Problem Set 1 Solutions

Alberto Ramírez

March 27, 2014

Problem 1

Part B: Log-Likelihood Function

From the homework assignment we know that the probability of observing y_i is given by $f_Y(y_i|x_i, \lambda) = \frac{exp(-\lambda)\lambda^{y_i}}{y_i!}$. We can thusly write the joint distribution of the given model (since we make the i.i.d. assumption) as:

$$f_Y(y_1, y_2, ...y_N|x, \theta) = \prod_{i=1}^{N} \frac{exp(-\lambda)\lambda^{y_i}}{y_i!} \tag{1}$$

Then taking the log of expression (1) we can obtain the log-likelihood function as:

$$L(\theta|y_i) = \ln f_Y(y_1, y_2, ...y_N|x, \theta) = \sum_{i=1}^{N}[-\lambda + y_i \ln \lambda - \ln y_i!] \tag{2}$$

This is the function to be maximized in Matlab, where we multiply the above by $\frac{1}{N}$ to take the average log-likelihood function.

(Optional) Part E: Proposed GMM Estimator

Since the poisson distribution has the first and second moments equal to the parameter λ, which we have parameterized through $\lambda = exp(-X\theta)$, specifying two moment conditions based on the sample moments yields a just identified case (since we have two parameters θ_1 and θ_2). From the mean we can define the first moment condition as:

$$m_\mu(\theta) = exp(-X\theta) - \frac{1}{n}\sum_{i=1}^{n} y_i$$

The second moment condition will then be given by the sample variance (define $\bar{y}$ as the sample mean):

$$m_{var}(\theta) = exp(-X\theta) - \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Then in matrix form we have our 2×1 moment vector as:

$$m_2(\theta) = \begin{bmatrix} exp(-X\theta) - \frac{1}{n}\sum_{i=1}^{n} y_i \\ exp(-X\theta) - \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{bmatrix}$$

So our GMM estimator, $\widehat{\theta}_{GMM}$, is defined as:

$$\widehat{\theta}_{GMM}(\widehat{W}) = \arg\min_{\theta} \; n \cdot m_2(\theta)'\widehat{W}m_2(\theta)$$

Problem 2

Part A: Weak Exogeneity

Suppose we misspecify the ARX(2) model with the ARX(1) model of $E(y_t|y_{t-1}, x_t) = \alpha + \rho y_{t-1} + \beta x_t$ from which we seek to estimate $y_t = \alpha + \rho_1 y_{t-1} + \beta x_t + \nu_t$. Then for consistency of the estimate of ρ, $\hat{\rho}$, we would need weak exogeneity to hold, implying $E(y_{t-1}\nu_t) = 0$. However, in misspecifying the ARX(2) model we know that ν_t in the ARX(1) estimated equation does not follow the same distribution of ε_t, and in fact is no longer white noise; specifically $\nu_t = \rho_2 y_{t-2} + \varepsilon_t$ and in Probelm 3 I will show the distribution that this follows and why it is not white noise. We can thus derive the expression for $E(y_{t-1}\nu_t)$:

$$E(y_{t-1}\nu_t) = E[y_{t-1} \cdot (\rho_2 y_{t-2} + \varepsilon_t)]$$

$$= \rho_2 E(y_{t-1}y_{t-2}) + \cancelto{0}{E(y_{t-1}\varepsilon_t)} = \rho_2 E(y_{t-1}y_{t-2})$$

where in the second line the term $E(y_{t-1}\varepsilon_t) = 0$ since ε_t is WN. This only leaves the term $\rho_2 E(y_{t-1}y_{t-2})$. We must show that the value of this term is not equal to 0. We can go one step further and actually calculate the value of this autocovariate (**optional**), which we expect to be positive as all the ρ_i of the DGP are positive.

Proof that $E(y_{t-1}y_{t-2}) \neq 0$

We will need to apply Time Series analysis to be able to finish the proof of this problem, which is facilitated by the fact that we have classical error.

First, we can rearrange the problem to make this more convenient. To do this we will implement the lag operator, L. For y_{t-1} we can write:

$$y_{t-1} = \alpha + \rho_1 y_{t-2} + \rho_2 y_{t-3} + \beta x_{t-1} + \varepsilon_{t-1} \Leftrightarrow y_{t-1} - \rho_1 y_{t-2} - \rho_2 y_{t-3} = \alpha + \beta x_{t-1} + \varepsilon_{t-1}$$

and we can obtain the similar representation for y_{t-2}. Now applying the lag operator, $\rho(L)$, we get:

$$(1 - \rho_1 L - \rho_2 L^2)y_{t-1} = \alpha + \beta x_{t-1} + \varepsilon_{t-1}$$

$$(1 - \rho_1 L - \rho_2 L^2)y_{t-2} = \alpha + \beta x_{t-2} + \varepsilon_{t-2}$$

