Journal of Business & Economic Statistics
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/ubes20

Flexible Approximation of Subjective Expectations Using Probability Questions

Charles Bellemare [a] , Luc Bissonnette [b] & Sabine Kröger [c]

[a] Département d'Économique, Université Laval, , Pavillon J.-A.-DeSève, 1025, avenue des Sciences-Humaines, Québec (Québec), G1V 0A6, Canada

[b] Department of Econometrics and OR, Tilburg University, Warandelaan 2, PO Box 90153, 5000, LE Tilburg, The Netherlands

[c] Département d'Économique, Université Laval, Pavillon J.-A.-DeSève, 1025, avenue des Sciences-Humaines, Québec (Québec), G1V 0A6, Canada

Version of record first published: 06 Apr 2012

To cite this article: Charles Bellemare, Luc Bissonnette & Sabine Kröger (2012): Flexible Approximation of Subjective Expectations Using Probability Questions, Journal of Business & Economic Statistics, 30:1, 125-131

To link to this article: http://dx.doi.org/10.1198/jbes.2011.09053

Flexible Approximation of Subjective Expectations Using Probability Questions

Charles BELLEMARE
Département d'Économique, Université Laval, Pavillon J.-A.-DeSève, 1025, avenue des Sciences-Humaines, Québec (Québec) G1V 0A6, Canada (*cbellemare@ecn.ulaval.ca*)

Luc BISSONNETTE
Department of Econometrics and OR, Tilburg University, Warandelaan 2, PO Box 90153, 5000 LE Tilburg, The Netherlands (*l.bissonnette@uvt.nl*)

Sabine KRÖGER
Département d'Économique, Université Laval, Pavillon J.-A.-DeSève, 1025, avenue des Sciences-Humaines, Québec (Québec) G1V 0A6, Canada (*skroger@ecn.ulaval.ca*)

We propose a flexible method to approximate the subjective cumulative distribution function of an economic agent about the future realization of a continuous random variable. The method can closely approximate a wide variety of distributions while maintaining weak assumptions on the shape of distribution functions. We show how moments and quantiles of general functions of the random variable can be computed analytically and/or numerically. We illustrate the method by revisiting the determinants of income expectations in the United States. A Monte Carlo analysis suggests that a quantile-based flexible approach can be used to successfully deal with censoring and possible rounding levels present in the data. Finally, our analysis suggests that the performance of our flexible approach matches that of a correctly specified parametric approach and is clearly better than that of a misspecified parametric approach.

KEY WORDS: Approximation of subjective probability distribution; Elicitation of probabilities; Spline interpolation.

1. INTRODUCTION

The measurement of subjective expectations has proven useful for eliciting knowledge of economic agents and experts on the future realization of various economic variables (e.g., Dominitz and Manski 1997; Engelberg, Manski, and Williams 2009) and improving the empirical content of stochastic models of choice under uncertainty (Bellemare, Kröger, and van Soest 2008; Delavande 2008). It has been advocated that the measurement of expectations should proceed by first measuring subjective probability distributions. In particular, there is growing evidence that agents reveal different points of their subjective distribution (mean, median, or other quantiles) when asked for their best point prediction of a future event (see Manski 2004 for a review). Thus, deriving expectations from probability distributions can improve interpersonal comparisons while providing more information on the uncertainty faced by respondents.

Up to now, two approaches have been used to make inferences on subjective distributions. The first approach is parametric and assumes that the subjective distribution of a respondent is drawn from a parametric distribution (e.g., a normal or lognormal distribution) that depends on a finite number of unknown parameters. As with most parametric approaches, misspecification of the underlying distribution may lead to biased forecasts and inferences. The second approach is fully nonparametric, placing no restriction of the nature or shape of subjective distributions. This approach overcomes potentials biases due to misspecification of the underlying distribution at the expense of providing set rather than point identification of the functionals of interest.

In this paper, we present a flexible method that yields point identification of the distribution function of a respondent while maintaining weak assumptions on the shape of the underlying distribution. The flexible approach builds on cubic spline interpolation, which requires only that the underlying distributions be twice differentiable on their support. Moreover, the estimation by cubic splines involves solving a system of linear equations. Thus our flexible approach provides a simple analytical solution for the estimated function. Cubic splines are well-known interpolation methods (see, e.g., Judd 1998); however, to the best of our knowledge, they have not been applied to fit individual specific cumulative distribution functions using subjective expectations data. The closest work using interpolation methods to fit a cumulative distribution function is that of Kriström (1990), who estimated the population-level distribution of willingness to pay for an environmental good using linear interpolation and aggregated survey responses to valuation questions.

We illustrate our approach by revisiting the determinants of expectations concerning future income using data from the Survey of Economic Expectations (SEE). These data are characterized by high levels of censoring and potential rounding. Censoring occurs when individuals report a nonzero probability that the future outcome will fall outside the range of potential values spanned by the probability questions. The parametric approach maintains sufficiently strong distributional assumptions

Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1198/jbes.2011.09053

to deal with censoring. In contrast, the flexible approach maintains weaker distributional assumptions. As a result, estimated moments will be affected by censoring. To overcome this problem, we propose a quantile-based flexible approach that uses the estimated median as a measure of central tendency and the estimated interquartile range (IQR) as a measure of respondent uncertainty. We compare estimators of the determinants of expectations and uncertainty using both a specific parametric approach and our quantile-based flexible approach. We find that both approaches provide similar results for most determinants of future income, suggesting that the distributional assumptions chosen to implement the parametric approach are reasonable.

