ECON 899b: Problem Set 2

Katherine Kwok*

November 2021

Overview: For this assignment, the goal is program different methods of numerical integration to evaluate the choice probabilities and log-likelihood in a multinomial probit model. In particular, we use the quadrature method, GHK method, and accept/reject method to evaluate the choice probabilities.

Set Up: Our goal is to evaluate the likelihood for individuals to repay their loans in different periods, using a sample from the National Survey of Mortgage Originations Public Use File. There are two related outcome variables $T_i \in \{1, 2, 3, 4\}$, the period of the loan, and $Y_i t$, whether the loan was prepaid at the end of period t. The relationship between the variables are as follows:

$$T_i = \begin{cases} 1 & \text{If } Y_{i0} = 1 \\ 2 & \text{If } Y_{i0} = 0 \text{ and } Y_{i1} = 1 \\ 3 & \text{If } Y_{i0} = 0 \text{ and } Y_{i1} = 0 \text{ and } Y_{i2} = 1 \\ 3 & \text{If } Y_{i0} = 0 \text{ and } Y_{i1} = 0 \text{ and } Y_{i2} = 0 \end{cases} \tag{1}$$

And we assume that at each period t, $Y_{it} = 1$ if $\alpha_t + X_i\beta + Z + it\gamma + \epsilon_{it} > 0$, where X_i is a vector of time-invariant borrower characteristics, and Z_{it} is a vector of time-varying borrower characteristics. We further impose that $\epsilon_{i0} \sim N(0, \sigma_0^2)$ where $\sigma_0^2 = \frac{1}{(1-\rho)^2}$, and $\epsilon_{it} = \rho\epsilon_{it-1} + \eta_{it}$ for all $t > 1$ and $\eta_{it} \sim N(0, 1)$.

The likelihood for holding a loan for T_i period(s) is given in the handout. These choice probabilities are expressed in terms of $\alpha_t + X_i\beta + Z + it\gamma + \epsilon_{it}$. Essentially, we can use the CDFs and PDFs of the standard normal distribution, integrated between negative infinity and the condition associated with each choice, to find the likelihoods.

Numerical Integration Methods: The attached code file "helper_functions.jl" contains the programs for quadrature method, GHK method, and accept/reject method. I have written a brief sketch of my approach to each method below:

1. **Quadrature Method:** For this method, I use the definition of choice probabilities in the handouts and the KPU sparse-grid nodes and weights (at precision 20). For the likelihood of $T_i = 1$, I only need to evaluate the CDF of the standard normal distribution at $(-\alpha_0 - X_i\beta - Z_{it}\gamma)/\sigma_0$.

   For the likelihoods of $T_i \in \{2, 3, 4\}$, I need to use the KPU sparse-grids. The likelihood of $T_i = 2$ requires a single integration between negative infinity and $\alpha_0 + X_i\beta + Z_{it}\gamma$, while the likelihoods of $T_i \in \{3, 4\}$ require double integrations over negative infinity and $\alpha_0 + X_i\beta + Z_{it}\gamma$, negative and infinity and $\alpha_1 + X_i\beta + Z_{it}\gamma$.

*I collaborated with Anya Tarascina and Claire Kim on this assignment.

Briefly, the steps for $T_i \in \{2,3,4\}$ are: First, transform the grid points from (0,1) into the appropriate range of the integration. Second, take the transformed grid point and plug it into the CDF and density functions as defined for each choice probability. Then, multiply the product from the second step with the KPU weights and sum. The weighted sum is the choice probability for $T_i \in \{2,3,4\}$.

2. **GHK Method:** For this method, the idea is to utilize the nested structure of the error terms for $t \in \{0,1,2\}$. I sequentially draw the error terms and define the choice probabilities as follows:

   - Compute $\Phi_{i0} = \Pr(\epsilon_{i0} < \alpha_0 - X_i\beta - Z_{it}\gamma)$, and draw ϵ^r_{i0} from the truncated standard normal distribution $\Phi((\alpha_0 - X_i\beta - Z_{it})/\sigma)$. The choice probability $Pr(T_i = 1|X_i, Z_{it}, \theta) = \Phi_{i0}$.
   - Draw η^r_{i1} from the truncated standard normal distribution $\Phi(\alpha_1 - X_i\beta - Z_{it} - \rho\epsilon^r_{i0})$. Compute $\Phi_{i1} = \Pr(\eta_{i1} < \alpha_1 - X_i\beta - Z_{it}\gamma - \rho\epsilon^r_{i0})$. Calculate $\epsilon^r_{i1} = \rho\epsilon^r_{i0} + \eta^r_{i1}$. The choice probability $Pr(T_i = 2|X_i, Z_{it}, \theta) = (1 - \Phi_{i0})\Phi_{i1}$.
   - Compute $\Phi_{i2} = \Pr(\eta_{i2} < \alpha_2 - X_i\beta - Z_{it}\gamma - \rho\epsilon^r_{i1})$. The choice probability $Pr(T_i = 3|X_i, Z_{it}, \theta) = (1 - \Phi_{i0})(1 - \Phi_{i1})\Phi_{i2}$.
   - The choice probability $Pr(T_i = 4|X_i, Z_{it}, \theta) = (1 - \Phi_{i0})(1 - \Phi_{i1})(1 - \Phi_{i2})$.

