

# ECON 623

## Problem Set 4

Due in class on Tuesday November 29, 2011

John Rust

I. Consider a *deterministic* (i.e. non-random) sequence of functions  $\{f_n(\theta)\}$  that are continuous functions of  $\theta$  for  $\theta \in \Theta \subset R^k$ . Suppose  $\{f_n(\theta)\}$  converge *pointwise* to a function  $f(\theta)$  for each  $\theta \in \Theta$ . Suppose  $\theta^*$  is the unique maximizer of  $f(\theta)$  for  $\theta \in \Theta$  and that  $\theta_n^*$  is a maximizer of  $f_n(\theta)$  for  $\theta \in \Theta$ . Will  $\{\theta_n^*\}$  converge to  $\theta^*$ ? If not, provide a counterexample, if so, provide a general proof.

- A. Suppose we strengthen the concept of convergence to *uniform convergence* and assume that  $\{f_n\}$  converges uniformly to  $f$ . Can you prove whether or not  $\{\theta_n^*\}$  converges to  $\theta^*$  in this case?
- B. The *Arzela-Ascoli Theorem* states that  $\{f_n\}$  converges uniformly to  $f$  if and only if the sequence  $\{f_n\}$  is *uniformly equicontinuous*. Define what this means and give an example of a sequence of functions that is uniformly equicontinuous and one that is not equicontinuous.
- C. Suppose we know that  $\{f_n\}$  is *uniformly Lipschitz continuous*, i.e.  $\exists L < \infty$  such that  $\forall n$  we have

$$|f_n(\theta) - f_n(\theta')| \leq L\|\theta - \theta'\| \quad (1)$$

then show that the sequence  $\{f_n\}$  is uniformly equicontinuous.

- D. Use result C. and Taylor series to show that if each  $f_n$  in the sequence  $\{f_n\}$  is continuously differentiable in  $\theta$  and we have

$$\sup_n \left\| \frac{\partial}{\partial \theta} f_n \right\| \leq K < \infty \quad (2)$$

then the sequence  $\{f_n\}$  is uniformly equicontinuous.

- E. Prove that if  $f(\theta)$  is uniquely maximized at  $\theta^*$  in the compact domain  $\Theta \subset R^k$  and if  $\{f_n\}$  converges uniformly to  $f$  and if  $\hat{\theta}_n$  maximizes  $f_n$  over  $\theta \in \Theta$  then  $\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta^*$ .
- F. Go back through parts A to F and discuss how your answers and analysis changes if  $\{f_n\}$  is a *random* sequence of functions (i.e. each  $f_n$  may also depend on one or more random variables/vectors that make it a random function of  $\theta$ ) that converges *uniformly in  $\theta$  to a non-random function  $f(\theta)$  with probability 1*. Do the results listed above continue to hold? If the addition of randomness changes any of the key results, discuss why, if not, discuss why it is that randomness of the  $\{f_n\}$  sequence does not affect our conclusions provided we qualify all conclusions with the words, “with probability 1” (or “almost surely”). In particular, write a definition of the generalization of the concept of uniform equicontinuity of the functions  $\{f_n\}$  in the stochastic case. This generalized concept of uniform equicontinuity of sequences of random functions is known in economics and statistics as *stochastic equicontinuity*.

II. Consider the linear regression model

$$\tilde{y} = \tilde{X}\beta^* + \tilde{\varepsilon} \quad (3)$$

where  $\tilde{\varepsilon} \sim N(0, \sigma^2)$  and  $\tilde{X}$  is a random  $K \times 1$  vector with density  $f(X, \gamma^*)$ , and  $\gamma$  is a  $J \times 1$  vector of unknown parameters for the density of  $\tilde{X}$ , and we assume that  $\tilde{\varepsilon}$  and  $\tilde{X}$  are independently distributed. Assume we have access to a sample  $(y_1, X_1, \dots, y_N, X_N)$  of  $N$  IID draws from this “data generating process” and we are interested in estimating the unknown parameter vector  $\theta^* = (\sigma^2, \beta^*, \gamma^*)$ .