In the final part of the article, we present a Monte Carlo analysis designed to measure the impact of censoring and rounding on estimates of the determinants of expectations. We focus on comparing the performance of our flexible approach with that of a correctly specified parametric approach as well as an incorrectly specified parametric approach. We find that the flexible approach generates unbiased estimates of the determinants of expectations. This result holds when we introduce censoring and rounding levels believed to be present in the data. Moreover, the performance of the flexible approach is comparable to that of the correctly specified parametric approach but clearly outperforms the incorrectly specified parametric approach that we consider.

2. A FLEXIBLE APPROACH

Our objective is to approximate the subjective probability distribution $F_i(z) = \Pr_i(Z \le z)$ of a respondent i using his or her answers to J probability questions of the type "what is the percent chance that Z is less than or equal to z_j?," where $z_1 < z_2 < \cdots < z_J$ are threshold values. Thus the J data points available to make inferences on $F_i(z)$ are $\{(z_1, F_i(z_1)), \ldots, (z_J, F_i(z_J))\}$, where $0 \le F_i(z_j) \le 1$ denotes the probability statement to a question with threshold z_j. Censoring occurs when $F_i(z_1) > 0$ and/or $1 - F_i(z_J) > 0$. This implies that some probability mass is not contained within the interval $[z_1, z_J]$.

We propose to use the available data to approximate the subjective cumulative distribution function $F_i(z)$ using cubic spline interpolation. A cubic spline is a piecewise polynomial function defined on $J-1$ intervals, $[z_1, z_2], \ldots, [z_{J-1}, z_J]$. On each interval, the function $F_i(z)$ is assumed to be given by a polynomial $a_j + b_j z + c_j z^2 + d_j z^3$, where (a_j, b_j, c_j, d_j) are the interval-specific polynomial coefficients. The spline approximation of the function $F_i(z)$ is constructed by simply connecting the different polynomials at the relevant threshold values. The set $\{(a_j, b_j, c_j, d_j) : j = 1, \ldots, J-1\}$ contains the $4(J-1)$ unknown polynomial coefficients to be estimated. Exploiting continuity at the endpoints and interior thresholds provides $2J-2$ equations

$$F_i(z_j) = a_j + b_j z_j + c_j z_j^2 + d_j z_j^3 \quad \text{for } j = 2, \ldots, J-1,$$

$$F_i(z_j) = a_{j+1} + b_{j+1} z_j + c_{j+1} z_j^2 + d_{j+1} z_j^3 \quad \text{for } j = 2, \ldots, J-1,$$

$$F_i(z_1) = a_1 + b_1 z_1 + c_1 z_1^2 + d_1 z_1^3,$$

$$F_i(z_J) = a_{J-1} + b_{J-1} z_J + c_{J-1} z_J^2 + d_{J-1} z_J^3.$$

Next, restrictions that the first and second derivatives of $F_i(\cdot)$ agree at the interior thresholds generate $2J-4$ additional equations

$$b_j + 2c_j z_j + 3d_j z_j^2 = b_{j+1} + 2c_{j+1} z_j + 3d_{j+1} z_j^2 \quad \text{for } j = 2, \ldots, J-1,$$

$$2c_j + 6d_j z_j = 2c_{j+1} + 6d_{j+1} z_j \quad \text{for } j = 2, \ldots, J-1.$$

Two more conditions, so-called "boundary conditions" at the endpoints, are needed to estimate the $4(J-1)$ polynomial coefficients of the cubic spline. There is very little guidance in the literature to chose these boundary conditions. Here we chose to impose that $F_i''(z_1) = F_i''(z_J) = 0$, yielding what is known in the literature as a natural cubic spline (see Judd 1998). Thus restrictions on the derivatives and the boundary conditions generate a system of $4(J-1)$ linear equations that can be solved for the $4(J-1)$ unknown parameters. We experimented with boundary conditions restricting the first derivative at both endpoints $F_i'(z_1) = F_i'(z_J) = 0$ or by mixing restrictions on first and second derivatives [e.g., setting $F_i''(z_1) = F_i'(z_J) = 0$ or $F_i'(z_1) = F_i''(z_J) = 0$]. We found that these changes had only minor effects on the estimated splines. We also experimented with linear and quadratic splines and found the cubic spline approximation to be superior. We did not find that increasing the order of the spline further increased the quality of the approximation. Thus, we use natural cubic splines throughout the rest of the article.

