Written Exam at the Department of Economics
Winter 2019–20

Advanced Microeconometrics

Retake Exam

— Suggested Answers —

Problem 1

Consider the following discrete choice model for a sample of individuals $i = 1, \ldots, N$:

$$y_i = \underset{j \in \{1,\ldots,J\}}{\arg\max} \{u_{ij}\}, \qquad u_{ij} = v_{ij} + \varepsilon_{ij}, \qquad v_{ij} = x_{ij}'\beta + w_i'\gamma_j \tag{1}$$

where $y_i \in \{1, \ldots, J\}$ is the alternative chosen by individual i and u_{ij} denotes the utility derived from choosing alternative j for individual i. Utility, u_{ij}, is composed of an observed and deterministic part of utility, v_{ij}, and unobserved and random component of utility, ε_{ij}.

The observed part of utility $v_{ij} = x_{ij}'\beta + w_i'\gamma_j$ depends on a vector of choice specific observed explanatory variables, x_{ij}, that vary with both individuals i and choice alternative j, and a vector of observed characteristics, w_i, only specific to the decision maker i.

Assume further that ε_{ij} is iid extreme value type 1 distributed with location parameter $\mu = 0$ and scale parameter σ, such that conditional choice probabilities has the logit form

$$\Pr(y_i = j \mid v_{i1}, \ldots, v_{iJ}; \sigma) = \frac{\exp(v_{ij}/\sigma)}{\sum_{k=1}^{J} \exp(v_{ik}/\sigma)}$$

Question 1.1: Discuss the identification of this model and explain if the scale and level of utility are identified from the choice data. In particular, discuss whether parameters β, γ_j and σ^2 are identified given the observable

variables $\{y_i, x_{i1}, \ldots, x_{iJ}, w_i\}$. If some parameters are unidentified, discuss which normalizations that are necessary to achieve identification.

Suggested answer

If we plug the expressions for v_{ij} into the conditional choice probability

$$\Pr(y_i = j|x_{i1}, ..., x_{iJ}, w_1, ..., w_J; \sigma, \beta, \gamma) = \frac{\exp(x'_{ij}\beta/\sigma + w'_i\lambda_j/\sigma)}{\sum_{k=1}^{J} \exp(x'_{ik}\beta/\sigma + w'_i\lambda_k/\sigma)}$$

For discrete choice models in general, we need to fix both the scale and level of utility. As should be clear from the above expression, we see that only the ratios - β/σ and λ/σ - are identified, since we can multiply the parameters by any constant c without affecting these ratios. Hence, the parameters γ, β and σ are not separable identified and we need some normalization of the parameters for identification. For instance we can normalize σ to 1.

Since only differences in utility matter for the optimal choices, we also need to fix the level of utility for one alternative. The utilities can be arbitrarily shifted by adding a constant K without changing the choice probabilities. For the logit model this gives

$$\Pr(y_i = j|v_{i1}, ..., v_{iJ}; \sigma) = \frac{\exp(v_{ij}/\sigma + K)}{\sum_{k=1}^{J} \exp(v_{ik}/\sigma + K)} = \frac{\exp(K)\exp(v_{ij}/\sigma)}{\exp(K)\sum_{k=1}^{J} \exp(v_{ik}/\sigma)} = \frac{\exp(v_{ij}/\sigma)}{\sum_{k=1}^{J} \exp(v_{ik}/\sigma)}$$

Usually we normalize the the level of utility for some reference alternative to zero, say $j = 1$, by excluding any intercepts, β_0, in β and setting $\gamma_1 = 0$.

Finally note that γ is identified through variation in choices, y_i and individual characteristics, w_i, and β is identified through variation in y_i and through variation in x_{ij} over alternatives.

Question 1.2: Given the model outlined above, derive the corresponding log-likelihood function for a random sample of N observations $\{y_i, x_{i1}, \ldots, x_{iJ}, w_i\}_{i=1}^{N}$, and describe how to obtain Maximum Likelihood estimates of the parameters of interest.

