Econ 551b Econometrics II
Suggested Solution to Problem Set 4

Prof. John Rust, Hiu Man Chan

Spring, 1999

QUESTION 1 First order condition satisfied by the MLE is

$$\nabla_\theta l_N(\widehat{\theta}_N) = \frac{1}{N}\sum_{i=1}^N \nabla_\theta \log f(x_i|\widehat{\theta}_N) = 0$$

Taylor series expansion around $\theta_N = \theta^* + \frac{\delta}{N}$ yields

$$0 = \frac{1}{N}\sum_{i=1}^N \nabla_\theta \log f(x_i|\theta_N) + \frac{1}{N}\sum_{i=1}^N \nabla_{\theta\theta} \log f(x_i|\overline{\theta})(\widehat{\theta}_N - \theta_N)$$

where $\overline{\theta}$ is between θ_N and $\widehat{\theta}_N$. Rearranging yields

$$\sqrt{N}(\widehat{\theta}_N - \theta_N) = \left[\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\theta} \log f(x_i|\overline{\theta})\right]^{-1} \frac{1}{\sqrt{N}}\sum_{i=1}^N \nabla_\theta \log f(x_i|\theta_N).$$

For the first term, $\theta^* + \frac{\delta}{N} \to \theta^*$ implies $\widehat{\theta}_N \xrightarrow{p} \theta^*$ and $\overline{\theta} \xrightarrow{p} \theta^*$, and by LLN and continuous mapping theorem, we have

$$\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\theta} \log f(x_i|\overline{\theta}) \xrightarrow{p} E\left[\nabla_{\theta\theta} \log f(x_i|\theta^*)\right] = \mathcal{I}(\theta^*).$$

For the second term, by CLT for *triangular arrays.* (See *Note* below).

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \nabla_\theta \log f(x_i|\theta_N) \xrightarrow{d} N\left[0, \mathcal{I}(\theta^*)\right].$$

Finally by Slutsky theorem,

$$\sqrt{N}\left(\widehat{\theta}_N - \theta_N\right) \xrightarrow{d} N\left[0, \mathcal{I}^{-1}(\theta^*)\right].$$

Therefore MLE is regular.

To establish this result, we need (i) $\theta^* \in \text{int}(\Theta)$ so that $\theta_N \in \Theta$, (ii) $\{x_i\}$ are independent, and (iii) Lindeberg condition
For all $\varepsilon > 0$,

$$\lim_{N\to\infty} S_N^{-2} \sum_{i=1}^N \int_{|y_{iN}|>\varepsilon S_N} y_{iN}^2 f(x_i|\theta_N) dx_i = 0$$

where

$$\begin{aligned} y_{iN} &= \nabla_\theta \log f(x_i|\theta_N) \\ S_N^2 &= \sum_{i=1}^N Var(y_{iN}) \end{aligned}$$

in order to apply CLT for triangular array $\nabla_\theta \log f(x_i|\theta_N)$.

Note: The CLT needed above for the second term is not the usual sort of CLT, but a special CLT for *triangular arrays.*

Definition: A *triangular array* is a doubly indexed sequence of random vectors $\{y_{i,N}\}$ where the index i runs from $1, \ldots, N$ for each N and the index N runs from $1, \ldots, \infty$. Furthermore for each N, the sequence $\{y_{1,N}, \ldots, y_{N,N}\}$ is *IID.*

Note that the random vectors $\{\partial f(x_i|\theta_N)/\partial\theta\}$ are a triangular array, since for each N the random variables $\{x_i\}$, $i = 1, \ldots, N$ is an *IID* sequence from the density $f(\cdot|\theta_N)$. It follows that $\{\partial f(x_i|\theta_N)/\partial\theta\}$, $i = 1, \ldots, N$ is also an *IID* sequence, and thus the sequence $\{y_{i,N}\} = \{\partial f(x_i|\theta_N)/\partial\theta\}$ is a triangular array of random vectors.

We now sketch the proof of a central limit theorem for triangular arrays of random variables. This proof can then be extended to a proof of the CLT for triangular arrays of random vectors via the *Cramer-Wold device,* i.e. if $\{y_{i,N}\}$ is a triangular array of random vectors in R^K and we can show that for each non-random vector $a \in R^K$ that $\{a' y_{i,N}\}$ is a triangular array of random variables that satisfy the CLT with asymptotic distribution $N(0, a'\Sigma a)$ where Σ is a $K \times K$ covariance matrix, then we can conclude that the original sequence $\{y_{i,N}\}$ satisfies a (vector) CLT with asymptotic distribution $N(0, \Sigma)$.

CLT for Triangular Arrays of Random Variables Consider a triangular array of random variables $\{y_{i,N}\}$ satisfying $E\{y_{i,N}\} = 0$ and $\text{var}(y_{i,N}) = \sigma_N^2$. If $\lim_{N\to\infty} \sigma_N^2 = \sigma^2 > 0$ then we have:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N y_{i,N} \underset{D}{\Longrightarrow} N(0, \sigma^2).$$

Proof: (sketch) We show the moment generating function for the normalized sum $\frac{1}{\sqrt{N}} \sum_{i=1}^N y_{i,N}$ converges to the moment generating for a $N(0, \sigma^2)$, i.e. $\exp\{t^2\sigma^2/2\}$. We have

$$E\left\{\exp\{\frac{t}{\sqrt{N}} \sum_{i=1}^N y_{i,N}\}\right\} = \prod_{i=1}^N M_{y_{i,N}}(t/\sqrt{N}), \tag{1}$$

where $M_{y_{i,N}}(t)$ is the moment generating function for $y_{i,N}$. We can do a Taylor series approximation of this function about 0:

$$M_{y_{i,N}}(t/\sqrt{N}) = 1 + M'_{y_{i,N}}(0)(t/\sqrt{N} - 0) + \frac{1}{2} M''_{y_{i,N}}(\bar{t}_N)(t/\sqrt{N} - 0)^2, \tag{2}$$

where $\bar{t}_N$ is a point on the line segment joining $t/\sqrt{N}$ and 0. Note that by the basic property of moment generating functions we have $M'_{y_{i,N}}(0) = E\{y_{i,N}\} = 0$, and $M''_{y_{i,N}}(0) = E\{y_{i,N}^2\} = \sigma_N^2$. Furthermore the assumption that $\lim_{N\to\infty} \sigma_N^2 = \sigma^2$ implies that

$$\lim_{N\to\infty} M''_{y_{i,N}}(\bar{t}_N) = \sigma^2.$$

Now substituting equation (2) into equation (1) we obtain

$$E\left\{\exp\{\frac{t}{\sqrt{N}}\sum_{i=1}^{N} y_{i,N}\}\right\} = \left[1+\frac{1}{2}M''_{y_{i,N}}(\bar{t}_N)t^2/N\right]^N.$$

