

ECON 623
Solutions to In Class Part of the Final Exam (worth 50% of final exam grade)

Do Two out of Three questions below. Question IV is an optional BONUS QUESTION.
 (if you do well on the bonus question you get next higher grade in Econ 623 if you are near a grade borderline, e.g. if you are B+ the extra credit can move your class grade to an A, etc.)

John Rust
 Fall 2008

I. An *autogression* can be written as

$$y_t = a + by_{t-1} + \varepsilon_t \quad (1)$$

where $\{\varepsilon_t\}$ is an *i.i.d* stochastic process that is also independent of $\{y_t\}$ and $b \in (0, 1)$ is known as the “autoregressive coefficient”. Suppose the process is “initialized” at $y_0 = 0$ at $t = 0$, and thus $y_0 = 0$ is the “initial condition” for the autoregression.

1. (10 points) What is the “one step ahead” forecast for this process, i.e. what is $E\{y_{t+1}|y_t\}$?

Answer: $E\{y_{t+1}|y_t\} = E\{a + by_t + \varepsilon_{t+1}|y_t\} = a + by_t$, since $E\{\varepsilon_{t+1}|y_t\} = 0$ given that ε_{t+1} is independent of $(\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_1)$ and y_t is given by

$$\begin{aligned} y_1 &= a + \varepsilon_1 \\ y_2 &= a + by_1 = a + b(a + b\varepsilon_1) + \varepsilon_2 \\ y_3 &= a + by_2 = a + ba + b^2a + b^2\varepsilon_1 + b\varepsilon_2 + \varepsilon_3 \\ \dots &= \dots \\ y_t &= a \sum_{s=0}^{t-1} b^s + \sum_{s=0}^{t-1} b^s \varepsilon_{t-s} + b^t y_0 \end{aligned} \quad (2)$$

so y_t is a function of only $(\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_1)$ and hence ε_{t+1} is independent of y_t .

2. (10 points) What is the “k step ahead” forecast for this process, i.e. what is $E\{y_{t+k}|y_t\}$?

Answer: Using the Law of Iterated Expectations repeatedly, we calculate first with $k = 2$,

$$E\{y_{t+k}|y_t\} = E\{y_{t+2}|y_t\} = E\{E\{y_{t+2}|y_{t+1}\}|y_t\} = E\{a + by_{t+1}|y_t\} = a + bE\{y_{t+1}|y_t\} = a + ba + b^2y_t. \quad (3)$$

Or for arbitrary $k > 0$ we have

$$E\{y_{t+k}|y_t\} = a \sum_{s=0}^{k-1} b^s + b^k y_t. \quad (4)$$

3. (20 points) Write formulas for

$$\lim_{t \rightarrow \infty} E\{y_t|y_0\} \quad (5)$$

and

$$\lim_{t \rightarrow \infty} \text{var}\{y_t|y_0\} \quad (6)$$

Answer The first one is easy given the answer in part 2 above, taking limits we have

$$\lim_{t \rightarrow \infty} E\{y_t|y_0\} = \lim_{t \rightarrow \infty} a \sum_{s=0}^{t-1} b^s = \frac{a}{1-b}, \quad (7)$$

when $b \in (0, 1)$ since in this case we have a convergent geometric series in powers of b . The variance is a bit harder but note that since $y+0=0$, the conditional variance $\text{var}(y_t|y_0)$ is the same as the unconditional variance of y_t , and using the representation of y_t as a moving average of $\{\epsilon_t\}$ in equation (??), and using the fact that $\{\epsilon_t\}$ is an *IID* sequence, we see via a direct calculation that

$$\text{var}(y_t|y_0) = \text{var}(y_t) = \text{var}\left(\sum_{s=0}^{t-1} b^s \epsilon_{t-s}\right) = \sigma^2 \sum_{s=0}^{t-1} b^{2s}. \quad (8)$$

Taking the limit, and again noticing that we have a geometric series but now in multiples of b^2 , we have

$$\lim_{t \rightarrow \infty} \text{var}(y_t|y_0) = \frac{\sigma^2}{1-b^2}. \quad (9)$$

4. (20 points) Suppose that $\{\epsilon_t\}$ is an *IID* Gaussian process (i.e. where each $\epsilon_t \sim N(0, \sigma^2)$). Does y_t converge in probability to any value? If so, which value does it converge to?

Answer Clearly, since the conditional variance of y_t is not converging to 0, y_t cannot be converging in probability to any constant value.

5. (20 points) Does y_t converge in distribution? If so, which distribution does it converge to as $t \rightarrow \infty$?