Since we have the values of ρ_i and these values are such that $|\rho_i| < 1$ we know the process is covariance-stationary.[1]

Now we can define the characteristic polynomial of the process' lag operator to find the roots (we substitute in the values of ρ_i and switch L to z - we make this final substitution

[1] This simply means that the value of any autocovariance is not a function of the date, only the lag length between dates - other conditions must be satisfied, such as finite variance and a mean that is not time dependent; however, that $|\rho_i| < 1$ is a sufficient condition for stationarity of the process as this implies the two previously stated conditions.

as L is an operator and strictly speaking is not a variable). We require for stationarity that the roots z are outside the unit circle, $|z| > 1$:

$$\theta(z) = 1 - 0.5z - 0.4z^2 = 0 \Leftrightarrow z_1 = -2.325 \text{ and } z_2 = 1.075$$

We also, we know from time series analysis that we can factorize the lag polynomial:

$$(1 - \rho_1 L - \rho_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L)$$

Substituting z for L in the above factorization we can see the characteristic polynomial is equal to the factorization:

$$\theta(z) = 1 - 0.5z - 0.4z^2 = (1 - \lambda_1 z)(1 - \lambda_2 z) = 0$$

and trivially this implies that the values of z that equate the characteristic polynomial to 0 are either:

$$z = \frac{1}{\lambda_1} \text{ or } z = \frac{1}{\lambda_2}$$

which obviously implies that z and λ are inverses. Importantly, these λs are the eigenvalues of the system (and because of the inverse nature we have to have that the λs are inside the unit circle for a stationary process, that is $|\lambda_i| < 1$). Therefore we have that:

$$\lambda_1 = -0.43 \text{ and } \lambda_2 = 0.93$$

Since we have all eigenvalues within the unit circle we know that the inverse of our lag operator exists and is well defined (has a closed form). It can be shown that this inverse is:

$$(1 - \rho_1 L - \rho_2 L^2)^{-1} = \psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \psi_4 L^4 + ... = \psi(L)$$

Further, because we have distinct eigenvalues we will have closed form solutions of the coefficients ψ_j! It can be shown that these coefficients take on the functional forms:

$$\psi_j = c_1 \lambda_1^j + c_2 \lambda_2^j \quad \text{where we have: } c_1 = \frac{\lambda_1}{\lambda_1 - \lambda_2} \quad c_2 = \frac{\lambda_2}{\lambda_2 - \lambda_1} \tag{3}$$

We can now generate the MAX(∞) representation of our ARX(2) process by applying the inverse of the lag operator:[2]

$$y_{t-1} = \psi(L)[\alpha + \beta x_{t-1} + \varepsilon_{t-1}] \tag{4}$$

$$y_{t-2} = \psi(L)[\alpha + \beta x_{t-2} + \varepsilon_{t-2}] \tag{5}$$

where in the above we get that:

$$\psi(L)\alpha = \frac{\alpha}{1 - \rho_1 - \rho_2} = \mu \text{ the unconditional mean of the process}$$

Now taking the unconditional expectation,[3] $E(y_{t-1}y_{t-2})$, of (6) and (7) we get:

$$E(y_{t-1}y_{t-2}) = E[(\mu + \beta\psi(L)x_{t-1} + \psi(L)\varepsilon_{t-1})(\mu + \beta\psi(L)x_{t-2} + \psi(L)\varepsilon_{t-2})] - \mu^2$$

[2] We know that any zero-mean, covariance stationary AR(p) process can be represented as an MA(∞) process. This theorem is known as Wold's decomposition.

[3] If we instead subtract $\beta\psi(L)x_{t-i}$ for $i = 1, 2$ from both sides of (6) and (7), then our autocovariance would be conditional on the exogenous term; hence we would instead be calculating $E(y_{t-1}y_{t-2}|x_{t-1}, x_{t-2})$. This would not detract from the conclusion of our proof, but the value of this expectation would obviously be different from the unconditional one we are deriving.

where μ^2 is the only surviving term from $E(y_{t-1})E(y_{t-2})$ in the RHS of the above covariance equation. Then the unconditional expectation is given by:

$$\begin{aligned}
E(y_{t-1}y_{t-2}) =& \cancel{\mu^2} + \beta^2 E[(\psi_0 x_{t-1} + \psi_1 x_{t-2} + ...)(\psi_0 x_{t-2} + \psi_1 x_{t-3} + ...)] + \\
& E[(\psi_0 \varepsilon_{t-1} + \psi_1 \varepsilon_{t-2} + ...)(\psi_0 \varepsilon_{t-2} + \psi_1 \varepsilon_{t-3} + ...)] - \cancel{\mu^2} \\
=& \beta^2 [E(\psi_0 \psi_1 x_{t-2}^2) + E(\psi_1 \psi_2 x_{t-3}^2) + ... + E(\psi_j \psi_{j+1} x_{t-(j+2)}^2) + ...] + \\
& E(\psi_0 \psi_1 \varepsilon_{t-2}^2) + E(\psi_1 \psi_2 \varepsilon_{t-3}^2) + ... + E(\psi_j \psi_{j+1} \varepsilon_{t-(j+2)}^2) + ... \\
=& (\psi_0\psi_1 + \psi_1\psi_2 + ... + \psi_j\psi_{j+1} + ...)(\beta^2\sigma_x^2 + \sigma_\varepsilon^2) \\
=& (\beta^2\sigma_x^2 + \sigma_\varepsilon^2)\sum_{j=0}^{\infty} \psi_j\psi_{j+1} > 0
\end{aligned}$$

