C Approximation Error: Normal Versus Log-normal

When computing subjective income distributions using either normal or log-normal distributions, we have only used data on the median (C^3_{it}) and the difference between first and third quartiles ($C^4_{it} - C^2_{it}$ or C^4_{it}/C^2_{it}). Hence, for either the normal and log-normal distributions, the three quartiles reported in the data (C^2_{it}, C^3_{it}, C^4_{it}) will not partition the support of the subjective income distribution into four segments that each have a probability of .25, unless the distributional assumption is exactly correct. Therefore, we evaluate the validity of a particular distributional assumption using the loss function:

$$AE(D) = \frac{1}{N}\sum_{i=1}^{N}[(F(C^3_{it};D) - F(C^2_{it};D) - 0.25)^2 + (F(C^4_{it};D) - F(C^3_{it};D) - 0.25)^2], \quad (24)$$

where $F(w;D)$ is the cdf of the distribution computed using distributional assumption D.

Using the same sample as in Section 3, we compute the value of $AE(D)$ for $D =$ normal and $D =$ log-normal. We find that $AE(\text{normal}) = 0.0101$ and $AE(\text{log-normal}) = 0.0103$. Hence, we conclude that the fit of the two distributions is quite similar with, if anything, the normal having a slightly better fit.

D Expressing $E(W_{it})$ as a weighted sum of $E(W_{it}|G_{it} = 2.00)$, $E(W_{it}|G_{it} = 3.00)$, and $E(W_{it}|G_{it} = 3.75)$

We show that $E(W_{it})$ can be expressed as a weighted sum of $E(W_{it}|G_{it} = 2.00)$, $E(W_{it}|G_{it} = 3.00)$, and $E(W_{it}|G_{it} = 3.75)$. For the ease of notation, we write $E(W_{it}|G_{it} = g_{it})$ as $E(W_{it}|g_{it})$. Hence,

$$\begin{aligned}
E(W_{it}) &= E_{G_{it}}(E(W_{it}|G_{it})) = \int_2^4 E(W_{it}|g_{it})dF_{G_{it}}(g_{it}) \\
&= \int_2^3 [E(W_{it}|2.00) + \frac{E(W_{it}|3.00) - E(W_{it}|2.00)}{3.00 - 2.00}(g_{it} - 2)]dF_{G_{it}}(g_{it}) \\
&+ \int_3^4 [E(W_{it}|3.00) + \frac{E(W_{it}|3.75) - E(W_{it}|3.00)}{3.75 - 3.00}(g_{it} - 3)]dF_{G_{it}}(g_{it}) \\
&= \int_2^3 [E(W_{it}|2.00)(1 - \frac{g_{it} - 2}{3.00 - 2.00}) + E(W_{it}|3.00)\frac{g_{it} - 2}{3.00 - 2.00}]dF_{G_{it}}(g_{it}) \\
&+ \int_3^4 [E(W_{it}|3.00)(1 - \frac{g_{it} - 3}{3.75 - 3.00}) + E(W_{it}|3.75)\frac{g_{it} - 3}{3.75 - 3.00}]dF_{G_{it}}(g_{it}) \\
&= \sum_G \lambda^G_{it} E(W_{it}|G) \qquad\qquad G = 2.00,\ 3.00 \text{ or } 3.75, \qquad (25)
\end{aligned}$$

where $\lambda_i^{2.00} = \int_2^3 (3 - g_{it}) dF_{G_{it}}(g_{it})$, $\lambda_i^{3.00} = \int_2^3 (g_{it} - 2) dF_{G_{it}}(g_{it}) + \int_3^4 (1 - \frac{g_{it}-3}{0.75}) dF_{G_{it}}(g_{it})$ and $\lambda_i^{3.75} = \int_3^4 \frac{g_{it}-3}{0.75} dF_{G_{it}}(g_{it})$.

E Magnitude of the Measurement Error

In this section, we show that equation (12), along with additional assumptions, implies equation (13). Recall that equation (12) states:

$$\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it}) = \varsigma_i - \sum_{g_{it}} \lambda_i^{g_{it}} \varsigma_i^{g_{it}}. \qquad \text{(12 revisited)}$$

