I. General Estimation Apparatus

a. General Overview

Typically we think of the true β_0 as satisfying some population moments: $E\left(f(y_t, x_t, \beta_0)\right) = 0$ for SOME function f.

Then, we use the analogy principle to estimate based on sample: $L_T\left(\hat{\beta}\right) = 0$ gives us the finite sample estimate.

If we pick l_t (*moment equations*) *s.t.*:

1. l_t *~ niid*

2. l_t *and* S_T *satisfies "regularity conditions": (dom. functions for* l_t *and* $\frac{\partial l_t}{\partial \theta}$ *to ensure uniform convergence so we can pass* $p\lim s$)

   a) (*in scalar case*) $l_t(\theta)$ *is a continuous function of* θ *satisfying* $|l_t^2| \le h_t(y_t)\ \forall\ \theta$ *s.t.* $E|h_t|^{1+\delta} < 0$ *for some* δ

   and $S_T(\theta) = \frac{1}{T}\frac{\partial L_T}{\partial \theta} = \frac{1}{T}\sum_t \frac{\partial l_t}{\partial \theta}$ *is continuous in* θ *satisfying* $\left|\frac{\partial l_t}{\partial \theta}\right| \le g_t(y)\ \forall\ \theta$ *s.t.* $E|g_t|^{1+\delta} < \infty$ *for some* δ

   b) (*in vector case*) $l_t(\theta)$ *is a continuous function of* θ *satisfying* $|\lambda_1' l_t l_t' \lambda_2| \le h_t(y_t)\ \forall\ \theta$ *s.t.* $E|h_t|^{1+\delta} < \infty$ *for some* δ

   and $S_T(\theta) = \frac{1}{T}\frac{\partial L_T}{\partial \theta} = \frac{1}{T}\sum_t \frac{\partial l_t}{\partial \theta}$ *is continuous in* θ *satisfying* $\left|\lambda_1' \frac{\partial l_t}{\partial \theta} \lambda_2\right| \le g_t(y)\ \forall\ \theta$ *s.t.* $E|g_t|^{1+\delta} < 0$ *for some* δ

3. $E\left(l_t\left(\theta_0\right)\right) = 0\ \forall\ t$ (*which implies* $E(L_T(\theta_0)) = 0$)

Then, if we pick $\hat{\theta}$ *s.t.* $L_T(\hat{\theta}) = 0$ (*i.e. the* $\hat{\theta}$ *that solves the sys of eqns* $L_T(\theta) = \frac{1}{T}\sum_t l_t(\theta) = 0$),

(*with* $\dim L_T = \dim l_t = \dim \theta$ *so we can solve system exactly, k eqns and k unknowns*),

then $\hat{\theta}$ *is consistent for true* θ_0, *with* $L_T(\theta_0) \xrightarrow{P} 0$ (*equivalently,* $E(l_t(\theta_0)) = 0$, *and* $\hat{\theta} \to_D N(\theta_0, S\left(\theta_0\right)^{-1} V S\left(\theta_0\right)^{-1})$

Thus, for all our estimators, as long as we can find the relevant quantities, then the estimator that minimizes the relevant distance function is consistent for the true parameter, and we know immediately consistency and asymptotic variance-covariance.

b. Asymptotic Variance assuming all assumptions satisfied

$$\boxed{S\left(\theta_0\right)^{-1} V^{-1} S\left(\theta_0\right)^{-1}}$$

$$V = A\mathrm{var}\left(\sqrt{T} L_T\left(\theta_0\right)\right) = E\left[l_t\left(\theta_0\right) l_t'\left(\theta_0\right)\right]$$

$$S\left(\theta_0\right) = P\lim\left(\frac{\partial}{\partial \theta} L_T\left(\theta_0\right)\right) = E\left(\frac{\partial}{\partial \theta} l_t\left(\theta_0\right)\right)$$

c. Finite Sample Approximation of the Asymptotic Distribution:

$$\hat{\theta} \sim_A N\left(\theta_0, \frac{1}{T}\hat{S}\left(\hat{\theta}\right)^{-1}\hat{V}\left(\hat{\theta}\right)\hat{S}\left(\hat{\theta}\right)^{-1}\right)$$

d. Sampling Error Formula and Identification Condition:

From multivariate mean value theorem, $L_T\left(\hat{\theta}\right) = L_T\left(\theta_0\right) + S\left(\tilde{\theta}\right)\left(\hat{\theta} - \theta_0\right) = 0$ *for some* $\tilde{\theta}$ *convex combination of* $\hat{\theta}$ *and* θ_0, we obtain:

$$\hat{\theta} - \theta_0 = \left(-S\left(\tilde{\theta}\right)\right)^{-1} L_T(\theta_0)$$

Note that $P\lim - S\left(\theta_0\right)$ invertible is the identification condition! (Otherwise there will be nontrivial linear combinations of $p\lim \hat{\theta} - \theta_0$ that give us $L_T = 0$ at the limit. Then we don't have unique minimization of the problem; i.e. parameter not identified.

Note2: This also tells us that

II. Extremum Estimators: Estimators and Distance Functions

a. Estimators are defined by the distance functions they minimize

All the estimators we will consider minimize a quadratic form: $Q_T = \frac{1}{T}e'Me$

The question is, how do we pick an M to optimize efficiency.

GLS: $Q_T = \frac{1}{T}e'Me$ **where** $e \sim (0, \Phi)$ **(Then, by Gauss-Markov the most efficient estimator in this class uses M = $\alpha\Phi^{-1}$ → GLS)**[1]

OLS: $Q_T = \frac{1}{T}e'e$

GMM: $Q_T = H_T^{'}MH_T = (ve)'M(ve) = e'\left(v'Mv\right)e$ **(The most efficient of this class is also a GLS estimator because it requires M to be the Var(H_T))**

Clearly: GMM is a special case of GLS in that for any M in GMM, we can find a weighting matrix M in the GLS framework that will get us the same estimator.

The **efficiently picked M**, $E\left(h_t(\theta_0)h_t^{'}(\theta_0)\right)^{-1}$ (variance of the moment conditions), **is the GLS version of the estimator and the "efficient GMM" estimator.**.

Note: H_T is a reduction in dimensionality that is necessary when there is overidentification.

(Everything we do can be thought of in this GMM framework. But remember, the "efficient" GMM is a GLS estimator)

ML: $Q_T = \frac{1}{T}\sum_t \log f_t(y_t, x_t; \beta)$ *where* f_t *is the density function of the observed data*

[1] Using this weighing matrix, we have $\Phi^{-1/2}e \sim (0, I)$. Then the setup is same as Gauss-Markov and we obtain a BLUE estimator.

b. Distance function and L_T

The estimator solves the system of equations given by: $\frac{\partial Q_T\left(\hat{\theta}\right)}{\partial \theta} = L_T(\hat{\theta}) = 0$

Note: The way we have constructed Q_T and L_T, L_T is always just identified (dim L_T = K = dim θ)
(In least squares problems this is not an issue. Just check FOC of the problem.)
(In the case of GMM, H_T reduces the dimensionality to make "just identified". In case of IV, weighting matrix does not matter because G() is square and assumed to be invertible)

c. Note on Feasibility of GLS

- e'Me:If our metric is e'Me, then in order for the optimal M to be estimable we need to make assumptions about the functional form of the heteroskedasticity (if any), to get **Feasible GLS**
- H_T'MH_T: In this case, the optimal M is estimable in the sample since the variance of H_T is some weighted average of the errors.

(But heteroskedastic errors in either case are always estimable. That is, even if I stick in non-optimal GLS weighting matrix, I can always get consistent estimates of the heteroskedastic-robust standard errors)

III. Lt, lt, St

a. For Single Equation Estimators

	$Q_T(\beta)$	$L_T(\beta)$	$S_T(\beta)$	Avar $\sqrt{T}L_T(\beta_0) = V(\beta_0)$	Plim $S_T(\beta_0)$
OLS	$\frac{1}{T}\sum_t \left(y_t - x_t'\beta\right)^2$	$\frac{1}{T}\sum_t -2x_t\left(y_t - x_t'\beta\right)$	$\frac{1}{T}\sum_t 2x_t x_t'$	$4E(\varepsilon_t^2 x_t x_t')$	$2E(x_t x_t')$
NLLS (NL in x's)	$\frac{1}{T}\sum_t \left(y_t - f_t(x_t,\beta)\right)^2$	$\frac{1}{T}\sum_t -2\frac{\partial f_t}{\partial \beta'}\left(y_t - f_t(x_t,\beta)\right)$	$\frac{1}{T}\sum_t 2\frac{\partial^2 f_t}{\partial\beta\partial\beta'}(y_t - f_t)$ $+\frac{1}{T}\sum_t 2\frac{\partial f_t}{\partial\beta}\frac{\partial f_t}{\partial\beta'}$	$4E\left(\varepsilon_t^2 \frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)$	$2E\left(\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)$
NLLS (NL in x's and y's) $f_t = e_t$	$\frac{1}{T}\sum_t \left(f(y_t,x_t,\beta)\right)^2$	$\frac{1}{T}\sum_t 2\frac{\partial f}{\partial\beta'} f(y_t,x_t,\beta)$	$\frac{1}{T}\sum_t 2\frac{\partial^2 f_t}{\partial\beta\partial\beta'} f_t$ $+\frac{1}{T}\sum_t 2\frac{\partial f_t}{\partial\beta}\frac{\partial f_t}{\partial\beta'}$	$4E\left(\begin{matrix} f_0(y_t,x_t,\beta_0)^2 * \\ \frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta'}\end{matrix}\right)$	$2E\left(\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta'}\right)$
GLS	$\frac{1}{T}e'Me$ $e = \left(y_1 - x_1'\beta,\ldots,y_T - x_T'\beta\right)$	$\frac{2}{T}\frac{\partial e}{\partial\beta'}Me$	$\frac{2}{T}\left(\frac{\partial^2 e}{\partial\beta\partial\beta'}Me\right)+\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}M\frac{\partial e}{\partial\beta}\right)$	$p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}Mee'M'\frac{\partial e}{\partial\beta}$ $= p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}\Phi^{-1}\Phi\Phi^{-1}\frac{\partial e}{\partial\beta}$ $= p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}\Phi^{-1}\frac{\partial e}{\partial\beta}$	$p\lim\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}M\frac{\partial e}{\partial\beta}\right)$ $p\lim\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}\Phi^{-1}\frac{\partial e}{\partial\beta}\right)$
NGLS	$\frac{1}{T}e'Me$ $e = \left(y_1 - f_1,\ldots,y_T - f_1\right)$	$\frac{2}{T}\frac{\partial e}{\partial\beta'}Me$	$\frac{2}{T}\left(\frac{\partial^2 e}{\partial\beta\partial\beta'}Me\right)+\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}M\frac{\partial e}{\partial\beta}\right)$	$p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}Mee'M'\frac{\partial e}{\partial\beta}$ $= p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}\Phi^{-1}\Phi\Phi^{-1}\frac{\partial e}{\partial\beta}$ $= p\lim\frac{4}{T}\frac{\partial e}{\partial\beta'}\Phi^{-1}\frac{\partial e}{\partial\beta}$	$p\lim\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}M\frac{\partial e}{\partial\beta}\right)$ $p\lim\frac{2}{T}\left(\frac{\partial e}{\partial\beta'}\Phi^{-1}\frac{\partial e}{\partial\beta}\right)$

	$Q_T(\beta)$	$L_T(\beta)$	$S_T(\beta)$	Avar $\sqrt{T}L_T(\beta_0) = V(\beta_0)$	Plim $S_T(\beta_0)$
2SLS GMM With Weight Matrix W'W^{-1}	$H_T'\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}H_T =$ $\left[\frac{1}{T}W'(Y-Z\beta)\right]'\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}$ $\left[\frac{1}{T}W'(Y-Z\beta)\right]$	$2\frac{\partial H_t}{\partial\beta'}\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}H_T =$ $-\frac{2}{T}Z'W\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}\left[\frac{1}{T}W'(Y-Z\beta)\right] =$ $\frac{1}{T}\sum_t -2\left(\frac{1}{T}Z'W\right)\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}w_t\left(y_t - z_t'\beta\right)$	(ignoring first term, which goes to 0) $2\frac{\partial H_t}{\partial\beta'}\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}\frac{\partial H_t}{\partial\beta}$ $=-\frac{2}{T}Z'W\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}\frac{1}{T}W'Z$ $=\frac{1}{T}\sum_t -2\left(\frac{1}{T}Z'W\right)\left(\frac{\hat\sigma^2}{T}W'W\right)^{-1}w_t z_t'$	$\frac{4}{p\lim\hat\sigma^2}\Sigma_{wz}'\Sigma_{ww}^{-1}E\left(\varepsilon_t^2 w_t w_t'\right)\Sigma_{wz}\Sigma_{ww}^{-1}\Sigma_{wz}$	$-2\Sigma_{wz}'\Sigma_{ww}^{-1}\Sigma_{wz}$
N2SLS (NL in x's)	$\frac{1}{T}H_T'(W'W)^{-1}H_T =$ $\frac{1}{T}\left[W'(Y-f(X,\beta))\right]'(W'W)^{-1}$ $\left[W'(Y-f(X,\beta))\right]$	$2\frac{\partial H_t}{\partial\beta'}\left(\frac{1}{T}W'W\right)^{-1}H_T =$ $-2\left[\frac{1}{T}W'\frac{\partial f}{\partial\beta}\right]'\left(\frac{1}{T}W'W\right)^{-1}\left[W'(Y-Z\beta)\right] =$ $\frac{1}{T}\sum_t -2G(\beta)\left(\frac{1}{T}W'W\right)^{-1}w_t\left(y_t - f_t(\beta)\right)$	(ignoring first term, which goes to 0) $2\frac{\partial H_t}{\partial\beta'}\left(\frac{1}{T}W'W\right)^{-1}\frac{\partial H_t}{\partial\beta}$ $=-2G(\beta)'\left(\frac{1}{T}W'W\right)^{-1}G(\beta)$	$4G_0ME\left(\varepsilon_t^2 w_t w_t'\right)MG_0'$	$-2G_0'MG_0 =$ $-2G_0'E(w_t w_t')^{-1}G_0$
N2SLS (NL in x's)	$\frac{1}{T}H_T'(W'W)^{-1}H_T =$ $\left[W'f(Y,X,\beta)\right]'\left(\frac{1}{T}W'W\right)^{-1}$ $\left[W'f(Y,X,\beta)\right]$	$2\frac{\partial H_t}{\partial\beta'}\left(\frac{1}{T}W'W\right)^{-1}H_T =$ $-2\left[\frac{1}{T}W'\frac{\partial f}{\partial\beta}\right]'\left(\frac{1}{T}W'W\right)^{-1}\left[W'(Y-Z\beta)\right] =$ $\frac{1}{T}\sum_t -2G(\beta)\left(\frac{1}{T}W'W\right)^{-1}w_t\left(y_t - f_t(\beta)\right)$	(ignoring first term, which goes to 0) $2\frac{\partial H_t}{\partial\beta'}\left(\frac{1}{T}W'W\right)^{-1}\frac{\partial H_t}{\partial\beta}$ $=-2G(\beta)'\left(\frac{1}{T}W'W\right)^{-1}G(\beta)$	$4G_0E(w_t w_t')^{-1}E\left(\varepsilon_t^2 w_t w_t'\right)$ $E(w_t w_t')^{-1}G_0'$	$-2G_0'E(w_t w_t')^{-1}G_0$

b. Multiple Equation Estimators

2 Representations

1. $Y = X\delta + U$: $\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} = \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_M \end{bmatrix}$ $\quad U \sim \left(0, \Omega_{MxM} \otimes I_T\right)$

$$Z = \begin{bmatrix} \vec{z}'_{1(1xK)} \\ \vdots \\ \vec{z}'_{n(1xK)} \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nK} \end{bmatrix}_{TxK} \quad X = \begin{bmatrix} X_{1(TxD_1)} & & \\ & \ddots & \\ & & X_{M(TxD_M)} \end{bmatrix}_{TMx\sum_m D_m} \quad X_m = \begin{bmatrix} \vec{x}'_m(1) \\ \vdots \\ \vec{x}_m(T) \end{bmatrix}_{TxD_m} \quad \delta_{\sum_m K_m} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_M \end{bmatrix} \quad Y = \begin{bmatrix} y_{1(Tx1)} \\ \vdots \\ y_{M(Tx1)} \end{bmatrix}_{TMx1} \quad y_m = \begin{bmatrix} y_m(1) \\ \vdots \\ y_m(T) \end{bmatrix}_{Tx1} \quad u = \begin{bmatrix} u_{1(Tx1)} \\ \vdots \\ u_{M(Tx1)} \end{bmatrix}_{TMx1} \quad u_m = \begin{bmatrix} u_m(1) \\ \vdots \\ u_m(T) \end{bmatrix}_{Tx1}$$

2. Contemporaneous Form (each data point is a vector):

$$y(t) = \underline{x}(t)'\delta + u(t) \;:\; \begin{pmatrix} y_1(t) \\ \vdots \\ y_M(t) \end{pmatrix} = \begin{pmatrix} x_1(t)' & & \\ & \ddots & \\ & & x_M(t)' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix} + \begin{pmatrix} u_1(t) \\ \vdots \\ u_M(t) \end{pmatrix} \quad \textit{for } t = 1,...,T \qquad u(t) \sim \left(0, \Omega_{MxM}\right)$$

(For 2SLS/3SLS) $\underline{w}(t) = \begin{pmatrix} w_1(t)_{k_1x1} & & \\ & \ddots & \\ & & w_M(t)_{k_Mx1} \end{pmatrix} = \left(I_M \otimes w(t)\right)$ *if same instruments across equations* $\frac{1}{T}W'W = \frac{1}{T}\sum w(t)w(t)'$

Note on Var-Cov Matrix: These are contemporaneous cross-equation covariances: **homoskedasticity is ACROSS observations. There can be heteroskedasticity across equations** (e.g. $\phi_{11}, \phi_{22}, \phi_{33}$ can all be different)

$$E(u(t)u(t)' \mid \underline{x}(t)) = \Omega_{3x3} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix}, \qquad E(UU' \mid X) = \left(\Omega_{3x3} \otimes I_T\right) = \begin{pmatrix} \phi_{11} & & & \phi_{12} & & & \phi_{13} & & \\ & \ddots & & & \ddots & & & \ddots & \\ & & \phi_{11} & & & \phi_{12} & & & \phi_{13} \\ \phi_{12} & & & \phi_{22} & & & \phi_{13} & & \\ & \ddots & & & \ddots & & & \ddots & \\ & & \phi_{12} & & & \phi_{22} & & & \phi_{13} \\ \phi_{13} & & & \phi_{23} & & & \phi_{33} & & \\ & \ddots & & & \ddots & & & \ddots & \\ & & \phi_{13} & & & \phi_{23} & & & \phi_{33} \end{pmatrix}$$

	$Q_T(\beta)$	$L_T(\beta)$	$S_T(\beta)$	Avar $\sqrt{T}L_T(\beta_0) = V(\beta_0)$	Plim $S_T(\beta_0)$
OLS	$\frac{1}{T}U'U = \frac{1}{T}\sum_t u(t)'u(t)$	$\frac{1}{T}\sum_t -2\underline{x}(t)\left(y(t) - \underline{x}'(t)\delta\right)$	$\frac{2}{T}\sum_t \underline{x}(t)\underline{x}'(t)$	$4E(\underline{x}(t)\Omega\underline{x}(t)')$	$2E\left(\underline{x}(t)\underline{x}(t)'\right)$
SUR/ JGLS **(QML) (assuming homo across obs)**	$\frac{1}{T}(Y - X\beta)'\left(\hat{\Omega}^{-1}\otimes I\right)(Y - X\beta) =$ $\frac{1}{T}\sum_t \left(y(t) - \underline{x}(t)'\beta\right)'\hat{\Omega}^{-1}\left(y(t) - \underline{x}(t)'\beta\right)$	$\frac{-2}{T}\sum_t \underline{x}(t)\hat{\Omega}^{-1}_{MxM}\left(y(t) - \underline{x}(t)'\beta\right)$	$\frac{2}{T}\sum_t \underline{x}(t)\hat{\Omega}^{-1}_{MxM}\underline{x}(t)' =$ $\frac{2}{T}X'\left(\hat{\Omega}^{-1}\otimes I_T\right)X$	$4E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)$	$2E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)$ $=_A \frac{2}{T}X'\left(\hat{\Omega}\otimes I_T\right)X$
NLSUR/ NJGLS (assuming homo across obs)	$\frac{1}{T}\sum_t \left(y(t) - f(\underline{x}(t),\beta)\right)'\hat{\Omega}^{-1}\left(y(t) - f(\underline{x}(t),\beta)\right)$	$\frac{-2}{T}\sum_t \frac{\partial f_t}{\partial\beta'}\hat{\Omega}^{-1}\left(y(t) - f(\underline{x}(t),\beta)\right)$	$\frac{2}{T}\sum_t \frac{\partial f_t(x(t),\beta)}{\partial\beta'}\hat{\Omega}^{-1}\frac{\partial f_t(x(t),\beta)}{\partial\beta}$	$4E\frac{\partial f(x(t),\beta_0)}{\partial\beta'}\Omega^{-1}\frac{\partial f(x(t),\beta_0)}{\partial\beta'}$	$2E\frac{\partial f(x(t),\beta_0)}{\partial\beta'}\Omega^{-1}\frac{\partial f(x(t),\beta_0)}{\partial\beta'}$
MLE (no endog. Y's only endog vars)	$\frac{1}{T}\sum \log f(y_t)$ $f(y_t) = (2\pi)^{-m/2}\lvert\Omega\rvert^{-1/2}$ $\exp\left(-\frac{1}{2}(y_t - X_t'\beta)'\Omega^{-1}(y_t - X_t'\beta)\right)$	$\frac{\partial Q_T}{\partial(\beta,\Omega)} = \frac{1}{T}\sum l_t$			
ME 2SLS NOTE: HERE, W is just nxK, it's the NORMAL form	$H_T'\left(I_M\otimes\frac{1}{T}W'W\right)^{-1}H_T =$ $\left[\frac{1}{T}(I_M\otimes W')(Y - X\beta)\right]'\left(I_M\otimes\left(\frac{1}{T}W'W\right)^{-1}\right)$ $\left[\frac{1}{T}(I_M\otimes W')(Y - X\beta)\right] =$ $\boxed{\frac{1}{T}(Y - X\beta)'\left(I_M\otimes W\left(W'W\right)^{-1}W\right)(Y - X\beta)} =$ $\left[\frac{1}{T}\sum \underline{w}(t)\left(y(t) - \underline{x}(t)\beta\right)\right]'\left(\begin{matrix} I_M\otimes \\ \left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\end{matrix}\right)$ $\left[\frac{1}{T}\sum \underline{w}(t)\left(y(t) - \underline{x}(t)\beta\right)\right]$	$\frac{-2}{T}X'(I_M\otimes P_W)(Y - X\beta) =$ $-2\left[\frac{1}{T}\sum \underline{w}(t)\underline{x}(t)'\right]'$ $\left(I_M\otimes\left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\right)$ $\left[\frac{1}{T}\sum \underline{w}(t)\left(y(t) - \underline{x}(t)'\beta\right)\right]$	$\frac{2}{T}X'(I_M\otimes P_W)X =$ $2\left[\frac{1}{T}\sum \underline{w}(t)\underline{x}(t)'\right]'$ $\left(I_M\otimes\left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\right)$ $\left[\frac{1}{T}\sum \underline{w}(t)\underline{x}(t)'\right]$	$4E\left[\underline{w}(t)\underline{x}(t)'\right]$ $\left(I_M\otimes E\left(\underline{w}(t)\underline{w}(t)'\right)^{-1}\right)$ $\left(\Omega\otimes E(w(t)w(t)')\right)$ $\left(I_M\otimes E\left(\underline{w}(t)\underline{w}(t)'\right)^{-1}\right)$ $E\left[\underline{w}(t)\underline{x}(t)'\right] =$ $4E\left[\underline{w}(t)\underline{x}(t)'\right]$ $\left(\Omega\otimes E(w(t)w(t)')^{-1}\right)$ $E\left[\underline{w}(t)\underline{x}(t)'\right]$ (*) Avar wont simplify unless $\Omega = \sigma^2$ So, need robust s.e. unless we have conditional homosk.	$2E\left(\underline{w}(t)\underline{x}(t)'\right)'$ $\left(I_M\otimes E\left(w(t)w(t)'\right)^{-1}\right)$ $E\left(\underline{w}(t)\underline{x}(t)'\right)$

