

ECON 640: Final Exam Review Notes

Heteroskedasticity, Serial Correlation, Nonlinear Models, IV and GMM

Dr. Yang Liang

Contents

1 Post-OLS World: Heteroskedasticity and Serial Correlation	2
1.1 Baseline OLS setup and homoskedasticity	2
1.2 Heteroskedasticity: definition and consequences	2
1.3 Generalized Least Squares (GLS)	3
1.4 White heteroskedasticity-robust standard errors	3
1.5 Serial correlation / autocorrelation	4
1.6 Correcting for serial correlation: Newey–West and cluster-robust	4
1.6.1 Newey–West HAC standard errors	4
1.6.2 Cluster-robust standard errors	5
2 Nonlinear Models, MLE, and Logarithmic Specifications	5
2.1 Logarithms and percentage changes	5
2.2 Common log models and interpretations	5
2.3 Introduction to Maximum Likelihood Estimation (MLE)	6
2.3.1 Example: Normal mean	6
2.4 MLE for the linear regression (OLS as MLE)	7
2.5 Logit model and its MLE	7
2.6 Marginal effects in the logit model	7
3 Instrumental Variables (IV) and Generalized Method of Moments (GMM)	8
3.1 Structural vs. reduced-form models and endogeneity	8
3.2 Sources of endogeneity	8
3.3 Basic IV: assumptions and estimator	9
3.4 Potential outcomes and LATE intuition	9
3.5 IV and OLS as GMM	10
3.5.1 OLS as GMM	10
3.5.2 GLS as GMM	11
3.5.3 IV / 2SLS as GMM	11
3.6 Efficient GMM: two-step procedure and J -test	11
3.7 Summary: unifying OLS, GLS, and IV	13

1 Post-OLS World: Heteroskedasticity and Serial Correlation

1.1 Baseline OLS setup and homoskedasticity

Consider the linear regression model

$$y = X\beta + u,$$

where y is $n \times 1$, X is $n \times k$ with full column rank, β is $k \times 1$, and u is $n \times 1$.

The OLS estimator is

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y.$$

Under the classical assumptions:

- $\mathbb{E}[u | X] = 0$ (exogeneity),
- $\text{Var}(u | X) = \sigma^2 I_n$ (homoskedasticity and no correlation across observations),

we have

$$\mathbb{E}[\hat{\beta}_{\text{OLS}} | X] = \beta, \quad \text{Var}(\hat{\beta}_{\text{OLS}} | X) = \sigma^2(X'X)^{-1}.$$

1.2 Heteroskedasticity: definition and consequences

Definition. Heteroskedasticity means that the conditional variance of the error is not constant:

$$\text{Var}(u_i | X) = \sigma_i^2 \neq \sigma^2.$$

In matrix form,

$$\text{Var}(u | X) = \Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

Then OLS is still

$$\hat{\beta}_{\text{OLS}} = \beta + (X'X)^{-1}X'u,$$

so as long as $\mathbb{E}[u | X] = 0$, OLS remains *unbiased* and *consistent*. However,

$$\text{Var}(\hat{\beta}_{\text{OLS}} | X) = (X'X)^{-1}X'\Omega X(X'X)^{-1} = (X'X)^{-1} \left(\sum_{i=1}^n \sigma_i^2 x_i x_i' \right) (X'X)^{-1},$$

which differs from the homoskedastic formula $\sigma^2(X'X)^{-1}$. OLS is no longer efficient (not BLUE) and the usual homoskedastic standard errors are incorrect.

1.3 Generalized Least Squares (GLS)

Assume Ω is *known*, positive definite. The idea of GLS is to transform the model so that the transformed error has identity covariance.