Absent censoring, moments can be directly estimated from the fitted subjective cumulative distribution function. In particular, the λth noncentral moment of Z can be computed analytically using

$$\widehat{\mathbf{E}}_i(Z^\lambda) = \sum_{j=1}^{J-1} \left[\frac{\widehat{b}_j z^{\lambda+1}}{\lambda+1} + \frac{2\widehat{c}_j z^{\lambda+2}}{\lambda+2} + \frac{3\widehat{d}_j z^{\lambda+3}}{\lambda+3} \Bigg|_{z_j}^{z_{j+1}} \right]. \tag{1}$$

Approximating $\mathbf{E}_i(h(Z)) = \int h(z)\, dF_i(z)$ of a general function $h(\cdot)$ is slightly more complicated. In such cases, numerical integration can be performed by quadrature or simulation using $\widehat{F}_i(z)$. Similarly, quantiles can be obtained numerically by inverting $\widehat{F}_i(z)$. Quantiles are especially useful in the presence of censoring, which occurs when survey respondents report a nonzero probability that Z will fall below z_1 and/or above z_J. In such cases, relevant medians can be used as a measure of central tendency, and the interquartile range (IQR) can be computed as a measure of subjective uncertainty as long as $F_i(z_1)$ and $1 - F_i(z_J)$ are less than or equal to 0.25.

We illustrate the flexible approach by fitting three different distributions: a symmetric standard normal, an asymmetric chi-squared distribution with 3 degrees of freedom, and a bimodal distribution (with modes at $\pi/2$ and $5\pi/2$). The density of the bimodal distribution is given by $\frac{\sin(z)+1}{A}$ over the $[0, 3\pi]$ interval, where $A = 2 + 3\pi$ ensures that the function integrates to 1 over its domain. We fitted each cumulative distribution function using between four and six data points equally spaced between 3 and -3 for the normal distribution, between 0 and 8 for the chi-squared distribution, and between 0 and 3π for the bimodal distribution. The results are reported in Figure 1. As expected, the goodness of fit increases with the number of data points for all three interpolations. A slight approximation error

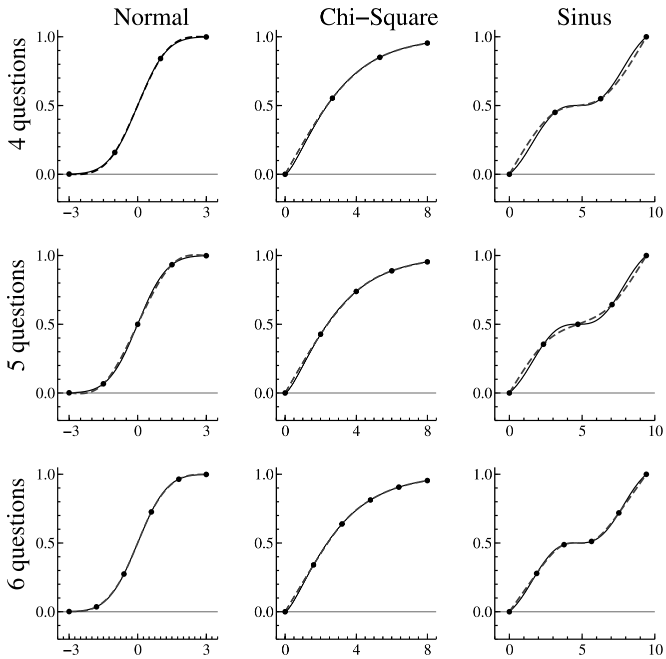

Figure 1. Fitted normal, chi-squared, and sinus distributions using cubic spline interpolations with four to six data points (questions). The solid lines represent the true distributions. The dashed lines represent the fitted distributions using the data points (dark points).

remains in the lower hand of the distribution when the number of data points is increased from four to six. Finally, we find that the approach has more difficulty fitting the bimodal distribution than the other two distributions. In contrast, the interpolation manages to provide a very good fit of the distribution with five or more data points.

Monotonicity

Cubic spline interpolation can produce oscillations that can cause the estimated distribution function to be nonmonotonically increasing. This is particularly problematic when estimating quantiles by inverting $\widehat{F}_i(\cdot)$ to obtain a unique solution. Perhaps the simplest and most effective way to correct for these oscillations is to use the Hyman filter (Hyman 1983). This filter works in two steps. In a first step, define $\widehat{f}_i'(z_j)$ as the estimated value of the first derivative of the spline function at the threshold z_j. Next, define $S_{i-1/2} = (\widehat{F}_i(z_j) - \widehat{F}_i(z_{j-1}))/(z_j - z_{j-1})$ and $S_{i+1/2} = (\widehat{F}_i(z_{j+1}) - \widehat{F}_i(z_j))/(z_{j+1} - z_j)$ as the left-side slope connecting with the previous threshold, $(\widehat{F}_i(z_{j-1}), z_{j-1})$, and the right-side slope connecting with the threshold $(\widehat{F}_i(z_{j+1}), z_{j+1})$. de Boor and Swartz (1977) have shown that if an estimated function satisfies the criteria

$$0 \le \widehat{f}_i'(z_j) \le 3\min\big(S_{i-1/2}, S_{i+1/2}\big), \tag{2}$$

then it is monotone on the interval $[z_j, z_{j+1}]$. Thus the criteria (2) can be used to identify all points where the monotonicity condition is violated. In a second step, the condition of the equality of the second derivatives at each of the thresholds where monotonicity is violated is replaced by

$$\widehat{f}_i'(z_j) = \min\big[\max(0, \widehat{f}_i'(z_j)), 3\min\big(S_{i-1/2}, S_{i+1/2}\big)\big].$$

Hyman (1983) compared his filter approach to correct for nonmonotonicity with various alternative spline methods (e.g., Akima splines) and found that cubic spline interpolation coupled with his filter is the most effective method (in a mean squared error sense) to impose monotonicity on an estimated function.

