Spring 2001 John Rust
Economics 551b 37 Hillhouse, Rm. 27

SOLUTIONS TO FINAL EXAM
April 27, 2001

Part I: 15 minutes, 15 points. Answer all questions below:

1. Suppose $\{\tilde{X}_1, \ldots, \tilde{X}_N\}$ are *IID* draws from a $N(\mu, \sigma^2)$ distribution (i.e. a normal distribution with mean μ and variance σ^2). Consider the estimator $\hat{\theta}_N$ defined by:

$$\hat{\theta}_N = \left(\frac{1}{N} \sum_{i=1}^{N} \tilde{X}_i \right)^2 \tag{1}$$

Which of the following statements are true and which are false?

To answer this, note that the sample mean of the N *IID/* observations $\{\tilde{X}_1, \ldots, \tilde{X}_N\}$, $\overline{X}_N$ is distributed as $N(\mu, \sigma^2/N)$. Then $\hat{\theta}_N$ is the square of $\overline{X}_N$ and is thus a non-central χ^2 random variable. Its expectation is

$$E\{\hat{\theta}_N\} = \left(\mu^2 + \frac{\sigma^2}{N} \right) \tag{2}$$

and its variance is

$$\text{var}(\hat{\theta}_N) = E\{\hat{\theta}_N^2\} - [E\{\hat{\theta}_N\}]^2 = E\{\overline{X}_N^4\} - [E\{\overline{X}_N^2\}]^2 = \left(\frac{2\sigma^2}{N^2} + \frac{4\mu^2\sigma^2}{N} \right). \tag{3}$$

Thus it is clear that $\hat{\theta}_N$ converges in probability to μ^2 and is an upward biased estimator of μ^2. These conclusions would follow even if the X_i's were not normally distributed. In that case we would use the continuous mapping theorem and the fact that $\hat{\theta}_N$ is a continuous function (x^2) of the sample mean $\overline{X}_N$, and thus, $\hat{\theta}_N$ converges in probability to μ^2 is a simple application of the continuous mapping theorem. Also since the function x^2 is convex, Jensen's inequality can be used to show that

$$E\{\hat{\theta}_N\} = E\{[\overline{X}_N]^2\} > [E\{\overline{X}_N\}]^2 = \mu^2. \tag{4}$$

From these results the following true and false answers should now be obvious:

A. $\hat{\theta}_N$ is a consistent estimator of σ^2. False.

B. $\hat{\theta}_N$ is an unbiased estimator of σ^2. False.

C. $\hat{\theta}_N$ is a consistent estimator of μ. False.

D. $\hat{\theta}_N$ is an unbiased estimator of μ. False.

E. $\hat{\theta}_N$ is a consistent estimator of μ^2. True.

F. $\hat{\theta}_N$ is an unbiased estimator of μ^2. False.

G. $\hat{\theta}_N$ is an upward biased estimator of μ^2. True.

H. $\hat{\theta}_N$ is a downward biased estimator of μ^2. False.

2. Consider estimation of the linear model

$$y = X\beta + \epsilon \tag{5}$$

based on N *IID* observations $\{y_i, X_i\}$ where X_i is a $K \times 1$ vector of independent variables and y_i is a 1×1 scalar independent variable. Mark each of the following statements as true or false:

A. The Gauss-Markov Theorem proves that the ordinary least squares estimator (OLS) it BLUE (Best Linear Unbiased Estimator). True.

B. The Gauss-Markov Theorem requires that the error term in the regression ϵ be normally distributed with mean 0 and variance σ^2. False.

C. The Gauss-Markov Theorem does not apply if the true regression function does not equal $X\beta$, i.e. if $E\{y|X\} \neq X\beta$. True.

D. The Gauss-Markov Theorem does not apply if there is heteroscedasticity. True.

E. The Gauss-Markov Theorem does not apply if the error term has a non-normal distribution. False.

F. The maximum likelihood estimator of β is more efficient than the OLS estimator of β. True.

G. The OLS estimator of β will be unbiased only if the error terms are distributed independently of X and have mean 0. False.

H. The maximum likelihood estimator of β is the same as OLS only in the case where ϵ is normally distributed. True.

I. The OLS estimator will be a consistent estimator of β even if the error term ϵ is not normal and even if there is heteroscedasticity. True.

J. The OLS estimator of the asymptotic covariance matrix for β, $\hat{\sigma}^2(X'X/N)^{-1}$ (where $\hat{\sigma}^2$ is the sample variance of the estimated residuals $\hat{\epsilon}_i = y_i - X_i\hat{\beta}$) is a consistent estimator regardless of whether ϵ is normally distributed or not. True.

K. The OLS estimator of the asymptotic covariance matrix for β, $\hat{\sigma}^2(X'X/N)^{-1}$ (where $\hat{\sigma}^2$ is the sample variance of the estimated residuals $\hat{\epsilon}_i = y_i - X_i\hat{\beta}$) is a consistent estimator regardless of whether there is heteroscedasticity in ϵ. False.

L. If the distribution of ϵ is double exponential, i.e. if $f(\epsilon) = \exp\{-|\epsilon|/\sigma\}/(2\sigma)$, the maximum likelihood estimator of β is the Least Absolute Deviations estimator and it is asymptotically efficient relative to the OLS estimator. True.

M. The OLS estimator cannot be used if the regression function is misspecified, i.e. if the true regression function $E\{y|X\} \neq X\beta$. False.

N. The OLS estimator will be inconsistent if ϵ and X are correlated. True.

O. The OLS estimator will be inconsistent if the dependent variable y is truncated, i.e. if the dependent variable is actually determined by the relation

$$y = \max[0, X\beta + \epsilon] \tag{6}$$

True.

P. The OLS estimator is inconsistent if ϵ has a Cauchy distribution, i.e. if the density of ϵ is given by

$$f(\epsilon) = \frac{1}{\pi(1+\epsilon^2)} \tag{7}$$

True.

Q. The 2-stage least squares estimator is a better estimator than the OLS estimator because it has two stages and is therefore twice as efficient. False.

R. If the set of instrumental variables W and the set of regressors X in the linear model coincide, then 2 stage least squares estimator of β is the same as the OLS estimator of β. True.

Part II: 30 minutes, 30 points. Answer 2 of the following 6 questions below.

QUESTION 1 (Probability question) Suppose $\tilde{Z}$ is a $K \times 1$ random vector with a multivariate $N(0,I)$ distribution, i.e. $E\{\tilde{Z}\} = 0$ where 0 is a $K \times 1$ vector of zeros and $E\{\tilde{Z}\tilde{Z}'\} = I$ where I is the $K \times K$ identity matrix. Let M be a $K \times K$ idempotent matrix, i.e. a matrix that satisfies

$$M^2 = M * M = M \tag{8}$$

Show that

$$\tilde{Z}'M\tilde{Z} \sim \chi^2(J) \tag{9}$$

where $\chi^2(J)$ denotes a chi-squared random variable with J degrees of freedom and $J = \text{rank}(M)$. **Hint:** Use the fact that M has a singular value decomposition, i.e.

$$M = XDX' \tag{10}$$

where $X'X = I$ and D is a diagonal matrix whose diagonal elements are equal to either 1 or 0.

Answer: Let X be the $K \times K$ orthonormal matrix in the singular value decompostion of the idempotent matrix M. Since $X'X = I$, it follows that $\tilde{W} \equiv X'\tilde{Z}$ is $N(0,I)$. Thus, $\tilde{Z}'M\tilde{Z}$ can be rewritten as $(X'\tilde{Z})'D(X'\tilde{Z})$. Since D is a diagonal matrix with J 1's $K-J$ 0's on its main diagonal, it follows that $(X'\tilde{Z})'D(X'\tilde{Z}) = \tilde{W}'D\tilde{W}$ algebraically the sum of J *IDD* $N(0,1)$ random variables and thus has a $\chi^2(J)$ distribution. That is, assuming without loss of generality that the first J elements of the diagonal of D are 1's and the remaining $K-J$ elements are 0's we have

$$\tilde{Z}'M\tilde{Z} = (X'\tilde{Z})'D(X'\tilde{Z}) = \tilde{W}'D\tilde{W} = \tilde{W}_1 + \cdots + \tilde{W}_J. \tag{11}$$

Since the $\{\tilde{W}_j\}$ are *IID* $N(0,1)$'s, the result follows.

QUESTION 2 (Markov Processes)

A. (10%) Are Markov processes of any use in econometrics? Describe some examples of how Markov processes are used in econometrics such as providing models of serially dependent data, as a framework for establishing convergence of estimators and proving laws of large numbers, central limit theorems, etc. and as computational tool for doing simulations.

B. (10%) What is a random walk? Is a random walk always a Markov process? If not, provide a counter-example.

C. (40%) What is the ergodic or invariant distribution of a Markov process? Do all Markov processes have invariant distributions? If not, provide a counterexample of a Markov process that doesn't have an invariant distribution. Can a Markov process have more than 1 invariant distribution? If so, give an example.

D. (40%) Consider the discrete Markov process $\{X_t\} = \{1, 2, 3\}$ with transition probability

$$
\begin{aligned}
P\{X_{t+1} = 1|X_t = 1\} &= \frac{1}{2} \quad P\{X_{t+1} = 2|X_t = 1\} = \frac{1}{3} \quad P\{X_{t+1} = 3|X_t = 1\} = \frac{1}{6} \\
P\{X_{t+1} = 1|X_t = 2\} &= \frac{3}{4} \quad P\{X_{t+1} = 3|X_t = 2\} = \frac{1}{4} \quad P\{X_{t+1} = 2|X_t = 3\} = 1
\end{aligned}
$$

Does this process have an invariant distribution? If so, find all of them.

ANSWERS:

A. Markov processes play a major role in econometrics, since they it provides one of the simplest yet most general frameworks for modeling temporal dependence. Markov processes are used extensively in time series econometrics, since there are laws of large numbers and central limit theorems that apply to very general classes of Markov processes that satisfy a "geometric ergodicity" condition. Markov processes are also used extensively in Gibbs Sampling, which is a technique for simulating draws from a posterior distribution in econometric models where the posterior has no convenient analytical solution.

B. A random walk $\{X_t\}$ is a special type of Markov process that is represented as

$$X_t = X_{t-1} + \epsilon_t, \tag{12}$$

where $\{\epsilon_t\}$ is an *IID* process that is independent of $\{X_t\}$. If $E\{\epsilon_t\} > 0$ the random walk has *positive drift* and if $E\{\epsilon_t\} < 0$ it has *negative drift.* A random walk is always a Markov process, since X_{t-1} is a sufficient statistic for determining the probability distribution of X_t, and previous values $\{X_{t-2}, X_{t-3}, \ldots,\}$ are irrelevant. If F is the CDF for ϵ_t, then the Markov transition probability for $\{X_t\}$ is given by

$$\Pr\{X_t \leq x'|X_{t-1} = x\} = F(x' - x). \tag{13}$$

C. If a Markov Process has a transition probabilty $P(x'|x)$, then its invariant distribution Π is defined by

$$\Pi(x') = \int P(x'|x)\Pi(dx). \tag{14}$$

What this equation says is that if $X_t \sim \Pi(x)$ (i.e. X_t is distributed according to the probability distribution Π), then X_{t+1} is also distributed according to this same probability distribution. Not all Markov processes have invariant distributions. A random walk does not have an invariant distribution, i.e. there is no solution to the equation (14) above. To see this, note in particular that due to the independence between X_{t-1} and ϵ_t we have

$$\operatorname{var}(X_t) = \operatorname{var}(X_{t-1} + \epsilon_t) = \operatorname{var}(X_{t-1}) + \operatorname{var}(\epsilon_t) > \operatorname{var}(X_{t-1}), \tag{15}$$

so that regardless of what distribution X_{t-1} has, it is impossible for X_t to have this same distribution.