Now we use a result from calculus (which can be proved by taking logs and appealing to *L'Hôpital's rule*) that if $\lim_{N\to\infty} a_N = a$ then $\lim_{N\to\infty}[1+a_N/N]^N = \exp a$. Using this result we obtain:

$$\lim_{N\to\infty} E\left\{\exp\{\frac{t}{\sqrt{N}}\sum_{i=1}^{N} y_{i,N}\}\right\} = \lim_{N\to\infty}\left[1+\frac{1}{2}M''_{y_{i,N}}(\bar{t}_N)t^2/N\right]^N = \exp\{\sigma^2t^2/2\}. \tag{3}$$

Equation (3) shows that the limit of the limit of the moment generating function of the normalized sum $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} y_{i,N}$ equals the moment generating function for a $N(0,\sigma^2)$ random variable. By appealing to *Bochner's theorem* (which actually requires characteristic functions, which is simply the moment generating function evaluated at ti instead of t, where $i=\sqrt{-1}$), it follows that the normalized sum $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} y_{i,N}$ converges in distribution to a $N(0,\sigma^2)$ random variable, completing the proof of the CLT for triangular arrays.

QUESTION 2 Twice differentiating $\ln f(x|\theta)$, we get:

$$\frac{\partial^2 \ln f(x|\theta)}{\partial\theta\partial\theta'} = \frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta\partial\theta'} - \frac{\partial \ln f(x|\theta)}{\partial\theta}\frac{\partial \ln f(x|\theta)}{\partial\theta'}$$

Hence,

$$E[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta\partial\theta'}] = E[\frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta\partial\theta'}] - E[\frac{\partial \ln f(x|\theta)}{\partial\theta}\frac{\partial \ln f(x|\theta)}{\partial\theta'}]$$

And we know

$$\begin{aligned}
E[\frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta\partial\theta'}] &= \int \frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta\partial\theta'}f(x|\theta)dx \\
&= \int \frac{\partial^2 f(x|\theta)}{\partial\theta\partial\theta'}dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta'}\int f(x|\theta)dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta'}1 \\
&= 0
\end{aligned}$$

where the third equality holds by assuming that the regularity conditions for interchanging integration and differentiation operations hold.
Therefore,

$$E[\frac{\partial \ln f(x|\theta)}{\partial\theta}\frac{\partial \ln f(x|\theta)}{\partial\theta'}] = -E[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta\partial\theta'}]$$

QUESTION 3 Let's assume that all the six assumptions given in your class notes hold. To find the asymptotic distribution of the NLLS estimator, we first establish its consistency.
Let

$$Q(\theta) = E(\frac{1}{2}(Y_i - f(X_i|\theta))^2)$$

and

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{N}\frac{1}{2}(Y_i - f(X_i|\theta))^2$$

Then $\hat{\theta} = \arg\min Q_n(\theta) \rightarrow_p \theta^o = \arg\min Q(\theta)$.
Now since

$$\begin{aligned}
Q(\theta) &= \frac{1}{2}E\left((Y_i - g(X_i) + g(X_i) - f(X_i|\theta))^2\right) \\
&= \frac{1}{2}\left[E\left((Y_i - g(X_i))^2\right) + E\left((g(X_i) - f(X_i|\theta))^2\right) + 2E\left((Y_i - g(X_i))(g(X_i) - f(X_i|\theta)\right)\right] \\
&= \frac{1}{2}\left[E\left((Y_i - g(X_i))^2\right) + E\left((g(X_i) - f(X_i|\theta))^2\right)\right]
\end{aligned}$$

and the first term is independent of θ. Therefore

$$\theta^o = \arg\min E((g(X_i) - f(X_i|\theta))^2)$$

Under correct specification, it is clear that $\theta^o = \theta^*$, so $\hat{\theta} \rightarrow_p \theta^*$. Under misspecification, the NLLS estimator converges to a point θ^o that gives the best mean squared approximation in the parametric family of $f(X_i|\theta)$ to $g(X_i)$.
Next, using the fact that $\frac{\partial Q_n(\hat{\theta})}{\partial\theta} \approx 0$ (in more accurate term, $\frac{\partial Q_n(\hat{\theta})}{\partial\theta} = o_p(n^{-1/2})$), perform Taylor Series Expansion about $\theta = \theta^o$:

$$\frac{\partial Q_n(\hat{\theta})}{\partial\theta} = 0 = \frac{\partial Q_n(\theta^o)}{\partial\theta} + \frac{\partial^2 Q_n(\overline{\theta})}{\partial\theta\partial\theta'}(\hat{\theta} - \theta^o)$$

With $\overline{\theta} \rightarrow_p \theta^o$. Rearranging the terms, we get

$$\sqrt{n}(\hat{\theta} - \theta^o) = -[\frac{\partial^2 Q_n(\overline{\theta})}{\partial\theta\partial\theta'}]^{-1}\sqrt{n}\frac{\partial Q_n(\theta^o)}{\partial\theta}$$

Note that the data is iid, and that

$$\begin{aligned}
E\left(\frac{\partial}{\partial\theta}Q_n(\theta^o)\right) &= E\left(\frac{\partial}{\partial\theta}\frac{1}{2}(Y_i - f(X_i|\theta^o))^2\right) \\
&= \frac{\partial}{\partial\theta}E\left(\frac{1}{2}(Y_i - f(X_i|\theta^o))^2\right) \\
&= \frac{\partial}{\partial\theta}Q(\theta^o) = 0 \\
\frac{\partial}{\partial\theta}Q_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i|\theta))\frac{\partial}{\partial\theta}f(X_i|\theta) \\
\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}(\frac{\partial}{\partial\theta}f(X_i|\theta)\frac{\partial}{\partial\theta'}f(X_i|\theta) - (Y_i - f(X_i|\theta))\frac{\partial^2}{\partial\theta\partial\theta'}f(X_i|\theta))
\end{aligned}$$