	$Q_T(\beta)$	$L_T(\beta)$	$S_T(\beta)$	Avar $\sqrt{T}L_T(\beta_0) = V(\beta_0)$	Plim $S_T(\beta_0)$
ME 3SLS	$Q_T = H_T'\left(\hat{\Omega}\otimes\frac{1}{T}W'W\right)^{-1}H_T$ $\left[\frac{1}{T}(I_M\otimes W')(Y-Z\beta)\right]'\left(\hat{\Omega}^{-1}\otimes\left(\frac{1}{T}W'W\right)^{-1}\right)$ $\left[\frac{1}{T}(I_M\otimes W')(Y-Z\beta)\right] =$ $\boxed{\frac{1}{T}(Y-X\beta)'\left(\hat{\Omega}^{-1}\otimes W(W'W)^{-1}W\right)(Y-X\beta)} =$ $\left[\frac{1}{T}\sum_t \underline{w}(t)\left(y(t)-\underline{x}(t)'\beta\right)\right]'\left(\begin{matrix}\hat{\Omega}^{-1}\otimes\\ \left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\end{matrix}\right)$ $\left[\frac{1}{T}\sum_t \underline{w}(t)\left(y(t)-\underline{x}(t)'\beta\right)\right]$	$-\frac{2}{T}X'\left(\hat{\Omega}^{-1}\otimes P_W\right)(Y-X\beta)$ $-2\left[\frac{1}{T}\sum_t \underline{w}(t)\underline{x}(t)'\right]'\left(\begin{matrix}\hat{\Omega}^{-1}\otimes\\ \left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\end{matrix}\right)$ $\left[\frac{1}{T}\sum_t \underline{w}(t)\left(y(t)-\underline{x}(t)'\beta\right)\right]$	$\frac{2}{T}X'\left(\hat{\Omega}^{-1}\otimes P_W\right)X =$ $2\left[\frac{1}{T}\sum_t \underline{w}(t)\underline{x}(t)'\right]'\left(\begin{matrix}\hat{\Omega}^{-1}\otimes\\ \left(\frac{1}{T}\sum w(t)w(t)'\right)^{-1}\end{matrix}\right)$ $\left[\frac{1}{T}\sum_t \underline{w}(t)\underline{x}(t)'\right]$	$-4E\left(\underline{w}(t)\underline{x}(t)'\right)'$ $\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)$ $E(\underline{w}(t)\underline{x}(t)')$ (see below for algebra steps (*))	$-2E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E(w(t)w(t)')^{-1}\right)$ $E\left(\underline{w}(t)\underline{x}(t)'\right)$

(*)Algebra for 3SLS/3SLS V(β_0): $\frac{1}{T}\sum_t \underline{w}(t)e(t)e(t)'\underline{w}(t)' \Rightarrow_P E\left[(1\otimes\underline{w}(t))(e(t)\otimes 1)(e(t)'\otimes 1)(1\otimes\underline{w}(t)')\right] = E\left[(e(t)e(t)'\otimes\underline{w}(t)\underline{w}(t)')\right] = (\Omega\otimes E(\underline{w}(t)\underline{w}(t)'))$

Note: Moment conditions in systems of equations are: $\begin{bmatrix} E\left(W_{kxT}'U_{1(Tx1)}\right)\\ \vdots\\ E\left(W_{kxT}'U_{m(Tx1)}\right)\end{bmatrix} = E\left[(I_m\otimes Z')_{mk\times mT}U_{mT\times 1}\right] = 0$ *or* $\begin{bmatrix} E\left(W_{kx1}(t)u_1(t)_{1x1}\right)\\ \vdots\\ E\left(W_{kx1}(t)u_m(t)_{1x1}\right)\end{bmatrix}_{mkx1} = E\left[(I_m\otimes W(t))_{mk\times m}U(t)_{m\times 1}\right]_{nkx1} = 0$

Thus, by analogy principle, sample moment conditions are: $\frac{1}{T}\sum_t^T (I_m\otimes W(t))_{mk\times m}\left[Y(t)_{mx1}-\underline{x}(t)'\delta\right]_{m\times 1}$

"Efficient" weighting matrix is variance of moment conditions, i.e.

$$E\left[(I_m\otimes W(t))_{mk\times m}U(t)_{m\times 1}U(t)_{1xm}'\left(I_m\otimes W(t)'\right)_{mxmk}\right]^{-1}_{mkxmk} = E\left[(I_m\otimes W(t))_{mk\times m}\Omega_{mxm}\left(I_m\otimes W(t)'\right)_{mxmk}\right]^{-1}_{mkxmk}$$

$$= E\left[(I_m\otimes W(t))_{mk\times m}(\Omega_{mxm}\otimes 1)\left(I_m\otimes W(t)'\right)_{mxmk}\right]^{-1}_{mkxmk} = E\left[\left(\Omega\otimes W(t)W(t)'\right)\right]^{-1}_{mkxmk} = \boxed{E\left(\Omega^{-1}\otimes\left[W(t)W(t)'\right]^{-1}\right)_{mkxmk}}$$

Omega = I, then $M_{2SLS} = \left(I_m\otimes\left(\frac{1}{T}W'W\right)^{-1}\right)$ *is efficient GMM and asymptotically equivalent to* 3*SLS*

Thus, if

Generally, 3*SLS is efficient GMM :* $M_{3SLS} = \left(\hat{\Omega}\otimes\left(\frac{1}{T}W'W\right)^{-1}\right)$

Note: $S_T\left(\hat{\beta}\right)$ does not have the 2 in it, because $L_T\left(\hat{\beta}\right)=0 \Rightarrow \frac{1}{T}\sum_t -2x_t\left(y_t - x_t'\beta\right)=0 \Leftrightarrow \frac{1}{T}\sum_t -x_t\left(y_t - x_t'\beta\right)=0$

So, the one that we USE for estimation never has the 2 or 4 attached in front.

IV. Avar and its sample Estimate

Estimator	Avar = $S_0^{-1}VS_0^{-1}$ (Robust)	Sample Var Estimate $S_T\left(\hat{\beta}\right)^{-1}\hat{V}_T\left(\hat{\beta}\right)S_T\left(\hat{\beta}\right)^{-1}$ (Robust)	Avar $\pm\alpha S_0^{-1}$ (Homoskdas)	Sample Var Est. $S_T\left(\hat{\beta}\right)^{-1}\hat{V}_T\left(\hat{\beta}\right)S_T\left(\hat{\beta}\right)^{-1}$ (Robust)
OLS	$E\left(x_t x_t'\right)^{-1}E\left(\varepsilon^2 x_t x_t'\right)E\left(x_t x_t'\right)^{-1}$	$\frac{1}{T}\left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{T}X'\hat{\varepsilon}\hat{\varepsilon}'X\right)\left(\frac{1}{T}X'X\right)^{-1}$	$\sigma^2 E\left(x_t x_t'\right)^{-1}$	$\frac{1}{T}\hat{\sigma}^2\left(\frac{1}{T}X'X\right)^{-1}$
NLLS (parameters nonlinear in x's)	$E\left(\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)^{-1}E\left(\varepsilon_t^2\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)E\left(\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)^{-1}$	$\frac{1}{T}\left(\frac{1}{T}\sum_t\frac{\partial f_t(\hat{\beta})}{\partial\beta}\frac{\partial f_t(\hat{\beta})}{\partial\beta'}\right)^{-1}\left(\frac{1}{T}\sum_t\hat{\varepsilon}_t^2\frac{\partial f_t(\hat{\beta})}{\partial\beta}\frac{\partial f_t(\hat{\beta})}{\partial\beta'}\right)\left(\frac{1}{T}\sum_t\frac{\partial f_t(\hat{\beta})}{\partial\beta}\frac{\partial f_t(\hat{\beta})}{\partial\beta'}\right)^{-1}$	$\sigma^2 E\left(\frac{\partial f(\beta_0)}{\partial\beta}\frac{\partial f(\beta_0)}{\partial\beta}'\right)^{-1}$	$\hat{\sigma}^2\frac{1}{T}\sum_t\left(\frac{\partial f(\hat{\beta})}{\partial\beta}\frac{\partial f(\hat{\beta})}{\partial\beta'}\right)^{-1}$
NLLS (parameters nonlinear in y'x and x's. only in e's)	$E\left(\frac{\partial f}{\partial\beta}\frac{\partial f}{\partial\beta'}\right)^{-1}E\left(f_0(y_t,x_t,\beta_0)^2\frac{\partial f}{\partial\beta}\frac{\partial f}{\partial\beta'}\right)E\left(\frac{\partial f}{\partial\beta}\frac{\partial f}{\partial\beta'}\right)^{-1}$	$\frac{1}{T}\sum_t 2\left(\frac{\partial f}{\partial\beta}\right)' f(y_t,x_t,\beta)$	$\sigma^2 E\left(\frac{\partial f}{\partial\beta}\frac{\partial f}{\partial\beta'}\right)^{-1}$	$\hat{\sigma}^2\frac{1}{T}\left(\sum_t\frac{\partial f}{\partial\beta}\frac{\partial f}{\partial\beta'}\right)^{-1}$
2SLS	$\left(G_0'MG_0\right)^{-1}G_0'ME(\varepsilon_t^2 w_t w_t')$ $MG_0\left(G_0'MG_0\right)^{-1}=$ $\left(\Sigma_{WZ}'\Sigma_{WW}\Sigma_{WZ}\right)^{-1}\Sigma_{WZ}'\Sigma_{WW}E(\varepsilon_t^2 w_t w_t')$ $\Sigma_{WW}\Sigma_{WZ}\left(\Sigma_{WZ}'\Sigma_{WW}\Sigma_{WZ}'\right)^{-1}$ *Here,* $G_0=E(w_t'z_t)=\Sigma_{WZ}$	$\left(S_{WZ}'S_{WW}S_{WZ}\right)^{-1}S_{WZ}'S_{WW}\hat{S}$ $S_{WW}S_{WZ}\left(S_{WZ}'S_{WW}S_{WZ}'\right)^{-1}$ *Here,* $\hat{S}=\frac{1}{T}\sum_t\hat{\varepsilon}_t^2 w_t w_t'$	$\sigma^2\left(G_0'MG_0\right)^{-1}=$ $\sigma^2\left(\Sigma_{WZ}'\Sigma_{WW}\Sigma_{WZ}\right)^{-1}$ *Here,* $G_0=E(w_t'z_t)=\Sigma_{WZ}$	$\hat{\sigma}^2\left(S_{WZ}'S_{WW}S_{WZ}\right)^{-1}$
N2SLS	$\left(G_0E(w_t w_t')^{-1}G_0'\right)^{-1}G_0E(w_t w_t')^{-1}$ $E\left(\varepsilon_t^2 w_t w_t'\right)$ $E(w_t w_t')^{-1}G_0'\left(G_0E(w_t w_t')^{-1}G_0'\right)^{-1}$	$\left(\hat{G}S_{WW}^{-1}\hat{G}'\right)^{-1}\hat{G}S_{WW}^{-1}\hat{S}S_{WW}^{-1}\hat{G}'$ $\left(\hat{G}S_{WW}^{-1}\hat{G}'\right)^{-1}$	$\sigma^2\left(G_0E(w_t w_t')^{-1}G_0'\right)^{-1}$	$\hat{\sigma}^2\left(\hat{G}S_{WW}^{-1}\hat{G}'\right)^{-1}$

a. <u>When is heteroskedastic variance/covariance matrix not consistently estimable?</u>

OLS/NLS: If heteroskedastic, we can estimate the robust var-cov matrix because it is a weighted average of the variance of each e(t).

GLS: For "optimal" weighting matrix, need to make assumptions about the functional form of V. Cannot estimate each e(t) since we only have 1 observation.
If weighing matrix is wrong, then we still have consistency, but again would need to use robust standard errors.

GMM/2SLS: Can estimate the variance of the moment conditions bc it is a weighted average of the heteroskedastic errors.

ML: Does not incorporate heteroskedasticity; in principle we can include it in the maximization problem, but then becomes infeasible.

SUR/JGLS: (when there's heteroskedasticity across observations) Cannot estimate the efficient weighting matrix feasibly (this is like the GLS equivalent in the ME setup). Need to make functional form assumptions about the form of the heteroskedasticity.
If the weighting matrix is wrong (for example, if we assume conditional homoskedasticity across observations), then consistency is still ok. But to get correct standard errors we need to use ROBUST standard errors.

3SLS: Always estimable for the same reasons why GMM/2SLS works.

FIML/LIML: Cannot handle heteroskedasticity across observations/does not incorporate it. In principle we can include it in the maximization problem, but have to program it yourself.

Note: Clearly, homoskedastic var/covariance matrices are always consistently estimable (under the normal assumptions)

V. Estimators: Explicit Forms, Efficiency, and Comparison

Estimator	Form		Estimator	Form
OLS	$\beta_{OLS} = (X'X)^{-1}X'Y$		**OLS**	$\beta_{OLS} = (X'X)^{-1}X'Y$
NLLS (parameters nonlinear in x's)	No Explicit Form		**SUR/JGLS**	$\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat{S}^{-1}s_{ZY}$ $=\left[X'\left(\hat{\Sigma}^{-1}\otimes I_n\right)X\right]^{-1}X'\left(\hat{\Sigma}^{-1}\otimes I_n\right)Y$
NLLS (parameters nonlinear in y'x and x's. only in e's)	No Explicit Form		**2SLS**	$\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat{S}^{-1}s_{ZY}$ $=\left[X'(I\otimes P_Z)X\right]^{-1}X'(I\otimes P_Z)Y$
GLS	$\beta_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y$		**3SLS**	$\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat{S}^{-1}s_{ZY}$ $=\left[X'\left(\hat{\Sigma}^{-1}\otimes P_Z\right)X\right]^{-1}X'\left(\hat{\Sigma}^{-1}\otimes P_Z\right)Y$
2SLS	$\delta_{2SLS} = \left(S_{ZX}'S_{ZZ}^{-1}S_{ZX}\right)^{-1}S_{ZX}'S_{ZZ}^{-1}s_{ZY}$			
N2SLS	No Explicit Form			
IV	$\delta_{IV} = S_{ZX}^{-1}s_{ZY}$			
NLIV	No Explicit Form			

VI. Identification Condition for estimators: $p\lim S(\tilde{\theta}) = p\lim S(\theta_0) = p\lim \frac{\partial L_T(\theta_0)}{\partial \theta'}$ **Invertible**

a. Least Squares

$$p\lim S(\theta_0) = p\lim\left(\frac{1}{T}\sum_t 2x_t x_t'\right) \text{ invertible} \Leftrightarrow \boxed{E(x_t x_t')} \text{ invertible}$$

b. NLS

$$p\lim S(\theta_0) = 2E\left(\frac{\partial f(\beta_0)}{\partial \beta}\frac{\partial f(\beta_0)}{\partial \beta}'\right) \text{ invertible} \Leftrightarrow \boxed{E\left(\frac{\partial f(\beta_0)}{\partial \beta}\frac{\partial f(\beta_0)}{\partial \beta}'\right)} \text{ invertible}$$

c. 2SLS (this condition is true for GMM estimators in general)

$p\lim S(\theta_0) = -2\Sigma_{wz}'\Sigma_{ww}^{-1}\Sigma_{wz}$ *invertible* $\Leftrightarrow \Sigma_{wz} = \boxed{E(w_t z_t')}$ *is full column rank* (*b.c. # of regressors* ≤ *# instruments*) (rank condition)

d. NL2SLS

$$p\lim S(\theta_0) = -2G_0'E(w_t w_t')^{-1}G_0 \text{ invertible} \Leftrightarrow G_0 = E\left(\frac{\partial H_t(\beta_0)}{\partial \beta}\right) \text{ is full column rank (rank condition)}$$

e. Identification for System of Equations

Typically we use the fact that each equation is identified.

But, each equation can be not identified individually, but as a system if we have cross equation restrictions we can identify the parameters. (see 271A PS3 Ch4 An. Ex #5.

f. Another (equivalent) way to think about it: Are population moment conditions uniquely satisfied?

$P\lim L_T(\beta) \neq P\lim L_T(\delta) \;\; \forall \beta \neq \delta$

Thus, for OLS, $E\left[x_t'\left(y_t - x_t'\beta\right)\right] \neq E\left[x_t'\left(y_t - x_t'\delta\right)\right] \Leftrightarrow E\left(x_t x_t'\right)(\beta - \delta) \neq 0 \;\; for\ \beta \neq \delta \Leftrightarrow E\left(x_t x_t'\right)$ in*vertible*

- Check: Are there multiple parameter values that can minimize the distance function/ give us the same answers?
- Check: Are there transformations of the parameters such that we can isolate each one?

Nonlinear Examples:

1. $y_t = \beta\gamma + \gamma^2 z_t + u_t$: β, γ not identified. We can't identify the sign of g and therefore not the sign of b. Thus, if $(\hat{\beta}, \hat{\gamma})$ minimizes SSR of the model, so does $(-\hat{\beta}, -\hat{\gamma})$
2. $x_t = \beta_1 + \beta_2 z_t^{\beta_3}$: β_3 not identified if $\beta_2 = 0$, because in this case any β_3 can minimize the SSR.
   β_1, β_2 not identified since then the model becomes $x_t = \beta_1 + \beta_2$, so we can't distinguish between the 2 coefficients in the constant term.

VII. Comparison of Estimators

First: In order to compare efficiency, we must have the **SAME MODEL**! Otherwise the comparison does not make sense.

a. Principle of GLS

$\hat{\theta} \sim_A N\left(\theta_0, \frac{1}{T}\hat{S}\left(\hat{\theta}\right)^{-1}\hat{V}\left(\hat{\theta}\right)\hat{S}\left(\hat{\theta}\right)^{-1}\right)$. GLS is about how to optimally pick M s.t. $\frac{1}{T}\hat{S}\left(\hat{\theta}\right)^{-1}\hat{V}\left(\hat{\theta}\right)\hat{S}\left(\hat{\theta}\right)^{-1}$ reduces to $\alpha S\left(\theta_0\right)^{-1}$

For least squares estimators defined by $Q_T = e'Me \;\; e \sim (0,\Phi), \;\; pick \;\; M = \alpha\Phi^{-1}$ (Gauss Markov – this is FINITE sample variance)

For estimator defined by $Q_T = \left(v'e\right)' M \left(v'e\right) = H_T M H_T \;\; \sqrt{T}H_T \Rightarrow (0,\Omega), \;\; pick \;\; M = \alpha\Omega^{-1}$ (Efficient GMM – this is asymptotic variance)

(Note: If you are an "optimal" LS estimator in finite sample, i.e. for any n, you must also be asymptotically efficient, i.e. for large n)

Example:

i. **OLS**: If we have
   1. Linearity : y = xb + e
   2. Exogeneity $E(e_t \mid X) = 0$
   3. No Multicollinearity
   4. Conditional Homoskedasticity
   5. No correlation between observations

Then, OLS is **BLUE by Gauss Markov** (which says that if above assumptions hold, OLS is BLUE, and by below, it's the "optimal" GLS.)

In this case, $Y = X'\beta + \varepsilon, \;\; \varepsilon \sim (0,\sigma^2 I)$ and I proportional to inverse of variance covariance matrix, thus optimal estimator defined by **Q_T = e'Ie = e'e**

ii. **GLS**: If we have
   1. Linearity : y = xb + e
   2. Exogeneity $E(e_t \mid X) = 0$
   3. No Multicollinearity

   But…

   Conditional heteroskedasticity and/or Autocorrelated errors

In this case, $Y = X'\beta + \varepsilon, \;\; \varepsilon \sim (0,\Phi)$ **, then pick M = $\alpha\Phi^{-1}$**

iii. **GMM**: If we have

$Q_T = \left(v'e\right)' M \left(v'e\right) = H_T M H_T \;\; \sqrt{T}H_T \Rightarrow (0,\Omega), \;\; pick \;\; M \Rightarrow_P \alpha\Omega^{-1}$ **(Inverse of variance of moment conditions)**

b. Which estimators are desired, given both consistent?

Want OLS/GLS over 2SLS because by Gauss-Markov, they are BLUE.
Want NLS over N2SLS
Want ME3SLS over ME2SLS (provided that they are not the same)
Want JGLS/SUR over 3SLS and 2SLS by Gauss Markov.

Note: Generally, the "GLS" estimator is preferred to "IV" estimator if we do not have endogeneity problems, so that both are consistent (b.c. of Gauss-Markov, GLS is most efficient in its class).

Intuition: In 2SLS, we predict the endogenous variable based on a set of instruments. This prediction is noisy. But the "BEST" set of instrument to predict the variable is the variable itself. So if the variable is exogenous, then OLS most efficient! No reduction of dimensionality necessary in the first step.

c. Which estimator is more efficient than which? When?

- **OLS vs. GLS:**
  - Under conditional homoskedasticity and no autocorrelation, OLS is BLUE.
  - Under conditional heteroskedasticity and/or serial correlation, GLS is BLUE ($\varepsilon \sim (0,\Phi)$, *then* $M = \Phi^{-1}$).

Multiple Equation (assuming no cross equation restrictions)

- **SUR/JGLS vs. OLS:**
  SUR is generally more efficient than OLS (multiple equation) UNLESS, **GIVEN NO CROSS EQUATION RESTRICTIONS**:
  i. The variance-covariance matrix of errors for equations $\Omega = E(u(t)u(t)')$ is diagonal: i.e. $Cov(e_j(t), e_i(t)) = 0 \ \forall \ i \neq j$ → equations are not correlated!
  ii. $\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}_3 = \ldots$ (i.e. all regressors are the same) → model is just identified (See 272PS3 #1(i))
  (in this case, both estimators are just multivariate regression which is just equation by equation OLS)
  (unless there are restrictions within or across equations, in which case, LS would not take them into account and SUR would, and estimators would be different and SUR more efficient.) (HW3#1(i))

- **2SLS vs. 3SLS:**
  3SLS is generally more efficient (asymptotically) than 2SLS UNLESS (GIVEN NO CROSS EQUATION RESTRICTIONS):
  i. The variance-covariance matrix of errors for equations $\Phi = E(u(t)u(t)')$ is diagonal, i.e. $Cov(e_j(t), e_i(t)) = 0 \ \forall \ i \neq j$ → equations are not correlated!
  ii. All equations are just identified: Then, both estimators are both defined by the solution to the system of equations, i.e. equation by equation I.V.:

$$H_T(\delta) = \frac{1}{T}(I \otimes W')(Y - X\delta) = \begin{bmatrix} \frac{1}{T}W'(y_1 - X_1\delta_1) \\ \vdots \\ \frac{1}{T}W'(y_M - X_M\delta_M) \end{bmatrix} = 0 \Rightarrow \hat{\delta} = \begin{bmatrix} \hat{\delta}_1^{IV} \\ \vdots \\ \hat{\delta}_M^{IV} \end{bmatrix} = \begin{bmatrix} (W'X_1)^{-1}W'y_1 \\ \vdots \\ (W'X_M)^{-1}W'y_M \end{bmatrix}$$

  In (i) estimators are asymptotically the same, and (ii), they're the same in finite same as well!

  2SLS with no cross equation restrictions is just equation by equation 2SLS: Think of it as the IV version of the multiple equation OLS estimator**. It does not take advantage of cross-equation correlations because weighting matrix is** $I_M \otimes E\left(\underline{w}(t)\underline{w}(t)'\right)^{-1}$.
  3SLS, on this other hand, is a "special case" of 2SLS, with a different weighting matrix such that it takes advantage of cross equation correlations.

  $ME\ \ OLS : ME\ \ 2SLS :: SUR : 3SLS$: SUR/3SLS are the "GLS" versions of the estimators in their class, under conditional homoskedasticity.