Taking the variance of both sides, we have:

$$\begin{aligned}
var(\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it})) &= var(\varsigma_i - \sum_{g_{it}} \lambda_i^{g_{it}} \varsigma_i^{g_{it}}) \\
&= var(\varsigma_i) + \sum_{g_{it}} var(\lambda_i^{g_{it}} \varsigma_i^{g_{it}}) && \text{(independence of MEs)} \\
&= var(\varsigma_i) + \sum_{g_{it}} E((\lambda_i^{g_{it}})^2) E((\varsigma_i^{g_{it}})^2) - (E(\lambda_i^{g_{it}}) E(\varsigma_i^{g_{it}}))^2 \\
& && (\lambda_i^{g_{it}} \perp\!\!\!\perp \varsigma_i^{g_{it}}) \\
&= var(\varsigma_i) + \sum_{g_{it}} E((\lambda_i^{g_{it}})^2) var(\varsigma_i^{g_{it}}) \\
& && (E(\varsigma_i) = 0 \text{ and } E(\varsigma_i^{g_{it}}) = 0) \\
&= var(\varsigma_i)[1 + \sum_{g_{it}} E((\lambda_i^{g_{it}})^2)]. && (var(\varsigma_i) = var(\varsigma_i^{g_{it}}))
\end{aligned}$$

Therefore,

$$var(\varsigma_i) = \frac{var(\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it}))}{1 + \sum_{g_{it}} E((\lambda_i^{g_{it}})^2)}. \qquad \text{(13 revisited)}$$

F Taking into Account Interpolation Errors

In Section 3.2.2, we note that interpolation error could be introduced into our computations because it is necessary to interpolate the means of subjective income distributions conditional on values of GPA other than 2.00, 3.00 or 3.75. In addition, errors can be introduced because it is necessary to compute distributions of final GPA from data. In this appendix, we show that taking into account these errors would lead to a smaller value of $var(\varsigma_i)$, implying a larger estimate of our measure of true heterogeneity.

We start by describing how we incorporate both types of errors into our analysis. With respect to the potential error introduced during the computation of the distribution of final GPA, we denote $F_{G_{it}}(g_{it})$ and $\widetilde{F}_{G_{it}}(g_{it})$ as the true CDF and the computed CDF of

G_{it}, respectively. We allow the CDFs to potentially differ from each other and denote the difference as $F^{\Delta}_{G_{it}}(g_{it}) = \widetilde{F}_{G_{it}}(g_{it}) - F_{G_{it}}(g_{it})$.

For ease of notation, we denote a vector that includes $(E(W_{it}|G_{it} = 2.00), E(W_{it}|G_{it} = 3.00), E(W_{it}|G_{it} = 3.75))$ as $\mathbf{E^W_{G_{it}}}$, and a vector that includes $(\widetilde{E}(W_{it}|G_{it} = 2.00), \widetilde{E}(W_{it}|G_{it} = 3.00), \widetilde{E}(W_{it}|G_{it} = 3.75))$ as $\widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}$. The interpolation approach that we use to compute the mean of subjective income distributions conditional on values of GPA other than 2.00, 3.00, or 3.75 is essentially a mapping from $\widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}$ to $\widetilde{E}(W_{it}|G_{it} = g_{it})$, $g_{it} \neq 2.00, 3.00, 3.75$. We denote this mapping as $\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}})$. Note that the difference between the computed value of the conditional mean, $\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}})$, and the true value of conditional mean, $E(W_{it}|G_{it} = g_{it})$, is a result of both the measurement error, $\widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}} - \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}} = (\varsigma_i^{2.00}, \varsigma_i^{3.00}, \varsigma_i^{3.75})$, and the interpolation error, $\widetilde{E}^W(g_{it}; \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}}) - E(W_{it}|G_{it} = g_{it})$.

The mean of subjective income distribution computed using Approach 2, $\widetilde{E^2}(W_{it})$, is then given by,

$$
\begin{aligned}
\widetilde{E^2}(W_{it}) &= \int_2^4 \widetilde{E}(W_{it}|G_{it} = g_{it}) d\widetilde{F}_{G_{it}}(g_{it}) = \int_2^4 \widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}) d\widetilde{F}_{G_{it}}(g_{it}) \\
&= \int_2^4 E(W_{it}|G_{it} = g_{it}) d\widetilde{F}_{G_{it}}(g_{it}) + \int_2^4 (\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}) - E(W_{it}|G_{it} = g_{it})) d\widetilde{F}_{G_{it}}(g_{it}) \\
&= \int_2^4 E(W_{it}|G_{it} = g_{it}) dF_{G_{it}}(g_{it}) + \int_2^4 E(W_{it}|G_{it} = g_{it}) dF^{\Delta}_{G_{it}}(g_{it}) \\
&\quad + \int_2^4 (\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}) - E(W_{it}|G_{it} = g_{it})) d\widetilde{F}_{G_{it}}(g_{it}) \\
&= E(W_{it}) + \int_2^4 E(W_{it}|G_{it} = g_{it}) dF^{\Delta}_{G_{it}}(g_{it}) + \int_2^4 (\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}) - \widetilde{E}^W(g_{it}; \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}})) d\widetilde{F}_{G_{it}}(g_{it}) \\
&\quad + \int_2^4 (\widetilde{E}^W(g_{it}; \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}}) - E(W_{it}|G_{it} = g_{it})) d\widetilde{F}_{G_{it}}(g_{it}) \qquad (26)
\end{aligned}
$$