Let

$$P = \Omega^{-1/2},$$

the symmetric square root of Ω^{-1} . Premultiply the model:

$$Py = PX\beta + Pu.$$

Define transformed variables:

$$y^* = Py, \quad X^* = PX, \quad u^* = Pu.$$

Then

$$y^* = X^*\beta + u^*,$$

with

$$\text{Var}(u^* | X) = P\Omega P' = I_n.$$

Applying OLS to the transformed model yields the GLS estimator:

$$\hat{\beta}_{\text{GLS}} = (X^{*\prime} X^*)^{-1} X^{*\prime} y^* = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y.$$

Its variance is

$$\text{Var}(\hat{\beta}_{\text{GLS}} | X) = (X' \Omega^{-1} X)^{-1},$$

which is more efficient than OLS whenever $\Omega \neq \sigma^2 I$.

In practice, Ω is unknown, so we use *Feasible GLS* (FGLS): estimate Ω from an initial OLS regression (using residuals) and plug in $\hat{\Omega}$.

1.4 White heteroskedasticity-robust standard errors

Starting from

$$\text{Var}(\hat{\beta}_{\text{OLS}} | X) = (X' X)^{-1} X' \Omega X (X' X)^{-1} = (X' X)^{-1} \left(\sum_{i=1}^n \sigma_i^2 x_i x_i' \right) (X' X)^{-1},$$

we replace σ_i^2 with the squared OLS residuals \hat{u}_i^2 to obtain the *White* (heteroskedasticity-consistent) variance estimator:

$$\widehat{\text{Var}}(\hat{\beta}_{\text{OLS}})_{\text{White}} = (X' X)^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 x_i x_i' \right) (X' X)^{-1}.$$

In matrix notation, define $\hat{U} = \text{diag}(\hat{u}_1^2, \dots, \hat{u}_n^2)$, then

$$\widehat{\text{Var}}(\hat{\beta}_{\text{OLS}})_{\text{White}} = (X' X)^{-1} X' \hat{U} X (X' X)^{-1}.$$

Intuition: the contribution of each observation to the variance is scaled by \hat{u}_i^2 . Observations with large residuals are treated as having more “noisy” information about the parameters.

1.5 Serial correlation / autocorrelation

In many applications (time series or panel data), the errors are correlated across time (or within clusters). A simple time series model:

$$y_t = x_t' \beta + u_t, \quad t = 1, \dots, T,$$

with AR(1) errors:

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad \varepsilon_t \text{ i.i.d.}$$

Serial correlation implies that

$$\text{Cov}(u_t, u_{t-s}) \neq 0 \quad \text{for some } s \neq 0.$$

Consequences:

- If x_t is exogenous and the regressors do not depend on u_t , OLS remains unbiased and consistent.
- However, the usual OLS variance formula assumes no correlation across errors, so the standard errors are incorrect and inference is misleading.
- OLS is inefficient compared to estimators that exploit the structure of $\text{Var}(u)$.

1.6 Correcting for serial correlation: Newey–West and cluster-robust

1.6.1 Newey–West HAC standard errors

Newey–West (1987) standard errors are *Heteroskedasticity and Autocorrelation Consistent* (HAC). They adjust the variance of $\hat{\beta}_{\text{OLS}}$ to allow for both heteroskedasticity and serial correlation up to some lag q :

$$\widehat{\text{Var}}(\hat{\beta})_{\text{NW}} = (X'X)^{-1} \left(\sum_{h=-q}^q w_h \hat{\Gamma}_h \right) (X'X)^{-1},$$

where:

- $\hat{\Gamma}_h = \sum_{t=|h|+1}^T x_t \hat{u}_t \hat{u}_{t-|h|} x_{t-|h|}'$ is the sample autocovariance of the moment $x_t \hat{u}_t$ at lag h ,
- w_h are weights (e.g., Bartlett weights $w_h = 1 - |h|/(q + 1)$) that downweight higher lags.

Special case $q = 0$ reduces to White's estimator (heteroskedasticity only).

