Spring 1997 John Rust
Economics 551b 37 Hillhouse, Rm. 27

MIDTERM EXAM: SOLUTIONS

QUESTION 1 Seemingly unrelated regression estimator is identical to OLS estimator when regressors of all equations are identical. (Therefore you can check your SUR program by comparing SUR estimates and OLS estimates. Now, you have your own SUR program which you can use to estimate models with different regressors(!).)

We know that GLS estimator of SUR model is unbiased and efficient and can be written as

$$\widehat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

where

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}, \qquad X_i = \begin{bmatrix} X_{i1} & & 0 \\ & \ddots & \\ 0 & & X_{ip} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Omega & & 0 \\ & \ddots & \\ 0 & & \Omega \end{bmatrix} = I_N \otimes \Omega$$

If we introduce the regressor matrices stacked by equation by equation instead of observation by observation (see solution of Q4 in Problem set 3), this estimator can be rewritten as:

$$\widehat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

where

$$X = \begin{bmatrix} X_{(1)} & & 0 \\ & \ddots & \\ 0 & & X_{(p)} \end{bmatrix}, \qquad X_{(j)} = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{Nj} \end{bmatrix}.$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 I_N & & \sigma_{1p} I_N \\ & \ddots & \\ \sigma_{p1} I_N & & \sigma_p^2 I_N \end{bmatrix} = \Omega \otimes I_N$$

Suppose we have identical regressors, $X_{(1)} = \cdots = X_{(p)} = Z$.

$$X = \begin{bmatrix} Z & & 0 \\ & \ddots & \\ 0 & & Z \end{bmatrix} = I_p \otimes Z$$

Then

$$
\begin{aligned}
\widehat{\beta}_{GLS} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\
&= \Big((I_p \otimes Z)'(\Omega \otimes I_N)^{-1}(I_p \otimes Z)\Big)^{-1}(I_p \otimes Z)'(\Omega \otimes I_N)^{-1}y \\
&= \Big((I_p \otimes Z')(\Omega^{-1} \otimes I_N)(I_p \otimes Z)\Big)^{-1}(I_p \otimes Z')(\Omega^{-1} \otimes I_N)y \\
&= (\Omega^{-1} \otimes Z'Z)^{-1}(\Omega^{-1} \otimes Z')y \\
&= \Big((\Omega \otimes (Z'Z)^{-1}\Big)(\Omega^{-1} \otimes Z')y \\
&= (I_p \otimes (Z'Z)^{-1}Z')y \\
&= \begin{bmatrix} (Z'Z)^{-1}Z' & & 0 \\ & \ddots & \\ 0 & & (Z'Z)^{-1}Z' \end{bmatrix} \begin{bmatrix} y_{(1)} \\ \vdots \\ y_{(p)} \end{bmatrix} = \begin{bmatrix} (Z'Z)^{-1}Z'y_{(1)} \\ \vdots \\ (Z'Z)^{-1}Z'y_{(p)} \end{bmatrix} \\
&= \widehat{\beta}_{OLS}
\end{aligned}
$$

1. $\gamma = \frac{\alpha_i}{1-\beta_i}$ should be constant among equations.
2. SUR-GLS estimator or OLS estimator equation by equation from argument above.
3. Construct sample covariance matrix from OLS residuals.
4. See the calculation above.
5. I apologize for asking you to do Gibbs sampling for the full set of 63 securities: this isn't really feasible on most computers in a reasonable amount of time. With 63 securities it takes special programming to be able to run the Gibbs sampling program without using a great deal of memory space. Even when you do this the program still takes quite a bit of cpu time for the ordinary Pentium cpus in the Statlab. However if you choose a few securities from the data set, say 3 or 4 stocks, then Gibbs sampling runs reasonably quickly. Any student who just did this part using just a subset of securities in `stockdat` received full credit. All students who even attempted to do this part received generous partial credit, and nearly full credit if you developed a computer program to carry out the Gibbs sampling that looked correct and included it in your anwer, regardless of whether it actually worked for the problem using all 63 securities. I was more interested in getting students to write code for this problem rather than seeing actual output since the output will differ from person to person by the nature of the Gibbs sampling algorithm. I have written a program `surgibbs.gpr` that carries out Gibbs sampling for seemingly unrelated regression models and have the code set up to do Gibbs sampling with 3 stocks from `stockdat` so you can see how it runs. The program comes with 2 procedures, `set_hp.g` and `get_dat.g` which you can modify to use the software for other problems, including an artificial problem where you know the true β and Ω parameters in advance (this is useful as a check that the program is coded correctly).