Suggested answer

The likelihood contribution is given by the conditional choice probability for the chosen alternatives and measures how likely a given realized observation is as a function of the parameters

$$l_i(\sigma, \beta, \gamma) = \prod_{k=1}^{J} \Pr(y_i = k | v_{i1}, ..., v_{iJ}; \sigma, \beta, \gamma)^{\mathbb{1}\{y_i = j\}}$$

The corresponding likelihood function for a random sample of individual is

$$L(\sigma, \beta, \gamma) = \prod_{i=1}^{N} l_i(\sigma, \beta, \gamma)$$

and the corresponding log-likelihood function is given by

$$\begin{aligned}
\log L(\sigma, \beta, \gamma) &= \sum_{i=1}^{N} \sum_{k=1}^{J} \log Pr(y_i = k | v_{i1}, ..., v_{iJ}; \sigma, \beta, \gamma) \cdot \mathbb{1}\{y_i = j\} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{J} (x'_{ik}\beta/\sigma + w'_i\gamma_k/\sigma \\
&\quad - \log\{\sum_{l=1}^{J} \exp\{x'_{il}\beta/\sigma + w_i\gamma_l/\sigma\}\}) \cdot \mathbb{1}\{y_i = j\}
\end{aligned}$$

To to obtain the Maximum Likelihood Estimator (MLE) we want to maximize the log-likelihood subject to the normalizations mentioned above, ie. $\sigma = 1$, $\gamma_1 = 0$ and $\beta_0 = 0$. Hence, we want to maximize the likelihood function with respect to the vector of parameters $\theta = (\beta, \gamma)$ with $\gamma = (\beta_1, .., \beta_K)$ and $\gamma = (\gamma, 2, .., \gamma_J)$.

$$\hat{\theta}_{MLE} = \arg\max \log L(1, \beta, \gamma)$$

The optimization problem can easily be solved by standard gradient based Newton type solvers since the likelihood function for logit is differentiable and globally concave in parameters θ. Since it is a likelihood function we want to maximize, we can use the BHHH algorithm that exploits that the outer-product of the gradient can be used as an approximation of the hessian due to the well known information identity.

Question 1.3: Let $p_{ij} \equiv \Pr(y_i = j \mid w_i, x_{i1}, \ldots, x_{iJ})$ and show that the partial effects of the observed outcome with respect to marginal changes in covariates x_{ik} and w_i can be written as

$$\frac{\partial p_{ij}}{\partial x_{ik}} = p_{ij} \left(\mathbb{1}\{k = j\} - p_{ik} \right) \beta/\sigma$$

and

$$\frac{\partial p_{ij}}{\partial w_i} = p_{ij} \left(\gamma_j/\sigma - \sum_{l=1}^{J} p_{il}\gamma_l/\sigma \right)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

Discuss also whether the partial effects are identified if σ is unidentified?

Suggested answer

Differentiate the conditional choice probabilities, p_{ij} with respect to x_{ik} :

$$\begin{aligned}
\frac{\partial p_{ij}}{\partial x_{ik}} &= \frac{e^{v_{ij}/\sigma}\mathbb{1}\{k = j\}\beta/\sigma(\sum_{l=1}^{J} e^{v_{il}/\sigma})}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} - \frac{e^{v_{ij}/\sigma}e^{v_{ik}/\sigma}\beta/\sigma}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} \\
&= \beta/\sigma\mathbb{1}\{k = j\}\frac{e^{v_{ij}/\sigma}}{\sum_{l=1}^{J} e^{v_{il}/\sigma}} - \beta/\sigma\frac{e^{v_{ij}/\sigma}e^{v_{ik}/\sigma}}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} \\
&= \beta/\sigma p_{ij}\mathbb{1}\{k = j\} - \beta/\sigma p_{ij}p_{ik} \\
&= p_{ij}(\mathbb{1}\{k = j\} - p_{ik})\beta/\sigma
\end{aligned}$$

Similar differentiate p_{ij} with respect to w_i:

$$\begin{aligned}
\frac{\partial p_{ij}}{\partial w_i} &= \frac{e^{v_{ij}/\sigma}\gamma_j/\sigma(\sum_{l=1}^{J} e^{v_{il}/\sigma})}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} - \frac{e^{v_{ij}/\sigma}(\sum_{l=1}^{J} e^{v_{il}/\sigma}\gamma_l/\sigma)}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} \\
&= \gamma_j/\sigma\frac{e^{v_{ij}/\sigma}}{\sum_{l=1}^{J} e^{v_{il}/\sigma}} - \frac{e^{v_{ij}/\sigma}(\sum_{l=1}^{J} \gamma_l/\sigma e^{v_{il}/\sigma})}{(\sum_{l=1}^{J} e^{v_{il}/\sigma})^2} \\
&= \gamma_j/\sigma p_{ij} - p_{ij}\sum_{l=1}^{J} p_{il}\gamma_l/\sigma \\
&= p_{ij}(\gamma_j/\sigma - \sum_{l=1}^{J} p_{il}\gamma_l/\sigma)
\end{aligned}$$

From these we see that the marginal effects of x_{ij} and w_i are identified even though σ is not separately identified, since they only depends on the ratios β/σ and γ/σ which are identified.