Therefore we can apply Uniform Law of Large Numbers to obtain

$$\frac{\partial^2 Q_n(\overline{\theta})}{\partial\theta\partial\theta'} \rightarrow_p E\left(\frac{\partial}{\partial\theta}f(X_i|\theta^o)\frac{\partial}{\partial\theta'}f(X_i|\theta^o)\right) - E\left((Y_i - f(X_i|\theta^o))\frac{\partial^2}{\partial\theta\partial\theta'}f(X_i|\theta^o)\right) \equiv A(\theta^o)$$

And by Central Limit Theorem,

$$\sqrt{n}\frac{\partial Q_n(\theta^o)}{\partial\theta} \to_d N\left(0, E\left((Y_i - f(X_i|\theta^o))^2 \frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right)\right) \equiv N(0, B(\theta^o))$$

As a result, by Slutsky Theorem,

$$\sqrt{n}(\hat{\theta} - \theta^o) \to_d N\left(0, A(\theta^o)^{-1} B(\theta^o) A(\theta^o)^{-1}\right)$$

Under correct specification, $A(\theta^o)$ can be simplified since with $E(\epsilon_i|X_i) = 0$,

$$E\left((Y_i - f(X_i|\theta^o))\frac{\partial^2}{\partial\theta\partial\theta'} f(X_i|\theta^o)\right) = E\left(\epsilon_i \frac{\partial^2}{\partial\theta\partial\theta'} f(X_i|\theta^o)\right) = 0$$

Therfore,

$$A(\theta^o) = \frac{\partial^2 Q_n(\overline{\theta})}{\partial\theta\partial\theta'} \to_p E\left(\frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right)$$

Furthermore, as ϵ_i are iid and independent of X_i, we have

$$\begin{aligned}
B(\theta^o) &= E\left((Y_i - f(X_i|\theta^o))^2 \frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right) \\
&= E\left(\epsilon_i^2\right) E\left(\frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right) \\
&= \sigma^2 E\left(\frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right)
\end{aligned}$$

Where σ^2 is the variance of ϵ_i.
With cancellations we finally get:

$$\sqrt{n}(\hat{\theta} - \theta^o) \to_d N\left(0, \sigma^2 E\left(\frac{\partial}{\partial\theta} f(X_i|\theta^o) \frac{\partial}{\partial\theta'} f(X_i|\theta^o)\right)^{-1}\right)$$

QUESTION 4

(a) We are asked to compute maximum likelihood estimates of the parameter vector $(\beta,\ \sigma^2,\ \mu,\ \delta^2)$, in the model given by:

$$y_i = \exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^4\right) + \varepsilon_i \tag{4}$$

where we are instructed to believe that (y_i, x_i) are *IID* draws from $x_i \sim N(\mu, \delta^2)$ and $\varepsilon_i \sim N(0, \sigma^2)$. Actually you were mislead: the errors ϵ_i are heteroscedastic, so conditional on x_i we have $\tilde{\epsilon}_i \sim N(0, \sigma^2(x_i))$ where $\sigma^2(x) = \exp\{-4 - .2x^2\}$. So you will be estimating a misspecified model, and later in Econ 551 we will discuss test statistics which are capable of detecting this misspecification. In the meantime your job is to calculate the MLE's of the parameters, $(\beta,\ \sigma^2,\ \mu,\ \delta^2)$. The first step is to write down the likelihood function $L_N(\theta)$ for the data, $\{(y_i, x_i)\}_{i=1}^N$. In general "brute force" maximization of $L_N(\theta)$ may not be a good idea: it might be better to try a "divide and conquer" strategy. Note that the the joint density of (y, x), $f(y, x|\theta)$, is a product of a conditional likelihood of y given x, $g(y|x, \beta, \sigma^2)$, times the marginal density of x, $h(x|\mu, \delta^2)$. It is easy to see that this factorization or *separability* in the joint likelihood enables us to compute the MLEs for the (β, σ^2) parameters and the (μ, δ^2) parameters independently. It also implies a *block diagonality* property which enable us to show that the asymptotic distributions of these parameters are independent.

$$\begin{aligned}
L_N(\theta) &= \frac{1}{N}\sum_{i=1}^N \ln f\left(y_i, x_i|\left(\beta,\ \sigma^2,\ \mu,\ \delta^2\right)\right) \\
&= \frac{1}{N}\sum_{i=1}^N \ln g\left(y_i|x_i\beta,\ \sigma^2\right) + \frac{1}{N}\sum_{i=1}^N \ln h\left(x|\mu, \delta^2\right) \\
&= \left\{-\frac{1}{2}\ln\sigma^2 - \frac{1}{2N\sigma^2}\sum_{i=1}^n [y_i - \exp(X_i\beta)]^2\right\} \\
&\quad + \left\{-\frac{1}{2}\ln\delta^2 - \frac{1}{2N\delta^2}\sum_{i=1}^n [x_i - \mu]^2\right\} - \frac{1}{2}\ln(2\pi) \\
&= L_N^c(\beta, \sigma^2) + L_N^m(\mu, \delta^2),
\end{aligned}$$

where L_N^c denotes the conditional likelihood of the $\{y_i\}$'s given the $\{x_i\}$'s, L_N^m denotes the marginal likelihood of the $\{x_i\}$'s, $X_i = (1, x_i, x_i^2, x_i^3)$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$. The separability of parameters in the third and forth equation allows us to break the estimation problem into two subproblems, which ordinarily makes the programming considerably easier and computations considerably faster:

(1) calculate the MLE's for $(\beta,\ \sigma^2)$ from the conditional likelihood $L_N^c(\beta, \sigma^2)$:

$$\max_{(\beta,\ \sigma^2)} L_N^c(\beta, \sigma^2) = -\frac{1}{2}\ln\sigma^2 - \frac{1}{N}\sum_{i=1}^N \frac{[y_i - \exp(X_i\beta)]^2}{2\sigma^2}, \tag{5}$$

with FOC's

$$\begin{aligned}
\frac{\partial L_N^c}{\partial\beta}(\beta, \sigma^2) &= \frac{1}{N\sigma^2}\sum_{i=1}^N [y_i - \exp(X_i\beta)]\exp(X_i\beta)\, X_i' = 0 \\
\frac{\partial L_N^c}{\partial\sigma^2}(\beta, \sigma^2) &= \frac{-1}{2\sigma^2} + \frac{1}{2N\sigma^4}\sum_{i=1}^N [y_i - \exp(X_i\beta)]^2 = 0,
\end{aligned}$$

Note that there is further separability in this first subproblem: the FOC for β is the same as the FOC for nonlinear least squares (NLS) estimation of β in equation (2) above, ignoring σ^2 since it doesn't affect the solution for $\hat{\beta}$. Once we have computed the NLS estimate of $\hat{\beta}$, we use the second equation to compute $\hat{\sigma}^2$ as the sample variance of the estimated NLS residuals.

(2) calculate the MLE's for (μ, δ^2) from the marginal likelihood

$$\max_{(\mu,\,\delta^2)} L_N^m(\mu, \delta^2) \;\; -\frac{1}{2}\ln\delta^2 - \frac{1}{2N\delta^2}\sum_{i=1}^{N}[x_i - \mu]^2, \tag{6}$$

with FOC's:

$$\begin{aligned} \frac{\partial}{\partial\mu} L_N^m(\mu, \delta^2) &= \frac{1}{N\delta^2}\sum_{i=1}^{n}[x_i - \mu] = 0 \\ \frac{\partial}{\partial\delta} L_N^m(\mu, \delta^2) &= -\frac{1}{2\delta^2} + \frac{1}{2N\delta^4}\sum_{i=1}^{n}[x_i - \mu]^2 = 0 \end{aligned}$$

respectively. Using attached Gauss code `nlreg.gpr` for computing NLS estimates (the sum of squared errors, derivatives and hessian are coded in the `eval.g` procedure) we are able to numerically solve for a vector $\widehat{\beta}$ that sets the FOC for β given in equation (2) above to zero. There are closed-form expressions for the MLEs for the remaining parameters:

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \exp\left(X_i\hat{\beta}\right)\right]^2, \\ \widehat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \widehat{\delta}^2 &= \frac{1}{N}\sum_{i=1}^{N}[x_i - \widehat{\mu}]^2. \end{aligned}$$

Table 1: Maximum Likelihood Estimates of θ using `data1.asc`

Parameter	β_0	β_1	β_2	β_3	σ^2	μ	δ^2
True Value	-3.000	2.000	-4.000	-0.010	0.015	0.000	1.000
MLE	-3.259	2.228	-2.571	-0.971	0.015	-0.047	0.974
Std. Dev.	0.194	0.809	0.836	0.304	.00055	0.025	0.036

Figures 1 and 2 below plot the true and estimated regression functions for this problem. Figure 1 plots the data points also: we see that both the true and estimated regression functions are generally quite close to each other and both go through the middle of the "data cloud". However we see that the MLE gives a substantially downward biased estimate of β_3^* and this causes the estimated regression function to make big divergences from the true regression function at extreme high and low values of x, say $|x| > 4$. However since there are very few high or low values of x in the sample, the NLS and MLE are not able to "penalize" this divergence: the MLE sets $\hat{\beta}_3 = -.97$ since it helps fit the data around $x = 0$ where most of the data points are. Figure 2 provides a blow up of Figure 1 without the data points to show you how the estimated regression function diverges from the truth near $x = 0$. Overall, despite the misspecification of the heteroscedasticity, the MLE and NLS estimators seem to do a pretty good job of uncovering the true regression function, at least for those x's where we have sufficient data. Note that the data plot indicates heteroscedasticity, since the variance of the data around the regression function is bigger in the middle of the graph (near $x = 0$) than at large positive or negative values of x.

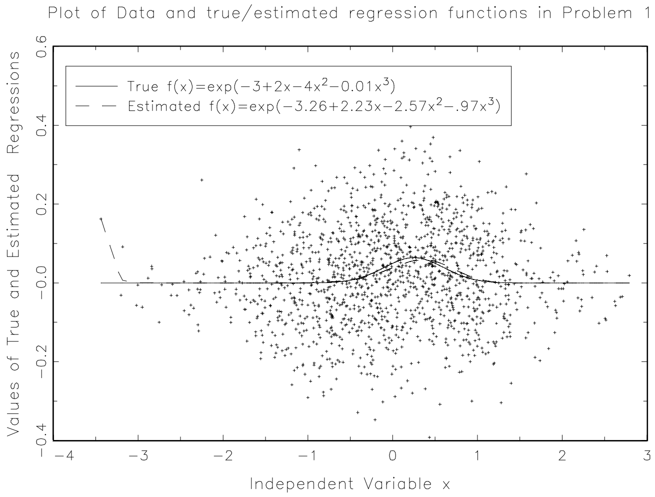

Figure 1: Plot of `data1.asc` and true and estimated regression functions

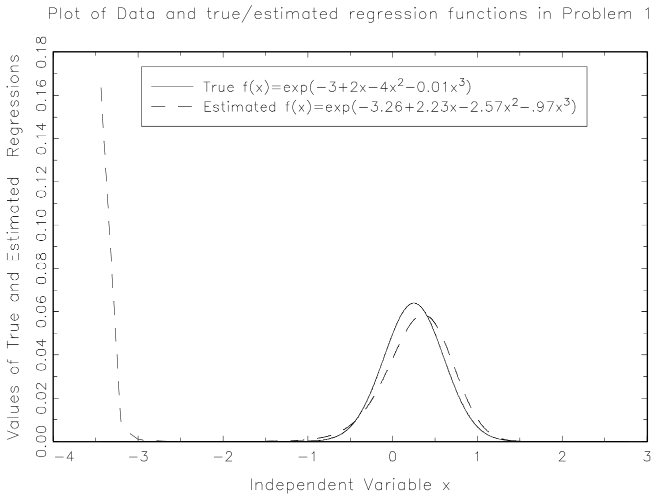

Figure 2: Blow up of true and estimated regression functions

Since the true model used to generate `data1.asc` has heteroscedastic and not homoscedastic error terms $\{\epsilon_i\}$ as assumed here, it is easy to show that the MLE $\hat{\sigma}^2$ of the misspecified model converges to the expectation of $\sigma^2(x) = \exp\{-4 - .2x^2\}$. We can calculate $E\{\sigma^2(\tilde{x})\}$ as follows:

$$
\begin{aligned}
E\{\sigma^2(x)\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\{-4 - .2x^2\} \exp\{-\frac{x^2}{2}\} dx \\
&= \frac{\exp\{-4\}}{\sqrt{1.4}} \frac{\sqrt{1.4}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\{-\frac{\sqrt{1.4}x^2}{2}\} dx \\
&= \frac{\exp\{-4\}}{\sqrt{1.4}} = .015479540.
\end{aligned}
$$

We leave it to you to show that even though the model is misspecified, the MLE $\hat{\sigma}^2$ converges to $E\{\sigma^2(\tilde{x})\}$ as $N \to \infty$. Indeed in this case we find that $\hat{\sigma}^2 = 0.015$, which happens to be almost exactly equal to the "true" value.

Recall that the asymptotic distribution of the standardized MLE estimator is given by:

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \underset{d}{\Longrightarrow} N\left(0, H\left(\theta^*\right)^{-1} \mathcal{I}\left(\theta^*\right) \mathcal{H}\left(\theta^*\right)^{-\infty}\right), \tag{7}$$

where $H\left(\theta^*\right)$ is the Hessian and $\mathcal{I}\left(\theta^*\right)$ is the information matrix (both evaluated at θ^*). However the likelihood is misspecified in this case (due to heteroscedasticity), and it is easy to verify that the equality of $\mathcal{I}\left(\theta^*\right) = -\mathcal{H}\left(\theta^*\right)$ does not hold, so the correct asymptotic covariance matrix is given by the White "misspecification consistent" formula in equation (4) rather than by the inverse of the information matrix $\mathcal{I}\left(\theta^*\right)$. For example, consider the (β, β) block of $\mathcal{I}(\theta^*)$, or $\mathcal{I}_{\beta\beta}(\theta^*)$:

$$\mathcal{I}_{\beta\beta}(\theta^*) = E\left\{\frac{\sigma^2(x)\exp\{X\beta^*\}XX'}{\sigma^4}\right\} \neq -E\left\{-\frac{\exp\{2X\beta^*\}XX'}{\sigma^2}\right\} = -H_{\beta\beta}(\theta^*)$$

We see that a sufficient condition for the two expressions above to equal each other is $\sigma^2(x) = \sigma^2$ for all x, i.e. for the model to be homoscedastic as you were asked to assume. The failure of this equality can be a basis for a specification test statistic that can detect model misspecification which we will discuss in more detail below. Note that even despite the misspecification, the separability property implies that the Hessian is a block diagonal matrix:

$$E\left[\begin{array}{cc} \frac{\partial^2 \ln g\left(y|x;\,(\beta,\,\sigma^2)\right)}{\partial(\beta,\sigma^2)\partial(\beta,\sigma^2)'} & 0 \\ 0 & \frac{\partial^2 \ln h\left(x;\,(\mu,\,\delta^2)\right)}{\partial(\mu,\delta^2)\partial(\mu,\delta^2)'} \end{array}\right]$$

$$= \left[\begin{array}{cccc} \frac{1}{\sigma^2}E\left(Z_i X_i\right) & 0 & 0 & 0 \\ 0 & -\frac{1}{2\sigma^4} & 0 & 0 \\ 0 & 0 & -\frac{1}{\delta^2} & 0 \\ 0 & 0 & 0 & -\frac{1}{2\delta^4} \end{array}\right],$$

where $Z_i \equiv \left[y_i - 2\exp\left(X_i\beta\right)\right]\exp\left(X_i\beta\right)X_i'$. It is also easy to verify that despite the misspecification of the heteroscedasicity, the information matrix $\mathcal{I}(\theta^*)$ is still block diagonal. Block diagonality of $\mathcal{I}(\theta^*)$ and $H(\theta^*)$ implies that the covariance matrix of $\sqrt{N}(\hat{\theta} - \theta^*)$ is block diagonal. One can see further block diagonality between β and σ^2 and between μ and δ^2. This block diagonality is not just a consequence of the separability in both the marginal and conditional likelihood in the parameters describing the mean or conditional mean (μ and β, respectively) and the variance parameters (δ^2 and σ^2, respectively). You should verify through direct calculation that this block diagonality is a result of the symmetry of the normal distribution, which implies that $E\{\epsilon^3\} = 0$, where the conditional distribution of $\epsilon = (y - \exp\{X\beta\})$ given X is $N(0, \sigma^2(x))$, and similarly, the distribution of $\eta = (x - \mu^*) \sim N(0, \delta^2)$, which implies that $E\{\eta^k\} = 0$ for any positive odd integer k. If the normal distribution were not symmetric the block diagonality property wouldn't hold.

Using the block diagonality property, it is easy to compute the standard errors of the full parameter vector, $\hat{\theta}$. The covariance matrix of $\hat{\beta}$ is given by $1/N$ times the upper (β, β) block of $H(\theta^*)^{-1}\mathcal{I}(\theta^*)H(\theta^*)^{-1}$ and it is easy to verify that this is the same as the covariance matrix for the nonlinear least squares estimator for β^* (which is also numerically identical to the MLE) that is output from the `nlreg.gpr` program. By working with equation (4), you can show that the estimated variance of σ^2 is given by $\text{var}(\hat{\sigma}^2) = [\hat{\mu}_4 - \hat{\sigma}^4]/N$ where $\hat{\mu}_4$ is the sample analog of the fourth central moment of $\tilde{X}$:

$$\hat{\mu}_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^4.$$

There is a similar formula for the estimated variance of $\hat{\delta}^2$. We have $\text{var}(\hat{\delta}^2) = 1.948/1500$, so the estimated standard error of $\hat{\delta}^2$ is 0.036. The estimated standard error of $\hat{\sigma}^2$ is $\text{std}(\hat{\sigma}^2) = 0.0005499$.