- **LIML vs. (single equation) 2SLS:**
  - They're both k class estimators. 2SLS is a k class estimator with k = 1 (p. 541). So when the equation is just identified (k = 1), LIML = 2SLS numerically.
  - LIML and 2SLS have same asymptotic distribution, so we cannot prefer one over the other on asymptotic grounds
  - In finite sample, LIML has invariance property while 2SLS does not (which makes LIML more desirable)
  - Other literature suggest that LIML should be preferred in finite sample over 2SLS

- **FIML vs. 3SLS:**
  They are asymptotically equivalent. However, LIML has invariance property that 3SLS (and 2SLS don't).

- **SUR vs. FIML/LIML:**
  If there is no endogeneity problems on the RHS, then 3SLS is SUR. Which implies that SUR and FIML/LIML estimation of the system are equivalent!

Multiple Equation (with cross equation restrictions): Generally, **with cross-equation restrictions, estimating in multiple equation is better**!

- Cross equation restrictions invariably gives us over-identification, and therefore we can obtain efficiency gains by exploiting the over-identification.
- If a system is just-identified when we impose the cross-equation restrictions, then each equation is not identified individually, in which case, we NEED to use multiple equation system!

VIII. LS vs. SUR and ME 2SLS vs. 3SLS (WHEN THERE IS NO CROSS EQUATION RESTRICTIONS)

1. ME OLS

a. ME OLS without cross equation restrictions (each equation has different parameter) = equation by equation OLS

$$Q_T = \frac{1}{T}U'U = \frac{1}{T}\left[\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} - \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix}\right]^T\left[\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} - \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix}\right] = \frac{1}{T}\begin{pmatrix} Y_1 - X_1\beta_1 \\ Y_2 - X_2\beta_2 \\ \vdots \\ Y_M - X_M\beta_M \end{pmatrix}^T\begin{pmatrix} Y_1 - X_1\beta_1 \\ Y_2 - X_2\beta_2 \\ \vdots \\ Y_M - X_M\beta_M \end{pmatrix}$$

$$= \frac{1}{T}\left[\left(Y_1 - X_1\beta_1\right)^T\left(Y_1 - X_1\beta_1\right) + \ldots + \left(Y_M - X_M\beta_M\right)^T\left(Y_M - X_M\beta_M\right)\right]$$

$$= \frac{1}{T}\left[Q_{1T} + \ldots + Q_{MT}\right] = \frac{1}{T}\left[\frac{SSR_1}{T} + \ldots + \frac{SSR_M}{T}\right]$$

So, if all equations have different parameters, when we take FOC with respect to each parameter, it's just doing equation by equation OLS.

b. Under conditional homoskedasticity across observations, if true var-cov matrix is diagonal, then ME OLS is efficient

$$L_T = \begin{bmatrix} \left.\frac{\partial Q_{1T}}{\partial \beta_1'}\right|_{1xK_1} \\ \vdots \\ \left.\frac{\partial Q_{MT}}{\partial \beta_M'}\right|_{1xK_M} \end{bmatrix} = \begin{bmatrix} \frac{-2}{T} X_1'(Y_1 - X_1\beta_1) \\ \vdots \\ \frac{-2}{T} X_M'(Y_M - X_M\beta_M) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_t -2x_{1t}\left(y_{1t} - x_{1t}'\beta_1\right) \\ \vdots \\ \frac{1}{T}\sum_t -2x_{Mt}\left(y_{1t} - x_{Mt}'\beta_M\right) \end{bmatrix}$$

$$A\text{var}\left(\sqrt{T}L_T\right) = V = -4\begin{bmatrix} E(x_{1t}u_{1t}^2 x_{1t}') & & E(x_t u_{1t} u_{Mt} x_{Mt}') \\ & \ddots & \\ E(x_t u_{Mt} u_{1t} x_{Mt}') & & E(x_{1t}u_{Mt}^2 x_{Mt}') \end{bmatrix} = -4\begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}') & & \sigma_{M1}E(x_t x_{Mt}') \\ & \ddots & \\ \sigma_{1M}E(x_t x_{Mt}') & & \sigma_M^2 E(x_{1t}x_{Mt}') \end{bmatrix} \text{ under Homo}$$

$$\text{If } \Omega \text{ diagonal, i.e.} \sigma_{ij} = 0 \text{ for } i \neq j, V = \begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}') & & \\ & \ddots & \\ & & \sigma_M^2 E(x_{1t}x_{Mt}') \end{bmatrix}$$

$$\text{and and } A\text{var}(\hat{\beta}) = S^{-1}V\left(S^{-1}\right)' = \begin{bmatrix} E(x_{1t}x_{1t}')^{-1} & & \\ & \ddots & \\ & & E(x_{1t}x_{1t}')^{-1} \end{bmatrix}\begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}') & & \\ & \ddots & \\ & & \sigma_M^2 E(x_{1t}x_{Mt}') \end{bmatrix}\begin{bmatrix} E(x_{1t}x_{1t}')^{-1} & & \\ & \ddots & \\ & & E(x_{1t}x_{1t}')^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}')^{-1} & & \\ & \ddots & \\ & & \sigma_M^2 E(x_{1t}x_{Mt}')^{-1} \end{bmatrix} = \begin{bmatrix} A\text{var}^*(\hat{\beta}_1) & & \\ & \ddots & \\ & & A\text{var}^*(\hat{\beta}_M) \end{bmatrix} \text{ where } A\text{var}^*(\hat{\beta}_i) \text{ is the single-equation OLS avar for } \hat{\beta}_i$$

c. If all equations just identified, $\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{X}_3 = \ldots$, then OLS = GLS in finite sample! (see below)
(In this case, the LS and JGLS/SUR estimator become the Multivariate Regression estimator)

d. If there is conditional heteroskedasticity, then nether OLS/JGLS are efficient!

2. JGLS/SUR

a. Under conditional homoskedasticity across observations, JGLS is efficient
(see below)

b. Under conditional homoskedasticity across observations, and given diagonal var-cov matrix, JGLS = OLS (no cross equation restrictions)

$$Q_T = \frac{1}{T}\sum_t \left(y(t) - \underline{x}(t)'\beta\right)' \hat{\Omega}^{-1}\left(y(t) - \underline{x}(t)'\beta\right),\ L_T = \frac{-2}{T}\sum_t \underline{x}(t)\hat{\Omega}^{-1}\left(y(t) - \underline{x}(t)'\beta\right)$$

$$A\text{var}\left(\sqrt{T}L_T\right) = V = E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)$$

$$Var(\hat{\beta}_{SUR}) = S^{-1}VS^{-1} = E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)^{-1} E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)^{-1}$$

$$= E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)^{-1}$$

$$\text{If } \Omega \text{ diagonal, } E\left(\underline{x}(t)\Omega^{-1}\underline{x}(t)'\right)^{-1} = E\left(\begin{bmatrix} x_{1t} & & \\ & \ddots & \\ & & x_{Mt}\end{bmatrix}\begin{bmatrix} \frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_M^2}\end{bmatrix}\begin{bmatrix} x_{1t}' & & \\ & \ddots & \\ & & x_{Mt}'\end{bmatrix}\right)^{-1} = E\left(\begin{bmatrix} x_{1t} & & \\ & \ddots & \\ & & x_{Mt}\end{bmatrix}\begin{bmatrix} \frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_M^2}\end{bmatrix}\begin{bmatrix} x_{1t}' & & \\ & \ddots & \\ & & x_{Mt}'\end{bmatrix}\right)^{-1}$$

$$= \begin{bmatrix} \frac{1}{\sigma_1^2}E(x_{1t}x_{1t}') & & \\ & \ddots & \\ & & \frac{1}{\sigma_M^2}E(x_{Mt}x_{Mt}')\end{bmatrix}^{-1} = \begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}')^{-1} & & \\ & \ddots & \\ & & \sigma_M^2 E(x_{Mt}x_{Mt}')^{-1}\end{bmatrix} = VAR - COV\ MATRIX\ OF\ OLS!$$

c. If all equations just identified, X_1=X_2=… (i.e. all regressors the same in each equation), then JGLS = OLS in finite sample as well and equal efficiency

$$\text{Then, } X = \left(I_M \otimes X_1\right)$$

$$\text{Thus, } \hat{\beta}_{GLS} = \left[X'V^{-1}X\right]^{-1}\left[X'V^{-1}Y\right] = \left[\left(I_M \otimes X_1\right)'\left(\Omega^{-1}\otimes I\right)\left(I_M \otimes X_1\right)\right]^{-1}\left[\left(I_M \otimes X_1\right)'\left(\Omega^{-1}\otimes I\right)Y\right]$$

$$= \left(\Omega \otimes \left[X_1'X_1\right]^{-1}\right)\left(\Omega^{-1}\otimes X_1'\right)Y = \left(I \otimes \left[X_1'X_1\right]^{-1}X_1'\right)Y$$

$$= \begin{pmatrix} \left[X_1'X_1\right]^{-1}X_1' & & \\ & \ddots & \\ & & \left[X_1'X_1\right]^{-1}X_1'\end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_M\end{pmatrix}$$

$$= \begin{pmatrix} \left[X_1'X_1\right]^{-1}X_1'Y_1 \\ \vdots \\ \left[X_1'X_1\right]^{-1}X_1'Y_M\end{pmatrix}$$

d. If heteroskedastic across observations as well, then no "robust" S.E. available. Need to model the heteroskedasticity for FGLS.

e. Quasi-ML interpretation of JGLS
See MaCurdy notes. But as long as there is no nonlinearity wrt the LHS variable, JGLS = QML.

3. ME 2SLS

a. ME2SLS without cross equation restrictions (each equation has different parameter) = equation by equation 2SLS

$$Q_T = \boxed{\frac{1}{T}(Y - X\beta)'\left(I_M \otimes W(W'W)^{-1}W\right)(Y - X\beta)} = \frac{1}{T}\left[\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} - \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_M \end{bmatrix}\right]' \begin{bmatrix} P_W & & \\ & \ddots & \\ & & P_W \end{bmatrix}\left[\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{pmatrix} - \begin{bmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{bmatrix}\begin{bmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_M \end{bmatrix}\right]$$

$$= \frac{1}{T}\begin{bmatrix} Y_1 - X_1\beta_1 \\ \\ Y_M - X_M\beta_M \end{bmatrix}' \begin{bmatrix} P_W & & \\ & \ddots & \\ & & P_W \end{bmatrix}\begin{bmatrix} Y_1 - X_1\beta_1 \\ \\ Y_M - X_M\beta_M \end{bmatrix} = \frac{1}{T}\left[(Y_1 - X_1\beta_1)' P_W (Y_1 - X_1\beta_1) + \ldots + (Y_M - X_M\beta_M)' P_W (Y_M - X_M\beta_M)\right]$$

$$= \frac{1}{T}\left[\left(W'(Y_1 - X_1\beta_1)\right)'(W'W)^{-1}\left(W'(Y_1 - X_1\beta_1)\right) + \ldots + \left(W'(Y_M - X_M\beta_M)\right)'(W'W)^{-1}\left(W'(Y_M - X_M\beta_M)\right)\right]$$

$$= Q_{1T} + \ldots + Q_{MT}$$

Therefore, if all the equations have different parameters, i.e. no cross-equation restrictions, then when we take FOC with respect to each equation's parameters, we are just performing the single-equation problem by itself!

b. Under homoskedasticity across observations, if true Var-Cov is diagonal, then we have "efficient" GMM

$$L_T = \frac{\partial Q_T}{\partial \beta'} = \begin{bmatrix} \left.\frac{\partial Q_{1T}}{\partial \beta_1'}\right|_{1xK_1} \\ \vdots \\ \left.\frac{\partial Q_{MT}}{\partial \beta_M'}\right|_{1xK_M} \end{bmatrix} = \begin{bmatrix} \frac{-2}{T} X_1' P_W (Y_1 - X_1\beta_1) \\ \vdots \\ \frac{-2}{T} X_M' P_W (Y_M - X_M\beta_M) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_t -2\left(\frac{1}{T}X_1'W\right)\left(\frac{1}{T}W'W\right)^{-1} w_t\left(y_t - x_{1t}'\beta_1\right) \\ \vdots \\ \frac{1}{T}\sum_t -2\left(\frac{1}{T}X_M'W\right)\left(\frac{1}{T}W'W\right)^{-1} w_t\left(y_t - x_{Mt}'\beta_M\right) \end{bmatrix}$$

$$A\mathrm{var}\left(\sqrt{T}L_T\right) = V = \begin{bmatrix} E\left(w_t x_{1t}'\right)' E\left(w_t w_t'\right)^{-1} E(w_t u_{1t}^2 w_t') E\left(w_t w_t'\right)^{-1} E\left(w_t x_{1t}'\right) & & E\left(w_t x_{1t}\right)' E\left(w_t w_t'\right)^{-1} E(w_t u_{1t} u_{Mt} w_t') E\left(w_t w_t'\right)^{-1} E\left(w_t x_{Mt}'\right) \\ & \ddots & \\ & & E\left(w_t x_{Mt}\right)' E\left(w_t w_t'\right)^{-1} E(w_t u_{Mt}^2 w_t') E\left(w_t w_t'\right)^{-1} E\left(w_t x_{Mt}'\right) \end{bmatrix}$$

$$\textit{Under hom}o = \begin{bmatrix} \sigma_1^2 E\left(w_t x_{1t}'\right)' E\left(w_t w_t'\right)^{-1} E\left(w_t x_{1t}'\right) & & \sigma_{1M} E\left(x_{1t} w_t'\right) E\left(w_t w_t'\right)^{-1} E\left(w_t x_{Mt}'\right) \\ & \ddots & \\ \sigma_{M1} E\left(w_t x_{1t}'\right)' E\left(w_t w_t'\right)^{-1} E\left(w_t x_{Mt}'\right) & & \sigma_{MM} E\left(x_{Mt} w_t'\right) E\left(w_t w_t'\right)^{-1} E\left(w_t x_{Mt}'\right) \end{bmatrix} = \begin{bmatrix} E\left(w_t x_{1t}'\right) & & \\ & \ddots & \\ & & E\left(w_t x_{Mt}'\right) \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 \Sigma_{ww}^{-1} & & \sigma_{1M} \Sigma_{ww}^{-1} \\ & \ddots & \\ \sigma_{M1} \Sigma_{ww}^{-1} & & \sigma_M^2 \Sigma_{ww}^{-1} \end{bmatrix} \begin{bmatrix} E\left(w_t x_{1t}'\right) & & \\ & \ddots & \\ & & E\left(w_t x_{Mt}'\right) \end{bmatrix}$$

$$= 4 \boxed{E\left(\underline{w}(t)\underline{x}(t)'\right)' \left(\Omega \otimes \Sigma_{ww}^{-1}\right) E\left(\underline{w}(t)\underline{x}(t)'\right)}$$

$$\textit{IF DIAGONAL} \quad \Omega, V = \begin{bmatrix} \sigma_1^2 \Sigma_{wx_1}' \Sigma_{ww}^{-1} \Sigma_{wx_1} & & \\ & \ddots & \\ & & \sigma_M^2 \Sigma_{wx_M}' \Sigma_{ww}^{-1} \Sigma_{wx_M} \end{bmatrix},$$

$$\textit{and } A\text{var}(\hat{\beta}) = S^{-1} V \left(S^{-1}\right)' = \begin{bmatrix} \left(\Sigma_{wx_1}' \Sigma_{ww}^{-1} \Sigma_{wx_1}\right)^{-1} & & \\ & \ddots & \\ & & \left(\Sigma_{wx_M}' \Sigma_{ww}^{-1} \Sigma_{wx_M}\right)^{-1} \end{bmatrix} \begin{bmatrix} \sigma_1^2 \Sigma_{wx_1}' \Sigma_{ww}^{-1} \Sigma_{wx_1} & & \\ & \ddots & \\ & & \sigma_M^2 \Sigma_{wx_M}' \Sigma_{ww}^{-1} \Sigma_{wx_M} \end{bmatrix} \begin{bmatrix} \left(\Sigma_{wx_1}' \Sigma_{ww}^{-1} \Sigma_{wx_1}\right)^{-1} & & \\ & \ddots & \\ & & \left(\Sigma_{wx_M}' \Sigma_{ww}^{-1} \Sigma_{wx_M}\right)^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 \left(\Sigma_{wx_1}' \Sigma_{ww}^{-1} \Sigma_{wx_1}\right)^{-1} & & \\ & \ddots & \\ & & \sigma_M^2 \left(\Sigma_{wx_M}' \Sigma_{ww}^{-1} \Sigma_{wx_M}\right)^{-1} \end{bmatrix} = \begin{bmatrix} A\text{var}^*(\hat{\beta}_1) & & \\ & \ddots & \\ & & A\text{var}^*(\hat{\beta}_M) \end{bmatrix} \quad \textit{where} \quad A\text{var}^*(\hat{\beta}_i) \ \textit{is the} \ \text{sin}\,\textit{gle} - \textit{equation 2SLS a var for} \ \hat{\beta}_i$$

(Note: In this case, 2SLS may do better in finite sample given our a priori knowledge of the diagonal matrix. Asymptotically same, though.)

c. If all equations just identified, then both estimators are both defined by the solution to the system of equations, i.e. equation by equation I.V.

$$H_T(\delta) = \frac{1}{T}\left(I \otimes W'\right)(Y - X\delta) = \begin{bmatrix} \frac{1}{T} W'(y_1 - X_1\delta_1) \\ \vdots \\ \frac{1}{T} W'(y_M - X_M\delta_M) \end{bmatrix} = 0 \Rightarrow \hat{\delta} = \begin{bmatrix} \hat{\delta}_1^{IV} \\ \vdots \\ \hat{\delta}_M^{IV} \end{bmatrix} = \begin{bmatrix} \left(W'X_1\right)^{-1} W' y_1 \\ \vdots \\ \left(W'X_M\right)^{-1} W' y_M \end{bmatrix}$$

i.e. both estimators are exactly the same in finite sample.

d. If there is conditional heteroskedasticity across observations, then neither is efficient GMM and need robust standard errors

e. QML version of ME 2SLS: LIML

Finite sample estimates not exactly the same, but asymptotically the same distribution.

4. 3SLS

a. Under Conditional Homoskedasticity across equations, 3SLS is efficient GMM
(see below: Var-Cov matrix reduces as long as we have conditional homoskedasticity)

b. If Var-Cov diagonal, then 3SLS =$_A$ 2SLS = Efficient GMM

$$A\text{var}\left(\sqrt{T}L_T\right) = V = \boxed{E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes\Sigma_{ww}^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)} \quad under \ \text{hom}o$$

$$A\text{var}(\hat{\beta}) = S^{-1}V\left(S^{-1}\right)' = \left[E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)\right]^{-1}\left[E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)\right]\left[E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)\right]^{-1}$$

$$= \left[E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)\right]^{-1}$$

$$IF \ \ DIAGONAL \ \ \Omega, \ then \ A\text{var}(\hat{\beta}) = S^{-1} = \left[E\left(\underline{w}(t)\underline{x}(t)'\right)'\left(\Omega^{-1}\otimes E\left(w(t)w(t)'\right)^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)'\right)\right]^{-1}$$

$$= \begin{bmatrix} \left(\frac{1}{\sigma_1^2}\Sigma_{wx_1}'\Sigma_{ww}^{-1}\Sigma_{wx_1}\right)^{-1} & & \\ & \ddots & \\ & & \left(\frac{1}{\sigma_M^2}\Sigma_{wx_M}'\Sigma_{ww}^{-1}\Sigma_{wx_M}\right)^{-1} \end{bmatrix} = \begin{bmatrix} \sigma_1^2\left(\Sigma_{wx_1}'\Sigma_{ww}^{-1}\Sigma_{wx_1}\right)^{-1} & & \\ & \ddots & \\ & & \sigma_M^2\left(\Sigma_{wx_M}'\Sigma_{ww}^{-1}\Sigma_{wx_M}\right)^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} A\text{var}^*(\hat{\beta}_1) & & \\ & \ddots & \\ & & A\text{var}^*(\hat{\beta}_M) \end{bmatrix} \quad where \ \ A\text{var}^*(\hat{\beta}_i) \ \ is \ the \ \text{sin}gle-equation \ 2SLS \ a\text{var} \ \ for \ \hat{\beta}_i$$

Thus, as long as diagonal var-cov matrix, then 3SLS and 2SLS are asymptotically equivalent and efficient GMM.

c. If all equations just identified, then both estimators are both defined by the solution to the system of equations, i.e. equation by equation I.V.

$$H_T(\delta) = \frac{1}{T}(I\otimes W')(Y - X\delta) = \begin{bmatrix} \frac{1}{T}W'(y_1 - X_1\delta_1) \\ \vdots \\ \frac{1}{T}W'(y_M - X_M\delta_M) \end{bmatrix} = 0 \Rightarrow \hat{\delta} = \begin{bmatrix} \hat{\delta}_1^{IV} \\ \vdots \\ \hat{\delta}_M^{IV} \end{bmatrix} = \begin{bmatrix} (W'X_1)^{-1}W'y_1 \\ \vdots \\ (W'X_M)^{-1}W'y_M \end{bmatrix}$$

i.e. both estimators are exactly the same in finite sample.

d. If there is conditional heteroskedasticity across observations, then neither is efficient GMM and need robust standard errors

e. QML version of 3SLS: FIML
Again, different finite sample estimates, but asymptotically the same distribution.

5. Comparing all Multiple Equation Estimators

- No endogeneity, diagonal var-cov matrix: 3SLS = JGLS = FIML and, if no cross equation restrictions = OLS = 2SLS
- No endogeneity, non-diagonal var-cov matrix: 3SLS = JGLS = FIML

IX. Consistency of Parameter Estimates Issues: When/ what do we need for estimators to be consistent?

If $L_T(\theta_0)$ L(theta) doesn't ALL go to 0, then all estimates are off! Why? $\hat{\theta}-\theta=\left[S(\theta_0)\right]^{-1}L_T(\theta_0)\Rightarrow p\lim\left(\hat{\theta}-\theta\right)=\left[p\lim S(\theta_0)\right]^{-1}p\lim L_T(\theta_0)$. If one element of $p\lim L_T(\theta_0)$ non-0, can't generally have consistency. Unless by SOME luck that $\left[p\lim S(\theta_0)\right]^{-1}$ is such that the nonzero element does not enter. NOT GENERALLY TRUE. (PSI)

What do we need for consistency?

General

a. Least Squares / GMM estimators:
   i. Regularity conditions to guarantee uniform convergence
   ii. Need to show $L_T(\theta_0)\xrightarrow{P}0$
b. MLE:
   i. Only need regularity conditions to guarantee uniform convergence

   We get $L_T(\theta_0)\xrightarrow{P}0$ for free because

$$\frac{\partial}{\partial\theta}\left(\frac{1}{T}\sum_t\log L(x_t,\theta)\right)\rightarrow_P E\left(\frac{\partial\log L(x_t,\theta)}{\partial\theta}\right)=E\left(\frac{1}{L(x_t,\theta)}\frac{\partial L(x_t,\theta)}{\partial\theta}\right)=\int_{-\infty}^{\infty}\frac{1}{L(x_t,\theta)}\frac{\partial L(x_t,\theta)}{\partial\theta}L(x_t,\theta)\,dx_t=\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}L(x_t,\theta)\,dx_t=\frac{\partial}{\partial\theta}(1)=0$$

c. QMLE for NLS:
   i. Need regularity conditions to guarantee uniform convergence
   ii. Need linearity in endogenous variables
   (If not, then either do MLE and assume distribution to be true, or do N2SLS, or LIML/FIML.

N2SLS

Can't put fitted values in the "first stage". There is no "two stage" here, a misnomer. If we put fitted values in they are not consistent. Why?

$$\text{"True" } Q_T=\left[\frac{1}{T}W'\left(Y-f(X,\beta)\right)\right]'\left(\frac{1}{T}W'W\right)^{-1}\left[\frac{1}{T}W'\left(Y-f(X,\beta)\right)\right]\ \neq\ Q_T^*=\left[\frac{1}{T}\left(Y-f(P_WX,\beta)\right)\right]'\left[\frac{1}{T}\left(Y-f(P_WX,\beta)\right)\right]$$

If we have sample selection problem: $E(\varepsilon_t\mid\delta=1)\neq0$, because it's a function of the x's!

