Written Exam at the Department of Economics
Winter 2018–19

Advanced Microeconometrics

Final Exam

— Suggested Answers —

Problem 1

Consider the following random utility model, for a sample of N individuals:

$$y_i = \underset{j\in\{0,1\}}{\arg\max}\{u_{ij}\}, \qquad i = 1,\dots,N, \tag{1}$$

$$u_{ij} = x_i'\beta_j + \varepsilon_{ij}, \qquad \text{for } j = 0,1, \tag{2}$$

where the explanatory variables contained in the $K \times 1$ vector x_i influence each level of utility through the $K \times 1$ vector of regression coefficients β_j, for $j = 0, 1$.

The error terms are assumed to be independent and identically distributed across observations, with a joint normal distribution:

$$\begin{pmatrix}\varepsilon_{i0}\\ \varepsilon_{i1}\end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{bmatrix}\right). \tag{3}$$

Question 1.1: Show that the probability of choosing alternative 1, for each individual $i = 1,\dots,N$, is equal to

$$\Pr(y_i = 1 \mid x_i, \theta) = \Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}}\right), \tag{4}$$

where $\theta = (\beta_0', \beta_1', \sigma_0^2, \sigma_1^2, \sigma_{01})'$, and $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution.

Suggested answer

Alternative 1 will be chosen if the corresponding utility u_{i1} is larger than the utility associated with the other alternative u_{i0}:

$$\begin{aligned}
\Pr(y_i = 1 \mid x_i, \theta) &= \Pr(u_{i1} > u_{i0} \mid x_i)\,, \\
&= \Pr(x_i'\beta_1 + \varepsilon_{i1} > x_i'\beta_0 + \varepsilon_{i0} \mid x_i, \theta)\,, \\
&= \Pr(\varepsilon_{i0} - \varepsilon_{i1} < x_i'(\beta_1 - \beta_0) \mid x_i, \theta)\,, \\
&= \Pr\left(\frac{\varepsilon_{i0} - \varepsilon_{i1}}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} < \frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \mid x_i, \theta \right)\,, \\
&= \Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \right)\,.
\end{aligned}$$

Given the joint normality of the error terms assumed in Eq. (3), the difference of these two random variables follows a normal distribution, $\varepsilon_{i0} - \varepsilon_{i1} \sim \mathcal{N}(0, \sigma_0^2 + \sigma_1^2 - 2\sigma_{01})$. Therefore, rescaling this difference by $\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}$ provides a random variable following the standard normal distribution. The last line is then obtained from the fact that if $Z \sim \mathcal{N}(0,1)$, then $\Pr(Z < z) = \Phi(z)$.

Question 1.2: Using Eq. (4), derive the corresponding log-likelihood function of the model for the whole sample of N individuals.

Suggested answer

The likelihood function is derived using the expression of the density function, which is equal, for an individual i, to

$$f(y_i \mid x_i, \theta) = \prod_{j=0}^{1} \Pr(y_i = j \mid x_i, \theta)^{\mathbb{1}\{y_i = j\}}\,.$$

This density is derived by considering all possible values taken by y_i, using the probability in Eq. (4).

Given the independence of the error terms assumed across individuals, the likelihood function is equal to the product of the individual contributions to the likelihood:

$$\begin{aligned}
L_N(\theta; y, x) &= \prod_{i=1}^{N} \prod_{j=0}^{1} \Pr(y_i = j \mid x_i, \theta)^{\mathbb{1}\{y_i=j\}}, \\
&= \prod_{i=1}^{N} \left[1 - \Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \right) \right]^{\mathbb{1}\{y_i=0\}} \\
&\qquad \times \left[\Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \right) \right]^{\mathbb{1}\{y_i=1\}},
\end{aligned}$$

for the vector of parameters $\theta = (\beta_0', \beta_1', \sigma_0^2, \sigma_1^2, \sigma_{01})'$, where $\mathbb{1}\{\cdot\}$ is the indicator function that is equal to 1 if the corresponding condition is fulfilled, to 0 otherwise.

The corresponding log-likelihood function requested in the question is obtained as

$$\begin{aligned}
\mathcal{L}_N(\theta; y, x) &= \ln L_N(\theta; y, x), \\
&= \sum_{i=1}^{N} \left\{ \mathbb{1}\{y_i = 0\} \ln \left[1 - \Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \right) \right] \right. \\
&\qquad \left. + \mathbb{1}\{y_i = 1\} \ln \Phi\left(\frac{x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}} \right) \right\}.
\end{aligned}$$

Question 1.3: Discuss the identification of the model. In particular, explain precisely which parameter(s) can be identified, and which restrictions, if any, are required to achieve identification.