- A. Write down the natural logarithm of the *full likelihood function* for the observed data  $(y_1, X_1, \dots, y_N, X_N)$  divided by  $N$  and call this  $L_N(\theta)$ .
- B. In many studies, the analyst is not so interested in the parameters  $\gamma^*$  of the distribution of covariates  $f(X, \gamma^*)$  and is really only interested in the parameters  $(\sigma^2, \beta)$ . The (average of the log) of the *partial likelihood function* is given by

$$L_N(\sigma^2, \beta) = \frac{1}{N} \sum_{i=1}^N \log(f(y_i | X_i, \sigma^2, \beta)) \quad (4)$$

where  $f(y|X, \sigma^2, \beta)$  is the conditional density of  $\tilde{y}$  given  $\tilde{X} = X$  which is easy to show is a normal distribution with mean  $X\beta$  and variance  $\sigma^2$ . Show that even though we estimate  $(\sigma^2, \beta)$  using this partial likelihood function (and thus ignore the parameter  $\gamma$  and information contained in the  $X_i$  observations that is reflected in the full likelihood function above), that computing the maximum likelihood estimates for  $(\sigma^2, \beta)$  by maximizing the partial likelihood function above results in the *same parameter estimates* as would be obtained if we maximized the full likelihood function in part A above.

- C. Outline the argument for why the *partial maximum likelihood estimates*  $(\hat{\sigma}^2, \hat{\beta})$  from part B will be consistent and asymptotically normally distributed even though we are ignoring the parameter  $\gamma$  and the information contained in the density  $f(X, \gamma)$ . Do you think that ignoring this extra information will come at a cost in terms of *reduced asymptotic efficiency* of the partial maximum likelihood estimates of  $(\sigma^2, \beta^*)$ ?
- D. Compute the *Information Matrix*  $I$  for the partial likelihood function and show that it is *block diagonal* between the parameters  $\sigma^2$  and  $\beta^*$ . That is, show that if  $I_{\theta^2, \beta^* j}$  is the element of the Information matrix corresponding to the pair of parameters  $(\sigma^2, \beta_j^*)$  that we have  $I_{\theta^2, \beta^* j} = 0, j = 1, \dots, K$ . Use this result to show that this implies that the asymptotic covariance matrix implies zero covariance between  $\hat{\sigma}^2$  and  $\hat{\beta}$  asymptotically, i.e.  $\text{cov}(\hat{\sigma}^2, \hat{\beta}_j) \rightarrow 0, j = 1, \dots, K$ , where  $\text{cov}(\hat{\sigma}^2, \hat{\beta}_j)$  is the asymptotic covariance between  $\hat{\sigma}^2$  and  $\hat{\beta}_j$ .
- E. For the full likelihood function, show that there is also block diagonality for the parameters  $\sigma^2, \beta$  and  $\gamma$ , i.e. we have for the following elements of the full information matrix  $I$

$$\begin{aligned} I_{\sigma^2, \beta_k^*} &= 0, & k &= 1, \dots, K \\ I_{\sigma^2, \gamma_j^*} &= 0, & j &= 1, \dots, J \\ I_{\beta_k^*, \gamma_j^*} &= 0, & j &= 1, \dots, J, k = 1, \dots, K \end{aligned}$$

Use this to show that there is no loss in asymptotic efficiency in estimating  $(\sigma^2, \beta^*)$  using the partial likelihood function instead of the full likelihood function, i.e. both of these maximum likelihood estimators achieve the Cramer-Rao lower bound for  $(\sigma^2, \beta^*)$ .