D. The transition probability matrix P for this process is given by the following 3×3 matrix

$$P = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 3/4 & 0 & 1/4 \\ 0 & 1 & 0 \end{bmatrix} \tag{16}$$

The invariant probability is the solution Π to the 3×3 system of equations

$$\Pi = \Pi P \tag{17}$$

We can write this out as

$$\begin{aligned} \pi_1 &= \tfrac{1}{2}\pi_1 + \tfrac{3}{4}\pi_2 + 0\pi_3 \\ \pi_2 &= \tfrac{1}{3}\pi_1 + 0\pi_2 + \pi_3 \\ \pi_3 &= \tfrac{1}{6}\pi_1 + \tfrac{1}{4}\pi_2 + 0\pi_3 \end{aligned} \tag{18}$$

You can verify the the unique non-zero solution to the above system of equations is $(\pi_1, \pi_2, \pi_3) = (1/2, 1/3, 1/6)$, i.e. the unique invariant distribution is the same as the first row of P.

QUESTION 3 (Consistency of M-estimator) Consider an M-estimator defined by:

$$\widehat{\theta}_N = \arg\max_{\theta \in \Theta} Q_N(\theta).$$

Suppose following two conditions are given
(i) (Identification) For all $\varepsilon > 0$

$$Q(\theta^*) > \sup_{\theta \notin B(\theta^*, \varepsilon)} Q(\theta)$$

where $B(\theta^*, \varepsilon) = \{\theta \in R^k | \|\theta - \theta^*\| < \epsilon\}$.
(ii) (Uniform Convergence)

$$\sup_{\theta \in \Theta} |Q_N(\theta) - Q(\theta)| \xrightarrow{p} 0.$$

Prove consistency of the estimator by showing

$$P\left(\widehat{\theta}_N \notin B(\theta^*, \varepsilon)\right) \to 0.$$

ANSWER Uniform convergence in probability can be stated formally as follows: for any $\delta > 0$ we have

$$\lim_{N\to\infty} \Pr\left\{\sup_{\theta\in\Theta} |Q_N(\theta) - Q(\theta)| < \delta\right\} = 1. \tag{19}$$

Now, given any $\varepsilon > 0$, define δ by

$$\delta \equiv Q(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q(\theta) \tag{20}$$

The identification assumption implies that $\delta > 0$. Now, we want to show that for any $\varepsilon > 0$ we have

$$\lim_{N\to\infty} \Pr\left\{\widehat{\theta}_N \notin B(\theta^*,\varepsilon)\right\} = 0. \tag{21}$$

Notice if $\widehat{\theta}_N \notin B(\theta^*,\varepsilon)$ then we have

$$Q_N(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) \leq 0. \tag{22}$$

So it is sufficient to show that uniform convergence implies that

$$\lim_{N\to\infty} \Pr\left\{Q_N(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) \leq 0\right\} = 0. \tag{23}$$

Using the δ defined in equation (20) and the definition of uniform convergence in probability in equation (19), we have

$$\lim_{N\to\infty} \Pr\left\{\sup_{\theta\in\Theta} |Q_N(\theta) - Q(\theta)| < \delta/3\right\} = 1. \tag{24}$$

Thus, for N sufficiently large, the following inequalities will hold with probability arbitrarily close to 1,

$$\begin{array}{rcl} Q_N(\theta^*) & > & Q(\theta^*) - \delta/3 \\ \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) & < & \sup_{\theta\notin B(\theta^*,\varepsilon)} Q(\theta) + \delta/3 \end{array} \tag{25}$$

Combining the above inequalities, it follows that the following inequality will hold with probability arbitrarily close to 1 for N sufficiently large:

$$Q_N(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) > Q(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q(\theta) - \frac{2\delta}{3} = \frac{\delta}{3}. \tag{26}$$

This implies that

$$\lim_{N\to\infty} \Pr\left\{Q_N(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) > \frac{1\delta}{3}\right\} = 1. \tag{27}$$

Since δ is arbitrary, this implies that

$$\lim_{N\to\infty} \Pr\left\{Q_N(\theta^*) - \sup_{\theta\notin B(\theta^*,\varepsilon)} Q_N(\theta) \leq 0\right\} = 0. \tag{28}$$

Since the event that $\widehat{\theta}_N \notin B(\theta^*, \varepsilon)$ is a subset of the event that $Q_N(\theta^*) - \sup_{\theta \notin B(\theta^*, \varepsilon)} Q_N(\theta) \leq 0$, it follows that the limit in equation (19) holds, i.e. $\hat{\theta}_N$ is a consistent estimator of θ^*.

QUESTION 4 (Time series question) Suppose $\{X_t\}$ is an ARMA(p,q) process, i.e.

$$A(L)X_t = B(L)\epsilon_t$$

where $A(L)$ is a q^{th} order lag-polynomial

$$A(L) = \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \cdots + \alpha_q L^q$$

and $B(L)$ is a p^{th} order lag-polynomial

$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_p L^p$$

and the lag-operator L^k is defined by

$$L^k X_t = X_{t-k}$$

and $\{\epsilon_t\}$ is a white-noise process, $E\{\epsilon_t\} = 0$ and ($\text{cov}(\epsilon_t, \epsilon_s)$= 0 if $t \neq s$, $= \sigma^2$ if $t = s$).

A. (30%) Write down the autocovariance and spectral density functions for this process.

B. (30%) Show that if $p = 0$ an autoregression of X_t on q lags of itself provides a consistent estimate of $(\alpha_0/\sigma, \ldots, \alpha_q/\sigma)$. Is the autoregression still consistent if $p > 0$?

C. (40%) Assume that a central limit theorem holds, i.e. the distribution of normalized sums of $\{X_t\}$ to converge in distribution to a normal random variable. Write down an expression for the variance of the limiting normal distribution.

ANSWERS

A. The answer to this question is very complicated if you attempt to proceed via direct calculation (although it can be done), but it much easier if you use the concept of a *z-transform* and the *covariance generating function* $G(z)$ of the scalar process $\{X_t\}$. The answer is that spectral density function $f(\lambda)$ for the $\{X_t\}$ process is given by

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|B(e^{-i\lambda})|^2}{|A(e^{-i\lambda})|^2} \tag{29}$$

provided the characteristic polynomial $A(z) = 0$ has no roots on the unit circle. The autocovariances of the $\{X_t\}$ process can then be derived from the spectral density via the formula

$$\text{cov}(X_t, X_{t+k}) = \gamma_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda k} d\lambda. \tag{30}$$

Answering this question presumes a basic familiarity with Fourier transform technology. I repeat the basics of this below.

Given a sequence of real numbers $\{\psi_k\}$ where k ranges from $-\infty, \ldots, \infty$ the z-transform $G(z)$ is defined by

$$G(z) = \sum_{k=-\infty}^{\infty} \psi_k z^k \tag{31}$$

where z is a complex variable satisfying $r^{-1} < |z| < r$ for some $r > 1$. The autocovariance generating function is then just the z-transform of the autocovariance sequence $\{\gamma_k\}$:

$$G(z) \equiv \sum_{k=-\infty}^{\infty} \gamma_k z^k \tag{32}$$

where $\gamma_k = \text{cov}(X_t, X_{t+k}) = E\{X_t X_{t+k}\}$. Thus, if we can find a representation for $G(z)$, we can pick off the autocovariances γ_k as the coefficient of z^k of the power series representation for $G(z)$. Alternatively we can define the *spectral density* $f(\lambda)$ for the $\{X_t\}$ process by

$$f(\lambda) = \sum_{k=-\infty}^{\infty} \gamma_k e^{-i\lambda k} \tag{33}$$

where $i = \sqrt{-1}$. Note that for by the standard properties of Fourier series, we can uncover the autocovariance γ_k by the formula:

$$\gamma_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda k} d\lambda. \tag{34}$$

This is due to the fact that the sequence of complex valued functions $\{e^{i\lambda k}\}$ mapping $[-\pi, \pi]$ to the unit circle in the complex plane are an orthogonal sequence under the complex inner product for complex-valued functions mapping $[-\pi, \pi] \to C$ defined by:

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(\lambda) \overline{g}(\lambda) d\lambda, \tag{35}$$

where $\overline{g}(\lambda)$ is the complex conjugate of $g(\lambda)$. Since the complex conjugate of $e^{i\lambda k}$ is $e^{-i\lambda k}$ we have

$$\langle e^{i\lambda j}, e^{i\lambda k} \rangle \equiv \int_{-\pi}^{\pi} e^{i\lambda j} e^{-i\lambda k} d\lambda = \int_{-\pi}^{\pi} e^{i\lambda(j-k)} d\lambda. \tag{36}$$

Clearly if $j = k$ then we have

$$\langle e^{i\lambda j}, e^{i\lambda k} \rangle = \int_{-\pi}^{\pi} d\lambda = 2\pi, \tag{37}$$

but if $j \neq k$ we have, using the identity $e^{i\lambda} = \cos(\lambda) + i \sin(\lambda)$,

$$\langle e^{i\lambda j}, e^{i\lambda k} \rangle = \int_{-\pi}^{\pi} \cos(\lambda(j-k)) + i \sin(\lambda(j-k)) d\lambda = 0. \tag{38}$$

since $\sin(k\lambda)$ and $\cos(k\lambda)$ are periodic functions for any non-zero integer k, their integrals over the interval $[-\pi, \pi]$ are zero. Thus, since $\{e^{i\lambda k}\}$ is an orthogonal family, γ_k is essentially the k^{th} regression coefficient if we "regress" the spectral density function against the sequence of orthonormal basis functions $\{e^{i\lambda k}\}$,

$$\langle f(\lambda), e^{i\lambda k} \rangle = \int_{-\pi}^{\pi} f(\lambda) e^{i\lambda k} d\lambda = \int_{-\pi}^{\pi} \sum_{j=-\infty}^{\infty} \gamma_k e^{-i\lambda j} e^{i\lambda k} d\lambda = 2\pi \gamma_k. \tag{39}$$

Solving the above equation for γ_k results in the Fourier inversion formula in equation (34) above. Note also that the spectral density is related to the covariance generating function by the identity

$$f(\lambda) = G(e^{-i\lambda}), \tag{40}$$

so the problem reduces to finding an expression for the covariance generating function for an ARMA(p,q) process. Assume that the *characteristic polynomial* $A(z)$ has no roots on the unit circle, i.e. there is no complex number z with $|z| = z\overline{z} = 1$ such that $A(z) = 0$. In this case it can be show (see Theorem 3.1.3 of Brockwell and Davis, 1991), that the ARMA process $\{X_t\}$ has an infinite moving average representation:

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \epsilon_{t-j}, \tag{41}$$

where ψ_j is the j^{th} coefficient in the power series representation of the z-transform of the $\{\psi_j\}$ sequence, where the z-transform $\Psi(z)$ is given by

$$\Psi(z) = B(z)A(z)^{-1}. \tag{42}$$

However covariance generating function for an infinite MA process (41) can be derived as follows:

$$\gamma_k = \mathrm{cov}(X_{t+k}, X_t) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|k|} \tag{43}$$

Thus, the autocovariance generating function is given by

$$\begin{aligned} G(z) &= \sum_{k=-\infty}^{\infty} \gamma_k z^k \\ &= \sigma^2 \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|k|} \\ &= \sigma^2 \left[\sum_{j=-\infty}^{\infty} \psi_j^2 + \sum_{k=1}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+k} (z^k + z^{-k}) \right] \\ &= \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j z^j \right) \left(\sum_{k=-\infty}^{\infty} \psi_k z^{-k} \right) \\ &= \sigma^2 \Psi(z) \Psi(z^{-1}). \end{aligned} \tag{44}$$

However using the fact that $\Psi(z) = B(z)A(z)^{-1}$ it follows that

$$G(z) = \sigma^2 \frac{B(z)B(z^{-1})}{A(z)A(z^{-1})} \tag{45}$$

Substituting $z = e^{-i\lambda}$ we obtain

$$f(\lambda) = G(e^{i\lambda}) = \sigma^2 \frac{B(e^{i\lambda})B(e^{-i\lambda})}{A(e^{i\lambda})A(e^{-i\lambda})} = \sigma^2 \frac{|B(e^{i\lambda})|^2}{|A(e^{i\lambda})|^2} \tag{46}$$

since $A(e^{-i\lambda}) = \overline{A}(e^{i\lambda})$ and thus $A(e^{i\lambda})A(e^{-i\lambda}) = |A(e^{-i\lambda})|^2$ and similarly for B.