Suggested answer

The model parameters θ only enter the (log-)likelihood function derived previously through the probability in Eq. (4). Therefore, identification can be achieved by ensuring that it is not possible to change the values of any of the parameters without affecting this probability—i.e., without changing the corresponding (log-)likelihood.

Several problems may affect identification in this model. First, it appears for any $K \times 1$ real vector c, it is possible to define $\widetilde{\beta}_0 = \beta_0 + c$ and $\widetilde{\beta}_1 = \beta_1 + c$, without changing the probability in Eq. (4), since it depends on $\beta_1 - \beta_0 = \widetilde{\beta}_1 - \widetilde{\beta}_0$. Therefore, only $\gamma \equiv \beta_1 - \beta_0$ can be identified in this model. One possible solution is to fix $\beta_0 = 0$. Second, it is possible to rescale the latent utilities without affecting the likelihood (setting $\beta_0 = 0$ does not solve this problem). This can be seen, for example, by multiplying the covariance matrix of the error terms defined in Eq. (3) by a constant d^2 (i.e., multiply each element of this covariance matrix by d^2), and at the same time multiply the vector of (normalized) regression coefficients γ by d, for any $d > 0$. This transformation does not change the probability in Eq. (4):

$$\Pr(y_i = 1 \mid x_i) = \Phi\left(\frac{x_i'(d\gamma)}{\sqrt{d^2\sigma_0^2 + d^2\sigma_1^2 - 2d^2\sigma_{01}}}\right) = \Phi\left(\frac{x_i'\gamma}{\sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{01}}}\right).$$

The only way to solve this problem is to fix the variance of the difference of the error terms, $\mathrm{V}[\varepsilon_{i0} - \varepsilon_{i1}] = \sigma_0^2 + \sigma_1^2 - 2\sigma_{01}$, to a constant. This requires to fix these three parameters.

Question 1.4: Using the identification strategy discussed in Question 1.3, show that this random utility model with two alternatives can be expressed as a standard probit model. State the corresponding probit model as part of your answer.

Suggested answer

Following up on the previous discussion about identification, it appears that this multinomial probit model with only two alternatives boils down to a standard probit model when $\beta_0 = 0$ and the covariance matrix of the error terms is fixed, for example to $0.5 \times I_2$, where I_2 is the identity matrix of dimension 2, to obtain $\mathrm{V}[\varepsilon_{i0} - \varepsilon_{i1}] = \sigma_0^2 + \sigma_1^2 - 2\sigma_{01} = 1$ (note that σ_0^2, σ_1^2 and σ_{01} could be fixed to different values to produce a unit variance of the difference of the error terms).

The corresponding probit model can be expressed as follows, where the parameters are mapped explicitly to the original model in Eq. (1):

$$\begin{aligned} y_i &= \mathbb{1}\{y_i^\star > 0\}\,, && i = 1, \ldots, N, \\ y_i^\star &\equiv u_{i1} - u_{i0} = x_i'\gamma + u_i, && \\ \gamma &\equiv \beta_1 - \beta_0, && \\ u_i &\equiv \varepsilon_{i1} - \varepsilon_{i0} \overset{iid}{\sim} \mathcal{N}(0,1)\,. && \end{aligned}$$

Problem 2

Consider the following linear regression model with two scalar regressors x_1 and x_2, for $i = 1, \ldots, N$:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}\big(0, \tau^{-1}\big)\,, \tag{5}$$

where the precision parameter $\tau = 1/\sigma^2$ is the inverse of the variance of the error term. The observations are collected in the vectors $y = (y_1, \ldots, y_N)'$, $x_1 = (x_{11}, \ldots, x_{1N})'$ and $x_2 = (x_{21}, \ldots, x_{2N})'$.