(then orthogonality condition will not hold, AND we can't instrument: the error term does not have mean = 0, therefore no matter what instrument you put in there we cannot fix this problem. All we can do is ML, make assumption about distribution of errors. Including inverse mills ratio gives us consistency.

X. Consistency of Standard Errors Issues: When are Standard Errors of estimates unreliable?

When standard errors are inconsistent we cannot do hypothesis testing. (e.g. Wald test, needs estimator and standard errors be consistent)

a. Fitted Values on RHS of Regression Model
   i. Heckman 2-Step
   ii. Any 2-step estimation procedure
b. Non-Robust S.E.'s when errors are heteroskedastic
c. When we have autocorrelated errors. Or more precisely, when x_te_t autocorrelated, then we need a HAC estimator!
(Note: This is also true in multiple equation systems, when the within-equation errors are autocorrelated)

XI. Quasi Maximum Liklihood

a. Estimation

i. Why we care

Maximum likelihood estimators are the most efficient in a very broad "regular" class of estimators.

ii. For "least squares" estimators, assuming normality (quasi-max-like) gets us the same estimators as in a GLS problem (CHECK THE CASES WHEN THEY ARE EQUIVALENT)

Example1: OLS with heteroscedasticity

$y_t = x_t'\beta + e_t \quad and \quad assume\ e_t \mid x_t \sim_{niid} N(0,\sigma_t^2) \Rightarrow y_t \sim N(x_t'\beta, \sigma_t^2)$

$Then,\ f_t(y_t \mid \beta) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\left\{-\frac{1}{2\sigma_t^2}(y_t - x_t'\beta)^2\right\} \Rightarrow \log f_t = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_t^2\right) - \frac{1}{2\sigma_t^2}(y_t - x_t'\beta)^2$

$$Q_T = \frac{1}{T}\sum_t \log f_t = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_t \log\left(\sigma_t^2\right) + \sum_t \frac{1}{2\sigma_t^2}(y_t - x_t'\beta)^2$$

$$L_T\left(\hat{\beta}\right) = \frac{\partial Q_T\left(\hat{\beta}\right)}{\partial \beta} = \sum_t \frac{1}{\sigma_t^2} - x_t\left(y_t - x_t'\hat{\beta}\right) = 0$$

$$\hat{\beta}_{ML} = (X'V^{-1}X)\left(X'V^{-1}Y\right) = \hat{\beta}_{GLS} \quad where\ V = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_T^2 \end{pmatrix}$$ (*THIS IS THE WLS ESTIMATOR*!)

$$\text{Re}call,\ by\ in\text{var}iance\ \hat{\sigma}_{ML}^2 = \frac{1}{T}\sum\left(y_t - x_t'\hat{\beta}_{ML}\right)^2$$

Here, we run into the feasibility issue, since the error variance for each observation is not known. Functional form needs to be assumed.

Example 2: OLS with homoskedasticity

$$Q_T = \frac{1}{T}\sum_t \log f_t = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma^2\right) + \frac{1}{T}\sum_t \frac{1}{2\sigma^2}(y_t - x_t'\beta)^2$$

$$L_T\left(\hat{\beta}\right) = \frac{\partial Q_T\left(\hat{\beta}\right)}{\partial \beta} = \frac{1}{\sigma^2}\sum_t -x_t\left(y_t - x_t'\hat{\beta}\right) = 0 \Rightarrow \hat{\beta}_{ML} = (X'X)\left(X'Y\right) = \hat{\beta}_{OLS}$$

$$\text{Re}call,\ by\ in\text{var}iance\ \hat{\sigma}_{ML}^2 = \frac{1}{T}\sum\left(y_t - x_t'\hat{\beta}_{ML}\right)^2 = \frac{1}{T}SSR$$

Takeaway: People assume normality not because they believe normality, but because the assumption of normality does not matter, and it gets us to the same estimators and allows us to claim optimality in a very broad class of estimators. Also, computationally easier.

b. Change of Variables Formula: Useful for Finding the Density function

Suppose $u \sim f_u$, *and* $y = g(u)$ *for* g *monotone, and* g^{-1} *continuously differentiable.*

Then, $f_v(.) = f_u(g^{-1}(y))abs(|J|)$ *where* $|J| = \det\left(\frac{\partial g^{-1}(y)}{\partial y}\right)$

c. QML and Instrumental Variables: LIML and FIML (QML version of 2SLS and 3SLS)

$$\left.\begin{aligned} Y_1 &= \beta_0 + \beta_1 Y_2 + \varepsilon \\ Y_2 &= \pi_0 + \pi_1 '\mathbf{Z} + u \end{aligned}\right\} = \qquad \left.\begin{aligned} Y_1 - \beta_2 Y_2 &= \beta_0 + \varepsilon \\ Y_2 &= -\pi_0 + \pi_1 '\mathbf{Z}_{(mx1)} + u \end{aligned}\right\} \Rightarrow \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}_{2x2} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{2x1} = \begin{bmatrix} \beta_0 & \mathbf{0} \\ \pi_0 & \pi_1' \end{bmatrix}_{2xm} \begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix}_{mx1} + \begin{bmatrix} \varepsilon \\ u \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{2x1} = \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \beta_0 & \mathbf{0} \\ \pi_0 & \pi_1' \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix} + \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \varepsilon \\ u \end{bmatrix}$$

Or, $Y_t = \Gamma^{-1} B X_t + \Gamma^{-1} v_t$

Suppose $v_t \sim_{iid} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Omega\right) \Rightarrow f_v = \frac{1}{2\pi^{m/2}|\Omega|} \exp\left\{-\frac{1}{2}(v_t)'\Omega^{-1}(v_t)\right\}$

Let $g(v_t) = \Gamma^{-1} B X_t + \Gamma^{-1} v_t \Rightarrow g^{-1}(Y_t) = \Gamma Y_t - B X_t$ *and* $\frac{\partial g^{-1}(Y_t)}{\partial Y_t} = \Gamma$

Then, by change of $\text{var}:\ f_Y = \frac{1}{(2\pi)^{m/2}|\Omega|^{1/2}} \exp\left\{-\frac{1}{2}(\Gamma Y_t - B X_t)'\Omega^{-1}(\Gamma Y_t - B X_t)\right\} \|\Gamma\|$

Thus,

$$Q_T = \frac{1}{T}\sum_t \log f(Y_t; B, \Omega) = -\frac{m}{2}\log(2\pi) - \log|\Omega|^{\frac{1}{2}} + \log\|\Gamma\| - \frac{1}{2T}\sum_t (\Gamma Y_t - B X_t)'\Omega^{-1}(\Gamma Y_t - B X_t)$$

Note: $\|\mathbf{\Gamma}\|$ may not be treated as a constant, it might have parameters in there (though in this particular case, the determinant is 0).
Note2: If there is heteroskedasticity, then we need to write our own program / model the heteroskedasticity.
Note: To be able to do this, we NEED linearity of the parameters in the endogenous variables. Otherwise we cannot write the problem as matrix multiplication.

d. Validity of Hypothesis Testing (without Normality Assumption)

Continuing from above example, we show that the LR can be used (when there is homoskedasticity). Normality assumption does not matter!

$$\therefore Q_T = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR}{T}\right) + \frac{1}{2T\frac{SSR}{T}}\sum_t \left(y_t - x_t'\hat{\beta}\right)^2$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR}{T}\right) + 1$$

Thus, in likelihood ratio test, since *we know* TQ_T *satisfies the "property"*,

$$2T\left(Q_{T,U} - Q_{T,C}\right) = 2T\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR_U}{T}\right) + 1 + \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{SSR_C}{T}\right) - 1\right)$$

$$= T\left(\log\left(\frac{SSR_C}{T}\right) - \log\left(\frac{SSR_U}{T}\right)\right) = T\left(\log(SSR_C) - \log(SSR_U)\right)$$

$$= T\left(\log\left(\frac{SSR_U}{SSR_C}\right)\right)$$

Note: We showed in homework that this is a proper transformation of Q_T =e'e such that the "property" is satisfied. So, the **log-likelihood of the model constitutes a valid distance function test**!

Note2: **This works for LS/NLLS estimator with homoskedastic errors (when we assume normality). (See HW2 #1 (iii))**

e. When does "QUASI" not work? When is Normality a critical assumption?

- **If in single equation or in SUR we have nonlinearity of parameter in y(t), then the transformation has a non-constant Jacobian and QMLE is NOT the same as (J)NLS! (here we need to assume normality to believe our results)**
- When the assumptions under which we do the estimation is not valid. For example, if we do MLE on linear equation with no endogeneity, assuming homoskedasticity, we get OLS. But, if in fact there is heteroskedasticity, then this is no longer valid. Because the "optimal" estimator is WLS, and the "condition" no longer holds.
- In a **probit**, or any application in **censored regression, sample selection, dummy endogenous variables via ML estimation problems**, we need to assume that the distributional assumptions are correct, then we can use the above statistic – here the normality assumption IS important!

f. Heteroskedasticity and QML: Infeasible unless we know form of heteroskedasticity

As in GLS, if we don't know the form of heteroskedasticity, we cannot do ML estimation, it is infeasible; need to program on your own / model heteroskedasticity.

g. Endogeneity and QML

Suppose we have a single equation model with endogeneity problem.

- **If we believe in normality:** Then we can completely specify the distribution of the data (and identification condition is the kulbac-liebler), then we can estimate the parameters of the model. So, we do ML estimation (**NOT QUASI**).

- **If we do not believe in normality**: We cannot do "quasi" ml. Instead We could do LIML, i.e. the (quasi) ml counterpart to 2SLS. It doesn't necessarily give us the same estimates in finite sample as 2SLS, but **asymptotically they're the same**. The reason they're not the same is because once you take the FOC in the FIML problem, we don't get the same objective function back, there's a det|Gamma*sigma*gamma| term, and the gamma matrix has parameters wrapped in there that affects the maximization problem. In this case, I think "quasi" always works. In fact, in my notes for 3/2/07, I wrote in the first line: FIML/LIML are quasi max. likelihood applied to system of equations. But I guess we know that 2SLS and LIML have same asymptotic distribution, and 3SLS and FIML have same asymptotic distribution, so if one of them is consistent, then both of them are. So intuitively (without really knowing how to show the "why" part except wildly waving hands), it seems like the normality assumption does not matter. But of course we could always just use 2SLS. The reason why people might prefer LIML to 2SLS though (says Hiyashi) is that LIML seems to have better finite sample properties.

XII. Hypothesis Testing

a. Wald Test/T-Test (NEED CONSISTENCY OF ESTIMATOR)

From estimation we know:

$\hat{\beta} - \beta_0 \sim_A N(0, A)$ *In OLS with* hom*oskedasticity,* $A = \frac{1}{T}\hat{\sigma}^2 (X'X)^{-1}$, *OLS with robust SE,* $A = \frac{1}{T}\left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{T}X'\hat{\varepsilon}\hat{\varepsilon}'X\right)\left(\frac{1}{T}X'X\right)^{-1}$

$H_0 : a(\beta_0) = 0 \quad H_a : Not\ H_0$

Then, by delta method, $\sin ce\ \sqrt{T}\left(\hat{\beta} - \beta_0\right) \Rightarrow_D N(0, p\lim TA)$

$\sqrt{T}\left(a(\hat{\beta}) - a(\beta_0)\right) = \sqrt{T}a(\hat{\beta}) \Rightarrow_D N_q\left(0, \frac{\partial a(\beta_0)}{\partial \beta'} p\lim TA \frac{\partial a(\beta_0)}{\partial \beta}\right)$ *under the null* (*assu*$\min g\ R = \frac{\partial a(\beta_0)}{\partial \beta'}$ *full row rank*)

Thus,

$$a(\hat{\beta}) \sim_A N_q\left(0, \underset{qxk}{\frac{\partial a(\hat{\beta})}{\partial \beta'}} \underset{kxk}{A} \underset{kxq}{\frac{\partial a(\hat{\beta})}{\partial \beta}}\right)$$

$$W \equiv \left(a(\hat{\beta})\right)'\left(\hat{R}A\hat{R}\right)^{-1}\left(a(\hat{\beta})\right) \sim_A ChiSq(q)$$

Note: T-test is when q = 1. Then we can do normal approximation.

b. Distance Function Test

Distance function needs to satisfy

$$A\,\mathrm{var}\left(\sqrt{T}\left.\frac{\partial Q_T}{\partial \beta}\right|_{\beta_0}\right) = A\,\mathrm{var}\left(\sqrt{T}H_T(\beta_0)\right) = E(l_t(\beta_0)l_t(\beta_0)') \underset{A}{=} \pm\left.\frac{\partial^2 Q_T}{\partial\beta\partial\beta'}\right|_{\beta_0} = E\left(\left.\frac{\partial l_t}{\partial \beta}\right|_{\beta_0}\right)$$

If "property" satisfied for Q_T, then,

$$\mathbf{2T(Q_{T,C} - Q_{T,U}) \sim_a ChiSq(q)}$$

KEY: THE DISTANCE FUNCTION MUST REMAIN THE SAME. SO IF WE MAKE ANY TRANSFORMATIONS, MUST MAKE THE SAME TRANSFORMATION IN ORDER FOR THE DISTANCE FUNCTION TEST TO BE VALID.

- For LS type estimators, we always need homoskedasticity.
  OLS/NLS: (Assuming Homoskedasticity) – For proof, see HW2 #1 (iii).

  Why Need Adjustment: $4\sigma^2 E(x_t x_t') \neq 2E(x_t x_t')$

  Original: $Q_T = \frac{1}{T} e'e$

  Adjustment: $Q_T^* = \frac{1}{2\hat{\sigma}_u^2} Q_T$. *Then,* $2T\left(Q_{T,r}^* - Q_{T,u}^*\right) = 2T\left(\frac{Q_{T,r}}{2\hat{\sigma}_u^2} - \frac{Q_{T,u}}{2\hat{\sigma}_u^2}\right) = \frac{1}{\hat{\sigma}_u^2}\left(SSR_R - SSR_U\right) \sim Chi - Sq(q)$

- For Single Equation GMM-Type estimators, distance function test valid when the estimator is the "efficient" GMM estimator
  - 2SLS/NL2SLS (HW2 #2 iv b): (Need conditional homoskedasticity – under this condition we know 2SLS is efficient GMM)

    Original: $Q_T = H_T M_T H_T$ *with* $M_T = \left(\frac{1}{T} W'W\right)^{-1}$, $H_T = \frac{1}{T} W'e$

    Adjustment: $Q_T^* = \frac{1}{2\hat{\sigma}^2} Q_T$. *Then,* $2T\left(Q_{T,r}^* - Q_{T,u}^*\right) = 2T\left(\frac{Q_{T,r}}{2\hat{\sigma}^2} - \frac{Q_{T,u}}{2\hat{\sigma}^2}\right) = \frac{1}{\hat{\sigma}^2}\left(SSR_R - SSR_U\right) \sim Chi - Sq(q)$ NL2SLS

  Note: If no homoskedasticity, we can pick a different GMM estimator, using the efficient weighting matrix. Then, distance function test will be valid again.

- The "efficient" GMM estimator. Only need to divide by 2.

- For M.E. GLS Estimators:
  - OLS: (Need conditional homoskedasticity across observations, diagonal and no endogeneity)

    Original: $Q_T = \frac{1}{T}\left[Q_{1T} + ... + Q_{MT}\right] = \frac{1}{T}\left[\frac{SSR_1}{T} + ... + \frac{SSR_M}{T}\right]$

    Adjustment: $Q_T^* = \frac{Q_{1T}}{2\hat{\sigma}_{1,U}^2} + ... + \frac{Q_{MT}}{2\hat{\sigma}_{M,U}^2}$. *Then,* $2T\left(Q_{T,r}^* - Q_{T,u}^*\right) = \left(\left[\frac{SSR_{1,R}}{\hat{\sigma}_{1,U}^2} + ... + \frac{SSR_{M,R}}{\hat{\sigma}_{M,U}^2}\right] - \left[\frac{SSR_{1,U}}{\hat{\sigma}_{1,U}^2} + ... + \frac{SSR_{M,U}}{\hat{\sigma}_{M,U}^2}\right]\right) \sim Chi - Sq(q)$

    Why this adjustment?

    $$If\ \Omega\ diagonal,\ i.e. \sigma_{ij} = 0\ for\ i \neq j, V = 4\begin{bmatrix} \sigma_1^2 E(x_{1t}x_{1t}') & & \\ & \ddots & \\ & & \sigma_M^2 E(x_{1t}x_{Mt}') \end{bmatrix}\ and\ A\mathrm{var}(\hat{\beta}) = S^{-1}V\left(S^{-1}\right)' = 2\begin{bmatrix} E(x_{1t}x_{1t}')^{-1} & & \\ & \ddots & \\ & & E(x_{1t}x_{1t}')^{-1} \end{bmatrix}$$

  - JGLS/SUR: (Need conditional homoskedasticity across observations/within equations, and no endogeneity)

    Original: $Q_T = \frac{1}{T}\left(Y - X\beta\right)'\left(\hat{\Omega}^{-1} \otimes I\right)\left(Y - X\beta\right)$

    Adjustment: $Q_T^* = \frac{1}{2} Q_T$. *Then,* $2T\left(\frac{1}{2} Q_{T,r} - \frac{1}{2} Q_{T,u}\right) = T\left(Q_{T,r} - Q_{T,u}\right) \sim Chi - Sq(q)$ (Note: This is NOT SSR, but the minimized Q_T)

    If we believe Phi is diagonal, then we can do as above, since JGLS output gives us SSR's for each equation.

- For M.E. GMM estimators
  - 2SLS (Need conditional homoskedasticity across observations/within equations, $\boxed{\Phi = Diagonal}$, and no covariance across equations (In this case, the estimator is same as 3SLS and is efficient GMM)

    Original: $Q_T = H_T^{'}\left(I_M \otimes \frac{1}{T} W'W\right)^{-1} H_T = Q_{1T} + \ldots + Q_{MT}$

    Adjustment: $Q_T^* = \frac{Q_{1T}}{2\hat{\sigma}_{1,U}^2} + \ldots + \frac{Q_{MT}}{2\hat{\sigma}_{M,U}^2}$. $Then,\ 2T\left(Q_{T,r}^* - Q_{T,u}^*\right) = T\left(\left[\frac{Q_{1T,R}}{\hat{\sigma}_{1,U}^2} + \ldots + \frac{Q_{MT,R}}{\hat{\sigma}_{M,U}^2}\right] - \left[\frac{Q_{1T,U}}{\hat{\sigma}_{1,U}^2} + \ldots + \frac{Q_{MT,U}}{\hat{\sigma}_{M,U}^2}\right]\right) \sim Chi - Sq(q)$

    Why this adjustment:

    $$A\mathrm{var}\left(\sqrt{T}L_T\right) = V = 4\boxed{E\left(\underline{w}(t)\underline{x}(t)^{'}\right)^{'}\left(\Omega \otimes \Sigma_{ww}^{-1}\right)E\left(\underline{w}(t)\underline{x}(t)^{'}\right)} = \begin{bmatrix} E\left(w_t\, x_{1t}^{'}\right) & & \\ & \ddots & \\ & & E\left(w_t\, x_{Mt}^{'}\right) \end{bmatrix}^T \begin{bmatrix} \sigma_1^2\Sigma_{ww}^{-1} & & \sigma_{1M}\Sigma_{ww}^{-1} \\ & \ddots & \\ \sigma_{M1}\Sigma_{ww}^{-1} & & \sigma_M^2\Sigma_{ww}^{-1} \end{bmatrix} \begin{bmatrix} E\left(w_t\, x_{1t}^{'}\right) & & \\ & \ddots & \\ & & E\left(w_t\, x_{Mt}^{'}\right) \end{bmatrix}$$

    $$P\lim S = 2\begin{bmatrix} \left(\Sigma_{wx_1}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_1}\right) & & \\ & \ddots & \\ & & \left(\Sigma_{wx_M}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_M}\right) \end{bmatrix}$$

    $$IF\ \ DIAGONAL\ \ \Omega,\ V = 4\begin{bmatrix} \sigma_1^2\Sigma_{wx_1}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_1} & & \\ & \ddots & \\ & & \sigma_M^2\Sigma_{wx_M}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_M} \end{bmatrix}.$$

    $$Thus,\ need\ to\ adjust\ each\ Q_{iT}\ \ by\ \frac{1}{2\sigma_i},\ \ V = \begin{bmatrix} \frac{1}{\sigma_1^2}\Sigma_{wx_1}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_M^2}\Sigma_{wx_M}^{'}\Sigma_{ww}^{-1}\Sigma_{wx_M} \end{bmatrix} = P\lim S$$

  - 3SLS (Need conditional homoskedasticity across observations/within equations, but heteroskedasticity across equations/different error variances in diff equations is ok)
    Original:

    Adjustment: $Q_T^* = \frac{1}{2}Q_T$. $Then,\ 2T\left(\frac{1}{2}Q_{T,r} - \frac{1}{2}Q_{T,u}\right) = T\left(Q_{T,r} - Q_{T,u}\right) \sim Chi - Sq(q)$ (Note: This is NOT SSR, but the minimized Q_T)

    Be Careful: 3SLS usually doesn't have same weighting matrix across restricted vs. unrestricted models!
    Unlike in JGLS, we cannot do the 2SLS test here since we are not given Q_T for each.

- For single equation ML estimators:
  - Need conditional homoskedasticity (if have homoskedastic model) or heteroskedasticity (if heteroskedastic model) Then, use LR test: **$2(LogLike_U - LogLike_C) \sim Chi\text{-}Sq(q)$.**
- For M.E. ML estimators (ML/FIML/LIML)
  - Need conditional homoskedasticity (if have homoskedastic model) or heteroskedasticity (if heteroskedastic model) Then, use LR test: **$2(LogLike_U - LogLike_C) \sim Chi\text{-}Sq(q)$.**

- For Probit Model
  - If estimated by LR, then, assuming we believe in normality, then use LR test.
  - Can also use NLS to estimate with heteroskedastic errors, but cannot do distance function test.

c. (Quasi) Likelihood Ratio Test

- In a LR Test, null is: restricted model and unrestricted model are the same.
- As long as the assumptions about the variance of errors is correct, then normality assumption does not matter and we can estimate the estimate and use the LR test without adjustment: **$2(LogLik_U - LogLik_R) \sim ChiSq(q)$**
  Why? BC the LR boils down to a proper transformation of the distance function that satisfies the property. So normality assumption does not matter.

$$\therefore Q_T = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR}{T}\right) + \frac{1}{2T\frac{SSR}{T}}\sum_t \left(y_t - x_t'\hat{\beta}\right)^2$$

$$= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR}{T}\right) + 1$$

Thus, in likelihood ratio test, since we know TQ_T satisfies the "property",

$$2T\left(Q_{T,U} - Q_{T,C}\right) = 2T\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{SSR_U}{T}\right) + 1 + \frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\frac{SSR_C}{T}\right) - 1\right)$$

$$= T\left(\log\left(\frac{SSR_C}{T}\right) - \log\left(\frac{SSR_U}{T}\right)\right) = T\left(\log(SSR_C) - \log(SSR_U)\right)$$

$$= T\left(\log\left(\frac{SSR_U}{SSR_C}\right)\right)$$

Note: We showed in homework that this is a proper transformation of Q_T =e'e such that the "property" is satisfied. So, the **log-likelihood of the model constitutes a valid distance function test**!

Note2: **This works for LS/NLLS estimator with homoskedastic errors (when we assume normality). (See HW2 #1 (iii))**

When does it not work?
When the assumptions under which we do the estimation is not valid. For example, if we do MLE on linear equation with no endogeneity, assuming homoskedasticity, we get OLS. But, if in fact there is heteroskedasticity, then this is no longer valid. Because the "optimal" estimator is WLS, and the "condition" no longer holds.