Answer However it should be clear that y_t does converge in distribution. For any t the calculations above show that $y_t \sim N(\mu_t, \sigma_t^2)$ where

$$\begin{aligned} \mu_t &= a \sum_{s=0}^{t-1} b^s \\ \sigma_t^2 &= \sigma^2 \sum_{s=0}^{t-1} b^{2s} \end{aligned} \quad (10)$$

Since both μ_t and σ_t^2 converge (as deterministic sequences), it follows that y_t converges in distribution to $N(\mu_\infty, \sigma_\infty^2)$ where

$$\begin{aligned} \mu_\infty &= \frac{a}{1-b} \\ \sigma_\infty^2 &= \frac{\sigma^2}{1-b^2} \end{aligned} \quad (11)$$

Note that $\{y_t\}$ is a special type of *Markov process* since the conditional distribution of y_{t+1} given $(y_t, y_{t-1}, \dots, y_0)$ depends only on the most recent value y_t , and not the other values $(y_{t-1}, y_{t-2}, \dots, y_0)$. This is the so-called *Markov property*. Let $f(y_{t+1}|y_t)$ denote the conditional probability density for y_{t+1} given y_t . It is easy to see that when $\{\epsilon_t\}$ is an *IID* Gaussian process that $y_{t+1} \sim N(a + by_t, \sigma^2)$. Thus $f(y_{t+1}|y_t)$ is a normal density with mean $a + by_t$ and variance σ^2 . Let ϕ be the density of a

normal distribution with mean μ_∞ and variance σ_∞^2 given in equation (??) above. Then ϕ can be shown to be an *invariant distribution* for the Markov process $\{y_t\}$, that is

$$\phi(y) = \int_{-\infty}^{+\infty} f(y|x)\phi(x)dx. \quad (12)$$

The interpretation of the invariant distribution is that if a Markov process is *initialized* by choosing y_0 as a draw from this invariant distribution ϕ , then in every subsequent period t , the marginal distribution of y_t will also have this same normal distribution ϕ too. This implies that the Markov process $\{y_t\}$ is *strictly stationary* in this case. On the other hand, if the initial condition is a particular constant, such as $y_0 = 0$ or some other fixed value, or if y_0 is drawn from some other distribution other than ϕ , then the process $\{y_t\}$ is no longer strictly stationary. The initial condition has an effect on the process, but this effect dies out geometrically fast, i.e. at rate b^t which tends to 0 as $t \rightarrow \infty$. One can show that because the effect of the initial condition does die out geometrically, the marginal distribution of y_t converges to ϕ as $t \rightarrow \infty$. In fact this is what we have already shown above, y_t converges in distribution to ϕ as $t \rightarrow \infty$.

6. (10 points) Does the initial condition y_0 matter for any of your conclusions above? What about the parameters (a, b, σ^2) ?

Answer Clearly, from the answer to part 5 above, the initial condition y_0 has no effect on the limiting normal distribution of y_t , and it dies out geometrically fast. However the limiting normal distribution does depend on the three parameters (a, b, σ^2) as we can see in equation (??) above.

7. How are your conclusions above affected in the case where $b = 1$?

Answer When $b = 1$ then $\{y_t\}$ becomes a *random walk* (with *drift* when $a \neq 0$), and we can see from equation (??) that the mean and variance of y_t , μ_t and σ_t^2 are given by

$$\begin{aligned} \mu_t &= at \\ \sigma_t^2 &= \sigma^2 t \end{aligned} \quad (13)$$

so $\lim_{t \rightarrow \infty} \sigma_t^2 = \infty$. Further, from the moving average representation for $\{y_t\}$ in equation (??) above, the initial condition y_0 no longer dies out geometrically as $t \rightarrow \infty$. Instead the initial condition is *persistent and permanent*. Some people think of the stock market as following a random walk and the ε_t are “return shocks”. If the stock market really does evolve according to a random walk, then a large market decline such as we have currently experienced, is *permanent* in the sense that the effect of this negative shock does not die out over time. If the drift is zero, $a = 0$, then $\{y_t\}$ is also a *martingale*, that is, it satisfies the property

$$E\{y_{t+k}|y_t\} = y_t, \quad \text{w.p.1} \quad (14)$$

So if the stock market is a random walk, and if there is no drift, then the martingale property holds, and this says if the Dow has fallen below 9000 then we expect in all future years the Dow will remain below 9000! Depressing isn't it? Let's hope the stock market really isn't a random walk with no drift as so many people claim it to be.

II. Consider the standard linear regression model

$$y_i = X_i\beta^* + \varepsilon_i, \quad i = 1, \dots, N \quad (15)$$

where initially we follow the “old fashioned” textbooks in econometrics (e.g. Peter Schmidt, *Econometrics* 1976) and assume that the X_i are $k \times 1$ *non-stochastic* regressors and the error terms in the regression are *IID* $N(0, \sigma^2)$ random variables.