   I repeat the algorithm above for 100 simulations, and then compute the average choice probabilities for each observation in the data set.

3. **Accept/Reject Method:** For this method, the idea is to repeatedly make random draws of the error terms and check whether they fall in the bounds corresponding to each choice. The steps I followed are as follows:

   - Randomly draw 100 $\epsilon_{i0}, \eta_{i1}, \eta_{i2}$ values from the uniform distribution over (0,1). Rather than using the random draw, an alternative is to use Halton sequences.
   - Using the inverse CDF of the normal and standard normal distributions, convert $\epsilon_{i0}, \eta_{i1}, \eta_{i2}$ and compute $\epsilon_{i1} = \rho\epsilon_{i0} + \eta_{i1}$, $\epsilon_{i2} = \rho\epsilon_{i1} + \eta_{i2}$.
   - Using similar conditions as with the quadrature method, check if the drawn $\epsilon_{i0}, \epsilon_{i1}, \epsilon_{i2}$ are within the appropriate ranges for the choices associated with $T_i = 1, 2, 3, 4$.
   - Calculate the choice probabilities as number of accepted draws divided by total number of draws.

Findings: The attached code file "main_program.jl" runs the three numerical integration methods using the given paramter values, and then runs maximum likelihood estimation of the probability of repaying loans in different periods using the quadrature method. My results are summarized in the tables below.

Table 1 shows the summary statistics for the choice probabilities regarding loan periods (T = 1, 2, 3, 4). Basically, I select the choice probabilities based on the observed choice for a given individual (e.g. select $Pr(T_i = 1)$ if individual had observed loan period $T_i = 1$). Then, I compute the mean, minimum, 25th percentile, median, 75th percentile, and maximum for a given method. The mean, 25th percentile, median, and 75th percentile are quite similar across all three methods. However, the simulation-based methods are farther off from the quadrature results, potentially because I used random draws, rather than the Halton method to sample the simulation data.

Table 2 provides another summary of the predicted choice probabilities from running the three numerical integration methods. It appears that the predicted probabilities for T= 2 and T = 3 from the GHK and Accept/Reject methods are lower than that of Quadrature method. However, the predicted probabilities for T = 1 and T = 4 are very similar.

Table 1: Summary Statistics Choice Probabilities by Method

statistic	quadrature	GHK	accept reject
mean	0.57043	0.54897	0.5527
minimum	0.10441	0.04655	0.0
q25	0.5687	0.51535	0.54
median	0.67192	0.66321	0.66
q75	0.71652	0.72414	0.72
maximum	0.75975	0.78	0.89

Notes: This table provides the summary statistics for choice probabilities for the observed choice of an individual in the dataset. For instance, if an individual is observed with T = 4, then I only select their probability of choosing T = 4 in computing the summary statistics.

Table 2: Average Choice Probabilities by Method

choice probabilities	quadrature	GHK	accept reject
Probability of $T_i = 1$	0.13535	0.13535	0.13527
Probability of $T_i = 2$	0.20073	0.10348	0.11355
Probability of $T_i = 3$	0.15597	0.08783	0.071
Probability of $T_i = 4$	0.68029	0.67334	0.68019

Notes: This table summarizes the choice probabilities for each loan period based on individual characteristics and given parameter values, across all individuals in the data set (not subsetting by observed choice).

For the coefficients, my estimates for i_close_0 (1 - i_open_0) are much larger than the STATA results. My estimates for the next two choices seem relatively closer, though they are still not exactly matching the STATA results either. If I had more time, I would look into this, to see what is causing the issue.

Table 3: Probit Coefficients Estimated by MLE

coefficients	i_close_0	i_close_1	i_close_2
α_0	204.915	0.177	0.011
α_1	-1.0	-0.600	-0.968
α_2	-1.0	-1.0	-0.872
score_0	1484.565	3.502	1.0299
rate_spread	26.799	0.157	0.0459
i_large_loan	51.441	0.063	0.0209
i_medium_loan	91.225	0.218	0.055
i_refinance	80.767	0.165	0.062
age_r	96.878	0.251	0.076
cltv	169.480	0.526	0.154
dti	78.088	0.257	0.077
cu	400.916	1.145	0.338
first_mort_r	361.501	0.844	0.258
i_FHA	95.065	0.447	0.141
i_open_year2	41.859	0.104	0.036
i_open_year3	44.042	0.113	0.037
i_open_year4	41.283	0.168	0.038
i_open_year5	37.638	0.119	0.041
score_0	1484.864	3.802	1.330
score_1	1469.622	3.397	1.218
score_2	1478.289	3.435	1.208
ρ	-5317.752	-0.240	0.323