Note: Normality assumption does not matter because it reduces down to a GLS problem with a distance function test. Remember, "normality" is assumed because we are exploiting the structure of the pdf of a normal, that allows us to reduce the problem to a GLS problem. But computationally,

ML is faster and easier. So, we don't assume normality because we believe in normality, we assume normality because it is more efficient and gets us to the same place if we were to use the standard quadratic form / GLS distance function.

On the other hand, in a probit, we need to assume that the distributional assumptions are correct, then we can use the above statistic – here the normality assumption IS important!

d. (Specification Test) Wu-Hausman-Durbin Endogeneity Test

- **Why we care?** Again, if there is no endogeneity, then we would prefer to use OLS because of its greater efficiency (2SLS requires a first step approximation).
- **Idea:** Under the null that Y_2 exogenous, W(t) = {x(t),w_2(t)} valid instruments/exogenous, then, if I put any fitted Y_2 or v and it enters, the equation, then it either means it should have been included in the model. If Y_2 is endogenous/correlated with error, then it must be endogenous through $\hat{v}(t)$, since $Y_2(t) = \hat{Y}_2(t) + \hat{v}(t)$ and given valid instruments w(t), $\hat{Y}_2(t)$ is also orthogonal to the error term (since E(f(W)e(t)) = E(f(W)E(e(t)|W))=0). Therefore, if $\hat{v}(t)$ enters, that means it should have been included in the model in the first place. If fitted Y_2 enters controlling for Y_2, then it must be that there is some information that in Y2 correlated with the errors that should be part of the model.
- **Why it works (6 lines for comp): NEED TO MAINTAIN ASSUMPTION THAT THE OTHER REGRESSORS ARE EXOGENOUS**
  NEED OLS PROJECTION, and NEED LINEARITY!

  Model: $Y_1(t) = Y_2(t)\gamma_1 + x(t)'\beta + e(t)$

  H_0: Y_2 exogenous (and that w(t) is exogenous / w_1(t) overidentifying restrictions do not enter the model) H_a: Y_2 endogenous

  Assume: $w(t) = (w_1(t)\ x(t))'$ are valid instruments/exogenous

  Let $Y_2(t) = \hat{Y}_2(t) + \hat{v}(t)$ *where* $\hat{Y}_2(t) = f_t(W)$ (*CAN BE ANY* $f(W)$, *but typically orthogonal projection*)

  Then, $Y_1(t) = \hat{Y}_2(t)\gamma_1 + \hat{v}(t)\gamma_2 + x(t)'\beta + e(t)$

  Under null, Y_2 *exogenous* $\Rightarrow Cov(\hat{v}(t), e(t)) = Cov\left(Y_2(t) - \hat{Y}_2(t), e(t)\right) = E(Y_2(t)\varepsilon(t)) - E(f_t(W)\varepsilon(t)) = 0$ (since *w's and* Y_2 *assumed to be exogenous*)

  By projection theorem $\Rightarrow \text{cov}\left(\hat{Y}_2(t), \hat{v}(t)\right) = 0$, *and* $\text{cov}\left(x(t), \hat{v}(t)\right) = 0$ (since $x(t)$ *in* $w(t)$).

  Thus, coefficient on $\hat{v}(t)$ *does not depend / affect on other regressors / their coeffs and no endogeneity bias, and consistent for*:

$$\frac{Cov\left(\hat{v}(t), Y_1(t)\right)}{Var(\hat{v}(t))} = \frac{Cov\left(\hat{v}(t), Y_2(t)\gamma_1 + x(t)'\beta + e(t)\right)}{Var(\hat{v}(t))} = \frac{\gamma_1 Cov\left(\hat{v}(t), \hat{Y}_2(t) + \hat{v}(t)\right)}{Var(\hat{v}(t))} = \gamma_1$$

  Thus, if $Y_2(t)$ *exogenous,* $H_0 : \gamma_1 = \gamma_2$

  2 *Tests*:

$$Y_1(t) = \hat{Y}_2(t)\gamma_1 + \left(Y_1(t) - \hat{Y}_2(t)\right)\gamma_2 + x(t)'\beta + e(t) \quad or \quad Y_1(t) = \left(Y_2(t) - \hat{v}(t)\right)\gamma_1 + \hat{v}(t)\gamma_2 + x(t)'\beta + e(t)$$

$$\Rightarrow (1)\ Y_1(t) = Y_2\gamma_2 + \left(\gamma_1 - \gamma_2\right)\hat{Y}_2(t) + x(t)'\beta + e(t) \quad or \quad (2)\ Y_1(t) = Y_2(t)\gamma_1 + \left(\gamma_2 - \gamma_1\right)\hat{v}(t) + x(t)'\beta + e(t)$$

- **Variants: Doing DHW in 2SLS (projecting what on what??)**

- **Including the fitted residuals gives us consistency**: Why? Because if Y_2 is in fact endogenous, then it is correlated with error through $\hat{v}(t)$. Thus, now that we are controlling for the part of Y_2 that is endogenous, now we would have consistency: $Y_1(t) = Y_2(t)\gamma_1 + \left(\gamma_2 - \gamma_1\right)\hat{v}(t) + x(t)'\beta + e(t)$
- **Testing for endogeneity of multiple regressors?** Project the suspected endogenous variables onto the instruments, include them in the regression and do a joint test / Wald Test.

- **Note on Rejection and Overidentifying Restrictions:**
  If we reject, it could be because Y_2 is not exogenous (maintaining that x(t) exogenous). However, if our null hypothesis is that $x_2(t)$ is exogenous (i.e. the extra, overidentifying restrictions), maintaining Y_2 and $x_1(t)$ is exogenous, then the test would be exactly the same but instead we would reject $x_2(t)$ as valid instruments.

  If our null is that y2(t) is exogenous, and w(t) = w(t) = {x(t), w2(t)} are valid instruments. Then, controlling for y2(t) and x(t), w(t) should not be correlated with y(t), or some function of w(t) (i.e. y-hat) should not be correlated with

  If we change null hypothesis to, y2(t) endogenous, and w(t) = {x(t), w2(t)} valid instruments. Then, controlling for y2(t) and x(t), w(t) should not be correlated with y(t), or any function of it. Thus if it enters the model, we reject the null.

- **DON'T DO**:
  - **Don't plug yhat(t) into a nonlinear specification**
  - **Don't estimate yhat(t) using a nonlinear specification**

e. (Specification Test) Test for Overidentifying restrictions (i.e. testing whether an instrument is valid)

- **"Overidentifying"**: Recall, for IV estimation we need at least as many instruments as regressors. In this case, the model is exactly identified. If we have more instruments than regressors, then we have overidentification. We want to test whether these overidentifying restrictions, i.e. the extra instruments, are valid instruments.

  First, in order to do the test, must have over-identified system. Taking out the instruments in question, the valid instruments must make the system AT LEAST just identified. (Otherwise we would have extreme Multicollinearity).
  If not, and all I have is 1 instrument, then all I can do is stick w1(t) on the right hand side. But we don't know how to interpret the results if significant. We don't know if it's that w1 is not a valid instrument or that Y2 is endogenous.

  If we have over-identification (at least 2 extra instruments)
  Testing 1 Instrument:
  $H_0 : w_2(t)$ *a valid instrument* (maintaining that y2 is endogenous, and **x**, w1 valid instruments)
  We run 2SLS on $y(t) = x_t^{'}\beta + \gamma\, y_{2t} + w_{2t} + \varepsilon_t$**,** fitting Y2 with **all instruments** $W_t = \{\mathbf{x}_t, w_{1t}, w_{2t}\}$.
  Then, under the null, since a valid instrument, w2 only affects y_1 through y_2, so should not enter the equation separately, controlling for $\hat{y}_2$, since all its explanatory power is included in $\hat{y}_2$. But, if controlling for $\hat{y}_2$, w2 still enters, then it must be still a significant factor in determining y(t) and should have been in the model in the first place. We reject the null if enters.

  (Note: Again, if we reject the null, then it could be that w2 not a valid instrument, or that the maintained assumption is wrong. We can't separate them!)

f. (Specification Test) Hausman Test: Test for Comparing Different Estimators (a more efficient to a less efficient) of the <u>Same Model (Another Way to Test Endogeneity)</u>

Use: Generally used to test a less efficient estimator, but one that is robust to mis-specification (e.g. robust to endogeneity/heteroskedasticity), against a more efficient estimator that is not robust.

(i.e. Both estimators are consistent under the null, and one is more efficient under the null; only ONE is consistent under the alternative hypothesis)

To compare a model estimated with a more efficient (and consistent under the null but not under alternative) to a less efficient estimator under the null (**SAME REGRESSORS**) (and consistent under the null and alternative), and testing whether they are significantly different from each other (i.e. 2SLS vs. OLS or even 3SLS vs. OLS), use the following Wald test:

H_0: x(t) is exogenous, both OLS and 2SLS are consistent and OLS is asymptotically efficient. So $b_{ols}=b_{2sls}$ (plims)
H_a: x(t) is endogenous, 2SLS is consistent but OLS is inconsistent.

$$\sqrt{T}\left[\left(\hat{\beta}^L-\hat{\beta}^E\right)-\left(\beta^L-\beta^E\right)\right]\Rightarrow_D N(0,Var\left(\beta^L\right)+Var\left(\beta^E\right)-2Cov\left(\beta^L,\beta^E\right))$$

$$Thus, under\ null, \sqrt{T}\left[\left(\hat{\beta}^L-\hat{\beta}^E\right)-\left(\beta^L-\beta^E\right)\right]\Rightarrow_D N(0,Var\left(\beta^L\right)+Var\left(\beta^E\right)-2Cov\left(\beta^L,\beta^E\right))=N(0,Var\left(\beta^L\right)-Var\left(\beta^E\right))$$

by Hausman (Cov bt estimates of a more and less efficient estimators given the same X's is just the var *of the efficient estimate)*

$$Thus,\left(\hat{\beta}^L-\hat{\beta}^E\right)'\left(Var\left(\beta^L\right)-Var\left(\beta^E\right)\right)^{-1}\left(\hat{\beta}^L-\hat{\beta}^E\right)\Rightarrow_D ChiSq(k)\quad k=\#\ of\ \ parameters\ in\ \beta/\#\ of\ \ regressors$$

We approximate with:

$$\boxed{W=\left(\hat{\beta}_{2SLS}-\hat{\beta}_{OLS}\right)'\left(Var(\beta_{2SLS})-Var\left(\hat{\beta}_{OLS}\right)\right)^{-1}\left(\hat{\beta}_{2SLS}-\hat{\beta}_{OLS}\right)\sim_{Approx} ChiSq(k)}$$

Generally, denote E as the more efficient estimator and L as the less efficient, then,

$$\boxed{W=\left(\hat{\beta}^L-\hat{\beta}^E\right)'\left(Var(\beta^L)-Var\left(\hat{\beta}^E\right)\right)^{-1}\left(\hat{\beta}^L-\hat{\beta}^E\right)\sim ChiSq(?)}$$

Note2: If the models contain different X's we should instead compute a different specification test statistic as MaCurdy proposed in lecture
Note3: Why do we care about this test? Because if there is no significant difference between the 2, we prefer OLS because of better efficiency (BLUE)!

g. Test for Autocorrelation (correlated errors)

Motivation: " If there true errors are autocorrelated, then this can be detected through the autocorrelation of the residuals.

$H_0:\varepsilon(t)$ *not autocorrelated* $\quad H_a:\varepsilon(t)$ *is* $AR(p)$ *or* $\varepsilon(t)$ *is* $MA(q)$

This is a version of the endogeneity test. Take fitted residuals, stick it on the RHS and see if it enters. If so, then reject the null that errors are not correlated.
The idea is the same: under the null that errors are not correlated, the lag residuals should not be correlated with should not enter the model.

h. Heteroskedasticity and Hypothesis Testing

Heteroskedasticity across observations rules out distance function test and lagrange multiplier test for those estimators that are most efficient under conditional homoskedasticity (i.e. (N)LS (if no endogeneity), (N)2SLS, (N)3SLS). This is because the "property" will be violated, since the estimators are no longer the most efficient estimators under these assumptions. **In these cases, we can still do Wald Test with robust standard errors!**

Heteroskedasticity does not prevent us from doing distance/LM tests, as long as we pick the right estimators – we need to pick the most efficient estimator under the assumptions (i.e. WLS, efficient GMM). Then, we can apply distance and LM tests.

i. <u>Hypothesis Testing for Structural Breaks / Population groups</u>

Restricted model: Run model not allowing coefficients to differ across subpopulation to get SSR(r)
Unrestricted model 1: Run model but have interaction terms with each of the regressors to get SSR(u) OR (equivalently) with dummies for each coefficient
Unrestricted model 2: Run original model twice, once for each population (to get SSR1u and SSR2u) (*)
Then, it is equivalent testing these restrictions by:

SSR(r) – SSR(u) = SSR(r) – (SSR1(u) + SSR2(u))

However, in order for us to do this test via distance function test, key assumption is conditional homoskedasticity across groups!

Unrestricted model 1 and unrestricted model 2 will have the SAME coefficient estimates and SAME total SSR! If homoskedastic within and across groups, then I can get asymptotically equivalent var-cov matrix by stacking the separately estimated one. If heteroskedastic errors across or within group, then stacking the robust standard error is asymptotically equivalent to robust standard errors of (*)

Unrestricted model 2 is equivalent to a partitioned regression that stacks as follows (See Appendix for How Dummy Variable Works)
Model (*): $y_t = \beta_0 D_1 + \beta_1 x_t D_1 + \beta_2 z_t D_1 + \beta_3 D_2 + \beta_4 x_t D_2 + \beta_5 z_t D_2 + \varepsilon_t$ **<u>EQUIVALENT TO</u>** $y_t = \beta_0 + D_2 + \beta_1 x_t + \beta_2 x_t D_2 + \beta_3 z_t + \beta_4 z_t D_2 + \varepsilon_t$ on 2 groups
With asymptotically equivalent standard errors. Thus, I can just construct the estimates and SE's by stacking the group by group equation, and conduct inference!

$$Y_{nx1} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} \quad X_{nx(2k)} = \begin{pmatrix} \mathbf{X}^{(1)}_{(n/2)xk} & 0 \\ 0 & \mathbf{X}^{(2)}_{(n/2)xk} \end{pmatrix} \quad \varepsilon_{nx1} = \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{pmatrix}$$

Running regression separately gives us

$$\hat{\beta}_1 = \left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)\prime}\mathbf{y}^{(1)} \quad and \quad \hat{\beta}_2 = \left(\mathbf{X}^{(2)\prime}\mathbf{X}^{(2)}\right)^{-1}\mathbf{X}^{(2)\prime}\mathbf{y}^{(2)}$$

$$SSR_1 = \left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right)'\left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right) \quad and \quad SSR_2 = \left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)'\left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)$$

Running regression with dummies gives us:

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

$$SSR = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) = \begin{pmatrix} \mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1 \\ \mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2 \end{pmatrix}'\begin{pmatrix} \mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1 \\ \mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2 \end{pmatrix}$$

$$= \left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right)'\left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right) + \left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)'\left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)$$

$$= SSR_1 + SSR_2$$

Note: We get the same coefficient estimates and the same SSR. BUT we get different Standard Errors in Standard OLS output!

This is because Standard OLS S.E. are calculated from $\frac{1}{T}\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ *where* $\hat{\sigma}^2 = \frac{SSR}{T}$ *or in software* $\frac{SSR}{T-K-1}$ (*both consistent*)

So, in the pooled regression, $\hat{\sigma}^2$ estimated using data from ALL groups, assuming that homoskedasticity ACROSS groups. Where as in separate regressions, $\hat{\sigma}^2$ is estimated using JUST data from that group.

So, if there is heteroskedasticity across groups, then running the pooled regression will give wrong standard errors.
If this is the case (and assuming homoskedasticity within groups), either

1. **Run separate regressions (in which case, we get right homoskedastic variance of error estimation for each group)**

If there is heteroskedasticity across and within groups,

1. **Run pooled regression but use Robust S.E. (but can't do distance function test) and run separate equations with robust s.e. (will get asymptotically the same s.e.'s**

j. "Test the impact of X": Take dervatives of y with respect to X and plug in related values for test.

BE CAREFUL WITH NONLINEAR SPECIFICATIONS: NOT JUST THE COEFFICIENT!

$d_1 = \Phi(\beta_0 + \beta_1 x_1) + v$

e.g. $\text{Im}\, pact\ of\ x_1\ on\ d_1\ (probability\ of\ 1)\ is: \frac{\partial d_1}{\partial x_1} = \phi(\beta_0 + \beta_1 x_1)\beta_1$

$Thus,\ H_0: \frac{\partial d_1}{\partial x_1} = 0 \Leftrightarrow \phi(\beta_0 + \beta_1 x_1)\beta_1 = 0 \Leftrightarrow \beta_1 = 0\ (\text{since}\ \phi(.) \neq 0\ bc\ pdf\ of\ normal)$

k. "Average effect of X": Plug in values for X and test.

XIII. Endogeneity

a. What does it mean to be endogenous?

Interpretation of Error Term: $y_t = x_t'\beta + \varepsilon_t$ **→ Error term are the (unobserved/not-included) FACTORS outside of x(t) that influence/determine y(t).**

Endogeneity: Thus, endogeneity of x(t) means that it is correlated with unobserved factors, or factors not included in the model that affect y(t).

b. Endogeneity in a system of simultaneous equations: To talk about endogeneity from "simultaneity", need to have economic model to back it up.
Key MaCurdy Point: We need an economic model to talk about what is endogenous and what is not endogenous in a simultaneous equation model.

- Example 1: Suppose we have a system $Q = P\gamma_1 + X_1\beta_1 + e_1$ and $P = X_2\beta_2 + e_2$. If our model is that price is exogenously moved, then it must be that Q is endogenous (assuming X_1 also exogenous).

c. Statistical Sources of Endogeneity

i. e(t) are autocorrelated
Ex:

$y_t = y_{t-1}\delta + x_t\beta + e_t,\ e_t = \rho e_{t-1} + u_t\ \ with\ u_t \sim_{iid} (0, \sigma^2)$

$Then,\ y_{t-1}\ endogenous\ \text{since}\ Cov(y_{t-1}, e_t) = Cov(y_{t-2}\delta + x_{t-1}\beta + e_{t-1}, \rho e_{t-1} + u_t) = \delta\rho Cov(y_{t-2}, e_{t-1}) + \rho Var(e_{t-1}) \neq 0$

Solution: **Durbin's Method (See "time series" section)**

ii. Errors in Variables / Measurement Error
Ex:

$y_t = x_t\beta + e_t\ but\ only\ observe\ x_t^* = x_t + u_t$

$Then,\ true\ eqn\ becomes: y_t = x_t^*\beta + (\beta u_t + e_t)\ \ and\ we\ have\ endogeneity\ with\ unobserved\ error$

Solution: **Instrument for x^* (Again, need economic model to justify what is a valid instrument for x*)**

d. Omitted Variables Bias

i. If the included variables are not correlated with the omitted variable, then there is no bias. Otherwise, there is a bias.

ii. Can we sign the bias? Generally no.

Simple Regressions

- If we ran a simple regression, and there is 1 omitted variable, then we can sign the bias provided that the omitted variable is correlated with the included variable.

$$Model: y = \beta_0 + \beta_1 x_1 + \varepsilon \quad True: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$p\lim \hat{\beta}_1 = \frac{Cov(y, x_1)}{Var(x_1)} = \frac{Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, x_1)}{Var(x_1)} = \beta_1 + \frac{\beta_2 Cov(x_1, x_2)}{Var(x_1)}$$

- If we ran a simple regression, and there are 2 (or more) omitted variables, then we can't (generally) sign the bias, provided that the omitted variables are correlated with the included variable. (We CAN sign if we know that the covariances and the coefficients all have the same sign. If not, we can't because we don't know the magnitudes).

$$Model: y = \beta_0 + \beta_1 x_1 + \varepsilon \quad True: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$p\lim \hat{\beta}_1 = \frac{Cov(y, x_1)}{Var(x_1)} = \frac{Cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, x_1)}{Var(x_1)} = \beta_1 + \frac{\beta_2 Cov(x_1, x_2) + \beta_3 Cov(x_1, x_3)}{Var(x_1)}$$

Multiple Regressions

- If we ran a multiple regression, and there is 1 omitted variable, then we can't the bias provided that the omitted variable is correlated with included variables, or if the omitted variable is correlated with at least one of the included variables, and the included variables are correlated with each other.

Suppose we regress $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$$\gamma = Var(\mathbf{x})^{-1} Cov(\mathbf{x}, y) = \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) \\ Cov(x_1, x_2) & Var(x_2) \end{bmatrix}^{-1} \begin{bmatrix} Cov(x_1, y) \\ Cov(x_2, y) \end{bmatrix}$$

$$= \frac{1}{Var(x_1)Var(x_2) - Cov(x_1, x_2)^2} \begin{bmatrix} Var(x_2) & -Cov(x_1, x_2) \\ -Cov(x_1, x_2) & Var(x_1) \end{bmatrix} \begin{bmatrix} Cov(x_1, y) \\ Cov(x_2, y) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{Var(x_2)Cov(x_1, y) - Cov(x_1, x_2)Cov(x_2, y)}{Var(x_1)Var(x_2) - Cov(x_1, x_2)^2} \\ \frac{Var(x_1)Cov(x_2, y) - Cov(x_1, x_2)Cov(x_1, y)}{Var(x_1)Var(x_2) - Cov(x_1, x_2)^2} \end{bmatrix} = \begin{bmatrix} \frac{Cov(x_1, y) - \frac{Cov(x_1, x_2)}{Var(x_2)} Cov(x_2, y)}{Var(x_1) - \frac{Cov(x_1, x_2)}{Var(x_2)} Cov(x_1, x_2)} \\ \frac{Cov(x_2, y) - \frac{Cov(x_1, x_2)}{Var(x_1)} Cov(x_1, y)}{Var(x_2) - \frac{Cov(x_1, x_2)}{Var(x_1)} Cov(x_1, x_2)} \end{bmatrix}$$

$$p\lim\hat{\beta}_1 = \frac{Cov(x_1,y) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_2,y)}{Var(x_1) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_1,x_2)}$$

$$= \frac{Cov(x_1,\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_2,\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)}{Var(x_1) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_1,x_2)}$$

$$= \frac{\beta_1\,\mathrm{var}(x_1) + b_3\,\mathrm{cov}(x_1,x_3) - \frac{Cov(x_1,x_2)}{Var(x_2)}b_1Cov(x1,x2) - b_3\left(\frac{Cov(x_1,x_2)}{\mathrm{var}(x2)}\mathrm{cov}(x2,x3)\right)}{Var(x_1) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_1,x_2)}$$

$$= \frac{b1\left(Var(x_1) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_1,x_2) + b3\left(Cov(x_1,x_3) - \frac{Cov(x_1,x_2)}{\mathrm{var}(x2)}\mathrm{cov}(x2,x3)\right)\right)}{Var(x_1) - \frac{Cov(x_1,x_2)}{Var(x_2)}Cov(x_1,x_2)}$$

So, here even if x1 and x2 are not correlated, as long as x1 and x3 are correlated, x1 and x2 are correlated, and the signs aren't always the same, then we can't tell!

iii. Figuring out the omitted variables bias using regression outputs

What is the effect of including a variable in a regression specification?

Suppose we want to know the effect of including Z in the model: $y_t = x_t'\beta + \varepsilon_t$

Then, we can run: $z_t = x_t'\alpha + u_t$

Then, if we have output from

$y_t = x_t'\beta + \delta\hat{u}_t + \varepsilon_t = x_t'\beta + \delta\left(z_t - x_t'\hat{\alpha}\right) + \varepsilon_t = x_t'\left(\beta - \hat{\alpha}\delta\right) + \delta z_t + \varepsilon_t$

Then, δ is the coefficient on z_t, and effect on each regressor is given by $\left(\beta - \hat{\alpha}\delta\right)$

e. Dummy Endogenous Variables

Suppose our model is: $y_2 = x_2\beta_2 + \delta y_1 + y_1(v_1 - v_0) + v_0$ *where y_1 is Indicator, so δy_1 gives diff intercept and $y_1(v_1 - v_0) + v_0$ is error*

Problem: $E\left[v_i \mid x_2, z\right] = 0$ *but* $E\left[v_i \mid x_2, y_1\right] = y_1E\left[\left(v_1 - v_0\right) \mid x_2, y_1\right] + E\left(v_0 \mid x_2, y_1\right) \neq 0$

Solution:

If $v_1 = v_0 = v$, then can use I.V.