1. (10 points) What is the distribution of y_i under this assumption?

Answer $y_i \sim N(X_i\beta^*, \sigma^2)$

2. (10 points) Define the “ordinary least squares estimator” OLS $\hat{\beta}$ based on the N observations $\{(y_i, X_i)\}$, $i = 1, \dots, N$ and show that the OLS estimator has the property that it is *linear* in y (where y is the $N \times 1$ vector of “stacked” dependent variables) and an *unbiased* estimator of β^* .

Answer $\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i=1}^N (y_i - X_i\beta)^2$. Using matrix notation, we have

$$\hat{\beta} = (X'X)^{-1}X'y \quad (16)$$

where X is the $N \times k$ matrix of regressors, and y is the $N \times 1$ vector of dependent variables. Thus, $\hat{\beta}$ is clearly a linear function of y , since it is the product of the $k \times N$ matrix $(X'X)^{-1}X'$ and the $N \times 1$ vector y .

3. (20 points) State and prove the *Gauss Markov Theorem*.

Answer The Gauss Markov Theorem states that the OLS estimator is the *best, linear unbiased estimator* of β^* . That is, it has the smallest covariance matrix among all linear and unbiased estimators. Since $y = X\beta^* + \varepsilon$ (where ε is the $N \times 1$ vector of residual terms), it is easy to verify that the OLS estimator is unbiased,

$$E\{\hat{\beta}\} = E\{(X'X)^{-1}X'y\} = E\{(X'X)^{-1}X'(X\beta^* + \varepsilon)\} = \beta^* \quad (17)$$

since $E\{(X'X)^{-1}X'\varepsilon\} = (X'X)^{-1}X'E\{\varepsilon\} = 0$ since the X matrix are non-stochastic (and thus independent of ε). The covariance matrix of $\hat{\beta}$ can be calculated to be $\sigma^2(X'X)^{-1}$. Now consider any other linear unbiased estimator of β^* , i.e an estimator of the form Ly for some $k \times N$ matrix L (which might depend on X). The unbiased property requires $LX = I$, and calculating the covariance matrix of this estimator we find that it is $\sigma^2(LL')$. So we need to show that the matrix $\Omega = [LL' - (X'X)^{-1}]$ is positive semi-definite to complete the proof. To show that Ω is positive semi-definite, note that we can rewrite it as

$$\begin{aligned} \Omega &= L[I - X(X'X)^{-1}X']L' \\ &= LML' \\ &= LMML' \\ &= B'B \end{aligned} \quad (18)$$

where $B = ML'$, and $M = [I - X(X'X)^{-1}X']$ is an idempotent matrix that is often called the “residual matrix” since $My = y - \hat{y} = e$ where e is the $N \times 1$ matrix of regression residuals and $\hat{y} = X\hat{\beta}$ is the regression prediction of the y vector. Clearly any matrix that can be written as BB' is symmetric and positive semi-definite (if this is not immediately obvious why, make sure you know how to prove this last step yourself).

4. (20 points) Does the Gauss Markov Theorem continue to hold if we relax the assumptions that the X_i 's are non-stochastic and the ε_i 's are $N(0, \sigma^2)$? If so, show how your proof above can be modified to allow for stochastic X_i 's and non-normal ε_i 's. What are the weakest general restrictions would you have to place on the distribution of (y_i, X_i) in order for the Gauss Markov Theorem to continue to hold in this case?

Answer The Gauss-Markov Theorem holds under considerably weaker assumptions than those stated in part 1 of this problem, i.e. the “old fashioned textbook assumptions”. In particular the error terms in the regression need not be normally distributed and the regressors can be stochastic rather than deterministic. A more general sufficient condition for the Gauss-Markov Theorem to continue to hold is that (y_i, X_i) are *IID* and that

$$E\{y_i|X_i\} = X_i\beta^* \quad (\text{true regression is linear}) \quad (19)$$

and

$$\text{var}(y_i|X_i) = \sigma^2 \quad (\text{homoscedasticity}) \quad (20)$$

This implies that $E\{\varepsilon_i|X_i\} = 0$, where $\varepsilon_i = y_i - X_i\beta^*$. Under this assumption, you can repeat the steps of the Gauss-Markov Theorem, but use the Law of Iterated expectations in the places where previously you used the much stronger assumption that the X matrix was non-stochastic. So in particular, to show that the OLS estimator is unbiased, note that

$$E\{\hat{\beta}\} = E\{(X'X)^{-1}X'y\} = \beta^* + E\{(X'X)^{-1}X'\varepsilon\} \quad (21)$$