where $E(\varepsilon_{t-j}\varepsilon_{t-k}) = 0$ and$E(x_{t-j}x_{t-k}) = 0 \ \forall j \neq k$ since ε_t is white noise and $x_t \sim i.i.\mathcal{N}(0,1)$. This shows that $E(y_{t-1}y_{t-2}) \neq 0$ since $\psi_j > 0 \ \forall j$ as defined in (5), and obviously $\psi_j \to 0$ as $j \to \infty$. We can see this immediately because $|\lambda_2| > |\lambda_1|$ implies that $\lambda_1 \to 0$ as $j \to \infty$ at a faster rate than λ_2. This means the largest eigenvalue dominates the process so that in our case ψ_j is always positive. Thus the ARX(1) process does not satisfy weak exogeneity. Further, given that we know the eigenvalues of the process we may go one step further and actually calculate the value of this expectation.

Optional: Calculation of $E(y_{t-1}y_{t-2})$ via the MAX(∞) Representation

We will first work with the MAX(∞) representation first since we have already derived the infinite series. Substituting in the values of ψ_j from (5) we obtain:

$$\begin{aligned}
E(y_{t-1}y_{t-2}) =& (\beta^2\sigma_x^2 + \sigma_\varepsilon^2)\sum_{j=0}^{\infty}\psi_j\psi_{j+1} = (\beta^2\sigma_x^2 + \sigma_\varepsilon^2)\sum_{j=0}^{\infty}(c_1\lambda_1^j + c_2\lambda_2^j)(c_1\lambda_1^{j+1} + c_2\lambda_2^{j+1}) \\
=& (\beta^2\sigma_x^2 + \sigma_\varepsilon^2)\sum_{j=0}^{\infty}\left(\frac{\lambda_1^{j+1}}{\lambda_1 - \lambda_2} + \frac{\lambda_2^{j+1}}{\lambda_2 - \lambda_1}\right)\left(\frac{\lambda_1^{j+2}}{\lambda_1 - \lambda_2} + \frac{\lambda_2^{j+2}}{\lambda_2 - \lambda_1}\right) \\
=& (\beta^2\sigma_x^2 + \sigma_\varepsilon^2)\sum_{j=0}^{\infty}\left[\frac{\lambda_1^{2j+3}}{(\lambda_1 - \lambda_2)^2} + \frac{\lambda_2^{2j+3}}{(\lambda_2 - \lambda_1)^2} - \frac{\lambda_1^{j+1}\lambda_2^{j+2} + \lambda_2^{j+1}\lambda_1^{j+2}}{(\lambda_1 - \lambda_2)^2}\right] \\
=& \frac{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}{(\lambda_1 - \lambda_2)^2}\sum_{j=0}^{\infty}\left[\lambda_1^{2j+3} + \lambda_2^{2j+3} - \lambda_1^{j+1}\lambda_2^{j+2} - \lambda_2^{j+1}\lambda_1^{j+2}\right] \\
=& \frac{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}{(\lambda_1 - \lambda_2)^2}\sum_{j=0}^{\infty}\left[\lambda_1^{2j+3} + \lambda_2^{2j+3} - (\lambda_1\lambda_2)^j(\lambda_1\lambda_2^2 + \lambda_2\lambda_1^2)\right] \\
=& \frac{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}{(\lambda_1 - \lambda_2)^2}\left[\lambda_1^3\sum_{j=0}^{\infty}\lambda_1^{2j} + \lambda_2^3\sum_{j=0}^{\infty}\lambda_2^{2j} - (\lambda_1\lambda_2^2 + \lambda_2\lambda_1^2)\sum_{j=0}^{\infty}(\lambda_1\lambda_2)^j\right] \\
=& \frac{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}{(\lambda_1 - \lambda_2)^2}\left[\frac{\lambda_1^3}{1 - \lambda_1^2} + \frac{\lambda_2^3}{1 - \lambda_2^2} - \frac{\lambda_1\lambda_2^2 + \lambda_2\lambda_1^2}{1 - \lambda_1\lambda_2}\right]
\end{aligned}$$

Since we know the actual values of the eigenvalues (and that $\sigma_\varepsilon^2 = 1$, $\sigma_x^2 = 1$, and $\beta = 2$) we can substitute these in and obtain the value of the unconditional autocovariance:

$$E(y_{t-1}y_{t-2}) = 16.217 \neq 0$$

As we expected the above autocovariate is positive.