The parameters of the model $\theta = (\beta_1, \beta_2, \tau)'$ are assumed to be a priori independent, such that $p(\theta) = p(\beta_1, \beta_2, \tau) = p(\beta_1)p(\beta_2)p(\tau)$. An improper prior is assumed on the regression coefficients, and a Gamma distribution on the precision parameter:

$$p(\beta_1) \propto 1, \qquad p(\beta_2) \propto 1, \qquad \tau \sim \mathcal{G}(a_0, b_0)\,, \tag{6}$$

with $a_0 > 0$ and $b_0 > 0$, where the probability density function of the Gamma distribution is

$$p(\tau \mid a_0, b_0) = \frac{1}{\Gamma(a_0) b_0^{a_0}} \tau^{a_0 - 1} \exp\left\{ -\frac{\tau}{b_0} \right\}, \tag{7}$$

with $\Gamma(\cdot)$ denoting the Gamma function.

Question 2.1: Without deriving any conditional distributions, outline the steps of a Gibbs sampler that can be implemented to draw the three parameters of the model iteratively (i.e., in three different steps).

Be as precise as possible in the description of the sampler.

Suggested answer

Initialize the sampler by assigning starting values to the parameters $\beta_2^{(0)}$ and $\tau^{(0)}$ (note that β_1 will be updated first, hence no starting value required for this parameter). Starting values can be fixed (user-defined) or random (e.g., from the prior), or specified using OLS estimates.

For each MCMC iteration $t = 1, \dots, T$, cycle through the following steps:

(1) Sample $\beta_1^{(t)}$ from $p(\beta_1 \mid y, x_1, x_2, \beta_2^{(t-1)}, \tau^{(t-1)})$.

(2) Sample $\beta_2^{(t)}$ from $p(\beta_2 \mid y, x_1, x_2, \beta_1^{(t)}, \tau^{(t-1)})$.

(3) Sample $\tau^{(t)}$ from $p(\tau \mid y, x_1, x_2, \beta_1^{(t)}, \beta_2^{(t)})$.

The total number of MCMC iterations T should be chosen such that practical convegence to the stationary distribution is achieved. The first T_0 iterations, when the sampler has not yet reached stationarity, should be discarded (burn-in period).

Question 2.2: Derive the conditional distribution $p(\tau \mid y, x_1, x_2, \beta_1, \beta_2)$.

Explain the concept of natural conjugacy, and explain why the Gamma distribution assumed on τ is (or is not) a natural conjugate prior in this model.

Suggested answer

This conditional distribution is derived by applying Bayes' theorem:

$$p(\tau \mid y, x_1, x_2, \beta_1, \beta_2) \propto p(y \mid x_1, x_2, \beta_1, \beta_2, \tau) p(\tau).$$

The likelihood function is obtained from the standard linear regression model, where the precision parameter $\tau = 1/\sigma^2$ is used instead of the variance σ^2. The kernel of this likelihood function, with respect to τ, is derived as:

$$
\begin{aligned}
p(y \mid x_1, x_2, \beta_1, \beta_2, \tau) &= \prod_{i=1}^{N} (2\pi)^{-1/2} \tau^{1/2} \exp\left\{ -\frac{\tau}{2} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right\}, \\
&= (2\pi)^{-N/2} \tau^{N/2} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{N} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right\}, \\
&\propto \tau^{N/2} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{N} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right\}.
\end{aligned}
$$

The kernel of the prior is obtained from Eq. (7):

$$
p(\tau) \propto \tau^{a_0 - 1} \exp\left\{ -\frac{\tau}{b_0} \right\}.
$$

Combining the kernels of the likelihood and of the prior provides:

$$
\begin{aligned}
&p(\tau \mid y, x_1, x_2, \beta_1, \beta_2) \\
&\quad \propto \tau^{N/2} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{N} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right\} \tau^{a_0 - 1} \exp\left\{ -\frac{\tau}{b_0} \right\}, \\
&\quad \propto \tau^{a_0 + N/2 - 1} \exp\left\{ -\tau \left(\frac{1}{b_0} + \frac{1}{2} \sum_{i=1}^{N} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right) \right\},
\end{aligned}
$$

which appears to be the kernel of the following Gamma distribution:

$$
\tau \mid y, x_1, x_2, \beta_1, \beta_2 \sim \mathcal{G}\left(a_0 + \frac{N}{2}, \left[\frac{1}{b_0} + \frac{1}{2} \sum_{i=1}^{N} (y_i - x_{1i}\beta_1 - x_{2i}\beta_2)^2 \right]^{-1} \right).
$$

A prior distribution is said to be a *natural conjugate* if the resulting posterior distribution belongs to the same family of distribution. Since we assumed a Gamma prior distribution on τ and obtained a Gamma distribution for its conditional distribution, we can conclude that the Gamma distribution is a natural conjugate prior for τ in this model.

