

ECON 623
Solutions to Problem Set 4
 John Rust and Pablo Salanis Macarias

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

Answer As we discussed in class, pointwise convergence of functions is not sufficient to insure that $\lim_{n \rightarrow \infty} \theta_n^* = \theta^*$. For example, let $\Theta = [0, 1]$ (so $k = 1$) and let $f(x) = 2x$ so $\theta^* = \operatorname{argmax}_{\theta \in \Theta} f(\theta) = 1$. Define the sequence of functions $\{f_n\}$ by

$$f_n(x) = \max[f(x), t_n(x)], \tag{1}$$

where the functions $t_n(x)$ are the “tent functions” given by

$$t_n(x) = \begin{cases} nx & \text{if } x \in [0, 1/n] \\ 1 - n(x - 1/n) & \text{if } x \in (1/n, 2/n] \\ 0 & \text{if } x \in (2/n, 1] \end{cases} \tag{2}$$

Then it is easy to show that $\lim_{n \rightarrow \infty} t_n(x) = 0$ for each $x \in [0, 1]$ so $\{t_n\}$ converges pointwise to the zero function, but not uniformly (since $\max_{x \in [0, 1]} t_n(x) = 1$ for all $n \geq 1$). Further it is easy to see that $\theta_n^* = \operatorname{argmax}_{x \in [0, 1]} f_n(x) \rightarrow 0$, (i.e. the values of θ that maximize the $\{f_n\}$ functions converge to 0, and not to $\theta^* = 1$, the value of θ that maximizes the limiting function $f(x) = 2x$).

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 θ in this case?

Answer Yes, the result can be proved when $\{f_n\}$ converges to f uniformly. We use a proof by contradiction. Suppose that θ_0 is a limit point of the sequence $\{\theta_n^*\}$ (the sequence could have multiple limit points, and the argument below will hold for any one of them). Suppose that θ_0 is not a maximizer of the limiting function $f(\theta)$. Then it is the case that if θ^* is a maximizer of $f(\theta)$ we must have $f(\theta^*) > f(\theta_0)$. Let $\varepsilon = f(\theta^*) - f(\theta_0)$. We now show that this results in a contradiction. You can get an idea of why a contradiction must result from the schematic convergence diagram

$$\begin{array}{ccc} f_n(\theta_n^*) & \geq & f_n(\theta^*) \\ \downarrow & & \downarrow \\ f(\theta_0) & \geq & f(\theta^*) \end{array} \tag{3}$$

Uniform convergence of the $\{f_n\}$ implies that if $\{\theta_{n_j}^*\}$ is any subsequence of $\{\theta_n^*\}$ that converges to θ_0 , then $\lim_{j \rightarrow \infty} f_{n_j}(\theta_{n_j}^*) = f(\theta_0)$ (you should be able to show this using the *Triangle inequality* for norms). This is what we mean by the first downward pointing arrow in the schematic above. Also, since uniform convergence implies pointwise convergence, we have $\lim_{n \rightarrow \infty} f_n(\theta^*) = f(\theta^*)$. This is indicated by the 2nd downward arrow in the schematic diagram above. These convergence relations thus suggest that it must be the case that $f(\theta_0) \geq f(\theta^*)$, i.e. that θ_0 is also a maximizer of the limiting function f . To prove this rigorously, and account for the possibility that the sequence $\{\theta_n^*\}$ might

have multiple limit points, let $\{\theta_{n_j}^*\}$ be any subsequence of $\{\theta_n^*\}$ satisfying $\lim_{j \rightarrow \infty} \theta_{n_j}^* = \theta_0$. By uniform convergence, we can choose an \bar{J} sufficiently large so that

$$|f_{n_j}(\theta_{n_j}^*) - f(\theta_0)| < \varepsilon/3, \quad (4)$$

where $\varepsilon = f(\theta^*) - f(\theta_0) > 0$ is the amount which the maximum of f in the set Θ exceeds $f(\theta_0)$. To see this, use the Triangle inequality to get

$$\begin{aligned} |f_n(\theta_n^*) - f(\theta_0)| &= |f_n(\theta_n^*) - f(\theta_n^*) + f(\theta_n^*) - f(\theta_0)| \\ &\leq |f_n(\theta_n^*) - f(\theta_n^*)| + |f(\theta_n^*) - f(\theta_0)| \\ &\leq \|f_n(\theta_n^*) - f(\theta_n^*)\| + |f(\theta_n^*) - f(\theta_0)| \end{aligned} \quad (5)$$

where $\|f\| = \sup_{\theta \in \Theta} |f(\theta)|$. Since uniform convergence implies that $\lim_{n \rightarrow \infty} \|f_n - f\| = 0$, it means that the first term in the inequality (5) can be made as small as desired when n is sufficiently large, and since the $\{f_n\}$ are continuous, it is not hard to show that uniform convergence implies that the limit function f must also be continuous. Thus, for any sequence $\{\theta_n^*\}$ that converges to θ_0 , it follows that the second term on the right hand side of inequality (5) can be made as small as desired. Thus, if $\{\theta_{n_j}^*\}$ is a subsequence of $\{\theta_n^*\}$ that converges to θ_0 , the same reasoning can be shown that there must exist a \bar{J} sufficiently large that inequality (4) must hold for all $j \geq \bar{J}$.

