ECON6300/7320/8300

Advanced Microeconometrics

Review of Multiple Regression and M-estimation

Christiern Rose

University of Queensland

Lecture 2

Features of microeconometrics (1)

- Data pertain to firms, individuals, households, etc
- Focus on "outcomes", and relationships linking outcomes to actions of individuals
  - earnings = f(hours worked, **years of education**, gender, expereince, institutions)
- Heterogeneity of economic subjects' preferences, constraints, goals etc. explicitly acknowledged (no "representative agent" assumption)
- Noisy data, large samples,
- Economic factors supplemented by social, spatial, temporal interdependence

Features of microeconometrics (2)

- Sources of data:
  - Surveys (Govt/private); cross section or longitudinal (panel)
  - Census
  - Administrative data (by-product: Tax related, health related, program related)
  - Natural experiments
  - Designed experiments
  - Randomized trials with controls
- Type of data impacts method and model used in analysis

Features of microeconometrics (3)

- Measures of "outcomes"
  - Continuous (e.g. earnings)
  - Discrete (binary or multinomial chpoice as in discrete choice) or integers-valued (number of doctor visits)
  - Partially observed/censored (hours of work)
  - Proportions or intervals

- Type of measure may affect the choice of model used
- Many types of regression models

This Lecture

1. Understand the role of regression
2. Review of the basic regression analysis results in matrix notation (Main reference: W. Greene)
3. Review features of regression analysis
4. Review of the scope and limitations of regression model; consider causal parameters and treatment effects
5. Compare causal ("structural") and non-causal ("reduced form") regression models
6. Move on to the topic of m-Estimation

General set-up and notation

- Data: $(\mathbf{y} : (N \times 1), \mathbf{X} : (N \times K))$
- Regression model in matrix notation: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
- A joint unknown population distribution of data: $f(\mathbf{y}, \mathbf{X}; \theta^0)$, where both f and θ^0 are unknown
- Three approaches:
  1. Fully parametric: assume f is given, θ^0 is finite dimensional but unknown
  2. Semi-parametric: assume that θ^0 is finite dimensional but unknown we can specify some moment functions for y, e.g. $E[y|X]$, or $Var[y|X]$ and we do not want to make assumptions about the distribution $f(.)$
  3. Nonparametric: assume that θ^0 is infinite dimensional, and we want to estimate the relation between y and X without making a parametric assumption about $f(.)$

Definition and notation

- θ^0 : vector of mean and variance parameters in the relationships to be estimated
- $\widehat{\theta}$: the estimator of θ^0 based on sample of observations from the population of interest.
- In general $\widehat{\theta} \neq \theta^0$; $(\widehat{\theta} - \theta^0)$: sampling error has a statistical distribution
- Ideally the distribution of $\widehat{\theta}$ is centered on θ^0 (**unbiased estimator**) with high precision (**efficiency property**), and a known distribution, to support statistical inference (**probability statements** and **hypothesis testing**).
- Consistency means $\widehat{\theta} \xrightarrow{p} \theta^0$.

General approach to estimation and inference

- Model specification and identification
  - Which specification/restrictions are reasonable?
  - Can the parameter θ^0 be recovered given infinite data?
- Correct model specification or correct specification of key components of the model given the data we have available is **necessary** for consistency
- Qualification: All models are necessarily misspecified as they are simplifications
- True specifications vs Pseudo-true specifications

- Under additional assumptions the estimators are asymptotically normally distributed,
  - i.e. the sampling distribution is well approximated by the multivariate normal in large samples:

$$\widehat{\theta} \overset{a}{\sim} \mathcal{N}[\theta, \mathrm{V}[\widehat{\theta}]]$$

    where $\mathrm{V}[\widehat{\theta}]$ denotes the (asymptotic) variance-covariance matrix of the estimator (VCE).
- Efficient estimators are consistent and have smaller variance, and VCE ($\widehat{\mathrm{V}[\widehat{\theta}]}$).