Question 1.4: Derive the odds ratio p_{ik}/p_{il} and show that this ratio of choice probabilities between alternatives k and l does not depend on x_{ij} for any alternative j other than k and l. Discuss the implications of this.

Suggested answer

The odds ratio is

$$\begin{aligned}\frac{p_{ik}}{p_{il}} &= \frac{e^{v_{ik}/\sigma}/(\sum_{j=1}^{J} e^{v_{ij}/\sigma})}{e^{v_{il}/\sigma}/(\sum_{j=1}^{J} e^{v_{ij}/\sigma})} \\ &= \frac{e^{v_{ik}/\sigma}}{e^{v_{il}/\sigma}} \\ &= e^{(v_{ik}-v_{il})/\sigma} \\ &= e^{((x_{ik}-x_{il})'\beta+w_i(\gamma_k-\gamma_l))/\sigma}\end{aligned}$$

So the odds ratio between alternative k and alternative l is independent of x_{ij} for any $j \neq k, l$. Hence, any changes in x_{ij}, $j \neq k, l$ will have the same proportional affect on p_{ik} as on p_{il} and thereby not affect the odds ratio. This reflects the independence of irrelevant alternatives (IIA) property of the logit model, which implies restrictive and often unrealistic substitution patterns in the model. One example that illustrate this is the well known red bus - blue bus problem.

Question 1.5: You work with a co-author on a residential choice model where you try to model in what region $y_i \in \{1, \ldots, J\}$ that households choose to locate. You are interested in predicting the effect on location choices from a counterfactual change in attributes of a particular region (such as house prices, school quality, crime rates, pollution, etc.). Your co-author is concerned that the substitution patterns imposed by the logit model are too restrictive and suggests that you instead work with the probit model where ε_{ij} is multivariate normal.

Why is your co-author concerned, and how could probit potentially help to address this issue?

Suggested answer

The coauthor is concerned with the IIA property already mentioned. In the present context, the IIA property implies for instance that an increase

in house prices of region j will increase the demand in terms of choice probabilities for any other region $k \neq j$ with the same proportion. These substitution patterns are often unrealistic. For instance you would anticipate that an increase in the house prices in region j would have the largest effect on houses closes to that region.

The probit model does not suffer from the IIA property, since it allows us to specify a covariance matrix for the error term ε_{ij} that implies correlation across alternatives (residential regions), $j = 1, .., J$; and thereby more flexible spatial substitution patterns. By estimating the parameters of the (appropriately normalized) covariance matrix in the probit model, we can learn about the substitution patterns that we do not capture in the deterministic part of utility, but are nevertheless observed in the data.

Compared to the logit model, the multinomial probit is much harder to solve and estimate since we cannot express the choice probabilities in closed form. More about this below.

Problem 2

We now consider the Probit model, which has the same structure as above except that ε_{ij} in Eq. (1) now follows a multivariate normal distribution. Specifically, ε_{ij} is an element in the J dimensional vector $\varepsilon_i = (\varepsilon_{i1}, .., \varepsilon_{iJ})'$, where $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

Question 2.1: To estimate the parameters of the probit model, your co-author suggests that you construct an estimator based on the following optimization problem:

$$\widehat{\theta} = \arg\max_{\theta} \left[\frac{1}{N} \sum_{i=1}^{N} \ln \widehat{f}(y_i \mid z_i, u_{iM}; \theta) \right] \tag{2}$$

where u_{iM} is a sample of M random draws $u_{iM} = \{u_i^{(1)}, \ldots, u_i^{(M)}\}$ from the standard normal distribution, for each $i = 1, \ldots, N$. Here we have defined $z_i = \{x_{i1}, \ldots, x_{iJ}, w_i\}$ to simplify notation.

Describe the principle of the estimation method your co-author is referring to. As part of your answer, you are expected to provide and justify a possible expression of $\widehat{f}(y_i \mid z_i, u_{iM}; \theta)$, and to outline the steps of the corresponding estimation approach.