Similarly, by uniform convergence, we can choose \bar{J} sufficiently large so that when $j \geq \bar{J}$ we have

$$|f_{n_j}(\theta^*) - f(\theta^*)| < \varepsilon/3 \quad (6)$$

since uniform convergence implies pointwise convergence. Now inequality (4) implies the following inequality holds

$$f(\theta_0) \geq f_{n_j}(\theta_{n_j}^*) - \frac{\varepsilon}{3} \quad (7)$$

for $j \geq \bar{J}$. But since $f(\theta_{n_j}^*) \geq f(\theta^*)$ (since $\theta_{n_j}^*$ is a maximizer of f_{n_j} whereas θ^* is not necessarily a maximizer of f_{n_j}), we have

$$f(\theta_0) \geq f_{n_j}(\theta^*) - \frac{\varepsilon}{3} \quad (8)$$

for $j \geq \bar{J}$. Similarly inequality (6) implies the following inequality

$$f_{n_j}(\theta^*) \geq f(\theta^*) - \frac{\varepsilon}{3} \quad (9)$$

Combining inequalities (8) and (9) we get the following inequality

$$f(\theta_0) \geq f(\theta^*) - \frac{2\varepsilon}{3} \quad (10)$$

Inequality (10) can be written as

$$\frac{2}{3}\varepsilon \geq f(\theta^*) - f(\theta_0) \quad (11)$$

However this is a contradiction, since if θ_0 is not a maximizer of f and θ^* is a maximizer of f , we defined ε as the difference between $f(\theta^*)$ and $f(\theta_0)$, i.e.

$$f^*(\theta^*) - f(\theta_0) = \varepsilon \quad (12)$$

But inequality (11) is clearly inconsistent with equation (12) and results in a contradiction. We conclude that the hypothesis that $f(\theta_0) < f(\theta^*)$ cannot hold because it results in a contradiction. It follows that $f(\theta_0) \geq f(\theta^*)$, i.e. the limit point θ_0 is a maximizer of the limiting function f .

- 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.

Answer A sequence of functions $\{f_n\}$ (or more generally a set of functions \mathcal{F}) is *uniformly equicontinuous* if and only if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for all $f \in \{f_n\}$ (or all $f \in \mathcal{F}$) we have

$$|f(\theta) - f(\theta')| < \varepsilon \quad \text{if } \|\theta - \theta'\| < \delta \quad (13)$$

Note that if \mathcal{F} consists of the singleton set $\{f\}$, then uniform equicontinuity just reduces to the notion of an *equicontinuous function*. So *uniform equicontinuity* just means that every function in the set \mathcal{F} must be equicontinuous and the inequality (13) holds for each $f \in \mathcal{F}$. An example of a non-equicontinuous set of functions is the sequence of tent functions $\{f_n\}$ in the first answer of this problem. They are not equicontinuous because the “tent” part implies a jump from 0 to 1 in a distance of only $1/n$ and thus there is no fixed value of $\delta < 0$ that insures that the jump is less than any small positive number ε for *all* values of n . An example of an equicontinuous set of functions is the sequence $\{f_n\}$ defined by

$$f_n(x) = \begin{cases} \frac{2x}{n} & \text{if } x \in [0, 1/2] \\ \frac{2(1-x)}{n} & \text{if } x \in (1/2, 1] \end{cases} \quad (14)$$

These are also “tent functions” but ones that converge uniformly to the zero function. It should not be hard to verify directly that this sequence is uniformly equicontinuous since the maximum absolute value of the slope of any f_n is $\frac{1}{2}$. Parts C and D below can be used to show that $\{f_n\}$ is *uniformly Lipschitz* with a Lipschitz constant of $L = \frac{1}{2}$.

- C. Suppose we know that $\{f_n\}$ is *uniformly Lipschitz continuous*, i.e. $\exists L < \infty$ such that $\forall n$ we have

$$|f_n(\theta) - f_n(\theta')| \leq L\|\theta - \theta'\| \quad (15)$$

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

Answer Using the definition in inequality (13) in part B, for any $\varepsilon > 0$ let $\delta = \varepsilon/L > 0$. Then clearly if $\|\theta - \theta'\| < \delta$ we have that $|f_n(\theta) - f_n(\theta')| < \varepsilon$ for all $n \geq 1$, so it follows that $\{f_n\}$ is a uniformly equicontinuous sequence of functions.

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

$$\sup_n \left\| \frac{\partial}{\partial \theta} f_n \right\| \leq K < \infty \quad (16)$$

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

Answer By the Mean Value Theorem (or Taylor’s Theorem without remainder) we have

$$f_n(\theta) - f_n(\theta') = \frac{\partial}{\partial \theta} f_n(\hat{\theta})(\theta - \theta') \quad (17)$$

where $\hat{\theta}$ is an element of Θ . The properties of norms imply

$$|f_n(\theta) - f_n(\theta')| = \left\| \frac{\partial}{\partial \theta} f_n(\hat{\theta}) \right\| \|\theta - \theta'\| \quad (18)$$

Define the constant L by

$$L = \sup_n \left\| \frac{\partial}{\partial \theta} f_n(\theta) \right\| \quad (19)$$

Then equations (17) and (19) imply the (uniform) Lipschitz inequality condition (15) of Part C, and hence it follows from Part C that $\{f_n\}$ is uniformly equicontinuous.

- E. Prove that if $f(\theta)$ is uniquely maximized at θ^* in the compact domain $\Theta \subset \mathbb{R}^k$ and if $\{f_n\}$ converges uniformly to f and if $\hat{\theta}_n$ maximizes f_n over $\theta \in \Theta$ then $\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta^*$.

Answer We already proved that if $\{\hat{\theta}_n\}$ is a sequence such that for each n , $\hat{\theta}_n$ is a maximizer of f_n , then if $\{f_n\}$ converges uniformly to f , then any limit point of the sequence $\{\hat{\theta}_n\}$ is a maximizer of the limiting function f . If the maximizer of f is the unique point $\theta^* \in \mathbb{R}^k$, then this implies any limit point of the sequence $\{\hat{\theta}_n\}$ converges to θ^* . Since any sequence of points in a compact set must have at least one limit point, and since all limit points must be equal to θ^* , the unique maximizer of f , it follows that the sequence of maximizers of $\{f_n\}$, $\{\hat{\theta}_n\}$ converges to θ^* , i.e. that

$$\lim_{n \rightarrow \infty} \hat{\theta}_n = \theta^* \quad (20)$$

- 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 θ) that converges *uniformly in θ 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*.