Optional: Calculation of $E(y_{t-1}y_{t-2})$ via the ARX(2) Representation

Yet simpler, one can show the the autocovariances of an ARX(2) process follows the second-order difference equation shown below (the proof is simple and can be obtained from Hamilton):

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} \text{ for j=1,2,3,....}$$

with the variance of an ARX(2) process given by:

$$\gamma_0 = \frac{(1-\phi_2)\sigma^2}{(1+\phi_2)\left[(1-\phi_2)^2 - \phi_1^2\right]}$$

where in our case $\sigma^2 = \beta^2\sigma_x^2 + \sigma_\varepsilon^2$. So the variance of our process is $\gamma_0 = 19.4805$. Then for $E(y_{t-1}y_{t-2})$ we have that the distance in time between observations is given by $|t-1-(t-2)| = 1 = j$, our lag length. We can then write:

$$E(y_{t-1}y_{t-2}) = \gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_{-1}$$

Since the process is stationary we immediately have that $\gamma_{-1} = \gamma_1$. Substituting this in the above equation we obtain the functional form of our autocovariance:

$$\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1 \Leftrightarrow \gamma_1 = \frac{\phi_1}{1-\phi_2}\gamma_0 \Leftrightarrow \gamma_1 = \frac{\phi_1(\beta^2\sigma_x^2 + \sigma_\varepsilon^2)}{(1+\phi_2)\left[(1-\phi_2)^2 - \phi_1^2\right]}$$

Finally, substituting in $\phi_1 = \rho_1$, $\phi_2 = \rho_2$, along with $\beta^2\sigma_x^2 + \sigma_\varepsilon^2 = 5$, in the above equation (and the values of these parameters) we get the value of the autocovariate:

$$E(y_{t-1}y_{t-2}) = \gamma_1 = 16.2338 \neq 0$$

Which corresponds to the value obtained from the MAX(∞) representation. If we divide this autocovariate by the variance we obtain the autocorrelation of the first lag, $\varrho_1 = \frac{\gamma_1}{\gamma_0} = 0.8324$.

Part B: Consistency of IV Estimation of ARX(1) Model with Lags of $\mathbf{x_t}$

For IV estimation define the following concatenated $N \times 3$ matrices of instruments $Z = [1, x_t, x_{t-1}]$ and data $X = [1, y_{t-1}, x_t]$, as well as the 3×1 column vector of parameters as $\delta = (\alpha, \rho, \beta)'$.[4] Suppose that our Z and X matrices are such that $E(Z'X)$ is of full column rank and just identified. Then we may apply the method of moments and use the IV estimator, $\hat{\delta}_{IV}$:[5]

$$\hat{\delta}_{IV} = (Z'X)^{-1}Z'Y$$

Consistency requires that in the limit the expectation of the sampling error is 0, i.e. $p\lim_{T\to\infty}(\hat{\delta}_{IV} - \delta) = E(\hat{\delta}_{IV} - \delta_{IV}) = 0$. We can derive the sampling error of this estimator as:

$$\hat{\delta}_{IV} - \delta = (Z'X)^{-1}Z'Y - \delta = (Z'X)^{-1}Z' \cdot (X\delta + \nu) - \delta$$

[4] Note that orthogonality conditions that include a constant as a regressor have the unconditional expectation $E(\nu_t) = 0$. Since our orthogonality condition is about the instrument (which includes a constant) we get the unconditional $E(\nu_t) = 0$ as well.

[5] If we had considered more than three instruments then we would not be able to use the IV estimator and would have to work instead with the Generalized Method of Moments, GMM.

$$
\begin{aligned}
&= \cancelto{I_g}{(Z'X)^{-1}Z'X}\ \delta + (Z'X)^{-1}Z'\nu - \delta = \delta + (Z'X)^{-1}Z'\nu - \delta \\
&= (Z'X)^{-1}Z'\nu = \left(\frac{Z'X}{T}\right)^{-1}\frac{Z'\nu}{T} = \left(\frac{\sum_{t=1}^{T} x_t z_t'}{T}\right)^{-1}\frac{\sum_{t=1}^{T} z_t \nu_t}{T}
\end{aligned}
$$

Now applying the weak law of large numbers[6], we have that:

$$
\underset{T\to\infty}{p\mathrm{lim}}\left(\frac{\sum_{t=1}^{T} x_t z_t'}{T}\right) = E(x_t z_t')\ ,\ \underset{T\to\infty}{p\mathrm{lim}}\frac{\sum_{t=1}^{T} z_t \nu_t}{T} = E(z_t \nu_t)
$$