where ε is the $N \times 1$ vector whose i^{th} component is $\varepsilon_i = y_i - X_i\beta^*$. So to prove that OLS is an unbiased estimator of β^* it is sufficient to show that $E\{(X'X)^{-1}X'\varepsilon\} = 0$. But by the Law of Iterated Expectations, we have

$$E\{(X'X)^{-1}X'\varepsilon\} = E\{E\{(X'X)^{-1}X'\varepsilon|X\}\} = 0 \quad (22)$$

where the result that $E\{(X'X)^{-1}X'\varepsilon|X\} = 0$ follows from the assumptions that the data are *IID* and the true regression function is linear, which implies that component by component, we have $E\{\varepsilon_i|X\} = 0$. The homoscedasticity assumption is needed to derive the standard formula for the conditional covariance matrix of the OLS estimator, $\text{var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$. The statement of the Gauss-Markov needs to be modified so that “best linear unbiased estimator” means the one with the smallest *conditional* covariance matrix.

5. (10 points) Define what the R^2 is for the regression and provide sufficient conditions for $0 \leq R^2 \leq 1$. Are there circumstances where R^2 could over outside the unit interval? If so, provide an example where this could happen. Is it the case that adding additional explanatory variables to a regression always increases the R^2 ? If so, provide a proof, otherwise provide a counterexample where the addition of an explanatory variable decreases the R^2 .

Answer The standard definition of R^2 is

$$R^2 = \frac{\text{SSR}}{\text{SST}} \quad (23)$$

where $\text{SST} = \|y - \bar{y}\|^2$ and $\text{SSR} = \|\hat{y} - \bar{y}\|^2$, where y is the $N \times 1$ vector of dependent variables, $\hat{y} = X(X'X)^{-1}X'y$ are the regression predictions of y , and \bar{y} is an $N \times 1$ vector, each of whose

components equals the mean of y . Clearly $R^2 \geq 0$, but R^2 can be greater than 1 if the regression does not contain a constant term. For example, suppose $N = 2$ and there is only one regressor x . and we have

$$y = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad x = \begin{bmatrix} 3 \\ -4 \end{bmatrix} \quad (24)$$

You can verify that in this case we have $\hat{\beta} = -\frac{1}{5}$ and

$$\hat{y} = \begin{bmatrix} -\frac{3}{5} \\ \frac{4}{5} \end{bmatrix} \quad (25)$$

You can calculate that in this case $SST = \frac{1}{2}$ and $SSR = 1.3$, so $R^2 = 2.6$ in this case. If the regression contains a constant term, then $R^2 \leq 1$ since we have

$$SST = SSR + SSE \quad (26)$$

where $SSE = \|e\|^2$, where $e = y - \hat{y}$. If you do not know how to show the above identity, note that $\|e\|^2 = \langle e, e \rangle$, and you expand SST as follows

$$SST = \|y - \bar{y}\|^2 = \langle y - \bar{y}, y - \bar{y} \rangle = \langle y - \hat{y} + \hat{y} - \bar{y}, y - \hat{y} + \hat{y} - \bar{y} \rangle = \langle e + \hat{y} - \bar{y}, e + \hat{y} - \bar{y} \rangle \quad (27)$$

Using the orthogonality of the residual vector e and \hat{y} (which is a linear combination of the columns of the X matrix), it follows that $\langle e, \hat{y} \rangle = 0$. Further, if there is a constant term in the regression, then e is orthogonal to the constant term, and thus orthogonal to \bar{y} , so $\langle e, \bar{y} \rangle = 0$. Using these orthogonality conditions, it follows that $\langle e, \hat{y} - \bar{y} \rangle = \langle y - \hat{y}, \hat{y} - \bar{y} \rangle = 0$, and from this the identity that $SST = SSR + SSE$ follows. Finally, when a constant term is included in the regression, it is necessarily the case that adding additional explanatory variables must increase R^2 . This follows because adding an additional explanatory variable necessarily reduces SSE, and using the identity, we can rewrite R^2 as

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \quad (28)$$

and since adding a regressor to the regression does not affect SST but reduces SSE, it follows that this increases R^2 . However if there is not a constant in the regression, adding a regressor does not necessarily increase R^2 even though it is still true that adding a regressor reduces SSE. This is because the identity in equation (??) no longer holds in the absence of a constant term, so it is possible to construct examples where R^2 is initially larger than 1 and adding another regressor reduces R^2 to below 1. I leave it to you to try some examples (such as extending the example I provided above where $R^2 > 1$) to show this is possible.