Answer Basically, in each of the parts above we make exactly the same arguments except in all the appropriate places we make the qualification “for almost all” or “with probability 1”. There are also “in probability” generalizations of parts A to E above, but the “almost sure” or “with probability 1” versions are the easiest to state. For example, stochastic equicontinuity can be stated as

Definition A sequence of random functions $\{f_n\}$ which are measurable functions $f_n(\omega, \theta)$ where $\omega \in \Omega$ and for each $\theta \in \Theta$ f_n is a measurable function of ω on the probability space (Ω, \mathcal{F}, P) and for almost all $\omega \in \Omega$ $f_n(\omega, \theta)$ is a continuous function of $\theta \in \Theta$ where Θ is a compact subset of a *metric space* with distance $d(\theta, \theta')$ is *stochastically equicontinuous* if for each $\varepsilon > 0$ there is a $\delta > 0$ such that for all n we have

$$|f_n(\omega, \theta) - f_n(\omega, \theta')| < \varepsilon \quad \text{if } d(\theta, \theta') < \delta \quad (21)$$

for $\omega \in \Gamma \subseteq \Omega$ satisfying $P(\Gamma) = 1$.

II. Consider the linear regression model

$$\tilde{y} = \tilde{X}\beta^* + \tilde{\varepsilon} \quad (22)$$

where $\tilde{\varepsilon} \sim N(0, \sigma^2)$ and \tilde{X} is a random $K \times 1$ vector with density $f(X, \gamma^*)$, and γ 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 $L_N(\theta)$.

Answer

$$L_N(\theta) = \frac{1}{N} \sum_{i=1}^N \log (\phi(y_i|X_i, \beta, \sigma^2) f(X_i, \gamma)) \quad (23)$$

where $\phi(y|X, \beta, \sigma^2)$ is the conditional Gaussian density

$$\phi(y|X, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(y - X\beta)^2/2\sigma^2) \quad (24)$$

- B. In many studies, the analyst is not so interested in the parameters γ^* of the distribution of covariates $f(X, \gamma^*)$ and is really only interested in the parameters (σ^2, β) . The (average of the log) of the *partial likelihood function* is given by

$$L_N(\sigma^2, \beta) = \frac{1}{N} \sum_{i=1}^N \log(f(y_i|X_i, \sigma^2, \beta)) \quad (25)$$

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 σ^2 . Show that even though we estimate (σ^2, β) using this partial likelihood function (and thus ignore the parameter γ and information contained in the X_i observations that is reflected in the full likelihood function above), that computing the maximum likelihood estimates for (σ^2, β) 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.

Answer By the basic property of the log function, $\log(A \times B) = \log(A) + \log(B)$ we have

$$L_N(\theta) = L_N(\beta, \sigma^2) + L_N(\gamma) \quad (26)$$

where

$$\begin{aligned} L_N(\beta, \sigma^2) &= \frac{1}{N} \sum_{i=1}^N \log (\phi(y_i|X_i, \beta, \sigma^2)) \\ L_N(\gamma) &= \frac{1}{N} \sum_{i=1}^N \log (f(X_i, \gamma)) \end{aligned} \quad (27)$$

The functions $L_N(\beta, \sigma^2)$ and $L_N(\gamma)$ are called *partial likelihood functions* whereas $L_N(\theta)$ is called the *full likelihood function*. Clearly any $(\hat{\beta}, \hat{\sigma}^2)$ that maximizes the partial likelihood $L_N(\beta, \sigma^2)$ and any $\hat{\gamma}$ that maximizes $L_N(\gamma)$ also maximize the full likelihood function $L_N(\theta)$ due to the multiplicative separability of the functions and the property of the log function.

- 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 γ 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 (σ^2, β^*) ?

Answer Since the (partial) maximum likelihood estimator is numerically identical to the (full) maximum likelihood estimators for (β, σ^2) and γ , the maximization of the two partial likelihoods $L_N(\beta, \sigma^2)$ and $L_N(\gamma)$ separately will result in consistent and asymptotically efficient parameter estimates of (β, σ^2) and γ and the information matrix in this case will be *block diagonal*. The key exception to this general statement is if there happen to be *overlapping parameters* such as β parameters that are components of γ or vice versa. Then, which it can be shown that maximization of the partial likelihood function will still result in *consistent* parameter estimates, it will also be true generally that the resulting partial likelihood estimator will be *less efficient* than the full information maximum likelihood estimator. For example maximizing only $L_N(\beta, \sigma^2)$ over (β, σ^2) will result in consistent but less efficient parameter estimates compared to maximizing the full likelihood function $L_N(\beta, \sigma^2, \gamma)$ if some of the β parameters or σ^2 also enter into the γ parameters. The information matrix will not be block-diagonal in this case and there will be a *loss of information* and hence a *loss of asymptotic efficiency* (i.e. higher asymptotic variance) from maximizing the partial likelihood $L_N(\beta, \sigma^2)$ instead of the full likelihood function.

- D. Compute the *Information Matrix* I for the partial likelihood function and show that it is *block diagonal* between the parameters σ^2 and β^* . That is, show that if $I_{\sigma^2, \beta^* j}$ is the element of the Information matrix corresponding to the pair of parameters (σ^2, β_j^*) that we have $I_{\sigma^2, \beta^* j} = 0$, $j = 1, \dots, K$. 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 $\hat{\beta}_j$.

Answer The i^{th} in the summation expression for the partial log-likelihood function for (β, σ^2) in equation (27) is

$$\log(\phi(y_i|X_i, \beta, \sigma^2)) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - (y_i - X_i\beta)^2 / (2\sigma^2) \quad (28)$$

The partial derivatives of this with respect to σ^2 and β are

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \log(\phi(y_i|X_i, \beta, \sigma^2)) &= -\frac{1}{2\sigma^2} + (y_i - X_i\beta)^2 / 2\sigma^4 \\ \frac{\partial}{\partial \beta} \log(\phi(y_i|X_i, \beta, \sigma^2)) &= -X_i'(y_i - X_i\beta) / \sigma^2 \end{aligned} \quad (29)$$