(b) If the model is correctly specified, we know that the information equality will hold which implies that:

$$H(\theta^*)^{-1}\mathcal{I}(\theta^*)\mathcal{H}(\theta^*)^{-\infty} = \mathcal{I}^{-\infty}(\theta^*). \tag{8}$$

We compare White's misspecification consistent estimate to the inverse of information:

$$\hat{H}\left(\hat{\theta}_{ML}\right)^{-1}\hat{\mathcal{I}}\left(\hat{\theta}_{ML}\right)\hat{H}\left(\hat{\theta}_{ML}\right)^{-1} = \begin{bmatrix} 0.0374 & -0.0872 & 0.0255 & 0.0157 \\ -0.0872 & 0.6543 & -0.5643 & -0.2215 \\ 0.0255 & -0.5643 & 0.6998 & 0.2518 \\ 0.0157 & -0.2215 & 0.2518 & 0.0923 \end{bmatrix}$$

$$\hat{\mathcal{I}}^{-1}\left(\hat{\theta}_{ML}\right) = \begin{bmatrix} 114.9 & -254.7 & 9.0 & 95.8 \\ -254.7 & 2155.5 & -2045.8 & -810.5 \\ 9.0 & -2045.8 & 3765.1 & 6.2 \\ 95.8 & -810.5 & 6.2 & 1721.5 \end{bmatrix},$$

where all the estimates are given from the results of eval.g. Following White ("Maximum Likelihood Estimation of Misspecified Models" *Econometrica* 1982) we can construct a formal hypothesis test statistic using the difference between the estimates of the upper diagonal elements of $H(\theta^*)$ and the corresponding elements of $\mathcal{I}(\theta^*)$. This statistic should be small if the null hypothesis of correct specification is true (since $-H(\theta^*) = \mathcal{I}(\theta^*)$ in that case), and large if the model is misspecified (since it is not necessarily true that $-H(\theta^*) = \mathcal{I}(\theta^*)$ is the model is misspecified). The large difference in the two difference estimates of the covariance matrix for $\hat{\beta}$ suggests that $-H(\theta^*)$ and $\mathcal{I}(\theta^*)$ are different, and hence that the model is misspecified. However we did not actually compute the actual test statistic to see at what level of significance null would actually be rejected (i.e. to compute the marginal signficance level of the test statistic) since we didn't expect you to know about this particular specification test statistic at this stage of the course.

(c) If the true model were *log-linear*, i.e. if

$$y_i = \exp\{X_i\beta\}\exp\{\epsilon_i\}$$

where $\epsilon_i \sim N(0, \sigma^2)$, then it is easy to see that the $\{y_i\}$ are conditionally lognormally distributed so it is valid to take log transformation and estimate (β, σ^2) by OLS:

$$\ln(y_i) = X_i\beta + \epsilon_i$$

. It would not difficult to show that the OLS estimates of the log-linearized model are actually the maximum likelihood estimates of the original lognormal specification (make sure you understand this by writing down the lognormal likelihood function and verifying what we just said is true)! However the error term for the specification of Model I in part (a) is additive and not multiplicative, and so the log transformation is generally not valid. Indeed, there is a positive probability of observing negative realizations of y_i, something that has zero probability under the lognormal specification. Thus, to even do OLS one must screen out all (y_i, x_i) pairs where y_i is negative, something that is generally not a good idea. When one does OLS on this nonrandomly selected subsample, it is not surprising that the estimates are highly biased:

Table 2: Comparison of MLE and OLS estimates of β

Parameter	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$
True values	-3.000	2.000	-4.000	-0.010
Model I (MLE)	-3.259	2.228	-2.570	-0.971
Model II (OLS)	-2.543	0.118	-0.152	-0.033

We can show analytically why the OLS estimates for Model II will be inconsistent when the true model is Model I with additive normal errors rather than multiplicative lognormal errors. After screening out negative y's, the OLS estimator solves:

$$\widehat{\beta}_{OLS} = \underset{\beta \in R^4}{argmin} \frac{1}{N} \sum_{i=1}^{N} \left[\ln(y_i) - X_i\beta\right]^2 I\left\{y_i > 0\right\}. \tag{9}$$

As $N \to \infty$, we can show that the right hand side of the above equation converges uniformly to

$$E\left\{\left[\ln(y) - X\beta\right]^2 I\{y > 0\}\right\}.$$

In general the β that minimizes the expression above is not the same as the β that minimizes

$$E\{[y - \exp(X\beta)]^2\},$$

which is the true β^* when the conditional mean function $\exp\{X\beta\}$ is correctly specified. Therefore, we conclude that the probability limits for $\widehat{\beta}$'s from the two different models(Model I and Model II) are not the same.

(d) Figure 3 below plots the squared residuals $\hat{\epsilon}_i^2 = (y_i - \exp\{X_i\hat{\beta}\})^2$ from the MLE/NLS estimation results in part (a). The figure also plots the results of the following nonlinear regression:

$$\hat{\epsilon}_i^2 = \exp\{X_i\gamma\} + \eta_i, \tag{10}$$

where as before $X_i = (1, x_i, x_i^2, x_i^3)$ and γ is a conformable 4×1 parameter vector. We see substantial evidence of heteroscedasticity, confirming our earlier visual impression in looking at the data in figure 1. The estimated conditional variance function looks concave and symmetric w.r.t. y-axis, almost like a normal density. Table 3 summarizes the estimation results. According to table 3, the regression coefficient for the quadratic term is significant and seems to dominate the form of the heteroscedasticity plotted in figure 3. Figure 4 does a blow up, plotting the estimated and true conditional variance functions.

