SUPPLEMENT TO "BARGAINING WITH ASYMMETRIC INFORMATION: AN EMPIRICAL STUDY OF PLEA NEGOTIATIONS"

(*Econometrica*, Vol. 85, No. 2, March 2017, 419–452)

BY BERNARDO S. SILVEIRA

APPENDIX B: ESTIMATION APPENDIX

B.1. *Estimation of the Model Primitives*

THE PRIMITIVES TO BE ESTIMATED ARE THE COST PARAMETERS α_d, β_d, α_p, and β_p; the distribution of defendants' types $F(\cdot)$; and the distribution of trial sentences, characterized by $\nu(z)$ and $g(\cdot|Z=z)$.

For $z \in \{l, h\}$, I can trivially estimate $\nu(z)$ by using the empirical probabilities that $\Psi = 0$, conditional on $Z = z$. To recover the other primitives, I follow the steps outlined in the proof of Proposition 2, which shows how to recover $F(\cdot)$ and $g(\cdot|Z=z)$, based on the parameters α_d, β_d, α_p, and β_p. Equation (A.1) holds for all $t \in [\underline{t}, \bar{t}]$, and thus defines a system of infinitely many equations. Notice that, besides α_d, β_d, α_p, and β_p, the system contains the parameter μ, which captures the behavior of $F(\cdot)$ for values of θ lower than $\tilde{\theta}(\underline{t})$.[46] Let $\boldsymbol{\omega} \equiv [\alpha_d\ \beta_d\ \alpha_p\ \beta_p\ \mu]$ be the vector of all unknown variables in (A.1). The model is overidentified, so I estimate $\boldsymbol{\omega}$ by maximum likelihood.

Specifically, let $\Omega \equiv \Re^4_{++} \times [0, 1]$ be the space of possible values for $\boldsymbol{\omega}$, and consider $\ddot{\omega} \in \Omega$. From $\hat{s}(\cdot)$ and (5.10), I numerically obtain $\ddot{\theta}(\cdot; \ddot{\omega})$, the function $\tilde{\theta}(\cdot)$ consistent with $\ddot{\omega}$. Using (5.11) and (5.12), I then obtain $\ddot{f}(\cdot|\ddot{\omega})$, the density function $f(\cdot)$ consistent with $\ddot{\omega}$. Similarly, from (5.13) and the estimated density $\hat{b}(\cdot|\Psi = 1, Z = z)$, I numerically compute $\ddot{g}(\cdot|Z = z; \ddot{\omega})$, the density $g(\cdot|Z=z)$ consistent with $\ddot{\omega}$. Using $\ddot{f}(\cdot|\ddot{\omega})$ and $\ddot{g}(\cdot|Z = z; \ddot{\omega})$, I obtain the likelihood that $\Psi = 3$, given Z, and consistently with $\ddot{\omega}$. Such likelihood is

$$\ddot{P}[\Psi = 3|Z = z; \ddot{\omega}] = \int_{[\hat{\underline{t}}, \hat{\bar{t}}]} \int_{\underline{\theta}}^{\ddot{\theta}(t;\ddot{\omega})} (1 - x)\ddot{f}(x|\ddot{\omega})\ddot{g}(t|Z = z; \ddot{\omega})\, dx\, dt.$$

From (5.4), I can compute the likelihood that $\Psi = 1$, given Z, and consistently with $\ddot{\omega}$. This likelihood is given by

$$\ddot{P}[\Psi = 1|Z = z; \ddot{\omega}] = \int_{[\hat{\underline{t}}, \hat{\bar{t}}]} 1 - \ddot{F}\big[\ddot{\theta}(t; \ddot{\omega})|\ddot{\omega}\big]\ddot{g}(t|Z = z; \ddot{\omega})\, dt,$$

where $\ddot{F}[\cdot|\ddot{\omega}]$ is a CDF obtained from $\ddot{f}(\cdot|\ddot{\omega})$. From (5.7) and (5.8), the likelihood that $T = t$ and $\Psi = 2$, given Z, and consistently with $\ddot{\omega}$, is

$$\ddot{P}[\Psi = 2|Z = z; \ddot{\omega}]\ddot{g}(t|\Psi = 2, Z = z; \ddot{\omega}) = \int_{\underline{\theta}}^{\ddot{\theta}(t;\ddot{\omega})} x\ddot{f}(x|\ddot{\omega})\, dx\ddot{g}(t|Z = z; \ddot{\omega}).$$

