HOW WELL DO SELECTION MODELS PERFORM? ASSESSING THE ACURACY OF ART AUCTION PRE-SALE ESTIMATES

Binbing Yu and Joseph L. Gastwirth

National Institute of Health and The George Washington University

This note presents the supplementary materials for the analysis of art auction data.

S1. Percentages of final bids falling below, within and above the predicted intervals

Let $G = P - L$ be the half range of the predicted interval $[L, U]$. In Table S1.1 we show the percentages of items whose highest bids were below, within or above the interval $[P - d \times G, P + d \times G]$, where d is a multiplier increasing the width of the original predicted interval $[L, U]$. From the first line ($d = 1$) corresponding to the original prediction interval, one observes that the only 20.4%-32.7% and 33.6%-48.2% of the highest bids fall within the predicted interval for all items and the sold items, respectively. This suggests that the auctioneers under-estimate the variability of the bids. Even when $d = 1.75$, which nearly doubles the width of the prediction interval, the percentages of highest bids for all items and for the sold items remain below 50% and 64%, respectively. This suggests that the prediction errors have a "heavy" tail and the selection model should be modified appropriately.

S2. Calculation of the response probability in the selection models

The Newton-Raphson method was used to obtain the maximum likelihood estimates of the parameters (θ, ψ) for the loglikelihood function of the selection model. We present the calculation of response probability $P(S_i = 1|X_i)$ for normal selection model and t_ν selection model. Here we let $\beta = (\beta_0, \beta_1), \gamma = (\gamma_0, \gamma_1)$ and $\mathbf{X}_i = (1, X_i)^T$.

S2.1. Response probability $P(S_i = 1|X_i)$ for normal and t_ν selection models

Table S1.1: The percentages of the highest bids below, within and above the predicted interval by different inflation factors d

		All items	Only sold items
Sale #	Factor d	(Below, Within, Above)	(Below, Within, Above)
3850	1.00	(45.1, **32.7**, 22.2)	(19.0, **48.2**, 32.7)
	1.25	(41.7, **36.6**, 21.7)	(16.1, **52.0**, 31.9)
	1.75	(30.7, **49.6**, 19.7)	(7.3, **63.7**, 28.9)
6371	1.00	(66.1, **20.4**, 13.4)	(43.6, **33.6**, 22.7)
	1.25	(62.9, **23.7**, 13.4)	(38.2, **39.1**, 22.7)
	1.75	(52.7, **35.5**, 11.8)	(26.4, **53.6**, 20.0)
8990	1.00	(43.2, **33.1**, 23.7)	(32.9, **39.1**, 28.0)
	1.25	(40.7, **35.5**, 23.7)	(30.6, **41.3**, 28.0)
	1.75	(32.2, **48.0**, 19.8)	(21.4, **55.3**, 23.3)
9028	1.00	(54.6, **31.7**, 13.7)	(40.8, **41.4**, 17.8)
	1.25	(54.1, **32.7**, 13.2)	(40.8, **42.0**, 17.2)
	1.75	(40.0, **49.8**, 10.2)	(24.8, **61.8**, 13.4)
9038	1.00	(36.3, **34.5**, 29.2)	(26.0, **40.0**, 34.0)
	1.25	(34.8, **36.3**, 28.9)	(24.4, **42.0**, 33.6)
	1.75	(25.7, **49.7**, 24.5)	(15.6, **55.9**, 28.5)

Lemma S2.1. In normal selection models, the probability of response ($S_i = 1$) given X_i is,

$$P(S_i = 1|X_i) = \Phi\Big(\frac{(\gamma+\delta\beta)^T\mathbf{X}_i}{\sqrt{1+(\delta\sigma)^2}}\Big) \tag{S2.1}$$

Proof of Lemma S2.1. In normal selection model, $P(S_i = 1|X_i) = \Phi(\alpha^T\mathbf{X}_i)$. According to the reparametrization following Equation (3.3) in the paper, $\alpha = (\gamma+\delta\beta)\sqrt{1-\rho^2} = (\gamma+\beta\delta)/\sqrt{1+(\delta\sigma)^2}$, so

$$P(S_i = 1|X_i) = \int P(S_i = 1|X_i, y)\phi(y|\beta^T\mathbf{X}_i, \sigma^2)dy = \Phi\Big(\frac{(\gamma+\delta\beta)^T\mathbf{X}_i}{\sqrt{1+(\delta\sigma)^2}}\Big).$$

