Online Appendix B to Measuring Segmentation in the Financial News Market: Model Checking using Simulated Data

November 4, 2019

To provide more intuition on how well the model estimates the three unobserved variables of interest—even in the presence of a selection mechanism—I simulate a hundred samples of a simplified structure, fit the latent model to them, and perform regressions using the true and the estimated audience and EA characteristics. For each sample, I simulate the following simplified version of equation (1) in the main paper:

$$\textit{Tone}_{i,j} = \textit{Room}_j \times \textit{Outl}_i + \textit{News}_j \tag{1}$$

The three latent variables are simulated as follows: Each sample has 100 outlets, for which I randomly choose a *positive* or *negative* audience. If an outlet has a positive audience it has a 0.5 in outlet-specific slant ($\textit{Outl}_i$). Likewise, if an outlet has a negative audience, it has -0.5 in outlet-specific slant ($\textit{Outl}_i$). Each sample has 200 EAs, for which I first randomly assign a certain *Size* (uniformly distributed between 0 and 1). *Size* is meant as a catch-all term for other determinants that can influence coverage selection and that are not part of the Bayesian model equation. $\textit{News}_j$ is then computed as $X_j + 0.4 \times \textit{Size}_j$ where $X_j \sim U(-1, 1)$. To induce some correlation between room for interpretation and economic news, I compute $\textit{Room}_j$ as $Z_j + 0.3 \times |\textit{News}_j|$, with $Z_j \sim U(0, 1)$ and rescaling the result to lie between 0 and 1 for all 200 EAs.

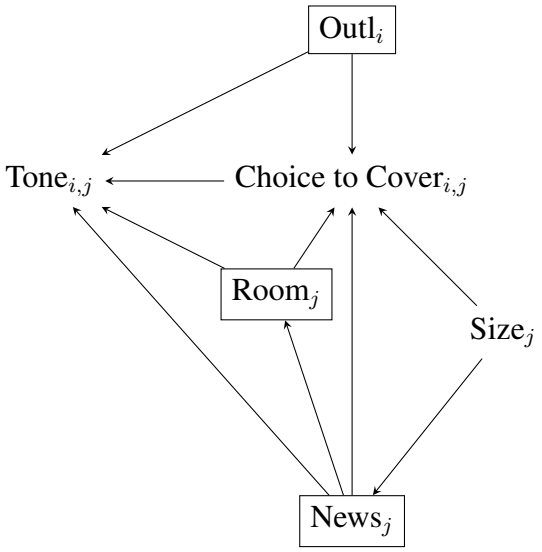
Next, for each outlet and each EA, I compute the probability that the outlet covers the EA using

a simple logistic regression, similar to the form used in the main tests:

$$\begin{aligned} \textit{Covered}_{i,j} &\sim \text{Bin}(1, p_{i,j}) \\ \text{logit}(p_{i,j}) &= -2 + \textit{Size}_j + \textit{Room}_j + |\textit{News}_j| - \textit{Outl}_i \cdot \textit{Room}_j \end{aligned} \tag{2}$$

This setup serves two purposes. The first is to show that even though the decision to cover an EA depends on a determinant (*Size*) that is not in the measurement equation, this endogeneous selection does not affect inferences in the tone equation as long as *News* is measured. Essentially *Room* and *News* are EA-fixed effects and will block all endogeneous selection. This can also be seen in the following causal graph.

Figure 1: Simulation Setup



In this simulation *Size* affects *Tone* via the choice to cover and via its impact on the amount of economic news. Since both are controlled for in the tone model (choice to cover equals 1 and *News* is estimated), there is no free path through which *Size* can still influence *Tone* and thus cause a correlated omitted variable bias. Said differently, conditional on coverage choice, room for interpretation, and *News*, *Size* does not affect tone.

Of course, the issue is that except for Size and Tone, none of the variables are observable. They have to be estimated. To do so I use the latent model developed in Sec. 3. I do this for every sample and compare the estimated variables with the true ones. Table 1 shows descriptive

comparisons between the true variables and the estimated versions using the Bayesian model for all 100 samples. The Panel A shows that the estimated variables seem to always capture the true variables well (demonstrated by correlation coefficients close to one). Panel B and C further show that the estimated economic news closely track the true values as well.

Comparing true and estimated averages and standard deviations for outlet characteristics and room for interpretation shows the impact of the identifying assumption. By assumption, the model scales the standard deviation of the outlet characteristics to be one, instead of the true 0.5. To compensate, the variation in estimated room for interpretation is scaled by the model to be half of what it's true value is. But, even though the true variation of the components of the product cannot be recovered, this is not needed for the purposes of this study. To make statements about the magnitude of segmentation, we need to compare it to the variation in undisputed news $News_j$. For this comparison, the outlet characteristics need to be scaled by the average room for interpretation, which "fixes" the scale.