If not, IV can't solve $E(v_i \mid x_2, y_1) \neq 0$.

DO MLE.

(see Neale's Notes)

XIV. Exogeneity and Orthogonality: Definitions, Instruments, Usefulness

Def: $\mathbf{z}_t$ exogenous if $E(\varepsilon_t \mid \mathbf{Z}) = 0 \rightarrow \mathrm{E}(E(\mathbf{z}_t\varepsilon_t) = E(\mathbf{z}_t E(\varepsilon_t \mid \mathbf{Z})) = 0$

Def: $\mathbf{z}_t$ first moment independent if $E(\varepsilon_t \mid \mathbf{z}_t) = 0 \rightarrow \mathrm{E}(E(\mathbf{z}_t\varepsilon_t) = E(\mathbf{z}_t E(\varepsilon_t \mid \mathbf{z}_t)) = 0$

Def: $\mathbf{z}_t$ is orthogonal to error / predetermined if $E(\varepsilon_t \mathbf{z}_t) = 0$

a. Implications: Any function of z_t is orthogonal to error

Any function of $\mathbf{z}_t$ is orthogonal to the error term: $E\left(f(\mathbf{z}_t)\varepsilon_t\right) = E\left(f(\mathbf{z}_t)E(\varepsilon_t \mid \mathbf{z}_t)\right) = 0$

- **In 2SLS, we don't really need to use a linear projection in the first step. It can be any function of z (e.g. $\Phi(\mathbf{z}_t\text{'}\mathbf{b})$).** Then, the fitted value of the endogenous variable is orthogonal to the error term, and we get consistency.
- **We can use any function of z as an instrument, since any function of z is orthogonal to the error term!**

XV. Instruments: How many and Justification.

It's important to remember that when we start talking what's endogenous and what's an instrument, we need to have an economic model in order to justify these things.

a. What we require for instruments

$E(\varepsilon_t \mid z_t) = 0$ and correlated with the endogenous variables. Note: This implies $E(z_t\varepsilon_t) = 0$ (orthogonality, a "pseudo" generality) and that instrument is only correlated to the LHS variable through the endogenous variable.

b. Implications: Any function of the instrument satisfies orthogonality

$E(f(z_t)\varepsilon_t) = E\left(f(z_t)E(z_t \mid \varepsilon_t)\right) = 0$

Thus, **if we have valid instruments, we can ALWAYS do at least as good by adding functions of these instruments.**

Example: Suppose we're going to estimate by 2SLS.

Suppose population model is: $y_{1t} = \beta_0 + \beta_1 y_{2t} + \beta_2 x_t + \varepsilon_t$ with y_2 endogenous.

Suppose we believe the following: $y_{2t} = \delta_0 + \delta_1 z_{1t} + \delta_2 z_{2t} + \delta_2 z_{3t} + u_t$ where the z's are valid instruments (and y is correlated with e through u)

Then, I can do AT LEAST as well by including all the cross products of the instruments, allowing for nonlinearities (which might allow us to predict y better) by using… :

$$y_{2t} = \delta_0 + \delta_1 z_{1t} + \delta_2 z_{2t} + \delta_2 z_{3t} + \delta_3 z_{1t} z_{2t} + \delta_4 z_{1t} z_{3t} + \delta_5 z_{2t} z_{3t} + \delta_6 z_{1t} z_{2t} z_{3t} + u_t$$

XVI. Simultaneous Equations, Structural and Reduced Form

Def: When we say a simultaneous equation, we mean the regressors and error term are related to each other through a system of simultaneous equations (e.g. demand and supply equation)

Reduced Form: An equation in reduced form presents endogenous variable as a function of exogenous variables only. The reduced form of an econometric model has been rearranged algebraically so that each endogenous variable is on the left side of one equation, and only predetermined variables (exogenous variables and lagged endogenous variables) are on the right side.

Structural Form: An equation in structural form presents one endogenous variable as a function of exogenous and endogenous variables.

(Working's) Simultaneous Equations Model for Market Equilibrium

Setup: The "true" relationship between demand and supply of coffee is modeled as follows

Demand Equation: $q_i^d = \alpha_0 + \alpha_1 p_i + u_i$ (u_i represents factors that influence coffee demand other than price)

Supply Equation: $q_i^s = \beta_0 + \beta_1 p_i + v_i$ (v_i represents factors that influence coffee supply other than price)

Market Equilibrium: $q_i^d = q_i^s$

Note: We assume $E(u_i) = 0$ and $E(v_i) = 0$ (if not, include nonzero means in the intercepts)

Endogeneity: Here, the regressor p_i is **endogenous/not predetermined**, i.e. not orthogonal to the (contemporaneous) error term, and therefore does not satisfy the orthogonality condition that

$E(p_i \cdot u_i) = 0 \Leftrightarrow \text{cov}(p_i,u_i) = 0$ and $E(p_i \cdot v_i) = 0 \Leftrightarrow \text{cov}(p_i,v_i) = 0$

The endogeneity in this example arises from the face that price is a function of both error terms u_i and v_i, which is a result of market equilibrium.

To see endogeneity, treat the 3 equations as a system of simultaneous equations and solve for p_i and q_i

$$q_i^d = q_i^s \Rightarrow \alpha_0 + \alpha_1 p_i + u_i = \beta_0 + \beta_1 p_i + v_i \Rightarrow (\alpha_1 - \beta_1)p_i = (\beta_0 - \alpha_0) + (v_i - u_i) \Rightarrow p_i = \frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}$$

$$\text{So, } Cov(p_i,u_i) = Cov\left(\frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}, u_i\right) = \frac{1}{(\alpha_1 - \beta_1)} Cov((v_i - u_i), u_i) = \frac{1}{(\alpha_1 - \beta_1)}(Cov(v_i,u_i) - Var(u_i)) = \frac{-Var(u_i)}{(\alpha_1 - \beta_1)} \quad (\text{Since } Cov(v_i,u_i) = 0 \text{ by assumption})$$

$$Cov(p_i,v_i) = Cov\left(\frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}, v_i\right) = \frac{1}{(\alpha_1 - \beta_1)} Cov((v_i - u_i), v_i) = \frac{1}{(\alpha_1 - \beta_1)}(Var(v_i) - Cov(v_i,u_i)) = \frac{Var(v_i)}{(\alpha_1 - \beta_1)} \quad (\text{Since } Cov(v_i,u_i) = 0 \text{ by assumption})$$

Therefore, $\text{cov}(p_i,u_i) = 0$ and $\text{cov}(p_i,v_i) = 0$ iff $\text{Var}(u_i) = 0$ and $\text{Var}(v_i) = 0$ respectively.

Not possible (except in the extreme case when, for example, there are no other factors that shift demand, so $u_i = 0$)!

What is the Endogeneity Bias?

When we regress observed quantity on a constant and price, we neither estimate the demand nor supply curve because price is endogenous in both equations.

Recall that the **OLS estimator is consistent for the least squares projection coefficients**: in this case, the least squares projection of (true) q_i on a constant and (true) p_i gives a coefficient of p_i given by $\text{Cov}(p_i,q_i)/\text{Var}(p_i)$

Suppose we observe $\{q_i, p_i\}$ and **we regress q_i on a constant and p_i, what is it that we estimate?**

OLS estimate of the price coefficient $\hat{\alpha}_1$ (from the demand equation) is consistent for:

$$\xrightarrow{P} \frac{Cov(p_i,q_i)}{Var(p_i)} = \frac{Cov(p_i,\alpha_0 + \alpha_1 p_i + u_i)}{Var(p_i)} = \frac{Cov(p_i,\alpha_1 p_i + u_i)}{Var(p_i)} = \frac{\alpha_1 Var(p_i) + Cov(u_i,p_i)}{Var(p_i)} = \alpha_1 + \frac{Cov(u_i,p_i)}{Var(p_i)}$$

$$\text{Asymptotic Bias} = \frac{Cov(u_i, p_i)}{Var(p_i)}$$

OLS estimate of the price coefficient $\hat{\beta}_1$ (from the supply equation)

$$\xrightarrow{P} \frac{Cov(p_i, q_i)}{Var(p_i)} = \frac{Cov(p_i, \beta_0 + \beta_1 p_i + u_i)}{Var(p_i)} = \frac{Cov(p_i, \beta_1 p_i + v_i)}{Var(p_i)} = \frac{\beta_1 Var(p_i) + Cov(v_i, p_i)}{Var(p_i)} = \beta_1 + \frac{Cov(v_i, p_i)}{Var(p_i)}$$

$$\text{Asymptotic Bias} = \frac{Cov(v_i, p_i)}{Var(p_i)}$$

Since $Cov(p_i, u_i) \neq 0$ *and* $Cov(p_i, v_i) \neq 0$, therefore **endogeneity bias/simultaneous equation bias/simultaneity bias** exists! (bc regressor and error term are related to each other through a system of simultaneous equations).
So, **OLS estimator is not consistent for either α_1 or β_1.**

Solution: Instrumental Variables and 2 Stage Least Squares

The reason why demand curve nor supply curve can be consistently estimated because we cannot infer from the data whether the observed changes in price and quantity is due to a shift in demand or supply. Therefore, we might be able to estimate the demand/supply if some of the factors that shift the supply/demand curves are observable.

Def: A predetermined variable (predetermined in the system) that is correlated with the endogenous regressor is called an **instrumental variable** or **instrument**. Sometimes we call it a **valid instrument** to emphasize that the correlation with the endogenous regressor is not 0.

Observable Exogenous Supply Shifters:
Given "appropriate" observable supply shifters (Instrument), we can estimate demand and supply!
Suppose the supply shifter v_i can be divided into an observable factor x_i and an unobservable factor ξ_1 with Cov($x_i.\xi_1$)=0[2]
→ Supply Equation: $q_i^s = \beta_0 + \beta_1 p_i + \beta_2 x_i + \xi_i$

Suppose further that the observed supply shifter x_i is predetermined in the demand equation, i.e. uncorrelated with the error term u_i (e.g. think of x_i is the temperature in coffee growing regions). **If the temperature (x_i) is uncorrelated with the unobserved factors that shift demand (u_i), i.e. temperature (x_i) is an instrument (for the demand equation), it would be possible to extract from observed price movements a component that is related to the temperature (i.e. the observed supply shifter) but uncorrelated with the demand shifter. Then, we can estimate the demand curve by examining the relationship between coffee consumption and that component of price.**

[2] This decomposition is always possible by the projection theorem. v_i can be expressed as the projection onto the space spanned by x_i and the orthogonal complement (remember, v_i includes all factors that affect supply, so by definition has at least as many dimensions than x_i). i.e. If the least squares projection of v_i on a constant and x_i is E*(v_i | 1 x_i) = $\gamma_0 + \beta_2 x_i$. Define $\xi_i = v_i - \gamma_0 + \beta_2 x_i$. By definition, ξ_i is orthogonal to x_i and E(ξ_i) = 0, therefore ξ_i, x_i uncorrelated. Substituting this into the original supply equation, and combining the intercept terms we get the resulting expression.

XVII. Time Series: Empirical Techniques and Concepts

a. Autocorrelation in (observed) Errors (See section XVI)

From Hayashi: Our assumption (2.5) that $\{g_i\}$ is MDS → no serial correlation in $g_i = x_i \cdot \varepsilon_i$. So, while we COULD have autocorrelated errors (provided that x_i does not contain constant term) in this model, and still have the MDS assumptions hold. Since we don't observe true errors, however, we cannot know for sure. Therefore, **observing autocorrelated sample errors is enough evidence that the MDS assumption MAY not hold in reality**!

If x_i contains a constant term (as almost always), however, MDS assumption implies that the (true) errors are not correlated.

$$E(g_t \mid g_{t-1}, g_{t-2}, \ldots) = 0 \Rightarrow E(x_t \varepsilon_t \mid g_{t-1}, g_{t-2}, \ldots) = 0$$

$$If\ \ x_t = (1, z_t)',\ \ for\ example \Rightarrow E(\varepsilon_t \mid g_{t-1}, g_{t-2}, \ldots) = 0 \Rightarrow E(\varepsilon_t \varepsilon_{t-j}) = E(\varepsilon_{t-j} E(\varepsilon_t \mid g_{t-1}, g_{t-2}, \ldots)) = 0$$

So if we observe autocorrelated sample errors, again this is evidence that MDS assumption (2.5) does not hold, and **we cannot invoke the usual CLT** on **x(t)e(t)!**

Testing for Serial Correlation in Errors

- Box Pierce Q
- Ljung-Box Q

If we can estimate the form of the autocorrelation consistently, then we can use a FGLS correction (e.g. Prais-Whinston correction). But if we go ahead with OLS anyways, then we have to use a different standard error estimate (VARHAC).

b. Durbin's Method: We can get rid of AR autocorrelated errors by ρ-transforming the system

Suppoe we want to estimate time series: $y(t) = x(t)'\beta + e(t)\ and\ e(t) = \rho e(t-1) + u(t)$

ρ-transformation:

AR(1) Example

Use:

1) $\triangle C_t = \beta_1 + \beta_2 \triangle GDP_t + U_t$

2) $U_t = \rho_1 U_{t-1} + \rho_2 U_{t-1} + \varepsilon_t$

3) $\triangle C_{t-1} = \beta_1 + \beta_2 \triangle GDP_{t-1} + U_{t-1}$

4) $\triangle C_{t-2} = \beta_1 + \beta_2 \triangle GDP_{t-2} + U_{t-2}$

We Get:

$$\begin{aligned}\triangle C_t &= \beta_1 + \beta_2 \triangle GDP_t + \rho U_{t-1} + \varepsilon_t = \beta_1 + \beta_2 \triangle GDP_t + \rho_1 U_{t-1} + \rho_2 U_{t-2} + \varepsilon_t \\ &= \beta_1 + \beta_2 \triangle GDP_t + \rho_1 (\triangle C_{t-1} - \beta_1 - \beta_2 \triangle GDP_{t-1}) + \rho_2 (\triangle C_{t-2} - \beta_1 - \beta_2 \triangle GDP_{t-2}) + \varepsilon_t \\ &= \beta_1 (1 - \rho_1 - \rho_2) + \beta_2 \triangle GDP_t - \beta_2 \rho_1 \triangle GDP_{t-1} - \beta_2 \rho_2 \triangle GDP_{t-2} + \rho_1 \triangle C_{t-1} + \rho_2 \triangle C_{t-2} + \varepsilon_t\end{aligned}$$

Use:

1) $y(t) = x(t)'\beta + e(t)$

2) $e(t) = \rho e(t-1) + U(t) \qquad U(t)\ iid$

3) $y(t-1) = x(t-1)'\beta + e(t-1)$

We Get:

From 1)+2): $y(t) = x(t)'\beta + \rho e(t-1) + U(t)$

From 3) : $y(t) = x(t)'\beta + \rho\left[y(t-1) - x(t-1)'\beta\right] + U(t)$

We Estimate the ρ *– transformed system*:

$$\boxed{y(t) = \rho y(t-1) + x(t)'\beta - x(t-1)'\beta\rho + U(t)}$$

AR(2) Example

Use:

1) $\triangle C_t = \beta_1 + \beta_2 \triangle GDP_t + U_t$

2) $U_t = \rho_1 U_{t-1} + \rho_2 U_{t-1} + \varepsilon_t$

3) $\triangle C_{t-1} = \beta_1 + \beta_2 \triangle GDP_{t-1} + U_{t-1}$

4) $\triangle C_{t-2} = \beta_1 + \beta_2 \triangle GDP_{t-2} + U_{t-2}$

We Get:

$$\begin{aligned}\triangle C_t &= \beta_1 + \beta_2 \triangle GDP_t + \rho U_{t-1} + \varepsilon_t = \beta_1 + \beta_2 \triangle GDP_t + \rho_1 U_{t-1} + \rho_2 U_{t-2} + \varepsilon_t \\ &= \beta_1 + \beta_2 \triangle GDP_t + \rho_1(\triangle C_{t-1} - \beta_1 - \beta_2 \triangle GDP_{t-1}) + \rho_2(\triangle C_{t-2} - \beta_1 - \beta_2 \triangle GDP_{t-2}) + \varepsilon_t \\ &= \beta_1(1 - \rho_1 - \rho_2) + \beta_2 \triangle GDP_t - \beta_2\rho_1 \triangle GDP_{t-1} - \beta_2\rho_2 \triangle GDP_{t-2} + \rho_1 \triangle C_{t-1} + \rho_2 \triangle C_{t-2} + \varepsilon_t\end{aligned}$$

c. Durbin-Watson Statistic: Testing Serial Correlation in Error Term

- $\dfrac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2}$

- **Why is "2" a Critical Number?**

  We can find the plim as follows: $\dfrac{\frac{1}{T}\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\frac{1}{T}\sum_{i=2}^{n} e_i^2} = \dfrac{\frac{1}{T}\sum_{i=2}^{n} e_i^2 - 2\frac{1}{T}\sum_{i=2}^{n} e_i e_{i-1} + \frac{1}{T}\sum_{i=2}^{n} e_{i-1}^2}{\frac{1}{T}\sum_{i=2}^{n} e_i^2} \to_P \dfrac{2\sigma^2 - 2Cov(e_i, e_{i-1})}{\sigma^2} = 2 - \dfrac{2}{\sigma^2}Cov(e_i, e_{i-1})$

  Thus, under the null, if no autocorrelation, then plim is 2! Otherwise smaller.

d. Autocovariance, Autocorrelation, Sample Autoovariance, and Sample Autocorrelation

- **τ Order Autocovariance:** $\gamma(\tau) = Cov(y_t, y_{t-j})$

- **τ Order Autocorrelation Coefficient**: $\rho(\tau) = \dfrac{\gamma(\tau)}{\gamma(0)}$

e. Sample Autocovariance of τ order: $\hat{\gamma}(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T}(y_t - \bar{y}_T)(y_{t-\tau} - \bar{y}_T)$ *where* $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$

f. Sample Autocorrelation of τ order: $\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)} = \frac{\frac{1}{T}\sum_{t=\tau+1}^{T}(y_t - \bar{y}_T)(y_{t-\tau} - \bar{y}_T)}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_T)^2}$

g. Correllelogram

A correllelogram plots values of the same autocorrelation of many orders. $\hat{\rho}(\tau) = \frac{\hat{\gamma}(\tau)}{\hat{\gamma}(0)} = \frac{\frac{1}{T}\sum_{t=\tau+1}^{T}(y_t - \bar{y}_T)(y_{t-\tau} - \bar{y}_T)}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y}_T)^2}$

h. ACF and PACF

- **ACF:** The Autocorrelation at the given lag. The ACF will vary between -1 and +1, with values near ±1 indicating stronger correlation.

$$r_m = \frac{\sum_{i=1}^{n-m}(x_i - \bar{x})(x_{i+m} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- **PACF:** The Partial Autocorrelation at the given lag. The PACF will vary between -1 and +1, with values near ±1 indicating stronger correlation. The PACF removes the effect of shorter lag autocorrelation from the correlation estimate at longer lags. This estimate is only valid to one decimal place.

$$\Phi_{mm} = \frac{r_m - \sum_{j=1}^{m-1}\Phi_{m-1,j}\, r_{m-1}}{1 - \sum_{j=1}^{m-1}\Phi_{m-1,j}\, r_j}$$

(Think of this as a multiple regression coefficient: controlling for the effect/correlation of shorter lags to y_t, what is the correlation of y_t and y_{t-j}.)

i. ARMA and relationship to ACF/PACF: How do we detect which order ARMA?
(Think of PACF having to do with AR coefficients, i.e. the multivariate regression coefficients, and ACF having to do with the univariate regression coefficients, which is easier understood when we convert each process to an MA)

- **Pure White Noise**: For a pure white noise process autocorrelations should be 0.

- **Pure AR**: A pure autoregressive process (AR(p)) will have a **ACF that is steadily declining (but does not cut off) and a pacf that cuts off to 0 at some point (IF AR(p), should cut off at p+1)**
  Why? $AR(p): y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$. Think of PACF as the multivariate regression coefficient. So we expect an AR(p) to have P significant values in the PACF. But ACF are the univariate regression coefficients, which may remain significant for a long time because it can be written as a MA(inf) (if the process satisfies the stationarity/stability condition), so any y_t will be correlated with any y_{t-j} for any j, because they will have overlapping e_t terms.

- **Pure MA**: A pure MA (MA(q)) process will have a ACF that cuts off to 0 at some point and a PACF that steadily declines (but does not cut off). (If MA(q), should cut off at q+1).
  Why? $AR(p): y_t = \mu + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \phi_q \varepsilon_{t-q}$. So, autocorrelation between any y_t and y_{t-j} should end for j= q+1. So, we expect MA(q) to have q significant values in the ACF. Furthermore, since we can write MA(q) as an AR(inf) (if satisfies regularity conditions), and therefore should be steadily declining (by absolute summability of the coefficients).

- **Stationary Process**: A stationary process will have a ACF whose ACF declines to 0 as the number of lags increases (by absolute summability of coefficients that is necessary to make the linear process a stationary process).

- **Random Walk**: If random walk, then the first difference is white noise, i.e. not autocorrelated. Thus ACF and PACF should both be 0 → no autocorrelation.

j. Estimating AR Parameters
We can estimate AR coefficients consistently using OLS. This is why AR is preferred.

k. Estimating MA Parameters
Not as clear what's going on with MA.

REMEMBER, (as long as e(t)'s are iid white noise, process is STRICTLY stationary and ergodic), and Ergodic Theorem still works here and we have a law of large numbers (as long as process is stationary). Play with limits to see what you can get to be consistent.

l. Testing for Correlated errors
Stick estimated residuals on RHS and see if it enters.
Why does this work? $y(t) = x(t)'\beta + [e(t) + \rho e(t-1)]$
$H_0 : \rho = 0,\ maintaining\ that\ x(t)\ not\ correlated\ with\ errors$
Under the null hypothesis, e(t) is not correlated, and we maintain that regressors are orthogonal to the error term (so that OLS estimates are consistent.
Thus, when I include fitted lag residual, it is consistent for the true lag residual, and since e(t-1) not correlated with regressors, does not affect the coefficients on the regressors. Thus, if the coefficient on the fitted lag-residual term enters, it means that it should have been included in the model, i.e. lag residual is a factor that determines y(t), and therefore we reject the null that the residuals are not autocorrelated!

m. Lagged Dependent Variable as Regressor with Autocorrelated Errors

If regressors include a lagged dependent variable, and we have autocorrelated errors, OLS estimator will be biased, inconsistent, and inefficient. (because the lagged dependent variable is correlated with the autocorrelated errors!)

Ex:

$y_t = x_t'\beta + \gamma y_{t-1} + \varepsilon_t \quad \varepsilon_t = u_t + \rho u_{t-1}$

$Cov(y_{t-1}, \varepsilon_t) = Cov(x_{t-1}'\beta + \gamma y_{t-2} + u_{t-1} + \rho u_{t-2}, u_t + \rho u_{t-1}) \neq 0$

n. Lagged Dependent Variable/Regressors as Instruments: When can we do it?

Depending on the type of the autocorrelation in errors, we can use lagged variables as instruments.

$MODEL: \; y_t = x_t'\beta + \delta y_{t-1} + \varepsilon_t$

Example1: Errors are MA(1) $\varepsilon_t = u_t + \theta u_{t-1}$

Then, y(t-1) endogenous, but we can use y(t-2) as an instrument since not correlated!

Example2: Errors are AR(1) $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$

Then, y(t-1) endogenous, and we CANNOT use y(t-2) as an instrument because AR(1) = MA(inf), so y(t-2) also correlated with e(t)!