QUESTION 2 Let N_k be the number of times which outcome k occured in all sample.

$$N_k = \sum_{i=1}^{N} I\{x_i = k\}, \qquad \sum_{k=1}^{K} N_k = N$$

With

$$f(x_i|\theta) = \prod_{k=1}^{K} \theta_k^{I\{x_i=k\}},$$

the joint distribution of $\{\tilde{X}_1, \ldots, \tilde{X}_N\}$ can be written as

$$f(x_1, \ldots, x_N|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}.$$

1. Since log likelihood function is

$$\begin{aligned} l_N(\theta) &= \log f(x_1, \ldots, x_N|\theta) \\ &= \sum_{k=1}^{K} N_k \log \theta_k \\ &= \sum_{k=1}^{K-1} N_k \log \theta_k + \left(N - \sum_{k=1}^{K-1} N_k\right) \log\left(1 - \sum_{k=1}^{K-1} \theta_k\right), \end{aligned}$$

MLE $\widehat{\theta}_k$(for $k = 1, \ldots, K-1$) can be derived from

$$\frac{\partial}{\partial \theta_k} l_N(\theta) = \frac{N_k}{\theta_k} - \frac{(N - N_1 - \cdots - N_{K-1})}{(1 - \theta_1 - \cdots - \theta_{K-1})} = 0.$$

The solution to this system is

$$\widehat{\theta}_k = \frac{N_k}{N}.$$

2.Since

$$E\left[N_k\right] = \sum_{i=1}^{N} EI\{x_i = k\} = \sum_{i=1}^{N} \theta_k = N\theta_k$$

$$E\left[\widehat{\theta}_k\right] = E\left[\frac{N_k}{N}\right] = \frac{1}{N} E\left[N_k\right] = \frac{1}{N} N\theta_k = \theta_k.$$

Therefore $\widehat{\theta}_k$ is an unbiased estimator.
3.

$$\begin{aligned} E\left[N_k^2\right] &= E\left[\left(\sum_{i=1}^{N} I\{x_i = k\}\right)^2\right] \\ &= \sum_{i=1}^{N} EI\{x_i = k\}^2 + \sum_{i=1}^{N}\sum_{j \neq i}^{N} EI\{x_i = k\} I\{x_j = k\} \\ &= N\theta_k + N(N-1)\theta_k^2. \end{aligned}$$

$$E\left[\widehat{\theta}_k^2\right] = \frac{1}{N^2} E\left[N_k^2\right] = \frac{\theta_k}{N} + \frac{(N-1)\theta_k^2}{N}.$$

$$
\begin{aligned}
\mathrm{var}(\widehat{\theta}_k) &= E\left[\widehat{\theta}_k^2\right] - \left\{E\left[\widehat{\theta}_k\right]\right\}^2 \\
&= \frac{\theta_k}{N} + \frac{(N-1)\theta_k^2}{N} - \theta_k^2 = \frac{1}{N}(1-\theta_k)\theta_k
\end{aligned}
$$

For $k \neq l$

$$
\begin{aligned}
E\left[N_k N_l\right] &= E\left[\left(\sum_{i=1}^{N} I\{x_i = k\}\right)\left(\sum_{i=1}^{N} I\{x_i = l\}\right)\right] \\
&= \sum_{i=1}^{N} EI\{x_i = k\}I\{x_i = l\} + \sum_{i=1}^{N}\sum_{j\neq i}^{N} EI\{x_i = k\}I\{x_j = l\} \\
&= 0 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} EI\{x_i = k\}EI\{x_j = l\} \\
&= N(N-1)\theta_k\theta_l.
\end{aligned}
$$