- In many (most) cases large sample (normal) distribution of $\hat{\theta}$ is the best we can do. Hence inference on $\hat{\theta}$ is based on distributions derived from the normal
- Test statistics based on (asymptotic) normal results include t-test, F-test, chi-square test

- Standard errors of the parameter estimates are obtained from $\widehat{\mathrm{V}}[\hat{\theta}]$.
- Different assumptions about the data generating process (d.g.p.), such as heteroskedasticity, can lead to different VCE.

OLS regression

- Linear regression estimated by least squares can be regraded as semi-parametric
- Goal: to estimate the linear conditional mean function

$$\mathrm{E}\left[y|\mathbf{x}\right] = \mathbf{x}'\beta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K, \qquad (1)$$

  where usually an intercept is included so $x_1 = 1$.
- $\mathrm{E}[y|\mathbf{x}]$ is of direct interest if goal is prediction based on $\mathbf{x}'\beta$
- Econometrics interested in marginal effects (e.g. price change on quantity transacted): $\frac{\partial\, \mathrm{E}\,[y|\mathbf{x}]}{\partial x_j} = \beta_j$.
- The linear regression has two components, conditional mean and the error

$$\begin{aligned} y_i &= \mathrm{E}[\, y_i|\mathbf{x}_i] \;+\; u_i && (2)\\ y_i &= \mathbf{x}_i'\beta + \; u_i, \quad i = 1, \ldots, N. && (3) \end{aligned}$$

OLS (1)

- The objective function is the sum of squared errors, $Q_N(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{N}(y_i - \mathbf{x}_i'\beta)^2$ which is minimized w.r.t. β
- Solving FOC (first oder conditions) using calculus methods yields the OLS solution: $\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$
- Matrix notation provides a very compact way to represent estimator and variance matrix formulas that involve sums of products and cross-products.
- $\mathbf{y}$: $N \times 1$ column vector with i^{th} entry y_i, $\mathbf{X}$: $N \times K$ regressor matrix $\mathbf{X}$ to have i^{th} row $\mathbf{x}_i'$.
- Convention is that all vectors as column vectors, with transpose if row vectors are desired.

OLS (2)

- The OLS estimator can be written in matrix or mixed matrix-scalar notation:

$$
\begin{aligned}
\widehat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i \\
&= \begin{bmatrix}
\sum_{i=1}^{N} x_{1i}^2 & \sum_{i=1}^{N} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{N} x_{1i}x_{Ki} \\
\sum_{i=1}^{N} x_{2i}x_{1i} & \sum_{i=1}^{N} x_{1i}^2 & & \vdots \\
 & & \ddots & \\
\sum_{i=1}^{N} x_{Ki}x_{1i} & & \cdots & \sum_{i=1}^{N} x_{1i}^2
\end{bmatrix}^{-1} \\
&\quad \times \begin{bmatrix}
\sum_{i=1}^{N} x_{1i}y_i \\
\sum_{i=1}^{N} x_{2i}y_i \\
\vdots \\
\sum_{i=1}^{N} x_{Ki}y_i
\end{bmatrix}.
\end{aligned}
$$

Properties of OLS estimator

- Properties of any estimator depend on assumptions about the data generating process (d.g.p).
- For the linear regression model this reduces to assumptions about the regression error u_i.
- As a starting point in regression analysis it is typical to assume:

1. $\mathrm{E}[u_i|\mathbf{x}_i] = \mathbf{0}$ (exogeneity).
2. $\mathrm{E}[u_i^2|\mathbf{x}_i] = \sigma^2$ (conditional homoskedasticity).
3. $\mathrm{E}[u_i u_j|\mathbf{x}_i, \mathbf{x}_j] = \mathbf{0}, \; i \neq j$, (conditional uncorrelatedness).
4. $u \sim i.i.d.\ N[0, \sigma^2]$, (not essential for estimation but often added for simplicity)

Properties of OLS estimator (2)

- Assumption 1 is essential for consistent estimation of β, and implies that the conditional mean given in (1) is correctly specified.
- It also implies linearity and no omitted variables. Linearity in variables can be relaxed.