A second important point is that, for purposes of examining the role of room for interpretation or heterogeneity in outlet characteristics, the scale is also not crucial. This is demonstrated in Panel A of Table 1 and Table 2. As the correlations in Panel A show, the model predicts which outlet is positive and which is negative with perfect accuracy and the estimates are in general almost perfectly correlated with the true value. In addition, the first two columns of Table 2 report average coefficients and standard deviation of the coefficients from 100 coverage choice regressions of the form of Eq. 2. Column 1 shows results using the true values for *Frame*, *News*, and *Room*, while column 2 uses the estimated variable values from the Bayesian model. The only difference in coefficients is the twice as large coefficient on *Room*, which is due to the scale of the estimate version of *Room* being half of the true one. This difference disappears, once all variables are standardized to have standard deviation one.

A third important aspect is illustrated in columns 3 to 7 in Table 2. Including or Excluding *Size* into the Tone regression (or the estimation) does not matter, even though *Size* is an endogenous coverage determinant. Comparing column 3 and column 4 shows that *Size*, even though being a coverage determinant is superfluous. Column 5 shows that not including *News* leads to distortions however. The results serve to highlight that a selection mechanism (e.g., coverage selection based on size) can be ignored as long as undisputed news are controlled for (e.g., via earnings announcement fixed effects). Column 6 to 7 illustrate that this also holds when using the estimated versions of *Room*, *News*, and *Outl*. The average coefficients across all 100 simulated samples are essentially

identical to the ones in column 3 and 4.

Table 1: Measurement model results based on simulated data

Variable	N	Mean	StD	Min	P05	P25	Med	P75	P95	Max
Panel A: True vs estimated average magnitude of framing										
Subset of negative outlets:										
True Frame	100	-0.50	0.00	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50	-0.50
Estimated Frame	100	-1.00	0.09	-1.27	-1.16	-1.05	-0.98	-0.94	-0.87	-0.73
Subset of positive outlets:										
True Frame	100	0.50	0.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Estimated Frame	100	1.00	0.10	0.78	0.85	0.94	0.99	1.05	1.14	1.36
Panel B: Pearson correlation true vs estimated variables										
News	100	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Frame	100	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Room	100	0.99	0.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Panel C: True vs estimated standard deviation of variables										
True News	100	0.59	0.02	0.55	0.56	0.58	0.59	0.60	0.62	0.65
Estimated News	100	0.59	0.02	0.55	0.56	0.58	0.59	0.60	0.62	0.65
True Frame	100	0.50	0.00	0.48	0.49	0.50	0.50	0.50	0.50	0.50
Estimated Frame	100	1.00	0.01	0.98	0.99	1.00	1.00	1.00	1.01	1.01
True Room	100	0.24	0.01	0.21	0.22	0.23	0.24	0.25	0.26	0.27
Estimated Room	100	0.12	0.01	0.10	0.11	0.12	0.12	0.12	0.13	0.13

Table 1 shows descriptive statistics from fitting the Bayesian model to 100 simulated samples of fake data. Each sample consists of 100 outlets and 200 EAs. The fake tone data has the form: $Tone_{i,j} = RoomInt_j \times Frame_i + EconNews_j$ where i denotes the outlet and j denotes the EA. *News* is the undisputed economic news, *Room* is the room for interpretation in news, and *Frame* is the amount of framing (towards good or bad tone) that an outlet adds.

Table 2: Measurement model results based on simulated data

	True	*Estimated*	*True*	*True*	*True*	*Estimated*	*Estimated*
	Pr(Covered)	Pr(Covered)	Tone	Tone	Tone	Tone	Tone
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(Intercept)	−2.009	−2.003	−0.000	0.000	0.539	−0.001	−0.001
	(0.058)	(0.062)	(0.003)	(0.003)	(0.114)	(0.000)	(0.000)
|News|	1.004	0.991					
	(0.047)	(0.048)					
Frame	−0.005	0.000	−0.001	−0.001	0.016	−0.008	−0.008
	(0.077)	(0.039)	(0.005)	(0.005)	(0.030)	(0.001)	(0.001)
Room	1.016	2.021	−0.000	−0.000	−0.515	0.005	0.005
	(0.066)	(0.183)	(0.004)	(0.004)	(0.184)	(0.001)	(0.001)
Room:Frame	−0.992	−0.995	1.001	1.001	0.999	1.024	1.024
	(0.129)	(0.132)	(0.009)	(0.009)	(0.045)	(0.003)	(0.003)
Size	0.999	0.998	0.001			−0.000	
	(0.053)	(0.055)	(0.004)			(0.000)	
News			1.000	1.000		1.001	1.001
			(0.002)	(0.002)		(0.000)	(0.000)

Table 2 shows average coefficients from ols regressions to 100 simulated samples of fake data. Each sample consists of 100 outlets and 200 EAs. *News* is the undisputed economic news, *Room* is the room for interpretation in news, and *Frame* is the amount of framing (towards good or bad tone) that an outlet adds, *Size* is firm size, used as a catch-all for endogenous coverage determinants.