XVIII. Regression and Autocorrelated Errors

a. 2 Reasons for Autocorrelation

- **Serial correlation in the error term:** We can fix by Durbin's Method and Cochrane-Orcutt.
- **Omitted variables with time components:** Can't fix with above. Need to test if we have misspecification or serial correlation in errors.

b. How it affects OLS/GMM, M.E. GMM

Standard OLS/GMM will still be consistent. However, standard errors will be underestimated/t-stat overestimated, and OLS no longer efficient.

c. Methods

- If we know the form of the autocorrelation, and we can estimate ρ consistently, then we can use a GLS estimator (e.g. Prais-Whinston correction) (271B HW#1, #5 HW#2, #4) (Note: We can also do this in a Panel set up in multiple equation systems as well.)

However, GLS requires us to know $\Omega = \text{var}(\varepsilon \mid X)$, and is typically infeasible. Futhermore, when we have to estimate $\Omega = \text{var}(\varepsilon \mid X)$, GLS may not dominate OLS!

- If we go ahead with OLS/GMM estimation, where $g_t = x_t \varepsilon_t$ is autocorrelated, then by CLT for stationary, ergodic processes with Gordin conditions (i.e. the the autocorrelation has to meet these restrictions in order for us to invoke this CLT. This replaces the usual MDS CLT):

$$\frac{1}{\sqrt{n}} \sum x_t \varepsilon_t \Rightarrow_D N\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right) \quad where \; \gamma_j = Cov(g_t, g_{t-k}) = Cov(x_t \varepsilon_t, x_{t-k} \varepsilon_{t-k})$$

or more generally,

$$\sqrt{n} \sum z_t \varepsilon_t \rightarrow_D N(0, S), \; S = \sum_{j=-\infty}^{\infty} \Gamma_j = \Gamma_0 + \sum_{j=1}^{\infty} \left(\Gamma_j + \Gamma_j'\right) \quad where \; \Gamma_j = E\left(g_t g_{t-j}'\right) \quad for \; (j = 0, \pm 1, \pm 2, ...) = j-th \; auto\text{cov} \; matrix$$

Thus, to get a consistent estimator for the true variance/covariance matrix, when there is heteroskedasticity and autocorrelation of unknown form, we need HAC estimators such as Newey-West.

- Use panel setup: estimate equations by time.
  Be CAREFUL: Many times we fix autocorrelation problems in panel setup by stacking the equations by time. However, in the cases where we index equations by something other than time (for example, we can estimate a system of demand equations for different firms), and each equation contains observations indexed by time. Then, in that scenario, if we have autocorrelated errors within equation, we will need to use HAC estimators for M.E. estimators!

XIX. Discrete Dependent Variables, modeling $P(d = 1 \mid x(t))$: Probit, Logit, Multinomial Logit (restrictive but useful in descriptiveness)

Motivation: d is an indicator (work/no work). We model d = 0 if y(t) < 0 and d = 1 if y(t) > 0, for some variable y(t) (ex. Marginal propensity to work).

So, we model $y(t) = x(t)'\beta + \varepsilon(t)$ (e.g. model the marginal propensity to work)

Want to know: $P(d = 1 \mid x(t)) = P(y(t) > 0 \mid x(t)) = P(\varepsilon(t) > -x(t)'\beta) = 1 - P(\varepsilon(t) \leq -x(t)'\beta) = 1 - F(-x(t)'\beta)$ **where F(.) is the CDF of the error term**

3 standard distributional assumptions about the error term

a. Linear Probability Model

Linear probility assumes: F(.) = Uniform(-.5,.5)

Thus, $P(\delta = 1 \mid x(t)) = 1 - (.5 - x(t)'\beta) = .5 + x(t)'\beta$

Note: People don't use this because it can give us probabilities outside [0,1]. We COULD fix this by constrained maximization; but it's easier to probit or logit.

b. Probit Model: (d is Bernoulli(p)): $P(\delta(t) = 1 \mid x(t), \beta, \theta) = \Phi(x_t'\beta)$, so model is $\delta(t) = \Phi(x(t)'\beta) + v(t)$

Probit assumes: F(.) = Φ(.) (standard normal)

Thus, $P(\delta(t) = 1 \mid x(t)) = 1 - \Phi(-x(t)'\beta) = \Phi(x(t)'\beta)$ *by symmetry of std normal* $= \int_{-\infty}^{x(t)'\beta} \phi(x)\, dx$

ii. Estimation of Probit:

- (Typically) β estimated by **MLE**:

$$L(\delta(t) \mid x(t), \beta, \text{parameters of } f) \equiv L(\delta(t)) = \left[P(\delta(t) = 1 \mid x(t), \beta, \theta)\right]^{\delta(t)} \left[P(\delta(t) = 0 \mid x(t), \beta, \theta)\right]^{1-\delta(t)} = \left(\Phi(x(t)'\beta)\right)^{\delta(t)} \left(1 - \Phi(x(t)'\beta)\right)^{1-\delta(t)}$$

$$\hat{\beta} = \arg\max Q_T(\beta) = \arg\max \frac{1}{T} \sum_t LogL = \arg\max \frac{1}{T} \sum_t \delta(t) \log\left[\Phi(x(t)'\beta)\right] + \left(1 - \delta(t)\right) \log\left[1 - \Phi(x(t)'\beta)\right]$$

$$FOC: L_T = \frac{1}{T} \sum \delta(t) x(t) \frac{\phi(x(t)'\beta)}{\Phi(x(t)'\beta)} + \left(1 - \delta(t)\right) x(t) \frac{-\phi(x(t)'\beta)}{1 - \Phi(x(t)'\beta)}$$

by Fund.Th.of .Calc: $\int_{x=.}^{x=f(b)} g(x)\, dx = \frac{\partial f(b)}{\partial x} g(f(b))$

- **(Equivalently) β estimated by NLS (with robust SE)**:

  We can write regression model as: $\delta(t) = E(\delta(t) \mid x(t), \beta, \theta) + v(t) = P(\delta(t) = 1 \mid x(t), \beta, \theta) + v(t) = \Phi(x(t)'\beta) + v(t)$

  Once we write it in this form, we can estimate by NLS. If we weight observations by the variance of v(t), i.e. Φ(1−Φ), then we obtain MLE estimate.

Why?

Claim: $v(t)$ *conditionally heteroskedastic.*

$$Var(v_t \mid x_t) = E(v_t^2 \mid x_t) = E\left(\left[\delta_t - \Phi(x_t'\beta)\right]^2 \mid x_t\right) = E\left(\delta_t^2 - 2\delta_t\Phi(x_t'\beta) + \Phi(x_t'\beta)^2 \mid x_t\right) = E\left(\delta_t^2 \mid x_t\right) - 2E\left(\delta_t \mid x_t\right)\Phi(x_t'\beta) + \Phi(x_t'\beta)^2$$

$$= E\left(\delta_t \mid x_t\right) - 2E\left(\delta_t \mid x_t\right)\Phi(x_t'\beta) + \Phi(x_t'\beta)^2 = \Phi(x_t'\beta) - 2\Phi(x_t'\beta)^2 + \Phi(x_t'\beta)^2$$

$$= \boxed{\Phi(x_t'\beta)\left(1 - \Phi(x_t'\beta)\right)}$$

Thus, by GLS, optimally we weight by the inverse of the square-root of variance. (i.e. weighted nonlinear least squares)

- Probit with a different Cutoff? Does not matter (as long as we have a constant term in the regression model)!
  Let $\mathbf{y(t) = x(t)'\beta + \varepsilon(t) = \beta_0 + x(t)'\beta^* + \varepsilon(t)}$
  Define now I(t) > 0 if y(t) > K

$$\delta(t) = E\left(\delta(t) \mid x(t), \beta, \theta\right) + v(t) = P\left(\delta(t) = 1 \mid x(t), \beta, \theta\right) + v(t) = P\left(y(t) > K \mid x(t)\right) + v(t) = P(\varepsilon(t) > -\beta_0 - x'(t)\beta + \mathrm{K})$$
$$= P(\varepsilon(t) > -\left[\beta_0 - K\right] - x'(t)\beta^*) = 1 - \Phi\left(-\beta_0^* - x'(t)\beta^*\right) = \Phi\left(x'(t)\beta^*\right)$$

  If the model does not have a constant term, then we need to be very careful!

$$\delta(t) = E\left(\delta(t) \mid x(t), \beta, \theta\right) + v(t) = P\left(\delta(t) = 1 \mid x(t), \beta, \theta\right) + v(t) = P\left(y(t) > K \mid x(t)\right) + v(t) = P(\varepsilon(t) > -x'(t)\beta + \mathrm{K})$$
$$= 1 - \Phi\left(K - x'(t)\beta\right) = \Phi\left(x'(t)\beta - K\right)$$

c. Logit Model: (d is Bernoulli(p))

Logit assumes: F(.) = logistic = $\frac{1}{1+e^{x(t)'\beta}}$ (Note: Logistic is characterized by its CDF)

Thus, $\boxed{P\left(\delta(t) = 1 \mid x(t)\right) = 1 - \frac{1}{1+e^{x(t)'\beta}} = \frac{e^{x(t)'\beta}}{1+e^{x(t)'\beta}} \Leftrightarrow \ln\left(\frac{p_i}{1-p_i}\right) = x(t)'\beta}$

Interpretation:

- log odds ratio: Note that the log odds ratio can range from 0-inf to inf. For p = .5, odds ratio = 1 (1 success to 1 failure). For p = .8, odds ratio = 4. Etc…
- 1 Unit change in β changes log odds ratio by proportionally by…

d. Multinomial Logit (d_j binomial for j=n states: this is a SUR estimation, essentially)

Setup: $d_j(t) = \begin{cases} 1 \; if \; y_j(t) \geq y_k(t) \; \forall k \\ 0 \; if \; not \end{cases}$ We model $y_j(t) = x_j(t)'\beta_j + u_j(t)$ for j = 1,…n states

Multinomial logit assumes:

1. $u_1, ..., u_m$ *iid* log *Weibull* : $F(u) = e^{-e^{-u}}$
2. n = m: # states = # equations

3. $$P(\delta(t)=1 \mid x(t)) = \frac{e^{-x_j(t)'\beta_j}}{\sum_k 1+e^{-x_k(t)'\beta_k}}$$

Note: For n = 2 and m = 1, this is a logit.

Caution: This model receives a lot of attention from economists. But to make it interpretable is difficult. The iid assumption is restrictive and unrealistic in most applications.

e. Guide for recognizing which problems:

If want parameters of Pr(d(t)=1|x(t)), then it's a probit/logit/lin probability model problem.

XX. Limited Dependent Variables Models, modeling Pr(y*,d): Sample Selection, Tobit(1-equation), and Gen. Tobit and Heckman 2-Step (2 Equations)

a. 2 Types of Problems: Truncated and Censored Dependent Variable

Motivation: Limited dependent variable models are designed to handle samples that have been truncated or censored in some way.

Censoring: A sample has been censored if no observations have been systematically excluded, but some of the information contained in them has been suppressed.
We observe the SAME values above a certain y. (e.g. for income > $100,000, we only observe $100,000)
Thus, the probability distribution of the data has an atom at the censoring point.

Example: Households with incomes in excess of $100,000 always report $100,000 in the data.

Truncation: A sample has been truncated if some observations that should have been there have been systematically excluded from the sample. (SAMPLE SELECTION). We only observe y when y > c. (i.e. we only observe a sub-population)

Example: a sample of households with incomes under $100,000 necessarily excludes all households with incomes over that level. This is not a random sample of all households. So, if the dependent variable is income, or something correlated with income, then results using the truncated sample could potentially be quite misleading.

The following models attempt to "fix" truncated, censored dependent variable problems.

b. Tobit: Model for Censoring (uses 1 equation). Makes distributional assumption on the above setup (It's like Probit in sample selection setup)

Setup: Suppose our model is $y(t) = x(t)'\beta + u(t)$, and we only observe $y^*(t) = \begin{cases} y(t) & when\ y(t) > 0 \\ 0 & when\ y(t) \le 0 \end{cases}$

That is, our censoring point is 0, so any y(t) < 0 we only observe y*(t) = 0

Then, we want to estimate the model: $y^*(t) = \begin{cases} x(t)'\beta & when\ x(t)'\beta > 0 : \delta(t) = 1 \\ 0 & when\ x(t)'\beta \le 0 : \delta(t) = 0 \end{cases}$

Problem: We can't just run OLS here because $y^*(t) = X'\beta + u_2$ and $E(u(t) \mid d=1) = E(u(t) \mid y \ge 0) = E(u(t) \mid u(t) \ge x(t)'\beta) = $ *function of* $x(t)'\beta \ne 0$!

Likelihood Function: $L\left(y^*(t), \delta(t)\right) = L\left(y^*(t) \mid \delta(t)\right) L\left(\delta(t)\right) = \left[L\left(y^*(t) \mid \delta(t)=1\right) L\left(\delta(t)=1\right)\right]^{\delta(t)} \left[L\left(y^*(t) \mid \delta(t)=0\right) L\left(\delta(t)=0\right)\right]^{1-\delta(t)} = \left[g(y(t))\right]^{\delta(t)} \left[P\left(\delta=0\right)\right]^{1-\delta(t)}$

Note: g(y(t)) is the density of y(t). $g(y^*(t) \mid \delta=1) = \frac{g(y^*(t), \delta=1)}{P(\delta=1)} = \frac{g(y(t))}{P(\delta=1)}$ *and* $g(y^*(t) \mid \delta=0) = 1$ *at* $y^*(t) = 0$

Distance Function: $Q_T = \frac{1}{T}\sum_t \ln L\left(y^*(t),\delta(t)\right)$ *where* $\ln L\left(y^*(t),\delta(t)\right) = \delta(t)\ln g(y(t)) + (1-\delta(t))\ln P(\delta(t)=0)$

Classic Tobit Assumption: $u(t) \sim N(0,\sigma^2)$

Thus,

- $P\left(d(t)=0\right) = P(y(t)\le 0) = P\left(u(t)\le -x(t)'\beta\right) = P\left(\frac{u(t)}{\sigma} \le \frac{-x(t)'\beta}{\sigma}\right) = \Phi\left(\frac{-x(t)'\beta}{\sigma}\right)$
- $L\left(y^*(t) \mid \delta(t)=1\right) = L(y(t) \mid y(t)>0) = \frac{L(y(t), y(t)>0)}{\Pr(y(t)>0)} = \frac{\frac{1}{\sigma}\phi\left(\frac{y(t)-x(t)'\beta}{\sigma}\right)}{\Pr(y(t)>0)} \Rightarrow L\left(y^*(t) \mid \delta(t)=1\right)L(\delta(t)=1) = \frac{1}{\sigma}\phi\left(\frac{y(t)-x(t)'\beta}{\sigma}\right)$

So,

$$L\left(y^*(t),\delta(t)\right) = \left[\frac{1}{\sigma}\phi\left(\frac{y(t)-x(t)'\beta}{\sigma}\right)\right]^{\delta(t)}\left[\Phi\left(\frac{-x(t)'\beta}{\sigma}\right)\right]^{1-\delta(t)}$$

$$LogLik\left(Y^* \mid \delta\right) = \sum \frac{\delta(t)}{\sigma}\phi\left(\frac{y(t)-x(t)'\beta}{\sigma}\right) + \sum\left(1-\delta(t)\right)\Phi\left(\frac{-x(t)'\beta}{\sigma}\right)$$

c. Generalized Tobit, Mills Ratio, and the Heckman 2-Step: Sample Selection (For a certain type of truncated data where we use 2 equations here)

Motivation: Truncation is frequently based not on the value of the dependent variable, but rather on the value of another variable that is correlated with it.

Example: People may choose to enter the labor force only if their market wage exceeds their reservation wage. Then a sample of people who are in the labor force will exclude those whose reservation wage exceeds their market wage → we say the sample was **selected** on the basis of the difference between market and reservation wages, and the problem that this selection causes is **sample selection bias**.

Use 2 equations to solve the sample selection problem:
Eqn 2 is the one we care about, Eqn 1 is where the selection comes from (the one we use to get inverse mills).

i. **Setup:** Prototype 2 equation system – 2 states and 2 equations

$y_1(t) = x_1(t)'\beta_1 + u_1(t)$

$y_2(t) = x_2(t)'\beta_2 + u_2(t)$

$\delta(t) = \begin{cases} 1 \text{ if } y_1 > 0 \\ 0 \text{ if } y_1 \le 0 \end{cases}$ *AND* $\delta(t) = \begin{cases} 1 & \text{if } y_2 \text{ observed} \\ 0 \text{ if } y_2 \text{ not observed} \end{cases}$

Want to estimate β_2 *accounting for the sample selection problem.*

Example: Consider women and work: d = 1 if woman works, d = 0 if woman does not work.

(1) $LnWage = \beta_0 + \beta_1 Educ + \beta_2 Age + \beta_3 AFQT + \beta_4 Race + \beta_5 Age^2 + \beta_6 Age * Educ + \beta_7 AFQT * Educ + U_W$

(2) $Weeks = \alpha_0 + \alpha_1 Educ + \alpha_2 Age + \alpha_3 AFQT + \alpha_4 Race + \alpha_5 Age^2 + \alpha_6 Age * Educ + \alpha_7 AFQT * Educ + \alpha_8 FamInc + \alpha_9 Married + U_{WK}$

We want to estimate the determinants of log weekly wages accounting for the sample selection problem, i.e. we only see data on wages when people work! (d(t) = 1 if Wks(t) > 0, and d(t) =0 if Wks < 0.

How do we estimate β_2 consistently?

Steps: NEED TO FIND THE DISTRIBUTION of the DATA (y*(t), d = 1) to get MLE

First, want $f(u_2(t) \mid \delta(t)=1) = f(y_2(t) - x_2(t)'\beta_2 \mid \delta(t)=1)$

From Bayes' Rule,

$$f(\delta(t)=1 \mid u_2(t)) = f(x_1(t)'\beta_1 + u_1(t) > 0 \mid u_2(t)) = f(u_1(t) > -x_1(t)'\beta_1 \mid u_2(t)) = \int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t) \mid u_2(t))\ du_1$$

$$f(u_2(t) \mid \delta(t)=1) = \frac{P(\delta(t)=1 \mid u_2(t))\, f(u_2(t))}{P(\delta(t)=1)} = \frac{\int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t) \mid u_2(t))\, f(u_2(t))\ du_1}{P(\delta(t)=1)} = \frac{\int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t), u_2(t))\ du_1}{P(\delta(t)=1)}$$

Second, define observed $y_2 : y_2^*(t) = \delta(t) y_2(t)$

Then,

$$g(y_2^*(t) \mid \delta(t)=1) = g(\delta(t) y_2(t) \mid \delta(t)=1) = g(y_2(t) \mid \delta(t)=1) = g(x_2(t)'\beta_2 + u_2(t) \mid \delta(t)=1) = \frac{\int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t), y_2(t) - x_2(t)\beta_2)\ du_1}{P(u_1(t) \geq -x_1(t)\beta_1)}$$

Thus,

1. $g(y_2^*(t) \mid \delta(t)=1) = \dfrac{\int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t), y_2(t) - x_2(t)\beta_2)\ du_1}{P(u_1(t) \geq -x_1(t)\beta_1)}$
2. $g(y_2^*(t) \mid \delta(t)=0) = 1 \ \text{ at } \ y_2^* = 0$

From above, we can construct the Joint Likelihood of the Whole Sample / Data:

$$L\left(y_2^*(t), \delta_1(t)\right) = L\left(y_2^*(t) \mid \delta_1(t)\right) P(\delta_1(t)) = \left[L\left(y_2^*(t) \mid \delta_1(t)=1\right) P(\delta_1(t)=1)\right]^{\delta(t)} \left[L\left(y_2^*(t) \mid \delta_1(t)=0\right) P(\delta_1(t)=0)\right]^{1-\delta(t)}$$

$$\boxed{L\left(y_2^*(t), \delta_1(t)\right) = \left[\int_{-x_1(t)'\beta_1}^{\infty} f(u_1(t), y_2(t) - x_2(t)'\beta_2)\ du_1\right]^{\delta(t)} \left[P(\delta_1(t)=0)\right]^{1-\delta(t)}}$$

ii. **3 Estimation Methods: MLE using Generalized Tobit, Inverse Mills Ratio with NLS, Inverse Mills with Heckman 2-Step**

1. MLE Using Generalized Tobit:

Using above formulation, and assuming that $\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right) = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma \end{pmatrix}\right)$

(See Davison&Mac p.543 for full likelihood function (15.55))
(THIS IS WHAT TSP ASSUMES, correlations are between equations)

Here, S.E.'s are valid, and we can do hypothesis testing.
But Be Careful: There is no "Quasi" maximum likelihood here. If we don't believe in the distributional assumptions, then all the estimates will be off.

Note: The ML estimation will give you MLE's for coefficients in both equations AND the parameters of the model (i.e. the variance/covariance matrix)

Sample output (from empirical exercise 3)

Recall our 2 equations: (1) is the equation of interest

(1) $LnWage = \beta_0 + \beta_1 Educ + \beta_2 Age + \beta_3 AFQT + \beta_4 Race + \beta_5 Age^2 + \beta_6 Age * Educ + \beta_7 AFQT * Educ + U_W$

(2) $Weeks = \alpha_0 + \alpha_1 Educ + \alpha_2 Age + \alpha_3 AFQT + \alpha_4 Race + \alpha_5 Age^2 + \alpha_6 Age * Educ + \alpha_7 AFQT * Educ + \alpha_8 FamInc + \alpha_9 Married + U_{WK}$

```
    Probit Dependent variable: WEEKS

Regression Dependent variable: LNW

 Number of observations = 2117  Schwarz B.I.C. = 2905.25
Number of positive obs. = 1643  Log likelihood = -2825.92
Fraction of positive obs. =    0.776098
```

Parameter	Estimate	Standard Error	t-statistic		P-value	
C	**7.77797**	**5.38658**	**1.44395**		**[.149]**	**FROM EQUATION (II)**
EDUC	**1.31557**	**.409760**	**3.21059**	******	**[.001]**	
AGE	**-.555829**	**.369271**	**-1.50520**		**[.132]**	
AFQT	**.413845**	**.083689**	**4.94501**	******	**[.000]**	
RACE	**.027723**	**.088639**	**.312767**		**[.754]**	
AGE2	**.010399**	**.634250E-02**	**1.63960**		**[.101]**	
AGEEDUC	**-.042265**	**.013907**	**-3.03903**	******	**[.002]**	
AFQTEDUC	**-.131502**	**.031459**	**-4.18011**	******	**[.000]**	
FAMINC	**.918107E-05**	**.842562E-06**	**10.8966**	******	**[.000]**	
MARRIED	**-.323996**	**.049589**	**-6.53363**	******	**[.000]**	
C	.614371	3.90318	.157402		[.875]	FROM EQUATION (I)
EDUC	-.430345	.280935	-1.53183		[.126]	
AGE	.364569	.265251	1.37443		[.169]	
AFQT	-.081473	.067750	-1.20255		[.229]	
RACE	.035260	.069266	.509060		[.611]	
AGE2	-.655086E-02	.452041E-02	-1.44918		[.147]	
AGEEDUC	.018483	.964145E-02	1.91704		[.055]	
AFQTEDUC	.066805	.024255	2.75431	**	[.006]	
want to test coeff → SIGMA	**.934002**	**.018582**	**50.2627**	******	**[.000]**	**FROM VAR-COV MATRIX**
want to test coeff → RHO	**-.937001**	**.809481E-02**	**-115.753**	******	**[.000]**	

Test $H_0 : \sigma_{12} = \rho\sigma = 0$

Heckman 2 Step and Inverse Mills Ratio

We can alternatively estimate β_2 by rewriting the above as:

$$y_2(t) = x_2(t)'\beta + u_2(t) \Rightarrow E(y_2(t) \mid x_2(t), \delta = 1) = x_2(t)'\beta_2 + E(u_2(t) \mid \delta(t) = 1) = x_2(t)'\beta_2 + E(u_2(t) \mid y_1(t) > 0)$$
$$= x_2(t)'\beta_2 + E_\theta\left(u_2(t) \mid u_1(t) > -x_1(t)'\beta_1\right)$$
$$= x_2(t)'\beta_2 + b\left(x_1(t)\beta_1, \theta\right) \quad \textit{where } b \textit{ is some function of } x_1, \beta_1, \textit{and } \theta \ (\textit{parameter of distribution})$$

Note: b is a non-0 quantity as long as u_1 and u_2 are not independent of each other.