Following steps similar to those in Section D, we can show that:

$$
\int_2^4 (\widetilde{E}^W(g_{it}; \widetilde{\mathbf{E}}^{\mathbf{W}}_{\mathbf{G_{it}}}) - \widetilde{E}^W(g_{it}; \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}})) d\widetilde{F}_{G_{it}}(g_{it}) = \sum_{g_{it}} \widetilde{\lambda}_i^{g_{it}} \varsigma_i^{g_{it}}, \qquad g_{it} = 2.00, 3.00 \text{ or } 3.75, \qquad (27)
$$

where $\widetilde{\lambda}_i^{2.00} = \int_2^3 (3 - g_{it}) d\widetilde{F}_{G_{it}}(g_{it})$, $\widetilde{\lambda}_i^{3.00} = \int_2^3 (g_{it} - 2) d\widetilde{F}_{G_{it}}(g_{it}) + \int_3^4 (1 - \frac{g_{it}-3}{0.75}) d\widetilde{F}_{G_{it}}(g_{it})$ and $\widetilde{\lambda}_i^{3.75} = \int_3^4 \frac{g_{it}-3}{0.75} d\widetilde{F}_{G_{it}}(g_{it})$.

Denoting $\Delta_{it} \equiv \int_2^4 E(W_{it}|G_{it} = g_{it}) dF^{\Delta}_{G_{it}}(g_{it}) + \int_2^4 (\widetilde{E}^W(g_{it}; \mathbf{E}^{\mathbf{W}}_{\mathbf{G_{it}}}) - E(W_{it}|G_{it} = g_{it})) d\widetilde{F}_{G_{it}}(g_{it})$, equation (26) can be written as:

$$
\widetilde{E^2}(W_{it}) = E(W_{it}) + \sum_{g_{it}} \widetilde{\lambda}_i^{g_{it}} \varsigma_i^{g_{it}} + \Delta_{it} \qquad g_{it} = 2.00, 3.00 \text{ or } 3.75. \qquad (28)
$$

Taking the difference between the mean computed using Approach 1 and the mean computed using Approach 2, we obtain:

$$\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it}) = \varsigma_i - \sum_{g_{it}} \widetilde{\lambda}_i^{g_{it}} \varsigma_i^{g_{it}} - \Delta_{it} \qquad g_{it} = 2.00,\ 3.00 \text{ or } 3.75. \tag{29}$$

Recall that ς_i and $\varsigma_i^{g_{it}}$, $g_{it} = 2.00$, 3.00 or 3.75, are, by assumption, independent of other factors. Hence, they are independent of Δ_{it} since none of them show up in the expression of Δ_{it}. Taking the variance of both sides of equation (29), we find:

$$\begin{aligned} var(\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it})) &= var(\varsigma_i - \sum_{g_{it}} \widetilde{\lambda}_i^{g_{it}} \varsigma_i^{g_{it}}) + var(\Delta_{it}) \\ &= var(\varsigma_i)[1 + \sum_{g_{it}} E((\widetilde{\lambda}_i^{g_{it}})^2)] + var(\Delta_{it}) \\ &\geq var(\varsigma_i)[1 + \sum_{g_{it}} E((\widetilde{\lambda}_i^{g_{it}})^2)]. \end{aligned} \tag{30}$$

Therefore,

$$var(\varsigma_i) \leq \frac{var(\widetilde{E^1}(W_{it}) - \widetilde{E^2}(W_{it}))}{1 + \sum_{g_{it}} E((\widetilde{\lambda}_i^{g_{it}})^2)}. \tag{31}$$

Since both $\widetilde{\lambda}_i^{g_{it}}$ in this section and $\lambda_i^{g_{it}}$ in Section 3.2.2 are computed using the same distribution of G_{it} (we assume that there is no error in the distribution of G_{it} in Section 3.2.2), they are numerically identical. Thus, the right side of equation (31) is numerically identical to the right side of equation (13). As a result, equation (31) shows that our estimates of $var(\varsigma_i)$ reported in Table 4 should be considered as upper bounds for the true value of $var(\varsigma_i)$.