And by Slutzky's theorem, given that we have $E(x_t z_t')$ is square and of full column rank (and so invertible):

$$
\underset{T\to\infty}{p\mathrm{lim}}\left(\frac{\sum_{t=1}^{T} x_t z_t'}{T}\right)^{-1} = E(x_t z_t')^{-1}
$$

Finally, Cramer's theorem gives the following result:

$$
\underset{T\to\infty}{p\mathrm{lim}}(\hat{\delta}_{IV} - \delta) = E(x_t z_t')^{-1}E(z_t \nu_t) = 0 \Leftrightarrow \underset{T\to\infty}{p\mathrm{lim}}\,\hat{\delta}_{IV} = \delta
$$

Where we have used that valid instruments satisfy weak exogeneity so that $E(z_t \nu_t) = 0$.

Part D: Discussion of Results

Several things of note are occurring as a result of the Monte-Carlo simulation.

- *Small vs. Large sample sizes for ARX(1)-IV, ARX(2)-OLS, and ARX(2)-IV estimation*:

  Comparing histograms of $n = 30$ vs. $n = 1000$ we can see the role that asymptotics play in consistency of estimation. For the case of $n = 30$ the histograms of $\hat{\rho}_i - \rho_i$ are not centered around 0 exactly for the ARX(1)-IV, ARX(2)-OLS, and ARX(2)-IV, all of which should be consistently estimated according to theory; rather the histograms are centered slightly negative of 0, indicating that at low sample sizes the estimates are slightly downward biased. Increasing the sample size to $n = 1000$ we observe the histograms centered at 0 almost exactly (still there is a slight downward bias - this is characteristic of OLS estimation of time series data). This demonstrates that both OLS and IV estimation are consistent asymptotically, and may produce biased estimates at low sample sizes; given that these are extremum estimators this should not be surprising. From the dispersion of the histograms we can see that OLS estimation of the properly specified ARX(2) process is more efficient even in low sample sizes compared to IV estimation of both the ARX(1) and ARX(2) specifications (discussed below).

- *Estimation of Misspecified ARX(1) Model with OLS*:

  Irrespective of sample size we observe the histogram of $\hat{\rho}_1 - \rho_1$ centered at around 0.32, indicating the estimate of this misspecified model has an upward bias. We expect this to occur as a result of the omitted variable bias in misspecifying the dynamics of the model, which leads to the violation of weak exogeneity.

[6] Specifically, since in our vector of z_t we have x and lags of x exogenous to the error term in expectation, we use the WLLN for martingale sequences.

- *Standard Errors - Precision of Estimation*

  The results of the estimation are shown in the tables below with HAC standard errors when appropriate.

Table 1: Parameter Point Estimates and Standard Errors

N=1000	ARX(1) - OLS	ARX(1) - IV	ARX(2) - OLS	ARX(2) - IV	ARX(1) - GMM
ρ_1	0.8413	0.4732	0.5126	0.4930	0.4726
	(0.0108)	(0.0334)*	(0.0147)	(0.0177)	(0.0335)*
ρ_2			0.3974	0.3974	
			(0.0140)	(0.0186)	

Standard errors are robust for heteroskedasticity. *HAC robust

Table 2: MC Parameter Estimates and Standard Errors

N=1000	ARX(1) - OLS	ARX(1) - IV	ARX(2) - OLS	ARX(2) - IV	ARX(1) - GMM
ρ_1	0.8274	0.4921	0,4990	0.4991	0.4939
	(0.0160)	(0.0315)	(0.0130)	(0.0157)	(0.0317)
ρ_2			0.3997	0.3997	
			(0.0133)	(0.0178)	

Simulation conducted with 1000 realizations of the time series.

As can be seen for the parameters of interest (ρ_1 and ρ_2) the estimation of the properly specified ARX(2) process (estimated with OLS) yields the most precise estimates, evidenced by the standard errors. Since this is the true model we expect the estimated asymptotic variance to be the most efficient. Further, the ARX(2)-IV estimates have a slightly lower precision than the OLS estimate, indicating that although IV estimation yields consistent estimates these are not efficient (this is especially visible in comparing the histograms). Moreover, comparing estimates of ρ_1 across the estimation procedures we see that the ARX(1)-IV estimation has the greatest drop in precision for the estimate compared to the ARX(2)-OLS and ARX(2)-IV methods; that is, obtaining the consistent estimate for ρ_1 comes with a sacrifice of efficiency when conducting IV estimation. Although not presented here, a similar table for $n = 30$ would yield the same general result concerning the decreasing precision in estimation across the methodologies and specifications. And finally comparing the point estimates of the coefficients from the single realization of the time series (Table 1) with the results from the Monte-Carlo simulation of 1000 realizations of the time series (Table 2) we can see that, in general, the Monte Carlo simulation generates estimates that are much closer to the true parameters of the DGP with greater precision in the estimation.