- F. Now suppose we believe there is potential *heteroscedasticity* in the regression, so that instead of a constant variance  $\sigma^2$ , we now believe that the  $\tilde{\epsilon}_i$  random variables have a *conditional variance* of the form

$$\text{var}\{\tilde{\epsilon}|X\} = \sigma^2(X, \alpha^*), \quad (5)$$

where  $\alpha^*$  is a  $H \times 1$  vector of unknown parameters to be estimated along with  $\beta^*$  (assume given the result in part E above, we are not interested in estimating  $\gamma^*$  and since there is no penalty for ignoring it and using the partial likelihood, you are justified to continue to do this also in this part too). Thus, we are now estimated in estimating  $\theta^* = (\alpha^*, \beta^*)$ . For example a common specification is to use the exponential function,

$$\sigma^2(X, \alpha^*) = \exp\{X\alpha^*\} \quad (6)$$

as this ensures that the variance is always positive. Show that the maximum likelihood estimate of  $\beta^*$  is a type of *weighted least squares* and that the observations are weighted by the *inverse of the conditional standard deviations*, i.e. the weight on the  $i^{\text{th}}$  observation in the weighted regression is  $w_i = 1/\sigma^2(X, \hat{\alpha})$  where  $\hat{\alpha}$  is the maximum likelihood estimate of  $\alpha^*$ .

- G. Suppose you ignored the heteroscedasticity and just estimated  $\beta^*$  by OLS. Would the OLS estimator be consistent? If so, what is the “efficiency penalty” of ignoring the heteroscedasticity in the  $\tilde{\varepsilon}_i$  relative to the most efficient (maximum likelihood) estimator? (Hint: to answer this, compare the asymptotic covariance matrix for OLS to the Cramer-Rao lower bound for  $\beta^*$  and show by a direct argument that the OLS asymptotic covariance matrix is “larger” than the asymptotic covariance matrix for the MLE, in the sense that the difference between the two covariances,  $D$  given by

$$D = \sigma^2 [E\{\tilde{X}'\tilde{X}\}]^{-1} - [I_{\beta^*, \beta^*}]^{-1} \quad (7)$$

is a positive semi-definite matrix.

- H. Suppose you cannot find a program to maximize the partial likelihood function when there is the exponential specification of conditional heteroscedasticity, i.e.  $\sigma^2(X, \alpha) = \exp\{X\alpha\}$  but I propose the following three step procedure as an alternative:

1. Do OLS and using the first stage OLS estimates for  $\hat{\beta}_{\text{OLS}}$ , compute the estimated squared residuals,  $\hat{u}_i^2 = (y_i - X\hat{\beta}_{\text{OLS}})^2, i = 1, \dots, N$ .
2. Do a regression with the log of these squared residuals as the dependent variables,

$$\log(\hat{u}_i^2) = X_i\alpha + v_i, \quad i = 1, \dots, N \quad (8)$$

to obtain the OLS estimates of  $\hat{\alpha}_{\text{OLS}}$ .

3. Using the OLS estimates  $\hat{\alpha}_{\text{OLS}}$  do a weighted least squares to obtain “second stage” estimates for  $\hat{\beta}_{\text{3swls}}$

$$\hat{\beta} = \underset{\beta \in R^K}{\operatorname{argmin}} \sum_{i=1}^N \left( \frac{y_i - X_i\beta}{\exp\{X_i\hat{\alpha}_{\text{OLS}}/2\}} \right)^2 \quad (9)$$

Discuss the pros and cons of this three step approach relative to maximum likelihood. Compare the asymptotic efficiencies of the three step “feasible weighted least squares” estimator of  $\beta$ ,  $\hat{\beta}_{\text{3swls}}$  to the maximum likelihood estimator,  $\hat{\beta}_{\text{MLE}}$ .

- I. The file `regression.out` contains a  $1000 \times 3$  data matrix where the first column contains  $N = 1000$  observations on the dependent variable  $y_i$ , and the second two columns contain  $X_{i,1}$  and  $X_{i,2}$ , two explanatory variables in the regression of interest

$$y_i = \beta_1 + \beta_2 X_{i,1} + \beta_3 X_{i,2} + \beta_4 X_{i,1} X_{i,2} + \beta_5 X_{i,1}^2 + \beta_6 X_{i,2}^2 + \varepsilon_i \quad (10)$$

where

$$\sigma^2(X_i, \alpha) = \exp\{\alpha_1 + \alpha_2 X_{i,1} + \alpha_3 X_{i,2} + \alpha_4 X_{i,1} X_{i,2} + \alpha_5 X_{i,1}^2 + \alpha_6 X_{i,2}^2\} \quad (11)$$

Estimate  $(\beta, \alpha)$  using both the partial maximum likelihood approach and the 3 stage estimator discussed in part H above and compare the point estimates and approximate standard errors (computed from the estimated asymptotic normal distribution, but adjusted for sample size  $N$ ). For each of these estimates, conduct a Chi-square test of the hypothesis  $H_0 : \alpha_l = 0, l = 2, \dots, 6$ . For each estimator (MLE and 3 step weighted least squares) report the *marginal significance level* of the test of this hypothesis.