It is not hard to show that at the true parameter values, $\theta^* = (\beta^*, \sigma^{*2})$ that the expectations of the gradient of the partial log likelihood function (29) is equal to zero

$$\begin{aligned} E \left\{ \frac{\partial}{\partial \sigma^2} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right\} &= -\frac{1}{2\sigma^2} + E \{ (y_i - X_i\beta) / 2\sigma^4 \} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^2} = 0, \\ E \left\{ \frac{\partial}{\partial \beta} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right\} &= -E \{ X_i'(y_i - X_i\beta) / \sigma^2 \} = -E \{ X_i'\epsilon \} / \sigma^2 = 0. \end{aligned} \quad (30)$$

Now consider the Information matrix, I , which can be partitioned as

$$I = \begin{bmatrix} I_{\beta\beta'} & I_{\beta\sigma^2} \\ I_{\sigma^2\beta'} & I_{\sigma^2\sigma^2} \end{bmatrix} \quad (31)$$

We want to show that the off-diagonal block of the partitioned information matrix is 0, i.e. that $I_{\beta\sigma^2} = 0$. Writing this out explicitly, we have

$$\begin{aligned} I_{\beta\sigma^2} &= E \left\{ \left[\frac{\partial}{\partial \beta} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right] \left[\frac{\partial}{\partial \sigma^2} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right] \right\} \\ &= \frac{1}{4\sigma^4} E \{ X_i'(y_i - X_i\beta) \} - \frac{1}{4\sigma^6} E \{ X_i'(y_i - X_i\beta)^3 \} \\ &= 0 \end{aligned} \quad (32)$$

We used the Law of Iterated Expectations to derive equation (32) and specifically $\varepsilon_i = (y_i - X_i\beta^*)$

$$\begin{aligned} E \{ X_i'(y_i - X_i\beta^*) \} &= E \{ X_i'\varepsilon_i \} \\ &= E \{ X_i' E \{ \varepsilon_i | X_i \} \} \\ &= E \{ X_i' 0 \} \\ &= 0. \end{aligned} \quad (33)$$

Similarly, we have

$$\begin{aligned} E \{ X_i'(y_i - X_i\beta^*)^3 \} &= E \{ X_i'\varepsilon_i^3 \} \\ &= E \{ X_i' E \{ \varepsilon_i^3 | X_i \} \} \\ &= E \{ X_i' 0 \} \\ &= 0, \end{aligned} \quad (34)$$

since the conditional expectation of $E \{ \varepsilon_i^3 | X_i \}$ is zero since ε_i^3 is an odd function and the Gaussian density is an even (symmetric) density about zero, so the expectation of an even times an odd function results in an odd function (antisymmetric about 0) that has expectation zero.

E. For the full likelihood function, show that there is also block diagonality for the parameters σ^2 , β and γ , i.e. we have for the following elements of the full information matrix I

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

Use this to show that there is no loss in asymptotic efficiency in estimating (σ^2, β^*) 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 (σ^2, β^*) .

Answer We won't do all of the various combinations here since we already showed block diagonality between σ^2 and β in the partial likelihood function. We will show that $I_{\gamma\beta} = 0$ also by appealing to

the Law of Iterated Expectations.

$$\begin{aligned}
I_{\gamma\beta} &= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] \left[\frac{\partial}{\partial \beta} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right] \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] E \left\{ \left[\frac{\partial}{\partial \beta} \log(\phi(y_i|X_i, \beta, \sigma^2)) \right] |X_i \right\} \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] [E \{-X_i'(y_i - X_i\beta)/2\sigma^2 |X_i\}] \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] [-X_i' E \{(y_i - X_i\beta) |X_i\} / 2\sigma^2] \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] [-X_i' E \{\epsilon_i |X_i\}] / 2\sigma^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \gamma} \log(f(X_i, \gamma)) \right] [-X_i' 0] / 2\sigma^2 \right\} \\
&= 0.
\end{aligned} \tag{35}$$

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

$$\text{var}\{\tilde{\epsilon}|X\} = \sigma^2(X, \alpha^*), \tag{36}$$

where α^* is a $H \times 1$ vector of unknown parameters to be estimated along with β^* (assume given the result in part E above, we are not interested in estimating γ^* 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,

$$\sigma^2(X, \alpha^*) = \exp(X\alpha^*) \tag{37}$$

as this ensures that the variance is always positive. Show that the maximum likelihood estimate of β^* 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^{th} observation in the weighted regression is $w_i = 1/\sqrt{\sigma^2(X, \hat{\alpha})} = \exp(-X\hat{\alpha}/2)$ where $\hat{\alpha}$ is the maximum likelihood estimate of α^* .

Answer The partial loglikelihood with heteroscedasticity is

$$\begin{aligned}
L_N(\beta, \alpha) &= \frac{1}{N} \sum_{i=1}^N \log(\phi(y_i|X_i, \beta, \alpha)) \\
&= \frac{1}{N} \sum_{i=1}^N \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} X_i \alpha - (y_i - X_i \beta)^2 / 2 \exp(X_i \alpha) \right]
\end{aligned} \tag{38}$$

Let $\hat{\alpha}$ be the maximum likelihood estimate of α . It is easy to see from the partial log likelihood function (38) that $\hat{\beta}$ is the solution to a weighted least squares problem with $1/\exp(X_i \hat{\alpha})$ constituting the weights

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^N (y_i - X_i \beta)^2 / \exp(X_i \hat{\alpha}). \tag{39}$$

This is a weighted least squares problem.

- G. Suppose you ignored the heteroscedasticity and just estimated β^* by OLS. Would the OLS estimator be consistent? If so, what is the “efficiency penalty” of ignoring the heteroscedasticity in the $\tilde{\epsilon}_i$ 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 β^* 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

$$D = \sigma^2 [E\{\tilde{X}'\tilde{X}\}]^{-1} - [I_{\beta^*\beta^*}]^{-1} \quad (40)$$

is a positive semi-definite matrix.