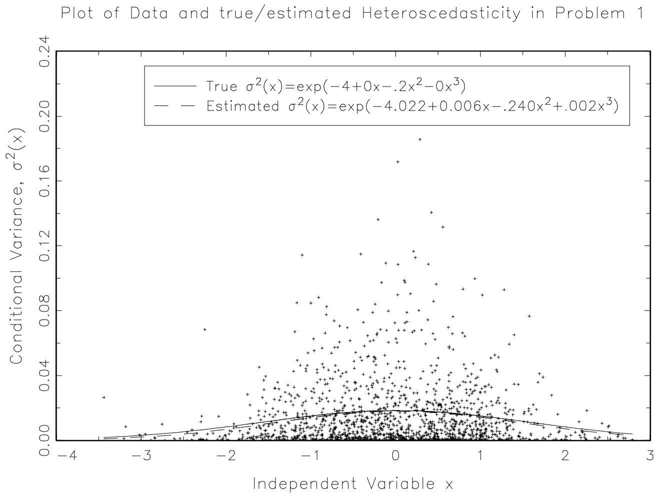

Figure 3: Estimated Squared Residuals and True vs. Estimated $\sigma^2(x)$ Functions

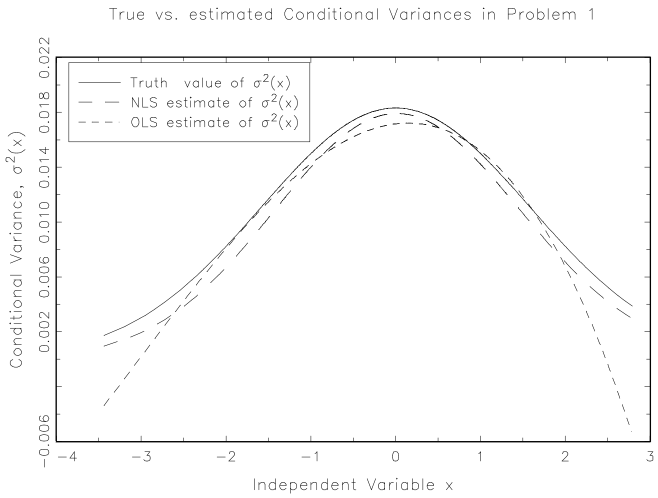

Figure 4: Comparison of True, NLS, and OLS Estimates of $\sigma^2(x)$ Functions

Table 3: NLS Estimates of γ using squared residuals from first stage as data

Parameter	γ_0	γ_1	γ_2	γ_3
True Value	-4.0	0.0	-0.2	0.0
NLS estimate	-4.022	0.006	-0.240	0.002
Standard Deviation	0.050	0.069	0.039	0.025

Some students may have used simple OLS, estimating a specification like

$$\hat{\epsilon}_i^2 = X_i\gamma + \eta_i \tag{11}$$

rather than the exponential specification in equation (7). This is also OK since we weren't specific about what type of tool to use to check for heteroscedasticity. The only disadvantage of the OLS specification in (8) over the exponential specification in (7) is that the latter doesn't guarantee that $\hat{\sigma}^2(x) \geq 0$ for all x. However we find that even in the linear specification most of the predicted $\hat{\sigma}^2(x_i)$ values are indeed positive. Figure 4 compares the predicted values of $\sigma^2(x)$ using both specifications, and we can see that the negative predicted values of $\sigma^2(x)$ occur at the extreme high and low values of the observed x's.

(e) Now we consider full information maximum likelihood (FIML) estimation of Model III, which is the same as model I, but with an exponential specification for conditional heteroscedasticity. Thus the joint density for (y, x) is given by:

$$y_i = \exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3\right) + \varepsilon_i \tag{12}$$

where $\varepsilon_i \sim N\left(0, \exp\left(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2\right)\right)$, and $x_i \sim N\left(\mu, \delta^2\right)$. We want to consider simultaneous or "full information maximum likelihood" (FIML) estimation of the parameter vector $\theta = (\beta, \gamma, \mu, \delta^2)$. The log-likelihood function $L_N(\theta)$ is given by:

$$\begin{aligned}
L_N(\theta) &= \frac{1}{N}\sum_{i=1}^{N} \ln f\left(y_i, x_i;\ (\beta,\ \gamma)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \ln g\left(y|x;\ (\beta,\ \gamma)\right) + \frac{1}{N}\sum_{i=1}^{N} \ln h\left(x|(\mu,\ \delta^2)\right) \\
&= \left\{\frac{-1}{2N}\sum_{i=1}^{N} X_i\gamma - \frac{1}{2N}\sum_{i=1}^{N} \frac{\left[y_i - \exp\left(X_i\beta\right)\right]^2}{\exp\left(X_i\gamma\right)}\right\} \\
&\quad + \left\{-\frac{N}{2}\ln\delta^2 - \frac{1}{2\delta^2}\sum_{i=1}^{N}\left[x_i - \mu\right]^2\right\} - \ln(2\pi) \\
&\equiv L_N^c(\beta, \gamma) + L_N^m(\mu, \delta^2).
\end{aligned}$$

where $X_i = \left(1, x_i, x_i^2, x_i^3\right)$ and $\gamma = \left(\gamma_0, \gamma_1, \gamma_2, \gamma_3\right)'$. The gradients of $L_N(\theta)$ with respect to $(\ \beta,\ \gamma)$ are:

$$\begin{aligned}
\frac{\partial}{\partial\beta} L_N^c(\beta, \gamma) &= -\frac{1}{N}\sum_{i=1}^{N} \frac{\left[y_i - \exp\left(X_i\beta\right)\right]\exp\left(X_i\beta\right)X_i'}{\exp\left(X_i\gamma\right)} \\
\frac{\partial}{\partial\gamma} L_N^c(\beta, \gamma) &= -\frac{1}{2N}\sum_{i=1}^{N} X_i' + \frac{1}{2N}\sum_{i=1}^{N} \frac{\left[y_i - \exp\left(X_i\beta\right)\right]^2}{\exp\left(X_i\gamma\right)} X_i'.
\end{aligned}$$