B. When $q = 0$ we can write the ARMA representation for $\{X_t\}$ in autoregressive form:

$$X_t = \frac{\alpha_1}{\alpha_0} X_{t-1} + \cdots \frac{\alpha_q}{\alpha_0} X_{t-q} + \frac{\beta_0}{\alpha_0} \varepsilon_t \tag{47}$$

Since $\{\varepsilon_t\}$ is serially uncorrelated, and since X_{t-j} depends only on lagged values $(\varepsilon_{t-j}, \varepsilon_{t-j-1}, \ldots)$, it follows that $\mathrm{cov}(\varepsilon_t, X_{t-j}) = 0$ so the coefficients α_j/α_0 and the error variance $\beta_0^2\sigma^2/\alpha_0^2$ in the above equation can be consistently estimated by OLS. We cannot identify all the parameters unless we make an identifying normalization on the variance of the white noise process such as $\sigma^2 = 1$, or normalize $\beta_0 = 1$. Suppose we make the latter normalization. Then the variance of the estimated residuals provides a consistent estimator of σ^2/α_0, and then dividing the estimated regression coefficient for the j^{th} lag of X_t in the above autoregression by the square root of the estimated variance of the residuals provides a consistent estimator of α_j/α_0.

C. Since $E\{X_t\} = 0$ then under suitable mixing conditions a central limit theorem will hold, i.e.

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \underset{d}{\Rightarrow} N(0, \Omega) \tag{48}$$

where Ω is the long run variance given by

$$\Omega = \sum_{j=-\infty}^{\infty} \gamma_j \tag{49}$$

where $\gamma_j = \mathrm{cov}(X_t, X_{t+j})$ is the autocovariance at lag j, which can be derived from the spectral density function computed in part A.

QUESTION 5 (Empirical question) Assume that shoppers always choose only a single brand of canned tuna fish from the available selection of K alternative brands of tuna fish each time they go shopping at a supermarket. Assume initially that the (true) probability that the decision-maker chooses brand k is the same for everybody and is given by θ_k^*, $k = 1, \ldots, K$. Marketing researchers would like to learn more about these choice probabilities, $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$ and spend a great deal of money sampling shoppers' actual tuna fish choices in order to try to estimate these probabilities. Suppose the Chicken of the Sea Tuna company undertook a survey of N shoppers and for each shopper shopping at a particular supermarket with a fixed set of K brands of tuna fish, recorded the brand b_i chosen by shopper i, $i = 1, \ldots, N$. Thus, $b_1 = 2$ denotes the observation that consumer 1 chose tuna brand 2, and $b_4 = K$ denotes the observation that consumer 4 chose tuna brand K, etc.

A. (10%) Without doing any estimation, are there any general restrictions that you can place on the $K \times 1$ parameter vector θ^*?

Answer: we must have $\theta_j^* \geq 0$ and $\sum_{j=1}^{K} \theta_j^* = 1$.

B. (10%) Is it reasonable to suppose that θ_k^* is the same for everyone? Describe several factors that could lead different people to have different probabilities of purchasing different brands of tuna. If you were a consultant to Chicken of the Sea, what additional data would you recommend that they collect in order to better predict the probabilities that consumers buy various brands of tuna? Describe how you would use this data once it was collected.

Answer: no, it is quite unreasonable to assume that everyone has the same purchase probability. People of different ages, income levels, ethnic backgrounds and so forth are likely to have different tastes for tuna. Also, Chicken of the Sea is just one of many different brands of tuna and the prices of the competing brands and observed characteristics of the competing brands (such as whether the tuna is packed in oil or water, the consistency of the tuna, and other characteristics) affects the probability a given consumer will choose Chicken of the Sea. Let the vector of observed characteristics for the K brands be given by the $L \times K$ matrix $Z = (Z_1, \ldots, Z_K)$ (i.e. there are L observed characteristics for each of the K different brands). Let the characteristics of the j^{th} household be denoted by the $M \times 1$ vector X_j. Then a model that reflects observed heterogeneity and the competing brand characteristics would result in the following general form of the *conditional probability* that household j will choose brand k from the set of competing tuna brands offered in the store at time of purchase, $\Pr(k|X_j, Z)$. An example of a model of consumer choice behavior that results in a specific functional form for $\Pr(k|X_j, Z)$ is the *multinomial logit model.* This is a model derived from a model of utility maximization where the utility of choosing brand k is given by $u(X_j, Z_k, \theta) + \epsilon_k$, where $(\epsilon_1, \ldots, \epsilon_K)$ are unobserved factors affecting household j's decision, and are assumed to have a Type III extreme value distribution. In this case, the implied formula for $\Pr(k|X_j, Z)$ is given by

$$\Pr(k|X_j, Z, \theta, \sigma) = \frac{\exp\{u(X_j, Z_k, \theta)/\sigma\}}{\sum_{k'=1}^{K} \exp\{u(X_j, Z_{k'}, \theta)/\sigma\}} \tag{50}$$

where σ is the scale parameter in the marginal distribution of ϵ_k. Thus, given data $(X_1, \ldots, X_N)$ on the characteristics of N consumers, and their choices of tuna $(d_1, \ldots, d_N)$ and the observed characteristics Z, we could estimate the parameter vector θ by maximum likelihood using the log-likelihood function $L_N(\theta)$ given by

$$L_N(\theta) = \frac{1}{N} \sum_{j=1}^{N} \log\left(\Pr(d_j|X_j, Z, \theta, \sigma)\right). \tag{51}$$

and the estimated model could be used to predict how the probabilities of purchasing different brands of tuna (and the predicted aggregate market shares) change in response to changes in prices or observed characteristics of the different brands of tuna.

C. (20%) Using the observations $\{b_1, \ldots, b_N\}$ on the observed brand choices of the sample of N shoppers, write down an estimator for θ^* (under the assumption that the "true" brand choice probabilities θ^* are the same for everyone). Is your estimator unbiased?

Answer: In the simpler case where there are no characteristics X_j or product attributes Z, the the choice probability can be represented by a single parameter, $\Pr(k|X_j, Z, \theta) = \theta_k$. These θ_k are also the observed market shares since everyone is homogeneous. The market share for brand k can be estimated in this sample as the fraction of the N people who choose brand k,

$$s_k = \frac{1}{N} \sum_{i=1}^{N} I\{b_i = k\} \tag{52}$$

Thus if s_k is the observed market share for product k, then we can estimate θ_k by $\hat{\theta}_k = s_k$.

D. (20%) What is the maximum likelihood estimator of θ^*? Is the maximum likelihood estimator unbiased?

Answer: The likelihood function in this case can be written as

$$L_N(\theta) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} I\{b_j = k\}\log(\theta_k). \tag{53}$$

subject to the constraint that $1 = \theta_1 + \cdots + \theta_K$. Introducing a lagrange multiplier λ for this constraint, the lagrangian for the likelihood function is

$$\mathcal{L}(\theta,\lambda) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} I\{b_j = k\}\log(\theta_k) + \lambda(1 - \sum_{k=1}^{K}\theta_k) \tag{54}$$

The first order conditions are

$$\frac{1}{N}\sum_{i=1}^{N}\frac{I\{b_i = k\}}{\theta_k} - \lambda = 0. \tag{55}$$

Solving this for $\hat{\theta}_k$ and substituting this into the constraint, we can solve for λ, obtaining $\hat{\lambda} = 1$. The resulting estimator is the same as the intuitive market share estimator given above, i.e.

$$\hat{\theta}_k = s_k \tag{56}$$

If the data $\{b_1, \ldots, b_N\}$ are really *IID* and the "representative consumer" model is really correct, then $\hat{\theta}_k$ is an unbiased estimator of θ_k^* since

$$E\{\hat{\theta}_k\} = \frac{1}{N}\sum_{i=1}^{N} E\left\{I\{b_i = k\}\right\} = \theta_k^* \tag{57}$$

since the random variable $I\{b_i = k\}$ is a bernoulli random variable which equals 1 with probability θ_k^* and 0 with probability $1 - \theta_k^*$.

E. (40%) Suppose Chicken of the Sea Tuna company also collected data on the prices $\{p_1, \ldots, p_K\}$ that the supermarket charged for each of the K different brands of tuna fish. Suppose someone proposed that the probability of buying brand j was a function of the prices of all the various brands of tuna, $\theta_j^*(p_1, \ldots, p_K)$, given by:

$$\theta_j^*(p_1, \ldots, p_K) = \frac{\exp\{\beta_j + \alpha p_j\}}{\sum_{k=1}^{K}\exp\{\beta_k + \alpha p_k\}}$$

Describe in general terms how to estimate the parameters $(\alpha, \beta_1, \ldots, \beta_K)$. If $\alpha > 0$, does an increase in p_j decrease or increase the probability that a shopper would buy brand j?

Answer: This answer was already discussed in the answer to part B. The model is a special case of the more general multinomial logit model discussed in the answer to part B. In this case the implicit utility function only depends on the single characteristic of brand k, namely its price p_k and the other characteristics of the brand are implicitly captured in the brand-specific dummy variable β_k. Since now consumer-level characteristics enter the model, the utility function is given by

$$u(X_j, Z_k, \theta) = \beta_k + \alpha p_k \tag{58}$$

where $\theta = (\beta_1, \ldots, \beta_K, \alpha)$. If $\alpha > 0$ then the utility of brand k increases in the price of the brand k, an economically counter-intuitive result. This sugggests that the probability of purchasing brand k is an increasing function of p_k and this can be verified by computing

$$\frac{\partial \Pr}{\partial p_K}(k|p_1, \ldots, p_K, \theta) = \alpha \Pr(k|p_1, \ldots, p_K, \theta)[1 - \Pr(k|p_1, \ldots, p_K, \theta)] > 0. \tag{59}$$

QUESTION 6 (Regression question) Let (y_t, x_t) be *IID* observations from a regression model

$$y_t = \beta x_t + \epsilon_t$$

where y_t, x_t, and ϵ_t are all scalars. Suppose that ϵ_t is normally distributed with $E\{\epsilon_t|x_t\} = 0$, but $\text{var}(\epsilon_t|x_t) = \sigma^2|x_t|^\theta$. Consider the following two estimators for β^*:

$$\hat{\beta}_T^1 = \frac{\sum_{t=1}^T y_t}{\sum_{t=1}^T x_t}$$

$$\hat{\beta}_T^2 = \frac{\sum_{t=1}^T x_t y_t}{\sum_{t=1}^T x_t^2}$$

A. (20%) Are these two estimators consistent estimators of β^*? Which estimator is more efficient when: 1) if we know *a priori* that $\theta^* = 0$, and 2) we don't know θ^*? Explain your reasoning for full credit.