Note $E\left[I\{x_i = k\}I\{x_i = l\}\right] = 0$ since at least one of $I\{x_i = k\}$ and $I\{x_i = l\}$ must be zero.

$$
E\left[\widehat{\theta}_k\widehat{\theta}_l\right] = \frac{1}{N^2}E\left[N_k N_l\right] = \frac{(N-1)\theta_k\theta_l}{N}
$$

$$
\begin{aligned}
\mathrm{cov}(\widehat{\theta}_k, \widehat{\theta}_l) &= E\left[\widehat{\theta}_k\widehat{\theta}_l\right] - E\left[\widehat{\theta}_k\right]E\left[\widehat{\theta}_l\right] \\
&= \frac{(N-1)\theta_k\theta_l}{N} - \theta_k\theta_l = -\frac{1}{N}\theta_k\theta_l
\end{aligned}
$$

Therefore

$$
\Sigma = \mathrm{cov}(\widehat{\theta}) = \frac{1}{N}\begin{bmatrix}
\theta_1(1-\theta_1) & -\theta_1\theta_2 & \cdots & -\theta_1\theta_{K-1} \\
-\theta_1\theta_2 & \theta_2(1-\theta_2) & \cdots & -\theta_2\theta_{K-1} \\
\cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots \\
-\theta_1\theta_{K-1} & \theta_2\theta_{K-1} & \cdots & \theta_{K-1}(1-\theta_{K-1})
\end{bmatrix}.
$$

4.
For $k = l$

$$
\frac{\partial^2}{\partial\theta_k^2} l_N(\theta) = -\frac{N_k}{\theta_k^2} - \frac{(N - N_1 - \cdots - N_{K-1})}{(1-\theta_1-\cdots-\theta_{K-1})^2}
$$

$$
\begin{aligned}
E\left[\frac{\partial^2}{\partial\theta_k^2} l_N(\theta)\right] &= -\frac{N\theta_k}{\theta_k^2} - \frac{(N - N\theta_1 - \cdots - N\theta_{K-1})}{(1-\theta_1-\cdots-\theta_{K-1})^2} \\
&= -\frac{N}{\theta_k} - \frac{N}{(1-\theta_1-\cdots-\theta_{K-1})}
\end{aligned}
$$

For $k \neq l$

$$
\frac{\partial^2}{\partial\theta_k\partial\theta_l} l_N(\theta) = -\frac{(N - N_1 - \cdots - N_{K-1})}{(1-\theta_1-\cdots-\theta_{K-1})^2}
$$

$$
\begin{aligned}
E\left[\frac{\partial^2}{\partial\theta_k\partial\theta_l} l_N(\theta)\right] &= -\frac{(N - N\theta_1 - \cdots - N\theta_{K-1})}{(1-\theta_1-\cdots-\theta_{K-1})^2} \\
&= -\frac{N}{(1-\theta_1-\cdots-\theta_{K-1})}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_N(\theta) &= -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} l_N(\theta)\right] \\
&= N\begin{bmatrix} \frac{1}{\theta_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \frac{1}{\theta_{K-2}} & 0 \\ 0 & \cdots & \cdots & \frac{1}{\theta_{K-1}} \end{bmatrix} \\
&\quad + N\frac{1}{(1-\theta_1-\theta_2-\cdots-\theta_{K-1})}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ 1 & \cdots & \cdots & 1 \\ 1 & \cdots & \cdots & 1 \\ 1 & 1 & \cdots & 1 \end{bmatrix}.
\end{aligned}
$$

5. MLE can be proved to be efficient if it has a variance equal to Cramer-Rao lower bound. You can easily show $\Sigma = \mathcal{I}_N^{-1}(\theta)$ by verifying $\Sigma \times \mathcal{I}_N(\theta) = I$.