3. REVISITING EXPECTATIONS OF FUTURE INCOME

In this section we illustrate the flexible approach by revisiting data on income expectations that were previously analyzed in a parametric setting by Dominitz (2001). Data are taken from the 1994–1995 SEE administered through WISCON, a national telephone survey conducted by the University of Wisconsin Survey Center. We focus on the following survey question:

What do you think is the percent chance (or chances out of 100) that your own total income, before taxes, will be under \$$z_j$ (in the next 12 months)?

For each respondent, four initial thresholds z_j were selected based on self-reported minimal and maximal values for their income support. Respondents could then be asked one or two additional questions based on their four answers. A detailed description of the branching algorithm to determine the income level or additional questions was presented by Dominitz (2001). We observe between four and six data points for each of 1249 respondents in the SEE aged 25–59 who were active in the labor force at the time they answered the SEE and who provided all of the information for our analysis.

Figure 2 documents the extent of censoring in these data by plotting the sample distributions of $F_i(z_{1,i})$ and $F_i(z_{J,i})$. We find that only 44% of respondents have uncensored distributions at the lower end [$F_i(z_{1,i}) = 0$ in the left panel], whereas 66% of respondents have uncensored distributions at the upper hand [$F_i(z_{J,i}) = 1$ in the right panel]. Only 37% of all sample respondents have uncensored distributions at both ends, a proportion too low to perform meaningful inferences using predicted moments. We deal with censoring by using the median as the measure of central tendency and the IQR as the measure of dispersion. Note that a small subsample of respondents have $F_i(z_{1,i})$ or $1 - F_i(z_{J,i})$ exceeding 0.25 and (to a lesser extent) exceeding 0.5; thus the estimated medians and/or IQR of respondents in this subsample are potentially biased. We report a Monte Carlo analysis to assess how such biases affect the measurement of the determinants of expectations.

We compared estimates using our proposed quantile-based flexible approach with those of a parametric approach applied to the same data. The parametric approach involves fitting the best lognormal distribution when sufficient data points are available. Respondents who state at most one value of $F_i(z_{j,i})$ that differs from 0 or 1 are fitted with the best *log-triangular* distribution, following the procedure of Engelberg, Manski, and Williams (2009).

We applied the Hyman filter for 850 respondents (68%) to correct for nonmonotonicity of the cumulative distribution function predicted by the flexible approach. Figure 3 presents a scatterplot of the predicted medians (left panel) and IQR (right panel) using both approaches (the flexible on the horizontal axis and the parametric on the vertical axis). We found similar predicted medians for both approaches, with predictions scattering

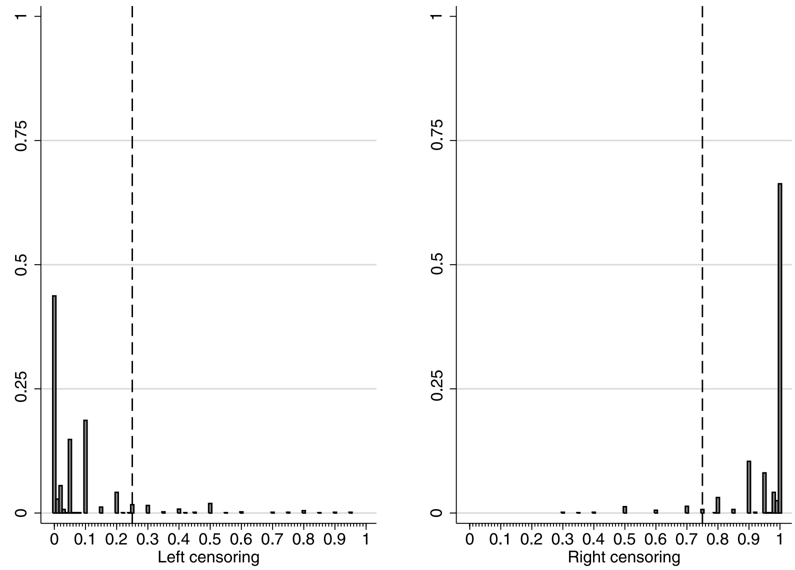

Figure 2. Distribution of $F_i(z_{1,i})$ (left) and $F_i(z_{J,i})$ (right) in the SEE data ($N = 1249$). Dashed lines are at 0.25 (left) and 0.75 (right).

closely and relatively equally below and above the 45-degree line. More important differences emerge when looking at the predicted IQR. There the flexible method tends to predict higher dispersions (74.5% of predicted IQR are below the 45-degree line).

We next estimated linear models for both approaches using the predicted medians and IQR as dependent variables and using a set of independent variables including realized income in the last year, basic demographic characteristics, employment status, and education level (using no high school diploma as the reference class). The first column of Table 1 presents some sample descriptive statistics of these variables. We estimated our models using the ordinary least squares (OLS) estimator with robust standard errors. Results are presented in subsequent columns of Table 1.