Answer: Both estimators are consistent estimators of β^*. To see this note that by dividing the numerator and denominator of $\hat{\beta}_T^1$ and applying the Law of Large Numbers we obtain

$$\hat{\beta}_T^1 \underset{p}{\longrightarrow} \frac{E\{y\}}{E\{x\}} = \frac{\beta^* E\{x\}}{E\{x\}} = \beta^*. \tag{60}$$

The second estimator is the OLS estimator and it is also a consistent estimator of β

$$\hat{\beta}_T^2 \underset{p}{\longrightarrow} \frac{E\{xy\}}{E\{x^2\}} = \frac{\beta^* E\{x^2\}}{E\{x^2\}} = \beta^*. \tag{61}$$

When $\theta^* = 0$ the Gauss-Markov Theorem applies and the OLS estimator is the best linear unbiased estimator of β^*. It is also the maximum likelihood estimator when the errors are normally distributed, and so is asymptotically efficient in the class of all (potentially nonlinear) regular estimators of β^*. We can derive the asymptotic efficiency of $\hat{\beta}_T^2$ relative to $\hat{\beta}_T^1$ through a simple application of the central limit theorem. We have

$$\sqrt{T}(\hat{\beta}_T^1 - \beta^*) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T \epsilon_t}{\frac{1}{T} \sum_{t=1}^T x_t} \underset{d}{\Longrightarrow} \frac{\tilde{Z}}{E\{x\}} \sim N\left(0, \frac{\sigma^2}{E\{x\}^2}\right). \tag{62}$$

where $\tilde{Z} \sim N(0, \sigma^2)$. Similarly, the asymptotic distribution of the OLS estimator $\hat{\beta}_T^2$ is given by

$$\sqrt{T}(\hat{\beta}_T^2 - \beta^*) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t \epsilon_t}{\frac{1}{T} \sum_{t=1}^T x_t^2} \underset{d}{\Longrightarrow} \frac{\tilde{W}}{E\{x^2\}} \sim N\left(0, \frac{\sigma^2}{E\{x^2\}}\right). \tag{63}$$

where $\tilde{W} \sim N(0, \sigma^2 E\{x^2\})$. If the variance of $\tilde{x}$ is positive we have

$$\begin{aligned} \text{var}(\tilde{x}) &= E\{x^2\} - E\{x\}^2 > 0 \\ &\Longrightarrow E\{x^2\} > E\{x\}^2. \end{aligned} \tag{64}$$

This implies that the asymptotic variance of $\hat{\beta}_T^1$ is greater than the asymptotic variance of $\hat{\beta}_T^2$.

In the case where we don't know θ we can repeat the calculations given above, but the asymptotic distributions of the two estimators will depend on the unknown parameter θ^*. In particular, when $\theta^* \neq 0$ the unconditional variance of ϵ_t is given by

$$\text{var}(\epsilon_t) = E\{\text{var}(\epsilon_t | x_t)\} = \sigma^2 E\{|x|^{\theta^*}\}. \tag{65}$$

This implies that

$$\sqrt{T}(\hat{\beta}_T^1 - \beta^*) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T \epsilon_t}{\frac{1}{T} \sum_{t=1}^T x_t} \underset{d}{\Longrightarrow} \frac{\tilde{Z}}{E\{x\}} \sim N\left(0, \frac{\sigma^2 E\{|x|^{\theta^*}\}}{[E\{x\}]^2}\right). \tag{66}$$

since with heteroscedasticity, the random variable $\tilde{Z}$, the asymptotic distribution of $1/\sqrt{T} \sum_{t=1}^T \epsilon_t$, is $N(0, \sigma^2 E\{|x|^{\theta^*}\})$ instead of $N(0, \sigma^2)$. Similarly we have

$$\sqrt{T}(\hat{\beta}_T^2 - \beta^*) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t \epsilon_t}{\frac{1}{T} \sum_{t=1}^T x_t^2} \underset{d}{\Longrightarrow} \frac{\tilde{W}}{E\{x^2\}} \sim N\left(0, \frac{\sigma^2 E\{x^2 |x|^{\theta^*}\}}{[E\{x^2\}]^2}\right). \tag{67}$$

In this case, which of the two estimators $\hat{\beta}_T^1$ or $\hat{\beta}_T^2$ is more efficient depends on the value of θ^*.

B. (20%) Write down an asymptotically optimal estimator for β^* if we know the value of θ^* *a priori.*

   Answer: If we know θ^* we can do maximum likelihood using the conditional density of y given x given by

$$f(y|x, \beta, \theta^*) = \frac{1}{\sqrt{2\pi}\sigma |x|^{\theta^*/2}} \exp\left\{\frac{-(y - x\beta)^2}{|x|^{\theta^*}}\right\}. \tag{68}$$

   The maximum likeihood estimator in this case can be easily shown to be a form of weighted least squares:

$$\hat{\beta}_T = \underset{\beta \in R}{argmin} \sum_{t=1}^T \frac{(y_t - x_t\beta)^2}{|x_t|^{\theta^*}}. \tag{69}$$

C. (20%) Write down an asymptotically optimal estimator for (β^*, θ^*) if we don't know the value of θ^* *a priori.*

Answer: If we don't know θ^* *a priori* we can still use the likelihood function given in part B to estimate (β, θ) jointly. The maximum likelihood estimator for β can also be cast as a weighted least squares estimator, but in the case where θ^* is not known we replace θ^* in formula (69) by $\theta(\beta)$, where this is the unique solution to

$$\sum_{t=1}^{T} \log(|x_t|) = \sum_{t=1}^{T} \frac{(y_t - x_t\beta)^2 \log(|x_t|)}{|x_t|^{\theta}}. \tag{70}$$

The maximum likelihood estimator for θ is then given by $\theta(\hat{\beta}_T)$ where $\hat{\beta}_T$ is the weighted least squares estimator given above.

D. (20%) Describe the feasible GLS estimator for (β^*, θ^*). Is the feasible GLS estimator asymptotically efficient?

Answer: The feasible GLS estimator is based on an initial inefficient estimator $\hat{\beta}_T$ of β^* which is used to construct estimated residuals $\hat{\epsilon}_t = (y_t - x_t\hat{\beta}_T)$ and from these an estimator for θ^*. If we could observe the true residuals we could estimate θ^* via the following nonlinear regression of ϵ_t^2 on x_t

$$\epsilon_t^2 = \sigma^2 |x_t|^{\theta^*} + u_t \tag{71}$$

where $E\{u_t|x_t\} = 0$. This suggests that it should be possible to estimate θ^* using the estimated residuals $\{\hat{\epsilon}_t\}$ as follows

$$\hat{\theta}_T = \underset{\theta\in R, \sigma^2>0}{argmin} \sum_{t=1}^{T} (\hat{\epsilon}_t^2 - \sigma^2 |x_t|^{\theta})^2. \tag{72}$$

It can be shown that if the initial estimator $\hat{\beta}_T$ is $\sqrt{T}$-consistent, then the nonlinear least squares estimator for θ^* given above will also be $\sqrt{T}$-consistent, and that the following three step, feasible GLS estimator for β^* will be asymptotically efficient:

$$\hat{\beta}_T^{\mathrm{f}} = \underset{\beta\in R}{argmin} \sum_{t=1}^{T} \frac{(y_t - x_t\beta)^2}{|x_t|^{\hat{\theta}_T}}. \tag{73}$$

E. (20%) How would your answers to parts A to D change if you didn't know the distribution of ϵ_t was normal?

Answer: The answer to part A is unchanged. However if we don't know the form of the conditional distribution of ϵ_t given x_t, we can't write down a likelihood function that will determine the asymptotically optimal estimator for β^*, regardless of whether we know θ^* or not. Thus, there is no immediate answer to parts B and C. In part D we can still do the same feasible GLS estimator, and while it is possible to show that this is asymptotically efficient relative to OLS, it is not clear that it is asymptotically optimal. There is a possibility of doing *adaptive estimation,* i.e. of using a first stage inefficient estimator of β^* to construct estimated residuals $\hat{\epsilon}_t$ and then using these estimated residuals to try to estimate the conditional density $f(\epsilon|x)$ non-parametrically. Then using this nonparamtric distribution we could do maximum likelihood. Unfortunately the known results for this sort of adaptive estimation procedure requires that the error term ϵ_t be independent of x_t. However if there is heteroscedasticity, then ϵ_t will not be independent of x_t and adaptive estimation may not be possible. In this case the most efficient possible estimator can be ascertained by deriving the *semi-parametric efficiency bound* for the parameter of interest β, where the conditional density $f(\epsilon|x)$ is treated as a non-parametric "nuisance parameter". However this goes far beyond what I expected students to write in the answer to this exam.

Part III (60 minutes, 55 points). Do 1 out of the 4 questions below.

QUESTION 1 (Hypothesis testing) Consider the GMM estimator with *IID* data, i.e the observations $\{y_i, x_i\}$ are independent and identically distributed using the moment condition $H(\theta) = E\{h(\tilde{y}, \tilde{x}, \theta)\}$, where h is a $J \times 1$ vector of moment conditions and θ is a $K \times 1$ vector of parameters to be estimated. Assume that the moment conditions are correctly specified, i.e. assume there is a unique θ^* such that $H(\theta^*) = 0$. Show that in the overidentified case $(J > K)$ that the minimized value of the GMM criterion function is asymptotically χ^2 with $J - K$ degrees of freedom:

$$N H_N(\hat{\theta}_N)'[\hat{\Omega}_N]^{-1} H_N(\hat{\theta}_N) \underset{d}{\Longrightarrow} \chi^2(J-K), \tag{74}$$

where H_N is a $J \times 1$ vector of moment conditions, θ is a $K \times 1$ vector of parameters, $\chi^2(J-K)$ is a Chi-squared random variable with $J - K$ degrees of freedom,

$$\hat{\theta}_N = \underset{\theta \in \Theta}{argmin}\, H_N(\theta)'[\hat{\Omega}_N]^{-1} H_N(\theta),$$

$$H_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} h(y_i, x_i, \theta),$$

and $\hat{\Omega}_N$ is a consistent estimator of Ω given by

$$\Omega = E\{h(\tilde{y}, \tilde{x}, \theta^*) h(\tilde{y}, \tilde{x}, \theta^*)'\}.$$

Hint: Use Taylor series expansions to provide a formula for $\sqrt{N}(\hat{\theta}_N - \theta^*)$ from the first order condition for $\hat{\theta}_N$

$$\nabla H_N(\hat{\theta}_N)' \hat{\Omega}_N^{-1} H_N(\hat{\theta}_N) = 0 \tag{75}$$

and a Taylor series expansion of $H_N(\hat{\theta}_N)$ about θ^*

$$H_N(\hat{\theta}_N) = H_N(\theta^*) + \nabla H_N(\tilde{\theta}_N)(\hat{\theta}_N - \theta^*) \tag{76}$$

where

$$\nabla H_N(\theta) \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\partial h}{\partial \theta}(y_i, x_i, \theta) \tag{77}$$

is the $(J \times K)$ matrix of partial derivatives of the moment conditions $H_N(\theta)$ with respect to θ and $\tilde{\theta}_N$ is a vector each of whose elements are on the line segment joining the corresponding components of $\hat{\theta}_N$ and θ^*. Use the above two equations to derive the following formula for $H_N(\hat{\theta}_N)$

$$H_N(\hat{\theta}_N) = M_N H_N(\theta^*) \tag{78}$$

where

$$M_N = \left[I - \nabla H_N(\hat{\theta}_N)[\nabla H_N(\hat{\theta}_N)'\hat{\Omega}_N^{-1}\nabla H_N(\tilde{\theta}_N)]^{-1}\nabla H_N(\hat{\theta}_N)'\hat{\Omega}_N^{-1}\right]. \tag{79}$$

Show that with probability 1 we have $M_N \to M$ where M is a $(J \times J)$ idempotent matrix. Then using this result, and using the Central Limit Theorem to show that

$$\sqrt{N} H_N(\theta^*) \underset{d}{\Longrightarrow} N(0, \Omega), \tag{80}$$

and using the probability result from Question 0 of Part II, show that the minimized value of the GMM criterion function does indeed converge in distribution to a $\chi^2(J-K)$ random variable as claimed in equation (74).