Hint (for the very detailed answer): You can always obtain draws from the multivariate normal by rescaling a J-vector $u_i^{(m)}$ of independent draws from the standard normal distribution. In particular, $\varepsilon_i^{(m)} = u_i^{(m)}\mathbf{L} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{L}$ is the lower triangular Cholesky matrix such that $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}'$.

Suggested answer

[Note that there was a typo in the optimization problem stated in Eq. (2) in the original exam set. The argmin operator should be replaced by the argmax operator. We have corrected this in Eq. (2) above. Any confusion originating from this typo has been taken into account when grading.]

The estimation method the colleague is referring to is the *Method of Simulated likelihood* (MSL) estimator.

The MSL estimator finds the parameters that maximizes log of the simulated likelihood function given our sample of random draws, u_{iM}. Hence, a natural choice for $\hat{f}$ would be based on simulated conditional choice probabilities

$$\hat{p}_j(z_i, u_{iM}; \theta) = 1/M \sum_{m=1}^{M} \mathbb{1}\{ \underset{k \in (1,2,\ldots,J)}{\arg\max} \{x_{ij}'\beta + w_i'\gamma_j + \varepsilon_{ik}^{(m)}\} = j\}\}$$

where $\varepsilon_{ik}^{(m)}$ is the k'th element of the vector $\varepsilon_i^{(m)} = u_i^{(m)}L$ containing the J correlated error terms.

For an individual i that chooses alternative y_i the simulated likelihood then becomes

$$\hat{f}(y_i|z_i, u_{iM}; \theta) = \prod_{j=1}^{J} \{\hat{p}_j(z_i, u_{iM}; \theta)\}^{\mathbb{1}\{y_i=j\}}$$

In order to evaluate the objective function of the maximization problem

given by eq. (2) we do the following four steps: (1) for each individual i, take M random samples of the J dimensional vector $u_i^{(m)}$ drawn from the the standard normal distribution, (2) calculate the corresponding correlated error terms, $\varepsilon_i^{(m)}$, (3) calculate $\hat{f}$ for each individual., (4) aggregate over i to obtain the objective function.

Note that for each evaluation of the objective function, as we search over the parameter space, we keep the simulation draws, $u_i^{(m)}$, fixed.

The parameters are the structural parameters mentioned above and the parameters indexing the variance-covariance matrix Σ (when appropriately normalized by fixing at least one element).

Question 2.2: How do you recommend to choose the number of random draws M in Question 2.1? In particular, explain how this number affects the bias of the estimator. Is the estimator consistent for a fixed number of draws, e.q. $M = 1$. *(no derivations expected).*

Suggested answer

The MSL estimator contains a simulation bias due to the log transformation of the simulated probabilities. Suppose $\hat{p}_{ij}(\theta) \equiv \hat{p}_j(z_i, u_{iM}; \theta)$ is an unbiased simulator of the true choice probabilities $p_{ij}(\theta)$, so that $\mathbb{E}_m[\hat{p}_{ij}(\theta)] = p_{ij}(\theta)$, where the expectation is over draws used in the simulation. However, since the log operator is a nonlinear transformation, $\log(p_{ij}^{(m)}(\theta))$ is not an unbiased simulator for $\log(p_{ij}(\theta))$. The bias in the simulator of $\log(p_{ij}(\theta))$ translate into bias in the MSL estimator. This bias diminish as more draws are used in the simulation. MSL is consistent when M increase at any rate with N and is asymptotically equivalent to MLE if M increase at a rate faster than $\sqrt{N}$. However, MSL is biased and inconsistent for any fixed M, e.g. when $M = 1$.

Question 2.3: How would you modify the optimization problem in Eq. (2) to implement an estimator $\widehat{\theta}$ with the remarkable property that it is consistent for $M = 1$? Describe *briefly* the corresponding estimation approach.

Suggested answer

We could apply the *Method of Simulated Moments* (MSM) estimator. A possible choice would be the following moment conditions

$$q(y_i, z_{ij}; \theta) = [\mathbb{1}\{y_i = j\} - 1/M \sum_{m=1}^{M} p_{ij}^{(m)}(\theta)] z_{ij}$$

where we have defined $z_{ij} = \{x_{ij}, w_i\}$ to simplify notation. The resulting objective function is given by

$$Q(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} [\mathbb{1}\{y_i = j\} - 1/M \sum_{m=1}^{M} p_{ij}^{(m)}(\theta)] z_{ij} = 0$$

The important feature of this estimator is that $p_{ij}^{(m)}$ enters the objective function linearly. As a result, if $p_{ij}^{(m)}$ is unbiased subsimulator for p_{ij} then $[\mathbb{1}\{y_i = j\} - 1/M \sum_{m=1}^{M} p_{ij}^{(m)}(\theta)] z_{ij}$ is unbiased for $[\mathbb{1}\{y_i = j\} - p_{ij}(\theta)] z_{ij}$. Since there is no simulation bias in the estimation condition, the MSM estimator is consistent, even for $M = 1$.