Overall, inferences using the flexible and parametric approaches are similar, suggesting that the assumption of expected income following a lognormal distribution is reasonable. Only small differences emerge. For instance, the flexible approach predicts that women and Hispanics expect significantly lower average median future income. Both methods yield different results concerning the effect of unemployment on the uncertainty of future income and the income uncertainty faced by African-Americans. Results of the parametric approach suggest that the currently unemployed face significantly lower income uncertainty, whereas results of the flexible approach indicate that the previously unemployed have significantly higher income uncertainty. The parametric approach finds that African-American respondents have significantly greater income uncertainty, whereas this effect is smaller and insignificant using the flexible approach.

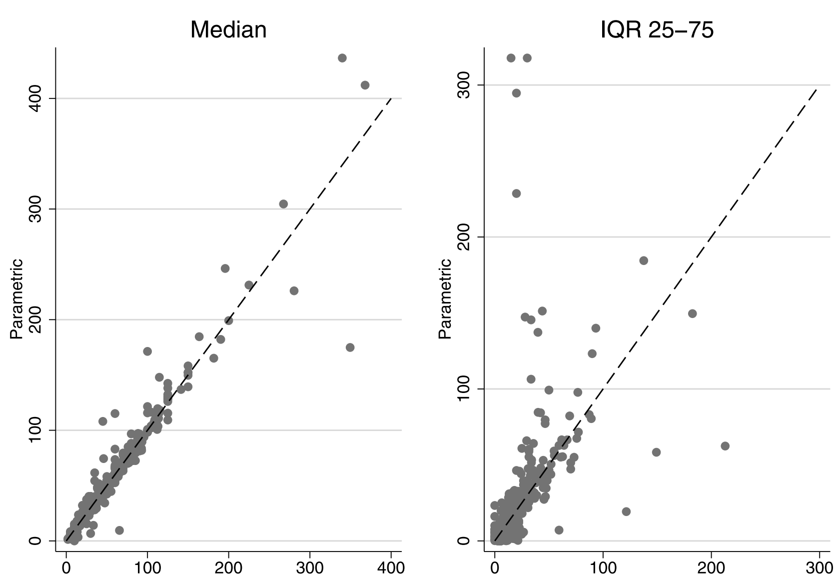

Figure 3. Scatterplot of estimated medians (left) and IQRs (right) of subjective income expectations, with either the parametric (vertical axis) or flexible (horizontal axis) method ($N = 1249$). The dashed line represents the 45-degree line.

Table 1. Determinants of subjective medians and IQR in the SEE using the parametric and flexible approaches

	Desc. stat.	Median		IQR 25–75	
	Mean (SD)	Param.	Flexible	Param.	Flexible
Income	36.592	0.841***	0.828***	0.225***	0.295***
	(32.999)	(0.063)	(0.052)	(0.039)	(0.040)
Self-employed	0.123	4.201**	3.206*	12.910***	9.579***
	(0.329)	(2.019)	(1.887)	(2.446)	(1.364)
Currently unemployed	0.059	0.272	−0.336	−4.031***	−1.291
	(0.236)	(1.897)	(1.431)	(1.271)	(1.140)
Previously unemployed	0.117	−3.956***	−3.532***	2.332	4.137***
	(0.321)	(1.315)	(1.193)	(1.694)	(1.446)
Female	0.467	−2.217	−2.756**	0.808	0.474
	(0.499)	(1.360)	(1.246)	(1.301)	(0.820)
Partner	0.651	0.440	1.226	−1.909	−0.357
	(0.477)	(0.867)	(0.757)	(1.421)	(0.720)
Age	39.001	−0.008	0.036	−0.216***	−0.159***
	(9.155)	(0.049)	(0.041)	(0.068)	(0.045)
White	0.877	−0.113	−0.340	−0.415	−1.624
	(0.329)	(1.539)	(1.459)	(1.659)	(1.360)
Black	0.067	0.516	−0.933	8.640**	2.222
	(0.251)	(2.258)	(2.001)	(4.227)	(1.685)
Hispanic	0.031	−3.080	−4.672**	4.296	−0.529
	(0.174)	(2.475)	(2.088)	(6.392)	(2.182)
High school diploma	0.164	−0.746	−0.813	−3.205	0.149
	(0.371)	(2.400)	(2.306)	(4.290)	(1.428)
Att. college w/o graduating	0.431	0.859	0.871	−2.028	0.619
	(0.495)	(2.344)	(2.255)	(4.202)	(1.355)
College graduate	0.365	5.139**	5.309**	−2.649	−1.326
	(0.482)	(2.598)	(2.449)	(4.172)	(1.534)
Constant		7.454**	6.117**	13.626***	7.019***
		(3.183)	(3.046)	(4.972)	(2.277)
R^2		0.775	0.804	0.146	0.421
N		1249	1249	1249	1249

NOTE: Standard errors are in parentheses (Eicker–White used in OLS estimation). * Significant at 10% level. ** Significant at 5% level. *** Significant at 1% level.