Answer Recall that the asymptotic covariance matrix of the maximum likelihood estimator of β^* is the $K \times K$ matrix Ω_{mle} given by

$$\Omega_{mle} = [I_{\beta^*\beta^*}]^{-1}. \quad (41)$$

This is the “Cramer-Rao lower bound” and we will compute an explicit formula for this lower bound below. Note that in equation (40) the formula $\sigma^2[E\{\tilde{X}'\tilde{X}\}]^{-1}$ is *not* the correct asymptotic covariance matrix for the OLS estimator in the presence of heteroscedasticity. The correct asymptotic covariance matrix Ω_{ols} is

$$\text{var}\{\sqrt{N}(\hat{\beta}_{ols} - \beta^*)\} = \Omega_{ols} = [E\{\tilde{X}'\tilde{X}\}]^{-1} E\{\exp(\tilde{X}'\alpha^*)\tilde{X}'\tilde{X}\}[E\{\tilde{X}'\tilde{X}\}]^{-1}. \quad (42)$$

Only in the special case of *homoscedasticity*, i.e. where $\text{var}(\epsilon|X) = \sigma^2(X, \alpha^*) = \exp(\tilde{X}'\alpha^*) = \sigma^2$, is it the case that the asymptotic covariance matrix of $\sqrt{N}(\hat{\beta} - \beta^*)$ is $\sigma^2[E\{\tilde{X}'\tilde{X}\}]^{-1}$. But then, as we show below, the right hand side of inequality (40), $I_{\beta^*\beta^*}$ will also equal $\sigma^2[E\{\tilde{X}'\tilde{X}\}]^{-1}$. In this case $D = 0$, which is trivially positive semi-definite.

When there is heteroscedasticity, i.e. when it is not the case that $\exp(\tilde{X}'\alpha^*)$ equals a single value σ^2 with probability 1, the inequality we need to show is actually

$$D = [E\{\tilde{X}'\tilde{X}\}]^{-1} [E\{\exp(\tilde{X}'\alpha^*)\tilde{X}'\tilde{X}\}] [E\{\tilde{X}'\tilde{X}\}]^{-1} - [I_{\beta^*\beta^*}]^{-1} \geq 0 \quad (43)$$

where 0 is interpreted as a $K \times K$ matrix of zeros, and $D \geq 0$ is just a short hand for saying that the matrix D is positive semi-definite. We want to show that the D matrix in the latter, correct, formula (43) for the difference between the asymptotic variance-covariance matrix of the OLS estimator and that of the maximum likelihood estimator is positive semi-definite when there is heteroscedasticity. This implies that the OLS estimator will be consistent but less efficient than the maximum likelihood estimator in the presence of heteroscedasticity. First, we derive an explicit formula for $I_{\beta^*\beta^*}$. We have

$$\begin{aligned} I_{\beta^*\beta^*} &= E\{\tilde{X}'\epsilon^2\tilde{X}/\exp(2\tilde{X}'\alpha^*)\} \\ &= E\{E\{\epsilon^2|\tilde{X}\}\tilde{X}'\tilde{X}/\exp(2\tilde{X}'\alpha^*)\} \\ &= E\{\exp(\tilde{X}'\alpha^*)\tilde{X}'\tilde{X}/\exp(2\tilde{X}'\alpha^*)\} \\ &= E\{\exp(-\tilde{X}'\alpha^*)\tilde{X}'\tilde{X}\} \end{aligned} \quad (44)$$

Since the information matrix is block-diagonal (as we showed above), the covariance matrix of the maximum likelihood estimator, $\hat{\beta}$, is the inverse of $I_{\beta^*\beta^*}$. This is equals (leaving out the i subscripts)

$[E\{\exp(-\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}]^{-1}$. By the Cramer-Rao Theorem, maximum likelihood is asymptotically efficient, so it follows that the asymptotic covariance matrix of the OLS estimator, which is not efficient when there is heteroscedasticity, is greater than the asymptotic covariance of the maximum likelihood estimator, so the difference matrix D in equation (43) is positive semi-definite.

To show this via a direct argument, consider the asymptotic distribution of the following $2K \times 1$ vector

$$\begin{bmatrix} \sqrt{N}(\hat{\beta}_{ols} - \beta^*) \\ \sqrt{N}\frac{\partial}{\partial \beta}L_N(\beta^*, \alpha^*) \end{bmatrix} \implies N(0, \Gamma) \quad (45)$$

where $L_N(\beta^*, \alpha^*)$ is the partial log-likelihood function (38) and Γ is the $2K \times 2K$ matrix

$$\Gamma = \begin{bmatrix} [E\{\tilde{X}'\tilde{X}\}]^{-1}[E\{\exp(\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}][E\{\tilde{X}'\tilde{X}\}]^{-1} & I \\ I & E\{\exp(-\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\} \end{bmatrix} \quad (46)$$

We will show why the off-diagonal blocks of Γ equal the $K \times K$ identity matrix I shortly. But for now, focus on the main implication: since Γ is a covariance matrix, it is positive semi-definite. Given any $\gamma \in R^K$, define the vector $\lambda \in R^{2K}$ by

$$\lambda = \begin{bmatrix} \gamma \\ [-E\{\exp(-\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}]^{-1}\gamma \end{bmatrix} \quad (47)$$

Then since Γ is positive semi-definite we have $\lambda'\Gamma\lambda \geq 0$. Writing out this quadratic form, we have

$$\begin{aligned} \lambda'\Gamma\lambda &= \gamma' [E\{\tilde{X}'\tilde{X}\}]^{-1}[E\{\exp(\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}][E\{\tilde{X}'\tilde{X}\}]^{-1} - [E\{\exp(\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}]^{-1} \gamma \\ &= \gamma'D\gamma \geq 0. \end{aligned} \quad (48)$$

Since γ is an arbitrary vector in R^K , it follows that the difference in the OLS asymptotic covariance matrix and the Cramer-Rao lower bound D in equation (43) is positive semidefinite, as predicted by the Cramer-Rao Theorem.