Lemma S2.2 For the selection model with a t_ν error distribution, the probability of response is

$$P(S_i = 1|X_i) = \int_0^\infty \Phi\Big(\frac{(\gamma+\delta\beta)^T\mathbf{X}_i}{\sqrt{1+\nu(\delta\sigma)^2/(2z)}}\Big)\frac{z^{\frac{\nu-2}{2}}\exp(-z)}{\Gamma(\frac{\nu}{2})}dz. \tag{S2.2}$$

Proof of Lemma S2.2 Because t_ν distribution is a mixture of a normal distribution and inverse χ^2 distribution (Box and Tiao, 1973, eq. 2.7.21), i.e.,

$$f_t(y|\mu,\sigma^2;\nu) = \int_0^\infty \phi(y|\mu,\sigma^2/u)f_\nu(u)du,$$

where $\phi(y|\mu,\sigma^2/u)$ is the density of a normal distribution and $f_\nu(u) = \frac{\nu(\nu u)^{\frac{\nu-2}{2}}\exp(-\frac{\nu u}{2})}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}$, the response probability can be written as

$$\begin{aligned} P(S_i = 1|X_i;\theta,\psi) &= \int_{-\infty}^\infty P(S_i = 1|X_i, y)f_t(y|\beta^T\mathbf{X}_i, \sigma^2;\nu)dy \\ &= \int_{-\infty}^\infty\int_0^\infty P(S_i = 1|X_i, y)\phi(y|\beta^T\mathbf{X}_i, \sigma^2/u)f_\nu(u)dudy \end{aligned}$$

By interchanging the order of integration, according to **Lemma S2.1**, this is equivalent to

$$\int_0^\infty \Phi\Big(\frac{(\gamma+\delta\beta)^T\mathbf{X}_i}{\sqrt{1+(\delta\sigma)^2/u}}\Big)f_\nu(u)du$$

Letting $u = \frac{2z}{u}$ in $f_\nu(u)$, we obtain Equation (S2.2). Note that Equation (S2.2) can be alternatively expressed by the cdf a Student's t distribution (Lemma 1 of Azzalini and Capitaino (2003), p. 380).

S2.2. Approximation of response probability for t_ν selection models

For selection models using t_ν distribution, the response probability (S2.2) can be approximated using Gauss-Laguerre Integration (Abramowitz and Stegun, 1964),

$$\int_0^\infty \exp(-z)g(z)dz \approx \omega_k g(z_k)$$

where $\{\omega_k, k = 1..n\}$ and $\{z_k, k = 1..n\}$ are the weights and abscissas of a n points approximation. Hence,

$$P(S_i = 1|X_i) \approx \sum_{k=1}^{n} \Phi\Big(\frac{(\gamma + \delta\beta)^T \mathbf{X}_i}{\sqrt{1 + \nu(\delta\sigma)^2/(2z_k)}}\Big)\frac{\omega_k z_k^{\frac{\nu-2}{2}}}{\Gamma(\frac{\nu}{2})}.$$

S2.3. Prediction of the final bids of the unsold items

For the normal selection model (see **Lemma S2.1**),

$$P(S_i = 0) = 1 - \Phi\Big\{\frac{(\gamma_0 + \delta\beta_0) + (\gamma_1 + \delta\beta_1)X_i}{\sqrt{1 + (\delta\sigma)^2}}\Big\},$$

and

$$E\{A_i I(S_i = 0)\} = \exp(\beta_0 + \beta_1 X_i + \frac{\sigma^2}{2}) \times \Big[1 - \Phi\Big\{\frac{\gamma_0 + \delta(\beta_0 + \sigma^2) + (\gamma_1 + \delta\beta_1)X_i}{\sqrt{1 + (\delta\sigma)^2}}\Big\}\Big].$$

For the selection model when the errors follow the t_2 distribution, the imputed value can be evaluated numerically (see **Lemma S2.2**).

References

Abramowitz, M., and Stegun, I.A. (1964). *Handbook of Mathematical Functions*, Applied Mathematics Series, Volume 55 (Washington: National Bureau of Standards; reprinted 1968 by Dover Publications, New York).

Azzalini, A., and Capitaino, A. (2003). Distribution generated by perturbation of symmetry with emphasis on a multivariate skewed t distribution. *J. Roy. Statist. Soc. Ser. B* **65,** 367–389.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis,* Reading, Mass: Addison-Wesley.

Laboratory of Epidemiology, Demography and Biometry
National Institute on Aging
National Institutes of Health, Bethesda, MD 20904, U.S.A.

E-mail: yubi@mail.nih.gov

Department of Statistics

The George Washington University, Washington DC 20052, U.S.A.

E-mail: jlgast@gwu.edu