3.1 Monte Carlo Analysis

We conducted a Monte Carlo analysis to assess how censoring and possible rounding in the SEE income data can affect the results in our application. Our analysis focuses on comparing the performance of our proposed flexible approach with the performance of the parametric approach, using both correctly specified and misspecified distribution functions for the parametric approach. We begin by specifying the data-generating process of medians med_i and interquartile ranges IQR_i

$$\text{med}_i = \theta_0 + \theta_1 x_{1i} + \varepsilon_i, \tag{3}$$

$$\text{IQR}_i = \gamma_0 + \gamma_1 x_{2i} + \eta_i, \tag{4}$$

where x_{1i} and x_{2i} are two determinants, and where ε_i and η_i denote homoscedastic unobserved heterogeneity. Our objective is to analyze the properties of the OLS estimator of $(\theta_0, \theta_1, \gamma_0, \gamma_1)'$ in the presence of censoring and rounding. To proceed, we specify (3) and (4) as equations generating quantiles of a Kumaraswamy distribution defined over the [0, 1] interval with parameters $(\alpha_i \geq 0, \beta_i \geq 0)$. The Kumaraswamy distribution is sufficiently flexible to accommodate a wide range of symmetric and asymmetric distributions of potential outcomes (Kumaraswamy 1980). For example, $(\alpha_i = 2, \beta_i = 2)$ implies a symmetric distribution centered at 0.5, whereas $(\alpha_i = 1, \beta_i = 5)$ produces a severely left-skewed distribution with mode at 0.2. We specify our data-generating process in the following way. First, values of x_{1i} and x_{2i} are drawn from a uniform distribution on the $[-0.5, 0.5]$ interval, whereas values of ε_i and η_i are each drawn from a standard normal distribution with mean 0 truncated to the $[-0.1, 0.1]$ interval. Finally, we set $(\theta_0 = 0.5, \theta_1 = 0.3, \gamma_0 = 0.5, \gamma_1 = 0.3)$. These data-generating processes force both med_i and IQR_i to lie within $[0.25, 0.75]$. We next present in detail the steps performed in our Monte Carlo simulations. Our analysis of the flexible and parametric approaches differs only with respect to step 4.

Step 1. Draw $(\text{med}_i, \text{IQR}_i)$ for $i = 1, 2, \ldots, N$ using equations (3) and (4).

Step 2. Compute for each i the parameters (α_i, β_i) corresponding Kumaraswamy distribution by numerically solving the following system of equations:

$$\mathrm{med}_i = Q_{0.5}(\alpha_i, \beta_i), \tag{5}$$

$$\mathrm{IQR}_i = Q_{0.75}(\alpha_i, \beta_i) - Q_{0.25}(\alpha_i, \beta_i) \tag{6}$$

such that $Q_\kappa(\alpha_i, \beta_i) = F^{-1}(\kappa; \alpha_i, \beta_i)$ where $F^{-1}(\cdot)$ denotes the inverse mapping of the Kumaraswamy cumulative distribution function $F(x) = 1 - (1 - x^{\alpha_i})^{\beta_i}$ evaluated at $0 \leq \kappa \leq 1$ with parameters (α_i, β_i).

Step 3. Generate points $\{z_{j,i} : j = 2, \ldots, J-1\}$ using a branching algorithm inspired by our empirical application. In particular, respondents with $\mathrm{med}_i \leq 0.42$ are assigned the vector of thresholds $(0, 0.125, 0.25, 0.4, 0.7, 1)$, those with $0.42 < \mathrm{med}_i < 0.59$ are assigned thresholds $(0, 0.2, 0.4, 0.6, 0.8, 1)$, and those with $\mathrm{med}_i \geq 0.59$ are assigned thresholds $(0, 0.3, 0.6, 0.75, 0.875, 1)$. As with our empirical application, this algorithm assumes that prior information about the location of the distribution is used to generate thresholds. Then the cumulative probabilities $F(z_{j,i}; \alpha_i, \beta_i)$ are computed at all $z_{j,i}$ values.

Step 4 (Flexible approach). Compute estimates $\widehat{\mathrm{med}}_i$ and $\widehat{\mathrm{IQR}}_i$ using the flexible approach.

Step 4 (Parametric approach). Compute the value of $\boldsymbol{\delta}$ that minimizes the following loss function:

$$\widehat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \sum (\Pr(Z \leq z_{j,i}; \boldsymbol{\delta}) - F_i(z_{j,i}, r))^2,$$

where the summation is over the data points of respondent i and $\Pr(Z \leq z_{j,i}; \boldsymbol{\delta})$ denotes a parametric cumulative distribution function with an unknown vector of parameters $\boldsymbol{\delta}$. We consider the correctly specified case where $\Pr(Z \leq z_{j,i}; \boldsymbol{\delta})$ is correctly chosen to be the Kumaraswamy distribution with parameters $\boldsymbol{\delta} = [\alpha_i, \beta_i]$. We also consider a misspecified case where $\Pr(Y \leq z_{j,i}; \boldsymbol{\delta})$ is chosen to be the Normal distribution with mean τ_i and variance γ_i^2. We compute estimates $\widehat{\mathrm{med}}_i$ and $\widehat{\mathrm{IQR}}_i$ using $\widehat{\boldsymbol{\delta}}$.