1.6.2 Cluster-robust standard errors

In many panel-data settings, we have clusters (e.g., states, firms, individuals) indexed by $s = 1, \dots, S$, each with n_s observations. The model is

$$y_{it} = x'_{it}\beta + u_{it}, \quad i \in \text{cluster } s.$$

We allow arbitrary correlation of u_{it} within a cluster s , but assume independence across clusters. Let X_s be the stacked regressor matrix for cluster s and \hat{u}_s the stacked residual vector.

The cluster-robust variance estimator is:

$$\widehat{\text{Var}}(\hat{\beta})_{\text{cluster}} = (X'X)^{-1} \left(\sum_{s=1}^S X'_s \hat{u}_s \hat{u}'_s X_s \right) (X'X)^{-1}.$$

Comparison:

- White: allows heteroskedasticity, but assumes independence across all observations.
- Cluster-robust: allows heteroskedasticity and arbitrary dependence *within clusters*.
- Newey–West: designed for serial correlation over time in a single series (or panel with time dimension), with dependence decaying with lag up to q .

2 Nonlinear Models, MLE, and Logarithmic Specifications

2.1 Logarithms and percentage changes

Useful log rules:

$$\ln\left(\frac{1}{x}\right) = -\ln x, \quad \ln(ax) = \ln a + \ln x, \quad \ln\left(\frac{x}{a}\right) = \ln x - \ln a, \quad \ln(x^d) = d \ln x.$$

For small Δx ,

$$\ln(x + \Delta x) - \ln x \approx \frac{\Delta x}{x},$$

so a small relative change $\Delta x/x$ approximates the change in $\ln x$.

2.2 Common log models and interpretations

Case 1: Linear-log model.

$$y_i = \beta_0 + \beta_1 \ln x_i + u_i.$$

A 1% increase in x is approximately a change $\Delta \ln x \approx 0.01$, so

$$\Delta y \approx \beta_1 \Delta \ln x \approx 0.01 \beta_1.$$

Thus, a 1% increase in x is associated with an approximate change in y of $\beta_1/100$ units.

Case 2: Log-linear model.

$$\ln y_i = \beta_0 + \beta_1 x_i + u_i.$$

A one-unit increase in x implies

$$\Delta \ln y \approx \beta_1 \quad \Rightarrow \quad \frac{\Delta y}{y} \approx \beta_1,$$

so β_1 is approximately the *proportional* change in y for a one-unit increase in x . In percentage terms, a one-unit increase in x is associated with an approximate $100\beta_1\%$ change in y .

A more exact interpretation uses

$$\frac{y_2}{y_1} = e^{\beta_1} \quad \Rightarrow \quad \% \Delta y \approx 100 \cdot (e^{\beta_1} - 1).$$

Case 3: Log-log model.

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + u_i.$$

Then

$$\beta_1 = \frac{\partial \ln y}{\partial \ln x}$$

is an *elasticity*: the percentage change in y induced by a 1% change in x . For example, $\beta_1 = 0.8$ means a 1% increase in x is associated with a 0.8% increase in y .

2.3 Introduction to Maximum Likelihood Estimation (MLE)

Suppose W_i are i.i.d. with density $f(w_i; \theta)$, where θ is a parameter vector. The likelihood for a sample $\{w_i\}_{i=1}^n$ is

$$L(\theta) = \prod_{i=1}^n f(w_i; \theta),$$

and the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^n \ln f(w_i; \theta).$$

The MLE is

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \ell(\theta).$$

2.3.1 Example: Normal mean

If $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2 , the log-likelihood (up to constants) is

$$\ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.$$

Maximizing this is equivalent to minimizing $\sum(y_i - \mu)^2$, so

$$\hat{\mu} = \bar{y}.$$

2.4 MLE for the linear regression (OLS as MLE)

Consider

$$y_i = \alpha + \beta x_i + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

The joint density of y_i given x_i is

$$f(y_i | x_i; \alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right).$$

The log-likelihood (up to constants) is

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

For fixed σ^2 , maximizing ℓ with respect to (α, β) is equivalent to minimizing the sum of squared residuals. Thus, the MLE for (α, β) coincides with OLS. The MLE for σ^2 is the sample variance of residuals.