G Joint Decomposition

In Section 4.1 and Section 4.2, we estimated the fraction of total initial uncertainty that is explained by uncertainty about GPA and major, respectively. In this appendix we explain how to examine how much of total initial income uncertainty is due to uncertainty about both of the two factors combined.

We start by decomposing total income uncertainty into the contribution of uncertainty about both final GPA and major and the contribution of uncertainty about other factors, following an equation similar to Equation (4) and Equation (15):

$$\begin{aligned} var(W_{it}) &= var_{G_{it},M_{it}}(E(W_{it}|G_{it},M_{it})) + E_{G_{it},M_{it}}(var(W_{it}|G_{it},M_{it})) \\ &= \{var_{G_{it}}[E_{M_{it}|G_{it}}(E(W_{it}|G_{it},M_{it}))] + E_{G_{it}}[var_{M_{it}|G_{it}}(E(W_{it}|G_{it},M_{it}))]\} + E_{G_{it},M_{it}}(var(W_{it}|G_{it},M_{it})) \\ &= \{var_{G_{it}}(E(W_{it}|G_{it})) + E_{G_{it}}[var_{M_{it}|G_{it}}(E(W_{it}|G_{it},M_{it}))]\} + E_{G_{it}}[E_{M_{it}|G_{it}}(var(W_{it}|G_{it},M_{it}))]. \end{aligned} \tag{32}$$

The sum of the two terms in the fancy bracket corresponds to the contribution of uncertainty about both final GPA and major to total initial income uncertainty, while the last term corresponds to the contribution of uncertainty about all other factors. Analogous to Equation (14) and Equation (16), we define the contribution of final GPA and major to total income uncertainty as follows:

$$R_{it}^{GM} = \frac{var_{G_{it}}(E(W_{it}|G_{it})) + E_{G_{it}}[var_{M_{it}|G_{it}}(E(W_{it}|G_{it}, M_{it}))]}{\{var_{G_{it}}(E(W_{it}|G_{it})) + E_{G_{it}}[var_{M_{it}|G_{it}}(E(W_{it}|G_{it}, M_{it}))]\} + E_{G_{it}}[E_{M_{it}|G_{it}}(var(W_{it}|G_{it}, M_{it}))]}. \tag{33}$$

G.1 Estimation

We focus on the time of entrance ($t = 0$). In order to compute the joint contribution of final GPA and major, we need to compute all three terms on the RHS of Equation (32). The first term can be computed using exactly the same method as in Section 3.1.2. We now explain how to estimate the second and third term on the RHS.

Note that we can compute $E(W_{i0}|G_{i0})$ and $var(W_{i0}|G_{i0})$ for $G_{i0} = 2.00, 3.00, 3.75$. Hence, if we have data on the distribution of $M_{i0}|G_{i0}$, we can apply the method detailed in Section 4.2 to estimate $E(W_{i0}|G_{i0}, M_{i0})$ and $var(W_{i0}|G_{i0}, M_{i0})$ for all M_{i0} and $G_{i0} = 2.00, 3.00, 3.75$ and compute $var_{M_{i0}|G_{i0}}(E(W_{i0}|G_{i0}, M_{i0}))$ and $E_{M_{i0}|G_{i0}}(var(W_{i0}|G_{i0}, M_{i0}))$ for $G_{i0} = 2.00, 3.00, 3.75$. Then, we can interpolate their values at other realizations of G_{i0} ($G_{i0} \neq 2.00, 3.00, 3.75$) and compute $E_{G_{i0}}[var_{M_{i0}|G_{i0}}(E(W_{i0}|G_{i0}, M_{i0}))]$ and $E_{G_{i0}}[E_{M_{i0}|G_{i0}}(var(W_{i0}|G_{i0}, M_{i0}$ using a simulation-based method.