ANSWER: The hint provides most of the answer. Plugging the Taylor series expansion for $H_N(\hat{\theta}_N)$ given in equation (76) into the GMM first order condition given in equation (75) and solving for $(\hat{\theta}_N - \theta^*)$ we obtain

$$\hat{\theta}_N - \theta^* = -\left[\nabla H_N(\hat{\theta}_N)\hat{\Omega}_N^{-1}\nabla H_N(\tilde{\theta}_N)\right]^{-1}\nabla H_N(\hat{\theta}_N)\hat{\Omega}_N^{-1} H_N(\theta^*). \tag{81}$$

Substituting the above expression for $\hat{\theta}_N - \theta^*$ back into the Taylor series expansion for $H_N(\hat{\theta}_N)$ in equation (76) we obtain the representation for $H_N(\hat{\theta}_N)$ given in equations (78) and (79). Now we can write the optimized value of the GMM objective function as

$$\begin{aligned} H_N(\hat{\theta}_N)'[\hat{\Omega}_N^{-1}]^{-1}H_N(\hat{\theta}_N) &= H_N(\theta^*)'M_N'\hat{\Omega}_N^{-1}M_N H_N(\theta^*) \\ &= H_N(\theta^*)'\Omega^{-1/2}\Omega^{1/2}M_N'\hat{\Omega}_N^{-1/2}\hat{\Omega}_N^{-1/2}M_N\Omega^{1/2}\Omega^{-1/2}H_N(\theta^*) \end{aligned} \tag{82}$$

Now, since $\Omega = E\{h(\tilde{y},\tilde{x},\theta^*)h(\tilde{y},\tilde{x},\theta^*)'\}$, it follows from the Central Limit Theorem that

$$\sqrt{N} H_N(\theta^*) \underset{d}{\Longrightarrow} N(0, \Omega). \tag{83}$$

so that

$$\sqrt{N}\Omega^{-1/2} H_N(\theta^*) \underset{d}{\Longrightarrow} N(0, I), \tag{84}$$

where I is the $J \times J$ identity matrix. Now consider the matrix in the middle of the expansion of the quadratic form in equation (82). We have

$$\hat{\Omega}_N^{-1/2} M_N \Omega^{1/2} \underset{p}{\longrightarrow} \Omega^{-1/2} M \Omega^{1/2} \equiv Q, \tag{85}$$

where

$$Q = \left[I - \Omega^{-1/2}\nabla H(\theta^*)[\nabla H(\theta^*)'\Omega^{-1}\nabla H(\theta^*)]^{-1}\nabla H(\theta^*)'\Omega^{-1/2}\right], \tag{86}$$

and where

$$M = \left[I - \nabla H(\theta^*)[\nabla H(\theta^*)'\Omega^{-1}\nabla H(\theta^*)]^{-1}\nabla H(\theta^*)'\Omega^{-1}\right], \tag{87}$$

and where $\nabla H(\theta^*) = E\{\partial h(\tilde{y},\tilde{x},\theta^*)/\partial\theta'\}$. It is straightfoward to verify that the matrix Q in equation (86) is symmetric and idempotent. Thus, we have

$$N H_N(\hat{\theta}_N)'[\hat{\Omega}_N^{-1}]^{-1}H_N(\hat{\theta}_N) \underset{d}{\Longrightarrow} [Q\tilde{Z}]'[Q\tilde{Z}] = \tilde{Z}'Q\tilde{Z}, \tag{88}$$

where $\tilde{Z} \sim N(0, I)$. By the probability result in Question 1 of Part II, it follows that $\tilde{Z}'Q\tilde{Z} \sim \chi^2(\mathrm{rank}(Q))$. However we have $\mathrm{rank}(Q) = \mathrm{rank}(M)$, and $\mathrm{rank}(M) \leq J - K$ due to the fact that

$$M\nabla H(\theta^*) = 0, \tag{89}$$

where 0 denotes a $J \times K$ matrix of zeros, as can be verified by multiplying $\nabla H(\theta^*)$ on both sides of equation (87). However since $Q = I - R$ where R is given by

$$R = \Omega^{-1/2}\nabla H(\theta^*)[\nabla H(\theta^*)'\Omega^{-1}\nabla H(\theta^*)]^{-1}\nabla H(\theta^*)'\Omega^{-1/2} \tag{90}$$

and $\text{rank}(R) \leq K$, it follows that $\text{rank}(Q) \geq J - K$. Combining these two inequalities we have $\text{rank}(Q) = J - K$ and we conclude that we have established the result tat

$$N H_N(\hat{\theta}_N)\hat{\Omega}_N^{-1}H_N(\theta_N) \underset{d}{\Longrightarrow} \chi^2(J-K). \tag{91}$$

QUESTION 2 (Consistency of Bayesian posterior) Consider a Bayesian who has observes *IID* data $(X_1, \ldots, X_N)$, where $f(x|\theta)$ is the likelihood for a single observation, and $p(\theta)$ is the prior density over an unknown finite-dimensional parameter $\theta \in R^K$.

A. (10%) Use Bayes Rule to derive a formula for the posterior density of θ given $(X_1, \ldots, X_N)$.

Answer: The posterior is given by

$$f(\theta|X_1, \ldots, X_N) = \frac{\prod_{i=1}^N f(X_i|\theta)p(\theta)}{\int \prod_{i=1}^N f(X_i|\theta)p(\theta)d\theta}. \tag{92}$$

B. (20%) Let $P(\theta \in A|X_1, \ldots, X_N\}$ be the posterior probability θ is in some set $A \subset \Theta$ given the first N observations. Show that this posterior probability satisfies the *Law of iterated expectations:*

$$E\{P(\theta \in A|X_1, \ldots, X_{N+1})|X_1, \ldots, X_N\} = P(\theta \in A|X_1, \ldots, X_N).$$

Answer: The formula for the posterior probability that $\theta \in A$ given $(X_1, \ldots, X_N)$ is just the expectation of the indicator function $I\{\theta \in A\}$ with respect to the posterior density for θ given above. That is,

$$P(\theta \in A|X_1, \ldots, X_N) = \frac{\int I\{\theta \in A\}\prod_{i=1}^N f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^N f(X_i|\theta)p(\theta)d\theta}. \tag{93}$$

Similarly, we have

$$P(\theta \in A|X_1, \ldots, X_N, X_{N+1}) = \frac{\int I\{\theta \in A\}\prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}. \tag{94}$$

Now, to compute the conditional expectation $E\{P(\theta \in A|X_1, \ldots, X_{N+1})|X_1, \ldots, X_N\}$ we note that the appropriate density to use is our posterior belief about X_{N+1} given $(X_1, \ldots, X_N)$. This conditional density can be derived using the posterior for θ

$$\begin{aligned} f(X_{N+1}|X_1, \ldots, X_N) &= \int f(X_{N+1}|\theta)f(\theta|X_1, \ldots, X_N)d\theta \\ &= \frac{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^N f(X_i|\theta)p(\theta)d\theta}. \end{aligned} \tag{95}$$

Thus, $E\{P(\theta \in A|X_1, \ldots, X_{N+1})|X_1, \ldots, X_N\}$ is given by

$$\int_{X_{N+1}} \frac{\int_\theta I\{\theta \in A\}\prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta} f(X_{N+1}|X_1, \ldots, X_N)dX_{N+1}. \tag{96}$$

Using the formula for $f(X_{N+1}|X_1,\dots,X_N)$ given in equation (95) we get

$$\begin{aligned}
&\int_{X_{N+1}} \frac{\int_\theta I\{\theta \in A\} \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta} f(X_{N+1}|X_1,\dots,X_N)dX_{N+1} \\
= &\int_{X_{N+1}} \frac{\int_\theta I\{\theta \in A\} \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta} \frac{\int \prod_{i=1}^{N+1} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta} dX_{N+1} \\
= &\int_{X_{N+1}} \frac{\int_\theta I\{\theta \in A\} f(X_{N+1}|\theta) \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta} dX_{N+1} \\
= &\int_\theta \frac{\int_{X_{N+1}} f(X_{N+1}|\theta)dX_{N+1} I\{\theta \in A\} \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta} \\
= &\int_\theta \frac{I\{\theta \in A\} \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta}{\int \prod_{i=1}^{N} f(X_i|\theta)p(\theta)d\theta} \\
= &P(\theta \in A|X_1,\dots,X_N). \qquad (97)
\end{aligned}$$

C. (20%) A *martingale* is a stochastic process $\{\tilde{Z}_t\}$ that satisfies $E\left\{\tilde{Z}_{t+1}|\mathcal{I}_t\right\} = \tilde{Z}_t$, where $\mathcal{I}_t$ denotes the information set at time t and includes knowledge of all past Z_t's up to time t, $\mathcal{I}_t \supset (\tilde{Z}_1,\dots,\tilde{Z}_t)$. Use the result in part A to show that the process $\{\tilde{Z}_t\}$ where $\tilde{Z}_t = P(\theta \in A|\tilde{X}_1 \dots, X_t)$ is a martingale. (We are interested in martingales because the *Martingale Convergence Theorem* can be used to show that if θ is finite-dimensional, then the posterior distribution converges with probability 1 to a point mass on the true value of θ generating the observations $\{X_i\}$. But you don't have to know anything about this to answer this question.)

The Law of the Iterated Expectations argument above is the proof that the $\{Z_t\}$ process, $Z_t \equiv P(\theta \in A|X_1,\dots,X_t)$, is a martingale. That is, if we let $\mathcal{I}_t = (X_1,\dots,X_t)$, then we have

$$E\{Z_{t+1}|\mathcal{I}_t\} = E\{P(\theta \in A|X_1,\dots,X_{t+1})|X_1,\dots,X_t\}. \qquad (98)$$

The Law of Iterated Expectations result above establishes that

$$E\{P(\theta \in A|X_1,\dots,X_{t+1})|X_1,\dots,X_t\} = P(\theta \in A|X_1,\dots,X_t), \qquad (99)$$

from which we conclude that the posterior probability process is a martingale.

D. (50%) Suppose that if θ is restricted to the K-dimensional simplex, $\theta = (\theta_1,\dots,\theta_K)$ with $\theta_i \in (0,1)$, $i = 1,\dots,K$, $1 = \sum_{i=1}^K \theta_i$, and the distribution of X_i given θ is multinomial with parameter θ, i.e.

$$Pr\{X_i = k\} = \theta_k, \quad k = 1,\dots,K.$$

Suppose the prior distribution over θ, $p(\theta)$ is *Dirichlet* with parameter α:

$$p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_K)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}$$

where both $\theta_i > 0$ and $\alpha_i > 0$, $i = 1,\dots,K$. Compute the posterior distribution and show 1) the posterior is also Dirichlet (i.e. the Dirichlet is a conjugate family), and show directly that as $N \to \infty$ that the posterior distribution converges to a point mass on the true parameter θ generating the data.