Three different data sets with $N = 100$ observations and different levels of correlation between the two regressors, $\rho \equiv \mathrm{corr}(x_1, x_2) \in \{0.5, 0.9, 0.99\}$, are generated from the model specified in Eq. (5).

The three-step Gibbs sampler outlined in Question 2.1 is run for 1,000 iterations on each of these three data sets, with the same prior specification in the three cases. The trace plots and autocorrelograms of the parameter β_1 are displayed in Fig. 2.1 for the three cases.

Question 2.3: Match each of the three cases shown in Fig. 2.1 to the three data sets (i.e., the three different values of ρ). Explain intuitively the differences observed between these three MCMC outputs.

Suggested answer

The three MCMC outputs show different levels of autocorrelations, which imply different speeds of convergence as well as different levels of mixing for the three Markov chains. This is due to the correlation ρ between the two regressors x_1 and x_2 used in the data generating process. The Gibbs sampler is implemented in three steps, where each regression coefficient is updated conditionally on the other one (e.g., β_1 given β_2, and β_2 given β_1). As a consequence, the higher the correlation between the two regressors, the higher the autocorrelation of the Markov chain, as the sampler becomes more "sticky" and slower in exploring the whole parameter space because of this correlation. Therefore, we can deduce that case 1 corresponds to $\rho = 0.99$ (largest autocorrelations), case 2 to $\rho = 0.9$ and case 3 to $\rho = 0.5$ (smallest autocorrelations).

Question 2.4: Explain precisely if and how you can use the random draws of β_1 shown in Fig. 2.1 to draw posterior inference about this parameter in each of the three cases.

If these draws cannot be used, what would you have to change in the implementation of the Gibbs sampler to be able to do posterior inference on the parameters?

Suggested answer

To be able to use the random draws of these Markov chains for posterior inference, the first iterations should be discarded as burn-in period to make sure the results do not depend on the initialization of the model (starting values). For cases 2 and 3, a small burn-in period of 20 iterations would be enough, as convergence is very fast. For case 1, convergence is much slower, and a larger number of iterations should be discarded (at least 200).

In all three cases, the number of random draws (after burn-in) might be too small to allow precise posterior inference. This is because Monte Carlo integration relies on a law of large numbers, therefore the larger the number of random draws, the better the approximation. This might be especially problematic is cases 1 and 2, where autocorrelations are large, and more draws would be required for the posterior sample to be representative of the target distribution.

One solution would be to increase the number of MCMC iterations, in order to decrease Monte Carlo error. Another alternative would be to modify the Gibbs sampler. For instance, instead of sampling β_1 and β_2 sequentially, it is possible to sample them jointly from $p(\beta_1, \beta_2 \mid y, x_1, x_2, \tau)$. This joint sampling would reduce the autocorrelation a lot, thereby improving mixing.

[Note: You do not need any additional information about the data generating process, about the prior specification or about the configuration of the Gibbs sampler to answer the last two questions.]