[46] The system also implicitly contains $\pi \equiv \int_{\underline{\theta}}^{\tilde{\theta}(\underline{t})} xf(x)\, dx$. Like μ, the parameter π depends on the behavior of $F(\cdot)$ outside of the range of the support over which this function is identified. In estimating the model, I set $\pi = 0$. My empirical results suggest that μ is very close to 1. Since $\pi \leq 1 - \mu$, setting $\pi = 0$ is unlikely to make any practical difference in the estimation.

 DOI: 10.3982/ECTA12974

I am ready to define an observation's likelihood contribution. I consider the likelihood, conditional on $\Psi \neq 0$.[47] Let W_i be the data corresponding to observation i.[48] Given $\ddot{\omega}$, the likelihood contribution of an observation i, conditional on $\Psi_i \neq 0$, is

$$\begin{aligned} l(\ddot{\omega}, W_i) = {} & \ddot{P}[\Psi = 1 | Z = z; \ddot{\omega}]^{\mathbb{1}\{\Psi_i = 1\}} \\ & \times \left\{ \ddot{P}[\Psi = 2 | Z = z; \ddot{\omega}] \ddot{g}(t_i | \Psi = 2, Z = z; \ddot{\omega}) \right\}^{\mathbb{1}\{\Psi_i = 2\}} \\ & \times \ddot{P}[\Psi = 3 | Z = z; \ddot{\omega}]^{\mathbb{1}\{\Psi_i = 3\}}. \end{aligned} \tag{B.1}$$

I obtain an estimate $\hat{\omega}$ for ω by performing a numerical search to find the parameters that maximize the sum of the logarithms of $l(\ddot{\omega}, W_i)$ over all observations for which $\psi \neq 0$.[49] Finally, estimates for $g(\cdot | Z = z)$ and $f(\cdot)$ are defined by $\ddot{g}(\cdot | Z = z; \hat{\omega})$ and $\ddot{f}(\cdot | Z = z; \hat{\omega})$, respectively.

B.2. *Observed Heterogeneity*

I divide the observations in the data into a finite number of covariate groups and implement the estimator described in Section 5 separately for each one of them. The first step of the estimator consists of computing two types of conditional densities: that of trial sentences, conditional on a conviction at trial, and that of settlement offers, conditional on a plea bargain. These conditional densities must be estimated for cases under the responsibility of both lenient and harsh judges. Therefore, for each one of the covariate groups under consideration in my analysis, I must estimate four conditional densities. I use the smoothing method by Li and Racine (2007), which I briefly describe below. Notice that the notation employed in this part of the Supplemental Material differs from that of the rest of the paper.

Let Y be a univariate continuous random variable and X an r-dimensional discrete random variable. Denote by $f(\cdot)$, $g(\cdot)$, and $\mu(\cdot)$ the joint density of (X, Y) and the marginal densities of Y and X, respectively. For each dimension s of X, let c_s be the number of values in the support of X_s and λ_s be a real number between zero and $(c_s - 1)/c_s$. Define the vector $\lambda = (\lambda_1, \ldots, \lambda_r)$ and consider the following estimators of $f(\cdot)$ and $\mu(\cdot)$:

$$\hat{f}(x, y) = n^{-1} \sum_{i=1}^{n} L(x, X_i, \lambda) k_{h_0}(y - Y_i) \quad \text{and}$$

$$\hat{\mu}(x) = n^{-1} \sum_{i=1}^{n} L(x, X_i \lambda),$$

where n is the sample size, $k_{h_0}(\cdot)$ is a kernel function with bandwidth h_0, and

$$L(x, X_i \lambda) = \prod_{s=1}^{r} \left[\lambda_s / (c_s - 1) \right]^{\mathbb{1}(X_{is} \neq x_s)} (1 - \lambda_s)^{\mathbb{1}(X_{is} = x_s)}.$$

[47] Notice that the empirical probability that $\Psi = 0$ is useful only for identifying $\nu(z)$.