Answer: The Dirichlet-Multinomial combination is a *conjugate family of distributions.* That is, if the prior distribution is Dirichlet with prior hyperparameters $(\alpha_1, \ldots, \alpha_K)$ and the data are generated by a multinomial with K mutually exclusive outcomes, then the posterior distribution after observing N *IID* draws from the multinomial is also Dirichlet with parameter $(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$ where

$$n_k = \sum_{i=1}^{N} I\{X_i = k\} \tag{100}$$

By the Law of Large Numbers we have that

$$\frac{n_k}{N} = \frac{1}{N}\sum_{i=1}^{N} I\{X_i = k\} \underset{p}{\longrightarrow} E\{I\{X_i = k\}\} = \theta_k^*. \tag{101}$$

We prove the consistency of the posterior by showing that for any $\theta \neq \theta^*$ we have with probability 1

$$\lim_{N\to\infty} \log\left(\frac{p(\theta^*|X_1, \ldots, X_N)}{p(\theta|X_1, \ldots, X_N)}\right) \longrightarrow \infty. \tag{102}$$

This implies that the limiting posterior puts infinitely more weight on the event that $\theta = \theta^*$ than on any other possible value for θ. Dividing by N and taking limits we have

$$\lim_{N\to\infty} \frac{1}{N} \log\left(\frac{p(\theta^*|X_1, \ldots, X_N)}{p(\theta|X_1, \ldots, X_N)}\right) = \sum_{k=1}^{K} \frac{(\alpha_k + n_k)}{N}\left[\log(\theta_k^*) - \log(\theta_k)\right] \underset{p}{\longrightarrow} \sum_{k=1}^{K} \theta_k^* \left[\log(\theta_k^*) - \log(\theta_k)\right]. \tag{103}$$

However by the Information Inequality we have

$$\sum_{k=1}^{K} \theta_k^* \left[\log(\theta_k^*) - \log(\theta_k)\right] > 0. \tag{104}$$

This result implies that with probability 1

$$\lim_{N\to\infty} \log\left(\frac{p(\theta^*|X_1, \ldots, X_N)}{p(\theta|X_1, \ldots, X_N)}\right) = \lim_{N\to\infty} N\left[\frac{1}{N}\log\left(\frac{p(\theta^*|X_1, \ldots, X_N)}{p(\theta|X_1, \ldots, X_N)}\right)\right] \longrightarrow \infty, \tag{105}$$

since the latter term converges with probability 1 to a positive quantity.

Another way to see the result is to note that if the $K \times 1$ vector $\tilde{\theta}$ has a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_K)$ then

$$E\{\tilde{\theta}_j\} = \frac{\alpha_j}{\sum_{k=1}^{K} \alpha_k}, \tag{106}$$

and

$$\operatorname{var}(\tilde{\theta}_i) = \frac{\alpha_j \left(\sum_{k=1}^{K} \alpha_k - \alpha_j\right)}{\left(\sum_{k=1}^{K} \alpha_k\right)^2 \left(\sum_{k=1}^{K} \alpha_k + 1\right)}. \tag{107}$$

Since the posterior distribution is Dirichlet with parameter $(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$, we can divide the numerator and denominator of the expression for $E\{\theta_j | X_1, \ldots, X_N\}$ by N and use the Law of Large Numbers to show that in the limit with probability 1 we have

$$E\{\tilde{\theta}_j | X_1 \ldots, X_N\} = \frac{\alpha_j + n_j}{\sum_{k=1}^K (\alpha_k + n_k)} = \frac{\alpha_j/N + n_j/N}{\sum_{k=1}^K (\alpha_k/N + n_k/N)} \longrightarrow \frac{\theta_j^*}{\sum_{k=1}^K \theta_k^*} = \theta_k^*. \tag{108}$$

Via a similar sort of calculation, we can show that the conditional variance $\mathrm{var}(\tilde{\theta}_j | X_1, \ldots, X_N\}$ converges to zero since we have

$$\begin{aligned}
\mathrm{var}(\tilde{\theta}_i | X_1, \ldots, X_N) &= \frac{(\alpha_j + n_j)\left(\sum_{k=1}^K (\alpha_k + n_k) - (\alpha_j + n_j)\right)}{\left(\sum_{k=1}^K (\alpha_k + n_k)\right)^2 \left(\sum_{k=1}^K (\alpha_k + n_k) + 1\right)} \\
&= \frac{(\alpha_j/N + n_j/N)\left(\sum_{k=1}^K (\alpha_k/N + n_k/N) - (\alpha_j/N + n_j/N)\right)}{N^2 \left(\sum_{k=1}^K (\alpha_k/N + n_k/N)\right)^2 \left(\sum_{k=1}^K (\alpha_k/N + n_k/N) + 1\right)}
\end{aligned} \tag{109}$$

and the numerator of the latter expression converges with probability 1 to $\theta_j^*(1 - \theta_j^*)$ but the denominator converges to $+\infty$ with probability 1.

QUESTION 3 Consider the *random utility model:*

$$\tilde{u}_d = v_d + \tilde{\epsilon}_d, \quad d = 1, \ldots, D \tag{110}$$

where $\tilde{u}_d$ is a decision-maker's payoff or utility for selecting alternative d from a set containing D possible alternatives (we assume that the individual only chooses one item). The term v_d is known as the deterministic or *strict utility* from alternative d and the error term $\tilde{\epsilon}_d$ is the random component of utility. In empirical applications v_d is often specified as

$$v_d = X_d \beta \tag{111}$$

where X_d is a vector of observed covariates and β is a vector of coefficients determining the agent's utility to be estimated. The interpretation is that X_d represents a vector of characteristics of the decision-maker and alternative d that are observable by the econometrician and ϵ_d represents characteristics of the agent and alternative d that affect the utility of choosing alternative d which are unobserved by the econometrician. Define the agent's *decision rule* $\delta(\epsilon_1, \ldots, \epsilon_D)$ by:

$$\delta(\epsilon) = argmax_{d=1,\ldots,D} \left[v_d + \tilde{\epsilon}_d\right] \tag{112}$$

i.e. $\delta(\epsilon)$ is the optimal choice for an agent whose unobserved utility components are $\epsilon = (\epsilon_1, \ldots, \epsilon_D)$. Then the agent's *choice probability* $P\{d|X\}$ is given by:

$$P\{d|X\} = \int I\{d = \delta(\epsilon)\} f(\epsilon|X) d\epsilon \tag{113}$$

where $X = (X_1, \ldots, X_D)$ is the vector of observed characteristics of the agent and the D alternatives and $f(\epsilon|X)$ is the conditional density function of the random components of utility given the values of observed components X, and $I\{\delta(\epsilon) = d\}$ is the *indicator function* given by $I\{\delta(\epsilon) = d\} = 1$ if $\delta(\epsilon) = d$ and 0 otherwise. Note that the integral above is actually a multivariate integral over the D components of $\epsilon = (\epsilon_1, \ldots, \epsilon_D)$, and simply represents the probability that the values of the vector of unobserved utilities ϵ lead the agent to choose alternative d.

Definition: The *Social Surplus Function* $U(v_1,\ldots,v_D,X)$ is given by:

$$U(v_1,\ldots,v_D,X) = E\left\{\max_{d=1,\ldots,D}[v_d+\tilde{\epsilon}_d]\Big|X\right\} = \int_{\epsilon_1}\cdots\int_{\epsilon_D}\max_{d=1,\ldots,D}[v_d+\epsilon_d]f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D \tag{114}$$

The Social Surplus function is the expected maximized utility of the agent.[1]

A. (50%) Prove the *Williams-Daly-Zachary Theorem:*

$$\frac{\partial U}{\partial v_d}(v_1,\ldots,v_D,X) = P\{d|X\} \tag{115}$$

and discuss its relationship to *Roy's Identity.*

Hint: Interchange the differentiation and expectation operations when computing $\partial U/\partial v_d$:

$$\begin{aligned}\frac{\partial U}{\partial v_d}(v_1,\ldots,v_D,X) &= \partial/\partial v_d\int_{\epsilon_1}\cdots\int_{\epsilon_D}\max_{d=1,\ldots,D}[v_d+\epsilon_d]f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D\\ &= \int_{\epsilon_1}\cdots\int_{\epsilon_D}\partial/\partial v_d\max_{d=1,\ldots,D}[v_d+\epsilon_d]f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D\end{aligned}$$

and show that

$$\frac{\partial}{\partial v_d}\max_{d=1,\ldots,D}[v_d+\epsilon_d] = I\{d=\delta(\epsilon)\}.$$

Answer: The hint gives away most of the answer. We simply appeal to the Lebesgue Dominated Convergence Theorem to justify the interchange of integration and differentiation operators. As long as the distribution of the $\{\epsilon_d\}$'s has a density, the derivative

$$\partial/\partial v_d\max_{d=1,\ldots,D}[v_d+\epsilon_d] = I\{d=\delta(\epsilon)\}. \tag{116}$$

exists almost everywhere with respect to this density and is bounded by 1, so that the Lebesgue Dominated Convergence Theorem applies. It is easy to see why the partial derivative of $\max_{d=1,\ldots,D}[v_d+\epsilon_d]$ equals the indicator function $I\{d=\delta(\epsilon)\}$: if this function equals 1 then alternative d yields the highest utility and we have

$$v_d+\epsilon_d > v_{d'}+\epsilon_{d'} \quad \forall d'\neq d$$

Thus, $v_d+\epsilon_d = \max_{d'=1,\ldots,D}[v_{d'}+\epsilon_{d'}]$ and we have $\partial/\partial v_d\max_{d=1,\ldots,D}[v_d+\epsilon_d]=1$ when $I\{d=\delta(\epsilon)\}=1$. However when $I\{d=\delta(\epsilon)\}=0$, then alternative d is not the utility maximizing choice, so that $\max_{d'=1,\ldots,D}[v_{d'}+\epsilon_{d'}] > v_d+\epsilon_d$. It follows that we have $\partial/\partial v_d\max_{d=1,\ldots,D}[v_d+\epsilon_d]=0$ when $I\{d=\delta(\epsilon)\}=0$ so that the identity claimed in (116)

[1] If we think of an economy consisting of a population of agents each with their own observed vector of utilities ϵ and $f(\epsilon|X)$ is the density function representing the distribution of these "types" in the population, then $U(v_1,\ldots,v_D,X)$ represents the indirect or maximized utility of a typical person in the population. This is the reason U is referred to as a Social Surplus Function.

holds with probability 1, and so via the Lebesgue Dominated Convergence Theorem we have

$$\begin{aligned}
\frac{\partial U}{\partial v_d}(v_1,\ldots,v_D,X) &= \frac{\partial}{\partial v_d}\int_{\epsilon_1}\cdots\int_{\epsilon_D}\max_{d=1,\ldots,D}[v_d+\epsilon_d]f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D \\
&= \int_{\epsilon_1}\cdots\int_{\epsilon_D}\frac{\partial}{\partial v_d}\max_{d=1,\ldots,D}[v_d+\epsilon_d]f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D \\
&= \int_{\epsilon_1}\cdots\int_{\epsilon_D} I\{d=\delta(\epsilon)\}f(\epsilon_1,\ldots,\epsilon_D|X)d\epsilon_1\cdots d\epsilon_D \\
&= P\{d|X\}. \qquad (117)
\end{aligned}$$

B. (50%) Consider the special case of the random utility model when $\epsilon=(\epsilon_1,\ldots,\epsilon_D)$ has a multivariate (Type I) *extreme value distribution:*

$$f(\epsilon|X)=\prod_{d=1}^{D}\exp\{-\epsilon_d\}\exp\{-\exp\{-\epsilon_d\}\}. \qquad (118)$$

Show that the conditional choice probability $P\{d|X\}$ is given by the *multinomial logit formula:*

$$P\{d|X\}=\frac{\exp\{v_d/\sigma\}}{\sum_{d'=1}^{D}\exp\{v_{d'}/\sigma\}}. \qquad (119)$$

Hint 1: Use the Williams-Daly-Zachary Theorem, showing that in the case of the extreme value distribution (118) the Social Surplus function is given by

$$U(v_1,\ldots,v_D,X)=\sigma\gamma+\sigma\log\left[\sum_{d=1}^{D}\exp\{v_d/\sigma\}\right]. \qquad (120)$$

where $\gamma=.577216\ldots$ is Euler's constant.