Standard assumptions about b → Inverse Mills Ratio:

Assuming $\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right) = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma \end{pmatrix}\right)$, then

$$b(x_1(t), \beta_1, \theta) = E(u_2(t) \mid \delta = 1) = \int_{-\infty}^{\infty} u_2(t) f\left(u_2(t) \mid \delta(t) = 1\right) du_2 = \sigma_{12} E\left(u_1(t) \mid u_1(t) > -x_1(t)\beta_1\right) \textit{ under bivariate normal assumption}$$
$$= \sigma_{12} \frac{\phi\left(x_1(t)\beta_1\right)}{\Phi\left(x_1(t)\beta_1\right)}$$
$$= \sigma_{12}\Lambda(x_1(t)\beta_1)$$

From the above transformation of the orginal model, now we can estimate by running a censored regression on the d = 1 sample!

$\boxed{y_2(t) = x_2(t)'\beta_2 + u_2(t) = x_2(t)\beta_2 + \sigma_{12}\Lambda(x_1(t)'\beta_1) + v_2(t) \quad \textit{where } v_2(t) = u_2(t) - \sigma_{12}\Lambda(x_1(t)'\beta_1)}$ **(notice: error term now will be heteroskedastic)**

2 estimation procedures using this setup:

2. **NLS** ($v_2(t)$ heteroskedastic) with robust S.E. : We can do hypothesis testing using this method!
   Caution: If we allow Xb to be too flexible, we might run into identification problems / multicollinearity for the Mills Ratio.

3. **Heckman 2 Step:**
   a. Run probit to estimate $\hat{\beta}_1$ and calculate $\hat{\Lambda}(x_1(t)'\beta)$
   b. Run OLS (with **robust s.e.**) on : $y_2(t) = x_2(t)\beta_2 + \sigma_{12}\hat{\Lambda}(x_1(t)'\hat{\beta}_1) + v_2(t)$ **(can be run in 2SLS too if there is endogenous variable problem)**

d. Hypothesis Testing and the Heckman 2-Step: What you can and cannot do, and Solutions.

Cannot: This gives us consistent estimates for b_2 and σ_{12}, but we cannot do hypothesis testing on b_2 because the standard errors are incorrect due to the first step estimation! (Takeaway: People don't use this in practice!)

Can: We could, however, do testing on σ_{12}: under the null, there is no sample selection problem, and $\sigma_{12} = 0$. The significance of this coefficient tells us whether the sample selection/ truncation matters (i.e. whether the truncated/selected sample is different from the population sample.

Solution: Instead, the generalized Tobit ML estimation and NLS with robust standard errors

e. Censored/Truncated Regression Problem and I.V.

One of MaCurdy's Rants: I.V. does not solve a lot of problems. It does not solve the censored regression problem because no matter what instrument you use, you cannot solve $E(u_2(t) \mid \delta = 1) \neq 0$.

Explanation: The problem is that the mean of $u_2(t)$ conditional on y(t) > c is not zero. This is because $y(t) > c \Leftrightarrow x(t)'\beta + u(t) > c \Leftrightarrow u(t) > c - x(t)'\beta$.

Since observation t will only be included in sample when u(t) > c – x(t)'b, thus, for observations that are in this sample, E(u(t)) > c – x(t)'b not 0! It is a function of the x's!

f. Inverse Mills Ratio when Cut-Off not 0

Suppose now we observe data if $y_1 > c$

$$y_1(t) = x_1(t)'\beta_1 + u_1(t)$$

$$y_2(t) = x_2(t)'\beta_2 + u_2(t)$$

$$\delta(t) = \begin{cases} 1 \text{ if } y_1 > c > 0 \\ 0 \text{ if } y_1 \le c \end{cases} \quad AND \quad \delta(t) = \begin{cases} 1 & \text{if } y_2 \text{ observed} \\ 0 \text{ if } y_2 \text{ not observed} \end{cases}$$

Want to estimate β_2 accounting for the sample selection problem.

$$y_2(t) = x_2(t)'\beta + u_2(t) \Rightarrow E(y_2(t) \mid x_2(t), \delta = 1) = x_2(t)'\beta_2 + E(u_2(t) \mid \delta(t) = 1) = x_2(t)'\beta_2 + E(u_2(t) \mid y_1(t) > c)$$

$$= x_2(t)'\beta_2 + \boxed{E_\theta\left(u_2(t) \mid u_1(t) > c - x_1(t)'\beta_1\right)}$$

$$= x_2(t)'\beta_2 + b\left(x_1(t)\beta_1, \theta\right) \quad \text{where } b \text{ is some function of } x_1, \beta_1, \text{and } \theta \text{ (parameter of distribution)}$$

Assuming $\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right)$, then

$$b(x_1(t), \beta_1, \theta) = E(u_2(t) \mid \delta = 1) = \int_{-\infty}^{\infty} u_2(t) f\left(u_2(t) \mid \delta(t) = 1\right) du_2 = \sigma_{12} E\left(u_1(t) \mid u_1(t) > c - x_1(t)\beta_1\right) \text{ under bivariate normal assumption}$$

$$= \boxed{\sigma_{12} \frac{\phi\left(x_1(t)\beta_1\right)}{\Phi\left(x_1(t)\beta_1 - c\right)}} \quad (\text{because } P(u_1(t) > c - x_1(t)\beta_1) = 1 - \Phi(c - x_1(t)\beta_1) = \Phi(x_1(t)\beta_1 - c)$$

$$= \sigma_{12}\Lambda(x_1(t)\beta_1)$$

XXI. Group Regressions / Dummy Variables / Error Variance (Hayashi 78)

Idea: We can combine regressions for different population groups in 1 using dummies

Suppose we have Blacks, Whites, Asians, Hispanics, Others.

Using blacks as reference group, we specify:

$$y = \beta_0 + \beta_1 White + \beta_2 Asian + \beta_3 Other + \beta_4 School + \beta_5 White * School + \beta_6 Asian * School + \beta_7 Other * School$$

Then, for each group:

Blacks: $y = \beta_0 + \beta_4 School$

White: $y = (\beta_0 + \beta_1) + (\beta_4 + \beta_5) School$

Asians: $y = (\beta_0 + \beta_1) + (\beta_4 + \beta_6) School$

Other: $y = (\beta_0 + \beta_3) + (\beta_4 + \beta_7) School$

XXII. Fitted Values: THINGS TO BE CAREFUL ABOUT

- **Whenever we put fitted values on RHS, standard errors are off! Cannot do valid inference because variance from first stage estimation not taken into account.**
- **No fitted values in nonlinear specifications: gives us inconsistent estimators (because the distance functions are not the same) and can't do inference (because standard errors do not account for first stage estimation)**
- **N2SLS is not done in 2 stages. Don't put fitted values in the distance function for a 1st stage. (see the n2sls portion)**
- **Including fitted value from endogeneity test gives us consistent estimates.**
- **Putting the fitted inverse mills ratio "corrects" for sample selection bias, and gives us consistent estimates.**
- **In 2SLS, we can fit values using any function of the instruments (see the exogeneity section)**

XXIII. ANOVA Table For STATA/SAS/TSP

a. Components of ANOVA: SST, SSR, SSM

$$\underbrace{\sum_{i=1}^{N}(y_i-\bar{y})^2}_{(SST)}=\underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}_{(SSR)}+\underbrace{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}_{(SSM)}$$

$$\hat{\sigma}^2=\frac{1}{N-K-1}\sum_i(y_i-\hat{y}_i)^2$$ **(Estimate used for S.E. when conditional homoskedasticity assumed/ non-robust SE)**

b. Different Outputs

STATA

```
      Source |       SS       df       MS              Number of obs =     145  (N)
-------------+------------------------------           F(  4,   140) =  437.69  (Wald Test, all coeffs = 0)
       Model |  269.514813     4  67.3787034           Prob > F      =  0.0000
    Residual |  21.5520098   140  .153942927 (σ̂²)      R-squared     =  0.9260  (SSM/SST)
-------------+------------------------------           Adj R-squared =  0.9238
       Total |  291.066823   144  2.02129738           Root MSE      =  .39236   σ̂
```

(SSR) (SST) (Sample Variance of y)

SAS

```
                          Number of Observations Read         145    (N)
                          Number of Observations Used         145    (N)


                                  Analysis of Variance

                                              Sum of           Mean
            Source                    DF       Squares         Square    F Value    Pr > F

     (SSM)  Model                       4     269.51481       67.37870     437.69    <.0001
     (SSR)  Error                     140      21.55201        0.15394     (Wald, all coeffs = 0)
     (SST)  Corrected Total           144     291.06682          (σ̂²)

                     (σ̂) Root MSE              0.39236    R-Square     0.9260 (SSM/SST)
                          Dependent Mean        1.72466    Adj R-Sq     0.9238
                          Coeff Var            22.74969
```

XXIV. Log-Specifications (semi-log/log-log): Interpreting Coefficients

Log Wage = a + b Log Age + e: $\frac{\partial \log Wage}{\partial \log Age} = \frac{\partial Wage}{Wage}\frac{Age}{\partial Age} = \frac{\partial Wage / Wage}{\partial Age / Age} = \beta \Rightarrow 100\frac{\partial Wage}{Wage} = \beta\left(100\frac{\partial Wage}{Wage}\right)$: if b = .33, then 1% increase in age→.33 % increase in wage

Wage = a + b Log Age + e: $\frac{\partial Wage}{\partial \log Age} = \frac{\partial Wage}{\partial Age}\frac{\partial Age}{\partial \log Age} = \frac{\partial Wage}{\partial Age / Age} = \beta \Rightarrow \partial Wage = \beta\left(\frac{\partial Age}{Age}\right)$: if b = .33, then 1% increase in age → increase in wage of .01*.33 = .0033

LogWage = a + b Age + e: $\frac{\partial \log Wage}{\partial Age} = \frac{\partial Wage}{Wage}\frac{1}{\partial Age} = \beta \Rightarrow \frac{\partial Wage}{Wage} = \beta(\partial Age)$: if b = .33, then 1 year increase in age → (1*.33 *100)% = 33% increase in wages

XXV. Appendix

a. How Dummy Variables Work

i. Dummies / interaction on all variables

Suppose we run a regression $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$ with dummies for 2 groups: $y_t = \beta_0 D_1 + \beta_1 x_t D_1 + \beta_2 z_t D_1 + \beta_3 D_2 + \beta_4 x_t D_2 + \beta_5 z_t D_2 + \varepsilon_t$ (*)

Then, we can run tests on the model (*) by stacking the results of the original model!

We can construct the X matrix according to the above equations as follows:

$$Y_{nx1} = \begin{pmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{pmatrix} \quad X_{nx(2k)} = \begin{pmatrix} \mathbf{X}^{(1)}_{(n/2)xk} & 0 \\ 0 & \mathbf{X}^{(2)}_{(n/2)xk} \end{pmatrix} \quad \varepsilon_{nx1} = \begin{pmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{pmatrix}$$

Running regression with dummies gives us:

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \quad \textit{with std.err} \quad \hat{\sigma}^2 \left(X'X\right)^{-1} = \frac{1}{T}\frac{SSR}{T}\begin{pmatrix} \left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1} & 0 \\ 0 & \left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1} \end{pmatrix} = \frac{1}{T}\begin{pmatrix} \frac{SSR_1 + SSR_2}{T}\left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1} & 0 \\ 0 & \frac{SSR_1 + SSR_2}{T}\left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1} \end{pmatrix}$$

$$\textit{robust std.err}\left(X'X\right)^{-1}\left(X'BX\right)\left(X'X\right)^{-1} = \begin{pmatrix} \left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1} & 0 \\ 0 & \left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1} \end{pmatrix}\begin{pmatrix} \left[\mathbf{X}^{(1)'}\mathbf{B}^{(1)}\mathbf{X}^{(1)}\right] & 0 \\ 0 & \left[\mathbf{X}^{(2)'}\mathbf{B}^{(1)}\mathbf{X}^{(2)}\right] \end{pmatrix}\begin{pmatrix} \left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1} & 0 \\ 0 & \left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1}\left[\mathbf{X}^{(1)'}\mathbf{B}^{(1)}\mathbf{X}^{(1)}\right]\left[\mathbf{X}^{(1)'}\mathbf{X}^{(1)}\right]^{-1} & 0 \\ 0 & \left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1}\left[\mathbf{X}^{(2)'}\mathbf{B}^{(1)}\mathbf{X}^{(2)}\right]\left[\mathbf{X}^{(2)'}\mathbf{X}^{(2)}\right]^{-1} \end{pmatrix}$$

$$\textit{Here, } B = \begin{bmatrix} \hat{\varepsilon}_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\varepsilon}_n^2 \end{bmatrix} = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}$$

$$SSR = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) = \begin{pmatrix} \mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1 \\ \mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2 \end{pmatrix}'\begin{pmatrix} \mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1 \\ \mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2 \end{pmatrix}$$

$$= \left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right)'\left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right) + \left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)'\left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)$$

$$= SSR_1 + SSR_2$$

Running regression separately gives us

$\hat{\beta}_1 = \left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)\prime}\mathbf{y}^{(1)}$ *and* $\hat{\beta}_2 = \left(\mathbf{X}^{(2)\prime}\mathbf{X}^{(2)}\right)^{-1}\mathbf{X}^{(2)\prime}\mathbf{y}^{(2)}$

$std.err: \hat{\sigma}_1^2\left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1} = \frac{SSR_1}{T_1}\left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1}$ *and* $\hat{\sigma}_2^2\left(\mathbf{X}^{(2)\prime}\mathbf{X}^{(2)}\right)^{-1} = \frac{SSR_2}{T_2}\left(\mathbf{X}^{(2)\prime}\mathbf{X}^{(2)}\right)^{-1}$

$Robust: \left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1}\left(\mathbf{X}^{(1)\prime}\mathbf{B}^{(1)}\mathbf{X}^{(1)}\right)\left(\mathbf{X}^{(1)\prime}\mathbf{B}^{(1)}\mathbf{X}^{(1)}\right)^{-1}$ *and* $\left(\mathbf{X}^{(2)\prime}\mathbf{X}^{(2)}\right)^{-1}\left(\mathbf{X}^{(2)\prime}\mathbf{B}^{(2)}\mathbf{X}^{(2)}\right)\left(\mathbf{X}^{(2)\prime}\mathbf{B}^{(2)}\mathbf{X}^{(2)}\right)^{-1}$

$SSR_1 = \left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right)'\left(\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\hat{\beta}_1\right)$ *and* $SSR_2 = \left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)'\left(\mathbf{y}^{(2)} - \mathbf{X}^{(2)}\hat{\beta}_2\right)$

Note: We get the same coefficient estimates and the same SSR. BUT we get different (non-robust) Standard Errors in Standard OLS output! But same robust standard errors!

This is because Standard OLS S.E. are calculated from $\frac{1}{T}\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ *where* $\hat{\sigma}^2 = \frac{SSR}{T}$ *or in software* $\frac{SSR}{T-K-1}$ (*both consistent*)

So, in the pooled regression, $\hat{\sigma}^2$ **estimated using data from ALL groups, assuming that homoskedasticity ACROSS groups. Where as in separate regressions,** $\hat{\sigma}^2$ **is estimated using JUST data from that group.**

Hypothesis Testing

Thus, there is heteroskedasticity across groups, then running the pooled regression will give wrong standard errors, and we cannot do a valid distance function test! But, we can do a Wald Test using robust standard errors!

This is <u>equivalent</u> to the specification: $y_t = \beta_0 + D_2 + \beta_1 x_t + \beta_2 x_t D_2 + \beta_3 z_t + \beta_4 z_t D_2 + \varepsilon_t$ (which gives the "incremental" effect).

ii. Dummies/ interation on a subset of variables

Suppose regression equation is: $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$

Again, 2 groups (blacks and whites)

Suppose we have dummies/interactions for the intercept and for z, xo the model becomes: $y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 z_i + \beta_4 z_i D_i + \varepsilon_i$

Again this is equivalent to $y_i = \beta_0 D_1 + \beta_1 D_2 + \beta_2 x_i + \beta_3 z_i D_1 + \beta_4 z_i D_2 + \varepsilon_i$

According to the latter, X is partitioned to be: $\begin{pmatrix} 1 & 0 & x_{11} & z_{11} & \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \vdots & z_{1J} & \\ 0 & 1 & \vdots & & z_{1K} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & x_{1N} & & z_{1N} \end{pmatrix}$. According to the former, the partition is: $\begin{pmatrix} 1 & 0 & x_{11} & z_{11} & \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \vdots & z_{1J} & \\ 1 & 1 & \vdots & z_{1K} & z_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & x_{1N} & z_{1N} & z_{1N} \end{pmatrix}$

iii. Simple example about why the 2 dummy specifications are the same: 1. $y_t = \beta_0 D_1 + \beta_1 D_2 \Leftrightarrow$ 2. $y_t = \beta_0 + \beta_1 D_2$

1. $\hat{\beta} = \begin{pmatrix} \bar{\mathbf{y}}^{(1)} \\ \bar{\mathbf{y}}^{(2)} \end{pmatrix}$ from work shown earlier… $SSR = \left(Y - X'\hat{\beta}\right)'\left(Y - X'\hat{\beta}\right) = \sum_i \left(x_i^{(1)} - \bar{X}^{(1)}\right)^2 + \sum_i \left(x_i^{(2)} - \bar{X}^{(2)}\right)^2$

$$\mathbf{X} = \begin{pmatrix} 1 & \\ \vdots & \\ \vdots & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \left(\begin{pmatrix} 1 & \cdots & \cdots & \cdots & 1 \\ & & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & \\ \vdots & \\ \vdots & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \cdots & \cdots & \cdots & 1 \\ & & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} N & N_2 \\ N_2 & N_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{all} y_i \\ \sum_{Group2} y_i \end{pmatrix} = \frac{1}{NN_2 - N_2^2} \begin{pmatrix} N_2 & -N_2 \\ -N_2 & N \end{pmatrix} \begin{pmatrix} \sum_{all} y_i \\ \sum_{Group2} y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{N_2(N-N_2)} \left[N_2 \sum_{all} y_i - N_2 \sum_{Group2} y_i \right] \\ \frac{1}{N_2(N-N_2)} \left[N \sum_{Group2} y_i - N_2 \sum_{all} y_i \right] \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{N_1}\left[\sum_{all} y_i - \sum_{Group2} y_i \right] \\ \frac{1}{N_2 N_1}\left[(N_1 + N_2) \sum_{Group2} y_i - N_2 \sum_{all} y_i \right] \end{pmatrix} = \begin{pmatrix} \frac{1}{N_1} \sum_{Group1} y_i \\ \frac{1}{N_2} \sum_{Group2} y_i + \frac{1}{N_1} \sum_{Group2} y_i - \frac{1}{N_1} \sum_{all} y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{N_1} \sum_{Group1} y_i \\ \frac{1}{N_2} \sum_{Group2} y_i - \left(\frac{1}{N_1} \sum_{all} y_i - \frac{1}{N_1} \sum_{Group2} y_i \right) \end{pmatrix} = \begin{pmatrix} \frac{1}{N_1} \sum_{Group1} y_i \\ \frac{1}{N_2} \sum_{Group2} y_i - \frac{1}{N_1} \sum_{Group1} y_i \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{y}}^{(1)} \\ \bar{\mathbf{y}}^{(2)} - \bar{\mathbf{y}}^{(1)} \end{pmatrix}$$

So, the 2 specifications are equivalent (in terms of coefficient estimates and SSR)!
Again, Standard Errors will be different!

b. Important Distributions for Hypothesis Testing

i. Sampling from the Normal Distribution: Properties of the Sample Mean and Sample Variance[3]

[3] Proof of (c):

Let $X_1,...,X_n$ *iid* $N(\mu,\sigma^2)$ *and let* $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ *and* $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$. *Then,*

a. $\overline{X}$ *and* S^2 *are independent RV*. (*Basu's Theorem*)

b. $\overline{X}\sim N(\mu,\sigma^2/n)$

c. $(n-1)S^2/\sigma^2\sim ChiSq(n-1)$

ii. Chi – Square Distribution

(*a*) *If* Z *is a* $N(0,1)$ *RV, then* $Z^2\sim ChiSq(1)$ → *Square of a standard normal RV is a chi-squared RV*

(*b*) *If* $X_1,...,X_n$ *are independent and* $X_i\sim ChiSq(p_i)$, *then* $X_1+...+X_n\sim ChiSq(p_1+...+p_n)$ → *Independent chi-squared RV's add to a chi- sq RV, and then degrees of freedom also add.*

iii. T Distribution

Def: *Let Z have a normal distribution with mean 0 and variance 1. Let V have a chi-square distribution with v degrees of freedom. Suppose that Z and V are independent. Then,*

$$T=\frac{Z}{\sqrt{V/v}}\sim t_v$$

Application to random sample from normal distribution:

What is the motivation behind T? Recall, given iid sample, from CLT, we know that $Z=\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$.

T distribution arises when the population standard deviation σ is unknown and has to be estimated from the data S_n.

Th:. *Let* X_i *iid* $N(\mu,\sigma^2)$. *Let* $\overline{X}_n=(X_1+...+X_n)/n$ *and* $S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$. *Then,* $T=\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}}\sim t_{n-1}$ [4]

T-Distribution and F Distribution:

The square of a value of t with v degrees of freedom is distributed as F with 1 and v degrees of freedom.

---

[4] If σ unknown, however, $T=\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}}\sim t_{n-1}$ bc $\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}}=\frac{(\overline{X}_n-\mu)/\sigma}{S_n/\sigma\sqrt{n}}=\frac{(\overline{X}_n-\mu)/(\sigma/\sqrt{n})}{(\sqrt{n-1}S_n/\sigma)/\sqrt{n-1}}$

$\frac{(\overline{X}_n-\mu)/(\sigma/\sqrt{n})}{(\sqrt{n-1}S_n/\sigma)/\sqrt{n-1}}$, $(\overline{X}_n-\mu)/(\sigma/\sqrt{n})\sim N(0,1)=Z$, $(\sqrt{n-1}S_n/\sigma)\sim ChiSq(n-1)$.

So, by above, $T\sim t_{n-1}$

iv. F-Ratio/Distribution

Def: *Let U ~ ChiSq(u) RV, V ~ ChiSq(v), U and V independent RV. Then,*

$$F = \frac{U/u}{V/v} \sim F_{u,v}$$

Theorems:

(*a*) *If* $X \sim F_{u,v}$, *then* $1/X \sim F_{v,u}$ → *Reciprocal of an F RV is again an F RV.*

(*b*) *If* $X \sim t_q$, *then* $X^2 \sim F_{1,q}$ *(Recall, from above def, the numerator of t RV is a Z,* $\therefore T^2 = \frac{Z^2}{V/v} \sim F_{1,v}$ *)*

(*c*) *If* $X \sim F_{u,v}$, *then* $(u/v)X/(1+(u/v)X) \sim beta(u/2, v/2)$

Application to random sample from normal distribution:

$H_0 : R\hat{\beta} = r \quad K : R\hat{\beta} \neq r$

Then, for linear Gaussian regression model, the Wald statistic has a known finite sample distribution

$$W = \frac{(R\hat{\beta} - r)' \left[R; (Z'Z)^{-1} R\right]^{-1} (R\hat{\beta} - r)/\#r}{SSR/n-2} = d = N(\#r, n-2)$$

More specifically, if parameter of interest is 2-dimentional,

$H_0 : \hat{\beta} = \tilde{\beta} \quad K : \hat{\beta} \neq \tilde{\beta}$ (i.e. R = I_2)

Then,

$$W = \frac{(\hat{\beta} - \tilde{\beta})' (Z'Z)(\hat{\beta} - \tilde{\beta})/2}{SSR} = d = F(2, n-2)$$

If not a Gaussian model, we only know the asymptotic distribution → chi-squared