We have one loose end to tie down. As promised above, we need to calculate the asymptotic covariance matrix for $\sqrt{N}(\hat{\beta}_{ols} - \beta^*)$ and $\sqrt{N}\frac{\partial}{\partial \beta}L_N(\beta^*, \alpha^*)$ and show it equals the $K \times K$ identity matrix. We can write

$$\begin{bmatrix} \sqrt{N}(\hat{\beta}_{ols} - \beta^*) \\ \sqrt{N}\frac{\partial}{\partial \beta}L_N(\beta^*, \alpha^*) \end{bmatrix} = \begin{bmatrix} [\frac{1}{N}\sum_{i=1}^N X_i'X_i]^{-1} [\frac{1}{\sqrt{N}}\sum_{i=1}^N X_i'\epsilon_i] \\ \frac{1}{\sqrt{N}}\sum_{i=1}^N \exp(-X_i\alpha^*)X_i'\epsilon_i \end{bmatrix} \quad (49)$$

We apply the Central Limit Theorem to the independently distributed random vectors entering the normalized sums in the expression on the right hand side of equation (49) to get

$$\begin{bmatrix} \sqrt{N}(\hat{\beta}_{ols} - \beta^*) \\ \sqrt{N}\frac{\partial}{\partial \beta}L_N(\beta^*, \alpha^*) \end{bmatrix} \implies \tilde{Z} \sim N(0, \Gamma) \quad (50)$$

where 0 is a $2K \times 1$ vector of zeros and Γ is a $2K \times 2K$ variance-covariance matrix that can be partitioned as

$$\Gamma = \begin{bmatrix} \Omega_{ols} & C \\ C & I_{\beta^*\beta^*} \end{bmatrix} \quad (51)$$

The $K \times K$ matrix C in equation (51) is the asymptotic covariance of $\sqrt{N}(\hat{\beta}_{ols} - \beta^*)$ and $\sqrt{N} \frac{\partial}{\partial \beta} L_N(\beta^*, \alpha^*)$. We can compute this covariance as the covariance of the terms in the normalized sums in equation (49) above. That is,

$$\begin{aligned}
C &= \text{cov}([E\{\tilde{X}'\tilde{X}\}]^{-1}\tilde{X}'\varepsilon, \exp(-\tilde{X}\alpha^*)\tilde{X}'\varepsilon) \\
&= E\{[E\{\tilde{X}'\tilde{X}\}]^{-1}[\tilde{X}'\varepsilon^2 \exp(-\tilde{X}\alpha^*)\tilde{X}]\} \\
&= [E\{\tilde{X}'\tilde{X}\}]^{-1} [E\{E\{\varepsilon^2|\tilde{X}\} \exp(-\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}] \\
&= [E\{\tilde{X}'\tilde{X}\}]^{-1} [E\{\exp(\tilde{X}\alpha^*) \exp(-\tilde{X}\alpha^*)\tilde{X}'\tilde{X}\}] \\
&= [E\{\tilde{X}'\tilde{X}\}]^{-1} [E\{\tilde{X}'\tilde{X}\}] \\
&= I.
\end{aligned} \tag{52}$$

The argument we used above to show that OLS is inefficient relative to the maximum likelihood estimator (which achieves the Cramer-Rao lower bound) is a general one and is not specific to this example of the normal regression model with heteroscedasticity. Though I certainly did not expect this as part of the answer, here is how the argument works in the general case.

Cramer-Rao Lower Bound Consider IID observations $\{x_i\}$ from a density $f(x, \theta^*)$ that satisfies certain regularity conditions, particularly that θ^* is an interior point of a compact parameter space $\Theta \subset \mathbb{R}^K$ and that $f(x, \theta)$ is twice continuously differentiable with respect to θ for almost all x , and that all relevant moments of $\log(f(\tilde{x}, \theta))$ exist including the expectation of the gradient $\frac{\partial}{\partial \theta} f(\tilde{x}, \theta)$ and the hessian $\frac{\partial^2}{\partial \theta \partial \theta'} f(\tilde{x}, \theta)$ for all $\theta \in \text{int}(\Theta)$. Let $\hat{\theta}_N$ be any CAN estimator of θ^* based on N IID observations $\{x_i\}_{i=1}^N$ from $f(x, \theta^*)$ (where CAN is an abbreviation for Consistent, Asymptotically Normal) then if Ω is the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}_N - \theta^*)$ we have

$$D = \Omega - [I_{\theta\theta'}]^{-1} \geq 0 \tag{53}$$

i.e. the difference between the asymptotic covariance matrix for $\sqrt{N}(\hat{\theta}_N - \theta^*)$ and the maximum likelihood estimator $\sqrt{N}(\hat{\theta}_{N,mle} - \theta^*)$ is a positive semidefinite matrix D .

Proof As we did above, we want to show that

$$\begin{bmatrix} \sqrt{N}(\hat{\theta}_n - \theta^*) \\ \sqrt{N} \frac{\partial}{\partial \theta} L_N(\theta^*) \end{bmatrix} \implies \tilde{Z} \sim N(0, \Gamma) \tag{54}$$

where $L_N(\theta)$ is the log-likelihood function

$$L_N(\theta) = \frac{1}{N} \sum_{i=1}^N \log(f(x_i, \theta)) \tag{55}$$

and 0 is a $2K \times 1$ vector of zeros, and Γ is a $2K \times 2K$ variance-covariance matrix that can be partitioned as

$$\Gamma = \begin{bmatrix} \Omega & C \\ C & I_{\theta^*\theta^*} \end{bmatrix}. \tag{56}$$

If we can show that $C_N = \text{cov}(\sqrt{N}(\hat{\theta}_N - \theta^*), \frac{\partial}{\partial \theta} L_N(\theta^*)) \rightarrow I$ as $N \rightarrow \infty$, then we can apply the same argument as we did above in the normal regression with heteroscedasticity (see equation (48))

to show that $D \geq 0$. So to show that $C = I$, first note that since $\hat{\theta}_N$ is a consistent estimator by assumption, it must *converge in mean* to θ for any $\theta \in \Theta$. That is, we have

$$\lim_{N \rightarrow \infty} E\{\hat{\theta}_N|\theta\} = \theta. \quad (57)$$

To make sure you understand formula (57), note that we use the conditional expectation notation to reflect the possibility that there are different possible values of the “true parameter” θ^* . If $\hat{\theta}_N$ is a CAN estimator of a parameter θ^* , its expectation will generally depend on the value of the true parameter θ^* . We want to consider estimators that “work” regardless of what the value of the true parameter θ^* might be, and thus in the discussion below we drop the $*$ superscript and simply treat θ as the “true” parameter value. That is, we assume that the data we observe are *IID* draws from the density $f(x, \theta)$. Thus we can write $E\{\hat{\theta}_N|\theta\}$ as a more explicit notation that emphasizes the dependence of the expectation of $\hat{\theta}_N$ on value of θ .