2.5 Logit model and its MLE

Let $Y_i \in \{0, 1\}$ and X_i be a regressor (vector). In the logit model,

$$P(Y_i = 1 | X_i) = p_i = \Lambda(X'_i \beta) = \frac{1}{1 + \exp(-X'_i \beta)}.$$

The Bernoulli likelihood for observation i is

$$P(Y_i = y_i | X_i) = p_i^{y_i} (1 - p_i)^{1-y_i}.$$

The sample log-likelihood is

$$\ell(\beta) = \sum_{i=1}^n [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)].$$

Differentiating,

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n X_i (y_i - p_i).$$

Setting $\partial \ell / \partial \beta = 0$ yields a nonlinear system in β ; no closed-form solution exists. Numerical optimization (e.g., Newton–Raphson, BFGS) is used to obtain $\hat{\beta}$.

2.6 Marginal effects in the logit model

For a scalar regressor x and

$$p(x) = \Lambda(\alpha + \beta x),$$

the marginal effect of x on the probability is

$$\frac{\partial p(x)}{\partial x} = \frac{\partial \Lambda(\alpha + \beta x)}{\partial (\alpha + \beta x)} \cdot \beta = \Lambda(\alpha + \beta x)(1 - \Lambda(\alpha + \beta x))\beta = p(x)(1 - p(x))\beta.$$

Key properties:

- The marginal effect depends on both β and $p(x)$. It is largest in magnitude when $p(x) = 0.5$, where $p(1 - p)$ is maximized ($= 0.25$). Then

$$\left. \frac{\partial p}{\partial x} \right|_{p=0.5} = 0.25 \beta.$$

- At extreme probabilities (close to 0 or 1), marginal effects are small.

Average marginal effects (AME). Given an estimate $\hat{\beta}$, we often compute

$$\widehat{\text{AME}} = \frac{1}{n} \sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i)\hat{\beta},$$

where $\hat{p}_i = \Lambda(X'_i \hat{\beta})$. AMEs summarize the average impact of a one-unit change in x on $P(Y = 1)$ across the sample.

Another approach is to compute the marginal effect at specific covariate values (e.g., at the sample means of X).

3 Instrumental Variables (IV) and Generalized Method of Moments (GMM)

3.1 Structural vs. reduced-form models and endogeneity

Structural models. Structural equations are derived from economic theory and have causal interpretations (e.g. demand, supply, production):

$$Q^d = \alpha_0 - \alpha_1 P + u^d, \quad Q^s = \beta_0 + \beta_1 P + u^s.$$

At equilibrium $Q^d = Q^s = Q$; price and quantity are determined simultaneously.

Reduced-form models. Reduced-form equations express endogenous variables (e.g. P, Q) as functions of exogenous variables and shocks:

$$P = \pi_0 + \pi_1 Z + v_P, \quad Q = \gamma_0 + \gamma_1 Z + v_Q,$$

where Z are exogenous instruments (e.g. cost shifters).

3.2 Sources of endogeneity

OLS requires $\mathbb{E}[X'u] = 0$. This fails when regressors are correlated with the error term. Common sources:

1. **Simultaneity.** Y and X are determined together in a system (e.g., price and quantity in supply–demand). The regressor X is correlated with the structural error.

2. **Reverse causality.** Y affects X : for example, health and income; healthier individuals may earn more, so income is correlated with the error in the health equation.
3. **Omitted variable bias.** A relevant variable affects both X and Y but is omitted from the regression (e.g., ability in a wage regression).
4. **Measurement error.** Observed regressor $X^* = X + v$ is noisy; classical measurement error biases OLS towards zero.