The gradients for L_N with respect to $(\ \mu,\ \delta^2)$ are the same as in part (a). The hessian matrix for L_N^c with respect to $(\ \beta,\ \gamma)$ is given by:

$$\begin{aligned}
\frac{\partial^2}{\partial\beta\partial\beta'} L_N^c(\beta, \gamma) &= \frac{1}{N}\sum_{i=1}^{N} \left[y_i - 2\exp\left(X_i\beta\right)\right]\exp\left\{X_i(\beta - \gamma)\right\} X_i' X_i \\
\frac{\partial^2}{\partial\beta\gamma'} L_N^c(\beta, \gamma) &= -\frac{1}{N}\sum_{i=1}^{N} \left[y_i - \exp(X_i\beta)\right]\exp\{X_i(\beta - \gamma)\} X_i' X_i \\
\frac{\partial^2}{\partial\gamma\partial\gamma'} L_N^c(\beta, \gamma) &= -\frac{1}{N}\sum_{i=1}^{N} \frac{\left[y_i - \exp\left(X_i\beta\right)\right]^2}{2\exp\left(X_i\gamma\right)} X_i' X_i.
\end{aligned}$$

It is easy to verify (using the law of iterated expectations) that when $(\beta, \gamma) = (\beta^*, \gamma^*)$, the expectation of $\partial^2 L_N^c(\beta, \gamma)/\partial\beta\partial\gamma' = 0$, i.e. we have block diagonality between the β and γ parameters (assuming the model is correctly specified). Similarly one can verify that the (β, γ) block of the information matrix $\mathcal{I}$ is zero. This implies that the asymptotic covariance between the maximum likelihood estimates $\hat{\beta}$ and $\hat{\gamma}$ is zero, so they are asymptotically independently distributed. This independence suggests the following 2-step procedure to obtain initial consistent estimates of (β^*, γ^*): 1) estimate β by NLS (see attached Gauss code `eval_nls.g` and shell program `nlreg.gpr`), 2) use the estimated squared residuals ϵ_i^2 to estimate the γ parameters by NLS using the exponential specification in equation (7). We did this using the same `eval_nls.g` procedure we used for step 1, with a slight modification of `nlreg.gpr` to substitute $\{\hat{\epsilon}_i^2\}$ instead of $\{y_i\}$ as the dependent variable in the regression.

However we can do even better than this. We can do a 3rd step, weighted NLS or feasible generalized least squares (FGLS) estimation of β using the estimated conditional variance $\hat{\sigma}^2(x_i) = \exp\{X_i\hat{\gamma}\}$ from step 2 as weights. The procedure `eval_fgls.g` provides the code to do the FGLS estimation. Due to the block diagonality property, it is not hard to show that the FGLS estimates of β^* have the same asymptotic distribution as the MLE: i.e. FGLS is asymptotically efficient in this case. To see this, note that the gradient and hessian of $L_N^c(\beta, \gamma)$ with respect to β is the same as the gradient and hessian for the following FGLS criterion function:

$$\hat{\beta}_{\text{fgls}} = \underset{\beta \in R^4}{argmax} -\frac{1}{N}\sum_{i=1}^{N} \frac{[y_i - \exp\{X_i\beta\}]^2}{\exp\{X_i\hat{\gamma}\}} \tag{13}$$

We know that the block diagonality property implies that as long as $\hat{\gamma}$ is any consistent estimator of γ^* that a solution $\hat{\beta}$ to the FOC $\partial L_N^c(\beta, \hat{\gamma})/\partial\beta = 0$ is asymptotically efficient. But since this is also the FOC for the FGLS estimator (10), it follows that $\hat{\beta}_{\text{fgls}}$ is also an asymptotically efficient estimator, i.e. it attains not only the Chamberlain efficiency bound for condition moment restrictions, but the Cramer-Rao lower bound as well. It is not hard to show that these two bounds coincide in this case: make sure you understand this by verifying the equality yourself.

It is not apparent that the $\hat{\gamma}$ obtained from the 4th step of our estimation procedure which regresses the squared residuals from the FGLS estimation in step 3 on $\exp\{X_i\gamma\}$ will be asymptotically efficient since the first order condition for γ from maximizing $L_N^c(\hat{\beta}, \gamma)$ with respect to γ does not appear to be the same as the FOC for γ from the nonlinear regression in step 4 of our suggested estimation procedure. So to get fully efficient estimates, we can use $(\hat{\beta}_{\text{fgls}}, \hat{\gamma}_{\text{fgls}})$ as starting values for direct FIML estimation of the full parameter vector (β^*, γ^*) using $L_N^c(\beta, \gamma)$. The procedure `eval_fiml.g` and the shell program `mle.gpr` implement full maximum likelihood estimation of Model III (note we have also provided the procedure `hesschk.g` to allow you to compare numerical and analytically calculated values of the hessian matrix, verifying that the analytic formulas for the hessian matrix given above are correct). This full model is rather delicate and we were unable to get it to converge starting from $(\beta, \gamma) = 0$. However we had no problems with convergence starting from $(\hat{\beta}_{\text{fgls}}, \hat{\gamma}_{\text{fgls}})$. Table 4 below compares the FGLS and MLE estimates.

Table 4: Comparison of MLE and FGLS Estimates of Model III

Parameter	Truth	MLE	Std. Error	FGLS	Std. Error
β_0	-3.000	-3.256	0.190	-3.256	0.190
β_1	2.000	2.209	0.777	2.200	0.789
β_2	-4.000	-2.558	0.819	-2.548	0.833
β_3	-0.010	-0.966	0.295	-0.962	0.300
γ_0	-4.000	-4.012	0.044	-4.022	0.050
γ_1	0.000	0.030	0.053	0.006	0.069
γ_2	-0.200	-0.254	0.026	-0.239	0.039
γ_3	0.000	-0.010	0.014	0.002	0.025

We see that the FIML and FGLS estimates of β are very close to each other and the standard errors are nearly identical, as we would expect from the theoretical result shown above that the FGLS estimator of β is asymptotically efficient. There are more significant differences in the FIML and FGLS estimates of γ. In particular the standard errors of the FGLS estimates are significantly larger than the FIML estimates of γ, which suggests that the FGLS estimates are not asymptotically efficient. Students should be able to verify that this is the case by deriving analytic formulas for the asymptotic covariance matrix for the MLE and FGLS estimators of γ.