- (10 points) The OLS estimator is an example of a *semi-parametric estimator* in the sense that under fairly general conditions it will be a consistent and asymptotically normal estimator for the parametric part of the model (i.e. (β^*, σ^2)) even though we do not specify that the error terms ϵ_i have a parametric distribution such as $N(0, \sigma^2)$. Provide the weakest conditions you can for (y_i, X_i) (but assuming they are *IID* random variables) under which OLS will be consistent and asymptotically normal and derive the asymptotic distribution for the OLS estimator in this case.

Answer While it is hard to say what the absolute *weakest* conditions would be (and proving the absolute weakest conditions would turn this into a unrealistically difficult problem, requiring very arcane

assumptions and verifications, which was not what I was asking of you). I would accept any answer that had conditions that were *sufficiently weaker* than the “old fashioned textbook assumptions”. Thus, sufficiently weak assumptions would be those where 1) β^* is uniquely defined in the limit to the least squares problem

$$\beta^* = \underset{\beta}{\operatorname{argmin}} E\{(y - X\beta)^2\} \quad (29)$$

A sufficient condition for this is that the $k \times k$ matrix $E\{X'X\}$ exists and is invertible. Then the only other condition we need is the weakest condition for a law of large numbers to hold. This requires something just a bit stronger than the existence of second moments for the random vector (y, X) . I am not asking for specifics on these conditions since it is rather technical, but assuming that a Law of Large numbers holds and results in the almost sure (with probability 1) convergence or convergence in probability of the following quantities:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N X_i' X_i &\rightarrow E\{X'X\} \\ \frac{1}{N} \sum_{i=1}^N X_i' y_i &\rightarrow E\{X'y\} \end{aligned} \quad (30)$$

(again the conditions for convergence in probability are slightly weaker than convergence with probability 1, but worrying about these are arcane details is not something I expected you to do in this question). The Slutsky theorem implies that

$$\hat{\beta} = \left[\frac{1}{N} \sum_{i=1}^N X_i' X_i \right]^{-1} \left[\frac{1}{N} \sum_{i=1}^N X_i' y_i \right] \rightarrow [E\{X'X\}]^{-1} [E\{X'y\}] = \beta^* \quad (31)$$

Asymptotic normality can be proven under somewhat stronger conditions on the moments necessary to for a Central Limit Theorem to be applied to

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N X_i' \varepsilon_i \implies N(0, \Omega) \quad (32)$$

where $\varepsilon_i = y_i - X_i \beta^*$ and Ω is the covariance matrix

$$\Omega = E\{\varepsilon^2 X'X\} \quad (33)$$

so sufficient conditions on the random vector (y, X) need to be imposed to ensure the existence of Ω and that a central limit theorem holds in (??). I have not emphasized technical details in my half of the course, just for an understanding of the big picture, and so I did not take off any points if you failed to put in the right technical conditions for a CLT to hold, or got those details wrong. Instead I was looking for evidence that you understand the big picture and that the OLS estimator will be consistent and asymptotically normal for β^* even if the true conditional expectation is not equal to $X\beta^*$, and even if the error terms are not normal, and so forth. All we need is that the *population regression coefficients* β^* defined from the solution to the limiting or population least squares problem (??) exists and are uniquely defined, and then sufficient conditions for the “analogy principle” to hold, so that the sample analog of the limiting population regression coefficients, the OLS estimator $\hat{\beta}$ will converge to the limiting population value β^* as $N \rightarrow \infty$, both in probability,

and that a suitably normalized OLS coefficient vector, $\sqrt{N} [\hat{\beta} - \beta^*]$ will converge in distribution to $N(0, \Omega)$ under very weak conditions, even when the regression is “misspecified” (i.e. when the true conditional expectation is not linear, i.e. $E\{y|X\} \neq X\beta^*$).

7. (20 points) Define the concept of *asymptotic efficiency* and provide sufficient conditions for which the OLS estimator is asymptotically efficient. If the residuals are independent of X_i and have a double exponential distribution, i.e. where ε_i in equation (??) above has a density $f(\varepsilon|\sigma)$ given by

$$f(\varepsilon|\sigma) = \frac{1}{2\sigma} \exp\{-|\varepsilon|/\sigma\} \quad (34)$$

will OLS be asymptotically efficient estimator of β^* in this case? If not, describe an asymptotically efficient estimator in this case. Finally suppose ε_i is independent of X_i and has a *mixed normal distribution* i.e. the CDF of ε_i , $F(\varepsilon)$, is given by

$$F(\varepsilon) = \alpha\Phi(\varepsilon/\sigma_1) + (1 - \alpha)\Phi(\varepsilon/\sigma_2), \quad (35)$$

where Φ is the standard Normal CDF and $\alpha \in (0, 1)$. Is OLS asymptotically efficient estimator of β^* in this case? If not, describe an asymptotically efficient estimator, write a formula for the asymptotic covariance matrix of this efficient estimator and compare it to the covariance matrix of the OLS estimator and show the difference between the latter and former is a positive semi-definite matrix.