Problem 3

Consider the following MATLAB functions:

```
function [p] = f1(V)
    p = exp(V)./sum(exp(V));
end

function [p, y] = f2(V, M)
    J=numel(V);
    U = V - evrnd(0,1,M,J);
    [Vmax, y] = max(U,[],2);
    for j=1:J
        p(j)=mean(y==j);
    end
end
```

and the following piece of code:

```
rng(123);
V=[0,1,2,3];
M=10000;

fprintf('f1 = ');
fprintf('%10.4f ', f1(V));
fprintf('\n');

fprintf('f2 = ');
fprintf('%10.4f ', f2(V, M));
fprintf('\n');
```

which produces the following output:

```
f1 =     0.0321     0.0871     0.2369     0.6439
f2 =     0.0306     0.0897     0.2306     0.6491
```

Question 3.1: Express in mathematical terms what these two functions do. You should just provide a few equations to answer this question. Be explicit about the notation.

[Note: The `MATLAB` *function* `evrnd(mu, sigma, m, n)` *produces a* $m \times n$ *matrix of random draws from the type 1 extreme value distribution with location parameter* `mu` *and scale parameter* `sigma`*. MATLAB returns the version suitable for modeling minima rather than maxima. We need the mirror image of this distribution which is why we take the negative value]*

Suggested answer

The first function $p = f1(V)$ calculates the vector of closed form conditional choice probabilities, p, of the logit model

$$p_j = \frac{e^{V_j}}{\sum_{k=1}^{J} e^{V_k}}$$

for a given vector of alternative-specific deterministic utility parts, V.

The second function $[p, y] = f2(V, M)$ outputs a vector of simulated conditional choice probabilities p and a vector of simulated choices y for a given a vector of alternative-specific deterministic utility parts, V, and a given constant M that specifies the number simulations used.

The function first simulates a $J \times M$ matrix of alternative-specific utilities given by the sum of the deterministic utility part and random utility part, where the random utility part is drawn from the extreme value distribution.

$$U_j^{(m)} = V_j + \varepsilon_j^{(m)}, \quad \varepsilon_j^{(m)} \sim EV(0, 1)$$

Secondly the function finds the alternative $y^{(m)}$ that gives the highest amount of utility for each simulation, m

$$y^{(m)} = \underset{j \in (1,2,\ldots,J)}{\arg\max} \{U_j^{(m)}\}$$

Finally the function finds the simulated alternative-specific conditional

choice probabilities, p_j

$$p_j = 1/M \sum_{m=1}^{M} \mathbb{1}\{y^{(m)} = j\}$$

Hence, p is a vector containing J conditional choice probabilities and y is a vector containing M simulated choices.

Question 3.2: Explain why the output of `f1()` and `f2()` look similar.

Suggested answer

Function $f1()$ calculates the closed form conditional choice probabilities of the logit model, derived from the random utility model, where the random utility part is assumed to follow the extreme value distribution.

Function $f2(\,)$ calculates the simulated conditional choice probabilities from the random utility model where the random utility part is simulated by the extreme value distribution.

This implies that the two functions only differs with respect to how the random utility part is integrated out. In the closed form expression for the conditional choice probabilities the random utility part is integrated out analytically. In contrast the random utility part is integrated out nummerically when calculating the simulated the conditional choice probabilities by applying Monte Carlo integrations. Hence, as the number of simulations, M, grows the simulated conditional choice probabilities should by the law of large numbers converge to the closed form conditional choice probabilities.

$$1/M \sum_{m=1}^{M} \mathbb{1}\{V_j + u_{ij}^{(m)} = j\} \xrightarrow[M\to\infty]{} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \mathbb{1}\{\arg\max_k V_k + \varepsilon_k = j\} d\varepsilon_1, ..., d\varepsilon_J$$

Since $M = 10,000$ in the provided piece of code, we would expect these conditional choice probabilities to be very similar, which is also the case in the shown output.