We do impose additional regularity conditions to insure that regardless of what the true value θ might be, it should be *estimable*. Another way of saying this is that we need to impose another key regularity condition, namely, that as we vary the true parameter θ , it should always remain *identified*. This is mathematically equivalent to the requirement that the expectation of the random function of theta $\log(f(\tilde{x}, \theta))$ is *uniquely* maximized at a value of θ that equals the true parameter θ regardless of which $\theta \in \Theta$ we happen to pick, i.e.

$$\forall \theta \in \Theta \quad \theta = \underset{\theta' \in \Theta}{\operatorname{argmax}} E\{\log(f(\tilde{x}, \theta'))|\theta\} = \underset{\theta' \in \Theta}{\operatorname{argmax}} \int \log(f(x, \theta')) f(x, \theta) dx \quad (58)$$

and

$$\forall \theta' \in \Theta \quad \text{if } \theta' \neq \theta \quad \text{then } E\{\log(f(\tilde{x}, \theta'))|\theta\} < E\{\log(f(\tilde{x}, \theta)|\theta)\}. \quad (59)$$

Let’s assume that the identification does hold, we now consider taking the gradient of $E\{\hat{\theta}_N|\theta\}$. If we take the gradient of $E\{\hat{\theta}_N|\theta\}$ with respect to θ on both sides of equation (57) above, we get

$$\begin{aligned} \frac{\partial}{\partial \theta} \left[\lim_{N \rightarrow \infty} E\{\hat{\theta}_N|\theta\} \right] &= \frac{\partial}{\partial \theta} \theta \\ \lim_{N \rightarrow \infty} \left[\frac{\partial}{\partial \theta} E\{\hat{\theta}_N|\theta\} \right] &= I. \end{aligned} \quad (60)$$

We now show that

$$C_N = \operatorname{cov}(\sqrt{N}(\hat{\theta}_N - \theta), \sqrt{N} \frac{\partial}{\partial \theta} L_N(\theta)) = \left[\frac{\partial}{\partial \theta} E\{\hat{\theta}_N|\theta\} \right], \quad (61)$$

so equation (60) implies that $\lim_{N \rightarrow \infty} C_N = C = I$, which establishes the result. So we only have to show that equation (61) holds. Note that $E\{\hat{\theta}_N\}$ can be written explicitly as

$$E\{\hat{\theta}_N\} = \int \cdots \int \left[\hat{\theta}_N(x_1, \dots, x_N) \prod_{i=1}^N f(x_i, \theta) \right] dx_1 dx_2 \cdots dx_N, \quad (62)$$

where we have written the estimator as a function of the data $\hat{\theta}_N(x_1, \dots, x_N)$ to emphasize that any estimator is just some (measurable) function of the data, and given that $\{x_i\}$ are *IID*/ observations

from the density $f(x, \theta)$, the joint density of $\{x_i\}$ is just $\prod_{i=1}^N f(x_i, \theta)$. Now taking the gradient of $E\{\hat{\theta}_N\}$ with respect to θ we have

$$\frac{\partial}{\partial \theta} E\{\hat{\theta}_N\} = \int \cdots \int \hat{\theta}_N(x_1, \dots, x_N) \left[\frac{\partial}{\partial \theta} \prod_{i=1}^N f(x_i, \theta) \right] dx_1 \cdots dx_N. \quad (63)$$

the partial derivative of this expectation with respect to the hypothesized true parameter value θ . Equation (63) is an explicit formula for this derivative. Now we write

$$\prod_{i=1}^N f(x_i, \theta) = \exp \left(\sum_{i=1}^N \log(f(x_i, \theta)) \right) \quad (64)$$

So we have

$$\frac{\partial}{\partial \theta} \prod_{i=1}^N f(x_i, \theta) = \left[\sum_{i=1}^N \frac{\partial}{\partial \theta} \log(f(x_i, \theta)) \right] \left[\prod_{i=1}^N f(x_i, \theta) \right] dx_1 \cdots dx_N. \quad (65)$$

So using equations (60) and (63) we obtain

$$\begin{aligned} \frac{\partial}{\partial \theta} E\{\hat{\theta}_N | \theta\} &= \int \cdots \int \hat{\theta}_N(x_1, \dots, x_N) \left[\frac{\partial}{\partial \theta} \prod_{i=1}^N f(x_i, \theta) \right] dx_1 \cdots dx_N \\ &= \int \cdots \int \left[\sum_{i=1}^N \frac{\partial}{\partial \theta} \log(f(x_i, \theta)) \right] \left[\prod_{i=1}^N f(x_i, \theta) \right] dx_1 \cdots dx_N \\ &= NE \left\{ \hat{\theta}_N \frac{\partial}{\partial \theta} L_N(\theta) | \theta \right\} \\ &= E \left\{ \sqrt{n} \hat{\theta}_N \sqrt{N} \frac{\partial}{\partial \theta} L_N(\theta) | \theta \right\} \\ &= E \left\{ \sqrt{n} (\hat{\theta}_N - \theta) \sqrt{N} \frac{\partial}{\partial \theta} L_N(\theta) | \theta \right\} \\ &= \text{cov}(\sqrt{N}(\hat{\theta}_N - \theta), \sqrt{N} \frac{\partial}{\partial \theta} L_N(\theta)). \end{aligned} \quad (66)$$