Answer The asymptotic efficiency of an estimator is (within the class of all consistent asymptotically normal estimators of a known parameter vector θ^*), the estimator that has the smallest asymptotic covariance matrix. Under fairly weak regularity conditions, but under the strong and important assumption that the statistical model is *correctly specified*, the maximum likelihood estimator can be shown to be asymptotically efficient, and its asymptotic covariance matrix, equal to the inverse of the *information matrix* and also known as the *Cramer-Rao lower bound* implies that maximum likelihood estimation achieves asymptotic efficiency with an asymptotic covariance matrix equal to the Cramer-Rao lower bound. There is an extension of the notion of efficiency due to LeCam and others for any “locally asymptotically normal” families of statistical models that can handle estimators that have non-normal asymptotic distributions, and this more general theory still arrives at the same conclusion that the maximum likelihood estimator is asymptotically efficient, and any other estimator that is consistent and has a limiting asymptotic distribution, even if not normal, will be one that can be represented as the “limiting distribution of the maximum likelihood estimator plus independent, additive “noise”, and hence would be dominated by the maximum likelihood estimator by any decision maker who has a convex loss function for deviations from the estimated parameter and the true values. A sufficient condition for OLS to be asymptotically normal is that the conditional distribution of y given X is normal with mean $X\beta^*$ and variance σ^2 . Then as we showed in class, the OLS estimator is the same as the maximum likelihood estimator and hence is asymptotically efficient. For the case where the residuals have a double exponential distribution as in equation (??), it is straightforward to see that the maximum likelihood estimator is equivalent to the *least absolute deviations* (LAD) estimator, and so OLS is not asymptotically efficient in this case — the LAD estimator is. In the case of the mixed normal distribution we can define the maximum likelihood estimator for $\theta = (\beta, \alpha, \sigma_1, \sigma_2)$ as follows

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \log(f(y_i - X_i\beta|\alpha, \sigma_1, \sigma_2)) \quad (36)$$

where $f(\varepsilon|\alpha, \sigma_1, \sigma_2)$ is the density corresponding to the mixed normal CDF in equation (??) above. Specifically, we have

$$f(\varepsilon|\alpha, \sigma_1, \sigma_2) = \alpha\phi(\varepsilon/\sigma_1)/\sigma_1 + (1 - \alpha)\phi(\varepsilon/\sigma_2)/\sigma_2 \quad (37)$$

where $\phi(\varepsilon)$ is the standard normal density,

$$\phi(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp\{-\varepsilon^2/2\} \quad (38)$$

It is not hard to verify that the maximum likelihood estimator is different than OLS, and since maximum likelihood is asymptotically efficient, the OLS is asymptotically inefficient in the case of a mixed normal error distribution. Since the error term is independent of the X vector, the asymptotic covariance matrix of the OLS estimator is

$$\Omega = \sigma^2 E\{X'X\} \quad (39)$$

where σ^2 is the variance of the mixed normal error term ε which is

$$\sigma^2 = \alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2. \quad (40)$$

For the maximum likelihood estimator, the asymptotic covariance matrix is the submatrix of the inverse of the information matrix that corresponds to the β vector. Recall that the information matrix is defined as

$$I = E \left\{ \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \theta} \right] \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \theta} \right]' \right\} \quad (41)$$

It is tedious but you should be able to show (using properties of odd and even functions, and specifically that with respect to the mixed normal density the expectation of any odd function is zero) that the information matrix is *block diagonal* for the “ β block” (corresponding to the β parameters) with respect to the other three θ parameters, $(\alpha, \sigma_1, \sigma_2)$. Specifically, you can show that

$$\begin{aligned} I_{\beta, \alpha} &= E \left\{ \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \beta} \right] \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \alpha} \right]' \right\} = \mathbf{0} \\ I_{\beta, \sigma_1} &= E \left\{ \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \beta} \right] \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \sigma_1} \right]' \right\} = \mathbf{0} \\ I_{\beta, \sigma_2} &= E \left\{ \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \beta} \right] \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \sigma_2} \right]' \right\} = \mathbf{0} \end{aligned} \quad (42)$$

This implies that the asymptotic covariance matrix for the β coefficients in the maximum likelihood estimator is

$$\Sigma_{\beta, \beta} = [I_{\beta, \beta}]^{-1} = \left[E \left\{ \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \beta} \right] \left[\frac{\partial \log(f(y - X\beta|\alpha, \sigma_1, \sigma_2))}{\partial \beta} \right]' \right\} \right]^{-1} \quad (43)$$