6. Run the program `dirichlet.gpr` (which in turn calls a procedure `setparm.g`) available on the Econ 551 Web page. This program does the calculations required in part 6, i.e. it compares that posterior probability that $\tilde{\theta}$ (a draw from the Dirichlet posterior) is within a ball of radius .01 of the true parameter θ^*. I calculate this probability by simulation. Since this posterior probability is simulated, and since the data used to form the posterior is simulated, it makes no sense to report numbers here since they vary from run to run. The important part of this exercise is to verify that the simulated probability from the exact posterior is very close to the probability calculated from the normal approximation to the posterior.

7. First, notice that the full $K \times 1$ vector $\lambda = (\lambda_1, \ldots, \lambda_K)$ is not identified: you can add a constant μ to each component of λ and the probability in (5) will be unchanged. Therefore I impose an arbitrary normalization that $\lambda_K = 0$ and our problem reduces to the estimation of the unrestricted $(K-1) \times 1$ vector $(\lambda_1, \ldots, \lambda_{K-1})$. Another arbitrary but convenient normalization is:

$$
1 = \sum_{k=1}^{K} \exp\{\lambda_k\}.
$$

Under the first normalization, one can show that there is a one to one mapping between the $K-1$ λ parameters and the $K-1$ θ parameters. Thus, the by the invariance of maximum likelihood, it is easy to see that $\hat{\lambda}$ will be given by the unique solution to the $K-1$ system of equations

$$
\hat{\theta}_k = \frac{\exp\{\hat{\lambda}_k\}}{1 + \sum_{k'=1}^{K-1} \exp\{\hat{\lambda}_{k'}\}}.
$$

Under the second normalization we can get an explicit representation: $\hat{\lambda}_k = \exp\{\hat{\theta}_k\}$. Since $\hat{\theta}_k$ is an unbiased estimator as was verified in part 2 above, it follows from Jensen's inequality that $\hat{\lambda}_k = \exp\{\hat{\theta}_k\}$ will be a downward biased estimator of $\lambda_k^* = \log(\theta_k^*)$. Due to the presence of bias and the arbitrariness of the identifying normalization, it is difficult to determine whether the MLE $\hat{\lambda}_N$ attains the generalized Cramer-Rao lower bound in finite samples. However we know that this is a regular problem, so the MLE is asymptotically unbiased and efficient, and so does attain the Cramer-Rao lower bound asymptotically. Under the first identifying normalization, $\lambda_K = 0$, it is easy to calculate the information matrix in this case and verify that it is the same as the *inverse* of the information matrix for θ. This should not be surprising, since the λ parameters are an inverse transformation of the θ parameters in this case. I present the derivation below. The information matrix $I(\lambda^*)$ is given by:

$$I(\lambda^*) = \sum_{k=1}^{K} \frac{\partial}{\partial \lambda} \log p_k(\lambda^*) \frac{\partial}{\partial \lambda'} \log p_k(\lambda^*) p_k(\lambda^*).$$

But we have

$$\log p_k(\lambda^*) = \lambda_k^* - \log\left[1 + \sum_{j=1}^{K-1} \exp\{\lambda_j^*\}\right],$$

and

$$\frac{\partial}{\partial \lambda} \log p_k(\lambda^*) = e_k - p(\lambda^*),$$

where e_k is a $(K-1) \times 1$ vector of zeros with a 1 in the k^{th} place. Substituting this into the equation for $I(\lambda^*)$ above we get

$$I(\lambda^*) = \operatorname{diag}\left[p(\lambda^*)\right] - p(\lambda^*)p(\lambda^*)'.$$

You can verify that this matrix is the same as the covariance matrix for $\hat{\theta} = p(\hat{\lambda})$ given above, and as already demonstrated, this covariance matrix is the inverse of the information matrix for θ. So the information matrix for λ is the inverse of the information matrix for θ.