**III.** The file `adaptreg.out` contains a  $5000 \times 6$  data matrix. The first four columns are dependent variables  $(y_1, \dots, y_4)$  in four identical regressions

$$y_i = a + b * x_1 + c * x_2 + \varepsilon_i \quad (12)$$

where  $y_i$  is the dependent variable in the regression and  $\varepsilon_i$  is the error term in the regression equation. I generated the  $\varepsilon_i$  ( $i = 1, \dots, 4$ ) from four potentially different unknown densities  $f_i(\varepsilon)$  which your job is to try to determine. I also want you to estimate, *as efficiently as you possibly can*, the three unknown regression coefficients  $\theta = (a, b, c)$ . For simplicity I have used the same  $x_1$  and  $x_2$  covariates in each regression and the only thing that changes is the error terms. The error terms  $\varepsilon_i$  were generated independently of the  $(x_1, x_2)$  values and are thus *IID* random variables.

1. Estimate the four regressions separate by OLS and compute the covariance matrices in each case. In which of the four cases are the variances of  $(a, b, c)$  the lowest?
2. Using the residuals from the four OLS regressions, compute four non-parametric densities for the  $\varepsilon_i$  and plot them at 500 equally spaced points on the interval  $[-8, 8]$ . Do any of the error terms appear to be normally distributed?
3. Try to guess the distributions  $f_i(\varepsilon)$  that I used to generate the error terms in the four cases.
4. Using the calculated non-parametric densities from part 2, compute four separate second stage *adaptive maximum likelihood estimates* of  $\theta = (a, b, c)$ . Compare your estimated variances of the  $\theta$  coefficients in this second stage to the variances of your OLS estimates in part 1 above. In which case are these variances smaller?
5. Suppose I allow you to *pool* the data from the four regressions into one *seemingly unrelated regression* with  $N = 4 * 5000 = 20000$  observations. Describe how to get the most efficient possible estimates of  $\theta$  by pooling your data, but taking into consideration that the error terms in the four regressions may be different, and thus, in the pooled data, the error terms may be *heteroscedastic*. Using your proposed method of efficient pooling of the data, compute your estimates of  $\theta$  and their estimated variances and compare them to your results in parts 1 and 4 above.
6. Suppose I told you that the four densities that I used to generate the error terms were 1) normal, 2) double exponential (Laplace), 3) uniform, and 4) triangular (where the latter two distributions have support on the interval  $[-8, 8]$ ). Describe whether you could use this information to obtain even more efficient estimates than would be possible using only the information given in part 5 above.

**IV.** Consider a simple *structural simultaneous equations model* of equilibrium in a commodity market, such as corn. Suppose we believe that demand for corn is a linear equation of the form

$$\begin{aligned} q_d &= a_d - b_d p + \varepsilon_d \\ q_s &= a_s + b_s p + \varepsilon_s \end{aligned} \tag{13}$$

In these equations,  $q_d$  is the quantity of corn demanded (in aggregate) by consumers and intermediaries (including ethanol demand), and  $q_s$  is the amount of domestic (and foreign) corn supplied to the U.S. market, and  $p$  is the market price of corn (we assume there is only a single market for a single “homogenous” quantity “bushel of corn”). Suppose that the error terms have a bivariate normal distribution where  $\varepsilon_d$  is independent of  $\varepsilon_s$  and both have mean zero and a diagonal covariance matrix, with each having the same variance  $\sigma^2$ . Thus the unknown parameter vector to be estimated is  $\theta = (a_d, b_d, a_s, b_s, \sigma^2)'$ , a  $5 \times 1$  vector of unknown “structural parameters”.