- Assumptions (2)-(3) determine the form of the VCE of $\widehat{\beta}$.
- Assumptions 1-3 lead to $\widehat{\beta}$ being *asymptotically* normally distributed with **default estimator** of the VCE

$$\widehat{V}_{\text{default}}[\widehat{\beta}] = s^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{4}$$

where $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\beta}$ and $s^2 = (N-K)^{-1}\sum_i \widehat{u}_i^2$.

Properties of OLS estimator (3)

- Under assumptions 1-3 (with or without 4), $\widehat{\beta}$ has asymptotic normal distribution (assuming no perfect collinearity).
- $\widehat{\beta}$ converges in probability to β and s^2 to σ^2
- Under assumptions 1-4 $\widehat{\beta}_i/se(\widehat{\beta}_i)$ are exactly *t*-distributed.
- Assumption 4 is not always assumed. If not it is common to continue to use the t-distribution for hypothesis testing (as opposed to the standard normal), hoping that it provides a better finite sample approximation.
- If assumptions 2-3 are relaxed, OLS is no longer efficient.

Heteroskedasticity-robust standard errors

- If assumption 1 holds, but 2 or 3 do not, we have heteroskedastic or dependent errors.
- Then variance estimated using the standard formula is wrong

- A heteroskedasticity-robust estimator, of the correct formula of the VCE of the OLS estimator is

$$\widehat{V}_{\text{robust}}[\widehat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{N}{N-K}\sum_i \widehat{u}_i^2 \mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}. \quad (5)$$

- Note that we now have a correction factor.

- For cross-section data the above "**robust estimator**" is widely used as the default variance matrix estimate in most applied work
- In Stata a robust estimate of the VCE is obtained using the `vce(robust)` option of the `regress` command Related better options are `vce(hc2)` and `vce(hc3)`

Objectives of econometric model

1. Data description and summary of associations between variables (including data mining)
2. Conditional prediction and policy analysis, prospective and retrospective
   – Simulation of counter-factual scenarios to address "what if" type questions
   – Analysis of interventions, both actual and hypothetical
3. Estimation of causal ("structural", "key") parameters
   – Inference about structural parameters and interdependence between endogenous variables
4. Empirical confirmation or refutation of hypotheses regarding microeconomic behavior.

When assumptions fail

- "All models are lies but they get us closer to the truth."
- A specified/assumed model is a "pseudo-true" model, our approximation to the unknown d.g.p.

- Goal: Get the best estimates of the assumed model (usually an approximation)
- Use diagnostic checks to see if the approximation can be improved

Common failures

- Omitted variable bias (OVB) is unavoidable. So how to interpret OLS estimates?
- Suppose the correct regression is $\mathbf{y} = \mathbf{x}\beta + \mathbf{z}\gamma + \mathbf{u}$ but $\mathbf{z}$ is incorrectly omitted.
  - Consequences - Modeling objectives 2-4 are affected but not necessarily 1!
  - $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. is biased as $E[\widehat{\beta}|\mathbf{X}, \mathbf{Z}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma$ where the second term measures the bias
  - $\widehat{\beta}$ suffers from confounding (i.e., its value depends on $\mathbf{Z}\gamma$) and β is not identified
  - However, $\mathbf{x}\widehat{\beta}$ may still be useful for conditional prediction

Some common misspecifications

- Potentially a very long list. The most important are:
  1. Omitted variables (unobserved factors that affect economic behavior - e.g., business confidence)
  2. Misspecified functional forms (departures from linearity)
  3. Ignoring endogenous regressors
  4. Ignoring measurement errors in regressors
  5. Ignoring violations of "classical" assumptions (heteroskedasticity, serial and cross section dependence)

Regression diagnostics and tests

Usual to apply diagnostic checks of model specification

- A standard modeling cycle has four steps:

specification $\rightarrow$ estimation $\rightarrow$ diagnostics $\rightarrow$ re-estimation