Step 5. Estimate the following equations:

$$\widehat{\mathrm{med}}_i = \theta_0 + \theta_1 x_i + \overline{\varepsilon}_i, \tag{7}$$

$$\widehat{\mathrm{IQR}}_i = \gamma_0 + \gamma_1 x_i + \overline{\eta}_i, \tag{8}$$

where $\overline{\varepsilon}_i = \varepsilon_i + \widehat{\mathrm{med}}_i - \mathrm{med}_i$ and $\overline{\eta}_i = \eta_i + \widehat{\mathrm{IQR}}_i - \mathrm{IQR}_i$. Equations (7) and (8) are identical to equations (3) and (4), except that the true medians and IQRs are replaced by approximated values generated using either the parametric approach or the flexible approach. Estimated values $(\widehat{\theta}_0, \widehat{\theta}_1)'$ and $(\widehat{\gamma}_0, \widehat{\gamma}_1)'$ are saved. We repeat steps 1–5 to for 10,000 samples of size 100.

The foregoing five steps generate our baseline results without censoring or rounding. To analyze the effects of rounding, we replace the probabilities $F(z_j; \alpha_i, \beta_i)$ in step 3 by the closest of the following numbers: 0, 1, 2, 3, 5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 65, 70, 75, 80, 85, 90, 95, 97, 98, 99, or 100. This sequence closely matches the probability responses in our application. It also is one of the main rounding patterns discussed in the literature (see Manski and Molinari 2010).

To analyze the effects of censoring, we randomly draw for each i a pair of censoring levels from below and from above using the empirical distribution of censoring levels presented in Figure 2. Let c_i^0 and c_i^1 denote these censoring levels. We then rescale the thresholds assigned in step 3 such that $z_{1,i} = Q_{c_i^0}(\alpha_i, \beta_i)$ and $z_{J,i} = Q_{1-c_i^1}(\alpha_i, \beta_i)$.

We evaluate the performance of the flexible and parametric approaches with rounding and censoring by computing parameter and standard error biases. Parameter bias is computed using $(\frac{1}{S}\sum_{s=1}^{S} \widehat{\phi}^s - \phi)/\phi$, where $\phi \in \{\theta_0, \theta_1, \gamma_0, \gamma_1\}$ are the true values and $\widehat{\phi}^s$ denotes the estimated parameter in simulation $s \leq S = 10{,}000$. We also compute the percent bias of the estimated standard errors using $(\frac{1}{S}\sum_{s=1}^{S} \mathrm{se}(\widehat{\phi}^s) - \mathrm{sd}(\widehat{\phi}^s))/\,\mathrm{sd}(\widehat{\phi}^s)$, where $\mathrm{sd}(\widehat{\phi}^s)$ denotes the standard deviation of all $\widehat{\phi}^s$ and $\mathrm{se}(\widehat{\phi}^s)$ denotes the standard error predicted using the covariance matrix of the OLS estimator with homoscedasticity $(\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Thus we report the percent difference between the average standard error predicted by the OLS estimator and the actual standard deviation of the estimates over the 10,000 simulations.

Table 2 presents the results. We see that under the baseline scenario (no censoring or rounding), both parameter and standard error biases are small and negligible for the flexible and correctly specified parametric approaches. Of note, these results also hold when censoring and rounding levels believed to be present in our data are incorporated in the analysis. This suggests that results of our empirical application are robust to censoring and possible rounding in the data. We also find that our flexible approach clearly outperforms the misspecified parametric approach based on the erroneous assumption

Table 2. Lines $(\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\gamma}_0, \widehat{\gamma}_1)$ present the corresponding parameter biases computed using $(\frac{1}{S}\sum_{s=1}^{S} \widehat{\phi}^s - \phi)/\phi$, where $\phi \in \{\gamma_0, \gamma_1, \theta_0, \theta_1\}$ are the true values and $\widehat{\phi}^s$ denotes the estimated parameter in simulation $s \leq S = 10{,}000$. Lines std$(\cdot)$ present the percent bias of the estimated standard errors of the corresponding estimated parameter using $(\frac{1}{S}\sum_{s=1}^{S} \mathrm{se}(\widehat{\phi}^s) - \mathrm{sd}(\widehat{\phi}^s))/\,\mathrm{sd}(\widehat{\phi}^s)$, where $\mathrm{sd}(\widehat{\phi}^s)$ denotes the standard deviation of all $\widehat{\phi}^s$, and where $\mathrm{se}(\widehat{\phi}^s)$ denotes the standard error predicted using the covariance matrix of the OLS estimator with homoscedasticity $(\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