This covariance matrix will be smaller (in a positive definite sense) than the OLS asymptotic covariance matrix for β , Ω in equation (??). Given the independence of the error term and the regressors X one can show that $\Sigma_{\beta, \beta}$ takes the form

$$\Sigma_{\beta, \beta} = \lambda E\{X'X\} \quad (44)$$

where

$$\lambda^{-1} = E \left\{ \left[\frac{f'(\varepsilon|\alpha, \sigma_1, \sigma_2)}{f(\varepsilon|\alpha, \sigma_1, \sigma_2)} \right]^2 \right\}. \quad (45)$$

With some effort, you should be able to show that $\lambda < \sigma^2$, and hence the maximum likelihood estimator of the regression coefficients with mixed normal error terms is more efficient asymptotically than the OLS estimator. I did not expect students to get into a detailed proof of this on the exam: just noting the simple fact that the maximum likelihood estimator is different from OLS and results in a smaller asymptotic covariance matrix was sufficient to get full credit for this part.

III. Consider the standard linear regression model

$$y_i = X_i \beta^* + \varepsilon_i, \quad i = 1, \dots, N \quad (46)$$

where (y_i, X_i) are assumed to be *i.i.d.* random vectors, with y_i a 1×1 random variable and X_i a $K \times 1$ random vector for $i = 1, \dots, N$ but where ε_i (a 1×1 random variable) is *not* independent of X_i .

1. Suppose that $E\{\varepsilon_i|X_i\} \neq 0$, but that $E\{\varepsilon_i X_i\} = 0$ where 0 is a $k \times 1$ vector of 0's. Will the OLS estimator still be a consistent estimator for β^* in this case?
2. Suppose that $E\{\varepsilon_i X_i\} \neq 0$ but we can observe a vector Z_i with dimension $J \times 1$ for which $E\{\varepsilon_i Z_i\} = 0$. Is it possible to construct a consistent estimator of β^* in this case? If so, describe the assumptions you would need to construct a consistent estimator for β^* and describe the most efficient estimator of ε_i that you can think of for β^* under these conditions, but where you do not necessarily know the distribution for ε_i .
3. Suppose we know that ε_i is normally distributed and satisfies $E\{\varepsilon_i|X_i\} = 0$ but $\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i)$ where $\sigma^2(X_i)$ is a *known* function of X_i . Show that the OLS estimator is a consistent estimator for β^* in this case and derive the asymptotic covariance matrix for the OLS estimator. Then, derive a more efficient estimator for β^* and derive its covariance matrix. Show that the difference between the asymptotic covariance matrix for OLS and this other estimator that you derive is positive semi-definite, thereby confirming that the OLS estimator is consistent but inefficient in this case.
4. Continue to assume that ε_i is normally distributed, but now weaken the previous assumption that its conditional variance $\text{var}(\varepsilon_i|X_i)$ is a known function of X_i . Now assume that $\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i|\theta)$ where θ is a vector of unknown parameters that could potentially be estimated. Describe the most efficient estimator for β^* in this case and contrast how this estimator differs from the most efficient estimator for β^* that you could think of in part 3 above. Will it always be the case that the most efficient estimator for β^* when we do not know the form of conditional heteroscedasticity exactly (i.e. in the case where $\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i|\theta)$ versus the case where $\text{var}(\varepsilon_i|X_i) = \sigma^2(X_i)$ is a completely known function of X_i) will have lower asymptotic efficiency due to our lack of knowledge of the precise form of conditional heteroscedasticity? If so, sketch how you would prove this, if not, sketch a counterexample where not knowing θ actually improves the efficiency of the estimator of β^* .
5. Suppose that we *know nothing* about the form of conditional heteroscedasticity, i.e. that $\sigma^2(X_i)$ is a completely unknown function to us, except the general restriction that $\sigma^2(X)$ is a smooth function of X that is uniformly bounded for all $X \in R^K$ and ε_i has uniformly bounded 4th moments (so we don't have to worry about problems of unbounded moments and laws of large numbers and central limit theorems can apply). Does this lack of knowledge about the form of heteroscedasticity prevent us from being able to derive the asymptotic covariance matrix for the OLS estimator? Does it prevent us from being able to find a more efficient estimator than OLS?