We may specifically attribute the decrease in the precision of the ARX(1)-IV process to the estimate of $\widehat{Avar(\hat{\delta}_{IV})}$. The estimation of the asymptotic variance requires the matrix of estimated fourth moments $\hat{\Omega} = Z'\hat{\nu}\hat{\nu}'Z$, where we can show that due to omitted variable bias we inflate the error term through $\nu_t = \rho_2 y_{t-2} + \varepsilon_t$. Therefore the sum of squared residuals increases necessarily as a result, inflating the estimation of $\widehat{Avar(\hat{\delta}_{IV})}$.

Based on these results we may state for this problem that

$$\widehat{Avar(\hat{\delta})}_{ARX(2)-OLS} \leq \widehat{Avar(\hat{\delta}_{IV})}_{ARX(2)} \leq \widehat{Avar(\hat{\delta}_{IV})}_{ARX(1)}$$

Problem 3

Part A: Define the IV Estimator as a GMM Estimator

Recall that we had exact identification for the ARX(1) process in defining the IV estimator (otherwise we would not have been able to utilize the Method of Moments). For a general representation, we may define the following moment conditions (of which we will have 3 in total, since we have three parameters to estimate):

$$m_3(\hat{\delta}) \equiv Z'(Y - X\hat{\delta})$$

such that this moment condition satisfies $E[m_3(\hat{\delta})] = 0$. Then we can define the following equation to minimize in Matlab, as was done in part E of problem 1:

$$J(\hat{\delta}, \hat{W}) = m_3(\hat{\delta})'\hat{W}m_3(\hat{\delta})$$

Now we need to define the weighting matrix $\hat{W}$ of *dim* 3×3. Since we will be conducting efficient 2-step GMM estimation, we may start with the identity matrix I_3. In the second step we update the weighting matrix with the efficient weighting matrix obtained through the consistant estimate of δ from step 1:

$$\underset{T\to\infty}{plim}\, \hat{W} = (\hat{\Omega})^{-1} \text{ , } \hat{\Omega} \equiv \left(\frac{1}{T}\right)\sum_{t=1}^{T} \hat{\nu}_t^2 z_t z_t' = Z'\hat{\nu}\hat{\nu}'Z \text{ (in matrix form)} \tag{6}$$

where $\hat{\nu}$ is the estimated residual given the estimate $\hat{\delta}$ from step 1. Note that reliable estimation of this matrix of fourth moments requires large sample sizes. So when $n = 30$ efficiency is not guaranteed; see the discussion in Hayashi, page 215 for more information.

Part B: Asymptotic Distribution of the Estimator

The sampling error of GMM can be shown to be (in matrix form):

$$\hat{\delta}(\hat{W}) - \delta = (X'Z\hat{W}Z'X)^{-1}X'Z\hat{W}Z'\nu$$

Multiplying by $\sqrt{n}$ on both sides and applying Slutzky's and Cramer's Theorem, then a central limit theorem we obtain the following (asymptotic) result:

$$\sqrt{n}\left(\hat{\delta}(\widehat{W}) - \delta\right) \to_d \mathcal{N}\left(0, n \cdot (D_\infty \widehat{W}_\infty D_\infty')^{-1} D_\infty \widehat{W}_\infty \Omega_\infty \widehat{W}_\infty D_\infty' (D_\infty \widehat{W}_\infty D_\infty')^{-1}\right) \tag{7}$$

where we have used $\Omega = Z'\nu\nu'Z$ in the above. Then substituting in (9) the estimation of Ω, $\widehat{\Omega}$, for the efficient weighting matrix as defined in (8), we can simplify this variance to:

$$\widehat{Avar\left(\hat{\delta}(\hat{\Omega}^{-1})\right)} = \left(X'Z\hat{\Omega}^{-1}Z'X\right)^{-1}$$

These terms may be estimated directly and will give the standard errors of the parameter estimates. (Notice that here all the n - or in our case, T - terms cancel out when generating the sample counterpart of (9).)

The HAC AVAR Estimator We were given that our error is classical (white noise), and presumably we do not need to correct for autocorrelation/serial correlation. Further, we were given an independant assumption on the distribution of x_t, which is why we were able to use one lag of x_t as an instrument. However, we do have a problem with the IV, GMM, and also the OLS estimation of the ARX(1) model, since ν_t no longer satisfies white noise assumptions: it is trivial to show that

$$\nu_t \sim \mathcal{N}\left[0, \rho_2^2\gamma_0 + \sigma_\varepsilon^2\right]$$

where γ_0 is as derived in problem 2 and $E(\nu_t) = \rho_2 E(y_{t-2}) = \rho_2\mu = 0$ since $\mu = 0$ (as we derived in problem 2). We then have $E(\nu_t\nu_{t-s}) = \rho_2^2\gamma_s \neq 0$, where γ_s is a generalization of the result we found in Problem 2 for the autocovariance of y. Recall that white noise requires $E(\varepsilon_t\varepsilon_{t-s}) = 0$. This failure implies there are autocorrelations in the error terms for the estimations of the ARX(1) model. We then implement the following procedure using the Newey-West (or Kernal) estimator. Estimate the residuals $\hat{\nu}$ and generate the diagnal matrix $\hat{\nu}\hat{\nu}'$. Call this matrix D. The heteroskedastically robust variance is then given by:

$$\widehat{Avar\left(\hat{\delta}(\hat{\Omega}^{-1})\right)}_H = \left(X'Z(Z'DZ)^{-1}Z'X\right)^{-1}$$