In deriving equation (66) we used the result that $E\{\frac{\partial}{\partial \theta} L_N(\theta) | \theta\} = 0$. This follows from the identification assumption that θ is the unique maximizer over all $\theta \in \Theta$ of the function $g(\theta') = E\{\log(\tilde{x}, \theta') | \theta\}$. We also used the result that $0 = E\{\theta \frac{\partial}{\partial \theta} L_N(\theta) | \theta\}$. This result holds because the expectation is over the random variables (x_1, \dots, x_N) , so that θ factors out as follows

$$\begin{aligned} E \left\{ \theta \frac{\partial}{\partial \theta} L_N(\theta) | \theta \right\} &= \theta E \left\{ \frac{\partial}{\partial \theta} L_N(\theta) | \theta \right\} \\ &= \theta \cdot 0 \\ &= 0, \end{aligned} \quad (67)$$

since as we already noted, $\forall \theta \in \text{int}(\Theta)$ we have $E\{\log(f(\tilde{x}, \theta') | \theta)\}$ is uniquely maximized at an interior point of Θ , so this implies that $E\{\frac{\partial}{\partial \theta} \log(f(\tilde{x}, \theta) | \theta)\} = 0$ which implies also that $E\{\frac{\partial}{\partial \theta} L_N(\theta) | \theta\} = 0$.

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

$$\log(\hat{u}_i^2) = X_i\alpha + v_i, \quad i = 1, \dots, N \quad (68)$$

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

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

$$\hat{\beta} = \underset{\beta \in R^K}{\operatorname{argmin}} \sum_{i=1}^N \left(\frac{y_i - X_i\beta}{\exp(X_i\hat{\alpha}_{OLS}/2)} \right)^2 \quad (69)$$

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 β , $\hat{\beta}_{3swls}$ to the maximum likelihood estimator, $\hat{\beta}_{mle}$.

Answer The three step procedure outlined above will generally not produce consistent estimates of α^* and will generate inefficient of β^* . The reason is that the regression (68) will generally not produce consistent estimates of α^* and without consistent estimates of α^* the third stage weighted least squares regression (69) will not be using the right weights to achieve the Cramer-Rao lower bound.

Why does the regression (68) result in inconsistent parameter estimates of α^* ? It is not due to the use of the (log of) the squared regression residuals \hat{u}_i instead of the log of the square of the true regression error terms ε_i as the dependent variable in equation (68). Even if we were using the log of squares of the true error terms ε_i , we show that a regression will result in inconsistent parameter estimates of α^* . Note that the following regression equation *does* hold

$$\varepsilon_i^2 = \exp(X_i\alpha^*) + \eta_i \quad (70)$$

where $E\{\eta_i|X_i\} = 0$. This holds by construction since we assumed that under the true data generating mechanism, the conditional variance of y_i , which is also the conditional variance of ε_i is $\exp(X_i\alpha^*)$. But since $E\{\varepsilon_i|X_i\} = 0$, it follows that the conditional variance of ε_i is equal to $E\{\varepsilon_i^2|X_i\}$. Then equation (70) is just telling us that the realized squared error term ε_i equals its conditional expectation $\exp(X_i\alpha^*)$ plus an error term η_i , and by the properties of conditional expectations and the Law of Iterated Expectations, it follows that $E\{\eta_i|X_i\} = 0$.

However it is invalid to take logs of both sides of equation (70) and proclaim that $E\{\log(\varepsilon_i^2)|X_i\} = X_i\alpha^*$, ignoring the presence of the error term η_i . In fact, we can use Jensen’s inequality to show that

$$E\{\log(\varepsilon_i^2)|X_i\} < X_i\alpha^* \quad (71)$$

so it follows that even in the best case where we were able to use the true error terms ε_i , the second stage regression in equation (68) will generally not be consistent for α^* .

However if we did the *nonlinear regression* below using the squares of the first stage regression residuals $\{\hat{u}_i\}$, then we can show that this nonlinear regression generally will result in consistent estimates of α^* .

$$\hat{u}_i^2 = \exp(X_i\alpha^*) + e_i \quad (72)$$

- I. The file `regression.out` contains a 1000×3 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

$$y_i = \beta_1 + \beta_2 X_{i,1} + \beta_3 X_{i,2} + \beta_4 X_{i,1} X_{i,2} + \beta^5 X_{i,1}^2 + \beta_6 X_{i,2}^2 + \varepsilon_i \quad (73)$$

where

$$\sigma^2(X_i, \alpha) = \exp(\alpha_1 + \alpha_2 X_{i,1} + \alpha_3 X_{i,2} + \alpha_4 X_{i,1} X_{i,2} + \alpha_5 X_{i,1}^2 + \alpha_6 X_{i,2}^2) \quad (74)$$

Estimate (β, α) 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 5000×6 data matrix. The first four columns are dependent variables (y_1, \dots, y_4) in four identical regressions

$$y_i = a + b * x_1 + c * x_2 + \varepsilon_i \quad (75)$$

where y_i is the dependent variable in the regression and ε_i is the error term in the regression equation. I generated the ε_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 ε_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 ε_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 θ 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 θ 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 θ 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

$$\begin{aligned} q_d &= a_d - b_d p + \varepsilon_d \\ q_s &= a_s + b_s p + \varepsilon_s \end{aligned} \tag{76}$$

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 “homogeneous” quantity “bushel of corn”). Suppose that the error terms have a bivariate normal distribution where ε_d is independent of ε_s and both have mean zero and a diagonal covariance matrix, with each having the same variance σ^2 . Thus the unknown parameter vector to be estimated is $\theta = (a_d, b_d, a_s, b_s, \sigma^2)'$, a 5×1 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 θ) is *identified* if the expected log likelihood for the observed data is uniquely maximized at a “true” parameter vector θ^* . Otherwise it is *unidentified* (including *partially identified*), if there is a set of parameters and not just a unique value θ^* 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 2×2 covariance matrix of $(\varepsilon_d, \varepsilon_s)$. Then instead of 5 unknown parameters, show there are 7 unknown parameters in θ 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 9×1 vector of unknown “structural parameters” in the supply/demand specification given below

$$\begin{aligned} q_d &= a_d - b_d p + c_d y + \varepsilon_d \\ q_s &= a_s + b_s p + c_s r + \varepsilon_s \end{aligned} \tag{77}$$

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?