- Diagnostic checks involve testing a restricted model against a less restricted model

  - Ex. 1: fewer regressors vs. more regressors (e.g. F-tests)
  - Ex. 2: homoskedastic errors vs. heteroskedastic errors (e.g. tests of homoskedasticity)
  - Ex. 3: nonlinear regression vs. linear regression (tests of nonlinearity)

  - Ex. 4: serially independent errors vs. dependent errors (tests of serial correlation)
- Regression is almost always followed by **postregression analysis** involving diagnostics

Structural vs reduced form models

- Very highly **structured models** derived from detailed specification of: underlying economic behavior; institutional set-up, constraints and administrative information; statistical and functional form assumptions, assumptions of agent's optimizing behavior

- **Reduced form** studies which aim to uncover correlations and associations among variables
- Hybrid models that have some elements of structural models but do not necessarily assume optimizing behavior.

An example of Mincerian earnings regression

$$\ln E = \beta_0 + \beta_1 yreduc + \beta_3 age + \beta_3 occ + \mathbf{x}'\gamma + \varepsilon$$

1. Does this regression equation (with perhaps a small number of regressors) provide a good fit to the sample data? Is the fit improved by adding age^2 to the regression? [**Data description**]
2. Is the regression equation a good predictor of earnings at different ages and occupations? [**Conditional prediction**]
3. What does the regression say about the rate of return to an extra year of education? [**Structural or causal parameter**]
4. Can the regression be used to explain the sources of earnings differential between male and female workers? [**Counterfactual scenario**]

- These seemingly different objectives are connected, but may imply differences in emphasis on various aspects of modeling

Regression decomposition - an example of counterfactual analysis

- Consider the problem of explaining male-female earnings differential

$$
\begin{aligned}
Y_i^g &= \beta_0^g + \sum x_{ik}\beta_k^g + \varepsilon_i^g, \quad g = M, F \\
\widehat{\Delta} &= \bar{Y}^F - \bar{Y}^M \\
&= (\widehat{\beta}_0^F - \widehat{\beta}_0^M) + \sum_{k=1}^{K} (\widehat{\beta}_k^F - \widehat{\beta}_k^M)\bar{x}_k^F + \sum_{k=1}^{K} (\bar{x}_k^F - \bar{x}_k^M)\widehat{\beta}_k^M + R
\end{aligned}
$$

- This is counterfactual analysis as it answers the question: what if certain differentials were equalized?

m-Estimation

- We consider the very extensive topic of *m*-estimation. Almost all estimation methods used in this class are special cases of m-estimation.

- Examples: Least squares (LS); generalized least squares (GLS); generalized method of moments (GMM); maximum likelihood (ML); quantile regression (QR)

- Objective: Introduce key and useful asymptotic properties of m-estimators

Basic set-up and notation

Definitions

We define an **m-estimator** $\widehat{\theta}$ of the $q \times 1$ parameter vector θ is an estimator that maximizes an objective function that is a sum or average of N sub-functions

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} q(y_i, \mathbf{x}_i, \theta), \tag{6}$$

where $q(\cdot)$ is a scalar function, y_i is the dependent variable, $\mathbf{x}_i$ is a regressor vector (of exogenous variables) and we assume conditional independence over i

- Common properties of $q(\cdot)$ - continuity and differentiability w.r.t. θ
- m-estimation typically involves minimizing or maximizing a specified objective function defined in terms of data and unknown population parameters.

m-estimation

Definition

The estimator $\widehat{\theta}$ that is the solution to the first-order conditions $\partial Q_N(\theta)/\partial\theta|_{\widehat{\theta}} = \mathbf{0}$, or equivalently

$$\frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial q(y_i, \mathbf{x}_i, \theta)}{\partial\theta}\right|_{\widehat{\theta}} = \mathbf{0}. \tag{7}$$

is an m-estimator. It is a system of q **estimating equations** in q unknowns that does not necessarily have a closed-form solution for $\widehat{\theta}$ in terms of data $(\mathbf{y}_i, \mathbf{x}_i, i = 1, \ldots, N)$.

