimates since the effect of covariates will change as they deviate from their sample means.

We hope that our comparisons will give researchers some guidance on how to approach the analysis of these data in the future, and whether the inferences drawn in some prior studies should be treated with caution. However, we recognize that our evidence only applies to one of the four distributional categories that we identified earlier. We leave the question of how best to conduct regression analyses of the other categories of proportional data to future research as they raise additional statistical issues to those considered in this study.

Acknowledgements

The authors wish to thank J Herman, J Prisbrey and D Webbink for constructive comments on earlier drafts of this paper. The authors especially wish to thank B Marx (the editor), an anonymous associate editor and our referees for helping to improve our paper by their suggestions.

References

Aitchison J (1986) *The statistical analysis of compositional data.* New York, NY: Chapman and Hall.

Albarran A, Umphrey D (1994) Marketing cable and pay cable services: impact of ethnicity, viewing motivations, and program types. *Journal of Media Economics*, **7**, 47–58.

Atkin D (1992) A profile of cable subscribership: the role of audience satisfaction variables. *Telematics and Informatics*, **9**, 53–60.

Atkinson A (1985) *Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis.* New York: Oxford University Press.

Barclay MJ, Smith CW (1995) The determinants of corporate leverage and dividend policies. *Journal of Applied Corporate Finance*, **7**, 4–19.

Barndorff-Nielsen OE, Jorgensen B (1991) Some parametric models on the simplex. *Journal of Multivariate Analysis*, **39**, 106–16.

Cox C (1996) Nonlinear quasi-likelihood models: applications to continuous proportions. *Computational Statistics & Data Analysis*, **21**, 449–61.

Demsetz H, Lehn K (1985) The structure of corporate ownership: causes and consequences. *Journal of Political Economy*, **93**, 1155–77.

DeSarbo, W, Ramaswamy V, Lenk P (1993) A latent class procedure for the structural analysis of two-way compositional data. *Journal of Classification*, **10**, 159–93.

Dyke GV, Patterson HD (1952) Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.

Fahrmeir L, Tutz G (1994) *Multivariate statistical modelling based on generalized linear models.* New York: Springer-Verlag.

Federal Communications Commission (1993) FCC 93-177, Report and Order and Further Notice of Proposed Rule Making, MM Docker 92-266 (3 May 1993), 6134.

Federal Communications Commission (1994) FCC 94-38, Second Order on Reconsideration, Fourth Report and Order and Fifth Notice of Proposed Rulemaking, MM Docket 92-266 (30 March 1994), 4277.

Godfrey L (1988) *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches.* New York: Cambridge University Press.

Hermalin B, Wallace N (1994) The determinants of efficiency and solvency in savings and loans. *The RAND Journal of Economics*, **25**, 361–81.

Hviid M, Villadsen B (1995) Beta distributed market shares in a spatial model with an application to the market for audit services. *Review of Industrial Organization*, **10**, 737–47.

Johnson NL, Kotz S, Balakrishnan N (1995) *Continuous univariate distributions*, volumes 1 and 2. New York: John Wiley & Sons, Inc.

Jorgensen B (1983) Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19–28.

Jorgensen B (1997) *The theory of dispersion models.* New York: Chapman & Hall.

Kmenta J (1986) *Elements of econometrics*, 2nd edition. New York: Macmillan Publishing.

LaRose R, Atkin D (1988) Satisfaction, demographic, and media environment predictors of cable subscription. *Journal of Broadcasting and Electronic Media*, **32**, 403–14.

Maddala GS (1991) A perspective on the use of limited-dependent and qualitative variables models in accounting research. *The Accounting Review*, **66**, 786–807.

McCullagh P, Nelder JA (1989) *Generalized linear models.* New York: Chapman & Hall.

McDonald JB, Xu YJ (1995) A generalization of the beta distribution with applications. *Journal of Econometrics*, **66**, 133–52.

McQuarrie A, Tsai C (1998) *Regression and time series model selection.* New Jersey: World Scientific Publishing Company.

Mehran H (1995) Executive compensation structure, ownership, and firm performance. *Journal of Financial Economics*, **38**, 163–84.

Mielke PW (1975) Convenient beta distribution likelihood techniques for describing and comparing meterological data. *Journal of Applied Meteorology*, **14**, 985–90.

Mittelhammer RC (1996) *Mathematical statistics for economics and business.* New York: Springer-Verlag.

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**, 619–32.

Park R (1972) Prospects for cable in the 100 largest television markets. *Bell Journal of Economics and Management Science*, **3**, 130–50.

Reagan J, Ducey R, Bernstein J (1985) Local predictors of basic and pay cable subscription. *Journalism Quarterly*, **59**, 397–400.

Sparkes V, Kang N (1986) Public reactions to cable television: time in the diffusion process. *Journal of Broadcasting*, **27**, 163–75.

Scott, D (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–10

Titterington DM (1989) Logistic-normal distribution. In Kotz S, Johson NL, Reed CB eds. *Encyclopedia of statistical sciences* (supplementary volume). New York: John Wiley & Sons, pp. 90–91.

Webb GK (1983) *The economics of cable television.* Lexington, MA: Lexington Books.