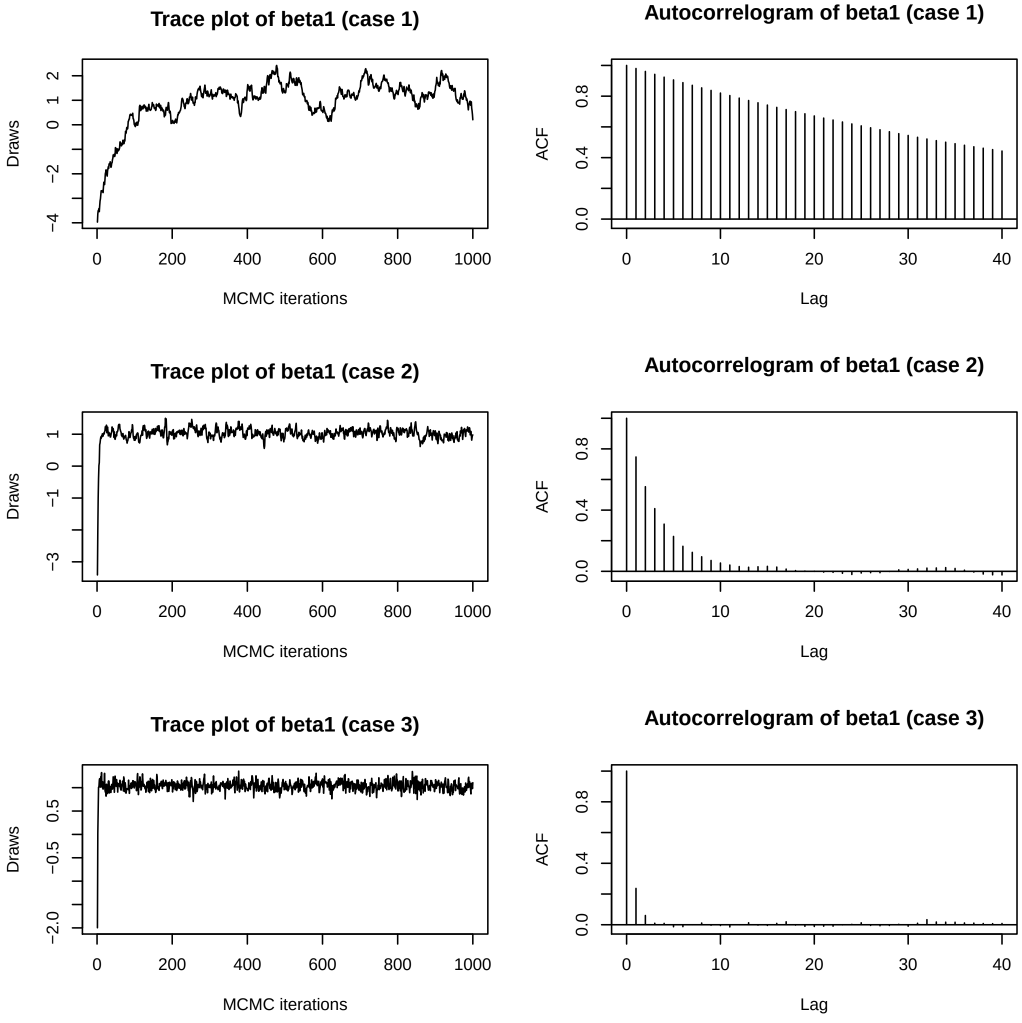

Figure 2.1: Trace plots and autocorrelograms of the parameter β_1 for the three different cases.

Problem 3

Consider the following `MATLAB` function:

```
function [y,X,theta] = gen_data(N,J,K)
    theta = 2*rand(K,J-1)-1;
    X = [ones(N,1), sqrt(2).*randn(N,K-1)];
    V = [zeros(N,1), X*theta];
    E = gevinv(rand(N,J));
    U = V + E;
    [maxU,y] = max(U,[],2);
end
```

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

[Note: The `MATLAB` function `gevinv()` computes the inverse of the CDF of the standard Gumbel distribution (type 1 extreme value distribution).]

Suggested answer

One possible solution (note that alternative representations are possible), for a sample of N individuals, J alternatives and K regressors:

$$
\begin{aligned}
y_i &= \underset{j\in\{1,\ldots,J\}}{\arg\max}\{u_{ij}\}, && \text{for } i = 1,\ldots,N,\\
u_{ij} &= x_i'\theta_j + e_{ij}, && \text{for } j = 1,\ldots,J,\\
e_{ij} &\sim \mathcal{G}umbel(0,1), &&\\
x_i &= (1, x_{i2},\ldots,x_{iK})', &&\\
x_{ik} &\sim \mathcal{N}(0,2)\,, && \text{for } k = 1,\ldots,K,\\
\theta_j &= (\theta_{j1},\ldots,\theta_{jK}), &&\\
\theta_{1k} &= 0, && \text{for } k = 1,\ldots,K,\\
\theta_{jk} &\sim \mathcal{U}(-1,1)\,, && \text{for } j = 2,\ldots,J.
\end{aligned}
$$

Question 3.2: Describe briefly the econometric model that can be used to fit the corresponding data, as well as an estimation method that can be implemented to estimate the unknown parameters θ.

Suggested answer

The data generated by this MATLAB function correspond to a multinomial logit model: The error terms of the model follow the standard Gumbel distribution, and the regressors are fixed across alternatives for all individuals, with corresponding regression coefficients that vary across alternatives. This model has N observations, J different alternatives, and K regressors (where the first one is an intercept term).

The multinomial logit model can be estimated with maximum likelihood estimation.