	Flexible			Correct parametric			Misspecified parametric		
	Baseline	Rounding	Censoring	Baseline	Rounding	Censoring	Baseline	Rounding	Censoring
$\widehat{\theta}_0$	0.000	−0.000	−0.000	0.000	0.000	−0.000	−0.001	−0.001	−0.002
$\widehat{\theta}_1$	−0.008	−0.002	−0.007	0.001	0.001	−0.001	−0.241	−0.240	−0.223
$\widehat{\gamma}_0$	−0.010	−0.013	−0.008	0.000	−0.002	−0.000	−0.171	−0.173	−0.146
$\widehat{\gamma}_1$	−0.011	−0.016	−0.018	−0.000	0.001	−0.000	−0.430	−0.431	−0.367
std$(\widehat{\theta}_0)$	−0.002	0.007	−0.011	−0.002	0.002	−0.011	−0.003	0.004	−0.013
std$(\widehat{\theta}_1)$	−0.011	−0.005	−0.011	−0.015	−0.014	−0.014	−0.002	−0.003	−0.009
std$(\widehat{\gamma}_0)$	0.003	−0.006	0.005	0.001	−0.007	0.004	0.002	−0.005	0.003
std$(\widehat{\gamma}_1)$	0.040	0.046	0.035	−0.004	−0.002	−0.001	−0.027	−0.020	−0.030

that distributions are normal in the population. There parameter bias is substantial: -24% for $\widehat{\theta}_1$, -17% for $\widehat{\gamma}_0$, and -43% for $\widehat{\gamma}_1$. These biases are not affected by censoring and rounding.

4. CONCLUSION AND DISCUSSION

Our Monte Carlo analysis suggests that the quantile-based flexible approach is robust to levels of rounding discussed in the literature and can accommodate censoring levels present in our data. We found that the flexible approach is comparable to a (first-best) correctly specified parametric approach in terms of bias and efficiency. Moreover, it clearly outperforms the misspecified parametric approach that we consider. We interpret these results as an indication that the flexible approach represents a potentially useful alternative to the existing parametric approach when researchers have little prior knowledge of the shape of the underlying distributions.

The flexible approach has three limitations. First, it lacks a distribution theory which would allow one to make inferences on individual specific distribution functions. This limitation might not pose a significant problem in practice, given that research on subjective expectations has focused on making statistical inferences on the determinants on expectations rather than on individual distribution functions. A second limitation is that moments are biased in the presence of censoring. This is expected because the flexible approach maintains weak assumptions on the shape of the distribution, thereby preventing extrapolation outside of the support spanned by the probability questions. Finally, our quantile-based flexible approach can accommodate only moderate levels of censoring.

Greater levels of censoring can be dealt with in several ways. The first and simplest way is to drop observations with excessive censoring. Though simple, this approach may introduce selection biases if the observations dropped represent a nonrandom subset of observations. A second way is to revert back to the parametric approach and maintain stronger distributional assumptions. Although this would allow accounting for censoring in the data, adopting a fully parametric approach introduces possible specification biases. Our analysis suggests that such biases can be sizeable. Finally, the survey design could be improved by designing probability questions to gather information on a larger range of possible outcomes. The flexible approach could then be used to make inferences while maintaining weaker assumptions on the underlying distributions.

ACKNOWLEDGMENTS

The authors thank Arthur van Soest, participants at the ESA meetings in Nottingham and Tucson, two referees, an associate editor, and the editor for comments. Financial support was provided by the FQRSC. An OX code with files implementing all the procedures discussed in this paper can be downloaded at *http://www.ecn.ulaval.ca/charles.bellemare/*.

[Received February 2009. Revised April 2011.]

REFERENCES

Bellemare, C., Kröger, S., and van Soest, A. (2008), "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities," *Econometrica*, 76, 815–839. [125]

de Boor, C., and Swartz, B. (1977), "Piecewise Monotone Interpolation," *Journal of Approximation Theory*, 21 (4), 411–416. [127]

Delavande, A. (2008), "Pill, Patch or Shot? Subjective Expectations and Birth Control Choice," *International Economic Review*, 49, 999–1042. [125]

Dominitz, J. (2001), "Estimation of Income Expectations Models Using Expectations and Realization Data," *Journal of Econometrics*, 102, 165–195. [127]

Dominitz, J., and Manski, C. F. (1997), "Using Expectations Data to Study Subjective Income Expectations," *Journal of the American Statistical Association*, 92, 855–867. [125]

Engelberg, J., Manski, C., and Williams, J. (2009), "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business & Economic Statistics*, 27, 30–41. [125,127]

Hyman, J. M. (1983), "Accurate Monotonicity Preserving Cubic Interpolation," *SIAM Journal on Scientific Computing*, 4 (4), 645–654. [127]

Judd, K. L. (1998), *Numerical Methods in Economics*, Cambridge, MA: MIT Press. [125,126]

Kristrom, B. (1990), "A Non-Parametric Approach to the Estimation of Welfare Measures in Discrete Response Valuation Studies," *Land Economics*, 66, 135–139. [125]

Kumaraswamy, P. (1980), "A Generalized Probability Density Function for Double-Bounded Random Processes," *Journal of Hydrology*, 46, 79–88. [129]

Manski, C. F. (2004), "Measuring Expectations," *Econometrica*, 72 (5), 1329–1376. [125]

Manski, C. F., and Molinari, F. (2010), "Rounding Probabilistic Responses in Surveys," *Journal of Business & Economic Statistics*, 28, 219–231. [130]