Then the Newey-West estimator may be obtained by defining the matrix $C = \hat{\nu}_t\hat{\nu}_{t-1}'$ similar to the matrix D. Next define the *dim* 3×3 matrix $\hat{\Gamma} = Z'CZ$. Now using an appropriate bandwidth, q,[7] we can estimate the HAC variance-covariance matrix:

$$\widehat{Avar\left(\hat{\delta}(\hat{\Omega}^{-1})\right)}_{HAC} = \widehat{Avar\left(\hat{\delta}(\hat{\Omega}^{-1})\right)}_H + \sum_{j=1}^{q}\left(1 - \frac{j}{q+1}\right)\left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right)$$

Problem 5

We must verify the missing steps to show that $n \cdot m_n(\hat{\theta})'\hat{\Omega}^{-1}m_n(\hat{\theta}) \rightarrow_d \chi^2_\nu$, $\nu = (g - K)$ where the degrees of freedom, $(g - K)$, are the g instruments and K regressors.

From the lecture notes we know that the annihilator matrix, P, which we must show is idempotent and has $trace(P) = g - K$, is defined as:

$$P = I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}$$

By idempotent we mean that a matrix is characterized by: 1) $P = PP$; and 2) $P = P'$. We must first check these two properties of the above defined P matrix.

- $P = PP$:

$$PP = (I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2})(I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2})$$

[7] The bandwidth is the chosen lag length for constructing the weights of the autocovariances. The bandwidth q should then grow in proportion to the size of the dataset, at a rate $T^{1/4}$, where consistency requires that $\frac{q}{T^{1/4}} \rightarrow 0$ as $T \rightarrow \infty$. In practice, however, we have finite data sets. Stock & Watson (2003) suggest a rule of thumb for calculating q, $q = 0.75T^{1/3}$, and taking the integer value (that is, round down).

$$= I_g I_g - 2I_g\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}+$$
$$\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}$$
$$= \overset{I_g}{\cancel{I_g I_g}} - 2I_g\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}+$$
$$\Omega_\infty^{-1/2}D_\infty'\cancel{(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}\Omega_\infty^{-1/2}D_\infty'}(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}$$
$$= I_g - 2I_g\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2} + \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}$$
$$= I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2} = P$$

- $P = P'$:

$$P' = (I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2})' = (I_g)' - (\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2})'$$
$$= I_g - \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2} = P$$

Hence the matrix is idempotent.

To show that the matrix has $trace(P) = g - K$ we first make the following definition:

$$A = \Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}D_\infty\Omega_\infty^{-1/2}$$

And we proceed as follows:

$$trace(P) = trace(I_g - A) = trace(I_g) - trace(A)$$

Since I_g is an identity matrix of $dim\ g \times g$ representing the instruments[8]:

$$trace(I_g) = g$$

Next, using the properties of a trace we can rearrange $trace(A)$ as follows[9]:

$$trace(A) = trace\left(\underbrace{\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}}_{C}\underbrace{D_\infty\Omega_\infty^{-1/2}}_{B}\right)$$

$$= trace\left(\underbrace{D_\infty\Omega_\infty^{-1/2}}_{B}\underbrace{\Omega_\infty^{-1/2}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}}_{C}\right) = trace\left(\overset{I_K}{\cancel{D_\infty\Omega_\infty^{-1}D_\infty'(D_\infty\Omega_\infty^{-1}D_\infty')^{-1}}}\right) = K$$

This shows that $trace(P) = g - K$.