Hint 2: To derive equation (120) show that the extreme value family is *max-stable:* i.e. if $(\epsilon_1,\ldots,\epsilon_D)$ are *IID* extreme value random variables, then $\max_d\{\epsilon_d\}$ also has an extreme value distribution. Also use the fact that the expectation of a single extreme value random variable with location parameter α and scale parameter σ is given by:

$$E\{\tilde\epsilon\}=\int_{-\infty}^{+\infty}\epsilon\exp\{-\epsilon\}\exp\{-\exp\{-\epsilon\}\}d\epsilon=\alpha+\sigma\gamma, \qquad (121)$$

and the CDF is given by

$$F(x|\alpha,\sigma)=P\{\tilde\epsilon\le x|\alpha,\sigma\}=\exp\left\{-\exp\left\{\frac{-(x-\alpha)}{\sigma}\right\}\right\}. \qquad (122)$$

Hint 3: Let $(\epsilon_1,\ldots,\epsilon_D)$ be *INID* (independent, non-identically distributed) extreme value random variables with location parameters $(\alpha_1,\ldots,\alpha_D)$ and common scale parameter σ. Show that this family is max-stable by proving that $\max(\epsilon_1,\ldots,\epsilon_D)$ is an extreme value random variable with scale parameter σ and location parameter

$$\alpha=\sigma\log\left[\sum_{d=1}^{D}\exp\{\alpha_d/\sigma\}\right] \qquad (123)$$

Answer: Once again, the hints are virtually the entire answer to the problem. By hint 1, if the Social Surplus function is given by equation (120) then by the Williams-Daly-Zachary Theorem we have

$$P\{d|X\} = \frac{\partial}{\partial v_d}\left[\sigma\gamma + \sigma \log\left[\sum_{d=1}^{D} \exp\{v_d/\sigma\}\right]\right] = \frac{\exp\{v_d/\sigma\}}{\sum_{d'=1}^{D}\exp\{v_{d'}/\sigma\}}. \tag{124}$$

Now to show that the Social Surplus function has the form given in equation (120), we use the fact that if $\{\epsilon_d\}$ are independent random variables, we have following formula for the probability distribution of the random variable $\max_{d=1,\ldots,D}[v_d+\epsilon_d]$:

$$\Pr\left\{\max_{d=1,\ldots,D}[v_d+\epsilon_d] \le x\right\} = \prod_{d=1}^{D} \Pr\left\{v_d+\epsilon_d \le x\right\}. \tag{125}$$

Now, let ϵ_d have a Type III extreme value value distribution with location parameter $\alpha_d = 0$ and scale parameter $\sigma > 0$. Then it is easy to see that $v_d + \epsilon_d$ is also a Type III extreme value random variate with location parameter v_d and scale parameter σ. That is, the family of independent Type III extreme distributions is max-stable. Plugging in the formula for the Type III extreme value distribution from equation (122) into the formula for the CDF of $\max_{d=1,\ldots,D}[v_d+\epsilon_d]$ given above, we find that

$$\Pr\left\{\max_{d=1,\ldots,D}[v_d+\epsilon_d] \le x\right\} = \exp\left\{-\exp\left\{\frac{-(x-\alpha)}{\sigma}\right\}\right\}, \tag{126}$$

where the location parameter is given by the log-sum formula in equation (123). The form of the Social Surplus Function in equation (120) then follows from the formula for the expectation of an extreme value random variate in equation (121), and formula (123) for the location parameter of the maximum of a collection of independent Type III extreme random variables, i.e.

$$U(v_1,\ldots,v_D,X) \equiv E\left\{\max_{d=1,\ldots,D}[v_d+\epsilon_d]\right\} = \sigma\gamma + \alpha = \sigma\gamma + \sigma\log\left[\sum_{d=1}^{D}\exp\{v_d/\sigma\}\right]. \tag{127}$$

QUESTION 4 (Latent Variable Models) The *Binary Probit Model* can be viewed as a simple type of latent variable model. There is an underlying linear regression model

$$\tilde{z} = X\beta^* + \epsilon \tag{128}$$

but where the dependent variable $\tilde{z}$ is *latent,* i.e. it is not observed by the econometrician. Instead we observe the dependent variable y given by

$$y = \begin{cases} 1 & \text{if} \quad \tilde{z} > 0 \\ 0 & \text{if} \quad \tilde{z} \le 0 \end{cases} \tag{129}$$

1. (5%) Assume that the error term $\epsilon \sim N(0,\sigma^2)$. Show that the scale of β^* and the parameter σ^2 is not simultaneously identified and therefore without loss of generality we can normalize $\sigma^2 = 1$ and interpret the estimated β coefficients as being the true coefficients β^* divided by σ:

$$\beta = \frac{\beta^*}{\sigma}. \tag{130}$$

Answer: Notice that if $\lambda > 0$ is an arbitrary positive constant, then if we divide both sides of equation (128) by λ, the probability distribution for the observed dependent variable has not changed since we have

$$\tilde{z} > 0 \Longleftrightarrow \frac{\tilde{z}}{\lambda} > 0. \tag{131}$$

Thus the model with latent variable $\tilde{z}/\lambda$ is *observationally equivalent* to the model with the latent variable $\tilde{z}$. If we normalize the variance of ϵ to 1, this is equivalent to dividing $\tilde{z}$ by the standard deviation σ of the underlying "true" ϵ variable, so that our estimates of β should be interpreted as being estimates of β/σ.

2. (10%) Derive the conditional probability $\Pr\{y = 1|X\}$ in terms of X, β and the standard normal CDF, Φ and use this probability to write down the likelihood function for N *IID* observations of pairs $\{(y_i, X_i)\}, i = 1, \ldots, N$.

   Answer: We have

$$\Pr\{y = 1|X, \beta^*\} = \Pr\{\tilde{z} > 0\} = \Pr\{X\beta^* + \epsilon > 0\} = \Pr\{-\epsilon < X\beta^*\} = \Phi(X\beta^*), \tag{132}$$

   where Φ is the CDF of a $N(0,1)$ random variable, and we used the fact that if $\epsilon \sim N(0,1)$ then $-\epsilon \sim N(0,1)$. Using this formula, the likelihood for N observations $\{y_i, X_i\}$ is given by

$$L(\beta) = \prod_{i=1}^{N} [\Phi(X_i\beta)]^{y_i} [1 - \Phi(X_i\beta)]^{(1-y_i)}. \tag{133}$$

3. (20%) Show that β can be consistently estimated by nonlinear least squares by writing down the least squares problem and sketching a proof for its consistency.

   We observe that y satisfies the following nonlinear regression equation:

$$y = \Phi(X\beta^*) + \xi, \tag{134}$$

   where $E\{\xi|X\} = 0$. To see this, note that conditional on X the residual ξ takes on two possible values. If $y = 1$, which occurs with probability $\Phi(X\beta^*)$, then $\xi = 1 - \Phi(X\beta^*)$. If $y = 0$, which occurs with probability $1 - \Phi(X\beta^*)$, then $\xi = -\Phi(X\beta^*)$. Thus we have the conditional expectation is given by

$$E\{\xi|X\} = [1 - \Phi(X\beta^*)]\Phi(X\beta^*) - \Phi(X\beta^*)[1 - \Phi(X\beta^*)] = 0. \tag{135}$$

   Thus, since the conditional expectation of y is given by the parametric function $\Phi(X\beta^*)$ it follows from the general results on the consistency of nonlinear least squares that the nonlinear least squares estimator

$$\hat{\beta}_N^n = \underset{\beta \in R^k}{argmin} \sum_{i=1}^{N} [y_i - \Phi(X_i\beta)]^2 \tag{136}$$

   will be a consistent estimator of β^*.

4. (20%) Derive the asymptotic distribution of the maximum likelihood estimator by providing an analytical formula for the asymptotic covariance matrix of the MLE estimator $\hat{\beta}_N$

Hint: This is the inverse of the information matrix $\mathcal{I}$. Derive a formula for $\mathcal{I}$ in terms of Φ, X and β and possibly other terms.

Answer: We know that if the model is correctly specified and basic regularity conditions hold, that the maximum likelihood estimator, β_N^M, is consistent and asymptotically normally distributed with

$$\sqrt{N}[\hat{\beta}_N^m - \beta^*] \underset{d}{\Longrightarrow} N(0, \mathcal{I}^{-1}), \tag{137}$$

where $\mathcal{I}$ is the *Information Matrix* given by

$$\mathcal{I} = E\{\frac{\partial}{\partial\beta}\log f(y|X,\beta^*)\frac{\partial}{\partial\beta'}\log f(y|X,\beta^*)\}. \tag{138}$$

In the case of the probit model we have

$$\log f(y|X,\beta^*) = y\log(\Phi(X\beta)) + (1-y)\log(1-\Phi(X\beta)), \tag{139}$$

and so we have

$$\frac{\partial}{\partial\beta}\log f(y|X,\beta^*) = \frac{y\phi(X\beta)X}{\Phi(X\beta)} - \frac{(1-y)\phi(X\beta)X}{(1-\Phi(X\beta))} \tag{140}$$

where

$$\phi(X\beta) = \Phi'(X\beta) = \frac{1}{\sqrt{2\pi}}\exp\{-(X\beta)^2/2\}. \tag{141}$$

Using this formula it is not hard to see that

$$\begin{aligned}\mathcal{I} &= E\left\{\left[\frac{1}{\Phi(X\beta^*)} + \frac{1}{[1-\Phi(X\beta^*)]}\right]\phi^2(X\beta^*)XX'\right\} \\ &= E\left\{\left[\frac{\phi^2(X\beta^*)XX'}{\Phi(X\beta^*)[1-\Phi(X\beta^*)]}\right]\right\}.\end{aligned} \tag{142}$$

5. (20%) Derive the asymptotic distribution of the nonlinear least squares estimator and compare it to the maximum likelihood estimator. Is the nonlinear least squares estimator asymptotically inefficient?