6. Now suppose that ε_i is independent of X_i but has a density $f(\varepsilon|\theta)$ that depends on a vector of unknown parameters and for any θ ε_i has mean zero and has finite 4th moments (and thus finite variance too). Will OLS be consistent and asymptotically normal under this situation? When will it be asymptotically efficient (i.e. what are sufficient conditions on $f(\varepsilon|\theta)$ for OLS to be asymptotically efficient)? When OLS is not asymptotically efficient, describe an estimator (i.e. write down an equation for this alternative estimator, even if the estimator for β^* is only implicitly defined) that would be more efficient than OLS.
7. Now suppose that ε_i is independent of X_i but the density of ε_i is unknown to us, except that we know that ε_i has mean zero and has finite 4th moments (and thus finite variance too). Will OLS still be asymptotically normal in this case? Is it possible to find a more efficient estimator than OLS in this case? (Hint: Consider the following “two step estimator”. In step 1, we do OLS and using the OLS estimator, we construct the N residuals, $\hat{\varepsilon}_i = y_i - X_i\hat{\beta}$, $i = 1, \dots, N$. Then using these estimated residuals we construct a *non-parametric kernel density estimator of the unknown density function for ε_i , $f(\varepsilon)$* . Then in step 2, we use this nonparametrically estimated density function, $\hat{f}(\varepsilon)$ as a basis for a *second step maximum likelihood estimator*. This two step process is known as *adaptive estimation* and statisticians have proven that the two step adaptive estimator is as efficient as maximum likelihood assuming that we *knew* what the density of ε_i was in the first place, that is, it is as efficient as what we would get if we knew the density $f(\varepsilon)$ from the outset and thus did not have to estimate it.) Derive a formula for the asymptotic covariance matrix for this two-step adaptive estimator for β^* and show that if $\varepsilon_i \sim N(0, \sigma^2)$ that the adaptive estimator is no more efficient than OLS but if ε_i has a double exponential distribution (see equation (??) above), then the adaptive estimator is strictly more asymptotically efficient than OLS. **BONUS (but not required):** write a formula for the non-parametric kernel density estimator of $f(\varepsilon)$ using the estimated residuals from the first stage regression, $\hat{\varepsilon}_i$, $i = 1, \dots, N$ as the “data” for this non-parametric estimator).

IV. BONUS EXTRA CREDIT QUESTION (not required) Suppose we are trying to estimate an unknown regression function

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, N \quad (47)$$

where the observations (y_i, x_i) are *IID*, where $E\{\varepsilon_i|x_i\} = 0$, and x_i is a single scalar random variable that is uniformly distributed on the $[0, 1]$ interval and our interest is to try to estimate what $f(.5)$ is. Consider the following very simple “naive estimator” of $f(.5)$: we take a small interval of width $2h$ around the point $x = .5$, so we consider the interval $[.5 - h, .5 + h]$ where h is some small number such as $h = .001$. Then we form our estimator of $f(.5)$ by simply taking the average of all of the y_i ’s for which the corresponding x_i ’s lie in this interval, i.e. we only include the y_i ’s in our average provided that $x_i \in [.5 - h, .5 + h]$, otherwise we discard these observations. Thus, we are simply “selecting a subsample near $x = .5$ ” and then just computing a simple sample mean of the y_i ’s for this subsample.

1. Write an explicit expression for this estimator using indicator functions, i.e. $I\{.5 - h \leq x_i \leq .5 + h\}$.
2. Initially assume that h is *fixed* (as $N \rightarrow \infty$). For this fixed h , write an equation for the probability limit of this estimator as $N \rightarrow \infty$ with h remaining fixed.
3. Assume that $f(x)$ is a smooth function (i.e. is continuously differentiable). Use the Mean Value Theorem from calculus to derive an error bound between $f(.5)$ and the probability limit that you have calculated in part 2 above.

4. Show that the error bound you have calculated above depends on h and tends to zero as $h \downarrow 0$.
5. Now, assume that we allow h to tend to 0 as the sample size increases, i.e. we choose a rule $h(N)$ satisfying $\lim_{N \rightarrow \infty} h(N) = 0$. Discuss the difficulties you might encounter if you try to repeat what you did in step 2 (when h was fixed, independent of N) but for the case where $h(N) \rightarrow 0$ as $N \rightarrow \infty$. Will this probability limit always exist and be equal to $f(.5)$ no matter how fast or slow $h(N)$ tends to zero as $N \rightarrow \infty$? Try to guess the rate as which $h(N)$ should go to zero so that this estimator will be a consistent estimator for $f(.5)$.
6. Show that this estimator is actually a special case of a *non-parametric kernel estimator* of $f(x)$ for a particular choice of *kernel* and that h is the *bandwidth parameter* for this kernel density estimator. The naive estimator we have constructed is sometimes called the *histogram estimator* since we are using a histogram centered at $x = .5$ to screen out the x_i 's that do not lie in this "bin" about $x = .5$ when we form our average of the y_i 's.