QUESTION 3

$$\begin{aligned}
|H_N(\hat{\theta}) - H(\theta^*)| &= |H_N(\hat{\theta}) - H(\hat{\theta}) + H(\hat{\theta}) - H(\theta^*)| \\
&\leq |H_N(\hat{\theta}) - H(\hat{\theta})| + |H(\hat{\theta}) - H(\theta^*)| \\
&\leq \sup_{\theta \in \Theta} |H_N(\theta) - H(\theta)| + |H(\hat{\theta}) - H(\theta^*)|
\end{aligned}$$

For the first term in the last inequality, $\sup_{\theta\in\Theta} |H_N(\theta) - H(\theta)| \to 0$ with probability 1 by assumption. For the second term, H continuous and $\hat{\theta} \to \theta^*$ with probability 1, together implies $|H(\hat{\theta}) - H(\theta^*)| \to 0$ with probability 1 by continuous mapping theorem.

Therefore $H_N(\hat{\theta}) \to H(\theta^*)$ with probability 1.

QUESTION 4. The Econ 551 web has *Gauss* programs for computing maximum likelihood estimates of the multinomial logit model plus the shell programs for running maximum likelihood. I have also posted `evalbprob.g` which is the procedure for computing the likelihood and derivatives for the biniomial probit model. You can run each of these programs to get the maximum likelihood estimates for the models. I have posted the estimation results for each of these programs on the Econ 551 web page.

A. The shell program for the estimation is `tnlest.gpr` the procedure for calculating likelihood and derivatives of the alternative specific version of the trinomial logit model is `evalmnl1.g` and the output of the program is in the file `data1.est` all on the Econ 551 web page. The file `data1.est` presents estimation results for the full sample of 2,000 data points and the subsample of the first 1,500 data points. The coefficient estimates for both samples are similar, and the standard errors are higher for the the 1,500 subsample as expected. The question about the R^2 leads us into hypothesis testing. A χ^2 goodness of fit test statistic would be appropriate here. But I didn't expect students to know anything about this since we haven't covered this topic at this point in Econ 551. But to anticipate, if the model is correctly specified then we should have the regression equation:

$$I\{\tilde{d}_i = j\} = P(j|X_i, \theta^*) + \epsilon_i$$

That is, $I\{\tilde{d}_i = j\}$ is an indicator for the event that the decision $\tilde{d}_i$ taken by person i is alternative j should equal the choice probability $P(j|X_j, \theta^*)$ plus an error term, ϵ_i. Since $P(j|X_j, \theta^*)$ is the conditional expectation of $I\{\tilde{d}_i = j\}$, the error term must have mean zero. Therefore we can "test" the model both in sample and out of sample by computing the mean prediction errors, and seeing how close they are to zero:

$$\tau_N^* = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[I\{d_i = j\} - P(j|X_j, \theta^*) \right].$$

Under standard regularity conditions, if the null hypothesis that the model is correctly specified, $\tau_N^* \underset{d}{\Longrightarrow} \tilde{Z} \sim N(0, \sigma^2)$ where σ^2 is the unconditional variance of ϵ_i. This can be used to form a test statistic. However we need to replace the unknown true θ^* with the maximum likelihood estimate $\hat{\theta}_N$ and then determine the asymptotic distribution of the corresponding $\hat{\tau}_N$ statistic, doing an "Amemiya correction" for the fact that we are using an estimated value $\hat{\theta}_N$ to construct $\hat{\tau}_N$ and we need to account for this extra estimation error in the derivation of the asymptotic distribution of the test statistic. Once we do this, we can use the τ statistic as a "moment condition" and set up a hypothesis test to see how well this moment condition is satisfied. Indeed, since $E\{X\epsilon\} = E\{XE\{\epsilon|X\}\} = 0$, we can construct an entire family of moment conditions that should be zero if the model is correctly specified. We can actually construct an entire vector of such test statistics, one for each alternative $j = 1, \ldots, J$ and for each instrument X that we use in forming the moment conditions. Later in Econ 551 we will show how to test all these moment conditions simultaneously using a single Chi-square goodness of fit statistic. This statistic would be the analog of the R^2. It is important to note that we should expect the model to fit better within sample than out-of-sample since maximum likelihood is designed to choose parameters to fit a given sample of data as well as possible. In fact, with a full set of alternative specific dummies, it is easy to show that the mean prediction error for each alternative is identically zero within sample, i.e. $\hat{\tau}_N = 0$. However $\hat{\tau}_N$ is not necessarily zero if we use out of sample data, i.e. if we estimate $\hat{\theta}_N$ using $\{d_i, X_i\}$ for $i = 1, \ldots, N$ and then construct $\hat{\tau}_N$ using $\{d_i, X_i\}$ for $i = N+1, \ldots, N+M$. The program `tnlest.gpr` presents the mean prediction errors (by alternative) both in-sample and out-of-sample (where in sample $\hat{\tau}_N$ was constructed from the first $N = 1500$ observations and the out-of-sample $\hat{\tau}_M$ was constructed from $\hat{\theta}_N$ and the remaining $M = 500$ observations in the data set. The results verify that the mean prediction errors are zero (modulo rounding error) within sample, but are non-zero out of sample. Later in the course we will post software for doing χ^2 specification tests using the residuals from the estimated choice model.