Unfortunately, the distribution of $M_{i0}|G_{i0}$ is not directly available in the data. To deal with this issue, we propose a method to estimate it using data on the unconditional distribution of M_{i0}, P_{ij0}, the distribution of G_{i0}, $F_{G_{i0}}(g_{i0})$ and the expectation of $G_{i0}|M_{i0}$, $E(G_{i0}|M_{i0})$.[30]

Denote the conditional probability of major, $\text{Prob}(M_{i0} = j|G_{i0} = g_{i0})$, as $P_{ij0}^{C}(g_{i0})$. Furthermore, we assume that $P_{ij0}^{C}(g_{i0})$ has the following form:

$$P_{ij0}^{C}(g_{i0}; \rho_{i10}^{0}, \rho_{i10}^{1}, ...) = \frac{exp(\rho_{ij0}^{0} + \rho_{ij0}^{1} g_{i0})}{\sum_{j'} exp(\rho_{ij'0}^{0} + \rho_{ij'0}^{1} g_{i0})}, \tag{34}$$

where ρ_{i70}^{0} and ρ_{i70}^{1} are normalized to 0. This leaves us $2 \times (7 - 1) = 12$ parameters to estimate. Note that this specification actually corresponds to the case where final major is determined by a multinomial logistic model with final GPA as the regressor.

[30] More precisely, what we observe in the data (Question 5 in Appendix A) is the conditional expectation of semester GPA, $E(G_{i0}^{k}|M_{i0})$, instead of the conditional expectation of final GPA, $E(G_{i0}|M_{i0})$. The two would be identical if there does not exist a GPA minimum requirement for graduation. In practice, because most students believe that receiving grades less than the minimum is highly unlikely (and do not think they will drop out), in this section we simply approximate $E(G_{i0}|M_{i0})$ by $E(G_{i0}^{k}|M_{i0})$.

We start by writing $E(G_{i0}|M_{i0})$ as a function of P_{ij0}, $F_{G_{i0}}(g_{i0})$ and $P^C_{ij0}(g_{i0})$.

$$\begin{aligned} E(G_{i0}|M_{i0}) &= \int g_{i0} dF_{G_{i0}|M_{i0}}(g_{i0}) \\ &= \int \frac{P^C_{ij0}(g_{i0})}{P_{ij0}} g_{i0} dF_{G_{i0}}(g_{i0}), \end{aligned} \tag{35}$$

where the second line follows from the Bayes rule.

We can rearrange the terms in Equation (35) to derive an expression for P_{ij0}:

$$\begin{aligned} P_{ij0} &= \frac{1}{E(G_{i0}|M_{i0})} \int P^C_{ij0}(g_{i0}) g_{i0} dF_{G_{i0}}(g_{i0}) \\ &= \frac{1}{E(G_{i0}|M_{i0})} \int \frac{exp(\rho^0_{ij0} + \rho^1_{ij0} g_{i0})}{\sum_{j'} exp(\rho^0_{ij'0} + \rho^1_{ij'0} g_{i0})} g_{i0} dF_{G_{i0}}(g_{i0}). \end{aligned} \tag{36}$$

Note that, by definition, P_{ij0} also satisfies the following equation:

$$\begin{aligned} P_{ij0} &= \int P^C_{ij0}(g_{i0}) dF_{G_{i0}}(g_{i0}) \\ &= \int \frac{exp(\rho^0_{ij0} + \rho^1_{ij0} g_{i0})}{\sum_{j'} exp(\rho^0_{ij'0} + \rho^1_{ij'0} g_{i0})} dF_{G_{i0}}(g_{i0}). \end{aligned} \tag{37}$$

Equation (36) and (37) allow us to express P_{ij0} as two different functions of $(\rho^0_{ij0}, \rho^1_{ij0})$, $j = 1, 2, 3, ..., 7$. We label them as $\widetilde{P}^1_{ij0}(\cdot)$ and $\widetilde{P}^2_{ij0}(\cdot)$, respectively. We then define the estimator of $(\rho^0_{ij0}, \rho^1_{ij0})$, $j = 1, 2, 3, ..., 7$ to be the minimizer of the sum of squared differences between P_{ij0} and $\widetilde{P}^1_{ij0}(\rho^0_{i10}, \rho^1_{i10}, ...)$ and between P_{ij0} and $\widetilde{P}^2_{ij0}(\rho^0_{i10}, \rho^1_{i10}, ...)$. Formally, we have:

$$\{\widehat{\rho}^0_{i10}, \widehat{\rho}^1_{i10}, ...\} \equiv argmin \sum_{q=1}^{2} \sum_{j=1}^{7} [\widetilde{P}^q_{ij0}(\rho^0_{i10}, \rho^1_{i10}, ...) - P_{ij0}]^2. \tag{38}$$

Once $\{\widehat{\rho}^0_{i10}, \widehat{\rho}^1_{i10}, ...\}$ are estimated, we can use Equation (34) to compute the distribution of $M_{i0}|G_{i0}$ for any realization of G_{i0} and compute the three terms in Equation (32) in the way described in the second paragraph of this subsection.