   Answer: The first order condition for the nonlinear least squares estimator $\hat{\beta}_N$ is given by:

$$0 = \frac{1}{N}\sum_{i=1}^{N}[y_i - \Phi(X_i\hat{\beta}_N)]\phi(X_i\hat{\beta}_N)X_i. \tag{143}$$

   Expanding this first order condition in a Taylor series about β^* we obtain

$$\begin{aligned}0 &= \frac{1}{N}\sum_{i=1}^{N}[y_i - \Phi(X_i\beta^*)]\phi(X_i\beta^*)X_i \\ &- \left[\frac{1}{N}\sum_{i=1}^{N}\phi^2(X_i\tilde{\beta}_N)X_iX_i' - [y_i - \Phi(X_i\tilde{\beta}_N)]\phi'(X_i\tilde{\beta}_N)X_iX_i'\right](\hat{\beta}_N - \beta^*).\end{aligned} \tag{144}$$

where $\tilde{\beta}_N$ is a vector each of who coordinates are on the line segment joining the corresponding components of $\hat{\beta}_N$ and β^*. Solving the above equation for $\sqrt{N}(\hat{\beta} - \beta^*)$ we obtain

$$\begin{aligned}\sqrt{N}(\hat{\beta}_N - \beta^*) &= \left[\frac{1}{N}\sum_{i=1}^{N}\phi^2(X_i\tilde{\beta}_N)X_iX_i' - \frac{1}{N}\sum_{i=1}^{N}[y_i - \Phi(X_i\tilde{\beta}_N)]\phi'(X_i\tilde{\beta}_N)X_iX_i'\right]^{-1} \\ &\times \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[y_i - \Phi(X_i\beta^*)]\phi(X_i\beta^*)X_i\right]. \end{aligned} \tag{145}$$

Applying the Central Limit Theorem to the second term in brackets in the above equation we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[y_i - \Phi(X_i\beta^*)]\phi(X_i\beta^*)X_i] \underset{d}{\Longrightarrow} N(0,\Omega), \tag{146}$$

where Ω is given by

$$\begin{aligned}\Omega &= E\left\{\left[[1-\Phi(X\beta^*)]^2\Phi(X\beta^*) + [\Phi(X\beta^*)]^2[1-\Phi(X\beta^*)]\right]\phi^2(X\beta^*)XX'\right\} \\ &= E\left\{[1-\Phi^2(X\beta^*)]\phi^2(X\beta^*)XX'\right\}. \end{aligned} \tag{147}$$

Appealing to the uniform strong law of large numbers, we can show that the other term in equation (145) converges to the following limiting value with probability 1:

$$\left[\frac{1}{N}\sum_{i=1}^{N}\phi^2(X_i\tilde{\beta}_N)X_iX_i' - [y_i - \Phi(X_i\tilde{\beta}_N)]\phi'(X_i\tilde{\beta}_N)X_iX_i'\right] \longrightarrow \Sigma \tag{148}$$

where

$$\Sigma = E\left\{\phi^2(X\beta^*)XX'\right\}. \tag{149}$$

It follows that the asymptotic distribution of the nonlinear least squares estimator is given by

$$\sqrt{N}[\hat{\beta}_N - \beta^*] \underset{d}{\Longrightarrow} N(0, \Sigma^{-1}\Omega\Sigma^{-1}). \tag{150}$$

Since the maximum likelihood estimator is an asymptotically efficient estimator and the nonlinear least squares estimator is a potentially inefficient estimator, we have

$$\mathcal{I}^{-1} \leq \Sigma^{-1}\Omega\Sigma^{-1}. \tag{151}$$

To see that the inequality is strict in general, consider the special case where there is a degenerate distribution with only one possible X vector. Then turning the above inequality around we want to show that

$$\mathcal{I} > \Sigma\Omega^{-1}\Sigma. \tag{152}$$

However when the distribution of X is degenerate we have

$$\mathcal{I} = \frac{\phi^2(X\beta^*)XX'}{\Phi(X\beta^*)[1-\Phi(X\beta^*)]}. \tag{153}$$

Similarly we have

$$\Sigma\Omega^{-1}\Sigma = \frac{\phi^2(X\beta^*)XX'}{[1-\Phi^2(X\beta^*)]}. \tag{154}$$

However since

$$\frac{1}{\Phi(X\beta^*)[1-\Phi(X\beta^*)]} > \frac{1}{[1-\Phi^2(X\beta^*)]}, \tag{155}$$

it follows that $\mathcal{I} > \Sigma\Omega^{-1}\Sigma$, so that the nonlinear least squares estimator will generally be strictly asymptotically inefficient in comparison to the maximum likelihood estimator.

6. (25%) Show that the nonlinear least squares estimator of β is subject to heteroscedasticity by deriving an explicit formula for the conditional variance of the error term in the nonlinear regression formulation of the estimation problem. Can you form a more efficient estimator by correcting for this heteroscedasticity in a two stage feasible GLS procedure (i.e. in stage 1 computing an initial consistent, but inefficient estimator of β by ordinary nonlinear least squares and in stage two using this initial consistent estimator to correct for the heteroscedasticity and using the stage two estimator of β as the feasible GLS estimator)? If so, is this feasible GLS procedure asymptotically efficient? If you believe so, provide a sketch of the derviation of the asymptotic distribution of the feasible GLS estimator. Otherwise provide a counterexample or a sketch of an argument why you believe the feasible GLS procedure is asymptotically inefficient relative to the maximum likelihood estimator.

   Answer: There is heteroscedasticity in the nonlinear regression formulation of the probit estimation problem in (134) since we have

$$\mathrm{var}(\xi|X) = E\{\xi^2|X\} = [1-\Phi(X\beta^*)]^2\Phi(X\beta^*) + [\Phi(X\beta^*)]^2[1-\Phi(X\beta^*)]. \tag{156}$$

   Now suppose we do an initial first step nonlinear least squares estimation to obtain an initial $\sqrt{N}$-consistent estimator $\hat{\beta}_N$ and then use this to construct a second stage weighted nonlinear least squares problem as follows:

$$\hat{\beta}_N^g = \underset{\beta\in R^k}{argmin}\ \frac{1}{N}\sum_{i=1}^N \frac{[y_i-\Phi(X_i\beta)]^2}{[1-\Phi(X\hat{\beta}_N)]^2\Phi(X\hat{\beta}_N)+[\Phi(X\hat{\beta}_N)]^2[1-\Phi(X\hat{\beta}_N)]}. \tag{157}$$

   It turns out that this two stage, feasible GLS estimator has the same asymptotic distribution as maximum likelihood, i.e. it is an asymptotically efficient estimator. It is easiest to see this result by assuming first that we know the exact form of the heteroscedasticity, i.e. in the denominator of the second stage we weight the observations by the inverse of the exact conditional heteroscedasticity givein in equation (156). Then repeating the Taylor series expansion argument that we used to derive the asymptotic distribution of the unweighted nonlinear least squares estimator, it is not difficult to show that

$$\begin{aligned}\sqrt{N}(\hat{\beta}_N-\beta^*) &= \left[\frac{1}{N}\sum_{i=1}^N\frac{\phi^2(X_i\tilde{\beta}_N)X_iX_i'}{E\{\xi_i^2|X_i\}} - \frac{1}{N}\sum_{i=1}^N\frac{[y_i-\Phi(X_i\tilde{\beta}_N)]\phi'(X_i\tilde{\beta}_N)X_iX_i'}{E\{\xi_i^2|X_i\}}\right]^{-1} \\ &\times \left[\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{[y_i-\Phi(X_i\beta^*)]\phi(X_i\beta^*)X_i}{E\{u_i^2|X_i\}}\right].\end{aligned} \tag{158}$$

   Once again, appealing to the Central Limit Theorem, we can show that the second term in equation (158) converges in distribution to

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N\frac{[y_i-\Phi(X_i\beta^*)]\phi(X_i\beta^*)X_i}{E\{u_i^2|X_i\}} \underset{d}{\Longrightarrow} N(0,\Omega), \tag{159}$$

where in the GLS case Ω is given by

$$\begin{aligned}\Omega &= E\left\{\frac{\phi^2(X\beta^*)XX'}{[[1-\Phi(X\beta^*)]^2\Phi(X\beta^*)+[\Phi(X\beta^*)]^2[1-\Phi(X\beta^*)]]}\right\}\\ &= E\left\{\frac{\phi^2(X\beta^*)XX'}{\Phi(X\beta^*)[1-\Phi(X\beta^*)]}\right\}\\ &= \mathcal{I}.\end{aligned} \tag{160}$$

Similarly, we can show that the other term in equation (158) converges with probability 1 to the matrix Σ,

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi^2(X_i\tilde{\beta}_N)X_iX_i' - [y_i-\Phi(X_i\tilde{\beta}_N)]\phi'(X_i\tilde{\beta}_N)X_iX_i'}{E\{\xi_i^2|X_i\}}\right] \longrightarrow \Sigma \tag{161}$$

where we also have $\Sigma = \mathcal{I}$. Thus, the GLS estimator converges in distribution to

$$\sqrt{N}[\hat{\beta}_N^f - \beta^*] \underset{d}{\Longrightarrow} N(0,\Sigma^{-1}\Omega\Sigma^{-1}) = N(0,\mathcal{I}^{-1}), \tag{162}$$

so the GLS estimator is asymptotically efficient. To show that the feasible GLS estimator (i.e. the one using the estimated conditional variance as weights instead of weighting by the true conditional variance) has this same distribution is a rather tedious exercise in the properties of uniform convergence and will be omitted.

Final Comment: I note that the GMM efficiency bound for the conditional moment restriction

$$H(\beta^*|X) = E\{h(\tilde{y},\tilde{X},\beta^*)|\tilde{X}=X)\} \tag{163}$$

coincides with $\mathcal{I}^{-1}$ when $h(\tilde{y},\tilde{X},\beta) = \tilde{y}-\Phi(\tilde{X}\beta)$. To see this, recall that the GMM bound for conditional moment restrictions is given by

$$\left[E\left\{\nabla H(\beta^*|X)\Omega^{-1}(X)\nabla H(\beta^*|X)'\right\}\right]^{-1}, \tag{164}$$

where

$$\Omega(X) = E\{h(\tilde{y},\tilde{X},\beta^*)h(\tilde{y},\tilde{X},\beta^*)'|\tilde{X}=X\} \tag{165}$$

and

$$\nabla H(\beta^*|X) = E\left\{\frac{\partial}{\partial\beta}h(\tilde{y},\tilde{X},\beta^*)\beta'|\tilde{X}=X\right\}. \tag{166}$$

In the case where $h(\tilde{y},\tilde{X},\beta) = \tilde{y}-\Phi(\tilde{X}\beta)$ we have

$$\Omega(X) = E\{\xi^2|X\} = [1-\Phi(X\beta^*)]^2\Phi(X\beta^*)+[\Phi(X\beta^*)]^2[1-\Phi(X\beta^*)], \tag{167}$$

that is, $\Omega(X)$ is just the conditional heteroscedasticity of the residuals in the nonlinear regression formulation of the probit problem. Also, we have

$$\nabla H(\beta^*|X) = -\phi(X\beta^*)X. \tag{168}$$

Plugging these into the matrix in the inside of the expectation of the GMM bound we have

$$\nabla H(\beta^*|X)\Omega^{-1}(X)\nabla H(\beta^*|X)' = \frac{\phi^2(X\beta^*)XX'}{\Phi(X\beta^*)[1-\Phi(X\beta^*)]}. \tag{169}$$

Taking expectations with respect to X and comparing to the formula for the information matrix in equation (138) we see that

$$E\left\{\nabla H(\beta^*|X)\Omega^{-1}(X)\nabla H(\beta^*|X)'\right\} = \mathcal{I}. \tag{170}$$

Since the GMM bound is the inverse of this matrix, it equals the inverse of the information matrix, $\mathcal{I}^{-1}$, and hence is the same as the (asymptotic) Cramér-Rao lower bound.