B The Econ 551 Web page has a shell program `bpest.gpr` and corresponding procedure `evalbp.g` to carry out the maximum likelihood estimation of the binary probit model. The estimation output is in the file `data2.est`. The true coefficient vector that was used to generate the data was $\beta^* = (1, -.5, 1)$ (the coefficients for the first alternative were normalized to zero) and the covariance matrix of the 2×1 error vector is given by

$$\Omega = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}$$

C The Econ 551 Web page has a Gauss file `surgibbs.gpr` which carrys out the Gibbs sampling for the seemingly unrelated regression model. You can use this program to do Gibbs sampling on the data-augmented version of the probit model (i.e. where we treat utilities as observed) and structure the algorithm as in Rossi and McCulloch *Journal of Econometrics* (1994). I gave generous partial credit to any student who attempted to do this problem, and full credit to any student who developed computer code that looks correct and included it with the answer. By the way, the true coefficient vector used to generate the data was $\beta^* = (1, -.5, 1, -1, 2, -.8)$ where these 6 coefficients are for the constant and slopes of the 2 X variables for alternatives 2 and 3, and the coefficients for alternative 1 were normalized to zero. The covariance matrix for the error terms in the probit model is given by:

$$\Omega = \begin{bmatrix} 1 & .2 & 0 \\ .2 & 1 & -.1 \\ 0 & -.1 & 1 \end{bmatrix}$$

QUESTION 5. The fact that $\mathrm{med}(\epsilon_t|X_t) = 0$ implies that $\mathrm{med}(y_t|X_t) = X_t\beta^*$. But from earlier in the semester we know that the median is the solution to the minimization problem

$$\mathrm{med}(\tilde{y}) = \underset{\mu}{argmin}\, E\{|\tilde{y} - \mu|\}.$$

Conditioning on X it follows that $X\beta^*$ is the solution to the problem

$$X\beta^* = \underset{\mu}{argmin}\, E_X\{|\tilde{y} - \mu|\},$$

so unconditionally we have:

$$\beta^* = \underset{\beta}{argmin}\, E\{|\tilde{y} - \tilde{X}\beta|\}.$$

Under assumptions of no multicollinearity between the columns of $\tilde{X}$, β^* will be the unique solution to the above minimization problem and hence is uniquely identified. Assume the observations (y_t, X_t) are *IID*. By the uniform strong law of large numbers we have with probability 1:

$$\sup_{\beta\in\Theta} \left\| \frac{1}{T}\sum_{t=1}^{T} |y_t - X_t\beta| - E\{|\tilde{y} - \tilde{X}\beta|\}\right\| \to 0,$$

where Θ is a compact set in R^k containing β^*. Since $E\{|\tilde{y}-\tilde{X}\beta\|\}$ is uniquely minimized at β^*, it follows that $\hat{\beta}_T \to \beta^*$ with probability 1. For further details on the asymptotic properties of the LAD estimator (including a derivation of the asymptotic distribution of the LAD estimator), see Koenker and Bassett, "Regression Quantiles" in the January 1978 *Econometrica.*