1. If we impose the maintained hypothesis that the corn market is *in equilibrium*, can you write down a likelihood function for the observations, if we have  $T$  time series observations on  $(q_{d,t}, q_{s,t}, p_t)$  where we assume equilibria in successive years are independent of each other and  $q_t = q_{d,t} = q_{s,t}$  is the aggregate quantity of corn produced and consumed each year?
2. The structural model (and the associated parameter vector  $\theta$ ) is *identified* if the expected log likelihood for the observed data is uniquely maximized at a “true” parameter vector  $\theta^*$ . Otherwise it is *unidentified* (including *partially identified*), if there is a set of parameters and not just a unique value  $\theta^*$  that maximizes the expected log likelihood function. In the latter case, the set of  $\theta$  that maximize the expected log likelihood function are said to be *observationally equivalent*. Is structural model of equilibrium in the corn market identified in this case?.
3. What if we allowed a non-diagonal covariance matrix for  $(\varepsilon_d, \varepsilon_s)$ ? Then we are trying to estimate the upper diagonal of the  $2 \times 2$  covariance matrix of  $(\varepsilon_d, \varepsilon_s)$ . Then instead of 5 unknown parameters, show there are 7 unknown parameters in  $\theta$  to be estimated. Is the model identified in this case?.
4. Suppose we extend the model as follows: rainfall levels  $r$  are known to affect supply but not aggregate demand, and per capita income  $y$  is known to affect demand but not supply. Now the parameter vector is  $\theta = (a_d, b_d, c_d, a_s, b_s, c_s, \sigma_d^2, \sigma_s^2, \sigma_{d,s})'$ , a  $9 \times 1$  vector of unknown “structural parameters” in the supply/demand specification given below

$$\begin{aligned} q_d &= a_d - b_d p + c_d y + \varepsilon_d \\ q_s &= a_s + b_s p + c_s r + \varepsilon_s \end{aligned} \tag{14}$$

Are these parameters identified in this case. Can you think of a simpler *instrumental variables* strategy for estimating the parameters? What would be the relevant *instruments* do use the IV/regression approach?

**V.** The file **x.dat** contains 20,000 observations from an unknown density that I would like you to try to estimate. Download these observations and see if you can infer what density I used to generate the x.dat file. To make your life easier I have provided a link to some Matlab code, `kdensity.m` and `denplot.m`, that will compute and plot a non-parametric kernel density estimate at 1000 points along an interval  $[a, b]$  where  $a$  and  $b$  are values you specify. In this problem, I am willing to tell you that the support of this

unknown (to you) density is  $[0,2]$ , so use `denplot.m` to plot the density at 1000 points equally spaced over the interval  $[0,2]$  and plot the results (denplot includes *Matlab* code that plots the computed density, so if you have *Matlab*, you can just use the `denplot` function to plot the density, you would call it as `denplot('x','x',0,2)`).

1. Define what we mean by a *kernel density estimator* of an unknown density  $f(x)$  at a point  $x$ . Write a formula for your estimator,  $\hat{f}(x)$  and define what is meant by the *bandwidth parameter*. Compare the kernel density estimate  $\hat{f}(x)$  with a naive *histogram estimator* of  $f(x)$ .
2. The `kdensity.m` programs assumes a Gaussian kernel and the Silverman “rule of thumb” choice of bandwidth,  $h$ . Show how the results vary by recomputing the density using a bandwidth 50% of the size computed by the Silverman rule, and 200% of this value.
3. Show how the results are affected if you use the Epanechnikov kernel instead of a Gaussian kernel that `denplot.m` uses (you can go to Wikipedia for the definition of the Epanechnikov kernel).
4. Suppose I tell you that I used a piece-wise linear density on the interval  $[0,2]$  to generate the  $x$  data. Suppose I also tell you that there are at most 4 segments to this piecewise linear density. Can you determine a more efficient way to estimate the unknown density than the nonparametric kernel density estimator?
5. Suppose I told you that the density I used might possibly be *discontinuous* as a function of  $x$ , at least at a finite number of points in the interval  $[0,2]$ . How would this knowledge affect your answers to the questions above?