To show that given these properties we obtain our desired result, consider that in the class notes we are also given that:

$$\sqrt{n}\hat{\Omega}^{-1/2}m_n(\hat{\theta}) \rightarrow_d \mathcal{N}(0, I_g) \text{ (the standard normal)}$$

and further given the following quadratic form:

$$\mathcal{N}(0, I_g)' \cdot P \cdot \mathcal{N}(0, I_g)$$
$$= \left(\sqrt{n}\hat{\Omega}^{-1/2}m_n(\hat{\theta})\right)' P \left(\sqrt{n}\hat{\Omega}^{-1/2}m_n(\hat{\theta})\right) = n \cdot m_n(\hat{\delta})'\hat{\Omega}^{-1/2}P\hat{\Omega}^{-1/2}m_n(\hat{\theta})$$

we can apply the following property:

A quadratic form $x'Ax$ where $x \sim \mathcal{N}(0, 1)$ and A is idempotent implies $x'Ax \sim \chi^2_\nu$ with $\nu = trace(A)$

Therefore, we have that:

$$n \cdot \left(m_n(\hat{\delta})'\hat{\Omega}^{-1/2}\right)\left(\hat{\Omega}^{-1/2}m_n(\hat{\theta})\right) = n \cdot m_n(\hat{\theta})'\hat{\Omega}^{-1}m_n(\hat{\theta}) \rightarrow_d \chi^2_\nu\ ,\ \nu = tace(P) = (g - K)$$

[8] by the property that the trace equals the sum of the diagnol elements in a sqare matrix

[9] $trace(CB) = trace(BC)$

Problem 6

We must provide an explanation of how to estimate each quantity that appears in the test statistic that we choose to test the linear restrictions given.

Supposing we use an arbitrary weighting matrix W_∞ such that the asymptotic distribution of our GMM estimator is:

$$\sqrt{n}(\hat{\theta}-\theta) \to_d \mathcal{N}\left[0, n\cdot(D_\infty W_\infty D'_\infty)^{-1}D_\infty W_\infty \Omega_\infty W_\infty D'_\infty (D_\infty W_\infty D'_\infty)^{-1}\right]$$

then the asymptotic estimate for our variance-covariance matrix given a consistent estimate of θ, $\hat{\hat{\theta}}(W_\infty)$, will be given by:

$$\widehat{Avar\left(\hat{\theta}(W_\infty)\right)} = (D_\infty W_\infty D'_\infty)^{-1}D_\infty W_\infty \widehat{\Omega}_\infty W_\infty D'_\infty (D_\infty W_\infty D'_\infty)^{-1} \tag{8}$$

With the above estimate we wish to conduct the inference test on the hypotheses: $H_o: R\theta = r$ versus $H_A: R\theta \neq r$, where $dim\ r = q\times 1$ vector of total restrictions and $dim\ R = q\times K$ (q restrictions × K regressors).

Before moving onto the proper test statistic, we will first go over the estimation of (10) above. We know that a consistent estimate of D_∞ will be given by $X'Z$, where X is the $n\times K$ design matrix of regressors and Z is the $n\times g$ matrix of exogenous (i.e. $E(z_i\cdot\varepsilon_i)=0$) variables, where we require that $E(X'Z)\neq 0$ and $X'Z$ is of full rank. Then given our arbitrary W_∞ matrix we can estimate $\widehat{\Omega}_\infty$ from any consistent estimate $\hat{\theta}(W_\infty)$.

Moving now to the test statistic to evaluate the hypotheses, the discussion of proposition 3.8 in Hayashi (page 222 and 223) gives us an insight as to why we are given in the hint to use the Wald statistic (as opposed to the LR statistic, our other option). Since we are dealing with linear restrictions the proposition tells us that the two statistics are numerically equivalent. However, the LR statistic is asymptotically χ^2 only when the efficiency condition $p\lim W_\infty = \widehat{\Omega}_\infty^{-1}$ is satisfied. As we are using an arbitrary weighting matrix rather than this efficient weighting matrix, the LR is not guaranteed to be asymptotically χ^2. The Wald statistic, on the other hand, is asymptotically χ^2 even if W_∞ does not satisfy this efficiency condition.

We know from the class notes that the form of the Wald statistic for linear restrictions is easily shown to be[10]:

$$W = \left(R\hat{\delta}(W_\infty)-r\right)'\left[R\widehat{Avar\left(\hat{\delta}(W_\infty)\right)}R'\right]^{-1}\left(R\hat{\delta}(W_\infty)-r\right) \to_d \chi^2(dim\ r) \tag{9}$$

Since we have already described how to estimate the asymptotic variance-covariance matrix in (6), the Wald test statistic presented in (7) is straightforward to calculate as we know a priori the forms of R and r.

Interpreting the result of this statistic is straightforward. The test statistic will tell us for a given sampling error, $\hat{\delta}(W_\infty)-\delta$, whether the restrictions are satisfied under the null hypothesis for a level α test, comparing with the known distribution $\chi^2(dim\ r)$. If we obtain that $W \leq \chi^2(dim\ r)$ then we have evidence that the restrictions fall under the null (and thus we fail to reject it). If, however, we find that $W > \chi^2(dim\ r)$ then with our given variance of the sampling error we have evidence that we may reject the null in favor of the alternative.

[10] See page 489 in Hayashi for a thorough derivation of this statistic using the taylor expansion of the sampling error and the mean value theorem around $\bar{\theta}$, where $\bar{\theta}\in[\hat{\theta},\theta]$.