- The term m-estimator is interpreted as an abbreviation for **maximum-likelihood-like estimator**.
- Many econometricians define an m-estimator as optimizing over a sum of terms, as in (6).
- Other authors define an m-estimator as solutions of equations such as (7).
- Examples: MLE, GMM, OLS, NLS

Property	Algebraic formula
Objective Function	$Q_N(\theta) \; = N^{-1}\sum_i q(y_i, \mathbf{x}_i, \theta)$ is maximized wrt θ
Examples	MLE: $q_i = \ln f(y_i \vert \mathbf{x}_i, \theta)$ is the log-density NLS: $q_i = -(y_i - g(\mathbf{x}_i, \theta))^2$ is minus the squared error MM: $q_i = [(y_i - g(\mathbf{x}_i, \theta))\mathbf{x}_i'\mathbf{x}_i(\mathbf{y}_i - \mathbf{g}(\mathbf{x}_i, \theta))]$
First-order conditions	$\partial Q_N(\theta)/\partial\theta = N^{-1}\sum_{i=1}^{N} \partial q(y_i, \mathbf{x}_i, \theta)/\partial\theta\vert_{\widehat{\theta}} = \mathbf{0}.$

Example

Univariate distribution : $y_i \quad (i = 1, \ldots, N)$ is a $1/0$ binary variable generated by a Bernoulli trial with parameter π which is the target parameter of interest

Method	Objective Function	First order condition
OLS	$Q_N = \frac{1}{N} \sum_{i=1}^{N} (y_i - \pi)^2$	$\frac{1}{N} \sum_{i=1}^{N} (y_i - \pi) = 0$
ML	$Q_N = \frac{1}{N} \prod f(y_i; \pi) = \frac{1}{N} \prod \pi^{y_i} (1-\pi)^{1-y_i}$	$\frac{1}{N} \sum_{i=1}^{N} (y_i - \pi) = 0$
MM	$Q_N = \frac{1}{N} \sum_{i=1}^{N} (y_i - \pi)[\pi(1-\pi)]^{-1}(y_i - \pi)$	$\frac{1}{N} \sum_{i=1}^{N} (y_i - \pi) = 0$

Variance estimation for m-estimators

- For all m-estimators we can obtain the expression for the stochastic error of the estimator.
- We can then derive the expression for the asymptotic variance of the estimator.
- Two approaches are possible
  1. Derive the variance expression assuming that the errors are i.i.d. (restrictive)
  2. Derive the variance expression assuming that the errors are heteroskedastic or serially correlated (less restrictive).
- The second approach yields robust variance estimator relative to the i.i.d. case
- Example of least squares is given below.

Standard vs robust variance estimation

Standard version	**Robust version (two-step)**
Assume u_i are i.i.d.; V[$\mathbf{u}$|$\mathbf{X}$]=$\sigma^2\mathbf{I}_N$	Assume u_i are not i.i.d; V[$\mathbf{u}$|$\mathbf{X}$]=$\Omega \neq \sigma^2\mathbf{I}_N$
$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$	
$= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$	
$\widehat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$	
$\mathrm{V}[\widehat{\beta}\|\mathbf{X}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\|\mathbf{X}]$	$\mathrm{V}[\widehat{\beta}\|\mathbf{X}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\|\mathbf{X}]$
$\mathrm{V}[\widehat{\beta}\|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$\mathrm{V}[\widehat{\beta}\|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$
$\widehat{\sigma}^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/(N-K)$	$\widehat{(\mathbf{X}'\Omega\mathbf{X})} = \sum_{i=1}^{N} \mathbf{x}_i\widehat{u}_i^2\mathbf{x}_i'$
$\widehat{V}[\widehat{\beta}] = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$	$\widehat{V}[\widehat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}(\sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{x_i}\widehat{u}_{\mathbf{i}}^{2}\mathbf{x_i}')(\mathbf{X}'\mathbf{X})^{-1}$

LLS Properties

- Given the i.i.d. assumption and exogeneity of regressors LS estimator is unbiased (and consistent) and efficient. (Gauss-Markov Theorem.)