3.3 Basic IV: assumptions and estimator

Consider

$$Y_i = \beta X_i + u_i.$$

Suppose we observe an instrument Z_i satisfying:

- **Relevance:** $\text{Cov}(Z_i, X_i) \neq 0$,
- **Exogeneity (exclusion):** $\text{Cov}(Z_i, u_i) = 0$ (equivalently, Z_i affects Y_i only through X_i).

The IV moment condition is

$$\mathbb{E}[Z_i(Y_i - \beta X_i)] = 0.$$

Solving for β :

$$\beta_{\text{IV}} = \frac{\mathbb{E}[Z_i Y_i]}{\mathbb{E}[Z_i X_i]} = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, X_i)}.$$

If Z_i is binary, this reduces to a Wald estimator:

$$\beta_{\text{IV}} = \frac{\mathbb{E}[Y_i | Z_i = 1] - \mathbb{E}[Y_i | Z_i = 0]}{\mathbb{E}[X_i | Z_i = 1] - \mathbb{E}[X_i | Z_i = 0]}.$$

3.4 Potential outcomes and LATE intuition

Under a potential outcomes framework with a binary treatment D and binary instrument Z :

$$Y_i(1), Y_i(0), \quad D_i(1), D_i(0),$$

we classify individuals as:

- *Always-takers:* $D_i(1) = 1, D_i(0) = 1$,
- *Never-takers:* $D_i(1) = 0, D_i(0) = 0$,
- *Compliers:* $D_i(1) = 1, D_i(0) = 0$,
- *Defiers:* $D_i(1) = 0, D_i(0) = 1$ (ruled out by monotonicity).

Under:

- Independence of Z and potential outcomes,
- Exclusion (instrument affects Y only via D),
- Monotonicity (no defiers),

the IV estimand identifies the *Local Average Treatment Effect* (LATE):

$$\beta_{\text{IV}} = \mathbb{E}[Y_i(1) - Y_i(0) \mid \text{compliers}].$$

3.5 IV and OLS as GMM

GMM is built on moment conditions

$$\mathbb{E}[m(W_i, \theta_0)] = 0,$$

where $m(W_i, \theta)$ is a $q \times 1$ vector of functions and $q \geq k = \dim(\theta)$.

Define sample moments

$$g_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(W_i, \theta),$$

and a positive definite weighting matrix W_n . The GMM estimator is

$$\hat{\theta}_{\text{GMM}} = \arg \min_{\theta} g_n(\theta)' W_n g_n(\theta).$$

3.5.1 OLS as GMM

For the linear model $y_i = x_i' \beta + u_i$ with $\mathbb{E}[u_i \mid x_i] = 0$, the moment condition is

$$\mathbb{E}[x_i(y_i - x_i' \beta)] = 0.$$

Define $m_i(\beta) = x_i(y_i - x_i' \beta)$, so

$$g_n(\beta) = \frac{1}{n} X'(Y - X\beta).$$

Choosing

$$W_n = \left(\frac{X'X}{n} \right)^{-1},$$

the GMM criterion is

$$Q_n(\beta) = g_n(\beta)' W_n g_n(\beta) = (Y - X\beta)' X (X'X)^{-1} X' (Y - X\beta),$$

which is minimized at the OLS estimator

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1} X' Y.$$

3.5.2 GLS as GMM

With heteroskedastic or correlated errors and known Ω ,

$$\text{Var}(u \mid X) = \Omega, \quad \mathbb{E}[X'u] = 0,$$

the moment condition is still $\mathbb{E}[X'(Y - X\beta)] = 0$, but the optimal weight matrix (in terms of efficiency) is $W = (X'\Omega X)^{-1}$, leading to the GLS estimator

$$\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

3.5.3 IV / 2SLS as GMM

For IV, we have instruments Z and moment condition

$$\mathbb{E}[Z'(Y - X\beta)] = 0.$$

The sample moment is

$$g_n(\beta) = \frac{1}{n}Z'(Y - X\beta).$$

The GMM objective is

$$Q_n(\beta) = g_n(\beta)'Wg_n(\beta) = (Y - X\beta)'ZWZ'(Y - X\beta).$$

The first-order condition yields

$$\hat{\beta}_{\text{GMM}} = (X'ZWZ'X)^{-1}X'ZWZ'Y.$$

Special cases:

- If $W = (Z'Z/n)^{-1}$, then

$$\hat{\beta}_{\text{GMM}} = (X'P_ZX)^{-1}X'P_ZY, \quad P_Z = Z(Z'Z)^{-1}Z',$$

which is the 2SLS estimator.

- If W is chosen as the inverse of the covariance matrix of the moments,

$$\Omega_m = \mathbb{E}[m(W_i, \theta_0)m(W_i, \theta_0)'],$$

we obtain *efficient GMM*.

3.6 Efficient GMM: two-step procedure and J -test

Two-step efficient GMM.

1. **Step 1 (initial estimate).** Choose a simple weighting matrix $W^{(0)}$, e.g. $W^{(0)} = I_q$ or $(Z'Z/n)^{-1}$, and compute a preliminary estimate

$$\tilde{\theta} = \arg \min_{\theta} g_n(\theta)'W^{(0)}g_n(\theta).$$

2. **Step 2 (estimate optimal weight).** Estimate the covariance matrix of the moments at $\tilde{\theta}$:

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n m(W_i, \tilde{\theta}) m(W_i, \tilde{\theta})'.$$

Set the optimal weight

$$\hat{W} = \hat{\Omega}^{-1}.$$

3. **Step 3 (efficient estimate).** Re-estimate

$$\hat{\theta}_{\text{GMM}} = \arg \min_{\theta} g_n(\theta)' \hat{W} g_n(\theta).$$

Iterating the last two steps does not improve asymptotic efficiency beyond the two-step estimator.

Asymptotic variance. Let

$$G = \mathbb{E} \left[\frac{\partial m(W_i, \theta_0)}{\partial \theta'} \right], \quad \Omega = \mathbb{E}[m(W_i, \theta_0)m(W_i, \theta_0)'].$$

Then under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{\text{GMM}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G' W \Omega W G)^{-1} G' W \Omega W G (G' W G)^{-1}).$$

The variance is minimized when $W = \Omega^{-1}$, yielding

$$\text{Avar}(\hat{\theta}_{\text{opt}}) = (G' \Omega^{-1} G)^{-1}.$$

Over-identification test (Sargan–Hansen J -test). If the model is over-identified ($q > k$), there are more moments than parameters. The J -statistic is

$$J = n g_n(\hat{\theta}_{\text{GMM}})' \hat{W} g_n(\hat{\theta}_{\text{GMM}}).$$

Under the null that all moment conditions are correct, asymptotically

$$J \sim \chi_{q-k}^2.$$

Interpretation:

- Small J : moments are consistent with the data; instruments are jointly valid.
- Large J : reject over-identifying restrictions; some instruments or model assumptions may be invalid.

3.7 Summary: unifying OLS, GLS, and IV

All three can be viewed as GMM estimators:

$$\hat{\beta}_{\text{GMM}} = (X'Z W Z' X)^{-1} X' Z W Z' Y,$$

with different choices of Z (instruments) and W (weighting matrix):

- **OLS:** $Z = X$, $W = (X'X/n)^{-1}$; moments $\mathbb{E}[X'(Y - X\beta)] = 0$.
- **GLS:** $Z = X$, $W = \Omega^{-1}$, where Ω is the error covariance; moments same as OLS but weighted optimally for heteroskedastic/correlated errors.
- **IV / 2SLS:** Z are external instruments, $W = (Z'Z/n)^{-1}$ (2SLS) or Ω^{-1} for efficient GMM IV.

This GMM perspective emphasizes that estimation is about choosing:

1. Which moment conditions (which Z),
2. How to weight them (which W).