[48] That is, if $\Psi_i \in \{1, 3\}$, W_i consists of z_i and ψ_i, the realizations of Z_i and Ψ_i. If $\Psi_i = 2$, W_i also includes t_i, the realization of T_i. Notice that I do not take into account the realization s_i of S_i, which is observed when $\Psi_i = 1$. That is because the likelihood of $S = s$, given $\Psi = 1$ and $Z = z$, is simply $\hat{b}(z | \Psi = 1, Z = z)$, which does not depend on $\ddot{\omega}$.

[49] I constrain $\ddot{\alpha}_d$ and $\ddot{\beta}_d$ to satisfy the conditions in footnote 32.

Finally, define the estimator of the conditional density $g(y|x)$ as

$$\hat{g}(y|x) = \hat{f}(x, y)/\hat{\mu}(x).$$

Notice that $\hat{g}(y|x)$ is obtained using all observations in the data—even those in which $X \neq x$. These observations are weighted down, relative to the ones satisfying $X = x$. The weights are given by the vector $\lambda = (\lambda_1, \ldots, \lambda_r)$. In one extreme case, λ_s is zero for all s, and $\hat{g}(y|x)$ is calculated employing only observations such that the realization of X is x. In the other extreme case, $\lambda_s = (c_s - 1)/c_s$ for all s, and $\hat{g}(y|x)$ becomes the estimate of $g(\cdot)$, the unconditional density of Y. The vector λ can be regarded as a collection of smoothing parameters—one for each dimension of X. Together, λ and h_0 determine the extent to which points away from (y, x) affect $\hat{g}(y|x)$. As argued by Li and Racine (2007), positive values of λ increase the finite sample bias of $\hat{g}(y|x)$ but also reduce its variance, with an ambiguous effect on the mean squared error.

The greatest challenge in implementing this estimator, therefore, is the choice of the smoothing parameters λ and h_0. In my application, I follow Li and Racine (2007) and select λ by maximum likelihood cross-validation. For any given sample size and any covariate dimension c, this method aims to select relatively large values of λ_c if the distribution of Y is not largely affected by variations in X_c, and small values of λ_c if the distribution of Y varies considerably with X_c. Moreover, the selected values of λ_c tend to decrease as the sample size increases.

For each covariate group, I estimate four conditional densities. Using the notation of Li and Racine's estimator presented above, Y may represent four random variables: trial sentences assigned by lenient judges, trial sentences assigned by harsh judges, settlement offers made under lenient judges, and settlement offers made under harsh judges. The discrete random variable X refers to the covariates used to divide the data into groups.[50] This random variable has the following five dimensions: (i) defendant's gender (male or female), (ii) defendant's race (African-American or non-African-American), (iii) the type of defense counsel (public defender, court-assigned attorney, or privately-held attorney), (iv) the length of the defendant's criminal record (short or long, as defined in Section 6), and (v) Superior Court division (numbers one to eight). The function $k_{h_0}(\cdot)$ is the Epanechnikov kernel.

Table X contains the smoothing parameters λ obtained by maximum likelihood cross-validation for each of the four conditional densities of my analysis. Notice that, for every covariate c, lambda must belong to the interval $[0, (c_s - 1)/c_s]$, where c_s is the covariate's support. The upper endpoints of this interval are shown in the last column of the table. All the selected smoothing parameters are far away from these endpoints, suggesting that the covariates under consideration are important in explaining the distributions of trial sentences and settlement offers. In particular, the smoothing parameters associated with the defendant's previous criminal record are very close to zero. The parameters associated with race are also relatively low—ranging from 0.05 to 0.12. The gender parameters are larger for the densities of trial sentences than for those of settlement offers, which can be explained by the larger sample sizes used to compute the latter.

As explained in Section 5, the supports of trial sentences and settlement offers are bounded, which complicates the estimation of the conditional densities described above.

[50]To be sure, I estimate four conditional densities. The densities of trial sentences are conditional on a conviction at trial, and those of settlement offers are conditional on a plea bargain. Besides conditioning on the case outcome, I estimate these densities conditioning on five covariates. In the notation of this supplement, X refers only to these covariates.