- The linear predictor $E[y|\mathbf{X} = \mathbf{X}_f] = \mathbf{X}_f\widehat{\beta}$ is also the optimal predictor (unbiased and efficient).
- The i.i.d. assumption is violated if errors are heteroskedastic, or serially correlated in which case,

$$V[\mathbf{u}] = \Omega \neq \sigma^2\mathbf{I}_N$$

Two possible structures for N=5

$$\Omega_{N\times N} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 \end{bmatrix}$$

$$\Omega_{N\times N} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & 0 & 0 \\ 0 & \sigma_{23} & \sigma_3^2 & \sigma_{34} & 0 \\ 0 & 0 & \sigma_{34} & \sigma_4^2 & \sigma_{45} \\ 0 & 0 & 0 & \sigma_{45} & \sigma_5^2 \end{bmatrix}$$

Properties of OLS vs. GLS

- Then OLS is consistent but GLS estimator is more efficient.
- Two alternatives are: (i) use feasible two-step GLS, or (ii) use the robustified estimator of $\widehat{V}[\widehat{\beta}]$, which requires fewer assumptions.

- The robustified variance estimator is the "sandwich estimator" which can be computed in two steps.
- The idea behind robust variance estimator can be extended to other M-estimators.

Generalized Least Squares Estimator

GLS	**FGLS**
$\widehat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{y}$	$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ consistent
$\widehat{\beta} = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}(\mathbf{X}\beta + \mathbf{u})$	Assume $\Omega = \Omega(\theta)$;
$= \beta + (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}\Omega^{-1}\mathbf{u}$	θ can be consistently estimated
$\widehat{\beta} - \beta = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}\Omega^{-1}\mathbf{u}$	given $\widehat{\beta}$
V$[\widehat{\beta}\vert\mathbf{X},\Omega]$ =E$[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'\vert\mathbf{X}, \Omega]$	$\widehat{\Omega} = \Omega(\widehat{\theta})$
V$[\widehat{\beta}\vert\mathbf{X},\Omega] = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$	V$[\widehat{\beta}\vert\mathbf{X},\widehat{\Omega}] = (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}$
$\widehat{V}[\widehat{\beta}] = (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}$	$\widehat{\beta}_{FGLS} \xrightarrow{p} \widehat{\beta}_{GLS}$ b/c $\widehat{\Omega} \xrightarrow{p} \Omega$

Why m-estimation?

- Large sample optimality of m-estimators
- Consistency and asymptotic normality

Property	**Algebraic formula**
Consistency	Is plim $Q_N(\theta)$ maximized at $\theta = \theta_0$?
Consistency (informal)	Does $E\left[\partial q(y_i, \mathbf{x}_i, \theta)/\partial\theta\big|_{\theta_0}\right] = \mathbf{0}$?
Limit Distribution	$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}]$
	$\mathbf{A}_0 = \text{ plim } N^{-1}\sum_{i=1}^{N} \partial^2 q_i(\theta)/\partial\theta\partial\theta'\big|_{\theta_0}$
	$\mathbf{B}_0 = \text{plim} N^{-1}\sum_{i=1}^{N} \partial q_i/\partial\theta \times \partial q_i/\partial\theta'\big|_{\theta_0}$.
Asymptotic Distribution	$\widehat{\theta} \overset{a}{\sim} \mathcal{N}[\theta_0, N^{-1}\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{-1}]$
	$\widehat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N} \partial^2 q_i(\theta)/\partial\theta\partial\theta'\big|_{\widehat{\theta}}$
	$\widehat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N} \partial q_i/\partial\theta \times \partial q_i/\partial\theta'\big|_{\widehat{\theta}}$