TABLE X

CONDITIONAL DENSITY ESTIMATORS—COVARIATES' SMOOTHING PARAMETERS[a]

	Conditional Density Estimator				
	Trial Sentences		Offers		
Covariate	Lenient	Harsh	Lenient	Harsh	Upper Endpoint
Gender	0.39	0.11	0.03	0.03	0.50
Race	0.12	0.05	0.09	0.09	0.50
Counsel	0.23	0.18	0.20	0.25	0.67
Record	0.02	0.03	0.01	0.00	0.50
Division	0.41	0.33	0.44	0.41	0.88

[a]Covariates' smoothing parameters selected by maximum likelihood cross-validation. The parameters are used for smoothing across covariate groups in the kernel estimation of conditional densities of trial sentences and settlement offers.

I use a boundary correction proposed by Karunamuni and Zhang (2008). Using the notation of this Supplemental Material, the approach consists of reflecting a transformation of the data near the boundary of Y. The reflected data points have the same x as the corresponding observations in the original data set, but y is modified. The estimator uses separate bandwidths h_0 for points near the boundary and away from it. Differently from the naive reflection of the untransformed data, this method allows the partial derivative of $g(y|x)$ with respect to y to be different from zero at the boundary of the support. See Karunamuni and Zhang (2008) for details.

Table XI reports the bandwidths h_0 for points away from the boundary, which are computed using Silverman's "rule-of-thumb" (Silverman (1986)). The bandwidths for trial sentences are 21.83 months (lenient judges) and 25.31 months (harsh judges). Those for settlement offers are 6.59 months (lenient judges) and 7.37 months (harsh judges). The larger bandwidths for trial sentences reflect the relative scarcity of cases that result in an incarceration conviction at trial.

B.3. *Standard Errors*

I use 1200 bootstrap samples for each group to compute standard errors for the parameters reported in Table VI. For each such sample, I estimate the densities of trial sentences and settlement offers using the same bandwidths and smoothing parameters employed in the main data. There are two main issues with this procedure. First, I do not offer a proof of the validity of the bootstrap for my estimator. Subsampling methods (Politis, Romano, and Wolf (1999)) are more robust than the bootstrap, but, to apply these methods, the

TABLE XI

CONDITIONAL DENSITY ESTIMATORS—TRIAL SENTENCES AND SETTLEMENT OFFERS' BANDWIDTHS[a]

	Trial Sentences		Settlement Offers	
	Lenient	Harsh	Lenient	Harsh
Bandwidth	21.83	25.31	6.59	7.37

[a]Bandwidths selected by Silverman's "rule-of-thumb" (Silverman (1986)). Measured in months.

convergence rate of the estimator must be known. The second issue is that, for part of the bootstrap samples, the last step of the estimation procedure—that is, obtaining maximum likelihood estimates for α_d, β_d, α_p, β_p, and μ—becomes computationally too costly. This is the case whenever the estimated settlement offer function is too convex. I do not implement the last estimation step for these samples.[51] Thus, the standard deviations reported in Table VI may slightly overstate the actual precision of my estimator.

REFERENCES

KARUNAMUNI, R. J., AND S. ZHANG (2008): "Some Improvements on a Boundary Corrected Kernel Density Estimator," *Statistics & Probability Letters*, 78, 499–507. [4]

LI, Q., AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice* (First Ed.). Princeton, NJ: Princeton University Press. [2,3]

POLITIS, D. N., J. P. ROMANO, AND M. WOLF (1999): *Subsampling* (First Ed.). New York: Springer. [4]

SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall. [4]

Olin Business School, Washington University in St. Louis, One Brookings Drive, Campus Box 1133, St. Louis, MO 63130, U.S.A.; silveira@wustl.edu.

Co-editor Liran Einav handled this manuscript.

Manuscript received 10 November, 2014; final version accepted 22 September, 2016; available online 31 October, 2016.

[51] More precisely, I drop from my analysis every bootstrap sample in which the coefficient associated with the last C-spline basis is greater than 4. As a reference, using the main data, I estimate this coefficient to be 2.43 for covariate group one and 1.27 for group two. This procedure eliminates 13.83% and 32.50% of the 1200 bootstrap samples for groups one and two, respectively.